ΕΘΝΙΚΟ ΜΕΤΣΟΒΕΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΜΕΤΡΩΝ ΠΛΗΡΟΦΟΡΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΧΑΤΖΗΖΑΧΑΡΙΑΣ

ΕΠΙΒΛΕΠΩΝ: ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ
ΚΑΘΗΓΗΤΗΣ ΕΜΠ

ΑΘΗΝΑ ΙΟΥΛΙΟΣ 2014

“One could perhaps describe the situation by saying that God is a mathematician of a very high order, and He used very advanced mathematics in constructing the universe”

Paul Dirac

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα εργασία, αποτελεί το επιστέγασμα μιας πεντάχρονης επίπονης προσπάθειας, η οποία εν τέλει στέφθηκε με επιτυχία με την επιτυχή ολοκλήρωση των σπουδών μου στο ΕΜΠ.

Πρωτίστως θα ήθελα να ευχαριστήσω τον καθηγητή της ΣΕΜΦΕ κύριο Χρήστο Κουκουβίνο, ο οποίος μου έδωσε την ευκαιρία να δουλέψω μαζί του πάνω σε ένα τόσο ενδιαφέρον θέμα, καθώς και για την γενικότερη συνεισφορά του καθ' όλη τη διάρκεια των σπουδών μου.

Πάνω από όλα όμως, θα ήθελα να ευχαριστήσω τους γονείς μου, Παναγιώτη και Μαρία, οι οποίοι μου εμφύσησαν από μικρό την αγάπη προς την Ελλάδα, αποτελώντας στην ουσία τον κύριο παράγοντα που με ώθησε να σπουδάσω τελικά στην Αθήνα και να ζήσω αυτή την ανεπανάληπτη εμπειρία της φοιτητικής ζωής εδώ και οι οποίοι στάθηκαν δίπλα μου όλα αυτά τα χρόνια στηρίζοντας με κάθε δυνατό τρόπο τις προσπάθειές μου.

Κωνσταντίνος Χατζηζαχαρίας

ΠΕΡΙΛΗΨΗ

Αναμφίβολα, ένα από τα σημαντικότερα προβλήματα που υπάρχουν όσον αφορά την ταξινόμηση ενός συνόλου, έχει να κάνει με τον μεγάλο αριθμό δεδομένων από τα οποία αυτό αποτελείται, πράγμα το οποίο δυσχεραίνει σε μεγάλο βαθμό την όλη διαδικασία, καθιστώντας την ταξινόμηση, πολλές φορές αναποτελεσματική, με την εμφάνιση μεγάλου αριθμού σφαλμάτων. Προκειμένου να επιλυθεί το πρόβλημα αυτό, εφαρμόζονται στα σύνολα δεδομένων, διάφορες τεχνικές επιλογής χαρακτηριστικών, οι οποίες χρησιμοποιούνται για την επιλογή ενός μικρού υποσυνόλου αποτελούμενο από τα σημαντικότερα και πιο χρήσιμα χαρακτηριστικά, με αποτέλεσμα το νέο αυτό σύνολο να μπορεί να ταξινομηθεί με ακρίβεια. Υπάρχουν πάρα πολλές τέτοιες μεθόδοι επιλογής βέλτιστου υποσυνόλου με την απομόνωση των καλύτερων χαρακτηριστικών από το σύνολο. Στην εργασία αυτή, επικεντρωνόμαστε αποκλειστικά στην ανάλυση μεθόδων επιλογής χαρακτηριστικών, οι οποίες ειναι βασισμένες αποκλειστικά πάνω σε μέτρα πληροφορίας όπως την εντροπία και την υπό συνθήκη αμοιβαία πληροφορία.

Συγκεκριμένα, στο πρώτο κεφάλαιο, γίνεται μια λεπτομερής ιστορική αναδρομή, για τη θεωρία πληροφορίας και για το πως αυτή κατέληξε να θεωρείται μετρήσιμο μέγεθος. Γίνεται ιδιαίτερη αναφορά, στο θεμελιωτή της Claude Shannon, ενώ στο τέλος του κεφαλαίου περιγράφεται συνοπτικά η μέθοδος με την οποία η πληροφορία μεταφέρεται από τον μεταδότη στον παραλήπτη της.

Στο δεύτερο κεφάλαιο, αναλύονται τα σημαντικότερα μέτρα πληροφορίας που υπάρχουν. Ιδιαίτερη έμφαση, δίνεται στο βασικότερο από αυτά, την εντροπία κατά Shannon, πάνω στην οποία βασίζονται και τα υπόλοιπα μέτρα. Επίσης, αναλύονται διάφορες μαθηματικές σχέσεις, μεταξύ των μέτρων αυτών, χρήσιμες για τη συνέχεια της εργασίας, οι οποίες δείχνουν ακριβώς το πόσο αλληλένδετα είναι αυτά τα μέτρα μεταξύ τους.

Το τρίτο κεφάλαιο είναι αποκλειστικά αφιερωμένο πάνω στην επιλογή χαρακτηριστικών. Αναλύονται συγκεκριμένα τα 2 στάδια τα οποία περιλαμβάνει η διαδικασία αυτή, ενώ παράλληλα, παρουσιάζονται οι 3 κατηγορίες (filter, wrapper, embedded), πάνω στις οποίες χωρίζονται οι διάφορες μεθόδοι. Στο τέλος, αναφέρονται επιγραμματικά, κάποιες συγκεκριμένες μεθόδοι επιλογής χαρακτηριστικών, οι οποίες θα μας απασχολήσουν σε πειραματικές συγκρίσεις που γίνονται στο πέμπτο κεφάλαιο.

Στο τέταρτο κεφάλαιο, γίνεται αναφορά στις μηχανές εκμάθησης καθώς και στη διαδικασία της ταξινόμησης δεδομένων. Παράλληλα αναλύονται και σχολιάζονται οι σημαντικότεροι ταξινομητές που υπάρχουν.

Τέλος, στο πέμπτο και βασικότερο κεφάλαιο, αναλύονται λεπτομερώς, 3 από τις σημαντικότερες μεθόδους επιλογής χαρακτηριστικών βασισμένες σε μέτρα πληροφορίας που υπάρχουν. Συγκεκριμένα, περιγράφονται και αναλύονται οι μεθόδοι mMIFS, mMIFS-U και CMIM. Στο τέλος κάθε υποκεφαλαίου, περιγράφονται τα πειράματα που έγιναν, προκειμένου να υπολογίστεί η αποδοτικότητα των μεθόδων αυτών. Τα αποτελέσματα των πειραμάτων αυτών παρουσιάζονται σχηματικά και σχολιάζονται εκτενώς.

ABSTRACT

There is no doubt that one of the biggest problems in dataset classification, concerns the large amount of data the set may have. In such a case, classification is quite often ineffective with substantial errors. To solve this problem, it is quite common to use some very specific feature selection methods, in order to decrease the amount of our data, thus determining a small subset consisting of the most significant data. In this way, the new subset can then be classified without any problem. There is a large variety of feature selection methods to be chosen from. In this paper we focus on feature selection methods based on information measures such as entropy and conditional mutual information.

Chapter 1, includes a detailed review of the history of information theory. It introduces Claude Shannon, the "father" of information theory and discusses his contribution to this field and also the way he manages to measure information. At the end of the chapter there is a short description about the way information is transferred from the transmitter to the receiver.

In Chapter 2, there is an analysis of the most important information measures with the focus on entropy, the most basic of these, and on which all other information measures are based. Furthermore some important mathematical equations which combine these information measures are presented.

Chapter 3 is dedicated to feature selection. More specifically, it presents the 2 stages of this procedure and describes the 3 categories (filter, embedded, wrapper) which feature selection methods are divided into. The end of this chapter briefly introduces some special feature selection methods, which will be use in Chapter 5 in the experiments present therein.

Chapter 4 deals with classification problems. It discusses machine learning and also defines the meaning of classification. Furthermore there is extensive discussion on some of the most important classifiers used.

Finally, in the fifth and most important chapter we present three feature selection methods based on information measures. More specifically, this chapter deals with mRMR, CMIM and mMIFS-U methods. At the end of each section there is a description of some experiments that were carried out in order to compute the efficiency of these methods. The results of these experiments are presented in graphic form and conclusions are given.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΘΕΩΡΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Η Θεωρία πληροφορίας είναι το τμήμα των εφαρμοσμένων μαθηματικών που ασχολείται με την ποσοτικοποίηση της πληροφορίας. Αναπτύχθηκε από τον Claude Shannon στη προσπάθεια του να ανακαλύψει τα θεμελιώδη όρια της επεξεργασίας σήματος σε εφαρμογές όπως η συμπίεση, η αποθήκευση και η μεταφορά δεδομένων. Από τη θεμελίωση της, η θεωρία αυτή έχει διευρυνθεί σε μεγάλο βαθμό, ώστε σήμερα να βρίσκει εφαρμογές σε πολλούς άλλους τομείς, όπως στην επαγωγική στατιστική, 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.

Το βασικό μέτρο της πληροφορίας είναι η εντροπία πληροφοριών και εκφράζεται συνήθως μετρώντας το μέσο αριθμό των bits που χρειάζονται για να αποθηκευθεί ή να μεταβιβαστεί ένα σύμβολο σε ένα μήνυμα. Η εντροπία πληροφοριών ποσοτικοποιεί την αβεβαιότητα που εμπλέκεται στην πρόβλεψη της τιμής μίας τυχαίας μεταβλητής.

Τα πιο θεμελιώδη αποτελέσματα από την θεωρία πληροφορίας, είναι το θεώρημα του Shannon για τον πηγαίο κώδικα, το οποίο καθιερώνει ότι, κατά μέσο όρο, ο αριθμός των δυαδικών ψηφίων (bits) που χρειάζονται για να αναπαρασταθεί το αποτέλεσμα ενός αβέβαιου γεγονότος δίνεται από την εντροπία πληροφοριών του, καθώς και η θεωρία του, για τα θορυβώδη κανάλια μεταφοράς, η οποία αναφέρει ότι η αξιόπιστη επικοινωνία είναι δυνατή σε θορυβώδη κανάλια με την προϋπόθεση ότι ο συντελεστής της επικοινωνίας είναι κάτω από ένα ορισμένο όριο.

Η θεωρία πληροφοριών είναι ένας ευρύς κλάδος, με εξίσου ευρείες και «βαθιές» εφαρμογές, όπως προαναφέρθηκε. Εντούτοις ο τομέας στον οποίο βρίσκει την πιο πλατειά αποδοχή και είναι κάτι περισσότερο από απαραίτητη είναι αυτός της Θεωρίας κωδικοποίησης. Τέλος κάποιοι άλλοι πιο σύγχρονοι τομείς στους οποίους χρησιμοποιείται η Θεωρία της Πληροφορίας είναι πάνω στην Ανάκτηση Πληροφοριών, στην συλλογή πληροφοριών, ακόμα και στην σύνθεση μουσικής.

1.1 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ

Αναμφίβολα η ημερομηνία που αποτελεί ορόσημο για την θεωρία της πληροφορίας, δεν μπορεί να είναι άλλη από την 30 Απριλίου 1916. Ήταν η ημέρα που ήρθε στη ζωή ο Claude Shannon, (30 Απριλίου 1916 – 24 Φεβρουαρίου 2001), ένας σπουδαίος Αμερικανός επιστήμονας, ο οποίος έβαλε τα θεμέλια στην ανάδειξη και την κατανόηση του προβλήματος της πληροφορίας, καθιστώντας την μετρήσιμο μέγεθος. Δεν είναι άλλωστε τυχαίο που ο Σάνον, έμεινε στην ιστορία του

επιστημονικού-και οχι μόνο- κόσμου σαν ο «πατέρας» της Θεωρίας της Πληροφορίας.

Μέχρι και το 1930, οι δημοσιεύεσεις πάνω στη Θεωρία της Πληροφορίας ήταν από ελάχιστες και πολύ περιορισμένες. Δύο ήταν η πιο σημαντικές από αυτές. Η πρώτη ήταν η εργασία που παρουσιάστηκε από τον Harvy Nyquist το 1924 με τίτλο "Certain Factors Affecting Telegraph Speed". Μέσα σε αυτή, περιέγραφε το τρόπο με τον οποίο θα μπορούσαν τα μηνύματα (χαρακτήρες), να σταλούν με ένα τηλεγραφο με τη μέγιστη ταχύτητα, αλλά χωρίς παραμόρφωση.

Η δεύτερη σημαντική μελέτη ήταν αυτή που δημοσιεύτηκε το 1928 από τον Αμερικανό Ralph Hartley ο οποίος μέσα από τη διατριβή του με τίτλο "Transmission of Information" έγινε ο πρώτος που προσπάθησε να ορίσει τη πληροφορία σαν μια μετρήσιμη ποσότητα. Υποθέσε ότι για κάθε σύμβολο ενός μηνύματος υπάρχει η δυνατότητα για n επιλογές. Για παράδειγμα, θεωρώντας ένα μήνυμα k συμβόλων υπάρχουν n^k πιθανά διακεκριμένα μηνύματα. Με βάση αυτή του την υπόθεση, όρισε το ποσό της πληροφορίας, να ισούται με το λογάριθμο του αριθμού των διακεκριμένων μηνυμάτων. Δηλαδή σύμφωνα με το Hartley η πληροφορία είναι ίση με

$$H_{\mathrm{H}}(n^{\mathrm{k}}) = \log(n^{\mathrm{k}}) = \mathrm{k}\log(\mathrm{n}),$$

Προφανώς βάση της πιο πάνω εξίσωσης, για μηνύματα μήκους 1 η πληροφορία θα είναι ίση με

$$\mathrm{H_H}(\mathrm{n}^1) = \log(\mathrm{n})$$

και συνεπώς πρόκυπτει εύκολα η σχέση

$$\mathrm{H_H}(n^{\mathrm{k}}) = \mathrm{k}\,\mathrm{H_H}(\mathrm{n}^1)$$

Η πιο πάνω ισότητα μπορεί επίσης, να ερμηνευτεί διαισθητικά με την υπόθεση, ότι ένα μήνυμα που αποτελείται από k σύμβολα, θα έχει και k-φορές περισσότερη πληροφορία από ένα μήνυμα το οποίο αποτελείται μόνο από ένα σύμβολο. Η μόνη συνάρτηση, η οποία ικανοποιεί την εξίσωση αυτή δηλαδή την

$$\mathrm{g}(n^{\mathrm{k}}) = \mathrm{k}\,\mathrm{g}(\mathrm{n})$$

είναι η

$$\mathrm{g}(\mathrm{n}) = \log(\mathrm{n})$$

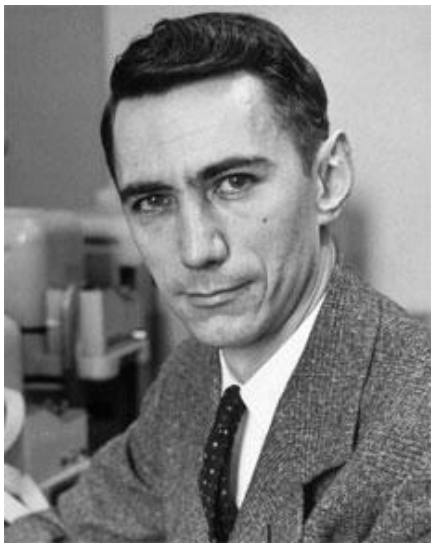
εξ΄ου και η εμφάνιση του λογαρίθμου στον ορισμό που έδωσε ο Hartley.

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 δεν προνόησε καθόλου, για το γεγονός ότι τα n σύμβολα, μπορεί να έχουν άνισες πιθανότητες εμφάνισης, ή ότι μπορεί να υπάρχει μια εξάρτηση μεταξύ των k διαδοχικών συμβόλων.

Τα προβλήματα αυτά, λύθηκαν, όταν εμφανίστηκε στο προσκήνιο ο Claude Shannon , ο οποίος κατάφερε να επεκτείνει τις θεωρίες των Nyquist και Hartley, καταφέρνοντας έτσι να θεμελιώσει την έννοια της πληροφορίας, όπως είναι σήμερα γνωστή.

1.2 Η ΘΕΩΡΙΑ ΤΟΥ SHANNON

Όλα ξεκίνησαν γύρω στο 1936, όταν ο πρόεδρος του τμήματος Μηχανολογίας στο φημισμένο πανεπιστήμιο του MIT, στο οποίο ο Σάνον έκανε το μεταπτυχιακό του, τον όρισε υπεύθυνο για τη λειτουργία μιας δύσχρηστης υπολογιστικής συσκευής, που είχε κατασκευάσει ο ίδιος και είχε ονομάσει «διαφορικό αναλυτή». Προκειμένου να βελτιώσει τη συσκευή αυτή, ο Σάνον, άρχισε να σκέφτεται διάφορους τρόπους αντικατάστασης των δύσχρηστων μηχανικών μερών της με ηλεκτρικά κυκλώματα.



Claude Elwood Shannon (1916-2001)

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 οπόιος 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 χαρακτηρίζεται από πολλούς σαν μία από τις κορυφαίες του 20^{ου} 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 κρυπτο / αποκρυπτογράφηση που απόκτησε κατά τη διάρκεια του 2^{ου} 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.

Αναμφίβολα η πιο σημαντική στιγμή στην καριέρα του, η οποία έμελλε να αλλάξει εξ ολοκληρου το τρόπο με τον οποίο βλέπουμε τα πράγματα εμείς σήμερα, ήταν το 1948 όταν ο Σάνον δημοσίευσε την αξεπέραστη εργασία του, με τίτλο « A Mathematical Theory of Communication». Ήταν ο πρώτος που έκανε μια ολοκληρωτική μαθηματική απόπειρα θεμελίωσης της Θεωρίας της Πληροφορίας. Στις σελίδες αυτής της εργασίας, γίνεται λόγος για πρώτη φορά για μια μονάδα μέτρησης της πληροφορίας, το δυαδικό ψηφίο (binary digit) που συντμήθηκε αργότερα από επιστήμονες του χώρου αρχικά σε binit και στη συνέχεια στο γνωστό μας 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 μετρο 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 ενα πιθανοθεωρητικό πείραμα A, σε ένα δειγματικό χώρο X, να ισούται με

$$H(A) = -\sum_{i=1}^{n} pi \log pi$$

όπου με pi συμβολίζεται η πιθανότητα του ενδεχομένου $xi \in X$

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

Ρ να γίνει ένα γεγονός, τόση περισσότερη ποσότητα πληροφορίας συνοδεύει την πραγματοποίηση του. Και αντίστροφα, αν η πιθανότητα πραγματοποίησης ενός γεγονότος είναι μεγάλη, τότε η πληροφορία που «κουβαλάει» το γεγονός αυτό είναι μικρή.

Για παράδειγμα, αν κάποιος ακούσει πως «Αύριο θα βρέχει στο κέντρο της Αθήνας», το μήνυμα αυτό, έχει μεγάλη πληροφορία, γιατί είναι ένα αβέβαιο γεγονός. Αν όμως ακούσει κάποιος πως «στην Ευρώπη αύριο θα βρέχει», τότε το κείμενο αυτό έχει πολύ μικρή πληροφορία. Γιατί στο μήνυμα αυτό η πιθανότητα να βρέχει κάπου στην Ευρώπη είναι πολύ μεγάλη, ίσως αγγίζει και το 100%.

Γεγονός είναι ότι η θεωρία της πληροφορίας που όπως διατυπώθηκε από τον Claude Shannon ξεκίνησε την ψηφιακή επανάσταση που οδήγησε στην ανάμυξη 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.

Από τη στιγμή που διατύπωσε τα θεωρήματά του, η Φύση δεν μπορούσε πια να ιδωθεί μόνο σαν ύλη και ενέργεια. Μία τρίτη συνιστώσα προστέθηκε στην προσπάθεια εξήγησης του κόσμου: **η πληροφορία**.

1.3 Η ΕΝΝΟΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

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, κάτι καινούργιο, δηλαδή κάποιος γίνεται κοινωνός ενός γεγονότος, βάζοντας έτσι τέλος στη άγνοια του και εξαφανίζει την αβεβαιότητα για κάτι.

Η λέξη αυτή χρησιμοποιείται με 2 έννοιες σήμερα. Η πρώτη, ταυτίζεται με αυτό που ονομάζεται «είδηση», ενώ η δεύτερη έννοια συνδέεται με το «μαθηματικό μέγεθος», το οποίο έρχεται σε αντίθεση με την αμφιβολία που υπάρχει για κάτι. Αυτή η διφορούμενη ερμηνεία, πολλές φορές οδηγεί σε σύγχυση, τόσο που υπάρχουν ολόκληρες μελέτες οι οποίες καλύπτουν την πρώτη ή τη δευτερη έννοια, αγνοώντας επιδεικτικά την άλλη.

Η έννοια της «είδησης», αποτελεί στην ουσία ένα στοιχείο, πάνω στο οποίο ενσωματώνεται μια πληροφορία, για τις πράξεις, τις σκέψεις ή τα γεγονότα που έγιναν. Είδηση, ονομάζεται η περιγραφή, ένός τέτοιου γεγονότος πραγματικού ή φανταστικού, αποδεδειγμένου ή μη. Το τελευταίο στην ουσία εννοεί ένα ψίθυρο 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.

Τρία είναι τα βασικά στοιχεία που πρέπει να διακρίνει κάποιος σε μια είδηση. Καταρχήν, το πρώτο είναι η μορφή που θα έχει αυτή, καθώς μεταδίδεται μέσα από ένα μέσο επικοινωνίας (κανάλι). Δεύτερον, πρέπει να είναι γνωστό, το μέγεθος της αξίας, που έχει, τόσο για τον αποστολέα της, όσο και για τον δέκτη που θα την παραλάβει. Είναι προφανές, ότι καμιά πληροφορία δεν μπορεί να είναι το ίδιο σημαντική και για τους 2 αλλά και ούτε τελείως ασυσχέτιστη. Τέλος, το τρίτο στοιχείο που χαρακτηρίζει μια είδηση, είναι αυτό του «καινούργιου». Μια πληροφορία η οποία διαδίδεται για αρκετό διάστημα, στην πραγματικότητα καταλήγει στο σημείο, να μην προκαλεί «σημαντική» εντύπωση, γιατί δημιουργεί μια τάση εθισμού σε αυτή.

Όσο αφορά τη δευτερη έννοια που έχει η πληροφορία, μια ολόκληρη επιστήμη έχει αναπτυχθεί γύρω από αυτή, η οποία ασχολείται με προβλήματα συλλογής, μετάδοσης, αποθήκευσης, επεξεργασίας και υπολογισμού της. Το κύριότερο πρόβλημα το οποίο έπρεπε να αντιμετωπιστεί από τους επιστήμονες που ανάπτυξαν αυτή τη θεωρία, ήταν το πρόβλημα της ταχύτητας στη μετάδοση της, 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.

Κατά την ανάπτυξη της θεωρίας της πληροφορίας, εγκαταλήφθηκε τελείως η «ανθρώπινη» σημασία της. Για παράδειγμα η φράση «2 καλά αυτοκίνητα», δεν είναι τίποτα άλλο για τους ειδικούς, παρά ένα σύνολο αποτελούμενο από 15 χαρακτήρες για τους οποίους υπάρχουν 2 κενές θέσεις μετά τον $1^{ο}$ και το $5^{ο}$ σύμβολο. Το περιεχόμενο δεν παίζει κανένα απολύτο 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 το ίδιο, σαν 2 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 τη ταχύτητα μετάδοσης του μηνύματος, ανεξαρτήτως από τη σημασία που μπορεί να έχει αυτό.

1.3.1 ΒΑΣΙΚΕΣ ΜΟΡΦΕΣ ΠΛΗΡΟΦΟΡΙΑΣ:

- Ισοδυναμες πληροφορίες: αυτές οι οποίες έχουν την ίδια μορφή ψυχολογικής επίδρασης στο «πομπό»
- Διακρινομενες πληροφορίες: οι μη ισοδύναμες
- Απολυτα νέες πληροφορίες: αυτές οι οποίες διακρίνονται από κάθε γνωστή πληροφορία
- Εμμένουσες πληροφορίες: αυτές οι οποίες μπορούν να αποκομιστούν από μια περιοχή του χώρου, αρκετό καιρό μετά την εκδήλωσή τους(πχ φωτογραφίες, σχέδια, CD)
- Μεταβατικές πληροφορίες: αυτές οι οποίες δεν εκδηλώνονται σε μια περιοχή του χώρου παρά για ένα μόνο σύντομο χρονικό διάστημα(πχ τηλεφωνικές συνδέσεις)
- Ασυνεχείς πληροφορίες: αποτελούνται από ξεχωριστά στοιχεία στο χώρο(πχ τηλεγραφηματα)
- Συνεχείς πληροφορίες: τα στοιχεία από τα οποία αποτελείται είναι απολύτως συνεχή μεταξύ τους τοποθετημένα το ένα δίπλα στο άλλο.

1.4 Η ΔΙΑΔΙΚΑΣΙΑ ΜΕΤΑΦΟΡΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Το κλασσικό παράδειγμα της θεωρίας της πληροφορίας είναι το μηχανικό πρόβλημα της μετάδοσης της πληροφορίας μέσα από έναν θορυβώδες κανάλι. Η πληροφορία για να μεταδοθεί, ξεκινά από ένα **μεταδότη** (πομπό) και μέσω ενός διαύλου επικοινωνίας, καταλήγει στο **παραλήπτη** (δέκτη), όπως φαίνεται και στο πιο κάτω σχήμα.

Σχήμα 1: Διαδικασία μετάδοσης της πληροφορίας

Το θεμελιώδες πρόβλημα, γύρω από αυτή τη μεταφορά, έγκειται στο γεγονός, ότι είναι πιθανόν να εμφανιστούν κατά τη διάρκεια της, σφάλματα ή παραμορφώσεις, οι οποίες να οφείλονται κατά κύριο λόγο στο **θόρυβο (noise)** που μπορεί να υπάρχει στο δίαυλο επικοινωνίας. Η μεταφορά της πληροφορίας, πρέπει να γίνεται χωρίς σφάλματα, τουλάχιστο στο βαθμό που να ικανοποιεί τις απαιτήσεις που έχει ο δέκτης της. Η διακίνηση επομένως, πρέπει να είναι ή χωρίς λάθη, ή να είναι τόσο καλη, έτσι ώστε να μπορούν να πραγματοποιηθούν και ορισμένα σφάλματα, χωρίς ιδιαίτερο κόστος.

Στην πραγματικότητα, κάτι τέτοιο, δηλαδή η μετάδοση της πληροφορίας, χωρίς σφάλματα, δεν είναι εφικτή τις πλείστες φορές. Η μόνη παρέμβαση που μπορεί να γίνει, είναι όσον αφορά την επιλογή του μέσου μετάδοσης το οποίο να ικανοποιεί ορισμένες οριακές συνθήκες που είναι απαραίτητες. Το σύστημα επικοινωνίας, πρέπει να μεταδίδει τη πληροφορία που παράγεται από την πηγή στο προορισμό της με όσο το δυνατόν μεγαλύτερη ακρίβεια.

1.4.1 Η ΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΤΗ ΜΕΤΑΦΟΡΑ ΤΗΣ:

Προκειμένου να μπορεί να μεταδοθεί η πληροφορία, δέχεται μια συγκεκριμένη επεξεργασία. Αρχικά μέσα από τον πομπό, προκειμένου να καταστεί η πληροφορία κατάλληλη για μεταφορά, δια μέσου ενός διαύλου επικοινωνίας, ακολουθείται μια τυποποιημένη διαδικασία, η οποία διαιρείται στα πιο κάτω 4 στάδια.

1. Πρώτα από όλα γίνεται ένας έλεγχος της πληροφορίας που πρόκειται να μεταδοθεί, προκειμένου να διαπιστωθεί πιο μέρος της είναι σχετικό με το πρόβλημα. Το μέρος που δεν χρειάζεται, διαγράφεται. Αυτή η διαδικασία ονομάζεται data reduction.
2. Γίνεται μια κατάλληλη επεξεργασία της πληροφορίας που παραμένει, προκειμένου να αναπαριστάσται με όσο το δυνατόν περισσότερο συμπαγή τρόπο. Η συμπύκνωση αυτή των δεδομένων λέγεται data compression.
3. Στο $3^{ο}$ στάδιο, ασφαλίζεται η παραγόμενη πληροφορία προκειμένου να εμποδίστει μια πιθανή παράνομη χρήση της. Πολλές φορές για καλύτερη ασφάλεια η πληροφορία κρυπτογραφείται.
4. Τέλος προστατεύεται η πληροφορία έναντι πιθανών σφαλμάτων τα οποία μπορεί να εμφανιστούν στο δίαυλο μεταφοράς. Αυτό γίνεται κατορθωτό με την προσθήκη επιπλέον πληροφορίας, η οποία είναι δυνατον να χρησιμοποιηθεί αργότερο προκειμένου να ξανακατασκευάσει την αρχική.

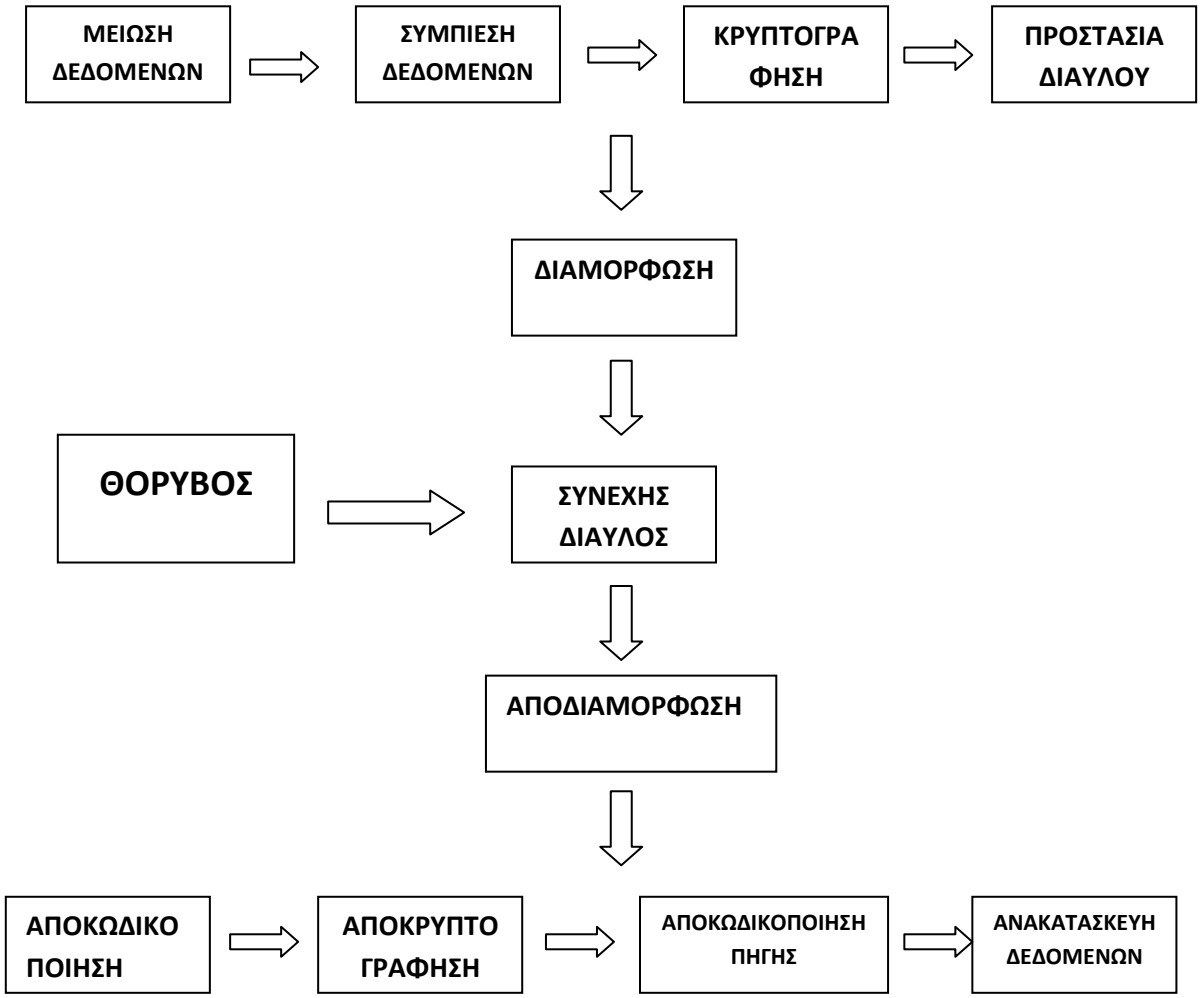
Μέσω αυτών των σταδιων, η πληροφορία, παραδίδεται από τον πομπο στο δίαυλο. Εκεί, τα σύμβολα του μηνύματος διαμορφώνονται σε κατάλληλα σήματα, προκειμένου να καταστεί δυνατή η μετάδοσή τους μέσω του συνεχούς διαύλου. Στο

τέλος της μετάδοσης, πρέπει να υποστούν αποδιαμόρφωση (demodulation), έτσι ώστε να μετατραπούν και πάλι σε σύμβολα.

Λόγω του θορύβου που υπάρχει, είναι δυνατό τα σήματα να υποστούν παραμόρφωση, για αυτό και υπάρχει το ενδεχόμενο ορισμένα σύμβολα να μετατραπούν σε κάποια άλλα λανθασμένα, κατά τη μετάδοση. Μετα την ολοκλήρωση αυτής της διαδικασίας, η πληροφορία ελέγχεται για σφάλματα και όπου είναι δυνατόν, αυτά διορθώνονται, μέσω της χρήσης της αποκωδικοποίησης του διαύλου (channel decoding).

Τέλος, ο δέκτης, προκειμένου να παραδώσει την πληροφορία σωστά στον προορισμό της, την αποκρυπτογραφεί και ακολούθως την αποκωδικοποιεί. Το μήνυμα, μεταδιδεται στην επιθυμητή του μορφή, μέσω της ανακατασκευής των δεδομένων του (data reconstruction).

Η διαδικασία αυτή, παρουσιάζεται αναλυτικά από το παρακάτω σχήμα



Σχήμα 2: Σχηματική περιγραφή της διαδικασίας μεταφοράς της πληροφορίας και όλων των σταδίων που αυτή περιλαμβάνει.

2. ΜΕΤΡΑ ΠΛΗΡΟΦΟΡΙΑΣ

Στη θεωρία της πληροφορίας, όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, η πληροφορία, θεωρείται μετρήσιμο μέγεθος. Για αυτό ακριβώς, αναπτύχθηκαν συγκεκριμένα μέτρα τα οποία ειναι υπέυθυνα για την μετρησή αυτής. Τα κυριότερα από αυτά είναι:

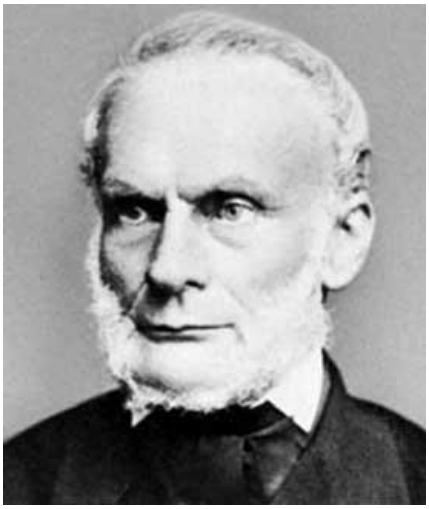
1. Η **εντροπία** (entropy), η οποία μετράει την πληροφορία που φέρει μια τυχαία μεταβλητή X
2. Η **σχετική εντροπία** (relative entropy), η οποία μετράει την ομοιότητα των X και Y.
3. Η **κοινή εντροπία** (joint entropy), η οποία μετράει τη συνολική πληροφορία των X και Y.
4. Η **υπό συνθήκη εντροπία** (conditional entropy), η οποία μετράει την πληροφορία του X, όταν η Y είναι γνωστή και αντιστρόφως.
5. Η **αμοιβαια πληροφορία** (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 2 μεταβλητών X, Y όταν είναι γνωστή μια τρίτη μεταβλητή Z.
7. Η **υπο συνθήκη σχετική εντροπία**(conditional relative entropy)

Πιο κάτω παρουσιάζονται αναλυτικά κάθε μια από αυτές τις έννοιες.

2.1 ΕΝΤΡΟΠΙΑ(ENTROPY)

Μια από τις πιο δύσκολες και αφηρημένες έννοιες που ορίστηκαν ποτέ είναι αυτή της εντροπίας. Μια έννοια, η οποία για τους περισσότερους είναι κάτι αόριστο ή και κάτι το περίεργο, αλλά της οποίας τα αποτελέσματα βιώνουν όλοι οι ανθρώποι καθημερινά.

Η λέξη εντροπία χρησιμοποιήθηκε για πρώτη φορά, 1850 από το Γερμανό φυσικό Rundolf Clausius στο πλαίσιο των μελετών του για τη θερμοδυναμική. Ο Clausius σχημάτισε τον όρο από τις ελληνικές λέξεις "εν" και "τροπή" (με την έννοια της μετατροπής). Δεν ήξερε ότι ήδη υπήρχε τέτοια λέξη στην ελληνική γλώσσα ("ντροπή"), και έτσι έφτιαξε τη λέξη "entropie" κατά αντιστοιχία με τη λέξη "energie" θέλοντας να περιγράψει αυτό που θεωρούσε ότι είναι "εκφυλισμός" της ενέργειας κατά τη διαδικασία της μετατροπής της από τη μία μορφή στην άλλη.



Rudolf Julius Clausius (1822-1888)

Σύμφωνα με τους νόμους της φυσικής, για να παραχθει έργο, πρέπει να μετατραπεί ένα είδος ενέργειας σε ένα άλλο. Σε ένα ενεργειακώς κλειστό σύστημα, το οποίο δεν ανταλλάσσει ενέργεια με το περιβάλλον, η συνολική ποσότητα ενέργειας που περιέχει παραμένει σταθερή. Δεν είναι όμως όλη αυτή η ενέργεια "χρήσιμη". Μπορεί κάποιο κομμάτι της να μπορεί 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. Στην περίπτωση που η εντροπία ενός συστήματος είναι ίση με το μηδέν, τότε όλη η ενέργεια που είναι διαθέσιμη στο 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.

Όταν το 1948 ο Σάνον, μέσω της «θρυλικής», πλέον εργασίας του «A Mathematical Theory of Communication». κατάφερε να ποσοτικοποιήσει την πληροφορία, βρέθηκε μπροστά στο δίλλημα του τι όνομα θα έδινε στο νέο αυτό μέτρο. Σύμφωνα με το μύθο, σε μια συνομιλία που είχε με ένα εξίσου διάσημο μαθηματικό, τον Τζον φον Νόιμαν (Janos Neumann) , ο δεύτερος του είχε πει συγκεκριμένα.

«*Πρέπει να το ονομάσεις εντροπία για δύο λόγους: Πρώτον, η συνάρτηση αυτή χρησιμοποιείται ήδη στη θερμοδυναμική με το ίδιο όνομα. Δεύτερο, και σημαντικότερο, ο περισσότερος κόσμος δεν γνωρίζει τι πραγματικά είναι η εντροπία, και αν χρησιμοποιείς τον όρο εντροπία σε ένα αντεπιχείρημα θα κερδίζεις πάντα*».

Έτσι πήρε το ονομά της, η ποσότητα που μετρούσε τη πληροφορία. Η ονομασία δεν ήταν εντελώς αυθαίρετη, αφού η εντροπία στη θεωρία πληροφοριών, παρουσιάζει όντως πολλές ομοιότητες με την εντροπία όπως αυτή ορίστηκε στη θερμοδυναμική απο τους Gibbs και Boltzmann

2.1.1 ΕΝΤΡΟΠΙΑ ΚΑΤΑ SHANNON

Η εντροπία στη θεωρία πληροφοριών είναι στην ουσία το «μετρο αβεβαιότητας» που διακατέχει το σύστημα. Ο Shannon είδε πως 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 αυτή. Με λίγα λόγια η εντροπία είναι υπεύθυνη για τη μέτρηση του βαθμού αταξίας και αυξάνεται ανάλογα με την πιθανότητα, ενώ η πληροφορία ουσιαστικά μετράει το βαθμό της τάξης μέσα στη δομή του μηνύματος και ελαττώνεται όσο αυξάνεται η πιθανότητα. Αυτός είναι βασικά και ο λόγος που συχνά αναφέρεται η πληροφορία Α σαν η αρνητική εντροπία Η, δηλαδη ισχύει ότι Α= -Η.

Πιο απλά, η πληροφορία ενός συστήματος αποτελεί μέτρο της εσωτερικής του τάξης, αντίστροφο δηλαδή της αταξίας του. Ορίζοντας σαν το μέτρο της αταξίας ενός συστήματος, την εντροπία, η πληροφορία ενός συστήματος είναι αντιστρόφως ανάλογη της.

Ο Shannon μοντελοποίησε την πληροφορία σαν μία σειρά από γεγονότα που συμβαίνουν με συγκεκριμένες πιθανότητες, κάτι που ήρθε σε πλήρη αντίθεση με το πώς ο άνθρωπος αντιλαμβάνεται την πληροφορία στην καθημερινή ζωή. Έπιπλέον ο Shannon, έθεσε τις κάτωθι αξιωματικές συνθήκες, τις οποίες πρέπει να πληρεί κάθε μέτρηση της πληροφορίας:

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, τόσο περισσότερη πληροφορία περιέχει.

2. Η H(P) είναι μία συνεχής συνάρτηση ως προς τη πιθανότητα p.

3. Ισχύει η αρχή της προσθετικότητας, δηλαδή

$$H(p_i q_i, \dots, p_i q_m, \dots, p_n q_1, \dots, p_n q_m) = H(p_i, \dots p_n) + H(q_1, \dots, q_m)$$

   Ο όρος προσθετικότητα εννοείται για δύο δειγματικούς χώρους Χ και Υ, όπου τα ενδεχόμενα του ενός χώρου είναι ανεξάρτητα από του άλλου.

4. Η H(P), γίνεται μέγιστη, όταν όλες οι πιθανότητες είναι ίσες. Αυτό αντιστοιχεί με την κατάσταση στην οποία υπάρχει μέγιστη αβεβαιότητα. Αντίστοιχα, η H(P), γίνεται ελάχιστη, όταν ένα ενδεχόμενο έχει πιθανότητα ίση με 1.

5. Η H(P), είναι συμμετρική, δηλαδή η διάταξη των πιθανοτήτων p1,…,pn δεν έχει επίδραση στη τιμή της H(P).

2.1.2 ΟΡΙΣΜΟΣ ΤΗΣ ΕΝΤΡΟΠΙΑΣ

Έχοντας τα πιο πάνω κατά νου, ορίζουμε την εντροπία μιας διακριτής τυχαίας μεταβλητής Χ, η οποία λαμβάνει πεπερασμένο αριθμό πιθανών τιμών $x_1, x_2, \dots, x_n$ με αντίστοιχες πιθανότητες $p_1 p_2, \dots, p_n$, έτσι ώστε $p_i > 0$ και $\sum_{i=1}^{n} pi = 1$, ως

$$H_b(X) = \sum_{i=1}^{n} pi\ log_b pi$$

Η βάση b του λογάριθμου, συνήθως λαμβάνει τη τιμή 2, και η εντροπία μετριέται σε bits ενώ όταν η βάση b ισούται με e, τότε η εντροπία μετριέται σε nats. Επιπλέον ισχύει η σύμβαση ότι 0log0=0, η οποία προκύπτει από τον ορισμό της συνέχειας καθώς ισχύει ότι $\mathrm{x}\log\mathrm{x} \to 0$ καθώς $x \to 0$.

Γενικά θα πρέπει να τονιστεί ότι η εντροπία, ορίζεται συναρτήσει της κατανομής της τυχαίας μεταβλητής X. Δεν εξαρτάται όμως από τις πραγματικές τιμές που παίρνει η X, αλλά μόνο από τις πιθανότητες που έχουν αυτές.

Ένας σχετικός ορισμός με αυτό της εντροπίας, είναι αυτός της «προσδοκιας» (expectation), η οποία συμβολίζεται με το γράμμα Ε. Η προσδοκία Ε, μετρά τη προσδοκούμενη τιμή της τυχαίας μεταβλητής g(X), όταν η $X \sim p(x)$. Η τιμή της δίνεται από τον τύπο

$$E_P\ \ g(x) = \sum_{x \in X} g(x)p(x)$$

Ιδιαίτερο ενδιαφέρον, παρουσιάζει ο πιο πάνω ορισμός, στην ειδική περίπτωση που ισχύει ότι $g(X) = log\frac{1}{p(x)}$

Έχοντας αυτό κατά νου, η εντροπία μπορεί επιπλέον, να ερμηνευτεί και σαν η προσδοκόμενη τιμή του $log\frac{1}{p(X)}$, όταν η X κατανέμεται σύμφωνα με τη σ.μ.π *p(x)*, δηλαδή σαν,

$$H(X) = E_P log\frac{1}{p(X)}$$

Ο πιο πάνω ορισμός της εντροπίας, είναι πιο κοντά στον ορισμό της εντροπίας στη θερμοδυναμική.

Για την εντροπία, ισχύουν 2 ικανές και αναγκαίες συνθήκες:

Καταρχήν, παίρνει πάντοτε θετικές τιμές, δηλαδή

$$H(X) \geq 0$$

Αυτό προκύπτει άμεσα από το ορισμό της πιθανότητας και του λογαρίθμου αφού ισχύει ότι

$$0 \leq p(x) \leq 1 \quad \kappa\alpha\iota \quad log\frac{1}{p(x)} \geq 0$$

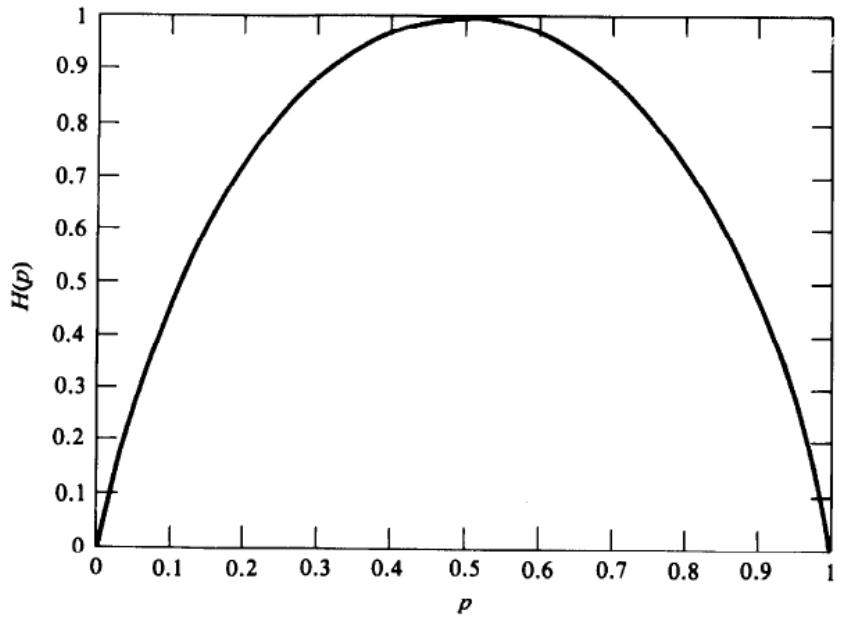
Η δεύτερη συνθήκη είναι ότι

$$H_b(X) = log_b a \, H_a(X)$$

κάτι που προκύπτει άμεσα, μέσα από τη γνωστή σχέση

$$log_b p = log_b a \, log_a p$$

Η δεύτερη συνθήκη, στην ουσία υπονοεί ότι εαν κάποιος θελήσει να αλλάξει τη βάση του αλγόριθμου για την εντροπία, αυτό μπορεί να γίνει κατορθωτό, απλά πολλαπλασιάζοντάς τη σχέση με μια κατάλληλη μεταβλητή.



Σχήμα 2.1: Γραφική αναπαράσταση της σχέσης μεταξύ της εντροπίας H() και της πιθανότητας p. Όταν η πιθανότητα να συμβεί ένα γεγονός είναι μηδενική ή αντίστοιχα ίση με 1, τότε η εντροπία ισούται με μηδέν. Αντίστοιχα, η εντροπία για ένα γεγονός παίρνει τη μέγιστη τιμή της, όταν η πιθανότητα για ένα γεγονός ισούται ακριβώς με ½.

Ένα ενδεικτικό παράδειγμα, προκειμένου να κατανοηθεί καλύτερα ο τρόπος υπολογισμού της εντροπίας, είναι το ακόλουθο. Έστω η διακριτή μεταβλητή *X* που παίρνει τις κάτωθι τιμές

$$X=\begin{cases} a \ \ με\ πιθανότητα\ \frac{1}{2} \\ b\ με\ πιθανότητα \frac{1}{4} \\ c\ με\ πιθανότητα \frac{1}{8} \\ d\ με\ πιθανότητα \frac{1}{8} \end{cases}$$

Τότε η εντροπία H(X) της X θα είναι ίση με

$$\mathrm{H(X)} = -½ \log ½ - ¼ \log ¼ - 1/8 \log 1/8 - 1/8 \log 1/8 = 7/4 \text{ bits}$$

2.2 ΣΧΕΤΙΚΗ ΕΝΤΡΟΠΙΑ (RELATIVE ENTROPY)

Η σχετική εντροπία ή αλλίως απόκλιση κατά Kullback–Leibler, είναι ένα μέτρο που που υπολογίζει την διαφόρα μεταξύ 2 κατανομών, συγκρίνοντας μια «πραγματική» κατανομή έστω την p(x) και μια «αυθαιρετη», την q(x). Πιο απλά, είναι ένα μέτρο που υπολογίζει την απόκλιση από την πραγματική τιμή, όταν για παράδειγμα ειπωθεί ότι η κατανομή είναι q ενώ στην πραγματικότητα είναι p.

Συγκεκριμένα, η σχετική εντροπία $D(p|q)$, μεταξύ 2 συναρτήσεων μαζας πιθανότητας p(x) και q(x) ισούται με

$$\mathrm{D(p||q)} = -\sum_{\mathrm{x}\in\mathrm{X}} \mathrm{p(x)} \log \mathrm{q(x)} + \mathrm{H(X)}$$

$$= \sum_{\mathrm{x}\in\mathrm{X}} \mathrm{p(x)} \log \frac{p(\mathrm{X})}{\mathrm{q(X)}}$$

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

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. Μπορεί κάποιος να την αντιληφθεί καλύτερα σαν μια «απόσταση» μεταξύ 2 κατανομών.

Η σχετική εντροπία είναι μια έννοια πολυ 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.

2.3 ΚΟΙΝΗ ΕΝΤΡΟΠΙΑ (JOINT ENTROPY)

Η κοινή εντορπία H(X,Y), μεταξύ 2 τυχαίων μεταβλητών X,Y, δηλαδή η εντροπία της σύνθετης κατανομής τους p(x, y) ορίζεται σαν

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y) log_2[p(x,y)]$$

Όταν οι κατανομές X, Y είναι ανεξάρτητες μεταξύ τους, τότε ισχύει ότι

$$\mathrm{H(X,Y)} \;=\; \mathrm{H(X)} + \;\mathrm{H(Y)}$$

ενώ όταν οι μεταβλητές είναι πάνω από 2, δηλ είναι $X_1,\ldots,X_n$, η κοινή εντροπία ισούται με

$$H(X_1,\ldots,X_n) = -\sum_{x_1}\ldots\sum_{x_n} p(x_1,\ldots,x_n)\; log_2[p(x_1,\ldots,x_n)]$$

2.4 ΥΠΟ ΣΥΝΘΗΚΗ ΕΝΤΡΟΠΙΑ (CONDITIONAL ENTROPY)

Η υπό συνθήκη εντροπία H(Y|X), καθορίζει το ποσό της πληροφορίας, που χρειάζεται προκειμένου να υπολογιστεί μια τυχαία μεταβλητή Y, όταν είναι γνωστή η τιμή μιας άλλης μεταβλητής, της X, θεωρώντας, πάντα ότι $(X,Y)\sim p(x,y)$. Με άλλα λόγια η υπο συνθήκη εντροπία μετρά την αβεβαιότητα της Y όταν είναι γνωστή η X μεταβλητη. Πιο συγκεκριμένα η H(Y|X) ισούται με

$$H(Y|X) \;\equiv \sum_{x\in\mathcal{X}} p(x)\, H(Y|X=x)$$
$$= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\, \log\, p(y|x)$$

$$
\begin{aligned}
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ \log\ p(y|x)\\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\, p(y|x)\\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)}.\\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x)}{p(x,y)}.
\end{aligned}
$$

Προφανώς, στην περίπτωση που η τυχαία μεταβλητή Y, καθορίζεται τελείως μέσω της X, ισχύει ότι $H(Y|X)=0$, ενώ όταν οι δυο μεταβλητές, X, Y είναι τελείως ανεξάρτητες μεταξύ τους, ισχύει ότι H(Y|X)=H(Y).

Όσον αφορά την φυσικότητα με την οποία ορίζονται τόσο η κοινή όσο και η υπο συνθήκη εντροπία, αυτή διαφαίνεται από το γεγονός ότι η εντροπία ενός ζεύγους τυχαίων μεταβλητών (X,Y), ισούται με το άθροισμα της εντροπίας της πρώτης μεταβλητής με την υπό συνθήκη εντροπία της δεύτερης μεταβλητής.

Πράγματι μέσω του κανόνα της αλύσίδας αποδεικνύεται ότι

$$\mathrm{H(X,Y)} \ = -\sum_{\mathrm{x\in X}}\sum_{\mathrm{y\in Y}} p(x,y)\ \log \mathrm{p(x,y)}$$

$$= -\sum_{\mathrm{x\in X}}\sum_{\mathrm{y\in Y}} p(x,y)\ log\ \mathrm{p(x)p(y|x)}$$

$$= -\sum_{\mathrm{x\in X}}\sum_{\mathrm{y\in Y}} p(x,y)\ log\ \mathrm{p(x)} - \sum_{\mathrm{x\in X}}\sum_{\mathrm{y\in Y}} p(x,y)\ log\ \mathrm{p(y|x)}$$

$$= -\sum_{\mathrm{x\in X}} \mathrm{p(x)}\log \mathrm{p(x)}\ - \sum_{\mathrm{x\in X}}\sum_{\mathrm{y\in Y}} p(x,y)\ log\ \mathrm{p(y|x)}$$

$$= H(X) + \ H(Y|X)$$

Ισοδύναμα ισχύει ότι

$$log\ p(X,Y)\ =\ log\ p(X)\ +\ log\ p(\ Y|X)$$

2.5 ΑΜΟΙΒΑΙΑ ΠΛΗΡΟΦΟΡΙΑ (MUTUAL INFORMATION)

Η αμοιβαία πληροφορία I(X;Y) είναι το μέτρο που μετρά την πληροφορία που μοιράζονται 2 τυχαίες μεταβλητές μεταξύ τους, υπολογίζει δηλαδή το πόσο μπορεί

η γνώση για την δεύτερη μεταβλητή να μειώσει την αβεβαιότητα για την πρώτη. Το μέτρο αυτό ουσιαστικά βασίζεται στην αμοιβαία εξάρτηση που υπάρχει μεταξύ των 2 μεταβλητών. Πιο συγκεκριμένα, έστω ότι δίνονται 2 τυχαίες μεταβλητές καθώς και η από κοινου σ.μ.π τους p(x, y), όπως και οι περιθώριες κατανομές πιθανότητας των X, Y p(x) και p(y) αντίστοιχα. Τότε η αμοιβαία πληροφορία θα ισούται με τη σχετική εντροπία μεταξύ της από κοινου σ.μ.π p(x,y) και των p(x),p(y), δηλαδή

$$I(\mathrm{X};\mathrm{Y}) = \sum_{x\in X}\sum_{y\in Y} \mathrm{p}(\mathrm{x},\mathrm{y}) \log \frac{\mathrm{p}(\mathrm{x},\mathrm{y})}{\mathrm{p}(\mathrm{x})\mathrm{p}(\mathrm{y})}$$

$$= D(p(x,y)||p(x)p(y))$$

$$= E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}$$

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

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

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

Στην περίπτωση τώρα, που η X, είναι μια ντερερμινιστική συνάρτηση της Y(ή το αντίθετο), 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 το Y(και το αντίστροφο). Σε αυτή την περίπτωση, η αμοιβαία πληροφορία, θα είναι ίση με την αβεβαιότητα που περιέχει μόνη της η Y(ή η X αντίστοιχα). Δηλαδή θα είναι ίση με την εντροπία της Y.

Η αμοιβαία πληροφορία, είναι ένα μέτρο εξάρτησης που βασίζεται στην από κοινού κατανομή των X, Y, δηλαδή στηρίζεται στο γεγονός ότι οι 2 μεταβλητές θα έχουν έστω κάποια συσχέτιση μεταξύ τους.

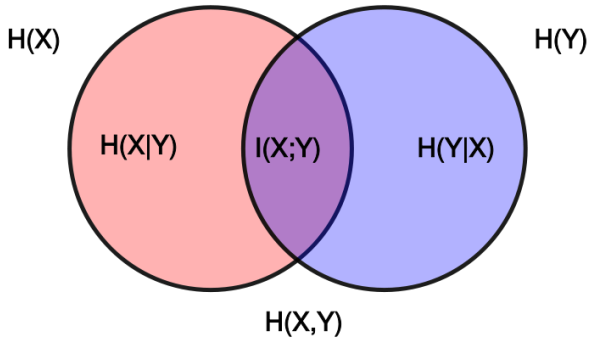
Όταν οι τυχαίες μεταβλητές X, Y είναι τελείως ανεξάρτητες, τότε I(X;Y)=0. Αυτό προκύπτει από το γεγονός ότι εαν, X,Y ανεξάρτητες, τότε

$$\mathrm{p}(\mathrm{x},\mathrm{y}) = \mathrm{p}(\mathrm{x})\ \mathrm{p}(\mathrm{y}) \qquad \text{και άρα}$$

$$\log\frac{\mathrm{p(x,y)}}{p(x)p(y)} = log1 = 0$$

Τέλος 2 αναγκαίες και ικανές συνθήκες για την αμοιβαία πληροφορία είναι οι εξής:

- Είναι πάντοτε θετική δηλαδη I(X;Y)≥ 0
- Είναι συμμετρική δηλαδή ισχύει πάντα ότι I(X;Y)=I(Y;X)



Σχήμα 2.2: Σχηματική απεικόνιση της σχέσεων μεταξύ της αμοιβαίας πληροφορίας I(X;Y), των εντροπιων H(X) H(Y) και της κοινής εντροπίας H(X,Y) και των υπο συνθήκη εντροπιών H(X\|Y) και H(Y\|X) για 2 συσχετισμένα σύνολα X,Y.

2.6 ΥΠΟ ΣΥΝΘΗΚΗ ΑΜΟΙΒΑΙΑ ΠΛΗΡΟΦΟΡΙΑ (CONDITIONAL MUTUAL INFORMATION)

Η υπό συνθήκη αμοιβαία πληροφορία, είναι ένα από τα πιο βασικά μέτρα στην θεωρία πληροφοριών. Στην ουσία μετρά την αναμενόμενη τιμή 2 τυχαίων μεταβλητών X,Y όταν είναι γνωστή η μια 3^{η} μεταβλητη Z.

Γενικά ορίζεται σαν

$$I(X;Y|Z) = \mathbb{E}_Z\big(I(X;Y)|Z\big) = \sum_{z\in Z} p_Z(z) \sum_{y\in Y}\sum_{x\in X} p_{X,Y|Z}(x,y|z)\log\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)},$$

$$I(X;Y|Z) = \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} p_{X,Y,Z}(x,y,z)\log\frac{p_Z(z)p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)}.$$

$$I(X;Y|Z) = H(X,Z)+H(Y,Z)-H(X,Y,Z)-H(Z) = H(X|Z)-H(X|Y,Z)$$

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.

Δηλαδή θα μπορούσε να ισχύει

$$I(X1, X2; Y1, Y2|Z1; Z2)$$

2.7 ΥΠΟ ΣΥΝΘΗΚΗ ΣΧΕΤΙΚΗ ΕΝΤΡΟΠΙΑ (CONDITIONAL RELATIVE ENTROPY)

Η υπό συνθήκη σχετική εντροπία, ορίζεται σαν το γινόμενο όλων των σχετικών εντροπιών, μεταξύ των υπο συνθήκη συναρτήσεων μάζας πιθανότητας $p(y|x)$, $q(y|x)$, με την συνάρτηση μάζας πιθανότητας p(x). πιο συγκεκριμένα ορίζεται σαν

$$D\big(p(y|x)||q(y|x)\big) = \sum_x p(x) \sum_y p(y|x) log \frac{p(y|x)}{q(y|x)}$$

2.8 ΣΧΕΣΕΙΣ ΜΕΤΑΞΥ ΤΩΝ ΔΙΑΦΟΡΩΝ ΜΕΤΡΩΝ ΠΛΗΡΟΦΟΡΙΑΣ

ΣΧΕΣΗ ΜΕΤΑΞΥ ΕΝΤΡΟΠΙΑΣ ΚΑΙ ΑΜΟΙΒΑΙΑΣ ΠΛΗΡΟΦΟΡΙΑΣ:

Η αμοιβαία πληροφορία I(X;Y) μπορεί επιπλέον να γραφεί και σαν

$$I(X;Y) = \sum_{x,y} \left(\mathrm{p(x,y)} \ \log \frac{p(x,y)}{p(x)p(y)} \right)$$

$$I(X;Y) = \sum_{x,y} \left(\mathrm{p(x,y)} \ \log \frac{p(x|y)}{p(x)} \right)$$

$$= \sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)$$

$$= \sum_{x,y} p(x,y) \log p(x) - (- \sum_{x,y} p(x,y) \log p(x|y))$$

$$I(X;Y) = H(X) - H(X|Y)$$

Δηλαδή η I(X;Y) μετρά την αβεβαιότητας του X, όταν είναι γνωστό το υ

Λόγω της συμμετρικότητας της αμοιβαίας πληροφορίας I(X;Y), η πιο πάνω σχέση μπορεί να γραφτεί και σαν

$$I(X;Y) = H(Y) - H(Y|X)$$

όση δηλαδή πληροφορία δίνει το X για το Y, τόση δίνει και το Y για το X

Όπως δείχθηκε από τον κανόνα της αλυσίδας ισχύει ότι

$$H(X,Y) = H(X) + H(Y|X)$$

και από αυτό προκύπτει η σχέση

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Τέλος, η αμοιβαία πληροφορία μιας τυχαίας μεταβλητής μαζί με τον εαυτό της, θα ισούται με την εντροπία της. Δηλαδή,

$$I(X;X) = H(X) - H(X|X) = H(X)$$

<u>ΑΜΟΙΒΑΙΑ ΠΛΗΡΟΦΟΡΙΑ ΚΑΙ ΕΝΤΡΟΠΙΑ:</u>

Από τα πιο πάνω, εύκολα μπορούν να εξαχθούν οι παρακάτω σχέσεις για την αμοιβαία πληροφορία και την εντροπία.

$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y) = H(Y) - H(Y|X)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I(X;Y) = I(Y;X)$$

ΚΑΝΟΝΑΣ ΑΛΥΣΙΔΑΣ ΓΙΑ ΤΗΝ ΕΝΤΡΟΠΙΑ:

$$H(X_1, X_2) = H(X_1) + H(X_2)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1)$$

$$= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1)$$

…….

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \dots, X_1)$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1)$$

Εναλλακτική απόδειξη θεωρώντας πως $p(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \dots, x_1)$

τότε

$$H(X_1, X_2, \dots, X_n) = -\sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n)\, log p(x_1, x_2, \dots, x_n)$$

$$= -\sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n)\, log \prod_{i=1}^{n} p(x_i|x_{i-1}, \dots, x_1)$$

$$= -\sum_{x_1, x_2, \dots, x_n} \sum_{i=1}^{n} p(x_1, x_2, \dots, x_n) \log p(x_i|x_{i-1}, \dots, x_1)$$

$$= -\sum_{i=1}^{n} \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_i|x_{i-1}, \dots, x_1)$$

$$\sum_{i=1}^{n} \sum_{x_1, x_2, \dots, x_i} p(x_1, x_2, \dots, x_n) \log p(x_i|x_{i-1}, \dots, x_1)$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1)$$

ΚΑΝΟΝΑΣ ΑΛΥΣΙΔΑΣ ΓΙΑ ΤΗΝ ΑΜΟΙΒΑΙΑ ΠΛΗΡΟΦΟΡΙΑ:

$$I(X_1, X_2, \dots, X_n; Y) = H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n|Y)$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1, Y)$$

$$= \sum I(X_i; Y|X_1, X_2, \dots, X_{i-1})$$

ΚΑΝΟΝΑΣ ΑΛΥΣΙΔΑΣ ΓΙΑ ΤΗ ΣΧΕΤΙΚΗ ΕΝΤΡΟΠΙΑ

$$D\big(p(y|x)||q(y|x)\big) = \sum_x \sum_y p(x,y)\ log\frac{p(x,y)}{q(x,y)}$$

$$= \sum_x \sum_y p(x,y)\ log\frac{p(x)p(y|x)}{q(x)q(y|x)}$$

$$= \sum_x \sum_y p(x,y)\ log\frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y)\ log\frac{p(y|x)}{q(y|x)}$$

$$= D(p(x)||q(x)) + D(p(y|x||q(y|x))$$

3. ΕΠΙΛΟΓΕΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Ειναι γεγονός, ότι κατά την διαρκεια των τελευταιων 100 χρόνων συντελεστηκαν στον τομεα της επιστημης και της τεχνολογίας, κοσμοιστορικές αλλαγές. Τεχνολογικά προηγμένες μεθόδοι, μας δίνουν σήμερα, τη δυνατότητα να καταγράφουμε δεδομένα με ολοένα αυξανόμενο ρυθμό. Πλέον η πρόσβαση και η απόκτηση της πληροφορίας, εχει καταστεί πιο εύκολη από ποτέ. Για παράδειγμα ενω το 1997, ο αριθμός των χαρακτηριστικών που βρισκόταν στη διάθεση ενός σχεδιαστή, για τη δημιουργία ενός συστήματος ταξινόμησης, δεν ξεπερνούσε τα 40, πλέον η κατάσταση έχει αλλάξει δραματικά, με τους αριθμούς των χαρακτηριστικών στις μέρες μας, να φθάνουν σε μερικές εκατοντάδες χιλιάδες. Επιπλέον ένα μεγάλο μέρος των δεδομένων μπορεί να αφορά άχρηστες πληροφορίες, πληροφορίες που δεν παρουσιάζουν ενδιαφέρον σε σχέση με το πρόβλημα που μελετάται. Ο τεράστιος όγκος δεδομένων που παράγεται επηρεάζει αρνητικά την ικανότητα του ερευνητή να τα μελετήσει, να εξάγει συμπεράσματα και να λύσει διάφορα προβλήματα.

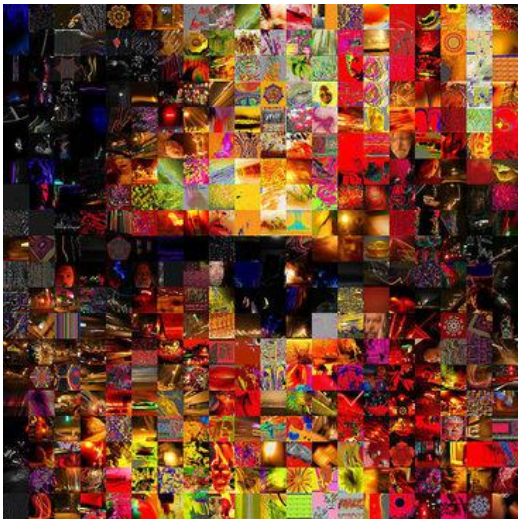
Προκειμένου να αντιληφθεί κάποιος καλύτερα την πιο πάνω κατάσταση, παρουσιάζεται το ακόλουθο πρόβλημα. Έστω ότι ένας ασθενής, χρειάζεται να κάνει κάποιες συγκεκριμένες εξετάσεις, προκειμένου να διαπιστωθεί η πιθανότητα που υπάρχει να εμφανίσει κάποιο είδος καρδιοπάθεια. Στην προκειμένη περίπτωση ο γιατρός που θα κάνει τη διάγνωση, θα ανατρέξει σε μια βάση δεδομένων για το συγκεκριμένο άτομο, έτσι ώστε να μπορέσει να εξάγει τα συμπεράσματά του. Τον ενδιαφέρει να μελετήσει, χαρακτηριστικά του ασθενούς, όπως το ύψος, το βάρος, η αρτηριακή πίεση, τα επίπεδα χοληστερίνης κ.α. Παρόλα αυτά, στη βάση δεδομένων, μπορεί να υπάρχουν και άλλα χαρακτηριστικά του ασθενή, όπως ο τόπος καταγωγής, η οικογενειακή κατάσταση και το επάγγελμά του, χαρακτηριστικά τα οποία και προφανώς, είναι άχρηστα χωρίς καμία χρησιμότητα, σε σχέση με το πρόβλημα που μελετάμε. Ο γιατρός, στο τέλος θα τα αγνοήσει φυσικά, αφού όμως πρώτα ξοδέψει κάποιο συγκεκριμένο χρόνο, προκειμένου να τα απομονώσει μόνο τα στοιχεία που χρειάζεται.

Λαμβάνοντας υπόψη τα πιο πάνω προβλήματα και προκειμένου αυτά να αντιμετωπιστούν αποτελεσματικά, τα τελευταία χρόνια έχουν αναπτυχθεί διάφορες μεθόδοι μείωσης του όγκου των χαρακτηριστικών. Αυτές οι μεθόδοι χωρίζονται σε 2 μεγάλες κατηγορίες, στην εξαγωγή χαρακτηριστικών (feature extraction) μέσω της οποίας τα υπάρχοντα δεδομένα μετασχηματίζονται σε ένα χώρο μικρότερων διαστάσεων και στην επιλογή χαρακτηριστικών (feature selection). Σε αυτό το κεφάλαιο επικεντρωνόμστε στην δεύτερη κατηγορία μεθόδων.

Η επιλογή χαρακτηριστικών, επίσης γνωστή και ως επιλογή μεταβλητών, είναι η διαδικασία, κατά την οποία, δοσμένων ενός συνόλου D που αποτελείται από M

χαρακτηριστικά D={Xi, i=1,…,M} και μιας μεταβλητής C, στόχος είναι να επιλεγεί απο το διάστημα D, ένα υποσύνολο που να αποτελείται από m χαρακτηριστικά, όπου m<<M, το οποίο θα χαρακτηρίζει με ένα βέλτιστο τρόπο τη μεταβλητή C, χωρίς να γίνει ο οποιοσδήποτε μετασχηματισμός στα δεδομένα. Η βασική ιδέα της επιλογής χαρακτηριστικών, είναι ότι για να σχεδιάστεί ένα οποιοδήποτε μοντέλο, είναι καλύτερο να απομονωθεί ένα υποσύνολο χαρακτηριστικών, στο οποίο θα δοθεί έμφαση, αντί να χρησιμοποιηθούν όλα τα διαθέσιμα χαρακτηριστικά, πολλά από τα οποία δεν παρέχουν χρήσιμες πληροφορίες, ή παρέχουν πληροφορίες οι οποιες ήδη υπάρχουν.

Προφανώς και η αξιολόγηση ενός χαρακτηριστικού ως λίγο ή πολύ χρήσιμου δεν είναι εύκολη διαδικασία και αποτελεί το κύριο αντικείμενο μελέτης στο πρόβλημα της επιλογής χαρακτηριστικών. Η δυσκολία αυτής της επιλογής οφείλεται κυρίως σε δύο παράγοντες.



Σχηματική αναπαράσταση ενός συνόλου χαρακτηριστικών, όπου η μείωση του πλήθους τους φαντάζει αναγκαία

Πρώτον, το πλήθος υποσυνόλων που θα μπορούσαν να επιλεγούν αυξάνεται εκθετικά σε σχέση με τον αριθμό των χαρακτηριστικών του αρχικού συνόλου. Ακόμη και αν υποθέσει κάποιος 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 παρά μόνο στην πράξη μέσω της χρήσης ευρετικών τεχνικών επιλέγεται τελικά ένα υποσύνολο που αναμένεται ότι θα οδηγήσει σε καλή απόδοση τον αλγόριθμο μάθησης που θα το χρησιμοποιήσει.

Τα οφέλη από τη χρήση της επιλογής χαρακτηριστικών, είναι πολλά και σημαντικά. Χάρη σε αυτή, μειώνεται το μέγεθος του υπο κατασκευή μοντέλου με αποτέλεσμα να είναι πιο εύκολη η κατανόησή του. Προφανώς και κανένας άνθρωπος δεν μπορεί να ερμηνεύσει μια απεικόνιση χιλιάδων μεταβλητών που παράχθηκε από έναν αλγόριθμο μάθησης με επίβλεψη, ίσως όμως μπορεί να ερμηνεύσει μία απεικόνιση π.χ. είκοσι μεταβλητών. Ακόμη, λόγω του ότι υπάρχουν λιγότερα δεδομένα, επιτυγχάνεται καλύτερη απόδοση του μοντέλου και επιπλέον ελάττωση του χρόνου

κατασκευής του. Τέλος, χάρη στην επιλογή χαρακτηριστικών, αποφεύγεται η υπερπροσαρμογή των δεδομένων με αποτέλεσμα να είναι πιο εύκολη η γενίκευση των αποτελεσμάτων.

Η επιλογή χαρακτηριστικών, μπορεί να χρησιμοποιηθεί σε πάρα πολλούς τομείς, ακόμη και της καθημερινότητας, προκειμένου να ταξινομηθούν διάφορες κατηγορίες προιόντων και αντικειμένων. Επιπλέον οι μεθόδοι επιλογής χαρακτηριστικών, συχνά χρησιμοποιούνται και σε τομείς, όπου υπάρχουν πολλά χαρακτηριστικά και συγκρτικά λίγα δείγματα. Όταν για παράδειγμα στόχος είναι ο διαχωρισμός των υγειών ασθενών, από αυτούς με καρκίνο, πρόβλημα όπου οι παρατηρήσεις (ασθενείς) οι οποίες είναι διαθέσιμες για δοκιμή, δεν υπερβαίνουν τις 100, αντίθετα με τις μεταβλητές (χαρακτηριστικά) ο αριθμός των οποίων κυμαίνεται από 6000-60000, η επιλογή χαρακτηριστικών καθίσταται αναγκαία.

3.1 ΔΙΑΔΙΚΑΣΙΑ ΔΗΜΙΟΥΡΓΙΑΣ ΒΕΛΤΙΣΤΟΥ ΥΠΟΣΥΝΟΛΟΥ

Η διαδικασία επιλογής, ενός υποσυνόλου χαρακτηριστικών, προκειμένου να μειωθεί ο αριμός των χαρακτηριστικών και να επιτευχθεί το βέλτιστο αποτέλεσμα περιλαμβάνει 2 φάσεις:

- Αρχικά, ελλατώνεται ο αριθμός των χαρακτηριστικών, με την απόρριψη εκείνων που φέρουν τη λιγότερη πληροφορία, χρησιμοποιώντας τις βαθμωτές τεχνικές επιλογής χαρακτηριστικών.
- Ακολούθως, εξετάζονται τα χαρακτηριστικά που έμειναν, σε διάφορους συνδυασμούς και επιλέγεται ο καλύτερος προκειμένου να καταλήξουμε στο βέλτιστο δυνατό υποσύνολο χαρακτηριστικών.

3.1.1 ΒΑΘΜΩΤΕΣ ΤΕΧΝΙΚΕΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Κατά τη βαθμωτή επιλογή χαρακτηριστικών, αφού πρώτα επιλεγεί το κριτήριο που θα εφαρμοστεί, ακολούθως τα χαρακτηριστικά ιεραρχούνται σε φθίνουσα σειρά και υπολογίζεται η ετεροσυσχέτιση του πρώτου στη σειρά με τα υπόλοιπα. Θα πρέπει να τονιστεί ότι οι τεχνικές που αναλύονται παρακάτω, δεν λαμβάνουν υπόψη τους τις συσχετίσεις μεταξύ των χαρακτηριστικών και επιπλέον δεν αξιοποιούν τον συντελεστή ετεροσυσχέτισης. Τρείς είναι οι κυριότερες βαθμωτές τεχνικές επιλογής χαρακτηριστικών, οι οποίες και παρουσιάζονται παρακάτω.

(Α) ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ t-test:

Πρόκειται για την πιο δημοφιλή μέθοδο που ακολουθείται, ιδιαίτερα όταν τα χαρακτηριστικά ακολουθούν κανονική κατανομή. Η βασική ιδέα της μεθόδου t-test,

είναι να ελεγθεί εαν η μέση τιμή του χαρακτηριστικού, διαφέρει από κλάση σε κλάση σε σημαντικό βαθμό.

Στόχος δηλαδή είναι να γίνει έλεγχος της πιο κάτω υπόθεσης

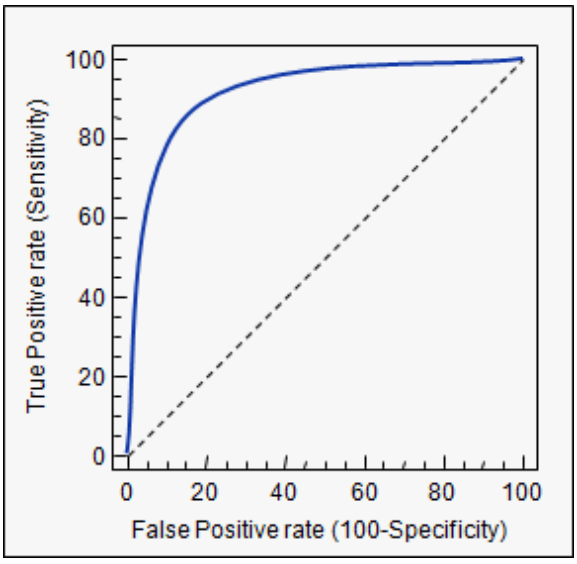
H_0: τα χαρακτηριστικό έχει την ίδια μέση τιμή σε κάθε κλάση

H_1: το χαρακτηριστικό έχει διαφορετική μέση τιμή σε κάθε κλάση

Στην περίπτωση που ισχύει η μηδενική υπόθεση, το χαρακτηριστικό απορρίπτεται λόγω του ότι είναι δύσκολο να χωριστούν τα δεδομένα του σε κλάσεις. Αντίθετα σε περίπτωση που ισχύει η εναλλακτική, επιλέγεται το χαρακτηριστικό, αφού θα είναι εύκολη η διάκριση του στις διάφορες κλάσεις.

(Β) Η ΚΑΜΠΥΛΗ ΛΕΙΤΟΥΡΓΙΚΟΥ ΧΑΡΑΚΤΗΡΙΚΟΥ ΔΕΚΤΗ (RECEIVER OPERATING CHARACTERISTIC-ROC)

Η ROC, χρησιμοποιείται σε περίπτωση που στον έλεγχο t-test, οι μέσες τιμές των χαρακτηριστικών, βρίσκονται κοντά μεταξύ τους, με αποτέλεσμα η πληροφορία αυτή να μην είναι επαρκής προκειμένου να καθοριστεί εαν η ταξινόμηση θα γίνει καλά ή όχι. Μέσω της ROC, μπορούν να εξαχθούν πληροφορίες σχετικά με την επικάλυψη που υπάρχει μεταξύ των διαφόρων κλάσεων, αφού η μέθοδος αυτή, μπορεί να ποσοτικοποιήσει μια περιοχή μεταξύ 2 καμπυλών η οποία ονομάζεται εμβαδόν κάτω από την ROC καμπύλη (Area Under the receiver operating Curve - AUC).



Σχήμα 3.1: Μια ενδεικτική ROC καμπύλη

(Γ) ΛΟΓΟΣ ΔΙΑΚΡΙΣΗΣ ΤΟΥ FISHER:

Ένα ικανοποιητικό κριτήριο, που χρησιμοποιείται για τν ποσοτικοποίηση της διακριτής ικανότητας ενός χαρακτηριστικού. Ο λόγος Fisher, είναι ανεξάρτητος της κατανομής που ακολουθεί η κάθε κλάση και ορίζεται σαν

$$FDR = \sum_{i} \sum_{i \neq j} \frac{(\mu_i - \mu_j)^2}{{\sigma_i}^2 + {\sigma_j}^2}$$

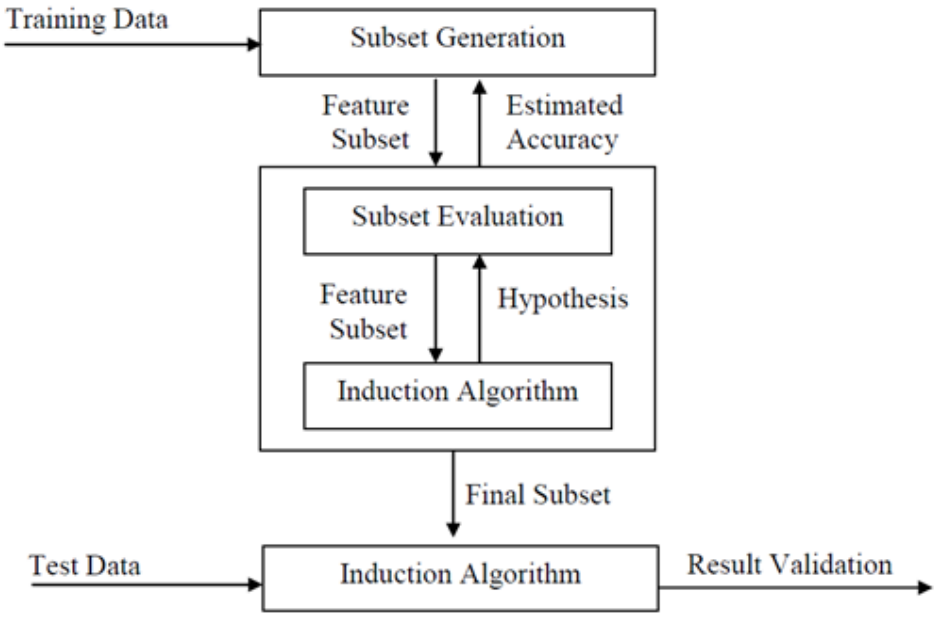
3.1.2 ΔΙΑΔΙΚΑΣΙΑ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Ένας αλγόριθμος επιλογής χαρακτηριστικών, μπορεί να θεωρηθεί, σαν μια εξειδικευμένη μηχανή αναζήτησης, η οποία προτείνει βέλτιστα υποσύνολα. Θεωρητικά μιλώντας, η απλούστερη μέθοδος που μπορεί να θεωρήσει κάποιος, είναι να δοκιμάσει κάθε δυνατό υποσύνολο χαρακτηριστικών, προκειμένου να καταλήξει στο βέλτιστο, αυτό δηλαδή με το μικρότερο σφάλμα. Αυτό όμως στην πράξη είναι απίστευτα εξαντλητικό και χρονοβόρο, χωρίς να δίνει πάντα τα επιθυμητά αποτελέσματα.

Γι αυτό ακριβώς το λόγο επιλέγονται συνήθως διάφορες άλλες μεθόδοι. Μια τυπική διαδικασία επιλογής χαρακτηριστικών, αποτελείται από 2 φάσεις, αυτή της επιλογής των χαρακτηριστικών και αυτή της αξιολόγησής 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 υποψηφίου 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 τρέχων σύνολο των επιλεγμένων χρακτηριστικών, είναι αρκετά καλό ή όχι.

Η τεχνική επιλογής χαρακτηριστικών, μπορεί να χρησιμοποιηθεί είτε σε προβλήματα ομαδοποίησης είτε σε προβλήματα ταξινόμησης. Σε αυτά της ταξινόμησης, ανάλογα με το πως και το πότε αξιολογείται η χρησιμότητα των υποψηφίων προς επιλογή χαρακτηριστικών οι μεθόδοι που ακολουθούνται μπορούν να χωριστούν σε 3 κατηγορίες, στα **φίλτρα** (filters), στις **wrapper** και στις **embedded** μεθόδους.



Σχήμα 3.2: Σχηματική απεικόνιση της διαδικασίας επιλογής χαρακτηριστικών

3.2 FILTER ΜΕΘΟΔΟΙ

Οι filter(φίλτρα) μεθόδοι βασίζονται στην έννοια της συνάφειας μεταξύ χαρακτηριστικών και κλάσης, δηλαδή σε αυτή τη κατηγορία μεθόδων ανήκουν όσοι αλγόριθμοι δεν βασίζονται σε κάποιο ταξινομητή προκειμένου να εκτιμήσουν την ποιότητα ενός υποσυνόλου χαρακτηριστικών, αντίθετα χρησιμοποιώντας στατιστικά μέτρα προσπαθούν να εντοπίσουν συναφή χαρακτηριστικά. Οι μεθόδοι αυτοί, χωρίζονται σε δύο βασικές κατηγορίες, τις μονοπαραγοντικές (univariate) και τις πολυπαραγοντικές (multivariate) μεθόδους.

3.2.1 ΜΟΝΟΠΑΡΑΓΟΝΤΙΚΕΣ ΜΕΘΟΔΟΙ

Οι μονοπαραγοντικές (univariate) μεθόδοι, αξιολογούν κάθε χαρακτηριστικό ξεχωριστά, με βάση τη συσχέτιση του με την κλάση. Δηλαδή όσο μεγαλύτερη είναι αυτή, τόσο πιο καλό θεωρείται το χαρακτηριστικό. Ακολούθως, επιλέγονται τα Ν πιο συσχετισμένα χαρακτηριστικά, όπου το Ν καθορίζεται ανάλογα με την περίσταση. Η κυριότερη αδυναμία αυτών των μεθόδων, είναι η εμφάνιση φαινομένων πλεονασμού, δηλαδή περιπτώσεις όπου επιλέγονται περιττά χαρακτηριστικά, δηλαδή χαρακτηριστικά τα οποία 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 επείδη, κάθε χαρακτηριστικό εξετάζεται μονο του, χωρίς να λαμβάνονται υπόψη τα άλλα που έχουν ήδη επιλεγεί.

Η συσχέτιση μεταξύ των χαρακτηριστικών, μετριέται μέσω διαφόρων κριτηρίων. Τα σημαντικότερα από αυτά, είναι το κριτήριο του Fisher καθώς και η μέτρηση της αμοιβαίας τους πληροφορίας $I(X;Y)$, η οποία μπορεί να ανιχνέυσει και τις γραμμικές εξαρτήσεις μεταξύ των μεταβλητών.

Το κριτήριο του Fisher για 2 κλάσεις δίνεται από τον τύπο

$$w_i = \frac{(\mu_{i_1} - \mu_{i_2})^2}{({\sigma_{i_1}}^2 - {\sigma_{i_2}}^2)}$$

όπου

μ_{i_1}, μ_{i_2} = οι μέσες τιμές του i-οστού χαρακτηριστικού για τα παραδείγματα της 1ης και 2ης κλάσης αντίστοιχα

$\sigma_{i_1}, \sigma_{i_2}$= οι τυπικές αποκλείσεις του i-οστού χαρακτηριστικού για τα παραδείγματα της 1ης και 2ης κλάσης αντίστοιχα

Στην περίπτωση που η τιμή του w_i είναι μεγάλη, σημαίνει ότι τα παραδείγματα διαφέρουν σημαντικά μεταξύ τους ως προς το i χαρακτηριστικό.

3.2.2 ΠΟΛΥΠΑΡΑΓΟΝΤΙΚΕΣ ΜΕΘΟΔΟΙ

Οι πολυπαραγοντικές (multivariate) μεθόδοι αξιολογούν τα χαρακτηριστικά, λαμβάνοντας υπόψη τους πάντα και την παρουσία των άλλων χαρακτηριστικών, προκειμένου να αποφευχθεί η επιλογή περιττών στοιχείων. Με λίγα λογια, οι πολυπαραγοντικές μεθόδοι, επικεντρώνονται σε 2 στόχους όσο αφορά την επιλογή ενός υποσυνόλου, πρώτον αυτό να περιέχει χαρακτηριστικά που έχουν μεγάλη συσχέτιση σε σχέση με την κλάση και ταυτόχρονα, τα χαρακτηριστικά αυτά 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. Η κύρια μέθοδος αυτού του τύπου είναι η mRMR, η οποία βασίζεται στην αμοιβαία πληροφορία και η οποία θα περιγραφεί αναλυτικά σε επόμενο κεφάλαιο.

Η κύρια διαφορά των μεθόδων φίλτρου, σε σχέση με τις wrapper μεθόδους, είναι ότι εδώ κριτήριο δεν θεωρείται κάποιο από τα μέτρα διαχωρισιμότητας των κλάσεων, αλλά η ίδια η απόδοση του ταξινομητή. Επιπλέον οι μεθόδοι φίλτρου είναι πιο γρήγορες σε εκτέλεση αφού δεν περιλαμβάνουν επαναλήψεις, ενώ μπορούν και πιο εύκολα να γενικευτούν, αφού δεν βασίζονται σε ένα συγκεκριμένο ταξινομητή.

3.3 WRAPPER ΜΕΘΟΔΟΙ:

Στην κατηγορία των wrapper μεθόδων εντάσσονται όλοι οι αλγόριθμοι επιλογής χαρακτηριστικών που χρησιμοποιούν την ακρίβεια ταξινόμησης ως κριτήριο αξιολόγησης των υποσυνόλων. Η αξιολόγηση με βάση την απόδοση του ταξινομητή απαιτεί και την κατασκευή του ταξινομητή για κάθε υποσύνολο χαρακτηριστικών που εξετάζεται με αρνητικό επακόλουθο το αυξημένο υπολογιστικό κόστος σε σχέση με τις πιο εξελιγμένες embedded μεθόδους ή τα φίλτρα. Παρόλα αυτά οι wrapper έχουν και κάποια συγκεκριμένα πλεονεκτήματα που κάνουν τη χρήση τους ελκυστική.
Η διαδικασία που ακολουθεί μια τυπική wrapper μέθοδος είναι η εξής. Αρχικά χωρίζει τα δεδομένα εκπαίδευσης σε 2 νεα σύνολα, το σύνολο εκαπίδευσης (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 αυτά τα δεδομένα αυτά ταξινομούνται, είναι το κριτήριο αξιολόγησης των wrapper μεθόδων για ένα οποιοδήποτε υποψήφιο σύνολο χαρακτηριστικών.

Αυτές οι μεθόδοι, αντιμετωπίζουν μεγάλες δυσκολίες, όταν ο αριθμός των δεδομένων εκπαίδευσης είναι μικρός. Σε αυτή τη περίπτωση, δεν μπορούν να σχηματιστούν ικανοποιητικά σε μέγεθος σύνολα, πράγμα καταστροφικό, αφού στη περίπτωση που το σύνολο εκπαίδευσης είναι μικρό, δεν μπορεί να γίνει καλή εκπαίδευση του ταξινομητή, ενώ όταν το σύνολο επικύρωσης είναι μικρό, δεν μπορεί να γίνει αξιόπιστη εκτίμηση όσο αφορά την ακρίβεια στη ταξινόμηση.

Το πρόβλημα αυτό έρχεται να το λύσει η χρήση της μεθόδου cross-validation χάρη στην οποία αποφεύγεται και η υπερπροσαρμογη (over-fitting) του συνόλου εκπαίδευσης. Σύμφωνα με αυτή τη τεχνική, τα δεδομένα εκπαίδευσης, χωρίζονται σε k υποσύνολα ξένα μεταξύ τους. Στη συνέχεια ο ταξινόμητης εκπαιδεύεται με τα k-1 από αυτά, αφήνοντας το τελευταίο να παίξει το ρόλο του συνόλου επικύρωσης. Η διαδικασία αυτή επαναλαμβάνεται k-φορές, ούτως ώστε όλα τα υποσύνολα να παίξουν το ρόλο του συνόλου επικύρωσης. Τέλος το κριτήριο αξιολόγησης υπολογίζεται βρίσκοντας το μέσο όρο της ακρίβειας όλων των συνόλων επικύρωσης.

Όπως αναφέρθηκε και στην εισαγωγή, οι wrapper μεθόδοι παρουσιάζουν συγκεκριμένα πλεονεκτήματα που τις καθιστούν εκλυστικές στη χρήση. Ένα από αυτά είναι το γεγονός ότι μπορούν να χρησιμοποιηθούν με οποιοδήποτε ταξινομητή, καθώς δεν εξαρτώνται από το τρόπο λειτουργίας αυτών, αλλά τους χρησιμοποιούν μόνο για την αξιολόγηση των υποψηφίων συνόλων χαρακτηριστικών.

Το κύριο πλεονέκτημα τους ομως, είναι ότι λαμβάνουν σοβαρά υπόψη τους την επαγωγική μεροληψία (inductive bias) του ταξινομητή. Η επαγωγική μεροληψία, είναι το σύνολο 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. Οι wrapper μεθόδοι, λαμβάνουν υπόψη ότι κάθε ταξινομητής έχει τα δικά του ιδιαίτερα χαρακτηριστικά και έτσι απεικονίζει με το δικό του ξεχωριστό τρόπο, κάθε είσοδο(input) που δέχεται σε έξοδο(output), χωρίς να σημαίνει κατ' ανάγκη ότι το καλύτερο σύνολο που σχηματίζει ένας ταξινομητής, θα είναι το καλύτερο και για τους υπόλοιπους.

Αντίστοιχα, το κύριο μειονέκτημα των μεθόδων wrapper, είναι το μεγάλο υπολογιστικό κόστος που έχουν, κόστος που οφείλεται κυρίως στο γεγονός ότι για να αξιολογηθεί ένα υποσύνολο, πρέπει πρώτα να εκπαιδευτεί ο ταξινομητής και ακολούθως να μετρηθεί η απόδοση του συνόλου επικύρωσης, πράγμα το οποίο καθιστά τη διαδικασία ιδιαίτερη χρονοβόρα και αργή.

Επιπλέον ένα άλλο μειονέκτημα, είναι ότι υπάρχουν φορές όπου επιλέγονται υποσύνολα θεωρητικά βέλτιστα, πάντα με βάση την ακρίβεια που δίνουν στη ταξινόμηση του συνόλου επικύρωσης τους, τα οποία όμως αποδεικνύεται ότι είναι κακά, καθιστώντας έτσι την όλη διαδικασία μη αξιόπιστη. Αυτό συμβαίνει κυρίως λόγω του μεγάλου όγκου των υποσυνόλων που εξετάζονται, πράγμα το οποίο μπορεί να οδηγήσει στον εντοπισμό ενός συνόλου το οποίο να έχει καλή απόδοση στο σύνολο επικύρωσης χωρίς όμως να έχει καλή ικανότητα γενικά. Αυτό το πρόβλημα, γίνεται πιο έντονο στις περιπτώσεις όπου το σύνολο των δεδομένων εκπαίδευσης είναι πολύ μικρό.

Οι κύριες wrapper μεθόδοι είναι οι forward selection και η backward selection method οι οποίες περιγράφονται πιο κάτω.

3.3.1 ΜΕΘΟΔΟΣ ΤΗΣ ΠΡΟΣ ΤΑ ΕΜΠΡΟΣ ΕΠΙΛΟΓΗΣ (FORWARD SELECTION)

Ο αλγόριθμος χτίζει αυξητικά το υποσύνολο των επιλεγμένων χαρακτηριστικών. Αρχικά το υποσύνολο επιλεγμένων χαρακτηριστικών είναι κενό. Σε ένα τυπικό 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 ένα από πριν επιλεγμένο κριτήριο, ενσωματώνεται στο τρέχον υποσύνολο. 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.

Συγκεκριμένα ο στόχος είναι η επιλογή ενός υποσυνόλου που να αποτελείται από m χαρακτηριστικά, μέσα από ένα μεγαλύτερο σύνολο Sn. Αρχικά, ορίζεται το σύνολο των χαρακτηριστικών, να είναι κενό. 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 γίνεται η μεγαλύτερη μείωση στο σφάλμα της ταξινόμησης. Το σύνολο που περιέχει αυτό το χαρακτηριστικό ονομάζεται Z_1.

Η μέθοδος συνεχίζεται για τον εντοπισμό του δεύτερου χαρακτηριστικού, του X_2, από το σύνολο $\{S_n\text{-}Z_1\}$, για το οποίο το υποσύνολο $Z_2=\{Z_1,X_2\}$, οδηγεί στη μεγαλύτερη μείωση του σφάλματος. Αυτή η διαδικασία συνεχίζεται ακριβώς με τον ίδιο τρόπο, μέχρις ότου, το σφάλμα αρχίσει να αυξάνεται αντί να μειώνεται, δηλαδή όταν, έστω στην k-οστή επιλογή χαρακτηριστικού, ισχύει ότι $e_{k+1}>e_k$. Τότε σταματά η διαδικασία. Πρέπει να τονιστεί ότι σε περίπτωση ισότητας, δηλ. στην περίπτωση που ισχύει $e_{k+1}=e_k$, στην k-οστή επιλογή, η μέθοδος συνεχίζεται κανονικά. Οταν ικανοποιηθεί η συνθήκη τερματισμού, τα χαρακτηριστικά που τελικά επιλέγονται, πρέπει να ισούνται με τη διάσταση που είχε το υποσύνολο, όταν επιτεύχθηκε για πρώτη φορά το χαμηλότερο σφάλμα.

Για παράδειγμα, αν με τη μέθοδο αυτή, τα σφάλματα ταξινόμησης ανα βήμα είναι τα ακόλουθα, έστω {9,7,6,3,3,3,7}, παρόλο που η διαδικασία σταματά στο 6^o βήμα, εντούτοις, τα χαρακτηριστικά που τελικά επιλέγονται για να αποτελούν το τελικό υποσύνολο, θα είναι μόνο τα πρώτα 4.

3.3.2 ΜΕΘΟΔΟΣ ΤΗΣ ΠΡΟΣ ΤΑ ΠΙΣΩ ΕΠΙΛΟΓΗΣ (BACKWARD SELECTION)

Η μέθοδος αυτή, είναι η αντίθετη της προς τα εμπρός επιλογής χαρακτηριστικών. Εδώ, η διαδικασία, ξεκινά έχοντας όλες τις μεταβλητές διαθέσιμες. Κάθε μεταβλητή ελέγχεται μέσω ενός κριτηρίου συγκρίσεως, βάση του οποίου εξετάζεται πως θα είναι το μοντέλο χωρίς αυτή τη μεταβλητή. Αυτές οι οποίες συμβάλουν στη καλυτέρευση του μοντέλου με την απομάκρυνση τους, διαγράφονται. Η διαδικασία σταματά, όταν το μοντέλο δεν μπορεί να βελτιωθεί περετέρω.

Συγκεκριμένα, η μέθοδος αυτή, αποκλείει σε κάθε βήμα ένα χαρακτηριστικό για το οποίο, το υποσύνολο S_{k-1} που προκύπτει χωρίς αυτό, θα έχει σφάλμα ταξινόμησης e_{k-1}, που είναι μικρότερο από το σφάλμα e_k, που είχε το υποσύνολο S_k το οποίο περιελάμβανε αυτό το χαρακτηριστικό. Λόγω του ότι όλα τα χαρακτηριστικά του συνόλου S_k, θεωρούνται υποψήφια για διαγραφή, σχηματίζονται k διαφορετικά υποσύνολα S_{k-1}, για το οποία υπολογίζεται και το αντίστοιχο σφάλμα τους. Αν για κάθε διαμορφωμένο σύνολο S_{k-1}, το e_{k-1}, είναι μεγαλύτερο του e_k, τότε δεν υπάρχει βελτίωση του μοντέλου και επομένως ούτε και μεγαλύτερη ακρίβεια στην ταξινόμηση, δηλαδή κάθε χαρακτηριστικό του συνόλου S_k είναι χρήσιμο και έτσι η

μέθοδος backward τερματίζεται. Σε περίπτωση που δεν συμβαίνει αυτό, τότε από το σύνολο των k διαμορφομένων S_{k-1} επιλέγεται σαν καινούργιο σύνολο χαρακτηριστικών, αυτό που έχει το μικρότερο σφάλμα. Η διαδικασία συνεχίζεται μέχρι να ικανοποιηθεί η συνθήκη τερματισμού που αναφέρθηκε πιο πριν.

3.4 EMBEDDED ΜΕΘΟΔΟΙ

Στην κατηγορία αυτή ανήκουν οι μεθόδοι που έχουν σχεδιαστεί με στόχο να δουλεύουν μόνο σε συνεργασία με ένα ταξινομητή συγκεκριμένου τύπου. Σε αντίθεση με τις wrapper που απλώς χρησιμοποιούν την έξοδο ενός ταξινομητή, οι μέθοδοι αυτές επιλέγουν χαρακτηριστικά με βάση το πως επηρεάζεται κάποια συνάρτηση κόστους που εμπλέκεται στη διαδικασία εκπαίδευσης του ταξινομητή.

Μέσω αυτής της ενσωμάτωσης της επιλογής χαρακτηριστικών στη διαδικασία εκπαίδευσης προκύπτουν διάφορα πλεονεκτήματα όπως μεγάλο κέρδος σε υπολογιστικό κόστος. Επίσης, οι embedded μέθοδοι καταφέρνουν να κάνουν καλύτερη χρήση των διαθέσιμων δεδομένων αφού δεν υπάρχει η ανάγκη αυτά να χωριστούν σε σύνολα εκπαίδευσης και επικύρωσης. Ακόμη υπερτερούν των filter μεθόδων στο γεγονός, ότι μπορούν να λαμβάνουν υπόψη τους, την επαγωγική μεροληψία, όπως και οι wrapper. Οι πιο σημαντικές από αυτές τις μεθόδους, έχουν σχεδιαστεί έτσι ώστε να λειτουργούν σε συνδυασμό με την Μηχανή Διανυσματικής Υποστήριξης *(*support vector machine-SVM) για την εκτέλεσή τους. *Οι κύριοι* embedded μεθόδοι είναι τα δέντρα αποφάσεων (decision trees), η μέθοδος της ελαχιστης απολυτης συστολης και επιλογης φορέα (least absolute shrinkage and selection operator-LASSO) και η μέθοδος του τυχαιου πολυωνυμικου λογαριθμου (random multinomial logit-RMNL).

3.5 ΔΙΑΦΟΡΕΣ ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Πιο κάτω παρουσιάζονται επιγραμματικά κάποιες από τις πιο γνωστές μεθόδους επιλογής χαρακτηριστικών.

3.5.1 ΤΥΧΑΙΑ ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ(RANDOM FEATURE SELECTION)

Η τυχαία επιλογή χαρακτηριστικών, θεωρείται η πιο τετριμμένη μέθοδος επιλογής. Βασίζεται συγκεκριμένα στην ομοιόμορφη τυχαία επιλογή ενός δείγματος δεδομένων χωρίς επανάληψη. Μέσω αυτής της διαδικασίας, επιλέγονται χαρακτηριστικά, ανεξάρτητα μεταξύ τους, τα οποία όμως δεν είναι πάντα αυτά που παρέχουν την περισσότερη πληροφορία, με συνέπεια να οδηγούμαστε σε φτωχά αποτελέσματα, εφόσον τις πλείστες φορές, μονο ένα μικρό μέρος των επιλεγμένων χαρακτηριστικών, παρέχει ουσιαστικά πληροφορίες για την προς πρόβλεψη κλάση.

3.5.2 ΜΕΘΟΔΟΣ ΜΕΓΙΣΤΟΠΟΙΗΣΗΣ ΑΜΟΙΒΑΙΑΣ ΠΛΗΡΟΦΟΡΙΑΣ

Η μέθοδος της μεγιστοποίησης της αμοιβαίας πληροφορίας(mutual information maximization-MIM), σχεδιάστηκε προκειμένου να ξεπεράσει την αδυναμία που προκύπτει, όταν η επιλογή χαρακτηριστικών γίνεται τυχαία. Συγκεκριμένα η MIM επιλέγει k χαρακτηριστικά, $\square_1, \square_2, .. \square_\square$ από ένα σύνολο $\square_1, \square_2, .. \square_\square$, όπου n>>k. Τα χαρακτηριστικά που επιλέγονται είναι αυτά που μεγιστοποιούν την αμοιβαία πληροφορία $\square(\square; \square_\square)$, μεταξύ αυτών και της κλάσης που πρέπει να προβλέψουν. Παρόλα αυτά η MIM, δεν διασφαλίζει την χαλαρή εξάρτηση μεταξύ των χαρακτηριστικών που επιλέγονται, με αποτέλεσμα, να μπορεί να οδηγήσει σε φαινόμενα πλεονασμού χαρακτηριστικών και την επιλογή χαρακτηριστικών ασήμαντων χωρίς να δίνουν καμία επιπλέον πληροφορία.

3.5.3 ΜΕΘΟΔΟΣ ΓΡΗΓΟΡΗΣ ΣΥΣΧΕΤΙΣΗΣ ΒΑΣΙΣΜΕΝΗ ΣΕ ΦΙΛΤΡΑ

Η μέθοδος γρηγορης συσχετισης βασισμενη σε φιλτρα (Fast Correlation-Based Filter-FCBF) είναι μια πολυπαραγοντική μέθοδος επιλογής χαρακτηριστικών, η οποία ξεκινά την αναζήτηση από το πλήρες σύνολο χαρακτηριστικών και ακολούθως χρησιμοποιώντας τη συμμετρική αβεβαιότητα υπολογίζει την εξάρτηση των χαρακτηριστικών, βρίσκοντας το καλύτερο υποσύνολο μέσω χρήση της προς τα πίσω wrapper μεθόδου επιλογής χαρακτηριστικών.

Η συμμετρική αβεβαιότητα (symmetrical uncertainty) είναι ένα κανονικοποιημένο θεωρητικό μέτρο πληροφορίας, που βασίζεται στην εντροπία και την υπο συνθήκη εντροπία, προκειμένου να υπολογίσει την εξάρτηση μεταξύ των χαρακτηριστικών. Για να υπολογιστεί η SU, θα πρέπει τα χαρακτηριστικά να είναι διακριτά, ενώ για τα συνεχή χαρακτηριστικά, γίνεται μια κατάλληλη διακριτικοποίηση σε αυτά προκειμένου να υπολογιστούν και αυτά. Η συμμετρική αβεβαιότητα δίνεται από τον τύπο

$$SU(X,Y) = 2\,[\frac{H(X) - H(X|Y)}{H(X) + H(Y)}]$$

Γενικά η FCBF είναι μια μέθοδος επιλογής του βέλτιστου υποσυνόλου, η οποία είναι βασισμένη στη συσχέτιση μεταξύ των χαρακτηριστικών και η οποία αποδίδει γενικά πολύ γρηγορότερα σε σχέση με άλλες μεθόδους. Τέλος η FCBF, έχει ένα εσωτερικό κριτήριο διακοπής, το οποίο τερματίζει τη διαδικασία όταν δεν υπάρχουν πλέον άλλα χαρακτηριστικά για να αποκλειστούν.

3.5.4 C4.5 ΔΥΑΔΙΚΑ ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ:

Τα δυαδικα δεντρα(binary trees), είναι μια άλλη μέθοδος που χρησιμοποιείται για επιλογή χαρακτηριστικών. Προτάθηκε για πρώτη φορά από τους Ratanamahatana και Gunopulos το 2003 και είναι βασισμένη στη δημιουργία πολλών δυαδικών δέντρών και ακολούθως στη βαθμολόγηση των χαρακτηριστικών, σύμφωνα με τον αριθμό των φορών που αυτά εμφανίζονται στους κορυφαίους κόμβους. Η τεχνική αυτή όπως προτείνεται και στη βιβλιογραφία, είναι μια μέθοδος φιλτρου η οποία δίνει πολύ καλά αποτελέσματα σε συνδυασμό με τον ταξινομητή naïve Bayes και επιπλέον ειναι ένα πολύ καλό παράδειγμα ενός συστήματος ικανού να εντοπίζει τη στατιστική εξάρτηση μεταξύ περισσοτέρων από 2 χαρακτηριστικών, εφόσον η επιλογή ενός χαρακτηριστικού σε δυαδικό δέντρο εξαρτάται από τη στατιστική συμπεριφορά του σε συνδυασμό πάντα με τα χαρακτηριστικά που επιλέχθσαν πριν από αυτό.

Ένας αποδοτικός τρόπος να αυξηθεί η αποδοτικότητα αυτής της μεθόδου, επιτυγχάνεται με τη χρησιμοποίηση τυχαίων υποσυνόλων χαρακτηριστικών αντί υποσυνόλων από εκπαιδευτικά παραδείγματα, όπως συμβαίνει στις μεθόδους bagging.

4. ΔΙΑΔΙΚΑΣΙΑ ΤΑΞΙΝΟΜΗΣΗΣ

4.1 ΕΚΜΑΘΗΣΗ ΜΗΧΑΝΩΝ:

Η εκμάθηση μηχανών, είναι μια περιοχή της τεχνητής νοημοσύνης, η οποία έχει σχέση με αλγόριθμους αλλά και μεθόδους οι οποίες επιτρέπουν στους υπολογιστές να «μαθαίνουν». Πιο συγκεκριμένα, σκοπός της εκμάθησης μηχανών είναι να καταφέρει να καταστήσει μια μηχανή έμπειρη βελτιώνοντας έτσι την αποτελεσματικότητά της για μια συγκεκριμένη λειτουργία. Η λογική μια μηχανής εκμάθησης είναι να δίνει την τιμή yi μιας συνάρτησης (άγνωστη προς εμάς) που αντιστοιχεί σε δοσμένο σημείο xi .

Προκειμένου να το πετύχει αυτό η μηχανή θα πρέπει να εκπαιδευτεί από πριν πάνω σε έτοιμα δεδομένα. Για αυτό ακριβώς το σκοπό, υπάρχουν τα δεδομένα εκπαίδευσης (training data), τα οποία αυτό ακριβώς το ρόλο παίζουν.

Δοσμένου ενός συνόλου δεδομένων εκπαίδευσης $D = \{(x_1, y_1), \dots, (x_n, y_n)$, δηλαδή ενός συνόλου όπου για κάθε $x_i \in R^n$, είναι γνωστή η κλαση y_i, στην οποία αυτό ανήκει, η μηχανή εκπαιδεύεται στο να κατανοήσει τη σχέση που υπάρχει μεταξύ των δύο, έτσι ώστε να μπορεί στο μέλλον από μόνη της να ταξινομήσει κάθε άγνωστο χαρακτηριστικό x_m, στην κλάση που αυτό ανήκει. Τα x_i σε ένα σύνολο δεδομένων εκπαίδευσης λέγονται και πρότυπα εκπαίδευσης, ενώ τα y_i στόχοι εκπαίδευσης αντίστοιχα.

Στο τέλος αυτής της διαδικασίας, αναπτύσεται ένας αλγόριθμος ο οποίος λέγεται **ταξινομητής** και είναι αυτός που θα χρησιμοποιηθεί στο μέλλον, προκειμένου να ταξινομήσει κάθε άγνωστη ποσότητα.

Οι αλγόριθμοι εκμάθησης μηχανών, χωρίζονται ανάλογα με τον τρόπο που λειτουργούν σε τρεις μεγάλες κατηγορίες, τη επιβλεπόμενη, τη μη επιβλεπόμενη και την ενισχυτική μάθηση.

- Στην **επιβλεπόμενη** μάθηση, δίνονται τα δεδομένα εκπαίδευσης (input) το καθένα από τα οποία είναι τοποθετημένο στη σωστή έξοδο(output). Ακολούθως μέσω αυτών, γενικοποιείται το μοντέλο χρησιμοποιώντας αυτά τα δεδομένα όσο καλύτερα μπορεί για να τοποθετήσει τα νέα δεδομένα όσο καλύτερα μπορεί.
- Στη **μη επιβλεπόμενη** μάθηση, δίνονται μόνο τα δεδομένα εκπαίδευσης, χωρίς επιπλέον πληροφορίες, και ο ταξινομητής τα χωρίζει σε ομάδες (clusters) με βάση κάποιο κριτήριο. Ακολούθως προσπαθεί να εντοπίσει

εναλλακτικά πρότυπα, προκειμένου να μπορέσει να καθορίσει το σωστό output, για τα νέα δεδομένα.

- Στην **ενισχυτική** μάθηση οι έξοδοι του ταξινομητή αξιολογούνται από ένα συνολικό δείκτη συμπεριφοράς και η αξιολόγηση αυτή ανατροφοδοτείται στον ταξινομητή έτσι, ώστε επιθυμητές συμπεριφορές να ενισχύονται, ενώ ανεπιθύμητες να αποτρέπονται.

4.2 ΤΑΞΙΝΟΜΗΣΗ:

Ταξινόμηση, είναι το πρόβλημα κατά το οποίο, δοσμένου ενός συνόλου δεδομένων $X = \{x_1, x_2, \dots, x_n\}$, και ενός συνόλου κλάσεων $C = \{c_1, c_2, \dots, c_m\}$, πρέπει να καθοριστεί που ανήκει κάθε ένα από τα στοιχεία του συνόλου Χ. Η κατηγοροποίηση αυτών των δεδομένων, ορίζεται μέσω της απεικόνισης $f: X \to C$, που σημαίνει ότι κάθε χαρακτηριστικό x_i, αντιστοιχεί σε μια κλάση c_j.

Η ταξινόμηση είναι μια από τις πιο βασικές τεχνικές που χρησιμοποιούνται, με σκοπό την κατάταξη δεδομένων. Η διαδικασία που ακολουθείται βασίζεται στην εξέταση ενός αντικειμένου, τη μελέτη των ιδιαίτερων χαρακτηριστικών αυτού και ακολούθως στην αντιστοίχησή του με την κατάλληλη από μια σειρά κλάσεων. Τα χαρακτηριστικά, αναλύονται με βάση κάποιες επιμερους ιδιότητες που έχουν, γνωστές και ως ερμηνευτικές μεταβλητές (explanatory variables). Αυτές οι ιδιότητες μπορεί να είναι κατηγορικές (πχ Α,Β,ΑΒ,Ο ομάδα αίματος), ακέραιες τιμές ή και διάφορα μεγέθη(πχ small, medium, large).

Με πιο απλά λόγια, η ταξινόμηση, βρίσκει τους συσχετισμούς, μεταξύ ενός συνόλου δεδομένων και στη συνέχεια τα χωρίζει ανάλογα με τις κοινές τους ιδιότητες σε διάφορες κλάσεις, σύμφωνα με ένα προκαθορισμένο μοντέλο. Οι αλγόριθμοι οι οποίοι χρησιμοποιούνται για τη ταξινόμηση, ονομάζονται ταξινομητές. Σε όρους μηχανικής μάθησης, η ταξινόμηση θεωρείται μια περίπτωση επιβλεπόμενης μάθησης, αφού δίνεται πάντα ένα σύνολο εκπαιδευτικών χαρακτηριστικών προκειμένου να βοηθήσουν στη διαδικασία.

Τέλος η ταξινόμηση σχετίζεται άμεσα με την επιλογή χαρακτηριστικών, αφού στις περιπτώσεις όπου ο αριθμός των δειγμάτων του εκπαιδευτικού συνόλου είναι μικρός σε σχέση με τον αριθμό των χαρακτηρστικών που πρέπει να κατηγοριοποιηθούν, μια αποτελεσματική ταξινόμηση, μπορεί να επιτευχθεί, μόνο μέσω της επιλογής ενός κατάλληλου υποσυνόλου από αυτά, διαδικασία η οποία γίνεται μέσω της επιλογής χαρακτηριστικών.

4.3 ΤΑΞΙΝΟΜΗΤΕΣ:

Ταξινομητές ονομάζονται μια πλειάδα μεθόδων που χρησιμοποιούνται προκειμένου να ταξινομηθούν διάφορα δεδομένα. Ο ταξινομητής είναι στην ουσία ένας αλγόριθμος που δέχεται ως είσοδο (input) άγνωστα δεδομένα (πρότυπα) και επιχειρεί κατά την έξοδο (output), να τα κατατάξει στην κατηγορία (κλάση) που ανήκει το καθένα.

Κάθε αλγόριθμος ταξινόμησης χωρίζεται εν γένει σε δύο στάδια: το στάδιο εκπαίδευσης ή μάθησης και το στάδιο ελέγχου.

- Κατά το **στάδιο μάθησης**, χρησιμοποιείται ένα μέρος του δείγματος προκειμένου να «εκπαιδευτεί» ο ταξινομητής, δηλαδή να βρεθούν οι βέλτιστες παράμετροι για τη λειτουργία του.
- Κατά το **στάδιο ελέγχου**, ελέγχεται η απόδοση του ταξινομητή σε δεδομένα διαφορετικά από αυτά του συνόλου εκπαίδευσης.

Πιο κάτω αναλύονται συνοπτικά μερικοί από τους κυριότερους αλγόριθμους ταξινόμησης που υπάρχουν, οι οποίοι είναι οι

- Naïve Bayes
- LDA
- SVM
- K-NN
- Perceptron
- AdaBoost

4.3.1 NAIVE BAYES:

Οικογένεια απλών γραμμικών ταξινομητών πιθανότητας, δηλαδή ταξινομητών οι οποίοι είναι δυνατό να προβλέπουν την κατανομή πιθανότητας ενός συνόλου κλάσεων και που παρέχουν ικανοποιητική ταξινόμηση δίνοντας επίσης το βαθμό βεβαιότητας αυτής, καθιστώντας τους έτσι πολύ χρήσιμους, ιδιαίτερα όταν εφαρμόζονται σε μεγάλα συστήματα.

Οι ταξινομητες Naive Bayes, βασίζονται στο Θεώρημα Bayes, καθώς και στην υπόθεση ότι τα χαρακτηρστικά που ταξινομούνται, δεν έχουν καμία απολύτως σχέση μεταξύ τους (είναι εντελώς ανεξάρτητα). Η επιλεγμένη κλάση Y εξαρτάται από το πρόσημο της σχέσης

$$f(x_1, \dots, x_n) = log \frac{P\ (Y = 1 | X_{\nu(1)} = x_{\nu(1)}, \dots, X_{\nu(K)} = x_{\nu(K)})}{P\ (Y = 0 | X_{\nu(1)} = x_{\nu(1)}, \dots, X_{\nu(K)} = x_{\nu(K)})}$$

όπου τα $X_{\nu(1)}, \dots, X_{\nu(K)}$, είναι τα χαρακτηριστικά που ήδη επιλέχθηκαν.

Θεωρώντας ότι, τα $X_{\nu(.)}$, είναι ανεξάρτητα μεταξύ τους, δεδομένης της κλάσης Y, και επιπλέον ότι το $a = log \frac{P(Y=1)}{P(Y-0)}$, η πιο πάνω σχέση γράφεται σαν

$$f(x_1, \dots, x_n) = log \frac{\prod_{k=1}^{N} P(X_{\nu(K)} = x_{\nu(K)} | Y = 1)}{\prod_{k=1}^{N} P(\ X_{\nu(K)} = x_{\nu(K)} | Y = 0)} + a$$

$$= \sum_{k=1}^{N} log \frac{P(X_{\nu(K)} = x_{\nu(K)} | Y = 1)}{P(X_{\nu(K)} = x_{\nu(K)} | Y = 0)} + \alpha$$

$$= \sum_{k=1}^{N} \{log \frac{P\big(X_{\nu(K)} = 1 \big| Y = 1\big)\ P(X_{\nu(K)} = 0 | Y = 0)}{P\big(X_{\nu(K)} = 1 \big| Y = 0\big)\ P(X_{\nu(K)} = 0 | Y = 1)} \} X_{\nu(K)} + \ b$$

Βάση της πιο πάνω σχέσης μπορούν να εκτιμηθούν και τα διανύσματα βαρών $(\omega_1, \dots, \omega_N)$ της σχέσης

$$f(x_1, \dots, x_n) = \sum_{k=1}^{N} \omega_k X_{\nu(k)} \ + b$$

μέσω της σχέσης

$$\omega_k = log \frac{P\big(X_{\nu(K)} = 1 \big| Y = 1\big)\ P(X_{\nu(K)} = 0 | Y = 0)}{P\big(X_{\nu(K)} = 1 \big| Y = 0\big)\ P(X_{\nu(K)} = 0 | Y = 1)}$$

Σημειωτέον ότι η πόλωση b για τη κατάταξη ενός διανύσματος, δεδομένου του ω_k, εκτιμάται εμπειρικά με στόχο να μειώσει το σφάλμα στα δεδομένα εκπαίδευσης.

Ο ταξινομητής Naïve Bayes, είναι ιδιαίτερα δημοφιλής στην Κατηγοροποίηση Κειμένων, στην οποία χρησιμοποιεί την συχνότητα με την οποία χρησιμοποιούνται οι λέξεις σαν χαρακτηριστικό. Ένας τέτοιος ταξινομητής συμπεριφέρεται με τη λογική ότι η σημασία ένός συγκεκριμένου χαρακτηριστικού, δεν σχετίζεται με την παρουσία ή όχι άλλων χαρακτηριστικών στο σύνολο. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί μήλο εάν είναι κόκκινο, στρογγυλό και έχει διάμετρο 3’ ίντζες, αφού ο ταξινομητής θεωρεί ότι κάθε ένα από αυτά τα χαρακτηριστικά συνεισφέρει

σημαντικά σε αυτή την υπόθεση, ανεξαρτήτως, αν έχει ή όχι και τα υπόλοιπα χαρακτηριστικά για να είναι όντως μήλο.

Τέλος το μεγάλο πλεονέκτημα αυτού του ταξινομητή, είναι ο μικρός αριθμός των δεδομένων εκπαίδευσης (training data) που απαιτεί, προκειμένου να εκτιμήσει τις παραμέτρους (μεσο και διασπορα), οι οποίες είναι απαραίτητες στην ταξινόμηση. Αυτό καθιστά τον πιο πάνω ταξινομητή ιδιαίτερα αποδοτικό σε πολλά σύνολα δεδομένων, με αποτέλεσμα να θεωρείται ακόμα και ισοδύναμος με άλλους ταξινομητές πιο σύγχρονους απο αυτόν.

4.3.2 ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ (SVM)

Οι ταξινομητές αυτοί, είναι στην ουσία μοντέλα μάθησης με επίβλεψη, που σχετιζονται με αλγόριθμους μάθησης, οι οποίοι αναλύουν δεδομένα και αναγνωρίζουν πρότυπα. Χρησιμοποιούνται σε προβλήματα ταξινόμησης, στη προσέγγιση της μορφής της συνάρτησης, καθώς και στην ανάλυση παλλινδρόμησης. Είναι από τις πιο σύγχρονες τεχνικές ταξινόμησης που ακολουθούνται αφού δημιουργήθηκαν μόλις το 1995, από τον Vapnik, προκύπτοντας μέσα από τη στατική θεωρία εκμάθησης(statistical learning theory), που ο ίδιος ανάπτυξε.

Ο αλγόριθμος SVM (support vector machine), είναι ένας μη γραμμικός δυαδικός ταξινομητής, που βασίζεται στις μεθόδους του πυρήνα Kernel και μπορεί να χρησιμοποιηθεί τόσο για προβλήματα 2 κλάσεων, όσο και πολλαπλών.

Συγκεκριμένα, σε κάθε πρόβλημα, δίνεται ενα σύνολο απο N δεδομένα εκπαίδευσης τα οποία αναπαρίστανται με τη μορφή $D = \{(x_i, y_i), i = 1, \dots, N\}$. Τα y_i, παίρνουν τις τιμές (+1,-1), ανάλογα με την κλαση που ανήκει το δοθέν σημείο x_i. Στόχος των δεδομένων εκπαίδευσης, είναι να βοηθήσουν τον SVM, να κατανοήσει τη σχέση μεταξύ, σημείου-κλάσης, έτσι ώστε να μπορεί από μόνος του να κατατάξει ένα οποιοδήποτε σημείο στη σωστή θέση.

Προκειμένου να αποφασίσει ο SVM σε ποια κλάση θα ανήκει κάθε νέο σημείο, θα πρέπει πρώτα να βρεθεί το όριο της κλάσης, δηλαδή να βρεθεί μια νοητή γραμμή (αντίστοιχα επίπεδο αν $x_\iota \in R^3$ ή υπερεπίπεδο για $x_\iota \in R^v$ με v>3) η οποία να μπορεί να καθορίσει ανάλογα με τη θέση που βρίσκεται το σημείο, που αυτό ανήκει. Λόγω του ότι μπορούν να βρεθούν πληθώρα τέτοιων ευθειών, οι οποίες να μπορούν να διαχωρίζουν τα x_i, στόχος είναι να βρεθεί η βέλτιστη, η οποία θα είναι αυτή που θα βρίσκεται όσο πιο μακριά γίνεται από τα στοιχεία των κλάσεων. Ο καλύτερος διαχωρισμός, επιτυγχάνεται από το υπερεπίπεδο το οποίο θα έχει τη μεγαλύτερη απόσταση από τη πιο κοντινή κουκκίδα κάθε κλάσης, αφού ισχύει ότι όσο πιο μεγάλο είναι το περιθώριο μεταξύ των κατηγοριών, τόσο μικρότερη είναι και η πιθανότητα για ένα γενικό σφάλμα στη ταξινόμηση.

Ο SVM προκειμένου να διευκολύνει τη διαδικασία, παρουσιάζει τα δεδομένα με τη μορφή κουκκίδων στο χώρο. Επιπλέον, προκειμένου να κρατηθούν οι υπολογισμοί σε ένα λογικό επίπεδο, οι απεικονίσεις που γίνονται από το SVM, σχεδιάζονται έτσι ώστε να διασφαλίζουν ότι όλα τα χαρακτηριστικά θα υπολογίζονται απλά και εύκολα αφου θα ορίζονται όλα με τον ίδιο τρόπο.

Γενικά τα προβλήματα που αντιμετωπίζουν οι SVM, μπορούν να χωριστούν σε 3 κατηγορίες, τις **πλήρως διαχωρισμένες κλάσεις**, τις **μη απόλυτα διαχωρίσιμες κλάσεις** και τις **μη διαχωρίσιμες κλάσεις**.

ΠΛΗΡΩΣ ΔΙΑΧΩΡΙΣΙΜΕΣ ΚΛΑΣΕΙΣ:

Ονομάζονται οι κλάσεις, για τις οποίες τα σημεία εκπαίδευσης, μπορούν να χωριστούν πλήρως από ένα υπερεπίπεδο.

Συγκεκριμένα δοθέντος ενός συνόλου $(x_i, y_i), όπου\ x_i \in R, y_i \in (-1,1), i = 1, \dots, n,$ στόχος είναι να βρεθεί το βέλτιστος διαχωρισμός των δεδομένων μέσω ενός υπερεπιπέδου, το οποίο γράφεται στη μορφή

$$\vec{w}\vec{x} + b = 0,\ \ οπου$$

$\vec{x} =$ σημεία του συνόλου εκπαίδευσης

$\vec{w} =$ το διάνυσμα προσανατολισμού της διαχωρ. ευθείας

$b =$ μια σταθερά

Ορίζοντας σαν p_1, μια μικρή απόσταση από το υπερεπίπεδο στην πλευρά της κλάσης +1 και p_2, την αντίστοιχη απόσταση από την κλάση -1, τότε αφού το βέλτιστο υπερεπίπεδο θεωρείται αυτό το οποίο θα απέχει περισσότερο από τις 2 κλάσεις ταυτόχρονα, στόχος είναι η μεγιστοποίηση της απόστασης (p_1,p_2), απόσταση η οποία ονομάζεται και διάκενος(margin).

Για την επίλυση του πιο πάνω προβλήματος ορίζονται οι εξισώσεις

$$\vec{w}\vec{x} + b = +1\ \ y = +1$$

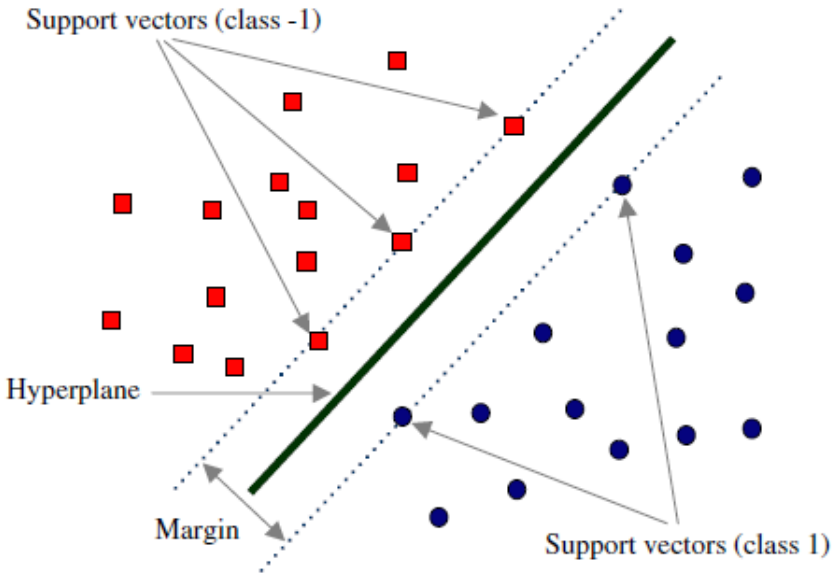
$$\vec{w}\vec{x} + b = -1\ \ y = -1$$

οι οποίες απεικονίζουν 2 ευθείες που απέχουν p_1 και p_2, αντίστοιχα από τη διαχωριστική γραμμή(υπερεπίπεδο).

Προφανώς όλα τα σημεία για τα οποία ισχύει ότι $\vec{w}\vec{x} + b \geq +1$ θα ανήκουν στην κλάση +1 ενώ αυτά για τα οποία θα ισχύει ότι $\vec{w}\vec{x} + b \leq -1$ θα ανήκουν στην κλάση -1.

Οι περιορισμοί αυτοί γράφονται στη γενική μορφή σαν

$$y_i(\vec{w}\vec{x_i} + b) - 1 \geq 0 \;\; \forall\, i = 1, \dots, n$$



Σχήμα 4.1: Απεικόνιση των δεδομένων 2 πλήρως διαχωρίσιμων κλάσεων, μαζί με τον διάκενο (margin) και τα ακραία τους σημεία (support vectors)

Η απόσταση των (p_1,p_2), γράφεται και σαν $\frac{2}{||\vec{w}||}$. Ζητούμενο είναι η μεγιστοποίηση αυτής της ποσότητας. Δηλαδή για να επιτευχθεί ο μέγιστος διάκενος θα πρέπει να ελαχιστοποιηθεί η ποσότητα

$$\frac{1}{2}\,||\vec{w}|| = \frac{1}{2}\sum_{i=1}^{n} w_i w_i$$

Η πιο πάνω εξίσωση, μπορεί να επιλυθεί με τη βοήθεια των πολλαπλασιαστών Lagrange, οι οποίοι συμβολίζονται με α. Μετά από πράξεις, το τελικό πρόβλημα θα έχει τη μορφή

$$\max Lp = \sum a_i - \frac{1}{2}\sum a_i a_j y_i y_j x_i x_j$$

$$\sum a_i y_i = 0$$

$$a_i \geq 0$$

Όπου οι a_i, παίρνουν μηδενικές τιμές για τα ακραία σημεία(support vectors)

Τέλος, για την ταξινόμηση του συνόλου των δεδομένων στις 2 κατηγορίες, χρησιμοποιείται η συνάρτηση απόφασης

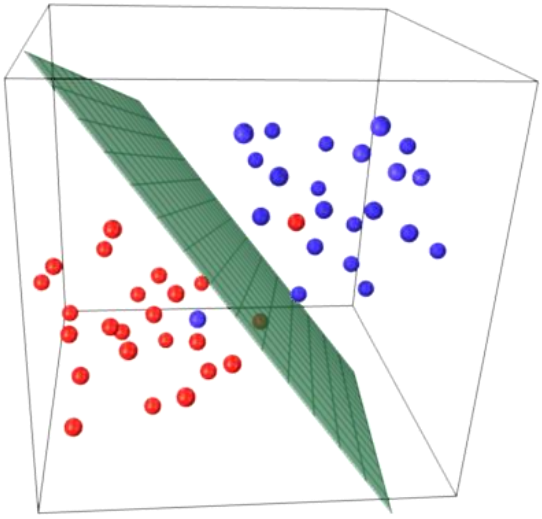
$$f = (\vec{w}\,\vec{x} + b),$$

Για αυτή τη συνάρτηση, ελέγχεται το πρόσημο της, δηλαδή το sign(f). Αν αν το πρόσημο της συνάρτησης είναι θετικό, τότε ανήκει στην κλάση +1 ενώ αν είναι αρνητικό, ανήκει στην κλάση -1.

ΜΗ ΑΠΟΛΥΤΑ ΓΡΑΜΜΙΚΑ ΔΙΑΧΩΡΙΣΙΜΕΣ ΚΛΑΣΕΙΣ:

Σε αυτή τη κατηγορία, ανήκουν τα προβλήματα, για τα οποία τα σημεία εκπαίδευσης ανήκουν σε κλάσεις οι οποίες όμως δεν διαχωρίζονται πλήρως(δηλαδη έχουν μια περιοχή που είναι κοινή).

Επειδή, υπάρχει μια περιοχή για την οποία υπάρχουν σημεία και των 2 κλάσεων, δεν μπορεί να οριστεί ευθεία διαχωριστική γραμμή. Μια λύση θα ήταν να χαραχθεί μια γραμμή(αντιστ. υπερεπίπεδο), η οποία να περνά μέσα από αυτά τα σημεία. Κάτι τετοιο όμως θα ήταν ασύμφορο, αφού θα μείωνε την απόσταση μεταξύ των 2 κλάσεων και επιπλέον θα προκαλούσε αύξηση της πολυπλοκότητας της μεθόδου.



Σχήμα 4.2: Απεικόνιση 2 κλάσεων που δεν μπορούν να διαχωριστούν πλήρως. Φαίνονται κάποια ακραία σημεία τα οποία παραμένουν σε λάθος κλάση και δεν παίζουν σημαντικό ρόλο στο καθορισμό του υπερεπιπέδου

Για αυτό, μια καλύτερη λύση που προτείνεται είναι η χαλάρωση των περιορισμών των εξισώσεων,

$$\vec{w}\vec{x} + b \geq +1$$

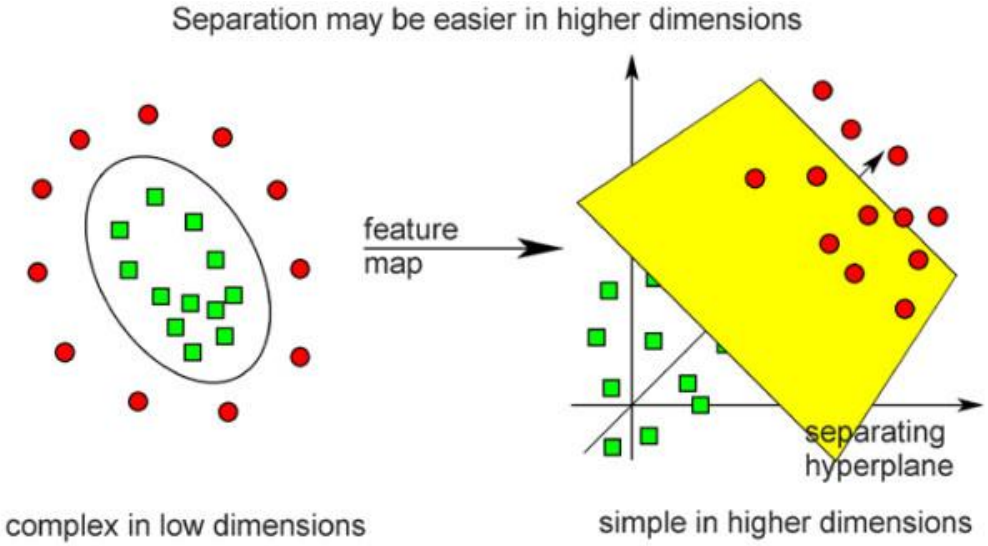
$$\vec{w}\vec{x} + b \leq -1.$$

που χρησιμοποιηθήκαν πιο πριν. Συγκεκριμένα, τα σημεία τα οποία βρίσκονται στην κοινή περιοχή, αυτή δηλαδή που δεν μπορεί να χωριστεί σε κλάσεις, ορίζονται έτσι ώστε να μην παίζουν πολυ βασικό ρόλο, δεν παίζουν δηλαδή των ρόλο των ακραίων σημείων και δεν διαδραματίζουν έτσι πρωτεύοντα ρόλο για τον ορισμό του υπερεπιπέδου.

ΜΗ ΔΙΑΧΩΡΙΣΙΜΕΣ ΚΛΑΣΕΙΣ:

Τα προβλήματα, για τα οποία, τα δεδομένα δεν μπορούν να διαχωριστούν από κανένα υπερεπίπεδο. Συμβαίνει σε αρκετές περιπτώσεις, όταν π.χ τα δεδομένα της μιας κλάσης είναι κατανεμημένα γύρω από τα δεδομένα της άλλης). Είναι η πιο κοινή περίπτωση των προβλημάτων που καλείται να επιλύσει η SVM.

Σε αυτές τις περιπτώσεις, η κατάσταση αντιμετωπίζεται μόνο μέσω του διαχωρισμού των κλάσεων μέσω μιας άλλης μορφής γραμμής, π.χ καμπύλης. Λόγω του ότι, όπως ειπώθηκε και πιο πάνω μέσω της αναζήτησης διαφορετικών τύπων γραμμών, αυξάνεται η πολυπλοκότητα της μεθόδου, στόχος είναι η επίλυση του προβλήματος, με τη χρήση των υπάρχουσων εξισώσεων μέσω των κατάλληλων μετασχηματισμών, αφού πλέον θα αναφέρονται σε άλλο είδος γραμμής.



Σχήμα 4.3: Παράδειγμα μη διαχωρίσιμων κλάσεων και απεικόνιση αυτών σε περισσότερες διαστάσεις ώστε να καταστεί εφικτός ο διαχωρισμός του από ένα υπερεπίπεδο

Αρχικά απεικονίζονται τα δεδομένα σε ένα χώρο μεγαλύτερων διαστάσεων, έτσι ώστε να μπορούν να διαχωριστούν από κάποιο υπερεπίπεδο και να γίνουν έτσι οι κλάσεις γραμμικά διαχωρίσιμες. Η απεικόνιση Z του χώρου R^k στο χώρο H, ο οποίος είναι ένας οποιοσδήποτε χώρος στον οποίο μπορεί να γίνει ένας πλήρης διαχωρισμός των δεδομένων, έχει την εξής μορφή.

$$Z: R^k \rightarrow H$$

Ακολουθώντας την ίδια διαδικασία που έγινε στις πλήρως διαχωρίσιμες κλάσεις, εφαρμόζονται και εδώ οι ίδιες σχέσεις του γραμμικού διαχωρισμού, αυτή τη φορά όμως στον νέο χωρο H. Σημειωτέον, ότι αυτή η μέθοδος μπορεί να εφαρμοστεί και στις μη πλήρως διαχωρίσιμες κλάσεις. Το πρόβλημα επομένως ανάγεται στην ακόλουθη μορφή

$$max: Ld = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j=1}^{n} a_i a_j y_i y_j < x_i x_j >$$

$$\sum_{i=1}^{n} a_i y_i = 0$$

$$a_i \geq 0$$

Ο όρος $< x_i x_j >$ είναι το αποτέλεσμα του εσωτερικού γινομένου των σημείων στο χώρο μεγαλύτερων διαστάσεων. Δηλαδή ισούται με $< x_i x_j > = K(x_i, x_j)$ όπου K μια συνάρτηση που μας δίνει το αποτέλεσμα του εσωτερικού γινομένου στο χώρο H χωρίς να είναι απαραίτητο να γνωρίζουμε το χώρο αυτό. Η συνάρτηση K ονομάζεται συνάρτηση κελύφους (Kernel function). Ενδεικτικά αναφέρεται, ότι στην περίπτωση που ισχύει ότι $K(x_i, x_j) = x_i x_j$, τότε το πρόβλημα συμπίπτει με αυτό στις πλήρως διαχωρίσιμες κλάσεις.

Τέλος, όπως και πριν η κατάταξη των σημείων εξαρτάται και πάλι από το πρόσημο της ακόλουθης εξίσωσης, η οποία μετασχηματίζεται αφού αυτή τη φορά η κατάταξη γίνεται στο χώρο H και έχει έτσι τη μορφή

$$f = sign\,(\sum y_i a_i K(x, x_i) + b^*$$

Όπου αν η f έχει θετικό πρόσημο, τότε ανήκει στην +1 κλάση, ενώ αν έχει αρνητικό, ανήκει στην -1 κλάση.

4.3.3 ΓΡΑΜΜΙΚΗ ΔΙΑΚΡΙΤΗ ΑΝΑΛΥΣΗ (LDA)

Ο LDA (linear discriminant analysis) είναι ένας από τους πιο απλούς ταξινομητές που υπάρχουν, με χρήση τόσο στη στατιστική όσο και στην εκμάθηση μηχανών. Χρησιμοποιείται προκειμένου να εντοπίσει ένα γραμμικό συνδυασμό απο χαρακτηριστικά, προκειμένου να τα χαρακτηρίσει ή να τα διαχωρίσει σε 2 ή και περισσότερες κλάσεις. Το αποτέλεσμα που προκύπτει, συνήθως χρησιμοποιείται για τη μείωση της διάστασης των δεδομένων πριν να γίνει η ταξινόμησή τους. Συγκεκριμένα χρησιμοποιώντας ως δείγμα εκμάθησης, ένα σύνολο εναλλακτικών δραστηριοτήτων των οποίων η ταξινόμηση τους είναι γνωστή και σκοπός είναι η ανάπτυξη μιας σειράς διακριτών συναρτήσεων οι οποίες να μεγιστοποιούν τη διακύμανση μεταξύ των κλάσεων.

LDA ΓΙΑ 2 ΚΛΑΣΕΙΣ:

Στην περίπτωση που υπάρχουν 2 κλάσεις, η μέθοδος LDA, προσεγγίζει το πρόβλημα ως εξής. Αρχικά, υποθέτει οι συναρτήσεις πυκνότητας πιθανότητας $p(\vec{x}|y = 0)\ και\ p(\vec{x}|y = 1)$ κατανέμονται κανονικά μαζί με τη μέση τιμή και τη διασπορά τους $(\vec{\mu_o}, \sigma_{y=0})$, $(\vec{\mu_1}, \sigma_{y=1})$ αντίστοιχα. Μέσω αυτής της υπόθεσης και εφόσον ισχύει ότι

$$\frac{\vec{x}-\overrightarrow{\mu_0}}{\sigma_0}\,(\vec{x}-\overrightarrow{\mu_0})+\ln|\sigma_0|-\frac{\vec{x}-\overrightarrow{\square_1}}{\square_1}\left(\vec{\square}-\overrightarrow{\square_1}\right)-\ln|\square_1|<\square$$

όπου Τ, ένα άνω κατώφλι, τότε τα σημεία ανήκουν στη 2η κλάση, αλλιώς ανήκουν στην 1η κλάση.

Η LDA, κάνει επιπλέον την πρόβλεψη ότι οι διασπορές είναι πανομοιότυπες, δηλαδή ότι ισχύει $\sigma_0=\sigma_1=\sigma$. Απο αυτό η πιο πάνω σχέση μετατρέπεται στη σχέση

$$\vec{w}\ \vec{x}>c$$

Για κάποια σταθερά c, όπου

$$w\propto\sigma(\overrightarrow{\mu_1}-\overrightarrow{\mu_0})$$

Δηλαδή με λίγα λόγια, το κριτήριο που υπολογίζει το κατά ποσο μια μεταβλητή x_i, ανήκει σε μια κλάση y, είναι στη ουσία μια συνάρτηση του πιο πάνω συνδυασμού των ήδη γνωστών μεταβλητών.

LDA ΓΙΑ ΠΟΛΛΕΣ ΚΛΑΣΕΙΣ:

Σε περίπτωση που οι κλάσεις είναι περισσότερες από 2, με μέση τιμή μ_ι και διασπορά να είναι κοινή, έστω σ, τότε η μεταβολή μεταξύ των κλασεων μπορεί να καθοριστεί από τη δειγματική διασπορά των μέσων των κλάσεων σ_b, δηλαδή από τον τύπο

$$\sigma_\beta=\frac{1}{N}\sum_{\iota=1}^{N}(\mu_i-\mu)(\mu_\iota-\mu)^T$$

όπου μ είναι η μέση τιμή όλων των μέσων τιμών της κλάσης

Επιπλέον, ο διαχωρισμός μεταξύ των κλάσεων, στην διευθυνση του ω, μπορεί να οριστεί από τον τύπο

$$S=\frac{\omega^T\sigma_\beta\vec{\omega}}{\overrightarrow{\omega^\tau}\,\sigma\,\vec{\omega}}$$

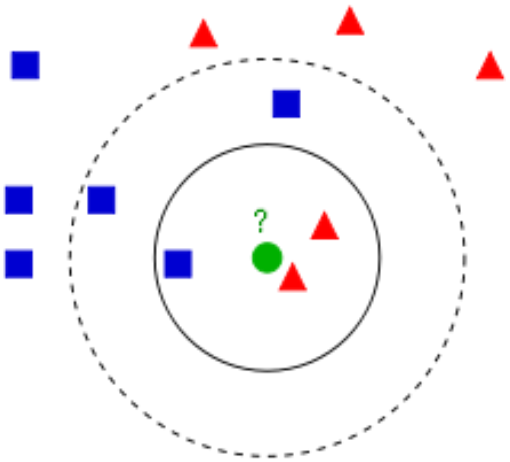
Ο πιο πάνω τύπος, σημαίνει ότι οποτε το $\vec{\omega}$, είναι ένα ιδιοδυάνυσμα του σ_β, τότε ο διαχωρισμός θα είναι ίσος με την αντίστοιχη ιδιοτιμή του.

4.3.4 ΑΛΓΟΡΙΘΜΟΣ Κ-ΠΛΗΣΙΕΣΤΕΡΩΝ ΓΕΙΤΟΝΩΝ (k-NN)

Ο αλγόριθμος k-NN (k-nearest neighbors algorithm) θεωρείται ο απλούστερος μη-παραμετρικός αλγόριθμος μάθησης. Αυτός ο αλγόριθμος, δεν κάνει προβλέψεις για τις υποκειμενικές κατανομές δεδομένων, πράγμα πολυ χρήσιμο, αφού σε πραγματικές συνθήκες τα περισσότερα αληθινά δεδομένα, δεν ακολουθούν τις θεωρητικές προβλέψεις. Επιπλέον ο k-NN, δεν κάνει χρήση των δεδομένων εκπαίδευσης προκειμένου να κάνει οποιεσδήποτε γενικεύσεις, δηλαδή δεν περνά από τη φάση εκπαίδευσης, όπως συμβαίνει με τους άλλους ταξινομητές. Αντίθετα τα δεδομένα εκπαίδευσης, χρησιμοποιούνται μόνο κατά τη διάρκεια των ελέγχων, όχι πιο πριν. Εντούτοις, παρόλο που συμβαίνει αυτό, το υπολογιστικό κόστος αυτής της μεθόδου είναι ιδιαίτερα υψηλό, τόσο σε χρόνο όσο και στην απαιτούμενη μνήμη που χρειάζεται.

Τα δεδομένα εκπαίδευσης, είναι διανύσματα ενός πολυδιάστατου χώρου χαρακτηριστικών των οποίων η κλάση στην οποία ανήκουν είναι γνωστή εκ των προτέρων. Κατά τη φάση της εκπαίδευσης, τα δεδομένα αυτά αποθηκεύονται για να χρησιμοποιηθούν στη συνέχεια. Στην επόμενη φάση, δηλαδή αυτή της ταξινόμησης, ορίζεται από πριν ένας σταθερός αριθμός k, ο οποίος είναι πάντοτε θετικός και ο οποίος είναι υπεύθυνος για να αποφασίσει πόσοι γείτονες, μπορούν να επηρεάσουν το ταξινομητή. Ακολούθως, εκτελούνται τα παρακάτω βήματα.

- Αρχικά, υπολογίζονται οι αποστάσεις του προτύπου προς ταξινόμηση από κάθε πρότυπο εκπαίδευσης,
- Ακολούθως οι αποστάσεις ταξινομούνται και επιλέγονται τα k πρότυπα με τις μικρότερες αποστάσεις από το πρότυπο εισόδου (οι λεγόμενοι k πλησιέστεροι γείτονες)
- Το πρότυπο εισόδου ταξινομείται στην κλάση στην οποία ανήκει η πλειοψηφία των k πλησιέστερων γειτόνων, δηλαδή σε αυτή η οποία είναι η πιο κοινή μεταξύ των χαρακτηριστικών που βρίσκονται πιο κοντα σε αυτό.



Σχήμα 4.4: χαρακτηριστικό παράδειγμα ταξινόμησης με την χρήση της μεθόδου k-NN. Το δείγμα δοκιμής (πρασινο) πρόκειται να τοποθετηθεί είτε στην κόκκινη ομάδα είτε στην μπλε. Στην περίπτωση που k=3(συνεχής κύκλος), τότε θα ταξινομηθεί στην κόκκινη, εφόσον υπάρχουν 2 τρίγωνα και 1 τετράγωνο, ενώ εάν k=5(διακεκομμένος κύκλος), τότε θα ταξινομηθεί στην μπλε ομάδα(3 τετράγωνα με μονο 2 τρίγωνα)

Για τον υπολογισμό της απόστασης, μπορεί να χρησιμοποιηθεί οποιοδήποτε μέτρο. Ωστόσο, τα ευρέως χρησιμοποιούμενα μέτρα απόστασης $d(x, y)$ είναι:

Η ευκλείδια απόσταση $d(x,y) = \sqrt{(x-y)^T(x-y)}$

Η απόσταση Minkowski $d(x,y) = \sum_{i=1}^{n}[(|x_i - y_i|^p)^{\frac{1}{p}}]$

Η απόσταση Mahalanobis $d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$

Όπου S= η συνδιασπορά των δεδομένων εκπαίδευσης

Η απόσταση του συνημιτόνου $d(x,y) = 1 - \frac{x^T y}{||x||||y||}$

Η απόσταση συσχέτισης $d(x,y) = 1 - \frac{(x-\bar{x})^T(y-\bar{y})}{||x-\bar{x}||||y-\bar{y}||}$

Η απόσταση Chebyshev $d(x,y) = max_i|x_i - y_i|$

Η καλύτερη επιλογή του k εξαρτάται κάθε φορά, από το σύνολο των δεδομένων που υπάρχουν. Γενικά όμως, όσο μεγαλύτερο είναι τόσο καλύτερη είναι η

ταξινόμηση. Εν γένει, αύξηση του k μέχρι ενός σημείου βελτιώνει την ακρίβεια του αλγορίθμου. Με περαιτέρω αύξηση η απόδοση χειροτερεύει, διότι συμπεριλαμβάνονται ψήφοι αρκετά «μακρινών» γειτόνων. Όπως γίνεται κατανοητό, η παράμετρος k χρειάζεται να μην είναι πολλαπλάσιο του αριθμού των κλάσεων, ούτως ώστε να αποφεύγονται ενδεχόμενες «ισοπαλίες». . Στη περίπτωση των 2 κλάσεων, το k επιλέγεται να είναι περιττός αριθμός, ενώ στη περίπτωση που υπάρχουν λίγα δεδομένα, συχνά επιλέγεται το k=1. Σε αυτή τη περίπτωση ο αλγοριθμος k-NN, λέγεται και nearest neighbor algorithm.

Παρόλα τα πλεονεκτήματα της k-NN, υπαρχουν και σοβαρά μειονεκτήματα. Ο ταξινομητής k-NN υποφέρει από την κατάρα της διαστασιμότητας και την απουσία συστηματικού τρόπου για τον προσδιορισμό του εκάστοτε καταλλήλου μέτρου απόστασης. Όπως έχει ήδη εξηγηθεί, αν η διάσταση των προτύπων είναι μεγάλη, απαιτείται πολύ μεγάλος αριθμός προτύπων εκπαίδευσης, ώστε αυτά να είναι όσο το δυνατόν πυκνότερα στο χώρο για να υπάρξει ικανοποιητικό αποτέλεσμα. Επίσης, δεν έχουν όλες οι διαστάσεις την ίδια διαγνωστική αξία, επομένως αποστάσεις που δίνουν ίδια έμφαση σε όλες τις διαστάσεις, όπως η ευκλείδεια ή η Minkowski ενδεχομένως να μην είναι κατάλληλες για τη μέτρηση αυτής.

4.3.5 PERCEPTRON

Είναι ένας γραμμικός ταξινομητής, κατ' ακρίβεια ένα είδος τεχνητού νευρωνικού δικτύου, που εφευρέθηκε το 1957 στο Αεροναυτικό Εργαστήριο του Κορνελλ, από τον Φρανκ Ροσενβαλτ (Frank Rosenbiatt). Το Perceptron, παρόλο που όταν δημιουργήθηκε ήταν αρκετά υποσχόμενο, εν τέλει αποδείχθηκε ότι δεν μπορούσε να χρησιμοποιηθεί στην ταξινόμηση, παρά μόνο σε προβλήματα 2 κλάσεων.
Συγκεκριμένα, μπορεί να εκτιμήσει τους συντελεστές $(\omega_1, \dots, \omega_N)$, διορθώνοντας τους, εφόσον τα δεδομένα εκπαίδευσης, είναι λανθασμένα ταξινομημένα. Πιο συγκεκριμένα, όσο υπάρχει κάποιο λάθος στην ταξινόμηση των δειγμάτων, το χαρακτηριστικό διάνυσμα, προστίθεται στο κανονικό διάνυσμα, εφόσον είναι θετικό, ενω διαφορετικά αφαιρείται. Δηλαδή εκτελείται το πρόβλημα

$$f(x_1, \dots, x_n) \begin{cases} 1 \ \varepsilon\varphi\acute{o}\sigma o\nu \ \sum_{k=1}^{N} \omega_k X_{\nu(k)} \ + b \ > 0 \\ 0 \qquad\qquad\qquad\qquad \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

όπου η σταθερά b, θεωρείται πάντα θετική.

Στην περίπτωση που το σύνολο εκπαίδευσης είναι γραμμικά διαχωρίσιμο, τότε η διαδικασία, συγκλίνει και μπορεί εύκολα να οριοθετηθεί ο αριθμός των επαναλήψεων που γίνονται. Στην αντίθετη περίπτωση, η διαδικασία αυτή τερματίζεται μετά από ενα πεπερασμένο αριθμό επαναλήψεων.

4.3.6 ADABOOST(ADAPTIVE BOOSTING)

Ένας ευρέως διαδεδομένος ταξινομητής , ο οποίος συχνά αναφέρεται και ως ο καλύτερος μη κλασσικός ταξινομητής, είναι ο AdaBoost, ο οποίος βασίζεται στη μέθοδο Boosting. Η μέθοδος αυτή είναι στην ουσία ένας αλγόριθμος, ο οποίος επιλέγει και συνδυάζει διάφορους ταξινομητές, συνήθως τους πιο αδύνατους(weak) αυτούς δηλαδή με το μεγαλύτερο σφάλμα στη ταξινόμηση, προκειμένου να δημιουργήσει ένα καινούργιο ταξινομητή, δυνατότερο και με μεγαλύτερη ακρίβεια.

Ο AdaBoost, που αναπτύχθηκε απο τους Freund και Schapire το 1996, είναι ο πιο γνωστός αλγόριθμος των μεθόδων boosting. Ο ταξινομητής αυτός λύνει επίσης ικανοποιητικά ένα από τα βασικότερα προβλήματα που αντιμετωπίζουν οι μηχανές εκμάθησης, την κατάρα της διαστασιμότητας, όπου κάθε δείγμα μπορεί να περιέχει ένα τεράστιο αριθμό από υποψήφια χαρακτηριστικά. Ο AdaBoost, σε αντίθεση με άλλους ταξινομητές, όπως τα νευρωνικά δίκτυα και ο SVM, κατά τη διάρκεια της εκπαίδευσης του, επιλέγει μόνο τα χαρακτηριστικά εκείνα, τα οποία είναι γνωστό ότι θα βελτιώσουν την προβλεπτική ικανότητα του μοντέλου, μειώνοντας έτσι τη διάσταση και επιπλέον βελτιώνοντας το χρόνο εκτέλεσης της διαδικασίας, εφόσον τα άσχετα χαρακτηριστικά, δεν χρειάζεται να υπολογιστούν.

Λόγω του ότι η μέθοδος AdaBoost, υποφέρει από φαινόμενα υπερπροσαρμογής (overfitting), στις περιπτώσεις όπου τα δεδομένα που υπάρχουν είναι θορυβώδη, χρησιμοποιείται μια παραλλαγή της μεθόδου, η $\text{AdaBoost}_{\text{reg}}$ (Ratsch 1998) η οποία επανορθώνει τη κλασσική μέθοδο, περιθωριοποιώντας τα δείγματα τα οποία επηρεάζουν σε μεγάλο βαθμό τη διαδικασία εκπαίδευσης.

Τέλος, επισημαίνεται ότι η AdaBoost, δεν συνδυάζεται με καμία μέθοδο επιλογής χαρακτηριστικών, εφόσον, μπορεί από μόνος του να κάνει τόσο την επιλογή των βέλτιστων χαρακτηριστικών, όσο και την εκτίμηση τους, προεκιμένου να τα ταξινομήσει. Αυτό το επιτυγχάνει, απλά συγκρίνοντας τα επιλεγμένα χαρακτηριστικά $X_1, X_2, \dots, X_k$ με κάποιο άλλο κανόνα ταξινόμησης αντί να τα ενσωματώσει με τα βάρη τους $\omega_1, \omega_2, \dots, \omega_k$, όπως αυτά υπολογίστηκαν μέσω της διαδικασίας boosting.

5. ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕΣΩ ΜΕΤΡΩΝ ΠΛΗΡΟΦΟΡΙΑΣ

5.1 Η mRMR ΜΕΘΟΔΟΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Οπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, η διαδικασία της επιλογής χαρακτηριστικών, αποτελεί μια πολυ σημαντική παράμετρο, στα προβλήματα ταξινόμησης. Πολλές μεθόδοι μπορούν να χρησιμοποιηθούν για αυτή τη διαδικασία. Μια από τις πιο δημοφιλείς μεθόδους η οποία χρησιμοποιείται και η οποία βασίζεται πάνω στη αμοιβαία πληροφορία, είναι η μέθοδος mRMR (minimal redundancy-maximal relevance), δηλαδή η μέθοδος της μέγιστης συνάφειας και του ελάχιστου πλεονασμού (Hanchuan Peng, Fuhui Long, and Chris Ding 2005)

Η μέθοδος αυτή, δημιουργήθηκε, προκειμένου να καλυψει το πρόβλημα που υπήρχε όσο αφορά μια άλλη μέθοδο, αυτή του κριτηρίου της μέγιστης εξάρτησης(Max-Dependency), που και αυτή με τη σειρά της βασίζεται στη αμοιβαία πληροφορία. Η μέθοδος αυτή εμφάνιζε ορισμένες δυσκολίες στην άμεση εφαρμογή της. Αυτά ακριβώς τα προβλήματα, ήρθε να καλύψει η mRMR, η οποία είναι ισοδύναμη με αυτή, ίσως και αποδοτικότερη, κάτι που αποδεικνύεται και πειραματικά μέσω της σύγκρισης των 2 διαδικασιών.

5.1.1 ΕΙΣΑΓΩΓΗ

Οπως είναι γνωστό, ο εντοπισμός των σημαντικότερων χαρακτηριστικών από ένα σύνολο δεδομένων, είναι αναγκαίος στη προσπάθεια για ελαχιστοποίηση του σφάλματος στη ταξινόμηση. Δοθέντος ενός προβλήματος, για το οποίο θα πρέπει να ορίστούν τα «βελτιστα» χαρακτηριστικά του, είναι απαραίτητο να οριστεί ένας αλγόριθμος, που να έχει τη δυνατότητα να επιλέξει το καλύτερο υποσύνολο. Η συνθήκη των βέλτιστων χαρακτηριστικών, είναι συχνά ταυτόσιμη με το ελάχιστο σφάλμα στη ταξινόμηση.

Σε καταστάσεις, στις οποίες, οι ταξινομητές δεν είναι συγκεκριμένοι, προκειμένου να επιτευχθεί το ελάχιστο σφάλμα, είναι αναγκαίο να υπάρχει η μέγιστη δυνατή εξάρτηση της κλάσης(target class), με την κατανομή των δεδομένων στο υποσύνολο R^m.

Η πιο δημοφιλής προσέγγιση, προκειμένου να επιτευχθεί η μέγιστη εξάρτηση (Max-Dependency) είναι μέσω της μεθόδου επιλογής χαρακτηριστικών με τη μέγιστη συνάφεια (Max-Relevance). Η συνάφεια συνήθως ορίζεται με όρους αμοιβαίας πληροφορίας ή συσχέτισης, που είναι και οι πιο δημοφιλείς τρόποι, που χρησιμοποιούνται για να καθοριστεί η εξάρτηση στις μεταβλητές.

Στο κριτήριο της μέγιστης συνάφειας, επιλέγονται τα m χαρακτηριστικά Xi, τα οποία έχουν τη μεγαλύτερη αμοιβαία πληροφορία με τη επιλεγμένη κλάση c, κάτι που συνεπάγεται ότι θα έχουν και τη μεγαλύτερη εξάρτηση με τη κλάση αυτή. Συχνά παρατηρείται το φαινόμενο, κατα τη διαδοχική αναζήτηση των χαρακτηριστικών να επιλέγονται τα m κορυφαία χαρακτηριστικά, αυτά δηλαδή που ήρθαν πρώτα κατά την μεμονωμένη αναζήτηση σαν τα m χαρακτηριστικά που βελτιστοποιούν τη ταξινόμηση. Αυτό είναι πρόβλημα, γιατί στη μέθοδο της επιλογής χαρακτηριστικών, είναι ευρέως αποδεκτό, ότι οι συνδυασμοί μεμονωμένων καλών χαρακτηριστικών δεν οδηγούν απαραιτήτως σε μια καλή ταξινόμηση.

Επιπλέον, λόγω του μεγάλου όγκου δεδομένων, που συνήθως επιλέγεται με αυτές τις μεθόδους, κάποιοι ερευνητές έχουν μελετήσει συγκεκριμένους τρόπους, προκειμένου να περιορίσουν την περίσσεια στα χαρακτηριστικά που επιλέγονται και να πετύχουν τον ελάχιστο πλεονασμό(min-Redundancy) σε αυτα. Η περιορισμός των χαρακτηριστικών δεν επιφέρει κανένα απολύτως πρόβλημα στην εξάρτηση τους, αντιθέτως, είναι δυνατό κατά τη διαδοχική αναζήτηση αγαθών, η από κοινού εξάρτηση των χαρακτηριστικών με την κλάση c, να μεγιστοποιείται και αντίστοιχα ο πλεονασμός μεταξύ των χαρακτηριστικών να μειώνεται.

Στόχος αυτού του κεφαλαίου είναι η καταρχήν παρουσίαση του κριτηρίου της μέγιστης εξάρτησης καθώς και της μεθόδου mRMR, μια θεωρητική ανάλυση των δύο και παρουσίαση κάποιων ενδεικτικών αποτελεσμάτων, που προέκυψαν από την πειραματική σύγκριση τους. Επιπλέον, σκοπός είναι να δείχθεί πως είναι εφικτό, η μέθοδος mRMR, να συνδυαστεί και με άλλες μεθόδους επιλογής χαρακτηριστικών σχηματίζοντας ένα αλγόριθμο 2 σταδίων, ο οποίος δίνει ένα συμπαγές υποσύνολο που αποτελείται από τα καλύτερα χαρακτηριστικά με πολύ χαμηλό υπολογιστικό κόστος.

5.1.2 ΤΟ ΚΡΙΤΗΡΙΟ ΤΗΣ ΜΕΓΙΣΤΗΣ ΕΞΑΡΤΗΣΗΣ

Χρησιμοποιώντας όρους από τη θεωρία πληροφορίας, το ιδανικό υποσύνολο χαρακτηριστικών που θα επιλεγεί πρέπει να είναι αυτό που έχει τη μέγιστη αμοιβαία πληροφορία I( ) με την κλάση c που μας ενδιαφέρει, δηλαδή αυτό από το οποίο η κατηγορία έχει τη μεγαλύτερη εξάρτηση. Το υποσύνολο χαρακτηριστικών που πρεπει να επιλεγεί είναι αυτό που μεγιστοποιεί την ποσότητα:

$$\max D(S,c) \qquad D = I(\{x_i, i = 1, \dots, m\}; c)$$

Προφανώς όταν m>1, η αναζήτηση γίνεται προσθέτοντας ένα χαρακτηριστικό κάθε φορά. Δοσμένου δηλαδή ενός συνόλου με m-1 χαρακτηριστικά, το m-χαρακτηριστικό, μπορεί να οριστεί σαν αυτό το οποίο προσφέρει τη μεγαλύτερη αύξηση της $I(S_m; c)$ η οποία ορίζεται ως εξής

$$I(S_m; c) = \int\int p(S_m, c) log \frac{p(S_m, c)}{p(S_m)p(c)} \, dS_m dc$$

$$I(S_m; c) = \int\int p(S_{m-1}, x_m, c) log \frac{p(S_{m-1}, x_m, c)}{p(S_{m-1}, x_m)p(c)} \, dS_{m-1} dx_m dc$$

$$= \int \dots \int p(x_1, x_2, \dots, x_m, c) log \frac{p(x_1, x_2, \dots, x_m, c)}{p(x_1, x_2, \dots, x_m)p(c)} \, dx_1 \dots dx_m dc$$

οπου $S_m = \{x_1, x_2, \dots, x_m\}$: το υποσύνολο χαρακτηριστικών που επιλέγεται

c= η κλαση (target class)

$p(S_m)$= η από κοινού κατανομή των $x_1, x_2, \dots, x_m$

$p(S_m, c)$= η από κοινού κατανομή των $x_1, x_2, \dots, x_m$ $και$ c

Παρά τη μεγάλη θεωρητική αξία που έχει το κριτήριο της μέγιστης εξάρτησης, συνήθως, είναι δύσκολο να προσδιοριστούν με ακρίβεια οι ποσότητες $p(S_m, \mathrm{c})$ $και$ $p(S_m)$ και επομένως η αμοιβαία πληροφορία $I(S_m; c)$. Αυτό οφείλεται στις εξής δυσκολίες που παρουσιάζονται. Καταρχήν, η εκτίμηση της πυκνότητας πολλών μεταβλητών, συχνά απαιτεί και τον υπολογισμό του αντιστρόφου πίνακα συνδιασποράς κάτι που αρκετές φορές μπορεί να μην είναι εφικτό.

Επιπλέον ένα άλλο πρόβλημα, είναι το γεγονός ότι ο αριθμός των δειγμάτων που υπάρχει τις πλείστες φορές είναι σχετικά μικρός. Αυτό συμβαίνει γιατί ο αριθμός παραδειγμάτων που απαιτούνται για την εκτίμηση των από κοινού κατανομών $p(S_m, \mathrm{c})$ $και$ $p(S_m)$ αυξάνεται εκθετικά ως προς των αριθμό των χαρακτηριστικών Sm. Για παράδειγμα ας υποτεθεί ότι ζητείται η εκτίμηση της κατανομής $p(x_k)$ και ότι τα $(x_1, \dots, x_k, \dots, x_m)$ είναι διακριτά χαρακτηριστικά με δυαδικό πεδίο ορισμού. Υπάρχουν 2^m διαφορετικές τιμές που μπορεί να πάρει το x_k, χρειάζεται επομένως ο

υπολογισμός 2^m πιθανοτήτων. Ο αριθμός αυτός θα είναι πολύ μεγαλύτερος από το πλήθος των διαθέσιμων παραδειγμάτων εκτός και αν το m είναι πολύ μικρό.

Τέλος ένα άλλο πρόβλημα που παρουσιάζεται σε αυτή τη μέθοδο είναι ο μεγάλος χρόνος που χρειάζεται προκειμένου να γίνουν οι διάφοροι υπολογισμοί, με αποτέλεσμα να την καθιστά ιδιαίτερη αργή.

Παρόλο που το κριτήριο της μέγιστης εξάρτησης, είναι καλό στις περιπτώσεις όπου απαιτείται η επιλογή λίγων χαρακτηριστικών από ένα μεγάλο σύνολο, δεν είναι κατάλληλο για τις περιπτώσεις όπου χρειάζεται να επιτευχθεί ταξινόμηση με υψηλή ακρίβεια. Για αυτούς ακριβώς τους λόγους μια εναλλακτική πρόταση, είναι ο συνδυασμός του κριτηρίου της μέγιστης συνάφειας (max relevance) και του ελάχιστου πλεονασμού (min redundancy), προκειμένου να κατασκευαστεί μια νέα μέθοδος πιο αποτελεσματική, η mRMR.

Εφόσον η αμοιβαία πληροφορία $I(S_m; c)$ μεταξύ των χαρακτηριστικών και της κλάσης, δεν μπορεί να υπολογιστεί, πρέπει να χρησιμοποιηθεί κάποια άλλη προσέγγιση ώστε να βρεθεί και να επιλεγεί το υποσύνολο από το οποίο η κλάση να έχει τη μεγαλύτερη εξάρτηση. Η μονοπαραγοντική (univariate) προσέγγιση στο πρόβλημα αυτό είναι να υπολογιστεί η εξάρτηση της κλάσης από κάθε χαρακτηριστικό ξεχωριστά και εν συνεχεία να επιλεγούν τα m χαρακτηριστικά από τα οποία υπάρχει η μεγαλύτερη εξάρτηση. Δηλαδή στόχος είναι να βρεθεί ένα υποσύνολο S_p μέσω της σχέσης

$$S_p = \arg\max \{ \sum_{x_i \in S_m} I(x_i; c) \}$$

Το βασικό πρόβλημα της μονοπαραγοντικής αυτής προσέγγισης είναι ότι επιλέγεται μεγάλος αριθμός περιττών χαρακτηριστικών. Αν ένα χαρακτηριστικό επιλέγεται γιατί έχει υψηλή αμοιβαία πληροφορία με την κλάση, τότε χαρακτηριστικά πολύ όμοια με αυτό θα έχουν επίσης υψηλή αμοιβαία πληροφορία με την κλάση και θα επιλεγούν και αυτά. Όμως ένα σύνολο που αποτελείται απο χαρακτηριστικά που έχουν πολλές ομοιότητες μεταξύ τους προσφέρει λίγη μόνο περισσότερη πληροφορία για την κλάση από αυτήν που προσφέρουν μερικά μόνο χαρακτηριστικά του συνόλου. Ακραίο αλλά όχι απίθανο παράδειγμα είναι η επιλογή δύο χαρακτηριστικών που έχουν ίδιες μεταξύ τους τιμές σε κάθε παράδειγμα του συνόλου εκπαίδευσης.

Ίσως να είναι χρήσιμη η επιλογή ενός εκ των δύο αλλά η γνώση του δεύτερου δεν προσφέρει κανένα κέρδος σε πληροφορία. Ένα καλύτερο υποσύνολο

χαρακτηριστικών μπορεί να προκύψει αν επιλέγονται χαρακτηριστικά που έχουν μεν μεγάλη συνάφεια με την κλάση, αλλά την ίδια στιγμή είναι μεταξύ τους όσο το δυνατόν ανόμοια. Δυστυχώς, δεν υπάρχει κάποιος προφανής τρόπος για το πώς μπορεί να μετρηθεί ο βαθμός στον οποίο τα χαρακτηριστικά του υποσυνόλου S είναι μεταξύ τους όμοια. Σαν εναλλακτική λύση χρησιμοποιείται η μέση τιμή της ομοιότητας μεταξύ όλων των πιθανών ζευγών από χαρακτηριστικά του S. Η ποσότητα αυτή θα αναφέρεται στο εξής ως περιττή πληροφορία (redundancy) του S και συμβολίζεται με W(S)

Σε αυτή την ιδέα βασίζεται η μέθοδος mRMR που παρουσιάζεται στην επόμενη ενότητα.

5.1.3 Η ΜΕΘΟΔΟΣ mRMR

Η μέθοδος mRMR, δεν βασίζεται στην επιλογή χαρακτηριστικών με βάση την ανεξαρτησία που υπάρχει μεταξύ τους. Αντιθέτως προσπαθεί να επιλέξει χαρακτηριστικά τα οποία ελαχιστοποιούν τον πλεονασμό και μεγιστοποιούν τη συνάφεια που υπάρχει μεταξύ τους. Άλλωστε, στη πράξη, για πραγματικά δεδομένα, συνήθως ένα σύνολο από χαρακτηριστικά, τελείως ανεξάρτητα μεταξύ τους, δεν οδηγεί και σε τόσο καλά αποτελέσματα. Αντίθετα, είναι δυνατόν, χαρακτηριστικά με από κοινού επίδραση, να οδηγούν σε πολύ καλά αποτελέσματα. Επιπλέον, είναι δυνατόν να μειωθεί άμεσα ο πλεονασμός των χαρακτηριστικών, απλά υπολογίζοντας την αμοιβαία πληροφορία μεταξύ των συνεχών χαρακτηριστικών μεταβλητών.

Συγκεκριμένα, η μέθοδος mRMR εφαρμόζει μια στοιχειώδη επιλογή πρώτης τάξης, προκειμένου να δημιουργήσει ένα υποψήφιο σύνολο χαρακτηριστικών, κάτι που διευκολύνει αφάνταστα την εφαρμογή σε αυτό το σύνολο των μεθόδων wrapper, προκειμένου να καταλήξουν σε συμπαγή υποσύνολα χαρακτηριστικών με μια βέλτιστη ακρίβεια στη ταξινόμηση. Ο mRMR, είναι ιδιαίτερα χρήσιμος στην περίπτωση που έχουμε να κάνουμε με χαρακτηριστικά μεγάλου μεγέθους, είτε όταν έχουμε να αντιμετωπίσουμε προβλήματα επιλογής μεταβλητών στα οποία υπάρχουν χιλιάδες υποψήφια χαρακτηριστικά. Χαρακτηριστικό παράδειγμα τέτοιας περίπτωσης, είναι το πρόβλημα της επιλογής γονιδίων.

Η μέθοδος mRMR θέτει δύο συνθήκες οι οποίες πρέπει να ικανοποιούνται από ένα υποσύνολο χαρακτηριστικών:

i) τα χαρακτηριστικά του υποσυνόλου πρέπει να έχουν όσο το δυνατόν μεγαλύτερη συνάφεια με την κατηγορία(max-relevance)

ii) τα χαρακτηριστικά του υποσυνόλου πρέπει να είναι όσο το δυνατόν ανόμοια μεταξύ τους. (min-redundancy)

Η καλυτερη προσέγγιση της $D(S,c)$, γίνεται μέσω του κριτηρίου της μέγιστης συνάφειας, το οποίο βασίζεται στον υπολογισμό της μέσης τιμής όλων των αμοιβαίων πληροφοριών, μεταξύ των χαρακτηριστικών Xi και της κλάσης c. Δηλαδή η μέγιστη συνάφεια $\max D(S,c)$ υπολογίζεται από το τύπο

$$D(S,c) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i\,;c)$$

Όπου $S = \{x_1, x_2, \dots, x_m\}$: ένα υποσύνολο χαρακτηριστικών

c= η κλάση

Οσο αφορά την ελαχιστοποίηση των όμοιων χαρακτηριστικών, προκειμένου να επιτευχθεί ο ελάχιστος πλεονασμός, μετριέται η ομοιότητα μεταξύ 2 χαρακτηριστικών X_i $και$ X_j χρησιμοποιώντας την αμοιβαία πληροφορία $I(X_i;X_j)$ και ελαχιστοποιώντας την ποσότητα R(S) που δίνεται από τη σχέση

$$\min R(S) \qquad R = \frac{1}{|S|^2}\sum_{x_i x_j \in S} I(X_{i,}X_j)$$

Η μέθοδος mRMR, είναι στην ουσία ο συνδυασμός των 2 πιο πάνω κριτηρίων. Στόχος της είναι η εύρεση ενός υποσυνόλου Φ(D,R) που να ελαχιστοποιεί την περιττή πληροφορία R(S) και ταυτόχρονα να μεγιστοποιεί τη συνάφεια της D(S). Κατά κανόνα η αύξηση της συνάφειας συνοδεύεται από αύξηση της περιττής πληροφορίας και έτσι δεν υπάρχει μοναδικό υποσύνολο που να υπερέχει έναντι των άλλων με βάση και τα δύο κριτήρια. Το Φ(D,R) ορίζεται μέσω του ακόλουθου τύπου

$$\max \Phi(D,R) \qquad \acute{o}\pi o\upsilon \qquad \Phi = D(S,c) - R(S)$$

Προκειμένου η διαδικασία να γίνει υπολογιστικά εφικτή, το υποσύνολο σχηματίζεται χρησιμοποιώντας μία forward μέθοδο άπληστης αναζήτησης. Αρχικά

επιλέγεται το πιο συναφές χαρακτηριστικό (αυτό με τη μεγαλύτερη αμοιβαία πληροφορία με την κλάση). Στη συνέχεια, σε κάθε επανάληψη επιλέγεται ένα χαρακτηριστικό που έχει μεγάλη συνάφεια με την κλάση και ταυτόχρονα μικρή ομοιότητα με τα ήδη επιλεγμένα χαρακτηριστικά. Τροποποιώντας κατάλληλα την πιο πάνω εξίσωση δημιουργείται μία συνάρτηση που αντί να αξιολογεί υποσύνολα χαρακτηριστικών, αξιολογεί τα χαρακτηριστικά μεμονωμένα. Έστω ότι έχουν ήδη επιλεγεί τα m-1 και αναζητείται από το σύνολο { X- S_{m-1}}, το m-οστό χαρακτηριστικό το οποίο μεγιστοποιεί την Φ(.), τότε αυτό μπορεί να βρεθεί μεγιστοποιώντας την ακόλουθη συνάρτηση.

$$\max_{x_i \in X - S_{m-1}} [\, I(x_j\,; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_i; x_j)]$$

Ένας ενδεικτικός αλγόριθμος που να συνοψίζει τη μέθοδο mRMR, όπου στόχος είναι η δημιουργία ενός συνόλου S που να αποτελείται από m βέλτιστα χαρακτηριστικά είναι ο ακόλουθος:

$S = \{\,\}$

$X^* = \max\{\, I(X; c)\}$

$S = S \cup X^*$

$\alpha\nu\ m \geq 2\ \tau\acute{o}\tau\varepsilon\ \mu\acute{\varepsilon}\chi\rho\iota\ m = M\ \tau\acute{o}\tau\varepsilon$

$$X^* = \max_{x_i \in X - S_{m-1}} [\, I(x_j\,; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_i; x_j)]$$

$S = S \cup X^*$

$TELOS$

5.1.4 ΙΣΟΔΥΝΑΜΙΑ mRMR ΚΑΙ ΚΡΙΤΗΡΙΟΥ ΜΕΓΙΣΤΗΣ ΕΞΑΡΤΗΣΗΣ

Οταν επιλέγεται 1 χαρακτηριστικό κάθε φορά, δηλαδή η επιλογή, είναι διαδοχική πρώτης τάξεως (first order), τότε η μέθοδος mRMR, είναι ισοδύναμη με το κριτήριο της μέγιστης εξάρτησης.

Πράγματι, έστω ότι έχει ήδη επιλεγει το σύνολο S_{m-1}, δηλαδή το σύνολο των *m-1* χαρακτηριστικών που είναι πιο σημαντικά και αναζητά κάποιος το m-οστό χαρακτηριστικό από το σύνολο $X - S_{m-1}$.

Εφόσον, η εξάρτηση D αντιπροσοπεύεται από την αμοιβαία πληροφορία, δηλαδή $D = I(S_m; c)$ όπου $S_m = \{ S_{m-1}, x_m \}$ είναι μια πολυμεταβλητή. Από τον ορισμό της αμοιβαίας πληροφορίας

$$I(S_m; c) = H(c) + H(S_m) - H(S_m, c)$$

$$= H(c) + H(S_{m-1}, x_m) - H(S_{m-1}, x_m, c) \quad (1)$$

όπου H(.), είναι η εντροπία των αντίστοιχων μεταβλητών

ορίζεται η ποσότητα $J(S_m) = J(x_1, x_2, \dots, x_m)$ σαν

$$J(x_1, x_2, \dots, x_m) = \int \dots \int p(x_1, x_2, \dots, x_m) log \frac{p(x_1, x_2, \dots, x_m)}{p(x_1)p(x_2), \dots, p(x_m)} dx_1 dx_2, \dots, dx_m) \quad (2)$$

Παρομοίως ορίζεται και η ποσότητα $J(S_m, c) = J(x_1, x_2, \dots, x_m, c)$ ως

$$J(x_1, x_2, \dots, x_m, c) =$$
$$\int \dots \int p(x_1, x_2, \dots, x_m, c) log \frac{p(x_1, x_2, \dots, x_m, c)}{p(x_1)p(x_2), \dots, p(x_m)p(c)} dx_1 dx_2 \dots dx_m dc \quad (3)$$

Από τις (2), (3) προκύπτει άμεσα ότι

$$H(S_{m-1}, x_m) = H(S_m) = \sum_{i=1}^{m} H(x_i) - J(S_m) \quad (4)$$

$$H(S_{m-1}, x_m, c) = H(S_m, c) = H(c) + \sum_{i=1}^{m} H(x_i) - J(S_m, c) \quad (5)$$

Αντικαθιστώντας, αυτούς τους όρους, με τους αντίστοιχους στην εξίσωση (1), προκύπτει ότι

$$J(S_m; c) = J(S_m, c) - J(S_m)$$

$$= J(S_{m-1}, x_m, c) - J(S_{m-1}, x_m) \quad (6)$$

Προφανώς η μέγιστη εξάρτηση, είναι ισοδύναμη με την πιο πάνω εξίσωση, μεγιστοποιώντας το πρώτο όρο και ελαχιστοποιώντας συγχρόνως τον δεύτερο.

ΥΠΟΛΟΓΙΣΜΟΣ ΚΑΤΩΤΑΤΟΥ ΟΡΙΟΥ:

Θεωρώντας ότι ισχύει η ανισότητα $\log(z) \leq z - 1$ η σχέση (2) μπορεί να γραφτεί ως

$$-J(x_1, x_2, \dots, x_m)$$
$$= \int \dots \int p(x_1, x_2, \dots, x_m) log \frac{p(x_1, x_2, \dots, x_m)}{p(x_{1)} p(x_2), \dots, p(x_m)} \, dx_1 dx_2, \dots, dx_m)$$
$$\leq \int \dots \int p(x_1, x_2, \dots, x_m)[\frac{p(x_1), \dots, p(x_m)}{p(x_1, x_2, \dots, x_m)} - 1] \, dx_1 dx_2, \dots, dx_m)$$
$$= \int \dots \int p(x_1) \dots p(x_m) dx_1 dx_2 \dots dx_m - \int \dots \int p(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m$$

$$= 1 - 1 = 0 \qquad (7)$$

Προφανώς και η ελάχιστη τιμή, επιτυγχάνεται, όταν όλες οι μεταβλητές είναι ανεξάρτητες μεταξύ τους, δηλαδή όταν $p(x_1, x_2, \dots, x_m) = p(x_1), \dots, p(x_m)$. Αφού έχουν ήδη επιλεγεί m-1 χαρακτηριστικά, αυτή η συνθήκη ανεξαρτησίας, σημαίνει ότι η αμοιβαία πληροφορία, μεταξύ του x_m και οποιουδήποτε χαρακτηριστικού x_i (ι=1,...,m-1) θα είναι η ελάχιστη. Αυτό είναι και το κριτηριο του Ελάχιστου Πλεονασμου

ΕΥΡΕΣΗ ΑΝΩ ΟΡΙΟΥ ΤΟΥ ΠΡΩΤΟΥ ΟΡΟΥ(UPPER BOUND):

Καταρχήν υπολογίζεται το άνω όριο για την γενική μορφη $J(y_1, y_2, .., y_n)$ σαν

$$J(y_1, y_2, .., y_n) = \dots = \dots \leq \sum_{i=1}^{n-1} H(y_i) \qquad (8)$$

Η (8) μπορεί εύκολα να επεχταθεί σαν

$$J(y_1, y_2, .., y_n) \leq \min\{ \sum_{i=2}^{n} H(y_i), \sum_{i=1, i\neq 2}^{n} H(y_i), \dots, \sum_{i=1, i\neq n-1}^{n} H(y_i) \sum_{i=1}^{n-1} H(y_i)$$

Το μέγιστο για τον πρώτο όρο της εξίσωσης (6), $J(S_{m-1}, x_m, c)$ επιτυγχάνεται όταν όλες οι μεταβλητές είναι στο μέγιστο εξαρτημένες μεταξύ τους. Αυτό σημαίνει ότι η x_m θα έχει τη μέγιστη δυνατή εξάρτηση με την κατηγορία c. Αυτό είναι το κριτηριο της max-relevance

ΔΙΑΦΟΡΕΣ ΤΩΝ 2 ΜΕΘΟΔΩΝ

Παρόλο που μαθηματικά μπορεί κάποιος να καταλήξει μέσω της μιας μεθόδου στη άλλη, εντούτοις αυτές παρουσιάζουν κάποιες διαφορές. Καταρχήν, κάποιος μπορεί να οδηγηθεί στη μέγιστη εξάρτηση μέσω της μεγιστοποίησης του J(Sm,c). Η διαφορά μεταξύ της μεθόδου mRMR και του κριτηρίου της μέγιστης εξάρτησης, είναι ότι ενώ στη μέγιστη εξάρτηση, η συσχέτιση μεταξύ των κατανεμημένων δεδομένων στο υποσύνολο R^m και της κλάσης c, έπαιζε πολύ σημαντικό ρόλο, κάτι τέτοιο δεν συμβαίνει στη μεθοδο mRMR, κάτι που φαίνεται συγκρίνωντας τις εξισώσεις (3) και (10). Επιπλέον, δεν μπορεί να παραγνωριστεί το γεγονός ότι, ενω στη μέγιστη εξάρτηση, υπολογίζονται οι ποσότητες p(X1,…,Xm) και p(X1,….Xm, c), η mRMR αποφεύγει κάτι τέτοιο υπολογίζοντας μόνο τις ποσότητες p(Xi,Xj) και p(Xi,c), κάτι το οποίο είναι προφανώς πολύ πιο εύκολο να γίνει και επιπλέον δίνει πολύ πιο ακριβή αποτελέσματα.

5.1.5 ΑΛΓΟΡΙΘΜΟΙ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ 2 ΣΤΑΔΙΩΝ

Όπως αναφέρθηκε και στην εισαγωγή, ένα από τα κυριότερα προβλήματα που υπάρχει σε ένα πρόβλημα επιλογής του καλύτερου υποσυνόλου, έχει να κάνει με τον καθορισμό του βέλτιστου αριθμού των χαρακτηριστικών που θα πρέπει να περιέχει το τελικό υποσύνολο.

Η μέθοδος mRMR που παρουσίαστηκε πιο πάνω είναι μια μέθοδος στοιχειώδης πρώτης επιλογής χαρακτηριστικών, χωρίς την δυνατότητα να αφαιρεί περιττά χαρακτηριστικά, από αυτά που έχει ήδη επιλέξει. Αυτό όμως που μπορεί να κάνει, είναι να συνδυάζεται μαζί με άλλες μεθόδους επιλογής χαρακτηριστικών, οι οποίες σε δεύτερο στάδιο θα μπορούσαν να κάνουν αυτή τη δουλειά.

Συγκεκριμένα, στο πρώτο στάδιο με τη χρήση της μεθόδου mRMR, καθορίζεται ένα υποψήφιο σύνολο χαρακτηριστικών. Με τη βοήθεια του μοντέλου επαλήθευσης cross-validation, υπολογίζεται το σφάλμα στη ταξινόμηση και στη συνέχεια απομονωνεται μια σταθερή περιοχή Ω στην οποία το σφάλμα είναι μικρό. Ο βέλτιστος αριθμός των χαρακτηριστικών, που θα έχει το υποσύνολο, θα εξαρτάται από αυτή τη περιοχή Ω. Συγκεκριμένα ακολουθούνται τα παρακάτω βήματα:

1. Με τη χρήση της μεθόδου mRMR, επιλέγονται n διαδοχικά χαρακτηριστικά, από ένα σύνολο δεδομένων X. Αυτή η επιλογή, οδηγεί σε n διαδοχικά υποσύνολα χαρακτηριστικών, έστω τα $S_1 \subset S_2 \subset \ldots \subset S_{n-1} \subset S_n$
2. Γίνεται σύγκριση όλων των υποσυνόλων χαρακτηριστικών που επιλέχθησαν μεταξύ τους, με σκοπό τη δημιουργία ενός νέου συνόλου το οποίο θα αποτελείται από αυτά στα οποία το σφάλμα είναι μικρό (δηλ. έχει μικρή μέση τιμή και μικρή διασπορά). Έστω ότι το νέο σύνολο υποσυνόλων το οποίο ονομάζεται Ω, θα αποτελείται από k υποσύνολα.
3. Από τα υποσύνολα που ανήκουν στο Ω, υπολογίζεται αυτό με το μικρότερο σφάλμα ταξινόμησης e_k. Το υποψήφιο σύνολο χαρακτηριστικών που αναζητείται, προκειμένου να θεωρείται βέλτιστο, θα πρέπει να έχει μέγεθος, ίσο με το υποσύνολο k για το οποίο αντιστοιχεί το μικρότερο σφάλμα.

Υπάρχουν αρκετά εξεζητημένες μεθόδοι, οι οποίες θα μπορούσαν να συνδυαστούν με τη μέθοδο mRMR σε δεύτερη φάση, προκειμένου, να δημιουργήσουν συμπαγή σύνολα χαρακτηριστικών, αφαιρώντας χαρακτηριστικά από τα ήδη επιλεγμένα. Συγκεκριμένα, με τη χρήση της διαδικασίας mRMR σε πρώτο στάδιο, δημιουργείται ένα μικρό σύνολο από υποψήφια χαρακτηριστικά, για τα οποία στη συνέχεια εφαρμόζονται σε αυτά άλλες μεθόδοι, οι οποίες συνήθως wrapper, προκειμένου να τελειοποιήσουν το αποτέλεσμα. Δύο από τους κύριους αλγορίθμους που μπορούν να χρησιμοποιηθούν, είναι οι wrapper μεθόδοι της προς τα πίσω επιλογής (backward selection) και της προς τα εμπρος επιλογής(forward selection) που περιγράφηκαν αναλυτικά σε προηγούμενο κεφάλαιο.

Ο λόγος που χρησιμοποιούνται wrapper μεθόδοι σε δεύτερο στάδιο έχει να κάνει με την ίδια την λειτουργία της mRMR μεθόδου. Είναι πολλές φορές πιθανόν, τα διαφορα χαρακτηριστικά που επιλέγονται από τη μέθοδο mRMR, να μην οδηγούν σε τόσο καλά αποτελέσματα, δηλαδή, να μην καταφέρνουν να μειώσουν σημαντικά το σφάλμα στη ταξινόμηση ή ακόμη και να μην μπορούν να το περιορίσουν καθόλου. Υπάρχουν ακόμη και περιπτώσεις όπου η επιλογή ενός χαρακτηριστικού να οδηγεί σε αύξηση του σφάλματος, δηλαδη να υπαρχουν διακυμάνσεις στο σφάλμα όσο αυξάνεται ο αριθμός των χαρακτηριστικών που επιλέγεται. Αυτό μπορεί να οφείλεται σε πολλούς λόγους. Καταρχήν, μια πιθανή αιτία, είναι το γεγονός ότι κάποια από τα πρόσθετα χαρακτηριστικά είναι «noisy», δηλαδή παράγουν θόρυβο. Μια δεύτερη αιτία, αφορά τη μέθοδο «cross-validation» που χρησιμοποιείται και η οποία είναι δυνατό να ευθύνεται και αυτή στις διακυμάνσεις που υπάρχουν πάνω στη καμπύλη του σφάλματος. Τέλος μια άλλη σημαντική αιτία

για αυτό το γεγονός, είναι το γεγονός ότι η μέθοδος mRMR, χρησιμοποιεί τη εξίσωση Φ=D-R, η οποία αφαιρεί τους όρους που επιλέγονται με τη μέγιστη συνάφεια από αυτούς που επιλέγονται από τη ελάχιστο πλεονασμό. Αυτό μπορεί να οδηγήσει σε φαινόμενα κατά τα οποία ένα χαρακτηριστικό το οποίο να έιναι να ουσιαστικά περιττό, να παρουσιάζει ταυτόχρονα μεγάλο ενδιαφέρον όσο αφορά τη συνάφεια του με αποτέλεσμα να επιλέγεται τελικά ανάμεσα στα κορυφαία χαρακτηριστικά. Ακριβώς λόγω του πιο πάνω φαινομένου, για να περιοριστεί το σφάλμα στη ταξινόμηση, χρησιμοποιούνται σε 2^{η} φάση και άλλες μεθόδοι επιλογής χαρακτηριστικών, προκειμένου να αφαιρέσουν όσα χαρακτηριστικά επιλέχθησαν σε πρώτη φάση και τα οποία δεν χρειάζονται είτε γιατί είναι περιττά είτε γιατί δεν προσφέρουν κάτι ουσιαστικό στη μείωση του σφάλματος.

5.1.6 ΤΟ ΚΡΙΤΗΡΙΟ «RM-CHARACTERISTIC»

Μια από τις κύριες μεθόδους που χρησιμοποιείται προκειμένου να συγκριθούν 2 σύνολα μεταξύ τους είναι το κριτήριο RM-characteristic, το οποίο περιγράφεται ως εξής. Δοσμένων 2 συνόλων, έστω Sn^1 και Sn^2, τα οποία έχουν n χαρακτηριστικά το καθένα, το Sn^1 λέγεται ότι είναι πιο «χαρακτηριστικό» από το Sn^2 όταν και μόνο όταν το σφάλμα ταξινόμησης για το πρώτο σύνολο, είναι μικρότερο από αυτό του δεύτερου συνόλου. Ο πιο πάνω ορισμός μπορεί να επεκταθεί και στα υποσύνολα αυτών των συνόλων.

Πιο συγκεκριμένα, έστω ότι εφαρμόζεται μια Α μέθοδος επιλογής χαρακτηριστικών για 2 σύνολα δεδομένων, με αποτέλεσμα να δημιουργηθούν μια σειρά από υποσύνολα για το κάθε ένα από αυτά, δηλαδή Sn^1: $S_1{}^1cS_2{}^1c. . .cS_n{}^1$ και $S_n{}^2$: $S_1{}^2cS_2{}^2c$ … $.S_n{}^2$. Τότε το $S_n{}^1$ είναι RM-characteristic (recursively more) δηλαδή αναδρομικά πιο χαρακτηριστικό από το $S_n{}^2$ αν για κάθε $k\in\Omega$, όπου $\Omega=\{1.. . .,n\}$ το σφάλμα ταξινόμησης είναι πάντα μικρότερο από το αντίστοιχο του $S_n{}^2$.

Εφόσον είναι πολλές φορές ανεπαρκές, να συγκρίνεται αποκλειστικά το σφάλμα ταξινόμησης για συγκεκριμένο αριθμό χαρακτηριστικών 2 συνόλων, προκειμένου να καθοριστεί πιο από αυτά είναι ανώτερο, είναι προτιμότερο να χρησιμοποιείται αυτό το κριτήριο. Με τη σύγκριση των 2 συνόλων χαρακτηριστικών μπορούν να συγκριθούν αποτελεσματικά οι 2 μεθόδοι επιλογής χαρακτηριστικών που τα δημιούργησαν. Για παράδειγμα, έστω ότι από 2 διαφορετικές μεθόδους επιλογής χαρακτηριστικών, τις F^1 και F^2, δημιουργούνται 2 σύνολα, ένα για κάθε μέθοδο. Εάν αποδειχθεί ότι το σύνολο που παράχθηκε από την F^1 είναι RM-characteristic σε σχέση με το σύνολο που δημιουργήθηκε από την F^2, τότε μπορεί να εξαχθεί το συμπέρασμα ότι η 1^{η} μέθοδος επιλογής χαρακτηριστικών είναι καλύτερη από τη δεύτερη.

Σύγκριση μεθόδου mRMR και του κριτηρίου μέγιστης συνάφειας με βάση το RM-characteristic:

Η σύγκριση μπορεί να γίνει με 3 διαφορετικούς τρόπους

1. ΑΜΕΣΗ ΣΥΓΚΡΙΣΗ: Με τη χρήση των 2 μεθόδων επιλέγουμε n διαδοχικά υποσύνολα χαρακτηριστικών έστω S_1C S_2 C,. . . ,S_n. Για αυτά τα υποσύνολα βρίσκουμε το σφάλμα ταξινόμησης . Εαν το σφάλμα είναι μικρότερο για περισσότερα υποσύνολα που δημιουργήθηκαν με την mRMR μέθοδο, τότε αυτή είναι RM-characteristic σε σχέση με τη μέθοδο της μέγιστης συνάφειας και επομένως μπορεί να θεωρηθεί ανώτερη για την διαδοχική επιλογή χαρακτηριστικών.(και αντίστροφα)

2. ΧΡΗΣΗ ΑΛΛΩΝ ΜΕΘΟΔΩΝ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ: Δημιουργούνται όπως και πιο πάνω n διαδοχικά υποσύνολα χαρακτηριστικών έστω S_1C S_2 C,. . . ,S_n. Ακολούθως, εφαρμόζονται σε αυτά τα υποσύνολα η wrapper μέθοδος backward selection προκειμένου να γίνουν κάποιες βελτιώσεις και να απομακρυνθούν κάποια περιττά χαρακτηριστικά. Εν συνεχεία, υπολογίζονται και πάλι τα σφάλματα ταξινόμησης και ανάλογα με το ποια μέθοδος παράγει τα καλύτερα αποτελέσματα σε σχέση με την άλλη θεωρείται «RM-characteristic" και άρα ανώτερη.

3. ΧΡΗΣΗ ΠΕΡΙΣΣΟΤΕΡΩΝ ΜΕΘΟΔΩΝ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΣΕ 2η ΦΑΣΗ: Όπως και πριν, με τη χρήση της mRMR καθώς και του κριτηρίου μέγιστης συνάφειας δημιουργούνται n διαδοχικά υποσύνολα χαρακτηριστικών τα S_1C S_2 C,. . . ,S_n. Σε δεύτερο στάδιο, εφαρμόζονται σε αυτά τα υποσύνολα τόσο η backward όσο και η forward μέθοδοι καθώς και άλλες μεθόδοι wrapper. Μετά την εφαρμογή όλων των μεθόδων αν τα αποτελέσματα συγκλίνουν στο ότι η mRMR είναι RM-characteristic σε σχέση με τη μέθοδο μέγιστης συνάφειας, ή το αντίστροφο, τότε μπορούμε να πούμε σχεδόν με σιγουριά ότι αυτό όντως ισχύει.

Πρακτικά, δεν είναι πάντοτε δυνατό να γίνεται κάθε φορά η σύγκριση των σφαλμάτων ταξινόμησης για 2 σύνολα S_n^1 και S_n^2. Μπορεί όμως να γίνει ένας ικανοποιητικός αριθμός συγκρίσεων ο οποίος δεν θα αφήνει αμφιβολίες για το συμπέρασμα που εξάγεται. Για παράδειγμα, εάν, μετά από τις πιο πάνω συγκρίσεις καταλήξει κάποιος ότι 90% των συγκρίσεων συμφωνούν ότι $e_k^1<e_k^2$ τότε δεν υπάρχει αμφιβολία ότι το S_n^1 είναι RM-characteristic σε σχέση με το S_n^2.

5.1.6 ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Οι συγκρίσεις για τις 2 μεθόδους από τους Hanchuan Peng, Fuhui Long, and Chris Ding , έγιναν τόσο για διακριτά, όσο και για συνεχή συνολα δεδομένων καθώς και για πληθώρα διαφορετικών συνδυασμών. Συγκεκριμένα και εφόσον η μέθοδος mRMR δεν απαιτεί τη χρήση συγκεκριμένων ταξινομητών χρησιμοποιηθήκαν 3 διαφορετικοί ταξινομητές (Naive Bayes, SVM, LDA) καθώς και 4 διαφορετικά σύνολα δεδομένων(HDR, ARR, NCI, LYM) τα οποία παρουσιαζονται αναλυτικά στον πιο κάτω πίνακα.

Data set	HDR MultiFeat		Arrhythmia		NCI		Lymphoma	
Acronym	HDR		ARR		NCI		LYM	
Source	UCI [28], Duin et al [6][14][13]		UCI [28]		Ross et al [26] Scherf et al [27]		Alizadeh et al [1]	
Raw data type	Continuous							
Experimental data type	Discrete				Continuous			
Processing method	Binarize at μ		Discretize at μ±σ to be 3-state		z-score (mean value 0, standard deviation 1)			
# Variable	649		278		9703		4026	
# Sample	2000		420		60		96	
# Class	10		2		9		9	
Class	Name	# Sample	Name	# Sample	Name	# Sample	Name	# Sample
C1	0	200	Normal	237	NSCLC	9	DLBCL	46
C2	1	200	Abnormal	183	Renal	9	CLL	11
C3	2	200			Breast	8	ABB	10
C4	3	200			Melanoma	8	FL	9
C5	4	200			Colon	7	RAT	6
C6	5	200			Leukemia	6	TCL	6
C7	6	200			Ovarian	6	RBB	4
C8	7	200			CNS	5	GCB	2
C9	8	200			Prostate	2	LNT	2
C10	9	200						
Testing method	10-fold CV				LOOCV			

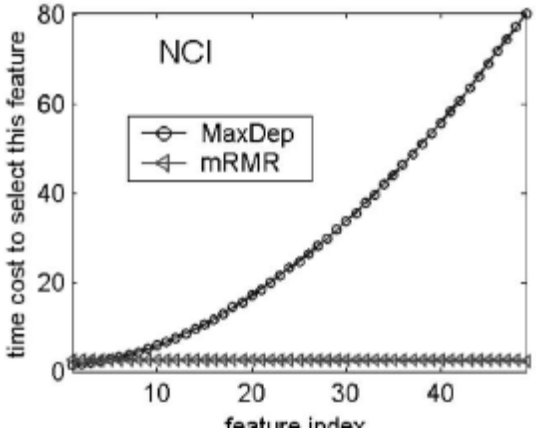
Πίνακας 5.1: Ο πίνακας αναλύει συνοπτικά τα 4 σύνολα δεδομένων που χρησιμοποιούνται στα πειράματα που έγιναν. Τα δεδομένα έχουν συλλεχθεί από κάποιες αληθινές καταστάσεις.

Τα 2 πρώτα σύνολα δεδομένων (HDR, ARR), έχουν υποστεί διακριτικοποίηση ώστε να μπορούν να χρησιμποιηθούν για διακριτές μεταβλητές, ενώ τα 2 τελευταία (NCI, LYM) χρησιμοποιούνται μόνο για συνεχή επιλογή χαρακτηριστικών. Επιπλέον για τα HDR και ARR σύνολα δεδομένων χρησιμοποιείται η cross-validation τεχνική για 10 επαναλήψεις, ενώ για τα NCI και LYM χρησιμοποιείται η τεχνική LOOCV (leave-one-out cross validation). Η ποικιλία αυτή στους τρόπους πειραματικών συγκρίσεων, συμβάλλει στη παρουσίαση μιας ολοκληρωμένης μελέτης και επιπλέον βοηθά στη σιγουριά των αποτελεσμάτων που λαμβάνονται, αφού γίνονται κάτω από πολλές διαφορετικές συνθήκες. Πιο κάτω παρουσιάζονται και σχολιάζονται τα αποτελέσματα από τις συγκρίσεις που έγιναν.

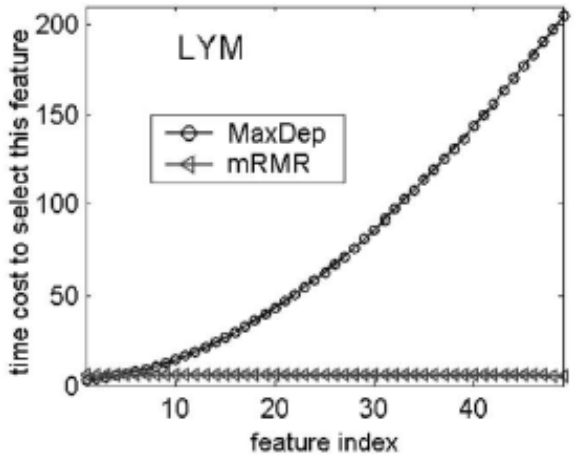
ΣΥΓΚΡΙΣΗ mRMR ΚΑΙ ΜΕΓΙΣΤΗΣ ΕΞΑΡΤΗΣΗΣ:

Η σύγκριση των δυο μεθόδων γίνεται τόσο οσο αφορά την πολυπλοκότητα που απαιτεί ή κάθε μια στην επιλογή των χαρακτηριστικών όσο και στην ακρίβεια που επιτυγχάνεται στην τελική επιλογή των χαρακτηριστικών.

Προκειμένου να υπολογιστεί η πολυπλοκότητα κάθε μεθόδου, υπολογίστηκε ο χρόνος(σε δευτερόλεπτα), που χρειάστηκει η κάθε μέθοδος, προκειμένου να επιλέξει 50 βέλτιστα χαρακτηριστικά, βασισμένη στην εκτίμηση της αμοιβαίας πληροφορίας τους. Οι μετρήσεις έγιναν για τα συνεχή σύνολα δεδομένων NCI και LYM με τα αποτελέσματα να φαίνονται στις πιο κάτω γραφικές παραστάσεις.



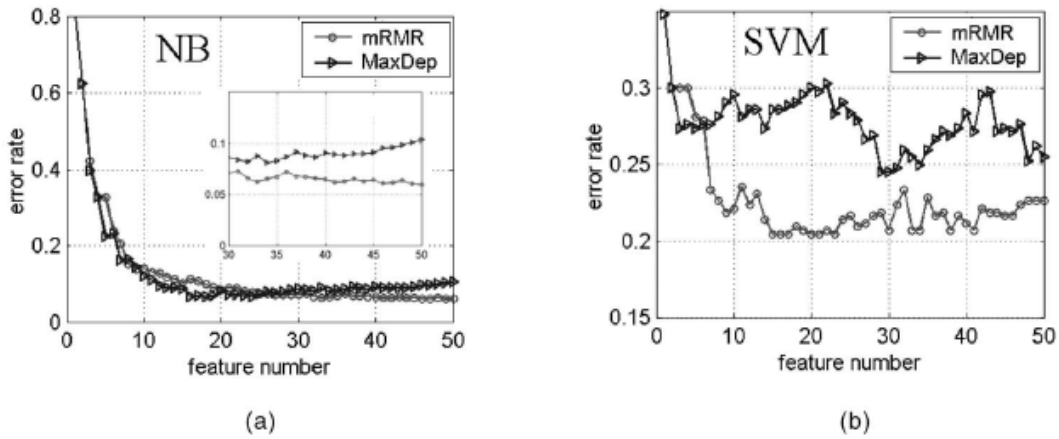
Σχήμα 5.1: Ο χρόνος που χρειάστηκε για να επιλεγούν τα χαρακτηριστικά από το σύνολο δεδομένων NCI για τις 2 μεθόδους



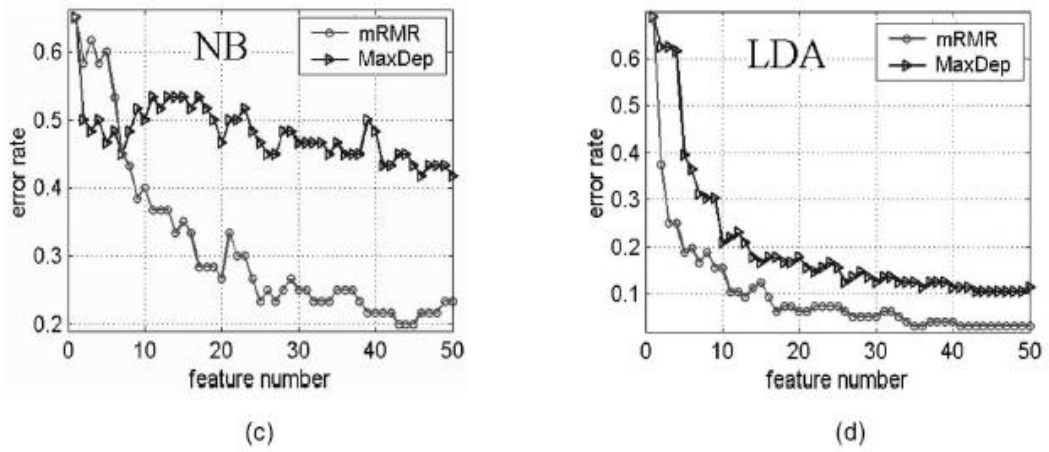
Σχήμα 5.2: Ο χρόνος που χρειάστηκε για να επιλεγούν τα χαρακτηριστικά από το σύνολο δεδομένων LYM για τις 2 μεθόδους

Τα αποτελέσματα, όπως φαίνονται και από τις γραφικές παραστάσεις, δείχνουν ξεκάθαρα ότι η mRMR μέθοδος υπερτερεί αισθητά της μεθόδου μέγιστης εξάρτησης και στα 2 σύνολα δεδομένων, LYM και NCI. Για παράδειγμα, στα NCI δεδομένα, για την επιλογή 20 χαρακτηριστικών η μέθοδος μέγιστης εξάρτησης 20 περίπου δευτερόλεπτα την ίδια στιγμή που η mRMR χρειάζεται λιγότερ από 3. Αντίστοιχα στο σύνολο δεδομένων LYM, ενώ η mRMR δεν χρειάζεται σε καμία στιγμή, περισσότερα από 2 δευτερόλεπτα για να εντοπίσει τα κορυφαία χαρακτηριστικά, η Max-Dep χρειάζεται μέχρι και 200 δευτερόλεπτα μέχρι να βρει τα 50 κορυφαία χαρακτηριστικά.

Όσον αφορά την ακρίβεια στην ταξινόμηση, αυτή εξετάζεται χρησιμοποιόντας τα τεσσερα σύνολα δεδομένων αλλά και τους 3 ταξινομητές. Συγκεκριμένα, μετριέται το σφάλμα ενόσω αυξάνεται ο αριθμός των χαρακτηριστικών.



(a) (b)



Σχήμα 5.3: Σύγκριση της ακρίβειας ταξινόμησης για τις μεθόδους mRMR και MaxDep. (a) Σύγκριση για τα HDR δεδομένα με τη χρήση του naïve Bayes ταξινομητή (b) Σύγκριση για τα ARR δεδομένα με τη χρήση του ταξινομητή SVM (c) Σύγκριση για τα δεδομένα NCI με τη χρήση του naïve Bayes (d) Σύγκριση για τα LYM δεδομένα με τη χρήση του LDA ταξινομητή.

Για τα HDR δεδομένα με τη χρήση του NB ταξινομητή, εξάγονται κάποια πολυ σημαντικά συμπεράσματα. Καταρχήν, η συνολική απόδοση και των 2 μεθόδων είναι σχεδόν παρόμοια. Το σφάλμα πέφτει κατακόρυφα, με την επιλογή των πρώτων 10 χαρακτηριστικών, ενώ στη συνέχεια σχεδόν σταθεροποιείται, σε χαμηλά πάντα πλ αίσια. Η διαφορά των 2 μεθόδων έγκειται στο γεγονός ότι ενώ για την mRMR το σφάλμα συνεχώς μειώνεται μέχρι να σταθεροποιηθεί από ενα σημείο και μετά, αντίθετα στη μέθοδο της μέγιστης εξάρτησης, φαίνεται ξεκάθαρα ότι το σφάλμα ενω συνεχώς μειώνεται, σε κάποια στιγμή, συνήθως μετά την επιλογή ήδη αρκετών χαρακτηριστικών, αυξάνεται και πάλι. Αυτό οδηγεί στο συμπέρασμα, ότι η επιλογή πολλών χαρακτηριστικών με αυτή τη μέθοδο οδηγεί σε κακή ταξινόμηση.

Όσο αφορά τα υπόλοιπα σύνολα δεδομένων, παρατηρείται μια ξεκάθαρη υπεροχή της mRMR μεθόδου. Αρχικά για τα ARR δεδομένα με τον SVM ταξινομητή, το σφάλμα αυξομειώνεται και με τις 2 μεθόδους χωρίς σε καμια περίπτωση να δίνει ικανοποιητικά αποτελέσματα. Εντούτοις, η μέθοδος maxDep, δίνει χαμηλότερο σφάλμα, μονο για την επιλογή των πρώτων 3-5 χαρακτηριστικών αντίθετα με την mRMR που έχει σταθερά μικρότερο σφάλμα στη συνέχεια. Για τα NCI δεδομένα, που ταξινομήθηκαν με την βοηθεια του NB ταξινομητή, όπως και πριν, η maxDep, ενώ ξεκινά καλύτερα έχοντας μικρότερη τιμή στο σφάλμα για τα πρώτα 7 χαρακτηριστικά, στη συνέχεια το σφάλμα όχι μόνο δεν είναι χαμηλότερο από αυτό που δίνει η mRMR μέθοδος, αλλά αντίθετα αυξάνεται φθάνοντας στο σημείο να είναι σχεδόν διπλάσιο από το δικό της.

Τέλος για τα LYM δεδομένα, εδώ υπάρχει μια ξεκάθαρη υπεροχή της mRMR καθόλη τη διάρκεια της επιλογής, φθάνοντας στο σημείο να έχει σχεδόν μηδαμινό σφάλμα όταν επιλέγονται και τα 50 χαρακτηριστικά. Αντίστοιχα με την μέθοδο της μεγιστης εξάρτησης, ενώ παίρνονται ικανοποιητικά αποτελέσματα με το σφάλμα συνεχώς να ελλατώνεται, εντούτοις δεν καταφέρνει σε καμία περίπτωση να έχει καλυτερη απόδοση από την mRMR.

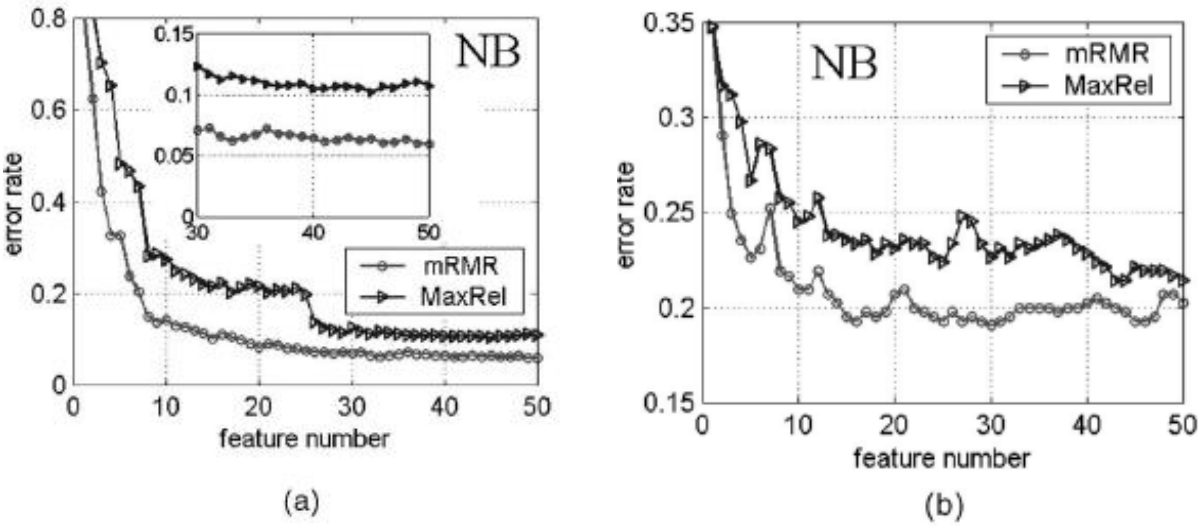
Ο λόγος που η μέθοδος mRMR αποδίδει καλύτερα από την μέθοδο της μέγιστης εξάρτησης, και κυριώς όταν ο αριθμός των χαρακτηριστικών είναι μεγάλος είναι επειδή σε χώρους πολλών διαστάσεων όπως εδώ η εκτίμηση της αμοιβαίας πληροφορίας είναι ελάχιστα αξιόπιστη. Αυτό το φαινόμενο είναι πιο έντονο σε συνεχείς μεταβλητές. Ο λόγος που η διαφορά των 2 μεθόδων, δεν είναι τόσο έντονη στο HDR σύνολο δεδομένων, όσο είναι στα άλλα 3 σύνολα, οφείλεται στο γεγονός ότι το HDR, έχει πολύ μεγαλύτερο αριθμό δειγμάτων εκπαίδευσης, με αποτέλεσμα η ακρίβεια της αμοιβαίας πληροφορίας, να μην μειώνεται τόσο γρήγορα όσο συμβαίνει στα άλλα σύνολα.

Γενικά, είναι εύκολο να διαπιστωθεί ότι είναι προτιμότερο να χρησιμοποιείται η mRMR μέθοδος, παρά η μέθοδος της μεγιστης εξάρτησης αφού δίνει καλύτερα αποτελέσματα, τόσο στη ακρίβεια της ταξινόμησης αλλά και στο χρόνο που απαιτείται προκειμένου να γίνει η επιλογή των χαρακτηριστικών.

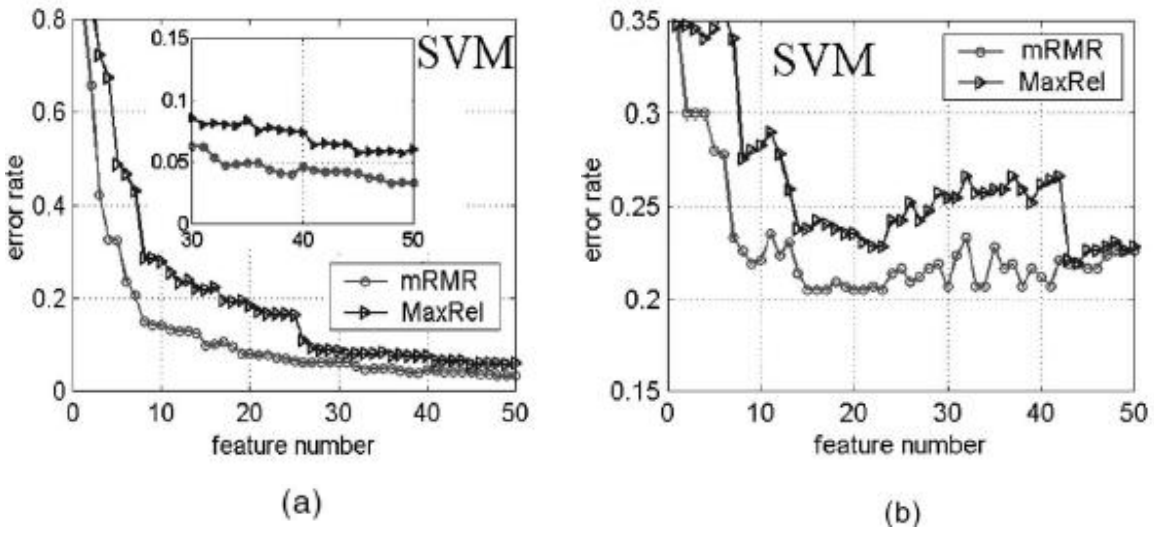
ΣΥΓΚΡΙΣΗ mRMR ΚΑΙ ΜΕΓΙΣΤΗΣ ΣΥΝΑΦΕΙΑΣ:

Λόγω του ότι η mRMR μέθοδος έχει περίπου την ίδια υπολογιστική πολυπλοκότητα με τη μέθοδο της μέγιστης συνάφειας σε αυτή τη σύγκριση, εξετάζεται μόνο η διαφορά που έχουν όσο αφορά τη ακρίβεια στη ταξινόμηση. Οι συγκρίσεις, γίνονται ξεχωριστά για διακριτές μεταβλητές τις οποίες έχουν τα σύνολα HDR, ARR και για συνεχείς, δηλαδή για τα σύνολα δεδομένων LYN ΚΑΙ NCI.

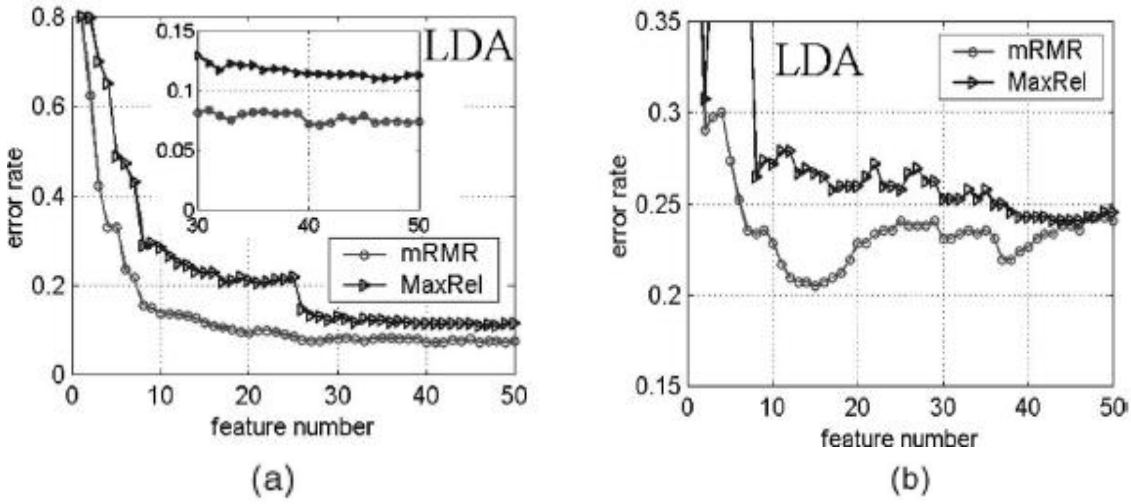
ΣΥΓΚΡΙΣΗ ΓΙΑ ΔΙΑΚΡΙΤΑ ΣΥΝΟΛΑ:



Σχήμα 5.4: Σύγκριση του σφάλματος των μεθόδων mRMR και MaxRel με τον ταξινομητή naïve Bayes (a) χρήση του συνόλου δεδομένων HDR (b) χρήση του συνόλου δεδομένων ARR



Σχήμα 5.5: Σύγκριση του σφάλματος των μεθόδων mRMR και MaxRel με τον ταξινομητή SVM (a) χρήση του συνόλου δεδομένων HDR (b) χρήση του συνόλου δεδομένων ARR



Σχήμα 5.6: Σύγκριση του σφάλματος των μεθόδων mRMR και MaxRel με τον ταξινομητή LDA (a) χρήση του συνόλου δεδομένων HDR (b) χρήση του συνόλου δεδομένων ARR

Για τα HDR δεδομένα, τόσο και για τους 3 ταξινομητές(NB, SVM, LDA) που χρησιμοποιήθηκαν, προκειμένου να συγκρίθούν οι μεθόδοι, φαίνεται ξεκάθαρα, ότι η mRMR υπερισχύει κατά κράτος της μεθόδου μέγιστης συνάφειας και στις 3 περιπτώσεις. Μπορεί επομένως να ειπωθεί ότι η mRMR είναι RM-characteristic σε σχέση με την max-Relevance μέθοδο. Γενικά όμως και οι 2 μεθόδοι τείνουν να δίνουν πολύ καλά αποτελέσματα όσο αφορά το σφάλμα στη ταξινόμηση. Συγκεκριμένα και για την επιλογη των 50 χαρακτηριστικών το σφάλμα της mRMR είναι 6%, 3.5% και 7%, ενώ της μεθόδου της μέγιστης συνάφειας 10%, 5.5% και 11% για τους 3 ταξινομητές NB, SVM και LDA αντίστοιχα.

Όσο αφορά τα ARR δεδομένα, και εδώ τα χαρακτηριστικά που επιλέχθηκαν μέσω της mRMR μεθόδου υπερέχουν αισθητά αυτών της max-Relevance, μέχρι και τη στιγμή που αρχίζουν να πλησιάζουν τα 50. Εκεί, οι δυο μεθόδοι παρουσιάζουν περίπου το ίδιο σφάλμα στη ταξινόμηση και για τους 3 ταξινομητές που χρησιμοποιούνται. Παρολα αυτά, ανεξαρτήτως αυτού η mRMR αποδίδει και εδώ καλύτερα, αφού 15 mRMR χαρακτηριστικά, δίνουν καλύτερη ακρίβεια στην ταξινόμηση παρά 50 Max-Rel χαρακτηριστικά.

ΣΥΓΚΡΙΣΗ ΓΙΑ ΣΥΝΕΧΗ ΣΥΝΟΛΑ:

Classifier	m / Method	1	5	10	15	20	25	30	35	40	45	50
NB	MaxRel	65.00	51.67	51.67	45.00	46.67	43.33	41.67	38.33	36.67	33.33	36.67
	mRMR	65.00	60.00	40.00	35.00	26.67	23.33	25.00	25.00	21.67	20.00	23.33
SVM	MaxRel	98.33	46.67	55.00	50.00	45.00	55.00	41.67	35.00	38.33	35.00	36.67
	mRMR	98.33	70.00	58.33	48.33	40.00	31.67	31.67	31.67	26.67	23.33	23.33
LDA	MaxRel	73.33	60.00	60.00	50.00	46.67	46.67	41.67	36.67	38.33	41.67	40.00
	mRMR	73.33	66.67	50.00	53.33	45.00	33.33	35.00	35.00	33.33	30.00	30.00

Πίνακας 5.2: Σύγκριση της ακρίβειας ταξινόμησης για τις μεθόδους mRMR και MaxRel με τη χρήση του συνόλου δεδομένων NCI

Για το NCI σύνολο δεδομένων, η μέθοδος mRMR αποδίδει και πάλι καλύτερα από τη μέθοδο της μεγιστης συνάφειας και για τους 3 ταξινομητές. Με εξαίρεση κάποιες περιπτώσεις, που η maxRel έχει μικρότερο σφάλμα στη ταξινόμηση, κυριώς όταν έχουν επιλεγεί κάτω από 15 χαρακτηριστικά, γενικά όσο τα χαρακτηριστικά αυξάνονται, το σφάλμα είναι παντού μικρότερο για τα χαρακτηριστικά που επιλεχθησαν από τη mRMR.

Classifier	m / Method	1	5	10	15	20	25	30	35	40	45	50
NB	MaxRel	72.92	25.00	15.63	13.54	13.54	12.50	13.54	12.50	11.46	11.46	10.42
	mRMR	72.92	17.71	16.67	10.42	11.46	9.38	10.42	9.38	9.38	7.29	8.33
SVM	MaxRel	42.71	27.08	21.88	21.88	18.75	16.67	14.58	14.58	15.63	11.46	12.50
	mRMR	42.71	11.46	10.42	7.29	5.21	7.29	7.29	5.21	5.21	5.21	4.17
LDA	MaxRel	68.75	32.29	22.92	23.96	23.96	21.88	22.92	17.71	16.67	16.67	15.63
	mRMR	68.75	18.75	15.63	12.50	6.25	7.29	5.21	3.13	4.17	3.13	3.13

Πίνακας 5.3: Σύγκριση της ακρίβειας ταξινόμησης για τις μεθόδους mRMR και MaxRel με τη χρήση του συνόλου δεδομένων LYM

Για τα LYM δεδομένα, η διαφορά μεταξύ των 2 μεθόδων είναι πολύ μεγάλη σε όλες τις περιπτώσεις. Ενδεικτικό αυτού, το σφάλμα για τη μέθοδο μέγιστης συνάφειας όταν επιλέγονται και τα 50 χαρακτηριστικά για το SVM ταξινόμητη, ειναι το τριπλάσιο σε σχέση με αυτό που δίνει η mRMR. Γενικά εύκολα διαπιστώνεται ότι στις περιπτώσεις των συνεχών χαρακτηριστικών έιναι και πάλι προτιμότερο να επιλέγονται αυτά με τη χρήση της μεθόδου mRMR.

ΣΥΓΚΡΙΣΗ ΣΥΜΠΑΓΩΝ ΥΠΟΣΥΝΟΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ(mRMR vs MaxRel):

Μια αλλη ιδιαίτερα ενδιαφέρουσα σύγκριση που μπορεί να γίνει, είναι πάνω στα χαρακτηριστικά που προκύπτουν από τη mRMR μέθοδο και τη μέθοδο της μέγιστης συνάφειας, μετά και την εφαρμογή σε αυτά των wrapper μεθόδων της προς τα εμπρος(forward) και της προς πίσω(backwrard) επιλογής χαρακτηριστικών. Είναι γνωστό ότι με αυτές τις μεθόδους αφαιρούνται κάποια επιπλέον «κακά» ή περιττά χαρακτηριστικά, επομένως προκύπτει ένα βέλτιστο υποσύνολο από το υποψήφιο σύνολο χαρακτηριστικών που αρχικά επιλέχθηκε, από το οποίο μπορούν να εξαχθούν χρήσιμα συμπεράσματα.

Συγκεκριμένα, σε περίπτωση που η mRMR μέθοδος είναι RM-characteristic σε σχέση με τη μέθοδο μέγιστης συνάφειας (ή το αντίθετο) τότε είναι λογικό ότι από αυτή θα πάρουμε ένα πιο καλό υποσύνολο χαρακτηριστικών , μετά την εφαρμογή των wrapper μεθόδων πάνω στα σύνολα χαρακτηριστικών που επιλέχθηκαν σε πρώτη φάση με τις 2 αυτές μεθόδους, δηλαδή ένα υποσύνολο με πιο μικρό σφάλμα ταξινόμησης.

Όπως φάνηκε και στις προηγούμενες συγκρίσεις, τα πρώτα 50 χαρακτηριστικά που επιλέχθηκαν μέσω των 2 αυτών μεθόδων οδηγούν σε καλά αποτελέσματα, όσο αφορα το σφάλμα και στα 4 σύνολα δεδομένων που έχουμε. Εφαρμόζονται σε αυτά και οι 2 wrapper μεθόδοι, με τη χρήση και των 3 διαθέσιμων ταξινομητών (NB, SVM,LDA).

Τα αποτελέσματα από αυτές τις συγκρίσεις, φαίνονται στο πιο κάτω πίνακα.

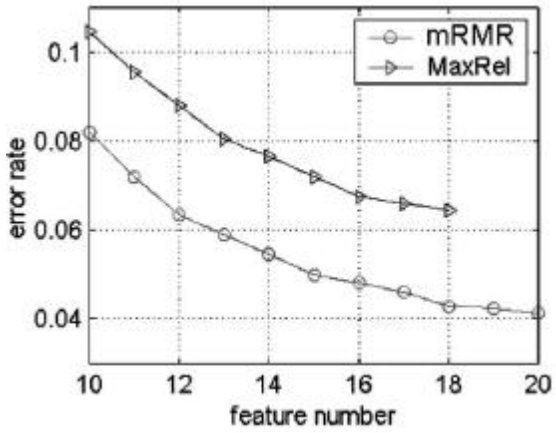
Data set	Wrapper	NB		SVM		LDA	
		MaxRel	mRMR	MaxRel	mRMR	MaxRel	mRMR
HDR	Forward	6.45	3.20	5.50	3.45	6.80	4.05
	Backward	5.95	3.10	4.55	2.85	6.90	4.00
ARR	Forward	18.81	17.86	20.95	19.29	19.76	18.10
	Backward	23.10	17.86	20.48	19.52	20.48	18.33
NCI	Forward	26.67	13.33	25.00	21.67	31.67	33.33
	Backward	20.00	15.00	18.33	13.33	30.00	25.00
LYM	Forward	6.25	5.21	6.25	2.08	6.25	2.08
	Backward	5.21	3.13	3.13	3.13	6.25	3.13

Πίνακας 5.4: Σύγκριση τους σφάλματος ταξινόμησης(αναφορά του μικρότερου σφάλματος που προκύπτει) για τις μεθόδους mRMR και MaxRel με τη χρήση των wrapper μεθόδων forward και backward selection για τα 4 σύνολα δεδομένων

Πιο συγκεκριμένα, εύκολα μπορεί να διαπιστώσει κανείς, πως το σφάλμα το χαρακτηριστικών που επιλέχθηκαν αρχικά μέσω της μεθόδου mRMR είναι πάντα μικρότερο από αυτό των χαρακτηριστικών που επιλέχθησαν με τη μέθοδο maxRel, και για τους 3 ταξινομητές και για τα 4 σύνολα δεδομένων, και για τις 2 wrapper μεθόδους με ελάχιστες μόνο εξαιρέσεις.

Πιο συγκεκριμένα, μόνο για το NCI σύνολο δεδομένων, με τη forward μέθοδο στο οποίο χρησιμοποιείται ο LDA ταξινόμητης, η μέθοδος μέγιστης συνάφειας υπερτερεί της mRMR, με το σφάλμα των 2 να είναι 31.67% και 33.33% αντίστοιχα.

Μια πιο αναλυτική σύγκριση φαίνεται στη τη πιο κάτω γραφική παράσταση. Εδώ έγινε μια πρώτη επιλογή από το HDR πακέτο δεδομένων, με τις 2 μεθόδους, χρησιμοποιώντας τον NB ταξινομητη χαρακτηριστικών. Ακολούθως χρησιμοποιήθηκε σε αυτά τα σύνολα η forward selection, wrapper μεθοδος. Όπως φαίνεται ξεκάθαρα, το σφάλμα για το υποσύνολο των mRMR-χαρακτηριστικών, είναι παντού μικρότερο από αυτό των maxRel-χαρακτηριστικών. Μπορεί επομένως να ειπωθεί με βεβαιότητα, ότι τα υποψήφια χαρακτηριστικά που επιλέγονται από τη mRMR μπορούν να παράξουν υποσύνολα στα οποία γίνεται πιο εύκολα η ταξινόμηση.



Σχήμα 5.7: Γραφική παράσταση για τις μεθόδους mRMR και MaxRel για το σύνολο δεδομένων HDR με τη χρήση του naïve Bayes ταξινομητή

Εν κατακλείδι, είναι φανερό, από όλες τις συγκρίσεις που έγιναν, ότι η mRMR μέθοδος είναι καλύτερη και αποδοτικότερη από την maxDep, καθώς και την maxRel, δίνοντας μικρότερο σφάλμα ταξινόμησης. Επομένως, αποτελεί μια σίγουρη λύση στην επιλογή χαρακτηριστικών.

5.2 Η CMIM ΜΕΘΟΔΟΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Σε αυτό το κεφάλαιο, αναλύεται και εξετάζεται μια μέθοδος επιλογής χαρακτηριστικών η οποία βασίζεται στην υπο συνθήκη αμοιβαία πληροφορία και η οποία προτάθηκε από τον Francois Fleuret το 2003.

Η μέθοδος αυτή αναπτύχθηκε προκειμένου να βελτιώσει τις μεθόδους επιλογής χαρακτηριστικών, που είχαν ήδη αναπτυχθεί τα προηγούμενα χρόνια. Συγκεκριμένα, ήδη υπάρχουσες μεθόδοι, όπως η βαθμολογία Φισερ(fisher score by Furey 2000), το τεστ Kolmogorov-Smitnov καθώς και η συσχέτιση Pearson(Miyahara και Pazzani 2000), δεν μπορούσαν να εξασφαλίσουν ότι τα χαρακτηριστικά τα οποία που επιλέγονταν θα είχαν όσο το δυνατόν, λιγοτερη εξάρτηση μεταξύ τους, με αποτέλεσμα να παρουσιάζονται φαινόμενα πλεονασμού με την επιλογή περιττών και άχρηστων χαρακτηριστικών.

Ακόμη και το κριτήριο που προτάθηκε το 2003 από τους Katanamahatana και Gunopulos, το οποίο είναι βασισμένο στα δέντρα αποφάσεων, παρόλο που έδινε ικανοποιητικές λύσεις στο πιο πάνω πρόβλημα, εντούτοις οδηγούσε σε μεγάλα υπολογιστικά κόστη καθώς και σε φαινόμενα υπερπροσαρμογής (overfitting) των δεδομένων.

Η μέθοδος του Fleuret, βασίζεται στην επιλογή χαρακτηριστικών τα οποία έχουν τη μέγιστη δυνατή πληροφορία με την προβλεπόμενη κλάση και τα οποία ακολούθως συγκρίνονται με τα ήδη επιλεγμένα χαρακτηριστικά, εξασφαλίζοντας έτσι ότι η τελική επιλογή θα αποτελείται απο χαρακτηριστικά ανα 2 ανεξάρτητα όπου το κάθε ένα από αυτά τα παρέχει μεγάλα ποσά πληροφόρίας. Η διαδικασία αυτή, αποφεύγει την επιλογή χαρακτηριστικών τα οποία είναι όμοια με τα ήδη επιλεγμένα, ακόμη και αν θεωρούνται μεμονωμένα ισχυρά, εφόσον δεν μεταφέρουν καθόλου επιπλέον πληροφορία στην κλάση. Με λίγα λόγια, η CMIM, εξασφαλίζει μια καλή λύση στο δίλημμα μεταξύ ανεξαρτησίας και διακρισης των χαρακτηριστικών.

Μια παρόμοια μέθοδος με την CMIM, είναι αυτή της γρήγορης συσχέτισης βασισμένης στις μεθόδους φίλτρου(Yu και Liu, 2003). Αυτή η μέθοδος επιλέγει χαρακτηριστικά τα οποία έχουν υψηλή συσχέτιση με την κλάση, προκειμένου να διαπιστωθεί εαν είναι λιγότερο ή περισσότερο σχετικά με τα ήδη επιλεγμένα χαρακτηριστικά. Παρόλο που οι 2 μεθόδοι μοιάζουν μεταξύ τους, εντούτοις, η δεύτερη, δεν βασίζεται σε μια μοναδική διαδικασία προκειμένου να πετυγχαίνει κάθε φορά, τόσο την ανεξαρτησία μεταξύ των χαρακτηριστικών, όσο και ψηλή πληροφόρηση για την κλάση, με αποτέλεσμα, να παρασύρεται σε καταστάσεις στις οποίες η εξάρτηση μεταξύ των χαρακτηριστικών και της κλάσης εμφανίζεται υπό όρους.

Για την καλύτερη κατανόηση, από εδώ και πέρα υιοθετούνται οι ακόλουθοι συμβολισμοί για τα δεδομένα μας

Y= συμβολίζεται μια δυαδική τυχαία μεταβλητή που αντιπροσωπεύει την κλάση που εξετάζεται

$\{X_1, X_2, \dots, X_N\}$ = το σύνολο όλων των διαθέσιμων δυαδικών χαρακτηριστικών

$\{X_{N(1)}X_{N(2)}, \dots, X_{N(K)}\}$ = το σύνολο των K χαρακτηριστικών που επιλέγονται κατά τη διαδικασία της επιλογής χαρακτηριστικών, όπου K<<N

$F = \{(x^{(1)}, y^{(1)}), \dots., (x^{(T)}, y^{(T)})$ = το σύνολο των εκπαιδευτικών χαρακτηριστικών, μαζί με την κλάση στην οποία ανήκει το κάθε ένα από αυτά.

5.2.1 ΜΕΓΙΣΤΟΠΟΙΗΣΗ ΤΗΣ ΥΠΟ ΣΥΝΘΗΚΗΣ ΑΜΟΙΒΑΙΑΣ ΠΛΗΡΟΦΟΡΙΑΣ

Όπως είναι γνωστό, ο κύριος στόχος της επιλογής χαρακτηριστικών, είναι η απομόνωση ενός υποσυνόλου χαρακτηριστικών, τα οποία θα φέρουν τη μεγαλύτερη δυνατή πληροφορηση για την κλάση που μας ενδιαφέρει. Η βέλτιστη λύση, θα ήταν η επιλογή ενός συνόλου από K-χαρακτηριστικά $\{X_1, X_2, \dots, X_K\}$, από το σύνολο $\{X_i : i = 1, \dots, N\}$, όπου N>>K, τα οποία θα ελαχιστοποιούν την αβεβαιότητα για την κλάση Y που μας ενδιαφέρει, δηλαδή θα ελαχιστοποιούν την ποσότητα $H(Y|X_1, \dots, X_K)$.

Φυσικά κάτι τέτοιο δεν είναι καθόλου ρεαλιστικό αφού για να καταστεί εφικτό θα πρέπει να γίνουν 2^{K+1} υπολογισμοί πιθανοτήτων, προκειμένου να βρεθούν τα βέλτιστα χαρακτηριστικά. Ακόμη και αν βρισκόταν ένας τρόπος, κατι τέτοιο να συμβεί, αυτή η εκτίμηση θα ήταν υπολογιστικά πολύ χρονοβόρα και ανούσια να γίνει.

Μια διαφορετική προσέγγιση, θα ήταν να γίνει μια τετριμμένη τυχαία επιλογή η οποία θα εξασφαλίζει σε ένα μεγάλο βαθμό, την ανεξαρτησία μεταξύ των χαρακτηριστικών, μέσω της ισοδύναμης επιλογής διαφορετικών τύπων από αυτά. Εντούτοις κάτι τέτοιο, δεν μπορεί εξ ορισμού να οδηγήσει σε υψηλές αποδόσεις όσο αφορά την προγνωστική ικανότητα της μεθόδου αυτής. Η μεγαλύτερη αδυναμία, εντοπίζεται στο γεγονός ότι παρόλο που τα χαρακτηριστικά που θα επιλέγονται θα φέρουν μεγάλες ποσότητες πληροφορίας για την κλάση, εντούτοις θα αυξάνουν και σε μεγάλο βαθμό τον πλεονασμό αυτής. Για παράδειγμα, ναι μεν μπορεί να επιλέγονται χαρακτηριστικά διαφόρων τύπων, κανείς όμως δεν μπορεί να διασφαλίσει ότι δεν υπάρχει η πιθανότητα πολλά παρόμοια χαρακτηριστικά με μεγάλη προγνωστική αξία(δηλαδη που φέρουν πολλή πληροφορία), να είναι όλα της ίδια μορφής, να ανήκουν δηλαδή στην ίδια οικογένεια.

Ο αλγόριθμος που θα προταθεί παρακάτω, βασίζεται σε μια ενδιάμεση λύση. Συγκεκριμένα, η προσέγγιση που γίνεται, στηρίζεται σε μια λύση ισορροπίας για το δίλημμα μεταξύ της ανεξαρτησίας και της μεμονωμένης δύναμης των χαρακτηριστικών, επιτυγχάνοντας βέλτιστα αποτελέσματα με το να συγκρίνει κάθε χαρακτηριστικό με αυτά τα οποία έχουν ήδη επιλεγεί. Ένα χαρακτηριστικό Xr, θεωρείται καλό μόνο όταν η υπο συνθήκη αμοιβαία πληροφορία(CMI) σε σχέση με την κλάση Y , $I(Y, X_r|X)$, είναι σχετικά μεγάλη για κάθε γνωστό χαρακτηριστικό X, το οποίο έχει ήδη επιλεγεί. Δηλαδή το X_r, θεωρείται καλό μόνο εφόσον φέρει ικανοποιητική ποσότητα πληροφορίας για την κλάση Y και ταυτόχρονα την πληροφορία αυτή δεν την κουβαλά κανένα από τα ήδη επιλεγμένα χαρακτηριστικά. Αυτή η διαδικασία, μπορεί να συνοψιστεί μεσω του ακόλουθου αλγόριθμου.

$$N(1) = arg\max_n I(Y; X_n) \quad (1)$$

$\forall k, \quad 1 \leq k \leq K$

$$N(k+1) = arg\max_n\{\min_{d \leq k} I(Y; X_n|X_{N(d)}) = A(n,k) \quad (2)$$

Η τιμή της $I(Y, X_n|X_{N(d)})$, είναι χαμηλή, μόνο εαν η X_n, δεν φέρει επαρκή ποσότητα πληροφορίας για τη Y, είτε όταν αυτή η πληροφορία υπάρχει ήδη στα $X_{N(d)}$ χαρακτηριστικά που έχουν επιλεγεί, ή τουλάχιστον σε ένα από αυτά.

Επιλέγοντας μεσω της 2ης εξίσωσης το χαρακτηριστικό με τη μέγιστη $A(n,k)$, εξασφαλίζεται ότι το νέο χαρακτηριστικό θα παρέχει σημαντική πληροφορία για την κλάση και ταυτόχρονα θα είναι ανεξάρτητο σε σχέση με τα προηγούμενα που διαλέχτηκαν. Αυτός ο υπολογισμός μπορεί να γίνει με ακρίβεια για κάθε χαρακτηριστικό μέσω της εκτίμησης της κατανομής δυαδικών μεταβλητών και επιπλέον, παρά το φαινομενικά υψηλό υπολογιστικό κόστος που φαίνεται να έχει η όλη διαδικασία, αυτή μπορεί να εφαρμοστεί με πολύ αποτελεσματικό τρόπο.

Η μέθοδος αυτή είναι βασισμένη στο κριτήριο των Koller και Sahami, που ανέπτυξαν το 1996, το οποίο αναφέρεται στη δυνατότητα χρήσης του στρώματος Markov, προκειμένου να βοηθήσει στον εντοπισμό των χαρακτηριστικών τα οποία μπορούν να αφαιρεθούν χωρίς να βλάψουν σε καμία περίπτωση την απόδοση της διαδικασίας της ταξινόμησης. Το κριτήριο λέει, ότι μια υποομάδα απο M χαρακτηριστικά, αποτελεί ένα στρώμα για ένα τυχαίο χαρακτηριστικό X_i αν και μόνο αν το χαρακτηριστικό αυτό, είναι υπό συνθήκη ανεξάρτητο από τα άλλα χαρακτηρστικά και την κλάση που πρόκειται να κάνει πρόβλεψη, δοσμένου του M.

Η CMIM, μέθοδος στηρίζεται σε μια προσέγγιση αυτού του κριτηρίου. Συγκεκριμένα θεωρεί ότι οι οικογένειες των M χαρακτηριστικών είναι στην ουσία ένα μοναδικό χαρακτηριστικό ήδη επιλεγμένο. Έτσι ένα χαρακτηριστικό X, μπορεί να απορριφθεί,

εαν υπάρχει ήδη επιλεγμένο κάποιο άλλο χαρακτηριστικό X_v, έτσι ώστε τα Χ και Υ, να είναι υπο συνθήκη ανεξάρτητα δεδομένου του X_v. Η πιο πάνω έννοια μπορεί να εκφραστεί με μαθηματικά σύμβολα ως ακολούθως

$$\exists k, \tau\varepsilon\tau o\iota o\ \omega\sigma\tau\varepsilon\ \ I\left(Y; X\middle|X_{N(k)}\right) = 0 \qquad \text{ή και σαν}$$

$$\min_k I(Y; X|\, X_{N(k)}) = 0$$

αφού η αμοιβαία πληροφορία είναι θετική.

5.2.2 Η ΜΕΘΟΔΟΣ CMIM

Προκειμένου να είναι σαφής ο αλγόριθμος για την μέθοδο CMIM, ορίζεται ένας απλός και αποτελεσματικός τρόπος υπολογισμού των ποσοτήτων $I(Y; X_n|X_m)$ και $I(Y; X_n)$. Όπως είναι γνωστό, από προηγούμενο κεφάλαιο, η αμοιβαία πληροφορία μεταξύ της κλάσης Υ και μιας μεταβλητής X_n ορίζεται σαν

$$I(Y; X_n) = H(Y) + H(X_n) - H(Y, X_n)$$

ενώ η υπο συνθήκη αμοιβαία πληροφορία μεταξύ αυτών, όταν είναι γνωστή μια άλλη μεταβλητή έστω η X_m, ορίζεται σαν

$$I(Y; X_n|X_m) = H(Y|X_m) - H(Y|X_n, X_m)$$

$$= H(Y, X_m) - H(X_m) - H(Y, X_n, X_m) + H(X_n, X_m)$$

Οι σχέσεις $H(Y)$, $H(Y, X_n)$ και $H(Y, X_n, X_m)$ μπορούν να εκτιμηθούν μέσω του συνόλου των δεδομένων εκπαίδευσης, με προσθαφαίρεση των εκτιμήσεων της εντροπίας για οικογένειες αποτελούμενες από 1-3 μεταβλητές. Έστω τώρα ότι τα x, y, z είναι 3 δυαδικά διανύσματα και οι u, v, w 3 δυαδικές τιμές. Βάση αυτών και λαμβάνοντας υπόψη ότι $x^{(t)} \in \{0,1\}^T$ είναι ένα δυαδικό δυάνυσμα χαρακτηριστικών που αντιστοιχεί στο t-οστό δείγμα των εκπαιδευτικών δεδομένων ενώ το $x_n{}^{(t)} \in \{0,1\}^T$, είναι το n-οστό χαρακτηριστικό του t δείγματος των εκπαιδευτικών δεδομένων ορίζονται οι ακόλουθες σχέσεις

$$n_u(x) = ||\{t: x^{(t)} = u\}||$$

$$n_{u,v}(x, y) = ||\{t: x^{(t)} = u, y^{(t)} = v\}||$$

$$n_{u,v,w}(x, y, z) = ||\{t: x^{(t)} = u, y^{(t)} = v, z^{(t)} = w\}||$$

Βάση αυτών και ορίζοντας επιπλέον την πόσοτητα ξ(x) σαν

$$\xi(x) = \frac{x}{T}\log(x) \quad \forall x$$

αλλά και μέσω της σύμβασης ότι $\xi(0) = 0$ ορίζονται οι σχέσεις

$$H(Y) = \log(T) - \sum_{u\in\{0,1\}} \xi(n_u(y))$$

$$H(Y, X_n) = \log(T) - \sum_{u,v\ \in\{0,1\}^2} \xi(n_{u,v}(y, x_n))$$

$$H(Y, X_n, X_m) = \log(T) - \sum_{u,v,w\ \in\{0,1\}^3} \xi(n_{u,v,w}(y, x_n, x_m))$$

Οι υπολογισμοί αυτών των ποσοτήτων βασίζονται πάνω στη μέτρηση του αριθμού των εμφανίσεων συγκεκριμένων μοτίβων δυαδικών ψηφίων(bits), σε οικογένειες αποτελούμενες από 1-3 διανύσματα, καθώς και στην αξιολόγηση των συναρτήσεων ξ( ), σε ακέραιες τιμές μεταξύ 0-T (όπου T ο αριθμός των δειγμάτων εκπαίδευσης).

Ο πιο δύσκολος υπολογισμός είναι φυσικά ο πρώτος, και οι εκτιμήσεις για τα $n_u, n_{u,v}, n_{u,v,w}$ τα οποία αναλύονται σε συνδέσμους δυαδικών διανυσμάτων οι οποίοι υπολογίζονται σε bits. Μια λύση η οποία θα βοηθούσε τη διαδικασία υπολογισμού τους, θα ήταν μέσω ενός πίνακα αναφοράς ο οποίος θα βοηθούσε στη μέτρηση του αριθμού των bits σε ομάδες bytes καθώς και υπολογισμό των συνδέσμων σε σύνολα των 32 bits. Ο πίνακας αναφοράς, θα μπορούσε επιπλέον να χρησιμοποιηθεί για την αξιολόγηση του ξ, αφού οι τιμές αυτού δίνονται σε ακέραιες τιμές.

<u>ΚΑΝΟΝΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ CMIM:</u>

Ο μέθοδος CMIM, στην απλή της μορφή, έχει ως εξής. Αρχικά μέσω της συνάρτησης mut_inf υπολογιζεται η αμοιβαία πληροφορία $I(Y; X_n)$ η οποία αρχικοποιεί το πρόβλημα, δίνοντας τη τιμή αυτής στη συνάρτηση s[n]. Ακολούθως, ο αλγόριθμος σε κάθε επανάληψη βρίσκει το χαρακτηριστικό N(k) με την ψηλότερη τιμή και ανανεώνει κάθε φορά τη τιμή του s[n], το οποίο παίρνει την ελάχιστη τιμή για την ποσότητα $I(Y; X_n|X_m)$ η οποία υπολογίζεται μέσω της συνάρτησης cond_mut_inf, όπου η συνάρτηση $s[n] = \min_{l\le k} I(Y; X_n|\, X_{N(l)}$. Το υπολογιστικό κόστος του αλγορίθμου αυτού υπολογίζεται να είναι $O(K \times N \times T)$, όπου $K \times N$ οι κλησεις

της μεθόδου cond_mut_inf, κάθε μια από τις οποίες κοστίζει $O(T)$. Αναλύτικά η μέθοδος που ακολουθείται, έχει ως εξής

$$for\ n = 1 \ldots . N \quad do$$

$$s[n] \leftarrow mut\_\inf(n)$$

$$for\ k = 1 \ldots . K \quad do$$

$$nu[k] = arg \max_{n} s[n]$$

$$for\ n = 1 \ldots N \quad do$$

$$s[n] \leftarrow \min(s[n], cond\_mut\_\inf(n, nu[k]))$$

ΓΡΗΓΟΡΟΣ ΑΛΓΟΡΙΘΜΟΣ CMIM:

Ο γρήγορος αλγόριθμος CMIM, βασίζεται στο γεγονός ότι το σκορ των διανυσμάτων καθώς προχωρά η διαδικασία, μπορεί μόνο να μειώνεται με αποτέλεσμα τα κακά σκορ, μπορούν να αποφευχθούν να ανανεώνονται. Έστω διαισθητικά, ας φανταστεί κάποιος ένα σύνολο που αποτελείται από χαρακτηριστικά πανομοιότυπα μεταξύ τους. Επιλέγοντας ένα από αυτά, τα υπόλοιπα καθίστανται αχρείαστα μέχρι και το τέλος όλων των υπολογισμών αφού δεν έχουν να προσφέρουν τίποτα περισσότερο. Το γεγονός αυτό μπορεί να εντοπιστεί στην αρχή όταν και το σκορ είναι χαμηλό και θα παραμείνει χαμηλό, εφόσον μπορεί μόνο να μειώνεται.

Συγκεκριμένα ο γρήγορος αλγόριθμος, αποθηκεύει για κάθε χαρακτηριστικό Xn, μια τιμή ps[n], η οποία είναι η ελάχιστη από όλες τις τιμές της υπό συνθήκης πληροφορίας όπως αυτή διατυπώθηκε στην εξίσωση (2) του αλγορίθμου για τη μεγιστοποίηση της αμοιβαίας πληροφορίας. Επιπλέον προστίθεται στον αλγόριθμο, ένας επιπλέον παράγοντας, ο m[n], ο οποίος, φέρει κάθε φορά το περιεχόμενο του τελευταίου χαρακτηριστικού που επιλέχθηκε και λήφθηκε υπόψη για τον υπολογισμό του ps[n] το οποίο ορίζεται σαν

$$ps[n] = \min_{l \le m[n]} I(Y; X_n | X_{N(l)}$$

Σε κάθε επανάληψη, ο αλγόριθμος εξετάζει όλα τα χαρακτηριστικά και αναβαθμίζει το σκορ του ps[n], μονο εφόσον το καλύτερο χαρακτηρστικό στην επανάληψη αυτή, δεν είναι καλύτερο από το προηγούμενο, εφόσον το τελικό σκορ μπορεί μόνο να μειώνεται σε κάθε επανάληψη. Για παράδειγμα, εαν το πρώτο βέλτιστο χαρακτηρστικό ισούται με 0.02 και το 2^{o} έχει τιμή 0.01, δεν χρειάζεται να γίνει αναβάθμιση στη τιμή, αφού είναι καλύτερο, έχει μικρότερη τιμή Η μέθοδος που ακολουθείται με αυτό τον αλγόριθμο δίνει ακριβώς τα ίδια αποτελέσματα με την κανονική μέθοδο, που περιγράφηκε πιο πριν. Συγκεκριμένα, η διαφορά του είναι ότι τρέχει ανάμεσα στα υποψήφια χαρακτηριστικά, χωρίς όμως να υπολογίζει την CMI, μεταξύ ενός υποψηφίου προς επιλογή χαρακτηριστικού και της κλάσης, δοσμένων των τελευταίων επιλεγμένων χαρακτηριστικών, εφόσον η τιμή του υποψηφίου αυτού χαρακτηριστικού, είναι πιο χαμηλή από την ήδη επιλεγμένη ως βέλτιστη s^* με αποτέλεσμα να μειώνει έτσι τον αριθμό των πράξεων. Ο γρήγορος αλγόριθμος CMIM έχει την ακόλουθη μορφή.

$$for\ n = 1 \ldots . N \quad do$$

$$ps[n] \leftarrow mut\_\inf(n)$$

$$m[n] \leftarrow 0$$

$$for\ k = 1 \ldots . K \quad do$$

$$s^* \leftarrow 0$$

$$for\ n = 1 \ldots N \ \ do$$

$$while\ ps[n] > s^*\ and\ \ m[n] < k - 1 \quad do$$

$$m[n] \leftarrow m[n] + 1$$

$$ps[n] \leftarrow \min(ps[n], cond\_mut\_\inf(n, nu\big[m[n]\big]))$$

$$if\ ps[n] > s^*\ \ then$$

$$s^* \leftarrow ps[n]$$

$$nu[k] \leftarrow n$$

5.2.3 ΣΥΓΚΡΙΣΗ ΜΕ ΑΛΛΕΣ ΜΕΘΟΔΟΥΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ:

Προκειμένου να αποδειχθεί η χρησιμότητα της μεθόδου CMIM, καθώς και η ικανότητα της στη ταξινόμηση, αυτή συγκρίθηκε από τον Francois Fleuret, μαζί με άλλες μεθόδους επιλογής χαρακτηριστικών και ακολούθως με τη χρήση διαφόρων ταξινομητών, μετρήθηκε το προκύπτον σφάλμα για την ταξινόμηση καθώς και η ταχύτητα της κάθε μεθόδου. Συγκεκριμένα, οι πειραματικές διαδικασίες έγιναν με

τη χρήση 2 διαφορετικών συνόλων δεδομένων, το πρώτο από τα οποία αφορούσε ένα σύνολο εικόνων και την ταξινόμηση αυτών σε 2 κατηγορίες, ενώ το δεύτερο, ένα σύνολο διαφόρων ενώσεων και την αξιολόγηση αυτών για την ικανότητα τους να δεσμεύσουν μια περιοχή με θρομβίνη. Και στις 2 περιπτώσεις, τα δεδομένα που επιλέχθησαν, χωρίστηκαν σε 2 ομάδες, στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής, προκειμένου να βοηθήσουν στη διεκπαιρέωση της διαδικασίας ταξινόμησης. Όπως θα φανεί και στη συνέχεια, πράγματι, η CMIM, πετυγχαίνει υψηλές αποδόσεις αποδεικνύοντας το λόγο για τον οποίο θεωρείται από τις πιο καλές μεθόδους επιλογής χαρακτηριστικών.

5.2.3.1 ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ

ΣΥΝΟΛΟ ΦΩΤΟΓΡΑΦΙΩΝ:

Το πρώτο πείραμα, αφορούσε τη ταξινόμηση εικόνων, ένα κλασσικό δηλαδή πρόβλημα αναγνώρισης το οποίο προσπαθεί να προβλέψει τη πραγματική κατηγορία στην οποία μια εικόνα ανήκει. Στην προκειμένη, υπήρχαν 2 κατηγορίες, φωτογραφίες προσώπων και φωτογραφίες εξωφύλλου. Για αυτή τη πειραματική διαδικασία, χρησιμοποιήθηκαν δεδομένα τα οποία επιλέχτηκαν από το διαδύκτιο. Ειδικότερα για κάθε πρόσωπο σε κάθε σκηνή, παράχθηκαν 10 μικρές διαφορετικές εικόνες προσώπου μεσω περιστροφής, μετάφρασης ή αλλαγής της κλίμακας τους, έτσι ώστε να τυχαιοποιηθεί η στάση τους, ενώ για τις φωτογραφίες εξωφύλλου, συγκεντρώθηκαν σκηνές από δάση, δημόσια κτίρια και κατασκευές, μέσα από τις οποίες απομονώθηκαν συγκεκριμένες εικόνες. Συνολικά, επιλέχτηκαν 14,268 φωτογραφίες προσώπων και 14,800 φωτογραφίες εξωφύλλου ως δεδομένα εκπαίδευσης και αντίστοιχα 5,202 με 5,584 ως δεδομένα δοκιμής.

Κάθε εικόνα είχε μέγεθος 28x28 pixels και 256 επίπεδα grayscale, ενώ συγκεκριμένα για τις φωτογραφίες προσώπου, αυτές επεξεργάστηκαν κατάλληλα έτσι ώστε η απόσταση μεταξύ των ματιών να είναι 10-12 pixels και η κλίση τους από -20 μέχρι 20 βαθμούς. Για κάθε πείραμα, τόσο το σύνολο εκπαίδευσης όσο και το σύνολο δοκιμής, αποτελούνταν από 500 εικόνες προσώπων και εξωφύλλου.

ΣΥΝΟΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΜΟΡΙΑΚΗΣ ΒΙΟ-ΔΡΑΣΤΗΡΙΟΤΗΤΑΣ:

Το 2^{ο} σύνολο δεδομένων, ήταν βασισμένο σε 1,909 διαφορετικές ενώσεις, οι οποίες δοκιμάστηκαν για την ικανότητά τους να δεσμεύουν μια περιοχή θρομβίνης. Το πρόβλημα αυτό, ήταν αντίστοιχο με ένα πρόβλημα ελέγχου φαρμάκων, το οποίο προσπαθούσε να προβλέψει ποια μόρια θα κατάφερναν να επιτύχουν το καλύτερο

αποτέλεσμα. Συγκεκριμένα κάθε ένωση είχε μια δυαδική κλάση η οποία έπαιρνε τις τιμές ενεργή-ανενεργή, καθώς και 139,351 χαρακτηριστικά, η σημασία των οποίων είναι άγνωστη, αν και αυτά παραμένουν συνεπή με τα δείγματα.

Προκειμένου να καταστεί εφικτή η χρήση των διαφόρων τεχνικών στις συγκρίσεις που θα γίνουν, ο αριθμός των δυαδικών χαρακτηριστικών, μειώθηκε στα 2,500 με μια τυχαία μέθοδο επιλογής αυτών, αφού τυχόν χρήση κλασσικών μεθόδων επιλογής, θα απαιτούσαν τεράστιο υπολογιστικό χρόνο καθιστώντας το πρόβλημα δυσεπιλυτο. Όλα τα πειράματα έγιναν με τη χρήση του cross-validation, σε 25 επαναλήψεις. Σε κάθε γύρο, επιλέγονταν 100 δείγματα σαν δεδομένα δοκιμής και τα υπόλοιπα χρησιμοποιούνταν σαν δεδομένα εκπαίδευσης. Εφόσον όμως το ισοζύγιο μεταξύ θετικών και αρνητικών δειγμάτων δεν είναι ισορροπημένο (42 θετικά, 1869 αρνητικά) το σφάλμα ισορροπίας χρησιμοποιήθηκε τόσο για τα δεδομένα εκπαίδευσης όσο και για αυτά της δοκιμής.

5.2.3.2 ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Συνολικά, έγινα 3 πειραματικές διαδικασίες, οι πρώτες 2 με τη χρήση του συνόλου φωτογραφιών και η $3^{η}$ με τη χρήση των ενώσεων για πρόβλεψη της βιο-δραστηριότητας. Σε όλα τα πειράματα, έγιναν διάφοροι συνδυασμοί μεταξύ των μεθόδων επιλογής χαρακτηριστικών, CMIM, FCBF(Fast Correlation Based Filter), C4,5 Binary Trees(δυαδικά δεντρα αποφασεων), Random feature selection (τυχαια επιλογη χαρακτηριστικων), MIM feature selection (Mutual information Maximization) και AdaBoost καθώς και των ταξινομητών k-NN, SVM, naïve Bayesian και Perceptron.

Προκειμένου να υπολογιστεί η στατιστική σημαντικότητα στις συγκρίσεις που έγιναν, έγινε μια εκτίμηση του ποσοστού του σφάλματος από τα σύνολα δεδομένων καθώς και της διασποράς αυτών των εκτιμήσεων, με την υπόθεση πάντα ότι τα δείγματα, είναι πάντα ανεξάρτητα καθώς και ισοδύναμα κατανεμημένα. Συγκεκριμένα μετρήθηκαν και στις 3 πειραματικές διαδικασίες το σφάλμα εκπαίδευσης, το σφάλμα ταξινόμησης καθώς και η ποσότητα $\frac{e^{*}-e}{\sqrt{\sigma_{e^{*}-\sigma_e}}}$, όπου e^{*} είναι το σφάλμα της μεθόδου αναφοράς CMIM συνδυασμένη με τον ταξινομητή naïve Bayesian, ενώ τα $\sigma_{e^{*}}$ και σ_e είναι οι εκτιμήσεις της διασποράς για τα αντίστοιχα σφάλματα.

Για την πρώτο πείραμα, το πρόβλημα, ήταν η επιλογή μέσω μιας μεθόδου χαρακηριστικών, ενός συνόλου 50 εικόνων και ακολούθως η ταξινόμηση αυτών, ο υπολογισμός του σφάλματος και η μέτρηση της ταχύτητας της κάθε μεθόδου. Η παράμετρος σ του Γκαουσιανού πυρήνα, υπολογίστηκε ξεχωριστά για κάθε μέθοδο επιλογής χαρακτηριστικών μέσω της βελτιστοποιήσης του σφάλματος δοκιμής σε 25 γύρους. Επιπλέον τα σφάλματα ταξινόμησης και δοκιμής εκτιμήθηκαν

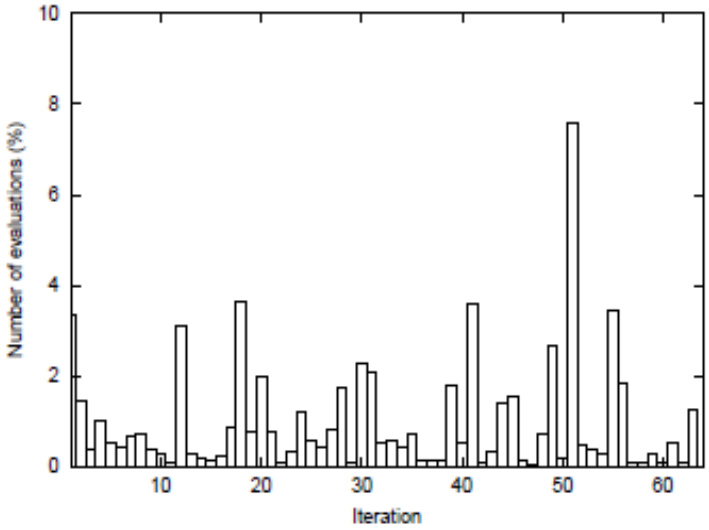
ξεχωριστά, μέσω άλλων 25 γύρων δοκιμών με τη χρήση του cross-validation. Παρόλο που τα δεδομένα μπορεί να υπόφεραν από over-fitting και over-estimate, εντούτοις αυτά τα προβλήματα θεωρούνται αμελητέα, λαμβάνοντας υπόψη το μεγάλο μέγεθος των δεδομένων που υπήρχαν.

Η δεύτερη πειραματική διαδικασία, έγινε προκειμένου να ελεγχθεί η ευρωστία του συνδυασμού της CMIM μεθόδου με τον ταξινομητή naïve Bayes. Ακριβώς για αυτό το σκοπό, εδώ χρησιμοποιήθηκαν θορυβώδη(noisy) δεδομένα εκπαίδευσης, τα οποία ως γνωστό είναι προβληματικά όσον αφορά συστήματα boosting. Αυτό επιτεύχθηκε μέσω της τυχαίας αλλαγής κατά 5% στις ετικέτες εκπαίδευσης, κάτι το οποίο είναι αρκετά ρεαλιστικό, λαμβάνοντας κάποιος υπόψη ότι είναι πολύ πιθανόν κάποια δεδομένα εκπαίδευσης να τοποθετηθούν σε λάθος κατηγορία. Αυτή η κατάσταση δυσκολεύει αρκετά τις μηχανές εκμάθησης οι οποίες λαμβάνουν σημαντικά υπόψη τους τις ακραίες τιμές, εφόσον ενα ποσοστό 5% αυτών των τιμών είναι κατανεμημένο ομοιόμοργα μεταξύ των δεδομένων εκπαίδευσης.

Τέλος η τρίτη πειραματική διαδικασία είναι βασισμένη στο σύνολο των ενώσεων οι οποίες προβλέπουν τη μοριακή βιο-δραστηριότητα και ασχολείται με τη τελική επιλογή 10 χαρακτηριστικών μεταξύ των δεδομένων. Αυτή η διαδικασία ήταν και η πιο δύσκολη για να αναλυθεί, εφόσον τα χαρακτηριστικά με τα οποία έχουμε να κάνουμε, είναι κατα κύριο λόγο άγνωστα. Επιπλέον λόγω του ότι όπως αναφέρθηκε και πιο πάνω ο πληθυσμός τους δεν είναι ισορροπημένος, οι μεθόδοι που εφαρμόζονται είναι ευαίσθητες σε φαινόμενα over-fitting (υπερπροσαρμογής).

5.2.3.3 ΤΑΧΥΤΗΤΑ ΜΕΘΟΔΩΝ

Όπως ειπώθηκε πιο πάνω, για τη ταξινόμηση των φωτογραφιών επιλέγονται 50 χαρακτηριστικά μεταξύ των 43,904 που υπάρχουν με τη βοήθεια ενός συνόλου 500 εκπαιδευτικών δεδομένων. Η κανονική εφαρμογή της CMIM με τον naïve Bayesian, απαιτεί χρόνο 18800ms, προκειμένου να επιτύχει αυτή την επιλογή δεδομένων σε ένα συνηθισμένο υπολογιστή , ενώ αντίστοιχα η γρήγορη εκδοχή αυτής για την ίδια ακριβώς επιλογή δεδομένων, απαιτεί χρόνο μόνο 255ms. Αντίστοιχα για το σύνολο δεδομένων για τη θρομβίνη, όπου στόχος είναι η επιλογή 10 χαρακτηριστικών απο τα 139,351, με ένα σύνολο δεδομένων εκπαίδευσης αποτελούμενο από 1909 δείγματα, ο υπολογιστικός χρόνος είναι 156,545ms για την κανονική εκτέλεση, ενώ με τη γρήγορη εκτέλεση ο χρόνος αυτός είναι μόνο 1401ms.



Σχήμα 5.8: απεικονίζει για το σύνολο δεδομένων προσώπων, το ποσοστό (%) των αξιολογήσεων σε κάθε βήμα της διαδικασίας επιλογής των χαρακτηριστικών

Η δραματική αυτή μείωση στο χρόνο που χρειάζεται για να αποδόσει ο αλγόριθμος μπορεί να εξηγηθεί απλά κοιτάζοντας τον αριθμό των κλησεων της cond_mut_inf στον κανονικό και στον γρήγορο αλγόριθμο, με τη διαφορά για την επιλογή των φωτογραφιών να είναι από 4,346,496 στις 54,928 κλήσεις, ενώ για τα δείγματα θρομβίνης η διαφορά να είναι από 13,795,749 στις 62,125.

5.2.3.4 ΣΧΟΛΙΑΣΜΟΣ ΠΕΙΡΑΜΑΤΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Πιο κάτω, παρουσιάζονται τα αποτελέσματα όλων των συγκρίσεων που έγιναν και για τις 3 πειραματικές διαδικασίες. Είναι ξεκάθαρο, ότι η CMIM μέθοδος είναι μια απο τις καλύτερες που υπάρχουν για επιλογή χρακτηριστικών, αφού ακόμη και συνδυασμένη με τον απλό ταξινομητή naïve Bayesian, καταλαμβάνει την $4^{η}$, $3^{η}$ και $1^{η}$ θέση αντίστοιχα για τα 3 πειράματα που έγιναν σε σύγκριση με τους υπόλοιπους 26 συνδυασμούς που υπάρχουν μεταξυ μεθόδου επιλογής και ταξινομητή.

Classifier	Training error	Test error (e)	$\frac{e^* - e}{\sqrt{\sigma_{e^*} + \sigma_e}}$
CMIM + SVM	0.53%	1.12%	−2.77
AdaBoost feature selection + SVM	0%	1.21%	−2.11
AdaBoost	0%	1.45%	−0.45
CMIM feature selection + naive Bayesian	0.52%	1.52%	–
CMIM feature selection + k-NN	0%	1.69%	1.07
AdaBoost feature selection + k-NN	0%	1.71%	1.19
FCBF feature selection + SVM	0.75%	1.85%	2.02
FCBF feature selection + naive Bayesian	1.28%	2.13%	3.60
CMIM feature selection + perceptron	0%	2.28%	4.40
AdaBoost feature selection + perceptron	0%	2.46%	5.32
C4.5 feature selection + SVM	0.73%	2.58%	5.91
FCBF feature selection + k-NN	0%	2.75%	6.73
C4.5 feature selection + perceptron	0%	3.26%	9.02
C4.5 feature selection + naive Bayesian	1.4%	3.28%	9.11
FCBF feature selection + perceptron	0%	3.50%	10.03
AdaBoost feature selection + naive Bayesian	0.4%	3.51%	10.06
C4.5 feature selection + k-NN	0%	3.57%	10.31
MIM + SVM	3.26%	5.67%	17.73
MIM feature selection + perceptron	3.56%	8.28%	25.06
MIM feature selection + naive Bayesian	5.58%	8.54%	25.72
MIM feature selection + k-NN	0.23%	8.99%	26.84
Random feature selection + SVM	9.04%	11.86%	33.44
Random feature selection + perceptron	13.36%	17.45%	44.66
Random feature selection + k-NN	0.30%	21.54%	52.18
Random feature selection + naive Bayesian	21.69%	24.77%	57.93

Πίνακας 5.5: Απεικονίζονται τα σφάλματα εκπαίδευσης και δοκιμής καθώς και η διαφορά μεταξύ του σφάλματος e^* και του σφάλματος δοκιμής e για όλους τους συνδυασμούς μεθόδων επιλογής χαρακτηριστικών και ταξινομητών για το σύνολο των φωτογραφιών προσώπων και εξωφύλλων

Classifier	Training error	Test error (e)	$\frac{e^{*}-e}{\sqrt{\sigma_{e^{*}}+\sigma_{e}}}$
CMIM + SVM	5.68%	1.37%	−3.59
FCBF feature selection + SVM	6.02%	1.49%	−2.79
CMIM feature selection + naive Bayesian	5.06%	1.95%	–
FCBF feature selection + naive Bayesian	5.38%	2.39%	2.38
C4.5 + SVM	5.57%	2.99%	5.30
AdaBoost$_{reg}$ (optimized on test set)	3.80%	3.06%	5.61
C4.5 feature selection + naive Bayesian	6.14%	3.62%	8.03
AdaBoost feature selection + SVM	4.39%	4.18%	10.25
CMIM feature selection + k-NN	0.08%	5.36%	14.42
MIM + SVM	7.85%	5.87%	16.07
AdaBoost	0.58%	6.33%	17.48
C4.5 feature selection + k-NN	0.71%	6.34%	17.52
FCBF feature selection + k-NN	0.87%	6.50%	17.99
AdaBoost feature selection + k-NN	0.39%	7.20%	20.02
AdaBoost feature selection + perceptron	0.12%	8.23%	22.82
MIM feature selection + naive Bayesian	9.47%	8.59%	23.75
CMIM feature selection + perceptron	7.36%	9.32%	25.60
FCBF feature selection + naive Bayesian	8.20%	9.33%	25.62
AdaBoost feature selection + naive Bayesian	10.28%	9.46%	25.94
C4.5 feature selection + perceptron	7.58%	11.06%	29.71
Random + SVM	13.00%	12.19%	32.23
MIM feature selection + k-NN	2.92%	11.46%	30.61
MIM feature selection + perceptron	11.53%	13.12%	34.23
Random feature selection + perceptron	19.47%	20.58%	48.75
Random feature selection + k-NN	1.43%	24.77%	56.29
Random feature selection + naive Bayesian	24.13%	24.99%	56.68

Πίνακας 5.6: Απεικονίζονται τα σφάλματα εκπαίδευσης και δοκιμής καθώς και η διαφορά μεταξύ του σφάλματος e^{*} και του σφάλματος δοκιμής e για όλους τους συνδυασμούς μεθόδων επιλογής χαρακτηριστικών και ταξινομητών για το σύνολο των φωτογραφιών προσώπων και εξωφύλλων όπου εδώ οι ετικέτες εκπαίδευσης έχουν αλλάξει κατά ένα ποσοστό 5%.

Classifier	Training error	Test error (e)	$\frac{e^* - e}{\sqrt{\sigma_{e^*} + \sigma_e}}$
CMIM feature selection + naive Bayesian	10.45%	11.72%	–
AdaBoost feature selection + SVM	9.35%	12.99%	1.36
AdaBoost feature selection + naive Bayesian	10.29%	13.60%	1.99
AdaBoost$_{reg}$ (optimized on test set)	9.48%	13.64%	2.04
CMIM + SVM	13.21%	13.65%	2.05
AdaBoost	9.49%	13.76%	2.16
C4.5 feature selection + naive Bayesian	9.22%	13.90%	2.31
C4.5 + SVM	8.72%	17.34%	5.65
CMIM feature selection + k-NN	17.17%	18.77%	6.97
FCBF feature selection + naive Bayesian	13.62%	19.22%	7.37
FCBF feature selection + SVM	13.39%	23.14%	10.76
MIM feature selection + naive Bayesian	21.53%	23.35%	10.94
CMIM feature selection + perceptron	20.31%	23.51%	11.08
C4.5 feature selection + perceptron	12.86%	23.88%	11.38
MIM + SVM	24.65%	25.75%	12.93
FCBF feature selection + perceptron	21.98%	27.06%	13.98
FCBF feature selection + k-NN	19.28%	27.94%	14.68
Random + SVM	30.10%	30.92%	17.05
C4.5 feature selection + k-NN	24.18%	34.11%	19.54
Random feature selection + naive Bayesian	39.32%	40.13%	24.23
MIM feature selection + perceptron	32.70%	40.27%	24.34
Random feature selection + perceptron	43.61%	45.68%	28.63
Random feature selection + k-NN	45.09%	47.29%	29.94
MIM feature selection + k-NN	50.00%	50.00%	32.19

Πίνακας 5.7: Απεικονίζονται τα σφάλματα εκπαίδευσης και δοκιμής καθώς και η διαφορά μεταξύ του σφάλματος e^* και του σφάλματος δοκιμής e για όλους τους συνδυασμούς μεθόδων επιλογής χαρακτηριστικών και ταξινομητών για τη διαδικασία επιλογής 10 χαρακτηριστικών από το σύνολο δεδομένων ανίχνευσης θρομβίνης.

Στον 1^{ο} πίνακα όπου παρουσιάζονται τα αποτελέσματα για την πρώτη πειραματική διαδικασία, το καλύτερο σκορ, επιτυχάνονται με τη χρήση της μεθόδου CMIM, μαζί με τον ταξινομητή SVM, ακολουθούμενο από τη AdaBoost μαζί με τον SVM. Επίσης είναι ξεκάθαρο ότι η CMIM μέθοδος είναι η καλύτερη, συνδυασμένη με όλους τους ταξινομητές σε σχέση με τις υπόλοιπες μεθόδους που χρησιμοποιούνται.

Στον 2^{ο} πίνακα, φαίνεται και πάλι η υπεροχή της CMIM, αφού έρχεται πρώτη μαζί με τον SVM ταξινομητή, ακολουθούμενη από την FCBF συνδυασμένη και αυτή με τον SVM. Εδώ η CMIM, υπερταιρεί με τους ταξινομητές k-NN, SVM και naïve Bayesian κάτι που δεν συμβαίνει με τον perceptron, όπου έρχεται δεύτερη μετά την AdaBoost. Επιπλέον εδώ όλες οι μεθόδοι που είναι βασισμένες στην αίσθηση(perceptron) ή στην ενίσχυση(boosting), έχουν ψηλές τιμές σφάλματος εφόσον αυτές είναι ευαίσθητες στις ακραίες τιμές, αντίθετα με τις άλλες τεχνικες ταξινόμησης όπως οι SVM, naïve Bayesian και AdaBoost οι οποίες και μπορούν να προστατεύονται από το over-fitting για αυτό και καταλαμβάνουν τις πρώτες 8 θέσεις στον πίνακα.

Τέλος στον 3^{ο} πίνακα, η καλύτερη μέθοδος είναι η CMIM με τον naïve Bayesian ταξινομητή με δεύτερη καλύτερη την AdaBoost με τον SVM. Εδώ η CMIM, δεν υπερταιρεί σε όλους τους ταξινομητές, εφόσον με τον SVM, η AdaBoost μέθοδος επιλογής χαρακτηριστικών, αποδίδει καλύτερα με αυτόν. Τέλος παρατηρείται ότι τα σφάλματα εδώ, είναι ιδιαίτερα ψηλά, κάτι που ειπώθηκε και πιο πάνω, εφόσον τα δεδομένα τα οποία είναι διαθέσιμα προς επιλογή, δεν είναι ισορροπημένα και επιπλέον περιέχουν χαρακτηριστικά πλείστα από τα οποία ειναι άγνωστα.

Συμπερασματικά, η CMIM είναι μια πολύ απλή μέθοδος επιλογής χαρακτηριστικών η οποία αποδίδει καλύτερα στην ταξινόμηση, σε σύγκριση με άλλες μεθόδους και μάλιστα σε πολύ ικανοποιητικό βαθμό, με μικρό σφάλμα και επίσης με μεγάλη ταχύτητα, εφόσον απαιτεί χρόνο εκπαίδευσης, μόνο μερικά εκατοστά του δευτερολέπτου. Η μέθοδος αυτή, μπορεί να χρησιμοποιηθεί ικανοποιητικά, σε εφαρμογές για τις οποίες απαιτείται η εκπαίδευση ενός μεγάλου συνόλου ταξινομητών, εφόσον μπορεί να συνδυαστεί με πληθώρα από αυτούς.

5.3 Η mMIFS-U ΜΕΘΟΔΟΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Σε αυτό το κεφάλαιο, προτείνεται ένας εύκολος και αποδοτικός τρόπος διαδοχικής επιλογής χαρακτηριστικών, προκειμένου να βρεθεί ένα κατάλληλο υποσύνολο από αυτά, για το πρόβλημα της ταξινόμησης. Η μέθοδος αυτή, χρησιμοποιεί μια νέα προσέγγιση της υπό συνθήκη αμοιβαίας πληροφορίας, μεταξύ των υποψηφίων προς επιλογή χαρακτηριστικών και των κλάσεων, η οποία βασίζεται σε ένα σύνολο από ήδη επιλεγμένα χαρακτηριστικά τα οποία ανήκουν από πριν σε κάποια από τις κλάσεις που υπάρχουν και χρησιμοποιούνται σαν ένα ανεξάρτητο κριτήριο ταξινόμησης των νέων χαρακτηριστικών.

Ο αλγόριθμος mMIFS-U που προτείνεται από τους Novovicova, Somol Haindl και Pudil , εφαρμόζεται κυρίως σε προβλήματα κατηγοριοποίησης κειμένων, δινοντας πολύ καλά αποτελέσματα, πράγμα το οποίο αποδεικνύεται στο τέλος συγκρίνοντας τον με τη μέθοδο MIFS του Battini καθώς και τη MIFS-U των Kwak και Choi.

5.3.1 ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ

Στα πλείστα προβλήματα ταξινόμησης, η επιλογή χαρακτηριστικών πριν από αυτή είναι σχεδόν απαραίτητη, κάτι που μπορεί να περιγραφεί και από την ακόλουθη κατάσταση. Έστω ότι υπάρχει το εξής σύνολο από m χαρακτηριστικά $X = \{X_1, X_2, \dots, X_m\}$ το οποίο σχετίζεται με το σύνολο των n κλάσεων $D = \{c_1, c_2, \dots, c_n\}$. Δοσμένου ενός συνόλου δεδομένων εκπαιδευσης(training data), στόχος είναι με τη βοήθεια αυτών να γίνει η σωστή ταξινόμηση των χαρακτηριστικών στη σωστή κλάση από το σύνολο D αυτών.

Στη περίπτωση που το σύνολο των δεδομένων εκπαίδευσης είναι μικρό σε σχέση με το σύνολο των χαρακτηριστικών Χ, παρατηρούνται φαινόμενα επιλογής περιττών ή και άσχετων χαρακτηριστικών, λόγω του ότι οι ταξινομητές δεν μπορούν να αποδώσουν αποτελεσματικά πάσχοντας από φαινόμενα υπερπροσαρμογής. Ακριβώς για αυτό το σκοπό είναι αναγκαία η μείωση του αριθμού των χαρακτηριστικών. Αυτό μπορεί να επιτευχθεί μέσω της επιλογής από το σύνολο των Χ χαρακτηριστικών, ενός υποσυνόλου από k χαρακτηριστικά, τα οποία θα αντιπροσωπεύουν τον επιθυμητό αριθμό έτσι ώστε να γίνει σωστά η ταξινόμηση.

Στόχος της επιλογής χαρακτηριστικών στη ταξινόμηση, είναι να επιτευχθεί η μέγιστη δυνατή τιμή της αμοιβαίας πληροφορίας I(C,X), με το μικρότερο δυνατό υποσύνολο χαρακτηριστικών. Με την επιλογή αυτών που μεγιστοποιούν την I(C,S), όπου ScX τα υποψήφια προς επιλογή χαρακτηριστικά, μπορούν να συγκριθούν με τα χαρακτηριστικά τα οποία έχουν ήδη επιλεγεί έτσι ώστε να παραμείνουν στο τέλος αυτά τα οποία σχετίζονται πιο πολυ με την κλάση.

Πιο συγκεκριμένα, από το σύνολο X, επιλέγεται το κατάλληλο υποσύνολο ScX, το οποίο αποτελείται από k<m, χαρακτηριστικά με στόχο να ελαχιστοποιησει τη υπό συνθήκη εντροπία H(C|S), δηλαδή να μεγιστοποιήσει τη αμοιβαία πληροφορία I(C,S), μεταξύ του υποσυνόλου των χαρακτηριστικών S και της κλάσης C.

5.3.2 ΜΕΘΟΔΟΙ ΤΑΞΙΝΟΜΗΣΗΣ MIFS ΚΑΙ MIFS-U

Πρακτικά, είναι ασύμφορο, αν όχι σχεδόν αδύνατο να υπολογιστεί η αμοιβαία πληροφορία μεταξύ όλων των υποψηφίων χαρακτηριστικών και της κλάσης. Η υλοποίηση αυτής της μεθόδου απαιτεί τεράστιους υπολογισμούς οι οποίοι ακόμη και με τη διαδοχική προς τα εμπρός αναζήτηση είναι υπολογιστικά πολύ ακριβοί. Ακριβώς για αυτούς τους λόγους έχουν αναπτυχθεί παρεμφερείς μεθόδοι για τον υπολογισμό της I(C,S). Τέτοιες μεθόδοι είναι οι αλγόριθμοι MIFS και MIFS-U οι οποίες αναπτύχθηκαν από τους Roberto Battiti το 1994 και τους Kwak και Choi το 1999 αντίστοιχα.

Έστω ότι είναι ήδη γνωστό ένα υποσύνολο από ήδη επιλεγμένα χαρακτηριστικά, το S. Προκειμένου να επιλεγεί ένα χαρακτηριστικό Xi από το σύνολο X/S των χαρακτηριστικών που δεν έχουν ακόμη επιλεγεί, θα πρέπει η ποσότητα της πληροφορίας μεταξύ αυτού και της κλάσης C, χωρίς να ληφθεί υπόψη η πληροφορία που δίνεται από τα ήδη επιλεγμένα χαρακτηριστικά S, να είναι η μέγιστη μεταξύ όλων των υποψηφίων προς επιλογή χαρακτηριστικών που ανήκουν σε αυτό το σύνολο. Με αυτό το τρόπο μεγιστοποιείται η υπο συνθήκη αμοιβαία πληροφορία I(C,Xi|S) δοσμένου του συνόλου S.

Αυτό που έκαναν οι Battiti ,Kwak και Choi ήταν να υπολογίσουν μόνο την πληροφορία μεταξύ της κλάσης και του υποψηφίου χαρακτηριστικου I(C;Xi) καθώς και την I(Xi;Xp), αυτή δηλαδή μεταξύ του υποψηφίου χαρακτηριστικού και ενός ακόμη χαρακτηριστικού Xp που ανήκει στα ήδη επιλεγμένα XpЄS.

ΕΚΤΙΜΗΣΗ ΤΗΣ I(C,Xi|S) ΑΠΟ ΤΟΝ ΑΛΓΟΡΙΘΜΟ MIFS:

$$I_{Battiti}(C, X_i|S) = I(C; X_i) - \beta \sum_{X_p \in S} I(X_P; X_i)$$

Μια παραλλαγή αυτού είναι η εκτίμηση των Kwak και Choi, με την υπόθεση ότι η κλάση C, δεν αλάζει το λόγο μεταξύ της εντροπίας του Xp και της αμοιβαίας πληροφορίας των Xp και Xi.

ΕΚΤΙΜΗΣΗ ΤΗΣ I(C,Xi|S) ΑΠΟ ΤΟΝ ΑΛΓΟΡΙΜΘΟ MIFS-U:

$$I_{Kwak}(C, X_i|S) = I(C; X_i) - \beta \sum_{X_p \in S} \frac{I(C; X_p)}{H(X_p)} I(X_P; X_i)$$

Και στις 2 περιπτώσεις, ο δεύτερος όρος του δεξιού μέρους, χρησιμοποιείται προκειμένου να καθορίσει το μέγεθος της περιττής πληροφορίας σε σχέση με την κλάση C, μεταξύ του υποψηφίου προς επιλογή χαρακτηριστικού Xi και των ήδη επιλεγμένων χαρακτηριστικών. Τέλος το β, είναι μια παράμετρος η οποία χρησιμοποιείται προκειμένου να ελέγχει τα χαρακτηριστικά που θεωρούνται περιττά, μεταξύ αυτών που είναι μόνα τους. Ο καθορισμός του γίνεται πειραματικά, ανάλογα με τη μέθοδο ταξινόμησης που χρησιμοποιείται.

5.3.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΥΠΟ ΣΥΝΘΗΚΗ ΑΜΟΙΒΑΙΑΣ ΠΛΗΡΟΦΟΡΙΑΣ $I(C; X_i|X_p)$:

Η υπο συνθήκη αμοιβαία πληροφορία όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, στην περίπτωση που επιδίωξη είναι η μείωση της αβεβαιότητας της κλάσης C και του υποψηφίου χαρακτηριστικού Xi, όταν το Xp είναι γνωστό ισούται με

$$I(C; X_i|X_P) = H(X_i|X_p) - H(X_i|C, X_i)$$

Επιπλέον η $I(C; X_i|X_P)$, μπορεί να γραφτεί και σαν

$$I(C; X_i|X_P) = I(C; X_i) - I(X_i; X_p) + I(X_i, X_p|C) \quad \alpha\varphi o\acute{\upsilon}$$

$$\begin{aligned} & I(C; X_i) - I(X_i; X_p) + I(X_i, X_p|C) \\ &= H(C) - H(C|X_i) - H(X_i) + H(X_i|X_p) - H(X_i|C) - H(X_i|X_p, C) \\ &= H(X_i|X_p) - H(X_i|X_p, C) + H(C) - H(C|X_i) - H(X_i) + H(X_i|C) \\ &= I(C, X_i) - I(C, X_i) + H(X_i|X_p) - H(X_i|X_p, C) \\ &= I(C; X_i|X_P) \qquad (1) \end{aligned}$$

Ο συντελεστής της συνάφειας μεταξύ των Xi και Xp, μπορεί να ερμηνευτεί σαν η σχετική μείωση της αβεβαιότητας για το Xi, όταν το Xp είναι γνωστό και ορίζεται

σαν ο λόγος της αμοιβαίας πληροφορίας $I(X_i; X_p)$ με την εντροπία της γνωστής μεταβλητής Xp, δηλαδή σαν

$$CU_{X_iX_p} = \frac{I(X_i; X_p)}{H(X_p)} = \left(1 - \frac{H(X_p|X_i)}{H(X_p)}\right) \quad (2)$$

Ο πιο πάνω τύπος ονομάζεται μέτρο συσχέτισης ή και συντελεστής της αβεβαιότητας μεταξυ των 2 μεταβλητών.

Έστω τώρα ότι η κλάση C, δεν μεταβάλλει το λόγο μεταξύ της εντροπίας του Xp και της αμοιβαίας πληροφορίας μεταξύ των Xi και Xp, δηλαδή έστω ότι ισχύει

$$\frac{H(X_p|C)}{I(X_i; X_p|C)} = \frac{H(X_p)}{I(X_i; X_p)} \quad (3)$$

Τότε η υπο συνθήκη αμοιβαία πληροφορία $I(C; X_i|X_p)$ θα ισούται με

$$I(C; X_i|X_p) = I(C; X_i) - [I(X_i; X_p) - I(X_i, X_p|C)]$$

$$= I(C; X_i) - [CU_{X_iX_p} H(X_p) - I(X_i, X_p|C)]$$

$$\text{όπου } I(X_i, X_p|C) = CU_{X_iX_p} H(X_p|C) \qquad \text{από (2) και (3)}$$

$$\text{αρα } I(C; X_i|X_p) = I(C; X_i) - CU_{X_iX_p} I(C; X_p) \quad (4)$$

Βάση της εξίσωσης (4), προτείνεται η ακόλουθη εκτίμηση $I_1(C; X_i|X_p)$ για την $I(C; X_i|X_p)$ η οποία θα έχει την εξής μορφή

$$I_1(C; X_i|X_p) = I(C; X_i) - \max_{X_p \in S} CU_{X_iX_p} I(C; X_p) \quad (5)$$

Με απλά λόγια αυτό σημαίνει ότι το καλύτερο χαρακτηριστικό σε κάθε βήμα του προς τα εμπρός αλγορίθμου διαδοχικής αναζήτησης θα βρίσκεται μεγιστοποιώντας την πιο πάνω ποσότητα, δηλαδή

$$X^+ = arg \max_{X_i \in X/S} \{I(C; X_i) - \max_{X_p \in S} CU_{X_iX_p} I(C; X_p)\} \quad (6)$$

5.3.4 Ο ΑΛΓΟΡΙΘΜΟΣ mMIFS-U:

Ο αλγόριθμος διαδοχικής προς τα εμπρός αναζήτησης mMIFS-U (Novovicova, Somol, Pudil 2004) βασίζεται στην εκτίμηση της ποσότητας της αμοιβαίας πληροφορίας υπο συνθήκη, όπως αυτή υπολογίζεται από την εξίσωση (5). Πιο συγκεκριμένα, η υλοποίησή του ακολουθεί τα παρακάτω βήματα

ΑΡΧΙΚΟΠΟΙΗΣΗ

Θέτω S=”empty set”, και επιπλέον Χ=” αρχικό σύνολο όλων των Μ χαρακτηριστικών “

ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΥΠΟΛΟΓΙΣΜΟΣ

Για όλα τα χαρακτηριστικά Xi Є Χ υπολόγισε την $I(C; X_i)$

ΕΠΙΛΟΓΗ ΠΡΩΤΟΥ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ

Εντοπισμός του χαρακτηριστικού Χ*Є Χ το οποίο μεγιστοποιεί την $I(C; X_i)$

Θέτω $X = X \setminus \{X^*\}$ και S=$\{X^*\}$

ΑΠΛΗΣΤΗ ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Επανάλαβε μέχρι να επιλεγεί ο επιθυμητός αριθμός των χαρακτηριστικών

- Υπολογισμός εντροπίας

Για όλα τα $X_p \in S$ υπολόγισε την εντροπία $H(X_p)$ στη περίπτωση που δεν δίνεται ήδη

- Υπολογισμός της αμοιβαίας πληροφορίας μεταξύ όλων των χαρακτηριστικών

Για όλα τα χαρακτηριστικά (X_i, X_p) όπου $X_i \in X$ και $X_p \in S$, υπολόγισε την $I(X_i; X_p)$, εάν δεν είναι ήδη διαθέσιμη

- Επιλογή του επόμενου χαρακτηριστικού

Βρίσκω το επόμενο χαρακτηριστικό $X^+ \in X$ σύμφωνα με τη σχέση (6)

Θέτω $X = X \setminus \{X^*\}$, όπου $S = S \cup \{X^*\}$

5.3.5 ΣΥΓΚΡΙΣΗ ΜΕ ΑΛΛΕΣ ΜΕΘΟΔΟΥΣ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ:

Προκειμένου να διαπιστωθεί η απόδοση της μεθόδου mMIFS-U, διενεργήθηκαν από τους Novovicova, Somol, Haindl και Pudil μια σειρά από πειραματικές συγκρίσεις με άλλες παρόμοιες μεθόδους, για τη κατηγοριοποίηση ενός κειμένου. Πιο κάτω αναλύονται συνοπτικά, διάφορες έννοιες, οι οποίες χρησιμοποιήθηκαν στην πειραματική διαδικασία. Ακολούθως παρουσιάζονται τα πειραματικά αποτελέσματα σε γραφικές παραστάσεις.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΟΥ

Είναι η διαδικασία που ακολουθείται για την ταξινόμηση ενός αρχείου, γραμμένου στη μητρική γλώσσα σε μια ή και περισσότερες θεματικές ενότητες, δηλαδή της κατάταξης κάθε x_i αρχείου στη σωστή κατηγορία c_j, όπου κάθε αρχείο $x_i = \{x_i{}^1, x_i{}^2, \dots., x_i{}^p\}$, αντιπροσωπεύεται από ενα σύνολο χαρακτηριστικών. Η ταξινόμηση αυτή βασίζεται σε ένα σύνολο εκπαιδευτικών δεδομένων τα οποία ανήκουν σε προκαθορισμένη κατηγορία. Υπάρχουν πολλών ειδών αρχεία, επιστημονικά αρθρα, κριτικές ταινιών, διαφημίσεις κτλ. Το είδος καθορίζεται συγκεκριμένα, από τον τρόπο με τον οποίο αυτό δημιουργήθηκε, τον τρόπο που επεξεργάστηκε, τη γλώσσα που χρησιμοποιεί και το είδος του κοινού στο οποίο απευθύνεται.

ΣΥΝΟΛΟ REUTERS

Το σύνολο Reuters είναι μια συλλογή από κείμενα, η οποία μαζί και με τις προηγούμενες παραλαγγές της, έχει καταστεί σημείο αναφοράς στην Κατηγοριοποίηση Κειμένου τα τελευταία χρόνια. Συγκεκριμένα, είναι ένα σύνολο, αποτελούμενο από 21578 ιστορίες, οι οποίες εμφανίστηκαν στις ειδήσεις του Reuters κατά το έτος 1987. Τα δεδομένα αυτά συλλέχθηκαν και κατηγοροιοποιήθηκαν από την Carnegie Group Inc και την Reuters Ltd σε 135 θεματικές ενότητες, οι οποίες αφορούσαν περισσότερο την οικονομία και την επιχειρηματικότητα.

Η συλλογή αυτή, περιλαμβάνει διάφορα επιμέρους χαρακτηριστικά, τα οποία την έχουν καταστήσει αυτή τη στιγμή τη πιο γνωστή και χρησιμοποιούμενη συλλογή για κατηγοριοποίηση κειμένου. Καταρχήν κάθε κείμενο, μπορεί να ανήκει σε περισσότερες από μια κατηγορίες(κλάσεις). Επίσης το σύνολο των κατηγοριών δεν εξαντλείται, αφού υπάρχουν πολλά αρχεία τα οποία δεν ανήκουν σε καμία από αυτές. Τέλος η κατανομή των κειμένων διαμέσου των κλάσεων είναι ασύμμετρη. Για παράδειγμα, ενώ υπάρχουν κατηγορίες με ελάχιστα στοιχεία υπάρχουν άλλες που περιλαμβάνουν χιλιάδες. Λόγω του τελευταίου χαρακτηριστικού, οι

κατηγορίες με λίγα αρχεία, δυσκολεύουν τις ταξινομήσεις που βασίζονται σε τεχνικές μηχανικές μάθησης, αφού η κατασκευή του ταξινομητή καθίσταται αρκετά δύσκολη.

Το κύριο πρόβλημα που αντιμετώπιζε αυτή η συλλογή, ήταν το γεγονός, ότι οι διαφοροι ερευνητές που το χρησιμοποιούσαν, λόγω της ασάφειας που υπήρχε έφθασαν στο σημείο να χρησιμοποιούν διαφορετικά σύνολα εκπαίδευσης κάθε φορά. Το πρόβλημα αυτό έγινε αντιληπτό, όταν οι ίδιοι οι ερευνητές προσπάθησαν να αφαιρέσουν κάποια αρχεία, τα οποία δεν είχαν συγκεκριμένη ετικέτα, καθώς υπήρχαν διάφορες απόψεις για το που αυτά ανήκαν. Για να λυθεί αυτό το θέμα, πλέον καθορίζεται επακριβώς που ανήκει κάθε αρχείο, με το να διευκρινίζεται πλήρως τι πρέπει να περιλαμβάνει για να ανήκει κάπου.

Τέλος μια τεχνική που αναπτύχθηκε προκειμένου να καταστήσει δυνατή τη σύγκριση μεταξύ διαφόρων πειραματικών αποτελεσμάτων, είναι ο διαχωρισμός (split) των αρχείων, σε σύνολα εκπαίδευσης και σε σύνολα δοκιμής. Η μέθοδος που χρησιμοποιείται για αυτό το διαχωρισμό είναι η ModApte, η οποία περιλαμβάνει 9603 δεδομένα εκπαίδευσης και επίσης 3299 δεδομένα δοκιμής χωρισμένα σε 135 κλάσεις. Για τις συγκρίσεις που γίνονται εδώ, χρησιμοποιήθηκαν μόνο 90 κλάσεις για τις οποίες υπαρχει και δεδομένο εκπαίδευσης και δοκιμής.

Η προεργασία του κειμένου περιελάμβανε την απομάκρυνση όλων των στοιχείων που δεν ήταν γράμματα, όπως τελείες και κόμματα, μετατροπή των κεφαλαίων γραμμάτων σε μικρά και αγνόηση όλων των λέξεων οι οποίες περιείχαν μη αλφαριθμητικούς χαρακτήρες. Κάθε λέξη αντικαταστάθηκε από τη αντίστοιχη μορφολογική της ρίζα και επιπλέον αφαιρέθηκε κάθε λέξη με λιγότερο από 3 αναφορές. Το αποτέλεσμα ήταν να προκύψει ένα λεξιλόγιο με μέγεθος 7487 λέξεις.

ΜΕΘΟΔΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΤΗΣ ΕΠΙΔΟΣΗΣ ΤΑΞΙΝΟΜΗΣΗΣ

Προκειμένου να διαπιστωθεί εαν το αποτέλεσμα που τελικά εξάχθηκε, συμφωνεί με αυτό το οποίο προβλεποταν ή υπολογιζόταν ότι θα έιναι, γίνεται μια μελέτη της διαδικασίας που ακολουθήθηκε και αξιολόγηση της απόδοσής της. Στο πείραμα που έγινε, για την εκτίμηση της ακρίβειας της ταξινόμησης πολλαπλών κλάσεων, χρησιμοποιήθηκαν 2 συγκεκριμένα μέτρα αξιολόγησης το μέτρο precision (γνωστό και ως μέτρο θετικής προγνωστικής αξίας) και το μέτρο recall (ανάκληση). Ειδικότερα οι εκτιμήσεις έγιναν με τη χρήση των τύπων

$$prec = \frac{\sum_{j=1}^{N} TP_j}{\sum_{j=1}^{N} (TP_j + FP_j)}$$

$$rec = \frac{\sum_{j=1}^{N} TP_j}{\sum_{j=1}^{N} (TP_j + FN_j)}$$

όπου TP(true negative) είναι τα αρχεία τα οποία άνηκαν στη κλάση c_j και τοποθετήθηκαν εκεί, FP(false positive) είναι τα αρχεία τα οποία δεν ανήκαν στην κλάση c_j αλλά τοποθετήθηκαν εκεί και τέλος FN(false negative) τα αρχεία που άνηκαν στην κλάση c_j αλλά δεν τοποθετήθηκαν εκεί.

ΠΟΛΥΚΑΤΗΓΟΡΙΚΗ ΤΑΞΙΝΟΜΗΣΗ:

Η πολυκατηγορική ταξινόμηση είναι μια παραλλαγή του προβλήματος ταξινόμησης, όπου για κάθε παράδειγμα εκχωρούνται πολλαπλές ετικέτες. Βασικά με πιο απλά λόγια είναι το πρόβλημα, για το οποίο πρέπει να βρεθεί ένα μοντέλο το οποίο να αντιστοιχεί δεδομένα x με κατηγορίες y. Υπάρχουν 2 κύριες μεθόδοι για αυτό το πρόβλημα

Hard classification: Η μέθοδος μετασχηματισμού του προβλήματος, προτείνει το μετασχηματισμό του προβλήματος σε ένα σύνολο από δυαδικά προβλήματα ταξινόμησης, τα οποία μπορούν να επιλυθούν χρησιμοποιώντας απλούς ταξινομητές κλάσεων.

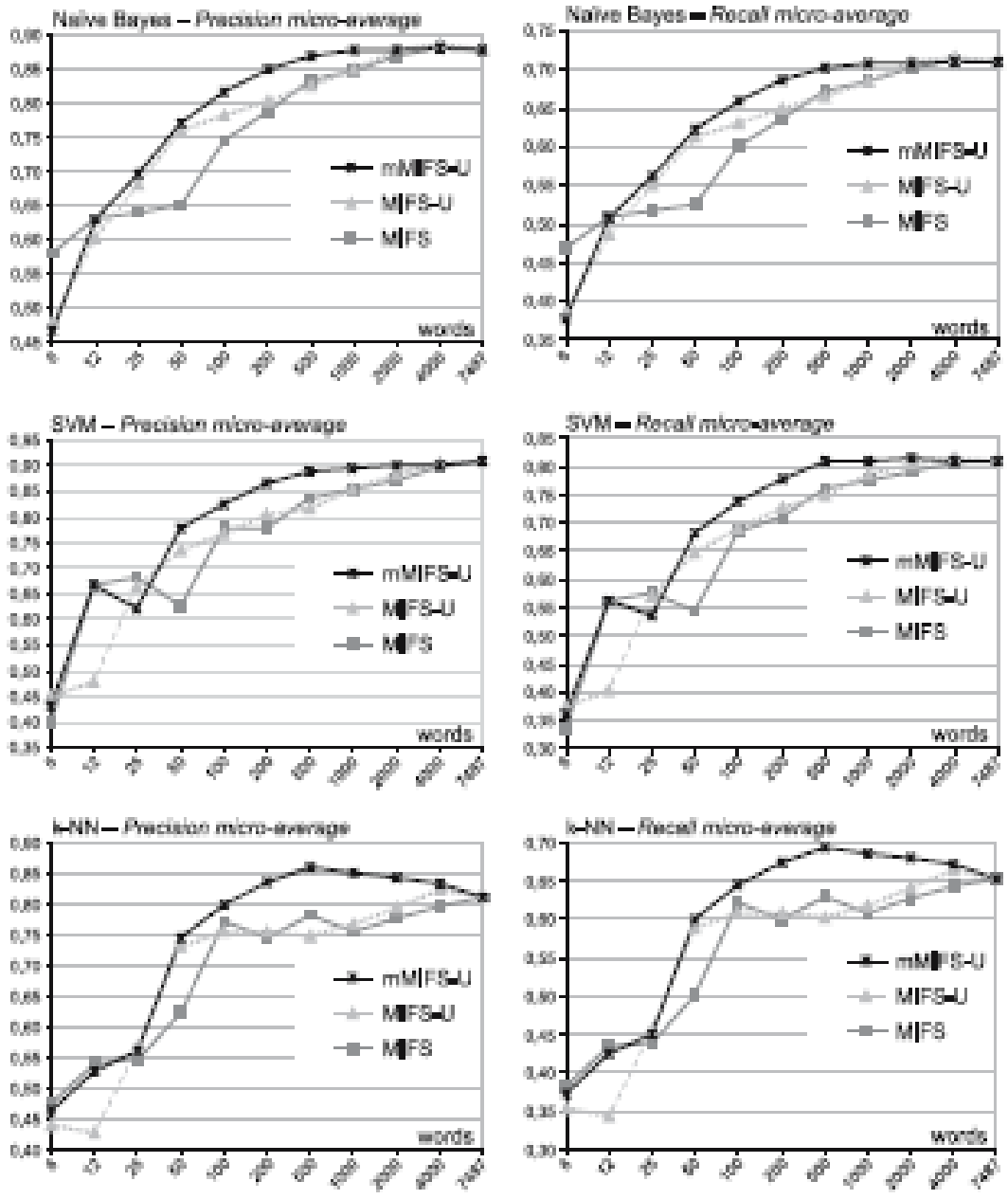
Ranking classification: Η δεύτερη μέθοδος, προσαρμογής αλγορίθμου, προκρίνει τη προσαρμογή των αλγορίθμων έτσι ώστε να κάνουν κατευθείαν την πολυκατηγορική ταξινόμηση, δηλαδή μετατρέπουν το πρόβλημα, σε ένα απλούστερο, για να αντιμετωπιστεί συνολικά.

THRESHOLDING:

Η μέθοδος thresholding παίζει καταλυτικό ρόλο όσο αφορά το τελικό αποτέλεσμα στην ταξινόμηση. Υπάρχουν 3 δημοφιλείς μεθόδοι thresholding από τις οποίες δεν είναι εύκολο να φανεί πια από αυτές είναι η βέλτιστη.

- η RCut(rank-based thresholding) η οποία ταξινομεί τις κατηγορίες βάση του αποτελέσματος που δίνουν, τοποθετώντας την ένδειξη “YES” στις t κορυφαίες από αυτές, όπου $t \in [1, \dots, m]$, m είναι ο συνολικός αριθμός των κατηγοριών
- η SCut (score- based local optimization) η οποία ταξινομεί τα δεδομένα δοκιμής βάση της απόδοσης τους και βάζει την ένδειξη “YES” σε κάθε ένα από τα k_j κορυφαία σε βαθμολογία από αυτά, όπου k ο αριθμός των αρχείων που ανήκουν στην κατηγορία C_j.

- η PCut (proportion-based assignments) επαληθεύει το σύνολο των αρχείων, συντονίζοντας την μέδοδο threshold μέχρι να επιτευχθεί η βέλτιστη απόδοση του ταξινομητή σε κάθε κατηγορία.



Σχήμα 5.9: Απεικονίζονται γραφικά οι συγκρίσεις μεταξύ των μεθόδων επιλογής χαρακτηριστικών mMIFS-U, MIFS-U, MIFS, για το Reuters σύνολο δεδομένων με τη χρήση του Apte split RCut-thresholding, όπου στις αριστερές γραφικές παραστάσεις χρησιμοποιείται η precision μέθοδος αξιολόγησης της ταξινόμησης, ενώ για τις δεξιά, χρησιμοποιείται η recall μέθοδος. Η οριζόντια γραμμή μετρά τον αριθμό των λέξεων που ταξινομούνται ενώ η κάθετη γραμμή μετρά την ακρίβεια που επιτυγχάνεται στην ταξινόμηση.

Στα πειράματα που έγιναν, η mMIFS-U, συγκρίθηκε μαζί με τις μεθόδους MIFS και MIFS-U οι οποίες αναφέρθηκαν πιο πάνω, για τη μείωση του μεγέθους ενός συνόλου λέξεων έστω $V = \{w_1, \dots, w_N\}$ αποτελούμενο από N λέξεις. Ακολούθως με τη χρήση 3 διαφορετικών ταξινομητών, του SVM, του k-NN και του Naïve Bayes τα δεδομένα που έμειναν κατηγοριοποιήθηκαν ανάλογα με το περιεχόμενο τους. Για την εξέταση των πιο πάνω αλγορίθμων, χρησιμοποιήθηκε συγκεκριμένα το σύνολο Reuters-21578. Χρησιμοποιήθηκαν τόσο η precison όσο και η recall μέθοδοι αξιολόγησης της επίδοσης της ταξινόμησης, καθώς και η RCut μέθοδος για το thresholding.

ΑΞΙΟΛΟΓΗΣΗ ΠΕΙΡΑΜΑΤΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ:

Ταξινομώντας τα δεδομένα Reuters που προήλθαν και από τις 3 μεθόδους, την mMIFS-U, MIFS-U, MIFS με τη χρήση και των 3 ταξινομητών όσο και των 2 κριτηρίων αξιολόγησης της ακρίβειας της μεθόδου, φαίνεται περίτρανα και από τα πιο πάνω γραφήματα (σχήμα 5.9) ότι η διαδοχική προς τα εμπρός μέθοδος mMIFS-U υπερέχει αισθητά των άλλων 2 και για τους 3 διαφορετικούς ταξινομητές στην ακρίβεια της πολυκατηγορικής ταξινομήσης.

5.4 ΓΕΝΙΚΟ ΣΥΜΠΕΡΑΣΜΑ

Πιο πάνω παρουσιάστηκαν και αναλύθηκαν εκτενώς, 3 από τις πιο σύγχρονες μεθόδους επιλογής χαρακτηριστικών που υπάρχουν βασισμένες σε μέτρα πληροφορίας. Πιο συγκεκριμένα και οι 3 μεθόδους, χρησιμοποιούν την αμοιβαία πληροφορία, ενώ οι CMIM και mMIFS-U χρησιμοποιούν επιπλέον και την υπο συνθήκη αμοιβαία πληροφορία για την κατασκευή του αλγορίθμου τους. Α

πό τις πειραματικές διαδικασίες που έγιναν και παρουσιάστηκαν πιο πάνω, είναι φανερό ότι οι mRMR, CMIM και mMIFS-U μεθόδοι υπερταιρούν αισθητά από αντίστοιχες μεθόδους, παρόμοιες με αυτές, τόσο στην ταχύτητα επιλογής των χαρακτηριστικών, όσο και στον περιορισμό του σφάλματος στην ταξινόμηση. Αυτό που προκαλέι τη μεγαλύτερη εντύπωση από όλα, είναι το γεγονός ότι επιτυγχάνουν μικρότερα σφάλματα από τις άλλες μεθόδους, με οποιοδήποτε ταξινομητή και αν συνδυαστούν, για οποιοδήποτε σύνολο δεδομένων και αν δοκιμαστούν και επιπλέον για οποιοδήποτε κριτήριο και αν μετρηθούν.

Εν κατακλείδι και οι 3 αυτές μεθόδοι, δημιουργήθηκαν με σκοπό να καλύψουν κάποιες ατέλεις ή και να τελειοποιήσουν ορισμένα χαρακτηριστικά από μεθόδους

που ήδη προυπήρχαν. Τελικά, όχι μόνο μπορούν να χρησιμοποιηθούν επάξια με τις μεθόδους πάνω στις οποίες στηρίχτηκε η κατασκευή τους, αλλά είναι ικανές να τις αντικαταστήσουν τελείως, προσφέρωντας αξιόπιστες, λειτουργικές και οικονομικές λύσεις σε ένα θεμελιώδες πρόβλημα, όπως είναι αυτό της επιλογής χαρακτηριστικών, από ένα σύνολο.

REFERENCES

[1] Χρ.Κουκουβίνος Α.Παπαιωάννου Βιβλίο «Θεωρια Πληροφοριών και Κωδίκων» (2003)

[2] Aftab, Cheung, Kim, Thakkar, Yeddanapudi PROJECT HISTORY MIT 6933-Final Paper Information Theory and the Digital Revolution- "Information Theory and the Digital Age" (2010)

[3] Ahmed Al Ani, Mohamed Deriche IEEE 1051-4651 pages 82-85 "Feature Selection using a Mutual Information Based Measure" (2002)

[4] B.Senliol, G.Gulgezen, L.Yu, Z.Cataltepe "Fast Correlation based Filter (FCBF) with a Different Search Strategy" (2008)

[5] Benjamin Auffarth, Maite Lopez, Jesus Gerquides ICDM 10th Industrial Conference on Advances in Data Mining pages 248-262 "Comparison of Redundancy and Relevance Measures for Feature Selection in Tissue Classification of CT images" (2010)

[6] C.A.Ratanamahatana, D.Gunopulos Applied Artificial Intelligence vol.17(5-6) pages 475-487 "Feature Selection for the naïve Bayesian classifier using decision trees" (2003)

[7] C.Ding, H.Peng Journal of Bioingormatics and Computational Biology vol.3 No.2 pages 185-205 "Minimum Redundancy Feature Selection from Microarray Gene Expression Data" (2004)

[8] C.J.C. Burges Knowledge Discovery and Data Mining vol.2 pages 1-43 "A Tutorial on Support Vector Machines for Pattern Recognition" (1998)

[9] F. Debole, F Sebastiani Journal of the American Society for Information Science and Technology vol. 56(2) pages 584-596 "An Analysis of the Relative Difficulty of Reuters-21578 Subsets" (2004)

[10] Francois Fleuret Journal of Machine Learning Research 5 pages 1531-1555 "Fast Binary Feature Selection with Conditional Mutual Information" (2004)

[11] Francois Fleuret Technical Report RR-4941 INRIA "Binary Feature Selection with Conditional Mutual Information" (2003)

[12] H. Liu, L, Yu Journal IEEE Transactions on Knowledge and Data Enginnering ISSUE 4 VOL.17 pages 491-502 “Toward Integrating Feature Selection Algorithms for Classification and Clustering” (2005)

[13] H.Liu, X.Wu, S.Zhang CICM’11 20th ACM International Conference on Information and Knowledge Management pages 979-984 “Feature Selection using Hierarchical Feature Clustering” (2011)

[14] Hanchuan Peng, Fuhui Long, Chris Ding IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 27, no. 8 “Feature Selection based on Mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy” (2005)

[15] Inder Jeet Taneja BOOK “ Generalized Information Measures and their Applications” (2001)

[16] Isabelle Guyon, A. Elisseeff Journal of Machine Learning Research 3 pages 1157-1182 “An Introduction to Variable and Feature Selection” (2003)

[17] J.Novovicova, P.Somol, M.Haindl, P.Pudil CIARP-2007 Proceedings of the Congrress on pattern recognition 12th Iberoamerican Conference LNCS-4756 pages 417-426 “Conditional Mutual Information Based Feature Selection for Classification Task” (2007)

[18] L.B Goncalves, J Leonardo Ribeiro Macrini, Pesqui. Oper. Vol 31 no.3 “Renyi entropy and Cauchy-Schwartz Mutual Information applied to MIFS-U rariable selection algorithm: A comparative study” (2011)

[19] L.Yu, H.Liu Proceedings of the International Conference on Maching Learning pages 856-863 “Feature Selection for High Dimensional Data- A Fast Correlation-Based Filter Solution” (2003)

[20] Luis Talavera IDAS’05 Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis pages 440-451 “An evaluation of filter and wrapper methods for Feature Selection in Categorical Clustering”

[21] M Ikonomakis, S. Kotsiantis, V. Tampakas WSEAS Transactions on Computers ISSUE 8 vol.4 pages 966-974 “ Text Classification Using Machine Learnig Techniques” (2005)

[22] N.X Vinh, J Epps, J Bailey Journal of Machine Learning Research 11 pages 2837-2854, " Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance" (2010)

[23] P.Langley, W.Iba, K.Thompson Proceedings of AAAI-92 pages 223-228 Proceedings "An Analysis of Bayesian Classifiers" (1992)

[24] R. Renner, U Maurer Proceedings IEEE International Symposium on page 364 "About the Mutual (Conditional) Information" (2002)

[25] R.Kohavi, G.H.John Artifical Intelligence pages 273-324 "Wrappers for Feature subset Selection" (1996)

[26] S.Balakrishnama, A.Ganapathiraju Institude for Signal anf Information Processing "Linear Discriminant Analysis- A Brief Tutorial" (1998)

[27] Thomas M. Cover, Joy A. Thomas "Elements of Information Theory" Chapter 2 "Entropy, Relative Entropy and Mutual Information" (1991)

[28] Z Zhaug, Edwin R. Hancock SIMBAD LNCS-7005 pages 235-249 "Mutual Information Criteria for Feature Selection" (2011)