AIAI 2009

5TH IFIP CONFERENCE ON ARTIFICIAL INTELLIGENCE APPLICATIONS & INNOVATIONS

WORKSHOPS PROCEEDINGS

AIAI-2009 Workshops Organization

Workshop Chair

Nick Bassiliades

Aristotle University of Thessaloniki, Greece

Workshop on Biomedical Informatics and Intelligent Approaches in the Support of Genomic Medicine (BMIINT) *24th April 2009*

Organizers: **George Potamias, Vassilis Moustakis**

Workshop on Artificial Intelligence Approaches for Biometric Template Creation and Multibiometrics Fusion (ArtIBio) *24th April 2009*

Organizers: **Nicolas Tsapatsoulis, Bernadette Dorizzi, Anixi Antonakoudi, Constantinos Pattichis**

2nd Workshop on Artificial Intelligence Techniques in Software Engineering (AISEW 2009) *25th April 2009*

Organizers: **Ioannis Stamelos, Michalis Vazirgiannis**

Workshop on Artificial Intelligence Applications in Environmental Protection (AIAEP) *25th April 2009*

Organizers: **Mihaela Oprea, Nick Bassiliades**

April 23-25, 2009

Capsis Hotel Conference Centre

Thessaloniki, Greece

Organizing Institutions

ifip — International Federation for Information Processing

University of Macedonia

Aristotle University of Thessaloniki

Democritus University of Thrace

Information

http://delab.csd.auth.gr/aiai2009

Introduction

The ever expanding abundance of information and computing power enables researchers and users to tackle highly interesting issues, such as applications providing personalized access and interactivity to multimodal information based on user preferences and semantic concepts or human-machine interface systems utilizing information on the affective state of the user. The general focus of the AIAI conference is to provide insights on how AI can be implemented in real world applications.

During the main AIAI-2009 conference, four (4) Workshops, on various specific AI application areas, were held, with parallel sessions:

- Workshop on Biomedical Informatics and Intelligent Approaches in the Support of Genomic Medicine (BMIINT)
- Workshop on Artificial Intelligence Approaches for Biometric Template Creation and Multibiometrics Fusion (ArtIBio)
- 2nd Workshop on Artificial Intelligence Techniques in Software Engineering (AISEW 2009)
- Workshop on Artificial Intelligence Applications in Environmental Protection (AIAEP)

This volume contains papers selected for presentation at the Workshops of the 5th IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI 2009) being held from 23rd till 25th of April, in Thessaloniki, Greece. The Workshops were held on 24th (BMIINT, ArtIBio) and 25th of April (AISEW 2009, AIAEP). The IFIP AIAI 2009 conference is co-organized by the Aristotle University of Thessaloniki, by the University of Macedonia Thessaloniki and by the Democritus University of Thrace. AIAI 2009 is the official conference of the WG12.5 "Artificial Intelligence Applications" working group of IFIP TC12 the International Federation for Information Processing Technical Committee on Artificial Intelligence (AI).

It is a conference growing and maintaining high standards of quality. The purpose of the 5th IFIP AIAI Conference is to bring together researchers, engineers and practitioners interested in the technical advances and business / industrial ap-plications of intelligent systems. AIAI 2009 is not only focused in providing in-sights on how AI can be implemented in real world applications, but it also covers innovative methods, tools and ideas of AI on architectural and algorithmic level.

The response to call for workhsop proposals was satisfactory. Five workshop proposals were received; finally, four of them managed to receive a critical mass of submitted papers and make it to the AIAI-2009 conference.

I would like to express my thanks to the main AIAI-2009 Conference Organizers, General Co-Chairs Professors Ioannis Vlahavas and Max Bramer, Program Committee chair, Associate Professor L. Iliadis, and Organizing Committee chair Professor Yannis Manolopoulos, for their trust in organizing the workshops and their crucial help and advises. Most of all, special thanks are due to the organizers of the workshops, namely, George Potamias and Vassilis Moustakis, for BMIINT, Nicolas Tsapatsoulis, Bernadette Dorizzi, Anixi Antonakoudi and Constantinos Pattichis, for ArtIBio, Ioannis Stamelos and Michalis Vazirgiannis, for AISEW 2009, and Mihaela Oprea, for AIAEP.

The wide range of topics and high level of contributions guarantees a very successful set of workshops. I express my special thanks to all who have contributed to the organization and scientific contents of this set of workshops, first to the authors and reviewers of the papers, as well as the members of the Program and Organization Committees.

Nick Bassiliades
AIAI-2009 Workshop Chair

Table of Contents

Workshop on Biomedical Informatics and Intelligent Approaches in the Support of Genomic Medicine (BMIINT)

1. BMIINT Editorial 1-3
   George Potamias, Vassilis Moustakis
2. COMBIOMED: A Cooperative Thematic Research Network on COMputational BIOMEDicine in Spain 4-11
   Fernando Martin-Sanchez, Victoria Lopez-Alonso, Isabel Hermosilla-Gimeno, Guillermo Lopez-Campos, and the COMBIOMED Network
3. Homogenising access to heterogeneous biomedical data sources 12-23
   Erwin Bonsma and Jeroen Vrijnsen
4. Unification of heterogeneous data towards the prediction of oral cancer reoccurrence 24-35
   Konstantinos P. Exarchos, Yorgos Goletsis, Dimitrios I. Fotiadis
5. Building a System for Advancing Clinico-Genomic Trials on Cancer 36-47
   Stelios Sfakianakis, Norbert Graf, Alexander Hoppe, Stefan Ruping, Dennis Wegener, Lefteris Koumakis, George Zacharioudakis
6. Discovery of Genotype-to-Phenotype Associations: A Grid-enabled Scientific Workflow Setting 48-59
   Lefteris Koumakis, Stelios Sfakianakis, Vassilis Moustakis, George Potamias
7. Cross-platform integration of transcriptomics data 60-71
   Georgia Tsiliki, Marina Ioannou, Dimitris Kafetzopoulos
8. Method for relating inter-patient gene copy numbers variations with gene expression via gene influence networks 72-87
   Sylvain Blachon, Gautier Stoll, Carito Guziolowski, Andrei Zimovyev, Emanuel Barillot, Anne Siegel, Ovidiu Radulescu

9. Revealing Disease Mechanisms via Coupling Molecular Pathways Scaffolds and Microarrays: A Study on the Wilm's Tumor Disease 88-99
Alexandros Kanterakis, Vassilis Moustakis, Dimitris Kafetzopoulos, George Potamias
10. A Simple Algorithm Implementation for Pattern-Matching with Bounded Gaps in Genomic and Proteomic Sequences, on the Grid EGEE Platform, using an intuitive User Interface 100-109
K.I. Vegoudakis, K.G. Margaritis, N. Maglaveras

Workshop on Artificial Intelligence Approaches for Biometric Template Creation and Multibiometrics Fusion (ArtIBio)

11. Gaussian Mixture Model Coupled with Independent Component Analysis for Palmprint Verification 110-117
R. Raghavendra, Bernadette Dorizzi, Ashok Rao, Hemantha G Kumar
12. Support Vector Machines for Dynamic Biometric Handwriting Classification 118-125
Tobias Scheidat, Marcus Leich, Mark Alexander, Claus Vielhauer
13. Applied surveillance using biometrics on Agents infrastructures 126-133
Manolis Sardis, Vasilis Anagnostopoulos, Nikos Doulamis
14. Unsupervised Human Members Tracking Based on an Silhouette Detection and Analysis Scheme 134-141
Costas Panagiotakis, Anastasios Doulamis
15. Facial Biometric Templates and Aging: Problems and Challenges for Artificial Intelligence 142-149
Andreas Lanitis
16. POLYBIO Multibiometrics Database: Contents, description and interfacing platform 150-157
Anixi Antonakoudi, Anastasis Kounoudes, Zenonas Theodosiou
17. Palm geometry biometrics: A score-based fusion approach 158-167

Nicolas Tsapatsoulis, Constatinos Pattichis

2nd Workshop on Artificial Intelligence Techniques in Software Engineering (AISEW 2009)

18. Semantic annotation, publication and discovery of Java software components: an integrated approach 168-178
Zinon Zygkostiotis, Dimitris Dranidis, Dimitris Kourtesis
19. Quality Classifiers for Open Source Software Repositories 179-188
George Tsatsaronis, Maria Halkidi, Emmanouel Giakoumakis
20. A probabilistic Approach for Change Impact Prediction in Object-Oriented Systems..... 189-200
M.K. Abdi, H. Lounis, H. Sahraoui
21. Improving Evolutionary Test Data Generation with the Aid of Symbolic Execution 201-210
M. Papadakis, N. Malevris
22. Reliable Confidence Intervals for Software Effort Estimation 211-220
Harris Papadopoulos, Efi Papatheocharous and Andreas S. Andreou
23. Bootstrap confidence intervals for regression error characteristic curves evaluating the prediction error of software cost estimation models 221-230
Nikolaos Mittas, Lefteris Angelis
24. Approaching Software Cost Estimation Using an Entropy-Based Fuzzy k-Modes Clustering Algorithm..... 231-241
Efi Papatheocharous, Andreas S. Andreou

Workshop on Artificial Intelligence Applications in Environmental Protection (AIAEP)

25. AIAEP Introduction 242-245
Mihaela Oprea, Nick Bassiliades
26. Artificial Intelligence Applications in the Atmospheric Environment: Status and Future Trends (Invited Talk) 246-247

Kostas D. Karatzas
27. Hydrological Neural Modeling aided by Support Vector Machines 248-259
L.S. Iliadis, S.I. Spartalis
28. A Data Mining System for Estimating a Largesize Matrix for the Environmental Accounting..... 260-269
Ting Yu, Manfred Lenzen, Blanca Gallego, John Debenham
29. Abductive Reasoning in Environmental Decision Support Systems..... 270-279
Franz Wotawa, Ignasi Rodriguez-Roda, Joaquim Comas
30. Supporting Decision Making in Maritime Environmental Protection with a Knowledge-based Education and Awareness Approach 280-290
Konstantinos Kotis, Andreas Papasalouros, Nikitas Nikitakos
31. An Environmental Diagnosis Expert System..... 291-302
Mihaela Oprea, Daniel Dunea, pages
32. Autonomous Inspection of Complex Environments by Means of Semantic Techniques 303-310
M. Ziegenmeyer and K. Uhl and J.M. Zöllner, R. Dillmann
33. Use of AI Techniques for Residential Fire Detection in Wireless Sensor Networks 311-321
Majid Bahrepour, Nirvana Meratnia, Paul J. M. Havinga
34. Validation of a knowledge-based risk model for biological foaming in anaerobic digestion simulation 322-332
J. Dalmau, J. Comas, I. Rodríguez-Roda, E. Latrille, J.P. Steyer
35. Estimation of the permeability of granular soils using neuro-fuzzy system..... 333-342
A. Sezer, A.B. Göktepe, S. Altun

BMIINT Editorial

SCENE. Entering the post-genomics epoch a new challenging mission is posted: to bring innovative biomedical research findings directly to the clinic and the bedside. The mission could be accomplished by *intersecting* the clinical, biological and information sciences. Endeavour is inspired by the needs raised by genomic and personalised medicine; it targets the *in-silico* biology domain; and, it is enabled by the transition to *interdisciplinary* principles and orientation. Transition is guided by the **'-***omics*' revolution as realised by advances in transcriptomics, proteomics, metabolomics, physiomics *and* nanomedicine. The '-omics' levels extends traditional clinical data models and medical decision making in two sides: on the one side to include genotypes, and on the other to include (when appropriate) '-omics' findings - the phenotypes. Yet integration is not easy simply because at both ends (genotype, phenotype) the amount of available data is immense, and complexity of processes is high. At the up-stream (or research) level scientists may be interested in the investigation between genotype-phenotype information to form *associations* and *patterns* of disease and *susceptibility* indices. At the middle-stream (industrial level) technology and service providers are interested in embedding research advances into concrete products and services via intelligent devices. Device impact is further enhanced by advances in the '*nano'* field, as expressed by nanomaterials, nanomedicine and nanoinformatics. At the down-stream (clinical theatre) level healthcare professionals (and patients as well) look forward to apply new technology on the vision of continuously improving social welfare decision making**.**

Actors & **Director.** The primary item in the respective multidisciplinary R&D agenda is *translational research* with *Biomedical Informatics* (BMI) as the driver, and *Artificial Intelligence* (AI) called to provide the needed analytical and decision-making machinery. **BMIINT** liaisons the clinical, biology, core- and Bioinformatics fields, and provides a forum for the presentation of advances at the conjunction of BMI and AI components and procedures. It takes an interdisciplinary perspective and focuses on theory, methods, techniques, tools, systems and services, which support integration, management and intelligent analysis of respective bio-related information and data sources. Emphasis is placed on the *repositioning* of methods and techniques from other domains of application into the BMIINT frontier

Plot(s)**.** BMIINT includes nine papers, which report on R&D work with broad coverage of BMIINT context. Article by Martin-Sanchez and colleagues, reports the organizational aspects and scientific aspects in relation to the COMBIOMED: A Cooperative Thematic Research Network on COMputational BIOMEDicine in Spain. The network focuses on gene-disease associations, pharmainformatics and decision support systems at the point of care. Work reported by Bonsma and Vrijnsen; Exarchos, Goletsis and Fotiadis focus on heterogeneous data integration. Bonsma and Vrijnsen provide an enrichment of the OGSA-DAI web-services framework to access medical images and microarray data; work carried out in the context of Advanced Clinico-Genomic Trials in Cancer (ACGT) project (FP6-IST-2005-026996). Exarchos et al., integrate clinical, imaging and genomic data sources to induce reliable biomarkers for the progression of oral cancer – work relates to the NeoMark project (FP7-ICT-2007-224483). Related to the two aforementioned papers is the work presented by Sfakianakis and colleagues as well by Koumakis and colleagues. Sfakianakis and colleagues endeavour on the

daily work of clinicians and bio-statisticians, develop usability criteria and propose a front-end interface layer for a Grid based architecture able to support huge computational tasks. The work of Koumakis and colleagues takes a step backwards and captures scientific-workflow design and operation specifics with due regard to a clinical scenario that achieves the seamless integration of clinico-genetic heterogeneous data sources, and the discovery of indicative SNP-phenotype associations and predictive models. Both articles draw from the ACGT project experience while Koumakis and colleagues work relates also to the GEN2PHEN project (co-funded via the European Commission, Health theme, project number 200754). Remaining articles focus on more specific BMIINT subjects. Tsiliki and colleagues overview requirements and context of cross-platform integration; the subject has immense interest, given the multitude of microarray platforms. Blachon and colleagues as well as Kanterakis and colleagues bring into BMIINT a systems biology flavor. Blachon and colleagues work is on the Ewing sarcoma; using a comparative genomic hybridization array they present data collection and preprocessing procedures and then move on to a gene influence network to model discovery of links between gene copy number variations and expression level variations – work relates to the SITCON project (from the ANR BIOSYS-2006 program). Kanterakis and colleagues reports on a methodology that couples gene-regulatory networks and microarray gene-expression data to reveal and identify molecular paths that differentiate between different disease phenotypes (with targets to the Wilms tunor domain) – work also relates to the ACGT project. Finally, Vegoudakis and colleagues report on an interface that supports patterns matching over genomic and proteomic sequences on a Grid based system – work relates to the EGEE project, co funded by the European Commission.

Elias Siores, University of Bolton, UK
Srdjan Stankovic, University of Belgrad, YU
Ioannis Tollis, University of Crete & FORTH / Institute of Computer Science (ICS), GR
Oswaldo Trelles, University of Malaga, ES
Ioannis Tsamardinos, University of Crete & FORTH-ICS, GR
Manolis Tsiknakis, FORTH / Institute of Computer Science (ICS), GR
Michael Zervakis, Technical University of Crete, GR

The Chairs of AIAI 2009/BMIINT

Dr. George Potamias, Institute of Computer Science, BMI Lab, FORTH – Heraklion, Greece

Prof. Vassilis Moustakis, Institute of Computer Science, BMI Lab, FORTH & Technical University of Crete, Department of Production Engineering and Management, Management Systems Laboratory, Chania, Greece

Crete, April 2009

ACKNOWLEDGEMENT: The Chairs want to thank the members of the Scientific Committee for their support. In addition, organization of BMIINT is partially supported by the **ACTION-Grid** EU project (FP7 ICT 224176), as well by the **ACGT** and **GEN2PHEN** projects (as mentioned in the preceding text).

COMBIOMED: A Cooperative Thematic Research Network on COMputational BIOMEDicine in Spain

Fernando Martin-Sanchez, Victoria Lopez-Alonso, Isabel Hermosilla-Gimeno, Guillermo Lopez-Campos, and the COMBIOMED Network.

Medical Bioinformatics Department, Institute of Health "Carlos III", Madrid, Spain

fms@isciii.es

Abstract

The Cooperative Thematic Research Network on Computational Biomedicine COMBIOMED was approved in the last call for Thematic Networks in Health Research within the Spanish National Plan for Scientific Research, Development and Technological Innovation and it is funded for the period 2008-2011. The COMBIOMED Network is currently addressing various aspects that range from basic to applied research in science for the development of methods and tools to solve problems in biomedical science in the context of personalized medicine. This paper describes and analyses the organizational aspects and scientific areas in which this network has been focused (gene-disease association, pharmainformatics and decision support systems at the point of care). At the same time, COMBIOMED aims to play a central role in the education of researchers and in the training of health professionals in techniques for the processing of biomedical information.

1. Introduction

The COMBIOMED Network continues the work initiated by INBIOMED, the Cooperative Thematic Research Network on Biomedical Informatics (2004-2007) that developed a platform for the storage, integration and analysis of clinical, genetic, and epidemiological data and images focused on the investigation of complex diseases [1]. Computational biomedicine represents the interface between biomedical and computer sciences. It provides an inclusive environment for a better understanding of the biological processes that take place in each of the levels of organization of living organisms and the intricate network interactions between

them. One of the objectives of COMBIOMED is to establish contacts and to collaborate with the most relevant international initiatives in the field, such as the National Centers for Biomedical Computing [2], or the Biomedical Informatics Cores of the Clinical Science and Translational Awards [3] funded by the National Institutes of Health (NIH).

2. COMBIOMED: description and organization

Several of the 12 groups participating in COMBIOMED have previously participated and led some of the European Networks of Excellence (NoE) on biomedical informatics, bioinformatics and systems biology, such as INFOBIOMED [4] and BIOSAPIENS [5]. Previous experience in Spanish initiatives (INBIOMED [6], INB [7]) has been crucial to the creation and development of those European initiatives.

The design of the network, shown in Figure 1, consists of the following levels:

- Coordination and management.
- Computational aspects, which serve as instrumental support to the network including software and middleware, hardware, GRID, algorithms, and programming.
- Coordination of the work carried out by the research groups. This level addresses aspects such as data and text mining, clinical decision-making, electronic health records, image processing, disease simulation, and biomedical ontologies that help manage and integrate chemical, genetic, environmental, clinical and imaging data.
- Horizontal activities affecting all groups and lines of work. Particular attention is paid to the connection of the network with the scientific community and society (integrated Knowledge Management, education and training, communication, dissemination, Quality and Safety) (Figure 2).

3. COMBIOMED: scientific areas

COMBIOMED focuses on three scientific research areas: gene-disease association (Disease-omics), pharmainformatics and decision support systems at the point of care (Info-POC).

3.1 Gene disease association (Disease-omics)

The study of the molecular causes of disease and individual genetic variations allows deepening into personalized medicine [8] by developing safer and more efficient preventive, diagnostic and therapeutic solutions. The scientific community needs more advanced computational resources (functional analysis of genes and proteins in the context of genetic variability, alternative splicing…) [9], access to specific comparative genomic information (genomic data visualization) and prediction of the effects of individual mutations (SNPs) in the pathways and macromolecular complexes with the consequent implications in the associated diseases.

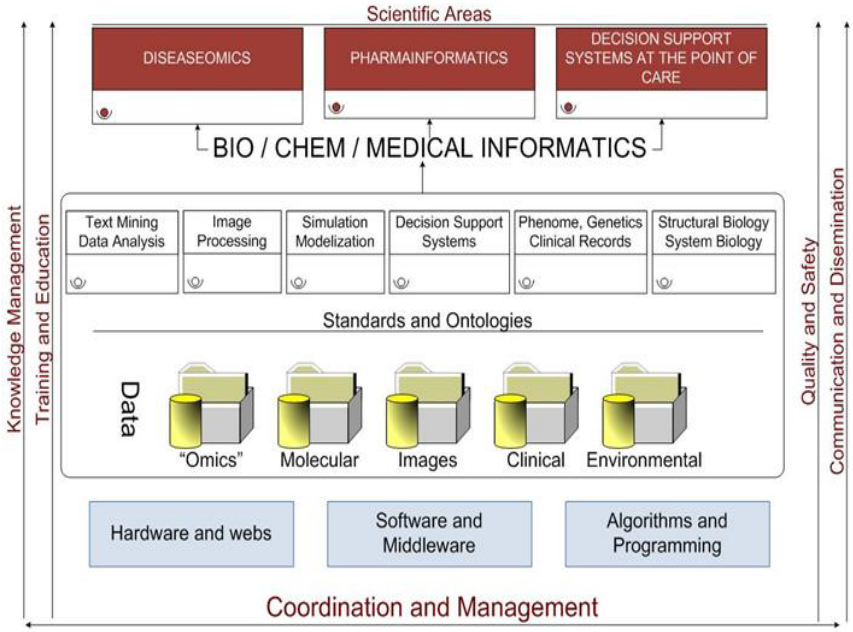


Figure 1. Graphical representation of the structure of the COMBIOMED Network

COMBIOMED works on these computational challenges in genotype-phenotype association and genomic epidemiology studies, to advance the understanding and modeling of the influence of environmental and genetic factors in the development of diseases. The network is using modules already developed by the National Institute of Bioinformatics (INB) to connect new methods that will be made available as Web services. This will help develop specific solutions for the analysis of genomic and clinical data.

The network is also developing systems that facilitate access to textual information about gene-disease relationships using automated information extrac-

tion methods and natural language processing with specific applications to problems of biomedical importance [10].

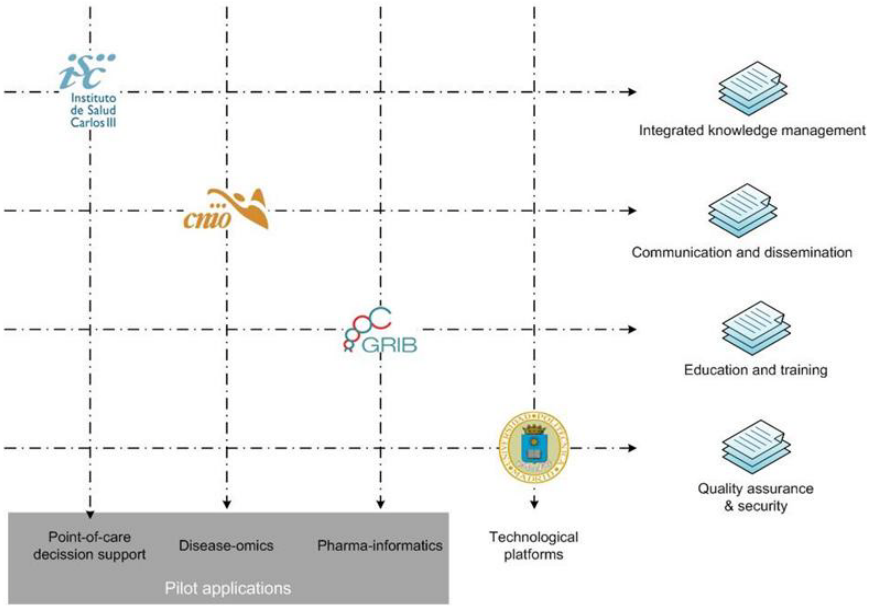


Figure 2. Organisation of scientific and horizontal activities within COMBIOMED

3.2 Pharma-informatics

The discovery and development of drugs is an area of great significance for human health, and at the same time it is an area of great socio-economic importance because it gives its "raison d'etre" to an industry which business is highly knowledge-intensive. Biomedical research in general and the R & D of drugs in particular, generate enormous amounts of data that require sophisticated computational tools for their management and analysis in order to extract the knowledge they need. This is one of the main reasons for the emergence of a new field of scientific activity that includes disciplines such as Computational Biology, and Biomedical Informatics.

Pharmaceutical research labs were pioneers in identifying the need and usefulness of computational approaches for the management and exploitation of the data generated in pre-clinical and clinical research. They are aware that certain

computational methods and their associated software can perform simulations and predictions that save time and investment in the development of drugs [11]. Computational approaches in systems biology are facilitating the management, visualization and development of predictive and descriptive mathematical models on interaction networks between biomolecular entities. This information is generated in the experimental laboratory largely based on the use of microarrays technologies [12].

Virtual screening and computer simulation techniques are very useful for the selection and testing of compounds to be considered in the initial stages of the design of a new drug. Moreover, the pharmacological and toxicological knowledge accumulated on the different groups of compounds allows for the development of quantitative models that can be used to perform *in-silico* prediction studies of the pharmacological and toxicological behavior of compounds not synthesized or tested.

Information technology also plays an important role in areas such as the management and exploitation of data from clinical trials. In addition, physiological advanced simulation techniques may allow the study of the behavior of organs of different individuals when exposed to drugs with different properties.

In coordination with the INB and the Spanish Technological Platform of Innovative Medicines [13], the COMBIOMED Network is developing technological solutions to facilitate the advancement of biomedical knowledge management geared towards the development of pharmaceutical R & D in all its stages.

3.3 Decision support systems at the point of care (INFO-POC)

In recent decades medical practice has sought greater integration of scientific knowledge in its routine. The tremendous growth of scientific knowledge and technological innovation requires the development of solutions that allow the use of a large amount of information in the clinical decision-making process. Within this context, Computational Biomedicine promotes the combination of techniques such as Medical Informatics (MI), bioinformatics (BI) and computing in the development of new methods and standards for clinical and biomolecular data integration and analysis [14]. At the same time, they facilitate a new approach that has as its overall objective to create a new integrated research framework for the development of diagnostic methods within the context of genomic medicine in the so-called "point of care".

The COMBIOMED network proposes the common research line of INFO-POC to carry out computational developments to represent and analyze clinical and biomedical knowledge at the point of patient care (POC). The collaboration between the diverse groups of the COMBIOMED network makes possible a continuous exchange of information and tools.

The network will support decision-making processes in a context of miniaturization of diagnostic systems and accessibility to information about molecular causes of diseases. This context is in line with recent trends on Convergent Technologies NBIC (Nano, Bio, Info and Cogno) with the objective of contributing to the development of a line of intelligent and miniaturized systems to be used at the point of care.

The availability and applicability of new technologies at the point of care could be a key incentive for translational research which may also imply a reduc-

Figure 3. A conceptual diagram of the design of the INFO-POC pilot.

tion in the time devoted to decision-making.

The DNA microarray technology and the Bioinformatics tools that allow microarray data storage, management and analysis have enabled the development of diagnostic tests for complex diseases [15]. In addition to the biomolecular results obtained through these miniaturized point-of-care test systems there exists the requirement of placing molecular data (i.e. mutations in a gene, sequences of DNA, proteins…) in context, through the recovery of relevant information from reference databases (in silico), and its interpretation by implementing systems to support the diagnosis process (in info). The enormous complexity of cellular processes (metabolism, signal transduction, gene expression, and so on) needs the

development of new computational models and simulations to understand their behavior overall. The recent boost of systems biology and computational cell biology reflects this fact. The design of new computer-based methods in the Semantic Web for data recovery can contribute to the representation and computational analysis of biological knowledge at the POC. The knowledge generated will be integrated into computerized protocols for the diagnosis, treatment and management of patients (Figure 3).

The combination of bioinformatics and biomedical computing tools will facilitate the development of diagnostic models, supported by new standards. These tools need to be linked by using standard medical terminologies and coding with clear semantics to facilitate the effective implementation within clinical information systems.

Conclusions

The creation of the COMBIOMED Network represents a national and international reference in biomedical computing, which aims to provide solutions to the computational challenges posed by basic and translational research, and clinical practice in the context of the new personalized medicine. The most relevant research groups in Spain are cooperating to develop methods, systems, applications and pilot projects and to yield educational recommendations to promote biomedical computing research in the next years.

More specifically, computational developments within the COMBIOMED Network allow advancing in the representation and analysis of clinical and biomolecular knowledge, and the joint research will enable the new generation of miniaturized systems to support decision making with obvious clinical applications in health at the point of care.

Acknowledgments

The Cooperative Research Network COMBIOMED is funded by the Institute of Health "Carlos III", Madrid, Spain. The leaders of the network research groups are: F. Martin, A. Valencia, R. Guigó, F. Sanz, M. Orozco, A. Pazos, G. Bueno, L. Pardo, P. Larrañaga, O. Coltell, V. Maojo, M. Robles and M.I. Loza.

References

[1] López V, et al. INBIOMED: a platform for the integration and sharing of genetic, clinical and epidemiological data oriented to biomedical research. 4th IEEE Intern. Symp. on BioInf.. and BioEng. 2004;222-226.

[2] NCRR webpage available at: http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/ Accessed on 02/26/2009.

[3] NIH BISTI webpage available at: http://www.bisti.nih.gov/ncbc/ accessed on 02/26/2009.

[4] The INFOBIOMED Network of Excellence webpage available at http:// www.infobiomed.org Accessed on 02/26/2009.

[5] BIOSAPIENS Network of Excellence webpage available at http://www.biosapiens.info Accessed on 02/26/2009.

[6] Cooperative Thematic Research Network of Biomedical Informatics INBIOMED web page available at www.inbiomed.retics.net/ Accessed on 02/26/2009

[7] Bioinformatics National Institute INB webpage available at http://www.inab.org/ Accessed on 02/26/2009.

[8] Sadee W, Dai Z. Pharmacogenetics/ genomics and personalized medicine. Hum Mol Genet. 2005;14: 207-214.

[9] Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo G. Are splicing mutations the most frequent cause of hereditary disease?. FEBS Lett. 2005:28:1900-1903.

[10] Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. Genome Biol. 2005;6(7):224.

[11] Jorgensen WL. The many roles of computation in drug discovery. Science. 2004;303:1813-1818.

[12] Butcher EC, Berg EL, Kunkel EJ. Systems Biology in drug discovery. Nat. Biotechnol. 2004;22:1253-1259.

[13] The Spanish Technological Platform for Innovative Medicines. http://www.medicamentos-innovadores.org/. Accessed on 02/26/2009.

[14] Alonso-Calvo R, Maojo V, Billhardt H, Martin-Sanchez F, García-Remesal M, and Pérez-Rey D. An Agent- and Ontology-based System for Integrating Public Gene, Protein and Disease Databases. J. Biomed. Inform. 2007;40(1):17-29.

[15] Vissers LE, Veltman JA, van Kessel AG, Brunner HG. Identification of disease genes by whole genome CGH arrays. Hum Mol Genet. 2005. 2: 215-223.

Homogenising access to heterogeneous biomedical data sources

Erwin Bonsma and Jeroen Vrijnsen

Philips Research, High Tech Campus 37, 5656 AE Eindhoven, The Netherlands

Abstract This paper reports our experiences of developing data access services in the context of the ACGT project. The paper documents two aspects of the work that we carried out. First the focus is on the problem of how to best provide a syntactically homogeneous data access interface for a set of heterogeneous data sources. We describe related work, outline the approach we have taken, and report our findings. The second part of this paper documents integration issues that we encountered when realizing the data access services. Choices with regards to realization have significant impact on the time and effort that is needed to develop and maintain the services and our experiences may provide useful guidance to others wanting to develop similar functionality.

Introduction

The work reported here has been carried out in the context of the ACGT (Advancing Clinico-genomic Trials on Cancer) project. The aim of ACGT is to develop open-source, semantic and grid-based technologies in support of post-genomic clinical trials in cancer research [1]. One of the main challenges in carrying out post-genomic research is to efficiently manage and retrieve all relevant data. Carrying out a post-genomic clinical trial involves the collection and storage of a wide variety of data, including: clinical data collected on Case Report Forms (e.g. symptoms, histology, administered treatment, treatment response), imaging data (e.g. X-Ray, CT, MR, Ultrasound), and genomic data (e.g. microarray data). Next to that there are many public biomedical databases that are relevant. These store information about gene and protein sequences, pathways, genomic variation, microarray experiments, medical literature, tumour antigens, protein domains, metabolites, etc. Biomedical researchers currently have to use many different tools and web interfaces to find and extract the data that is relevant to their clinical research. Providing seamless and integrated access to clinical, genetic and image databases would therefore greatly facilitate post-genomic research.

In order to provide seamless access to a heterogeneous set of databases syntactic and semantic integration needs to take place. Syntactic data integration handles differences in the formats and mechanisms of data access, whereas semantic integration deals with the meaning of information; it must handle the fact that information can be represented in different ways, using different terms and identifiers.

With regards to syntactic heterogeneities, the main areas where databases differ are:

- access protocols, e.g. SOAP/HTTP, DICOM, JDBC,
- data formats, e.g. different formatting of date values,
- message formats, e.g. XML, HTML, protocol-specific, and
- query mechanisms, e.g. SQL, literal matching, keyword-based search, or protocol-specific.

An example of a query mechanism specific to the biomedical domain is BLAST [2], which is used by sequence databases. Matching is approximate and parameters can be specified controlling the accuracy and speed of matching. A completely different query mechanism is needed to access medical image data, which is standardised using the DICOM protocol [3]. DICOM does not allow complex queries, as it does not intend to provide a generalized database query mechanism [4]. The baseline query functionality is very basic, and the optional extended query functionality is still limited and eccentric.

Semantic integration in ACGT is handled using Query Translation, carried out by a semantic mediator that uses a Local as View approach. It accepts queries expressed in the ACGT Master Ontology, divides them in sub-queries, and translates each to the ontology used by the underlying database. The remainder of this paper focusses on the syntactic integration of data sources. For details about the semantic integration approach, please refer to [5].

Related work

Syntactically homogeneous access to distributed data sources is typically provided by way of wrappers [6, 7, 8, 9]. One of the main challenges in building wrappers is the variation in the query functionality of the underlying data sources [10]. Data sources may not only use different data models and syntactically different query mechanisms, but their query capabilities can differ as well. This makes it difficult to support a common query language, an essential step towards syntactic homogeneity. There are two extreme approaches [7]. A highly expressive common query language can be chosen. This, however, makes it difficult to implement wrappers for sources with primitive query capabilities. Furthermore, if the wrappers are used by a mediator, it means that query decomposition, subquery scheduling and result composition may be done by both; the mediator must be able to decompose queries across multiple data sources and a wrapper for a data source must be able to decompose a complex query into simpler ones that the data source

can handle. This means duplication of implementation effort but also leads to overall sub-optimal query execution performance. On the other hand, if a very basic common query language is chosen, significant and unnecessary performance penalties are introduced as the capabilities of the underlying data sources are not effectively used.

As neither approach is ideal, an intermediate solution is proposed in [7]. A powerful common query language is chosen, but wrappers may choose to only support a subset of the queries, based on the capabilities of the underlying data source. Each wrapper describes the queries it supports using the Relational Query Description Language (RQDL) developed for this purpose. An RQDL specification consists of a set of query templates that represent parameterized queries that are supported. RQDL uses a context-free grammar to describe arbitrarily large sets of templates. Templates can be schema-independent as well as schema-dependent. Benefits of this approach are that wrappers can provide and expose query functionality that better corresponds to that of the underlying data source. A drawback is the increased complexity associated with interpreting and reasoning about the query capabilities of each source, but feasibility is demonstrated by the Capabilities-Based Rewriter, described in the same paper, that uses the wrappers and produces query execution plans in reasonable time.

A more recent example, applied in practice to life sciences data, is given by DiscoveryLink [8], a database middleware system for extracting data from multiple sources in response to a single query. The system consists of two parts: a wrapper architecture, and a query optimizer. SQL is used as the common query language for the wrappers, but wrappers may only support a subset of SQL. In the simplest case, a wrapper retrieves a projection over all rows in a given table. Wrappers can, however, also indicate that they support filtering conditions, or joins, and if so, how many. The paper proposes to involve wrappers in the query optimization process. Wrappers are asked for estimates on the query execution time and expected size of the result set for different sub queries. The query optimizer will use this information when deciding how to decompose the query. It requires efficient communication between the query optimizer and the wrapper, which is made possible because wrappers are shared-libraries, co-located with the query optimizer.

EDUTELLA [11] uses an RDF query language that has various language levels, with increasing functionality. The basic level supports RDF graph matching, the level above that adds disjunction, and the use of recursion in queries is added at even higher levels. Support for aggregation is an optional feature, orthogonal to these levels. Wrappers can support the level of the query language that best fits the query capabilities of their data source.

It is generally recognized that writing wrappers requires significant programming effort, and as a result significant research efforts have been devoted to automating parts of this (see e.g. [6], [9]). In general, automation is focused on a subset of the different data sources, e.g. sources with a web interface [12].

Approach

We identified the following functional requirements for the data access services. Firstly, they should provide a uniform data access interface. This includes uniformity of transport protocol, message syntax, query language, and data format. Secondly, they should export the structure of the database, using a common data model, together with possible query limitations of the data source. Clients of the web service require this information for constructing queries. Thirdly, they should enforce the data source access policy, and audit access to data sources. For post-genomic clinical trial data, there exist strict legal and ethical requirements that need to be adhered to.

A common query language is needed to achieve a uniform interface. It needs to meet various requirements. Firstly, it must be sufficiently expressive; it should support the types of queries that clinicians and biomedical researchers want to carry out. Secondly, it must be attainable, with acceptable effort, to map the query language to those used by the various data sources that need to be accessed. Thirdly, it must be convenient to use the query language for semantic mediation, the next step of the data integration process. Fourthly, it should be a community accepted standard. This ensures that there are sufficient support tools available, such as parsing and query engines, and also increases the possibilities for our approach to be eventually widely adopted. We have chosen SPARQL [13] as the query language, as it satisfies all these requirements.

Web Services have been chosen as the common interface technology within ACGT, as this technology suits the distributed nature of the project with respect to the data, computing resources, and development teams. For the data access services we decided additionally to use OGSA-DAI, a Web Services framework for data access [14]. It uses an activity framework that enables flexible service invocation, and re-use of common data access functionality. The results of queries will be returned using the SPARQL Query Results XML Format [15], which is the natural choice given the web services context and the use of SPARQL.

To meet the second requirement each data access service exports its schema using RDF Schema [16]. This is the standard way to describe RDF data sources, which is how the data sources appear given that SPARQL is used.

Access to each data source is controlled by integrating the data access service into the ACGT security infrastructure. Authentication is credential-based and delegation of credentials between services is supported. Authorization is controlled centrally and authorization decisions are, amongst others, based on membership to virtual organizations, which can be created as required.

Implementation

We have implemented data access services for three data source types: relational databases, medical image databases, and microarray databases. These databases have been chosen after careful review of requirements; they are considered the most important in the context of post-genomic clinical trials given the data-mining scenarios that were identified during the requirements-gathering process.

Figure 1 shows the data access services in the context of the data analysis architecture. The workflow enactor carries out data-mining workflows. It uses the semantic mediator for retrieving data. The latter accepts queries expressed in the ACGT Master Ontology, and converts them to the local ontology of the data source that is queried. The query results are converted in the opposite direction. Before a data access service handles a query, it checks whether or not the user is authorized to access the data source by contacting the authorization server. The data access services handle SPARQL queries from the semantic mediator. Additionally, they may also be contacted directly by the workflow enactor. This is the case for retrieval of image and assay files, which do not require semantic mediation. The requested data is typically not returned to the workflow enactor, but delivered to file at a specified temporary storage location. The workflow enactor receives the unique identifiers for files that have been created, which it can forward to the data-mining service so that the latter can retrieve and analyse the data.

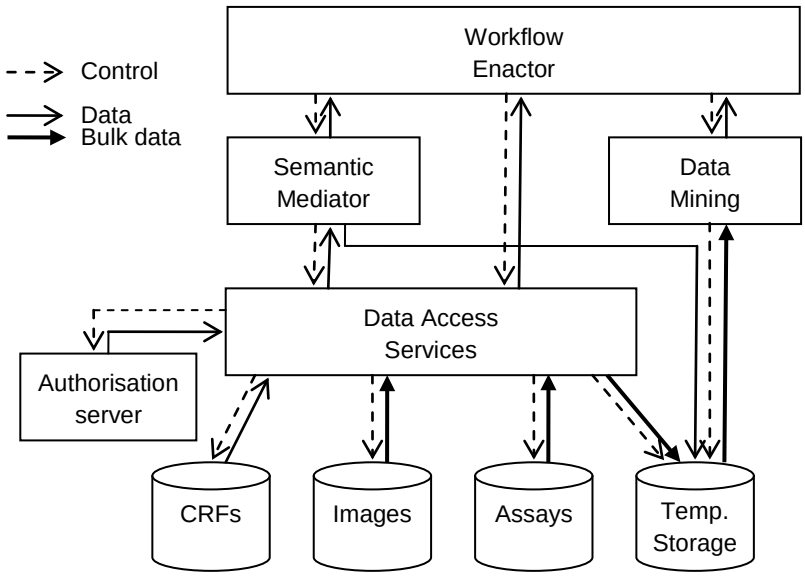


Fig. 1. The data analysis architecture of ACGT

There are two relevant aspects with regards to the terminology we use. First of all, we use the term “data access service” to refer to a class of services, e.g. the DICOM data access service, as well as for referring to specific instances, e.g. the data access service for DICOM database X. The distinction should always be apparent from the context. Secondly, each data access service is not actually a standalone web service. Within the OGSA-DAI framework multiple data access services are deployed as different data resources within a single OGSA-DAI web service. This has implications for the addressing of the data access services, but is not important for the remainder of this paper.

Query functionality

For the implementation of the query functionality for relational databases it is necessary to translate queries from SPARQL to SQL. For this, we are using the Open Source package D2RQ [17]. It can wrap a relational database into a virtual, read-only Jena RDF graph [18], rewrite SPARQL queries and Jena API calls into application-datamodel-specific SQL queries, and transform the returned data ino RDF triples. We therefore only had to integrate this functionality into the OGSA-DAI activity framework.

Realizing a data access service for medical image databases requires more effort. First of all, custom code is needed to implement the query translation. As the DICOM information model maps naturally to RDF, it is relatively straightforward to express DICOM queries in SPARQL. However, the DICOM standard only provides limited query functionality, which means that only a subset of syntactically valid SPARQL queries can be expressed as DICOM queries. For the initial implementation, we only support SPARQL queries that can either be directly converted to a DICOM query, or that can be handled using a single DICOM query combined with filters at the data access service that do not require temporary storage of query results (i.e. any query match that is returned by a DICOM server is either immediately discarded, or after optional conversion, immediately returned to the client). This way, the data access service does not need to store intermediate results, and implementation is significantly simplified. Figure 2 shows an example of a supported SPARQL query for a DICOM image repository.

For the medical image data access service, image retrieval functionality was also added; the ability to query the image metadata is of limited use if the actual images cannot be retrieved. The retrieval functionality has been implemented using OGSA-DAI’s activity framework so that it can be invoked in various ways. For example, a single request message can be used to query the image metadata, and to asynchronously retrieve and deliver the corresponding images.

```
PREFIX dicom: <http://example.philips.com/dicom/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?name ?dob ?studyId ?studyDescr
WHERE {
  ?patient dicom:PatientsName ?name ;
           dicom:PatientsBirthDate ?dob .
  ?study dicom:Patient ?patient ;
         dicom:StudyID ?studyId .
  OPTIONAL {
    ?study dicom:StudyDescription ?studyDescr .
  }
  FILTER ( ?dob >= "1970-01-01"^^xsd:date &&
           ?dob <  "1980-01-01"^^xsd:date )
}
```

Fig. 2. Example of DICOM query expressed using SPARQL.

Our third data access service provides access to the BASE database, a database for storing the results of microarray analysis [19]. The data access service interacts with the BASE database by way of a Web Service interface. The current implementation of the data access service provides retrieval of assay files, given their unique identifiers. More advanced query functionality is not provided, as this has not been needed yet. Typically assay files are obtained by first querying the clinical data, e.g. for all patients with an ER-negative tumor that responded positively to treatment, and next retrieving the corresponding assay files from BASE.

Miscellaneous functionality

Due to the heterogeneity of the data sources, each data access service requires code that is specific to its type of data source. However, the different data access services also need to provide common functionality, which offers the opportunity for code reuse. The main mechanism by which the OGSA-DAI platform encourages the reuse of code is through its activity framework [14]. Requests from clients to an OGSA-DAI data resource can contain multiple activities, linked together into a pipeline. For example, the first activity may comprise a query to a DICOM server for a set of image identifiers. A second activity may extract the identifiers and retrieve the corresponding images. The images may be fed to a third activity, which compresses the image data, and feeds the resulting archive file to a fourth activity, which delivers the archive to a specified FTP server. Al-

though the interface of the query activities for the different data access services is typically the same (thus providing a homogeneous interface), their implementation is typically highly dependent on the type of data source that is queried. Activities further in the pipeline are typically more generic and their implementation may be reused by multiple data access services. The OGSA-DAI platform comes with a large set of generic activities, but we developed additional ones for use by our data access services. One example is an activity for delivering files to the Gridge Data Management System [20], which we use for temporary storage, using myProxy certificates for authentication. Another activity can calculate checksums for data-streams. It can be used for testing service functionality after changes to the implementation, as well as for carrying out periodic liveness tests of a running service. We also extended the default ZIP activity so that it can pack multiple files into a single archive.

Integration experiences

The realization of the data access services requires integration of a large number of third-party software libraries. There are two reasons why a large number of third-party packages is needed: firstly, the complexity of the software stack associated with (grid-based) Web Services, the interface standard chosen in ACGT, and secondly, the heterogeneity of the underlying databases, which typically each have their own sets of standards and APIs associated with them. The software stack for the data access services consists of the following layers:

- data access services
- OGSA-DAI
- Globus
- Tomcat

The lowest layer is the Tomcat web service container, which hosts the web services. Globus sits on top of Tomcat; it is used for implementing the certificate-based security framework. The layer above that consists of OGSA-DAI. It provides a modular, activity-based data access framework for use by the layer above it. The top layer consists of the data access services which handle query and result transformation, and data retrieval and storage for the supported data sources. Each class of data access service depends on various third-party libraries for its implementation. For example, the relational data access service uses D2RQ [17] to translate SPARQL queries to SQL, which in turn uses Jena [18]. The DICOM data access service uses Jena as well, together with dcm4che [21] for accessing DICOM servers. The BASE data access service uses client-code provided by the BASE developers for accessing their BASE web service.

Given this setup, one of the biggest problems is managing the depencies between all different third-party software packages that are used. This is especially

challenging because all data service resources are deployed within the same OGSA-DAI instantiation and third-party packages (deployed as Java jar files) that are needed by only one or a few of the data access services are visible to all. This can lead to dependency conflicts between data access services that are otherwise independent. A pair of data access services that can individually be deployed successfully inside an OGSA-DAI instantiation may not necessarily be successfully deployed alongside each other.

Three concrete issues that we encountered may help to illustrate the types of integration problems this gives. Firstly, after we had discovered and reported a bug in a third party library we used (Jena), we could not deploy the release that included the fix, as this new release was incompatible with another third party library (D2RQ) that we were using

Secondly, we have experienced problems deploying compiled and packaged code provided by other partners in ACGT, which was due to a slight incompatibility in an underlying third-party library (Axis) provided by the version of the (Globus-based) web service container that was used. Fortunately, the incompatibility did not exist at the source code level, so rebuilding the code with the third-party libraries of the container where the service was to be deployed fixed the problem.

We encountered a third problem after we had upgraded the web services container, which was required to fix a depency conflict. One of our services would now hang when handling requests. As it turned out, this was due to a change of the third-party library implementing the JavaMail API, which resided three layers below our code. It was due to a more strict implementation of the JavaMail API, which in turn revealed a bug in another third-party library (Axiom), which relied on a more lenient interpretation of the API's contract in order to function correctly.

It is worth pointing out that in all three cases, the fact that source code was available for all third-party software components greatly helped in tracking down and solving the problem.

Discussion

We have implemented OGSA-DAI data access services for three types of data sources: relational databases, medical image databases and a micro-array database. The main research question is how to best provide a syntactically homogeneous interface, and a key question is the query language that is used. We have chosen SPARQL as the common query language and have demonstrated that it can be successfully applied to relational databases and DICOM image databases.

For the relational databases, the SPARQL language does not support all features offered by the query language of the data source, SQL. For instance, it does not support aggregation of data (averaging, summation, counting, etc). So aggregation needs to be performed at the client-side, even though the underlying data-

base supports it directly, which negatively affects performance. The actual use of the system by the end users will clarify whether this is a problem that needs to be addressed.

For medical image databases, SPARQL is more expressive than the query support provided by the DICOM protocol. For this reason, the data access service does not support all queries. These limitations are currently described as text, but should be expressed in a more formal manner, so that other services and applications can interpret these and handle accordingly. In order to select a suitable formal framework for this, we need to thoroughly review the capabilities and limitations of all relevant data sources.

A capability-restricted data access architecture has the advantage that it is easier to develop data access services for data sources; as a consequence, new data sources can be integrated much more quickly. It may, however, complicate applications and services that use the data access services. A higher level data access service may therefore be introduced that hides query restrictions of the underlying services. This generic service would decompose queries for a specific data access service as need be, store the intermediate results, and join these to produce the final answer. This would facilitate implementation of the semantic mediator, while incurring a slight performance penalty. However, this higher-level data access service may also carry out generic optimizations such as caching of query results, resulting in performance gains.

Another open issue is how to provide text-based query functionality. There are many public biomedical databases where part of the data is free text. Examples are descriptions of microarray experiments (e.g. in GEO [22] and ArrayExpress [23]), descriptions of gene and protein functions (e.g. in UniProt [24] and EntrezGene [25]), and abstracts and titles of medical publications (e.g. in PubMed [26]). Although most databases provide keyword-based functionality for querying data, this method of searching is not directly supported by SPARQL, so it is not immediately obvious how to extend the current data access services interface to support this functionality. One approach would be to add a separate text-based query interface for data sources that support this. This exposes more details of the underlying data source, resulting in a less homogeneous interface. This is undesirable but may be unavoidable in practice. However, there is a more important question that needs to be answered first: how should querying of text data be handled by the semantic layer? This is an important question as it determines the query interface that is available to end-users, but answering it falls outside the scope of this paper.

To give an impression of the overhead caused by the use of data access services, compared to direct interaction with the databases, we can report the results of performance experiments that we have carried out. The amount of overhead depends on various factors, including the complexity of the query, the amount of results that are returned, and the underlying database. For simple queries the performance degration could be as much as a factor hundred (in particular for the relational database, which responds very quickly). For more complex queries the

overhead decreased significantly, down to a factor of two (for the DICOM database). Overhead was similarly low for retrieval of bulk image and microarray data, but high for retrieval of bulk data that is returned in the XML response message. The latter is due to limitations of the API for constructing the response message, which needs to be constructed entirely in memory before it can be sent to the client.

Finally, many of the problems encountered when deploying data access services for heterogeneous data sources are of a practical nature. For reasons of scalability, all data access services are deployed in the same web services container. This implies however that they run inside the same virtual machine, which can lead to unexpected conflicts. The complexity of the grid-based web services stack in combination with the need to use many third-party libraries, each with their own dependencies and particular implementations of part of the web services stack, makes it a challenge to resolve dependency conflicts.

Acknowledgments This research has been supported by the Advanced Clinico-Genomic Trials in Cancer (ACGT) project (FP6 IST-2005-026996) funded by the European Commission. Thanks to the anonymous reviewer for the useful feedback.

References

[1] Tsiknakis M, Kafetzopoulos D, Potamias G, et al (2006) Building a European Biomedical Grid on Cancer: The ACGT Integrated Project. In: Studies in Health Technology and Informatics, Volume 120

[2] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). "Basic local alignment search tool". J Mol Biol 215 (3): 403–410

[3] Bidgood Jr WD, Horii SC, Prior FW, van Syckle DE (1997) Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. J Am Med Inf Assoc 4(3).

[4] National Electrical Manufacturers Association (2004) Digital Imaging and Communications in Medicine (DICOM), Part 4: Service Class Specifications, Annex C: Query/Retrieve Service Class, 27-74

[5] Martín L, Bonsma E, Potamias G, et al (2007) Data Access and Management in ACGT: Tools to solve syntactic and semantic heterogeneities between clinical and image databases. In: LNCS 4802, 24-33.

[6] Papakonstantinou Y, Gupta A, Garcia-Molina H, Ullman J (1995) A query translation scheme for rapid implementation of wrappers, In: LNCS 1013, 319-344

[7] Papakonstantinou Y, Gupta A, Haas L (1998) Capabilities-based query rewriting in mediator systems, Distrib and Parallel Databases, 6:73-110

[8] Haas LM, Schwarz PM, Kodali P, et al (2001) DiscoveryLink: A system for integrated access to life sciences data sources. IBM Syst J, 20(2):489-511

[9] Thiran P, Hainaut JL, Houben GJ (2005) Database Wrappers Development: Towards Automatic Generation, In: Proc 9th Eur Conf Softw Maintenance and Reengineering, 207-216

[10] Hernandez T, Kambhampati S (2004) Integration of biological sources: Current systems and challenges. ACM SIGMOD Record 33 (3), 51-60

[11] Nejdl W, Wolf B, Qu C, et al (2002) EDUTELLA: A P2P networking infrastructure based on RDF. In: Proc 11th Int World Wide Web Conf (WWW2002)

[12] Liu L, Zhang J, Han W, et al (2005) XWRAPComposer: A Multi-page data extraction service for bio-computing applications. In: Proc 2005 IEEE Int Conf on Serv Comput, 271-278

[13] Prud'hommeaux E, Seaborne A (2007) SPARQL Query Language for RDF, W3C Candidate Recommendation, http://www.w3.org/TR/2007/CR-rdf-sparql-query-20070614/. Accessed 25 Feb 2009
[14] Antonioletti M, et al, The design and implementation of grid database services in OGSA-DAI. In: Concurrency and Computation: Practice and Experience, 17(2-4):357-376
[15] Becket D, Broekstra J (2008) SPARQL Query Results XML Format. W3C Recommendation. http://www.w3.org/TR/rdf-sparql-XMLres/. Accessed 30 Mar 2009
[16] Brickley D, Guha RV (2004) RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation. http http://www.w3.org/TR/rdf-schema/. Accessed 30 Mar 2009
[17] Bizer C, Seaborne A (2004) D2RQ – Treating non-RDF databases as virtual RDF graphs. In: Proc 3rd Int Semantic Web Conf (ISWC2004)
[18] Carroll JJ, Dickinson I, Dollin C, et al (2003) Jena: Implementing the semantic web recommendations. Technical Report HPL-2003
[19] Saal LH, Troein C, Vallon-Christersson J, et al (2002) BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data. Genome Biology 2002 3(8): software0003.1-0003.6
[20] Pukacki J, Nabrzyski J, et al (2006) Programming Grid Applications with Gridge, Comp Methods in Sci and Technol 12(1):47-68
[21] Open Source Clinical Image and Object Management, http://www.dcm4che.org/. Accessed 11 Feb 2009
[22] Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/ Accessed 11 Feb 2009
[23] ArrayExpress, http://www.ebi.ac.uk/microarray-as/ae/. Accessed 11 Feb 2009
[24] Uniprot, http://www.uniprot.org/. Accessed 11 Feb 2009
[25] Entrez Gene, http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene. Accessed 11 Feb 2009
[26] PubMed, http://www.ncbi.nlm.nih.gov/pubmed/. Accessed 11 Feb 2009

Unification of heterogeneous data towards the prediction of oral cancer reoccurrence

Konstantinos P. Exarchos[1,2], Yorgos Goletsis[3], Dimitrios I. Fotiadis[1,*]

[1] Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, GREECE
[2] Dept. of Medical Physics, Medical School, University of Ioannina, Ioannina, GREECE
[3] Dept. of Economics, University of Ioannina, Ioannina, GREECE

*Corresponding author

kexarcho@cc.uoi.gr, goletsis@cc.uoi.gr, fotiadis@cs.uoi.gr

Abstract. Oral cancer is the predominant neoplasm of the head and neck. Annually, more than 500.000 new cases of oral cancer are reported, worldwide. After the initial treatment of cancer and its complete disappearance, a state called remission, reoccurrence rates still remain quite high and the early identification of such relapses is a matter of great importance. Up to now, several approaches have been proposed for this purpose yielding however, unsatisfactory results. This is mainly attributed to the fragmented nature of these studies which took into account only a limited subset of the factors involved in the development and reoccurrence of oral cancer. In this work we propose a unified and orchestrated approach based on Dynamic Bayesian Networks (DBNs) for the prediction of oral cancer reoccurrence after the disease has reached remission. Several heterogeneous data sources featuring clinical, imaging and genomic information are assembled and analyzed over time, in order to procure new and informative biomarkers which correlate with the progression of the disease and identify early potential relapses (local or metastatic) of the disease.

Keywords: oral cancer, dynamic Bayesian networks, reoccurrence, disease modeling

Introduction

Oral cancer refers to the cancer that arises in the head and neck region, i.e. in any part of the oral cavity or oropharynx. Oral cancer constitutes the eighth most common cancer in the worldwide cancer incidence ranking; more than half million patients are diagnosed with oral squamous-cell carcinoma worldwide every year [1]. Oral cancer is highly related to the sex of the patient, with men facing twice the risk of being diagnosed with oral cancer than women. Research has revealed several risk factors associated with the development of oral cancer. Smoking and excessive consumption

of alcohol, and especially the combination of the two, constitute predominant risk factors for developing oral cancer. Moreover, sun exposure is another important risk factor, particularly for the cancer of the lip [1]. Some studies have also suggested that infection with the human pappilomavirus (HPV) is associated with oral cancer, especially with occurrences in the back of the mouth (oropharynx, base of tongue, tonsillar pillars and crypt, as well as the tonsils themselves) [2].

Cancer cells can spread to other adjacent parts of the neck, the lungs or elsewhere in the body. A common metastasis occurs in the neck lymph nodes through the lymphatic system which helps the cancer cells spread. Although nowadays, the continuous improvements in treatment protocols of cancer have achieved high rates of successful disease disappearance [3], there is a critical stage for the disease evolvement after the treatment called remission; during this stage there is no clinical, laboratory or imaging evidence of the neoplastic mass and the patient is considered cancer free. Nevertheless, even at this point some "invisible" disease particles might still be present leading to a potential spread or metastasis of the disease. Specifically, in terms of oral cancer, locoregional reoccurrence rates after the disease has reached remission have been reported in the range of 25-48%; such high figures can be justified given the deeply infiltrative nature of these tumors, as well as, the significant potential for occult neck metastasis [4].

The reoccurrence rates for oral cancer are quite high and they also suffer from poor prognosis, which can be partly attributed to histologically unfavorable features [4]. Moreover, patients with oral cavity cancer have to deal with the impact of the disease and its treatment on their physical appearance and on the ability to eat and speak, and subsequently with a significant decrease of the quality of life. Hence, early detection of reoccurrence might prove very beneficial [5]. Currently implemented methods aiming to predict oral cancer reoccurrence after the disease has reached remission, have reported quite inadequate results. Although several factors have been associated with the reoccurrence of oral cancer, such as age, site and stage of the primary tumor as well as histological features, they have not been studied altogether in a collective study. Moreover, especially in the molecular basis of the disease, currently available biomarkers are limited in number and efficiency [6, 7]. The efficient combination of the already known ones will greatly benefit the accurate stratification of the patients in terms of staging.

In the general framework of disease prognosis and modeling, several diverse approaches have been proposed in the literature. Most of them involve a prognostic model which implements a risk score depicting the progression of the disease and the general condition of the patient. Based on this score, simple decision rules are used to stratify the patients into several risk categories [8, 9]. More recent approaches utilize advanced machine learning algorithms, such as Artificial Neural Networks (ANNs) or Support Vector Machines (SVMs) which accept as input several variables and provide prediction about the desired outcome. However, most of these approaches use a "black-box" architecture and thus do not provide adequate reasoning about the decision [10, 11]. In addition, it is very cumbersome, if not infeasible to represent properly temporal problems using these algorithms. These issues pose significant limitations for the acceptability of the produced decision systems both by the medical community and the patients. In the case of oral cancer, and cancer in general, the physicians are extremely interested in knowing if, when and why a reoccurrence will

appear. Hence, especially for the problem under consideration (i.e. oral cancer reoccurrence prediction) it is very important to provide sufficient justification about the prediction, but also to introduce the time dimension in the modeling procedure.

In this work, we present an efficient framework in order to systematically study and analyze the factors associated with the reoccurrence of oral cancer, after the remission of the disease. This objective involves the integration of heterogeneous clinical, imaging and genomic data, thus facilitating the multiscale and multilevel modeling of the disease progression over time. Due to the constantly evolving nature of the disease, we employ DBNs, which efficiently cope with temporal causalities, thus, identifying the timing of a potential reoccurrence. Moreover, the intuitive design of DBNs allows for comprehensible decisions coupled with adequate justification. The multitude of gathered data is likely to uncover the evolution and development of the disease during remission, thus assisting the monitoring of patients after treatment, but also contribute towards the accurate stratification of patients in terms of staging. Knowing in advance the progression of the disease, i.e. identifying groups of patients with higher/lower risk of reoccurrence is a key factor towards the determination of the most proper treatment.

Materials and Methods

Clinical scenario

In order to clarify the steps of our study, a clinical scenario is employed which is shown in Figure 1. Initially a patient is diagnosed with cancer through traditional clinical procedures. At this point the physician gathers the required data in order to extract the baseline profile and the patient is treated properly. After the physician's therapeutic intervention, the patient either reaches complete remission or particles of the cancer tissue still remain intact. In the latter case the patients do not qualify for the purposes of our study, whereas from the patients in complete remission, where the cancer is no longer visible, data are further collected, forming the post-treatment profile. Afterwards, and during a two year time span, data are collected from the patient regularly (i.e. scheduled visits are planned for months 1, 3, 6, 9, 12, 15 and 18 after treatment) in order to formulate as a personalized follow-up signature, which is being constantly analyzed. The choice of the follow-up period was determined by the fact that a reoccurrence is most likely to appear in a two year period after the initial treatment. The purpose of this analysis is to stratify the patients in two clusters: i) low risk of disease reoccurrence and ii) high risk of reoccurrence. Hence, we are able to dully identify relapses of the disease and adjust the follow-up treatment accordingly.

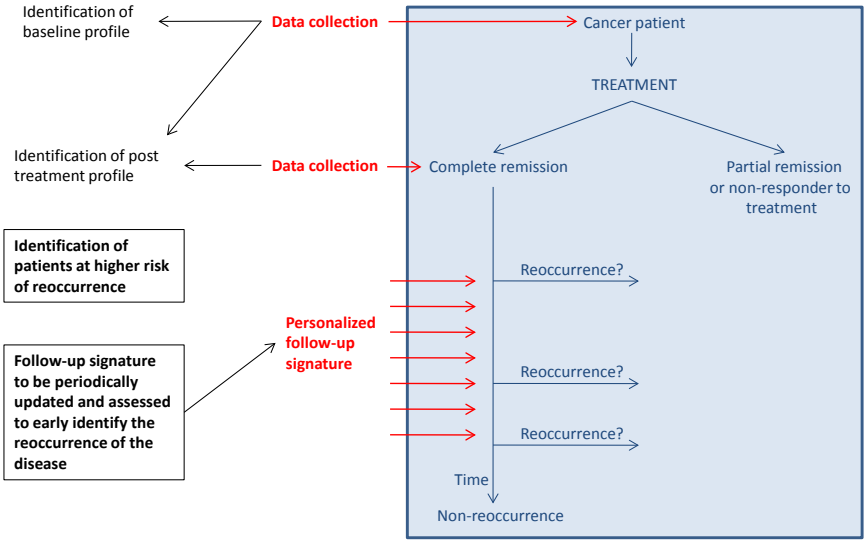


Figure 1: Clinical scenario employed in our study.

Data collection

The progress of the disease in a total of 150 patients with oral squamous cell carcinoma is evaluated during the present study. The cases are collected from two major clinical centers which reside in Italy and Spain. According to available literature 70-80% of these patients are expected to achieve complete remission of the disease after treatment, and an approximate 30-40% of them will develop a reoccurrence of the cancer. Relapses during a two-year time span are marked, as well as the timing of the relapse, and the patients are grouped in two categories, the relapsers and the non-relapsers, which we aim to discriminate by studying and analyzing a multitude of heterogeneous data.

Due to the complex nature of cancer, a major challenge towards its diagnosis and treatment is to formulate a collective approach in order to "frame" every possible aspect. For this purpose we propose a holistic approach which involves the integration and analysis of multiscale and multilevel data. Specifically, clinical, imaging and genomic data are assembled ranging in the scale of dimension and localization. The employment and careful analysis of the above heterogeneous data is likely to reveal the interactions which take place during oral cancer onset and progression. Consequently, the data collected from every patient will comprise the following information:

- Clinical data from health records and standard laboratory markers, histological data from tumor mass specimen

- High throughput genomic data from tumor tissue specimens and circulating cells, profiling gene expression at whole genome level by oligo-RNA microarrays
- Imaging data of the prime tumor mass (and secondary localizations if present)

All these data will be efficiently integrated into a single repository formulating the basis of our study. The data involved in the present study along with the specific techniques employed for their manipulation and analysis are described in detail in the sections that follow.

Clinical

For the diagnosis and monitoring of patients with oral cancer the following types of clinical data are assembled:

- Anamnesis
- Demographics
- Risk factor
- Tumor clinical aspect
- TNM staging
- N characteristics

Anamnesis refers to the detailed medical review of the patient's past health state. Detailed information about the patient's past health problems, general health state, family medical history, oral cancer risk factors and symptoms is gathered in order to establish the diagnosis. Demographic data along with several risk factors are also assembled in order to aid the diagnosis. Next the tumor's clinicopathological stage and developmental phase are evaluated. The most common staging system used for oral cancer is the TNM system. Moreover, several markers have been proven to affect the patient's response to adjuvant and neo-adjuvant treatments [12, 13]. In the present study we compile an extensive list containing all these clinical factors in order to perform a collective study of their relation with oral cancer progression and treatment efficacy. All these data, which comprise the clinical data associated with oral cancer, are thoroughly analyzed for the purposes of the present study.

Genomic

Current advances in the field of genomics have enormously facilitated the thorough analysis of gene expression within cells and tissues. Hence, we are able to extract important information about the interactions and biological pathways which take place during cancer evolution. The framework of the present work employs oligonucleotide and complementary DNA arrays in order to unravel the molecular basis of oral cancer. Nucleic acid arrays have rapidly become a popular investigational tool for cancer biologists, towards the identification of robust genetic biomarkers, thus, shedding considerable light into the complexity of the disease.

Systematic analysis of gene expression data is likely to yield potential tumor markers, or reliable combinations of biomarkers, that can be afterwards used in the daily practice for the diagnosis and monitoring of carcinoma of the head and neck.

Gene expression data come from a feature extraction (FE) file. An FE file is a tab delimited text file comprising of expression values (Log2-ratio data), raw intensity data, background information, metadata regarding the experiment and the scanning settings, gene annotation, etc. A typical FE file is shown in Figure 2.

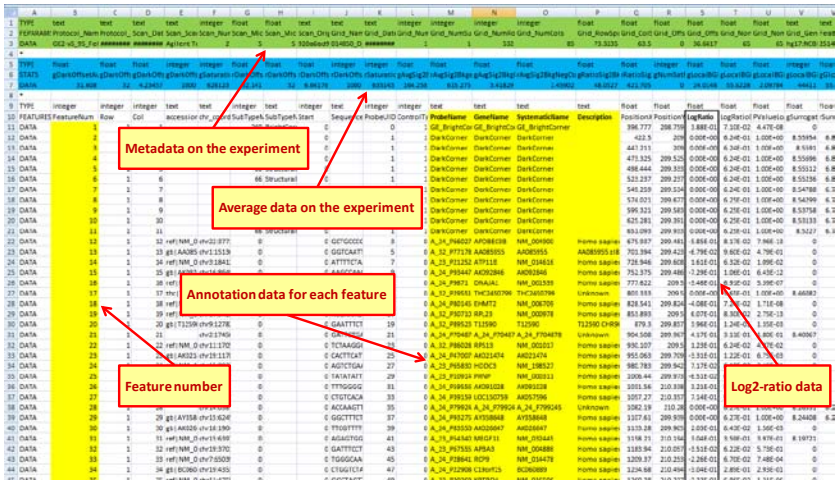


Figure 2: Typical entities extracted from a microarray experiment.

In the present study, all microarray experiments are conducted using the same platform, the same array design and the same FE procedure, in order to minimize the risk of possible sources of variability in the data, other than biological variability.

Especially for genomic data, a preprocessing stage is necessary for enhancing the quality of the data. After obtaining the gene expression data from the microarray experiments the duplicate and control features are eliminated. Control features are negative and positive control elements usually represented by empty features or spots that are hybridized independently from the original sample. Whereas, duplicate features are probes corresponding to a gene or a known internal control sequence which are printed more than once in the array, usually in random positions. They are used to verify the internal consistency of the data and the regional quality of the hybridization. Furthermore, data with high variability, too low signal and genes with a large number of missing values, constituting unreliable expression levels are carefully filtered out.

The overall flowchart for the basic preprocessing of the gene expression data is shown in Figure 3.

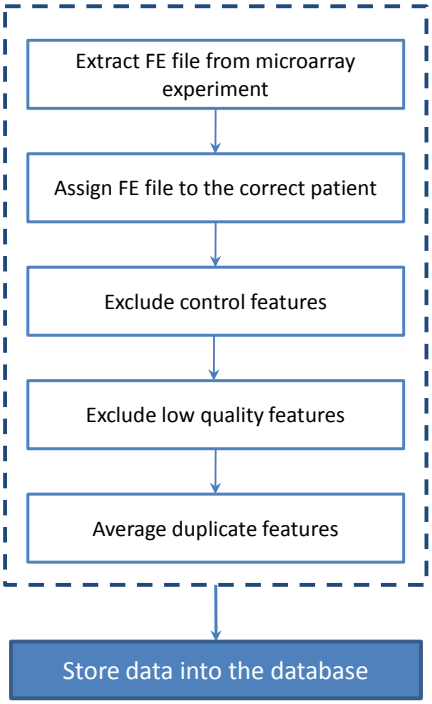


Figure 3: Preprocessing of the gene expression data.

Imaging

Image data from the cancerous tissue can reveal certain significant characteristics of the localization and progress of the disease. The present study employs MRI and CT images. The manipulation of the employed images involves the following main steps, which are also depicted in the flowchart of Figure 4.

- Image preprocessing
- Definition of regions of interest (ROIs)
- Extraction and selection of features
- Classification of the selected ROI

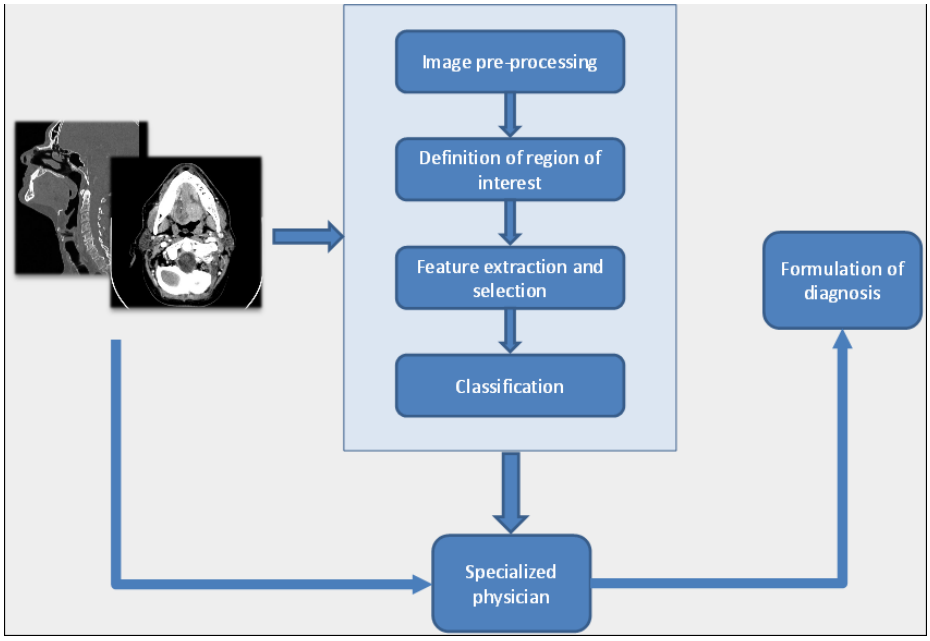


Figure 4: Image data analysis and manipulation.

Initially, the images need to be preprocessed properly in order to improve their quality to facilitate the overall image analysis procedure. The most common types of imaging data contamination are noise and artefacts. Noise causes random distortion in the data, and although several approaches have been proposed in the literature (e.g. application of filters), it is still quite difficult to remove it, due to its random nature. On the other hand, artefacts usually involve more deterministic perturbations of the data, hence it is easier to detect and omit them. These problems can be attributed to several factors such as human error, measuring device limitations, etc. Other types of image preprocessing involve edge enhancement (e.g. unsharping, wavelet transformation), image contrast enhancement (histogram equalization) and image standardization.

In the next step, we detect regions of interest (ROIs), i.e. regions of the preprocessed image bearing enhanced role for our purposes. For the initial approximate definition of some ROIs, a specialized radiologist pinpoints sites of interest, i.e. tumor center, lymph nodes or potential infiltrations. Moreover, automatic methods are also be employed for the detection of ROIs. Active contour models are often employed for automatic definition and tracking of anatomical contours in 2D medical images due to their ability to approximate accurately the random shape of organ boundaries. Seeded region growing is another example of semi-automatic method widely used for the definition of ROIs in medical images.

Afterwards, several features are extracted from the ROIs in order to uniquely characterize the image itself or structures contained in the data. Some of these features represent quantitative measurements with certain physical meaning, that a specialized physician must take into account in order to formulate the diagnosis. However, in some cases features with no apparent physical meaning can be extracted

due to their enhanced discriminative potential. The most common features employed for the analysis of medical images are: pixel based features, texture features, shape features (transformation dependent and transformation independent). Specifically, in the present study, the following features are calculated from each ROI:

- Six (6) features from first order statistics
- Forty eight (48) features from spatial gray-level dependencies matrix
- Twenty (20) features from gray-level differences matrix
- Twelve (12) features from Law's texture energy measurements and
- Three (3) features from fractal dimension measurements

Additional features describing specific properties of the image under consideration are assessed, such as tumor volume, periosteal infiltration, etc. All features extracted during this stage are deposited in a collective repository along with the genomic and clinical data.

Dynamic Bayesian Networks (DBNs)

In the present study we employ DBNs in order to early identify potential relapses of the disease, during the period of remission. As it is described in the clinical scenario, a snapshot of the patient's medical condition is acquired during every predefined follow-up with the doctor. By exploiting the information of history snapshots we aim to model the progression of the disease in the future. The proposed prognostic model is based on DBNs, which are temporal extensions of Bayesian Networks (BNs.) [14]. A BN can be described as $B = (G, P)$ where G is a directed acyclic graph, where the nodes correspond to a set of random variables $\mathbf{X} = \{x_1, x_2, ..., x_N\}$, and P is a joint probability distribution of variables in $\mathbf{X}$, which factorizes as:

$$P(\mathbf{X}) = \prod_{i=1}^{N} P(x_i \mid \pi_G(x_i)) \quad (1)$$

where $\pi_G(x)$ denotes the parents of x in G. A DBN can be defined as a pair $DB = (B_0, B_{trans})$ where B_0 is a BN, defining the prior $P(\mathbf{X_0})$ and B_{trans} is a two-slice temporal BN (2TBN) which defines $P(\mathbf{X_t} \mid \mathbf{X_{t-1}})$. The semantics of a DBN can be defined by "unrolling" the 2TBN until we have T time-slices. The resulting joint distribution is given by:

$$P(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X}_T) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(x_i^t \mid \pi(x_i^t)) \quad (2)$$

In order to build a model that successfully evaluates the current state or predicts a state in the future (next time slice), we need to train both the structure of the DBN (G_0, G_t) and the parameters of the conditional probability distributions, using both expert knowledge as a prior model and experimental data to get a more accurate posterior model. After the training procedure, we obtain a model as the one shown in Figure 5. By providing some evidence to the model, we are able to compute the

probability of any variable for every time slice (i.e. in any predefined follow-up visit), including of course the probability for reoccurrence.

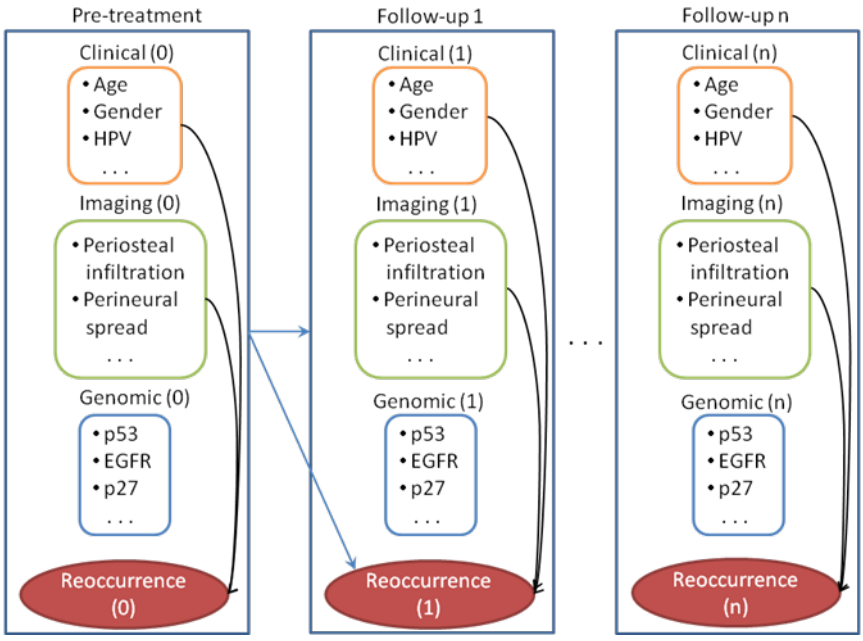


Figure 5: Provisional architecture of the employed DBN model.

For the development of the DBN two implementations have been explored. In the first implementation every source of data is used separately, in order to build a distinct DBN, specifically tailored for a certain type of data. Consequently, three DBNs are developed and their outputs are combined using a meta-classification function (Figure 6(a)). In the second, all sources of data are employed altogether in order to develop a single DBN (Figure 6(b)). However, in both implementations, the contribution and feedback from a specialized doctor, during the DBN construction, is substantial. The two implementations are depicted in Figure 6.

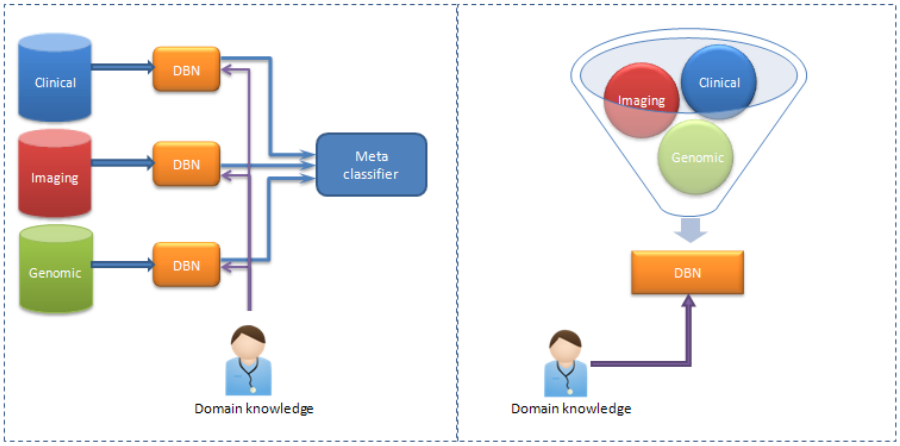


Figure 6: Analysis schemes: (a) multiple DBNs, (b) single DBN.

As this work is currently under development, detailed testing of both implementations depicted in the figure above is needed so as to assess the potential of each one. The assessment will be done using an annotated dataset, covering the two-years follow-up data, which is currently being populated.

Discussion & conclusions

In the present study we propose an advanced framework which implements heterogeneous sources of data towards the prediction of oral cancer reoccurrence in patients that have reached remission. A large amount of clinical, genomic and imaging features are analyzed in order to extract biomarkers that are highly associated with relapses of oral cancer. Thus, we overcome a major limitation of similar studies in the field that employ only a confined subset of features that are associated with oral cancer. Another significant challenge is to capture the disease progression over time. For this purpose we employ DBNs, which are specifically designed to represent temporal causalities. The inclusion of the time dimension is very important as most doctors are interested – even with a rough approximation – in the timing of the reoccurrence. Furthermore, DBNs are able to provide reasoning for the reported decisions, thanks to their transparent architecture. This characteristic is very appealing, if not prerequisite by the medical community. Hence, not only we are able to predict a certain outcome but also to gain insight about the rationale of every decision. In overall, the currently proposed framework contributes significantly towards the monitoring of oral cancer evolvement since it can answer if, when and why a reoccurrence might appear.

Acknowledgements

This work is part funded by the European Commission NeoMark project (FP7-ICT-2007-224483) – ICT enabled prediction of cancer reoccurrence.

References

1. Haddad, R.I., Shin, D.M.: Recent advances in head and neck cancer. The New England journal of medicine **359** (2008) 1143-1154
2. Mork, J., Lie, A.K., Glattre, E., Hallmans, G., Jellum, E., Koskela, P., Moller, B., Pukkala, E., Schiller, J.T., Youngman, L., Lehtinen, M., Dillner, J.: Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. The New England journal of medicine **344** (2001) 1125-1131
3. Forastiere, A., Weber, R., Ang, K.: Treatment of head and neck cancer. The New England journal of medicine **358** (2008) 1076; author reply 1077-1078
4. Godden, D.R., Ribeiro, N.F., Hassanein, K., Langton, S.G.: Recurrent neck disease in oral cancer. J Oral Maxillofac Surg **60** (2002) 748-753; discussion753-745
5. Sciubba, J.J.: Oral cancer. The importance of early diagnosis and treatment. American journal of clinical dermatology **2** (2001) 239-251
6. D'Silva, N.J., Ward, B.B.: Tissue biomarkers for diagnosis & management of oral squamous cell carcinoma. The Alpha omegan **100** (2007) 182-189
7. Lippman, S.M., Hong, W.K.: Molecular markers of the risk of oral cancer. The New England journal of medicine **344** (2001) 1323-1326
8. Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., Bergner, M., Bastos, P.G., Sirio, C.A., Murphy, D.J., Lotring, T., Damiano, A., et al.: The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest **100** (1991) 1619-1636
9. Le Gall, J.R., Lemeshow, S., Saulnier, F.: A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. Jama **270** (1993) 2957-2963
10. Cruz, J.A., Wishart, D.S.: Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics **2** (2006) 59-78
11. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine **34** (2005) 113-127
12. Woolgar, J.A., Rogers, S., West, C.R., Errington, R.D., Brown, J.S., Vaughan, E.D.: Survival and patterns of recurrence in 200 oral cancer patients treated by radical surgery and neck dissection. Oral oncology **35** (1999) 257-265
13. Woolgar, J.A., Scott, J., Vaughan, E.D., Brown, J.S., West, C.R., Rogers, S.: Survival, metastasis and recurrence of oral cancer in relation to pathological features. Annals of the Royal College of Surgeons of England **77** (1995) 325-331
14. Murphy, K.P.: Dynamic Bayesian Networks: Representation, Inference and Learning. UNIVERSITY OF CALIFORNIA (2002)

Building a System for Advancing Clinico-Genomic Trials on Cancer

Stelios Sfakianakis, Norbert Graf, Alexander Hoppe, Stefan Rüping, Dennis Wegener, and Lefteris Koumakis

Abstract The analysis of clinico-genomic data poses complex computational problems. In the project ACGT, a grid-based software system to support clinicians and bio-statisticians in their daily work is being developed. Starting with a detailed user requirements analysis, and with the continuous integration of usability analysis in the development process, the project strives to develop an architecture that will substantially improve the way clinico-genomic trials are conducted today. In this paper, results of the initial requirements analysis and approaches to address these requirements are presented. We also discuss the importance of appropriate metadata to tailor the system to the needs of the users.

1 Introduction

The goal of the Advancing Clinico-Genomics Trials on Cancer (ACGT[1]) project is to develop an open-source and open access IT infrastructure that provides the biomedical research community with the tools needed to integrate complex clinical information and make a concrete step towards the tailorization of treatment to the patient[4]. The necessity of such an environment is evident today more than ever due to the recent advancements in

Stelios Sfakianakis · Lefteris Koumakis
Institute of Computer Science, FORTH, Greece, e-mail: {ssfak,koumakis}@ics.forth.gr

Norbert Graf · Alexander Hoppe
University Hospital of Saarland, Paediatric Haematology and Oncology, D-66421 Homburg, Germany, e-mail: {norbert.graf,alexander.hoppe}@uniklinikum-saarland.de

Stefan Rüping · Dennis Wegener
Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany, e-mail: {stefan.rueping,dennis.wegener}@iais.fraunhofer.de

[1] http://www.eu-acgt.org

high throughput genomics and post-genomics technologies. These technologies yield an enormous pool of data that needs to be managed, analysed, correlated, and comprehended for the treatment of diseases like cancer and for the benefit of the community at large.

In this paper we discuss how the clinical requirements for such an environment can be addressed in a large-scale system and how appropriate meta data can be used to achieve satisfaction of the end user. The rest of the paper is structured as follows: Section 2 gives an overview over the user requirements analysis that was conducted in ACGT and highlight the main challenges. Section 3 introduces the main ACGT architecture, and in Section 4 some new approaches to address the user requirements are presented. Section 5 concludes.

2 The End-User View

Treatment and survival of patients with cancer is increasing steadily for most age groups as shown in Fig. 1 which gives the average annual percentage change over a period of 10 years. One of the most important reasons for this success story is the enrollment of patients in prospective clinical trials. Nevertheless, for most clinical trials in cancer, the number of patients recruited is much lesser than the number of eligible patients. In adults only 5% of cancer patients are participating in such trials. Therefore, higher rates are of utmost importance, especially in those cancers with still a dismal prognosis. To achieve this goal it is necessary to facilitate the building and running of clinical trials and to attract more patients to participate. In addition, the improvement in molecular biology has to be taken into account to create more clinico-genomic trials.

Recent advances in methods and technologies in molecular biology have resulted in an explosion of information and knowledge about cancer and its treatment. As a result, our ability to characterize and understand the various forms of cancer is growing exponentially. Information arising from post-genomics research and combined genetic and clinical trials on one hand, and advances from high-performance computing and informatics on the other, is rapidly providing the medical and scientific community with an enormous opportunity to improve prognosis of patients with cancer by individualizing treatment. To achieve this goal, a unifying platform is needed that has the capacity to process this huge amount of multi-level and heterogeneous data in a standardized way. Multi-level data collection within clinico-genomic trials and interdisciplinary analysis by clinicians, molecular biologists and others involved in life science is mandatory to further improve the outcome of cancer patients. It is essential to merge the research results of biomolecular findings, imaging studies and clinical data of patients and to enable users to easily join, analyze and share even great amounts of data. To provide a functional

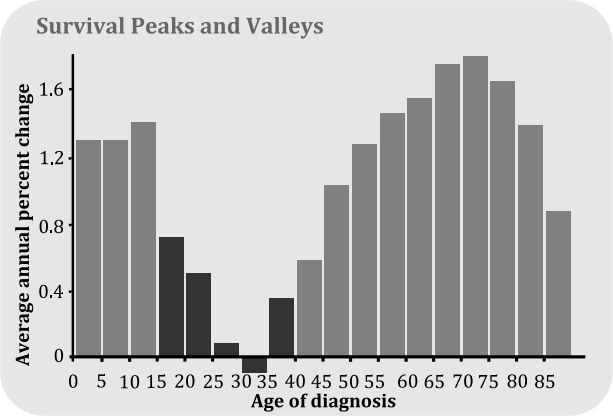


Fig. 1 Average annual change in survival in patients with cancer [5].

and user-friendly platform it is of utmost importance that the development of such a platform is user-driven and evaluated by end users right from the planning and development phase. Tools and software developed within ACGT are based on the user's needs and have to be in accordance with ethical and legal requirements of the European Community.

The project has selected indicative Clinical Trials on Cancer, namely breast cancer, pediatric nephroblastoma and in-silico modeling and simulation of tumor growth and response to treatment, for the initial requirements gathering activity. Since ACGT sees the requirements engineering process as a structured set of activities which will lead to the production of the final system requirements, an iterative requirements engineering process has been adopted, mainly based on scenarios and prototyping. Inputs to the requirements engineering process are information about existing systems, user and stakeholder needs, organizational standards, regulations and other domain information. As clinico-genomic trials are in the center of ACGT a Clinical Trial Management System is of utmost importance, to collect clinical, biomedical, imaging and other trial specific and relevant data. ACGT will provide such a tool, called ObTiMA [7, 8], whose functionality includes administrative and scientific aspects of clinico-genomic trials.

It has to be stressed that such a complex platform as ACGT, dealing with extremely sensitive data (patient data) and used by many different, sometimes multi-role, end-users, having different needs and requirements, a Data Protection Framework for ACGT is mandatory. This is based on the anonymization of patient data, the informed consent from participating patients and the binding of partners/centers by contracts to the ACGT policies and procedures and will ensure compliance with the Data Protection regulations. In addition and from an ethical point of view it is strongly demanded to

let patients participate in and have a measure of influence over the processing of their genetic data.

To assure the success and functionality of the ACGT environment end-users are involved in every step of the development process.

2.1 Main Requirements

In summary, the user requirements that have been identified by the clinical experts for the ACGT environment can be divided into the following aspects:

- Appropriateness: the data analysis environment should provide the appropriate tools and services to support users in the state-of-the-art scientific analysis of biomedical data. Section 4.1 introduces the use of the GridR component [9], a "gridified" version of the well-known R statistical software, which is a de-facto standard for many kinds of biomedical data analysis.
- Extensibility and reusability: the platform should be easily extensible to new tasks and existing solutions should be easily reusable and transferable to similar problems. Extensibility is addressed in Section 4.1, while an important aspect of reusability, namely quality control, is described in Section 4.5
- Performance: the system must be performant enough to facilitate large analysis and optimization tasks, which calls for an efficient use of the grid architecture. Challenges exists not only because of the size of the data sets (see Section 4.2), but also from their complexity and heterogeneity (Section 4.3), which is a result of the distributed nature of pan-european clinical trials.
- Security: The system must be secure and protect the privacy of the involved patients. This is discussed in Section 4.4.
- Usability: the system should be easy to use for inexperienced users, but also provide a powerful interface for experts. Usability is best achieved by a continuous process of evaluation and optimization. In Section 4.6, approaches to automatically identify parts of the system that require a high amount of attention are discussed.

3 The ACGT Architecture

The complexity and the diversity of user requirements have a strong impact on the design of the ACGT architecture. It is evident that a multidisciplinary and multiparadigm approach is necessary in order to deal with these requirements. For these reasons the ACGT platform is designed according to the

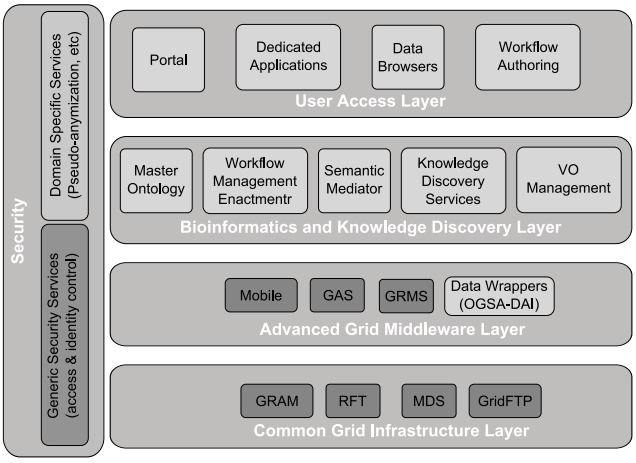


Fig. 2 The ACGT layered architecture

following technologies and standards: Service Oriented Architecture (Web Services), the grid, and the Semantic Web. In essence, the grid provides the computational and data storage infrastructure, the general security framework, the virtual organization abstraction and relevant user management mechanisms etc. The machine to machine communication is performed via XML programmatic interfaces over web transport protocols, which are commonly referred as Web Services interfaces. Finally the Semantic Web adds the knowledge representation mechanisms through the means of OWL ontologies, the implementation-neutral query facilities with the SPARQL "universal" query language and the associated query interfaces.

The adopted architecture for ACGT is shown in Fig. 2. A layered approach has been followed for providing different levels of abstraction and a classification of functionality into groups of homologous software entities. In this approach we consider the security services and components to be pervasive throughout ACGT so as to provide both for the user management, access rights management and enforcement, and trust bindings that are facilitated by the grid and domain specific security requirements like pseudonymization. Apart from the security requirements, the grid infrastructure and other services are located in the first (lowest) two layers: the Common Grid Layer and the Advanced Grid Middleware Layer. The upper layer is where the user access services, such as the portal and the visualization tools, reside. Finally, the Bioinformatics and Knowledge Discovery Services are the "workhorse" of ACGT and the corresponding layer is where the majority of ACGT specific services lie.

4 Addressing the User Requirements in ACGT

In the following, we will try to give a short overview of how to integrate the user requirements of Section 2 into the ACGT grid architecture.

4.1 Extensibility

The requirement for extensibility is very important in the context of grid-enabled data mining [11]. Especially, in order to keep track with new scientific developments, it is crucial to be able to quickly integrate new analysis services or algorithms into a data mining platform. Related to the ACGT environment, extensibility denotes the possibility of extending the environment at the workflow level, at the service level, or at the algorithm level. In order to deal with such requirements we have found that the use of metadata descriptions and the ontology based integration of the ACGT platform components provides a future proof approach to extensibility. In the following we will introduce GridR [9] as an example to demonstrate how the ACGT system can easily be extended by new services and algorithms.

GridR is an analysis tool based on the statistical environment R [3] that allows using the collection of methodologies available as R packages in a grid environment. The aim of GridR is to provide a powerful framework for the analysis of clinico-genomic trials involving large amount of data (e.g. microarray-based clinical trials). The GridR service (see Fig. 3) combines the wide spectrum of methods available in R with an effective distributed grid data management system (DMS) and efficient execution supported by a grid resource management system (GRMS), see [10]. In this fashion, users can make efficient use of distributed, parallel computational resources in their R scripts, while all the technical details are hidden from them. The R code to be executed can be given directly by the user in the form of a script, but in order to increase the possibility of distributing and re-using code, the intended way to execute R code is by storing it in a metadata repository, such that it becomes available to the whole system. Technically, an R function or script f thus becomes an $f-$service. Consequently, users who prefer to work on the workflow level and not edit their own code can make use of available R scripts and even all the single R functions in R libraries in their workflows.

Along these lines new algorithms can be "gridified" and be seamlessly integrated with the rest of the ACGT grid environment without a need for changing the service's or the R script's implementation.

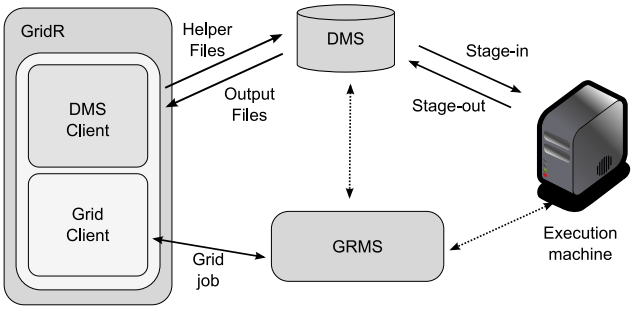


Fig. 3 GridR Architecture

4.2 Large Data Sets

Data is the most valuable asset of ACGT and therefore the platform should be able to manage big data sets in an efficient and secure way. The storage and the transfer of the data is of particular importance and something that should be taken care in a uniform way in the whole ACGT environment.

Data storage requirements are addressed by the grid infrastructure and the ACGT "Data Grid", which controls the sharing and the management of large amounts of distributed data. However, an additional issue has to do with the protocols, infrastructure, and policies for moving these large data sets to the processing nodes where the data analysis is performed. In some cases the grid infrastructure could be employed so that instead of moving the data around, the processing tasks, by the means of grid job submission and scheduling services, are transferred where the data reside. Nevertheless, the majority of services and data processing tools in ACGT are implemented as XML Web Services that are accessible through the network. Being a text format, XML is well known for its unfriendliness for transferring binary data. There are a couple of solutions for this ranging from encoding the binary data in hexadecimal or, most often, in Base64 text format, to using "attachments" in the SOAP messages. Nevertheless these approaches impose additional processing and bandwidth costs and so we opt for another option, which is to transmit *references to data* as part of the Web Services interaction while the data itself can be transferred through "out of band" channels, e.g. by the means of GridFTP. This approach offers the advantage of "quicker" XML interactions, easier and more performant service composition since there is no need to "get" (download) a huge binary data set in order to "give" (upload) it to another service, identity of the data so that they can associated with metadata through their references, etc.

4.3 Complex Data

A particular characteristic of data analysis in clinical trials is that the data used in a statistical analysis can be very heterogeneous and dynamic, meaning that many different tools and approaches may be necessary to analyze the data, but also that intermediate results may become invalid as the trial progresses and more data becomes available. The situation is further exacerbated when improved interactivity can result in changing workflows on the fly and cloning running workflows to explore alternatives in parallel. This involves the risk that the user becomes overwhelmed by the enormous amount of information and choices that are available to him. Hence, approaches to help the user better deal with the possibilities of the system are necessary.

For these reasons, a hyperlinked presentation of information has been proposed as a tool for better supporting the collaboration in scientific communities [6]. In essence, provenance information can be viewed as a graph of services invocations, with edges representing several types of lineage and provenance. For example, relationships such as "produced-by", "part-of", "derived-from", "input-of" etc. can be modeled this way. Each entity (e.g. service, data) is identified by an HTTP URI to provide identification, retrieval, and linking facilities for constructing a web of data and metadata in accordance with the Semantic Web vision [1]. Therefore in ACGT we aim to employ the semantic web technology in order to facilitate the tasks of both the users and the intelligent knowledge extraction services. Users are able to navigate to the information graph formed by the casual and other relationships between and among services and data just by following the hyperlinking paradigm that was popularized by the World Wide Web. On the other hand, semantic web enabled software entities are empowered to take advantage of the semantically rich content and to draw conclusions and knowledge based on the referenced ontologies.

4.4 Security

The sensitivity of the patient data requires a strong security framework to provide enough safety nets in order to maintain privacy, confidentiality, and integrity. The grid middleware already supports much of the necessary infrastructure, in terms of certificate based Grid Security Infrastructure (GSI), the Virtual Organization (VO) abstraction and the user credential management, and the Grid Authorization Services (GAS). In ACGT this "system level" security is complemented by "domain specific" mechanisms like *pseudomymization* that permits the identification of patient specific information without revealing the true person identity. All data is anomymized before their entry in ACGT and even during their analysis all the processing tasks are audited and authorized based on the end users' identity[2].

An interesting assertion about security of services can be made when using tool repositories as described in Section 4.1: with a standard web service, which can be deployed anywhere, it is principally impossible to give technical guarantees about which code is executed, as only the interface of the service is given and standardized. In general this is a desired property of web services, however, when considering data security, this means that external (legal) measures have to be taken to prohibit the service owner from disclosing information about the data. With the use of tool repositories, it can be guaranteed by a central instance, that the code in the repository has been reviewed and is secure to use, because the code that is being executed is directly transported to the execution site from the repository. In addition, this shipment of algorithms allows to analyze the data on a secure site, without needing to transport sensitive data.

4.5 Quality Control

In dynamic, distributed, and heterogeneous environments with multiple actors and complex use cases it is important to have a continuous validation of the different functional components. Therefore an ACGT validation and testing infrastructure is required to constantly monitor the ACGT services and report any malfunctions. This infrastructure for the automatic testing and validation of ACGT workflows and services is useful both for the initial decision making process about the acceptance of a new service and for the monitoring the status of the ACGT services as a whole. The status of ACGT services and workflows is checked with respect to the following criteria:

- Liveness, i.e. that it's "alive" and normally operating
- Correctness, i.e. that it delivers the correct results
- Performance, i.e. that it responds in a timely fashion

A number of tests are developed as scripts for each service according to these criteria. These tests are of course service and workflow specific because different components have different notions of correctness or performance. Nevertheless all of them are given some sample input data and parameters and based on this information they validate the target services according to the services' interface and functionality. The tests are stored centrally and re-evaluated periodically.

The advantage of this testing scenario is that even complex, user-defined workflows can be tested periodically, such that a single user can be notified if a workflow (i.e. a scientific experiment) of hers fails to meet the expected results. In this way, not only software quality, but also the quality of published, clinically relevant findings can be controlled.

4.6 Usability

Much thought in the ACGT project is given to the usability of the final software, including a formal usability analysis and end user integration throughout the runtime of the project to guarantee that the software will meet the requirements of the end users. Usability analysis is an important, but very time consuming process. In this section, we will present some approaches on how to improve the usability of the system using information present in the system's meta data.

The idea is that workflow execution statistics can be gathered together with other meta data and put into relation with the user's content with the system. For example, an analysis of workflows which are often canceled can provide the system's administrator with valuable information on how help users to select a better workflow. A statistic of the execution time of different services can help a developer to choose which services to optimize. A list of often used services can help new users to select good services.

Hyper-linking between meta data, workflow templates and workflow statistics also allow for a more complex reasoning of the users intent. One example could be as follows: the user executes different workflows on a data set, or variants thereof. From the meta data of the data set the system finds out that all the variants of the data set point to the same basic data set (e.g. a trial) and hence can reason that all the workflow executions belong together. It can then search the database of historic workflow executions to see whether a similar groups of workflow have been executed by another user. If this is the case, it is reasonable to assume that both users try to solve a similar problem, and hence the best workflow of the old user can be suggested to the new user. Of course, privacy aspects have to be considered in this kind of scenario.

5 Conclusions

There are a number of projects that aim at developing grid-based infrastructure for post-genomic cancer clinical trials, the most advanced of which are NCI's caBIG[2] in the USA and CancerGrid[3] in the UK. The overall approach in those projects is somewhat different from the one in ACGT. In caBIG, the bottom-up, technology-oriented, approach was chosen, in which the focus was put on the integration of a large number of analysis tools but with weak concern on data privacy issues. CancerGrid on the other hand addresses the very needs of the British clinical community. In contrast, the goal of the ACGT project is develop a pan-european system that is driven by current demands

[2] Cancer Biomedical Informatics Grid, https://cabig.nci.nih.gov/

[3] http://www.cancergrid.org/

from clinical practice. With two on-going international clinical trials actually conducted in the framework of the project, the approach is top-down, with clinicians' and biomedical data analysts' needs at the heart of all technical decisions, considering data privacy issues as central as data analysis needs.

In this user driven endeavor the technical concerns raised by the multiplicity and heterogeneity of user requirements demand state of the art methodologies and technologies. In the ACGT work plan the employment of ontologies and metadata annotations and the realization of intelligent higher level services are the primary implementation targets. Finally, in the realization of this environment, we aspire that the users are also involved. Guided and facilitated by the infrastructure, they can actively participate by creating and sharing information and knowledge. Only this way the ACGT is enriched and improved to become a really useful scientific tool.

Acknowledgments:

The authors wish to thank the ACGT consortium for their contributions and various ideas on which the ACGT project was developed. The ACGT project is funded by the European Commission (FP6/2004/IST-026996).

References

1. Berners-Lee, T.: Linked Data Design Issue. http://www.w3.org/DesignIssues/-LinkedData.html
2. B. Claerhout, N. Forgo, T. Krügel, M. Arning, and G. De Moor (2008): A Data Protection Framework for Transeuropean genetic research projects. Studies in health technology and informatics, vol. 141, 67.
3. R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org
4. Rüping, S., Sfakianakis, S., Tsiknakis, M. (2007): Extending Workflow Management for Knowledge Discovery in Clinico-Genomic Data. Proc. Healthgrid 2007: From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences 183–193
5. Couzin, J. (2007): Survival in young adults with Cancer. Science **317**
6. Goble, C.,Gocho, O., Alper, P., De Roure, D. (2006): e-Science and the Semantic Web: A Symbiotic Relationship. 9th Int. Conf. on Discovery Science (DS2006), Springer , 1–12.
7. Graf, N., Weiler, G., Brochhausen, M., Scherer, F., Hoppe, A., Tsiknakis, M., Kiefer, S., Aran Lunzer, Yuzuru Tanaka (2007) : The importance of an ontology based clinical data management system (OCDMS) for clinico-genomic trials in ACGT. SIOP 2007, 39th International Society Of Pediatric Oncology Annual Meeting, Mumbai, India
8. Weiler, G., Brochhausen, M., Graf, N., Hoppe, A., Schera, F., Kiefer, S. (2007): Ontology Based Data Management Systems for post-genomic clinical Trials within an European Grid Infrastructure for Cancer Research. SIOP 2007, 39th International Society Of Pediatric Oncology Annual Meeting, Mumbai, India

9. Wegener, D., Sengstag, T., Sfakianakis, S., Rüping, S., Assi, A. (2007): GridR: An R-based grid-enabled tool for data analysis in ACGT clinico-genomic trials. Proc. 3rd Intl. Conf. on e-Science and Grid Computing (eScience 2007), Bangalore, India
10. Pukacki, J., Kosiedowski, M., Mikolajczak, R., Adamski, M., Grabowski, P., Jankowski, M., Kupczyk, M., Mazurek, C., Meyer, N., Nabrzyski, J., Piontek, T., Russell, M., Stroinski, M., Wolski, M. (2006): Programming Grid Applications with Gridge. Computational Methods in Science and Technology **12**.
11. Wegener, D., May, M. (2007): Extensibility of Grid-Enabled Data Mining Platforms: A Case Study. Proc. of the 5th International Workshop on Data Mining Standards, Services and Platforms, San Jose, California, USA, 13–22.

Discovery of Genotype-to-Phenotype Associations: A Grid-enabled Scientific Workflow Setting

Lefteris Koumakis, Stelios Sfakianakis, Vassilis Moustakis, and George Potamias

Institute of Computer Science, FORTH

Abstract. The heterogeneity and scale of the data generated by high throughput genotyping association studies calls for seamless access to respective distributed data sources. Toward this end the utilization of state of the art data resource management and integration methodologies such as Grid and Web Services is of paramount importance for the realization of efficient and secure knowledge discovery scenarios. In this paper we present a Grid-enabled Genotype to Phenotype scenario (GG2P) realized by a respective scientific workflow. GG2P supports seamless integration of clinico-genetic heterogeneous data sources, and the discovery of indicative and predictive clinico-genetic models. GG2P integrates distributed (publicly available) genotyping databases (ArrayExpress) and utilizes specific data-mining techniques for feature selection – all wrapped around custom-made Web Services. GG2P was applied on a whole-genome SNP-genotyping experiment (breast cancer vs. normal/control phenotypes). A set of about 100 discriminant SNPs were induced, and classification performance was very high. The biological relevance of the findings is strongly supported by the relevant literature.

1 Introduction

Scientific community experiences an increasing need for efficient data management and analysis tools and there is an unprecedented demand for extraction and processing of knowledge. This is more than evident in the domain of bioinformatics since the beginning of the "genomic revolution". After the completion of the Human Genome Project and the emergence of high throughput technologies (DNA microarrays, high-density SNP genotyping, mass spectrometry etc) a vast amount of biological data are being produced on a daily basis. This has raised the expectation of extracting valuable knowledge for post-genomic personalized disease treatment. Therefore new challenges for the data analysis and knowledge discovery processes are introduced.

Knowledge Discovery and Data Mining are the most prominent methods and tools for the state of the art scientific discovery. Requirements for biological data management are very demanding due to size and complexity, quality properties (missing values or noisy data are frequent), and inherent domain heterogeneity. These new requirements have given rise to modern software engineering methodologies and tools, such as Grid (Foster 2003) and Web Services (Curbera et al 2002). These new technologies aim to provide the means for building sound data integration, management and processing frameworks.

This paper presents an integrated scenario to support seamless access and analysis of Single Nucleotide Polymorphisms (SNP) genotype data, as produced by relative SNP genotyping platforms. Effort is cast toward the discovery of reliable and predictive multi-SNP profiles being able to distinguish between different phenotypes. The employed data-mining technique is founded on a novel feature selection algorithm. The whole approach is realized in a Grid-enabled scientific (BPEL-compliant – BPEL stands for Business Process Execution Language) workflow editor and enactment environment, and presents an integrated scenario aiming to support Grid-enabled Genotype-to-Phenotype (GG2P) association studies. In particular, GG2P seamlessly accesses and gets phenotypic and genotypic SNP data; analyzes them; and presents results (e.g, the most discriminant and descriptive SNPs) in an appropriately devised html file with links to the Ensembl genome browser.

2 Enabling Technology

With the completion of the human genome and the entrance into the post-genomic era the large amount of data produced makes difficult to extract and evaluate the hidden information without the aid of advanced data analysis techniques. Data mining has successfully provided solutions for finding information from data in many fields including bioinformatics. Many problems in science and industry have been addressed by data mining methods and algorithms such as clustering, classification, association rules and feature selection. In particular, feature selection is a common technique for gene/SNP feature reduction and selection in bioinformatics. It is based on data mining technique for selecting a subset of relevant features and building robust predictive models. The main idea is to choose a subset of input features by eliminating those that exhibit limited predictive performance. Feature selection can significantly improve the comprehensibility of the resulted classifier models and support the development of models that generalizes better to unseen cases.

The heterogeneity and scale of clinico-genetic data raises the demand for: (a) seamless access and integration of relevant information and data sources, and (b) availability of powerful and reliable data analysis operations, tools and services. The challenge calls for the utilization and appropriate customization of high performing *Grid*-enabled infrastructures and Web technology - as presented by *Web*

Services, and *Scientific Workflows* environments. Smooth harmonization of these technologies and flexible orchestration of services present a promising approach for the support of integrated genotype-to-phenotype association studies.

Grid technology. Grid computing (Foster 2003) is a general term used to describe both hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities. Grid has emerged as the response to the need for coordinated resource *sharing* and problem solving in dynamic, multi-institutional *virtual organizations*. Sharing of computers, software, data, and other resources is the primary concern of Grid architectures. In a modern service oriented architecture the Grid defines the general security framework (e.g. the authentication of the users and services), the virtual organization abstraction, the user management mechanisms, authorization definition and enforcement, etc. It provides both the computational and the data storage infrastructure, which is required for the seamless management and processing of large data sets.

*Semantic and Knowledge Grid*s. Semantic Grid presents a Grid computing approach in which information, resources and data processing services are employed with the use of semantics and respective data models. It facilitates the discovery, automated linkage and smooth harmonization of services. In a Semantic Web analogy, Semantic Grids can be defined as "*extensions of current Grids in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation*" (De Route et al 2005). Encapsulation of Web Science and knowledge-oriented technologies in Grid-enabled infrastructures represents a flexible knowledge-driven environment referred as the Knowledge Grid (Zhuge 2004). In their layered architecture organization, Knowledge Grids define and form an additional layer, which supports implementation of higher level and distributed knowledge discovery services on a virtual interconnected environment of shared computational and data analysis resources. This setting permits and enables: automated discovery of resources; representation, creation and management of statistical and data mining processes; and composition of existing data and processing resources in 'compound services packages' (Cannataro and Talia 2003).

Web services. The Web Services suite of standards presents the most popular and successful integration methodology approach. Based on Web Services standards the machine-machine communication is performed via XML programmatic interfaces over web transport protocols (e.g., SOAP), which are specified using the Web Service Definition Language (WSDL) (Curbera et al 2002). These common data representation and service specification formats, when properly deployed, enable the integration of heterogeneous and geographically disparate software systems. Web Services enhance and support the development of distributed, multi-participant, and interoperable systems that can be utilized in the combination of services and their reuse as processing steps into more complex high level scenarios, commonly referred as workflows.

Scientific workflows. The Workflow Management Coalition (WFMC, www.wfmc.org) defines a workflow as "the automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules". A workflow consists of all the steps and the orchestration of a set of activities that should be executed in order to deliver an output or achieve a larger and sophisticated goal. In essence a workflow can be abstracted as a composite service, e.g. a service that is composed by other services that are orchestrated in order to perform some higher level functionality. The (potentially parallel) steps (tasks) that a workflow follows may exhibit different degrees of complexity, and are usually connected in a non-linear way, formulating a directed acyclic graph (DAG). A Workflow Management System defines, manages and executes workflows through the execution of software that is driven by a computer representation of the workflow logic (Deelman et al 2006, Fox and Gannon 2006).

In addition to the business oriented use cases, workflows have a lot of potential in scientific areas as well. In a lot of scientific sectors, the demand is put not only on the computational power but on the complex structure of the inter-dependable tasks to be performed. Sophisticated problem-solving engages a variety of inter-depended data analysis tasks and analytical tools, e.g., pre-processing and re-formatting of heterogeneous datasets into formats suitable as input to other analytic process. Moreover, large-scale scientific computations involve much of intervention, as in the case of the interpretation of intermediate results by domain experts. But, at some stage of the process just normal personnel could be engaged. So, the rights and roles of involved persons should be explicitly defined. In addition, the computational environment itself is heterogeneous, ranging from supercomputers to clusters of personal computers. So, there is a need to model and explicitly define the engaged computational nodes and networks. Scientific workflows are introduced as an amalgamation of scientific problem-solving and traditional workflow techniques. They have been proposed as a mechanism for coordinating processes, tools, and people for scientific problem solving purposes and aim to support "coarse-granularity, long-lived, complex, heterogeneous, scientific computations" (Singh and Vouk 1997).

To assist the bioinformatics community in building complex scientific workflows, and in the context of the EU FP6 integrated project (www.eu-acgt.org), the ACGT Workflow Editor and Enactment Environment (WEEE) have been designed and developed (Sfakianakis et al 2009). WEEE is a Web-based graphical tool that allows users to combine different Web Services into complex workflows, and it is accessible through the ACGT Portal. It supports searching and browsing of a Web Services repository and of respective data sources, as well as their orchestration and composition through an intuitive and user friendly graphical interface. Created workflows can be stored in user spaces and can be later retrieved and edited. So, new versions of them can be easily produced. Designed workflows can be executed in a remote machine or even in a cluster of machines in the Grid. In this way there is no burden imposed on the user's local

machine since the majority of computation and data transfer of the intermediate results are take place in the Grid where the services are executed. Publication and sharing of the workflows are also supported so that the user community can exchange information and users benefit from each other's research. WEEE is based on the BPEL (Arkin et al 2005) workflow standard and supports the BPEL representation of complex bioinformatics workflows.

The ACGT Grid environment is supported by the Gridge toolkit (www.gridge.org/) – an open source software platform, compatible with the Globus toolkit (www.globus.org) aimed to help users to deploy ready-to-use grid middleware services and create productive Grid infrastructures. All Gridge Toolkit software components have been integrated together and form a consistent distributed system following the same interface specification rules, license, and quality assurance and testing (Pukacki et al 2006).

The GG2P scenario presented in this paper is enabled by the smooth integration of components from the aforementioned technologies. GG2P aims to seamlessly integrate and mine distributed and heterogeneous clinical and genotype data sources using: (i) existing public-domain and custom-made Web Services for accessing remote and distributed genotype and phenotype data sources, and for downloading the targeted experiments and the respective data annotation (XML) files; (ii) specially devised Web Services to extract relevant information and raw data, including appropriate data pre-processing and re-formatting operations; and (iii) specially suited for G2P association studies data mining processes wrapped as Web Services. In addition, the results (profiles of specific SNPs) are automatically linked with state-of-the-art genome browsers (e.g., Ensembl), and are appropriately visualized.

3 The GG2P scenario

An SNP is a single base substitution of one nucleotide with another. With high-throughput SNP genotyping platforms massive genotyping data may be produced for individual samples (i.e., diseased, treated or, control). It is known that a category of diseases are associated to a single SNP or gene (also known as monogenic diseases). In general, a single SNP or gene is not informative because a disease may be caused by completely different modifications of alternative pathways in which each SNP makes only a small contribution. Most of the complex diseases, including cancer, are characterized by groups of genes with a number of susceptible genes interacting with each other. It's important to search for multiple SNP profiles - among a huge number of them, that not only associate with a disease but exhibit a high discrimination power between different phenotypic classes. The GG2P scenario aims exactly towards this direction with the relevant literature started to include similar approaches (Nunkesser et al 2007, Zhou and Wang 2007, Schwender et al 2008). The steps followed by the corresponding scientific workflow are presented and described in the sequel.

Data access and retrieval. Using Web Services from the European Bioinformatics Institute's (EBI) repository (http://www.ebi.ac.uk/Tools/webservices/) we access and extract phenotypic and genotypic data from public experiments. Specifically, using specific ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) Web Services we may get information about a specific experiment or, get information about relevant experiments using keywords. The complete SNP array dataset used in this study is available on the NCBI GEO database under accession no. GSE3743. The dataset refers to a genotyping experiment of 78 sample hybridizations performed on the Affymetrix GeneChip Human Mapping 10K Array Xba 131 (Mapping10K_Xba131) array design. The raw data file includes 78 transformed and/or normalized data files. The hybridized samples concern breast cancer (BRCA) and normal (CTRL) cases. More information about the dataset can be found at (Richardson et al 2006). Note that GG2P could be easily customized to work with other experiments and respective datasets.

Data mediation. The response of ArrayExpress web service is an XML file with links to phenotypic (via the 'sdrf' tag) and genotype (via the 'fgem' or 'raw' tags) experimental data (see Fig 3.1 for a sample of the XML response file). We utilized a special parser to extract the needed information from the XML file.

```
<experiment total-assays="78" total-samples="78" total="1" revision="080925" ver-
sion="1.1">
<experiment>
<id>1627324147</id>
<accession>E-GEOD-3743</accession>
<name>Genotyping of human breast tumors</name>
<samples>78</samples>
…
<files>
<raw celcount="78" count="78" name="E-GEOD-3743.raw.zip"/>
<fgem count="78" name="E-GEOD-3743.processed.zip"/>
<idf name="E-GEOD-3743.idf.txt"/>
<sdrf name="E-GEOD-3743.sdrf.txt"/>
<biosamples>
<png name="E-GEOD-3743.biosamples.png"/>
<svg name="E-GEOD-3743.biosamples.svg"/>
</biosamples>
</files>
</experiment>
</experiments>
```

Fig. 3.1. Part of Web Service XML response file (from ArrayExpress)

The parser locates the 'samples', 'sdrf' and 'fgem' tags. The 'samples' tag identifies the number of included samples/hybridizations, and the 'sdrf' tag points to the respective file with description of each hybridization. From the 'fgem' tag we may identify and download the SNP profiles of the respective experiment's samples. It is essential to align phenotypic classes with the respective samples'/hybridizations' genotype data, and form a unified dataset to be analyzed. We employ a natural-language mechanism, enabled by specific ontologies and controlled vocabularies (Potamias et al 2005). The result is a homogenized and appropriately formatted file (with phenotype class annotations and respective genotype data), which serves as input to a specific analytical process.

Data preprocessing. Depending on the data and the data mining algorithm, the formed data file may need extra processing. For example, many algorithms can handle only nominal values. In such a case, and if the data comes with continuous feature values, we have to discretize them. Furthermore, as genotype profiling platforms (like Affymetrix) produce too many 'NoCalls', one may be also interested to reduce these 'missing values' utilizing an appropriate data pre-processing process. After the needed pre-processing are performed, the 'filtered' dataset is transformed into the ARFF format - a de facto standard for machine learning. ARFF supported by the Weka machine learning package (http://www.cs.waikato.ac.nz/ml/weka/) (Witten and Frank 2005).

Data analysis. A variety of existing data mining algorithms exists in the public domain (e.g., Weka, R-package/Bioconductor, BioMoby). Here we rely on a feature reduction and selection approach. Dimensionality reduction and feature selection is a well-known and addressed issue in machine learning and data mining (Guyon and Elisseeff 2003). We are interested on the identification of SNP-phenotypic class associations, and on respective discrimination/classification models. The profiles of these SNPs are able to distinguish between particular pre-classified patient samples. Core operations of this process are implemented in the MineGene gene selection system, and their Web Services deployment (Potamias et al 2004, Potamias et al 2006).

3.1 GG2P in action

For the realization of GG2P scenario we used part of the ACGT Grid infrastructure – the Data Management System, the service repository and the workflow editing and execution environment. The Data Management System (DMS) is a secured and distributed file system over the Grid. The service repository gives access rights as well as metadata information about the available services. The workflow editor is a Web2 application and, as already mentioned, the workflow enactor is a BPEL-compliant application installed in a Grid node. Fig. 3.2 introduces the GG2P knowledge discovery scenario as implemented in the context of the ACGT WEEE workflow editing and execution environment. The Web Servic-

es (not shaded shapes in the workflow area of Fig. 3.2) are registered in the ACGT services repository.

The ACGT environment requires authorization from the DMS and the services repository. DMS grand permissions to user's account in the Grid and services repository give access to available services. Then the user composes and draws he desired workflow. At the next step the editor translates (or compile) the graphical workflow into BPEL. Finally, the enactment of the workflow may start. The first web service takes as input a query (first, from left, shaded shape of Fig. 3.2) and returns an XML file with information about all the related to the query experiments in the EBI ArrayExpress repository. For the specific scenario we used a query with the keywords "homo sapiens" & "breast cancer" & "genotype" & "affymetrix" & "Mapping10K_Xba131".

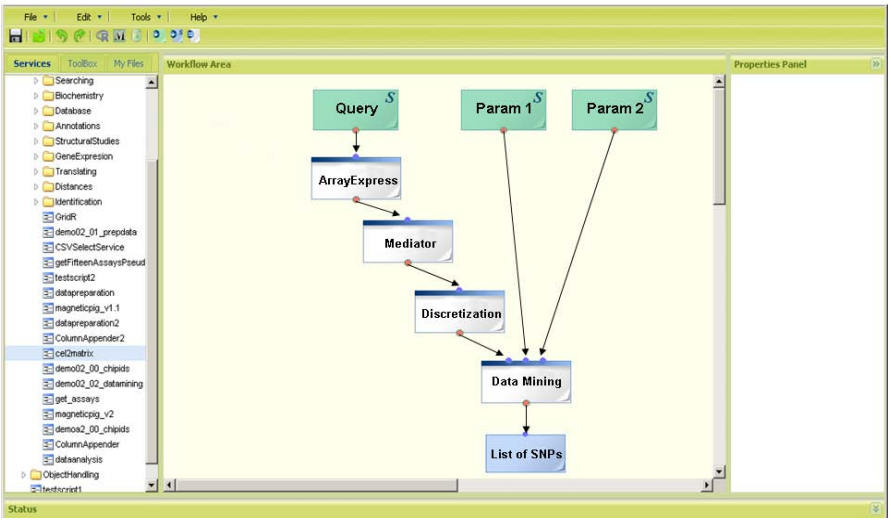


Fig. 3.2. The GG2P scientific workflow as implemented in ACGT's Workflow Editor and Enactment Environment (WEEE). Web services include: ArrayExpress, Mediator, Discretization, and Data Mining. Services are activated by a Query (top part). Deployment of Data Mining also needs specification of parameters ('Param 1' and 'Param 2')

The second service (Mediator) takes as input the repository's XML response file and creates the homogenized file with the clinical and genotype data. The generated file is stored in DMS at the user's account. The next service (Discretization) discretizes and transforms the experiment data to arff format. Discretization service retrieves the data from DMS and stores the arff-formatted data back to the DMS. The final service implements the (two-valued) SNP feature selection algorithm. The service again retrieves data from DMS and stores the results in the DMS. Then, after the editor requests the results from the DMS, SNP annotations and links to the Ensembl genome browser are automatically assigned to the se-

lected SNPs. Finally, an html file is formed and is used for the visualization of results (see Fig. 4.1).

4 Results and Discussion

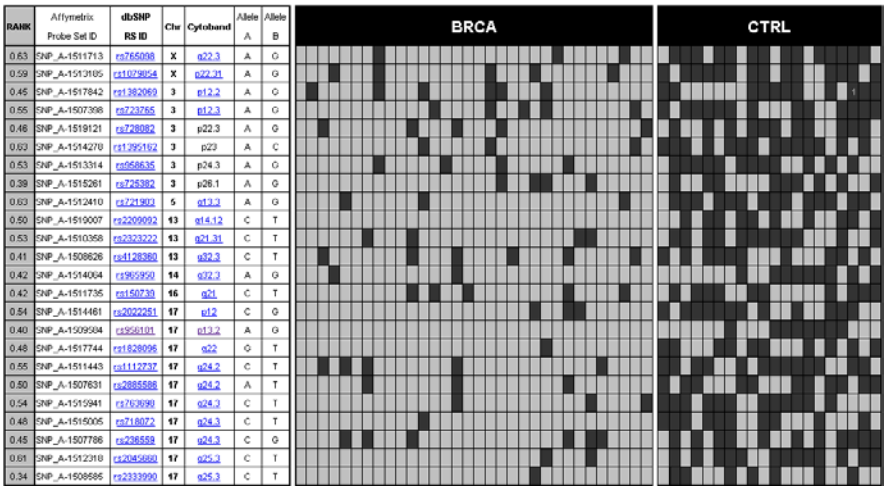
The Affymetrix SNP genotyping platforms produce processed data files where, each SNP receives three different values: AA and BB that represent paternal or maternal homozygosity statuses, respectively, and AB for heterozygosity ones. The '0' and '1' nominal values are assigned to the AA/BB and AB SNP feature values, respectively. This results into a two-valued feature representation space. In this setting a set of SNPs could be considered as an ideal discriminator between two different phenotypic classes if it displays the '0' value for all sample cases in one class and the '1' value for all sample cases in the other class. From the total of the 78 sample cases included in the target SNP genotyping experiment we excluded the ones that have more than 10% of missing 'NoCall' values, resulting into a dataset of 36 BRCA and 36 CTRL cases.

For the target BRCA vs. CTRL study, the execution of the GG2P scientific workflow resulted into a set of about 100 most discriminant SNPs. With these SNPs the following highly performing figures are achieved: 96.2% accuracy, 92.2% sensitivity, 96.2% specificity, and 0.979 ROC/AUC.

Fig. 4.1 visualizes just the top 24 of them with the highest ranks (for those sample cases with no 'NoCall' SNP values) sorted by their chromosomal location. The first column shows the discrimination power (the rank) for each SNP (as calculated by MineGenes' core feature selection process). The second column shows the Affymetrix code name for the probe that represents the respective SNP. The third column displays the corresponding code, namely: dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/). The dbSNP - SNP databases, represent a widely used public-domain archive for a broad collection of SNPs as well as small genomic insertion/deletions (indels) and is hosted at the National Center for Biotechnology Information (NCBI). The next three columns display information about the genomic region of the respective SNP: column four the chromosomal location; column five the cytoband, and columns five and six the nucleotide allele variations for the two (paternal/maternal) alleles. The last column shows the nearest gene present in the corresponding SNP's genomic physical position.

All hyperlinks are automatically assigned to the respective items by consulting the annotation files provided by Affymetrix. When clicking on a specific cytoband one is transferred to the respective visualization screen of the Ensembl genome browser (www.ensembl.org). So, inspection of results and further investigation is enabled and supported. In Fig 4.1 one may also observe and contrast the SNP characteristic profile patterns between BRCA and CTRL cases, respectively - gray and dark shaded cells represent homozygosity ('AA/BB') and heterozygosity ('AB') statuses, respectively.

The main observation is that the homozygosity patterns are dominant in the BRCA cases - a finding which is consistent with the **Loss of Heterozygosity** (LOH) situation in pathogenic situations. LOH in a cell represents the loss of regular function of one of the gene's alleles when the other allele is inactive. In oncology, LOH refers to somatic mutations and occurs when the offsping's functional allele is inactivated by the mutation. In such situations, normal tumor suppressor functionality is inactivated and tumorigenesis events are almost certain.



RANK	Affymetrix Probe Set ID	dbSNP RS ID	Chr	Cytoband	Allele A	Allele B
0.63	SNP_A-1511713	rs765098	X	q22.3	A	G
0.59	SNP_A-1513105	rs1079854	X	q22.31	A	G
0.45	SNP_A-1517842	rs1382069	3	p12.2	A	G
0.55	SNP_A-1507398	rs723765	3	p12.3	A	G
0.46	SNP_A-1519121	rs728082	3	p22.3	A	G
0.60	SNP_A-1514278	rs1395162	3	p23	A	C
0.53	SNP_A-1513314	rs968635	3	p24.3	A	G
0.39	SNP_A-1515261	rs725382	3	p26.1	A	G
0.60	SNP_A-1512410	rs721903	5	q13.3	A	G
0.50	SNP_A-1519007	rs2209092	13	q14.12	C	T
0.53	SNP_A-1510358	rs2323222	13	q21.31	C	T
0.41	SNP_A-1508626	rs4128380	13	q32.3	C	T
0.42	SNP_A-1514064	rs1965950	14	q32.3	A	G
0.42	SNP_A-1511735	rs150730	16	q21	C	T
0.54	SNP_A-1514461	rs2022251	17	p12	C	G
0.40	SNP_A-1509584	rs956101	17	p13.2	A	G
0.48	SNP_A-1517744	rs1828096	17	q22	G	T
0.55	SNP_A-1511443	rs1112737	17	q24.2	C	T
0.50	SNP_A-1507631	rs2885588	17	q24.2	A	T
0.54	SNP_A-1515941	rs763690	17	q24.3	C	T
0.48	SNP_A-1515005	rs718072	17	q24.3	C	T
0.45	SNP_A-1507786	rs236559	17	q24.3	C	G
0.61	SNP_A-1512318	rs2045880	17	q25.3	C	T
0.34	SNP_A-1508686	rs2333990	17	q25.3	C	T

Fig. 4.1. The induced most discriminant and highest ranked BRCA vs. CTRL SNPs (for the ArrayExpress E-GEOD-3743 genotyping experiment) – gray shaded and dark shaded cells indicate homozygosity and heterozygocity statuses, respectively. It can be easily observed that LOH (Loss Of Heterozygosity) patterns dominate the BRCA cases

We further examined the biological relevance of the findings, i.e., does the identified and most discriminant SNPs relate to LOH and breast cancer situations. Literature search provide us with strong evidence for that. We refer to just two indicative SNPs in cytobands 17p13.2 1nd 17p12 (both highly ranked). Chromosome 17p is among the most frequently deleted regions in a variety of human malignancies including breast cancer. In (Seitz et al 2001) the localization of a putative tumour suppressor gene (TSG) at 17p13, distal to the TP53 (the most indicative tumor suppressor) gene, was further refined for breast carcinomas. It was found that 73% (37 of 51) of the breast tumors exhibited loss of heterozygosity (LOH) at one or more loci at 17p13. The allelic loss patterns of these tumours suggest the presence of at least seven commonly deleted regions on 17p13. The three most frequently deleted regions were mapped at chromosomal location 17p13.3 - 17p13.2. Furthermore, the data suggest that different subsets of LOH in this region are associated with more aggressive tumor behavior. Additional evidence for the association between the 17p13 genomic region and breast cancer are also reported in (Mao et al 2005) and (Ellsworth 2003). Similar findings are re-

ported for the 17p12 region. In (Shen et al 2000) sixty-three markers are reported that display ≥25% LOH, with the highest values being observed on 17p12 (48.4% for the well, and ~87% for the poorly differentiated breast tumor cases).

5 Conclusions and Future Work

We presented an integrated methodology that enables the discovery of genotype-to-phenotype associations and predictive models, and supports G2P association studies. The methodology is realized in the context of the GG2P scenario being implemented with the aid of Web Services and Scientific Workflows and operating in a grid environment. In particular the ACGT (EU FP6 integrated project) Grid infrastructure and its WEEE workflow editing and enactment environment were utilized.

The GG2P workflow was executed on an indicative SNP genotyping experiment (from the ArrayExpress repository) that concerns the hybridization breast cancer and normal/control tissue samples. We were able to identify about 100 indicative SNPs that exhibit contrasted homozygosity / heterozygosity profiles, and achieve highly discriminant performance figures for the respective phenotypic classes. The most highly ranked SNPs exhibit clear loss of heterozigosity patterns, a common situation in tumorgenesis. Literature searches provide strong evidence about the biological relevance of the findings – the respective SNP's genomic regions are strongly association with characteristic breast cancer phenotypes.

Our immediate R&D plans, among other, include: experimentation with other public-domain genotyping experiments, and enrichment of GG2P and its workflow realization with other data-mining techniques (e.g., clustering, association rules mining etc).

Acknowledgements: This work is partially supported by the European Commission's Sixth and Seventh Framework Programme in the context of the ACGT (FP6-2005-IP-026996) and GEN2PHEN (FP7 HEALTH-F4-2007-200754) Integrated projects, respectively.

References

Arkin A et al (Eds.) (2005) Web services business process execution language. Version 2.0. http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html Accessed 8 March 2009

Cannataro M, Talia D (2003) The Knowledge Grid. Communications of the ACM 46(1):89 –93, 2003

Curbera F et al (2002) Unraveling the web services web: An Introduction to SOAP, WSDL, and UDDI. IEEE Internet Computing 6(2):86–93

De Roure D, Jennings NR, Shadbolt NR (2005) The Semantic Grid: Past, Present, and Future. Proceedings of the IEEE 93(3):669–681

Deelman E, Zhao Z, Belloum A (Eds.) (2006) Scientific Programming Journal, special issue on workflows to support large-scale science 14(3–4)
Ellsworth EE et al (2003) High-Throughput Loss of Heterozygosity Mapping in 26 Commonly Deleted Regions in Breast Cancer. Cancer Epidemiology Biomarkers & Prevention 12:915–919
Foster I (2003) The Grid: Computing without bounds. Scientific American 288(4):60–67.
Fox G, Gannon D. (Eds.) (2006) Concurrency and Computation: Practice and Experience, Special Issue on Workflow in Grid Systems 18(10)
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. JMLR: Special Issue on Variable and Feature Selection 3:1157–1182
Mao X et al (2005) Genetic losses in breast cancer: toward an integrated molecular cytogenetic map. Cancer Genetics and Cytogenetics 160(2):141–151
Nunkesser R, Bernholt T, Schwender H, et al (2007) Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. Bioinformatics 23(24): 3280–3288
Potamias G, Koumakis L, Moustakis VM (2005) Enhancing Web Based Services by Coupling Document Classification with User Profile. International Conference on Computer as a Tool (EURCON 2005) 1:205–208
Potamias G, May M. Ruping S (2006) Grid-based Knowledge Discovery in Clinico-Genomic Data. Lecture Notes in Bioinformatics (LNBI) 4345:219–230
Potamias G., Koumakis L, Moustakis V (2004) Gene Selection via Discretized Gene Expression Profiles and Greedy Feature-Elimination. Lecture Notes in Artificial Intelligence (LNAI) 3025:256–266, (2004)
Pukacki J et al (2006) Programming Grid Applications with Gridge, Computational Methods in Science and Technology 12(1):47–68
Richardson A et al (2006) X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 9(2):121–32
Schwender H, Ickstadt K, Rahnenführer J (2008) Classification with high-dimensional genetic data: assigning patients and genetic features to known classes. Biometrical Journal 50(6):91 – 926
Seitz S et al (2001) Detailed deletion mapping in sporadic breast cancer at chromosomal region 17p13 distal to the TP53 gene: association with clinicopathological parameters. Journal of pathology 194(3):318–326
Sfakianakis S, et al (2009) Web-based Authoring and Secure Enactment of Bioinformatics Workflows. 4th International Workshop on Workflow Management (ICWM2009), Geneva, Switzerland
Shen C-Y et al (2000) Genome-wide Search for Loss of Heterozygosity Using Laser Capture Microdissected Tissue of Breast Carcinoma: An Implication for Mutator Phenotype and Breast Cancer Pathogenesis. Cancer Research 60:3884–3892
Singh MP, Vouk MA (1997) Scientific workflows: Scientific computing meets transactional workflows.
http://people.engr.ncsu.edu/mpsingh/papers/databases/workflows/sciworkflows.html Accessed 8 March 2009
Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques (2nd Edition), Morgan Kaufmann, San Francisco
Zhou N, Wang L (2007) Effective selection of informative SNPs and classification on the HapMap genotype data. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid= 2245981. Accessed 8 march 2009
Zhuge H (2004) The Knowledge Grid. World Scientific Singapore

Cross-platform integration of transcriptomics data

Georgia Tsiliki, Marina Ioannou and Dimitris Kafetzopoulos

Abstract An increasing number of studies have profiled gene expressions in tumor specimens using distinct microarray platforms and analysis techniques. With the accumulating amount of microarray data, one of the most challenging tasks is to develop robust statistical models to integrate the findings. This article reviews some recent studies on the field. We also study the intensity similarities between data sets derived from various platforms, after appropriate rescaling of the measurements. We found that intensity and fold-change variability similarities between different platform measurements can assist the analysis of independent data sets and can produce comparable results with those obtained for the independent data set alone.

1 Introduction

With the increasing availability of published microarray data sets there is a need to develop approaches for validating and integrating results across multiple studies. The overlap of gene expression signatures of various studies is very small, for example between the "Amsterdam" signature [23] and the "Rotterdam" signature [24], mainly due to the small sample sizes of individual studies and error measurements. A major concern in the "meta-analysis" of DNA microarrays is the lack of a single standard experimental platform for data generation. The microarray technologies

Georgia Tsiliki
FORTH, Institute of Molecular Biology & Biotechnology, P.O. Box 1385, 711 10, Heraklion, Greece, e-mail: tsiliki@imbb.forth.gr

Marina Ioannou
FORTH, Institute of Molecular Biology & Biotechnology, P.O. Box 1385, 711 10, Heraklion, Greece, e-mail: mioannou@imbb.forth.gr

Dimitris Kafetzopoulos
FORTH, Institute of Molecular Biology & Biotechnology, P.O. Box 1385, 711 10, Heraklion, Greece, e-mail: kafetzo@imbb.forth.gr

currently in use differ in how DNA sequences are laid on the array, the length of these sequences, splicing variations and the number of samples measured in each hybridization. As a result, an important source of technological variability in gene expression measurements is the platform used.

The increasing number and availability of large-scale gene expression studies of human and other organisms provide strong motivation for cross-study analyses that combine existing and/or new data sets. In a cross-study analysis, the data, relevant test statistics or conclusions of several studies are combined. Several studies have compared measurements across platforms [9] and reported their findings in terms of reproducibility of results, power increase of studies, validation of gene signature results [10] [8] [11] [25]. The MAQC Quality Control Consortium, the FDAs Critical Path Initiative, NCIs caBIG and others are implementing procedures that will broadly enhance data quality. The MAQC consortium have reported that proper sample preparation is sufficient to dramatically enhance multi-lab and multi-platform correlations [16].

However, combining data from different expression studies and possibly different gene expression platforms poses a number of statistical difficulties due to the different processing facilities. As a consequence, measurements from different platforms cannot be directly combined. Identifying and removing such systematic effects is the primary statistical challenge in cross-study analysis. We note that technological differences between studies may be confounded with biological differences arising from the choice of patient cohorts (e.g. age, gender or ethnicity). In many cases, technological artifacts are dominant, though care should be taken to verify this, and one can hope to remove them while leaving biological information intact.

Here we briefly review some recent techniques to minimize error measurements and safely combine results of studies which address the same biological questions. Furthermore, we evaluate how the direct use of intensity data from independent data sets and platforms, can facilitate the statistical analysis of other microarray studies. An advantage of such an approach is that the same methodology can be used and the measurement errors can be controlled in the same way for all data sets. Our scope is to demonstrate that the power of any statistical conclusions can be retained when the data is enhanced with external data from various platforms. For that purpose our working example is the classification of ER samples in a breast cancer data set.

1.1 Recent literature review

In general, it will make sense to combine data sets of studies which address the same questions, or, experiments with some sufficiently similar aspects so that one can hope to make better inference from the whole than from the experiment separately. However, in order to compare experiments that are performed on different gene expression platforms, the first thing one should look at is how to link oligonucleotide probe sets, spotted sequences, and other microarray features. Typically, a sequence-specific identifier (GenBank accession number) serves as a reference to the array

probe sequences. Thus, the first step in a cross-study analysis would be to identify a subset of genes which are consistently measured across platforms. The next step would be to derive for each individual data set numerically comparable quantities from the expression values of genes in the common list by applying specific data transformation and normalization methods.

The most simple approach to integrate data would be to sample standardize and gene median center each available data set, and then combine data sets. More systematic approaches have been proposed for integration of findings from multiple studies using different array technologies. Particularly, according to [14], there are several potential approaches to cross-study analysis, depending on what information is being synthesized. Existing studies either combine information from primary statistics (such as t-statistics or p-values) [13] [19] or secondary statistics (such as gene lists) that are derived from the individual studies [3]. Additionally, other approaches to meta-analysis of gene-expression data are considered by [4] [15] [12], which directly integrate the data and then proceed with the analysis.

[22] proposed optimization methods for cross-laboratory and cross-platform microarray expression data, based on three simple and often employed techniques to identify discrepancy in expression data sets. They created an experimental design that compared three functionally different normal tissues: human liver, lung and spleen. Particularly, they reported that when precision, biological interpretation and multiple platform data sets were considered together, they allowed for better selection of genes with respect to a particular outcome. They considered precision and sensitivity measurements which were useful in finding the minimal detectable fold-change and raw performance values for an array platform. Also, Gene Ontology and pathway analyses were considered, which were thought to be a valuable way of examining and comparing the actual biological interpretation. Differences in pathways indicated consistency problems which could be quantified by counting the differentially expressed genes between platforms that moved in different directions.

Along these lines, [25] integrated three independent microarray gene expression data sets for breast cancer and identified a structured prognostic signature consisting of 112 genes organized into 80 pair-wise expression comparisons. The method used for integration of data sets was based on the ranks of the expression values within each sample first introduced in [5]. Since the features were rank-based, data normalization was not necessary before data integration.

A cross-study normalization method called *XPN* was suggested by [14], which based on identifying homogeneous groups of genes and samples in the combined data. Specifically, they employed k-means clustering independently to genes and samples of the combined data to identify blocks (or clusters) in the data. Then, each gene expression value was a scaled and shifted block mean plus noise. Their model assumed that the samples of each available study fall roughly into one of the statistically homogenous sample groups identified, and that each group was defined by an associated gene profile that was constant within each of the estimated gene groups. They examined three existing breast cancer data sets and reported that XPN successfully preserved biological information according to ER prediction error rates while removing systematic differences between platforms.

The reliability of gene expression across three previously published breast cancer studies was evaluated by [4]. They compared the strength of evidence of gene to phenotype associations across studies and combined effects across studies. Their methods are implemented by [2] on an R package (www.r-project.org) library called MergeMaid (http://www.bioconductor.org/packages/2.2/bioc/html/MergeMaid.html). They defined a reliability score and set a threshold via permutations to distinguish which were the "reliable" genes in two study experiments, i.e. the genes consistently measured in all studies. For multi-study experiments they considered an alternative interclass correlation coefficient per gene. Finally, they used a between studies combined effect based on the first eigenvector of a *principal component analysis* (PCA) of each study, to determine the genes that are associated with the phenotype.

In order to account for inter-study variation, [3] suggested an "effect size" model for multiple microarray studies. They defined effect sizes as standardized indexes measuring the magnitude of a treatment or covariate effect. They suggested the use of a fixed-effects model (FEM) or a random-effects model (REM) (or alternatively a hierarchical Bayesian model) depending on the homogeneity of study effects. Finally, they measured the statistical significance of their combined results by permutation tests and FDR calculations. Many of their methods are implemented in GeneMeta R package library (http://www.bioconductor.org/packages/2.2/bioc/html/GeneMeta.html).

Finally, an interesting approach is that by [15] who applied a two-stage Bayesian mixture modeling strategy to analyze four independent breast cancer microarray studies derived from different microarray platforms (spotted cDNAs, Affymetrix GeneChip, and inkjet oligonucleotides). They derived an inter-study validated 90-gene "meta-signature" predictive of breast cancer recurrence. Their analysis was based on the signed conditional probabilities of differential expression as introduced in [12]. Particularly, [12] proposed a Bayesian mixture model transformation of DNA microarray data with potential features applicable to meta-analysis of microarray studies, although they employed them in the context of molecular classification. The basic idea was to estimate the platform independent probability of over-expression, under-expression or baseline expression for gene sample combinations given the observed expression measurements. Along these lines, [15] reported that the use of the specific probability measures increased the power of statistical analysis by increasing the sample size.

There is a great challenge to compare and integrate results across independent microarray studies. Meta-analysis studies sometimes produce comparable results even under different logics. Although all approaches, normalization or combination of secondary results, have their merits, here we proceed with studying the effects of scaling existing measurements from various platforms as that was suggested by [12]. An important selection criterion for data integration is the measurement correlations between platforms [18]. Nonetheless, a large number of genes might be lost when looking at the correlation due to different levels of noise between platforms. We find that rescaling of measurements should be able to prevent that.

2 Integrate findings

Here we suggest some characteristics of the data that need to be accounted for when assimilating results from different studies, and evaluate them in independent data sets. Particularly, we consider the "translation" procedure of values as that was first suggested by [12] and it was employed by [15] on the same content. They estimated probabilities of over-expression, under-expression and baseline expression, and translated the intensity measurements into a probability of differential expression. The new probability scale can make comparisons between platforms on a unified scale rather than using gene-specific summaries. For an analytic description of the method see [15].

We use the four data sets also considered by [15], namely the [20], [21], the [23], [7] data sets. The first two studies are cDNA microarray studies, the third is an injekt oligonucleotide array study and the fourth an Affymetrix GeneChip study. The data sets consist of 305 breast cancer samples in total and $2,555$ common genes. The study-specific breast cancer prognosis signatures have been previously reported to have a small overlap. [15] suggested that combination of the four in a probability scale derives a 90 gene meta-signature which is strongly associated with survival in breast cancer patients. We study their approach in terms of the sample's ER status categorization. Furthermore, we suggest a few modifications which seem to strengthen our results in an independent data set produced with homemade two-colour spotted arrays from Qiagen V3 human library. All results presented here are with respect to that independent data set which consists of $34,772$ 70mer probes and 29 samples (18 ER+ and 11 ER− samples). We refer to that data set as $Data_1$ from here onwards.

Measurements for all four data sets [20] [21] [23] [7] considered here are on the so called "poe" scale [15] and vary in the interval $[0,1]$. Our scope is to measure the accuracy of sample classification with respect to their ER status by using simple statistical measures. For that reason, we only consider t-test calculations and Ward's hierarchical clustering with euclidean distance. We avoid comparing our results with those derived when studies are considered individually, since those finding are based on a more advanced statistical methodology. Thus, our scope is to compare the ER classification outcome in $Data_1$ samples when it is assisted by external data and under the same statistical methodology.

2.1 ER signatures when combining data sets

If we consider all 304 samples (one sample from [23] data set had an unknown ER status and was excluded from further analysis), we find a set of 272 genes adequate to distinguish the two classes (ER+, ER−). From those we found 75 common with $Data_1$. There are some common genes with those reported by [23], for example, for ER categorization. Particularly, [23] reported a set of 550 genes, from which 223 are common with $Data_1$. However, only 12 genes are common between the

two list and can be found in $Data_1$. In Figure 1 we can see the two ER signatures. An interesting observation is that both appear to have two mis-calssification errors. We apply agglomerative hierarchical clustering algorithm to expression ratios using Euclidean distance metric and Ward clustering algorithm [13].

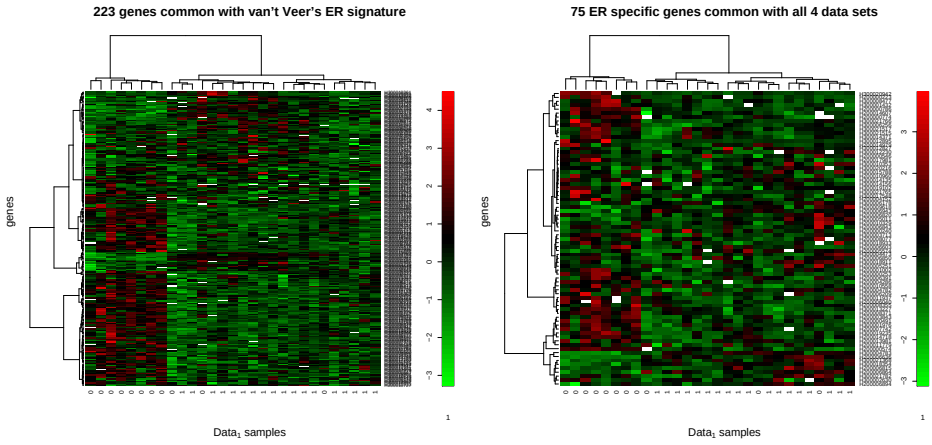


Fig. 1 The ER statistically significant genes reported by van't Veer and those found when we considered the combined four data sets. Results are shown on $Data_1$.

Alternatively, if we consider the whole of $Data_1$ and apply the same methodology as before, we find 279 genes able to statistically distinguish ER. We refer to those results as the *Intrinsic Model* results. However, it would be interesting to consider only the genes of $Data_1$ which are common with the [20] [21] [23] [7] data sets. In this case, 120 statistically significant genes are able to distinguish the two ER classes. Those results are refer to as the *Starting Model* results.

2.2 Intensity and fold-change similarities

Many times the intensity measurements vary between platforms for their common probes. That variability could indicate platform specific effects, or even random noise due to experiment conditions. In this subsection, we study how that variability can affect an ER derived signature which is based on many platforms. For that reason, we consider only probes that appear to have "similar" values across the four data sets in terms of magnitude on the "poe" scale. Particularly, since we are interested on ER classification, we search for genes with similar intensity behaviour in separately ER+ and ER− samples.

We employ Kruskal-Wallis rank sum tests [6, p.115] per gene, to test the null hypothesis that the location parameters of the distribution of ER+ and ER− samples

are the same in each of the four data sets. The alternative is that they differ in at least one. We consider only genes with high p-values for both ER+ and ER− samples, which based on the test give evidence for accepting the null. The left hand-side plot of Figure 2 shows the 44 genes that appear to have the same location distribution parameters for both ER+ and ER− samples across the four data sets. For the right hand-side plot we consider 100 permutations per gene and finally report only 65 genes with significant permutation based empirical p-values with respect to ER status. We can observe that the mis-classification errors are three in both cases, however, permutation procedure is inferior in terms of the number of genes included.

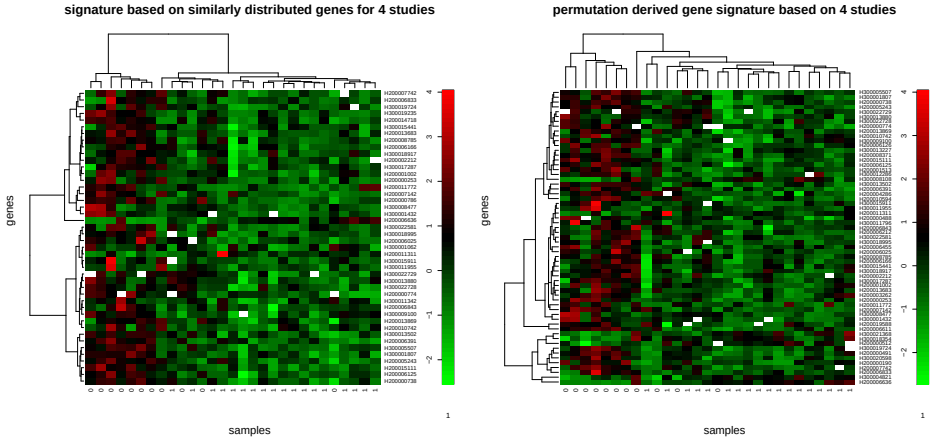


Fig. 2 The ER statistically significant genes among those with same location distribution parameters for the four data sets, as reported by Kruskal-Wallis sum rank test without or with data permutation techniques. Results are shown on $Data_1$.

Another characteristic of the data is the fold-change behaviour between the ER+ and ER− samples. When we consider genes with the same amount of fold-change variability across the four data sets, we find that 24 genes, common for the four data sets and $Data_1$, could distinguish the two ER classes. The genes were selected to have the same fold-change levels for the four platform measurements examined here. In Figure 3 we can see that the two ER classes can be well distinguished and in this case.

2.3 Results

In order to evaluate the approaches suggested before and account for statistical sampling error, we employ *multiclass bootstrap resampling* techniques and estimate via probabilistic measures whether clusters of the original data found by hierarchical

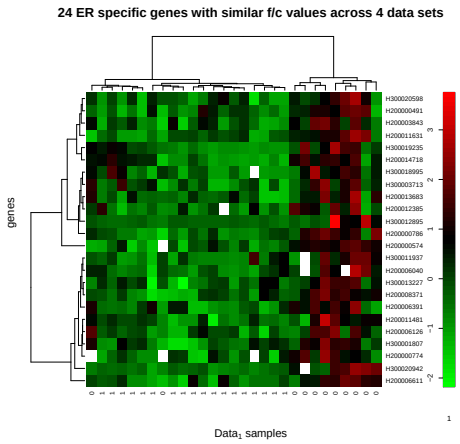


Fig. 3 The ER statistically significant genes with similar fold-change levels among the four data sets examined. Results are shown on $Data_1$.

clustering are strongly supported by the data. For that reason we calculate two types of p-values as they are defined in [17]; the *approximately unbiased* (AU) p-value and the *bootstrap probability* (BP) value. AU p-value is computed by multiscale bootstrap resampling and is thought to be a better approximation to unbiased p-value than BP value which is computed by normal bootstrap resampling. However, the AU p-values themselves include sampling error, since they are also computed by a limited number of bootstrap samples. The null hypothesis in this case is that the clusters of the data are observed by chance. Clusters with AU p-values higher than 95% are strongly supported by data, i.e. those clusters do not seem to be caused by sampling error, but may stably be observed if we increase the number of observations.

[17] suggested that 10 sample sizes for each data set should be examined. Along these lines, we consider sample sizes equal to the $r' = \{0.49, 0.6, 0.69, 0.8, 0.89, 1.0, 1.09, 1.2, 1.29, 1.4\}$ percentages of the original sample size. For each sample size we generate $10,000$ bootstrap samples. For each bootstrap sample, we apply hierarchical clustering to obtain the sets of bootstrap replications of dendrograms and compute the BP for observing each cluster. Finally, we estimate AU p-values by fitting a regression model to the BP values calculated for each cluster and each sample. For an analytic description of the method see [17].

In Table 1 we report the AU and BP values for the approaches already mentioned for the two major clusters of the data C_0 and C_1, where C_0 mostly contains ER$-$ samples and C_1 mostly contains ER$+$ samples. We also report the frequency of misclassified samples in C_0 and C_1, and the number of statistically significant genes with respect to ER status. Note the decrease in the number of significant genes because of mapping when information from combined data is used. The results in the first

row of the table (Intrinsic Model) correspond to clustering results after t-test calculations are directly employed to $Data_1$, whereas, results in the second raw (Starting Model) refer to $Data_1$ but only to its common genes with the four data sets. We consider those values as a baseline for comparison with other approaches suggested here. The results in the third row (Simple Model) correspond to clustering results derived from the four merged data sets. Particularly, we found the significant genes, with respect to the ER status, when the four data sets were considered together and after *Benjamini-Hochberg* correction was applied, and applied our finding to $Data_1$. The Fold-change variability, Kruskal-Wallis (K-W), K-W with Permutations results correspond to methods presented in section 2.2.

Table 1 We report the AU and BP values from bootstraping, the frequency of mis-classified samples and the number of statistically significant genes with respect to ER status. For each variable the two values corresponds to clusters C_0 and C_1, respectively. K-W corresponds to Kruskal-Wallis method. Results are reported for $Data_1$.

Approach	AU (%)	BP (%)	Mis-classifications Freq.	Num. Genes
Intrinsic Model	88 - 83	69 - 52	0 - 0.182	279
Starting Model	89 - 88	38 - 34	0 - 0.182	120
Simple Model	75 - 79	10 - 8	0.111 - 0.15	75
Fold-change Variability	93 - 95	34 - 34	0 - 0.25	24
K-W	76 - 79	27 - 10	0 - 0.182	44
K-W with Permutations	86 - 78	12 - 7	0 - 0.182	65

We can observe that the K-W and Simple Model results have similar AU p-values, although the number of significant genes is higher in the second case. However, they both have smaller AU p-values compared to the Starting Model. Better results in terms of AU p-values and number of genes, can be observed in the case of permutation sampling with Kruskal-Wallis tests. The number of genes increases from 44 to 65 and the AU p-values are elevated supporting the alternative hypothesis that C_0 and C_1 clusters are not observed by chance. However, Fold-change Variability results exhibit the highest AU p-values, although the number of significant genes is small compared to that of the other approaches. The mis-classification frequency is relatively small in all cases, whereas the BP values are variable compared to the AU.

To prove the power of a high number of independent data sets used, in Table 2 we focus on the fold-change variability results but for only three data sets ([23] [20] [21]) and two data sets ([23] [21]) chosen at random from the four. We can observe that our results benefit in terms of AU p-values when information from more data sets is used.

Table 2 We report the AU and BP values from bootstraping, the frequency of mis-classified samples and the number of statistically significant genes with respect to ER status. For each variable the two values corresponds to clusters C_0 and C_1, respectively. K-W corresponds to Kruskal-Wallis method. Results are reported for $Data_1$.

Approach	AU (%)	BP (%)	Mis-classifications Freq.	Num. Genes
Fold-change Variability	93 - 95	34 - 34	0 - 0.25	24
F-c V 3 data sets	70 - 75	25 - 20	0 - 0.182	29
F-c V 2 data sets	63 - 69	20 - 19	0 - 0.143	31

3 Conclusions

We considered how information from studies using various platforms can facilitate the search for significant genes with respect to the categorization of ER samples. Our analysis focused on ER status classification although other parameters, binary or continuous such as breast cancer prognosis, could be studied. An obvious limitation of such approaches is the restriction of the study to only annotated common probes.

We studied the effect of rescaling measurements from four platforms to a common scale and use the information obtained by that data. We employed resampling techniques to minimize sampling error and variability introduced by the different platforms. Our results were compared to those obtained from direct analysis of data, and were thought to be able to describe properties of independent data sets. Particularly, we found that an important property in such kind of analyses is the fold-change variability of common probes across various studies. The performance of K-W analysis was also comparable to that of direct analysis, when data was enhanced with permutations. In all cases, gain in terms of AU p-values resulted in loss of some genes. Overall, we showed that knowledge from numerous data sets produced under the same biological question, can greatly assist the statistical analysis of independent data sets.

References

1. Cope, L., Garrett-Mayer, E.S., Gabrielson, E., Parmigiani, G.: The Integrative Correlation Coefficient: a Measure of Cross-study Reproducibility for Gene Expression Array Data. Working paper, Johns Hopkins University, Dept. of Biostatistics (2007)
2. Cope, L., Zhong, X., Garrett-Mayer, E.S., Parmigiani, G.: MergeMaid: R Tools for Merging and Cross-Study Validation of Gene Expression Data. Working paper, Johns Hopkins University, Dept. of Biostatistics (2004)
3. Choi, J.K., Yu, U., Kim, S., Yoo, O.J.: Combining multiple microarray studies and modeling interstudy variation. Bioinformatics **19**, i84–i90 (2003)
4. Garrett-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., Gabrielson, E.: Cross-study validation and combined analysis of gene expression microarray data. Biostatistics **9(2)**, 333–354 (2008)

5. Geman, D., d'Avignon, C., Naiman, D.Q., Winslow, R.L.: Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. Statistical Applications in Genetics and Molecular Biology **3**, 19 (2004)
6. Hollander, M., Wolfe, D.A.: Nonparametric Statistical Methods. New York: John Wiley & Sons (1973)
7. Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.H., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., Chen, C.M., West, M., Nevins, J.R., Huang, A.T.: Gene expression predictors of breast cancer outcomes. Lancet **361**, 1590-1596 (2003)
8. Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S.C., Hoffman, E., Jedlicka, A.E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S.Q., Yu, W.: Multiple-laboratory comparison of microarray platforms. Nature Methods **2(5)**, 329–330 (2005)
9. Patterson, T.A., Lobenhofer, E.K., Fulmer-Smentek, S.B., Collins, P.J., Chu, T.M., Bao, W., Fang, H., Kawasaki, E.S., Hager, J., Tikhonova, I.R., Walker, S.J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J.C., Tong, W., Shi, L., Wolfinger, R.D.: Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nature Biotechnology **24(9)**, 1140–1150 (2006)
10. Miron, M., Woody, O.Z., Marcil, A., Murie, C., Sladek, R., Nadon, R.: A methodology for global validation of microarray experiments. BMC Bioinformatics **7**, 333–352 (2006)
11. Members of Toxicogenomics Research Consortium: Standardizing global gene expression analysis between laboratories and across platforms. Nature Methods **2(5)**, (2005)
12. Parmigiani, G., Garrett, E.S., Anbazhagan, R., Gabrielson, E.: A statistical framework for expression-based molecular classification in cancer. J. R. Statist. Soc. B **64(4)**, 717–736 (2002)
13. Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., Chinnaiyan, A.M.: Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer. Cancer Research **62**, 4427–4433 (2002)
14. Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M., Nobel, A.B.: Merging two gene-expression studies via cross-platform normalization. Bioinformatics **24(9)**, 1154–1160 (2008)
15. Shen, R., Ghosh, D., Chinnaiyan, A.M.: Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics **5**, 94 (2004)
16. Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S. et al.: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology **24**, 1151–1161 (2006)
17. Shimodaira, H.: Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. The Annals of Statistics **32 (6)**, 2616–2641 (2004)
18. Shippy, R., Sendera, T.J., Lockner, R., Palaniappan, C., Kaysser-Kranich, T., Watts, G., Alsobrook, J.: Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. BMC Genomics **5**, 61 (2004)
19. Smith, D.D., Sætrom, P., Snøve Jr, O., Lundberg, C., Rivas, G.E., Glackin, C., Larson, G.P.: Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. BMC Bioinformatics **9**, 63 (2008)
20. Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Eystein Lonning, P., Borresen-Dale, A.L.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Pnas **98 (19)**, 10869–10874 (2001)
21. Sotiriou, C., Neo, S.Y., McShane, L.M., Korn, E.L., Long, P.M., Jazaeri, A., Martiat, P., Fox, S.B., Harris, A.L., Liu, E.T.: Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Pnas **100 (18)**, 10393–10398 (2003)
22. Stafford, P., Brun, M.: Three methods for optimization of cross-laboratory and cross-platform microarray expression data. Nucleic Acids Research **35**, 10 (2007)

23. van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Letters to Nature **415**, 530–536 (2002)
24. Wang, Y., Klijn, J.G.M., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M.J.J., Atkins,D., Foekens, J.A.: Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet **365**, 671–679 (2005)
25. Xu, L., Tan, A.C., Winslow, R.L., Geman, D.: Merging microarray data from separate breast cancer studies provides a robust prognostic test. BMC Bioinformatics **9**, 125 (2008)

Method for relating inter-patient gene copy numbers variations with gene expression via gene influence networks

Sylvain Blachon, Gautier Stoll, Carito Guziolowski, Andrei Zinovyev, Emmanuel Barillot, Anne Siegel and Ovidiu Radulescu

Abstract During tumorigenesis, genetic aberrations arise and may deeply affect the tumoral cell physiology. It has been partially demonstrated that an increase of genes copy numbers induces higher expression; but this effect is less clear for small genetic modifications. To study it, we propose a systems biology approach that enables the integration of CGH and expression data together with an influence graph derived from biological knowledge. This work is based on 3 key ideas. 1) Inter-individual variations in gene copy number and in expression allow to attack tumor varability and ultimately adresses the problem of individual-centered therapeutics. 2) Confronting post-genomic data to known regulations is a good way to check the soundness and limits of current knowledge. 3) The abstraction level of qualitative modeling allows integration of heterogeneous data sources. We tested this approach on Ewing tumor data. It allowed the definition of new biological hypotheses that were assessed by random permutation of the initial data sets.

Sylvain Blachon
IRISA, UMR 6074 CNRS/INRIA/Universite Rennes 1, campus de Beaulieu, F, 35042 Rennes Cedex, France e-mail: sylvain.blachon@irisa.fr

Gautier Stoll,
UMR 900 INSERM/Mines ParisTech/Institut Curie, 26 rue d'Ulm 75248 PARIS cedex 05, France

Carito Guziolowski
IRISA

Andrei Zinovyev
UMR 900 INSERM/Mines ParisTech/Institut Curie

Emmanuel Barillot
UMR 900 INSERM/Mines ParisTech/Institut Curie

Anne Siegel
IRISA

Ovidiu Radulescu
IRMAR , UMR 6625 CNRS/Universite Rennes 1, Campus de Beaulieu, 35042 Rennes cedex, France

Abreviations : *GCNv* = Gene Copy Number Variation ; *ELv* = Expression Level Variation; ES = Ewing Sarcoma

1 Introduction

Relating genomic instabilities to gene expression is a difficult challenge which is not yet completely resolved. The biological hypothesis is that a gene amplified genomically (in tumor cells for example) induces a higher expression.

Relating gene expression profiles to gene copy numbers was mostly performed using correlation analyzes, in order to find candidate genes serving as markers or as potential targets for therapy [12].

However, these correlation analyzes cannot explain all gene behaviours : in the best case, 50% of them can be explained [7, 10]. This proportion is much weaker for tumors that have less instabilities (like Ewing sarcoma) than more common tumors, like breast cancers. Hence, on those tumors, it appears difficult to extract relevant global properties to relate CGH data to tumor outcomes [8, 16, 11, 1] or to gene expression [3, 15].

We proposed new method for the study of genomic instabilities in tumors, based on the systems biology approach. In this approach, we include the biological processes that regulate transcription through the dynamics of one or several networks of interacting molecules. In such a model genes, transcripts and proteins are network components. The simple process one-gene-one-transcript-one-protein is replaced by a more global point of view involving all the connections among the network components.

In order to deal with small genetic modification, we adopt a more mechanistic approach to genetic variability via a network model. Genetic variability, having nowadays interesting perspectives in personalized medicine, has been addressed by various biologists since Darwin. The idea that interaction between genes can modulate the effects of this variability can be traced back to Conrad Waddington, whose chreods can be interpreted as representations of the "elastic" response of gene networks. Here, gene-gene interactions can stabilize the effect of genetic variability. However, this can be done only up to a certain extent, as some variability is necessarily persistent. The persistent variation is not entirely random, it bares information on the network.

Based on these ideas, a framework was conceived to address the following questions :

- how are gene copy number (*GCN*) and expression level (*EL*) variations related?
- is a (theoretical) gene regulation model consistent with (real life) observed variations in patient pairs? If so, what is the *GCN* contribution to consistency?

This framework is based on qualitative reasoning which formalizes biological interactions [13] and is efficient [1] on large scale regulatory networks [5, 17].

[1] see also the web interface : http://www.irisa.fr/symbiose/bioquali

We applied our methods to Ewing sarcoma: it is a pediatric bone tumors that originate from a translocation t(11;22)(q24;q12), producing a chimeric gene : *EWS-FLI1* [2]. This chimeric gene is thought to act as an aberant transcription factor. A set of target genes that can be either activated or repressed has been already discovered [14].

2 Available data and model for Ewing sarcoma

2.1 Data description and preprocessing

A home made Comparative Genomic Hybridization (CGH) array was built in-house at Institut Curie (3920 probes 60 bp long covering all the chromosomes). CGH data were produced on a set of 47 tumors, including 7 cell lines.

Among the 39 remaining tumors [2], 12 were diagnosed as metastatic tumors before analysis and therapy. After analysis and therapy, on the 27 remaining tumors, 10 evolved in metastasis while 17 remained localized tumors.

An Affymetrix U133A chip was used to measure expression levels on the biopsies of patient tumors. Microarray data were normalized by the GC-RMA technique.

Breakpoint detection on CGH data was performed using GLAD algorithm[6]. GLAD allows CGH level smoothing in a given genomic region flanked by two breakpoints.

2.2 Model description

A gene regulation model involving 130 genes, including $EWS-FLI1$, was designed within SITCON project [4]. The genes/pathways included in this network model were indentified by analysis of transcriptome time series on Ewing cell lines. The logical connections between genes are based on 1) scientific literature and 2) manually curation of TRANSPATH [9] database. Main tumor phetnypes are included in this network: cell cycle regulation, apoptosis and cell migration.

3 Systems biology method for analyzing genetic and expression variations

Our methodology aims at confronting pairwise variations with the ($EWS-FLI1$) gene network model described above.

[2] One was too noisy and was discarded from the analysis

In order to cope with genetic variability in gene networks, we represent differences between individuals as perturbations of the network. Biologically, the hypothesis is that data obtained on a cell population coming from a tumor biopsie reflect a molecular steady state. For a patient pair, the whole set of observed variations describes the qualitative differences between the two steady states. This can be coherently done in a framework that was first introduced in [13], based on interaction graphs and qualitative equations.

3.1 Qualitative equations

Consider a network of n interacting components. The interaction model is the digraph $G = (V,E)$, $V = \{1,\ldots,n\}$. There is an edge $j \to i \in E$ if j influences the production of i. Edges are labelled by a sign $\{+,-\}$ which indicates whether j activates or represses the production of i. Let us denote by $sign(\delta X_i) \in \{\mathbf{+},\mathbf{-},\mathbf{?}\}$ the sign of the variation of i between two conditions, and by $sign(j \to i) \in \{\mathbf{+},\mathbf{-}\}$ the sign of the edge $j \to i$ in the interaction graph.

For every predecessor j of i, $sign(j \to i) * sign(X_j)$ provides the sign of the *influence variation* of j on the species i. Notice that this can be either positive (increased activation or decreased repression) or negative (decreased activation or increased repression). Then, the constraints that the network imposes on the variations can be expressed as qualitative equations:

$$sign(\delta X_i) \approx \sum_{j \to i} sign(j \to i) sign(\delta X_j). \tag{1}$$

The sign algebra is summarized in the following table.

$$\begin{array}{l|l|l|l} \mathbf{+}+\mathbf{-}=\mathbf{?} & \mathbf{+}+\mathbf{+}=\mathbf{+} & \mathbf{+}\times\mathbf{-}=\mathbf{-} & \mathbf{+}\times\mathbf{+}=\mathbf{+} \\ \mathbf{-}+\mathbf{-}=\mathbf{-} & \mathbf{-}\times\mathbf{-}=\mathbf{+} & \mathbf{?}+\mathbf{?}=\mathbf{?} & \mathbf{?}\times\mathbf{?}=\mathbf{?} \\ \mathbf{?}+\mathbf{-}=\mathbf{?} & \mathbf{?}+\mathbf{+}=\mathbf{?} & \mathbf{?}\times\mathbf{-}=\mathbf{?} & \mathbf{?}\times\mathbf{+}=\mathbf{?} \end{array}$$

$$\mathbf{+} \not\approx \mathbf{-} \mid \mathbf{?} \approx \mathbf{+} \mid \mathbf{?} \approx \mathbf{-}$$

3.2 Taking into account genetic variations

In order to take into account the genetic variability of the patients we introduced new qualitative variables representing, for a given pair of patients, the GCN variations. The corresponding nodes in the interaction digraph will be called "gene nodes". There is one gene node for each gene considered and in our analysis we kept a set of 126 genes. The remaining nodes are either mRNA or protein nodes occurring in the $(EWS-FLI1)$ network.

The **central hypothesis** here is that gene nodes act directly and positively on the mRNA nodes in the network.

To summarize, the interaction model contains:

1. gene nodes : the sign stems from *GCN* variation between two patients,
2. mRNA nodes : the sign stems from *EL* variation between two patients,
3. proteins : the sign stems from protein activity variation between two patients.

GCN variations and *EL* variations come from CGH and microarray data. The protein activity variations remain unknown but can be predicted thanks to our formalism.

3.3 Encoding variations

For each gene k, we define $GCNv^k_{i,j} = CGH(i,k) - CGH(j,k)$, where $CGH(i,k)$ is the CGH level of the gene k in the patient i smoothed by GLAD algorithm. When $\mid GCNv^k_{i,j} \mid > 0,2$ the variation is considered as significant [6].

Similarly, for gene expression variation $ELv^k_{i,j} = EL(i,k) - EL(j,k)$, where $EL(i,k)$ is the mean expression level measured by Affymetrix probes corresponding to the gene k in the patient i. To evaluate the significance of the variation, a Student test was used on the set of probesets measuring $EL(i,k)$ with an alpha risk of 5%.

Both for gene and mRNA nodes, significant variations are encoded **+** or **–**. The **?** sign is used for nodes that are undetermined at various steps of our calculations.

3.4 Consistency analysis

For each pair of patients, we solve the system of qualitative equations (1), augmented by the information on signs coming from data. If there are solutions, the system is declared compatible. In case of compatibility some nodes have the same unique sign in each one of the many possible solutions. The unique signs of these nodes (called hard components) are predictions of the model. By this, the signs on protein nodes are predicted.

If no solution can be found, a localization of the source of conflict is attempted by subsystem analysis. First, all local violations (meaning that at least one equation (1) is violated by data information) are declared “local inconsistencies”. All locally inconsistent patterns have the same structure : one node together with its predecessors. All the other situations are declared “global inconsistencies”. Globally inconsistent patterns are more complex (they contain at least two nodes with their respective predecessors).

Notice that testing the consistency and looking for sources of conflicts is actually a NP-hard question. It appears that the topology of the network allows to

handle these questions. We used decision diagrams, a data structure meant to represent functions on finite domains; it is widely used for the verification of circuits or network protocols. Using such a compact representation of the set of solutions, we proposed efficient algorithms for computing solutions of the systems, predictions, and other properties of a qualitative system [17].

3.5 Monte Carlo estimates for statistical significance of consistency

Consistency could occur also by chance. In order to estimate the significance of consistency results, we used random perturbations and Monte Carlo estimates of the mean numbers of pairs of patients for which random data is consistent with the network.

For a pair of patients (i,j), let us note:

1. $C^+_{i,j}$ and $C^-_{i,j}$ the set of genes for which the gene copy numbers vary positively, resp. negatively, between the patients i and j.
2. $E^+_{i,j}$ and $E^-_{i,j}$ the set of genes for which the gene expressions vary positively, resp. negatively, between the patients i and j.

Straightforwardly, $C^+_{i,j} \cap C^-_{i,j} = \oslash$ and $E^+_{i,j} \cap E^-_{i,j} = \oslash$

The qualitative equations (1) were solved with $N = 1000$ data sets (each data set contains $P(P-1)/2$ patient comparisons, where P is the number of patients) produced by randomly permuting the elements contained in $C^+_{i,j}$, $C^-_{i,j}$, $E^+_{i,j}$ and $E^-_{i,j}$.

For each random dataset, consistency was tested. In case of consistency, predictions on network nodes were computed. Each random dataset r is consistent N^C_r times, locally inconsistent N^{LI}_r times and globally inconsistent N^{GI}_r times. Note that $N^C_r + N^{LI}_r + N^{GI}_r = P(P-1)/2$.

The distributions of N^C, N^{LI} and N^{GI} provide the estimates for the number of consistent and inconsistent pairs with random data we are looking for.

4 Results

In this section, we apply our method to Ewing sarcoma data. We show results for a couple of questions - the first concerning the relation between *GCNv* and *ELv* ; the second concerning the model consistency tests and the impact of *GCNv* on them.

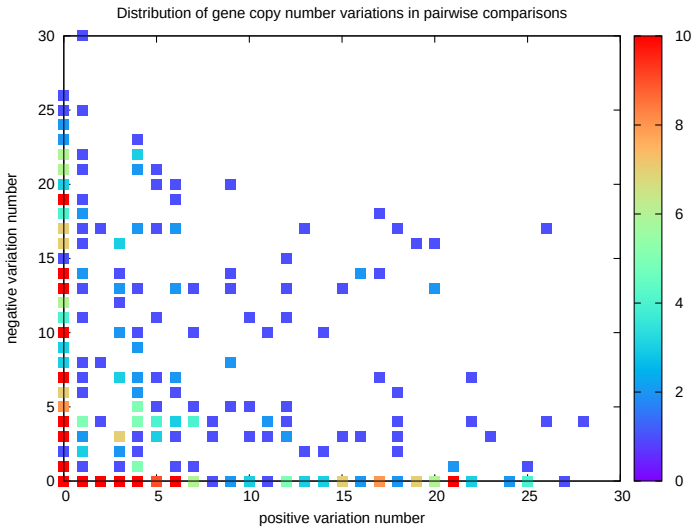


Fig. 1 Repartition of the $CGHv^+$ and $CGHv^-$ cardinals in the 741 patient pairs. Each point represents at least one patient pair - the number of patient pair is color-coded according to the palette on the right. Observe that a lot of patient pairs have few $CGHv$. The higher peak is the point (0,0) on which 164 patient pairs aggregate.

4.1 Discovering links between gene copy number variations and expression level variations

How are Gene Copy Number variations (*GCNv*) and Expression Level variations (*ELv*) related ?

GCNv and *ELv* were evaluated for each gene and each patient pair. There are 39 patients, thus 741 patient pairs.

First, the *Figure 1* and *Figure 2* show the repartition of the variation numbers in the patient pairs.

It is striking to see the difference in variation repartitions. *GCNv* are less frequent but also mainly distributed along the *x* and *y* axis. This is coherent with the relative genome stability of ES.

In spite of this general trend, some ES can exhibit a high number of *GCNA*s. When these unstable tumors are compared to rather stable tumors, imbalances are favored in one sense rather than the other, giving this picture with most of the pairs around the 0,0 point and distributed along the *x* and *y* axis.

A different picture can be observed with expression data. Variations are more frequent and distributed in a larger area, showing a rather homogeneous variability of *ELv* among the patient pairs.

From these figures, one can imagine that *GCNv* and *ELv* are independent variables.

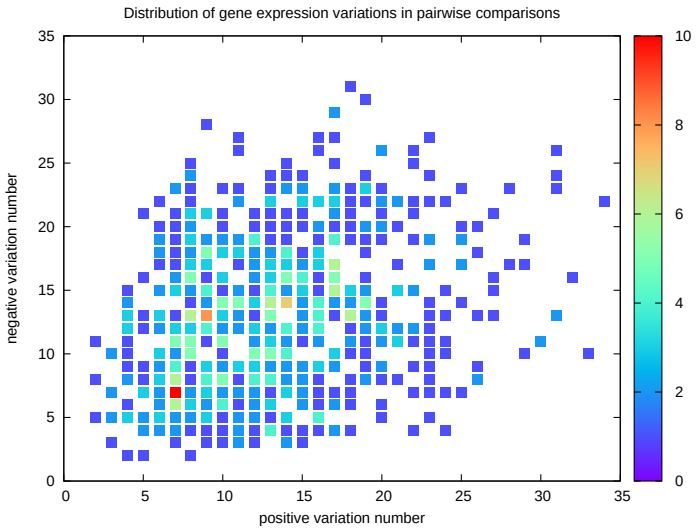


Fig. 2 Repartition of the ELv^+ and ELv^- cardinals in the 741 patient pairs. Each point represents at least one patient pair - the number of patient pair is color-coded according to the palette on the right. The distribution is much different than the *vCGH* one.

To verify this, a Pearson χ^2 independence test was performed under the null hypothesis that *GCNv* and *ELv* are two independent variables. The repartitions and contributions to the χ^2 score are shown in *Table 1*.

χ^2	*GCNv* / *ELv*	0	+	-	?
Observed	0	53117	2197	2598	14447
	+	6969	294	573	1948
	-	6781	550	283	2127
	?	1194	38	32	218
Expected	0	52547	2386	2701	14523
	+	7132	323	365	1964
	-	7101	321	364	1955
	?	1080	49	55	297
Contribution to Chi^2	0	2,59	15,0	3,98	0,40
	+	3,74	2,54	118,09	0,13
	-	14,4	162,91	17,9	15,10
	?	11,96	2,42	9,84	21,23

Table 1 Pearson χ^2 Independence test. $\chi^2 = 402,2 \gg 27,88$, the χ^2 value for 9 freedom degrees with an alpha risk of 0,001. The major contribution is given by situations where a gene changes in an anti-correlated way in *GCN* and in *EL*. The "?" sign corresponds to cases when a probe signal in at least one experiment is too noisy to assess the variation sign.

The χ^2 statistics equals 402, much greater than the value for an alpha risk of 0,1%., with 9 freedom degrees. The null hypothesis can be confidently rejected. This confirms that *GCNA*s can affect transcription in ways that can be investigated by comparing pairs of patients.

Moreover, the major contribution occurs when *GCNv* and *ELv* have opposite signs. This is even more striking on the whole gene set (χ^2 = 88761 ; contribution of *GCNv* and *ELv* having opposite sign = 83,7%). This was clearly unexpected and it is highly counter-intuitive.

We will show that our qualitative reasoning method allows to find explanations to this surprising phenomenon.

4.2 Checking the EWS-FLI1 regulation model

To address this issue, we used a systems biology approach based on the qualitative analysis to confront an interaction model with CGH and expression data.

The consistency analysis raw results are shown on *Figure 3*.

We were concerned with two main questions:

- In which proportion is the EWS-FLI1 network model consistent with real data? In cases of inconsistency, what does this tell about the model?
- Is information contained in CGH data useful to uncover regulations ?

4.2.1 Explaining inconsistencies

On real data including *GCNv* influences, the model is consistent with the data in 317 patient pairs (42,8% of the 741). Additionally, 314 (42,4%) local inconsistencies and 110 (14,8%) global inconsistencies were found.

Understanding the incompatibility sources may help to focus on the model weaknesses. First, it is necessary to analyze the sources of local inconsistencies, the most numerous ones.

All the local inconsistencies have the same origin : a patient pair (i,j) where there is at least one gene k for which : $GCNv(i,j,k) = -ELv(i,j,k)$. We call this an *anticorrelated variation*.

This is not a rare case: from the *Table 1*, there are 1123 cases in the 741 patient pairs. They are spread in 367 patient pairs and involve 67 genes.

Hence, 367 local inconsistencies were expected. This means that 50 pairs that were expected to be locally inconsistent were explained by the model.

More precisely, on the 67 genes that are involved in anticorrelated variations, 23 are never involved in local consistencies This is due to the presence in the network model of at least one transcription regulation on those genes. Those explained locally inconsistent influences appear 414 (36,9%).

In other words, local inconsistencies point to the lack of transcription regulations in the model. Adding them can potentially remove all local inconsistencies.

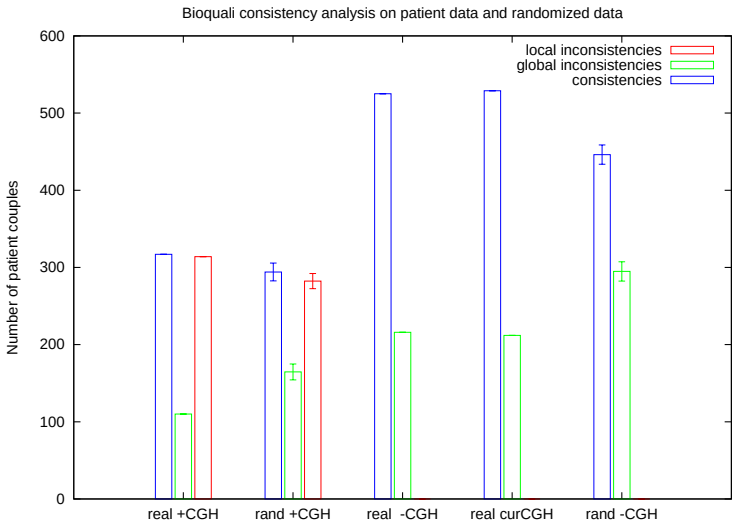


Fig. 3 Consistency global results on patient data and randomized data. $real+CGH$ means that the model was confronted to patient data including all *GCNv* influences. $rand+CGH$ means that the model was confronted to randomized data including all random *GCNv* influences. $real-CGH$ means that the model was confronted to *ELv* observed on patient data without *GCNv* influences. *realcurCGH* means that the model was confronted to patient data without the unexplained anti-correlated *vCGH / ELv*. $rand-CGH$ means that the model was confronted to randomized data without random *GCNv* influences.

The global inconsistencies point to other model weaknesses. Unfortunately, our solver is not able to localize the whole set of inconsistent subgraphs (see section 2)[3].

Only 33 influences are involved in the inconsistent set we obtained. All of themare implied in transcription regulation, including: $TP53$, $E2F1$ and $EWS-FLI1$.Therefore, our method allows to focus on subgraphs of a complex model that need refinement to become consistent with observations.

4.2.2 CGH data increase the consistency between the data and the model

To assess what relevant information is contained in CGH data, it can be useful to :

- reproduce the consistency analysis without taking into account the anticorrelated variations that remain unexplained by the model ;
- compare the result to the consistency analysis when *GCN* influences are removed (by taking into account only *ELv*, hence discarding CGH information).

This analysis gave the results shown in *Figure 3*. Without CGH information, the model is consistent with *ELv* alone in 525 (70,9%) patient pairs. Using CGH

[3] Notice that a more powerful implementation of constraints solver, with Answer Set Programming, will be soon available to overcome this technical problem.

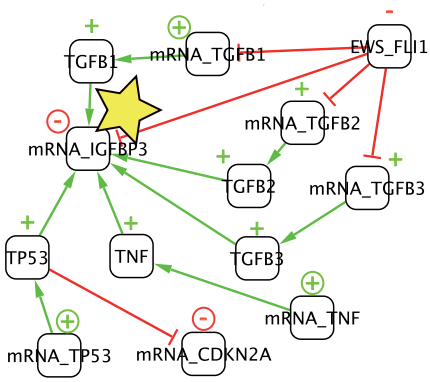


Fig. 4 Inconsistent subgraph without *GCNv* influences on the patient pair (EW57,EW58). Observed signs of gene expression variation are circled. The star points to the inconsistency localization, here the *mRNA_IGFBP*3 node.

information on the genes having transcription regulators, the model is consistent with *ELv* and *GCNv* in 528 (71,3%) patient pairs.

In 3 cases, the model becomes consistent thanks to information from CGH data.

To understand better the impact of *GCNv* influences, the 3 inconsistent subgraphs were represented using Cytoscape software. The *Figures 4* and 5 show an example of this analysis.

On this example, the inconsistency is localized on the *mRNA_IGFBP*3 node. $IGFBP3$ is positively targeted by a set of regulators, except $EWS-FLI1$ that was shown to inhibit $IGFBP3$. However, between the two patients, given the observations, *mRNA_IGFBP*3 should be activated. This is not the case, producing the inconsistency.

The $IGFBP3GCNv$ proposes an explanation to this phenomenon : due to its negative variation between the two patients, the negative $IGFBP3ELv$ can be understood. A biological interpretation for such a pattern can be: despite the positive signals arising from various regulators, the difference in gene copy number is sufficient to decrease $IGFBP3EL$.

Obviously, this hypothesis must be confirmed by experimental validation.

To conclude, CGH data bring information on local variations that have an influence on the $EWS-FLI1$ network model.

4.2.3 Assessing the statistical quality of consistency frequencies

In order to assess the quality of consistency tests, a randomization of input data was performed using 1000 random permutation on *GCNv* and *ELv* for each patient pair. The 741000 data sets with *GCNv* were confronted to the EWS-FLI1 model; the same analysis was repeated on the same datasets without considering *GCN* influences.

Consistency analyzes with and without *GCN* influences were performed to be compared to results on real datasets. Results are exposed in *Figure 3*. This shows that there are less consistencies and proportionally more inconsistencies on random data than on real data.

The distribution of consistency frequency distribution obtained for the 1000 datasets including *GCN* influences follows a normal distribution ($\mu = 279$, $\sigma = 10,1$ - Kolmogorov-Smirnov normality test value = 0,0289 $<$ 0,0386 , the bilateral value for an alpha risk of 5%).

Given such a distribution, the probability to obtain a consistency frequency equal to or greater than 317[4] equals 3,79%.

Similarly, the distribution of the consistency frequency distribution obtained on the 1000 datasets without *GCN* influences (see *Figure 6*). The distribution follows a normal distribution ($\mu = 446$, $\sigma = 12,5$ - Kolmogorov-Smirnov normality test value = 0,0251 $<$ 0,0386 , the bilateral value for an alpha risk of 5%).

Given such a distribution, the probability to obtain a consistency frequency equal to or greater than 525 [5] equals $1,31.10^{-10}$. This probability is even lower for the real data set using *GCN* explained by the model[6].

This proves that one can trust the consistency frequency obtained on real data sets.

However, it is surprising to observe such a high number of consistent cases on randomized data sets. We are currently investigating the reasons. Two hypotheses motivate us :

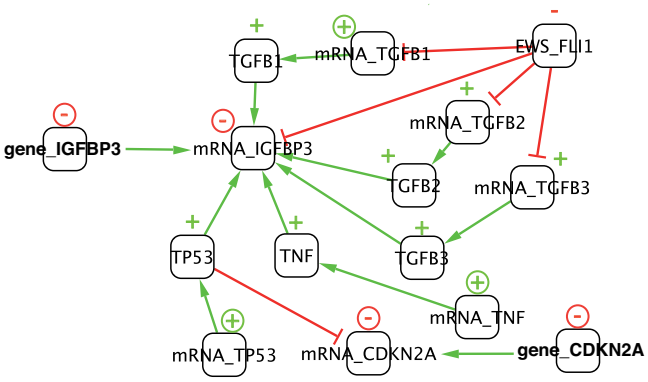


Fig. 5 Subgraph with *GCN* influences on patient pair (EW57,EW58). Observed signs of gene expression variation are circled. The *IGFBP3GCN* varies negatively between the two patients and resolves what was previously an inconsistency between the model and *ELv* alone.

[4] the consistency frequency obtained on real data set without *GCN*.

[5] The consistency frequency obtained on real data set without *GCN*.

[6] The consistency frequency obtained on real data set with the *GCN* influences explained by the model equals 528.

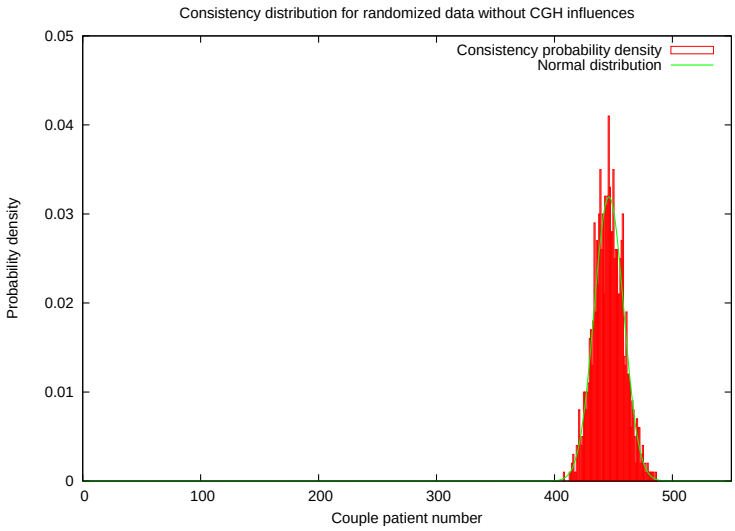


Fig. 6 Consistency frequency distribution for randomized data without *vCGH* influences. The distribution follows a normal distribution ($\mu = 446$, $\sigma = 12,5$). It must be compared to the consistency number obtained on real data without *GCNv* influences (525) and with *GCNv* explained by the model (528).

- the network topology may be robust to random variations;
- there is an effect of the number of constraints imposed by observations - as there is a variability in $\mid GCNv \mid$ and $\mid ELv \mid$ as shown on *Figure 1 and 2*.

A simple way is to test wether a significant correlation exists between $\mid GCNv(i,j) \mid$ and of $\mid GCNv(i,j) \mid$ and the consistency for each patient pair (i,j).

5 Discussion

Given the difficulties to analyze CGH data on ES tumors, we propose to change of paradigm and propose a systems biology approach dedicated to invesitgate the inter-patient variability simultaneously at genomic and transcriptomic levels and their compatibility with a EWS-FLI1 gene regulation network.

This study addresses two main issues: 1) the link between CGH and expression pairwise variations in 39 ES; 2) the consistency between a $EWS-FLI1$ model and these variations.

To handle the first question, interesting representations of patient pairs as functions of *GCNv* and of *ELv* cardinalities were produced. It appears that the patient pair distribution following *GCNv* is highly different from its counterpart following *ELv*. This shows that patient pairwise comparisons exhibit different transcriptomic

and genomic variability patterns. One could be mistaken in interpreting this as the result of an independence between the variables *GCNv* and *ELv*.

The χ^2 independence test states that these two variables are undoubtely related. This agrees with biological intuition: when a gene copy number increases, the gene expression level is expected to increase - and *vice versa*.

However, surprisingly, the major dependency contribution comes from anticorrelated variations. This is true for the whole set of measured genes. This suggests the existence of a feedback regulation of genes present in altered regions that counteracts *GCN* imbalances.

The $EWS-FLI1$ network model is able to deal in part with these anticorrelated variations : no gene having at least one transcriptional regulation appears in local inconsistencies. On the contrary, if a "deficient gene" does not have a transcription regulation, it will be involved in a local inconsistency if its *GCNv* and *ELv* appear anticorrelated. Thus, our method points to the model incompleteness. Adding missing transcription regulations can potentially remove all local inconsistencies.

To answer the second question, the compatibility of the EWS-FLI1 model with the pairwise genomic and transcriptomic variations was verified. It appears that the model is consistent with expression data in more than 70% of the cases after having silenced the "deficient gene" *GCNv* influence.

3 cases that were inconsistent using *ELv* alone become consistent. The analysis of these inconsistent subgraphs shows that some *ELv* that were unexplained by the model could be explained by local *GCN* variations. This suggests that local *GCN* variations carry valuable information that can propagate through the interactions. This result validates the capacity to investigate such local effects of *GCNv* by our approach.

Finally, we compared our results to 1000 randomized datasets. It appears that the consistency frequencies obtained on real datasets cannot be obtained by chance.

Another intriguing phenomenon appeared during this latter analysis: we did not expect so high consistency frequencies on randomized datasets. We are currently studying wether this is related to the genomic and transcriptomic variation number or whether this is a consequence of the intrinsic network robustness.

Biological system robustness may be the key to understand apparent contradictions in experimental data on ES. Let us consider the following paradox: the existence of a general trend that relates genetic instabilities to worst prognosis is opposed to the difficulty of finding repeated and specific genetic disorders linked to tumor outcomes.

If we consider that genetic instability acquisition is a stochastic process, stemming from a disturbed DNA repair machinery, it is likely that the largest part of genetic disorders have no individual effect on the cell physiology. This may result from a negative feedback control.

In the same time, it is also possible that, in exceptional cases, a specific gene disorder manages to overcome feedback and have visible effects on cell physiology. $EWS-FLI1$ itself is an extreme example.

As future work, we intend to use our method to detect these exceptional cases. We already proved that in a very limited number of cases (3 on 216 inconsistencies) the information on *GCNv* carried by specific genes can explain an unusual network behavior.

The novel hypothesis that outcomes from this work is that genetic disorder accumulation can have a global impact by increasing the probability that a specific gene disorder has consequences on a stabilized network. We expect that such events will be found more frequently in metastatic tumors than in non metastatic ones.

6 Acknowledgements

We acknowledge the financial support from ANR BIOSYS-2006 program (SITCON project) and the PIC Bioinformatique et Biostatistiques from Institut Curie. AZ, EB and GS are members of the team "Systems Biology of Cancer", équipe labellisée par la Ligue Nationale Contre le Cancer. We thank Philippe Hupé and Philippe La Rosa for help and disponibility for using Vamp. We also thank Michel Le Borgne and Jacques Nicolas for stimuling discussions and the bioinformatics GenOuest platform for their invaluable help to overcome technical problems and use the Bioquali application efficiently. SB's work was supported by a INRIA-French ministry postdoc grant.

References

1. S Brisset, G Schleiermacher, M Peter, A Mairal, O Oberlin, O Delattre, and A Aurias. Cgh analysis of secondary genetic changes in ewing tumors: correlation with metastatic disease in a series of 43 cases. *Cancer Genet Cytogenet*, 130(1):57–61, Oct 2001.
2. O Delattre, J Zucman, B Plougastel, C Desmaze, T Melot, M Peter, H Kovar, I Joubert, P de Jong, and G Rouleau. Gene fusion with an ets dna-binding domain caused by chromosome translocation in human tumours. *Nature*, 359(6391):162–5, Sep 1992.
3. B I Ferreira, J Alonso, J Carrillo, F Acquadro, C Largo, J Suela, M R Teixeira, N Cerveira, A Molares, G Goméz-López, A Pestaña, A Sastre, P Garcia-Miguel, and J C Cigudosa. Array cgh and gene-expression profiling reveals distinct genomic instability patterns associated with dna repair and cell-cycle checkpoint pathways in ewing's sarcoma. *Oncogene*, 27(14):2084–90, Mar 2008.
4. Stoll G., Zinovyev A., Tirode F., Laud-Duval K., Delattre O., and Barillot E. Model-based approach for analysis of transcriptome perturbation reveals ewing oncogene interaction network. In *International Conference on Systems Biology, poster*, Long Beach, CA, USA, 2007.
5. Carito Guziolowski, P Veber, Michel Le Borgne, Ovidiu Radulescu, and Anne Siegel. Checking consistency between expression data and large scale regulatory networks: a case study. *The Journal of Biological Physics and Chemistry*, 7(2):37–43, 2007.
6. Philippe Hupé, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi, and Emmanuel Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–22, Dec 2004.
7. Elizabeth Hyman, Päivikki Kauraniemi, Sampsa Hautaniemi, Maija Wolf, Spyro Mousses, Ester Rozenblum, Markus Ringnér, Guido Sauter, Outi Monni, Abdel Elkahloun, Olli-P Kallion-

iemi, and Anne Kallioniemi. Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Res*, 62(21):6240–5, Nov 2002.

8. S Knuutila, G Armengol, AM Bjorkqvist, W ElRifai, ML Larramendy, O Monni, and J Szymanska. Comparative genomic hybridization study on pooled dnas from tumors of one clinical-pathological entity. *Cancer Genetics and Cytogenetics*, 100(1):25–30, Jan 1998.
9. Mathias Krull, Susanne Pistor, Nico Voss, Alexander Kel, Ingmar Reuter, Deborah Kronenberg, Holger Michael, Knut Schwarzer, Anatolij Potapov, Claudia Choi, Olga Kel-Margoulis, and Edgar Wingender. Transpath: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Research*, 34(Database issue):D546–51, Jan 2006.
10. Hyunju Lee, Sek Won Kong, and Peter J Park. Integrative analysis reveals the direct and indirect interactions between dna copy number aberrations and gene expression changes. *Bioinformatics*, 24(7):889–96, Apr 2008.
11. T Ozaki, M Paulussen, C Poremba, C Brinkschmidt, J Rerin, S Ahrens, C Hoffmann, A Hillmann, D Wai, K L Schaefer, W Boecker, H Juergens, W Winkelmann, and B Dockhorn-Dworniczak. Genetic imbalances revealed by comparative genomic hybridization in ewing tumors. *Genes Chromosomes Cancer*, 32(2):164–71, Oct 2001.
12. Jonathan R Pollack, Therese Sørlie, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA*, 99(20):12963–8, Oct 2002.
13. O Radulescu, S Lagarrigue, A Siegel, P Veber, and M Le .... Topology and linear response of interaction networks in molecular biology. *Journal of The Royal Society Interface*, Jan 2006.
14. N Riggi and I Stamenkovic. The biology of ewing sarcoma. *Cancer Letters*, Jan 2007.
15. S Savola, F Nardi, K Scotlandi, P Picci, and S Knuutila. Microdeletions in 9p21.3 induce false negative results in cdkn2a fish analysis of ewing sarcoma. *Cytogenet Genome Res*, 119(1-2):21–6, Jan 2007.
16. M Tarkkanen, S Kiuru-Kuhlefelt, C Blomqvist, G Armengol, T Bohling, T Ekfors, M Virolainen, P Lindholm, O Monge, P Picci, S Knuutila, and I Elomaa. Clinical correlations of genetic changes by comparative genomic hybridization in ewing sarcoma and related tumors. *Cancer Genetics and Cytogenetics*, 114(1):35–41, Jan 1999.
17. Philippe Veber, Carito Guziolowski, Michel Le Borgne, Ovidiu Radulescu, and Anne Siegel. Inferring the role of transcription factors in regulatory networks. *BMC Bioinformatics*, 9:228, Jan 2008.

Revealing Disease Mechanisms via Coupling Molecular Pathways Scaffolds and Microarrays: A Study on the Wilm's Tumor Disease

Alexandros Kanterakis[1], Vassilis Moustakis[1,3], Dimitris Kafetzopoulos[2], and George Potamias[1]

[1]Institute of Computer Science, FORTH, Greece
[2]Institute of Molecular Biology and Biotechnology, FORTH, Greece
[3]Department of Production Engineering and Management, TUC, Chania, Greece

Abstract. Moving towards the realization of genomic data in clinical practice, and following an individualized healthcare approach, the function and regulation of genes has to be deciphered and manifested. This is even more possible after the later advances in the area of molecular biology and biotechnology that have brought vast amount of invaluable data to the disposal of researchers. Two of the most significant forms of data come form microarray gene expression sources, and gene interactions sources – as encoded in Gene Regulatory Pathways (GRPs). The usual computational task involving microarray experiments is the gene selection procedure while, GRPs are used mainly for data annotation. In this study we present a novel perception of these resources. Initially we locate all functional paths encoded in GRPs and we try to assess which of them are compatible with the gene-expression values of samples that belong to different clinical categories (diseases and phenotypes). Then we apply usual feature selection techniques to identify the paths that discriminate between the different clinical phenotypes providing a paradigm shift over the usual gene selection approaches. The differential ability of the selected paths is evaluated and their biological relevance is assessed. The whole approach was applied on the Wilm's tumor domain with very good and indicative results.

1. Introduction

The interdisciplinary research field of molecular biology and bioinformatics is continuously enriched by the advances in many areas such as sequence analysis, genome annotation and analysis of gene and protein regulation. These advances have brought to the present the post-genomic era where, as the basic knowledge is tamed, we are mainly seek for methods that integrate various and heterogeneous

types of established biological knowledge. The major question to deal with relates to the regulation of the function of genes, targeting the ways that this function affects the overall phenotype of a living organism. A first step towards this is the combined processing of clinical and genomic data. Clinical data as the explicit expression of the phenotypic features of an organism should be integrated with the genomic data that represent the genotypic signature of the organism. This effort can help researchers to gain insights about the role of gene function in pathology, locate the risks and susceptibilities of each unique person and thus, provide individualized healthcare [1].

From a biology point of view, the goal is to provide a systematic, genome-scale view of genes interactions and functionality [2]. The advantage of this approach is that it can identify emergent properties of the underlying molecular system as a 'whole' – an endeavor of limited success if targeted genes, reactions or even molecular pathways are studied in isolation [3]. Individuals show different phenotypes for the same disease – they respond differently to drugs and sometimes the effects are unpredictable. Many of the genes examined in early clinico-genomic studies were linked to single-gene traits, but further advances engage the elucidation of multi-gene determinants of drug response. Differences in the individuals' background DNA code but mainly, differences in the underlying gene *regulation* mechanisms alter the expression or function of proteins being targeted by drugs, contribute significantly to variation in the responses of individuals. The challenge is to accelerate our understanding of the molecular mechanisms of these variations and to produce targeted individualized therapies.

In this paper we present an integrated methodology that couples and 'amalgamates' knowledge and data from both Gene Regulatory Pathway (GRP) and Microarray (MA) gene-expression sources. The methodology comprises two main parts. In the first part we decompose a number of targeted pathways – pathways involved in particular disease phenotype, into all possible functioning paths (i.e., part of a molecular pathway). Then, by introducing gene expression knowledge from a MA experiment we rank all paths according to the 'compatibility difference' they exhibit among the samples of different clinical phenotypes. In the second part of the methodology we substitute genes with paths and gene expression with compatibility value ranks. At the end we apply feature selection techniques to identify those functional paths that differentiate between the targeted phenotypic classes, and we assess their prediction power (classification) performance. As a proof of concept we apply the technique on a microarray experiment that targets the *Wilms' tumor* (WT; nephroblastoma) disease. We were able to identify significant paths in various molecular pathways that reveal distinct mechanisms between different WT phenotypes. The targeted WT phenotypes concern the tumor grade histological feature. Results are discussed about their biological relevance.

A preliminary implementation of the methodology is made in a system called MinePath. MinePath aims to uncover potential gene-regulatory 'fingerprints' and mechanisms that govern the molecular and regulatory profiles of diseases.

2. MAs and GRNs as sources of biomedical knowledge

2.1. Microarrays

Microarrays [4], [5] are devices able to measure simultaneously the expression of thousands of genes, revolutionizing the areas of molecular diagnostics and prognostics. A number of pioneering studies have been conducted that profile the expression-level of genes for various types of cancers such as leukemia, breast cancer and other tumors [6], [7]. The aim is to add molecular characteristics to the classification of diseases so that diagnostic procedures are enhanced and prognostic predictions are improved. These studies demonstrate the great potential and power of gene-expression profiling in the identification and prediction of various disease phenotypes and prognostic disease factors.

Gene-expression data analysis depends on Gene Expression Data Mining (GEDM) technology, and the involved data analysis is based on two basic approaches: (a) hypothesis testing - to investigate the induction or perturbation of a biological process that leads to predicted results, and (b) knowledge discovery - to detect underlying hidden-regularities in biological data. For the latter, one of the major challenges is gene-selection [8], [9]. Possible prognostic genes for disease outcome, including response to treatment and disease recurrence, are then selected to compose the molecular signature (gene-markers) of the targeted disease.

2.2. Gene Regulatory Pathways

GRPs are network structures that depict the interaction of DNA segments during the transcription of genes into mRNA. The prominent and vital role of GRPs in the study of various biology processes is a major sector in contemporary biology research, where numerous thorough studies have been conducted and reported [10], [11]. From a computational point of view, GRPs can be conceived as analogue of biochemical computers that regulate the level of expression of target genes [12]. Each network has inputs, usually proteins or transcription factors that initiate the network function. The outputs are usually certain proteins (encoded by specific genes). The network by itself acts as a mechanism that determines cellular behavior where the nodes are genes and edges are functions that represent the molecular reactions between the nodes. These functions can be perceived as Boolean functions, where nodes have only two possible states ("on" and "off"), and the whole network being represented as a simple directed graph [13]. The notion of

GRPs is by itself an abstraction of the underlying chemical dynamics of the cell, thus the expectation of high reliability in terms of modeling is limited. It is indicative that most of the relations in known and established GRPs have been derived from laborious and extensive laboratory experiments and careful study of the existing biochemical literature. Thus GRPs are far from being complete, at least with respect to their ability to capture and model all the internal cell dynamics of complex living organisms.

Current efforts focus on the reconstruction of GRPs by exploring gene-expression data. For example in [14] it is reported that network topologies, as extracted from gene co-expression events, could discover motifs and regulatory hubs that can characterize the entire cellular states and guide further pharmaceutical research. Very few methods of gene regulatory inference are considered superior, mainly because of the intrinsically noisy property of the data, 'the curse of dimensionality', and the lack of knowledge about the 'true' underlying structure of the networks.

The study of the function, structure and evolution of GRPs in combination with microarray gene-expression profiles and data is essential for contemporary biology research. First of all, researchers have uncovered a multitude of biological facts, such as protein properties and genome sequences. But this alone is not sufficient to interpret biological systems and understand their robustness, which is one of the fundamental properties of living systems at different levels [15]. This is mainly because cell, tissues, organs, organisms or any other biological systems defined by evolution are essentially complex physicochemical systems. They consist of numerous dynamic networks of biochemical reactions and signaling interactions between active cellular components. This cellular complexity has made it difficult to build a complete understanding of cellular machinery to achieve a specific purpose [16]. To circumvent this complexity microarrays and molecular networks can be combined in order to document and support the detected and predicted interactions [17]. The advances and tools that each discipline carries can be integrated in a holistic and generic perspective so that the chaotic complexity of biology networks can be 'screened' and traced down.

2.3. Coupling MAs and GRPs

Microarray experiments involve more variables (genes) than samples (patients). This fact, leads to results with poor biological significance. There is an open debate whether we should concentrate on gathering more data or on building new algorithms in order to improve biological significance. Simon et al. in [18] published a very strict criticism on common pitfalls on microarray data mining while in [19] comments about the bias in the gene selection procedure are presented. Moreover, due to limitations in DNA microarray technology higher differential expressions of a gene do not necessarily reflect a greater likelihood of the gene being related to a disease and therefore, focusing only on the candidate genes with

the highest differential expressions might not be the optimal procedure [20]. Another significant aspect is the noisy content microarray experiments. Appropriate statistical analysis of noisy data is very important in order to obtain meaningful biological information [21], [22]. Evidence on this is given by the fact that different methods produce gene-marker lists that are strikingly different [23]. As a result, and because the immature state of microarray technology, reproducibility of microarray experiments and the accompanied statistical prediction models are pretty low, except when protocols are uniformly and strictly followed [24], [25].

In the light of the aforementioned observations and in order to overcome the posted limitations we have to consider MA-based gene-expression profiles just as an instance of biological information, strongly connected - rather than isolated, from other sources of related biological knowledge. In other words, gene-expression profiles should be examined, explored and interpreted not as 'static' but as instances of the underlying regulatory framework, as encoded by established and known GRPs.

3. Methodology

Existing GRPs databases provide us with widely utilized networks of proved molecular validity. The most known are network that describe important cellular processes such as cell-cycle, apoptosis, signaling, and regulation of important growth factors. Online public repositories contain a variety of information that includes not only the network per se but links and rich annotations for the respective nodes (genes) and edge (regulation). In the current study we utilize the KEGG pathways repository. KEGG provides a format representation standardized by its own markup description language (KGML).

The gene regulatory relations we consider are restricted to what might be observed in a microarray experiment: a change in the expression of a regulator gene modulates the expression of a target gene mainly via protein-DNA interactions. In other words, there are genes that causally regulate other genes. A change in the expression of these genes might change dramatically the behavior of the whole network. The identification and prediction of such changes is a challenging task in bioinformatics. Moreover, we have to identify real, true networks and use them as scaffolds [26] to methods that infer gene regulatory networks out from gene expression data. This approach can aim several areas of biology research such as genomic medicine [27], microarray data mining [28] and phylogenetic analysis [29]. We have implemented our approach on coupling GRPs and MA data in a system called MinePath.

3.1. Pathway decomposition

MinePath relies on a novel approach for GRP processing that takes into account all possible functional interactions of the network, the network's *scaffolds*. The

different GRP scaffolds correspond to the different functional paths that can be followed during the regulation of a target gene.

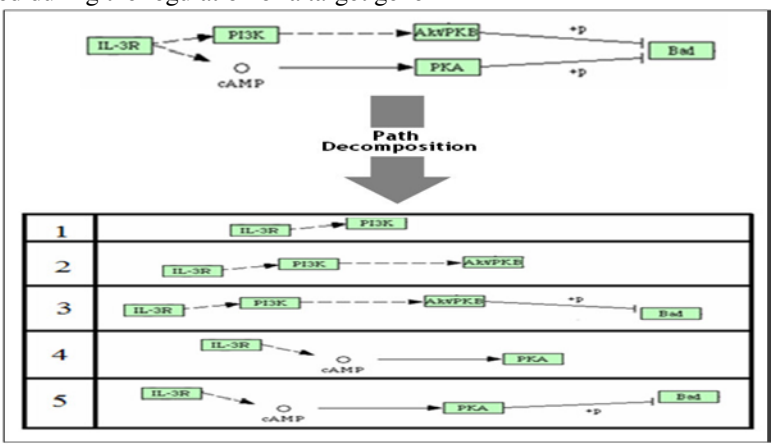


Fig. 3.1. Funtion-path decomposition – the GRP scaffolds: Top: A target part of the KEGG cell-cycle GRP; Bottom: The five decomposed fictional paths (scaffolds) for the targeted path part – all possible functional routes taking place during network regulation machinery.

Different GRPs are downloaded from the KEGG repository. With an XML parser (based on the specifications of the KGML representation of GRPs) we obtain all the internal network semantics (see next sub-section). In a subsequent step, all possible and functional network paths are extracted as exemplified in Fig 3.1. Each functional path is annotated with the possible valid values according to Kauffman's principles that follow a binary setting: each gene in a functional path can be either 'ON' or 'OFF'. According to Kauffman [13], the following functional gene regulatory semantics apply: (a) the network is a directed graph with genes (inputs and outputs) being the graph nodes and the edges between them representing the causal (regulatory) links between them; (b) each node can be in one of the two states: 'ON' or 'OFF': 'ON' and 'OFF' states correspond to the gene being expressed (i.e., the respective substance being present) or not expressed, respectively; and (c) time is viewed as proceeding in discrete steps - at each step the new state of a node is a Boolean function of the prior states of the nodes with arrows pointing towards it.

Since the regulation-edge connecting two genes defines explicitly the possible values of each gene, we can set all possible state-values that a gene may take in a path. Thus, each extracted path contains not only the relevant sub-graph but the state-values of the involved genes as well. The only requirement concerns the following assumption: for a path being functional it should be 'active' during the GRP regulation process; in other words we assume that all genes in a path are functionally active. For example, assume the functional path A → B ('→' is an activation/expression regulatory relation). If gene A is on an 'OFF' state then, gene B is not allowed to be on an 'ON' state - B could become 'ON' only and only if it

is activated/expressed by another gene in a different functional path (e.g., C → B). The assumption follows a 'closed world assumption', that is: if what we know is just the 'A → B' gene-gene interaction then, B could be activated only from A; if A is inactive there is no causal evidence for B being active. If we had allowed non-functional genes to have arbitrary values then the significant paths would be more likely to be 'noisy' rather than exhibiting some form of biological importance.

After parsing the targeted GRPs, the involved genes are stored in a database that acts as a repository for future reference. Through this repository we can query paths being parts of target GRPs, GRPs that contain specific genes or target a specific regulatory relation. Moreover, the stored paths can be combined and analyzed in the view of specific microarray experiments and respective gene-expression sample profiles. As the database repository contain and retrieves functional paths from a variety of different GRPs (e.g., cell-cycle, apoptosis etc), we may combine different molecular pathways and networks – a major need for molecular biology and a big challenge for systems biology and contemporary bioinformatics research.

3.2. Combining gene-expression profiles and functional paths

The next step is to locate microarray experiments and respective gene-expression data for which we expect (suspect) the targeted GRPs play an important role. For example the cell-cycle and apoptosis GRPs play an important role in tumorgenesis and cancer progression.

With a gene-expression/functional-path *matching* operation, the valid and most prominent GRP functional paths are identified. These paths uncover and present potential underlying gene regulatory mechanisms that govern the gene-expression profile of the samples under investigation. Such a discovery may guide to the finer classification of samples as well as to the re-classification of diseases, providing the most prominent molecular evident for that.

3.3. Matching GRP paths with MA data

The samples of a binary transformed (discretized) gene-expression matrix are matched against targeted molecular pathways and respective GRP functional paths (retrieved form the described repository). We follow a gene-expression value discretization process presented elsewhere (please refer to [9]). As already exemplified, GRP and MA gene-expression data matching aims to differentiate GRP paths and identify the most prominent functional paths for the given samples. In other words, the quest is for the paths that exhibit high matching scores for one of phenotypic class and low matching scores for another. This is a paradigm shift from mining for genes with differential expression to mining for subparts of GRP with

differential function. The algorithm for differential path identification is inherently simple (see Fig. 3.3).

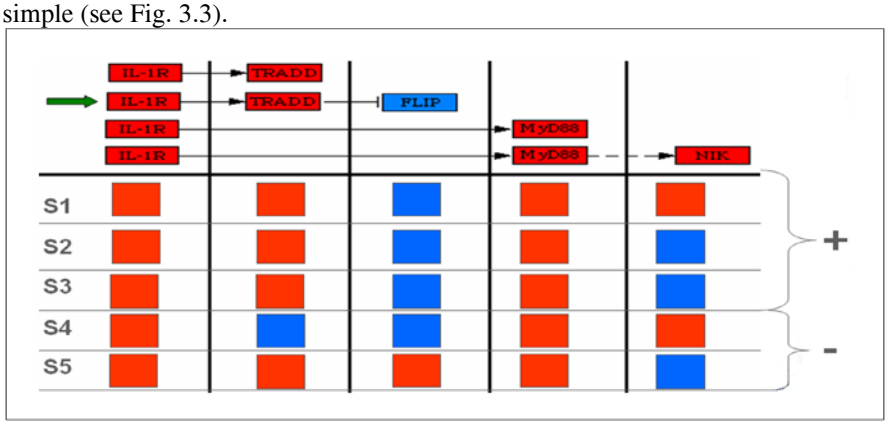


Fig. 3.3. Matching Functional-paths (scaffolds) and gene-expression profiles: Samples S1, S2, S3 belong to the '+' class and samples S4, S5 belong to the '-' class. The first path (IL-1R → TRADD) satisfies samples 1,2,3,5. Second path (IL-1R → TRADD ⊣ FLIP) satisfies samples S1, S2, S3. Third path satisfies all samples and the forth path doesn't satisfy any sample. The green arrow indicates that the second path yields the maximum differential power and it contains a potential function differentiation since it is consisted only with samples that belong to the '+' class. ('→': activation; ' ⊣': inhibition).

– For each path we compute the number of samples that is consistent for each disease phenotypic class. Suppose that there are S_1 and S_2 samples belong to the two classes, respectively. Assume that path P_i is consistent with $S_{i;1}$ and $S_{i;2}$ samples form the first and second class, respectively. Formula 1,

$$\left|\left(S_{i;1} / S_1\right) - \left(S_{i;2} / S_2\right)\right| \quad (1)$$

computes the *differential power* of the specific path with respect to the two classes. Ranking of paths according to formula 1 provides the most differentiating and prominent GRP functional paths for the respective disease phenotypes. These paths present evidential molecular mechanisms that govern the disease itself, its type, its state or other targeted disease phenotypes (e.g., positive or negative response to specific drug treatment). The formula can be enriched so that longer consistent paths acquire stronger power. It can also be relaxed so that 'consistent' is a continuous indicator rather than a Boolean value. Finally we may introduce 'unknown' values for missing and erroneous gene expression values.

4. Revealing Regulating Wilm's tumor Molecular Mechanisms

The presented MA-GRP coupling methodology was applied on a study for expression profiling of the Wilm's tumor (WT, nephroblatoma) disease [30]. In the orig-

inal publication the researchers report new candidate genes for various WT clinical phenotypes.

WT samples were divided according to the *histological risk grade* ('low/intermediate' and 'high'), relapse of tumor ('no', 'yes'), survival ('relapse-free' and 'death'), metastasis ('no', 'yes') and response to chemotherapy ('good', 'poor'). The results presented in this paper focus on the histological risk grade as a target WT phenotype. In the original published study a set of 20 differentially expressed genes are reported for this WT phenotype [30].

From the ArrayExpress online microarray experiments' repository (http://www.ebi.ac.uk/microarray-as/ae/) we downloaded the expression values and the clinical annotation of 138 samples from the WT study - 108 of them being classified as histology risk 'low/intermediate', and 30 as 'high'. For this study we targeted 17 GRPs – the selection was made on the basis of their susceptibility and incrimination to the WT disease and on established biological and clinical knowledge of their involvement in cell regulatory tumor growth mechanisms. The path decomposition process resulted into a total of 8937 functional paths. Most of these paths didn't show any special differential ability over the samples. In order to identify the significant paths the matching gene-expression formula (formula 1 presented in section 3.3) was applied. A threshold value of 0.5 was set to filter-out not differential paths (the threshold was fixed after experimentation with various cut-off values). Filtering resulted into a set of 87 functional-paths for further exploration.

The next step was to find the most relevant and discriminant functional-paths, and build a classifier able to distinguish between the two phenotypic classes - 'high' and 'low' (including 'intermediate' samples) histological risk grade, respectively. The whole dataset is presented as a binary-{0,1} array-matrix of 87 lines for functional paths, and 138 columns for samples. The value "1' in the position *i,j* of this array means that the *i* path is 'active' for sample *j*. Active means that all genes that comprise this path are either 'ON' or 'OFF' according the interaction relationships of the genes of the path. Respectively, a '0' value means that the genes involved in the path do not exhibit the same value as the expression value of the respective sample. The array-matrix can be seen as an indicator of which paths are functional on which samples. Furthermore, it comprises a resemblance of normal gene-expression matrices - instead of genes being either active or inactive, according to their expression over different kind of samples, we have paths being functional or non-functional over the same set of samples. This gives us the ability to apply whatever feature selection processes to select the most relevant and discriminant functional-paths. For this, we rely on a feature/gene-selection algorithmic process presented in [9].

Initially a Wilcoxon rank-sum test ($p<0.005$) was applied that reduced the functional-paths from 87 to 54. Then, ranking and selection of the most discriminant functional-paths was performed – ranking is based on an information-theoretic entropic formula, and selection encompasses a naïve Bayes classification process

[9]. The whole process resulted into a complex of four discriminant and indicative functional-paths (see Fig. 4.)

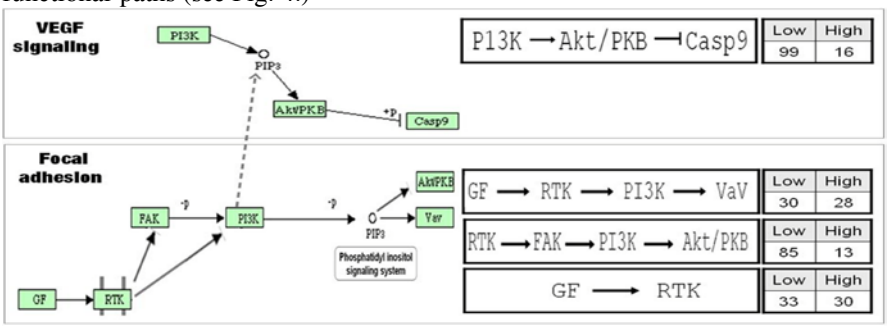


Fig 4. Indicative 'low'/'high' histological risk functional-paths for the WT disease: The GRP name, its KEGG graph representation, the identified functional-paths, and the coverage/discrimination statistics, e.g., (99,16) in the upper part of the figure, indicates that the specific path [P13K → Akt/PKB -| Casp9] covers 99 'low' and 16 'high' risk samples, respectively. The presence of 'P13K' in both GRPs is quite interesting for the biology of the WT disease, and respective therapeutic targets.

The four selected functional-paths are involved in two different GRPs: one from 'VEGF signaling', and three from 'Focal adhesion'. The three functional-paths from 'Focal adhesion' are subparts of the whole 'Focal adhesion' GRP.

We performed a Leave One Cross Validation (LOOCV) procedure in order to assess the discrimination/classification performance that these paths exhibit (note that each functional-path is now considered as a feature). A 95% LOOCV (131/138 samples) accuracy figure was achieved when, the fitness (i.e., train vs. train) figure was inferior, 91% (126/138, 8 misses for 'high' classified as 'low', and 4 misses for the inverse case). This finding is quite interesting: beside the high accuracy performance data 'overfitting' is reduced. This is a strong indication for the high relevance of the four identified functional-paths (at least for the available dataset).

In addition, we applied the same feature/gene selection algorithmic process on the original gene-sample matrix (i.e., the normal gene-selection setting for microarray gene-expression profiles). This resulted into the same LOOCV accuracy and in 89% (123/128) fitness (17 genes selected) accuracy figures. A potential speculation on this finding is the following: with the presence of thousands of genes and of a limited number of samples, gene-selection processes are lean to overfitting events. We believe that this explains the diversity of results produced by available gene-selection techniques, their instability on different population (for the same disease) cohorts, and the inability to relate statistical significance with biological relevance. In contrast, the introduced methodology is able to identify not just discriminant gene-markers but, discriminant, indicative and 'stable' gene-regulatory mechanisms that govern disease phenotypes and clinical manifestations.

In a preliminary attempt to find biological evidence for our findings we focus on the involvement of 'P13K' and 'AkT'gene-products in both identified GRPs (see Fig. 4). The related literature report the 'P13K/AkT' complex to be implicated in WT disease, as well as it is the main component of WT therapeutic targets

[31]. Certainly, further biological validation of the approach is needed, a task for future research.

Conclussions

We have presented an integrated methodology for the coupling of both GRPs and MA gene expression profiles. In the heart of the methodology is the decomposition of GRPs into functional-paths (or, scaffolds), the matching of these paths with samples' gene expression profiles, and the application of feature selection techniques for the identification of the most relevant and discriminant ones.

Application of the methodology on gene-expression data for the Wilms' tumor disease showed that: we can identify a limited number of functional-paths that exhibit significantly differential behavior between different WT phenotypes ('low/intermediate' vs. 'high' histological grade risk). The findings provide valuable insights for further research over the function and role of the involved genes and their underlying regulatory machinery.

Among others, our on-going and future R&D work include: (a) further experimentation with various real-world microarray studies and different GRP targets (accompanied with the evaluation of results form molecular biology and clinical research experts); (b) extension of path decomposition to multiple GRPs; (c) elaboration on more sophisticated path/gene-expression profile matching formulas and operations; (d) incorporation of different gene nomenclatures in order to cope with microarray experiments from different platforms and nomenclatures; and (e) porting of the whole methodology in a Web-Services and scientific workflow environment.

Acknowledgements: This work was supported by the European Commission's Sixth Framework Programme in the context of the ACGT (FP6-2005-IP-026996) Integrated project.

References

[1] J. Bell, 'Predicting disease using genomics', Nature 429, 453-456 (2004).
[2] T. Ideker, T. Galitski and L. Hood, 'A new approach to decoding life: systems biology', Annu Rev Genomics Hum Genet, 2, 343-372 (2001).
[3] F.S. Collins, E.D. Green, A. E. Guttmacher and M. S. Guyer, 'A Vision for the Future of Genomics Research', Nature, 422(6934), 835-847 (2003).
[4] H.F. Friend, 'How DNA microarrays and expression profiling will affect clinical practice', Br Med J., 319, 1-2 (1999).
[5] D.E. Bassett, M.B. Eisen, and M.S. Boguski, 'Gene expression informatics: it's all in your mine', Nature Genetics, 21(Supplement 1), 51-55 (1999).
[6] T.R. Golub et al., 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', Science, 286, 531-537 (1999).
[7] L.J. van 't Veer et al., 'Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer', Nature, 415, 530-536 (2002).
[8] M.E. Troyanskaya, M.E. Garber, P.O. Brown, D. Botstein, and R.B. Altman, 'Nonparametric methods for identifying differentially expressed genes in microarray data', Bioinformatics, 18 (11), 1454-1461 (2002).

[9] G. Potamias, L. Koumakis and V. Moustakis, 'Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination', LNAI, 3025, 256-266 (2004).
[10] J. M. Bower and H.Bolouri, Computational Modeling of Genetic and Biochemical Networks, Computational Molecular Biology Series, MIT Press, 2001.
[11] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, Molecular Biology of the Cell, Garland Science, New York, 2002.
[12] Arkin and J. Ross, 'Computational functions in biochemical reaction networks', Biophys J., 67(2), 560-578 (1994).
[13] S. A. Kauffman, The Origins of Order: Self-Organization and Selection in Evolution, Oxford Univ. Press, New York, 1993
[14] N.M. Babu, N.M. Luscombe, L. Aravind, M. Gerstein and S.A. Teichmann, 'Structure and evolution of transcriptional regulatory networks', Curr. Opin. Struct. Biol., 14, 283-291 (2004).
[15] H. Kitan, 'Robustness from top to bottom', Nat. Genet.,38, 133 (2006).
[16] H. Kitano, 'Systems biology: a brief overview', Science, 295(5560), 1662-1664 (2002).
[17] K. Kwoh and P. Y. Ng, 'Network analysis approach for biology', Cell. Mol. Life Sci., 64, 1739-1751 (2007).
[18] R. Simon, M. D. Radmacher, K. Dobbin and L. M. McShane, 'Pitfalls in the Use of DNA Microarray Data for Diagnostic Classification', Journal of the National Cancer Institute, 95(1), 14-18, (2003).
[19] Ambroise and G. J. McLachlan, 'Selection bias in gene extraction on the basis of microarray gene-expression data', PNAS, 99(10), 6562-6566, (2002).
[20] S. Draghici, S. Sellamuthu and P. Khatri, 'Babel's tower revisited: a universal resource for cross-referencing across annotation databases', Bioinformatics, 22(23), 2934-2939 (2006).
[21] D.K. Slonim, 'From pattern to pathways: gene expression data analysis comes of age', Nature Genetics, 32, 502-508 (2002).
[22] J. Quackenbush, 'Computational Analysis of Microarray Data', Nature Reviews Genetics, 2, 418-427 (2001).
[23] W. Pan, 'A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments', Bioinformatics, 18(4), 546-554 (2002).
[24] MTRC: Members of the Toxicogenomics Research Consortium, 'Standardizing global gene expression analysis between laboratories and across platforms', Nature Methods, 2, 351-356 (2005).
[25] Robert et al. 'Robust interlaboratory reproducibility of a gene expression signature measurement consistent with the needs of a new generation of diagnostic tools', BMC Genomics, 8:148 (2007).
[26] T. Ideker and D. Lauffenburger, ,Building with a scaffold: emerging strategies for high- to low-level cellular modeling', Trends in Biotechnology, 21(6), 255-262 (2003).
[27] M.A. Hoffman, 'The genome-enabled electronic medical record', Journal of Biomedical Informatics, 40(1), 44-46 (2007).
[28] P. Jares, 'DNA Microarray Applications in Functional Genomics', Ultrastructural Pathology, 30, 209-219, (2006).
[29] R. Jothi, T. M Przytycka and L. Aravind, 'Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment', BMC Bioinformatics, 8:173 (2007).
[30] Zirn B, Hartmann O, Samans B, Krause M, Wittman S, Mertens F, Graf N, Eilers M, Gessler M. Expression profiling of Wilms tumor reveals new candidate genes for different clinical conditions. Int. J. Cancer: 118, 1954-1962 (2006).
[31] International Society of Paediatric Oncology - SIOP Education Book. http://www.siop.nl/content/files/SIOP_Educational_Book_2008.pdf Acessed 8 march, 2009.

A Simple Algorithm Implementation for Pattern-Matching with Bounded Gaps in Genomic and Proteomic Sequences, on the Grid EGEE Platform, using an intuitive User Interface

Vegoudakis K. I., Margaritis K. G., Maglaveras N.

Abstract In the last decade an unprecedented development in bioinformatics has been observed. An extremely high number of organisms have been sequenced and included in genomic databases. The huge amount of data produced needs to be stored and processed for further analysis. Scientists, have researched algorithms for finding complicated patterns in DNA sequences, but there is a need for computational power and large storage systems, in order to implement specific algorithms in as many as possible DNA sequences stored in large Databases and save the results produced for future use.

Recently, a number of pattern matching algorithms that allowing gaps have been introduced and Grid is an emerging techology that seems to be helpful in this kind of biomedical research. In this paper, we present our effort towards the construction of a user friendly Interface for accessing the Grid EGEE platform. The Interface is a specific one and was built to perform Pattern-Matching with Bounded Gaps in Genomic and Proteomic Sequences. The algorithm and the Interface is tested with small and large DNA and protein sequences as an input, downloaded from a gene and protein repository, NCBI (National Center for Biotechnology Information) and the gathered results produced are presented.

Vegoudakis Konstantinos
Parallel and Distributed Processing Laboratory,
Department of Applied Informatics, University of Macedonia, 156 Egnatia str., P.O. Box 1591, 54006 Thessaloniki, Greece, e-mail: kostasve@uom.gr

Margaritis Konstantinos
Parallel and Distributed Processing Laboratory,
Department of Applied Informatics, University of Macedonia, 156 Egnatia str., P.O. Box 1591, 54006 Thessaloniki, Greece, e-mail: kmarg@uom.gr

Maglaveras Nikos
Lab of Medical Informatics, Health Sciences Faculty, Aristotle University, AUTH PB 323, 54124 Thessaloniki, Greece, e-mail: nicmag@med.auth.gr

1 Introduction

Bioinformatics is playing an increasingly large role in the study of fundamental biomedical problems and represents a new, growing area of science that uses computational approaches to answer biological questions [1]. The explosion of sequence and structural information available to researchers need to be processed and stored in special supercomputing infrastructures and storage systems respectively. Grid is an emerging technology for distributed computing in advanced science and engineering and was firstly developed as a concept to enable resource sharing within scientific collaborations [2]. Although, great investments have been made the previous years worlwide towards the constuction of Grid Infrastructures, the Grid community expects to evaluate the necessity and the current potential of using the Grid [3].

Grid technologies have enormous potential for heavy computational and storage demanding applications but a few of them have been implemented and executed in the context of Grid computing. There are a lot of reasons of the slow take up of Grid computing but the most substantial one is the lack of an existing simple unique framework and easy to use User Interface for executing applications either biomedical or not. For example the glite WMS (Workload Management System) demands the experience of using Unix-like scripts in order to interact with the Grid infrastructure.

The potential of Grid computing in healthcare has been examined and worked out by the HealthGrid initiative [4] according to which the prospects of Grid Computing are deployed e.g. computational models of systems/organs, pharmaceutical research, large-scale epidemiological studies and so on. The scale and the complexity of the Euroeopan EGEE (Enabling Grid for E-sciencE) project and the launch and the expanding of biomedical applications in it, is described in [5]. It is worth mentioning the BIOMED Virtual Organization (VO) which is hosted on the EGEE project and created in the context of the BioinfoGRID project, whose purpose is to promote the Bioinformatics applications for life science, in order to carry out research based on the Grid networking technology [6].

User-friendly access to the grid environment is of great importance for the scientific community. In particular a lot of effort has been made to construct an interface which is easy to use. The g-Eclipse project aims to build an integrated workbench framework to access the power of existing Grid infrastructures. g-Eclipse tries to ease the execution of the existing applications that are needed to be executed in the Grid environment and it provides tools for the customization of the Grid users' applications and management of the Grid resources [7]. GuiGen is a comprehensive set of tools for creating customized graphical user interfaces and was originally designed for the use in computational grids [8]. In addition, P-GRADE provides a high-level graphical environment to develop parallel applications transparently both for parallel systems and the Grid and it also supports workflow definition and coordinated multi-job execution for the Grid. One of the main advantages of P-GRADE is that the user does not have to learn the different APIs for parallel systems and the Grid. The current version of P-GRADE supports the interactive execution of parallel

programs both on Globus-2 and Condor Grids [9]. However, the above mentioned approaches still require technical knowledge from the end users.

In this work, which is a sequence of previous work on the research of an intuitive user-friendly and generic User Interface [10], [11], either with the help of XML or not, another user-friendly interface for Pattern-Matching with Bounded Gaps in genomic sequences is presented. This inteface is a specific one and it is used only for submission of certain type of jobs, those for string pattern matching with bounded gaps. The user has the ability to interact with the UI for the job submission, job status retrieval, upload/download of DNA and protein sequence files that are needed, etc.

The problem of fast searching of patterns that contain Classes of characters and Bounded size Gaps (CBG) in text occurs in various fields and the most important one, is protein matching. The design of two new practical CBG practical algorithms that are faster and simpler than all regular expression search techniques are described in [12]. In Crochemore et. al. [13], algorithms for several versions of approximate string pattern mathching with gaps are presented. Further restrictions to the gaps are introduced in [14] with lower and upper bounds restrictions on the gaps. The user interface created, implements these algorithms on the Grid platform.

The applicability of the user interface is examined via a set of jobs. Different DNA and Protein sequences were used to search in them for different patterns with bounded gaps. It has to be noted that especially large DNA sequences need a lot of computing time and storage in order to execute. For this reason the interface built, enables the user to submit multiple executions of jobs with string pattern matching with bounded gaps. This method is known as parameter study and it is often met in biomedical applications.

More precisely, the user interface is expected to utilize existing bioinformatics applications on available grid testbed (such as EGEE, NGS, etc), [15]. In its final form it will require only Java. As a result installation and configuration of specific operating system and grid middleware toolkit will not be necessary. The user interface current production status, simplifies the job submission process on EGEE grid infrastructure, which is not a trivial task from a biologist's end-user point of view. The parameter study as mentioned above, gives the biologist an assistance in his work, as he gains time, retrieving all the results from the experiments in a reasonable amount of time.

This paper is structured as follows. First, the methodology adopted is presented, providing the necessary implementation and introductory details about the string pattern matching problem with bounded gaps and the creation of the UI for the Grid. Then, the results obtained from the execution of multiple jobs on the Grid environment, are presented. Finally, our future research perspectives and the conclusions drawn are discussed.

2 Methodology

2.1 Basic Definition of the String Matching Problem with Bounded Gaps

The adoption of two uniformly Alphabets [14] $\Sigma_{DNA} = \{A,C,G,T\}$ and $\Sigma_{Protein} = \{A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$ is needed in order to define the string pattern matching problem with bounded gaps. Each letter is standing for the first letter of the chemical name of the nucleotide in a DNA sequence and each letter in the protein alphabet represents the amino acid abbreviation for each protein in a protein sequence.

Let X be a string drawn from Σ_{DNA}. X represents an array X[1..n] of n$\geq$ 0 symbols, where n=length(X) denotes the lentgh of the string X. X[i] denotes the ith symbol of X. In addition, X[i..j] denotes the substring of X contained between ith and the jth symbol. Given a text T of length n and a pattern P of length m, an occurence with `b`-bounded gaps of P in T is an increasing sequence of indices $i_1, i_2, ..., i_m$ such that (i) $1 \leq i_1$ and $i_m = i \leq n$ and (ii) $i_{h+1} - i_h \leq \mathtt{b} + 1$, for h=1,2,...,m-1. $P \trianglelefteq_b^i T$ means that P has an occurence with `b`-bounded gaps that terminates at position i in text T. In the same way, an occurence with `a`-bounded gaps of P in T is an increasing sequence of indices $i_1, i_2, ..., i_m$ such that (i) $1 \leq i_1$ and $i_m = i \leq n$ and (ii) $i_{h+1} - i_h \geq \mathtt{a} + 1$, for h=1,2,...,m-1. $P \trianglelefteq_a^i T$ means that P has an occurence with `a`-bounded gaps that terminates at position i in text T. Finally, an occurence with `(a,b)`-bounded gaps of P in T is an increasing sequence of indices $i_1, i_2, ..., i_m$ such that (i) $1 \leq i_1$ and $i_m = i \leq n$ and (ii) $\mathtt{a} + 1 \leq i_{h+1} - i_h \leq \mathtt{b} + 1$, for h=1,2,...,m-1. $P \trianglelefteq_{(a,b)}^i T$ means that P has an occurence with `(a,b)`-bounded gaps that terminates at position i in text T.

In [14], four algorithms are described. The interface for the Grid implements specifically the string pattern matching problem with `b` - bounded gaps, `a`-bounded gaps and `(a,b)`-bounded gaps. The three implemented algorithms for the solution of the string pattern matching problems with bounded gaps are described as follows:

Problem 1 (upper bounded gaps) Given a text T of length n, a pattern P of length m and a positive integer `b`, the STRING PATTERN MATCHING PROBLEM WITH `b`-BOUNDED GAPS is to find all positions j in T such that $P \trianglelefteq_b^i T$, for $1 \leq j \leq n$.

Problem 2 (lower bounded gaps) Given a text T of length n, a pattern P of length m and a positive integer `a`, the STRING PATTERN MATCHING PROBLEM WITH `a`-BOUNDED GAPS is to find all positions j in T such that $P \trianglelefteq_a^i T$, for $1 \leq j \leq n$.

Problem 3 (lower & upper bounded gaps) Given a text T of length n, a pattern P of length m and positive integers `(a,b)`, the STRING PATTERN MATCHING PROBLEM WITH `(a,b)`-BOUNDED GAPS is to find all positions j in T such that $P \trianglelefteq_{(a,b)}^i T$, for $1 \leq j \leq n$.

In addition, the fourth algorithm mentioned in [14] is not implemented due to its nature of expecting many parameter, that is, `a` and/or `b` restrictions on every possible gap between every nucleotide. The specific interface could be expanded with Regular Expressions or with the help of a special parser for the input of the fourth algorithm. An other feature Regular Expressions can provide, is finding occurences of patterns with bounded gaps using the IUPAC nucleotide code.

In other words, from a biologist's point of view, Problems 1, 2 and 3 find all positions in a DNA or protein sequence T with at most `b` gaps, at least `a` gaps and between `a and b` gaps respectively, between every two nucleptides in the pattern specified. Setting `a`=0 and `b`=0 turns the problem in finding a specific pattern in a DNA sequence. The problems described above need plenty of time to run with large DNA sequences in length. For this reason, two complementary options are examined: Grid computing which offers the ability to run computationally and storage intensive applications [15] and a GUI friendly enough, in order to submit multiple instances of the algorithm with different `a` and/or `b`. As a result, a biologist can benefit from the fact of retrieving his/her result in a reasonable amount of time.

2.2 The Implementation of the User Interface

The GUI in Fig. 1 incorporates all the necessary steps for submitting and managing a job in a Grid environment. The control panel (Submit job/s, Status, Create Proxy, Save Job/s, Load Job/s, Cancel Job/s) which offers a user-friendly job management, also exists in the specific User Interface designed for string pattern matching with Bounded gaps.

EGEE was the Grid infrastructure that our interface utilized and gLite 3.1 [16] was the necessary middleware for accessing the Grid platform. The GUI was developed in Java and WMProxy API (Application Programming Interface) was used for submitting, cancelling and retrieving the output. WMProxy is implemented as a web service. A web service allows us to take advantage of the benefits of the web, not only to provide information, but also to offer services to a greater community of possible users [17]. Within the bioinformatics community, an average end-user might need to access and use hundreds of databases and tools on a given day [18].

The creation of a VOMS (Virtual Organization Membership Service) proxy certificate for accessing the Grid and the status of the submitted jobs are handled by the gLite user interface via the use of java. For data uploading the gsiFTP client [11] was used for uploading the C++ implementation of the string pattern matching with bounded gaps algorithm.

All the scripts and the JDL (Job Description Language) files that are needed for the submission of the jobs are generated automatically via the GUI. This automization saves time and makes the submission of jobs for the naive user simpler enough. The only parameters the user has to fill in, are the number of the jobs, if he wants to perform a parameter study and then the names of the error and output files. Through

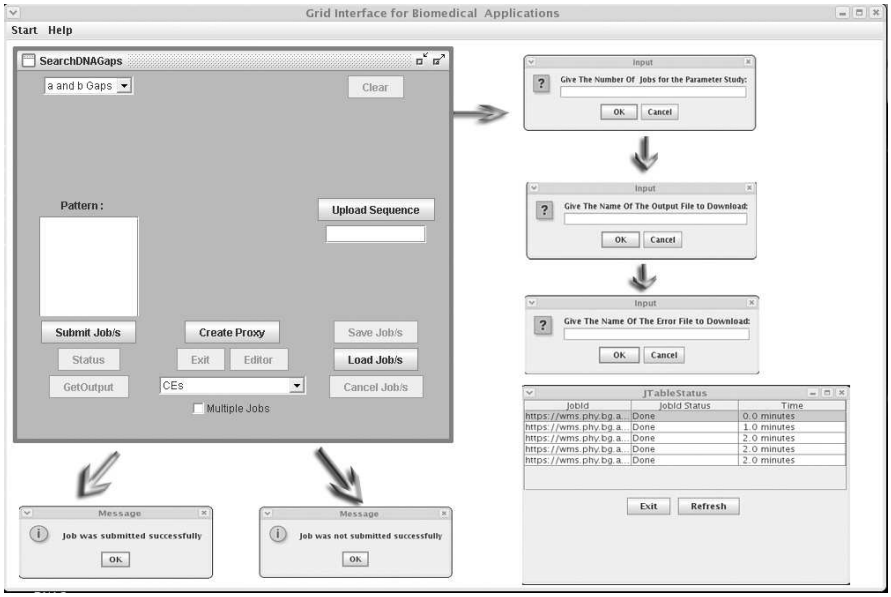


Fig. 1 The specific User Interface, SearchDNAGaps that implements and eases the string pattern matching with bounded Gaps

a list box, the user has also the abilty to choose from a list of predifined CEs (Computing Element), or a random one.

The uploading of the DNA or protein sequences in fasta format is performed, via the use of the GUI. It has to be mentioned that each fasta file should contain only one DNA sequence as an input, because in this implementation different patterns are utilized and searched for occurrences in the DNA sequence. The user is enabled to choose one of the three problems for string pattern matching with bounded gaps mentioned in Sect. 2.1.

An another important element of the specific user interface, is the submission and the management of multiple jobs (Status, Save, Load, Cancel, Get ouput).As a result, the parameter study for different values of the string pattern matching algorithm with bounded gaps (`a,b, a and b`) can be performed and retrieval of output is faster and efficient.

3 Test Case and Results

Our algorithm implementation was tested on the Grid EGEE platform with different DNA sequences that differ in size, downloaded randomly from NCBI. Figure 2[1], illustrates these results. Twelve different DNA sequences ranged from 3626 to

[1] Arabidopsis2 and Arabidopsis3 are segments of the Arabidopsis DNA sequence

30.423.563 base pairs and one small Pattern, ATGCGCG, were used as an input. For each DNA sequence, `a`, `b`, and `(a,b)` parameters were initilized with the same values five times. The values of `a,b` were chosen randomly, according to the three problems described in Sect. 2.1. The results are illustrated in Fig. 2 and one can easily draw the conclusion that the execution time of a job for a string pattern matching with bounded gaps depends on the length of the DNA sequences. More precisely, by keepping the pattern unchanged, if the length of a DNA sequence gets over the eigth million base pairs, then the execution time grows rapidly. For this reason, Grid can be used with multiple job submission and different parameter values, with large DNA sequences in length. Also, the small deviation in execution times for different `a` and/or `b` is due to the fact that CEs were chosen randomly.

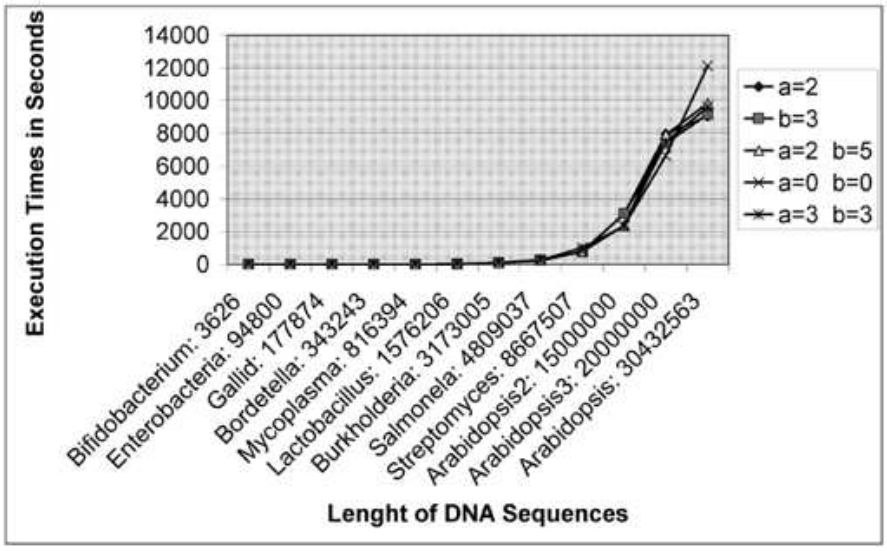


Fig. 2 Execution Times in seconds of the string pattern matching algorithm with bounded gaps for different DNA sequences and (a, b, a and b) parameters

As far as Protein sequences concerned, it has to be noted that they are rather small in length. Titin protein with gi number (genInfo identifier), gi|108861911 was used for testing. The results produced were as they were expected to be. Due to the small sizes of protein sequences, the execution times are rather small, only five seconds. Today, personal computers are very fast and they can handle small DNA or protein sequences efficiently as Fig. 2 shows. The use of Grid is strongly recomended for large sequences, that require hours for the execution of a simple job of string pattern matching.

4 Discussion - Conclusions

In this paper, a specific user-friendly interface was developped for the execution of specific jobs, those for string pattern matching with bounded gaps in DNA or Protein sequences. The rationale of this work is to present the Grid infrastrucure as a mean, that can execute computationally and storage exhausive applications, in the concept of a parameter study approach, by using the multible job submission feature of the proposed specific GUI. Table 1 summarizes and gives a brief explanation of each of the technical terms, projects, and resources appeared in the paper.

The applicability and the potentiality of the proposed specific GUI was illustrated by testing it with different DNA sequences that vary in the size of length. Protein sequences are rather small in size and the algorithm needs only seconds to be executed, so personal computers can be used for this kind of string pattern matching.

The implementation of the algorithm takes as input only one sequence at each job and examines for bounded gaps. In the future, the algorithm and the Grid interface will be changed to query a large number of different sequences saved in fasta format, at each job. Additionally, the combination of the string pattern matching with bounded gaps algorithm, with approximate string matching algorithms [19], will allow to find the positions of a text where a given patterns occurs, allowing a number of ”errors” in the matches with bounded gaps. The parallelization of the algorithm with MPI (Message Passing Interface) and the use of multiple submission or job workflows could expand the features of the GUI, facilitating the user to submit fewer jobs.

The current implementation requires the installation of the gLite middleware in the computer hosting the GUI. This limitation is expected to be solved in the near future and we working on that. Finally, the need of a mobile Graphical User Interface for the Grid using the Java techologies, is in our nearest expectations and we are looking to present one in the nearest future.

Table 1 A brief explanation of each of the technical terms, projects, and resources appeared in the paper

Term	Brief Explanation	URL resource
gLite WMS	gLite Workload Management System	http://glite.web.cern.ch/glite/packages/R3.1/deployment/glite-WMS/glite-WMS.asp
gLite UI	gLite User Interface	http://glite.web.cern.ch/glite/packages/R3.1/deployment/glite-UI/glite-UI.asp
EGEE	Enabling Grids for E-sciencE	http://www.eu-egee.org/
NGS	The National Grid Service	http://www.grid-support.ac.uk/
NCBI	The National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
HealthGrid	HealthGrid Community	http://community.healthgrid.org/
BioinfoGRID	The BioinfoGRID project	http://www.bioinfogrid.eu/
g-Eclipse	g-Eclipse project	http://www.geclipse.org/
GuiGen	GuiGen is a comprehensive set of tools for creating customized graphical user interfaces (GUIs)	http://www.zib.de/schintke/guigen/ index.en.html
P-GRADE	Parallel Grid Run-time and Application Development Environment	http://www.p-grade.hu/
API	Application programming interface	http://en.wikipedia.org/wiki/API
Globus Toolkit	The Globus Toolkit is an open source software toolkit used for building Grid systems and applications	http://www.globus.org/
Condor Project	The Condor Project	http://www.cs.wisc.edu/condor/
CBG algorithms	Classes of characters and Bounded size Gaps algorithms	Navarro G, Raffinot M (2001)
Regular Expressions	Regexp (for short) is a special text string for describing a search pattern	http://java.sun.com/docs/books/tutorial/essential/regex/
IUPAC	International Union of Pure and Applied Chemistry	http://www.bioinformatics.org/sms/ iupac.html
GUI	Graphical User Interface	http://infovis.cs.vt.edu/GUI/java/
WMProxy	WMProxy is a new component to access the gLite Workload Management System (WMS)	http://trinity.datamat.it/projects/EGEE/wiki/wiki.php?n=WMProxyService.AboutWMProxyService
VOMS	Virtual Organization Membership Service	http://www.globus.org/grid_software/security/voms.php
gsiFTP client	A client developed by using GridFTP (GSI enabled FTP) protocol for the file transfers	http://www-unix.globus.org/cog/distribution/1.1/API.html
JDL	Job Description Language	https://edms.cern.ch/file/590869/1/EGEE-JRA1-TEC-590869-JDL-Attributes-v0-8.pdf
MPI	The Message Passing Interface standard	http://www.mcs.anl.gov/research/projects/mpi/

References

1. Baxevanis A. D. , Francis Ouellette B. F.: Bioinformatics and the Internet. In: BIOINFORMATICS: A Practical Guide to the Analysis of Genes and proteins. SECOND EDITION. pp 1-17 John Wiley & Sons, New York, (2001)
2. Foster I., Kesselman C.: Concepts and Architecture. In: The Grid: Blueprint for a New Computing Infrastructure. Elsevier, San Francisco, (2004)
3. Goble C.: The Grid needs you! Enlist now. In: On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, ser. LNCS, R. Meersman et al. Eds. Berlin Heidelberg: *Springer-Verlag* , vol. 2888, pp. 589 - 600, (2003)
4. Breton V et al (2005) The Healthgrid white paper, HealthGrid Proc., Oxford, UK, pp 249 - 321
5. Gagliardi F. et al.: Building an infrastructure for scientific Grid computing: status and goals of the EGEE project, Phil. Trans. R. Soc. A (2005) doi:10.1098/rsta.2005.1603
6. Milanesi L et al (2007) White paper: guidelines and recommendations for the scientific community based on the experience and the results gained from the BioinfoGRID project, http://www.bioinfogird.eu
7. g-Eclipse at http://www.geclipse.eu/
8. Reinefeld A, Stuben H, Schintke F, Din G (2002) GuiGen: A toolset for creating customized interfaces for grid user communities. Future Generation Computer Systems 18(8): 1075 - 1084
9. Kacsuk P., Dozsa G., Kovacs J., Lovas R., Podhorszki N., Balaton Z. and Gombas G. (2003) P-GRADE: A Grid Programming Environment. Journal of Grid Computing 1: 171197
10. Vegoudakis K., Koutkias V., Malousi A., Chouvarda I. and Maglaveras N. A generic Grid interface and execution framework for biomedical applications. BioInformatics and BioEngineering, BIBE 2008. 8th IEEE International Conference.
11. Vegoudakis K., Koutkias V., Malousi A., Chouvarda I. and Maglaveras N. Towards Userfriendly Interfacing of Biomedical Applications with the Grid: A Paradigm with SVM Optimization for Gene Prediction, in Proc. of the 4th European Congress for Medical and Biomedical Engineering 2008 (eMBEC2008), 23-27 November, Antwerp, 2008.
12. Navarro G, Raffinot M (2001) Fast and simple character classes and bounded gaps pattern matching, with application to protein searching. Proceedings of the fifth annual international conference on Computational biology, Montreal, Quebec, Canada Pages: 231 - 240
13. Crochemore M, Makris C, Rytter W, Tsakalidis A, Tsichlas K (2002) Approximate String Matching with Gaps. Nordic Journal of computing, 9(2002): 54 - 65
14. Pinzon YJ, Wang S (2005) Simple algorithm for pattern-matching with bounded gaps in genomic sequences. In Proceedings of ICNAAM05, pages 827 - 831
15. Kransogor N., Shah A.A., Barthel D., Lukasiak P. and Blazewicz J., January 2008. Web and Grid Technologies in Bioinformatics, Computational and Systems Biology: A Review. Current Bioinformatics, 3, pp. 10 - 31(22)
16. gLite documentation at http://glite.web.cern.ch/glite/documentation/
17. Avellino G. et al, 2006. Flexible job submission using web services: The gLite WMProxy experince. 15th International Conference on Computing In High Energy and Nuclear Physics, Mumbai, India, pp.831 - 835
18. Curcin V., Ghanem M. and Guo Y., 2005. Web services in the life sciences. Drug discovery today, 10(12), pp. 865-871.
19. Navarro G., 2001. A Guided Tour to Approximate String Matching. ACM Computing Surveys, 33(1), pp. 31 - 88

Gaussian Mixture Model Coupled with Independent Component Analysis for Palmprint Verification

Raghavendra.R, Bernadette Dorizzi, Ashok Rao, Hemantha Kumar G

Abstract In this paper we present a new scheme for Palmprint verification. The proposed method can be viewed as a combination of Gaussian Mixture Model (GMM) followed by Independent Component Analysis (ICA I and ICA II) applied directly on the pixels. This approach follows the path opened by previous works making use of GMM followed by Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) projection methods for face recognition and is expected to be efficient to tackle major variations in the data. Extensive experiments have been carried out on PolyU palmprint database. We show that ICA I performs better than PCA, LDA and ICA II in the non Mixture Model case, while in the Mixture Model case, ICA II MM outperforms the three other Mixture Models (PCA MM, LDA MM, ICA I MM). Moreover, using artificially corrupted data (noisy palmprints), we show the robustness of Mixture Model approaches to noise, a property which can be very interesting in realistic situations.

1 Introductiong

Gaussian Mixture Model (GMM) has continued to receive a lot of attention over years [2]. GMM provides a mathematical approach to statistical modeling of a wide variety of random phenomena and also provides semi parametric framework

Raghavendra.R
Institut TELECOM, TELECOM and Management SudParis, Evry, France and University of Mysore,India.

Bernadette Dorizzi
Institut TELECOM, TELECOM and Management SudParis, Evry, France.

Ashok Rao
Mentor, C.I.T, Gubbi, India.

Hemantha Kumar G
University of Mysore, Mysore, India

to model an unknown probability density distribution shapes. The combination of GMM with linear projection schemes is becoming popular as it provides more than one transformation matrix. Here the idea is to capture the variability in data while keeping a linear projection method for feature extraction. Two powerful approaches making use of GMM followed by linear projection methods such as Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) are currently in vogue [3][4] for face verification. In this paper, we want to investigate two other projection methods ICA I and ICA II on the different transformation matrices obtained using GMM. Indeed ICA will allow one to represent data in terms of statistically independent variables, it also captures higher order statistics (phase spectrum) with a set of non orthogonal vector basis. We chose to experiment this approach on palmprint because this biometric has several advantages compared with other hand biometrics such as fingerprints [1]. It also presents some variability in illumination by example which is of interest to be tackled using GMM models.

Indeed palmprint contains more information than fingerprint and this way it is expected to be more distinctive. Palmprint can be captured using low resolution devices (as low as 200 dpi), it is invariant with time and is widely accepted by end users. Palmprint contains many features like geometry features, wrinkles, ridges, principal lines etc. In the literature[1], there have been many approaches proposed for palmprint Verification/Identification. Kong et al.[1], used a texture based approach where 2D Gabor filter is used, which have shown better performance than Principal line extraction method. As palmprint contains more distinctive features such as principal lines and wrinkles, much of the work is reported on extraction of the principal lines. Zhang et al.[5], extracted palm lines using twelve templates. Han et al.[6], proposed principal line extraction using Sobel and morphological operation. Hu et al.[7], uses Principal Component Analysis (PCA), PCA and Linear Discriminate Analysis (LDA), PCA and Locality Preserving Projection (LPP) and 2DLPP for palmprint recognition. They show that 2DLPP outperform all other methods. Zang et al.[1], use Fourier transform to extract frequency domain features of palmprint and obtain improved result. Shang et al.[8], proposed the use of Fast ICA on palmprint images. Both ICA I and ICA II architectures are explored and finally the classification is done using RBF neural networks.

In this paper, we propose two novel schemes called ICA I Mixture Model (ICA I MM) and ICA II Mixture Model (ICA II MM) for building a palmprint verification system. As our concern is mainly to show the improvement that can be expected from the joint use of GMMs and ICAs, we work directly on the pixels and do not perform any prior feature extraction on the images. This way, our system will not be optimal compared to state of the art, but we can nevertheless perform some experiments showing the superiority of our approach compared to non mixture models. We also show improvement in performance using ICA I or ICA II instead of PCA or LDA associated to MM. Extensive experiments are carried out on PolyU palmprint database to prove the efficacy of proposed scheme.

The rest of the paper is organized as follows: Section 2 describes the palmprint verification system designed using the proposed methods, Section 3 describes Ex-

perimental set up, Section 4 describes results and discussion and Section 5 draws the conclusion.

2 Proposed Method

Figure 1 shows the block diagram of our palmprint verification system. The proposed system contains three important steps: Preprocessing, Feature extraction and Classification.

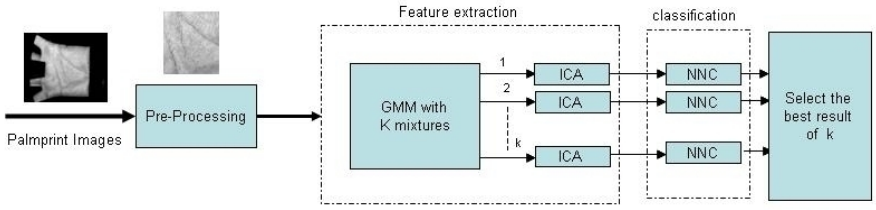


Fig. 1 Block Diagram of Proposed Palmprint Veification System

2.1 Preprocessing

Before feature extraction, it is necessary to obtain a sub-image from the captured palmprint image, and to eliminate the variations caused by rotation and translation. The sub image is also called ROI. This region contains prominent information about palmprint such as: Principal lines, Ridges, Wrinkles etc. Hence, quality of ROI has a great influence on the accuracy of the palmprint verification system. In this paper, we use the method given in [1] to extract the ROI.

2.2 Feature Extraction

This section describes the proposed methods namely ICA I Mixture model(ICA I MM) and ICA II Mixture Model(ICA II MM) for palmprint verification. In our feature extraction model, we first learn the different Gaussian mixtures as done classically in the literature[2]: in practice after determining the right number of mixtures, we use the EM algorithm in order to learn the GMM parameters using the learning(training) database as defined in Section 3. Usually, GMM is used for likelihood estimation of data, but in our approach we use GMM to obtain different transformation (covariance) matrices. Thus, a GMM with k mixtures gives k different transfor-

mation matrices. For each transformation matrix, we determine the associated ICA transformation(ICA I & ICA II separately).

2.2.1 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is a statistical signal processing technique and belongs to a class of Blind Source Separation (BSS) methods for separating data into underlying information components. According to literature [9], there are two ways in which ICA architecture can be implemented in image recognition task. In Architecture I (ICA I) input images in X are considered as a linear mixture of image of a statistically independent basis S combined with an unknown mixing matrix M. The ICA I algorithm learns the weight matrix w that corresponds to the coefficients of the linear mixture [9]. Architecture II (ICA II) finds the statistically independent coefficients for input data. The source separation is performed on the pixels, and each row of learned weight matrix w is an image of M, the inverse matrix of w contains the basis images in the column [9]. In practice, ICA II separates the data taking into account higher statistics while ICA I addresses the variation up to second order statistics. Details of ICA I & ICA II are discussed in [9].

2.3 Classification

After projecting the k transformation matrices using ICA (ICA I & ICA II separately), we do the classification using Nearest Neighbor Classifier (NNC). That is, for each test and training images, we calculate k distances (k is the number of mixtures) using NNC and, at the end, we select the transformation matrix that gives the minimal distance (see Figure 1).

2.4 Model Order

Before using the Mixture model, one has to determine the number of mixture components i.e number of mixtures. Choosing few components may not accurately model the distinguishing features present in the considered modality (i.e palmprint). Also, choosing too many components may over fit the data and reduce the performance and also result in excessive computational complexity both in training and classification. In our experiments, we find the model order by cross validation. Given a training dataset, we evaluate the performance over different numbers of mixture components. We then select the number of mixture components which give the best performance.

3 Experimental set up

The proposed algorithms are validated on polyU palmprint database[1]. The polyU palmprint database contains 7752 gray scale images corresponding to 386 different palm images. Each palm contains twenty samples, out of which ten samples are taken in first session and another ten are taken in second session. The average interval between the first and second session is two months. From this database, we have selected palmprint images corresponding to 150 different users. For training, we selected a total of six samples such that three samples are taken from first session and remaining three are taken from second sessions. Thus, we have 900 palm images for training. For testing, we have chosen 2 samples from each session resulting to a total of 600 palmprint images. We conducted two different experiments:

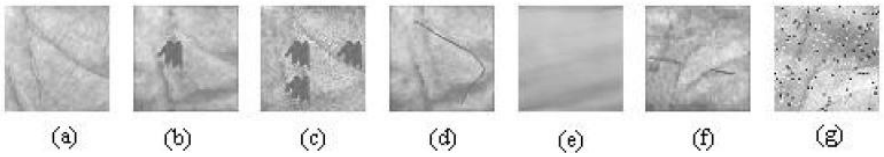


Fig. 2 (a) Clean palmprint (b)-(g). Noisy Palmprints

1. Experiment on clean data (Database I): We use good quality samples to evaluate the performance of proposed algorithms.
2. Experiment on Noisy data (Database II): We use low quality(noisy) samples to evaluate the performance of proposed algorithms. For obtaining low quality samples we artificially introduce five different types of noise such as smudging, lines across & along principal lines, salt & pepper and blurring effect. These noises simulate two different real time conditions. First, blurring of the palmprint image and salt & pepper noise could indicate improper maintenance of the equipment. Second, smudged palmprints and lines indicate improper biometric traits presented by the user. These noisy palmprints along with clean palmprint are shown in Figure 2.

4 Results and Discussion

This section describes the results obtained using proposed algorithms on two different palmprint databases. On each of these databases the results of proposed algorithms are compared with other algorithms such as LDA Mixture model[4], PCA Mixture Model[3] and also with non mixture model approaches (PCA, LDA, ICA I & ICA II).

4.1 Results on Database I

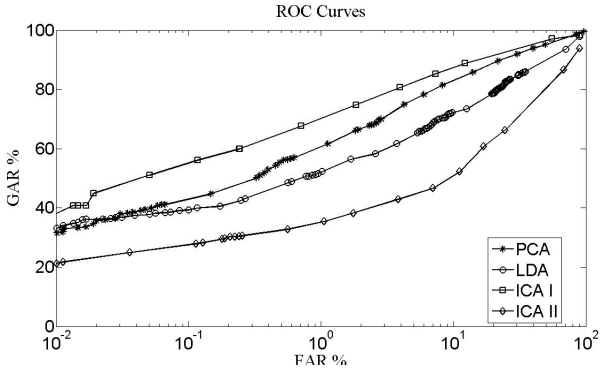


Fig. 3 Performance of Non Mixture Model Approaches on Database I

Figure 3 shows the performance of Non-Mixture Model approaches on clean palmprint samples(Database I). It can be observed that ICA I outperforms all other methods with $GAR = 39.86\%$ at $FAR = 0.01\%$. Figure 4 shows the performance of proposed ICA I MM and ICA II MM along with LDA MM and PCA MM. Here, the best performance is observed for ICA II MM with $GAR = 62.14\%$ at $FAR = 0.01\%$. In order to explain why ICA II is performing better with GMM while it is the worst method in the Non Mixture Model case, we suggest the following interpretation. We think that the GMM will effectively model the higher order statistics while ICA II is able to properly address these higher order statistics presented by GMM. Thus, the combination of GMM followed by ICA II will capture higher order stastistical information that is richer than information used in PCA MM, LDA MM & ICA I MM. For this reason, the proposed ICA II MM shows best performance over the other methods. It is also observed from Figures 3 & 4 that performance of mixture model based approaches are better than Non Mixture Model approaches(average increase of about 22.7% in GAR at $FAR = 0.01\%$).

4.2 Results on Database II

The robustness to noise of the proposed algorithm is validated on Database II. Figure 5 and 6 show the performance of Non Mixture Model along with Mixture Model approaches on the noisy database(Database II). Similar to what occurred on Database I in the Non Mixture Model case, here also ICA I gives the best result with $GAR = 24.2\%$ at $FAR = 0.01\%$ (see Figure 5). Figure 6 shows the performance of proposed ICA I MM and ICA II MM along with LDA MM and PCA MM. Here, it is observed that, ICA II MM outperforms other methods with $GAR = 39.3\%$ at

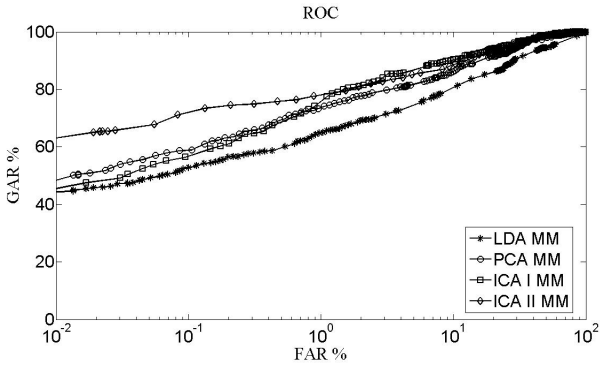


Fig. 4 Performance of Mixture Model Approaches on Database I

$FAR = 0.01\%$ as was already the case for non noisy data (Database I). However the gap in performance (in terms of GAR) between all the algorithms (in both Mixture Model & Non Mixture Model approaches) at low FAR (at $FAR = 0.01\%$) is less improved than in Database I. Let us note, however, a global degradation of the noisy

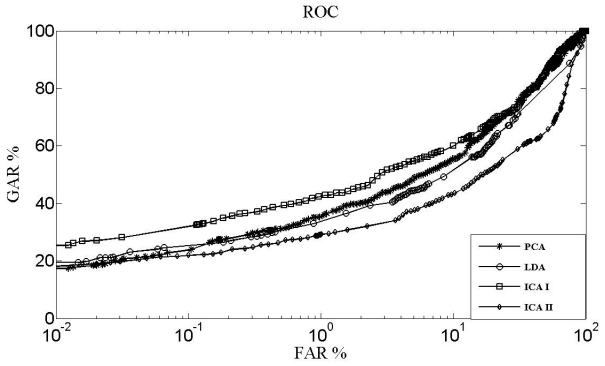


Fig. 5 Performance of Non Mixture Model Approaches on Database II

results compared to those of the clean case for both Mixture Model & Non Mixture Model approaches. Note that the use of Mixture Model based Approaches on noisy data gives a high level of performance (decrease of only roughly 20% as compared to clean case). This is very important in practical application.

5 Conclusion

In this paper we propose two methods for feature extraction namely ICA I MM and ICA II MM. We have conducted extensive experiments on both clean and noisy

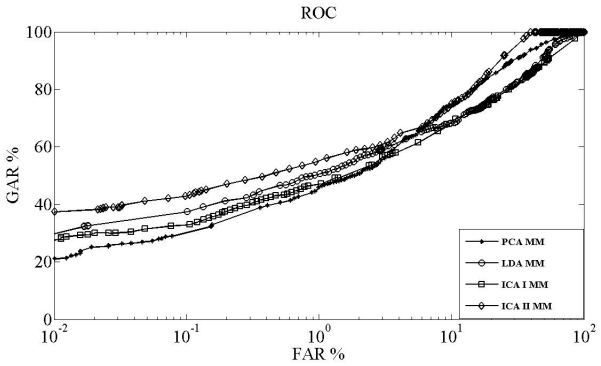


Fig. 6 Performance of Mixture Model Approaches on Database II

databases to prove the efficacy of the proposed methods. Experimental results on both type of data show that ICA II MM out performs the three other mixture model approaches(PCA MM, LDA MM & ICA I MM) which is due to the appropriate use of higher order statistics present in the data. Our experiments also show that Mixture Model approaches out performs Non Mixture Model approaches as they use more than one transformation matrices. This is specially true in the noisy case which is important in practice.

References

1. David D, Zhang.: Palmprint Authentication. Springer-Verlag, 2004.
2. Geoffrey J. McLachlan, David Peel.: Finite Mixture Models. Wiley Series in Probability and Statistics, 2000
3. Hyun-Chul, Kim., Daijin, Kim., Sung Yang, Bang.: Face recognition using the mixture-of-eigenfaces method. Pattern Recognition Letters, 23, 1549-1558 (2002)
4. Hyun-Chul, Kim., Daijin, Kim., Sung Yang Bang.: Face recognition using LDA mixture model. Pattern Recognition Letters , 24, 2815-2812 (2003)
5. Zhang,D., Kong,W., You,J., Wong,M.: Online palmprint identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25, 9, 1041-1050 (2003)
6. Han,C C., Cheng, H L., Lin, C L., Fan, C K.: Personal authentication using palmprint features. Pattern Recognition, 37, 10, 371-381 (2003)
7. Dewen, Hu., Guiyu, Feng., Zongtan, Zhou.: Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition. Pattern Recognition, 40, 339-342, (2007)
8. Li, Shang., De-Shuang, Huang.,Ji-Xiang, Du., Zhi-Kai, Huang.: Palmprint Recognition Using Fast ICA and Radial Basis Probabilistic Neural Network. Nerocomputing, 69, 1782-1786 (2006).
9. Marian Stewart, Bartlett.,Javier, R Movrllan.,Terrence, J Sejnowski.: Face Recognition by Independent Component Analysis. IEEE Transaction on neural networks, 13 06, 1450-1464 (2002)

Support Vector Machines for Dynamic Biometric Handwriting Classification

Tobias Scheidat, Marcus Leich, Mark Alexander, and Claus Vielhauer

Abstract Biometric user authentication is a recent topic in the area of computer security. This paper presents a machine learning approach to single modality user authentication. Here support vector machines (SVM) are employed to classify dynamic handwriting samples. The general goal of SVMs is to carry out binary classifications and/or to handle multiple class problems using a combination of different SVMs. Here a multi-class SVM is proposed to execute verification as well as identification of persons based on their handwriting using a given PIN and a freely chosen PIN. In the best case (trade-off for all rates) for verification using the free PIN a false acceptance rate (FAR) of 0.0083 and an attacker acceptance rate (AAR) of 0.0241 are determined while the false rejection rate (FRR) yields zero. In identification mode using the free PIN, we observe a FRR of 0.0083 and an attacker identification rate (AIR) of 0.2195 at a false identification rate (FIR) level of zero in our experiments.

1 Introduction

The authentication of persons and information plays an important role in information technology. Mostly, user authentication is based on one or combinations of the three factors: secret knowledge, personal possession and/or biometrics. While knowledge and possession provides possibilities to hand over to unauthorized per-

T. Scheidat, M. Leich, M. Alexander, C. Vielhauer
University of Magdeburg, Department of Computer Science, Advanced Multimedia and Security Lab, Universitätsplatz 2, D-39106 Magdeburg, Germany
e-mail: tobias.scheidat@iti.cs.uni-magdeburg.de, {marcus.leich, mark.alexander}@st.ovgu.de, claus.vielhauer@iti.cs.uni-magdeburg.de

C. Vielhauer
Brandenburg University of Applied Science, Department of Informatics and Media, Magdeburger Str. 50, D-14770 Brandenburg, Germany e-mail: claus.vielhauer@fh-brandenburg.de

sons with or without intent or can be lost, biometrics is linked to physical or behavioral characteristics of a person. On the other side, biometric systems have to handle a certain fuzziness of the data of individual persons (intra class variability) and alikeness of data of different persons (inter class similarity). To overcome these drawbacks in a sufficient manner, a great variety of biometric authentication methods were proposed. Following, a small selection of such methods is given without neglecting other publications.

An overview on performance of machine learning techniques in biometrics has been published by Abreu and Fairhurst [1]. Here eight machine learning techniques have been used for classification of fingerprint and signature samples. Mean and standard deviation for absolute error counts are provided for performance evaluation. SVMs among neural net classifiers reached the lowest classification error rates.

A different approach is taken by Fuentes et al.[3]. Instead of using a SVM directly for classification it is used to fuse matching scores of two expert system for online signature verification. The idea behind this concept is to combine the unique strengths of both systems into a single one. The system operates on a test set which partly included skilled forgeries and reaches FAR values of 0.046 and FRR values of 0.083.

In this paper a multi-class SVM is suggested and evaluated for dynamic handwriting verification and identification. The evaluation shows very promising results based on a database of 30 writers with regard to the measures used, false acceptance rate (FAR), attacker acceptance rate (AAR) and false rejection rate (FRR) for verification and false identification rate (FIR), FRR and attacker identification rate (AIR) for identification. Because of the limited number of samples available for testing and because of the different nature of performance measures employed (error rates vs. mean of error counts) and samples semantic results are of limited comparability to those of Abreu and Fairhurst [1] and Fuentes et al.[3].

This paper is structured as follows: The next section describes fundamentals of support vector machines and the configuration of the suggested SVM for handwriting verification and identification. In section 3 the evaluation setup, methodology and results are presented and discussed, while the forth section concludes the contribution and provides an overview of future work.

2 Materials and Methods

This section provides an introduction to SVMs and their use for multi-class classification. Additionally details of the features extracted from the handwriting samples are provided. These features represent the components of the sample vectors presented to the SVM for training as well as for later classification. This section concludes with a discussion of possible problems that can occur during SVM training and classification and presents an approach to overcome these difficulties.

Support Vector Machine: In their basic form SVMs are limited to solving binary classification problems for linear separable classes. These limitations can be overcome by using sophisticated kernel functions that enable the SVM to solve more complex binary classification problems. Additionally several SVMs can be combined to solve multi-class problems.

Since in biometric systems the feature vectors for several persons are not expected to be linear separable, the well-known radial basis function kernel is chosen. Furthermore the *one-against-all* approach [4] for implemantation of multi-class is implemented in the following way: For each person p of the N persons that are to be enrolled we train one SVM using the enrolment samples of p as positive samples and all other other enrolment samples of the remaining $N-1$ persons as negative samples. Consequently, after the enrolment process the entire systems consists of N SVMs, one for each enrolled person.

Based on this system structure it is easily possible to devise an identification and verification procedure for new samples from users that try to authenticate on the system. In the verification scenario a user tries to be verified as an enrolled person. The sample from this user is simply presented to the matching SVM. The user is then accepted or rejected based on the SVM output. The identification process works similarly. Here the sample of the user is presented to all SVMs. If no SVM accepts the sample, the user is rejected. If only one SVM accepts the sample, the user is identified as the corresponding person. If more than one SVM accepts the sample, the result is ambiguous which again leads to the rejection of the user.

For the experiments described in this paper the LIBSVM[1] [2] as pre-existing implementation of the SVM classification algorithm is chosen.

Features: During the data acquisition a sequence of five physical values is sampled by the handwriting sensor time dependently. These values are the X and Y position, the pen tip pressure and the angles azimuth and altitude. Based on these values for each handwritten samples a set of 103 (first 69 features are described in [6], features 70-103 are based on current work) statistical features is determined, which represents the corresponding sample. These feature sets are used as input for the evaluation of the authentication performance of the SVM system.

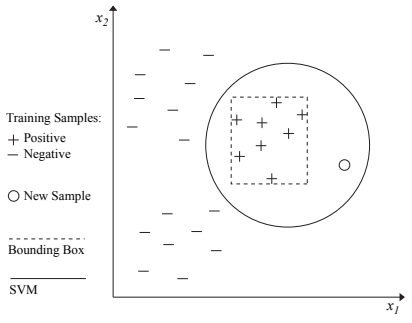


Fig. 1 Bounding box for samples of one person. The SVM classifies the new sample as positive while the bounding box rejects the unknown data.

[1] Version 2.87

These statistical features are partly based on dynamic (e.g. writing time, pen down time, min/max velocity in X or Y direction) but also on static (e.g. aspect ratio, intersections of the writing trace with itself or helper lines) characteristics of the sampled handwritten data. Some features or groups of them are identical or quite similar to 8 out of the 18 features used by Abreu and Fairhurst in [1].

Supporting Bounding Box: Even if the SVMs reach excellent classification rates for samples of trained persons the system might still by prone to false-classification of samples from not enrolled persons. This is due to the fact that during SVM training the feature space is not exhaustively covered with negative samples in regions where no positive samples exist. These regions can be assigned to any class by the SVM without affecting the reclassification rate of the training samples. Fig. 1 illustrates the problem.

To achieve a higher rejection rate for samples of unknown persons it was decided to apply a simple bounding box heuristic. A hyper-bounding box is computed for each person based on the training sample set. The box simply consists of the minimum and maximum values observed for each component of the training samples. A sample is rejected by the bounding box if its components are not within the corresponding ranges of the bounding box. An SVM with bounding box heuristic accepts a sample only if both the SVM and the bounding box accept the sample.

3 Evaluation

This section provides details about the data set and the training and test procedures the experiments where conducted with. Based on that, an analysis of the observed error rates is presented.

Data Set: During data acquisition test persons were ask to donate 8 samples each per semantic class. In case of dynamic handwriting, semantics are alternative written contents to the signature. In this work 30 persons and the semantics *given PIN* and *free PIN* were chosen. While the *given PIN* consists of the default sequence of five digits (77993), the *free PIN* can be freely chosen by the writer under the restriction to write exactly five digits, too.

The values of all samples have been scaled to $[0, 1]$ separately for each semantic using the minimum and maximum values of each feature observed for all persons.

Methodology: For the experiments it was chosen to randomly split the samples of the 30 persons available into two equally sized groups. The first group consists of the 15 persons that are to be enrolled in the system (enrolment group). The second group consists of the remaining 15 persons (attacker group). This group is used to simulate blind attacks of not enrolled persons. The samples of all persons are stored in lists in which each sample has a fixed position. The position of the samples in these lists are randomised one time before all experiments take place.

For a single SVM parameter combination the following test procedure is undertaken: For each person a training and test sample set is determined by selecting two index values m and n with $m \neq n$ as index to the sample list of that person. The samples at the positions m and n are added to the test set of the person, samples at all other positions are added to the training set. m and n are equal for all persons.

After all training and test sets have been created one SVM is trained for each person using this person's training samples as positive samples and all other persons training samples as negative samples. The resulting set of SVMs now represents a multi-class classifier as described in 2.

Using the test samples of all persons the performance measures of this classifier can be determined. For verification each test sample is presented to the system 15 times, each time using a different person as claimed identity. One of these tests can produce at most one false reject (claimed person is the same as the actual sample origin and the sample gets falsely rejected) and at most 14 false accepts (claimed person a different from the sample origin, but the sample is accepted). For identification the sample is simply presented to the entire system. This single test produces at most one false identification (sample is identified as belonging to a different person than the sample origin) or at most one false reject (person is rejected, though the sample origin is a enrolled person).

Additionally all samples of the attacker group are presented to the system. For verification one sample can cause at most 15 false accepts (attacker accepts) while for identification one sample can produce at most one false accept (attacker accept/identification).

The entire procedure (starting from the selection of m and n) is repeated 28 times, so that all possible combinations for m and n are iterated. Over all these experiments all false accepts/identifications, false rejects and attacker accepts/identifications are summed up for verification and identification respectively. The corresponding error rates: FAR/FIR, FRR, and AAR/AIR are computed by dividing the summed errors by the number of maximum number of times this particular type of error can occur. Consequently all are normalised to $[0,1]$. Additionally for identification $FIR + FRR \leq 1$ holds true.

The error rates just described are computed for various SVM parameter combinations. In the chosen experiment setup, we vary the SVM cost parameter C ($[2,194]$, stepsize: 8) and the radial basis function kernel parameter γ ($[0.0002, 0.0194]$, stepsize: 0.0008). In previous tests these parameters proved to be most influential to the classification performance, being most sensitive within the tested intervals. In their effects on error rates these parameters behave similar to the threshold value in distance based classifiers. However, the behaviour of these parameters is not monotonic as it is often the case for threshold values. All remaining LIBSVM parameters were left at their default values.

Results: As presented in the preceding section each parameter combination $(C;\gamma)$ used for training/testing can be evaluated using the three measures FAR/FIR, FRR, AAR/AIR. These 3 dimensional tuples are from now on referred to as *operating points* (OP). The following rates have been determined without using the bounding

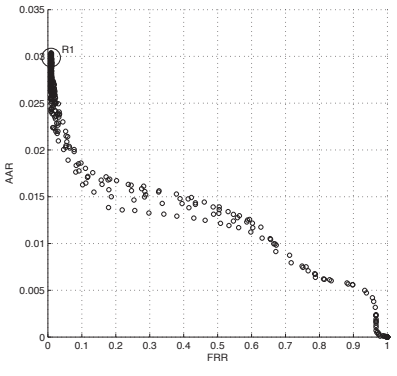


Fig. 2 FRR-AAR-projection of all observed operating points of the verification system (*free PIN* semantic). Region R1 contains all points with FAR = 0.000085. For all other operating points FAR is zero.

box heuristic: For verification most parameter combinations tested (74%) reach an FAR of exactly zero. All other cases reveal a very low FAR of 0.000085. Since the system does not reach equally low FRR values it is impossible to provide an equal error rate (EER) according to the traditional definition. However a modified EER based on AAR and FRR can be estimated if the OPs are projected onto the FRR-AAR plane as depicted in Fig. 2. In this representation the modified EER is about 0.025. As can also be seen the tested parameter combinations are capable of constructing quite effective SVMs (very low FRR) as well as extremely over-fitted SVMs (FRR=1 while still retaining an FAR of 0). The reason for this tendency to over-fitting is the massive class imbalance when one SVM is only trained with positive samples of one person and negative samples of all remaining person. This class imbalance also abets the division of the feature space into disjoint regions in which only samples of a specific persons are accepted.

This behaviour is also reflected in the AAR which converges to zero for high FRR values (over-fitting) and reaches values around 0.03 for very low FRR values. It should also be noted that the highest AAR values are reached for operating points with an FAR not equal to 0 (see region R1 in Fig. 2).

In identification mode the system behaves similar to the verification mode. Differences can be found in the FIR values, which are now consistently zero and in slightly higher AIR values. The reason for the lowered FIR is the stronger rejection criteria which rejects a sample if it is accepted by more than one SVM. Obviously the few false rejects observed during verification testing are the result of two SVMs accepting the same sample. Similarly, because of this stronger rejection criteria the AIR values for identification should be lower than AAR for verification. However the opposite is the case, AIR values are consistently larger for identification. The reason for this is the already mentioned almost disjoint separation of the feature space among SVMs. This separation is the reason that, for verification, a sample causes only one out of 14 possible false accepts, while for identification, the same sample causes one false accept out of one possible false accept which leads to a much higher influence to the AIR.

Table 1 depicts exemplary OP for identification and verification using *free PIN* and *given PIN* semantic. The two OPs chosen for each semantic and authentication

Table 1 Selected operating points

Semantic	Verification				Identification			
	$(C;\gamma)$	FAR	FRR	AAR	$(C;\gamma)$	FIR	FRR	AIR
free PIN	$(26, 0.0154)^a$	0	.0095	.0241	$(114, 0.0018)^a$	0	.0905	.0924
free PIN	$(34, 0.0194)^b$	0	.0083	.0254	$(34, 0.0098)^b$	0	.0083	.2195
given PIN	$(90, 0.0194)^a$	.0016	.0560	.0309	$(58, 0.0066)^a$	0	.1524	.235
given PIN	$(90, 0.0194)^b$	.0016	.0560	.0309	$(154, 0.0194)^b$	.0012	.0738	.2933

[a] parameters for OP with minimum distance to origin
[b] parameters for OP with minimum FRR (if several OPs with the lowest FRR exist, the one with the lowest AAR/AIR is displayed)

mode are the ones with the smallest Euclidean distance to the origin[2] and the ones with the smallest observed FRR.

For the *given PIN* the measured rates are generally higher then the rates for the *free PIN* semantic. Obviously the features described in 2 have better discriminatory properties if they are extracted from samples that differ in content and writing style instead of writing style alone.

It has to be noted that a variety of parameter combinations can lead to satisfying results depending on which error rate is preferred to be minimised. The values of these parameters are also dependant in the data set used for testing as can easily be seen in the varying coordinates in Table 1.

The bounding box approach introduced in 2 performs as expected. At the cost of a very high FRR it is able to reduce the AAR/AIR rate. However, if the size of the bounding box is increased, FRR values drop and AAR/AIR values converge to the previously observed levels.

In previous work [5] a verification algorithm using biometric hashing based on the first 69 features used here has been described. For the semantic given PIN an EER of 0.0832 was determined. It has to be noted, that the underlying database and test methodology are not identical to those used here. However, a trend towards change for the better can be indicated by the usage of the SVM and an enhanced feature set for verification performance.

4 Conclusion and Suggestions for Future Work

The SVM classifier described in this paper is capable of reliable identification of 15 persons who are enrolled using a freely chosen PIN. Because of the tendency to over-fitting, the system exhibits very low FAR/FIR values while producing a still acceptable FRR under 1%. If the enrolled persons are restricted to using the same PIN, system performance drops notably, though this drop affect mostly the FRR which in certain use cases may sill within acceptable ranges.

[2] the origin corresponds to the OP with $FIR = FRR = AAR = 0$ or $FIR = FAR = AIR = 0$

If a more balanced FAR/FIR-FRR-ratio is desired the ν-SVM is an option to consider, since the ν parameter provides information on the class imbalance and hopefully eliminate the tendency to over-fittng. However the elimination of over-fitting might result in an increased AAR/AIR rate, since the positive regions in the feature space are likely to grow in area and thus leave more room for acceptance of samples that otherwise would have been rejected.

The simple bounding box approach suggested in this paper proves to be not suitable to reliable distinguish samples of an enrolled person from samples of a not enrolled person. More sophisticated approaches might be able to reduce the AAR/AIR without extreme effect on the FRR.

Another approach to consider is the one-class SVM which relies solely on positive samples for training. After training the SVM returns whether a presented sample fits the learned distribution. Not being exposed to a majority of negative samples is likely to prevent excessive over-fitting. The total lack of negative samples might lead to an increased FAR/FIR, though.

Furthermore the experiments described in this paper (as well as future experiments based on this work) are to be conducted on a larger data set to ensure statistical significance of the results. It is also considered to adapt the test environment to that of Abreu and Fairhurst [1] to allow direct comparison. Seeking for a deployment-ready authentication system the influence of user count, authentication type, and sample semantic on the parameters of optimal OPs have to be analysed.

Acknowledgements This work is partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project WritingPrint). We would particularly like to thank Jana Dittmann for her support and fruitful discussions in the context of our work.

References

1. Abreu, M.C.C., Fairhurst, M.C.: An empirical comparison of individual machine learning techniques in signature and fingerprint classification. In: B.A.M. Schouten, N.C. Juul, A. Drygajlo, M. Tistarelli (eds.) BIOID, *Lecture Notes in Computer Science*, vol. 5372, pp. 130–139. Springer (2008)
2. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001). Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
3. Fuentes, M., Garcia-Salicetti, S., Dorizzi, B.: On-line signature verification: Fusion of a hidden markov model and a neural network via a support vector machine. In: IWFHR '02: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02), p. 253. IEEE Computer Society, Washington, DC, USA (2002)
4. Hsu, C.W., Lin, C.J.: A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks **13**(2), 415–425 (2002)
5. Scheidat, T., Vielhauer, C., Dittmann, J.: Single-semantic multi-instance fusion of handwriting based biometric authentication systems. In: Proceedings IEEE International Conference on Image Processing (ICIP 2007), pp. 393–396 (2007)
6. Vielhauer, C.: Biometric User Authentication for IT Security: From Fundamentals to Handwriting. Springer, New York (2006)

Applied Surveillance using Biometrics on Agents Infrastructures

Manolis Sardis, Vasilis Anagnostopoulos, Nikos Doulamis

National Technical University of Athens, Department of Telecommunications & Software Engineering, 15773 Zografou Campus, Athens, Greece,
WWW home page: http://www.ece.ntua.gr
emails: {sardis, vanag}@telecom.ntua.gr, ndoulam@cs.ntua.gr

Abstract. The biometric identification of individuals has many safety and security applications in several fields of major interest. This can range from healthcare to commerce and from defense to banking systems. Current biometric technologies are hardly suitable for remote online massive use. This paper presents an innovative approach to design advanced biometric systems for personal identification, recognition and monitoring, with high level of privacy and scalability. The proposed system consists of a modular design using distributed systems, multi agent technologies, adaptable techniques that allow the creation of a surveillance infrastructure based on biometrics sensors. The proposed infrastructure is characterized by scalability and adaptability on sensors (cameras) number and signal manipulation through the multi agent infrastructure. System results can be extended to link more enterprises and application domains that need person identification and related events for access control on critical resources.

1 Introduction

Safety and security airports, public buildings, and private domain areas, like corporate buildings, is of paramount importance nowadays. Consequently we need secure ways to identify human actions. The starting point of unpredictable unsecured actions is usually the entrance on the specific sensitive place, where we need surveillance. In order to avoid such conditions we propose through this work an agent infrastructure that achieves practical human surveillance using camera sensors and biometrics technologies.

The application of advanced techniques for person recognition and identification is directed to verify if a person belongs or not to a prefixed group, denying or granting the access to places, information and services. Even though specific suspicious persons do not need to be identified or recognized, early detection of suspicious behaviors is applied to recognize malicious intentions in advance and to

prevent dangerous acts. Authentication is testing if an individual has access to proceed.

In the last years using biometric techniques to develop services for surveillance was the common approach [1]. Biometrics is the science and technology of measuring and analyzing human body characteristics. Systems based on this technology are pattern recognition systems that acquire biometric data, a set of characteristics, and a mechanism to compare these features against a collection of previously acquired templates. There is evidence that this procedure can be automated by the use of Haar wavelets in a supervised learning manner [11] and this is one of the motivations behind this paper.

Authentication is a simple one-to-one verification process where an acquired pattern is compared against a single stored template in order to determine whether the user is the individual who claims to be. Identification is a one to many process where the pattern at hand is compared against a database of multiple stored templates in order to establish the identity. Most of the biometric devices are operating in *authentication mode* using the following two steps: (a) an identity is requested from a database where participants templates have been previously stored, (b) then by presenting a live biometric sample for comparison the intelligent system provides or not the related labels and outputs.

The *identification mode* is working directly by doing a search in the database based on the acquired biometric sample. In general the biometric framework should be practical in terms of performance and acceptability and circumvention. Data fusion techniques should be incorporated in order to eliminate the biometric system errors [2]. Face analysis techniques [3] present the best results for identification and recognition. The system examines the faces and attempts to find visually similar matches in its stored database visitors. The agent technology controls the related events from the GUI to the external systems connected on the infrastructure or the end users/operators. The operator is able to provide feedback to the infrastructure in the case of misclassified visitors. The face recognizer is working online and is able to learn online. Face recognition under unconstrained conditions is very challenging. The purpose of the visitor identification infrastructure is to provide not only access control but a whole adaptive solution for security using surveillance and biometrics techniques. These solutions are extremely useful for enterprises to increase the domestic and worldwide markets.

The paper is structured as follows: Section 2 analyzes the design considerations of the proposed infrastructure. Section 3 proposes multi-agent technology combined with biometrics sensors as a solution. In Sections 4 and 5 we present the system design and related use cases for the system evaluation. Finally, in section 6 conclusions and remarks close the paper structure giving future research topics.

2 Design considerations

To design ambient intelligence environments, many methodologies and techniques have to be merged together originating many approaches present in recent literature [4]. From the wide range of the possible solutions any ambient intelligent environment is characterized by the following goals and technologies.

- Adaptability, the infrastructure is reactive to new habits, behaviors and needs from the operators and the working environment.
- Context awareness, the infrastructure has the ability to recognize people and the situational context.
- Flexibility, for the operators to customize the control outputs from the infrastructure to other systems connected on it.
- Extensibility, for the surveillance objects and their attributes that can be implemented and controlled by the infrastructure.

The above characteristics are achieved by the use of the artificial intelligence and the proposed multi agent platform. System independence and transparency is achieved by the use of the biometrics sensors that bring the face recognition signals into the infrastructure and are able to verify the correspondence between people requesting access and authorized accounts, which have been stored during the learning procedures from the infrastructure, in system storage module.

Finally, the automatic validation and recognition allows the infrastructure to adapt related services and the according environmental parameters to the system operators' requests and preferences of the recognized objects/persons, during the surveillance. The following paragraphs are analyzing the above infrastructure in more details.

3 Biometrics and multi agents

The proposed infrastructure has been implemented using a distributed collection of agents [6]. The number sensors /cameras provides adaptability to different conditions and environments as biometric characteristics that characterize the identification, recognition and monitoring of the surveillance objects, allow the increase of identification and recognition probability percentage. Biometric integraability is a relevant issue in advanced system design. Also system adaptability is supported by allowing the participation of different agents, which manipulate different algorithms and techniques for surveillance, in the proposed infrastructure allowing the inclusion and deployment of more advanced and new solutions, without changing the overall system architecture and structure. Based on "Fig. 1" the *Surveillance Input* agent represents the biometric sensor that is used for the surveillance. In our case study the system is based on cameras. The *Agent for recognition* uses algorithms for the processing of the data input. With the help of

Evaluator Agent and *Agent for Event creation* the system does the recognition and identification of the objects/humans and prepares the related *events* that will trigger the graphical user interface of the infrastructure. The role of the *Output controller* is to manipulate the output data to other external systems that have been integrated in the proposed infrastructure, either as an extension or as an integrated part of the system. The agent community needs a coordinator of all agents, and this is performed by the *Agent Network Controller*.

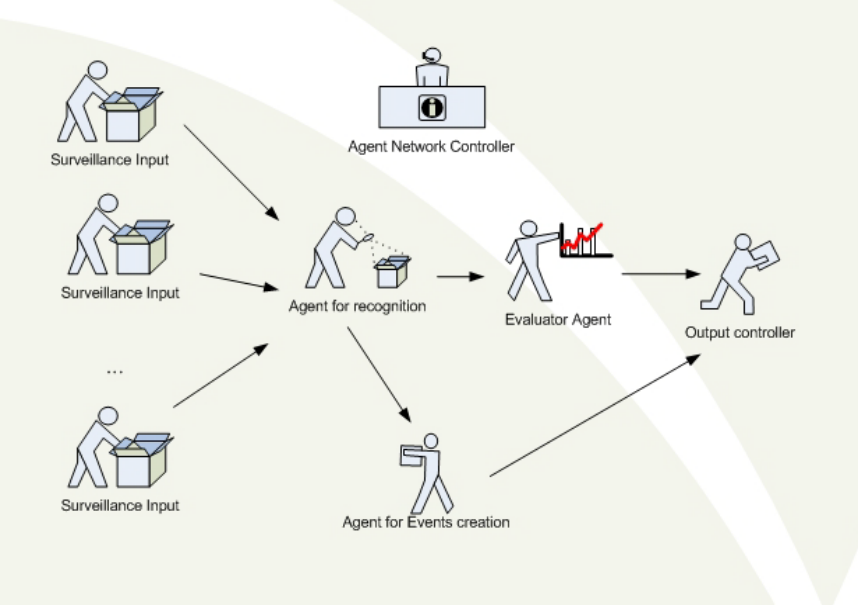


Fig. 1, Multi Agent Infrastructure

The parallel use of biometric sensors and algorithms using agents to control them provides distributed information processing [5] through a modularly scalable software architecture. This architecture supports the integration of heterogeneous and legacy systems by encapsulation in related agents.

The face detection scope is to examine a camera image and extract regions containing a face. In this infrastructure the face detector described in [8] by Viola and Jones has been employed. This detector scans square regions at various scales in the image and extracts features in the form of Haar wavelets to determine whether the region contains a human face. The selection of this detector from a number of other published detectors, like [9] and [10] pool of detectors (references), was based on acceptable detection accuracy requirements, its readily available implementation in the OpenCV library and finally its suitability for real-time processing. Different camera positions generate images for various face orientations for increasing accuracy. However, in this specific application there are physical constraints as to the distance of the human from the camera and to the orienta-

tion of the face. Two cameras suffice to have acceptable performance. Moreover the feedback from the system, guides the user to correct his placing with respect to the camera.

The face recognition procedures rely on a previously acquired collection of several images per visitor, in a controlled environment, and fast modeling of a variety of illumination conditions in the form of normalized color spaces. The output from the multi-agent recognition module will trigger the graphical user interface, giving the possibility to the infrastructure operators to provide feedback to the user and fine tuning of the system.

4 System Architecture

The proposed infrastructure has been implemented using a distributed collection of agents based on the architecture of "Fig. 2".

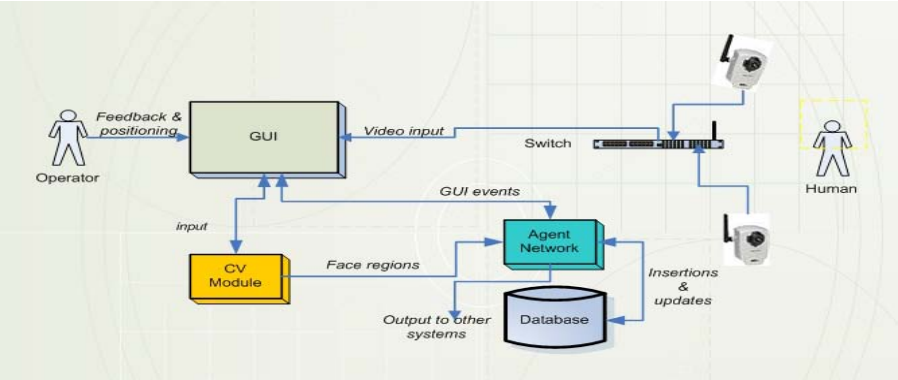


Fig. 2, System Architecture

The architecture we adopted for the problem at hand is comprised by various modules which we will analyze now. We have implemented a module for the Axis 207w cameras which grabs the provided JPEG pictures. The video input is controlled by the GUI which also presents the frames to the operator. This GUI also sends the frames to the computer vision module, for face region detection and extraction. In order to speed up the procedure, the regions are sent directly to the multi-agent infrastructure module for recognition. In addition to the mentioned role this module also controls database access. Successful recognitions are fed back to the GUI in the form of a series of matches ordered by decreasing probability via an event mechanism.

In order to optimize the interaction diagram for performance we exploit the hierarchical nature of the Haar wavelets in our setup. We didn't choose a complex methodology like in [7], since our problem is simpler. By their definition each newer level detects details in half the previous level. For this reason, since our im-

ages are rectangular each new level generates four new sub images. Our network is a graph of identical peers. These peers take as input two images of power of 2 side lengths, wavelet-tranform them and compute inner products. An inner product is also performed in order to detect similarity between the two transforms. If they are greater than a threshold, an agent forwards the four equally sized parts of the image (SouthEast, SouthWest, NorthEast and NorthWest) to its neighbors ("Fig. 3"). We define neighbors as the agents having number with residual equal to the initiator minus one mod 4. We design our system around 16 agents and our algorithm works in depth 4 per side. For this reason, the Evaluator expects 16 messages at depth 4 per stored image. The percentage of them that arrives at the evaluator is the probability of success.

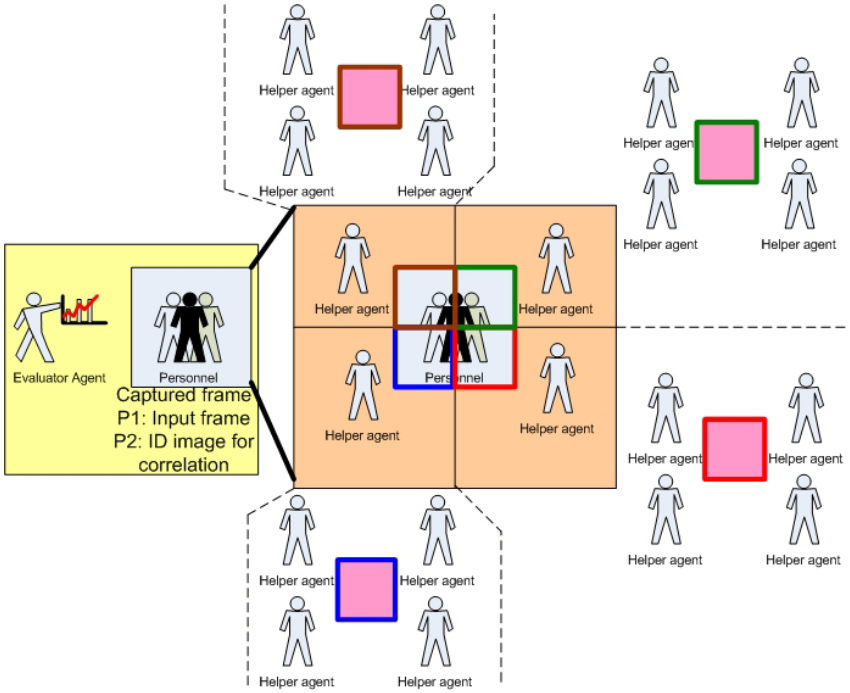


Fig. 3, Agents subsystem decomposition

For speedier calculations the corresponding transforms of the images can be pre-computed and retrieved by the orientation and depth pair of the region in the images. In our case, we do not scan the whole image via translations because this step is already done by the face detector. Moreover we make our images normalized to 64x64 resolution which is clearly a power of 2. There is over completeness in our design because the transforms of the lower levels are used in the higher levels too. But this redundancy is unavoidable in order not to drop information and have well defined inner products. We also achieve a nesting of subspaces from coarser to finer information.

5 System design and case studies

The proposed infrastructure has been implemented using a Java-based system implemented on the JADE platform. The prototype design is presented in "Fig. 4". "Fig. 4" shows a typical example of the application of our system on real-world data. The captured images have size 640x480 pixels with zero compression. They are resized to half their size for presentation purposes. The frame grabber is implemented in C++ and the GUI in TCL/TK 8.5.6. For the agent-based system we selected the well-documented JADE platform and for face detection we used the stock implementation of the OpenCV library. The communication between the GUI and JADE was accomplished with SOAP messaging. The database stores the Haar features and the normalized 64x64 sized images. We used the mingw/msys compiler system and jdk 1.6.11.

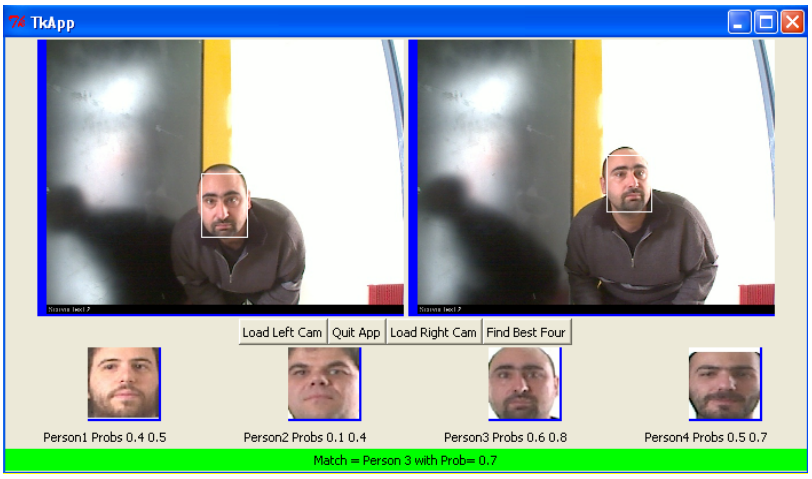


Fig. 4, Prototype system

The functionality is the following: two buttons load images from the two-camera system. The button *Find Best Four* initiates the procedure of face detection and recognition and finds the four best matches in the database by resizing to 64x64, extracting features and using a Euclidean distance. In order to reduce false positives, we select face detected regions that have high matching in the database. In this way, we eliminate spurious face regions that have negligible resemblance to the already stored data due to noise. This is a typical problem with the above mentioned Haar detector. The existing data were extracted manually from the operator after the recording of 18 persons in 4 poses with 100 copies for the same person-pose pair. We also stored some negative results from the entrance without any humans to increase robustness. In a latter version we should try to automate this procedure possibly by selecting another face detection algorithm or exploiting further real biometric information. We used the already trained detector of OpenCV but the cascade was run as a pipeline on different agents.

6 Conclusions

This paper proposes an infrastructure for intelligent environments and biometrics. The surveillance tasks are distributed on multi agents and resources are shared between heterogeneous services. Our intent is to further extend our prototype and experimentation by considering as future goals the integration of new services mediated by other categories of adaptive agents and to perform an evaluation between different surveillance algorithms implemented by the proposed infrastructure.

References

[1] Delac, K., Grgic, M., "A Survey of Biometric Recognition Methods", *46-th International Symposium Electronics in Marine*, ELMAR-2004, ISBN 953-7044-02-5, Zadar, Croatia, 16-18 June 2004, pp. 184-193, (2004)

[2] Ross, A., Jain, A., "Information fusion in biometrics", *Pattern Recognition Letters*, No. 24, Elsevier Science, pp. 2115-2125, (2003)

[3] Pigeon, S., Vandendorpe, L., "Image-based multimodal face authentication", *Signal Processing* Volume:69, issue: 1, August 31, pp. 59-79, (1998)

[4] Leca, R. Groza, V., "Online Personal Identification Agent", *IEEE International Workshop on Measurement Systems for Homeland Security, Contraband Detection and Personal Security*, Orlando, FL, USA, 29-30 March, IMS 2005, (2005)

[5] You, J., Zhang, D., Cao, J., Minyi, G., "Parallel biometrics computing using mobile agents", *Parallel Processing, 2003. Proceedings. 2003 International Conference on*, 6-9 Oct., pp. 305-312, (2003)

[6] Sterritt, R., Garity, G., Hanna, E., O'Hagan, P., "Autonomic Agents for Survivable Security Systems", 1st *IFIP Workshop on Trusted and Autonomic Ubiquitous and Embedded Systems (TAUES 2005)*, at EUC'05, Nagasaki, Japan, 6-9th December, in "LNCS 3823", (2005)

[7] Swiniarski, W. R., "An Application of Rough Sets and Haar Wavelets to Face Recognition", in *Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing,* Springer-Verlag,, pp. 561-568, http://portal.acm.org/citation.cfm?id=646472.692649, (2001)

[8] Belhumeur, N.P., Hespanha, P.J., Kriegman, J.D., "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," IEEE Transactions on *Pattern Analysis and Machine Intelligence*, 19, no. 7, pp. 711-720, (1997)

[9] Rowley A. H., Baluja, S., Kanade, T., "Neural Network-Based Face Detection," *IEEE Transactions On Pattern Analysis and Machine intelligence 20*, pp. 23--38, doi:10.1.1.110.5546, (1998)

[10] Viola, P., Jones, M., "Robust Real-time Object Detection", *International Journal of Computer Vision*, doi:10.1.1.23.2751, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.2751, (2001)

[11] Papageorgiou C,, Poggio T. "A trainable system for object detection", *International Journal of Computer Vision,* vol 38, No. 1, June 2000, pp. 15-33.

Unsupervised Human Members Tracking Based on an Silhouette Detection and Analysis Scheme

Costas Panagiotakis and Anastasios Doulamis

Abstract In this paper, an unsupervised, automatic video human members(human body and parts) tracking algorithm is proposed based on a propabilictic human silhouette detection method and a geometric human silhouette analysis algorithm. First, the human silhouette is estimated using the following scheme : 1) an face detection, and 2) a human body detection based on biometric 2-D templates. Next, the human members are recognized and 18 major points, located on the human body and parts, are detected using a 2-D biometric model. The proposed method is executed automatically in each frame of the video sequence.

1 Introduction

Human motion analysis has many applications in many areas, such as analysis of athletic events, surveillance, content-based image storage and retrieval. The main scientific challenges in human motion analysis are to detect, track and identify people and to recognize the human activity [1]. The detection and tracking algorithms are challenged by occluding and fast/complicated moving objects, as well as illumination changes. A combination of human shape-motion features estimation, silhouette analysis, skin color detection, template matching, 2–D/3–D human modeling, background modeling have been used on human detection and tracking systems. There are model based approaches and systems using Shape-From-Silhouette methods to detect and track the human in 2D [2] or 3D space [3].

Several works have been proposed recently in the literature for detecting video actions and activities [2]. In [4], a Hidden Markov models are used for identifying

Costas Panagiotakis
Department of Computer Science, University of Crete, Heraklion, Greece, e-mail: cpanag@csd.uoc.gr

Anastasios Doulamis
Technical University of Crete, Greece, e-mail: adoulam@cs.ntua.gr

human activities in video streams. In addition, in [5] human activity is identified from a video sequence. The activity is represented as a set of pose and velocity vectors for the major body parts (hands, legs, and torso) and then stored in a set of multidimensional hash tables. On the contrary, in [6], stochastic algorithms are exploited to detect visual activities and interactions in a video sequence.

In this paper, a novel approach is presented for automatic human detection and tracking in video sequences. The proposed architecture consists of two main modules. The first refers to the automatic identification of humans regardless of their motion actions, background complexity and possible background movement. This part is based on biometric feature of human color. The second module aims at automatically extracting points of interest from the human based on 2-D biometric human model. The points refer to salient parts of persons, such as the head, shoulders, etc. In this way, we are able to identify complex human activities, by considering the point motion in time. For example, we are able to detect particular gestures of humans, the direction of their movement, the actions of the head, e.g., nodding, and velocity of human movement, etc.

The rest of the paper is organized as follows. Section 2 presents the proposed silhouette estimation algorithm. Section 3 describes the human silhouette analysis method. Finally, Sections 4, 5 provide experimental results and the discussion.

2 Human Content Identification

Various methods and algorithms have been proposed in the literature over the years for human face detection, ranging from edge map projections to recent techniques using generalized symmetric operators [7]. The eigentemplate approach to the detection of facial features has been proposed in [8], while in [9] the face pixels are localized by modeling the human face with an elliptic shape. In our approach, the two-chrominance components of pixels are used for performing the human face detection task efficiently, while simultaneously exploiting information available in the bit stream of MPEG-coded images. This is due to the fact that the distribution of the two-chrominance components, corresponding to a human face, are located in a very small region of the color space as has been shown in [10]. Thus, blocks of a color image x, whose respective chrominance values are located at this small region, can be considered as face blocks. On the contrary, blocks of chrominance with values located far from this region correspond to non-face blocks.

Let us denote by $q(B_i) = [u(B_i)v(B_i)]^T$ a 2-dimensional vector containing the average chrominance components, $u(B_i)v(B_i)$, for the B_i block. Then, the histogram of the chrominance values, corresponding to the face area, is modeled by a Gaussian probability density function (pdf). Therefore, the probability of a block, say B_j, belonging to the face class, say Ω_f, is given by the following equation

$$P(q(B_j)|\Omega_f) = \frac{e^{-\frac{1}{2}(q(B_i)-\mu_f)^T \cdot S_f^{-1} \cdot (q(B_i)-\mu_f)}}{(2\pi)^{\frac{N}{2}} |S_f|^{\frac{1}{2}}} \quad (1)$$

where μ_f and S_f are the mean vector and variance matrix of the pdf respectively. The parameters of 1 can be estimated based on several training data of face images and using the maximum likelihood algorithm. Equation 1 indicates that an image block B_i belongs to the face area, if the respective probability of its chrominance values, $P(q(B_i)|\Omega_f)$ is high. Instead, blocks with a low probability $P(q(B_i)|\Omega_f)$ are classified as non face blocks. In our case, a confidence interval of 80% has been used to discriminate face and non face blocks. Therefore, a binary mask M is formed, with size $\frac{N_1}{8} x \frac{N_2}{8}$ pixels; a pixel unit with value equal to one indicates a face block, while a zero value indicates a non face one.

This method does not, however, exploit any geometric information about the human face. Thus, it is possible that some non face blocks are classified as face ones. This is, for example, the case of blocks that have similar chrominance properties to ones belonging to face regions, e.g., human hands. For this reason, an iterative technique is applied to the binary mask M in order to localize the segment that corresponds to the face region. First, the morphological erosion operator is applied to image M, using a small rectangular structuring element; then, the number of non connected objects in the filtered mask is computed. In case that the number of objects is greater than one, a new morphological filtering is applied using, however, a greater structuring element. This procedure iterates until the number of objects gets equal to one. Then, the segment of M, which overlaps to the segment of the final filtered mask, is considered as a face region. In this case, a binary mask, say M_f, is formed, in which pixels with value equal to one correspond to the face segment, while zero values indicate the other areas.

2.1 Human Body Detection

Human body detection is next performed, exploiting information provided by the previous face detection module. In particular, the human body is localized using a probabilistic model, the parameters of which are estimated according to the center, height and width of the face region, denoted as $c_f = [c_x\ c_y]^T$, d_f and h_f respectively. Let us also denote by $r(B_i) = [r_x(B_i)\ r_y(B_i)]^T$ the distance between the ith block, B_i, and the origin, with $r_x(B_i)$ and $r_y(B_i)$ the respective x and y coordinates.

Since humans are usually located in standing position, a square rectangular is adopted in our case for modeling human body. We assume independence from the x and y location. Thus, block B_i belongs to the human body class, say Ω_b, if

$$P(r(B_i)|\Omega_b) = A \cdot [(1 - \frac{r_x(B_i) - \mu_x}{w_x}), (1 - \frac{r_y(B_i) - \mu_y}{w_y})] \quad (2)$$

where μ_x, μ_y express the parameters of the human body location model; these are calculated based on information derived from the face detection task, taking into account the relationship between human face and body. In our simulations, the

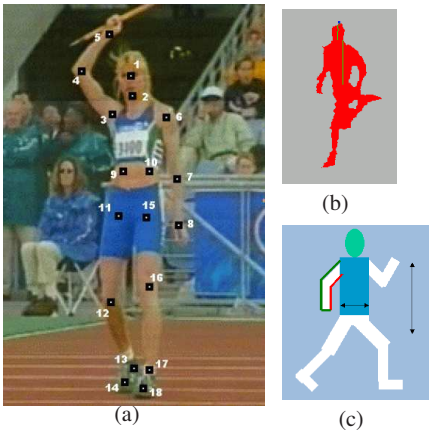


Fig. 1 **(a)** The 18 major human points. **(b)** An example of successful execution of the end of head (blue point) localization method. **(c)** The left arm can be distinguished from the main body because the proportion of red boundary pixels is high.

parameters in 2 are estimated with respect to the face region as follows $\mu_x = c_x$, $\mu_y = c_y + \frac{h_f}{2}$, $\sigma_x = d_f$, $\sigma_y = \frac{h_f}{2}$.

Similarly, A,w_x and w_y are appropriate parameters that control the decay of the probability.

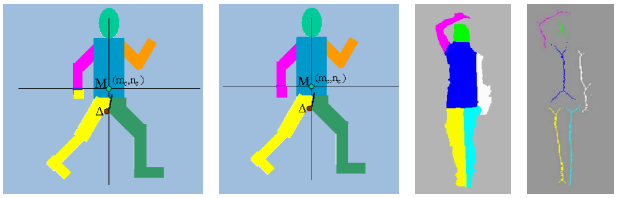
The human face and body detection modules provide an initial estimation of the foreground object. In particular, all blocks that have been classified either to the face or body classes are included in the initial estimate of the foreground object. Similarly, a background set, is created containing blocks of the image which are classified with high confidence to the background class.

3 Human Silhouette Analysis

3.1 Human Members Recognition

In this section we examine the human body members recognition method. The human body is divided into the following six members: head, main body, left leg, right leg, left arm and right arm (Figure 2(c)). The human silhouette pixels will be classified to one of the above members. The human parts are detected based on their geometric/biometric information. The member recognition algorithm is sequential. The more "visible" members are computed first in order to decrease the search space of others.

First, the major human body axis is determined using central moments. The silhouette is rotated according to the major axis. The rotation center is the mass center



(a) (b) (c) (d)

Fig. 2 **(a)** The initial approximation. **(b)** The final segmentation result. **(c)** The human members estimation. **(d)** The human members' skeletons.

of the human. Now, the human is located vertically. Next, we estimate the end point of the head (X_d, Y_d) using the following iterative method (Figure 1(b)). The (X_d, Y_d) point is initialized as the human body mass center (m_c, n_c) and it is changed dynamically by getting the mean of human points that belongs to the current line. So, the above method can be characterized as dynamic mean method.

Dynamic Mean Method

$X_d = m_c, \quad Y_d = n_c$
repeat {
$\quad X_d = X_d - 1, \quad Y_d = E_{(X_d, y) \in Man}(y)$
} until $(X_d - 1, Y_d) \in Background$

The head region will be placed in a rectangle defined by the point (X_d, Y_d) as the end of the head. The maximum height and width of the head are proportional to the human height. The down limits of the head region, are determined by the first local minimum of the left and right horizontal silhouette projections.

The main body region is computed using an iterative algorithm similar to the end point of head estimation method. The maximum height and width of the main body are proportional to the human height. Using the human boundary, it can be determined if the arms can be distinguished from the main body (visible arm) and if the legs are distinguished. The rule is the following: if the proportion of the boundary pixels whose the closest background pixel is on the right, computed in an area where is located the left arm, exceeds a threshold, then the left arm can be distinguished from the main body. The above value is the ratio between the red pixels and the red plus green pixels of Figure 1(c). This knowledge helps in definition of the main body limits.

The legs and arms regions estimation can be done in the same time. An initial approximation (Figure 2(a)) is computed using the mass center of the human and the border point D that discriminates the left and right leg. It is possible that a regions than six, that is the number of the human members, will be created. For these cases, we determine first which are the fault regions using a criterion based on the surface and the mass center of the regions. Next, we join the faults regions to the regions which have the most common boundary points with the faults regions. In the Figure 2(b), the false left leg region is correctly classified to left arm member.

3.2 Major Human Points Estimation

In the second stage, the 18 major human points are estimated using the human members segmentation. This method is based on a geometric/biometric 2-D model for each major human point. The method uses the skeleton of each human member (Figure 2(d)), as the joint points belong in the skeletons. The skeleton is defined as the set of points whose distance from the nearest boundary is locally maximum. The 18 major human points are computed sequentially. The easier defined points are computed first in order to decrease the search space of others.

First, the center of the head is computed as the mass center of the head region. The neck point is defined as the mean of the boundary points between head and main body region. The two shoulders points are computed by minimizing an appropriate function F. The function domain is an isosceles triangle whose vertex is the neck point and its base vertices are the two shoulder points. The function is minimized when the triangle base is maximized and the triangle height is minimized at the same time.

The 18 major points formulations are defined in Figure 1(a) under the proposed biometric 2-D model. The points (9), (10) of the main body are computed using the main body height and width. The points (11), (15) of the main body are defined by the mean of boundary between main body region and left or right leg region respectively.

Concerning the legs' points, the ankle point A is computed first. We compute the farthest point B of skeleton points from point (9) using one line segment that should belongs to silhouette. The ankle point is defined as the farthest point, of the not visited skeleton points from B using one line segment. The knee point K is estimated by minimizing the function $G(X)$ which is defined by the following equation. The constant 0.2 of equation 3 has been estimated using our experimental dataset. Let F be the point (9) of the main body. Let the function $d(X,AF)$ be the minimum distance of point X from the line segment AF.

$$G(X) = (|XF| - |XA|)^2 - 0.2 \cdot d^2(X,AF) \tag{3}$$

If the point X is located close to the middle of AF and close to the knee angle at the same time, then the proposed function $G(.)$ will be minimized. Finally, the end of leg point E is computed using the knee and ankle points. The E point is defined as the skeleton point close to ankle point, whose distance from the knee point is maximum. In each arm, we have to compute two points, the elbow point and the end of arm point which are estimated similarly with the knee and ankle point, respectively.

4 Results

The proposed algorithm have been tested in several sequences. The silhouette estimation method (first stage) gives in most of the frames accurate results. Concerning

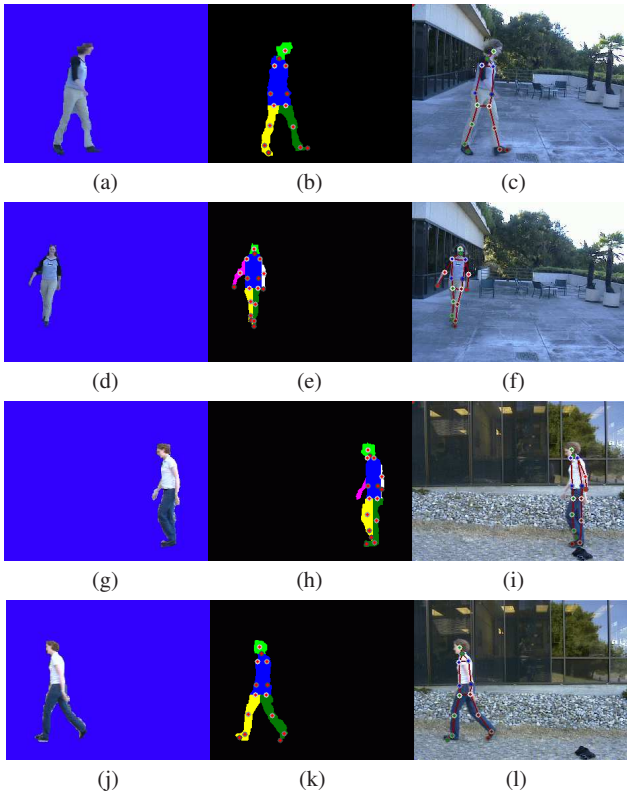


Fig. 3 Results of the three stages. The estimated silhouettes (first column), the human members recognition (second column) and the 18 major human points computation (third column).

the next two stages, the human body members are successfully recognized in the well estimated silhouettes with high accuracy. The mean error in head, main body major points estimation is about 2% of the total human height, while the arms - legs major points are estimated with less than 200% of the previous error. The above errors are related with the quality of the estimated silhouette, when the accuracy of the estimated silhouette is high the mean error in the estimation of each major human point is less than 1% of the total human height. In Figure 3, some results of the three stages are shown. The complexity of the total algorithm is $O(N)$[1].

[1] N = # pixels of the human. If we did not use skeletons we would have a complexity of $O(N\sqrt{N})$.

5 Conclusion

In this paper, an unsupervised, automatic human members and 18 major human points tracking algorithm is proposed using an adaptable - extended face detection based method and a geometric human silhouette analysis algorithm. The adaptive behavior of the first stage is very important in such dynamically changing environments, where object properties frequently vary through time. The silhouette analysis algorithm is color independent and it detects the major human points without tracking them based on 2-D geometric human model.

In order to decrease the computation cost, the proposed method locates the major points sequentially, where the location of one feature influences the location of the rest. Sometimes, such methods can produce totally erroneous results in the case that a failure occurs in the starting stages. We decrease this probability by starting from the most "visible" parts and well defined points. Moreover, if in some frame the algorithm fails (in some points/parts), the system will not lose his stability, since in the next frame human detection (not tracking) will be performed.

An extension of the proposed methodology may include the estimation of static (chromatic-biometric) features of the human members and the human-activity recognition (walking, sitting, running, etc). Security system and statistics analysis of human motion systems could be based on our method.

References

1. J. Aggarwal and S. Park, "Human motion: Modeling and recognition of actions and interactions," in *3DPVT04*, 2004, pp. 640–647.
2. C. Panagiotakis, E. Ramasso, G. Tziritas, M. Rombaut, and D. Pellerin, "Shape-motion based athlete tracking for multilevel action recognition," in *Proc. of AMDO 2006*, 2006, pp. 385–394.
3. K.M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time: Part ii: Applications to human modeling and markerless motion tracking," *Int. Journal of Computer Vision*, vol. 63, no. 3, pp. 225–245, 2005.
4. M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 844–851, 2000.
5. J. Ben-Arie, Zhiqian Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Trans. on PAMI*, vol. 24, no. 8, pp. 1091–1104, 2002.
6. Y.A. Ivanov and A.F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 852–872, 2000.
7. D. Reisfeld, H. Wolfson, and Y. Yeshurum, "Detection of interest points using symmetry," in *Proc. of Inter. Conf. Coputer Vision*, 1990.
8. B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *Pattern Analysis Machine Intelligent*, vol. 19, no. 3, pp. 696–710, 1996.
9. A. Eleftheriadis and A. Jacquin, "Automatic face location for model-assisted rate control in h.261 compatible coding of video," *Signal Processing Image Communication*, vol. 7, pp. 435–455, 1995.
10. H. Wang and Shih-Fu Chang, "A highly efficient system for automatic face region detection in mpeg video sequences," *IEEE Trans. on Circuits and Syst. for Video Technol,special issue on Multimedia Systems and Technologies*, 1997.

Facial Biometric Templates and Aging: Problems and Challenges for Artificial Intelligence

Andreas Lanitis

Department of Multimedia and Graphic Arts, Cyprus University of Technology

P.O Box 50329, Lemesos, 3066, Cyprus

andreas.lanitis@cut.ac.cy

Abstract. The performance of face recognition and/or authentication systems is greatly affected by within-person variations encountered in human faces. Within person facial variations distort the appearance of faces leading to inconsistencies between facial features stored in templates and features derived from a face, of the corresponding subject, captured at a different instance. For this reason a considerable amount of effort has been devoted to the development of methods for eliminating within person-variations during face recognition/authentication. Among all types of within-person variations encountered, aging-related variations display unique characteristics that make the process of dealing with this type of variation a challenging task. In this paper we describe experiments that enable the quantification of the effects of aging on the performance of face recognition systems. We also review typical approaches that aim to eliminate the effects of aging in face recognition and outline future research directions for this area.

1 Introduction

Researchers working in this field of face-based authentication aim to define biometric templates containing discriminatory features that are least affected by within-person types of variations in order to enable accurate identity verification [20]. Examples of within-person facial variations include variations due to different image acquisition conditions, different expressions, face orientation, occluding structures and aging. Researchers who compared the impact of within-person variations in different biometric templates [5, 9] conclude that facial templates display reduced permanence when compared to other widely used biometric modalities (i.e iris and fingerprints) hence it is highly important to develop methods for eliminating within-person variations in face recognition/authentication algorithms. Among all types of appearance variations, aging related facial variation displays unique

characteristics that require customized attempts for dealing with the problem. One solution to the problem of aging is the frequent update of facial templates through a data acquisition process [19]. However, template updates can be performed only if all subjects whose faces are stored in a database are regularly available and willing to provide up to date face images.

During the recent years an increased research activity in the area of facial aging simulation is recorded [4, 6, 7, 8, 10, 12, 14, 15, 16, 17]. This interest is attributed to the potential of using age modeling techniques for a number of different applications that includes the prediction of the current appearance of missing persons [6], age estimation [4, 7], age-specific human computer interaction and the development of age invariant face authentication systems [10]. Research in facial aging is backed up by the existence of two publicly available face datasets [8, 15] dedicated for use in conjunction with studies related to face aging.

This paper aims to formulate the challenges involved in developing age-invariant face templates. Along these lines we briefly describe the physiological mechanisms that cause age-related variations on human faces and discuss the difficulties involved in dealing with aging variation. We also present results of an experimental evaluation that aims to quantify the effect of aging on face recognition. Challenges associated with the problem of aging are presented and possible ways for dealing with this problem are suggested. The discussion of this issue takes the form of a short review of the topic where different approaches to the problem and possible directions for future research activities are outlined.

2 Face Aging Process

The appearance of a human face is affected considerably by the aging process. Examples of aging effects on faces are shown in figure 1. Facial aging is mainly attributed to bone movement and growth, and skin related deformations. Aging related appearance variation due to bone growth usually takes place during childhood and puberty whereas skin related effects mainly appear in older subjects. Skin related effects are associated with the introduction of wrinkles caused by reduced skin elasticity and reduction of muscle strength [3, 12].

Figure 1: Face images displaying aging variation. Each row shows images of the same individual (Images from the FG-NET Aging Database [8]).

Dealing with aging variation is a challenging task because the aging process has the following unique characteristics:

Control: Unlike other types of variation (i.e face orientation) the effects of aging cannot be controlled and/or reversed. This characteristic of aging variation implies that (1) it is not possible to rely on the cooperation of a subject for eliminating aging variation during face image capture and (2) the process of collecting training data suitable for studying the effects of aging requires long time intervals.

Diversity of Aging Variation: Both the rate of aging and the type of age-related effects differ for different individuals. Typically diverse aging effects are encountered in subjects of different ethnic origins and different genders. External factors may also lead to diversities in the aging pattern adopted by different individuals. Such factors include health conditions, lifestyle, psychological conditions, climatic related factors and deliberate attempts to intervene with the aging process through the use of anti-aging products [3] or cosmetic surgeries. As a result common aging patterns cannot be applied successfully to all subjects.

3 Effects of Aging on Face Recognition

In this section we discuss the results of an experimental evaluation that aims to quantify the effects of aging on face recognition. As part of the experiment we use images from the FG-NET Aging Database [8] for training and testing face recognition systems in cases that we deal with significant aging variation. For the needs of our experiments we have used images from the FG-NET Aging database that contains 1002 images from 82 subjects. The database was divided in the following groups:

Group A and Group B: Each group contains half of the images of each of the 82 subjects. The separation of images was done in such way so that both groups contain similar distribution of ages for each subject. On average for each subject the

age difference between samples in Groups A and B is 1.6 years, the minimum difference is 0 years and the maximum difference is 5 years.

Group C and Group D: Each group contains half of the images of each of the 82 subjects. However in this case Groups C and D contain the samples of each subject corresponding to younger and older ages respectively. The separation between younger and older faces of each subject is done based on the mean age among all samples belonging to the subject in question. On average for each subject the age difference between samples in Groups C and D is 15 years, the minimum difference is 6 years and the maximum difference is 34 years.

Figure 2 shows typical samples belonging to groups A,B,C and C for a one subject from the dataset.

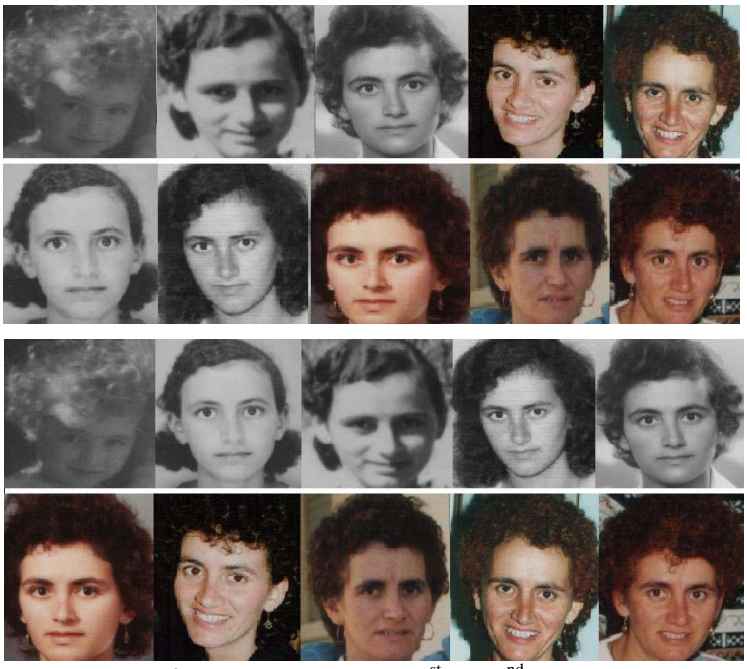


Figure 2: Samples from Groups A and B (1st and 2nd row) and Groups C and D (3rd and 4th row). Age distributions of samples in groups A and B are similar whereas age distributions of samples in groups C and D are distinctively different.

Face templates used in the experiments correspond to different facial parts so that it is feasible to compare the effects of aging for different parts. In particular templates derived from the overall internal facial region, upper face and lower face are used in our experiments (see figure 3). For the needs of the experiments facial templates are constructed based on a low dimensional Active Appearance Model-based representation [2] of the faces in the training and test sets. As part of the ex-

periments two classification methods are used – the first method is based on a shortest Mahalanobis distance classifier and second is based on a Support Vector Machine Classifier. In both cases classifiers are trained using images from the train sets (sets A and C) and tested using sets B and D respectively. In our results (see table 1) we quote the reduction of the correct recognition rate when dealing with extreme aging variation (test set D) when compared to the case that we deal with reduced aging variation (test set B).

Figure 3: Face templates used in our experiments

	Reduction in Recognition Performance (Percentage units)	
	Shorter Distance Classifier	Support Vector Machine
Upper Face	13	14
Lower Face	8	8
Internal Face	15	16

Table 1: Reduction of correct recognition when dealing with extreme aging variation, compared to the case that we deal with reduced aging variation

According to the results the following conclusions can be derived.

- On average the performance of face recognition drops by about 12% when dealing with faces with different age distributions than the ones in the training set. Therefore it is of utmost importance to deal with aging variation.
- The effects of aging have greater impact when using face templates that include the upper face region. When using templates based on the lower part of the face aging effects cause smaller decrease in the recognition performance. However, the regions affected more intensively by aging are at the same time the most discriminatory regions.

The results of this preliminary experiment suggest that researchers, who work in the area of face recognition/verification, should run age invariance tests in order to assess the robustness of their approaches to aging variation. As part of this exercise along with the standard performance evaluation metrics that are usually quoted, the ability of an algorithm to deal with aging variation also needs to be specified. The framework used in the experiment described above could form the basis of developing standardized age-invariance test methodologies.

4 Discussion

We presented an overview where we outlined how aging affects human faces and we presented experimental results that demonstrate that face aging can affect dramatically the performance of face recognition systems. In order to deal with this problem we need to use techniques capable of simulating/eliminating aging effects on biometric facial templates. Aging simulation can be performed based on data driven approaches or methods based on modeling physiological mechanisms that cause the process of aging.

Data Driven Approaches: Data driven approaches rely on the analysis of aging datasets so that aging patterns are defined and used as the basis for simulating aging effects on previously unseen face templates. This method requires suitable datasets that contain face images of the same individual at different ages. Ideally such samples should be normalized so that only aging variation is observed among the samples. Data driven age modeling approaches described in the literature include the use of machine learning algorithms for defining aging patterns [4, 6, 15, 17] which can later be applied to novel faces in an attempt to simulate age effects. A different approach to the problem involves the use of statistical modeling for establishing the distributions of faces belonging to different age groups enabling in that way the application of aging effects by forcing a face to move closer to a target age distribution [8, 16].
From a different perspective data driven approaches can be used for the definition of age invariant facial representations. For example Ling et al [10] suggest that a face representation based on differences in gradient orientation displays increased tolerance to aging variation. These findings were verified in a scenario involving face verification experiments in the cases that the age between faces in a pair differ by up to 10 years.

Modeling Physiological Aging Mechanisms: This approach involves the use of complex mathematical models of physiological mechanisms that can be used for simulating aging effects on human faces and/or face templates. The process of defining such models requires deep understanding of the physical nature of the process of aging so that this process can be artificially reproduced. Thompson [18] was among the first researchers who proposed the use of mathematical models for modeling the growth of biological organisms. Based on this proposition several researchers attempted to derive functions that can be used for simulating the process of aging on 2D or 3D face outlines [11, 13]. Ideally physiological models of aging should take into account different aspects of personal characteristics in order to be able to produce aging simulations customized for different individuals.

As an alternative to the two categories of approaches mentioned above, it is possible to adopt combined data-driven and model-based approaches. In this respect the parameters of physiological models of aging are optimized through a machine-learning process that operates on suitable aging datasets [1, 14].

5 Conclusions

The aging process causes significant alterations on faces affecting in that way the long term performance of face authentication systems. The solution to this problem is to develop smart systems that will be able to modify biometric facial templates in an attempt to simulate aging effects ensuring in that way that face templates are always consistent with the current facial appearance of a subject. This task is extremely difficult because aging in combination with external factors that influence the process of aging, cause compounded effects that are difficult to predict and model. For this reason modeling aging variation requires state of the art Artificial Intelligence techniques that will be able to deal with this highly demanding problem. Issues that need to be addressed in order to address the problem include the establishment of accurate person specific aging patterns that take into account the possible effect of external factors and the definition of facial representations that include discriminatory but at the same time age invariant features.

References

[1] L.Boissieux, G. Kiss, N. Magnenat-Thalmann and P. Kalra. "Simulation of Skin Aging and Wrinkles with Cosmetics Insight". Computer Animation and Simulation, pp.15-27, 2000.

[2] T.F.Cootes, G.J. Edwards and C.J. Taylor. "Active Appearance Models". IEEE Transactions of Pattern Analysis and Machine Intelligence, 23, pp 681-685, 2001.

[3] N. Dayan (ed), "Skin Aging Handbook: Market Perspectives, Pharmacology, Formulation, and Evaluation Techniques", Andrew William Press, 2008.

[4] X. Geng, ZH. Zhou and K. Smith-Miles. "Automatic Age Estimation Based on Facial Aging Patterns". IEEE Transactions of Pattern Analysis and Machine Intelligence, 29(12), 2234-2240, 2007.

[5] A. Jain A, A. Ross and S. Prabhakar. "An Introduction to Biometric Recognition". IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics', 14(1), 2004.

[6] A. Lanitis, C.J. Taylor and T.F. Cootes, "Toward Automatic Simulation Of Aging Effects on Face Images". IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol 24, no 4, pp 442-455, 2002.

[7] A. Lanitis, C. Draganova and C. Christodoulou. "Comparing Different Classifiers for Automatic Age Estimation". IEEE Transactions On Systems Man and Cybernetics, Part B, 34(1), pp 621-629, 2004.

[8] A. Lanitis. "Comparative Evaluation of Automatic Age Progression Methodologies". EURASIP Journal on Advances in Signal Processing, Article ID 239480, 2008.

[9] S. Latifi and N. Solayappan. "A Survey of Unimodal Biometric Methods, Security and Management", pp 57-63, 2006.

[10] H. Ling, S. Soatto, N. Ramanathan, D.W. Jacobs. "A Study of Face Recognition as People Age". IEEE 11th International Conference on Computer Vision, pp 1-8, 2007.
[11] L.S. Mark and J.T.Todd. "The Perception of Growth in Three Dimensions". Perception and Psychophysics Vol 33, No. 2, pp 193-196, 1983.
[12] E.Patterson, S. Sethuram, M. Albert, K. Ricanek, M. King. "Aspects of Age Variation in Facial Morphology Affecting Biometrics". IEEE International Conference on Biometrics: Theory, Applications, and Systems, pp 1-6, 2007.
[13] J.B. Pittenger and R.E Shaw. "Aging Faces as Viscal-Elastic Events: Implications for a Theory of Nonrigid Shape Perception". Journal of Experimental Psychology: Human Perception and Performance 1(4), pp 374-382, 1975.
[14] N. Ramanathan and R. Chellappa. "Modeling Age Progression in Young Faces". Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2006.
[15] K. Ricanek and T. Tesafaye. "MORPH A Longitudinal Image Database of Normal Adult Age-Progression". Procs. of the 7th IEEE International Conference on Automatic Face and Gesture Recognition, 2006.
[16] C.M. Scandrett (née Hill), C.J. Solomon and S.J. Gibson. "A person-specific, rigorous aging model of the human face". Pattern Recognition Letters, Vol 27, no 15, pp 1776-1787, 2006.
[17] J. Suo, F. Min, S. Zhu, S. Shan, and X. Chen. "A Multi-Resolution Dynamic Model for Face Aging Simulation". IEEE Conference on Computer Vision and Pattern Recognition, pp 1–8, 2007.
[18] D.W. Thompson, "On Growth and Form", Cambridge University Press, 1961.
[19] U. Uludag, A. Ross and A.Jain. "Biometric template selection and update: a case study in Fingerprints", Pattern Recognition 31(7), pp 1533-1542, 2004.
[20] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face recognition: A literature survey", ACM Computing Surveys, pp. 399-458, 2003.

POLYBIO Multibiometrics Database: Contents, description and interfacing platform

Anixi Antonakoudi[1], Anastasis Kounoudes[2] and Zenonas Theodosiou[3]

[1] Philips College, 4-6 Lamias str. 2001, P.O. Box 28008, 2090 Nicosia, Cyprus

[2]SignalGeneriX Ltd, Arch. Leontiou A', Maximos Court B', P.O. Box 51341, 3504 Limassol, Cyprus

[3] Cyprus University Technology, Arch. Kyprianos Kyprianos, P.O. Box 50329, 3603 Limassol, Cyprus,

Abstract. Biometrics is an automated authentication mechanism that allows the identification or verification of individual based on unique physiological and behavioural characteristics. The combination of two or more biometric technologies in one application, better known as a multimodal biometric system, can provide enhanced security. Apart from the sound choice of fusion methodologies for the combination of single modality biometrics, the success of such multimodal biometric systems significantly relies on the availability of biometric databases, through which the validation of these systems is made possible. This paper presents a new multimodal database, acquired in the framework of the POLYBIO project funded by the Cyprus Research Promotion Foundation (CRPF). The database consists of fingerprint images captured via an optical sensor, frontal and side views of still and video face images as well as the outside surface of the human palm from two web-camera sensors, and a series of voice utterances recorded with the use of a distant array microphone. The POLYBIO database includes real multimodal and multi-biometric data from 45 individuals acquired in just a single session. In this contribution, the novel platform for data acquisition and combination - through an integrated device - of the four aforementioned single biometric modalities is described and the protocols used for this purpose as well as the contents of the database and its statistics are presented.

1 Introduction

In recent years, there has been a growth in the research of biometric systems, mainly due to the increasing pressure exerted on many Western countries to increase their counter terrorism measures and legislation. Such systems can alleviate problems that plague traditional verification methods such as passwords or identification cards [1] and can be widely used for either access to physical locations such as airports and commercial buildings or for accessing remotely sensitive in-

formation via the World Wide Web. Furthermore, the use and combination of more that one modality can offer enhanced security and can exploit to a higher degree the uniqueness, universality, permanence – to a certain extend - and collectability of each of the single combined biometrics.
Although many biometric databases of single biometric traits such as fingerprint, face and voice have been developed that enabled the growth of unimodal biometric systems [2], the progress on multimodal systems is still prohibited due to the lack of reliable multi-biometric databases. The few publicly available multimodal databases consist of only matching scores produced by several biometric systems operating on different modalities [3]. Therefore, although these databases encourage the research on multimodal fusion, they do not allow further research on different types of fusion other than the scores matching of different systems. Moreover, the absence of more than two or three important traits of the same individual in a single database is another limitation of the existing multimodal biometric databases.

The creation of multimodal databases implies a certain degree of difficulty and challenges in the following manner: the design of an integrated platform for multimodal biometric data acquisition is a complex multidisciplinary approach since it combines many different methodologies for the extraction of biometric traits to be performed under a unified framework. Moreover, the procedure of data collection is highly resource- and time- consuming as it requires a significant number of test subjects that need to cooperate for the collection of their biometric traits, a process that requires a long period of time. Last but not least, the legal issues concerning the collection of biometric data are to be taken into serious consideration as this subject is highly controversial [2]. However, due to the integrated efforts of all participants involved in the POLYBIO project, most of the aforementioned difficulties have been overcome. The presented multimodal database includes fingerprint, palm, voice and still and video face images from a significant number of individuals that have been collected through a novel integrated platform. The reasons behind the selection of these specific biometric traits in the current contribution was the creation of a multimodal , multi-biometric system that offers reliable results that can be interpreted and applied in real time without increasing the user annoyance (e.g. iris scan) [4]. Increased user annoyance has been observed over the use of fingerprints mainly due to its association with criminal prosecution but this limitation has been overcome due to the increasing and popular use of systems based on fingerprint recognition for granting access to personal computer and laptops. In the sections that follow, the methodology for data collection as well as the platform for data acquisition are described and the contents of the resulting database and their statistics are presented.

2 Multibiometric Data Acquisition

The multibiometric data were collected through the use of an integrated platform that was created in the framework of project POLYBIO [5]. The scenario in the acquisition process was an office room where the acquisition hardware and software could be operated by a system supervisor, guiding the steps of the test subjects through the data collection procedure. Environmental conditions such as lighting or background noise were not controlled so as to simulate a realistic situation. The data acquisition system is depicted in Fig1 and is composed of two main components: (a) the multimodal biometric sensor hardware and (b) the data acquisition software. The hardware part consists of four separate sub-systems namely an array microphone for the recording of speech (1), a front –facing USB web camera for the capturing of still and video face images (2), a USB optical fingerprint sensor (3) and a down-facing USB web-camera accompanied by two lighting units and a black board panel with six positioning pins for palm image acquisition (4).

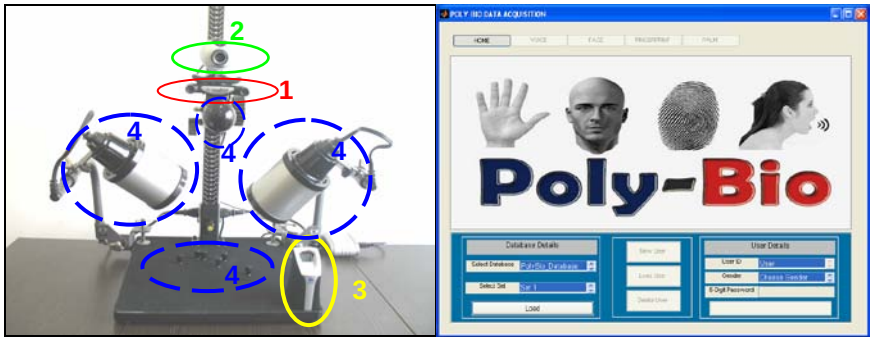
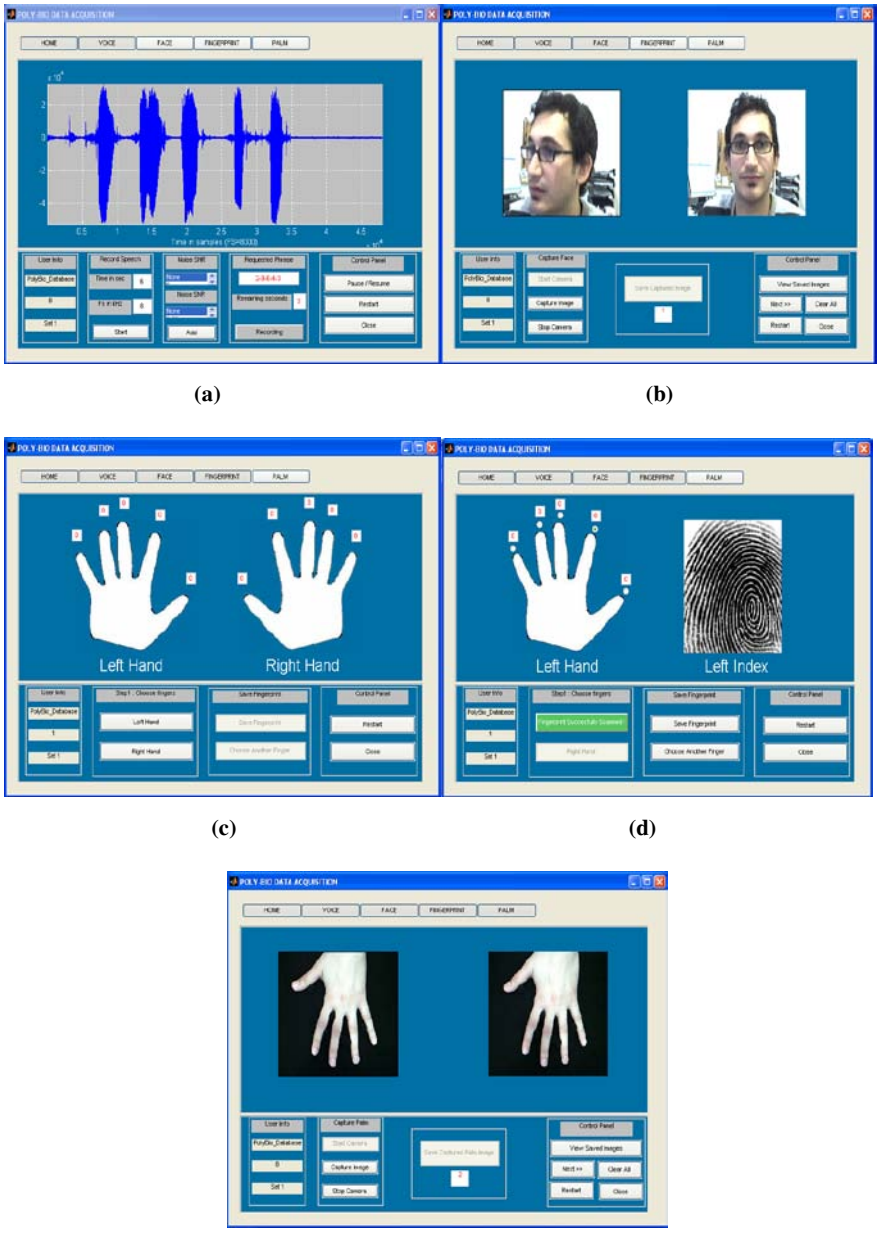


Fig. 1. (a) Hardware Component of Multibiometric Data Acquisition System. (b) Data Acquisition Software

The software part provides a user-friendly Graphical User Interface that allows the registration of every new user by choosing the person's gender and by assigning him or her with a five digit password. Furthermore, it can re-load any existing user for further capturing of the individual's biometric traits. Once a specific user has been selected or registered, the software provides four different interactive panels depicted in Fig. 2. (a) – (e) for the acquisition of the subject's biometric data. For each subject, the information collected for each of the modalities considered are given below through the following description:

Voice – Before the initiation of the recording procedure, the system provides the user with a choice on the sampling frequency of the recorded sample and its duration. The system prompts the test subject to pronounce a sequence of five digits in English.



(a) **(b)**

(c) **(d)**

(e)

Fig. 2. (a) Voice panel, (b) face panel, (c)-(d) fingerprint panels, (e) palm panel

This process is automated and the sequence of numbers appears in the white panel as shown in Fig. 2.(a). A clock counting backwards notifies the user when the recording begins. Once the sequence has been recorded, the speech wave appears on the axes shown at the top of Fig. 2.(a). The supervisor, sitting next to the test subject checks the recorded sequence through the use of a set of headsets connected to the array microphone to verify that no significant level of background noise has corrupted the recording. In case a sequence has to be repeated, the system provides a manual recording facility. This procedure is repeated twenty two times. Therefore, a total of twenty two five-digit sequences are collected for each individual in the database, one of which is his/her unique password that was used by the system for the user registration procedure and ten random sequences repeated on two times (Fig. 1.(b)).

Face –The process of collecting face images is completed with the aid of the interactive panel shown in Fig. 2.(b). Once the front facing camera of the system has been activated, a live streaming of the desired data appears on the left window. Once the supervisor finds the appropriate pose of the test subject, a snapshot is taken and is shown on the right hand window of Fig. 2.(b). If the still image is satisfactory, it is saved in the database. This procedure is repeated four times and a total of **four still images** are being collected for each individual. The system also captures a **live video** with duration of 20 seconds of each person so as to acquire more face angles, since the user is prompted by the system supervisor to turn his/her head left and right.

Fingerprint – Fig. 2.(c) and (d) depict the interfaces for selecting and capturing fingerprint images from the USB optical sensor. The panel shown in Fig. 2.(c) aids the individual to select the hand to be recorded (left or right). Once the hand has been selected, the fingerprint collection process is initiated. The supervisor chooses the finger to be scanned, thumb, index or middle and makes the appropriate selection in Fig. 2.(d). The user then prompts the subject to place the selected finger on the USB optical device and initiates the capturing of the fingerprint image. Once the captured image appears on the panel and is deemed satisfactory, the image is saved in the database. This process is repeated for four times for each finger and for both hands. Hence, the collected data are four images of the thumb, four of the index, four of the middle, four of the ring and four of the little finger for the left and right hand respectively, i.e. a total of forty scanned images for each individual.

Palm – The procedure of the palm image collection is similar to that of the face image collection and is completed with the aid of the panel shown in Fig. 2(e). The user is prompted by the supervisor to place his/her left hand facing downwards on the black board panel with the six positioning pins. The facing downwards camera is activated and the palm images are taken, which if they are considered by the system supervisor of good quality, they are stored in the database. A total of four palm images are collected for each individual.

Table 1 presents more information about the images acquired for face, palm and fingerprint modality. Other personal data that was acquired for all participants and stored independently from the database was gender, name, age and nationality. Moreover, donors wearing glasses had to remove them for half of the images of the face capture, so that facial samples of them without glasses -in case they wore contact lenses – would exist.

Table 1. Image Information

Modality	Height	Width	Format	Type
Face	240	320	Jpg	RGB
Palm	240	320	Jpg	RGB
Fingerprint	292	248	Jpg	Grayscale

Since biometric data is considered "personal data" defined as such by the corresponding regulation for the European requirements on the protection of individuals with regard to the processing and movement of personal data [6], all participants have willingly signed a consent agreement. This agreement ensured that these sensitive pieces of information should only be used and processed for the purposes of the current research project and that the movement of this material will be confined within the members of the participants of this project for further processing. Furthermore, all research outcomes that result from this data will be presented in an anonymous way.

3 Description of POLYBIO database

The panels shown in Fig. 2(b), (d) and (e) contain typical images for the face, fingerprint, and palm biometric traits respectively. The recorded speech utterances are shown in Fig. 2(a) only in terms of the speech waveform but there are saved in the database as a wav files.

Although a considerable effort was made to create a robust database through the use of the specialized hardware and software described in Section 2, and although a human supervisor was present at all times, the possibility of software and/or human errors was not completely eliminated. After initial acquisition of the biometric traits, all samples were verified by a human expert. The ones that were non-compliant with the acquisition protocol were discarded according to the following set of strict rules:

1 - All facial and palm samples should be four for each registered user. If any of these samples are missing the particular subject is rejected.

2 - The fingerprint images that should be acquired for each user are four scanned fingerprint images for all fingers of both hands (left and right). If any of these images is missing, the subject is rejected.

3 – The total number of speech utterances should be twenty-two five digit numbers pronounced in the English language. The above number results from the repetition two series of five-digit number utterances pronounced twice and the single utterance of five digits (also pronounced twice) that constitutes the individual password that was initially assigned to each participant during his /her enrollment to the database system. Errors in the pronunciation of the speech sequences can be corrected by the system supervisor during the acquisition procedure by manual re-entrance of the missing or incorrect sequence. Any future discovery of a missing or erroneous speech sample results in the discarding of the corresponding donor.

The following figure (Fig.3.) show the statistics of the biometric data and population acquired.

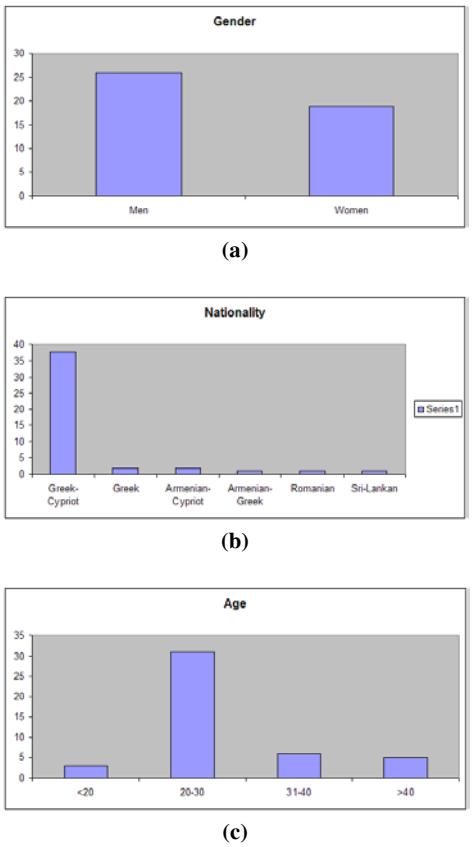


Fig. 3. Statistics of the biometric data, (a) gender, (b) nationality, (c) ages

4 Conclusions and Further Work

This paper presents a new multimodal database, containing biometric features of forty five individuals. The database consists of fingerprint images captured via an optical sensor, frontal and side views of still and video face images as well as the outside surface of the human palm from two web-camera sensors, and a series of voice utterances recorded with the use of a distant array microphone, resulting to a total of 2160 images and 990 pronounced speech utterances. The samples were collected with the use of a novel platform for biometric data acquisition and specific protocols were followed for the sound selection of all samples that were to be finally stored into the database. A variable number of test subjects were selected with a wide range of different characteristics in terms of age and nationality and pronunciation while the gender was kept almost in equal amounts. However, an increase in the number of test subjects with different characteristics is considered of high importance from the members of the project consortium, as it will add value to the diversity of the current database, making it an indispensible tool in applications where accurate user identification and recognition is required.

5 Acknowledgements

This work was undertaken in the framework of the POLYBIO (Multibiometric Security System) project funded by the Cyprus Research Promotion Foundation (CRPF) under the contract PLHRO /0506/04.

References

[1] Ross, A., Jain, A. K.: Identification Information fusion in biometrics.
Pattern Recognition Letters 24, 2115–2125 (2003)
[2] J. Wayman, A. Jain, D. Maltoni, D. Maio (Eds.), Biometric Systems:
Technology, Design and Performance Evaluation, Springer, 2005.
[3] N. Poh, S. Bengio, Database, protocols and tools for evaluating score-level fusion
algorithms in biometric authentication, Pattern Recognition 39 (2006) 223–233.
[4] L. Ma, T. Tan, Y. Wang, Y. Wang, D. Zhang, "Personal Identification based on Iris Texture Analysis", IEEE Trans on Pattern Analysis and Machine Intelligence, vol. 25, no.12, pp. 1519-1533, Dec. 2003.
[5] POLYBIO website, http://polybio.signalgenerix.com/
[6] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data: http://www.cdt.org/privacy/eudirective/EU_Directive_.html

Palm geometry biometrics: A score-based fusion approach

Nicolas Tsapatsoulis and Constantinos Pattichis

Abstract In this paper we present an identification and authentication system based on hand geometry. First, we examine the performance of three different methods that are based on hand silhouette and on binarized hand images and then we investigate approaches for combining the feature vectors, identification-verification scores, and individual acceptance-rejections decisions taken by using each one of the proposed methods individually. The proposed system has been tested on the POLYBIO hand database which consists of 180 hand images from 45 individuals. The experiments show that fusion of feature vectors results in a slightly better performance in both identification and authentication tests while combination of scores and decisions leads to a significant improvement in authentication performance and minor improvement in identification.

1 Introduction

Biometrics technology aims to identify biological and behavioral features that are considered unique to a person and use them for authentication control in accessing secured places or devices. Biometric authentication systems are based on various modalities such as hand, iris, fingerprints, voice and face [7]. Fingerprints are by far the most widely used biometric for identification while iris is used for authentication control for large populations (i.e. at airports instead of using passports). Facial images are used in passport control for identification but the actual test involves the comparison of the photograph on the passport with the one stored in the database;

Nicolas Tsapatsoulis
Cyprus University of Technology, 31 Archbishop Kyprianos Str.,
CY-3036, Limassol, Cyprus, e-mail: nicolas.tsapatsoulis@cut.ac.cy

Constantinos Pattichis
University of Cyprus, 75 Kallipoleos Str.,
CY-1678, Nicosia, Cyprus, e-mail: pattichi@cs.ucy.ac.cy

that is, it stills based on something that one carries (passport) than something related with her/his live instance (face in this particular example). Speech recognition based authentication systems prove to be cost effective for simple access control implementations but cannot be used in high security zones. Finally, behavioral biometrics is an emerging technology not mature enough yet to be used in real life authentication or identification tasks.

Hand-geometry based authentication systems are gaining importance because they provide a good compromise between performance, cost of implementation and intrusiveness for security applications involving low to medium user population. Unfortunately as the user population increases the efficiency of hand-geometry based systems decreases [8]. However, combination with other forms of biometrics like fingerprint and palmprint is easy and can significantly increase the confidence levels in both identification and authentication procedures. The major advantage of hand geometry verification systems is the ease of image acquisition compared to the other biometric modalities. The acquisition system simply requires a properly placed camera that can get the image of the hand. Additional advantages of hand geometry systems include user-friendliness, non intrusiveness, and low template storage cost.

Hand geometry authentication systems based either on hand silhouette [8] [15] or on measurements extracted from palm and hand [14] [10]. The latter systems lead, in general, to better authentication performance but they are very sensitive to the localization of hand-extreme points based on which measurements are recorded [5]. Hand silhouette based systems, on the other hand, are much more robust, they have a compact mathematical representation and require less pro-processing effort.

In this paper we present three methods for hand geometry based identification and authentication. The two of them are based on hand silhouettes and involve Fourier descriptors and power spectrum estimation respectively, while the third uses the region and contour shape descriptors, proposed in MPEG-7 framework [6], extracted using the binarized hand image. In addition simple area measurements extracted from the binarized hand image are also examined for comparison purposes. In a further step we investigate feature, score and decision based fusion of the abovementioned methods in order to increase the authentication and identification rates.

The paper is organized as follows: In Section 2 we present the hand image acquisition system. Section 3 is devoted to the description of the individual methods for hand geometry authentication and identification. The proposed fusion methodology is explained in Section 4. In Section 5 we present the evaluation protocol we have employed along with extended experimental results. Finally conclusions are drawn and further work hints are given in Section 6.

2 Image acquisition

The multibiometric data were collected through the use of an integrated platform that was created in the framework of project POLYBIO [9]. The scenario in the

acquisition process was an office room where the acquisition hardware and software could be operated by a system supervisor, guiding the steps of the test subjects through the data collection procedure. Environmental conditions such as lighting or background noise were not controlled so as to simulate a realistic situation. The data acquisition system is depicted in Figure 1.

The process of collecting hand images is completed with the aid of the interactive panel shown in Figure 2. The user is prompted by the supervisor to place his/her left hand facing down-wards on the black board panel with the six positioning pins (pegs). The facing down-wards camera is activated and the palm images are taken, which if they are considered by the system supervisor of good quality, they are stored in the database. A total of four palm images are collected for each individual. The acquired palm images are of 240 pixels length x 320 pixels width, color ones (RGB model) and they are compressed using the JPEG compression scheme (quality 80%). More information on the palm image acquisition procedure can be fount at [1].

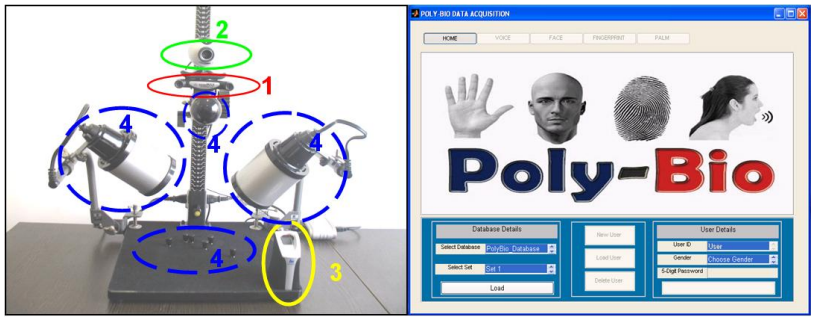


Fig. 1 The POLYBIO multibiometric data acquisition system

Fig. 2 Interactive panel for hand image acquisition

3 Biometric template creation

The biometric templates that are created from the hand images are based either on the palm contour or on the palm area. In both cases binarization of hand images is required. We used for this purpose a simple threshold approach. The threshold is obtained using the Otsu's method [11]. In a subsequent step morphological processing (the closing operator was applied) is adopted in order to fill in holes within the palm area.

3.1 The Fourier descriptor template

The first template was created using the Fourier descriptors of the palm contour. Let us consider the palm contour as a function of a complex variable $z(n) = x(n) + jy(n)$, $n = 0, 1, ..., M-1$, where M is the number of contour points and $(x(i), y(i))$ are the $2D$ coordinates of the i-th point. By taking the Fourier expansion of $z(n)$ we get:

$$a(k) = \sum_{m=0}^{M-1} z(m) \cdot e^{\frac{-2j\pi km}{M}}, \quad 0 \leq k \leq M-1 \tag{1}$$

The Fourier descriptors are the normalized amplitude coefficients of the Fourier series:

$$F_d(k) = \frac{a(k)}{\|\mathbf{a}\|}, \quad \|\mathbf{a}\| = [F_d(1)\ F_d(2) ...\ F_d(M)] \tag{2}$$

The Fourier descriptors actually indicate the frequencies of the curve changes along the contour. For denoising purposes palm contour is first approximated using 64 Fourier coefficients.

3.2 The power spectrum template

Assuming second order stationarity an approximation of palm contour's autocorrelation function is given by:

$$R(k) = \frac{1}{(M-k) \cdot \|\sigma\|} \sum_{m=0}^{M-1-k} (z(m) - \mu) \cdot (z(m+k) - \mu), \quad k < M-1 \tag{3}$$

where μ denotes the coordinates of contour's centroid and σ is the variance of contour's coordinates.

Taking the discrete Fourier transform of autocorrelation series lead us to an estimation of the contour's power spectrum:

$$PSD(k) = \sum_{m=0}^{M-1} R(m) \cdot e^{\frac{-2j\pi km}{M}}, \quad 0 \leq k \leq M-1 \tag{4}$$

The magnitude of PSD coefficients is used to describe the contour (only the coefficients with high energy value are used).

3.3 Area measurements

The following measurements of the binarized hand image were used to create another simple template:

1. *Ratio of Minor to Major axis length*: The minor to major axis length is the length (in pixels) of the minor /major axis of the ellipse that has the same normalized second central moments as the palm area.
2. *Solidity*: The proportion of the pixels in the convex hull that are also in the region (ratio of Area / ConvexArea).
3. *Extent*: It represents the pixels in the bounding box that are also in the palm region. It is computed as the Area divided by the area of the bounding box.
4. *Ratio of area to image dimensions*: Area refers to the actual number of pixels in palm region.
5. *Eccentricity*: Corresponds to the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1 (0 and 1 are degenerate cases; an ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment).
6. *Ratio of area to image dimensions*: It is computed as $\frac{Perimeter \times \pi}{Area}$
7. *Equivalent diameter / Major axis length*: Equivalent diameter is the diameter of a circle with the same area as the palm region. It is computed as $\sqrt{\frac{4 \cdot Area}{\pi}}$.

3.4 The MPEG-7 visual descriptor template

MPEG-7 visual descriptors include the color, texture and shape descriptor. A total of 22 different kind of features are included, nine for color, eight for texture and five for shape. The various feature types are shown in Table 1. In the third column of this Table is indicated whether or not the corresponding feature type is used in holistic image and/or object description. The number of features shown in the fourth column in most cases is not fixed and depends on user choice; we indicate there the settings in our implementation.

Four different templates were created from the hand image, corresponding to the Color Layout (CL) descriptor, the Contour Shape (CS) descriptor, the Region Shape (RS) descriptor, and the Edge Histogram (EH) descriptor. The features of these descriptors were computed using the MPEG-7 experimentation model [12].

Table 1 MPEG-7 visual descriptors used to create hand image templates

Descriptor	*Type*	*# of features*	*Usage level*	*Comments*
Color	DC coefficient of DCT (Y channel)	1	Both	Part of the Color Layout descriptor
	DC coefficient of DCT (Cb channel)	1	Both	Part of the Color Layout descriptor
	DC coefficient of DCT (Cr channel)	1	Both	Part of the Color Layout descriptor
	AC coefficients of DCT (Y channel)	5	Both	Part of the Color Layout descriptor
	AC coefficients of DCT (Cb channel)	2	Both	Part of the Color Layout descriptor
	AC coefficients of DCT (Cr channel)	2	Both	Part of the Color Layout descriptor
	Dominant colors	Varies	Both	Includes color value, percentage and variance
	Scalable color	16	Both	
	Structure	32	Both	They used in both holistic image and image segment description
Texture	Intensity average	1	Both	Part of the Homogeneous Texture descriptor
	Intensity standard deviation	1	Both	Part of the Homogeneous Texture descriptor
	Energy distribution	30	Both	Part of the Homogeneous Texture descriptor
	Deviation of energy's distribution	30	Both	Part of the Homogeneous Texture descriptor
	Regularity	1	Both	Part of the Texture Browsing descriptor
	Direction	1 or 2	Both	Part of the Texture Browsing descriptor
	Scale	1 or 2	Both	Part of the Texture Browsing descriptor
	Edge histogram	80	Both	Includes the spatial distribution of five types of edges
Shape	Region shape	35	Segment	A set of angular radial transform coefficients
	Global curvature	2	Both	Part of the Contour Shape descriptor
	Prototype curvature	2	Both	Part of the Contour Shape descriptor
	Highest peak	1	Both	Part of the Contour Shape descriptor
	Curvature peaks	Varies	Both	Describes curvature peaks in term of amplitude and distance from highest peak

4 Fusion methodologies

Two basic fusion methodologies were examined in order to identify whether or not fusion of different templates (even from the same modality) can enhance the performance of a hand-based biometric system. We first examined feature based fusion; that is the feature vectors of templates created using the previous methods where concatenated in various combinations (see also Table b 2).

Score based fusion was performed in two steps: We first normalized the scores achieved using the various templates by dividing with the highest threshold for each method so as the thresholds to lie in the interval $[0\ 1]$. Normalization is very important because non-normalized scores lead to performance lower to that obtained by feature based fusion. The second step includes weighting of scores so as the template with the better performance to contribute more in the total score. As in feature based fusion various combinations were examined. The results are summarized in Table 2

5 Evaluation protocol and experimental results

Evaluation was based on the POLYBIO multimodal biometric database [1] which contains samples from voice, face, palm and fingerprint for 45 individuals. Four data capture sessions were stored for each biometric. In our experiments we used the palm images of this database. Three images per individual were used for training and one for testing.

Let us denote with $\mathbf{f}_j^k$ the j-th ($j = 1,...,3$) palm feature vector of the k-th subject ($k = 1,2,...,N$). This feature vector is obtained with one of the methods described

in Section 3. We also denote with $\mathbf{y}^k$ the feature vector used for testing. Due to the limited number of training instances per subject (i.e., three) we consider as the biometric template of the k-th subject the matrix:

$$\mathbf{F}^k = [\mathbf{f}_1^k \ \ \mathbf{f}_2^k \ \ \mathbf{f}_3^k] \tag{5}$$

It is obvious that many different templates can be constructed depending on the number of training vectors. Gaussian models and Neural Network representations are among the most popular approaches for template construction and user modeling. In our case we have implicitly consider that all training instances serve as Support Vectors [3].

For each subject we also define a threshold:

$$T^k = max_{i \neq j}(||\mathbf{f}_i^k - \mathbf{f}_j^k||) \tag{6}$$

False Rejection (FR) and False Acceptance (FA) are then defined as:

$$FR \Leftarrow min_j(||\mathbf{y}^k - \mathbf{f}_j^k||) > T^k \tag{7}$$

$$FA \Leftarrow min_{j,\ l \neq k}(||\mathbf{y}^l - \mathbf{f}_j^k||) < T^k \tag{8}$$

We evaluated the palm biometric by using a four folder cross validation approach. Three instances per subject were randomly selected and used as training patterns while the fourth was used for testing. We repeated this process for 20 cycles and for each one of the individual and feature based fusion methods. The average results are shown in the Table 2. In this table it is also shown the results of the score based fusion approach for the combination of several types of features.

We used two widely known evaluation metrics: EER (equal error rate, i.e. FA = FR) and Identification Error (IE). An identification error occurs in cases where the best matching stored template does not belong to the individual that attempts to enter the system.

Among all individual template methods the Color Layout (CL) performs better in both IE (5.71%) and EER (6.29%). Disappointing results were obtained from the Contour Shape (CS) descriptor althiugh this descriptor was defined for retrieval of image objects. In contrary Edge Histogram (EH) descriptor provides also satisfactory rates although was defined for texture (and not object) description.

In feature based fusion the best results were obtained by concatenating the Color Layout, Contour Shape and Edge Histogram descriptors. This is a quite logical result because these descriptor provide complementary information (color, contour and texture). If we take into account the dimensionality of fused template then excellent results are also obtained by combining the Color Layout and Contour Shape descriptors.

In score-based fusion several combinations lead to satisfactory results. The best results in terms of IE (0%) were obtained using a combination of Color Layout, Contour Shape and Contour Shape descriptors. At the same time a combination

of Color Layout, Contour Shape and Edge Histogram descriptors leads to the best performance in terms of EER (2.12%).

Table 2 Evaluation results for the individual and fusion based methods in terms of identification error (IE) and equal error rate (EER)

Method	*Features*	*# of features*	*IE (%)*	*EER (%)*
FD	Fourier descriptors	8	20.00	14.15
FD	Fourier descriptors	16	11.43	14.61
FD	Fourier descriptors	32	12.14	16.62
PS	Power spectrum coefficients of contour	8	20.71	14.18
PS	Power spectrum coefficients of contour	16	13.57	13.84
PS	Power spectrum coefficients of contour	24	15.00	16.10
AF	Area related features	7	10.00	7.30
CL	MPEG-7: Color layout descriptor	12	5.71	6.29
CS	MPEG-7: Contour shape descriptor	5	43.57	15.48
RS	MPEG-7: Region shape descriptor	35	12.86	11.75
EH	MPEG-7: Edge histogram descriptor	80	10.71	7.16
FF1	Concatenated FD and PS	32	10.00	14.89
FF2	Concatenated FD and AF	23	6.43	9.93
FF3	Concatenated PS and AF	23	4.29	8.06
FF4	Concatenated CL and CS	17	1.43	3.87
FF5	Concatenated CL and RS	47	3.57	4.91
FF6	Concatenated CL and EH	92	3.57	3.33
FF7	Concatenated CS and RS	40	8.57	8.36
FF8	Concatenated CS and EH	85	5.00	5.76
FF9	Concatenated RS and EH	115	4.29	5.79
FF10	Concatenated FD, PS and AF	39	5.71	10.46
FF11	Concatenated CL, CS and RS	52	3.57	4.20
FF12	Concatenated CL, CS and EH	97	1.43	2.06
FF13	Concatenated CL, RS and EH	127	1.43	2.28
FF14	Concatenated CS, RS and EH	120	3.57	4.82
SF1	Score fusion of FD and PS	32	9.29	14.73
SF2	Score fusion of FD and AF	23	4.29	7.47
SF3	Score fusion of PS and AF	23	4.29	6.81
SF4	Score fusion of CL and CS	17	0.71	4.16
SF5	Score fusion of CL and RS	47	0.00	4.55
SF6	Score fusion of CL and EH	92	1.43	2.85
SF7	Score fusion of CS and RS	40	9.29	7.78
SF8	Score fusion of CS and EH	85	5.00	5.32
SF9	Score fusion of RS and EH	115	4.29	5.09
SF10	Score fusion of FD, PS and AF	39	2.86	7.30
SF11	Score fusion of CL, CS and RS	52	0.00	2.73
SF12	Score fusion of CL, CS and EH	97	0.71	2.12
SF13	Score fusion of CL, RS and EH	127	0.71	2.66
SF14	Score fusion of CS, RS and EH	120	2.86	4.52

6 Conclusion

In this work, we have presented an experimental study on palm geometry verification. The performance of several feature types including Fourier descriptors, power spectrum coefficients of palm's contour, area related measurements, and MPEG-7 visual descriptors was investigated. In addition both feature based and score based fusion was examined. Evaluation was based on 180 palm images obtained by 45 different users. The results indicate that: (1) Score based fusion provides the best results both in terms of equal error rate (EER) and identification error (IE), (2) both score based and feature based fusion lead to much better results than single method approaches, (3) Non-contour features, like the MPEG-7 color layout and edge histogram descriptors enhance the performance of the system but their robustness needs to be re-evaluated on data (hand images) obtained during different time periods, and (4) the best result is obtained by combining three MPEG-7 descriptors (color layout, contour shape, region shape) using score based fusion.

Future work includes the evaluation of the proposed score based fusion method on a larger dataset. We plan to use the data of the Biosecure Network of Excellence [2]. This network has been promoting since 2004 the development of biometric reference systems and reference databases. In addition decision based fusion and alternative score based fusion methodologies will be examined.

Acknowledgment. This work was undertaken in the framework of the POLYBIO (Multibiometric Security System) project funded by the Cyprus Research Promotion Foundation (CRPF) under the contract PLHRO /0506/04.

References

1. Antonakoudi A., Kounoudes A., Theodosiou Z. (2009). POLYBIO Multibiometrics database: Contents, description and interfacing. In: Proceedings of the Workshop on Artificial Intelligence Approaches for Biometric Template Creation and Multibiometrics Fusion (ArtIBio), Thessaloniki, Greece.
2. BioSecure: Biometrics for Secure Authentication http://biosecure.it-sudparis.eu/AB/. Cited 14 April 2009
3. Burges C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2: 121 - 167.
4. Duda R. O., Hart P. E., Stork D. G. (2004). Pattern Classification. Wiley-Interscience, 2nd edition.
5. Gonzales-Marcos A., Sanchez-Reillo R., Sanchez-Avila C. (2000) Biometric identification through hand geometry measurements. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10):1168-1171
6. ISO/IEC 15938-3:2001 Information Technology - Multimedia Content Description Interface - Part 3: Visual, Version 1.
7. Jain A. K., Ross A., Prabakar S. (2004). An introduction to biometric recognition. IEEE Transactions on on Circuits and Systems for Video Technology 14(1):4-20.
8. Konukoglu E., Yoruk E., Sankur B., Darbon J. (2006). Shape based hand recognition. IEEE Transactions on Image Processing 15(7):1803-1815.

9. Kounoudes A., Tsapatsoulis N., Theodosiou Z., Milis M. (2008). POLYBIO: Multimodal Biometric Data Acquisition Platform and Security System. Lecture Notes In Computer Science, 5372/2008:216-227.
10. Kumar A., Wong D. C. M., Shen H. C., Jain A. K. (2006). Personal authentication using hand images. Pattern Recognition Letters 27(13):1478-1486
11. Otsu N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on System, Man and Cybernetics 9:62-66
12. MPEG-7 Visual Experimentation Model (XM), Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4063, Mar. 2001.
13. Witten I. H., Frank E. (2005). Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.
14. Wong L., Shi P. (2002). Peg-free hand geometry recognition using hierarchical geometry and shape matching. In: Proceedings of IAPR Workshop on Machine Vision Applications, Nara, Japan, pp. 281-284
15. Yoruk E., Dutagaci H., Sankur B. (2006). Hand biometrics. Image and Vision Computing 24(5):483-497

Semantic Annotation, Publication, and Discovery of Java Software Components: An Integrated Approach

Zinon Zygkostiotis[1], Dimitris Dranidis[1, 2], Dimitrios Kourtesis[2]

[1] *Computer Science Department, CITY College,*
Affiliated Institution of the University of Sheffield,
Tsimiski 13, 54624 Thessaloniki, Greece
zzygkostiotis@city.academic.gr, dranidis@city.academic.gr

[2] *South East European Research Centre (SEERC),*
Research Centre of the University of Sheffield and CITY College
Mitropoleos 17, 54624, Thessaloniki, Greece
dkourtesis@seerc.org

Abstract: Component-based software development has matured into standard practice in software engineering. Among the advantages of reusing software modules are lower costs, faster development, more manageable code, increased productivity, and improved software quality. As the number of available software components has grown, so has the need for effective component search and retrieval. Traditional search approaches, such as keyword matching, have proved ineffective when applied to software components. Applying a semantically-enhanced approach to component classification, publication, and discovery can greatly increase the efficiency of searching and retrieving software components. This has been already applied in the context of Web technologies, and Web services in particular, in the frame of Semantic Web Services research. This paper examines the similarities between software components and Web services and adapts an existing Semantic Web Service publication and discovery solution into a software component annotation and discovery tool which is implemented as an Eclipse plug-in.

1. Introduction

The advent of rapid application development has led to an ever increasing emphasis on software reuse. Component-Based Software Development (CBSD) emphasises the reuse of existing code from either in-house repositories or 3rd party vendors, and has been shown to result in lower development costs, faster time-to-market, more effective maintenance and application upgrade, increased programmer productivity, and improved overall software quality [4, 18].

With software components being stored in code repositories, private or public ones, these repositories can potentially become extremely large. As they grow in

number and in size, so does the need to be able to search them effectively and retrieve component information and specifications. To enable this, there needs to be a standard way of representing this component-related information, thus facilitating Computer-Aided Software Engineering (CASE) tools in the discovery and retrieval of relevant results.

A number of search solutions have been proposed, developed and implemented for this purpose to date [20, 19], ranging from basic keyword searches to more advanced methods such as signature and behaviour matching using formal logic-based techniques. Traditional search approaches, such as keyword matching, are effective when searching Web pages and text documents. However, they have proven to be very inefficient when applied to software components. One of the reasons for this is that it is extremely difficult to convey sufficiently expressive domain-related information through a component's name or description.

The use of Semantic Web technologies in the annotation of software component information has enormous potential in achieving better targeted searches and more meaningful and accurate search results [20, 19, 14, 2]. Adding machine-processable semantic information to components and publishing this information in a standard way would make the classification, search and retrieval of components more effective, and would thus enable greater utilisation and easier integration of the vast number of software components currently available.

This paper presents an integrated approach for the annotation, publication and discovery of reusable Java software components through the use of Semantic Web technologies. We propose a method for annotating Java source code using domain ontologies that have been encoded in OWL-DL [10], and a means to publish and subsequently discover the resulting semantic descriptions using a semantic registry which employs Description Logic (DL) reasoning to perform matchmaking among software component advertisements and requests. Our approach has been validated through the development of a fully functional plug-in for the Eclipse IDE that supports all three facets of the approach: Java code annotation, publication of semantic descriptions, and search of software components. Our approach and implementation builds on earlier research work in the area of semantically-enhanced publication and discovery of Web Services, and relies on an existing open source semantic service registry for publication and discovery.

This paper is organised as follows. In Section 2 we look at various approaches for description and discovery of software components, explore similarities between software components and Web services, and report on a recently developed approach for publication and discovery of Web services with a semantically-enhanced service registry. Section 3 describes how this system can be adapted for use with software components and details an Eclipse plug-in developed for this purpose. We also provide an overview of the semantic matchmaking process when searching for software components, and outline the benefits this approach can bring over the use of traditional keyword-based and signature-based retrieval methods. Section 4 concludes the paper with a summary of the main points in this work.

2. Background of the Approach and Related Work

The basis of Component-Based Software Engineering (CBSE), also referred to as Component-Based Software Development (CBSD), is that certain functions and parts of large software systems appear numerous times within the system; therefore they should only be written once and not repeatedly throughout the application. Encompassing the required functionality into pluggable components and defining interfaces independent of any domain details allows components to be reused.

2.1 Software Component Retrieval Approaches

When the idea of software componentisation was first proposed by McIlroy in 1968 [11], it was recognized that a key requirement would be the indexing and retrieval of components. Currently, there is no universal agreement as to what information is required to describe software components such that they can be effectively retrieved. Existing code repository implementations tend to use proprietary or non-standard syntax and semantics for component descriptions and indexing, often employing quite elaborate classification schemes [3, 9, 8]. This inevitably makes searching in different code repositories even more difficult.

Mili et al. [12] group the types of search used for software component retrieval into four basic categories: simple keyword-based text searches, faceted classification and retrieval, signature matching and behaviour matching. Other research has classified all or some of these types into similar categories, such as Ostertag et al. [15], who also describe methods for free-text keyword searching and faceted index searching.

Keyword-based searching is the simplest approach to implement and is the one that the majority of search engines use. The successful retrieval of relevant components using this method is highly dependent on the original names given to the components and cannot take into account such information as relationships between components, their execution context and synonymous keywords. A software component retrieval scheme based on this approach is described in [13].

Faceted classification and retrieval involves extracting keywords from component descriptions and documentation and arranging this information into a predefined classification scheme or taxonomy. Although such an approach has been shown to be quite effective in the retrieval of relevant search results [15], it is only effective if the components fit into the classification scheme being used. Hence, a significant effort is required to maintain such classification schemes.

The signature matching approach, such as that described in [9], is rather detached from the application domain in that it attempts to describe components based on input and output parameters, creating a signature based on a mathematical algorithm. However, components having matching signatures are not guaran-

teed to be related. Behavioural matching extends signature matching somewhat in that it attempts to also describe the particular behaviour of a component. According to Suguraman et al. [19], both these approaches are cumbersome and inefficient.

2.2 Software Components vs. Services, and Semantic Retrieval

The similarities and differences of software components and Web services is regularly discussed throughout much of the literature in the field of CBSE. In [1], Breivold and Larsson provide a comparison framework for component-based and service-oriented software engineering and discuss the research efforts that have been done in combining the strengths of the two.

Yao et al. [20] suggest that there is little difference between software components and Web services, going as far as saying that a reusable component is in fact a service, and on this basis, the description and matching technologies employed for Semantic Web Services are just as applicable to software component description, classification and retrieval. Korthaus et al. [6] also investigate the use of Semantic Web technologies in the field of CBSE, and argue that CBSE can greatly benefit from the use of existing Semantic Web technologies for component classification, publication and discovery.

Paar [16] describes a Microsoft Visual Studio add-in developed for annotating C# source code with semantic information using ontologies encoded in DAML (the precursor of the OWL Web Ontology Language which is now a W3C standard) and WSDL (Web Service Description Language). The system annotates C# source code with references to ontology concepts and then extracts this information, converting it into a specially-adapted and semantically-extended form of WSDL. This WSDL file can then be used to advertise the component in much the same way as one would do for Web services.

Similarly, Yao et al. [20] describe a semantics-based approach to component classification and retrieval where software components are annotated with DAML ontologies. The component annotations and user queries are described in a WSDL format and then translated into "conceptual graphs", which are then used in their semantic matchmaking process. They also employ a software component repository based on the UDDI standard. However, one of the limitations they discuss was the lack of semantic support in both WSDL and UDDI.

2.3 The FUSION Semantic Registry

The use of Semantic Web technologies to represent Web service properties and the introduction of semantic matchmaking functionality in service registries (primarily UDDI) has been the focus of several works in recent years, generally within the field of Semantic Web Services (SWS) research. Kourtesis et al. [7]

present an approach that is focused at automating the evaluation of Web service integrability on the basis of the input and output messages that are defined in the service's interface. The approach has been applied during the development of the FUSION Semantic Registry, a semantically-enhanced service registry utilised in research project FUSION and released as open source software.

The aim in integrability-oriented service matchmaking within the FUSION Semantic Registry is to detect if interoperability at the level of data can be guaranteed among an advertised service and its prospective consumer, such that proper data flow and communication can take place. In plain terms, we seek to ensure that the data that the consumer is able to provide upon invocation are sufficient with regard to the input data that the advertised service expects to receive, and conversely, the output data that the advertised service produces are sufficient with regard to the data that the consumer expects to receive. This relates directly to the notions of covariance and contravariance applied in the context of function subtyping and safe substitution, which have been studied in detail within type-theory and object-oriented programming research [17].

In order to represent the functional and non-functional service properties that are of interest for matchmaking one needs to create a *Functional Profile* and define its key attributes in terms of references to an ontology encoded in OWL-DL. A *Functional Profile* is expressed as an OWL class with three types of object properties: *hasCategory* (representing the service's functional categorisation), *hasInput* (representing the service's set of input data parameters) and *hasOutput* (representing the service's set of output data parameters). There must always be one category declared, and zero or more inputs and outputs.

The purpose of describing services as OWL-based functional profiles is to enable semantic matchmaking among service advertisements and requests. When a service provider publishes a service advertisement, the service's profile is stored in the registry as an *Advertisement Functional Profile (AFP).* During the discovery process the requestor constructs a profile describing the desired service, i.e. a *Request Functional Profile (RFP).* Matchmaking among the two is performed through subsumption checking with a Description Logics reasoner (Pellet). Details on the publication and discovery algorithms are provided in [7].

3. Semantic Annotation, Publication and Discovery

In this paper we propose to build on the FUSION Semantic Registry infrastructure for classification, publication and retrieval of reusable software components. The following subsections detail how this can be realised in the form of an Eclipse plug-in for Java source code. Section 3.1 describes how components are described through ontology-based annotations. Section 3.2 provides an overview of the Eclipse plug-in and the functionality it offers. Finally, section 3.3 looks at the way in semantic matchmaking process is carried out within the Semantic Registry.

3.1 Semantic Annotations for Java (SA-Java)

In order to adapt the publication and discovery mechanisms of the FUSION Semantic Registry to cater for software components, we need a method of describing components similar to the one used for describing services, i.e. we need software component functional profiles. A Software component profile contains all information required to semantically describe, publish and search for reusable Java code in our approach. The information it holds is outlined in Table 1.

Table 1. Attributes of Advertisement Software Component Profiles

Attribute	Description
Identifier	A name given to the component for readability. It plays no part in semantic publication or discovery process but allows the component to be found via keyword-based search.
Description	A free-text description of the component. As with the identifier, it plays no part in the semantic publication or discovery process.
Category	The URI of an OWL concept describing the category to which the user has chosen to classify the component.
Inputs	The URIs of one or more OWL concepts describing the inputs the component expects. If this field is empty, the component does not require inputs.
Outputs	The URIs of one or more OWL concepts describing the outputs the component returns. If this field is empty, the component does not return any values.
RepositoryURI	The location where the component or component source code can be found. This can be a local or network file system location, Web URL or CVS/SVN repository location.

Advertisement Software Component Profiles are constructed automatically by the registry at the time of publication. Part of the information that is required for constructing them (Identifier, Description, and RepositoryURI) is obtained from the user, while the rest (Category, Inputs, and Outputs) is obtained by parsing the source code and retrieving semantic annotations placed there by the developer.

The method that we employ for source code annotation makes use of the standard annotation facility that was introduced by Sun with the release of Java 5.0 [5]. This allows adding metadata to code elements such as package declarations, type declarations, constructors, methods, fields, parameters, and variable declarations. The Java 5.0 platform comes with some predefined annotation types, but also allows developers to define their own types.

For the needs of our approach we have defined three types of annotations. The *Category* annotation is used to classify a Java class or method with regard to an ontological concept describing its purpose or application domain. For example, if a class contained methods and functions related to cash transactions from ATM machines, then the category annotation would provide a reference to an OWL concept describing this. Similarly, the *Input* and *Output* annotations are used to classify the inputs and outputs in terms of domain objects. The code snippet below

shows how a Java method annotated in this way could look.

```
@SAJavaCategory("http://www.seerc.org/onto.owl#ATM_Services")
public
@SAJavaOutput("http://www.seerc.org/onto.owl#Loggon_Confirmation")
boolean logon (
      @SAJavaInput("http://www.seerc.org/onto.owl#Card_Number")
      String userId,
      @SAJavaInput("http://www.seerc.org/onto.owl#PIN")
      String password )
{
    // method body
}
```

3.2 SA-Java Eclipse Plug-In

Our approach comes with a tool that interfaces with the FUSION Semantic Registry and supports Java source code annotation, publication of semantic descriptions, and search and retrieval of software components. The tool was developed as a plug-in for the popular Eclipse IDE and offers two separate views: annotation of source code is provided by the *Annotator view*, while component publication and discovery is provided by the *Semantic Registry view*. A screenshot of the SA-Java plug-in in the Eclipse workbench is shown in Figure 1.

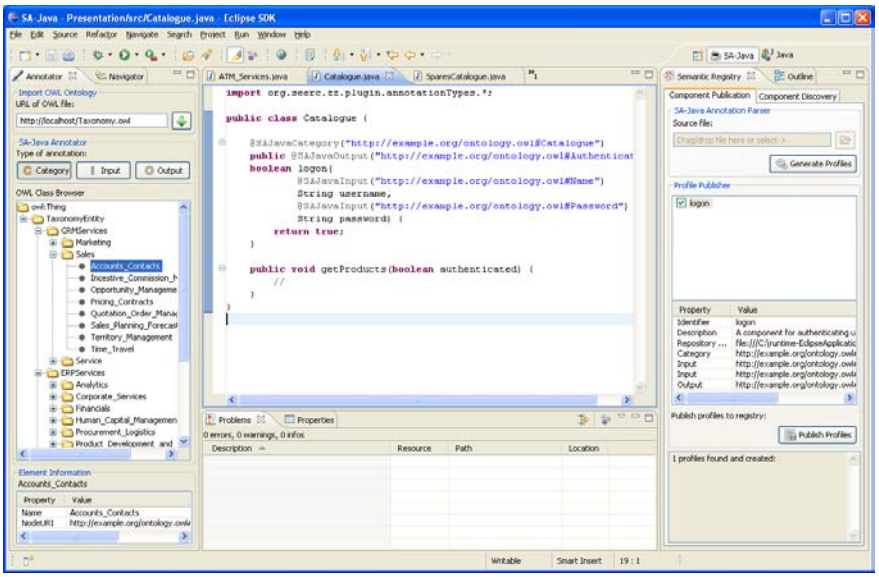


Figure 1. The SA-Java Plug-In and its different views

In order to annotate a Java source file with semantic information, an OWL file is first loaded into a browser in the Annotator view where the classes described in the OWL file can be examined. The user then selects the type of annotation required and, by *dragging and dropping* the OWL class directly from the browser to a point in the source code, the respective annotation is added.

Publication involves parsing an already annotated source file and creating candidate profiles. The plug-in scans a file for annotations and creates candidate profiles based on the annotated information that is found. A wizard is presented to the user who can examine the generated candidate profiles in turn, and edit or add any necessary information. On completion of the wizard, all generated candidate profiles are imported into the Semantic Registry view. The user can then select which of the candidate profiles should be actually published.

Component discovery is accomplished by creating Request Software Component Profiles. It is the profiles themselves that are used as search parameters rather than the traditional keyword text based approach most people are familiar with. The Registry view has a Profile Builder for this purpose where users can create software profiles which can then be sent to the registry for semantic matching. Any component profiles found in the registry that match the one sent as the search parameter are returned.

3.3 Semantic Matchmaking of Components

To illustrate the semantic matching process employed in the FUSION Semantic Registry, we use the examples of three methods whose profiles are detailed below.

Advertised profile of Class A *logon* method:
hasCategory: http://www.seerc.org/onto.owl#**Catalogue**
hasInput: http://www.seerc.org/onto.owl#**Name**
hasInput: http://www.seerc.org/onto.owl#**Password**
hasOutput: http://www.seerc.org/onto.owl#**Authenticated**

Advertised profile of Class B *signin* method:
hasCategory: http://www.seerc.org/onto.owl#SparesCatalogue
hasInput: http://www.seerc.org/onto.owl#Name
hasInput: http://www.seerc.org/onto.owl#Password
hasOutput: http://www.seerc.org/onto.owl#Authenticated

Advertised profile of Class C *logon* method:
hasCategory: http://www.seerc.org/onto.owl#ShoppingCart
hasInput: http://www.seerc.org/onto.owl#Name
hasInput: http://www.seerc.org/onto.owl#Password
hasOutput: http://www.seerc.org/onto.owl#Authenticated

Note the usage of OWL concepts for describing the different parts of the profiles. For instance, the category of the profile for the logon method of Class A has the value of *Catalogue*, which signifies that the profile either models or is related to a Catalogue object in the domain. The other two profiles have been categorized

as *SparesCatalogue* and *ShoppingCart*. For the purpose of this discussion let us assume that there exists an OWL-encoded taxonomy hierarchy in which the *SparesCatalogue* concept is defined as a subclass of *Catalogue*, whereas the *ShoppingCart* concept is a sibling class of *Catalogue.*

One thing we can observe in the example profiles is that all three require the same basic information as arguments and return the same result, regardless of the names they have been given in the method declarations and, perhaps more significantly, of the Java data types used for the arguments. Semantic annotation and profile generation makes no distinction between the Java data types used for elements, only what they represent.

As described earlier, to carry out a search using this approach, a request profile must be created that describes the required component. For example, we might be interested in finding any component that provides a method which accepts a username and a password as arguments and returns a response whether authentication has been successful, as in the following request profile:

Request Profile 1

```
hasCategory: http://www.seerc.org/onto.owl#Thing
hasInput:    http://www.seerc.org/onto.owl#Name
hasInput:    http://www.seerc.org/onto.owl#Password
hasOutput:   http://www.seerc.org/onto.owl#Authenticated
```

Searching using this profile would return all three of the advertised classes. This is because all three require two arguments that represent usernames and passwords (regardless of the Java data types used) and return a response indicating whether authentication has been successful. Also, their categorizations are all subclasses of the *Thing* concept (note that owl:Thing is the top concept in every OWL ontology and by definition subsumes every other possible concept).

We could modify the above profile to search for components that could be modelled as *Catalogue* objects. For this, we would need to replace the *hasCategory* URI with "http://www.seerc.org/onto.owl#Catalogue". This time, only classes A and B would be returned as matching, as they have been categorized as either of type Catalogue, or *SparesCatalogue* (which is a subclass of *Catalogue*). Class C's categorization is unrelated and so it would be excluded from the returned results.

The above example is a simple illustration of the semantic matching process based on the category classification of the profiles. The same procedure is applied when matching other elements of the profile, that is, the inputs and outputs, with even more interesting results. For example, the *Name* concept would also match any concepts that are a subclass of *Name*. When developers construct request profiles, what they are specifying is the number and types of inputs they can provide and the number and types of outputs they expect. In other words, they require a component that is related to a specific category and can return at least the outputs requested given at most the inputs that can be provided. To illustrate this, let us examine the following request profile.

Request Profile 2
hasCategory: http://www.seerc.org/onto.owl#SparesCatalogue
hasInput: http://www.seerc.org/onto.owl#Name
hasInput: http://www.seerc.org/onto.owl#Password
hasInput: http://www.seerc.org/onto.owl#EmployeeId
hasOutput: http://www.seerc.org/onto.owl#Authenticated

Searching using this profile would still return Class B even though the request profile has an extra input. This is because Class B can still provide the required output with only two of the three inputs the requestor is able to provide. Class B can therefore be utilized, and the extra input, *EmployeeId*, could be ignored. The opposite, however, is not true. Take, for example, the following request profile:

Request Profile 3
hasCategory: http://www.seerc.org/onto.owl#SparesCatalogue
hasInput: http://www.seerc.org/onto.owl#Password
hasOutput: http://www.seerc.org/onto.owl#Authenticated

This would return none of the three advertised classes. This is because they all require at least two inputs but the requestor here states that only one can be provided. Therefore, the components would not have enough information with which to carry out their tasks. The matching procedure with the outputs is similar but reversed. In this case, there is a match if the advertised profile can provide at least the outputs required by the requestor – any others can be ignored.

Hence we can see that applying a semantics-based approach to component search and retrieval is far more effective than traditional search approaches. Keyword-based and signature-based matching approaches cannot distinguish between components that display the same name/different functionality or different name/same functionality properties. Applying semantics not only goes a long way in solving this problem, but can also match components that can fulfil a request even if they are not a direct match.

4. Conclusions

The work presented has shown how Semantic Web technologies can be applied to CBSE, in particular, to the annotation, publication and discovery of software components. We proposed a method for annotating Java source code using domain ontologies encoded in OWL-DL, and a means to publish and subsequently discover the resulting semantic descriptions using a semantic registry which employs DL reasoning to perform matchmaking among advertisements and requests. Our approach is supported by a fully functional plug-in for the Eclipse IDE that supports annotation, publication and discovery of components, and is shown to offer significant benefits for retrieval of software components over the use of traditional approaches such as keyword- or signature-based matching.

References

1. Breivold H.P., Larsson M. (2007). Component-Based and Service-Oriented Software Engineering: Key Concepts and Principles. Proceedings of the 33rd EUROMICRO Conference on Software Engineering and Advanced Applications, pp.13-20.
2. Dong J.S. (2004). Software Modeling Techniques and the Semantic Web. Proceedings of the 26th International Conference on Software Engineering, pp. 724-725.
3. Graubmann P., Roshchin M. (2006). Semantic Annotation of Software Components. Proceedings of the 32nd EUROMICRO Conference on Software Engineering and Advanced Applications, pp. 46-53.
4. Haines G., Carney D., Foreman J. (2007). Component-Based Software Development / COTS Integration. Carnegie Mellon Software Engineering Institute, Pittsburgh.
5. JDK 5.0 Documentation. (2004). JDK 5.0 Developer's Guide: Annotations. Sun Microsystems Inc. http://java.sun.com/j2se/1.5.0/docs/guide/language/annotations.html.
6. Korthaus A., Schwind M., Seedorf S. (2007). Leveraging Semantic Web Technologies for Business Component Specification. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol., no. 2, pp. 130-141.
7. Kourtesis D., Paraskakis I. (2008). Combining SAWSDL, OWL-DL and UDDI for Semantically Enhanced Web Service Discovery. Proceedings of the 5th European Semantic Web Conference, LNCS 5021, pp. 614-628.
8. Lee J., Kim J., Shin G. (2003). Facilitating Reuse of Software Components using Repository Technology. Proceedings of the 10th Asia-Pacific Software Engineering Conference, pp. 136-142.
9. Luqi, Guo J. (1999). Toward Automated Retrieval for a Software Component Repository. Proceedings of the 6th Symposium on Engineering of Computer-Based Systems (ECBS '99), pp. 99-105.
10. McGuinness D.L., van Harmelen F. (2004). OWL Web Ontology Language Overview, W3C Recommendation
11. McIlroy D. (1968). Mass-Produced Software Components. Proceedings of the 1st International Conference on Software Engineering, Garmisch Pattenkirchen, Germany, pp. 88-98.
12. Mili R., Mili A., Mittermeir R.T. (1998). A Survey of Software Storage and Retrieval, Annals of Software Engineering, vol. 5, no. 2, pp. 349-414.
13. Mili A., Mittermeir R. (1994). Storing and Retrieving Software Components: a Refinement Based System. Proceedings of the 16th International Conference on Software Engineering (ICSE-16), pp. 91-100.
14. Oberle D., Eberhart A., Staab S., Volz R. (2004). Developing and Managing Software Components in an Ontology-Based Application Server. Proceedings of the 5th International Middleware Conference, LNCS 3321, pp. 459-477.
15. Ostertag E., Hendler J., Prieto-Diaz R., Braun C. (1992). Computing Similarity in a Re-Use Library System - an AI Approach. ACM Transactions on Software Engineering and Methodology, vol. 1, no. 3, pp. 205-228.
16. Paar A. (2003). Semantic Software Engineering Tools. OOPSLA '03, Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications, pp. 90-91.
17. Simons A.J.H. (2002). The Theory of Classification, Part 4: Object Types and Subtyping. Journal of Object Technology, vol. 1, no. 5. pp. 27-35.
18. Szyperski C. (1998). Component Software: Beyond Object-Oriented Programming: Addison-Wesley.
19. Sugumaran V., Storey, V.C. (2003). A Semantic-Based Approach to Component Retrieval. SIGMIS Database, vol. 34, no. 3, pp. 8-24.
20. Yao H., Etzkorn L. (2004). Towards a Semantic-based Approach for Software Reusable Component Classification and Retrieval. Proceedings of the 42nd Annual Southeast Regional Conference, Huntsville, Alabama, pp. 110-115.

Quality Classifiers for Open Source Software Repositories

George Tsatsaronis, Maria Halkidi, and Emmanouel A. Giakoumakis

Abstract Open Source Software (OSS) often relies on large repositories, like SourceForge, for initial incubation. The OSS repositories offer a large variety of meta-data providing interesting information about projects and their success. In this paper we propose a data mining approach for training classifiers on the OSS meta-data provided by such data repositories. The classifiers learn to predict the successful continuation of an OSS project. The 'successfulness' of projects is defined in terms of the classifier confidence with which it predicts that they could be ported in popular OSS projects (such as FreeBSD, Gentoo Portage).

1 Introduction

Initial open source software (OSS) projects rely on large repositories for hosting and distribution until they become independent. A huge amount of project meta-data is collected and maintained in such software repositories providing useful information about projects and their success. In this paper we propose a data mining approach that processes the meta-data contained in such OSS repositories. The proposed approach aims at the construction of a classifier that is trained on the meta-data of existing projects and predicts the successful continuation of any given OSS. The *successfulness* of a project is defined with regard to the confidence level of the classifier which predicts that this project will be ported in widely used OSS projects (e.g. FreeBSD). We argue that the classifier decision, along with its confidence level, can be incorporated into known models of software success, like the model of DeLone

George Tsatsaronis
Department of Informatics, Athens University of Economics and Business, 76, Patission Str., Athens, e-mail: gbt@aueb.gr

Maria Halkidi
Department of Technology Education and Digital Systems, University of Piraeus, e-mail: mhalk@unipi.gr

Emmanouel A. Giakoumakis
Department of Informatics, Athens University of Economics and Business, 76, Patission Str., Athens, e-mail: mgia@aueb.gr

and McLean [4], or its reexamination and expansion for OSS software, by Crowston et al. [3]. We have experimentally evaluated the proposed mining approach in the SourceForge and the FreshMeat OSS project meta-data released from the Floss project. We also evaluated the importance of the underlying features using Information Gain and Chi-Square. The results of this study report high F-Measures for the classifiers based on the most important FLOSS features.

The proposed approach consists of two main steps: a) data collection and preprocessing, b) training classifiers. The meta-data in OSS repositories are usually provided in formats that are not suitable for mining. Thus, one of the most important elements in the proposed approach is the *pre-processing procedure*. Techniques such as *parsing, crawling* and *feature selection* are used to collect data from the FLOSS project, which contains crawled projects from important OSS repositories, like SourceForge and FreshMeat. We also discuss methods that can be used to train classifiers on the stored project data.

The rest of the paper is organized as follows. Section 2 discusses related work in mining OSS projects and related models for measuring information systems' success. Section 3 presents the data mining approach for extracting knowledge from OSS repositories. Section 4 provides experimental results and analysis. Section 5 concludes and gives further insight to possible future work.

2 Related Work

Data Mining in Software Engineering. Data mining is widely used for supporting industrial scale software maintenance, debugging and testing. An approach that exploits classification methods to analyze logical bugs is proposed in [7]. This work treats program executions as software behavior graphs and develops a method to integrate closed graph mining and SVM classification in order to isolate suspicious regions of non-crashing bugs. A semi-automated strategy for classifying software failures is presented in [9]. This approach is based on the idea that if m failures are observed over some period during which the software is executed, it is likely that these failures are due to a substantially smaller number of distinct defects. A predictive model for software maintenance using data and text mining techniques is proposed in [10]. To construct the model, they use data collected from more than 100.000 open source software projects lying in the SourceForge portal. Using SAS Enterprise Miner and SAS Text Miner, they focused on collecting values for variables concerning maintenance costs and effort from OSS projects, like Mean Time to Recover (MTTR) an error. They clustered the remaining projects based on their descriptions, in order to discover the most important categories of OSS projects lying in the SourceForge database. Finally, they used the SAS Enterprise Miner to train classifiers on the MTTR class variable. The reported results highlight interesting correlations between the class variable and the number of downloads, the use of mail messages and the project age. There is also a number of other works [12, 1, 5, 9, 7] that use classification methods in software engineering, assisting

with its main tasks (development, debugging, maintenance, testing).

Models of OSS Success. The most popular model for measuring information systems' (IS) success is the one proposed by DeLone and McLean [4]. They actually introduce six interrelated factors of success: 1) system quality, 2) information quality, 3) use, 4) user satisfaction, 5) individual impact, and 6) organizational impact. Based on this approach, Seddon [11] reexamined the factors that can measure success and concluded that the related factors are system quality, information quality, perceived usefulness, user satisfaction and IS use. Based on those approaches, a number of measures that can be used to assess the success in FLOSS is presented by Crowston et al. in [3]. These measures are defined based on the results of a statistical analysis applied to a subject of project data in FLOSS. Specifically the empirical study was based on a subset of SourceForge projects. In this paper we propose another such measure, that can be added to the *use* and the *user satisfaction* factors of the proposed models of success.

3 Software Success Classification Approach

We introduce a classification approach that adopts data mining techniques in order to extract useful information from OSS repositories and further analyze it to predict softwares' successful continuation. Based on that, our approach aims at constructing a metric that assesses software success and which can be considered as an implementation of the factors *use* and *user satisfaction* to the model of DeLone and McLean [4] for measuring IS success. An expansion of this model has been proposed by Crowston et al. [3] for OSS software, according to which the data lying in FLOSS from SourceFourge are mapped to potential measures of OSS success. The following metric can be incorporated to the *System use* process phase, as this is described in [3]. We have developed the proposed metric taking into account the FLOSS data for both SourceForge and FreshMeat repositories.

3.1 OSS Porting Classification Metric

Without loss of generality we consider that OSS software ported to widely used open source operating systems is widely accepted as useful and significant by the majority of users. Then we claim that a project is considered to be *successful*, from the point of view of popularity and user satisfaction, if it is selected to be ported in two of the most popular open source operating systems, namely FreeBSD and Gentoo. Based on this, we construct classifiers, and we use the confidence of their classification output as a metric of OSS successfulness. The main modules of our approach can be summarized as follows: (a) *Data collection and pre-processing*. The OSS repositories maintain huge amount of OSS meta-data that provide useful information about the hosted projects. This module refers to methods used to collect data from OSS repositories and properly pre-process them for data mining.(b) *Software classification*. This module provides the methods to train classifiers for

predicting projects' course over time, trained on the collected OSS meta-data. The classifiers are built to predict the successful continuation of an OSS based on a specific set of project features. Based on the trained classifiers, new project releases (or unclassified projects) can be classified with regard to their successfulness (as this has been defined above).

Data Collection and Preprocessing Techniques

A large amount of data are collected and maintained in OSS repositories. These data contain useful information about projects that we aim to analyze so as to extract interesting knowledge and make inferences about future course of projects. However the data are provided in various formats that in most of the cases are not suitable for data mining. Thus, a pre-processing procedure is needed before data mining is applied to the available data. Specifically, techniques such as *crawling* and *feature selection* are used to collect project data from various OSS portals and select those that contain interesting information for further analysis. *Crawling* usually refers to the process of browsing the World Wide Web in a methodological and automated manner. An extensive analysis of a crawling mechanism is provided by Chakrabarti in Chapter 2 of [2]. We used a smaller crawling mechanism to browse the Web pages of the two open source operating systems (FreeBSD, Gentoo) and collect further information about the projects existing in SourceForge and FreshMeat for later processing. Many of the programming techniques used are described in [8]. For the processing of the SourceForge and the FreshMeat data, we used the FLOSS data and index. *Feature selection* is the technique of selecting a subset of relevant features for creating robust learning models. In our approach, feature selection techniques assist in a two-fold manner. Firstly, they assist the mining procedure with further analyzing the project data crawled. Thus, they provide an image of all the features being used. Secondly they assist in selecting the features that are more useful with regards to the final classification, which in our case is to predict whether an OSS software is successful enough to be ported in FreeBSD and Gentoo Portage. As feature selection criteria, we have adopted Information Gain (IG) and Chi-Squared (x^2).

According to the IG criterion, the expected reduction of uncertainty in guessing the class variable, once the feature value is known, needs to be measured. This is expressed through measuring the expected reduction in entropy, once the feature value is known, given by equation 3. While for the case of discrete value features equations 1-2 apply, in the case of the continuous value features, the domain of each feature variable X is divided into many subintervals of a given equal length and each X_i is true iff X belongs to the corresponding interval. Let S be the set of s data objects. Considering a set of m classes C_i, the expected information needed to classify a given object is defined as follows:

$$I(s_1,\ldots,s_m) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad (1)$$

where s_i is the number of data objects in S labeled c_i and p_i is the probability that an object belongs to class c_i, $p_i = s_i/s$. Let attribute A have v distinct values $\{\alpha_1, \alpha_2, \ldots, \alpha_v\}$. The attribute A can be used to partition S in v subsets $\{S_i\}_{i=1}^{v}$,

where S_i contains those objects in S that have value α_i for A. The entropy or expected information based on the partitioning of S into subsets by A is given by:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} - \sum_{i=1}^{m} p_{ij} log_2(p_{ij}), \ldots, s_{mj}) \quad (2)$$

where s_{ij} is the number of data objects of class C_i in a subset S_j, and $p_{ij} = \frac{s_{ij}}{|S_j|}$ is the probability that an object in S_j belongs to class C_i. The encoding information that would be gained by branching on A is:

$$Gain(A) = I(s_{1j}, \ldots, s_{mj}) - E(A) \quad (3)$$

Chi-squared (x^2) is often used to measure the association between two features in a contingency table. In the case of a binary classification problem it can be used to measure the association between an input feature and the class variable, as shown in the following equation:

$$x^2 = \sum_{i=1}^{r} \left(\frac{(n_{iP} - \mu_{iP})^2}{\mu_{iP}} + \frac{(n_{iN} - \mu_{iN})^2}{\mu_{iN}}\right) \quad (4)$$

where r are all the possible values of the examined feature, N and P are the two classes (Negative and Positive respectively), n_{iP} and n_{iN} are the number of instances belonging to class P or N respectively and have the value i of the examined feature, and $\mu_{ij} = \frac{n_{*j} n_{i*}}{n}$ with n being the number of instances, $i = 1, \ldots, r$ and $j = P, N$. In this work we use both IG and chi-square to measure the significance of the stored features contained in the OSS projects' meta-data, as crawled by the FLOSS project.

Software Classification

In order to train classifiers that can predict the degree of successfulness of an OSS lying in FreshMeat or SourceForge, the most important factor is to define the projects that belong to the classes of *successful* and *unsuccessful* projects. This can be the training set of projects that can be used for training the classifiers. The approach we adopt is to define as *successful* the OSS projects that have been ported in both FreeBSD and Gentoo Portage. This heuristic criterion satisfies part of the *System use* process phase of the model of successfulness defined in [3], namely *Interest*, *Number of Users* and *User Satisfaction*. We concluded to that criterion after having crawled the FreeBSD, Gentoo Portage and the Debian Popularity Contest, aiming to find their common set of projects with the considered FLOSS projects. FreeBSD Ports[1] is a package management system for the FreeBSD operating system and it contains all projects that are ported to FreeBSD. Portage[2] is also a package management system but it contains projects ported to Gentoo Linux. Furthermore, Debian Popularity Contest[3] is a project which attempts to map the usage of Debian packages. For the Gentoo Portage, we have used the latest snapshot[4] in order to determine the projects that are ported into it. From the Debian Popularity Contest we have collected for each package of Debian the number of users who installed

[1] `http://www.freebsd.org/`

[2] `http://www.gentoo.org/`

[3] `http://popcon.debian.org/`

[4] `http://gd.tuwien.ac.at/opsys/linux/gentoo/snapshots/`

it, the number of users who use it frequently, the number of users who do not use it frequently, the number of users who upgraded to the latest version of the software and the number of users that did not publish enough information. After having carefully analyzed all these data and examined several criteria that can be used as determining *successful* and *unsuccessful* projects, we concluded that the aforementioned criterion (porting of an OSS project in both FreeBSD and Gentoo Portage) expressed better the *System Use* as discussed earlier, and also provided better experimental results that other criteria examined.

Having determined the *successfulness* criterion, it is straightforward to construct classifiers taking into account the features offered by SourceForge and FreshMeat (an analysis of the features follows in the next section). Classification is a two step process:

1. *Training step.* A classification model is built describing a predefined set of classes (i.e *Successful*, *UnSuccessful*). The model is trained by analyzing a set of data whose classification (class labels) is known (training data set).
2. *Classification step*. First the predictive accuracy of the trained model (classifier) is estimated. If it is acceptable the classifier is used to classify project instances for which the class label is unknown.

Since each repository (FreshMeat, SourceForge) maintains different attributes for the hosted projects, a classifier per repository must be developed. Based on our approach any of the widely used classification algorithms can be used to analyze the training set of projects and construct the software classification model. In our current work, we have adopted *Naive Bayesian, Decision trees* and *Support Vector Machines* classification methods to train three different classifiers based on the available set of project data. For each of these classifiers, the successfulness metric is their confidence level of the classified instance. For the SVM, it is the distance of the considered project feature vector, from the hyperplane separating negative from positive instances. For the NB (Naive Bayesian) classifier, it is the probability that the considered project feature vector belongs to the *successful* class. Finally, for the decision trees (i.e. the C4.5 classifier) it is the frequency of occurrence of the training examples that are in the *successful* class, following the branch of the tree with attribute values of the examined instance. The experimental study in section 4 shows the performance of all three used classifiers for both FreshMeat and SourceForge.

3.2 Features Description

In this work we have used project data that have been collected and are available from the FLOSSMole project [5]. Specifically we work with the releases of SourceForge and FreshMeat project data indexed by the FLOSS repository. Below we present the main attributes (features) of the project data in the FLOSS repository that are used for the classification process. Also we indicate which portal (FreshMeat - FM or SourceForge - SF) maintains each attribute for the projects it hosts.

[5] `http://flossmole.sourceforge.net/`

1. *Project License (FM and SF):* Project's license type (such as GPL, LGP, BSD License, Freeware, Shareware etc)
2. *Vitality Score* (FM) which is defined as $VitalityScore = \frac{V*T_0}{T_n}$
   where V is the number of the project's versions, T_0 is the time elapsed since first project upload (usually counted in days), and T_n is the time elapsed since latest version upload.
3. *Popularity Score (FM):* Represents project's popularity based on:
   - Users' visits in project's URL, i.e. URL hits(further referred to as a). They refer to the visits in project's original web site, and not at FreshMeat's site for the project.
   - Users' visits in FreshMeat's project site (b).
   - Users' subscriptions (c).

   Then, the popularity score is measured as: $PopularityScore = \sqrt{(a+b)\cdot(c+1)}$
4. *Rating (FM):* Any subscribed user can rate a project. Given 20 or more user ratings, the project engages a ratings based ranking. Rating is measured as a means of a weighted ranking (WR), which is computed as follows:
   $$WR = (v(v+m))R + (m(v+m))\cdot C \tag{5}$$
   where: R = average (mean) rating for the project from users = (Rating)
   v = number of votes for the project = (votes)
   m = minimum votes required (currently 20)
   C = the mean vote across the whole report
5. *Subscriptions (FM):* Number of subscribed users in a project.
6. *Developers* (FM and SF): Number of developers per project and other information about them.
7. *Target audience / End Users* (SF): The target group of the resulting software.
8. *Executing operating system* (SF): The operating system in which the resulting project software can execute.
9. *DBMS Environment/Technology* (SF): For the projects using a DBMS, this is the used DBMS environment, or technology in general.
10. *Programming Language* (SF): The programming language in which the project software is written.
11. *Number of downloads* (SF).
12. *Interface* (SF): The interface of the project (Library, Command Line, Web, GUI etc).
13. *Natural language* (SF): The natural language of the project (English, France, etc).
14. *Topic* (SF): The topic of the project (Databases, Network Administration).
15. *Registration Date* (FM and SF): The date that the project was inserted into the portal.
16. *Days since Registration Date* (FM and SF): Days elapsed since the registration date.

17. *Project status* (SF): The status of the project (such as Beta, Planning, Production/Stable, Alpha, Pre-Alpha etc)
18. *Number of project donators* (SF).
19. *Rank* (SF): A ranking of the projects produced by SourceForge, considering downloads and days since registration date.

4 Experimental Evaluation

Experimental Setup. The projects that we considered for our experiments numbered 112915 for SourceForge and 41908 for FreshMeat. For both portals, we found the projects which are available from FreeBSD Ports and Gentoo Portage in order to label them as *successful*. For our experiments, we used 10-fold cross validation on three classifiers (decision trees, Naive Bayes, and SVM). For learning decision trees, we used the J48 algorithm that is WEKA's [13] implementation of the C4.5 (an extension of the basic ID3 algorithm) decision tree learner. We also used WEKA for the Naive Bayesian classifier. For support vector machines (SVMs) we used Joachim's SVM^{light} [6]. For each of the two data sets (FreshMeat and SourceForge), we trained all three classifiers on all of the features described. For the features evaluation IG and chi-square was used. For the evaluation of the classifiers, we measured precision, recall and F-Measure. The goal of this experiment is to prove the value of the OSS portals' meta-data, and consequently the value of the proposed success metric that is based on training classifiers, as an additional factor of OSS success.

Features Analysis. Feature selection requires analysis of all considered features. We have computed IG and chi-square values of all features using 10-fold cross validation, based on the used criterion. The measurements for the IG and the chi-square criteria are shown in Table 1 for the FreshMeat and the SourceForge repositories. The results are shown in decreasing order of importance based on the IG measure. From the results obtained we note primarily that the produced feature rankings based

	FreshMeat	
	Information Gain	Chi-Square
Popularity	**0.324**	**1793.254**
Subscriptions	**0.319**	**1756.487**
Vitality	**0.238**	**1781.661**
#Rating	**0.219**	**1253.699**
Rating	**0.189**	**111.144**
Days	**0.107**	**620.316**
Developers	**0.05**	**269.701**
License	**0.033**	**187.66**

	SourceForge	
	Information Gain	Chi-Square
Downloads	**0.199**	**759.95**
Rank	**0.153**	**595.337**
OS	**0.145**	**558.826**
Language	**0.138**	**533.17**
Days	**0.119**	**469.172**
Status	**0.095**	**381.716**
Interface	**0.078**	**317.054**
Developers	**0.075**	**298.099**
Users	**0.072**	**276.279**
License	**0.03**	**112.715**
DBMS	**0.01**	**5.548**
Donors	**0**	**0**

Table 1 Information Gain and Chi-Square for the FreshMeat and SourceForge Features.

on the two measures (IG and chi-square) are exactly the same in the case of SourceForge, while in the case of FreshMeat the only discrepancy is that the third feature according to IG is ranked second according to chi-square. This shows that the selected criterion of successfulness (porting of an OSS to FreeBSD and Portage) produces a stable ranking of features for both IG and chi-square. Regarding the features' importance, in the case of the FreshMeat data, the top 5 features proved to be *popularity*, *subscriptions*, *vitality*, *number of ratings* and *ratings*, while in the case of the SourceForge data, these are *downloads*, *rank*, *OS*, *language* and *days*. The IG drops dramatically for the rest features. At this point we must note that the top ranked features according to both measures include the features we were expecting to rank high, based on previously proposed models [3]. The feature ranking can be used to decrease the classifiers' model size. In the next section we show that the learned models can be reduced to only considering the top 5 features for each repository, without important decrease in performance.

Classifiers Evaluation. In order to measure the classifier's performance without introducing subset selection or feature selection bias, we have used 10-fold cross validation. The results from the 10 folds are averaged to produce a single estimation. We use *F-measure* to estimate the quality of each classifier. *F-measure* is defined as the harmonic mean between a classifier's precision and recall. All three measures were computed as follows:

$$Recall = \frac{TruePos}{TruePos + FalseN}, Precision = \frac{TruePos}{TruePos + FalsePos} \tag{6}$$

$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{7}$$

where *TruePos* is the number of the actual *successful* projects classified as *successful*, *FalseN* is the number of the actual *successful* projects classified as *unsuccessful* and *FalsePos* is the number of the actual *unsuccessful* projects classified as *successful*. Table 2 shows Precision, Recall and F-Measure values for the 10-fold cross validation execution of the J.48, Naive Bayes and SVM classifiers in the FreshMeat and the SourceForge data sets respectively. SVMs managed overall the top F-Measure compared to J.48 and the NB classifiers. For the FreshMeat data set, the classifiers reached an F-Measure of around 75%, with a precision reaching 88% for the SVMs. For the SourceForge data set, the classifiers reached an F-Measure of the same level with an overall smaller precision from the FreshMeat data set. In general, the classifiers performance for the SourceForge data set is smaller than in

	FreshMeat			SourceForge		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
SVM All Features	**0.88**	**0.57**	**0.69**	**0.67**	**0.75**	**0.7**
SVM Top-5 Features	**0.62**	**0.96**	**0.75**	**0.66**	**0.73**	**0.69**
C4.5 All Features	**0.79**	**0.81**	**0.79**	**0.77**	**0.71**	**0.73**
C4.5 Top-5 Features	**0.77**	**0.78**	**0.77**	**0.75**	**0.7**	**0.72**
NB All Features	**0.76**	**0.83**	**0.79**	**0.81**	**0.78**	**0.79**
NB Top-5 Features	**0.74**	**0.84**	**0.78**	**0.79**	**0.76**	**0.77**

Table 2 Precision (P), Recall (R) and FMeasure (F1) for SVM, C4.5 and NB in the FreshMeat and SourceForge data sets.

FreshMeat, depicting that FreshMeat's features are more descriptive for the used criterion of porting. This is also verified from the IG and chi-square feature values in Table 1. Overall, the proposed metric can predict whether a project will be ported into FreeBSD and Gentoo Portage, with high F-Measure.

5 Conclusions and Future Work

In this paper we propose a new Open Source Software (OSS) successfulness metric, that is based on the development of classifiers which predict the porting of an OSS into the FreeBSD and Gentoo Portage open source operating systems. We have evaluated the proposed metric by measuring the performance of Support Vector Machines (SVM), Decision Trees (C4.5) and Naive Bayes classifiers constructed on the features contained for all projects in FreshMeat and SourceForge. We also conducted an analysis of the features' importance and we experimentally show that the classifiers obtain similar performance if the top-5 features are kept, instead of all, which also include the most important features according to previous related work. As a future work we aim at combining heuristic criteria and/or manually annotated projects to enrich the training procedure with instances of better quality. We also aim at stacking classifiers for the purpose of boosting the classifier's performance.

Acknowledgments. We wish to thank D. Drosos for his assistance with the experimental study. This work was supported by EU Commision, FP6, contract No IST-5-033331 (SQO-OSS).

References

1. Bowring, J., Rehg, J., M.J., H.: Acive learning for automatic classification of software behavior. ISSTA (2004)
2. Chakrabarti, S.: Mining the Web. Morgan Kaufmann (2003)
3. Crowston, W., Hoison, J., Annabi, H.: Information systems success in free and open source software development: Theory and measures. Software Process Improvement and Practice, Vol. 11, pp. 123–148 (2006)
4. DeLone, W., McLean, E.: Information systems success: The quest for the dependent variable. Information Systems Research, Vol. 3(1), pp. 60–95 (2006)
5. Francis, P., Leon, D., Minch, M., Podguraki, A.: Tree-based method for classifying software failures. In: International Symposium on Software Reliability Engineering (2004)
6. Joachims, T.: Making large-scale SVM learning practical. Advances in Kernel methods - support vector learning. B. Scholkopf, C. Burges and A. Smola (ed.), MIT-Press. (1999)
7. Liu, C., Yan, X., Yu, H., Han, J., Yu, P.: Identifying reasons for software changes using historic databases. In: SDM (2006)
8. Loton, T.: Web Content Mining with Java. Wiley (2002)
9. Podgurski, A., Masri, W., McCleese, Y., Minch, M., Sun, J., Wang, B., Masri, W.: Automated support for classifying software failure reports. In: ICSE (2003)
10. Raza, U., Tretter, M.J.: Predicting software outcomes using data mining and text mining. In: SAS Global Forum (2007)
11. Seddon, P.: A respecification and extension of the delone and mclean model of is success. Information Systems Research, Vol. 8(3), pp. 240–253 (1997)
12. Williams, C., Hollingsworth, J.: Automating mining of source code repositories to improve bug finding techniques. IEEE Transactions on Software Engineering 31(6):466–480. (2005)
13. Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)

A Probabilistic Approach for Change Impact Prediction in Object-Oriented Systems

M.K. Abdi*, H. Lounis**, H. Sahraoui*

** Département d'Informatique et de Recherche Opérationnelle,*
Université de Montréal,
CP 6128 succ Centre-Ville, Montréal QC H3C 3J7, Canada
{abdimust, sahraouh}@iro.umontreal.ca
*** Département d'Informatique, Université du Québec à Montréal*
Case postale 8888, succursale Centre-ville, Montréal QC H3C 3P, Canada
lounis.hakim@uqam.ca

Abstract Several non probabilistic approaches were proposed in the literature to analyze and predict change impact in Object-Oriented (OO) systems. Different aspects were considered in these studies and several experiments were conducted to check some hypotheses. However, causality relation between software internal attributes and change impact still misses convincing explanations. In this paper, we propose a probabilistic approach using Bayesian networks to answer to this problematic of change impact analysis and prediction in OO systems. The built probabilistic model is tested on data extracted from a real system. The running of different scenarios on the network, globally confirm results already found in previous studies.

1 Introduction

Systems modification is a difficult task that has an impact on systems becoming [24]. Change effects must be considered. A small change can have considerable and unexpected effects on the system. Risks incurred during a modification are related to the consequence of a given change impact. When modularity is adequately used, it limits the effects relating to changes. Nevertheless, change impacts are subtle and difficult to discover; designers and maintainers need mechanisms to analyze changes and to know how they are propagated in the whole system.

The main motivation of our work is to improve the maintenance of object-oriented systems, and to intervene more specifically on change impact analysis. By identifying the potential impact of a modification, one reduces the risk to deal with expensive and unpredictable changes. Consequently, we try to give more explanations on real and responsible factors for change impact and its evolution. Among several models of representation, Bayesian networks (BNs) constitutes a particular quantitative approach which can integrate uncertainty within reasoning [20] offering thus explanations that are close to reality. Moreover, with BNs, it is

also possible to exploit experts' judgements to anticipate predictions, in our case, on change impact. In addition, BNs have the capacity of incremental training on data. This is true as well for parameters training as for structure training, facilitating the model evolution. This characteristic will contribute to the improvement of Bayesian network structure and parameters, by the acquisition of new data.

In this paper, section 2 presents various works related to change impact analysis. Our approach is presented in the third section. We start by presenting the principal stages of our approach, followed by a short recall on BNs. Then we illustrate gradually how to build the graph (BN) within the framework of our experimentation. After that, we explain the parameters assignment (probabilities) to network nodes. Section 4 concerns network execution and results discussion. Finally, our work perspectives are discussed in the conclusion.

2 Related works

Several studies were conducted on change impact. Thus, Han [12] developed an approach for computing change impact on design and implementation documents. This approach considers the original representation of software artefacts (classes) rather than a model of extracted system separately. The artefacts dependencies imply inheritance, aggregation and association. Furthermore, impacts are not defined in a formal way. On another side, Lindvall [19] identified the most common and frequent changes in C++, so that the change models can be specified to help developers to envisage the future needs. In [4], Antoniol and al. predicted evolving object-oriented systems size starting from the analysis of the classes impacted by a change request. They predicted changes size in terms of added/modified lines of code. Kung and al [16], interested by regression testing, developed a change impact model based on three links: inheritance, association, and, aggregation. They also defined formal algorithms to calculate all the impacted classes including ripple effects. Lee and Offutt examined in [17] and [18] the effects of encapsulation, inheritance, and polymorphism on change impact; they also proposed algorithms for calculating the complete impact of changes made in a given class. However, some changes, implying for instance inheritance and aggregation, were not completely covered by their algorithms.

In [7], impact analysis was made to reduce the costs and duration of regression tests. The study was made starting from a dependence graph. Briand and al. in [5], tried to see if coupling measures, capturing all kinds of collaboration between classes, can help to analyze change impact. This study, (i) showed that some coupling metrics, related to aggregation and invocation, are connected to ripple effect, and, (ii), it allows performing dependence analysis and reducing impact analysis effort. In [6], [14] and [15], a change impact model was defined at an abstract level, to study the changeability of object-oriented systems. The adopted approach uses characteristic properties of OO systems design (complexity, cohesion, coupling, etc.), measured by metrics, to predict changeability. According to a different perspective, Sahraoui and al. studied in [22] the impact of refactoring on structure and thus on structural metrics. This study made it possible

to determine the refactorings that can improve or deteriorate certain structural properties. Recently, in [1], [2], and [3], the authors also showed that coupling, measured by some metrics, influences change impact.

On the other hand, in [9], Fenton and Neil show well the advantages of the causal-modelling approach using Bayesian networks compared to the naive regression-based approach. In other works [10], [11], and [21], they also prove through case studies that Bayesian nets can provide relevant predictions, as well as incorporating the inevitable uncertainty, reliance on expert judgement, and incomplete information that are pervasive in software engineering. In this work, we try to explore this way of research and thus show the advantages of probabilistic approach using Bayesian nets compared to the approach adopted in our former work [1], [2], and [3].

In the following section, we present our approach (proposition) by explaining its different stages.

3 Proposition

The main stages of our approach are the following:

1- Graph structure construction (BN) starting from practical knowledge (empirical studies)
2- Parameters affectation (node probability table, fuzzy logic).
3- Bayesian inference (algorithms, tools)
4- Results

The first two stages are explained in the present section while the two last stages are presented in section 4. In order to facilitate the comprehension of used concepts in our approach, a recall on the basic concepts of Bayesian networks is essential.

3.1 Recall on Bayesian networks

BNs are based on the Bayes theorem. This theorem describes the relations which exist between simple and conditional probabilities. If A and B are two events and if we know the probability of A, of B and B knowing A, the Bayes theorem allows to determine the probability of A knowing B:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

BNs are the result of a merging between graph theory and probability theory [20]. A BN is a causal graph where:

- Nodes represent random variables. A random variable has some states, for example “Yes” and “No”, and a distribution probability for these states, where the sum of probabilities of all states must be equal to 1. Thus, a BN model is in conformity with the standard axioms of probability theory.

- Oriented edges define causal relations between nodes. An edge goes from a parent node towards a child node. Parent nodes which affect the same child node must be independent variables. Each node is related to a Node Probability Table (NPT), which models uncertain relation between the node and its parents. Tables of conditional probability related to BN nodes determine the force of the graph bonds and are used to calculate the distribution probability of each node in BN. This is carried out by specifying the conditional probability of a node knowing all its parents: p(X | A, B), X being the child node of A and B. If a node has no parent, a probability table would be associated for this node. Usually, NPTs are generally created by using a mixture of empirical data with experts judgement. In this causal graph, the cause and effect relationships between the variables are not deterministic, but probabilistic. Thus, observation of a cause or several causes doesn't involve systematically the effect or effects which depend on them, but modifies only the probability of observing them. The particular interest of BNs is to hold account as well of experts knowledge (in the graph or its structure) as of experiments contained in data (parameters).

3.2 Graph construction (BN)

Generally, the BN construction is done in two stages: produce the suitable graph because this model is sensitive to the type of applied reasoning, then affect probability values to network nodes [20]. The affectation of these values is done according to domain experts or starting from empirical studies. At this level, it is important to check that the parents nodes which affect the same child node are independent variables. Moreover, in order to respect the BNs construction formalism, during the introduction of bonds between nodes, it is necessary to check the absence of cycles between network nodes.

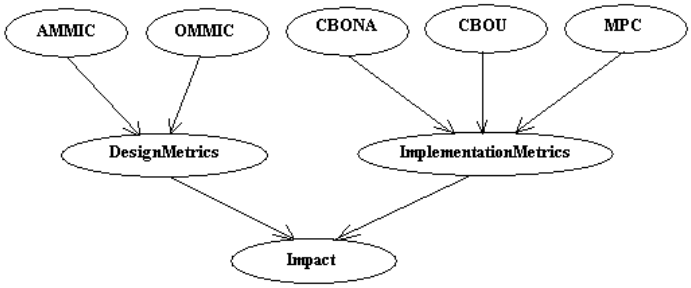


Figure 1. Change impact network

As already stated (in section 2), we checked in [1], [2], and [3], the hypothesis claiming that coupling influences change impact in an object-oriented systems. However, if we consider at the same time all metrics measuring the various facets of coupling between classes, the BN construction is likely to be hard and its structure complex. In addition, the results affirm that among the ten selected metric (see table 1), measuring this architectural property, five metrics are

effectively relevant to change impact. Some of these metrics are regarded as design metrics (*AMMIC* and *OMMIC*), others are considered as implementation metrics (*MPC*, *CBOU*, and *CBONA*). The figure 1 above presents the graph expressing this knowledge in the form of a BN. Let us note that in a BN, the relation between parents and child nodes are causal (case of *Impact* node) or definitional (case of *DesignMetrics* node).

Metrics	Definition
RFC	Response For a Class: number of methods called upon in response to a message.
MPC	Message Passing Coupling: number of messages sent by a class in direction of the other classes of the system.
CBOU	CBO Using: refers to the classes used by the target class.
CBOIUB	CBO Is Used By: refers to the classes using the target class.
CBO	Coupling Between Object: number of classes with which a class is coupled.
CBONA	CBO No Ancestors: CBO without considering the classes ancestors.
AMMIC	Ancestors Method–Method Import Coupling: number of parents classes with which a class has an interaction of the method-method type and a coupling of the type IC.
OMMIC	Others Method–Method Import Coupling: number of classes (others that super classes and subclasses) with which a class has an interaction of the method-method type and a coupling of the type IC.
DMMEC	Descendants Method–Method Export Coupling: number of subclasses with which a class has an interaction of the method-method type and a coupling of the type EC.
OMMEC	Others Method–Method Export Coupling: number of classes (others that super classes and subclasses) with which a class has an interaction of the method-method type and a coupling of the type EC.

Table 1. The selected coupling metrics

3.3 Parameters affectation

To affect probabilities to the nodes, it is necessary to distinguish two types of variable in BN: entry variables and intermediate variables. The entry nodes probabilities are directly deduced from measurements of these variables starting from a given test system. In our case, we chose a program analysis toolbox system, called BOAP, and, developed at the computer science research center of

Montreal (CRIM) [8]. It is a set of integrated software tools, which allow an expert to evaluate some software qualities, e.g., conceptual or structural weaknesses, too complex instructions, etc. We considered the BOAP system in its version 1.1.0; it is written in Java and contains 394 classes. The metric considered in this work are extracted from this system.

Entry nodes. In our network (figure 1), the entry nodes represent the different metrics. All these entry variables are quantitative variables which have measurable numerical values. The number of possible values for these variables can be infinite. That depends of course on the considered test system. In order to facilitate the probabilities definition, these variables are initially transformed into discrete variables having a limited number of values. This transformation can be accomplished by application of fuzzy logic. Indeed, the fuzzy partitioning process replaces the various values of a metric by a set of functions which represent the membership degree (or adhesion) of each value to the various fuzzy labels (often "small", "average" and "large"). The fuzzy partitioning generalizes the regrouping methods by groups allowing a value to be partially classified in one or more groups at the same time. The adhesion or the value membership is distributed in all groups. However, empirically, we can determine the optimal number of groups with statistics known under the name of Dunn partition coefficient F_k. This coefficient indicates us how to gather with a better way a data set in various groups [23]. The more the Dunn coefficient is high, the more the fuzzy subsets coincide classical logic sets. Therefore, the optimal number of groups is that which maximizes F_k. The Dunn partition coefficient is calculated according to the formula:

$$F_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{g=1}^{k} u_{ig}^2$$

N being the full number of observations (data), g the index for a group, k the number of groups and u_{ig} the value or the membership degree of a given object to a group.

Table 2 presents the results of fuzzy partitioning with 2 and 3 groups for the *AMMIC* metric. These results show that with two groups the Dunn coefficient is 0.8171413 and with three groups it is equal to 0.7768965. Therefore, for this metric, the partitioning in two groups is retained. Moreover, it is the same number of groups which was retained following the fuzzy partitioning tests for the four others metric. We used for that the statistics software S-plus (version 8.0) [13].

Table 3 gives an example of NPT for *AMMIC* node. It is about an example of value measured (equal to 25) for the *AMMIC* metric. To this value correspond two membership degrees (0.4349570 and 0.56504302) in the two fuzzy subsets. These membership degrees constitute the probabilities which are used to define the NPT of *AMMIC* node.

*** Fuzzy Partitioning ***

```
Membership coefficients:
numeric matrix: 394 rows, 2 columns.
        [,1]       [,2]
 1 0.9873814 0.01261863
 2 0.9873814 0.01261863
 3 0.9873814 0.01261863
 …       …          …

392 0.9873814 0.01261863
393 0.9873814 0.01261863
394 0.9873814 0.01261863

Coefficients:
 dunn_coeff normalized
  0.8171413  0.6342827
```

```
Membership coefficients:
numeric matrix: 394 rows, 3 columns.
         [,1]        [,2]        [,3]
 1 0.99358023 0.004435426 0.001984347
 2 0.99358023 0.004435426 0.001984347
 3 0.99358023 0.004435426 0.001984347
 …        …           …           …

392 0.9935802 0.004435427 0.001984347
393 0.9935802 0.004435427 0.001984347
394 0.9935802 0.004435427 0.001984347

Coefficients:
 dunn_coeff normalized
  0.7768965  0.6653447
```

Table 2. Example of fuzzy partitioning for *AMMIC*

Small	0.43
Large	0.57

Table 3. The NPT of *AMMIC* entry node

Intermediate nodes. The intermediate nodes are not directly measurable. They are defined or influenced by their parent nodes. For each intermediate node Cc which has possible values $\{V_{c1}, \ldots V_{ck}, \ldots V_{cn}\}$ and has parents $\{C_{p1}, \ldots C_{pi}, \ldots C_{pm}\}$ with possible values $\{V_{ci1}, \ldots V_{cij}, \ldots V_{cil}\}$, we need to define a table which gives the probabilities for all possible combinations of values:

$$P\,(V_{ck} \mid V_{p1j}, \ldots, V_{pmj})$$

These probability values can be adjusted by using machine learning starting from the sample data or the treated cases. A parent can influence positively or negatively his child nodes. The probability distributions are affected according to the importance or the weight of each parent for the child node. At the beginning, to derive NPT it is necessary to consider the weight of each parent node in definition or influence of its child node. For that, NPTs are initially given starting from studies in the field and experts opinions. For instance, the *DesignMetrics* variable is defined by its two parents *AMMIC* and *OMMIC*. It is a question of finding the conditional probability of *DesignMetrics* node: p (*DesignMetrics*| *AMMIC*, *OMMIC*). However, like the relation between the parent nodes *AMMIC* and *OMMIC* and their child node *DesignMetrics* is definitional, the strong presence of these metrics also defines the strong presence of *DesignMetrics*. A possible scenario for the *DesignMetrics* node NPT is presented in table 4:

AMMIC	Small		Large	
OMMIC	Small	Large	Small	Large
Oui	0.2	0.4	0.4	0.8
Non	0.8	0.6	0.6	0.2

Table 4. The *DesignMetrics* intermediate node NPT

A reasoning which can be applied is the following: if the number of classes (others that super-classes and subclasses) with which this class has an importation interaction of the method-method type is small (*AMMIC* small), and the number of parents classes with which this class has an importation interaction of the method-method type is small also (*OMMIC* small), the design metrics presence probability in such a system is weak or small. Therefore, the probability of the state "Yes" in the probability table of *DesignMetrics* node can be 20%. Conversely, if *AMMIC* is large, and *OMMIC* is large also, the probability of the state "Yes" of *DesignMetrics* node can be 80%. It is important to recall here that there are obviously other metrics (other than those considered in this study) and which are defined like design metrics or implementation metrics, and consequently, can positively or negatively influence change impact.

4 Bayesian Network execution

Once the graph structure and all NPTs are defined, we can proceed with the Bayesian inference. It results an update of conditional probabilities of all nodes. We have used the BNJ (Bayesian Network tools in Java) environment to achieve this goal. BNJ is a set of open source software tools intended for research and development by using graphic probabilities models. It is written in Java and is available on the web[1].

Let us recall that our experimentation was made on the BOAP test system (version 1.1.0) which contains 394 classes, or 394 instances. For the network execution, we will randomly choose an instance from which we take the metric values corresponding to entry nodes. As soon as the probabilities distributions are updated for introduced values, we will have an estimate in the form of probability for the various states assignated to the *Impact* node (figure 2).

Having affected three states «Weak», «Average», and «Strong» to the *Impact* node, and with the used input data (see figure above), we can conclude that the change impact has 43% of probability of being "Strong". The possibility of processing scenarios of the form « what will occur if… ?», that Bayesian networks offer, allows to identify potential problems and actions to be undertaken for improvement.

1. http://bnj.sourceforge.net/

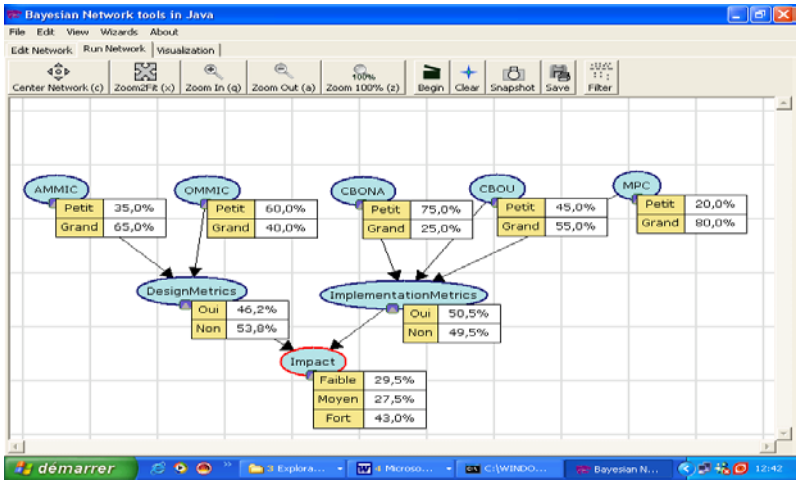


Figure 2. Change impact network after scenario 1

The scenario 2 execution shows that by decreasing the metrics values *CBONA* and *CBOU*, change impact weakens more (its probability of being "Weak" grows from 29,5% to 34,8%). Conversely, the scenario 3 execution shows that by increasing the *CBONA* and *CBOU* metrics values; the change impact becomes increasingly strong. The probability of the «Strong» state moves from 37,8% to 47,4%. Figure 3 illustrates this result.

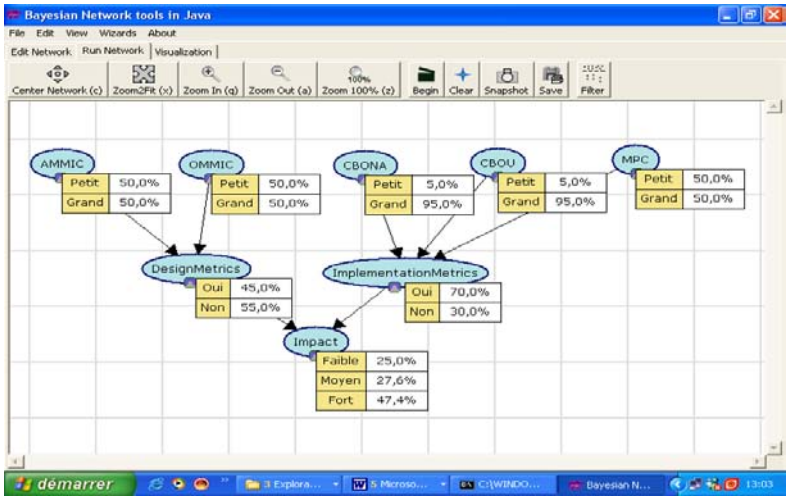


Figure 3. Change impact network after scenario 3

Finally, the last scenario execution shows that by maintaining the values of *CBONA* and *CBOU* metrics and by increasing the *AMMIC* one, change impact becomes little stronger. The probability of the «Strong» state grows from 37,8% to 41,5%.

Discussion

Results obtained in the second and third scenarios confirm those already found in our former work [1], [2], and [3], by using a non probabilistic approach (see respectively rule 1 and rule 2 of figure 4). For example, the scenario 3 result expressing that *CBONA* and *CBOU* metrics influence positively change impact, corresponds to the result illustrated by the causality rule 2 [2]:

Rule 1 : CBONA $\leq$ 3.5
CBOU $\leq$ 0.5
$\rightarrow$ impact: Weak (0.46)

Rule 2 : CBONA > 3.5
CBOU > 36.5
$\rightarrow$ impact: Strong (0.48)

Rule 3 : CBONA $\leq$ 3.5
CBOU $\in$]0.5,1.5]
AMMIC $\leq$ 0.5
$\rightarrow$ impact: Weak (0.54)

Rule 4 : CBONA $\leq$ 3.5
CBOU $\in$]0.5,1.5]
AMMIC > 0.5
$\rightarrow$ impact: Weak (0.76)

Figure 4. Causality rules examples

On the other hand, the scenario 4 result does not confirm one of our results (see rules 3 and 4 of figure 4) found before in [1] and [2]. Indeed, by maintaining the *CBONA* and *CBOU* metrics values small, and by increasing the *AMMIC* value, change impact does not become more weak. Its probability of being "Weak" was 34,8% then it was reduced to 30,8% whereas in theory, it must increase. In our opinion, that could be explained by the fact that change impact can be positively or negatively influenced by other metrics, other than those considered in the present study, or also, by other factors, like system size, complexity, etc.

5 Conclusion

We proposed in this article a probabilistic approach using Bayesian networks to analyze and predict change impact in object-oriented systems. A thorough study and a general synthesis of various former works dealing with this subject were initially essential. To verify our approach, we took again a correlation hypothesis between coupling and change impact already verified in former works. The experimentation was made on BOAP system. It contains 394 classes. The results of our empirical studies ([1], [2] and [3]) were useful for the graph structure construction (Bayesian Network). Thereafter, we defined the NPTs of entry and intermediate nodes. We used fuzzy logic to derive probabilities values starting from a set of measures (variables values or entry nodes).

The network execution and the creation of several scenarios enabled us to make predictions on change impact. The results of the second and third scenario confirmed results already found with other non probabilistic approach. On the other hand, the results of the fourth scenario contradict one of our results found before [2]. That leads us to search a hypothesis explaining this last result.

Finally, we are in the process of considering further experiments on other systems by including other coupling measurements, other architectural properties, or other factors which could supplement or better explain this causality relation.

References

[1] M.K Abdi, H. Lounis, H. Sahraoui: "Using Coupling Metrics for Change Impact Analysis in Object-Oriented Systems" In QAOOSE 2006 Proceedings, 10th ECOOP Workshop on Quantitative Approaches in Object-Oriented Software Engineering, 3 July 2006 - Nantes, France

[2] M.K Abdi, H. Lounis, H. Sahraoui: "Analyzing Change Impact in Object-Oriented Systems " In proceedings of the 32nd EUROMICRO Software Engineering and Advanced Applications Conference, Cavtat/Dubrovnik (Croatia), August 29-September 1, 2006.

[3] M.K Abdi, H. Lounis, H. Sahraoui, M.K Rahmouni : "Vers une approche d'analyse de l'impact du changement dans un système à objets", dans revue "L'Objet", volume 13 – N° 1/2007, Pages 147-169, Éditions Hermès.

[4] G. Antoniol, G.Canfora, A. D. Lucia, "Estimating the size of changes for evolving Object-Oriented Systems : a Case Study" *in Proceedings of the 6th International Software Metrics Symposium*, pages 250-258, Boca Raton, Florida, Nov 1999

[5] L.C. Briand, J. Wüst, H. Lounis, "Using Coupling Measurement for Impact Analysis in Object-Oriented Systems" *in proceedings of the International Conference on Software Maintenance ICSM'99,* Oxford, England, August 30 – September 3, 1999.

[6] M.A. Chaumun, H. Kabaili, R.K. Keller and F. Lustman. *"A Change Impact Model for Changeability Assessment in Object-Oriented Software Systems".* In Proceedings of the Third Euromicro Working Conference on Software Maintenance and Reengineering CSMR'99, pages 130-138, Amsterdam, The Netherlands, March 1999.

[7] R. Cantave, Abstractions via un modèle générique d'application orientée objet, Master's thesis, Université Laval, Canada, Avril 2001

[8] E. Alikacem, H. Snoussi, *"BOAP 1.1.0 : Manuel d'utilisation",* CRIM, Janvier 2002.

[9] N.E. Fenton and M. Neil, "Software Metrics: Roadmap", in 'The Future of Software Engineering' (Editor: Anthony Finkelstein) 22nd International Conference on Software Engineering, ACM Press ISBN 1-58113-253-0, pp.357-370, 2000.

[10] N.E. Fenton and M. Neil, "The Jury Observation Fallacy and the use of Bayesian Networks to present Probabilistic Legal Arguments", Mathematics Today (Bulletin of the IMA, 36(6)), 180-187, 2000.

[11] N.E. Fenton and M. Neil, "Making Decisions: Using Bayesian Nets and MCDA", Knowledge-Based Systems 14, 307-325, 2001.

[12] J. Han, "Supporting Impact Analysis and Change Propagation in Software Engineering Environments" *in Proceedings of the STEP'97*, London, England, pages 172-182, July 1997.

[13] Insightful Corporation, Seattle, WA, *S-PLUS® 8 for Windows® User's Guide*, Copyright © 1987-2007.

[14] H. Kabaili, R.K. Keller, F. Lustman, and G. Saint-Denis. Class Cohesion Revisited: An Empirical Study on Industrial Systems. In *Proceedings of the Workshop on Quantitative Approaches in Object-Oriented Software Engineering*, pages 29-38, Cannes, France, June 2000

[15] H. Kabaili, *"Changeabilité des logiciels orientés objet : propriétés architecturales et indicateurs de qualité"*, PhD thesis, Université de Montréal, Canada, Janvier, 2002

[16] D.C. Kung, J. Gao, P. Hsia, J. Lin, Y. Toyoshima, "Class firewall, test order, and regression testing of object-oriented programs" *in Journal of Object-Oriented Programming*, Vol. 8, No. 2, pages 51-65, May 1995.

[17] M. Lee, A.J. Offutt, "Algorithmic Analysis of the Impact of Changes to Object-Oriented Software" *in Proceedings of the ICSM'96*, pages 171-184, 1996.

[18] M. Lee, Change Impact Analysis for Object-Oriented Software, PhD thesis, George Mason University, Virginia, USA, 1998

[19] M. Lindvall, "Measurment of change : Stable and Change-Prone Constructs in a commercial C++ System" *in Proceedings of the 6th International Software Metrics Symposium,* pages 40-49, Boca Raton, Florida, Nov 1999

[20] P. Naïm, P. Wuillemin, P. Leray, O. Pourret, A. Becker, "Réseaux bayésiens", Edition Eyrolles, 2004

[21] Neil M, Fenton NE, Nielsen L, "Building large-scale Bayesian Networks", The Knowledge Engineering Review, 15(3), 257-284, 2000.

[22] H.A. Sahraoui, R. Godin, T. Miceli, "Can metrics help to bridge the gap between the improvement of OO design quality and its automation ?", *in Proceedings of the International Conference on Software Maintenance (ICSM'00)*, 2000

[23] E. Trauwaert, On the meaning of Dunn's partition coefficient for fuzzy clusters, *Fuzzy Sets and Systems,* Vol.25, No 2, pp 217-242, 1988.

[24] N. Wilde, R. Huitt, "Maintenance support for object-oriented programs" i*n IEEE Transactions on Software Engineering,* Vol. 18, Issue 12, Pages 1038–1044, Dec 1992.

Improving Evolutionary Test Data Generation with the Aid of Symbolic Execution

M. Papadakis[1] and N. Malevris[1]

Abstract Recently, search based techniques have received great attention as a means of automating the test data generation activity. On the contrary, more traditional methods that automate the test data generation usually employ symbolic execution by incorporating a path generation phase and constraint solvers to produce the sought test data. In this paper, the benefits of both schools of thought are bridged in an attempt to investigate whether a mixed strategy approach could be employed when evaluating a coverage criterion. To this effect, a strategy that uses symbolic execution and dynamic domain reduction in order to enhance the initial population and approximately prune the search space considered by evolutionary based methods is proposed. This suggestion is also put under a number of tests which clearly show a dramatic improvement of its effectiveness. This suggests that the combination of evolutionary based and symbolic execution approaches can be beneficial toward, automating the generation of test data.

1. Introduction

It is a well known fact that the cost of software testing can reach 50% or even 60% of the total software development cost. In order to reduce the cost overhead, a lot of effort has been put by the software engineering community, in an attempt to automate the testing activity and thus reduce the overall software development cost. The usual way to evaluate the test thoroughness of a piece of software is to establish a collection of testing requirements that must be fulfilled when the software is executed against test cases, through a number of test coverage criteria that guide and evaluate the effectiveness of the test data generated.

Test data generation techniques based on Genetic Algorithms (GAs) have been adopted in the literature [7, 16] and a number of researchers have proposed vari-

[1] Athens University of Economics and Business {mpapad, ngm}@aueb.gr

ous algorithms based on both local [2, 16] and global searches [7, 8, 16] with an appropriate adaptation of their fitness function according to the coverage criterion considered. Evolutionary Algorithms (EA) for test data generation approaches have been widely studied in the literature [2, 7, 8, 13, 16] although a number of unresolved research issues still remain [7, 13], mainly on the efficiency and effectiveness of these approaches.

A hybrid approach to test data generation that integrates GAs and symbolic execution in a complementary way is proposed. Existing attempts [18, 19] use structural methods and EAs in an effort to facilitate the testing exercise. The purpose of this study is to guide the input domain selection in such a way so as to improve and make more efficient the EA. Symbolic execution is used not as a complimentary tool but as a yardstick, towards this purpose. It is used to guide the test data selection rather than evaluating a suggested path. In this respect, the proposed method have no common philosophy with the refs above [13, 18, 19]. Our goal is to demonstrate that information from path testing and symbolic execution can be used effectively in a combined way, in order to improve the efficiency and effectiveness of the EA. The proposed work stems from suggestions in the literature about the reduction of the search space [9, 12] and the observation that candidate solutions near the target ones can be efficiently refined through local search algorithms. We argue that the fulfillment of these two directions can be achieved through path testing in a systematic and automated way, relying on the ESPM method (hereafter called "Y&M") suggested by Yates and Malevris [4]. Through this approach a set of linearly independent candidate paths is produced, while all candidate paths have a high probability of being feasible. Here, these two attributes of the Y&M method are exploited in order to refine the initial population set of the EA. In order to reduce the search space when targeting to a particular branch [4], the assumption of the Y&M method is utilized which suggests that a feasible path has higher probability to be contained in a set of the first *k*-shortest paths. Based on this assumption, which was substantiated in [4], some irrelevant variables are pruned away and domain reduction is performed by using the DDR procedure proposed in [5], over the common constraints that constitute a subset of the constraints derived from the selected path set. The reduced domain space then forms the search space of the EA.

Our approach has been empirically investigated with reference: a) the impact of the initial population enhancement; b) the impact of the input domain reduction procedure; and c) the overall improvement of the proposed approach, clearly showing a dramatic improvement of the evolutionary approaches when guided by path testing.

2 Incorporated methods

Symbolic Execution: The symbolic evaluation process [1] of a program consists of assigning symbolic values to variables in order to deduce an abstract alge-

braic representation of the program's computations and representation. This technique is based on the selection of paths from its control flow graph and the computation of symbolic states. The symbolic state of a path forms a mapping from input variables to symbolic values and a set of constraints called *path conditions* over those symbolic values. Path conditions represent a set of constraints called *symbolic expressions* that form the computations performed over the selected path. Solving the path conditions results in test data which if input to the selected path, this will be executed. If the path condition has no solution the path is infeasible.

In this paper, the Y&M method [4] was chosen because of its flexibility and its ability to generate likely to be feasible paths. The method can be detailed as: ***Step1:*** Generate a set of program paths, whose constituent paths each involve a minimum number of predicates, and cover the target elements of code unit under test. ***Step2:*** Symbolically execute the current path set and determine the achieved coverage. ***Step3:*** Select uncovered elements and generate alternative path sets giving priority to those containing a lower number of predicates. Steps 3 and 2 are repeated *k* times or until the target coverage is achieved.

Evolutionary Algorithms: A GA test data generator employs a genetic algorithm as its primary search engine in seeking suitable test data according to a target test adequacy criterion. The basic steps of a GA are the following [7, 13]: ***Step1***. Create an initial population of candidate solutions. ***Step2***. Compute the fitness values of each of these candidates. ***Step3***. Select all the candidates that have the fitness values on or above a threshold. ***Step4***. Perturbate each of these selected candidates using genetic operators.

These steps, except the initialization step, are repeated until the coverage goal is met or until the time limit has been reached. Before the use of a GA we need to define the following: Some domain-dependent attributes, a representation of the problem solution in terms of genetic inputs (chromosomes), the fitness function, the candidate selection methods (chromosomes reproduction) and the genetic operators.

```
Input: Program P, Target Coverage, Time_Limit
Output: A set of test Inputs
Graph ← GenerateGraph(P);
Tests ← RandomInputs();
K = 1;
repeat
1    Paths ← GenerateY&M_Paths(Graph, k);
2    Tests ← Symbolic Execution ( Paths );
3    Targs ← GenTargets(Graph, Tests, Targs);
4    Tests ← EvolutionaryTesting(Targs);
5    for all ( Targs not covered )
         DomainReduction (Targs);
     K = K + 1;
until (Coverage==Target OR Time<Time_Limit)
return;.
```

Fig. 1. Proposed Algorithm

3 Approach Description

The motivation behind our work is the combination of static and dynamic approaches namely symbolic execution and evolutionary testing, in order to improve their performance. Our framework tries to guide the search along two directions. First by enhancing the initial population with results from established testing techniques such as [4] and second, by using the same method to dynamically and ap-

proximately reduce the input search space. Our approach uses *k* paths from the path set of the Y&M method [4] in order to prune away all the irrelevant input variables and uses the common parts of those paths to define a common set of constraints. The algorithm in Figure 1 outlines the proposed test data generation scheme. Given a program P, the algorithm iteratively computes sets of inputs based on both symbolic execution and evolutionary testing in order to achieve the targeting coverage. The algorithm loop as in figure 1 contains three basic steps (lines 1-5 of fig. 1). (a) Starts by generating and executing symbolically the current set of paths according to the Y&M method. The resulting test data enhance the randomly generated initial population of the evolutionary algorithm. (b) The test generation process continues with the evolutionary algorithm phase, where newly test inputs are constructed. (c) The final step of the loop is the domain reduction process according to uncovered branches. The algorithm terminates when the time limit is exceeded or the coverage goal is met.

Symbolic execution: The Symbolic execution process phase is carried out based on the Y&M method for a given parameter value *k* see [4] for details. Here we use some approximations in order to overcome, abstract away, unhandled expressions and generate test data. Note that solutions will be refined from the evolutionary algorithm so we only need to generate solutions close to the target ones. Unhandled expressions such as non linear constraints are ignored or replaced with new symbolic values. Constructs such as pointer values will be approximated through simple expressions containing only (in)equality constraints. The resulting path conditions are then passed to the constraint solver for feasibility check. If the considered path condition is found infeasible, the path is marked as infeasible. Otherwise, test data will be produced which enhance the current population of the evolutionary algorithms.

Evolutionary testing: The evolutionary algorithm phase takes the instrumented version of the program under test, the program's CFG and evolutionary parameters, and produces tests according to the evolutionary method used. In the present study we use the SGA&D[2] as used in [7, 16]. In our study we evaluate, based on symbolic execution, the improvement made over these two algorithms by considering two approaches (approach 1, 2) and a third which combines previous two (approach 3).

Standard GA (approach 1) setup: ***Representation:*** A chromosome as a bit string representing test cases. ***Evaluation function:*** Each chromosome is evaluated simply by using the program's CFG and accounting the number of decisions covered. ***Selection:*** A new population is formed by selecting the best chromosomes in terms of coverage and ***Recombination:*** Performing mutation and crossover operations.

Differential GA (approach 2) setup: ***Representation:*** A chromosome as a bit string representing test cases. ***Evaluation function:*** Each chromosome is evaluated according to its percentage coverage contribution. ***Selection:*** New population is formed by selecting the best chromosomes in terms of coverage and ***Re-***

[2] Standard Genetic Algorithm and Differential

combination: For each test case A={a_1, a_2, … a_n} the algorithm selects randomly two different mates B={b_1, b_2, … b_n} and C={c_1, c_2, … c_n} from the population. Then according to probability (Change Factor) it selects input values (a_i) for alteration. For all selected values (a_i) the algorithm calculates new ones based on the formula: $a_i' = a_i + \alpha (b_i - c_i)$ according to α (Factor). If the resulting A′={ a_1', a_2', … a_n'} solution performs better based on the objective function, the solution A is replaced by A′.

Hybrid approach of Standard and Differential GA (approach 3) setup*:* This algorithm forms the combination of the two preceding ones by applying the standard one first and then the differential one by using the final population of the standard algorithm as the initial population to the differential one.

Domain reduction: The goal of the proposed approach is to guide the search by reducing the search space and thus improve its application as stated in [9]. The main idea behind the reduction algorithm is to use the set of paths of Y&M method in order to dynamically define a domain approximation of the targeting branch. Through our research in path testing we have observed that many paths that covering particular branch share many parts. This observation helps as to reduce the search space by using domain reduction approaches as in [5] only for the common parts of the Y&M path set. We are interested about the common parts of the paths and so we eliminate all non common conditions. Additionally we detect irrelevant variables as those not used in any path condition of our set. The common set of constraints forms the targeting reduced domain approximated by using a similar to the Dynamic Domain Reduction method [5], alternative to the constraint propagation as used in [14]. The main assumption behind the approach is the same with the Y&M method: at least one feasible path is contained in a set of selected k-shortest paths and so reducing the search space based on them directs the search to the contained feasible paths.

Our approach reduces the search space using binary search on input domain space based on the concerned constraints. For each constraint in the set, the algorithm reduces appropriately the domain space based on the variables, their comparison use in the constraint set and their assigned domain space. The domains of input variables are split in half *(binary search)* based on the constraint comparison.

4 Empirical Setup and Results

In order to evaluate the proposed approach we have implemented a semi automated prototype tool. Our tool consists of the *symbolic execution module*, the *domain reduction module* and the *evolutionary testing module*.

Test Subjects. We used a set of thirteen java and C programs for our experiment. This set of test objectives contains some common programs used in test data generation studies [5, 7, 16] such as the Triangle classification, the Quadratic formula, the Euclidian algorithm, Binary Search, Bubble sort and the Calendar. The

other programs were taken from study [10]. Table 1 shows details of the pro grams.

Evaluation Procedure. The equipment and standards used in our experiments include:

1. Adopt the GA parameters for: **Approach 1:** (Standard GA) infinite MaxGens, MaxTime 30 Sec and MaxPopulation = 1000. **Approach 2:** (Differential GA) Infinite MaxGens, MaxTime 30 Sec and MaxPopulation = 1000, factor="0.5" changefactor="0.5". **Approach 3:** (Both Standard and Differential GA) same setting with MaxTime 20 Sec for each of the two approaches.
2. Run the prototype tool on a Core 2 Duo 1.66GHz with 2 GB Ram computer running the Windows Vista operating system.

Table 1. Subject programs

Test Object	Lines of Code	No. of Branches
Triangle	40	46
Quadratic	12	6
Triangle2	38	35
Euclidian	9	10
Binary search	20	17
Bubble sort	14	11
Calendar	22	12
Insert	26	20
Dbll	86	78
C prog One	45	23
C prog Two	32	17
C prog Three	64	31
C prog Four	28	30

3. For every approach and in order to avoid any form of bias from random effects we introduced 3 experimental cases under which we ran every experiment 10 times and compared their average values. ***For case 1:*** First run the EA in isolation and then the algorithm for parameter k = 1 (i.e. include steps 1, 2, 3, 4 of the proposed algorithm, fig. 1).***For case 2:*** First run the EA in isolation and then the domain reduction procedure followed by the evolutionary algorithm (i.e. include steps 4, 5 of the proposed algorithm fig. 1). ***For case 3:*** Run the proposed algorithm for k = 60 and for MaxTime = 30s. (i.e. include all algorithm steps from 1-5, fig. 1)

Case 1. In the current case we tried to examine the impact of population initialization through path testing on evolutionary based test data generation. In order to simulate and observe the behavior of GAs through the guidance from the symbolic execution process we initialized the population using random testing and symbolic execution based only on the basic path set produced from the use of the Y&M method (parameter value *k=1*).

Table 2 records the average coverage achieved within the same time limit for both the original evolutionary algorithm (Before) and the proposed approach (After) which uses randomly generated initial population (Before) and the proposed approach which enhance the initial population (After). The results of this experiment show that, for the programs used, the initial population enhancement improved the evolutionary algorithm effectiveness by on average 9.82% for Standard GA, 8.49% for Differential and 7.83% for their combination.

Table 2: Branch Coverage achievement before and after initial population enhancement.

Test Object	Cov. Before/After (Approach1)	Cov. Before/After (Approach2)	Cov. Before/After (Approach3)
Triangle	61.06% / 82.33%	57.87% / 86.59%	62.55%/ 87.74%
Quadratic	100% /100%	100% /100%	100% /100%
Triangle2	63.06% / 97.22%	60.83% / 98.89%	65.56 / 97.78%
Euclidian	*100% /100%*	*100% /100%*	*100% /100%*
Binary search	*71.655% /100%*	*98.647%/100%*	*98.647%/100%*
Bubble sort	*100% /100%*	*100% /100%*	*100% /100%*
Calendar	*100% /100%*	*100% /100%*	*100% /100%*
Insert	*100% /100%*	*100% /100%*	*100% /100%*
Dbll	*79.23% / 81.54%*	*84.74 % / 88.52%*	*88.52% / 96%*
C prog One	*72.17% / 73.91%*	*70.87% / 73.91%*	*73.91%/ 73.91%*
C prog Two	*52.94% / 88.24%*	*52.94% / 86.47%*	*52.94%/ 85.29%*
C prog Three	*84.52% / 89.03%*	*90.32% / 92.26%*	*89.03%/ 92.26%*
C prog Four	*76.67% / 76.67%*	*76.67% / 76.67%*	*76.67%/ 76.67%*
Average	**81.64%/91.46% (+9.82%)**	**84.07%/92.56% (+8.49%)**	**85.22%/93.05% (+7.83%)**

Case 2. In the current case we tried to examine the impact of domain reduction on the performance of GA by considering the domain reduction process described in section 3. We used k-value: 60 (number of selected paths). In order to simulate and observe the behavior of GAs through the guidance of the reduced domain we concentrated only on those branches that were left uncovered from the evolutionary algorithm on the initial search space. Table 3 records the average coverage improvement in the same time limit over the initial approach.

Table 3: Branch Coverage improvement achieved through domain reduction.

Test Object	Cov. Improvement (Approach1)	Cov. Improvement (Approach2)	Cov. Improvement (Approach3)
Triangle	*4.3%*	*11.2%*	*12.1%*
Triangle2	*1.2%*	*5.2%*	*5.6%*
Binary search	*0%*	*0%*	*0%*
Dbll	*0%*	*0%*	*0%*
C prog One	*1.8%*	*2.9%*	*0%*
C prog Two	*1.4%*	*2.3%*	*2.3%*
C prog Three	*3.2%*	*-1.2%*	*1.7%*
C prog Four	*5.2%*	*6.8%*	*5.8%*
Average	***1.55%***	***2.47%***	***2.50%***

The results of this experiment show that, the domain reduction procedure improves the evolutionary algorithm effectiveness by on average 1.55% for Standard GA, 2.47% for Differential and 2.50% for their combination.

Case 3. In the current case we tried to examine the overall improvement in terms of coverage of the proposed approach (proposed algorithm figure 1) over the evolutionary testing. The experiments where conducted within the same time limit (30s) for all approaches. Table 4 records the average coverage achieved in the same time limit for the evolutionary methods (Approaches 1, 2 and 3). Based on the positive findings of cases 1 and 2, case 3 incorporates these results as they are reflected by the algorithm in Figure 1.

Table 4: Branch Coverage achievement within the same time limit.

Test Object	Cov. GA/proposed (Approach1)	Cov. GA/ proposed (Approach2)	Cov. GA/ proposed (Approach3)
Triangle	*61.06% / 100%*	*57.87% / 100%*	*62.55% / 100%*
Quadratic	*100% /100%*	*100% /100%*	*100% /100%*
Triangle2	*63.06% / 100%*	*60.83% / 100%*	*65.56 / 100%*
Euclidian	*100% /100%*	*100% /100%*	*100% /100%*
Binary search	*71.655% /100%*	*98.647%/100%*	*98.647%/100%*
Bubble sort	*100% /100%*	*100% /100%*	*100% /100%*
Calendar	*100% /100%*	*100% /100%*	*100% /100%*
Insert	*100% /100%*	*100% /100%*	*100% /100%*
Dbll	*79.23% / 100%*	*84.74 % / 100%*	*88.52% / 100%*
C prog One	*72.17%/76.27%*	*70.87%/76.27%*	*73.91%/76.27%*
C prog Two	*52.94% / 100%*	*52.94% / 100%*	*52.94% / 100%*
C prog Three	*84.52% / 100%*	*90.32% / 100%*	*89.03% / 100%*
C prog Four	*76.67% / 97%*	*76.67% / 97%*	*76.67% / 97%*
Average	***81.64%/98% (+16.36%)***	***84.07%/98% (+13.93%)***	***85.22%/98% (+12.78%)***

The results of this experiment show that, for the test programs considered, the proposed algorithm in figure 1 outperforms the EA by on average 16.36% for Standard GA, 13.93% for Differential and 12.78% for their combination.

5 Related Work

Generally there has been a considerable amount of work in the area of automatic test generation based on both symbolic execution [1, 3, 5] and evolutionary techniques [2, 7, 8, 11, 13, 15, 16, 17]. Closest to our research is the work by Xie [6] where two automated tools, one for evolutionary testing and one for symbolic execution have been integrated in order to improve the structural testing of object-

oriented programs. They propose a framework that attempts to improve the coverage by targeting to sequences of method calls through evolutionary testing and to their coverage improvement by symbolic execution. The integration is based only on static input initialization from one tool to the other and not by simultaneously and dynamically completing one another as in our approach. In [12], another similar to [6] approach that combines static analysis and search based methods is proposed. The main objective of this work is a theoretical and empirical evaluation of the effects of domain reduction through irrelevant variable removal. In their empirical study the authors used static analysis based on program slicing in order to identify and discard the irrelevant variables. In [2] a search reduction was made considering the path coverage criterion. Each input variable was ranked according to influence graph constructed using dynamic data flow information. The variable value would remain unchanged if it was likely to impact segments that were currently being traversed correctly or if it did not affect the path. In [16], a framework that utilizes two optimization algorithms, the Batch-Optimistic and the Close-Up is proposed. This approach uses a domain control mechanism witch starts with small domain spaces and modifies its boundaries to larger ones at subsequent phases. Nevertheless, they did not guide the domain reduction through symbolic execution. In [11] Andreou et al, propose a specially designed genetic algorithm for data flow criteria which relies on the data flow graph.

6 Conclusions and Future work

This paper presented a test data generation technique that integrates symbolic execution and GAs in an effective and complementary way. The motivation behind our proposal is to combine both static and dynamic analysis techniques in order to complement each other and result in an improved generation process. This technique tries to guide the search of EA through symbolic execution and domain reduction, via two main directions. First by using symbolic execution on a limited set of paths and second by identifying common constraints that form a reduced domain. We also described the results obtained for evaluating the effectiveness of our approach, by using both guidance directions in isolation and in combination.

Here, the attempt has been to investigate whether symbolic evaluation can assist an existing genetic algorithmic approach rather than to be compared against it. Under such circumstances the feeling of the authors is that the amelioration in the results presented here will hold with different GAs. The choice of the most effective genetic algorithm to be used as a basis in the enhancement process, is however a matter of future research.

In all three experiments with a set of programs, three different GAs were used (basically two as the third is a combination of the other two). In all experiments, the support offered by symbolic execution via path generation, improved the coverage ability on two counts. First by increasing the achieved coverage by 8.7% on average as in case 1, while for cases 2 and 3 the improvement was on average

again 2.17 and 14.4% respectively and second by improving on the run time as it provided a reduction in the domain from which sample values ought to be generated. These two activities do highlight the strength of the proposed method i.e. of combining symbolic execution with the employment of a search based technique in an attempt to evaluate a coverage criterion when structurally testing a piece of software. Future work is directed towards conducting more experiments in order to statistically validate the claims of the present findings. Series of experiments are also planned to determine the optimal use of symbolic execution and the domain reduction process.

References

1. King, J. C. (1976). Symbolic execution and program testing. *Commun ACM* **19**(7), 385-394.
2. Korel, B. (1990). Automated Software Test Data Generation. *IEEE Trans. Softw. Eng*. **16**(8), 870-879.
3. Koutsikas, C., Malevris, N. (2001). A Unified Symbolic Execution System. *AICCSA* 466-469.
4. Yates, D. F., Malevris, N. (1989). Reducing the Effects of Infeasible Paths in Branch Testing, *in proc of Symposium on Testing, Analysis, and Verification,* 48-54.
5. Offutt, A. J., Jin, Z., Pan, J. (1999). The Dynamic Domain Reduction Procedure for Test Data Generation. *Softw., Pract. Exper.* **29**(2), 167-193.
6. Inkumsah, K., Xie, T. (2008). Improving structural testing of object-oriented programs via integrating evolutionary testing and symbolic execution. *In Proc. ASE*, 297–306.
7. McGraw, G., Michael, C., Schatz, M. (2001). Generating software test data by evolution. *IEEE Trans. Softw. Eng*, **27**(12), 1085-1110.
8. Pargas R., Harrold M., Peck R. (1999). Test-data generation using genetic algorithms. *Software Testing, Verification and Reliability*, **9**(4), 263-282.
9. Chen, S., Smith, S. (1999). Improving genetic algorithms by search space reductions. *GECCO*, 135-140.
10. Malevris, N., Yates, D. F. (2006). The collateral coverage of data flow criteria when branch testing. *Information & Software Technology* **48**(8), 676-686.
11. Andreou, A. S., Economides, K. A., Sofokleous, A. A. (2007). An Automatic Software Test-Data Generation Scheme Based on Data Flow Criteria and Genetic Algorithms. *CIT* 867-872
12. Harman, M., Hassoun, Y., Lakhotia, K., McMinn, P., Wegener, J. (2007). The impact of input domain reduction on searchbased test data generation. *ACM FSE*, 155-164.
13. McMinn, P. (2004). Search-based software test data generation: A survey. *Software Testing, Verification and Reliability*, **14**(2), 105-156.
14. Hentenryck, P., Saraswat, V., Deville, Y., (1998). Design, implementation and evaluation of the constraint language cc(fd) . *Journal of Logic Programming*, **37**, 139-164.
15. Xanthakis, S., Ellis, C., Skourlas, C., Gall, A. L., Katsikas, S., Karapoulios, K. (1992). Application of genetic algorithms to software testing *ICSEA*, 625-636.
16. Sofokleous, A. A., Andreou, A. S., (2008). Automatic, evolutionary test data generation for dynamic software testing. *Journal of Systems and Software* **81**(11), 1883-1898.
17. Ferguson, R., Korel, B. (1996). The chaining approach for software test data generation. *IEEE Trans. Softw. Eng,* **5**(1), 63-86.
18. Girgis, M. R., (2005). Automatic Test Data Generation for Data Flow Testing Using a Genetic Algorithm, *Jucs*, **11**(6), 898-915.
19. Ahmed, M. A., Hermadi, I. (2008). GA-based multiple paths test data generator, *Computers and Operations Research* **35**(10), 3107-3124.

Reliable Confidence Intervals for Software Effort Estimation

Harris Papadopoulos, Efi Papatheocharous and Andreas S. Andreou

Abstract This paper deals with the problem of software effort estimation through the use of a new machine learning technique for producing reliable confidence measures in predictions. More specifically, we propose the use of Conformal Predictors (CPs), a novel type of prediction algorithms, as a means for providing effort estimations for software projects in the form of predictive intervals according to a specified confidence level. Our approach is based on the well-known Ridge Regression technique, but instead of the simple effort estimates produced by the original method, it produces predictive intervals that satisfy a given confidence level. The results obtained using the proposed algorithm on the COCOMO, Desharnais and ISBSG datasets suggest a quite successful performance obtaining reliable predictive intervals which are narrow enough to be useful in practice.

1 Introduction

Accurate software cost estimation has always been a challenging subject impelling intensive research from the software engineering community for many years now [11]. Especially over the last 50 years researchers have been working to improve the estimation accuracy of the methods proposed and provide

Harris Papadopoulos
Computer Science and Engineering Department, Frederick University, 7 Y. Frederickou St., Palouriotisa, Nicosia 1036, Cyprus. e-mail: H.Papadopoulos@frederick.ac.cy

Efi Papatheocharous
Department of Computer Science, University of Cyprus, 75 Kallipoleos str., CY1678 Nicosia, Cyprus e-mail: efi.papatheocharous@cs.ucy.ac.cy

Andreas S. Andreou
Department of Computer Science, University of Cyprus, 75 Kallipoleos str., CY1678 Nicosia, Cyprus e-mail: aandreou@cs.ucy.ac.cy

a competitive edge for software companies and managers. An accurate estimate, especially from the early stages of the software project life-cycle, could provide more efficient management over the whole software project resources and processes. Such estimates, even for well-planned projects, are hard to make and thus it may prove helpful to handle the uncertainty of the estimates through an effort prediction interval. Nevertheless, the difficulty to produce intervals for effort estimation reflecting a new project remains due to the high level of complexity and uniqueness of the software process. Estimating robust confidence intervals of the required software costs, as well as selecting and assessing the most suitable cost drivers for describing the escalating behavior of effort, both remain difficult issues to tackle and are constantly at the forefront right from the initiation of a project and until the system is delivered.

In this work, we aspire to tackle the uncertainty of the software process and the volatility of leading cost factors in the development environment by producing confidence intervals for software effort. Most cost models and techniques proposed in literature provide a single estimate for effort and usually have poor generalization ability [8]. Our approach attempts to promote reliable confidence intervals to reach better solutions in such approximation problems so as to alleviate the deficiencies of the techniques proposed so far and to address the problem in a possibly more effective and practical manner.

The present paper proposes the use of a novel machine learning technique, called Conformal Prediction (CP) [24], which can be used to produce predictive intervals, or regions, that satisfy a required level of confidence. Conformal Predictors (CPs) are based on conventional machine learning algorithms and transform the output of these algorithms from point to confidence interval predictions. The most important property of CPs is that they are well calibrated, meaning that in the long run the predictive intervals they produce for some confidence level $1-\delta$ will not contain the true label of an example with a relative frequency of at most δ [16]. Furthermore, this is achieved without assuming anything more than that the data are distributed independently by the same probability distribution (i.i.d.), which is the typical assumption of most of the machine learning methods. In this paper we use the Ridge Regression Conformal Predictor (RRCP), which, as its name suggests, is based on the well known Ridge Regression (RR) algorithm. We applied RRCP to three empirical cost datasets, namely the COCOMO, the Desharnais and the ISBSG, each including a set of different cost factors measured over a series of completed software projects in the past.

The rest of this paper is organised as follows: Section 2 presents a brief literature overview of data driven cost estimation and outlines similar machine learning attempts reported in literature. Section 3 presents the background theory behind RR and CP, while section 4 provides a description of the experiments and associated results produced using the aforementioned datasets. Finally, section 5 summarises the findings of the paper and suggests future research steps.

2 Brief Literature Overview

Recently, many machine learning approaches have been investigated in the field of software effort estimation, such as neural networks [20], neuro-fuzzy approaches [9], support vector regression [17], genetic algorithms [1] and rule induction algorithms [14] with quite satisfactory results. Machine learning techniques use empirical data from past projects to build a model that can then be employed to predict the effort of a new project. Although such data-driven cost estimation methods employed in empirical software engineering studies report quite encouraging results, all these techniques suffer from lack of uncertainty value consideration. Most existing methods for software effort estimation only provide a single value as their estimation for the effort of a given project, without any associated information about how "good" this estimation may be. Yet, the provision of an interval which will include the effort of a project at a given level of confidence would have been much more informative [8]. For this reason, recent studies such as [2, 8, 12, 13], employed different techniques for producing predictive intervals.

A relatively new development in the area of Machine Learning is the introduction of a novel technique, called Conformal Prediction [24], that can be used for complementing the predictions of traditional algorithms with provably valid measures of confidence. Therefore, this technique is ideal for dealing with the problem of confidence interval prediction for software cost estimation. What we call in this paper CP was first proposed in [6] and then greatly improved in [23]. In both [6] and [23] the base algorithm used was support vector machine. Slightly later CP was applied to other algorithms such as Ridge Regression [15] and k-nearest neighbours for both classification [21] and regression [19]. At the same time, in an effort to improve the computational efficiency of CPs, a modification of the original approach was developed, called Inductive Conformal Prediction [18]. Since then CP has been applied successfully to problems such as the early detection of ovarian cancer [7] and the classification of leukaemia subtypes [3].

3 Ridge Regression Conformal Predictor

We first briefly describe the Conformal Prediction (CP) framework and then focus on its application to the dual form Ridge Regression (RR) algorithm [22]. RR is an improvement of the classical Least Squares technique and its one of the most widely used regression algorithms. The main reason we chose the dual form RR method is that it uses kernel functions to allow the construction of non-linear regressions without having to carry out expensive computations in a high dimensional feature space.

We are interested in making a prediction for the label of an example x_l, based on a set of training examples $\{(x_1, y_1), \ldots, (x_{l-1}, y_{l-1})\}$; where each

$x_i \in \mathbb{R}^d$ is the vector of attributes for example i and $y_i \in \mathbb{R}$ is the label of that example. Our only assumption is that all (x_i, y_i), $i = 1, 2, \ldots$, are i.i.d. CP considers every possible label $\tilde{y}$ of the new example x_l and assigns a value α_i to each pair (x_i, y_i) of the extended set

$$\{(x_1, y_1), \ldots, (x_l, \tilde{y})\} \tag{1}$$

that indicates how strange, or non-conforming, that pair is for the rest of the examples in the same set. This value, called the *non-conformity score* of the pair (x_i, y_i), is calculated using a traditional machine learning algorithm, called the *underlying algorithm* of the corresponding CP. More specifically, the non-conformity score of a pair (x_i, y_i) is the degree of disagreement between the actual label y_i and the prediction $\hat{y}_i$ of the underlying algorithm, after being trained on (1). The function used for measuring this degree of disagreement is called the *non-conformity measure* of the CP.

The non-conformity score α_l is then compared to the non-conformity scores of all other examples to find out how unusual $(x_l, \tilde{y})$ is according to the non-conformity measure used. This is achieved with the function

$$p((x_1, y_1), \ldots, (x_l, \tilde{y})) = \frac{\#\{i = 1, \ldots, l : \alpha_i \geq \alpha_l\}}{l}, \tag{2}$$

the output of which is called the p-value of $\tilde{y}$, also denoted as $p(\tilde{y})$, since this is the only unknown value in (1). An important property of (2) is that $\forall \delta \in [0, 1]$ and for all probability distributions P on Z,

$$P^l\{((x_1, y_1), \ldots, (x_l, y_l)) : p(y_l) \leq \delta\} \leq \delta; \tag{3}$$

a proof can be found in [16]. This makes it a valid test of randomness with respect to the i.i.d. model. According to this property, if $p(\tilde{y})$ is under some very low threshold, say 0.05, this means that $\tilde{y}$ is highly unlikely as the probability of such an event is at most 5% if (1) is i.i.d. Assuming it were possible to calculate the p-value of every possible label following the above procedure, then we could exclude all labels with a p-value under some very low threshold, or *significance level*, δ and have at most δ chance of being wrong. Consequently, given a confidence level $1 - \delta$ a regression CP outputs the set

$$\{\tilde{y} : p(\tilde{y}) > \delta\}, \tag{4}$$

in other words the set of all labels that have a p-value greater than δ. Of course it is impossible to explicitly consider every possible label $\tilde{y} \in \mathbb{R}$, so the RRCP follows a different approach which allows it to compute (4) efficiently. This approach, proposed in [15], is described in the next few paragraphs.

To approximate a set of examples the well known Ridge Regression procedure recommends finding the w which minimizes the function

$$a\|w\|^2 + \sum_{i=1}^{l}(y_i - w \cdot x_i)^2, \tag{5}$$

where a is a positive constant, called the *ridge parameter*. Notice that RR includes Least Squares as a special case (by setting $a = 0$). The RR prediction $\hat{y}_t$ for an example x_t is then $\hat{y}_t = w \cdot x_t$.

According to [22], the dual form RR formula for predicting the label y_t of a new input vector x_t is

$$Y(K + aI)^{-1}k, \tag{6}$$

where $Y = (y_1, \ldots, y_l)$ is the vector consisting of the labels of the examples in the training set, K is the $l \times l$ matrix of dot products of the input vectors $x_1, \ldots, x_l$ of those examples,

$$K_{j,i} = \mathcal{K}(x_j, x_i), \quad j = 1, \ldots, l, \quad i = 1, \ldots, l, \tag{7}$$

k is the vector of dot products of x_t and the input vectors of the training examples,

$$k_i := \mathcal{K}(x_i, x_t), \quad i = 1, \ldots, l, \tag{8}$$

and $\mathcal{K}(x, x')$ is the kernel function, which returns the dot product of the vectors x and x' in some feature space.

The non-conformity measure used by RRCP is

$$\alpha_i = |y_i - \hat{y}_i|, \tag{9}$$

where $\hat{y}_i$ is the prediction of the RR procedure for x_i based on the examples in (1). Using (6) the vector of all non-conformity scores $(\alpha_1, \ldots, \alpha_l)$ of the examples in (1) can be written in matrix form as

$$|Y - Y(K + aI)^{-1}K| = |Y(I - (K + aI)^{-1}K)|. \tag{10}$$

Furthermore $Y = (y_1, \ldots, y_{l-1}, 0) + (0, \ldots, 0, \tilde{y})$ and so the vector of non-conformity scores can be represented as $|A + B\tilde{y}|$ where

$$A = (y_1, \ldots, y_{l-1}, 0)(I - (K + aI)^{-1}K) \tag{11}$$

and

$$B = (0, \ldots, 0, \tilde{y})(I - (K + aI)^{-1}K). \tag{12}$$

Notice that now each $\alpha_i = \alpha_i(\tilde{y})$ varies piecewise-linearly as we change $\tilde{y}$. Therefore the p-value $p(\tilde{y})$ (defined by (2)) corresponding to $\tilde{y}$ can only change at the points where $\alpha_i(\tilde{y}) - \alpha_l(\tilde{y})$ changes sign for some $i = 1, \ldots, l-1$. This means that instead of having to calculate the p-value of every possible $\tilde{y}$, we can calculate the set of points $\tilde{y}$ on the real line that have a p-value $p(\tilde{y})$ greater than the given significance level δ, leading to a feasible prediction algorithm. A detailed description of this algorithm is given in [15], while a more efficient version can be found in [24].

4 Experiments and Results

We applied the Ridge Regression CP algorithm to three popular software effort estimation datasets the COCOMO 81 (COCOMO) the Desharnais and the ISBSG. The COCOMO [4] dataset contains information about 63 software projects from different applications. Each project is described by 17 cost attributes. The second dataset, Desharnais [5], includes observations for more than 80 systems developed by a Canadian Software Development House at the end of 1980. The third dataset, ISBSG [10] was obtained from the International Software Benchmarking Standards Group (ISBSG, Repository Data Release 9) and contains an analysis of software project costs for a group of projects. The projects come from a broad cross section of industry and range in size, effort, platform, language and development technique data. The release of the dataset used contains 100 characteristics and 3024 project data grouped in categories. Of those only 16 attributes (cost factors) were considered in our experiments, in which the dependent variable was the Full-Cycle Work Effort. In order to select these 16 attributes we performed several steps to clean, homogenize and codify the dataset. Firstly, we omitted columns describing attributes that may not be measured early in the development life cycle (i.e. are not available until the project concludes). Secondly, we removed those columns that had more than 40% of the total number of records filled with blank or unknown values. Thirdly, we kept only "qualitative" projects assessed as A, or B by the ISBSG reviewers and consistently reporting unique and easy to interpret information. Furthermore, all the numerical columns were cleaned from null values (row filtering), while we defined new categories to describe different but logically similar sample values of categorical columns, thus merging and homogenizing similar pieces of information into new categories (e.g. Oracle v7, Oracle 8.0 into Oracle). Finally, in order to be used with RR, every categorical (including multi-valued) attribute was replaced with n binary attributes, one for each category indicating whether the attribute belongs to that category or not. The final ISBSG dataset after this processing contained 467 records.

Before conducting our experiments the numerical attributes of all datasets were transformed to their natural logarithm values and then normalised to a minimum value of 0 and a maximum value of 1 so that all attributes lie in the same range and thus have the same impact. The effort values supplied to the algorithm were also log transformed, but the outputs produced were transformed back to the original effort scale before being compared with the actual values. Therefore all results reported here are on the original scale of efforts.

Our experiments followed a 10 fold cross-validation process. Each data set was split randomly into 10 parts of almost equal size and our tests were repeated 10 times, each time using one of the 10 parts as the testing set and the remaining 9 as the training set. This procedure was repeated 100 times for each dataset and the results reported here are over all runs. The kernel

used for computing the matrix K in (7) and the vector k in (8) was the RBF kernel, which is defined as

$$\mathcal{K}(x_j, x_i) = exp(-\frac{\|x_j - x_i\|^2}{2\gamma^2}). \quad (13)$$

The parameter γ of the RBF kernel as well as the ridge parameter a in (6) are typically determined by trial and error. Thus for γ we tried the values 2 to 7 with increments of 0.5 and for a the values $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$.

We first applied the original RR approach to each dataset in order to check its performance on the data in question. The performance of the method was evaluated using the Mean Magnitude of Relative Error ($MMRE$), which is one of the most widely used metrics for evaluating the accuracy of cost estimation models. The Magnitude of Relative Error for a test project i was calculated as

$$MRE_i = \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \quad (14)$$

where y_i is the actual effort for project i and $\hat{y}_i$ is the predicted effort produced by (6) with x_i as the new input vector. The MRE of all projects in each of the 10 parts of every dataset was calculated using only the projects in the remaining 9 parts as training examples. This process resulted in one MRE value for each project, and since it was repeated 100 times, the $MMRE$ for each dataset was calculated as the mean of all $100N$ MREs where N is the number of projects in that dataset. Table 1 reports the best results obtained together with the γ value and the ridge parameter a that were used. The results of this table suggest that a considerable accuracy in predicting software effort values has been achieved by the base method, which is comparable to what the international literature has to offer using similar techniques up until today.

After the encouraging results obtained with the original Ridge Regression method we applied the RRCP to the three datasets. More specifically, RRCP was applied for the 95%, 90% and 80% confidence levels to every project in each of the 10 parts of the dataset, using as training set the projects in the remaining 9 parts. This process resulted in three predictive intervals for every project, one for each of the three confidence levels. Again the same process was repreated 100 times and thus resulted in $100N$ predictive intervals for

Dataset	γ	a	$MMRE$
COCOMO	5.5	0.001	0.4221
Desharnais	5	0.05	0.3454
ISBSG	3.5	0.1	0.5969

Table 1 The best results of the original Ridge Regression method and the parameters used.

Data Set	Median Width			Median Relative Width			Percentage of Errors (%)		
	80%	90%	95%	80%	90%	95%	80%	90%	95%
COCOMO	170.8	278.8	383.4	1.3593	2.2075	3.1105	19.11	8.68	3.68
Desharnais	4549	5579	6396	1.2588	1.5462	1.7698	19.03	9.06	4.38
ISBSG	4451	6571	9126	1.6821	2.4639	3.3747	20.01	10.02	5.01

Table 2 The tightness and reliability results of the Ridge Regression CP on the three data sets.

each confidence measure, where N is the number of projects in the dataset in question. The parameters used for these experiments are the same as the ones used for the original RR method, given in table 1.

Table 2 reports the results of the RRCP in terms of the tightness and reliability of the produced predictive intervals. The first part of the table reports the median widths of the intervals for each of the three confidence levels (95%, 90% and 80%). In addition to the width of each predictive interval, we also calculated its Relative Width (RW) as $RW_i = w_i/y_i$ where w_i is the width of the predictive interval produced for the example x_i and y_i is its actual effort value. The median values of the relative widths can be found in the second part of the table. We chose to report the median values of both the widths and the relative widths instead of the means so as to avoid the strong impact of a few extremely large or extremely small intervals. In the third and final part of table 2 we check the reliability of the obtained predictive intervals. This is performed by reporting the percentage of examples for which the true effort value is not inside the interval output by the RRCP. In effect this checks empirically the validity of the predictive intervals produced. The percentages reported here are either below or almost equal to the required significance level.

The enhancement of the original method with CP offers a much more informative scenery for a project manager to decide how to distribute effort as he/she is provided with lower and upper estimation limits instead of single effort prediction values, something that enables him/her to plan according to worst and best case scenarios. It is worth to note that the median widths of the predictive intervals reported in Table 2 for the 95% confidence level correspond to 3.36%, 27.34% and 6.09% of the whole range of efforts of the COCOMO, Desharnais and ISBSG datasets respectively. This provides a strong indication that the intervals produced are narrow enough to be useful in practice.

5 Conclusions

We have proposed the use of the Ridge Regression Conformal Predictor for obtaining reliable software effort confidence intervals. Confidence intervals

make an estimate much more informative than point predictions since they indicate the level of uncertainty of the estimate; the bigger the interval for a given project the higher the uncertainty of the estimate. This allows for different treatment of estimates, so that the uncertain estimates are given further thought or more careful planning. The main advantages of CPs over other techniques that produce confidence intervals are that (a) the confidence intervals produced by CPs are provably valid, (b) they do not require any extra assumptions other than i.i.d. and (c) they produce a different confidence interval for each individual project, the size of which reflects how difficult the effort of the project is to estimate.

The RRCP was applied on three empirical cost datasets each including a set of cost factors measured over a series of completed software projects. The results obtained by the proposed approach demonstrate that it can produce predictive intervals that are well-calibrated and narrow enough to be useful to project managers. Our plans for future work include utilising Conformal Prediction to produce confidence intervals for specific project cost descriptive variables, such as size, complexity and duration and by using historical data samples to attempt to associate these variables with effort. Furthermore, the approach followed in this paper may also be applied in conjunction with other algorithms reported in literature to perform well in predicting effort, such as neural networks and genetic algorithms. In this case the Ridge Regression part will be substituted by other predictive models to form the basis for Conformal Prediction. Finally, genetic algorithms can be utilized for the selection of the most appropriate subset of cost factors to be used as inputs to the CP so as to improve even further the results reported here.

References

1. Andreou, A., Papatheocharous, E.: Tools in Artificial Intelligence, chap. 1. Computational Intelligence in Software Cost Estimation: Evolving conditional sets of effort value ranges, pp. 1–20. I-Tech Education and Publication KG, Vienna, Austria (2008). URL http://intechweb.org/downloadpdf.php?id=5277
2. Angelis, L., Stamelos, I.: A simulation tool for efficient analogy based cost estimation. Empirical software engineering **5**, 35–68 (2000)
3. Bellotti, T., Luo, Z., Gammerman, A., Delft, F.W.V., Saha, V.: Qualified predictions for microarray and proteomics pattern diagnostics with confidence machines. International Journal of Neural Systems **15**(4), 247–258 (2005)
4. Boehm, B.: Software Engineering Economics. Prentice Hall (1981)
5. Desharnais, J.M.: Analyse statistique de la productivite des projects de development en informatique a partir de la technique de points de fonction. MSc. Thesis. Montreal Universite du Quebec (1988)
6. Gammerman, A., Vapnik, V., Vovk, V.: Learning by transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 148–156. Morgan Kaufmann, San Francisco, CA (1998)
7. Gammerman, A., Vovk, V., Burford, B., Nouretdinov, I., Luo, Z., Chervonenkis, A., Waterfield, M., Cramer, R., Tempst, P., Villanueva, J., Kabir, M.,

Camuzeaux, S., Timms, J., Menon, U., Jacobs, I.: Serum proteomic abnormality predating screen detection of ovarian cancer. The Computer Journal, doi:10.1093/comjnl/bxn021 (2008)
8. Gruschke, T., Jørgensen, M.: The role of outcome feedback in improving the uncertainty assessment of software development effort estimates. ACM Transactions of Software Engineering Methodology **17**, 1–35 (2008)
9. Huang, X., Ho, D., Ren, J., Capretz, L.: Improving the COCOMO model using a neuro-fuzzy approach. Applied Soft Computing **7**, 9–40 (2007)
10. International Software Benchmarking Standards Group: The ISBSG estimating, benchmarking & research suite release 9 (2005). URL http://www.isbsg.org/
11. Jørgensen, M., Shepperd, M.: A systematic review of software development cost estimation studies. IEEE Transactions on Software Engineering **33**(1), 33–53 (2007)
12. Jørgensen, M., Sjøberg, D.: An effort prediction interval approach based on the empirical distribution of previous estimation accuracy. Information Software and Technology **45**, 123–136 (2003)
13. Jørgensen, M., Teigen, K., Moløkke, K.: Better sure than safe? overconfidence in judgment based software development effort prediction intervals. Systems and Software **70**, 79–93 (2004)
14. Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., Webster, S.: An investigation of machine learning based prediction systems. Journal of Systems and Software **53**(1), 23–49 (2000)
15. Nouretdinov, I., Melluish, T., Vovk, V.: Ridge regression confidence machine. In: Proceedings of the 18th International Conference on Machine Learning (ICML'01), pp. 385–392. Morgan Kaufmann, San Francisco, CA (2001)
16. Nouretdinov, I., Vovk, V., Vyugin, M.V., Gammerman, A.: Pattern recognition and density estimation under the general i.i.d. assumption. In: Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory, *Lecture Notes in Computer Science*, vol. 2111, pp. 337–353. Springer (2001)
17. Oliveira, A.: Estimation of software projecs effort with support vector regression. Neurocomputing **69**(13-15), 1749–1753 (2006)
18. Papadopoulos, H.: Tools in Artificial Intelligence, chap. 18. Inductive Conformal Prediction: Theory and Application to Neural Networks, pp. 315–330. I-Tech, Vienna, Austria (2008). URL http://intechweb.org/downloadpdf.php?id=5294
19. Papadopoulos, H., Gammerman, A., Vovk, V.: Normalized nonconformity measures for regression conformal prediction. In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008), pp. 64–69. ACTA Press (2008)
20. Papatheocharous, E., Andreou, A.: Software cost estimation using artificial neural networks with inputs selection. In: Proceedings of the 9th International Conference on Enterprise Information Systems, pp. 398–407. Madeira, Funchal (2007)
21. Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive confidence machines for pattern recognition. In: Proceedings of the 13th European Conference on Machine Learning (ECML'02), *Lecture Notes in Computer Science*, vol. 2430, pp. 381–390. Springer (2002)
22. Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: Proceedings of the 15th International Conference on Machine Learning (ICML'98), pp. 515–521. Morgan Kaufmann, San Francisco, CA (1998)
23. Saunders, C., Gammerman, A., Vovk, V.: Transduction with confidence and credibility. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, vol. 2, pp. 722–726. Morgan Kaufmann, Los Altos, CA (1999)
24. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)

Bootstrap Confidence Intervals for Regression Error Characteristic Curves Evaluating the Prediction Error of Software Cost Estimation Models

Nikolaos Mittas, Lefteris Angelis

Department of Informatics, Aristotle University of Thessaloniki 54124, Thessaloniki GREECE, e-mail: *{nmittas, lef}@csd.auth.gr}*

Abstract

The importance of Software Cost Estimation at the early stages of the development life cycle is clearly portrayed by the utilization of several algorithmic and artificial intelligence models and methods, appeared so far in the literature. Despite the several comparison studies, there seems to be a discrepancy in choosing the best prediction technique between them. Additionally, the large variation of accuracy measures used in the comparison procedure constitutes an inhibitory factor which complicates the decision-making. In this paper, we further extend the utilization of Regression Error Characteristic analysis, a powerful visualization tool with interesting geometrical properties in order to obtain Confidence Intervals for the entire distribution of error functions. As there are certain limitations due to the small-sized and heavily skewed datasets and error functions, we utilize a simulation technique, namely the bootstrap method in order to evaluate the standard error and bias of the accuracy measures, whereas bootstrap confidence intervals are constructed for the Regression Error Characteristic curves. The tool can be applied to any cost estimation situation in order to study the behavior of comparative statistical or artificial intelligence methods and test the significance of difference between models.

1 Introduction

A crucial issue and an open problem which attracts the interest of researchers in software engineering is the ability to build accurate prediction models in order to estimate the cost of a forthcoming project. Due to this fact, a large amount of studies is towards this direction evaluating the performance of different *Software*

Cost Estimation (SCE) methods and models [1]. Although there is an obligation for a project manager to select the "best" prediction technique, there seems to be no global answer for all kinds of data. Furthermore, the wide variety of the proposed approaches that diversifies from expert judgment techniques to algorithmic and machine learning models renders the task of the selection extremely difficult. According to [2], the main reason for the contradictory results is the lack of standardization in software research methodology which leads to heterogeneous sampling, measurement and reporting techniques and the appropriateness of the prediction techniques on the available data.

The situation becomes much more complicated when we consider the divergent opinions about which accuracy measures are most appropriate in order to compare the predictions obtained by alternative models. Although a lot of accuracy indicators have been proposed in the literature and used in practice so far [3], there is a confusion of what different statistics really measure [4].

From all of the aforementioned, it is clear that the validation of prediction methods and the selection of the most appropriate model is a critical and non-trivial procedure. As we remarked in [5], a single error measure is just a statistic, i.e. a value computed from a sample (mean, median or percentage) and as such contains significant variability. Hence, when we compare models based solely on a single value we take the risk to consider as significant a difference which in fact may be not so significant. For these reasons, the determination of the "best" prediction technique has to be based on a more formal comparison procedure through inferential statistical approaches. On the other hand, in some circumstances, traditional methods might lead to erroneous inference when the dataset is considerably small and skewed or when the parametric assumptions do not hold. Thus, the utilization of resampling techniques is proposed for the selection of the best prediction technique.

In this paper, we extend our previous study [6] in which we presented the *Regression Error Characteristic* (REC) curves and the benefits from using them for the visual comparison of prediction models. Specifically, we propose a bootstrap method for the construction of *Confidence Intervals* (CIs) for REC curves, so as to test graphically the significance of the difference between two prediction techniques. The bootstrap method is the most appropriate, as the sample of errors is non-normally distributed, heavily skewed and usually of small size. By utilizing bootstrap, we illustrate how the selection of the best model can be accomplished with graphical means. Moreover, by providing bootstrap estimates, such as standard error and bias, we show how indicators of accuracy can be affected by the small software samples.

The rest of the paper is organized as follows: In Section 2, we present the bootstrap method. In Section 3, we describe the methodology followed for the construction of bootstrap REC curves. In Section 4, we present the experimental results obtained by the application of bootstrap REC curves on a real dataset. Finally, in Section 5, we conclude by discussing the results and by providing some directions for future research.

2 The bootstrap method

The comparison of prediction models is usually based on a validation procedure where various functions of errors are evaluated from the actual Y_A and the estimated Y_E cost. This results in a "point estimation" for the unknown accuracy i.e. a single value, which is computed from a particular sample coming from an practically infinite and unknown population.

Bootstrap is a simulation technique that can be used in order to extract and explore the sample distribution of a statistic [7]. We use here the most known version, the non-parametric bootstrap, which is based entirely on the empirical distribution of the dataset, without any assumption on the population. In general, the technique is to use a random sample $\mathbf{x} = (x_1, ..., x_n)$ from which we draw a large number (say *B*) of bootstrap samples by sampling with replacement in order to make statistical inference about an unknown population parameter θ (mean, median, percentage, etc.). The sample statistic $\hat{\theta}$ is a point estimator of the parameter θ (for details on the method see [5]).

The approximate distribution obtained by bootstrap can be used for computing the standard error, the bias and the CIs for the population parameter θ. In our case the random sample consists of the prediction errors obtained by a certain method. The goal is to utilize the bootstrap distributions in order to construct CI for REC curves and test whether a prediction technique provides better results than a comparative model for a certain accuracy estimator.

The simplest way to construct a $(1-\alpha)\times 100\%$ CI is the bootstrap empirical percentile method. First, from the empirical distribution containing all the θ^{*i} values ($i = 1,2,...,B$) obtained from the bootstrap samples, we compute the values $\theta^*_{a/2}$ and $\theta^*_{1-a/2}$ corresponding to the $100(a/2)$-th and the $100(1-a/2)$-th percentiles. Then, the bootstrap percentile CI is simply given by

$$[\theta^*_{a/2},\ \theta^*_{1-a/2}] \tag{1}$$

Two typical measures of accuracy for $\hat{\theta}$ is the *standard error* (SE) (Eq. 2) and the *bias* (Eq. 3) of the estimator that can be also estimated by the bootstrap samples by

$$SE_{boot} = \sqrt{\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{\mathrm{B}-1}} \quad \text{where} \quad \hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B}\hat{\theta}^*(b)}{B} \tag{2}$$

$$BIAS_{boot} = \hat{\theta}^*(\cdot) - \hat{\theta} \tag{3}$$

Suppose now that we wish to evaluate the prediction performance of a model (ModelA) on a specific SCE dataset. Suppose also that we obtain predictions using the well-known method of *jackknife* (or *hold-one-out)*, i.e. we estimate the cost of each one of the projects in the dataset using a model constructed by all the other projects. After applying the model on the dataset, we obtain by the jackknife method one sample of error expressions which are values of continuous variables. Based on these samples, we have to draw conclusions concerning their means, medians, percentages or certain percentiles of their distributions, so they can be utilized as the basis for the extraction of bootstrap replicates in order to evaluate the CI, SE and bias of the prediction model.

3 Bootstrapping Regression Error Characteristic curves

REC curves were introduced for comparison purposes of SCE models in [6], where it was pointed out that their utilization can be proved quite beneficial since they reinforce the knowledge of project managers obtained either by single accuracy indicators or by comparisons through formal statistical comparisons. Their most important feature is their ability to present easily accuracy results to non-experts and support the decision-making.

More precisely, a REC curve is a two-dimensional plot where the x-axis represents the error tolerance (i.e. all possible values of) of a predefined error measure and the y-axis represents the accuracy of a prediction model. *Accuracy* is defined as the percentage of projects that are predicted within the error tolerance e. An important feature is that REC curves are very informative since they take into account the whole error distribution, and not just a single statistic of the errors, providing information about extreme points, bias and other characteristics.

REC curves have interesting geometrical characteristics. The most significant one is that commonly used measures of the distribution such as the median or certain percentiles of errors can be estimated by exploiting the shape of a REC curve. In Fig. 1 (a), we see the REC curve of a hypothetical prediction model. The horizontal reference line from 0.5 intersects the REC curve in a point which corresponds to $e = 0.32$ (vertical reference line). This means that 50% of projects have an error smaller than 0.32 which is the median of errors. Similarly, we can evaluate other measures, as for example the well-known pred25.

Based on the bootstrap distributions of error functions, we can easily construct a 95% CI using the bootstrap empirical percentile method in order to draw conclusions regarding the predictive performance of a model. For example, in Fig. 1 (b), we can see the 95% CI of the hypothetical model for the entire distribution of errors. Suppose now that one wishes to know how confident should feel about the accuracy of the constructed model which results in a median error 0.32, that is to estimate a lower and upper bound for this median. Utilizing the bootstrap REC curves, the practitioner should be 95% confident that the unknown parameter for

the median population error varies within the interval [0.22, 0.43]. The same procedure can be followed in order to graphically compare alternative prediction models by constructing the REC 95% CI curves for each model. When the 95% CIs of models do not have an overlapping point, this means that there is a statistically significant different between the predictive performance of the two comparative models. Hence, REC curves provide an easily interpretable visualization technique for the complicated task of the selection of the "best" prediction model on a specific dataset.

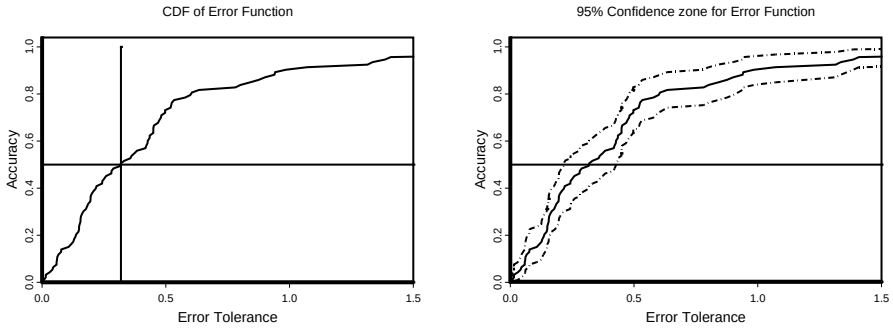


Fig. 1 (a) REC curve example for the evaluation of median error (b) 95% CI of error distribution

4 Experimentation

As the scope of this study is the investigation of bootstrap approach for the construction of REC 95% CIs, we have to choose the prediction techniques to deal with. There are two approaches that have attracted the research interest and have been extensively studied [2], namely the Regression Analysis and the Estimation by Analogy (EbA). The predictive accuracy of the models is usually based on two measures of "local" errors. More specifically, we use the *Magnitude of Relative Error* ($MRE = |\text{actual effort} - \text{estimated effort}| / \text{actual effort}$) and the *Absolute Error* ($AE = |\text{actual effort} - \text{estimated effort}|$) obtained by the jackknife validation of each model. These samples of errors can be utilized for the evaluation of the overall predictive accuracy of each model through well-known statistics (such as the mean and the median) [5]. Furthermore, the sample of errors constitutes the basis for the construction of REC curves and the extraction of bootstrap replicates of the proposed bootstrap method.

The dataset used for experimentation contains 63 projects from a commercial Finnish bank [8]. In order to fit the Regression model, we had to follow all the preliminary analysis for the dataset concerning the transformations and the concatenation of the variables [9]. A Stepwise Regression procedure was then applied to determine the variables having a significant impact on the response variable.

As EbA is free of assumptions, we used all the original variables for building the model, whereas the analogue projects were found through a special dissimilarity coefficient suggested by [10] that takes into account the mixed-type variables. The statistic for the combination of the efforts of neighbor projects was the arithmetic mean, whereas the number of analogies was decided by a calibration procedure, was three.

The REC curves for each of the local accuracy measures obtained by the two comparative models are presented in Fig. 2 (a) and (b). As the REC curves (MRE and AE) for the LS model are always above the corresponding REC curves of EbA, we can infer that LS dominates. Generally, a prediction model performs well if the REC curve climbs rapidly towards the upper left corner. REC curves can also identify extreme errors. When these outliers are present, the top of the REC curve is flat and does not reach 1 until the error tolerance is high. In our plots, we limit the range of the x-axis not to include the extremely high error values for better illustration of the figures. For example, in Figure 2, we can see that both the MRE and AE REC curves for EbA do not reach 1. This fact is a consequence of the presence of few projects producing extremely high values of errors.

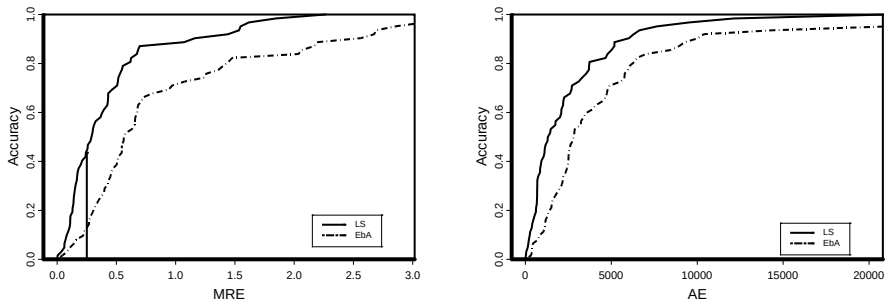


Fig. 2 (a) MRE and (b) AE REC curves for the comparative models

Method	LS		EbA	
Measure	MMRE (%)	MdMRE (%)	MMRE (%)	MdMRE (%)
Actual	45.37	29.38	99.41	56.98
$Estimate_{boot}$	45.41	29.47	98.73	58.88
SE_{boot}	6.09	5.26	12.70	5.79
$Bias_{boot}$	0.04	0.09	-0.68	1.90
95% CI	[34.46, 57.93]	[19.98, 42.56]	[75.65, 124.08]	[50.28, 67.95]
Measure	MAE	MdAE	MAE	MdAE
Actual	2624.16	1373.77	5410.47	2834.33
$Estimate_{boot}$	2624.73	1454.91	5424.04	2952.89
SE_{boot}	430.22	345.81	1006.20	444.48
$Bias_{boot}$	0.57	81.14	13.57	118.56
95% CI	[1908.07, 3558.92]	[922.69, 2150.84]	[3759.62, 7592.91]	[2462, 4193.67]

Table 1 Accuracy measures for the comparative models

The vertical line that crosses the x-axis at 0.25 can be used for the estimation of the pred25 accuracy measures that are based on the MREs of a model (Fig. 2a). More precisely, the pred25 is defined as the percentage of projects with $MRE \leq 0.25$. The aforementioned accuracy measure can be easily evaluated by getting the accuracy of a model at 0.25 error tolerance. It is clear that LS also dominates in terms of pred25 since its value is very close to 0.44 (or 44%), whereas the corresponding value for EbA model is not higher than 0.12 (or 12%).

The general results of the predictive accuracy of the two comparative models are presented in Table 1. It is obvious that LS outperforms EbA in terms of all the accuracy measures. Hence, the conclusions derived from the inspection of REC curves are verified by the accuracy measures that evaluate the prediction performance of the comparative models through certain statistics.

At this point from the bootstrap replicates of MREs (or AEs), we can construct a lower and upper bound (dash line) for each point of the REC 95% CIs (Fig. 3). Moreover, we also report the SE and bias evaluated through the bootstrap technique for each of the comparative models (Table 1). For example, we can observe that the mean (or $\text{Estimate}_{\text{boot}}$) for all the MMRE (Mean MRE) replicates is 45.41%, which is a value very close to the estimated MMRE through the jackknife procedure on the initial dataset and for this reason the bias can be considered low (0.04%). Another interesting issue arisen from the evaluation of the bootstrap accuracy measures is that MMRE and MAE (Mean AE) for EbA present extremely high values of SE compared to the corresponding estimates of the other indicators of error.

Although the REC 95% CIs are very informative, since we can assess a confidence zone for each percentile of the distribution of errors, we cannot draw conclusions for the predictive performance of the alternative models because they do not have one common basis for the comparison procedure. Indeed, the x-axis for the LS model (Fig. 3a) varies from 0 to 1.5, whereas EbA seems to have extremely higher values of error with the maximum value to be higher than 3 (Fig. 3b). This fact is also verified by the inspection of the initial REC curves (Fig. 2), where the LS model climbs more rapidly on the left corner of the graph meaning lower values of errors. The findings are also similar for the case of AEs (Fig. 3c and 3d).

In order to compare the overall predictive performance of the alternative models, we can use the Wilcoxon signed rank test, which constitutes a non-parametric procedure testing whether there is a significant difference between the medians of two paired samples. Alternatively, we propose the utilization of bootstrap REC curves for the identification of significant differences between the medians of the models. Having in mind that we wish to compare the medians of LS and EbA models, we can easily exploit the geometry of REC 95% CIs for 0.50 accuracy value.

As we can observe from Fig. 4a, the 95% CI for the MdMRE (Median MRE) of LS varies within the interval [19.98%, 42.56%], whereas for EbA (dash line) the corresponding interval diversifies within the interval [50.28%, 67.95%]. More importantly, it is obvious from the inspection of the geometry that the two CIs do not

have an overlapping point which means that there is a statistically significant difference between the alternative models. This is also the case regarding the comparison of MdAE (Median AE) (Fig. 4b). More specifically, the 95% CI for LS constructed through the bootstrap replicates of MdAE varies within the interval [922.69, 2150.84], whereas for EbA model within the interval [2462, 4193.67], indicating a statistical significant difference. In order to verify the effectiveness of bootstrap REC curves to graphically detect the differences between the comparative models, we also make use of the Wilcoxon sign rank test for matched pairs. All pair-wise tests have p-values smaller than 0.05 revealing that the differences observed in Fig. 4 are in fact statistically significant.

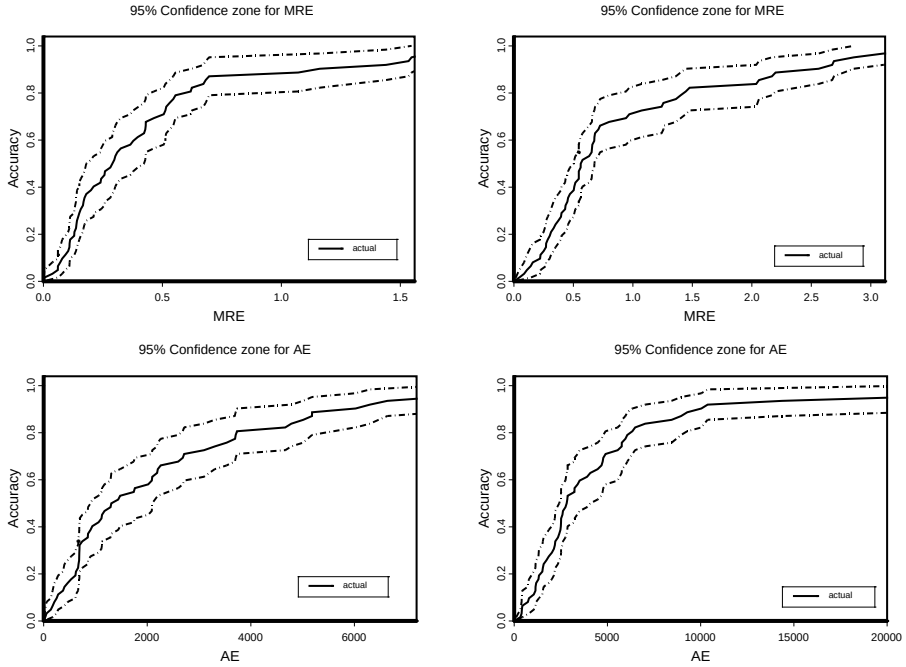


Fig. 3 (a-c) LS and (b-d) EbA REC 95% CI curves for the comparative models

The bootstrap REC curves can also be utilized for the construction of CIs for the pred25 accuracy measure, whereas a hypothesis test can also be conducted for the comparison purposes. Contrary to the MdAE, the pred25 CIs are evaluated by drawing a reference vertical line from 0.25 and from the intersecting points of the REC curve's lower and upper bound, a horizontal line to meet the accuracy axis. Hence, the graphical tool for performing statistical tests for the pred-measures that are essentially percentages and have not been considered yet in formal comparisons, constitutes an easily interpretable manner to assess the predictive power of different models. In Fig. 5, we can notice that the 95% CI for LS varies within the

interval [30.65%, 56.45%] and does not present an overlapping point compared to the EbA model that diversifies within the interval [4.84%, 19.35%], so there is a statistically significant difference between the alternative models regarding the pred25 accuracy measure.

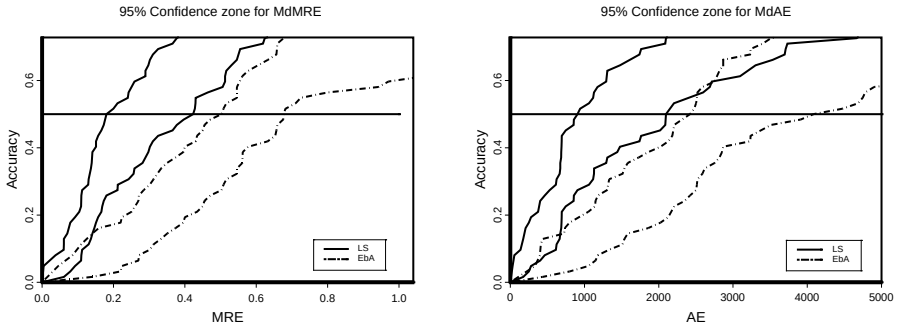


Fig. 4 (a) MdMRE and (b) MdAE comparison for the comparative models

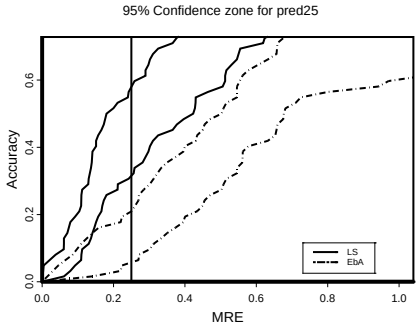


Fig. 5 pred25 comparison for the comparative models

5 Conclusions

In this paper, we deal with the critical task of the selection of the "best" model for a specific Software Cost Estimation dataset with completed projects. More specifically, we extend the utilization of Regression Error Characteristic curves that constitutes an easily interpretable tool, by the construction of bootstrap Confidence Intervals for different error functions.

As the plethora of comparative studies concerning the selection of the "best" model reveals contradictory results, the goal of this paper is to further extend the research on this area. Our intention is not to determine the superiority of either Regression analysis or Estimation by Analogy methods, but rather to facilitate the project managers with a visualization tool that contributes to the systematic com-

parisons of any kind of prediction methods either statistical or artificial intelligence. Moreover, the most important problem that a practitioner has to be faced with is the small-sized and heavily skewed samples of projects and the unavailability of new data in Software Cost Estimation area. This limitation can be resolved by the utilization of bootstrap resampling techniques.

Summarizing our findings, we can conclude that the REC curves for all the expressions of error we studied show for this specific dataset that LS outperforms EbA and is the most plausible choice for predicting the effort of a forth-coming project. The most important here is that the conclusions obtained by a simple visual comparison through REC curves constructed by the jackknife samples of errors. In addition, the most essential advantage provided by this study, is that we enhance the comparison procedure through the construction of bootstrap CIs for the entire distributions of error functions. In this way, a practitioner is able to assess the benefits for each of the comparative models through the examination of certain percentiles of errors. Furthermore, we also provide a graphical tool to test the statistical significance of the differences between the comparative models for common accuracy measures, like MRE, pred25 and AE through geometrical characteristics and properties of the bootstrap REC curves. Finally, as shown in our experiments, the statistical tests comparing the samples of errors, confirm the visual results, in the sense that each time the difference between two prediction error samples is significant, this is clearly shown by the bootstrap 95% CIs of REC curves.

References

[1] Jorgensen, M., Shepperd, M.J. (2007). A systematic review of software development cost estimation studies. *IEEE Transactions on Software Engineering*, 33(1), 33-53.

[2] Mair, C.M., Shepperd, M.J. (2005). The consistency of empirical comparisons of regression and analogy-based software project cost prediction. *IEEE Proceedings International Symposium on Empirical Software Engineering, (ISESE)*, 509-518.

[3] Foss, T., Stensrud, E., Kitchenham, B., Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. *IEEE Transactions on Software Engineering*, 29(11), 985-995.

[4] Kitchenham, B.A., Pickard, L., MacDonell, S., Shepperd, M. (2001). What accuracy statistics really measure. *IEE Proceedings-Software*, 148(3), 81-85.

[5] Mittas, N., Angelis, L. (2008). Comparing cost prediction models by resampling techniques. *Journal of Systems and Software*. 81(5), 616-632.

[6] Mittas, N., Angelis, L. (2008). Comparing software cost prediction models by a visualization tool. *Proceedings of the IEEE 34th Euromicro Conference on Software Engineering and Advanced Applications*, 433-440.

[7] Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

[8] Maxwell, K. (2002) *Applied statistics for software managers*. Prentice-Hall, PTR.

[9] Sentas, P., Angelis, L., Stamelos, I., Bleris, G. (2005). Software productivity and effort prediction with ordinal regression. *Information and Software Technology*, 47, 17-29.

[10] Kaufman L, Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. New York, John Wiley.

Approaching Software Cost Estimation Using an Entropy-Based Fuzzy *k*-Modes Clustering Algorithm

Efi Papatheocharous and Andreas S. Andreou

University of Cyprus, Department of Computer Science,

75 Kallipoleos str., CY1678 Nicosia, Cyprus

efi.papatheocharous@cs.ucy.ac.cy, aandreou@cs.ucy.ac.cy

Abstract

A new software cost estimation approach is proposed in this paper, which attempts to cluster empirical, non-homogenous project data samples via an entropy-based fuzzy *k*-modes clustering algorithm. The target is to identify groups of projects sharing similar characteristics in terms of cost attributes or descriptors, and utilise this grouping information to provide estimations of the effort needed for a new project that is classified in a certain group. The effort estimates produced address the uncertainty and fuzziness of the clustering process by yielding interval predictions based on the mean and standard deviation of the samples having strong membership within a cluster. Empirical validation of the proposed methodology was conducted using a filtered version of the ISBSG dataset and yielded encouraging results both in terms of practical usage of the clustered groups and of approximating effectively project costs.

1. Introduction

Software cost estimation involves the process to foresee the total costs spent during the development of a software product based on several factors, called 'cost drivers', and mostly relate with the product to be developed, the engineering process followed and the people engaged in the process. During the last few decades the main cost driver attracting most of the research interest is development effort (typically measured in person-months) [11]. Various attempts have been made over the years to model the correlation between cost drivers essentially utilising project size and duration, and the actual and predicted effort for a project, without a comprehensive solution, as development effort estimation is also affected by project-specific factors, which cannot be easily in-

cluded in a cost model [1]. The generation of a cost model usually faces serious difficulties due to lack of accurate definitions of the factors involved [9], the nature of the collected data, the non-deterministic range of possible values for the categorical data and the rich number of missing values observed for many projects [1].

The aim of this work is to approximate the issue of software effort estimation by performing a certain type of homogenization and clustering on historical project cost samples. To this end, the ISBSG R9 dataset [8] was employed, which contains an adequate number of past project recordings, but at the same time it suffers from non-homogeneity in terms of recording methods used, counting approaches and interpretation of key project characteristics. Our goal is to identify clusters of similar projects that are sufficiently close to each other so as to use their descriptive characteristics (i.e. cost attributes) for classifying a new project in a certain cluster. The latter is performed according to how close the new project is to the centre of the cluster using a similarity distance. Once this is done the attempt focuses on exploiting the transformation of the effort of the projects participating in the cluster for providing an estimate for the new project. Clustering in our case is performed by a simple and quite promising algorithm, namely the Entropy-Based Fuzzy *k*-Modes Clustering Algorithm, while the transformation of the participating effort sample in a cluster follows a nearest-neighbors approach. The closely-related clustered projects are utilised to provide effort prediction intervals of minimum width, related to the mean and standard deviation values of the respective effort. Our results thus far suggest that the proposed approach may be considered successful enough as it is able to provide estimations with an accuracy of around 77% on average, while homogenization of data via clustering seems to lead to significantly improved estimations compared to using the dataset as it is.

The rest of this paper is organised as follows: Section 2 presents a brief literature overview, while section 3 outlines the algorithm used to extract the project clusters. Section 4 presents the experimental process followed and provides a description of the dataset, along with some preprocessing activities performed. Section 5 discusses the experimental results obtained and finally, section 6, summarises the findings of the paper and suggests future research steps.

2. Literature Overview

Estimation methods reported in the software engineering literature of the last 30 years may be classified into the following categories: Expert judgment, Algorithmic and Machine Learning. The latest developments in Machine Learning techniques mostly combine concepts and notions from the area of Soft Computing to form cost estimators or predictors, while a large part of ongoing research concerns data-driven techniques. Data-driven cost estimation is a widely used class of estimation techniques that rely on past project data values

related to factors affecting costs that are combined in some way attempting to estimate the actual effort level. Researchers suggest that data-driven techniques applied in conjunction with a multiple set of techniques on different subsets of data may produce a range of estimated values instead of crisp values and reduce the inaccuracy degree involved in the estimation [10, 13]. Consequently, the notion of prediction interval, as reported in [4], is a minimum-maximum range of values for the effort estimates, attached with a confidence level with which the actual value of the effort is included in the range.

Analogy-based estimation is a widely adopted method in software cost estimation that identifies analogous projects to the one under estimation and uses their data to derive an estimate [10]. The similarity measures between pairs of projects are critical for identifying the most appropriate historical data from which the estimation will be generated. Usually the similarity measures are selected empirically using jackknife-like procedures. Typically, the measures that identify the most similar projects in the majority of the cases are considered as the appropriate ones to use and are applied in every new estimation procedure. However, there are situations where default similarity measures may not be the most appropriate ones.

Clustering in general seeks to organise data samples into several subsets by employing a variety of techniques. There are several types of clustering methods, and in particular for software cost estimation fuzzy clustering techniques were examined yielding better figures of adjustment than their crisp equivalents [2]. Various tools and models have been developed proposing that data mining and computational intelligent techniques may be utilised to assist automatic clustering algorithms in finding distinct subsets of highly related concepts in a more efficient manner. In this study, we aim to combine such notions from data-driven, analogy and fuzzy clustering techniques, to deal with the lack of homogeneity present in historical data and introduce improved cost estimates lying within ranges of values. In addition, this study aims to investigate the effect of a set of contributing factors to effort (including numerical and categorical in nature) for clustering, while the proposed approach is utilised to determine suitable groups of software projects for building effort estimation models. In [7] the authors emphasize the importance of establishing homogeneity of the data in an effort estimation model and investigate the effect of clustering in the ISBSG repository. The empirical experiments conducted showed that the estimation accuracy obtained using clustered data is not significantly different compared to that of the ordinary least squares method or using the original data without clustering.

3. Entropy-based Fuzzy *k*-modes Clustering Algorithm

Entropy-based clustering [14] essentially groups similar data samples into clusters based on their entropy values. The goal is to determine the number of clus-

ters present in the set and identify their centres by traversing the dataset only once. Data samples with many surrounding samples have total entropy values lower than the rest and may be considered as candidates for representing their clusters. A new cluster is initially formed with the sample defined as the cluster centre and then is allocated data samples that have a similarity value higher than parameter β which represents the similarity threshold [14]. The *k*-modes algorithm was introduced in [6] and was extended to include fuzzy elements to account for uncertainty data samples [5], where the dissimilarity function is altered to a simple matching of the attributes describing the samples in the dataset and thus is not based on the Euclidean distance. In addition, in the fuzzy version of the algorithm the cluster centres are defined by the modal value of each attribute instead of the mean value and their computation relies on the assignment of the most frequent category of each attribute as the representative of the cluster.

Let $X_1 = [x_{11}, x_{12},\ldots, x_{1m}]$ and $X_2 = [x_{21}, x_{22},\ldots,x_{2m}]$ be two data samples of a dataset described by *m* attributes. The dissimilarity between the two samples, $d(X_1, X_2)$, is given by:

$$d(X_1, X_2) = \sum_{j=1}^{m} \delta(x_{1j}, x_{2j}) \tag{1}$$

where:

$$\delta(x_{1j}, x_{2j}) = \begin{cases} 0, & x_{1j} = x_{2j} \\ 1, & x_{1j} \neq x_{2j} \end{cases} \tag{2}$$

The dissimilarity function in equation (1) is then used to (re)assign a data sample to a cluster. Accordingly, in the case of the hard *k*-modes algorithm, if object X_i yields the shortest distance with centre Z_l in a given iteration, this is represented by setting the value at the nearest cluster to 1 and the values at the rest of the clusters to 0 in the partition matrix *W*. Formally, for $\alpha = 1$:

$$\hat{w}_{li} = \begin{cases} 1, & \text{if } d(Z_l, X_i) \leq d(Z_h, X_i), \quad 1 \leq h \leq k \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In the case of the fuzzy *k*-modes algorithm, for $\alpha > 1$, the partition matrix *W* is given by:

$$\hat{w}_{li} = \begin{cases} 1, & \text{if } X_i = Z_l \\ 0, & \text{if } X_i = Z_h,\ h \neq l \\ \dfrac{1}{\displaystyle\sum_{h=1}^{k} \left[\frac{d(Z_l, X_i)}{d(Z_h, X_i)} \right]^{1/(\alpha-1)}}, & \text{if } X_i \neq Z_l \text{ and } X_i \neq Z_h,\ 1 \leq h \leq k \end{cases} \tag{4}$$

for $1 \leq l \leq k$, $1 \leq i \leq n$. This means that if a data sample has exactly the same attribute values with a particular cluster centre, then it will be assigned fully to that cluster and not at all to the rest. Otherwise, the data sample will be

characterised by a membership degree for each cluster denoting its partial membership in the cluster [12].

4. Experimental Approach

This section describes the proposed methodology for effort estimation which involves the following five-steps: (i) data preparation, (ii) entropy-based clustering, (iii) fuzzy *k*-modes clustering, (iv) selection of groups of suitable projects, and (v) investigation of effort prediction within the retrieved software projects. The selection of the main attributes to experiment with was based on a step-wise attempt to clean, homogenize and obtain a satisfactory portion of the ISBSG dataset described by both numerical and categorical attributes. A hold-out sample technique was used in each experiment repetition with 75% of the project samples used for performing the clustering of the data (i.e. training set) and 25% being utilised during the evaluation (i.e. testing set).

4.1. Dataset description, cleaning and fuzzification

The dataset utilised in the experiments is obtained from the International Software Benchmarking Standards Group (ISBSG R9) [8]. This dataset contains an analysis of multi-organisational, multi-application domain and multi-environment software project cost data. The initial release of the dataset used contains 100 characteristics and 3024 project data grouped in categories describing data quality, project size, effort, productivity, schedule, software quality, architecture, documents and techniques, project and product attributes.

The dataset is rich in samples but may be considered biased and fairly heterogeneous, having many inconsistent or null project values. To alleviate this problem a large part of data was removed, especially in cases where the data reported was considered irrelevant to cost prediction, or where the values or technique used to gather or report the values were found inadequate according to directions issued by the ISBSG. Secondly, the dataset went through a series of preprocessing steps for selecting attributes according to some data pruning principles which led to a clean, consistent, categorical dataset, as all numerical attributes underwent a fuzzy transformation [15] to host linguistic values.

The fuzzy transformation of the numerical attributes was performed by determining the degree to which they belong to each of the appropriate fuzzy sets via membership functions [15]. For each numerical cost attribute variables m_i, n_i, a_i and b_i were calculated ($1 \leq i \leq n$, and n is the number of linguistic terms in the classification table being analyzed) according to equations (5)-(8) and after following the fuzzification illustrated in Figure 1 [3].

$$m_i = \min \textit{value of linguistic term } T_i \textit{ in classification table} \quad (5)$$

$$n_i = \frac{m_i + m_{i+1}}{2} \quad (6)$$

$$a_i = n_{i-1} \quad (7)$$

$$b_i = m_{i+1} \quad (8)$$

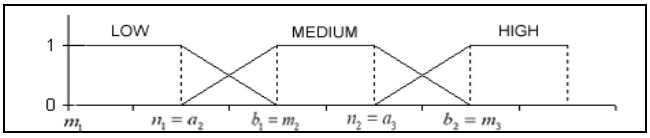


Fig. 1. Fuzzification values of numerical attributes

The filtered ISBSG dataset that was finally utilised in the experiments comprised 50 columns and 424 rows with the following project attributes: Count Approach, Adjusted Function Points, Project Elapsed Time, Implementation Year (a new column extracted from the column previously named Implementation Date in the original ISBSG dataset), Development Type, Organization Type, Development Technique, Functional Sizing Technique, Development Platform, Language Type, Primary Programming Language, Database System, Recording Method, Resource Level, Max Team Size and Average Team Size. The dependent variable was the Full-Cycle Work Effort which was also a newly formed column from the original ISBSG dataset, containing only the summary work effort values accounted for all the development phases and not adjusted to include parts of the effort values from phases that were not measured.

4.2. Fuzzy k-modes clustering and effort estimation

Clustering was performed using the training set of data samples and the Entropy-based algorithm that defined the cluster centres as suggested by [12] and was briefly described in the previous sections. The algorithm computes the number of clusters k with their respective initial cluster centres that will be used by the fuzzy k-modes algorithm. Finally, the fuzziness exponent α defines the level of fuzziness that will be adopted by the clustering process.

At the validation step, we aimed to isolate smaller areas within each cluster which conform to the new project in question. The attributes of the new project are matched against the final cluster centres produced as a result of the previous clustering procedure isolating the nearest centre using the partition matrix W as explained before. Thus, we retrieve the most similar projects from the repository by calculating the membership degree of each project in the cluster for which its centre is closer to the new project. Then, a cut-off limit is used to reduce the selected set of projects that should respond to a similarity measure, called φ applied for the surrounding projects, which represents a value for level of confidence. The cut-off limit is constructed by defining an upper and lower

bound based on the value of the new project's membership degree to the closest cluster centre. For experimentation purposes, this was set to ±10% meaning that, for instance, if a new project was assigned a membership degree of 60% in the search cluster, then the projects retrieved would have membership degrees between 50%-70%. The confidence level, which ensures that only the closest projects falling between the +10% and -10% radius distance from the new project are selected, was set to the minimum values of 75% and 85% similarity threshold degree (φ) respectively.

The final step of the methodology is to relate the derived fuzzy clusters of specific degrees with effort predictions and overall assess the areas selected within the clusters produced. The membership and similarity parameters mentioned above essentially filter out the dissimilar and irrelevant projects to the new one; the mean effort value and standard deviation of the actual effort values of the projects kept is then computed. The predicted effort value of the new project is estimated to lie within the range [*mean effort value (mean)* ± *standard deviation (std)*]. One exception to the aforementioned range is the case where the standard deviation is greater than the mean, in which we take the lower bound of the interval to be equal to zero. In general, our aim is to offer bounded estimation intervals of the minimum possible width, rather than single point value predictions, yielding more general estimates on one hand, but of a more informative nature on the other, and somehow with encapsulation of the inherent estimation uncertainty. Additionally, we attempt to assess the relative accuracy of estimation intervals by using the validation set of data mentioned earlier (testing) and measure the percentage of the projects in this set that have their estimated effort values lying within the range [*mean−std*, *mean+std*]. We call this the Hit Ratio (*HR*) of the corresponding estimation process and we report it in the results section that follows. Additionally, we try to evaluate the reliability of our approach by comparing in percentage terms the interval size calculated (reported as width) with the Overall Size (*OS*) and the Cluster Size (*CS*) computed using the actual minimum and maximum effort values contained in the overall training and the clustered samples respectively. These two supplementary metrics essentially measure how much shrinking of the effort estimation interval the method has achieved relatively to the "worst" case, which is the width of the initial available set of projects and to the intermediate stage where projects are filtered via clustering and therefore their range of values to use for estimations is narrowed. This *OS* and *CS* metrics assist in evaluating how good our estimation intervals really are.

5. Experiments and Results

The results of applying the entropy-based and fuzzy *k*-modes algorithm on the preprocessed data, as previously described, are presented in this section. Firstly, we experimented with the entropy algorithm to locate the cluster cen-

tres (k) and subsequently we implemented hard clustering. Secondly, we applied the fuzzy k-modes algorithm, produced the fuzzy clustering results and studied the influence of the dataset to the parameters. Experiments were carried out with variations of the ISBSG dataset as follows: Experimental dataset *EDS1* included all available project characteristics plus the effort; all project characteristics excluding effort constituted *EDS2*; removing the outliers from *EDS1* and *EDS2* based on the box plots of the effort sample values resulted datasets *EDS3* and *EDS4* respectively; finally, using *EDS3* and adjusting the weight of the effort variable to reach the dominant significance level of 51% in the clustering process compared to the rest of the attributes, produced dataset *EDS5*. Similarity parameter β and fuzzy exponent α were varied, taking values from the sets {0.3, 0.4, 0.5, 0.55, 0.6, 0.7, 0.8, 0.9} and {1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8} respectively. Table 1 summarizes the best results obtained with respect to the width of the estimation (or prediction) interval and the hit ratio.

Table 1. Results obtained with the fuzzy k-modes algorithm using various experimental datasets (EDS)

EDS	β	α	k	φ	OS (%)	CS (%)	HR (%)	mean effort	std effort	width
1	0.55	1.5	42	0.75	7.80	49.09	38.68	12727.77	5843.59	11687.19
2	0.7	1.2	95	0.75	10.16	49.39	50.94	7885.80	7614.47	15228.94
3	0.4	1.4	6	0.85	2.26	67.68	36.17	1711.81	1695.34	3390.68
4	0.3	1.8	3	0.85	2.39	67.45	28.72	1931.39	1788.68	3577.36
5	0.8	1.7	25	0.75	1.60	37.92	76.60	2030.93	1198.76	2397.53

The results reported indicate relatively large prediction intervals in most of the cases, except *EDS1* and *EDS5,* with standard deviations being lower than the means in all cases. Moreover, a mediocre hit ratio performance is observed, which amounts to approximately 30-40% hits for *EDS1, EDS3* and *EDS4* and slightly over 50% hits for the *EDS2*. The accuracy of the predicted effort values is significantly improved in the *EDS5* case; the hit ratio is quite high suggesting that estimations produced lay within the calculated width in nearly 77% of the cases. It is worth noticing that when the effort attribute participates in a dataset performance is improved (cases *EDS1* and *EDS3* in comparison with *EDS2* and *EDS4* respectively). This outcome suggests that the effect of previous values for the attribute being estimated leads to forming better clusters. One may argue that the participation of effort samples in the clustering process may bias results, but this is not true; past effort values are treated by the algorithm as descriptors of the behavior of effort in relation with the rest of the participating factors. Hence, what effort samples offer is essentially a way to map cost factors onto the effort attribute and form knowledge about how effort evolves. Additionally, the narrower widths obtained with *EDS3* and *EDS4* confirm that when extreme values are removed from the datasets the estimation performance is again improved. Overall, *EDS5* yielded the most promising results and was thus further analyzed through additional experimentation reported in Table 2.

The results reported in Table 2 indicate that the *std* of the effort values is consistently lower than the mean value reported, while the prediction intervals yielded are quite low and thus more useful to project managers. Furthermore, independently to the fuzzy *k*-modes parameters tested, the overall diversity of the results throughout the dataset is small. Recalling that *EDS5* involves all available attributes including the effort variable, the latter acting both as a filter for outliers and at the same time playing a decisive role in the clustering process as the most significant attribute, it is obvious and to a large extent logical, that practically this variable improves clustering results in terms of projects homogenization. The dataset indicates a significantly better picture with relatively narrower widths, while the best results achieved a spread of approximately 2398 man-hours (mh), with a mean effort value of 2031mh and a corresponding standard deviation of 1199mh. The *HR* degree in relation to both *OS* and *CS* degrees reported suggest that clustering data in small segments has been achieved: The derived interval in the best case is 17% of the initial and 38% of the clustered one. Another observation worthy of mentioning is that the best results consistently suggest k=25 as the "optimal" number of clusters, while parameters $\beta$and αassume the values of 0.8 and 1.7 respectively.

Table 2. Further results obtained after experimentation with *EDS5* (φ=0.75)

β	α	k	OS (%)	CS (%)	HR(%)	mean effort	std effort	width
0.8	1.7	25	1,60	37.92	76.60	2030.93	1198.76	2397.53
0.8	1.5	25	1,88	39.69	76.60	2294.23	1406.38	2812.76
0.8	1.8	25	1,92	38.74	45.74	2206.57	1433.75	2867.49
0.9	1.7	104	1,93	50.20	62.77	2625.31	1440.06	2880.12
0.9	1.5	104	1,94	77.48	62.77	2647.77	1452.07	2904.14

At this point we should mention that we attempted to compare our findings with the results of a simple *k*-nearest-neighbors (*k*-nn) algorithm. Preliminary *k*-nn results exhibited larger intervals (widths), which may be considered inferior to those of our approach, with better *HR* values as expected. Due to space limitations, though, these results will not be presented here.

The basic assumption under investigation in the present paper was that homogenizing samples in distinct clusters that share common values for certain cost factors contributes to achieving successful effort estimations. The results above lead us to infer that this assumption is partly supported; one has to be cautious, though, as regards generalization of this argument as this was not the case for all datasets used, at least to the extent to which small estimation intervals were produced. This, of course, may be the result of a number of causes which should be further investigated as part of our future work, examining the effect each cause may have on clustering, and hence the associated effort estimation processes. For example, one possible cause may be the fact that resemblance of a project with a cluster centre used to assign the former as a member of that

cluster with degree *r* is measured only in terms of how many factors are identical, not which exact factors are matched. Thus, this should be further analyzed and assessed so as to contribute to improving the estimation process.

6. Conclusions

A new methodology has been presented in this paper which attempted to improve the means for selecting clusters of project data from a large repository to address the problem of software cost estimation. Specifically, the proposed methodology employed entropy-based and fuzzy *k*-modes clustering to suggest an innovative project clustering for the ISBSG R9 repository and obtain effort estimation (prediction) intervals for new projects based on the similarity of cost attributes. The methodology identifies clusters of similar projects and then classifies a new project in a certain cluster according to its resemblance with the cluster centre. Projects in this cluster which are closely-related within a specified degree of resemblance to the new project are isolated and then their effort values are utilised to provide an estimation interval for the effort of the new project.

Our ultimate goal was to apply an already successful clustering algorithm and reduce the heterogeneous nature of our data repository, something which was performed successfully. The clustering of the projects in homogeneous groups according to their specific characteristics may be considered a small novel step forward in the area of software cost estimation where the attribute space is multi-dimensional. Even though it would be extremely useful to exploit such information provided by the clusters formed and achieve improved effort predictions, as targeted by this paper, we may not claim that the results obtained are optimal. After performing and evaluating a preliminary set of experiments conducted it became evident that there is ample room to improve the results of the algorithm possibly using better encoding and parameter set-up. Finally, as regards the clustered projects achieved by the method, they could be proven more valuable in estimating effort if they were utilised by other techniques and be employed as an intermediate input to other cost models performing point estimations. Examples of such techniques that could possibly work better when provided with clustered data rather than the original ones are regression, inductive learners, decision trees etc. Thus, such approximations could capture more efficiently correlations among various parameters of the project other than effort, such as productivity, schedule, team size etc. Our future research plans will address the above and consider examining how processed and clustered datasets may be studied in a homogeneous setting allowing dependencies between cost factors to be brought to light. To this end, hybrid forms of cost models may be employed, having the clustering module as the feeding platform of the input values satisfying certain cost attribute characteristics and a cost model for refining the estimation intervals by applying further processing either in a data-

driven, quantitative form (e.g. prediction with artificial neural networks), or in a qualitative manner (e.g. fuzzy cognitive maps or influence diagrams).

References

[1] Angelis, L., Stamelos, I., and Morisio, M.: Building A Software Cost Estimation Model Based On Categorical Data. Proceedings of the 7th International Symposium on Software Metrics, IEEE Computer Society, p. 4 (2001)

[2] Aroba, J., Cuadrado-Gallego, J.J., Sicilia, M., Ramos, I. and García-Barriocanal, E.: Segmented software cost estimation models based on fuzzy clustering. Journal of Systems and Software, Vol. 81, pp. 1944-1950 (2008)

[3] Braz, M.R., Vergilio, S.R: Using Fuzzy Theory for Effort Estimation of Object-Oriented Software. In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. pp. 196-201 (2004)

[4] Gruschke, T.M., Jørgensen, M.: The role of outcome feedback in improving the uncertainty assessment of software development effort estimates. In ACM Transactions of Software Engineering Methodology. Vol. 17, pp. 1-35 (2008)

[5] Huang, Z. and Ng, M.K.: A Fuzzy k-Modes Algorithm for Clustering Categorical Data. IEEE Transactions on Fuzzy Systems, Vol. 7, No. 4, pp. 446-452 (1999)

[6] Huang, Z.: Extensions to the k-Means Algorithm for Clustering Large Datasets with Categorical Values. Data Mining and Knowledge Discovery, Vol. 2, No. 3, pp. 283-304 (1998)

[7] Huang, S.-J., Chiu, N.-H and Liu, Y.-J: A comparative evaluation on the accuracies of software effort estimates from clustered data. Information of Software Technology 50, pp.879-888 (2008)

[8] International Software Benchmarking Standards Group (ISBSG), Estimating, Benchmarking & Research Suite Release 9, ISBSG, Victoria, (2005) http://www.isbsg.org/

[9] Kitchenham, B., Hughes, R., Linkman, S.: Modeling software measurement data. IEEE Transactions on Software Engineering 27 (9), 788–803 (2001)

[10] MacDonell, S.G., Shepperd, M.J.: Combining Techniques to Optimize Effort Predictions in Software Project Management. Journal of Systems and Software, Vol. 66, No. 2, Elsevier, Amsterdam, pp. 91-98 (2003)

[11] Sommerville, I.: Software Engineering. Addison-Wesley Longman Publishing Co., Inc. (2007)

[12] Tsekouras, G.E., Papageorgiou, D., Kotsiantis, S., Kalloniatis, C. and Pintelas, P.: Fuzzy Clustering of Categorical Attributes and its Use in Analyzing Cultural Data. International Journal of Computing Intelligence (WASET), Vol. 1, No. 2, pp. 123-127, ISSN 1304-2386 (2005)

[13] Xu, Z., Khoshgoftaar, T. M.: Identification of Fuzzy Models of Software Cost Estimation. Fuzzy Sets and Systems, Vol. 145, No. 1, Elsevier, pp.141-163 (2004)

[14] Yao, J., Dash, M., Tan, S.T., Liu, H.: Entropy-based Fuzzy Clustering and Fuzzy Modeling. Fuzzy Sets and Systems, Vol. 113, No. 3, pp. 381-388 (2000)

[15] Zadeh, L.A.: Fuzzy Set. Information and Control, Vol. 8, pp. 338-353 (1965)

AIAEP Introduction

We are glad to introduce the IFIP AIAI 2009 Workshop on *Artificial Intelligence Applications in Environmental Protection* (AIAEP'2009). This is the first edition of this scientific forum under the IFIP framework.

The evolution of the climatic changes from the recent years has raised new challenges for the researchers in the environmental sciences that has to find new solutions to the environmental problems. Artificial intelligence can provide efficient tools for most environmental problems of monitoring, analysis, interpretation, forecasting, management, and control. The development of interdisciplinary research groups between specialists from environmental sciences and artificial intelligence can lead to new ideas and innovative applications in the area of environmental protection systems. We hope that AIAEP'2009 will be a good opportunity for the researchers to exchange ideas and to initiate new collaborations in this interdiciplinary area, as both, Artificial intelligence and Environmental protection are domains of strategic interest.

For this first edition, we received 12 papers and after a reviewing process, 9 papers were selected based on the judgement of two referees from the organising committee. The technical issues addressed by the selected papers for this workshop come from all environmental fields: water resources management, environmental accounting, wastewater treatment, maritime pollution, oil spill pollution, soil erosion risk assessment, air pollution dispersion assessment, detection and classification of waste on irregular terrains, residential fire detection, biological foaming in anaerobic digestion simulation and estimation of permeability of granular soil.

The artificial intelligence tools and techniques that were applied range from data mining, abductive reasoning, decision support systems, knowledge-based systems, expert systems, planning, autonomous

inspection robots, semantic techniques, fuzzy logic to artificial neural networks and neuro-fuzzy systems.

Acknowledgements

We would like to express our acknowledgements to many people and organisations: IFIP AIAI 2009 Tecnical and Workshop Programme Committees, IFIP, the organizers of the AIAI 2009, Prof. I. Vlahavas, Prof. I. Manolopoulos, and Assist. Prof. L. Iliadis.

The co-chairs like to express their thanks to the Programme Committee and the referees for their invaluable help and advice during the reviewing process. It has been a pleasure to work with them. Also, we like to thank to all the authors for their effort on writing the papers and for sharing their research work of applying Artificial intelligence in Environmental protection. Without their works the AIAEP'2009 workshop could not have been possible.

Mihaela Oprea[1] and Nick Bassiliades[2]

[1]Department of Informatics
University Petroleum-Gas of Ploiesti
Romania.
mihaela@upg-ploiesti.ro

[2]Department of Informatics
Aristotle University of Thessaloniki
Greece
nbassili@csd.auth.gr

Organization Chairs

- Mihaela Oprea, *University Petroleum-Gas of Ploiesti, Department of Informatics, Romania*

- Nick Bassiliades, *Aristotle University of Thessaloniki, Department of Informatics, Greece*

Programme Committee

- Nick Bassiliades, *Aristotle University of Thessaloniki, Department of Informatics, Thessaloniki, Greece* (nbassili@csd.auth.gr)

- Vladan Devedzic, *Department of Information Systems and Technologies, University of Belgrade, Serbia and Montenegro* (devedzic@sbb.rs)

- Ioannis Douros, *Laboratory of Heat Transfer and Environmental Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece* (jdouros@aix.meng.auth.gr)

- Karina Gibert, *Statistics and Operations Research Department, Technical University of Catalonia, Barcelona. Catalonia, Spain* (karina.gibert@upc.edu)

- Tony Jakeman, *The Fenner School of Environment and Society, The Australian National University, Canberra, Australia* (tony.jakeman@anu.edu.au)

- Mihaela Oprea, *Department of Informatics, University Petroleum-Gas of Ploiesti, Ploiesti, Romania* (m_oprea@yahoo.com)

- Ignasi Rodríguez-Roda, *Laboratory of Chemical and Environmental Engineering (LEQUIA). University of Girona, Girona. Catalonia, Spain* (ignasi@lequia.udg.es)

- Miquel Sànchez-Marrè, *Knowledge Engineering & Machine Learning Group, Technical University of Catalonia, Barcelona, Catalonia, Spain* (miquel@lsi.upc.edu)

- Athanassios Tsadiras, *Department of Information Technology, Alexander Technology Educational Institute (ATEI) of Thessaloniki, Greece* (tsadiras@it.teithe.gr)

- Grigorios Tsoumakas, *Department of Informatics, Aristotle University of Thessaloniki, Greece* (greg@csd.auth.gr)

- Franz Wotawa, *Institute for Software Technology, Graz University of Technology, Austria* (wotawa@ist.TUGraz.at)

Artificial Intelligence Applications in the Atmospheric Environment: Status and Future Trends

Kostas D. Karatzas

Informatics Applications and Systems Group, Dept. of Mechanical Engineering, Aristotle University of Thessaloniki, Greece.

Extended Abstract

The problem of assessing, managing and forecasting air pollution (AP) has been in the top of the environmental agenda for decades, and contemporary urban life has made this problem more intense and severe in terms of quality of life degradation. A number of computational methods have been employed in an effort to model and simulate air quality (AQ). Air pollution is related to various substances, is affected by physical and chemical mechanisms of various spatial and temporal scales, and is regulated in terms of target values that are different to each other. Thus, AP requires for computational and knowledge management tools that are able to deal with its complex (and exiting from the scientific point of view) nature. Moreover, such methods should be able to deal with missing observation data, data of mixed nature (be it nominal, categorical, binary or other), and imitate the behavior and the "intelligence" of the phenomena that need to be modeled and simulated. This means that deterministic modeling, employing fluid mechanics, atmospheric chemistry and physics (the "traditional way for modeling AQ") are not able to "catch" all the aspects of the AP problem. Other methods should be employed, that are able to deal with knowledge extraction and management, and are able to map knowledge into the "intelligence" of the algorithms that they apply. On this basis, Artificial Intelligence should be used. This is a thesis recognized already from the 90ties, where the first sets of scientific publications in areas like neural networks and fussy logic have appeared with applications in AQ.

In the first era of AI applications to AP problems, a variety of methods was tested, with the aim to investigate existing observation data (available in the form of time series), and forecast parameters of interest (i.e. concentrations). Such methods included Artificial Neural Networks, Decision Trees, and Fuzzy Logic algorithms (to name some of the most representative ones). Then, the effort was also to extract knowledge and thus reveal interrelationships between parameters of interest (i.e. meteorological dependencies of concentrations, relations between pollutants,

etc). On this basis, methods like principal component analysis were also applied. In a step to further investigate relationships, understand mechanisms of pollution and forecast the behavior of pollutants, additional (supervised and unsupervised) algorithms of AI were employed. These included methods like Self Organizing Maps and Clustering. In parallel, there was an effort to formerly represent the knowledge that is being made available (or is being extracted) from the AQ field. This is where AI communicated to software engineering methods like Ontologies and Semantic Web languages.

Future trends involve the automation of the processes of knowledge extraction and representation, in order to deal with data being available sporadically and dynamically, as in the case of a mobile user. In addition, contemporary, personalized electronic information services, call for methods that may support environment related decision on the basis of "negotiations" made on the fly (by applying AI methods like agents). Moreover, research work is undertaken on the methods applied for feature selection in order to "feed" the AI algorithms. This selection is now based on the usage of various methods that screen all available (or the most promising) feature combinations, with the aid of genetic algorithms, information gain criteria and multiple regression schemes (to name just some of the existing or emerging methods in this area). In this presentation, issues related to the past, present and future of AI applications in the AQ field are going to be addressed.

Hydrological Neural Modeling aided by Support Vector Machines

L. S. Iliadis [1], S.I. Spartalis[2]

[1] *Department of Forestry & Management of the Environment & Natural Resources liliadis@fmenr.duth.gr*
[2] *Department of Production Engineering & Management, School of Engineering sspart@pme.duth.gr Democritus University of Thrace, Greece*

Abstract: This paper aims in the construction of an Artificial Neural Network model that performs successful estimation of the Maximum Water Supply (m^3 /sec) and of the Special Water Flow (m^3/sec*Km^2) using several topographic, meteorological and morphometric parameters as independent variables. Support Vector Machines applying specific Kernel algorithms [9] are used to determine the Error or Loss of the Neural Network model and to enhance its ability to generalize. Data come from the Greek island of Thasos, which is located in the North-Eastern part of the Aegean sea. As a matter of fact, this manuscript can be considered as a specific case study, but its modeling design principles and its Error minimization methods can be applied in a wide range of research studies and applications

Keywords: Artificial Neural Networks, Support Vector Machines, Error Minimization, Water Resources management

1. Introduction

Effective water resources management is one of the most crucial environmental challenges of our times. Extreme flood incidents happen more and more often all over the world and they cause immense damages in infrastructure and in human lives.

This research effort has two distinct orientations of equal importance. First it deals with the development of Artificial Neural Network (ANN) modeling of both average annual maximum water supply (m^3/sec) (AAMWS) and special water flow (SWF) (m^3/sec*Km^2) for the torrential watersheds of a wide mountainous area. It also deals with the supportive role of *Soft Computing* methods like e-Regression Support Vector Machines towards the evaluation of neural network models. From this point of view this specific contribution can be considered as having an innovative role towards ANN development and evaluation.

Another key aspect is the fact that it is performed by using structural data (remaining unchanged over-time) and only few actual real time measurements. Under this perspective this modeling effort can be considered cost and time effective and it can provide invaluable assistance towards the design of flood protection and prevention policy. Existing methods like the one of Gavrilovic [6] [23] concentrate mostly in the load of sediments and secondly in the influence of various structural

and dynamic factors. The equation of Gavrilovic is very useful and widely used but it cannot respond to sudden changes in the morphometric background of a torrential stream. For example if the percentage of forest cover is suddenly reduced dramatically due to a major forest fire, the Gavrilovic equation absorbs this change during the first years and it requires a longer period of time to show a significant difference.

1.1. Literature Review

Estimation of both AAMWS and SWF is a very important process, as it can serve for the rough calculation of existing water supplies and for the evaluation of the potential Torrential Risk for each mountainous watershed.

ANN have been used effectively in various research projects concerning water protection and water management in Europe, in USA and in many other countries. An ANN using Fuzzy Logic has been developed in the Netherlands for the control of water levels in polder areas [14]. Another ANN that performs river flow forecasting has been developed also in the Netherlands [3]. Also in USA (US Department of Energy) and in Europe, ANN have been developed for the prediction of stream-water quality [1],[15]or towards water-management in general. Wastewater and water management and protection models using ANN have also been developed recently [7],[20]. UNESCO has also funded several research projects using ANN and fuzzy logic for urban water management and flood risk evaluation [25][10]. Finally neural networks have been used for the prediction and forecasting of water resources variables [15].

2. The Environmental Problem of floods

2.1. Data and research area

Actual data corresponding to the eight input parameters described above were gathered from the *twenty* most important mountainous streams of the Thasos island. The *average annual rain height* corresponds to a period of *thirty* years. Thus, thirty data records were structured for each torrential stream, raising the total number of existing data vectors to *six hundred*. Except of the rain height and forest cover, the rest of the input parameters are structural and they do not change significantly overtime. A small sample of the available data that were used in this research is presented in the following table 1.

In the case of water resources modeling the input data are points in a space R^n and the output represents points in the well-known plain R^k where k =2 in this case, since the output parameters are two.

According to [18] there are eight main parameters that influence the AAMWS and the SWF and for which there are available data in the Greek repositories. So the input space can be considered as vectors of the R^8 space. The eight input data are the *average altitude*, the *average slope*, the *average annual rain height*, the percentage of *forest cover*, the percentage of *compact geological forms*, the *area*, the total *number of pipes* and the *total length of the pipes* of the torrential watersheds of Thasos island.

Table 1: A small sample of the selected data records

Torrential Stream	*Average Watershed Altitude*	*Average Watershed Slope*	*Average watershed annual rain-height*	*Percentage of forest cover of watershed*	*Percentage of watershed compact geological forms*
	m	*%*	*Mm*	*%*	*%*
Thasos	433.21	35.61	1059.40	81.47	87.81
Panagia	261.14	28.50	919.03	56.68	78.97
Potamia	574.74	49.93	1151.40	88.83	84.89
Skala Potamias	535.44	53.31	1126.10	56.14	76.55
Mavrou Lakou	378.19	50.84	1023.20	88.08	85.15
Kleisidiou	330.64	35.67	975.11	36.49	100.02
Thimonias	325.41	35.10	971.88	93.74	96.83

The two output parameters of the model are the AAMWS and the corresponding SWF which is estimated [13] by the following formula 1 where F is the area of the mountainous watershed in Km^2 and a is a parameter determined by water management specialists. The parameter a takes values in the closed interval [0.6,1]. SWF is a normalized version (per km^2) of the mean water supply (MWS).

$$q_{max} = a \frac{32}{0.5 + \sqrt{F}} \qquad (1)$$

3. Theoretical Background

3.1. ANN

Modern ANNs are rooted in many disciplines, like neurosciences, mathematics, statistics, physics and engineering. They find many successful applications in such diverse fields as modeling, time series analysis, pattern recognition and signal processing, due to their ability to learn from input data with or without a teacher.
Their computing power is achieved through their massively parallel distributed structure and their ability to learn and therefore generalize [8]. Generalization refers to their ability to produce reasonable output for inputs not encountered during the training process [8]. ANN consist of units called neurons whose computing ability is typically restricted to a rule for combining input signals and an activation rule that takes the combined input to calculate the output signal [2].
The proper design of ANN requires the development of various topologies using numerous optimization algorithms and transfer functions, before determining the optimal one. We have experimented thoroughly towards this direction. Sixty percent of the actual data records corresponding to the eight independent parameters,

have been used as input in the *training* phase and the other 40% were used as input in the *testing* phase of the ANN in order to evaluate the generalization ability.

3.2. Support Vector Machines

The *support vector machines* (SVM) or optimal margin classifiers are new type of learning algorithms based on statistical learning theory proposed by Vapnick [27]. SVM are used not only for classification but for regression (functional approximation) problems as well [4],[28]. When used for classification, a SVM algorithm creates a *Hyperplane* that separates the data into two classes. Given training examples labeled either "yes" or "no", a maximum-margin hyperplane is identified which splits the "yes" from the "no" training examples, such that the distance between the hyperplane and the closest examples (the margin) is maximized. It is a fact that SVM can handle multiple continuous and categorical variables. To construct an optimal hyperplane, SVM employees an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, SVM models can be classified into various distinct groups:

- Classification SVM Type 1 (also known as C-SVM classification)
- Classification SVM Type 2 (also known as nu-SVM classification)
- Regression SVM (also known as epsilon-SVM regression)

In this project, an *epsilon-regression* SVM (ERSVM) was applied. In an ERSVM one has to estimate the functional dependence of the dependent variable y on a set of independent variables x. Here an **ε-tube** is constructed that determines the loss degree of the regression. It assumes, like other regression problems, that the relationship between the independent and dependent variables is given by a deterministic function f plus the addition of some additive noise: $y = f(x) + \text{noise}$ (2)

The task is then to find a functional form for **f** that can correctly predict new cases that the SVM has not been presented with before. This can be achieved by training the SVM model on a sample set. In our case the ANN plays the role of the predicting function. In fact the learning machine is given $\boldsymbol{l}$ training data from which it attempts to learn the input-output relationship (which may have the form of dependency, mapping or function) f(x). In this case the training data set has the form $D = \{(X_i, Y_i) \in R^n \times R \mid i \in \{1,\ldots,l\}\}$ which means that it contains one pair of values (X_i, Y_i). It should be specified that the inputs X are n-dimensional Vectors that belong to R^n and the output Y of the system are continuous values [12]. The SVM uses approximating functions of the form $f(x, w) = \sum_{i=1}^{N} w_i \phi_i(x)$ (3) where the functions $\phi_i(x)$ are called features [12]. In regression problems like the one we are facing in this study, typically some measures of error approximation are used.

3.3. Applications of RSVM to ANN evaluation

Let's suppose that we have an ANN with an input vector **X**, a bias weight vector **b**, a *hidden layer* weights matrix **V**, and an output weight vector $\mathbf{w}^T$. More specifically: $X=[X_1,X_2,X_3,\ldots\ldots,X_n]^T$ (4) $b=[b_1,b_2,b_3,\ldots\ldots,b_j]^T$ (5) $w=[w_1,w_2,w_3,\ldots\ldots,w_j,w_{j+1}]^T$ (6)

$$V=\begin{bmatrix} u_{11} & \dots & u_{1i} & \dots & u_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ u_{j1} & \dots & u_{ji} & \dots & u_{jn} \\ \dots & \dots & \dots & \dots & \dots \\ u_{J1} & \dots & u_{Ji} & \dots & u_{Jn} \end{bmatrix} \quad (7)$$

The output layer neurons may have linear activation functions (for regression problems) or *sigmoid* for classification or pattern recognition tasks. The approximation scheme for such an ANN is shown in formula 8 [12]:

$$o(X,V,w,b) = F(X,V,w,b) = \sum_{j=1}^{J} w_j \sigma_j (v_j^T x + b_j) \quad (8)$$

The construction of a SVM algorithm for regression is actually a problem of minimizing the empirical risk R_{emp}^{e} and $\|W\|^2$ at the same time. Actually the problem is the estimation of a linear regression hyperplane f(x,w) =W^Tx+b. This can be achieved by minimizing the quantity R given by formula 9 [12]:

$$R = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{I} |y_i| - f(x_i, w)_e\right) \quad (9)$$

Here Vapnick's ε-insensitivity loss function replaces squared error and C ~1/ε.

In fact Vapnick introduced a general type of error (loss) function the linear loss function with ε-insensitivity zone which is given by the following formula 10 [12].

$$|y - f(x,w)|_\varepsilon = \begin{cases} 0 \dots if |y - f(x,w)| \le \varepsilon \\ |y - f(x,w)| - \varepsilon \dots otherwise \end{cases} \quad (10)$$

According to this algorithm the loss equals to 0 if the difference between the predicted by the ANN value f(x,w) and the measured value is less than ε. In this way function 10 defines a ε-tube. In the cases with the ANN predicted value within the tube, the loss (or error) is zero. In all other cases the loss equals to the magnitude of the difference between the ANN output value and the radius ε of the tube. If ξ and ξ^* are given by the following formulas 11 and 12, then formula 13 presents the final value of risk R that has to be minimized [27],[12].

The following equations 11 and 12 $|y - f(x,w)| - \varepsilon = \xi$ (11) $|y - f(x,w)| - \varepsilon = \xi^*$ (12) are valid for all of the data that are located above the tube.

For all of the data that are located below the tube equation 13 is true $R_{w,\xi,\xi^*} = \left[\frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{I}\xi + \sum_{i=1}^{I}\xi^*\right)\right]$ (13)

In the above formula 13 the following constrains should stand.

$y_i - w^T x_i - b \le \varepsilon + \xi$ (14) $w^T x_i + b - y_i \le \varepsilon + \xi^*$ (14)

$\xi \geq 0 \;\; i = 1, I$ (15) $\xi^* \geq 0 \;\; i = 1, I$ (16)

The following figure 1 clearly shows the case of an ε-tube and the parameters involved.

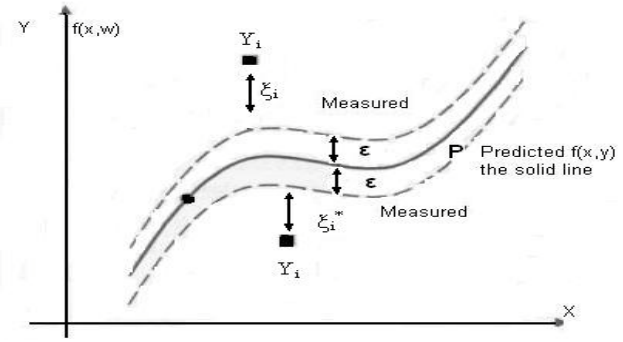


Figure 1: The parameters used in one dimensional ERSVM

In the previous formulae the ξ and ξ* are slack variables for measurements "*above*" and "*below*" an ε-tube. Both slack variables are positive values. Finally it should be clarified that the constant C which influences a trade-off between an approximation error and the weights vector norm $\| W \|$ is a design parameter whose value is chosen by the user. Also the norm $\| W \| = \sqrt{w_1^2 + \ldots + w_n^2}$ where W is the weight vector. An increase in C penalizes larger errors (large ξ and ξ*) and so it leads to a decrease in approximation error [12].

4. Hydrological Neural Model

Training is the process of ANN development, where the connection weights are adapted or modified in response to stimuli being presented at the input buffer and optionally to the output buffer. The *hidden layer* is the place where the data is being processed and it may consist of one or two sub-layers.

The *feed forward* network structure with input, output and HSL varying from 1 to 2, applying various optimization algorithms were tried in the training phase. We have experimented with the *back propagation* (BP) optimization algorithm (introduced by Rummelhart et al.) [21],[22][24][26], the *general regression* ANN architecture, the *modular ANN* topology [5] and the *radial basis function (RBF)* model. Our experimentations included the employment of various *transfer functions*.

M*odular ANN*s were proposed by Jacobs, Jordan, Nowlan and Hinton [19]. They consist of a group of BP networks referred to as "*local experts*" competing to learn different aspects of a problem. A "*gating network*" controls the competition and learns to assign different parts of the data space to different networks. When only one ANN is appropriate for a given problem, the *gating network* tends to favor it [19]. RBF are ANN having an internal representation of hidden neurons which are "*radially symmetric*". For a neuron to be *radially symmetric* it needs to have the following three constituents: a) A center which is a vector in the input space that is typically stored in the weight vector from the input layer to the pattern unit. b) A

distance measure to determine how far an input vector stands from the center. In this case standard Euclidean distance is used. c) A transfer function (using a single variable) which determines the output of the Neuron by mapping the output of the distance function.

The "*root mean square error"* (RMS Error) and the Confusion Matrix (CM) were used to check the validity of the ANN. The CM is roughly a kin to a scatter diagram, with the x-axis representing the desired output and the y-axis representing the actual output. The main difference from a scatter diagram is that the CM breaks the diagram into a grid. Each grid square is called a bin. Each output from the probe point produces a count within one of the bins [19]. For example if the probe point produces an output of 0.6 and the desired output was 0.5 then the bin around the intersection of 0.6 from the y-axis and 0.5 from the x-axis receives a count. These counts are displayed by a bar within the bin and the bar grows as counts are received. The bin receiving the most counts is shown at full height, while all of the other bins are scaled in relation to it [19]. The network with the optimal configuration must have the bins (the cells in each matrix) on the diagonal from the lower left to the upper right. Also more sophisticated Soft Computing tools were also applied to examine and determine the validity of the optimal ANN. An important aspect of the CM is that the value of the vertical axis of the produced histogram is the *common mean correlation* (CMC) *coefficient* of the desired (d) and the predicted output (y) across the Epoch. The CMC is calculated by the following formula (2):

$$CMC = \frac{\sum \left(d_i - \bar{d}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(d_i - \bar{d}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}, \qquad \bar{d} = \frac{1}{E}\sum_{1}^{n} d_i \qquad \text{and} \qquad \bar{y} = \frac{1}{E}\sum_{1}^{n} y_i$$

(17)

It should be clarified that i ranges from 1 to n (the number of cases in the data training set) and E is the Epoch size which is the number of sets of training data presented to the ANN learning cycles between weight updates.

5. Experimental Results

The first and the most important thing that had to be determined was the topology of the ANN and its optimization and transfer functions. Thus various experimentations were performed using various numbers of hidden sub-layers with various numbers of neurons used in each sub-layer.

The six hundred data records were divided in two subsets, the training and the testing one. The division was performed in a totally random manner. The training set included the 60% of the data and the testing the rest 40%. Training has shown that the optimal Artificial Neural Network (OANN) was the one that used BP optimization algorithm with the "*tangent hyperbolic"* activation function (mapping into the range –1.0 to 1.0) and with the "*extended delta bar delta"* (ExtDBD) training rule [11], [17] as the transfer function. It consisted of eight neurons in the input layer, only one hidden sub-layer with nine neurons and two neurons in the output layer. The RMS Error in the training phase was 0.0045 and the R^2=0.9997. It should be clarified here that the ExtDBD is a *heuristic* technique that has been

successful in a number of application areas and it uses termed momentum. A term is added to the standard weight change, which is proportional to the previous weight change. In this way good general trends are reinforced and oscillations are damped [19]. The random number seed was kept constant before each training round and the learning coefficient ratio was kept at 1. The following figure 2, shows the architecture of the OANN.

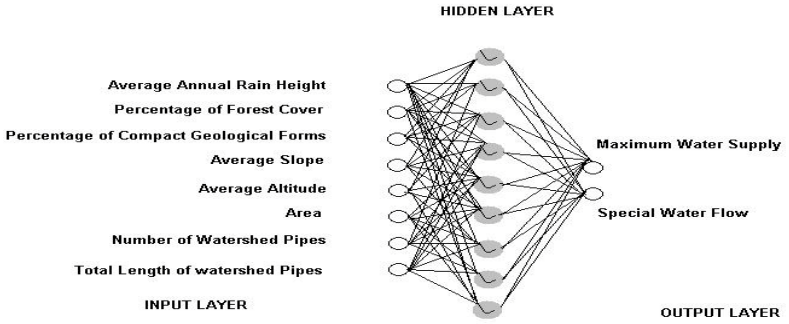


Figure 2: Architecture of the optimal ANN

Every time an ANN model is developed we need to ensure that the data are compatible to the learning algorithms and to the transfer functions applied. When a value coming into a processing element (PE) is beyond the PEs transfer function range, that PE is said to be saturated and of course the function produces dummy values [19]. The tangent hyperbolic function that is used in this specific research is mapping into the range –1.0 to 1.0 but it accepts values only between -3 and +3. Saturation occurs when a PE's summation value (the sum of the inputs times the weights) exceeds this range. So the input data were normalized by dividing them properly (by 1000) so that they will not exceed the acceptable range. This provided a good performance of the transfer function.

After the determination of the ANN structure and topology, the testing process was performed to prove the potential ability of the ANN to generalize with first time seen data. The results of the testing phase are the following. The RMS Error was 0.1404 and the R^2 =0.9763. Two confusion matrices were developed for the two output parameters. Both confusion matrices had all of their cells located very close to the diagonal from the lower left to the upper right. The CMC in the case of the *maximum water flow* was equal to 0.9654 whereas in the case of the Special Water flow it was equal to 0.9872. The following figure 3 shows the two confusion matrices corresponding to the two output parameters. It is clear that the ANN shows a good performance with unseen data and also it retains a good and simple structure which is also very important.

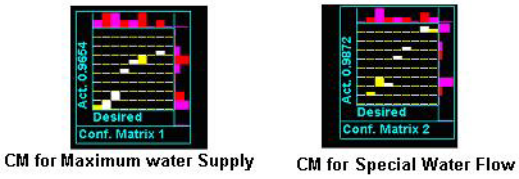


Figure 3: The Confusion Matrices in the Testing Phase

A complicated ANN with many hidden sub-layers or with a large number of hidden neurons does not guarantee good level of generalization capacity. The specific ANN that was developed here has a simple architecture and this is an important benefit. Though the generalization ability of the ANN has been demonstrated an ERSVM kernel was used to determine the risk of the ANN model.

5.1. Using the SVM to determine the ANN with the minimum Loss

The output of the ANN (with first time seen data) and the actual experimental values (coming from direct measurements in the most important torrential streams of the island) were used as input to the RSVM. Also the weights of the output Layer of the BP ANN were input to the SVM to determine the actual risk according to formula 13, [27],[29].

In both cases the actual measured values were subtracted from the ANN output ones in order to estimate the maximum value. For the case of AAMWS ε was defined to be equal to 4.5 almost equal to the maximum difference between the predicted and the measured values. This value of defines the width of a specific tube. Based on formulae 11 and 12 and on the values of the weight vector *W* of the developed ANN, $\sum_{i=1}^{l} \xi = 4.6608$ whereas $\sum_{i=1}^{l} \xi^* = 4.924$ where constant C was defined to have a low value of. 0.1.

Based on these values and using formula 13, the final estimation for the value of risk R for AAMWS was equal to 0.204.

The following estimations were performed for the SWF. According to formulae 11 and 12 and using the values of the weight vector *W* of the developed ANN, the $\sum_{i=1}^{l} \xi = 3.289$ whereas $\sum_{i=1}^{l} \xi^* = 0.935$ where C=0.1 and ε was defined to be equal to 1.4 almost equal to the maximum difference between the predicted and the measured values.

Finally the risk R for the Special Water Flow was estimated using formula 13 to have a value of 0.1998. In both of the cases the risk was quite low showing that the ANN prediction values include quite low risk. This means that the degree of loss is quite low for the ANN predicted values, compared to the actual ones.

Table 2: The BP ANN output compared to actual measurements in Testing

Special Water Flow			Average Annual Maximum Water Supply		
Actual Experimental Measurements y	ANN Output f(X,w) for SWF	Tube width value ε for SWF	Actual Experimental Measurements y	ANN Output f(X,w) for AAMWS	Tube width value ε for AAMWS
4.97	3.754	1.4	180.94	196.01	4.5
11.07	11.562		81.69	80.84	
5.11	4.11		176.36	191.628	
9.11	10.588		98.84	94.279	
4.64	3.355		190.77	205.605	
10.45	11.745		86.44	83.275	
14.28	13.206		63.28	66.413	
7	6.349		128.9	140.593	
11.27	11.117		80.07	80.146	
6.59	5.458		136.54	154.573	

6. Conclusion

The developed OANN for the case of Thasos island, has proven its ability to estimate the AAMWS and the SWF successfully and reliably and most of all its ability to generalize with first time seen data. This has been proven not only by the instruments used in the Training and Testing phases but also from the low risk estimation that was done by an ERSVM. Water management experts should consider the estimations of the ANN for each watershed in order to design specific measures and the proper actions towards an orthological management of the existing water supplies in the mountainous watersheds. In this way the Torrential Risk and the Risk of water lack can be reduced.

Thus not only a useful and reliable tool has been developed for the water management of Thasos but also the potential use of ANN on a wider scale has been proven. More similar applications can be developed for other areas with available data. The modeling methodology is innovative and it uses a lot of modern instruments and algorithms that evaluate its performance from various perspectives.

Of course we can not claim that this model has covered all of the parameters determining the AAMWS and its corresponding SWF. With this volume of data (twenty torrential streams concerning a period of thirty years) using the eight mentioned input parameters this study can be considered as a successful preliminary one. The results of this research effort have a limited scope for the area of Thasos, but the modeling methodology can be applied in any place. However it can only be considered a preliminary one. A future effort with more data series concerning

perhaps more variables, will determine beyond any doubt the final structures of the proper ANN.

References

1. Bowers J.A., Shedrow C.B., (2000) Predicting Stream Water Quality using Artificial Neural Networks. Westinghouse Savannah River Company, SRS Ecology Environmental Information Document, MS-2000-00112. Savannah River Site, Aiken, SC 29808, US Department of Energy
2. Callan R., (1999) The Essence of Neural Networks. Prentice Hall, UK
3. Dibike Y., Solomatine D., (1999) River Flow Forecasting using Artificial Neural Networks European Geophysical Society (EGS) XXIV General Assembly, The Hague, The Netherlands
4. Drucker, H., C.J.C. Burges, L. Kaufman, A Smola, and V.N. Vapnik, (1997) Support Vector Regression Machines Advances in neural information processing systems Volume 9, 155-161. Cambridge, MA:MIT Press
5. Gaupe, D., (1997) Principles of artificial neural networks World Scientific. Singapore
6. Gavrilovic Sl. (1972) Inzenjering o bujicnim tovoklima i eroziji. Beograd
7. Hamed M. M., Khalafallah M. G., and Hassanien E. A., 2004. Prediction of wastewater treatment plant performance using artificial neural networks. Environmental Modelling and Software, Elsevier Science, October 2004, Vol. 19, Issue 10, pp. 919-928.
8. Haykin S., (1999) Neural Networks A Comprehensive Foundation. Prentice Hall,USA.
9. Herbrich R., (2002) Learning Kernel Classifiers: Theory and Algorithms. MIT Press, Cambridge, MA.
10. Iliadis L., Maris F., Marinos D., (2004) A Decision Support System using Fuzzy relations for the estimation of long-term torrential risk of mountainous watersheds: The case of river Evros. Proceedings
ICNAAM 2004 Conference, Chalkis, Greece
11. Jacobs, R.A. (1988) Increased rates of convergence through learning rate adaption. Neural Networks 1:295-307.
12. Kecman V., (2001) Learning and Soft Computing. MIT Press. London England.
13. Kotoulas D., (2001) Hydrology and Hydraulics of natural Environment, pp. 223. Aristoteles University of Thessaloniki, Greece.
14. Lobbrecht A.H., Solomatine D.P., (1999) Control of water levels in polder areas using neural networks and fuzzy adaptive systems. Water Industry Systems: Modelling and Optimization Applications, Vol, 1. pp. 509-518. Research Studies Press. Baldock, UK
15. Maier H. R., Dandy G. C., (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications

Environmental Modelling and Software, Volume 15, Issue 1, January 2000, pp. 101-124.
16. Maier H. R., Morgan N., Chow C. W. K., (2004) Use of ANN for predicting optimal alum doses and treated water quality parameters Environmental Modelling & Software, Volume 19, Issue 5, May 2004, pp. 485-494.
17. Minai, A.A. and R.D. Wiliams. (1990) Acceleration of Back-Propagation through Learning Rate and Momentum Adaption. International Joint Conference on Neural Networks. Volume I, pp. 676- 679.
18. Maris, L. Iliadis, (2007) "A computer system using two membership functions and T-Norms for the calculation of mountainous watersheds torrential risk: The case of lakes Trixonida and Lisimaxia" Book Chapter Book Series: "Developments in Plant and Soil Sciences" Book Title: "Eco-and Ground Bio-Engineering: The Use of Vegetation to Improve Slope Stability" Volume 103 pp. 247-254 Springer Netherlands
19. Neuralware, (2001) Getting started. A tutorial for Neuralworks Professional II/PLUS. Carnegie, PA, USA
20. Partalas I., Tsoumakas G., Hatzikos E., Vlahavas I., 2008. Greedy Regression Ensemble Selection: Theory and an Application to water Quality. Information Sciences 178 (20) pp. 3867-3879
21. Rummelhart D.E., Hinton G.E., Wiliams, R.J., (1985) Learning Internal Rep resentations by Error Propagation. Institute for Cognitive Science Report 8506. San Diego, University of California
22. Rummelhart D.E., Hinton G.E., Wiliams R.J., (1986) Learning Internal Repre sentations by Error Propagation. Parallel Distributed Processing, Explorations in the Microstructure of Cognition. Vol 1, Foundations. Cambridge, MA: MIT Press.
23. A. Tazioli, (2009) Evaluation of erosion in equipped basins preliminary results of a comparison between the Gavrilovic model and direct measurements of sedi ment transport. Environmental Geology. Volume 56, number 5, pp. 825-831, Springer Berlin
24. Thamarai S., Malmathanraj R., (2005) Missile Defense and Interceptor Alloca tion by Modifed Bionet Neural Network. Proceedings of the 9th International Con ference of EANN, Lille France pp. 299-306.
25. UNESCO (1998-2000). http://www.unesco-ihe.org/hi/projects.htm
26. Van Looy S., Meeus J., Wyns B., Vander Cruyssen B., Boullart L., Keyser F. (2005) Comparison of Machine Learning models for prediction of dose increase in patients with rheumatoid arthritis. Proceedings of the 9th International Conference of Engineering Applications of Neural Networks, Lille France pp. 189-196.
27. Vapnik V.N., (1995) The nature of statistical learning theory. New York: Springer Verlag
28. Vapnik V.N., S. Golowich, and A. Smola, (1997) Support Vector method for functuion approximation regression estimation and signal processing. In Advances in neural information processing systems. Vol.9. Cambridge, MA:MIT Press.
29. Vapnik V.N., (1998) Statistical learning Theory. New York Wiley.

A Data Mining System for Estimating a Large-size Matrix in Environmental Accounting

Ting Yu, Manfred Lenzen, Blanca Gallego and John Debenham

Centre for Integrated Sustainability Analysis, University of Sydney, NSW 2006, Australia

Faculty of Information Technology, University of Technology, Sydney, NSW 2007, Australia

Abstract: This paper presents a data mining system being capable of automatically estimating and updating a large-size matrix for environmental accounting. Environmental accounting addresses how to correctly measure greenhouse gas emission of an organization. Among the various environmental accounting methods, the Economic Input-Output Life Cycle Assessment (EIO-LCA) method uses information about industry transactions-purchases of materials by one industry from other industries, and the information about direct environmental emissions of industries, to estimate the total emissions throughout the whole supply chain. The core engine of the EIO-LCA is the input-output model which is in the format of a matrix. This system aims to estimate the large-size input-output model and consists of a series of components with the purposes of data retrieval, data integration, data mining, and model presentation. This unique system is able to interpret and follow users' XML-based scripts, retrieve data from various sources and integrate them for the following data mining components. The data mining component is based on a unique mining algorithm which constructs the matrix from the historical data and the local data simultaneously. This unique data mining algorithm runs over the parallel computer to enable the system to estimate a matrix of the size up to 3700-by-3700. The result demonstrates the acceptable accuracy by comparing a part of the multipliers with the multipliers calculated by the matrix constructed by the surveys. The accuracy of the estimation directly impacts the quality of environmental accounting.

1. Introduction

Environmental protection has caught more and more attention, while the climate is becoming more unpredictable. Instead of a particular protection technique such as water cleaning, environmental accounting brings environmental costs to the attention of corporate stakeholders who may be able and motivated to identify ways of

reducing or avoiding those costs while at the same time improving environmental quality.

In order to report the environmental cost of the activity of an organization, environmental accounting requires proper methodologies to correctly measure the environmental impact such as greenhouse gas emission [1]. There are several accounting approaches of measuring emission, such as auditing and *triple bottom line methods (TBL)* [2]. The TBL method captures an expanded spectrum of values and criteria for measuring organizational (and societal) success: *economic, ecological and social*. With the ratification of the United Nations TBL standard for urban and community accounting in early 2007, this became the dominant approach to public sector full cost accounting [3]. The traditional accounting method only measures success regarding the economic, but ignore other two. A *life cycle assessment (LCA)* is the investigation and valuation of the environmental impacts of a given product or service caused or necessitated by its existence, and an evaluation of the environmental impacts of a product or process over its entire life cycle. Environmental life cycle assessment is often thought of as "cradle to grave" and therefore as the most complete accounting of the environmental costs and benefits of a product or service [4]. Among the various LCA methods, the Economic Input-Output Life Cycle Assessment (EIO-LCA) method uses information about industry transactions - purchases of materials by one industry from other industries, and the information about direct environmental emissions of industries, to estimate the total emissions throughout the supply chain [4]. In the EIO-LCA method, the input-output model is the key engine. The input-output model simply uses a matrix representing the intra-industry flows and the flow between industrial sections and consumption or the flow between the value-added section and the industrial section. Because the economic constantly evolves, the input-output model needs to be updated at least annually to reflect the new circumstance. Unfortunately, in most countries such as Australia, the input-output model is only constructed every 3-4 years, because the large amount of monetary and human cost is involved. The Centre for Integrated Sustainability Analysis (ISA), University of Sydney, is developing a data mining system to estimate and update the input-output model at different level on a regular basis.

The past decades have seen the booming supply of data from various sources, and large amounts of data regarding the environment and economic can be accessed. Unavoidably, data from various sources has various structures and ways of represent their underlying meaning. It is a time-consuming process to restructure the various types of data into a single structure and estimate the matrix. In many cases, this kind of integration and matrix estimation operation becomes a daily routing task in order to keep the information up to date. The proposed system aims to automate the whole process and reduces the manual intervention and much human's involvement.

2. System Design

The whole system consists of functional components: data retrieval, data integration, data mining and model presentation. The row data is retrieved from various data sources, and restructured and integrated into a data mining model. Then the data mining model is fed into the data mining algorithm and consequently solved by the optimization engine. The result from the data mining algorithm is the final result that is an estimated matrix.

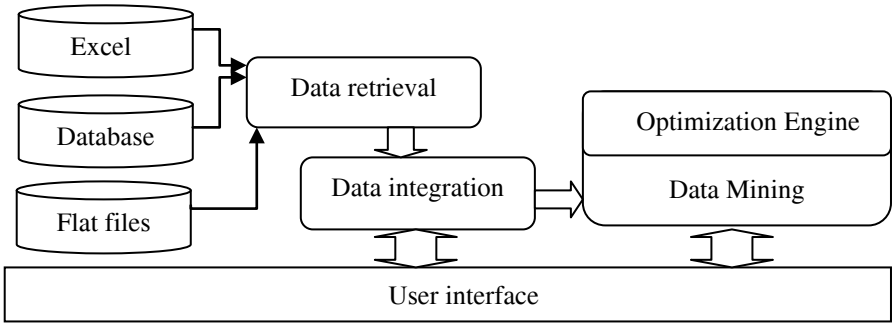


Fig. 1. Structure of the System

The data retrieval component acts as interfaces to all types of datasets including greenhouse gas emission measurement, macro and micro economic data that are stored in various formats such as Excel files, databases etc. The data integration component unifies these heterogeneous datasets to a single format, integrates and restructures the data retrieved by the previous component and presents the result for data mining. The data mining component is the core of the whole system. In this component, a unique data mining algorithm is designed to estimate the matrix.

3. Data Integration

The data integration component includes two main sub-modules: the structure builder and the model constructor. The structure builder constructs the tree structure that we will discuss in detail later, and the model constructor constructs the mining model for the following data mining component. Within the model constructor, there are two processes to restructure the data: 1) require the interfaces to retrieval data from various sources and integrate them, and 2) restructure and assign the meaning to the data according to the previous tree structure and users' specification and populate the mining model.

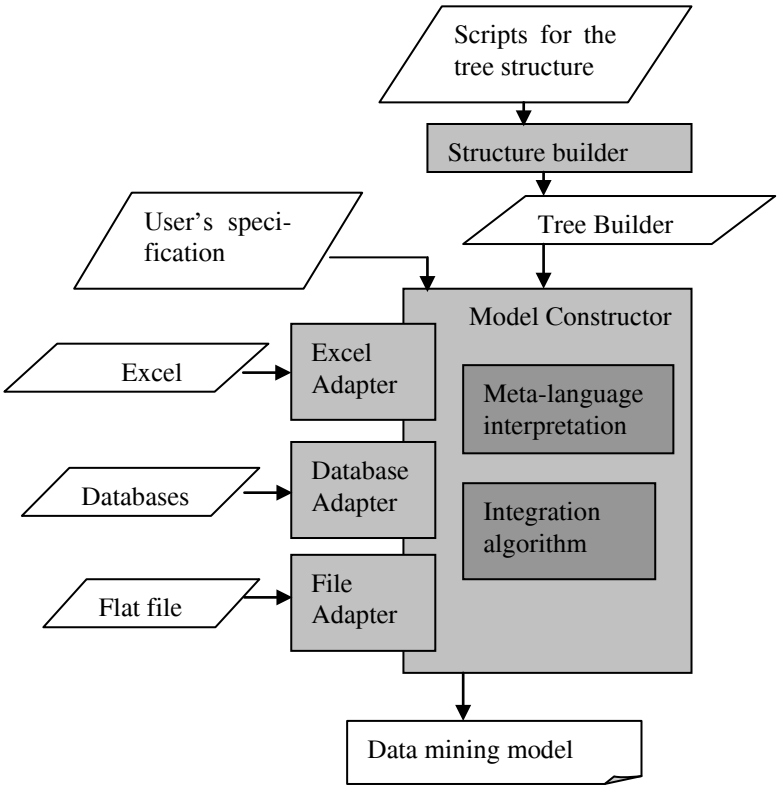


Fig 2. Data Integration Component

The first step is to construct the tree structure. The tree structure is pre-required for restructuring data collected from various sources. An example of the tree structure (See Fig 3) is a three-level tree representing the Australian Economic, one branch of which represents the sheep industry section within the New South Wales, a state of Australia. If the numerical indices are employed instead of their names, the sheep industry section within the New South Wales, a state of Australia can be written in [1,1,1] which means the first leaf in the first branch of the first tree.

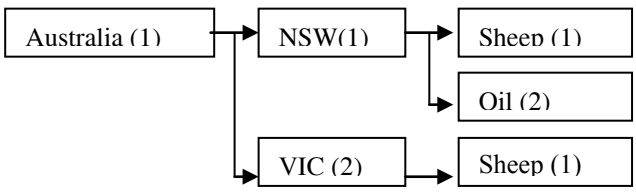


Fig 3, An Example of Tree Structure

The row and column of the matrix is defined by this tree structures, thereby the matrix is defined by the tree structures. The tree structure is unnecessarily with three levels. For example, a matrix (see Figure 4) can be organized by one three-

level tree at the row side and one two-level tree at the column side. The coordinate of one entry, say X_1, can be defined as by [1,1,1] at the row side and [1,1] at the column side. That means the entry, X_1, defined by a three-level tree structure and a two-level tree structure at the column side. The tree structure is crucial to assign the meaning to the data retrieved from various sources, since the coordinates of entries are completely determined by it.

			China (1)	
			Shoe (1)	Retail (2)
Australia (1)	NSW (1)	Sheep (1)	$X_1 = 0.23$	X_2
		Oil (2)	X_3	X_4
	VIC (2)	Sheep (1)	X_5	X_6
		Oil (2)	X_7	X_8

Fig 4, An Example of the Matrix Defined by the 3-level Tree and the 2-level Tree

Considering the difference between applications, a dynamical structure of resultant matrix provides the flexibility to expand this software system to different application. On the other hand, the flexibility of the structure makes the system to be available to various level of implementation. For example, there is huge difference between the structures of resultant matrix at the national and at the corporate level, as the operations within a corporate are much simpler than those of a nation in the most cases. The dynamic of the structure is introduced by a multi-tree structure like Figure 3.

Considering the complexity of the model, a Meta language is introduced to provide users' an easy way to organize their data. The Meta language must be compact and accurate to make the description to be readable and useful. It is unrealistic to write hundred thousands of code to describe a single model at a daily base. The meta language we create is based on the coordinate of the valuable in the resultant matrix. For example, the coordinates of one entry is written as [1, 1, 1 -> 1, 1]. The value of this entry X_1 is indicated as the (0.23) [1, 1, 1 -> 1, 1] (See Figure 4). The system will fill the 0.23 in the cell with the coordinate [1, 1, 1] at the row side and [1, 1] at the column side. Consequently, this script indicates that 0.23m dollar worth of sheep products are transferred to the shoe industry in China. Some other notations are also included in order to improve the flexibility and efficiency of the Meta language. Therefore, users' specification is a set of XML-based files including some scripts written in the meta language. This kind of XML-based file indicates the system where to find the data source, how to retrieval the desired information from these sources and where to allocate the data into the optimization model.

4. Temporal-Spatial Mining with Conflict Information

The data mining component is the core engine of the whole system. In this component, a unique data mining algorithm is designed to estimate the matrix. This mining algorithm utilizes two types of information: the historical information which contains the temporal patterns between matrices of previous years, and the local information within the current year. For example, this local information can be the total output of the given industry within the current year, or the total greenhouse emission of the given industry. The simplified version of the mining algorithm can be written in the format of an optimization model as below:

$$Min[\frac{dis(X-\overline{X})}{\varepsilon_1}+\sum\frac{e_i^2}{\varepsilon_{i+1}}], \text{ subject to: } GX+E=C \quad (1)$$

where:

X is the target matrix to be estimated , $\overline{X}$ is the matrix of the previous year, E is a vector of the error components $[e_1,...,e_i]^T$

dis is a distance metric which quantifies the difference between two matrices.

G is the coefficient matrix for the local constraints

C is the right-hand side value for the local constraints.

As the *dis* metric has many variety, the one used in this paper is the $\sum(X_i-\overline{X}_i)^2$.

The idea here is to minimize the difference between the target matrix and the matrix of the previous year, while the target matrix satisfies with the local regional information to some degree. For example, if the total export of the sheep industry from Australia to China is known as c_1, then $GX+E=C$ can be $[1,1]\begin{bmatrix}X_1\\X_2\end{bmatrix}+e_1=c_1$. The element e_i in E represents the difference between the real value and estimate value, for example, $e_1=c_1-[1,1]\begin{bmatrix}X_1\\X_2\end{bmatrix}$. The reason why it is introduced is to solve the *conflicting information*. Very often the data collected from different sources is inconsistent between each other, and even conflicting. For example, the number of export of Australian Sheep industry reported by the Australian government may be not consistent with the number of import from Australian Sheep by the Chinese government. Here e_i is introduced to balance the influence between the conflicting information, and reaches a tradeoff between the conflicting information.

This mining algorithm assumes the temporal stability, which assumes the industry structure of a certain region keeps constant or has very few changes within the given time period. This assumption is often required to be verified for long time

period. Within the short time periods, dramatic change of the industry structure is relatively rare.

Traditional data mining only employs the single data mining techniques, such as the temporal model or spatial model. But two-dimensional data mining algorithm can integrate two types of the data mining models, thereby maximally utilize the available information. In our case, $dis(X - \overline{X})$ models the temporal information of the input-output models between years, and $GX + E = C$ models the spatial or other type of regional information of the input-output model within the current year.

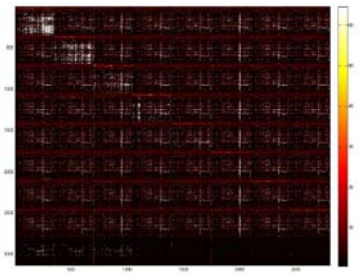
The reason why the temporal-spatial mining algorithm is suitable to this system is due to the unique characteristics of the datasets that the system aims to process. The datasets often contain the temporal patterns between years, such as the trend of the carbon emission of certain industry sections, and also much spatial information regarding the total emission within a certain region such as national total emission and state total emission. On the other hand, it is very common that either of datasets is not comprehensive and imperfect and even the conflicts between the datasets exist. Thereby, the mining algorithm is required to consolidate the conflicted datasets to uncover underlying models.

5. Experimental Results

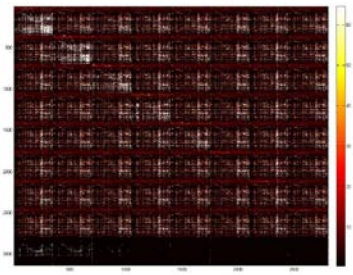
The direct evaluation of a large-size matrix is a rather difficult task. A thousand-by-thousand matrix contains up to ten million numbers. Simple measurements such as the sum do not make too much sense, as the important deviation is submerged by the total deviation which normally is far larger than the individual ones. The key criterion here is the distribution or the interrelationship between the entries X_i within the matrix: whether the matrix reflects the true underlying industry structure, not necessary the exactly right value, at least the right ratios. During the experiment, the coefficient ε_1 in the equation (1) is tuned to fit the data properly. Here three examples are presented to demonstrate the effects of the tuning.

Darker the color is, the smaller the value of the entry will be. From three pictures (See Figure 5), while ε_1 is set smaller, the mining algorithm pushes the model toward the first part of the equation (1).

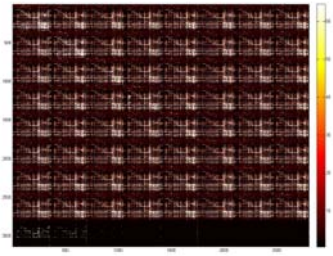
Often, we estimate the result of experiments by two methods: direct comparison and indirect comparison between the multipliers of the matrices. The comparison between a part of the resulting input-output table and the available survey data examines the quality of the result of the experiment under the microscope, but it hardly gives the overall quality.



with the coefficient as $\varepsilon_1 = 0.01$



with the coefficient as $\varepsilon_1 = 0.1$



with the coefficient $\varepsilon_1 = 1$

Figure 5: Three pictures by the turning the coefficient ε_1

The multipliers in the input-output framework reflect the aggregated impacts of the final demand changes on the upstream industries [5]. The information contained by the multipliers is very similar to the sensitivity analysis in general statistics. The general formula of constructing the multipliers is:

$$M = D(I - A)^{-1}$$

where M is the multiplier, I is the identity matrix, D is the change in the final demand, and A is the technique coefficients matrix, each entry of which is $X_i / \sum_{i=1}^{n} X_i$. Here, X_i is a value from the matrix estimated by the equation (1). More detailed explanation is available at [5]. This multiplier counts the impact of the change on the whole upstream industries, and not only the direct supply of the final demand. Any deviation occurring in the upstream industries from the underlying true structure will be amplified and reflected on the multipliers. Thereby, the multipliers send an indirect warning signal to imply the deviation occurring on the upstream.

Here a part of the multipliers are used to measure the quality of the resulting matrix. This matrix aims to calculate the total water usage of the different industries in Australia. A part of the data is collected from the Water Account reports pro-

duced by the Australian Bureau of Statistics [6]. In the following tables, the direct intensity indicates the direct usage of the water by the industry, and total multipliers aggregate the whole upstream water usage of the industry.

Industry	Water Usage (ML)	Final Demand (K$)	Direct Intensity	Total Multiplier
Sheep and lambs	385360	2198275	0.175306192	0.229353737
Wheat	795403	4442246	0.179059807	0.27047284
Barley	193433	1080896	0.178962619	0.235765397
Beef cattle	1557332	8872487	0.175528219	0.265691956
Untreated milk & Dairy cattle	2275602	3687201	1.233958	1.46699422
Pigs	150566	847785	0.177604301	0.273531332
Poultry & Eggs	312984	1811428	0.288054	0.47927187
Sugar cane	1269012	346329	3.664307556	3.720540595
Vegetables & Fruit	862027	3747712	0.262444	0.3157365
Ginned cotton	2120	2534832	0.000836372	0.297221617
Softwoods	141702	809463	0.175060473	0.236982027
Hardwoods	53954	307576	0.175416808	0.239130312
Forestry	150577	860234	0.175046587	0.229861225
Black coal	159409	18603943	0.008572854	0.10262849

Table 1: Estimated Multipliers

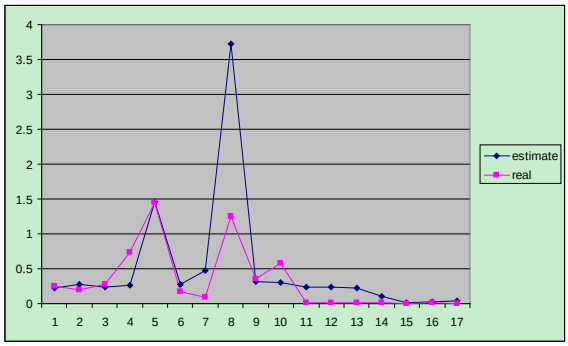


Figure 6: Comparison between two series of multipliers

From the above plot (see Figure 6) comparing the two series of the multipliers, two series basically follow the same pattern, which indicate the industry structure is estimated properly. However the estimated multipliers are more volatile than the true underlying multipliers. This phenomenon indicates the estimated multipliers amplify the errors introduced to the upstream industries.

6. Conclusion

To the best of our knowledge, this system is the first data mining system for estimating the input-output tables for environmental accounting. The unique characteristics of the data for environmental accounting determine the data mining system must be capable of dealing the temporal and spatial data simultaneously. At the same time, the large size of the estimated matrix makes it a difficult task to check the quality of the matrix. This paper presents a completed data mining system starting from data collection to data mining and presentation. According to the result of the experiments, the system successfully produces the input-output tables for the triple bottom line methods (TBL) in environmental accounting. This system makes environmental accounting a rather easy task without a huge amount of work to collect and update both data and model. Before this system, this kind of collection and updating work costs months of work, but now it takes only a few days with the consistent quality.

There are still many places to be further investigated. For example, the mining algorithm can incorporate more historical data by including the data of a few previous years instead of the data of the immediate previous year in the current model. It requires much larger computational ability, and we are investigating the algorithms over the parallel computing which will bring enormous power to process extremely large datasets.

Acknowledgement:
Authors would like to thanks our colleagues Dr Chris Dey, Jeremy Badcock and Jodie Gonzalez Jennings for contribution.

Reference:

1. Peters, G., et al., *Towards a deeper and broader ecological footprint.* Engineering Sustainability, 2007.
2. Foran, B., M. Lenzen, and C. Dey, *Balancing Act: A Triple Bottom Line Analysis of the Australian Economy*. 2005: CSIRO and the University of Sydney.
3. Brown, D., J. Dillard, and R.S. Marshall, *Triple Bottom Line: A Business Metaphor for a Social Construct* in *Understanding the Social Dimension of Sustainability*. 2006, Taylor & Francis Group.
4. Hendrickson, C.T., L.B. Lave, and H.S. Matthews, *Environmental Life Cycle Assessment of Goods and Services: An Input-Output Approach*. 2006 Resources for the Future.
5. Miller, R.E. and P.D. Blair, *Input-output Analysis, Foundations and Extensions*. 1985, Englewood Cliffs, New Jersey: Prentice-Hall Inc.
6. *4610.0 - Water Account, Australia*. 2004-05, The Australian Bureau of Statistics: Canberra.

Abductive Reasoning in Environmental Decision Support Systems

Franz Wotawa, Ignasi Rodriguez-Roda, and Joaquim Comas

Abstract Decision trees and rule-based system including variants based on propositional and fuzzy logic have been the method of choice in many applications of environmental decision support systems. Reasons are the ease of use, the capability of representing uncertainty, and the fast computation of results at runtime when using decision trees or other similar means for knowledge representation. Unfortunately there are drawbacks related with these modeling paradigms. For example, the cause-effect relationships between quantities are not captured correctly. The resulting model is well appropriated for a certain purpose but can hardly be re-used. Moreover, maintaining the knowledge base can be an intricate task. In this paper we focus on the problems related with decision trees in the context of environmental decision support systems using an example from the domain. We further present abductive reasoning as an alternative for modeling and show how these technique can be easily implemented using existing techniques.
Keywords: environmental decision support systems, abductive reasoning, modelling

1 Introduction

Decision support systems gain importance. This holds especially in the environmental domain where decisions have to be drawn but where knowledge is not commonly accessible and not easy to obtain. In most cases environmental decision support systems like [1, 2] are based on decision trees, rule-based systems, case-based systems, or fuzzy logic. Although, these methodologies have been successfully used in practical applications, e.g., [16], they have some drawbacks. One drawback is that the

Franz Wotawa
Technische Universität Graz, Institute for Software Technology, Inffeldgasse 16b/2, A-8010 Graz, e-mail: wotawa@ist.tugraz.at

Ignasi Rodriguez-Roda and Joaquim Comas
Laboratory of Environmental and Chemical Engineering University of Girona, E-17071 Girona, Catalonia (Spain) e-mail: {ignasi.rodriguezroda,joaquim.comas}@udg.edu

models used are different from the models used in physics or chemistry. Hence, it is necessary to rewrite the model in order to fit the purpose. On the other hand there is the advantage that decisions can be easily obtained from the models and thus explaining the reasons behind a decision is easy.

In order to make modeling straightforward without loosing the capabilities of providing a decision and the reasons behind in an easy way, model-based reasoning has been developed. In model-based reasoning a model is directly used to provide solutions for a problem like diagnosis. The model itself should be as close as possible to the original model. Model-based reasoning has been successfully applied to diagnosis [17]. There are applications also in the environmental domain. See for example the work by Struss and colleagues [12, 13, 18, 14]. All described applications are based on consistency-based diagnosis where models of the correct behavior have to be available. However, there is another methodology for model-based reasoning which relies on abductive reasoning. In abductive reasoning models of the faulty behavior have to be formalized in order to get an explanation for a given problem. In the environmental domain the faulty behavior is usually known as well as their consequences. Hence, abductive reasoning seems to provide a good foundation for environmental decision support systems.

In this paper, we discuss the problems behind decision trees and other similar methods used for modeling in detail by means of using a knowledge-based model to detect the risk of microbiology-related solids separation problems, which is one of the main critical perturbation in the biological treatment of wastewater. Afterwards we introduce abductive reasoning and present an algorithm that allows for computing minimal explanations. Finally, we give a brief presentation of our implementation and a conclusion.

2 Problem description

In this section we discuss modeling using decision trees in detail. We use Comas et al. [2] decision tree model that is used to predict the risk of bulking, foaming, and rising sludge, microbiology-related operational problems when simulating biological wastewater treatment. Beside the decision trees for the different types of risks the authors give a verbal explanation and tables where the involved parameters and their corresponding risks are related to each other. In order to be more comprehensive we focus on only one simplified phenomenon, i.e., the risk of foaming. According to Comas et al. [2] the risk of foaming is influenced by the food-to-microorganism ratio (F/M_fed), the sludge residence time (SRT), the dissolved oxygen concentration (DO), and the ratio between readily biodegradable substrate concentration (S_s) and the slowly biodegradable substrate concentration (X_s). The verbally given explanations relate these parameters to the growth of certain microorganism, i.e., *Nocardioforms* and *Microthrix parvecilla*, which cause two different types of foaming. The given explanation is modeled using decision trees. Figure 1 depicts the given decision tree. Moreover, the authors also specify the behavior by means of a table.

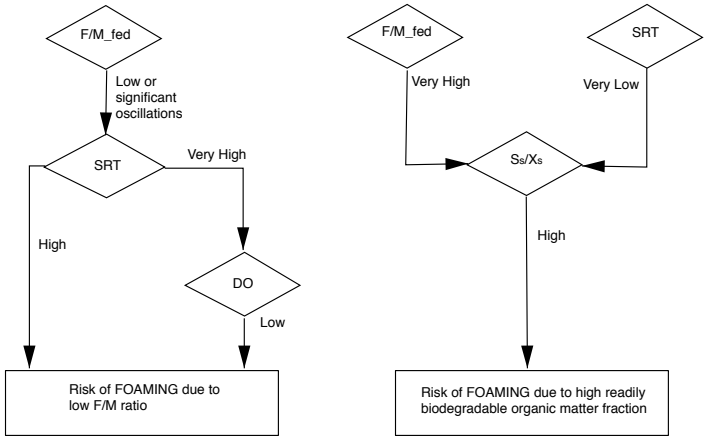


Fig. 1 Decision tree for evaluating the risk of foaming taken from [2]

In the given table each entry specifies a qualitative value for the risk of foaming given a qualitative value for a parameter. In order to be self contained the table representing the knowledge stored in the left decision tree of Figure 1 is given in Table 1 ignoring the DO parameter. Beside the fact that the DO parameter is not represented it is interesting to note that the authors use a qualitative representation for the parameters and the risk assessment. To obtain a qualitative representation of a parameter value requires an additional step. In the case of Comas et al. [2] a fuzzification step is used. After obtaining a qualitative value for the risk defuzzification can be used to finally receive a quantitative risk value.

Table 1 Table representing the knowledge behind foaming due to low F/M

	F/M_fed			
SRT	Low	Normal	High	Very High
Very Low	Low	Low	Low	Low
Low	Low	Low	Low	Low
Normal	Moderate	Low	Low	Low
High	High	Moderate	Low	Low
Very High	High	Moderate	Low	Low

Given the table and the decision trees it becomes obvious that the original decision tree does not represent the cases where the risk assessment is at the level moderate or low. Moreover, the decision trees from Figure 1 should be put together

in order to represent the whole knowledge at once. Both of these weaknesses of the original decision tree model from [2] can be easily incorporated in a decision tree by introducing new leaf nodes for the different sort of risk assessments and new connections. However, there are other issues which hardly can be tackled in such a simple way.

Consider the SRT values given in Table 1. If SRT is known to be low the foaming risk is low and there is no need to know any value for F/M anymore. Hence, the size of the decision tree, which corresponds to the number of decisions, depends on the ordering of the decision. Algorithms for decision tree induction take care of the size of the decision tree by selecting the optimal ordering of decisions to be taken. But even in the case of optimal orderings a decision tree requires to answer decisions that need not to be answered always. Hence, decision trees cannot adapt to certain situations. We term this problem as *poor flexibility* problem.

Another drawback of decision trees is that the absence of knowledge cannot be handled appropriately. In particular it is hardly possible to conclude any default knowledge. For example, it might be useful to assume that there is a low risk of foaming in case we have no knowledge at all. Such reasoning can only be represented if adding a new decision asking whether there is knowledge available or not. However, in this case the knowledge that no knowledge is there is now explicit available, which is not the same in default reasoning where something can be conclude unless it contradicts given knowledge. Therefore, decision trees *lack handling default reasoning* in case of unavailable knowledge.

Since decision trees are constructed in a way that supports a certain task, they usually do not represent the whole knowledge available. For example, in the waste water treatment plant domain there is knowledge regarding the growth of certain microorganisms as function of parameters. In the decision tree this knowledge is not represented. Instead only decisions regarding parameter values are represented. Decision trees represent a sequence of questions to be answered by the user in order to distinguish final conclusions. Hence, it is not necessary to represent all cause-effect relationships, which do not support this task. The fact that decision trees are *not complete* representations of models for a domain is a problem for extracting knowledge to be generally used.

Because of the mentioned problem of decision trees, i.e. the:

- the poor flexibility,
- the absence of default reasoning capabilities, and
- the incompleteness

they are usually hard to maintain. In many cases small changes of the underlying theory causes huge changes of the decision tree. This might not be a problem in cases where either the domain knowledge is small or not very much elaborated, or where the decision tree is automatically extracted from a set of data. In the latter case the decision trees help to extract useful information regarding relationships of certain parameters.

Modeling in the domain of natural sciences is a difficult and time consuming task. In many cases especially when explanations are important the models repre-

sent *cause-effects* relationships because effects can be explained in terms of their causes. In most cases decision trees do not capture cause-effect relationships. Instead decisions are based on effects whereas the leaf nodes represent the causes. Hence, reasoning is applied from effects to causes and not vice versa. Note that in our example application, i.e., the risk assessment, someone can argue that the decision trees indeed represent the reasoning from causes to effects although, the used causes are not necessarily the root causes from which all explanations start. A more typical example of the successful use of decision trees for modeling is discussed in [1] where cause-effect reasoning is not handled correctly.

Given the fact that decision trees are very successfully used in practice a question arises. Is cause-effect modeling essential in practice? No, this is not the case, and even from a more philosophical perspective, causality is not the only way of expressing explanations in science (see for example [15]). However, as already mentioned causality-based models are usually available when it comes to explanations. Hence, when using a modeling paradigm like decision trees, which is not able to represent causality in all cases, we have to convert the cause-effect models into models allowing for reasoning from effects to causes. This conversion is a tedious task because it requires the elimination of multiple explanations. We always are able to eliminate multiple explanations. For this purpose distinguishing observations, i.e., decisions, have to be introduced. In summary, we can say that it is not inevitable for successful practical use to represent cause-effect reasoning in the model but this requires additional effort in order to eliminate ambiguities in explanation.

In the next section, we discuss an alternative reasoning schema, i.e., abductive reasoning, which allows using models representing cause-effect relationship directly and thus avoiding the mentioned problems.

3 Abductive reasoning

Given the problems regarding modeling using decision trees or rule-based systems, which we discussed in the previous section, we now focus on a different modeling paradigm. In abductive reasoning the causes are inferred from a logical model representing cause-effect relationships. Therefore, the logical model is most closely to models available. Note that for example in medicine the available model describe causes, i.e., diseases, and their effects, i.e. symptoms but the medical doctors have to conclude the disease from the available symptoms. Hence, medical doctors always use abductive diagnosis. The formalization of this process including therapy is discussed in [11]. Wotawa [19] describes the application of abductive reasoning in the environmental domain and focuses on effects. In particular [19] introduces an algorithm for computing the next optimal observation necessary to reduce possible explanations. The underlying ideas came from work on consistency-based diagnosis, i.e., [17, 8, 9]. The difference between consistency-based diagnosis and abductive diagnosis is that the former uses the correct behavior only whereas the latter has

knowledge regarding the behavior in case of faults. Console et al. [4, 3] formally prove the relationship between abductive and consistency-based diagnosis.

In this section, we focus on the implementation of abductive reasoning. The idea is to rely on well-known algorithms. In particular we show how assumption-based truth maintenance systems [5, 6] can be used for computing abductive explanations for given effects. Before introducing the algorithm we briefly give the basic definitions. For a more detailed explanation we refer the interested reader to [19]. We start with the definition of knowledge base.

Definition 1 (Knowledge base (KB)). A knowledge base (KB) is a tuple (A, Hyp, Th) where A denotes the set of propositional variables, $Hyp \subseteq A$ the set of hypothesis, and Th the set of horn clause sentences over A.

In the definition of KB hypotheses corresponds directly to causes such that for every cause there is a propositional variable that allows to hypothesis about the truth value of the cause. Hence, we use the terms hypotheses and causes in an interchangeable way. Having knowledge about a system and some observations we are interested in finding explanations. This leads naturally to the definition of abduction.

Definition 2 (PHCAP). Given a knowledge base (A, Hyp, Th) and a set of observations $Obs \subseteq A$ then the tuple (A, Hyp, Th, Obs) forms a propositional horn clause abduction problem (PHCAP).

Given a PHCAP we are interested in a solution. Hence, similarly to [11] we define solutions as follows:

Definition 3 (Diagnosis; Solution of a PHCAP). Given a PHCAP (A, Hyp, Th, Obs). A set $\Delta \subseteq Hyp$ is a solution if and only if $\Delta \cup Th \models Obs$ and $\Delta \cup Th \not\models \bot$. A solution Δ is parsimonious or minimal if and only if no set $\Delta' \subset \Delta$ is a solution.

In this definition diagnoses need not to be minimal or parsimonious. In most practical cases only minimal diagnoses or minimal explanations for given effects are of interest. Hence, from here on we assume that all diagnoses are minimal diagnoses if not specified explicitly.

Example 1. We use the rightmost decision tree from Figure 1 and model the knowledge represented there as KB. We use `fm_fed`, `srt`, `do` to represent the involved variables. `foaming_risk` is used to represent risk of foaming. The values of the variables are given in parantheses. The horn clauses for representing the knowledge can be formulated as follows:

`fm_fed(low)` $\wedge$ `srt(high)` $\rightarrow$ `foaming_risk(high)`

`fm_fed(low)` $\wedge$ `srt(very_high)` $\wedge$ `do(low)` $\rightarrow$ `foaming_risk(high)`

This model is not complete because there are currently no hypothesis specified, which are of interest to explain a certain observation. In this example we are interested in explaining the assessment of risk. Hence, we introduce the hypotheses `FM_fed_L`, `SRT_H`, `SRT_VH`, `DO_L` that represents certain values of the involved variables. Extending the KB with information regarding hypothesis requires to add

the following rules:
FM_fed_L → fm_fed(low)
SRT_H → srt(high)
SRT_VH → srt(very_high)
DO_L → do(low)
Moreover, from Table 1 we might conclude that a SRT value that is low or very low always leads to a low risk. This holds also for F/M_fed in case of being high or very high. Such knowledge can also be introduced in a similar way: FM_fed_H → fm_fed(high)
FM_fed_VH → fm_fed(very_high)
SRT_L → srt(low)
SRT_VL → srt(very_low)
fm_fed(high) → foaming_risk(low)
fm_fed(very_high) → foaming_risk(low)
srt(low) → foaming_risk(low)
srt(very_low) → foaming_risk(low)
What is missing to complete the KB is information regarding inconsistencies. Obviously it can never be the case that a variable has different values assigned at the same time. Hence, we introduce rules like foaming_risk(high) ∧ foaming_risk(low) → ⊥ to KB where ⊥ represents the contradiction, i.e., an always wrong proposition.

¿From the KB we can abductively conclude the following explanations, i.e., diagnoses, from the observation foaming_risk(high): { FM_fed_H, SRT_H }, { FM_fed_H, SRT_VH, DO_L }. Both diagnoses are parsimonious. { FM_fed_H, SRT_H, DO_L } is also an explanation but not minimal. { FM_fed_H, SRT_H, SRT_L } is not a diagnosis because it leads to an inconsistency.

Abductive reasoning, i.e., providing a parsimonious explanations for observations given a KB, can be implemented by checking all subsets of the hypotheses whether they allow inferring the observations in a consistent way. This unfortunately is not effective in practice. Another way is to rely on available systems and algorithm for reasoning based on explanations. Assumption-based truth maintenance systems (ATMS) [5, 6] can be used for this purpose. An algorithm implementing the ATMS has been provided by de Kleer [7]. Many improvements for computing solutions based on ATMS has been suggested including [10]. An ATMS also works on KB defined in this paper when using the word assumptions instead of hypothesis. The ATMS works on a graph representation of KB. Assumptions, propositions, and the contradiction are represented as nodes. The contradiction is named NoGood in terms of the AMTS. The rules are represented as set of connections between nodes. Every node in an ATMS has a label. The label comprises all sets of assumptions from which the corresponding proposition can be derived.

Example 2. The hypothesis FM_fed_H is represented by a node. The label of this node is a set comprising the hypothesis because this node is only true if the hypothesis, i.e., the assumption, is assumed to be true. The label of the proposi-

tion `fm_fed(high)` comprises the set `FM_fed_H` because of rule `FM_fed_H` $\rightarrow$ `fm_fed(high).`

The task of the ATMS is to ensure consistency. This is done by changing the label of nodes. Each node has a label, i.e., a set of sets of assumptions from which the node can be inferred and from which the contradicting node cannot be inferred. The latter requirement causes an ATMS algorithm to remove elements from the label that also lead to the NoGood. Hence, an abductive explanation for a single proposition is an element of the label of the corresponding node. Because of the ATMS these elements provide a consistent explanation and fulfill the definition of abductive diagnosis. The only thing that remains now is the extension to the case where we have a set of observations to be explained. This extension can be easily done by adding a rule where the left side is the conjunction of all observations and the right side is a new proposition `explanation` not used in the KB. Hence, the label of `explanation` provides all abductive diagnoses for the given PHCAP. The following algorithm for computing all parsimonious abductive explanations relies on this generalization.

Algorithm abductiveExplanations
Input: A PHCAP (A, Hyp, Th, Obs)
Output: All minimal diagnoses

1. Store Th in an ATMS
2. Add the rule $\bigwedge_{o \in Obs} o \rightarrow$ `explain` to the ATMS where `explain` is not an element of A.
3. Return the label of `explain` as result.

4 Implementation

We have implemented an abductive reasoning system based on an ATMS using the programming language Java. Figure 2 depicts the main window of our implementation where the user can edit, save, or load a KB. Instead of $\wedge$, $\rightarrow$, and $\bot$ a comma ',' '->', and 'false' are used. To distinguish hypothesis from ordinary propositions the former start with a capitalized character. Moreover, every rule has to be ended with a period '.'. The KB given in Figure 2 is the one discussed in Example 1. Figure 3 shows the window where the results are presented to the user. The label of the NoGood as well as the label of the node `foaming_risk(high)` are given. For the latter we obtain the result also given in Example 1. The labels of the other nodes can be obtained by expanding the nodes.

Beside the given small abductive theory we tested out implementation on other KBs having from 10 to about 50 rules and from about 4 to 12 hypotheses. For all examples, the running time was less than 10 ms on a standard notebook. Because of the fact that the computational complexity of the underlying problem is exponential, we do not expect to be able to handle larger systems comprising hundreds of hypotheses. However, in the environmental domain the number of hypotheses

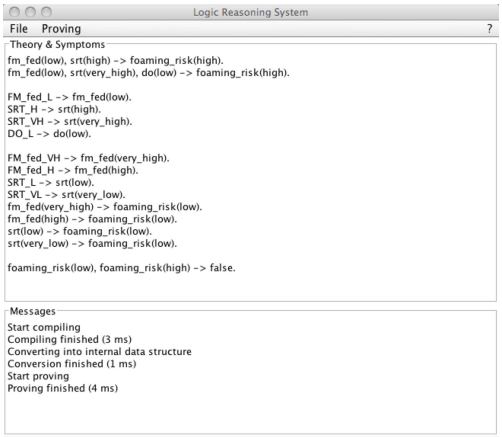


Fig. 2 The main GUI of our implementation

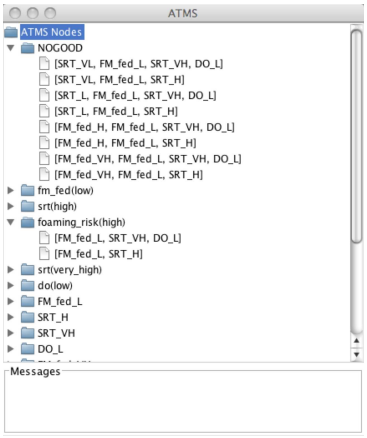


Fig. 3 The GUI providing the obtained results

is expected not to be too high. It is worth noting that the current implementation does neither use the most well elaborated ATMS algorithm nor is itself optimized. Hence, for the desired domain and given today's computational power, the proposed methodology seems to be appropriated.

5 Conclusion

The purpose of this paper is twofold. First, it discusses problems related to currently used techniques for environmental decision support systems, which often rely on decision trees, rule-bases, or fuzzy logic. Problems are the missing flexibility, failing to model default reasoning, and incompleteness. These problems may not impact a certain application. However, they prevent the models to be adapted and used in other applications. Moreover, in most cases creating and maintaining such models is not as easy than expected. Second, the paper provides a solution to the mentioned problem. In particular we propose the use of abductive reasoning as basis for environmental decision support systems. In abductive reasoning models based on cause-effect relationships can be directly used. Moreover, default reasoning is also possible.

¿From a practical view the use of abductive reasoning for applications has been limited because of the unavailability of tools that can be used by people who are not expert in logical-based modeling. Currently, there is an implementation available but the development of models might still not be that easy. In the future we want to focus on usability regarding modeling. Moreover, the whole process of decision making has to be captured by an implementation. Hence, getting more information in a smart way has to be ensured. Again we leave this topic for future research.

Acknowledgements The work described in this paper was partially funded by Austrian Academic Exchange Service (OeAD) under contract ES 19/2008.

References

1. Comas J, Rodriguez-Roda I et al (2003) A knowledge-based approach to the deflocculation problem: integrating on-line, off-line, and heuristic information. *Water Research*, 37:2377–2387.
2. Comas J, Rodriguez-Roda I et al (2008) Risk assessment modelling of microbiology-related solids separation problems in active sludge systems. Environmental Modelling and Software **23** 1250–1261
3. Console L, Dupré D. T et al (1991) On the relationship between abduction and deduction. *Journal of Logic and Computation*, 1(5):661–690.
4. Console L, Torasso P (1990) Integrating models of correct behavior into abductive diagnosis. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 160–166, Stockholm. Pitman Publishing.
5. De Kleer J (1986) An assumption-based TMS. *Artificial Intelligence*, 28:127–162.
6. De Kleer J (1986) Problem solving with the ATMS. *Artificial Intelligence*, 28:197–224.
7. De Kleer J (1988) A general labeling algorithm for assumption-based truth maintenance. In *Proceedings AAAI*, pages 188–192, Saint Paul, Minnesota, Morgan Kaufmann.
8. De Kleer J, Williams B. C (1987) Diagnosing multiple faults. *Artificial Intelligence*, 32(1):97–130.
9. De Kleer J, Mackworth A. K et al (1992) Characterizing diagnosis and systems. *Artificial Intelligence*, 56.
10. Forbus K. D, De Kleer J (1988) Focusing the ATMS. In *Proceedings AAAI*, pages 193–198, Saint Paul, Minnesota, Morgan Kaufmann.
11. Friedrich G, Gottlob G et al (1990) Hypothesis classification, abductive diagnosis and therapy. In *Proceedings of the International Workshop on Expert Systems in Engineering*, Vienna. Springer Verlag, Lecture Notes in Artificial Intelligence, Vo. 462.
12. Heller U, Struss P (1996) Transformation of Qualitative Dynamic Models – Application in Hydro-Ecology. In *Proceedings of the 10th International Workshop on Qualitative Reasoning*, pages 83–92. AAAI Press.
13. Heller U, Struss P (1997) Conceptual Modeling in the Environmental Domain. In *Proceedings of the 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics*, volume 6, pages 147–152, Berlin, Germany.
14. Heller U, Struss P (1998) Diagnosis and therapy recognition for ecosystems - usage of model-based diagnosis techniques. In *12th International Symposium Computer Science for Environmental Protection (UI-98)*, Bremen.
15. Okasha S (2002) Philosophy of Science - A Very Short Introduction. Oxford University Press.
16. Poch M, Comas J et al (2004) Designing and building real environmental decision support systems. Environmental Modelling and Software 19(9):857–873.
17. Reiter R (1987) A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95.
18. Struss P (1998) Artificial intelligence for nature - why knowledge representation and problem solving should play a key role in environmental decision support. In O. Herzog and A. Günter, editors, *KI-98: Advances in Artificial Intelligence, LNAI 1504*. Springer Verlag, Berlin.
19. Wotawa F (2009) On the use of abduction as an alternative to decision trees in environmental decision support systems In *1st International Workshop on Intelligent Systems for Environmental Engineering and EcoInformatics (i-Seek)*, Fukuoka, Japan.

Supporting Decision Making in Maritime Environmental Protection with a Knowledge-based Education and Awareness Approach

Konstantinos Kotis

University of the Aegean, Dept. of Information and Communication Systems Eng., AI-Lab, Karlovassi, Greece, kotis@aegean.gr

Andreas Papasalouros

University of the Aegean, Dept. of Mathematics, Karlovassi, Greece, andpapas@aegean.gr

Nikitas Nikitakos

University of the Aegean, Dept. of Department of Shipping Trade and Transport, Chios, Greece, nnik@aegean.gr

Abstract In this paper we present an approach which aims to support learners in general and environmental decision makers in particular, towards effective decision making in maritime environmental pollution via education and awareness of specific maritime environmental pollution policies. We build on previous work concerning the automatic construction of multiple-choice questions from ontologies (automatic assessment) and extend it by integrating if-then rules towards building an environmental knowledge base for maritime pollution. Preliminary evaluation of this work is conducted with a prototype environmental pollution (focused on maritime pollution with oil) ontology in OWL and example rules in SWRL for capturing knowledge related to diagnosis, response and environmental-change events of oil spill pollution.

Introduction

Expert systems for environmental pollution have been around some time (e.g. Meech and Veiga 1997; Ceccaroni et al 2004; Harzikos et al 2008; Karatzas and

Kaltsatos 2007). AI researchers have been working on this topic integrating also new technologies coming from the Semantic Web e.g. (Ceccaroni et al 2004). Work has been done on SWRL to support decision making in knowledge bases for other domains such as Transportation (Gang et al 2008) or Dental domain (Seon and Hong-Gee 2006). Although decision making seems to be well supported on this area, to the best of our knowledge there isn't much that have been done to support environmental decision makers via education and awareness. Intelligent Tutoring Systems (ITS) provide direct customized instruction or feedback to learners whilst performing a task implementing "learning by doing". ITS have been recently proved proper candidates for tackling such issues, using technological advances of Artificial Intelligence techniques in the service of environmental awareness/education and decision making support.

ITS's consist of four different subsystems or modules: the interface module, the expert module, the student module, and the tutor module[1]. The interface module provides the means for the student (learner more generally) to interact with the ITS, usually through a graphical user interface and sometimes through a rich simulation of the task domain the student is learning (e.g., controlling a power plant or performing a medical operation). The expert module references an expert or domain model containing a description of the knowledge or behaviors that represent expertise in the subject-matter domain the ITS is teaching -- often an expert system or cognitive model. An example would be the kind of diagnostic and subsequent corrective actions an expert engineer takes when confronted with an oil pollution alarm at sea. The student module uses a student model containing descriptions of student knowledge or behaviors, including his *misconceptions* and *knowledge gaps*. An apprentice technician might, for instance, not know that an oil spill of 200 tones in a small area of sea surface is not a major oil spill event (knowledge gap) or he may believe that the designated area of oil spoil is small and no action is needed (misconception). A mismatch between a student's behavior or knowledge and the expert's presumed behavior or knowledge is signaled to the tutor module, which subsequently takes corrective action, such as providing feedback or remedial instruction. To be able to do this, it needs information about what a human tutor in such situations would do i.e. the tutor model (Koedinger and Corbett 2006).

An ITS is only as effective as the various models it relies on to adequately model expert, student and tutor knowledge and behavior[1]. Thus, building an ITS needs careful preparation in terms of describing the knowledge and possible behaviors of experts, students and tutors. This description needs to be done in a formal language in order that the ITS may process the information and draw inferences, automatically generating new knowledge as feedback or instructions. Therefore the knowledge contained in the models should be organized and linked to an inference engine. It is through the latter's interaction with the descriptive data that tu-

[1] http://en.wikipedia.org/wiki/Intelligent_tutoring_system

torial feedback is generated in order to support environmental decision making for diagnosis of environmental damage and selection of appropriate responses/actions.

In this paper we propose a built-up on our previous work concerning the proposal of an e-learning approach towards the development of an ITS which automatically constructs multiple-choice questions from any domain ontology. Such built-up is considered as an extension of the OWL knowledge base by integrating SWRL rules. SWRL (W3C 2004b) is a Rule based ontology language, allowing users to take advantage of inferencing new knowledge from existing OWL knowledge bases, towards an OWL/SWRL-based process. We use the maritime environmental pollution as an evaluation domain by representing knowledge needed to capture diagnosis, response and environmental-change events of oil pollution. Such domain is encoded in a prototype OWL ontology and is used in combination to SWRL rules to represent policies and decision making of environmental protection.

SWRL has been developed in order to extend OWL language expressivity, based on a combination of the OWL-DL and OWL Lite sublanguages of the OWL Web Ontology Language (W3C 2004a) with the Unary/Binary Datalog RuleML sublanguages of the Rule Markup Language. SWRL describes the knowledge of OWL ontology by highly abstract syntax expression, which realized the combination between the Horn-like rules and OWL Knowledge Base ($SHOIN(D)=\Sigma$). We use SWRL to formally express productive and deductive rules for diagnosis and response (diagnose and react) policies, in cases where OWL itself is not enough (we refer to the generic example of "parent(?x,?y) $\wedge$ brother(?y,?z) $\Rightarrow$ uncle(?x,?z)" rule) (W3C 2004b) and the additional expressivity power of SWRL is preferred (closer to human way of representing knowledge and easy way of deducing conclusions). The resulted combined knowledge base (Σ, P) is an integration of $SHOIN(D) = \Sigma$ and a finite set of rules P.

To the best of our knowledge, although some work has been done towards using SWRL in teaching strategies e.g. (Wang et al 2005), there is not any previous work that seamlessly, and in an automatic fashion, integrates an OWL-DL/SWRL knowledge base with an learning approach to support environmental decision making via education and awareness. In this paper we present a work-in-progress approach which utilizes an environmental ontology and rules (ITS expert model), a set of strategies for identifying the semantics of evaluation material in the form of multiple choice questionnaires (ITS teaching module) and a set of simple techniques for natural language generation (ITS interface module).

In the current version of the proposed approach, no student module is available, thus personalization or complex interaction with students (decision makers) is not supported. We conjecture that the approach can be used by beginners in the environmental pollution decision making domain. Such users do not need to be familiar with the underlining technology of ontologies and knowledge bases, and more important, they do not need to be experts in the domain of environmental pollution. Users must have obtained basic knowledge from text documents or oral

presentations related to the domain prior to their questionnaire-based assessment. Such basic knowledge is asserted in the knowledge base manually (currently by knowledge engineers in collaboration with domain experts). Automated population of the ontology with facts is out of the scope of this work.

The "EnvOPol" Knowledge Base

A knowledge base is a collection of models, stored facts and rules that can be used for problem solving. The "EnvOPol" knowledge base (built for experimentation reasons) integrates a prototype ontology concerning environmental pollution, focusing on maritime pollution by oil. The knowledge has been acquired from Web resources related to sea pollution Factsheets[2], consulting also the hierarchical description of environmental entities provided by the Eionet GEMET thesaurus[3]. Furthermore, domain experts and ontology engineers that have been participating in the experiment contributed their knowledge either informally or formally using ontology engineering tool Protégé[4] ver. 3.4, partially following the ontology engineering methodology HCOME (Kotis and Vouros 2006). An OWL-DL version of the prototype ontology may be viewed at http://www.icsd.aegean.gr/kotis/Ontologies/oilPollution.owl. OWL-DL language was selected due to the maximum expressiveness possible while retaining computational completeness (all conclusions are guaranteed to be computed), decidability (all computations will finish in finite time), and the availability of practical reasoning. Also, OWL-DL is a W3C standard language for Web Documents and applications. Due to space limitations we provide only semantics for a subset of the conceptualizations, in order to be able for readers to follow the examples (model, facts and rules) presented in this paper. A simple hierarchical caption of the ontology is presented in Figure 1.

A main concept is the oil pollution event *(oil_pollution_event* $\sqsubseteq$ *Event),* which may be of any type, based mainly on its severity importance (currently we have conceptualize *disastrous*, *significant* and *minor* events). Disastrous oil pollution events (*pollution_event_Disastrous_oil_spill* $\sqsubseteq$ *oil_pollution_event*) are defined as events that concern a large region of oil spill, and the severity of their oil spill and the severity of their spill volume is characterized as disastrous (*(oil_spill_region_size_on_photo* $\ni$ *"large")* $\sqcap$ *(has_oil_spill_volume_severity* $\ni$ *oil_spill_volume_severity_disastrous)* $\sqcap$ *(has_recovery_time_severity* $\ni$ *recovery_time_severity_disastrous)).*

2 http://www.ypte.org.uk/environmental-facts.php
3 http://eionet.eu.int/GEMET
4 http://protege.stanford.edu/

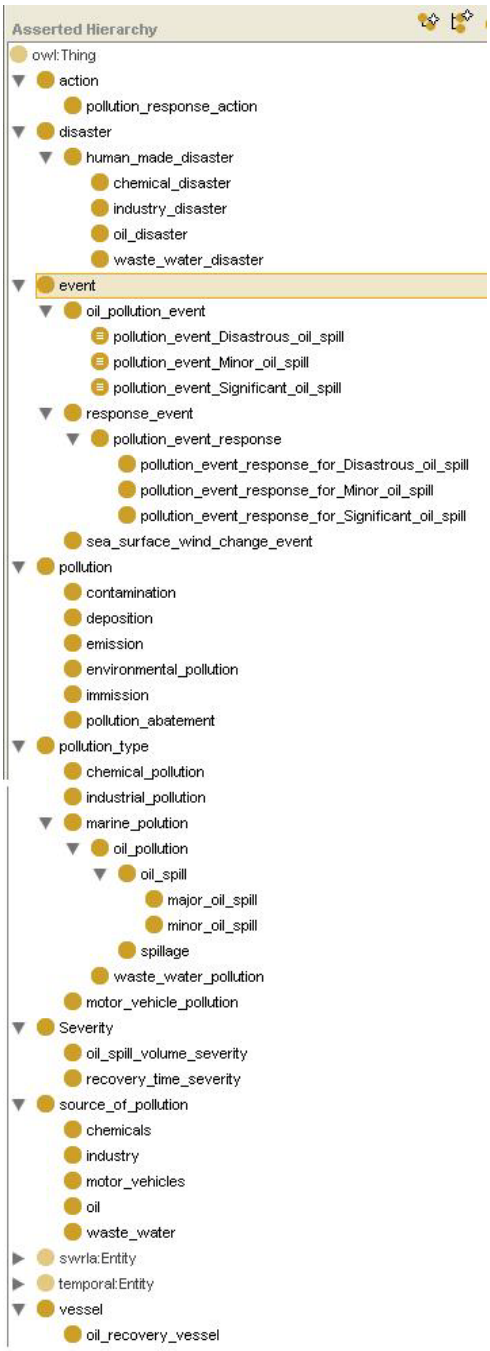

Asserted Hierarchy
owl:Thing
action
pollution_response_action
disaster
human_made_disaster
chemical_disaster
industry_disaster
oil_disaster
waste_water_disaster
event
oil_pollution_event
pollution_event_Disastrous_oil_spill
pollution_event_Minor_oil_spill
pollution_event_Significant_oil_spill
response_event
pollution_event_response
pollution_event_response_for_Disastrous_oil_spill
pollution_event_response_for_Minor_oil_spill
pollution_event_response_for_Significant_oil_spill
sea_surface_wind_change_event
pollution
contamination
deposition
emission
environmental_pollution
immission
pollution_abatement
pollution_type
chemical_pollution
industrial_pollution
marine_polution
oil_pollution
oil_spill
major_oil_spill
minor_oil_spill
spillage
waste_water_pollution
motor_vehicle_pollution
Severity
oil_spill_volume_severity
recovery_time_severity
source_of_pollution
chemicals
industry
motor_vehicles
oil
waste_water
swrla:Entity
temporal:Entity
vessel
oil_recovery_vessel

Fig. 1. A hierarchical caption of the ontology taken from Protégé tool

Similarly we define minor and significant oil spill pollution events. The severity of oil spill volume and of the recovery time are primitive classes that classify severity individual objects created for different measurements of recovery time (measured in years) or oil volume (measured in tonnes) respectively. For instance, the *recovery_time_severity_disastrous* individual object describes (with its properties inherited by the related class) the time needed to recover from an event with a disastrous severity i.e. *min_severity_value* property with a value of 100.

A response to an oil pollution event is described as another type of event *(pollution_event_response* $\sqsubseteq$ *Event).* Based on the severity of a pollution event, we distinguish different types of responses, each one initiating different actions for recovery *(* $\forall$*initiate_action. pollution_response_action).* Each response event is related to pollution event e.g. a pollution event response for a disastrous oil spill concerns a pollution event of a disastrous oil spill *(* $\forall$*concerns_event.pollution_event_Disastrous_oil_spill).* An inverse build-in OWL property (*inverseOf(concerns_event, concerns_response))* ensure that events and responses are related in both directions.

Finally, in order to be able to experiment with reasoning related to environmental change knowledge, another type of event is represented, the event of a wind change on the sea surface (*sea_surface_wind_change_event* $\sqsubseteq$ *Event*). Such an event is recorded by sensor input however in our case only simulation data is used for experimentation. Individuals of this event type are different recordings of sea surface wind speed (datatype property with allowed values of "low", "medium", "high") at specific time and date of a specific location.

Using the OWL-DL axioms specified in the KB, we are able not only to assert specific oil pollution events that are fully identified (and assign a specific pollution event response) but also to infer new events by computing inferred types. The inference of such knowledge is achieved via a reasoning mechanism (Pellet 1.5[5]) and the proper design of defined classes (necessary and sufficient conditions). However, as already stated, the "EnvOPol" knowledge base was extended with deductive and production rules in order to represent knowledge for diagnosis and response (diagnose and react) using the SWRL formalism. Some example rules are provided below:

Example rule set A: ("*discover which oil pollution events are disastrous based on their severity and oil spill size region on a satellite photo*" and "*retrieve the responses available for such a disastrous event*"):

1. oil_pollution_event(?e)
   $\wedge$ has_oil_spill_volume_severity(?e, oil_spill_volume_severity_disastrous)
   $\wedge$ has_recovery_time_severity(?e, recovery_time_severity_disastrous)

[5] http://clarkparsia.com/pellet/

∧ oil_spill_region_size_on_photo(?e, "large")
→ pollution_event_Disastrous_oil_spill(?e)

2. pollution_event_Disastrous_oil_spill(?e)
   ∧ pollution_event_response_for_Disastrous_oil_spill(?r)
   → concerns_response(?e, ?r)

3. concerns_response(?e, ?r) → sqwrl:selectDistinct(?r)

Example rule set B: ("discover which oil pollution events are minor based on their severity and oil spill size region on a satellite photo" and *"select those which need to be upgraded to disastrous because of a sea surface wind change event with specific characteristics")*:

1. oil_pollution_event(?e)
   ∧ has_oil_spill_volume_severity(?e, oil_spill_volume_severity_minor)
   ∧ has_recovery_time_severity(?e, recovery_time_severity_minor)
   ∧ oil_spill_region_size_on_photo(?e, "small")
   → pollution_event_Minor_oil_spill(?e)

2. sea_surface_wind_change_event(?w) ∧ time(?w, ?wTime) ∧ date(?w, ?wDate)
   ∧ location(?w, ?wLocation) ∧ pollution_event_Minor_oil_spill(?e)
   ∧ time(?e, ?eTime) ∧ date(?e, ?eDate) ∧ location(?e, ?eLocation)
   ∧ windSpeed(?w, ?sNew) ∧ windSpeed(?e, ?sOld)
   ∧ swrlb:notEqual(?sNew, ?sOld) ∧ swrlb:matches(?sNew, "high")
   ∧ swrlb:lessThanOrEqual(?eDate, ?wDate)
   ∧ swrlb:lessThanOrEqual(?eTime, ?wTime)
   ∧ swrlb:matches(?eLocation, ?wLocation)
   → sqwrl:selectDistinct(?e) ∧ upgrade_to_disastrous(?e, "true")

In this human-readable syntax, a rule has the form: *antecedent* → *consequent,* where both *antecedent* and *consequent* are conjunctions of atoms written $a_1 \wedge ... \wedge a_n$.

The "QuGAR-OWL" ITS approach

QuGAR-OWL (*A*utomatic *G*eneration of *Qu*estion items from *R*ules and *OWL* ontologies) is an e-learning approach towards an ITS that generates multiple choice questionnaires from populated OWL ontologies in an automatic fashion (Papasalouros et al 2008). The approach utilizes ontologies that represent both domain and multimedia knowledge. Multimedia questionnaires are currently restricted to items with images. For evaluation and experimental purposes we have produced results with a number of domain ontologies for text-based questionnaires. The approach is open to any source of knowledge that can be mapped to OWL semantics

and of course to any source that already uses OWL semantics to represent its knowledge. Heterogeneous and distributed domain-specific knowledge can also be automatically transformed in a QuGAR-OWL-generated questionnaire, given that there is an OWL model that these resources can be mapped to (and aligned).

Certain strategies have been identified and used for selecting the correct answers in question items, as well for selecting distractors (Kehoe 1995). The selected strategies are analytically presented in (Papasalouros et al 2008). Below we provide a simple strategy and a related example question automatically generated for the maritime environmental pollution ontology.

- Strategy A (text-based):

Choose individuals which are not members of a given class, provided that they are members of one of its superclasses. More specifically, if *A(a)* for some *a,* then correct answer is: *A(a)*. For the distractors selection, we assume that *B* is a superclass of *A*. Then, if *B(b)*, $b \neq a$ and *b* is not an individual of *A*, then *A(b)* is a distractor.

- Generated Question A:

Which of the following sentences is true?

A. PERM01 is a pollution event response for Minor oil spill. (C)
B. PERS01 is a pollution event response for Minor oil spill. (D)
C. PERD01 is a pollution event response for Minor oil spill. (D)
D. PERD02 a pollution event response for Minor oil spill. (D)

In the above, only choice A is a correct answer, indicated with (C), since PERM01 is an individual of ontology class *pollution_event_response_for_Minor_oil_spill.* The other choices, indicated with a (D), are distractors, containing individuals which belong to disjoint sibling classes of the above class (OWL disjointWith axiom has been utilized).

Preliminary work on extending QuGAR-OWL approach to handle rules also (specifically SWRL rules) used with problem solving related domains such as the environmental protection/pollution domain, proves that it can be used as a support tool for improving the effectiveness of decision making via education and awareness of diagnosis/response policies. More specifically, we identify a number of new strategies that extend our previous work with text-based and multimedia-based strategies. In this paper we present the first two rule-based strategies (Strategy B and Strategy C).

- Strategy B (rule-based):

Given that $d1 \wedge d2 \wedge ... \wedge dm \rightarrow v1 \wedge v2 \wedge ... \wedge vk$ is a rule in the knowledge base, where x is a variable and C is a class, and one of the atoms $v1,v2,...,vk$ in the head of the rule is in the form C(x), then a multiple choice question item can be formed

as follows: The rule provides the semantics for the correct answer and distractors are selected among *disjoint siblings* of or among *subclasses* of C. As an example we assume that the following rule exists in the knowledge base:

1. oil_pollution_event(?e)
   ∧ has_oil_spill_volume_severity(?e, oil_spill_volume_severity_disastrous)
   ∧ has_recovery_time_severity(?e, recovery_time_severity_disastrous)
   ∧ oil_spill_region_size_on_photo(?e, "large")
   → pollution_event_Disastrous_oil_spill(?e)

Based on concept *pollution_event_Disastrous_oil_spill,* which appears in the head of the above rule, this strategy generates question items as in the following example.

- Generated Question B:

If an oil pollution event has disastrous oil spill volume severity and disastrous recovery time and large region size on photo, then the pollution event is a(n):

A. Disastrous oil spill pollution event (C)
B. Oil spill pollution event
C. Minor oil spill pollution event (D)
D. Significant oil spill pollution event (D)

In the above example, the correct answer is indicated by (C), while the wrong answers (distractors) are indicated by (D) (for presentation reasons only in the paper).

- Strategy C (rule-based):

For a rule in the form $d1 \wedge d2 \wedge ... \wedge dm \rightarrow v1 \wedge v2 \wedge ... \wedge vk$, if one of the atoms $d1,d2,...,dk$ in the body of the rule is in the form C(x), where x is a variable and C is a class, then generate a sentence based on the rule as correct answer. Distractors are generated by substituting C with one of its *super-classes* or one of its *disjoint siblings*.

As an example, classes *pollution_event_Disastrous_oil_spill(?x)* and *pollution_event_response_for_Disastrous_oil_spill(?y)* appear as atoms in the head of the following rule:

1. pollution_event_Disastrous_oil_spill(?e)
   ∧ pollution_event_response_for_Disastrous_oil_spill(?r)
   → concerns_response(?e, ?r)

- Generated Question C:

Which of the following is correct?

A. A disastrous pollution oil spill event concerns a disastrous pollution oil spill event response. (C)
B. A pollution oil spill event concerns a disastrous pollution oil spill event response (D)

C. A disastrous pollution oil spill event concerns a pollution oil spill event response (D).
D. A minor oil spill event concerns a pollution oil spill event response (D).

In current version of QuGAR-OWL, natural language generation is based on the names of ontology classes and properties, provided that they follow certain conventions. Future work should tackle the problem of generating natural language items from domain-specific OWL and SWRL semantics with further study of OWL-to-NLG techniques (e.g. the work presented in Karakatsiotis et al (2007)).

Conclusion and Future Work

In this paper, building on our previous work on ITS, we present preliminary results of novice and original work towards a) a maritime environmental pollution knowledge base (model, facts, rules), b) the extension of ITS to handle rules for the automatic generation of multiple choice questions, c) the use of the proposed ITS extension to support decision making via education and awareness in the domain of maritime environmental protection. Since this is a work in progress, we need to implement and evaluate the rule-based question generation strategies within the prototype intelligent tutoring system. Furthermore, issues such as interaction and feedback should be explored since currently we only consider interaction within the task of capturing multimedia knowledge by annotating images, and we generate feedback only from the correct/wrong answers. In the current version of the tool, no student module is available, thus personalization or complex interaction with students is not supported. Furthermore, users must obtain basic knowledge from text documents or oral presentations related to the domain. Such basic knowledge is asserted in the knowledge base manually (currently by knowledge engineers). Future work concerns the active participation of decision makers in the knowledge base development process, following a human-centered and collaborative ontology engineering approach supported by Wiki-based argumentation technology. Finally, the problem of generating natural language items from domain-specific OWL and SWRL semantics should be tackled with further study of OWL-to-NLG techniques.

References

Ceccaroni L, Cortes U, Sanchez-Marre M (2004) OntoWEDSS: augmenting environmental decision-support systems with ontologies. Environmental Modelling & Software 19, pp. 785–797

Gang C, Qingyun D, Hongli M (2008) The Design and Implementation of Ontology and Rules Based Knowledge Base for Transportation. International Conference on Computer Science and Software Engineering, pp.1035-1038

Hatzikos EV, Tsoumakas G, Tzanis G et al (2008) An Empirical Study on Sea Water Quality Prediction. Knowledge Based Systems 21(6):471-478

Karakatsiotis G, Galanis D, Lampouras G et al (2008) NaturalOWL: Generating Texts from OWL Ontologies in Protege and in Second Life. System demonstration, 18th European Conference on Artificial Intelligence, Patras, Greece

Karatzas K, Kaltsatos S (2007) Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. Simulation Modelling Practice and Theory, 15(10):1310-1319

Kehoe J (1995) Writing multiple-choice test items. Practical Assessment, Research & Evaluation. Vol. 4 No. 9, retrieved February 2009 from http://pareonline.net/getvn.asp?v=4&n=9

Koedinger R, Corbett A (2006) Cognitive Tutors: Technology bringing learning science to the classroom. In Sawyer, K., The Cambridge Handbook of the Learning Sciences, Cambridge University Press, pp. 61–78

Kotis K, Vouros AG (2006) Human-Centered Ontology Engineering: the HCOME Methodology. International Journal of Knowledge and Information Systems (KAIS), 10(1): 109-131

Meech A, Veiga M (1997) Predicting the impact of mercury pollution with a fuzzy expert system. Systems, Man, and Cybernetics, Vol. 2, IEEE press, pp.1056–1061

Papasalouros A, Kotis K, Kanaris K (2008) Automatic generation of multiple-choice questions from domain ontologies. IADIS e-Learning 2008 (eL 2008), Amsterdam

Seon P, Hong-Gee K (2006) Dental Decision Making on Missing Tooth Represented in an Ontology and Rules. Springer Berlin / Heidelberg 2006, Vol. 4185, pp. 322-328

W3C (2004a) OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-features/

W3C (2004b) SWRL: A Semantic Web Rule Language Combining OWL and RuleML. http://www.w3.org/Submission/SWRL/

Wang E, Kashani L, Kim YS (2005) Teaching Strategies Ontology Using SWRL Rules. International Conference on Computers in Education (ICCE), Singapore

An Environmental Diagnosis Expert System

Mihaela Oprea, Daniel Dunea

University Petroleum-Gas of Ploiesti, Department of Informatics,
Bd. Bucuresti nr. 39, Ploiesti, 100680, Romania

Abstract. The paper presents an expert system, SBC-MEDIU, developed` for environmental diagnosis. Two modules of the system are discussed in detail: the module for air pollution analysis and dispersion assessments, SBC-AIR, and the module for soil erosion risk assessments, SBC-SOIL. Both modules provide at the end of the analysis the environmental diagnosis result and the associated allert code. Some experimental results obtained so far are also described.

1 Introduction

An efficient environmental management system has to include software tools for air, water and soil pollution diagnosis. In the recent years, several artificial intelligence (AI) tehniques were applied to environmental diagnosis (such as knowledge-based systems, expert systems, case-based reasoning, artificial neural networks, data mining). Some systems recently reported in the literature, that are based on AI techniques, could be found in [1], [8] and [12]. Usually, these systems are specialized either to air, or water or soil analysis. Most of the systems are dedicated to air and water analysis, and few of them to soil analysis. Our research work involved the integration of all three main components of the environment analysis (air, water and soil) into an environmental diagnosis expert system. The reason of developing an integrated system is the dependency that can appear between air, water and soil pollution. Due to the high complexity of such a system we have concentrated our efforts to the inclusion of the expert knowledge for some types of environmental diagnosis. We have developed an expert system, SBC-MEDIU, that has three modules, SBC-AIR, SBC-SOIL, and SBC-WATER, for air, soil and water pollution analysis. In this paper we shall focus on two modules of the system, SBC-AIR, for air pollution analysis and dispersion assessments, and SBC-SOIL, for soil erosion risk assessments. The module SBC-WATER makes a surface water pollution analysis and is described in [9]. The expert system can be used as an educational tool for students, and also, as a decision support tool for the local Environmental Agencies.

2 The architecture of the expert system

The expert system SBC-MEDIU has a modular architecture, presented in Figure 1. The three modules of the system are SBC-AIR, SBC-WATER and SBC-SOIL, each

having a knowledge base with specific expert knowledge represented under the form of production rules. All modules are using the inference engine of VP-Expert, an expert system generator [2]. The purpose of module SBC-AIR is to make a diagnosis of the air pollution, taking into account different parameters such as the air pollutants concentrations, and some meteorological data. Also, this module is doing air pollution dispersion assessments. The module SBC-WATER makes surface water pollution diagnosis, while the module SBC-SOIL is doing soil erosion risk assessments.

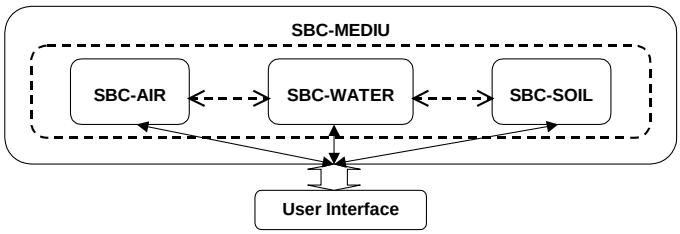


Fig. 1. The modular architecture of the expert system SBC-MEDIU.

Figure 2 shows the basic components of each module of the system SBC-MEDIU: a knowledge base, the inference engine, and the databases with standards (for a clean air, water and soil), and timeseries of specific measurements (meteorological, concentrations of air pollutants, water pollutants, and soil pollutants etc).

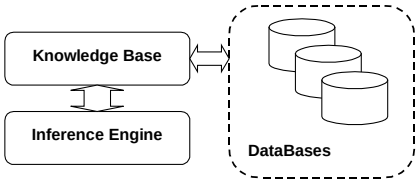


Fig. 2. The basic components of each module of SBC-MEDIU.

3 The module SBC-AIR

The analysis made by the module SBC-AIR has the following parts: (1) the air quality index assessment, (2) the air pollution analysis, and (3) the air pollution dispersion assessment. The knowledge base used in the air pollution analysis includes rules from the knowledge base of the expert system DIAGNOZA_MEDIU [7]. In this section we describe parts (1) and (3) of the analysis.

Air Quality Index

A comprehensive environmental assessment system can facilitate tracking and benchmarking of the air quality performance, providing a tool for measuring any continuous improvement [6]. Legislation, through regulations and norms, establishes the ambient concentrations of pollutant to which the receptor is limited. Air quality criteria delineate the effects of air pollution.Standards for air pollution (e.g. Romanian O.M. 592/2002) are concentrations over a given time period that are considered to be

acceptable in the light of what is known about the effects of each pollutant on health and on the environment. They can also be used as a benchmark to see if air pollution is getting better or worse. An exceedence of a standard is a period of time where the concentration is higher than that set down by the standard. In order to make useful comparisons between pollutants, for which the standards may be expressed in terms of different averaging times, the number of days on which an exceedence has been recorded must be reported by the expert system. After making specific inferences, the expert system provides reliable answers (solutions /results /decisions) to solve air pollution aspects based on the information identified in the knowledge base. SBC-AIR established the air quality indicator using recorded imissions pollutant concentrations (15 minutes - sulfur dioxide, hourly - ozone, nitrogen dioxide and 24 hours values - PM10 fraction), introduced by the user, providing warning capabilities related to the potential impact of air pollution on sensible individuals: (1) *Green Code* – levels 1, 2 and 3 - Effects are not noticed by individuals sensitive to air pollutants; (2) *Yellow Code* - levels 4, 5 and 6 - Medium effects, may be noticed amongst sensitive individuals, do not require intervention; (3) *Red Code* - levels 7, 8 and 9 - Significant effects may be noticed by sensitive individuals and action to avoid or reduce these effects may be needed; (4) *Maximum Code* - level 10 - The effects on sensitive individuals described for 'High' levels of pollution are worsening - Maximum Alert threshold. Figure 3 shows a screenshot of the SBC-AIR module run.

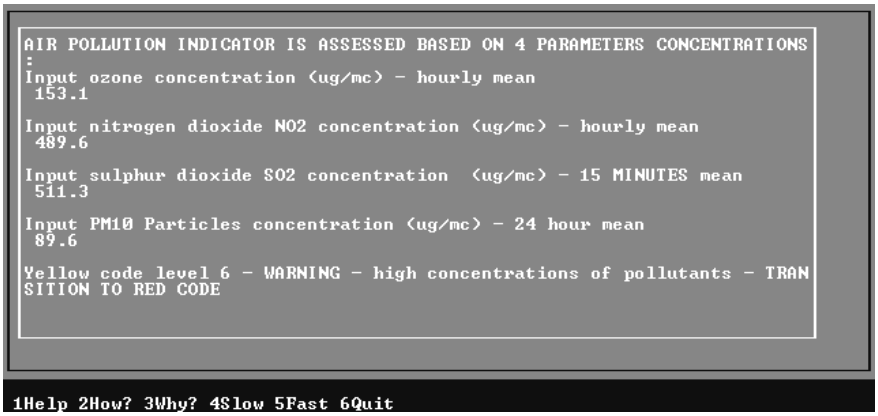


Fig. 3. SBC-AIR quality index result: Yellow code level 6.

Air pollution dispersion

SBC-AIR required the integration of a simple tool assessing what happens to pollutants in the atmosphere after they are discharged from stationary emission sources. For the present, the stochastically based Gaussian type model is the most useful in modeling for regulatory control of pollutants [13]. The Gaussian plume model provided rough estimates of pollutant ground level concentrations (imissions) in the absence of monitored data, allowing Air quality index calculations. Therefore, the final objective of the calculations using SBC-AIR capabilities was to determine if an emission will result in ground ambient concentrations which exceed air quality standards that have been set by reference to air quality criteria.

Local meteorological processes and topography control the amount of pollution as it spreads and reaches ground level. Inversions are the principal meteorological factor present when air pollution situations occur, such as: Surface or Radiation Inversions, Evaporation Inversion, Advection Inversion and Subsidence Inversion [13]. Pollutants are transported by wind and turbulence, and they may undergo chemical transformations before being deposited on the earth's surface [10]. The way in which atmospheric characteristics affect the concentration of air pollutants after they leave the source can be viewed in three stages: effective emission height, bulk transport of the pollutants and dispersion of the pollutants. Several factors affect the plume, including the effective height (*H*) of emission, which is a measure of how high the pollutants are emitted into the atmosphere directly above the source. The height is dependent upon source characteristics and atmospheric conditions. The turbulence caused by the air flow over the surface and by possible instability governs the diffusion of the plume contents. Visible plumes are indicators of stability conditions. Five special models have been observed and classified by the following names: looping (unstable), coning (stable), fanning (slight stable), fumigation (moderately stable) and lofting (dispersal upward). Recognition of these conditions is helpful to the modeler and in gaining an additional understanding of dispersion of pollutants [14]. The major factors that characterize the emission source are: composition, concentration, and density, velocity of emission, temperature of emission, pressure of emission, the diameter of emitting stack or pipe, and the effective height of emission.

Knowing the location of the source relative to the receptor and its characteristics would allow calculating the concentration at a particular downwind receptor using a dispersion model. A Gaussian mathematical model was used, which incorporates source-related factors and meteorological factors to estimate pollutant concentration from the stationary sources. The model is applicable to continuous sources of gases and particulates less than 10 μm in diameter estimating the plume concentrations over horizontal distances of 10^2 to 10^4 m [6].

$$C(x,y,z,H) = \frac{Q}{2\pi\sigma_y\sigma_z U} \cdot \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \cdot \left[\exp\left(-\frac{(z-H)^2}{2\sigma_z^2}\right) + \exp\left(-\frac{(z+H)^2}{2\sigma_z^2}\right)\right] \quad (\mathbf{1})$$

where: *C(x, y ,z, H)* - pollutant concentration at any point in the plume (μg m^{-3});

Q – emission rate of pollution from the source (μg s^{-1}) ;

H - the effective height of the pollution source, function of the chimney height, its diameter, speed and temperature of gas exhaustion, and air layering;

σ_y, σ_z – horizontal and vertical standard deviations of the pollutant concentration distributions in the y and z directions;

U - the wind speed measured at the height of the source (m s^{-1}).

The Gaussian model (equation 1) and related additional algorithms, were used to allow SBC-AIR to compute ground level concentrations at any receptor point (X_o,Y_o) in a polluted region resulting from each of the isolated sources in the emission inventory. Despite its limitations in building a complex image of the bulk transport of pollutants and their chemical transformations before being deposited on the earth's surface, and the building downwash and surrounding topographical features effects, the system provided preliminary information of main air pollution dispersion factors status and trends based on real data from different emission sources in the Dâmboviţa County. We have selected as exemplification the emission source of Doiceşti power

plant stack (1074.6 and 392.5 $m^3 s^{-1}$) from the 17 stationary sources with mass flows greater than 4 $m^3 s^{-1}$.

The first step of the program is to compute lateral and vertical dispersion coefficients for each atmospheric stability category: Very Unstable, Moderately Unstable, Slight unstable, Neutral, Somewhat stable, and Stable. The user is asked to input the distance from the stationary source (km) of the point required to assess the pollutant concentration. Plume shapes are comprehensive indicators to estimate and select the atmospheric stability class. A screenshot of this step run is shown in Fig. 4.

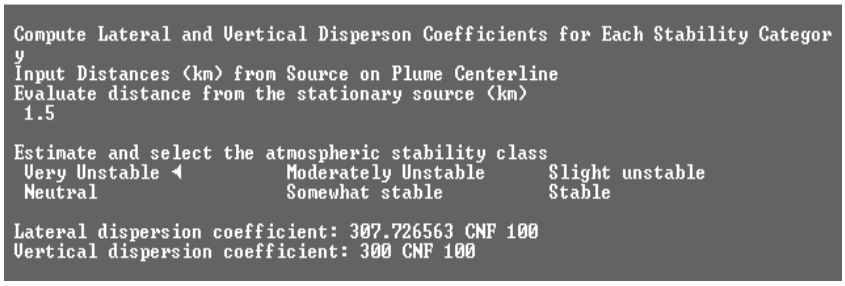


Fig. 4. SBC-AIR screenshot of the first step run.

The second step involves the introduction of the emission source characteristics such as: gas exit velocity (m/s), stack diameter (m), gas exit temperature and ambient temperature in Kelvin degrees. A screenshot of this step run is shown in Figure 5.

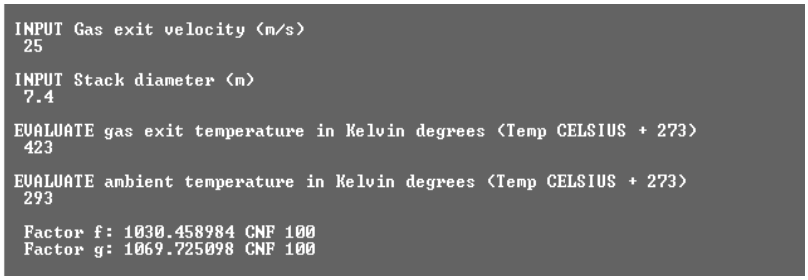


Fig. 5. SBC-AIR screenshot of the second step run.

Several more parameters are required to be introduced: wind velocity measured at 10 meter height (m/s), the stack height of the emission stationary source (m), and the source emission rate (g/s). Consequently, SBC-AIR computes the ground level concentrations (e.g. 208.5 μg m^{-3}) at any receptor point from the emission source on plume centerline in the selected region resulting from each of the isolated sources in the emission inventory. These steps are shown in Figure 6.

If the main composition of the emission from the analyzed stack is known, by comparing the result of the estimated concentration of ground-level pollution on plume centerline at selected distance from source with the standard limit values, SBC-AIR provides the estimation of the pollutant exceeding. Aggregation of the results obtained for different distances permits the graphical visualization of the pollutant dispersion (see Figure 7).

```
Input wind velocity measured at 10 meter height (m/s)
 11

INPUT Stack height of the emission stationary source (m)
 200

STACK HEIGHT Effect (m): 554.239990 CNF 100
INPUT Emission rate (g/s)
 3500

Estimated Concentration of Ground-Level Pollution (mmg/m3) on Plume Centerline
 at Selected Distance (km) from Source:
208.504517 CNF 100
Select the main composition of the emission from the analyzed stack
 SO2                    NOx ◄                   PM10
 CO

IF THE INPUTTED EMISSION WAS NITROGEN OXIDES THEN THE POLLUTION EXCEEDED
 STANDARD LIMITS FOR HOURLY VALUES - Warning!
Exceeding of the pollutant concentration compared to standard limits (mmg/m3):
 8.504517 CNF 100
```

Fig. 6. SBC_AIR screenshot of the last step run.

Smoke plume from stack is often trapped in the radiation inversion layer at night and then brought to the ground in fumigation during morning hours. This converts into high ground-level concentration. With moderately unstable condition, pollution is transported downward toward ground level. In the looping pattern situation, the sinusoidal path may bring the plume content to ground level close to the emitting source. Figure 7 highlights this situation showing a maximum ground concentration at 1.5 km far from the power plant stack. This is a dangerous situation for the inner-locality residents suffering from asthma or other respiratory illnesses. In this case, the wind speed regime does not have a significant influence on the ground concentration.

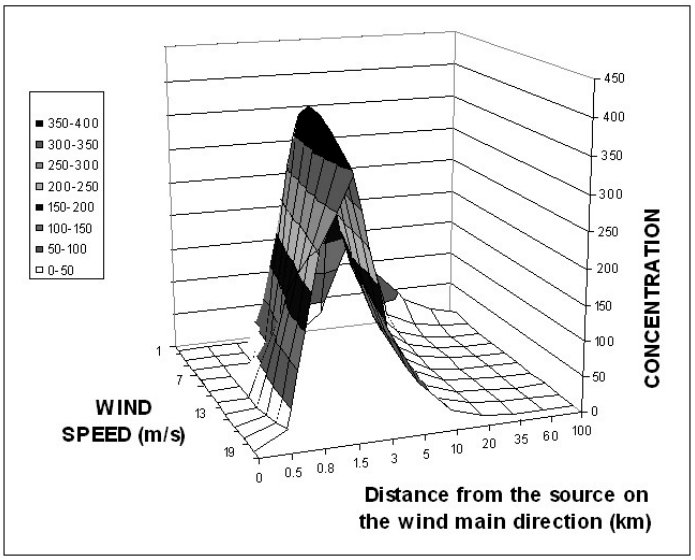


Fig. 7. Pollutant ground concentration (μg m^{-3}) in unstable atmospheric conditions for Doiceşti power plant stack using SBC-AIR dispersion assessment.

4 The module SBC-SOIL

Soil erosion is a major environmental threat to the sustainability and productive capacity of agriculture [4]. During the last 40 years, nearly one-third of the world's arable land has been lost by erosion and continues to be lost at a rate of more than 10 million hectares per year [11]. Average rates for soil loss have been estimated at 17 tones/ha per year in the United States and Europe, and 30-40 tones/ha per year in Asia, Africa and South America, mainly due to inadequate agricultural land use [11]. A recent review on erosion models and the quality of spatial predictions [3], states the great difficulties associated with calibrating and validating spatially distributed soil erosion models. It is mentioned that this is due to the large spatial and temporal variability of soil erosion phenomena and the uncertainty associated with the input parameter values used in models to predict these processes. Jetten et al. conclude that the construction of more complete and therefore more complex models will not overcome this problem; rather the situation may be improved by using more spatial information for model calibration and validation, and by using models that describe the dominant processes operating in a given landscape.

SBC-SOIL is a simplified assessment tool for soil erosion risk that assists novice application users in evaluating environmental problems for various local weather and geo-morphological conditions. It was developed by using the elements of universal erosion formula: pluvial intensity (aggression), soil erosion capacity, vegetation and cropping system influence, versants characteristics and adding surface runoff effect (soil permeability and surface slope). Validation and verification steps were done according a special attention to the input and control variables definition, interface conditions definition, rule base structure, inference rules design and their transformation to actions – consequently, the expert decision elaboration.

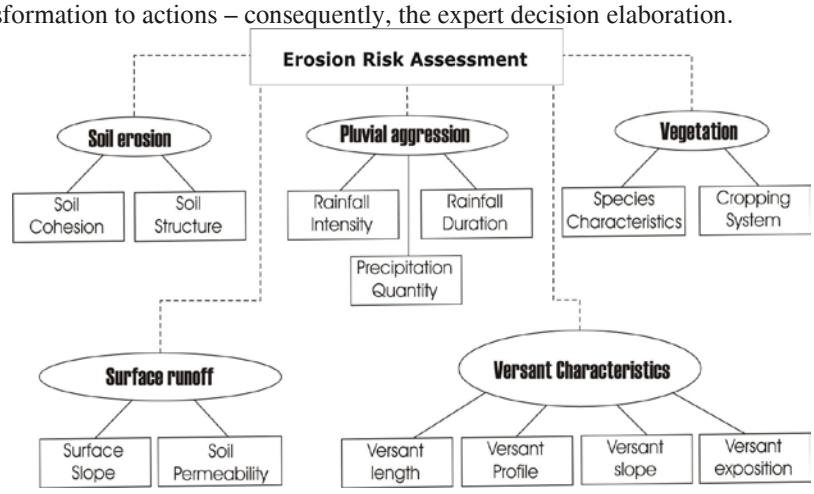


Fig. 8. The decision tree for heuristic assessment of global erosion risk of a specific site.

Figure 8 highlights the SBC-SOIL inner structure, consequently the decision tree that provides the evaluation of global erosion risk from a specific location based on the risk analysis of the individual components. The development of SBC-SOIL has

relied on the elements that constitute the erosion universal formula, adding surface runoff effects on erosion. Production rules were conceived for the following components: soil erosion capacity (9 rules), pluvial aggression (10 rules), vegetation and cropping system influence (13 rules) and versant characteristics (10 rules), adding surface runoff risk estimation (9 rules). The final rule set that evaluates the global erosion risk contains 5 rules. We show the run of SBC-SOIL step by step.

1) *Soil erosion* capacity is influenced by factors such as the structural soil aggregates dimension and water stability, granular structure, soil volume weight etc. Two factors were considered to empirically estimate this parameter: particles cohesion and soil structure. Figure 9 shows the screenshot of the system run for soil erosion analysis.

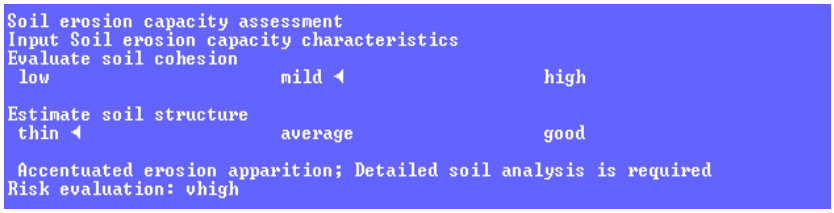


Fig. 9. Soil erosion analysis – screenshot of system run.

2) *Pluvial aggression* is assessed using three variables according to end-user estimation, as follows: rainfall intensity, rainfall duration and precipitation quantity. Figure 10 shows a screenshot of pluvial aggression analysis.

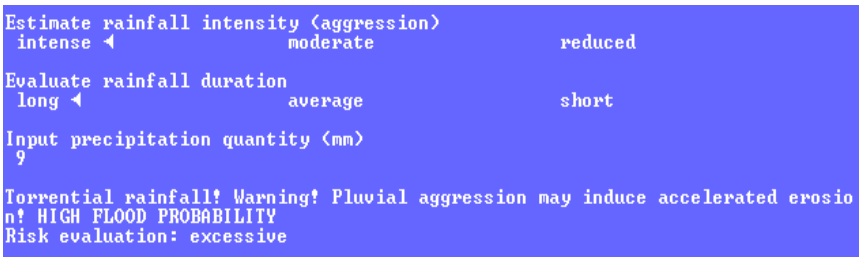


Fig. 10. Pluvial aggression analysis – screenshot of system run.

3) *The vegetation and cropping system* or vegetation arrangements influence considered the grouping of species by erosion sensibility, and cropping system by erosion protection degree. Figure 11 shows a screenshot of vegetation analysis.

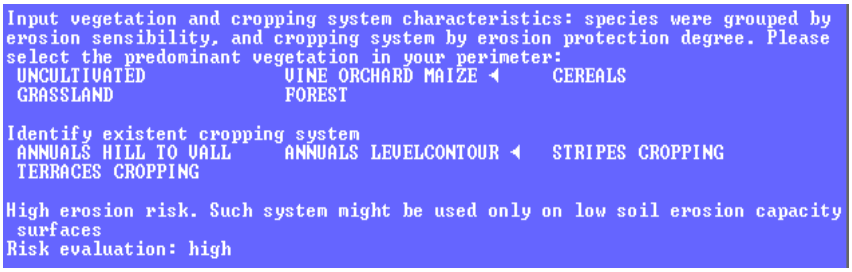


Fig. 11. Vegetation analysis – screenshot of system run.

4) *Versant characteristics effect* on erosion risk was estimated based on versant length, slope, profile form and exposition. Figure 12 shows a screenshot of versant characteristics analysis.

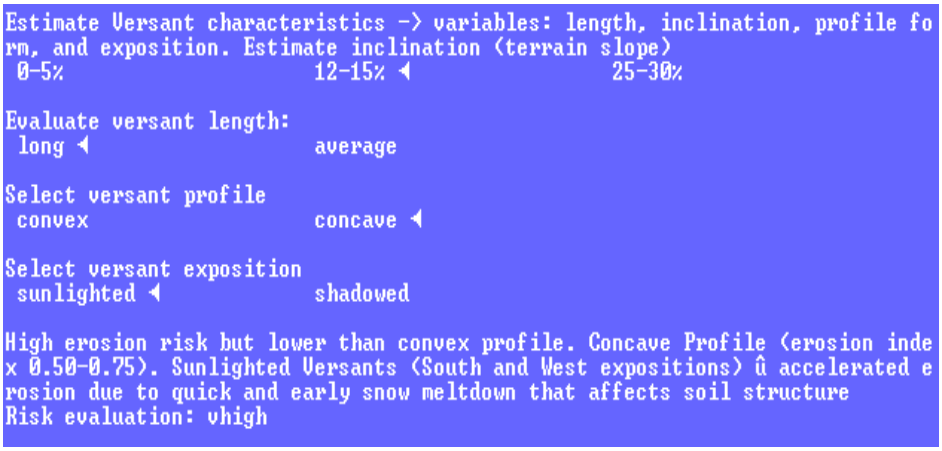


Fig. 12. Versant characteristics analysis – screenshot of system run.

5) *The Surface Runoff Class* site characteristic determined from the relationship of the soil permeability class and field slope was adapted from the Soil Survey Manual (1993) and was included in SBC-SOIL in order to increase the system complexity. The erosion risk is established based on runoff classification from soil permeability category that characterize the infiltration process and surface slope. Table 1 shows the erosion risk assessment, while the decision table for the surface runoff effect on erosion risk is given in Table 2.

Table 1. Erosion Risk assessment in SBC-SOIL based on runoff classification from soil permeability category and surface slope, and the decision table of extracted specific rules.

SLOPE	Very rapid	Moderately Rapid and Rapid	Moderately Slow and Moderate	Slow	Very Slow
Concave	N	N	N	N	N
<1	N	N	N	L	M
1-5	N	VL	L	M	H
5-10	VL	L	M	H	VH
10-20	VL	L	M	H	VH
>20	L	M	H	VH	VH

Soil permeability Classes (mm/h):
Very slow < 0.15 mm/h
Slow 0.15-0.5 mm/h
Moderately slow 0.5-1.5 mm/h
Moderate 1.5 – 5 mm/h
Moderately rapid 5-15 mm/h

RISK: *N-negligible*
M-medium
VL – very low
L - low
H- high
VH-very high

Table 2. Decision table for the surface runoff effect on erosion risk.

RULE	SURFACE SLOPE	SOIL PERMEABILITY	RISK
A1	REDUCED	SLOW	**NEGLIGIBLE**
A2	AVERAGE	SLOW	**MODERATE – RELATIVELY HIGH**
A3	ACCENTUATED	SLOW	**HIGH-VERY HIGH**
A4	REDUCED	MODERATE	**NEGLIGIBLE**
A5	AVERAGE	MODERATE	**MODERATE**
A6	ACCENTUATED	MODERATE	**RELATIVELY HIGH**
A4	REDUCED	RAPID	**NEGLIGIBLE**
A5	AVERAGE	RAPID	**REDUCED**
A6	ACCENTUATED	RAPID	**MODERATE**

Figure 13 shows a screenshot of the system run for surface runoff analysis.

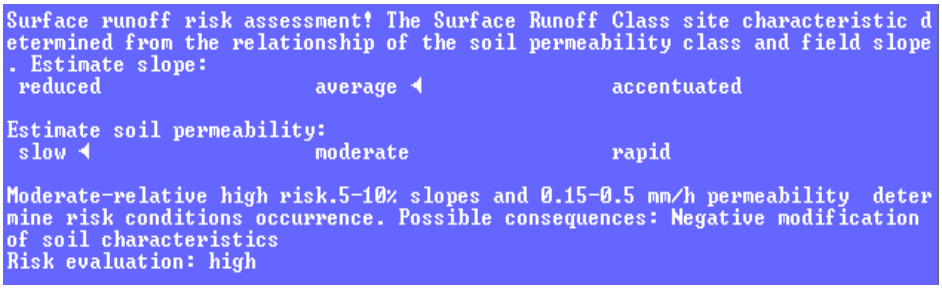


Fig. 13. Surface runoff analysis – screenshot of system run.

6) SBC-SOIL provides the final analysis giving comprehensive answers concerning the global erosion risk and brief descriptions of the negative consequences which might occur. Several color codes were considered to facilitate the understanding of final results: (1) *RED CODE* – Maximum Risk; (2) *ORANGE CODE* – Critical Risk; (3) *YELLOW CODE* - High Risk (with two levels, 1 and 2); (4) *GREEN CODE* – Moderate or reduced Risk; (5) *IDEAL CONDITIONS* – Absence of Risk or minimal Risk.

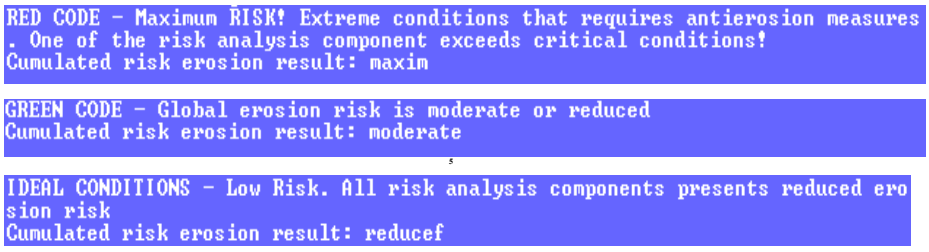


Fig. 14. Expert messages containing the final estimation of erosion risk using artificial reasoning.

Table 3 shows the decision table used by the module SBC-SOIL for the global evaluation of erosion risk.

Table 3. Decision table for the global evaluation of erosion risk.

		6_1	6_2	6_3	6_4	6_5
Operator		**OR**	**OR**	**AND**	**AND**	**OR**
Soil erosion capacity	erod	excessive	vhigh	high	reduced/vreduced	high
Pluvial aggression	stand	excessive	vhigh	high	reduced/vreduced	high
Vegetation	veg	excessive	vhigh	high	reduced/vreduced	high
Versant characteristics	versant	excessive	vhigh	high	reduced/vreduced	high
Surface runoff	risk	excessive	vhigh	high	reduced/vreduced	high
Global risk assessment	**scop**	*maxim*	*excessive*	*high*	*reduced*	*high*
	code Display	Red	Orange	Yellow level 1	IDEAL CONDITIONS	Yellow level 2 **/ELSE** *moderate* Green code

The rules from the knowledge base of module SBC-SOIL were generated from the decision tables 1, 2, 3, as well as from the existing speciality literature [3], [4] and [5].

The module SBC-SOIL facilitates the understanding of the complex relationships among the main factors that are responsible for the apparition, development and acceleration of various surfaces erosion process. It has proved to be a versatile diagnosis tool for the global erosion risk, providing the user with the ability to perform *a posteriori*, detailed analysis of the main erosion factors that presents high risks, using SBC-SOIL resources from its knowledge base.

5 Conclusion and future work

Environmental components (air-water-soil) are characterized by the high complexity of the involved processes, which are difficult to be translated into deterministic models. The expert system SBC_MEDIU provides an integrated decision support tool for environmental management, solving problems such as water quality, air pollution and soil erosion analysis. The system can be used as an educational tool for the students that study the environmental protection domain.

At the level of local Environmental Protection Agencies, SBC-MEDIU can be integrated in a Local Monitoring Plan to provide useful information for decisional support and reliable answers to:

- standards/regulations demands concerning qualitative and quantitative aspects;
- trends of environmental quality modification due to various factors;
- impact of deterioration on ecosystems;
- and the efficiency of strategies and management action for pollution control.

The preliminary use of this system might also facilitate the environmental monitoring-network design, control strategy evaluation or control-technology evaluation.

As a future work we shall verify the performance of the knowledge bases on more scenarios, with interdependencies between air, water and soil pollution. Also, we will consider the increasing of the system complexity to become an alert system serving to signal when potential pollution is high for water, air or soil, requiring interaction between control agencies and emitters. The system will serve to locate areas of expected high concentration for correlation with health effects, or to identify environmental pollution issues.

Acknowledgement. The research reported in this paper was partially funded by the Romanian Education and Research Ministry and the National Council of Academic Research (UEFISCSU-CNCSIS) under the postdoctoral research programme CEEX19-1533/2006.

References

1. Bellamy P.H., Jones R.J.A., Identifying Changes in Soil Quality: Contamination and Organic Matter Decline, A. Ebel and T. Davitashvili (editors), *Air, Water and Soil Quality Modelling for Risk and Impact Assessment*, Springer (2007) 271-279.
2. Friedrich S., Gargano M., *Expert system design and development using VP-expert*, London, Wiley (1989).
3. Jetten, V., Govers, G., Hessel, R., Erosion models: quality of spatial predictions, *Hydrological Processes* 17 (2003) 887-900.
4. Gavrilescu E., Olteanu I., *Environmental Quality. Analysis methods (soil)*, in Romanian, Universitaria Publishing House, Craiova (2003).
5. Mureşan D. et al., *Irrigations, draining, and soil erosion control*, in Romanian, Didactical and Pedagogical Publishing House, Bucureşti (1992).
6. Nicolescu C., Gorghiu G., Dunea D., Buruleanu L., Moise V., Mapping Air Quality: An Assessment of the Pollutants Dispersion in Inhabited Areas to Predict and Manage Environmental Risks, WSEAS Transactions on Environment and Development, Volume 4 (2008) 1078-1088.
7. Oprea M., A case study of knowledge modelling in an air pollution control decision support system, *AiCommunications*, IOS Press, 18(4) (2005) 293-303.
8. Oprea M., Sànchez-Marrè M. (editors), *Proceedings of the International Workshop Binding Environmental Sciences and Artificial Intelligence*, 4th edition, Valencia, Spain (2004).
9. Oprea M., Dunea D., Modelling a Surface Water Pollution Analysis System with a Knowledge-based Approach, *Proceedings of EMCSR 2008*, Vienna, vol. 2 (2008) 585-590.
10. Pepper I.L., Gerba C.P., Brusseau M.L., (editors), Environmental and Pollution Sciences, 2nd edition, Academic Press, Elsevier (2006).
11. Pimentel, D., Harvey, C., Resosudarmo, P., Sinclair, K., Kurz, D., McNair, M., Crist, S., Shpritz, L., Fitton, L., Saffouri, R., Blair, R., Environmental and economic costs of soil erosion and conservation benefits, *Science* 267 (1995) 1117-1123.
12. Platt U., Air Pollution Monitoring Systems – Past - Present - Future, Y.J. Kim and U. Platt (editors), Advanced Environmental Monitoring, Springer (2008) 3-20.
13. Schnelle K.B., Brown C. A., Air Pollution Control Technology Handbook, CRC Press LLC (2002).
14. Vogmahadlek Ch., Satayopas B., Applicability of RAMS for a Simulation to Provide Inputs to an Air Quality Model: Modeling Evaluation and Sensitivity Test, WSEAS Transactions on Environment and Development, 8(3) (2007) 129-138.

Autonomous Inspection Of Complex Environments by Means of Semantic Techniques

M. Ziegenmeyer and K. Uhl and J.M. Zöllner and R. Dillmann

Abstract The autonomous inspection of complex environments is a challenging task. An autonomous inspection robot should actively examine entities of interest (EOIs), e.g. defects, and should perform additional inspection actions until the data analysis results reach an appropriate level of confidence. In this paper a semantic approach for inspection planning, plan execution, assessment of the data analysis results, decision making and replanning is proposed. The main idea is to incorporate human expert knowledge via a semantic inspection model. For the experimental evaluation of this approach the detection and classification of waste on irregular terrains with the hexapod walking machine LAURON is chosen. First preliminary simulation results are presented.

1 Introduction

The inspection of complex environments like sewers, pipelines, power transmission lines or dams is a challenging task for autonomous inspection robots.

Recently, there has been a lot of research in this area. The approaches can be roughly categorized into two categories. First, the hardware design and the control of the inspection robot itself are considered, e.g. Nassiraei et al. [5]. Second, appropriate sensor systems, their automatic placement and the corresponding data analysis components are examined, e.g. Duran et al. [2].

However, there exist only few integrated approaches aiming at fully autonomous inspection systems. In [1] the Onboard Autonomous Science Investigation System (OASIS) is described. OASIS is designed to operate onboard a planetary rover identifying and reacting to serendipitous science opportunities. It analyzes data the rover gathers during traverses, and then prioritizes the data for transmission back to earth

M. Ziegenmeyer, K. Uhl, J. M. Zöllner, R. Dillmann
FZI Forschungszentrum Informatik, Intelligent Systems and Production Engineering (ISPE),
Haid-und-Neu-Str. 10-14, D-76131 Karlsruhe, Germany, e-mail: {lastname}@fzi.de

based on criteria set by the science team. OASIS is also searching for specific targets it has been told to find. If one of these targets is found, it is identified as a new science opportunity and is sent to the planning and scheduling component. A continuous planning approach [3] is used to iteratively adjust the plan as new goals occur, while ensuring that resource and other operation constraints are met. The expert knowledge for identifying science opportunities is provided to the system by means of algorithms for feature extraction from images, analyzing the gathered data and prioritizing rocks.

Today, there exist only few semantic approaches regarding autonomous inspection missions. The authors of [6] present an approach for autonomous mission plan recovery for maintaining operability of unmanned underwater vehicles. The approach uses ontology reasoning in order to orient the planning algorithms adapting the mission plan of the vehicle. It can handle uncertainty and action scheduling in order to maximize mission efficiency and minimize mission failures due to external unexpected factors. In one of the simulation scenarios smart AUVs with fully autonomous inspection methods are briefly mentioned, otherwise nothing is stated on the on-line assessment of inspection data for mission planning and decision making.

Nevertheless, a semantic inspection approach offers several advantages. On the one hand, easy system extensibility and maintenance is achieved by the explicit separation of knowledge representation and execution control. On the other hand, the human comprehension of the system decisions is improved significantly. Moreover, the usability of the system is increased by allowing the user to communicate with the system on a semantic level.

In this paper we investigate a semantic approach for inspection planning, plan execution, assessment of the data analysis results, decision making and replanning. The main idea is to incorporate human expert knowledge via a semantic inspection model.

2 Semantic Inspection Approach

The proposed semantic inspection approach comprises a mission control architecture which is outlined in Sect. 2.1. At the core of this mission control architecture a knowledge base containing all knowledge relevant to the execution of inspection missions with autonomous service robots is located. The knowledge base is described in Sect. 2.2. The autonomous inspection process consisting of inspection planning, plan execution, assessment of the data analysis results, decision making and replanning is presented in Sect. 2.3.

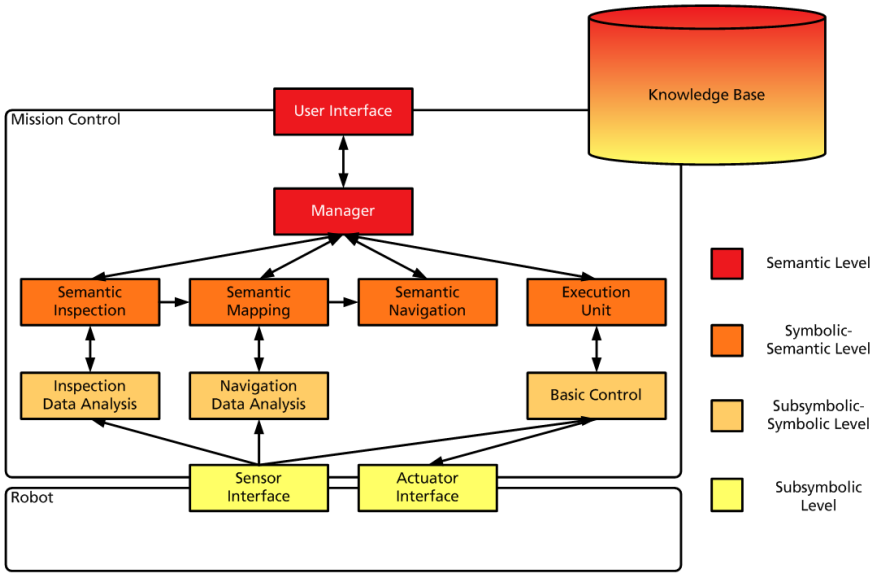


Fig. 1 The mission control architecture.

2.1 Architecture

The components of the mission control are arranged in an hierarchical architecture which consists of four distinguished levels. The four levels depend on the type of data that is processed and are depicted in Fig. 1. The mission control architecture has been implemented with *MCA2* [7] - a modular, network-transparent and real-time capable C++ framework for controlling robots. In the following, the individual components are described briefly.

Inspection Data Analysis The *Inspection Data Analysis* continuously reads data from the inspection sensors and searches for EOIs. If an EOI is detected, the corresponding region of the sensor data is segmented. For the segmented region features are computed which are used for classification. The *Inspection Data Analysis* is stateless and takes only the current measurement into account.

Navigation Data Analysis The *Navigation Data Analysis* continuously reads data from the navigation sensors. It locates and classifies regions in this sensor data and determines their parameters. Like the *Inspection Data Analysis* the *Navigation Data Analysis* is stateless and takes only the current measurement into account.

Semantic Inspection The *Semantic Inspection* receives abstract inspection goals from the *Manager*, computes appropriate plans to achieve those goals and passes them back to the *Manager*. Moreover, it performs a temporal fusion of the individual inspection data analysis results and assesses them. Depending on these results and

based on the semantic inspection model, it proposes to the *Manager* whether and how a found EOI should be examined further.

Semantic Mapping The *Semantic Mapping* temporally fuses the data from the *Navigation Data Analysis* and computes respectively updates the semantic region map of the environment. Moreover, the EOIs found by the *Semantic Inspection* are registered within the semantic region map.

Semantic Navigation The *Semantic Navigation* receives abstract locomotion goals from the *Manager*, computes plans by means of the semantic region map to achieve those goals, and passes them back to the *Manager*.

Manager The *Manager* is the highest level control and decision component. It decomposes the given mission goals into inspection and navigation subgoals and passes them to the *Semantic Inspection* and the *Semantic Navigation* for planning. It fuses the resulting subplans, passes them to the *Execution Unit* and coordinates and monitors their execution.

Execution Unit The *Execution Unit* receives plans from the *Manager*. It decomposes these plans into individual actions, passes them to the *Basic Control* and monitors their execution.

Basic Control The *Basic Control* receives a single symbolic action or a set of parallel actions from the *Execution Unit* at a time. These are passed as subsymbolic commands to the sensor and actor interfaces of the robot platform and their execution is monitored.

2.2 Knowledge Base

The knowledge base consists of several ontologies which model the concepts and contexts required for the semantic inspection control in a general form (terminological box, T-Box), and concrete instances of concepts and relations which represent the current state of the world (assertional box, A-Box). The T-Box of the knowledge base is organized in three abstraction layers: a basic ontology, a core ontology and a specific domain ontology (cf. Fig. 2). The ontologies are realized with *OWL-DL*[1]. As framework for managing the ontologies and for reasoning processes regarding the ontologies *KAON2*[2] is used, which supports the SHIQ(D) subset of OWL-DL. For this paper, the description of the knowledge base concentrates on the mission and inspection subontologies of the core ontology.

Mission Subontology The core concept of the mission subontology is the plan. The structure of plans is modeled after so-called *Flexible Programs* [4]. A plan is represented as a tree of nodes. There are three types of nodes: branching nodes, action

[1] OWL-DL: http://www.w3.org/TR/owl-features/

[2] KAON2: http://kaon2.semanticweb.org/

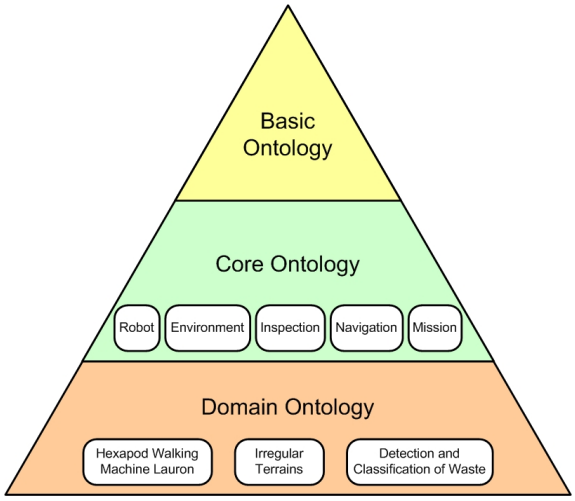


Fig. 2 The structure of the knowledge base. The basic ontology contains fundamental concepts like parameter, timestamp, condition, function, and data type. The core ontology includes robot, environment, inspection, navigation and mission subontologies modeling the central concepts and relations of the particular fields. The domain ontology contains application specific subontologies.

nodes and planning nodes. Each node contains a unique identifier Id, a precondition C_{pre}, a runtime condition C_{rt}, a postcondition C_{post}, a rating function R, and a success measure S. All inner nodes of a plan are branching nodes. They structure the plan into sequential and parallel parts. Therefore, they contain seats arranged in parallel groups. For each seat there can be several candidate nodes. The leaf nodes of a plan are either action or planning nodes. Action nodes contain elementary actions and planning nodes comprise subgoals.

Inspection Subontology The key concepts of the inspection subontology are the entity of interest (EOI) class and the inspection method.

An EOI class contains information about appropriate inspection methods for the detection and analysis of EOIs of a particular type. Moreover, it contains knowledge about characteristic features and potential locations as well as information about potential confusions with other EOI classes.

An inspection method consists of appropriate elementary actions for the detection and analysis of certain EOI classes. This comprises actions for acquiring sensor measurements, preprocessing sensor measurements, sensor data fusion, segmentation of potential EOIs, feature computation and classification. Moreover, each inspection method contains a reliability function and criteria for the assessment of the results.

2.3 Autonomous Inspection

An autonomous inspection robot should actively examine entities of interest (EOIs), e.g. defects. If the data analysis results are uncertain, additional inspection actions, e.g. activating a special sensor, approaching the EOI from a different perspective

or employing a different data analysis algorithm, should be taken to increase the confidence of the results. The selection of these actions should be driven by the assessment of the individual circumstances.

Therefore, the inspection of complex environments should occur in cycles of inspection planning, plan execution, assessment of the data analysis results, decision making and replanning.

Inspection Planning We decided to choose an hierarchical approach for inspection planning: Complex inspection goals are recursively decomposed by the *Semantic Inspection* into simpler subgoals until the subgoals can be solved with elementary actions. The knowledge necessary for decomposing goals into subgoals is stored within the knowledge base in form of the available flexible program nodes. For each goal to be achieved a corresponding root node for a flexible program is selected. Based on the current situation stored in the knowledge base the inspection planner then decomposes this root node into an executable flexible program.

The knowledge about EOI classes and inspection methods is used to compute plans which gradually increase the classification confidence of an EOI. The plans can also contain planning nodes with navigation goals, which are used to change the robot position or to reposition sensors. The navigation goals are passed to the *Semantic Navigation* which decomposes them into executable flexible subprograms. The semantic navigation approach will be described in a future paper.

Plan Execution During plan execution the *Execution Unit* processes the given flexible programs by a depth-first strategy. The processing state of each node can be virtual (not yet visited), instantiated (candidates chosen) and finished (fully processed). The selection of candidate nodes takes place by checking the preconditions and evaluating the rating functions of the respective candidates. Both the precondition checks and the evaluation of the rating functions are based on the current situation stored in the A-Box of the knowledge base. Action nodes trigger elementary actions which are executed by the *Basic Control* until the postcondition is reached or the runtime condition is no longer satisfied. Planning nodes initiate replanning processes for subgoals.

Assessment of the Data Analysis Results For the assessment of the inspection data analysis results by the *Semantic Inspection* an assignment between previously found EOIs and current EOIs has to be conducted. This is based on the world coordinates of the EOIs and the EOI hypotheses. For EOIs assigned to previously found EOIs a temporal fusion of the hypotheses has to be performed. This is achieved by means of Bayesian networks and incorporates the reliabilities of the used inspection methods as well as other factors, e.g. the sensor resolution.

Decision Making In case of uncertainty regarding the data analysis results of an EOI a decision has to be made by the *Semantic Inspection* whether and how to proceed with the inspection of the EOI. Here we use a probabilistic approach in form of Bayesian decision networks. The available decision options correspond to goals stored in the knowledge base. Moreover, the different goals are prioritized according to the current inspection goals and criteria stored in the knowledge base.

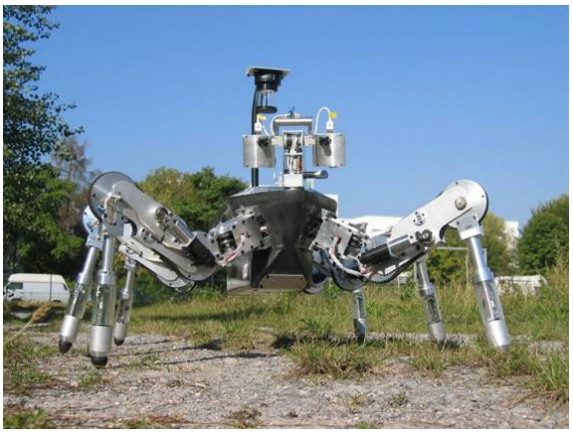


Fig. 3 The six-legged walking machine LAURON IVc is equipped with appropriate sensors for localization, navigation and perception of its environment, e.g. a stereo camera system and a 3D time-of-flight camera on a pan-tilt unit. Moreover, an extensive behavior repertoire for locomotion and navigation exists.

Replanning The new inspection goals from the decision making step are integrated into the overall plan by the *Semantic Inspection* according to their priorities and resource constraints by reinvoking the inspection planning process.

3 Preliminary Results

To be able to conduct experiments and evaluate the proposed semantic inspection approach an appropriate robotic platform and inspection scenario has to be chosen. For this purpose the hexapod walking machine LAURON IVc (cf. Fig. 3) is used. As inspection scenario the detection and classification of different kinds of waste on irregular terrains like river and channel banks, seashores, countryside areas such as dunes or forests, or areas along the highways, is chosen. The vision is to equip the front legs of the next LAURON generation with simple waste-grippers, extend the working area of the legs by an additional degree of freedom and place a garbage container on the back of the machine.

While the full realization of the mission control system for the proposed inspection scenario is still work in progress, a simulation environment has been established for early testing. The simulation environment contains a model of LAURON and is based upon the existing behavior repertoire for locomotion and navigation. It enables testing of fully implemented components together with component stubs. The component stubs are realized as question/answer methods for simulating the desired functionality, which can be used in an interactive as well as an automated way.

Several systematic experiments were conducted to validate the different components. First, the planning process and the suitability of the expert knowledge defined in the knowledge base were verified. Therefore, different inspection goals were passed to the inspection planner for decomposition. Second, the execution of the generated flexible programs in case of errors (e.g. malfunction of a sensor) was analyzed. Third, the assessment of the data analysis results and the decision mak-

ing process were validated by simulating different inspection situations and data analysis results.

The results of these first functional tests were promising and showed the principal feasibility of the proposed semantic inspection approach. Nevertheless, more simulation and especially real-world experiments have to be done to detect potential improvements of the proposed approach.

4 Conclusion and Future Work

In this paper a semantic approach for inspection planning, plan execution, assessment of the data analysis results, decision making and replanning was presented. For the experimental evaluation of the proposed approach the detection and classification of waste on irregular terrains with the hexapod walking machine LAURON was chosen. First preliminary simulation results were presented.

Future work will focus on the further realization of the mission control system for the proposed inspection scenario to allow for real field tests. Moreover, an appropriate user interface for semantic interaction with the inspection control system will be developed. Finally, learning capabilities for self optimizing the resource usage, the data analysis process, the planning process and the decision making process will be investigated.

References

1. R. Castano, T. Estlin, D. Gaines, C. Chouinard, B. Bomstein, R.C. Anderson, M. Burl, D. Thompson, A. Castano, and M. Judd. Onboard autonomous rover science. In *Aerospace Conference, 2007 IEEE*, pages 1–13, 2007.
2. O. Duran, K. Althoefer, and L. D. Seneviratne. Automated pipe defect detection and categorization using camera/laser-based profiler and artificial neural network. In *Automation Science and Engineering, IEEE Transactions on*, volume 4, pages 118–126, January 2007.
3. T. Estlin, D. Gaines, C. Chouinard, R. Castano, B. Bornstein, M. Judd, I. Nesnas, and R. Anderson. Increased mars rover autonomy using AI planning, scheduling and execution. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 4911–4918, 2007.
4. S. Knoop, S. R. Schmidt-Rohr, and R. Dillmann. A flexible task knowledge representation for service robots. In *The 9th International Conference on Intelligent Autonomous Systems (IAS-9)*, 2006.
5. A. A. F. Nassiraei, Y. Kawamura, A. Ahrary, Y. Mikuriya, and K. Ishii. Concept and design of a fully autonomous sewer pipe inspection mobile robot "KANTARO". In *Robotics and Automation, 2007 IEEE International Conference on*, pages 136–143, Roma, April 2007.
6. Pedro Patron, Emilio Miguelanez, Yvan R. Petillot, and David M. Lane. Fault tolerant adaptive mission planning with semantic knowledge representation for autonomous underwater vehicles. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 2593–2598, 2008.
7. K. Uhl and M. Ziegenmeyer. MCA2 – an extensible modular framework for robot control applications. In *The 10th International Conference on Climbing and Walking Robots (CLAWAR 2007)*, 2007.

Use of AI Techniques for Residential Fire Detection in Wireless Sensor Networks

Majid Bahrepour, Nirvana Meratnia, Paul J. M. Havinga

(m.bahrepour, n.meratnia, p.j.m.havinga)@utwente.nl

Pervasive Systems Research Group, Twente University, the Netherlands

Abstract: Early residential fire detection is important for prompt extinguishing and reducing damages and life losses. To detect fire, one or a combination of sensors and a detection algorithm are needed. The sensors might be part of a wireless sensor network (WSN) or work independently. The previous research in the area of fire detection using WSN has paid little or no attention to investigate the optimal set of sensors as well as use of learning mechanisms and Artificial Intelligence (AI) techniques. They have only made some assumptions on what might be considered as appropriate sensor or an arbitrary AI technique has been used. By closing the gap between traditional fire detection techniques and modern wireless sensor network capabilities, in this paper we present a guideline on choosing the most optimal sensor combinations for accurate residential fire detection. Additionally, applicability of a feed forward neural network (FFNN) and Naïve Bayes Classifier is investigated and results in terms of detection rate and computational complexity are analyzed.

1 Introduction

Fires may take place in various environments, such as residential places, forests or open spaces. The easiest way to detect a fire at residential places is using the smoke detectors or any other similar sensors, which are usually sensitive to ionization or obscuration [1]. The problem with such detectors is that they are prone to false alarms. This means that in noisy conditions, such as smoking a cigarette or toasting a bread, a fire alarm may be generated wrongly [2, 3].

Generally, to reduce false alarms and perform fire detection accurately, two approaches are used [4]. The first approach uses one type of sensor and conducts the fire detection by a complex algorithm. An example of this approach is the work presented in [5], which uses a flame detection sensor and a fuzzy-wavelet classifier. In contrast, the second approach uses multiple sensors and performs the detec-

tion by a simple mathematical operation. The work presented in [2] is an example of the second approach, which uses CO and ionization (ION) sensors and a simple mathematic operation. Some researchers also tried to combine both approaches by using multiple sensors and an appropriate algorithm. The work presented in [6], which uses a feed forward neural network (FFNN) and four sensors, i.e., temperature, ION, CO and photoelectric, and their rising rates to discriminate fires from nuisance sources, is an example of the combined approach.

In recent studies, Wireless Sensor Networks (WSN) has also been proposed for fire detection [7-14]. In this type of research, fire detection in residential areas as well as forests and mines are considered as applications for WSN.

Although there are many achievements in the area of fire detection (in terms of selecting optimal sensors and algorithms) using individual sensors in general, these achievements often have not made their way into the WSN field. In this paper, we aim at bringing knowledge of already established fields of AI (because of their learning process, reasonable accuracy and computational cost) and fire detection into the WSN field.

The rest of this paper is structured as follows. Section 2 briefly reviews previous contributions for fire detection using WSN. In Section 3, our proposed fire detection technique is introduced. Section 4 reports the experimental results. Finally, some conclusions and future plans are given in Section 5.

2. Literature Review

In this section, contributions of WSN for fire detection are briefly surveyed. A more complete literature review on this matter can be found in our technical report [4].

Yu et al. [13] used the National Fire Danger Rating System (NFDRS) for forest fire detection. NFDRS inputs four sensory information (humidity, temperature, smoke and windy speed) and generates a fire-likelihood index. The contribution of this study is the function of a feed-forward neural network for data aggregation and reducing communication overhead.

Lu Zhiping et al. [14] proposed a forest fire detection approach using WSN. Their system is composed of some sensor nodes, gateway(s) and task manager(s). Each sensor node is equipped with temperature and humidity sensors. After obtaining sensory information at sensor nodes, the data is fused at the gateways and data analysis and decision making tasks are conducted by the task manager nodes.

In [7], the author incorporated Fire Weather Index (FWI) and a novel k-coverage algorithm to detect forest fires. K-coverage algorithm monitors each point by using k or more sensor nodes to improve fault tolerance. Therefore, some sensors can be put in standby mode to extend network lifetime. Although there are many algorithms to find the minimum number of sensors to be used, they are

usually NP complete problems [12]. The proposed k-coverage solution proved that it can prolong the network life time.

Zervas et al. proposed a sensor network approach for early fire detection of open spaces such as jungles and urban areas [15]. They incorporated a temperature sensor and maximum likelihood algorithm to fuse sensory information. Their proposed system architecture is composed of (1) sensing subsystem, (2) computing subsystem, and (3) localized alerting subsystem. The author concluded the applicability of their approach for early fire detection.

A skyline approach for early forest fire detection is proposed in [10]. Skyline is built using greater values, i.e., those sensor readings with large temperature and high wind speed. Only data on skyline are sent to a sink to be used for fire detection. Sink processes the data according to the suggested algorithm and results in a fast and energy efficient forest fire detection.

Marin-Perianu et al. proposed a distributed fuzzy inference engine, called D-FLER, for event detection using WSN [9]. They considered fire as an event utilizing smoke and temperature sensors. D-FLER combines individual sensor inputs with neighborhood observation using a distributed fuzzy logic engine. The prototype of their work was implemented in practice using Ambient μNode 2.0 platform [16].

3. Proposed Fire Detection Approach

By looking at the previous work on fire detection using WSN, we can conclude that, use of WSN for fire detection can be improved in two directions. The first direction is to use more sensors in combination and conduct sensor fusion. This can lead to more accurate fire detection by incorporating more than one sensor [6]. The second direction is to use more intelligent detection algorithms such as AI approaches, as fires and nuisances have a distinct pattern.

In WSN research community, selection of sensors was often carried out randomly or assumption-basely. Although temperature sensors are probably the simplest and the most obvious sensors for fire detection, studying various sources in this field reveals that all researchers agree on the fact that it alone is not a suitable indicator for fires and gas concentration sensors result in a better fire detection and discriminating fire and noise sources [3,6]

In our approach, we adapt the optimal sensor set from [6] and use temperature, ionization, photoelectric and CO sensors. We assume that every node in the WSN contains all the required sensors. In this case, communication overhead between neighboring nodes is avoided and each sensor node can detect fire locally by itself.

To achieve this goal, sensor nodes need a computationally cheap, yet, efficient algorithm to conduct fire detection in a (near) real-time manner. For this reason, we propose to use FFNN and Naïve Bayes classifier. Subsections 3.1-3.3 provide information about these classifiers and the reasons why they are helpful for WSN.

3.1 Feed Forward Neural Network (FFNN)

The artificial neural network (ANN) is a mathematical model or computational model based upon biological neural networks. It is composed of an interconnected group of artificial neurons and processes information using a connectionist approach for computation [17]. Feed forward neural network (FFNN) is a sort of the neural networks, in which each layer is fed by its back layer [18]. FFNN consists of one input layer, one or more hidden layers and one output layer. Fig 1 shows the FFNN's architecture.

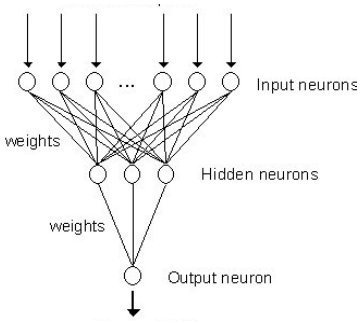


Fig. 1: Architecture of a Neural Network

The challenge of such networks is finding the weights. The process of finding the appropriate weights, which is called 'learning', can be carried out by some algorithms such as gradient descent (GD) approach.

3.2 Naïve Bayes Classifiers

A Naïve Bayes classifier uses Bayesian statistics and Bayes' theorem to find the probability of each instance belonging to a specific class. It is called Naïve because of emphasizing on independency of the assumptions. To find the probability of belongingness of each instant to a specific class, Eq. 2 can be used. Eq. 2, expresses the probability of an example $E = (x_1, x_2, ..., x_n)$ belonging to class c [19].

$$p(c \mid E) = \frac{p(E \mid c)p(c)}{p(E)} \tag{1}$$

3.3 Advantages of the FFNN and Naïve Bayes Classifier for WSN

The main advantage of the FFNN for WSN is its ease to be programmed into a sensor node. Let us assume to have an FFNN with three neurons in input layer, two neurons in hidden layer and a neuron in output layer. The weights can be

found by the GD learning algorithm. Then, we might have a network similar to Fig 2. Another advantage of the FFNN is its parallel capability, which means parameters used in Eq. 2 can be calculated independently and in parallel.

This network can be easily programmed into sensor nodes using Eq. 1. Evaluating this mathematical formula in form of a business rule is computationally very cheap and appropriate for resource constraint sensor nodes. This equation can be extended to more neurons and layers but the idea is the same. Eq. 2 formulates the network in a form of mathematical model. One should note that each neuron passes the sum of product (SOP) of the previous layer. In some networks SOP is given to a non-linear function such as tangent and transformation is a nonlinear one that makes Eq. 2 slightly different.

$$Output = \left\lfloor W_{3,1} \times \sum_{j=1}^{3} (W_{1,j} \times I_j) \right\rfloor + \left\lfloor W_{3,2} \times \sum_{j=1}^{3} (W_{2,j} \times I_j) \right\rfloor \quad (2)$$

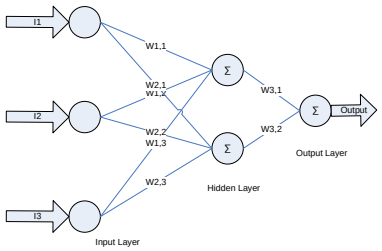


Fig. 2: An FFNN with three neurons in Input, two neurons in hidden layer, and one neuron in output layer along with their corresponding weights.

Naïve Bayes classifier is also easy to implement. The most time-consuming part is how to compute $p(E \mid c)$ in Eq. 1. This probability calculation is important to make the classifier more accurate. In basic literatures of pattern recognition or machine learning, it is proposed that this probability can be estimated by some standard data distribution such as Gaussian or Poisson [20].

To do a more accurate probability calculation, we can divide data into some intervals and count the data frequency within that interval. The new instances are also partitioned to the same intervals for finding the probability of each feature to be in that class.

To clarify the method, suppose we have the following data for ten samples in two classes A, B:

$$A = [8,7,2,4,6,9,8,9,1,3]$$

$$B = [1,1,1,3,3,5,8,4,2,2]$$

Then we divide these data into two intervals. Two intervals were chosen to simplify the example however the number of intervals are arbitrary. Therefore,

those numbers less than five are allocated in the first interval, i.e., interval i_1, and the rest in the second interval, interval i_2.

Table 1. Classes and their Probability

	i_1 $(x<5)$	i_2 $(x\geq 5)$
P_A	0.4	0.6
P_B	0.8	0.2

Now, let us assume to have an instance $x_1 = 3$ that should be classified into either Class *A* or Class *B*. It can easily be discovered that 3 belongs to the first interval, i_1, as it is less than 5. Then by looking at the probability table, Table 1, this can be seen that the probability of belongingness to class *B*, is higher ($P_B = 0.8 > P_A = 0.4$). Therefore we classify x_1 to class *B*.

This method of classification is also considerable for WSN because this is based upon a table which can be computed offline. Thereafter, this table is programmed into a sensor node and a simple algorithm inside the sensor nodes searches the table for the higher probable class.

In the next section the empirical results for both approaches is presented and also compared with a recent study.

4. Empirical Results

To evaluate the proposed approach, a set of data were obtained and a number of experiments were conducted. Subsection 4.1 describes the dataset, while Subsection 4.2 reports and compares the final results.

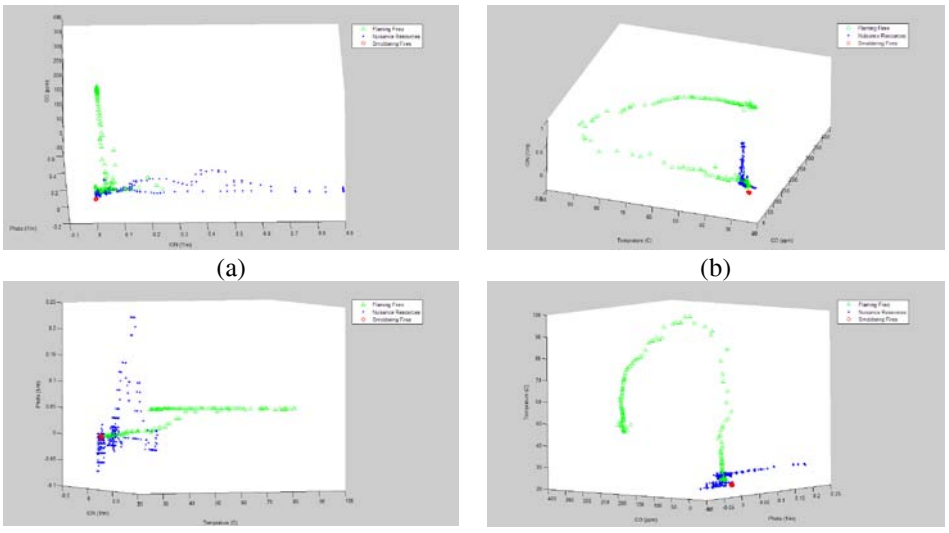
4.1 Dataset

A set of data were obtained from NIST website (http://smokealarm.nist.gov/). To identify smoldering fire data, flaming fire data is combined with noise. Therefore, two smoldering fire dataset (SDC31, SDC40), two flaming fire dataset (SDC10, SDC14) and two nuisance resource dataset (MHN06, MHN16) were merged together. Totally 1400 data records were prepared, all having same units. Fig. 3 displays the data in 3D space. The goal is to make a classifier that can separate these data and classify them into their respective class, i.e., fire and noise.

4.2 Experimental results

The data were given to both classifiers and the results were obtained. To perform a cross validation, 1400 data records were divided to a 1000 training data

and a 400 test data. All data were randomly mixed and given to the classifiers. Each test repeated 10 times and the average accuracy rate by changing the classifiers' parameters is reported in Tables 2-3. Table 4 provides a general comparison of our approach with a recent study, in which a distributed fuzzy system was proposed for residential fire detection using WSN [9]



(a) (b)

(c) (d)

Fig. 3: Fire and noise data. (a) Ion, Photo and CO (b) Ion, temperature and CO (c) Temperature, Ion and Photo (d) Photo, CO and Temperature

For simulation of the proposed approaches Matlab® 7.1 was used. A two pass smoothing filter for a preprocessor was also applied that was adapted from [6].

Table 2. Empirical Results for Naïve Bayes Classifier

Number of Intervals	10	100	300	600	1000
Accuracy	32.15%	63.15%	96.425%	98.675%	100%

Table 3. Empirical Results for FFNN

Number of Neurons in the Hidden Layer	5	10	20	50
Accuracy Rate	97.495%	98.45%	93%	90.1%

Table 4. Comparing the Empirical Results with D-FLER [9]

Best Result	Naïve Bayes	Neural Network	D-FLER[9]
Accuracy Rate	100%	98.45%	98.67%

4.2 Computation Complexity Consideration

To compare these three approaches, not only the accuracy but also computation complexity is of significant importance, as they need to be implemented on tiny resource constraint sensor nodes.

4.2.1 FFNN's Computation Complexity

The most expensive part of the FFNN computationally is the training phase. Since we consider that FFNN is trained once and then is programmed into the sensor nodes, its computation complexity is negligible.

The computation complexity of a FFNN with m neurons in its input layer (number of features), n neurons in hidden layer, and p neurons in output layer is shown in Eq. (3).

$$O_{FFNN} = O(m \times n \times p) \tag{3}$$

In this calculation the multiplication operator is considered as the key for computation complexity calculation.

4.2.2 Naïve Bayes's Computation Complexity

The most expensive part of the Naïve Bayes classifier computationally is making the probability table. We assume that this probability table is made once and then is programmed into the sensor nodes. In this case computation complexity is calculated for search process only, which is more expensive. Computation complexity for the Naïve Bayes is calculated based on the Eq. (4), where m is number of features, i is number of classes, and j is number of intervals.

$$O_{NaiveBayes} = O(m \times i \times j) \tag{4}$$

4.2.3 D-FLER's Computation Complexity

Defining the fuzzy rules and membership functions represent the most complicated part of the fuzzy inference engine design. Assuming that these are programmed into the sensor nodes, the time complexity of the fuzzy inference engine

is calculated based on the Eq. (5), where m is the number of membership functions per input, i is the number of inputs, r is the number of rules, o is the number of outputs (in the particular case of fire detection, o =1).

$$O_{D-FLER} = O(m \times i \times r \times o) \quad (5)$$

As shown in [9], the actual execution time can be greatly influenced by the specific defuzzification method chosen, to the extent that the number of outputs o can become the determinant factor.

4.2.4 Computation Complexity Comparison

Comparing computation complexity of the FFNN, the Naïve Bayes classifier, and fuzzy logic approaches shows that they are all product of three terms and if all variables have the same values it is a non-linear equation of power 3.

Table 5. Computation Complexity Comparison

	Naïve Bayes	Neural Network	D-FLER[9]
Computation Complexity	$O(i \times j \times m)$	$O(m \times n \times p)$	$O(m \times i \times r \times o)$

5. Conclusion

Wireless Sensor Networks may be deployed in many places thus they have different requirements. According to their scenarios each sensor node is either equipped with all the appropriate sensors or just a sub set of them. Fire in WSN is considered as an event; therefore event detection techniques are used for its detection. In this study, the optimal set of four sensors, i.e., temperature, ionization, photoelectric and CO, were adapted from [6] and two fire detection techniques based on the FFNN and the Naïve Bayes classifier were proposed to detect fire on each node locally. To carry out the detection task the sensory information is given to a classifier. The computation complexity and accuracy rate of each of these techniques and a comparison between them and a recent study, called D-FLER [9] based on fuzzy logic were presented. Results show that while all the three have similar computation complexity, the Naïve Bayes classifier can achieve a better accuracy and has a lower communication overhead (since it is centralized assuming all the sensors are present at the sensor node). However, in case just a sub set of sensors is present at each sensor node, D-FLER has the advantage. This is a good guideline to choose a proper technique for a particular scenario (centralized versus distributed) in mind.

Acknowledgment

Authors would like to thank Mihai Marin-Perianu for his contribution on complexity calculation of the D-FLER approach, and Thomas Cleary for his valuable input regarding unit conversions in preparing the data set. This work is supported by the EU's Seventh Framework Programme, Aware project and the SENSEI project.

Reference:

1. Brain, M. *How Smoke Detectors Work* 2000 [cited; Available from: http://home.howstuffworks.com/smoke1.htm.
2. Gottuk, D.T., et al., *Advanced fire detection using multi-signature alarm algorithms.* Fire Safety Journal, 2002. **37**(4): p. 381-394
3. Milke, J.A. *Using Multiple Sensors for Discriminating Fire Detection*. in *Fire Suppression and Detection Research Application Symposium*. 1999: National Fire Protection Research Foundation
4. Bahrepour, M., N. Meratnia, and P.J.M. Havinga, *Automatic Fire Detection: A Survey from Wireless Sensor Network Perspective.* . 2008, Centre for Telematics and Information Technology, University of Twente: Enschede.
5. Thuillard, M. *Application of Fuzzy Wavelets and Wavelets in Soft Computing Illustrated with the Example of Fire Detectors*. in *Wavelet Applications VII*. 2000.
6. Cestari, L.A., C. Worrell, and J.A. Milke, *Advanced Fire Detection Algorithms Using Data from the Home Smoke Detector Project.* Fire Safety Journal, 2005. **40**: p. 1-28.
7. Bagheri, M., *Efficient K-Coverage Algorithms for Wireless Sensor Networks and Their Applications to Early Detection of Forest Fires*, in *Computing Science*. 2007, SIMON FRASER UNIVERSITY. p. 75.
8. Bernardo, L., et al. *A Fire Monitoring Application for Scattered Wireless Sensor Networks: A peer-to-peer cross-layering approach* in *International Conference on Wireless Information Networks and Systems (WINSYS'07)*. 2007. Barcelona, Spain.
9. Marin-Perianu, M. and P. Havinga, *D-FLER – A Distributed Fuzzy Logic Engine for Rule-Based Wireless Sensor Networks* Lecture Notes in Computer Science. Vol. 4836. 2008, Heidelberg: Springer Berlin.
10. Pripužic, K., H. Belani, and M. Vukovic, *Early Forest Fire Detection with Sensor Networks: Sliding Window Skylines Approach*. 2008, White Paper: University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Telecommunications.

11. Tan, W., et al. *Mine Fire Detection System Based on Wireless Sensor Network*. in *Information Acquisition, 2007. ICIA '07. International Conference on*. 2007.
12. Yang, S., et al., *On Connected Multiple Point Coverage in Wireless Sensor Networks* International Journal of Wireless Information Networks, 2006. **13**(4): p. 289-301.
13. Yu, L., N. Wang, and X. Meng. *Real-time forest fire detection with wireless sensor networks*. in *Wireless Communications, Networking and Mobile Computing*. 2005.
14. Zhiping, L., et al., *The Design of Wireless Sensor Networks for Forest Fire Monitoring System*. 2006, White Paper: School of Electronics and Information, Hangzhou Dianzi University.
15. Zervas, E., et al. *Fire Detection in the Urban Rural Interface through Fusion Techniques*. in *Mobile Adhoc and Sensor Systems (MASS 2007)*. 2007.
16. Hofmeijer, T., et al., *AmbientRT - Real Time System Software Support for Data Centric Sensor Networks.* Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2004: p. 61-66.
17. Wikipedia, *Neural network*, Wikipedia.
18. Mehrotra, K., C.K. Mohan, and S. Ranka, *Elements of Artificial Neural Networks*. 1996, MIT Press.
19. Zhang, H. *The Optimality of Naive Bayes*. in *Seventeenth Florida Artificial Intelligence Research Society Conference*. 2004: AAAI Press.
20. Alpaydin, E., *Introduction to Machine Learning* 2004: MIT Press.

Validation of a knowledge-based risk model for biological foaming in anaerobic digestion simulation

Dalmau[a] J., Comas[a] J., Rodríguez-Roda[a] I., Latrille[b] E. and Steyer[b] J.P.

[a]Laboratory of Chemical and Environmental Engineering (LEQUIA), University of Girona, Faculty of Sciences, Campus Montilivi s/n, 17071 Girona, Catalonia, Spain. (E-mail: jordi@lequia.udg.cat; quim@lequia.udg.cat; ignasi@lequia.udg.cat)
[b]INRA, UR050, Laboratoire de Biotechnologie de l'Environnement, Avenue des Étangs, 11100 Narbonne, France. (E-mail: latrille@supagro.inra.fr; steyer@supagro.inra.fr)

Abstract Anaerobic digestion (AD) is a complex biological system which can be affected by several operational problems. Among them, biological foaming is one of the most difficult to deal with. It has many effects, such as causing gas pipe clogging and probe failures, and it can affect mixing devices, etc. Since the mechanisms involved in biological foaming development are not fully understood, it is not included in standard anaerobic digestion models. For this reason, a knowledge-based risk model to determine the suitable conditions for the development of biological foaming during AD simulation was developed. The resulting knowledge-based system, based on organic loading rate and its daily variation, was experimentally validated using real data from a fully instrumented pilot plant (1 m^3 upflow fixed bed digester). Results show a good correlation between the knowledge-based risk model and the estimated biological foaming risk from real data.

Keywords: Anaerobic digestion, foaming, fuzzy logic, knowledge-based systems, validation**.**

1. INTRODUCTION

Activated sludge processes are complex biological systems in which organic matter and nutrients (nitrogen and phosphorous) are removed from wastewater. The system consists of an aeration tank where oxygen is selectively supplied and it is used by the microbial consortia (i.e. biomass and/or sludge) to grow and reproduce by consuming the substrate (i.e. pollutants) present in the wastewater.

The system also includes a secondary settler in which the treated water is separated from the biomass. From the bottom of the clarifier a fraction of the activated sludge is returned to the reactor in order to maintain the biomass constant in the reactor. To prevent overgrowth of the biomass in the system, a small fraction of the sludge is wasted from the system. This fraction represents a significant cost for the activated sludge process, since further treatment is required.

The most common alternative for sludge treatment is AD, as well as for wastewater with high contents of organic matter. In this process (Figure 2), the organic matter

(e.g. sludge coming from activated sludge treatment) is biologically degraded in a digester in absence of oxygen. AD advantages are numerous since they provide a treatment for highly loaded wastewater, low sludge production and production of energy in form of methane. AD processes are, like in the activated sludge, very complex biological systems since a huge amount of microbial species are involved in the process.

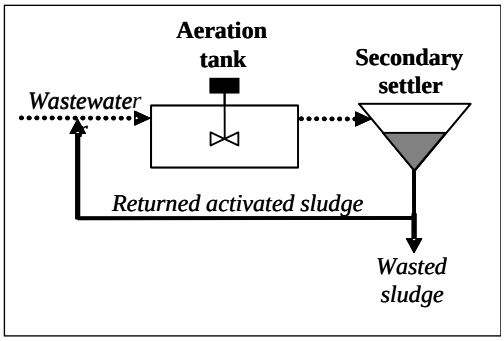


Figure 1. Activated Sludge System.

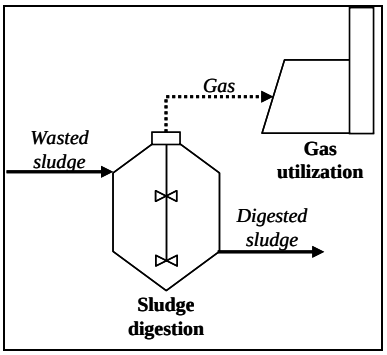


Figure 2. Anaerobic digestion System

Within this complexity some bacteria can have its own growth promoted by certain conditions which can cause imbalances in the digester in the form of a thick foam blanket. According to Pagilla *et al.* (1997), consequences of biological foaming are numerous:

- Blockage of gas mixing devices.
- Inversion of digester solids profiles.
- Foam binding of recirculation pumps.
- Fouling of gas collection pipes (due to entrapped foam solids).
- Foam penetration between floating covers and digester walls.
- Decrease of the digestion efficiency.

There is not yet a complete agreement on the parameters that favours conditions for foaming forming bacteria. Some authors state that a proper control of the feeding

will prevent excessive foaming to appear (Massart *et al.*, 2006; Schaffer *et al.*, 2006). Others state that pre-treatment of the feeding is necessary to avoid foaming appearing (Barjenbruch and Kopplow 2003; Elliott and Mahmoot 2007). Besides, some claim that the presences of some filamentous bacteria (e.g. *M. Parvicella*, *Nocardia amarae* …) in the activated sludge system are the cause for foaming problems in the anaerobic digester (Pagilla *et al.*, 1997; Westlund *et al.*, 1998). Precisely all these uncertainty about the causes of biological foaming hinders the development of a mechanistic model to assess the biological foaming appearance.

Knowledge-based systems have proven to be appropriate tools to deal with complex processes like those involving microbiology-related problems in activated sludge systems (Comas *et al.*, 2003; Poch *et al.*, 2004). Specifically fuzzy logic has been successfully applied to a variety of systems. For instance, in Lardon *et al.* (2005) is applied to several AD operational imbalances and, in Carrasco *et al.* (2004) it is shown how a fuzzy system is able to control and diagnose acidification states in an anaerobic digester.

When building knowledge-based systems, the selection of input variables and the study of the data related to the problem under study is important in order to get a reliable system. For this reason, a previous variable selection was performed to a set of data from a pilot plant in order to find the most relevant input variables for the knowledge-based system developed. The knowledge gained with the variable selection together with the heuristic knowledge present in the literature led to the development of a knowledge-based AD risk model implemented in fuzzy logic to assess favorable conditions for biological foaming in simulation. The rationale behind this risk model was that the deterministic modelling of some WWTP simulation scenarios, although performing better regarding economic and environmental issues, can induce a higher risk of biological foaming.

The aim of this paper is to test the performance of the developed AD risk model with real data from a pilot plant. The paper is structured as follows; first the variable selection method is explained together with a brief summary of the AD risk model. Then the validation section illustrates and discusses the performance of the AD risk model validation with real data and, finally, some conclusions are drawn.

2. DEVELOPMENT OF THE AD RISK MODEL

To select the most relevant variables a wrapper approach with a hill-climbing elimination strategy (Kohavi and John, 1997) was used. The same methodology was used in Dalmau *et al.* (2007) in order to find the most relevant variables for acidogenic states in anaerobic digestion. Afterwards in Dalmau *et al.* (2008), the same approach was applied to biological foaming in AD which is, as commented above, a more challenging issue.

A home-made neural network toolbox for static models for use in MATLAB 5.3 or higher was used. Two layers were chosen in all the ANN architectures: a hidden layer of neurons with sigmoid transfer functions and an output layer with linear transfer functions for outputs. The initialization method was performed using the Nguyen-Widrow algorithm option, which initializes the weights with random values, later selecting their probability distributions to make all neurons active for the expected data

ranges (Nguyen and Widrow, 1990). It also provides automatic data scaling and weights conversion. Bayesian regularisation is used to prevent over-fitting.

Figure 3 depicts the methodology used that starts with the ten times training of the reference ANN with all the variables. Its average Root Square Mean Error (RSME) is calculated and stored as the reference error. Next, one input variable is removed and a new ANN (ANN1 in figure 3) is trained ten times without it. This last step is repeated for each input variable ending up with *n* ANNs 1, one for each removed input variable with their related average RSME 1. Whenever a relevant variable is removed, the average RSME 1 of the related ANN 1 will increase with respect to the average reference error. On the other hand, whenever a non-relevant variable is removed the RSME 1 of the related ANN1 will decrease. Therefore, the variables which RSME 1 is higher than the reference error are selected as relevant variables.

Among relevant variables the one with the higher RSME 1 is selected first and a new ANN (ANN 2 this time) is trained ten times again using it as the only input. If the related average RSME (RSME 2) is higher than the average reference error no improvement is found, so the variable with the second higher average RSME 1 is selected and a new ANN 2 is trained (ten times as well) with both variables, and again, its average RSME 2 is compared with the reference. This iterative process is repeated until an average RSME 2 lower than the average reference RSME is obtained.

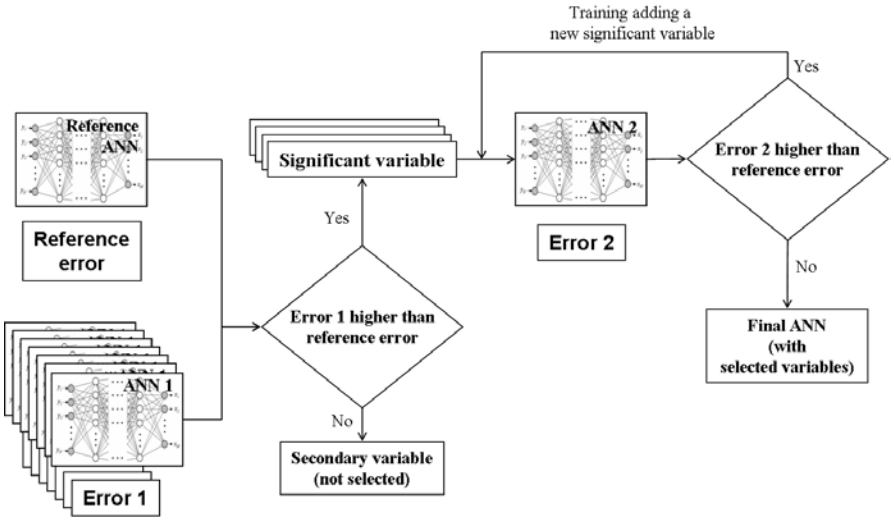


Figure 3. Methodology schema, based on Dalmau *et al.* (2007).

Experimental data used were obtained from a pilot plant from LBE of the INRA, France. Overall, a set of 8133 data was used for the variable selection. Among all variables a first selection was done based on the common variables which are available in real plants. Some others were not selected for instance, temperatures since it is usually constant so it will be difficult to extract information from its profile. Input variables involved in this study were: inflow rate and pH in the influent flow rate; vola-

tile fatty acids concentration, total organic carbon and pH in the digester and, carbon dioxide and methane percentage in the gas phase. As output, biological foaming appearance (foaming index) in the digester was used, based on the heuristic knowledge provided by the experts. It was noticed that when foaming appeared in the digester high variations of the gas flow rate and pressure coincided due to the slug release of gas bubbles trapped inside the foam. It is important to point that even though foaming can be estimated this way; this is an approach to study variables influence or relation. This approach it cannot be used in simulation because biological foaming is not currently modelled so the results of the simulation cannot reflect its effects on the gas flowrate and pressure variations.

Eventually, as shown in figure 4, the variables with RSME higher than the reference error (i.e. relevant variables) were: total organic carbon in the digester, the carbon dioxide and methane percentage in the gas phase and, the inflow rate and the pH in the inflow rate. The relevance of gas-related variables (i.e. carbon dioxide and methane percentage) can be due to the approach taken to determine foaming. According to Zhao and Viraraghavan (2004) high carbon dioxide production is representative of poor digestion that may lead to foaming, but in general is representative of general process imbalance but no related to a specific cause. So, taking a look to the other variables, precisely total organic carbon in the digester and inflow rate, the results can be related to some statements present in the literature. In Massart *et al.* (2006) it is stated that inconsistent feeding in the digester is one of the causes for foaming. Feeding is related to Organic Loading Rate (OLR), which is related to the inflow rate and the amount of sludge feed to the digester (Metcalf and Eddy, 2003) related at the same time to the organic matter present in the digester (total organic carbon in the digester).

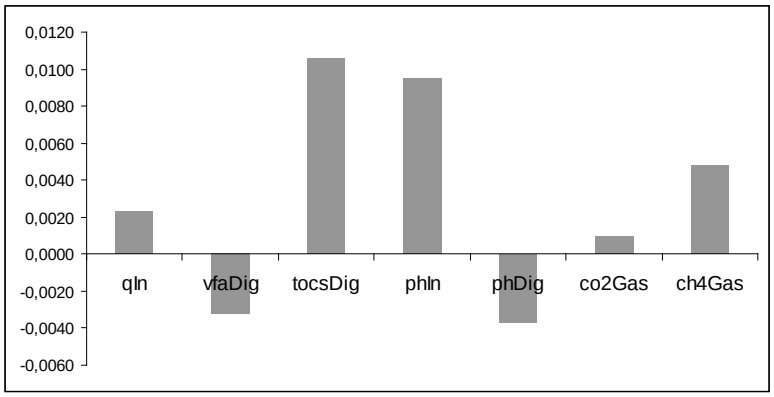


Figure 4. Difference between the RMSE and Reference error for each variable. From Dalmau *et al.* (2008).

3. AD RISK MODEL

To develop the risk model the relevant variables selected previously from real data were compared with the knowledge present in the experiences from the bibliography. As seen in the previous section, some coincidences were found. Finally, the combina-

tion of OLR and its variation were selected as inputs of the model. As a problem of biological origin, the presence of some filamentous bacteria (mainly *M. Parvicella*) in the anaerobic digester's inflow rate is also relevant regarding biological foaming so it was also taken into account in the AD Risk Model. This input is obtained from the risk model developed by Comas *et al.* (2008). This model used heuristic knowledge to evaluate simulation results and look for suitable conditions for the development of microbiology-related settling problems (i.e. bulking, foaming and rising sludge) in the AS system. More specifically, the AD risk model uses as input the risk of foaming related to *M. parvicella* which cause foaming in the AS (FAS risk) system as well. The basic knowledge base is presented in table 3.1. For a low FAS risk, as OLR and its variation (OLRvar) increase, the risk of foaming increases as well. Since the pilot plant treated diluted industrial distillery wastewater and was not sludge from an activated sludge system, the FAS risk was considered to be low in the validation step. Thus, the knowledge base for higher FAS risks is not presented here. However, further details on the AD Risk Model will appear in Dalmau *et al.* (2009).

Table 3.1. Knowledge base of the AD Risk Model for low FAS risk.

		OLR (kg VS·m^{-3}·d-1)				
		Very Low	**Low**	**Medium**	**High**	**Very High**
OLR var (%)	**Low**	Low	Low	Medium	Medium	High
	Medium	Low	Medium	Medium	High	High
	High	Medium	Medium	High	High	High

The AD risk model is used to assess the risk of biological foaming in AD simulation. As an example of the AD risk model performance, figure 5 shows a profile of the risk of biological foaming within the benchmark simulation model Nº2 (BSM2; Jeppsson *et al.*, 2007), however, it is out of the scope of this paper to discuss its performance. FAD risk stands for risk of biological foaming in AD. The x-axis contains one-year simulation time from July 1st. The seasonal effect of FAS (according to Hug *et al.*, 2006) can be noticed in the profile influencing the FAD risk. Although OLR is oscillating it remains in a constant range, however when OLRvar decreases (end of first summer period and middle winter) it is reflected in the FAD risk. Despite the AD risk model was developed and it can represent the dynamics of biological foaming in a simulated anaerobic digester, it was not validated yet with real data.

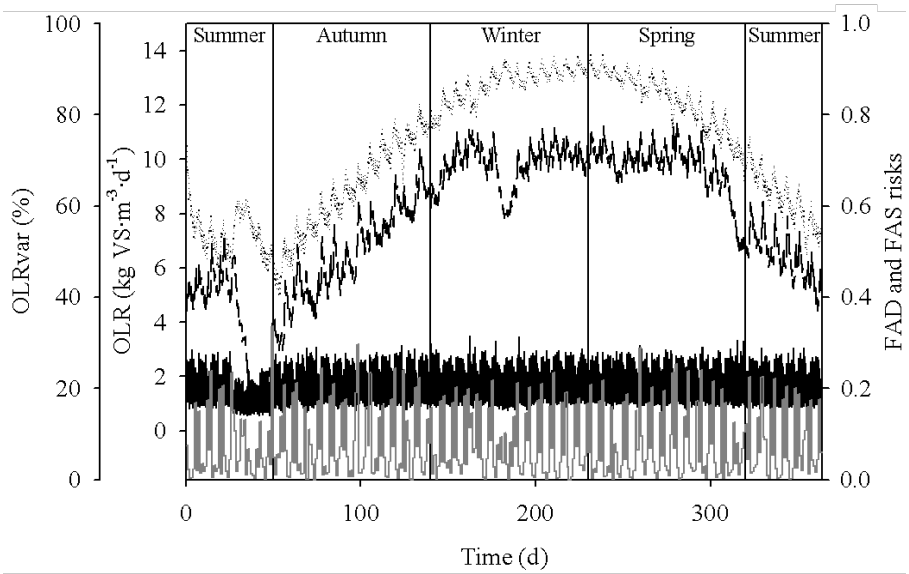


Figure 5. Simulated results of the FAD risk model for the open loop case for a one-year simulation (from July 1st). OLR (solid line); OLRvar (grey line); FAS risk (dotted black line); FAD risk (dashed line).

4. VALIDATION OF THE AD RISK MODEL

The AD risk model was developed to be implemented to BSM2, so previous to its validation with real data it was necessary to make some assumptions.

OLR and its daily variation have to be calculated from the variables measured in the pilot plant. OLR for the AD risk model is calculated as shown in Eq. 1.

$$OLR = \frac{VS}{HRT} \qquad \text{Eq.1}$$

where,
VS= Volatile Solids $kg \cdot L^{-1}$
HRT= Hydraulic Retention Time (d)

HRT in days is obtained from Eq. 2

$$HRT = \frac{V}{qIn \cdot 24} \qquad \text{Eq. 2}$$

where,
V= pilot plant volume (1000 L.)
qIn= inflow rate in $(L \cdot h^{-1})$

Since AD risk model OLR calculation is based on VS, it was necessary to transform the measured COD (tocsDig) into VS. According to Metcalf and Eddy (2003) for untreated wastewater the biological oxygen demand/total organic carbon ratio (BOD/TOC) is between 1.2 and 2.0 (1.6 was taken as the average of the rank), there is also a relation between BOD and chemical oxygen demand (COD) from 0.3 to 0.8 (0.55 was taken as the average of the rank). Therefore, putting together both ratios, COD can be expressed as a function of TOC (tocsDig in our case; Eq. 3).

$$COD = 2.9 * tocsDig \quad \text{Eq. 3}$$

where,
tocsDig: total organic carbon in the digester ($mg \cdot L^{-1}$)
COD in mg COD $\cdot$ L^{-1}.

In Copp (2002) it is pointed that there is a relation between total suspended solids (TSS) and COD from particulate compounds (Eq. 4).

$$TSS = 0.75 \cdot COD_p \quad \text{Eq. 4}$$

where,
COD_p in mg COD $\cdot$ L^{-1}.

A last assumption is made in order to simplify the conversion supposing that all the TSS can be accounted as VS. This way, we consider all the COD measured in the pilot plant can be degraded as it was VS. Thus, from Eq 3. and Eq 4. we get Eq. 5.

$$VS = \frac{2.175 \cdot tocsDig}{1000} \quad \text{Eq. 5}$$

where,
VS in $kg \cdot L^{-1}$

Data gathered during approximately almost three months was used to validate the AD risk model. Figure 6 shows the profile for both the simulated biological foaming risk (SFR) from the AD risk model and the foaming index estimated from real data (FR).

From figure 6 some aspects can be pointed. First of all, reasonable good fitting is achieved (RMSE=0.06). Secondly, it becomes clear that there are two differentiated periods, approximately the first month and the last two months. The first period is marked for an apparent stability of the process with a good coincidence between SFR and FR (both showing low foaming risk), whereas the second shows much more oscillations and peaks revealing a probably more unstable period. In this last period, in some specific points (i.e. around days 33, 43 and 58) there are some divergences in which the model shows relatively high foaming risk when the real data show low risk of foaming. It is important to note that the inherent uncertainty of the mechanisms of foaming that hinders the development of mechanistic models cannot be included in the AD Risk Model. This can be the main reason behind the main differences in the

validation results. Nevertheless, the general trends of the instability are indeed detected by the AD risk model allowing it to assess operational conditions of the anaerobic digester that can favour biological foaming.

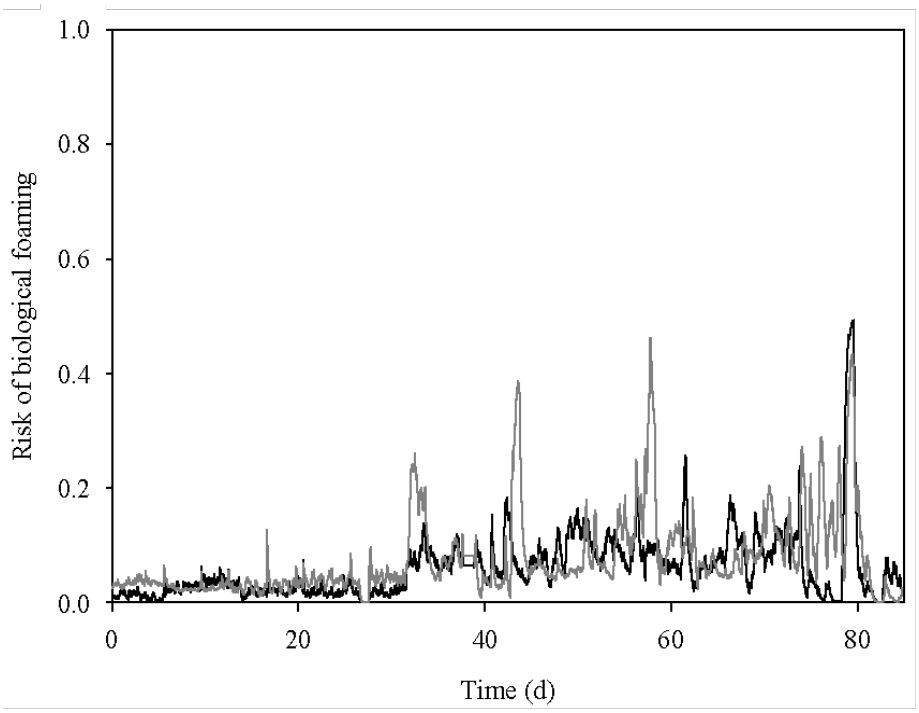


Figure 6. SFR (black line) versus FR (grey line).

5. CONCLUSIONS

The AD risk model has been validated using real data from a pilot plant. Real data has been adapted to the AD risk model and the results show that quite a good fitting of the data can be achieved, showing that the AD risk model is able to represent the general conditions of an anaerobic digester regarding biological foaming. However, further validation with real data from an anaerobic digester treating sludge from an activated sludge system would be of interest since it would allow to consider the effect of filamentous bacteria in the AD feed.

ACKNOWLEDGEMENTS

Authors wish to thank Dr. Cristian Trelea from the Institut National Agronomique Paris-Grignon for developing and providing us with the Neural Network toolbox. The work was also partially supported by the Spanish Ministry of Education and Sciences project DPI2006-15707-C02-01.

REFERENCES

Barjenbruch M and Kopplow O (2003) Enzymatic, mechanical and thermal pretreatment of surplus sludge. Adv Environ Res 7:715-720.

Carrasco E F, Rodríguez J, Puñal A, Roca E and Lema J M (2004) Diagnosis of acidification states in an anaerobic wastewater treatment plant using a fuzzy-based expert system. Control Eng Pract, 12:59-64.

Comas J, Rodríguez-Roda I, Poch M, Gernaey K V, Rosen C and Jeppsson U (2008) Risk assessment modelling of microbiology-related solids separation problems in activated sludge systems. Environ Modell Softw 23(10-11):1250-1261.

Comas J, Rodríguez-Roda I, Sánchez-Marrè M, Cortés U, Freixó A, Arráez J and Poch M (2003) A knowledge-based approach to the defloculation problem: integrating on-line, off-line, and heuristic information. Wat Res, 37(10):2377-2387.

Copp J B (2002) The COST Simulation Benchmark – Description and Simulator Manual. Office for Official Publications of the European Communities, Luxembourg, ISBN 92-894-1658-0.

Dalmau J, Comas J, Rodríguez-Roda I, Pagilla K and Steyer J P (2009) Risk model development and simulation for foaming in anaerobic digestion. Wat Res, (Submitted).

Dalmau J, Comas J, Rodríguez-Roda I, Latrille E and Steyer J P (2008) A neural network approach to select the most relevant variables for foaming in anaerobic digestion. Proceedings of the International Congress on Environmental Modelling and Software. Barcelona, Spain, 7-10 July. ISBN: 978-84-7653-074-0

Dalmau J, Comas J, Rodríguez-Roda I, Latrille E and Steyer J P (2007) Using artificial neural networks for selecting relevant information in anaerobic digestion, Paper presented at 11th IWA Specialitst Conference on Anaerobic Digestion (AD11), Brisbane, Australia, September 23-27.

Elliot A and Mahmood T (2007) Pretreatment technologies for advancing anaerobic digestion of pulp and paper biotreatment residues. Wat Res, 41:4273-4286.

Hug T, Gujer W and Siegrist H (2006) Modelling seasonal dynamics of "*Microthrix parvicella*". Water Sci Technol, 54(1):189-198.

Jeppsson U, Pons M-N, Nopens I, Alex J, Copp J, Gernaey K V, Rosen C, Steyer JP and Vanrolleghem P A (2007) Benchmark simulation model no 2: general protocol and exploratory case studies. Water Sci Technol 56(8):67-78.

Kohavi R, John G H (1997) Wrappers for feature subset selection. Art Intell 97:273-324.

Lardon L, Puñal A, Martínez J A, Steyer J P (2005) Modular expert system for the diagnosis of operating conditions of industrial anaerobic digestion plants. Water Sci Technol 52(1-2):427-433.

Massart N, Bates R, Corning B and Neun G (2006) Design and operational considerations to avoid excessive anaerobic digester foaming. 79th Water Environmental Federation Technical Exhibition and Conference.

Metcalf and Eddy (2003) Wastewater engineering: Treatment and reuse. Fourth ed., McGraw-Hill, New York, USA.

Nguyen D; Widrow B (1990) Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. Proceedings of Int. Joint Conf. Neural Networks. Washington DC, USA, January 15-19, 21-26.

Pagilla K R, Kraney K C, Kido W H (1997) Causes and effects of foaming in anaerobic sludge digesters. Water Sci Technol, 36(6-7):463-470.

Poch M, Comas J, Rodríguez-Roda I, Sánchez-Marrè M, Cortés U (2004) Designing and building real environmental decision support systems. Environ Modell Softw 19(9):857-873.

Schafer P, Uhte W, Newman G (2006) Improved anaerobic digestion feed systems and concepts. Proceedings of the 80th Annual Water Environment Federation Technical Exhibition and Conference; Dallas, Texas, USA, Oct 21-25; Water Environment Federation: Alexandria, Virginia, 374-392.

Westlund A D, Hagland E and Rothman M (1998) Foaming in anaerobic digesters caused by *Microthrix parvicella.* Water Sci Technol 37(4-5), 51-55.

Zhao H W and Viraraghavan T (2004) Analysis of the performance of an anaerobic digestion system at the Regina wastewater treatment plant. Biores Technol, 95:301-307.

Estimation of the permeability of granular soils using neuro-fuzzy system

A. Sezer[1], A.B. Göktepe[2], S. Altun[3]

Department of Civil Engineering, Ege University, Izmir, Turkey

Abstract

Determination of the permeability coefficient is crucial for the solution of several geotechnical engineering problems such as modeling of underground flow, determination of the hydraulic properties of leachate water in waste disposal areas, calculation of the compressibility, and so on. Constant head permeability test, which is usually performed for the determination of the permeability, is easy to apply; however, it is not easy to obtain undisturbed sand specimens from field. Therefore, the tests are usually employed on specimens having similar relative densities to those from the field. An alternative approach to permeability tests for granular soils is the prediction of permeability levels in terms of a number of particle size distribution and shape parameters. Although these methods are capable of making reasonable predictions for permeability coefficient, they have certain limitations. In this study, the approximation ability of neuro-fuzzy systems is utilized for the prediction of the permeability coefficient. Permeability test results on 20 different types of granular soils are used to generate a database to train adaptive neuro-fuzzy inference system (ANFIS), which is considered to predict the results of eight different permeability tests. It is concluded that ANFIS structure is superior in the prediction of permeability tests considering particle shape and grain size distribution information.

[1] Address: Ege Universitesi Insaat Muhendisligi Bolumu, 35100, Bornova, Izmir, Turkey
e-mail: alper.sezer@ege.edu.tr

[2] Address: Ege Universitesi Insaat Muhendisligi Bolumu, 35100, Bornova, Izmir, Turkey
e-mail: abgoktepe@gmail.com

[2] Address: Ege Universitesi Insaat Muhendisligi Bolumu, 35100, Bornova, Izmir, Turkey
e-mail: selim.altun@ege.edu.tr

1. Introduction

Permeability is considered one of the most important parameters in soil mechanics. Basically, it is defined by the quantity of water passing through a soil medium in a certain period, and is determined by in-situ and laboratory tests. In common practice, the permeability coefficient is usually obtained by constant-head permeability test, and is utilized in filtration-drainage, settlement, and stability calculations. These problems are extremely important for environmental aspects such as waste water management, slope stability control, erosion, and structural failure related with the ground settlement issues. In this respect, empirical equations are utilized to predict these parameters; however, these equations have certain limitations and uncertainties. In addition to the incorporation ability of the past experiences regarding to these obscure parameters, neuro-fuzzy systems enable the engineers to predict the unknown parameters belonging to these problems with its superior approximation abilities.

Although constant head permeability tests don't take much time to perform, the relationship between permeability parameter and a number of grain size distribution parameters have been investigated by several researchers (Seelheim, 1880; Hazen, 1892; Slichter, 1898; Freeze and Cherry, 1979; Carrier III, 2003; Chapuis, 2004). In Table 1, the empirical equations presented by a number of researchers are given. In addition, advantages, disadvantages and limitations of these studies are also included in the table. These formulas are capable of estimating the permeability with a reasonable precision. Therefore, an alternative and more precise technique is developed in this study using adaptive neuro-fuzzy inference methodology. Using the database obtained by several permeability tests, which include a number of particle shape and grain size distribution parameters, permeability coefficient of sands are modeled with the methodology. It should be mentioned that, because grain size distribution is the main factor affecting the void ratio and relative density of soils, regarding to the knowledge that global void ratio is primarily effective on the permeability of granular soils, the training and testing databases have similar gradations.

2. Materials and Methods

The conceived ANFIS model comprises following input parameters: a) D_{10} (the diameter which finer material is equal to 10% of the total by weight), b) D_{60} (the diameter which finer material is equal to 60% of the total by weight), c) mean roundness, d) mean sphericity, e) global void ratio (e), f) maximum void ratio (e_{max}), and g) total fractal dimension. D_{10} and D_{60} parameters are extracted from the grain size distribution of the sands. The roundness, the sphericity and total fractal dimension parameter definitions are given in another study (Sezer et al., 2008). Soils used in this study consist of 100 % sand, and the physical properties

of the sands are given in Table 2. The void ratios of the sands are extracted from the specimens prepared under Standard Proctor densification level. Sands are arranged in similar gradations to compare their permeability coefficients in relation with their shape characteristics. Three groups of the sands (tabulated as B, L and A in Table 2) are crushed materials and two of them (R and S) are natural materials. The sands are separated into four grain size distributions, and labeled as 1, 2, 3, and 4, which representing coarse, medium, fine and well gradation, respectively. As can be derived from Table 2, three of the sands of each origin (1, 2 and 3) are poorly graded and the last sand group (4) is well graded. In order to ensure 100 % coarse material inclusion, the soils in this study are washed under Sieve No.120. Moreover, gravel sized or bigger particles are not included; therefore, all the grains are passed the No.4 sieve. Detailed explanation on the origin of the soils can be found in elsewhere (Sezer, 2008).

Table 1. Empirical equations manifested for permeability prediction of soils

Researcher / Organization	Equation	Limitations, Advantages / Disadvantages
Hazen (1892)	$k = C_H d_{10}^2$	Effective diameter changes between 0.1 and 30 mm (Hazen, 1892; Carrier III, 2003).
Kenney et al. (1984)	$k = (0.05 \sim 1)\,\mathrm{d}_5^2$	D=0.074-25.4 mm and C_U=1.04-12.
Breyer-(Kresic, 1998)	$k = 6\times10^{-2}\times\frac{g}{v}\times\log\left(\frac{500}{C_u}\right)\times d_{10}^2$	C_u= 1~20, d_{10}= 0.06~0.6 mm.
Slichter (1898)	$k = \frac{g}{v}\times n^{3.287}\times d_{10}^2$	best suited for soils with d_{10} =0.01 ~5 mm (Vukovic & Soro, 1992)
Chapuis (2004)	$k = 1.5\times d_{10}^2\times\frac{e^3}{1+e}\times\frac{1+e_{\max}}{e_{\max}^3}$	N/A
NAVFAC (Chapuis et al. , 1989)	$k = 10^{1.291e-0.6435}\, d_{10}^{\,10^{0.5504-0.2937e}}$	e=0.3~0.7; d_{10}=0.10~2.0 mm; C_u=2 ~12; and d_{10}/d_5 >1.4
Terzaghi- (Odong, 2007)	$k = 0.0084\times\frac{g}{v}\times\left[\frac{n-0.13}{(1-n)^{1/3}}\right]^2\times d_{10}^2$	The selected average value of 0.0084 is actually a classification coefficient typically ranging between 0.0061 and 0.00107.
USBR- (Vukovic and Soro, 1992)	$k = 0.048\times\frac{g}{v}\times d_{20}^{0.3}\times d_{10}^2$	Gives the best results when C_u is lower than 5 (Cheng and Chen, 2007)
Alyamani and Şen (1993)	$k = 1.5046*(I_0 + 0.025*(d_{50}-d_{10}))^2$	The method is more accurate for well-graded sample (Odong, 2007).
Kozeny-Carman (1956)	$k = 0.083\times\frac{g}{v}\times\left[\frac{n^3}{(1-n)^2}\right]\times d_{10}^2$	d_{10}<3 mm., for granular soils, the inertia term is not taken into account (Carrier III, 2003).

For investigating the effect of particle shape on the permeability of soils, constant head permeability tests are employed on the soil specimens in accordance with ASTM D2434-68 standard. The combination permeameter is utilized to perform

permeability tests on specimens of 31.65 cm^2 cross-section and of 10-11 cm in length, which were prepared at Proctor density level.

Permeabilities of granular soils (k) are computed in accordance with D'arcy's Law:

$$Q = k\frac{h}{\ell}a \qquad (1)$$

where Q is the discharge, a is the cross sectional area of the specimen, h is the hydraulic load on the specimen and ℓ is the length of the specimen. The test apparatus is given in Figure 1.

Table 2. Physical properties of the soils used in this study (Sezer, 2008).

Origin	Sand type	C_u	C_c	Gravel %	Sand %	Silt-Clay %
Limestone (L)	1	1.55	0.94	0	100	0
	2	1.61	1.03	0	100	0
	3	1.67	0.90	0	100	0
	4	8.00	1.22	0	100	0
Basalt (B)	1	1.55	0.94	0	100	0
	2	1.61	1.03	0	100	0
	3	1.67	0.90	0	100	0
	4	11.82	1.18	0	100	0
Andesit (A)	1	1.55	0.94	0	100	0
	2	1.61	1.03	0	100	0
	3	2.33	0.76	0	100	0
	4	6.00	1.03	0	100	0
River Sand (R)	1	1.55	0.94	0	100	0
	2	1.61	1.03	0	100	0
	3	1.67	0.90	0	100	0
	4	8.50	1.98	0	100	0
Shore sand (S)	1	1.55	0.94	0	100	0
	2	1.61	1.03	0	100	0
	3	1.82	0.92	0	100	0
	4	6.68	1.04	0	100	0

The permeability test results and the corresponding empirical equation outcomes are given in Figure 3. The coefficient of permeability (k) of coarse grained sands is higher, in comparison with medium and fine sands. Least coefficients are observed for well graded soils. Analyzing the results given in Figure 3, it can be concluded that the given formulas are capable of estimating the *k* parameter to a reasonable degree. Nevertheless, the outcomes of different equations are still far from the equality. The Breyer formula is not taken into account for the predictions in coarse and fine sands, where USBR formula is not used for the permeability estimation of well graded sands (Vukovic and Soro, 1992). The investigations, which are graphically demonstrated in Figure 3, indicate that the outcomes of Chapuis and Slichter methods are in harmony with the test results. Investigating the test results on medium sands (Figure 3b), it can be stressed that NAVFAC

formula rearranged by Chapuis et al. (1989) and Kozeny-Carman (1956) methods are quite successful. Furthermore, the test results given in Figure 3c indicate that Chapuis method best predicts the permeability coefficient of fine sands.

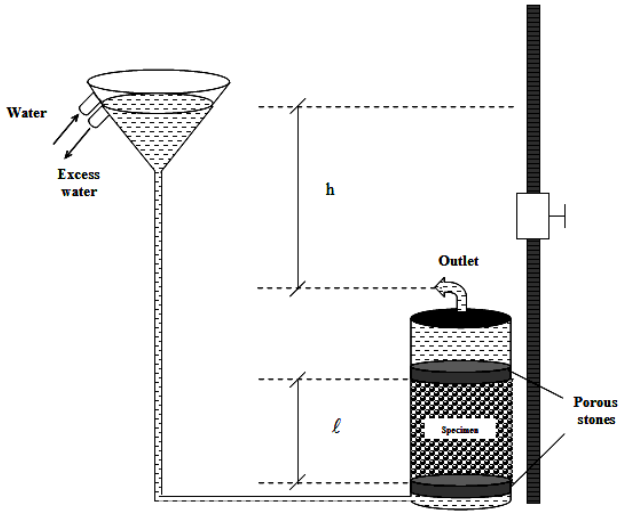


Fig. 1. The combination permeameter- constant head situation**.**

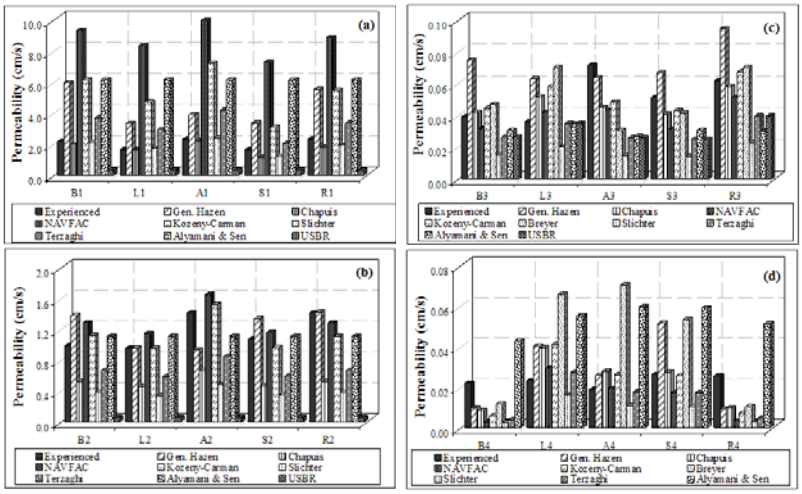


Figure 3. Permeability test results and the outcomes of the empirical formulae on a) Coarse b) Medium c) Fine d) Well graded sands

For all poor graded sands, the USBR method underestimated the permeability test results. Alyamani & Sen method has underestimated permeability in coarse grained sands, while the method overestimates the permeability coefficients in fine grained sands. Nevertheless, despite the good predictions in poor graded sands, Chapuis (2004), Alyamani & Sen (1993) and USBR methods give similar permeability coefficient predictions. In this manner, great care should be taken for the use of these formulas in permeability prediction. Moreover, for well graded soils, the estimates are far from proposing a formula (Figure 3d).

As can be seen from Figure 3, the differences in the results and the limitations on the empirical formulae encourage seeking for a new alternative approach for the permeability prediction. In this investigation, ANFIS is considered to be a possible and plausible alternative to traditional techniques.

3. Adaptive neuro-fuzzy inference system and the model

Fuzzy sets have the advantage of use of the "partial belonging concept", instead of crisp belongingness. The membership value of a data point is the measure of the belongingness of the point to any set. The Adaptive Neuro-Fuzzy Inference System (ANFIS) is emerged as a powerful technique which couples Artificial Neural Network (ANN) and Fuzzy Inference System (FIS) methodologies (Jang and Sun; 1995). The NN-based learning technique and a fuzzy inference methodology which using membership function and a fuzzy rule-base philosophies, is used to establish nonlinear relationships between input and output spaces (Jang, 1993; Jang and Sun, 1995). Therefore, the ANFIS is capable of organizing self-structures in terms of the rules and membership functions with the help of input and output data patterns.

In this study, the model is setup using backpropagation learning algorithm and Sugeno-type fuzzy inference system. Details of Sugeno type neurofuzzy structure can be found elsewhere (Jang, 1993).

Consider a Sugeno type of fuzzy system based on the rules:

Rule 1. If x is A_1 and y is B_1, then $f_1=p_1x+q_1y+r_1$

Rule 2. If x is A_2 and y is B_2, then $f_2=p_2x+q_2y+r_2$

Premises results can be computed by :

$$w_1 = \mu_{A_1}(x)\mu_{B_1}(y); \quad w_2 = \mu_{A_2}(x)\mu_{B_2}(y) \tag{2}$$

The weighted average may be calculated as:

$$f = \frac{w_1 f_1 + w_2 f_2}{f_1 + f_2} \tag{3}$$

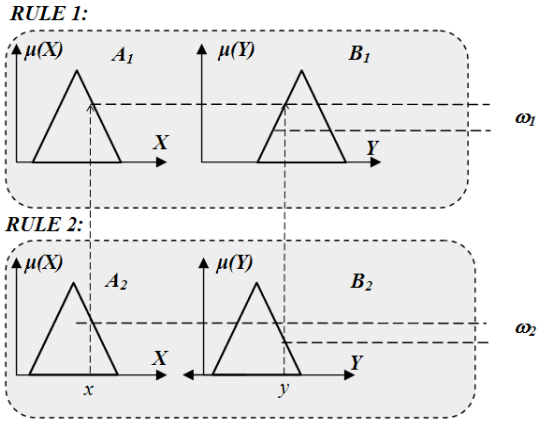


Figure 4. Sugeno-fuzzy model with two rules.

Definition of this parameter into diverse phases can be formulated by:

$$\overline{w}_i = \frac{w_i}{w_1 + w_2} \tag{4}$$

Then output can be obtained as:

$$f = \overline{w}_1 f_1 + \overline{w}_2 f_2 \tag{5}$$

Figure 5 demonstrates the corresponding ANFIS architecture to the rules in Figure 4. The circular and square nodes represent the fixed and learnt nodes, respectively.

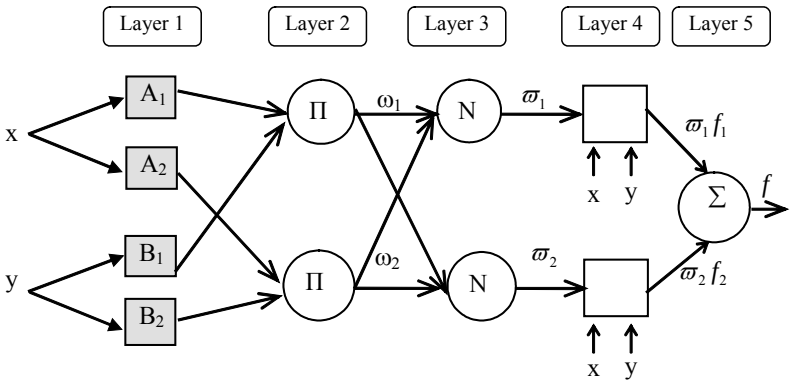


Figure 5. The ANFIS architecture for a two rule Sugeno system

Since the lack of space, no further explanation on the ANFIS calculation will be given here. Detailed explanation is given in Jang (1993). As mentioned before, the input parameters of the model are: D_{10}, D_{60}, r, s, e, e_{max} and D_{tot}, while the output is the permeability coefficient (k). Focusing on the three dimensional surface graphs of the two selected input and one output, it is possible to conclude that, there are nonlinear relationship among the parameters, and coefficient of permeability is highly effected by high levels of mean sphericity and void ratio (Figure 6). Similar behaviors are observed at other three-dimensional surface plots; nevertheless, no other surface graph is presented here for space problem.

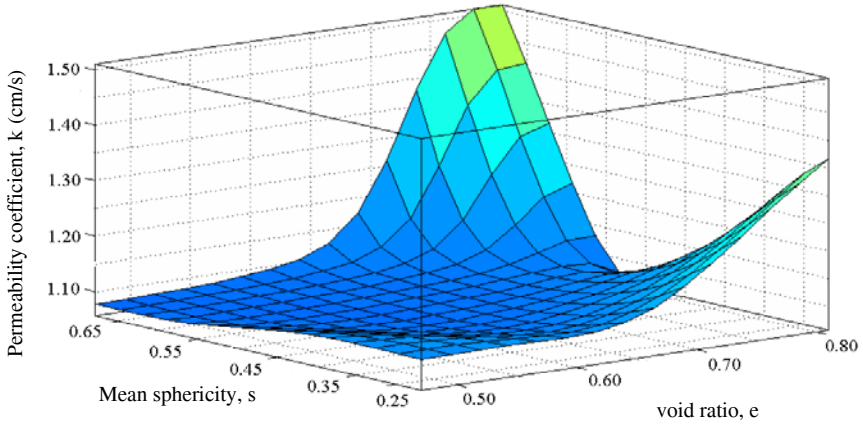


Figure 6. The surface graph of the output against two input parameters.

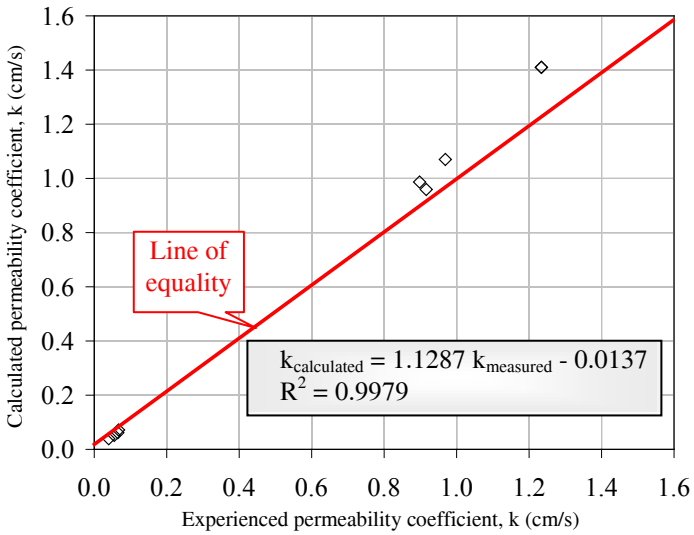


Figure 7. The comparison of calculated and experienced permeability coefficients.

The testing of model is employed on data having similar grain size distribution. The calculated and permeability test results are given below. Extremely high R^2 value of 0.9979 approves the permeability coefficient modeling capability of the model (Figure 7).

4. Conclusions

In this investigation, ANFIS methodology is applied on uniform and well graded sands in order to propose an alternative approach for the permeability coefficient estimation. Regarding to the 28 tests on prepared sands, permeability tests are conducted to find out the hydraulic conductivities of these soils. Empirical permeability coefficients are discussed in summary for the evaluation of alternative methods for permeability prediction in the literature. Moreover, graphs illustrating the success of these equations are drawn to question the success of these methods. It is concluded that, these methods are capable but not sufficient for correct prediction of k parameter, in terms of their narrow prediction level and limitations in application of grain size distribution. Adaptive neuro-fuzzy inference system is used for k parameter prediction, using a number of particle shape and grain size distribution parameters. It is also derived that, ANFIS is successful in permeability prediction of soils having similar gradation. It is clear that the modeling ability of the ANFIS model strictly depends on the data, i.e. grain size distribution and the shape of the sand grains. On the other hand, it should be noted that larger database development is essential for accurate prediction of *k* parameter for a wide-range grain sizes (including gravels and sands of different median sizes).

Acknowledgements

The authors are thankful to TUBITAK, Ege University Research Funds, and Ebiltem for their support in the image processing part of this study with 105M013, 2005MUH021 / 2007MUH012 and 2006BIL001 projects, respectively.

References

Alyamani MS, Sen Z (1993) Determination of hydraulic conductivity from grain-size distribution curves. Ground Water 31: 551-555.

ASTM Standard D2434-68 (2006) Standard Test Method for Permeability of Granular Soils (Constant Head). ASTM International, West Conshohocken, PA, DOI: 10.1520/D2434-68R06, www.astm.org.

Carman PC (1956) Flow of gases through porous media. Butterworths Scientific Publications, London.

Carrier III WD (2003) Goodbye, Hazen, Hello, Kozeny-Carman. *ASCE* J Geotech Geoenviron 129: 1054-1056.

Chapuis RP, Gill DE, Baass K (1989) Laboratory permeability tests on sand: influence of the compaction method on anisotropy. Can Geotech J 26: 614-622.

Chapuis RP (2004) Predicting the Saturated Hydraulic Conductivity of Sand and Gravel Using Effective Diameter and Void Ratio. Can Geotech J 41: 787-795.

Freeze RA, Cherry JA (1979) Groundwater. Prentice Hall, New Jersey.

Hazen A (1892) Some physical properties of sand and gravel, with special reference to their use in filtration. Massachusettes State Board of Health, 24th Annual Report, Boston.

Jang JSR (1993) ANFIS: Adaptive-network-based fuzzy inference systems. IEEE T Syst Man Cyb 23: 665–85.

Jang JSR, Sun CT (1995) Neuro-fuzzy modeling and control. P IEEE 83: 378–405.

Kenney TC, Lau D, Ofoegbu GI (1984) Permeability of compacted granular materials. Can Geotech J 21: 726-729.

Kresic N (1998). Quantitative Solutions in Hydrogeology and Groundwater Modeling. Lewis Publishers, Florida.

NAVFAC (1974). Design Manual- Soil Mechanics, Foundations and Earth Structures (DM-7). US Department of Navy Printing Office, Washington.

Odong J (2007) Evaluation of the empirical formulae for determination of hydraulic conductivity based on grain size analysis. J Am Sci 3: 54-60.

Seelheim F (1880) Methoden zur bestimmung der durschlassigkeit des bodens. Z analy chem 19: 387-402.

Sezer A (2008). Determination of microstructural properties of different types of soils by image processing techniques. Doctor of Philosophy dissertation, Ege University, Turkey.

Sezer A, Altun S, Göktepe AB, Erdogan D (2008) The correlation between CBR Strength and fractal dimensions of sands", The 12th International Conference of International Association for Computer Methods and Advances in Geomechanics (IACMAG), Goa – India, 1-6 October, *CD-ROM.*

Slichter CS (1898). Theoretical investigation of the motion of ground waters. 19th Annual Report. U.S. Geology Survey, USA.

Vukovic M., Soro A (1992) Determination of Hydraulic Conductivity of Porous Media from Grain-Size Composition. Water Resources Publications, USA.