ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Επιλογή Χαρακτηριστικών για Ταξινόμηση με τη βοήθεια Μέτρων Πληροφορίας

(Feature Selection for Classification using Measures of Information Theory)

ΜΗΤΡΟΠΟΥΛΟΥ ΑΙΚΑΤΕΡΙΝΗ

Επιβλέπων: Κουκουβίνος Χρήστος
Καθηγητής Ε.Μ.Π.

Αθήνα, 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Επιλογή Χαρακτηριστικών για Ταξινόμηση με τη βοήθεια Μέτρων Πληροφορίας

ΜΗΤΡΟΠΟΥΛΟΥ ΑΙΚΑΤΕΡΙΝΗ

Αθήνα, 2016

Στον πατέρα μου, που με έκανε να αγαπήσω τα Μαθηματικά.

Σε ευχαριστώ.

Πρόλογος

Η Θεωρία Πληροφοριών αποτελεί τον κλάδο των Μαθηματικών που περιγράφει τρόπους ποσοτικοποίησης, διαχείρισης και απεικόνισης της αβεβαιότητας της πληροφορίας. Στην καθαρή της μορφή αποτελεί ανακάλυψη των μηχανικών, με τις πλέον περίβλεπτες πρακτικές επιτυχίες της να αναφέρονται στην εκπομπή των εικόνων της έγχρωμης τηλεόρασης, στη σχεδίαση των συστημάτων ραντάρ έγκαιρης προειδοποίησης και στην ανάκτηση ανέπαφων μηνυμάτων από μακρινά διαστημόπλοια. Παρουσιάστηκε στον κόσμο με τη μορφή δύο εργασιών του Claude Shannon της Bell Telephone Laboratories, που δημοσιεύτηκαν στο Bell System Technical Journal τον Ιούλιο και τον Οκτώβριο του 1948. Στην ουσία οι εργασίες αποτελούνται από μία σειρά από θεωρήματα που έχουν να κάνουν με το πρόβλημα της αποστολής μηνυμάτων από μια θέση σε μια άλλη, γρήγορα, οικονομικά και αποτελεσματικά. Ωστόσο, η πιο συναρπαστική συνέπεια της μελέτης του Shannon έγκειται στο γεγονός ότι στάθηκε ικανή να καταστήσει την έννοια της πληροφορίας τόσο λογική και σαφή, που να μπορεί να τοποθετηθεί μέσα σε ένα τυπικό πλαίσιο ιδεών. Χειριζόμενος την πληροφορία με όρους σαφώς καθορισμένους αλλά και τελείως αφηρημένους, ο Shannon μπόρεσε να τη γενικεύσει, εδραιώνοντας νόμους που ισχύουν όχι μόνο για λίγους τύπους πληροφοριών αλλά για όλα τα είδη και παντού, και μπορούν να χρησιμοποιηθούν για την έρευνα οποιουδήποτε συστήματος που εκπέμπει ένα μήνυμα από μια θέση σε μια άλλη.

Από την θεμελίωση των βάσεων από τον Claude Shannon το 1948 μέχρι σήμερα, έχει επιτευχθεί η διείσδυσή της Θεωρίας Πληροφοριών σε όλους σχεδόν τους τομείς της επιστήμης και της τεχνολογίας. Η Θεωρία Πληροφοριών διαδραματίζει σημαντικό ρόλο στην διαμόρφωση των θεωριών μετάδοσης, αποθήκευσης και συμπίεσης δεδομένων με σημαντικές συνεισφορές –ενδεικτικά-στα πεδία της φυσικής, της στατιστικής, της επιστήμης των υπολογιστών, της κυβερνητικής, της βιολογίας και των οικονομικών.

Περιεχόμενα

Περιεχόμενα Σχημάτων

Περιεχόμενα Πινάκων

Περίληψη

Αντικείμενο της παρούσας Διπλωματικής Εργασίας αποτελεί η ανάδειξη της χρησιμότητας της Θεωρίας Πληροφοριών στο αντικείμενο της επιλογής χαρακτηριστικών. Αυτό επιτυγχάνεται με την παρουσίαση ορισμένων μεθόδων επιλογής χαρακτηριστικών που εφαρμόζονται σε σύνολα δεδομένων, προκειμένου να γίνει η ορθή ταξινόμησή τους σε κατηγορίες και την αξιολόγησή τους. Μια σύνοψη του περιεχομένου των Κεφαλαίων που περιλαμβάνονται στην εργασία, έχει ως εξής:

Στο Κεφάλαιο 1 γίνεται μια σύντομη εισαγωγή στη Θεωρία Πληροφοριών καθώς και μια αναλυτική ιστορική αναδρομή των κυριότερων σταθμών από τις απαρχές της μέχρι σήμερα. Επίσης, παρουσιάζεται η βασική έννοια της πληροφορίας, παρατίθενται ορισμένες από τις βασικότερες εφαρμογές της σε διάφορα επιστημονικά πεδία και γίνεται αναφορά στους τρόπους μεταφοράς πληροφοριών μέσω των δικτύων μετάδοσης.

Στο Κεφάλαιο 2 διατυπώνονται και αναλύονται οι βασικότερες έννοιες που αφορούν την ανάπτυξη της Θεωρίας Πληροφοριών, με έμφαση στα μέτρα πληροφορίας. Παράλληλα, γίνεται αναφορά στο απαραίτητο μαθηματικό υπόβαθρο κάθε μέτρου, ώστε να γίνει κατανοητή τόσο η χρησιμότητά τους όσο και αλληλοσυσχέτιση που εμφανίζουν.

Στο Κεφάλαιο 3 γίνεται λεπτομερής ανάλυση της διαδικασίας επιλογής χαρακτηριστικών και περιγραφή των επιμέρους σταδίων της. Ακόμη, παρουσιάζονται οι γνωστότερες κατηγορίες μεθόδων επιλογής χαρακτηριστικών, οι οποίες είναι οι μέθοδοι φίλτρων (filter), οι μέθοδοι περιτυλίγματος (wrapper), και οι ενσωματωμένες μέθοδοι (embedded).

Το Κεφάλαιο 4 αποτελεί το κυριότερο κομμάτι της Διπλωματικής Εργασίας, καθώς σε αυτό γίνεται εκτενής αναφορά και ανάλυση των εξής αλγορίθμων επιλογής χαρακτηριστικών: της υβριδικής μεθόδου Φίλτρου/Περιτυλίγματος, της μεθόδου mMIFS-U και των μεθόδων NMIFS, GAMIFS. Η δομή παρουσίασης των παραπάνω μεθόδων περιλαμβάνει την ανάλυση των μαθηματικών σχέσεων που απαιτούνται για την

κατανόησή τους, την περιγραφή της αλγοριθμικής διαδικασίας της καθεμιάς, καθώς και την παράθεση των πειραματικών δεδομένων από τη εφαρμογή τους σε διάφορα σύνολα δεδομένων, σε συνδυασμό με την εξαγωγή συμπερασμάτων που προέκυψαν από τη σύγκριση τους.

Abstract

The subject of this Diploma Dissertation is the emergence of the usefulness of Information Theory regarding feature selection. This is accomplished by presenting some feature selection methods applied to data sets, in order to classify them correctly into categories and evaluate them. A summary of the contents of each Chapter contained in this Diploma Dissertation is as follows:

Chapter 1 provides a brief introduction to Information Theory, as well as a detailed historical overview of the main milestones from its beginnings to this day. Also, the basic concept of information is presented, along with some of its key applications in various disciplines and references to the ways of transferring information via transmission networks.

In Chapter 2 the basic concepts related to the development of Information Theory are formulated and analyzed, with emphasis on information measures. Furthermore, reference is made to the necessary mathematical background of every measure, in order to understand both their usefulness and correlation between them.

Chapter 3 gives a detailed analysis of the feature selection process and a description of its individual stages. Moreover, the most popular categories of feature selection methods are presented, i.e. the filter methods, the wrapper methods, and the embedded methods.

Chapter 4 is the main part of this Diploma Dissertation, as it comprises a comprehensive report and analysis of the following feature selection algorithms: the Hybrid Filter/Wrapper method, the mMIFS-U method and the NMIFS and GAMIFS methods. The presentation structure of each method includes the analysis of all the mathematical relations necessary for their understanding, the description of each one's algorithmic process, as well as the exposition of the corresponding experimental data produced by the application of each method to various data sets, followed by conclusions that resulted from their comparison.

Ευχαριστίες

Θα ήθελα πρωτίστως να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Χρήστο Κουκουβίνο για την εξαίρετη συνεργασία που είχαμε καθ'όλη τη διάρκεια εκπόνησης της παρούσας Διπλωματικής Εργασίας. Επιπλέον, θα ήθελα να ευχαριστήσω την Διδακτορική φοιτήτρια Κρυσταλλένια Δρόσου για την πολύτιμη συμβολή της στην περάτωση της εργασίας, καθώς υπήρξε πάντα διαθέσιμη και ιδιαίτερα συνεργάσιμη.

Οι ευχαριστίες επεκτείνονται σε δύο άτομα: στον φίλο και καθηγητή μου κ. Αναστάσιο Μαυραγάνη, ο οποίος ήταν δίπλα μου καθ'όλη την διάρκεια των σπουδών μου, προσφέροντάς μου αμέριστη στήριξη και βοήθεια, καθώς και στον πολύ στενό μου φίλο Δημήτρη Παπαδόπουλο που ήταν στο πλευρό μου αδιάκοπα και ακούραστα όλα αυτά τα χρόνια.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, την μητέρα μου που φρόντισε να μου παρέχει τα απαραίτητα ώστε να ολοκληρώσω τις σπουδές μου και τον πατέρα μου, ο οποίος δεν είναι εν ζωή, ωστόσο αποτέλεσε και θα αποτελεί για πάντα το πρότυπό μου καθώς φρόντισε να μου μεταδώσει την αγάπη του για το αντικείμενο των Μαθηματικών και με ενέπνευσε να ακολουθήσω αυτό το μονοπάτι. Μου μετέδιδε πάντα απλόχερα και με περίσσεια αγάπη τις γνώσεις του, εμπνέοντας μου ένα αίσθημα σιγουριάς και αυτοπεποίθησης που με ακολουθεί μέχρι και σήμερα.

Μητροπούλου Αικατερίνη

Εθνικό Μετσόβιο Πολυτεχνείο.
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Αθήνα, 2016

Κεφάλαιο 1:

Εισαγωγή

1.1 Γενικά

Η προσπάθεια του ανθρώπου να επιτεύξει την αποτελεσματική και ταυτόχρονα ταχεία μετάδοση των πληροφοριών ξεκίνησε ήδη από τα πρώτα στάδια της ύπαρξής του και συγκεκριμένα από την εποχή που δημιουργήθηκαν οι πρωταρχικές κοινωνικές ομάδες (νομάδες, φυλές, οικισμοί, κράτη). Τα κωδικοποιημένα ηχητικά σήματα των πρωτόγονων λαών, τα οπτικά σήματα, όπως οι γνωστές φρυκτωρίες, (τα οποία ήταν σήματα από βουνό σε βουνό με την βοήθεια των οποίων μεταδόθηκε η καταστροφή της Τροίας και η επιστροφή του Αγαμέμνονα), καθώς και ο καπνός αποτελούν χαρακτηριστικά παραδείγματα αυτής της προσπάθειας του ανθρώπινου είδους.

Από αυτήν την πρώτη προσπάθεια κάνοντας ένα μεγάλο άλμα, χιλιετιών, στη σήμερον ημέρα, σημειώθηκε ραγδαία ανάπτυξη των μέσων καταγραφής, αποθήκευσης, επεξεργασίας και μετάδοσης δεδομένων, καθώς επίσης και των συναφών τεχνολογιών επικοινωνίας, με κομβικότερα σημεία την εφεύρεση του τηλεφώνου, της τηλεόρασης και των δικτύων υπολογιστών, η οποία αποτελεί χαρακτηριστικό κάθε σύγχρονης κοινωνίας. Κοινός τόπος αυτών των τεχνολογικών επιτευγμάτων είναι η ανάγκη για ακριβή, ταχεία, ασφαλή και οικονομική αποθήκευση και μετάδοση της πληροφορίας. Η ανάπτυξη μεθόδων για την ικανοποίηση των παραπάνω αναγκών οδήγησε αναπόφευκτα στην μαθηματική θεμελίωση της έννοιας της πληροφορίας και στη δημιουργία της επιστήμης που ονομάζεται “Θεωρία Πληροφορίας”.

Η Θεωρία Πληροφορίας είναι το τμήμα των εφαρμοσμένων μαθηματικών που ασχολείται με την έννοια της πληροφορίας, την ποσοτικοποίησή της, τα μέτρα και τις

εφαρμογές της. Πιο συγκεκριμένα, στα θέματα που απασχολούν τη Θεωρία Πληροφορίας συγκαταλέγονται η ποσοτικοποίηση της αβεβαιότητας στην πρόβλεψή της πληροφορίας (ή εντροπία) και οι μονάδες μέτρησης αυτής, η ροή της πληροφορίας, η κωδικοποίησή της, η συλλογή και η ανάκτησή της καθώς και η κατασκευή συστημάτων επεξεργασίας και μετάδοσής της. Τα όρια της Θεωρίας Πληροφοριών είναι αρκετά ασαφή. Η θεωρία αυτή επικαλύπτει κατά ένα μεγάλο μέρος τη Θεωρία Επικοινωνίας, όμως είναι περισσότερο προσανατολισμένη στην επεξεργασία και τη μετάδοση των πληροφοριών και λιγότερο στις λεπτομερείς λειτουργίες των συσκευών που χρησιμοποιούνται στα δίκτυα επικοινωνίας.

Η έννοια της πληροφορίας αποτελούσε αφηρημένη έννοια πριν από τα μέσα του εικοστού αιώνα. Επομένως, οποιαδήποτε προσπάθεια εξαγωγής νόμων που διέπουν την πληροφορία ήταν αρχικά αδύνατη. Ωστόσο,μετά τη μαθηματική θεμελίωσή της, η θεωρία αυτή έχει διευρυνθεί σε μεγάλο βαθμό, ώστε σήμερα να βρίσκει εφαρμογές σε πολλούς άλλους τομείς, όπως στην Επαγωγική Στατιστική, στην Επιλογή Χαρακτηριστικών, στην επεξεργασία της φυσικής γλώσσας, σε δίκτυα εκτός των δικτύων επικοινωνίας, στη Νευροβιολογία, στην εξέλιξη και τη λειτουργία μοριακών κωδικών στην Οικολογία, στη Θερμική Φυσική, στους κβαντικούς υπολογιστές, στην ανίχνευση λογοκλοπής καθώς και σε άλλες μορφές ανάλυσης δεδομένων. Πρωταρχικό ρόλο στην θεμελίωσή της διαδραμάτισε η εφαρμογή της στην Θεωρία Κωδικοποίησης, η οποία έδωσε και το αρχικό έναυσμα για την ανάπτυξή της. Είναι λοιπόν εμφανές ότι η Θεωρία Πληροφοριών είναι ένας ευρύς κλάδος, με εξίσου ευρείες και “βαθιές” εφαρμογές, όπως προαναφέρθηκε.

1.2 Ιστορική Αναδρομή

Πρώτος ο Hartley το 1928 προσπάθησε να προσδιορίσει την έννοια “ποσότητα πληροφορίας”. Υποστήριξε πως η “πληροφορία” προκύπτει από την διαδοχική επιλογή συμβόλων ή λέξεων από ένα δοσμένο “αλφάβητο” ή λεξιλόγιο, προκειμένου να οικοδομηθεί ένα μήνυμα (κείμενο) με κάποιο νόημα (τάξη, λογική). Ας υποθέσουμε ότι σχηματίζουμε λέξεις ή μηνύματα αποτελούμενα από ν σύμβολα από ένα αλφάβητο Ν

συμβόλων. Τότε μπορούμε να επιλέξουμε N^v διαφορετικές λέξεις. Συγκεκριμένα ο Hartley όρισε την ποσότητα πληροφορίας (ή πληροφορικό περιεχόμενο) ως το δεκαδικό λογάριθμο του πλήθους των διαφορετικών λέξεων που μπορούν να σχηματιστούν, αποτελούμενες από ένα δεδομένο πλήθος συμβόλων. Στην περίπτωση μηνυμάτων μήκους κ συμβόλων από ένα αλφάβητο με N σύμβολα, η ποσότητα πληροφορίας είναι ίση με

$$H(N^k) = \log(N^k) = k \cdot \log(N) \quad . \qquad 1.1$$

Για μηνύματα μήκους 1 συμβόλου, από το ανωτέρω αλφάβητο, η ποσότητα πληροφορίας είναι

$$H(N^1) = \log(N) \quad . \qquad 1.2$$

Οι ανωτέρω σχέσεις ανταποκρίνονται στη διαίσθησή μας ότι η ποσότητα πληροφορίας ενός μηνύματος αποτελούμενου από κ σύμβολα θα πρέπει να είναι κ φορές μεγαλύτερη από αυτή ενός μηνύματος που αποτελείται από 1 σύμβολο. Αυτός είναι, άλλωστε, ο λόγος που επιλέχθηκε η λογαριθμική συνάρτηση στον ορισμό της ποσότητας πληροφορίας, αφού πληροί τη σχέση

$$f(x^y) = y\, f(x) \quad . \qquad 1.3$$

Με βάση του λογαρίθμου το 10, η μονάδα της ποσότητας πληροφορίας είναι η decit (decimal unit) ή Hartley. Αν χρησιμοποιήσουμε φυσικό λογάριθμο, η μονάδα είναι το nat (natural unit). Εξετάζοντας ως παράδειγμα το σχηματισμό μηνυμάτων μήκους ενός συμβόλου από ένα αλφάβητο αποτελούμενο από 10 σύμβολα, η ποσότητα πληροφορίας κάθε μηνύματος είναι ίση με

$$H(N^1) = \log_{10}(10) = 1\,\text{decit} \quad . \qquad 1.4$$

Με βάση του λογαρίθμου το 2, η μονάδα της ποσότητας πληροφορίας καλείται bit (binary unit). Αν τώρα εξετάσουμε ως παράδειγμα το σχηματισμό μηνυμάτων μήκους ενός συμβόλου από ένα αλφάβητο αποτελούμενο από δύο σύμβολα, τότε η ποσότητα πληροφορίας είναι

$$H(N^1) = \log_2(N) = \log_2(2) = 1\,\text{bit} \quad . \qquad 1.5$$

Επιπλέον ο λογάριθμος, εξασφάλιζε ότι το ποσό της πληροφορίας θα αυξάνεται καθώς

θα αυξάνεται ο αριθμός των συμβόλων N, κάτι που διαισθητικά έμοιαζε λογικό. Παρόλα αυτά, η εξίσωση δεν ήταν σωστή. Το λάθος του Hartley, στην προσέγγιση που έκανε, ήταν το γεγονός, ότι επηρεάστηκε από το νόμο που, σχεδόν ταυτόχρονα, είχαν διατυπώσει ο Nyquist στις Ηνωμένες Πολιτείες της Αμερικής και ο Kupfmuller στη Γερμανία, το 1924. Σύμφωνα με αυτό το νόμο, η μετάδοση σημάτων τηλεγράφου σ' ένα δεδομένο ρυθμό απαιτεί ένα καθορισμένο εύρος συχνοτήτων. Ο Hartley στον ορισμό του δε διακρίνει διαφορετικές πιθανότητες για τα σύμβολα που απαρτίζουν το αλφάβητο, θεωρεί την επιλογή καθενός εξ αυτών κατά το σχηματισμό ενός μηνύματος ως ίσης πιθανότητας γεγονός.

Αντίθετα, ο Shannon (1916-2001) εισήγαγε την έννοια της πιθανότητας στον ορισμό της ποσότητας πληροφορίας και έθεσε τις βάσεις της σύγχρονης Θεωρίας Πληροφορίας. Η επιλογή κάθε συμβόλου συνδέεται με κάποια, στη γενική περίπτωση, διαφορετική πιθανότητα. Έτσι, ο ορισμός του Hartley είναι μια ειδική περίπτωση του ορισμού του Shannon για την ποσότητα πληροφορίας.

Το 1936, ο πρόεδρος του τμήματος Μηχανολογίας στο φημισμένο πανεπιστήμιο του MIT, στο οποίο ο Shannon έκανε το μεταπτυχιακό του, τον όρισε υπεύθυνο για τη λειτουργία μιας δύσχρηστης υπολογιστικής συσκευής, που είχε κατασκευάσει ο ίδιος και είχε ονομάσει "διαφορικό αναλυτή", και ο οποίος ήταν ένας υπολογιστής αποτελούμενος από μηχανικά μέρη και χρησίμευε για την επίλυση σύνθετων εξισώσεων. Προκειμένου να βελτιώσει τη συσκευή αυτή, ο Shannon, άρχισε να σκέφτεται διάφορους τρόπους αντικατάστασης των δύσχρηστων μηχανικών μερών της με ηλεκτρικά κυκλώματα. Βασισμένος στη Θεωρία του Boole, σύμφωνα με την οποία, όλα τα προβλήματα είναι δυνατόν να λυθούν με τη χρήση 2 μόλις συμβόλων, του 0 και του 1, προσπάθησε να εφαρμόσει αυτή τη προσέγγιση στα ηλεκτρικά διακοπτόμενα κυκλώματα, συμβολίζοντας με 1 το διακόπτη ο οποίος ενεργοποιείται και με 0 το διακόπτη ο οποίος ήταν ανενεργός. Υποστήριξε επίσης ότι οι διακόπτες αυτοί θα μπορούσαν να συνδέονται με τρόπο που να τους επιτρέπει να εκτελούν και πιο πολύπλοκες πράξεις, προτείνοντας πέρα από τις απλές δηλώσεις "ναι" και "όχι", τη χρήση του "και", του "ή" και του "δεν". Τα συμπεράσματα της έρευνας του τα παρουσίασε μέσα από τη διατριβή του με τίτλο "A Symbolic Analysis of Relay and Switching Circuits", η οποία δημοσιεύτηκε το 1937 και θεωρείται σαν μία από τις κορυφαίες του 20ου αιώνα, αφού σε αυτή διατύπωσε επίσης

την άποψη ότι η διπλή έλικα του DNA, δεν είναι τίποτα άλλο παρά ένα πληροφοριακό σύστημα.

Γενικά ο Shannon πίστευε ότι η πληροφορία δεν διέφερε από οποιοδήποτε άλλο μέγεθος και κατά συνέπεια ήταν δυνατός ο χειρισμός της από μηχανές. Εφαρμόζοντας τα αποτελέσματα των προηγούμενων ερευνών του στο πρόβλημα που είχε να αντιμετωπίσει, χρησιμοποίησε και πάλι τη λογική του Boole, καθώς και την εμπειρία του στην κρύπτο / αποκρυπτογράφηση που απόκτησε κατά τη διάρκεια του 2ου Παγκοσμίου πολέμου, προκειμένου να αναπτύξει ένα μοντέλο που θα απλοποιούσε όσο το δυνατόν περισσότερο την πληροφορία. Κατασκεύασε έτσι, ένα δυαδικό σύστημα από δυνατότητες επιλογής ναι/όχι που μπορούσε να αντιπροσωπεύεται από δυαδικό κώδικα 1/0. Πρότεινε επίσης την προσθήκη στην πληροφορία μιας σειράς από ειδικούς κώδικες κατά τη διάρκεια της μετάδοσής της, με στόχο να ελαχιστοποιούνται τα παράσιτα (θόρυβος) που είχαν ως αποτέλεσμα την αλλοίωσή της.

Το 1948, με την δημοσίευση της εργασίας του Shannon, η οποία έγινε σε συνεργασία με τον Weaver με τίτλο “A Mathematical Theory of Communication”, και η οποία δημοσιεύτηκε στο Bell System Technical Journal, γεννήθηκε μια νέα επιστημονική περιοχή, η Θεωρία της Πληροφορίας ή Θεωρία Πληροφοριών. Στόχος αυτής της επιστημονικής περιοχής είναι η θεμελίωση εννοιών και θεωρημάτων που επιτρέπουν τη μαθηματική περιγραφή της διαδικασίας της επικοινωνίας. Με αυτόν τον τρόπο, η μετάδοση πληροφοριών μπορεί να αναλυθεί με μαθηματική αυστηρότητα και ακρίβεια, ενώ σε ένα επόμενο βήμα καθίσταται δυνατόν να σχεδιαστούν καλύτερα συστήματα επικοινωνιών. Η νέα θεωρία, η οποία είναι βασισμένη στη Στατιστική, τη Θεωρία Πιθανοτήτων και την Άλγεβρα είναι ικανή να απαντήσει με μαθηματική ακρίβεια σε ερωτήματα που σχετίζονται με τη βέλτιστη συμπίεση των δεδομένων, την περιγραφή των διαύλων επικοινωνίας, την κωδικοποίηση των μηνυμάτων πληροφορίας, το ρυθμό μετάδοσης των πληροφοριών σε περιβάλλον θορύβου, την κρυπτογράφηση κ.α..

Γεγονός είναι ότι η Θεωρία της Πληροφορίας που όπως διατυπώθηκε από τον Claude Shannon ξεκίνησε την ψηφιακή επανάσταση που οδήγησε στην ανάπτυξη και την εδραίωση νέων μέσων επικοινωνίας – μεταξύ των οποίων και το Internet. Χρησιμοποιήθηκε επίσης για να λυθούν γρίφοι σε γνωστικούς τομείς τόσο διαφορετικούς μεταξύ τους όσο η Πληροφορική, η Γενετική Μηχανική, τα Νευρωνικά

Συστήματα, η Γλωσσολογία, η Φωνητική, η Ψυχολογία και τα Οικονομικά Μεταξύ άλλων άνοιξε νέους δρόμους στη μελέτη του Χάους και έφερε το Διάστημα πιο κοντά στον άνθρωπο.

Στην εργασία του Shannon εισάγεται για πρώτη φορά μια μονάδα μέτρησης της πληροφορίας, το δυαδικό ψηφίο (binary digit), που συντμήθηκε αργότερα αρχικά σε binit και στη συνέχεια στο γνωστό bit. Συγκεκριμένα, ο Shannon , κατανόησε ότι η πληροφορία για ένα γεγονός, είχε άμεση σχέση με τη πιθανότητα του, καταφέρνοντας πρώτος να συνδέσει τις 2 έννοιες μεταξύ τους. Σχετικά με τη μέτρηση της πληροφορίας όπως την όρισε ο Hartley, πρότεινε ότι αυτή μπορεί να χρησιμοποιηθεί, σαν μέτρο πληροφορίας, με την υπόθεση όμως ότι όλα τα σύμβολα θα έχουν ίδια πιθανότητα εμφάνισης.

Για τη γενική περίπτωση, όρισε την μέση πληροφορία H(A) που μπορεί να φέρει ένα πιθανοθεωρητικό πείραμα A, σε ένα δειγματικό χώρο X, να ισούται με

$$H(A) = -\sum_{i=1}^{n} p_i \cdot \log(p_i) \quad , \qquad 1.6$$

όπου με p_i συμβολίζεται η πιθανότητα του ενδεχομένου $x_i \in X$.

Η σύνδεση που έκανε ο Shannon , μεταξύ της πληροφορίας και της πιθανότητας είναι στην πραγματικότητα πολύ λογική αφού η πληροφορία συνδέεται με την πιθανότητα μέσω της έννοιας της αβεβαιότητας. Όσο μικρότερη είναι η πιθανότητα P να συμβεί ένα γεγονός, τόση περισσότερη ποσότητα πληροφορίας συνοδεύει την πραγματοποίηση του. Και αντίστροφα, αν η πιθανότητα πραγματοποίησης ενός γεγονότος είναι μεγάλη, τότε η πληροφορία που μεταφέρει το γεγονός αυτό είναι μικρή. Για παράδειγμα, αν κάποιος πει πως “Θα περάσει ένα μπλε αυτοκίνητο τώρα.”, το μήνυμα αυτό, έχει μεγάλη πληροφορία, γιατί είναι ένα αβέβαιο γεγονός. Αν όμως πει κάποιος πως “Θα περάσει ένα μπλε αυτοκίνητο μέσα στην μέρα.”, τότε το κείμενο αυτό έχει πολύ μικρή πληροφορία. Γιατί στο μήνυμα αυτό η πιθανότητα να περάσει κάποια στιγμή ένα μπλε αυτοκίνητο είναι πολύ μεγάλη, ίσως αγγίζει και το 100%.

1.3 Βασικές Έννοιες

Η βασική έννοια της Θεωρίας της Πληροφορίας, είναι η ίδια η έννοια της πληροφορίας. Η λέξη πληροφορία αναφέρεται σαν μια αλληλουχία συμβόλων, που είτε καταγράφονται είτε μεταδίδονται και μπορεί να ερμηνευτεί ως ένα μήνυμα, το οποίο μεταφέρει κάποια γνώση για κάποιο αντικείμενο, κάτι καινούργιο σχετικά με αυτό, βάζοντας έτσι τέλος στη άγνοια και εξαφανίζοντας την αβεβαιότητα που υπάρχει. Το μήνυμα αυτό μπορεί να επηρεάσει ένα δυναμικό σύστημα το οποίο έχει την ικανότητα να την επεξεργαστεί.

Αναφέρουμε ότι ο όρος “πληροφορία” στην σύγχρονη κοινωνία έχει δύο έννοιες. Η πρώτη είναι η έννοια του “μαθηματικού μεγέθους”, όπως αυτό ειπώθηκε παραπάνω. Η δεύτερη είναι αυτή που χρησιμοποιείται ευρέως στην κοινή γλώσσα και έχει συνυφαστεί με την λέξη “είδηση”. Είναι, δηλαδή, γεγονότα και απόψεις που προσφέρονται και λαμβάνονται από έμβια όντα, μέσα μαζικής επικοινωνίας, ηλεκτρονικούς υπολογιστές (κυρίως μέσω διαδικτύου) και από πάσης φύσεως παρατηρήσιμα φαινόμενα στο περιβάλλον. Τα γεγονότα ενδέχεται να είναι είτε πραγματικά είτε φανταστικά, αποδεδειγμένα ή μη. Ουσιαστικά, αυτό σημαίνει ότι πληροφορία μπορεί να αποτελέσει και ένας ψίθυρος ή μια διάδοση. Ο μηχανισμός διάδοσης των ψιθύρων, μπορεί να είναι τυχαίος ή μη. Γενικά η λέξη πληροφορία στην είδηση, υποδηλώνει τη γνώση που προέρχεται από το εξωτερικό περιβάλλον, έχει μια καθορισμένη μορφή και προσφέρει κάτι καινούργιο, ή τουλάχιστον ένα μέρος της είναι νέο.

Η διαχείριση της πληροφορίας γίνεται μεταδίδοντάς την, αποθηκεύοντάς την ή με την επεξεργασία της ώστε να τη μεταβάλλουμε και να παραχθεί νέα πληροφορία. Οι θεωρητικοί C. S. Pierce (1839 – 1914) και C. W. Morris (1901 – 1979) έκαναν διάκριση της πληροφορίας σε τρεις κατηγορίες ανάλογα με το επίπεδο στο οποίο αξιολογείται η αλληλεπίδρασή της με το σύστημα πομπού – δέκτη. Αυτές οι τρεις κατηγορίες είναι η συντακτική πληροφορία που σχετίζεται με τα σύμβολα και τις σχέσεις μεταξύ αυτών, από τα οποία αποτελούνται τα μηνύματα, η σημασιολογική πληροφορία που σχετίζεται με τη σημασία και η πραγματική που σχετίζεται με τη χρήση και τη δυνατή επίπτωση των μηνυμάτων. Έτσι, ενώ ο συντακτικός τύπος της πληροφορίας αναφέρεται στη μορφή, ο σημασιολογικός και ο πραγματικός αναφέρονται στο περιεχόμενο.

Πιο αναλυτικά, στο συντακτικό επίπεδο η πληροφορία αξιολογείται σύμφωνα με τους επίσημους δεσμούς μεταξύ των διαφόρων στοιχείων που την συνθέτουν, τους κανόνες που διέπουν τον κώδικα επικοινωνίας, την χωρητικότητα των διαύλων επικοινωνίας και το σχεδιασμό συστημάτων και μεθόδων κωδικοποίησης για την μετάδοση, επεξεργασία και αποθήκευσή της.

Το εννοιολογικό επίπεδο αφορά με το πώς διαμορφώνεται η έννοια – το νόημα της πληροφορίας. Όσον αφορά τις πληροφορίες που δίνονται σε φυσική γλώσσα, αυτές βασίζονται στις συμφωνημένες, γραπτές ή άγραφες, πολιτισμικές, πολιτιστικές, ηθικές ή απλά συμφωνημένες συμβάσεις που κάνουν μεταξύ τους τα μέλη μιας ομάδας ανθρώπων. Στις διάφορες εννοιολογικές μονάδες, που μπορεί να είναι για παράδειγμα οι λέξεις που στοιχειοθετούν μια πρόταση, έχει προσδοθεί μια περισσότερο ή λιγότερο ακριβής ή ελεύθερη έννοια. Στην περίπτωση των πιο τεχνικών ή μηχανοποιημένων γλωσσών, όπως είναι οι κώδικες των υπολογιστών, το νόημα των εννοιολογικών μονάδων είναι αμφιμονοσήμαντα ορισμένο βάσει των ιδιοτήτων που παρουσιάζουν και των λειτουργιών που μπορούν να εφαρμόσουν. Τέλος, σε άλλη περίπτωση, όπως αυτήν της μουσικής ως εννοιολογικές πληροφορίες μπορούν να νοηθούν οι συναισθηματικές καταστάσεις που παρατηρούνται σε κατάλληλους αποδέκτες – ακροατές.

Σύμφωνα με τον Shannon οι έννοιες, τεχνολογικά τουλάχιστον, δεν είναι προαπαιτούμενο για τη σωστή επεξεργασία της σύνταξης της πληροφορίας, παρ'όλα αυτά μπορούν να φανούν χρήσιμες σε περιπτώσεις όπως η συμπίεση δεδομένων (με στόχο την αύξηση της αποδοτικότητας της συμπίεσης).

Ολοκληρώνοντας, στο πραγματικό επίπεδο η πληροφορία σχετίζεται με την αξία της χρησιμότητάς της. Το επίπεδο καθορίζεται σε μεγάλο βαθμό από το υπόβαθρο και το περιβάλλον του λήπτη, με αποτέλεσμα να επηρεάζεται σημαντικά από οικονομικούς, πολιτικούς, κοινωνικούς ή/και ψυχολογικούς παράγοντες, ενώ πολλές φορές αποτελεί συνάρτηση του χρόνου αφού πολλές πληροφορίες που φτάνουν με καθυστέρηση δεν αξίζουν πια, ενώ οι πληροφορίες που φτάνουν εγκαίρως, όπως και οι σωστές προβλέψεις, μπορεί να είναι εξαιρετικά πολύτιμες.

Αυτή η διαβάθμιση της πληροφορίας επιτρέπει τη διαχείριση της σε κάθε επίπεδο ξεχωριστά και άρα πιο εξειδικευμένα. Για παράδειγμα επιτρέπεται η διαχείριση της πληροφορίας (μεταφορά, επεξεργασία, αποθήκευση) στο συντακτικό επίπεδο χωρίς να

είναι απαραίτητη η γνώση του εννοιολογικού περιεχομένου της πληροφορίας. Ακόμη, έχουμε τη δυνατότητα να διαχειριζόμαστε τις πληροφορίες βάσει του εννοιολογικού της περιεχομένου χωρίς να λαμβάνουμε υπόψη την πρακτική τους χρησιμότητα.

Ας εξετάσουμε, στη συνέχεια, τις ακόλουθες προτάσεις για να αποσαφηνίσουμε αυτές τις έννοιες.

- ➔ Ο Κώστας είδε να περνάνε δύο λεωφορεία έξω από την στάση Κατεχάκη του μετρό.
- ➔ Έξω από την στάση του μετρό Κατεχάκη, ο Κώστας είδε να περνάνε δύο λεωφορεία.
- ➔ Αύριο στην Αθήνα θα είναι βροχερή μέρα.
- ➔ Αύριο στο κέντρο της Αθήνας θα βρέχει καταρρακτωδώς, στα δυτικά προάστια θα ψιχαλίζει μόνο, στα βόρεια προάστια θα ρίξει χαλάζι, ενώ στα νότια και ανατολικά προάστια θα έχει μόνο συννεφιά.

Οι δύο πρώτες προτάσεις ενώ διαφοροποιούνται ως προς τη σύνταξη, είναι ταυτόσημες ως προς τη σημασία, προσφέρουν δηλαδή την ίδια πληροφόρηση. Αντίθετα, οι δύο τελευταίες προτάσεις διαφέρουν όχι μόνο ως προς τη σύνταξη αλλά και ως προς το περιεχόμενό τους. Η τέταρτη πρόταση ενώ αναφέρεται στο ίδιο γεγονός με την τρίτη πρόταση, είναι πιο λεπτομερής από αυτήν και επομένως προσφέρει περισσότερη πληροφόρηση. Η πραγματική διάσταση της πληροφορίας εξαρτάται κυρίως από το δεδομένο γενικό πλαίσιο. Δηλαδή, η σημασία της τρίτης και της τέταρτης πρότασης είναι σημαντική και ενδιαφέρουσα μόνο για όσους βρίσκονται στην Αθήνα ή ενδεχομένως θέλουν να έρθουν στην Αθήνα την αυριανή μέρα, και όχι για κάποιους που βρίσκονται παραδείγματος χάριν στην Καλαμάτα. Ιδιαίτερα, η ακρίβεια της τέταρτης πρότασης μπορεί να καθορίσει επιλογές, πιθανόν, των εργαζομένων σχετικά με το πώς θα μετακινηθούν αύριο, ή ακόμα και τι θα φορέσουν, ή των καθηγητών των σχολείων για το αν θα πάνε κάποιο περίπατο οι μαθητές την αυριανή μέρα, και πολλά άλλα.

Όπως θα αναφερθεί λεπτομερέστερα και στη συνέχεια, η Θεωρία Πληροφορίας ασχολείται με τη συντακτική πληροφορία, δηλαδή η πληροφορία εξαρτάται από την πιθανότητα εμφάνισης που έχουν τα διάφορα μηνύματα και όχι από τη σημασία τους. Το περιεχόμενο τους δεν παίζει κανένα απολύτως ρόλο. Κάπως έτσι αναπτύχθηκε αυτή η επιστήμη, σύμφωνα με την οποία 1000 σύμβολα τοποθετημένα, με την ίδια σειρά που έχουν τοποθετηθεί άλλα 1000 σύμβολα, αντιμετωπίζονται ακριβώς με τον ίδιο τρόπο,

σαν δύο σύνολα του ίδιου μεγέθους, ανεξάρτητα αν το πρώτο σύνολο προέρχεται από την Ιλιάδα του Ομήρου και το δεύτερο από ένα βιβλίο μαγειρικών συνταγών. Η τάση που επικρατεί στη Θεωρία της Πληροφορίας, σχετίζεται πλέον αποκλειστικά με την ταχύτητα μετάδοσης του μηνύματος, ανεξαρτήτως από τη σημασία που μπορεί να έχει αυτό.

Ο λόγος είναι ότι το κυριότερο πρόβλημα το οποίο έπρεπε να αντιμετωπιστεί από τους επιστήμονες που ανάπτυξαν αυτή τη θεωρία, ήταν το πρόβλημα της ταχύτητας στη μετάδοση της πληροφορίας, μέσω των διαφόρων συσκευών και μηχανημάτων, με βάση τους τεχνικούς και φυσικούς περιορισμούς που υπήρχαν. Προκειμένου να αντιμετωπίσουν αποτελεσματικά αυτά τα προβλήματα, αποφάσισαν να μετρούν την πληροφορία μόνο ποσοτικά και όχι ποιοτικά. Επομένως, η κάθε πληροφορία που μεταδίδεται, αντιμετωπίζεται μόνο σαν μια αλληλουχία συμβόλων τοποθετημένων σε μια σειρά για το οποία το μόνο που ενδιαφέρει είναι το μέγεθος και τίποτα περισσότερο. Έτσι καταλήγουμε στους παρακάτω ορισμούς όσον αφορά δύο πληροφορίες:

(1) Ισοδύναμες πληροφορίες: οι πληροφορίες οι οποίες έχουν την ίδια μορφή ψυχολογικής επίδρασης στο “πομπό”

(2) Διακρινόμενες πληροφορίες: οι πληροφορίες οι οποίες είναι μη ισοδύναμες – έχουν διαφορετική μορφή ψυχολογικής επίδρασης στο “πομπό”

(3) Απόλυτα νέες πληροφορίες: αυτές οι οποίες διακρίνονται από κάθε γνωστή πληροφορία

(4) Εμμένουσες πληροφορίες: αυτές οι οποίες μπορούν να αποκομιστούν από μια περιοχή του χώρου, αρκετό καιρό μετά την εκδήλωσή τους (π.χ. φωτογραφίες, σχέδια, CD)

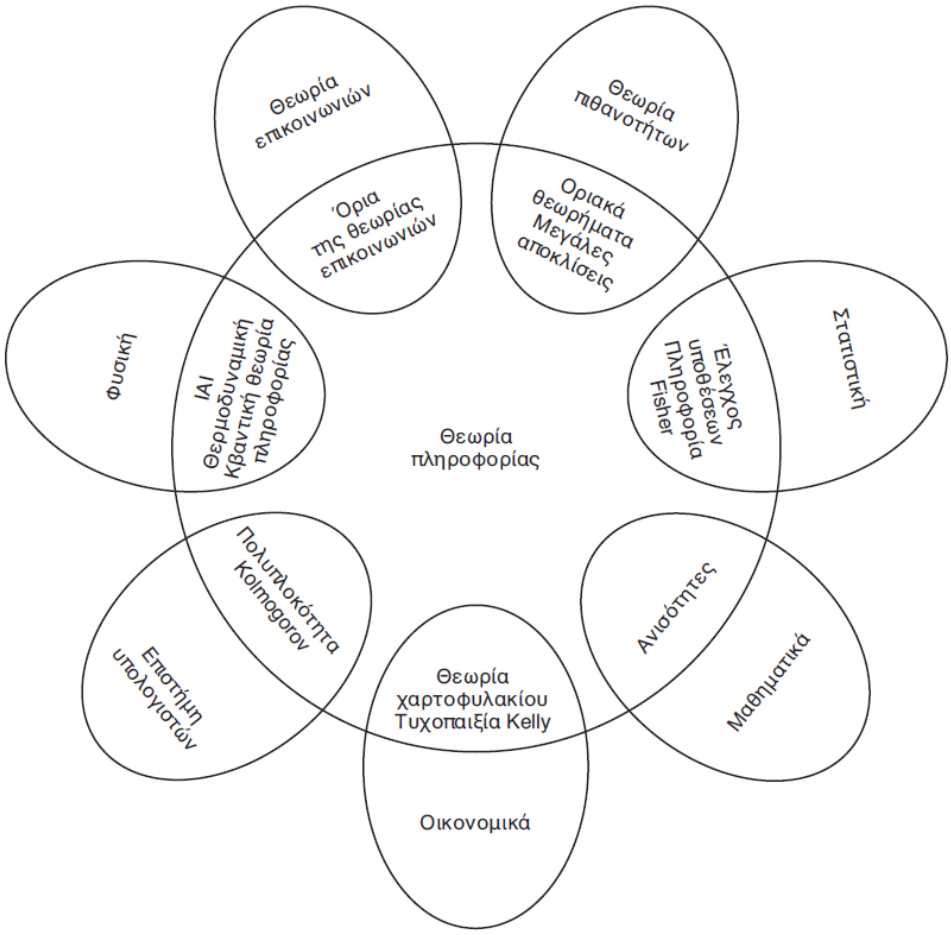
(5) Μεταβατικές πληροφορίες: αυτές οι οποίες δεν εκδηλώνονται σε μια περιοχή του χώρου παρά για ένα μόνο σύντομο χρονικό διάστημα (π.χ. τηλεφωνικές συνδέσεις)

(6) Ασυνεχείς πληροφορίες: αυτές που αποτελούνται από ξεχωριστά στοιχεία στο χώρο (π.χ. τηλεγραφήματα)

(7) Συνεχείς πληροφορίες: οι πληροφορίες που τα στοιχεία από τα οποία αποτελούνται είναι απολύτως συνεχή μεταξύ τους τοποθετημένα το ένα δίπλα στο άλλο.

1.4 Εφαρμογές

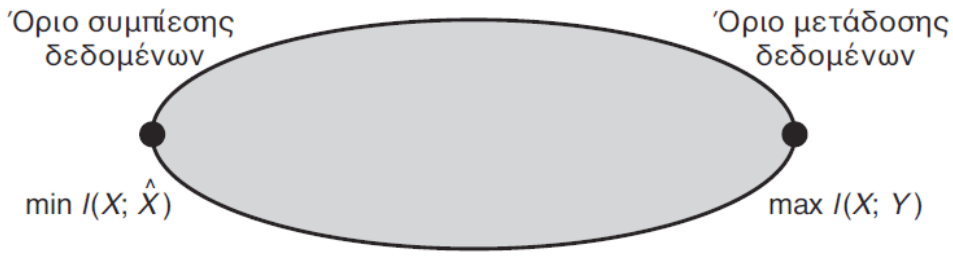
Όπως διατυπώθηκε προηγουμένως, οι συνεισφορές της Θεωρίας Πληροφοριών σε πολλά επιστημονικά πεδία με χαρακτηριστικότερα τη Στατιστική Φυσική (Θερμοδυναμική), την Επιστήμη των Υπολογιστών (πολυπλοκότητα Kolmogorov ή αλγοριθμική πολυπλοκότητα), τη Στατιστική Συμπερασματολογία (ξυράφι του Όκκαμ), τις Πιθανότητες και τη Στατιστική (εκθέτες σφάλματος για τον βέλτιστο έλεγχο υποθέσεων και εκτίμηση) είναι θεμελιώδεις. Αυτό φαίνεται στο παρακάτω σχήμα (Σχήμα 1.1). Στη συνέχεια, ακολουθεί λεπτομερέστερη ανάλυση της συσχέτισης του κάθε πεδίου με την επιστήμη της Θεωρίας Πληροφοριών.



Σχήμα 1.1 Σχέση της Θεωρίας Πληροφορίας με άλλα γνωστικά αντικείμενα

Ηλεκτρομηχανική (Θεωρία Επικοινωνιών). Στις αρχές της δεκαετίας του 1940 εθεωρείτο

ότι είναι αδύνατο να σταλεί πληροφορία με θετικό ρυθμό και αμελητέα πιθανότητα σφάλματος. Ο Shannon εξέπληξε τους επιστήμονες που ασχολούνταν με τη Θεωρία Επικοινωνιών αποδεικνύοντας ότι η πιθανότητα σφάλματος μπορούσε να γίνει σχεδόν μηδενική για κάθε ρυθμό επικοινωνίας μικρότερο από τη χωρητικότητα του διαύλου. Η χωρητικότητα μπορεί να υπολογιστεί με απλό τρόπο από τα χαρακτηριστικά του διαύλου που αφορούν τον θόρυβο. Ο Shannon υποστήριξε επιπλέον ότι τυχαίες διεργασίες, όπως η μουσική και η ομιλία, έχουν ένα επίπεδο πολυπλοκότητας ανεπίδεκτο περαιτέρω μείωσης και ότι το σήμα δεν μπορεί να συμπιεστεί περισσότερο από όσο επιτρέπει αυτό το επίπεδο. Σε αυτή την ποσότητα έδωσε το όνομα εντροπία, έχοντας υπόψη του την παράλληλη χρήση του όρου στη Θερμοδυναμική, και έδειξε ότι αν η εντροπία της πηγής είναι μικρότερη από τη χωρητικότητα του διαύλου, τότε ασυμπτωτικά μπορεί να επιτευχθεί επικοινωνία άνευ σφαλμάτων.



Σχήμα 1.2 Η Θεωρία Πληροφορίας αντιπροσωπεύει τα δύο ακραία σημεία της Θεωρίας Επικοινωνιών

Όπως φαίνεται στο Σχήμα 1.2, σήμερα η Θεωρία Πληροφορίας αντιπροσωπεύει τα ακραία σημεία του συνόλου όλων των δυνατών τεχνικών επικοινωνίας. Στο ένα άκρο του συνόλου των δυνατών τεχνικών επικοινωνίας βρίσκεται το ελάχιστο $I(X;\hat{X})$ της συμπίεσης δεδομένων. Σε όλες τις τεχνικές συμπίεσης δεδομένων ο ρυθμός περιγραφής πρέπει να είναι τουλάχιστον ίσος με αυτό το ελάχιστο. Στο άλλο άκρο βρίσκεται το μέγιστο $I(X;Y)$ της μετάδοσης δεδομένων, που είναι γνωστό ως χωρητικότητα διαύλου. Συνεπώς, όλες οι τεχνικές διαμόρφωσης και συμπίεσης δεδομένων βρίσκονται μεταξύ αυτών των ορίων.

Η Θεωρία Πληροφορίας προτείνει επίσης τρόπους για την επίτευξη αυτών των εσχάτων ορίων που διέπουν την επικοινωνία. Ωστόσο, αυτές οι θεωρητικά βέλτιστες τεχνικές επικοινωνίας, όσο κομψές και αν είναι, πιθανόν να μην είναι καθόλου πρακτικές από

υπολογιστικής πλευράς. Το γεγονός ότι χρησιμοποιούμε κάποιες απλές τεχνικές διαμόρφωσης και αποδιαμόρφωσης, αντί της τυχαίας κωδικοποίησης και του κανόνα αποκωδικοποίησης, βάσει του πλησιέστερου γείτονα που προτείνεται στην απόδειξη του Shannon για το θεώρημα χωρητικότητας διαύλου, οφείλεται απλώς και μόνο στην υπολογιστική εφικτότητά τους. Η πρόοδος στα ολοκληρωμένα κυκλώματα και τον σχεδιασμό κωδίκων μας έχει επιτρέψει να απολαύσουμε κάποια από τα οφέλη που προβλέπει η θεωρία του Shannon. Η υπολογιστική πρακτικότητα τελικά επιτεύχθηκε με την έλευση των κωδίκων turbo. Ένα καλό παράδειγμα εφαρμογής των ιδεών της Θεωρίας Πληροφορίας είναι η χρήση των κωδίκων διόρθωσης σφαλμάτων στους συμπαγείς δίσκους και στα DVD. Η πρόσφατη ερευνητική δραστηριότητα πάνω στις πλευρές της Θεωρίας Πληροφορίας που αφορούν τις επικοινωνίες έχει επικεντρωθεί στη Δικτυακή Θεωρία Πληροφορίας: τη θεωρία των ταυτόχρονων ρυθμών μετάδοσης από πολλούς πομπούς σε πολλούς δέκτες παρουσία παρεμβολής και θορύβου. Κάποιες από τις σχέσεις αλληλεξάρτησης των ρυθμών μεταξύ πομπών και δεκτών είναι μη αναμενόμενες, αλλά όλες τους έχουν μια μαθηματική απλότητα. Ωστόσο, δεν έχει βρεθεί ακόμα κάποια ενιαία θεωρία.

Επιστήμη των Υπολογιστών (πολυπλοκότητα Kolmogorov). Οι Kolmogorov, Chaitin και Solomonoff διατύπωσαν την ιδέα ότι η πολυπλοκότητα μιας συμβολοσειράς δεδομένων μπορεί να οριστεί μέσω του μήκους του μικρότερου δυνατού δυαδικού προγράμματος υπολογιστή που υπολογίζει τη συγκεκριμένη συμβολοσειρά. Δηλαδή η πολυπλοκότητα είναι το ελάχιστο μήκος περιγραφής. Όπως αποδεικνύεται, αυτός ο ορισμός της πολυπλοκότητας είναι καθολικός, δηλαδή ανεξάρτητος από τον εκάστοτε υπολογιστή, και θεμελιώδους σημασίας. Επομένως, η πολυπλοκότητα Kolmogorov θέτει τις βάσεις για τη Θεωρία της Περιγραφικής Πολυπλοκότητας. Είναι μάλιστα ιδιαίτερα ευχάριστο το γεγονός ότι η πολυπλοκότητα Kolmogorov είναι προσεγγιστικά ίση με την εντροπία Shannon Η, αν η ακολουθία λαμβάνεται τυχαία σύμφωνα με μια κατανομή που έχει εντροπία Η. Άρα, η σχέση μεταξύ Θεωρίας Πληροφορίας και πολυπλοκότητας Kolmogorov είναι τέλεια. Πράγματι, η πολυπλοκότητα Kolmogorov θεωρείται πιο θεμελιώδης από την εντροπία Shannon. Είναι η υπέρτατη συμπίεση δεδομένων και οδηγεί σε έναν λογικά συνεπή τρόπο εξαγωγής συμπερασμάτων. Υπάρχει μια ευχάριστη συμπληρωματική σχέση μεταξύ της αλγοριθμικής και της υπολογιστικής πολυπλοκότητας. Μπορούμε να φανταστούμε την υπολογιστική πολυπλοκότητα (χρονική

πολυπλοκότητα) και την πολυπλοκότητα Kolmogorov (μήκος προγράμματος ή περιγραφική πολυπλοκότητα) σαν δύο άξονες που αντιστοιχούν ο πρώτος στον χρόνο εκτέλεσης και ο δεύτερος στο μήκος ενός προγράμματος. Η πολυπλοκότητα Kolmogorov επικεντρώνεται στην ελαχιστοποίηση ως προς τον δεύτερο άξονα, ενώ η υπολογιστική πολυπλοκότητα στην ελαχιστοποίηση ως προς τον πρώτο άξονα. Η ταυτόχρονη ελαχιστοποίηση και των δύο δεν έχει μελετηθεί ακόμη αρκετά.

Φυσική (Θερμοδυναμική). Η εντροπία και ο δεύτερος νόμος της Θερμοδυναμικής γεννήθηκαν στους κόλπους της Στατιστικής Μηχανικής. Η εντροπία είναι ένα μέγεθος που αυξάνεται πάντα. Μεταξύ άλλων, ο δεύτερος νόμος της θερμοδυναμικής μας επιτρέπει να απορρίψουμε κάθε ισχυρισμό περί αεικίνητων μηχανών.

Μαθηματικά (Θεωρία Πιθανοτήτων και Στατιστική). Οι θεμελιώδεις ποσότητες της Θεωρίας Πληροφορίας –η εντροπία, η σχετική εντροπία και η αμοιβαία πληροφορία– ορίζονται ως συναρτησιοειδή κατανομών πιθανότητας. Χαρακτηρίζουν τη συμπεριφορά μακρών ακολουθιών τυχαίων μεταβλητών και μας επιτρέπουν να εκτιμούμε τις πιθανότητες σπάνιων ενδεχομένων (Θεωρία Μεγάλων Αποκλίσεων) και να βρίσκουμε τον καλύτερο εκθέτη σφάλματος κατά τον έλεγχο υποθέσεων.

Φιλοσοφία της Επιστήμης (ξυράφι του Όκκαμ). Ο Γουλιέλμος του Όκκαμ είχε πει: “Τα αίτια δεν πρέπει να πολλαπλασιάζονται περισσότερο από όσο είναι απαραίτητο”, με άλλα λόγια: “Η απλούστερη εξήγηση είναι η καλύτερη”. Οι Solomonoff και Chaitin υποστήριξαν ότι αν πάρουμε έναν σταθμισμένο συνδυασμό όλων των προγραμμάτων που εξηγούν κάποια δεδομένα και παρατηρήσουμε την επόμενη έξοδό τους, τότε έχουμε μια καθολικά καλή μέθοδο πρόβλεψης. Επιπλέον, αυτή η διαδικασία εξαγωγής συμπερασμάτων είναι αποτελεσματική σε πολλά προβλήματα που δεν μπορούμε να χειριστούμε μέσω της στατιστικής. Για παράδειγμα, αυτή η μέθοδος θα προβλέψει τελικά τα επόμενα ψηφία του αριθμού π. Αν την εφαρμόσουμε στις ρίψεις ενός κέρματος που φέρνει γράμματα με πιθανότητα $p=0,7$ και κορόνα με πιθανότητα $q=0,3$, θα μπορέσουμε να συμπεράνουμε το αποτέλεσμα της επόμενης ρίψης. Αν την εφαρμόσουμε στη χρηματιστηριακή αγορά, θα πρέπει ουσιαστικά να βρει όλους τους “νόμους” της χρηματιστηριακής αγοράς και να τους προεκβάλλει κατά βέλτιστο τρόπο. Θεωρητικά, χρησιμοποιώντας μια τέτοια μέθοδο θα μπορούσαμε να ανακαλύψουμε και τους νόμους της φυσικής του Νεύτωνα. Ασφαλώς, η συγκεκριμένη μέθοδος εξαγωγής

συμπερασμάτων δεν είναι καθόλου πρακτική, διότι για να απορριφθούν όλα τα προγράμματα υπολογιστή που δεν καταφέρνουν να παραγάγουν τα υπάρχοντα δεδομένα απαιτείται απαγορευτικά πολύς χρόνος. Θα προβλέπαμε τι πρόκειται να συμβεί αύριο ύστερα από εκατό χρόνια.

Οικονομικά (επενδύσεις). Οι επαναλαμβανόμενες επενδύσεις σε μια στάσιμη χρηματιστηριακή αγορά έχουν ως αποτέλεσμα την εκθετική μεγέθυνση του πλούτου. Ο ρυθμός μεγέθυνσης του πλούτου είναι μια δυϊκή ποσότητα του ρυθμού εντροπίας της χρηματιστηριακής αγοράς. Οι αντιστοιχίες μεταξύ της Θεωρίας Βέλτιστων Επενδύσεων σε μια χρηματιστηριακή αγορά και της Θεωρίας Πληροφορίας είναι εντυπωσιακές.

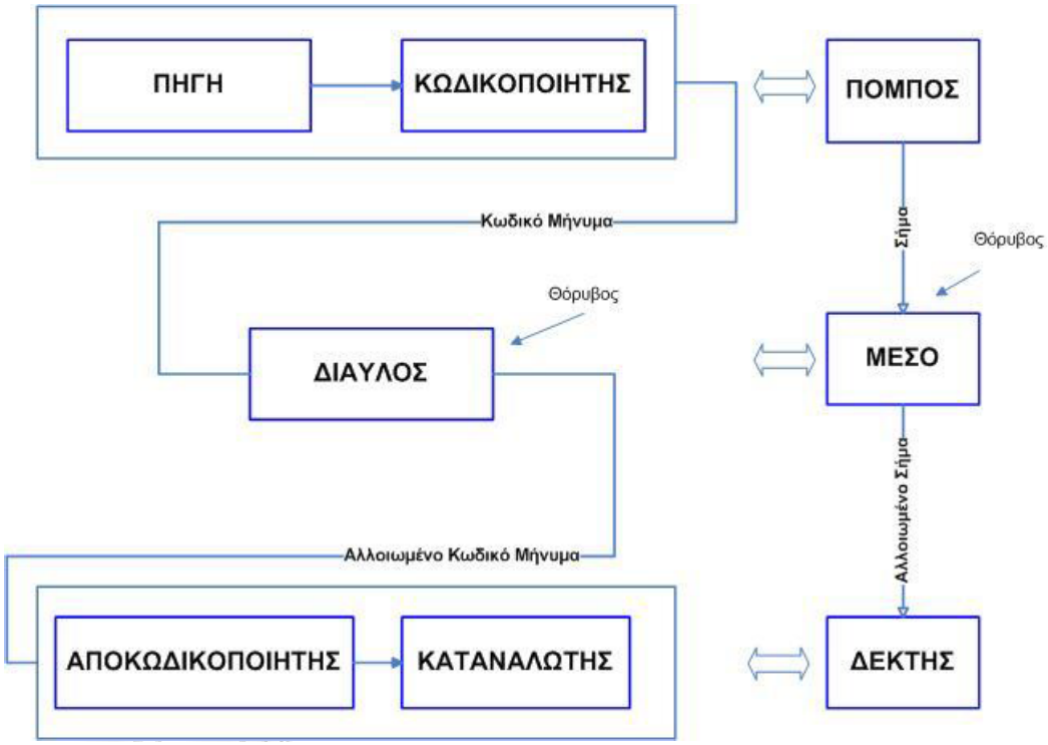
Υπολογισμός και Επικοινωνία. Καθώς κατασκευάζουμε ολοένα και μεγαλύτερους υπολογιστές από ολοένα και μικρότερα μέρη, βρισκόμαστε αντιμέτωποι με ένα όριο που αφορά τόσο τις υπολογιστικές όσο και τις επικοινωνιακές δυνατότητες. Οι επικοινωνιακές δυνατότητες επιβάλλουν περιορισμούς στις υπολογιστικές δυνατότητες και ανάποδα οι υπολογιστικές δυνατότητες επιβάλλουν περιορισμούς στις επικοινωνιακές δυνατότητες. Αυτά τα δύο συνυφαίνονται, και έτσι όλες οι ανακαλύψεις της Θεωρίας Επικοινωνιών μέσω της Θεωρίας Πληροφορίας έχουν άμεση επίδραση στη Θεωρία Υπολογισμού.

1.5 Μεταφορά της Πληροφορίας

Σε κάθε επικοινωνιακή διεργασία που είναι σε εξέλιξη, λαμβάνει χώρα ροή πληροφορίας μεταξύ ενός αποστολέα και ενός αποδέκτη. Η πληροφορία αυτή μπορεί να έχει διάφορες μορφές αναλόγως με την περίπτωση, όπως ηλεκτρισμού, μουσικής, λέξεων ή ακόμα και εικόνων. Η μεταφορά της πληροφορίας επιτυγχάνεται, στη γενική περίπτωση, με τη βοήθεια ενός δικτύου μετάδοσης. Έτσι, τα βασικά μέρη ενός επικοινωνιακού μοντέλου είναι ο αποστολέας ή πηγή πληροφορίας, το κανάλι ή δίκτυο μετάδοσης και ο παραλήπτης ή προορισμός αυτής.

Η αποθήκευση της πληροφορίας παίζει σήμερα σημαντικό ρόλο. Αν και κατά κανόνα δεν είναι ζήτημα μετάδοσης, μπορεί ωστόσο να περιγραφεί ως μέρος του καναλιού ή δικτύου μετάδοσης. Η πληροφορία κατά τη μετάδοσή της μπορεί να αλλοιωθεί από την

επενέργεια του θορύβου πάνω στο κανάλι. Ένα στοιχειώδες σύστημα επικοινωνίας στα πλαίσια της Θεωρίας της Πληροφορίας αντιστοιχεί στο κλασσικό τηλεπικοινωνιακό σύστημα πομπού, διαύλου και δέκτη, όπως απεικονίζεται στο παρακάτω σχήμα (Σχήμα 1.3).



Σχήμα 1.3 Γενικό διάγραμμα συστήματος τηλεπικοινωνίας

Ο πομπός αποτελείται από την πηγή πληροφορίας και τον κωδικοποιητή. Η πληροφορία παράγεται στην πηγή (πληροφορίας) και οργανώνεται σε μηνύματα πληροφορίας, τα οποία στη συνέχεια μετατρέπονται σε κωδικά μηνύματα. Ο δίαυλος επικοινωνίας, ο οποίος είναι στην ουσία το μέσο που παρεμβάλλεται μεταξύ του πομπού και του δέκτη, διοχετεύει την κωδικοποιημένη πληροφορία στο σημείο προορισμού. Όπως αναφέρθηκε και παραπάνω όταν η πληροφορία διαπερνά το δίαυλο, είναι δυνατόν να αλλοιωθεί λόγω της παρουσίας θορύβου. Η πληροφορία λαμβάνεται από το δέκτη, όπου αρχικά αποκωδικοποιείται και στη συνέχεια παρουσιάζεται στον προορισμό της.

Η μεταφορά της πληροφορίας θα πρέπει να είναι, ως ένα βαθμό, χωρίς σφάλματα. Γι'αυτό πρέπει να είναι δυνατή η διόρθωση σφαλμάτων ή η μεταφορά να είναι τόσο καλή, ώστε να μην υπεισέρχονται παρά μόνο ασήμαντα σφάλματα που είναι ανεκτά. Μια τέλεια, δηλαδή χωρίς σφάλματα, μεταφορά δεν είναι δυνατή για σήματα ομιλίας,

μουσικής ή video. Μπορούν μόνο να τεθούν απαιτήσεις ως προς το μέγεθος της απόκλισης του σήματος που λαμβάνει ο αποδέκτης από το σήμα που έχει αποστείλει ο μεταδότης. Η απαιτούμενη ποιότητα μεταφοράς της πληροφορίας οδηγεί στη επιλογή κατάλληλου μέσου μεταφοράς ή καναλιού και επιβάλλει οριακές συνθήκες προσαρμογής του καναλιού στον αποστολέα και τον παραλήπτη. Μια από τις σημαντικές επιδιώξεις σχεδιαστών επικοινωνιακών συστημάτων είναι η ελαχιστοποίηση των απωλειών πληροφορίας στο κανάλι και η βέλτιστη επανάκτηση πληροφορίας που έχει προσβληθεί από θόρυβο. Για την επίτευξη της επιδίωξης αυτής χρησιμοποιούνται τεχνικές κωδικοποίησης στην πλευρά του αποστολέα και αντίστοιχες τεχνικές αποκωδικοποίησης στην πλευρά του αποδέκτη. Λαμβάνοντας υπόψη την κωδικοποίηση και την αποκωδικοποίηση, οδηγούμαστε στην γενική δομή ενός επικοινωνιακού μοντέλου όπως αυτό παρουσιάστηκε πιο πάνω στο Σχήμα 1.3.

Ακολουθεί μια αναλυτική περιγραφή των εννοιών του συστήματος επικοινωνίας από την άποψη των τηλεπικοινωνιών.

Επικοινωνία είναι κάθε διαδικασία μεταφοράς πληροφορίας μεταξύ δύο σημείων του χώρου – χρόνου (π.χ. τηλεφωνική συνδιάλεξη).

Πηγή πληροφορίας είναι το τμήμα του συστήματος επικοινωνίας που παράγει πληροφορία με τη μορφή συμβόλων (π.χ. δελτίο καιρού). Η πληροφορία προσαρτάται στα σύμβολα με κριτήριο τη πιθανότητα εμφάνισης τους στην έξοδο της πηγής πληροφορίας.

Αλφάβητο είναι το σύνολο των συμβόλων που χρησιμοποιεί η πηγή πληροφορίας (π.χ. αριθμοί, γράμματα, διαγράμματα, χάρτες).

Λέξη Πληροφορίας είναι η βραχεία διάταξη συμβόλων πληροφορίας (π.χ. λέξη αποτελούμενη από γράμματα, όπως σταθμός).

Μήνυμα Πληροφορίας είναι η διάταξη των λέξεων πληροφορίας (π.χ. μια πρόταση αποτελούμενη από λέξεις, όπως ο σιδηροδρομικός σταθμός είναι συνέχεια ανοικτός).

Κωδικοποίηση είναι η αντικατάσταση των συμβόλων πληροφορίας από άλλα (κωδικά) σύμβολα με αντικειμενικό σκοπό τη βελτιστοποίηση της επικοινωνίας (π.χ. αντικατάσταση γραμμάτων από τελείες και παύλες κατά τον κώδικα Morse). Η κωδικοποιημένη πληροφορία οργανώνεται επίσης σε επιμέρους κωδικές λέξεις και

κωδικά μηνύματα.

Κώδικας είναι κάθε τεχνική κωδικοποίηση. Το σύνολο των κωδικών συμβόλων είναι το αλφάβητο του κώδικα. Η αμφιμονοσήμαντη απεικόνιση συμβόλων, λέξεων και μηνυμάτων πληροφορίας σε κωδικά σύμβολα, κωδικές λέξεις και κωδικά μηνύματα είναι το κλειδί του κώδικα. Έστω για παράδειγμα ότι έχουμε το κωδικό αλφάβητο Γ={0,1}. Η δυαδική κωδικοποίηση είναι η συνηθέστερη επιλογή στα ψηφιακά συστήματα επικοινωνίας. Με βάση αυτό το κωδικό αλφάβητο, ας υποθέσουμε ότι για την πηγή πληροφορίας με αλφάβητο Α={α_1,α_2,α_3} υπάρχουν οι τρεις παρακάτω δυαδικοί κώδικες (Σχήμα 1.4):

$\alpha_1 \rightarrow 0$	$\alpha_1 \rightarrow 00$	$\alpha_1 \rightarrow 0$
$\alpha_2 \rightarrow 1$	$\alpha_2 \rightarrow 01$	$\alpha_2 \rightarrow 11$
$\alpha_3 \rightarrow 1$	$\alpha_3 \rightarrow 10$	$\alpha_3 \rightarrow 10$
I	II	III

Σχήμα 1.4 Δυαδικοί Κώδικες

Ο κώδικας I είναι απεικόνιση του Α στο Γ, ο κώδικας II είναι απεικόνιση του Α στο Γ^2 ενώ ο κώδικας III είναι απεικόνιση του Α στο $\Gamma\cup\Gamma^2$. Οι δύο πρώτοι κώδικες έχουν σταθερό αριθμό δυαδικών συμβόλων (ισομήκεις) ενώ ο τρίτος μεταβλητό (δεν είναι ισομήκης).

Απαραίτητο στοιχείο του πομπού είναι ο μεταλλάκτης που μετατρέπει το κωδικοποιημένο μήνυμα σε σήμα, δηλαδή μορφή κατάλληλη για μετάδοση (π.χ. σειρά

ηλεκτρικών παλμών). Το σήμα αποτελεί τον υλικό φορέα της πληροφορίας.

Δίαυλος Πληροφορίας ή Κανάλι είναι η αλυσίδα μέσων και συσκευών (π.χ. καλώδια, κυματοδηγοί, οπτικές ίνες) που μεταδίδουν το σήμα με την αποτυπωμένη σε αυτό πληροφορία.

Χωρητικότητα Διαύλου Πληροφορίας είναι ο μέγιστος αριθμός μετάδοσης πληροφορίας (π.χ. το τηλέτυπο μεταδίδει 10 λέξεις / sec). Καθορίζει το χρόνο και το κόστος που απαιτούνται για τη μετάδοση μηνύματος ή το πλήθος των μηνυμάτων που είναι δυνατό να διοχετεύει ταυτόχρονα ο δίαυλος πληροφορίας.

Θόρυβος είναι κάθε ανεξέλεγκτη παρεμβολή του περιβάλλοντος του διαύλου που προκαλεί αλλοίωση του σήματος και συνεπώς σφάλματα μετάδοσης (απώλεια πληροφορίας). Συνήθως στα κανάλια επικοινωνίας υπάρχουν διαφόρων ειδών θόρυβοι όπως ο θερμικός θόρυβος, ο κρουστικός θόρυβος, ο θόρυβος περιβάλλοντος ή η παρεμβολή ομιλίας από άλλες γραμμές (κανάλια). Μέχρι το 1948 ο τηλεπικοινωνιακός μηχανικός επιδίωκε την προστασία του σήματος από τον θόρυβο, δηλαδή την πιστή αναπαραγωγή του σήματος στο δέκτη. Με την ωρίμανση της Θεωρίας Πληροφορίας, το ενδιαφέρον μετατοπίστηκε στην πιστή αναπαραγωγή του μηνύματος πληροφορίας που είναι αποτυπωμένο το σήμα. Σύγχρονα τηλεπικοινωνιακά συστήματα εξασφαλίζουν αξιόπιστη ροή πληροφορίας με σήμα βαθιά θαμμένο σε θόρυβο.

Αποκωδικοποιητής αντιπροσωπεύει την επεξεργασία που γίνεται στο σήμα που προκύπτει στην έξοδο του καναλιού προκειμένου να αναπαραχθεί ένα όσο το δυνατόν πιστότερο αντίγραφο του σήματος στην έξοδο της πηγής πληροφορίας.

Κεφάλαιο 2:

Μέτρα Πληροφορίας

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο εισάγουμε τους περισσότερους από τους βασικούς ορισμούς που απαιτούνται για την ανάπτυξη της Θεωρίας Πληροφοριών. Αφού διατυπωθούν, θα αναλυθούν εκτενέστερα οι μεταξύ τους σχέσεις και ερμηνείες. Στη θεωρία της πληροφορίας, όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, η πληροφορία θεωρείται μετρήσιμο μέγεθος. Η έννοια της πληροφορίας, βέβαια, είναι πολύ ευρεία για να καλυφθεί πλήρως από έναν και μόνο ορισμό. Για αυτό ακριβώς, αναπτύχθηκαν συγκεκριμένα μέτρα τα οποία είναι υπεύθυνα για την μέτρηση αυτής.

Για κάθε κατανομή πιθανότητας θα ορίσουμε μια ποσότητα που ονομάζεται εντροπία. Η εντροπία αποτιμά τον μέσο όρο πληροφορίας που φέρει μια τυχαία μεταβλητή, και έχει πολλές ιδιότητες που συμφωνούν με όσα θα αναμέναμε διαισθητικά από ένα μέτρο πληροφορίας. Επεκτείνοντας αυτή την έννοια, θα ορίσουμε και την έννοια της αμοιβαίας πληροφορίας, η οποία είναι ένα μέτρο της ποσότητας πληροφορίας που περιέχει μια τυχαία μεταβλητή σχετικά με κάποια άλλη. Υπό αυτό το πρίσμα, η εντροπία είναι η αυτοπληροφορία μιας τυχαίας μεταβλητής. Η αμοιβαία πληροφορία είναι ειδική περίπτωση μιας γενικότερης ποσότητας που ονομάζεται σχετική εντροπία, η οποία είναι ένα μέτρο της απόστασης μεταξύ δύο κατανομών πιθανότητας. Θα ορίσουμε και μερικές άλλες σχετικές έννοιες. Όλες αυτές οι ποσότητες σχετίζονται στενά μεταξύ τους, και έχουν ορισμένες απλές κοινές ιδιότητες, μερικές από τις οποίες θα αποδείξουμε σε αυτό το κεφάλαιο.

2.2 Ορισμοί

Τα κυριότερα μέτρα πληροφορίας είναι τα παρακάτω:

(1) Η **εντροπία** (entropy), η οποία μετράει την μέση πληροφορία που φέρει μια τυχαία μεταβλητή X

(2) Η **σχετική εντροπία** (relative entropy), η οποία μετράει την ομοιότητα των X και Y.

(3) Η **κοινή εντροπία** (joint entropy), η οποία μετράει τη συνολική πληροφορία των X και Y.

(4) Η **δεσμευμένη** ή **υπό συνθήκη εντροπία** (conditional entropy), η οποία μετράει την πληροφορία του X, όταν η Y είναι γνωστή και αντιστρόφως.

(5) Η **αμοιβαία πληροφορία** ή **διαπληροφορία** (mutual information) μετρά την μείωση της αβεβαιότητας για το X, όταν είναι γνωστή η Y μεταβλητή.

(6) Η **υπο συνθήκη αμοιβαία πληροφορία** (conditional mutual information) η οποία μετρά την αναμενόμενη αμοιβαία πληροφορία μεταξύ δύο μεταβλητών X, Y όταν είναι γνωστή μια τρίτη μεταβλητή Z.

(7) Η **υπο συνθήκη σχετική εντροπία** (conditional relative entropy) η οποία μετράει το σταθμισμένο άθροισμα των σχετικών εντροπιών των δεσμευμένων κατανομών των μεταβλητών X, Y για τις διάφορες τιμές του Y.

Πιο κάτω παρουσιάζονται αναλυτικά κάθε μια από αυτές τις έννοιες.

2.3 Εντροπία

Η πρώτη και ίσως και η δυσκολότερη και πιο αφηρημένη έννοια που ορίστηκε είναι αυτή της εντροπίας. Η λέξη εντροπία είναι σύνθετη και προέρχεται από τις λέξεις “εν” και “τροπή” και ουσιαστικά σημαίνει εσωτερική αλλαγή ή αλλαγή εντός ενός συστήματος. Πρώτη φορά χρησιμοποιήθηκε το 1850 από το Γερμανό φυσικό Rudolf Clausius στο πλαίσιο των μελετών του για τη θερμοδυναμική, ο οποίος ήθελε να περιγράψει τον εκφυλισμό της ενέργειας κατά τη διαδικασία της μετατροπής της από μια μορφή σε μια

άλλη μέσα σε ένα ενεργειακώς κλειστό σύστημα, το οποίο δεν ανταλλάσσει ενέργεια με το περιβάλλον. Επειδή όμως δεν είναι όλη η διαθέσημη ενέργεια χρήσιμη, δηλαδή κάποιο κομμάτι της μπορεί μεν να χρησιμοποιηθεί για την παραγωγή έργου (δηλ. ενέργειας), ωστόσο ένα ποσοστό της είναι άχρηστο, η ποσότητα της εντροπίας, σύμφωνα με τον Clausius, εκφράζει ακριβώς αυτό το ποσοστό που δεν μπορεί να χρησιμοποιηθεί. Στην περίπτωση που η εντροπία ενός κλειστού συστήματος είναι ίση με το μηδέν, τότε όλη η ενέργεια που είναι διαθέσιμη σε αυτό το σύστημα, μπορεί να χρησιμοποιηθεί για την παραγωγή έργου. Σε κάθε άλλη περίπτωση η "χρήσιμη" ενέργεια ισούται με τη συνολική ενέργεια του συστήματος μείον ένα ποσοστό της που εκφράζεται από την εντροπία του συγκεκριμένου συστήματος.

Το 1948 ο Shannon, μέσω της γνωστής, πλέον εργασίας του “A Mathematical Theory of Communication”, καταφέρνει να ποσοτικοποιήσει την πληροφορία, και επηρεασμένος από τον Janos Neumann αποφασίζει να ονομάσει την ποσότητα που μετρούσε την πληροφορία εντροπία.

Η εντροπία στη θεωρία πληροφοριών είναι στην ουσία το μέτρο αβεβαιότητας που διακατέχει το σύστημα. Ο Shannon παρατήρησε το εξής φαινόμενο. Όσος λιγότερος θόρυβος παράγεται σε ένα μοντέλο επικοινωνίας, πομπού και δέκτη, τόση περισσότερη πληροφορία μεταδίδει. Και αντιστρόφως, όσο αυξάνεται η αταξία (θόρυβος) ενός συστήματος τόσο λιγότερη πληροφορία μεταδίδει αυτό. Θα μπορούσαμε να πούμε πως η πληροφορία του συστήματος αποτελεί μέτρο της εσωτερικής του τάξης (δηλ. αντιστρόφως ανάλογη με την αταξία). Αλλά η εντροπία είναι το μέτρο της αταξίας ενός συστήματος, άρα η πληροφορία είναι αντιστρόφως ανάλογη της εντροπίας. Αυτός είναι βασικά και ο λόγος που συχνά αναφέρεται η πληροφορία A σαν η αρνητική εντροπία H, δηλαδή ισχύει ότι $A=-H$. Άρα έχουμε τον εξής ορισμό για την εντροπία:

Αν X είναι μια διακριτή τυχαία μεταβλητή με δειγματικό χώρο $X=\{x_1,x_2,..,x_n\}$ και συνάρτηση μάζας πιθανότητας p_i , με $p_i>0$ και $\sum_{i=1}^{n} p_i=1$, τότε η μέση ποσότητα πληροφορίας (ή μέση πληροφορία) της X, $H(X)$, δίνεται από τη σχέση:

$$H_b(X)=-\sum_{i=1}^{n} p_i \log_b(p_i) \quad . \qquad 2.1$$

Η μέση πληροφορία ονομάζεται και εντροπία (όπως είπαμε και στο προηγούμενο κεφάλαιο).

Η βάση b του λογαρίθμου, συνήθως λαμβάνει τη τιμή 2, και η εντροπία σε αυτήν την περίπτωση μετριέται σε bits ενώ όταν η βάση b ισούται με e, τότε η εντροπία μετριέται σε nats. Επιπλέον, ισχύει η σύμβαση ότι $0 \cdot \log(0) = 0$, η οποία προκύπτει από τον ορισμό της συνέχειας καθώς ισχύει ότι $x \cdot \log(x) \to 0$ καθώς $x \to 0$. Σημειωτέον ότι η εντροπία ορίζεται συναρτήσει της κατανομής της τυχαίας μεταβλητής X. Δεν εξαρτάται όμως από τις πραγματικές τιμές που παίρνει η μεταβλητή X, αλλά μόνο από τις πιθανότητες που έχουν οι τιμές αυτές.

Όπως μπορούμε να συνάγουμε και από τον ορισμό της εντροπίας, η ποσότητα πληροφορίας (ή αλλιώς το πληροφορικό περιεχόμενο) ενός γεγονότος x_i της τυχαίας μεταβλητής X είναι ίση με το αρνητικό του λογαρίθμου της πιθανότητας εμφάνισής που έχει, p_i , δηλαδή ίσο με $(-\log(p_i))$. Επομένως, η ποσότητα πληροφορίας ενός γεγονότος x_i είναι αντιστρόφως ανάλογη της πιθανότητας εμφάνισής του.

Οι ιδιότητες της μέσης (ποσότητας) πληροφορίας, που παράλληλα έχουν τεθεί και ως απαιτήσεις κατά τον ορισμό της, δηλαδή κατά την αναζήτηση από τον Shannon και άλλους ερευνητές της κατάλληλης συνάρτησης, διακρίνονται στις πέντε ακόλουθες:

(1) Το ποσό της πληροφορίας σε ένα γεγονός x εξαρτάται μόνον από την πιθανότητά του. Αυτή είναι μία φυσική απαίτηση, μιας και όσο πιο απίθανο είναι ένα γεγονός να πραγματοποιηθεί, τόσο περισσότερη πληροφορία περιέχει.

(2) Η μέση πληροφορία $H(X)$ είναι συνεχής ως προς το p.

(3) Η εντροπία είναι προσθετική. Η ιδιότητα αυτή αναφέρεται στην περίπτωση κατά την οποία δύο τυχαίες μεταβλητές X και Y, οι οποίες είναι ανεξάρτητες μεταξύ τους, συνδυάζονται. Τότε, για τη συνδυασμένη ποσότητα πληροφορίας των δύο μεταβλητών ισχύει $H(X,Y) = H(X) + H(Y)$.

(4) Η εντροπία $H(X)$ παίρνει τη μέγιστη τιμή της όταν όλα τα ενδεχόμενα είναι ισοπίθανα. Τότε, η αβεβαιότητα είναι η μέγιστη δυνατή και, κατά συνέπεια, η επιλογή ενός μηνύματος προσφέρει τη μέγιστη δυνατή μέση πληροφορία. Αντίθετα, η $H(X)$, γίνεται ελάχιστη, όταν ένα ενδεχόμενο έχει πιθανότητα ίση με 1.

(5) Η μέση πληροφορία $H(X)$ είναι συμμετρική, δηλαδή η διάταξη των πιθανοτήτων δεν την επηρεάζει. Έτσι, διαφορετικές τυχαίες μεταβλητές με κατανομές πιθανοτήτων που προέρχονται από μεταθέσεις της ίδιας κατανομής πιθανοτήτων έχουν ίση εντροπία. Σε ορισμένες περιπτώσεις, ακόμα και διαφορετικές κατανομές πιθανοτήτων οδηγούν στην ίδια μέση ποσότητα πληροφορίας.

Ένας σχετικός ορισμός με αυτό της εντροπίας, είναι αυτός της προσδοκίας (expectation). Η προσδοκία (αναμενόμενη τιμή) συμβολίζεται με E, και μετρά τη προσδοκώμενη τιμή της τυχαίας μεταβλητής $g(X)$, όταν η $X \sim p(x)$. Η τιμή της δίνεται από τον τύπο:

$$E_p g(x) = \sum_{x \in X} g(x) \cdot p(x) \quad , \qquad 2.2$$

και απλούστερα γράφεται και ως $E_g(X)$, όταν η συνάρτηση μάζας πιθανότητας εννοείται από τα συμφραζόμενα. Ιδιαίτερο ενδιαφέρον, παρουσιάζει επίσης η αυτοαναφορική αναμενόμενη τιμή της $g(X)$ ως προς την $p(x)$ όταν για αυτές τις δύο τις συνδέει η σχέση $g(X) = \log(\frac{1}{p(X)})$. Σε αυτήν την περίπτωση η εντροπία της X μπορεί επίσης να ερμηνευτεί ως η αναμενόμενη τιμή της τυχαίας μεταβλητής $\log(\frac{1}{p(X)})$, όπου η X λαμβάνεται σύμφωνα με τη συνάρτηση μάζας πιθανότητας $p(x)$. Συνεπώς,

$$H(X) = E_p \log(\frac{1}{p(X)}) \quad . \qquad 2.3$$

Αυτός ο ορισμός της εντροπίας συνδέεται με τον ορισμό της εντροπίας στη θερμοδυναμική.

Για τον ορισμό της εντροπίας πρέπει να καθοριστούν οι δύο ιδιότητες που είναι ικανές και αναγκαίες συνθήκες για να να είναι ορθός.

(1) Παίρνει πάντα θετικές τιμές, δηλαδή $H(X) \geq 0$.

(2) Πρέπει να ισχύει ότι $H_b(X) = \log_b(a) \cdot H_a(X)$.

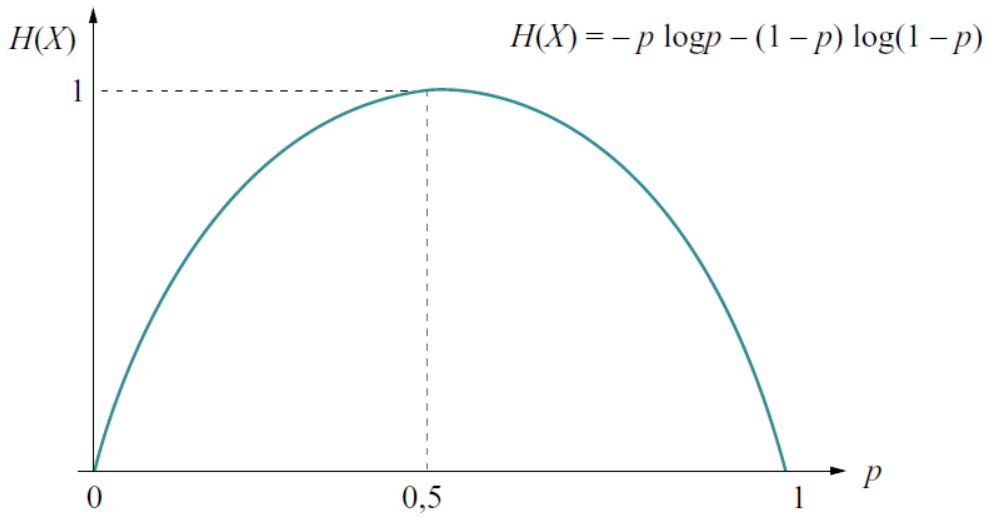
Η πρώτη ιδιότητα προκύπτει άμεσα από τους ορισμούς της πιθανότητας και του λογαρίθμου, βάση των οποίων πρέπει να ισχύει $0 \leq p(x) \leq 1$ και $\log(\frac{1}{p(x)}) \geq 0$, ενώ η

δεύτερη ιδιότητα αποδεικνύεται άμεσα από τη γνωστή σχέση που ισχύει στους λογαρίθμους, $\log_b(p)=\log_b(a)\cdot\log_a(p)$ και στην ουσιαστικά αυτό που λέει είναι ότι, είναι επιτρεπτό να αλλάξουμε τη βάση του λογαρίθμου για την εντροπία, απλά πολλαπλασιάζοντας τη σχέση με τον κατάλληλο συντελεστή.

Στην απλούστερη περίπτωση μιας διακριτής τυχαίας μεταβλητής X, η οποία έχει μόνο δύο ενδεχόμενα, (π.χ. εκπομπή ενός από δύο δυνατά μηνύματα) και οι πιθανότητες αυτών ισούται με p και $(1-p)$, αντίστοιχα, ο τύπος της εντροπίας θα είναι:

$$H(X)=-p\cdot\log(p)-(1-p)\cdot\log(1-p) \quad . \qquad 2.4$$

Στο Σχήμα 2.1 φαίνεται η γραφική παράσταση της συμπεριφοράς της μέσης ποσότητας πληροφορίας ως συνάρτηση της πιθανότητας p. (Η μονάδα μέτρησης της μέσης ποσότητας πληροφορίας είναι το bit, δηλαδή ο λογάριθμος είναι με βάση το 2.)



Σχήμα 2.1 Η μέση ποσότητα πληροφορίας ως συνάρτηση της p

Παρατηρούμε στη γραφική παράσταση (Σχήμα 2.1) ότι η μέση πληροφορία παίρνει τη μέγιστη τιμή, που ισούται με ένα, όταν τα δύο γεγονότα μπορούν να συμβούν με την ίδια πιθανότητα, δηλαδή $p=\frac{1}{2}$. Από την άλλη πλευρά, αν $p=1$ ή $p=0$, τότε η εντροπία είναι 0, αφού το τελικό αποτέλεσμα (η έκβαση του πειράματος) είναι βέβαιο.

Σε ένα άλλο λίγο πιο πολύπλοκο παράδειγμα όπου, έστω ότι έχουμε έναν αγώνα αυτοκινήτων, στον οποίο συμμετέχουν οκτώ αυτοκίνητα με πιθανότητες νίκης για το

καθένα από αυτά $(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64})$. Η εντροπία του αγώνα υπολογίζεται ως εξής:

$$H(X)=-\frac{1}{2}\cdot\log(\frac{1}{2})-\frac{1}{4}\cdot\log(\frac{1}{4})-\frac{1}{8}\cdot\log(\frac{1}{8})-\frac{1}{16}\cdot\log(\frac{1}{16})-\frac{1}{64}\cdot\log(\frac{1}{64})$$
$$-\frac{1}{64}\cdot\log(\frac{1}{64})-\frac{1}{64}\cdot\log(\frac{1}{64})-\frac{1}{64}\cdot\log(\frac{1}{64})=2\,\text{bits}\ .$$

2.4 Σχετική Εντροπία

Η σχετική εντροπία ή αλλιώς απόκλιση κατά Kullback–Leibler, είναι ένα μέτρο της απόστασης μεταξύ δύο κατανομών. Στη στατιστική, εμφανίζεται ως αναμενόμενος λογάριθμος του λόγου πιθανοφανειών. Η σχετική εντροπία $D(p||q)$ είναι ένα μέτρο του πόσο άστοχο είναι να θεωρήσουμε ότι η κατανομή είναι η q όταν η πραγματική κατανομή είναι η p. Για παράδειγμα, αν γνωρίζαμε την πραγματική κατανομή p της τυχαίας μεταβλητής, θα μπορούσαμε να κατασκευάσουμε έναν κώδικα με μέσο μήκος περιγραφής $H(p)$. Αν αντ' αυτού χρησιμοποιούσαμε τον κώδικα για μια κατανομή q, θα χρειαζόμασταν κατά μέσο όρο $H(p)+D(p||q)$ για να περιγράψουμε την τυχαία μεταβλητή.

Συγκεκριμένα, η σχετική εντροπία $D(p||q)$, μεταξύ δύο συναρτήσεων μάζας πιθανότητας $p(x)$ και $q(x)$ ισούται με:

$$\begin{aligned} D(p||q) &= \sum_{x\in X} p(x)\cdot\log(q(x))+H(X) \\ &= \sum_{x\in X} p(x)\cdot\log(\frac{p(x)}{q(x)}) \\ &= E_p\log(\frac{p(X)}{q(X)})\ . \end{aligned} \qquad 2.5$$

Η εντροπία του Shannon είναι μια τυχαία ειδική περίπτωση της σχετικής εντροπίας. Πράγματι η εντροπία του Shannon, μιας τυχαίας μεταβλητής, είναι η σχετική εντροπία αυτής ως προς μια κατάσταση που γνωρίζουμε με απόλυτη βεβαιότητα (μεταβλητή Y), δηλαδή $H(X)=H(X|Y)$, όπου $P(Y=y)=1$, για κάποια τιμή του Y.

Η σχετική εντροπία είναι πάντα μη αρνητική και ισούται με μηδέν αν και μόνο αν ισχύει

ότι $p=q$. Παρόλο που συγκαταλέγεται στη σχετική λίστα των μέτρων πληροφορίας, η σχετική εντροπία, δεν είναι ένα πραγματικό μέτρο. Αυτό γιατί δεν είναι συμμετρική, δηλαδή η διαφορά του p από το q δεν ισούται με τη διαφορά του q από το p και επιπλέον δεν ικανοποιεί την τριγωνική ανισότητα. Είναι χρήσιμο κάποιος να την αντιληφθεί καλύτερα σαν μια "απόσταση" μεταξύ δύο κατανομών παρόλο που δεν είναι μια πραγματική απόσταση.

Η σχετική εντροπία είναι μια έννοια πολύ μεγάλης σημασίας για την κλασσική στατιστική μηχανική του Gibbs και χρησιμοποιείται πολύ συχνά και στην Κβαντική Θεωρία πληροφορίας διότι πολλά σημαντικά αποτελέσματα της τελευταίας βασίζονται στη μονοτονία της. Είναι λοιπόν η κατάλληλη έκφραση της πληροφορίας, αφού η απροσδιοριστία μιας μεταβλητής, μετριέται πάντα σε σχέση με μια άλλη μεταβλητή.

2.5 Κοινή Εντροπία

Σε πολλές περιπτώσεις μας ενδιαφέρει να εξετάσουμε την ποσότητα πληροφορίας που περιέχει ένας συνδυασμός δύο τυχαίων μεταβλητών, δηλαδή ενός πειράματος το οποίο αποτελείται από δύο υποπειράματα. Έστω ένα τυχαίο πείραμα (X,Y) έχει ως δυνατά αποτελέσματα όλους τους δυνατούς συνδυασμούς των αποτελεσμάτων των δύο υποπειραμάτων του $X=\{x_1,x_2,..,x_n\}$ και $Y=\{y_1,y_2,..,y_m\}$, επομένως έχει το δειγματοχώρο:

$$(X,Y)=\{(x_1,y_1),(x_1,y_2),...,(x_1,y_m),...,(x_n,y_1),(x_n,y_2),...,(x_n,y_m)\} \quad . \qquad 2.6$$

Η κατανομή πιθανοτήτων δίνεται από:

$$P=\{p(x_1,y_1),p(x_1,y_2),...,p(x_1,y_m),...,p(x_n,y_1),p(x_n,y_2),...,p(x_n,y_m)\} \quad . \qquad 2.7$$

Αν (X,Y) είναι ένα τυχαίο πείραμα, το οποίο έχει δισδιάστατο δειγματοχώρο και κατανομή πιθανοτήτων αυτή που αναφέραμε παραπάνω, τότε η συνδυασμένη πληροφορία του $H(X,Y)$ ορίζεται ως η μέση τιμή:

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_i)\cdot\log(p(x_i,y_i)) \quad . \qquad 2.8$$

Η παραπάνω σχέση μπορεί να γραφτεί εναλλακτικά και μέσω της αναμενόμενης τιμής Ε ως:

$$H(X,Y) = -E_p \log(p(X,Y)) \quad . \tag{2.9}$$

Όταν οι κατανομές είναι ανεξάρτητες μεταξύ τους, τότε ισχύει ότι:

$$H(X,Y) = H(X) + H(Y) \quad . \tag{2.10}$$

Αφού οι δύο τυχαίες μεταβλητές είναι ανεξάρτητες, ισχύει ότι $p_{ij} = p(x_i) \cdot p(y_j)$ και έτσι έχουμε:

$$\begin{aligned}
H(X,Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \cdot \log(p_{ij}) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_i \cdot p_j \cdot \log(p_i \cdot p_j) \\
&= -\sum_{i=1}^{n} p_i \sum_{j=1}^{m} (\log(p_i) + \log(p_j)) \\
&= -\sum_{i=1}^{n} p_i \sum_{j=1}^{m} (p_j \cdot \log(p_i)) - \sum_{i=1}^{n} p_i \sum_{j=1}^{m} (p_j \cdot \log(p_j)) \\
&= -\sum_{i=1}^{n} p_i \cdot \log(p_i) \sum_{j=1}^{m} p_j - \sum_{i=1}^{n} p_i \sum_{j=1}^{m} (p_j \cdot \log(p_j)) \\
&= -\sum_{i=1}^{n} p_i \cdot \log(p_i) - \sum_{j=1}^{m} (p_j \cdot \log(p_j)) \\
&= H(X) + H(Y) \quad .
\end{aligned} \tag{2.11}$$

Ο ορισμός της μέσης ποσότητας πληροφορίας $H(X,Y)$ μπορεί να επεκταθεί και για περισσότερες από δύο διαστάσεις, έστω n, δηλαδή $H(X_1, \ldots, X_n)$. Σε κάθε περίπτωση λαμβάνουμε υπόψη όλους τους δυνατούς συνδυασμούς αποτελεσμάτων και, εφόσον γνωρίζουμε τις πιθανότητες αυτών, μπορούμε να υπολογίσουμε τη συνδυασμένη ποσότητα πληροφορίας. Άρα όταν οι μεταβλητές είναι παραπάνω από δύο, έστω n, δηλαδή έχουμε τις μεταβλητές $X_1, \ldots, X_n$, η κοινή εντροπία ισούται με:

$$\begin{aligned}
H(X_1, \ldots, X_n) &= -\sum_{X_1} \ldots \sum_{X_n} p(x_1, \ldots, x_n) \cdot \log(p(x_1, \ldots, x_n)) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \quad .
\end{aligned} \tag{2.12}$$

Αυτό αποδεικνύεται ως εξής:

$$
\begin{aligned}
H(X_1, X_2, .. X_n) &= -\sum_{x_1, x_2, .. x_n} p(x_1, x_2, .., x_n) \cdot \log(p(x_1, x_2, .., x_n)) \\
&= -\sum_{x_1, x_2, .. x_n} p(x_1, x_2, .., x_n) \cdot \log(\prod_{i=1}^{n} p(x_i | x_{i-1}, .., x_1)) \\
&= -\sum_{x_1, x_2, .. x_n} \sum_{i=1}^{n} p(x_1, x_2, .., x_n) \cdot \log(p(x_i | x_{i-1}, .., x_1)) \\
&= -\sum_{i=1}^{n} \sum_{x_1, x_2, .. x_n} p(x_1, x_2, .., x_n) \cdot \log(p(x_i | x_{i-1}, .., x_1)) \\
&= -\sum_{i=1}^{n} \sum_{x_1, x_2, .. x_i} p(x_1, x_2, .., x_i) \cdot \log(p(x_i | x_{i-1}, .., x_1)) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1}, .., X_1) \ .
\end{aligned}
\tag{2.13}
$$

όπου γράφουμε $p(x_1, .., x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1}, .., x_1)$. Όσον αφορά το δεύτερο μέρος του τύπου εφαρμόζουμε κατ' επανάληψη τον κανόνα για το ανάπτυγμα της εντροπίας για δύο μεταβλητές, οπότε έχουμε:

$$
H(X_1, X_2) = H(X_1) + H(X_2 | X_1) \tag{2.14}
$$

$$
\begin{aligned}
H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3 | X_1) \\
&= H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1)
\end{aligned}
\tag{2.15}
$$

$$
\vdots
$$

$$
\begin{aligned}
H(X_1, X_2, .., X_n) &= H(X_1) + H(X_2 | X_1) + \ldots + H(X_n | X_{n-1}, .., X_1) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots X_1) \ .
\end{aligned}
\tag{2.16}
$$

2.6 Δεσμευμένη Εντροπία

Σε αρκετές περιπτώσεις, αναλόγως το πρόβλημα, μας ενδιαφέρει να υπολογίσουμε την ποσότητα πληροφορίας μιας τυχαίας μεταβλητής, X, όταν είναι γνωστό το αποτέλεσμα μιας δεύτερης τυχαίας μεταβλητής, Y, θεωρώντας πάντα δεδομένο ότι για τις δύο μεταβλητές ισχύει $(X, Y) \sim p(x, y)$. Η ποσότητα αυτή, καλείται είτε δεσμευμένη ή υπό

συνθήκη ποσότητα πληροφορίας της μεταβλητής X ως προς την μεταβλητή Y και συμβολίζεται με $H(X|Y)$. Με άλλα λόγια η υπό συνθήκη εντροπία μετρά την αβεβαιότητα της τυχαίας μεταβλητής X όταν είναι γνωστή η Y μεταβλητή. Η αναλυτική της μορφή είναι:

$$\begin{aligned} H(X|Y) &= \sum_{y\in Y} p(y)\cdot H(X|Y=y) \\ &= -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\cdot \log(p(x|y)) \\ &= -\sum_{y\in Y}\sum_{x\in X} p(y,x)\cdot \log(p(x|y)) \\ &= -\sum_{y\in Y, x\in X} p(y,x)\cdot \log(p(x|y)) \\ &= -\sum_{y\in Y, x\in X} p(y,x)\cdot \log\left(\frac{p(y,x)}{p(y)}\right) \\ &= \sum_{y\in Y, x\in X} p(y,x)\cdot \log\left(\frac{p(y)}{p(y,x)}\right) \ . \end{aligned} \qquad 2.17$$

Εναλλακτικά γράφεται και ως:

$$\begin{aligned} H(X|Y) &= \sum_{y\in Y} p(y)\cdot H(X|Y=y) \\ &= -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\cdot \log(p(x|y)) \\ &= -\sum_{y\in Y}\sum_{x\in X} p(y,x)\cdot \log(p(x|y)) \\ &= -E_p \log(p(X|Y)) \ . \end{aligned} \qquad 2.18$$

Όπως θα προσδοκούσε κάποιος, στην περίπτωση όπου, η μία τυχαία μεταβλητή, X, καθορίζεται πλήρως μέσω της άλλης τυχαίας μεταβλητής Y, ισχύει ότι η υπό συνθήκη ποσότητα πληροφορίας της πρώτης μεταβλητής, X, ως προς την δεύτερη μεταβλητή, Y, ισούται με μηδέν, δηλαδή, $H(X|Y)=0$. Αντίστοιχα είναι πάλι αναμενόμενο, ότι στην περίπτωση που οι δυο τυχαίες μεταβλητές, X, Y είναι τελείως ανεξάρτητες μεταξύ τους, για την δεσμευμένη ποσότητα πληροφορίας να ισχύει ότι $H(X|Y)=H(X)$, δηλαδή να είναι ίση με την εντροπία της τυχαίας μεταβλητής X. Οι δύο ισότητες που περιγράφονται παραπάνω αποδεικνύονται σχεδόν άμεσα από τον τύπο της δεσμευμένης εντροπίας, αλλά είναι και διαισθητικά σωστές με βάση αυτό που θα περίμενε κανείς από τον ορισμό της.

Αναφέρεται σε αυτό το σημείο, ότι η φυσικότητα του ορισμού της από κοινού εντροπίας και της δεσμευμένης εντροπίας αποκαλύπτεται από το γεγονός ότι η εντροπία ενός

ζεύγους τυχαίων μεταβλητών, έστω X,Y , είναι η εντροπία της μιας μεταβλητής, Y, προστιθέμενη με τη δεσμευμένη εντροπία αυτής της μεταβλητής, Y, ως προς την μεταβλητή X. Δηλαδή:

$$H(Y,X)=H(Y)+H(X|Y) \quad , \qquad 2.19$$

και λόγω συμμετρίας ισχύει επίσης ότι:

$$H(X,Y)=H(X)+H(Y|X) \quad . \qquad 2.20$$

Αυτό αποδεικνύεται ακολούθως:

$$\begin{aligned} H(Y,X) &= -\sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(y,x)) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(y)\cdot p(x|y)) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(y)) - \sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(x|y)) \\ &= -\sum_{y\in Y} p(y)\cdot\log(p(y)) - \sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(x|y) \\ &= H(Y)+H(X|Y) \quad . \end{aligned} \qquad 2.21$$

Ισοδύναμα, μπορούμε να γράψουμε:

$$\log(p(Y,X))=\log(p(Y))+\log(p(X|Y)) \quad . \qquad 2.22$$

2.7 Αμοιβαία Πληροφορία

Η αμοιβαία πληροφορία $I(X;Y)$ είναι ένα μέγεθος που μετράει την ποσότητα της πληροφορίας που μια τυχαία μεταβλητή περιέχει για μια άλλη τυχαία μεταβλητή. Με άλλα λόγια, υπολογίζει σε ποιον βαθμό μπορεί η γνώση που σημειώνεται για την δεύτερη μεταβλητή να μειώσει την αβεβαιότητα που υπάρχει για την πρώτη μεταβλητή. Το μέτρο αυτό ουσιαστικά, βασίζεται στην αμοιβαία εξάρτηση που υπάρχει μεταξύ των δύο μεταβλητών. Πιο συγκεκριμένα, έστω ότι δίνονται δύο τυχαίες μεταβλητές X,Y , για τις οποίες η από κοινού σ.μ.π. $p(x,y)$ είναι γνωστή, καθώς επίσης είναι γνωστές και οι αντίστοιχες περιθώριες κατανομές πιθανότητας των X,Y , $p(x)$ και $p(y)$. Τότε η ποσότητα της αμοιβαίας πληροφορίας $I(X;Y)$ αυτών των δύο μεταβλητών θα είναι ίση με τη σχετική εντροπία μεταξύ της από κοινού σ.μ.π $p(x,y)$ και των περιθώριων κατανομών πιθανότητας $p(x)$ και $p(y)$, δηλαδή:

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right) \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log\left(\frac{p(x|y)}{p(x)}\right) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(p(x)) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(p(x|y)) \\
&= \sum_{x\in X} p(x)\cdot\log(p(x)) - \left(-\sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(p(x|y))\right) \\
&= H(X) - H(X|Y) \ . \qquad 2.23
\end{aligned}
$$

Λόγω συμμετρίας έπεται επίσης ότι:

$$I(X;Y) = H(Y) - H(Y|X) \ , \qquad 2.24$$

Άρα, όταν είναι γνωστή η μεταβλητή X, δίνει για την μεταβλητή Y τόση πληροφορία, όση δίνει αντίστοιχα και η μεταβλητή Y όταν είναι αυτή γνωστή, για την μεταβλητή X. Ένας διαφορετικός τρόπος γραφής της αμοιβαίας πληροφορίας είναι ο παρακάτω:

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right) \\
&= D(p(x,y)\|p(x)\cdot p(y)) \\
&= E_{p(X,Y)}\log\left(\frac{p(X,Y)}{p(X)\cdot p(Y)}\right) \ . \qquad 2.25
\end{aligned}
$$

Όπως αποδείχτηκε σε παραπάνω ενότητα (Ενότητα 2.6), για την κοινή εντροπία δύο τυχαίων μεταβλητών X,Y ισχύει ότι $H(X,Y) = H(X) + H(Y|X)$, επομένως μετά από πράξεις προκύπτει και η σχέση:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \ . \qquad 2.26$$

Τέλος παρατηρούμε ότι:

$$I(X;X) = H(X) - H(X|X) = H(X) \ . \qquad 2.27$$

Άρα προκύπτει ότι η αμοιβαία πληροφορία που περιέχει μια τυχαία μεταβλητή με τον ίδιο της τον εαυτό, ισούται στην πραγματικότητα με την ίδια την εντροπία της τυχαίας μεταβλητής. Αυτός είναι άλλωστε και ο λόγος, για τον οποίο μερικές φορές η εντροπία μιας τυχαίας μεταβλητής αποκαλείται και αυτοπληροφορία. Παρατηρούμε επίσης, πως η αμοιβαία ποσότητα πληροφορίας δύο ανεξάρτητων τυχαίων μεταβλητών X,Y είναι $I(X;Y) = 0$. Αυτό το συμπέρασμα προκύπτει άμεσα από το γεγονός ότι όταν οι μεταβλητές X,Y είναι ανεξάρτητες μεταξύ τους, τότε για την από κοινού σ.μ.π. ισχύει η

εξίσωση $p(x,y)=p(x)\cdot p(y)$ και άρα για τον λογάριθμο προκύπτει $\log(\frac{p(x,y)}{p(x)\cdot p(y)})=\log(1)=0$, οπότε θα έχουμε:

$$\begin{aligned} I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(\frac{p(x,y)}{p(x)\cdot p(y)}) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(1) \\ &= 0 \quad . \end{aligned} \tag{2.28}$$

Από την άλλη πλευρά, αν βλέπουμε ότι η X είναι μια ντετερμινιστική συνάρτηση της Y (ή το αντίθετο), δηλαδή $H(X|Y)=0$, τότε όλη την πληροφορία την οποία μεταφέρει το X, θα την μοιράζεται με το Y. Με λίγα λόγια γνωρίζοντας το X, θα μπορεί κάποιος να καθορίσει πλήρως το Y (και το αντίστροφο). Σε αυτή την περίπτωση, η αμοιβαία πληροφορία, θα είναι ίση με την αβεβαιότητα που περιέχει μόνη της η μεταβλητή Y (ή η μεταβλητή X αντίστοιχα). Επομένως θα είναι ισούται με την εντροπία της μεταβλητής Y ή αντίστοιχα την εντροπία της μεταβλητής X, δηλαδή $I(X;Y)=H(X)=H(Y)$.

Σε περίπτωση που οι τυχαίες μεταβλητές X,Y είναι συνεχείς τότε η $I(X;Y)$ θα ισούται με:

$$I(X;Y)=\iint_{Y,X} p(x,y)\cdot\log(\frac{p(x,y)}{p(x)\cdot p(y)})dxdy \quad , \tag{2.29}$$

όπου $p(x,y)$ είναι η σ.π.π. των τυχαίων μεταβλητών (X,Y) και $p(x)$, $p(y)$ οι περιθώριες κατανομές τους αντίστοιχα.

Τέλος δύο αναγκαίες και ικανές συνθήκες για την αμοιβαία πληροφορία είναι οι εξής:

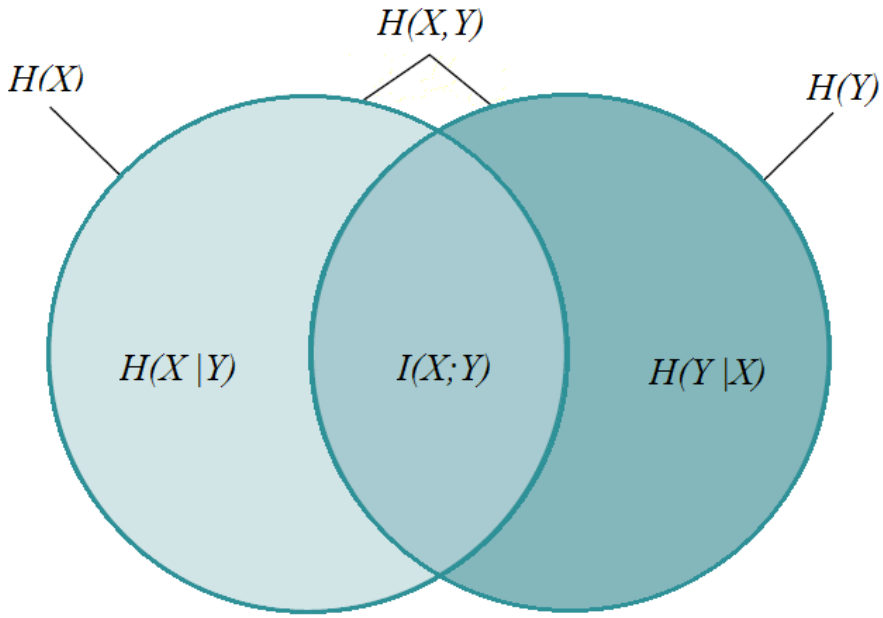
(1) Είναι πάντοτε θετική, δηλαδή $I(X;Y)\geq 0$.

(2) Είναι συμμετρική, δηλαδή ισχύει πάντα ότι $I(X;Y)=I(Y;X)$.

Η δεύτερη αναφέρθηκε και παραπάνω.

Η σχέση μεταξύ των μέτρων $H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, $H(Y|X)$ και $I(X;Y)$ που αναφέρθηκαν στα προηγούμενα κεφάλαια, μπορεί επίσης να αναπαρασταθεί μέσω ενός διαγράμματος Venn (Σχήμα 2.2), και παράλληλα να γίνει και καλύτερα κατανοητή. Παρατηρούμε ότι η αμοιβαία πληροφορία $I(X;Y)$ των δύο τυχαίων μεταβλητών X,Y , αντιστοιχεί στην τομή της πληροφορίας που περιέχει η τυχαία μεταβλητή X ($H(X)$) με

την πληροφορία που περιέχει η τυχαία μεταβλητή Y ($H(Y)$), ενώ η κοινή εντροπία αυτών των δύο ($H(X,Y)$) αντιστοιχεί στην ένωση τους. Τέλος η δεσμευμένη εντροπία κάθε τυχαίας μεταβλητής ως προς την άλλη μεταβλητή ($H(X|Y)$ και $H(Y|X)$) αντιστοιχεί στο αντίστοιχο σχετικό συμπλήρωμα κάθε πληροφορίας σε σχέση με την άλλη.



Σχήμα 2.2 Σχέσεις μεταξύ δύο μέτρων ποσότητας πληροφορίας

Η αμοιβαία πληροφορία ικανοποιεί τον παρακάτω κανόνα αλυσίδας:

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i-1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1) \quad . \qquad 2.30$$

Η απόδειξη του κανόνα είναι:

$$\begin{aligned} I(X_1, X_2, \ldots, X_n; Y) &= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y) \\ &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y) \\ &= \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \ldots X_{i-1}) \quad . \end{aligned} \qquad 2.31$$

2.8 Υπό Συνθήκη Αμοιβαία Πληροφορία

Η υπό συνθήκη αμοιβαία πληροφορία είναι επίσης ένα από τα πιο βασικά μέτρα στην θεωρία πληροφοριών και την συμβολίζουμε με $I_z(X;Y)$. Η ποσότητα αυτή μπορεί να ερμηνευτεί ως η μείωση της αβεβαιότητας της τυχαίας μεταβλητής X λόγω της γνώσης που υπάρχει για τη τυχαία μεταβλητή Y, με δεδομένο ότι έχει ήδη παρατηρηθεί μια τρίτη τυχαία μεταβλητή Z. Γενικά ορίζεται ως:

$$\begin{aligned} I_z(X;Y) &= I(X;Y|Z) = E_z(I(X;Y|Z)) \\ &= \sum_{z\in Z} p_z(z) \sum_{y\in Y}\sum_{x\in X} p_{X,Y|Z}(x,y|z)\cdot \log\left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)\cdot p_{Y|Z}(y|z)}\right) \\ &= \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} p_{X,Y,Z}(x,y,z)\cdot \log\left(\frac{p_z(z)\cdot p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)\cdot p_{Y,Z}(y,z)}\right) . \end{aligned} \qquad 2.32$$

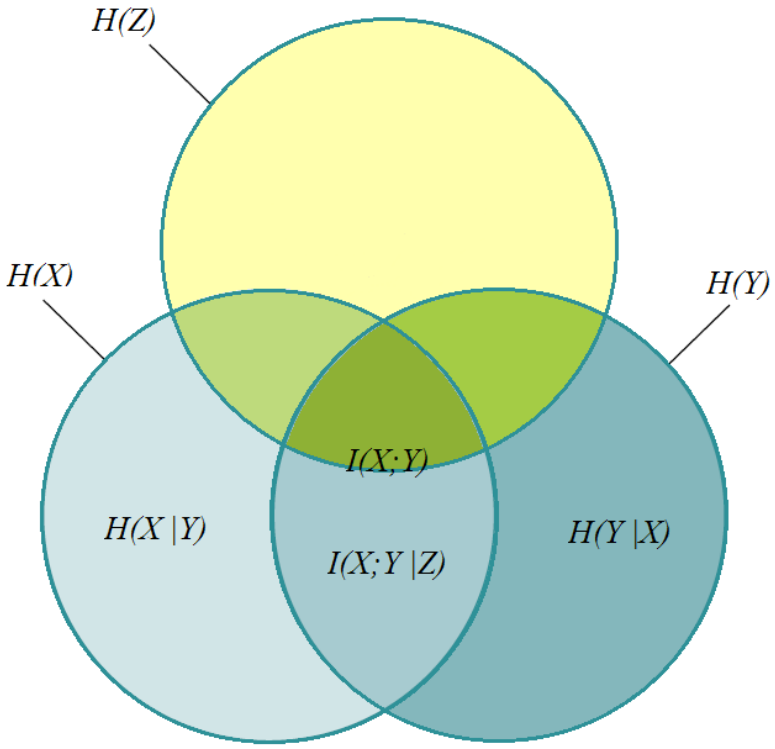
Εναλλακτικά, η υπό συνθήκη αμοιβαία πληροφορία μπορεί να εκφραστεί μέσω της προσδοκίας ως:

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} p_{X,Y,Z}(x,y,z)\cdot \log\left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)\cdot p_{Y|Z}(y|z)}\right) \\ &= E_{p(x,y,z)} \log\left(\frac{p(X,Y|Z)}{p(X|Z)\cdot p(Y|Z)}\right) . \end{aligned} \qquad 2.33$$

Τέλος, εύκολα αποδεικνύεται ότι μπορεί να γραφτεί και με την παρακάτω μορφή:

$$I(X;Y|Z) = H(X,Z)+H(Y,Z)-H(X,Y,Z)-H(Z) = H(X|Z)-H(X|Y,Z) . \qquad 2.34$$

Και σε αυτήν την περίπτωση, όπου έχουμε τρεις τυχαίες μεταβλητές X,Y,Z υπάρχει η δυνατότητα να αναπαραστήσουμε τις σχέσεις μεταξύ των διαφόρων μέτρων που τις περιγράφουν $H(X)$, $H(Y)$, $H(Z)$, $H(X|Y)$, $H(Y|X)$, $I(X;Y)$ και $I(X;Y|Z)$ συνολικά με ένα διάγραμμα Venn (Σχήμα 2.3). Ουσιαστικά επεκτείνουμε το Σχήμα 2.2 του προηγούμενου κεφαλαίου, που παρουσιάζει αυτές τις σχέσεις στην περίπτωση έχουμε μόνο δύο μεταβλητές, σε τρεις μεταβλητές. Αντίστοιχα εδώ παρατηρούμε ότι η υπό συνθήκη αμοιβαία πληροφορία $I_z(X;Y)$ αντιστοιχεί στην τομή της πληροφορίας που περιέχει η τυχαία μεταβλητή X με την πληροφορία που περιέχει η τυχαία μεταβλητή Y μεμονωμένα, μείον την τομή της πληροφορίας που περιέχουν και οι τρεις τυχαίες μεταβλητές X,Y,Z μαζί.



Σχήμα 2.3 Σχέσεις μεταξύ τριών μέτρων ποσότητας πληροφορίας

Για την υπό συνθήκη αμοιβαία πληροφορία, οι μεταβλητές X,Y,Z δεν είναι αναγκαίο να αντιπροσωπεύουν αποκλειστικά επιμέρους τυχαίες μεταβλητές αλλά θα μπορούσαν επίσης να αντιπροσωπεύουν την από κοινού κατανομή κάθε συνδυασμού τυχαίων μεταβλητών, οι οποίες όμως να ορίζονται στο ίδιο χώρο πιθανοτήτων. Με άλλα λόγια θα μπορούσε να ισχύει $I(X_1,X_2;Y_1,Y_2|Z_1;Z_2)$.

2.9 Υπό Συνθήκη Σχετική Εντροπία

Αν θεωρήσουμε ότι οι $p(x,y)$ και $q(x,y)$ είναι οι από κοινού συναρτήσεις μάζας πιθανότητας, τότε η δεσμευμένη σχετική εντροπία $D(p(y|x)\|q(y|x))$ αυτών είναι ο μέσος όρος των σχετικών εντροπιών μεταξύ των δύο δεσμευμένων συναρτήσεων μάζας πιθανότητας $p(y|x)$ και $q(y|x)$ υπολογισμένος ως προς τη συνάρτηση μάζας πιθανότητας $p(x)$. Συγκεκριμένα,

$$D(p(y|x)||q(y|x)) = \sum_x p(x) \sum_y p(y|x) \cdot \log\left(\frac{p(y|x)}{q(y|x)}\right) = E_{p(x,y)} \log\left(\frac{p(Y|X)}{q(Y|X)}\right) \quad . \qquad 2.35$$

Ο παραπάνω συμβολισμός για την από κοινού σχετική εντροπία δεν είναι πλήρης, διότι δεν αναφέρει την κατανομή $p(x)$ της δεσμεύουσας τυχαίας μεταβλητής. Συνήθως όμως εννοείται από τα συμφραζόμενα.

Η σχετική εντροπία, μεταξύ δύο από κοινού κατανομών ενός ζεύγους τυχαίων μεταβλητών μπορεί να εκφραστεί ως άθροισμα μιας σχετικής εντροπίας και μιας δεσμευμένης σχετικής εντροπίας. Αυτός θεωρείται και ο κανόνας αλυσίδας για την σχετική εντροπία. Ο τύπος είναι:

$$D(p(y,x)||q(y,x)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \quad . \qquad 2.36$$

Και η αντίστοιχη απόδειξη:

$$\begin{aligned} D(p(y,x)||q(y,x)) &= \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(x,y)}{q(x,y)}\right) \\ &= \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(x) \cdot p(y|x)}{q(x) \cdot q(y|x)}\right) \\ &= \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(x)}{q(x)}\right) + \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(y|x)}{q(y|x)}\right) \\ &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \quad . \end{aligned} \qquad 2.37$$

Κεφάλαιο 3:

Επιλογή Χαρακτηριστικών

3.1 Εισαγωγή

Καθώς τα τεχνολογικά μέσα εξελίσσονται με την πάροδο του χρόνου, η απόκτηση πληροφορίας γίνεται συνεχώς ευκολότερη. Αποτέλεσμα αυτού του γεγονότος είναι τα δεδομένα να καταγράφονται ολοένα και με αυξανόμενο ρυθμό. Αυτός ο τεράστιος όγκος δεδομένων που παράγεται επηρεάζει αρνητικά την προσπάθειά μας να μελετήσουμε ένα πρόβλημα, να εξάγουμε συμπεράσματα και τελικά να καταφέρουμε να το λύσουμε. Ο λόγος είναι ότι μεγάλο μέρος των δεδομένων που έχουμε στην διάθεσή μας, μπορεί να αφορά άχρηστες πληροφορίες, με την έννοια ότι μπορεί να είναι άσχετες με το υπό μελέτη πρόβλημα.

Σε ένα, σχετικό και με το αντικείμενο την παρούσας εργασίας, παράδειγμα που θα βοηθήσει τον αναγνώστη να καταλάβει το ζητούμενο, θα μπορούσαμε να υποθέσουμε ότι έχουμε έναν αλγόριθμο που θέλει να ταξινομήσει εικόνες σε διάφορες κατηγορίες αναλόγως με το τι αναπαριστούν. Κάποια από τα χαρακτηριστικά που θα του δώσουμε για τις εικόνες μπορεί να είναι το χρώμα της εικόνας που πιάνει το μεγαλύτερο μέρος της, το πόσες διαφορετικές ομάδες χρωμάτων έχει, τις διαστάσεις της, το πότε δημιουργήθηκε, το αν είναι φωτογραφία ή προϊόν επεξεργασίας, το που τραβήχτηκε, ή αν το σχέδιο της μοιάζει με κάποιο από την βάση δεδομένων. Κάποια από αυτά τα χαρακτηριστικά όπως τα δύο πρώτα και το τελευταίο μπορεί να είναι χρήσιμα για να πούμε τι απεικονίζει, και κάποια άλλα όπως οι διαστάσεις ή η ημερομηνία δημιουργίας της να είναι άχρηστα σε σχέση με το πρόβλημα που μελετάται, και τελικά να αγνοηθούν αφού όμως έχει σπαταληθεί χρόνος στην εξέταση της χρησιμότητάς τους.

Σε πολλά προβλήματα, λοιπόν, του πραγματικού κόσμου όπως σε εφαρμογές αναγνώρισης προτύπων, στατιστικής ανάλυσης προσδιορισμού συναρτήσεων από πεπερασμένα σύνολα δεδομένων ακόμη και σε ιατρικές εφαρμογές όπου το ζητούμενο είναι αν κάποιος ασθενής πάσχει από μια συγκεκριμένη ασθένεια κ.ά., προκύπτει η αναγκαιότητα διαχείρισης και επεξεργασίας τεράστιας ποσότητας πληροφορίας αποδοτικά και γρήγορα. Αυτό σημαίνει ότι, πριν προβούμε σε οποιουδήποτε είδους εργασία πάνω στον ακατέργαστο όγκο μετρήσεων που ενδεχομένως διαθέτουμε, πρέπει να προσδιορίσουμε με όσο το δυνατόν βέλτιστο τρόπο τις μεταβλητές που περιγράφουν με επάρκεια το πρόβλημα.

Οι τεχνικές που συνθέτουν το πεδίο ελάττωσης διαστάσεων μπορούν να καταταχθούν σε διάφορες κατηγορίες αναλόγως ποιο κριτήριο λαμβάνεται υπόψιν. Για παράδειγμα με βάση το πλήθος των μεταβλητών που πρόκειται να μειωθούν χωρίζονται σε “αυστηρή ελάττωση” (hard reduction) κατά την οποία έχουμε δραστική μείωση των αρχικών διαστάσεων, “χαλαρή ελάττωση” (soft reduction) που αφορά σε ήπια μείωση της αρχικής διάστασης και τέλος σε οπτικοποίηση (visualization) κατά την οποία τα δεδομένα μέτριων διαστάσεων προβάλλονται σε δύο ή τρεις διαστάσεις. Ένας άλλος τρόπος διαχωρισμού είναι αν θεωρείται ο χρόνος σαν μεταβλητή ή όχι, όπου έχουμε τις στατικές (static) και τις χρονικά εξαρτημένες (time-dependent) τεχνικές. Ακόμα ανάλογα το είδος των δεδομένων μπορούν να διακριθούν σε συνεχείς και διακριτές.

Ωστόσο, οι τεχνικές ελάττωσης διαστάσεων συνηθίζεται να ταξινομούνται σύμφωνα με τη γενική λειτουργία που επιτελούν διότι είναι πιο χρήσιμο. Σύμφωνα με αυτό το κριτήριο, διακρίνονται στις τεχνικές εξαγωγής χαρακτηριστικών (feature extraction) που στόχος του είναι η δημιουργία νέων χαρακτηριστικών μέσω μετασχηματισμών του αρχικού χώρου των ακατέργαστων δεδομένων, και σε αυτές της επιλογής χαρακτηριστικών (feature selection) που επιλέγουν τα πιο αντιπροσωπευτικά χαρακτηριστικά από τα υπάρχοντα στον αρχικό χώρο.

Η κύρια ιδέα της επιλογής χαρακτηριστικών είναι η επιλογή ενός υποσυνόλου μεταβλητών εξαλείφοντας εκείνα τα στοιχεία με μικρή ή περιττή πληροφορία. Με αυτόν τον τρόπο μειώνεται η διάσταση των δεδομένων, αφού τα περιττά διανύσματα δεν συμμετέχουν στις περαιτέρω διαδικασίες. Πιο συγκεκριμένα θα μπορούσαμε να δώσουμε τον παρακάτω ορισμό:

Δεδομένου ενός D-διάστατου συνόλου N δειγμάτων $\{\mathbf{x}_i\}_{i=1}^{N}$ με άγνωστη κατανομή πιθανότητας, στόχος είναι να προσδιοριστεί μια νέα d-διάστατη απεικόνιση ώστε να διατηρείται η βασική δομή των δεδομένων με το ελάχιστο δυνατό σφάλμα αναπαράστασης.

Συνοπτικά η σημασία των διαδικασιών επιλογής χαρακτηριστικών αναλύεται στις ακόλουθες διαστάσεις (Kira and Rendell, 1992):

(1) Πιθανή μείωση του θορύβου στα δεδομένα, η οποία οφείλεται στην ύπαρξη χαρακτηριστικών που δεν παρέχουν αξιόπιστη πληροφορία.

(2) Περιορισμός του υπολογιστικού φόρτου που απαιτείται για την υλοποίηση της ανάλυσης και την ανάπτυξη βέλτιστων υποδειγμάτων.

(3) Απλοποίηση των αναπτυσσόμενων υποδειγμάτων, καθώς υποδείγματα που εξετάζουν περιορισμένη πληροφορία έχουν πιο απλή μορφή και συνεπώς μπορούν να ερμηνευτούν πιο εύκολα.

(4) Μείωση του χρόνου και του κόστους της χρήσης των υποδειγμάτων, καθώς περιορίζεται η ποσότητα της πληροφορίας που πρέπει να είναι διαθέσιμη για τη χρήση τους.

Από την άλλη, η δυσκολία του προβλήματος της επιλογής ενός υποσυνόλου χαρακτηριστικών οφείλεται σε δύο κυρίως λόγους. Πρώτον, το πλήθος υποσυνόλων που θα μπορούσαν να επιλεγούν αυξάνεται εκθετικά σε σχέση με τον αριθμό των χαρακτηριστικών του αρχικού συνόλου. Αν υποθέσουμε ότι δίνεται ένα κριτήριο αξιολόγησης υποσυνόλων, η εύρεση του καλύτερου υποσυνόλου ως προς αυτό δεν είναι υπολογιστικά εφικτή. Δεύτερον, η ποιότητα ενός υποσυνόλου εξαρτάται από πολλούς παράγοντες και έτσι δεν μπορεί να οριστεί εύκολα ένα αντικειμενικό κριτήριο αξιολόγησης. Με άλλα λόγια δεν υπάρχει τρόπος να αποτιμηθεί με ακρίβεια η ποιότητα ενός υποσυνόλου, αντίθετα στην πράξη με χρήση ευρετικών (heuristic) τεχνικών επιλέγεται τελικά ένα υποσύνολο που αναμένεται ότι θα οδηγήσει σε καλή απόδοση τον αλγόριθμο μάθησης που θα το χρησιμοποιήσει.

Η τεχνική της επιλογής χαρακτηριστικών μπορεί να χρησιμοποιηθεί είτε σε προβλήματα ομαδοποίησης, είτε σε προβλήματα ταξινόμησης. Όσον αφορά την ταξινόμηση, ζητούμενο είναι ο εντοπισμός εκείνων των χαρακτηριστικών που είναι απαραίτητα για τη

σωστή κατάταξη των αντικειμένων σε κατηγορίες. Απώτερος στόχος είναι να χρησιμοποιηθεί μόνο αυτό το υποσύνολο χρήσιμων χαρακτηριστικών κατά τη διαδικασία της ταξινόμησης και να αγνοηθούν τα υπόλοιπα χαρακτηριστικά, άσχετα ή περιττά. Θεωρητικά, εφόσον αγνοηθούν μόνο άσχετα ή περιττά χαρακτηριστικά, η απόδοση του ταξινομητή που θα κατασκευαστεί δεν θα είναι χειρότερη από την απόδοση που θα επιτυγχανόταν αν χρησιμοποιούνταν όλα τα χαρακτηριστικά. Πολλές φορές στην πράξη, η απόδοση όχι μόνο δεν χειροτερεύει αλλά βελτιώνεται σημαντικά γιατί χάρη στη μείωση της διάστασης αμβλύνεται το φαινόμενο της υπερπροσαρμογής (overfitting). Υπενθυμίζεται ότι υπερπροσαρμογή εμφανίζεται όταν μια απεικόνιση προσαρμόζεται πολύ καλά στα δεδομένα εκπαίδευσης, αλλά έχει κακή ικανότητα γενίκευσης, με αποτέλεσμα να επηρεάζεται από τον τυχαίο θόρυβο που υπάρχει σε αυτά.

Στη διεθνή βιβλιογραφία έχουν προταθεί διάφορες μεθοδολογίες στο πρόβλημα της επιλογής χαρακτηριστικών. Πολλές από τις μεθοδολογίες αυτές είναι άρρηκτα συνδεδεμένες με συγκεκριμένες τεχνικές ταξινόμησης, ενώ άλλες είναι γενικοί αλγόριθμοι οι οποίοι μπορούν να εφαρμοστούν ανεξαρτήτως του τρόπου που υλοποιείται η ταξινόμηση ενός υποδείγματος. Στις μέχρι σήμερα έρευνες, η αποτελεσματικότητα των μεθοδολογιών αυτών έχει ελεγχθεί για περιορισμένο αριθμό τεχνικών ταξινόμησης και σε περιορισμένα σύνολα δεδομένων. Τα στοιχεία αυτά καθιστούν δύσκολη την εξαγωγή ασφαλών συμπερασμάτων σχετικά με την πραγματική αποτελεσματικότητα των μεθοδολογιών επιλογής χαρακτηριστικών.

Γι'αυτό παρά τη σημαντική έρευνα που έχει πραγματοποιηθεί στο χώρο αυτό δεν υπάρχει μια ολοκληρωμένη έρευνα σχετικά με την αποτελεσματικότητα των προτεινόμενων μεθοδολογιών και αλγορίθμων. Μια τέτοια έρευνα πρέπει να λάβει υπόψη τις αλληλεπιδράσεις μεταξύ των διαδικασιών επιλογής χαρακτηριστικών και των τεχνικών που χρησιμοποιούνται για την ανάπτυξη υποδειγμάτων προς ταξινόμηση. Οι μέχρι σήμερα έρευνες περιορίζονται σε συγκεκριμένες τεχνικές ταξινόμησης. Δεδομένου, όμως, του πλήθους των διαφορετικών τεχνικών ταξινόμησης που είναι έως τώρα διαθέσιμες, είναι εμφανές ότι απαιτείται μια πιο ολοκληρωμένη ανάλυση.

Μπορεί κάποιος να φανταστεί έναν αλγόριθμο επιλογής χαρακτηριστικών σαν μια εξειδικευμένη μηχανή αναζήτησης, η οποία προτείνει βέλτιστα υποσύνολα. Αν και θεωρητικά η απλούστερη μέθοδος που μπορεί να εφαρμόσει κάποιος, είναι να

δοκιμάσει κάθε δυνατό υποσύνολο χαρακτηριστικών, προκειμένου να καταλήξει στο βέλτιστο, δηλαδή εκείνο με το μικρότερο σφάλμα, αυτό στην πράξη είναι απίστευτα εξαντλητικό και χρονοβόρο, χωρίς να δίνει πάντα τα επιθυμητά αποτελέσματα.

Γι'αυτό ακριβώς το λόγο επιλέγονται συνήθως διάφορες άλλες μέθοδοι. Μια τυπική διαδικασία επιλογής χαρακτηριστικών, αποτελείται από δύο φάσεις. Η πρώτη είναι η επιλογή των χαρακτηριστικών και η δεύτερη η αξιολόγησή τους και περιλαμβάνει τα ακόλουθα βήματα: Αρχικά, δημιουργία ενός υποψηφίου σετ που περιέχει ένα υποσύνολο από τα αρχικά χαρακτηριστικά μέσω ορισμένων στρατηγικών. Ακολούθως αξιολόγηση του υποψήφιου συνόλου και εκτίμηση της χρησιμότητας των χαρακτηριστικών στο σύνολο αυτό. Με βάση αυτή την αξιολόγηση, ορισμένα χαρακτηριστικά στο σετ μπορεί να απορριφθούν, ενώ κάποια άλλα μπορεί να προστεθούν. Τέλος, χρησιμοποιούνται ορισμένα κριτήρια διακοπής, προκειμένου να καθοριστεί εάν το τρέχον σύνολο των επιλεγμένων χαρακτηριστικών, είναι αρκετά καλό ή όχι.

Κάθε αλγόριθμος επιλογής χαρακτηριστικών ακολουθεί μια συγκεκριμένη στρατηγική προκειμένου να διερευνήσει το σύνολο των χαρακτηριστικών που του δίνεται ως είσοδος. Η στρατηγική διερεύνησης στοχεύει στον προσδιορισμό κατάλληλων συντελεστών στάθμισης $w_1, w_2, \ldots, w_m$ για τα m χαρακτηριστικά που του δίνονται ανάλογα με την αναμενόμενη συμβολή που θα έχουν στην ανάπτυξη ενός αξιόπιστου υποδείγματος ταξινόμησης. Οι συντελεστές στάθμισης μπορεί να είναι είτε πραγματικοί αριθμοί στο διάστημα $[0,1]$ αναπαριστώντας τη σημαντικότητα του κάθε χαρακτηριστικού ή να έχουν δυαδική μορφή $\{0,1\}$ ανάλογα με το εάν ένα χαρακτηριστικό επιλέγεται $(w_j=1)$ ή όχι $(w_j=0)$.

Οι πιο γνωστές μέθοδοι επιλογής υποσυνόλου χαρακτηριστικών είναι οι μέθοδοι φίλτρων (filter), οι μέθοδοι περιτυλίγματος (wrapper), και οι ενσωματωμένες μέθοδοι (embedded). Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στην πρώτη κατηγορία εφαρμόζονται πριν τη χρησιμοποίηση κάποιας τεχνικής ταξινόμησης και συνεπώς δεν επηρεάζονται από αυτή. Ουσιαστικά, οι αλγόριθμοι αυτής της κατηγορίας λειτουργούν ως φίλτρα για την απαλοιφή των μη σχετικών ή πλεοναστικών χαρακτηριστικών. Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στη δεύτερη κατηγορία χρησιμοποιούν τη μέθοδο ταξινόμησης ως μέρος της διαδικασίας

(John G.H., 1994). Ειδικότερα, βασιζόμενοι σε εμπρόσθιες, ανάστροφες ή τυχαίες διαδικασίες, οι αλγόριθμοι της κατηγορίας αυτής χρησιμοποιούν τη μέθοδο ταξινόμησης για την αξιολόγηση της αποτελεσματικότητας του συνόλου των χαρακτηριστικών που επιλέγονται. Τέλος, στην τρίτη κατηγορία περιλαμβάνονται αλγόριθμοι και τεχνικές, η εφαρμογή των οποίων είναι άμεσα συνδεδεμένη με μια συγκεκριμένη τεχνική ταξινόμησης.

3.2 Μέθοδοι Φίλτρου (filter)

Στην κατηγορία των φίλτρων ανήκουν όσοι αλγόριθμοι βασίζονται στην έννοια της συνάφειας μεταξύ χαρακτηριστικών και κλάσης, δηλαδή δεν βασίζονται σε κάποιο ταξινομητή προκειμένου να εκτιμήσουν την ποιότητα ενός υποσυνόλου χαρακτηριστικών, ενώ αντίθετα, με την χρήση στατιστικών μέτρων προσπαθούν να εντοπίσουν συναφή χαρακτηριστικά. Ουσιαστικά αξιολογούν τη σχετικότητα των χαρακτηριστικών ερευνώντας μόνο τις ιδιότητες των στοιχείων. Η πιο κοινή διαδικασία είναι βασισμένη στον υπολογισμό της σχετικότητας χαρακτηριστικών γνωρισμάτων. Τα χαρακτηριστικά γνωρίσματα με τη χαμηλή σχετικότητα αφαιρούνται. Τα υπόλοιπα χαρακτηριστικά γνωρίσματα χρησιμοποιούνται ως εισαγωγή στον αλγόριθμο ταξινόμησης.

Οι τεχνικές φίλτρων έχουν το πλεονέκτημα ότι είναι απλές και γρήγορες, είναι εφικτές από άποψη υπολογιστικής πολυπλοκότητας που εφαρμόζεται στα δεδομένα υψηλής διάστασης (όπως δεδομένα DNA, κείμενο κτλ) και επίσης είναι ανεξάρτητες από τον αλγόριθμο ταξινόμησης. Το τελευταίο πλεονέκτημα είναι πολύ σημαντικό, δεδομένου ότι κάποιος μπορεί να εφαρμόσει μια μέθοδο φίλτρων για να παράξει ένα βέλτιστο υποσύνολο χαρακτηριστικών γνωρισμάτων που μπορεί να αξιολογηθεί χρησιμοποιώντας διαφορετικούς ταξινομητές (classifiers).

Το κύριο μειονέκτημα των αλγορίθμων αυτών είναι ότι αγνοούν την αλληλεπίδραση που πιθανόν υπάρχει μεταξύ του συνόλου των χαρακτηριστικών που επιλέγεται και της τεχνικής ταξινόμησης που χρησιμοποιείται για την ανάπτυξη του υποδείγματος ταξινόμησης αφού κατά την διαδικασία της επιλογής των χαρακτηριστικών δεν υπάρχει

οποιαδήποτε αλληλεπίδραση με τον ταξινομητή. Αυτό σημαίνει ότι κάθε χαρακτηριστικό γνώρισμα ελέγχεται χωριστά και κατά συνέπεια οι εξαρτήσεις χαρακτηριστικών γνωρισμάτων αγνοούνται. Το γεγονός αυτό μπορεί να οδηγήσει στη χαμηλότερη απόδοση ταξινόμησης σε σύγκριση με άλλες τεχνικές επιλογής χαρακτηριστικών γνωρισμάτων. Σε μια προσπάθεια να ελεγχθούν οι εξαρτήσεις χαρακτηριστικών γνωρισμάτων πολλών μεταβλητών οι τεχνικές φίλτρων έχουν βελτιωθεί, στοχεύοντας στην ενσωμάτωση των εξαρτήσεων χαρακτηριστικών γνωρισμάτων μέχρι ενός ορισμένου βαθμού.

Οι μέθοδοι φίλτρου διακρίνονται σε δύο βασικές κατηγορίες, τις μονοπαραγοντικές μεθόδους (univariate) και τις πολυπαραγοντικές μεθόδους (multivariate).

Οι μονοπαραγοντικές (univariate) μέθοδοι, πρώτα αξιολογούν μεμονωμένα κάθε χαρακτηριστικό, με βάση τη συσχέτιση του με τις κλάσεις. Όσο μεγαλύτερη συσχέτιση υπάρχει, τόσο πιο χρήσιμο θεωρείται το χαρακτηριστικό. Ύστερα επιλέγονται τα k πιο συσχετισμένα χαρακτηριστικά, όπου το k καθορίζεται ανάλογα με την περίσταση. Η κυριότερη αδυναμία των μονοπαραγοντικών μεθόδων, είναι η εμφάνιση φαινομένων πλεονασμού, δηλαδή περιπτώσεις όπου επιλέγονται περιττά χαρακτηριστικά, με την έννοια ότι είναι όμοια μεταξύ τους και έτσι ο συνδυασμός τους δεν προσφέρει πολύ περισσότερη πληροφορία για την κατηγορία από αυτή που θα προσέφερε κάθε χαρακτηριστικό από μόνο του. Αυτό συμβαίνει κατά κύριο λόγο επειδή κάθε χαρακτηριστικό αξιολογείται ξεχωριστά, χωρίς να λαμβάνονται υπόψη τα άλλα που έχουν ήδη επιλεγεί.

Μερικά από τα κριτήρια που έχουν χρησιμοποιηθεί για τη μέτρηση της συσχέτισης είναι το κριτήριο του Fischer το οποίο μπορεί να χρησιμοποιηθεί σε προβλήματα δύο κατηγοριών, το F-test που μπορεί να χρησιμοποιηθεί για προβλήματα με K κατηγορίες και τέλος η μέτρηση της αμοιβαίας τους πληροφορίας $I(X;Y)$, η οποία μπορεί να ανιχνεύσει και τις γραμμικές εξαρτήσεις μεταξύ των μεταβλητών.

Το κριτήριο του Fischer για την συσχέτιση του i-οστού χαρακτηριστικού υπολογίζεται από τον τύπο

$$w_i=\frac{(\mu_{i1}-\mu_{i2})^2}{\sigma_{i1}^2+\sigma_{i2}^2} \quad , \qquad 3.1$$

όπου το μ_{i1} και το μ_{i2} στον αριθμητή αναφέρονται στη μέση τιμή του i-οστού χαρακτηριστικού για τα παραδείγματα της πρώτης και δεύτερης κατηγορίας αντίστοιχα, και κατά τα ίδια τα σ_{i1} και σ_{i2} στον παρονομαστή αναφέρονται στις τυπικές αποκλίσεις του i-οστου χαρακτηριστικού για τα παραδείγματα της πρώτης και δεύτερης κατηγορίας αντίστοιχα. Όταν έχουμε μεγάλη τιμή του βάρους w_i σημαίνει ότι τα παραδείγματα της πρώτης κατηγορίας διαφέρουν από τα παραδείγματα της δεύτερης ως προς το χαρακτηριστικό i σε σημαντικό βαθμό και επομένως το χαρακτηριστικό έχει ισχυρή συσχέτιση.

Το F-test υπολογίζεται από τον τύπο

$$w_j = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_j)^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2} \quad , \qquad 3.2$$

όπου $\bar{x}_{kj}$ είναι η μέση τιμή του χαρακτηριστικού j μόνο για τα παραδείγματα της κατηγορίας c_k , ενώ $\bar{x}_j$ είναι η μέση τιμή του χαρακτηριστικού j υπολογισμένη με βάση όλα τα παραδείγματα. Η έκφραση I(A) ισούται με ένα αν η πρόταση A είναι αληθής, διαφορετικά ισούται με μηδέν. Ο αριθμητής αυξάνεται όταν οι μέσες τιμές του χαρακτηριστικού j $\bar{x}_{kj}$ διαφέρουν μεταξύ τους για διαφορετικά k και άρα διαφέρουν και από τη μέση τιμή $\bar{x}_j$ του χαρακτηριστικού. Ο παρονομαστής μειώνεται όταν υπάρχει μικρή διακύμανση μεταξύ παραδειγμάτων της ίδιας κατηγορίας ως προς το χαρακτηριστικό j, δηλαδή τα x_{ij} έχουν πολύ κοντινές τιμές με το $\bar{x}_{kj}$.

Το μέτρο της αμοιβαίας πληροφορίας αναλύθηκε εκτενώς στο Κεφάλαιο 2 οπότε δεν θα αναφερθεί πάλι. Ο αναγνώστης καλείται να ανατρέξει στο Κεφάλαιο 2 σε περίπτωση που θέλει να το επαναφέρει στη μνήμη του.

Οι πολυπαραγοντικές (multivariate) μέθοδοι, σε αντίθεση με τις μονοπαραγοντικές, αξιολογούν τα χαρακτηριστικά λαμβάνοντας υπόψιν την παρουσία και των άλλων χαρακτηριστικών, προσπαθώντας έτσι να αποφύγουν την επιλογή περιττών χαρακτηριστικών και τα φαινόμενα πλεονασμού που έχει η πρώτη κατηγορία. Ουσιαστικά αυτό που κάνουν είναι να φτιάχνουν ένα βέλτιστο υποσύνολο επιλέγοντας χαρακτηριστικά που έχουν μεγάλη συσχέτιση σε σχέση με την κλάση (όπως και οι

μονοπαραγοντικές μέθοδοι) και παράλληλα ελέγχουν τα χαρακτηριστικά που επιλέγονται να είναι όσο το δυνατόν πιο ανόμοια μεταξύ τους. Τα υποσύνολα δηλαδή αξιολογούνται με βάση την περιεχόμενη πληροφορία, και τη σχετική ανεξαρτησία που έχουν μεταξύ τους.

3.3 Μέθοδοι Περιτυλίγματος (wrapper)

Ενώ οι μέθοδοι φίλτρων επιλέγουν χαρακτηριστικά γνωρίσματα βασισμένα στα κριτήρια επιλογής χωρίς χρησιμοποίηση ταξινομητή, οι μέθοδοι περιτυλίγματος ενσωματώνουν τους ταξινομητές (classifiers) μέσα στην αναζήτηση υποσυνόλων χαρακτηριστικών γνωρισμάτων. Στην κατηγορία αυτών των μεθόδων εντάσσονται όλοι οι αλγόριθμοι επιλογής χαρακτηριστικών που χρησιμοποιούν την ακρίβεια ταξινόμησης ως κριτήριο αξιολόγησης των διάφορων υποσυνόλων των χαρακτηριστικών.

Επιπλέον, στις μεθόδους περιτυλίγματος τα χαρακτηριστικά γνωρίσματα συνήθως αξιολογούνται σε ομάδες και όχι χωριστά. Όλα τα ήδη διαθέσιμα χαρακτηριστικά γνωρίσματα χρησιμοποιούνται για να παραγάγουν τα υποσύνολα χαρακτηριστικών γνωρισμάτων που αξιολογούνται. Για να βρεθεί το διάστημα όλων των υποσυνόλων χαρακτηριστικών γνωρισμάτων, ένας αλγόριθμος αναζήτησης είναι συνέχεια “τυλιγμένος” (wrapped) γύρω από το πρότυπο ταξινόμησης.

Εντούτοις, καθώς το διάστημα του υποσυνόλου χαρακτηριστικών γνωρισμάτων αυξάνεται εκθετικά με τον αριθμό χαρακτηριστικών γνωρισμάτων, χρησιμοποιούνται ευρετικές (heuristic) μέθοδοι αναζήτησης για να καθοδηγήσουν την αναζήτηση ενός βέλτιστου υποσυνόλου. Αυτές οι μέθοδοι αναζήτησης μπορούν να ταξινομηθούν στους αιτιοκρατικούς (deterministic) και τυχαίους (random) αλγορίθμους αναζήτησης. Πιο συγκεκριμένα στην πρώτη κατηγορία ανήκουν οι εμπρόσθιες και οι ανάστροφες διαδικασίες οι οποίες είναι και οι πιο συνηθισμένες, ενώ στην δεύτερη οι τυχαίες διαδικασίες.

Η διαδικασία που ακολουθεί μια τυπική μέθοδος περιτυλίγματος είναι η εξής: Αρχικά χωρίζει τα δεδομένα εκπαίδευσης σε 2 νέα σύνολα, το σύνολο εκπαίδευσης (training) και το σύνολο επικύρωσης (validation). Ακολούθως, διαγράφονται όσα χαρακτηριστικά δεν

ανήκουν στο υποψήφιο προς επιλογή υποσύνολο. Στη συνέχεια, ο ταξινομητής, εκπαιδεύεται με το τροποποιημένο σύνολο εκπαίδευσης και βάση αυτής της εκπαίδευσης κατατάσσει τα στοιχεία που ανήκουν στο τροποποιημένο σύνολο επικύρωσης σε μια σειρά. Η ακρίβεια με την οποία τα δεδομένα αυτά ταξινομούνται, είναι το κριτήριο αξιολόγησης των μεθόδων περιτυλίγματος για ένα οποιοδήποτε υποψήφιο σύνολο χαρακτηριστικών.

Βέβαια, η αξιολόγηση με βάση την απόδοση του ταξινομητή, η οποία απαιτεί την κατασκευή του ταξινομητή για κάθε ξεχωριστό υποσύνολο χαρακτηριστικών που εξετάζεται, έχει ως αρνητικό επακόλουθο το αυξημένο υπολογιστικό κόστος σε σχέση με τις πιο εξελιγμένες ενσωματωμένες μεθόδους ή τα φίλτρα. Το αυξημένο υπολογιστικό κόστος είναι ίσως και το πιο βασικό μειονέκτημα των μεθόδων περιτυλίγματος. Η αποτίμηση κάθε υποψήφιου υποσυνόλου, που συνεπάγεται την εκπαίδευση του ταξινομητή και ύστερα τη μέτρηση της απόδοσης του στο σύνολο επικύρωσης, είναι συνήθως χρονοβόρα διαδικασία και κάνει το υπολογιστικό κόστος ακόμα πιο υψηλό.

Ένα ακόμα από τα πιο βασικά μειονεκτήματα αυτών των μεθόδων, εμφανίζεται στην περίπτωση που τα διαθέσιμα παραδείγματα είναι λίγα. Τότε δεν υπάρχει η δυνατότητα να σχηματιστεί μεγάλο σύνολο επικύρωσης γιατί το σύνολο εκπαίδευσης γίνεται υπερβολικά μικρό και αντίστροφα. Και στις δύο περιπτώσεις υπάρχει πρόβλημα, πιο συγκεκριμένα, στη περίπτωση που το σύνολο εκπαίδευσης είναι μικρό, δεν μπορεί να γίνει καλή εκπαίδευση του ταξινομητή, ενώ στη περίπτωση που το σύνολο επικύρωσης είναι μικρό, δεν μπορεί να γίνει αξιόπιστη εκτίμηση όσο αφορά την ακρίβεια στη ταξινόμηση.

Στην περίπτωση που έχουμε μικρό σύνολο επικύρωσης, για την επίτευξη αξιόπιστων αποτελεσμάτων, χρησιμοποιούνται τεχνικές επαναληπτικής δειγματοληψίας (επαναδειγματοληψίας-resampling techniques), όπως το cross-validation (Stone, 1974) και το bootstrap (Efron, 1983) χάρη στις οποίες αποφεύγεται και η υπερπροσαρμογή (over-fitting) του συνόλου εκπαίδευσης, το οποίο είναι ένα πολύ συχνό φαινόμενο. Στην τεχνική cross-validation τα παραδείγματα εκπαίδευσης χωρίζονται σε k ξένα υποσύνολα (υποσύνολα παραδειγμάτων). Ο ταξινομητής εκπαιδεύεται στα παραδείγματα των k-1 υποσυνόλων, τα οποία παίζουν τον ρόλο του συνόλου εκπαίδευσης, ενώ ένα υποσύνολο παίζει το ρόλο του συνόλου επικύρωσης. Η ίδια διαδικασία επαναλαμβάνεται k φορές

έτσι ώστε κάθε υποσύνολο να παίξει το ρόλο του συνόλου επικύρωσης ακριβώς μία φορά. Ο μέσος όρος της ακρίβειας ταξινόμησης στα k διαφορετικά σύνολα επικύρωσης είναι το κριτήριο αξιολόγησης. Σε περίπτωση που χρησιμοποιείται κάποια τεχνική επαναληπτικής δειγματοληψίας όπως το cross-validation, το ήδη μεγάλο υπολογιστικό κόστος που υπήρχε, αυξάνεται περαιτέρω σε σημαντικό βαθμό.

Τέλος, στα μειονεκτήματα κατατάσσεται και ένα ακόμη πρόβλημα των μεθόδων περιτυλίγματος. Σε μερικές περιπτώσεις, η επιλογή θεωρητικά βέλτιστου υποσυνόλου με βάση την ακρίβεια ταξινόμησης μπορεί να αποδειχθεί τελείως αναξιόπιστη ως μέθοδος παρά το γεγονός ότι η ακρίβεια αυτή μετράται σε ένα ξεχωριστό σύνολο επικύρωσης που δεν χρησιμοποιείται κατά τη φάση της εκπαίδευσης. Αυτό συμβαίνει κυρίως λόγω του μεγάλου όγκου των υποσυνόλων που εξετάζονται, γεγονός που αυξάνει την πιθανότητα να βρεθεί τελικά από τύχη ένα υποσύνολο που δίνει πολύ καλή απόδοση στο σύνολο επικύρωσης, χωρίς όμως να έχει καλή ικανότητα γενίκευσης. Ενώ την ίδια στιγμή είναι πολύ πιθανό άλλα υποσύνολα με σημαντικά μικρότερη ακρίβεια ταξινόμησης στο σύνολο επικύρωσης να επιτυγχάνουν καλύτερη ικανότητα γενίκευσης. Το πρόβλημα αυτό είναι γενικά γνωστό ως το πρόβλημα πολλαπλών συγκρίσεων και γίνεται ακόμα πιο έντονο όταν το σύνολο με τα διαθέσιμα δεδομένα εκπαίδευσης είναι πολύ μικρό.

Παρόλα αυτά οι μέθοδοι περιτυλίγματος έχουν διάφορα πλεονεκτήματα που κάνουν τη χρήση τους ιδιαίτερα δελεαστική, ειδικά σε συνδυασμό με μεθόδους φίλτρων για την ανάπτυξη υβριδικών αλγορίθμων. Οι μεθοδολογίες περιτυλίγματος μπορούν να χρησιμοποιηθούν με οποιονδήποτε ταξινομητή καθώς δεν εξαρτώνται από τον τρόπο λειτουργίας τους, παρά μόνο χρησιμοποιούν την απόδοση τους για να αξιολογήσουν υποψήφια υποσύνολα χαρακτηριστικών.

Το ισχυρότερο επιχείρημα υπέρ της χρήσης των μεθοδολογιών περιτυλίγματος είναι ότι λαμβάνουν υπόψη την επαγωγική μεροληψία (inductive bias) του ταξινομητή. Κάθε ταξινομητής έχει τα δικά του ιδιαίτερα χαρακτηριστικά και τον δικό του τρόπο που απεικονίζει την είσοδο που δέχεται σε έξοδο. Η επαγωγική μεροληψία είναι το σύνολο όλων αυτών των υποθέσεων που κάνει ο ταξινομητής, στη περίπτωση που δεν υπάρχουν επαρκή στοιχεία, έτσι ώστε να μπορέσει να κατατάξει δεδομένα στη σωστή κατηγορία. Σε αυτή τη περίπτωση, εφόσον δεν υπάρχουν παρόμοια παραδείγματα στα δεδομένα εκπαίδευσης και ο ταξινομητής δεν είναι βέβαιος, το πρόβλημα δεν μπορεί να λυθεί

πλήρως. Αυτό σημαίνει ότι το καλύτερο υποσύνολο χαρακτηριστικών για έναν ταξινομητή δεν είναι απαραίτητα το καλύτερο υποσύνολο για έναν ταξινομητή άλλου τύπου. Η ακρίβεια ταξινόμησης είναι το πιο αξιόπιστο κριτήριο για να ελεγχθεί αν ένα υποσύνολο χαρακτηριστικών δουλεύει καλά σε συνδυασμό με έναν ταξινομητή.

Θεωρητικά η χρήση μεθοδολογιών περιτυλίγματος δίνει τη δυνατότητα ανακάλυψης αλληλεπιδράσεων μεταξύ χαρακτηριστικών κι αυτό γιατί τα χαρακτηριστικά δεν αξιολογούνται μεμονωμένα αλλά ως μέρη ενός υποσυνόλου. Φυσικά αν υπάρχουν χαρακτηριστικά που αλληλεπιδρούν, η ανακάλυψη τους εξαρτάται από το αν θα τύχει να βρεθούν στο ίδιο υποψήφιο υποσύνολο ώστε να αξιολογηθούν ως ομάδα. Τελικά αυτό είναι κάτι που εξαρτάται από το μηχανισμό αναζήτησης και κυρίως από το χρόνο που δίνεται στον αλγόριθμο για να εκτελεστεί.

Παρακάτω θα αναλύσουμε τις εμπρόσθιες (forward selection) και τις ανάστροφες (backward selection) διαδικασίες καθώς αυτές οι πιο συνηθέστερα χρησιμοποιούμενες.

Στη μέθοδο της προς τα μπρος επιλογής χαρακτηριστικών (forward selection) ξεκινάμε από το κενό σύνολο, στο οποίο σταδιακά προστίθενται χαρακτηριστικά. Ουσιαστικά το υποσύνολο των επιλεγμένων χαρακτηριστικών “χτίζεται” αυξητικά. Σε κάθε βήμα του αλγόριθμου, εξετάζονται όλα τα υποσύνολα που προκύπτουν από την προσθήκη ενός χαρακτηριστικού στο τρέχον υποσύνολο. Το χαρακτηριστικό που οδηγεί στη μεγαλύτερη αύξηση απόδοσης σύμφωνα με το κριτήριο αξιολόγησης Q της ποιότητας κάθε υποσυνόλου χαρακτηριστικών, ενσωματώνεται στο τρέχον υποσύνολο.

Ο στόχος είναι, πάντα, η επιλογή ενός υποσυνόλου που να αποτελείται από m χαρακτηριστικά, μέσα από ένα μεγαλύτερο σύνολο S_n ($m \leq n$). Αρχικά, ορίζεται το σύνολο των χαρακτηριστικών, να είναι το κενό σύνολο. Επίσης, ορίζεται το σφάλμα ταξινόμησης να ισούται με τον αριθμό των δειγμάτων που υπάρχουν, έστω N. Η διαδικασία ξεκινά με στόχο τον εντοπισμό του πρώτου χαρακτηριστικού, έστω X_1 , για το οποίο μεγιστοποιείται το κριτήριο απόδοσης Q. Το σύνολο που περιέχει αυτό το χαρακτηριστικό ονομάζεται Z_1 . Η μέθοδος συνεχίζεται για τον εντοπισμό του δεύτερου χαρακτηριστικού X_2 από το σύνολο $\{S_n - Z_1\}$, του τρίτου από το $\{S_n - Z_2\}$, κ.ο.κ. Σε κάθε επανάληψη η διαδικασία συνεχίζεται ακριβώς με τον ίδιο τρόπο, με δεδομένο ένα υποσύνολο χαρακτηριστικών $Z_\kappa \subset S_n$ το οποίο έχει επιλεχθεί στο στάδιο k (προηγούμενο

βήμα) της εμπρόσθιας διαδικασίας, το νέο υποσύνολο Z_{k+1} στο επόμενο στάδιο της (τρέχον βήμα) διαμορφώνεται έτσι ώστε το κριτήριο αποτίμησης Q να γίνεται μέγιστο, δηλαδή $Z_{k+1}=Z_k\cup\{X_i\in S_n\setminus Z_k | X_i=\text{argmax } Q(Z_k\cup X_i)\}$. Το κριτήριο που χρησιμοποιείται είναι το σφάλμα της ταξινόμησης. Όσο μεγαλύτερη μείωση έχουμε σε αυτό με την εισαγωγή ενός χαρακτηριστικού, τόσο βέλτιστα πληρείται το κριτήριο Q, και επιλέγεται το συγκεκριμένο χαρακτηριστικό.

Η επέκταση με την προσθήκη ενός χαρακτηριστικού κάθε φορά συνεχίζεται ωσότου ικανοποιηθεί κάποια συνθήκη τερματισμού. Συνήθως η επέκταση σταματάει όταν κανέναν από τα υποσύνολα δεν οδηγεί σε βελτίωση της απόδοσης (αύξηση του κριτηρίου Q) ή εναλλακτικά, όπως είναι προφανές, όλα τα χαρακτηριστικά έχουν επιλεχτεί. Μια πολύ αυστηρή συνθήκη τερματισμού μπορεί να οδηγήσει τον αλγόριθμο σε πρόωρο σταμάτημα. Ενώ μια πιο χαλαρή συνθήκη τερματισμού προβλέπει τη συνέχιση των επεκτάσεων εφόσον υπάρχει κάποιο υποσύνολο που οδηγεί σε απόδοση το ίδιο καλή ή καλύτερη από την έως τώρα και τερματισμό αν η απόδοση δεν βελτιωθεί μετά από m διαδοχικές επεκτάσεις. Αντίστοιχα εδώ το κριτήριο τερματισμού είναι το αν το σφάλμα αρχίσει να αυξάνεται αντί να μειώνεται, δηλαδή όταν, έστω στην k-οστή επανάληψη επιλογής του k χαρακτηριστικού, ισχύει ότι $e_{k+1}>e_k$ για όλα τα υποψήφια χαρακτηριστικά προς επιλογή. Τονίζεται ότι σε περίπτωση ισότητας, δηλαδή στην περίπτωση που ισχύει $e_{k+1}=e_k$, η μέθοδος συνεχίζεται κανονικά λόγω του ότι έχουμε "το ίδιο καλή" απόδοση όπως αναφέρθηκε και προηγουμένως.

Στη μέθοδο της προς τα πίσω επιλογής χαρακτηριστικών (backward selection) γίνεται η αντίθετη διαδικασία από την μέθοδο της προς τα εμπρός επιλογής. Πιο αναλυτικά, η διαδικασία ξεκινά έχοντας όλες τις μεταβλητές διαθέσιμες.

Στο αρχικό σύνολο συμμετέχουν όλα τα χαρακτηριστικά (αντί για το κενό σύνολο που είχαμε πριν), και σε κάθε επανάληψη αντί να προστίθεται ένα χαρακτηριστικό διαγράφεται ένα. Για τη διερεύνηση της σταδιακής απαλοιφής, εξετάζονται σε κάθε βήμα όλα τα δυνατά υποσύνολα που προκύπτουν από τη διαγραφή ενός χαρακτηριστικού από το τρέχον υποσύνολο, και τελικά διαγράφεται αυτό το χαρακτηριστικό, του οποίου η απουσία οδηγεί στη μεγαλύτερη απόδοση ως προς το κριτήριο αξιολόγησης Q. Άρα στην πρώτη επανάληψη αφαιρείται ένα χαρακτηριστικό

από το αρχικό σύνολο S_n και δημιουργείται το υποσύνολο S_{n-1} , στην δεύτερη δημιουργείται το S_{n-2} , κ.ο.κ.. Σε κάθε βήμα με δεδομένο ένα υποσύνολο χαρακτηριστικών $Z_k \subset S_n$ το οποίο έχει επιλεχθεί στο προηγούμενο στάδιο k της διαδικασίας, το νέο υποσύνολο $Z_{k+1} \subset Z_k$ στο παρόν στάδιο της διαμορφώνεται έτσι ώστε το αντίστοιχο κριτήριο αποτίμησης Q $Z_{k+1} = Z_k \setminus \{X_i \in Z_k | X_i = \text{argmax } Q(Z_k \setminus X_i)\}$ να μεγιστοποιείται. Και εδώ τον ρόλο του κριτηρίου Q παίζει η μείωση στο σφάλμα ταξινόμησης. Όσο μεγαλύτερη είναι αυτή τόσο μεγιστοποιείται το Q, οπότε κάθε φορά επιλέγεται το υποσύνολο που έχει το μικρότερο σφάλμα.

Η διαδικασία και στην προς τα πίσω επιλογή σταματά, όταν έχει μείνει μόνο ένα χαρακτηριστικό στο υποσύνολο, δηλαδή $Z_k = S_1$ και ουσιαστικά $|Z_k| = 1$, ή όταν το μοντέλο δεν μπορεί να βελτιωθεί περαιτέρω, δηλαδή, αν για κάθε διαμορφωμένο υποσύνολο Z_{k+1} το κριτήριο Q δεν αυξάνεται ούτε μένει ίδιο. Σε αυτήν την περίπτωση το σφάλμα e_{k+1} , είναι αυστηρά μεγαλύτερο του e_k οποιοδήποτε χαρακτηριστικό και αν διαγραφεί από το υποσύνολο, και άρα δεν υπάρχει βελτίωση του μοντέλου και επομένως ούτε και μεγαλύτερη ακρίβεια στην ταξινόμηση, που σημαίνει ότι όλα τα χαρακτηριστικά του υποσυνόλου Z_k είναι χρήσιμα. Τονίζεται ότι και εδώ, σε περίπτωση που υπάρχει ισότητα ($e_{k+1} = e_k$), η μέθοδος συνεχίζεται κανονικά λόγω του ότι έχουμε “το ίδιο καλή” απόδοση.

3.4 Ενσωματωμένες Μέθοδοι (embedded)

Στην κατηγορία αυτή συναντούμε μεθόδους παρόμοιας περίπου φιλοσοφίας με τις μεθόδους περιτυλίγματος. Εδώ ανήκουν οι μέθοδοι και οι τεχνικές η εφαρμογή των οποίων είναι άμεσα συνδεδεμένη με μια συγκεκριμένη τεχνική ταξινόμησης. Ουσιαστικά έχουν σχεδιαστεί με στόχο να δουλεύουν αποκλειστικά και μόνο σε συνεργασία με ένα ταξινομητή συγκεκριμένου τύπου και ποτέ μόνες τους. Σε αντίθεση με τις μεθόδους περιτυλίγματος που απλώς χρησιμοποιούν την έξοδο ενός ταξινομητή για αξιολόγηση, οι μέθοδοι αυτές επιλέγουν χαρακτηριστικά με βάση το πως επηρεάζεται κάποια συνάρτηση κόστους που εμπλέκεται στη διαδικασία εκπαίδευσης του ταξινομητή.

Ταυτόχρονα, δηλαδή, με την ανάπτυξη των υποδειγμάτων ταξινόμησης οι τεχνικές αυτές ενσωματώνουν στη δομή τους κατάλληλες διαδικασίες επιλογής των χαρακτηριστικών που συμμετέχουν στο τελικό υπόδειγμα.

Από αυτήν την ενσωμάτωση της διαδικασίας της επιλογής χαρακτηριστικών στη διαδικασία της εκπαίδευσης προκύπτουν διάφορα πλεονεκτήματα σε σχέση με τις μεθόδους περιτυλίγματος. Καταρχήν αυτή η κατηγορία μεθόδων έχει μεγάλο κέρδος σε υπολογιστικό κόστος συγκριτικά με την προηγούμενη. Άλλο ένα πλεονέκτημα είναι ότι, οι ενσωματωμένες μέθοδοι καταφέρνουν να κάνουν καλύτερη χρήση των διαθέσιμων δεδομένων (δεδομένα εκπαίδευσης) αφού δεν υπάρχει η ανάγκη αυτά να χωριστούν σε σύνολα εκπαίδευσης και επικύρωσης σε αντίθεση με τις μεθόδους περιτυλίγματος.

Όσον αφορά τις μεθόδους φίλτρων, και πάλι οι μέθοδοι αυτής της κατηγορίας υπερτερούν διότι έχουν το πλεονέκτημα, να λαμβάνουν υπόψη τους την επαγωγική μεροληψία, όπως και οι μέθοδοι περιτυλίγματος.

Κεφάλαιο 4:

Μέθοδοι Επιλογής Χαρακτηριστικών

Στο παρόν κεφάλαιο παρατίθενται οι εξής αλγόριθμοι επιλογής χαρακτηριστικών: η Υβριδική Μέθοδος Φίλτρου/Περιτυλίγματος (4.1), η mMIFS-U Μέθοδος Επιλογής Χαρακτηριστικών (4.2) και οι NMIFS, GAMIFS Μέθοδοι Επιλογής Χαρακτηριστικών (4.3). Για κάθε έναν αλγόριθμο επιλογής χαρακτηριστικών αρχικά δίνεται μια ανάλυση του μαθηματικού και υπολογιστικού υποβάθρου που απαιτείται, στη συνέχεια ακολουθεί η περιγραφή της αλγοριθμικής διαδικασίας και παρατίθενται τα σχετικά πειραματικά αποτελέσματα από την δοκιμή του και τελικά εξάγονται κάποια συμπεράσματα. Είναι αναμενόμενο ότι απαιτείται η ορθή επιλογή καθενός από τους προαναφερθέντες αλγορίθμους με γνώμονα τη φύση του προβλήματος, προκειμένου εκείνος να αποφέρει βέλτιστα αποτελέσματα.

4.1 Υβριδική Μέθοδος Φίλτρου/Περιτυλίγματος

4.1.1 Εισαγωγή

Στην επιβλεπόμενη μάθηση, όπως έχει ήδη αναφερθεί, πολλές φορές τα προβλήματα που είναι προς επίλυση έχουν πολλά χαρακτηριστικά, διάφορα από τα οποία μάλιστα, είναι άσχετα με το συγκεκριμένο πρόβλημα ταξινόμησης. Αυτά τα άσχετα ή περιττά χαρακτηριστικά περιπλέκουν τις διαδικασίες μάθησης. Η υψηλή διάσταση του συνόλου χαρακτηριστικών αφενός αυξάνει το πρόβλημα της υπερπροσαρμογής και αφετέρου μπορεί να οδηγήσει σε κακή ικανότητα γενίκευσης, απαιτώντας παράλληλα περισσότερο

χρόνο για την εκμάθηση.

Από τις τρεις πιο γνωστές κατηγορίες μεθόδων (φίλτρου, περιτυλίγματος, ενσωματωμένες μέθοδοι), οι μέθοδοι περιτυλίγματος έχουν ως βασικό μειονέκτημα το υψηλό κόστος και γι'αυτό αποφεύγεται η χρήση τους για μεγάλα σύνολα δεδομένων.

Σύγχρονες εφαρμογές στοχεύουν στην αντιμετώπιση αυτού του προβλήματος συνδυάζοντας μεθόδους φίλτρου με μεθόδους περιτυλίγματος, για να επιταχύνουν τον αλγόριθμο της μεθόδου περιτυλίγματος, μέσω προεπεξεργασίας με μέθοδο φίλτρου, ενώ παράλληλα περιορίζεται και η μεροληψία.

Η υβριδική μέθοδος Φίλτρου/Περιτυλίγματος (Schaffernicht E., 2011) που αναλύεται, χρησιμοποιεί την σταθμισμένη αμοιβαία πληροφορία ως στατιστικό μέτρο του φίλτρου για να αξιολογήσει τα χαρακτηριστικά σε μια προς τα εμπρός επιλογή.

4.1.2 Σταθμισμένη Αμοιβαία Πληροφορία

Η γνωστή αμοιβαία πληροφορία που αναφέρθηκε και στο Κεφάλαιο 2, αξιολογεί την εξάρτηση μεταξύ δύο τυχαίων μεταβλητών. Όσον αφορά την επιλογή χαρακτηριστικών, υπολογίζει την εξάρτηση ανάμεσα σε ένα χαρακτηριστικό F_i και μια μεταβλητή T, η οποία μπορεί να παίρνει τιμές ανάλογα με την κλάση που απευθυνόμαστε. Τα πεζά γράμματα στον τύπο χρησιμοποιούνται για τον συμβολισμό των διαφόρων τιμών που μπορούν να πάρουν αυτές οι τυχαίες μεταβλητές. Υπενθυμίζεται η σχέση της αμοιβαίας πληροφορίας:

$$I(F_i,T)=\int_{f_i}\int_{t} p(f_i,t)\cdot\log\left(\frac{p(f_i,t)}{p(f_i)\cdot p(t)}\right)dtdf_i \quad . \qquad 4.1$$

Η ιδέα της σταθμισμένης αμοιβαίας πληροφορίας (Guiasu S., 1977) δεν είναι ευρέως γνωστή διότι έχει περιορισμένο αριθμό πρακτικών εφαρμογών. Αναλυτικά ορίζεται ως:

$$wI(F_i;T;W)=\int_{f_i}\int_{t} w(s_j)\cdot p(f_i,t)\cdot\log\left(\frac{p(f_i,t)}{p(f_i)\cdot p(t)}\right)dtdf_i \quad . \qquad 4.2$$

Για κάθε τιμή s_j , όπου καθεμία από τις οποίες αναφέρεται σε έναν συγκεκριμένο

συνδυασμό ενός χαρακτηριστικού f_i και μιας κλάσης t, αντιστοιχείται ένα βάρος $w(s_j)$, το οποίο είναι μεγαλύτερο είτε ίσο του 0, δηλαδή $w(s_j) \geq 0$. Αυτό οδηγεί στην δημιουργία συγκεκριμένης σχετικότητας κάθε τιμής s_j με το μοντέλο, ανάλογης με το μέτρο της πληροφορίας που φέρει ένας συγκεκριμένος συνδυασμός. Στον αλγόριθμο θα χρησιμοποιηθεί για να σταθμιστεί η επίδραση των διάφορων συνδυασμών χαρακτηριστικού κλάσης.

Το κύριο πρόβλημα με την εξίσωση της αμοιβαίας πληροφορίας, είναι η εκτίμηση της απαιτούμενης σ.π.π. από τα διαθέσιμα δεδομένα. Η απλούστερη προσέγγιση, για τον υπολογισμό της αμοιβαίας πληροφορίας, είναι η χρήση ιστογραμμάτων για την εκτίμηση της σ.π.π. . Με αυτόν τον τρόπο απλοποιείται η διαδικασία του υπολογισμού, αφού χρησιμοποιούνται αθροίσματα, αντί για ολοκληρώματα ή μέθοδοι εκτίμησης πυκνότητας με πυρήνα (Khan S., 2007) (Schaffernicht E., 2010).

Ο υπολογισμός της σταθμισμένης αμοιβαίας πληροφορίας γίνεται άμεσα μόνο με κάποιες μεθόδους εκτίμησης. Αντί της παραπάνω εξίσωσης είναι ευκολότερο να χρησιμοποιηθεί η ακόλουθη σχέση:

$$wI(F_i; T; W) = \int_{f_i} \int_{t} w(s_j) \cdot p(f_i, t) \cdot \log\left(\frac{w(s_j) \cdot p(f_i, t)}{w(s_j) \cdot p(f_i) \cdot p(t)}\right) dt df_i \quad . \qquad 4.3$$

Ουσιαστικά, αυτή η σχέση υπολογίζει την σταθμισμένη αμοιβαία πληροφορία εισάγοντας μια συνάρτηση βάρους στις κατανομές πιθανότητας. Κάθε συνδυασμός συμβάλει στην σ.π.π. ανάλογα με το βάρος που του έχει τεθεί (οι συνδυασμοί με μηδενικό βάρος όπως είναι αναμενόμενο δεν συμβάλουν στην σ.π.π.).

Με την χρήση της τεχνικής του ιστογράμματος, πρώτα κατηγοριοποιούνται τα δεδομένα χωρίζοντας την κλίμακα τους σε ίσα διαστήματα και μετρώντας πόσες τιμές αντιστοιχούν σε κάθε διάστημα. Στη συνέχεια αντικαθιστώνται τα ολοκληρώματα στον τύπο της αμοιβαίας πληροφορίας με τα αθροίσματα των ράβδων του ιστογράμματος. Εάν χρησιμοποιηθεί αυτή η μέθοδος, δεν χρειάζεται να εισάγουμε βάρος. Ο λόγος είναι, ότι κάθε δείγμα δεν συμβάλει το ίδιο στην κλάση του και κατ' επέκταση στην πυκνότητα πιθανότητας, αλλά ανάλογα με το βάρος του.

4.1.3 Αλγόριθμος Υβριδικής Μεθόδου Φίλτρου/Περιτυλίγματος

Η βασική ιδέα του υβριδικού αλγορίθμου επιλογής χαρακτηριστικών που προτείνεται σε αυτήν την προσέγγιση είναι η εξής: Έστω ότι θεωρείται δεδομένος ένας ταξινομητής-προσεγγιστική συνάρτηση και το σφάλμα που δημιουργείται στην ταξινόμηση. Προκειμένου να επιλεχθεί το χαρακτηριστικό που θα εισαχθεί στο υποσύνολο των επιλεγμένων χαρακτηριστικών στο επόμενο βήμα του αλγορίθμου, έτσι ώστε να βελτιωθεί η απόδοση του ταξινομητή, χρησιμοποιούνται τα εσφαλμένα ταξινομημένα δείγματα και όχι όλα τα διαθέσιμα δείγματα. Αυτό επιτυγχάνεται σταθμίζοντας τα ταξινομημένα δείγματα που έχουν εισαχθεί σε σωστή και λάθος κλάση διαφορετικά στον υπολογισμό της αμοιβαίας πληροφορίας, εισάγοντας ουσιαστικά σε αυτά διαφορετικά βάρη.

Η απλούστερη περίπτωση ταξινόμησης που που μπορεί να συναντηθεί σε κάποιο πρόβλημα είναι όταν ένας ταξινομητής κάνει διακριτή ταξινόμηση των δεδομένων. Σε κάθε επανάληψη του αλγορίθμου που αναλύεται εδώ και θα παρουσιαστεί και αργότερα, όλα τα ταξινομημένα δείγματα που έχουν αντιστοιχιστεί στην σωστή κλάση δεν συμμετέχουν στον υπολογισμό της νέας αμοιβαίας πληροφορίας ανάμεσα στα υπολείποντα προς επιλογή χαρακτηριστικά και τις κλάσεις, αφού ναι μεν εισάγονται ως όροι στο άθροισμα, αλλά τους δίνεται μηδενικό βάρος. Μόνο τα λάθος ταξινομημένα δείγματα χρησιμοποιούνται σε αυτόν τον υπολογισμό, έχοντας μάλιστα ίσα βάρη μεταξύ τους στο άθροισμα.

Αντίθετα, στην περίπτωση που χρησιμοποιείται συνεχής συνάρτηση μεταβλητών, υπάρχει η δυνατότητα σε κάθε δείγμα να αντιστοιχίζεται διαφορετικό βάρος ανάλογα με το υπόλοιπο που έχει. Για παράδειγμα, ένα θετικό δείγμα το οποίο έχει ταξινομηθεί σωστά ως θετικό, αλλά είναι κοντά στο σύνορο με την κλάση των αρνητικών θα φέρει ένα υπόλοιπο διάφορο του μηδενός παρόλο που είναι στην σωστή κλάση. Ταυτόχρονα όμως, και η επιρροή του θα πρέπει να είναι μικρότερη από ένα άλλο δείγμα, το οποίο είναι από την λάθος μεριά του συνόρου απόφασης.

Παρακάτω ακολουθεί ο αλγόριθμος που περιγράφει την υβριδική μέθοδο φίλτρου περιτυλίγματος:

Algorithm 1. $S=wMI(X, Y)$

Input: data set of observations X and the corresponding labels Y (εισαγωγή των δειγμάτων X με όλα τα χαρακτηριστικά και της αντίστοιχης κλάσης Y)

Output: final feature set S and the final classifier (τελικό σύνολο χαρακτηριστικών και τελική ταξινόμηση)

$S \leftarrow 0$ (θέτουμε το σύνολο χαρακτηριστικών ως το κενό)

$W \leftarrow 1$ {Same weight for all samples} (θέτουμε βάρος 1 για όλους τους συνδυασμούς)

while stopping criterion not true **do** (έλεγχος συνθήκης τερματισμού)

$F_{max}=\text{argmax}_{F_i}[wI(F_i, T, W)]$ {Find feature with maximum weighted MI} (εύρεση χαρακτηριστικού με μέγιστη σταθμισμένη αμοιβαία πληροφορία)

$S \leftarrow S \cup F_{max}$ {Add feature to the subset} (εισαγωγή στο τελικό σύνολο χαρακτηριστικών)

$F \leftarrow F \setminus F_{max}$ {Remove feature from the candidate set} (αφαίρεση από το υποψήφιο σύνολο χαρακτηριστικών)

$Classifier \leftarrow$ **TRAINCLASSIFIER** (X, S, Y) (εκμάθηση ταξινομητή)

$Y' \leftarrow$ **APPLYCLASSIFIER** $(Classifier, X)$ (ταξινόμηση δειγμάτων X)

$W \leftarrow |Y-Y'|$ {Residual for each sample is the new weight.} (υπολογισμός υπολοίπου που τίθεται ως το νέο βάρος)

CHECKSTOPPINGCRITERION $()$

end while

Όπως φαίνεται και στον αλγόριθμο, χρησιμοποιήθηκε η σταθμισμένη αμοιβαία πληροφορία ως κριτήριο στα πλαίσια μιας προς τα εμπρός επιλογής χαρακτηριστικών. Αρχικά ορίζεται ένα κενό σύνολο χαρακτηριστικών. Έπειτα υπολογίζεται η σταθμισμένη αμοιβαία πληροφορία ανάμεσα σε όλα τα χαρακτηριστικά και όλες τις κλάσεις με ίσο βάρος για όλα τα δείγματα. Το χαρακτηριστικό με την μέγιστη σταθμισμένη αμοιβαία πληροφορία επιλέγεται για να εισέλθει στο σύνολο με έναν απλό αλγόριθμο κατάταξης σε φθίνουσα σειρά. Ύστερα ο ταξινομητής εκπαιδεύεται με αυτό το χαρακτηριστικό, και ταξινομεί όλα τα δείγματα. Για την εισαγωγή του βάρους σε κάθε δείγμα χρησιμοποιείται το υπόλοιπο του. Για την επιλογή του επόμενου χαρακτηριστικού μέσω της

σταθμισμένης αμοιβαίας πληροφορίας, χρησιμοποιείται το βάρος που μόλις υπολογίστηκε. Ο αλγόριθμος επαναλαμβάνεται, και στο επόμενο βήμα αφού επιλεχθεί το χαρακτηριστικό με την αμέσως επόμενη μέγιστη σταθμισμένη αμοιβαία πληροφορία, ο ταξινομητής ξαναεκπαιδεύεται και κάνει την νέα ταξινόμηση. Αυτή η διαδικασία συνεχίζεται έως ότου ικανοποιηθεί το κριτήριο τερματισμού.

Η μέθοδος μοιάζει με την βασική ιδέα του γνωστού αλγορίθμου AdaBoost (Freund Y., 1997) ο οποίος χρησιμοποιεί βάρη για να πετύχει καλύτερη ταξινόμηση. Όλα τα δείγματα που έχουν ταξινομηθεί λάθος θεωρούνται πιο σημαντικά στην επόμενη επανάληψη από εκείνα που ταξινομήθηκαν σωστά. Το σκεπτικό είναι ότι τα δείγματα που έχουν ταξινομηθεί σωστά, περιγράφονται επαρκώς από το υπάρχον υποσύνολο χαρακτηριστικών, και το ζητούμενο είναι να βρεθούν αυτά τα χαρακτηριστικά που περιγράφουν επαρκώς και τα λάθος ταξινομημένα δείγματα.

Το εύρος των τιμών Y, όταν αυτές είναι πραγματικές τιμές, είναι αυθαίρετο και δεν έχει σημασία καθώς δεν μας ενδιαφέρει η απόλυτη τιμή της σταθμισμένης αμοιβαίας πληροφορίας, αλλά η σχετική τιμή της σε σχέση με τα υπόλοιπα χαρακτηριστικά, τα οποία υπολογίζονται και αυτά με τα ίδια βάρη.

Παίρνοντας υπόψιν τους προσεγγιστικούς αλγόριθμους που υπάρχουν στην διεθνή βιβλιογραφία, όπως τον MLP, το παραπάνω δεν αποτελεί πρόβλημα. Το υπερεπίπεδο που διαχωρίζει τις κλάσεις υπολογίζεται ξανά σε κάθε βήμα, με την διαφορά ότι ο νέος υπόχωρος του χώρου όλων των χαρακτηριστικών, περιλαμβάνει το νέο χαρακτηριστικό που μπήκε στο υποσύνολο χαρακτηριστικών, και έτσι δημιουργείται σε αυτόν μια ακόμα διάσταση. Με την εισαγωγή της δημιουργούνται περισσότερες επιλογές για την επιλογή υπερεπιπέδου διαχωρισμού, ενώ παράλληλα το αποτέλεσμα του προηγούμενου βήματος, είναι πάντα και πάλι εφικτό.

Για ταξινομητές που χρησιμοποιούν συναρτήσεις υπολογισμού απόστασης όπως τα νευρωνικά δίκτυα RBF ή ταξινομητές του πλησιέστερου γείτονα, η κατάσταση είναι λίγο πιο περίπλοκη, ειδικά όταν οι χώροι είναι μικρής διάστασης. Η γειτονιά ενός δείγματος μπορεί να αλλάξει πολύ με την εισαγωγή ενός νέου χαρακτηριστικού. Προφανώς αυτό το πρόβλημα αμβλύνεται στην περίπτωση χώρου μεγάλης διάστασης αφού η εισαγωγή νέων χαρακτηριστικών επηρεάζει την γειτονιά του χαρακτηριστικού λιγότερο. Η αποδοτικότητα, βέβαια, του αλγορίθμου στα αρχικά στάδια μπορεί να μειωθεί λόγω

αυτού του φαινομένου. Όμως ο αλγόριθμος θα προσπαθήσει να το διορθώσει βασιζόμενος στα νέα υπόλοιπα που θα εισαχθούν ως βάρη, και θα διαλέξει χαρακτηριστικά που θα αντισταθμίσουν το πρόβλημα των νεοεισαχθέντων σφαλμάτων. Σαν συνέπεια, αυτή η διαδικασία θα δημιουργήσει ανωμαλία στα σφάλματα προκαλώντας διαδοχικά απότομες αυξήσεις και μειώσεις.

Η εύρεση καλού κριτηρίου τερματισμού μπορεί να αποδειχθεί δύσκολη, αλλά κρίσιμη, ειδικά στην περίπτωση των ταξινομητών που χρησιμοποιούν συναρτήσεις υπολογισμού απόστασης. Ο αλγόριθμος ξεκινάει με ένα άδειο υποσύνολο χαρακτηριστικών, και σε κάθε επανάληψη εισάγει ένα χαρακτηριστικό στο τελικό υποσύνολο του προηγούμενου βήματος, μέχρις ότου το πλήθος το χαρακτηριστικών φτάσει έναν επιθυμητό αριθμό ο οποίος έχει προκαθοριστεί, ή το προσεγγιστικό αποτέλεσμα δεν βελτιώνεται περαιτέρω. Στην περίπτωση που το νέο υποσύνολο βελτιώνει την απόδοση περισσότερο από ένα όριο ε, το οποίο μπορεί να πάρει και αρνητικές τιμές όταν η απόδοση μειώνεται, το νέο υποσύνολο επιβεβαιώνεται και ξεκινάει καινούρια επανάληψη. Διαφορετικά ο αλγόριθμος τερματίζεται. Άλλα πιθανά κριτήρια τερματισμού είναι να υπάρχει δοσμένος αριθμός βημάτων, που ισοδυναμεί με συγκεκριμένο αριθμό χαρακτηριστικών στο τελικό υποσύνολο, ή ένα συγκεκριμένο μέγιστο σφάλμα προσέγγισης στην τελική ταξινόμηση.

4.1.4 Παρόμοιοι Αλγόριθμοι

Η χρήση της αμοιβαίας πληροφορίας για την επιλογή χαρακτηριστικών είναι αρκετά συχνή. Εκτός από τις απλές βαθμωτές μεθόδους επιλογής χαρακτηριστικών (Torkkola K., 2006), οι οποίες δεν παίρνουν υπόψιν τις συσχετίσεις μεταξύ των χαρακτηριστικών και δεν διαχειρίζονται τα φαινόμενα πλεονασμού, αξιοσημείωτος είναι ο αλγόριθμος MIFS (Battiti R., 1994)και οι παραλλαγές του. Ο αλγόριθμος MIFS προσεγγίζει την κοινή αμοιβαία πληροφορία μέσω μόνο της πληροφορίας μεταξύ της κλάσης και του υποψηφίου χαρακτηριστικού και του αθροίσματος των αμοιβαίων πληροφοριών των χαρακτηριστικών ανά ζεύγη, δηλαδή κάθε ήδη επιλεγμένο χαρακτηριστικό με το υποψήφιο προς επιλογή χαρακτηριστικό. Αυτές οι μέθοδοι ανήκουν στην κατηγορία των φίλτρων και δεν λαμβάνουν υπόψη τους το αποτέλεσμα της μηχανής εκμάθησης, και ως

εκ τούτου, δεν εξαλείφουν την μεροληψία που εισάγεται.

Οι μέθοδοι περιτυλίγματος με τις εμπρόσθιες και τις ανάστροφες διαδικασίες αντιμετωπίζουν το φαινόμενο της μεροληψίας, όμως απαιτούν πολύ χρόνο για να καταλήξουν σε κάποιο αποτέλεσμα. Μερικές προτεινόμενες μέθοδοι, όπως οι floating search algorithms, αυξάνουν τον αριθμό των υποσυνόλων που εξετάζονται, συνδυάζοντας την προς τα εμπρός και την προς τα πίσω επιλογή χαρακτηριστικών, δηλαδή είτε εισάγοντας είτε διαγράφοντας πολλά χαρακτηριστικά σε ένα βήμα. Αυτό αυξάνει τον απαιτούμενο χρόνο ακόμη περισσότερο κάνοντάς αυτές της μεθόδους πολύ δύσχρηστες για μεγάλο αριθμό δεδομένων. Από την άλλη γίνεται προσπάθεια μείωσης των υπό εξέταση υποσυνόλων, χωρίς όμως αν είναι δυνατόν να χαθούν σημαντικά για το πρόβλημα υποσύνολα.

Ο συνδυασμός φίλτρων που βασίζονται στην αμοιβαία πληροφορία με μεθόδους περιτυλίγματος είναι μια προσέγγιση του προβλήματος. Μια μέθοδος (Peng H., 2005) που προτείνεται είναι η χρήση των μεθόδων περιτυλίγματος για το ξεδιάλυμα υποψήφιων υποσυνόλων που δημιουργούνται από κάποια αυξητική μέθοδο φίλτρου (incremental filter method) που βασίζεται στο μέτρο της αμοιβαίας πληροφορίας. Μια ακόμη μέθοδος (Van Dijck G., 2006) εφαρμόζει μια μέθοδο φίλτρου, που χρησιμοποιεί κριτήρια μέγιστης σχετικότητας και ελάχιστου πλεονασμού που βασίζονται στην αμοιβαία πληροφορία για να ελαττώσει τα χαρακτηριστικά των μεταβλητών εισόδου, έτσι ώστε ύστερα να τα εισάγει σε μια γενετική μέθοδο περιτυλίγματος (genetic wrapper search). Τα δέντρα Chow-Liu (Schaffernicht E., 2007) χρησιμοποιούνται για την μείωση των υποψήφιων χαρακτηριστικών σε κάθε ντετερμινιστική προς τα εμπρός επιλογή. Η δομή αυτών των δέντρων είναι μια προεπεξεργασία με μια μέθοδο παρόμοιας φιλοσοφίας με τις μεθόδους φίλτρου, η οποία χρησιμοποιεί το μέτρο της αμοιβαίας πληροφορίας. Όλες οι προαναφερθείσες μέθοδοι λειτουργούν υπολογίζοντας την αμοιβαία πληροφορία είτε μεταξύ των χαρακτηριστικών και των κλάσεων, είτε των χαρακτηριστικών μεταξύ τους, αλλά δεν συνυπολογίζουν το αποτέλεσμα του ταξινομητή.

Μια από τις λίγες μεθόδους που χρησιμοποιούν το αποτέλεσμα της μηχανής εκμάθησης για τον υπολογισμό της αμοιβαίας πληροφορίας στα πλαίσια της επιλογής χαρακτηριστικών, παρουσιάζεται από τον Torkkola (Torkkola K., 2003). Η ιδέα βασίζεται στη χρήση της θεωρίας πληροφοριών και υπολογίζει την τετραγωνική αμοιβαία

πληροφορία (quadratic mutual information) ανάμεσα στο αποτέλεσμα του μετασχηματισμού των δεδομένων και τον επιθυμητό στόχο για να προσαρμόσει τον μετασχηματισμό των χαρακτηριστικών βαθμωτά. Στόχος δηλαδή είναι η εύρεση του μετασχηματισμού που μεγιστοποιεί την αμοιβαία πληροφορία. Η προσέγγιση είναι μια μέθοδος φίλτρου ουσιαστικά, αφού η μηχανή εκμάθησης εκπαιδεύεται πάνω στον μετασχηματισμό των χαρακτηριστικών και δεν παρέχει αποτελέσματα ταξινόμησης των δειγμάτων.

Η πιο κοντινή προσέγγιση είναι ο αλγόριθμος RMI (Residual Mutual Information). Υπολογίζει την αμοιβαία πληροφορία ανάμεσα στις κλάσεις και το υπόλοιπο που δημιουργείται κατά την διάρκεια της εκμάθησης. Σε αυτήν την περίπτωση, το υπόλοιπο χρησιμοποιείται διαφορετικά. Θεωρείται ότι τα δείγματα που περιέχουν πληροφορίες σχετικά με το σφάλμα που θα δημιουργηθεί από την ταξινόμηση, μπορούν να χρησιμοποιηθούν για την ελάττωση του σφάλματος, και ως εκ τούτου επιλέγονται.

4.1.5 Πειραματικά Αποτελέσματα Υβριδικής Μεθόδου Φίλτρου/Περιτυλίγματος

Η αξιολόγηση του υβριδικού αλγορίθμου φίλτρου-περιτυλίγματος έγινε με εφαρμογή του σε σύγκριση με παρόμοιες μεθόδους, σε σύνολα δεδομένων από την βιβλιοθήκη του UCI (Newman D.J., 1998) και ένα άλλο μεγαλύτερο τεχνητό σύνολο δεδομένων, που δημιουργήθηκε με γνωστές ιδιότητες: 200 χαρακτηριστικά, 7 εγγενείς διαστάσεις, γραμμικές, μη γραμμικές και XOR συναρτησιακές εξαρτήσεις, πλεονασμούς και θόρυβο για όλες τις μεταβλητές. Ο ταξινομητής που χρησιμοποιήθηκε είναι του πλησιέστερου γείτονα για την μεταβλητή k να ισούται με τρία (k=3) και ένα πολυεπίπεδο νευρωνικό δίκτυο (Multi Layer Perceptron) με δύο κρυφά επίπεδα, με είκοσι και δέκα κρυφούς νευρώνες αντίστοιχα. Αυτοί οι ταξινομητές χρησιμοποιήθηκαν ως “μαύρα κουτιά” από τον αλγόριθμο και μετά την επιλογή κατάλληλου υποσυνόλου χαρακτηριστικών έπαιξαν τον ρόλο του τελικού ταξινομημένου συνόλου προς σύγκριση. Οι παράμετροι των ταξινομητών ορίστηκαν για όλα τα σύνολα, καθώς η επιλογή ταξινομητή δεν είναι το υπό εξέταση πρόβλημα. Γιαυτό και δεν υπήρχαν επιδιορθώσεις για τα διαφορετικά σύνολα δεδομένων και τους διαφορετικούς αλγόριθμους επιλογής χαρακτηριστικών, με

αποτέλεσμα να υπάρχει μεγάλη μεροληψία στο σφάλμα. Κατά την διάρκεια της επιλογής χαρακτηριστικών χρησιμοποιήθηκε η τεχνική cross-validation και χωρίστηκαν τα δεδομένα εκπαίδευσης σε τρία υποσύνολα, ενώ για την τελική αξιολόγηση της ίδιας της πρόβλεψης χρησιμοποιήθηκε η ίδια τεχνική αλλά έγινε διαίρεση σε δέκα υποσύνολα. Όλα τα προβλήματα, είναι προβλήματα δυαδικής ταξινόμησης, και γιαυτό ως μέτρο σύγκρισης χρησιμοποιήθηκε το μέσο σταθμισμένο σφάλμα (balanced error rate).

Η παρούσα μέθοδος, που χρησιμοποιεί την σταθμισμένη αμοιβαία πληροφορία, συγκρίθηκε με άλλους αλγόριθμους. Πιο συγκεκριμένα με τον αλγόριθμο προς τα εμπρός επιλογής (basic Sequential Forward Selection (SFS)), τον έναν αλγόριθμο που χρησιμοποιεί την απλή αμοιβαία πληροφορία (Mutual Information for Feature Selection (MIFS με $\beta=0,15$)), δέντρα Chow-Liu (CLT-FS) και έναν ακόμη αλγόριθμο που χρησιμοποιεί το υπόλοιπο (residual Mutual Information (RMI)). Τα αποτελέσματα για τον αλγόριθμο του πλησιέστερου γείτονα φαίνονται στον Πίνακα 4.1, και για τον MLP στον Πίνακα 4.2.

Πίνακας 4.1 Τα αποτελέσματα χρησιμοποιώντας την μέθοδο του πλησιέστερου γείτονα

Σύνολο Δεδομένων	Ionosphere	Spambase	GermanCredit	Breast Cancer	Artifical Data
Χαρακτηριστικά	34	57	24	30	200
Δείγματα	351	4601	1000	569	3000
Όλα	23.78(34/-)	10.84(57/-)	36.33(24/-)	3.55(30/-)	35.36(200/-)
SFS	12.04(5/189)	8.44(12/663)	31.61(7/164)	4.21(6/189)	16.20(8/1764)
MIFS	11.80(5/-)	8.65(19/-)	33.90(6/-)	4.36(5/-)	28.65(3/-)
CLT-FS	12.19(6/39)	15.97(6/76)	34.89(5/28)	4.42(4/30)	25.89(8/202)
RMI	13.82(5/6)	23.62(3/4)	35.45(5/6)	4.49 (3/4)	24.30(4/5)
WMI	11.57(5/6)	10.73(10/11)	33.31(8/9)	4.48(6/7)	29.52(5/6)

Στο παραπάνω πίνακα παρατίθενται τα αποτελέσματα της υβριδικής μεθόδου σε σύγκριση με τις υπόλοιπες μεθόδους, όταν αυτές χρησιμοποιήθηκαν σε συνδυασμό με τη μέθοδο ταξινόμησης του πλησιέστερου γείτονα για διάφορα σύνολα δεδομένων. Το σταθμισμένο σφάλμα αναγράφεται σε ποσοστό. Το πλήθος των επιλεγμένων

χαρακτηριστικών και των επαναλήψεων του ταξινομητή (ταξινόμηση και αξιολόγηση) είναι μέσα στην παρένθεση.

Πίνακας 4.2 Τα αποτελέσματα χρησιμοποιώντας την μέθοδο νευρωνικού δικτύου

Σύνολο Δεδομένων	Ionosphere	Spambase	GermanCredit	Breast Cancer	Artifical Data
Χαρακτηριστικά	34	57	24	30	200
Δείγματα	351	4601	1000	569	3000
Όλα	20.08(34/-)	13.81(57/-)	41.70(24/-)	13.78(30/-)	33.65(200/-)
SFS	18.47(3/130)	17.39(8/477)	39.06(4/110)	13.44(4/140)	20.11(7/1572)
MIFS	24.54(5/-)	16.29(19/-)	37.47(6/-)	12.48(5/-)	26.87(3/-)
CLT-FS	18.12(6/38)	17.26(9/97)	38.52(3/24)	9.37(8/37)	24.08(9/217)
RMI	17.08(5/6)	13.93(54/55)	39.73(15/16)	8.58(5/6)	21.74(6/7)
WMI	16.97(5/6)	16.41(9/10)	39.52 (6/7)	8.03(3/4)	19.29(6/7)

Στο παραπάνω πίνακα παρατίθενται τα αποτελέσματα της υβριδικής μεθόδου σε σύγκριση με τις υπόλοιπες μεθόδους, όταν αυτές χρησιμοποιήθηκαν σε συνδυασμό με τη μέθοδο ταξινόμησης του νευρωνικού δικτύου για διάφορα σύνολα δεδομένων. Το σταθμισμένο σφάλμα αναγράφεται και εδώ σε ποσοστό. Καθώς και το πλήθος των επιλεγμένων χαρακτηριστικών και των επαναλήψεων του ταξινομητή (ταξινόμηση και αξιολόγηση) είναι και εδώ μέσα στην παρένθεση.

Μελετώντας τα νούμερα στους παραπάνω πίνακες (Πίνακας 4.1, Πίνακας 4.2), συμπεραίνεται ότι η υπόθεση της καλής αποδοτικότητας του αλγορίθμου ήταν σωστή. Η αποδοτικότητα της υβριδικής μεθόδου, με τη χρήση του πλησιέστερου γείτονα, είναι η καλύτερη σε σύγκριση με τις υπόλοιπες μεθόδους. Όσον αφορά τον MLP, τα αποτελέσματα είναι ακόμη πιο ικανοποιητικά.

Ο αριθμός των απαραίτητων βημάτων για την προσαρμογή και την αξιολόγηση του ταξινομητή, είναι ο μικρότερος με εξαίρεση τον RMI και τον MIFS. Ειδικά στην περίπτωση που έχουμε μεγάλα σύνολα δεδομένων, η γραμμική εξάρτηση των χαρακτηριστικών είναι πιο ωφέλιμη σε σύγκριση με την τετραγωνική εξάρτηση του SFS και την λογαριθμική της CLT μεθόδου.

4.1.6 Συμπεράσματα

Ο υβριδικός αλγόριθμος φίλτρου-περιτυλίγματος στοχεύει στην ελαχιστοποίηση των επαναλήψεων εκπαίδευσης και αξιολόγησης του ταξινομητή κατά την διαδικασία επιλογής χαρακτηριστικών. Το πλήθος των επαναλήψεων είναι γραμμικό σε σχέση με τα επιλεγμένα χαρακτηριστικά μόνο στον αλγόριθμο WMI, ο οποίος παράλληλα πετυχαίνει ίδια και καλύτερη ακρίβεια από αυτήν των υπολοίπων μεθόδων περιτυλίγματος που έχουν περισσότερα βήματα.

Με βάση τα αποτελέσματα του πρώτου και του δεύτερου πίνακα εικάζεται ότι η χρήση του RMI είναι λιγότερο χρηστική στην περίπτωση της δυαδικής ταξινόμησης. Πιο πολύπλοκα προβλήματα μπορούν να επιλυθούν καλύτερα με τον RMI.

Παίρνοντας όλα τα στοιχεία υπόψιν είναι δύσκολο να καθοριστεί ποιος από τους δύο αλγόριθμους RMI και WMI είναι βέλτιστος συνολικά. Βέβαια μέχρι στιγμής είναι εμφανές ότι σε σχέση με τους υπόλοιπους αλγορίθμους ο WMI είναι προτιμότερος, ωστόσο απαιτούνται πιο εκτεταμένες μελέτες για να απαντηθεί οριστικά αυτή η ερώτηση.

4.2 mMIFS-U Μέθοδος Επιλογής Χαρακτηριστικών

4.2.1 Εισαγωγή

Η μέθοδος mMIFS-U – Modified Mutual Information for Feature Selection under Uniform Information Distribution (Novovicova J., 2007) ανήκει στην κατηγορία των μεθόδων φίλτρου. Ένα από τα προβλήματα αυτών των μεθόδων είναι το σύστημα αναζήτησης. Έχουν προταθεί διάφορες προσεγγίσεις του προβλήματος στη διεθνή βιβλιογραφία όπως η κλασσική αναζήτηση, η ευρετική αναζήτηση και η τυχαία αναζήτηση, που σκοπό έχουν την εύρεση του βέλτιστου συνδυασμού ανάμεσα στην ακρίβεια των αποτελεσμάτων και την υπολογιστική αποδοτικότητα. Πολλοί αλγόριθμοι φίλτρου αξιολογούν όλα τα χαρακτηριστικά ξεχωριστά σύμφωνα με κάποιο κριτήριο, τα ταξινομούν και διαλέγουν αυτά που είναι μεμονωμένα καλύτερα. Η επιλογή με αυτόν τον τρόπο δεν εξασφαλίζει

ότι τα χαρακτηριστικά είναι ασυσχέτιστα μεταξύ τους και μπορεί να οδηγήσει σε πλεονασμό και άρα σε υποσύνολο χαρακτηριστικών με λιγότερη πληροφορία.

Η μέθοδος mMIFS-U διαλέγει με επαναληπτικό τρόπο χαρακτηριστικά που μεγιστοποιούν την υπό συνθήκη αμοιβαία πληροφορία μεταξύ του εκάστοτε χαρακτηριστικού και της κλάσης, με δεδομένο κάθε προηγούμενο χαρακτηριστικό που έχει επιλεγεί. Η χρήση της υπό συνθήκης αμοιβαίας πληροφορίας εξασφαλίζει ότι τα χαρακτηριστικά που επιλέγονται έχουν υψηλή συσχέτιση με τις κλάσεις, και παράλληλα χαμηλή συσχέτιση με τα υπόλοιπα ήδη επιλεγμένα χαρακτηριστικά.

Τα πειράματα που έγιναν δείχνουν ότι αυτή η εκδοχή του αλγορίθμου φέρνει καλύτερα αποτελέσματα από τον αρχικό MIFS αλγόριθμο και επιπλέον φέρνει καλύτερα αποτελέσματα και από μια άλλη παραλλαγή του, τον MIFS-U (Kwak N.,2002).

4.2.2 Χρήση Αμοιβαίας Πληροφορίας

Όπως είναι γνωστό από την Θεωρία Πληροφορίας για την αμοιβαία πληροφορία ισχύει:

$$I(C,Y)=I(Y,C)=H(C)-H(C|Y)=\sum_{c\in C}\int_{y}\log\left(\frac{p(c,y)}{P(c)\cdot p(y)}\right)dy \quad , \qquad 4.4$$

όπου $P(c)$ είναι η πιθανότητα της κλάσης C, y είναι το διάνυσμα των χαρακτηριστικών Y, και $p(c,y)$ είναι η από κοινού σ.π.π. των C και Y.

Ο στόχος της επιλογής χαρακτηριστικών είναι να ελαχιστοποιήσει την αβεβαιότητα στην πρόβλεψη της κλάσης C όταν είναι γνωστά τα χαρακτηριστικά Y. Η εκμάθηση ενός ταξινομητή είναι ουσιαστικά η μεγιστοποίηση της αμοιβαίας πληροφορίας. Όσον αφορά την επιλογή χαρακτηριστικών με σκοπό την ταξινόμηση, το ζητούμενο είναι αυτή η μεγιστοποίηση να γίνει με το μικρότερο δυνατόν υποσύνολο χαρακτηριστικών.

Η αμοιβαία πληροφορία ανάμεσα στις κλάσεις και τα χαρακτηριστικά χρησιμοποιείται πολύ συχνά ως μέτρο σύγκρισης στην επιλογή χαρακτηριστικών, αφού μετράει την γενική εξάρτηση μεταξύ δύο μεταβλητών και όχι την συσχέτισή τους, και καθορίζει το άνω φράγμα της επίδοσης της θεωρητικής ταξινόμησης.

Ο υπολογισμός της αμοιβαίας πληροφορίας ανάμεσα σε όλα τα υποψήφια

χαρακτηριστικά και τις κλάσεις είναι πρακτικά αδύνατος, γι'αυτό γίνεται χρήση άπληστων αλγορίθμων επιλογής χαρακτηριστικών. Ακόμη και η απλή προς τα εμπρός επιλογή είναι υπολογιστικά πολύ ακριβή.

Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα έχουν προταθεί διαφορετικές μέθοδοι από τον Battiti (Battiti R., 1994), τον Kwak και τον Choi (Kwak N., 1999), (Kwak N., 2002). Έστω ένα σύνολο S, το οποίο είναι το υποσύνολο των ήδη επιλεγμένων χαρακτηριστικών, και το σύνολο $X \setminus S$ είναι το υποσύνολο των χαρακτηριστικών που δεν έχουν επιλεχθεί ακόμα και είναι υποψήφια προς επιλογή. Για το χαρακτηριστικό που πρόκειται να επιλεχθεί $X_i \in X \setminus S$, πρέπει να ισχύει ότι το ποσό της πληροφορίας σχετικά με την κλάση C που δίνεται μόνο από το νέο χαρακτηριστικό X_i , χωρίς την πληροφορία που δίνεται από τα ήδη επιλεγμένα χαρακτηριστικά του υποσυνόλου S, πρέπει να είναι το μέγιστο ανάμεσα στα υποψήφια προς επιλογή χαρακτηριστικά του $X \setminus S$. Γι'αυτό η υπό-συνθήκη αμοιβαία πληροφορία $I(C, X_i|S)$ των C και X_i δεδομένου του S μεγιστοποιείται. Αντί του υπολογισμού του $I(C, X_i|S)$, ο Battiti , ο Kwak και ο Choi υπολόγισαν μόνο το $I(C, X_i)$ και $I(X_S, X_i)$, $X_S \in S$.

Η μέθοδος υπολογισμού για το $I(C, X_i|S)$ στον αλγόριθμο MIFS του Battiti είναι:

$$I_{Battiti}(C, X_i|S) = I(C, X_i) - \beta \cdot \sum_{X_s \in S} I(X_S, X_i) \quad . \qquad 4.5$$

Ο Kwak και ο Choi βελτίωσαν αυτήν την μέθοδο στον αλγόριθμο MIFS-U, κάνοντας την υπόθεση ότι η κλάση C δεν αλλάζει τον λόγο μεταξύ της εντροπίας του X_S και της αμοιβαίας πληροφορίας ανάμεσα στα X_S και X_i .

$$I_{Kwak}(C, X_i|S) = I(C, X_i) - \beta \cdot \sum_{X_s \in S} \frac{I(C, X_S)}{H(X_S)} \cdot I(X_S, X_i) \quad . \qquad 4.6$$

Και στις δύο σχέσεις, ο δεύτερος όρος του δεξιού μέρους, χρησιμοποιείται προκειμένου να καθορίσει το μέγεθος της περιττής πληροφορίας ανάμεσα στο υποψήφιο χαρακτηριστικό X_i και τα ήδη επιλεγμένα χαρακτηριστικά σε σχέση με την κλάση C. Τέλος η παράμετρος β, χρησιμοποιείται σαν παράγοντας προκειμένου να ελέγχει την ποινή που δέχονται τα χαρακτηριστικά λόγω της περιττής πληροφορίας που περιέχουν. Έχει αποδειχθεί (Peng H., 2005) ότι μια κατάλληλη τιμή για το β, για τη μεγιστοποίηση

της αμοιβαίας πληροφορίας στην διαδοχική προς τα εμπρός επιλογή, είναι $\beta=\frac{1}{|S|}$, όπου $|S|$ είναι το πλήθος των χαρακτηριστικών στο υποσύνολο S.

4.2.3 Υπό-Συνθήκη Αμοιβαία Πληροφορία

Ο αλγόριθμος επιλογής χαρακτηριστικών mMIFS-U βασίζεται στον ορισμό της υπό-συνθήκη αμοιβαίας πληροφορίας $I(C,X_i|X_S)$ ως την μείωση της αβεβαιότητας για την κλάση C και το χαρακτηριστικό X_i όταν το X_S είναι γνωστό. Ήτοι:

$$I(C,X_i|X_S)=H(X_i|X_S)-H(X_i|C,X_S) \quad . \tag{4.7}$$

Η παραπάνω ικανοποιεί και τον κανόνα αλυσίδας:

$$I(C,X_i,X_S)=I(C,X_S)-I(C,X_i|X_S) \quad . \tag{4.8}$$

Για όλα τα υποψήφια προς επιλογή χαρακτηριστικά στον αλγόριθμο το $I(C,X_S)$ είναι κοινό και άρα δεν χρειάζεται να συγκριθεί. Έτσι ο αλγόριθμος βρίσκει το χαρακτηριστικό που μεγιστοποιεί το $I(C,X_i|X_S)$.

Η υπό-συνθήκη αμοιβαία πληροφορία $I(C,X_i|X_S)$ μπορεί εναλλακτικά να γραφτεί ως εξής:

$$I(C,X_i|X_S)=I(C,X_i)-[I(X_i,X_S)-I(X_i,X_S|C)] \tag{4.9}$$

Αυτό αποδεικνύεται ακολούθως:

$$\begin{aligned} I(C,X_i)-[I(X_i,X_S)-I(X_i,X_S|C)] &= H(C)-H(C|X_i)-[H(X_i)-H(X_i|X_S)]+H(X_i|C)-H(X_i|X_S,C) \\ &= H(C)-H(C|X_i)-H(X_i)+H(X_i|X_S)+H(X_i|C)-H(X_i|X_S,C) \\ &= H(X_i|X_S)-H(X_i|X_S,C)+H(C)-H(C|X_i)-[H(X_i)-H(X_i|C)] \\ &= I(C,X_i)-I(C,X_i)+H(X_i|X_S)-H(X_i|X_S,C) \\ &= I(C,X_i|X_S). \end{aligned} \tag{4.10}$$

Ο λόγος μεταξύ της αμοιβαίας πληροφορίας του υποψήφιου χαρακτηριστικού X_i και του επιλεγμένου χαρακτηριστικού X_S και της εντροπίας του X_S είναι ένα μέτρο συσχέτισης (γνωστό ως συντελεστής της αβεβαιότητας (coefficient of uncertainty)) μεταξύ των X_i και X_S (Cover T., 1991).

$$CU_{X_i,X_s}=\frac{I(X_i,X_s)}{H(X_s)}=\left(1-\frac{H(X_s|X_i)}{H(X_s)}\right) \quad , \tag{4.11}$$

για τον οποίο ισχύει $0\le CU_{X_i,X_s}\le 1$. Ο συντελεστής ισούται με 0 μόνο στην περίπτωση που τα X_i και X_s είναι ανεξάρτητα.

Υποθέτοντας ότι στην περίπτωση που υπό συνθήκη είναι η κλάση C, δεν αλλάζει ο λόγος της εντροπίας του X_s και της αμοιβαίας πληροφορίας των X_i και X_s , ισχύει το παρακάτω:

$$\frac{H(X_s|C)}{I(X_i,X_s|C)}=\frac{H(X_s)}{I(X_i,X_s)} \quad . \tag{4.12}$$

Και άρα η υπό-συνθήκη αμοιβαία πληροφορία $I(C,X_i|X_s)$ γράφεται και ως:

$$I(C,X_i|X_s)=I(C,X_i)-CU_{X_i,X_s}\cdot I(C,X_s) \quad . \tag{4.13}$$

Αυτό αποδεικνύεται λόγω της υπόθεσης που κάναμε (4.12) και του ορισμού του συντελεστή αβεβαιότητας (4.11) αφού θα ισχύει:

$$I(C,X_i|X_s)=CU_{X_i,X_s}\cdot H(X_s|C) \quad . \tag{4.14}$$

Μετά από πράξεις και κάνοντας χρήση αυτού του συμπεράσματος (4.14) και της προηγούμενης εξίσωσης που αποδείχθηκε για την υπό-συνθήκη αμοιβαία πληροφορία (4.6) βγαίνει άμεσα το ζητούμενο.

Από την εξίσωση 4.13 της υπό-συνθήκη αμοιβαίας πληροφορίας φαίνεται ότι το δεύτερο μέρος είναι η σταθμισμένη αμοιβαία πληροφορία $I(C,X_s)$ με το ισοδύναμο του βάρους για αυτήν να είναι ο συντελεστής αβεβαιότητας CU_{X_i,X_s} . Για απλούστευση των πράξεων έχει προταθεί η ακόλουθη τροποποίηση $\tilde{I}(C,X_i|S)$ για την εκτίμηση του $I(C,X_i|S)$:

$$\tilde{I}(C,X_i|S)=I(C,X_i)-\max_{X_s\in S}\left[CU_{X_i,X_s}\cdot I(C,X_s)\right] \quad . \tag{4.15}$$

Αυτό σημαίνει ότι το καλύτερο προς επιλογή χαρακτηριστικό στο επόμενο βήμα της διαδοχικής προς τα εμπρός αναζήτησης είναι αυτό που μεγιστοποιεί την ακόλουθη ποσότητα:

$$X^+=\arg\max_{X_i\in X\setminus S}\left\{I(C,X_i)-\max_{X_s\in S}\left[CU_{X_i,X_s}\cdot I(C,X_s)\right]\right\} \quad . \tag{4.16}$$

4.2.4 Αλγόριθμος mMIFS-U

Ο αλγόριθμος διαδοχικής προς τα εμπρός αναζήτησης mMIFS-U βασίζεται στην εκτίμηση της ποσότητας της υπό-συνθήκη αμοιβαίας πληροφορίας, όπως αυτή υπολογίζεται από την εξίσωση 4.15.

Algorithm 2. mMIFS-U

Initialization: Set S=”empty set”, set X=”initial set of all D features”. (αρχικοποίηση: S=κενό σύνολο, X=σύνολο με όλα τα χαρακτηριστικά.)

Pre-computation: For all features $X_i \in X$ compute $I(C, X_i)$. (Εκ των προτέρων υπολογισμός: Για όλα τα χαρακτηριστικά $X_i \in X$ υπολόγισε την $I(C, X_i)$.)

Selection of the first feature: Find feature $X^* \in X$ that maximizes $I(C, X_i)$;(Επιλογή πρώτου χαρακτηριστικού: Εντοπισμός του χαρακτηριστικού $X^* \in X$ το οποίο μεγιστοποιεί την $I(C, X_i)$;)

set $X = X \setminus \{X^*\}$, $S = \{X^*\}$.(ορισμός των $X = X \setminus \{X^*\}$, $S = \{X^*\}$.)

Greedy feature selection: Repeat until the desired number of features is selected. (Άπληστη επιλογή χαρακτηριστικών: επανάληψη μέχρι να επιλεγεί ο επιθυμητός αριθμός χαρακτηριστικών.)

Computation of entropy: For all $X_S \in S$ compute entropy $H(X_S)$,if it is not already available. (Υπολογισμός της εντροπίας: Για όλα τα $X_S \in S$ υπολόγισε την εντροπία $H(X_S)$ στη περίπτωση που δεν δίνεται ήδη.)

Computation of the MI between features: For all pairs of features X_i, X_S with $X_i \in X$, $X_S \in S$ compute $I(X_i, X_S)$, if it is not yet available. (Υπολογισμός της αμοιβαίας πληροφορίας μεταξύ όλων των χαρακτηριστικών: Για όλα τα χαρακτηριστικά X_i, X_S όπου $X_i \in X$, $X_S \in S$ υπολόγισε την $I(X_i, X_S)$, εάν δεν δίνεται ήδη.)

Selection of the next feature: Find feature $X^+ \in X$ according to formula 4.16. (Επιλογή του επόμενου χαρακτηριστικού: Εύρεση του επόμενου χαρακτηριστικού $X^+ \in X$ σύμφωνα με τη σχέση 4.16.)

Set $X = X \setminus \{X^+\}$, $S = S \cup \{X^+\}$.(ορισμός $X = X \setminus \{X^+\}$, $S = S \cup \{X^+\}$.

4.2.5 Πειραματικά Αποτελέσματα Αλγορίθμου mMIFS-U

Ο αλγόριθμος επιλογής χαρακτηριστικών εφαρμόστηκε με επιτυχία σε διάφορες κατηγορίες προβλημάτων, συμπεριλαμβανομένου και του προβλήματος κατηγοριοποίησης κειμένου (Forman G., 2003). Το πρόβλημα της κατηγοριοποίησης κειμένου (text categorization (TC)), το οποίο είναι γνωστό και ως ταξινόμηση κειμένου (text classification), αποτελεί τη διαδικασία ανάθεσης κειμένων τα οποία είναι γραμμένα σε φυσική γλώσσα, σε μια ή περισσότερες θεματικές ενότητες ανάλογα με το περιεχόμενό τους (κλάσεις), που ανήκουν σε ένα σύνολο C με $|C|$ το πλήθος των κλάσεων, δηλαδή $C=\{c_1,..,c_{|C|}\}$, και οι οποίες είναι προκαθορισμένες. Η κατασκευή ενός ταξινομητή κειμένων βασίζεται σε ένα σύνολο κειμένων εκπαίδευσης τα οποία ήδη ανήκουν σε κάποια προκαθορισμένη κλάση στο σύνολο C. Στην κατηγοριοποίηση κειμένων συνήθως χρησιμοποιείται μια αναπαράσταση που λέγεται bag-of-words, και κατά την οποία κάθε θέση στο διάνυσμα αναπαράστασης αποτελεί ένα χαρακτηριστικό και αντιστοιχεί σε μια συγκεκριμένη λέξη, πιο συγκεκριμένα στον αριθμό των εμφανίσεών της. Αυτός ο τρόπος αναπαράστασης οδηγεί στην δημιουργία χώρου χαρακτηριστικών πολύ υψηλών διαστάσεων, με αποτέλεσμα να είναι πολύ μεγάλοι για τις συμβατικές μεθόδους ταξινόμησης. Σε αυτόν τον τύπο προβλημάτων πιο συχνά για την ελάττωση των διαστάσεων του χώρου χρησιμοποιείται η επιλογή χαρακτηριστικών η οποία μπορεί να βασίζεται σε πολλά κριτήρια, και πιο συχνή είναι η χρήση μεθόδων βασισμένων σε φίλτρα.

Στις δοκιμές του αλγορίθμου χρησιμοποιήθηκαν οι τρεις διαδοχικές προς τα εμπρός μέθοδοι MIFS, MIFS-U, Mmifs-u που παρουσιάστηκαν προηγουμένως. Σε αυτήν την εφαρμογή οι μέθοδοι χρησιμοποιήθηκαν για την ελάττωση του μεγέθους του συνόλου του λεξιλογίου $V=\{w_1,..,w_{|V|}\}$, το οποίο έχει μέγεθος $|V|$ και ουσιαστικά αποτελείται από τις διαφορετικές λέξεις που εμφανίστηκαν μέσα στα κείμενα εκπαίδευσης. Μετά έγινε ταξινόμηση των κειμένων με τους ταξινομητές Naïve Bayes (McCallum A., 1998) που βασίζεται στο πολυωνυμικό μοντέλο, SVM (support vector machine-μηχανή διανυσματικής υποστήριξης) (Joachims T., 1998) και k-NN (k-Nearest Neighbor-k πλησιέστεροι γείτονες).

Το σύνολο δεδομένων που εξετάστηκε στα πειράματα για την αξιολόγηση του αλγορίθμου, είναι ένα πολύ συχνά χρησιμοποιούμενο, το Reuters-21578 το οποίο περιέχει συνολικά 21578 κείμενα. Όλα τα κείμενα προεπεξεργάστηκαν και τους αφαιρέθηκαν όλοι οι μη αλφαβητικοί χαρακτήρες, όπως τελείες, κόμματα, παρενθέσεις κ.λ.π. Επίσης όλα τα κεφαλαία γράμματα έγιναν μικρά, και μετά αφαιρέθηκαν όλες οι λέξεις που περιείχαν ψηφία ή μη αλφαριθμητικούς χαρακτήρες. Στη συνέχεια αφαιρέθηκαν και οι stop-words (αναφέρεται ότι η stop-words list είναι μια λίστα με λέξεις που είναι πολύ συχνά χρησιμοποιούμενες και δεν συμβάλουν στην εύρεση του είδους του κειμένου όπως π.χ. “και”, ”αλλά”, “όταν” κ.τ.λ.). Τέλος οι λέξεις αντικαταστάθηκαν με την μορφολογική τους ρίζα, και όσες από τις λέξεις εμφανίζονταν λιγότερο από τρεις φορές αφαιρέθηκαν. Το τελικό μέγεθος του λεξιλογίου ελαττώθηκε στις 7487 λέξεις. Χρησιμοποιήθηκε η τεχνική ModApte, κατά την οποία τα κείμενα του Reuters χωρίστηκαν σε δύο κατηγορίες. Η πρώτη είναι η κατηγορία των κειμένων εκπαίδευσης στην οποία εντάχθηκαν 9603 κείμενα, και η δεύτερη είναι η κατηγορία των κειμένων δοκιμής στην οποία εντάχθηκαν 3299 κείμενα. Αυτά τα κείμενα είναι χωρισμένα σε 135 κλάσεις οι οποίες αφορούν περισσότερο την οικονομία. Για τις συγκρίσεις που έγιναν, χρησιμοποιήθηκαν μόνο 90 κλάσεις από τις 135, και επιλέχθηκαν έτσι ώστε να υπάρχει τουλάχιστον ένα κείμενο εκπαίδευσης και ένα κείμενο δοκιμής σε αυτές.

Όλες οι μέθοδοι επιλογής χαρακτηριστικών εξετάστηκαν σε συνδυασμό με κάθε ένα από τους ακόλουθους ταξινομητές:

Naïve Bayes: Χρησιμοποιήθηκε το πολυωνυμικό μοντέλο. Η κλάση στην οποία κατατάσσεται το κείμενο d είναι αυτή που μεγιστοποιεί την εκ των υστέρων πιθανότητα (posterior probability) $P(c_j|d)$ καθεμιάς κλάσης για το κείμενο δοκιμής,

$$P(c_j|d) \propto P(c_j)\cdot\prod_{v}^{|V|} P(w_v|c_j)^{N_{iv}} \quad . \qquad 4.17$$

Στην εξίσωση ο όρος $P(c_j)$ είναι η εκ των προτέρων πιθανότητα (prior probability) της κλάσης c_j , ο όρος $P(w_v|c_j)$ είναι η πιθανότητα μια λέξη που επιλέχθηκε τυχαία σε ένα κείμενο της κλάσης c_j να είναι η w_v , και ο όρος N_{iv} είναι ο αριθμός εμφανίσεων της λέξης w_v σε ένα κείμενο d. Οι πιθανότητες των κλάσεων και των λέξεων εξομαλύνθηκαν

με Μπεϊζιανή εκτίμηση (Bayesian estimate) με την prior πιθανότητα των λέξεων, και Laplace εκτίμηση (Laplace estimate) αντίστοιχα.

Linear Support Vector Machine: Η μέθοδος SVM ορίζεται στον διανυσματικό χώρο, πάνω στον οποίο το πρόβλημα ταξινόμησης πρέπει να βρει το υπερεπίπεδο διαχωρισμού, το οποίο χωρίζει καλύτερα τα "σημεία των δεδομένων" της μιας κλάσης από την άλλη. Στην περίπτωση που τα δεδομένα είναι γραμμικά διαχωρίσιμα δηλαδή είναι πλήρως διαχωρίσιμα, το υπερεπίπεδο διαχωρισμού είναι αυτό με το οποίο επιτυγχάνεται η μεγιστοποίηση του κενού ανάμεσα στις δύο κλάσεις. Στην αναπαράσταση των κειμένων χρησιμοποιήθηκε η κανονικοποιημένη συχνότητα των λέξεων:

$$\text{tfidf}(w_i, d_j) = n(w_i, d_j) \cdot \log\left(\frac{|D|}{n(w_i)}\right) \quad , \qquad 4.18$$

όπου στην εξίσωση ο όρος $n(w_i)$ είναι ο αριθμός των κειμένων στο σύνολο D, τα οποία περιέχουν την λέξη w_i τουλάχιστον μια φορά, ενώ ο όρος $n(w_i, d_j)$ είναι ο αριθμός των εμφανίσεων της λέξης w_i στο κείμενο d_j .

K-Nearest Neighbor. Δοσμένου ενός κειμένου, ο αλγόριθμος κοιτάει τους κοντινότερους γείτονές του (κείμενα εκμάθησης), και χρησιμοποιεί τους k πιο κοντινούς σε αυτό, για να αναγνωρίσει σε ποια κλάση είναι οι περισσότεροι από αυτούς, και έτσι να κατηγοριοποιήσει και το νέο κείμενο. Η ομοιότητα που έχει το νέο κείμενο που ταξινομείται με καθένα από τους πλησιέστερους γείτονές του χρησιμοποιείται ως μέτρο βάρους. Δεδομένου ότι κάθε ένα από τα γειτονικά κείμενα ανήκει σε κάποια κλάση, τα επιμέρους αθροίσματα των βαρών που αντιστοιχήθηκαν στα τα γειτονικά κείμενα για κάθε κλάση, χρησιμοποιούνται για την αύξουσα ταξινόμηση των κλάσεων. Για την αναπαράσταση των κειμένων χρησιμοποιήθηκε η κανονικοποιημένη συχνότητα των λέξεων από την παραπάνω εξίσωση (4.18).

Προκειμένου να αξιολογηθεί η ακρίβεια της ταξινόμησης πολλαπλών κλάσεων (πολυκατηγορική ταξινόμηση), χρησιμοποιήθηκαν δύο τυπικά μέτρα αξιολόγησης πολυκατηγορικής ταξινόμησης. Το πρώτο είναι το μέτρο precision (γνωστό και ως μέτρο θετικής προγνωστικής αξίας) και το δεύτερο είναι το μέτρο recall (γνωστό και ως μέτρο ανάκλησης). Και και τα δύο μέτρα έγιναν micro-averaged. Πιο συγκεκριμένα οι εξισώσεις των μέτρων precision και recall για αυτήν την μέθοδο δίνονται παρακάτω:

$$\hat{\pi}_{mic} = \frac{\sum_{j=1}^{|C|} TP_j}{\sum_{j=1}^{|C|} (TP_j + FP_j)} \quad , \qquad 4.19$$

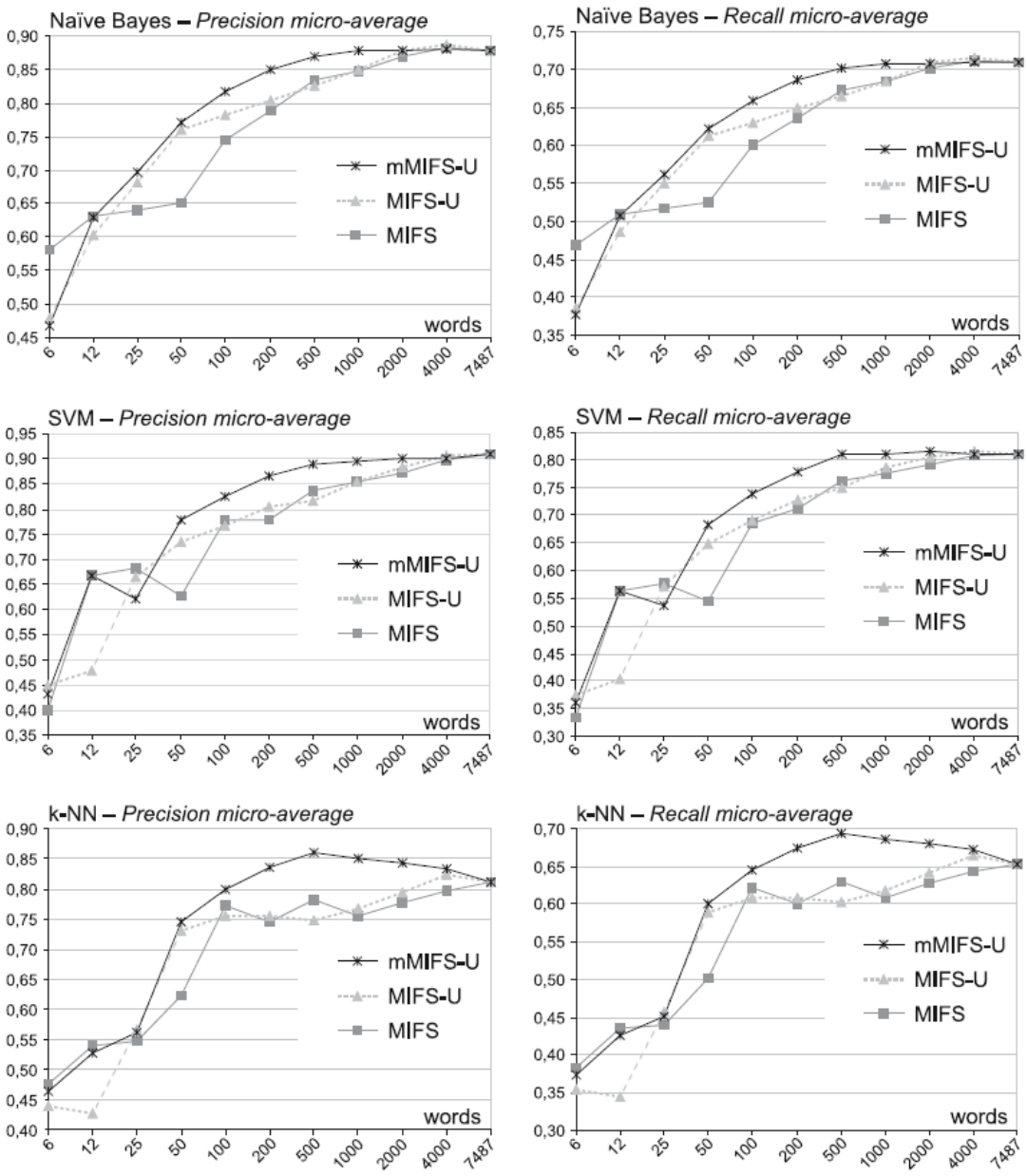
$$\hat{\rho}_{mic} = \frac{\sum_{j=1}^{|C|} TP_j}{\sum_{j=1}^{|C|} (TP_j + FN_j)} \quad , \qquad 4.20$$

Όπου ο όρος TP_j (true positive) αντιπροσωπεύει τα κείμενα τα οποία ταξινομήθηκαν σωστά στην κλάση c_j (δηλαδή ανήκαν σε αυτήν) και αντίστοιχα ο όρος FP_j (false positive) αυτά που ταξινομήθηκαν λανθασμένα στην κλάση c_j ενώ κανονικά ανήκαν σε κάποια άλλη. Τέλος ο όρος FN_j (false negative) αναφέρεται στα κείμενα τα οποία ταξινομήθηκαν λανθασμένα σε κάποια άλλη κλάση από την c_j , ενώ κανονικά ανήκαν σε αυτήν.

Υπάρχουν δύο κύριες παραλλαγές για το πρόβλημα της πολυκατηγορικής ταξινόμησης. Η πρώτη είναι η hard classification και η δεύτερη η ranking classification (Sebastiani F., 2002). Η hard classification αντιστοιχεί σε κάθε ζεύγος κειμένου-κλάσης d, c_j την δυαδική μεταβλητή YES (ναι) ή NO (όχι) ανάλογα με το αποτέλεσμα του ταξινομητή. Από την άλλη η ranking classification αντιστοιχεί στο κάθε ζεύγος κειμένου-κλάσης d, c_j μια πραγματική τιμή $\phi(d, c_j)$, η οποία συμβολίζει την απόφαση του ταξινομητή για το γεγονός ότι $d \in c_j$. Μετά ταξινομούνται σε αύξουσα σειρά όλες οι κλάσεις για το κείμενο d, σύμφωνα με την τιμή $\phi(d, c_j)$ που τους δόθηκε, και οι καλύτερες τ_j κλάσεις επιλέγονται για αυτό το κείμενο (τ_j είναι το κατώφλι (threshold) για την κλάση c_j). Για την εύρεση του τ_j έχουν προταθεί πολλοί αλγόριθμοι.

Στο σχήμα που ακολουθεί παρακάτω (Σχήμα 4.1) απεικονίζονται γραφικά σε έξι διαγράμματα οι επιδώσεις της ταξινόμησης με την μέθοδο mMIFS-U (η οποία είναι αυτή που αναλύθηκε και προτάθηκε προηγουμένως) σε σύγκριση με άλλες δύο μεθόδους επιλογής χαρακτηριστικών, τις MIFS-U και MIFS, για το σύνολο δεδομένων Reuters με 90 κλάσεις, στο οποίο έγινε χρήση των μεθόδων Apte split και Rcut-thresholding (οι μέθοδοι RCut, PCut και SCut (Yang Y., 2001) είναι ευρέως διαδεδομένες και χρησιμοποιούνται

πολύ συχνά στην ταξινόμηση).



Σχήμα 4.1 Επίδοση του ταξινομητή στα δεδομένα Reuters

Στις γραφικές παραστάσεις της αριστερά στήλης χρησιμοποιείται η precision micro-averaged μέθοδος αξιολόγησης της ταξινόμησης, ενώ σε αυτές της δεξιά στήλης, χρησιμοποιείται η recall micro-averaged μέθοδος. Ακόμη, στα διαγράμματα που είναι στην πρώτη γραμμή απεικονίζονται αυτά τα δύο μέτρα για τον ταξινομητή Naïve Bayes, ενώ στην δεύτερη γραμμή απεικονίζονται για τον ταξινομητή SVM (Support Vector Machine) και τέλος στην τρίτη γραμμή απεικονίζονται για τον k-NN (k-Nearest Neighbour). Τέλος επισημαίνεται ότι ο οριζόντιος άξονας σε κάθε διάγραμμα αντιστοιχεί

στο μέγεθος του λεξιλογίου που προέκυψε από την επιλογή χαρακτηριστικών για τα κείμενα που ταξινομούνται, ενώ ο κάθετος άξονας αντιστοιχεί στην ακρίβεια που επιτυγχάνεται στην ταξινόμηση.

Από την διαδικασία προέκυψε ότι η μέθοδος thresholding που χρησιμοποιείται παίζει καταλυτικό ρόλο όσον αφορά το τελικό αποτέλεσμα στην ταξινόμηση. Παρ'όλα αυτά δεν είναι εύκολο να επιλεγεί κάποια ως βέλτιστη. Τελικά χρησιμοποιήθηκε η RCut thresholding, η οποία ταξινομεί με αύξουσα σειρά τις κλάσεις για το κάθε κείμενο και βάζει την ταμπέλα ”YES” στις τ καλύτερες από αυτές, δηλαδή εκείνες με την υψηλότερη βαθμολογία. Όσον αφορά το threshold τ, υπάρχει μια παγκόσμια χρησιμοποιούμενη τιμή, η οποία είναι ουσιαστικά μια ακέραια τιμή για την οποία ισχύει $1 \leq \tau \leq |C|$ και χρησιμοποιείται για όλες τις κλάσεις. Εδώ χρησιμοποιήθηκε ως τιμή για το threshold ο μέσος όρος των κλάσεων που αντιστοιχούνται σε ένα κείμενο. Προκειμένου να αξιολογηθεί η τιμή για το κατώφλι, χρησιμοποιήθηκε ολόκληρο το σύνολο εκπαίδευσης.

Οι ταξινομητές Naïve Bayes και k-NN χρησιμοποιούνται πολύ συχνά στη ranking classification. Αντίθετα η SVM χρησιμοποιείται στη hard classification αφού σε κάθε κείμενο αντιστοιχείται μόνο μια κλάση, η οποία ξεχωρίζει σε σχέση με τις υπόλοιπες. Είναι πολύ πιθανό μάλιστα σε κάποιο κείμενο να μην αντιστοιχιστεί καμία κλάση. Σε αυτήν την περίπτωση το κείμενο επανατίθεται σε κάποια άλλη κλάση, η οποία μάλιστα είναι η καλύτερη δυνατή σύμφωνα με τον ταξινομητή SVM. Αυτή η τεχνική βελτιώνει το αποτέλεσμα της ταξινόμησης.

4.2.6 Συμπεράσματα

Παραπάνω αναλύθηκε ένας καινοτόμος αλγόριθμος διαδοχικής προς τα εμπρός επιλογής, ο οποίος βασίζεται στην εκτίμηση της υπο-συνθήκη αμοιβαίας πληροφορίας ανάμεσα στο υποψήφιο προς επιλογή χαρακτηριστικό και τις κλάσεις, δεδομένου ενός υποσυνόλου ήδη επιλεγμένων χαρακτηριστικών. Τα πειραματικά αποτελέσματα σε δεδομένα κειμένων που προήλθαν και από τις τρεις μεθόδους mMIFS-U, MIFS-U,και MIFS, σε συνδυασμό με τη χρήση και των τριών ταξινομητών Naïve Bayes, SVM και k-NN, καθώς και την αξιολόγησή τους και από τα δύο κριτήρια precision και recall αξιολόγησης

της ακρίβειας ταξινόμησης, δείχνουν ότι ο νέος τροποποιημένος MIFS-U αλγόριθμος διαδοχικής προς τα εμπρός επιλογής, mMIFS-U, έχει καλύτερα αποτελέσματα στην ταξινόμηση των δεδομένων Reuters αφού επιτυγχάνει μεγαλύτερη ακρίβεια για τα περισσότερα μεγέθη λεξιλογίου και ίση για λίγα από αυτά ανεξάρτητα του ταξινομητή.

4.3 NMIFS, GAMIFS Μέθοδοι Επιλογής Χαρακτηριστικών

4.3.1 Εισαγωγή

Οι αλγόριθμοι MIFS, MIFS-U, mRMR (Peng H., 2005) είναι όλοι αυξητικοί αλγόριθμοι οι οποίοι διαλέγουν ένα χαρακτηριστικό κάθε φορά. Σε κάθε επανάληψη κάποιο συγκεκριμένο κριτήριο μεγιστοποιείται το οποίο αφορά πάντα ένα χαρακτηριστικό και δεν αναφέρεται ποτέ σε κάποια ομάδα χαρακτηριστικών. Ωστόσο σε πολλά προβλήματα ταξινόμησης μια ομάδα χαρακτηριστικών που δρουν ταυτόχρονα μπορεί να είναι έχει υψηλή συσχέτιση με τις κλάσεις, ενώ κάθε χαρακτηριστικό μόνο του να είναι άσχετο. Εάν δεν λαμβάνεται καθόλου υπόψιν η επίδραση ομάδων χαρακτηριστικών στο πρόβλημα, σε αυτήν την περίπτωση όλα τα εκάστοτε χαρακτηριστικά θεωρούνται άσχετα με αυτό. Αυτό το φαινόμενο καλείται από την βιολογία, επίσταση (epistasis) (Mitchell M., 1996). Κατά συνέπεια, οι αλγόριθμοι επιλογής χαρακτηριστικών που αξιολογούν ένα χαρακτηριστικό την φορά δεν θα μπορέσουν να επιλέξουν το βέλτιστο υποσύνολο χαρακτηριστικών εάν η συνάρτηση ταξινόμησης εξαρτάται από ένα συνδυασμό δύο ή περισσότερων χαρακτηριστικών.

Ο αλγόριθμος NMIFS-Normalized Mutual Information for Feature Selection (Pablo A., 2009) που θα αναλυθεί παρακάτω είναι μια βελτίωση των μεθόδων MIFS, MIFS-U και mRMR. Λόγω του ότι είναι αυξητικός αλγόριθμος, είναι γρήγορος και αποτελεσματικός, αλλά η απόδοσή του πέφτει όταν στο πρόβλημα που επιλύεται υπάρχουν ομάδες χαρακτηριστικών που είναι σχετικές με αυτό, αλλά το καθένα χαρακτηριστικό μόνο του είναι άσχετο. Για αυτόν τον λόγο, προτείνεται και μια δεύτερη μέθοδος, η GAMIFS – Genetic Algorithm guided by Mutual Information for Feature Selection, η οποία είναι μια υβριδική μέθοδος φίλτρου/περιτυλίγματος που συνδυάζει έναν γενετικό αλγόριθμο με

τον NMIFS. Η μέθοδος GAMIFS μπορεί να βρει και μεμονωμένα χαρακτηριστικά που είναι σχετικά με τις κλάσεις, αλλά και ομάδες χαρακτηριστικών που είναι σχετικές.

4.3.2 Χρήση Αμοιβαίας Πληροφορίας

Αναφέρονται οι εξισώσεις της αμοιβαίας πληροφορίας για συνεχείς και διακριτές τυχαίες μεταβλητές. Έστω X, Y συνεχείς τυχαίες μεταβλητές με από κοινού σ.π.π. $p(x,y)$ και περιθώριες κατανομές $p(x)$ και $p(y)$ αντίστοιχα. Η αμοιβαία πληροφορία είναι:

$$I(X;Y)=\iint p(x,y)\cdot\log\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right)dxdy \quad . \tag{4.21}$$

Έστω τώρα X, Y διακριτές τυχαίες μεταβλητές που έχουν από κοινού σ.μ.π. $p(x,y)$ και περιθώριες κατανομές $p(x)$ και $p(y)$ αντίστοιχα, τότε η αμοιβαία πληροφορία είναι

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right) \quad . \tag{4.22}$$

Επειδή το κριτήριο επιλογής βασίζεται στον υπολογισμό της αμοιβαίας πληροφορίας ανάμεσα στα χαρακτηριστικά και τις κλάσεις, οι μέθοδοι επιλογής χαρακτηριστικών που βασίζονται στην αμοιβαία πληροφορία είναι πολύ ευαίσθητες στον υπολογισμό της. Ο υπολογισμός της απαιτεί την εκτίμηση τις σ.π.π. ή τις εντροπίες των δεδομένων. Ακόμη και μικρά σφάλματα μειώνουν την αποδοτικότητα της μεθόδου. Μια τεχνική για την προσέγγιση των σ.π.π. είναι η χρήση πυρήνων (Fukunaga K., 1990), η οποία συνδυάζει βασικές συναρτήσεις. Αυτές οι μέθοδοι λειτουργούν θέτοντας πάνω σε κάθε σημείο του χαρακτηριστικού μια βασική συνάρτηση, η οποία συνήθως είναι η Gauss. Η τελική εκτίμηση της σ.π.π. γίνεται από το άθροισμα όλων των συναρτήσεων. Η ποιότητα αυτής της μεθόδου είναι υψηλή καθώς επίσης και το κόστος όμως. Μια άλλη προσέγγιση είναι η χρήση ιστογραμμάτων (Fukunaga K., 1990), κατά την οποία ο χώρος χωρίζεται σε ίσα διαστήματα και μετρούνται τα στοιχεία που βρίσκονται σε κάθε διάστημα. Η επιλογή πλήθους των διαστημάτων είναι η κύρια αιτία σφαλμάτων, διότι πολύ μεγάλα διαστήματα μπορεί να εξομαλύνουν στην σ.π.π., ενώ πολύ μικρά δεν αντανακλούν την καμπύλη ομοιόμορφα και τείνουν να την κάνουν πολύ τραχιά, με πολλές διακυμάνσεις. Οι μέθοδοι που χρησιμοποιούν ιστογράμματα υπολογιστικά έχουν μικρό κόστος, αλλά

ενδέχεται να παράξουν μεγάλο σφάλμα εκτίμησης.

Ο Fraser (Fraser A. M., 1986) πρότεινε μια γρήγορη και αποδοτική μέθοδο για την εκτίμηση της αμοιβαίας πληροφορίας μεταξύ δύο τυχαίων μεταβλητών, η οποία κάνει χρήση προσαρμοσμένων ιστογραμμάτων. Ο αλγόριθμος είναι ένα ενδιάμεσο επίπεδο ανάμεσα στις μεθόδους που χρησιμοποιούν πυρήνες και τις μεθόδους που χρησιμοποιούν ιστογράμματα, καθώς η ακρίβεια των αποτελεσμάτων της μεθόδου του είναι καλύτερη από αυτή των κοινών ιστογραμμάτων, αλλά παράλληλα το ίδιο γρήγορη και αποδοτική.

Συνήθως οι μέθοδοι επιλογής χαρακτηριστικών που βασίζονται στην αμοιβαία πληροφορία εφαρμόζονται σε συνεχή χαρακτηριστικά. Παρ'όλα αυτά στα πραγματικά προβλήματα συνήθως συναντώνται και συνεχή και διακριτά χαρακτηριστικά. Γι'αυτό η μέθοδος που χρησιμοποιείται για την εκτίμηση της αμοιβαίας πληροφορίας πρέπει να είναι διαφορετική σε κάθε περίπτωση. Στον συγκεκριμένο αλγόριθμο χρησιμοποιείται μια επεκταμένη εκδοχή του αλγόριθμου του Fraser (Tesmer M., 2004) για συνεχείς μεταβλητές, ενώ για διακριτές χρησιμοποιούνται πίνακες συνάφειας.

Ο αλγόριθμος MIFS χρησιμοποιεί ως κριτήριο επιλογής την εξίσωση:

$$I(C;f_i)-\beta\cdot\sum_{f_s\in S} I(f_S;f_i) \quad . \tag{4.23}$$

Ο αλγόριθμος MIFS-U έχει την ακόλουθη παραλλαγή του κριτηρίου:

$$I(C;f_i)-\beta\cdot\sum_{f_s\in S}\frac{I(C;f_S)}{H(f_S)}\cdot I(f_S;f_i) \quad , \tag{4.24}$$

όπου $H(f_S)=-\sum_{f_s\in S} P(f_S)\cdot\log(P(f_S))$ είναι η εντροπία.

Το κριτήριο της μεθόδου mRMR, στον αυξητικό αλγόριθμο πρώτης τάξης είναι:

$$I(C;f_i)-\frac{1}{|S|}\cdot\sum_{f_s\in S} I(f_S;f_i) \quad , \tag{4.25}$$

όπου $|S|$ είναι η πληθικότητα του συνόλου S. Παρατηρείται ότι το κριτήριο της mRMR είναι το ίδιο με του MIFS όταν το β προσαρμοστεί ως $\beta=\frac{1}{|S|}$.

Σε όλες τις παραπάνω εξισώσεις ο αριστερός όρος $I(C;f_i)$ μετράει την σχετικότητα του χαρακτηριστικού που πρόκειται να προστεθεί στο υποσύνολο με τις κλάσεις, και ο δεξιός μετράει τον πλεονασμό του i-οστού χαρακτηριστικού σε σχέση με τα ήδη επιλεγμένα χαρακτηριστικά.

Το πρόβλημα των μεθόδων MIFS και MIFS-U είναι ότι ο αριστερά και ο δεξιά όρος των εξισώσεων (4.23) και (4.24) δεν είναι συγκρίσιμοι, αφού ο δεξιός όρος των εξισώσεων είναι ένα σωρευτικό άθροισμα το οποίο θα μεγαλώνει σε μέγεθος σε σχέση με τον αριστερό όρο καθώς το πλήθος των ήδη επιλεγμένων χαρακτηριστικών θα αυξάνεται. Όταν ο αριστερά όρος των εξισώσεων γίνεται αμελητέος σε σχέση με τον δεξιά, ο αλγόριθμος επιλογής χαρακτηριστικών αναγκάζεται να διαλέξει χαρακτηριστικά που είναι μη πλεονάζοντα σε σχέση με τα ήδη επιλεγμένα. Αυτό μπορεί να προκαλέσει την επιλογή άσχετων με το πρόβλημα χαρακτηριστικών νωρίτερα από ότι σχετικών. Το πρόβλημα λύνεται εν μέρη με το κριτήριο του αλγορίθμου mRMR (4.25) ο οποίος διαιρεί το άθροισμα με το πλήθος των χαρακτηριστικών του συνόλου S, $|S|$. Ακόμα ένα μειονέκτημα είναι ότι ο MIFS και ο MIFS-U βασίζονται στην παράμετρο β για να ελέγξουν την ποινή του πλεονασμού, αλλά στην πραγματικότητα η βέλτιστη τιμή της παραμέτρου εξαρτάται από το πρόβλημα. Επιπλέον, το κριτήριο του MIFS-U βασίζεται στην υπόθεση ότι ο λόγος της εντροπίας $H(f_S)$, του f_S και της αμοιβαίας πληροφορίας $I(f_i;f_S)$, ανάμεσα στο f_S και το f_i δεν αλλάζει όταν έχουμε υπό-συνθήκη την κλάση C, το οποίο όμως ισχύει μόνο για ομοιόμορφες κατανομές πιθανότητας.

Αναλυτικότερα, η αμοιβαία πληροφορία μπορεί να γραφτεί συναρτήσει της εντροπίας ως:

$$I(f_i;f_S)=H(f_i)-H(f_i|f_S)=H(f_S)-H(f_S|f_i) \quad . \qquad 4.26$$

Από την παραπάνω σχέση φαίνεται ότι η αμοιβαία πληροφορία μπορεί να πάρει τιμές στο διάστημα $0\le I(f_i;f_S)\le \min\{H(f_i),H(f_S)\}$. Αυτό συνεπάγεται ότι η αμοιβαία πληροφορία δύο τυχαίων μεταβλητών φράσσεται από πάνω από το ελάχιστο των εντροπιών τους. Καθώς η εντροπία ενός χαρακτηριστικού μπορεί να έχει μεγάλες διακυμάνσεις, αυτό το μέγεθος θα πρέπει να κανονικοποιηθεί πριν εφαρμοστεί σε κάποιο σύνολο χαρακτηριστικών. Η κανονικοποίηση αντισταθμίζει την μεροληψία της αμοιβαίας πληροφορίας για τα χαρακτηριστικά με μεγάλο εύρος τιμών, και φράσσει το

εύρος στο διάστημα $[0,1]$. Η μεροληψία της αμοιβαίας πληροφορίας είναι γνωστό πρόβλημα στα δέντρα απόφασης για την επιλογή χαρακτηριστικών.

Η κανονικοποιημένη αμοιβαία πληροφορία $NI(f_i;f_S)$, ανάμεσα στο υποψήφιο προς επιλογή χαρακτηριστικό f_i και ένα ήδη επιλεγμένο χαρακτηριστικό f_S με βάση το ελάχιστο της εντροπίας των δύο χαρακτηριστικών $\min\{H(f_i),H(f_S)\}$ ορίζεται ως:

$$NI(f_i;f_S)=\frac{I(f_i;f_S)}{\min\{H(f_i),H(f_S)\}} \quad . \qquad 4.27$$

Στον συγκεκριμένο αλγόριθμο θα χρησιμοποιηθεί ο μέσος όρος της κανονικοποιημένης αμοιβαίας πληροφορίας ως μέτρο αξιολόγησης για τον πλεονασμό ανάμεσα στο υποψήφιο προς επιλογή i-οστό χαρακτηριστικό και στο υποσύνολο των ήδη επιλεγμένων χαρακτηριστικών $S=\{f_S\}$, για $s=1,\ldots,|S|$, δηλαδή:

$$\frac{1}{|S|}\cdot\sum_{f_s\in S} NI(f_i;f_S) \quad . \qquad 4.28$$

όπου $|S|$ είναι η πληθικότητα του συνόλου S. Η τελευταία εξίσωση (4.28) είναι ένα συμμετρικό μέτρο συσχέτισης που παίρνει τιμές στο διάστημα $[0,1]$. Παίρνει την τιμή 0 στην περίπτωση που το υποψήφιο προς επιλογή χαρακτηριστικό f_i και το υποσύνολο S των ήδη επιλεγμένων χαρακτηριστικών είναι ανεξάρτητα, και την τιμή 1 όταν το χαρακτηριστικό f_i έχει μεγάλη συσχέτιση με το S.

Το κριτήριο επιλογής G, που θα χρησιμοποιηθεί στον αλγόριθμο NMIFS είναι το παρακάτω:

$$G=I(C;f_i)-\frac{1}{|S|}\cdot\sum_{f_s\in S} NI(f_i;f_S) \quad . \qquad 4.29$$

Ο δεξιός όρος της σχέσης (4.29) είναι το προσαρμοσμένο μέτρο ποινής του πλεονασμού, το οποίο αντιστοιχεί στον μέσο όρο της κανονικοποιημένης αμοιβαίας πληροφορίας ανάμεσα στο υποψήφιο προς επιλογή χαρακτηριστικό και το υποσύνολο S των ήδη επιλεγμένων χαρακτηριστικών. Σε αυτήν την εξίσωση για το κριτήριο δεν υπάρχει κάποια παράμετρος που να πρέπει να καθοριστεί από τον χρήστη σε αντίθεση με τα κριτήρια των αλγορίθμων MIFS και MIFS-U (4.23) και (4.24) που απαιτούν την εισαγωγή της παραμέτρου β.

4.3.3 Αλγόριθμος NMIFS

Παρατίθεται ο αλγόριθμος NMIFS, ο οποίος έχει ως κριτήριο επιλογής χαρακτηριστικών την εκτίμηση της ποσότητας της αμοιβαίας πληροφορίας, όπως αυτή υπολογίζεται από την εξίσωση 4.29.

Algorithm 3. NMIFS

Initialization: Set $F=\{f_i/i=1,...,N\}$, initial set of N features, and $S=\{\emptyset\}$, empty set. (ορισμός $F=\{f_i/i=1,...,N\}$, αρχικό σύνολο N χαρακτηριστικών, και $S=\{\emptyset\}$ κενό σύνολο.)

Calculation of the Mutual Information with respect to the classes: Calculate $I(f_i;C)$, for each $f_i \in F$. (Υπολογισμός της αμοιβαίας πληροφορίας σε σχέση με τις κλάσεις: Υπολογισμός του $I(f_i;C)$, για κάθε $f_i \in F$.)

Selection of the first feature: Find $\hat{f}_i = \max_{i=1,...,N}\{I(f_i;C)\}$. Set $F \leftarrow F \setminus \{\hat{f}_i\}$; set $S \leftarrow \{\hat{f}_i\}$. (Επιλογή πρώτου χαρακτηριστικού: Εύρεση του $\hat{f}_i = \max_{i=1,...,N}\{I(f_i;C)\}$. Ορισμός του $F \leftarrow F \setminus \{\hat{f}_i\}$; Ορισμός του $S \leftarrow \{\hat{f}_i\}$.)

Greedy feature selection: Repeat until $|S|=k$.(Άπληστη επιλογή χαρακτηριστικών: επανάληψη μέχρι $|S|=k$.)

Calculation of the Mutual Information between features: Calculate $I(f_i;f_S)$ for all pairs (f_i, f_S) , with $f_i \in F$ and $f_S \in S$ if it is not available. (Υπολογισμός της αμοιβαίας πληροφορίας ανάμεσα στα χαρακτηριστικά: Υπολογισμός του $I(f_i;f_S)$ για όλα τα ζεύγη $f_i;f_S$, με $f_i \in F$ και $f_S \in S$ εάν δεν δίνεται ήδη.)

Selection of the next feature: Select feature $f_i \in F$ that maximizes measure 4.29. (Επιλογή του επόμενου χαρακτηριστικού: Εύρεση του χαρακτηριστικού $f_i \in F$ που μεγιστοποιεί τη σχέση 4.29.)

Output: the set S containing the selected features. (Αποτέλεσμα: το σύνολο S που περιέχει τα επιλεγμένα χαρακτηριστικά.)

Το υπολογιστικό κόστος του NMIFS δίνεται από την ταυτόχρονη ταξινόμηση όλων των

ζευγών χαρακτηριστικών, η οποία απαιτείται όταν η εκτίμηση της αμοιβαίας πληροφορίας γίνεται με τον αλγόριθμο του Fraser. Η μέθοδος ταξινόμησης που επιλέχθηκε είναι ο αλγόριθμος ταξινόμησης σωρού (heapsort) (Press W., 1992), ο οποίος ακόμη και στην χειρότερη περίπτωση έχει πολυπλοκότητα $O(N \cdot \log N)$, όπου N είναι το πλήθος των χαρακτηριστικών. Επιπλέον ο NMIFS υπολογίζει και την εντροπία κάθε χαρακτηριστικού στο ίδιο βήμα που υπολογίζει την αμοιβαία πληροφορία αυτού του χαρακτηριστικού και της κλάσης $I(C; f_i)$. Γι'αυτόν τον λόγο ο NMIFS έχει την ίδια πολυπλοκότητα με τον MIFS, δηλαδή $O(N \cdot \log N)$.

4.3.4 Γενετικός Αλγόριθμος GAMIFS

Ο υβριδικός αλγόριθμος φίλτρου–περιτυλίγματος επιλογής χαρακτηριστικών έχει δύο μέρη: έναν γενετικό αλγόριθμο και ένα νευρωνικό δίκτυο. Ο γενετικός αλγόριθμος χρησιμοποιεί ντετερμινιστική επιλεκτική αντικατάσταση (deterministic crowding) (Mahfoud S. W., 1995). Αυτός ο αλγόριθμος χρησιμοποιεί στρατηγική διαμοιρασμού (niching), η οποία σε αντίθεση με τους απλούς γενετικούς αλγορίθμους μπορεί να βρει και να διατηρήσει πολλαπλές βέλτιστες λύσεις σε πολυτροπικά προβλήματα. Στην ντετερμινιστική επιλεκτική αντικατάσταση, όλα τα άτομα του πληθυσμού χωρίζονται τυχαία σε ζευγάρια και επανασυνδυάζονται, δηλαδή η πιθανότητα της διασταύρωσης είναι ένα. Εδώ έγινε χρήση της διωνυμικής διασταύρωσης επειδή δεν έχει μεροληψία θέσεως. Η μετάλλαξη είναι προαιρετική στην ντετερμινιστική επιλεκτική αντικατάσταση. Ο προκύπτον απόγονος μπαίνει σε μια διαδικασία επιλογής χαρακτηριστικών (tournament) με τον πιο κοντινό γονιό, η οποία έχει ως κριτήριο την απόσταση Hamming (Hamming distance). Ο νικητής αυτής της διαδικασίας αντιγράφεται στον νέο πληθυσμό της επόμενης γενιάς.

Για το πρόβλημα επιλογής χαρακτηριστικών, ένα υποσύνολο επιλεγμένων χαρακτηριστικών αναπαρίσταται ως δυαδικό διάνυσμα μήκους L (όπου L είναι ο συνολικός αριθμός χαρακτηριστικών του προβλήματος), το οποίο στην i-οστή θέση έχει την τιμή 1 εάν το i-στό χαρακτηριστικό περιλαμβάνεται στο υποσύνολο, και έχει την τιμή 0 εάν το i-στό χαρακτηριστικό δεν περιλαμβάνεται στο υποσύνολο. Προκειμένου να

εξεταστεί η απόδοση ενός ατόμου (χρωμόσωμα), τα δυαδικά διανύσματα εισάγονται σε έναν ταξινομητή MLP. Το μέγεθος της εισόδου έχει προκαθοριστεί να είναι L, και οι είσοδοι που αντιστοιχούν σε μη επιλεγμένα χαρακτηριστικά παίρνουν την τιμή 0. Η συνάρτηση απόδοσης περιλαμβάνει έναν όρο που απευθύνεται στην ακρίβεια ταξινόμησης και έναν όρο ποινής για τα άτομα που έχουν μεγάλο αριθμό χαρακτηριστικών. Η απόδοση ενός χρωμοσώματος c εκφράζεται ως:

$$J(c) = \text{accuracy}(c) - \lambda \cdot \left(\frac{\text{nfeatures}(c)}{L}\right) \quad , \qquad 4.30$$

όπου $\text{accuracy}(c)$ είναι το ποσοστό σφάλματος ανά μονάδα του ταξινομητή (ανά άτομο) όταν χρησιμοποιείται το υποσύνολο c και $\text{nfeatures}(c)$ είναι ο αριθμός των επιλεγμένων χαρακτηριστικών. Η παράμετρος λ ελέγχει το ισοζύγιο ανάμεσα στους δύο όρους της σχέσης (4.30). Για τον υπολογισμό του $\text{accuracy}(c)$, γίνεται εκμάθηση ενός ταξινομητή MLP τριών επιπέδων, προκειμένου να ελαχιστοποιηθεί το άθροισμα των τετραγωνικών σφαλμάτων με την χρήση ενός αλγόριθμου εκμάθησης, τον quasi-Newton δεύτερης τάξης, οπισθοδρομικής διάδοσης του σφάλματος (second order backpropagation quasi-Newton) (BPQ) (Saito K.,1997). Η ακρίβεια του ταξινομητή είναι το μέγιστο ποσοστό των σωστών ταξινομήσεων ανά μονάδα του ταξινομητή σε ένα σύνολο επικύρωσης. Ο αλγόριθμος εκμάθησης BPQ είναι πιο γρήγορος από τους αλγόριθμους πρώτης τάξης, και από πολλούς δεύτερης τάξης όπως ο Broydon-Fletcher-Goldfarb-Shanno (BFGS) (Saito K.,1997).

Θα δοθεί μια μέθοδος για την αρχικοποίηση του αρχικού πληθυσμού του γενετικού αλγορίθμου, με καλά σημεία εκκίνησης, η οποία χρησιμοποιεί για την αξιολόγηση χαρακτηριστικών την μέθοδο NMIFS. Επιπλέον, θα εισαχθεί ένας τελεστής μετάλλαξης που επίσης δουλεύει στον πλαίσιο του NMIFS, και επιταχύνει την σύγκλιση του αλγορίθμου, ο οποίος θα είναι ο μόνος που θα χρησιμοποιηθεί. Ο τελεστής επιτρέπει την εισαγωγή ενός σχετικού χαρακτηριστικού, ή την αφαίρεση ενός άσχετου χαρακτηριστικού από τα άτομα του πληθυσμού του γενετικού αλγορίθμου. Το μεταλλαγμένο άτομο πρώτα αξιολογείται για να επαληθευτεί εάν η μετάλλαξη βελτιώνει την ακρίβεια ταξινόμησης, και μόνο στην περίπτωση που η απόδοση του μεταλλαγμένου ατόμου είναι καλύτερη από το αρχικό άτομο, ολοκληρώνεται η μετάλλαξη. Στην αντίθετη περίπτωση η μετάλλαξη αναιρείται. Ο συγκεκριμένος τελεστής μπορεί να

χρησιμοποιηθεί σε συνδυασμό με οποιονδήποτε άλλο ταξινομητή εκτός του νευρωνικού δικτύου MLP, αφού ο μηχανισμός επιτάχυνσης της σύγκλισης του αλγορίθμου δεν εξαρτάται από τον τρόπο λειτουργίας του ταξινομητή.

Προκειμένου να γίνει η αρχικοποίηση του GAMIFS, ορίζεται ως P το μέγεθος του πληθυσμού και L το μήκος των ατόμων. Ο αλγόριθμος NMIFS χρησιμοποιείται για αξιολόγηση, έτσι ώστε να αρχικοποιηθεί ένα ποσοστό του πληθυσμού. Το ποσοστό αυτό του πληθυσμού, που θα αρχικοποιηθεί με αυτόν τον τρόπο, καθορίζεται μέσω μιας παραμέτρου ρ, για την οποία ισχύει ότι $\rho \in [0,1]$, και $\rho \cdot P$ θα οριστεί ως το πλήθος των ατόμων αυτής της κατηγορίας. Η παράμετρος $\theta \in [0,1]$ που καθορίζεται από τον χρήστη, επιτρέπει την επιλογή των καλύτερων $\theta \cdot L$ χαρακτηριστικών που προκύπτουν από τον NMIFS. Τα υπόλοιπα άτομα του πληθυσμού $(1-\rho) \cdot P$ θα αρχικοποιηθούν τυχαία.

Algorithm 4.1. Initialization of GAMIFS (with NMIFS)

initialize_nmifs () (αρχικοποίηση με τον nmifs)

{

Find the subset S of the best $\theta \cdot L$ features using NMIFS ranking ; (Εύρεση του υποσυνόλου S των καλύτερων $\theta \cdot L$ χαρακτηριστικών με χρήση του NMIFS ;)

Initialize $\rho \cdot P$ individuals by using NMIFS (Αρχικοποίηση των $\rho \cdot P$ ατόμων με χρήση του NMIFS) {

For all $f_i \in S$ (Για όλα τα $f_i \in S$) {

set the i-th bit to 1 ; (Ορισμός της τιμής 1 στο i-οστό ψηφίο ;)

}

Else (Αλλιώς) {

set the i-th bit randomly in $\{0,1\}$; (Ορισμός μιας τυχαίας τιμής από το $\{0,1\}$ στο i-οστό ψηφίο ;)

}

Initialize $(1-\rho) \cdot P$ individuals randomly ; (Αρχικοποίηση των $(1-\rho) \cdot P$ ατόμων τυχαία ;)

}

Για παράδειγμα, έστω $P=100$ το μέγεθος του πληθυσμού και $L=50$ το μήκος των ατόμων. Ακόμη έστω οι παράμετροι $\rho=0,3$ και $\theta=0,2$. Η διαδικασία αρχικοποίησης θα διαλέξει $\rho \cdot P=30$ άτομα να αρχικοποιηθούν με τον NMIFS. Τα ψηφία που αντιστοιχούν στα καλύτερα $\theta \cdot L=10$ χαρακτηριστικά που προκύπτουν από τον NMIFS θα λάβουν την τιμή 1, και τα υπόλοιπα θα λάβουν την τιμή 0 ή 1 τυχαία.

Όσον αφορά τον τελεστή μετάλλαξης, αφού το νόημα είναι να αξιοποιηθούν οι καλύτερες λύσεις, ο τελεστής μετάλλαξης εφαρμόζεται στο ποσοστό δ% των πρώτων καλύτερων ατόμων του πληθυσμού. Σε κάθε επανάληψη του αλγορίθμου επιλέγεται ένα άτομο και αξιολογείται εάν η απόδοσή του έχει μεγαλύτερη τιμή από την μέγιστη τιμή απόδοσης που αποκτήθηκε στην προηγούμενη γενιά πολλαπλασιασμένη με τον παράγοντα $1-\delta$.

Στην περίπτωση που αυτό ισχύει, εισάγεται ένα χαρακτηριστικό με πιθανότητα p_a στο άτομο, ή αφαιρείται ένα με πιθανότητα $1-p_a$ από αυτό. Όταν εισάγεται ένα χαρακτηριστικό, θεωρείται ότι τα $|S|$ ψηφία που έχουν σε αυτό το στάδιο την τιμή 1 στο άτομο, αντιστοιχούν στο υποσύνολο των χαρακτηριστικών που είναι ήδη επιλεγμένα ενώ τα ψηφία που έχουν την τιμή 0 αντιστοιχούν στα χαρακτηριστικά που δεν έχουν επιλεχθεί ακόμα. Ο NMIFS χρησιμοποιείται σε αυτό το στάδιο για να βρει πιο χαρακτηριστικό είναι το καλύτερο που θα μπορούσε να εισαχθεί στο υποσύνολο των επιλεγμένων χαρακτηριστικών, ανάμεσα φυσικά από τα χαρακτηριστικά που δεν έχουν επιλεγεί ακόμα και άρα δεν ανήκουν σε αυτό, δηλαδή έχουν την τιμή 0 στο άτομο. Αφού βρεθεί πιο χαρακτηριστικό είναι το πιο κατάλληλο, η τιμή του αντίστοιχου ψηφίου μέσα στο άτομο θα αλλάξει από το 0 στο 1.

Στην αντίθετη περίπτωση, αυτή της αφαίρεσης ενός χαρακτηριστικού, είτε το λιγότερο σχετικό χαρακτηριστικό στο άτομο αφαιρείται με πιθανότητα p_i , ή το περισσότερο πλεονάζον χαρακτηριστικό αφαιρείται με πιθανότητα $1-p_i$. Στην περίπτωση που αφαιρείται το περισσότερο πλεονάζον χαρακτηριστικό, αυτό θεωρείται ότι είναι εκείνο που παρουσιάζει την μεγαλύτερη αύξηση στο άθροισμα της αμοιβαίας πληροφορίας σε σχέση με τα υπόλοιπα $|S|-1$ εναπομείναντα χαρακτηριστικά.

Αφού ολοκληρωθεί η μετάλλαξη, ελέγχεται εάν η απόδοση του μεταλλαγμένου ατόμου είναι μεγαλύτερη από την απόδοση του αρχικού ατόμου. Εάν αυτό ισχύει, τότε το

τελευταίο αντικαθίσταται από το πρώτο στον πληθυσμό. Διαφορετικά, το αρχικό άτομο παραμένει στον πληθυσμό.

Οι τρεις τελεστές τοπικής αναζήτησης που αναφέρθηκαν παραπάνω εξηγούνται αναλυτικότερα. Αυτοί οι τελεστές ενεργούν σε ένα χρωμόσωμα c του πληθυσμού, τα ψηφία του οποίου λαμβάνουν την τιμή 1 όταν το χαρακτηριστικό ανήκει στο υποσύνολο των επιλεγμένων χαρακτηριστικών S, και λαμβάνουν την τιμή 0 όταν το χαρακτηριστικό δεν έχει επιλεγεί ακόμα, δηλαδή ανήκει στο $F \setminus S$, όπου F είναι το σύνολο όλων των χαρακτηριστικών.

add_nmifs: εισαγωγή του πιο πληροφοριακού χαρακτηριστικού ανάμεσα από τα χαρακτηριστικά που δεν έχουν επιλεχθεί ακόμα, δηλαδή

$$i^* = \text{argmax}_i \{ I(C; f_i) - \frac{1}{|S|} \cdot \sum_{f_s \in S} NI(f_s; f_i) \} \quad , \qquad 4.31$$

Όπου $f_i \in F \setminus S$. Ορισμός της τιμής 1 στο i^* −οστό ψηφίο του c και δημιουργία του νέου ατόμου $\bar{c}$.

remI_nmifs: Αφαίρεση του πιο άσχετου χαρακτηριστικού ανάμεσα σε αυτά του συνόλου των ήδη επιλεγμένων χαρακτηριστικών, δηλαδή

$$\hat{i} = \text{argmin}_S \{ I(C; f_S) \} \quad , \qquad 4.32$$

Όπου $f_S \in S$. Ορισμός της τιμής 0 στο $\hat{i}$ −οστό ψηφίο του c και δημιουργία του νέου ατόμου $\bar{c}$.

remR_nmifs: Αφαίρεση του πιο πλεονάζοντος χαρακτηριστικού ανάμεσα σε αυτά του συνόλου των ήδη επιλεγμένων χαρακτηριστικών, δηλαδή

$$\tilde{i} = \text{argmax}_S \{ \sum_{j=1, j \neq s}^{|S|} NI(f_j; f_s) \} \quad , \qquad 4.33$$

όπου $f_i, f_S \in S$. Ορισμός της τιμής 0 στο $\tilde{i}$ −οστό ψηφίο του c και δημιουργία του νέου ατόμου $\bar{c}$.

Η συνάρτηση $\text{rand}()$ δημιουργεί τυχαίους αριθμούς στο διάστημα $[0,1]$ με ομοιόμορφη κατανομή. Η διαδικασία μετάλλαξης λειτουργεί σε ένα δοσμένο χρωμόσωμα c του πληθυσμού, και παρουσιάζεται παρακάτω.

Algorithm 4.2. Mutation of GAMIFS (with NMIFS)

```
mutation_nmifs () (μετάλλαξη με τον nmifs)
{
   If  p_a≤rand()
       c̄=add_nmifs(c) ;
   elseif  p_i≤rand()
       c̄=remI_nmifs(c) ;
   Else
       c̄=remR_nmifs(c) ;
}
```

Θα παρουσιαστεί παρακάτω αναλυτικά ολόκληρος ο αλγόριθμος GAMIFS ο οποίος χρησιμοποιεί τους δύο αλγορίθμους (αρχικοποίησης και μετάλλαξης) που παρουσιάστηκαν προηγουμένως. Έστω c ένα άτομο του πληθυσμού, το οποίο έχει μήκος L και απόδοση $J(c)$. Έστω επίσης S το υποσύνολο των των ήδη επιλεγμένων χαρακτηριστικών στο χρωμόσωμα c (τα ψηφία που έχουν τον αριθμό 1) και $|S|$ η πληθικότητα του S. Ορίζεται το J_{max} να εκφράζει τη μέγιστη απόδοση που αποκτήθηκε στην διάρκεια της τελευταίας γενιάς και $1-\delta$ ο παράγοντας του J_{max} που αφορά την μετάλλαξη, όπως αυτός αναλύθηκε παραπάνω. Επιπλέον χρησιμοποιείται η απόσταση Hamming ανάμεσα σε δύο δυαδικά διανύσματα c_1 και c_2 για την αξιολόγηση της ομοιότητας ανάμεσα στους γονείς και τους μεταλλαγμένους απογόνους, η οποία συμβολίζεται ως $d(c_1,c_2)$. Σημειώνεται ότι αφού ο πληθυσμός P έχει μέγεθος n, η επανάληψη του αλγορίθμου εκτελείται $n/2$ φορές για κάθε γενιά, διότι χρησιμοποιεί κάθε φορά δύο άτομα του πληθυσμού.

Ο ακόλουθος αλγόριθμος αντιστοιχεί στον γενετικό αλγόριθμο με ντετερμινιστική επιλεκτική αντικατάσταση που βασίζεται στον NMIFS για την αρχικοποίηση και την μετάλλαξη.

Algorithm 4.3. GAMIFS

```
gamifs ()
{
   initialize_nmifs ;
   repeat {
      repeat n/2 times {
         select two parents p1 and p2 from P ;
          (c1 , c2)=crossover(p1 , p2) ;
         If J(c1)>(1−δ)·Jmax {
             c̄1=mutation_nmifs(c1) ;
            If J(c̄1)≤J(c1) , then c̄1=c1 ;
         }
         If J(c2)>(1−δ)·Jmax {
             c̄2=mutation_nmifs(c2) ;
            If J(c̄2)≤J(c2) , then c̄2=c2 ;
         }
         replace {
            If [d(p1 , c̄1)+d(p2 , c̄2)]≤[d(p1 , c̄2)+d(p2 , c̄1)] {
               If J(c̄1)>J(p1) , replace p1 by c̄1 ;
               If J(c̄2)>J(p2) , replace p2 by c̄2 ;
            }
            else {
               If J(c̄2)>J(p1) , replace p1 by c̄2 ;
               If J(c̄1)>J(p2) , replace p2 by c̄1 ;
            }
         }
      } / * end one generation * /
   } until (stopping condition)
}
```

4.3.5 Πειραματικά Αποτελέσματα Αλγορίθμων NMIFS, GAMIFS

Η απόδοση του αλγορίθμου NMIFS συγκρίθηκε με τα αποτελέσματα που έδωσαν οι αλγόριθμοι MIFS, MIFS-U και mRMR σε διαφορετικά τέσσερα σύνολα δεδομένων: ένα ομοιόμορφο σύνολο τεχνητών δεδομένων υπερκύβου, ένα σύνολο δεδομένων κυμματομορφής (Breiman L., 1984), ένα spambase σύνολο δεδομένων (Chow T.W., 2005) και τέλος ένα σύνολο δεδομένων από sonar (Newman D., 1998). Επιπλέον, ο NMIFS αξιολογήθηκε και σε ένα πρόβλημα χρονοσειρών, το Box and Jenkin's gas furnace, το οποίο αφορά την καύση ενός μείγματος αέρα-μεθανίου (Box G. E. P., 2003).

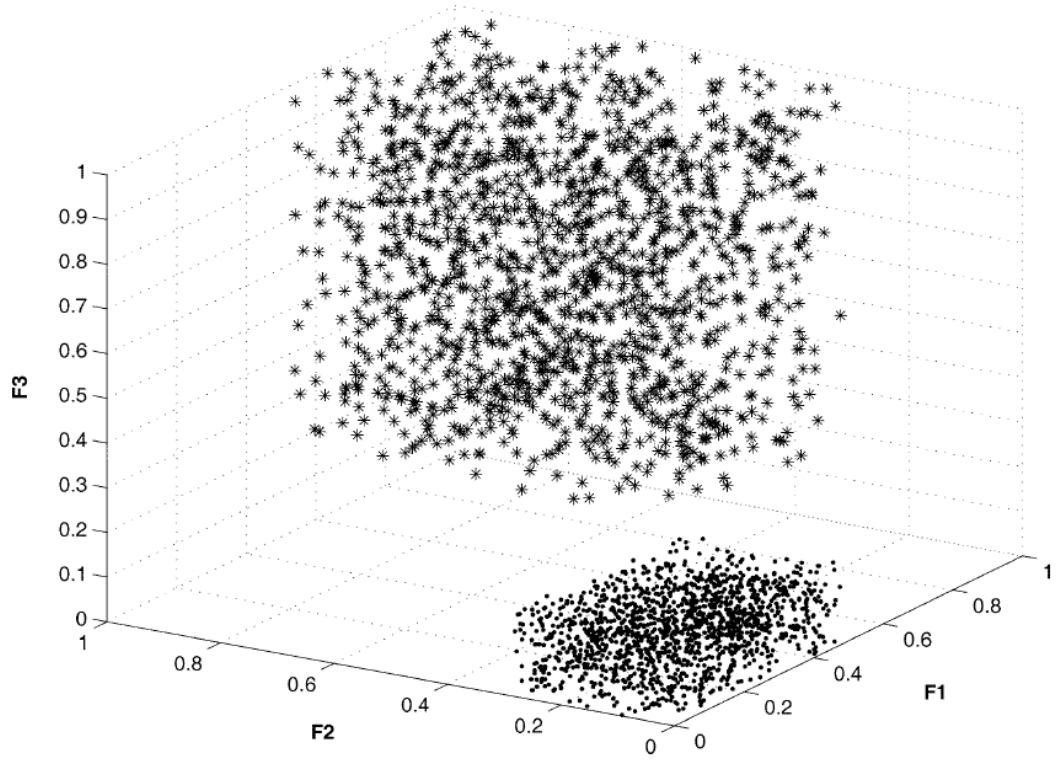
Σε όλες τις περιπτώσεις η αμοιβαία πληροφορία εκτιμήθηκε με την επεκταμένη εκδοχή του αλγόριθμου του Fraser στην περάτωση που το πρόβλημα αποτελούταν από συνεχή χαρακτηριστικά, και πίνακες συνάφειας στην περίπτωση που το πρόβλημα αποτελούταν από διακριτά χαρακτηριστικά. Η παράμετρος ελέγχου β που περιέχουν οι αλγόριθμοι MIFS και MIFS-U ορίστηκε να παίρνει τιμές στο διάστημα $[0,1]$ με βήμα 0,1 . Τα αποτελέσματα που περιείχαν τις καλύτερες τιμές για την παράμετρο ελέγχου β χρησιμοποιήθηκαν για την σύγκριση με τον NMIFS.

Τα υποσύνολα των χαρακτηριστικών αξιολογήθηκαν μέσω ενός MLP νευρωνικού δικτύου το οποίο αποτελούταν από ένα κρυφό επίπεδο το εκπαιδεύτηκε με τον BPQ μέσα από 200 epochs (όπου κάθε epoch είναι 200 επαναλήψεις του τρόπου διασταύρωσης MMX). Όλα τα αποτελέσματα που παρουσιάζονται παρακάτω είναι ο μέσος όρος δέκα δοκιμών με τυχαίες αρχικοποιήσεις. Όλα τα σύνολα δεδομένων εκτός από αυτό του sonar, χωρίστηκαν σε τρία μέρη: 50% των δεδομένων έγινε το σύνολο εκπαίδευσης, 25% το σύνολο επικύρωσης και 25% το σύνολο αξιολόγησης. Το σύνολο δεδομένων από sonar, λόγω μικρού μεγέθους των δειγμάτων (204) χωρίστηκε σε 50 δείγματα για εκπαίδευση, 50 για επικύρωση και 104 για αξιολόγηση. Η μεγαλύτερη τιμή σωστών ταξινομήσεων που αποκτήθηκε στο σύνολο επικύρωσης χρησιμοποιήθηκε για το κριτήριο τερματισμού. Επίσης για τα σύνολα δεδομένων που υπήρχαν πληροφορίες σχετικά με τις κλάσεις που ανήκαν αυτά, ο βέλτιστος αριθμός κρυφών μονάδων N_h , επιλέχθηκε με την εκτέλεση του BPQ για $N_h \in [1,20]$, και την επιλογή της τιμής με τα καλύτερα αποτελέσματα στο σύνολο επικύρωσης του MLP.

Η απόδοση του GAMIFS συγκρίθηκε με αυτή του NMIFS, του γενετικού αλγόριθμου ντετερμινιστικής επιλεκτικής αντικατάστασης χωρίς μετάλλαξη αλλά και του γενετικού αλγόριθμου ντετερμινιστικής επιλεκτικής αντικατάστασης με μετάλλαξη (Estévez P. A., 1998) σε τέσσερα διαφορετικά σύνολα δεδομένων: μη γραμμικά AND τεχνητά δεδομένα, δεδομένα κυμματομορφής του Breiman, spambase σύνολο δεδομένων και δεδομένα από sonar.

Στη παράμετρο λ στην συνάρτηση απόδοσης (4.30) δόθηκε η τιμή $0,1$, έτσι ώστε ο όρος που απευθύνεται στην ακρίβεια να είναι δέκα φορές πιο σημαντικός από τον όρο της ποινής. Το νόημα είναι ότι, όταν υπάρχουν δύο λύσεις με την ίδια ακρίβεια, αυτή με τον μικρότερο αριθμό χαρακτηριστικών να προτιμάται. Η απόδοση κάθε ατόμου αξιολογήθηκε τρεις φορές, και χρησιμοποιήθηκε η καλύτερη λύση. Έγιναν δέκα προσομοιώσεις με το τελικό υποσύνολο δεδομένων για κάθε γενετικό αλγόριθμο. Ο μέσος όρος των ταξινομήσεων χρησιμοποιήθηκε για την σύγκριση.

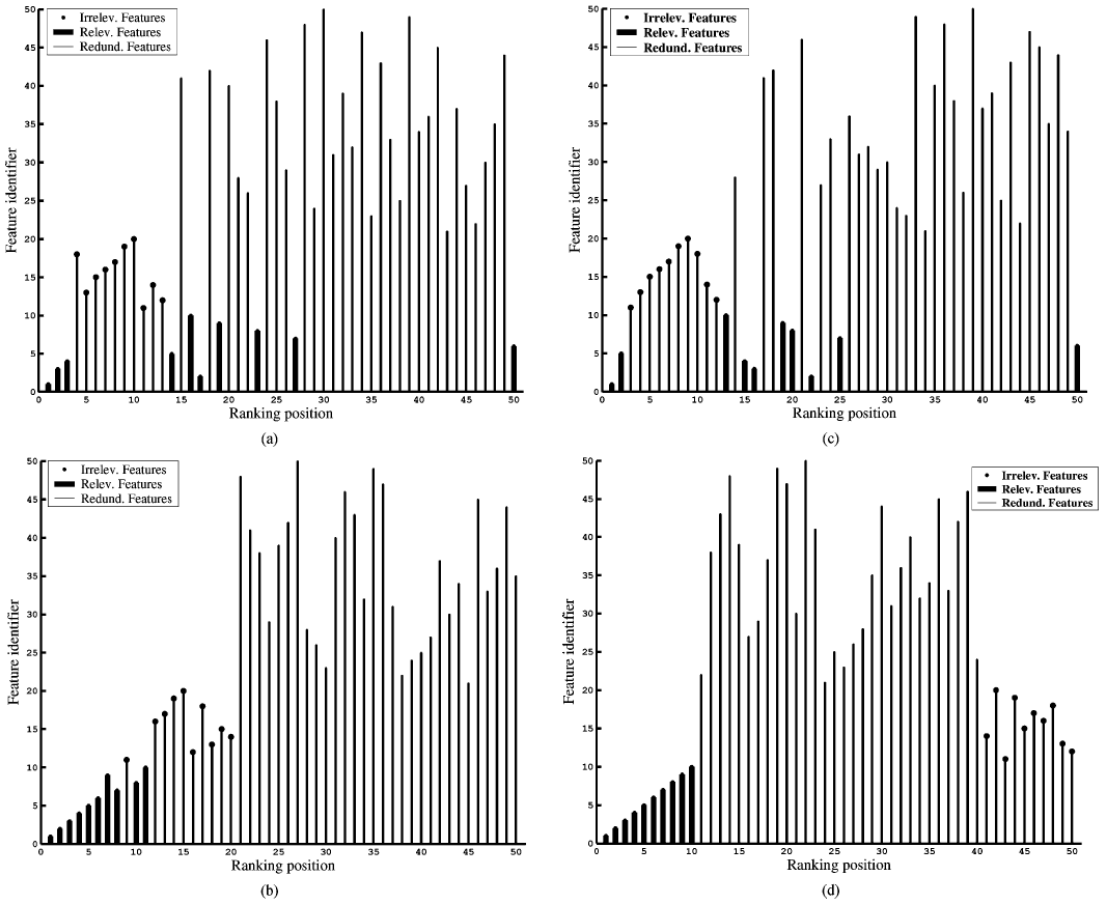
Στο τεχνητό σύνολο δεδομένων, η φύση κάθε χαρακτηριστικού (σχετικό, άσχετο, πλεονάζον) και η σειρά σημαντικότητας είναι γνωστή εξ αρχής. Το ζητούμενο είναι να βρεθούν τα σχετικά χαρακτηριστικά σε φθίνουσα σειρά σχετικότητας, όπου το πρώτο χαρακτηριστικό είναι το πιο σχετικό, το δεύτερο είναι το αμέσως επόμενο πιο σχετικό κ.ο.κ., ύστερα να βρεθούν τα πλεονάζοντα χαρακτηριστικά και τέλος τα άσχετα. Το πρόβλημα αυτό αποτελείται από δύο ομάδες των 500 σημείων που ανήκουν σε μια ομοιόμορφη κατανομή ενός υπερκύβου δέκα διαστάσεων $[0,1]^{10}$. Το σύνολο των σχετικών χαρακτηριστικών αποτελείται από δέκα χαρακτηριστικά, $(f_1, f_2, \ldots, f_{10})$, και ορίστηκε να είναι σε φθίνουσα σειρά σημαντικότητας. Θεωρήθηκε ότι ένα στοιχείο ανήκει στην κλάση C_1 εάν ισχύει για αυτό ότι $f_i \leq \gamma^{i-1} \cdot \alpha$ για $i=1,2,\ldots,10$, και στην κλάση C_2 διαφορετικά. Για $\alpha=0,5$ και $\gamma=0,8$, το πρώτο χαρακτηριστικό χωρίζει το διάστημα $[0,1]$ στις δύο κλάσεις C_1, C_2 στην τιμή $0,5$, το δεύτερο στην τιμή $0,4$, το τρίτο στην τιμή $0,32$ κ.ο.κ. όπως φαίνεται στο σχήμα παρακάτω (Σχήμα 4.2) το οποίο δείχνει μια 3-D εκδοχή του προβλήματος υπερκύβου, όπου με κουκίδες συμβολίζονται τα αντικείμενα της πρώτης κλάσης και με αστεράκια συμβολίζονται τα αντικείμενα της δεύτερης κλάσης. Είναι εμφανές ότι το χαρακτηριστικό f_1 διαχωρίζει τις κλάσεις καλύτερα από το f_2 , και με την σειρά του το f_2 καλύτερα από το f_3 .



Σχήμα 4.2 Τρισδιάστατη απεικόνιση του τεχνητού προβλήματος υπερκύβου

Το σύνολο δεδομένων του υπερκύβου αποτελείται από 50 χαρακτηριστικά: τα χαρακτηριστικά 1-10 είναι σχετικά, τα χαρακτηριστικά 11-20 είναι άσχετα και τα χαρακτηριστικά 21-50 είναι πλεονάζοντα. Τα τελευταία 30 είναι γραμμικοί συνδυασμοί των σχετικών χαρακτηριστικών με 10% πρόσθετο θόρυβο από Gauss κατανομή $N(0,1)$.

Το επόμενο σχήμα (Σχήμα 4.3) απεικονίζει τα αποτελέσματα των αλγορίθμων MIFS 4.3(a), MIFS-U 4.3(b), mRMR 4.3(c), NMIFS 4.3(d). Τα αποτελέσματα που απεικονίζονται για τον MIFS και τον MIFS-U αντιστοιχούν στα καλύτερα αποτελέσματα που προέκυψαν για τις διάφορες τιμές της μεταβλητής β στο διάστημα $[0,1]$. Ο άξονας x αναπαριστά την σειρά επιλογής (θέση στην αξιολόγηση των χαρακτηριστικών), ενώ ο άξονας y αναπαριστά τον αριθμό του χαρακτηριστικού (την πραγματική του θέση). Για παράδειγμα στη γραφική παράσταση 4.3(a) το χαρακτηριστικό με αριθμό 15 επιλέχθηκε από τον MIFS έκτο κατά σειρά και το χαρακτηριστικό με αριθμό 20 επιλέχθηκε δέκατο. Οι μπάρες που έχουν έντονη γραμμή αντιστοιχούν στα σχετικά χαρακτηριστικά, οι μπάρες με την κουκίδα στο άνω άκρο αντιστοιχούν στα άσχετα χαρακτηριστικά και τέλος οι μπάρες με λεπτή γραμμή αντιστοιχούν στα πλεονάζοντα χαρακτηριστικά.



Σχήμα 4.3 Επιλογή χαρακτηριστικών στο πρόβλημα του υπερκύβου για (a) MIFS με β=0,4 (b) MIFS-U με β=0,6 (c) mRMR (d) NMIFS

Οι αλγόριθμοι MIFS, MIFS-U και mRMR επιλέγουν τα πλεονάζοντα χαρακτηριστικά μετά τα άσχετα, και επίσης επιλέγουν άσχετα χαρακτηριστικά πριν από μερικά σχετικά. Αυτή η επίδραση είναι ισχυρότερη στον MIFS και στον mRMR (4.3(a), 4.3(c)), όπου όλα τα άσχετα χαρακτηριστικά επιλέχθηκαν πριν από τα εναπομείναντα επτά από τα δέκα και αντίστοιχα τα οκτώ από τα δέκα σχετικά χαρακτηριστικά. Ο MIFS-U 4.3(b) ενώ έχει καλύτερη συμπεριφορά από τον MIFS, δύο σχετικά χαρακτηριστικά επιλέχθηκαν μετά από ένα άσχετο. Από την άλλη ο NMIFS 4.3(d) επιλέγει όλα τα χαρακτηριστικά σε ιδανική σειρά: πρώτα το σύνολο με τα σχετικά χαρακτηριστικά σε αύξουσα σειρά μετά το σύνολο με τα πλεονάζοντα και τελευταία το σύνολο με τα άσχετα.

Το δεύτερο πείραμα έγινε στο πρόβλημα Box and Jenkin's gas furnace, το οποίο περιγράφει ένα φούρνο του οποίου η είσοδος $u(t)$ είναι η ροή αέρα μέσα στον φούρνο και η έξοδος $y(t)$ είναι η συγκέντρωση CO_2 στα αέρια που εξέρχονται από τον φούρνο. Στόχος του πειράματος ήταν η πρόβλεψη της συγκέντρωσης CO_2 της εξόδου με χρήση

των προηγούμενων τιμών εισόδου-εξόδου που ήταν ήδη γνωστές. Επινοήθηκαν δέκα υποψήφια χαρακτηριστικά ώστε να μπορέσει να δημιουργηθεί ένα μοντέλο πρόβλεψης για το πρόβλημα. Τα χαρακτηριστικά αυτά ουσιαστικά διακρίνονται σε δύο ομάδες. Έστω μια χρονική στιγμή t. Η πρώτη ομάδα αποτελείται από τους τέσσερις προηγούμενους χρόνους της εξόδου $y(t)$, δηλαδή τις τιμές $\{y(t-1), y(t-2), y(t-3), y(t-4)\}$ ενώ η δεύτερη ομάδα αποτελείται από άλλους έξι προηγούμενους χρόνους της εισόδου $u(t)$, δηλαδή τις τιμές $\{u(t-1), u(t-2), u(t-3), u(t-4), u(t-5), u(t-6)\}$. Ένας ταξινομητής MLP εκπαιδεύτηκε με είσοδο τα παραπάνω χαρακτηριστικά σε συνδυασμό με τις διάφορες μεθόδους επιλογής χαρακτηριστικών. Ο βέλτιστος αριθμός κρυμμένων επιπέδων καθορίστηκε εμπειρικά να είναι το τρία. Η αρχιτεκτονική του δικτύου MLP ορίστηκε ως $N_{in}-3-1$, όπου το N_{in} αντιστοιχεί στον αριθμό των επιλεγμένων χαρακτηριστικών, το τρία είναι ο αριθμός των κρυφών μονάδων και το ένα είναι ο αριθμός των μονάδων εξόδου.

Το συγκεκριμένο πρόβλημα έχει χρησιμοποιηθεί εκτενώς στην βιβλιογραφία για την αξιολόγηση και σύγκριση της απόδοσης μεθόδων επιλογής χαρακτηριστικών. Μερικές μέθοδοι ασαφούς λογικής επιλέγουν χαρακτηριστικά ελαχιστοποιώντας το σφάλμα που δημιουργείται στην εκτίμηση του ταξινομητή (Pedrycz W., 1984) (Sugeno M., 1993) (Tong R., 1980) (Xu C., 1987). Η εκμάθηση επιτυγχάνεται μέσω της εύρεσης ενός συνόλου ασαφών κανόνων, το οποίο θα επιτρέπει την λήψη καλής πρόγνωσης σχετικά με το αποτέλεσμα, χρησιμοποιώντας μόνο ένα υποσύνολο των υποψήφιων χαρακτηριστικών. Σημειώνεται ότι αυτές οι ασαφείς μέθοδοι ανήκουν στην κατηγορία των μεθόδων περιτυλίγματος, αφού η επιλογή των βέλτιστων χαρακτηριστικών γίνεται μέσω της εκπαίδευσης μοντέλου.

Τα αποτελέσματα του NMIFS και των ασαφών μεθόδων που αναφέρθηκαν παραπάνω συγκρίθηκαν μέσω των αξιολογήσεων των αποδόσεων του MLP, με είσοδο το αντίστοιχο υποσύνολο χαρακτηριστικών κάθε μεθόδου. Ως μέτρο αξιολόγησης ορίστηκε το κανονικοποιημένο μέσο τετραγωνικό σφάλμα (NMSE):

$$NMSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2} \quad , \qquad 4.34$$

Όπου ο όρος $\hat{y}_i$ αναφέρεται στην εκτιμημένη τιμή για το y (έξοδος του ταξινομητή MLP) και ο όρος $\bar{y}_i$ στην μέση τιμή των y. Όσο πιο κοντά βρίσκεται στο μηδέν η τιμή του NMSE, τόσο καλύτερη είναι η εκτίμηση, άρα και η ακρίβεια της ταξινόμησης. Όταν η τιμή του NMSE βρίσκεται κοντά στο ένα, τότε το μοντέλο εκτιμά μόνο την μέση τιμή των χρονοσειρών.

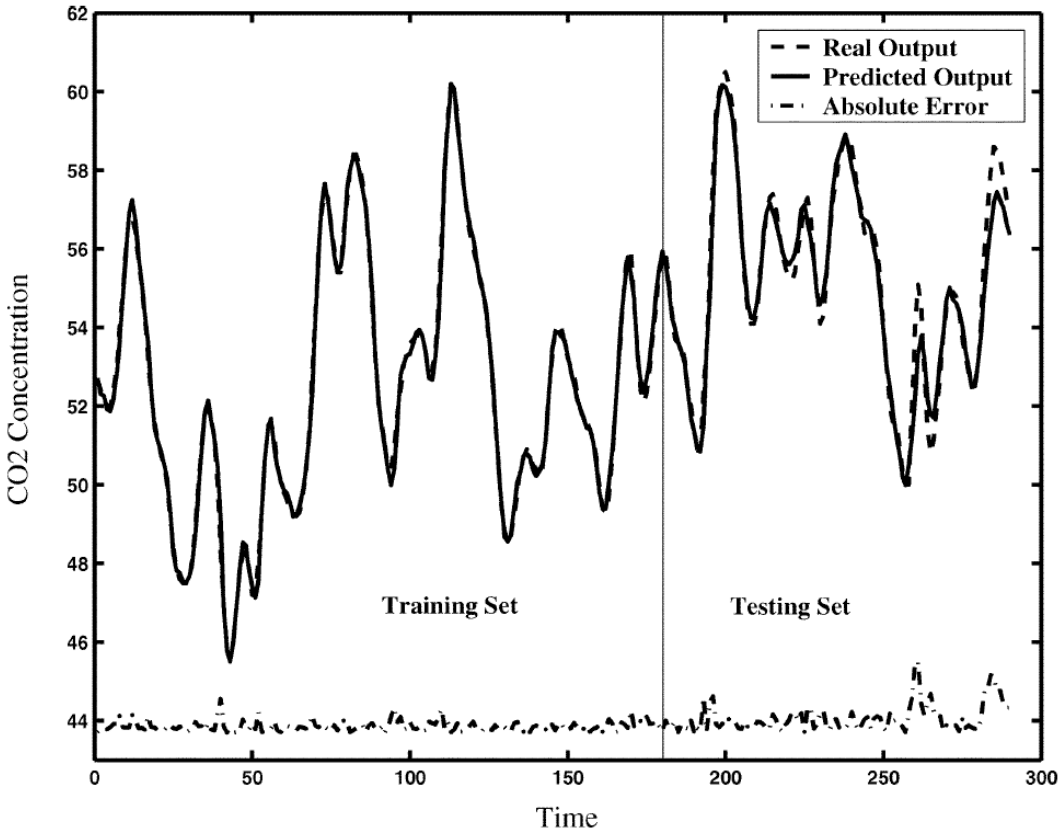
Πίνακας 4.3 Σφάλμα Εκτίμησης (NMSE) διαφόρων μεθόδων επιλογής χαρακτηριστικών για το σύνολο δεδομένων του gas furnace

Μέθοδος Επιλογής Χαρακτηριστικών	Πλήθος Χαρακτηριστικών	Επιλεγμένα Χαρακτηριστικά	NMSE
NMIFS, Pedrycz	2	$y(t-1),u(t-4)$	**0,068**
Tong, Xu	2	$y(t-1),u(t-4)$	0,068
MI	2	$y(t-1),u(t-5)$	0,094
NMIFS	3	$y(t-1),u(t-4),u(t-5)$	**0,049**
MI	3	$y(t-1),u(t-5),u(t-4)$	0,049
Sugeno	3	$y(t-1),u(t-4),u(t-3)$	0,062
NMIFS	4	$y(t-1),u(t-4),u(t-5),u(t-6)$	**0,042**
MI	4	$y(t-1),u(t-5),u(t-4),u(t-3)$	0,061

Ο Πίνακας 4.3 απεικονίζει το κανονικοποιημένο μέσο τετραγωνικό σφάλμα (NMSE) που λήφθηκε από τις διάφορες μεθόδους επιλογής χαρακτηριστικών σε συνδυασμό με διαφορετικά πλήθη επιλεγμένων χαρακτηριστικών.

Είναι αρκετά εμφανές ότι ο αλγόριθμος NMIFS ξεπέρασε σε απόδοση τον απλό MIFS σε όλες τις περιπτώσεις., εκτός από την περίπτωση που επιλέχθηκαν τρία χαρακτηριστικά, στην οποία είχαν την ίδια απόδοση. Ακόμη ο NMIFS έχει την ίδια απόδοση με τους αλγόριθμους του Pedrycz, Tong και Xu στην περίπτωση των δύο επιλεγμένων χαρακτηριστικών, και χαμηλότερο σφάλμα από τον αλγόριθμο του Sugeno στην περίπτωση των τριών επιλεγμένων χαρακτηριστικών. Παρά το γεγονός ότι ο αλγόριθμος NMIFS είναι μια μέθοδος φίλτρου, παρατηρείται πως η απόδοση του είναι το ίδιο καλή με τα μοντέλα της ασαφούς λογικής τα οποία ανήκουν στην κατηγορία των μεθόδων περιτυλίγματος.

Το σχήμα που ακολουθεί δείχνει τις πραγματικές και τις εκτιμημένες τιμές εξόδου όταν επιλέχθηκαν τέσσερα χαρακτηριστικά από τον NMIFS. Το μικρό απόλυτο σφάλμα φαίνεται στο κάτω μέρος του διαγράμματος.



Σχήμα 4.4 Εκτιμημένη τιμή εξόδου και απόλυτο σφάλμα με τα χαρακτηριστικά $(y(t-1), u(t-4), u(t-5), u(t-6))$ που επιλέχθηκαν από τον NMIFS για τα δεδομένα του gas furnace

Το μη γραμμικό AND πρόβλημα είναι επίσης ένα τεχνητό πρόβλημα που επινοήθηκε ώστε να αναδειχθεί μια περίπτωση όπου ο αλγόριθμος NMIFS και άλλοι αυξητικοί αλγόριθμοι αναζήτησης θα αποτύγχαναν. Η έκδοση του προβλήματος περιείχε δεκατέσσερα χαρακτηριστικά: τα πρώτα πέντε, (f_1-f_5), είναι τα άσχετα χαρακτηριστικά, τα επόμενα έξι, (f_6-f_{11}), είναι τα σχετικά χαρακτηριστικά και τέλος τα υπολείποντα τρία, $(f_{12}-f_{14})$, είναι τα πλεονάζοντα χαρακτηριστικά.

Κάθε πλεονάζον χαρακτηριστικό δημιουργήθηκε έτσι ώστε να είναι ουσιαστικά ένα αντίγραφο ενός σχετικού χαρακτηριστικού. Πιο συγκεκριμένα σε αυτό το παράδειγμα τα χαρακτηριστικά $f_{12}-f_{14}$ είναι αντίγραφα των τελευταίων τριών σχετικών

χαρακτηριστικών $f_9 - f_{11}$. Τα άσχετα χαρακτηριστικά δημιουργήθηκαν τυχαία από μια εκθετική κατανομή που έχει μέση τιμή το δέκα. Τα έξι σχετικά χαρακτηριστικά δημιουργήθηκαν από μια ομοιόμορφη κατανομή στο διάστημα $[-1,1]$. Η δουλειά αυτών είναι να καθορίζουν σε ποια από τις δύο κλάσεις C_1 , C_2 ανήκει ένα δείγμα x, σύμφωνα με την ακόλουθη μη γραμμική AND συνάρτηση.

---

nonlinear_AND ()

{

If $((f_6 \cdot f_7 \cdot f_8) > 0)$ AND $((f_9 + f_{10} + f_{11}) > 0)$, then

$x \in C_1$

If $((f_6 \cdot f_7 \cdot f_8) < 0)$ AND $((f_9 + f_{10} + f_{11}) < 0)$, then

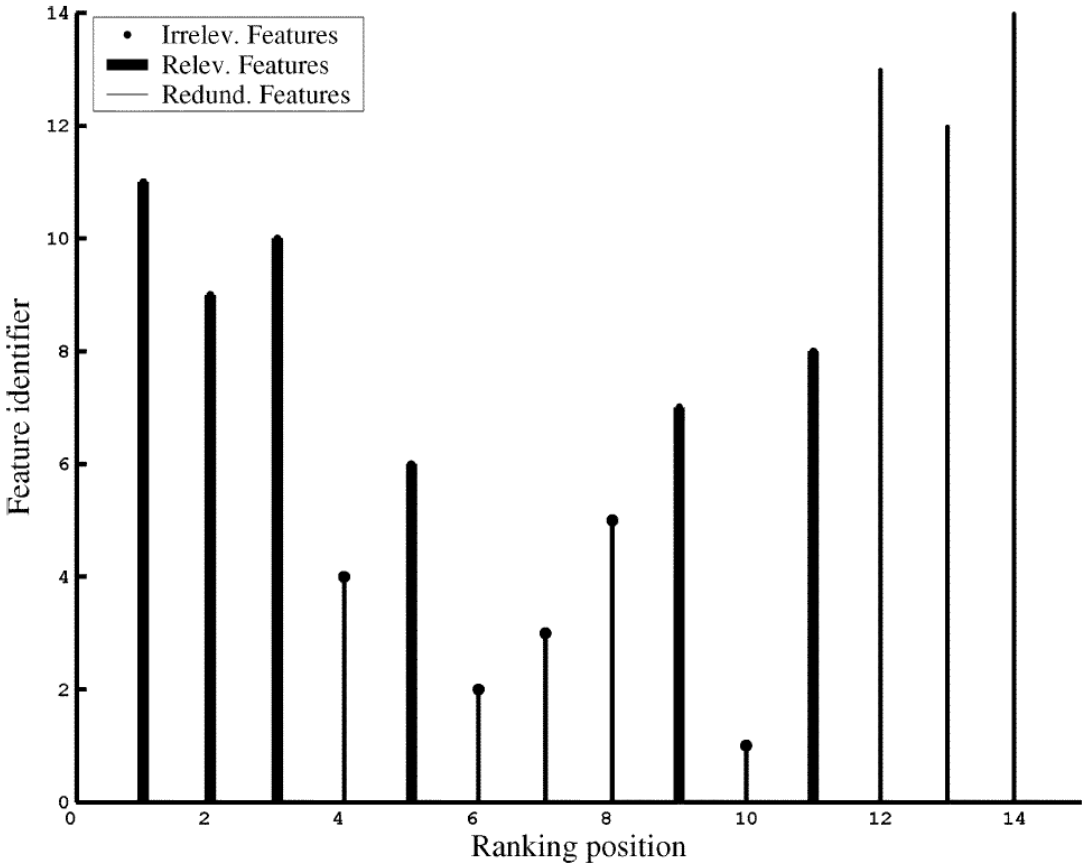
$x \in C_2$

}

Ο αλγόριθμος NMIFS δεν καταφέρνει να επιλύσει αυτό το πρόβλημα διότι τα χαρακτηριστικά f_6 , f_7 και f_8 δεν παρέχουν πληροφορία ξεχωριστά το καθένα, αλλά μόνο όταν συνδυάζονται σαν ομάδα. Επιπλέον, το μη γραμμικό πρόβλημα που σχηματίστηκε εδώ έχει οκτώ διαφορετικές βέλτιστες λύσεις, κάθε μία από τις οποίες αντιστοιχεί σε έναν από τους 2^3 δυνατούς συνδυασμούς των τριών αντιγράφων των χαρακτηριστικών. Ουσιαστικά κάθε συνδυασμός αντιγράφων μπορεί να αντικαταστήσει τα αντίστοιχα αρχικά σχετικά χαρακτηριστικά και να δημιουργήσει μια διαφορετική βέλτιστη λύση.

Το σχήμα που ακολουθεί (Σχήμα 4.5) δείχνει την διαδικασία επιλογής των χαρακτηριστικών που εκτέλεσε ο αλγόριθμος NMIFS με είσοδο το μη γραμμικό AND πρόβλημα που επινοήθηκε.

Παρόμοια με το Σχήμα 4.3 ο άξονας x αναπαριστά την σειρά που επιλέχθηκαν τα χαρακτηριστικά από τον αλγόριθμο NMIFS, ενώ ο άξονας y αναπαριστά τον πραγματικό αριθμό που δόθηκε στα χαρακτηριστικά με την δημιουργία του προβλήματος.

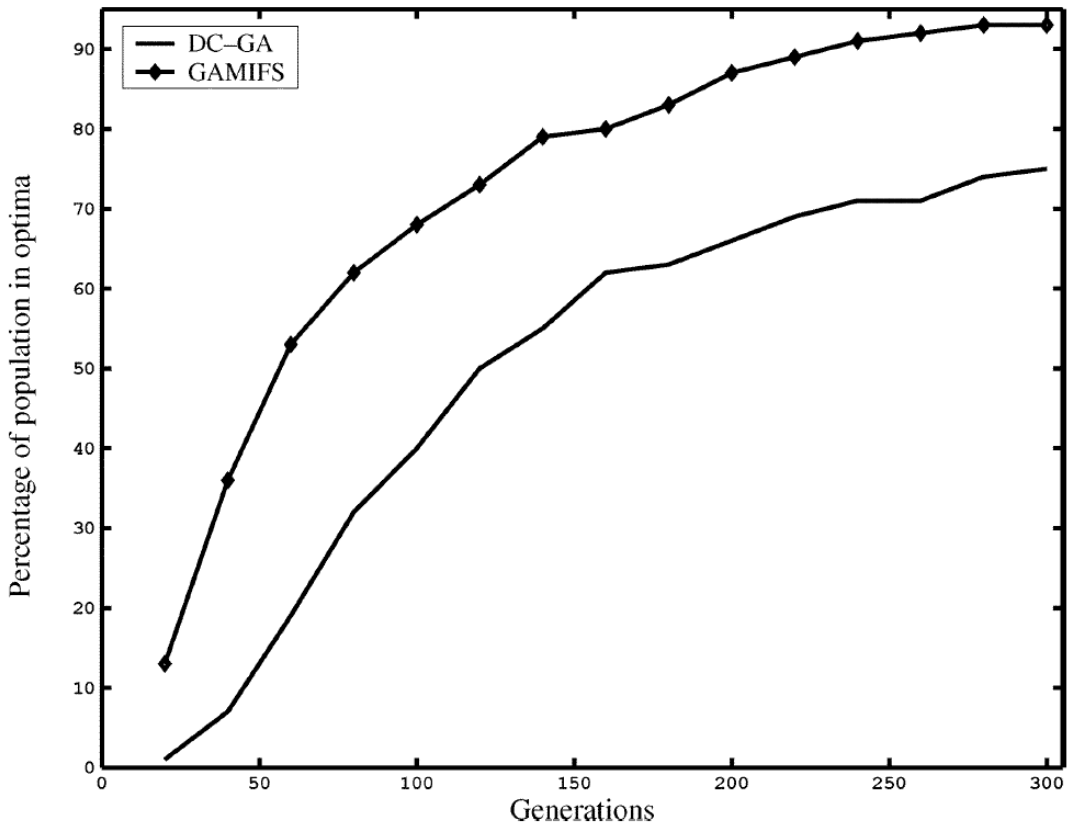


Σχήμα 4.5 Διαδικασία επιλογής χαρακτηριστικών για το μη γραμμικό AND πρόβλημα με τον αλγόριθμο NMIFS

Όπως φαίνεται στο σχήμα, τα σχετικά χαρακτηριστικά $f_9 - f_{11}$ τα οποία συνιστούν τον όρο του γραμμικού συνδυασμού στην μη γραμμική συνάρτηση AND επιλέγονται σωστά στις πρώτες θέσεις από τον αλγόριθμο. Αντίθετα, τα χαρακτηριστικά $f_6 - f_8$ που είναι επίσης σχετικά, και είναι αυτά που συμμετέχουν στον όρο του πολλαπλασιασμού της μη γραμμικής συνάρτησης AND επιλέγονται αφού έχουν ήδη επιλεχθεί πρώτα αρκετά άσχετα χαρακτηριστικά.

Από την άλλη ο αλγόριθμος GAMIFS σε όλες τις περιπτώσεις κατάφερε να βρει και να διατηρήσει και τις οκτώ βέλτιστες λύσεις από τους διαφορετικούς συνδυασμούς των έξι χαρακτηριστικών.

Το Σχήμα 4.6 που ακολουθεί δείχνει τον ρυθμό σύγκλισης του πληθυσμού των αλγορίθμων GAMIFS και DC-GA (γενετικός αλγόριθμος με ντετερμινιστική επιλεκτική αντικατάσταση) χωρίς μετάλλαξη και με τυχαία αρχικοποίηση, για τις οκτώ βέλτιστες λύσεις. Το οριακό κέρδος με την χρήση του GAMIFS είναι περίπου 40%.

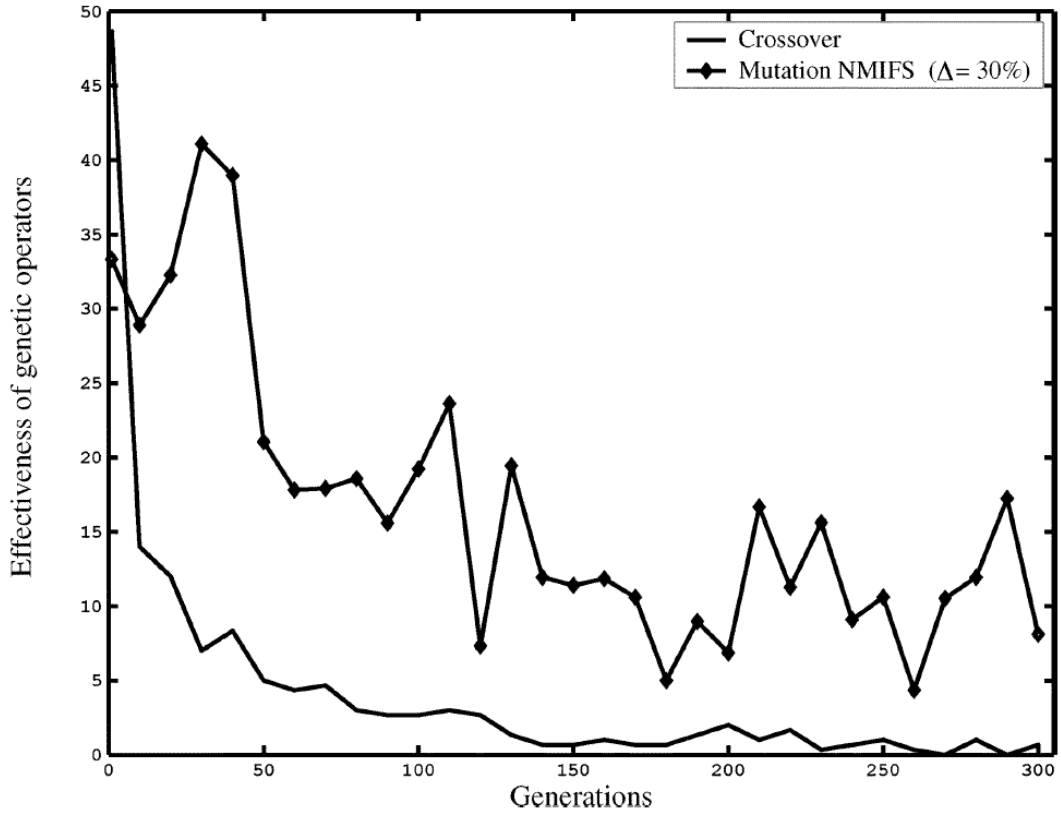


Σχήμα 4.6 Ρυθμός σύγκλισης του πληθυσμού για τις διάφορες βέλτιστες λύσεις συναρτήσει του πλήθους των γενιών του GAMIFS για το μη γραμμικό AND πρόβλημα

Η αποτελεσματικότητα του τελεστή μετάλλαξης και του τελεστή διασταύρωσης ορίζεται ως το ποσοστό των επιτυχημένων εφαρμογών που έχουν πάνω στα άτομα του πληθυσμού. Μια μετάλλαξη θεωρείται επιτυχής εάν η απόδοση του μεταλλαγμένου ατόμου είναι μεγαλύτερη από αυτή των γονιών του. Αντίστοιχα μια διασταύρωση θεωρείται επιτυχής όταν η απόδοση του απόγονου είναι μεγαλύτερη από των γονιών του. Είναι φανερό ότι όταν όλα ή ένα ποσοστό των ατόμων ενός πληθυσμού υπόκεινται στους τελεστές μετάλλαξης και διασταύρωσης, μερικά από αυτά βελτιώνουν την απόδοσή τους, ενώ μερικά άλλα μπορεί είτε να διατηρήσουν ίδια απόδοση είτε να τους μειωθεί, οπότε και σε αυτήν την περίπτωση ακυρώνεται η διαδικασία.

Στο Σχήμα 4.7 αναπαρίσταται η αποτελεσματικότητα του τελεστή διασταύρωσης και του τελεστή μετάλλαξης που αναλύθηκε προηγουμένως (με χρήση του NMIFS) συναρτήσει του πλήθους των γενιών. Η αποτελεσματικότητα του τελεστή μετάλλαξης είναι 15% κατά μέσο όρο, αλλά εφαρμόζεται μόνο στο καλύτερο 30% του πληθυσμού. Η

αποτελεσματικότητα του τελεστή διασταύρωσης πέφτει κάτω από το 5% μετά από 50 γενιές, αλλά εφαρμόζεται σε ολόκληρο των πληθυσμό.



Σχήμα 4.7 Αποτελεσματικότητα των τελεστών διασταύρωσης και μετάλλαξης συναρτήσει του αριθμού των γενιών για το μη γραμμικό AND πρόβλημα

Αφού οι λύσεις του μη γραμμικού AND συνόλου δεδομένων είναι ήδη γνωστές, το πρόβλημα χρησιμοποιήθηκε προκειμένου να βρεθεί ένα καλό σύνολο παραμέτρων για τον GAMIFS, το οποίο θα τον βοηθήσει να πετύχει μεγαλύτερο ρυθμό σύγκλισης. Το μέγεθος του πληθυσμού καθορίστηκε να είναι $P=300$ ώστε να αποφευχθεί πρόωρη σύγκλιση. Ο αριθμός γενιών καθορίστηκε να είναι $G=300$, προκειμένου τουλάχιστον το 75% του πληθυσμού να μπορέσει να συγκλίνει στις βέλτιστες λύσεις.

Οι παράμετροι ρ και θ, που χρησιμοποιούνται στην αρχικοποίηση με χρήση του NMIFS ορίστηκαν να διακυμαίνονται στο διάστημα $[0\ ,\ 0{,}5]$. Το κριτήριο που ορίστηκε για την επιλογή του καλύτερου συνδυασμού παραμέτρων ήταν η μεγαλύτερη αύξηση του αριθμού των ατόμων που περιείχαν οποιοδήποτε κομμάτι από τις λύσεις σε σχέση με τον γενετικό αλγόριθμο που έχει τυχαία αρχικοποίηση. Τα καλύτερα αποτελέσματα λήφθηκαν για τον συνδυασμό των παραμέτρων $\rho=0{,}15$ και $\theta=0{,}3$. Αυτές οι

παράμετροι πρέπει περιορίζονται ώστε να λαμβάνουν μικρές τιμές προκειμένου να διατηρηθεί ποικιλία μέσα στον πληθυσμό.

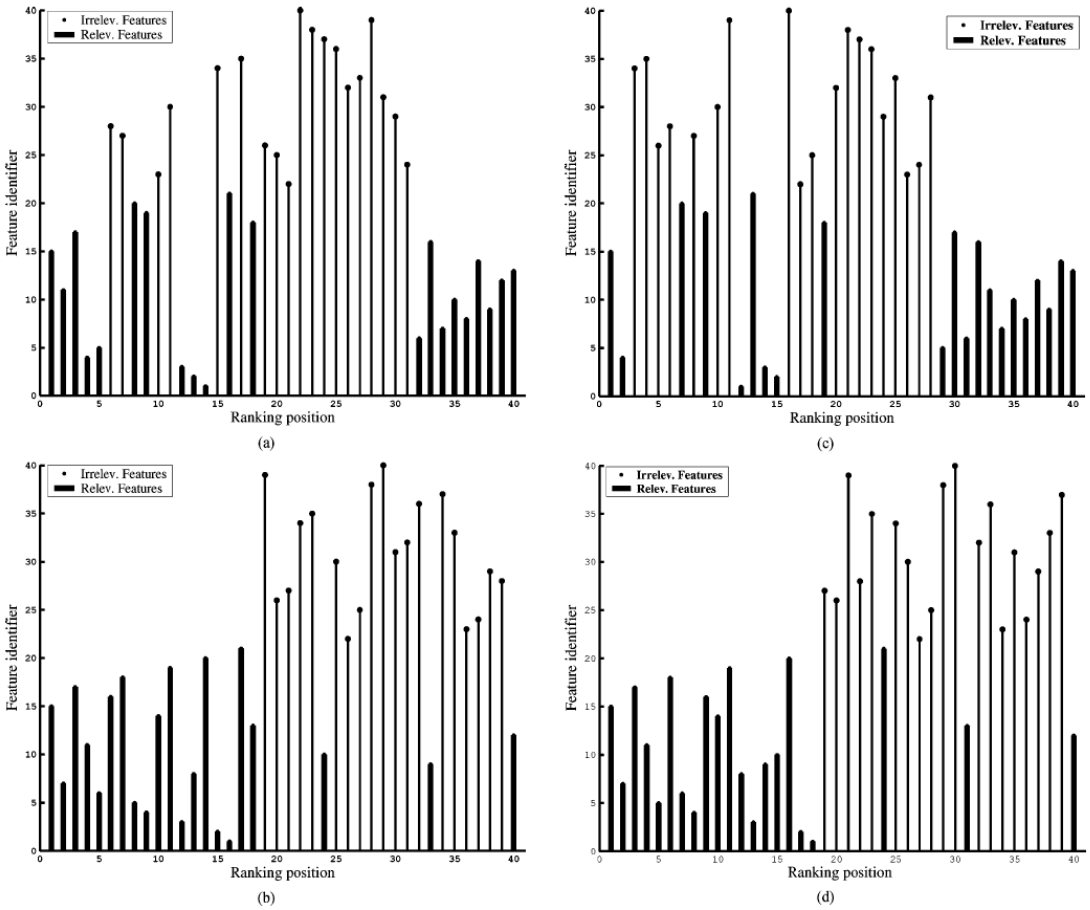
Επιπλέον, οι παράμετροι δ, p_a και p_i , που χρησιμοποιούνται στην μετάλλαξη με χρήση του NMIFS καθορίστηκαν ως ακολούθως: η παράμετρος δ ορίστηκε να διακυμαίνεται στο διάστημα $[0\ ,\ 0{,}5]$, ενώ οι παράμετροι p_a και p_i ορίστηκαν να διακυμαίνονται στο διάστημα $[0,1]$. Το κριτήριο που τέθηκε για την επιλογή του καλύτερου συνδυασμού παραμέτρων ήταν η αποτελεσματικότητα του τελεστή μετάλλαξης. Τα καλύτερα αποτελέσματα λήφθηκαν για τις τιμές των παραμέτρων $\delta=0{,}3$, $p_a=0{,}3$ και $p_i=0{,}5$. Αυτά τα αποτελέσματα συνιστούν ότι ο τελεστής μετάλλαξης τείνει να αφαιρεί ήδη επιλεγμένα χαρακτηριστικά από το υποσύνολο πιο συχνά απ'ότι εισάγει νέα χαρακτηριστικά σε αυτό. Ύστερα εκπαιδεύτηκε ένας ταξινομητής MLP με αρχιτεκτονική 14-6-1, δηλαδή με 14 εισόδους έξι κρυμμένες μονάδες και μία έξοδο, σε 100 epochs.

Με βάση τα αποτελέσματα του μη γραμμικού AND προβλήματος, οι παράμετροι του GAMIFS για προσομοιώσεις πάνω σε άλλα σύνολα δεδομένων οριστικοποιήθηκαν στις τιμές $P=100$, $G=200$, $\rho=0{,}15$, $\theta=0{,}3$, $\delta=0{,}3$, $p_a=0{,}3$ και τέλος $p_i=0{,}5$.

Το επόμενο πρόβλημα που εξετάστηκε είναι το σύνολο δεδομένων κυματομορφής του Breiman. Αυτό το σύνολο δεδομένων εισήγαγε ουσιαστικά το πρόβλημα αναγνώρισης του τύπου της κυματομορφής. Αρχικά επιλέγονται 21 σημεία ως δείγματα από κάθε μια από τις τρεις κυματομορφές τα οποία θα χρησιμοποιηθούν σαν χαρακτηριστικά. Μετά δημιουργούνται τρεις κλάσεις οι C_1 , C_2 , C_3 , από τυχαίους κυρτούς συνδυασμούς δύο κυματομορφών από τις τρεις $(1,2)$, $(1,3)$, $(2,3)$, αντίστοιχα. Στην περίπτωση που το πρόβλημα έχει θόρυβο, κάθε υπόδειγμα κυματομορφής αυξάνεται κατά 19 δείγματα (χαρακτηριστικά) τα οποία λαμβάνονται από την κανονική κατανομή $N(0,1)$. Η βάση δεδομένων του Breiman περιέχει 1000 υποδείγματα κυματομορφών (33% σε κάθε κλάση). Η ιδανική σειρά επιλογής είναι να επιλεγούν από τον αλγόριθμο πρώτα τα σχετικά χαρακτηριστικά 1-21, και μετά τα άσχετα 22-40.

Στο σχήμα που ακολουθεί (Σχήμα 4.8) απεικονίζονται τα αποτελέσματα που προήλθαν από τους αλγορίθμους MIFS, MIFS-U, mRMR και NMIFS με είσοδο το σύνολο δεδομένων του Breiman. Και εδώ ο άξονας x αναπαριστά την σειρά που επιλέχθηκαν τα χαρακτηριστικά από τον κάθε αλγόριθμο, ενώ ο άξονας y αναπαριστά τον πραγματικό

αριθμό που έχουν τα χαρακτηριστικά στο πρόβλημα.



Σχήμα 4.8 Επιλογή χαρακτηριστικών για το σύνολο δεδομένων του Breiman για (a) MIFS με β=0,3 (b) MIFS-U με β=0,4 (c) mRMR (d) NMIFS

Τα Σχήματα 4.8(a) και 4.8(c) δείχνουν ότι ο MIFS και ο mRMR είχαν χαμηλή απόδοση επιλέγοντας μόνο πέντε και δύο χαρακτηριστικά πρώτα, αντίστοιχα. Και οι δύο προσεγγίσεις αφήνουν τελευταία πάνω από τα μισά σχετικά χαρακτηριστικά. Στα Σχήματα 4.8(b) και 4.8(d) από την άλλη φαίνεται ότι και ο MIFS-U αλλά και ο NMIFS επιλέγουν 18 από τα 21 σχετικά χαρακτηριστικά στην σωστή σειρά. Σημειώνεται ότι για τον MIFS-U η βέλτιστη τιμή για την μεταβλητή β ($\beta = 0{,}6$) επιλέχθηκε από το διάστημα $[0,1]$. Αυτό ήταν εφικτό διότι στο σύνολο δεδομένων του Breiman είναι ένα τεχνητό πρόβλημα, και οι λύσεις είναι ήδη γνωστές.

Τα υποσύνολα χαρακτηριστικών που επιλέχθηκαν από τον NMIFS, τον GAMIFS, τον DC-GA χωρίς μετάλλαξη και τον DC-GA με μετάλλαξη (Estévez P. A., 1998) εισήχθησαν σε ένας MLP για να ελεγχθούν τα ποσοστά των ταξινομήσεων. Η αρχιτεκτονική του MLP

ήταν 40-15-3. Ο Πίνακας 4.4 δείχνει το ποσοστό των σωστών ταξινομήσεων για κάθε υποσύνολο χαρακτηριστικών που επιλέχθηκε από τους αλγόριθμους και χρησιμοποιήθηκε από τον MLP.

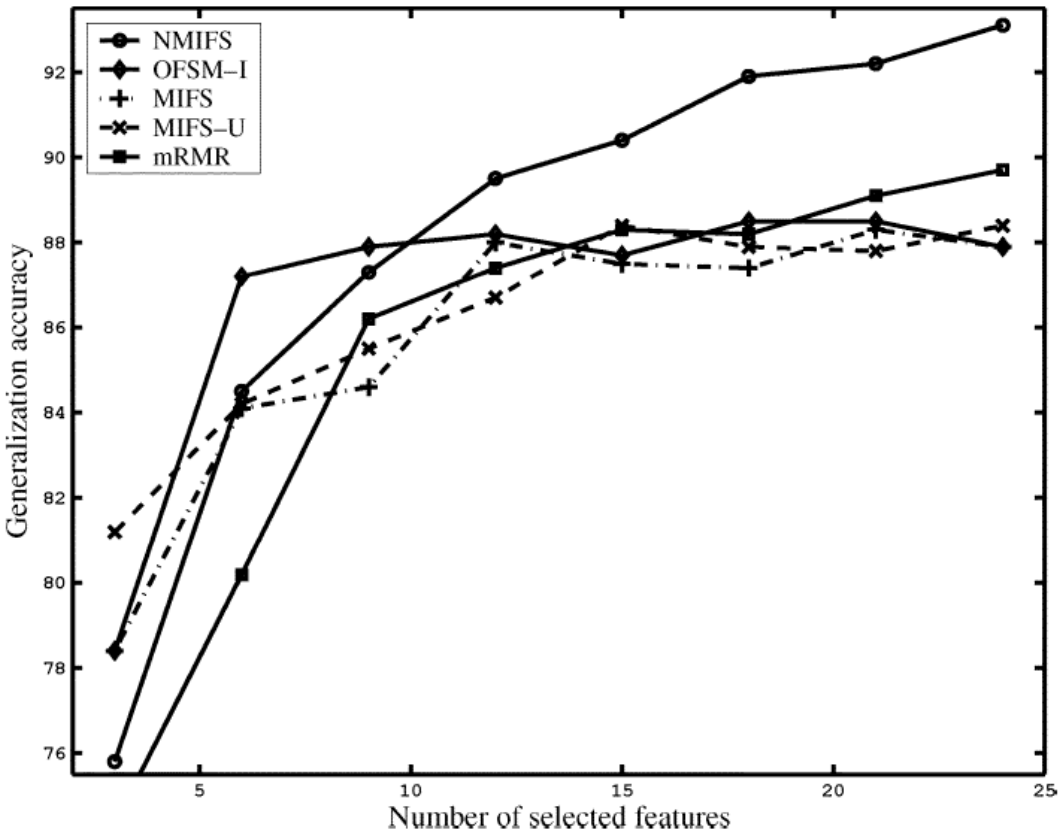
Πίνακας 4.4 Ποσοστά ταξινόμησης του MLP για το σύνολο δεδομένων του Breiman με είσοδο τα υποσύνολα χαρακτηριστικών που επιλέχθηκαν από διάφορες μεθόδους

Μέθοδος Επιλογής Χαρακτηριστικών	Πλήθος Χαρακτηριστικών	Ποσοστό Σωστών Ταξινομήσεων	P-value	Γενιές
GAMIFS	13	87,84	-	**89**
DC-GA (με μετάλλαξη)	13	86,4	0,005	92
DC-GA (χωρίς μετάλλαξη)	16	87,92	-	146
NMIFS	13	81,52	$3{,}03 \cdot 10^{-11}$	-
Όλα	40	83,32		

Για τα 13 χαρακτηριστικά που επιλέχθηκαν, ο GAMIFS έδωσε καλύτερη απόδοση από τον NMIFS και τον DC-GA με μετάλλαξη. Αυτή η διαφορά είναι στατιστικά σημαντική στο επίπεδο σημαντικότητας 0,01 σύμφωνα με το t-test. Αυτό φαίνεται και στην στήλη “P-value ”. Τα αποτελέσματα της ταξινόμησης του GAMFIS είναι καλύτερα ακόμα και από την περίπτωση που χρησιμοποιούνται όλα τα χαρακτηριστικά (40).

Το επόμενο πείραμα που θα αναλυθεί είναι αυτό που έγινε πάνω στα δεδομένα spambase. Η βάση δεδομένων spambase (Newman D., 1998) αποτελείται από 4.601 δείγματα (e-mail) και 57 χαρακτηριστικά για το καθένα, που διαχωρίζονται σε δύο κλάσεις (spam και όχι spam). Προκειμένου να αξιολογηθούν τα υποσύνολα των χαρακτηριστικών που δόθηκαν σαν αποτέλεσμα από διάφορες μεθόδους, εκπαιδεύτηκε πάνω σε αυτά ένας ταξινομητής MLP. Η αρχιτεκτονική που χρησιμοποιήθηκε ήταν 57-12-1. Οι δοκιμές έγιναν για διάφορα μεγέθη του υποσυνόλου χαρακτηριστικών, αναλυτικότερα επιλέχθηκαν οι αριθμοί 3, 6, 9, 12, 15, 18, 21 και 24 για το πλήθος των χαρακτηριστικών που έπρεπε να επιλέξει ο κάθε αλγόριθμος.

Το Σχήμα 4.9 αναπαριστά την ακρίβεια γενίκευσης του ταξινομητή MLP, όταν αυτός έχει ως είσοδο το υποσύνολο των χαρακτηριστικών που επέλεξαν οι μέθοδοι NMIFS, MIFS, MIFS-U, mRMR και OFS-MI.



Σχήμα 4.9 Ακρίβεια γενίκευσης ταξινομητή στο σύνολο δεδομένων spambase

Τα αποτελέσματα του OFS-MI πάνω στο σύνολο δεδομένων spambase αναπαράχθηκαν για λόγους σύγκρισης (Chow T. W., 2005). Για τον MIFS και τον MIFS-U επιλέχθηκε η καλύτερη τιμή της παραμέτρου β, αλλά αξίζει να αναφερθεί ότι για το διάστημα $[0{,}3\ ,\ 1{,}0]$ δεν εντοπίστηκαν σημαντικές διαφορές. Από το Σχήμα 4.9 φαίνεται ότι τα καλύτερα αποτελέσματα λήφθηκαν από τον αλγόριθμο NMIFS με 12 ή περισσότερα επιλεγμένα χαρακτηριστικά. Το σφάλμα της εσφαλμένης ταξινόμησης του NMIFS, όταν στο υποσύνολο επιλεγμένων χαρακτηριστικών υπάρχουν 24 χαρακτηριστικά, είναι κοντά στο 7%. Για λιγότερα από δέκα χαρακτηριστικά, ο OFS-MI έχει καλύτερα αποτελέσματα από τον NMIFS. Τέλος η απόδοση του NMIFS ξεπέρασε αυτήν των mRMR, MIFS και MIFS-U για όλα τα πλήθη χαρακτηριστικών, εκτός από την περίπτωση των τριών επιλεγμένων χαρακτηριστικών όπου ο MIFS-U είχε καλύτερη απόδοση από τον NMIFS.

Ο Πίνακας 4.5 που ακολουθεί δείχνει το ποσοστό των σωστών ταξινομήσεων με την χρήση των τριών επιλεγμένων χαρακτηριστικών από τις διάφορες μεθόδους, ως είσοδο στον MLP.

Πίνακας 4.5 Ποσοστά ταξινόμησης του MLP για το σύνολο δεδομένων spambase με είσοδο τα υποσύνολα χαρακτηριστικών που επιλέχθηκαν από διάφορες μεθόδους

Μέθοδος Επιλογής Χαρακτηριστικών	Πλήθος Χαρακτηριστικών	Ποσοστό Σωστών Ταξινομήσεων
GAMIFS	3	83,5
NMIFS	3	75,8
MIFS	3	78,4
MIFS-U	3	81,2
OFS-MI	3	78,4
Όλα	57	93,64

Είναι φανερό ότι ο GAMIFS είχε καλύτερα αποτελέσματα από τους NMIFS, MIFS, MIFS-U και OFS-MI.

Το τελευταίο πείραμα που έγινε είναι πάνω στο σύνολο δεδομένων από sonar (Newman D., 1998), το οποίο περιέχει 204 δείγματα από ενδείξεις sonar με 60 χαρακτηριστικά για κάθε δείγμα, όταν τέθηκε κοντά σε έναν μεταλλικό κύλινδρο και μια πέτρα. Η αρχιτεκτονική του ταξινομητή MLP ήταν 60-5-2.

Ο Πίνακας 4.6 δείχνει το ποσοστό των σωστών ταξινομήσεων που έκανε ο MLP όταν χρησιμοποίησε ως είσοδο τα χαρακτηριστικά που επιλέχθηκαν από τον NMIFS, τον mRMR, τον MIFS και τον MIFS-U.

Πίνακας 4.6 Ποσοστό σωστών ταξινομήσεων στο πείραμα πάνω στο σύνολο δεδομένων από sonar

Πλήθος Χαρακτηριστικών	NMIFS	mRMR	MIFS β=0,3	MIFS β=0,5	MIFS β=0,7	MIFS β=0,9	MIFS-U β=0,3	MIFS-U β=0,5	MIFS-U β=0,7	MIFS-U β=0,9
4	80,19	78,46	79,23	77,69	78,17	77,69	73,85	73,85	75,58	76,25
7	85,19	80,09	79,81	83,65	83,46	83,65	74,81	74,81	76,54	76,92
11	86,36	79,80	80,58	84,62	83,85	84,62	77,31	77,31	77,31	76,35
15	86,73	81,06	80,96	85,96	85,19	85,77	79,81	84,04	84,04	82,98
Όλα (60)	80,67									

Η απόδοση του NMIFS είναι καλύτερη από την απόδοση του mRMR, για όλα τα μεγέθη υποσυνόλων επιλεγμένων χαρακτηριστικών, καθώς επίσης και από του MIFS και του MIFS-U για όλους τους συνδυασμούς του μεγέθους του υποσυνόλου επιλεγμένων χαρακτηριστικών με τις δοκιμαστικές τιμές της παραμέτρου β όσον. Αξίζει επίσης να σημειωθεί ότι ο MIFS έδωσε καλύτερα αποτελέσματα από τον MIFS-U πάλι για όλα τα μεγέθη υποσυνόλων επιλεγμένων χαρακτηριστικών όταν τους εκχωρείται η ίδια τιμή για την παράμετρο β.

Επιπρόσθετα, ο Πίνακας 4.7 δείχνει το ποσοστό των σωστών ταξινομήσεων όταν χρησιμοποιήθηκαν ως είσοδος στον ταξινομητή MLP τα χαρακτηριστικά που επιλέχθηκαν από τον GAMIFS και άλλες μεθόδους.

Πίνακας 4.7 Σύγκριση του GAMIFS με άλλες μεθόδους στο σύνολο δεδομένων από sonar

Μέθοδος Επιλογής Χαρακτηριστικών	Πλήθος Χαρακτηριστικών	Ποσοστό Σωστών Ταξινομήσεων	P-value	Γενιές
GAMIFS	11	90,96	-	**186**
DC-GA (με μετάλλαξη)	14	90,38	-	104
DC-GA (χωρίς μετάλλαξη)	15	87,06	-	135
NMIFS	11	86,36	0,01	-
Όλα	60	80,67		

Ο GAMIFS ξεπέρασε και σε απόδοση τον DC-GA με μετάλλαξη και τον DC-GA χωρίς μετάλλαξη, βρίσκοντας την καλύτερη λύση με μικρότερο αριθμό επιλεγμένων χαρακτηριστικών. Τα αποτελέσματα της ταξινόμησης του GAMIFS και εδώ είναι καλύτερα ακόμα και από την περίπτωση που χρησιμοποιούνται όλα τα χαρακτηριστικά (60).

Στον πίνακα που ακολουθεί (Πίνακας 4.8) αναγράφονται οι χρόνοι εκτέλεσης του NMIFS και του GAMIFS σε διάφορα σύνολα δεδομένων. Προκειμένου να ελεγχθεί η επεκτασιμότητα των παραπάνω μεθόδων, συμπεριλήφθηκαν ακόμα τρία σύνολα δεδομένων, τα οποία έχουν μεγάλο αριθμό χαρακτηριστικών. Τα σύνολα προήλθαν από το feature selection challenge του συνεδρίου 2003 Neural Information Processing Systems (NIPS). Τα σύνολα που επιλέχθηκαν είναι το Madelon, το Gisette και το Arcene. Οι χρόνοι εκτέλεσης μετρήθηκαν σε σύστημα Pentium IV, 1, 8-GHz, 1-GB RAM.

Πίνακας 4.8 Χρόνοι εκτέλεσης των NMIFS και GAMIFS στα διάφορα σύνολα δεδομένων

Σύνολο Δεδομένων	Πλήθος Χαρακτηριστικών	Πλήθος Δειγμάτων	NMIFS	GAMIFS
Sonar	60	208	1 s	**3,6 hrs.**
Breiman	40	500	4 s	29,1 hrs.
Spambase	57	2.300	7 s	70 hrs.
Madelon	500	2.600	16 min.	-
Gisette	5.000	7.000	4,5 hrs.	-
Arcene	10.000	200	6 hrs.	-

Είναι φανερό ότι ο NMIFS μπορεί να εφαρμοστεί αποτελεσματικά σε σύνολα δεδομένων με περισσότερα από 10.000 χαρακτηριστικά. Ο χρόνος εκτέλεσης του NMIFS μπορεί να ελαττωθεί εάν χρησιμοποιηθεί μια γρηγορότερη μέθοδος εκτίμησης εντροπιών (Vasicek O., 1976).

Λόγω της αναπαράστασης που χρησιμοποιήθηκε για τον πληθυσμό του GAMIFS, το μέγεθος του χώρου αναζήτησης (των πιθανών ατόμων) είναι 2^L , όπου L είναι το πλήθος των χαρακτηριστικών. Γι'αυτόν τον λόγο, το πλήθος των χαρακτηριστικών που εισάγονται στον GAMIFS περιορίζεται να είναι μικρότερο του 100 ώστε να είναι υπολογιστικά αποδοτικός. Από την άλλη, στην περίπτωση που για παράδειγμα το πλήθος των γενιών είναι $G=200$ και ο πληθυσμός έχει μέγεθος $P=100$, ο συνολικός αριθμός συνδυασμών που θα αναζητηθούν από τον GAMIFS είναι $G\times P=20.000$, ένα μέγεθος που είναι μόλις ένα μικρό κλάσμα του χώρου αναζήτησης που προκύπτει για $L>20$. Σαν συνέπεια ο κατάλληλος χώρος αναζήτησης για τον GAMIFS είναι ανάμεσα στα 20 και 100 χαρακτηριστικά, αφού για $L<20$ ο χώρος αναζήτησης είναι συγκρίσιμος με τον αριθμό συνδυασμών του GAMIFS, και για $L>100$ ο χώρος αναζήτησης γίνεται απαγορευτικά μεγάλος. Μια διαφορετική προσέγγιση για τα μεγάλα σύνολα δεδομένων είναι να εφαρμοστεί πρώτα ο NMIFS, έτσι ώστε να ελαττωθεί ο αριθμός των χαρακτηριστικών περίπου στο 100, και μετά να εκτελεστεί ο GAMIFS πάνω στο επιλεγμένο υποσύνολο. Τέλος, επειδή το κομμάτι του GAMIFS που είναι υπολογιστικά πιο ακριβό, είναι αυτό του υπολογισμού της απόδοσης που απαιτεί την εκπαίδευση του νευρωνικού δικτύου MLP, μια άλλη επιλογή είναι η χρήση ενός απλούστερου και γρηγορότερου ταξινομητή.

4.3.6 Συμπεράσματα

Η μέθοδος επιλογής χαρακτηριστικών NMIFS που βασίζεται στην αμοιβαία πληροφορία, είναι μια βελτίωση των μεθόδων MIFS, MIFS-U και mRMR. Το μέτρο που χρησιμοποιήθηκε ως κριτήριο επιλογής είναι η κανονικοποιημένη αμοιβαία πληροφορία, η οποία μειώνει την μεροληψία προς τα χαρακτηριστικά με μεγάλο εύρος τιμών και περιορίζει το εύρος στο διάστημα $[0,1]$. Ακόμη ο NMIFS εξαλείφει την ανάγκη καθορισμού παραμέτρων από τον χρήστη, όπως είναι για παράδειγμα η παράμετρος β στον MIFS και τον MIFS-U. Αυτό είναι πολύ βοηθητικό στην πράξη διότι δεν υπάρχει σαφής τρόπος επιλογής βέλτιστης τιμής για την παράμετρο σε πραγματικά προβλήματα. Ο NMIFS είναι μια μέθοδος φίλτρου που επιλέγει τα βέλτιστα χαρακτηριστικά ανεξάρτητα από οποιαδήποτε μορφή εκμάθησης. Από τα πειράματα φάνηκε ότι ο NMIFS είχε καλύτερη απόδοση από τον MIFS, τον MIFS-U και τον mRMR σε διάφορα τεχνητά σύνολα δεδομένων αλλά και διάφορα γνωστά προβλήματα που χρησιμοποιούνται για την σύγκριση των αποτελεσμάτων των διαφόρων μεθόδων. Το μόνο πρόβλημα που διαφοροποιείται ως προς τα αποτελέσματά του είναι το σύνολο δεδομένων του Breiman, ακόμη και στο οποίο όμως ο NMIFS δεν είχε χειρότερη απόδοση, αλλά ίδια με τον MIFS-U, και επιπλέον ο NMIFS δεν χρειαζόταν προσδιορισμό της παραμέτρου β όπως ο MIFS-U. Όσον αφορά τον NMIFS συγκριτικά με τον mRMR, είναι εμφανές ότι η κανονικοποίηση της αμοιβαίας πληροφορίας έχει μεγάλο θετικό αντίκτυπο την απόδοση. Στο πρόβλημα χρονοσειρών του συνόλου δεδομένων gas furnace, ο NMIFS είχε είτε ίδια είτε καλύτερα αποτελέσματα από τα μοντέλα περιτυλίγματος ασαφούς λογικής.

Η δεύτερη μέθοδος επιλογής χαρακτηριστικών που περιγράφηκε, η GAMIFS, είναι μια υβριδική μέθοδος φίλτρου-περιτυλίγματος, η οποία συνδυάζει τα πλεονεκτήματα του NMIFS με γενετικούς αλγορίθμους. Η ακρίβεια του εκπαιδευμένου ταξινομητή MLP χρησιμοποιείται για να αξιολογηθεί η ποιότητα των υποσυνόλων χαρακτηριστικών, αντ'αυτού όμως μπορεί να χρησιμοποιηθεί οποιοσδήποτε ταξινομητής στο κομμάτι περιτυλίγματος του αλγορίθμου. Ο NMIFS χρησιμοποιείται εδώ σε δύο επίπεδα. Αφενός για να γίνει καλή αρχικοποίηση του γενετικού αλγορίθμου και αφετέρου στον τελεστή μετάλλαξης. Ο τελεστής μετάλλαξης επιτρέπει την εισαγωγή και απαλοιφή χαρακτηριστικών στα άτομα, με κριτήριο εισαγωγής το κριτήριο επιλογής του NMIFS, και

κριτήριο απαλοιφής το πιο πλεονάζον και το πιο άσχετο χαρακτηριστικό. Ο GAMIFS ξεπερνά τους περιορισμούς των αυξητικών αλγορίθμων αναζήτησης στους οποίους ανήκουν για και οι NMIFS, MIFS, MIFS-U και mRMR, όπως είναι η αδυναμία εύρεσης εξαρτήσεων ανάμεσα σε ομάδες χαρακτηριστικών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Κουκουβίνος, Χ. ≪(2003). Θεωρία Πληροφοριών και Κωδίκων≫.

[2] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning, *IEEE Transactions on Neural Networks,* vol. 5, Issue 4, p.p. 537-550.

[3] Box, G.E.P., Jenkins, G.M. (2003). Time Series Analysis. Cambridge, *Cambridge Univ. Press*.

[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stonem C.J. (1984). Classification and Regression Trees, *Chapman & Hall*.

[5] Chow, T.W., Huang, D. (2005). Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information, *IEEE Trans. Neural Netw.,* vol. 16, Issue 1, p.p. 213-224.

[6] Cover, T., Thomas, J. (1991). Elements of Information Theory, *1st edn., John Wiley & Sons*.

[7] Efron, B., Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, *The American Statistician*, vol. 37, Issue 1.

[8] Estévez, P.A., Caballero, R. (1998). A niching genetic algorithm for selecting features for neural networks classifiers, *Perspectives in Neural Computation (ICANN'98), Springer- Verlag,* p.p. 311-316.

[9] Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M. (2009). Normalized Mutual Information Feature Selection, *IEEE Transactions on Neural Networks*, vol. 20, Issue 2, p.p. 189-201.

[10] Forman, G. (2003). An experimental study of feature selection metrics for text categorization, *Journal of Machine Learning Research 3*, p.p. 1289-1305.

[11] Fraser, A.M., Swinney, H.L. (1986). Independent coordinates for strange attractors from mutual information, *Phys. Rev. A, Gen. Phys.,* vol. 33, Issue 2, p.p. 1134-1140.

[12] Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, *2nd ed*. *Academic Press*.

[13] Guiasu, S. (1977). Information Theory with Applications, *McGraw-Hill Inc.*

[14] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, *N´edellec, C., Rouveirol, C. (eds.) ECML 1998, LNCS, Springer,* vol. 1398, p.p. 137-142.

[15] John, G.H., Kohavi, R., Pfleger, K. (1994). Irrelevant feature and the subset selection problem, *Proceedings of the Eleventh International Conference in Machine Learning,* p.p. 121-129.

[16] Kakoyan, M. (2010). Επιλογή χαρακτηριστικών (feature selection) από βάσεις δεδομένων με φίλτρο αμοιβαίας πληροφορίας (mutual information filter), *ΤΕΙ Σερρών, Διπλωματική Εργασία*.

[17] Khan, S., Bandyopadhyay, S., Ganguly, A.R., Saigal, S., Erickson, D.J., Protopopescu, V., Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data, *Physical Review E 76*.

[18] Kira, K., Rendell, L. (1992). A Practical Approach to Feature Selection, *ML92 Proceedings of the ninth international workshop on Machine learning Pages,* p.p. 249-256.

[19] Kwak, N., Choi, C.H. (1999). Improved mutual information feature selector for neural networks in supervised learning, *Proceedings of the IJCNN 1999, 10th International Joint Conference on Neural Networks,* p.p. 1313-1318.

[20] Kwak, N., Choi, C.H. (2002). Input feature selection for classification problems, *IEEE Transactions on Neural Networks,* vol. 13, Issue 1, p.p. 143-159.

[21] Mahfoud, S.W. (1995). Niching methods for genetic algorithms, *Ph.D. dissertation, Dept. General Eng., Univ. Illinois at Urbana-Champaign*.

[22] McCallum, A., Nigam, K. (1998). A comparison of event models for naive Bayes text classification, *Proceedings of the AAAI-1998 Workshop on Learning for Text Categorization*.

[23] Menezes A., Van Oorschot, P.C., Vanstone, S.A. (1996). Handbook of Applied Cryptography, *CRC Press*.

[24] Mitchell, M. (1996). An Introduction to Genetic Algorithms. Cambridge, *MIT Press*.

[25] Newman, D.J., Hettich, S., Blake, S.L., Merz, C.J. (1998). UCI Repository of machine learning databases, *http://www.ics.uci.edu/~mlearn/MLRepository.html*.

[26] Novovicova, J., Somol, P., Haindl, M., Pudil, P. (2007). Conditional Mutual Information Based Feature Selection for Classification Task, *Progress in Pattern Recognition, Image Analysis and Applications, Proceedings of the 12th Iberoamericann Congress on Pattern Recognition*, *CIARP*, *Springer*, p.p. 417-426.

[27] Pedrycz, W. (1984). An identification algorithm in fuzzy relational systems, *Fuzzy Sets Syst.,* vol. 13, p.p. 153-167.

[28] Peng, H., Long, F., Ding, C. (2005). Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min Redundancy, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, p.p. 1226-1238.

[29] Press, W., Flannery, B., Teukolsky, S., Vetterling, W. (1992). Numerical Recipes in C, *2nd ed. Cambridge Univ. Press*.

[30] Saito, K., Nakano, R. (1997). Partial BFGS update and efficient step-length calculation for three-layer neural networks, *Neural Computation*, vol. 9, Issue 1, p.p. 123-141.

[31] Schaffernicht, E., Gross, H. (2011). Weighted Mutual Information for Feature Selection, *Proc. 21. Int. Conf. On Artificial Neural Networks (ICANN 2011), Springer,* p.p. 181-188.

[32] Schaffernicht, E., Kaltenhaeuser, R., Verma, S.S., Gross, H.M. (2010). On estimating mutual information for feature selection, *Diamantaras K., Duch W., Iliadis L.S. (eds.), ICANN 2010, LNCS, Springer,* vol. 6352, p.p. 362-367.

[33] Schaffernicht, E., Stephan, V., Gross, H.M. (2007). An efficient search strategy for feature selection using chow-liu trees, *de S´a, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007, LNCS, Springer,* vol. 4669, p.p. 190-199.

[34] Sebastiani, F. (2002). Machine learning in automated text categorization, *ACM*

Computing Surveys, vol. 34, Issue 1, p.p. 1-47.

[35] Shannon, C. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal 27,* vol. 3, p.p. 379-423.

[36] Stone, M. (1974) Cross-validation and multinomial prediction, *Biometrika 61,* vol. 3, p.p. 509-515.

[37] Sugeno, M., Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modeling, *IEEE Trans. Fuzzy Syst.,* vol. 1, Issue 1, p.p. 7-31.

[38] Tesmer, M., Estévez, P.A., (2004). AMIFS: Adaptive feature selection by using mutual information, *Proc. IEEE Int. Joint Conf. Neural Netw.,* p.p. 303-308.

[39] Tong, R. (1980). The evaluation of fuzzy models derived from experimental data, *Fuzzy Sets Syst.,* vol. 4, p.p. 1-12.

[40] Torkkola, K. (2003). Feature Extraction by Non Parametric Mutual Information Maximization, *Journal of Machine Learning Research 3,* p.p. 1415-1438.

[41] Torkkola, K. (2006). Information-Theoretic Methods, *Feature Extraction Foundations and Applications StudFuzz 207, Springer,* p.p. 167-185.

[42] Van Dijck, G., Van Hulle, M.M. (2006). Speeding up the wrapper feature subset selection in regression by mutual information relevance and redundancy analysis, *Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006, LNCS, Springer,* vol. 4131, p.p. 31-40.

[43] Vasicek, O. (1976). A test for normality based on sample entropy, *Roy, J., Statist. Soc. B,* vol. 31, p.p. 632-636.

[44] Xu, C., Yong, Z. (1987). Fuzzy model identification and self-learning for dynamic systems, *IEEE Trans. Syst. Man Cybern.,* vol. SMC-17, Issue 4, p.p. 683-689.

[45] Yang, Y. (2001). A study on thresholding strategies for text categorization, *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*.

[46] Δημητριάδης, Σ. (2010). Επιλογή χαρακτηριστικών με γενετικούς αλγορίθμους σε προβλήματα υπολογιστικής όρασης, *Πανεπιστήμιο Ιωαννίνων, Μεταπτυχιακή Εργασία*.

[47] Ζορκάδης, Β. (2002). Θεωρία Πληροφορίας και Κωδικοποίησης, *Τόμος Α', Ελληνικό Ανοικτό Πανεπιστήμιο*.

[48] Θεοχαρίδης, Γ. (2009). Αλγόριθμοι Ανάπτυξης Ταξινομητών Βασισμένων σε Ασαφείς Κανόνες, για την Ταξινόμηση Προτύπων, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Διπλωματική Εργασία*.

[49] Καραμάνου, Δ. (2011). Θεωρία Πληροφορίας ή Θεωρία Πληροφοριών Κανάλι Σύστημα, *Πανεπιστήμιο Πειραιά, Διπλωματική Εργασία*.

[50] Μπαστάς, Ν. (2007). Επιλογή χαρακτηριστικών και ταξινόμηση δεδομένων με χρήση νευρωνικών δικτύων RBF, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Διπλωματική Εργασία*.

[51] Παντελής, Α. (2010). Εξόρυξη γνώσης από δεδομένα με διατήρηση της ιδιωτικότητας χρησιμοποιώντας νευρωνικά δίκτυα RBF για οριζόντια κατατετμημένα δεδομένα σε περιβάλλον μη έμπιστων χρηστών, *Πανεπιστήμιο Αιγαίου, Διπλωματική Εργασία*.

[52] Πετρόχειλος, Ο. (2009). Επιλογή Χαρακτηριστικών για Προβλήματα Ταξινόμησης, *Πανεπιστήμιο Ιωαννίνων, Μεταπτυχιακή Εργασία*.

[53] Σαλάππα, Α. (2005). Αλγόριθμοι Επιλογής Χαρακτηριστικών σε Προβλήματα Ταξινόμησης: Μία Πειραματική Ανάλυση, *Πολυτεχνείο Κρήτης, Μεταπτυχιακή Εργασία*.