UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LUCAS PUGENS FERNANDES

A Clustering-based Approach to Identify *Petrofacies* from Petrographic Data

Thesis presented in partial fulfillment of the requirement for the degree of Master of Computing Science

Advisor: Prof. Dr. Mara Abel
Coadvisor: Prof. Dr. Joel Luís Carbonera

Porto Alegre
2020

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Pugens Fernandes, Lucas

A Clustering-based Approach to Identify Petrofacies from Petrographic Data / Lucas Pugens Fernandes. – Porto Alegre: PPGC da UFRGS, 2020.

57 f.:il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2020.

I. Abel, Mara. II. Luís Carbonera, Joel. III. Título.

ABSTRACT

In this work, we evaluate methods to automatically identify petrofacies from a dataset of thin section rocks description. Computationally, we face the task of petrofacies identification as an unsupervised problem; thus, we focus our efforts on the application of adapted classic clustering methods for the task. We developed a pre-processing approach using the domain ontology to reduce the feature-space, significantly improving the execution time and ease of interpretation, while maintaining the accuracy. All the data used during the experiments come from six real datasets extracted from multiple basins throughout the world. Our results show that the data pre-processing using a domain ontology can drastically reduce the feature-space and execution time while keeping the relevant information for the expert user. We develop a well-founded analysis of candidate algorithms, such as the classical K-Means to their combination with Wrapper Genetic Algorithms simultaneously selecting features and grouping clusters. The experiments suggest good promises in the automation of the petrofacies grouping task. However, there are challenging aspects for the application of clustering and feature selection in this domain, pointing to the need for new future research in this field.

Keywords: Reservoir petrofacies. Feature selection. Ontology. Clustering. Genetic algorithm.

Identificação Automatizada de *Petrofacies* Através de Clustering de Dados Petrográficos

RESUMO

Neste trabalho, nós avaliamos métodos para realizar a identificação automática de petrofácies a partir de um conjunto de dados de descrição de seções de rocha delgada. Computacionalmente, enfrentamos a tarefa de identificação de petrofácies como um problema de clustering; portanto, concentramos nossos esforços na aplicação de métodos de clustering para a solução do problema. Desenvolvemos um pré-processamento usando a ontologia de domínio para reduzir o espaço de atributos, melhorando significativamente o tempo de execução e a facilidade de interpretação, mantendo a precisão. Todos os dados usados durante os experimentos vêm de seis conjuntos de dados reais extraídos de várias bacias ao redor do mundo. Nossos resultados mostram que o pré-processamento de dados usando uma ontologia de domínio pode reduzir drasticamente o espaço de recursos e o tempo de execução, mantendo as informações relevantes para o usuário especialista. Desenvolvemos uma análise bem fundamentada dos algoritmos candidatos, como o K-Means clássico, para sua combinação com os algoritmos genéticos, selecionando simultaneamente atributos e agrupamento de petrofacies. Os experimentos sugerem promessas na automação da tarefa de agrupamento de petrofácies, no entanto, existem aspectos desafiadores para a aplicação de clustering e seleção de features neste domínio, apontando para a necessidade de futuras pesquisas neste campo.

Palavras-chave: Petrofacies de reservatório. Seleção de atributos. Agrupamento. Algoritmo genético.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AHC	Agglomerative Hierarchical Clustering
ARI	Adjusted Rand Index
CB	Campos Basin
DM	Data Mining
DHC	Divisive Hierarchical Clustering
EM	Equatorial Margin
GA	Genetic Algorithm
HC	Hierarchical Clustering
RI	Rand Index
PCA	Principal Component Analysis

CONTENTS

1 INTRODUCTION

In this work, we explore the application of clustering techniques to automate the task of identifying *petrofacies* in thin section rock samples. Reservoir *petrofacies* is an important conceptual tool applied by geology experts during the exploration phase of new reservoirs. It allows experts to elevate the abstraction level of the diverse rock segments found in the environment of a petroleum reservoir. However, *petrofacies* identification demands specialized knowledge and takes weeks to months to be applied appropriately.

The *petrofacies* identification is mainly based on petrographic descriptions of reservoir rocks, which describes and quantifies the mineralogical composition of the rock and texture aspects that influences the quality of the petroleum reservoir. These features are results of the provenance of sediments and the conditions of a rock deposition and lithification in that site. In this context, *petrofacies* is a pattern of features associated with the reservoir porosity that a petrographer identifies with the same aspect on multiple reservoir samples (DE ROS; GOLDBERG, 2007). Multiple petrofacies recognized over a reservoir can provide a useful abstraction of the reservoir composition in terms of reservoir quality. *Petrofacies* are not reusable for different reservoirs, making this a fundamentally unsupervised task.

Clustering analysis is the unsupervised branch of Machine Learning which aims to group related objects, commonly abstracted as points in space with coordinates composed by multiple feature values. Applications of clustering techniques range from server defense against multiple accesses with the same pattern, likely related to Denial of Service attacks (LEE *et al.*, 2008), to gene expression and its relation to health conditions (SORURI; SADRI; ZAHIRI, 2018). Clustering analysis differs from classification techniques mainly because the method does not require the definition of the classes beforehand, making clustering analysis an unsupervised task. We computationally formulate the task of automated petrofacies identification as a task of clustering the thin-section samples, where each cluster represents a *petrofacies* containing similar samples. The data naturally is composed of many features for a relatively small number of instances because the extraction of well cores is expensive and the experts analyze it at a microscopic level. Such nature turns this into a challenging task, as with the increase of features, the so-called “curse of dimensionality” plays a significant role in computational tasks, such as clustering and feature selection.

Our approach acknowledges that not all described petrographic features are relevant to the *petrofacies* separation process. It is then necessary to identify a suitable subspace (constituted by a subset of the complete set of features that describes the dataset) for performing the clustering process and identify meaningful clusters. A naive way of identifying such a subspace is adopting a brute force approach. However, the search space for exhaustive search in all subsets of features is prohibitively ample. Considering n as the number of features in the dataset, a brute force approach would have to test 2^n subspaces. Genetic Algorithms (GA) (MITCHELL, 1998) are commonly adopted for this kind of task, such as in (LIN *et al.*, 2013; LLETI *et al.*, 2004; MORITA *et al.*, 2003), reaching near-optimal solutions on large search-spaces, if it is given an appropriated fitness function. In this work, we implemented a GA for feature subset selection with a fitness function that performs Agglomerative Hierarchical Clustering (AHC) (HÄRDLE; SIMAR, 2007) for each individual, and we experiment it with multiple internal metrics as the fitness weight.

By observing the experts' application of the concepts, we identified well-founded criteria to select suitable algorithms for the problem, propose a feature selection and clustering algorithm based on Hierarchical Clustering (HC), GA, and a feature reengineering for datasets based on ontologies. We compare the proposed algorithm with classical approaches available in the literature. In this comparison, we considered six distinct datasets characterized by different sets of features. The classification (*petrofacies* assignment) of each example of such datasets is known, allowing the computation of external cluster evaluation metrics, such as the adjusted Rand index. According to the evaluation of geologist experts, our approach achieves meaningful geological results, and it provides a practical speed-up of the process while maintaining its coverage of different scenarios, requiring as input only the desired number of *petrofacies* and the dataset to be grouped. Optionally, the pre-processing step developed also requires an ontology describing the dataset features.

We collected the datasets explored from geological reports of real-world cases of petroleum exploration studies of sedimentary basins of South America and Europe. The studies include systematic petrographic descriptions captured using Petroledge® system (DE ROS; ABEL; GOLDBERG, 2007), an ontology-based software application for reservoir characterization. The ontology defines the qualities and domain of variation of the significant aspects of the rock, creating a standard format in which several types of reservoirs can be compared.

We also applied the ontology of Petroledge® (ABEL, 2001) to understand and extract meaningful numerical information from the raw data (e.g., groups of minerals that affect the

reservoir quality as a whole). In this way, we develop an approach that applies domain-specific knowledge directly to the dataset as a pre-processing step to re-engineer features, enhancing the quality of results.

This work was developed with the close support of geology experts to achieve useful insights into the field while analyzing the algorithm stability over multiple datasets representing different scenarios further described in Chapter 5.

The contribution of this research is the development and validation of a method to process geological data and extract *petrofacies* groups using ontologies as a pre-processing step. We found that our method, implemented as an algorithm, can extract meaningful *petrofacies* groups, greatly reduces the execution time, keeping similar scores on our quality metrics.

Chapter 2 provides the geological interpretation of the *petrofacies* identification process and describes the process as currently adopted in the industry. Chapter 3 describes the necessary knowledge about clustering analysis and related works. In Chapter 4, we present related works and their main contributions. Chapter 5 contains the methodology of development for the algorithm. Chapter 6 includes the experimental results that guided the purposed methodology during research as well as comparisons of the purposed method using classical algorithms, and Chapter 7 concludes this work with a general view of the contributions and the main lines of future work in the field.

2 RESERVOIR PETROFACIES

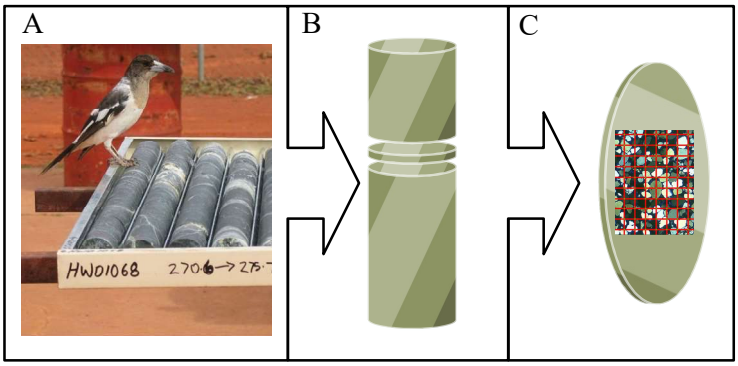
Petroleum exploration is an activity highly dependent on the planning of companies running exploration sites. As such, it is a significant concern for such companies to evaluate the economic potential of their assets. Petroleum companies assess the potential of a petroleum field based on three types of information: (1) the size of the geological body that contains petroleum that determines the potential amount of oil that can be extracted; (2) the porosity and permeability of the rock (generally mentioned as the *quality* of the reservoir) that defines the strategy and a realistic approximation to the amount of oil production; and (3) the price of petroleum barrel in the international market that, along with the quality of petroleum, defines the limits of investment required to put the field in production.

This work focuses on approaches that deal with the information in point (2), previously mentioned, help geologists to identify features of the rock responsible for the control of the porosity and permeability of a reservoir rock. Besides supporting the evaluation of porosity and permeability, petrofacies evaluation helps in defining strategies of production, since it helps to understand the internal heterogeneity of the reservoir rock. Our proposal utilizes petrographic qualitative and quantitative descriptions of reservoir rocks, which are the compositional and textural aspects of the rock described at the microscopic scale; aspects that are responsible for modifying the quality aspects of the reservoir.

In order to understand reservoir occurrence, geologists have developed conceptual tools to study and predict the quality of reservoirs. One such tool is the *petrofacies* concept proposed by de Ros and Goldberg (2007). A *petrofacies* is a group of petrographical features (composition and texture) visually recognized in thin sections of rocks and that show a particular signature in geophysical well logs. The *petrofacies* repeat themselves as a pattern through the reservoir associated with levels of porosity and permeability. This behavior allows the geologist to (1) define the internal variation of the porosity/permeability of the reservoir, and (2) extrapolate the scale of thin-section to the reservoir scale using the well log signature associated with a particular petrofacies.

The signature of petrofacies is particular to each sedimentary basin, which requires that the experts analyze and define new *petrofacies* for each reservoir occurrence. This characteristic excludes the possibility of using supervised model-learning techniques to learn the petrofacies in one basin and apply the model to some other occurrence for classifying the new samples.

Figure 1 – Process of extraction of thin sections. (A) shows a few samples of rock cores extracted from a well. (B) presents the cross-sectional cut of the rock cores. (C) shows the thin section sample extracted.

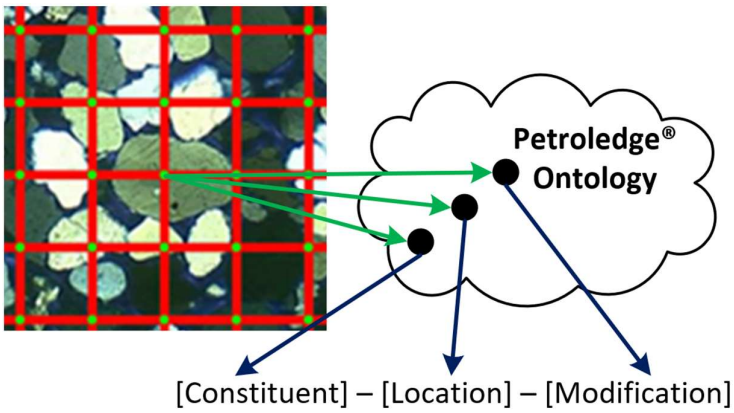


Sources: (A): Wikipedia[1], (B and C): Author.

The geologist performs a petrographic analysis on the optical microscopy of samples (thin sections Figure 1) extracted from sections of wellbore cores (contiguous rock sections extracted from a wellbore and organized in boxes, such as depicted in Figure 1) during the exploration phase of petroleum production. The petrographic analysis allows the qualitative (percentage) characterization of the composition (kinds of minerals that form the rock) and textural aspects that affect the quality of a reservoir. A systematic analysis of over 300 points equally distributed over the thin section (Figure 2) provides a quantitative distribution of the described features. The complete analysis of the set of points constitutes a sample description that will support the interpretation of the reservoir quality. Then, the geologist associates each resulting group of samples to the *petrofacies* that best reflects the features that influence the level of quality of the reservoir.

Experts define reservoir *petrofacies* by the combination of specific depositional structures, textures and primary composition, with dominant diagenetic processes. The combination of primary textural and compositional visual aspects with specific diagenetic processes and products correspond to defined ranges of porosity and permeability of petroleum reservoirs and contour rocks, while also showing a particular log and seismic signatures. shows the identification of three petrofacies of the Uerê Formation, Devonian of Solimões Basin, North of Brazil. The uniform original composition is strongly affected by the diageneses, developing three different levels of quality in the reservoir. The plot shows the characteristics of each petrofacies in terms of silica cement and intergranular volume. The quantitative analysis through point counting of the compositional and textural aspects abstracts visual into numerical and statistically relevant information about the rock under analysis.

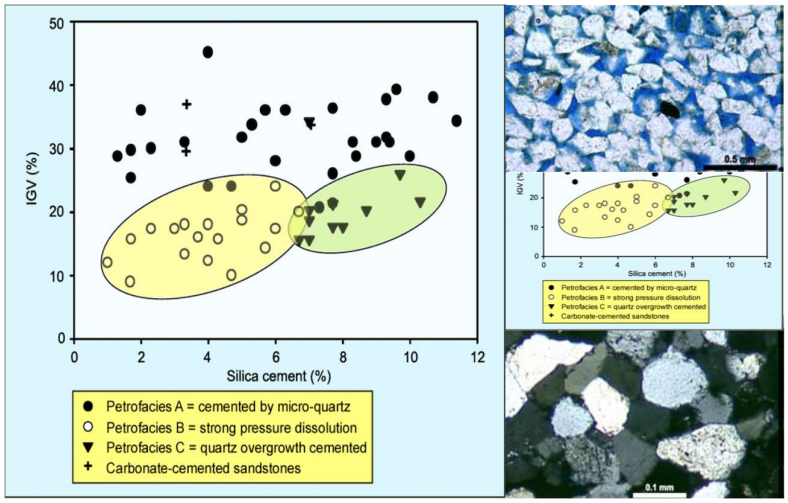
Figure 2 - Mapping of a primary constituent to its feature using the domain ontology.



Source: Author.

With the demand for increased speed in the exploration of new reservoirs, new software tools were developed to aid the experts during the collection and interpretation of the data. Petroledge®[2] (DE ROS; ABEL; GOLDBERG, 2007) is one such software tool that aids the experts to input, store, and generate interpretations about the data inserted through the point-counting system described above. The Petroledge system is implemented with a restricted vocabulary based on a domain ontology that allows the systematic translation of visual features of the rock to quantifiable petrographic features that can be examined using numerical AI methods.

Figure 3 – Reservoir petrofacies of Uerê sandstones represented in a diagram of intergranular volume by volume of silica.



Source: (DE ROS; GOLDBERG, 2007).

[2] Petroledge is a trademark of Endeeper Co.

Nowadays, experienced geologists separate petrofacies by manually exploring the vast space of features for identifying the features that represent the reservoir characteristics of the rock. Such work is time-consuming and error-prone since reservoir datasets consistently achieve hundreds of different features and experts inherently introduce bias during the analysis. To the best of our knowledge, no commercial tool offers an automated or aided *petrofacies* extraction and grouping process.

It is possible to notice the analogous relation of the petrofacies identification task and clustering techniques, where algorithms autonomously find an optimal or near-optimal solution for the best grouping of samples. Because of this similarity, we propose in this work an unsupervised approach for grouping petrofacies based on clustering techniques. It does not depend on pre-classified samples. The only required parameter is the number of desired petrofacies to be identified.

3 THEORETICAL FOUNDATION

Our work uses a combination of techniques from the field of Data Science and Knowledge Engineering combined to reach good and useful results presented through the rest of this dissertation. Knowledge Engineering was especially useful when dealing with this data to reach structural soundness while obtaining performance improvements when processing the data. Data Science offers us decades of research available when dealing with all kinds of datasets from all fields of knowledge.

This section firstly explains formally the clustering task as well as the metrics used through this work, followed by a discussion on criteria for selecting the best clustering approach for the task of petrofacies clustering. Lastly, we present a general introduction to the usage of metaheuristics for feature selection.

3.1 Clustering

Clustering techniques are applied to discover groups in data sets and to identify abstract structures that represent them, assuming only a statistical model of distribution, and no other prior knowledge about this data. These techniques enable users to reveal interesting patterns on the data structures and extract knowledge from large datasets by grouping (clustering) similar samples (See). They are especially useful when analyzing very high dimensional and sparse data, trying to identify the differences and similarities among the instances.

In this section, we use the notation of uppercase letters to refer to sets, and subscripted lowercase letters to represent individuals from such set, e.g., $a_i \in A$ and $b_i \notin A$. The indicator function $\mathbf{1}(\theta)$ is defined as 1 if the logic expression θ is true and 0 if θ is false.

Given the set $N = \{n_0, n_1, \dots, n_I\}$, each sample n_i represented by a set of Q numerical features $F = \{f_0, f_1, \dots, f_Q\}$, such samples are divided into $K = \{k_0, k_1, \dots, k_J\}$ groups. Lastly, $Y = \{y_0, y_1, \dots, y_I\}$ and $\hat{Y} = \{\hat{y}_0, \hat{y}_1, \dots, \hat{y}_I\}$ are the representation of the ground truth clustering (as manually made by the expert) and the clustering result from the algorithm, respectively. We define the notation $y_i = j$ and $\hat{y}_i = j$ as meaning that the golden and automatic clustering decided to assign the sample n_i to cluster k_j.

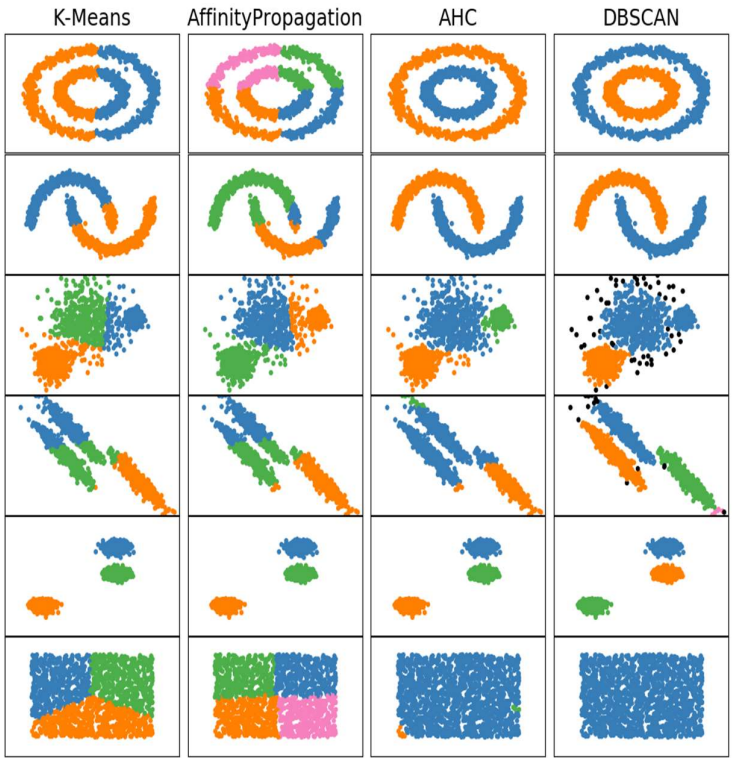
A cluster centroid – also called barycenter – G^N, is the average point calculated by the average of all samples' features in the set S s.t. $S \subset N$.

$$G^S = \frac{1}{|S|}\sum_{i=0}^{S} n_i \quad (1)$$

3.2 Clustering Algorithms

Families of clustering algorithms generally try to solve the clustering problem from different points of view, making different assumptions about what is a cluster. In order to select the best clustering algorithm, we must consider some requirements that are given by the intrinsic nature of data. Figure 4 shows the diversity of what can be considered clusters, as well as the characteristics of the clustering algorithms studied.

Figure 4 – Diversity of clusters and algorithms behaviors. Five synthetic 2D datasets are representing general cluster shapes. The point colors represent the assigned clusters. Black dots are considered outliers by the algorithm and were not assigned to any cluster.



Source: Adapted from Scikit-Learn documents[3].

K-Means is a well-known approach that is representative of the centroid-based clustering algorithms, where the objective is to form clusters with minimal square distance to the cluster

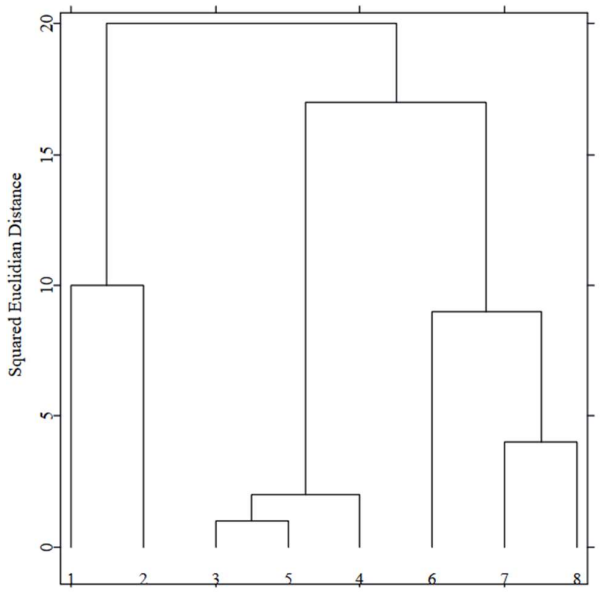
[3] https://scikit-learn.org/stable/modules/clustering.html

center (inertia value). Combined with smart initial centroid choices (K-Means++ from Arthur & Vassilvitskii, 2007), it tends to reach better results. K-Means works by looping two main steps until it reaches convergence (clusters assignments no longer change):

1. Assign each sample to the centroid with the smallest squared distance, e.g., Euclidean distance, to it;
2. Recalculate the position of the cluster's centroids using the average position of their assigned samples.

Generally speaking, K-Means try to minimize the global sum of distances from samples to their cluster's centroid, also referred to as inertia. K-Means is a greedy algorithm; consequently, it is non-optimal, i.e., it does not guarantee minimum inertia. However, multiple initializations are known to reduce the overall inertia. Dozens of initializations with random starting clusters tend to get close to optimal solutions. The solution with the best inertia value associated is selected as the final clustering solution.

Figure 5 – Dendrogram example of eight samples using single linkage and squared Euclidean distance.



Source: (HÄRDLE; SIMAR, 2007).

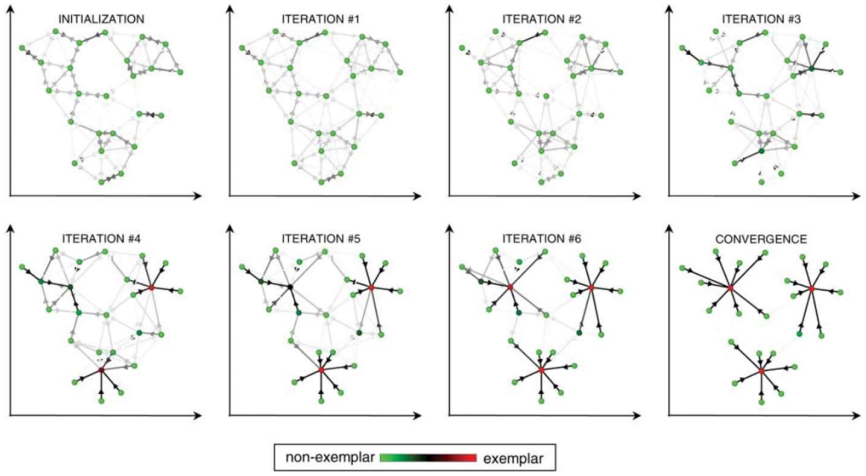
HC is based on hierarchical structures that connect all samples in the dataset. Those hierarchical structures, known as dendrograms (See Figure 5), are formed during the clustering process. HC algorithms are mainly divided into two classes according to the process of cluster formation. The Agglomerative HC (AHC) algorithm process starts with each sample belonging to a singleton cluster. The algorithm uses the distance metric d to calculate a distance matrix

between clusters. At the same time, the $linkage$ parameter decides the inter-cluster similarity, based on which the algorithm agglomerates the most similar clusters. The distance matrix is then updated and the algorithm repeats the process until there are K clusters. The Divisive HC (DHC) algorithm starts with all the samples belonging to the same cluster and the cluster is iteratively partitioned until the desired number of clusters is reached.

The $linkage$ criterion parameter of AHC algorithms can be the $average$ distance between clusters. The *minimum* distance between clusters ($single$). The *maximal* distance between clusters ($complete$) and $ward$, which group clusters that after the merge will have a minimal increase in their intra-cluster variance.

The Affinity Propagation algorithm from Frey and Dueck (2007a) simulates an "election" where data samples exchange two kinds of messages: $responsibility$ and $availability$. Responsibility messages from sample i to sample j show how good sample i is as a cluster center to sample j, while availability messages from sample i to sample j show how good the sample i is to integrate the possible cluster with the center in the sample j. Those messages are exchanged iteratively until the algorithm achieves convergence of clusters. Frey and Dueck (2007b) shows experiments for this algorithm in face recognition by clustering, gene clustering, text similarity, and airport routing.

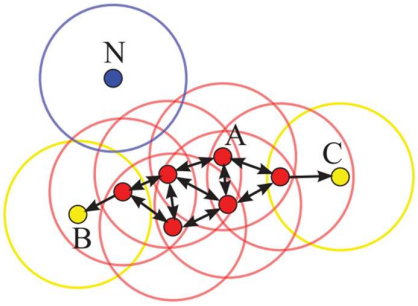
Figure 6 – Affinity propagation is illustrated for two-dimensional data points, where negative Euclidean distance (squared error) was used to measure similarity. Each point is colored according to the current evidence that it is a cluster center (exemplar). The darkness of the arrow directed from point i to point k corresponds to the strength of the transmitted message that point i belongs to exemplar point k.



Source: (FREY; DUECK, 2007a).

DBSCAN, from Schubert *et al.* (2017), is a density-based approach. This algorithm's main parameters are the radius ε and the $minPts$. The radius ε defines the maximum distance between points that are considered $density\ reachable$, while the $minPts$ is the minimum number of density reachable points a single point has to have to be considered a core point. A cluster is then defined as a set of density reachable core points with their directly reachable non-core points. Non-core points with no density reachable core points are considered outliers with no cluster defined. Figure 7 shows a simple example of a cluster formed by the DBSCAN algorithm, as well as visualization guides.

Figure 7 – DBSCAN clustering example using $minPts$ as 4. The circles are of radius ε. Arrows indicate points with density reachability. Core points are colored red, border points are yellow and outliers are blue.



Source: (SCHUBERT *et al.*, 2017).

3.3 Clustering Evaluation Metrics

The Cluster analysis field is also concerned with how to evaluate the quality of the clusters discovered in the data. There are two main kinds of metrics used to evaluate the clustering process: internal and external. Internal metrics assume one does not have access to Y, for it is not yet known. External metrics, on the other hand, show the quality of the generated clusters $\hat{Y}$ in comparison with Y.

3.3.1 External Metrics

Accuracy, one of the most widespread external metrics, is capable of measuring the quality of the automatic grouping $\hat{Y}$ when compared to the experts clustering Y:

$$accuracy(Y, \hat{Y}) = \frac{1}{I}\sum_{i=0}^{I} \mathbf{1}(\hat{y}_i = y_i) \qquad (2)$$

The accuracy value is, by definition, bound by [0,1], where higher values are better. However, accuracy is not well-suited for clustering applications, since the clusters defined in $\hat{Y}$

do not necessarily represent the semantic intended in Y. We also define $a = \sum_{i=1}^{I}\sum_{l=i}^{I} \mathbf{1}(\hat{y}_i = \hat{y}_l \,\&\, y_i = y_l)$, in other words, the number of sample pairs grouped in the same generated clusters, as well as the same golden cluster. Complementarily $b = \sum_{i=1}^{I}\sum_{l=i}^{I} \mathbf{1}(\hat{y}_i \neq \hat{y}_l \,\&\, y_i \neq y_l)$, as the number of sample pairs grouped in different generated clusters and different golden clusters. A measure of agreement between Y and $\hat{Y}$ (external metric) is given by Rand Index (RI):

$$RI(Y, \hat{Y}) = \frac{a + b}{C_2^I} \tag{3}$$

Where I is the number of samples in the dataset and C_2^I is the number of possible pairs in the dataset. Pure RI, however, has the drawback that in case the number of clusters is close to the number of samples, it does not guarantee that random cluster assignments will get lower values. The adjusted Rand index (ARI), a normalized by chance version of RI. ARI is defined by Hubert & Arabie, 1985 as:

$$ARI(Y, \hat{Y}) = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{4}$$

Where $E[RI]$ is the theoretically expected value of RI for the case of a random clustering $\hat{Y}$. ARI is bounded in the range $[-1,1]$, where near-zero values indicate a random clustering, negative values indicate disagreement and higher values show likeness between the ground truth and the generated clusters.

3.3.2 Internal Metrics

Internal metrics are any function that can generate a score of quality of a given cluster set without looking at the answer, e.g., how well separated are the clusters generated. It is, however, a quality measure only from a certain point of view, e.g., some concave cluster shapes are not well-defined by their inter-cluster distance. In what follows we describe the four internal metrics explored in this work.

- Silhouette (ROUSSEEUW, 1987)

One of the most common internal scores to measure clustering quality, it represents the average ratio of intra-cluster and outer-cluster distances. For each sample n_i, (ROUSSEEUW, 1987) defines the silhouette coefficient s_i as:

$$s_i = s(n_i) = \frac{\hat{d}_i - d_i}{\max(d_i, \hat{d}_i)} \quad (7)$$

Where d_i is the Euclidian average distance between the sample n_i and all other samples n_j with $i \neq j$ and $\hat{y}_i = \hat{y}_j$. Considering the group k_p as being the closest cluster on average to the cluster $k_{\hat{y}_i}$, $\hat{d}_i$ is the average distance between n_i and all other samples n_l, such that $i \neq j$ and $\hat{y}_l = p$. We define the clustering silhouette S as the arithmetic average of all s.

- PBM (PAKHIRA; BANDYOPADHYAY; MAULIK, 2004)

Given the cluster centroids G^j, for each cluster k_j, we firstly define D_B as the largest distance between two clusters centroids:

$$D_B = \max_{j<j'} d(G^j, G^{j'}) \quad (8)$$

We also define E_W and E_T as being respectively the sum of distances from each sample to its cluster centroid and the sum of distances of all samples of the dataset to the dataset centroid G^N. The PBM index (an acronym of the name of the authors) is then calculated as:

$$PBM = \left(\frac{1}{J} \times \frac{E_T}{E_W} \times D_B\right) \quad (9)$$

- GDI_{51} (BEZDEK; PAL, 1998)

GDIs, or the Generalized Dunn's Indices. The Dunn index (DUNN, 1974) is calculated as the ratio of the minimal inter-cluster distance d_{min} to the maximum intra-cluster distance :

$$Dunn = \frac{d_{min}}{d_{max}} \quad (10)$$

GIDs, however, use different methods to calculate inter-cluster distances for d_{min} and the intra-cluster distances for d_{max}. The authors propose a total of six different methods of calculating the inter-cluster distances δ and three methods of calculating the intra-cluster distances Δ, amounting to eighteen possible metrics. Those metrics are denoted by GDI_{uv}, where u and v denote the inter and intra-cluster distances used, respectively.

Note that $GDI_{11} = Dunn$, i.e., the first indexes denote the original Dunn index method. The intra-cluster distance $\Delta_1$is simply the Euclidean distance. Follows the equations for the best inter-cluster distance, δ_5:

$$\delta_5 = \frac{1}{\left|k_j + k_{j'}\right|}\left(\sum_{n_i \in k_j} d\left(n_i, G^j\right) + \sum_{n_i \in k_{j'}} d\left(n_i, G^{j'}\right)\right) \tag{11}$$

- SD Scattering (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001)

Given feature-wise variance vector γ^N for a group of samples N:

$$\gamma^N = \sum_{\hat{y} \in \hat{Y}} \{Var(f_{n,0}), Var(f_{n,1}), \dots, Var(f_{n,Q})\} \tag{12}$$

Where $f_{x,y}$ is the set of values of a feature f_x assigned to samples in the group y. The $SD\ Scattering$ is then calculated as:

$$SD_{scat} = \frac{\frac{1}{J}\sum_{j=1}^{J} \|\gamma^j\|}{\|\gamma^N\|} \tag{13}$$

In other words, the SD Scattering index is the ratio of variance inside clusters by the total variance in the dataset. This score decreases when clusters manage to reduce the total variance inside a dataset when compared to the entire dataset (lower values are better).

3.4 Genetic Algorithm

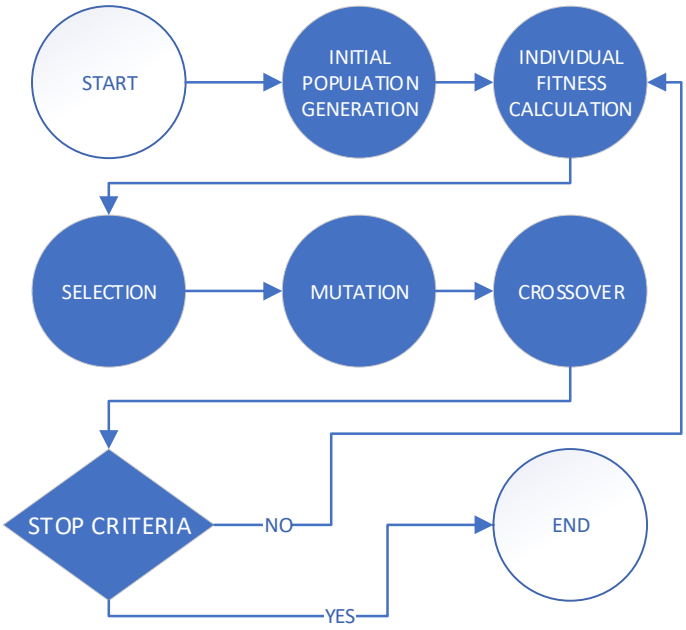
GAs are stochastic search algorithms inspired by the *survival of the fittest* notion from biology. The classical implementation of GA follows the structure shown in Figure 8. The general approach to implement these algorithms in real-case scenarios includes replacing and adapting the steps in Figure 8. We start from the *START* step, where the solution space needs to be encoded in the individual (also referred to as a solution or phenotype) representation $L = [l_0, l_1, \dots, l_N]$ where $l_n \in \{0,1\}$. It is important that the individual representation can be translated to a solution to the problem and consequently a fitness value for this solution. This data representation is essential for the use of most classical GA implementations and is analogous to the gene expression in genetic beings in a population – a gene is either active (1) or inactive (0).

The *INITIAL POPULATION GENERATION* will generate the starting population seed for the experiment. This step will select G individuals from a random uniform distribution to integrate the starting population.

The process of *INDIVIDUAL FITNESS CALCULATION* will necessarily apply the fitness function and assign one or more fitness score(s) for each individual. The chosen fitness function

will direct the algorithm when testing good solutions, i.e., it is the objective function to be optimized.

Figure 8 – General algorithm structure for GA.



Source: Author.

The *SELECTION* of the fittest individuals from the population can be a simple and straightforward selection of the best (highest fit score) half of the population, which will be able to pass their genes for the next generation. Another implementation for the selection step, called tournament selection, allows for the selection of individuals from the below-average half of the population. The tournament selection randomly selects u individuals from the population and the most fit between those survives for the next generation. These tournaments are repeated *G/2* times. This selection strategy increases the chances of overcoming complex search spaces where the worst solutions could surround the best solutions. Other implementations include the stochastic selection of individuals where their relative fitness in the population is set as the probability of being chosen.

The *MUTATION* is a step of GAs where small changes are randomly introduced in the selected population, again, improving the odds of escaping local minimums. Randomly flipping values from the L individual representation with some low probability is a typical implementation.

After that, the CROSSOVER will generate new individuals to integrate the population in the place of the ones discarded in the selection phase. The crossover process occurs between

two or more individuals randomly selected from the gene pool and the generated offspring is usually composed of a pair of individuals, where each one is closer to one parent.

Those operators are applied iteratively and the population is continuously updated until a defined *STOP CRITERIA* is reached. From the final population, the fittest solution is then selected and retrieved to the user.

Some strategies slightly change the structure shown in Figure 8 by altering the combination of mutation and crossover. The mutation step could generate new individuals instead of mutating the population in place and the crossover step skipped. Other implementations execute those operations with mutual exclusion in the population, mutating some while mating other individuals through the crossover.

3.5 Genetic Algorithms for Feature Selection

Metaheuristics are optimization algorithms with a partial search of the target-space. Metaheuristics use strategies of search (heuristics) connected, or not, to the domain of the search-space.

Researches base the feature selection algorithms on the notion that not all features have a real importance in the grouping of instances (e.g., Grouping cars by their type into a *hatchback, sedan, SUV* or *crossover* has nothing to do with their color). Some features can even be prejudicial to the clustering performance, depending on the semantic of wanted clusters. Literature works commonly refer to such features as noise features.

Metaheuristics are commonly used as high-level methods for optimization where an exhaustive search of the set of solutions is not feasible (in our case 2^Q, while Q averages to 738, as seen in Table 1). These methods, in general, leverage parallelism and heuristics in order to find close-to-optimal solutions. One metaheuristic analysis is defined in terms of how explorative the algorithm is, i.e., how good it is in escaping local minimums and how well it adapts to the general search space form of the optimization task.

The usage of GAs for feature selection is normally straightforward. Each individual is coded as a string of features that are either activated (1) or deactivated (0). After encoding the solution space, the next most important step to customize is the fitness function. An implementation of the fitness score on unsupervised tasks includes the clustering of the dataset using only the activated features for each individual. Internal clustering metrics of the generated clusters are used as the fitness score value of the algorithm. The rest of the GA implementation steps,

representing the search strategy of the algorithm is then formulated as any classic GA implementation.

The process of feature selection and calculation of the fitness score is given bellow in Equations 14 through 17:

The matrix representation of an individual in the algorithm can be given by:

$$W = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_n \end{bmatrix} \tag{14}$$

Where $W_x \in \{0,1\}$. The matrix representation of a dataset with n features and I individuals is given by:

$$D = \begin{bmatrix} D_{1,1} & \cdots & D_{1,n} \\ \vdots & \ddots & \vdots \\ D_{I,1} & \cdots & D_{I,n} \end{bmatrix} \tag{15}$$

The dataset D filtered by the features selected by the individual W is given by simple matrix multiplication:

$$D_W = D \times W^T \tag{16}$$

In the end, the fitness function is the value given by an internal metric applied over the clustering result of a given clustering algorithm using as input the filtered dataset D_W:

$$fitness = internal\_metric(clustering(D_W)) \tag{17}$$

In the literature, (Yang and Honavar (1998) developed a supervised method for feature selection using a wrapper Neural Network. This work had used a multi-objective function forming a tradeoff between the complexity of the network needed and the accuracy of the results using the selected features.

Abualigah, Khader and Al-Betar (2016) developed a new GA for unsupervised feature selection in text data. Their implementation – harmony search algorithm to solve the feature selection problem (FSHSTC) – successfully improves the overall clustering quality in tasks of identifying spam emails when compared to standard K-Means.(TABAKHI; MORADI; AKHLAGHIAN, 2014)

4 RELATED WORKS

This section is composed of references related to the current work. They are subdivided into three different subclasses. Section 4.1 presents works related to the process of feature selection using metaheuristics. Section 4.2 introduces literature about the use of ontologies in the processes of Data Mining (DM). Section 4.3 reviews studies on the automated identification of petrofacies.

4.1 Metaheuristics for Unsupervised Feature Selection

The use of stochastic search algorithms (e.g., GA, Swarm Optimization, Simulated Annealing) for unsupervised feature selection in high dimensional datasets has been a subject of research for more than two decades now and is applied to a variety of disciplines to reach a close-to-optimal subspace of features, improving results and performance of algorithms (BEZDEK *et al.*, 1994; BRILL; BROWN; MARTIN, 1992; ESTÉVEZ *et al.*, 2009; GUO; UHRIG, 1992).

The work of Lleti *et al.* (2004) presents a feature subset selection method, where instead of HC, the GA algorithm wraps a K-Means algorithm. Their work defines the number of clusters in an unsupervised manner using either the silhouette score or gap statistic to define the number of clusters.

Liu and Yu (2005), it is proposed a categorization framework of feature selection algorithms, where the present work fits into the categories of Random, Wrapper and Clustering on the axis of the search strategy, evaluation criteria and DM task, respectively. The authors of Xue *et al.* (2016) proposed a hierarchical classification of random feature selection algorithms where the current work will fall to the classes of GA, Wrapper Approach and Single Objective.

Many works present a similar approach to the one described here for feature selection by GA while using an internal metric as the objective function (KIM; STREET; MENCZER, 2000; MORITA *et al.*, 2003; SRIKRISHNA; ESWARA REDDY; SESHA SRINIVAS, 2013). These works use different fitness metrics, while Lin *et al.* (2013) also use the silhouette index as a fitness metric when combining feature selection and clustering to separate genetic material from different types of cancer.

Tabakhi, Moradi and Akhlaghian (2014) developed the unsupervised feature selection method based on ant colony optimization (UFSACO). Their work makes use of a fully connected graph structure where each node corresponds to a feature and the connected feature's

similarity gives their edges weights. They use the ant colony optimization algorithm to transverse this graph in either an exploitation strategy (greedy), which does not consider already visited nodes, or exploration strategy where all features can be revisited with a given probability. At the end of the process, the most visited nodes are selected as the feature subset.

The works mentioned above were studied for the development of the work proposed. The literature review and framework from Liu and Yu (2005) gave us a comprehensive view of the methods available, as well as standard nomenclature used in academic researches. The remaining works are proof of the feasibility of the solution proposed and gave insights on the possible architecture.

4.2 Ontology and Data Mining

Ontology is a logical theory, based on Philosophical principles, that expresses the commitment of a representational language with the intended models that the language vocabulary aims to represent (GUARINO; OBERLE; STAAB, 2009). When materialized through an artifact, an ontology is a formal model that represents the shared knowledge of a domain community. Ontology is currently a rising topic in information science, being largely used for solving problems as diverse as providing interoperability solutions in legacy applications, supporting knowledge-based systems, information retrieving methods, and enhancing data mining techniques like those that we describe in this work.

The ontology engineering process is of high importance when developing solutions for high-specialized fields. It allows for knowledge engineers to explicit shared knowledge with precision not previously seen, consequently improving the quality of developed solutions. In this work, we focus on the Petroledge® geological-driven ontology. However, it is important to notice that this approach has also proven to be effective in multiple high-specialized fields of knowledge.

Many works, such as (ESTÉVEZ *et al.*, 2009; KUO *et al.*, 2007; MITCHELL *et al.*, 2015; TIFFIN *et al.*, 2005), show that datasets described over well-founded domain ontologies can imbue DM explorations with expert knowledge. This find encourages our assumption that DM techniques can extract useful information from this richly described data.

Chen *et al.* (2015) uses an ontology-defined 2D emotional space from Zhang *et al.* (2011), combined with ML and signal analysis techniques to infer emotions directly from

electroencephalograms, generating interesting insights and enabling semantic queries over such datasets.

Other projects, such as NELL, from Mitchell *et al.* (2015), and Knowledge Vault, from Dong *et al.* (2014), try to automatically create and populate ontologies through the application of DM over large amounts of data. They follow the opposite direction of previous works by populating knowledge models from data, instead of using knowledge models to infer information about the data. The semantic quality of results from such works can be questionable since they are based mainly on statistics, which does not always represent truths about the world. Error and outliers can occur for many reasons, including errors on the source data and the expressing of emotions, such as irony, one of the main challenges to DM algorithms.

The well-established ontology of Petroledge was used as an example of our pre-processing methodology, reducing the feature-space while maintaining the semantic significance. To this end, the attributes and structures defined in this ontology were essential for developing a useful summarization of the data.

4.3 Automated *Petrofacies* Identification

Cevolani *et al.* (2013) have a similar approach for identifying *petrofacies* from thin section description data by employing K-Means and HC. The author, however, does not use feature selection as a means of simplifying outputs, relying on visualization techniques such as parallel coordinates and dendrograms to assist the geologist task. In Chapter 5, we compare the results of this work with the achievements of our method.

Our primary contribution in this work is a qualitatively validated approach that applies DM techniques over ontology-based thin section description datasets to identify petrofacies. Our ontology-based pre-processing method significantly reduces the feature space with little to no loss in the task performance. The access to such kind of quality data, with an unusually broad set of qualitative features, allows us to explore different methods of feature subset selection described in the literature.

5 CLUSTERING-BASED RESERVOIR PETROFACIES EXTRACTION

This chapter starts by presenting the research steps and development applied for the task of automatically identifying petrofacies, followed by an in-depth description of the final implementation, showing its features and advantages.

In this chapter, we discuss details that are necessary for applying clustering methods for petrofacies identification. Firstly, in Section 4.1, we describe the details of the datasets that constitute the focus of this work. In Section 4.2, we discuss the criteria that were considered for selecting the compared clustering approaches for petrofacies identification. Finally, in Section 4.3, we present a clustering-based method that was developed in this work for petrofacies identification.

5.1 Datasets Description

All the datasets used in this project are described using Petroledge ontology-based application for reservoir petrographic description. This ontology, described in Abel (2001), implements concepts to categorize mineral, texture, sedimentary structures and diagenetic modification in sedimentary rocks. It provides a total of 652 concepts.

This level of detail allows the expert to precisely describe every thin section point, such as the points shown in Figure 2. Consequently, this level of detail generates highly sparse datasets.

With the support of the Geoscience Institute from UFRGS, we were given access to six different datasets containing a total of 661 described thin-section samples and 75 known *petrofacies*. It is necessary to point out the data heterogeneity, i.e., *petrofacies* can frequently contain as few as one sample, or as much as twenty, with some outlier *petrofacies* reaching sixty-four and sixty-six samples (see Figure 9). Table 1 and present more statistics about the datasets subdivided into the scenarios explored in Section 5.3.1. The environmental and geographical features of those datasets are described as follows:

- **Campos Basin**, located offshore of the Rio de Janeiro and the Espírito Santo States, is a set of deep marine, turbiditic sandstone reservoirs and associated non-reservoir deposits from an oilfield in the Campos Basin, eastern Brazilian margin (MOHRIAK, 2003; WINTER *et al.*, 2007). The Campos Basin turbidite succession was deposited during the Upper Cretaceous period on sea bottom physiography generated by the deposition, tectonism and erosion of Lower Cretaceous salt and carbonate deposits. The reservoir sandstones are very porous, rich in feldspar grains,

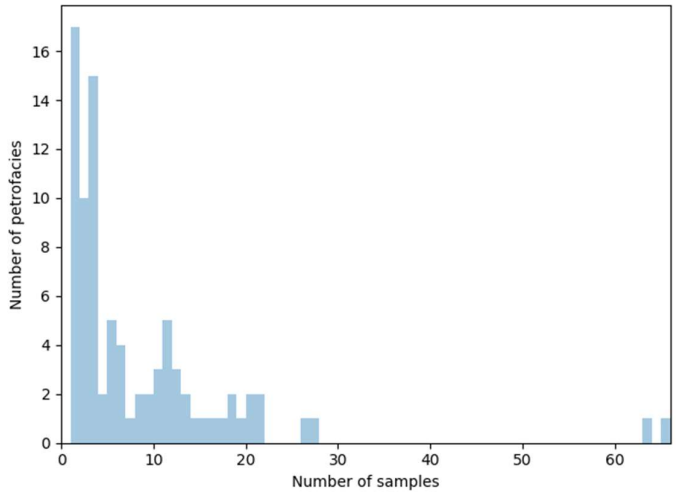
and with mild, limited diagenesis. Low-permeability associated deposits include finer-grained sandstones rich in mud intraclasts and glauconite, and hemipelagic mudrocks;

- **Carmopolis**, aptian of Sergipe-Alagoas Basin. Carmópolis Member sandstones and conglomerates were deposited by alluvial fans and fluvial systems during sea level downstands. The sandstones are rich in metamorphic and sedimentary lithic fragments and some constitute hybrid sandstones. Porosity was significantly preserved due to partial dolomite cementation and shallow oil emplacement. The primary composition does not show great variance, even so, thirteen reservoir petrofacies were defined based on cementation, compaction and porosity (SCHRANK; ALTENHOFEN; DE ROS, 2017);
- **Jequitinhonha**, located at the eastern Brazilian margin, comprehends siliciclastic sandstones, hybrid arenites and calcarenites from fluvio-deltaic, lacustrine and transitional deposits of Aptian-Albian age. The sandstones are dominantly arkoses, subordinately lithic arkoses and feldspathic litharenites, derived from uplifts of basement blocks during a rift context and present variable porosity. Compaction and diagenesis both impacted on reservoir quality. This along with facies diversity resulted in the definition of fourteen reservoir petrofacies (JARDIM; DE ROS; KETZER, 2011);
- **Equatorial Margin** refers to fluvial, estuarine, deltaic and shallow marine sandstones and associated finer-grained deposits from the Brazilian Equatorial Margin (JÚNIOR; COSTA; HASUI, 2008; MOHRIAK, 2003). The Equatorial Margin sandstones show a wide variation of porosity, owing mostly to the amounts of ductile low-grade metamorphic rock fragments, which were deformed during the burial and compaction of the succession;
- **Mucuri** is an aptian of the Espírito Santo Basin, southeast Brazilian margin. Mucuri mudstones, conglomerates and mostly sandstones were deposited by braided fluvial and coastal lacustrine systems. The sandstones are mainly arkose, poorly to well sorted, that generally present loose packing in the more coarse-grained sandstones and tight packing in the finer, coastal micaceous sandstones. Porosities found on the descriptions show significant variance. Primary composition, compaction degree, cementation and replacing of grains by clay minerals, calcite and pyrite, and grain

dissolution were the main aspects considered for the sixteen reservoir petrofacies defined for this dataset (ROCHA, 2018);

- **Western Peru** are tertiary deposits and comprises immature sandstones and hybrid sandstones rich in volcanic, metamorphic and sedimentary lithic fragments, deposited by deltaic and fluvial systems. The original porosity of the sands was intensively modified by diagenesis. Mechanical compaction was important in samples rich in ductile lithic fragments, which were converted into pseudomatrix. Twelve petrofacies were defined.

Figure 9 – Distribution of the number of thin-section samples into petrofacies.



Source: Author.

The datasets used in this work all represent instances of thin rock sections. Each entry in such a dataset describes multiple rock features observed by the specialist in a single thin section during the geological description phase (Figure 10). Experts can describe the features at multiple detail levels. They are hierarchically subdivided under the three compositional types – primary, diagenetic and porosity – which indicates the genesis of each mineral. Primary features describe the foundational rock mineralogy, with relative position and size inside the rock. Diagenetic features show us the chemical and physical changes undergone after the initial rock deposition, as well as relative position and size. Porosity features indicate porous structures observed in the rock, generated by dissolution of minerals, or physical fractures.

As seen in Figure 10, different features, numbers of features and range of values characterize each petrofacie. E.g., MudSand petrofacies (S-03 to S-07) are easily identified for presenting ~45% of its mineral matrix as primary feature *Siliciclastic mud matrix syndepositional – As*

intrabasinal constituent. However, CoarPor (S-32 to S-45) and FinPor (S-46 and S-47) petrofacies show almost identical feature values and are hardly statistically differentiable. Furthermore, some CoarPor samples deviate from the main group (S-32 to S-45) for having high values in feature 39, while low values in features 42 and 43. Experts can interpret such subtle combinations of features, using their domain knowledge, to identify petrofacies.

All six datasets shown in this work are raw data extracted directly from geology research on actual exploration projects. All this data is uniformly described and available from the Petroledge system. Such a consistent description of data collected through decades allows us to research algorithms and clustering techniques validated under multiple scenarios and locations.

Figure 10 – Dataset sample with omitted columns for readability. Row entries represent individual rock samples, while columns represent the different percentages of described features. The petrofacies column shows the specialist assignment for each entry.

Samples	01	02	27	33	39	40	41	42	43	45	46	47	49	Petrofacies	Cluster	Classification
S-01	7,667	0	0	0	0	0	0	1,67	0	0	0	0,33	23	SandMarl	1	Excellent
S-02	13,667	0	0	0	0,33	0,33	0	2	0	0	0	0,33	25,67	SandMarl		
S-03	0	0,33	1,67	0	0,67	5	0	4,67	0	0	0	2,33	47,67	MudSand	2	Excellent
S-04	0	0,33	3	0	1,67	2,67	0	8,33	0,33	0	0	2,33	46,33	MudSand		
S-05	1,333	0,67	2,33	0	2	3	0	7,33	0	0	0	2	43,33	MudSand		
S-07	6	0,33	1	0	1,33	2	0	7	0	0	0	2	41,67	MudSand		
S-09	9,333	0	0	0	0,33	4,67	1	21,67	1	4	5,33	0,33	0	CoarGlauCal	3	Regular
S-10	16	0	0	4	1,67	3	0	22,33	1,67	0,67	24,33	1,33	0	FinIntra		
S-12	13,841	0	0	1,38	1,38	2,42	0	29,76	3,46		11,76	0,69	0	FinIntra		
S-14	11	0	0	0	2,33	3	0	18	1,67	7,33	11,67	0,67	0	FinIntra		
S-16	6,667	0	0	1,33	4,67	4	0	28	0,67	1,33	5,33	0,67	0	FinIntra	4	Regular
S-17	8,333	0	0	2	6,67	6,33	0	25,33	0,67	2,67	6	0,67	0	FinIntra		
S-21	2	0	0	0	0,33	4,67	0	34	6,67	0,67	0	0,33	0	CoarCal	5	Good
S-22	3	0	0	0	0,33	4	1,33	25,33	13,33	0,33	3,67	0,33	0	CoarCal		
S-23	7,667	0	0	0	0,33	3	2,33	34	4,33	0,67	0	0	0	CoarSil		
S-32	35,667	0	0	0	3,33	0,67	0	24,67	8	1,33	0,67	0,33	0	CoarPor	6	Very Good
S-33	31	0	0	0	0,33	2,67	2,33	27,67	4,67	0,33	2,67	0	0	CoarPor		
S-36	35,333	0	0	0	1	4	0,33	24,33	3,67	0,33	0,33	0	0	CoarPor		
S-37	32,667	0	0	0	1,33	5,67	0,67	25	4,33	0,33	0,67	0	0	CoarPor		
S-41	25	0	0	0	0,33	1,33	0,67	29,33	10,67	0	0,67	0,67	0	CoarPor		
S-44	31	0	0	0	1,33	3,33	1,67	21,67	16	0,67	0,33	0	0	CoarPor		
S-45	31,667	0	0	0	1	3,67	2	19,33	15	0,67	0	0	0	CoarPor		
S-46	39	0	0	0	3,33	0,67	0	21,67	2	0	8	0,33	0	FinPor		
S-47	40	0	0	0	3	0,67	2,67	24,33	6	0,33	1	0,33	0	FinPor		
S-50	40,667	0	0	0	14,33	2,33	0,33	0	0	2,67	0,33	0,33	0	CoarPor	7	Very Good
S-51	33,333	0	0	0	12,67	2,33	0,33	0	0	0,67	0	0	0	CoarPor		

LEGEND

01 = Porosity
02 = [diagenetic]Albite - Ingrowth - Grain fracture-filling
27 = [diagenetic]Pyrite - Microcrystalline - Matrix-replacive
33 = [porosity]Shrinkage pore - Framework
39 = [primary]Detrital orthoclase - As monomineralic grain
40 = [primary]Detrital plagioclase - As monomineralic grain - Untwinned
41 = [primary]Detrital plagioclase - In plutonic rock fragment - Twinned
42 = [primary]Detrital quartz monocrystalline - As monomineralic grain
43= [primary]Detrital quartz monocrystalline - In plutonic rock fragment
45 = [primary]Glauconite peloid - As intrabasinal constituent
46 = [primary]Mud intraclast - As intrabasinal constituent
47 = [primary]Muscovite - As monomineralic grain
49 = [primary]Siliciclastic mud matrix syndepositional - As intrabasinal constituent

Source: Author.

5.2 Characterization of the Clustering Task

During this research, we experimented with multiple classic representative algorithms. This section will analyze the desired characteristics for different steps of the petrofacies grouping process. Jain (2010) lists nine main challenges formulated as questions that any researcher should answer during the development and application of clustering algorithms. Some of them

were defined and answered from the start of the research, while others were answered during development with the support of experiments and data. Namely, these nine questions and their objective answers, considering this work's context, are:

1. What is a cluster? A reservoir petrofacies.
2. Should the data be normalized? No. The data is already normalized in the percentage of sampling points, i.e., in the range $[0,1]$.
3. Does the data have any clustering tendency? Yes, results show that naïve algorithms can easily generate clusters resembling some petrofacies.
4. Does the data contain any outliers? Yes, the datasets show singleton petrofacies. These outliers, however, are relevant and cannot be filtered from the dataset.
5. Which clustering method should be used? Experimental results in Section 6.1 indicate different algorithms for different datasets. Some algorithms present relevant results setting a few parameters with feasible execution time. However, experts should individually experiment and analyze the results to select the clustering methods which make the most sense from a geological point of view.
6. How do we define the pair-wise similarity? Section 6.2 suggests that there is not much difference in results using either Euclidean or Manhattan distances. Other distance metrics have shown negligible results during experimentation.
7. What features should be used? It is not trivial to determine the relevant features for each dataset. However, we are able to infer from the results in Chapter 6 that our ontology-based feature pre-processing is able to reduce significantly the feature number while maintaining the relevant information for the petrofacies separation. A wrapper feature selection algorithm based on GA was developed and applied to all scenarios.
8. How many clusters are present in the data? Datasets vary depending on the number of samples and the nature of the locality being explored. It should be an empirically-chosen parameter.
9. Are the discovered clusters and partition valid? The clusters generated by the algorithms are preliminary results. Experts should always analyze clusters and features selected with domain-knowledge in order to validate and tune the results.

The following sections intend to answer the remaining questions proposed, as well as give an outline of the research development of this work.

5.3 A Clustering-based Method for Automated Petrofacies Identification

In this work, we explore how petrophysical datasets behave when clustering techniques are applied to identify petrofacies. In Chapter 6, we compare the generated petrofacies with the ones identified by experts.

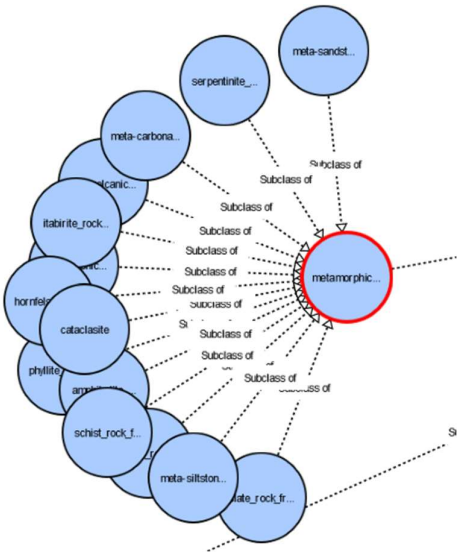
Through the behavior analysis of experts, we noticed that most of the main decisions taken when grouping petrofacies take into account the knowledge structure described by the domain ontology. To test this hypothesis and improve the algorithm results, we developed an ontology-driven feature pre-processing described in Section 5.3.1. The datasets are interpreted in two different scenarios: in the first scenario, the input features are '*raw*,' i.e., numerical values are directly extracted from the dataset as they were described and organized by the petrographers that produce the descriptions, without distinction of feature relevance or indication of feature agglomeration. In the second *Compositional-Locational (C-L)* scenario, the datasets are pre-processed taking into account the ontology and grouping suggested by experts.

5.3.1 Ontology-driven Feature Pre-processing

With the help of experts, we developed a new pre-processing grouping based on the domain ontology, where, in general, petrophysical features seen as relevant to the grouping of petrofacies kept their identities. While less relevant features, used in the definition of few petrofacies, were grouped and summed hierarchically until the reach of a hierarchical ontology seen as relevant. Using as input the raw features, as described by experts in the Petroledge system, with expert support, we derive two main types of features. Namely Compositional and Locational features.

Compositional features, which form the compositional scenarios, represent the sum of described features composed by the same mineral. For primary and diagenetic features, this is accomplished by taking the constituent and searching for its parent mineral in the ontology. For not having a mineral composition, porous features are grouped by the *other* tag. See Figure 11 for an example using the ontology.

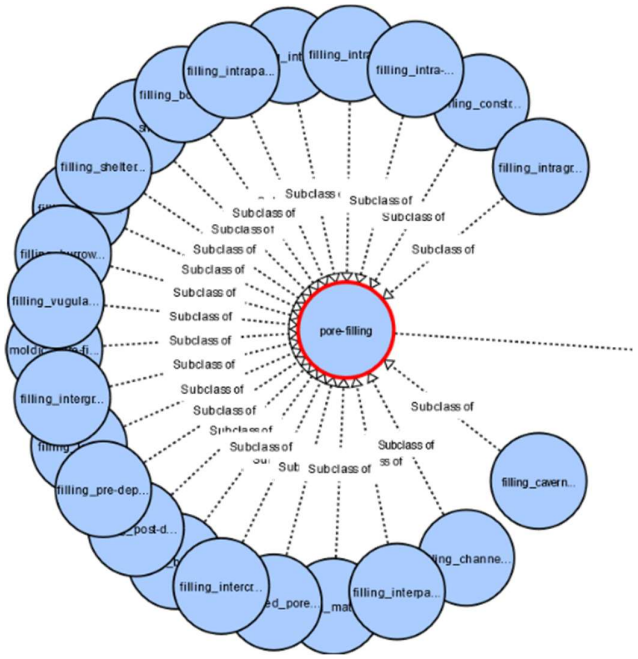
Figure 11 – *Metamorphic rock fragment* (highlighted in red) is a direct subclass of *primary grain constituent*. All of its subclasses are summed when generating compositional features.



Source: Author.

Locational features follow the location described by the expert in the ontology. Due to the age of some datasets ranging up to more than a decade, while the Petroledge system and ontology were still being updated, a few of the locations described in the dataset are not found in the current ontology. Those locations are considered their own group. Aside from those, the algorithm will look in the ontology for the parent concept. If the parent concept is either *primary*, *diagenetic* or *porosity location* (considered too broad), the location described turns into a locational feature; otherwise, the parent concept in the ontology is chosen as the locational feature. Figure 13 shows an example using the domain ontology.

Figure 12 – *Pore-filling* (highlighted in red) is a direct subclass of *diagenetic location*. All of its subclasses are summed when generating the locational features for diagenetic features.



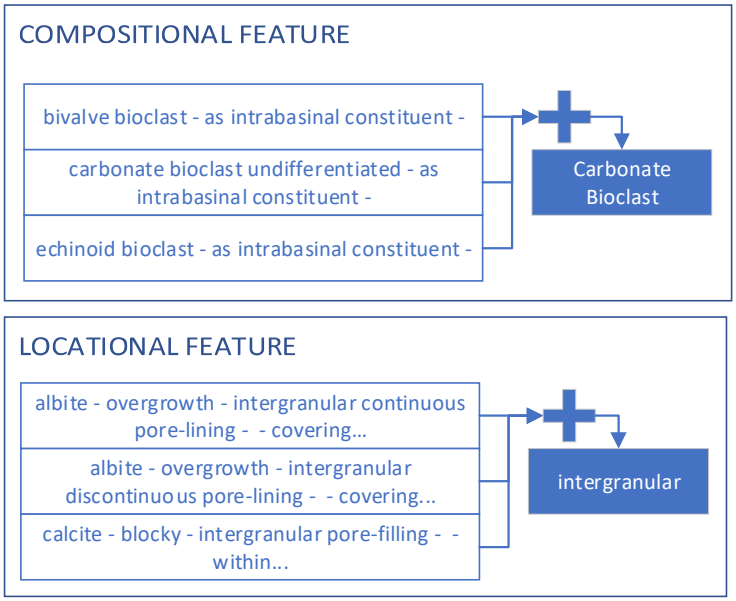
Source: Author.

Figure 13 shows a simplified example of the formation of two real pre-processed features. For the compositional feature, the ontology topology tells us that the primary constituent terms *bivalve bioclast*, *carbonate bioclast undifferentiated* and *echinoid bioclast* all are *carbonate bioclast*, so those features are grouped and summed for the compositional scenario. As for the locational feature, the ontology tells us that the location terms *intergranular continuous pore-lining*, *intergranular discontinuous pore-lining* and *intergranular pore-filling* are all subconcepts of *intergranular*.

The composition and location of features are considered essential attributes by experts. Experiments with different datasets (Chapter 6) also indicate that in this pre-processed format, fewer features (see Table 1) are needed to represent the clusters of petrofacies. Using only the compositional and locational features (C-L scenario) we see a considerable reduction of the sparsity of the data (see), further indicating the ontology-driven pre-processing ability of accurately summarizing the data. The algorithm uses the pre-processed features to reach results similar or even better than using the raw dataset with a significant execution time reduction across every experiment. The experiments shown in Chapter 6 divide the datasets into scenarios. Namely, the *Raw* scenario uses the raw features, such as exported directly from the

Petroledge dataset, the Compositional-Locational (C-L) scenarios are composed by the concatenation of the compositional and locational features.

Figure 13 – Example diagram showing the generation of one compositional and one locational feature: *Carbonate Bioclast* and *Intergranular*.



Source: Author.

Table 1 – Datasets descriptions. The raw features refer to the features extracted directly from the Petroledge system. The compositional and locational features are generated from the raw features using the methodology described in this section. Notice that the "*TOTAL*" row sums unique features presented across every dataset.

Dataset	Samples	Petrofacies	Raw Features	Compositional Features	Locational Features	Compositional + Locational Features
Campos Basin	53	10	457	29	15	44
Carmopolis	120	17	889	28	16	44
Jequitinhonha	66	14	312	22	10	32
Margem Equatorial	143	20	990	30	18	48
Mucuri	264	14	1329	32	19	51
Western Peru	61	12	448	28	14	42
TOTAL	**707**	**87**	**3526**	**36**	**22**	**61**

Source: Author.

Table 2 – Datasets sparsity of data through the scenarios, i.e., the percentage of zeroes in the dataset.

Dataset	Raw Sparsity	Compositional Sparsity	Locational Sparsity	Compositional + Locational Features
Campos Basin	91.47%	56.15%	50.59%	54.17%
Carmopolis	93.66%	54.70%	52.94%	54.04%
Jequitinhonha	90.61%	44.07%	35.55%	42.21%
Margem Equatorial	94.52%	53.41%	55.26%	54.13%
Mucuri	95.96%	61.88%	58.23%	60.49%
Western Peru	82.84%	36.59%	32.16%	35.07%
Average	**91.51%**	**51.13%**	**47.46%**	**41.22%**

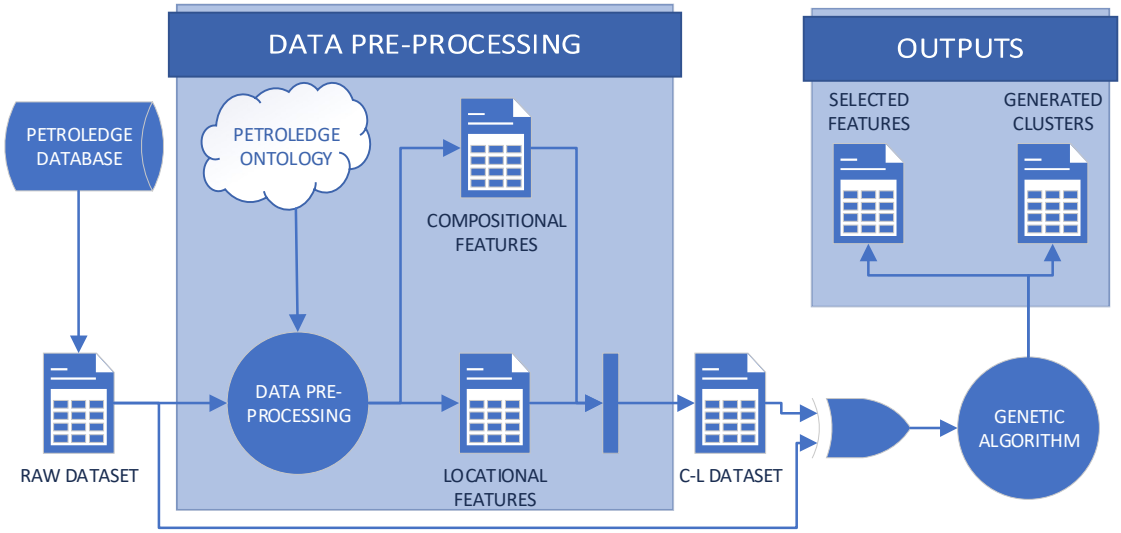
Source: Author.

It is important to notice that this pre-processing step can be generalized on new domains, with similarly structured datasets based on ontologies. In a general view, the "summarizing" pre-processing algorithm would need as input the ontology used on the dataset description, a set of collapsing rules, such as the "too broad location" seen above, and the proper dataset to be summarized. With a user-friendly interface and description language of these collapsing rules, an expert could dynamically execute experiments in order to tailor the pre-processing for each dataset environment. In this work, however, we apply the same collapsing rules for all the datasets. A clustering approach for petrofacies identification

In order to compare the quality of the new feature scenarios with the original feature set, we tested multiple algorithms, from K-Means to an algorithm that adopts GA for feature selection. We opted for a general and adaptable approach to enable an exploratory analysis of the datasets. Inspired by the expert behavior (see Section 5.2), this algorithm selects the features that best separate the petrofacies clusters.

The metaheuristic considers the search-space as the entire 2^n space containing every possible feature as either on or off. The algorithm starts by randomly activating or deactivating each of the n features for the initial population of size k. The selected clustering algorithm is executed k times, once for every individual in the population. An internal fitness metric is calculated for each of the k clustering results.

Figure 14 – Data flow of the implemented algorithm. Notice the exclusive or is intended to denote that either the raw dataset or the C-L dataset is used as input of the GA.



Source: Author.

The fitness metric indicates to the metaheuristic which of the generated solutions are the best to concentrate the search-space around. The algorithm will iteratively execute by selecting the most fit of the population and making small changes in the features selected to explore the search-space. Multiple internal fitness metrics were tested in Section 6.3. The algorithm consistently reaches the best results when using any of the four internal clustering metrics found, i.e., Dunn, GDI_{51}, SD Scattering and Silhouette.

Our algorithm uses the *VarOrCross* (referred to as *control parameters* p_c and p_m by Srinivas & Patnaik, 1994) algorithm to generate the descendant population. This algorithm will generate each individual for the new population by either the variation of a single individual of the parent generation or by crossing two individuals selected randomly from the parent generation. Our implementation will generate 80% of the individuals with the variation from a single parent and the remaining 20% by crossing parents.

Our mutation function is set to flip each of the chromosomes of the parent individual with a 5% chance. The uniform crossover function will randomly sample two individuals from the parent generation. The first of those parents will uniformly exchange chromosomes with the second with a 10% probability, producing a single offspring for the next generation that is 90% identical to the first parent and 10% to the second parent.

Our fitness function accepts an individual as a selection of features. It filters those features accordingly, runs the selected clustering algorithm over the dataset and uses the clusters

generated as input to one of the internal metrics (see Section 6.3 for a comparison of metrics). The output value of the internal metric is assigned to that individual as its fitness value.

As every individual, from all generations, are evaluated, the algorithm will always propagate the best individual forward to avoid losing this individual by a chance of selection or mutation. This process is called elitism. After all generations, the algorithm chooses the fittest individual as a result. It outputs the generated clusters of samples, as well as the features selected at the end of the process.

Figure 14 shows the high-level architecture developed. It represents the data-flow since the data export from the Petroledge system to the generated output. The entire software implementation is available at our GitHub repository[5].

[5] https://github.com/BDI-UFRGS/Lucas_Master

6 EXPERIMENTS

This chapter will describe the most relevant experiments realized and analysis of the results. All of the experiments were executed on all the six datasets and with the RAW and *C*-L scenarios. The scenario *RAW* indicates that the dataset suffered no pre-processing, coming straight from the Petroledge software. The scenario *C-L* indicates that the dataset used is the ontology-based feature reengineering process.

Due to the combinational nature of the algorithm parameters, we will show here the experiments that led to the best performance gains experimentally and enabled the algorithm to handle the input datasets.

All the experiments shown here were executed using an Intel® Core™ i5-5300U CPU operating at 2.30GHz and 12GB memory operating at 1600MHz. A reasonably standard configuration found on many notebooks of the market.

In Section 6.1, we compare the performance of the three selected clustering algorithms in the datasets available in different scenarios to select the best clustering algorithm used in the next sections. Section 6.2 compares different implementations of linkage and distance functions for the AHC algorithm. Section 6.3 compares different metrics as fitness functions for the GA. We present and discuss the timing-related results in Section 6.4. Finally, we discuss the overall most noticeable results presented in this Chapter in Section 6.5.

6.1 Clustering Algorithms Comparison

The first experiment, with its results shown in Table 3, does not use metaheuristics, it merely shows the ARI score reached by each clustering algorithm from the three described in Section 3.3.2. In order to be able to calculate the ARI, all clustering algorithms had their arguments set to match the number of petrofacies identified by the expert when grouping the petrofacies. The DBSCAN algorithm was not introduced in this comparison, as its' parameters were considered non-adaptable for new datasets in a predictable manner.

Table 3 - ARI comparison between different clustering methods over multiple datasets and scenarios. The average score of 10 runs for K-Means.

	Affinity Propagation	AHC	K-Means
C-L			
Campos Basin	0.55	0.44	0.56±0.01
Carmopolis	0.21	0.27	0.27±0.02
Jequitinhonha	0.45	0.40	0.41±0.04
Equatorial Margin	0.43	0.45	0.43±0.03
Mucuri	0.35	0.50	0.46±0.08
Western Peru	0.19	0.16	0.20±0.03
Avg.	**0.36**	**0.37**	**0.39**
RAW			
Campos Basin	0.38	0.56	0.51±0.07
Carmopolis	0.19	0.19	0.24±0.03
Jequitinhonha	0.29	0.31	0.28±0.04
Equatorial Margin	0.43	0.48	0.46±0.02
Mucuri	0.26	0.37	0.32±0.02
Western Peru	0.18	0.15	0.27±0.05
Avg.	**0.29**	**0.34**	**0.35**

Source: Author.

It follows the parameters set for each algorithm during their execution. These parameters were derived from tradeoffs between the best empirical results and acceptable execution times on the available hardware.

- **K-Means:**
  - K = Number of petrofacies in the dataset
  - $Number\ of\ starts$ = 100
- **AHC:**
  - K = Number of petrofacies in the dataset
  - d = Euclidian
  - $linkage$ = Ward
- **Affinity Propagation:**
  - $affinity$ = Individually tuned to get the number of petrofacies in the dataset with values between -400 and -7700

The results in Table 3 indicate that all three algorithms present similar results. However, the AHC algorithm is orders of magnitude faster than the K-Means algorithm with multiple initializations and the Affinity Propagation. All algorithms got similar quality results when working with the C-L scenarios, indicating the relevance of the Compositional and Locational features.

Since the Genetic metaheuristic will have to run hundreds of time-steps for hundreds of individuals in their populations, clustering time performance becomes relevant. The Affinity Propagation, K-Means and AHC present similar performances in terms of clustering quality for this task. In this sense, due to its ease of parametrization and determinism, AHC is a well-suited choice. The AHC is chosen as the clustering algorithm to generate the fitness metric from here onwards with no noticeable performance penalty.

6.2 Linkage and Distance Comparisons

As seen in Section 3.2, two critical parameters of AHC are its distance and linkage functions. The affinity function should receive two observations and return a similarity value of the two. In contrast, the linkage function will use the affinity function to determine which next clusters to merge. For the fitness function, we will use the silhouette coefficient for its simplicity to analyze the results. For the GA parameters, we set the population to 50 and 1000 generations, or 200 generations without fitness improvements. For the affinity function, we experiment with the common Euclidean and Manhattan distance functions. We also experiment with three linkage functions: Complete – the highest affinity value between two samples of two clusters; Single – the smallest affinity value between two samples of two clusters; and Ward, which choose to merge clusters that will minimize the global variance.

Table 4 shows the average ARIs obtained for every dataset in every scenario using the GA algorithm for feature selection and AHC with different affinities and linkages functions for the clustering of the datasets. We can see that the Ward linkage outperforms every other combination of affinity and linkage. For this reason, the next experiments will use the Ward linkage when using the AHC.

Table 4 – Average ARI of 10 runs for each of the combinations of affinity and linkage.

	Euclidean Affinity		Avg	Manhattan Affinity		Avg	Ward Linkage Euclidean
	Complete Linkage	Single Linkage		Complete Linkage	Single Linkage		
C-L							
Campos Basin	0.57±0.00	0.74±0.09	0.65	0.74±0.04	0.75±0.06	0.75	0.58±0.03
Carmopolis	0.27±0.02	0.02±0.00	0.15	0.27±0.03	0.03±0.06	0.15	0.26±0.01
Jequitinhonha	0.41±0.01	0.29±0.04	0.35	0.45±0.02	0.24±0.07	0.35	0.44±0.00
Equatorial Margin	0.45±0.06	0.37±0.04	0.41	0.47±0.03	0.38±0.02	0.42	0.49±0.02
Mucuri	0.12±0.00	0.11±0.02	0.12	0.13±0.03	0.11±0.01	0.12	0.46±0.09
Western Peru	0.21±0.03	0.12±0.02	0.17	0.21±0.03	0.13±0.02	0.17	0.21±0.02
Avg.	**0.34**	**0.28**	**0.31**	**0.38**	**0.27**	**0.33**	**0.41**
Raw							
Campos Basin	0.57±0.00	0.55±0.04	0.56	0.57±0.00	0.57±0.00	0.57	0.57±0.00
Carmopolis	0.23±0.02	0.02±0.00	0.13	0.22±0.02	0.03±0.02	0.13	0.26±0.02
Jequitinhonha	0.26±0.03	0.23±0.00	0.25	0.26±0.02	0.26±0.05	0.26	0.32±0.01
Equatorial Margin	0.49±0.04	0.51±0.03	0.50	0.52±0.01	0.50±0.02	0.51	0.52±0.01
Mucuri	0.07±0.00	0.06±0.01	0.07	0.08±0.01	0.06±0.00	0.07	0.10±0.00
Western Peru	0.22±0.03	0.18±0.02	0.20	0.25±0.05	0.21±0.02	0.23	0.26±0.04
Avg.	**0.31**	**0.26**	**0.29**	**0.32**	**0.27**	**0.29**	**0.34**

Source: Author.

6.3 Fitness Function Comparision

This section tries to determine the best available internal metric for usage as the fitness function of the feature selection GA described in Figure 8 and Figure 14. More than 30 internal metrics, drawn from the Scikit-Learn (PEDREGOSA *et al.*, 2011) and ClusterCrit (DESGRAUPES, 2013), were tested. The top four internal clustering metrics (described in Section 3.3.2) show their average ARI values presented in Table 5. We ran the experiments ten times using the Ward AHC, for 200 generations with a population of 100 individuals. For the GA parameters, again, we set the population to 50 and 1000 generations, or 200 generations without fitness improvements.

Table 5 – Average ARI of 10 runs of the GA algorithm when using the top-four internal clustering metrics as fitness functions.

	Dunn	GDI_{51}	SD Scattering	Silhouette
C-L				
Campos Basin	0.61±0.02	0.50±0.08	0.44±0.12	0.55±0
Carmopolis	0.19±0.17	0.17±0.02	0.16±0.05	0.23±0.05
Jequitinhonha	0.34±0.04	0.32±0.00	0.44±0.10	0.46±0.14
Equatorial Margin	0.19±0.20	0.38±0.05	0.39±0.01	0.36±0.05
Mucuri	0.11±0.19	0.33±0.08	0.32±0.00	0.36±0.02
Western Peru	0.16±0.08	0.17±0.03	0.20±0.10	0.14±0.00
Avg.	**0.28±0.22**	**0.31±0.13**	**0.33±0.13**	**0.35±0.15**
RAW				
Campos Basin	0.54±0.00	0.38±0.03	0.54±0.03	0.47±0.03
Carmopolis	0.18±0.06	0.22±0.00	0.17±0.01	0.15±0.05
Jequitinhonha	0.25±0.09	0.18±0.03	0.24±0.00	0.19±0.13
Equatorial Margin	0.49±0.02	0.41±0.02	0.46±0.01	0.40±0.01
Mucuri	0.17±0.13	0.33±0.03	0.34±0.04	0.17±0.05
Western Peru	0.10±0.07	0.17±0.04	0.18±0.04	0.16±0.02
Avg.	**0.31±0.19**	**0.28±0.10**	**0.32±0.14**	**0.26±0.14**

Source: Author.

The results above show us that none of those fitness metrics generate dominant ARI values. Through the interpretation of ARI, values above zero indicate concordance between the generated clustering and the ground truth. However, values below 0.5 can be considered as indicating weak correlation and thus, the algorithm alone is not able to generate the final petrofacies.

6.4 Timing Performance Comparison

With the usage of GA and the execution time turning into a relevant factor, Table 6 compares the execution times of the algorithms over the dataset's scenarios. We see an average of 5.47 times speedup of the execution time, while the previous results show that there is little to no performance loss when comparing the *C-L* and *Raw* scenarios. The timing of pre-processing steps is of less than a second for every dataset, so we believe it is not impactful in the overall timing results. The parametrization is the same as in Section 6.3.

Table 6 – Average execution time (in seconds) over 10 runs of the algorithm for each of the combinations of affinity and linkage.

	Euclidean Affinity		Avg	Manhattan Affinity		Avg	Ward Linkage
Row Labels	Complete Linkage	Single Linkage		Complete Linkage	Single Linkage		
C-L							
Campos Basin	21±6.0	20±4.9	21	18±2.6	22±5.6	20	19±5.9
Carmopolis	34±11.5	22±2.9	28	38±14.0	31±8.7	35	28±4.9
Jequitinhonha	21±3.9	19±5.2	20	20±3.6	23±5.1	21	21±5.2
Equatorial Margin	33±6.4	37±19.1	35	32±8.8	39±12.7	35	30±7.6
Mucuri	71±29.5	55±13.7	63	69±23.7	72±33.8	70	87±43.4
Western Peru	25±7.9	21±5.8	23	23±9.6	26±8.4	24	22±5.6
Avg.	**34**	**29**	**32**	**33**	**35**	**34**	**34**
RAW							
Campos Basin	37±9.0	42±14.3	39	42±6.6	37±6.9	39	35±6.4
Carmopolis	226±102.3	119±26.3	175	202±83.3	239±113.1	221	230±79.6
Jequitinhonha	48±14.3	37±6.9	43	70±24.7	56±15.2	63	42±11.9
Equatorial Margin	216±71.3	250±61.5	232	189±63.7	206±70.5	197	180±61.7
Mucuri	430±84.5	600±203.0	515	679±233.4	519±89.0	599	516±86.6
Western Peru	42±14.8	35±11.6	39	38±10.7	50±17.7	44	40±12.1
Avg.	**162**	**181**	**171**	**203**	**185**	**194**	**174**

Source: Author.

Even though the execution times shown here (in the order of seconds) seem not to make much difference when comparing to the manual identification of petrofacies (in the order of weeks), it is essential to notice that this execution would very likely be made many times by an expert experimenting with new parameters, mainly the number of clusters.

6.5 Discussion

We argue that it is essential to verify the generalization capability of any clustering algorithm. Our results show that the method developed behaves consistently well through multiple scenarios and datasets from the basins of different regions of South-America with fundamentally different reservoir environments and structures.

Through this chapter, we presented multiple scores and timing comparisons using different algorithms. When analyzed in conjunction, the broad trend is that the ontology-based pre-processing of the datasets presents no significant harmful effect on the quality of the results. This noticeable harmlessness, combined with the speedup in processing, indicates the ontology-based pre-processing as a promising method to improve processing times on large datasets

while keeping the data semantically significant. It is imperative, however, that an expert is involved in the process of defining the subtotals of interest for this task to avoid the loss of semantic meaning of the results obtained through this process.

7 CONCLUSION

We describe an implementation that directly uses a domain ontology to summarize datasets from thousands of features into a few dozen. Our experiments show that there is little to no loss in the task of automated petrofacies identification using the summarized data.

We explored clustering techniques applied to the petrofacies identification, identifying promising algorithms, developed a GA-based feature selection algorithm using the acquired knowledge and proposed a feature reengineering guided by the domain ontology that greatly reduces the dataset dimensionality and processing time.

In this work, the experiments suggest good promises in the automation of the petrofacies grouping task. To the best of our knowledge, no other work in the literature has presented such a complete and embracing view of the automation of the petrofacies grouping task. We were able to analyze six different datasets along the South-America and multiple rock formations.

The results indicate that there are challenging aspects for the application of clustering and feature selection in this domain, pointing to the need for new future research in this field. Geological datasets show much potential when analyzed correctly, making use of experts' knowledge, be it in the form of consultants or ontologies, to facilitate the understanding and analysis.

7.1 Contributions

Thanks to the availability and structure of the available datasets, our work was successful in providing a tool to reduce the complexity of the petrofacies grouping task for experts. Our pre-processing can significantly reduce the feature space explored by experts when working on new datasets. Our work provides evidence that the summarization of datasets with the use of a domain ontology can significantly reduce the processing time, increasing the speed of the iteration between collecting and interpreting data, while maintaining the clustering quality. This pre-processing step could be adapted on new domains with similar dataset structures, using three inputs: (1) the ontology describing the dataset, (2) a set of feature collapsing rules, and (3) the dataset which will be summarized.

We developed the experiments shown with the close support of domain experts to make well-founded decisions from the inception of this work to the experiment's significance. We can infer that the pre-processing, together with the feature selection algorithm, can significantly increase the iteration time of expert geology experts when proposing theories for the reservoir

structures. In the domain of petroleum exploration, the speed of this iteration can significantly reduce costs to exploration activities by avoiding the excavation of non-profitable wells.

7.2 Future Work

Other algorithms could prove best for this task and could enable the software automation to provide a more refined selection of features as output. We are interested in using subspace clustering techniques to provide more accurate clustering results to the end-user. Some promising works on subspace clustering are shown in (VIDAL, 2011).

We also find it necessary for future research on the GA parameters and how they affect the algorithm outcome. Our experiments indicate that the fixed population size and number of generations applied is enough to achieve convergence for the datasets at hand. However, we believe these parameters can be tailored for each new dataset, using more or fewer resources according to the number of features and samples.

A custom user interface could provide experts with the tools needed to compose petrofacies. The user would be able to tweak the results given from the algorithm in real-time and compose views to analyze the results in different feature-spaces, including the C-L space, offering an overview of the results and Raw space, offering a more detailed analysis of the grouped petrofacies.

The application of multi-objective optimization techniques, already widely implemented for GAs (DEB *et al.*, 2002), is a logical future step, once multiple objectives are identified and verified to be implicitly used during the petrofacies identification.

Big-data applications and datasets usually demand many hours of execution on data-center clusters to process all the data. Each new feature can potentially generate gigabytes of new data. Some business cases, such as search engines, need to retrain their algorithms on updated data constantly. If datasets such as these could be summarized for focused applications similarly, it could mean significant savings on computing power for those companies.

Future works on the domain will be followed by geology experts making use of the algorithm developed in order to help the petrofacies selection on new datasets. Interactive experiments will validate the usability of the developed software, as well as provide useful new insights on the developed processes.

REFERENCES

ABEL, M. **Estudo da perícia em petrografia sedimentar e sua importância para a engenharia de conhecimento**. 2001. *[s. l.]*, 2001. Disponível em: https://lume.ufrgs.br/handle/10183/2588. Acesso em: 7 maio. 2019.

ABUALIGAH, L. M.; KHADER, A. T.; AL-BETAR, M. A. Unsupervised feature selection technique based on genetic algorithm for improving the text clustering. *In*: 2016, **2016 7th international conference on computer science and information technology (CSIT)**. *[S. l.: s. n.]* p. 1–6.

ARTHUR, D.; VASSILVITSKII, S. k-means++: The advantages of careful seeding. *In*: 2007, **Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms**. *[S. l.: s. n.]* p. 1027–1035.

BEZDEK, J. C. *et al.* Genetic algorithm guided clustering. *In*: 1994, **Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence**. : IEEE, 1994. p. 34–39. Disponível em: https://doi.org/10.1109/ICEC.1994.350046. Acesso em: 15 out. 2018.

BEZDEK, J. C.; PAL, N. R. Some new indexes of cluster validity. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, *[S. l.]*, v. 28, n. 3, p. 301–315, 1998. Disponível em: https://doi.org/10.1109/3477.678624

BRILL, F. Z.; BROWN, D. E.; MARTIN, W. N. Fast generic selection of features for neural network classifiers. **IEEE Transactions on Neural Networks**, *[S. l.]*, v. 3, n. 2, p. 324–328, 1992.

CEVOLANI, J. T. *et al.* Computational methodology to study heterogeneities in petroleum reservoirs. *In*: 2013, **EAGE Annual Conference & Exhibition incorporating SPE Europec**. *[S. l.: s. n.]*

CHEN, J. *et al.* Electroencephalogram-based emotion assessment system using ontology and data mining techniques. **Applied Soft Computing**, *[S. l.]*, v. 30, p. 663–674, 2015. Disponível em: https://doi.org/10.1016/J.ASOC.2015.01.007. Acesso em: 16 out. 2018.

DE ROS, L. F.; ABEL, M.; GOLDBERG, K. **Advanced acquisition and management of petrographic information from reservoir rocks using the Petroledge System**. *[S. l.]*: American Association of Petroleum Geologists, 2007.

DE ROS, L. F.; GOLDBERG, K. Reservoir Petrofacies: A Tool for Quality Characterization and Prediction. **AAPG Annual Conference and Exhibition**, Long Beach, 2007. Disponível em: http://www.searchanddiscovery.com/documents/2007/07117deros/images/deros.pdf. Acesso em: 10 out. 2018.

DEB, K. *et al.* A fast and elitist multiobjective genetic algorithm: NSGA-II. **IEEE transactions on evolutionary computation**, *[S. l.]*, v. 6, n. 2, p. 182–197, 2002.

DESGRAUPES, B. clusterCrit: clustering indices. **R package version**, *[S. l.]*, v. 1, n. 3, p. 4–5, 2013.

DONG, X. *et al.* Knowledge vault. *In*: 2014, New York, New York, USA. **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14**. New York, New York, USA: ACM Press, 2014. p. 601–610. Disponível em: https://doi.org/10.1145/2623330.2623623. Acesso em: 3 out. 2018.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. **Journal of cybernetics**, *[S. l.]*, v. 4, n. 1, p. 95–104, 1974.

ESTÉVEZ, P. A. *et al.* Normalized mutual information feature selection. **IEEE Transactions on Neural Networks**, *[S. l.]*, v. 20, n. 2, p. 189–201, 2009.

FREY, B. J.; DUECK, D. Clustering by passing messages between data points. **science**, *[S. l.]*, v. 315, n. 5814, p. 972–976, 2007 a.

FREY, B. J.; DUECK, D. Clustering by Passing Messages Between Data Points. **Science**, *[S. l.]*, v. 315, n. 5814, p. 972–976, 2007 b. Disponível em: https://doi.org/10.1126/science.1136800. Acesso em: 9 jul. 2019.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? *In*: **Handbook on ontologies**. *[S. l.]*: Springer, 2009. p. 1–17. *E-book*.

GUO, Z.; UHRIG, R. E. Using genetic algorithms to select inputs for neural networks. *In*: 1992, **Combinations of Genetic Algorithms and Neural Networks, 1992., COGANN-92. International Workshop on**. *[S. l.: s. n.]* p. 223–234.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of intelligent information systems**, *[S. l.]*, v. 17, n. 2–3, p. 107–145, 2001.

HÄRDLE, W.; SIMAR, L. **Applied multivariate statistical analysis**. *[S. l.]*: Springer, 2007. v. 22007*E-book*.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, *[S. l.]*, v. 2, n. 1, p. 193–218, 1985.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, *[S. l.]*, v. 31, n. 8, p. 651–666, 2010. Disponível em: https://doi.org/10.1016/j.patrec.2009.09.011. Acesso em: 23 jun. 2019.

JARDIM, C. M.; DE ROS, L. F.; KETZER, J. M. Reservoir quality assessment and petrofacies of the Lower Cretaceous siliciclastic, carbonate and hybrid arenites from the Jequitinhonha Basin, Eastern Brazil. **Journal of Petroleum Geology**, *[S. l.]*, v. 34, n. 3, p. 305–335, 2011. Disponível em: https://doi.org/10.1111/j.1747-5457.2011.00507.x

JÚNIOR, A. V. S.; COSTA, J. B. S.; HASUI, Y. Evolução da margem atlântica equatorial do Brasil: Três fases distensivas. **Geociências (São Paulo)**, *[S. l.]*, v. 27, n. 4, p. 427–437, 2008.

KIM, Y.; STREET, W. N.; MENCZER, F. Feature selection in unsupervised learning via evolutionary search. *In*: 2000, New York, New York, USA. **Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00**. New York, New York, USA: ACM Press, 2000. p. 365–369. Disponível em: https://doi.org/10.1145/347090.347169. Acesso em: 10 out. 2018.

KUO, Y. *et al.* Domain ontology driven data mining: a medical case study. *In*: 2007, **Proceedings of the 2007 international workshop on Domain driven data mining**. *[S. l.: s. n.]* p. 11–17.

LEE, K. *et al.* DDoS attack detection method using cluster analysis. **Expert systems with applications**, *[S. l.]*, v. 34, n. 3, p. 1659–1665, 2008.

LIN, T.-C. *et al.* Classifying subtypes of acute lymphoblastic leukemia using silhouette statistics and genetic algorithms. **Gene**, *[S. l.]*, v. 518, n. 1, p. 159–163, 2013.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on knowledge and data engineering**, *[S. l.]*, v. 17, n. 4, p. 491–502, 2005.

LLETI, R. *et al.* Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. **Analytica Chimica Acta**, *[S. l.]*, v. 515, n. 1, p. 87–100, 2004.

MITCHELL, M. **An introduction to genetic algorithms**. *[S. l.]*: MIT Press, 1998. *E-book*. Disponível em: https://books.google.com.br/books?hl=en&lr=&id=0eznlz0TF-IC&oi=fnd&pg=PP9&dq=An+introduction+to+genetic+algorithms&ots=shlJ31YaPc&sig=J

YmC-LDC94DmtmE_P8PBo4MJJ3w#v=onepage&q&f=false. Acesso em: 10 out. 2018.

MITCHELL, T. *et al.* Never-Ending Learning. **Acl**, *[S. l.]*, v. 2, n. July, p. 1–4, 2015. Disponível em: https://doi.org/10.1007/s13398-014-0173-7.2. Acesso em: 2 out. 2018.

MOHRIAK, W. U. Bacias sedimentares da margem continental Brasileira. **Geologia, tectônica e recursos minerais do Brasil**, *[S. l.]*, v. 2003, p. 87–165, 2003.

MORITA, M. *et al.* Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition. *In*: 2003, **Seventh International Conference on Document Analysis and Recognition**. : IEEE Comput. Soc, 2003. p. 666–670. Disponível em: https://doi.org/10.1109/ICDAR.2003.1227746. Acesso em: 10 out. 2018.

PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, U. Validity index for crisp and fuzzy clusters. **Pattern recognition**, *[S. l.]*, v. 37, n. 3, p. 487–501, 2004.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of machine learning research**, *[S. l.]*, v. 12, n. Oct, p. 2825–2830, 2011.

ROCHA, E. C. da. **Definição de Petrofácies com Base Nas Principais Características Petrográficas que Controlam a Qualidade das Rochas Sedimentaries Como Reservatório de Hidrocarbonetos**. 2018. *[s. l.]*, 2018.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, *[S. l.]*, v. 20, p. 53–65, 1987. Disponível em: https://doi.org/10.1016/0377-0427(87)90125-7. Acesso em: 25 jul. 2017.

SCHRANK, A. B. S.; ALTENHOFEN, S. D.; DE ROS, L. F. Diagenetic Preservation and Modification of Porosity in Aptian Lithic Reservoirs from the Sergipe--alagoas Basin, Ne Brazil. **Journal of Sedimentary Research**, *[S. l.]*, v. 87, n. 11, p. 1156–1175, 2017.

SCHUBERT, E. *et al.* DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. **ACM Transactions on Database Systems (TODS)**, *[S. l.]*, v. 42, n. 3, p. 1–21, 2017.

SORURI, M.; SADRI, J.; ZAHIRI, S. H. Gene clustering with hidden Markov model optimized by PSO algorithm. **Pattern Analysis and Applications**, *[S. l.]*, v. 21, n. 4, p. 1121–1126, 2018.

SRIKRISHNA, A.; ESWARA REDDY, B.; SESHA SRINIVAS, V. **Automatic Feature Subset Selection using Genetic Algorithm for ClusteringInt. J. on Recent Trends in Engineering and Technology**. *[S. l.: s. n.]*. Disponível em:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.428.6828&rep=rep1&type=pdf. Acesso em: 10 out. 2018.

SRINIVAS, M.; PATNAIK, L. M. Genetic algorithms: A survey. **computer**, *[S. l.]*, v. 27, n. 6, p. 17–26, 1994.

TABAKHI, S.; MORADI, P.; AKHLAGHIAN, F. An unsupervised feature selection algorithm based on ant colony optimization. **Engineering Applications of Artificial Intelligence**, *[S. l.]*, v. 32, p. 112–123, 2014.

TIFFIN, N. *et al.* Integration of text-and data-mining using ontologies successfully selects disease gene candidates. **Nucleic acids research**, *[S. l.]*, v. 33, n. 5, p. 1544–1552, 2005.

VIDAL, R. Subspace clustering. **IEEE Signal Processing Magazine**, *[S. l.]*, v. 28, n. 2, p. 52–68, 2011.

WINTER, W. R. *et al.* Bacia de campos. **Boletim de Geociencias da PETROBRAS**, *[S. l.]*, v. 15, n. 2, p. 511–529, 2007.

XUE, B. *et al.* A Survey on Evolutionary Computation Approaches to Feature Selection. **IEEE Transactions on Evolutionary Computation**, *[S. l.]*, v. 20, n. 4, p. 606–626, 2016. Disponível em: https://doi.org/10.1109/TEVC.2015.2504420. Acesso em: 15 out. 2018.

YANG, J.; HONAVAR, V. Feature Subset Selection Using a Genetic Algorithm. **IEEE Intelligent Systems**, *[S. l.]*, v. 13, p. 44–49, 1998.

ZHANG, X. *et al.* Emotiono: An Ontology with Rule-Based Reasoning for Emotion Recognition. *In*: *[S. l.]*: Springer, Berlin, Heidelberg, 2011. p. 89–98. *E-book*. Disponível em: https://doi.org/10.1007/978-3-642-24958-7_11. Acesso em: 16 out. 2018.

APPENDIX <RESUMO EXPANDIDO>

Petrofacies de reservatório é definida como padrões de feições petrográficas que se repetem dentro de um mesmo reservatório de petróleo. As *petrofacies* possibilitam um alto nível de abstração sobre os milhares de dados e atributos acumulados durante a fase de extração e preparação dos dados. Apesar da comprovada utilidade deste conceito, o processo de identificação de petrofacies é, atualmente, um processo manual e lento, levando de semanas a meses.

Neste trabalho, nós estudamos um método de automação da identificação de petrofacies através de dados petrográficos extraídos de lâminas delgadas. Como petrofacies são definidas individualmente para reservatórios distintos, a tarefa de identificação é considerada não-supervisionada. Algoritmos de agrupamentos de amostras (*clustering*), largamente utilizados nesta categoria de problemas, combinados com técnicas de otimização para seleção de atrobutos constituem um dos principais focos de estudo neste trabalho. A seleção de atributos é também utilizada neste trabalho para manter a interpretabilidade sobre quais atributos mais influenciam no agrupamento das petrofacies.

Petroledge®, o software utilizado na descrição dos dados, emprega uma ontologia do domínio da geologia sedimentar para a coleta dos dados. Durante o acompanhamento do processo de identificação de petrofacies por especialistas, nós determinamos que um nível mais alto de abstração dos dados das lâminas delgadas pode ser promissor, enquanto utiliza menos poder de processamento.

Por representar ótimas caracteristicas de adaptibilidade para novos domínios e modularidade de componentes, Algoritmos Genéticos (AG) são utilizados aqui para a exploração do domínio de atributos de 2^Q possibilidades, dados Q diferentes atributos.

Tratando-se de um trabalho exploratório, nós ativemos os experimentos à algoritmos clássicos como representativos de diferentes políticas de agrupamento disponíveis na literatura.

Nós utilizamos seis conjuntos de dados (*datasets*) de seis bacias sedimentares distribuídas ao longo da costa marítima brasileira e peruana. Todos estes datasets foram descritos e suas petrofacies identificadas por especialistas. O nosso pré-processamento baseado em ontologias mantém a mesma semântica de descrição dos dados utilizada pelos especialistas, enquanto reduz significativamente o número de atributos (de ~ 800 para ~ 40), bem como a esparcidade (de 91% para 41% em média).

Todos os experimentos realizados utilizam dois cenários: *Crú*, ou seja, utilizando os atributos originais, assim como extraídos do banco de dados relacional do Petroledge® e *Composição-Localização* (*C-L*), utilizando os atributos abstraídos com o pré-processamento baseado em ontologias.

Primeiramente, realizamos a seleção do algoritmo de clustering que melhor se adapta aos dados. Na comparação dos algoritmos de *Affinity Propagation*, *Clustering Hierárquico Aglomerativo* (em inglês *AHC*) e *K-Means*, nós decidimos por continuar com o AHC, por apresentar os melhores resultados, sem incertezas e com o melhor desempenho com os datasets utilizados.

A seguir, selecionamos os melhores parâmetros de distanciamento e *linkage* aplicados em conjunto com o algoritmo de AG. Os resultados indicam os parâmetros de afinidade euclideana e *Ward-linkage* como os mais generalizáveis entre os datasets e cenários estudados.

Em seguida, após a comparação de dezenas de funções de aptidão do AG, os experimentos nos permitem determinar o coeficiente de silhueta como a métrica mais adequada para os dados disponíveis e promissora para novos datasets.

Finalmente, nós realizamos uma comparação temporal dos experimentos. É claramente visível uma aceleração média de 5.47 vezes com o cenário *C-L*, em comparação com o cenário *Crú.* Outra importante conclusão que temos ao analizar os resultados dos experimentos realizados é de que o pré-processamento baseado em ontologias apresenta benefícios, já que não afeta a qualidade dos resultados, enquanto reduz acelera o processamento.

Como trabalho futuro, o campo mais impactante em que esta pesquisa pode ser evoluída é na interface com usuários especialistas, realizando entrevistas formais e avaliação qualitativas da interface, bem como os resultados obtidos. No campo da computação, é necessário uma análise robusta de como o AG comporta-se com diferentes parâmetros de configuração. Novas e recentes técnicas de processamento não-supervisionado continuam surgindo. Campos de pesquisa como *subspace clustering* permitem, por exemplo, simultaneamente gerar os agrupamentos de dados e realizar a seleção de atributos.