ΤΜΉΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΏΝ ΚΑΙ ΜΗΧΑΝΙΚΏΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Τεχνικές ανάλυσης δεδομένων με classification και clustering»

Ονοματεπώνυμο

Μάρκου Γιώργος

Επιβλέπων καθηγητής: Ιωάννης Τζήμας

Πάτρα – Μάιος 2021

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Πάτρα, Ημερομηνία

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Ονοματεπώνυμο, Υπογραφή
2. Ονοματεπώνυμο, Υπογραφή
3. Ονοματεπώνυμο, Υπογραφή

Αφιέρωση

Η πτυχιακή αυτή εργασία είναι αφιερωμένη στους γονείς μου, Ανδρέα και Δήμητρα.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον κ. Αναστάσιο Δροσόπουλο νυν πρόεδρο του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου, τον πρώην πρόεδρο του τμήματος κ.Μιχάλη Παρασκευά καθώς και τον κ. Ιωάννη Τζήμα, καθηγητή του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου, για την πολύτιμη βοήθεια και καθοδήγηση που μου προσέφεραν καθ' όλη τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας αλλά και κατά την διάρκεια των σπουδών μου.

Περιεχόμενα

Λίστα Σχημάτων

Λίστα Πινάκων

Πρόλογος

Ο σκοπός αυτού του εγγράφου είναι να παρέχει μια εννοιολογική εισαγωγή στις τεχνικές στατιστικής ή μηχανικής μάθησης (ML) που κανονικά δεν θα εκτίθενται σε τέτοιες προσεγγίσεις κατά τη διάρκεια της τυπικής απαιτούμενης στατιστικής εκπαίδευσης. Η μηχανική εκμάθηση μπορεί να περιγραφεί ως μια μορφή στατιστικής ανάλυσης, συχνά ακόμη και χρησιμοποιώντας γνωστές και άγνωστες τεχνικές, που έχουν λίγο διαφορετική εστίαση από την παραδοσιακή αναλυτική πρακτική σε εφαρμοσμένους κλάδους. Η βασική ιδέα είναι ότι οι ευέλικτες, αυτόματες προσεγγίσεις χρησιμοποιούνται για την ανίχνευση μοτίβων στα δεδομένα, με πρωταρχική εστίαση στην πραγματοποίηση προβλέψεων για μελλοντικά δεδομένα.

Εάν κάποιος ερευνήσει τον αριθμό των διαθέσιμων τεχνικών στο ML χωρίς κανένα πλαίσιο, μπορεί κανείς εύκολα να κατακλυστεί όσον αφορά τον τεράστιο αριθμό προσεγγίσεων, καθώς και τις διάφορες τροποποιήσεις και παραλλαγές αυτών. Ωστόσο, οι ιδιαιτερότητες των τεχνικών δεν είναι τόσο σημαντικές όσο οι γενικότερες έννοιες που θα μπορούσαν να εφαρμοστούν στις περισσότερες ρυθμίσεις ML, και μάλιστα σε πολλές παραδοσιακές τεχνικές.

Όσον αφορά την προαπαιτούμενη γνώση, θα υποθέσω μια βασική εξοικείωση με τις αναλύσεις παλινδρόμησης που συνήθως παρουσιάζονται σε εφαρμοσμένους κλάδους. Όσον αφορά τον προγραμματισμό, κανένας δεν απαιτείται πραγματικά να ακολουθήσει το μεγαλύτερο μέρος του περιεχομένου εδώ.

Περίληψη

Η εξόρυξη δεδομένων και οι αλγόριθμοι μάθησης είναι ένας σημαντικός κλάδος της επιστήμης των υπολογιστών με αντικείμενο την ανακάλυψη ή εύρεση ή παραγωγή λειτουργικής γνώσης μέσω της ανάλυσης δεδομένων από μεγάλες αποθήκες δεδομένων και την εύρεση δομών που αναδεικνύουν την γνώση.

Οι εφαρμογές της εξόρυξης δεδομένων είναι ποίκιλες. Η εξόρυξη δεδομένων είναι η διαδικασία που περιλαμβάνει τον εντοπισμό καίριων και καινοτόμων βημάτων ή προτύπων, τα οποία παρουσιάζουν ενδιαφέρον, καθώς και τη δημιουργία περιγραφικών κατανοητών και προβλεπτικών μοντέλων από δεδομένα μεγάλης κλίμακας.

Σκοπός αυτής της εργασίας είναι να αναδείξει την σημαντικότητα της κατηγοριοποίησης και της ομαδοποίησης των δεδομένων είτε είναι μεγάλα δεδομένα είτε ένα σύνολο κάποιων εκατοντάδων γραμμών αποτελούμενα από έναν αριθμό στηλών ή χαρακτηριστικών(attributes).

Αρχικά, με τον όρο κατηγοριοποίηση (classification) αναφερόμαστε στην πρόβλεψη της ετικέτας μιας κατηγορίας για ένα καθορισμένο μη σημασμένο σημείο. Στην συγκεκριμένη ενότητα θα μελετήσουμε 2 παραδείγματα της πιθανοτητικής μεθοδολογίας για την κατηγοριοποίηση. Πρώτα θα μελετηθεί ο πλήρης κατηγοριοποιητής Bayes που χρησιμοποιεί το θεώρημα του Bayes για να προβλέψει ότι η ζητούμενη κατηγορία είναι εκείνη που μεγιστοποιεί την εκ των υστέρων πιθανότητα. Και στην συνέχεια θα δούμε και θα περιγράψουμε τον κατηγοροιοποιητή πλησιέστερων γειτόνων (nearest neighbors classifier), ο οποίος στηρίζεται σε μια μη παραμετρική μέθοδο για την εκτίμηση της πυκνότητας.

Επίσης, στα πλαίσια της παρούσας εργασίας αυτής έγινε μελέτη της μαθηματικής μοντελοποίησης και ανάλυσης αλγορίθμων. Επίσης, διερευνήθηκαν τα μαθηματικά μοντέλα που εφαρμόζονται στο συγκεκριμένο πεδίο και μελετήθηκαν διάφοροι αλγόριθμοι συσταδοποίησης.

Λέξεις κλειδιά: Κατηγοριοποίηση, Συσταδοποίηση, Μηχανική Μάθηση, Εξόρυξη Δεδομένων

Abstract

Data mining and learning algorithms are an important branch of computer science with the object of discovering or finding or producing functional knowledge through the analysis of data from large data warehouses and the finding of structures that highlight knowledge.

Data mining applications are diverse. Data mining is the process of involving key and innovative steps or patterns that are of interest, as well as creating descriptive comprehensible and predictive models from large-scale data.

The purpose of this paper is to highlight the importance of categorizing and grouping data whether it is large data or a set of hundreds of rows consisting of a number of columns or attributes.

Initially, the term classification refers to the prediction of a category label for a specified unmarked point. In this section we will study 2 examples of probabilistic methodology for categorization. The complete Bayes categorizer that uses Bayes' theorem to predict that the requested category is the one that maximizes the ex post probability will be studied first. And then we will see and describe the nearest neighbor's classifier, which is based on a non-parametric method for estimating density.

Also, in the context of the present work, a study of mathematical modeling and algorithm analysis was made. Also, the Mathematical models applied in the specific field were investigated and various clustering algorithms were studied.

Keywords: Classification, Clustering, Machine Learning, Data Mining

1 Εξόρυξη και ανάλυση δεδομένων

Στην σημερινή εποχή του 21ου αιώνα, που χαρακτηριστικά ο συγκεκριμένος αιώνας θεωρείται από πολλούς η «χρυσή» εποχή των δεδομένων. Υπάρχει ένας μεγάλος «όγκος δεδομένων», που εντοπίζεται σε αποθήκες (data warehouses) και βάσεις δεδομένων, ωστόσο δεν μπορεί να χρησιμοποιηθεί χωρίς να έχει προηγηθεί κάποιου είδους επεξεργασία [1]. Έτσι μέσω της εξόρυξης δεδομένων ή αλλιώς μέσω της εξόρυξης γνώσης από βάσεις δεδομένων προκύπτει η εύρεση της πληροφορίας ή των προτύπων από βάσεις δεδομένων με τη βοήθεια αλγορίθμων ομαδοποίησης (clustering) ή κατηγοριοποίησης (classification) [2] και των διαφόρων στατιστικών μέτρων, της τεχνητής νοημοσύνης (Artificial Intelligence), της μηχανικής μάθησης (Machine Learning) και των συστημάτων βάσεων δεδομένων (System of DB).

Προσπάθεια όλων των ερευνητών είναι η συλλογή μεγάλου όγκου πληροφορίας και εξαγωγή ασφαλών συμπερασμάτων για την πληροφορία που αποτυπώνεται για την χρησιμότητα που θα επέλθει αργότερα [3], ώστε να συνεισφέρει στην επιστημονική κοινότητα, η οποία έχει ως πρωταρχικό στόχο την βελτίωση των συνθηκών ζωής.

Το κύριο μέλημα της εξόρυξης δεδομένων είναι μέσω αλγορίθμων διαφόρων κατηγοριών είτε clustering είτε classification να δημιουργείται μια αυτόματη ανάλυση του μεγάλου όγκου δεδομένων που υπάρχει για την γρήγορη και εύκολη εξαγωγή κάποιου ιδιαίτερου χαρακτηριστικού ή προτύπου [4], εννοώντας ως πρότυπο έναν συνδυασμό χαρακτηριστικών ή στην γλώσσα των μαθηματικών ένα διάνυσμα με συνιστώσες, όπου κάθε συνιστώσα θα είναι και ένα χαρακτηριστικό.

Τα δεδομένα μπορούν συχνά να αναπαρασταθούν, συγκεκριμένα αποτελούμενα από όλα τα στοιχεία ή γενικά με την πιο αφαιρετική πληροφορία και χωρίς κάποια εμβάθυνση ή αλλιώς αφηρημένα με μια μήτρα δεδομένων διαστάσεων (n X m) [5], με γραμμές n και στήλες m, όπου οι γραμμές ή αλλιώς εγγραφές να αναπαριστούν οντότητες του συνόλου δεδομένων και οι στήλες τα χαρακτηριστικά (attributes) ή ιδιότητες που μας ενδιαφέρουν [6]. Σε κάθε γραμμή μιας μήτρας δεδομένων αποθηκεύονται αντίστοιχα και οι τιμές των γνωρισμάτων για μια καθορισμένη οντότητα. Όπως φαίνεται και στον παρακάτω πίνακα διακριτά τα στοιχεία

εγγραφές [7] αποτελούμενα από τα χαρακτηριστικά που το σύνολο των χαρακτηριστικών είναι ένα διάνυσμα που διακρίνει μοναδικά την κάθε εγγραφή του συνόλου δεδομένων.

Α/Α	Α	Β	Γ
1	1	2	1
2	2	2	1
3	3	1	2
4	3	3	2
5	4	3	2
6	4	1	2
:	:	:	:
136	6	6	2
137	4	1	1
138	4	1	2

Πίνακας 1: Πίνακας χαρακτηριστικών

Σε αυτό σημείο, να τονίσουμε ότι σε όλο τον κόσμο, η ύπαρξη των συνόλων δεδομένων δεν έχουν και απαραίτητα την μορφή μήτρας. Πιο συγκεκριμένα, στον χώρο της ιατρικής [8], όπου ο όγκος των δεδομένων είναι μεγάλος, άλλα και τα περισσότερα δεδομένα, φέρουν σημαντική πληροφορία που είναι δύσκολο να γίνει κατανοητή άμεσα από τους ειδικούς, παρά μόνο από την κατάλληλη μελέτη [9], μπορεί να γίνει ένας μετασχηματισμός που αποτελεί τα πρώτα βήματα της ανάλυσης δεδομένων και δεν είναι άλλη από την εξαγωγή χαρακτηριστικών. Έτσι, πληροφορία που θεωρείται περιττή και πράγματι είναι [10], τότε δεν λαμβάνεται υπόψιν στην συνέχεια της ανάλυσης ενός συνόλου δεδομένων.

1.1 Γνωρίσματα

Τα γνωρίσματα μπορούν να διακριθούν σε 2 κατηγορίες με βάση το πεδίο ορισμού τους [11] και ανάλογα με τον τύπο των τιμών που παίρνουν κατατάσσονται στην αντίστοιχη ομάδα γνωρισμάτων.

Σε **αριθμητικά** πεδία, μπορείτε να υπολογίσετε τα εξής [12]:

- Πλήθος—Υπολογίζει τον αριθμό τιμών που δεν είναι null. Μπορεί να χρησιμοποιηθεί σε αριθμητικά πεδία ή σε συμβολοσειρές. Το πλήθος των [null, 0, 1, 2] είναι 3.
- Άθροισμα—Το άθροισμα των αριθμητικών τιμών σε ένα πεδίο. Το άθροισμα των [null, null, 3] είναι 3.
- Μέση τιμή—Η μέση τιμή αριθμητικών τιμών. Η μέση τιμή των [0, 2, null] είναι 1.
- Ελάχιστο—Η ελάχιστη τιμή ενός αριθμητικού πεδίου. Το ελάχιστο των [0, 2, null] είναι 0.
- Μέγιστο—Η μέγιστη τιμή ενός αριθμητικού πεδίου. Η μέγιστη τιμή των [0, 2, null] είναι 2.
- Εύρος—Το εύρος ενός αριθμητικού πεδίου. Αυτό υπολογίζεται ως οι ελάχιστες τιμές που αφαιρούνται από την μέγιστη τιμή. Το εύρος των [0, null, 1] είναι 1. Το εύρος των [null, 4] είναι 0.
- Διακύμανση—Η διακύμανση ενός αριθμητικού πεδίου σε ένα ίχνος. Η διακύμανση του [1] είναι null. Η διακύμανση των [null, 1, 1, 1] είναι 1.
- Τυπική απόκλιση—Η τυπική απόκλιση ενός αριθμητικού πεδίου. Η τυπική απόκλιση του [1] είναι null. Η τυπική απόκλιση των [null, 1, 1, 1] είναι 1.

Ενώ από την άλλη πλευρά το **κατηγορικό** γνώρισμα έχει ως πεδίο ορισμού, εκείνες τις τιμές που αντιπροσωπεύουν ένα πλήθος συμβόλων ή γραμμάτων [13]. Για παράδειγμα, το φύλο (Sex), θα μπορούσε να αποτελέσει ένα κατηγορικό γνώρισμα με το πεδίο ορισμού του να είναι είτε M(Male) είτε F(Female).

Επίσης, να τονιστεί ότι 2 είναι οι τύποι των κατηγορικών γνωρισμάτων[14]:

Ονομαστικά (nominal): Οι τιμές του πεδίου ορισμού ενός γνωρίσματος δεν είναι διατεταγμένες και νόημα μπορούν να έχουν μόνο οι συγκρίσεις για ενδεχόμενη ισότητα. Πιο συγκεκριμένα, μπορούμε να ελέγχουμε αν η τιμή του γνωρίσματος για 2 δεδομένα είναι ίδια ή όχι.

Διατακτικά (ordinal): Οι τιμές του γνωρίσματος είναι διατεταγμένες, αυτό σημαίνει ότι επιτρέπεται τόσο η σύγκριση για ισότητα όσο και η σύγκριση για ανισότητες, αν και ενδέχεται και υπάρχει η πιθανότητα ο προσδιορισμός της ποσοτικής διαφοράς των τιμών να μην είναι εφικτός.

1.2 Πιθανοτική θεώρηση

Στο συγκεκριμένο κεφάλαιο θα παρουσιαστούν οι έννοιες της τυχαίας μεταβλητής, της συνάρτησης κατανομής [15], της συνάρτησης πυκνότητας πιθανότητας και άλλες [16], όπως θα μελετηθούν παρακάτω.

Η Θεωρία τών Πιθανοτήτων χρησιμοποιείται σε διάφορες εφαρμογές και καθώς η ορολογία δεν είναι κοινή ορισμένες φορές προκαλείται σύγχυση [17]. Οι παρακάτω όροι χρησιμοποιούνται για μη αθροιστική συνάρτηση κατανομής πιθανοτήτων [18-20]:

Συνάρτηση μάζας: Χρησιμοποιείται για διακριτές τυχαίες μεταβλητές (μη συνεχείς τιμές, δηλαδή χωρίς την έννοια του ορισμού «διαστήματος»).

Κατηγορική Κατανομή: Για διακριτές τυχαίες μεταβλητές με πεπερασμένο σύνολο τιμών.

Συνάρτηση Πυκνότητας : Χρησιμοποιείται για συνεχείς μεταβλητές, στις οποίες ορίζεται η έννοια του διαστήματος και των πραγματικών αριθμών.

Ωστόσο, οι προαναφερόμενοι όροι, δεν είναι και πλήρως εφαρμόσιμοι, ανάλογα με τα δεδομένα προς ανάλυση από τον αναλυτή, γιατί μπορεί να γίνεται αναφορά σε κατανομές όπως είναι οι αθροιστικές και μη αθροιστικές κατανομές.

Συνάρτηση κατανομής πιθανότητας: Είναι μια τυχαία μεταβλητή για να μπορέσουμε να περιγράψουμε τα ενδεχόμενα που μας ενδιαφέρουν σε έναν συγκεκριμένο χώρο, ο οποίος ορίζεται εντός του δειγματικού χώρου Ω, όπως είναι το διάγραμμα Venn. Στην συνέχεια με την συλλογή των διαφόρων στατιστικών μέτρων (όπως της ένωσης ή της τομής και των υπολοίπων πιθανοτικών μοντέλων) να υπολογιστεί η αντίστοιχη πιθανότητα.

Συνάρτηση Πιθανότητας: Χρησιμοποιείται για ένα τυχαίο πείραμα όπως για παράδειγμα είναι η ρίψη ενός νομίσματος για n φορές και έχουμε την μια τυχαία μεταβλητή Υ που λειτουργεί ως ο τρόπος μέτρησης του αριθμού των κεφαλών ή των κορωνών αντίστοιχα στις n ρίψεις.

Κατανομή πιθανότητας: Χρησιμοποιείται στις Πιθανότητες αλλά και την Στατιστική, καθώς παριστάνει την πιθανότητα για κάθε μετρήσιμο σύνολο ή υποσύνολο των διαφόρων πιθανών αποτελεσμάτων ενός τυχαίου πειράματος.

Επιπλέον τόσο στην θεωρία όσο και στην πράξη, συχνά προκύπτουν καταστάσεις κατά τις οποίες οι τυχαίες μεταβλητές που επρόκειτο να χρησιμοποιήσουμε είναι συνεχείς και όχι διακριτές [21], μπορούν δηλαδή να πάρουν οποιαδήποτε τιμή σε ένα δεδομένο διάστημα [22] (η τιμή αυτή εξαρτάται από το αποτέλεσμα του τυχαίου πειράματος).

1.3 Εξόρυξη δεδομένων

Η εξόρυξη δεδομένων περιλαμβάνει τους βασικούς αλγορίθμους που μας επιτρέπουν να εξάγουμε θεμελιώδεις πληροφορίες και γνώσεις σχετικά με τα σύνολα [23] δεδομένων που διαχειρίζονται από την επιστημονική κοινότητα της ανάλυσης δεδομένων [24]. Αποτελεί ένα πεδίο που καλύπτει πολλούς επιστημονικούς κλάδους, συνδυάζοντας έννοιες από συναφείς τομείς όπως τα συστήματα βάσεων δεδομένων, η στατιστική, η μηχανική μάθησης και η αναγνώριση μοτίβων.

Η αλγεβρική, η γεωμετρική και η πιθανοτητική θεώρηση των δεδομένων παίζουν καθοριστικό ρόλο στην εξόρυξη δεδομένων. Στην πραγματικότητα, η εξόρυξη δεδομένων αποτελεί κομμάτι μιας μεγαλύτερης διαδικασίας ανεύρεσης γνώσεων, [25] η οποία περιλαμβάνει τόσο προπαρασκευαστικές ενέργειες, όπως είναι η εξαγωγή χαρακτηριστικών όσο και τα βήματα που θα ακολουθήσει κανείς για την βασική ερμηνεία των δεδομένων [26].

Η διερευνητική ανάλυση δεδομένων αξιοποιεί με μεμονωμένο ή συνδυαστικό τρόπο τα αριθμητικά και κατηγορικά γνωρίσματα των δεδομένων, επιδιώκοντας να εξαγάγει βασικά χαρακτηριστικά από το σύνολο δεδομένων [27] με την βοήθεια στατιστικών που παρέχουν πληροφορίες για την κεντρικότητα, την διασπορά και οτιδήποτε μπορεί να φανεί χρήσιμο από τα στατιστικά μέτρα [28] που υπάρχουν μέγιστο ή ελάχιστο κ.ο.κ.

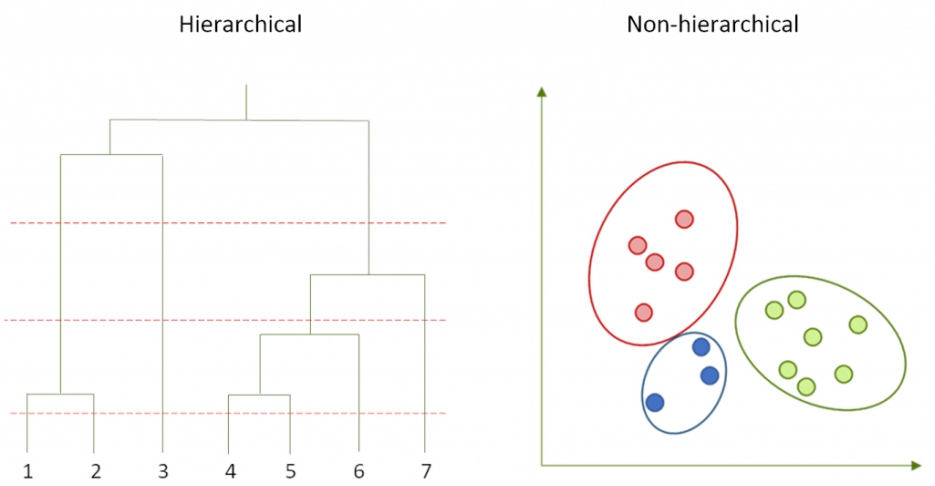
Η εξόρυξη συχνών μοτίβων αναφέρεται στο έργο της εξαγωγής διαφωτιστικών και χρήσιμων μοτίβων από τεράστια και σύνθετα σύνολα δεδομένων [29]. Στα μοτίβα περιλαμβάνονται σύνολα δεδομένων ή τιμών που είναι γνωρίσματα, που απλώς εμφανίζονται ταυτόχρονα και ονομάζονται στοιχειοσύνολα(itemsets), είναι πιο πολύπλοκα μοτίβα όπως οι ακολουθίες, όπου λαμβάνονται υπόψη ρητές σχέσεις προτεραιότητας και τα γραφήματα, όπου λαμβάνονται υπόψιν [30] ρητές αυθαίρετες σχέσεις μεταξύ σημείων.

Παραδείγματα εφαρμογών της Εξόρυξης Δεδομένων είναι [166] η ανάλυση συσχέτισης, η ταξινόμηση, η παλινδρόμηση, η ομαδοποίηση και άλλα. Με βάση τις μεθόδους τους, η εφαρμογή της εξόρυξης δεδομένων μπορεί να χωριστεί σε τρία τμήματα που είναι η εποπτευόμενη μάθηση, η μη εποπτευόμενη μάθηση και η ημι-εποπτευόμενη μάθηση [167] . *Η εποπτευόμενη μάθηση* χρησιμοποιεί την εξόρυξη δεδομένων όταν υπάρχει ένας επόπτης για να παρατηρήσει πώς λειτουργεί ο αλγόριθμος. *Η μη επιτηρούμενη μάθηση* λειτουργεί χωρίς να επιτηρείται από τον επόπτη, οπότε προέρχεται αποκλειστικά από τη δουλειά του υπολογιστή. Ένα από τα παραδείγματα της μη εποπτευόμενης μάθησης είναι η ομαδοποίηση. *Η ημι-εποπτευόμενη μάθηση* είναι η χρήση της εξόρυξης δεδομένων όπου λίγα δεδομένα επισημαίνονται [167].Γενικά, η εξόρυξη δεδομένων μπορεί να ταξινομηθεί ως περιγραφή και πρόβλεψη. Η διαδικασία της περιγραφής γίνεται για να βρεθεί ένα μοτίβο που

είναι κατανοητό, ενώ η διαδικασία πρόβλεψης είναι για την πρόβλεψη κατά βάσει των δεδομένων που χρησιμοποιούνται .

Η συσταδοποίηση ή ομαδοποίηση [142] είναι η διαδικασία της ομαδοποίησης ενός συνόλου αντικειμένων με τέτοιο τρόπο, ώστε αντικείμενα που βρίσκονται στην ίδια ομάδα (συστάδα ή cluster) να είναι πιο όμοια (με κάποιο κριτήριο) μεταξύ τους, απ' ότι με τα αντικείμενα των άλλων ομάδων [31]. Η διαφορά μεταξύ της συσταδοποίησης δεδομένων και της ταξινόμησης δεδομένων (data classification) είναι [32] οι ομάδες που κατά την ταξινόμηση θα τοποθετηθούν τα δεδομένα.

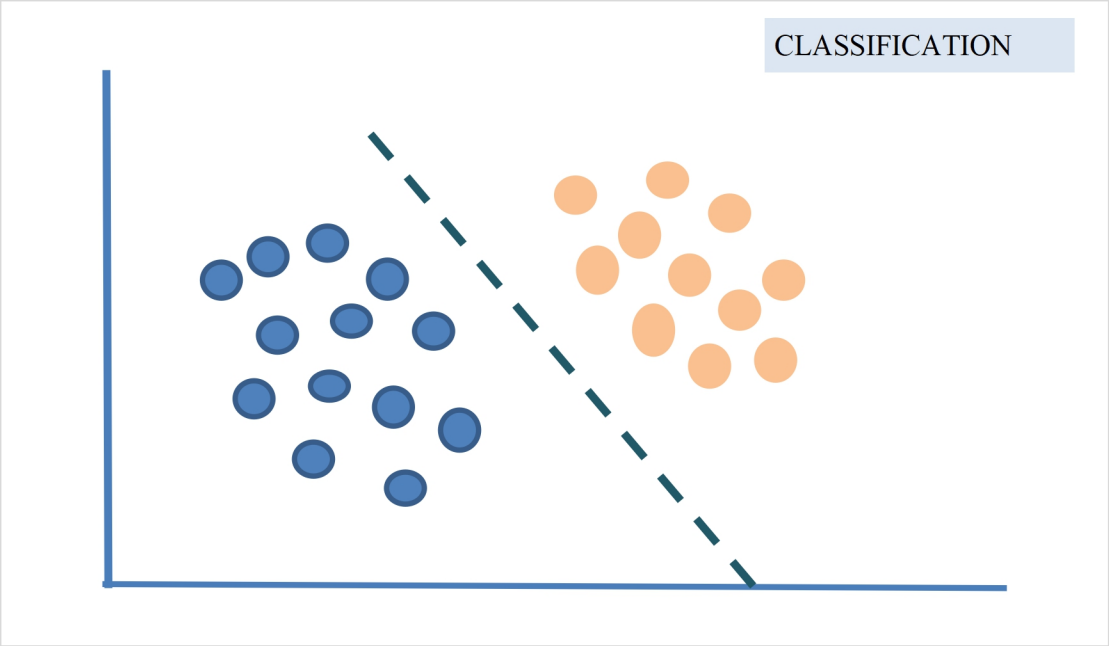
Επομένως, η εξόρυξη δεδομένων είναι η διαδικασία χειρισμού πληροφοριών από μια βάση δεδομένων η οποία δεν είναι άμεσα ορατή. Βιομηχανίες όπως η τραπεζική, η ασφάλιση και η ιατρική χρησιμοποιούν συνήθως την εξόρυξη δεδομένων για τη μείωση του κόστους, τη βελτίωση της έρευνας και την αύξηση των πωλήσεων. Οι τεχνικές ανάλυσης δεδομένων χρησιμοποιούνται [138] παραδοσιακά για τέτοιες εργασίες, όπως ανάλυση παλινδρόμησης, ανάλυση συμπλέγματος, αριθμητική ταξινόμηση, πολυδιάστατη ανάλυση, άλλες πολυπαραγοντικές στατιστικές μεθόδους, στοχαστικά μοντέλα, ανάλυση χρονοσειρών, μη γραμμικές τεχνικές εκτίμησης και άλλες. Η εξόρυξη δεδομένων [139] είναι μια επιστήμη που απαιτείται για την προώθηση της επιστήμης στον τομέα της τεχνητής νοημοσύνης και των στατιστικών και προβλέπεται να γίνει ένας εξαιρετικά επαναστατικός κλάδος της επιστήμης την επόμενη δεκαετία.



Σχήμα 1: Ιεραρχική και Μη-Ιεραρχική Διαχώριση

Επομένως, κάτι τέτοιο σημαίνει όπως παρατηρείται και από την παραπάνω εικόνα πώς εκ' των προτέρων είναι γνωστός ο αριθμός των ομάδων και αντίστοιχα τα διάφορα αντικείμενα ταξινομούνται εντός του εύρους των ομάδων [34] με βάσει την ομοιότητας τους και την συνάρτηση που εφαρμόστηκε για το όριο ομοιότητας των δεδομένων.

Κατηγοριοποίηση ονομάζεται η διαδικασία εκμάθησης μιας συνάρτησης στόχου (target function) f (μοντέλο) [35] κατά την οποία απεικονίζεται κάθε σύνολο γνωρισμάτων x σε μια από τις ήδη καθορισμένες ετικέτες κλάσεις y.



Σχήμα 2: Κατηγοριοποίηση

Η κατηγοριοποίηση αποτελεί αν όχι την πιο βασική κατηγορία ή μελέτη στην εξόρυξη ή την μηχανική μάθηση, άλλα συναποτελεί μια από τις πιο βασικές λειτουργίες στην εξόρυξη δεδομένων (data mining) με πλήθος εφαρμογών στον χώρο των οικονομικών [36] αλλά και της ιατρικής. Αποτελεί εργασία επιβλεπόμενης μάθησης (supervised learning), [37] που στόχεύει στην ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα ετικέτας ή στόχου (target attribute/class label) με ονομαστικές τιμές και σε ένα σύνολο άλλων γνωρισμάτων του συνόλου δεδομένων.

Ακόμη μια βασική μελέτη των συνόλων δεδομένων στην επιβλεπόμενη μάθηση είναι η Παλινδρόμηση [38], που έχει ώς στόχο την πρόβλεψη αριθμητικών τιμών.

Στην διαδικασία της κατηγοριοποίησης πραγματοποιείται ένας επαγωγικός αλγόριθμος και κατασκευάζεται ένα μοντέλο, έπειτα από την εξαγωγή χαρακτηριστικών [39], την μελέτη αλλά και την εφαρμογή αλγορίθμου για την εκπαίδευση του συνόλου δεδομένου με ένα συγκεκριμένο σύνολο δεδομένων, έτσι ώστε να μπορεί να εφαρμοστεί και σε πλήθος άλλων συνόλων δεδομένων. Η διαδικασία της κατηγοριοποίησης περιλαμβάνει τρία στάδια.

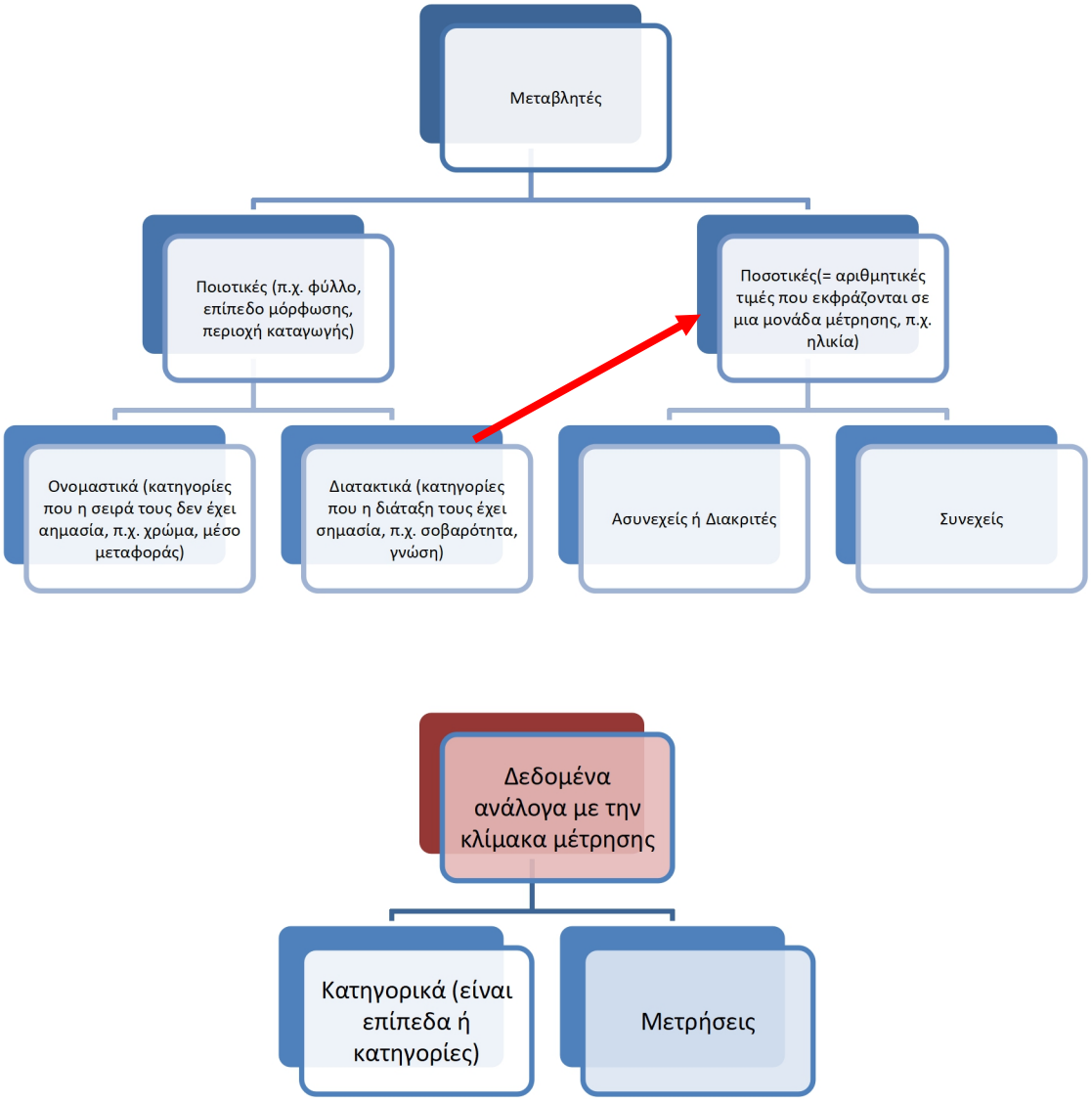
Επομένως, η συγκεκριμένη κατηγορία για την αναγνώριση προτύπων χρησιμοποιεί αρχικά μία διαδικασία προεπεξεργασίας των δεδομένων και έπειτα στη συνέχεια εφαρμόζεται ένας μηχανισμός εξαγωγής των χαρακτηριστικών που δεν ενδιαφέρουν έναν αναλυτή του οποιοδήποτε χώρου κάποια συγκεκριμένα χαρακτηριστικά [40], τα οποία δεν φέρουν κάποια σημαντική πληροφορία.Η μετέπειτα συνέχεια και το επόμενο βήμα που ακολουθεί είναι η εφαρμογή ενός αλγορίθμου ταξινόμησης [41] προκειμένου να εκπαιδευτεί με ένα σύνολο δεδομένων έτσι ώστε να μπορεί να εφαρμοστεί και με άλλα σύνολα δεδομένων. Επιπλέον θα πρέπει να τονιστεί ότι στο προηγούμενο βήμα που αναφέρθηκε, θα πρέπει να λαμβάνεται υπόψη να μην υπάρχει η πλήρη προσαρμογή του συγκεκριμένου αριθμού [42] πάνω στο σετ εκπαίδευσης αλλά να υπάρχει μία μέτρια απόδοση προς καλή, έτσι ώστε να μπορεί να εφαρμοστεί και σε ένα άλλο οποιοδήποτε σύνολο δεδομένων άγνωστο για τον αλγόριθμο ταξινόμησης που εκπαιδεύτηκε στο σύνολο εκπαίδευσης.

2 Εισαγωγή στην ανάλυση δεδομένων

Όταν είτε τα δεδομένα είτε τα μοντέλα περιλαμβάνουν συναρτήσεις, και όταν επιτρέπονται μόνο αδύναμες παραδοχές σχετικά με αυτές τις λειτουργίες όπως η ομαλότητα, [43] πρέπει να τροποποιηθούν γνωστές στατιστικές μέθοδοι και να αναπτυχθούν νέες προσεγγίσεις για να επωφεληθούν από αυτήν την ομαλότητα.

Έτσι λοιπόν η εισαγωγή στην ανάλυση δεδομένων αποτελεί μια πρόγευση στην περαιτέρω επέκταση τεχνικών, μεθόδων και αλγορίθμων που εφαρμόζονται προκειμένου να υπάρξει ένα αποτέλεσμα [44]. Αυτό το αποτέλεσμα μπορεί να χρησιμοποιηθεί για κάποια μελέτη ή κάποια βοήθεια κυρίως όπως είναι γνωστό στον ιατρικό κόσμο ώστε να αποφεύχθει απο κάποια

ιατρική μονάδα κάποιο λάθος, το οποίο μπορεί να στοιχήσει την ζωή ενός ανθρώπου κάτι το οποίο σίγουρα δεν επιθυμούν οι οργανισμοί ιατρικής φροντίδας.



Σχήμα 3: Διαχωρισμός μεταβλητών

Ωστόσο δεν είναι η εφαρμογή μόνο στον ιατρικό κόσμο είναι και στον οικονομικό κλάδο, στον οποίο επιθυμούν να γνωρίζουν την μελλοντική εξέλιξη των μετοχών και όλων των ποσοδεικτών [45]. Όλα τα προαναφερθέντα, έχουν μεγάλο βάθος ανάλυσης, αλλά σκοπός της παρούσας μελέτης μας είναι να αναδείξουμε την αρχή, που αποτελεί και το ήμισυ του παντός.

Πρέπει αρχικά κάποιος να αναγνωρίζει τα δεδομένα, να τα διακρίνει και να τα διαχωρίζει κατάλληλα, [46] ώστε και η συνέχεια της μελέτης να οδηγείται σε εξαχθέντα αποτελέσματα που μπορούν να υποστηριχθούν και να εξάγουν κάποια πολύτιμη πληροφορία για τον εκάστοτε οργανισμό που ενδιαφέρεται.

2.1 Αριθμητικά γνωρίσματα

Στα αριθμητικά γνωρίσματα θα δούμε τις βασικές στατιστικές μεθόδους για την διερευνητική ανάλυση δεδομένων.

Αρχικά διακρίνουμε την μονομεταβλητή ανάλυση που εστιάζει σε ένα μόνο γνώρισμα τη φορά και στην ουσία εκλαμβάνεται υπόψιν μόνο μια στήλη ενός συνόλου δεδομένων. Με ένα τέτοιο χαρακτηριστικό μπορούμε να εξάγουμε πολύτιμη πληροφορία που μπορεί να φαίνεται αρχικά σαν κάτι το οποίο είναι πολύ εύκολο να εφαρμοστεί, [47] όμως στην πραγματικότητα είναι το πρώτο βήμα και βασικό για να αναγνωρίζει κανείς τι είναι χαρακτηριστικό και τί αποτελεί την στήλη ενός συνόλου δεδομένων. Όπως προείπαμε, όταν λέμε για αριθμητικά γνωρίσματα, εννοούμε ότι μπορούν να εφαρμοστούν τύποι εύρεσης μέσης τιμής, εύρεσης του μέτρου διασποράς, [48] δηλαδή σε τι διαστήματα αριθμητικών τιμών εντοπίζουμε τις διάφορες τιμές ενός χαρακτηριστικού.

Αντίστοιχα και στην διμεταβλητή ανάλυση εστιάζει σε 2 γνωρίσματα και στην ουσία εκλαμβάνονται υπόψιν μόνο 2 στήλες ενός συνόλου δεδομένων [49]. Σε αυτή την περίπτωση έχουμε πιο πολύπλοκες μεθόδους, καθώς υπάρχει η έννοια της σύνθεσης των 2 στηλών και περισσότεροι υπολογισμοί για την εξαγωγή της πληροφορίας.

Επομένως γίνεται αντιληπτό ότι στην πολυμεταβλητή ανάλυση δεν είναι επιτρεπτή η άμεση επαφή του ερευνητή μιας μελέτης με τα δεδομένα της, [50] ταυτόχρονα τα αποτελέσματα της μελέτης δεν γίνονται κατανοητά εύκολα και άμεσα από τους αναγνώστες, καθώς η πλειοψηφία αυτών δεν είναι τόσο εξοικειωμένη με τα μαθηματικά μοντέλα τα οποία χρησιμοποιούνται. Επομένως [51], τόσο οι μη ειδικοί όσο και οι ίδιοι οι ερευνητές της μελέτης

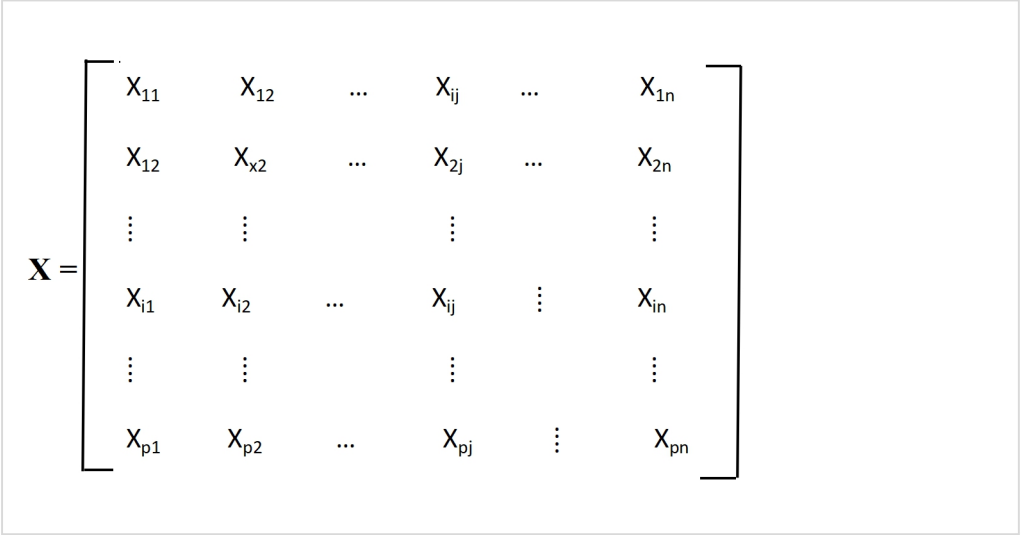
που παρόλο την άμεση επαφή τους με το αντικείμενο αντιλαμβάνονται πιο ευκολά και με μεγαλύτερη σαφήνεια τα δεδομένα όταν αυτά παρουσιάζονται οργανωμένα όπως με τη μορφή συχνοτήτων σε πίνακες (tabular analysis).

Άρα γίνεται αντιληπτό το γεγονός ότι η στατιστική ανάλυση πολυμεταβλητών δεδομένων είναι υπολογιστικά επίπονη και δύσκολα εξαγώγιμη [52] από τους ειδικούς. Για το λόγο αυτό, η ανάλυση αυτή γίνεται με την χρήση Η/Υ μέσω κατάλληλου λογισμικού ή έξυπνων αλγορίθμων μηχανικής μάθησης. Οι τεχνικές πολυμεταβλητών δεδομένων εφαρμόζονται π.χ. στην βιοστατιστική και στην φαρμακολογία, στα οικονομικά – χρηματοοικονομικά [53], στην εκπαίδευση για την μελέτη ενδιαφερόντων στατιστικών μέτρων για την μάζα των μελών που συναποτελούν την εκπαίδευση οποιαδήποτε βαθμίδας.

	Άτομο 1	Άτομο 2		Άτομο j		Άτομο n
Μεταβλητή 1:	x_{11}	x_{12}	…	x_{ij}	…	x_{1n}
Μεταβλητή 2:	x_{21}	x_{x2}	…	x_{2j}	…	x_{2n}
	…	…		…		…
Μεταβλητή i:	x_{i1}	x_{i2}		x_{ij}		x_{in}
	…	…		…		…
Μεταβλητή p:	x_{p1}	x_{p2}	…	x_{pj}	…	x_{pn}

Πίνακας 2: Μεταβλητές Δεδομένων

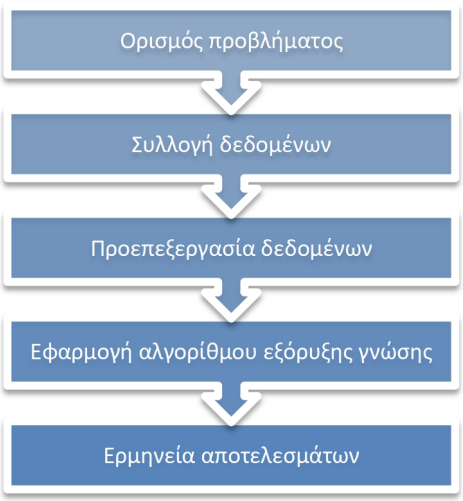
Παρακάτω παρουσιάζονται σε μορφή πίνακα για την καλύτερη κατανόηση των δεδομένων:

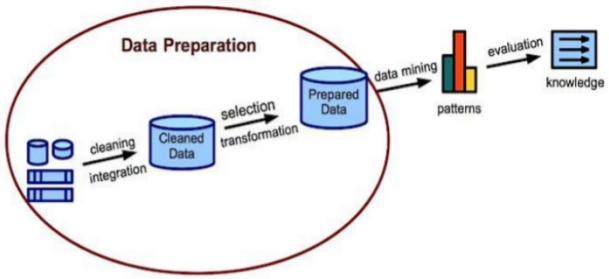


Πίνακας 3: Πίνακας χαρακτηριστικών

Επίσης σημαντικό ρόλο στην εξαγωγή γνώσης παίζει η κανονικοποίηση των δεδομένων. Η κανονικοποίηση αποτελεί μία διαδικασία μετατροπής των τιμών των διαφόρων χαρακτηριστικών που αποτελούν το σύνολο δεδομένων με σκοπό να μπορέσουμε να προσαρμόσουμε τις διάφορες τιμές [54] εντός ενός συγκεκριμένου εύρους τιμών. Για παράδειγμα μπορεί να έχουμε ένα σύνολο δεδομένων το οποίο να αποτελείται από χαρακτηριστικά στα οποία να έχουμε καταλήξει μετά από την ανάλυση και την εξαγωγή χαρακτηριστικών και να χρειάζεται να περαστούν αυτά τα δεδομένα από έναν αλγόριθμο ο οποίος [55], όμως δεν επιθυμεί μεγάλο εύρος τιμών για ένα συγκεκριμένο χαρακτηριστικό όπως είναι σε διάφορες οικονομικές εφαρμογές να παρατηρηθεί ένα μεγάλο εύρος τιμών. Έτσι, [56] αυτό μπορεί να κυμαίνεται σε ένα διάστημα το οποίο περιλαμβάνει χιλιάδες τιμές. Άρα γίνεται κατανοητό ότι για να μπορέσει ένας αλγόριθμος να δεχθεί δεδομένα και να εκπαιδευτεί καλύτερα πάνω σε συγκεκριμένα χαρακτηριστικά προσπαθούμε να περιορίσουμε το εύρος τιμών ενός συγκεκριμένου χαρακτηριστικού [57] ή και χαρακτηριστικών εντός ενός φυσιολογικού εύρους τιμών το οποίο πάντα καθορίζεται με βάση τις απαιτήσεις του εκάστοτε αλγόριθμου. Η πιο συνήθης περίπτωση είναι εκείνη κατά την οποία έχουμε να ορίζουμε τις διάφορες τιμές ενός χαρακτηριστικού για το πλήθος των εγγράφων εντός του εύρους 0 και 1.

Η Διαδικασία Εξόρυξης Γνώσης





Σχήμα 4: Εξόρυξη γνώσης

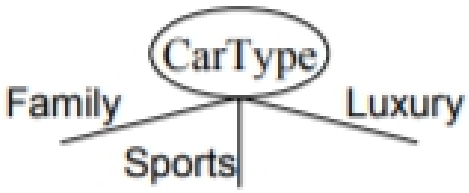
Προαναφέρθηκε ο ορισμός της έννοιας της κανονικοποίησης, παρακάτω θα δούμε τον όρο της προεπεξεργασίας των δεδομένων. Είναι ίσως από τα πιο σημαντικά βήματα με μεγάλη βαρύτητα στη διαδικασία εξόρυξης γνώσης και την αποτύπωση μοντέλου, καθώς έχουμε ότι, όπως αρχικά λαμβάνονται τα δεδομένα και διατυπώνεται ένα πρόβλημα επί ενός συνόλου δεδομένων και στη συνέχεια συλλέγονται και ακόμη περισσότερα δεδομένα για την καλύτερη εξαγωγή γνώσης στη συνέχεια περνάμε στην επεξεργασία των δεδομένων [58] η οποία περιλαμβάνει είτε τη διακριτοποίηση είτε την κανονικοποίηση. Όπως γίνεται αντιληπτό, η διακριτοποίηση είναι εκείνη η διαδικασία κατά την οποία ο ειδικός για την ανάλυση των δεδομένων θέτει στις αντίστοιχες τιμές των χαρακτηριστικών των διαφόρων εγγράφων που απαρτίζουν τα αντίστοιχα σύνολα, ονομαστικές τιμές [59] οι οποίες πιο συγκεκριμένα μπορούν να χρησιμοποιηθούν για την σύγκριση και την ομοιότητα μεταξύ τους. Ενώ από την άλλη [60], η κανονικοποίηση χρησιμοποιείται κυρίως σε συνεχείς μεταβλητές. Δηλαδή σε εκείνα τα χαρακτηριστικά τα οποία έχουν αποτελούμενες τιμές οι οποίες ορίζονται εντός ενός εύρους

τιμών και προσπαθούν να προσαρμοστούν σε μικρότερα εύρη τιμών για την καλύτερη όσο δυνατόν εφαρμογή των αλγορίθμων ταξινόμησης και εξόρυξης γνώσης.

2.2 Κατηγορικά γνωρίσματα

Ποιοτικές ή κατηγορικές λέγονται οι μεταβλητές των οποίων οι τιμές μπορούν να ταξινομηθούν σε κατηγορίες και δεν εκφράζουν απαραίτητα κάτι το μετρήσιμο (π.χ. ομά δα).

Στηρίζεται στην πολυμεταβλητή μεταχείριση των στοιχείων, λαμβάνοντας υπόψη πολλαπλές κατηγορικές μεταβλητές [61], δίχως να καταφεύγει δηλαδή στην κατά ζεύγη εξέταση των μεταβλητών. Ως αποτέλεσμα, είναι η δημιουργία γραφημάτων Χ-Υ των σημείων που προκύπτουν από τις στήλες και σειρές, με στόχο την ανεύρεση δομικών σχέσεων μεταξύ κατηγορικών μεταβλητών και παρατηρήσεων.



Σχήμα 5: Διαχωρισμός Car Type

Υπάρχουν διάφορες μέθοδοι ανάλυσης των κατηγορικών δεδομένων, σε μελέτη άρθρου παρατηρήσαμε και διαπιστώσαμε ότι η μέθοδος της Ανάλυσης Αντιστοιχιών (Correspondence Analysis-CA) χρησιμοποιείται με επιτυχία σε κατηγορικές μεταβλητές και βρίσκει πρόσφορο έδαφος στα πεδία εφαρμογών της κοινωνιολογίας, του marketing καθώς και της ψυχολογίας (Bourdieu, 1984, Greenacre, 2007). Την τελευταία εικοσιπενταετία όμως, η εφαρμογή της μεθόδου επεκτάθηκε και σε ποσοτικές μεταβλητές με το όνομα ανάλυση της αμοιβαίας

μεσοστάθμισης (Reciprocal Averaging-RA), με άμεση εφαρμογή κυρίως στις οικολογικές επιστήμες και σε θέματα εκτίμησης πληθυσμών σε σχέση με το περιβάλλον διαβίωσής τους (Hill, 1973).

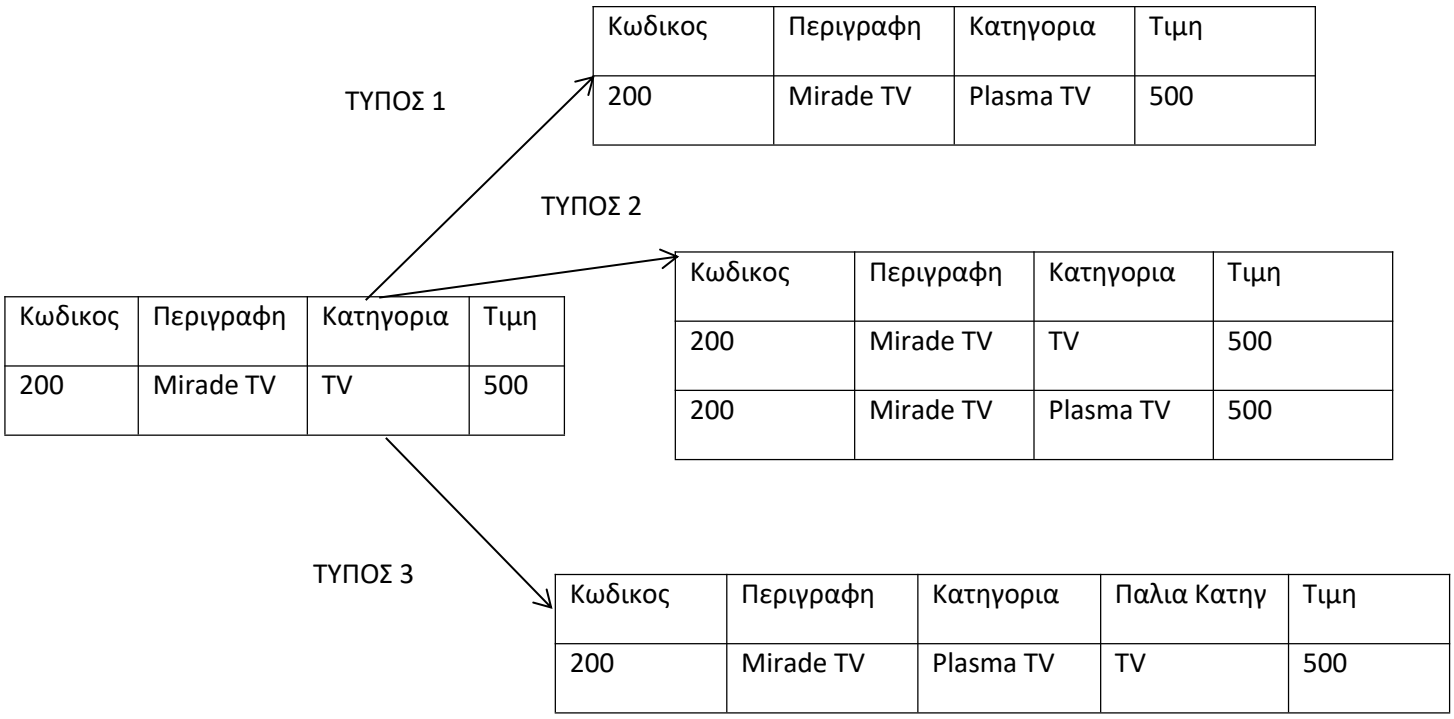
Τα δεδομένα ονομαστικής κλίμακας αναφέρονται και ως κατηγορικά δεδομένα (categorical data), κάνοντας λοιπόν τις αντίστοιχες μετρήσεις σε ένα ορισμένο αντικείμενο αυτό συνδέεται με μια ορισμένη κατηγορία ή τάξη (π.χ. άνδρας ή γυναίκα) [62]. Στις περιπτώσεις όπου τα δεδομένα ονομαστικής κλίμακας επιτρέπεται να λάβουν μόνο μία εκ' των δύο απο συγκεκριμένες τιμές, όπως άνδρες και γυναίκες, τότε ονομάζονται διχότομα ή δυαδικά (dichotomous, binary). Τα διχότομα δεδομένα –για παράδειγμα [63], η εμφάνιση ή όχι μιας πάθησης, θετικό ή αρνητικό αποτέλεσμα μιας εργαστηριακής δοκιμασίας κ.ά.– χρησιμοποιούνται συνεχώς στον υγειονομικό τομέα και γι' αυτό έχουν αναπτυχθεί πολλαπλές και ιδιαίτερες μέθοδοι ανάλυσης, όπως π.χ. η λογιστική παλινδρόμηση.

2.3 Πολυδιάστατα δεδομένα

Οι σύγχρονες επιχειρήσεις αλλά και μεγάλες εταιρείες πολυεθνικές παγκοσμίως, συλλέγουν δεδομένα τόσο εσωτερικά [64] εντός του οργανισμού τους όσο και από άλλες εξωτερικές πήγες. Επομένως γίνεται γνωστό ότι όλα αυτά τα δεδομένα στην ουσία ενώ διαθέτουν μεγάλη και πολύτιμη πληροφορία για τις εταιρίες και για την μετέπειτα συνέχεια στις μελλοντικές επενδύσεις του εκάστοτε οργανισμού θα πρέπει να υπάρξει η κατάλληλη εξαγωγή πληροφορίας και η σωστή λήψη αποφάσεων. Έτσι λοιπόν, θα πρέπει να είναι αρκετά λεπτομερής και ακριβής έτσι ώστε [65] από τις διάφορες πηγές η πληροφορία η οποία θα εξαχθεί αλλά και αυτή η οποία θα παραμείνει να είναι όσο το δυνατόν η καλύτερη [66], έτσι ώστε αρχικά να έχουμε λιγότερα δεδομένα ή πιο συγκεκριμένα λιγότερα χαρακτηριστικά από τη συλλογή των δεδομένων και κατά δεύτερον να μπορέσουμε να βελτιστοποιήσουμε τον αλγόριθμο με την εισαγωγή των επιθυμητών δεδομένων και την αντίστοιχη μορφή την οποία επιθυμεί ο αλγόριθμος [67]. Επομένως με αυτόν τον τρόπο θα εξαχθεί όσο το δυνατόν

καλύτερη πληροφορία η οποία θα διατυπωθεί ευανάγνωστα και καθαρά προς τους αναλυτές για τα μετέπειτα τους βήματα.

Ένα σημαντικό θέμα [68] που θα πρέπει να τονιστεί ως προς την επεξεργασία του μεγάλου όγκου δεδομένων είναι τα Συστήματα Επεξεργασίας Συναλλαγών (OLTP). Πρόκειται για πληροφοριακά συστήματα τα οποία καταγράφουν και επεξεργάζονται τις διάφορες συναλλαγές ενός οργανισμού ή μιάς επιχείρησης. Επίσης, οι αναλυτές και τα υψηλόβαθμα στελέχη των επιχειρήσεων χρησημοποιούν τα συστήματα OLAP [69] για διεξαγωγή αναλύσεων και έπειτα για τη λήψη αποφάσεων. Μέσω αυτών των συστημάτων εξασφαλίζουν την ταχεία, ευέλικτη πρόσβαση αλλά και την πολυδιάστατη επεξεργασία μεγάλων όγκων δεδομένων.



Σχήμα 6: Array παραδείγματα OLAP

Όπως αναφέρθηκε προηγουμένως ο ορός της πολυδιάστατης συλλογής των δεδομένων και επεξεργασίας αυτών είναι εκείνη η δυνατότητα με την οποία θα μπορέσει κάποιος να επιλέξει αυτή την πληροφορία και να τη χωρίσει στα κατάλληλα κομμάτια και τις αντίστοιχες

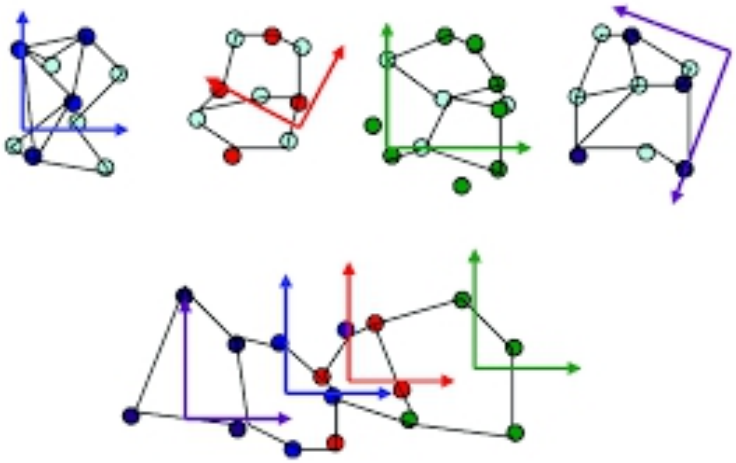
διαστάσεις [70]. Έτσι λοιπόν ένας αναλυτής ο οποίος θα πάει να δει μία σύνθετη πληροφορία διατυπωμένη σε ένα ερώτημα τότε θα χρειαστεί να ψάξει τμηματικά [71] σε κάθε συν αποτελούμενο υποσύνολο δεδομένων προκειμένου να μπορέσει να απαντήσει στο πολύπλοκο αλλά και σύνθετο ερώτημα. Το ερώτημα προς απάντηση θα χρειαστεί για να απαντηθεί την εξέταση αρκετών διαστάσεων του χώρου που πραγματοποιούν σύνολα δεδομένων. Για παράδειγμα [72], ένας χρήστης θα είχε την δυνατότητα να υποβάλλει ένα ερώτημα σε ένα σύστημα OLAP, ώστε να πληροφορηθεί για το συνολικό ύψος πωλήσεων είτε ανά κατηγορία προϊόντος είτε ανά γεωγραφική περιοχή .

Έτσι λοιπόν γίνεται αντιληπτό ότι με τον όρο πολυδιάστατα δεδομένα απαιτείται ένας μηχανισμός ή αλλιώς καλύτερα αισθητήρας των δεδομένων που συλλέγονται και στη συνέχεια να πραγματοποιείται μία προεπεξεργασία των δεδομένων με διάφορες μεθόδους όπως είναι ο καθαρισμός δεδομένων και έπειτα η ανίχνευση κατηγορικών ή αριθμητικών τιμών. Έπειτα πρέπει να εφαρμόζεται ο ορισμός της εξαγωγής χαρακτηριστικών κάτι που σημαίνει ότι στα πολυδιάστατα δεδομένα απαιτείται η συγκεκριμένη διαδικασία ως υποχρεωτική και όχι προαιρετική όπως ήταν σε περίπτωση ενός μικρού συνόλου δεδομένων, έτσι λοιπόν με την παραπάνω αναφορά θέλουμε να δώσουμε το νόημα ότι στα πολυδιάστατα δεδομένα θα συναντήσουμε αρκετά χαρακτηριστικά. Άρα θα έχουμε ένα σύνολο δεδομένων που θα αποτελείται από πολλές στήλες [73]. Επομένως, αυτό σημαίνει ότι θα πρέπει να κάνουμε εξαγωγή κάποιων χαρακτηριστικών να κρατήσουμε δηλαδή τα κυρίως σημαντικά χαρακτηριστικά και τα οποία μας ενδιαφέρουν να αναλύσουμε και έπειτα αυτά να τα εισάγουμε σε έναν αλγόριθμο ταξινόμησης προκειμένου να εξαχθούν τα κατάλληλα αποτελέσματα και να πραγματοποιηθεί η εκπαίδευση του συγκεκριμένου μοντέλου πάνω στα δεδομένα που συλλέχθηκαν, έπειτα να οδηγηθούμε στην συνεχή αξιολόγηση των δεδομένων ώστε να είναι ασφαλή τα συμπεράσματα στα οποία καταλήγουν τα μοντέλα και οι αλγόριθμοι που εφαρμόζονται πάνω στα σύνολα δεδομένων.

2.4 Μείωση διαστατικότητας

Στα προβλήματα κατηγοριοποίησης μηχανικής μάθησης, υπάρχουν συχνά πάρα πολλοί παράγοντες βάσει των οποίων γίνεται η τελική κατηγοριοποίηση. Αυτοί οι παράγοντες είναι βασικές μεταβλητές που ονομάζονται χαρακτηριστικά.

Τα προβλήματα στην ανάλυση δεδομένων είναι εκείνο το πλήθος δεδομένων το οποίο αποτελείται από έναν μεγάλο αριθμό χαρακτηριστικών που γίνονται ολοένα και περισσότερα σε περιοχές όπως είναι ο χώρος της βιοπληροφορικής [74] ή ο χώρος των χρηματοοικονομικών εφαρμογών. Επομένως σε μία τέτοια συγκεκριμένη κατάσταση κρίνεται απαραίτητη η μείωση της διάστασης των δεδομένων ώστε να μπορέσει να βελτιωθεί η αποδοτικότητα του αλγορίθμου [75] προς εφαρμογή για την εξαγωγή χρήσιμων πληροφοριών αλλά [76] και για την ακρίβεια των δεδομένων τα οποία θα χρησιμοποιηθούν για την εκπαίδευσή του αλγόριθμου και στη συνέχεια την εφαρμογή του σε διάφορα άγνωστα σύνολα δεδομένων.



Σχήμα 7: Νευρώνες εκπαίδευσης

Με αφορμή το παραπάνω έγινε ανάπτυξη διάφορων μεθόδων μείωσης διαστατικότητας [77] (Dimensionality Reduction Techniques). Η μείωση διαστάσεων (Dimensionality Reduction –DR)

ώς μέθοδος προσπαθεί να προβάλει ένα σύνολο από διανυσμάτων υψηλής διάστασης [78] σε ένα χώρο χαμηλότερης διάστασης. Στόχος αυτής της μεθόδου είναι να μείωσεί τις διαστάσεις των δεδομένων ενώ παράλληλα θα διατήρησουν αναλλοίωτες και τις ιδιοτήτες τους.

Πιο αναλυτικά οι Τεχνικές Μείωσης Διάστασης προβάλουν δεδομένα από τον αρχικό υψηλής διάστασης χώρο Rn σε έναν νέο χαμηλότερης διάστασης χώρο Rk (συνήθως k << n). Η μέθοδος αυτή είναι εξαιρετικά σημαντική και απαραίτητη για τους παρακάτω λόγους. Αρχικά, γιατί οι αποστάσεις μεταξύ των δεδομένων στον νέο ελαττωμένο χώρο [79] μπορούν να υπολογίστούν πιο γρήγορα συγκριτικά με τον αρχικό χώρο υψηλής διάστασης. Ταυτοχρόνως, γίνεται μείωση του μεγέθους του συνόλου δεδομένων και αποκαλύπτεται η δομή των δεδομένων [80], η οποία δεν είναι ορατή στον αρχικό πολυδιάστατο χώρο.

Η εφαρμογή της συγκεκριμένης τεχνικής που αφορά την μείωση της διάστασης του χώρου των χαρακτηριστικών που συναποτελούν το σύνολο δεδομένων, αποτελεί την κατάρα της μείωσης των διαστάσεων λόγω του γεγονότος ότι μπορεί να χαθεί κάποια πολύτιμη πληροφορία κατά τη διάρκεια μετάβασης [81] από έναν χώρο μεγάλων διαστάσεων σε έναν χώρο μικρότερων διαστάσεων. Προκειμένου να μπορέσει να αποτυπωθεί η ίδια πληροφορία [82], θα πρέπει να υπάρχει κατάλληλη και ταυτόχρονα ακριβής εφαρμογή του αλγορίθμου προκειμένου να μη χαθεί κάποια πληροφορία κατά τη διάρκεια της μείωσης της διάστασης του αρχικού συνόλου δεδομένων με την πληροφορία που διέθετε.

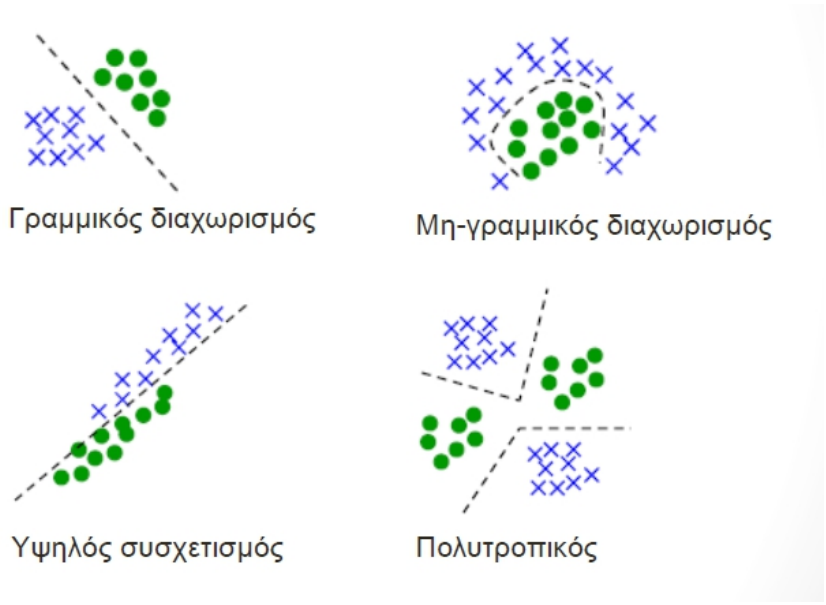
3 Κατηγοριοποίηση

Με τον όρο κατηγοριοποίηση (classification) αναφερόμαστε στην πρόβλεψη της ετικέτας μιας κατηγορίας (class label) [83] για ένα καθορισμένο μη σημασμένο σημείο η οποία είναι άγνωστη πριν διακρίνει το ένα αντικείμενο ένα άλλο με βάση τα χαρακτηριστικά. Τα δεδομένα [140] θα πρέπει να διαιρούνται με την κατάρτιση των δεδομένων, δηλαδή δεδομένων που θα συνδέονται και θα γνωρίζουν την ετικέτα κατηγορίας και τα δεδομένα δοκιμών που θα είναι τα δείγματα της δοκιμής για να ανακαλύψουν τις ετικέτες κατηγορίας. Σε αυτό το κεφάλαιο θα παρουσιαστούν η πιθανοτητική κατηγοριοποίηση που περιλαμβάνει τον κατηγοριοποιητή Bayes και επίσης την απλοποιημένη εκδοχή του κατηγοριοποιητή Bayes που δεν είναι άλλος κανείς εκτός από τον κατηγοριοποιητή Κ πλησιέστερων γειτόνων.

Στο δεύτερο σκέλος αυτού του κεφαλαίου θα δούμε τον κατηγοριοποιητή με δέντρα αποφάσεων decision tress [84] και τον σχετικό αλγόριθμο. Κλείνοντας το κεφάλαιο στην τρίτη ενότητα αυτού του κεφαλαίου και τελευταία θα αξιολογηθεί η κατηγοριοποίηση και ο τρόπος της διαδικασίας με την οποία πραγματοποιείται.

Ωστόσο, θα επικεντρωθούμε ακόμα λίγο στην έννοια της κατηγοριοποίησης, που είναι στην ουσία η ταξινόμηση των δειγμάτων στην σωστή κλάση [85] ή ομάδα αντικειμένων και το πρόβλημα κατάταξης ενός αντικειμένου σε μια κατηγορία (class).

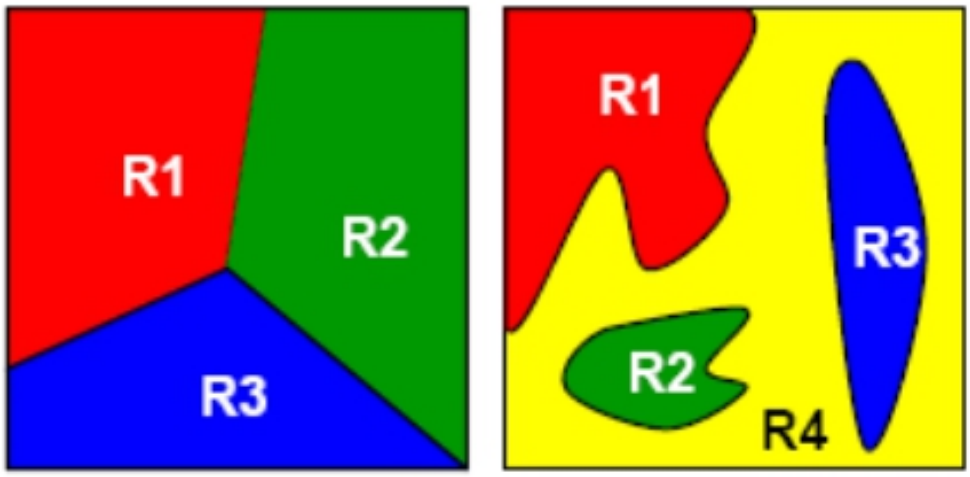
Ωστόσο, λόγω της εισαγωγής στις επόμενες ενότητες, θα χρειαστεί να προσδιορίσουμε κάποιες σημαντικές έννοιες ώστε [85] να γίνουν κατανοητές σε βάθος και όχι σε επιφανειακό επίπεδο[86], καθώς κρίνονται απαραίτητες για την συνέχεια της μελέτης και του προσδιορισμού των αλγορίθμων και ανάλυσης των διαφόρων τεχνικών κατηγοριοποίησης.



Σχήμα 8: Μορφές Διαχωρισμού

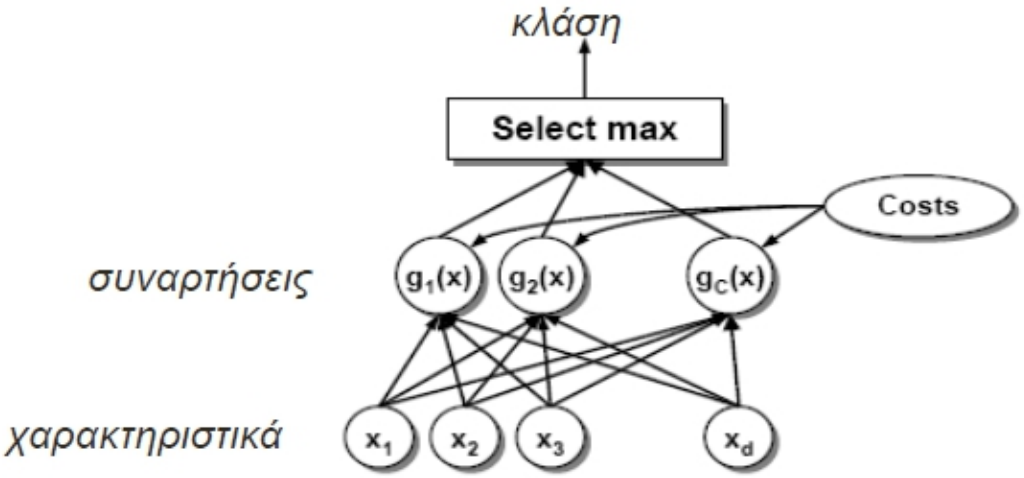
Αρχικά ως χαρακτηριστικά μπορεί να είναι συμβολικά [86] (για παράδειγμα χρώμα, μάρκα αυτοκινήτου) ή αριθμητικά (για παράδειγμα ύψος, βάρος). Ο συνδυασμός ορισμένων χαρακτηριστικών είναι το διάνυσμα χαρακτηριστικών (feature vector). Τό feature vector ορίζει τον ν-διάστατο χώρο και ονομάζεται χώρος χαρακτηριστικών [87] (feature space). Επίσης μια άλλη σημαντική έννοια που συνδέεται στενά με τα χαρακτηριστικά είναι η έννοια του προτύπου, που αποτελεί στην ουσία μια σύνθεση των χαρακτηριστικών και είναι κατά την διαδικασία της ταξινόμησης ένα πρότυπο ζεύγος μεταβλητών {χ, ω}, όπου (χ) είναι μια συλλογή χαρακτηριστικών (feature vector) και ως (ω) είναι η έννοια της παρατήρησης ή αλλιώς του ειδικού χαρακτηριστικού κλάσης (class label).

Ο ταξινομητής διαχωρίζει τον χώρο των χαρακτηριστικών σε συγκεκριμένες περιοχές απόφασης [88] (classes). Οι classes χωρίζονται με όρια απόφασης.



Σχήμα 9: Διαχωρισμός περιοχών

Έτσι λοιπόν γίνεται αντιληπτό ότι ένας classifier αντιπροσωπεύεται με ένα σύνολο διακριτών συναρτήσεων. Ένα διάνυσμα χαρακτηριστικών x καταχωρείται σε μια κλάση ωi αν gi(x)>gj(x).



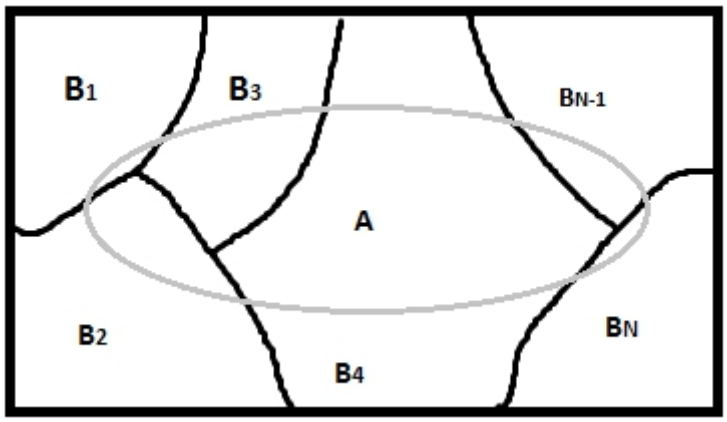
Σχήμα 10: Συναρτήσεις εισαγωγής σε χαρακτηριστικά

Επίσης βασικές μαθηματικές ή καλύτερα στατιστικές προσεγγίσεις που θα χρησιμοποιηθούν κυρίως σε αυτό το κεφάλαιο είναι η πιθανότητα υπό συνθήκη και το θεώρημα ολικής πιθανότητας. Παρακάτω αναλύονται ως εξής:

1) Αν Α και Β είναι δύο γεγονότα, η πιθανότητα του Α όταν ξέρουμε ότι το Β έχει ήδη συμβεί είναι:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

2) Έστω B1,B2,…,BN γεγονότα, χωρίς κοινά στοιχεία που η ένωση τους συμπίπτει με το χώρο δειγματοληψίας S, ονομάζονται διαμερισμός του S.



Σχήμα 11: Σύνολο Α

Ένα γεγονός Α μπορεί να εκφραστεί ως:

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \ldots \cup B_N) = (A \cap B_1) \cup (A \cap B_2) \cup \ldots (A \cap B_N)$$

Γι’ αυτό μπορεί να εκφραστεί ως εξής:

$$P[A]= P[A|B_1]P[B_1]+... P[A|B_N]P[B_N]= \sum_{k=1}^{N} P[A|B_k]P[B_K]$$

Η τιμή μιας εποπτευόμενης ταξινόμησης είναι συνήθως συνάρτηση της ακρίβειας. Ένας από τους βασικούς στόχους της ταξινόμησης είναι συχνά η επίτευξη υψηλής ακρίβειας [89], εάν είναι δυνατόν, έναν μικρό αριθμό εκπαιδευτικών δειγμάτων για να καταστεί η διαδικασία ταξινόμησης όσο το δυνατόν πιο χρήσιμη και οικονομική [90]. Ένας ελκυστικός ταξινομητής για αυτήν την εφαρμογή είναι μια μηχανή φορέα υποστήριξης (SVM) ,προτάθηκε αρχικά από τον Vladimir Vapnik για δυαδική ταξινόμηση το 1992. Βασίζεται σε μια μέθοδο μηχανικής εκμάθησης για τη θεωρία της στατιστικής μάθησης, ακολουθώντας τη θεωρία διάστασης VC και την αρχή ελαχιστοποίησης του δομικού κινδύνου.

Το SVM έχει καλή απόδοση γενίκευσης [156] και πλεονέκτημα όσον αφορά την επίλυση μικρού δείγματος, μη γραμμικής και υψηλής διάστασης αναγνώρισης προτύπων. Το SVM έχει εφαρμοστεί με επιτυχία σε ευρεία πεδία, όπως αναγνώριση προσώπου , ταξινόμηση κειμένου , αναγνώριση εικόνας, ανάκτηση πληροφοριών, ανίχνευση εισβολής, αναγνώριση φωνής.Οι ταξινομήσεις SVM είναι πιο ακριβείς από τις ευρέως χρησιμοποιούμενες εναλλακτικές λύσεις, όπως [91] η ταξινόμηση κατά μέγιστη πιθανότητα, το δέντρο αποφάσεων και οι προσεγγίσεις που βασίζονται σε νευρωνικά δίκτυα.

Ένα SVM στοχεύει στην προσαρμογή ενός βέλτιστου διαχωριστικού υπερπλάνου (OSH) [92] μεταξύ τάξεων εστιάζοντας στα δείγματα εκπαίδευσης που βρίσκονται στην άκρη των διανομών τάξης και των διανυσμάτων υποστήριξης. Το OSH είναι προσανατολισμένο έτσι ώστε [93] να τοποθετείται στη μέγιστη απόσταση μεταξύ των συνόλων διανυσμάτων υποστήριξης. Λόγω αυτού του προσανατολισμού, το SVM αναμένεται να γενικευτεί με μεγαλύτερη ακρίβεια σε αόρατες περιπτώσεις σε σχέση με ταξινομητές που στοχεύουν στην ελαχιστοποίηση του σφάλματος εκπαίδευσης [94] όπως τα νευρικά δίκτυα. Έτσι, με την ταξινόμηση SVM μόνο μερικά από τα δείγματα εκπαίδευσης που βρίσκονται στην άκρη των διανομών τάξης στο χώρο χαρακτηριστικών [95] (διανύσματα υποστήριξης) απαιτούνται για τη δημιουργία της επιφάνειας απόφασης σε αντίθεση με τους στατιστικούς ταξινομητές, όπως οι

ευρέως χρησιμοποιούμενοι ταξινομητές μέγιστης πιθανότητας στους οποίους όλες [96] οι προπονήσεις χρησιμοποιούνται για τον χαρακτηρισμό των τάξεων. Η συμβατική ταξινόμηση της μέγιστης πιθανότητας μπορεί, επομένως, να απαιτεί πολύ μεγαλύτερο μέγεθος δείγματος εκπαίδευσης από το SVM για την απόκτηση ακριβούς ταξινόμησης. Επιπλέον [97], μερικές φορές είναι δυνατόν να προσδιοριστούν οι πιο χρήσιμοι χώροι εκπαίδευσης για την παροχή διανυσμάτων υποστήριξης πριν από την ταξινόμηση [97]. Αυτή η δυνατότητα ακριβούς ταξινόμησης με βάση μικρά εκπαιδευτικά σύνολα σημαίνει ότι η υιοθέτηση της ταξινόμησης SVM [97] μπορεί να προσφέρει στον αναλυτή σημαντική εξοικονόμηση στην απόκτηση δεδομένων εκπαίδευσης.

Η δυαδική προσέγγιση SVM μπορεί, ωστόσο [98], να επεκταθεί για σενάρια πολλαπλών κλάσεων που συναντώνται συνήθως στην τηλεπισκόπηση. Αυτό επιτυγχάνεται γενικά με την αποσύνθεση του προβλήματος multiclass [98] σε μια σειρά δυαδικών αναλύσεων που μπορούν να αντιμετωπιστούν με ένα δυαδικό SVM ακολουθώντας είτε τις στρατηγικές one-against-one είτε one-against-all [99]. Εναλλακτικά, μπορεί επίσης [100] να πραγματοποιηθεί μια ταξινόμηση πολλαπλών κλάσεων με βάση μία μόνο βελτιστοποίηση. Ένα σημαντικό πλεονέκτημα του SVM multiclass [101] είναι ότι η ταξινόμηση σε όλες τις κατηγορίες γίνεται σε ένα μόνο βήμα. Αυτή η προσέγγιση είναι πολύ διαφορετική από αυτήν που υιοθετήθηκε στην ταξινόμηση [102] πολλαπλών κλάσεων με βάση το δυαδικό SVM. Με την τελευταία, μια σειρά αναλύσεων, ο αριθμός των οποίων είναι μια θετική συνάρτηση του αριθμού των τάξεων, πρέπει να πραγματοποιηθεί για να προκύψει η ταξινόμηση. Επιπλέον, μειώνοντας την ταξινόμηση σε ένα πρόβλημα βελτιστοποίησης [103], η προσέγγιση SVM multiclass μπορεί επίσης να απαιτεί λιγότερα διανύσματα υποστήριξης από μια ταξινόμηση πολλαπλών κλάσεων που βασίζεται σε δυαδικά SVM [104], αν και αυτό το δυναμικό σπάνια έχει διερευνηθεί.

Μία από τις κύριες προκλήσεις στο παραδοσιακό SVM είναι η επίλυση του προβλήματος προγραμματισμού (QPP) της υψηλής υπολογιστικής πολυπλοκότητας. Όταν το μέγεθος του σετ προπόνησης είναι L, η υπολογιστική πολυπλοκότητα του SVM είναι O (L^3). Προκειμένου να αποφευχθεί ο χρόνος προπόνησης είναι πολύ μεγάλος, προωθήστε το Least Squares Support

Vector Machine (LSSVM) [6] , v-Support Vector Machine (v-SVM) [7] , Proximal Support Vector Machine (PSVM) [8] , Twin Support Vector Machine (TSVM) [9] , ελαχίστου τετραγώνου δίδυμο Support Vector Machine (LSTSVM) [10] και ούτω καθεξής.

Το LSTSVM-Partial BT εκπαιδεύεται ταχύτερα από το OVA, OVO. Κατά τη διάρκεια της προπόνησης, ο αριθμός LSTSVM μειώνεται βήμα προς βήμα, χρειάζεται μόνο να κάνετε ερώτημα (K-1) / 2 LSTSVM για τη νέα παρουσία και δεν υπάρχει αδιαχώριστη περιοχή κατά τη διέλευση του δυαδικού δέντρου. Είναι ταχύτερο από το OVA, το OVO. Λόγω της ιεραρχικής δομής του δυαδικού δέντρου, θα εμφανιστεί η συσσώρευση σφάλματος και η ανισορροπία παρουσίας στη διαδικασία της προπόνησης.

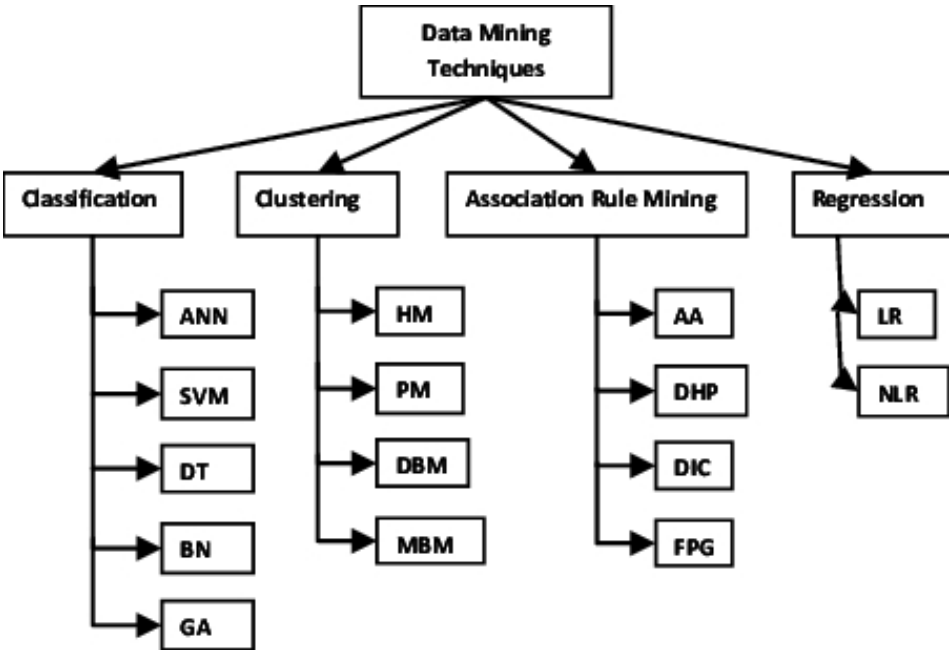
Με βάση την ταξινόμηση πολλαπλών κατηγοριών παρουσίας LSTSVM και μερικής εφαρμογής BT, επιλύθηκε αποτελεσματικότερα η συσσώρευση σφάλματος μερικής κατάταξης BT και η ανισορροπία παρουσίας. Η βασική αιτία της συσσώρευσης σφαλμάτων είναι η χαμηλή ακρίβεια ταξινόμησης κόμβων. Σε σύγκριση με το BSVM, βελτιώθηκε η ακρίβεια ταξινόμησης LSTSVM και μειώθηκε η πιθανότητα συσσώρευσης σφαλμάτων. Για πρόβλημα ανισορροπίας, το LSTSVM που χρησιμοποιεί διαφορετικές παραμέτρους ποινής c1, c2 αντιπροσωπεύει τη σημασία διαφορετικών συνόλων δεδομένων [157], δίνει σοβαρή τιμωρία μικρών συνόλων δεδομένων, για να αντιπροσωπεύσει τη σημασία των δειγμάτων.

3.1 Πιθανοτική κατηγοριοποίηση

Η Κατηγοριοποίηση (Classification) και η Παλινδρόμηση (Regression) έχουν πολλές ομοιότητες και ανήκουν στην επιβλεπόμενη μάθηση. Και στις δύο περιπτώσεις υπάρχει ένας κοινός στόχος ο οποίος είναι να προβλέψουν τις τιμές ενός γνωρίσματος μέσω της χρήσης άλλων γνωρισμάτων. Επίσης [105], και στις δύο περιπτώσεις η κατασκευή του μοντέλου γίνετε με τη χρήση και την επεξεργασία ενός συνόλου δεδομένων εκπαίδευσης. Ανάμεσα στην

κατηγοριοποίηση και στην παλινδρόμηση [106] υπάρχει μιά διαφορά όσων αφορά τον τύπο της εξαρτημένης μεταβλητής. Η παλινδρόμηση στοχεύει στην πρόβλεψη μιας εξαρτημένης μεταβλητής [107], η οποία περιέχει συνεχόμενες (αριθμητικές) τιμές.

Αντιθέτως, η κατηγοριοποίηση στοχεύει στη πρόβλεψη διακριτών ονομαστικών τιμών οι οποίες είναι συγκεκριμένες και γνωστές εξ' αρχής και ορίζουν την κλάση (κατηγορία) στην οποία ανήκει κάθε αντικείμενο. Έτσι σε προβλήματα κατηγοριοποίησης η εξαρτημένη μεταβλητή αναφέρεται και ώς γνώρισμα κλάσης.

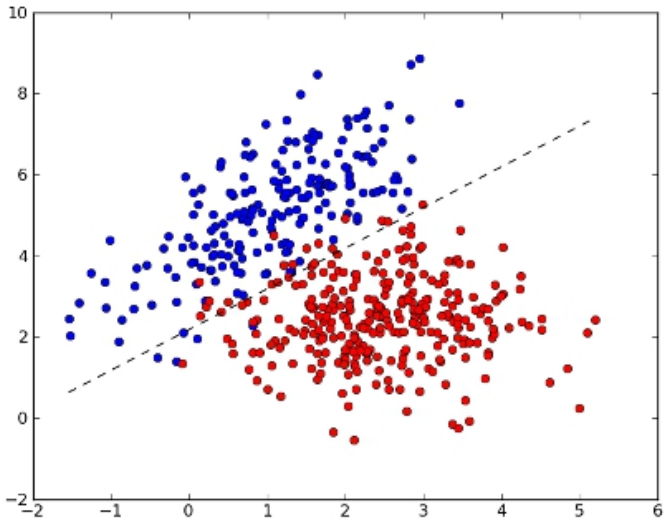


Σχήμα 12: Data Mining

Υπάρχει μεγάλη ποικιλία μεθόδων κατηγοριοποίησης και επαγωγικοί αλγόριθμοι, ορισμένες είναι και οι παρακάτω τα Δένδρα Αποφάσεων [108], τα Μπαϋεσιανά Δίκτυα, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron κλπ. Τα μοντέλα που δημιουργεί η κάθε μέθοδος διαφέρουν αναμεταξύ τους [108].

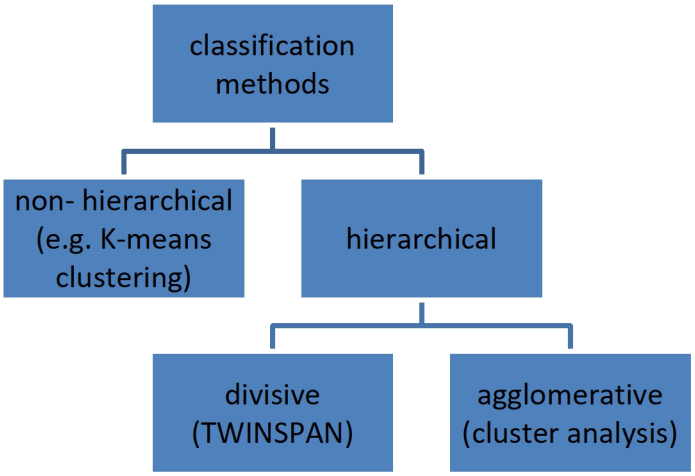
Ένα αντιπροσωπευτικό παράδειγμα τα μοντέλα Δένδρου Αποφάσεων, έχουν δενδρική δομή και διάφορους κόμβους όπου κάθε κόμβος πρακτικά είναι ένα σημείο ελέγχου σε κάποιο

γνώρισμα [108], κάθε κλάδος είναι ένα αποτέλεσμα του ελέγχου και κάθε φύλο είναι μια απόφαση κατηγοριοποίησης. Τα μοντέλα Μπαϋεσιανού Δικτύου είναι γράφοι κατευθυνόμενοι και ακυκλικοί ενώ υπάρχει κατανομή πιθανοτήτων συσχέτισης μεταξύ των μεταβλητών και κάθε κόμβος αντιστοιχεί σε μια μεταβλητή.



Σχήμα 13: Κατηγοριοποίηση

Τεχνικές ταξινόμησης για ανάλυση ιατρικής εικόνας και διάγνωση μέσω υπολογιστή καλύπτει τις πιο πρόσφατες εξελίξεις σχετικά με τον τρόπο εφαρμογής τεχνικών ταξινόμησης σε μια μεγάλη ποικιλία κλινικών εφαρμογών που είναι κατάλληλες για ερευνητές και βιοϊατρικούς μηχανικούς [109] στους τομείς της μηχανικής μάθησης, της βαθιάς μάθησης, της ανάλυσης δεδομένων, των δεδομένων σχεδιασμού συστημάτων διαχείρισης [110] και διάγνωσης μέσω υπολογιστή (CAD).



Σχήμα 14: Διαχωρισμός αλγορίθμων κατηγοριοποίησης

Επομένως τα βήματα που ακολουθούνται στην κατηγοριοποίηση, είναι αρχικά η συλλογή των δεδομένων και απάντηση αρχικά στην ερώτηση πόσα δείγματα χρειαζόμαστε για την ανάλυση μας, στην συνέχεια θα επιλεχθούν [111] τα επιθυμητά χαρακτηριστικά που επιθυμεί ο αναλυτής να χρησιμοποιήσει. Έπειτα, πραγματοποιείται η επιλογή του κατάλληλου μοντέλου [111] που μπορεί να είναι στατιστικό, νευρωνικό ή συντακτικό. Επιπλέον, γίνεται η κατάλληλη εκπαίδευση που μπορεί να πραγματοποιηθεί για την επιβλεπόμενη ή μη-επιβλεπόμενη μάθηση [112] και στο τέλος γίνεται η αξιολόγηση και η εκτίμηση της απόδοσης του εφαρμοζόμενου αλγορίθμου που υλοποιήθηκε.

Η μηχανή φορέα υποστήριξης (SVM) [113] είναι μια από τις πιο ισχυρές τεχνικές για εποπτευόμενη ταξινόμηση. Επιτυγχάνει την αντιστάθμιση μεταξύ ελαχιστοποίησης του σφάλματος εκπαίδευσης και μεγιστοποίησης του περιθωρίου διαχωρισμού με βάση τη θεωρία Vapnik-Chervonenkis [114] και την αρχή ελαχιστοποίησης του διαρθρωτικού κινδύνου. Αν και το SVM έχει δείξει καλή απόδοση γενίκευσης σε πολλές πραγματικές εφαρμογές, πάσχει από ορισμένα μειονεκτήματα, συμπεριλαμβανομένων μη προβολικών προβλέψεων και αυξημένης υπολογιστικής πολυπλοκότητας που προκαλείται από τον αυξημένο αριθμό φορέων

υποστήριξης όταν υπάρχει αντιμετώπιση προβλημάτων μεγάλης κλίμακας [115]. Σε πολλά σενάρια, είναι πιθανές οι πιθανότητες εξόδου επειδή είναι χρήσιμες για την εκτίμηση της εμπιστοσύνης ή της αβεβαιότητας στην πρόβλεψη. Για παράδειγμα, οι πληροφορίες αβεβαιότητας μπορούν να επηρεάσουν τη βέλτιστη θεραπεία των ασθενών [116] κατά τη λήψη ιατρικών αποφάσεων. Ο Platt [116] πρότεινε να δημιουργηθούν οπίσθιες εκτιμήσεις πιθανότητας μέσω της *εκ* των *υστέρων* προσαρμογής μιας σιγμοειδούς συνάρτησης στις εξόδους SVM. Ωστόσο [117], η έξοδος του σιγμοειδούς δεν είναι απαραίτητα καλή προσέγγιση της οπίσθιας πιθανότητας . Wu et αϊ. παρουσίασαν δύο μεθόδους για την απόκτηση εκτιμήσεων πιθανότητας πολλαπλών κλάσεων με ζεύγη ζεύξης. Στο λογισμικό LIBSVM [118], οι πιθανότητες ζευγαριού κατηγορίας υπολογίζονται πρώτα χρησιμοποιώντας μια βελτιωμένη εφαρμογή της μεθόδου του Platt . Στη συνέχεια, η δεύτερη προσέγγιση που προτείνεται χρησιμοποιείται για την απόκτηση πιθανολογικών εξόδων πολυμέσων SVM.

Το μηχάνημα συνάφειας φορέα (RVM) [119] είναι ένα μοντέλο εκμάθησης Bayesian , το οποίο δεν πάσχει από τους περιορισμούς του SVM. Το RVM εισάγει ένα μηδενικό μέσο όρο Gauss πριν από κάθε βάρος μοντέλου και μεγιστοποιεί την οριακή πιθανότητα χρησιμοποιώντας [120] μια διαδικασία μέγιστης πιθανότητας τύπου II. Ως αποτέλεσμα της αδυναμίας που προκαλεί προηγουμένως, η πλειονότητα των βαρών του μοντέλου κατανέμεται απότομα γύρω στο μηδέν. Ως εκ τούτου, αυτά τα βάρη κλαδεύονται και λαμβάνεται ένα αραιό μοντέλο. Για την αντιμετώπιση multiclass προβλημάτων, δύο εκδόσεις του multiclass RVM (mRVM $_1$ και mRVM $_2$) προτάθηκαν στο [120], με την εισαγωγή βοηθητικές μεταβλητές, οι οποίες οδηγούν φυσικά στην multinomial πιθανότητα probit για την εκτίμηση των πιθανοτήτων ένταξης κατηγορίας.

Ωστόσο, οι Chen et al. επεσήμανε ότι το RVM είναι ευαίσθητο στην παράμετρο του πυρήνα λόγω της ακατάλληλης διαμόρφωσης που υιοθετεί Gaussian μηδενικού μέσου όρου σε σχέση με τα βάρη τόσο για θετικές όσο και για αρνητικές τάξεις. Επομένως, ορισμένα δεδομένα εκπαίδευσης που ανήκουν σε θετική τάξη μπορεί να έχουν αρνητικά βάρη και το αντίστροφο, το οποίο οδηγεί σε υποβέλτιστες λύσεις. Για την αντιμετώπιση αυτού του προβλήματος [121], η πιθανοτική μηχανή ταξινόμησης φορέα (PCVM) υιοθετεί μια υπογεγραμμένη και

περικομμένη Gaussian πριν από κάθε βάρος, όπου το σύμβολο του προηγούμενου καθορίζεται από την ετικέτα κλάσης, δηλαδή, +1 ή -1. Η περικομμένη Gaussian προηγούμενη όχι μόνο περιορίζει το σημάδι των βαρών αλλά επίσης οδηγεί σε μια αραιή εκτίμηση των διανυσμάτων βάρους, και έτσι μειώνει την πολυπλοκότητα του μοντέλου. Επιπλέον, ένας αλγόριθμος προσδοκίας-μεγιστοποίησης (EM) [121] κλειστού τύπου χρησιμοποιείται για τη βελτιστοποίηση των παραμέτρων κατά τη διάρκεια της προπόνησης, γεγονός που βελτιώνει την ευρωστία των PCVM [121]. Ένα αποτελεσματικό PCVM προτάθηκε με τη διαδοχική προσθήκη ή διαγραφή βασικών συναρτήσεων σύμφωνα με τη μεγιστοποίηση της οριακής πιθανότητας, η οποία βελτίωσε την πολυπλοκότητα του χρόνου εκτέλεσης στο τετραγωνικό κόστος. Οι Schleif et al. [122] πρότεινε την προσέγγιση της μήτρας εισόδου χρησιμοποιώντας την προσέγγιση Nyström και απέκτησε γραμμικό χρόνο εκτέλεσης και πολυπλοκότητα μνήμης για το αυξητικό PCVM. Τα πειραματικά αποτελέσματα, απέδειξε ότι τα PCVM συνήθως ξεπέρασαν τα SVM και τα αρχικά RVM όσον αφορά την ακρίβεια της πρόβλεψης και την περιοχή κάτω από την καμπύλη του χαρακτηριστικού λειτουργίας του δέκτη (AUC). Ωστόσο, το PCVM είχε αρχικά σχεδιαστεί για δυαδική ταξινόμηση, επομένως δεν μπορεί να εφαρμοστεί άμεσα σε προβλήματα πολλών κλάδων.

3.1.1 Κατηγοριοποιητής Bayes

Ουσιώδη χαρακτηριστικά: είναι γνωρίσματα που λαμβάνονται από τα πρότυπα, και η διαδικασία της ταξινόμησης βασίζεται στις τιμές τους.

Τα ουσιώδη χαρακτηριστικά θα πρέπει να επιλέγονται έτσι ώστε οι τιμές που λαμβάνουν για τα διάφορα πρότυπα να έχουν μεγάλη **δια–κλασική** *(between class)* απόσταση και μικρή **ενδο–κλασική** (*within–class*) απόσταση[66]:

- ένας αριθμός χαρακτηριστικών γνωρισμάτων

$X_1,\ldots,X_l,$

αποτελούν το χαρακτηριστικό διανύσματα

$\underline{X} = [X_1,\ldots,X_l]^T \in R^l$

τα οποία αντιμετωπίζονται ως τυχαία διανύσματα

Κατά την διαδικασία της παρατήρησης ενός παραδείγματος εκμάθησης αυξάνεται ή μειώνετε κατά μεγάλο βαθμό [123] η εκτίμηση της πιθανότητας αν μια υπόθεση – πρόβλεψη είναι σωστή. Η προγενέστερη γνώση αλλά και η εμπειρία μπορεί να συνδυαστεί με τα στοιχεία που παρατηρήθηκαν για να καθορίσει την τελική πιθανότητα μιας υπόθεσης. Οι Μπεϋζιανοί μέθοδοι [124] έχουν τη δυνατότητα να προσαρμόσουν τις υποθέσεις που κάνουν στις πιθανολογικές προβλέψεις. Οι νέες περιπτώσεις δύναται να ταξινομηθούν με βάση τις προβλέψεις σε συνδυασμό με πολλαπλές υποθέσεις [125], που καθορίζονται από τις πιθανότητές τους. Παράλληλα παρέχουν πρότυπα για την λήψη της βέλτιστης απόφασης κατά την οποία μπορούν να μετρηθούν και άλλες πρακτικές μέθοδοι.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{(class - conditional probability)}$$

- P(A|B): Πιθανότητα να συμβεί το A δοθέντος του B
- P(A): Πιθανότητα να συμβεί το A
- P(B): Πιθανότητα να συμβεί το B
- P(B|A): Πιθανότητα να συμβεί το B δοθέντος του A

Το θεώρημα Bayes εκφράζεται ως:

$$P[\omega_1|x] = \frac{P[x|\omega j]\cdot P[\omega j]}{\sum_{i=1}^{d} P\,[x\,|\,\omega k]\cdot P[\omega k]} = \frac{P[x|\omega j]\cdot P[\omega j]}{P[x]}$$

- όπου ωj η κλάση j και x το διάνυσμα χαρακτηριστικών
- Ένας τυπικός κανόνας απόφασης είναι να επιλέγουμε την κλάση με τη μέγιστη P[ωj|x]
  - P[ωj] εκ των προτέρων πιθανότητα
  - P[ωj|x] εκ των υστέρων πιθανότητα
  - P[x|ωj] πιθανοφάνεια
  - P[x] σταθερά κανονικοποίησης

Ένας ταξινομητής naive Bayes υπολογίζει την class-conditional πιθανότητα θεωρώντας ότι τα χαρακτηριστικά γνωρίσματα είναι υπό όρους ανεξάρτητα μεταξύ τους, δεδομένης της ετικέτας κατηγορίας y.

Η υπό όρους υποθετική ανεξαρτησία μπορεί να δηλωθεί τυπικά ως εξής:

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y)$$

Όπου κάθε σύνολο χαρακτηριστικών γνωρισμάτων $\boldsymbol{X} = \{\boldsymbol{X}1, \boldsymbol{X}2, \ldots, \boldsymbol{X}d\}$ έχει d χαρακτηριστικά γνωρίσματα για κάθε εγγραφή.

Ο κανόνας ταξινόμησης Bayes (για δύο κλάσεις *M*=2). Έστω ότι τοποθετείται ως εξής, σύμφωνα με τον κανόνα που ακολουθεί και επίσης, ο ταξινομητής Bayes είναι άριστος ως προς την ελαχιστοποίηση της πιθανότητας σφάλματος κατάταξης.

$$\text{If } P(\omega_1|\underline{x}) > P(\omega_2|\underline{x}) \;\; \underline{x} \rightarrow \omega_1$$
$$\text{If } P(\omega_2|\underline{x}) > P(\omega_1|\underline{x}) \;\; \underline{x} \rightarrow \omega_2$$

Δεδομένου του θεωρήματος Bayes:

$$P(\omega_1|x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2|x) \rightarrow \frac{P(x|\omega_1)P(\omega_1)}{P(x)} \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \frac{P(x|\omega_2)P(\omega_2)}{P(x)}$$

Η P(x) μπορεί να απλοποιηθεί και μετά από ανακατάταξη της σχέσης προκύπτει ο λόγος πιθανοφάνειας Λ(x) και ο κανόνας απόφασης του Bayes:

$$\Lambda(x) = \frac{P(x|\omega_1)}{P(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_1)}{P(\omega_2)}$$

Ο Βέλτιστος Bayesian ταξινομητής, απλοποιείται σημαντικά όταν:

• Οι κλάσεις είναι ισοπίθανες.

• Τα χαρακτηριστικά σε όλες τις διαστάσεις ακολουθούν κανονική κατανομή για όλες τις κλάσεις.

Έτσι λοιπόν, ένας αφελής ταξινομητής Bayes είναι ένας απλός πιθανοτικός ταξινομητής που βασίζεται στην εφαρμογή του θεωρήματος του Bayes με ισχυρή ανεξαρτησία και υποθέσεις [126]. Αφήστε το *C* να είναι η τυχαία μεταβλητή που δηλώνει την κλάση μιας παρουσίας και το *X* να είναι ένα διάνυσμα τυχαίων μεταβλητών που υποδηλώνουν τις παρατηρούμενες τιμές χαρακτηριστικών. Αφήστε το *c* να είναι μια συγκεκριμένη ετικέτα κλάσης και το *x* αντιπροσωπεύει μια συγκεκριμένη παρατηρούμενη τιμή χαρακτηριστικού. Σύμφωνα με την υπόθεση της ανεξαρτησίας, αποδίδει X_1 , ... , X_n είναι όλα υπό όρους ανεξάρτητες μεταξύ τους, δεδομένου *C* . Η αξία αυτής της υπόθεσης είναι ότι απλοποιεί δραματικά την αναπαράσταση της υπό όρους πιθανότητας *P* (*X* | *C*), και το πρόβλημα της εκτίμησής του από τα δεδομένα εκπαίδευσης . Στην πραγματικότητα, η ακριβής εκτίμηση του *P* (*X* | *C*) απαιτεί συνήθως πολλά παραδείγματα. Για να δούμε γιατί, ας εξετάσουμε τον αριθμό των παραμέτρων που πρέπει να εκτιμήσουμε πότε το *C* είναι boolean και το *X* είναι ένα διάνυσμα *n* boolean χαρακτηριστικών. Σε αυτήν την περίπτωση, πρέπει να εκτιμηθεί το ακόλουθο σύνολο παραμέτρων:

$$\theta_{ij} \equiv P(X = x_i | \ C = c_j)$$

όπου το ευρετήριο *i* παίρνει 2^n πιθανές τιμές (μία για καθεμία από τις πιθανές διανυσματικές τιμές του *X*), και *j* παίρνει 2 πιθανές τιμές. Επομένως, πρέπει να εκτιμηθούν περίπου 2^{n+1} παράμετροι. Για να υπολογίσετε τον ακριβή αριθμό των απαιτούμενων παραμέτρων, σημειώστε για οποιοδήποτε σταθερό *j* , το άθροισμα *i* του ϑ_{ij} πρέπει να είναι ένα. Επομένως, για οποιαδήποτε συγκεκριμένη τιμή c_j , και για τις 2^n πιθανές τιμές του x_i , χρειαζόμαστε υπολογισμό μόνο $2^n - 1$ ανεξάρτητων παραμέτρων. Δεδομένων των δύο πιθανών τιμών για *C* πρέπει να εκτιμήσουμε συνολικά 2 ($2^n - 1$), όπως ϑ_{ij} παράμετροι για την εκμάθηση Bayesian ταξινομητές. Ο αφελής ταξινομητής Bayes, αντ' αυτού, μειώνει αυτήν την πολυπλοκότητα κάνοντας μια υπόθεση ανεξαρτησίας υπό όρους που μειώνει τον αριθμό των παραμέτρων που πρέπει να εκτιμηθούν, κατά τη μοντελοποίηση *P* (*X* | *C*), σχηματίστε το αρχικό 2 ($2^n - 1$) σε μόλις 2*n* . Επιπλέον, για την εκτίμηση του *P* (*C* | *X*), τα δεδομένα εκπαίδευσης μπορούν να χρησιμοποιηθούν για να μάθουν εκτιμήσεις των *P* (*X* | *C*) και *P*(*Γ*). Στη συνέχεια, νέα παραδείγματα *X* μπορούν να ταξινομηθούν χρησιμοποιώντας αυτές τις εκτιμώμενες κατανομές πιθανότητας, συν τον κανόνα Bayes. Αυτός ο τύπος ταξινομητή ονομάζεται *γενετικός* ταξινομητής, επειδή η κατανομή *P* (*X* | *C*) μπορεί να θεωρηθεί ότι περιγράφει τον τρόπο δημιουργίας τυχαίων παρουσιών *X που* εξαρτώνται από το χαρακτηριστικό στόχου *C*.

Εάν έχουμε μια δοκιμαστική περίπτωση *x* για ταξινόμηση, η πιθανότητα κάθε κλάσης δεδομένου του διανύσματος των παρατηρούμενων τιμών για τα προγνωστικά χαρακτηριστικά μπορεί να ληφθεί χρησιμοποιώντας το θεώρημα του Bayes:

$$P(C= c \mid X=x)= \frac{p(C= c)\, p(X=x|C=c)}{p(X=x)}$$

και μετά προβλέποντας την πιο πιθανή τάξη. Επειδή το συμβάν είναι ένας συνδυασμός εκχωρήσεων τιμών χαρακτηριστικών και λόγω της υπόθεσης ανεξαρτησίας υπό όρους, μπορεί να γραφτεί η ακόλουθη εξίσωση:

$$P(X=x|C=c)= \prod_i p(X_i=x_i|C=c)$$

που είναι αρκετά απλό να υπολογιστεί για την εκπαίδευση και τα δεδομένα δοκιμών.

Μια τυπική υπόθεση είναι ότι [127], σε κάθε κατηγορία, οι τιμές των αριθμητικών χαρακτηριστικών κατανέμονται κανονικά. Κάποιος μπορεί να αντιπροσωπεύει μια τέτοια κατανομή ως προς τη μέση και την τυπική απόκλιση, και μπορεί να υπολογιστεί η πιθανότητα μιας παρατηρούμενης τιμής από αυτές τις εκτιμήσεις.

Σε προηγούμενη μελέτη, εξετάστηκαν τρεις διαφορετικές τεχνικές ταξινόμησης (C4.5 tree tree classifier, Multi-Layer Perceptron τεχνητό νευρικό δίκτυο [128] και ο αφελής ταξινομητής Bayes) και διαπιστώθηκε ότι οι αφελείς Bayes έδωσαν τα πιο αξιόπιστα αποτελέσματα, παρόλο που οι απλοποιητικές παραδοχές [128] του παραβιάστηκαν έντονα από τα δεδομένα που αναλύθηκαν. Όπως και ο παραδοσιακός αφελής ταξινομητής Bayesian, ο νέος αλγόριθμος πρέπει να είναι ένα μοντέλο «white-box», στο οποίο ο λόγος για την κατάταξη στην ταξινόμηση [128] μπορεί να προσδιοριστεί ρητά εξετάζοντας το ίδιο το μοντέλο.

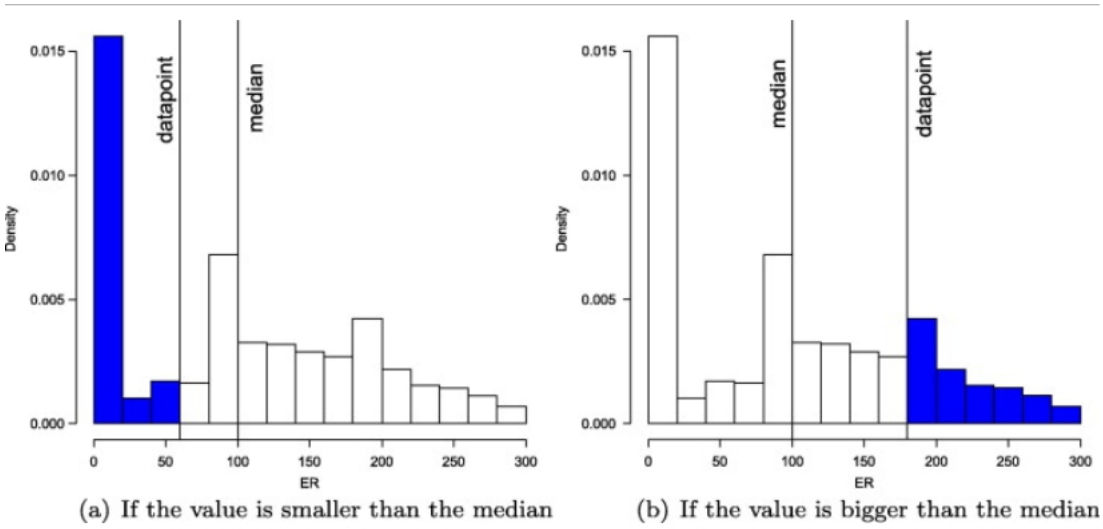
Η κύρια ιδέα του νέου αλγορίθμου είναι ότι όσο πιο κοντά είναι μια μεταβλητή τιμή στο διάμεσο της σε μια συγκεκριμένη κατηγορία [129], τόσο μεγαλύτερη είναι η πιθανότητα να αντιστοιχιστεί σε αυτήν τη συγκεκριμένη ομάδα. Αυτό είναι παρόμοιο με το παραδοσιακό αφελές Bayes, όπου χρησιμοποιείται ο μέσος όρος αντί για τη διάμεσο.

Στην αρχή του αλγορίθμου υπολογίστηκε η μέση τιμή κάθε δυνατότητας σε κάθε κατηγορία και τις πιθανότητες των *προηγούμενων* [129], οι οποίες ορίστηκαν ως ο λόγος μεταξύ του μεγέθους κάθε κατηγορίας (ως προς τον αριθμό των σημείων δεδομένων) και του συνολικού αριθμού των περιπτώσεων.

Το ακόλουθο βήμα είναι το κύριο μέρος της μεθόδου μας στην οποία υπολογίζονται οι μεμονωμένες πιθανότητες [130].

Για κάθε μεταβλητή, ελέγχουμε αν οι τιμές των μεμονωμένων μεταβλητών είναι μικρότερες ή μεγαλύτερες από τη μέση τιμή αυτής της μεταβλητής κατανομής σε κάθε τάξη. Εάν η τιμή είναι μικρότερη, υπολογίζουμε την περιοχή κάτω από το ιστόγραμμα που παραμένει στα αριστερά σε σχέση με την τιμή που αναλύεται . Εάν το ποσό είναι μεγαλύτερο, υπολογίζεται η

περιοχή στη δεξιά πλευρά, λαμβάνοντας υπόψη το τμήμα του ιστογράμματος που οριοθετείται από την τιμή και το μέγιστο. Στη συνέχεια, το ποσό που επιστρέφεται διαιρείται με το ήμισυ των συνολικών παρατηρήσεων, καθώς υποθέτουμε ότι η συνολική περιοχή κάτω από το ιστόγραμμα είναι ίση με μία.



Σχήμα 15: Ανάλυση απόδοσης

3.1.2 Κατηγοριοποιητής Κ πλησιέστερων γειτόνων

Το K-Nearest Neighbours είναι ένας από τους πιο βασικούς αλλά και ουσιαστικούς αλγόριθμους ταξινόμησης στη Μηχανική Εκμάθηση [131]. Ανήκει στον εποπτευόμενο τομέα μάθησης και βρίσκει έντονη εφαρμογή στην αναγνώριση προτύπων, στην εξόρυξη δεδομένων και στον εντοπισμό εισβολών.

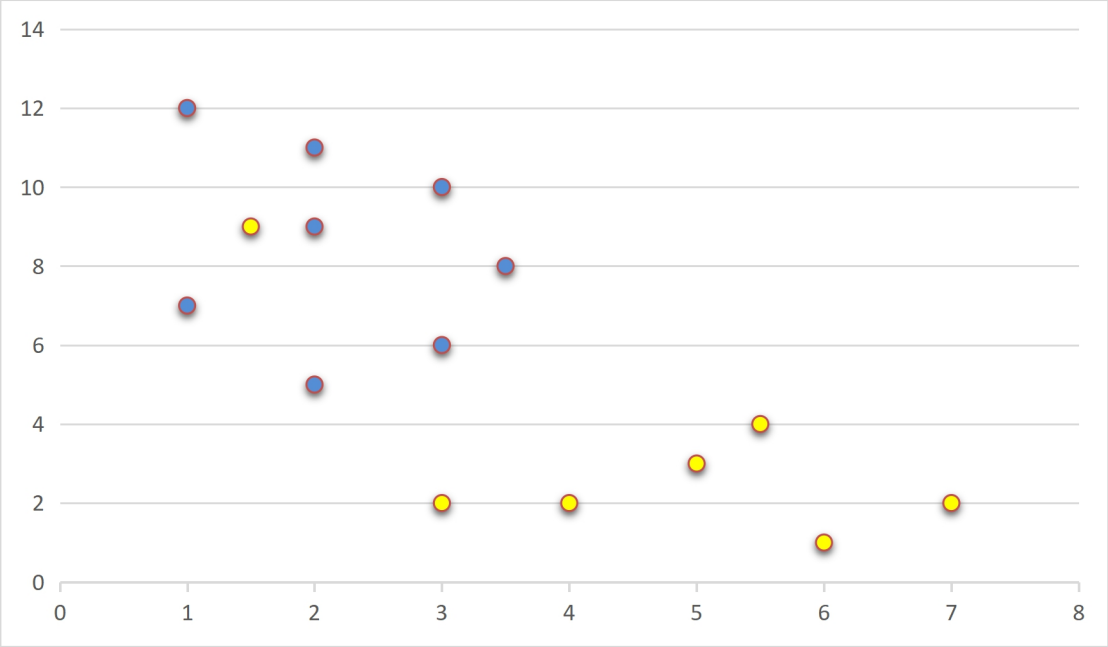
Είναι ευρέως διαθέσιμο σε σενάρια πραγματικής ζωής, δεδομένου ότι είναι μη παραμετρικό, πράγμα που σημαίνει ότι δεν κάνει υποκείμενες υποθέσεις σχετικά με την κατανομή δεδομένων (σε αντίθεση με άλλους αλγόριθμους όπως το GMM , οι οποίοι υποθέτουν μια Gaussian κατανομή των δεδομένων δεδομένων).

Η ταξινόμηση Κ-πλησιέστερου γείτονα (kNN) είναι μια από τις πιο θεμελιώδεις και απλές μεθόδους ταξινόμησης και θα πρέπει να είναι μία από τις πρώτες επιλογές για μια μελέτη ταξινόμησης [132] όταν υπάρχει λίγη ή καθόλου προηγούμενη γνώση σχετικά με τη διανομή

των δεδομένων. Η ταξινόμηση Κ-πλησιέστερου-γείτονα αναπτύχθηκε από την ανάγκη διεξαγωγής διακριτικής ανάλυσης όταν είναι άγνωστες ή [132] δύσκολο να προσδιοριστούν αξιόπιστες παραμετρικές εκτιμήσεις πυκνότητας πιθανότητας.

Μας δίνονται ορισμένα προηγούμενα δεδομένα (που ονομάζονται επίσης δεδομένα εκπαίδευσης) [133], τα οποία ταξινομούν τις συντεταγμένες σε ομάδες που προσδιορίζονται από ένα χαρακτηριστικό.

Για παράδειγμα, εξετάστε τον ακόλουθο πίνακα σημείων δεδομένων που περιέχει δύο δυνατότητες:



Σχήμα 16: Κατηγοριοποίηση στοιχείων

Για να επιλέξετε το Κ που είναι κατάλληλο για τα δεδομένα σας, εκτελούμε τον αλγόριθμο KNN αρκετές φορές με διαφορετικές τιμές του Κ και επιλέγουμε το Κ που μειώνει τον αριθμό των σφαλμάτων που συναντάμε [134], διατηρώντας παράλληλα την ικανότητα του αλγορίθμου να κάνει με ακρίβεια προβλέψεις όταν του δοθούν δεδομένα που δεν έχει ». είδατε πριν.

Ακολουθούν ορισμένα πράγματα που πρέπει να θυμάστε [134]:

1. Καθώς μειώνουμε την τιμή του K σε 1, οι προβλέψεις μας γίνονται λιγότερο σταθερές. Σκεφτείτε απλώς για ένα λεπτό, φανταστείτε το K = 1 και έχουμε ένα σημείο ερωτήματος που περιβάλλεται από πολλά μπλέ και ένα κίτρινο (σκέφτομαι την επάνω αριστερή γωνία της έγχρωμης πλοκής παραπάνω), αλλά το κίτρινο είναι ο μοναδικός πλησιέστερος γείτονας. Λογικά, πιστεύουμε ότι το σημείο ερωτήματος είναι πιθανότατα μπλέ, αλλά επειδή το K = 1, το KNN προβλέπει εσφαλμένα ότι το σημείο ερωτήματος είναι κίτρινο.

2. Αντίστροφα, καθώς μεγαλώνει τιμή του K, οι προβλέψεις μας γίνονται ολοένα και πιο σταθερές λόγω της πλειοψηφίας / μέσου όρου και, κατά συνέπεια, η πιθανότητα να κάνουμε πιο ακριβείς προβλέψεις αυξάνεται (έως ένα συγκεκριμένο σημείο). Τελικά, αρχίζουμε να βλέπουμε έναν αυξανόμενο αριθμό σφαλμάτων. Σε αυτό το σημείο ξέρουμε ότι έχουμε ωθήσει την τιμή του K πάρα πολύ.

3. Σε περιπτώσεις όπου επιλέγουμε την πλειοψηφία (π.χ. επιλέγοντας τη λειτουργία σε πρόβλημα ταξινόμησης) μεταξύ των ετικετών, συνήθως ορίζουμε το K έναν αριθμό τόσο όσο για να έχουμε ένα tiebreak.

Πλεονεκτήματα

1. Είναι καλός στο χειρισμό θορύβου, απλός και εύκολος στην εφαρμογή.

2. Δεν χρειάζεται να δημιουργήσετε ένα μοντέλο, να συντονίσετε πολλές παραμέτρους ή να κάνετε επιπλέον παραδοχές.

3. Ο αλγόριθμος είναι ευέλικτος. Μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση και αναζήτηση (όπως θα δούμε στην επόμενη ενότητα).

Μειονεκτήματα

1. Ο αλγόριθμος γίνεται σημαντικά πιο αργός καθώς αυξάνεται ο αριθμός των παραδειγμάτων ή / και των προβλέψεων / ανεξάρτητων μεταβλητών.

Το κύριο μειονέκτημα του KNN είναι να γίνει πολύ πιο αργός καθώς ο όγκος των δεδομένων αυξάνει, τον καθιστά μια μή πρακτική επιλογή [135] σε περιβάλλοντα όπου οι προβλέψεις

πρέπει να γίνουν γρήγορα. Επιπλέον, υπάρχουν ταχύτεροι αλγόριθμοι που μπορούν να παράγουν ακριβέστερα αποτελέσματα ταξινόμησης και παλινδρόμησης.

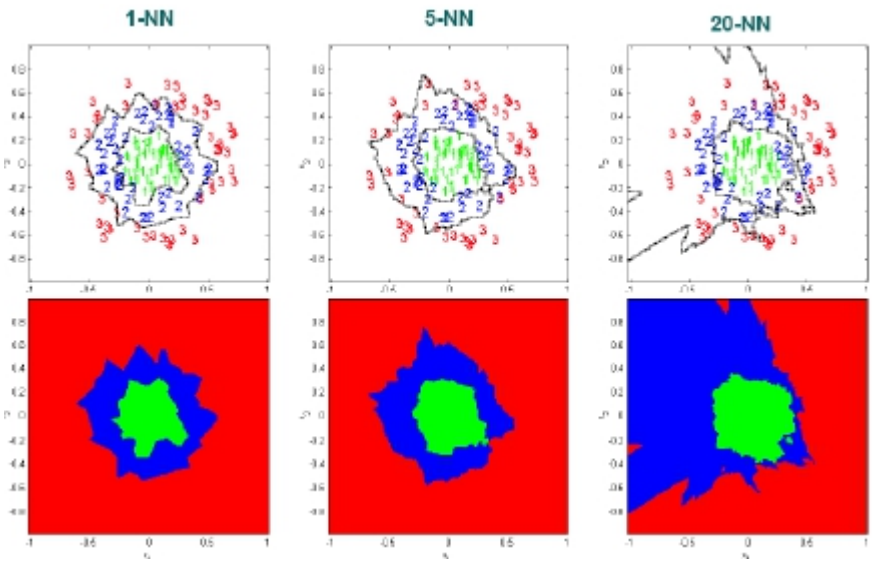
Ωστόσο, με την προϋπόθεση ότι έχετε επαρκείς υπολογιστικούς πόρους για να χειριστείτε γρήγορα τα δεδομένα που χρησιμοποιείτε για να κάνετε προβλέψεις [135], το KNN μπορεί να εξακολουθεί να είναι χρήσιμο στην επίλυση προβλημάτων που έχουν λύσεις που εξαρτώνται από τον εντοπισμό παρόμοιων αντικειμένων. Ένα παράδειγμα αυτού είναι η χρήση του αλγορίθμου KNN σε συστήματα σύστασης, μια εφαρμογή αναζήτησης KNN.

Ο k Nearest Neighbor Rule (kNN) είναι διαισθητική μέθοδος που ταξινομεί άγνωστα δείγματα [135] με βάσει την ομοιότητα τους με τα δείγματα εκπαίδευσης. Όταν έχουμε ώς δεδομένο ένα πρότυπο το οποίο είναι άγνωστο x(u) εντοπίζει τα k «κοντινότερα» δείγματα από τα δεδομένα εκπαίδευσης και αποδίδει την τιμή x(u) στην κλάση που εμφανίζεται πιο πολύ στο k-υποσύνολο.

Ο k-NN ανήκει στην κατηγορία των χαλαρών αλγορίθμων:

- Κατά την ταξινόμηση κάνει επεξεργασία τών δεδομένων εκπαίδευσης
- Αντοποκρίνεται στο αίτημα ταξινόμησης και συνδυάζει τα αποθηκευμένα δεδομένα εκπαίδευσης
- Δεν λαμβάνει υπόψη λογική ή άλλα αποτελέσματα.

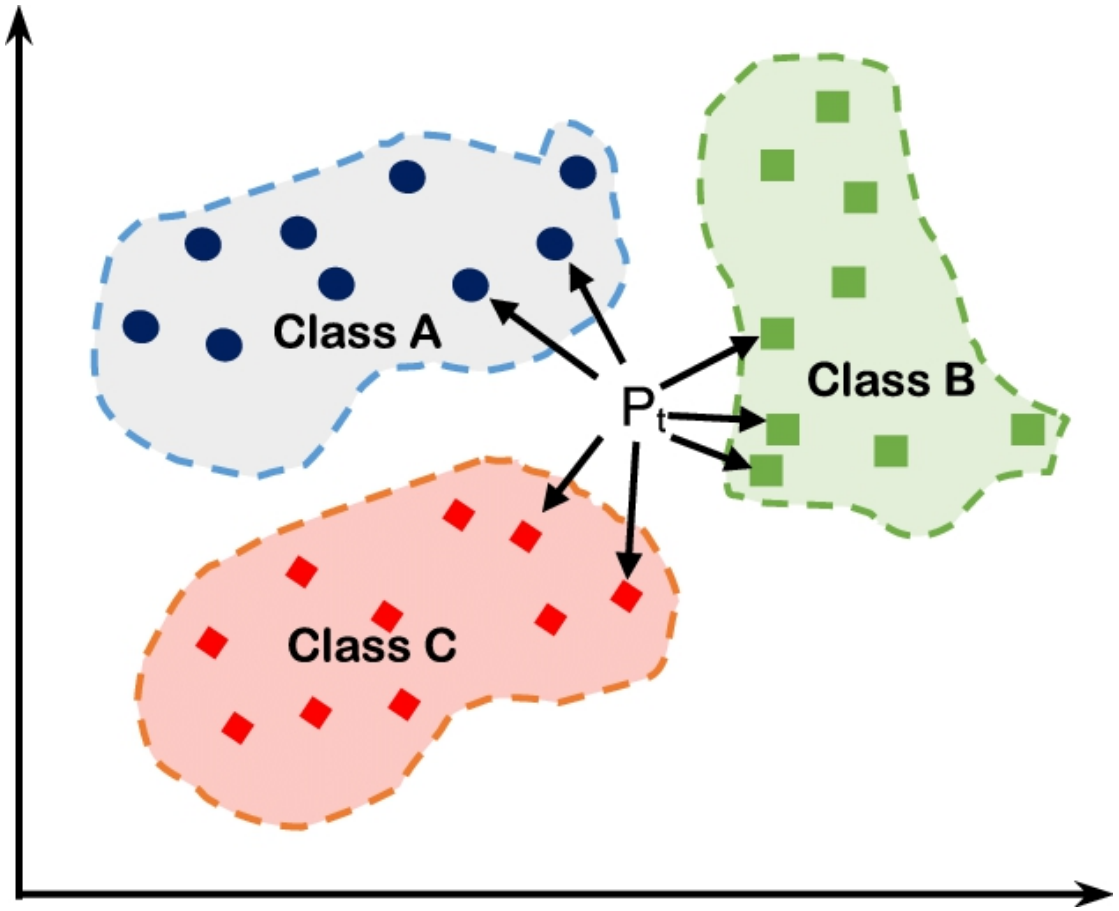
Τα Tradeoffs αλγορίθμων οι οποίοι κατατάσσονται στους χαλαρούς έχουν μικρότερο κόστος υπολογισμού [136] κατά την εκπαίδευση. Όμως οι απαιτήσεις αποθήκευσης και το κόστος υπολογισμού κατά την κλήση τους είναι σαφώς έχουν μεγαλύτερο.



Σχήμα 17: Εφαρμογή κ-NN

Έτσι συγκεντρωτικά τα χαρακτηριστικά του k-NN είναι ότι, έχει απλή υλοποίηση [136], έχει πολύ καλά αποτελέσματα για μεγάλο αριθμό δειγμάτων (N→∞). Από την άλλη όμως έχει μεγάλη απαίτηση σε αποθηκευτικό χώρο και υπολογιστικό κόστος στην κλήση [137]. Ακόμη έχει την περίπτωση όπου είναι ευάλωτος στην «κατάρα των πολυδιάστατων προβλημάτων».

Αν στη μονοδιάστατη περίπτωση (οι παρατηρήσεις είναι απλά μεμονωμένες τιμές και όχι διανύσματα) απαιτούνται N σημεία σε κάθε διάστημα εύρους h τότε στη πολυδιάστατη περίπτωση (διανύσματα μήκους l) απαιτούνται Nl σημεία για καλή εκτίμηση. Η εκθετική αύξηση των απαιτούμενων σημείων με την αύξηση της διάστασης του διανύσματος ονομάζεται κατάρα της διάστασης.



Σχήμα 18: Διαχωρισμός κ-NN

Έτσι, ένα μεγάλο k σημαίνει πιο ομαλές περιοχές αποφάσεων και δίνει πιο σωστές πιθανοτικά πληροφορίες [137], ωστόσο πολύ μεγάλο k μπορεί να χαλάσει την τοπικότητα της απόφασης και αυξάνει το υπολογιστικό κόστος.

Εκτός από την τεχνική ταξινόμησης KNN, υπάρχει επίσης μια τεχνική ταξινόμησης Modified K-Nearest Neighbor (MKNN) που προέρχεται από τον αλγόριθμο ταξινόμησης του KNN αυξάνοντας τον υπολογισμό της εγκυρότητας και της ψηφοφορίας βάρους [143]. Η εγκυρότητα χρησιμοποιείται για τον έλεγχο της εγκυρότητας των δεδομένων εκπαίδευσης και η δοκιμή θα βασίζεται στον αριθμό των γειτόνων σε όλα τα δείγματα δεδομένων εκπαίδευσης.Η ψηφοφορία βάρους χρησιμοποιείται για τον προσδιορισμό του υψηλότερου βάρους του υπολογισμού πολλαπλασιασμού μεταξύ εγκυρότητας με νέα δείγματα δεδομένων (δοκιμή δεδομένων) για τον προσδιορισμό της τελικής κατηγορίας προβλέψεων. Το MKNN

δημιουργήθηκε για να βελτιώσει την ακρίβεια του KNN προσθέτοντας έναν τύπο βήματος που είναι η βαθμίδα εγκυρότητας των δεδομένων εκπαίδευσης και της ψηφοφορίας βάρους [143]. Στο παρελθόν στο KNN, υπολογίστηκε μόνο η γειτονική απόσταση μεταξύ των δεδομένων εκπαίδευσης και των δεδομένων δοκιμών και προσδιορίστηκε η γειτονιά με βάση τον αριθμό K [143]. Ωστόσο, το MKNN θα εκτελέσει τη διαδικασία εγκυρότητας στα δεδομένα εκπαίδευσης προτού υπολογίσει τα δεδομένα εκπαίδευσης με δεδομένα δοκιμών, τότε το υψηλότερο βάρος ή ψηφοφορία [143] από τον πλησιέστερο γείτονα με βάση τον αριθμό του K ή των γειτόνων θα υπολογιστεί από το αποτέλεσμα πολλαπλασιασμού των δεδομένων εκπαίδευσης και των δεδομένων δοκιμών.

Γενικά, ο τύπος απόστασης Ευκλείδειας χρησιμοποιείται για τον προσδιορισμό της απόστασης μεταξύ δύο αντικειμένων εκπαίδευσης και δοκιμών.

$$\text{dxy}= \sqrt{\sum_{i=1}^{n} \quad (x_i - y_i)^2}$$

Χαρακτηριστικό παράδειγμα αποτελεί το εξής, ότι για να επιτύχει τους εθνικούς στόχους του ινδονησιακού κράτους, η κυβέρνηση της Ινδονησίας πρέπει να εκτελέσει τρεις κύριες λειτουργίες, δηλαδή τη λειτουργία ανάπτυξης, τη λειτουργία προστασίας και τη λειτουργία δημόσιας υπηρεσίας [145]. Κατά τη λειτουργία της λειτουργίας δημόσιας υπηρεσίας, η κυβέρνηση πρέπει να λάβει πληροφορίες σχετικά με τις καταγγελίες του κοινού και την κριτική για να κατευθύνει την κυβερνητική ανταπόκριση και την παροχή δημόσιων υπηρεσιών στον απαιτούμενο στόχο. Οι δημόσιες πληροφορίες που λαμβάνονται μέσω SMS ή ιστότοπων πρέπει να ταξινομηθούν σύμφωνα με κατηγορίες και να προωθηθούν στις σχετικές μονάδες εργασίας για να λάβουν άμεση παρακολούθηση. Διαφορετικές κατηγορίες και διαφορετικές μονάδες εργασίας μπορεί να έχουν διαφορετικές πληροφορίες. Οι ποικίλες ποσότητες πληροφοριών μεταξύ κατηγοριών και μονάδων εργασίας στην ταξινόμηση κειμένου προκαλούν ανισορροπία δεδομένων [146]. Για να αντιμετωπίσετε άνισα προβλήματα ταξινόμησης συνόλων δεδομένων, προτείνεται η χρήση του αλγόριθμου K-Nearest Neighbor (KNN) καθώς είναι καλύτερος από τους αλγόριθμους Bayesian και Support Vector Machine (SVM) [147]. Υπάρχει ένα πρόβλημα ασάφειας στην ταξινόμηση των πληροφοριών επειδή ο

καθένας υποβάλλει πληροφορίες ή παράπονα με διαφορετικό στυλ γραφής, παρόλο που μπορεί να σημαίνει το ίδιο. Οι ασαφείς αλγόριθμοι πιστεύεται ότι είναι αποτελεσματικοί για τον χειρισμό περιπτώσεων ασάφειας. Ο αλγόριθμος Fuzzy K-Nearest Neighbor (FKNN) είναι ένας αλγόριθμος βελτίωσης από το K-Nearest Neighbor (KNN) [148] όσον αφορά την πρόβλεψη συμμετοχής στην τάξη δοκιμής. Ο αλγόριθμος Fuzzy K-Nearest Neighbor μπορεί επίσης να ξεπεράσει την άνιση κατανομή εγγράφων ή ανισορροπημένων δεδομένων και τυχών ομοιόμορφα χαρακτηριστικά εγγράφων .

Η ταξινόμηση κειμένου χρησιμοποιώντας το K-Means αμφίδρομα αυξάνει την απόδοση ταξινόμησης σε μη ισορροπημένα σύνολα δεδομένων, ειδικά σε περιπτώσεις όπου μια ανισορροπία τάξης επηρεάζει έντονα την απόδοση ταξινόμησης με ποσοστό ακρίβειας 88% . Η ταξινόμηση για ημι-εποπτευόμενη εκμάθηση κειμένου χρησιμοποιείται ως αλγόριθμος μη ισορροπημένης ταξινόμησης με ποσοστό ακρίβειας 92%. Για να βελτιωθεί η ακρίβεια στα μη ισορροπημένα δεδομένα, χρησιμοποιείται προσέγγιση με βάση τη δομή Smote [141] έως και 95,5%.

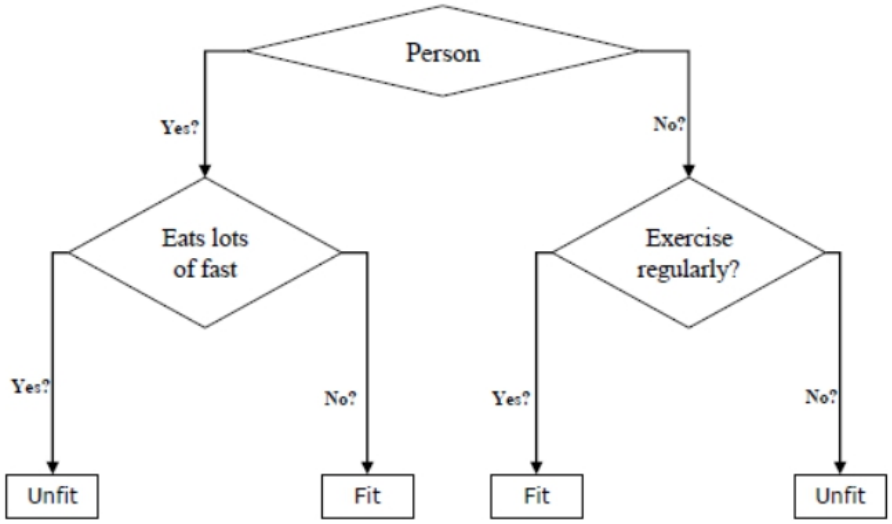
3.2 Κατηγοριοποιητής με δέντρο αποφάσεων

Το δέντρο αποφάσεων είναι το πιο ισχυρό και δημοφιλές εργαλείο ταξινόμησης και πρόβλεψης [144]. Πρόκειτε για ένα διάγραμμα ροής με διακλαδούμενη δομή δέντρου, όπου κάθε κόμβος αναπαριστά μια δοκιμή σε ένα χαρακτηριστικό, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα της δοκιμής και κάθε κόμβος φύλλων (κόμβος τερματικού) κρατά μια ετικέτα κλάσης.

Γενικά, η ανάλυση δένδρων αποφάσεων είναι ένα εργαλείο πρόβλεψης μοντελοποίησης που μπορεί να εφαρμοστεί σε πολλούς τομείς. Τα δέντρα αποφάσεων μπορούν να κατασκευαστούν [149] με μια αλγοριθμική προσέγγιση που μπορεί να χωρίσει το σύνολο δεδομένων με διαφορετικούς τρόπους με βάση διαφορετικές συνθήκες. Τα δέντρα

αποφάσεων είναι οι πιο ισχυροί αλγόριθμοι που εμπίπτουν στην κατηγορία εποπτευόμενων αλγορίθμων.

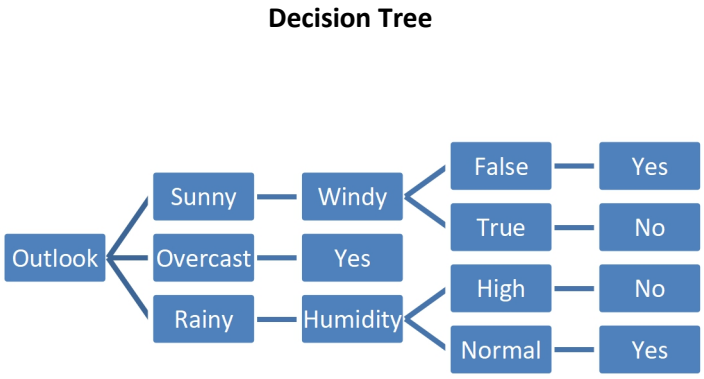
Μπορούν να χρησιμοποιηθούν για εργασίες ταξινόμησης και παλινδρόμησης. Οι δύο κύριες οντότητες ενός δέντρου είναι κόμβοι αποφάσεων, όπου τα δεδομένα χωρίζονται και φεύγουν [149] και οι κλάδοι όπου αντιπροσωπεύουν το αποτέλεσμα. Το παράδειγμα ενός δυαδικού δέντρου για την πρόβλεψη εάν ένα άτομο είναι κατάλληλο ή ανίκανο παρέχοντας διάφορες πληροφορίες όπως η ηλικία, οι διατροφικές συνήθειες και οι συνήθειες άσκησης, δίνεται παρακάτω:



Σχήμα 19: Αλγόριθμος δέντρων

Δημιουργούνται μοντέλα ταξινόμησης ή παλινδρόμησης με τη μορφή δέντρου. Μετατρέπει ένα σύνολο δεδομένων σε ολοένα και μικρότερα υποσύνολα ενώ ταυτόχρονα αναπτύσσεται σταδιακά το δέντρο αποφάσεων. Το τελικό αποτέλεσμα [150] είναι ένα δέντρο με κόμβους αποφάσεων και κόμβους φύλλων . Ένας κόμβος αποφάσεων (π.χ. Outlook) έχει δύο ή περισσότερους κλάδους (π.χ. Sunny, Overcast και Rainy). Ο κόμβος φύλλων (π.χ. Play) αναπαριστά μια απόφαση. Ο κορυφαίος κόμβος απόφασης σε ένα δέντρο που αντιστοιχεί στον καλύτερο προγνωστικό παράγοντα που ονομάζεται root node [150] . Τα δέντρα αποφάσεων μπορούν να χειριστούν τόσο κατηγορικά όσο και αριθμητικά δεδομένα.

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overoast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overoast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overoast	Mild	High	True	Yes
Overoast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Σχήμα 20: Decision Tree

Η ταξινόμηση εικόνας έχει αποτελέσει σημαντικό μέρος των πεδίων της τηλεπισκόπησης, της ανάλυσης εικόνας και της αναγνώρισης προτύπων. Η ταξινόμηση έχει επίσης θεωρηθεί ως μέσο συμπίεσης δεδομένων εικόνας μειώνοντας το μεγάλο εύρος DN σε πολλές φασματικές ζώνες σε μερικές κατηγορίες, σε μία μόνο εικόνα [151]. Η ταξινόμηση μειώνει αυτόν τον μεγάλο φασματικό χώρο σε σχετικά λίγες περιοχές και προφανώς οδηγεί σε απώλεια αριθμητικών πληροφοριών από την αρχική εικόνα. Οι παραδοσιακά χρησιμοποιούμενες μέθοδοι ταξινόμησης με βάση τα pixel [151] βασίζονται σε συμβατικές στατιστικές τεχνικές και δεν επιλύουν τη σύγχυση μεταξύ των τάξεων. Ως αποτέλεσμα, τα τελευταία χρόνια, έχουν προταθεί εναλλακτικές στρατηγικές, ιδίως η χρήση τεχνητών νευρικών δικτύων και δέντρων αποφάσεων, μεθόδων που προέρχονται από τη θεωρία των ασαφών συνόλων και την ενσωμάτωση δευτερευόντων πληροφοριών [152] όπως χαρακτηριστικά υφής, περιβάλλοντος και εδάφους. Η ταξινόμηση εικόνας υψηλής ανάλυσης παρουσιάζεται [152] βάσει ασαφών κανόνων με τη βοήθεια περιγραφών όπως: μορφή, υφή και σχέσεις μεταξύ αντικειμένων και υπο-αντικειμένων.

Πραγματοποιήθηκε μια μελέτη σε μια μικρή περιοχή που χρησιμοποιεί δεδομένα QuickBird για σύγκριση της αντικειμενοστραφής με προσέγγιση ταξινόμησης βάσει pixel [153]. Η αντικειμενοστραφής ταξινόμηση δεν ταξινομεί μεμονωμένα εικονοστοιχεία αλλά αντικείμενα που δημιουργήθηκαν κατά τη διαδικασία τμηματοποίησης πολλαπλών αναλύσεων, η οποία επιτρέπει τη χρήση, όχι μόνο φασματικών αποκρίσεων, αλλά και υφής, περιβάλλοντος και πληροφοριών από άλλα στρώματα αντικειμένων [153]. Οι ασαφείς λογικοί κανόνες εφαρμόζονται στην κατασκευή της ιεραρχίας της τάξης. Συγκρίνονται οι δύο τεχνικά και θεωρητικά διαφορετικές τεχνικές επεξεργασίας εικόνας που βασίζονται σε μεθόδους που αντλούν χωρικά σαφείς πολυκλίμακες πληροφορίες [153] με βάση τα συμφραζόμενα από μία ανάλυση εικόνων τηλεπισκόπησης.

Το δέντρο αποφάσεων δημιουργήθηκε χρησιμοποιώντας το λογισμικό δέντρου αποφάσεων See5 [154]. Το κύριο πλεονέκτημα του See5 είναι ότι μπορεί να μετατρέψει ένα δέντρο αποφάσεων σε κανόνες ταξινόμησης. Οι δορυφορικές εικόνες Thematic Mapper (TM) είναι σε μορφή BSQ [155] . Η δομή Peano Count Tree (P-tree) χρησιμοποιείται για την κατασκευή του ταξινομητή . Τα P-tree αντιπροσωπεύουν χωρικά δεδομένα bit-by-bit σε μια αναδρομική διάταξη τεταρτημορίων. Κάθε νέο συστατικό σε μια ροή χωρικών δεδομένων μετατρέπεται σε δέντρα P και στη συνέχεια προστίθεται στο σετ προπόνησης το συντομότερο δυνατό [154]. Τα P-trees [155] μπορούν να δημιουργηθούν αρκετά γρήγορα και μπορούν να θεωρηθούν ως "έτοιμο για εξόρυξη δεδομένων" και χωρίς απώλειες μορφή για την αποθήκευση χωρικών ή σχετικών δεδομένων. Το σύνολο δεδομένων εκπαίδευσης για το δέντρο αποφάσεων είναι διαφορετικό καθώς δεν είναι παραμετρικός ταξινομητής. Στο P-tree, το πρώτο βήμα αλλάζει τη μορφή δεδομένων σε μορφή bSQ [154]. Στη συνέχεια, το P-δέντρο κατασκευάστηκε εφαρμόζοντας τον αλγόριθμο κανονικού δέντρου αποφάσεων.

Υπάρχει έλλειψη στις μεθόδους ταξινόμησης που χρησιμοποιούν ταξινόμηση βάσει γνώσης χρησιμοποιώντας P-tree και αξιολόγηση ακρίβειας χρησιμοποιώντας εικόνα IRS 1D LISS III [155] χρησιμοποιώντας όλες τις τεχνικές ταξινόμησης. Η εφαρμογή της δομής δεδομένων P-tree για την ταξινόμηση είναι κυρίως για τη βελτίωση της ακρίβειας της ταξινόμησης. Ως εκ τούτου, το

έγγραφο αυτό δίνει έμφαση στην αξιολόγηση της ακρίβειας της ταξινόμησης εικόνας χρησιμοποιώντας τη δομή δεδομένων p-tree και μια συγκριτική ανάλυση της αντικειμενοστραφούς ταξινόμησης, της Γνωσιακής Βάσης Ταξινόμησης και του PTC (P-tree Classifier).

3.2.1 Αλγόριθμος δέντρου αποφάσεων

Ο βασικός αλγόριθμος για την κατασκευή δένδρων αποφάσεων που ονομάζεται ID3 από τον JR Quinlan [158], ο οποίος χρησιμοποιεί μια άπληστη τεχνική, άπληστη αναζήτηση μέσω του χώρου των πιθανών κλάδων. Το ID3 χρησιμοποιεί το Entropy και το Information Gain για να δημιουργήσει ένα δέντρο αποφάσεων. Στο μοντέλο ZeroR δεν υπάρχει πρόβλεψη, στο μοντέλο OneR προσπαθούμε να βρούμε [158] τον μοναδικό καλύτερο προγνωστικό παράγοντα, το αφελές Bayesian περιλαμβάνει όλους τους προγνωστικούς παράγοντες που χρησιμοποιούν τον κανόνα Bayes και τις παραδοχές ανεξαρτησίας μεταξύ των προβλέψεων, αλλά το δέντρο αποφάσεων περιλαμβάνει όλους τους προγνωστικούς παράγοντες με τις παραδοχές εξάρτησης μεταξύ των προβλέψεων.

Ένα δέντρο [158] μπορεί να *«μάθει»* διαχωρίζοντας το σύνολο πηγής σε υποσύνολα με βάση μια δοκιμή τιμής χαρακτηριστικού. Αυτή η διαδικασία επαναλαμβάνεται σε κάθε παράγωγο υποσύνολο με έναν αναδρομικό τρόπο που ονομάζεται *recursive partitioning* . Η αναδρομή ολοκληρώνεται όταν το υποσύνολο σε έναν κόμβο έχει την ίδια τιμή της μεταβλητής στόχου ή όταν ο διαχωρισμός δεν προσθέτει πλέον αξία στις προβλέψεις [159]. Η κατασκευή του ταξινομητή δέντρου αποφάσεων δεν απαιτεί καμία γνώση τομέα ή ρύθμιση παραμέτρων και επομένως είναι κατάλληλη για την ανακάλυψη διερευνητικών γνώσεων. Τα δέντρα απόφασης μπορούν να χειριστούν δεδομένα διαστάσεων. Γενικά [158], ο ταξινομητής δέντρων αποφάσεων έχει καλή ακρίβεια. Η επαγωγή δέντρων απόφασης είναι μια τυπική επαγωγική προσέγγιση για να μάθετε γνώσεις σχετικά με την ταξινόμηση.

Το δέντρο αποφάσεων στο παραπάνω σχήμα(20) ταξινομεί ένα συγκεκριμένο πρωί ανάλογα με το αν είναι κατάλληλο για παιχνίδι τένις και επιστροφή της κατάταξης που σχετίζεται με το

συγκεκριμένο φύλλο [158]. (Στην περίπτωση αυτή Ναι ή Όχι). Για παράδειγμα, η παρουσία

(Προοπτική = Βροχή, Θερμοκρασία = Καυτή, Υγρασία = Υψηλή, Άνεμος = Ισχυρή)

θα ταξινομηθεί στον αριστερότερο κλάδο αυτού του δέντρου αποφάσεων και επομένως θα ταξινομηθεί ως αρνητικό παράδειγμα.

Πλεονεκτήματα και αδυναμία προσέγγισης

του δέντρου αποφάσεων τα πλεονεκτήματα των μεθόδων δέντρων αποφάσεων είναι:

- Τα δέντρα απόφασης μπορούν να δημιουργήσουν κατανοητούς κανόνες.
- Τα δέντρα αποφάσεων εκτελούν ταξινόμηση χωρίς να απαιτούν πολύ υπολογισμό.
- Τα δέντρα αποφάσεων είναι σε θέση να χειρίζονται τόσο συνεχείς όσο και κατηγορηματικές μεταβλητές.
- Τα δέντρα αποφάσεων παρέχουν μια σαφή ένδειξη ποια πεδία είναι πιο σημαντικά για την πρόβλεψη ή την ταξινόμηση.

Οι αδυναμίες των μεθόδων δέντρων αποφάσεων:

- Τα δέντρα αποφάσεων είναι λιγότερο κατάλληλα για εργασίες εκτίμησης όπου ο στόχος είναι να προβλεφθεί η αξία ενός συνεχούς χαρακτηριστικού.
- Τα δέντρα αποφάσεων είναι επιρρεπή σε σφάλματα και σε προβλήματα ταξινόμησης με πολλές τάξεις και σχετικά μικρό αριθμό παραδειγμάτων εκπαίδευσης.
- Το δέντρο αποφάσεων μπορεί να είναι υπολογιστικά ακριβό στην προπόνηση. Η διαδικασία ανάπτυξης ενός δέντρου αποφάσεων είναι υπολογιστικά ακριβή. Σε κάθε κόμβο, κάθε υποψήφιο πεδίο διαχωρισμού πρέπει να ταξινομηθεί πριν να βρεθεί η καλύτερη διάσπασή του. Σε ορισμένους αλγόριθμους, χρησιμοποιούνται συνδυασμοί πεδίων και πρέπει να γίνει αναζήτηση για βέλτιστο συνδυασμό βαρών. Οι αλγόριθμοι κλαδέματος μπορεί επίσης να είναι δαπανηροί, καθώς πολλά υποψήφια υποδέντρα πρέπει να σχηματιστούν και να συγκριθούν.

Ο αλγόριθμος επαγωγής δέντρων αποφάσεων λειτουργεί αναδρομικά επιλέγοντας το καλύτερο χαρακτηριστικό για τη διάσπαση των δεδομένων και την επέκταση των κόμβων φύλλων του δέντρου έως ότου επιτευχθεί η διακοπή του κύκλου [159]. Η επιλογή της βέλτιστης κατάστασης δοκιμής διάσπασης καθορίζεται με σύγκριση της ακαθαρσίας των θυγατρικών κόμβων και εξαρτάται επίσης από το ποια μέτρηση προσμείξεων χρησιμοποιείται. Μετά τη δημιουργία του δέντρου αποφάσεων, μπορεί να πραγματοποιηθεί ένα βήμα κλαδέματος δέντρων για τη μείωση του μεγέθους του δέντρου αποφάσεων [160]. Τα δέντρα αποφάσεων που είναι πολύ μεγάλα είναι ευαίσθητα σε ένα φαινόμενο γνωστό ως υπερβολικό. Το κλάδεμα βοηθά κόβοντας τα κλαδιά του δέντρου initail με τρόπο που βελτιώνει την ικανότητα γενίκευσης του δέντρου αποφάσεων.

Ακολουθεί ένα παράδειγμα αναδρομικής συνάρτησης που χτίζει το δέντρο επιλέγοντας τα καλύτερα κριτήρια διαίρεσης για το δεδομένο σύνολο δεδομένων. Ονομάζεται με λίστα σειρών και, στη συνέχεια, βγαίνει σε κάθε στήλη (εκτός από την τελευταία, που έχει το αποτέλεσμα σε αυτήν), βρίσκει κάθε πιθανή τιμή για αυτήν τη στήλη και διαιρεί το σύνολο δεδομένων σε δύο νέα υποσύνολα [161]. Υπολογίζει τη μέση εντροπία μέσου όρου για κάθε ζεύγος νέων υποομάδων πολλαπλασιάζοντας την εντροπία κάθε συνόλου με το κλάσμα των στοιχείων που κατέληξαν σε κάθε σετ και θυμάται ποιο ζεύγος έχει τη χαμηλότερη εντροπία [162]. Εάν το καλύτερο ζεύγος υποομάδων δεν έχει χαμηλότερη σταθμισμένη μέση εντροπία από το τρέχον σετ, αυτός ο κλάδος τελειώνει και αποθηκεύονται οι μετρήσεις των πιθανών αποτελεσμάτων. Διαφορετικά, το buildtree καλείται σε κάθε σύνολο και προστίθενται στο δέντρο.

```
def buildtree (σειρές, skorf = εντροπία):
if len (σειρές) == 0: επιστροφή κωδικός απόφασης ()
current_score = scoref (σειρές)
# Ρυθμίστε μερικές μεταβλητές για να παρακολουθείτε τα καλύτερα κριτήρια
best_gain = 0,0
best_criteria = Κανένα
best_sets = Κανένα
στήλη_count = len (σειρές [0]) - 1
για col σε εύρος (0, στήλη_count):
# Δημιουργήστε τη λίστα διαφορετικών τιμών στο
# αυτήν τη στήλη
στήλες_τιμές = {}
για σειρά σε σειρές:
Τιμές στήλης [σειρά [στήλη]] = 1
# Τώρα δοκιμάστε να διαιρέσετε τις σειρές για κάθε τιμή
# σε αυτήν τη στήλη
για τιμή στη στήλη_values.keys ():
(set1, set2) = divideset (σειρές, στήλες, τιμή)
# Απόκτηση πληροφοριών
p = float (len (set1)) / len (σειρές)
κέρδος = current_score-p * scoref (set1) - (1-p) * scoref (set2)
εάν κέρδος> best_gain και len (set1)> 0 και len (set2)> 0:
best_gain = κέρδος
best_criteria = (στήλη, τιμή)
best_sets = (set1, set2)
# Δημιουργήστε τα υποκαταστήματα
εάν best_gain> 0:
trueBranch = buildtree (best_sets [0])
falseBranch = buildtree (best_sets [1])
Returnnnode απόφασης (col = best_criteria [0], value = best_criteria [1],
tb = trueBranch, fb = falseBranch)
αλλού:
κωδικός απόφασης επιστροφής (αποτελέσματα = unececounts (σειρές))
```

Σχήμα 21: Ψευδογλώσσα Decision Tree

Η εύρεση ενός βέλτιστου δέντρου αποφάσεων είναι ένα πλήρες πρόβλημα NP. Πολλοί αλγόριθμοι δέντρων αποφάσεων χρησιμοποιούν μια ευρετική προσέγγιση ή άπληστη στρατηγική για να καθοδηγήσουν την αναζήτησή τους στον τεράστιο χώρο υπόθεσης [163].

Η τεχνική ταξινόμησης είναι μια συστηματική προσέγγιση για τη δημιουργία μοντέλων ταξινόμησης από ένα σύνολο δεδομένων εισόδου. Για παράδειγμα, οι ταξινομητές δέντρων αποφάσεων [163], οι ταξινομητές βάσει κανόνων, τα νευρικά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης και οι αφελείς ταξινομητές Bayes είναι διαφορετικές τεχνικές για την επίλυση ενός προβλήματος ταξινόμησης. Κάθε τεχνική υιοθετεί έναν αλγόριθμο εκμάθησης για τον

προσδιορισμό ενός μοντέλου που ταιριάζει καλύτερα στο σχέδιο μεταξύ του συνόλου χαρακτηριστικών και της ετικέτας κλάσης των δεδομένων εισαγωγής. Επομένως, ένας βασικός στόχος του αλγορίθμου εκμάθησης είναι η δημιουργία μοντέλου πρόβλεψης που προβλέπει με ακρίβεια τις ετικέτες τάξης των προηγούμενων μη αναγνωρισμένων εγγραφών.

Το Decision Tree Classifier [163] είναι μια απλή και ευρέως χρησιμοποιούμενη τεχνική ταξινόμησης. Εφαρμόζει μια απλή ιδέα για την επίλυση του προβλήματος ταξινόμησης. Το Decision Tree Classifier θέτει μια σειρά προσεκτικά επεξεργασμένων ερωτήσεων σχετικά με τα χαρακτηριστικά του δοκιμαστικού αρχείου. Κάθε φορά που λαμβάνει μια απάντηση [164], υποβάλλεται μια ερώτηση παρακολούθησης έως ότου επιτευχθεί ένα συμπέρασμα σχετικά με την ετικέτα class του δίσκου.

Οι τεχνικές του δέντρου αποφάσεων κατασκευής είναι γενικά υπολογιστικά φθηνές, καθιστώντας δυνατή την ταχεία κατασκευή μοντέλων ακόμη και [165] όταν το μέγεθος του σετ κατάρτισης είναι πολύ μεγάλο. Επιπλέον, μόλις δημιουργηθεί ένα δέντρο αποφάσεων, η ταξινόμηση της δοκιμής είναι εξαιρετικά γρήγορη.

Παρακάτω, χρησιμοποιείται ο αλγόριθμος C4.5 [168], αυτός είναι ένας από τους τρόπους που χρησιμοποιούνται στο δέντρο αποφάσεων.

Ο αλγόριθμος C4.5 συμπληρώνεται το 1993 από τον Ross Quinlan, ο οποίος επίσης εισήγαγε τον αλγόριθμο ID3, την προηγούμενη έκδοση του C4.5.

$$\text{Gain (p,T)= Entropy(p)} - \sum_{i=1}^{n} (p_j \text{ x Entropy}(p_j))$$

where

$$\text{Entropy(P)= -} \sum_{i=1}^{n} (p_i \text{ x log}(p_i))$$

3.3 Αξιολόγηση κατηγοριοποίησης

Για την εκτίμηση της απόδοσης ενός ταξινομητή χρησιμοποιούμε διανύσματα, για τα οποία γνωρίζουμε την κλάση στην οποία ανήκουν (ground truthdata). Οι αποφάσεις που παίρνει ο ταξινομητής για τα διανύσματα αυτά τοποθετούνται στον πίνακα συγχύσεων (confusion matrix):

	Ταξινομήθηκαν		
		w1	w2
Ανήκουν	w1	x1	y1
	w2	y2	x2

Πίνακας 4: Confusion Matrix

Για να προχωρήσουμε στήν αξιολόγηση της ενός ταξινομητή η ενός μοντέλου ταξινόμησης σχετικά με την απόδοση τους [169] βασίζόμαστε στο test set, ο αριθμός των εγγραφών οι οποίες προβλέφθηκαν είτε σωστά είτε λάθος από τον ταξινομητή.

Οι **δείκτες επίδοσης** χρησιμοποιούνται για να μπορέσουμε να συγκρίνουμε τις αποδόσεις απο μοντέλο σε μοντέλο, ορισμένα παραδείγματα είναι η **ακρίβεια (accuracy)** και αντίστροφα η αποτίμηση σφάλματος (error rate) [170]. Παρακάτω παρουσιάζονται οι 2 προαναφερθείσες έννοιες της ακρίβειας και της αποτίμησης σφάλματος.

$$\text{Accuracy} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$\text{Error rate} = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions}$$

Έτσι ο πιο αποτελεσματικός ταξινομητής (με τίς πιό καλές προβλέψεις) είναι αυτός ο με το μεγαλύτερο **accuracy** και με το μικρότερο **error rate**. Επίσης, όταν η ταξινόμηση είναι δύο

κλάσεων, τότε υπάρχουν επιπλέον οι δείκτες επίδοσης **Sensitivity/recall(Ευαισθησίας)**και **Specificity(Ειδικότητας)**.

$$Sensitivity = \frac{TP}{TP+FN}$$

TP: true positive classified cases

TN: true negative classified cases

FP: false positive classified cases

$$Specificity = \frac{TN}{TN+FP}$$

FN: false negative classified cases

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy

Accuracy είναι αυτό που συνήθως σημαίνει, όταν χρησιμοποιούμε τον όρο ακρίβεια. Είναι ο λόγος του αριθμού των σωστών προβλέψεων προς τον συνολικό αριθμό των δειγμάτων εισόδου.

$$\text{Accuracy} = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Λειτουργεί καλά μόνο εάν υπάρχει ίσος αριθμός δειγμάτων που ανήκουν σε κάθε τάξη.

Για παράδειγμα, λάβετε υπόψη ότι υπάρχουν 96% δείγματα της κατηγορίας Α και 4% δείγματα της κατηγορίας Β στο εκπαιδευτικό μας σετ. Στη συνέχεια, το μοντέλο μας μπορεί εύκολα να αποκτήσει **ακρίβεια προπόνησης 96%** προβλέποντας απλώς κάθε δείγμα εκπαίδευσης που ανήκει στην τάξη Α.

Όταν το ίδιο μοντέλο δοκιμάζεται σε ένα σύνολο δοκιμών με δείγματα 65% της κατηγορίας Α και 35% δείγματα της κατηγορίας Β, τότε η **ακρίβεια** της **δοκιμής θα μειωθεί στο 65%.** Η

ακρίβεια δεν είναι μεγάλη, αλλά μας δίνει την ψευδή αίσθηση ότι επιτυγχάνουμε υψηλή ακρίβεια.

Το πραγματικό πρόβλημα προκύπτει, όταν το κόστος της εσφαλμένης ταξινόμησης των δειγμάτων δευτερεύουσας κατηγορίας είναι πολύ υψηλό. Εάν αντιμετωπίσουμε μια σπάνια όμως πολύ σοβαρή αρρώστια, η μή διάγνωση της ασθένιας ενός ασθενή έχει μεγαλύτερο κόστος και συνέπιες απο της αποστολή ενός υγιούς ατόμου σε περισσότερες και ίσως κατά μια ένοια περιττές εξετάσεις.

Λογαριθμική απώλεια

Logarithmic Loss ή Log Loss, λειτουργεί τιμωρώντας τις ψευδείς ταξινομήσεις. Λειτουργεί καλά για την ταξινόμηση πολλαπλών κατηγοριών. Όταν εργάζεστε με το Log Loss, ο ταξινομητής πρέπει να εκχωρήσει πιθανότητα σε κάθε τάξη για όλα τα δείγματα. Ας υποθέσουμε ότι υπάρχουν N δείγματα που ανήκουν σε τάξεις M και στη συνέχεια η απώλεια Log υπολογίζεται ως εξής:

$$\text{LogarithmicLoss} = \frac{-1}{N}\sum_{i=1}^{N}\ \sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

όπου,

y_ij, δείχνει αν το δείγμα ανήκει στην κλάση j ή όχι

p_ij, δείχνει την πιθανότητα του δείγματος i να ανήκει στην κλάση j

Το Log Loss δεν έχει άνω όριο και υπάρχει στο εύρος [0, ∞). Η απώλεια καταγραφής πλησιέστερα στο 0 υποδηλώνει υψηλότερη ακρίβεια, ενώ εάν η απώλεια καταγραφής είναι μακριά από το 0, τότε υποδεικνύει χαμηλότερη ακρίβεια.

Γενικά, η ελαχιστοποίηση της απώλειας καταγραφής δίνει μεγαλύτερη ακρίβεια για τον ταξινομητή.

Πίνακας σύγχυσης

Το Confusion Matrix όπως υποδηλώνει το όνομα μας δίνει έναν πίνακα ως έξοδο και περιγράφει την πλήρη απόδοση του μοντέλου.

Ας υποθέσουμε ότι έχουμε πρόβλημα δυαδικής ταξινόμησης. Έχουμε μερικά δείγματα που ανήκουν σε δύο τάξεις: ΝΑΙ ή ΟΧΙ. Επίσης, έχουμε τον δικό μας ταξινομητή που προβλέπει μια κλάση για ένα δεδομένο δείγμα εισόδου. Κατά τη δοκιμή του μοντέλου μας σε 165 δείγματα, έχουμε το ακόλουθο αποτέλεσμα.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Πίνακας 5: Πίνακας Σύγχυσης

Πίνακας σύγχυσης

Υπάρχουν 4 σημαντικοί όροι:

- **True Positives** : Οι περιπτώσεις στις οποίες προβλέψαμε ΝΑΙ και η πραγματική έξοδος ήταν επίσης ΝΑΙ.
- **True Negatives** : Οι περιπτώσεις στις οποίες προβλέψαμε ΟΧΙ και η πραγματική έξοδος ήταν ΟΧΙ.

- **False Positives** : Οι περιπτώσεις στις οποίες προβλέψαμε ΝΑΙ και η πραγματική έξοδος ήταν ΟΧΙ.
- **False Negatives** : Οι περιπτώσεις στις οποίες προβλέψαμε ΟΧΙ και η πραγματική παραγωγή ήταν ΝΑΙ.

Η ακρίβεια για τον πίνακα μπορεί να υπολογιστεί λαμβάνοντας μέση τιμή των τιμών που βρίσκονται στην **«κύρια διαγώνια»,** δηλαδή

$$\text{Accuracy} = \frac{TruePositive + TrueNegative}{TotalSample}$$

$$\text{Accuracy} = \frac{100+50}{165} = 0.91$$

Το Confusion Matrix αποτελεί τη βάση για τους άλλους τύπους μετρήσεων.

Περιοχή κάτω από την καμπύλη

Η περιοχή κάτω από την καμπύλη (AUC) είναι μια από τις πιο ευρέως χρησιμοποιούμενες μετρήσεις για αξιολόγηση. Χρησιμοποιείται για πρόβλημα δυαδικής ταξινόμησης. *Η AUC* ενός ταξινομητή ισούται με την πιθανότητα ότι ο ταξινομητής θα κατατάξει ένα τυχαία επιλεγμένο θετικό παράδειγμα υψηλότερο από ένα τυχαία επιλεγμένο αρνητικό παράδειγμα. Πριν από τον ορισμό της *AUC* , ας κατανοήσουμε δύο βασικούς όρους:

- **True Positive Rate (Sensitivity)** : Το True Positive Rate ορίζεται ως *TP / (FN + TP)* . Το True Positive Rate αντιστοιχεί στο ποσοστό θετικών σημείων δεδομένων που θεωρούνται σωστά ως θετικά, σε σχέση με όλα τα θετικά σημεία δεδομένων.

$$\text{TruePositiveRate} = \frac{TruePositive}{FalseNegative + TruePositive}$$

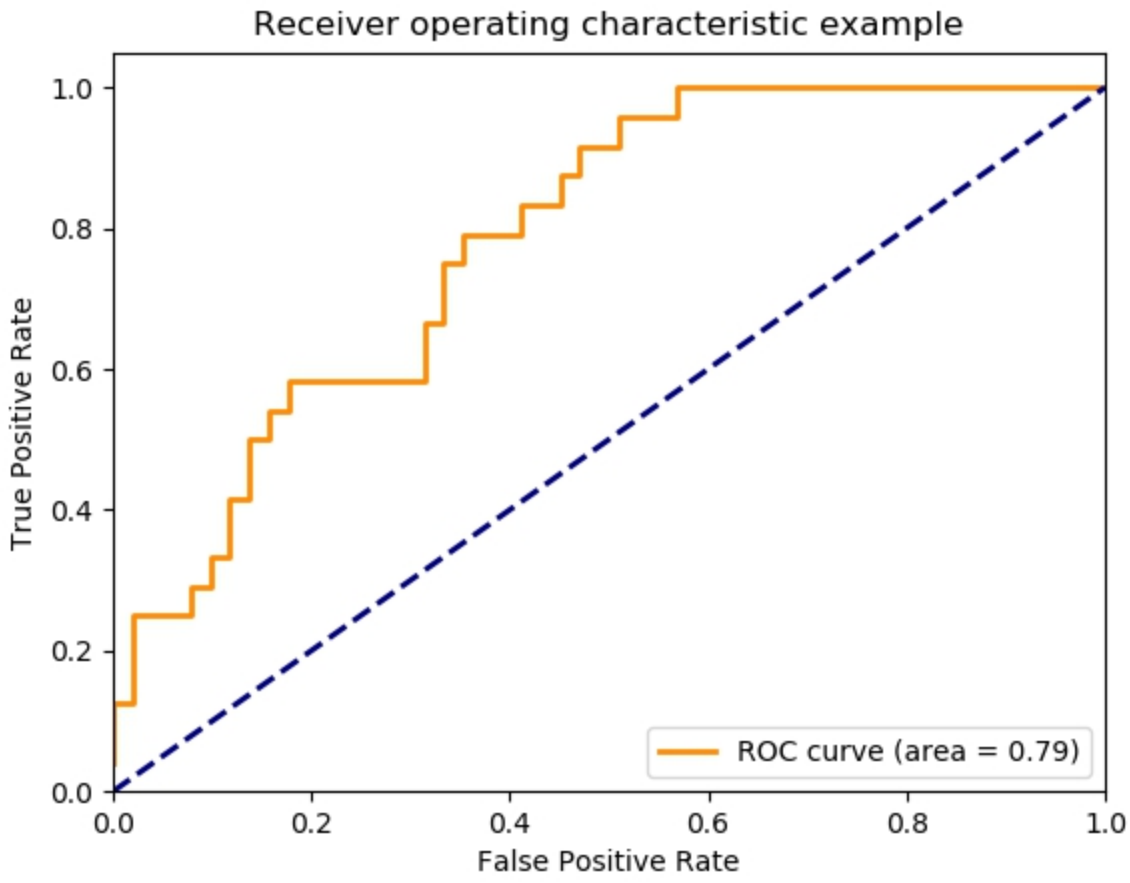
- **True Negative Rate (Ειδικότητα)** : Το True Negative Rate ορίζεται ως *TN / (FP + TN)* . Το False Positive Rate αντιστοιχεί στην αναλογία των αρνητικών σημείων δεδομένων που θεωρούνται σωστά ως αρνητικά, σε σχέση με όλα τα αρνητικά σημεία δεδομένων.

$$\text{TrueNegativeRate} = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

- **False Positive Rate** : False Positive Rate ορίζεται ως *FP / (FP + TN)* . Το False Positive Rate αντιστοιχεί στο ποσοστό των αρνητικών σημείων δεδομένων που λανθασμένα θεωρούνται θετικά, σε σχέση με όλα τα αρνητικά σημεία δεδομένων.

$$\text{FalsePositiveRate} = \frac{FaslePositive}{TrueNegative + FalsePositive}$$

Το False Positive Rate και το *True Positive Rate* έχουν και οι δύο τιμές στο εύρος **[0, 1]** . *Τα FPR* και *TPR* και τα δύο υπολογίζονται σε ποικίλες τιμές κατωφλίου όπως (0,00, 0,02, 0,04,…., 1,00) και σχεδιάζεται ένα γράφημα. *Το AUC* είναι η περιοχή κάτω από την καμπύλη του γραφικού *False Positive Rate έναντι True Positive Rate* σε διαφορετικά σημεία στο **[0, 1]** .



Σχήμα 22: ROC καμπύλη

Όπως είναι προφανές, η *AUC* έχει εύρος [0, 1]. Όσο μεγαλύτερη είναι η τιμή, τόσο καλύτερη είναι η απόδοση του μοντέλου μας.

Βαθμολογία F1

Το F1 Score χρησιμοποιείται για τη μέτρηση της ακρίβειας ενός τεστ

Η βαθμολογία F1 είναι η αρμονική μέση μεταξύ ακρίβειας και ανάκλησης. Το εύρος για το σκορ F1 είναι [0, 1]. Σας λέει πόσο ακριβής είναι ο ταξινομητής σας (πόσες περιπτώσεις ταξινομεί σωστά), καθώς και πόσο ισχυρός είναι (δεν χάνει σημαντικό αριθμό παρουσιών).

Υψηλή ακρίβεια αλλά χαμηλότερη ανάκληση, χάνει μεγάλο όγκο περιπτώσεων που δεν είναι εύκολο να ταξινομηθούν. Όσο πιο μεγάλη είναι η βαθμολογία F1, τόσο βελτιστοποιείτε η απόδοση του μοντέλου που έχουμε επιλέξει. Ο μαθηματικός τύπος, εκφράζεται ως εξής:

$$F1 = 2^* \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

Βαθμολογία F1

Η βαθμολογία F1 προσπαθεί να βρει την ισορροπία μεταξύ ακρίβειας και ανάκλησης.

- **Ακρίβεια:** Είναι ο αριθμός των σωστών θετικών αποτελεσμάτων δια του αριθμού των θετικών αποτελεσμάτων που προβλέπονται από τον ταξινομητή.

$$Precisions = \frac{TruePositives}{TrueNegatives + FalsePositives}$$

- **Ανάκληση:** Είναι ο αριθμός των σωστών θετικών αποτελεσμάτων διαιρούμενος με τον αριθμό ***όλων των*** σχετικών δειγμάτων (όλα τα δείγματα που θα έπρεπε να είχαν αναγνωριστεί ως θετικά).

$$Precisions = \frac{TruePositives}{TrueNegatives + FalseNegatives}$$

Μέσο απόλυτο σφάλμα

Το μέσο απόλυτο σφάλμα είναι ο μέσος όρος της διαφοράς μεταξύ των αρχικών τιμών και των προβλεπόμενων τιμών. Μας δίνει το μέτρο για το πόσο μακριά ήταν οι προβλέψεις από την πραγματική παραγωγή. Ωστόσο, δεν μας δίνουν καμία ιδέα για την κατεύθυνση του σφάλματος,δηλαδή αν είμαστε κάτω ή πάνω από την πρόβλεψη των δεδομένων. Μαθηματικά, αναπαρίσταται ως:

$$\text{MeanAbsoluteError} = \frac{1}{N}\sum_{j=1}^{N} \left|y_j - \hat{y}_j\right|$$

Σφάλμα μέσου τετραγώνου

Το μέσο τετράγωνο σφάλμα (MSE) είναι αρκετά ίδιο με το μέσο απόλυτο σφάλμα, με τη μόνη και σημαντική διαφορά πώς το MSE παίρνει το μέσο όρο του **τετραγώνου** της διαφοράς μεταξύ των αρχικών τιμών και των προβλεπόμενων τιμών. Το πλεονέκτημα του MSE είναι ότι είναι πιο εύκολο να υπολογιστεί η κλίση, ενώ το μέσο απόλυτο σφάλμα απαιτεί περίπλοκα γραμμικά εργαλεία προγραμματισμού για τον υπολογισμό της διαβάθμισης. Καθώς παίρνουμε το τετράγωνο του σφάλματος, η επίδραση των μεγαλύτερων σφαλμάτων γίνεται πιο έντονη και μικρότερη, επομένως το μοντέλο μπορεί τώρα να εστιάσει περισσότερο στα μεγαλύτερα σφάλματα.

$$\text{MeanSquaredError} = \frac{1}{N}\sum_{j=1}^{N} \left(y_j - \hat{y}_j\right)^2$$

4 Συσταδοποίηση

Στη συσταδοποίηση το πρόβλημα που αντιμετωπίζουμε είναι πως μας δίνεται ένα σύνολο δεδομένων χώρις τις ετικέτες και τίς κλάσεις που χρεαζόμαστε. Έτσι αναζητούμε έναν αλγόριθμο [171] ώστε να ομαδοποιήσει αυτό τον όγκο δεδομένων σε συστάδες οι οποίες να κάνουν με σωστό τρόπο τον διαχωρισμό των δεδομένων. Στην πράξη λοιπόν, οι συστάδες [172] αποτελούνται απο αντικείμενα τα οποία το κάθε ένα πρέπει να είναι πιο κοντα κάθε άλλο αντικείμενο της ίδιας συστάδας [172] απ' οτι σε οποιοδήποτε άλλο αντικείμενο που άνηκει σε διαφορετική συστάδα.

Δοθέντων:

- μιας ΒΔ $D=\{t_1, t_2, ..., t_n\}$ από εγγραφές,
- ενός μέτρου ομοιότητας $sim(t_i, t_j)$ μεταξύ δύο εγγραφών της ΒΔ και
- μιας ακέραιας τιμής k,

το **Πρόβλημα της Συσταδοποίησης** είναι η εύρεση μίας αντιστοίχισης $f : D \rightarrow \{1, ..., k\}$ όπου κάθε εγγραφή t_i της ΒΔ αντιστοιχίζεται σε μία συστάδα K_j, $1 \leq j \leq k$, έτσι ώστε:

- για κάθε εγγραφή η **ομοιότητα** μεταξύ αυτής και οποιασδήποτε εγγραφής από την ίδια συστάδα να είναι μεγαλύτερη από την ομοιότητα μεταξύ αυτής και οποιασδήποτε εγγραφής από άλλες συστάδες.

- Μία **Συστάδα**, K_j, περιέχει ακριβώς εκείνες τις πλειάδες που αντιστοιχίζονται σε αυτήν.

Μέρη συσταδοποίησης:

Αντιπροσώπευση δεδομένων με χαρακτηριστικά

- Μέσω χαρακτηριστικών ή επιλογής μιας υποομάδας δεδομένων.
- Είδη χαρακτηριστικών:
    - Ποσοτικά (quantitative), π.χ. αριθμητικές τιμές, διάρκεια
    - Ποιοτικά (qualitative), π.χ. χρώμα, ένταση ήχου
- Σημαντικό κομμάτι της διαδικασίας [173], κατά κύριο λόγο για χρονοσειρές.

- Η σωστή διαλογή χαρακτηριστικών [173] κάνει την ομαδοποίηση απλή και κατανοητή.
- Η λανθασμένη διαλογή χαρακτηριστικών κάνει την ομαδοποίηση πολύπλοκη [174] δίχως να αντιπροσωπεύει απολύτος κατανοητά τις φυσικές ομάδες των δεδομένων.

Υπολογισμός απόστασης χαρακτηριστικών

Υπολογισμός της ομοιότητας ή ανομοιότητας μεταξύ ζευγών δεδομένων μέσω μιας αριθμητικής απόστασης [175].

(1) Απόσταση Minkowski:

$$D(x_i, x_j)= (\sum_{k=1}^{d} | x_{i,k} - x_{j,k}|^p)^{1/p}$$

Όπου xi και xj: ανύσματα χαρακτηριστικών ή δεδομένων διαστάσεων d.

(2) Ευκλείδεια (Euclidean) : η πιο δημοφιλής, ειδική περίπτωση της απόστασης Minkowski

$$D(x_i, x_j)=\sqrt{\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k}\right)^2}$$

(3) Απόσταση Manhattan :Minkowski για p= 1

$$D(x_i, x_j)= \sum_{k=1}^{d} | x_{i,k} - x_{j,k}|$$

Τα ελαττώματα τών αποστάσεων Minkowski [176]: (ι) το χαρακτηριστικό το οποίο είναι μεγαλύτερο (σε πλάτος) τείνει να είναι ισχυρότερο των υπολοίπων (ιι) τα χαρακτηριστικά επηρεάζονται απο τις τιμές πλάτους, έτσι πρέπει αρχικά να κανονικοποιούνται - Ευκλίδεια απόσταση: είναι πιο κατάλληλη όταν τα δεδομένα σχηματίζουν απομονωμένες ομάδες

Απόσταση Mahalanobis:

$$D_M(x_i, x_j)=\sqrt{\left(x_i - x_j\right)\Sigma^{-1}\left(x_i - x_j\right)^T}$$

όπου $\sum -1$:πίνακας συνδιασποράς

Όταν Σ=I τότε έχουμε Ευκλίδεια απόσταση [176]. - Διαφέρει από την Ευκλίδεια απόσταση: (ι) λαμβάνει υπόψη τη συσχέτιση μεταξύ των δεδομένων, (ιι) δεν επηρεάζεται από το πλάτος των δεδομένων.

Ομαδοποίηση (ιεραρχική ή επιμεριστική)

Το clustering ή αλλιώς η ανάλυση συστάδων είναι η οργάνωση μιας συλλογής από δείγματα-στοιχεία σε συστάδες [177] με βάση κάποιο μέτρο ομοιότητας. Τα σημεία αναπαριστόνται σε έναν πολυδιάστατο χώρο [178] και συνήθως αναφέρονται ως διανύσματα τμών κάποιων μέτρων. Τα στοιχεία μίας ομάδας αναμεταξύ τους έχουν πολύ μεγαλύτερη ομοιότητα απ’ ότι με στοιχεία που ανήκου σε άλλες ομάδες.

Η διαδικασία της συσταδοποίησης κατατάσσεται στην μη επιβλέπουσα μάθηση [179] (unsupervised learning).Αντίστοιχα η κατηγοριοποίηση (supervised classification) ή επιβλεπόμενη μάθηση διαθέτει στοιχεία τα οποία είναι εξ’ αρχής ομαδοποιημένα και στην περίπτωση ενός νέου στοιχείου απλώς το εντάσουμε σε μία ήδη έτοιμη κλάση [179].Τα στοιχεία όπου έιναι εξ’ αρχής ομαδοποιημένα περιγράφουν και αποτελούν τις ομάδες – κλάσεις στις οποίες θα εντάξουμε νέα στοιχεία.

Αντίθετα στην μη επιβλεπόμενη μάθηση και συσταδοποίηση δίχως να έχουμε κάποια γνώση για τυχόν ομάδες που υπάρχουν εξ’ αρχής καλούμαστε να κάνουμε ομαδοποίηση των νέων στοιχείων σε λογικές κλάσσεις [179] μια διαδικασία σάφως πιο περίπλοκη. Συμπεραίνουμε λοιπόν πώς η συσταδοποίηση παράγεται απο τα δεδομένα και είναι απόλυτα οδηγούμενη από αυτά (data driven).

Επείσης οι τεχνικές συσταδοποίησης εφαρμόζονται [179] όταν τα δεδομένα μάς πρέπει να διαμοιραστούν σε φυσικές ομάδες [180] και πιθανώς αντικατοπτρίζουν έναν μηχανισμό που τα ομαδοποιεί βάσει κάποιων κοινών χαρακτηριστικών που παρουσιάζουν. Η συσταδοποίηση απαιτεί διαφορετικές τεχνικές από τις μεθόδους κατηγοριοποίησης και αυτοσχέτισης [181].

ΕΓΚΥΡΌΤΗΤΑ ΟΜΆΔΩΝ

Αναγκαία η εκτίμηση των ομάδων γιατί:

- Μέθοδοι συσταδοποίησης πάντοντε καταλήγουν σε μια ομαδοποίηση, ακόμα κι αν τα δεδομένα είναι θόρυβος και δεν υπάρχει καμιά φυσική ομαδοποίηση
- Διαφορετικές μεθόδοι πολύ πιθανόν να δώσουν διαφορετικές ομαδοποιήσεις

4.1 Ο αλγόριθμος Κ μέσων

Έστω ότι δίνεται μια συσταδοποίηση C = {C1, C2, … , CK}. Χρειαζόμαστε κάποια συνάρτηση βαθμολόγησης η οποία θα αξιολογεί την ποιότητα ή αλλιώς την καταλληλότητα της συσταδοποίησης [182]. Αυτή η συνάρτηση βαθμολόγησης βασίζεται στο άθροισμα των τετραγώνων των σφαλμάτων (sum of squared errors SSE) και ορίζεται ως:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}i)^2$$

test set predicted value actual value

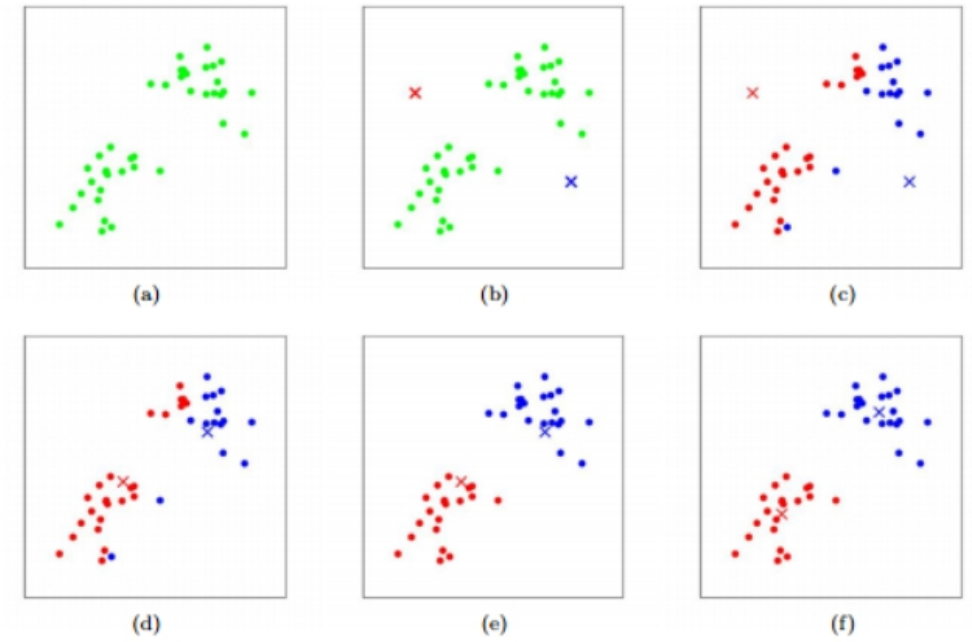
Πρωταρχικός στόχος είναι να βρεθεί εκείνη [183] η συσταδοποίηση που ελαχιστοποιεί τη βαθμολογία ή αλλιώς το σκορ του SSE.

Ο αλγόριθμος Κ μέσων χρησιμοποιεί μια άπληστη, ή αλλιώς πλεονεκτική (greedy), επαναληπτική τεχνική για να βρει μια συσταδοποίηση που ελαχιστοποιεί την αντικειμενική συνάρτηση SSE [183]. Κατά συνέπεια, μπορεί να συγκλίνει σε τοπικά βέλτιστα και όχι σε μια καθολικά βέλτιστη συσταδοποίηση.

Ο συγκεκριμένος αλγόριθμος καθορίζει τις αρχικές τιμές των μέσων για τις συστάδες παράγοντας [184] με τυχαίο τρόπο k σημεία στον χώρο δεδομένων. Για να επιτευχθεί αυτό συνήθως παράγεται μία τιμή με όμορφα τυχαίο τρόπο εντός του αντίστοιχου εύρους τιμών για κάθε διάσταση του συγκεκριμένου συνόλου δεδομένων που έχουμε. Κάθε επανάληψη του αλγόριθμου μέσον [184] αποτελείται από δύο βήματα όπου το πρώτο είναι η αντιστοίχιση σε συστάδες και το δεύτερο βήμα είναι η ενημέρωση του κέντρου βάρους.

Όσον αφορά τώρα την πλευρά της υπολογιστικής πολυπλοκότητας του συγκεκριμένου αριθμού δηλαδή των Κ μέσων, μπορούμε να διαπιστώσουμε το εξής ότι το Βήμα της αντιστοίχισης σε συστάδες απαιτεί έναν χρόνο O(knd) [185], Είδη για καθένα από τα n σημεία, πρέπει να υπολογίσουμε την απόστασή του από κάθε μία από n συστάδες κάτι που απαιτεί d πράξεις στις d διαστάσεις

Το K-Means βρίσκει τα καλύτερα κεντροειδή εναλλάσσοντας μεταξύ (1) εκχώρηση σημείων δεδομένων σε συστάδες με βάση τα τρέχοντα κεντροειδή (2) κεντρικά-κεντρικά [186] (σημεία που είναι το κέντρο ενός συμπλέγματος) με βάση την τρέχουσα εκχώρηση σημείων δεδομένων σε συστάδες.



Σχήμα 23: Τεχνική Κ μέσων

4.2 Ιεραρχική συσταδοποίηση

Στην συσσωρευτική ιεραρχική συσταδοποίηση ξεκινάμε με μία ξεχωριστή συστάδα για καθένα από τα n σημεία. Κατόπιν συγχωνεύουμε [187] επανειλημμένα τις δύο πλησιέστερες συστάδες που υπάρχουν και στη συνέχεια η διαδικασία σταματάει όταν όλα τα σημεία θα ανήκουν πλέον στην ίδια συστάδα [187]. Άρα θα κρέμονται από την ίδια ρίζα του συγκεκριμένου δέντρου που θα πραγματοποιηθεί για το δοθέν σύνολο δεδομένων[187]. Επομένως θέλουμε εμείς στο τελικό σύνολο να περιέχει μόνο μία συστάδα και επίσης έχουμε ότι το πλήθος των συστάδων μειώνεται κατά ένα σε κάθε βήμα εκτέλεσης του συγκεκριμένου αλγόριθμου της ιεραρχικής συστηματοποίησης [188], γιατί η διαδικασία αυτή τελικός παράγει μία ακολουθία n ένθετων συσταδοποιήσεων. Από την άλλη Μπορούμε να σταματήσουμε τη διαδικασία συγχώνευσης όταν θα υπάρχουν ακριβώς k εναπομείναντες συστάδες αν αυτό καθορίζεται ρητά από την περιγραφή.

Input:

D= {t_1,t_2, …., t_n} // Set of elements

A // Adjacency matrix showing distance between elements.

Output:

BE // Dendrogram represented as a set of ordered triples.

Το κύριο βήμα του αλγόριθμου είναι η εξακρίβωση του ζεύγους των συστάδων γειτόνων για τον υπολογισμό της απόστασης [189] μεταξύ δύο οποιονδήποτε συστάδων μπορούν να χρησιμοποιηθούν αρκετά μέτρα όπως είναι η απόσταση του μοναδικού συνδέσμου ή του πλήρους συνδέσμου είναι ο μέσος όρος των αποστάσεων της κάθε ομάδας [190].

Στην συσσωρευτική συσταδοποίηση πρέπει να υπολογίσουμε την απόσταση κάθε συστάδας από τις υπόλοιπες και αυτό γίνεται σε κάθε βήμα. Επομένως το πλήθος των συστάδων σε κάθε εκτέλεση μειώνεται κατά 1 [190]. Αρχικά απαιτείται χρόνος O(n^2) για την δημιουργία της μήτρας των αποστάσεων ανά ζεύγη εκτός και αν [190] η μήτρα παρέχεται ως είσοδος στον αλγόριθμο εκτέλεσης της σωρευτικής συστηματοποίησης.

Σε κάθε βήμα πρέπει να υπολογιστούν ξανά οι αποστάσεις της συγχωνευμένης συστάδας από τις υπόλοιπες ομάδες συστάδων,ενώ οι αποστάσεις μεταξύ των άλλων συστάδων παραμένουν ίδιες. Αυτό σημαίνει ότι στο βήμα t πρέπει να υπολογίσουμε O(n-t) αποστάσεις [191]. Η άλλη κυρία πράξη είναι η εύρεση του ζεύγους των πλησιέστερων συστάδων στη μήτρα αποστάσεων για αυτό αποθηκεύουμε τις n^2 αποστάσεις [192] σε μία δομή δεδομένων που ονομάζεται σωρός κάτι που μας επιτρέπει να βρούμε την ελάχιστη απόσταση σε χρόνο O(1), λόγω της ιδιότητας που υπάρχει στη συγκεκριμένη δομή δεδομένων [193]. Η δημιουργία του σωρού απαιτεί χρόνο O(n^2). Η ενημέρωση ή αλλιώς η διαγραφή των αποστάσεων [194] στη συγκεκριμένη δομή απαιτεί χρόνο O(logn) για κάθε τέτοια πράξη. Οπότε ο συνολικός χρόνος για όλα τα βήματα συγχώνευσης με βάση τη συγκεκριμένη δομή δεδομένων συσσωρευτική

συσταδοποίηση καταλήγουν την υπολογιστική πολυπλοκότητα της ιεραρχικής συστηματοποίησης να είναι O(n^2logn).

4.3 Αλγόριθμος DBSCAN

Στη συσταδοποίηση που βασίζεται στην πυκνότητα, δεν χρησιμοποιείται μόνο η απόσταση των σημείων στον προσδιορισμό των συστάδων, αλλά αξιοποιείται και η πυκνότητα των σημείων [195]. Ο συγκεκριμένος αλγόριθμος συσταδοποίησης χρησιμοποιεί την έννοια της πυκνότητας για την αναγνώριση συστάδων. Συγκεκριμένα [196], μπορούμε να ξεχωρίσουμε τρία βασικά βήματα, τα οποία ακολουθεί ο αλγόριθμος για να επιτύχει την επιθυμητή συσταδοποίηση. Τα βήματα αυτά θα φανούν καλύτερα όταν εξηγηθεί ο κώδικας.

Ο DBSCAN αλγόριθμος ομαδοποιεί τα στοιχεία βάση πυκνότητας [197], είναι αρκετά αποτελεσματικός,απλός και είναι ένας αντιπροσωπευτικός αλγόριθμος για την κατηγορία του. Παρακάτω θα αναλύσουμε τον συγκεκριμένο αλγόριθμο καθώς και την έννοια της πυκνότητας.

Algorithm 3. Reduce step for Distributed DBSCAN

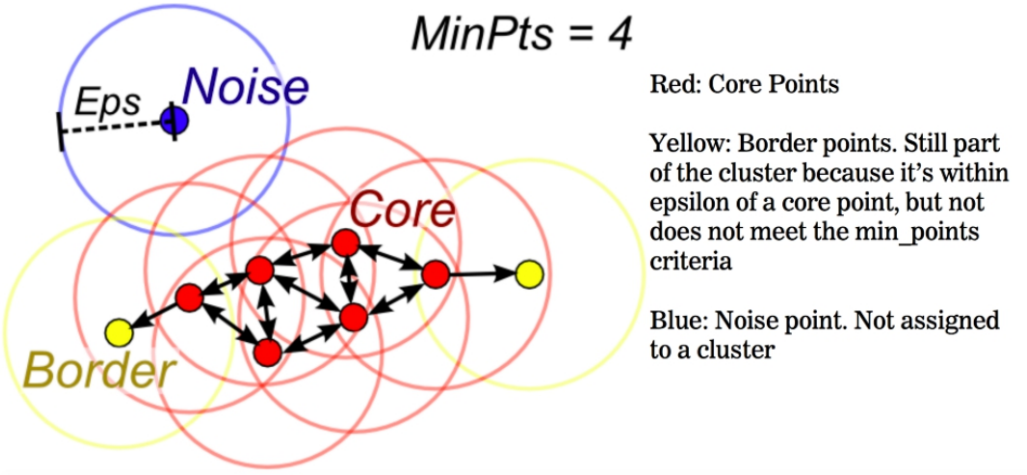
```
Data:
Clusters collected from nodes - CC,
distance - ε,
minimum number of points to create dense region - minPts
begin
    for each cluster C in CC do
        for each point P in C do
            if P is visited then
                Continue to next P
            end
            else
                mark P as visited
                nbrPts ⟵ points in ε-neighborhood of P that are not in C
                if sizeof(nbrPts) ≥ minPts then
                    Merge C with every cluster, to which points nbrPts belongs
                end
            end
        end
    end
end
```

Σχήμα 24: Αλγόριθμος DBSCAN

Υπάρχουν διαφόρων ειδών προσεγγίσεις με τον αλγόριθμο DBSCAN. Η πρώτη προσέγγιση θα γίνει με βάση το κέντρο με το οποίο ορίζεται ο DBSCAN αλγόριθμος [198]. Βάση αυτής λοιπόν, η πυκνότητα ορίζεται για ένα συγκεκριμένο στοιχείο του συνόλου δεδομένων, ως το πλήθος των δεδομένων που βρίσκονται μέσα σε μια συγκεκριμένη ακτίνα (έστω Eps) από το υπό εξέταση σημείο (συμπεριλαμβανομένου και του ίδιου του σημείου). Αν υποθέσουμε λοιπόν πως η η ακτίνα (ρ) έιναι ακόμα μεγαλύτερη τότε όλα τα στοιχεία του δεδομένου συνόλου θα έχουν πυκνότητα m (όπου είναι το πλήθος των στοιχείων του συνόλου) [199].Αντίστοιχα αν η ακτίνα είναι πολύ μικρή η πυκνότητα των σημείων θα είναι ίση με 1.

Μπορούμε να θεωρήσουμε ότι η μέθοδος DBSCAN είναι μία αναζήτηση για τις συνεκτικές συνιστώσες ενός γραφήματος του οποίου ισχύουν τα εξής ότι οι κορυφές αντιστοιχούν στα σημεία [198] ή αλλιώς πυρήνες του συνόλου δεδομένων και αφετέρου ότι υπάρχει μία μικρή κατευθυνόμενη ακμή ή αλλιώς πορεία που ενώνει τις δύο κορυφαίες [197] ή αλλιώς τα σημεία

πυρήνες του συνόλου δεδομένων αν η μεταξύ τους απόσταση είναι μικρότερη από ε, έτσι πιο συγκεκριμένα καθένα από τα δύο σημεία ανήκει στην ε-γειτονιά του άλλου σημείου. οι συνεκτικές συνιστώσες του συγκεκριμένου γραφήματος αντιστοιχούν στα σημεία πυρήνες [198] ή απλώς σημεία της κάθε συστάδας ακόμη κάθε σημείο πυρήνας του συνόλου δεδομένων ενσωματώνει συστάδα ή αλλιώς την ομάδα στην οποία ανήκει η τυχόν οριακά σημεία που ανήκουν στην περιφέρεια της γειτονιάς του.



Σχήμα 25: Μέθοδος DBSCAN

Μία σημαντική σημείωση για τον αλγόριθμο dbscan είναι ότι υπάρχει μία ευαισθησία στην επιλογή του σημείου ε [199], ειδικά στην περίπτωση που οι συστάδες έχουν διαφορετικές τιμές πυκνοτήτων. Επομένως εάν έχουμε ότι το ε είναι πολύ μικρό τότε οι πιο ωραίες συστάδες ,ομάδες αντικειμένων εγγράφων του συνόλου δεδομένων θα κατηγοριοποιηθούν ως θόρυβος σε άλλη περίπτωση που έχουμε ότι το ε [200] να είναι αρκετά μεγάλο τότε έχουμε πιο πυκνές συστάδες, ομάδες αντικειμένων και ενδέχεται να συγχωνευθούν και να γίνουνε μικρότερες σε πλήθος.

Συνοψίζοντας την περιγραφή για τον αλγόριθμο dbscan που αφορά τον υπολογισμό της ε-γειτονιάς για κάθε σημείο [200], έχουμε το εξής ότι αν στατικότητα του συνόλου δεδομένων δεν είναι πολύ μεγάλη τότε αυτό μπορεί να επιτευχθεί αποδοτικά και αποτελεσματικά με χρήση μίας χωρικής δομής αριθμοδεικτών σε χρόνο O(nlogn). Όταν η στατικότητα είναι αρκετά μεγάλη τότε ο χρόνος που απαιτείται είναι ίσος με O(n^2) για τον υπολογισμό της γειτονιάς

κάθε σημείο. Έτσι λοιπόν γίνεται αντιληπτό ότι η χειρότερη περίπτωση του αλγόριθμου dbscan έχει ως συνολική πολυπλοκότητα ίση με O(n^2).

4.4 Εγκυρότητα συσταδοποίησης

Υπάρχουν πολλές διαφορετικές μέθοδοι ανάλογα με το ζητούμενο τύπο των συστάδων ή ομάδων που θέλουμε να πραγματοποιήσουμε έλεγχο εγκυρότητας [201]. Η επαλήθευση της εγκυρότητας και αξιολόγηση των συσταδοποίησεων περιλαμβάνει τρεις βασικές εργασίες που πραγματοποιούνται και πρέπει να ακολουθηθούν. Η αποτίμηση της σταθεροποίησης για την αξιολόγηση της καταλληλότητας ή της εφαρμογής που μπορεί να επιτευχθεί [202], ή της ποιότητας της τακτοποίησης, στη συνέχεια για την σταθερότητα της σταθεροποίησης με στόχο τη σωστή ερμηνεία και την ευαισθησία που πρέπει να έχει το αποτέλεσμα της [203] και τρίτη παρατήρηση η οποία είναι και τελική είναι η τάση της ομαδοποίησης για την αξιολόγηση του κατά πόσο επιτεύχθηκε [204] η σωστή εφαρμογή στις ειδοποιήσεις και αν τα δεδομένα εμφανίζουν οποιαδήποτε εν γένει η προσαρμόσιμη δομή ομαδοποίησης.

Υπάρχει πληθώρα μέτρων και στατιστικών εγκυρότητας που έχουν υλοποιηθεί και προταθεί σε άλλες μελέτες και σχετικές εργασίες [205] που έχουν πραγματοποιηθεί και οι τύποι που μπορούμε να διακρίνουμε και να χωρίσουμε είναι οι εξής παρακάτω τρεις. Αρχικά είναι τα εξωτερικά μέτρα εγκυρότητας τα οποία χρησιμοποιούνται ως κριτήρια που δεν είναι εγγενή για το σύνολο δεδομένων δηλαδή μπορεί να έχουν τη μορφή πληροφοριών για τις συστάδες οι οποίες είναι γνωστές εκ των προτέρων η καθορίζονται από κάποιους ειδικούς του συγκεκριμένου χώρου από τον οποίον χτίζονται τα συγκεκριμένα σύνολα δεδομένων [206]. Στη συνέχεια μπορεί να έχουμε τα εσωτερικά μέτρα εγκυρότητας που στηρίζονται κυρίως σε

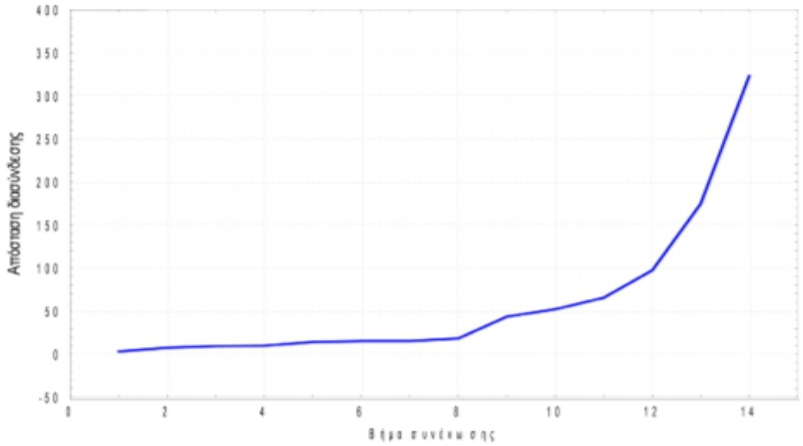
κριτήρια που προκύπτουν από τα ίδια δεδομένα ενός data set μπορεί δηλαδή να βασιστούμε σε κάποιες μετρικές συναρτήσεις αποστάσεις [207] εντός της ίδιας σύστασης ομάδας ή αλλιώς και μεταξύ διαφορετικών συστάδων για να ορίσουμε μέτρα που δείχνουν κατά πόσο συμπαγείς crs είναι οι συστάδες μεταξύ τους η και κατά πόσο διαχωρισμένες είναι δηλαδή να κοιτάξουμε τα σημεία απόστασης και διαφορών μεταξύ των συστάδων που σχηματίζονται εντός ενός συνόλου δεδομένων. Όσον αφορά [208] την 3η και τελευταία κατηγορία που είναι τα σχετικά μέτρα εγκυρότητας συγκρίνονται ευθέως με διαφορετικές μεθόδους συστηματοποίησης συνήθως αφορούν εκείνες που προκύπτουν από διαφορετικές ρυθμίσεις των παραμέτρων που δέχεται ο ίδιος ο αλγόριθμος.

α) Εσωτερική αξιολόγηση

Εφαρμόζεται αποκλειστικά στα στοιχεία της ταξινόμησης που μόλις σχηματίστηκαν, γι αυτο ονομάζεται εσωτερική. Ο αλγόριθμος που παράγει συστάδες με υψηλή ομοιότητα ανάμεσα στα στοιχεία της [206] και ταυτόχρονα με χαμηλή ομοιότητα μεταξύ των στοιχείων διαφορετικών συστάδων είναι αυτός ο οποίος θα κερδίσει και την μεγαλύτερη βαθμολογία. Το σημείο όπου πάντα μειονεκτεί αυτή η αξιλόγηση είναι ότι ενίοτε αποδίδονται υψηλοί βαθμοί επίδοσης δίχως όμως να παρέχεται σοβαρή πληροφόρηση από την ταξινόμηση [207] ακόμα ίσως ενδέχεται να αξιολογεί εσφαλμένα αλγόριθμους που μετρούν το ίδιο μοντέλο [208]. Πρακτικά η αξιολόγιση αυτή προτείνεται στίς περιπτώσεις που γίνεται σύγκριση μεταξύ αλγορίθμων και καποιός εξ’ αυτών αποδίδει καλύτερα από κάποιον άλλο [206], χωρίς όμως απαραίτητα να παράγει και πιο αξιόπιστα συγκριτικά αποτελέσματα. Προτείνονται λοιπόν οι παρακάτω δύο τεχνικές για την εκτίμηση της ποιότητας των αλγόριθμων ταξινόμησης:

• Ο δείκτης Davies-Bouldin (DB),

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$



Σχήμα 26: Δείκτης Davies-Bouldin

όπου n είναι ο αριθμός των συστάδων cx το κέντρο της συστάδας x, σx η μέση απόσταση των στοιχείων στη συστάδα x και d(ci , cj) η απόσταση μεταξύ των κέντρων ci και cj . Όταν λαμβάνουμε χαμηλές τιμές σημαίνει πως ο αλγόριθμος παράγει ένα σύνολο συστάδων με μεγάλη ομοιότητα ή αντίστοιχα όταν λαμβάνουμε υψηλές τιμές υπάρχει (χαμηλή ομοιότητα) μεταξύ των συστάδων.

β) Εξωτερική αξιολόγηση

Αξιολογούνται οι ταξινομήσεις με την συγκρότηση στοιχείων που δεν συμμετέχουν στην ταξινόμηση αλλά επιλέγονται από επιστήμονες ολκής [205] ώστε να εξυπηρετήσουν υπό μορφή ειδικών κλάσεων ορισμένα μέτρα αναφοράς της εγκυρότητας κάθε ταξινόμησης (benchmarks).

Ο δείκτης J του Jaccard [208], αντιπαραθέτει δύο ομάδες στοιχείων και λαμβανει ένα συγκεκριμένο εύρος τιμών 0-1. Οι τιμές που τείνουν στο 1 δηλώνουν ότι οι δύο συστάδες ταυτίζονται ενώ αντίστοιχα ώσες τιμές τείνουν πρός το 0 ότι δεν έχουν κοινά σημεία.

$$J = \frac{TP}{TP+FP+FN}$$

Επίλογος

Η μεθοδολογία ταξινόμησης και ομαδοποίησης είναι διαφορετική και το αποτέλεσμα που αναμένεται από τους αλγόριθμους τους διαφέρει επίσης. Με λίγα λόγια, τόσο η ταξινόμηση όσο και η ομαδοποίηση χρησιμοποιούνται για την αντιμετώπιση διαφορετικών προβλημάτων.

Η ταξινόμηση και η ομαδοποίηση είναι δύο αποτελεσματικές τεχνικές μηχανικής μάθησης που μπορείτε να χρησιμοποιήσετε για να βελτιώσετε τις επιχειρηματικές σας διαδικασίες. Παρόλο που αυτές οι διαδικασίες είναι παρόμοιες, μπορείτε να τις χρησιμοποιήσετε διαφορετικά για να κατανοήσετε τους αγοραστές σας και να βελτιώσετε την εμπειρία αγορών πελατών στο κατάστημά σας. Αναλύοντας, προσθέτοντας προφίλ και στοχεύοντας τους καταναλωτές σας χρησιμοποιώντας μηχανική εκμάθηση, θα δημιουργήσετε τελικά μια πιστή πελατειακή βάση και μια βελτιστοποιημένη απόδοση επένδυσης.

Συνεπώς, η **Ταξινόμηση** είναι ο αριθμός των τάξεων που είναι γνωστός. Επίσης, απαιτούνται δεδομένα εκπαίδευσης (συλλογή επισημασμένων παρουσιών). **Ακόμη,** με βάση τα δεδομένα εκπαίδευσης, το μοντέλο ταξινόμησης χρησιμοποιείται για την ταξινόμηση μελλοντικών παρουσιών σε ήδη καθορισμένες τάξεις. Οι δημοφιλείς αλγόριθμοι ταξινόμησης περιλαμβάνουν το Naive Bayes Classifier, Decision Trees και Random Forests.

Από την άλλη, η στην **Ομαδοποίηση** ο αριθμός των τάξεων είναι άγνωστος. Δεν απαιτούνται δεδομένα εκπαίδευσης. Η ομαδοποίηση χρησιμοποιείται για να κατανοήσει τα υπάρχοντα δεδομένα. Οι δημοφιλείς αλγόριθμοι που χρησιμοποιούνται για την ομαδοποίηση περιλαμβάνουν το K-Means, το Clustering Mean-Shift και τη χωρική ομαδοποίηση εφαρμογών με θόρυβο.

Παραθέσαμε επίσης τις προηγούμενες υποθέσεις που ενσωματώνει κάθε κατηγορία αλγορίθμων μηχανικής μάθησης. Με αυτόν τον τρόπο, θα μπορούσαμε να διατυπώσουμε μια λίστα ελέγχου με την οποία μπορούμε να συγκρίνουμε το σύνολο δεδομένων μας.

Αυτό, με τη σειρά του, μας επιτρέπει να προσδιορίσουμε αν πρέπει να χρησιμοποιήσουμε ταξινόμηση ή ομαδοποίηση για μια δεδομένη εργασία, σύμφωνα με τα χαρακτηριστικά της.

5 Βιβλιογραφία

[1] Tan, P. N., Steinbach, M. & Kumar, V. (2006). Introduction to Data Mining. Boston, MA: Pearson/AddisonWesley.

[2] Dunham, M. H. (2003). Data Mining: Introductory and Advanced Topics. Pearson Education, Upper Saddle River, N. J. Prentice Hall.

[3] k-means clustering. Ανακτήθηκε στις 27 Νοεμβρίου 2015, από: http://www.rdatamining.com/examples/kmeans-clustering

[4] Hierarchical Cluster Analysis. Ανακτήθηκε στις 27 Νοεμβρίου 2015, από: http://www.r-tutor.com/gpu-computing/clustering/hierarchical-cluster-analysis

[5] Data Mining Algorithms In R/Clustering/K-Means. Ανακτήθηκε στις 27 Νοεμβρίου 2015, από: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means DBSCAN. Ανακτήθηκε στις 29 Νοεμβρίου 2015, από: https://en.wikipedia.org/wiki/DBSCAN

[6] http://infolab.cs.unipi.gr/pre-eclass/courses/dwdm/slides/5-clustering.pdf

[7] http://www.ece.ucy.ac.cy/courses/ece480/lectures/HMY480_9.pdf

[8]https://www.mymanatee.org/arcgis_js_api/library/3.21/arcgis/esri/dijit/analysis/help/el/SummarizeAttributes_bd.html

[9] Data Mining Algorithms In R/Clustering/K-Means. Ανακτήθηκε στις 27 Νοεμβρίου 2015, από: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means

[10] Wikipedia (2015). Association Rule Learning. Ανακτήθηκε στις 27 Νοεμβρίου 2015, από: https://en.wikipedia.org/wiki/Association_rule_learning

[12] Hadoop. http://hadoop.apache.org

[13] HDFS. http://hadoop.apache.org/docs/stable1/hdfsdesign.html

[14] MapReduce. http://en.wikipedia.org/wiki/MapReduce

[15] Data, data everywhere, (2010), http://www.economist.com/node/15557443

[16] Hive: a warehousing solution over a map-reduce framework, Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., ... & Murthy, R, (2009), Proceedings of the VLDB Endowment, 2(2), 1626-1629

[17] Mining the Social Web, 2nd Ed., Matthew A. Russell, (2014), ISBN: 978-1-449-36761-9

[18] Mining Data from Twitter, Abhishanga Upadhyay, Luis Mao & Malavika Goda Krishna, (2014)

[19] Twitter Documentation, https://dev.twitter.com/overview/documentation

[20] OAuth 2 in Action, Justin Richer & Antonio Sanso, (2015), Manning Publications

[21] Apache NiFi Overview, Apache NiFi Team, (2014), https://nifi.apache.org/docs/nifi-docs/html/overview.html

[22] What is Apache Hadoop? (2012), https://www.oreilly.com/ideas/what-isapache-hadoop

[23] Hadoop, http://www.sas.com/en_us/insights/big-data/hadoop.html

[24] Welcome to Apache™ Hadoop®! (2014), https://hadoop.apache.org/

[25] Hadoop: What it is, how it works, and what it can do, (2011), https://www.oreilly.com/ideas/what-is-hadoop

[26] Apache Hadoop, http://hortonworks.com/hadoop/

[27] Apache Hadoop, (2011), https://en.wikipedia.org/wiki/Apache_Hadoop

[28] Apache Hadoop Basics, Hortonworks Inc., (2013)

[29] Hadoop: The Definitive Guide, Tom White, (2015)

[30] Ελένη Γολέμη.,(2010).Κρυπτογραφία & Εξόρυξη Δεδομένων.Ανακτήθηκε στις 16 Ιουλίου από http://nemertes.lis.upatras.gr/jspui/bitstream/10889/4791/1/ergasia-golemie.pdf

[31] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and AlgorithmsFree registration required. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

[32] Fayyad, Usama (1996). «From Data Mining to Knowledge Discovery in Databases» (PDF). Ανακτήθηκε στις 16 Ιουλίου 2012. Unknown parameter |coauthors= ignored (|author= suggested) (βοήθεια)

[33] Simmi Bagga., Dr. G.N. Singh., (2012). Applications of Data Mining.Ανακτήθηκε στις 19 Απριλίου ,2012 από http://www.ijsett.com/images/P5.pdf Αρχειοθετήθηκε 2016-11-23 στο Wayback Machine.

[34] Γούλου Ζωή.,(2010). Εφαρμογή μεθόδων εξόρυξης δεδομένων στη διαχείριση πελατειακών σχέσεων. Ανακτήθηκε στις 18 Ιουλίου από http://dspace.lib.uom.gr/bitstream/2159/14808/6/GoulouZoiMsc2012.pdf

[35] B. S. Everitt: The Cambridge Dictionary of Statistics, Cambridge University Press, Cambridge (3rd edition, 2006). ISBN 0-521-69027-7

[36] Bishop: Pattern Recognition and Machine Learning, Springer, ISBN 0-387-31073-8 den Dekker A. J., Sijbers J., (2014) "Data distributions in magnetic resonance images: a review", Physica Medica

[37] Resource Description Framework (RDF) Model and Syntax Specification (1999). Διαθέσιμο στο δικτυακό τόπο: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[38] C. Rogers, L. Laura, Sams Teach Yourself Java 6 in 21 Days, Fifth Edition (2007).

[39] Apache Jena Documentation. Διαθέσιμο στο δικτυακό τόπο: https://jena.apache.org/documentation/index.html

[40] C.D. Manning, M.Surdeanu, J.Bauer, J.Finkel, S.J.Bethard, D.McClosky, The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60 (2014).

[41] D.T.Pham, S.S.Dimov, C.D.Nguyen, Selection of K in K-means Clustering (2004). Διαθέσιμο στο δικτυακό τόπο: http://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf

[42] MATLAB Documentation (2015). Διαθέσιμο στο δικτυακό τόπο: http://www.mathworks.com/help/matlab/

[43] FFmpeg Documentation. Διαθέσιμο στο δικτυακό τόπο: http://ffmpeg.org/documentation.html

[44] The Neo4j Manual v2.2.3. Διαθέσιμο στο δικτυακό τόπο: http://neo4j.com/docs/2.2.3/

[45] Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition. Boston: Academic Press.

[46] Han, J., Kamber, M., & Pei, J. (2011). Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers.

[47] Heckerman, D. (1997). Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery, 1(1), 79-119. doi: 10.1023/A:1009730122752

[48] Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning, 20(3), 197-243. doi: 10.1007/bf00994016 Hwang, J., Lay, S., & Lippman, A. (1994). Nonparametric Multivariate Density Estimation: A Comparative Study. IEEE Transactions on Signal Processing, 42(10), 2795-2810. doi: 10.1109/78.324744

[49] Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 29(2), 119-127. doi: 10.2307/2986296 Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. Expert Systems with Applications, 32(4), 995-1003. doi: 10.1016/j.eswa.2006.02.016

[50] Koskivaara, E. (2004). Artificial Neural Networks in Analytical Review Procedures. Managerial Auditing Journal, 19(2), 191–223. doi: 10.1108/02686900410517821

[51] Loebbecke, J., Eining, M., & Willingham, J. (1989). Auditor's Experience with Material Irregularities: Frequency, Nature and Detectability. Auditing: A Journal of Practice and Theory, 9, 1-28.

[52] Loh, W. Y., & Shih, X. (1997). Split Selection Methods for Classification Trees. Statistica Sinika, 7, 815-840.

[53] Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. New York, NY: Springer + Business Media.

[54] Olaru, C., & Wehenkel, L. (2003). A Complete Fuzzy Decision Tree Technique. Fuzzy Sets and Systems, 138(2), 221-254. doi: 10.1016/S0165-0114(03)00089-7

[55] Parker, D. B. (1985). Learning-logic: Casting the Cortex of the Human Brain in Silicon. Technical Report TR-47. Boston, MA: Center for Computational Research in Economics and Management Science, MIT.

[56] Persons, O. (1995). Using Financial Statements Data to Identify Factors Associated with Fraudulent Financial Reporting. Journal of Applied Business Research, 11(3), 38-46.

[57] Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106. doi: 10.1007/bf00116251

[58] Quinlan, J. R. (1987). Simplifying Decision Trees. International Journal of Man-Machine Studies, 27(3), 221-234. doi: 10.1016/s0020-7373(87)80053-6

[59] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufman.

[60] Quinlan, J. R., & Rivest, R. L. (1989). Inferring Decision Trees Using the Minimum Description Length Principle. Information and Computation, 80(3), 227-248. doi: 10.1016/0890-5401(89)90010-2

[61] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-propagating Errors. Letters to Nature, 323(6088), 533-536. doi: 10.1038/323533a0

[62] Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Published by Addison Wesley.

[63] Programming Collective Intelligence, Toby Segaran, First Edition, Published by O' Reilly Media, Inc.

[64]http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html

[65]https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_decision_tree.htm

[66] https://www.saedsayad.com/decision_tree.htm

[67] https://www.geeksforgeeks.org/decision-tree/

[68] https://medium.com/swlh/decision-tree-classification-de64fc4d5aac

[69] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[70] https://www.geeksforgeeks.org/k-nearest-neighbours/

[71] https://repository.kallipos.gr/bitstream/11419/273/1/02_chapter_7.pdf

[72] Briem, G.J., Benediktsson, J.A., Sveinsson, J.R., 2002. “Multiple classifiers applied to multisource remote sensing data.”, IEEE Trans. Geosci. Remote Sens. 40, 2291–2299.

[73] Breiman L., 2001, Random Forests, Machine Learning, 45, 5-32.

[74] Campell B.J., 2002. Introduction to Remote Sensing. 3rd edition. Virginia

[75] Polytechnic Institute and State University. The Guilford Publications Press, New York.

[76] Colditz, R., 2015. “An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms.”, Remote Sens. 7, 9655.

[77] DeFries, R.S.; Foley, J.A.; Asner, G.P. Land-use choices: Balancing human needs and ecosystem function. Front. Ecol. Environ. 2004, 2, 249–257.

[78] Drusch, M. et al, 2012. ‘Sentinel-2: ESA’s optical high-resolution mission for {GMES} operational services.’ Remote Sens. Environ. 120, 25–36.

[79] Eitel, J.U. et al (2011), “Red-edge information from satellites improves early stress detection in a New Mexico conifer woodland.” Remote Sens. Environ., 115

[80] Friedl M.A. and Brodley C.E. (1997), “Decision Tree Classification of Land Cover from Remotely Sensed Data”, Remote Sensing of Environment, 61,399-409

[81] Galiano R. et al (2012), “An assessment of the effectiveness of a Random Forest classifier for land-cover classification.” ISPRS J. Photogramm. Remote Sens. 67, 93–104.

[82] Hastie T. et al. (2009), “The Elements of Statistical Learning : Data Mining, Inference, and Prediction”, Springer, Random Forest ; Ensemble Learning, 587-624

[83] Immitzer, M. et al (2016), “First experience with sentinel-2 data for crop and tree species classifications in Central Europe.” Remote Sens., 8, 166

[84] Lillesand T.M., & Kiefer R.W., (1994). Remote Sensing and Image Interpretation, John Wiley & Sons, New York, 3rd ed.

[85] Lillesand T.M. et al. (2004), “Remote Sensing and Image Interpretation ”, John Wiley & Sons, USA, Image Classification,550-552.

[86] Mandanici E. and Bitelli G., (2016), “Preliminary Comparison of Sentinel-2 and Landsat-8 Imagery for a Combined Use”, Remote Sens., 8, 1014. 402

[87] Olofsson et al., 2014, "Good practices for estimating area and assessing accuracy of land change", Remote Sensing of Envirinment, 148,42-57

[88] Pelletier C. et al, 2016, 'Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas', Remote Sensing of Environment 187 (2016) 156–168.

[89] Pesaresi, M. et al (2016). "Assessment of the added-value of sentinel-2 for detecting built-up areas." Remote Sens., 8, 299

[90] Pontius, R. G. (2000). "Quantification error versus location error in comparison of categorical maps.", Photogrammetric Engineering & Remote Sensing, 66, 1011–1016.

[91] Pontius, R. G., & Lippitt, C. D. (2006). "Can error explain map differences over time?", Cartography and Geographic Information Science, 33, 159–171Rogan, J., and D. Chen. (2004). "Remote Sensing Technology for Mapping and Monitoring Land-Cover and Land-use Change." Progress and Planning 61 (4): 301 - 325.

[92] Rokach L. (2009), "Ensemble-based classifiers", Artif Intell Rev (2010) 33, 1–

[93] Schuster, C. et al (2012). "Testing the red edge channel for improving land-use classifications based on high-resolution multi-spectral satellite data.", Int. J. Remote Sens., 33, 5583–5599

[94] Sharma R. et al (2013), "Decision tree approach for classification of remotely sensed satellite data using open source support", J. Earth Syst. Sci. 122, p 1237–1247

[95] Cormen, T. H. Leiserson, C. E. Rivest, R. L. Stein, C. (2001), Introduction to Algorithms (2nd ed.), MIT Press and McGraw-Hill.

[96] Ι. Μανωλόπουλος, Δομές δεδομένων, Εκδόσεις Art ofText (1998).

[97] Θ. Παπαθεοδώρου, Αλγόριθμοι: Εισαγωγικά Θέματα και Παραδείγματα, Εκδόσεις Πανεπιστημίου Πατρών, 1999.

[98] Hartigan J. A. & Wong M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society, Series C, 28(1), 100–108

[99] https://repository.kallipos.gr/bitstream/11419/2130/1/06_chapter05.pdf

[100] Lance G.N. & Williams W.T. (1967). A general theory of classification sorting strategies. I. hierarchical systems. Computer Journal, 9, 373-380.

[101] C. Cortes and V. Vapnik, "Support-vector network", *Machine Learning*, vol. 20, pp. 273-297, 1995. Show Context CrossRef Google Scholar

[102] X. X. Niu and C. Y. Suen, "A novel hybrid CNN-SVM classifier for recognizing handwritten digits", *Pattern Recognition*, vol. 45, pp. 1318-1325, 2012. Show Context CrossRef Google Scholar

[103] S. Kang, P. Kang, T. Ko, S. Cho, S. J. Rhee and K. S. Yu, "An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction", *Expert Systems with Applications*, vol. 42, no. 9, pp. 4265-4273, 2015. Show Context CrossRef Google Scholar

[104] A. Z. Ala'M, H. Faris and M. A. Hassonah, "Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts", *Knowledge-Based Systems*, vol. 153, pp. 91-104, 2018. Show Context Google Scholar

[105] M. Tipping, "Sparse Bayesian learning and the relevance vector machine", *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001. Show Context Google Scholar

[106] X. Qian, H. Huang, X. Chen and T. Huang, " Efficient construction of sparse radial basis function neural networks using L 1 -regularization ", *Neural Networks*, vol. 94, pp. 239-254, 2017. Show Context CrossRef Google Scholar

[107] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", *Advances in Large Margin Classifiers*, vol. 10, no. 3, 1999. Show Context Google Scholar

[108] T. F. Wu, C. J. Lin and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling", *Journal of Machine Learning Research*, vol. 5, pp. 975-1005, 2004. Show Context Google Scholar

[109] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011, [online] Available: http://www.csie.ntu.edu.tw/cjlin/libsvm. Show Context Access at ACM Google Scholar

[110] H. T. Lin, C. J. Lin and R. C. Weng, "A note on Platts probabilistic outputs for support vector machines", *Machine Learning*, vol. 68, no. 3, pp. 267-276, 2007. Show Context CrossRef Google Scholar

[111] T. Damoulas, M. A. Girolami, Y. Ying and C. Campbell, "Inferring sparse kernel combinations and relevance vectors: an application to subcellular localization of proteins", *Proceedings of the 7th International Conference on Machine Learning and Applications*, pp. 577-582, 2008. Show Context View Article Full Text: PDF (423KB) Google Scholar

[112] I. Psorakis, T. Damoulas and M. A. Girolami, "Multiclass relevance vector machines: sparsity and accuracy", *IEEE Trans. on Neural Networks*, vol. 21, no. 10, pp. 1588-1598, 2010, [online] Available: https://github.com/ipsorakis/mRVMs.git. Show Context View Article Full Text: PDF (1587KB) Google Scholar

[113] H. Chen, P. Tiňo and X. Yao, "Probabilistic classifcation vector machines", *IEEE Trans. on Neurlal Networks*, vol. 20, no. 6, pp. 901-914, 2009, [online] Available: http://staff.ustc.edu.cn/hchen/software.htm. Show Context Google Scholar

[114] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1-38, 1977. Show Context CrossRef Google Scholar

[115] H. Chen, P. Tiňo and X. Yao, "Efcient probabilistic classifcation vector machine with incremental basis function selection", *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 356-369, 2014. Show Context View Article Full Text: PDF (1801KB) Google Scholar

[116] D. W. Aha, "Tolerating noisy irrelevant and novel attributes in instance-based learning algorithms", *Int. J. Man-Machine Studies*, vol. 36, pp. 267-287, 1992. Show Context CrossRef Google Scholar

[117] C. Stanfill and D. Waltz, "Towards memory-based reasoning", *Communications of the ACM*, vol. 29, no. 12, pp. 1213-1228, 1986. Show Context Access at ACM Google Scholar

[118] D. W. Aha, D. Kibler and M. K. Albert, "Instance-based learning algorithms", *Machine Learning*, vol. 6, pp. 37-66, 1991. Show Context CrossRef Google Scholar

[119] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967. Show Context View Article Full Text: PDF (766KB) Google Scholar

[120] W. Lam, C.K. Keung and D. Liu, "Discovering Useful Concept Prototypes for Classification Based on Filtering and Abstraction", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1075-1090, 2002. Show Context View Article Full Text: PDF (946KB) Google Scholar

[121] C. Perou, T. Sorlie, M. Eisen, M. Van De Rijn, S. Jeffrey, C. Rees, J. Pollack, D. Ross, H. Johnsen, L. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. Zhu, P. Lonning, A. Borresen-Dale, P. Brown, D. Botstein **Molecular portraits of human breast tumours** Nature, 406 (2000), pp. 747-752 CrossRefView Record in ScopusGoogle Scholar

[122]J. Pollack, T. Sorlie, C. Perou, C. Rees, S. Jeffrey, P. Lonning, R. Tibshirani, D. Botstein, A. Borresen-Dale, P. Brown **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors** Proceedings of the National Academy of Sciences of the United States of America, 99 (2002), pp. 12963-12968 View Record in ScopusGoogle Scholar

[123]T. Sorlie, C. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. Eisen, M. Van De Rijn, S. Jeffrey, T. Thorsen, H. Quist, J. Matese, P. Brown, D. Botstein, P. Eystein Lonning, A. Borresen-Dale **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications** Proceedings of the National Academy of Sciences of the United States of America, 98 (2001), pp. 10869-10874 CrossRefView Record in ScopusGoogle Scholar

[124] L. Van'T Veer, H. Dai, M. Van De Vijver, Y. He, A. Hart, R. Bernards, S. Friend **Expression profiling predicts outcome in breast cancer** Breast Cancer Research, 5 (2003), pp. 57-58 View Record in ScopusGoogle Scholar

[125] M. Van De Vijver, Y. He, L. Van'T
Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts, M. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. Van Der
Velde, H. Bartelink, S. Rodenhuis, E. Rutgers, S. Friend, R. Bernards **A gene-expression signature as a predictor of survival in breast cancer** New England Journal of Medicine, 347 (2002), pp. 1999-2009 View Record in ScopusGoogle Scholar

[126] A. Naderi, A. Teschendorff, N. Barbosa-
Morais, S. Pinder, A. Green, D. Powe, J. Robertson, S. Aparicio, I. Ellis, J. Brenton, C. Caldas **A gene-expression signature to predict survival in breast cancer across independent data sets** Oncogene, 26 (2006), pp. 1507-1516 Google Scholar

[127] M. Galea, R. Blamey, C. Elston, I. Ellis **The Nottingham Prognostic Index in primary breast cancer** Breast Cancer Research and Treatment, 22 (1992), pp. 207-219 View Record in ScopusGoogle Scholar

[128] I. Evett, E. Spiehler **Rule induction in forensic science** KBS in Goverment, Online Publications (1987), pp. 107-118 View Record in ScopusGoogle Scholar

[129] S. Haberman, Generalized residuals for log-linear models, in: Proceedings of the 9th International Biometrics Conference, pp. 104–122. Google Scholar

[130] P. Royston **Algorithm as 181: The w test for normality** Applied Statistics, 31 (1982), pp. 176-180 CrossRefView Record in ScopusGoogle Scholar

[131] F. Harrell Jr., K. Lee, D. Mark **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors** Statistics in Medicine, 15 (1996), pp. 361-387 View Record in ScopusGoogle Scholar

[132] D.T. Larose, "Discovering Knowledge in Data An Introduction to Data Mining", *Wiley Interscience*, pp. 90-106. Show Context Google Scholar

[133] R. Agrawal, "K-Nearest Neighborn for Uncertain Data", *International Journal of Computer Applications (0975–8887)*, vol. 105, no. 11, pp. 13-16, 2014. Show Context Google Scholar

[134] Parvin Hamid, Alizadeth Hoseinali and M. Behrouz, "A Modification on K-Nearest Neighborn Classifier", *Global Journal of computer Science and Technolgy.*, vol. 10, no. 14, pp. 37-41, 2010. Show Context Google Scholar

[135] Parvin Hamid, Alizadeth Hoseinali, Minati, M. Behrouz and Bidgoli, "MKNN: Modified K-Nearest Neighborn", *Proceedings of the World Congress on Engineering and Computer Science WCECS.*, pp. 1-4, 2008, ISBN 978-988-98671-0-2. Show Context Google Scholar

[136] Aruma Singh, Smita Patel and Shukla, "Applying Modified K-Nearest Neighborn to Detect Threat in Collaborative Information Information Systems", *International Journal of Innovative Research in Science Engineering and Technology.*, vol. 3, no. 6, pp. 14141-14151, 2014. Show Context Google Scholar

[137] G. Haixiang, L. Yijing, L. Yanan, L. Xiao and L. Jinling, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification", *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176-193, Mar. 2016. Show Context CrossRef Google Scholar

[138] E. Khadangi and A. Bagheri, "Comparing MLP SVM and KNN for predicting trust between users in Facebook", *ICCKE 2013*, pp. 466-470, 2013. Show Context View Article Full Text: PDF (537KB) Google Scholar

[139] *A bi-directional sampling based on K-means method for imbalance text classification*, [online] Available: https://www.computer.org/csdl/proceedings-article/icis/2016/07550920/12OmNwO5M0m. Show Context Google Scholar

[140] D. Banerjee, G. Prabhat and R. Bhowal, "iCASSTLE : Imbalanced Classification Algorithm for Semi Supervised Text Learning", *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1012-1016, 2018. Show Context View Article Full Text: PDF (94KB) Google Scholar

[141] A. Mohasseb, M. Bader-El-Den, M. Cocea and H. Liu, "Improving Imbalanced Question Classification Using Structured Smote Based Approach", 2018. Show Context View Article Full Text: PDF (178KB) Google Scholar

[142] Han and M. Kamber, "Data Mining: Concepts and Techniques" in , Morgan Kaufmann Publishers, 2001. Show Context Google Scholar

[143] Magdalena Wrzesien and Karol Zaremski, "Mapping Warsaw from space application of object oriented approach to the analysis of urban structure", *Remote Sensing of Environment Laboratory Faculty of Geography and Regional Studies*, 2004. Show Context Google Scholar

[144] M. Pal and P. M. Mather, "Decision Tree Based Classification of Remotely Sensed data", 2001. Show Context Google Scholar

[145] J. R. Quinlan, *SeeS Manual*, 1997. Show Context Google Scholar

[146] Qiang Ding, William Perrizo and Victor Shi, "Integrating Query Processing and Data Mining in Relational DBMSs using P-Tree", 2003. Show Context Google Scholar

[147] Qin Ding, Maleq Khan, Amalendu Roy and William Perrizo, "The P-tree Algebra", 2003. Show Context Google Scholar

[148] Xiaoxia Sun, Jixian Zhang and Zhengjun Liu, "A comparison of object-oriented and pixel-based classification approaches using QUICKBIRD imagery" in , Beitaiping Rd, Haidian District, Beijing, China:Chinese Academy of Surveying and Mapping, no. 16, 2005. Show Context Google Scholar

[149] Sebastian Schiefer, Patrick Hostert, Alexander Damm and Ellen Diermayer, "Compression and Object-Oriented processing of Segmented Hyper-spectral image in ENVI", *Proceedings of 4th EARSeL Workshop on Imaging Spectroscopy*, 2005. Show Context Google Scholar

[150] B. Tso and P. Mather, "Classification Methods for Remotely Sensed Data" in , Taylor and Francis Inc, 2001. Show Context CrossRef Google Scholar

[151] Thomas Blaschke and Geoffrey J. Hay, "Object-oriented image analysis and scale-space: Theory and methods for modeling and evaluating multiscale landscape structure" in , Montreal:Department of Geography and Geoinformation, University of Salzburg, Austria, Geocomputing Laboratory, Department de geography, University de Montreal, 2005. Show Context Google Scholar

[152] YH Liu and YT Chen, "Face recognition using total margin-based adaptive fuzzy support vector machines[J]", *Neural Networks IEEE Transactions*, vol. 18, no. 1, pp. 178-192, 2007. Show Context Google Scholar

[153] TY Wang and HM Chang, "Fuzzy support vector machine for multi-class text categorization[J]", *Information Processing&Management*, vol. 43, no. 4, pp. 914-929, 2007. Show Context CrossRef Google Scholar

[154] Vikramjit Mitra, Chia-Jiu Wang and Satarupa Banerjee, "Text classification: A least square support vector machine approach[J]", *Applied Soft Computing*, vol. 7, no. 3, pp. 908-914, 2007. Show Context CrossRef Google Scholar

[155] J Li, N Allinsion, DC Tao and XL Liu, "Multi-training Support Vector Machine for Image Retrieval[J]", *ieee Transactions on Image Processing*, pp. 3597-3601, 2006. Show Context Google Scholar

[156] S Yu, YM Ye and FY Ma, "Accurate performance estimators for information retrieval based on span bound of support vector machine[J]", *Journal of Harbin Institute of Technology*, vol. 13, no. 1, pp. 113-117, 2006. Show Context Google Scholar

[157] Suykens. JAK and J Vandewalle, "Least squares support vector machine classifiers[J]", *NEURAL PROCESSING LETTERS*, vol. 9, no. 3, pp. 293-300, 1999. Show Context Google Scholar

[158] Q Wu and R Law, "An intelligent forecasting model based on robust wavelet v-support vector machine[J]", *EXPERT SYSTEMS WITH APPLICATIONS*, vol. 38, no. 5, pp. 4851-4859, 2011. Show Context CrossRef Google Scholar

[159] G. Fung and O.L. Mangasarian, "Proximal support vector machine classificers[C]", *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 77-86, 2001. Show Context Access at ACM Google Scholar

[160] Jayadeva, R. Khemchandani and S. chandra, "Twin support vector machines for pattern classification[J]", *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 5, pp. 905-910, 2007. Show Context View Article Full Text: PDF (1120KB) Google Scholar

[161] M.A. Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification[J]", *Expert Systems with Applications*, vol. 36, no. 4, pp. 7535-7543, 2009. Show Context CrossRef Google Scholar

[162] *Handbook of Pattern Recognition and Computer Vision*, pp. 61-124, 1993. Show Context Google Scholar

[163] T. Smith and M. Waterman, "New stratigraphic correlation techniques", *J. Geology*, vol. 88, pp. 451-457, 1980. Show Context CrossRef Google Scholar

[164] P. Smyth, "Clustering using Monte Carlo cross-validation", *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining*, pp. 126-133, 1996. Show Context Google Scholar

[165] "Clustering sequences with hidden Markov models" in Advances in Neural Information Processing, MA, Cambridge:MIT Press, vol. 9, pp. 648-654, 1997. Show Context Google Scholar

[166] P. Smyth, "Model selection for probabilistic clustering using cross validated likelihood", *Statist. Comput.*, vol. 10, pp. 63-72, 1998. Show Context CrossRef Google Scholar

[167] P. Smyth, "Probabilistic model-based clustering of multivariate and sequential data", *Proc. 7th Int. Workshop on Artificial Intelligence and Statistics*, pp. 299-304, 1999. Show Context Google Scholar

[168] P. Sneath, "The application of computers to taxonomy", *J. Gen. Microbiol.*, vol. 17, pp. 201-226, 1957. Show Context CrossRef Google Scholar

[169] P. Somervuo and T. Kohonen, "Clustering and visualization of large protein sequence databases by means of an extension of the self-organizing map", *LNAI 1967*, pp. 76-85, 2000. Show Context Google Scholar

[170] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons", *Biologiske Skrifter*, vol. 5, pp. 1-34, 1948. Show Context Google Scholar

[171] H. Späth (Spath), Cluster Analysis Algorithms, U.K., Chichester:Ellis Horwood, 1980. Show Context Google Scholar

[172] P. Spellman, G. Sherlock, M. Ma, V. Iyer, K. Anders, M. Eisen, et al., "Comprehensive identification of cell cycle-regulated genes of the Yeast Saccharomyces Cerevisiae by microarray hybridization", *Mol. Biol. Cell*, vol. 9, pp. 3273-3297, 1998. Show Context CrossRef Google Scholar

[173] M. Steinbach, G. Karypis and V. Kumar, "A comparison of document clustering techniques", *Tech. Rep. 00* , 2000. Show Context Google Scholar

[174] K. Stoffel and A. Belkoniene, " Parallel \$K\$ -means clustering for large data sets ", *Proc. EuroPar'99 Parallel Processing*, no. 1685, pp. 1451-1454, 1999. Show Context Google Scholar

[175] M. Su and H. Chang, "Fast self-organizing feature map algorithm", *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 721-733, May 2000. Show Context View Article Full Text: PDF (372KB) Google Scholar

[176] M. Su and C. Chou, " A modified version of the \$K\$ -means algorithm with a distance based on cluster symmetry ", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 674-680, Jun. 2001. Show Context View Article Full Text: PDF (794KB) Google Scholar

[177] R. Sun and C. Giles, "Sequence learning: Paradigms algorithms and applications", *LNAI 1828*, 2000. Show Context Google Scholar

[178] C. Sung and H. Jin, "A Tabu-search-based heuristic for clustering", *Pattern Recognit.*, vol. 33, pp. 849-858, 2000. Show Context CrossRef Google Scholar

[179] J. Kogan, C. Nicholas and M. Teboulle, "A Survey of Clustering Data Mining Techniques", *Berkhin P.*, pp. 25-71, 2006. Show Context Google Scholar

[180] A.K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010. Show Context CrossRef Google Scholar

[181] A. Ben Ayed, M. Ben Halima and A.M. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data", *6th International Conference of Soft Computing and Pattern Recognition (SoCPaR) Tunis Tunisia*, pp. 331-336, Aug 11-14, 2014. Show Context Google Scholar

[182] K. Kalti and M.A. Mahjoub, "Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm", *The International Arab Journal of Information Technology*, vol. 11, no. 1, Jan 2014. Show Context Google Scholar

[183] A.Y. Ng, M. Jordan and Y. Weiss, "On spectral clustering: Analysis and an algorithm", *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, no. 2, pp. 849-856, Dec 2001. Show Context Google Scholar

[184] E. W. Forgy, "Cluster analysis of multivariate data: efficiency vs interpretability of classifications", *Biometrics*, vol. 21, pp. 768-769, 1965. Show Context Google Scholar

[185] A. Ben Ayed, M. Ben Hammouda, M. Ben Halima and A.M. Alimi, "Random Forests based fuzzy Logic", *Ninth International Conference on Machine Vision (ICMV'2016) Nice France Nov 18-20 2016 International Society for Optics and Photonics SPIE*, vol. 10341, pp. 103412B-1-103412B-5. Show Context Google Scholar

[186] I. Lahmar, A. Ben Ayed, M. Ben Halima and A.M. Alimi, "Cluster Forest Based Fuzzy Logic for Massive Data Clustering", *Ninth International Conference on Machine Vision (ICMV'2016) Nice France*, vol. 10341, pp. 103412J-1-103412J-5, Nov 18-20, 2016. Show Context Google Scholar

[187] A. Ben Ayed, M. Ben Halima and A. M. Alimi, "Cluster Forests Based Fuzzy C-Means for Data Clustering", *9th International Conference on Computational Intelligence in Security for Information Systems San-Sebastian Spain*, pp. 564-573, Oct 19-21, 2016. Show Context Google Scholar

[188] Z.H. Zhou, "Ensemble methods: foundations and algorithms", *CRC press*, Jun 2012. Show Context Google Scholar

[189] H. Dhahri and A.M. Alimi, "The modified differential evolution and the RBF (MDE-RBF) neural network for time series prediction", *IEEE International Conference on Neural Networks 2006 Conference Proceedings*, pp. 2938. Show Context Google Scholar

[190] S. Bouaziz, H. Dhahri, A.M. Alimi and A. Abraham, "A hybrid learning algorithm for evolving Flexible Beta Basis Function Neural Tree Model", *Neurocomputing*, vol. 117, pp. 107-117, 2013. Show Context CrossRef Google Scholar

[191] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory", *Micro Machine and Human Science MHS'95. IEEE Proceedings of the Sixth International Symposium*, pp. 39-43, Oct 4, 1995. Show Context Google Scholar

[192] J.M. Mendel and B. J. Ri, "Type-2 fuzzy sets made simple", *IEEE Transactions on fuzzy systems 10.2*, pp. 117-127, 2002. Show Context View Article Full Text: PDF (407KB) Google Scholar

[193] M. Zouari, S. Cherif, H. Kammoun, H. Lajmi and A.M. Alimi, "Towards type-2 fuzzy rule base system for road choice", *15th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 243-248, Dec 14 2015. Show Context Google Scholar

[194] H. Liu, F. Zhao and L. Jiao, "Fuzzy spectral clustering with robust spatial information for image segmentation", *Applied Soft Computing*, vol. 12, no. 11, pp. 3636-3647, 2012. Show Context CrossRef Google Scholar

[195] J. Shi and J. Malik, "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000. Show Context View Article Full Text: PDF (2813KB) Google Scholar

[196] X. Wang, B. Qian and I. Davidson, "On constrained spectral clustering and its applications", *Data Mining and Knowledge Discovery*, vol. 28, pp. 1-30, Jan 2014.Show Context CrossRef Google Scholar

[197] U.V. Luxburg, "A tutorial on spectral clustering", *Statistics and computing*, vol. 17, no. 4, pp. 395-416, 2007. Show Context CrossRef Google Scholar

[198] Z. Li, C. Shang and Q. Shen, "Fuzzy-clustering embedded regression for predicting student academic performance", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 344-351, Jul 24, 2016. Show Context Google Scholar

[199] M. Xu, M. Guo, L. Shang and X. Jia, "Multi-value image segmentation based on FCM algorithm and Graph Cut Theory", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1333-1340, Jul 24, 2016. Show Context Google Scholar

[200] G. Chandrashekar and F. Sahin, "A survey on feature selection methods", *Computers & Electrical Engineering.*, vol. 40, no. 1, pp. 16-28, Jan 2014. Show Context CrossRef Google Scholar

[201] A. K. Jain, " Data clustering: 50 years beyond k k -means ", *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651-666, 2010. Show Context CrossRef Google Scholar

[202] U. von Luxburg, "A tutorial on spectral clustering", *Statist. Comput.*, vol. 17, no. 4, pp. 395-416, 2007. Show Context CrossRef Google Scholar

[203] W. Y. Chen, Y. Song, H. Bai, C. J. Lin and E. Y. Chang, "Parallel spectral clustering in distributed systems", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 568-586, 2011. Show Context View Article Full Text: PDF (3476KB) Google Scholar

[204] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation", *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1669-1680, Aug. 2015. Show Context View Article Full Text: PDF (1305KB) Google Scholar

[205] L. He, N. Ray, Y. Guan and H. Zhang, "Fast large-scale spectral clustering via explicit feature mapping", *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1058-1071, Mar. 2018. Show Context View Article Full Text: PDF (1919KB) Google Scholar

[206] H. Liu, R. Zhao, H. Fang, F. Cheng, Y. Fu and Y.-Y. Liu, "Entropy-based consensus clustering for patient stratification", *Bioinf.*, vol. 33, no. 17, pp. 2691-2698, 2017. Show Context CrossRef Google Scholar

[207] H. Liu, J. Wu, T. Liu, D. Tao and Y. Fu, "Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence", *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1129-1143, May 2017. Show Context View Article Full Text: PDF (599KB) Google Scholar

[208] D. Huang, J.-H. Lai and C.-D. Wang, "Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis", *Neurocomputing*, vol. 170, pp. 240-250, 2015. Show Context CrossRef Google Scholar