ΠΑΝΤΕΙΟΝ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΟΙΝΩΝΙΚΩΝ ΚΑΙ ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ

ΤΜΗΜΑ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΔΗΜΟΣΙΑ ΔΙΟΙΚΗΣΗ»

ΚΑΤΕΥΘΥΝΣΗ «ΟΙΚΟΝΟΜΙΚΗ ΕΠΙΣΤΗΜΗ»

Εξόρυξη Δεδομένων και εφαρμογές στην Οικονομική Επιστήμη

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βασιλική Ψυχαλοπούλου

Αθήνα, 2020

Τριμελής Επιτροπή

Μιχαήλ Φιλιππάκης, Αναπληρωτής Καθηγητής Πανεπιστημίου Πειραιά (Επιβλέπων)

Γεώργιος Πετράκος, Αναπληρωτής Καθηγητής Παντείου Πανεπιστημίου

Βασίλειος Κέφης, Καθηγητής Παντείου Πανεπιστημίου

Στην οικογένειά μου

Περιεχόμενα

Σχήματα

Εικόνες

Περίληψη

Σκοπός της εν λόγω εργασίας είναι η μελέτη και ανάλυση των τεχνικών Εξόρυξης Δεδομένων και των εφαρμογών τους στην Οικονομική Επιστήμη. Αρχικά, γίνεται αποσαφήνιση των εννοιών της Μηχανικής Μάθησης, της Ανακάλυψης Γνώσης και της Εξόρυξης Δεδομένων και κατόπιν, περιγράφονται η διαδικασία, οι μέθοδοι και τα βασικά εργαλεία για την Εξόρυξη Δεδομένων. Επιπλέον, αναδεικνύεται η σημασία της εφαρμογής τους στο ευρύ πεδίο των Οικονομικών. Στη συνέχεια, αναλύονται εκτενώς οι δύο πιο βασικές μέθοδοι Εξόρυξης Δεδομένων, η Κατηγοριοποίηση και η Συσταδοποίηση και παρουσιάζεται η λογική και τα βήματα των αλγορίθμων που εμπίπτουν στις συγκεκριμένες μεθόδους. Τέλος, γίνεται χρήση του λογισμικού MATLAB για την επίλυση ενός προβλήματος Κατηγοριοποίησης σχετικά με την θετική ή αρνητική απόκριση των πελατών μιας τράπεζας σε προθεσμιακές καταθέσεις. Τα αποτελέσματα έδειξαν ότι το μοντέλο που προέκυψε από τον αλγόριθμο Medium Tree είναι το πιο ακριβές και μπορεί να προβλέψει σωστά το 88,8% των απαντήσεων των πελατών.

Λέξεις-κλειδιά: Εξόρυξη Δεδομένων, Κατηγοριοποίηση, MATLAB, Οικονομική Επιστήμη, Συσταδοποίηση

Data mining and applications in Economic Science

Vasiliki Psychalopoulou

Abstract

The aim of this thesis is to study and analyze the Data Mining techniques and applications in economic Science. Initially, it clarifies the meanings of Machine Learning, Knowledge Discovery and Data Mining and, then describes the process, the methods and the basic tools for Data Mining. Moreover, it emerges the significance of the application of these techniques in the wide field of Economics. Subsequently, it analyze extensively Classification and Clustering , the two most important methods of Data Mining and shows the logic and steps of the algorithms that fall under these methods. Finally, it resolves a Classification problem with MATLAB software related with positive or negative response of bank's clients in term deposits. The results showed that the model which came up from the medium tree algorithm is the most accurate and able to predict right the 88,8% of the client's answers.

Keywords: Classification, Clustering, Data Mining, Economic Science, MATLAB

Εισαγωγή

Κατά την έλευση των τελευταίων δεκαετιών οι έννοιες «πληροφορία», «τεχνολογία», «επιστήμη», «οικονομία» έχουν αποκτήσει εξέχοντα ρόλο σε κάθε έκφανση του ανθρώπινου βίου. Η ραγδαία ανάπτυξη της τεχνολογίας και των τεχνολογικών επιστημών, ο τεράστιος όγκος της διαθέσιμης πληροφορίας, οι κοινωνικέ εξελίξεις σε κάθε επίπεδο και εξάπλωση των οικονομικών σχέσεων υπό το πρίσμα της παγκοσμιοποίησης έχουν συνυφανθεί μετατρέποντας τις σύγχρονες κοινωνίες σε «κοινωνίες της γνώσης». Σύμφωνα με τον Drucker (2000) «ο καθοριστικός συντελεστής της παραγωγής δεν είναι πια σήμερα το κεφάλαιο , ούτε το έδαφος, ούτε η εργασία. Είναι οι γνώσεις». Επομένως, οι σύγχρονες οικονομίες θεωρούνται «οικονομίες της γνώσης», διότι στηρίζονται στην παραγωγή και τη χρήση της γνώσης και της πληροφορίας. Ως εκ τούτου, η γνώση, βρίσκεται στον πυρήνα του κοινωνικού και οικονομικού φάσματος. Αν λοιπόν, μέχρι σήμερα το ζητούμενο ήταν η εξεύρεση της απαραίτητης πληροφορίας, τώρα πια είναι η σωστή διαχείριση και επεξεργασία αυτής της πληροφορίας προκειμένου να φτάσουμε στη γνώση.

Ο εντοπισμός κρυμμένης γνώσης βρίσκεται στο επίκεντρο ενός ευρύ φάσματος επιστημονικών κλάδων. Ένας από αυτούς τους κλάδους είναι και ο κλάδος των Οικονομικών. Η διοίκηση των επιχειρήσεων, η προώθηση των προϊόντων, η ικανοποίηση των καταναλωτών, το κόστος παραγωγής, ο εντοπισμός οικονομικής απάτης, ο έλεγχος χρηματοοικονομικών καταστάσεων, η δημιουργία προβλέψεων για τη βιωσιμότητα ενός οργανισμού είναι ορισμένα μόνο από τα σημεία ενδιαφέροντος για ανακάλυψη γνώσης που εντοπίζονται στο πεδίο αυτό.

Ώθηση στο ενδιαφέρον αυτό έχει δώσει η ραγδαία εξέλιξη της μηχανικής μάθησης. Η μηχανική μάθηση έχει αναπτύξει εργαλεία με τα οποία η διαχείριση τεράστιου όγκου δεδομένων γίνεται εύκολα επεξεργάσιμη και ικανή για την παραγωγή νέας γνώσης. Τα σημαντικότερα από αυτά τα εργαλεία συνοψίζονται στις τεχνικές Εξόρυξης Δεδομένων.

Η παρούσα εργασία έχει ως αντικείμενο την μελέτη των τεχνικών Εξόρυξης Δεδομένων και την χρησιμότητα τους στην Οικονομική Επιστήμη. Αναλυτικότερα, στο πρώτο κεφάλαιο γίνεται εισαγωγή στις έννοιες της Μηχανικής Μάθησης, της Ανακάλυψης Γνώσης και της Εξόρυξης Δεδομένων. Περιγράφονται η διαδικασία, οι μέθοδοι και το λογισμικό εξόρυξης δεδομένων. Επίσης, γίνονται αναφορές στην εφαρμογή τους σε διάφορους κλάδους της Οικονομίας. Στο δεύτερο και το τρίτο κεφάλαιο γίνεται εκτενέστερη περιγραφή των δυο πιο διαδεδομένων μεθόδων

εξόρυξης δεδομένων, της Κατηγοριοποίησης και της Συσταδοποίησης αντίστοιχα. Τέλος, στο τέταρτο κεφάλαιο παρουσιάζεται η επίλυση ενός απλού προβλήματος με τη μέθοδο της Κατηγοριοποίησης και χρήση του λογισμικού MATLAB.

Κεφάλαιο 1^{ο} Ανακάλυψη Γνώσης

1.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση (*Machine Learning*) είναι μέθοδος ανάλυσης δεδομένων η οποία, χρησιμοποιώντας αλγορίθμους που μαθαίνουν επανειλημμένα μέσα από δεδομένα, επιτρέπει στους υπολογιστές να βρίσκουν κρυμμένη γνώση χωρίς να χρειάζεται να προγραμματιστούν πλήρως για το που θα την βρουν. Παράγοντες όπως οι συνεχώς αναπτυσσόμενοι όγκοι διαθέσιμων δεδομένων, η φτηνότερη πια και μεγαλύτερη υπολογιστική ισχύς καθώς και η προσιτή οικονομικά αποθήκευση δεδομένων βοηθούν στην γρήγορη και αυτοματοποιημένη δημιουργία μοντέλων που μπορούν να αναλύσουν περισσότερα και πιο πολύπλοκα δεδομένα και να παράγουν ακριβή αποτελέσματα σε μεγαλύτερη κλίμακα. Η μηχανική μάθηση χωρίζεται σε δύο βασικές κατηγορίες, την επιβλεπόμενη μάθηση και την μη επιβλεπόμενη μάθηση.

Ο στόχος της επιβλεπόμενης μάθησης (*supervised learning*) είναι να δημιουργήσει ένα μοντέλο το οποίο θα κάνει προβλέψεις βασισμένο σε ενδείξεις με την παρουσία αβεβαιότητας. Ως προσαρμοστικός αλγόριθμος που είναι, αναγνωρίζει μοτίβα μέσα σε έναν όγκο δεδομένων και έτσι ένας υπολογιστής μαθαίνει μέσα από παρατηρήσεις. Συνεπώς, όσες περισσότερες παρατηρήσεις έχουμε τόσο βελτιώνεται και η ικανότητα πρόβλεψης του αλγορίθμου. Πιο συγκεκριμένα ένας αλγόριθμος επιβλεπόμενης μάθησης παίρνει μία ομάδα εισόδων και έχει και την αντίστοιχη ομάδα εξόδων και με βάση τις γνωστές αντιστοιχίες που ήδη έχει, αντιδράει αναλόγως σε νέα δεδομένα.

Αντιθέτως, στη μη επιβλεπόμενη μάθηση (*unsupervised learning*) έχουμε μόνο εισόδους και όχι εξόδους. Ως εκ τούτου, αυτό που προσπαθούμε είναι να βρεθεί μία σχέση, μία δομή μεταξύ των διαφορετικών εισόδων που έχουμε. Κυρίαρχη προσέγγιση μη επιβλεπόμενης μάθησης είναι η συσταδοποίηση, κατά την οποία βάζουμε τις εισόδους μας σε ομάδες και τα καινούρια δεδομένα που συναντάμε τα τοποθετούμε σε μία από τις ομάδες αυτές.

Αξίζει να σημειωθεί ότι η Μηχανική Μάθηση διαφέρει από την παραδοσιακή στατιστική μοντελοποίηση με διάφορους τρόπους. Η μηχανική μάθηση δημιουργεί δομικά συστήματα τα οποία μπορούν να μάθουν από τα δεδομένα και δίνει έμφαση στην βελτίωση και την απόδοση του συμπεράσματος. Αναφέρεται σε αυτοματοποιημένες διαδικασίες που διαχειρίζονται με επιτυχία μεγάλα σύνολα δεδομένων. Αντίθετα, η στατιστική μοντελοποίηση τονίζει τη θεωρία που βασίζεται

σε δοκιμασία υποθέσεων. Τα κλασικά στατιστικά εργαλεία εποπτεύονται από εκπαιδευμένο ερευνητή ο οποίος πρέπει να εξετάσει τα δεδομένα, την προέλευση τους τις ιδιότητες των εκτιμητών που θα χρησιμοποιήσει για την ανάλυση κλπ. Όπως είναι προφανές, δεν είναι δυνατό τέτοιες τεχνικές να εφαρμοστούν σε τεράστιο όγκο δεδομένων.

1.2 Εξόρυξη Δεδομένων

Ο όρος Εξόρυξη Δεδομένων (*Data Mining*) χρησιμοποιείται για να περιγράψει μια διαδικασία εξαγωγής πληροφοριών, μέσω της δημιουργίας προτύπων (*patterns*), από μεγάλες ποσότητες δεδομένων με τη χρήση κατάλληλων αλγορίθμων. Ανά τα χρόνια, έχουν χρησιμοποιηθεί διαφορετικοί ορισμοί για να περιγράψουν τη διαδικασία εξεύρεσης χρήσιμων μοντέλων στα δεδομένα όπως Εξαγωγή Γνώσης, Ανακάλυψη Πληροφοριών, Συλλογή Πληροφοριών, Ανασκαφή Δεδομένων, Διαδικασία Μοντελοποίησης Δεδομένων χωρίς να γνωρίσουν ιδιαίτερη απήχηση.

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (*Knowledge Discovery in Database* - KDD) είναι ένας ακόμη όρος, πληρέστερος θα μπορούσαμε να πούμε, που αναφέρεται στην ίδια διαδικασία. Σύμφωνα με τα παραπάνω, η εξόρυξη δεδομένων αναφέρεται στην εφαρμογή συγκεκριμένων αλγορίθμων για την εξαγωγή μοντέλων από τα δεδομένα. Ουσιαστικά πρόκειται για ένα μόνο στάδιο μιας ευρύτερης διαδικασίας, της KDD, για την ανακάλυψη χρήσιμης γνώσης από τα δεδομένα . Ο όρος KDD επινοήθηκε για να δοθεί έμφαση στο ότι η ανακάλυψη νέας γνώσης είναι το ζητούμενο αποτέλεσμα από τη διαδικασία επεξεργασίας των δεδομένων (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Ο πιο πρόσφατος ορισμός, Επιστήμη των Δεδομένων (*Data Science*), χρησιμοποιείται εναλλακτικά με την Εξόρυξη Δεδομένων και την Ανακάλυψη Γνώσης από Βάσεις Δεδομένων.

1.2.1 Εξόρυξη Δεδομένων και Ανακάλυψη Γνώσης

Όπως ήδη αναφέρθηκε, οι όροι Εξόρυξη Δεδομένων και Ανακάλυψη Γνώσης από Βάσεις Δεδομένων συχνά ταυτίζονται, παρόλο που αναφέρονται σε διαφορετικές διαδικασίες. Η KDD είναι μια μη τετριμμένη διαδικασία εντοπισμού έγκυρων, νέων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων στα δεδομένα (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Τα δεδομένα είναι ένα σύνολο γεγονότων και τα πρότυπα είναι εκφράσεις σε μια γλώσσα που περιγράφουν ένα υποσύνολο των

δεδομένων ή ένα μοντέλο ου μπορεί να εφαρμοστεί στα δεδομένα. Η KDD αναφέρεται ως μη τετριμμένη υπό την έννοια ότι δεν είναι ένας απλός προκαθορισμένος υπολογισμός, αλλά προϋποθέτει έρευνα και καταλήγει σε συμπεράσματα. Τέλος αποτελεί διαδικασία διότι περιλαμβάνει συγκεκριμένα στάδια (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Σύμφωνα με τους Brachman & Anand (1994), η ανακάλυψη της γνώσης είναι μια σύνθετη διαδικασία και απαιτεί την ανθρώπινη συμμετοχή. Επομένως, είναι πρωταρχικής σημασίας να προσπαθήσουμε να καταλάβουμε τι κάνει ο άνθρωπος που πραγματοποιεί αυτή τη διαδικασία. Σε μια σύντομη ανάλυση η KDD περιλαμβάνει τα εξής βήματα (Han, Kamber, & Pei, 2012):

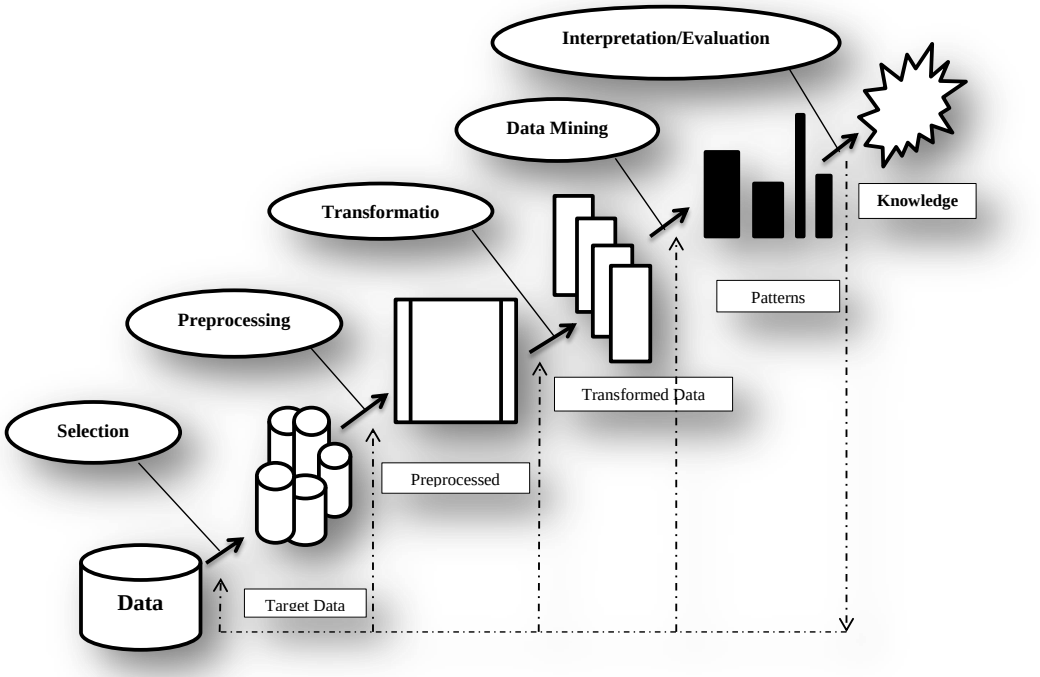
Βήμα 1. Συλλογή Δεδομένων (*Selection*). Το πρώτο βήμα της KDD είναι η ανάκτηση των δεδομένων που πρόκειται να αναλυθούν, από τη βάση δεδομένων. Τα δεδομένα αυτά είναι πιθανό να είναι ελλιπή ή θορυβώδη, προβλήματα που καλείται να επιλύσει το δεύτερο στάδιο.

Βήμα 2. Προ-επεξεργασία Δεδομένων (*Preprocessing*). Το στάδιο αυτό είναι εξαιρετικά σημαντικό για να εξασφαλιστεί η ποιότητα των αποτελεσμάτων. Περιλαμβάνει τον καθαρισμό και την ενσωμάτωση των δεδομένων για την εξάλειψη των θορύβων και των ασυνεπών δεδομένων.

Βήμα 3. Μετασχηματισμός Δεδομένων (*Transformation*). Το τρίτο στάδιο της KDD περιλαμβάνει τον μετασχηματισμό δεδομένων σε κατάλληλες μορφές για την εξόρυξη.

Βήμα 4. Εξόρυξη Δεδομένων (*Data Mining*) Στο βήμα αυτό γίνεται εφαρμογή αλγορίθμων για την δημιουργία ενός μοντέλου. Τα μοντέλα αυτά είναι συνήθως κατηγοριοποίησης ή πρόβλεψης.

Βήμα 5. Διερμηνεία και Αξιολόγηση (*Interpretation and Evaluation*). Στο στάδιο αυτό γίνεται η τελική διαδικασία της ερμηνείας των αποτελεσμάτων και της αξιολόγησης.



Σχήμα 1. Βήματα Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

1.2.2 Εξόρυξη Δεδομένων και Βάσεις Δεδομένων

Η Βάση Δεδομένων (*Database*) αποτελείται από μία συλλογή δεδομένων που σχετίζονται μεταξύ τους. Η πρόσβαση στα δεδομένα και η διαχείρισή τους υλοποιείται με τη χρήση ενός συνόλου προγραμμάτων ειδικού λογισμικού, γνωστό ως Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ). Οι Βάσεις Δεδομένων και τα προγράμματα λογισμικού συναποτελούν ένα Σύστημα Βάσης Δεδομένων. Οι Βάσεις Δεδομένων χωρίζονται σε κατηγορίες ανάλογα με τα μοντέλα δεδομένων που τις περιγράφουν. Τα μοντέλα δεδομένων που είναι βασισμένα στο αντικείμενο περιλαμβάνουν το μοντέλο οντοτήτων συσχετίσεων, το αντικειμενοστραφές μοντέλο κλπ. Σε αυτά που βασίζονται στις εγγραφές ανήκουν το ιεραρχικό, το δικτυακό και το σχεσιακό. Το σχεσιακό μοντέλο προσδιορίζει τις Σχεσιακές Βάσεις Δεδομένων.

Μια Σχεσιακή Βάση Δεδομένων είναι μια συλλογή από πίνακες, κάθε μια εκχωρημένη με ένα συγκεκριμένο όνομα. Οι πίνακες αναπαριστούν τις σχέσεις των δεδομένων και αποτελούνται από χαρακτηριστικά και πλειάδες. Τα χαρακτηριστικά των πινάκων παρουσιάζονται σε στήλες και οι πλειάδες, οι οποίες περιγράφονται από ένα σύνολο τιμών χαρακτηριστικών, σε γραμμές. Οι Σχεσιακές Βάσεις Δεδομένων είναι συχνά διαθέσιμες και είναι οι πιο σημαντικές αποθήκες πληροφορικής.

Επομένως αποτελούν μια σημαντική μορφή δεδομένων στη μελέτη της εξόρυξης δεδομένων.

Ένας ειδικός τύπος δεδομένων, η Αποθήκη Δεδομένων (*Data Warehouse*), είναι ένα αποθετήριο πληροφοριών που συλλέγονται από πολλαπλές πηγές και κατασκευάζονται μέσω διαδικασίας καθαρισμού, ολοκλήρωσης, μετασχηματισμού, φόρτωσης και ανανέωσης δεδομένων. Οι αποθήκες δεδομένων υποστηρίζουν την αναλυτική επεξεργασία των δεδομένων (*On-line analytical processing* – OLAP) συμβάλλοντας στη λήψη αποφάσεων. Οι παραδοσιακές βάσεις δεδομένων, από την άλλη πλευρά, υποστηρίζουν τα συστήματα επεξεργασίας συναλλαγών (*On-line Transactional processing* – OLTP) σε επίπεδο επιχειρησιακών λειτουργιών.

1.3 Διαδικασία Ανακάλυψης Γνώσης

1.3.1 Προσδιορισμός στόχου και επιλογή δεδομένων

Πρωταρχικός σκοπός στην διαδικασία ανακάλυψης γνώσης είναι ο προσδιορισμός ενός στόχου. Ο εκάστοτε αναλυτής θα προβεί στην εν λόγω διαδικασία στην προσπάθειά του να βρει νέες, χρήσιμες πληροφορίες που θα συμβάλουν στην επίλυση ενός προβλήματος, στην απάντηση κάποιου ερωτήματος. Επομένως, έχει σημασία να γίνει εξ αρχής ξεκάθαρο και κατανοητό το επιδιωκόμενο αποτέλεσμα. Σε δεύτερο στάδιο γίνεται η επιλογή των δεδομένων επί των οποίων θα γίνει η ανακάλυψη γνώσης. Ουσιαστικά πρόκειται για ένα σύνολο δεδομένων ή για ένα υποσύνολο αυτού.

1.3.2 Προ-επεξεργασία δεδομένων

Η προ-επεξεργασία των δεδομένων αποτελεί το σημαντικότερο στάδιο για την ανακάλυψη γνώσης διότι καθορίζει την ποιότητα του αποτελέσματος. Πρόκειται για το πιο απαιτητικό και χρονοβόρο μέρος της όλης διαδικασίας. Τρεις βασικές διεργασίες λαμβάνουν χώρα στην προ-επεξεργασία των δεδομένων, ο καθαρισμός, η ενοποίηση, ο μετασχηματισμός και η μείωση.

Καθαρισμός και Ενοποίηση. Στα δεδομένα που έχουν συλλεχθεί είναι πιθανό να μην υπάρχουν καταγεγραμμένες τιμές για ορισμένα χαρακτηριστικά ή να υπάρχουν καταγεγραμμένες λανθασμένες τιμές εξαιτίας διαφόρων παραγόντων, όπως δυσλειτουργικού εξοπλισμού καταγραφής ή προβλημάτων κατά την εγγραφή. Παρουσία τέτοιας περίπτωσης, υπάρχει κίνδυνος η ανάλυση να οδηγήσει σε μη

ποιοτικά αποτελέσματα. Τα προβλήματα αυτά, των ελλιπών και θορυβωδών δεδομένων, καλείται να επιλύσει το συγκεκριμένο στάδιο.

Ο καθαρισμός των δεδομένων περιλαμβάνει τη συμπλήρωση ελλιπών τιμών, τη διόρθωση ασυνεπειών στα δεδομένα καθώς και την αναγνώριση και εξομάλυνση ακραίων τιμών σε περίπτωση που περιέχουν θόρυβο. Η ενοποίηση έχει σαν στόχο να συνδυάσει δεδομένα από πολλές διαφορετικές πηγές ώστε να αφαιρεθούν περιττές πληροφορίες και να βελτιωθεί η όλη διαδικασία.

Μετασχηματισμός. Η επίτευξη θετικών αποτελεσμάτων από την ανακάλυψη νέας γνώσης προϋποθέτει την ύπαρξη και ανάλυση συγκρίσιμων δεδομένων. Βασικός στόχος του μετασχηματισμού είναι η δημιουργία τέτοιων δεδομένων. Ειδική μορφή μετασχηματισμού αποτελεί η διακριτοποίηση. Πρόκειται για την μετατροπή ενός εύρους συνεχών τιμών σε διακριτές.

Μείωση Δεδομένων. Μια ακόμη διαδικασία προ-επεξεργασίας αποτελεί η μείωση των δεδομένων. Στόχος είναι η δημιουργία μικρότερων συνόλων δεδομένων που θα απαιτούν λιγότερο χρόνο για επεξεργασία αλλά θα δίνουν εξίσου ικανοποιητικά αποτελέσματα.

1.3.3 Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων περιλαμβάνει τη δημιουργία μοντέλων ή τη διαμόρφωση προτύπων από τα παρατηρούμενα δεδομένα (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Αποτελεί το βασικότερο στάδιο διότι εδώ πραγματοποιείται ουσιαστικά η ανακάλυψη γνώσης. Δεν είναι τυχαίο άλλωστε που ο όρος είθισται να χρησιμοποιείται παραπέμποντας στο σύνολο της διαδικασίας. Σημείο αναφοράς στην εξόρυξη αποτελεί η εφαρμογή, συχνά επαναλαμβανόμενη, του κατάλληλου αλγορίθμου, για να εξαχθεί ένα μοντέλο το οποίο θα δίνει απαντήσεις για νέα δεδομένα.

Δύο τύποι μοντέλων προκύπτουν από την εξόρυξη δεδομένων, τα περιγραφικά (*descriptive*) και τα μοντέλα πρόβλεψης (*predictive*). Τα μοντέλα περιγραφής εξηγούν τη συμπεριφορά των υπαρχόντων δεδομένων ενώ τα μοντέλα πρόβλεψης, προβλέπουν τιμές για συγκεκριμένα χαρακτηριστικά που παρουσιάζουν κάποιο ενδιαφέρον.

1.3.4 Διερμηνεία και Αξιολόγηση

Στο τελικό στάδιο της διαδικασίας γίνεται, όπως είναι προφανές, η διερμηνεία και η αξιολόγηση των αποτελεσμάτων της ανάλυσης. Αξίζει να σημειωθεί, επειδή αρκετές φορές δημιουργείται σύγχυση, ότι δεν ερμηνεύονται τα μοντέλα που προκύπτουν από την εξόρυξη δεδομένων, αλλά τα νέα στοιχεία που σχετίζονται με τα δεδομένα.

1.4 Μέθοδοι Εξόρυξης Δεδομένων

Όπως ήδη αναφέρθηκε, η εξόρυξη δεδομένων έχει δύο συγκεκριμένους στόχους, την πρόβλεψη και την περιγραφή. Η περιγραφή παρέχει μια πολύ κατανοητή και εύκολα αξιοποιήσιμη αναπαράσταση των δεδομένων μέσω της ανακάλυψης προτύπων. Μια επιτυχημένη περιγραφή είναι ικανή να εξηγήσει σε μεγάλο βαθμό τη συμπεριφορά των δεδομένων. Η Συσταδοποίηση (*Clustering*), η Οπτικοποίηση (*Visualization*), οι Κανόνες Συσχέτισης (*Association Rules*) και τα Πρότυπα Ακολουθιών (*Sequential patterns*) είναι κάποιες από τις περιγραφικές μεθόδους. Η πρόβλεψη έχει σαν στόχο τον προσδιορισμό της μελλοντικής συμπεριφοράς κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον και βασίζονται στη συμπεριφορά άλλων μεταβλητών. Οι μέθοδοι πρόβλεψης είναι η Κατηγοριοποίηση (*Classification*) και η Παλινδρόμηση (*Regression*). Παρακάτω ακολουθεί μια παρουσίαση των μεθόδων αυτών (Χαλκίδη & Βαζιργιάννης, 2005).

1.4.1 Κατηγοριοποίηση

Η κατηγοριοποίηση αποτελεί μια από τις βασικές διεργασίες εξόρυξης δεδομένων. Σκοπός της είναι η ταξινόμηση ενός στοιχείου σε ένα σύνολο κατηγοριών που είναι ήδη καθορισμένες. Αναφέρεται στη δημιουργία ενός προτύπου που μπορεί να χρησιμοποιηθεί τόσο για την κατανόηση των υφιστάμενων δεδομένων όσο και για την πρόβλεψη του τρόπου με τον οποίο μελλοντικά δεδομένα θα συμπεριφερθούν. Σημειώνεται πως στη βιβλιογραφία οι όροι ταξινόμηση και κατηγοριοποίηση παρουσιάζονται ως συνώνυμοι.

Η κατηγοριοποίηση των νέων στοιχείων απαιτεί τη χρήση συγκεκριμένων τεχνικών. Οι δημοφιλέστερες τεχνικές κατηγοριοποίησης είναι τα Νευρωνικά Δίκτυα (*Neural Networks*), η Bayesian κατηγοριοποίηση, τα Δέντρα Αποφάσεων (*Decision Trees*), η κατηγοριοποίηση κοντινότερων γειτόνων και τα Support Vector Machines. (Χαλκίδη & Βαρζιγιάννης, 2005)

1.4.2 Παλινδρόμηση

Η παλινδρόμηση είναι μια λειτουργία που εκχωρεί υπάρχουσες τιμές σε μια ή περισσότερες μεταβλητές x_1, x_2,…, x_n , για να προβλέψει την τιμή μιας μεταβλητής y. Στην απλούστερη περίπτωση, χρησιμοποιούνται τυπικές στατιστικές τεχνικές όπως η γραμμική παλινδρόμηση (*linear regression*), που προσπαθεί να περιγράψει την σχέση των μεταβλητών x (ανεξάρτητη) και y (εξαρτημένη) με τη χρήση του υποδείγματος:

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_n x_{ni}$$

Στον πραγματικό κόσμο τα προβλήματα δεν είναι απλά γραμμικές προβολές προηγούμενων τιμών. Για παράδειγμα, ο όγκος πωλήσεων και οι τιμές ενός προϊόντος είναι δύσκολο να προβλεφθούν καθώς εξαρτώνται από πολύπλοκες αλληλεπιδράσεις πολλαπλών μεταβλητών. Επομένως, πιο σύνθετες τεχνικές απαιτούνται για την πρόβλεψη μελλοντικών τιμών.

1.4.3 Συσταδοποίηση

Η συσταδοποίηση αποτελεί μια από τις πιο χρήσιμες διεργασίες στην διαδικασία εξόρυξης γνώσης για τον προσδιορισμό προτύπων στα δεδομένα που αναλύονται. Πρόκειται για την τμηματοποίηση (*partitioning*) του συνόλου των δεδομένων σε συστάδες έτσι ώστε τα στοιχεία που ανήκουν σε μια συστάδα να παρουσιάζουν περισσότερες ομοιότητες μεταξύ τους απ’ ότι με τα στοιχεία των υπολοίπων συστάδων. Επομένως, βασικός σκοπός της διαδικασίας είναι να οδηγήσει στην οργάνωση προτύπων σε συστάδες οι οποίες θα μας επιτρέψουν να εξάγουμε χρήσιμα συμπεράσματα.

1.4.4 Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης αφορούν μια περιγραφική προσέγγιση για την αναζήτηση δεδομένων που μπορούν να βοηθήσουν στον εντοπισμό των σχέσεων μεταξύ τιμών σε μια βάση δεδομένων. Παρουσιάζουν έντονο ενδιαφέρον διότι εκφράζουν με συνοπτικό τρόπο, εύκολα κατανοητό από τον τελικό χρήστη, τις ενδεχομένως χρήσιμες πληροφορίες. Ειδικότερα, ανακαλύπτουν κρυμμένες συσχετίσεις μεταξύ των γνωρισμάτων σε κάποιο σύνολο δεδομένων. Οι πληροφορίες που μπορούν να ανακαλύψουν οι κανόνες συσχέτισης είναι πολύ σημαντικές και αφορούν διάφορους τομείς της ανθρώπινης δραστηριότητας.

Οι κανόνες συσχέτισης πρωτοεμφανίστηκαν για την ανάγκη ανάλυσης του «καλαθιού αγοράς» του καταναλωτή από τις υπεραγορές. Η τεχνολογική εξέλιξη, προσφέροντας τη δυνατότητα ηλεκτρονικής καταχώρησης των συναλλαγών κάθε καταναλωτή, οδήγησε στη συσσώρευση τεράστιου όγκου διαθέσιμης πληροφορίας σχετικά με τις προτιμήσεις των καταναλωτών. Ανέκυψε, λοιπόν, η ιδέα για αξιοποίηση αυτής της πληροφορίας.

Όπως ήδη αναφέρθηκε, τα καλάθια αγοράς των καταναλωτών εκφράζονται μέσω των κανόνων συσχέτισης. Πιο συγκεκριμένα, οι κανόνες αυτοί εμφανίζουν ένα αίτιο το οποίο συνδέουν με ένα αποτέλεσμα, είναι δηλαδή κανόνες «if-then» που συσχετίζουν αντικείμενα μεταξύ τους. Η συμβολή τους αναδείχθηκε ιδιαιτέρως χρήσιμη σε τομείς όπως η προώθηση προϊόντων, η διαχείριση αποθεμάτων και η τοποθέτηση των προϊόντων στα ράφια. (Χαλκίδη & Βαζιργιάννης, 2005)

1.4.5 Πρότυπα ακολουθιών

Η ανακάλυψη προτύπων ακολουθιών είναι παρόμοια διαδικασία με την εξαγωγή κανόνων συσχέτισης, υπό την έννοια ότι πρόκειται για συχνά εμφανιζόμενα πρότυπα που σχετίζονται με τον χρόνο. Τα ακολουθιακά δεδομένα είναι διαθέσιμα και χρησιμοποιούνται παντού στην καθημερινή ζωή. Χαρακτηριστικά παραδείγματα τέτοιων δεδομένων είναι τα κείμενα, τα δεδομένα καιρού, τα επιχειρησιακά δεδομένα, τα αρχεία ιατρικών ιστορικών και άλλα.

1.4.6 Οπτικοποίηση

Η γραφική απεικόνιση δίνει στον άνθρωπο τη δυνατότητα να αντιληφθεί καλύτερα τα δεδομένα συγκριτικά με την παρουσίασή τους σε μια λίστα αριθμών. Μερικές από τις κοινές και πολύ χρήσιμες γραφικές απεικονίσεις δεδομένων είναι τα ιστογράμματα, τα γραφήματα κιβωτίων που εμφανίζουν κατανομές τιμών, τα δισδιάστατα ή τρισδιάστατα scatter plot κλπ. Το πρόβλημα στη χρήση της απεικόνισης πηγάζει από το γεγονός ότι τα μοντέλα έχουν πολλές διαστάσεις ή μεταβλητές, αλλά περιορίζουμε την εμφάνιση αυτών των διαστάσεων σε μια δισδιάστατη οθόνη υπολογιστή ή χαρτί.

1.5 Λογισμικό Εξόρυξης Δεδομένων

Τα στοιχεία των βάσεων δεδομένων μπορούν να αναλυθούν μέσω διαφόρων προγραμμάτων λογισμικού. Τα πιο διαδεδομένα από αυτά είναι το Weka, η γλώσσα προγραμματισμού R και το MATLAB. Το Weka (Waikato Environment for

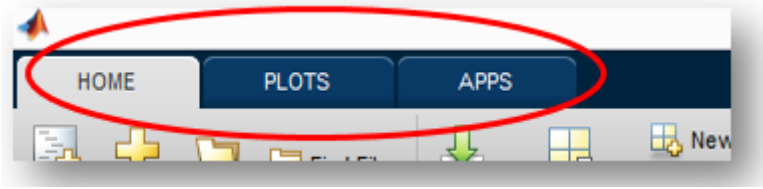
Knowledge Analysis) είναι μια δημοφιλής πλατφόρμα λογισμικού μηχανικής μάθησης που σχεδιάστηκε στο Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Ουσιαστικά, πρόκειται για μια συλλογή αλγορίθμων μηχανικής μάθησης για εργασίες εξόρυξης δεδομένων. Το Weka είναι ένα πρόγραμμα εξαιρετικά χρήσιμο για τον εκάστοτε ενδιαφερόμενο σε εργασίες εξόρυξης δεδομένων καθώς περιέχει εργαλεία για την προετοιμασία των δεδομένων, την ταξινόμηση, την παλινδρόμηση, την ομαδοποίηση, την εξόρυξη κανόνων συσχέτισης και την απεικόνιση. Η R είναι μια γλώσσα προγραμματισμού και ένα περιβάλλον που επιτρέπει στο χρήστη την υλοποίηση στατιστικών υπολογισμών και την κατασκευή γραφημάτων. Παρέχει μια ευρεία ποικιλία στατιστικών τεχνικών όπως, γραμμική και μη γραμμική μοντελοποίηση, κλασικές στατιστικές δοκιμές, ανάλυση χρονοσειρών, ταξινόμηση, ομαδοποίηση καθώς και γραφικών τεχνικών. Τέλος, το MATLAB, επειδή αποτελεί το εργαλείο που έχει χρησιμοποιηθεί στην παρούσα εργασία, αναλύεται εκτενέστερα παρακάτω.

1.5.1 MATLAB

Το MATLAB (MATrix LABoratory) είναι ένα λογισμικό πακέτο για αριθμητικούς υπολογισμούς και απεικονίσεις υψηλού επιπέδου. Παρέχει υπολογιστικό, οπτικό και προγραμματιστικό περιβάλλον με εκατοντάδες ενσωματωμένες συναρτήσεις οι οποίες αποτελούν ένα χρήσιμο εργαλείο για την επίλυση προβλημάτων γραμμικής άλγεβρας, ανάλυσης δεδομένων, απεικόνισης δεδομένων και συναρτήσεων σε χώρους δύο και τριών διαστάσεων, καθώς και την επίλυση πολλαπλών τύπων προβλημάτων διαφόρων επιστημονικών πεδίων. Ο χρήστης, ωστόσο, μπορεί να επεκτείνει την βιβλιοθήκη των έτοιμων συναρτήσεων, γράφοντας δικές του συναρτήσεις με την γλώσσα υψηλού επιπέδου που διαθέτει. Επιπρόσθετα, υπάρχουν πολλές διαθέσιμες εργαλειοθήκες-εφαρμογές (*toolboxes*), οι οποίες αποτελούνται από πακέτα εξειδικευμένων συναρτήσεων, γραμμένες για εφαρμογές όπως η επεξεργασία εικόνας, τα νευρωσικά δίκτυα, η ανάλυση δεδομένων, ο έλεγχος συστημάτων κ.λπ. Τα παραπάνω στοιχεία καταστούν το MATLAB ένα εξαιρετικό εργαλείο στον τομέα της εκπαίδευσης και της έρευνας.

Η διαδικασία εκκίνησης του MATLAB είναι συνοπτικά η εξής. Αφού πρώτα εγκατασταθεί στον υπολογιστή και εκτελεστεί από τον χρήστη, εμφανίζεται το βασικό παράθυρο αλληλεπίδρασής του, το οποίο αποτελείται από τρία υπο-παράθυρα: το Command Window, το Current Folder Window και το Workspace

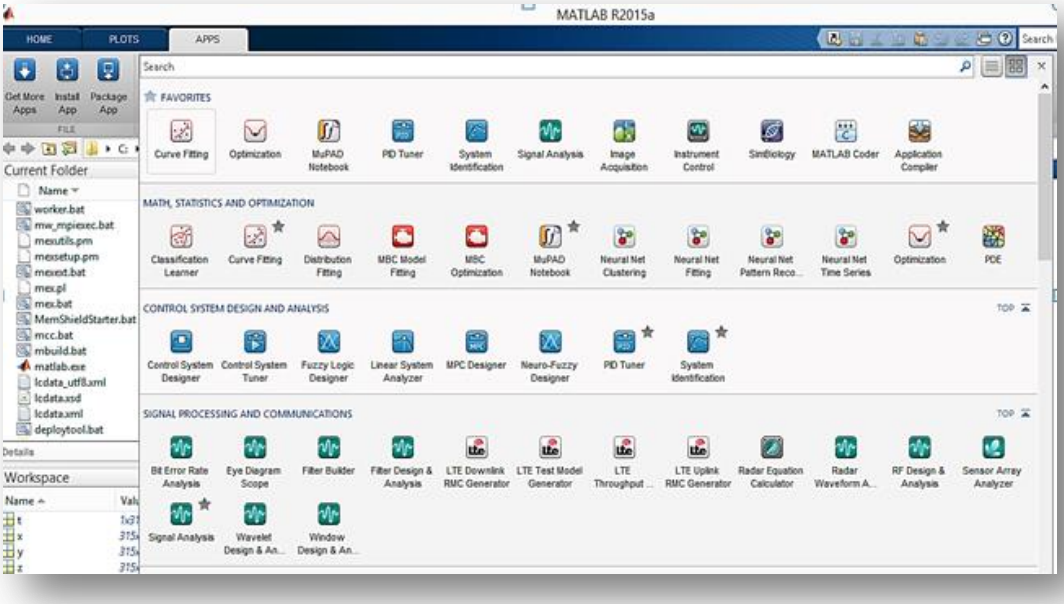
Window. Το Current Folder Window εμφανίζει όλα τα αρχεία του φακέλου στον οποίο εργαζόμαστε και το Workspace Window περιέχει πληροφορίες για όλες τις μεταβλητές της οποίες χρησιμοποιούμαι. Επιπλέον, το περιβάλλον του MATLAB περιλαμβάνει στο πάνω μέρος του τρείς καρτέλες εργαλειοθηκών. Οι καρτέλες αυτές είναι η HOME, η PLOTS και η APPS .



Εικόνα 1. MATLAB's Home

Η καρτέλα HOME αποτελεί την βασική καρτέλα-εργαλειοθήκη του MATLAB, περιέχει δηλαδή πληροφορίες, επιλογές και εργαλεία σχετικά με τα αρχεία, τις μεταβλητές, τον κώδικα, τον τρόπο προβολής του περιβάλλοντος, επιπρόσθετες πηγές, βιβλιοθήκες και άλλα. Η καρτέλα PLOTS αποτελεί ένα γρήγορο και εύκολο εργαλείο για την γραφική αναπαράσταση των μεταβλητών και των συναρτήσεων που χρησιμοποιούνται. Το παράθυρο εντολών (*Command Window*) αποτελεί το βασικό παράθυρο του προγράμματος. Χαρακτηριστικό γνώρισμα του είναι το σύμβολο “>>” το οποίο υποδεικνύει την γραμμή εκείνη στην οποία γράφουμε μία εντολή. Το παράθυρο εντολών αφενός δεν μπορεί να κλείσει μιας και αποτελεί το βασικό παράθυρο, αφετέρου μπορεί να είναι το μοναδικό ορατό παράθυρο προς τον χρήστη αν αυτός επιθυμεί να ελαχιστοποιήσει ή να κλείσει τα υπόλοιπα παράθυρα. Όλες οι εντολές του χρήστη γράφονται σε αυτό το παράθυρο.

Η καρτέλα APPS είναι ιδιαίτερα χρήσιμη για την διαδικασία εξόρυξης γνώσης. Μία εφαρμογή του MATLAB (*MATLAB app*) είναι ένα αυτόνομο πρόγραμμα με διεπαφή χρήστη, το οποίο αυτοματοποιεί μία εργασία ή έναν υπολογισμό. Το μόνο που πρέπει να κάνει ο χρήστης είναι να εισάγει τα δεδομένα τα οποία αφορούν την εργασία που θέλει να πραγματοποιήσει στο κατάλληλο πρόγραμμα. Τα APPS βρίσκονται στην τρίτη κεφαλίδα του MATLAB.



Εικόνα 2. MATLAB's APPS

Η εργαλειοθήκη Statistics and Machine Learning παρέχει συναρτήσεις και εφαρμογές για την περιγραφή, την ανάλυση και την μοντελοποίηση των δεδομένων. Η εφαρμογή Classification Learner αποτελεί κομμάτι της εργαλειοθήκης Statistics and Machine learning του MATLAB, και δίνει την δυνατότητα κατηγοριοποίησης δεδομένων και εκπαίδευσης μοντέλων με την βοήθεια της επιβλεπόμενης μάθησης. Η εφαρμογή καλύπτει μεγάλο εύρος των εργασιών της κατηγοριοποίησης όπως για παράδειγμα την επιλογή χαρακτηριστικών, την επιλογή μεθόδων κατηγοριοποίησης, την επιλογή τρόπου αξιολόγησης της ακρίβειας των μοντέλων καθώς και την εκπαίδευση και εξαγωγή ενός μοντέλου πρόβλεψης.

Η διαδικασία εισαγωγής των δεδομένων στην εφαρμογή είναι αρκετά απλή. Κατόπιν, δίνεται η δυνατότητα επιλογής των χαρακτηριστικών που θα στηριχθούμε για την πρόβλεψη και αυτών που θα θεωρήσουμε ως απόκριση. Στην συνέχεια, καλούμαστε να επιλέξουμε την μέθοδο (Validation Scheme) με την οποία θα εξετάζει την ακρίβεια του μοντέλου που χρησιμοποιούμε- εκπαιδεύουμε.

Μία μέθοδος κατηγοριοποίησης δεν έχει τα βέλτιστα αποτελέσματα για κάθε σύνολο δεδομένων. Η εύρεση του βέλτιστου κατηγοριοποιητή ενός συνόλου είναι μία διαδικασία συνεχών προσπαθειών και αποτυχιών. Το πακέτο Classification Learner App του MATLAB παρέχει ένα μεγάλο πλήθος κατηγοριοποιητών, το οποίο χωρίζεται σε τέσσερις τύπους μεθόδων κατηγοριοποίησης ως εξής:

- Decision Trees. Περιλαμβάνονται τρείς κατηγοριοποιητές (Complex Tree, Medium Tree, Simple Tree). Τα δέντρα απόφασης είναι εύκολα στην κατανόηση, με μικρή χρήση της μνήμης και παρέχουν γρήγορα αποτελέσματα τα οποία όμως μπορεί να έχουν χαμηλή προβλεπτική ακρίβεια. Η ανάπτυξη απλούστερων δέντρων οδηγεί σε αποφυγή του φαινομένου της υπερκάλυψης.
- Support Vector Machines. Περιλαμβάνονται έξι κατηγοριοποιητές (Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian, Medium Gaussian, Coarse Gaussian). Οι SVM's εφαρμόζονται σε δεδομένα τα οποία έχουν δύο οι περισσότερες κλάσεις και κατηγοριοποιούν τα δεδομένα βρίσκοντας ένα βέλτιστο υπερ-επίπεδο (hyperplane) το οποίο διαχωρίζει τα στοιχεία μίας κλάσης από τις υπόλοιπες. Οι απαιτήσεις σε χρόνο, χρήση της μνήμης και κατανόησης των αποτελεσμάτων διαφέρουν από κατηγοριοποιητή σε κατηγοριοποιητή.
- Nearest Neighbor Classifiers. Στην ομάδα αυτή περιλαμβάνονται έξι κατηγοριοποιητές (Fine KNN, Medium KNN, Coarse KNN, Cosine KNN, Cubic KNN, Weighted KNN). Τα μοντέλα που προκύπτουν συνήθως έχουν μεγάλη προβλεπτική ακρίβεια για δεδομένα μικρού μεγέθους με μικρό πλήθος διαστάσεων, μεγάλη χρήση της μνήμης και αποτελέσματα τα οποία είναι δύσκολα να ερμηνευτούν. Οι παραπάνω αλγόριθμοι στηρίζονται κυρίως στην ομοιότητα των χαρακτηριστικών για την κατηγοριοποίηση ενός νέου αντικειμένου.
- Ensemble Classifiers. Οι κατηγοριοποιητές της κατηγορίας αυτής είναι πέντε (Boosted Trees, Bagged Trees, Subspace Discriminant, Subspace KNN, RUSBoost Trees) και κάθε ένας απαρτίζετε από ένα σύνολο άλλων κατηγοριοποιητών. Τα αποτελέσματα ενός τέτοιου κατηγοριοποιητή προκύπτουν από την συγχώνευση των αποτελεσμάτων των μοντέλων που εκπαιδευτήκαν από το εκάστοτε σύνολο. Οι παραπάνω αλγόριθμοι είναι σχετικά αργοί εξαιτίας της εφαρμογής, εκπαίδευσης και συγχώνευσης πολλών μοντέλων και οι ποιότητα του αποτελέσματος εξαρτάται από την επιλογή του αλγορίθμου.

1.6 Εξόρυξη Δεδομένων και Οικονομία

Η αναζήτηση πληροφορίας και γνώσης αποτελούν κινητήριο μοχλό σε όλο το εύρος των διαδικασιών που απαντώνται στην Οικονομική Επιστήμη. Η διατήρηση τεράστιων βάσεων δεδομένων με πληροφορίες δυνητικά αξιοποιήσιμες και χρήσιμες για τη διερεύνηση οικονομικών ζητημάτων οδήγησε στην ανάγκη για την εφαρμογή των διαφόρων τεχνικών εξόρυξης γνώσης. Επομένως, η Εξόρυξη Δεδομένων βρίσκει εφαρμογή στην προσέγγιση οικονομικών ζητημάτων σε κρατικό επίπεδο, για παράδειγμα κατασκευάζοντας προβλέψεις σχετικά με την απασχόληση και την ανεργία, στον ακαδημαϊκό χώρο του οποίου άλλωστε αυτοσκοπό αποτελεί η δημιουργία γνώσης ανεξαρτήτως κλάδου και στις επιχειρήσεις.

Η επίτευξη των στόχων στις σύγχρονες επιχειρήσεις επιτάσσει την διαχείριση της γνώσης. Οι επιχειρήσεις που θα καταφέρουν νε το επιτύχουν σε μεγαλύτερο βαθμό θα μειώσουν τα κόστη, θα αυξήσουν την κερδοφορία, θα έχουν καλύτερη δυνατότητα προσαρμογής κ.λπ. και τελικά θα αποκομίσουν τα περισσότερα οφέλη. Επομένως, η εξόρυξη δεδομένων αποτελεί πολύτιμο εργαλείο για ενέργειες που σχετίζονται με το μάρκετινγκ και το μάνατζμεντ.

Ένας τρόπος με τον οποίο η Εξόρυξη Δεδομένων συμβάλει σε διαδικασίες επιχειρηματικής διαχείρισης, για να μειώσει τα κόστη παραδείγματος χάριν, βρίσκεται στην αρχή του κύκλου ζωής των προϊόντων, μέσω της έρευνας και της ανάπτυξης (Hongjun, et al., 2009). Όσον αφορά το μάρκετινγκ εντοπίζεται σε πλήθος ενεργειών όπως η προσέλκυση πελατών, η αναγνώρισης της καταναλωτικής συμπεριφοράς , η τμηματοποίηση της αγοράς κ.λπ. (Sabitha, Bhuvaneswari Amma, Annapoorani, & Balasubramanian, 2014).

Η εξόρυξη δεδομένων είναι ευρέως διαδεδομένη στον τραπεζικό κλάδο. Οικονομικά δεδομένα συλλέγονται κατά κύριο λόγο από τράπεζες και συνήθως πρόκειται για αξιόπιστα, ολοκληρωμένα και υψηλής ποιότητας δεδομένα. Στον κλάδο αυτό οι τεχνικές Εξόρυξης Δεδομένων ενδείκνυνται για τη διαχείριση του ρίσκου, τον εντοπισμό απάτης πιστωτικών καρτών (Chan, Fan, Prodromidis, & Stolfo, 1999), το ξέπλυμα χρήματος, τις προωθητικές ενέργειες.

Εξέχουσας σημασίας είναι οι συγκεκριμένες τεχνικές και σε θέματα που υπάγονται στον τομέα της λογιστικής και της ελεγκτικής. Δύο από τα σημαντικότερα προβλήματα που εντοπίζονται σε αυτόν τον τομέα είναι η πρόβλεψη επερχόμενης χρεοκοπίας (Olson, Delen, & Meng, 2012) και ο εντοπισμός παραποιημένων

χρηματοοικονομικών καταστάσεων (Kotsiantis, Koumanakos, Tzelepis, & Tampakas, 2006).

Τέλος, αξίζει να σημειωθεί ότι τεχνικές Εξόρυξης Δεδομένων εφαρμόζονται στον κλάδο των ασφαλίσεων (Devale & Kulkarni, 2012), του τουρισμού (Magnini, Honeycutt, & Hodge, 2003) και στο χρηματιστήριο (Barak & Modarres, 2015).

Κεφάλαιο 2^{ο} Κατηγοριοποίηση

Όπως ήδη αναφέρθηκε, η κατηγοριοποίηση αποτελεί μια από τις πιο βασικές εργασίες εξόρυξης γνώσης. Σκοπός της είναι η ανάθεση ενός στοιχείου σε ένα σύνολο κατηγοριών που έχει καθοριστεί ήδη. Έχει μελετηθεί εκτενώς στην μηχανική μάθηση, την αναγνώριση προτύπων και τη στατιστική.

Η διαδικασία χαρακτηρίζεται από ένα καθορισμένο σύνολο κατηγοριών και ένα σύνολο από προ-κατηγοριοποιημένα παραδείγματα. Ειδικότερα, περιλαμβάνει δύο βασικά βήματα την Εκμάθηση (*Learning*) και την Κατηγοριοποίηση (*Classification*). Κατά την Εκμάθηση τα δεδομένα εκπαίδευσης (*training data*) αναλύονται από έναν αλγόριθμο κατηγοριοποίησης για να κατασκευάσουν ένα μοντέλο που θα περιγράφει ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων. Τα στοιχεία του συνόλου εκπαίδευσης επιλέγονται τυχαία από ένα πληθυσμό δεδομένων και ανήκουν σε μια από τις προκαθορισμένες κατηγορίες. Στο επόμενο βήμα, αυτό της κατηγοριοποίησης, γίνεται ο υπολογισμός της ακρίβειας του μοντέλου με τη χρήση των δοκιμαστικών δεδομένων (*test data*). Το μοντέλο που προέκυψε κατηγοριοποιεί τα δοκιμαστικά παραδείγματα και κατόπιν τα δοκιμαστικά δεδομένα συγκρίνονται με τα δεδομένα που προέβλεψε το μοντέλο. Εάν η ακρίβεια του μοντέλου είναι αποδεκτή, το μοντέλο είναι κατάλληλο για την κατηγοριοποίηση μελλοντικών δεδομένων που δεν έχουν κατηγοριοποιηθεί.

2.1 Bayesian κατηγοριοποίηση

Βασίζεται στη στατιστική θεωρία κατηγοριοποίησης του Bayes. Θεωρούμε ένα δείγμα X το οποίο πρέπει να κατηγοριοποιηθεί σε μια από τις κατηγορίες C_1, C_2,…, C_n. Κάθε κατηγορία χαρακτηρίζεται από μια εκ των προτέρων πιθανότητα παρατήρησης της κλάσης C_i και το δείγμα X ανήκει σε μια κλάση C_i σύμφωνα με την υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας $p(X/C_i)$. Κατόπιν καθορίζουμε την εκ των υστέρων πιθανότητα $p(c_i/X)$ σύμφωνα με τον τύπο:

$$p(c_i/X) = \frac{p(ci/X)p(ci)}{p(X)}$$

Οι πιο γνωστοί Bayesian κατηγοριοποιητές είναι ο Naïve Bayesian κατηγοριοποιητής και τα Bayesian Belief Networks. Παρόλο που στη θεωρία, οι κατηγοριοποιητές

αυτοί έχουν ελάχιστο ποσοστό σφάλματος, στην πράξη αυτό δεν ισχύσει λόγω των υποθέσεων που είναι απαιτούμενες κατά τη χρήση τους.

2.1.1 Naïve Bayesian

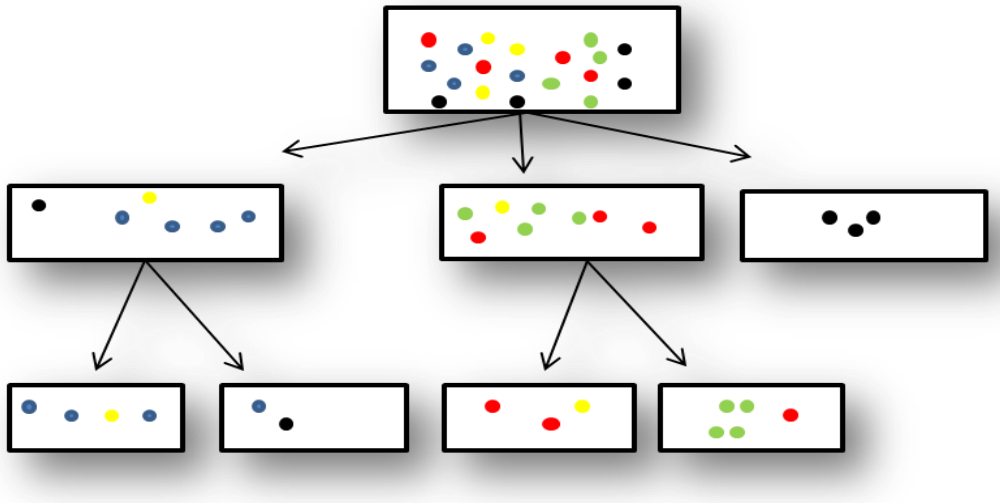
Ο συγκεκριμένος κατηγοριοποιητής μπορεί να προβλέψει ότι οι μετρήσεις που γίνονται στο δείγμα ανήκουν στην κατηγορία που έχει την μεγαλύτερη εκ των υστέρων πιθανότητα. Επομένως, στόχος είναι η μεγιστοποίηση της εκ των υστέρων υπόθεσης.

2.1.2 Bayesian Belief Networks

Τα Bayesian belief Networks λαμβάνουν υπόψη τις εξαρτήσεις που υπάρχουν μεταξύ των μεταβλητών. καθορίζονται από δύο στοιχεία, έναν κατευθυνόμενο ακυκλικό γράφο κι έναν πίνακα υπό συνθήκη πιθανότητας. Στον κατευθυνόμενο ακυκλικό γράφο κάθε κόμβος αντιπροσωπεύει μια τυχαία μεταβλητή και κάθε τόξο μια εξάρτηση πιθανοτήτων. Ένας κόμβος μπορεί να είναι είτε γονέας είτε απόγονος ανάλογα με τη θέση του και οι μεταβλητές είναι ανεξάρτητες από τους μη προγόνους τους. Ένας από τους κόμβους μπορεί να επιλεγεί ως έξοδος αντιπροσωπεύοντας τα γνωρίσματα μιας κατηγορίας.

2.2 Δέντρα απόφασης

Πρόκειται για τα δημοφιλέστερα μοντέλα κατηγοριοποίησης. Η κατασκευή του βασίζεται σε ένα σύνολο δεδομένων εκπαίδευσης που είναι ήδη κατηγοριοποιημένα. Οι κόμβοι και τα κλαδιά κάθε κόμβου προσδιορίζουν τον έλεγχο ενός γνωρίσματος και την πιθανή τιμή αυτού αντίστοιχα. Κάθε φύλλο του δέντρου αντικατοπτρίζει τις κατηγορίες που έχουν οριστεί. Η διαδικασία κατηγοριοποίησης ενός νέου δείγματος με βάση ένα δέντρο απόφασης, ξεκινά από τη ρίζα του δέντρου. Εξετάζοντας τα γνωρίσματα που καθορίζονται από αυτό τον κόμβο προσδιορίζονται στη συνέχεια οι εσωτερικοί κόμβοι, σε καθέναν από τους οποίους διενεργείται έλεγχος ικανοποίησης του δείγματος. Ο έλεγχος αυτός προσδιορίζει το κλαδί που θα διασχίσουμε στην συνέχεια καθώς και τον επόμενο κόμβο. Ο τελικός κόμβος είναι το φύλλο, που δείχνει σε ποια κατηγορία ανήκει το δείγμα (Χαλκίδη & Βαζιργιάννης, 2005).



Σχήμα 2. Δέντρο απόφασης

2.2.1 ID3

Ένας από τους βασικούς αλγορίθμους κατηγοριοποίησης θεωρείται ο αλγόριθμος ID3 με βάση τον οποίο μπορεί να κατασκευαστεί ένα δέντρο απόφασης. Η ολοκλήρωση της διαδικασίας απαιτεί συγκεκριμένα βήματα. Αρχικά, το δέντρο ξεκινάει με ένα κόμβο που αντιπροσωπεύει ολόκληρο το σύνολο των δεδομένων. Αν τα δείγματα εκπαίδευσης είναι της ίδιας κατηγορίας τότε ο κόμβος γίνεται φύλλο και προστίθεται η ετικέτα της κατηγορίας. Για τον διαχωρισμό των δειγμάτων στις κατηγορίες ο αλγόριθμος χρησιμοποιεί ένα μέτρο εντροπίας. Στη συνέχεια ο αλγόριθμος υπολογίζει το κέρδος πληροφορίας για κάθε γνώρισμα , ώστε να γίνει επιλογή των γνωρισμάτων που διαχωρίζουν καλύτερα τα δείγματα σε διαφορετικές κατηγορίες. Σαν γνώρισμα ελέγχου επιλέγετε αυτό με το μεγαλύτερο κέρδος πληροφορίας. Για το γνώρισμα ελέγχου δημιουργείται ένας κόμβος όσο δημιουργούνται κλαδιά για κάθε τιμή του. Το δείγμα δεδομένων διαχωρίζεται αναλόγως και ο αλγόριθμος εφαρμόζεται συνεχώς για τη μορφοποίηση ενός δέντρου. Η επαναλαμβανόμενη αυτή διαδικασία σταματάει όταν όλα τα δείγματα ανήκουν στην ίδια κατηγορία, ή αν δεν υπάρχουν άλλα γνωρίσματα για περαιτέρω διαχωρισμό του δείγματος, ή αν δεν υπάρχουν δείγματα που δεν έχουν κατηγοριοποιηθεί για το κλαδί ελέγχου. (Χαλκίδη & Βαζιργιάννης, 2005)

2.2.2 C4.5

Ο αλγόριθμος C4.5 αποτελεί ένα ακόμη αλγόριθμο για την κατασκευή δέντρων αποφάσεων. Η μέθοδος για την κατασκευή του δέντρου είναι σχετικά πιο απλή. Η σωστή λειτουργία του απαιτεί τη χρήση ολοκληρωμένων δεδομένων. Τα γνωρίσματα κάθε κόμβου του δέντρου είναι εφικτό να έχουν συνεχείς τιμές. Ωστόσο, ο αλγόριθμος αυτός κρίνεται ακατάλληλος για μεγάλα σύνολα δεδομένων, δεδομένου ότι παρουσιάζει πολύ μικρή ακρίβεια.

2.2.3 CART

Ο αλγόριθμος CART (Classification and Regression Trees), όπως γίνεται φανερό από το όνομά του, χρησιμοποιείται τόσο στην εφαρμογή μεθόδων ταξινόμησης όσο και στην εφαρμογή μεθόδων παλινδρόμησης. Ο αλγόριθμος αυτός απαντάται συχνά στην επίλυση ζητημάτων του πεδίου των οικονομικών. Πρόκειται για τη δημιουργία δέντρων απόφασης τόσο για συνεχείς όσο και για διακριτές τιμές. Το δέντρο αναπτύσσεται χωρίς να ανακόπτεται η διαδικασία. Όταν φτάσει στο μέγιστο σημείο, η διαδικασία κινείται προς τα πίσω διαιρώντας τα φύλλα και αφαιρώντας αυτά που συμβάλλουν λιγότερο στη διαδικασία εκπαίδευσης των δεδομένων. Ο μηχανισμός αυτός έχει σαν σκοπό την δημιουργία πολλών δέντρων. Το βέλτιστο δέντρο επιλέγεται κατόπιν διαδικασίας αξιολόγησης της απόδοσης.

2.2.4 SLIQ

Με βάση αυτή την προσέγγιση εφαρμόζεται ένα αρχικό στάδιο κατηγοριοποίησης των γνωρισμάτων. Αρχικά ορίζεται ένας κόμβος ως ρίζα του δέντρου. Για τον υπολογισμό του καλύτερου δυνατού διαχωρισμού γίνεται χρήσης της λίστας κατηγοριών. Αφού διαχωριστεί ένας κόμβος , οι είσοδοι της λίστας κατηγοριών τροποποιούνται για να υποδείξουν την λίστα στην οποία ανήκει η εγγραφή. Με τον αλγόριθμο αυτό γίνεται συχνή προσπέλαση των κατηγοριών με τυχαίο τρόπο και από τις δύο φάσεις της επαγωγής του δέντρου, οπότε πρέπει να βρίσκεται συνεχώς στη μνήμη για επίτευξη καλής απόδοσης. Το γεγονός αυτό περιορίζει το μέγεθος του συνόλου εκπαίδευσης.

2.2.5 SPRINT

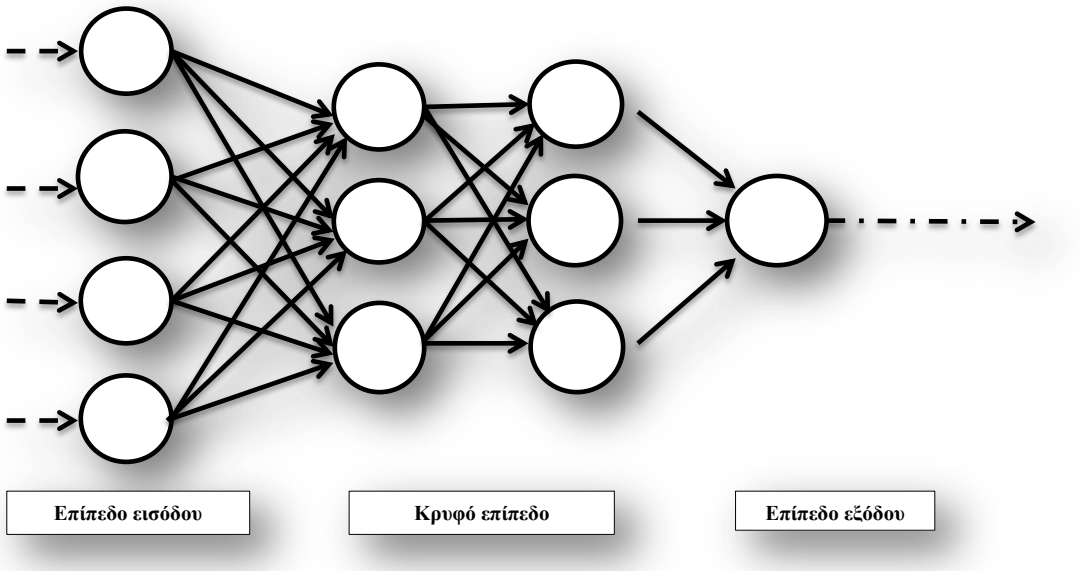
Ο αλγόριθμος αυτός είναι γρήγορος και μπορεί να χρησιμοποιηθεί σε οποιοδήποτε όγκο δεδομένων. Σε αντίθεση με άλλους αλγορίθμους, όπως ο SLIQ, δεν απαιτεί την διατήρηση στη μνήμη όλων των δεδομένων. Επιτρέπει τη χρήση πολλών επεξεργαστών για την δημιουργία ενός συνεπούς μοντέλου. Σύμφωνα με τον SPRINT, το σύνολο των γνωρισμάτων προ-κατηγοριοποιούνται και η κατηγοριοποίηση αυτή διατηρείται σε όλη τη διάρκεια του διαχωρισμού. Όσο το δέντρο αυξάνεται τα γνωρίσματα διαχωρίζονται μεταξύ των κόμβων.

2.3 Νευρωνικά Δίκτυα

Τα Νευρωνικά Δίκτυα αποτελούν μια πολύ συνηθισμένη προσέγγιση κατηγοριοποίησης. Αποτελούνται από «νευρώνες» που βασίζονται στη δομή των νευρώνων του εγκεφάλου. Επεξεργάζονται ένα στοιχείο κάθε φορά και η διαδικασία της μάθησης πραγματοποιείται συγκρίνοντας την κατηγοριοποίηση τους για μια αρχική εγγραφή, συνήθως αυθαίρετη, με την πραγματική κατηγοριοποίηση της εγγραφής που είναι γνωστή. Η διαδικασία είναι επαναληπτική καθώς τα λάθη από την κατηγοριοποίηση της αρχικής εγγραφής επανατροφοδοτούνται στο δίκτυο ώστε να χρησιμοποιηθούν για την τροποποίηση του αλγορίθμου τη δεύτερη φορά. Πιο συγκεκριμένα η διαδικασία περιλαμβάνει τα εξής βήματα:

- Αναγνώριση των χαρακτηριστικών εισόδου (*inputs*) και εξόδου (*outputs*).
- Κατασκευή δικτύου με κατάλληλη τοπολογία.
- Επιλογή του σωστού συνόλου εκπαίδευσης.
- Εκπαίδευση του δικτύου με βάση ένα αντιπροσωπευτικό σύνολο δεδομένων.
- Έλεγχος του δικτύου χρησιμοποιώντας ένα σύνολο ελέγχου ανεξάρτητο από το σύνολο εκπαίδευσης.

Το μοντέλο που παράγεται από την παραπάνω διαδικασία χρησιμοποιείται για να προβλέψει τις κατηγορίες (έξοδοι) των μη κατηγοριοποιημένων δεδομένων (είσοδοι).



Σχήμα 3. Νευρωνικό Δίκτυο

Το σχήμα 1 δείχνει τα επίπεδα οργάνωσης των νευρώνων. Το επίπεδο εισόδου (input layer) συνίσταται από τιμές ενός στοιχείου του συνόλου δεδομένων οι οποίες αποτελούν τις εισαγωγές στο επόμενο επίπεδο. Το επόμενο επίπεδο ονομάζεται κρυμμένο επίπεδο (hidden layer) και είναι εφικτό να υπάρξουν πολλά. Στο τελευταίο επίπεδο, το επίπεδο εξόδου (output layer), υπάρχει ένας κόμβος για κάθε κατηγορία. (Χαλκίδη & Βαζιργιάννης, 2005)

2.4 Η τεχνική των Κοντινότερων Γειτόνων

Η μέθοδος του κ κοντινότερου γείτονα Είναι μια απλή προσέγγιση κατηγοριοποίησης. Με τη μέθοδο αυτή τα στοιχεία κατηγοριοποιούνται χρησιμοποιώντας την πλειοψηφία μεταξύ των κατηγοριών από τα παραδείγματα που είναι τα πιο κοντινά σ αυτό που επιθυμούμαι να κατηγοριοποιηθεί. Η μέθοδος παράγει συνεχείς και επικαλυπτόμενες γειτονιές .Η τεχνική αυτή έχει μειονεκτήματα σε χώρους υψηλών διαστάσεων.

2.5 Support Vector Machines

Οι Support Vector Machines (SVM) έχουν πολλές εφαρμογές. Για να κατανοήσουμε τη διαδικασία, θεωρούμε ότι έχουμε n παρατηρήσεις. Κάθε παρατήρηση αποτελείται από ένα ζευγάρι x_i και μια ετικέτα της συσχετιζόμενης κατηγορίας y_j. Υποτίθεται ότι

υπάρχει κάποια άγνωστη κατανομή πιθανότητας με βάση την οποία τα στοιχεί παράγονται. Σκοπός είναι να μαθευτεί το σύνολο παραμέτρων μιας συνάρτησης έτσι ώστε, η συνάρτηση, να πραγματοποιεί την αντιστοίχιση $x_i \rightarrow y_i$. Μια συγκεκριμένη επιλογή συνόλου παραμέτρων καθορίζει μοναδικά την αντίστοιχη μηχανή εκπαίδευσης.

Οι SVM έχουν σαν στόχο την ελαχιστοποίηση του ανωτέρου ορίου του σφάλματος γενίκευσης. Επομένως, για την επίτευξη του στόχου, καταφεύγει στην εκμάθηση του συνόλου παραμέτρων της συνάρτησης ώστε να ικανοποιείται η ιδιότητα του μέγιστου περιθωρίου από τη μηχανή εκπαίδευσης.

Μια τέτοια μηχανή αντιστοιχεί τα δεδομένα σε ένα χώρο υψηλών διαστάσεων επιφέροντας υπολογιστικό και θεωρητικό κόστος μάθησης. Γι την αποφυγή της υπερκάλυψης, ανάμεσα σε πολλά υπερ-επίπεδα που μπορούν να διαχωρίσουν τα δεδομένα στο χώρο των χαρακτηριστικών , διαλέγει το μέγιστο υπερ-επίπεδο περιθωρίου. Η χρήση του μέγιστου επιπέδου οδηγεί σε έναν αλγόριθμο που μπορεί να μειωθεί σε ένα πρόβλημα βελτιστοποίησης. Προκειμένου να εκπαιδευτεί η μηχανή πρέπει να βρεθεί το μοναδικό ελάχιστο μιας συνάρτησης. Κατά συνέπεια οι μηχανές δεν έχουν το τοπικό πρόβλημα ελαχίστων που έχει επιπτώσεις σε πολλά σχήματα εκπαίδευσης.

2.6 Ασαφής κατηγοριοποίηση

Η αβεβαιότητα είναι ένα σημαντικό στοιχείο που συναντάται στην ανακάλυψη γνώσης. Παρόλα αυτά, δεν λαμβάνεται υπόψη σε καμία από τις προαναφερόμενες μεθόδους κατηγοριοποίησης. Οι κλασικές αυτές προσεγγίσεις είναι αυστηρές θα λέγαμε διότι ταξινομούν ένα αντικείμενο είτε σε μία κατηγορία είτε όχι.

Η εξαγωγή ασαφών κανόνων είναι μια προσπάθεια κατηγοριοποίησης προκειμένου να αναγνωριστεί κάθε κατηγορία δεδομένων και να αποφευχθεί η αβεβαιότητα. Σύμφωνα με τον Chiu (1997), η διαδικασία αυτή επιτρέπει τη διαμόρφωση σχέσεων στα δεδομένα σύμφωνα με κανόνες "if-then" που είναι κατανοητοί, επαληθεύσιμοι και επεκτείνονται εύκολα. Οι μέθοδοι εξόρυξης βασίζονται στην δημιουργία ομάδων στα δεδομένα. Κάθε ομάδα που αποκτάται αντιστοιχεί σε έναν ασαφή κανόνα ο οποίος συσχετίζει μια περιοχή στον χώρο εισόδου σε μια περιοχή στον χώρο εξόδου. Πιο απλά, το σύστημα χρησιμοποιεί ασαφή λογική για να καθορίσει την καλύτερη κατηγορία μέσα στην οποία μπορούν τα δεδομένα να κατηγοριοποιηθούν.

Κεφάλαιο 3^{ο} Συσταδοποίηση

Η συσταδοποίηση είναι μια διαδικασία σύμφωνα με την οποία δύνανται να ανακαλυφθούν συστάδες (*clusters*) και πρότυπα *(patterns)* τα οποία εμφανίζουν ενδιαφέρον και μπορούν να οδηγήσουν σε νέες πληροφορίες. Με τον όρο συστάδα αναφερόμαστε σε μία συλλογή αντικειμένων από τα δεδομένα, με βάση την ομοιότητα που παρουσιάζουν τα χαρακτηριστικά τους. Τα αντικείμενα μιας συστάδας συνήθως έχουν παρόμοια συμπεριφορά μεταξύ τους. Ουσιαστικά, πρόκειται για εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων.

Στην μηχανική μάθηση και την αναγνώριση προτύπων η διαδικασία μπορεί να βρεθεί με το όνομα μη εποπτευόμενη μάθηση. Ο όρος προκύπτει απ' το γεγονός ότι στην συσταδοποίηση οι κατηγορίες δεν είναι προκαθορισμένες ούτε υπάρχει κάποιο παράδειγμα που να υποδεικνύει ποιες από τις επιθυμητές σχέσεις μεταξύ των δεδομένων θα ήταν σωστές. Δεδομένου ότι, η κατηγοριοποίηση είναι η διαδικασία ανάθεσης ενός στοιχείου σε μια προκαθορισμένη κατηγορία, από την άλλη πλευρά η συσταδοποίηση παράγει τις αρχικές κατηγορίες στις οποίες οι τιμές ενός συνόλου δεδομένων μπορούν να ταξινομηθούν κατά την κατηγοριοποίηση.

Στα σύνολα δεδομένων είναι σύνηθες να επιλέγουμε ομάδες παρόμοιων χαρακτηριστικών. Η ομοιότητα των χαρακτηριστικών καθορίζεται πλήρως από τον σκοπό της εργασίας μας. Ένα τρόπος για να εκφράσουμε την ομοιότητα αποτελεί η χρήση μιας σειράς αποστάσεων μεταξύ ζευγών αντικειμένων. Ένα από τα βασικά συστατικά για την εκτέλεση μίας επιτυχημένης συσταδοποίησης είναι η ύπαρξη ποσοτήτων για κάθε αντικείμενο. Μέσω των ποσοτήτων αυτών και με την βοήθεια συναρτήσεων μέτρησης απόστασης μπορούμε να κατανοήσουμε αν δύο αντικείμενα είναι όμοια ή ανόμοια μεταξύ τους. Όπως γίνετε αντιληπτό οι ποσότητες αυτές για παρατηρήσεις που μοιάζουν μεταξύ τους, θα πρέπει να δίνουν πολύ μικρή τιμή στην απόσταση. Οι πιο γνωστοί τύποι απόστασης που χρησιμοποιούνται φαίνονται παρακάτω:

- Ευκλείδεια απόσταση, όπου i=1,2,…m, $d_{j,k} = \sqrt{\sum\left(X_{ij} - X_{ik}\right)^2}$
- Απόσταση Manhattan, όπου i=1,2,…m, $d_{j,k} = \sum\left|X_{ij} - X_{ik}\right|$
- Απόσταση Minkowski, όπου i=1,2,…m, $d_{j,k} = \sum\left(\left|X_{ij} - X_{ik}\right|^p\right)^{\frac{1}{p}}$

3.1 Διαδικασία συσταδοποίησης

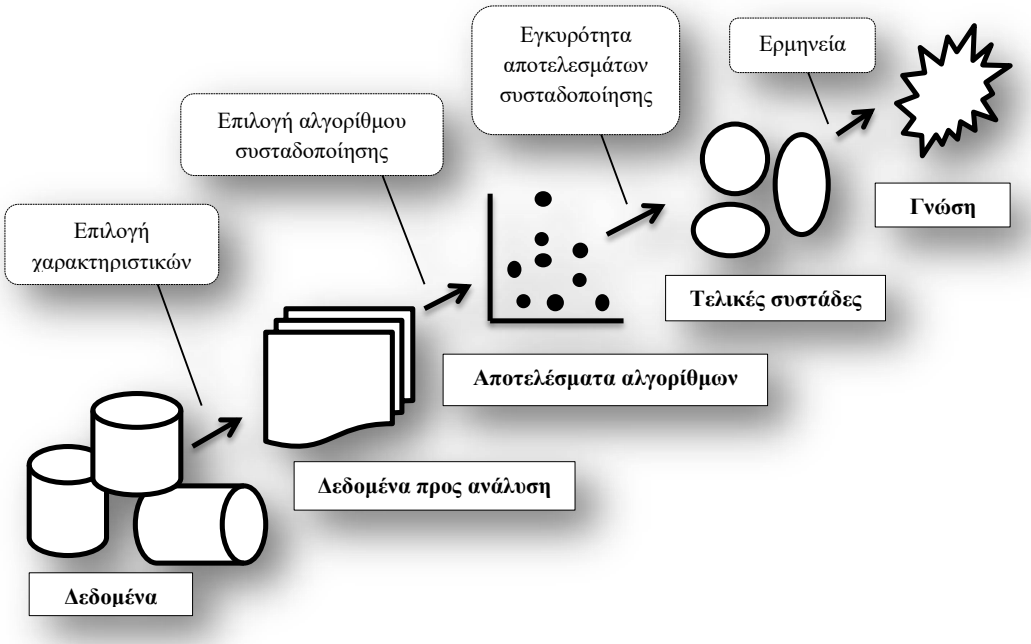
Ανάλογα με το κριτήριο που χρησιμοποιείται για την συσταδοποίηση, η διαδικασία μπορεί να οδηγήσει στην δημιουργία διαφόρων συστάδων για ένα συγκεκριμένο σύνολο δεδομένων. Για τον λόγο αυτό πριν την εφαρμογή της διαδικασίας είναι απαραίτητη η προ-επεξεργασία των δεδομένων. Τα βασικά βήματα της διαδικασίας της συσταδοποίησης αναλύονται ως εξής:

Επιλογή χαρακτηριστικών γνωρισμάτων. Για την εξαγωγή όσο το δυνατόν περισσότερων πληροφοριών είναι απαραίτητο να γίνει επιλογή των κατάλληλων γνωρισμάτων στα οποία θα εφαρμοστεί συσταδοποίηση.

Αλγόριθμος συσταδοποίησης. Αφορά την επιλογή του κατάλληλου αλγορίθμου που θα οδηγήσει σε ένα καλό σχήμα συσταδοποίησης, με βάση το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης.

Επικύρωση αποτελεσμάτων. Με την χρήση των κατάλληλων τεχνικών και κριτηρίων γίνεται η εξακρίβωση της ορθότητας των αποτελεσμάτων

Ερμηνεία αποτελεσμάτων. Η ερμηνεία των αποτελεσμάτων της συσταδοποίησης σε πολλές περιπτώσεις απαιτεί την ενσωμάτωση και άλλων στοιχείων για να καταλήξουμε στο σωστό συμπέρασμα.



Σχήμα 4. Διαδικασία Συσταδοποίησης

3.2 Απαιτήσεις ανάλυσης συστάδων

Η συσταδοποίηση αποτελεί ένα απαιτητικό πεδίο έρευνας. Παρακάτω αναπτύσσονται οι απαιτήσεις για την συσταδοποίηση ως εργαλείο εξόρυξης δεδομένων.

Επεκτασιμότητα. Οι περισσότεροι αλγόριθμοι ανάλυσης συστάδων είναι αποτελεσματικοί σε μικρά σύνολα δεδομένων (μερικές εκατοντάδες αντικείμενα). Ωστόσο, μία βάση μπορεί να περιέχει χιλιάδες ή ακόμη και εκατομμύρια εγγραφές. Η εφαρμογή ενός αλγορίθμου συσταδοποίησης σε ένα μεγάλο σύνολο δεδομένων μπορεί να οδηγήσει σε μεροληπτικά αποτελέσματα. Για τον λόγο αυτό οι αλγόριθμοι πρέπει να χαρακτηρίζονται από εξελιξιμότητα ώστε να μπορούν να είναι αποτελεσματικοί σε όλο και μεγαλύτερα σύνολα.

Διαχείριση διαφορετικού τύπου μεταβλητών. Πολλοί αλγόριθμοι είναι σχεδιασμένοι προκειμένου να χειρίζονται αριθμητικά δεδομένα. Δεν περιέχουν, όμως, όλα τα σύνολα αριθμητικά δεδομένα. Μπορεί να απαιτείται συσταδοποίηση σε σύνολα που περιέχουν κατηγορικά δεδομένα, δυαδικά δεδομένα, γραφήματα, εικόνες, πιθανότητες και έγγραφα ως δεδομένα.

Σχήμα συστάδας. Ο καθορισμός των συστάδων στους περισσότερους αλγορίθμους βασίζεται στην Ευκλείδεια απόσταση ή στην «απόσταση Manhattan». Ωστόσο, συνήθως, αυτά τα μέτρα απόστασης οδηγούν σε συστάδες όμοιες, σφαιρικού σχήματος και παρόμοιας πυκνότητας, το οποίο δεν είναι πάντα επιθυμητό, διότι αν σκεφτεί κανείς ότι μια συστάδα μπορεί να έχει την μορφή οποιουδήποτε αυθαίρετου σχήματος.

Εισαγωγή παραμέτρων: Είναι πιθανό, ένας αλγόριθμος συσταδοποίησης να απαιτεί την εισαγωγή κάποιων αρχικών παραμέτρων (π.χ. επιθυμητός αριθμός συστάδων που θέλουμε να δημιουργηθούν). Ωστόσο Όμως, οι προαπαιτούμενες παράμετροι εισόδου συχνά επηρεάζουν τα αποτελέσματα των αλγορίθμων ή δεν κατανοείται πλήρως η σημασία τους από των αλγόριθμο (π.χ. πολυδιάστατα αντικείμενα) γι αυτό και θα ήταν θεμιτό να αποφεύγονται.

Χειρισμός θορύβου: Τα περισσότερα «πραγματικά» δεδομένα (δηλαδή εννοώντας δεδομένα που προέρχονται απορροφάμε από τον πραγματικό κόσμο -περιβάλλον) περιέχουν έκτροπες, λανθασμένες ή άγνωστες τιμές. Αυτά τα δεδομένα χαρακτηρίζονται ως «δεδομένα με θόρυβο». Οι αλγόριθμοι συσταδοποίησης είναι «ευαίσθητοι» στον θόρυβο με αποτέλεσμα να παράγουν κακής ποιότητας συστάδες.

Συνεπώς, θέλουμε την δημιουργία μεθόδων συσταδοποίησης που μπορούν να διαχειριστούν δεδομένα με θόρυβο.

3.3 Μέθοδοι Συσταδοποίησης

Στην βιβλιογραφία εντοπίζονται πλήθος μεθόδων και αλγορίθμων συσταδοποίησης. Οι αλγόριθμοι αυτοί μπορούν να ταξινομηθούν σύμφωνα με, (i) τον τύπο δεδομένων που εισάγονται στον αλγόριθμο, (ii) τη μέθοδο που καθορίζει την συσταδοποίηση του συνόλου των δεδομένων και, τέλος, (iii) τις θεωρίες και τις θεμελιώδεις έννοιες στις οποίες είναι βασισμένες οι τεχνικές ανάλυσης συστάδων.

Στα πλαίσια της παρούσας εργασίας παρουσιάζονται οι βασικότεροι αλγόριθμοι συσταδοποίησης ταξινομημένοι σύμφωνα με την μέθοδο καθορισμού των συστάδων.

Ειδικότερα, μπορούν να ταξινομηθούν σε αλγορίθμους Διαιρετικής Συσταδοποίησης (*Partitional Clustering*), Ασαφής Συσταδοποίησης (*Fuzzy Clustering*), Μη Ασαφής Συσταδοποίησης (*Crisp Clustering*), Συσταδοποίησης βασισμένης στα Δίκτυα Kohonen (*Kohonen Net Clustering*), Ιεραρχικής Συσταδοποίησης (*Hierarchical Clustering*), Συσταδοποίησης βασισμένη στην πυκνότητα (*Density-based Clustering*) Συσταδοποίησης βασισμένη σε πλέγμα (*Grid-based Clustering*) και Συσταδοποίησης υποχώρων (*Subspace Clustering*).

3.3.1 Διαιρετική συσταδοποίηση.

Δεδομένου ενός συνόλου n αντικειμένων, μια διαιρετική μέθοδος δημιουργεί k διαμερίσεις του συνόλου, όπου κάθε διαμέριση ονομάζεται συστάδα με $k \leq n$. Βασίζετε δηλαδή στην αποσύνθεση του σύνολο δεδομένων σε k μη συσχετιζόμενες ομάδες, όπου κάθε ομάδα περιέχει ένα τουλάχιστον αντικείμενο. Οι περισσότεροι διαιρετικοί μέθοδοι είναι βασισμένοι σε μέτρα απόστασης ή αλλιώς μέτρα ομοιότητας και έχουν ως γενικό σκοπό την ελαχιστοποίηση των μέτρων ομοιότητας μεταξύ των δειγμάτων ή την μεγιστοποίηση των μέτρων ανομοιότητας μεταξύ των διαφορετικών συστάδων.

3.3.1.1 Αλγόριθμος k-MEANS

Ένας από τους συχνότερα χρησιμοποιούμενους αλγορίθμους συσταδοποίησης είναι ο αλγόριθμος k-MEANS. Ο συγκεκριμένος αλγόριθμος βασίζεται στην αποσύνθεση του συνόλου των δεδομένων για τη δημιουργία ασυσχέτιστων συστάδων και έχει σαν στόχο την ελαχιστοποίηση της μέσης τετραγωνικής απόστασης των δεδομένων από

τα πιο κοντινά κέντρα των συστάδων. Αναλυτικότερα, ο αλγόριθμος ξεκινά ορίζοντας τα αρχικά κέντρα των συστάδων, στη συνέχεια αναθέτει κάθε στοιχείο του συνόλου των δεδομένων στη συστάδα με το πλησιέστερο κέντρο και υπολογίζει ξανά τα κέντρα. Η διαδικασία αυτή, όπως γίνεται αντιληπτό, είναι επαναλαμβανόμενη και σταματά όταν τα κέντρα σταθεροποιηθούν.

Ο k-MEANS αποτελεί μια ευρέως αποδεκτή τεχνική συσταδοποίησης, η οποία χρησιμοποιείται αποτελεσματικά για συσταδοποίηση διαφόρων συνόλων σε διάφορα πεδία. Ωστόσο, υπάρχουν διάφορες εκδόσεις και παραλλαγές του k-MEANS. Οι παραλλαγές αυτές διαφέρουν κυρίως (i) στον τρόπο επιλογής των αρχικών k κέντρων των συστάδων, (ii) στον υπολογισμό της ομοιότητας, (iii) στον τρόπο υπολογισμού των νέων κέντρων συστάδων. Ορισμένες χαρακτηριστικές παραλλαγές του k-MEANS είναι: ο *ISODATA,* ο *Fuzzy C-Means*, ο *Sas Proc Fastclus*.

Η μέθοδος k-MEANS είναι εύκολη στην υλοποίηση και στην κατανόηση. Όμως υπάρχουν ορισμένοι περιορισμοί που πρέπει να ληφθούν υπόψη. Πιο συγκεκριμένα, η χρήση του αλγορίθμου k-MEANS περιορίζεται σε αριθμητικά δεδομένα. Για τον λόγο αυτό, εάν στο σύνολο δεδομένων μας παρουσιάζεται κατηγορικό χαρακτηριστικό τότε πρέπει είτε να το αγνοήσουμε είτε να μετατρέψουμε την τιμή του χαρακτηριστικού σε αριθμητική είτε να χρησιμοποιήσουμε έναν άλλον αλγόριθμο. Επίσης, ο k-MEANS είναι ευαίσθητος στην παρουσία έκτροπων παρατηρήσεων (*outliers*), αν αναλογιστούμε την συμβολή τους στον υπολογισμό της μέσης τιμής-κέντρα, καθώς και σε συστάδες που έχουν διαφορετικά μεγέθη. Τέλος, ο συγκεκριμένος αλγόριθμος, όπως και κάθε διαιρετικός αλγόριθμος, δημιουργεί συστάδες σφαιρικού σχήματος που δεν είναι πάντα επιθυμητό.

3.3.1.2 PAM

Ο PAM (Partitioning Around Medoids) αποτελεί μία από τις πιο γνωστές k-medoids μεθόδους. Για την εύρεση k συστάδων ο PAM καθορίζει ένα αντικείμενο αντιπρόσωπο για κάθε συστάδα. Τα αντικείμενα- αντιπρόσωποι καλούνται medoids και επιλέγονται μεταξύ αυτών που βρίσκονται πιο κοντά στα κέντρα των συστάδων. Μόλις επιλεγούν, κάθε μη επιλεγμένο αντικείμενο ομαδοποιείται στην συστάδα του medoid με το οποίο έχει τα περισσότερα κοινά χαρακτηριστικά. Η ποιότητα της συσταδοποίησης μετριέται με βάση τη μέση απόσταση ενός αντικειμένου από το medoid της συστάδας όπου ανήκει.

Η συσταδοποίηση με k-medoids αποτελεί βελτιστοποίηση των k-means μιας και ο αλγόριθμος k-MEANS επηρεάζεται από τις έκτροπες παρατηρήσεις, εξαιτίας της συνάρτησης που χρησιμοποιεί για την εύρεση του μέσου των συστάδων. Η ιδέα της μεθόδου είναι, αντί να επιλέγουμε την μέση τιμή (κέντρα) ως σημεία αναφοράς για μία συστάδα, να επιλέγουμε ένα αντικείμενο αντιπρόσωπο (τυχαία) για την αναπαράσταση της συστάδας, με κάθε συστάδα να έχει ένα αντικείμενο αντιπρόσωπο. Στην συνέχεια σε κάθε βήμα, εκτελείται μια ανταλλαγή αντικειμένων μέχρις ότου η ανταλλαγή αυτή οδηγήσει στην βελτίωση της ποιότητας της συσταδοποίησης. Επομένως, ο PAM είναι πιο αποτελεσματικός απέναντι σε σύνολα με θόρυβο ή έκτροπες παρατηρήσεις από ότι ο k-MEANS. Επιπρόσθετα, ο PAM δουλεύει ικανοποιητικά για μικρά σύνολα δεδομένων. Ωστόσο, για μεγάλα σύνολα κρίνετε μη αποδοτικός λόγω της μεγάλης πολυπλοκότητας του.

3.3.1.3 CLARA

Ο αλγόριθμος CLARA (Clustering Large Applications) σχεδιάστηκε προκειμένου να διαχειρίζονται μεγάλα σύνολα δεδομένων. Η βασική ιδέα είναι ότι για ένα μεγάλο σύνολο, αντί να εξετάσουμε όλο το σύνολο δεδομένων, παίρνουμε ένα δείγμα του και εφαρμόζουμε τον αλγόριθμο PAM. Πιο συγκεκριμένα, ο CLARA δέχεται ένα σύνολο δεδομένων ως είσοδο. Μέσω δειγματοληψίας (*sampling*), επιλέγετε ένα τυχαίο δείγμα του συνόλου στο οποίο εφαρμόζεται ο PAM. Στη συνέχεια, ο αλγόριθμος PAM θα εξετάσει και θα υπολογίσει τα βέλτιστα medoids του δείγματος. Ο CLARA δημιουργεί συστάδες από πολλαπλά τυχαία δείγματα και επιστρέφει την καλύτερη συσταδοποίηση ως έξοδο. Η βασική πρόταση αναφέρει ότι εάν το δείγμα είναι σχεδιασμένο με εντελώς τυχαίο τρόπο, τότε αναπαριστά ολόκληρο το σύνολο ικανοποιητικά και για το λόγο αυτόν τα αντικείμενα αντιπρόσωποι του δείγματος θα προσεγγίζουν τα medoids ολόκληρου του συνόλου δεδομένων.

Ο CLARA εφαρμόζει τον PAM μόνο σε δείγματα του συνόλου και έτσι η πολυπλοκότητα περιορίζεται σε κάθε επανάληψη. Η αποδοτικότητα του CLARA εξαρτάται από το μέγεθος του δείγματος. Παρατηρώντας ότι, ο PAM ψάχνει για τα καλύτερα medoids για κάθε δείγμα που του δίνεται, ενώ ο CLARA ψάχνει για τα καλύτερα k-medoids από όλα τα δείγματα, μπορούμε να διαπιστώσουμε ότι ο CLARA δεν θα μπορέσει να βρει την καλύτερη συσταδοποίηση εάν για παράδειγμα, ένα βέλτιστο medoid του συνόλου δεν επιλεχθεί κατά την εξαγωγή του δείγματος.

3.3.1.4. CLARANS

Ο αλγόριθμος CLARANS (Clustering Large Applications Based on Randomized Search) συνδυάζει τους αλγορίθμους PAM και CLARA. Ο CLARANS δημιουργεί ένα δείγμα από γείτονες σε κάθε βήμα της αναζήτησης του, με αποτέλεσμα να μην περιορίζεται η αναζήτηση τοπικά. Γείτονας (*neighbor*) συσταδοποίησης ονομάζεται η συσταδοποίηση που λαμβάνεται μετά τη αλλαγή ενός medoid. Ο αριθμός των γειτόνων που μπορούν να εξεταστούν ως medoids πιθανής λύσης καθορίζεται από την παράμετρο maxneighbor. Σε περίπτωση που εντοπιστεί ένας καλύτερος γείτονας, ο CLARANS μετακινείται στον κόμβο του γείτονα και η διαδικασία της αναζήτησης θα ξεκινήσει πάλι, με σκοπό να βρεθεί εάν αυτός ο κόμβος παράγει την καλύτερη λύση. Εάν βρεθεί ένα τοπικό βέλτιστο, ο CLARANS αρχίζει με έναν τυχαίο κόμβο για την αναζήτηση ενός νέου τοπικού βέλτιστου. Ο αριθμός των τοπικών βέλτιστων που θα αναζητήσει ο CLARANS καθορίζεται από μια παράμετρο που ονομάζεται numlocal.

Οι δύο παράμετροι, maxneighbor και numlocal, που εντοπίζονται στον αλγόριθμο, ανάλογα με τις τιμές που δέχονται επηρεάζουν τα αποτελέσματα του CLARANS. Για παράδειγμα για μεγάλο αριθμό maxneighbor ο CLARANS προσεγγίζει τον PAM και η αναζήτηση για την εύρεση τοπικού ελαχίστου έχει μεγαλύτερη διάρκεια. Γενικά, ο αλγόριθμος CLARANS έχει αποδειχθεί πιο αποδοτικός από τους αλγορίθμους CLARA και PAM.

3.3.2 Ιεραρχικοί Αλγόριθμοι

Οι ιεραρχικοί αλγόριθμοι συσταδοποίησης, εργάζονται ομαδοποιώντας τα δεδομένα σε συστάδες με ιεραρχική μορφή ή μορφή «δέντρου». Η αναπαράσταση των δεδομένων σε ιεραρχική μορφή είναι χρήσιμη για την σύνοψη και οπτικοποίηση των δεδομένων. Για παράδειγμα, ένας manager ανθρώπινου δυναμικού μίας εταιρίας, θα μπορούσε να οργανώσει τους εργαζομένους σε τμήματα ανάλογα με την ειδικότητα. Επιπλέον, θα μπορούσε να διαχωρίσει αυτά τα τμήματα σε μικρότερα υποτμήματα ανάλογα με την εμπειρία του προσωπικού. Όλα αυτά τα τμήματα αποτελούν μια ιεραρχική δομή μέσω τις οποίας θα μπορούσαμε να εξάγουμε χρήσιμες πληροφορίες. Επομένως, οι αλγόριθμοι συσταδοποίησης βασίζονται στην διάσπαση μεγαλύτερων συστάδων σε μικρότερες ή στην σύνδεση μικρότερων συστάδων σε μεγαλύτερες.

Οι ιεραρχικοί αλγόριθμοι συσταδοποίησης σύμφωνα με την μέθοδο που παράγουν τις συστάδες μπορούν περεταίρω να διαιρεθούν σε Συσσωρευτικούς Ιεραρχικούς αλγορίθμους (Agglomerative Hierarchical Clustering) και Διαιρετικούς Ιεραρχικούς αλγορίθμους (Divisive Hierarchical Clustering).

Οι συσσωρευτικοί ιεραρχικοί αλγόριθμοι, όπως καταλαβαίνει κανείς από την ονομασία, ενώνουν τις συστάδες μεταξύ τους. Αρχικά, θεωρούν κάθε αντικείμενο ως συστάδα και μειώνουν των αριθμό των συστάδων σε κάθε βήμα συγχωνεύοντας τις δύο πλησιέστερες συστάδες. Για να βρεθεί η ομοιότητα δύο συστάδων, χρησιμοποιείται ως κριτήριο κριτήρια η ελάχιστη, η μέγιστη, ή η μέση απόσταση μεταξύ των σημείων των δύο συστάδων.

Αντίθεση από τους συσσωρευτικούς αλγορίθμους, οι διαιρετικοί ιεραρχικοί αλγόριθμοι παράγουν σταδιακά σχήματα συσταδοποίησης αυξάνοντας τον αριθμό των συστάδων, οδηγώντας στον διαχωρισμό μιας συστάδας σε δύο. Αρχικά, θεωρούν ότι υπάρχει μία συστάδα η οποία περιλαμβάνει όλα τα αντικείμενα και στην συνέχει επιχειρούν τον διαχωρισμό σε ομοιογενείς ομάδες.

Βασικό ελάττωμα των ιεραρχικών μεθόδων αποτελεί το γεγονός ότι κάθε διαχωρισμός ή συσσώρευση που πραγματοποιείται δεν μπορεί να ανακληθεί. Το τελικό αποτέλεσμα του αλγορίθμου είναι ένα δένδρο από συστάδες το οποίο καλείται δενδρογράφημα και παρουσιάζει τον τρόπο που οι συστάδες σχετίζονται μεταξύ τους.

3.3.2.1 CURE

Ο αλγόριθμος CURE (Clustering Using Representatives) αποτελεί έναν ιεραρχικό αλγόριθμο συσταδοποίησης που έχει ως σκοπό την συγχώνευση των πλησιέστερων συστάδων, στοχεύοντας στην δημιουργία ομοιογενών ομάδων. Κύρια χαρακτηριστικά του αλγορίθμου είναι ότι μπορεί να αναγνωρίζει συστάδες αυθαίρετων σχημάτων, είναι αποδοτικός στην παρουσία outliers και οι απαιτήσεις του σε χώρο αποθήκευσης είναι ανάλογες του πλήθους των στοιχείων εισόδου.

Ο αλγόριθμος αρχικά λαμβάνει στοιχεία ως είσοδο, με κάθε στοιχείο να αναγνωρίζεται ως συστάδα. Σε κάθε βήμα που ακολουθεί συγχωνεύει τα πλησιέστερα ζευγάρια συστάδων. Για τον υπολογισμό της απόστασης μεταξύ δύο συστάδων, θεωρούμε για κάθε συστάδα c αντιπροσώπους , τα c πιο διάσπαρτα σημεία μέσα στην συστάδα). Οι c αντιπρόσωποι (*representatives*) είναι αυτοί που θα καθορίσουν το φυσικό σχήμα της συστάδας. Στην συνέχεια, μετακινούμε τα σημεία αυτά προς το μέσο της συστάδας κατά ποσοστό α. Αυτό έχει ως αποτέλεσμα να απαλειφθεί ο

θόρυβος και να μετριαστούν οι επιδράσεις των outliers. Η απόσταση μεταξύ δύο συστάδων προκύπτει από την απόσταση μεταξύ των πιο κοντινών αντιπροσώπων των δύο συστάδων.

Η παράμετρος α μπορεί να χρησιμοποιηθεί και για τον έλεγχο του σχήματος της συστάδας. Για μικρές τιμές του α, τα διάσπαρτα σημεία-αντιπρόσωποι δεν συρρικνώνονται αρκετά προς το κέντρο με αποτέλεσμα η συστάδα να μην έχει σφαιρικό σχήμα. Αντίθετα, για μεγάλες τιμές του α έχουμε την δημιουργία συμπαγών και σφαιρικών συστάδων.

Η πολυπλοκότητα του CURE εξαρτάται από τον αριθμό των στοιχείων εισόδου. Για τον λόγο αυτό, όταν τα στοιχεία εισόδου είναι πολλά ο αλγόριθμος δεν είναι άμεσα εφαρμόσιμος. Ο αλγόριθμος χειρίζεται μεγάλες βάσεις δεδομένων χρησιμοποιώντας τον συνδυασμό τυχαίας δειγματοποίησης (*sampling*) και τμηματοποίησης (*partitioning*). Επιλέγεται δηλαδή, ένα δείγμα του συνόλου των δεδομένων με τέτοιο τρόπο ώστε το δείγμα να είναι αντιπροσωπευτικό του συνόλου. Καθώς όμως ο διαχωρισμός μεταξύ των συστάδων μειώνεται και οι συστάδες γίνονται πιο πυκνές, απαιτούνται δείγματα μεγάλου μεγέθους για να διακριθούν με επιτυχία οι συστάδες που υπάρχουν σε ένα σύνολο. Άρα, πάλι η υπολογιστική πολυπλοκότητα του CURE θα αυξηθεί σημαντικά. Για τον λόγο αυτό προτείνεται ένα απλό σχήμα συσταδοποίησης όταν τα μεγέθη των δειγμάτων είναι πολύ μεγάλα.

3.3.2.2 BIRCH

Ο αλγόριθμος BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) σχεδιάστηκε για την συσταδοποίηση μεγάλων αριθμητικών συνόλων με την βοήθεια των μεθόδων ιεραρχικής συσταδοποίησης και άλλων όπως οι επαναληπτικές τμηματοποιήσεις. Με τον αλγόριθμο αυτό καταφέρνουμε να καταπολεμήσουμε τις δύο κυριότερες δυσκολίες που παρουσιάζουν οι συσσωρευτικοί αλγόριθμοι, την επεκτασιμότητα και την ανικανότητα αναίρεσης του προηγούμενου βήματος.

Ο BIRCH χρησιμοποιεί μια ιεραρχική δομή δεδομένων που καλείται CF-tree (Clustering Feature) για την τμηματοποίηση των στοιχείων του συνόλου δεδομένων με δυναμικό τρόπο. Μέσω των χαρακτηριστικών γνωρισμάτων (*CF*) μια συστάδας μπορούμε να υπολογίσουμε χρήσιμα στατιστικά χαρακτηριστικά όπως το κέντρο της, την ακτίνα της, την διάμετρο της και άλλα. Επομένως, για κάθε συστάδα έχουμε ένα διάνυσμα CF. .

Το CF-Tree είναι ένα height-balanced δέντρο, το οποίο καταχωρεί τα χαρακτηριστικά γνωρίσματα της συσταδοποίησης. Ένας κόμβος του δέντρου θα αντιπροσωπεύει ένα φύλλο ή την ρίζα ενός υπόδεντρου. Ένας κόμβος ο οποίος δεν είναι «φύλλο» περιέχει το άθροισμα των CF των παιδιών του. Ένα CF-Tree είναι βασισμένο σε δύο παραμέτρους, τον παράγοντα διακλαδισμού (branching factor) και το άνω όριο (threshold). Ο παράγοντας διακλαδισμού αναφέρεται στο μέγιστο επιτρεπτό πλήθος παιδιών ενός κόμβου, ενώ ο παράγοντας άνω ορίου αναφέρετε στην μέγιστη διάμετρο (ακτίνα) κάθε συστάδας.

Ο BRICH μπορεί να βρει μια καλή συσταδοποίηση με μία συνολική σάρωση των δεδομένων και στην συνέχεια να βελτιώσει την ποιότητα του με μερικές πρόσθετες σαρώσεις. Ωστόσο, το αποτέλεσμα δεν ανταποκρίνεται πλήρως σε αυτό που ο χρήστης αναλογίζετε ως συστάδα, δεδομένου ότι κάθε κόμβος στο CF-Tree μπορεί να κρατήσει έναν περιορισμένο αριθμό εισόδων εξαιτίας του μεγέθους του. Επιπλέον, εάν οι συστάδες δεν είναι σφαιρικές, ο BIRCH δεν λειτουργεί αποτελεσματικά μιας και χρησιμοποιεί μέτρα όπως η διάμετρος ή η ακτίνα για την οριοθέτηση των συστάδων. Τέλος, παρουσιάζει ευαισθησία στην σειρά των δεδομένων καθώς μπορεί να παράγει διαφορετικές συστάδες για διαφορετική σειρά των ίδιων δεδομένων εισόδου.

3.3.3 Αλγόριθμοι βασισμένοι στην πυκνότητα

Οι βασισμένοι στην πυκνότητα αλγόριθμοι θεωρούν τις συστάδες ως πυκνές περιοχές αντικειμένων οι οποίες χωρίζονται από περιοχές χαμηλής πυκνότητας. Για την αντιμετώπιση της δυσκολίας ανακάλυψης αυθαίρετων σχημάτων, οι αλγόριθμοι αυτοί ακολουθούν την στρατηγική βασισμένη στην πυκνότητα για την ανακάλυψη μη σφαιρικών συστάδων. Παρακάτω αναλύονται οι δυο αντιπροσωπευτικοί αλγόριθμοι, ο DBSCAN και ο DENCLUE.

3.3.3.1 DBSCAN

Ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise) βρίσκει αντικείμενα (πυρήνες) που έχουν γείτονες με μεγάλη πυκνότητα. Η γειτονιά κάθε αντικειμένου εκτίνεται σε ακτίνα Eps γύρω από το σύνολο και ο ελάχιστον αριθμός στοιχείων που μπορεί να έχει το σύνολο είναι MinPts. Ο αλγόριθμος, ενώνοντας τους πυρήνες με τους γείτονες τους δημιουργεί συστάδες. Στην συσταδοποίηση τα αντικείμενα διακρίνονται σε αντικείμενα πυρήνα και αντικείμενα όχι πυρήνα. Ο

αλγόριθμος DBSCAN δέχεται ως είσοδο το σύνολο δεδομένων και στην συνέχει απαιτεί από τον χρήστη τον καθορισμό των παραμέτρων Eps και MinPts που ορίζουν την ελάχιστη πυκνότητα για τη συσταδοποίηση.

Αρχικά, ο αλγόριθμος «μαρκάρει» ως «άγνωστα» όλα τα σημεία εισόδου. Στην συνέχεια, επιλέγει τυχαία ένα αντικείμενο, το «μαρκάρει» ως «γνωστό» και ελέγχει αν η γειτονιά του περιλαμβάνει τουλάχιστον MinPts. Αν το πλήθος των στοιχείων δεν είναι ικανοποιητικό για να οριστεί το αντικείμενο πυρήνας, «μαρκάρετε» ως θόρυβος. Τότε, ο DBSCAN λαμβάνει το επόμενο στοιχείο που δεν έχει επισκεφθεί από την βάση. Η διαδικασία τελειώνει όταν ο DBSCAN επισκεφθεί όλα τα αντικείμενα.

Παρόλο που ο αλγόριθμος μπορεί να βρει συστάδες με αυθαίρετα σχήματα έχει αρκετά προβλήματα. Το μεγάλο κόστος της εισαγωγής/εξαγωγής εξαιτίας της εφαρμογής της συσταδοποίησης απευθείας στο σύνολο των δεδομένων, με αποτέλεσμα να καθίσταται ασύμφορος για μεγάλα σύνολα. Επίσης, το γεγονός ότι επηρεάζεται από τις τιμές των παραμέτρων Eps και MinPts.

3.3.3.2 DENCLUE

Ο DENCLUE (DENsity-based CLUstEring) είναι ένας άλλος βασισμένος στην πυκνότητα αλγόριθμος. Η βασική ιδέα της προσέγγισης αυτής είναι η χρήση μη-παραμετρικών μεθόδων για την μοντελοποίηση της πυκνότητας των σημείων. Οι μη-παραμετρικές μέθοδοι δεν χρησιμοποιούν υποθέσεις σχετικά με τα δεδομένα (βλέπε Eps, MinPts στον DBSAN). Όπως είναι κατανοητό οι παράμετρος Eps και MinPts επηρεάζουν πλήρως την πυκνότητα μίας γειτονίας ενός αντικειμένου, διότι για μεγαλύτερες τιμές του Eps η τιμή της πυκνότητας για του αντικειμένου θα ήταν μεγαλύτερη. Για την αντιμετώπιση αυτού του προβλήματος ο DENCLUE χρησιμοποιεί την εκτίμηση πυκνότητας με πυρήνες. Μια τέτοια συνάρτηση όπως η συνάρτηση εκτίμησης της πυκνότητας με πυρήνες μπορεί να θεωρηθεί ως μία συνάρτηση επιρροής, δηλαδή μια συνάρτηση η οποία περιγράφει την επίδραση ενός σημείου από το σύνολο των δεδομένων στην γειτονιά του. Οι συστάδες μπορούν να προσδιοριστούν από τους density attractors. Ως density attractor θεωρούμε τα τοπικά μέγιστα της συνολικής συνάρτησης υπολογισμού της πυκνότητας.

Τα βασικά πλεονεκτήματα του DENCLUE είναι η καλή διαχείριση των συνόλων που περιέχουν θόρυβο και η ανακάλυψη συστάδων με περίεργες γεωμετρίες. Ο DENCLUE βασίζεται σε δύο παραμέτρους, την επίδραση ενός

στοιχείου στην γειτονιά του και την περιγραφή της σπουδαιότητας ενός density attractor.

3.3.4 Αλγόριθμοι βασισμένοι σε πλέγμα

Οι αλγόριθμοι που έχουν παρουσιαστεί μέχρις στιγμής στην εργασία κατευθύνονται από τα αντικείμενα-στοιχεία του συνόλου δεδομένων. Αντίθετα, οι αλγόριθμοι βασισμένοι σε πλέγμα έχουν ως κύριο οδηγό το διάστημα των χωρικών στοιχείων του υπό μελέτη συνόλου δεδομένων. Οι αλγόριθμοι αυτοί, καθορίζουν ένα σύνολο κυψελών, στο οποίο κάθε αντικείμενο αναθέτετε στην κατάλληλη κυψέλη και στην συνέχεια υπολογίζεται η πυκνότητα κάθε κυψέλης. Η πολυπλοκότητα αυτής της συσταδοποίησης δεν εξαρτάται από το μέγεθος του συνόλου των δεδομένων, αλλά από το πλήθος των «πυκνών» κελιών. Μερικοί διαδεδομένοι αλγόριθμοί βασισμένοι σε πλέγμα είναι οι STING και WaveCluster.

3.3.4.1 STING

Η μέθοδος STING (STatistical INformation Grid) είναι αντιπροσωπευτική της παραπάνω κατηγορίας. Κατά την εφαρμογή του STING, η χωρική περιοχή διαιρείται σε ορθογώνια κελιά. Δημιουργούνται διαφορετικά επίπεδα κελιών, όπου το κάθε επίπεδο αντιπροσωπεύει πληροφορίες συσταδοποίησης και τα οποία παράγουν μια ιεραρχική δομή. Κάθε κελί το οποίο βρίσκεται σε υψηλό επίπεδο διασπάται σε μικρότερα, με σκοπό την δημιουργία κελιών για τα χαμηλότερα επίπεδα. Οι στατιστικές πληροφορίες των χαρακτηριστικών κάθε κελιού όπως ο μέσος, η διακύμανση, το ελάχιστο ή μέγιστο και ο τύπος κατανομής υπολογίζονται από πριν και αποθηκεύονται ως στατιστικές παράμετροι. Οι στατιστικές αυτοί παράμετροι βοηθάνε στην διαδικασία των απαντήσεων στις ερωτήσεις της βάσης καθώς και σε άλλες εργασίες ανάλυσης δεδομένων.

Ο αλγόριθμος υποθέτει ότι τίθεται μια ερώτηση που μπορεί να απαντηθεί από την αποθηκευμένη στατιστική πληροφορία στην κατασκευασμένη δομή. Μια τέτοια ερώτηση μπορεί να είναι η εύρεση του εύρους των τιμών ενός φαινομένου, για παράδειγμα οι τιμές διαμερισμάτων, κοντά σε μια συγκεκριμένη περιοχή, για παράδειγμα στο κέντρο της πόλης. Οι στατιστικές παράμετροι μπορούν να χρησιμοποιηθούν σε μία «από πάνω προς τα κάτω» αναζήτηση μιας απάντησης σε μια ερώτηση που γίνετε σε μία βάση. Πιο συγκεκριμένα, στην αρχή αποφασίζετε από πιο επίπεδο της ιεραρχικής δομής θα ξεκινήσει η διαδικασία. Υποθέτουμε ότι το

αρχικό επίπεδο περιέχει έναν μικρό αριθμό κελιών. Για κάθε κελί στο επίπεδο που εργαζόμαστε, υπολογίζουμε το διάστημα εμπιστοσύνης με σκοπό να αντικατοπτρίσουμε τον βαθμό συσχέτισης ενός κελιού με την δοσμένη ερώτηση. Τα κελιά που θεωρούνται «άσχετα» με την ερώτηση απομακρύνονται από την διαδικασία. Στην συνέχεια «κατεβαίνουμε» κατά ένα επίπεδο στην ιεραρχική δομή και επαναλαμβάνουμε την ίδια διαδικασία σε εκείνα τα κελιά που αποτελούν τμήματα των προηγούμενων σχετιζόμενων με την ερώτηση κελιών (του προηγούμενου επιπέδου). Η διαδικασία επαναλαμβάνεται μέχρι το τελευταίο επίπεδο. Σε εκείνο το σημείο, εάν οι προδιαγραφές της απάντησης στην ερώτηση βρεθούν, επιστρέφεται ως έξοδος το διάστημα των συσχετιζόμενων κελιών.

3.3.4.2 WaveCluster

Ο WaveCluster είναι ένας αλγόριθμος βασισμένος σε πλέγμα καθώς και στην πυκνότητα, οπού χρησιμοποιεί wavelet μετασχηματισμούς για τον εντοπισμό των συστάδων. Ένας wavelet μετασχηματισμός είναι μια τεχνική επεξεργασίας σημάτων, η οποία αποσυνθέτει το σήμα σε διαφορετικές μπάντες συχνοτήτων. Η ιδέα για την χρήση των μετασχηματισμών πηγάζει από το γεγονός ότι τα σημεία-στοιχεία δεδομένων θεωρούνται k-διάστατα σήματα. Τα κομμάτια του σήματος που βρίσκονται στις υψηλές συχνότητες αντιστοιχούν στα σημεία που βρίσκονται στα σύνορα-όρια των συστάδων, ενώ τα κομμάτια στου σήματος που βρίσκονται στις χαμηλές συχνότητες αντιστοιχούν στα σημεία που βρίσκονται στο εσωτερικό των συστάδων. Εξετάζοντας τα σήματα σε διαφορετικές μπάντες συχνοτήτων, προκύπτουν διαφορετικές συσταδοποιήσεις.

3.3.5 Συσταδοποίηση υποχώρων

Η συσταδοποίηση υποχώρων εξετάζει τα προβλήματα που προκύπτουν από τις υψηλές διαστάσεις των δεδομένων. Αυτή η κατηγορία αλγορίθμων προσπαθεί να βρει τα υποσύνολα του αρχικού χώρου όπου η συσταδοποίηση δίνει πιο καλά αποτελέσματα. Η ανάλυση συστάδων έχει ως κύριο μέλημα την εύρεση ομάδων, οι οποίες αποτελούνται από όμοια αντικείμενα. Τα αντικείμενα αναπαρίστανται συνήθως ως διανύσματα μετρήσεων ή σημεία σε πολυδιάστατους χώρους. Η ομοιότητα των αντικειμένων συχνά υπολογίζεται από τα μέτρα απόστασης των διαφόρων διαστάσεων στο σύνολο των δεδομένων. Οι παραδοσιακοί αλγόριθμοι συσταδοποίησης λαμβάνουν υπόψη όλο το φάσμα διαστάσεων του συνόλου

δεδομένων με σκοπό την βέλτιστη περιγραφή και κατανόηση κάθε αντικειμένου του συνόλου. Ωστόσο, σε υψηλών διαστάσεων δεδομένα, πολλές από τις διαστάσεις, συνήθως, δεν σχετίζονται μεταξύ τους. Αυτές οι άσχετες μεταξύ τους διαστάσεις δημιουργούν προβλήματα στους αλγορίθμους. Ένα από τα προβλήματα που αντιμετωπίζουν οι βασισμένοι στα χαρακτηριστικά αλγόριθμοι είναι αυτό της υπερκάλυψης συστάδων μιας και είναι σύνηθες το φαινόμενο κατά το οποίο όταν βρισκόμαστε σε υψηλές διαστάσεις τα αντικείμενα ενός συνόλου δεδομένων να ισαπέχουν μεταξύ τους. Μία ακόμη δυσκολία που προσθέτουν οι πολλές διαστάσεις στους βασισμένους στα χαρακτηριστικά αλγορίθμους είναι η συρρίκνωση της σημασίας ενός μέτρου απόστασης, καθώς αυξάνουμε τον αριθμό των διαστάσεων.
Με σκοπό την εύρεση «σωστών» συστάδων, είναι απαραίτητη η απομάκρυνση των ασυσχέτιστων χαρακτηριστικών και η εφαρμογή αλγορίθμων σε περιοχές διαστάσεων που σχετίζονται μεταξύ τους. Δηλαδή, για τους παραπάνω λόγους είναι πιο ουσιαστικό, να ψάχνουμε για συστάδες σε συγκεκριμένα υποδιαστήματα του συνολικού διαστήματος δεδομένων.

3.3.5.1 CLIQUE

Ο CLIQUE (CLustering In QUEst) είναι ένας αλγόριθμος βασισμένος στην πυκνότητα ο οποίος ανακαλύπτει πυκνές περιοχές σε κάθε υποχώρο του συνολικού χώρου. Ο CLIQUE διαιρεί κάθε διάσταση του χώρο σε μη επικαλυπτόμενα, ίσα διαστήματα (κελιά). Στην συνέχεια, με την τεχνική ανακάλυψης συστάδων με βάση την πυκνότητα, διαχωρίζει τα πυκνά κελιά από τα υπόλοιπα. Ένα κελί είναι πυκνό, εάν το πλήθος των στοιχείων που βρίσκονται στην γειτονιά του υπερβαίνει ένα κατώτατο όριο.

Ο αλγόριθμος CLIQUE προχωρά από υποχώρους χαμηλότερης διάστασης σε υποχώρους υψηλότερης και ανακαλύπτει πυκνές περιοχές σε κάθε υποχώρο. Ο συνδυασμός των αντίστοιχων διαστημάτων ανά διάσταση βασίζεται στην ιδιότητα Apriori, η οποία βασίζετε σε πρότυπα συχνοτήτων και κανόνες συσχέτισης μεταξύ αντικειμένων. Για την συσχέτιση μεταξύ των συστάδων υποχώρων διαφορετικών διαστάσεων αναφέρουμε ότι, ένα k-διάστατο κελί $(k > 1)$ μπορεί να περιέχει τουλάχιστον λ σημεία αν και μόνο αν, κάθε κελί στον $(k - 1)$-διάστατο υποχώρο περιέχει τουλάχιστον λ σημεία.

Ο CLIQUE υλοποιείται σε δύο βήματα. Στο πρώτο βήμα, χωρίζει τον k-διάστατο χώρο σε κελιά με την διαίρεση κάθε διάστασης σε ίσο αριθμό και μήκος

διαστημάτων. Για ένα δεδομένο σύνολο διαστάσεων, ο συνδυασμός των αντίστοιχων διαστημάτων (ένα για κάθε διάσταση του συνόλου) καλείται μονάδα (*unit*) στον αντίστοιχο υποχώρο. Στην συνέχεια, βρίσκει όλες τις πυκνές μονάδες σε κάθε k-διάστατο υποχώρο μέσω της δημιουργίας των πυκνών (k-1)-διάστατων υποχώρων. Στο δεύτερο βήμα, ο αλγόριθμος χρησιμοποιεί τα πυκνά κελιά κάθε υποχώρου για την δημιουργία των συστάδων.

Ο αλγόριθμος βρίσκει αυτόματα υποχώρους υψηλών διαστάσεων οι οποίοι περιέχουν πυκνές συστάδες. Δεν επηρεάζεται από την σειρά εισαγωγής των στοιχείων και δεν θεωρεί καμία κανονική κατανομή για τα δεδομένα. Η πολυπλοκότητα του κλιμακώνετε γραμμικά ανάλογα με το μέγεθος των δεδομένων εισόδου και είναι εξελίξιμος όσο ο αριθμός των διαστάσεων των δεδομένων αυξάνετε. Ωστόσο, το μέγεθος των υποχώρων και το πλήθος σημείων που καθορίζουν την πυκνότηταέχουν μεγάλη επίδραση στην ποιότητα της συσταδοποίησης.

3.3.5.2 PROCLUS

Ο PROCLUS είναι ο πρώτος από πάνω προς τα κάτω (αναζήτηση υποχώρων) αλγόριθμος συσταδοποίησης. Η λειτουργία του είναι όμοια με αυτή του αλγορίθμου CLARANS, περιλαμβάνει δηλαδή σύνολα k αντιπροσώπων (*medoids*) και επαναληπτικές διαδικασίες με σκοπό την βελτίωση του συνόλου αντιπροσώπων. Ο PROCLUS αποτελείτε από τρείς φάσεις, την αρχικοποίηση, την επανάληψη και τη βελτίωση των συστάδων. Κατά την φάση της αρχικοποίησης, ο αλγόριθμος επιλέγει υποψήφια medoids τα οποία βρίσκονται σε μεγάλες αποστάσεις το ένα από το άλλο ,ώστε να αποφευχθεί η πιθανότητα δύο medoid να είναι αντιπρόσωποι της ίδιας ομάδας. Κατά την φάση της επανάληψης, αντικαθίστανται τα medoids που δεν ταιριάζουν στα υπό ανάλυση δεδομένα από το υπάρχον σύνολο αντιπροσώπων. Τέλος, για την εύρεση των διαστάσεων που επηρεάζουν περισσότερο κάθε συστάδα, επιλέγονται οι διαστάσεις κατά μήκος των οποίον τα σημεία έχουν τη μικρότερη μέση απόσταση (απόσταση Manhattan) από το τρέχον medoid.

3.3.6 Ασαφής συσταδοποίηση

Το ζήτημα της διαχείρισης της αβεβαιότητας, μας οδήγησε σε μια τεχνική συσταδοποίησης που χρησιμοποιεί έννοιες της ασαφούς λογικής και της θεωρίας ασαφών συνόλων, την ασαφής συσταδοποίηση. Υπάρχουν διάφορες προσεγγίσεις για

την επίλυση προβλημάτων μη ασαφής συσταδοποίησης με τους Fuzzy αλγορίθμους και τα πιθανοτικά μοντέλα να ξεχωρίζουν.

Οι αλγόριθμοι που έχουμε μελετήσει μέχρι στιγμής στην παρούσα εργασία έχουν ως αποτέλεσμα μη ασαφής συστάδες, που σημαίνει ότι ένα στοιχείο είτε ανήκει σε μία συστάδα είτε δεν ανήκει. Στην ασαφής συσταδοποίηση, ένα σημείο ανήκει σε κάθε συστάδα με κάποιο βάρος μεταξύ του 0 και του 1 (συχνά τα βάρη για κάθε σημείο έχουν άθροισμα 1). Δεδομένου ενός συνόλου αντικειμένων, $X = \{X_1, .., X_n\}$, ένα ασαφής σύνολο S είναι ένα υποσύνολο του συνόλου X το οποίο επιτρέπει σε κάθε αντικείμενο να έχει έναν βαθμό συσχέτισης μεταξύ του 0 και το 1. Εκφράζοντας την παραπάνω πρόταση με την μορφή συνάρτησης έχουμε: $F_s: X \rightarrow [0,1]$.

Ο διαδεδομένος ασαφής αλγόριθμος συσταδοποίησης Fuzzy C-Means (FCM), αποτελεί μια επέκταση του κλασικού αλγορίθμου k-MEANS για τις ασαφείς εφαρμογές. Ο FCM προσπαθεί να βρει το χαρακτηριστικό σημείο σε κάθε συστάδα, δηλαδή το κέντρο της συστάδας και έπειτα τον βαθμό συμμετοχής για κάθε αντικείμενο στην συστάδα.

Κεφάλαιο 4^{ο} Πρακτική Εφαρμογή

Στο παρόν κεφάλαιο αναπτύσσεται ένα απλό πρόβλημα εξόρυξης γνώσης και ανακάλυψης προτύπων σε ένα σύνολο δεδομένων. Τα δεδομένα τα οποία χρησιμοποιούνται στα πλαίσια της παρούσας εργασίας προέρχονται από την διαδικτυακή συλλογή UCI (Machine Learning Repository). Σημειώνεται ότι, η επεξεργασία των δεδομένων έγινε με τη χρήση του MATLAB. Κάτωθι, παρουσιάζεται αναλυτικά η διαδικασία.

4.1 Παρουσίαση δεδομένων

Τα στοιχεία που χρησιμοποιούνται σχετίζονται με τις εκστρατείες άμεσου μάρκετινγκ ενός Πορτογαλικού τραπεζικού ιδρύματος. Πιο συγκεκριμένα, οι ενέργειες αυτές βασίστηκαν σε τηλεφωνικές κλήσεις που είχαν ως σκοπό τη συλλογή πληροφοριών σχετικά με το αν ο πελάτης θα ανταποκριθεί θετικά ή όχι στην κατάθεση χρημάτων σε προθεσμιακό λογαριασμό. Συχνά, χρειάστηκαν περισσότερες από μία επαφές με τον ίδιο πελάτη.

Το σύνολο δεδομένων που επιλέχθηκε περιέχει δεκαέξι μεταβλητές εισόδου, εκ των οποίων οι εννέα είναι κατηγορικές (*categorical*) και οι επτά αριθμητικές (*numeric*) και μία μεταβλητή εξόδου για την οποία θα εκπαιδευτεί το μοντέλο[1].

Αναλυτικότερα, οι μεταβλητές αυτές έχουν ως εξής:

age: Ηλικία του πελάτη σε χρόνια (αριθμητική).

job: Είδος εργασιακής απασχόλησης π.χ. επιχειρηματίας, άνεργος κλπ (κατηγορική).

marital: Οικογενειακή κατάσταση π.χ. παντρεμένος, διαζευγμένος κλπ (κατηγορική).

education: Επίπεδο εκπαίδευσης π.χ. δευτεροβάθμια, άγνωστη κλπ (κατηγορική).

default: Αν έχει προεπιλεγμένη πίστωση ο κάθε πελάτης, δυαδικής μορφής με πιθανές απαντήσεις “ναι” και “όχι” (κατηγορική).

balance: Μέσο ετήσιο υπόλοιπο του δανείου μετρημένο σε ευρώ (αριθμητική).

housing: Αν έχει στεγαστικό δάνειο, δυαδικής μορφής με πιθανές απαντήσεις “ναι” και “όχι” (κατηγορική).

loan: Αν έχει προσωπικό δάνειο, δυαδικής μορφής με πιθανές απαντήσεις “ναι” και “όχι” (κατηγορική).

contact: Τύπος επικοινωνίας π.χ. άγνωστο, σταθερό κλπ (κατηγορική).

day: Τελευταία ημέρα επικοινωνίας του μήνα (αριθμητική).

[1] Τα στοιχεία αποτελούν μέρος dataset που αναφέρεται στο Moro, Cortez, & Rita (2014).

month: Τελευταίος μήνας επικοινωνίας του έτους (κατηγορική).

duration: Διάρκεια τελευταίας επικοινωνίας μετρημένη σε δευτερόλεπτα (αριθμητική).

campaign: Ο αριθμός των επαφών που πραγματοποιήθηκαν συνολικά κατά τη διάρκεια της καμπάνιας για τον συγκεκριμένο πελάτη (αριθμητική).

pdays: Ο αριθμός των ημερών που πέρασαν μετά την τελευταία επαφή με τον πελάτη που αφορούσε προηγούμενη καμπάνια (αριθμητική).

previous: Ο αριθμός των επαφών που πραγματοποιήθηκαν πριν από αυτή την καμπάνια στον συγκεκριμένο πελάτη (αριθμητική).

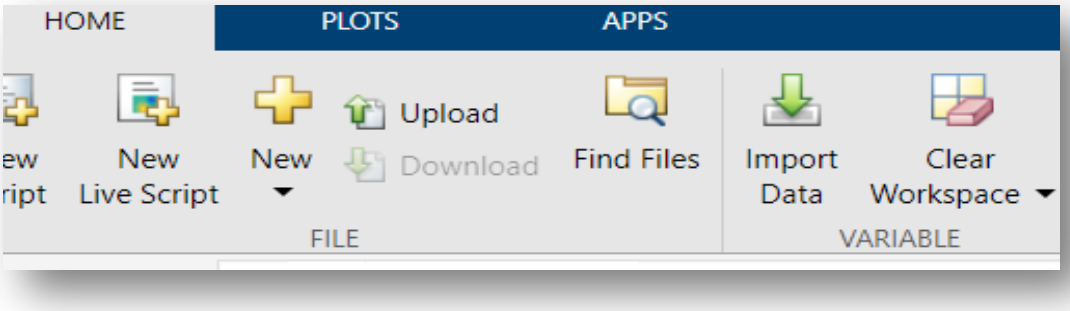
poutcome: Το αποτέλεσμα της προηγούμενης εκστρατείας μάρκετινγκ π.χ επιτυχία, άγνωστο κλπ (κατηγορική).

y: Πιθανή απάντηση του πελάτη για τη σύναψη συμβολαίου προθεσμιακής κατάθεσης δυαδικού τύπου «ναι» , «όχι» (κατηγορική).

Στόχος της ανάλυσης των συγκεκριμένων δεδομένων είναι η (σωστή) πρόβλεψη της απόφασης ενός πελάτη της τράπεζας για τη σύναψη συμβολαίου προθεσμιακής κατάθεσης.

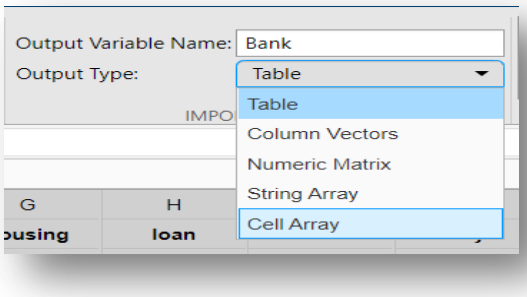
4.2 Επεξεργασία δεδομένων

Εφόσον τα δεδομένα «κατέβουν» στο τερματικό, ακολουθεί η εισαγωγή τους στο MATLAB. Η εισαγωγή μπορεί να γίνει εύκολα και γρήγορα με την χρήση του εργαλείου Import Data (εικόνα 3) το οποίο χρησιμεύει στην προβολή και εισαγωγή αρχείων που βρίσκονται στο τερματικό μας.



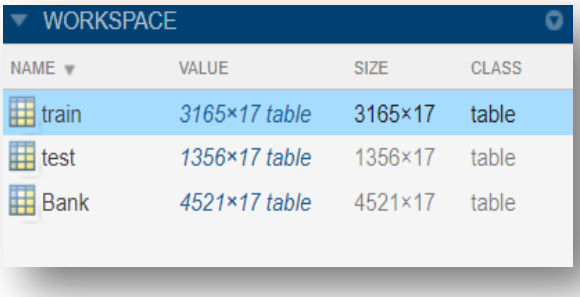
Εικόνα 3. Upload & Import Data

Το εργαλείο Import Data επιτρέπει την εισαγωγή δεδομένων σε διάφορες μορφές όπως Numeric Matrix, Column Vectors κλπ (εικ.2). Για την συγκεκριμένη εργασία επιλέγουμε την εισαγωγή των δεδομένων σε μορφή πίνακα (Table).



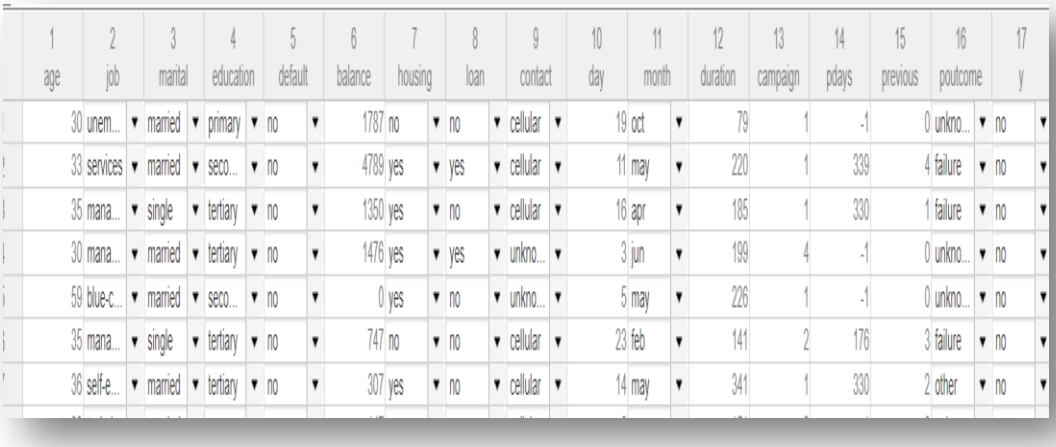
Εικόνα 4. Import Data Tool

Το σύνολο των δεδομένων θα πρέπει να είναι χωρισμένο σε train και test. Στην περίπτωσή που αυτό δεν συμβαίνει εξαρχής, θα πρέπει εμείς να χωρίσουμε τα δεδομένα με τυχαίο τρόπο. Στο παράθυρο Workspace βρίσκονται οι πίνακες με τους οποίους επιθυμούμε να εργαστούμε.



Εικόνα 5. Train & Test tables

Στην συνέχεια πρέπει να κατανοήσουμε τα δεδομένα μας, τις τιμές, τους τύπους, την συσχέτιση που έχουν με την μεταβλητή στόχο (y), ώστε να κρίνουμε ποιες μετατροπές θα εφαρμόσουμε και ποιες μεταβλητές θα χρησιμοποιήσουμε για την εκπαίδευση.



	1 age	2 job	3 marital	4 education	5 default	6 balance	7 housing	8 loan	9 contact	10 day	11 month	12 duration	13 campaign	14 pdays	15 previous	16 poutcome	17 y
	30	unem...	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unkno...	no
	33	services	married	seco...	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
	35	mana...	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
	30	mana...	married	tertiary	no	1476	yes	yes	unkno...	3	jun	199	4	-1	0	unkno...	no
	59	blue-c...	married	seco...	no	0	yes	no	unkno...	5	may	226	1	-1	0	unkno...	no
	35	mana...	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
	36	self-e...	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no

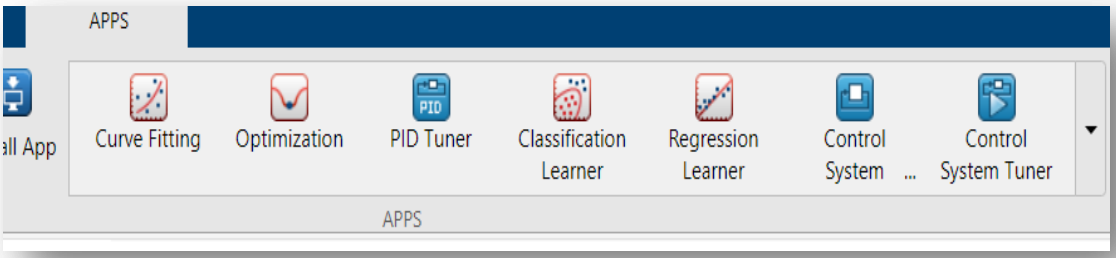
Εικόνα 6. Train Set

Η προ-επεξεργασία των δεδομένων είναι ένα σημαντικό βήμα στην διαδικασία εξόρυξης γνώσης. Με βάση ένα σύνολο το οποίο περιλαμβάνει θόρυβο, ημιτελή και ασυνεπή δεδομένα, τα αποτελέσματα της εξόρυξης γνώσης τείνουν στο να είναι ανακριβή και να στερούνται ποιότητας. Για το λόγο αυτό, χρησιμοποιούμε τεχνικές όπως ο καθαρισμός των δεδομένων, ο μετασχηματισμός των δεδομένων και η μείωση τους. Για χάριν ευκολίας, στην συγκεκριμένη περίπτωση δεν θα πραγματοποιηθούν αυτές οι διαδικασίες. Ορισμένα από τα γνωρίσματα, όπως το job,education,month κλπ είναι κατηγορικά και πρέπει να τα μετατρέψουμε σε αριθμητικά. Θα επιλέξουμε να μετατρέψουμε τις τιμές αυτές σε αριθμητικές με την βοήθεια της συνάρτησης double(). Θα έχουμε για παράδειγμα:

```
>> m_train=categorical(month_train);
>> month_train=double(month_train);
```

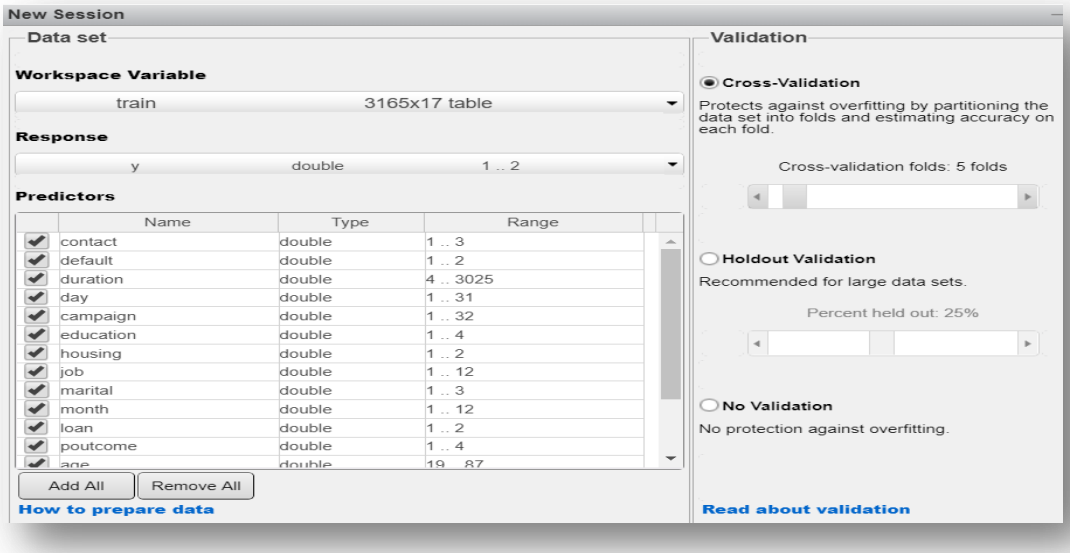
Εικόνα 7. Command Window 3

Στη συνέχεια εισάγουμε τα δεδομένα μας στην εφαρμογή Classification Learner του MATLAB. Κατά την έναρξη της εφαρμογής παρουσιάζεται το παρακάτω περιβάλλον:



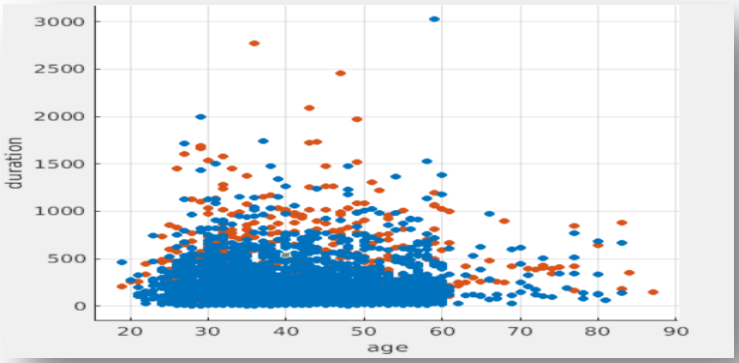
Εικόνα 8. Set Up Classification

Στο πρώτο βήμα, πρέπει να επιλεχθεί το dataset με το οποίο θέλουμε να εργαστούμε. Στην περίπτωση μας το σύνολο είναι το train_total. Το δεύτερο βήμα είναι αυτό στο οποίο ο χρήστης καλείται να κάνει τον διαχωρισμό των χαρακτηριστικών. Από τα χαρακτηριστικά του συνόλου πρέπει να επιλεχθούν εκείνα που θα χρησιμοποιηθούν για την πρόβλεψη (*Predictor*), εκείνα που θα ορίζουν την απάντηση (*Response*) καθώς και εκείνα που δεν θα χρησιμοποιηθούν στην εφαρμογή. Τέλος, η επιλογή της μεθόδου επικύρωσης(Validation Scheme) αφορά τον τρόπο εξέτασης της ακρίβειας του μοντέλου πρόβλεψης που θα αναπτυχθεί.

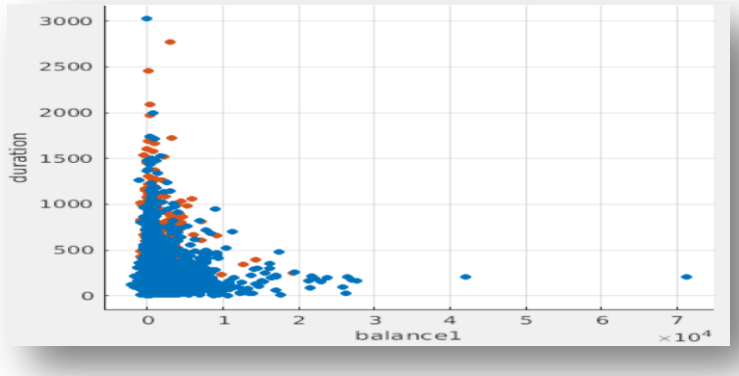


Εικόνα 9. Classification Learner App

Με την έναρξη της εφαρμογής εμφανίζεται το Scatter plot των χαρακτηριστικών. Το γράφημα αυτό έχει ως άξονες x και y χαρακτηριστικά που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου, και αναπαριστά τις απαντήσεις των πελατών της τράπεζας σε σχέση με δύο χαρακτηριστικά (εικόνα). Η επιλογή ή μη συγκεκριμένων χαρακτηριστικών αποτελεί μία καλή πρακτική για την αύξηση της ποιότητας των αποτελεσμάτων. Για τον λόγο αυτό, εναλλάσσουμε τις μεταβλητές στους άξονες ώστε να παρατηρήσουμε ποιος συνδυασμός μεταβλητών μας διαφοροποιεί καλύτερα τα αποτελέσματα. Για παράδειγμα, στην εικόνα 1 έχοντας στον οριζόντιο άξονα την μεταβλητή balance1 και στον κάθετο άξονα τη μεταβλητή duration παρατηρούμε ότι οι παρατηρήσεις είναι μπερδεμένες μεταξύ τους. Από την άλλη αν επιλέξουμε για τον οριζόντιο άξονα την μεταβλητή age και για τον κάθετο την μεταβλητή duration παρατηρούμε πιο ευδιάκριτα των διαχωρισμό των παρατηρήσεων «ναι» και «όχι».



Εικόνα 10. Scatter Plot 1



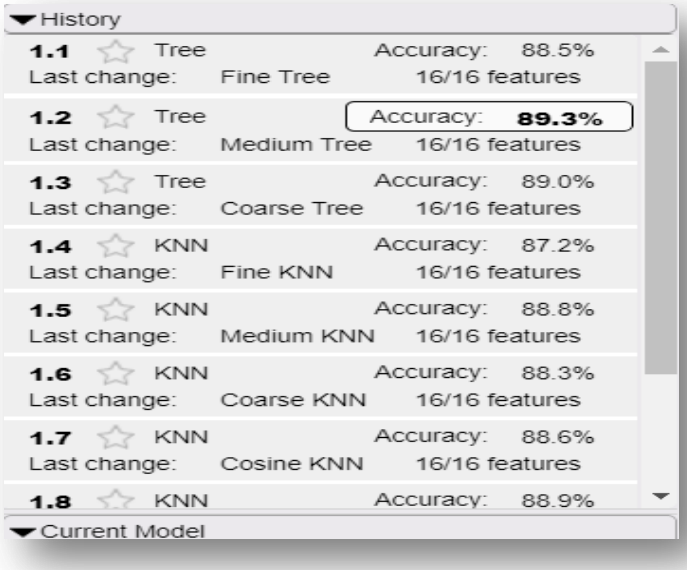
Εικόνα 11. Scatter Plot 2

Η επιλογή ή όχι αφαίρεσης ενός χαρακτηριστικού για την εύρεση του αποδοτικότερου μοντέλου μπορεί να γίνει εύκολα με το εργαλείο Feature Selection που βρίσκετε πάνω αριστερά στην εφαρμογή. Στο επόμενο βήμα γίνεται η επιλογή του μοντέλου με το οποίο θα εργαστούμε, δηλαδή του μοντέλου στο οποίο θα εισάγουμε τα δεδομένα, θα το εκπαιδεύσουμε και στην συνέχεια θα αναλύσουμε την αποτελεσματικότητα του με την βοήθεια του test dataset. Επιλέγουμε k-cross validation (k=5) για την επικύρωση των αποτελεσμάτων.

Είναι γεγονός ότι η επιλογή του καταλληλότερου κατηγοριοποιητή για την εκπαίδευση του μοντέλου απαιτεί εμπειρία και γνώσεις σχετικά με τους αλγορίθμους, τον τρόπο λειτουργίας τους και τα τεχνικά τους χαρακτηριστικά. Συχνά, η διαδικασία αυτή καταλήγει σε ένα loop πειραμάτων μέχρι την εύρεση του μοντέλου με τα καλύτερα αποτελέσματα. Υπάρχουνε πολλοί τρόποι να εκτιμήσουμε την ικανότητα ενός προβλεπτικού μοντέλου. Ο πιο διαδεδομένος είναι ο υπολογισμός της ακρίβειας της κατηγοριοποίησης. Η ακρίβεια της κατηγοριοποίησης είναι το ποσοστό των σωστών προβλέψεων- κατηγοριοποιήσεων στο σύνολο των προβλέψεων, δηλαδή

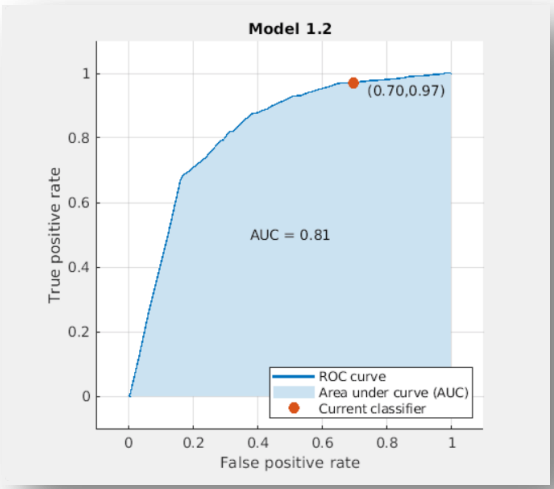
$$accuracy = \frac{correct\ predictions}{total\ predictions} * 100\ .$$

Το MALTAB για την απλούστευση της διαδικασίας αυτής, παρουσιάζει την ακρίβεια κάθε μοντέλου μετά την εκπαίδευση του. Στην περίπτωσή μας, έπειτα από την εκπαίδευση πολλών μοντέλων καταλήξαμε στο ότι ένας medium tree κατηγοριοποιητής μας δίνει το μεγαλύτερο ποσοστό ακρίβειας (εικόνα 9). Σε μία προσπάθεια περαιτέρω βελτίωσης του αποτελέσματος προσθαφαιρέθηκαν κάποια συγκεκριμένα χαρακτηριστικά αλλά δεν παρουσιάστηκε κάποια βελτίωση.



Εικόνα 12. Trained Models

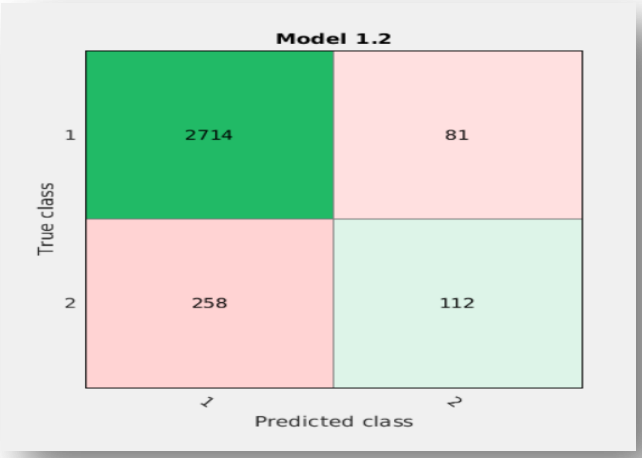
Επομένως, αποδοτικότερος κρίθηκε ο αλγόριθμος Medium Tree με AUC[2]=0,81, δηλαδή η καμπύλη ROC καλύπτει το 81% του γραφήματος.



Εικόνα 13. Καμπύλη ROC

[2] Ο AUC (Area Under Curve) είναι δείκτης που ισοδυναμεί με το εμβαδόν κάτω από την καμπύλη ROC και συνοψίζει την πληροφορία που παρέχει η καμπύλη. Όσο πιο κοντά βρίσκεται στο 1 τόσο πιο καλό είναι το αποτέλεσμα.

Ένας άλλος τρόπος παρουσίασης, κατανόησης και ερμηνείας των αποτελεσμάτων είναι το γράφημα Confusion Matrix. Στο συγκεκριμένο παράδειγμα το γράφημα που παρουσιάστηκε για τα αποτελέσματα του μοντέλου που εκπαιδεύσαμε ήταν το εξής:



Εικόνα 14. Confusion Matrix

Πιο αναλυτικά, το πάνω αριστερά κουτάκι αναφέρει ότι το μοντέλο πρόβλεψε σωστά 2714 παρατηρήσεις ως «όχι», ενώ το κάτω αριστερά ότι πρόβλεψε λάθος 258 παρατηρήσεις ως «όχι».

Μετά την εύρεση και επιλογή του κατάλληλου μοντέλου ακολουθεί η εξαγωγή του, με σκοπό την μελλοντική χρήση του με άλλα δεδομένα. Επιλέγουμε τον τύπο Export Model ο οποίος θα εμφανίσει το μοντέλο που εκπαιδεύσαμε στο Command Window του MATLAB μαζί με οδηγίες για την χρήση του. Κάνουμε export το μοντέλο με το όνομα “trainedModel2”.

```
Variables have been created in the base workspace.
Structure  'trainedModel2' exported from Classification Learner.
To make predictions on a new table, T:
  yfit = trainedModel2.predictFcn(T)
For more information, see How to predict using an exported model.
```

Εικόνα 15. Export Model

Στην συνέχεια κάνοντας χρήση της εντολής predict, η οποία δέχεται ως παραμέτρους το μοντέλο που εκπαιδεύσαμε και το test dataset, παράγουμε τα αποτελέσματα του μοντέλου.

4.3 Αποτελέσματα

Ύστερα από την επιλογή και εξαγωγή του μοντέλου που εκπαιδεύσαμε με την χρήση του Medium Tree κατηγοριοποιητή, εφαρμόστηκε η εντολή predict για να εξακριβωθεί η αποτελεσματικότητα του μοντέλου. Παρατηρήθηκε ότι το μοντέλο που εκπαιδεύσαμε προέβλεψε σωστά 1205 απαντήσεις από τις 1356 των δεδομένων δοκιμής. δηλαδή το 88,8% των απαντήσεων.

Συμπεράσματα

Η Εξόρυξη Δεδομένων αποτελεί ένα πολύ σημαντικό και χρήσιμο εργαλείο στην προσπάθεια μετατροπής τεράστιου όγκου ακατέργαστων δεδομένων σε πολύτιμη γνώση. Πληθώρα από πληροφορίες που παρέμεναν ανεπεξέργαστες στις βάσεις δεδομένων αποκτούν ουσία μέσα από τις διάφορες διεργασίες που προσφέρει η Μηχανική Μάθηση. Ιδιαίτερη είναι η σημασία τους σε κάθε ενέργεια που σχετίζεται με τον κλάδο των Οικονομικών. Χρηματοπιστωτικά ιδρύματα, παραγωγικές επιχειρήσεις και διαφόρων άλλων ειδών οικονομικοί οργανισμοί συλλέγουν αφθονία πληροφοριών που με σωστή ανάλυση γίνονται πλήρως αξιοποιήσιμες σε τομείς των Οικονομικών όπως είναι το μάρκετινγκ, το μάνατζμεντ, η ελεγκτική κ.λπ.

Κύριο σκοπό της παρούσας εργασίας αποτέλεσε η μελέτη των βασικών εννοιών της Εξόρυξης Δεδομένων. Για την πληρέστερη κατανόηση των εννοιών αυτών, πραγματοποιήθηκε πρακτική εφαρμογή των διαδικασιών εξόρυξης γνώσης σε δεδομένα προερχόμενα από τον κλάδο του μάρκετινγκ με τη χρήση του λογισμικού MATLAB. Συγκεκριμένα, δεδομένα που δημιουργήθηκαν στο πλαίσιο τηλεφωνικής καμπάνιας Πορτογαλικής τράπεζας σε πελάτες για τη συλλογή πληροφοριών σχετικά με το αν επιθυμούσαν ή όχι να καταθέσουν σε προθεσμιακό λογαριασμό. Επομένως, επιμέρους σκοπό αποτέλεσε η πρακτική εφαρμογή των εννοιών της Εξόρυξης Δεδομένων και η παρουσίαση του εργαλείου ανάλυσης.

Το αποτέλεσμα της πρακτικής εφαρμογής επιβεβαίωσε την χρησιμότητα της Εξόρυξης Δεδομένων. Επεξεργάστηκαν 4521 παρατηρήσεις με στόχο την πρόβλεψη. Δηλαδή, έγινε προσπάθεια εκπαίδευσης ενός μοντέλου κατηγοριοποίησης ικανού να προβλέψει σωστά αν ένας πελάτης της τράπεζας θα απαντούσε «ναι» ή «όχι» σε μια προθεσμιακή κατάθεση. Το μοντέλο έδειξε ότι μπορεί να προβλέψει αποτελεσματικά το 88.8% των απαντήσεων. Τέτοια μοντέλα μπορούν να εξαχθούν και να εφαρμοστούν για την πρόβλεψη νέων δεδομένων.

Αναφορές

UCI Machine Learning Repository. (2020). Ανάκτηση από Bank Marketing Data Set: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Barak, S., & Modarres, M. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications*.

Brachman, R., & Anand, T. (1994). The Process of Knowledge Discovery in Databases: A First Sketch. *AAAI Press*.

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed Data Mining in Credit Card Fraud Detection. *IEEE Intelligent Systems, 14*(6), σσ. 67-74.

Chiu, S. L. (1997). Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification. *Fuzzy Information Engineering: A Guided Tour of Applications*.

Devale, A., & Kulkarni, R. (2012). APPLICATIONS OF DATA MINING TECHNIQUES IN LIFE INSURANCE. *International Journal of Data Mining & Knowledge Management Process, 2*.

Drucker, P. (2000). *Μετακαπιταλιστική Κοινωνία.* Αθήνα: Gutenberg.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Conceptw and Techniques.* Morgan Kaufmann.

Hongjun, Y., Jianxin, C., Shihuan, T., Zhenkun, L., Yisong, Z., Luqi, H., και συν. (2009). New Drug R&D of Traditional Chinese Medicine: Role of Data Mining Approaches . *Journal of Biological Systems, 17*(3), σσ. 329-347.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*, σσ. 847-856.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2006). Forecasting Fraudulent Financial Statements using Data Minig. *International Journal of Computational Intelligence, 3*(2), σσ. 104-110.

Magnini, V. P., Honeycutt, E. D., & Hodge, J. a. (2003). Data Mining for Hotel Firms: Use and Limitations. *Cornell Hospitality Quarterly (CQ), 44*, σσ. 94-105.

Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, *62*, σσ. 22-31.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*, σσ. 464-473.

Sabitha, B., Bhuvaneswari Amma, N., Annapoorani, G., & Balasubramanian, P. (2014). Implementation of Data Mining Techniques to Perform Market Analysis. *International Journal of Innovative Research in Computer, 2*(11).

Χαλκίδη, Μ., & Βαζιργιάννης, Μ. (2005). *Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό.* Αθήνα: Τυπωθήτω.