ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΠΑΤΡΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΕΦΑΡΜΟΓΩΝ ΤΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΤΗΝ ΟΙΚΟΝΟΜΙΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΕΩΝ ΣΤΗΝ ΛΗΨΗ ΑΠΟΦΑΣΕΩΝ ΜΕ ΧΡΗΣΗ DATAMINING

STUDY CASES OF DECISION MAKING USING DATA MINING METHODS

ΒΟΖΙΚΗΣ ΧΡΗΣΤΟΣ **ΖΑΦΕΙΡΟΠΟΥΛΟΣ ΑΛΕΞΙΟΣ**

Επίβλεψη:
ΑΝΤΖΟΥΛΑΤΟΣ ΓΕΡΑΣΙΜΟΣ

ΑΜΑΛΙΑΔΑ 2012

ΑΝΤΙ-ΠΡΟΛΟΓΟΥ

Η παράγραφος αυτή αφιερώνεται σε όλους εκείνους που βοήθησαν, ώστε να ολοκληρωθεί αυτή η πτυχιακή διπλωματική εργασία. Ο πρώτος που θα θέλαμε να ευχαριστήσουμε είναι ο επιβλέποντας καθηγητής μας κ. Αντζουλάτος για την εμπιστοσύνη που μας έδειξε, δεδομένου των ιδιαιτεροτήτων που αντιμετωπίσαμε, την καλή θέληση για συνεργασία μεταξύ μας και την πολύτιμη βοήθεια και καθοδήγηση που μας παρείχε.

Κατόπιν θα θέλαμε να ευχαριστήσουμε τον πρώην επιβλέποντα καθηγητή μας κ. Παπαστεργίου Θωμά, ο οποίος αναγκάστηκε λόγο προσωπικών υποχρεώσεων να αποχωρίσει από την επίβλεψή της πτυχιακής εργασίας. Παρ' όλα αυτά η βοήθεια και καθοδήγηση του ακόμα και κατόπιν της αποχώρησης του από την επίβλεψή της πτυχιακής εργασίας.

Ένα μεγάλο ευχαριστώ πρέπει να αποδοθεί ακόμα στους φίλους και συνεργάτες: Ιωάννη Μεταξά και Χρύσανθο Δάρμα οι οποίοι θυσίασαν προσωπικό χρόνο προκειμένου να μας βοηθήσουν με συμβουλές υποδείξεις και υλικό.

Η διπλωματική αφιερώνεται στα προαναφερθέντα πρόσωπα, καθώς και σε όλους εκείνους οι οποίοι έχουν σκοπό της ζωής τους την γνώση και τη μάθηση.

Βοζίκης Χρήστος

Ζαφειρόπουλος Αλέξης

Αμαλιάδα, 2012

Περιεχόμενα

ΠΕΡΙΛΗΨΗ

Κύριος στόχος της εργασίας αυτής είναι να μυήσει μη εξοικειωμένους χρήστες στην εξόρυξη δεδομένων (data mining). Αυτό επιτυγχάνεται με την αναφορά βασικών μεθόδων επεξεργασίας δεδομένων καθώς και αλγορίθμων οι οποίοι μας βοηθούν στις τεχνικές εξόρυξης δεδομένων. Στο πρώτο κεφάλαιο, δίνεται ο ορισμός της ανακάλυψη γνώσης από δεδομένα, ο ορισμός της εξόρυξης δεδομένων και παραθέτονται πεδία στα οποία βρίσκουν εφαρμογή. Στο δεύτερο κεφάλαιο παραθέτονται οι τεχνικές και λειτουργίες εξόρυξης δεδομένων καθώς και τα χαρακτηριστικά της επιτυχημένης εξόρυξης. Το κεφάλαιο συνεχίζει αναφερόμενο στα μέτρα και συντελεστές δεδομένων, τα οποία εν δυνάμει επηρεάζουν την διαδικασία του data mining. Προχωρώντας στο τρίτο κεφάλαιο γίνεται εκτενής αναφορά στους κανόνες συσχέτισης, τις εφαρμογές που αυτοί βρίσκουν, τους κανόνες που τους διέπουν καθώς και μια εκτενής αναφορά στους αλγορίθμους που χρησιμοποιούνται. Στο κεφάλαιο αυτό υπάρχει και ανάλυση του αλγορίθμου A-priori. Τα κεφάλαια 4 και 5 ασχολούνται με τις λειτουργίες της ταξινόμησης και ομαδοποίησης αντίστοιχα, και αναλύονται οι βασικές κατηγορίες αλγορίθμων τους. Στο κεφάλαιο 6 αναλύονται μέτρα αξιολόγησης των αποτελεσμάτων που προκύπτουν από την λειτουργία των αλγορίθμων Εξόρυξης Δεδομένων. Στο κεφάλαιο 7 παρουσιάζονται οι οθόνες από το γραφικό περιβάλλον που υλοποιήθηκε με σκοπό ο χρήστης να μπορεί να εκτελέσει τους αλγόριθμους k-means και k-nn. Εν συνεχεία αναλύονται τα αποτελέσματα από την εκτέλεση των δύο αλγορίθμων πάνω σε τεχνητά και πραγματικά σύνολα δεδομένων, η ποιότητα των αποτελεσμάτων μετριέται με βάση δύο από τα μέτρα αξιολόγησης, που αναφέρονται στο κεφάλαιο 6.

SUMMARY

The main objective of this thesis is to initiate those users not acclimated to the uses of data mining. This is achieved with the report of basic methods of treatment of data as well as algorithms which help us in the techniques of data extraction. In the first chapter, is given the definition of discovery of knowledge from data, the definition of extraction of data and are illustrated fields in which it finds application. In the second chapter illustrated are, the techniques and operations of data mining extraction as well as the characteristics of achieved extraction. The chapter continues to report about the meters and factors of data, which potentially influence the process of data mining. Advancing in the third chapter becomes an extensive report in the rules of cross-correlation, the applications that these find, the rules that condition them as well as an extensive report in the algorithms that are used. This chapter also analyzes the A-priori algorithm. The chapters 4 and 5 deal with the operations of classification and clustering respectively. Also the basic categories of algorithms are presented in these chapters. Chapter 6 deals with the validation measures that can be used to measure the efficiency of the results of Data Mining algorithms. In Chapter 7 presents the screens in the Graphical User Interface that implemented in order for the user to execute the algorithms k-means and k-nn. It then analyzes the results of the execution of the two algorithms on artificial and real data sets, while the quality of the results measured by two of the evaluation measures that mention in chapter 6.

1. Εισαγωγή

Η ανακάλυψη γνώσης και η εξόρυξη δεδομένων είναι δύο ισχυρά εργαλεία ανάλυσης δεδομένων. Το data mining είναι ένα σημαντικό εργαλείο στην προσπάθεια του φιλτραρίσματος, εξόρυξης, η μετατροπής μεγάλων βάσεων δεδομένων σε στοχευμένα και περιεκτικά σύνολα πληροφοριών και την εύρεση κρυμμένων μοτίβων κατά την εύρεση γνώσης. Διάφοροι αλγόριθμοι μπορούν να χρησιμοποιηθούν για να ανασύρουν τις πληροφορίες που χρειάζονται για την επίλυση διαφόρων προβλημάτων της καθημερινής ζωής. Οι όροι «ανακάλυψη γνώσης» και «εξόρυξη δεδομένων» χρησιμοποιούνται για να περιγράψουν την εξαγωγή της υπονοούμενης, προηγουμένως άγνωστης και δυνητικά ωφέλιμης γνώσης από τα δεδομένα.

Η εξόρυξη δεδομένων είναι ένας διεπιστημονικός τομέας με πολλές λειτουργίες. Με αυτές τις λειτουργίες μπορούν να δημιουργηθούν μοντέλα εξόρυξης τα οποία περιγράφουν τα δεδομένα που θα χρησιμοποιηθούν όπως οι αλγόριθμοι συγκέντρωσης, οι κανόνες συσχέτισης κλπ.

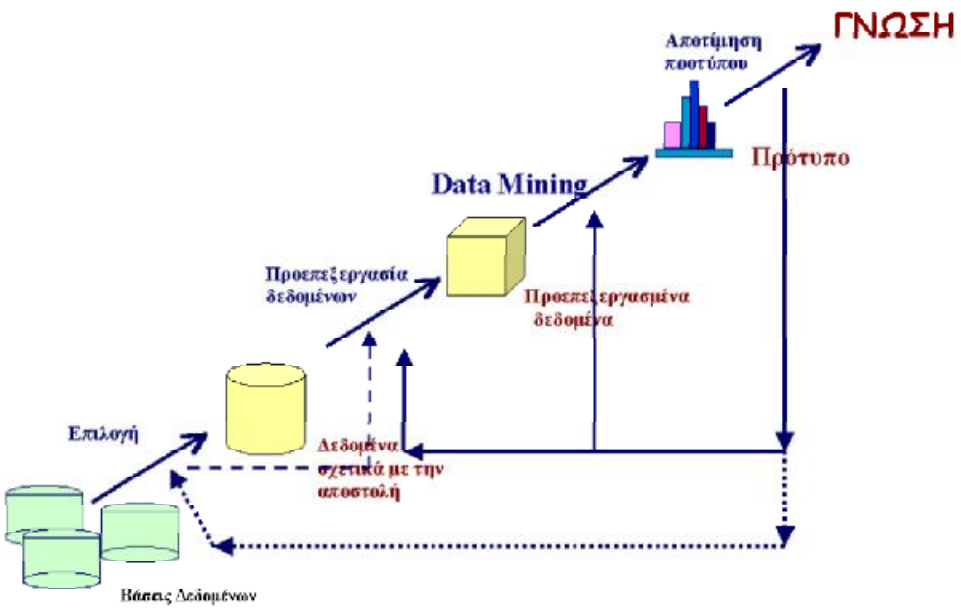
Η παρούσα εργασία με τίτλο μελέτη περιπτώσεων στη λήψη αποφάσεων με χρήση Data Mining, έχει σαν στόχο να αναδείξει τον βαθμό που μπορούν οι μέθοδοι Εξόρυξης Δεδομένων να αξιοποιήσουν δεδομένα για να εξάγουν χρήσιμα συμπεράσματα και γνώση. Στο σύνολο της εργασίας παρουσιάζεται η σημασία και η έννοια της τεχνικής Data Mining, περιγράφονται οι σημαντικότερες μέθοδοι (Ταξινόμηση, Ομαδοποίηση, Συσχέτιση) καθώς και πληροφορίες για τα διάφορα πεδία που εφαρμόζεται. Γίνεται εκτενής αναφορά στους κανόνες συσχέτισης ανάμεσα στα δεδομένα και τέλος με τη χρήση του αλγορίθμου K-NN (K nearest neighbors) καταλήγουμε σε χρήσιμα συμπεράσματα.

Μόλις γίνουν κατανοητές βασικές έννοιες της εξόρυξης δεδομένων, παρουσιάζονται κάποια παραδείγματα και κάποιοι αλγόριθμοι (υλοποιημένοι σε Matlab) ώστε να γίνει κατανοητή η ακριβής λειτουργία της ομαδοποίησης και της ταξινόμησης. Για τους αλγόριθμους επιλέχτηκε η Matlab, καθώς είναι εύκολη στην χρήση και στον προγραμματισμό, διαθέτει πληθώρα έτοιμων συναρτήσεων και πακέτων (toolboxes) τα οποία σχετίζονται ή / και μπορούν να χρησιμοποιηθούν για την ανάπτυξη νέων αλγορίθμων εξόρυξης δεδομένων. Τέλος μπορεί να χειρίζεται διαφορετικούς τύπους δεδομένων. Εάν το λογισμικό επεκταθεί μπορεί αν περιλαμβάνει και άλλους αλγορίθμους εξόρυξης δεδομένων.

1.1 Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Η εξόρυξη γνώσης από μεγάλες αποθήκες δεδομένων είναι πλέον ένα από τα βασικά ερευνητικά πεδία που απασχολεί σημαντικούς τομείς όπως είναι οι βάσεις δεδομένων και η στατιστική. Η αποδοτική διαχείριση μεγάλων βάσεων δεδομένων αποτελεί πλέον επιτακτική ανάγκη. Καθώς το μέγεθος της πληροφορίας είναι τόσο μεγάλο που η επεξεργασία της στο σύνολο της είναι πλέον ανέφικτη. Ο τεράστιος όγκος πληροφορίας μας οδήγησε λοιπόν στην ανάγκη για φιλτράρισμά της έτσι ώστε να εξαχθεί η λεγόμενη χρήσιμη γνώση. Η ανακάλυψη γνώσης από τις βάσεις δεδομένων (Knowledge discovery in data mining - KDD) αναφέρεται σε ολόκληρη την διαδικασία ανακάλυψης πολύτιμης γνώσης από αποθήκες δεδομένων (data warehouses). Σύμφωνα με τον ορισμό που δίνεται, *«η ανακάλυψη γνώσης από βάσεις δεδομένων είναι μια ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και τελικά κατανοητών προτύπων στα δεδομένα. Πρόκειται για μια επαναλαμβανόμενη ακολουθία από βήματα που περιλαμβάνει την συλλογή, την εξέταση και την μοντελοποίηση μεγάλων ποσοτήτων δεδομένων για την αποκάλυψη έως πρόσφατα άγνωστων προτύπων»*. Η Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων (KDD) είναι μια μη-τετριμμένη διαδικασία για την αναγνώριση έγκυρων, νέων, χρήσιμων και εύκολα κατανοητών προτύπων από τα δεδομένα. Εδώ πρέπει να σημειωθεί ότι η ανακάλυψη γνώσης αποτελεί έναν ταχέως αναπτυσσόμενο τομέα, η εξέλιξη του οποίου κατευθύνεται τόσο από ερευνητικά ενδιαφέροντα όσο και από ισχυρές πρακτικές, οικονομικές και επιστημονικές ανάγκες [3],[4].

Η εξόρυξη δεδομένων αποτελεί υποσύνολο της διαδικασίας KDD και μάλιστα βρίσκεται στον πυρήνα αυτής, θα αναφερθούμε διεξοδικότερα σε αυτήν στο Κεφάλαιο 2. Το παρακάτω απλό σχεδιάγραμμα απεικονίζει τις διαδικασίες που λαμβάνουν χώρα κατά την διαδικασία Ανακάλυψης Γνώσης.



Εικόνα 1 Η διαδικασία Εξεύρεσης γνώσης [5]

Τα βήματα της διαδικασίας KDD έχουν ως εξής [5]:

1 Ανάπτυξη και κατανόηση του πεδίου της εφαρμογής συμπεριλαμβανόμενης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα καθώς επίσης και των στόχων / προσδοκιών των τελικών χρηστών.

2 Δημιουργία του στοχευόμενου συνόλου δεδομένων (target data), το οποίο θα περιλαμβάνει τα δεδομένα από τα οποία πρόκειται να εξαχθεί η γνώση. Το βήμα αυτό είναι εξαιρετικά κρίσιμο καθώς η ποιότητα των δεδομένων επηρεάζει την απόδοση του συστήματος ανακάλυψης γνώσης.

3 Καθαρισμός και επεξεργασία των δεδομένων (data cleaning). Το βήμα αυτό περιλαμβάνει βασικές λειτουργίες όπως η απομάκρυνση του θορύβου, η αντιμετώπιση του προβλήματος των δεδομένων με ελλιπείς τιμές κ.α.

4 Μείωση της ποσότητας των δεδομένων (data reduction). Το βήμα αυτό περιλαμβάνει την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων του προβλήματος ανάλογα με τους στόχους της ανακάλυψης γνώσης, τη μείωση του πλήθους αυτών των χαρακτηριστικών κ.α.

5 Επιλογή των εργασιών εξόρυξης γνώσης (data mining) που θα χρησιμοποιηθούν για τις ανάγκες του προβλήματος, π.χ. ταξινόμηση, πρόβλεψη, ομαδοποίηση κ.α.

6 Επιλογή των αλγορίθμων εξόρυξης γνώσης (data mining) που θα χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Το βήμα αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου, την επιλογή των κατάλληλων παραμέτρων του μοντέλου κ.α.

7 Ερμηνεία των προτύπων που ανακαλύφθηκαν από την KDD διαδικασία – πιθανόν να χρειαστεί να επιστρέψουμε και πάλι σε κάποιο από τα παραπάνω βήματα.

8 Ενοποίηση της γνώσης που έχει εξαχθεί: ενσωμάτωση αυτής της γνώσης στο σύστημα ή απλά κοινοποίησή της με την κατάλληλη τεκμηρίωση στα ενδιαφερόμενα μέλη. Το βήμα αυτό περιλαμβάνει και έλεγχο συγκρούσεων με την γνώση που επικρατούσε πριν.

9 Παρουσίαση και χρήση της ανακαλυφθείσας γνώσης.

Ο εκάστοτε χρήστης μπορεί να επανέλθει σε οποιοδήποτε βήμα αν κάτι δεν πήγε καλά ή αν απλά δεν είναι ευχαριστημένος από κάποιο αποτέλεσμα όπως επίσης μπορεί να ξεκινήσει την όλη διαδικασία όχι από την αρχή της αλλά από οποιοδήποτε ενδιάμεσο βήμα.

Εδώ κρίνεται απαραίτητη η ανάλυση μερικών εννοιών του ορισμού προκειμένου να γίνει καλύτερη η κατανόηση του ορισμού και της λειτουργιάς των διαδικασιών KDD χρησιμοποιώντας ως παράδειγμα το στιγμιότυπο μιας τράπεζας με τρία πεδία ανά εγγραφή στη βάση δεδομένων της. Το χρέος, την κατάσταση του δανείου του πελάτη και τέλος το εισόδημα του πελάτη.

- Δεδομένα

Πρόκειται για ένα σύνολο παραδειγμάτων / στιγμιότυπων ενός προβλήματος που εμφανίζονται σε μια βάση δεδομένων. Στο παράδειγμά μας, η συλλογή εγγραφών από τη βάση δεδομένων μιας τράπεζας, όπου κάθε εγγραφή θα περιλάμβανε τρία πεδία (γνωρίσματα): το χρέος, το εισόδημα και την κατάσταση του δανείου των πελατών της τράπεζας.

- Εγκυρότητα

Τα πρότυπα που προκύπτουν από τη διαδικασία ανακάλυψης γνώσης θα πρέπει να ισχύουν με κάποιο βαθμό βεβαιότητας και για νέα, άγνωστα στιγμιότυπα του προβλήματος. Για παράδειγμα, αν στο πρότυπο που απεικονίζεται στο παρακάτω σχήμα το κατώφλι μετακινηθεί προς τα δεξιά τότε το μέτρο βεβαιότητας θα μειωθεί καθώς περισσότερα αποδεκτά μέχρι πρότινος δάνεια θα ανήκουν πλέον στην περιοχή των μη αποδεκτών δανείων.

- Κατανόηση

Τα πρότυπα θα πρέπει να είναι κατανοητά από τον ανθρώπινο παράγοντα, καθώς οι άνθρωποι είναι αυτοί που θα κληθούν να τα αξιοποιήσουν προκειμένου να εξάγουν χρήσιμα συμπεράσματα και να αποκτήσουν μια βαθύτερη κατανόηση των δεδομένων τους. Για την κατανόηση των προτύπων δε θα πρέπει να απαιτούνται εξειδικευμένες γνώσεις, αντιθέτως τα πρότυπα θα πρέπει

να είναι πλήρως κατανοητά και να βοηθούν ακόμη και μη ειδικούς στην εξαγωγή χρήσιμων συμπερασμάτων.

- Χρησιμότητα

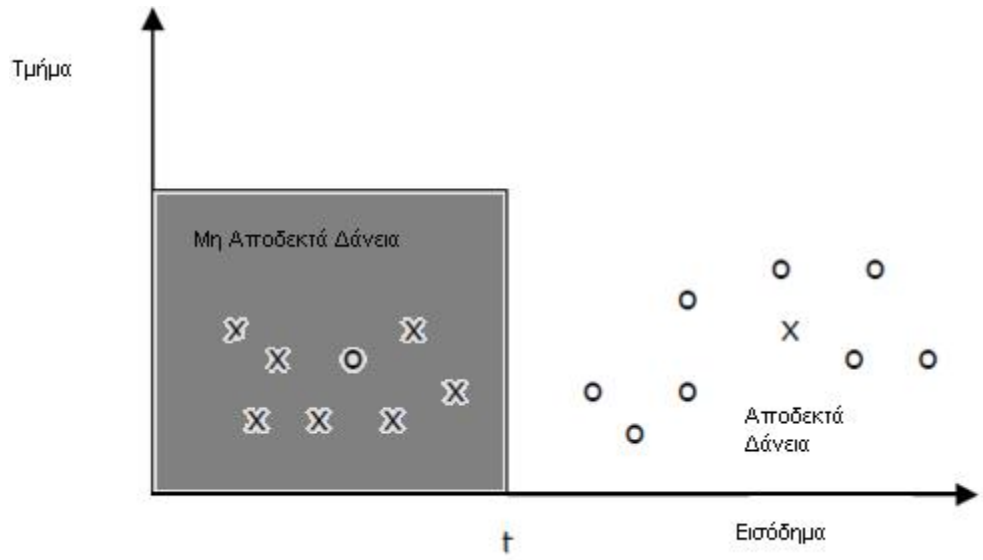
Τα πρότυπα θα πρέπει να είναι χρήσιμα, δηλαδή να οδηγούν σε κάποιες χρήσιμες ενέργειες. Για παράδειγμα, αν η τράπεζα εκμεταλλευτεί τους κανόνες απόφασης τους σχήματος, θα πρέπει να πετύχει αύξηση των κερδών της. Στο σχήμα του παραδείγματός μας, ο κανόνας απόφασης που προκύπτει είναι η έκδοση ή μη του δανείου ανάλογα με τη θέση της τιμής στο διάγραμμα.

- Πρότυπα (patterns)

Πρόκειται για εκφράσεις σε μια συγκεκριμένη γλώσσα οι οποίες περιγράφουν ένα υποσύνολο του συνόλου των παραδειγμάτων. Για παράδειγμα, η έκφραση ''Αν εισόδημα $< t$, τότε ο πελάτης δεν μπορεί να εξοφλήσει το δάνειο'', θα μπορούσε να είναι ένα πρότυπο για κάποιο κατάλληλο κατώφλι.

- Διαδικασία KDD

Πρόκειται για μια διαδικασία πολλών βημάτων που περιλαμβάνει την κατάλληλη προετοιμασία των δεδομένων, την αναζήτηση προτύπων και την αξιολόγηση της αποκτηθείσας γνώσης. Η KDD διαδικασία δεν είναι τετριμμένη καθώς εμπεριέχει κάποιο βαθμό αυτονομίας. Στο παράδειγμα του δανείου που αναφέραμε πιο πριν, ο υπολογισμός του μέσου όρου εισοδήματος του πελάτη αποτελεί πολύ χρήσιμο αποτέλεσμα, σε καμία όμως περίπτωση δεν αποτελεί ανακάλυψη γνώσης.



Εικόνα 2 Σχεδιάγραμμα απεικόνισης προτύπων για την έγκριση δανείου. [5]

2. Εξόρυξη Δεδομένων

2.1 Ορισμός Εξόρυξης Δεδομένων

Το ακατέργαστα δεδομένα δεν έχουν κανένα νόημα έως ότου επεξεργαστούν με κάποιο τρόπο. Με τις τεχνικές εξόρυξης δεδομένων νέες πληροφορίες μπορούν να παραχθούν από τα επεξεργασμένα δεδομένα. Ένας απέραντος όγκος δεδομένων εισέρχονται καθημερινά μέσω των διαφόρων δικτύων υπολογιστών στις βιομηχανίες και τις επιχειρήσεις. Για μια αλυσίδα σουπερμάρκετ τα δεδομένα πωλήσεων μπορούν να χρησιμοποιηθούν για να παραχθεί ένα μεγάλο ποσοστό πληροφοριών που είναι χρήσιμες στην ανάλυση αγοράς ώστε να προβλεφθεί η ζήτηση προϊόντος.

Χαρακτηριστικό παράδειγμα αποτελεί η μεγάλη Βρετανική αλυσίδα Tesco, η οποία το 1998 έκανε ένα μεγάλο βήμα ανοίγοντας υποκαταστήματα στις Ηνωμένες Πολιτείες. Η επιτυχία της αλυσίδας στη νέα αγορά ήρθε μέσα στους πρώτους έντεκα μήνες προκαλώντας την πρωτοκαθεδρία του παγκόσμιου κολοσσού που ονομάζετε Wal-Mart (WMT). Οι αναλυτές πιστεύουν ότι η τόσο σύντομη επιτυχία της Tesco, οφείλετε στις δυνατότητες του οργανισμού να επεξεργάζεται τα τεράστια ποσά δεδομένων που μαζεύει και να μεταφράζει αυτή τη γνώση σε πωλήσεις. Η Tesco χρησιμοποιεί τη γνώση που παράγει η Dunnhumby, μια βρετανική εταιρία εξόρυξης δεδομένων, για να διαχειρίζεται κάθε πτυχή της επιχείρησής της, από τη δημιουργία νέων σχεδίων καταστημάτων στην ανάπτυξη νέων προϊόντων και την πολιτική προώθησης νέων προϊόντων [50].

Η πρόοδος στα υπολογιστικά συστήματα και η εξέλιξη στην επικοινωνία έχει οδηγήσει στην δημιουργία μιας κοινωνίας ικανής να παρέχει διαρκώς νέες πληροφορίες. Το υλικό που συγκεντρώνεται καταγράφεται διαρκώς, με αποτέλεσμα τη δημιουργία τεράστιων βάσεων δεδομένων. Το ζήτημα που προκύπτει απ' όλα αυτά εντοπίζεται στο πώς μπορούμε να διαχειριστούμε όλες αυτές τις βάσεις δεδομένων και τις πληροφορίες που αποκτούμε. Όλα αυτά τα θέματα κατόπιν πολλών ετών μελέτης και σκέψης οδήγησαν στη δημιουργία της διαδικασίας της Εξόρυξης Δεδομένων (Data Mining). Πρόκειται για μία σειρά από μεθοδολογίες που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς ετερόκλητους κλάδους όπως: η οικονομία, η βιοστατιστική, η δημογραφία, η μετεωρολογία και η γεωλογία. Υπάρχουν πολλές και διαφορετικές απόψεις για το πώς θα μπορούσαμε να ορίσουμε την διαδικασία της εξόρυξης δεδομένων.

«Εξόρυξη Δεδομένων είναι η ανάλυση – συνήθως τεράστιων – παρατηρούμενων συνόλων δεδομένων, έτσι ώστε να βρεθούν οι μεταξύ τους σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων προκειμένου να προβεί σε αναλύσεις ανάλογα με το ζητούμενο στόχο του» [1].

Ένας δεύτερος ορισμός της διαδικασίας εξόρυξης γνώσης είναι ο εξής: *«Η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, προηγουμένως άγνωστης και δυνητικά ωφέλιμης, γνώσης από δεδομένα».*

Εναλλακτικά, συναντάται και ως *η επιστήμη της εξόρυξης χρήσιμης πληροφορίας από σύνολα ή βάσεις δεδομένων μεγάλου μεγέθους»* [2].

Η δήλωση των σχέσεων και η σύνοψη των στοιχείων στην οποία αναφέρεται ο πρώτος ορισμός, συχνά αναφέρεται ως μοντέλο ή πρότυπο. Βασικοί στόχοι της εξόρυξης δεδομένων είναι η περιγραφή και η πρόβλεψη. Δηλαδή, η αναγνώριση των προτύπων που επικρατούν σε ένα μεγάλο σύνολο δεδομένων και η δημιουργία προβλέψεων όσον αφορά τη μελλοντική αξία ή συμπεριφορά κάποιων μεταβλητών. Η αναγνώριση των προτύπων γίνεται μέσω γραμμικών εξισώσεων, κανόνων, διάκρισης σε συστάδες, απόδοσης γραφημάτων και δομών σε μορφή δέντρου, καθώς και επαναλαμβανόμενων προτύπων σε μορφή χρονοσειρών [1].

Αναφορικά με τη διαχείριση επιχειρηματικών πόρων (ERP), η εξόρυξη δεδομένων θεωρείται ως: *«η στατιστική και λογική ανάλυση εκτεταμένων συνόλων από δεδομένα συναλλαγών και εργασιών για τον εντοπισμό επαναλαμβανόμενων μοτίβων ή τάσεων που μπορούν να βοηθήσουν στη λήψη αποφάσεων»* [6].

Πρακτικά όμως τι σημαίνουν όλα αυτά; Αυτό μπορούμε να το απαντήσουμε με ένα απλό παράδειγμα.

Μία ναυτιλιακή εταιρία, αποφασίζει να εντάξει στα δρομολόγια των πλοίων της, δρομολόγια για τις άγονες γραμμές, καθώς το κράτος δίνει επιχορήγηση σε όποια εταιρία αναλάβει αυτά τα δρομολόγια. Έτσι ενισχύει τον στόλο της με επιπλέον πλοία, αποκλειστικά για αυτές τις μετακινήσεις, και επιπλέον προσωπικό. Δυστυχώς όμως για την εταιρία, τις περισσότερες από τις χειμερινές ημέρες υπάρχουν ισχυρότατοι άνεμοι που καθιστούν τον απόπλου των πλοίων απαγορευτικό. Τις ημέρες δε που ο καιρός επιτρέπει τον απόπλου, οι επιβάτες είναι λιγοστοί και στην πραγματικότητα το μόνο κέρδος της είναι από τις μεταφορές των αγαθών που μεταφέρει. Εκτός από το μηδαμινό κέρδος της εταιρίας και το κόστος για τα πλοία – προσωπικό, αναγκάζεται και πληρώνει αυτή χρήματα στο κράτος, καθώς δεν μπορεί να πραγματοποιήσει τα προκαθορισμένα – συμφωνημένα δρομολόγια. Αυτό έχει ως αποτέλεσμα η εταιρία να έχει ζημιές που διογκώνονται αντί για κέρδη.

Στην ίδια περίπου θέση βρίσκεται και ο κάθε επιχειρηματίας όταν επενδύει σε μια βάση δεδομένων, η οποία αποθηκεύει όλη τη δυνατή για την επιχείρησή του πληροφορία. Γνωρίζει καλά, πως είναι προφανώς χρήσιμο να διαθέτει και να διατηρεί αυτή την πληροφορία, όμως αυτό που δεν κατέχει είναι το πώς από τον αχανή όγκο δεδομένων μπορεί, αρχικά, να προκύψει χρήσιμη γνώση και στη συνέχεια με ποιό τρόπο αυτή τη γνώση μπορεί να οδηγήσει σε σημαντικά επιχειρηματικά οφέλη.

Η εξόρυξη δεδομένων βοηθά ακριβώς σε αυτό. Διαμορφώνει τις συνθήκες και τα στοιχεία από την βάση δεδομένων με σκοπό αυτός ο αχανής όγκος δεδομένων να ‘μεταφραστεί’ σε χρήσιμη πληροφορία η οποία θα δώσει κατευθύνσεις για διορθωτικές κινήσεις από το αρμόδιο τμήμα της επιχείρησης. Αν λοιπών απ τη συλλογή δεδομένων καιρικών φαινομένων είχε γίνει επεξεργασία θα μπορούσε η εταιρία να είχε προβεί σε άλλου είδους κινήσεις. Τέτοιες κινήσεις θα μπορούσαν να είναι είτε η χρήση των υπαρχόντων σκαφών για τα δρομολόγια αυτά και να γίνει επαναδιαπραγμάτευση της σύμβασης, είτε να μην αναληφθεί καθόλου η σύμβαση. Σε οποιαδήποτε περίπτωση όμως, αν

εκμεταλλευτεί κανείς έξυπνα τα δεδομένα και σωστά, μπορεί είτε αν αποκομίσει σημαντικά κέρδη, είτε να μην έχει ζημιά.

2.2 Χαρακτηριστικά επιτυχημένης εξόρυξης δεδομένων

Προκειμένου να πραγματοποιήσουμε αποτελεσματικά εξόρυξη δεδομένων, χρειάζεται να εξετάσουμε πρώτα τι χαρακτηριστικά πρέπει να έχει ένα τέτοιο σύστημα στην πράξη και τι προκλήσεις / ιδιαιτερότητες μπορεί να αντιμετωπίσει κάποιος κατά την ανάπτυξη τεχνικών εξόρυξης δεδομένων.

1. Διαχείριση διαφορετικών ειδών δεδομένων

Επειδή υπάρχουν πολλά είδη δεδομένων και βάσεων που χρησιμοποιούνται σε διάφορες εφαρμογές, κάποιος θα περίμενε ένα σύστημα ανακάλυψης γνώσης να είναι σε θέση να πραγματοποιεί αποτελεσματικά εξόρυξη δεδομένων σε διαφορετικά είδη δεδομένων. Αφού οι περισσότερες διαθέσιμες βάσεις είναι σχεσιακές, είναι κρίσιμο ένα σύστημα εξόρυξης δεδομένων να μπορεί να εργάζεται τόσο αποτελεσματικά όσο και αποδοτικά σε σχεσιακά δεδομένα. Επιπρόσθετα, πολλές βάσεις περιέχουν σύνθετους τύπους δεδομένων, όπως δομημένα δεδομένα και πολύπλοκα αντικείμενα, υπερκείμενο και πολυμεσικά δεδομένα, χωρικά, χρονικά και χώρο-χρονικά δεδομένα, δεδομένα συναλλαγών, οικονομικά δεδομένα κλπ. Έτσι ένα δυνατό σύστημα θα έπρεπε να είναι σε θέση να πραγματοποιεί αποτελεσματικά εξόρυξη δεδομένων σε τέτοια πολύπλοκα δεδομένα. Παρ' όλα αυτά, η ανομοιομορφία των τύπων δεδομένων και των διαφορετικών στόχων της εξόρυξης δεδομένων κάνουν μη ρεαλιστική την επιδίωξη και την προσδοκία για ένα ενιαίο σύστημα το οποίο θα μπορεί να διαχειρίζεται όλων των ειδών τα δεδομένα. [7]

2. Αποτελεσματικότητα και κλιμάκωση (scalability) των αλγορίθμων εξόρυξης δεδομένων

Προκειμένου να εξαγάγουμε αποτελεσματικά πληροφορίες και γνώση από τεράστιες ποσότητες δεδομένων, οι χρησιμοποιούμενοι αλγόριθμοι πρέπει να είναι αποτελεσματικοί και κυμαινόμενοι ως προς το χρόνο εκτέλεσης αναλόγως του μεγέθους των δεδομένων που εξεργάζονται . Αυτό σημαίνει, ότι ο χρόνος εκτέλεσης ενός αλγορίθμου εξόρυξης δεδομένων πρέπει να είναι προβλέψιμος και αποδεκτός σε μεγάλες βάσεις δεδομένων. Αλγόριθμοι με εκθετική ή ακόμα και μεσαίας τάξης πολυωνυμική πολυπλοκότητα δεν έχουν πρακτική χρήση [13].

3. Χρησιμότητα, βεβαιότητα και εκφραστικότητα αποτελεσμάτων εξόρυξης δεδομένων

Η γνώση που ανακαλύπτουμε θα πρέπει να απεικονίζει αποτελεσματικά τα περιεχόμενα της βάσης και να είναι χρήσιμη για συγκεκριμένες εφαρμογές. Οι ατέλειες θα πρέπει να εκφράζονται με διαφόρους τρόπους αλλά και μετρικές αβεβαιότητας, όπως για παράδειγμα με χρήση κατά προσέγγιση κανόνων ή ποσοτικών κανόνων. Ο θόρυβος και ειδικά δεδομένα θα πρέπει να διαχειρίζονται ορθά από τα συστήματα εξόρυξης δεδομένων. Αυτό επίσης μας παρακινεί για μια συστηματική μελέτη της ποιότητας της αποκαλυπτόμενης γνώσης, συμπεριλαμβανομένου του ενδιαφέροντος και της αξιοπιστίας, δημιουργώντας αναλυτικά μοντέλα και στατιστικά εργαλεία προσομοίωσης [13].

4. Διατύπωση διαφόρων ειδών αποτελεσμάτων εξόρυξης δεδομένων

Διάφορα είδη γνώσης μπορούν να αποκαλυφτούν από μεγάλες ποσότητες δεδομένων. Επίσης, κάποιος μπορεί να θέλει να εξετάσει την γνώση που έχει βρει από διαφορετικές απόψεις και να την παρουσιάσει με διαφορετικούς τρόπους. Αυτό απαιτεί να εκφράσουμε τόσο τις ανάγκες των απαιτήσεων της ίδιας της εξόρυξης δεδομένων όσο και την αποκαλυπτόμενη γνώση με γλώσσες υψηλού επιπέδου και γραφικά περιβάλλοντα έτσι ώστε το έργο του εκάστοτε χρήστη να μπορεί να προδιαγραφεί και από μη ειδικούς και η γνώση που ανακαλύπτουμε να είναι κατανοητή και άμεσα χρησιμοποιήσιμη από τους χρήστες. Αυτό επίσης προϋποθέτει το σύστημα ανακάλυψης γνώσης να υιοθετεί εκφραστικές τεχνικές αναπαράστασης γνώσης [13].

5. Ανακάλυψη γνώσης από διαφορετικές πηγές δεδομένων

Τα ευρέως διαδεδομένα τοπικά και ευρεία δίκτυα υπολογιστών, συμπεριλαμβανομένου του παγκόσμιου ιστού, συνδέουν πολλές πηγές δεδομένων από τεράστιες κατανεμημένες και ανομοιογενείς βάσεις. Η ανακάλυψη γνώσης από διαφορετικές πηγές μορφοποιημένων και μη δεδομένων με διαφορετικές υποστάσεις και έννοιες θέτει νέες προκλήσεις στην εξόρυξη δεδομένων. Από την άλλη, η εξόρυξη δεδομένων μπορεί να βοηθήσει στην αποκάλυψη των υψηλού επιπέδου ομοιοτήτων σε ανομοιογενείς βάσεις οι οποίες πολύ δύσκολα μπορούν να αποκαλυφθούν από απλά συστήματα δημιουργίας ερωτημάτων (query systems). Επιπρόσθετα, το τεράστιο μέγεθος της βάσης, ο μεγάλος καταμερισμός των δεδομένων, και η υπολογιστική πολυπλοκότητα ορισμένων μεθόδων εξόρυξης δεδομένων δρουν ως κίνητρα για την ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων εξόρυξης δεδομένων [13].

6. Διαφύλαξη της ιδιωτικότητας και της ασφάλειας των δεδομένων

Όταν τα δεδομένα μπορούν να εξετάζονται από διαφορετικές σκοπιές και σε διαφορετικά επίπεδα, διακυβεύεται ο σκοπός της προστασίας της ασφάλειας των δεδομένων και η διαφύλαξη τους από παραβίαση της ιδιωτικής ζωής. Είναι σημαντικό να εξετάσουμε πότε η ανακάλυψη γνώσης μπορεί να οδηγήσει σε τέτοιες περιπτώσεις, και τι μέτρα ασφαλείας μπορούν και θα πρέπει να προβλεφθούν και να υλοποιηθούν για την αποτροπή τυχόν αποκάλυψης ευαίσθητων πληροφοριών. Σημειωτέον ορισμένες από αυτές τις απαιτήσεις μπορεί να δημιουργούν αντικρουόμενους στόχους [41].

2.3 Τύποι και Κλίμακα Χαρακτηριστικών

Μέχρι σήμερα έχουν αναπτυχθεί, και συνεχίζουν να αναπτύσσονται, τεχνικές εξόρυξης δεδομένων, ενσωματώνοντας μεθόδους από διάφορα ερευνητικά πεδία (Μηχανική Μάθηση, Στατιστικά μοντέλα, Επαγωγή Κανόνων, Γραφική Οπτικοποίηση). Σύμφωνα με το είδος της γνώσης που εξάγεται και με τις ιδιαιτερότητες του προβλήματος που καλείται να αντιμετωπίσει, η εξόρυξη

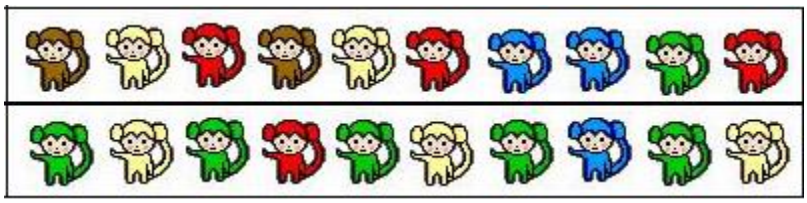
δεδομένων περιλαμβάνει διάφορες λειτουργίες, με πιο γνωστές την ταξινόμηση (classification), την ομαδοποίηση (clustering) και τη συσχέτιση (association). Προτού όμως προβούμε στην ανάλυση αυτών των λειτουργιών κρίνεται σκόπιμο να εξηγήσουμε στο μελετητή τι είδους κατηγορίες και σύνολα δεδομένων επεξεργαζόμαστε με τη λειτουργία της εξόρυξης δεδομένων προκειμένου να εξαγάγουμε σαφή και χρήσιμη γνώση και με ποιες διαδικασίες συλλέγονται αυτά [16].

Τα δεδομένα χαρακτηρίζονται από τον τύπο (type) και την κλίμακά (scale) τους. Επομένως, ένα χαρακτηριστικό μπορεί να λάβει τιμές είτε από ένα συνεχές διάστημα (υποσύνολο του χώρου R) και το αποκαλούμε συνεχές, είτε από ένα πεπερασμένο διακριτό σύνολο, οπότε καλείται διακριτό (discrete) ή πολλαπλών τιμών (multi- valued). Αν επιπλέον, το πεπερασμένο διακριτό σύνολο έχει δύο στοιχεία, τότε η μεταβλητή καλείται δυαδική (binary) ή διχοτόμος. Τα χαρακτηριστικά κατηγοριοποιούνται ανάλογα με τη κλίμακά τους [44][45][46]. Συγκεκριμένα, υπάρχουν τέσσερις κλίμακες χαρακτηριστικών:

Κατηγορικά δεδομένα (ordinal data / nominal data)

Ένα σύνολο δεδομένων ορίζεται ότι είναι κατηγορικό εάν οι τιμές ή οι παρατηρήσεις που ανήκουν σε αυτό, μπορούν να ταξινομηθούν σύμφωνα με χρήση κατηγοριών. Οι κατηγορίες πρέπει να επιλεχτούν προσεκτικά δεδομένου ότι μια κακή επιλογή μπορεί να οδηγήσει σε λάθος αποτελέσματα την έκβαση μιας έρευνας. Κάθε μεταβλητή πρέπει να ανήκει σε μια και μόνο μια κατηγορία και δεν πρέπει να υπάρχει αμφιβολία ως προς ποια. Παραδείγματα κατηγοριών που μπορούν να χρησιμοποιηθούν για τη μελέτη τέτοιου είδους δεδομένων είναι η ηλικία, το φύλο, το χρώμα. Τα κατηγορικά δεδομένα συνοψίζονται συχνά με την υποβολή εκθέσεων του ποσοστού των μεταβλητών που εμπίπτουν σε κάθε κατηγορία [16]. Ακολουθεί ένα απλό παράδειγμα.

Ο ακόλουθος πίνακας παρουσιάζει διαφορετικά χρωματισμένα παιχνίδια που τοποθετούνται σε δύο διαφορετικά ράφια.



Αυτά τα παιχνίδια μπορούν να ταξινομηθούν σύμφωνα με το χρώμα τους στις διαφορετικές κατηγορίες:

Χρώμα	**Αριθμός Παιχνιδιών**
Καφέ	2
Κίτρινο	5
Κόκκινο	4
Μπλε	3
Πράσινο	6

Αριθμητικά δεδομένα (Ratio-Scaled)

Τα αριθμητικά δεδομένα είναι δεδομένα που μπορούν να παρασταθούν μέσω πραγματικών και ακεραίων αριθμών. Τα αριθμητικά δεδομένα μπορούν να αναλυθούν χρησιμοποιώντας στατιστικές μεθόδους και τα αποτελέσματα μπορούν να επιδειχθούν χρησιμοποιώντας πίνακες, γραφήματα και διαγράμματα. Αν ο λόγος μεταξύ δύο τιμών ενός χαρακτηριστικού έχει νόημα τότε το χαρακτηριστικό ανήκει στη κατηγορία αυτή. Ένα παράδειγμα είναι το βάρος, αφού έχει νόημα να ειπωθεί πως ένα άτομο που ζυγίζει 80kgείναι δυο φορές πιο βαρύ από ένα άτομο που ζυγίζει 40kg.

Interval-Scaled

Αν για ένα χαρακτηριστικό, η διαφορά μεταξύ δύο τιμών έχει νόημα, ενώ ο λόγος τους δεν έχει, τότε το χαρακτηριστικό ανήκει στην κατηγορία αυτή. Ένα τυπικό παράδειγμα είναι η μέτρηση της θερμοκρασίας σε βαθμούς Κελσίου. Αν η θερμοκρασία στην Αθήνα και τη Κόρινθο είναι 10 και 20 βαθμούς αντίστοιχα, τότε το συμπέρασμα πως η θερμοκρασία στην Κόρινθο είναι 10 βαθμούς υψηλότερη απ την Αθήνα έχει νόημα. Παρ όλα αυτά δεν έχει νόημα να ειπωθεί πως η Κόρινθος είναι δύο φορές πιο ζεστή από την Αθήνα.

Είναι αξιοσημείωτο ότι κάθε κλίμακα περιέχει όλες τις ιδιότητες της προηγούμενης της. Δηλαδή, ένα ratio-scaledχαρακτηριστικό περιλαμβάνει όλες τις ιδιότητες των interval-scaled, ordinalκαι nominalχαρακτηριστικών. Ομοίως και ένα interval-scaledχαρακτηριστικό περιέχει τις ιδιότητες των ordinalκαι nominalχαρακτηριστικών και ούτω καθεξής.

2.4 Μέτρα εγγύτητας

Η επιλογή και ο υπολογισμός του μέτρου εγγύτητας (proximity measure) είναι πολύ βασικός για την διαδικασία ομαδοποίησης(θα αναφερθούμε στη συγκεκριμένη διαδικασία παρακάτω). Η επιλογή του μέτρου εγγύτητας είναι ένα ιδιαίτερα ενδιαφέρον πρόβλημα της ομαδοποίησης. Είναι εξαιρετικά σημαντικό ποια μέθοδο θα διαλέξουμε για να συγκρίνουμε τα δεδομένα μας. Στην συγκεκριμένη ενότητα θα δοθεί ορισμός για τα μέτρα εγγύτητας και πως αυτά ορίζονται, επίσης θα περιγράψουμε μερικά υπάρχοντα μέτρα εγγύτητας [14].

Στην πλειοψηφία τους οι τεχνικές ομαδοποίησης αρχίζουν με τον υπολογισμό ενός τετραγωνικού πίνακα ομοιότητας (similarity matrix) μεταξύ των οντοτήτων. Αυτός ο συμμετρικός πίνακας μας δίνει την ομοιότητα ή την ανομοιότητα μεταξύ όλων των οντοτήτων που λαμβάνουν μέρος στην ομαδοποίηση. Πολλές μέθοδοι ομαδοποίησης μπορούν να θεωρηθούν ως προσπάθειες να συνοψιστεί η πληροφορία που αναφέρεται στις σχέσεις μεταξύ των οντοτήτων και η οποία περιέχεται στον πίνακα ομοιότητας, έτσι ώστε οι σχέσεις αυτές να μπορούν να γίνουν ευκολότερα κατανοητές και να μπορούν να ερμηνευτούν πιο εύκολα. Είναι φανερό ότι η έξοδος ενός αλγορίθμου ομαδοποίησης,

οι ομάδες δηλαδή στις οποίες έχουν χωριστεί τα δεδομένα, θα είναι τόσης σημασίας όσης είναι και οι ομοιότητες και οι αποστάσεις εισόδου που δίνονται από τον πίνακα ομοιότητας [14].

2.4.1 Μέτρα ομοιότητας - Μέτρα ανομοιότητας

Τα μέτρα εγγύτητας χωρίζονται σε μέτρα ομοιότητας (similarity measures) και μέτρα ανομοιότητας (dissimilarity measures). Παρακάτω θα δώσουμε τους ορισμούς των μέτρων ομοιότητας και ανομοιότητας[15]. Δοθέντος λοιπόν ενός συνόλου δεδομένων X:

Ένα **μέτρο ομοιότητας** s στο X είναι μια συνάρτηση : X × X →ℜ όπου ℜ είναι το σύνολο των πραγματικών αριθμών, έτσι ώστε

$$\exists s_o \in \Re: -\infty < s(x,y) \leq s_o < +\infty, \qquad \forall x,y \in X$$

$$s(x,x) = s_o \,, \forall x \in X$$

και

$$s(x,y) = s(y,x), \forall x,y \in X \,.$$

Αν επιπλέον ισχύουν:

$$s(x,y) = s_o \quad \text{αν και μόνο αν} \quad x = y$$

και

$$s(x,y)s(y,z) \leq [s(x,y) + s(y,z)]s(y,z), \qquad \forall x,y,z \in X$$

τότε το μέτρο ομοιότητας s είναι μια μετρική (metric similarity measure).

Ένα **μέτρο ανομοιότητας** d στο X είναι μια συνάρτηση

$$d : X \times X \to \Re$$

έτσι ώστε

$$\exists\, d_o \in \Re: -\infty < d_o \leq d(x,y) < +\infty, \qquad \forall x,y \in X$$

$$d(x,x) = d_o$$

και

$$d(x,y) \leq d(x,y) + d(y,z), \qquad \forall x,y,z \in X$$

τότε το μέτρο ανομοιότητας d είναι μια μετρική (metric dissimilarity measure)

Από τα παραπάνω μπορούμε να βγάλουμε τα εξής συμπεράσματα:

• Τα μέτρα ομοιότητας μπορούν να λάβουν θετικές αλλά και αρνητικές τιμές.

• Η μέγιστη τιμή της ομοιότητας μεταξύ δυο διανυσμάτων του X επιτυγχάνεται όταν αυτά ταυτίζονται.

• Η ελάχιστη τιμή ανομοιότητας μεταξύ δυο διανυσμάτων του X επιτυγχάνεται όταν αυτά ταυτίζονται.

Γενικά θα μπορούσαμε να πούμε ότι τα μέτρα ομοιότητας είναι αντίθετα από τα μέτρα ανομοιότητας. Εύκολα μπορούμε να αποδείξουμε ότι αν το μέτρο ανομοιότητας d είναι μια μετρική, με $d(x,y) > 0, \forall x,y \in X$ τότε το μέτρο ομοιότητας $s = \frac{a}{d} \mu\varepsilon\ \alpha > 0$ είναι και αυτό μετρική. Επίσης εύκολα μπορούμε να αποδείξουμε ότι το μέτρο ομοιότητας $dmax - d$ είναι μια μετρική, όπου $dmax$ συμβολίζει την μέγιστη τιμή του d ανάμεσα σε όλα τα ζεύγη σημείων του X [15].

Θα μπορούσαμε να ομαδοποιήσουμε τα μέτρα εγγύτητας σε τέσσερις μεγάλες κατηγορίες:

1. Μέτρα απόστασης (Distance Measures).

2. Συντελεστές Σχέσης (Association Coefficients).

3. Συντελεστές Συσχέτισης (Correlation Coefficients).

Θα αναλύσουμε λίγο περισσότερο τα διαφορετικά είδη των μέτρων εγγύτητας παρακάτω.

2.4.2 Μέτρα απόστασης

Τα μέτρα απόστασης (distance measures) είναι πολύ διαδεδομένα και μπορούν να διακριθούν σε μέτρα ομοιότητας και μέτρα ανομοιότητας. Μια άλλη κατηγοριοποίηση σύμφωνα με το είδος των τιμών που συγκρίνουν είναι σε μέτρα πραγματικών τιμών και μέτρα διακριτών τιμών [16], [55].

Έστω X ο μονοδιάστατος χώρος των δεδομένων μας. Δυο οντότητες (περιπτώσεις) είναι ταυτόσημες όταν κάθε μια περιγράφεται μέσω μεταβλητών με την ίδια σημαντικότητα. Σ' αυτήν την περίπτωση το μέτρο ανομοιότητας είναι 0. Όταν το μέτρο ανομοιότητας έχει μικρή τιμή τότε οι οντότητές μας έχουν υψηλό βαθμό συσχέτισης και αντίστροφα όταν το μέτρο ανομοιότητας έχει μεγάλη τιμή η συσχέτιση των οντοτήτων είναι μικρότερη. Ακριβώς το αντίθετο συμβαίνει στα μέτρα ομοιότητας [16], [55].

2.4.3 Μέτρα Ανομοιότητας πραγματικών τιμών

Θα αναφερθούμε τώρα στα πιο γνωστά και στα πιο ευρέως χρησιμοποιούμενα μέτρα ανομοιότητας πραγματικών τιμών.

• Διατιμημένη (weighted) lp μετρική.

Η διατιμημένη lp μετρική δίνεται από τον τύπο:

$$d_p(x,y) = \left(\sum_{i=1}^{l} w_i \, |x_i - y_i|^p\right)^{1/p}$$

όπου xi και yi είναι η i −οστή συντεταγμένη των x και y με $i = 1,\dots,l$ και $wi \geq 0$ είναι ο i −οστός συντελεστής βάρους. Αν $wi = 1$ τότε η παραπάνω η παραπάνω μετρική ονομάζεται μη διατιμημένη μετρική lp ή μετρική Minkofski.

• Όταν το p λαμβάνει την τιμή 2 τότε η παραπάνω μετρική καλείται Ευκλείδεια απόσταση ή l2 μετρική.

$$d_2(x,y) = \left(\sum_{i=1}^{l} |x_i - y_i|^2\right)^{1/2}$$

Για την αποφυγή της τετραγωνικής ρίζας, η τιμή της απόστασης τετραγωνοποιείται και συμβολίζεται ως d^2.Η έκφραση αυτή αναφέρεται και ως Τετραγωνική Ευκλείδεια Απόσταση. Η διατιμημένη l2 μετρική μπορεί να γενικευτεί με τον ακόλουθο τρόπο:

$$d(x,y) = \sqrt{(x-y)^T B^T (x-y)}$$

όπου B είναι ένας συμμετρικός θετικά ορισμένος πίνακας. Όταν B=Σ^{-1}, όπου Σ είναι ο εντός των ομάδων πίνακας διασποράς / συνδιασποράς (pooled within /groups variance / co-variance), τότε το μέτρο αυτό καλείται Mahalanobis μετρική. Το μέτρο αυτό έχει το πλεονέκτημα έναντι των άλλων μέτρων, ότι επιτρέπει τις συσχετίσεις μεταξύ μεταβλητών. Όταν οι συσχετίσεις αυτές είναι μηδενικές τότε το μέτρο αυτό είναι ισοδύναμο με την τετραγωνική ευκλείδεια απόσταση [16],[17],[55].

• Όταν το p λάβει την τιμή 1 τότε λαμβάνεται η l_1διατιμημένη μετρική ή η

Manhattan (City - Block) μετρική.

$$d_1(x,y) = \left(\sum_{i=1}^{l} w_i \, |x_i - y_i|\right)$$

• Όταν το p λάβει την τιμή ∞ τότε λαμβάνεται η $l_{(\infty)}$ διατιμημένη μετρική:

$$d_\infty(x,y) = \max_{1 \leq i \leq l} w_i |x_i - y_i|$$

• Δυο άλλα μέτρα ανομοιότητας πραγματικών τιμών είναι τα ακόλουθα:

$$d_G(x,y) = -\log_{10}\left(1 - \frac{1}{l}\sum_{j=1}^{l}\frac{|x_j - y_j|}{b_j - a_j}\right)$$

όπου b_jκαι a_jείναι η μέγιστη και η ελάχιστη τιμή του j–οστού χαρακτηριστικού στα N διανύσματα του X, αντίστοιχα. Αποδεικνύεται ότι το μέτρο αυτό είναι μια μετρική. Αξίζει να σημειωθεί πως η τιμή του $d_g(x,y)$δεν εξαρτάται μόνο από τα δυο διανύσματα x και y αλλά από όλο το X. Έτσι αν $d_g(x,y)$είναι η απόσταση μεταξύ των δύο διανυσμάτων x και y στο X καιαν$d'_G\ (x,y)$είναι η απόσταση μεταξύ αυτών των ίδιων δυο διανυσμάτων στο σύνολο X, τότε εν γένει:

$$d_G(x,y) \neq d'_G(x,y)$$

Ένα άλλο μέτρο ανομοιότητας είναι:

$$d_Q(x,y) = \sqrt{\frac{1}{l}\sum_{j=1}^{l}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2}$$

Στο σημείο αυτό θα πρέπει ίσως να κάνουμε ένα μικρό σχόλιο για το πώς η Ευκλείδεια απόσταση επηρεάζεται αρνητικά από την αλλαγή στην κλίμακα των μεταβλητών. Ας δώσουμε ένα μικρό παράδειγμα. Έστω ότι οι τρεις περιπτώσεις A, B, C μετριούνται σε δυο μεταβλητές το βάρος σε λίβρες και το ύψος σε πόδια με τα ακόλουθα αποτελέσματα:

	Βάρος σε λίβρες	Ύψος σε πόδια
A	60	3.0
B	65	3.5
C	63	4.0

Οι Ευκλείδειες τους αποστάσεις είναι:

D^2_{AB}	25.25
D^2_{AC}	10.00
D^2_{BC}	04.25

Συνεπώς η περίπτωση A είναι στην δεύτερη φορά πιο κοντά στην περίπτωση B παρά στην C. Επιπλέον η ευκλείδεια απόσταση δεν διατηρεί ούτε και τις ταξινομήσεις απόστασης. Εξαιτίας αυτού, οι μεταβλητές συχνά τυποποιούνται πριν εφαρμοστεί η ευκλείδεια απόσταση, δηλαδή $Z_{ik} = \frac{x_{ik}}{\sigma_k}$, που σ_k

είναι η τυπική απόκλιση της κ – οστής μεταβλητής. Με τον τρόπο αυτό η ευκλείδεια μετρική διατηρεί τις σχετικές αποστάσεις [17].

2.4.4 Μέτρα ομοιότητας Πραγματικών τιμών

Θα ασχοληθούμε πολύ σύντομα, στο όνομα της πληρότητας, με μερικά μέτρα ομοιότητας μεταξύ πραγματικών τιμών.

• Εσωτερικό γινόμενο. Ορίζεται ως:

$$S_{inner}(x, y) = x^T y = \sum_{i=1}^{l} x_i \; yi$$

Στις περισσότερες περιπτώσεις το εσωτερικό γινόμενο χρησιμοποιείται όταν τα διανύσματα x και y είναι κανονικοποιημένα, έτσι ώστε να έχουν το ίδιο μήκος α. Στις περιπτώσεις αυτές το άνω και το κάτω φράγμα του εσωτερικού γινομένου είναι $+\alpha^2$ και $-\alpha^2$ αντίστοιχα και το εσωτερικό γινόμενο εξαρτάται αποκλειστικά από τη γωνία μεταξύ των διανυσμάτων.

• Μέτρα Tanimoto. Αυτό το μέτρο μπορεί να χρησιμοποιηθεί τόσο σε πραγματικά όσο και σε διακριτά διανύσματα. Ορίζεται ως εξής:

$$S_T(x, y) = \frac{x^T y}{\|x\|^2 + \|y\|^2 - x^T y}$$

Προσθέτοντας και αφαιρώντας τον όρο $x^T y$ από τον παρονομαστή της προηγούμενης συνάρτησης και μετά από μερικές αλγεβρικές πράξεις παίρνουμε εξής συνάρτηση:

$$S_T(x, y) = \frac{1}{1 + \frac{(x-y)^T(x-y)}{x^T y}}$$

• Ένα άλλο μέτρο ομοιότητας χρήσιμο σε συγκεκριμένες εφαρμογές δίνεται:

$$S_x(x, y) = 1 - \frac{d_2(x, y)}{\|x\| + \|y\|}$$

όπου το s (x, y) c παίρνει την μέγιστη τιμή του 1 όταν x = y και την ελάχιστη 0 όταν x = -y [17], [55].

2.4.5 Μέτρα ανομοιότητας Διακριτών τιμών

Έστω ότι οι συντεταγμένες των διανυσμάτων x ανήκουν σε ένα πεπερασμένο σύνολο $F = \{0, 1, \dots, k-1\}$, όπου k είναι ένα θετικός ακέραιος. Ας υποθέσουμε τώρα ότι η διάσταση των δεδομένων μας είναι l, τότε είναι φανερό πως υπάρχουν k^lδιανύσματα $x \; €f^l$. Θα μπορούσε κάποιος να θεωρήσει ότι τα διανύσματα αυτά είναι κορυφές ενός l–διάστατου πλέγματος [17], [55].

Έστω ότι $x \in f^l$ ο πίνακας:

$$A(x,y) = [a_{ij}] \qquad i,j = 0,1,\dots,k-1,$$

είναι ένας k×k πίνακας όπου το στοιχείο a_{ij} είναι το πλήθος των θέσεων όπου το πρώτο διάνυσμα έχει το σύμβολο i και το αντίστοιχο στοιχείο του δεύτερου διανύσματος έχει το σύμβολο j, όπου $i, j \in F$ τότε ο πίνακας αυτός καλείται contingency matrix. Για την καλύτερη κατανόηση του παραπάνω ορισμού ας δώσουμε ένα παράδειγμα [17].

Παράδειγμα:

Έστω $l = 6$, $k = 3$, και $x = [0,1,2,1,2,1]^T$ $y = [1,0,2,1,0,1]^T$ τότε ο πίνακας $A(x,y)$ είναι ο πίνακας:

$$A(x,y) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Διότι F = {0,1,2} και το στοιχείο a_{ij} του πίνακα ισούται με το πόσες φορές εμφανίζεται σε αντίστοιχες θέσεις στα διανύσματα x και y κάποιο συγκεκριμένο ζεύγος τιμών του συνόλου F. Παραδείγματος χάριν αν i = 1 και j = 1 τότε το $a_{1\,1}$ θα ισούται με 2 γιατί το ζεύγος τιμών (1,1) εμφανίζεται σε αντίστοιχες θέσεις στα διανύσματα x και y δύο φορές. Μια φορά στην θέση 4 και μια φορά στη θέση 6. Ας σημειωθεί εδώ ότι οι δείκτες του πίνακα A(x,y) αρχίζουν από το (1,0) στα δύο διανύσματα x και y αντίστοιχα. Το (0,1) εμφανίζεται μια φορά.

• **Απόσταση Hamming**. Η απόσταση Hamming ορίζεται ως το πλήθος των θέσεων όπου τα δυο διανύσματα διαφέρουν. Χρησιμοποιώντας τον πίνακα A που ορίσαμε λίγο παραπάνω μπορούμε να ορίσουμε την απόσταση Hamming πιο τυπικά ως εξής [17]:

$$d_H(x,y) = \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ j \neq i}}^{k-1} a_{ij}$$

Δηλαδή η απόσταση Hamming παριστάνει το άθροισμα όλων των στοιχείων του πίνακα A, πλην αυτών της κύριας διαγωνίου. Αυτό προφανώς το άθροισμα παριστάνει σε πόσα σημεία διαφέρουν τα δυο διανύσματα. Πράγματι το άθροισμα των στοιχείων της κύριας διαγωνίου του πίνακα A μας δίνει το πλήθος των αντίστοιχων θέσεων που τα δυο διανύσματα συμφωνούν. Συνεπώς το άθροισμα όλων των άλλων στοιχείων μας δίνει το πλήθος των θέσεων που υπάρχουν μη κοινά στοιχεία στα δυο διανύσματα [17].

• Η απόσταση l1. Μια άλλη απόσταση η οποία μπορεί να χρησιμοποιηθεί σε διακριτά διανύσματα είναι η απόσταση l1 η οποία έχει οριστεί παραπάνω.

Η απόσταση l1 και η απόσταση Hamming ταυτίζονται όταν τα δεδομένα έχουν δυαδικές τιμές. Είναι προφανές ότι τα μέτρα ανομοιότητας διακριτών τιμών μπορούν να χρησιμοποιηθούν άμεσα σαν μέτρα κατηγορικών δεδομένων. Δίνοντας μια διακριτή τιμή από το 0 ως το k-1 σε καθεμιά από τις

διαφορετικές τιμές κάθε μεταβλητής μας, μπορούμε να χρησιμοποιήσουμε άμεσα τις παραπάνω αποστάσεις. Η απόσταση Hamming είναι πιο κοντά στην φιλοσοφία των κατηγορικών δεδομένων αφού λαμβάνει υπ' όψιν της μόνο τα σημεία στα οποία διαφέρουν τα δυο διανύσματα χωρίς να ασχολείται με την αριθμητική της αντιστοίχησης των κατηγορικών δεδομένων σε φυσικούς αριθμούς. Αντίθετα η απόσταση l1 το κάνει αυτό και είναι ίσως λιγότερο κοντά στην φιλοσοφία των κατηγορικών δεδομένων [17].

2.4.6 Συντελεστές σχέσης.

Οι συντελεστές Σχέσης (Association Coefficients) χρησιμοποιούνται για να υπολογιστεί η ομοιότητα μεταξύ περιπτώσεων που περιγράφονται από δυαδικές μεταβλητές. Οι συντελεστές σχέσης παίρνουν τιμές μεταξύ 0 και 1. Οι συντελεστές αυτοί μπορούν να περιγραφούν μέσω ενός 2×2 πίνακα σχέσης (association table), στον οποίο το 1 ή το + αναφέρεται στην παρουσία μιας μεταβλητής ενώ το 0 ή το – αναφέρεται στην απουσία μιας μεταβλητής όπως φαίνεται στον παρακάτω πίνακα [17]:

		Περίπτωση i		
		+	-	
Περίπτωση j	+	a	b	a+b
	-	c	d	c+d
		a+c	b+d	p

Έστω δυο l – διάστατα διανύσματα x και y. Τα a, b , c και d αναφέρονται στο πόσες φορές εμφανίζεται το αντίστοιχο πρότυπο των δυαδικών μεταβλητών σ' αυτές τις δυο περιπτώσεις. Παραδείγματος χάριν το a μας δίνει το πλήθος των φορών εμφάνισης στις δυο περιπτώσεις δύο θετικών τιμών. Αντίστοιχα ισχύουν και για τα b, c, και d. Έχουν προταθεί πολλοί συντελεστές σχέσης κυρίως εξαιτίας της αβεβαιότητας στον τρόπο ενσωμάτωσης των αρνητικών ταιριασμάτων (δηλαδή των d του παραπάνω πίνακα) στον συντελεστή, καθώς επίσης και εξαιτίας του τρόπου αξιολόγησης των ταιριασμένων ζευγαριών των μεταβλητών. Είναι δυνατόν αυτά τα ζεύγη να είναι ισοδύναμα διατιμημένα ή και να έχουν διπλάσιο βάρος από ότι τα μη ταιριασμένα ζευγάρια ή και το αντίστροφο. Μερικοί συντελεστές αγνοούν τα αρνητικά ταιριάσματα όπως παραδείγματος χάριν ο συντελεστής (ii) στον παρακάτω πίνακα, ενώ άλλοι δίνουν περισσότερο βάρος στα ταιριασμένα ζευγάρια παρά στα μη ταιριασμένα όπως για παράδειγμα οι συντελεστές (iii) και (iv). Στον ακόλουθο πίνακα δίνουμε μερικούς συντελεστές συσχέτισης για δυαδικά δεδομένα [17],[18],[55]:

(i) $\frac{a+d}{p}$ (ii) $\frac{a}{a+b+c}$

(iii) $\frac{2a}{2a+b+c}$ (iv) $\frac{2(a+d)}{2(a+d)+b+c}$

(v) $\frac{a}{a + 2(b + c)}$ (vi) $\frac{a}{p}$

Ο συντελεστής (i) ονομάζεται Συντελεστής Απλού ταιριάσματος (Simple Matching Coefficient) και η ομοιότητα κυμαίνεται από 0 ως 1. Ο συντελεστής αυτός δεν μετασχηματίζεται εύκολα σε μετρική και λαμβάνει υπ' όψιν του την αρθρωτή απουσία μιας μεταβλητής. Το γεγονός αυτό οδηγεί κάποιες περιπτώσεις να εμφανίζονται πολύ όμοιες εξαιτίας κυρίως της ταυτόχρονης έλλειψης των ίδιων χαρακτηριστικών παρά εξαιτίας των όμοιων χαρακτηριστικών [1].Ο συντελεστής (ii) καλείται συντελεστής Jaccard και η ομοιότητα που απορρέει απ' αυτόν τον συντελεστή κυμαίνεται από 0 ως 1. Ο συντελεστής αυτός αποφεύγει τη χρήση της αρθρωτής απουσίας μιας μεταβλητής στον υπολογισμό της ομοιότητας [47].

Ας σημειώσουμε εδώ ότι διαφορετικοί συντελεστές σχέσης όταν εφαρμοστούν στο ίδιο σύνολο δεδομένων μπορεί να λάβουν εντελώς διαφορετικές τιμές. Ας δούμε ένα παράδειγμα όπου δυο περιπτώσεις χαρακτηρίζονται από την παρουσία ή την απουσία 10 μεταβλητών, όπως φαίνεται παρακάτω:

Μεταβλητές	1	2	3	4	5	6	7	8	9	10
Περίπτωση1	1	0	0	0	1	1	0	0	1	0
Περίπτωση1	0	0	0	0	1	0	0	1	1	0

Ο 2Χ2 πίνακας σχέσης για αυτές τις δύο περιπτώσεις, είναι ο εξής:

		Περίπτωση 1		
		1	0	
Περίπτωση 2	1	2	1	3
	0	2	5	7
		4	6	10

τότε εφαρμόζοντας τους συντελεστές (i) και (ii) θα λάβουμε:

(i)	0.70	(ii)	0.40
(iii)	0.50	(iv)	0.82
(v)	0.25	(vi)	0.20

Παρατηρούμε πάρα πολύ καθαρά την μεγάλη απόκλιση που έχουμε μεταξύ των διαφορών συντελεστών σχέσης όταν εφαρμόζονται πάνω στο ίδιο ζεύγος δεδομένων. Ο συντελεστής (iv) δίνει ομοιότητα 0.82 ενώ ο συντελεστής (vi) μας δίνει μια ομοιότητα 0.20.

Το γεγονός αυτό δεν θα ήταν τόσο σημαντικό αν όλοι οι συντελεστές ήταν αρθρωτά μονότονοι (jointly monotonic) δηλαδή, αν όλες οι τιμές των ζευγαριών των περιπτώσεων ενός συντελεστή μπορούσαν να ταξινομηθούν κατά τέτοιο τρόπο ώστε να σχηματίζουν μια μονότονη ακολουθία (μια ακολουθία δηλαδή που οι τιμές της είτε θα αυξάνουν είτε θα φθίνουν σ' όλο το μήκος της). Σ' αυτήν την περίπτωση τότε οι τιμές για τα ζεύγη που θα λαμβάνονταν μέσω ενός άλλου συντελεστή θα ήταν επίσης μονότονες. Αυτό όμως δεν ισχύει, γιατί αν υποθέσουμε ότι μια Τρίτη περίπτωση έχει την ακόλουθη μορφή για τις 10 δυαδικές μεταβλητές:

Μεταβλητές	1 2 3 4 5 6 7 8 9 10
Περίπτωση 1	0 0 0 0 0 0 0 1 0 0

Τότε εφαρμόζοντας τους συντελεστές (i) και (ii) θα λάβουμε:

Συντελεστής (i)	Συντελεστής (ii)
S_{12} =0.7	S_{12} =0.4
S_{13} =0.5	S_{13} =0.0
S_{23} =0.8	S_{23} =0.1

Όπως φαίνεται καθαρά οι συντελεστές δεν είναι αρθρωτά μονότονοι [17],[18]:

2.4.7 Συντελεστές συσχέτισης

Οι συντελεστές αυτοί καλούνται μερικές φορές και μέτρα γωνιών (angular measures) εξαιτίας της γεωμετρικής του ερμηνείας. Χρησιμοποιούνται πολύ συχνά στις κοινωνιολογικές επιστήμες. Ο πιο διαδεδομένος συντελεστής είναι ο product – moment συντελεστής συσχέτισης που προτάθηκε από τον Karl Pearson. Αρχικά, χρησιμοποιήθηκε σαν μέθοδος για την συσχέτιση μεταβλητών, αλλά έχει χρησιμοποιηθεί και στην ποσοτική κατηγοριοποίηση για συσχέτιση περιπτώσεων. Ο συντελεστής ορίζεται ως εξής:

$$r_{jk} = \frac{\sum(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum(x_{ij} - \bar{x}_j)^2 (x_{ik} - \bar{x}_k)^2}}$$

Οπου x_{ij}είναι η τιμή της μεταβλητής i για την περίπτωση j , $\bar{x}_j$ η μέση τιμή όλων των τιμών των μεταβλητών για την περίπτωση j [14,16]. Η τιμή του συντελεστή κυμαίνεται από -1 έως 1, το πρόσημο υποδηλώνει την κατεύθυνση της σχέσης. Το θετικό πρόσημο δείχνει πως η σχέση είναι θετική ενώ το αρνητικό πρόσημο μας δείχνει ότι η σχέση είναι αρνητική. Η απόλυτη τιμή του συντελεστή δηλώνει την σημαντικότητα (magnitude) της σχέσης [16].

Ο συντελεστής του Pearson είναι ένας δείκτης για την γραμμική σχέση μεταξύ δύο περιπτώσεων. Η υψηλή συσχέτιση μπορεί να λάβει χώρα μεταξύ των περιπτώσεων εφόσον οι μετρήσεις μιας περίπτωσης είναι σε γραμμική σχέση με την άλλη. Γραμμική σχέση (linear relationship) δεν σημαίνει ότι τα σημεία πέφτουν ακριβώς πάνω σε μια ευθεία γραμμή, αλλά ότι τα σημεία είναι τοποθετημένα γενικά κατά μήκος μιας γραμμής [14,16].

Ένα μειονέκτημα του συντελεστή του Pearson, είναι ότι η χρήση του για τον καθορισμό της συσχέτισης μεταξύ των περιπτώσεων δεν έχει κανένα στατιστικό νόημα, διότι απαιτεί τον υπολογισμό της μέσης τιμής μεταξύ διαφορετικών τύπων μεταβλητών και όχι τον υπολογισμό της μέσης τιμής κάθε μεταβλητής ανά περίπτωση. Ένας άλλος περιορισμός του συντελεστή είναι ότι συχνά αποτυγχάνει να ικανοποιήσει την τριγωνική ανισότητα [14,16].

2.5 Λειτουργίες εξόρυξης δεδομένων

1) Association Rules (**κανόνες συσχετίσεων).** Στην εξόρυξη δεδομένων, οι κανόνες συσχετίσεων χρησιμοποιούνται για να ανακαλύψουν τα στοιχεία που επανεμφανίζονται συχνά μέσα σε ένα σύνολο δεδομένων [19].Αποτελούνται από πολλαπλές ανεξάρτητες επιλογές δεδομένων (όπως η αγορά και οι συναλλαγές) και την ανακάλυψη των κανόνων τους, όπως η επίπτωση ή ο συσχετισμός, οι οποίοι αφορούν τα επανεμφανιζόμενα δεδομένα. Ερωτήσεις τύπου "εάν ένας πελάτης αγοράζει το Προϊόν Α, τι πιθανότητες υπάρχουν να αγοραστεί το Προϊόν Β" και "ποια προϊόντα ένας πελάτης θα αγοράσει ως συμπληρωματικά εάν αγοράσει τα προϊόντα Γ και Δ", απαντώνται με τη χρήση αλγορίθμων εύρεσης συσχετίσεων. Αυτή η εφαρμογή των κανόνων συσχέτισης είναι επίσης γνωστή ως market basket analysis (ανάλυση του καλαθιού αγοράς).Όπως με τις περισσότερες τεχνικές εξόρυξης δεδομένων, ο στόχος είναι να μειωθεί το ενδεχομένως τεράστιο ποσό των πληροφοριών σε ένα μικρό, κατανοητό σύνολο στατιστικά υποστηριγμένων πληροφοριών [2],[19],[20].

2) Cluster Analysis (ανάλυση συστάδων). Η ανάλυση συστάδων είναι μια συλλογή στατιστικών μεθόδων, η οποία προσδιορίζει τις ομάδες δεδομένων που συμπεριφέρονται ομοίως ή παρουσιάζουν παρόμοια χαρακτηριστικά. Η ανάλυση συστάδων είναι μια συλλογή στατιστικών μεθόδων, η οποία προσδιορίζει τις ομάδες δεδομένων που συμπεριφέρονται ομοίως ή παρουσιάζουν παρόμοια χαρακτηριστικά. Η ανάλυση συστάδων δεν είναι ένας συγκεκριμένος αλγόριθμος, αλλά ένας γενικός στόχος προς επίλυση. Μπορεί να επιλυθεί με διάφορους αλγορίθμους που διαφέρουν σημαντικά στην έννοια, του τι είναι συστάδα και πώς να τις βρουν αποτελεσματικά. Οι δημοφιλείς έννοιες των συστάδων περιλαμβάνουν ομάδες με χαμηλές αποστάσεις μεταξύ των μελών των συστάδων, πυκνοί τομείς του διαστήματος δεδομένων και διαστήματα ή ιδιαίτερες στατιστικές διανομές. Οι κατάλληλες τοποθετήσεις αλγορίθμου και παραμέτρου συγκέντρωσης (συμπεριλαμβανομένων των τιμών όπως η συνάρτηση απόστασης και το κατώτατο όριο πυκνότητας ή ο αριθμός αναμενόμενων συστάδων) εξαρτώνται από το μεμονωμένο σύνολο δεδομένων και την προοριζόμενη χρήση των αποτελεσμάτων. Ως εκ τούτου η ανάλυση συστάδων δεν είναι μια αυτόματη διαδικασία αλλά μια επαναληπτική διαδικασία της ανακάλυψης γνώσης που περιλαμβάνει λειτουργία

δοκιμής και αποτυχίας (trial and error). Θα είναι συχνά απαραίτητο να τροποποιηθεί η προεπεξεργασία και οι παράμετροι έως ότου επιτευχθεί το αποτέλεσμα σύμφωνα με τις επιθυμητές ιδιότητες [14],[2],[20].

3) Classification (κατηγοριοποίηση). Η κατηγοριοποίηση είναι μια λειτουργία εξόρυξης δεδομένων που κατατάσσει τα δεδομένα σε σύνολα κατηγοριών ή κλάσεων (classes). Ο στόχος της ταξινόμησης είναι να προβλεφθεί ακριβώς η κατηγορία στόχων για κάθε περίπτωση στα δεδομένα. Παραδείγματος χάριν, ένα πρότυπο ταξινόμησης θα μπορούσε να χρησιμοποιηθεί για να προσδιορίσει τους υποψηφίους δανείου σε κατηγορίες όπως χαμηλούς, μέσους, ή υψηλούς πιστωτικούς κινδύνους. Η διαδικασία ταξινόμησης αρχίζει με ένα σύνολο στοιχείων στο οποίο οι κατηγορίες δεδομένων είναι ήδη γνωστές. Στο προηγούμενο παράδειγμα, ένα πρότυπο ταξινόμησης που προβλέπει τον πιστωτικό κίνδυνο θα μπορούσε να αναπτυχθεί βασισμένο στα δεδομένα που παρατηρήθηκαν για πολλούς υποψηφίους δανείου για μια χρονική περίοδο. Εκτός από την ιστορική αξιολόγηση φερεγγυότητας, τα δεδομένα μπορεί να ακολουθήσουν την ιστορία απασχόλησης, την εγχώρια ιδιοκτησία ή το ενοίκιο, τα έτη κατοικίας, τον αριθμό και τύπο επενδύσεων, και ούτω καθεξής. Η αξιολόγηση φερεγγυότητας θα ήταν ο στόχος, οι άλλες ιδιότητες θα ήταν οι προβλέψιμες μεταβλητές (predictor variables) και τα δεδομένα κάθε πελάτη θα αποτελούσαν μια κλάση [1],[2],[20].

Περισσότερα για τους τύπους διαχείρισης και ανάλυσης δεδομένων, καθώς και των αλγορίθμων που χρησιμοποιούνται θα αναφερθούμε παρακάτω ξεχωριστά για τη κάθε μια απ' τις λειτουργίες εξόρυξης δεδομένων.

2.6 Μέτρα Αξιολόγησης Κανόνων

Η αξιολόγηση των κανόνων και η μέτρηση του πόσο ενδιαφέροντες είναι οι παραγόμενοι κανόνες είναι σημαντικός τομέας στην έρευνα εξόρυξης δεδομένων. Μέχρι τώρα δεν υπάρχει καμία διαδεδομένη συμφωνία για κανένα επίσημο καθορισμό για την αξιολόγηση των κανόνων. Με βάση την ποικιλομορφία των ορισμών που παρουσιάστηκαν μέχρι σήμερα, η αξιολόγηση μπορεί να οριστεί καλύτερα ως μια ευρεία έννοια που δίνει έμφαση στην περιεκτικότητα (conciseness), κάλυψη (coverage), αξιοπιστία (reliability),ιδιαιτερότητα (peculiarity), ποικιλομορφία (diversity), καινοτομία (novelty), πόσο απροσδόκητα (surprisingness), ωφελιμότητα (utility), και δυνατότητα εφαρμογής (actionability). Αυτά τα εννέα ειδικά κριτήρια χρησιμοποιούνται για να καθορίσουν εάν ένας κανόνας είναι ή όχι ενδιαφέρον [21].

1) Περιεκτικότητα (Conciseness). Ένας κανόνας μπορεί να θεωρηθεί περιεκτικός εάν περιέχει σχετικά λίγα χαρακτηριστικά (attributes) - ζευγάρια τιμών, ενώ ένα σύνολο κανόνων είναι περιεκτικό εάν περιέχει σχετικά λίγους κανόνες. Ένα περιεκτικός κανόνας ή ένα περιεκτικό σύνολο κανόνων μπορεί να γίνει εύκολα κατανοητό και να μπορεί να συγκρατηθεί στη μνήμη και έτσι προστίθεται ευκολότερα στη γνώση του χρήστη. Ένα μέτρο που μπορεί να ελαχιστοποιήσει το σύνολο των

κανόνων είναι η εμπιστοσύνη(confidence). Η τιμή της εμπιστοσύνης δείχνει πόσο αξιόπιστος είναι ένας κανόνας. Όσο υψηλότερη η τιμή του, τόσο μεγαλύτερη η πιθανότητα τα επικεφαλής δεδομένα να εμφανίζονται σε μια ομάδα εάν είναι γνωστό ότι όλα τα δεδομένα περιλαμβάνονται σε εκείνη την ομάδα.

2) Γενικότητα / κάλυψη (Generality / Coverage). Ένα σύνολο κανόνων είναι γενικό (general) εάν καλύπτει ένα σχετικά μεγάλο υποσύνολο ενός συνόλου δεδομένων. Η γενικότητα (ή κάλυψη) μετρά την περιεκτικότητα ενός κανόνα, δηλαδή εάν καλύπτει όλη τη γκάμα, όλο το σύνολο των δεδομένων. Εάν ένα πρότυπο χαρακτηρίζει περισσότερες πληροφορίες στο σύνολο δεδομένων, τείνει να είναι πιο ενδιαφέρον. Τα συχνά σύνολα αντικειμένων (frequent itemsets) είναι τα πιο μελετημένα γενικά πρότυπα στη βιβλιογραφία της εξόρυξης δεδομένων. Ένα σύνολο αντικειμένων είναι συχνό εάν η υποστήριξή (support)[E] του, το φράγμα των εγγραφών στο σύνολο δεδομένων που περιέχει το itemset, είναι πάνω από ένα δεδομένο κατώτατο όριο. Η υποστήριξη (support) και η εμπιστοσύνη (confidence) είναι μετρητές που καθαρίζουν κατά πόσο ένας κανόνας είναι ενδιαφέρον και χρησιμοποιούνται για αξιολόγηση των κανόνων. Απεικονίζουν αντίστοιχα τη χρησιμότητα και τη βεβαιότητα των εξαγόμενων κανόνων [14].

3) Αξιοπιστία (reliability). Ένας κανόνας είναι αξιόπιστος εάν η σχέση που περιγράφεται από τον κανόνα εμφανίζεται σε ένα υψηλό ποσοστό των εφαρμόσιμων περιπτώσεων. Παραδείγματος χάριν, ένας κανόνας συσχέτισης είναι αξιόπιστος εάν έχει υψηλή εμπιστοσύνη (confidence). Έχουν προταθεί πολλά μέτρα για εύρεση αξιοπιστίας των κανόνων συσχέτισης από διάφορους κλάδους όπως από τις πιθανότητες, τη στατιστική, ανάκτηση πληροφοριών [21].

4) Ιδιαιτερότητα (Peculiarity). Ένας κανόνας είναι ιδιαίτερος (peculiar) εάν είναι πολύ διαφορετικός από τους άλλους παραγόμενους κανόνες σύμφωνα με κάποιο κριτήριο απόστασης (distance). Οι ιδιαίτεροι κανόνες παράγονται από τα ιδιαίτερα δεδομένα (ή outliers), τα οποία είναι σχετικά λίγα σε αριθμό και σημαντικά διαφορετικά από τα υπόλοιπα δεδομένα. Τα ιδιαίτερα πρότυπα μπορεί να είναι άγνωστα στο χρήστη, κι επομένως ενδιαφέροντα [19].

5) Ποικιλομορφία (Diversity). Ένας κανόνας είναι ποικιλόμορφος εάν τα στοιχεία του διαφέρουν σημαντικά το ένα από το άλλο, ενώ ένα σύνολο κανόνων είναι ποικιλόμορφο εάν οι κανόνες του διαφέρουν σημαντικά ο ένας από τον άλλο. Η ποικιλομορφία είναι ένας κοινός παράγοντας για τη μέτρηση του πόσο σημαντικές είναι οι περιλήψεις. Μια περίληψη μπορεί να θεωρηθεί ποικιλόμορφη εάν η διανομή πιθανότητάς της είναι πολύ διαφορετική από την ομοιόμορφη διανομή. Μια ποικιλόμορφη περίληψη μπορεί να είναι ενδιαφέρουσα επειδή με την απουσία οποιασδήποτε σχετικής γνώσης, ένας χρήστης συνήθως υποθέτει ότι η ομοιόμορφη διανομή θα κρατήσει σε μια περίληψη. Σύμφωνα με αυτόν τον συλλογισμό, η πιο ποικιλόμορφη περίληψη είναι και η πιο ενδιαφέρουσα [21].

6) Καινοτομία (Novelty). Ένας κανόνας είναι καινοτόμος εάν δεν ήταν γνωστό πριν και δεν ήταν δυνατό να βγει ως συμπέρασμα από άλλους γνωστούς κανόνες. Κανένα σύστημα εξόρυξης δεδομένων δεν αντιπροσωπεύει όλα όσα ένας χρήστης ξέρει, και έτσι, η καινοτομία δεν μπορεί να μετρηθεί ρητά σε σχέση με τη γνώση του χρήστη. Ομοίως, κανένα γνωστό σύστημα εξόρυξης

δεδομένων δεν αντιπροσωπεύει ότι ο χρήστης δεν ξέρει, και επομένως, η καινοτομία δεν μπορεί να μετρηθεί ρητά σε σχέση με την άγνοια του χρήστη. Αντί αυτού, η καινοτομία μπορεί να ανιχνευτεί από τον χρήστη, είτε ρητά να προσδιορίζει ένα κανόνα ως νέο, είτε παρατηρεί ότι ένας κανόνας δεν έρχεται σε αντιπαράθεση με κανόνες που έχουν ανακαλυφθεί προγενέστερα . Στην τελευταία περίπτωση, τα πρότυπα που έχουν ανακαλυφθεί χρησιμοποιούνται ως προσέγγιση στη γνώση του χρήστη [21].

7) Εξαγόμενοι κανόνες (Surprisingness). Ένας κανόνας είναι έκπληξη (ή εξαγόμενος) εάν έρχεται σε αντίθεση με την υπάρχουσα γνώση ή τις προσδοκίες του χρήστη. Ένας κανόνας, που είναι εξαίρεση σε ένα πρότυπο που έχει ανακαλυφθεί ήδη, μπορεί επίσης να θεωρηθεί απροσδόκητος. Οι απροσδόκητοι κανόνες είναι επίσης ενδιαφέροντες επειδή προσδιορίζουν τις αποτυχίες στην προηγούμενη γνώση και μπορούν να προτείνουν μια πτυχή δεδομένων που χρειάζεται περαιτέρω μελέτη. Η διαφορά μεταξύ του απροσδόκητου και της καινοτομίας είναι ότι ένας κανόνας είναι καινούργιος και δεν έρχεται σε αντίφαση με οποιοδήποτε άλλο κανόνα που ήταν ήδη γνωστός στο χρήστη, ενώ ένας απροσδόκητος κανόνας έρχεται σε αντίφαση με την προηγούμενη γνώση ή τις προσδοκίες του χρήστη.

8) Ωφελιμότητα (Utility). Ένας κανόνας είναι ωφέλιμος εάν η χρήση του, από τον χρήστη, συμβάλλει στην επίτευξη ενός στόχου. Διαφορετικοί χρήστες μπορεί να έχουν διαφορετικούς στόχους σχετικά με τη γνώση που μπορεί να εξαχθεί από ένα σύνολο δεδομένων. Παραδείγματος χάριν, ένας χρήστης μπορεί να ενδιαφερθεί για την εύρεση όλων των πωλήσεων με ψηλό κέρδος από ένα σύνολο δεδομένων, ενώ άλλος μπορεί να ενδιαφέρεται για την εύρεση όλων το δοσοληψιών με μεγάλες αυξήσεις στις ακαθάριστες πωλήσεις. Αυτό το είδος ενδιαφέροντος είναι βασισμένο στις λειτουργίες χρησιμότητας που καθορίζονται από το χρήστη.

9) Δυνατότητα εφαρμογής(Actionability / Applicability). Ένας κανόνας είναι εφαρμόσιμος σε κάποια περιοχή εάν επιτρέπει τη λήψη απόφασης για μελλοντικές ενέργειες σε αυτήν την περιοχή [21].

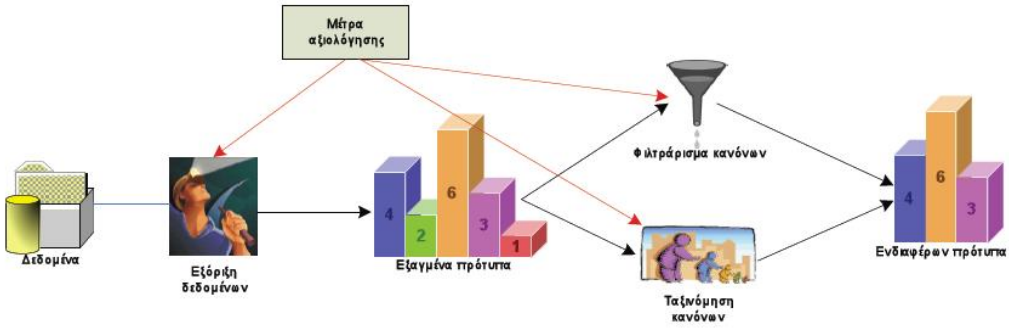
Αυτά τα εννέα κριτήρια μπορούν να ταξινομηθούν περαιτέρω σε τρεις κατηγορίες: αντικειμενικά, υποκειμενικά, και βασισμένα στη σημασιολογία. Ένα αντικειμενικό μέτρο είναι βασισμένο μόνο στα ακατέργαστα δεδομένα. Δεν απαιτείται καμία γνώση για το χρήστη ή την εφαρμογή. Τα περισσότερα αντικειμενικά μέτρα είναι βασισμένα στις θεωρίες πιθανοτήτων, τις στατιστικές, ή τη θεωρία πληροφοριών. Η περιεκτικότητα, η γενικότητα, η αξιοπιστία, η ιδιαιτερότητα, και η ποικιλομορφία εξαρτώνται μόνο από τα δεδομένα και τους κανόνες, και μπορούν έτσι να θεωρηθούν αντικειμενικά.

Ένα υποκειμενικό μέτρο λαμβάνει υπόψη και τα δεδομένα και το χρήστη των δεδομένων. Για να καθοριστεί ένα υποκειμενικό μέτρο, απαιτείται πρόσβαση στην περιοχή του χρήστη ή στο υπόβαθρο της γνώσης για τα δεδομένα. Αυτή η πρόσβαση μπορεί να ληφθεί με την αλληλεπίδραση με το χρήστη κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων ή με το να αντιπροσωπεύσει ρητά τη γνώση ή τις προσδοκίες του χρήστη. Η καινοτομία και τα απροσδόκητα πρότυπα εξαρτώνται από το χρήστη, καθώς επίσης και από τα δεδομένα και τους κανόνες, και ως εκ τούτου αυτά τα κριτήρια μπορούν να θεωρηθούν υποκειμενικά.

Ένα σημασιολογικό μέτρο εξετάζει τη σημασιολογία και τις εξηγήσεις των κανόνων. Επειδή τα σημασιολογικά μέτρα περιλαμβάνουν τη γνώση από το χρήστη, μερικοί ερευνητές τους θεωρούν σαν έναν ειδικό τύπο υποκειμενικού μέτρου. Η ωφελιμότητα και η δυνατότητα εφαρμογής εξαρτώνται από τη σημασιολογία των δεδομένων, και έτσι μπορούν να θεωρηθούν σημασιολογικά κριτήρια. Τα μέτρα που είναι βασισμένα στην ωφελιμότητα, όπου η σχετική σημασιολογία είναι οι χρησιμότητες των κανόνων, είναι ο πιο κοινός τύπος σημασιολογικού μέτρου. Για να χρησιμοποιηθεί μια προσέγγιση βασισμένη στη χρησιμότητα, ο χρήστης πρέπει να διευκρινίσει πρόσθετη γνώση για την περιοχή. Αντίθετα από τα υποκειμενικά μέτρα, όπου η γνώση είναι για τα ίδια τα δεδομένα και αντιπροσωπεύεται συνήθως με ένα σχήμα παρόμοιο με αυτό του ανακαλυμμένου προτύπου, η γνώση που απαιτείται για τα σημασιολογικά μέτρα δεν σχετίζει τη γνώση ή τις προσδοκίες του χρήστη με τα δεδομένα. Αντί αυτού, αντιπροσωπεύει μια λειτουργία χρησιμότητας που απεικονίζει τους στόχους του χρήστη. Παραδείγματος χάριν, ένας διευθυντής καταστημάτων θα προτιμήσει τους κανόνες συσχέτισης που αφορούν τα δεδομένα με ψηλό κέρδος παρά από εκείνους με την υψηλότερη στατιστική σημασία.

Για τον προσδιορισμό του πόσο ενδιαφέρον είναι ένας κανόνας, υπάρχουν τρεις μέθοδοι εκτέλεσης. Κατ' αρχάς, μπορούμε να ταξινομήσουμε κάθε κανόνα από το εάν είναι ενδιαφέρων ή όχι. Εν συνεχεία, μπορούμε να καθορίσουμε μια σχέση προτίμησης που καθορίζει εάν ένας κανόνας είναι πιο ενδιαφέρον από τον άλλο. Τέλος, να βαθμολογήσουμε τους κανόνες. Για την πρώτη ή τρίτη προσέγγιση, μπορούμε να καθορίσουμε ένα μέτρο βασισμένο στα προαναφερθέντα εννέα κριτήρια και να το χρησιμοποιήσουμε για να ξεχωρίσουμε τους ενδιαφέροντες και τους μη-ενδιαφέροντες κανόνες στην πρώτη προσέγγιση ή για να βαθμολογήσουμε τα πρότυπα στην τρίτη προσέγγιση.

Κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων, τα μέτρα αξιολόγησης που παρουσιάζουν ενδιαφέρουν είναι οι κανόνες, μπορούν να χρησιμοποιηθούν με τρεις τρόπους, τους οποίους καλούμε ρόλους των μέτρων. Το παρακάτω σχήμα παρουσιάζει αυτούς τους τρεις ρόλους. Κατ' αρχάς, τα μέτρα μπορούν να χρησιμοποιηθούν για να 'κλαδέψουν' τους κανόνες, που δεν είναι ενδιαφέροντες, κατά τη διάρκεια της διαδικασίας εξόρυξης ώστε να ελαχιστοποιηθεί ο χώρος αναζήτησης και να βελτιωθεί έτσι η αποδοτικότητα της εξόρυξης.



Εικόνα 4 Μέτρα αξιολόγησης κανόνων στη διαδικασία εξόρυξης δεδομένων [4]

2.7 Εφαρμογές εξόρυξης δεδομένων

Ένα απ' τα πιο δυνατά σημεία της εξόρυξης δεδομένων αντικατοπτρίζεται στην πληθώρα μεθόδων που μπορούν να χρησιμοποιηθούν για να λύσουν έναν μεγάλο αριθμό προβλημάτων .Το γεγονός αυτό επιτρέπει την χρήση της εξόρυξης δεδομένων σε πάρα πολλά διαφορετικά μεταξύ τους πεδία τα οποία βρίσκουν χρήσιμη την εισροή και επεξεργασία δεδομένων προκειμένου να παραχθούν χρήσιμα συμπεράσματα. Εφαρμογές της εξόρυξης δεδομένων μπορούμε να εντοπίσουμε σε εμπορικούς τομείς όπως για παράδειγμα σε αλυσίδες σουπερ-μάρκετ, τράπεζες, στα μέσα μαζικής ενημέρωσης και στην διαφήμιση. Γενικότερα με την εξόρυξη δεδομένων μπορούμε να προβλέψουμε συμπεριφορές και να εντοπίσουμε τάσεις και μοτίβα. Έτσι μπορούμε να δούμε την χρήση της, εκτός από τον τομέα της πληροφορικής, στην ιατρική, στην βιολογία αλλά και στις τηλεπικοινωνίες. Παρακάτω θα αναφερθούμε συνοπτικά σε ορισμένα πεδία και κλάδους τα οποία επωφελούνται σε σημαντικό βαθμό από τη χρήση διαδικασιών εξόρυξης δεδομένων.

2.7.1 Επιχειρήσεις και οργανισμοί

Η εξόρυξη δεδομένων στις διοικητικές εφαρμογές σχέσης πελατών μπορεί να συμβάλει σημαντικά στην κατώτατη ιεραρχική γραμμή. Αντί να έρθει σε επαφή τυχαία με μια προοπτική ή έναν πελάτη μέσω ενός τηλεφωνικού κέντρου ή την αποστολή του ταχυδρομείου, μια επιχείρηση μπορεί να συγκεντρώσει τις προσπάθειές της στις προοπτικές που προβλέπονται για να έχουν υψηλή πιθανότητα απάντησης σε μια προσφορά. Οι περιπλοκότερες μέθοδοι μπορούν να χρησιμοποιηθούν για να βελτιστοποιήσουν τους πόρους στις διαφημιστικές εκστρατείες έτσι ώστε κάποια μπορεί να προβλέψει σε ποιο κανάλι και ποια προσφορά ένα άτομο είναι πλέον πιθανό να αποκρίνεται πέρα από όλες τις πιθανές προσφορές. Επιπλέον, οι περίπλοκες εφαρμογές θα μπορούσαν να χρησιμοποιηθούν για να αυτοματοποιήσουν την αποστολή. Μόλις καθοριστούν τα αποτελέσματα από την εξόρυξη δεδομένων (πιθανή προοπτική/πελάτης και κανάλι/προσφορά), αυτή η "περίπλοκη διαδικασία" μπορεί είτε αυτόματα να στείλει μήνυμα ηλεκτρονικού ταχυδρομείο είτε μέσω κανονικού ταχυδρομείου. Τέλος, σε περιπτώσεις εάν πολλοί άνθρωποι κινηθούν διαφορετικά, πέρα από τις πιθανές προοπτικές χωρίς να δοθεί προσφορά, η διαμόρφωση ανόδου μπορεί να χρησιμοποιηθεί για να καθορίσει ποιοι άνθρωποι θα έχουν τη μέγιστη αύξηση στην απάντηση εάν τους δοθεί καινούρια προσφορά. Τα δεδομένα που συγκεντρώνονται μπορούν επίσης να χρησιμοποιηθούν για να ανακαλύψουν αυτόματα τα τμήματα ή τις ομάδες μέσα σε ένα σύνολο δεδομένων πελατών [6].

Στη βιομηχανία των επιχειρήσεων η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για την εξερεύνηση των τάσεων της αγοράς, να σχεδιαστούν σχέδια επενδύσεων και να εντοπισθούν ατέλειες στα λογιστικά συστήματα. Επιπροσθέτως, εφαρμόζεται στην καλυτέρευση των μεθόδων και διαδικασιών που χρησιμοποιούνται στις εκστρατείες marketing. Κατόπιν αναλύσεως των δεδομένων μια επιχείρηση μπορεί να προμηθεύσει τους πελάτες της με μια πιο επικεντρωμένη υποστήριξη

προϊόντος και παραθέτοντας συγκεκριμένες λύσεις στα προβλήματα που προκύπτουν να κρατήσει το βαθμό ικανοποίησης του πελατολόγιου της σε υψηλά επίπεδα [6].

Τα προγράμματα marketing κοστίζουν στους οργανισμούς ένα μεγάλο ποσοστό των χρημάτων που δαπανώνται/επενδύονται στο σχεδιασμό παραγωγής και διανομής της πρώτης ύλης. Αν η εκστρατεία δεν σχεδιαστεί με βάση συγκεκριμένους πελάτες, η αντίδραση στη προσφορά μπορεί να έχει αρνητικά αποτελέσματα όχι μόνο σε σχέση με τα έξοδα σχεδιασμού της εκστρατείας αλλά και με τη μορφή μικρής απορρόφησης του προϊόντος απ' την αγορά. Επιπροσθέτως εάν η διανομή του υλικού marketing δεν διαχειριστεί σωστά, η εκστρατεία δεν θα έχει την αναμενόμενη αποτελεσματικότητα. Ασυμβίβαστες ομάδες δεδομένων που περιλαμβάνουν λάθος ονόματα, διευθύνσεις οι οποίες δεν ισχύουν πια και ατελή πεδία εγγραφών, είναι λάθη που δημιουργούν προβλήματα μη απόκρισης στους στόχους του marketing και εκτεταμένα έξοδα εκστρατειών τα οποία δεν επιφέρουν αποτελέσματα λόγο κακών υπολογισμών. Με την ανάλυση δεδομένων που μπορούν να έχουν οι οργανισμοί μέσω των διαδικασιών εξόρυξης δεδομένων, παράγουν στοιχεία με τα οποία μπορούν να κάνουν διορθωτικές κινήσεις και γενικότερα να αναβαθμίσουν την εκστρατεία marketing [6].

Στα τμήματα ανθρωπίνων πόρων η εξόρυξη γνώσης μπορεί να είναι χρήσιμη, στον προσδιορισμό των χαρακτηριστικών των πιο επιτυχημένων υπαλλήλων τους. Πληροφορίες που λαμβάνονται, όπως τα πανεπιστήμια που φοίτησαν στο παρελθόν στελέχη τα οποία οδηγήθηκαν στην επιτυχία , μπορούν να βοηθήσουν την εστίαση του τμήματος ανθρωπίνων πόρων κατά τη διαδικασία διαλογής του κατάλληλου προσωπικού για πρόσληψη. Επιπλέον, οι στρατηγικές εφαρμογές διοίκησης επιχειρήσεων βοηθούν μια επιχείρηση να πλησιάσουν τους στρατηγικούς εταιρικούς μακροπρόθεσμους στόχους, όπως οι στόχοι μεριδίου κέρδους παραδείγματος χάριν και μέσω της ανάλυσης δεδομένων να λάβουν αποφάσεις που θα την φέρουν πιο κοντά στην ολοκλήρωση των στόχων αυτών [6].

2.7.2 Επιστήμη και εφαρμοσμένη μηχανική

Τα τελευταία χρόνια, η εξόρυξη δεδομένων έχει χρησιμοποιηθεί ευρέως στους τομείς της επιστήμης και της εφαρμοσμένης μηχανικής, όπως η βιοπληροφορική, η γενετική, η εφαρμοσμένη μηχανική ιατρική και η εκπαίδευση.

Στη μελέτη της ανθρώπινης γενετικής, ένας σημαντικός στόχος είναι να γίνει κατανοητή η σχέση χαρτογράφησης μεταξύ της διατομικής παραλλαγής στις ανθρώπινες ακολουθίες DNA και της μεταβλητότητας στην ευαισθησία ασθενειών. Με απλά λόγια, πρόκειται για την έρευνα του πώς οι αλλαγές σε μια ακολουθία DNA έχουν επιπτώσεις στο να εμφανίσουν άτομα ασθένειες όπως ο καρκίνος. Αυτό είναι πολύ σημαντικό για τη βελτίωση της διάγνωσης, την πρόληψη και τη θεραπεία των ασθενειών. Η μέθοδος εξόρυξης δεδομένων που χρησιμοποιείται για να εκτελέσει αυτόν τον στόχο είναι γνωστή ως multifactor dimensionality reduction (μείωση διαστατικότητας πολλαπλών ρόλων) [7].

Στη φαρμακοβιομηχανία. Στο κέντρο επιτήρησης παρενέργειας φαρμάκων ,της Uppsala ,από το 1998, έχει γίνει χρήση των μεθόδων εξόρυξης δεδομένων που καλύπτουν συνήθως την υποβολή εκθέσεων των σχεδίων ενδεικτικών των αναδυόμενων ζητημάτων ασφάλειας φαρμάκων στη σφαιρική βάση δεδομένων WHO η οποία αναλύει περί τα τεσσεράμισι εκατομμύρια πιθανών παρενεργειών φαρμάκων. Πρόσφατα, παρόμοια μεθοδολογία έχει αναπτυχθεί για να εξαγάγει μεγάλες συλλογές ηλεκτρονικών αρχείων υγείας συνδέοντας τις συνταγές φαρμάκων με τις ιατρικές διαγνώσεις[7].

Στη βιοπληροφορική. Τα τελευταία χρόνια, οι γρήγορες εξελίξεις στη γενετική έχουν παραγάγει έναν τεράστιο όγκο βιολογικών δεδομένων. Το να εξαγάγουμε συμπεράσματα από έναν τέτοιο μεγάλο όγκο δεδομένων απαιτεί πολύπλοκους υπολογισμούς. Η Βιοπληροφορική ή η υπολογιστική βιολογία, είναι η διεπιστημονική επιστήμη της ερμηνείας των βιολογικών στοιχείων που χρησιμοποιούν την τεχνολογία πληροφοριών και την πληροφορική. Η σημασία αυτού του νέου τομέα της έρευνας θα αυξηθεί δεδομένου ότι συνεχίζουμε να παράγουμε και να διαχειριζόμαστε μεγάλες ποσότητες βιολογικών δεδομένων. Ένας ιδιαίτερα ενεργός τομέας της έρευνας της βιοπληροφορικής, είναι η εφαρμογή και η ανάπτυξη των τεχνικών εξόρυξης δεδομένων ανάλυσης για να λυθούν τα βιολογικά προβλήματα. Τα παραδείγματα αυτού του τύπου ανάλυσης περιλαμβάνουν την πρωτεϊνική πρόβλεψη δομών, την ταξινόμηση γονιδίων, την ταξινόμηση καρκίνου βασισμένου στα microarray στοιχεία, συγκέντρωση των στοιχείων έκφρασης γονιδίων, στατιστική διαμόρφωση της πρωτεϊνικής αλληλεπίδρασης, κ.λπ. Επομένως, βλέπουμε ότι με τη χρήση της εξόρυξης δεδομένων η βιοπληροφορική μπορεί να αναλύσει τον τεράστιο όγκο δεδομένων της και να παράγει έγκυρα αποτελέσματα [7].

2.7.3 Συστήματα ERP(Enterprise Resource Planning)

Ο σχεδιασμός επιχειρηματικών πόρων (Enterprise Resource Planning) είναι ένα σύστημα βασισμένο σε ηλεκτρονικό υπολογιστή που χρησιμοποιείται για να διαχειριστεί τους εσωτερικούς (οικονομικούς, υλικό, ανθρώπινο δυναμικό) και εξωτερικούς πόρους (πελάτες, προμηθευτές) μιας επιχείρησης. Είναι μια αρχιτεκτονική λογισμικού της οποίας σκοπός είναι να διευκολυνθεί η ροή πληροφοριών μεταξύ όλων των επιχειρησιακών λειτουργιών μέσα στα όρια της οργάνωσης και να ρυθμιστούν οι συνδέσεις με τους εξωτερικούς συμμέτοχους (manage the connections to outside stake holders).Στηριγμένα σε μια κεντρική βάση δεδομένων και χρησιμοποιώντας μια κοινής πλατφόρμα υπολογισμού, είναι σχεδιασμένα με κύριο στόχο την επιτάχυνση των καθημερινών δραστηριοτήτων / συναλλαγών της επιχείρησης και ενσωματώνουν αρκετή "ευφυΐα" για να παράγουν ένα πλήθος αναφορών προς τη διοίκηση, οι αναφορές αυτές είναι κατά κανόνα χρονικά ή τοπικά εστιασμένες συνόψεις των αναλυτικών στοιχείων που τηρούνται στη βάση του ERP, π.χ., συγκεντρωτικά στοιχεία τριμήνων, γεωγραφικών περιοχών κ.λπ., είτε διαφορετικές όψεις τους. Με άλλα λόγια, δεν ενσωματώνουν αρκετή "νοημοσύνη" για να αναλύσουν τα δεδομένα και να προτείνουν μοντέλα δράσης ικανά να βοηθήσουν τη διοίκηση στη λήψη αποφάσεων.

Δεν προσθέτουν κάτι στη γνώση που χρειάζεται η επιχείρηση για να πάρει αποφάσεις για τις επόμενες κινήσεις της. Αυτή η "έλλειψη" ανάλυσης και δημιουργίας γνώσης αντιμετωπίζεται με τις διαδικασίες εξόρυξης δεδομένων. Οι διαδικασίες αυτές αναλύουν τα δεδομένα που διαθέτει το σύστημα ERP με σκοπό να παράγουν ωφέλιμη γνώση ώστε τα στελέχη οργανισμού να μπορέσουν να προβούν στις ενέργειες εκείνες οι οποίες θα τους επιτρέψουν να κάνουν διορθωτικές κινήσεις προς την επίτευξη των στόχων του οργανισμό [8].

2.7.4 Logistics (Εφοδιαστική)

Οι αναδυόμενες τεχνολογίες πληροφορικής και των εφαρμογών της, επέτρεψαν πολλές εταιρείες να αποκτήσουν τεράστιο όγκο δεδομένων για τους πελάτες, τους προμηθευτές και τους συνεργάτες τους. Ο τρόπος να διαχειριστεί κανείς την πολύ μεγάλη βάση δεδομένων στη διαχείριση της αλυσίδας εφοδιασμού είναι ένα πολύ σημαντικό ζήτημα. Ένα μεγάλο μέρος των δεδομένων είναι δυνατό να συγκεντρωθεί από τις πολλές δραστηριότητες κατά μήκος της εφοδιαστικής αλυσίδας. Τα στοιχεία αυτά πρέπει να αποθηκεύονται και να δέχονται την κατάλληλη επεξεργασία ώστε να είναι δυνατή η αναγνώριση τάσεων, πρακτικών ή ακόμη και γνώσης που διαφορετικά δεν είναι προφανής. Αυτή η γνώση μπορεί να οδηγήσει σε καλύτερο έλεγχο και διαχείριση της αλυσίδας. Η τεχνολογία data mining είναι μια διαδικασία που εφαρμόζει διαφορετικά εργαλεία ανάλυσης για να ανακαλύψει σχέσεις μεταξύ των δεδομένων ώστε να χρησιμοποιηθούν σε μελλοντικές προβλέψεις. Σε μια εφοδιαστική αλυσίδα λαμβάνουν μέρος πολλοί εταίροι με διαφορετικό ρόλο σε αυτήν ο καθένας με διαφορετικά δεδομένα εισόδου και εξόδου για τα πληροφοριακά τους συστήματα. Παρόλα αυτά, η συνεργασία όλων αυτών των μηχανισμών φέρει εις πέρας ένα γενικότερο έργο, αυτό της διακίνησης των προϊόντων από τη στιγμή που είναι πρώτες ύλες μέχρι την τελική πώληση ως έτοιμο προϊόν στον πελάτη. Έτσι, παρά το γεγονός ότι υπάρχουν πολλές ανομοιογένειες στα δεδομένα που χειρίζεται κάθε συμβαλλόμενο μέρος, όλα τα δεδομένα αναφέρονται στο ίδιο γενικότερο έργο. Αυτή η ασυμβατότητα είναι ένα από τα μεγαλύτερα προβλήματα στην καθημερινή επικοινωνία και μια επιτυχημένη αλυσίδα πρέπει να το ξεπεράσει ώστε να επιτύχει την μεγαλύτερη δυνατή ολοκλήρωση της. Αυτή η ολοκλήρωση μπορεί να επιτευχθεί σε πολλά επίπεδα όπως αυτό της ανταλλαγής δεδομένων και του ελέγχου. Οι εταίροι μιας εφοδιαστικής αλυσίδας διαθέτουν μια ή και περισσότερες βάσεις δεδομένων σχετικά με τις δραστηριότητες τους που αναφέρονται στην εφοδιαστική αλυσίδα. Ο συνδυασμός δεδομένων από τις διάφορες βάσεις δεδομένων και η αντίστοιχη επεξεργασία τους μπορεί να αναδείξει γνώση η οποία δεν ήταν προφανής. Οι τεχνικές εξόρυξης δεδομένων, μπορούν να βοηθήσουν τις στρατηγικές αποφάσεις αναλύοντας και αποκαλύπτοντας επιδόσεις και συμπεριφορές του παρελθόντος της επιχείρησης. Οι τεχνικές που μπορούν να χρησιμοποιηθούν είναι πολλές και ποικίλες, όπως η στατιστική ανάλυση, οι τεχνικές μοντελοποίησης, και άλλες τεχνικές βάσεων δεδομένων για να ανακαλυφθούν ασαφής, μέχρι τότε, σχέσεις μεταξύ των δεδομένων και των μοντέλων πρόβλεψης που χρησιμοποιούνται. Η διαδικασία που ακολουθείται κατά την εξόρυξη δεδομένων είναι άμεσα σχετιζόμενη με τον καθαρισμό και την ομογενοποίηση των δεδομένων από τις διαφορετικές πηγές – βάσεις δεδομένων μέχρι την δημιουργία μιας ενιαίας αποθήκης δεδομένων, στην

οποία εφαρμόζονται όλες εκείνες οι τεχνικές που μπορούν να αναδείξουν την γνώση πάνω στα δεδομένα. Οι εφαρμογές αυτών των τεχνικών μπορούν να εφαρμοστούν σε διαφορετικές λειτουργικές περιοχές της αλυσίδας όπως στις πωλήσεις, στις αγορές στις μεταφορές ή στην αποθήκευση. Από το κύκλωμα των πωλήσεων μπορεί να βρεθεί το ποια προϊόντα ζητούνται περισσότερο από τους πελάτες, ποιά από αυτά επιστρέφονται και για ποιο λόγο. Ταυτόχρονα, η ίδια πληροφορία μπορεί να καθοδηγήσει τη διοίκηση να αναζητήσει τους κατάλληλους προμηθευτές και να περιορίσει αυτούς που της παρέχουν μη κατάλληλα προϊόντα. Οι ίδιοι οι προμηθευτές ή ακόμη και οι παραγωγοί μπορούν να έχουν πληροφόρηση για τα προϊόντα που διαθέτουν, τα πλεονεκτήματα και τα μειονεκτήματα τους ώστε να μπορέσουν να διορθώσουν το συντομότερο δυνατό την παραγωγική τους διαδικασία [9].

2.7.5 Ηλεκτρονικό Εμπόριο

Η ηλεκτρονική αγορά είναι η διαδικασία με την οποία οι καταναλωτές αγοράζουν άμεσα αγαθά ή υπηρεσίες από ένα κατάστημα σε πραγματικό χρόνο από το διαδίκτυο, χωρίς κανένα ενδιάμεσο πωλητή. Όταν ένας καταναλωτής αγοράζει προϊόντα ή υπηρεσίες από μία εταιρία, η διαδικασία ονομάζεται B2C, δηλαδή Business to Costumers. Όταν μια εταιρία αγοράζει προϊόντα ή υπηρεσίες από μια άλλη επιχείρηση, η διαδικασία καλείται ενδοεπιχειρησιακές αγορές B2B –Business to Business. Τέλος υπάρχει άλλη μια μορφή ηλεκτρονικού εμπορίου, όταν το κράτος αγοράζει προϊόντα ή υπηρεσίες από μια επιχείρηση. Η διαδικασία ονομάζεται B2G, δηλαδή Business to Gonverment. Όπως γίνεται κατανοητό οι υποψήφιοι αγοραστές on-line αγορών μπορούν να βρίσκονται οπουδήποτε στον κόσμο και να παραγγέλνουν από μια πληθώρα διαφορετικών πραγμάτων. Η εξόρυξη δεδομένων θα βοηθούσε να τοποθετήσουμε τις αγορές σε γεωγραφικά τμήματα που αποτελούνται είτε από σύνολα κρατών, είτε από νομούς (αναλόγως ότι διευκολύνει την επιχείρηση) και σε προϊόντα που πωλούνται σε αυτά τα γεωγραφικά τμήματα. Έτσι η επιχείρηση μπορεί να κάνει επιχειρηματικές προβλέψεις και να μπορέσει να αναδιαρθρώσει καλύτερα το σύστημα διανομής της, τις αποθήκες και γενικότερα να μειώσει το κόστος προς όφελος της και φυσικά προς όφελος των καταναλωτών [10].

2.7.6 Λήψη αποφάσεων με πληροφοριακά συστήματα

Ένα ακόμα από τα δυνατά σημεία της εξόρυξης δεδομένων, είναι η λήψη αποφάσεων με πληροφοριακά συστήματα. Αυτό συμβαίνει γιατί η πληροφορία γίνεται γνώση, με αποτέλεσμα να αξιολογείται κατάλληλα από το εκάστοτε πληροφοριακό σύστημα και τον χειριστή του, έτσι ώστε να λαμβάνεται η βέλτιστη δυνατή απόφαση. Αξίζει να γίνει αναφορά στις 3 κυριότερες απεικονίσεις γνώσης οι οποίες χρησιμοποιούνται περισσότερο από τα πληροφοριακά συστήματα. Οι απεικονίσεις γνώσης λοιπόν που χρησιμοποιούνται περισσότερο για το εξαγόμενο αποτέλεσμα, είναι οι πίνακες απόφασης, τα δένδρα απόφασης και οι κανόνες απόφασης, στα οποία θα αναφερθούμε εκτενέστερα σε επόμενο κεφάλαιο [11].

2.7.7 Διακυβέρνηση

Με το να εξαγάγει δεδομένα, η κυβέρνηση μπορεί να πάρει πολλές εφαρμόσιμες πληροφορίες για να χαράξει τις επόμενες πολιτικές κινήσεις της, όπως, τη χρηματοδότηση, και την προστασία του περιβάλλοντος, την καταπολέμηση της εγκληματικότητας και την εθνική ασφάλεια. Ειδικά, υπάρχουν πολλές ερευνητικές περιπτώσεις για την εφαρμογή της τεχνολογίας εξόρυξης δεδομένων για να αποτραπεί το έγκλημα, όπως το πώς να γίνεται γρηγορότερα και σωστότερα η σύγκριση των φυσικών χαρακτηριστικών (δηλ. δακτυλικά αποτυπώματα, φωνές, πρόσωπα ή χέρια). Ο συνδυασμός της εξόρυξης δεδομένων και της βιομετρικής είναι αναμφισβήτητα μια εξαιρετικά πολύτιμη τεχνολογία που έχει τη μεγάλη βοήθεια στον προσδιορισμό εγκλήματος [12].

3. Κανόνες Συσχετίσεων (Association Rules)

3.1 Πως Λειτουργούν οι κανόνες συσχετίσεων

Οι κανόνες συσχετίσεως είναι μια λειτουργία του data mining η οποία ανακαλύπτει τις πιθανότητες επανεμφάνισης των στοιχειών μιας ομάδας. Οι σχέσεις μεταξύ επανεμφανιζόμενων στοιχείων εκφράζονται ως κανόνες συσχέτισης. Οι κανόνες συσχετίσεως είναι δηλώσεις τύπου if/then που βοηθούν στην αποκάλυψη των σχέσεων μεταξύ των φαινομενικά ανεξάρτητων δεδομένων σε μια σχεσιακή βάση δεδομένων ή άλλης αποθήκης πληροφοριών (data-warehouse) [19].

Αυτές οι τεχνικές επιτρέπουν στους αναλυτές και τους ερευνητές να αποκαλύψουν τα κρυμμένα πρότυπα (patterns) στα μεγάλα σύνολα δεδομένων. Ένα γενικό παράδειγμα εφαρμογής των κανόνων συσχετίσεων μπορεί να είναι το εξής: Πελάτες που παρήγγειλαν το προϊόν Α, συχνά παραγγέλνουν επίσης τα προϊόντα Β και Γ. Η εφαρμογή του αλγορίθμου A-Priori, μας επιτρέπει να επεξεργαστούμε γρήγορα τεράστια σύνολα δεδομένων για τέτοιες ενώσεις, βασισμένες σε προκαθορισμένες τιμές για την ανίχνευση των σχέσεων μεταξύ των δεδομένων. Ουσιαστικά οι κανόνες συσχετίσεων στην εξόρυξη δεδομένων χρησιμεύουν στις επιχειρήσεις για τον προσδιορισμό της συμπεριφοράς του πελάτη. Παίζουν ένα σημαντικό μέρος στην ανάλυση δεδομένων στο καλάθι αγορών, στη συγκέντρωση (clustering) προϊόντων και στις εφαρμογές cross και upselling στρατηγικών marketing [19].

Οι κανόνες συσχετίσεων έχουν δύο μέρη, ένα προηγούμενο (εάν / if) και μια συνέπεια (έπειτα / then). Το πρώτο μέρος (προηγούμενο) είναι ένα στοιχείο που βρίσκεται στα δεδομένα. Η συνέπεια είναι ένα στοιχείο που βρίσκεται σε σχέση με το πρώτο μέρος.

Οι κανόνες συσχετίσεων δημιουργούνται με την ανάλυση των δεδομένων για τα συχνά εμφανιζόμενα μέρη if / then και τη χρήση της υποστήριξης κριτηρίων εμπιστοσύνης για να προσδιοριστούν οι σημαντικότερες σχέσεις. Η υποστήριξη (support) είναι μια ένδειξη για το πόσο συχνά τα δεδομένα εμφανίζονται στη βάση δεδομένων. Η εμπιστοσύνη (confidence) δείχνει το πόσες φορές οι δηλώσεις if / then έχουν βρεθεί να είναι αληθινές. Γενικότερα μπορούμε να παραθέσουμε τους εξής ορισμούς :

Υποστήριξη (support) ενός κανόνα. Ορίζεται ως η υποστήριξη του συνόλου των αντικειμένων του κανόνα δηλαδή $sup(X \rightarrow Y) = sup\,(X \cup Y)$. Ουσιαστικά μας δείχνει πόσες φορές εμφανίζεται το σύνολο αντικειμένων του κανόνα [8].

Εμπιστοσύνη (confidence) ενός κανόνα. Ορίζεται ως $conf(X \rightarrow Y) = sup(X \cup Y)$ / $sup(X)$και καθορίζει τον αριθμό των εμφανίσεων του Yστις συναλλαγές (transactions) που περιέχουν το X. Ουσιαστικά μας δείχνει την ισχύ της συνεπαγωγής του κανόνα που εξετάζουμε [8].

Οι κανόνες συσχετίσεως συνήθως χρειάζονται για να ικανοποιήσουν μια καθορισμένη ως προς τον χρήστη ελάχιστη υποστήριξη (support) και μια καθορισμένη ως προς τον χρήστη ελάχιστη εμπιστοσύνη (confidence) συγχρόνως. Η δημιουργία κανόνων συσχετίσεως χωρίζετε συνήθως σε δύο χωριστά βήματα:

1.Εύρεση των πιο συχνά εμφανιζόμενων συνόλων αντικειμένων (frequent itemsets), δηλαδή των συνόλων που ικανοποιούν την απαίτηση για μεγάλη υποστήριξη.

2.Εξαγωγή των κανόνων συσχέτισης (association rules) που ικανοποιούν το κατώφλι (threshold) της ελάχιστης εμπιστοσύνης που έχουμε θέσει.

3.2 Είδη κανόνων συσχέτισης

Στην κλασσική ανάλυση προσπαθούμε να βρούμε κανόνες συσχέτισης X→Y. Η συσχέτιση αυτή έχει σκοπό να αναδείξει την πιθανότητα η ύπαρξη του συνόλου X να συνεπάγεται την ύπαρξη του συνόλου Y στις δοσοληψίες της βάσης δεδομένων. Για να γίνει κατανοητή η έννοια των κανόνων φέρνουμε ως παράδειγμα το καλάθι αγορών προϊόντων σε ένα πολυκατάστημα. Η στόχευση μας είναι να βρούμε ποια σύνολα προϊόντων θα οδηγήσουν στην αγορά άλλων προϊόντων. Ο κανόνας γάλα→ψωμί σημαίνει ότι η ύπαρξη γάλακτος συνεπάγεται την ύπαρξη ψωμιού μέσα στο καλάθι προϊόντων. Πρόκειται για μία αρκετά σημαντική πληροφορία η οποία όμως αποτελεί την μία όψη του νομίσματος. Τι σημαίνει αυτό; Ότι εξετάζουμε την συσχέτιση με θετικό τρόπο. Για το λόγο αυτό οι κανόνες της μορφής X→Y θα ονομάζονται στο εξής θετικοί κανόνες (Positive rules) [23].

Η άλλη όψη του νομίσματος έχει να κάνει με έναν διαφορετικό τρόπο συσχέτισης συνόλων αντικειμένων. Πιο συγκεκριμένα πρέπει να εξεταστεί κατά πόσο η ύπαρξη κάποιων συνόλων αντικειμένων σε μια δοσοληψία συνεπάγεται την απουσία κάποιων άλλων. Η συσχέτιση αυτή περιγράφεται από κανόνες της μορφής X→⌐Y όπου με το σύμβολο ⌐Y υποδηλώνουμε την απουσία του Y. Για να επιστρέψουμε στο παράδειγμα μας με το καλάθι αγορών ο κανόνας γάλα→⌐μπύρα αντιπροσωπεύει τους πελάτες που αγοράζουν γάλα και ταυτόχρονα δεν αγοράζουν μπύρα. Κάτι αντίστοιχο συμβαίνει και με τον ανάποδο τρόπο συσχέτισης, δηλαδή κατά πόσο η απουσία κάποιων αντικειμένων συνεπάγεται την ύπαρξη κάποιων άλλων μέσα στις συναλλαγές(transactions). Οι κανόνες εδώ είναι της μορφής ⌐X→Y. Οι δύο αυτές μορφές κανόνων, οι οποίες εμπεριέχουν με τον ένα ή τον άλλο τρόπο την ‘απουσία’ αντικειμένων, εξετάζουν την συσχέτιση με αρνητικό τρόπο. Για το λόγο αυτό οι κανόνες αυτοί θα ονομάζονται στο εξής αρνητικοί κανόνες (negative rules)[23].

3.3 Μεθοδολογία ανακάλυψης κανόνων συσχέτισης και ο συντελεστής συσχέτισης

Η μέχρι τώρα ανάλυση στην εύρεση κανόνων στηρίζεται στο ‘’δίδυμο’’ τιμών υποστήριξης-εμπιστοσύνης. Ανάλογα με το πόσο αυστηρά είναι τα κατώφλια (thresholds) που θα βάλουμε και για τις δύο τιμές, έχουμε και το ‘‘κόψιμο’’ υποψήφιων κανόνων από τα σύνολα συχνών αντικειμένων.

Αρχικά θεωρούμε μια μεγάλη βάση δεδομένων ενός συνόλου συναλλαγών. Κάθε συναλλαγή είναι μια λίστα από αντικείμενα – items (για παράδειγμα μια αγοραπωλησία ενός πελάτη). Στη συνέχεια βρίσκουμε όλους τους κανόνες οι οποίοι συσχετίζουν τα δεδομένα που παρουσιάζονται στο ένα σύνολο αντικειμένων με τα δεδομένα που παρουσιάζονται στο άλλο σύνολο δεδομένων. Δεν υπάρχουν περιορισμοί ως προς τον αριθμό των αντικειμένων στην υποστήριξη ή στην εμπιστοσύνη ενός κανόνα.

Παρόλα αυτά υπάρχουν περιπτώσεις όπου ακόμα και με αυστηρά κριτήρια, προκύπτουν σύνολα που δεν παρουσιάζουν μεγάλο ενδιαφέρον. Αυτό ακριβώς είναι που οδήγησε και σε έναν επιπλέον έλεγχο, χρησιμοποιώντας μια τιμή από το χώρο της στατιστικής. Αυτή η τιμή είναι ο συντελεστής συσχέτισης (Correlation Coefficient) [24].

Ο συντελεστής συσχέτισης είναι ένα μέγεθος που χρησιμοποιείται για να μετρήσουμε πόσο μεγάλη είναι η συσχέτιση μεταξύ δύο μεταβλητών. Έτσι αν έχουμε δύο μεταβλητές X, Y το μέτρο του συντελεστή δίνεται από την σχέση: [23]

$$\rho = \frac{cov(X,Y)}{\sigma X\, \sigma Y} \quad (1)$$

Όπου ο $cov(X,Y)$ αντιπροσωπεύει την συνδιασπορά των δύο μεταβλητών, ενώ $\sigma X\, \sigma Y$ αποτελούν τις τυπικές αποκλίσεις για τα X, Y αντίστοιχα.

Το εύρος τιμών του συντελεστή είναι στο διάστημα $[-1,1]$ με χαρακτηριστικές τιμές τις εξής:

•Τιμή μηδέν (0) σημαίνει ότι οι δύο μεταβλητές είναι ανεξάρτητες μεταξύ τους, δηλαδή δεν επηρεάζεται καθόλου η μία από την άλλη.

•Τιμή ένα (1) σημαίνει ότι δύο μεταβλητές έχουν τη μεγαλύτερη δυνατή εξάρτηση μεταξύ τους και μάλιστα ευθέως ανάλογη. Αυτό σημαίνει ότι όσο αυξάνεται / μειώνεται η X κατά το ίδιο ποσοστό αυξάνεται / μειώνεται και η Y.

•Τιμή μείον ένα (-1) σημαίνει και πάλι ότι οι δύο μεταβλητές έχουν τη μέγιστη δυνατή εξάρτηση μεταξύ τους. Αυτή τη φορά όμως είναι αντιστρόφως ανάλογη η αλλαγή στις τιμές τους. Δηλαδή όσο αυξάνεται/ μειώνεται η X τόσο μειώνεται/ αυξάνεται η Y.

Όπως προαναφέρθηκε το μέγεθος ρ αποτελεί στατιστικό μέγεθος και ως τέτοιο η τιμή του υπολογίζεται μέσα από άλλες στατιστικές τιμές(1). Επειδή όμως η στατιστική προσέγγιση ξεφεύγει από τα όρια και τους στόχους αυτής εργασίας πρέπει να βρούμε τρόπο υπολογισμού του συντελεστή

συσχέτισης μέσα από τις εμφανίσεις των X, Y. Εκτός από την παραπάνω στατιστική προσέγγιση είναι εφικτός ο τρόπος υπολογισμού του συντελεστή συσχέτισης μέσω του αριθμού εμφάνισης των X, Y [27]:

	Y	⌐Y	Σ
X	f11	f10	f1+
⌐X	f01	f00	f0+
Σ	f+1	f+0	N

Προτού δώσουμε τον ορισμό του συντελεστή συσχέτισης συναρτήσει των τιμών του πίνακα θα επεξηγήσουμε τι αντιπροσωπεύουν οι τιμές αυτές:

•Το f11 είναι ο αριθμός των εμφανίσεων του XY μέσα στις δοσοληψίες.

•Το f10 είναι ο αριθμός των εμφανίσεων του X⌐Y μέσα στις δοσοληψίες. (Υπενθυμίζουμε ότι το X⌐Y είναι ο αριθμός των δοσοληψιών που περιέχουν το X και δεν περιέχουν το Y).

•Το f00 αναφέρεται στον αριθμό των δοσοληψιών που δεν περιέχουν ούτε το X αλλά ούτε και το Y (⌐X⌐Y).

Από αυτά συνάγεται ότι η τιμή f1+ αποτελεί το άθροισμα των παραπάνω τιμών και αντιστοιχεί στον αριθμό των εμφανίσεων του X μέσα στις δοσοληψίες.

Η τιμή f0+ είναι το άθροισμα των παραπάνω τιμών και αντιστοιχεί στον αριθμό των εμφανίσεων των δοσοληψιών που δεν εμπεριέχουν το X

Αντιστοίχως έχουμε:

•Το f+1 είναι ο αριθμός των εμφανίσεων του Y μέσα στις δοσοληψίες

•Το f+0 είναι ο αριθμός των δοσοληψιών στα οποία δεν εμφανίζεταιο Y.

•Το N είναι ο συνολικός αριθμός των δοσοληψιών ο οποίος όπως φαίνεται από τον πίνακα αλλά και προκύπτει από την δυαδική λογική είναι ίσος με:

- Το άθροισμα f1++f0+. Είναι το άθροισμα των δοσοληψιών στα οποία εμφανίζεται το X και αυτών στα οποία δεν εμφανίζεται το X.
- Το άθροισμα f+1+f+0. Είναι το άθροισμα των δοσοληψιών στα οποία εμφανίζεται το Y και αυτών στα οποία δεν εμφανίζεται το Y.

Με βάση τις παραπάνω δίνουμε τον νέο ορισμό του συντελεστή συσχέτισης ως εξής [24],[25],[26],[27]:

$$\frac{(f11\, f00 \;+\; f10\, f01\,)}{sqrt\,[\,f\,+\,0\,f\,+\,1\,f1\,+\;f0+]}$$

3.4 Εφαρμογές κανόνων συσχέτισης

Ας θεωρήσουμε λοιπόν ότι βρισκόμαστε σε ένα σούπερ-μάρκετ και στο οποίο υπάρχει μια μεγάλη συλλογή από αντικείμενα (items). Κάθε επιχειρηματική απόφαση που πρέπει να παρθεί με στόχο την σωστή διαχείριση της επιχείρησης βασίζεται στην σωστή μελέτη των προϊόντων που διατίθενται προς πώληση, στην σωστή τοποθέτηση των εμπορευμάτων στα ράφια και πολλά άλλα [19].

Η ανάλυση των πιο πρόσφατων συναλλαγών δεδομένων είναι μια ευρέως χρησιμοποιούμενη προσέγγιση προκειμένου να παρθούν οι σωστές αποφάσεις. Μέχρι πρόσφατα, ωστόσο, μόνο τα στοιχεία σχετικά με τις συνολικές πωλήσεις σε συγκεκριμένο χρονικό διάστημα (μία μέρα, μία εβδομάδα, το μήνα, κλπ.) ήταν διαθέσιμα και προσβάσιμα από υπολογιστή. Με την τεχνολογία bar - code κατέστη δυνατή η αποθήκευση των λεγόμενων basket-data όπου συγκεκριμένα αποθηκεύονται τα αντικείμενα που αγοράστηκαν βασιζόμενοι σε κάθε συναλλαγή [19].

Οι συναλλαγές του τύπου basket-data δεν περιλαμβάνουν απαραίτητα αντικείμενα τα οποία κατ' ανάγκην αγοράστηκαν την ίδια χρονική περίοδο. Μπορεί να αποτελούνται και από αντικείμενα που αγοράζονται από κάποιον πελάτη κατά τη διάρκεια μιας χρονικής περιόδου. Τα παραδείγματα περιλαμβάνουν μηνιαίες αγορές από τα μέλη μιας λέσχης βιβλίου ή μιας επιχείρησης.

Πολλές οργανώσεις έχουν συλλέξει τεράστιες ποσότητες των δεδομένων αυτών. Αυτά τα σύνολα δεδομένων συνήθως αποθηκεύονται σε αποθήκες τριτογενούς μορφής και είναι πολύ δύσκολο να μεταφερθούν σε μεγάλα συστήματα βάσεων δεδομένων.

Ένας από τους βασικούς λόγους για την περιορισμένη επιτυχία των συστημάτων των βάσεων δεδομένων στον τομέα αυτό είναι ότι τα τρέχοντα συστήματα διαχείρισης βάσεων δεδομένων, δεν παρέχουν την απαραίτητη λειτουργικότητα για ένα χρήστη που ενδιαφέρεται να επωφεληθεί από αυτές τις πληροφορίες.

Το πρόβλημα το οποίο λοιπόν προκύπτει είναι το ακόλουθο: η εξόρυξη δεδομένων και πληροφοριών από μια μεγάλη συλλογή συναλλαγών basket-data με σκοπό την εύρεση κανόνων συσχέτισης μεταξύ ενός συνόλου αντικειμένων.

Ένα παράδειγμα ενός κανόνα συσχέτισης αποτελεί μπορεί να δηλωθεί το εξής[19]: {ότι το 90% των συναλλαγών που αγοράζουν ψωμί και βούτυρο }=>{αγοράζουν γάλα}.

Στον κανόνα αυτό το ψωμί και το βούτυρο είναι τα προηγηθέντα προϊόντα, ενώ το γάλα το συνεπακόλουθο προϊόν. Επίσης, η εμπιστοσύνη του κανόνα είναι 90% και εκφράζει την δύναμη του κανόνα (powerset). Στην περίπτωση αυτή, εκφράζει το ποσοστό των συναλλαγών που περιλαμβάνουν γάλα, δεδομένου ότι περιλαμβάνουν ψωμί και βούτυρο.

Οι Agrawal, Imielinski και Swami, χρησιμοποίησαν κανόνες συσχέτισης για την ανακάλυψη ομοιοτήτων μεταξύ προϊόντων σε μεγάλες βάσεις δεδομένων. Οι βάσεις δεδομένων αυτές

περιλάμβαναν τεράστιους όγκους δεδομένων από συναλλαγές πελατών, σε διάφορα σημεία πώλησης προϊόντων όπως υπεραγορές (supermarkets) [19].

Ακολουθεί ένα παράδειγμα του οποίου η δυναμική έγκειται στην βελτίωση των βάσεων δεδομένων και στην σωστή επεξεργασία των διαφόρων ερωτηματολογίων. Έστω λοιπόν ότι βρισκόμαστε σε ένα σούπερ-μάρκετ και ακολουθούμε τα παρακάτω βήματα:

Βήμα 1^{ο}. Να βρούμε όλους τους κανόνες οι οποίοι περιέχουν τη λέξη «Ζάχαρη» σαν συνεπακόλουθο προϊόν. Αυτοί οι κανόνες θα μπορούσαν να προωθήσουν τις πωλήσεις του καταστήματος σε «Ζάχαρη»

Βήμα 2^{ο}. Να βρούμε όλους τους κανόνες που περιέχουν τη λέξη «καφές» σαν προηγηθέν προϊόν. Αυτοί οι κανόνες θα μπορέσουν να καθορίσουν οι πωλήσεις ποιων προϊόντων θα επηρεαστούν αν το κατάστημα ελαττώσει σε μεγάλο βαθμό τις πωλήσεις των «καφές».

Βήμα 3^{ο}. Να βρούμε όλους τους κανόνες που περιέχουν σαν προηγηθέν προϊόν τη λέξη «Μπιφτέκια» και σαν επακόλουθο προϊόν τη λέξη «κέτσαπ».

Αυτού του είδους το ερώτημα μπορεί να διατυπωθεί εναλλακτικά σαν ένα ερώτημα για τα επιπρόσθετα προϊόντα τα οποία θα πρέπει να πωληθούν μαζί με τα «Μπιφτέκια» ώστε να είναι πολύ πιθανό ένα από αυτά να είναι και η «κέτσαπ».

Βήμα 4^{ο}. Να βρούμε όλους τους κανόνες συσχέτισης των προϊόντων τα οποία βρίσκονται στα ράφια Α και Β στο κατάστημα.

Οι κανόνες αυτοί μπορούν να βοηθήσουν στον καθορισμό του αν οι πωλήσεις του αντικειμένου στο ράφι Α σχετίζονται με τις πωλήσεις του αντικειμένου στο ράφι Β.

Βήμα 5^{ο}. Να βρούμε τους καλύτερους k κανόνες, που έχουν κάποιο προϊόν (π.χ. ζάχαρη) σαν συνεπακόλουθο προϊόν. Ο όρος καλύτερος κανόνας αναφέρεται στον κανόνα με το μεγαλύτερο ποσοστό εμπιστοσύνης ή την μεγαλύτερη συχνότητα εμφάνισης.

Κάνουμε χρήση μόνο αυτών των κανόνων οι οποίοι είναι και οι πλέον κατάλληλοι για λήψη αποφάσεων, αφού οι κανόνες αυτοί είναι οι πιο «βάσιμοι» και χαρακτηρίζουν το μεγαλύτερο μέρος των συναλλαγών (εφόσον είναι οι πιο συχνά εμφανιζόμενοι).

3.5 Αλγόριθμοι συσχετίσεων

Ο πιο διαδεδομένος αλγόριθμος της κατηγορίας αυτής καλείται **αλγόριθμος A-priori** και ο οποίος αφορά στην γνώση και εύρεση των κανόνων συσχέτισης. Έχει σχεδιαστεί με τέτοιο τρόπο ώστε να μπορεί να εφαρμοστεί σε βάσεις δεδομένων οι οποίες περιέχουν συναλλαγές (όπως για παράδειγμα σύνολα προϊόντων που αγοράστηκαν από πελάτες ή λεπτομέρειες για την συχνή επίσκεψη σε έναν δικτυακό τόπο). Άλλοι αλγόριθμοι είναι σχεδιασμένοι για να εφαρμόζονται, για την

εύρεση κανόνων συσχέτισης, σε δεδομένα στα οποία δεν εμπλέκεται κανενός είδους συναλλαγή ή σε δεδομένα στα οποία δεν υπάρχει συγκεκριμένο χρονοδιάγραμμα για παράδειγμα στην αλληλουχία DNA [19].

Όπως είναι σύνηθες στην εξόρυξη κανόνων συσχέτισης, δίνεται ένα δεδομένο σύνολο από itemsets[1], (για παράδειγμα σύνολα από πωλήσεις λιανικής όπου σε καθεμία δημιουργείται μια μεμονωμένη λίστα με τα προϊόντα τα οποία αγοράστηκαν) και σε αυτό το σύνολο ο αλγόριθμος επιχειρεί να βρει όλα τα υποσύνολα τα οποία είναι κοινά σε τουλάχιστον ένα ελάχιστο αριθμό c των itemsets.

3.5.1 Ο αλγόριθμος A-Priori

Ο A-priori προτάθηκε από τους R .Agrawal και R .Srikant το 1994 [19],[27],[29]. Στόχος τους ήταν η εξόρυξη των συνήθεις (frequent) itemsets για Boolean κανόνες συσχέτισης. Το όνομα βασίζεται στο γεγονός ότι ο αλγόριθμος χρησιμοποιεί ''προηγούμενη γνώση'' (prior knowledge) ιδιοτήτων των frequent itemsets. Επίσης, ο αλγόριθμος χρησιμοποιεί την προσέγγιση level-wise search, κατά την οποία k-itemsets χρησιμοποιούνται για την εύρεση k+1-itemsets. Πρόκειται για τον βασικό αλγόριθμο για εύρεση frequent itemsets. Τα συχνά itemsets μας είναι χρήσιμα, επειδή από αυτά προκύπτουν με κατάλληλες μεθόδους οι κανόνες συσχέτισης.

Το πρόβλημα της εύρεσης των κανόνων συσχέτισης που έχουν την επιθυμητή επιβεβαίωση και αξιοπιστία μπορεί να διαιρεθεί σε δυο υπο-προβλήματα:

- Εύρεση όλων των συνδυασμών των προϊόντων που έχουν επιβεβαίωση πάνω από την ελάχιστη επιβεβαίωση (minimum support). Όλοι αυτοί οι συνδυασμοί ονομάζονται μεγάλες λίστες από προϊόντα (largeitemsets) και όλοι οι υπόλοιποι συνδυασμοί μικρές λίστες από προϊόντα (small itemsets).

- Χρήση όλων των μεγάλων λιστών από προϊόντα για εξόρυξη των κανόνων συσχέτισης που ικανοποιούν την ελάχιστη αξιοπιστία. Για παράδειγμα, έστω τα ABCD και AB είναι μεγάλες λίστες από προϊόντα. Μπορούμε να καθορίσουμε αν ο κανόνας συσχέτισης AB=>CD ξεπερνά την ελάχιστη αξιοπιστία, υπολογίζοντας τη σχέση:

$$r = \frac{\upsilon\pi o\sigma\tau\acute{\eta}\rho\iota\xi\eta\ (ABCD)}{\upsilon\pi o\sigma\tau\acute{\eta}\rho\iota\xi\eta(AB)}$$

Αν $r \geq$ ελάχιστης αξιοπιστίας, τότε ο κανόνας συσχέτισης γίνεται αποδεκτός.

- Βρίσκουμε τα αγαθά που εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση (minimum support), δηλαδή το σύνολο L_1 =μεγάλες λίστες από 1 – αγαθό (large 1 itemsets)

[1] Itemsets: Ένα σύνολο αντικειμένων – κανόνων καλείται itemset. Κάθε Itemset αποτελείται από k-αντικείμενα και καλείται k-itemset. Ένα itemset μπορεί επίσης να θεωρηθεί και ένας συνδυασμός από αντικείμενα- items.

- Από κ=2 και όσο L_{k-1} δεν είναι κενό βρες το σύνολο Ck όλων των υποψήφιων μεγάλων λιστών από κ- αγαθά (candidatelarge k-itemsets) με βάση το L_{k-1} Βρες ποια από αυτά εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση και φτιάξε το σύνολο L_k = μεγάλες λίστες από κ- αγαθά.

- Για κάθε στοιχείο των $L_1, L_2, L_3, \dots, L_n$ βρες ποια ικανοποιούν την ελάχιστη αξιοπιστία (minimum confidence).

Στην συνέχεια χρησιμοποιούμε αυτά τα frequent itemsets για την παραγωγή κανόνων συσχέτισης.

Βήμα 1: joining (ένωση)

Όπως είπαμε και προηγουμένως, το σύνολο L_k θα βρεθεί από το σύνολο L_{k-1} . Για να βρεθεί το L_k, πρέπει πρώτα να βρεθεί το σύνολο C_k των υποψήφιων itemsets. Το σύνολο Ck θα βρεθεί αφού συνενωθεί (joined) το L_k με τον εαυτό του.

$L_k - 1$. join L_{k-1} .: Γίνεται μόνο στα itemsets του L_{k-1} . που είναι joinable.

Joinable: Δύο $k-1$-itemsets είναι joinable, εάν τα items τους μέχρι και το $k-2$ είναι τα ίδια.

Το k-itemset αυτό αποτελεί υποψήφιο k-itemset για το σύνολο L_k και θα μπει στο σύνολο C_k των υποψήφιων k-itemsets.

Παράδειγμα για $k = 4$: Για να βρούμε το L_4 από το L_3 θα κάνουμε την πράξη $L_3\ join\ L_3$: για κάθε itemset στο L_3 κοιτάμε αν είναι joinable με κάθε άλλο itemset στο L_3 (εκτός τον εαυτό του), δηλαδή αν έχουν τα ίδια items στις πρώτες 2 θέσεις. Αν δύο itemsets είναι joinable, τότε γίνονται joined και το 4-itemset που προκύπτει εισάγεται στο σύνολο C_4, αφού αποτελεί υποψήφιο 4-itemset για το L_4.

Βήμα 2: pruning (κλάδεμα)

Το σύνολο C_k από k-itemsets είναι υπερσύνολο του L_k, το οποίο θέλουμε να βρούμε. Ισχύει ότι: τα k-itemsets-μέλη του C_k μπορεί να είναι ή να μην είναι συχνά εμφανιζόμενα, αλλά όλα τα συχνά εμφανιζόμενα k-itemsets συμπεριλαμβάνονται στο C_k. Το ποια από τα itemsets αυτά είναι συχνά εμφανιζόμενα και ποια όχι, καθορίζεται με μια αναζήτηση στην Βάση Δεδομένων για κάθε itemset και μέτρηση του πλήθους των εμφανίσεων του κάθε ενός από αυτά. Συχνά εμφανιζόμενα είναι μόνο τα itemsets που έχουν αριθμό εμφανίσεων μεγαλύτερο ή ίσο του minimum support.

Επειδή το C μπορεί να είναι τεράστιο, γίνεται μια διαδικασία ελαχιστοποίησης των itemsets του. Χρησιμοποιείται η A-priori ιδιότητα:

«Οποιοδήποτε L_{k-1} itemset δεν είναι συχνά εμφανιζόμενο, δεν μπορεί να είναι υποσύνολο κάποιου συχνά εμφανιζομένου k-itemset»

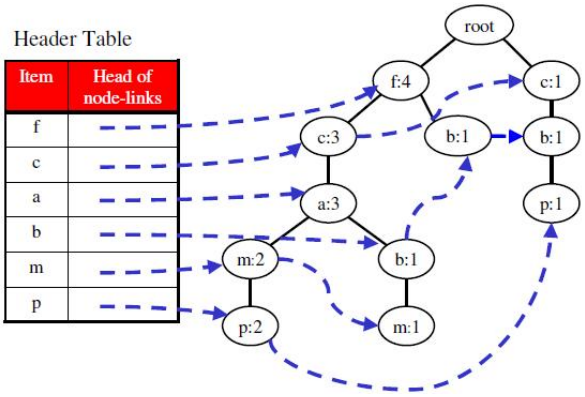
Έτσι, αν οποιοδήποτε από τα $k-1$-itemsets κάποιου υποψήφιου k-itemset στο C_k δεν υπάρχει στο L_{k-1} τότε το υποψήφιο k-itemset δεν μπορεί να είναι συχνά εμφανιζόμενο και αφαιρείται από το C_k. Με τον τρόπο αυτό, για κάθε τέτοιο k-itemset (που σίγουρα δεν είναι συχνά εμφανιζόμενο), γλιτώνουμε την προσπέλαση στον πίνακα συναλλαγών και την μέτρηση του αριθμού των εμφανίσεων του, δηλαδή τις πράξεις που θα κάναμε για να δούμε αν θα ικανοποιούσε το minimum support [28], [29].

Να τονίσουμε ότι με την διαδικασία κλαδέματος δεν προκύπτουν τα συχνά εμφανιζόμενα k-itemsets, αλλά τα k-itemsets που αποκλείεται να είναι συχνά εμφανιζόμενα. Έτσι, μετά το κλάδεμα, χρειάζεται η προσπέλαση της βάσης δεδομένων για κάθε ένα από τα εναπομείναντα itemsets στο C_k, ώστε να καθοριστεί αν ικανοποιούν το minimum support. Μόνο αν το ικανοποιούν θα είναι συχνά εμφανιζόμενα itemsets [19].

3.5.2 Ο αλγόριθμος FP-Growth

Ο αλγόριθμος FP-Growth (Frequent Pattern Growth) είναι ένας εναλλακτικός αλγόριθμος εξόρυξης κανόνων συσχέτισης, συνήθως γρηγορότερο από τον A-priori. Ο αλγόριθμος χρησιμοποιεί ένα προθεματικό δένδρο αναπαράστασης της συλλογής που λέγεται FP-δένδρο και έτσι καταφέρνει να ανακτά γρηγορότερα τις εγγραφές των συνόλων. Για την δημιουργία του FP-δένδρου, αφαιρεί από το σύνολο των εγγράφων όλα τα αντικείμενα που δεν είναι συχνά και αναδιατάσσει τα υπόλοιπα σε φθίνουσα διάταξη ως προς την τιμή της υποστήριξής τους. Στη συνέχεια εισάγει τις αναδιαταγμένες εγγραφές στο προθεματικό δένδρο, όπου κοινά προθέματα αναπαριστώνται με το ίδιο μονοπάτι. Τέλος, για την διευκόλυνση της αναζήτησης δημιουργεί συνδεδεμένες λίστες για κάθε αντικείμενο του δένδρου. Παράδειγμα ενός FP-δένδρου φαίνεται στην εικόνα 4 που έχει δημιουργηθεί βάση του παρακάτω πίνακα.

Items	Frequent Items
f, a, c, d, g, i, m, p	*f, c, a, m, p*
a, b, c, f, l, m, o	*f, c, a, b, m*
b, f, h, j, o	*f, b*
b, c, k, s, p	*c, b, p*
a, f, c, e, l , p, m, n	*f, c, a, m, p*



Εικόνα 5 Το FP-δένδρο αφού έχουν εισαχθεί όλες οι εγγραφές

Μόλις δημιουργηθεί το FP-δένδρο ο αλγόριθμος λειτουργεί αναδρομικά χρησιμοποιώντας το δένδρο για να βρει τα συχνά σύνολα. Αρχικά εξετάζει ξεχωριστά κάθε αντικείμενο κατά αύξουσα σειρά υποστήριξης και από την αντίστοιχη συνδεδεμένη λίστα βρίσκει τους κόμβους όπου εμφανίζεται το αντικείμενο αυτό, για να ακολουθήσει το μονοπάτι προς αυτούς τους κόμβους. Για κάθε μονοπάτι δίνει μια υποστήριξη ίση με τον αντίστοιχο αριθμό που έχει ο κόμβος στο FP-δένδρο. Στη συνέχεια συνδυάζει το υπό εξέταση αντικείμενο με τα μονοπάτια του για να βρει τα συχνά σύνολα.

4. Ταξινόμηση (classification)

4.1 Πως λειτουργεί η ταξινόμηση

Η ταξινόμηση ή κατηγοριοποίηση (classification) είναι η πιο γνωστή και πιο δημοφιλής λειτουργία εξόρυξης γνώσης. Πολλές εταιρίες του ιδιωτικού και του δημόσιου τομέα χρησιμοποιούν σε καθημερινή βάση συστήματα ταξινόμησης. Παραδείγματα τέτοιου είδους συστημάτων είναι τα συστήματα αναγνώρισης προτύπων, συστήματα ιατρικών διαγνώσεων, συστήματα έγκρισης δανείων και πιστωτικών καρτών, συστήματα ανίχνευσης λαθών σε βιομηχανικές εφαρμογές, συστήματα ταξινόμησης των τάσεων στην οικονομία κ.α. Για παράδειγμα όταν κάποιος προβλέπει μια ηλικία, στην ουσία επιλύει ένα πρόβλημα κατηγοριοποίησης**.**

Η βασική εργασία της ταξινόμησης είναι η δημιουργία ενός μοντέλου το οποίο θα χρησιμοποιείται για να κατηγοριοποιεί δεδομένα τα οποία δεν έχουμε κατηγοριοποιήσει. Η ταξινόμηση είναι μια διαδικασία δύο βημάτων:

1. Εκμάθηση (Learning). Με χρήση ενός μέρους των δεδομένων μας, τα οποία ονομάζονται δεδομένα εκπαίδευσης (training data), χτίζουμε ένα μοντέλο περιγράφοντας ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων.

2. Ταξινόμηση. Έχοντας το μοντέλο που προέκυψε από το προηγούμενο βήμα προσπαθούμε με χρήση δοκιμαστικών παραδειγμάτων (test samples) να επιβεβαιώσουμε την ακρίβεια του. Αν έχει τελικά μια αποδεκτή ακρίβεια τότε θα χρησιμοποιηθεί για την κατηγοριοποίηση νέων δεδομένων αλλά και δεδομένων τα οποία δεν ανήκουν σε κάποια κατηγορία.

Οι πιο γνωστές μέθοδοι ταξινόμησης είναι:

- Τα δέντρα απόφασης (decision trees) [31].
- Η μάθηση κατά Bayes(απλός / Naïve ταξινομητής Bayes).
- Τα νευρωνικά δίκτυα (neural networks).
- Η κατηγοριοποίηση κοντινότερων γειτόνων (nearest neighbor) [32],[33]
- Τα Support Vector Machines [34].

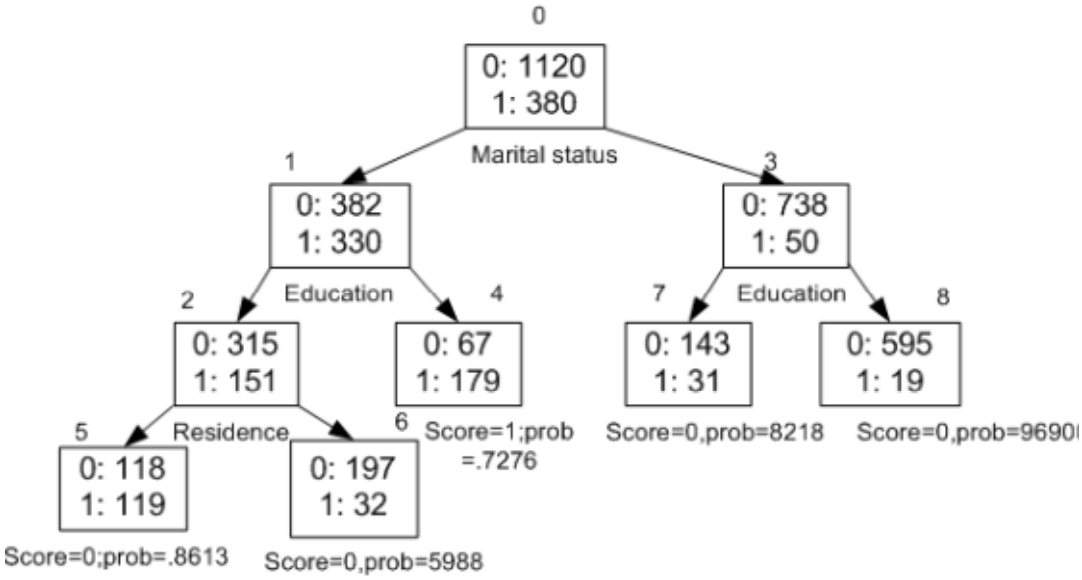
Αυτό που καλείται εποπτευμένη εκμάθηση είναι ο πιο θεμελιώδης στόχος στην εκμάθηση μηχανών. Στην εποπτευμένη εκμάθηση, έχουμε τα παραδείγματα κατάρτισης / εκπαίδευσης (training examples) και τα παραδείγματα δοκιμής (test examples). Ένα παράδειγμα κατάρτισης είναι ένα διαταγμένο ζευγάρι (x, y) όπου το x είναι μια περίπτωση (instance) και το y είναι μια τιμή.Ένα παράδειγμα δοκιμής είναι μια περίπτωση x με άγνωστη τιμή. Ο στόχος είναι να προβλεφθούν οι τιμές για τα παραδείγματα δοκιμής. Το όνομα «επόπτευση» προέρχεται από το γεγονός ότι κάποιος επόπτης έχει παράσχει τις τιμές για τα παραδείγματα κατάρτισης [43].

Υποθέτουμε ότι οι περιπτώσεις x είναι μέλη του συνόλου x, ενώ οι τιμές y είναι μέλη του συνόλου y. Κατόπιν ένας ταξινομητής είναι η οποιαδήποτε συνάρτηση: $f : x \to y$. Ένας εποπτευμένος αλγόριθμος μάθησης δεν θεωρείτε ταξινομητής. Αντιθέτως, είναι τα αποτελέσματα του αλγορίθμου είναι ο ταξινομητής. Μαθηματικά, ο εποπτευμένος αλγόριθμος εκμάθησης εκφράζεται με με μια συνάρτηση τύπου: $(x \times y)^n \to (x \to y)$ όπου το v είναι ο βασικός αριθμός (cardinality) του καταρτιζόμενου συνόλου.

4.2 Μέθοδοι κατηγοριοποίησης

4.2.1 Δέντρα απόφασης (decision trees)

Τα δένδρα ταξινόμησης / απόφασης είναι εργαλεία που βοηθάνε στην λήψη μιας απόφασης ή στην προσέγγιση μιας συνάρτησης που έχει ως έξοδο κάποια διακριτή τιμή. Στην μηχανική μάθηση ένα δένδρο ταξινόμησης είναι ένα μοντέλο πρόβλεψης που αντιστοιχεί τις παρατηρήσεις για κάποιο αντικείμενο με την έξοδο. Σε αυτά τα δένδρα τα φύλλα αναπαριστούν τις διακριτές ταξινομήσεις (εξόδους) του δένδρου ενώ κάθε κόμβος του δένδρου ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού. Κάθε κλαδί που ξεκινάει από ένα κόμβο αντιστοιχεί σε διαφορετική τιμή του χαρακτηριστικού που ελέγχει ο κόμβος και οδηγεί σε διαφορετικό φύλλο (έξοδο) [31].



Εικόνα 6 Δένδρο Απόφασης [20]

Ένας από τους λόγους που είναι δημοφιλής αυτή η τεχνική μάθησης είναι η ευκολία αναπαράστασης αλλά και κατανόησης της τεχνικής. Σημαντικό στοιχείο στα δένδρα ταξινόμησης είναι το πόσο εύκολα μπορεί να διαβαστεί και επεξηγηθεί το αποτέλεσμα που δίνει. Συνήθως τα αποτελέσματα των δένδρων ταξινόμησης μπορούν να αναπαρασταθούν με απλά μαθηματικά ή με ένα σύνολο κανόνων if-then και έτσι μπορούν εξηγήσουν γιατί δόθηκε το αποτέλεσμα που δόθηκε.

Ο πιο γνωστός αλγόριθμος της τεχνικής των δένδρων ταξινόμησης είναι ο ID3. Ο αλγόριθμος μπορεί να περιγραφεί σε τρία βήματα:

1. Βρες την ανεξάρτητη μεταβλητή η οποία αν χρησιμοποιηθεί ως κριτήριο διαχωρισμού των δεδομένων εκπαίδευσης θα οδηγήσει σε κόμβους κατά το δυνατό διαφορετικούς σε σχέση με την εξαρτημένη μεταβλητή.

Ο αλγόριθμος αυτός χρησιμοποιεί σαν κριτήριο επιλογής την εντροπία, η οποία παρέχει μία εκτίμηση όσον αφορά το βαθμό του σφάλματος που επιτελείται κάθε φορά κατά το χωρισμό του συνόλου εκπαίδευσης, βάσει του συγκεκριμένου πεδίου. Η εντροπία είναι ένα μέγεθος που χρησιμοποιείται στη Θεωρία της Πληροφορίας και έχει αρχικά προταθεί από τον Shannon.
Η εντροπία μπορεί να δοθεί από την εξίσωση:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_b p(xi)$$

όπου b είναι η βάση του λογάριθμου. Έτσι, το πεδίο με τη μικρότερη εντροπία χωρίζει καλύτερα το σύνολο εκπαίδευσης [48].

2. Κάνε τον διαχωρισμό

3. Επανέλαβε τη διαδικασία για κάθε έναν από τους κόμβους που προέκυψαν μέχρι να μην είναι δυνατός περαιτέρω διαχωρισμός.

Με αυτά τα βήματα κατασκευάζεται ένα δένδρο «άπληστα» βάσει ενός μηχανισμού διαχωρισμού. Συνήθως ο μηχανισμός διαχωρισμού είναι αυτός της εντροπίας της πληροφορίας ο οποίος επιλέγει εκείνη την ανεξάρτητη μεταβλητή που οδηγεί στο πιο συμπαγές δένδρο. Ο αλγόριθμος είναι ευριστικός γιατί δεν εγγυάται ότι το δένδρο που παράγεται είναι το μικρότερο δυνατό [31].

4.2.2 Μάθηση κατά Bayes

Η μάθηση κατά Bayes είναι μια από τις σημαντικότερες και τεχνικές μηχανικής μάθησης. Η διαφορά με τις άλλες τεχνικές είναι ότι βασίζεται σε μια διαφορετική εκδοχή του τι σημαίνει «μαθαίνω» από τα δεδομένα, χρησιμοποιεί πιθανότητες για να εκφράσει την αβεβαιότητα στην συνάρτηση που μαθαίνει. Η μάθηση δεν μπορεί να ξεκινήσει από το τίποτα, συνεπώς γίνεται κάποιου είδος παραδοχή για το μοντέλο που πρόκειται να «μάθουμε», με άλλα λόγια απαιτούνται κάποιες εκ των προτέρων πιθανότητες. Κατά την μάθηση αυτές οι πιθανότητες αναθεωρούνται και έτσι κάθε παράδειγμα μπορεί να μείωση ή να αύξηση την πιθανότητα μια υπόθεση να είναι σωστή. Αυτό δίνει την ευκαιρία στους αλγορίθμους που μαθαίνουν κατά Bayes να μην απορρίπτουν μια υπόθεση εάν δεν είναι σύμφωνη με τα παραδείγματα εκπαίδευσης. Αρνητικό στοιχείο της τεχνικής είναι η απαίτηση για την γνώση πολλών

τιμών πιθανοτήτων και όταν αυτές δεν είναι δυνατό να υπολογιστούν, υπολογίζονται κατ' εκτίμηση από παλιότερες υποθέσεις ή από εμπειρική γνώση [31].

Μια απλουστευμένη εκδοχή της κατά Bayes μάθησης είναι ο απλός (Naïve) ταξινομητής Bayes όπου θεωρούνται ότι όλα τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί ότι είναι ένα μήλο αν είναι κόκκινο, στρογγυλό και έχει διάμετρο περίπου 4 εκατοστά. Ακόμη και αν τα χαρακτηριστικά αυτά εξαρτώνται από την ύπαρξη των άλλων χαρακτηριστικών, ένας απλός ταξινομητής Bayes θεωρεί ότι όλα αυτά τα χαρακτηριστικά συμβάλουν ανεξάρτητα στην πιθανότητα ότι το φρούτο είναι ένα μήλο. Η ποσότητα P που περιγράφει έναν απλό ταξινομητή Bayes για ένα σύνολο παραδειγμάτων εκφράζει την πιθανότητα να είναι c η τιμή της εξαρτημένης μεταβλητής c με βάση τις τιμές $\chi = (X1, X2, \dots Xn)$των χαρακτηριστικών $X = (X1, X2, \dots Xn)$)και δίνεται από τη σχέση:$P(c|x) = P(c)\prod_i P(x_i|c)$

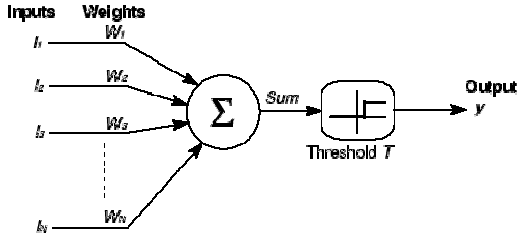
Όπου τα χαρακτηριστικά X_iθεωρούνται ανεξάρτητα μεταξύ τους. Ο υπολογισμός της παραπάνω ποσότητας για ένα σύνολο N παραδειγμάτων γίνεται με βάση τις σχέσεις:

- $P(c) = \frac{N(c)}{N}$
- $P(x_i|c)$ = , για χαρακτηριστικό με διακριτές τιμές,$\frac{N(\chi_i,c)}{N(c),}$
- $P(x_i|c) = g(x_i, \mu_c, \sigma_c^2)$,για χαρακτηριστικό με αριθμητικές τιμές,

όπου $N(c)$ είναι ο αριθμός των παραδειγμάτων που έχουν στην εξαρτημένη μεταβλητή την τιμή c, $N(x_i, c)$ είναι ο αριθμός των παραδειγμάτων που έχουν για το χαρακτηριστικό X_i και την εξαρτημένη μεταβλητή, τιμές x_iκαι c αντίστοιχα, και $g(x_i, \mu_c, \sigma_c^2)$, είναι η συνάρτηση πυκνότητας πιθανότητας Gauss με μέσο όρο μc και διασπορά σ_c για το χαρακτηριστικό X_i.

4.2.3 Νευρωνικά Δίκτυα

Η τεχνική μάθησης με νευρωνικά δίκτυα προσπαθεί να προσομοιώσει τον τρόπο μάθησης του ανθρώπινου εγκεφάλου. Ο ανθρώπινος εγκέφαλος έχει τον νευρώνα ως βασικό δομικό στοιχείο και αποτελείται από τον άξονα, τους δενδρίτες και τις συνάψεις. Ο άξονας είναι η πύλη εξόδου του νευρώνα και στέλνει σήματα προς άλλους νευρώνες, οι δενδρίτες είναι οι πύλες εισόδου του νευρώνα και οι συνάψεις είναι τα σημεία ένωσης μεταξύ του νευρώνα με δενδρίτες άλλων νευρώνων. Η μελέτη των νευρώνων και η προσπάθεια μαθηματικής μοντελοποίησης τους οδήγησαν στα τεχνικά νευρωνικά δίκτυα. Το μοντέλο McCulloch-Pitts είναι ένα απλό μοντέλο για την περιγραφή του νευρώνα [38].



Εικόνα 7 Μοντέλο McCulloch-Pitts

Τα είναι οι είσοδοι του νευρώνα και τα είναι τα συνοπτικά βάρη. Αν το άθροισμα είναι μεγαλύτερο από το κατώφλι τότε ο νευρώνας ενεργοποιείται διαφορετικά παραμένει αδρανής.Οπου είναι η έξοδος του νευρώνα, το άθροισμα των εισόδων πολλαπλασιασμένα με τα βάρη τους, το κατώφλι και f η βηματική συνάρτησης ενεργοποίησης. Υπάρχουν και άλλες μοντελοποιήσεις νευρώνων με σημαντικότερη διαφορά την συνάρτηση ενεργοποίησης του νευρώνα. Η πιο διαδεδομένη συνάρτηση ενεργοποίησης είναι η σιγμοειδής

.

Ένα νευρωνικό δίκτυο εκπαιδεύεται με ένα σύνολο από παραδείγματα που έχουν την είσοδο του νευρωνικού δικτύου αλλά και την έξοδο. Η έννοια της μάθησης και της εκπαίδευσης κρύβεται πίσω από τα βάρη του κάθε νευρώνα. Μετά από κάθε κύκλο εκπαίδευσης (εποχή) τα συναπτικά (synapses) βάρη μεταβάλλονται αναλόγως του σφάλματος που προέκυψε. Σφάλμα ορίζεται ως η διαφορά μεταξύ του στόχου του νευρωνικού δικτύου και της εξόδου του.

Τα νευρωνικά δίκτυα χρησιμοποιούνται κυρίως για ταξινόμηση και παρεμβολή και βρίσκουν μεγάλη εφαρμογή σε πραγματικά δεδομένα όπου υπάρχει θόρυβος. Είναι πολύ γρήγορα στην εφαρμογή της γνώσης που έχουν αποκτήσει και μπορούν να χρησιμοποιηθούν σε συστήματα πραγματικού χρόνου. Μειονέκτημα τους είναι ότι απαιτούν πολύ χρόνο στην εκπαίδευση και ότι δεν μπορεί να ερμηνευτεί η γνώση που αποκτούν από τον άνθρωπο.

4.2.4 Κατηγοριοποίηση μέσω του Αλγορίθμου των Κοντινότερων Γειτόνων (K-Nearest Neighbors)

Ο κοντινότερος γείτονας είναι ίσως ο πιο απλός αλγόριθμος για τη πρόβλεψη της κλάσης ενός παραδείγματος δοκιμής. Η φάση της εκμάθησης είναι απλή: Αποθηκεύεται κάθε παράδειγμα κατάρτισης, με την τιμή του. Για να γίνει μια πρόβλεψη σε ένα παράδειγμα δοκιμής, αρχικά υπολογίζετε η απόστασή του σε κάθε παράδειγμα κατάρτισης, κατόπιν, κρατούνται τα πιο κοντινά παραδείγματα κατάρτισης , όπου είναι ένας σταθερός ακέραιος. Βρίσκεται η τιμή που είναι η πιο κοινή μεταξύ αυτών των παραδειγμάτων. Αυτή η τιμή είναι η πρόβλεψη για αυτό το παράδειγμα δοκιμής. Σύμφωνα με τα παραπάνω η μέθοδος KNN είναι μια συνάρτηση τύπου:

).Μια συνάρτηση απόστασης έχει τον τύπο: [43].

Αυτή η βασική μέθοδος ονομάζετε αλγόριθμος KNN. Υπάρχουν δύο σημαντικές επιλογές στο σχεδιασμό του: Η τιμή του και η συνάρτηση της απόστασης που θα χρησιμοποιηθεί.Όταν υπάρχουν

δύο εναλλακτικές κατηγορίες, προκειμένου να αποφύγουμε δεσμούς / ενώσεις (ties)η πιο κοινή επιλογή για το K είναι ένας μικρός μονός ακέραιος αριθμός, παραδείγματος χάριν $K = 3$.Εάν υπάρχουν περισσότερες από δύο κλάσεις, δεσμοί είναι δυνατόν να δημιουργηθούν ακόμα και όταν το K είναι μονός αριθμός. Οι δεσμοί μπορούν επίσης να προκύψουν όταν οι δύο τιμές απόστασης είναι ίδιες. Μια εφαρμογή του KNN χρειάζεται έναν λογικό αλγόριθμο για να σπάσει τους δεσμούς, δεν υπάρχει καμία συναίνεση στον καλύτερο τρόπο να γίνει αυτό [43].

Όταν κάθε παράδειγμα είναι ένα καθορισμένου μήκους διάνυσμα πραγματικών αριθμών, η πιο κοινή συνάρτηση απόστασης είναι Ευκλείδεια απόσταση:

$$d(x,y) = \|x - y\| = \sqrt{(x-y)\cdot(x-y)} = \left(\sum_{i=1}^{m}(x_i - y_i)^2\right)^{\frac{1}{2}}$$

Όπου x και y είναι σημεία στο $x = \Re^m$ [43].

Ο αλγόριθμος Knn

Ο κοντινότερος γείτονας είναι ένας από τους δημοφιλέστερους κανόνες ταξινόμησης, αν και είναι μια παλαιά τεχνική. Δίνονται c κλάσεις ω_i, i =$1, 2, \dots\dots, c$ και ένα σημείο$x \in R^l$ και N σημεία εκμάθησης, x_i, $i = 1, 2, \dots\dots, N$, στο χώρο διάστασης l, με τις αντίστοιχες τιμές κλάσεων. Σε ένα σημείο x όπου η τιμή κλάσης του είναι άγνωστη, πρέπει να τοποθετηθεί το x σε μία απ τις κλάσεις του c. Ο κανόνας προκύπτει απ τα παρακάτω βήματα:

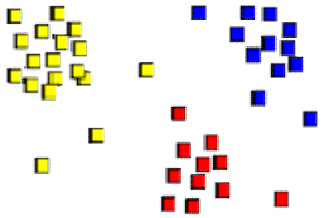
- Μεταξύ των σημείων εκμάθησης N ψάξε για τους k γείτονες όσο πιο κοντά στο x, με χρήση της συνάρτησης αποστάσεων (Euclidean, Manhattan, Mahalanobis). Η παράμετρος k ορίζεται από το χρήστη. Εδώ πρέπει να σημειώσουμε ότι δεν πρέπει να είναι πολλαπλάσιο του c. Δηλαδή για δύο κατηγορίες το kπρέπει να είναι ένας μονός αριθμός.
- Από τους K πιο κοντινούς γείτονες, ταχτοποίησε το νούμερο k_i από τα σημεία που ανήκουν στη κλάση ω_i. Προφανώς: $\sum_{i=1}^{c} k_i = k$.
- Όρισε το x στη κλάση ω_i για την οποία ισχύει $k_i > k_j$, $j \neq i$. Δηλαδή το x τοποθετείτε στη κλάση στην οποία ανήκουν οι περισσότεροι K κοντινότεροι γείτονες [46].

5. Ομαδοποίηση (clustering)

5.1 Τι είναι η ομαδοποίηση

Η ανάλυση συστάδων ή ομαδοποίηση είναι η λειτουργία της ανάθεσης ενός συνόλου αντικειμένων σε ομάδες (αποκαλούμενες συστάδες) έτσι ώστε τα αντικείμενα στην ίδια συστάδα να είναι πιο όμοια (υπό κάποια έννοια) το ένα με το άλλο, σε σχέση με αντικείμενα άλλων συστάδων. Η ομαδοποίηση είναι μια κύρια λειτουργία της εξόρυξης δεδομένων και μια κοινή τεχνική για την ανάλυση στατιστικών στοιχείων που χρησιμοποιείται σε πολλούς τομείς, συμπεριλαμβανομένης της εκμάθησης μηχανών, της αναγνώρισης σχεδίων, της ανάλυσης εικόνας, της ανάκτησης πληροφοριών, και της βιοπληροφορικής [35],[36].

Η ανάλυση κατά συστάδες αποσκοπεί στο διαχωρισμό μιας συλλογής από στοιχεία σε υποσύνολα έτσι ώστε να υπάρχει ομοιογένεια μέσα σε ένα υποσύνολο και ανομοιογένεια μεταξύ των στοιχείων που ανήκουν σε διαφορετικά υποσύνολα. Επιπροσθέτως μπορεί να αποσκοπεί στην ιεραρχική οργάνωση των συστάδων με την διαδοχική ομαδοποίηση αυτών, έτσι ώστε σε κάθε στάδιο της ιεραρχίας, οι συστάδες που ανήκουν στην ίδια ομάδα να είναι πιο όμοιες μεταξύ τους σε σχέση με αυτές που ανήκουν σε άλλη ομάδα [49].



Εικόνα 8 Το αποτέλεσμα μιας ανάλυσης συστάδων που παρουσιάζεται ως χρωματισμός των τετραγώνων σε τρεις συστάδες [49].

Εκτός του όρου της ομαδοποίησης, υπάρχουν διάφοροι όροι με παρόμοιες έννοιες συμπεριλαμβανομένης της αυτόματης ταξινόμησης, της αριθμητική ταξινομίας και της τυπολογικής ανάλυσης. Οι λεπτές διαφορές είναι συχνά στη χρήση των αποτελεσμάτων, ενώ στην εξόρυξη δεδομένων, οι προκύπτουσες ομάδες είναι το θέμα ενδιαφέροντος, στην αυτόματη ταξινόμηση πρώτιστα η ιδιότητα της διάκρισης τους είναι που παρουσιάζει το κύριο ενδιαφέρον [35].

5.2 Οι στόχοι της ομαδοποίησης

Ο στόχος της ομαδοποίησης είναι να καθοριστεί η εγγενής ομαδοποίηση σε ένα σύνολο μη διατιμημένων (unlabeled) δεδομένων. Αλλά πώς να αποφασίσουμε τι αποτελεί μια καλή συγκέντρωση; Μπορεί να αποδειχθεί ότι δεν υπάρχει κανένα απόλυτο «καλύτερο» κριτήριο που θα ήταν ανεξάρτητο από τον τελικό στόχο της ομαδοποίησης. Συνεπώς, είναι ο χρήστης που πρέπει να παρέχει αυτό το

κριτήριο, κατά τέτοιο τρόπο ώστε το αποτέλεσμα της ομαδοποίησης να ανταποκριθεί στις ανάγκες του. Για παράδειγμα, θα μπορούσαμε να ενδιαφερθούμε για την εύρεση των αντιπροσώπων για τις ομοιογενείς ομάδες (μείωση στοιχείων), για την εύρεση των «φυσικών συστάδων» και να περιγράψουμε τις άγνωστες ιδιότητές τους («physical» data types), για την εύρεση των χρήσιμων και κατάλληλων σχηματισμών ομάδας («χρήσιμες» κατηγορίες στοιχείων) ή για την εύρεση των ασυνήθιστων αντικειμένων στοιχείων (outlier ανίχνευση).

Πιθανές εφαρμογές

Οι αλγόριθμοι ομαδοποίησης μπορούν να εφαρμοστούν σε πολλούς τομείς, για παράδειγμα:

1)Μάρκετινγκ: εύρεση ομάδων Ν πελατών με παρόμοια συμπεριφορά δεδομένης μιας μεγάλης βάσης δεδομένων των στοιχείων πελατών που περιέχουν τις ιδιότητες και τα αρχεία αγοράς στο παρελθόν.

2)Βιολογία: Ταξινόμηση των δεδομένων φυτών και ζώων, και των χαρακτηριστικών γνωρισμάτων τους.

3)Βιβλιοθήκες: Παραγγελίες / δανεισμός βιβλίων [Z].

4) Σχεδιασμός Πόλης: Προσδιορισμός των οικοδομικών τετραγώνων σύμφωνα με τον τύπο κτηρίου, την αξία και τη γεωγραφική θέση τους.

5)Μελέτες σεισμού: Συγκέντρωση των παραχωρηθέντων επίκεντρων σεισμού για να προσδιορίσει τις επικίνδυνες ζώνες [16].

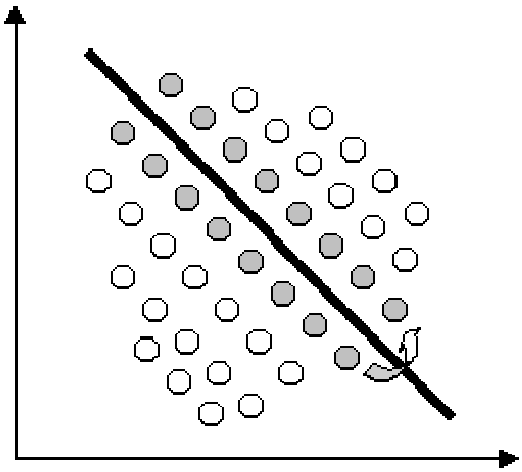
5.3 Αλγόριθμοι Ομαδοποίησης

Οι αλγόριθμοι ομαδοποίησης μπορούν να ταξινομηθούν στις παρακάτω κατηγορίες:

- Διαμεριστική ομαδοποίηση (Partitioning)
    - Αποκλειστική ομαδοποίηση.
    - Επικαλυπτική ομαδοποίηση.
- Ιεραρχική ομαδοποίηση.
- Πιθανοθεωρητική ομαδοποίηση

Στην πρώτη περίπτωση τα δεδομένα ομαδοποιούνται με έναν αποκλειστικό τρόπο, έτσι ώστε εάν ένα ορισμένο στοιχείο να ανήκει σε μια καθορισμένη συστάδα και έπειτα έτσι ώστε να μην μπορεί να συμπεριληφθεί σε μια άλλη συστάδα. Ένα απλό παράδειγμα αυτού, παρουσιάζεται στο σχέδιο 7, όπου ο χωρισμός των σημείων επιτυγχάνεται από μια ευθεία γραμμή σε ένα δισδιάστατο πεδίο. Αντίθετα ο δεύτερος τύπος, η επικάλυψη της συγκέντρωσης, χρησιμοποιεί συγκεχυμένα σύνολα για να συγκεντρώσει τα στοιχεία, έτσι ώστε κάθε σημείο μπορεί να ανήκει σε δύο ή περισσότερες συστάδες

με διαφορετικούς βαθμούς ιδιότητας μέλους. Σε αυτήν την περίπτωση, τα στοιχεία θα συνδεθούν σε μια κατάλληλη αξία ιδιότητας μέλους [4].



Σχέδιο 9 Αποκλειστική συγκέντρωση

Αντ' αυτού, ένας ιεραρχικός αλγόριθμος συγκέντρωσης είναι βασισμένος στην ένωση μεταξύ των δύο κοντινότερων συστάδων. Ο όρος αρχής πραγματοποιείται με τον καθορισμό κάθε στοιχείου ως συστάδα. Μετά από μερικές επαναλήψεις φθάνει στις τελικές επιθυμητές συστάδες. Τέλος, το τελευταίο είδος συγκέντρωσης χρησιμοποιεί μια απολύτως πιθανολογική προσέγγιση.

5.3.1 Διαμεριστικοί αλγόριθμοι

Οι διαμεριστικοί αλγόριθμοι ομαδοποίησης χωρίζονται σε υποκατηγορίες ανάλογα με το πώς χωρίζονται τα δεδομένα. Έτσι ο k-means και ο fuzzy k-means ανήκουν σε διαφορετικές κατηγορίες (ο k-means στην αποκλειστική ομαδοποίηση και ο fuzzy k-means στην επικαλυπτική). Επίσης αν ο αλγόριθμος μετρά αποστάσεις, ανήκει στην υποκατηγορία distance based, ενώ αν προσπαθεί να διαχωρίσει τον χώρο με βάση την πυκνότητα είναι density based.

5.3.1.1 K-means

Ο K-means είναι ίσως ο πιο διαδεδομένος επαναληπτικός αλγόριθμος ομαδοποίησης [56]. Εφαρμόζεται για την ομαδοποίηση ενός μεγάλου συνόλου αντικειμένων σε ομοιογενείς ομάδες. Όπως όλες οι επαναληπτικές μέθοδοι τμηματοποίησης, έτσι και ο k-means βασίζεται στην ιδέα της βελτιστοποίησης μιας συνάρτησης, η οποία είναι μια προσπάθεια να ερμηνευθεί με μαθηματικούς όρους η διαισθητική έννοια της ομάδας. Η συνάρτηση αυτή αναφέρεται ως κριτήριο της ομαδοποίησης (clustering criterion) ή ως συνάρτηση κόστους E η οποία δίνεται από τον ακόλουθο τύπο:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d(x_{ij} - \bar{x_i})$$

όπου:

k	Πλήθος των ομάδων
n_i	Πλήθος στοιχείων ομάδας C_i
x_{ij}	j-οστή περίπτωση της i-οστής ομάδας
$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$	Κέντρο της ομάδας C_i
$d(x, y) = \|x - y\|^2$	Είναι η τετραγωνική Ευκλείδεια απόσταση

Ο αλγόριθμος k-means αποτελείται από δύο φάσεις. Κατά την διάρκεια της πρώτης φάσης εκτελείται μια τμηματοποίηση των περιπτώσεων σε k ομάδες, ενώ κατά την διάρκεια της δεύτερης φάσης καθορίζεται η ποιότητα της τμηματοποίησης. Ο k-means είναι μια διαδικασία τεσσάρων βασικών βημάτων, αρχίζοντας από μία τυχαία αρχική τμηματοποίηση. Πιο συγκεκριμένα τα βήματα είναι τα ακόλουθα:

1. Επιλογή των αρχικών k κέντρων για τις k ομάδες.
2. Υπολογισμός της ανομοιότητας μεταξύ ενός αντικειμένου και του κέντρου μιας ομάδας.
3. Τοποθέτηση του αντικειμένου στην ομάδα, της οποίας το κέντρο είναι πιο κοντά στο αντικείμενο αυτό.
4. Ενημέρωση του κέντρου της ομάδας, έτσι ώστε να ελαχιστοποιηθεί η ανομοιότητα εντός της ομάδας.

   Εκτός από το πρώτο βήμα τα υπόλοιπα τρία εκτελούνται επαναληπτικά εώς ότου ο αλγόριθμος συγκλίνει.

Ο αλγόριθμος k-means έχει τις ακόλουθες σημαντικές ιδιότητες:

1. Η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι $O(tkmn)$ όπου m είναι το πλήθος των χαρακτηριστικών, n είναι το πλήθος των αντικειμένων, k είναι το σύνολο των ομάδων και t είναι το πλήθος των επαναλήψεων πάνω σε όλο το σύνολο δεδομένων. Συνήθως ισχύει ότι $k, m, t \ll n$. Στην ομαδοποίηση μεγάλων συνόλων δεδομένων ο αλγόριθμος k-means είναι πολύ γρηγορότερος από τους αρχικούς αλγόριθμους ομαδοποίησης, των οποίων η υπολογιστική πολυπλοκότητα είναι $O(n^2)$.
2. Συχνά τερματίζει σε ένα τοπικό βέλτιστο. Για να βρεθεί το ολικό βέλτιστο, υιοθετούνται τεχνικές όπως είναι οι γενετικοί αλγόριθμοι και το deterministic annealing, οι οποίες μπορούν να ενσωματωθούν με τον αλγόριθμο k-means.
3. Τα δεδομένα πρέπει να λαμβάνουν μόνο αριθμητικές τιμές, διότι ελαχιστοποιεί την συνάρτηση κόστους υπολογίζοντας τους μέσους των ομάδων.

4. Επειδή οι ομάδες, που ανακαλύπτει ο αλγόριθμος, έχουν κυρτά σχήματα, είναι δύσκολο να χρησιμοποιηθεί ο αλγόριθμος k-means για τον εντοπισμό μη κυρτών ομάδων.

Παρόλη την αποτελεσματικότητά του και τη διαδεδομένη χρήση του ο αλγόριθμος k-means έχει αρκετά μειονεκτήματα, τα σημαντικότερα από αυτά, είναι τα ακόλουθα:

1. Υποθέτει πως το πλήθος των ομάδων k σε μια βάση δεδομένων είναι εκ των προτέρων γνωστό, το οποίο δεν είναι απαραίτητα αληθές σε εφαρμογές του πραγματικού κόσμου.
2. Ως μια επαναληπτική μέθοδος, ο αλγόριθμος k-means είναι ιδιαίτερα ευαίσθητος στην επιλογή των αρχικών ομάδων. Επίσης επηρεάζεται αρκετά από την παρουσία θορύβου και απομακρυσμένων τιμών (outliers).
3. Συγκλίνει σε έναν τοπικό ελάχιστο, συχνά φτωχής ποιότητας, δηλαδή η λύση που προτείνει δεν είναι ικανοποιητική.
4. Ο αλγόριθμος θεωρεί τα k κέντρα ως αντιπροσώπους των δεδομένων κάθε ομάδας. Ωστόσο, είναι δυνατόν ο αριθμητικός μέσος να μην έχει καμία έγκυρη ερμηνεία.
5. Εξαιτίας του χρόνου που χρειάζεται να ολοκληρωθεί μια επανάληψη δεν μπορεί να χειρισθεί μεγάλες βάσεις δεδομένων γρήγορα. Επίσης, τα δεδομένα που μπορεί να επεξεργαστεί είναι μόνο αριθμητικά και όχι κατηγορικά.

5.3.1.2 Fuzzy C-Means Clustering (Συγκεχυμένη συγκέντρωση c-μέσων)

Ο αλγόριθμος FCM, είναι μια μέθοδος συγκέντρωσης που επιτρέπει ένα κομμάτι των δεδομένων να ανήκει σε δύο ή περισσότερες συστάδες. Αυτή η μέθοδος χρησιμοποιείται συχνά στην αναγνώριση πρότυπων. Είναι βασισμένο στην ελαχιστοποίηση της ακόλουθης συνάρτησης:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - c_j \right\|^2$$

$$1 \leq m < \infty$$

Όπου το m είναι οποιοσδήποτε πραγματικός αριθμός μεγαλύτερος από 1,Το U_{ij} είναι ο βαθμός ιδιότητας μέλους X_i στη συστάδα j, X_iείναι το i-στο των d-διαστάσεων μετρημένων στοιχείων, c_jείναι το κέντρο d-διάστασης της συστάδας, και ||*|| είναι οποιοσδήποτε κανόνας εκφράζει την ομοιότητα μεταξύ οποιωνδήποτε μετρημένων στοιχείων και του κέντρου. Ο συγκεχυμένος χωρισμός πραγματοποιείται μέσω μιας επαναληπτικής βελτιστοποίησης της αντικειμενικής συνάρτησης που παρουσιάζεται ανωτέρω, με την αναπροσαρμογή της ιδιότητας μέλους u_{ij}και η συστάδα στρέφεται c_j από:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}} \ ,$$

Αυτή η επανάληψη θα σταματήσει όταν ε όπου ε ένα κριτήριο λήξης μεταξύ 0 και 1, ενώ το K είναι τα βήματα επανάληψης. Αυτή η διαδικασία συγκλίνει σε ένα σημείο τοπικών ελαχίστων J_m.

Ο αλγόριθμος αποτελείται από τα ακόλουθα βήματα:

1) Εκκίνηση $U^{(0)}$

   Στο βήμα κ: Υπολόγισε τα κεντρικά διανύσματα: $C^{(k)}$ $[c_j]$ με $U^{(k)}$

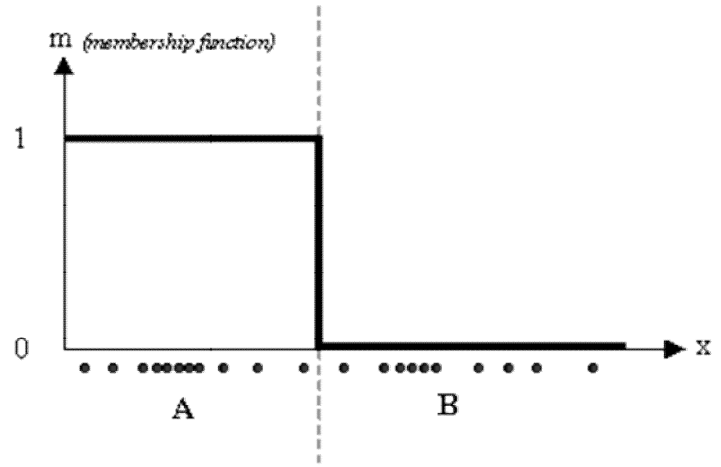
2) Ενημέρωσε:

3) Εάν $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ τότε σταμάτα διαφορετικά επιστροφή στο βήμα 2 [39].

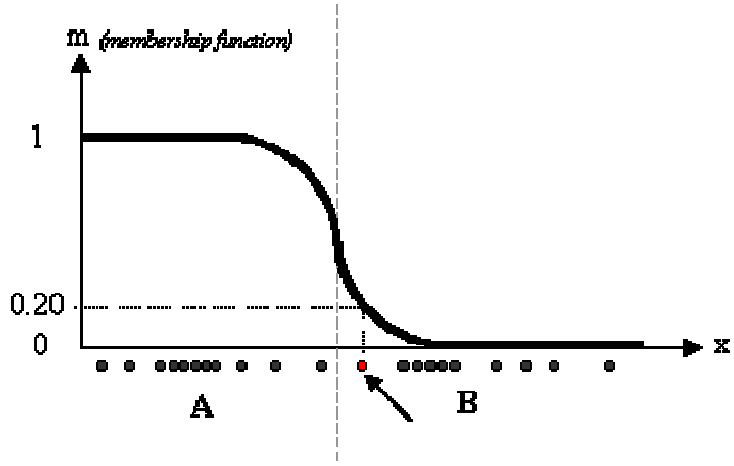
Παρατηρήσεις:

Τα δεδομένα είναι συνδεδεμένα σε κάθε συστάδα με τη βοήθεια μιας συνάρτησης ιδιότητας μέλους, η οποία αντιπροσωπεύει τη συγκεχυμένη συμπεριφορά αυτού του αλγορίθμου. Για να γίνει αυτό, πρέπει απλά να χτίσει μια κατάλληλη μήτρα(την ονομάζουμε) U της οποίας οι παράγοντες είναι αριθμοί μεταξύ 0 και 1, και αντιπροσωπεύουν το βαθμό ιδιότητας μέλους μεταξύ των δεδομένων και των κέντρων των συστάδων. Για καλύτερη κατανόηση, μπορούμε να εξετάσουμε αυτό το απλό μονοδιάστατο παράδειγμα. Λαμβάνοντας υπόψη ένα ορισμένο σύνολο δεδομένων, υποθέτουμε ότι διανέμονται σε έναν άξονα όπως στο παρακάτω σχήμα:

Εξετάζοντας την εικόνα, μπορούμε να προσδιορίσουμε δύο συστάδες στην εγγύτητα των δύο συγκεντρώσεων δεδομένων. Θα αναφερόμαστε σε αυτά ως `A και `B. Στην πρώτη προσέγγιση που παρουσιάστηκε -αλγόριθμος K-μέσων - συνδέσαμε κάθε δεδομένο με συγκεκριμένο κεντροειδές επομένως, αυτή η λειτουργία ιδιότητας μέλους μοιάζει έτσι:



Στην προσέγγιση FCM, αντ' αυτού, το ίδιο δεδομένο δεν ανήκει αποκλειστικά σε μια καλά καθορισμένη συστάδα, αλλά μπορεί να τοποθετηθεί με έναν μέσο τρόπο. Σε αυτήν την περίπτωση, η λειτουργία ιδιότητας μέλους ακολουθεί μια ομαλότερη γραμμή για να δείξει ότι κάθε στοιχείο μπορεί να ανήκει σε διάφορες συστάδες με τις διαφορετικές τιμές του συντελεστή ιδιότητας μέλους.



Στο παραπάνω σχέδιο, το στοιχείο που παρουσιάζεται ως κόκκινο χαρακτηρισμένο σημείο ανήκει περισσότερο στη συστάδα B παρά τη συστάδα A. Η τιμή 0.20 στον άξονα m το βαθμό ιδιότητας μέλους στο A για τέτοιο στοιχείο. Τώρα, αντί της χρήσης μιας γραφικής αντιπροσώπευσης, εισάγουμε το U μήτρα της οποίας παράγοντες είναι αυτοί που λαμβάνονται από τις λειτουργίες ιδιότητας μέλους:

$$U_{MC} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \end{bmatrix}_{(\alpha)} \qquad U_{MC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}_{(\beta)}$$

Ο αριθμός σειρών και στηλών εξαρτάται από πόσα δεδομένα και συστάδες εξετάζουμε. Ακριβέστερα έχουμε c = 2 στήλες (c = 2 συστάδες) και σειρές N, όπου το c είναι ο συνολικός αριθμός των συστάδων και το N είναι ο συνολικός αριθμός των στοιχείων. Το γενικό στοιχείο υποδεικνύετε ως εξής:u_{ij}.Στα παραδείγματα ανωτέρω έχουμε εξετάσει τις περιπτώσεις K-means (α) και FCM (β). Μπορούμε να παρατηρήσουμε ότι στην πρώτη περίπτωση (α) οι συντελεστές είναι πάντα ενωτικοί. Άλλες ιδιότητες παρουσιάζονται κατωτέρω:

$$u_{ij} \in [0,1] \quad \forall i,j$$

$$\sum_{j=1}^{c} u_{ik} = 1 \quad \forall i$$

$$0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall N[39]$$

5.3.2 Hierarchical clustering

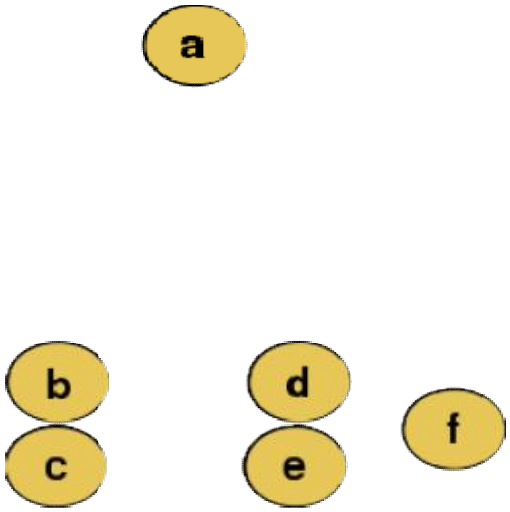
Η ιεραρχική συγκέντρωση δημιουργεί μια ιεραρχία των συστάδων που μπορεί να αντιπροσωπευθεί ως μια δομή δέντρων αποκαλούμενη δενδοδιάγραμμα .Η ρίζα του δέντρου αποτελείται από μια ενιαία συστάδα που περιέχει όλες τις παρατηρήσεις, και τα φύλλα αντιστοιχούν στις μεμονωμένες παρατηρήσεις.

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης είναι γενικά είτε συσσωρευτικοί, στους οποίους κάποιοι αρχίζουν στα φύλλα και συγχωνεύουν διαδοχικά τις συστάδες από κοινού ή διαχωριστικοί, στους οποίους κάποιοι αρχίζουν στη ρίζα και χωρίζουν κατ' επανάληψη τις συστάδες.

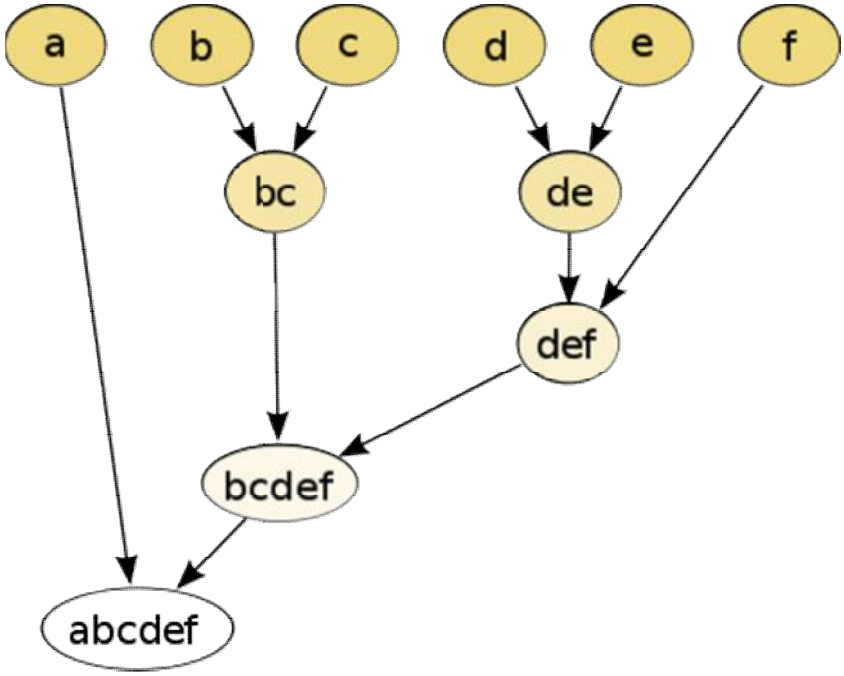
Οποιοσδήποτε έγκυρη μετρική μπορεί να χρησιμοποιηθεί ως μέτρο ομοιότητας μεταξύ των ζευγαριών των παρατηρήσεων. Η επιλογή της οποίας συγκεντρώνεται για να συγχωνεύσει ή να χωρίσει καθορίζεται από ένα κριτήριο συνδέσμων-αποστάσεων μεταξύ των παρατηρήσεων.

Η κοπή του δέντρου σε ένα δεδομένο ύψος θα δώσει μια συγκέντρωση ακριβείας σύμφωνα με τα επιλεγμένα δεδομένα. Στο ακόλουθο παράδειγμα η κοπή μετά από τη δεύτερη σειρά θα παραγάγει τις συστάδες {a} {bc} {de} {f}. Η κοπή μετά από την τρίτη σειρά θα παραγάγει τις συστάδες {a} {bc} {def}, που είναι μια πιο χονδροειδής συγκέντρωση, με έναν μικρότερο αριθμό μεγαλύτερων συστάδων.[37]

Παραδείγματος χάριν, υποθέτουμε ότι αυτό το στοιχείο πρόκειται να συγκεντρωθεί, και η Ευκλείδεια απόσταση είναι απόσταση μετρική.



Το δενδροδιάγραμμα ιεραρχικής ομαδοποίησης θα είναι ως εξής:



Αυτή η μέθοδος χτίζει την ιεραρχία από τα μεμονωμένα στοιχεία με σταδιακά να συγχωνεύσει τις συστάδες. Στο παράδειγμά μας, έχουμε έξι στοιχεία {a} {b} {c} {d} {e} και {f}. Το πρώτο βήμα είναι να αποφασιστούν ποια στοιχεία να συγχωνευτούν σε μια συστάδα. Συνήθως, θέλουμε να πάρουμε τα δύο πιο κοντινά στοιχεία, σύμφωνα με την επιλεγμένη απόσταση.

Ας υποθέσουμε ότι έχουμε συγχωνεύσει τα δύο πιο κοντινά στοιχεία β και γ, έχουμε τώρα τις ακόλουθες συστάδες {a}, {b, c}, {d}, {e} και {f}, και θέλουμε να τις συγχωνεύσουμε περαιτέρω. Για να

το κάνουμε, πρέπει να πάρουμε την απόσταση μεταξύ {a} και {bc}, και επομένως να καθορίσουμε την απόσταση μεταξύ δύο συστάδων. Συνήθως η απόσταση μεταξύ δύο συστάδων {A} και {B} είναι μια από την ακόλουθη:

•Η μέγιστη απόσταση μεταξύ των στοιχείων κάθε συστάδας (επίσης αποκαλούμενης πλήρη σύνδεσμο που συγκέντρωσης):

$$max\ \{d(x,y): x \in A, \quad y \in B\}$$

•Η ελάχιστη απόσταση μεταξύ των στοιχείων κάθε συστάδας (επίσης αποκαλούμενης ενιαίος-σύνδεσμος συγκέντρωσης):

$$min\ \{d(x,y):\ x \in A, \quad y \in B\}$$

•Η μέση απόσταση μεταξύ των στοιχείων κάθε συστάδας (επίσης αποκαλούμενης μέσος σύνδεσμος συγκέντρωσης):

$$\frac{1}{|A||B|}\sum_{x \in A}\sum_{y \in B} d(x,y)$$

•Η αύξηση στη διαφορά της συστάδα που συγχωνεύεται.

•Η πιθανότητα ότι οι συστάδες υποψηφίων προέρχονται από την ίδια συνάρτηση κατανομής.

Κάθε συσσώρευση εμφανίζεται σε μια μεγαλύτερη απόσταση μεταξύ των συστάδων από την προηγούμενη συσσώρευση και κάποιες μπορεί να αποφασίσουν να σταματήσουν τη διαδικασία, είτε όταν οι συστάδες βρίσκονται σε μεγάλες αποστάσεις και δεν μπορούν να συγχωνευτούν (κανόνας αποστάσεων) ή όταν υπάρχει ένας αρκετά μικρός αριθμός συστάδων (κριτήριο αριθμού).[37]

5.3.3 Mixture of Gaussians (Μίξη Γκαουσιανών)

Υπάρχει ένας άλλος τρόπος να εξεταστεί η συγκέντρωση των προβλημάτων: Μια πρότυπη συγκέντρωση, η οποία συνίσταται στη χρήση ορισμένων προτύπων για τις συστάδες και την προσπάθεια να βελτιστοποιηθεί η τακτοποίηση μεταξύ των στοιχείων και του προτύπου. Στην πράξη, κάθε συστάδα μπορεί να αντιπροσωπευθεί από μαθηματική άποψη από μια παραμετρική κατανομή, όπως έναν Γκαουσσιανό (συνεχής) ή ένα Poisson (διακριτό). Το ολόκληρο σύνολο δεδομένων επομένως διαμορφώνεται από ένα μίγμα αυτών των κατανομών. Μια μεμονωμένη κατανομή χρησιμοποίησε για να διαμορφώσει μια συγκεκριμένη συστάδα αναφέρεται συχνά ως αποτελούμενη constitutive κατανομή.

Ένα μίγμα πρότυπων έχει πολλές πιθανότητες να τείνει να έχει τα ακόλουθα γνωρίσματα:

• Οι constitutive κατανομές έχουν τις υψηλές «αιχμές» (τα στοιχεία σε μια συστάδα είναι σφιχτά).

• Το πρότυπο μιγμάτων «καλύπτει» τα στοιχεία καλά (τα κυρίαρχα σχέδια στα στοιχεία συλλαμβάνονται από τις constitutive κατανομές).

Κύρια πλεονεκτήματα της πρότυπης - βασισμένης συγκέντρωσης:

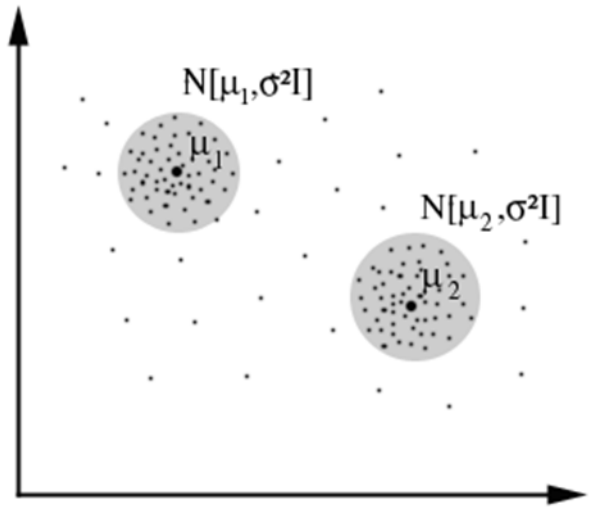
•Διαθέσιμες καλά μελετημένες στατιστικές τεχνικές συμπεράσματος .

•Ευελιξία στην επιλογή της συστατικής διανομή.

•Λήψη εκτιμήσεως πυκνότητας για κάθε συστάδα.

Μίγμα Γκαουσσιανών

Η πιο ευρύτατα χρησιμοποιημένη μέθοδος ομαδοποίησης αυτού του είδους είναι βασισμένη στην εκμάθηση ενός μίγματος Γκαουσσιανών: Μπορούμε πραγματικά να θεωρήσουμε τις συστάδες ως Γκαουσσιανές κατανομές που κεντροθετούνται στα βαρύκεντρα τους, όπως μπορούμε να δούμε στην παρακάτω εικόνα ο γκρίζος κύκλος αντιπροσωπεύει την πρώτη διαφορά της κατανομής:



Ο αλγόριθμος λειτουργεί κατά αυτόν τον τρόπο: [38]

1. Επιλέγει το συστατικό τυχαία με την πιθανότητα
2. Παίρνει δείγμα από το σημείο

5.4 Τάσεις στη διαδικασία Ομαδοποίησης

Αν και η ανάλυση συστάδων είναι υπαγόμενη της λεπτομερούς έρευνας για πολλά έτη και σε ποικίλες επιστήμες, υπάρχουν ακόμα αρκετά ανοικτά ερευνητικά ζητήματα. Συνοψίζουμε μερικές από τις πιο ενδιαφέρουσες τάσεις συγκέντρωσης ως εξής:

1. Ανακάλυψη και εύρεση των αντιπροσώπων των αυθαίρετα διαμορφωμένων συστάδων. Μια από τις απαιτήσεις στη συγκέντρωση είναι ο χειρισμός των αυθαίρετα διαμορφωμένων συστάδων. Εντούτοις, δεν υπάρχει καμία καθιερωμένη μέθοδος για να περιγράψει τη δομή των αυθαίρετα διαμορφωμένων συστάδων όπως καθορίζεται από έναν αλγόριθμο. Αν λάβουμε υπόψη το γεγονός ότι η συγκέντρωση είναι ένα σημαντικό εργαλείο για τη μείωση των δεδομένων, είναι σημαντικό να βρεθούν οι αρμόδιοι αντιπρόσωποι των συστάδων που περιγράφουν τη μορφή τους. Κατά συνέπεια, μπορούμε αποτελεσματικά να περιγράψουμε τα ελλοχεύοντα δεδομένα βασιζόμενοι στα αποτελέσματα της συγκέντρωσης, ενώ επιτυγχάνουμε μια σημαντική συμπίεση του τεράστιου ποσού των αποθηκευμένων δεδομένων (μείωση δεδομένων).

2. Συγκέντρωση μη-σημείου(non point clustering). Η μεγάλη πλειοψηφία των αλγορίθμων έχει εξετάσει μόνο τα αντικείμενα σημείου, αν και σε πολλές περιπτώσεις πρέπει να χειριστούμε τα σύνολα εκτεταμένων αντικειμένων όπως (υπέρ) - ορθογώνια. Κατά συνέπεια, μια μέθοδος που χειρίζεται αποτελεσματικά τα σύνολα μη-σημείου και ανακαλύπτει τις έμφυτες συστάδες που παρουσιάζονται μέσα τους, αποτελούν αντικείμενο περαιτέρω έρευνας με εφαρμογές σε διαφορετικές περιοχές (όπως οι χωρικές βάσεις δεδομένων / spatial databases, η ιατρική, η βιολογία).

3. Διαχείριση αβεβαιότητας στη διαδικασία συγκέντρωσης και την απεικόνιση των αποτελεσμάτων. Η πλειοψηφία της συγκέντρωσης των τεχνικών υποθέτει ότι τα όρια των συστάδων είναι σαφώς ορισμένα (crisp). Κατά συνέπεια κάθε σημείο δεδομένων μπορεί να ταξινομηθεί το πολύ σε μια συστάδα. Επιπλέον όλα τα σημεία που ταξινομούνται σε μια συστάδα, ανήκουν στο ίδιο με τον ίδιο βαθμό πεποίθησης (δηλ., όλες οι τιμές αντιμετωπίζονται εξίσου στη διαδικασία συγκέντρωσης). Το αποτέλεσμα είναι τέτοιο, ώστε σε μερικές περιπτώσεις «ενδιαφέροντα» δεδομένα ξεφεύγουν από τα όρια των συστάδων και έτσι δεν είναι ταξινομημένα καθόλου. Αυτό είναι απίθανο στην καθημερινή ζωή όπου μια μεταβλητή μπορεί να ταξινομηθεί σε περισσότερες από μια κατηγορίες. Κατά συνέπεια μια περαιτέρω κατεύθυνση εργασίας λαμβάνει υπόψη την αβεβαιότητα έμφυτη στα δεδομένα. Μια άλλη ενδιαφέρουσα κατεύθυνση είναι η μελέτη των τεχνικών που απεικονίζουν αποτελεσματικά πολυδιάστατες συστάδες που παίρνουν επίσης υπ όψιν τους χαρακτηριστικά γνωρίσματα αβεβαιότητας.

4. Επαυξητική συγκέντρωση (incremental clustering). Οι συστάδες σε ένα σύνολο δεδομένων, μπορούν να αλλάξουν ως εισαγωγές / ενημερώσεις (updates) καθώς και να γίνουν διαγραφές κατά το κύκλου ζωής τους. Κατόπιν είναι σαφές ότι υπάρχει μια ανάγκη να αξιολογηθεί το σχέδιο συγκέντρωσης που καθορίζεται για ένα σύνολο δεδομένων ώστε να ενημερωθεί έγκαιρα. Εντούτοις, είναι σημαντικό να χρησιμοποιηθούν οι πληροφορίες που κρύβονται στα προηγούμενα σχέδια συγκεντρώσεων ώστε αυτά να ενημερωθούν με έναν επαυξητικό τρόπο.

5. Περιορισμός - βασισμένη συγκέντρωση. Ανάλογα με την περιοχή εφαρμογής μπορούμε να εξετάσουμε τις διαφορετικές πτυχές συγκέντρωσης ως σημαντικότερες. Μπορεί να είναι σημαντικό να τονιστούν ή να αγνοηθούν μερικές πτυχές των δεδομένων σύμφωνα με τις απαιτήσεις της εξεταζόμενης εφαρμογής. Τα τελευταία χρόνια, υπάρχει μια τάση έτσι ώστε η ανάλυση συστάδων να είναι βασισμένη σε λιγότερες παραμέτρους αλλά με περισσότερους περιορισμούς. Αυτοί οι περιορισμοί μπορούν να υπάρξουν στο διάστημα δεδομένων ή στις ερωτήσεις των χρηστών. Κατόπιν μια διαδικασία συγκέντρωσης πρέπει να καθοριστεί ώστε να λάβουμε υπόψη τους περιορισμούς και να καθορίσουμε τις έμφυτες συστάδες που ταιριάζουν σε ένα σύνολο δεδομένων. [53]

6. Μέτρα αξιολόγησης

Ένα από τα σημαντικότερα ζητήματα είναι η αξιολόγηση των αποτελεσμάτων από την εφαρμογή των μεθόδων έτσι ώστε η γνώση που αποκαλύπτεται να ταιριάζει καλύτερα στα ελλοχεύοντα δεδομένα. Αυτό αποτελεί το κύριο αντικείμενο των μέτρων αξιολόγησης (validation measures). Στη συνέχεια συζητάμε τις θεμελιώδεις έννοιες αυτής της περιοχής ενώ παρουσιάζουμε τις διάφορες προσεγγίσεις ισχύος συστάδων που προτείνονται στη βιβλιογραφία [53].

6.1 Μέτρα Αξιολόγησης Ομαδοποίησης

Ο στόχος των μεθόδων ομαδοποίησης είναι να ανακαλυφθούν οι σημαντικές ομάδες οι οποίες είναι παρούσες σε ένα σύνολο δεδομένων. Γενικά, πρέπει να ψάξουν για τις συστάδες των οποίων τα μέλη είναι κοντά το ένα στο άλλο (με άλλα λόγια έχει έναν υψηλό βαθμό ομοιότητας) και καλά χωρισμένες. Ένα πρόβλημα που αντιμετωπίζουμε στη συγκέντρωση είναι να αποφασιστεί ο βέλτιστος αριθμός συστάδων που ταιριάζει σε ένα σύνολο δεδομένων.

Στις περισσότερες αξιολογήσεις οι αλγόριθμοι χρησιμοποιούν δεδομένα δύο διαστάσεων (2d) ώστε ο αναγνώστης να είναι σε θέση να ελέγξει οπτικά την ισχύ των αποτελεσμάτων (δηλ., πόσο καλά ο αλγόριθμος ομαδοποίησης ανακάλυψε τις συστάδες του συνόλου στοιχείων).

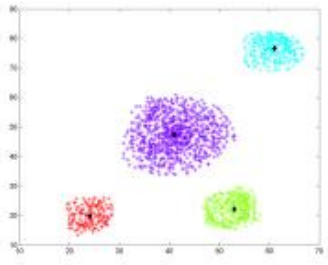
Είναι σαφές ότι η απεικόνιση των δεδομένων στο σύνολο είναι μια κρίσιμη επαλήθευση των αποτελεσμάτων. Στην περίπτωση μεγάλων πολυδιάστατων συνόλων δεδομένων, η αποτελεσματική απεικόνιση (π.χ. περισσότερες από τρεις διαστάσεις) του συνόλου δεδομένων είναι δύσκολη. Επιπλέον η αντίληψη για τις συστάδες που χρησιμοποιούν τα διαθέσιμα εργαλεία απεικόνισης είναι ένα δύσκολο έργο για τους ανθρώπους που δεν είναι εξοικειωμένοι με τα υψηλότερα διαστατικά διαστήματα.

Οι διάφοροι αλγόριθμοι συμπεριφέρονται σε έναν διαφορετικό τρόπο ανάλογα με:

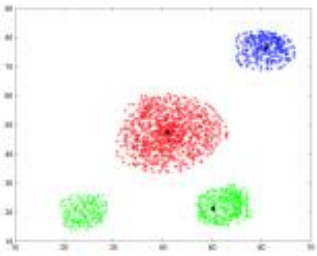
1. Τα χαρακτηριστικά γνωρίσματα του συνόλου στοιχείων (γεωμετρία και διανομή πυκνότητας των συστάδων).
2. οι τιμές παραμέτρων εισαγωγής.

Για παράδειγμα, υποθέστε το σύνολο δεδομένων στο σχέδιο (α).

Είναι προφανές ότι μπορούμε να ανακαλύψουμε τέσσερις συστάδες στο σύνολο δεδομένων. Εντούτοις, εάν εξετάζουμε έναν αλγόριθμο συγκέντρωσης (π.χ. K-Means) με ορισμένες τιμές παραμέτρου (στην περίπτωση των K-μέσων ο αριθμός συστάδων) ώστε να χωριστεί το σύνολο δεδομένων σε τρείς συστάδες, το αποτέλεσμα της συγκέντρωσης της διαδικασίας θα ήταν το σχέδιο συγκέντρωσης που παρουσιάζεται στον σχέδιο (β).



(α)



(β)

Στο παράδειγμά μας ο αλγόριθμος συγκέντρωσης (K-Means) βρήκε τις καλύτερες τρεις συστάδες στις οποίες το σύνολο στοιχείων μας θα μπορούσε να χωριστεί. Εντούτοις, αυτό δεν είναι ο βέλτιστος χωρισμός για το εξεταζόμενο σύνολο δεδομένων. Καθορίζουμε, εδώ, τον όρο «βέλτιστο» συγκεντρωμένο σχέδιο ως έκβαση του τρεξίματος ενός αλγορίθμου συγκέντρωσης (διαχωρισμό/ τμηματοποίηση) το οποίο καθιστά καλύτερα τα έμφυτα χωρίσματα του συνόλου δεδομένων. Είναι προφανές από το σχέδιο (β) που απεικονίζεται, ότι το πρότυπο δεν είναι το καλύτερο για το σύνολο δεδομένων μας δηλ., το πρότυπο συγκέντρωσης που παρουσιάζεται στην εικόνα (β) δεν ταιριάζει καλά με το σύνολο δεδομένων. Η βέλτιστη συγκέντρωση για το σύνολο δεδομένων μας θα είναι ένα σχέδιο με τέσσερις συστάδες.

Κατά συνέπεια, εάν στις παραμέτρους του αλγορίθμου συγκέντρωσης ορίζεται μια ανάρμοστη μεταβλητή, η μέθοδος συγκέντρωσης μπορεί να οδηγήσει σε ένα σχέδιο διαχωρισμού που δεν είναι βέλτιστο για το συγκεκριμένο σύνολο δεδομένων και να οδηγήσει σε λανθασμένες αποφάσεις.

6.2 Θεμελιώδεις έννοιες της αξιολόγησης συστάδων

Η διαδικασία αξιολόγησης των αποτελεσμάτων ενός αλγορίθμου συστηματοποίησης είναι γνωστή υπό τον όρο: αξιοπιστία ομάδων (cluster validity).Γενικά, υπάρχουν τρεις προσεγγίσεις για τη διερεύνηση της ισχύος συστάδων. Η πρώτη είναι βασισμένη σε εξωτερικά κριτήρια. Αυτό υπονοεί ότι αξιολογούμε τα αποτελέσματα ενός αλγορίθμου ομαδοποίησης βασισμένου σε μια προ-διευκρινισμένη δομή, η οποία επιβάλλεται σε ένα σύνολο δεδομένων και απεικονίζει τη διαίσθησή μας για τη συγκεντρωμένη δομή του συνόλου δεδομένων. Η δεύτερη προσέγγιση είναι βασισμένη σε εσωτερικά κριτήρια. Μπορούμε να αξιολογήσουμε τα αποτελέσματα ενός αλγορίθμου ομαδοποίησης σχετικά με τις ποσότητες διανυσμάτων που περιλαμβάνουν τα σύνολα δεδομένων που οι ίδιοι οι αλγόριθμοι θέτουν. Η τρίτη προσέγγιση ισχύος της αξιολόγησης της ομαδοποίησης είναι βασισμένη στα σχετικά κριτήρια. Εδώ η βασική ιδέα είναι η αξιολόγηση μιας συγκεντρωμένης δομής συγκρίνοντας τη με άλλα πρότυπα ομαδοποίησης, καταλήγοντας στον ίδιο αλγόριθμο αλλά με τις διαφορετικές τιμές παραμέτρου. Υπάρχουν δύο κριτήρια που προτείνονται για την αξιολόγηση της ομαδοποίησης και της επιλογής της βέλτιστης ομαδοποίησης [53]:

1. Η πυκνότητα, τα μέλη κάθε συστάδας πρέπει να είναι όσο το δυνατόν πιο κοντά το ένα στο άλλο. Ένα κοινό μέτρο της πυκνότητας είναι η διαφορά, η οποία πρέπει να ελαχιστοποιηθεί.

2. Ο διαχωρισμός, οι ίδιες συστάδες πρέπει να χωριστούν ευρέως κατά διαστήματα. Υπάρχουν τρεις κοινές προσεγγίσεις μετρώντας την απόσταση μεταξύ δύο διαφορετικών συστάδων:

    - Ενιαίος σύνδεσμος (single linkage): Μετρά την απόσταση μεταξύ των κοντινότερων μελών των συστάδων.

    - Πλήρης σύνδεσμος (Complete linkage): Μετρά την απόσταση μεταξύ των πιο απόμακρων μελών.

    - Σύγκριση κέντρων (Comparison of centroids): Μετρά την απόσταση μεταξύ των κέντρων των συστάδων.

Οι δύο πρώτες προσεγγίσεις είναι βασισμένες στις στατιστικές δοκιμές και το σημαντικό μειονέκτημά τους είναι το υψηλό υπολογιστικό κόστος τους. Επιπλέον, οι δείκτες σχετικοί με αυτές τις προσεγγίσεις στοχεύουν στη μέτρηση του βαθμού στον οποίο ένα σύνολο δεδομένων επιβεβαιώνει ένα διευκρινισμένο εκ των προτέρων πρότυπο. Η τρίτη προσέγγιση στοχεύει στην εύρεση του καλύτερου πρότυπου ομαδοποίησης όπου ένας αλγόριθμος συγκέντρωσης μπορεί να καθοριστεί κάτω από ορισμένες υποθέσεις και παραμέτρους.

6.3 Εξωτερικά κριτήρια

Με βάση τα εξωτερικά κριτήρια μπορούμε να εργαστούμε με δύο διαφορετικούς τρόπους. Αρχικά, μπορούμε να αξιολογήσουμε την προκύπτουσα δομή ομαδοποίησης c, συγκρίνοντας τη με μια ανεξάρτητη διαμέριση των στοιχείων P που φτιάχνονται σύμφωνα με τη διαίσθησή μας για τη συγκεντρωμένη δομή του συνόλου δεδομένων. Αφετέρου, μπορούμε να συγκρίνουμε τη μήτρα/ πίνακα εγγύτητας P, με τη διαμέριση P [53].

6.3.1 Σύγκριση του C με τη διαμέριση P

Θεωρούμε ότι το C ={C1,…,CM} είναι μια δομή ομαδοποίησης ενός συνόλου δεδομένων X και P = {P1,...,Ps} είναι μια καθορισμένη διαμέριση των δεδομένων.Αναφερόμαστε σε ένα ζευγάρι των σημείων (Xv, Xu) από το σύνολο δεδομένων που χρησιμοποιεί τους ακόλουθους όρους:

- SS: εάν και τα δύο σημεία ανήκουν στην ίδια συστάδα της δομής ομαδοποίησης C και στην ίδια ομάδα διαμέρισης P.
- SD: εάν τα σημεία ανήκουν στην ίδια συστάδα του C και στις διαφορετικές ομάδες του P.
- DS: εάν τα σημεία ανήκουν στις διαφορετικές συστάδες του Γ και στην ίδια ομάδα του P.
- DD: εάν και τα δύο σημεία ανήκουν σε διαφορετικές συστάδες του Γ και σε διαφορετικές ομάδες του P

Υποθέτουμε τώρα ότι το A, το B, το C και το D είναι ο αριθμός ζευγαριών SS, SD, DS και DD αντίστοιχα, τότε a+b+c+d =M το ποίο είναι ο μέγιστος αριθμός όλων των ζευγαριών στο σύνολο δεδομένων (που σημαίνει, M = N(N − 1)/2 όπου το N είναι ο συνολικός αριθμός των σημείων στο σύνολο δεδομένων).

Τώρα μπορούμε να καθορίσουμε τους ακόλουθους δείκτες για να μετρήσουμε το βαθμό ομοιότητας μεταξύ του C και του P:

- Στατιστική ακρών: R = (a + d) /M,
- Jaccard Coefficient: J = a/(a + b + c),

Οι ανωτέρω δύο δείκτες παίρνουν τις τιμές μεταξύ 0 και 1, και μεγιστοποιούνται όταν m = s.

Ένας άλλος δείκτης είναι:

- Ο δείκτης Mallows και Folkes

$$FM = a|\sqrt{m_1 m_2} = \sqrt{\frac{a}{a+b}\frac{a}{a+c}}$$

Όπου m1 = (a + b), m2 = (a + c).Για τους προηγούμενους τρεις δείκτες έχει αποδειχθεί ότι οι υψηλές αξίες των δεικτών δείχνουν τη μεγάλη ομοιότητα μεταξύ του C και του P. Όσο ψηλότερες οι τιμές αυτών των δεικτών είναι το τόσο παρόμοιο το C και το P είναι. Άλλοι δείκτες είναι:

- Huberts Γ στατιστική:

$$\Gamma = \left({}^{1}/_{M}\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} X(i,j)Y(i,j)$$

Οι υψηλές αξίες αυτού του δείκτη υποδεικνύουν μια ισχυρή ομοιότητα μεταξύ του X και του Y.

- Ομαλοποιημένη Γ στατιστική:

$$\bar{\Gamma} = \frac{\left[{}^{1}/_{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X(i,j)-\mu_x)(Y(i,j)-\mu_Y)\right]}{\sigma X \sigma Y}$$

Όπου καιY (i, j) είναι τα (i, j) στοιχεία των μητρών X, Y που αντίστοιχα πρέπει να συγκρίνουμε. Επίσης *μx*, *μy*, *σx, σy* είναι τα αντίστοιχα μέσα και οι διαφορές του X, μήτρες Y. Αυτός ο δείκτης παίρνει τις τιμές μεταξύ −1 και 1.

6.4 Εσωτερικά κριτήρια

Χρησιμοποιώντας αυτήν την προσέγγιση της ισχύος συστάδων ο στόχος μας είναι να αξιολογήσουμε συγκεντρωτικά, το αποτέλεσμα ενός αλγορίθμου που χρησιμοποιεί μόνο τις έμφυτες ποσότητες και χαρακτηριστικά γνωρίσματα στο σύνολο δεδομένων. Υπάρχουν δύο περιπτώσεις στις οποίες εφαρμόζουμε τα εσωτερικά κριτήρια της ισχύος συστάδων ανάλογα με τη συγκεντρωμένη δομή: α) ιεραρχία της συγκέντρωσης των προτύπων, και β) ενιαίο σχέδιο συγκέντρωσης.

6.4.1 Αξιοπιστία ιεραρχιών ομαδοποιήσεων

Ένα δενδροδιάγραμμα που παράγεται μέσω ενός ιεραρχικού αλγορίθμου ομαδοποίησης, μπορεί να αναπαρασταθεί μέσω ενός cophenetic πίνακα P_c. Θα καθοριστούν λοιπόν οι στατιστικοί δείκτες που μετρούν τον βαθμό ομοιότητας μεταξύ του cophenetic πίνακα P_c, ο οποίος παράγεται μέσω ενός συγκεκριμένου ιεραρχικού αλγόριθμου ομαδοποίησης, με τον πίνακα εγγύτητας P του X. Επειδή και οι δύο πίνακες είναι συμμετρικοί και έχουν τα διαγώνια στοιχεία τους ίσα με το μηδέν, μελετούνται μόνο τα $M = \frac{N(N-1)}{2}$ άνω διαγώνια στοιχεία των P_c και P. Ο δείκτης αυτός είναι γνωστός ως cophenetic συντελεστής συσχέτισης (CPCC). Μετρά την συσχέτιση μεταξύ των πινάκων P και P_c. Χρησιμοποιείται δε όταν τα δεδομένα είναι διαστήματα (interval) ή λόγοι (ratio). Ορίζεται ως: [53]

$$CPCC = \frac{\left(1/M\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}c_{ij} - \mu_P\mu_c}{\sqrt{\left[\left(1/M\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2\right]\left[\left(1/M\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_C^2\right]}}, -1 \le CPCC \le 1$$

Όπου $M = \frac{N(N-1)}{2}$ και το N είναι ο αριθμός σημείων στο διάγραμμα δεδομένων. Επίσης, μ_pκαι μ_cείναι οι μέσοι των μητρών P και Pc αντίστοιχα, και ορίζονται ως:

$$\mu_P = \left(1/M\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i,j),\ \mu_C = \left(1/M\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P_C(i,j)$$

Επιπλέον, οι $D_{i\,j}$, $C_{i\,j}$είναι τα (ι, j) στοιχεία των μητρών P και PC αντίστοιχα. Μια αξία του δείκτη κοντά στο 0 είναι μια ένδειξη μιας σημαντικής ομοιότητας μεταξύ των δύο μητρών.

6.5 Σχετικά κριτήρια

Η βάση των ανωτέρω περιγεγραμμένων μεθόδων επικύρωσης είναι η στατιστική δοκιμή. Κατά συνέπεια, το σημαντικότερο μειονέκτημα των τεχνικών βασισμένων στα εσωτερικά ή εξωτερικά κριτήρια είναι η υψηλή υπολογιστική ζήτησή τους. Μια διαφορετική προσέγγιση επικύρωσης συζητείται σε αυτό το τμήμα. Είναι βασισμένο στα σχετικά κριτήρια και δεν περιλαμβάνει τις στατιστικές δοκιμές. Η θεμελιώδης ιδέα αυτής της προσέγγισης είναι να επιλεχτεί το καλύτερο σχέδιο συγκέντρωσης ενός συνόλου καθορισμένων προτύπων σύμφωνα με ένα προ-διευκρινισμένο κριτήριο. Πιο συγκεκριμένα, το πρόβλημα μπορεί να οριστεί ως εξής [53]:

Θεωρούμε Palg το σύνολο παραμέτρων που συνδέονται με έναν συγκεκριμένο αλγόριθμο συγκέντρωσης (π.χ. ο αριθμός συστάδων nc). Μεταξύ των σχεδίων C_i συγκέντρωσης, $i = 1, \dots, nc$, που καθορίζεται από έναν συγκεκριμένο αλγόριθμο, για τις διαφορετικές τιμές των παραμέτρων σε Palg, επιλέγει αυτό που ταιριάζει καλυτέρα στο σύνολο δεδομένων.

Κατόπιν, μπορούμε να εξετάσουμε τις ακόλουθες περιπτώσεις του προβλήματος:

1. Το Palg δεν περιέχει τον αριθμό συστάδων, nc, ως παράμετρο. Σε αυτήν την περίπτωση, η επιλογή των βέλτιστων τιμών παραμέτρου περιγράφεται ως εξής: Τρέχουμε τον αλγόριθμο για ένα ευρύ φάσμα τιμών των παραμέτρων και επιλέγουμε τη μεγαλύτερη σειρά για την οποία η nc παραμένει σταθερά (συνήθως $nc \ll N$ (αριθμός περιπτώσεων / δεδομένων)). Κατόπιν επιλέγουμε ανάλογα με την περίπτωση τιμές των παραμέτρων Palg οι τιμές που αντιστοιχούν στη μέση αυτής της σειράς. Επίσης, αυτή η διαδικασία προσδιορίζει τον αριθμό συστάδων που κρύβονται κάτω από το σύνολο δεδομένων μας.

2. Το Palg περιέχει το nc ως παράμετρο. Η διαδικασία της ταυτοποίηση του καλύτερου σχεδίου ομαδοποίησης είναι βασισμένη σε έναν δείκτη ισχύος. Επιλέγοντας έναν κατάλληλο δείκτη απόδοσης, q, συνεχίζουμε με τα ακόλουθα βήματα:

    - Τρέχουμε τον αλγόριθμο συγκέντρωσης για όλες τις τιμές του nc μεταξύ ενός ελάχιστου $ncmin$ και ενός μέγιστου $ncmax$. Οι ελάχιστες και μέγιστες τιμές είναι καθορισμένο εκ των προτέρων από το χρήστη.

    - Για κάθε μια από τις τιμές του nc, τρέχουμε τους χρόνους αλγορίθμου r, χρησιμοποιώντας το διαφορετικό σύνολο τιμών για τις άλλες παραμέτρους του αλγορίθμου (π.χ. διαφορετικοί αρχικοί όροι).

    - Σχεδιάζουμε τις καλύτερες τιμές του δείκτη q που λαμβάνεται από κάθε nc ως λειτουργία του nc.

Βασισμένοι σε αυτό το γράφημα μπορούμε να προσδιορίσουμε το καλύτερο σχέδιο ομαδοποίησης. Πρέπει να τονίσουμε ότι υπάρχουν δύο προσεγγίσεις για τον καθορισμό της καλύτερης ομαδοποίησης ανάλογα με τη συμπεριφορά του q όσον αφορά το nc. Κατά συνέπεια, εάν ο δείκτης ισχύος δεν εκθέτει μια αυξανόμενη ή μειωμένο τάση ως nc αυξήσεις επιδιώκουμε το μέγιστο (ελάχιστο) του plot.Αφ' ενός, για τους δείκτες που αυξάνονται (μειώνονται) καθώς ο αριθμός συστάδων αυξάνεται ψάχνουμε για τις τιμές του nc στις οποίες εμφανίζεται μια σημαντική τοπική αλλαγή στην αξία του δείκτη. Αυτή η αλλαγή εμφανίζεται ως «γόνατο» στο plot και είναι μια ένδειξη του αριθμού των συστάδων που υπάρχει κάτω από το σύνολο δεδομένων. Επιπλέον, η απουσία ενός γονάτου μπορεί να είναι μια ένδειξη ότι το σύνολο δεδομένων δεν κατέχει καμία δομή συγκέντρωσης [53].

6.5.1 Αξιοπιστία Διαμεριστικών Αλγορίθμων Ομαδοποίησης

Δείκτης Dunn

Έστω ότι η συνάρτηση ανομοιότητας μεταξύ δύο ομάδων C_i και C_j δίνεται από την σχέση:

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_i} d(x, y),$$

diam(c) είναι η διάμετρος μιας συστάδας, η οποία μπορεί να θεωρηθεί ως μέτρο της διασποράς των συστάδων. Η διάμετρος της συστάδας c ορίζεται ως:

$$diam(C) = \max_{x,y \in C}(x, y)$$

Δηλαδή, η διάμετρος μιας ομάδας C είναι η απόσταση των δυο πιο μακρινών διανυσμάτων. Η διάμετρος $Diam(C)$, μπορεί να θεωρηθεί ως ένα μέτρο της διασποράς της C. Οπότε ο δείκτης Dunn για ένα συγκεκριμένο m ορίζεται ως εξής:

$$D_m = \min_{i=1,\dots,m} \left\{ \min_{j=i+1,\dots,m} \left(\frac{d(C_i, C_j)}{\max_{k=1,\dots,m} diam(C_k)} \right) \right\}$$

Είναι σαφές ότι εάν το σύνολο δεδομένων περιέχει τις συμπαγείς και καλά-χωρισμένες συστάδες, η απόσταση μεταξύ των συστάδων αναμένεται να είναι μεγάλη και η διάμετρος των συστάδων αναμένεται να είναι μικρή. Κατά συνέπεια, βασισμένοι στον καθορισμό δεικτών του Dunn, μπορούμε να καταλήξουμε στο συμπέρασμα ότι οι μεγάλες τιμές του δείκτη δείχνουν την παρουσία συμπαγών και καλά-χωρισμένων συστάδων.

Οι επιπτώσεις του δείκτη Dunn είναι:

1. Το μεγάλο χρονικό διάστημα που απαιτείται για τον υπολογισμό του.

2. Η ευαισθησία στην παρουσία θορύβου στα σύνολα δεδομένων δεδομένου ότι αυτά είναι πιθανό να αυξήσουν τις τιμές $diam(c)$

3. Τρεις δείκτες, προτείνονται οι PAL και Biswas (1997) που είναι πιο ανθεκτικοί στην παρουσία θορύβου. Είναι ευρέως γνωστοί ως οι δείκτες Dunn, δεδομένου ότι είναι βασισμένοι στο δείκτη Dunn. Επιπλέον, οι τρεις δείκτες για τον καθορισμό τους, χρησιμοποιούν έννοιες όπως: του ελάχιστα εκτεινόμενο δέντρο, (MST), η σχετική γραφική παράσταση γειτνίασης (RNG) και η γραφική παράσταση του Gabriel αντίστοιχα.

Δείκτης Davies-Bouldin (DB)

Το μέτρο R_{ij} ομοιότητας μεταξύ των συστάδων C_i και το C_jκαθορίζεται βασισμένο σε ένα μέτρο της διασποράς μιας συστάδας C_i και ένα μέτρο ανομοιότητας μεταξύ δύο clustersd_{ij}. Ο R_{ij} δείκτης καθορίζεται για να ικανοποιήσει τους ακόλουθους όρους:

1. $R_{i\,j} \geq 0$ $R_{i\,j} = R_{ij}$
2. if $s_i = 0$ and $s_{i\,j} = 0$ then $R_{i\,j} = 0$
3. If$s_{\,j} > s_k$and$d_{i\,j} = d_{i\,k}$ then$R_{i\,j} \geq R_{i\,k}$
4. if $s_{\,j} = s_k$ and $d_{i\,j} < d_{i\,k}$ then $R_{i\,j} < R_{i\,k}$

Αυτοί οι όροι δηλώνουν ότι R_{ij} είναι μη αρνητικό και συμμετρικό. Μια απλή επιλογή για τοR_{ij} που ικανοποιεί τους ανωτέρω όρους είναι οι Davies και Bouldin:

$R_{ij} = (s_i + s_j)/d_{ij}$

Κατόπιν ο δείκτης DB ορίζεται ως:

$$DB_{nc} = \frac{1}{nc}\sum_{i=1}^{nc} R_i$$

$$R_i = \max_{i=1,\ldots,nc, i\neq j} R_{ij}, i = 1, \ldots, nc$$

Είναι σαφές για τον ανωτέρω καθορισμό ότι DB_{nc}είναι η μέση ομοιότητα μεταξύ κάθε συστάδας C_i, i =1,…, nc και του πιο όμοιο του. Είναι επιθυμητό για τις συστάδες να έχουν την ελάχιστη πιθανή ομοιότητα η μια με την άλλη επομένως, επιδιώκουμε τις συγκεντρώσεις που ελαχιστοποιούν το DB.Ο DB_{nc} δείκτης δεν εκθέτει καμία τάση όσον αφορά τον αριθμό συστάδων και έτσι επιδιώκουμε την ελάχιστη αξία DB_{nc}στο plot της ενάντια του αριθμού συστάδων [53].

6.6 Άλλα Μέτρα Αξιολόγησης

Εκτός από τα παραπάνω μέτρα αξιολόγησης της ομαδοποίησης, υπάρχουν και μέτρα τα οποία χρησιμοποιούνται όταν είναι γνωστή η πραγματική κατηγορία στην οποία ανήκουν τα πρότυπα δεδομένα. Δύο από αυτά τα μέτρα, είναι η εντροπία (entropy) και η καθαρότητα (purity). Η εντροπία αντιπροσωπεύει την ανομοιότητα των στοιχείων που βρίσκονται σε ένα σύμπλεγμα (cluster). Όσο υψηλότερη είναι η ομοιογένεια υπάρχει στα σημεία (και τις ιδιότητες των σημείων), τόσο περισσότερο η τιμή της εντροπίας πλησιάζει το μηδέν. Ωστόσο, για να γίνει χρήση της λειτουργίας της εντροπίας, απαιτείται να προϋπάρχει η γνώση της πραγματικής ταξινόμησης – κατηγοριοποίησης των σημείων.

Έστω $C = \{C_1, C_2, \dots, C_k\}$ μια ομαδοποίηση που προκύπτει από έναν αλγόριθμο ομαδοποίησης και $L = \{L_1, L_2, \dots, L_m\}$ ο στόχος ταξινόμησης των προτύπων (pattern), τότε η **εντροπία** της κάθε ομάδας C_i ορίζεται ως:

$$H_i = -\sum_{j=1}^{m} P(x \in L_j / x \in C_i) \log P(x \in L_j / x \in C_i).$$

Για ένα δεδομένο σύνολο προτύπων (pattern), η εντροπία του συνόλου της ομαδοποίησης, είναι ο σταθμισμένος μέσος της εντροπίας του κάθε cluster [52].

Η εντροπία δίνει μια πιο σφαιρική εικόνα από ό, τι καθαρότητα αφού αντί να εξετάζει απλά τον αριθμό των αντικειμένων της κυρίαρχης κλάσης σε "περιλαμβάνεται" (in) και "δεν περιλαμβάνεται" (not in) , λαμβάνει το σύνολο της κατανομής υπόψη. Δεδομένου ότι ένα σύμπλεγμα με όλα τα αντικείμενα από την ίδια κατηγορία έχει εντροπία 0, ορίζουμε την βάση εντροπίας (entropy based) ως 1 μείον το [0,1] της κανονικοποιημένης εντροπίας. Η βάση εντροπίας της κάθε ομάδας ορίζεται ως:

$$\Phi^{(Entropy)}(C_l, k) = 1 - \sum_{h=1}^{g} -\frac{n_l{}^{(h)}}{n_l} \log_g \left(\frac{n_l^{(h)}}{n_l} \right)$$

Η συνολική μέτρηση της ποιότητας της εντροπίας συνεπάγεται ότι είναι [53]:

$$\Phi^{(Entropy)}(\lambda, k) = 1 + \frac{1}{n} \sum_{l=1}^{k} \sum_{h=1}^{g} n_l^{(h)} \log_g \left(\frac{n_l^{(h)}}{n_l} \right)$$

Η **καθαρότητα (purity)**, ορίζεται ως $r = \frac{1}{n} \sum_{i=1}^{k} a_i$ οπου το k δηλώνει τον αριθμό των ομάδων που βρέθηκαν στο σύνολο των δεδομένων και το a_i αντιπροσωπεύει τον αριθμό των μοντέλων της κλάσης στην οποία ανήκει η πλειοψηφία των σημείων i. Όσο μεγαλύτερη είναι η τιμή της καθαρότητας, τόσο καλύτερα έχει επιτευχθεί η ομαδοποίηση. ($0 \leq H \geq 1$, $0 \leq r \geq 1$) [52].

Η καθαρότητα μπορεί να ερμηνευθεί ως η ακρίβεια της ταξινόμησης με βάση την υπόθεση ότι όλα τα αντικείμενα μιας συστάδας που ταξινομούνται να είναι μέλη της κυρίαρχης κατηγορίας για την εν

λόγω συστάδα. Για ένα ενιαίο σύμπλεγμα, C, η καθαρότητα ορίζεται ως η σχέση του αριθμού των αντικειμένων στην κυρίαρχη κατηγορία με το συνολικό αριθμό των αντικειμένων:

$$\Phi^{(Purity)}(C_l, k) = \frac{1}{n_l}\max_h\left(n_l{}^{(h)}\right)$$

Για να αξιολογηθεί στο σύνολο της μια ομαδοποίηση, πρέπει να υπολογιστεί ο μέσος όρος της καθαρότητας της κάθε ομάδας σταθμισμένος ανάλογα με το μέγεθος της ομάδας.

$$\Phi^{(Purity)}(\lambda, k) = \frac{1}{n}\sum_{l=1}^{k}\max_h\left(n_l{}^{(h)}\right)$$

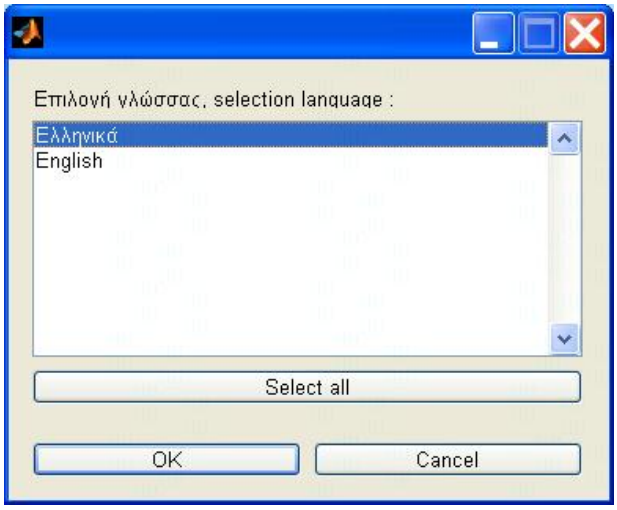
Και η καθαρότητα και η εντροπία είναι μεροληπτικές στο να ευνοούν έναν μεγάλο αριθμό ομάδων. Στην πραγματικότητα και για τα δύο αυτά κριτήρια, η παγκόσμια βέλτιστη τιμή επιτυγχάνεται όταν κάθε συστάδα έχει ένα αντικείμενο [53]

7. Υλοποίηση γραφικού περιβάλλοντος αλγορίθμων

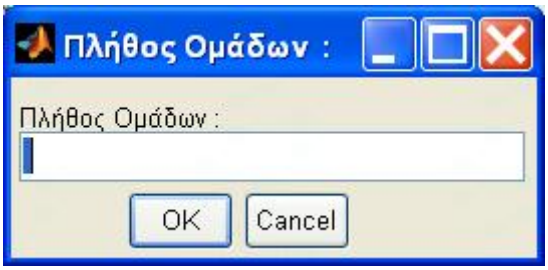
Σε αυτό το κεφάλαιο γίνεται εξέταση της συμπεριφοράς ενός γνωστού αλγόριθμου ομαδοποίησης (k-means) και ενός αλγόριθμου ταξινόμησης (k-nn) σε διάφορα σύνολα δεδομένων και με διαφορετικές παραμετροποιήσεις. Για τον λόγο αυτό, αναπτύξαμε ένα γραφικό περιβάλλον εύχρηστο προς τον απλό και μη εξοικειωμένο χρήστη, στο οποίο θα μπορεί να κάνει τις επιλογές του και να εξετάσει την συμπεριφορά των δυο αλγορίθμων γραφικά ή / και με χρήση δυο γνωστών μέτρων αξιολόγησης, της εντροπίας (Entropy) και της καθαρότητας (Purity).

7.1 Υλοποίηση αλγορίθμου k-means

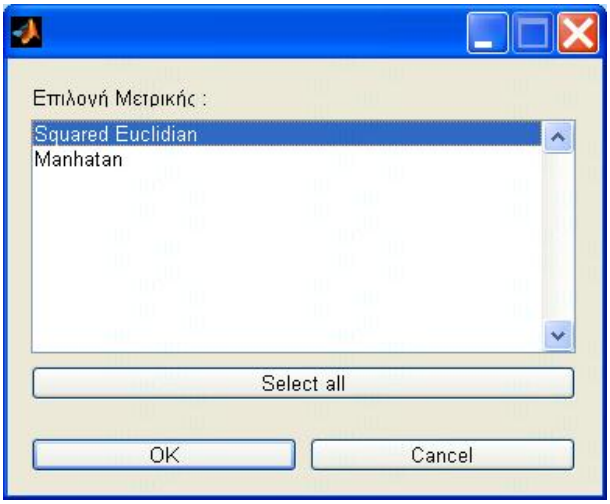
Επιλογή γλώσσας



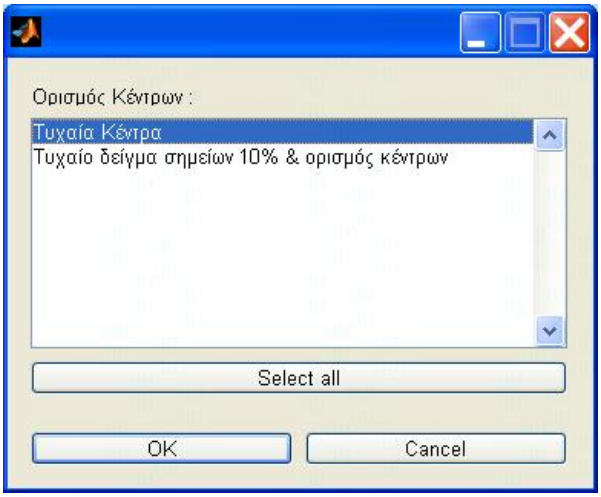
Αριθμός ομάδων που θέλουμε να ομαδοποιήσουμε τα στοιχεία



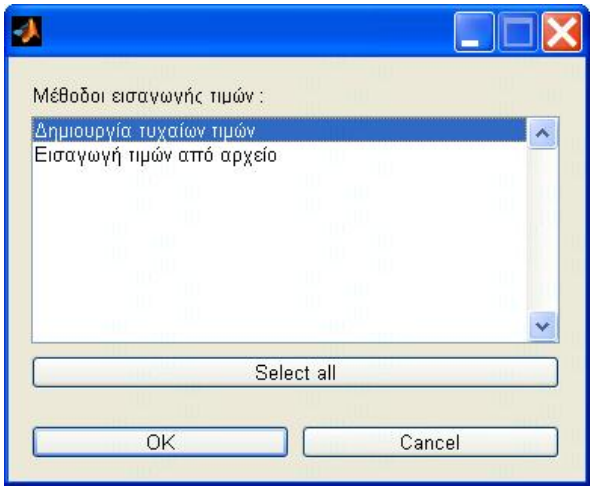
Επιθυμητή χρήση μετρικής για την ομαδοποίηση



Μέθοδος ορισμού των κέντρων των ομάδων



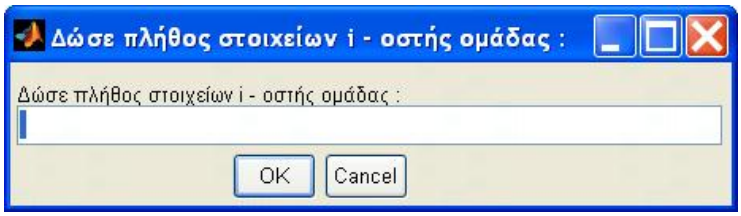
Μέθοδος εισαγωγής στοιχείων προς ομαδοποίηση



Σε αυτό το σημείο ο αλγόριθμος ακολουθεί διαφορετικές διαδρομές αναλόγως με την μέθοδο.

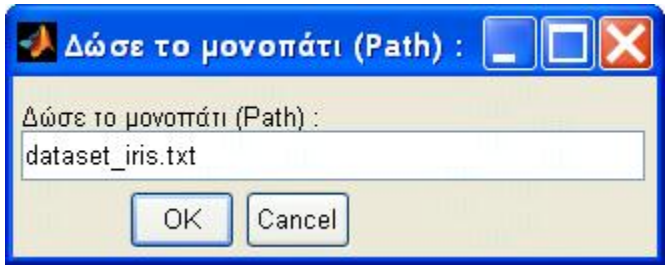
Δημιουργία τυχαίων τιμών:

Πληκτρολόγηση αριθμού στοιχείων που θα περιέχει η κάθε ομάδα
(το στάδιο επαναλαμβάνεται για όσες ομάδες έχουμε ζητήσει)



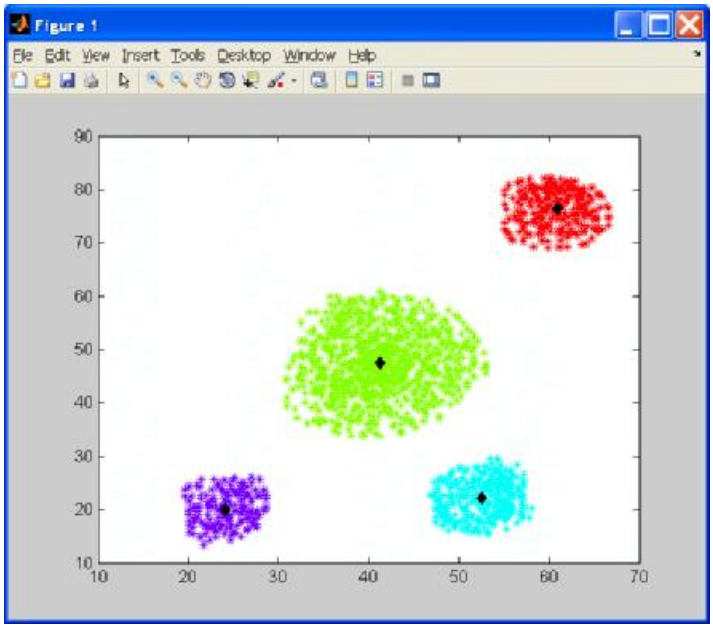
Εισαγωγή στοιχείων από αρχείο txt

Πληκτρολόγηση διαδρομής προσπέλασης που βρίσκεται το αρχείο

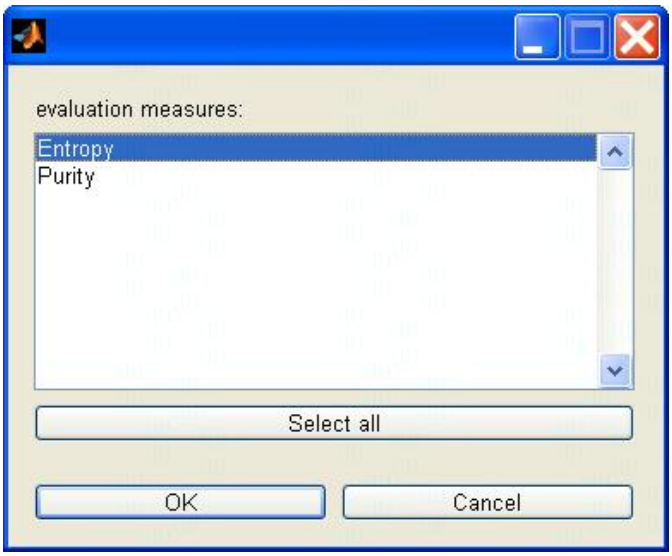


Ο αλγόριθμος σε αυτό το σημείο, για οποιαδήποτε μέθοδο εισαγωγής, συνεχίζει ως εξής:

Εμφάνιση του γραφήματος

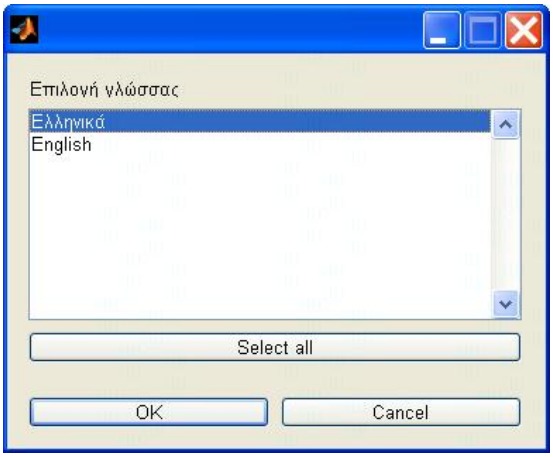


Επιλογή του μέτρου εγγύτητας που θα χρησιμοποιηθεί

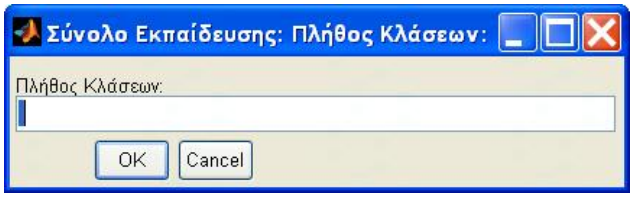

evaluation measures:
Entropy
Purity
Select all
OK
Cancel

7.2 Υλοποίηση αλγορίθμου k-nn

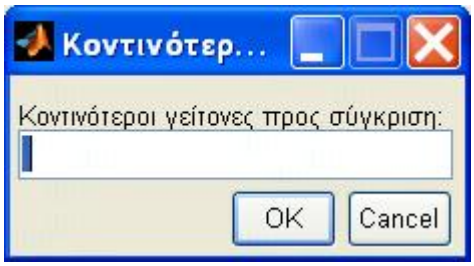
Επιλογή γλώσσας



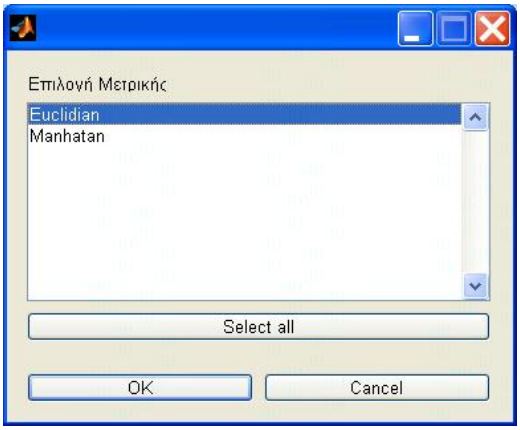
Αριθμός κλάσεων που θέλουμε να ταξινομήσουμε τα στοιχεία



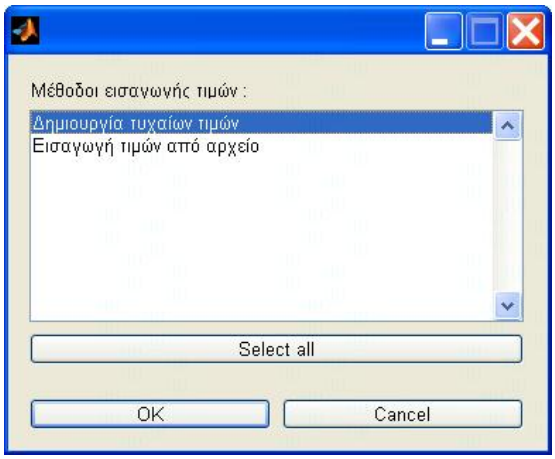
Επιλογή αριθμού γειτονικών στοιχείων προς σύγκριση



Επιθυμητή χρήση μετρικής για την ταξινόμηση



Μέθοδος εισαγωγής στοιχείων προς ταξινόμηση



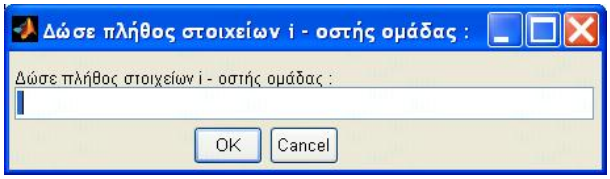
Σε αυτό το σημείο ο αλγόριθμος ακολουθεί διαφορετικές διαδρομές αναλόγως με την μέθοδο.

Δημιουργία τυχαίων τιμών:

Πληκτρολόγηση αριθμού στοιχείων που θα περιέχει η κάθε ομάδα

(το στάδιο επαναλαμβάνεται για όσες ομάδες έχουμε ζητήσει)

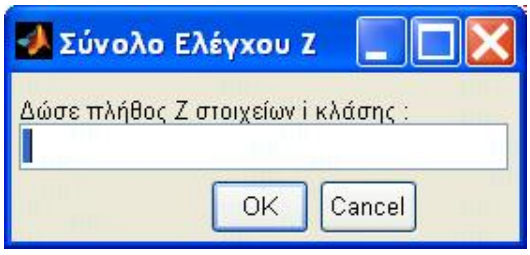
(train)



Πληκτρολόγηση αριθμού στοιχείων που θα περιέχει η κάθε ομάδα

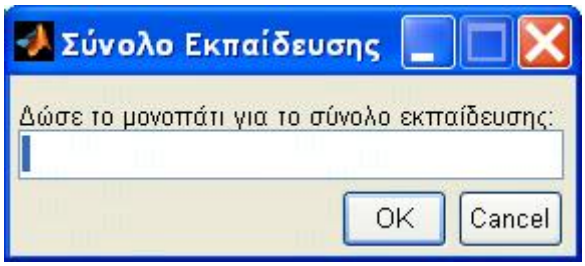
(το στάδιο επαναλαμβάνεται για όσες ομάδες έχουμε ζητήσει)

(test)

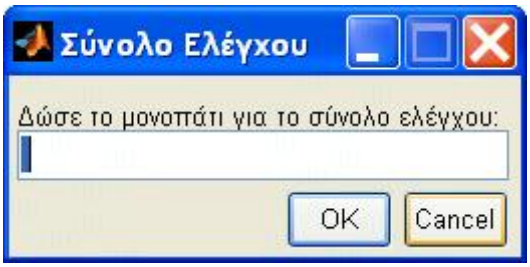


Εισαγωγή τιμών από αρχείο

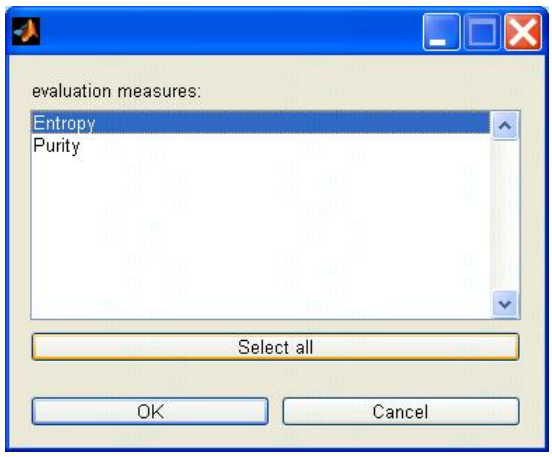
Πληκτρολόγηση διαδρομής προσπέλασης που βρίσκεται το αρχείο εκπαίδευσης



Πληκτρολόγηση διαδρομής προσπέλασης που βρίσκεται το αρχείο ελέγχου



Επιλογή του μέτρου εγγύτητας που θα χρησιμοποιηθεί



7.3 Αποτελέσματα αλγορίθμων

Στους πίνακες που παρουσιάζονται παρακάτω γίνεται αντιληπτός ο τρόπος που δουλεύουν οι αλγόριθμοι, όσον αφορά την ακεραιότητά τους, σε προκατασκευασμένες για αυτή την χρήση περιπτώσεις. Μελετάται λοιπόν η εντροπία και η καθαρότητα για τους 2 αλγόριθμους, και για το εύρος των διαφορετικών περιπτώσεων που αυτοί καλύπτουν (ορισμό κέντρων, επιλογή μετρικής, μέθοδοι εισαγωγής τιμής κ.ο.κ). Για να μελετηθούν και να αξιοποιηθούν το βέλτιστο οι τιμές των μέτρων, πρέπει να υπολογιστούν αρκετές φορές για κάθε συγκεκριμένη μέθοδο και στο τέλος να υπολογιστεί η μέση τιμή τους, κάτι που δεν ακολουθήθηκε στους συγκεκριμένους πίνακες, αλλά τα αποτελέσματα είναι αυτά που προκύπτουν μετά την πρώτη χρήση. Είναι σκόπιμο λοιπόν να μελετηθούν κάποιες ιδιαίτερες περιπτώσεις από τους πίνακες, στις οποίες το αποτέλεσμα παρουσιάζει διαφοροποίηση από το αναμενόμενο.

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την περιγραφή του αλγορίθμου k-means είναι τεχνητά σύνολα δεδομένων των οποίων οι ομάδες έχουν τυχαίο αυθαίρετο σχήμα. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν στον αλγόριθμο k-NN είναι τεχνητά σύνολα που προέρχονται από την κανονική κατανομή με διάφορες τιμές των παραμέτρων (μέση τιμή, συνδιασπορά) αλλά και ένα πραγματικό σύνολο δεδομένων προερχόμενο από την βάση δεδομένων UCI Machine Learning Repository [54].

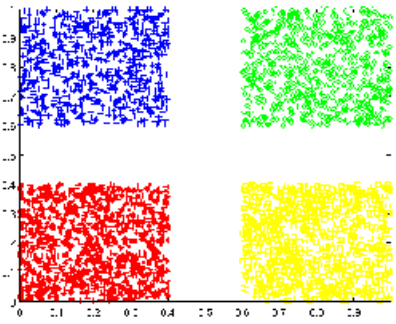
7.3.1 Αποτελέσματα k-means

Περιγραφή συνόλου δεδομένων

Τα χαρακτηριστικά των συνόλων είναι τα εξής:

1. Τα δεδομένα για τον πίνακα 1 (corners) είναι αυθαιρέτου σχήματος.
   Τα χαρακτηριστικά των συνόλων είναι τα εξής:

   Πλήθος στοιχείων 4000
   Κλάση 1: 1024
   Κλάση 2: 986
   Κλάση 3: 980
   Κλάση 4: 1010

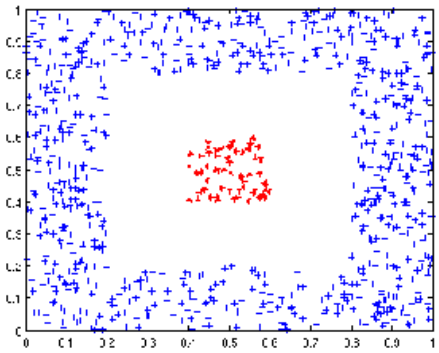


2. Τα δεδομένα για τον πίνακα 2 (nested) είναι αυθαιρέτου σχήματος.
   Τα χαρακτηριστικά των συνόλων είναι τα εξής:

Πλήθος στοιχείων 1000

Κλάση 1: 62

Κλάση 2: 938



3. Τα δεδομένα για τον πίνακα 3 (dset1) είναι αυθαιρέτου σχήματος.

Τα χαρακτηριστικά των συνόλων είναι τα εξής:

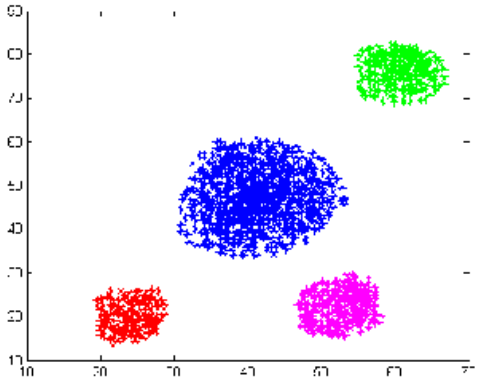
Πλήθος στοιχείων 1600

Κλάση 1: 859

Κλάση 2: 161

Κλάση 3: 265

Κλάση 4: 315



4. Τα δεδομένα για τον πίνακα 4 (dset2) είναι αυθαιρέτου σχήματος.

Τα χαρακτηριστικά των συνόλων είναι τα εξής:

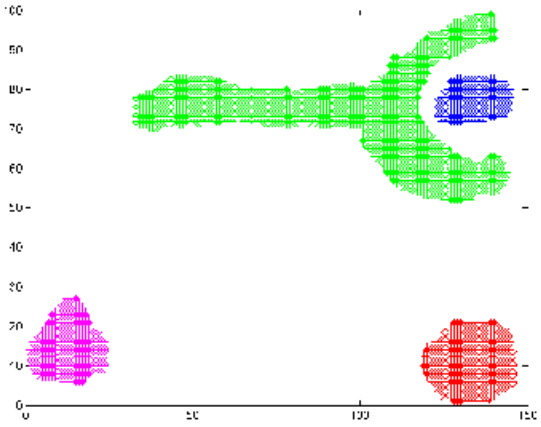
Πλήθος στοιχείων 2761

Κλάση 1: 361

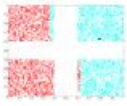
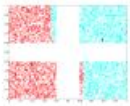
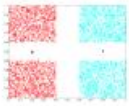
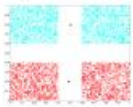
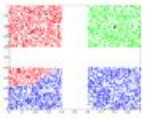
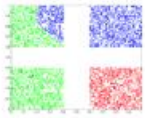
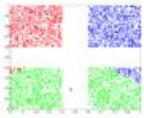
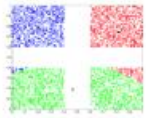
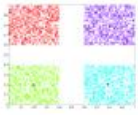
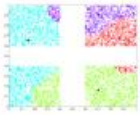
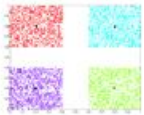
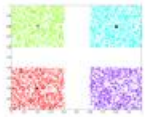
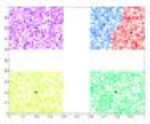
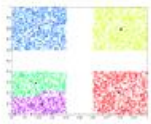
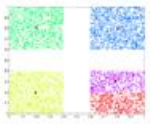
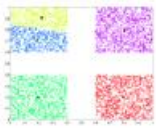
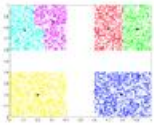
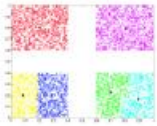
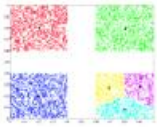
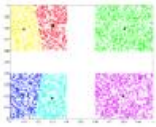
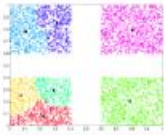
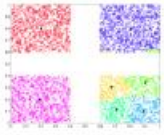
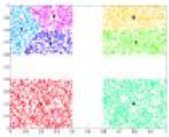
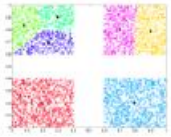
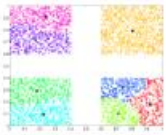
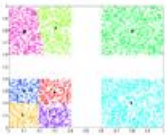
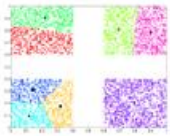
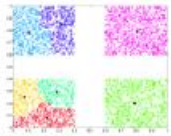
Κλάση 2: 480

Κλάση 3: 236

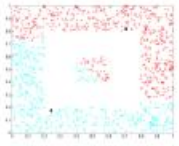
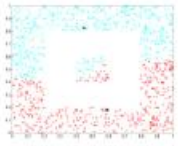
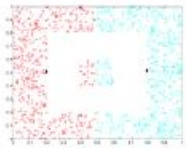
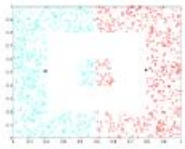
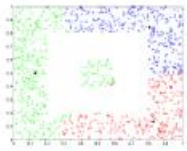
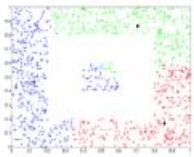
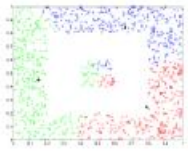
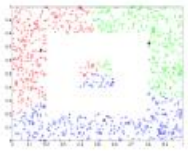
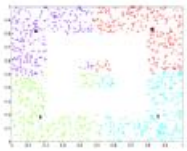
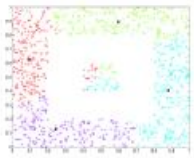
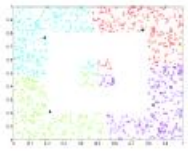
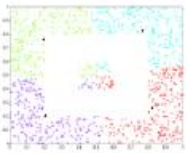
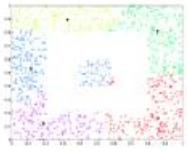
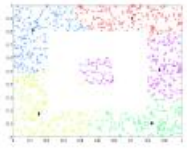
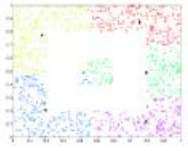
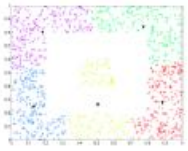
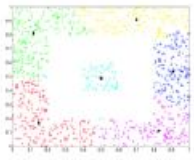
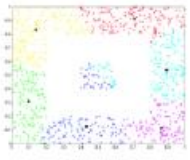
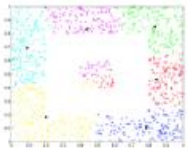
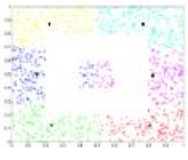
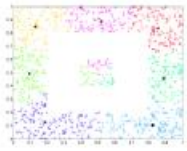
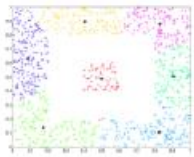
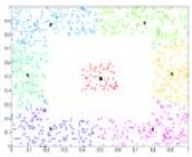
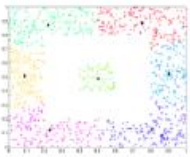
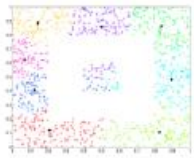
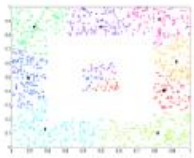
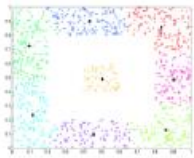
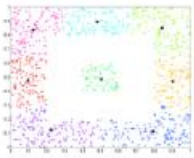
Κλάση 4: 1684



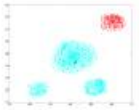
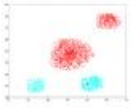
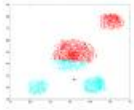
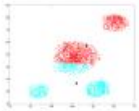
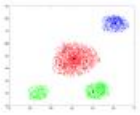
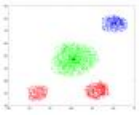
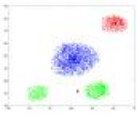
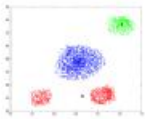
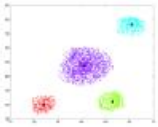
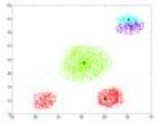
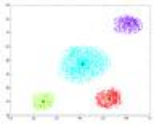
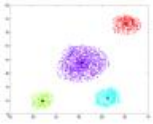
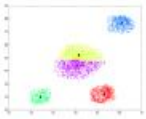
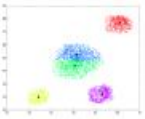
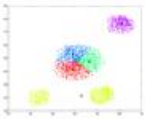
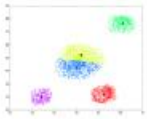
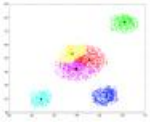
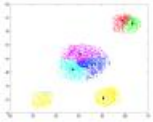
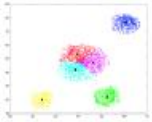
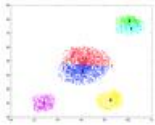
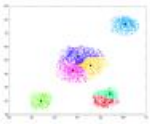
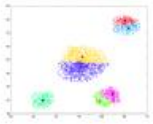
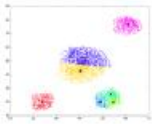
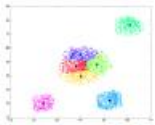
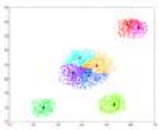
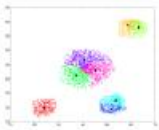
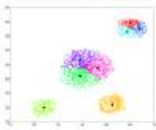
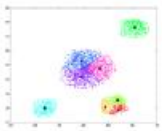
Ανάλυση αποτελεσμάτων

Corners								
ΠΛΗΘΟΣ ΟΜΑΔΩΝ	Manhattan				Euclidean			
	10%		TK		10%		TK	
	Entropy	Purity	Entropy	Purity	Entropy	Purity	Entropy	Purity
K=2	1.2584	0.5085	1.2584	0.5085	0.9998	0.5085	0.9998	0.5085
								
K=3	0.6708	0.7440	0.6861	0.7440	0.6534	0.7445	0.6473	0.7485
								
K=4	0	1	0.7928	0.7255	0	1	0	1
								
K=5	0	1	0.7751	0.7085	0	1	0	1
								
K=6	0	1	0.0172	0.9972	0	1	0	1
								
K=7	0	1	0.0062	0.9992	0	1	0	1
								
K=8	0	1	0	1	0	1	0	1
								

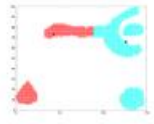
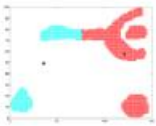
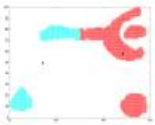
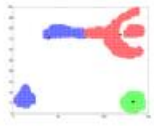
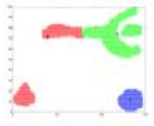
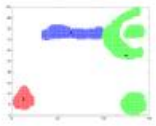
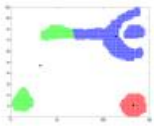
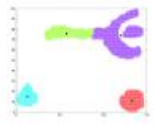
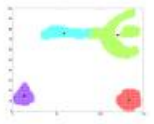
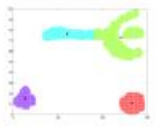
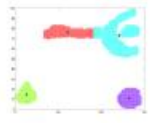
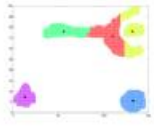
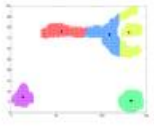
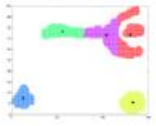
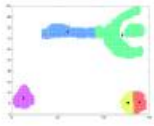
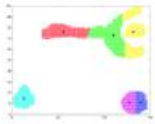
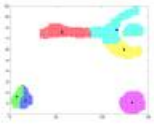
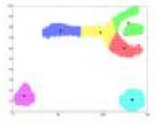
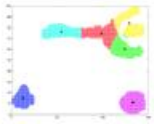
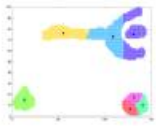
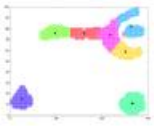
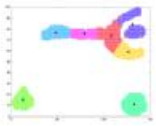
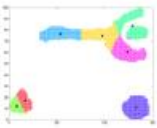
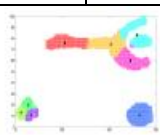
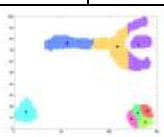
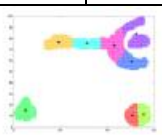
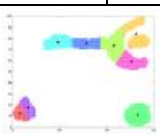
Πίνακας 1

Nested								
ΠΛΗΘΟΣ ΟΜΑΔΩΝ	Manhattan				Euclidean			
	10%		TK		10%		TK	
	Entropy	Purity	Entropy	Purity	Entropy	Purity	Entropy	Purity
K=2	0.3350	0.9380	0. 3353	0.9380	0.3328	0.9380	0.3328	0.9380
								
K=3	0.3070	0.9380	0.2851	0.9380	0.3333	0.9380	0.3333	0.9380
								
K=4	0.3283	0.9380	0.3283	0.9380	0.3310	0.9380	0.3310	0.9380
								
K=5	0.2431	0.9380	0.2077	0.9380	0.1516	0.9380	0.2494	0.9380
								
K=6	0.2234	0.9380	0.2384	0.9380	0.2550	0.9380	0.1269	0.9380
								
K=7	0.2539	0.9380	0.0128	0.9980	0	1	0.2539	0.9380
								
K=8	0.2432	0.9380	0.0128	0.9980	0.0074	0.9990	0.0074	0.9990
								

Πίνακας 2

dset1								
ΠΛΗΘΟΣ ΟΜΑΔΩΝ	Manhattan				Euclidean			
	10%		TK		10%		TK	
	Entropy	Purity	Entropy	Purity	Entropy	Purity	Entropy	Purity
K=2	1.0588	0.7025	0.8281	0.7338	1.1335	0.6169	1.1335	0.6169
								
K=3	0.2746	0.8994	0.2746	0.8994	0.3037	0.8956	0.2746	0.8994
								
K=4	0	1	0.2746	0.8994	0	1	0	1
								
K=5	0	1	0	1	0.2746	0.8994	0	1
								
K=6	0	1	0.3013	0.8994	0	1	0	1
								
K=7	0	1	0	1	0	1	0	1
								
K=8	0	1	0	1	0	1	0	1
								

Πίνακας 3

dset2								
ΠΛΗΘΟΣ ΟΜΑΔΩΝ	Manhattan				Euclidean			
	10%		TK		10%		TK	
	Entropy	Purity	Entropy	Purity	Entropy	Purity	Entropy	Purity
K=2	0.6741	0.7838	1.2020	0.6099	1.1991	0.6099	1.1991	0.6099
								
K=3	0.6380	0.7838	0.6360	0.7838	0.8466	0.7407	0.6142	0.7838
								
K=4	0.3260	0.9145	0.3260	0.9145	0.3241	0.9145	0.3241	0.9145
								
K=5	0.2177	0.9145	0.2265	0.9145	0.2433	0.9145	0.3241	0.9145
								
K=6	0.2187	0.9145	0.2971	0.9145	0.1938	0.9145	0.1908	0.9145
								
K=7	0.2265	0.9145	0.1867	0.9221	0.1661	0.9258	0.1938	0.9145
								
K=8	0.1959	0.9145	0.2265	0.9145	0.1661	0.9258	0.1661	0.9258
								

Πίνακας 4

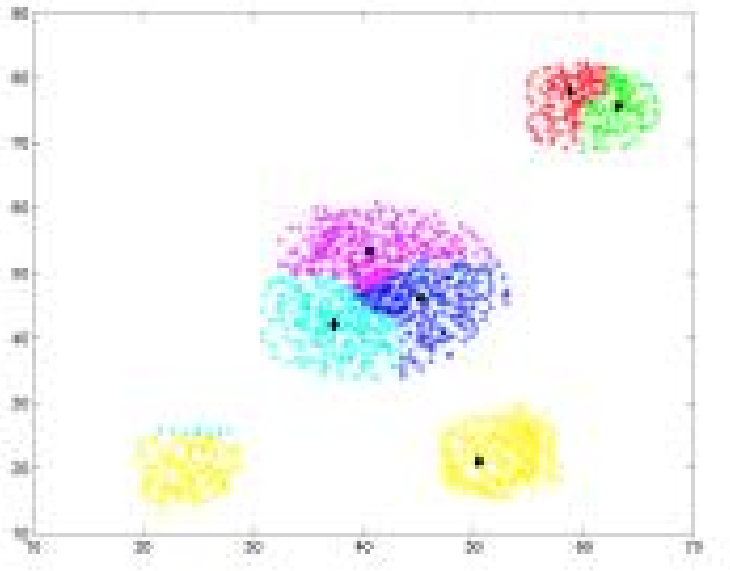
Πλήθος Ομάδων - number of clusters (1 , 2 , 3 , 4 , 5 , 6 , 7 , 8 , 9)

Μέθοδοι Ομαδοποίησης - clustering methods (Manhattan , Euclidean)

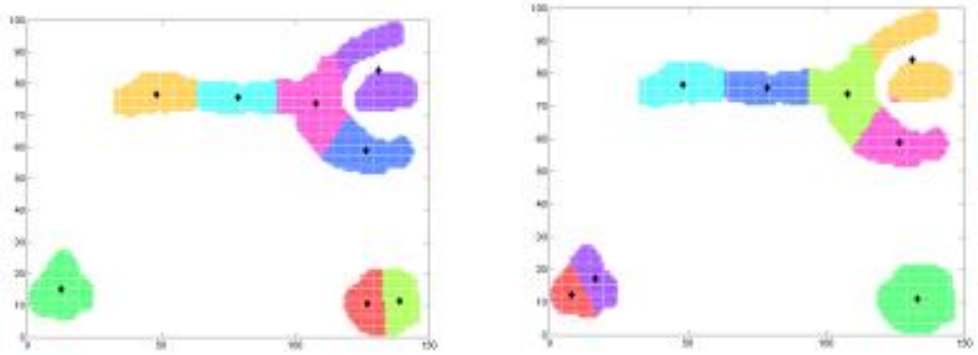
Μέθοδοι ορισμού κέντρου ομάδας (10% τυχαία στοιχεία και ορισμός κέντρου , τυχαίο κέντρο)

Μέτρα εγγύτητας - validation measures (Entropy , Purity)

Σχολιασμός αποτελεσμάτων



Το πρώτο παράδειγμα βρίσκεται στον πίνακα 3 και αφορά στην ομαδοποίηση με την μετρική Manhattan, λαμβάνοντας τυχαία αρχικά κέντρα και το σύνολο των ομάδων που ζητούνται να είναι 6. Στις 5 ομάδες, παρατηρούμε πώς η εντροπία παίρνει την τιμή 0 και το purity την τιμή 1. Η ομαδοποίηση θεωρείται επιτυχημένη. Με τις 6 ομάδες όμως, η εντροπία και το purity παίρνουν τις τιμές 0.3013 και 0.8994 αντίστοιχα. Αυτό συμβαίνει επειδή, λόγω των τυχαίων (αυθαίρετων) κέντρων, κάποιες ομάδες ενώνονται (φτιάχνεται μια ομάδα με στοιχεία με διαφορετικά labels) και κάποιες ομάδες χωρίζονται (φτιάχνονται περισσότερες από μια ομάδες με στοιχεία του ίδιου label). Στο γράφημα λοιπόν, από μια πρώτη ματιά φαίνεται πως υπάρχουν γαλάζια στοιχεία στην κίτρινη ομάδα. Αυτό από μόνο το όμως δεν δικαιολογεί την μεγάλη απόκλιση των δεικτών από την προηγούμενη ομαδοποίηση. Μπορεί να παρατηρηθεί πως και η κίτρινη ομάδα έχει δεδομένα που κανονικά άνηκαν σε δύο διαφορετικές κατηγορίες (ή πραγματικές ομάδες), τα οποία ομαδοποιούνται σύμφωνα με τον αλγόριθμο k-meansστην ίδια ομάδα. (Τα στοιχεία του κίτρινου cluster είναι προ-ομαδοποιημένα σε διαφορετικές ομάδες).



Το επόμενο παράδειγμα, βρίσκεται στον πίνακα 4 και αφορά στην ομαδοποίηση με την μετρική Euclidean, με τυχαία κέντρα και με δείγμα του 10% των στοιχείων ώστε να δημιουργηθούν κέντρα. Στις 8 ομάδες, παρατηρείται η χαμηλότερη τιμή που λαμβάνει η εντροπία και το purity. Γνωρίζοντας εκ των προτέρων τον αριθμό των «σωστών» ομάδων που είναι 4, αναμένονταν εκεί οι καλύτερες τιμές. Αυτό συμβαίνει, γιατί η εντροπία, όπως και το purity, δεν δείχνουν ενδιαφέρον για τον αριθμό των ομάδων και αν κάποια ομάδα χωριστεί σε 2 ή παραπάνω, μα για το αν κάποια ομάδα ομαδοποιηθεί με κάποια άλλη. Λόγω της ιδιαιτερότητας των στοιχείων που βρίσκονται στον πίνακα 4 και φαίνονται στα παραπάνω γραφήματα η μία ομάδα πάντα «μπερδεύεται» με στοιχεία άλλης. Αξίζει να αναφερθεί πως οι ομάδες μπορούν να εντοπιστούν «με το μάτι» σε όλες τις περιπτώσεις, λόγω της σύμπτυξης των στοιχείων, και της ύπαρξης κενού μεταξύ των «κανονικών» ομάδων.

Αυτό οφείλεται και στον αλγόριθμο ομαδοποίησης που χρησιμοποιείται. Ως γνωστόν, ο k-means ευνοεί τον εντοπισμό ομάδων οι οποίες έχουν σφαιρικό σχήμα και είναι καλά διαχωρισμένες μεταξύ τους, επομένως στο συγκεκριμένο σύνολο δεδομένων ακόμη και εάν ζητηθεί από τον αλγόριθμο k-means να εντοπίσει το πραγματικό πλήθος των ομάδων, αυτός θα αποτύχει.

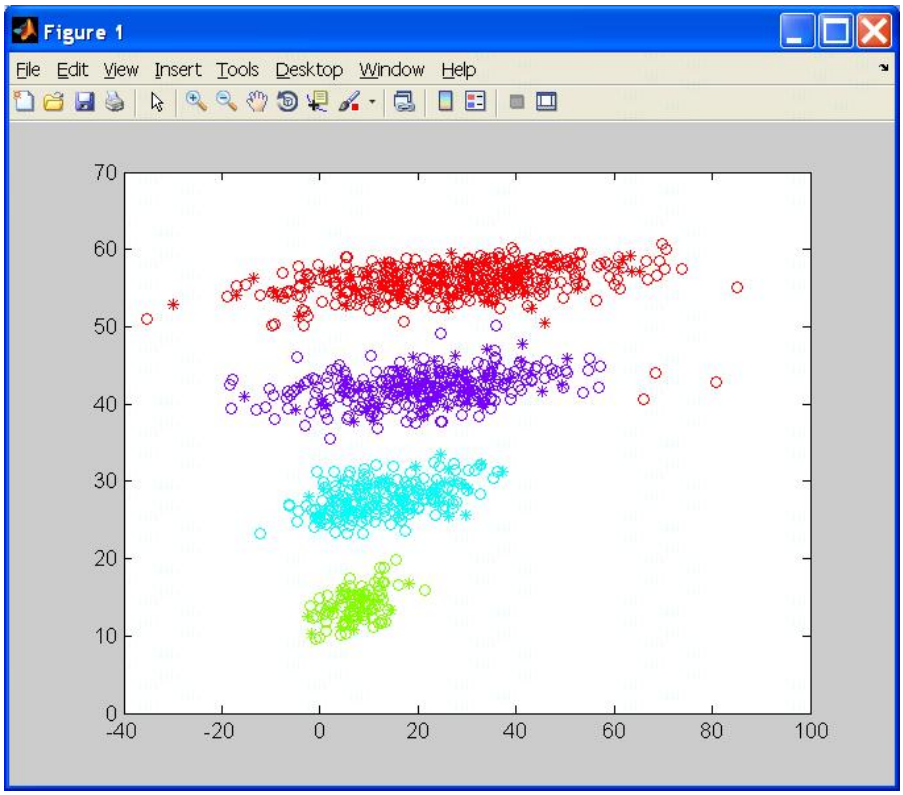
Ένα ακόμη παράδειγμα βρίσκεται στον πίνακα 2 και αφορά το σύνολο των περιπτώσεων (εξαιρουμένης της τελευταίας) της ομαδοποίησης με μετρική Manhattan και τον ορισμό των κέντρων τυχαίο. Και σε αυτή την περίπτωση, όπως και στην προηγούμενη, ο αριθμός των ομάδων ανέρχεται σε 4. Στην ομαδοποίηση των συγκεκριμένων στοιχείων στις άλλες περιπτώσεις, από τις τέσσερις ομάδες και μετά οι τιμές είναι οι αναμενόμενες, στο παράδειγμα που αναπτύσεται εδώ δεν συμβαίνει. Αυτό οφείλεται στο ότι η Manhattan εξετάζει διαφορετικά τις αποστάσεις των στοιχείων από τη Euclidean. Επίσης σε αυτή την περίπτωση της Manhattan τα κέντρα δημιουργούνται τυχαία ενώ εάν δημιουργούντουσαν έπειτα από εξέταση του 10% των στοιχείων του δείγματος, θα υπήρχαν καλύτερες τιμές (όπως φαίνεται στην πρώτη στήλη του πίνακα.)

7.3.2 Αποτελέσματα K-NN

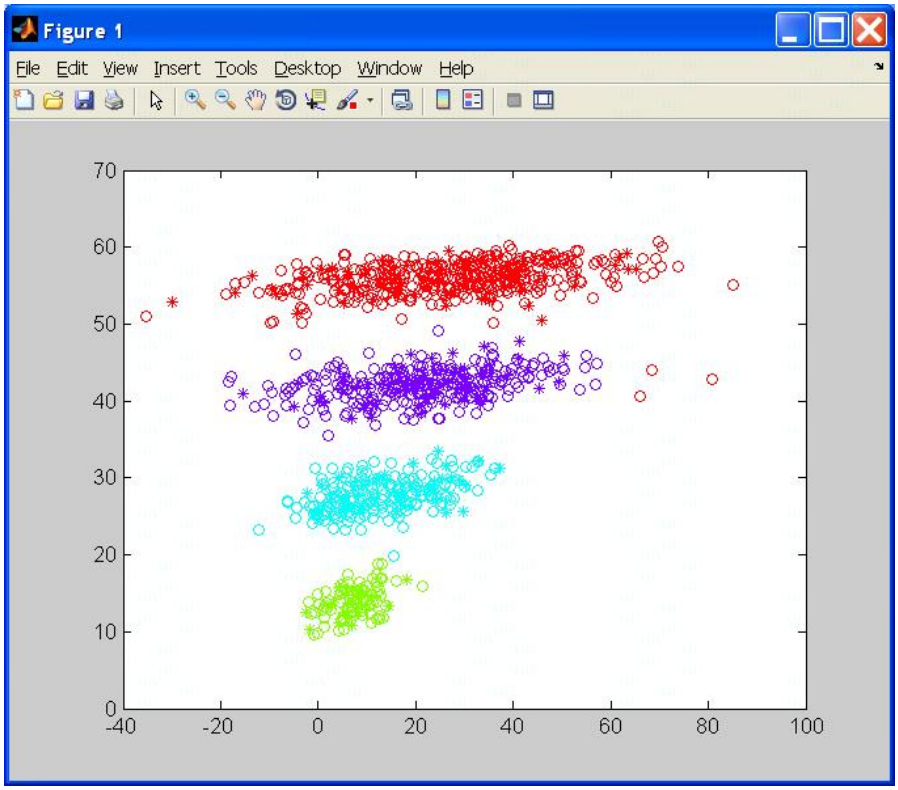
Το ακόλουθο σύνολο δεδομένων αποτελείται από 4 κλάσεις, τα τεχνητά σύνολα δεδομένων είναι 200 ενώ τα στοιχεία ελέγχου είναι 800. Οι κλάσεις έχουν ως εξής:
Κλάση 1: εκπαίδευσης: 20 ελέγχου: 80
Κλάση 2: εκπαίδευσης: 40 ελέγχου: 160
Κλάση 3: εκπαίδευσης: 60 ελέγχου: 240
Κλάση 4: εκπαίδευσης: 80 ελέγχου: 320
Τα στοιχεία δημιουργήθηκαν από τις κατανομές:

μ1=[7 14] S1=[5 .5; 0.8 2]
μ2=[14 28] S2=[10 .5; 1.6 2]
μ3=[21 42] S3=[15 .5; 2.4 2]
μ4=[28 56] S4=[20 .5; 3.2 2]
(ελέγχου)

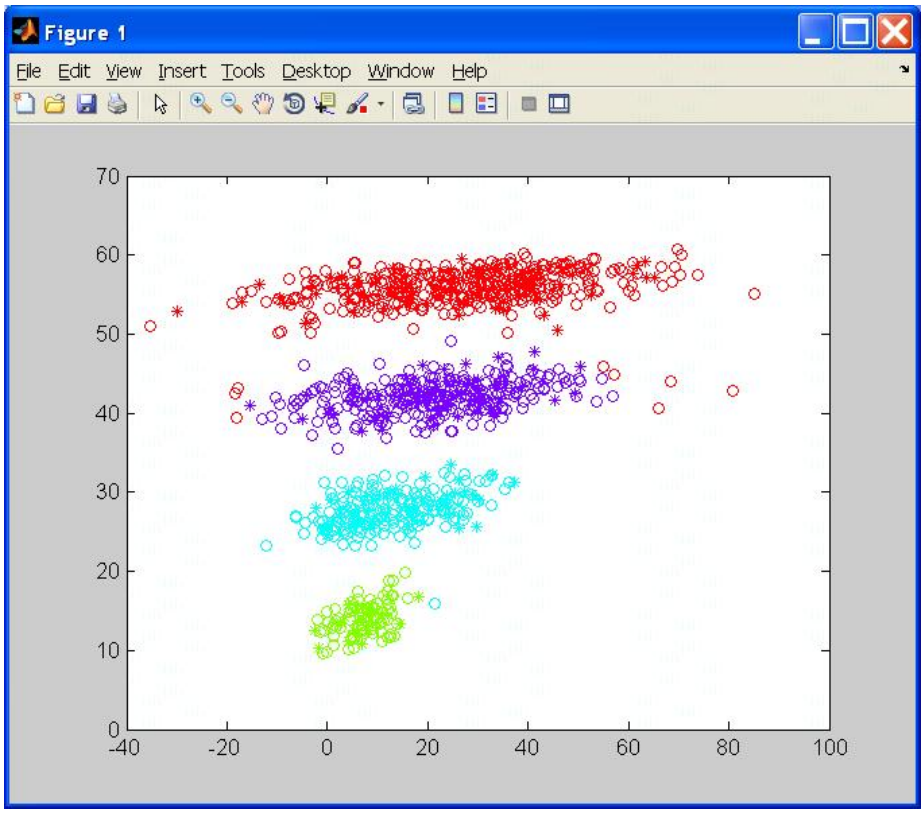
μ1=[7 14] S1=[5 .5; 0.8 2]
μ2=[14 28] S2=[10 .5; 1.6 2]
μ3=[21 42] S3=[15 .5; 2.4 2]
μ4=[28 56] S4=[20 .5; 3.2 2]
(εκπαίδευσης)



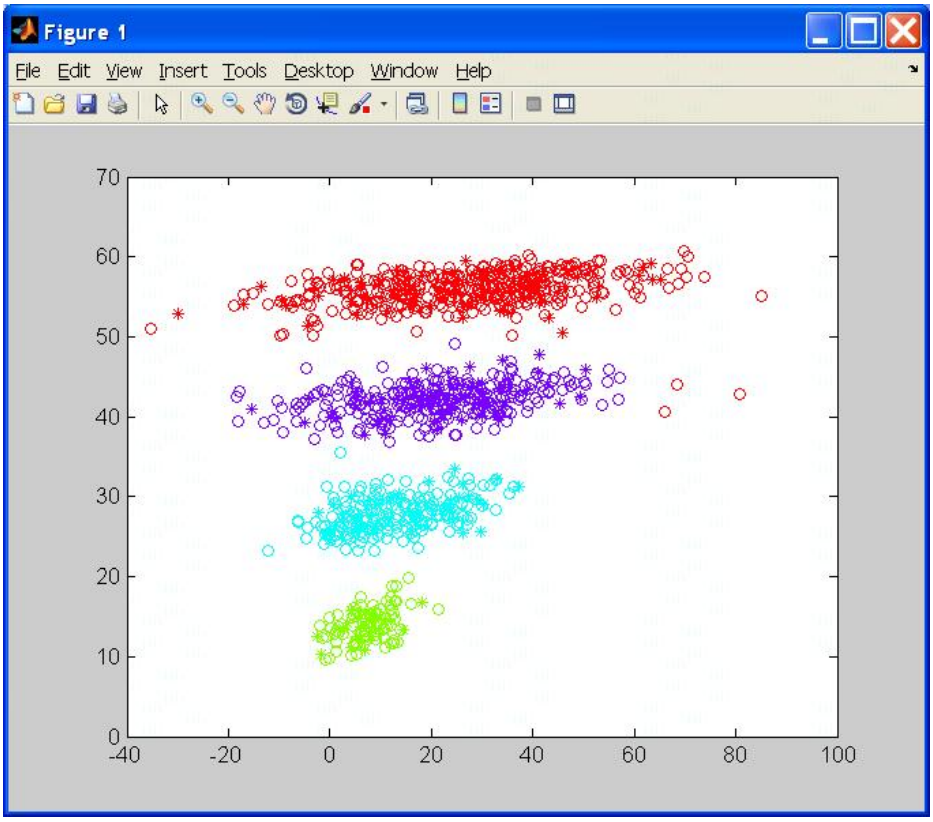
Euclidean k=5 Entropy:0.0515 Purity:0.9938



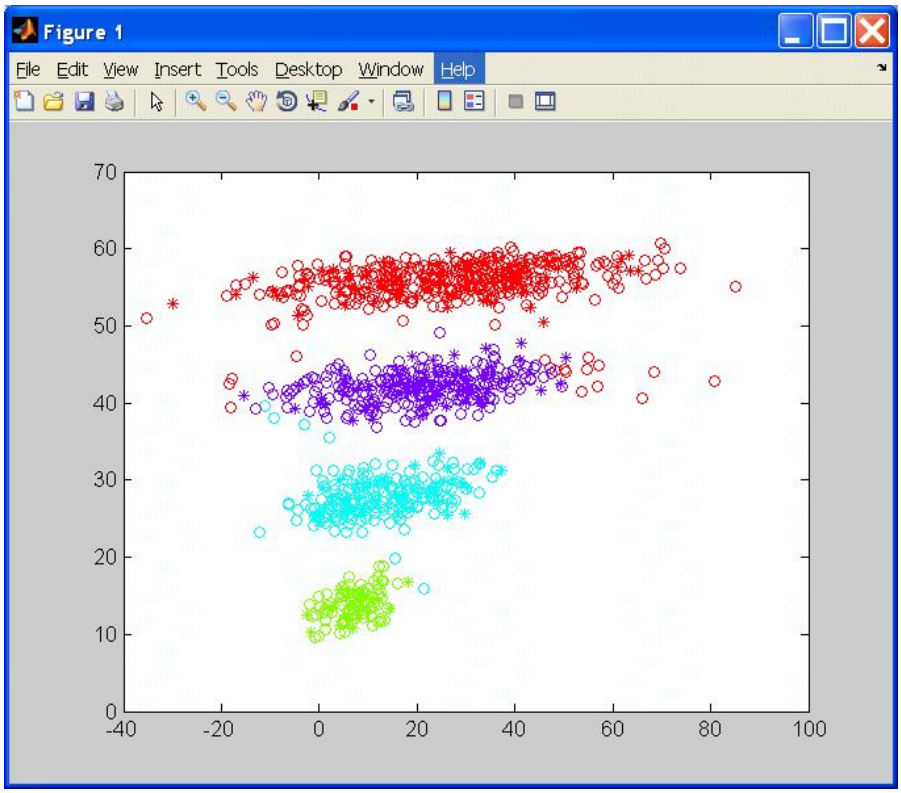
Manhattan k=5 Entropy: 0.0533 Purity:0.9938



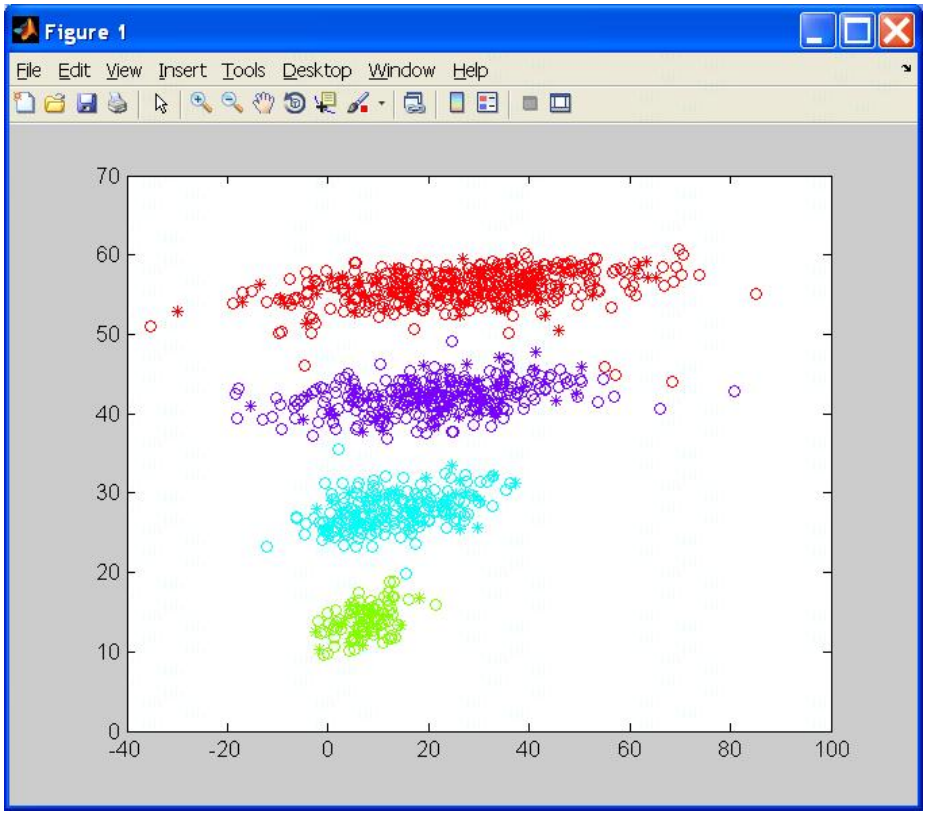
Euclidean k=10 Entropy: 0.0904 Purity: 0.9875



Manhattan k=10 Entropy: 0.0533 Purity: 0.9938



Euclidean k=20 Entropy: 0.1809 Purity: 0.9712



Manhattan k=20 Entropy: 0.0724 Purity: 0.9912

7.4 Πραγματικά σύνολα δεδομένων

Στον αλγόριθμο εισήχθησαν και πραγματικά σύνολα δεδομένων, οπότε και προέρχονται από αληθινά προβλήματα. Τα στοιχεία αφορούν στη Διαγνωστική Καρκίνου Στήθους Wisconsin (WDBC). Οι δημιουργοί ήταν οι:

- Dr. WilliamH. Wolberg, GeneralSurgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792 wolberg@eagle.surgery.wisc.edu
- W. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 street@cs.wisc.edu 608-262-6619
- Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 olvi@cs.wisc.edu

Τα αποτελέσματα έχουν ως εξής:
Γίνεται πρόβλεψη πεδίου. Δηλαδή ανάλογα με την πρόβλεψη – ταξινόμηση, πραγματοποιείται και η διάγνωση σε καλοήθης ή κακοήθης.
Τα σύνολα είναι γραμμικά διαχωρισμένα αφου χρησιμοποιηθούν και τα 30 εισαγόμενα χαρακτηριστικά γνωρίσματα. Η εκτιμώμενη ακρίβεια βρίσκεται στο 97.5% χρησιμοποιώντας επαναλαμβανόμενα 10-πλή διασταύρωση επικύρωσης. Ο ταξινομητής έχει πραγματοποιήσει σωστή διάγνωση για 176 διαδοχικά νέους ασθενείς μέχρι το Νοέμβριο του 1995.
Τα χαρακτηριστικά γνωρίσματα υπολογίζονται από μια ψηφιοποιημένη εικόνα μιας λεπτής βελόνας απορρόφησης (FNA) από ιστό στήθους. Περιγράφουν τα χαρακτηριστικά των πύρινων των κυττάρων που είναι παρόντα στην εικόνα. Μερικές εικόνες μπορούν να βρεθούν στη διεύθυνση http://www.cs.wisc.edu/~street/images/

Ο αριθμός των περιπτώσεων που περιλαμβάνονται ανέρχεται στις 569. Καθεμία από αυτές τις περιπτώσεις έχει 32 ιδιότητες. (ID, διάγνωση, 30 real-valued input features). Οι πληροφορίες ιδιοτήτων περιλαμβάνουν σαν πρώτη ιδιότητα τον αριθμό ταυτότητας (ID) της κάθε περίπτωσης. Σαν δεύτερη ιδιότητα περιλαμβάνεται η διάγνωση. Τέλος ακολουθούν οι 30 μετρήσεις (30 real-valued input features) για τις ιδιότητες 3-32.

Δέκα real-valued χαρακτηριστικά υπολογίζονται στο πύρινα κάθε κυττάρου.

a) εύρος (μέσος αποστάσεων από το κέντρο σε σημεία της περιμέτρου)

b) υφή (κανονική απόκλιση των μεταβλητών γκρίζας κλίμακας (gray-scale values))

c) περίμετρος

d) περιοχή

e) ομαλότητα (τοπική διαφοροποίηση σε εύρος αποστάσεων)

f) πυκνότητα

g) κοίλα (δριμύτητα της κοιλότητας των τμημάτων περιγράμματος)

h) κοίλα σημεία (αριθμός κοίλων τμημάτων των κοίλων)

i) συμμετρία

j) τμηματική (fractal) διάσταση("coastline approximation" – 1)

Οι μέσοι, το σφάλμα, και ο ‘’χειρότερος’’ ή μεγαλύτερος (μέσος των τριών μεγαλύτερων μεταβλητών) αυτών των χαρακτηριστικών υπολογίστηκαν σε κάθε εικόνα, αποτελώντας 30 χαρακτηριστικά. Για παράδειγμα το πεδίο 3 είναι η μέση ακτίνα, πεδίο 13 είναι ακτίνα SE, πεδίο 23 είναι η χειρότερη ακτίνα.

Όλες οι μεταβλητές καταγράφονται με 4 σημαντικά ψηφία.

Κατανομή κλάσης: 357 καλοήθης, 212 κακοήθης.

[54]

Ανάλυση αποτελεσμάτων

K-NN

Κοντινότεροι γείτονες	Manhattan		Euclidean	
	Entropy	Purity	Entropy	Purity
K=2	0.0000	1.0000	0.1450	0.9750
K=3	0.2417	0.9500	0.2417	0.9500
K=4	0.1450	0.9750	0.1450	0.9750
K=5	0.3212	0.9250	0.2417	0.9500
K=6	0.3212	0.9250	0.2417	0.9500
K=7	0.3900	0.9000	0.2417	0.9500
K=8	0.3212	0.9250	0.2417	0.9500
K=9	0.3900	0.9000	0.3212	0.9250

Συμπεράσματα

Παρότι η Εξόρυξη Δεδομένων μετράει μόλις λίγες δεκαετίες ύπαρξης, γνωρίζει μεγάλη άνθηση καθώς η χρήση της δεν περιορίζεται μα καλύπτει πολλούς διαφορετικούς τομείς (μαθηματικά - στατιστική, βιολογία - ιατρική, οικονομικά - marketing κ.ο.κ). Όσο λοιπόν εξελίσσεται η τεχνολογία των υπολογιστών, τόσο θα αυξάνει και η ευρύτερη χρήση της.

Η εξόρυξη δεδομένων έχει καταφέρει να επιλύσει αρκετά προβλήματα μα έχει να επιλύσει ακόμη περισσότερα καθώς υπάρχουν διάφορες προκλήσεις που χρειάζονται περισσότερη έρευνα. Κάποιες από αυτές τις προκλήσεις, είναι πως τα δεδομένα πλέον δεν είναι μόνο κείμενο και αριθμοί, αλλά μπορεί να είναι φωτογραφίες ή εικόνες, δεδομένα από χάρτες και άλλα, με αποτέλεσμα την μη ύπαρξη κατάλληλων συστημάτων εξόρυξης δεδομένων. Επιπροσθέτως με την όλο και πιο διαδεδομένη χρήση των υπολογιστών και του διαδικτύου, περισσότερες και πολυπλοκότερες βάσεις δεδομένων σχηματίζονται. Έτσι η απαίτηση είναι να δημιουργηθούν αλγόριθμοι εξόρυξης δεδομένων οι οποίοι να είναι και αποδοτικοί αλλά και να δίνουν απαντήσεις σε πραγματικό χρόνο.

Παρόλο που βοηθάει στην λήψη αποφάσεων καθώς μετά την επεξεργασία των δεδομένων είμαστε σε θέση να πάρουμε την καλύτερη δυνατή απόφαση (στο παράδειγμα της ταξινόμησης αν ο ασθενής θα ξεκινήσει χημειοθεραπείες, στο καλάθι αγοράς την τοποθέτηση των προϊόντων στα ράφια ή / και την διαφήμιση κάποιου προϊόντος ώστε να αυξηθούν και οι πωλήσεις κάποιου άλλου), ωστόσο δεν γίνεται να υπάρξει ένας ενιαίος αλγόριθμος για να ληφθούν – αξιολογηθούν διαφορετικές αποφάσεις. Χαρακτηριστικό παράδειγμα, ο αλγόριθμος ομαδοποίησης (k-means) που χρησιμοποιείται στην παρούσα πτυχιακή. Ομαδοποιεί στοιχεία ομάδων των οποίων οι ομάδες είναι σφαιρικές και σε απόσταση όπως στα dset1 δεδομένα (πίνακας 3), αλλά αποτυγχάνει να ομαδοποιήσει σωστά τις ομάδες για τα στοιχεία dset2 (πίνακας 4). Αν παρθεί σωστά η απόφαση για την επιλογή του καταλληλότερου αλγόριθμου εξόρυξης δεδομένων, τόσο καλύτερη θα είναι και η απεικόνιση της εξαγόμενης γνώσης, τόσο πιο επιτυχημένα και τα συμπεράσματα που θα προκύψουν.

Βιβλιογραφία

[1] W. Frawley and G. Piatetsky-Shapiro and C. Matheus.(1992): "Knowledge Discovery in Databases: An Overview".

[2] Kantardzic, Mehmed Data Mining(2003): "Concepts, Models, Methods, and Algorithms".

[3] Frawley William. F (1992): "Knowledge Discovery in Databases, An Overview".

[4] Fayyad (1996): "Advances in Knowledge Discovery and Data Mining". Microsoft Research,One Microsoft Way Redmond, WA 98052, USA.

[5] Ειρήνη Ντούτση (2003) :"Εξόρυξη γνώσης από ειδησεογραφικά δεδομένα και συσχετισμός με πραγματικά γεγονότα". Μεταπτυχιακή Εργασία.

[6] Ellen Monk, Bret Wagner (2006):"Concepts in Enterprise Resource Planning".Second Edition.

[7] Michigan State University."Genetic Algorithms Research and Applications Group (GARAGe)".

[8] Haughton(2003): "Review of Data Mining Software Packages".

[9] M.S. Chen, J. Han, and P.S. Yu. (1996): "Data Mining: An overview from Database Perspective, IEEE Transactions on Knowledge and Data Engineering", vol. 8, no. 6.

[10] Chaudhury, Abijit, Jean-Pierre Kuilboer (2002): "E-Business and e-Commerce Infrastructure".

[11] Cole, K., Fischer, O. and Saltzman, P. (1997):"Just-in-Time Knowledge Delivery".

[12] Jeffrey W. Seifert, (2007): “CRS Report For Congress Data mining and homeland security”.

[13] M.S. Chen, J. Han, and P.S. Yu (1996): "Data Mining: An overview from Database Perspective. Transactions on Knowledge and Data Engineering".

[14] EverittB.S. (1974): "Cluster Analysis". Journal of the American Statistical Association

[15] Slattery S, Craven M.Learning (1998): "To Exploit Document Relationships and Structure: The case for Relational Learning on the web, Working notes of learning from text and the web, Conf. on Automated Learning and Discovery". CONLAND.

[16] Seber G. A. F. (1984):"Multivariate Observations".

[17] Feldman R, Fresco M, Hirsh H. "Knowledge Management: A text Mining Approach, In proc. Of the 2nd Int. Conf. on practical Aspects of Knowledge Management".

[18] Dillon R. William, Goldstein Matthew (1984): "Multivariate Analysis Methods and Applications".

[19] R. Agrawal, T. Imielinski, A. Swami (1993):"Mining Association Rules Between Sets of Items in Large Databases".

[20] Oracle® Data Mining Concepts (2005-2008).

[21] Lıqiang Geng, Howard J. Hamilton. "Interestingness Measures for Data Mining: A Survey". University of Regina, Saskatchewan, Canada.

[22] Tan, Pang-Ning, Michael, Steinbach, Kumar, Vipin (2005):"Association Analysis: Basic Concepts and Algorithms".

[23] Maria-Luiza Antonie, Osmar R. Zaıane (2005):"Mining Positive and Negative Association Rules: An Approach for Confined Rules".

[24] Tan, P. Kumar, V. (2000):"Interestingness measures for association patterns: A perspective. In: Proc. of Workshop on Post processing in Machine Learning and Data Mining".

[25] Cohen, J. (1988):"Statistical power analysis for the behavioral sciences. Lawrence Erlbaum, New Jersey" (2nd edition).

[26] Savasere, A.Omiecinski, E., Navathe, S. (1998): "Mining for strong negative associations in a large database of customer transactions".

[27] Γουρδούλης Ιωάννης Πρόδρομος (2009) :"Αλγόριθμοι Εξόρυξης δεδομένων για χειρισμό πολλαπλών υποστηρίξεων και αρνητικών συσχετίσεων".

[28] http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf

[29] http://www.icaen.uiowa.edu/~comp/Public/Apriori.pdf

[30]http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput499/slides/Lect10/sld044.htm

[31] T. Mitchel (1997): "An Overview of Machine Learning". McGraw-Hill.

[32]J. Friedman (1994):"Flexible Metric Nearest Neighbor Classification, Technical Report". Department of Statistics, Stanford University.

[33] T. Cover, and P. Hart (1967): "Nearest Neighbor Pattern Classification, in IEEE Transactions on Information Theory".

[34] T. Joachims (1998): "Text Categorization with Support Vector Machines". In Proceedings of European Conference on Machine Learning.

[35] Osmar R. Zaïane (1999): "Principles of Knowledge Discovery in Databases".

[36] http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html

[37] Achtert, E.; Böhm, C.; Kröger, P.; Zimek, A. (2006): "Mining Hierarchies of Correlation Clusters".Proc. 18th International Conference on Scientific and Statistical Database Management.

[38] M. A. T. Figueiredoand A. K. Jain (2002): "Unsupervised Learning of Finite Mixture Models in IEEE Transactions on Pattern Analysis and Machine Intelligence".

[39] Bezdek, James C. (1981):"Pattern Recognition with Fuzzy Objective Function Algorithms."

[40] Walter, Scott (1999): "Minkowski, Mathematicians, and the Mathematical Theory of Relativity".

[41] http://en.wikipedia.org/wiki/Information_security

[42] T. Cover, P. Hart. (1967):"Nearest neighbor pattern classification". InIEEE Trans Inform Theory.

[43] Charles Elkan (2011): "Nearest Neighbor Classification".

[44] Sokal R. and Michener C. (1985): "A statistical method for evaluating systematic relationships", University of Kansas Scientific Bulletin, Vol 38 pp. 1409 - 1438.

[45] Jain A.K., Dubes R.C. (1988): "Algorithms for clustering data". Prentice Hall.

[46] Theodoridis Sergios, Koutroumbas Kostantinos (1999): "Pattern recognition". Academic Press.

[47] Dillon R. William, Goldstein Matthew (1984):"Multivariate Analysis Methods and Applications". John Wiley & Sons, Inc.

[48] C. E. Shannon (1948): "A mathematical theory of communication". Bell System Technical Journal. 27, pp.: 379-423, 623-656.

[49] http://en.wikipedia.org/wiki/Cluster_analysis

[50] http://www.usinessweek.com/globalbiz/content/dec2008/gb20081229_497909.htm?chan=globalbiz_europe+index+page_top+stories

[51] Αριστείδης Μελετίου (2005): "Ποιοτικοί Δείκτες Υπηρεσιών Βιβλιοθηκών και Διαχείριση Πόρων: Μεθοδολογίες Ανάλυσης και στρατηγικός σχεδιασμός". Πολυτεχνείο Κρήτης, Πολυτεχνειούπολη Χανιά

[52] Gerasimos S. Antzoulatos and Michael N. Vrahatis: "α-Clusterable Sets", in Lecture Notes in Artificial Intelligent 6911, pp 108-123, 2011.

[53] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis: " On Clustering Validation Techniques", Journal of Intelligent Information Systems, 17:2/3, 107–145, 2001

[54] http://archive.ics.uci.edu/ml/

[55] Αντζουλάτος Γεράσιμος (2002), “Εφαρμογές Αλγορίθμων και Έλεγχοι Αξιοπιστίας Ομαδοποίησης στην Αναγνώριση Προτύπων και στον Καθαρισμό Δεδομένων”, Διπλωματική Εργασία, ΜΔΕ «Μαθηματικά των Υπολογιστών και των Αποφάσεων», Πανεπιστήμιο Πατρών

[56] MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

Παραρτήματα

Ομαδοποίηση με χρήση του αλγόριθμου k-means

Ο παρακάτω αλγόριθμος παράγει το γραφικό περιβάλλον που παρουσιάστηκε παραπάνω και καλεί τις αντίστοιχες συναρτήσεις (datagen αν χρειάζεται δημιουργία τυχαίων δεδομένων). Στη συνέχεια ομαδοποιεί τα δεδομένα με την ήδη υπάρχουσα συνάρτηση του Matlab. Τέλος, αφού καλέσει τις αντίστοιχες συναρτήσεις, υπολογίζει το μέτρο της εντροπίας ή της καθαρότητας, για την αξιολόγηση της ομαδοποίησης.

```
%K_means 1.0.0
%made by Alex Zafeiropoulos & Chris Vozikis
%Special thanks to George Gounaris, Thomas Papastergiou & Gerasimos Antzoulatos!

function [gIdx, c] = main_kmeans

% K_MEANS    k-means clustring
%   IDX = k_means(X, K) partititions the N x P data matrix X into K
%   clusters through a fully vectorized algorithm, where N is the number of
%   data points and P is the number of dimensions (variables). The
%   partition minimizes the sum of point-to-cluster-centroid Euclidean
%   distances of all clusters. The returned N x 1 vector IDX contains the
%   cluster indices of each point.
%
%   To N Einai to pli8os twn upo eksetasi stoixeiwn
%   To omades einai o ari8mos twn omadopoiisewn pou 8eloume na kanoume
%   diladi se poses omades 8eloume na to xwrisoume...
%   IDX = k_means(X, C) works with the initial centroids, C, (K x P).
%
%   [IDX, C] = k_means(X, K) also returns the K cluster centroid locations
%   in the K x P matrix, C.

epilogesInit = {'Ελληνικά','English'};
[lang,~] = listdlg('PromptString','Επιλογή γλώσσας, selection language :',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
if lang == 1
   prompt = {'Πλήθος Ομάδων :'};
dlg_title = 'Πλήθος Ομάδων :';
```

```
num_lines = 1;
   def = {' '};
   options.Resize='on';
   options.WindowStyle='normal';
   Omad = inputdlg(prompt,dlg_title,num_lines,def,options);
   OM = str2num(Omad{1,1});
   md=0;
   choiceA=0 ;
   choiceB=0 ;
   epilogesInit = {'Squared Euclidian','Manhatan'};
   [choiceA,~] = listdlg('PromptString','Επιλογή Μετρικής :',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
if choiceA == 1
      val1='sqEuclidean';
elseif choiceA == 2
      val1='cityblock';
end
   epilogesInit = {'Τυχαία Κέντρα','Τυχαίο δείγμα σημείων 10% & ορισμός κέντρων'};
   [choiceB,~] = listdlg('PromptString','Ορισμός Κέντρων :',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
if choiceB == 1
      val2='sample';
elseif choiceB == 2
      val2='cluster';
end
   epilogesInit = {'Δημιουργία τυχαίων τιμών','Εισαγωγή τιμών από αρχείο'};
[md,~] = listdlg('PromptString','Μέθοδοι εισαγωγής τιμών :',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
if md==1
      [X,id,nc] = datagen(lang, OM);
      [rx,cx]=size(X);
```

```
        save('dataset.txt', 'X',  '-ascii' , '-double', '-tabs');
elseif md==2
        prompt = {'Δώσε το μονοπάτι (Path) :'};
dlg_title = 'Δώσε το μονοπάτι (Path) :';
num_lines = 1;
        def = {' '};
        options.Resize='on';
        options.WindowStyle='normal';
        path = inputdlg(prompt,dlg_title,num_lines,def,options);
        X_all = load(path{1});
        X = X_all(:,1:end-1);
        id = X_all(:,end);
        id = id+1;
end
    epilogesInit = {'Entropy','Purity'};
    [eval,~] = listdlg('PromptString','evaluation measures:',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);

    CLM = colormap(hsv(256));
    step = floor(256/OM);
    m=0;
    h=figure(1);
for i=1:OM
        plot( X(id(:)==i,1), X( id(:)==i,2 ), 'o', 'Color', CLM( m+step, :)  )
        hold on
        m = m + step;
end

elseif lang == 2
    prompt = {'Number of Clusters :'};
    dlg_title = 'Number of Clusters :';
    num_lines = 1;
    def = {''};
    options.Resize='on';
    options.WindowStyle='normal';
    Omad = inputdlg(prompt,dlg_title,num_lines,def,options);
    OM = str2num(Omad{1,1});
```

```
    md=0;
    choiceA=0 ;
    choiceB=0 ;
    epilogesInit = {'squared Euclidean','Manhatan'};
    [choiceA,~] = listdlg('PromptString','Metrical selection:',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
if choiceA == 1
        val1='sqEuclidean';
elseif choiceA == 2
        val1='cityblock';
end
    epilogesInit = {'random centroids','random sample 10% and then centroids'};
    [choiceB,~] = listdlg('PromptString','centroids :',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);

if choiceB == 1
        val2='sample';
elseif choiceB == 2
        val2='cluster';
end
    epilogesInit = {'random values','values from file'};
    [md,~] = listdlg('PromptString','method of input values :',...
'ListSize',[350 150],...
'SelectionMode','1, 2,',...
'ListString',epilogesInit,...
'InitialValue',1);
if md==1
        [X,id,nc] = datagen(lang,OM);
        [rx,cx]=size(X);
        save('dataset.txt', 'X',  '-ascii' , '-double', '-tabs');
elseif md==2
        prompt = {'Write path :'};
        dlg_title = 'Write path :';
```

```
        num_lines = 1;
        def = {' '};
        options.Resize='on';
        options.WindowStyle='normal';
        path = inputdlg(prompt,dlg_title,num_lines,def,options);
        X_all = load(path{1});

        X = X_all(:,1:end-1);
        id = X_all(:,end);
        id = id+1;
end
    epilogesInit = {'Entropy','Purity'};
    [eval,~] = listdlg('PromptString','evaluation measures:',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
    CLM = colormap(hsv(256));
    step = floor(256/OM);
    m=0;
    h=figure(1);
for i=1:OM
        plot( X(id(:)==i,1), X( id(:)==i,2 ), 'o', 'Color', CLM( m+step, :)  )
        hold on
        m = m + step;
end
end

% tic
[cidx, ctrs] = kmeans( X, OM, 'Distance', val1, 'Start', val2);
% toc

CLM = colormap(hsv(256));
step = floor(256/OM);

m=0;
for i=1:OM
    plot( X(cidx==i,1), X( cidx==i,2 ), '*', 'Color', CLM( m+step, :)  )
    hold on
    plot( ctrs(i,1), ctrs(i,2), 'kd',  'MarkerEdgeColor','k',  'MarkerFaceColor','k', 'MarkerSize',7)
```

```
    hold on
    m = m + step;
end

name = [ 'fig_clusters_' num2str(OM) '_' val1 '_' val2 '.png'];
saveas(h,name,'png')


[entropy total_clustered_points accuracy] = estimate_entropy_accuracy(id,cidx);

if eval==1
    disp ('entropy is: ');
    disp (entropy)
elseif eval==2
        disp ('Purity is: ');
        disp (accuracy)
end
end


XC = [X cidx];
save('data_cl.txt', 'XC', '-ascii', '-double')
end
```

Η συνάρτηση **datagen** δημιουργεί τυχαίες τιμές για τόσες ομάδες όσες ο χρήστης ζητήσει, με το αντίστοιχο πλήθος σημείων στην κάθε ομάδα.

```
function [X,id,nc] = datagen(lang,OM)

mu = [1 2];
X = []; id =[];

if lang==1
for i =1:OM

    prompt = {'Δώσε πλήθος στοιχείων i - οστής ομάδας :'};
    dlg_title =( 'Δώσε πλήθος στοιχείων i - οστής ομάδας :');
    num_lines = 1;
    def = {' '};
    options.Resize='on';
    options.WindowStyle='normal';
    nco = inputdlg(prompt,dlg_title,num_lines,def,options);
    nc(i) = str2num(nco{1,1});
    xi = rand(nc(i),2 );
    mu = 7*i*[1 2];
    Sigma = [5*i .5; 0.8*i 2];
    xi = repmat(mu,nc(i),1) + randn(nc(i),2)*Sigma;
    X = [X ; xi];

    id = [id ; i*ones(nc(i),1)];

end
elseif lang==2
for i =1:OM
    prompt = {'number of elements for i - th cluster :'};
    dlg_title = 'number of elements for i - th cluster :';
    num_lines = 1;
    def = {' '};
    options.Resize='on';
    options.WindowStyle='normal';
    nco = inputdlg(prompt,dlg_title,num_lines,def,options);
    nc(i) = str2num(nco{1,1});
    xi = rand(nc(i),2 );
    mu = 7*i*[1 2];
    Sigma = [5*i .5; 0.8*i 2];
```

```
        xi = repmat(mu,nc(i),1) + randn(nc(i),2)*Sigma;
        X = [X ; xi];

        id = [id ; i*ones(nc(i),1)];
end
end
    Y = [X id];
   save('data.txt', 'Y', '-ascii', '-double')
end
```

```
confidence
%function confusion(datafile,perffile,mode,params_no)
% idx: labels of the class that belong each data point (real
% classification)(id - k means, id - knn)
% perf: labels that are produced after the kmeans clustering (cidx -
% k-means, idZ -knn)
%
function [CFM] = conf(id,cidx)

class_no = max(id);

[rp cp] = size(cidx);

maxcidx = max(cidx);
confusion_matrix = zeros(maxcidx,class_no);

for j=1:cp
if (cidx(j) >= 0) % not an outlier
      confusion_matrix(cidx(j),id(j)) = confusion_matrix(cidx(j),id(j)) + 1;
end% if
end% for j

CFM = confusion_matrix;

tmp = sum(confusion_matrix);
identified_points = sum(tmp);
total_points = cp;
```

```
υπολογισμος εντροπιας-καθαροτητας
function [entropy total_clustered_points accuracy] = estimate_entropy_accuracy(id,cidx)

% Transpose the arrays to become a line-vector
id = id';
cidx = cidx';

CFM = conf(id,cidx);

[clusters classes] = size(CFM);

if (clusters == 0 || classes == 0)
   entropy = NaN;
   total_clustered_points = -5;
   accuracy = NaN;
return
end

entropy_vector = zeros(1,clusters);

entropy_sum = 0;  accuracy = 0;

[r num_points] = size(cidx);

for i=1:clusters
   sizes(i) = sum(CFM(i,:));
   accuracy = accuracy + max(CFM(i,:));

for j=1:classes
if (CFM(i,j) ~= 0)
        p = (CFM(i,j)./sizes(i));
        entropy_sum = entropy_sum - (p) * log2(p);
end
end
   entropy_vector(i) = entropy_vector(i) + entropy_sum;

   entropy_sum = 0;
end

accuracy = accuracy / num_points;
```

```
total_clustered_points = sum(sizes);


entropy = 0;
for i=1:1:clusters
   entropy = entropy + (sizes(i)./total_clustered_points)*entropy_vector(i)';
end
```

Ταξινόμηση με χρήση του k-nn

Ο παρακάτω αλγόριθμος παράγει το γραφικό περιβάλλον που παρουσιάστηκε παραπάνω και καλεί τις αντίστοιχες συναρτήσεις (datagen αν χρειάζεται δημιουργία τυχαίων δεδομένων). Στη συνέχεια ταξινομεί τα δεδομένα με την ήδη υπάρχουσα συνάρτηση του Matlab. Τέλος, αφού καλέσει τις αντίστοιχες συναρτήσεις, υπολογίζει το μέτρο της εντροπίας ή της καθαρότητας, για την αξιολόγηση της ταξινόμησης.

```
function main_knn_classifier

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% FUNCTION
%   [idZ (z)]=k_nn_classifier(X (Z),id,k, Z) (X)
% k-nearest neighbor classifier for c classes. The
% classification is based on a reference data set, X (Z), for which the class
% labels of its vectors are known.
%
% INPUT ARGUMENTS:
%  X (Z):  lxN1  matrix, whose i-th column corresponds to the
%       i-th reference vector.
%   id:  N1-dimensional vector whose i-th  component contains the
%       label of the class where the i-th reference vector belongs.
%   md:  the number of nearest neighbors of the reference set that are
%       taken into account for the classification of a given vector.
%   Z (X):  lxN matrix whose columns are the data vectors to be classfied.
%
% OUTPUT ARGUMENTS:
%   idZ:  N-dimensional vector whose i-th component contains the label
%       of the class where the i-th vector of Z (X) is assigned.
%
% Zafeiropoulos A. Vozikis X.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
md=0 ;
choiceA=0 ;
epilogesInit = {'Ελληνικά','English'};
[lang,~] = listdlg('PromptString','Επιλογή γλώσσας',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
```

```
'InitialValue',1);
if lang==1
        prompt = {'Πλήθος Κλάσεων:'};
dlg_title = 'Σύνολο Εκπαίδευσης: Πλήθος Κλάσεων:';
num_lines = 1;
        def = {''};
        Omad = inputdlg(prompt,dlg_title,num_lines,def);
OM = str2num(Omad{1,1});
        prompt = {'Κοντινότεροι γείτονες προς σύγκριση:'};
        dlg_title = 'Κοντινότεροι γείτονες προς σύγκριση:';
num_lines = 1;
        def = {' '};
        kont = inputdlg(prompt,dlg_title,num_lines,def);
        k = str2num(kont{1,1});
        epilogesInit = {'Euclidian','Manhatan'};
        [choiceA,~] = listdlg('PromptString','Επιλογή Μετρικής',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
        epilogesInit = {'Τυχαίες τιμές','Εισαγωγή τιμών απο αρχείο'};
[md,~] = listdlg('PromptString','Μέθοδοι εισαγωγής τιμών',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
if md==1
            [X,id,nc] = datagen(lang, OM);
            [rx,cx]=size(X);

        CLM = colormap(hsv(256));
        step = round(256/OM);
        m=0;
for i=1:OM
        plot( X(id(:)==i,1), X( id(:)==i,2 ), '*', 'Color', CLM( m+step, :)  )
        hold on
        m = m + step;
end
      [Z,idZ,ncz] = datagen1(lang,OM);
%       plot(Z(:,1), Z(:,2), 'k.')
```

```
elseif md==2
prompt = {'Δώσε το μονοπάτι για το σύνολο εκπαίδευσης:'};
            dlg_title = 'Σύνολο Εκπαίδευσης';
            num_lines = 1;
def = {' '};
            path = inputdlg(prompt,dlg_title,num_lines,def);
            X_all = load(path{1});
            X = X_all(:,1:end-1);
id = X_all(:,end);
            prompt = {'Δώσε το μονοπάτι για το σύνολο ελέγχου:'};
dlg_title = 'ΣύνολοΕλέγχου';
            num_lines = 1;
            def = {' '};
            path = inputdlg(prompt,dlg_title,num_lines,def);
            Z_all = load(path{1});
            Z = Z_all(:,1:end-1);
            idZ = Z_all(:,end);



end
elseif lang==2
        prompt = {'Number of Classes'};
        dlg_title = 'Number of Classes';
        num_lines = 1;
        def = {''};
        Omad = inputdlg(prompt,dlg_title,num_lines,def);
        OM = str2num(Omad{1,1});
        prompt = {'Define the number of nearest neighbors (k):'};
        dlg_title = 'Define the number of nearest neighbors (k):';
        num_lines = 1;
        def = {' '};
        kont = inputdlg(prompt,dlg_title,num_lines,def);
        k = str2num(kont{1,1});
        epilogesInit = {'Euclidian','Manhatan'};
        [choiceA,~] = listdlg('PromptString','Metrical Selection:',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
```

```
        epilogesInit = {'random values','values from file'};
        [md,~] = listdlg('PromptString','method of input values',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
if md==1
            [X,id,nc] = datagen(lang, OM);
            [rx,cx]=size(X);

        CLM = colormap(hsv(256));
        step = round(256/OM);
        m=0;
for i=1:OM
        plot( X(id(:)==i,1), X( id(:)==i,2 ), '*', 'Color', CLM( m+step, :)  )
        hold on
        m = m + step;
end
      [Z,idZ,ncz] = datagen1(lang,OM);
%       plot(Z(:,1), Z(:,2), 'k.')

elseif md==2
            prompt = {'Write path for train:'};
            dlg_title = 'Write path for train:';
            num_lines = 1;
            def = {' '};
            path = inputdlg(prompt,dlg_title,num_lines,def);
            X_all = load(path{1});
            X = X_all(:,1:end-1);
            id = X_all(:,end);
            prompt = {'Write path for test:'};
            dlg_title = 'Write path for test:';
            num_lines = 1;
            def = {' '};
            path = inputdlg(prompt,dlg_title,num_lines,def);
            Z_all = load(path{1});
            Z = Z_all(:,1:end-1);
            idZ = Z_all(:,end);
```

```
end

end
        epilogesInit = {'Entropy','Purity'};
        [eval,~] = listdlg('PromptString','evaluation measures:',...
'ListSize',[350 150],...
'SelectionMode','1, 2',...
'ListString',epilogesInit,...
'InitialValue',1);
       [knn_idZ]=knn_classifier(X,id,k,OM,Z,choiceA);
       [entropy total_clustered_points accuracy] = estimate_entropy_accuracy_knn(idZ,knn_idZ);
if eval==1
           disp('Entropy: ')
           disp (entropy)
elseif eval==2
              disp('Purity: ')
              disp (accuracy)
end
    CLM = colormap(hsv(256));
    step = floor(256/OM);
    m=0;
    h=figure(1);
    plot(Z(:,1), Z(:,2), 'k.')
for i=1:OM
        plot( X(id==i,1), X(id==i,2 ), '*', 'Color', CLM(m+step,:)  )
        hold on
        plot( Z(knn_idZ(:)==i,1), Z( knn_idZ(:)==i,2 ), 'o', 'Color', CLM( m+step, :)  )
        hold on
        m = m + step;
end
end
```

```
datagen (train)
function [X,id,nc] = datagen(lang,OM)

mu = [1 2];
X = []; id =[];
if lang==1
fori =1:OM
prompt = {'Δώσε πλήθος στοιχείων i - οστής κλάσης:'};
      dlg_title = 'Σύνολο Εκπαίδευσης X';
num_lines = 1;
      def = {' '};
      options.Resize='on';
      options.WindowStyle='normal';
      nco = inputdlg(prompt,dlg_title,num_lines,def,options);
      nc(i) = str2num(nco{1,1});
%nc(i) = input('Δώσεπλήθοςστοιχείων i - οστήςομάδας :');
      xi = rand(nc(i),2 );
      mu = 7*i*[1 2];
      Sigma = [5*i .5; 0.8*i 2];
      xi = repmat(mu,nc(i),1) + randn(nc(i),2)*Sigma;
      X = [X ; xi];
      id = [id ; i*ones(nc(i),1)];

end
elseif lang==2
for i =1:OM
      prompt = {'number of elements for i - th cluster :'};
      dlg_title = 'Training Set X';
      num_lines = 1;
      def = {' '};
      options.Resize='on';
      options.WindowStyle='normal';
      nco = inputdlg(prompt,dlg_title,num_lines,def,options);
      nc(i) = str2num(nco{1,1});
      xi = rand(nc(i),2 );
      mu = 7*i*[1 2];
      Sigma = [5*i .5; 0.8*i 2];
      xi = repmat(mu,nc(i),1) + randn(nc(i),2)*Sigma;
      X = [X ; xi];
      id = [id ; i*ones(nc(i),1)];
```

```
end
end

% Save traing set
Y = [X id];
save('train_data_knn.txt', 'Y', '-ascii', '-double')

end
```

```
datagen (test)
function [Z,idZ,ncz] = datagen1(lang, OM)

mu = [1 2];
Z = []; idZ =[];

if lang==1
for i =1:OM
      prompt = {'Δώσε πλήθος Z στοιχείων i κλάσης :'};
dlg_title = 'Σύνολο Ελέγχου Z';
num_lines = 1;
      def = {' '};
      options.Resize='on';
      options.WindowStyle='normal';
      nco = inputdlg(prompt,dlg_title,num_lines,def,options);
      ncz(i) = str2num(nco{1,1});

      zi = rand(ncz(i),2);
      mu = 7*i*[1 2];
      Sigma = [5*i .5; 0.8*i 2];
      zi = repmat(mu,ncz(i),1) + randn(ncz(i),2)*Sigma;

       Z = [Z ; zi];
       idZ = [idZ ; i*ones(ncz(i),1)];
end% end for loop

elseif lang==2
for i =1:OM
      prompt = {'Number of Z elements :'};
      dlg_title = 'Test set Z';
      num_lines = 1;
      def = {' '};
      options.Resize='on';
      options.WindowStyle='normal';
      nco = inputdlg(prompt,dlg_title,num_lines,def,options);
      ncz(i) = str2num(nco{1,1});

      zi = rand(ncz(i),2);
      mu = 7*i*[1 2];
      Sigma = [5*i .5; 0.8*i 2];
```

```
        zi = repmat(mu,ncz(i),1) + randn(ncz(i),2)*Sigma;
        Z = [Z ; zi];

        idZ = [idZ ; i*ones(ncz(i),1)];
end
end

    Y = [Z idZ];
    save('test_data_knn.txt', 'Y', '-ascii', '-double')

end
```

```
confidence knn
%function conf(id, idZ)
% id: labels of the class that belong each data point (real
% classification)(id - k means, id - knn)
% idZ: labels that are produced after the kmeans clustering (cidx -
% k-means, knn_idZ - knn)
%
function [CFM] = conf_knn(id,idZ)

class_no = max(id);

[rp cp] = size(idZ);

maxidZ = max(idZ);
confusion_matrix = zeros(maxidZ,class_no);

for j=1:cp
if (idZ(j) >= 0) % not an outlier
      confusion_matrix(idZ(j),id(j)) = confusion_matrix(idZ(j),id(j)) + 1;
end% if
end% for j

CFM = confusion_matrix;

tmp = sum(confusion_matrix);
identified_points = sum(tmp);
total_points = cp;
```

```
υπολογισμος εντροπιας καθαροτητας (entropypurity)
function [entropy total_clustered_points accuracy] = estimate_entropy_accuracy_knn(id,idZ)

% Transpose the arrays to become a line-vector
id = id';
idZ = idZ';

CFM = conf_knn(id,idZ);

[clusters classes] = size(CFM);

if (clusters == 0 || classes == 0)
   entropy = NaN;
   total_clustered_points = -5;
   accuracy = NaN;
return
end

entropy_vector = zeros(1,clusters);

entropy_sum = 0;  accuracy = 0;

[r num_points] = size(idZ);

for i=1:clusters
   sizes(i) = sum(CFM(i,:));
   accuracy = accuracy + max(CFM(i,:));

for j=1:classes
if (CFM(i,j) ~= 0)
        p = (CFM(i,j)./sizes(i));
        entropy_sum = entropy_sum - (p) * log2(p);
end
end
   entropy_vector(i) = entropy_vector(i) + entropy_sum;

   entropy_sum = 0;
end

accuracy = accuracy / num_points;
```

```
total_clustered_points = sum(sizes);


entropy = 0;
for i=1:1:clusters
   entropy = entropy + (sizes(i)./total_clustered_points)*entropy_vector(i)';
end
```