ΤΕΧΝΟΛΟΓΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΧΡΗΣΗ ΠΡΟΒΛΕΨΗΣ ΣΥΜΠΕΡΙΦΟΡΑΣ ΤΟΥ ΚΑΤΑΝΑΛΩΤΗ

ΟΝΟΜΑΤΕΠΩΝΥΜΑ: ΑΛΗ ΚΕΧΑΓΙΑ ΕΣΜΑ , ΚΟΥΤΣΟΥΚΟΥ ΑΘΑΝΑΣΙΑ , ΡΑΛΛΗΣ ΑΠΟΣΤΟΛΟΣ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ: ΠΑΠΑΔΟΠΟΥΛΟΣ ΔΗΜΗΤΡΙΟΣ

ΠΑΤΡΑ 2018

Περιεχόμενα

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία θα επιχειρήσουμε μια επισκόπηση στις τεχνικές εξόρυξης δεδομένων, που έχουν εφαρμογή στην οικονομία, για την εύρεση μοτίβων στην συμπεριφορά των καταναλωτών που να είναι χρήσιμα για μελλοντική εκμετάλλευση.

Αρχικά, στο πρώτο κεφάλαιο γίνεται παρουσίαση των στατιστικών εργαλείων της εξόρυξης δεδομένων, αλλά και ανάλυση της μεθόδου Market Basket Analysis που θα χρησιμοποιήσουμε για το πρακτικό μέρος.

Στο δεύτερο κεφάλαιο αναλύουμε τα βασικά στοιχεία της εξόρυξης δεδομένων, όπως η προεπεξεργασία και τα βασικά μοντέλα.

Στο τρίτο κεφάλαιο εισαγόμαστε σε πιο ειδικά θέματα της μεθόδου, όπως τα είδη των δεδομένων και η μορφή που πρέπει να έχουν ώστε να εξορυχθούν. Από αυτά μας ενδιαφέρουν περισσότερο τα δεδομένα συναλλαγών.

Στο πρακτικό μέρος γίνεται επίδειξη εξόρυξης δεδομένων για εύρεση κανόνων συσχέτισης σε δεδομένα αγορών μέσω του προγράμματος Weka. Τέλος, παρουσιάζονται τα συμπεράσματα και η βιβλιογραφία που χρησιμοποιήθηκε σε κάθε κεφάλαιο.

ABSTRACT

In this research we will attempt an overview of data mining techniques that can be applied in the economy to find patterns in consumer behavior that are useful for future exploitation.

Initially, the first chapter presents the statistical tools of data mining, but also an analysis of the Market Basket Analysis method that we will use for the practical part.

In the second chapter we analyze the key elements of data mining, such as pre-processing and basic models.

In the third chapter we are introducing more specific aspects of the method, such as the types of data and the form they need to have to be extracted. Of these, we are more interested in trading data.

The practical part demonstrates data mining to find correlation rules for purchase data through the Weka program. At the end, the conclusions and bibliography used in each chapter are presented.

ΚΕΦΑΛΑΙΟ 1: ΕΙΔΗ ΔΕΔΟΜΕΝΩΝ

1.1 Ποια είδη δεδομένων μπορούν να εξορυχτούν;

Ως γενική τεχνολογία, η εξόρυξη δεδομένων μπορεί να εφαρμοστεί σε οποιοδήποτε είδος δεδομένων, εφόσον τα δεδομένα έχουν νόημα για μια εφαρμογή-στόχο. Οι βασικότερες μορφές δεδομένων για εφαρμογές εξόρυξης είναι δεδομένα βάσης δεδομένων, δεδομένα αποθήκης δεδομένων και δεδομένα συναλλαγών. Οι έννοιες και οι τεχνικές που παρουσιάζονται σε αυτό το κεφάλαιο επικεντρώνονται σε τέτοια δεδομένα. Η εξόρυξη δεδομένων μπορεί επίσης να εφαρμοστεί σε άλλες μορφές δεδομένων (π.χ. ροές δεδομένων, δεδομένα παραγγελιών / ακολουθιών, δεδομένα γραφήματος ή δικτύου, χωρικά δεδομένα, δεδομένα κειμένου, δεδομένα πολυμέσων και WWW) (Glymour et al., 1997).

1.1.1 Δεδομένα βάσης δεδομένων

Ένα σύστημα βάσης δεδομένων, το οποίο ονομάζεται επίσης σύστημα διαχείρισης βάσεων δεδομένων (DBMS), αποτελείται από μια συλλογή αλληλένδετων δεδομένων, γνωστή ως βάση δεδομένων, και ένα σύνολο προγραμμάτων λογισμικού για τη διαχείριση και πρόσβαση στα δεδομένα. Τα προγράμματα λογισμικού παρέχουν μηχανισμούς για τον ορισμό δομών βάσεων δεδομένων και αποθήκευσης δεδομένων. Τα εργαλεία αυτά χρησιμοποιούνται για τον προσδιορισμό και τη διαχείριση ταυτόχρονων, κοινόχρηστων ή κατανεμημένων δεδομένων. Υπάρχουν και εργαλεία άμυνας του συστήματος που χρησιμοποιούνται για τη διασφάλιση της συνέπειας και της ασφάλειας των πληροφοριών που αποθηκεύονται κυρίως ενάντια στις προσπάθειες μη εξουσιοδοτημένης πρόσβασης.

Μια σχεσιακή βάση δεδομένων είναι μια συλλογή από πίνακες, κάθε ένας από τους οποίους έχει ένα μοναδικό όνομα. Κάθε πίνακας αποτελείται από ένα σύνολο χαρακτηριστικών (στήλες ή πεδία) και συνήθως αποθηκεύει ένα μεγάλο σύνολο πλειάδων (εγγραφές ή σειρές). Κάθε πλειάδα σε ένα σχεσιακό πίνακα αντιπροσωπεύει ένα αντικείμενο που αναγνωρίζεται από ένα μοναδικό χαρακτηριστικό και περιγράφεται από ένα σύνολο τιμών του χαρακτηριστικού. Ένα μοντέλο σημασιολογικών δεδομένων, όπως ένα μοντέλο δεδομένων οντότητας-σχέσης (ER), συχνά κατασκευάζεται για σχεσιακές βάσεις δεδομένων. Ένα μοντέλο δεδομένων ER αντιπροσωπεύει τη βάση δεδομένων ως σύνολο οντοτήτων και τις σχέσεις τους(Glymour et al., 1997).

Έστω ότι έχουμε μια σχεσιακή βάση δεδομένων για την AllElectronics. Η εταιρεία περιγράφεται από τους ακόλουθους πίνακες σχέσεων: πελάτης, στοιχείο, υπάλληλος και υποκατάστημα. Τα παρακάτω φαίνονται στο σχήμα 2.1.

customer	(*cust_ID, name, address, age, occupation, annual_income, credit_information, category, ...*)
item	(*item_ID, brand, category, type, price, place_made, supplier, cost, ...*)
employee	(*empl_ID, name, category, group, salary, commission, ...*)
branch	(*branch_ID, name, address, ...*)
purchases	(*trans_ID, cust_ID, empl_ID, date, time, method_paid, amount*)
items_sold	(*trans_ID, item_ID, qty*)
works_at	(*empl_ID, branch_ID*)

Σχήμα 1.1 Παράδειγμα σχεσιακής βάσης δεδομένων

Στο παράδειγμα αυτό ο πελάτης περιγράφεται από ένα σύνολο χαρακτηριστικών, συμπεριλαμβανομένων τον αριθμό ταυτότητας του πελάτη (Cust_ID), το όνομα του πελάτη, τη διεύθυνση, την ηλικία, το επάγγελμα, το ετήσιο εισόδημα και πιστωτικές πληροφορίες.

Επίσης, κάθε στοιχείο χρήσιμων σχέσεων, όπως ονόματα των εργαζομένων, και κλάδος παρουσιάζεται από ένα σύνολο δεδομένων που περιγράφουν τις ιδιότητες αυτών των οντοτήτων.

Οι πίνακες μπορεί επίσης να χρησιμοποιηθούν για να αντιπροσωπεύουν τις σχέσεις μεταξύ των ή μεταξύ πολλαπλών οντοτήτων. Στο παράδειγμά μας, οι τιμές αυτές περιλαμβάνουν τις αγορές (στοιχεία αγορών πελατών, δημιουργώντας μια συναλλαγή πώλησης που χειρίζεται ένας υπάλληλος), και το itemssold (παραθέτει στοιχεία για μια δεδομένη συναλλαγή). Επίσης έχουμε την κατηγορία custAD, δηλαδή χαρακτηριστικά πελάτη που περιέχει όνομα, διεύθυνση, ηλικία, επάγγελμα κτλ(Hand et al., 2001).

Τα σχεσιακά δεδομένα μπορούν να παρουσιαστούν γραμμένα σε μια ειδική γλώσσα (π.χ. SQL) ή με τη βοήθεια γραφικών διεπαφών χρήστη. Ένα δεδομένο μετατρέπεται σε ένα σύνολο σχεσιακών λειτουργιών, όπως η σύνδεση, η επιλογή και η προβολή, και στη συνέχεια βελτιστοποιείται για αποτελεσματική επεξεργασία. Ας υποθέσουμε ότι η εργασία μας είναι να αναλύσουμε τα δεδομένα της AllElectronics. Μέσω της χρήσης των σχεσιακών ερωτημάτων μπορούμε να λάβουμε πληροφορίες όπως μια λίστα όλων των αντικειμένων που πωλήθηκαν το τελευταίο τρίμηνο. Οι συσχετιστικές γλώσσες χρησιμοποιούν επίσης συναρτήσεις όπως το άθροισμα, το μέσο όρο, ο αριθμός, το μέγιστο και το ελάχιστο (minimum) .Η χρήση των αθροισμάτων σάς επιτρέπει να βρούμε απαντήσεις σε ερωτήματα όπως ποιες ήταν οι συνολικές πωλήσεις του τελευταίου μήνα, ομαδοποιημένες κατά υποκατάστημα ή πόσες πωλήσεις πραγματοποιήθηκαν τον μήνα Δεκέμβριο ή ποιος πωλητής είχε την υψηλότερη αποδοτικότητα.

Όταν εξάγουμε σχεσιακές βάσεις δεδομένων, μπορούμε να προχωρήσουμε περαιτέρω αναζητώντας τάσεις ή πρότυπα δεδομένων. Για παράδειγμα, τα συστήματα εξόρυξης δεδομένων μπορούν να αναλύσουν τα δεδομένα πελατών για να προβλέψουν τον πιστωτικό κίνδυνο νέων πελατών βάσει του εισοδήματός τους, της ηλικίας τους και των προηγούμενων πιστωτικών στοιχείων. Τα συστήματα εξόρυξης δεδομένων ενδέχεται επίσης να εντοπίζουν αποκλίσεις και στοιχεία για τις

πωλήσεις που απέχουν πολύ από τις αναμενόμενες σε σχέση με το προηγούμενο έτος. Τέτοιες αποκλίσεις μπορούν στη συνέχεια να διερευνηθούν περαιτέρω. Για παράδειγμα, η εξόρυξη δεδομένων μπορεί να ανακαλύψει ότι υπήρξε αλλαγή στη συσκευασία ενός στοιχείου ή σημαντική αύξηση της τιμής.

Οι σχεσιακές βάσεις δεδομένων είναι από τις πιο διαδεδομένες και πλουσιότερες αποθήκες πληροφοριών και ως εκ τούτου αποτελούν μια σημαντική μορφή δεδομένων στη μελέτη της εξόρυξης δεδομένων(Hand et al., 2001).

1.1.2 Αποθήκες δεδομένων

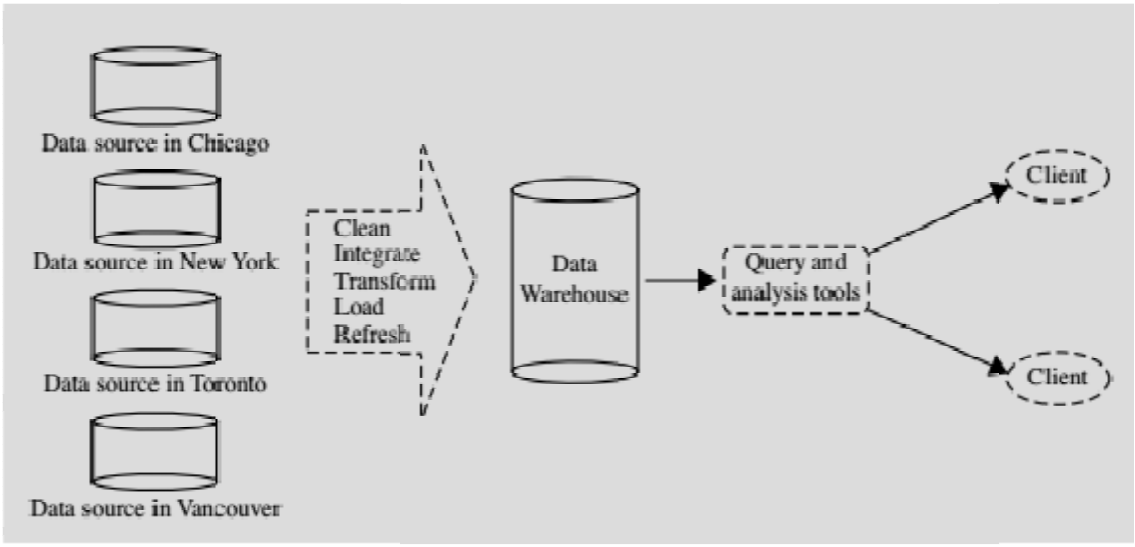
Ας υποθέσουμε ότι η AllElectronics είναι μια επιτυχημένη διεθνής εταιρεία με υποκαταστήματα σε όλο τον κόσμο. Κάθε υποκατάστημα έχει το δικό του σύνολο βάσεων δεδομένων. Ο πρόεδρος της AllElectronics ζήτησε να παράσχετε μια ανάλυση των πωλήσεων της εταιρείας ανά είδος είδους ανά κατάστημα για το τρίτο τρίμηνο. Πρόκειται για ένα δύσκολο έργο, ιδίως επειδή τα σχετικά δεδομένα διαδίδονται σε διάφορες βάσεις δεδομένων που βρίσκονται φυσικά σε πολυάριθμους χώρους.

Εάν η AllElectronics είχε μια αποθήκη δεδομένων, αυτή η εργασία θα ήταν εύκολη. Μια αποθήκη δεδομένων είναι ένα αποθετήριο πληροφοριών που συλλέγονται από πολλαπλές πηγές, αποθηκεύονται κάτω από ένα ενοποιημένο σύστημα και συνήθως καταχωρούνται σε έναν ενιαίο ιστότοπο. Οι αποθήκες δεδομένων κατασκευάζονται μέσω διαδικασίας καθαρισμού δεδομένων, ενσωμάτωσης δεδομένων, μετασχηματισμού δεδομένων, φόρτωσης δεδομένων και ανανέωσης περιοδικών δεδομένων.

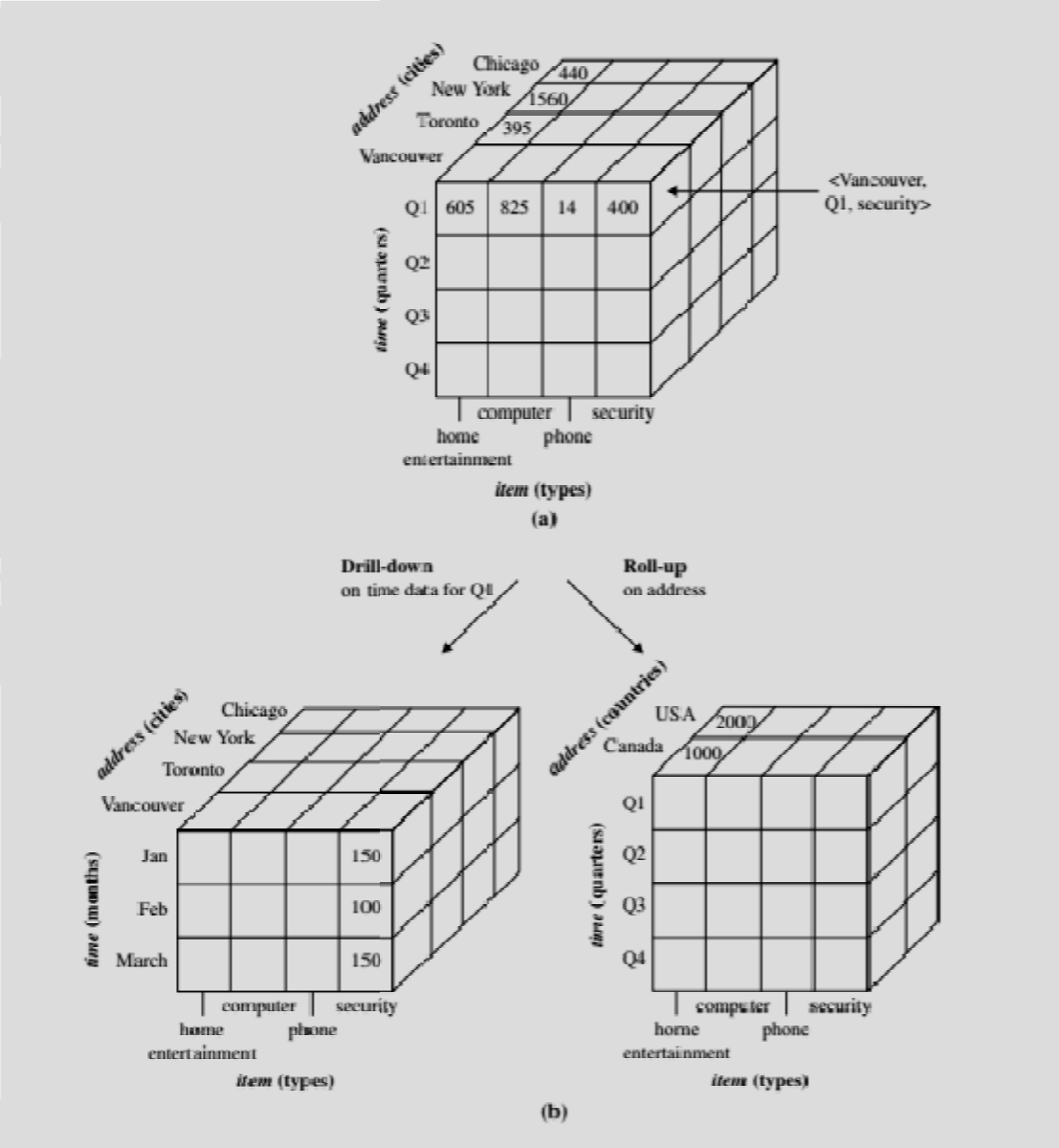
Για τη διευκόλυνση της λήψης αποφάσεων, τα δεδομένα σε μια αποθήκη δεδομένων είναι οργανωμένα γύρω από σημαντικά θέματα (π.χ. πελάτης, στοιχείο,

προμηθευτής και δραστηριότητα). Τα δεδομένα αποθηκεύονται για να παράσχουν πληροφορίες σε βάθος χρόνου, όπως στους τελευταίους 6 έως 12 μήνες, και συνήθως συνοψίζονται. Για παράδειγμα, αντί να αποθηκεύει τις λεπτομέρειες κάθε συναλλαγής πωλήσεων, η αποθήκη δεδομένων μπορεί να αποθηκεύει μια σύνοψη των συναλλαγών ανά είδος είδους για κάθε κατάστημα ή, συνοψισμένη σε υψηλότερο επίπεδο, για κάθε περιοχή πωλήσεων(Hastie et al., 2001).

Μια αποθήκη δεδομένων συνήθως διαμορφώνεται από μια πολυδιάστατη δομή δεδομένων, που ονομάζεται κύβος δεδομένων, όπου κάθε διάσταση αντιστοιχεί σε ένα χαρακτηριστικό ή σε ένα σύνολο χαρακτηριστικών στο σύστημα και κάθε κελί εμπεριέχει την τιμή κάποιου μέτρου.



Σχήμα 1.2 Παράδειγμα τρόπου λειτουργίας μιας αποθήκης δεδομένων

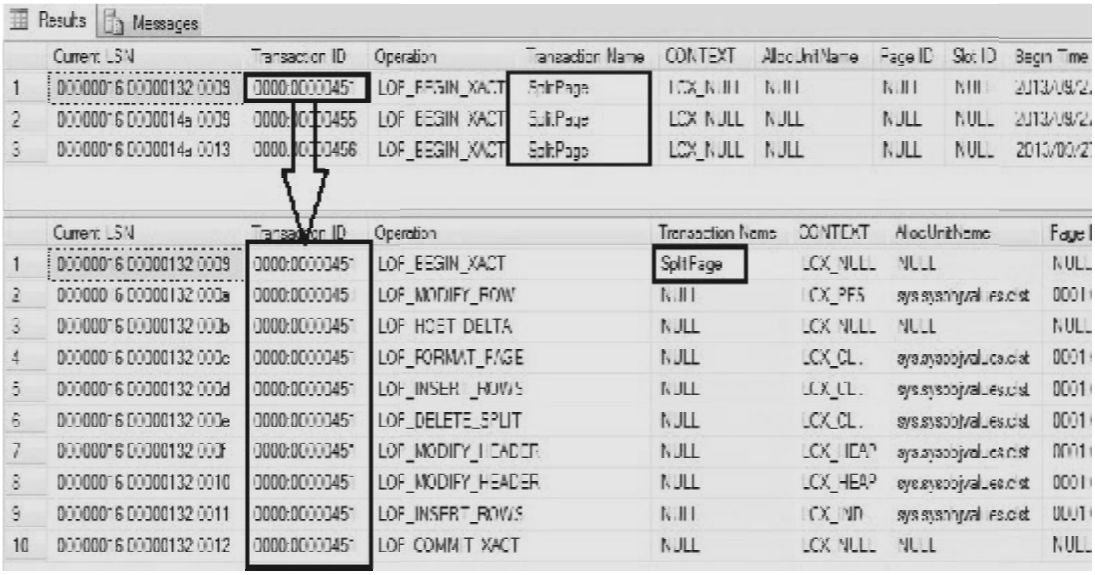


Σχήμα 1.3 Κύβοι δεδομένων

1.1.3 Δεδομένα συναλλαγών

Γενικά, κάθε εγγραφή σε μια βάση δεδομένων συναλλαγών καταγράφει μια συναλλαγή, όπως την αγορά ενός πελάτη, μια κράτηση πτήσης ή τα κλικ ενός

χρήστη σε μια ιστοσελίδα. Μια συναλλαγή τυπικά περιλαμβάνει έναν μοναδικό αριθμό ταυτότητας συναλλαγής (transJD) και έναν κατάλογο των στοιχείων που αποτελούν τη συναλλαγή, όπως τα στοιχεία που αγοράστηκαν στη συναλλαγή. Μια μεταβατική βάση δεδομένων μπορεί να έχει επιπλέον πίνακες, οι οποίοι περιέχουν άλλες πληροφορίες σχετικά με τις συναλλαγές, όπως περιγραφή στοιχείου, πληροφορίες σχετικά με τον πωλητή ή τον κλάδο κ.ο.κ. (Hastie et al., 2001).



Σχήμα 1.4 Βάση δεδομένων συναλλαγών

1.1.4 Άλλοι τύποι δεδομένων

Εκτός από τα δεδομένα σχεσιακής βάσης δεδομένων, τα δεδομένα αποθήκης δεδομένων και τα δεδομένα συναλλαγών, υπάρχουν πολλά άλλα είδη δεδομένων που έχουν ευπροσάρμοστες μορφές και δομές, και μάλλον διαφορετικές σημασίες. Αυτά τα είδη δεδομένων μπορούν να παρατηρηθούν σε πολλές εφαρμογές: δεδομένα χρόνου ή ακολουθίας (π.χ. ιστορικά αρχεία, δεδομένα χρηματιστηρίου και δεδομένα χρονοσειράς και βιολογικής ακολουθίας), ροές δεδομένων (π.χ. δεδομένα

παρακολούθησης βίντεο και αισθητήρων, (π.χ. σχεδιασμός κτιρίων, εξαρτημάτων συστημάτων ή ολοκληρωμένων κυκλωμάτων), δεδομένα πολυμέσων (συμπεριλαμβανομένων των δεδομένων κειμένου, εικόνων, βίντεο και ήχου), χωρικά δεδομένων, δικτυακά δεδομένα (π.χ. δεδομένα από κοινωνικά δίκτυα και δίκτυα πληροφόρησης), καθώς και πολλά άλλα είδη δεδομένων που μπορούμε να βρούμε στον παγκόσμιο ιστό. Αυτές οι εφαρμογές δημιουργούν νέες προκλήσεις, όπως το χειρισμό δεδομένων που μεταφέρουν ειδικές δομές (π.χ. αλληλουχίες, δέντρα, γραφήματα και δίκτυα) και συγκεκριμένη σημασιολογία (όπως η παραγγελία, η εικόνα, το περιεχόμενο ήχου και βίντεο και η συνδεσιμότητα) (Grabmeier & Rudolph, 2002) .

Μπορεί να παραχθεί ένα μεγάλο εύρος γνώσης και σημαντικών πληροφοριών από αυτά τα είδη δεδομένων. Όσον αφορά τα χρονικά δεδομένα, για παράδειγμα, μπορούμε να εξάγουμε τραπεζικά δεδομένα για τις μεταβαλλόμενες τάσεις, οι οποίες μπορούν να βοηθήσουν στον προγραμματισμό των τραπεζών σε συνάρτηση με τον όγκο της κίνησης των πελατών. Τα στοιχεία της χρηματιστηριακής αγοράς μπορούν να αποκαλύψουν τις τάσεις που θα μπορούσαν να βοηθήσουν να σχεδιάσει ο χρήστης επενδυτικές στρατηγικές.

Με τα χωρικά δεδομένα, μπορούμε να αναζητήσουμε μοτίβα που περιγράφουν τις μεταβολές στα ποσοστά φτώχειας με βάση τις αποστάσεις των πόλεων από τις μεγάλες εθνικές οδούς. Οι σχέσεις μεταξύ ενός συνόλου χωρικών αντικειμένων μπορούν να εξεταστούν προκειμένου να ανακαλυφθούν ποια υποσύνολα αντικειμένων είναι χωρικά αυτοσυσχετιζόμενα ή συσχετιζόμενα.

Με την εξόρυξη δεδομένων κειμένου, μπορούμε να εντοπίσουμε την εξέλιξη θεμάτων που ενδιαφέρουν σε κάποιον επιστημονικό τομέα. Με την εξόρυξη σχολίων χρηστών σχετικά με προϊόντα (τα οποία συχνά υποβάλλονται ως σύντομα μηνύματα κειμένου), μπορούμε να αξιολογήσουμε τα συναισθήματα των πελατών και να κατανοήσουμε πόσο καλά ένα προϊόν αφομοιώνεται από μια αγορά. Από τα δεδομένα πολυμέσων, μπορούμε να εξορύξουμε εικόνες για να εντοπίσουμε αντικείμενα και να τα ταξινομήσουμε θέτοντας σημασιολογικές ετικέτες.

Η εξόρυξη ιστού μπορεί να μας βοηθήσει να μάθουμε για την κατανομή συγκεκριμένων πληροφοριών στο WWW γενικότερα, να χαρακτηρίσουμε και να ταξινομήσουμε ιστοσελίδες και να αποκαλύψουμε τη δυναμική του ιστού και τις σχέσεις σύνδεσης μεταξύ διαφορετικών ιστοσελίδων, χρηστών, κοινοτήτων και δραστηριοτήτων στο διαδίκτυο.

Είναι σημαντικό να έχουμε κατά νου ότι σε πολλές εφαρμογές υπάρχουν πολλοί τύποι δεδομένων. Για παράδειγμα, στην εξόρυξη ιστού, συχνά υπάρχουν δεδομένα κειμένου και δεδομένα πολυμέσων (π.χ. εικόνες και βίντεο) σε ιστοσελίδες, δεδομένα γραφημάτων, όπως διαγράμματα ιστού, και δεδομένα χάρτη σε κάποιες τοποθεσίες Web. Στη βιοπληροφορική, γονιδιωματικές αλληλουχίες, βιολογικά δίκτυα και 3-D χωρικές δομές των γονιδιωμάτων μπορούν να συνυπάρχουν σε πολλά βιολογικά μόρια. Η εξόρυξη πολλαπλών και περίπλοκων δεδομένων συχνά οδηγεί σε καρποφόρα ευρήματα λόγω της εύρεσης συσχετίσεων μεταξύ φαινομενικά μη συσχετιζόμενων δεδομένων. Από την άλλη πλευρά, είναι επίσης δύσκολη διαδικασία στον εντοπισμό, στον καθαρισμό δεδομένων και την ενσωμάτωση των δεδομένων, καθώς και στην επεξεργασία τους, καταστάσεις που οφείλονται στις συχνά πολύπλοκες αλληλεπιδράσεις μεταξύ των πολλαπλών πηγών αυτών των δεδομένων.

Παρόλο που τα δεδομένα αυτά απαιτούν εξελιγμένες εγκαταστάσεις για αποτελεσματική αποθήκευση, ανάκτηση και ενημέρωση, παρέχουν επίσης εύφορο έδαφος και προκλήσεις έρευνας για την εξόρυξη δεδομένων.

1.2 Ποια είδη μοτίβων και δομών μπορούν να αναγνωριστούν;

Παρατηρήσαμε διάφορους τύπους αποθετηρίων δεδομένων και πληροφοριών στους οποίους μπορεί να εκτελεστεί η εξόρυξη δεδομένων. Ας εξετάσουμε τώρα τα είδη των μορφών που μπορούν να αναγνωριστούν.

Υπάρχουν διάφορα είδη μοτίβων που μπορούν να βρεθούν στις διαδικασίες της εξόρυξης δεδομένων. Γενικά, τα μοτίβα αυτά καθορίζονται από τα μοντέλα που χρησιμοποιούμε και έχουμε αναλύσει ήδη, και μπορούν να ταξινομηθούν σε δύο κατηγορίες: περιγραφικά και προγνωστικά(Fayyad et al., 1996).

ΚΕΦΑΛΑΙΟ 2: ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ (DATA MINING)

Εισαγωγή

Με την αύξηση της χρήσης του διαδικτύου εδώ και μια δεκαετία έχει αυξηθεί και ο αριθμός δεδομένων που αποθηκεύονται. Τα δεδομένα αυτά μπορούν να έχουν πολλές χρήσεις, και μια από αυτές είναι η πρόβλεψη της συμπεριφοράς των καταναλωτών.

Πιο ειδικά, στην περίπτωση μας είναι τα δεδομένα που καταγράφονται προέρχονται από αγορές σε καταστήματα λιανικής πώλησης. Τα δεδομένα αυτά συσσωρεύονται, αλλά σε παλαιότερες εποχές δεν υπήρχε η δυνατότητα επεξεργασίας των δεδομένων αυτών.

Κατά την διαδικασία εξόρυξης δεδομένων τα δεδομένα αποθηκεύονται ηλεκτρονικά και η κατηγοριοποίηση τους είναι αυτοματοποιημένη από υπολογιστή. Με αυτό τον τρόπο γίνεται αυτοματοποίηση της διαδικασίας, και εμείς πλέον καλούμαστε να βρούμε συσχετίσεις και μοτίβα εντός των δεδομένων αυτών.

Μία βάση δεδομένων με τις επιλογές των αγοραστών και μία βάση με τα profile τους μπορεί να δώσει εξαιρετικά χρήσιμες πληροφορίες για το πώς μπορεί να συνδεθεί ένα χαρακτηριστικό ενός αγοραστή με ένα προϊόν(Trnka, 2010).

Η τεχνική που θα χρησιμοποιήσουμε ονομάζεται Market Basket Analysis, και αναφέρεται στην χρήση δεδομένων αγορών καθημερινών προϊόντων για την εύρεση συσχετίσεων μεταξύ τους. Για παράδειγμα αν ένας αγοραστής επιλέγει ένα προϊόν Α, ποια είναι η πιθανότητα να επιλέξει στην ίδια αγορά και το προϊόν Β. Η γνώση της συσχέτισης μπορεί να βοηθήσει στην καλύτερη προώθηση των προϊόντων.

Ωστόσο, επειδή τα δεδομένα αυτά είναι δύσκολα προσβάσιμα, εμείς στην παρούσα εργασία θα παρουσιάσουμε τον τρόπο με τον οποίο λειτουργεί η συγκεκριμένη τεχνική και θα υλοποιήσουμε υποθετικό παράδειγμα.

2.1 Ορισμοί

Με τη εξόρυξη δεδομένων οι επιστήμονες και οι αναλυτές προσπαθούν να εξάγουν γνώση από ήδη καταχωρημένα δεδομένα. Υπάρχουν όμως αρκετοί πιο συγκεκριμένοι ορισμοί, που δίνουν σαφέστερα όρια σε αυτό που αποκαλούμε data mining. Στην πραγματικότητα η μεγαλύτερη συλλογή δεδομένων που έχει γίνει ποτέ, είναι μέσω ηλεκτρονικών συσκευών, και στην εξόρυξη δεδομένων χρησιμοποιούμε υπολογιστές για την επεξεργασία του τεράστιου όγκου δεδομένων που έχουν συσσωρευτεί(Trnka, 2010).

Η πληροφορία που μπορούμε να εξάγουμε από τα δεδομένα δεν προκύπτει αυτόματα.

Σύμφωνα με την Wikipedia "Εξόρυξη δεδομένων (ή ανακάλυψη γνώσης από βάσεις δεδομένων) είναι η εξεύρεση μιας (ενδιαφέρουσας, αυτονόητης, μη προφανής και πιθανόν χρήσιμης) πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις". Η εξόρυξη δεδομένων που αλλιώς λέγεται και ανακάλυψη γνώσης σε βάσεις δεδομένων ορίζεται ως η διαδικασία της ανακάλυψης χρήσιμων μοτίβων ή γνώση μέσα από πηγές δεδομένων όπως βάσεις δεδομένων, κείμενα, εικόνες και το Διαδίκτυο. Τα μοτίβα πρέπει να είναι επαληθευμένα, χρήσιμα και κατανοητά. Η εξόρυξη δεδομένων είναι ένα πολυσυλλεκτικό πεδίο που εμπεριέχει εκμάθηση μηχανών, στατιστική, εξόρυξη πληροφορίας, και αστικοποίηση. Τρεις όροι κλειδιά για την θεωρία της εξόρυξης δεδομένων, οι οποίοι πρέπει να αποσαφηνιστούν και να διαφοροποιηθεί η σημασία τους είναι τα δεδομένα, η πληροφορία και η γνώση:

Δεδομένα

Ως δεδομένο ονομάζουμε κάθε στοιχείο ή οντότητα το οποίο μπορεί να εισαχθεί σε ένα υπολογιστικό σύστημα προς επεξεργασία. Πλέον, υπάρχουν τεράστιες βάσεις δεδομένων στο διαδίκτυο από διαφορετικές ανθρώπινες δραστηριότητες που όταν μπουν σε επεξεργασία μπορούμε να εξάγουμε σημαντικά συμπεράσματα. Τέτοιες βάσεις δεδομένων μπορεί να έχουν:

- Επιχειρησιακά δεδομένα που είναι δεδομένα πωλήσεων, προτιμήσεων σε πωλήσεις, λογιστικής και οικονομικής φύσεως
- Τα μη επιχειρησιακά δεδομένα, όπως δεδομένα χρήσιμα για προβλέψεις ή χρήσιμα για υπολογισμό μακροοικονομικών δεικτών
- Τα μεταδεδομένα που είναι ένα είδος δεδομένων που περιγράφουν άλλα δεδομένα

Πληροφορία

Μετά την επεξεργασία των δεδομένων μπορούμε να λάβουμε συνδυαστικά αποτελέσματα που μπορούν με τη σειρά τους να χρησιμοποιηθούν για την παραγωγή γνώσης. Από τα δεδομένα μπορούμε να αναγνωρίσουμε:

- Μοτίβα (patterns)
- Συσχετίσεις (associations)
- Συνάφειες (relationships) (Weiss et al., 1991)

Για παράδειγμα, σε αυτή την εργασία θα προσπαθήσουμε να εξάγουμε συμπεράσματα πάνω στην ανάλυση των δεδομένων που λαμβάνουμε από αγορές. Πχ, ένα προϊόν Α μπορεί να βρίσκεται σε μια κοινή αγορά με το προϊόν Β σε πολύ μεγάλη πιθανότητα. Η πληροφορία αυτή μπορεί να μας οδηγήσει σε παραγωγή γνώσης πάνω στην συμπεριφορά του καταναλωτή.

Γνώση

Η ανάλυση των πληροφοριών μπορεί αν μας οδηγήσει με τη σειρά της στη γνώση. Για παράδειγμα με ανάλυση των πληροφοριών ότι υπάρχει συσχέτιση κάποιων χαρακτηριστικών ενός προϊόντος με κάποιο είδος πελάτη τότε μπορούμε να εξάγουμε γνώση για τις μελλοντικές πωλήσεις. Ακολούθως μπορεί να σχεδιαστεί ένα προϊόν με τα χαρακτηριστικά που ζητάει συγκεκριμένο καταναλωτικό κοινό(Wang et al., 2007).

2.2 Στάδια εξόρυξης δεδομένων

Πριν αναλύσουμε τα κύρια στάδια εξόρυξης δεδομένων θα αναλύσουμε τα στάδια εξεύρεσης κάποιας μορφής γνώσης από δεδομένων. Υπάρχουν τα παρακάτω γενικά βήματα, τα οποία δεν εφαρμόζονται απαραίτητα ωστόσο σε κάθε προσπάθεια ανεύρεσης γνώσης:

Αποθορυβοποίηση των δεδομένων

Αυτή διαδικασία γίνεται για την αφαίρεση του θορύβου, δηλαδή μη πραγματικών δεδομένων ή δεδομένων με τιμές που δεν ταιριάζουν με τις τιμές των υπόλοιπων που μπορούν να διαστρεβλώσουν τα αποτέλεσμα.

Ενσωμάτωση δεδομένων

Η διαδικασία αυτή αναφέρεται στην κατάσταση που τα δεδομένα δεν είναι σε μια κοινή βάση και δεν είναι ακόμα έτοιμα προς επεξεργασία. Ενσωματώνονται σε μια κοινή βάση και αποκτούν την ίδια μορφή.

Επιλογή των δεδομένων

Από τα διαθέσιμα δεδομένα ένα μεγάλο μέρος μπορεί να μην είναι χρήσιμα για την ανάλυση μας. Επιλέγονται τα δεδομένα που θα μπουν προς επεξεργασία για την ανεύρεση συσχετίσεων

Τροποποίηση δεδομένων

Για να εισαχθούν τα δεδομένα σε ένα υπολογιστικό σύστημα χρειάζεται να βρίσκονται σε μια κατάσταση που να μπορεί το σύστημα να τα αναγνωρίσει και να τα επεξεργαστεί.

Εξόρυξη δεδομένων

Είναι το σημαντικότερο από τα βήματα της διαδικασίας και αυτό γιατί στο συγκεκριμένο στάδιο, ποικίλες εξελιγμένες τεχνικές χρησιμοποιούνται για την εξαγωγή δυνητικά χρήσιμων προτύπων.

Αξιολόγηση προτύπων

Στο βήμα αυτό αναγνωρίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει συγκεκριμένων μέτρων αξιολόγησης (evaluation measures).

Αναπαράσταση γνώσης

Στο τελικό αυτό στάδιο, η γνώση που έχει ανακαλυφθεί παρουσιάζεται στον χρήστη, βοηθώντας τον έτσι να κατανοήσει και να ερμηνεύσει τα αποτελέσματα της εξόρυξης δεδομένων.

Κατά την εξόρυξη δεδομένων υπάρχουν πιο συγκεκριμένα στάδια που χρησιμοποιούμε τα οποία είναι περισσότερα ειδική φύσεως, και σχετίζονται με τον τρόπο που θα εκτελεστεί το πρακτικό μέρος της παρούσας εργασίας. Συνήθως λοιπόν, μπορούμε να ξεχωρίσουμε αυτά τα πέντε βασικά στάδια:

- Συλλογή δεδομένων
- Προ επεξεργασία δεδομένων
- Μετασχηματισμός δεδομένων
- Εξόρυξη δεδομένων
- Ερμηνεία και Αξιολόγηση

Κατά την ερμηνεία του Cross Industry Standard Process for Data Mining (CRISP-DM) υπάρχουν έξι και όχι πέντε διαφορετικά στάδια, που παραθέτουμε παρακάτω:

- Κατανόηση προβλήματος
- Κατανόηση μορφής των δεδομένων
- Προεργασία
- Μοντελοποίηση
- Αξιολόγηση

Τέλος, μπορούμε να χωρίσουμε τη διαδικασία σε τρία βασικά στάδια, όπως φαίνεται παρακάτω:

- Προ-επεξεργασία

- Εξόρυξη δεδομένων
- Επικύρωση αποτελέσματος(Xie Wen-xiu et al., 2010)

2.2.1 Προεργασία

Πριν την κύρια διαδικασία εξόρυξης δεδομένων θα πρέπει να ελέγξουμε το σύνολο δεδομένων που έχουμε. Στη διαδικασία που ονομάζουμε προεργασία εμπεριέχεται και η αποθορυβοποίηση των δεδομένων, δηλαδή η απόρριψη των τιμών αυτών που φαίνεται να μην είναι αληθείς η να μην ταιριάζουν με τα υπόλοιπα δεδομένα. Το σύνολο των δεομένων επίσης θα πρέπει να είναι αρκετά μεγάλο ώστε να υπάρχει αξιοπιστία στις συσχετίσεις που πιθανώς εξαχθούν, γιατί μικρό σύνολο δεδομένων μπορεί να μην είναι αρκετό για την εύρεση μιας συσχέτισης ή να εμφανίζονται συσχετίσεις τυχαίες με μικρή ισχύ(Σταυλιώτης, 2009).

2.2.2 Επεξεργασία των δεδομένων

Κατά την επεξεργασία των δεδομένων προσπαθούμε μέσω αλγορίθμων να αναγνώσουμε που βρίσκονται συσχετίσεις μεταξύ των δεδομένων. Σε αυτή τη διαδικασία μπορούμε να παρατηρήσουμε τις παρακάτω διεργασίες:

Ανίχνευση ανωμαλιών (Anomaly detection): Ο προσδιορισμός ασυνήθιστων εγγραφών δεδομένων, που μπορεί να παρουσιάζουν κάποιο ενδιαφέρον ή λάθη στα δεδομένα που απαιτούν περαιτέρω έρευνα.

Κατηγοριοποίηση: Είναι η διαδικασία γενίκευσης γνωστών δομών για την εφαρμογή της πάνω σε νέα δεδομένα. Για παράδειγμα, ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου ενδέχεται να προσπαθήσει να χαρακτηρίσει ένα μήνυμα ηλεκτρονικού ταχυδρομείου ως νόμιμο ή spam.

Συσταδιοποίηση: Πρόκειται για τη διαδικασία ανακάλυψης ομάδων και δομών στα δεδομένα που είναι «παρόμοια» κατά κάποιο τρόπο, χωρίς να χρησιμοποιούνται γνωστές δομές στα δεδομένα.

Ανάλυση συσχέτισης (Μοντέλο αλληλεξάρτησης): Αναζητήσεις για σχέσεις μεταξύ των μεταβλητών. Για παράδειγμα, ένα σούπερ μάρκετ μπορεί να συλλέξει δεδομένα που αφορούν της αγοραστικές συνήθεις των πελατών του. Χρησιμοποιώντας τους κανόνες συσχέτισης, μπορεί να υπολογίσει ποια προϊόντα αγοράζονται συνήθως μαζί και να χρησιμοποιήσει αυτή την πληροφορία για αγοραστικούς σκοπούς προς όφελος των πελατών του και του ίδιου.

Παλινδρόμηση: Προσπαθεί να βρει μία συνάρτηση που μοντελοποιεί τα δεδομένα με το λιγότερο δυνατό λάθος.

Σύνοψη: Παρέχει μια συμπαγέστερη αναπαράσταση των δεδομένων, συμπεριλαμβάνοντας την οπτικοποίηση και την παραγωγή κανόνων.

2.3 Στόχοι της εξόρυξης δεδομένων

Η εξόρυξη δεδομένων είναι μια τεχνική η οποία μπορεί να φανεί χρήσιμη σε ένα μεγάλο εύρος κλάδων της επιστήμης και της οικονομίας. Προσφάτως όμως απέκτησε μεγαλύτερο ερευνητικό ενδιαφέρον επειδή μπορεί να χρησιμοποιηθεί από επιχειρήσεις για την πρόβλεψη της συμπεριφοράς των καταναλωτών.

Οι μέθοδοι εξόρυξης γνώσης στοχεύουν στην ανακάλυψη στοιχειών που θα είναι χρήσιμα για τους οργανισμούς και τις επιχειρήσεις. Πληροφορίες για τυποποιημένες μορφές όπως για παράδειγμα, ότι υπάρχουν πελάτες που θα ψωνίσουν περισσότερο από δύο φορές σε περίοδο εκπτώσεων ή προσφορών, ή είναι πιθανό να αγοράσουν τουλάχιστον μια φορά κατά την διάρκεια των εορταστικών ημερών, Πάσχα και Χριστουγέννων, είτε για συσχετίσεις όπως όταν ένας πελάτης αγοράζει dvd player τότε πιθανότατα να αγοράσει και κάποια άλλη ηλεκτρονική συσκευή, μπορεί να

αποτελέσουν καθοριστικούς παράγοντες για την λήψη αποφάσεων όσον αφορά τη λειτουργία μιας εμπορικής επιχείρησης. Αυτό συμβαίνει επειδή μπορεί να ληφθούν αποφάσεις σχετικά με το ωράριο, το ύψος και τη διάρκεια των εκπτώσεων, ακόμη και για την τοποθέτηση των προϊόντων μέσα στα καταστήματα.

Παράλληλα τέτοιου είδους πληροφορίες χρησιμοποιούνται για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων ή και για τον σχεδιασμό διαφορετικών στρατηγικών μάρκετινγκ. Τα στελέχη της επιχείρησης, που είναι υπεύθυνα για την λήψη των αποφάσεων εκμεταλλεύονται τις δυνατότητες της εξόρυξης γνώσης και μετατρέπουν τις γνώσεις σε επιτυχή αποτελέσματα. Παρακάτω περιγράφονται και αναλύονται οι στόχοι της εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων έχει λοιπόν σαν βασικούς της στόχους την εφαρμογή τεχνικών πρόβλεψης και συμπεριφοράς τάσεων (prediction), την αναγνώριση, την περιγραφή (description) σε μεγάλες βάσεις δεδομένων (Fayyad et al, 1996,1996, Hegland, 2003), καθώς επίσης την ταξινόμηση και την βελτιστοποίηση των πόρων της. Ειδικότερα:

Πρόβλεψη: Περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Τα αποτελέσματα της εξόρυξης μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο, αλλά και αν το κλείσιμο μιας γραμμής παραγωγής θα είχε θετική επίδραση στις πωλήσεις. Συγχρόνως σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών φαινόμενων ίσως να οδηγούσε στην πρόβλεψη σεισμικής δραστηριότητας.

Αναγνώριση: Σε αυτή τη φάση οι τυποποιημένες μορφές των δεδομένων χρησιμοποιούνται για να δείξουν την ύπαρξη μιας δραστηριότητας ή ενός γεγονότος.

Περιγραφή: Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (descriptive data mining tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

Ταξινόμηση: Σε αυτό το στάδιο έχουμε διαχωρισμό των στοιχείων, με αποτέλεσμα να προκύπτουν διαφορετικές κατηγορίες ή κλάσεις. Για παράδειγμα, οι πελάτες ενός σούπερ μάρκετ είναι δυνατόν να χωριστούν σε παρορμητικούς, πιστούς ή αλλιώς όπως θα λέγαμε κανονικούς, σπάνιους και σε φίλους των εκπτώσεων και προσφορών. Κατά την ανάλυση των πωλήσεων αυτή η κατηγοριοποίηση χρησιμοποιείται για να ληφθούν αποφάσεις, ώστε να προσελκυστούν περισσότεροι πελάτες ανεξαρτήτως κατηγορίας.

Βελτιστοποίηση: Μεταξύ των άλλων σκοπός της εξόρυξης γνώσης είναι η βέλτιστη χρήση κάποιων πόρων κάτω από περιορισμούς. Τέτοιοι πόροι μπορεί να είναι ο χρόνος, ο χώρος, το χρήμα και η μεγιστοποίηση κάποιων μεγεθών, όπως είναι τα κέρδη είτε οι πωλήσεις. Σε αυτή την περίπτωση η εξόρυξη γνώσης έχει κοινά σημεία με την επιχειρησιακή έρευνα (Σταυλιώτης, 2009).

Εφαρμογές της εξόρυξης δεδομένων

Ιατρική

Τα τελευταία χρόνια, η εξόρυξη δεδομένων χρησιμοποιείται ευρέως στους τομείς της ιατρικής, όπως η βιοϊατρική ,το DNA ,η γενετική και η φαρμακευτική. Στον τομέα της γενετικής, ο σκοπός είναι να κατανοήσουμε την χαρτογράφηση της σχέσης μεταξύ της μεταβολής των ακολουθιών του ανθρώπινου DNA και την προδιάθεση στην αρρώστια. Η εξόρυξη δεδομένων είναι ένα σημαντικό εργαλείο που μπορεί να βοηθήσει στην βελτίωση της διάγνωσης, της πρόληψης και της θεραπείας των ασθενειών.

Εξαιτίας της αύξησης των βιοϊατρικών ερευνών, η μεγάλη κλίμακα γονιδιακών προτύπων και λειτουργιών πρέπει να εξετασθεί. Τα εργαλεία της εξόρυξης δεδομένων μπορούν να βοηθήσουν σε μεγάλο βαθμό για να μελετήσουμε την σύσταση του DNA και να βρούμε ποικίλα πρότυπα και λειτουργίες αυτού.

Ένας από τους κύριους στόχους που σχετίζεται με την ανάλυση δεδομένων του DNA είναι η σύγκριση ποικίλων ακολουθιών και η αναζήτηση ομοιοτήτων μεταξύ των δεδομένων του DNA. Η σύγκριση κυρίως περιλαμβάνει την γονιδιακή ακολουθία υγιών και βλαβερών ιστών για να βρει την διαφορά ανάμεσα σε αυτούς τους δύο τύπους. Αυτό μπορεί να επιτευχθεί ανακτώντας τις τάξεις υγιών αλλά και βλαβερών γονιδιακών ακολουθιών και μετά βρίσκοντας τις συχνά εμφανιζόμενες μορφές των δύο τάξεων. Αυτή η ανάλυση βοηθάει στο να βρίσκουμε τις ομοιότητες και τις διαφορές στις γενετικές ακολουθίες.

Στην βιοϊατρική, ερευνάται αν οι περισσότερες ασθένειες προκαλούνται από ένα συνδυασμό των γονιδίων. Η μέθοδος της συσχέτισης χρησιμοποιείται για να καθορίσει την συνύπαρξη ομάδων των γονιδίων και επίσης μπορούμε να εξετάσουμε την αλληλεπίδραση και την σχέση μεταξύ των γονιδίων.

Τα εργαλεία της οπτικοποίησης παίζουν επίσης ένα σημαντικό ρόλο στην εξόρυξη δεδομένων στην βιοϊατρική. Τα εργαλεία αυτά μπορούν να παρουσιάσουν πολύπλοκες δομές γονιδίων σε γράφους, δένδρα και αλυσίδες. Η οπτική παρουσίαση βοηθάει στην καλύτερη κατανόηση αυτών των δομών για ανακάλυψη γνώσης και εξερεύνηση των δεδομένων.

Υπάρχουν διάφοροι συνδυασμοί γονιδίων που συμβάλλουν στις ασθένειες, αλλά αυτά τα γονίδια ενεργοποιούνται σε διαφορετικά επίπεδα. Η ανάλυση μονοπατιού χρησιμοποιείται για να συνδέει διαφορετικά γονίδια με διαφορετικά στάδια κατά την εξέλιξη της ασθένειας. Η ανάλυση μονοπατιού διαδραματίζει ένα σπουδαίο ρόλο στην γενετική.

Οικονομία

Άλλος τομέας που εφαρμόζεται η εξόρυξη δεδομένων είναι η οικονομία. Τα οικονομικά δεδομένα κυρίως συλλέγονται από τράπεζες και από άλλους οικονομικούς οργανισμούς. Τα δεδομένα αυτά συνήθως είναι αξιόπιστα, ολοκληρωμένα και έχουν υψηλή ποιότητα και απαιτούν συστηματική μέθοδο για την ανάλυση αυτών. Η συνεισφορά της εξόρυξης δεδομένων στην επιστήμη της οικονομίας συναντάται στην συλλογή και κατανόηση των δεδομένων, στην βελτίωση δεδομένων (data refinement), στην δημιουργία και εκτίμηση ενός μοντέλου και στην ανάπτυξη αυτού. Η σωστή ανάλυση των οικονομικών δεδομένων μας διευκολύνει στο να παίρνουμε καλύτερες αποφάσεις ενεργώντας σύμφωνα με την ανάλυση της αγοράς. Τα εργαλεία και οι τεχνικές της εξόρυξης δεδομένων βοηθούν στο να αναλύσουμε τα οικονομικά δεδομένα με τους παρακάτω τρόπους:

Τα δεδομένα που συλλέγονται από διάφορα οικονομικά ινστιτούτα, όπως οι τράπεζες ,συγκεντρώνονται αρχικά στην αποθήκη δεδομένων (data warehouse). Οι τεχνικές της πολυδιάστατης ανάλυσης δεδομένων χρησιμοποιούνται για την ανάλυση τέτοιων δεδομένων που συλλέγονται στην αποθήκη δεδομένων για τις γενικές ιδιότητές του.

Μία άλλη εφαρμογή της εξόρυξης δεδομένων σχετίζεται με την πρόβλεψη αποπληρωμής δανείου και πολιτικές πίστωσης του πελάτη. Μέθοδοι της εξόρυξης όπως η επιλογή χαρακτηριστικών (feature selection) βοηθάει στην ταυτοποίηση ποικίλων χαρακτηριστικών όπως το επίπεδο εισοδήματος του πελάτη, την εξόφληση ανάλογα με τα έσοδα, την πιστωτική του ιστορία κτλ. Με την επεξεργασία αυτών των χαρακτηριστικών, η τράπεζα μπορεί να αποφασίσει για τις πολιτικές δανειοδότησης βάσει των σχετικά χαμηλών κινδύνων. Οι τεχνικές της συσταδοποίησης και της ταξινόμησης βοηθούν τα οικονομικά ινστιτούτα να ομαδοποιούν διάφορους πελάτες που έχουν κοινά χαρακτηριστικά. Η αποτελεσματική συσταδοποίηση και οι μέθοδοι φιλτραρίσματος βοηθούν τις τράπεζες να ταυτοποιούν μία ομάδα πελατών, να συσχετίζουν ένα νέο πελάτη με την παρούσα ομάδα και να τους παρέχουν κοινά οφέλη.

Τα εργαλεία της εξόρυξης δεδομένων βοηθούν τα οικονομικά ινστιτούτα να αναγνωρίζουν τις απάτες και τα εγκλήματα από παραποιημένα δεδομένα από τις διάφορες βάσεις δεδομένων και από το ιστορικό συναλλαγών που έγιναν από τους πελάτες. Οι τεχνικές οπτικοποίησης βοηθούν στην παρουσίαση δεδομένων με διαφορετικές μορφές, όπως γράφοι που βασίζονται σε συγκεκριμένα γνωρίσματα. Προβάλλοντας τα δεδομένα από διάφορες οπτικές γωνίες, η τράπεζα δύναται να διακρίνει τους πελάτες που έχουν επιχειρήσει παράνομες πράξεις και μετά μια λεπτομερής έρευνα αυτών των ύποπτων περιπτώσεων βοηθάει στην εξιχνίαση των απατών και των εγκλημάτων.

Τηλεπικοινωνία

Η τηλεπικοινωνιακή βιομηχανία αναπτύσσεται πολύ γρήγορα όπως και η τεχνολογία. Αυτές τις μέρες οι τηλεπικοινωνιακές υπηρεσίες έχουν επεκταθεί από τοπικές και μεγάλης απόστασης τηλεπικοινωνίες, στην χρήση φαξ, συσκευές τηλεειδοποίησης, κινητό τηλέφωνο, και ηλεκτρονικό ταχυδρομείο. Εξαιτίας των εξελίξεων στις τηλεπικοινωνιακές τεχνολογίες και για να δουλέψουν αποτελεσματικά αυτές οι τεχνολογίες, οι τεχνικές της εξόρυξης δεδομένων ενσωματώνονται σε αυτές τις τεχνολογίες για να παράγουν αποδοτικά αποτελέσματα. Η εξόρυξη δεδομένων βοηθάει στην διάκριση τηλεπικοινωνιακών προτύπων, καταπολέμησης παράνομων δραστηριοτήτων, και επίσης βοηθάει στην καλύτερη χρήση των πόρων και στη βελτίωση της ποιότητας των υπηρεσιών. Η εξόρυξη δεδομένων βελτιώνει τις τηλεπικοινωνιακές υπηρεσίες με τους εξής τρόπους:

Τα τηλεπικοινωνιακά δεδομένα που συλλέγονται, περιλαμβάνουν τον τύπο κλήσης, την τοποθεσία του καλούντος και του κληθέντος, τον χρόνο κλήσης, την διάρκεια κλήσης κλπ. Η πολυδιάστατη ανάλυση βοηθά στον προσδιορισμό και στην σύγκριση του φορτιού του συστήματος, κίνηση δεδομένων, και κέρδος κλπ. Η ανάλυση μπορεί να δείξει διαγράμματα και γράφους των πόρων του συστήματος, του προορισμού κλπ κάνοντας χρήση των εργαλείων οπτικοποίησης της εξόρυξης

δεδομένων. Τέτοια εργαλεία όπως η συσχετισμένη οπτικοποίηση και η συσταδοποίηση παρέχουν χρήσιμες υπηρεσίες στην ανάλυση των δεδομένων τηλεπικοινωνίας.

Το κυρίως πρόβλημα που αντιμετωπίστηκε από την βιομηχανία τηλεπικοινωνιών είναι οι παράνομες δραστηριότητες. Αυτές οι δραστηριότητες μπορεί να έχουν να κάνουν με σκόπιμες κλήσεις κατά την ώρα αιχμής, περιοδικές κλήσεις κ.α. με αποτέλεσμα να επιδρούν αρνητικά στην επίδοση του δικτύου επικοινωνιών. Μέθοδοι όπως η συσταδοποίηση και η ανάλυση ακραίων τιμών, συνεισφέρει στην ανίχνευση παράνομων προτύπων βελτιώνοντας την αποτελεσματικότητα των υπηρεσιών τηλεπικοινωνίας.

2.4 Παράδειγμα εξόρυξης δεδομένων στην οικονομία

Στο συγκεκριμένο υποκεφάλαιο θα αναλυθούν παραδείγματα χρήσης της εξόρυξης δεδομένων στην οικονομία και συγκεκριμένα στην τραπεζική. Τα τελευταία χρόνια, η δυνατότητα αποθήκευσης δεδομένων έχει αυξηθεί πάρα πολύ. Οι πληροφορίες που περιέχονται σε αυτά τα δεδομένα μπορεί να είναι πολύ σημαντικές. Είναι γνωστό ότι, για να ανταγωνιστούν αποτελεσματικά σε ολοένα και πιο ανταγωνιστικές παγκόσμιες αγορές, οι τράπεζες και οι άλλοι ισχυροί οικονομικοί οργανισμοί θα πρέπει να κατανοήσουν καλύτερα και το προφίλ των πελατών τους. Μια ξεκάθαρη προοπτική για τη συμπεριφορά και τα χαρακτηριστικά των πελατών προέρχεται από το οικονομικό ιστορικό τους. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν για να μπορέσουν οι τράπεζες να αποκτήσουν και να διατηρήσουν καλούς πελάτες, όπου οι καλοί πελάτες είναι οι πιο κερδοφόροι.

Η ανακάλυψη γνώσεων σε βάσεις δεδομένων, συχνά αποκαλούμενη εξόρυξη δεδομένων, είναι η συμπερίληψη γνώσης που κρύβεται μέσα σε μεγάλες συλλογές λειτουργικών δεδομένων.

Ο αριθμός των δεδομένων που συλλέχθηκαν από τις επιχειρήσεις αυξήθηκε ραγδαία τα τελευταία χρόνια. Οι υπάρχουσες τεχνικές ανάλυσης στατιστικών δεδομένων

δυσκολεύονται να αντιμετωπίσουν τους μεγάλους όγκους δεδομένων που είναι τώρα διαθέσιμα. Ούτε αξιοποιούν αποτελεσματικά την αυξημένη ισχύ επεξεργασίας που είναι τώρα διαθέσιμη. Αυτή η εκρηκτική ανάπτυξη έχει οδηγήσει στην ανάγκη για νέες τεχνικές και εργαλεία ανάλυσης δεδομένων προκειμένου να βρεθούν οι πληροφορίες που αποκρύπτονται στα δεδομένα αυτά. Συνεπώς, έχει προκύψει το ερευνητικό πεδίο της Αναγνώρισης Γνώσης σε Βάσεις Δεδομένων.

Η τράπεζα είναι ένας χώρος όπου συλλέγονται τεράστια ποσά δεδομένων. Αυτά τα δεδομένα μπορούν να δημιουργηθούν από συναλλαγές τραπεζικών λογαριασμών, αιτήσεις δανείων, αποπληρωμές δανείων, επιστροφές πιστωτικών καρτών κλπ. Υπάρχει η υποψία ότι σε αυτές τις τεράστιες λειτουργικές βάσεις δεδομένων υπάρχουν κρυφές πληροφορίες σχετικά με το οικονομικό προφίλ των πελατών και ότι αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν βελτίωση της απόδοσης της τράπεζας. Η εξόρυξη δεδομένων μπορεί να οδηγήσει σε αυτό που στην τραπεζική ονομάζεται HYPERBANK. Δηλαδή τραπεζική αυξημένης απόδοσης.

Ο στόχος του HYPERBANK είναι να ενσωματωθεί η επιχειρηματική μοντελοποίηση, η αποθήκευση δεδομένων, η εξόρυξη και οι υπολογισμοί υψηλής απόδοσης, ώστε να μπορέσουν οι τράπεζες να αυξήσουν την κερδοφορία βελτιώνοντας τη διαδικασία δημιουργίας προφίλ πελατών. Η εξόρυξη είναι μια επαναληπτική διαδικασία πολλαπλών σταδίων. Κάθε στάδιο απαιτεί τη χρήση ειδικών γνώσεων σχετικά με τον τομέα. Τα επιχειρηματικά μοντέλα είναι πηγές γνώσης του τομέα και έτσι η ενσωμάτωσή τους θα είναι αμοιβαία επωφελής: τα επιχειρηματικά μοντέλα της εφαρμογής θα παρέχουν εξειδικευμένες γνώσεις στη διαδικασία εξόρυξης και οι χρήσιμες γνώσεις που θα προκύψουν από τη διαδικασία εξόρυξης θα ανατροφοδοτήσουν τα επιχειρηματικά μοντέλα.

2.4.1 Η εξόρυξη στην τραπεζική

Ένα ευρέως αποδεκτό μοντέλο της διαδικασίας έχει τα εξής βήματα:

- Καθορισμός του στόχου της διαδικασίας - ανάπτυξη της κατανόησης του τομέα εφαρμογής και των στόχων του τελικού χρήστη.
- Επιλογή δεδομένων - επιλογή ενός συνόλου δεδομένων
- Προετοιμασία δεδομένων - αυτό μπορεί να περιλαμβάνει την αφαίρεση του θορύβου από τα δεδομένα, το χειρισμό λειμμένων πεδίων, τη χρήση μεθόδων μετασχηματισμού για τη μείωση του χώρου αναζήτησης, την απόκτηση νέων χαρακτηριστικών κ.λπ.
- Επιλέγουμε την εργασία εξόρυξης δεδομένων - αυτή η απόφαση μπορεί να εξαρτηθεί από τον στόχο της διαδικασίας, τον τύπο των διαθέσιμων δεδομένων (π.χ., μπορεί να παραγγελθεί) και τις διαθέσιμες τεχνικές. Οι εργασίες εξόρυξης δεδομένων περιλαμβάνουν την ανακάλυψη κανόνων σύνδεσης, την ανίχνευση διαδοχικών προτύπων, την ανακάλυψη παρόμοιων χρονικών αλληλουχιών, την πρόβλεψη μιας ταξινόμησης, την ανακάλυψη ομάδων και την πρόβλεψη αξιών.
- Επιλέγουμε τον αλγόριθμο εξόρυξης δεδομένων - μια εργασία εξόρυξης δεδομένων μπορεί να έχει περισσότερους από έναν διαθέσιμο αλγόριθμο. Η επιλογή του αλγορίθμου μέτρησης δεδομένων εξαρτάται από τον στόχο της διαδικασίας, δηλαδή εάν είναι προγνωστική, περιγραφική κ.λπ.

- Εξόρυξη δεδομένων - αναζήτηση μορφών γνώσης από το σύνολο δεδομένων.
- Ερμηνεία αποτελεσμάτων εξόρυξης – η παρουσίαση των αποτελεσμάτων εξόρυξης δεδομένων είναι σημαντική, καθώς και η αξιολόγηση είναι δύσκολη.
- Αξιολόγηση της ανακαλυφθείσας γνώσης – Επίσης γίνεται ενσωμάτωση των παραγόμενων γνώσεων στην οργάνωση .

Η εξόρυξη είναι μια επαναληπτική διαδικασία: τα αποτελέσματα ενός βήματος μπορεί να σημαίνουν ότι ένα προηγούμενο βήμα πρέπει να επανεξεταστεί. Αν και το βήμα εξόρυξης δεδομένων είναι συνήθως το πιο υπολογιστικά ακριβό, η ποιότητα των αποτελεσμάτων που επιτυγχάνεται από τη διαδικασία εξαρτάται σε μεγάλο βαθμό από τα υπόλοιπα στοιχεία. Οι επιλογές που έγιναν σε αυτά τα βήματα εξαρτώνται από την γνώση πεδίου του χρήστη και μπορούν να έχουν μεγάλη επίδραση στην ποιότητα του αποτελέσματος της διαδικασίας.

2.4.2 Πρακτική ανακάλυψη γνώσης

Σε αυτό το κεφάλαιο αναλύουμε ένα πείραμα εκτελέστηκε σε πάνω από 1 εκατομμύριο εγγραφές με πάνω από 50 ιδιότητες ανά εγγραφή. Περίπου το 50% των χαρακτηριστικών ήταν κατηγορηματικά με περισσότερες από δύο πιθανές τιμές, το 20% ήταν δυαδικό και το υπόλοιπο ήταν συνεχές. Τα δεδομένα φορτώθηκαν ως ένας και μόνο πίνακας για τη DB2 και τα εργαλεία εξόρυξης δεδομένων που χρησιμοποιήθηκαν ήταν το Intelligent Miner της IBM και το Profiler XpertRule από την Attar Software.

Ο στόχος της διαδικασίας ήταν να δημιουργηθεί ένα προγνωστικό και περιγραφικό μοντέλο πελατών σύμφωνα με κάποια αφηρημένα μέτρα. Επομένως, το πρώτο

καθήκον ήταν να καθορίσουν συγκεκριμένα αυτό το μέτρο και να αποφασίσουν για μια στρατηγική για τη διαδικασία.

2.4.3 Επιλογή δεδομένων και προετοιμασία δεδομένων

Δεν ήταν όλα τα αρχεία στη βάση δεδομένων σχετικά με το πείραμά. Το μέγεθος του συνόλου δεδομένων ήταν 286816 εγγραφές.

Ορισμένα χαρακτηριστικά περιείχαν κενές τιμές. Σε ορισμένες περιπτώσεις αυτές αντικαταστάθηκαν με την προεπιλεγμένη τιμή που υπήρχε για το χαρακτηριστικό αυτό, σε άλλες περιπτώσεις αποφασίστηκε μια προκαθορισμένη τιμή. Ο στόχος της διαδικασίας περιλαμβάνει την κατασκευή ενός μοντέλου σύμφωνα με κάποιο μέτρο. Δεν υπήρχε ένα μοναδικό χαρακτηριστικό στη βάση δεδομένων που να σχετίζεται άμεσα με αυτό το μέτρο και έτσι ένα νέο χαρακτηριστικό προστέθηκε στο σύνολο δεδομένων που προέκυψε από διάφορα υπάρχοντα χαρακτηριστικά.

Οι περισσότερες ιδιότητες που προστέθηκαν κατά τη διάρκεια του πειράματος όταν αποφασίστηκε ότι ήταν χρήσιμες. Ορισμένες από τις κατηγορικές ιδιότητες είχαν πολλές αξίες. Αυτό σήμαινε ότι τα αποτελέσματα εξόρυξης δεδομένων σχετικά με αυτά τα χαρακτηριστικά ήταν δύσκολο να ερμηνευτούν και έτσι αποφασίστηκε να μειωθεί ο αριθμός των διακεκριμένων κατηγοριών με την ομαδοποίηση τιμών χαρακτηριστικών. Η ομαδοποίηση πραγματοποιήθηκε είτε αυτόματα από το εργαλείο εξόρυξης δεδομένων είτε κατόπιν συμβουλών εμπειρογνωμόνων τομέα.

2.4.4 Επιλογή του αλγόριθμου εξόρυξης δεδομένων

Χρησιμοποιήθηκε ο αλγόριθμος επαγωγής του δέντρου αποφάσεων του XpertRule Profiler . Το Intelligent Miner προσφέρει μια ποικιλία αλγορίθμων για κάθε τεχνική. Κατά την ομαδοποίηση, λόγω του κυρίως κατηγορηματικού χαρακτήρα των δεδομένων, επιλέχθηκε ο δημογραφικός αλγόριθμος. Ο στόχος της διαδικασίας ήταν

να παραχθεί ένα περιγραφικό μοντέλο και έτσι ο αλγόριθμος ταξινόμησης δέντρων αποφάσεων επιλέχθηκε σε αντίθεση με μια προσέγγιση βασισμένη στο νευρικό δίκτυο, που δεν είναι πολύ περιγραφικό.

2.4.5 Εξόρυξη δεδομένων

Η φάση εξόρυξης δεδομένων είναι η πλέον ακριβή υπολογιστική και είναι εξαιρετικά επαναληπτική. Πρώτον, ο αλγόριθμος ομαδοποίησης χρησιμοποιήθηκε για την απεικόνιση των δεδομένων. Πολλά υποσύνολα των δεδομένων επιλέχθηκαν με το χαρακτηριστικό μέτρησης να είναι σταθμισμένο έντονα, πράγμα που σημαίνει ότι κάθε παραγόμενο σύμπλεγμα περιείχε όλες τις εγγραφές για μια συγκεκριμένη τιμή του χαρακτηριστικού μέτρου. Αυτό μας έδωσε περισσότερη «αίσθηση» για τα δεδομένα και, σε κάποιο βαθμό, καθοδηγούμενες επιλογές που έγιναν κατά τις ασκήσεις ομαδοποίησης και ταξινόμησης. Τόνισε επίσης ορισμένες ασυνέπειες στη βάση δεδομένων που έπρεπε να αντιμετωπιστούν και έδειξε την ανάγκη για ομαδοποίηση εντός ενός χαρακτηριστικού. Για τη ταξινόμηση του συνόλου δεδομένων σε σχέση με το χαρακτηριστικό μέτρησης χρησιμοποιήθηκε αλγόριθμος αποφάσεων. Αυτή η διαδικασία είναι αναγκαστικά διερευνητική και συνεπώς χρησιμοποιήθηκε μια προσέγγιση δοκιμής και σφάλματος. Το σύνολο χαρακτηριστικών που χρησιμοποιήθηκαν ως το ενεργό σύνολο δεδομένων, έναντι του οποίου η ιδιότητα μέτρου ήταν, αλλάζει διαρκώς. Οι εμπειρογνώμονες του τομέα ενημέρωσαν τη διαδικασία σχετικά με τα χαρακτηριστικά των τύπων που ήταν πιο «ενδιαφέροντα» από την άποψη ενός μοντέλου πρόβλεψης, δηλαδή των πληροφοριών που θα μπορούσε να χρησιμοποιήσει η τράπεζα για να αναλύσει τους πελάτες.

2.4.6 Ερμηνεία αποτελεσμάτων

Οι γραφικές μέθοδοι αναπαράστασης αποτελεσμάτων που χρησιμοποιούνται από το Intelligent Miner και το XpertRule Profiler είναι αρκετά ευανάγνωστες. Ορισμένα συμπεράσματα θα μπορούσαν να ερμηνευθούν από έναν μη εμπειρογνώμονα που χρησιμοποιεί γενικές γνώσεις ενώ άλλα αποτελέσματα έπρεπε να εξηγηθούν από έναν ειδικό με μια πιο οικεία γνώση των δεδομένων.

ΚΕΦΑΛΑΙΟ 3: ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗ

3.1 Εισαγωγή

Με την εφαρμογή της εξόρυξης δεδομένων στην πραγματικότητα χρησιμοποιούμε προηγμένα εργαλεία της στατιστικής με πολύ συγκεκριμένο τρόπο. Όπως ήδη αναφέραμε μέσω ης εξόρυξης δεδομένων προσπαθούμε να λάβουμε γνώση από πληροφορίες που λαμβάνονται από έναν τεράστιο αριθμό δεδομένων, που δεν μπορούν να επεξεργαστούν από έναν άνθρωπο, αλλά η επεξεργασία τους γίνεται με χρήση ηλεκτρονικών υπολογιστών (Glymour et al., 1997).

Μέσω της στατιστικής επιστήμης επεξεργαζόμαστε τον τεράστιο όγκο των δεδομένων αυτών που μας επιτρέπει να κατηγοριοποιηθούν σε διαχειρίσιμες ομάδες. Η διαδικασία αυτή απαιτεί τη χρήση ειδικών λογισμικών, ή ειδικών προγραμμάτων που δημιουργούνται εξαρχής από τον διαχειριστή των δεδομένων.

Υπάρχουν όμως και διαφορές στην στατιστική προσέγγιση επεξεργασίας των δεδομένων σε σχέση με την εξόρυξη των δεδομένων. Η εξόρυξη δεν στοχεύει μόνο σε κατηγοριοποίηση των δεδομένων αλλά σε εξαγωγή πληροφορίας και γνώσης.

Στην στατιστική επεξεργασία πολλές φορές εύκολα κατανοητά δεδομένα μπορούν να χρησιμοποιηθούν για την εξαγωγή πρόβλεψης, ενώ στην εξόρυξη δεδομένων ο πρώτος στόχος είναι η κατανόηση των δεδομένων και των πληροφοριών που περιέχουν(Casella and Berger, 2002).

Συνεπώς μπορούμε να καταλήξουμε αρχικά στο εξής συμπέρασμα:

Η στατιστική επεξεργασία των δεδομένων έχει ως πρώτο στόχο την παραγωγή προβλεπτικών μοντέλων

Η εξόρυξη δεδομένων έχει ως πρώτο στόχο τη παραγωγή προτύπων για την μελλοντική χρήση παρόμοιων δεδομένων

Στην εξόρυξη δεδομένων αντί για την κατασκευή ενός συνολικού μοντέλου όπου περιλαμβάνονται όλες οι ενδιαφέρουσες μεταβλητές, προτιμώνται τα απλά και κατανοητά μοντέλα, υπό μορφή κανόνων, δέντρων, γραφημάτων κ.λπ.

Ένας αλγόριθμος εξόρυξη δεδομένων έχει ως στόχο τη δήλωση ισχυρισμών για τοπικές εξαρτήσεις μεταξύ μεταβλητών. Έτσι, για να θεωρηθεί επιτυχημένος ένας αλγόριθμος, απαιτείται να έχει ιδιαίτερη υπολογιστική ικανότητα, χωρίς όμως να ξεχνάμε τη χρήση της Στατιστικής. Δηλαδή, η στατιστική επιστήμη αποτελεί τον πυρήνα όλων αυτών των διεργασιών. Στην ενότητα που ακολουθεί παρουσιάζουμε τις βασικότερες έννοιες και θέματα της Στατιστικής που σχετίζονται με την έννοια της εξόρυξη δεδομένων (Σταυλιώτης, 2008).

3.2 Εργαλεία στατιστικής στην εξόρυξη δεδομένων

Σε αυτό το υποκεφάλαιο θα παρουσιάσουμε τα κύρια προαπαιτούμενα γνώσης στατιστικής που θα χρησιμοποιηθούν στην εξόρυξη δεδομένων. Τέτοιες εφαρμογές της στατιστικής είναι η χρήση κατανομής πιθανότητας, ι τυχαίες μεταβλητές κτλ. Θα χρησιμοποιήσουμε επίσης την έννοια της ανεξαρτησίας τυχαίας μεταβλητής, και ιδίως εργαλεία εύρεσης εξάρτησης. Η τεχνική Market Basket Analysis για παράδειγμα είναι ακριβώς η προσπάθεια εύρεσης συσχετίσεων μεταξύ προϊόντων που αγοράζονται από καταναλωτές. Η συσχέτιση μεταξύ προϊόντων μας δίνει τη δυνατότητα να καταλάβουμε με ποιόν τρόπο για παράδειγμα μπορεί να γίνεται η ταυτόχρονη ή ακόλουθη διαφήμιση προϊόντων στο διαδίκτυο.

Αφού μπορέσουμε και κατηγοριοποιήσουμε τα δεδομένα σε ομάδες μπορούμε να χρησιμοποιήσουμε τις οικογένειες δεδομένων για την παραγωγή προβλεπτικών μοντέλων και για εξαγωγή χρήσιμων συμπερασμάτων (Cox, 2006).

Φυσικά, όπως σε κάθε περίπτωση δεδομένων που δεν εμπεριέχουν όλο τον πληθυσμό, θα πρέπει να γίνει στατιστική εκτίμηση ώστε το δείγμα να είναι όσο πιο κοντά είναι δυνατόν στα χαρακτηριστικά του πληθυσμού. Η αβεβαιότητα είναι ο πιο σημαντικός παράγοντας όταν κάνουμε προσέγγιση από έναν δείγμα σε έναν

πληθυσμό. Η αβεβαιότητα μπορεί να μειωθεί με προσοχή του εκτιμητή, αλλά υπάρχουν και πιο ειδικές στατιστικές τεχνικές για τη μείωσης της .Μια τεχνική είναι η επαναδειγματοληψία και η προσομοίωση (Ross, 1997).

Επιπλέον, ένας ακόμη στόχος της στατιστικής έρευνας είναι η μείωση των αναγκαίων υποθέσεων για να θεωρηθεί καλή μια εκτίμηση. Αυτό σημαίνει ότι επιθυμούμε να έχουμε έναν όσο γίνεται εύρωστο εκτιμητή (Huber, 1981). Ένας σημαντικός ερευνητής που άλλαξε το επιστημονικό πεδίο της στατιστικής είναι ο Bayes, ο οποίος δημιούργησε μια διαφορετική προσέγγιση για την εκτίμηση της αβεβαιότητας και τη μείωση της. Κατά τη μέθοδο αυτή γίνεται συνδυαστική χρήση μοντέλων, ώστε να λάβουμε το μοντέλο με τη μικρότερη δυνατή αβεβαιότητα. Κατόπιν λαμβάνουμε ως εκτίμηση τον σταθμισμένο μέσον όρο των αποτελεσμάτων που μας δίνει κάθε μοντέλο. Η εφαρμογή αυτή καλείται Bayesian model averaging και οδηγεί σε βελτίωση της προβλεπτικής ικανότητας (Bernardo and Smith, 1994).

Κατά τη εξόρυξη δεδομένων ι περισσότερες από τις διεργασίες, όπως η λήψη των δεδομένων, η επεξεργασία και η κατασκευή μοντέλου γίνεται αυτοματοποιημένα. Συνεπώς είναι εξαιρετικά σημαντικό να μην υπάρχουν σφάλματα τα οποία θα αναπαράγονται σε όλη τη διαδικασία δίνοντας μας λανθασμένα αποτελέσματα. Έτσι, ο υπολογισμός των πιθανών σφαλμάτων χρήζει ιδιαίτερης σημασίας. Αυτό μπορεί να απαιτεί Monte Carlo ανάλυση (Gentle, 2002). Παρατηρούμε ότι, αρχικά, οι αναλυτές δεδομένων αναγκάζονταν να αποφύγουν την ανάλυση περίπλοκων Μπεϋζιανών μοντέλων και τον υπολογισμό σύνθετων πιθανοφανειών. Αυτή η επιλογή οφειλόταν στη δυσχέρεια διενέργειας υπολογισμών.

3.3 Έλεγχος υποθέσεων (Hypothesis testing)

Κάθε στατιστικός έλεγχος πρέπει να υπάγεται σε έλεγχο των υποθέσεων που έγιναν. Οι υποθέσεις πολλές φορές σχετίζονται με το πόσο αντιπροσωπευτικό είναι το δείγμα σε σχέση με τον πληθυσμό. Έστω ότι έχουμε μια υπόθεση η οποία ορίζεται από μια παράμετρο. Η υπόθεση μπορεί να έχει μηδενική τιμή ή μια εναλλακτική

τιμή, που ονομάζονται Η0 και Η1 αντίστοιχα. Οι υποθέσεις αυτές καλύπτουν όλο το φάσμα των πιθανών τιμών τα παραμέτρου (Aczel, 1989).

Με την εφαρμογή ενός κανόνα για την απόφαση μπορούμε να δεχτούμε αν μία τιμή που λαμβάνει η συνάρτηση είναι αποδεκτή ή όχι, ανάλογα αν ικανοποιείται η τιμή της παραμέτρου.

Ο έλεγχος υποθέσεων είναι μια μέθοδος που από πολλούς ερευνητές χαρακτηρίζεται μονόπλευρη και οδηγεί πολλές φορές σε κανονικοποιημένα αποτελέσματα που δεν αποτυπώνουν το πραγματικό εύρος τιμών. Για αυτό το λόγο πολλές φορές η μέθοδος θεωρείται ασυνεπής. Ο κάθε έλεγχος υποθέσεων έχει ένα επίπεδο σημαντικότητας που ονομάζουμε α, το οποίο θα πρέπει να μειώνεται όσο αυξάνεται το μέγεθος του δείγματος, δηλαδή προσεγγίζει σταδιακά καλύτερα τον πληθυσμό. Προφανώς όταν το δείγμα αυξηθεί τόσο που γίνει ο πληθυσμός τα επίπεδο σημαντικότητας του ελέγχου είναι 0.

Μπορούμε επίσης να χρησιμοποιήσουμε πάνω από έναν έλεγχο υποθέσεων α, και να τους συνδυάσουμε και να καταλήξουμε σε έναν νέο έλεγχο υποθέσεων επιπέδου επίσης α. Η συγκεκριμένη άποψη παρήχθη από τον Ιταλό μαθηματικό Carlo Emilio Bonferroni και βασίζεται στην ανισότητα Bonferroni (Casella and Berger, 2002).

Έστω λοιπόν ότι έχουμε αριθμο m ελέγχων υποθέσεων για μια συγκεκριμένη παράμετρο, από τους οποίους ο κάθε έλεγχο υπόθεσης έχει βαθμό επίπεδο α. Τότε με βάση την διόρθωση Bonferroni (Bonferroni correction), ο βαθμός σημαντικότητας κάθε ελέγχου υποθέσεων γ ορίζεται από την απλή σχέση:

$$\gamma=\alpha/m$$

Μια εναλλακτική μέθοδος που μπορεί να χρησιμοποιηθεί είναι η μέθοδος S που αναπτύχθηκε από τον Scheffe το 1959. Σε περίπτωση που απαιτείται έλεγχος πολλαπλών υποθέσεων μπορούμε να χρησιμοποιήσουμε τους κανόνες του Miller.

3.4 Αξιολόγηση μοντέλων (Model scoring)

Μετά την δημιουργία κάποιου μοντέλου θα πρέπει να γίνει μία συγκριτική αξιολόγηση των μοντέλων για την επιλογή του κατάλληλου. Για αυτό το λόγο έχουν δημιουργηθεί κάποιες κλίμακες αξιολόγησης τω μοντέλων και των υποθέσεων. Οι περισσότερο ευρέως χρησιμοποιούμενοι κανόνες αξιολόγησης μοντέλων είναι αυτοί των Akaike ή AIC (Akaike, 1974) και Bayes (Schwarz, 1978). Για το κριτήριο AIC υπάρχει ο τύπος:

$$AIC = 2logL + 2q$$

όπου q το πλήθος των παραμέτρων του μοντέλου και L η πιθανοφάνεια υπολογισμένη στον εκτιμητή μέγιστης πιθανοφάνειας του υπό εκτίμηση μοντέλου (Maimon and Rokach, 2005).

Δεύτερο κριτήριο που θα παρουσιάσουμε είναι το κριτήριο του Bayes. Χρησιμοποιείται για μεγάλα δείγματα και δίνεται από τον τύπο:

$$BIC = -2 \cdot log\,(L) + q \cdot log\,(n)$$

όπου n το μέγεθος του δείγματος και q, L ομοίως με παραπάνω.

Πολλές φορές στη εξόρυξη δεδομένων χρειάζεται να χρησιμοποιήσουμε ένα ήδη γνωστό δείγμα για να προβλέψουμε τα χαρακτηριστικά ενός άγνωστου δείγματος. Με αυτό τον τρόπο μπορεί να γίνει εφαρμογή γνωστών αλγορίθμων σε άγνωστα δείγματα. Για να γίνει ωστόσο μια τέτοια θεώρηση θα πρέπει να υποθέσουμε ότι και

τα δύο δείγματα ακολουθούν τις ίδιες κατανομές πιθανότητας για τις παραμέτρους που εξετάζουμε. Αυτή η διαδικασία, όπως κάθε πρόβλεψη, εμπεριέχει κάποια αβεβαιότητα, η οποία θα πρέπει να μετρηθεί ώστε να γνωρίζουμε έστω και προσεγγιστικά την αξιοπιστία της μεθόδου. Κατά τη διαδικασία εξόρυξης δεδομένων υπάρχουν βάσεις δεδομένων για κατανομές πιθανότητας σε διάφορες διεργασίας, από τις οποίες ο διαχειριστής τω δεδομένων καλείται να επιλέξει ή να δημιουργήσει εξ αρχής.

3.5 Συνεργασία και σύγκριση μεταξύ στατιστικής και εξόρυξης δεδομένων

Όπως ήδη αναφέραμε η εξόρυξη δεδομένων χρησιμοποιεί σε πολύ μεγάλο βαθμό τις εφαρμογές της στατιστικής. Ωστόσο υπάρχουν και κάποιες σημαντικές διαφορές μεταξύ των δύο επιστημονικών τομέων. Παραθέτουμε τι απόψεις των Hand και συν. (2001), όπως συνοψίζονται στο έργο του Σταυλιώτη (2008).

3.5.1 Πλεονεκτήματα της εξόρυξης δεδομένων

Η μεγαλύτερη διαφορά μεταξύ τω δύο μεθόδων είναι το μέγεθος των βάσεων δεδομένων που μπορούν να διαχειριστούν. Ταυτόχρονα είναι και το μεγαλύτερο πλεονέκτημα της εξόρυξης δεδομένων. Η στατιστική μπορεί μέσω κάποιων προγραμμάτων να διαχειριστεί δεδομένα της τάξεως των εκατοντάδων χιλιάδων τιμών, ως ακραία περίπτωση. Η εξόρυξη δεδομένων σαν διαδικασία μπορεί να διαχειρίζεται εκατομμύρια ή και δισεκατομμύρια σημεία για παρόμοιες διαδικασίες.

Η στατιστική επίσης αντιμετωπίζει και πρόβλημα στην διαχείριση τεράστιων συνόλων δεδομένων στην περίπτωση πολλών μεταβλητών. Όπως καταλαβαίνουμε, η περίπτωση ύπαρξης πολλών μεταβλητών σε ένα σύνολο δεδομένων οδηγεί στην δημιουργία επιπρόσθετων περιορισμών στην αρχική επιλογή μοντέλου από έναν ειδικό. Στα πλαίσια της ΕΔ, ένα σύνολο δεδομένων δεν είναι αυτό που βλέπει απλά ένας στατιστικός, δηλαδή μερικές γραμμές που παρουσιάζουν τα αντικείμενα και

κάποιες στήλες όπου δίνονται οι μεταβλητές. Για παράδειγμα, η επιλογή ενός τυχαίου δείγματος από ένα σύνολο δεδομένων μπορεί να μην είναι μια εύκολη υπόθεση, όπως θα θεωρούσε ένας στατιστικός. Στην ΕΔ, ένα αρχείο μπορεί να αποθηκεύεται ταυτόχρονα σε πολλές μηχανές, διαιρεμένο σε μέρη.

Τέλος, εκτός από το πρόβλημα κατά την αύξηση των μεταβλητών, δεν πρέπει να ξεχνάμε τα σύνολα δεδομένων που αναπτύσσονται διαρκώς. Ας σκεφτούμε, για παράδειγμα την καταγραφή εισερχομένων κλήσεων στο τμήμα τηλεφωνικής εξυπηρέτησης ή την καταγραφή κατανάλωσης ηλεκτρικού ρεύματος. Αρχεία σαν αυτά πολλαπλασιάζουν διαρκώς το μέγεθός τους και προκαλούν αλλαγές στη φύση του προβλήματος και την αναζήτηση λύσης. Στην περίπτωση αυτή, η εξόρυξη δεδομένων μπορεί να βοηθήσει(Casella and Berger, 2002).

3.5.2 Πλεονεκτήματα της στατιστικής

Στην προηγούμενη υποενότητα αναφέραμε τα προβλήματα που ανακύπτουν σε μια στατιστική εφαρμογή όταν το μέγεθος ενός συνόλου δεδομένων είναι αρκετά μεγάλο ή αυξάνει διαρκώς. Όμως, δε μπορούμε να παραβλέψουμε τις αδυναμίες της εξόρυξης δεδομένων, που αποτελούν σημεία υπεροχής της στατιστικής.

Όπως σχολιάσαμε και στο πρώτο κεφάλαιο, η εξόρυξη δεδομένων είναι μια δευτερεύουσα διαδικασία ανάλυσης δεδομένων. Αυτό σημαίνει ότι τα προς ανάλυση δεδομένα είχαν συλλεχθεί αρχικά για κάποιο άλλο σκοπό. Το προσόν, λοιπόν, της στατιστικής είναι ότι αποτελεί την πρωτογενή ανάλυση για τα δεδομένα. Δηλαδή, τα δεδομένα συλλέγονται ύστερα από τη διαμόρφωση συγκεκριμένων ερωτημάτων, ενώ στη συνέχεια αναλύονται ώστε να απαντηθούν τα ερωτήματα αυτά(Cox, D.R. and Hinkley, 1974).

Στην πραγματικότητα, η στατιστική συλλέγει δεδομένα με τη διενέργεια πιο έγκυρων μεθόδων, όπως του Πειραματικού Σχεδιασμού (Experimental Design). Το πρόβλημα που αντιμετωπίζει η εξόρυξη δεδομένων είναι ότι όταν επιχειρείται η επίλυση ενός ζητήματος μέσω της ανάλυσης δεδομένων που δεν είχαν συλλεχθεί για

το ζήτημα αυτό, τότε μπορεί να μην προκύψει το ιδανικό αποτέλεσμα. Μπορεί, δηλαδή, να μην υπάρξει το κατάλληλο ταίριασμα των δεδομένων στο συγκεκριμένο ζήτημα.

Πέρα από τον τρόπο συλλογής των δεδομένων, τα μεγάλα σύνολα δεδομένων που χειρίζεται η εξόρυξη δεδομένων αντιμετωπίζουν και άλλα σοβαρά προβλήματα. Για παράδειγμα, ένα τεράστιο σύνολο δεδομένων μπορεί να περιέχει ελλείπουσες τιμές (missing values), θόρυβο (noise), ή «φθαρμένα» (corrupted) στοιχεία.

Κλείνοντας, αξίζει να αναφέρουμε ότι η εξόρυξη δεδομένων συμβαδίζει με τις κλασσικές τεχνικές διερευνητικής ανάλυσης δεδομένων της στατιστικής, αλλά είναι σε θέση να αντιμετωπίζει και άλλα ζητήματα. Ένα πολύ μεγάλο ή ασυνήθιστο (μη παραδοσιακό) σύνολο δεδομένων δε μπορεί να διαχειριστεί τόσο εύκολα από μια στατιστική εφαρμογή.

3.6 Εισαγωγή στους κανόνες συσχέτισης

Κατά την εξόρυξη δεδομένων μπορούμε να εξάγουμε πολύ χρήσιμα συμπεράσματα από την σχέσεις που θα παρατηρήσουμε μεταξύ των δεδομένων. Οι κανόνες συσχέτισης χρησιμοποιούνται για να ελέγξουν πιθανή σχέση μεταξύ τεράστιων ποσοτήτων δεδομένων. Στην οικονομία και στη βιομηχανία είναι εξαιρετικά χρήσιμη διαδικασία, ειδικά για τη πρόβλεψη της συμπεριφορά των καταναλωτών. Το πιο χαρακτηριστικό παράδειγμα είναι ο έλεγχος πιθανής συσχέτισης μεταξύ δυο προϊόντων. Αν ένα προϊόν Α αγοράζεται το ίδιο συχνά ή τις ίδιες ημέρες με το προϊόν Β τότε η επιχείρηση μπορεί να εξάγει συμπεράσματα για το μάρκετινγκ που θα χρησιμοποιήσει στην προώθηση των δυο προϊόντων.

Για παράδειγμα έστω ότι κάθε προϊόν σε ένα μαγαζί λιανικής πώλησης αντιπροσωπεύεται από μια δυαδική μεταβλητή που εξετάζει αν το προϊόν υπάρχει ή όχι. Τότε το σύνολο των αγορών μπορεί να παρασταθεί με ένα άνυσμα με τις μεταβλητές ύπαρξης ή μη των προϊόντων. Ο έλεγχος συσχέτισης μεταξύ

διαφορετικών ανυσμάτων μπορεί να μας δώσει συμπεράσματα για τη συσχέτιση παραπάνω από ενός προϊόντος.

Για την αξιολόγηση των κανόνων συσχέτισης και την εύρεση του βαθμό ενδιαφέροντος χρησιμοποιούμε την υποστήριξη (support) και την εμπιστοσύνη (confidence).

Οι κανόνες συσχέτισης είναι διαφορετικοί από τους κανόνες κατηγοριοποίησης διότι στους κανόνες συσχέτισης δεν υπάρχει συγκεκριμένη κατηγοριοποίηση αλλά η πρόβλεψη γίνεται για κάθε πιθανό χαρακτηριστικό και για παραπάνω από μία τιμές αυτού του χαρακτηριστικού. Λόγο αυτού του γεγονότος υπάρχουν πάνω από ένα κανόνες συσχέτισης και η δυσκολία είναι να χρησιμοποιηθούν αυτοί που είναι οι πιο χρήσιμοι. Οι κανόνες συσχέτισης συνήθως περιορίζονται σε αυτούς που ισχύουν σε κάποιο ελάχιστο αριθμό παραδειγμάτων π.χ. για το 80% του συνόλου δεδομένων και έχουν μεγαλύτερο από ένα ασφαλές μικρότερο επίπεδο ακριβείας π.χ. 95%.

Ακόμη και τότε είναι πάρα πολλοί και πρέπει να ελέγχονται όλοι για το ποιοι παράγουν νόημα. Οι κανόνες συσχέτισης συνήθως περιέχουν μόνο μη αριθμητικά χαρακτηριστικά. Η είσοδος σε ένα σχήμα εκπαίδευσης είναι ένα σύνολο από instances. Τα instances είναι τα πράγματα από τα οποία πρέπει να εξαχθούν συμπεράσματα. Κάθε instance είναι ένα ανεξάρτητο παράδειγμα από το concept για το οποίο γίνεται η εκπαίδευση. Κάθε σύνολο δεδομένων αντιπροσωπεύεται από ένα πίνακα από instances με κάποια χαρακτηριστικά τα οποία σε όρους βάσεων δεδομένων αντιπροσωπεύουν μία συσχέτιση ή ένα flat file. Κάθε ανεξάρτητο instance το οποίο αποτελεί την είσοδο σε μία εκμάθηση μηχανής χαρακτηρίζεται από τιμές σε ένα προκαθορισμένο πεδίο χαρακτηριστικών τα οποία ονομάζονται attributes. Μία δυσκολία που προκύπτει είναι όταν κάποια Instances που αναφέρονται στο ίδιο concept δεν έχουν τα ίδια χαρακτηριστικά με τα άλλα. Για παράδειγμα κάποια οχήματα μεταφοράς έχουν ρόδες ενώ κάποια όπως τα πλοία όχι. Στην περίπτωση αυτή χρησιμοποιούμε μια ένδειξη που σημαίνει "αυτό το χαρακτηριστικό δεν υπάρχει για το συγκεκριμένο instance". Υπάρχουν 2 μεγάλα ήδη χαρακτηριστικών τα οποία χωρίζονται σε 4 μικρότερα(Bartik, 2009).

Τα 2 ήδη είναι τα arithmetic τα οποία είναι συνεχή και αριθμητικά και τα nominal τα οποία παίρνουν τιμές από ένα προκαθορισμένο σύνολο τιμών. Τα 4 ήδη χαρακτηριστικών είναι τα nominal, ordinal, interval και ratio. Οι nominal τιμές δεν είναι συγκρίσιμες μεταξύ τους π.χ. ηλιόλουστος, βροχερός. Οι ordinal είναι π.χ. ζεστός>δροσερός>κρύος κτλ. Οι τιμές interval έχουν τιμές που εκτός από συγκρίσιμες είναι και ποσοτικές όπως οι τιμές θερμοκρασίας Κελσίου 20, 22 κτλ. Τέλος οι τιμές ratio είναι αυτές που δεν περιέχουν εξ ορισμού το μηδέν. Όπως για παράδειγμα η απόσταση ανάμεσα από 2 αμάξια(Bartik, 2009).

3.7 Μέτρα Αξιολόγησης Κανόνων

Οι κανόνες που παρήχθησαν με τις διεργασίες στατιστικής ή εξόρυξης δεδομένων πρέπει εν συνεχεία να αξιολογηθούν. Γενικά δεν υπάρχουν συγκεκριμένοι τρόποι αξιολόγησης των κανόνων αλλά χρησιμοποιούμε τους παρακάτω παράγοντες:

Περιεκτικότητα (Conciseness)

Μπορούμε να θεωρήσουμε ότι ένας κανόνας έχει περιεκτικότητα όταν περιέχει λίγα ζευγάρια τιμών, και ένα σύνολο κανόνων περιεκτικό αν περιέχει με τη σειρά του λίγους κανόνες. Η περιεκτικότητα βοηθάει στην ευκολότερη κατανόηση από τον χρήστη.

Γενικότητα (Generality)

Η γενικότητα είναι το μέτρο του ποιου μέρος του συνόλου των δεδομένων καλύπτει ο κανόνας. Όσο πιο γενικό θεωρείται ένα σύνολο κανόνων τόσο πιο ενδιαφέρον θεωρείται. Θεωρούμε συχνό ένα σύνολο αν η υποστήριξή του, το φράγμα των εγγραφών στο σύνολο δεδομένων που περιέχει το itemset, είναι πάνω από ένα δεδομένο κατώτατο όριο.

Αξιοπιστία

Αξιοπιστία ενός κανόνα σημαίνει ότι αν τον εφαρμόσουμε σε ένα μεγάλο σύνολο δεδομένων ή σε πολλά σύνολα δεδομένων έχει ικανοποιητικά αποτελέσματα. Δηλαδή, αν ένας κανόνας συσχέτισης δύο παραγόντων εφαρμόζεται σε πολλά δεδομένα και παρατηρείται ικανοποίηση του κανόνα θεωρείται αξιόπιστος. Έχουν προταθεί αρκετές μέθοδοι μέτρησης της αξιοπιστίας ενός κανόνα.

Αξιοπιστία ενός κανόνα σημαίνει ότι αν τον εφαρμόσουμε σε ένα μεγάλο σύνολο δεδομένων ή σε πολλά σύνολα δεδομένων έχει ικανοποιητικά αποτελέσματα. Δηλαδή, αν ένας κανόνας συσχέτισης δύο παραγόντων εφαρμόζεται σε πολλά δεδομένα και παρατηρείται ικανοποίηση του κανόνα θεωρείται αξιόπιστος. Έχουν προταθεί αρκετές μέθοδοι μέτρησης της αξιοπιστίας ενός κανόνα.

Ιδιαιτερότητα (Peculiarity)

Ιδιαίτερος θεωρείται ο κανόνας που διαφέρει από τους άλλους παραγόμενους κανόνες σύμφωνα με κάποιο κριτήρια απόστασης. Οι ιδιαίτεροι κανόνες προέρχονται πολλές φορές από ιδιαίτερα δεδομένα, δηλαδή λίγα δεδομένα και διαφορετικά από το μεγαλύτερο μέρος δεδομένων.

Ποικιλομορφία (Diversity)

Αν τα στοιχεία ενός κανόνα είναι διαφέρουν αρκετά, και αν οι κανόνες ενός συνόλου κανόνων τότε θεωρούμε ότι υπάρχει ποικιλομορφία στον κανόνα και στο σύνολο κανόνων αντίστοιχα. Η ποικιλομορφία αυξάνει το ενδιαφέρον του κανόνα ή του συνόλου κανόνων.

Καινοτομία (Novelty)

Υπάρχει δυσκολία μέτρησης της καινοτομίας, γιατί εξαρτάται και από την γνώση του χρήστη του κανόνα. Γενικά η καινοτομία είναι το μέτρο διαφορετικότητας ενός κανόνα σε σχέση με τους άλλους κανόνες ενός παρεμφερούς αντικειμένου.

Surprisingness

Αυτός ο παράγοντας αξιολόγησης μας δείχνει πόσο απροσδόκητος είναι ένας κανόνας. Δηλαδή από τη συσχέτιση δεδομένων που αναμέναμε, η συσχέτιση που εξάγεται είναι μη αναμενόμενη. Οι απροσδόκητοι κανόνες έχουν αυξημένο ενδιαφέρον επειδή αναγκάζουν σε αναθεώρηση τους ήδη υπάρχοντες κανόνες ενός αντικειμένου.

Ωφελιμότητα (Utility)

Θεωρούμε ότι ένας κανόνας είναι ωφέλιμος με βάση το πόσο βοηθάει στην επίτευξη κάποιου σκοπού. Βέβαια μπορεί ένας κανόνας να είναι χρήσιμος και σε διαφορετικό χρήστη και πεδίο από τον αρχικό του στόχο.

Εφαρμοσιμότητα (Actionability)

Ένας κανόνας είναι εφαρμόσιμος σε κάποια περιοχή εάν επιτρέπει τη λήψη απόφασης για μελλοντικές ενέργειες σε αυτήν την περιοχή(Bartik, 2009).

3.8 Market Basket Analysis

Με την τεράστια ποσότητα δεδομένων που συλλέγονται και αποθηκεύονται συνεχώς σε βάσεις δεδομένων, αρκετές βιομηχανίες ενδιαφέρονται για τους κανόνες συσχέτισης από τις βάσεις δεδομένων τους. Για παράδειγμα, οι σχέσεις σύνδεσης μεταξύ μεγάλων ποσοτήτων δεδομένων από επιχειρηματικές συναλλαγές μπορεί να

βοηθήσουν στο σχεδιασμό καταλόγων, το cross-marketing, την ανάλυση lossleader, και διάφορες διαδικασίες λήψης αποφάσεων των επιχειρήσεων.

Ένα τυπικό παράδειγμα χρήσης των δεδομένων είναι η ανάλυση καλαθιού αγοράς. Αυτή η μέθοδος εξετάζει τα πρότυπα αγοράς πελατών, προσδιορίζοντας συσχετισμούς μεταξύ των διαφόρων στοιχείων που τοποθετούν οι πελάτες στα καλάθια αγορών τους. Η ταυτοποίηση των συσχετίσεων αυτών μπορεί να βοηθήσει τους λιανοπωλητές να επεκτείνουν το μάρκετινγκ τους αποκτώντας γνώσεις σχετικά με τα στοιχεία που αγοράζονται συχνά από κοινού από τους πελάτες. Είναι χρήσιμο να εξετάσουμε την αγοραστική συμπεριφορά των πελατών και να βοηθήσουμε στην αύξηση των πωλήσεων και στην εξοικονόμηση αποθεμάτων εστιάζοντας στα δεδομένα των συναλλαγών των σημείου πώλησης. Αυτό είναι ένα ευρύ πεδίο για τους ερευνητές να αναπτύξουν έναν καλύτερο αλγόριθμο εξόρυξης δεδομένων(Trnka, 2010).

Η πλειονότητα των επιχειρηματικών αναγνωρισμένων οργανισμών έχουν συγκεντρώσει μαζικές πληροφορίες από τους πελάτες τους εδώ και δεκαετίες. Με τις εφαρμογές ηλεκτρονικού εμπορίου να αυξάνονται γρήγορα, οι οργανισμοί θα έχουν μια τεράστια ποσότητα των δεδομένων που δεν είχαν εδώ και χρόνια. Το Data Mining, που ονομάζεται επίσης ως Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων (KDD), υπάρχει για να καθορίσει τις τάσεις, τα πρότυπα, τους συσχετισμούς, και τις ανωμαλίες σε αυτές τις βάσεις δεδομένων που μπορούν να βοηθήσουν να δημιουργηθούν ακριβείς μελλοντικές προβλέψεις για επιχειρηματικές αποφάσεις(Trnka, 2010).

Το Market Basket Analysis είναι ένα από τα πιο σημαντικά πεδία εφαρμογής της Εξόρυξης Δεδομένων. Υπό την προϋπόθεση ότι ένα σύνολο συναλλαγών πελατών είναι καταχωρημένο σε στοιχεία, η κύρια πρόθεση είναι να προσδιοριστούν οι συσχετισμοί μεταξύ των πωλήσεων

Είναι γεγονός ότι όλα τα στελέχη σε κάθε είδους κατάστημα ή πολυκατάστημα θα ήθελαν να αποκτήσουν γνώσεις σχετικά με την αγοραστική συμπεριφορά των πελατών. Αυτό το σύστημα ανάλυσης καλαθιού αγοράς θα βοηθήσει τα στελέχη να κατανοήσουν τα σύνολα των στοιχείων που είναι οι πελάτες πιθανό να αγοράσουν.

Η ανάλυση αυτή μπορεί να πραγματοποιηθεί σε όλα τα στοιχεία των καταστημάτων λιανικής πώλησης και των συναλλαγών πελατών. Αυτά τα αποτελέσματα θα τα καθοδηγήσουν τους υπεύθυνους να σχεδιάσουν το μάρκετινγκ ή τη διαφήμιση. Για παράδειγμα, η ανάλυση καλαθιού αγοράς μπορεί να βοηθήσει επίσης τους διαχειριστές να προτείνουν νέα ρύθμιση στον τρόπο που γίνεται η διάταξη των προϊόντων στο κατάστημα. Βάσει αυτής της ανάλυσης, τα στοιχεία που αγοράζονται τακτικά μαζί μπορούν να τοποθετηθούν σε εγγύτητα με σκοπό την περαιτέρω προώθηση της πώλησης τέτοιων αντικειμένων μαζί. Εάν οι καταναλωτές που αγοράζουν υπολογιστές πιθανόν αγοράσουν και λογισμικό antivirus την ίδια στιγμή, η επιχείρηση πρέπει να τοποθετήσει το λογισμικό στην οθόνη του υπολογιστή που είναι προς πώληση. Είναι μια κίνηση που πιθανώς να συμβάλει στην ενίσχυση των πωλήσεων και των δύο αυτών στοιχείων.

Η εξόρυξη συσχετίσεων ανακαλύπτει όλους τους κανόνες που προσφέρονται στη βάση δεδομένων που εξασφαλίζουν κάποια ελάχιστη αξιοπιστία και ελάχιστο περιορισμό εμπιστοσύνης. Στην περίπτωση της εξόρυξης κανόνων συσχέτισης, ο στόχος δεν είναι προκαθορισμένος, ενώ εξόρυξη κανόνων ταξινόμησης υπάρχει μόνο με προκαθορισμένο στόχο(Trnka, 2010).

3.8.1 Βασικές πρακτικές στην ανάλυση καλαθιού αγοράς

Η Market Basket Analysis (MBA) μπορεί να αποδειχθεί πολύ χρήσιμη βοήθεια για τις κύριες προκλήσεις που αντιμετωπίζουν σήμερα οι έμποροι λιανικής πώλησης, απαντώντας σε μια σειρά εμπορικών ζητημάτων. Οι κορυφαίοι λιανοπωλητές χρησιμοποιούν MBA για να κάνουν τις επιχειρήσεις τους πιο προβλέψιμες και κερδοφόρες με τον εντοπισμό συγγενειών των προϊόντων στην πάροδο του χρόνου. Οι συσχετίσεις συναλλαγών μπορούν να αλλάζουν συνεχώς, και σε ένα μικρό μέγεθος καλαθιού αγοράς (3-5 τεμάχια), ή και σε μεγαλύτερο μέγεθος καλαθιού (καλάθι αγοράς παντοπωλείου με 15 -20 στοιχεία).

Μπορούμε να χαρακτηρίσουμε δύο επίπεδα δυνατοτήτων καλαθιού αγοράς που χρησιμοποιούνται από τους λιανοπωλητές:

Οι βασικές τεχνικές MBA. Αυτή η μέθοδος επιτρέπει στους λιανοπωλητές να δουν το μέγεθος και το περιεχόμενο του καλαθιού της αγοράς και να αναγνωρίσουν τις βασικές συγγένειες, αλλά δεν επιτρέπει τη διαδραστική εξερεύνηση των δεδομένων.

Προηγμένες τεχνικές MBA. Αυτή η μέθοδος προσφέρει πιο προηγμένες δυνατότητες αλληλεπίδρασης με τα δεδομένα της συναλλαγής για να ανακαλύψετε μοτίβα, συγγένειες και συσχετίσεις. Επίσης, βοηθά τους λιανοπωλητές να δουλέψουν πιο σκληρά, πιο έξυπνα, και σε πραγματικό χρόνο(Θεοδωρίδης & Πελέκης, 2011).

3.8.2 Πλεονεκτήματα που προσφέρει η ανάλυση καλαθιού αγοράς

1. Περισσότερο κερδοφόρα διαφήμιση. Οι έμποροι λιανικής πώλησης χρησιμοποιούν MBA για να κάνουν πιο στοχευμένη διαφήμιση και προωθητικές ενέργειες ώστε οι αγοραστές να ανταποκρίνονται σε διαφορετικές προσφορές. Για παράδειγμα, η τεχνική MBA μπορεί να βοηθήσει τους λιανοπωλητές να αποφευχθούν άσκοπες εκπτώσεις, όταν και όπου οι εκπτώσεις δεν αυξάνουν συνολικά το περιθώριο μικρού κέρδους. Οι έμποροι λιανικής πώλησης, επίσης, θέλουν να διαχωρίσουν τις τάσεις των πωλήσεων από την επίδραση της διαφήμισης για να καταλάβουν τη μετατόπιση των εσόδων.

2. Η ακριβέστερη στόχευση των προσφορών βελτιώνει την απόδοση της επένδυσης (ROI). Το MBA χρησιμοποιείται για τη βελτιστοποίηση των καμπανιών και των προωθήσεων για περιθώρια κέρδους και αύξηση

πωλήσεων με ακριβέστερη στόχευση. Για παράδειγμα, η αυξημένη ακρίβεια στη στόχευση στις προσφορές οδηγεί σε υψηλότερα ποσοστά κέρδους και επιτρέπει την πρόβλεψη των προτιμήσεων ώστε να προωθείται το κατάλληλο μείγμα προϊόντος στο σωστό πελάτη, τη σωστή στιγμή.

3. Ανάλυση των προτιμήσεων των καταναλωτών σε βάθος χρόνου. Η χρήση MBA για μεγάλο διάστημα επιτρέπει στους λιανοπωλητές να παρατηρούν την αγοραστική συμπεριφορά των πελατών στην πάροδο του χρόνου, αξιοποιώντας αυτή τη γνώση για την καλύτερη κατανόηση των πελατών τους. Οι έμποροι λιανικής πώλησης χρησιμοποιούν τις τεχνικές MBA για να λάβουν τα δεδομένα του κύκλου ζωής του πελάτη, έτσι ώστε να μπορούν να αναλύσουν τη συμπεριφορά της αγοραστικής ζωής του πελάτη, όπως η συχνότητα αγορών ή η περίοδος αυξημένων αγορών. Για παράδειγμα, ένας πωλητής παιχνιδιών εξήγησε ότι δεν έχει νόημα να πωλεί μηχανές βιντεοπαιχνιδιών (με πολύ μικρά περιθώρια) εκτός εάν ο πελάτης αγοράζει επίσης αξεσουάρ και το λογισμικό του παιχνιδιού (με υψηλά περιθώρια κέρδους).

4. Προσέλκυση μεγαλύτερης κυκλοφορίας στο κατάστημα. Οι έμποροι λιανικής πώλησης μπορούν να κατανοήσουν καλύτερα ποια προϊόντα και προσφορές θα φέρουν περισσότερους πελάτες στο κατάστημα συσχετίζοντας σε MBA την προσέλκυση πελατών με τις προσφορές ή με τα προϊόντα βοτρίνας.

5. Μπορεί να αυξηθεί το μέγεθος και η αξία του καλαθιού αγοράς. Με τα δεδομένα της πιστωτικής κάρτας, οι έμποροι λιανικής μπορούν να δουν πόσες φορές ο πελάτης ήταν στο κατάστημα και τα περιεχόμενα του καλαθιού του, και στη συνέχεια, να αξιοποιήσει αυτή τη γνώση με στόχο την αύξηση του μεγέθους του καλαθιού. Με το MBA, μπορούν να εντοπίζουν

και να στοχεύουν προωθητικές ενέργειες σε πελάτες οι οποίοι, για παράδειγμα, αγοράζουν όλες τις ανάγκες τους εκτός από συγκεκριμένα προϊόντα.

6. Ο πωλητής μπορεί να χρησιμοποιήσει ελεγχόμενα την αγορά σαν εργαστήριο. Ορισμένοι έμποροι λιανικής πώλησης χρησιμοποιούν MBA για τον προσδιορισμό της αξίας του μάρκετινγκ σε μια επίλεκτη «ομάδα καταναλωτών» των καταστημάτων, και, στη συνέχεια, εκτελούν την ανάλυση σε ένα άλλο «testgroup» των καταστημάτων.

7. Μπορεί να καθοριστεί το σημείο των τέλειων τιμών για ένα κατάστημα. Σήμερα, με τη χρήση παραδοσιακών εργαλείων συλλογής πληροφοριών, η βελτιστοποίηση των τιμών μπορεί να πάρει δύο ή τρεις εβδομάδες. Οι έμποροι λιανικής πώλησης θέλουν να είναι σε θέση να χρησιμοποιούν on-demand MBA ώστε να κάνει αυτές τις αποφάσεις σε σχεδόν πραγματικό χρόνο

8. Προσαρμογή των προϊόντων στα χαρακτηριστικά των καταναλωτών. Η κάθε επιχείρηση θα πρέπει να γνωρίζει τα χαρακτηριστικά ενός πληθυσμού στον οποία καλείται να πωλήσει ένα προϊόν. ΟΙ προτιμήσεις ποικίλλουν ανάλογα με το κλίμα, τη μόδα ή τα δημογραφικά στοιχεία, όπως το εισόδημα, η ηλικία ή η αστική τάξη έναντι των αγροτών.

9. Βελτιστοποιημένη διάταξη καταστημάτων. Οι έμποροι λιανικής πώλησης χρησιμοποιούν επίσης MBA για τη βελτίωση του σχεδιασμού χώρου(Θεοδωρίδης & Πελέκης, 2011).

3.8.3 Πως ενδυναμώνεται η επιχείρηση μέσω χρήσης MBA

Οι έμποροι πρέπει να δουν τις τάσεις πιο μακροπρόθεσμα για να αποφασίσουν πόσο να αγοράσουν και πώς κάποιο προϊόν ταιριάζει στο επιχειρηματικό μοντέλο. Εδώ είναι μερικοί τρόποι που κορυφαίοι λιανοπωλητές χρησιμοποιούν MBA.

Οι λιανοπωλητές χρειάζονται καλύτερα εργαλεία για να βοηθήσουν τους σχεδιαστές και τους εμπόρους. Ο προγραμματισμός είναι ένα από τα ταχύτερα αναπτυσσόμενα μέρη του λιανικού εμπορίου. Ένας καλός σχεδιασμός ρωτά: «Πόσο θα πρέπει να αγοράσουν οι καταναλωτές, πώς θα πάμε για να εμφανίζεται ένα προϊόν και ποια είναι η διάρκεια του κύκλου ζωής του προϊόντος» ή «Μπορούμε να πουλήσουμε αρκετά γρήγορα για να καλυφθεί το ύψος των δαπανών;». Αυτά τα υψηλής αξίας ερωτήματα και άλλες αποφάσεις υψηλού κινδύνου μπορούν να απαντηθούν ικανοποιητικά με τη χρήση MBA.

Η ανάλυση MBA δίνει τη δυνατότητα στους εμπόρους να αγοράζουν πιο έξυπνα και να ενισχύουν τη διαπραγματευτική θέση τους με τους προμηθευτές, παρέχοντας τους εμπόρους καλύτερες πληροφορίες σχετικά με την αγοραστική συμπεριφορά των πελατών. Ενώ ορισμένοι λιανοπωλητές προτιμούν να περιορίζουν τη χρήση του MBA στους σχεδιαστές, οι λιανοπωλητές πειραματίζονται όλο και περισσότερο με την παροχή στους εμπόρους δομημένων αλλά εύχρηστων εργαλείων MBA. Οι έμποροι επικεντρώνονται στην ανάγκη για την αγορά των αποθεμάτων, το σχεδιασμό και την διάθεση, αλλά μπορεί επίσης να ασχολούνται με τη διαφήμιση και προωθητικές ενέργειες. Η τεχνική MBA μπορεί να βοηθήσει στη βελτίωση των

αποφάσεων διάθεσης και αποθήκευσης, αλλά και στην καλύτερη κατανόηση της εποχικής ζήτησης(Trnka, 2010).

3.8.4 Προβλήματα της μεθόδου

Όλες οι τεχνικές έχουν τα δικά τους πλεονεκτήματα και μειονεκτήματα. Αυτή η ενότητα παρέχει μερικά από τα μειονεκτήματα των αλγορίθμων και τεχνικές για την υπέρβαση αυτών των δυσκολιών.

Μεταξύ των μεθόδων που συζητούνται για την εξόρυξη δεδομένων, ο αλγόριθμος apriori θεωρείται ο καλύτερος για την εξόρυξη κανόνων συσχετίσεων. Όμως υπάρχουν διάφορες δυσκολίες που αντιμετωπίζει ο αλγόριθμος apriori. Οι διάφορες δυσκολίες που αντιμετωπίζει ο αλγόριθμος apriori είναι:

1. Σαρώνει τη βάση δεδομένων πολλές φορές. Αυτό δημιουργεί την πρόσθετη εργασία για την αναζήτηση της βάσης δεδομένων. Επομένως, η βάση δεδομένων πρέπει να αποθηκεύει τεράστιο αριθμό δεδομένων, που πιθανώς να μην είναι απαραίτητα. Αυτό έχει ως αποτέλεσμα την έλλειψη μνήμης για την αποθήκευση αυτών των πρόσθετων δεδομένων. Επίσης, το φορτίο I / O δεν είναι επαρκές και παίρνει πολύ χρόνο για την επεξεργασία. Αυτό έχει ως αποτέλεσμα πολύ χαμηλή απόδοση.

2. Ένα συχνό στοιχείο με μεγαλύτερη απόσταση εμφάνισης, οδηγεί σε σημαντική αύξηση του χρόνου υπολογισμού.

3. Απαιτείται γενικότερη βελτίωση του αλγορίθμου.

Αυτά τα μειονεκτήματα μπορούν να ξεπεραστούν με την αποτελεσματική τροποποίηση του αλγορίθμου apriori. Η πολυπλοκότητα χρόνου για την εκτέλεση του Apriori μπορεί να λυθεί με τη χρήση ταχύτερου τροποποιημένου apriori αλγόριθμου. Για να συμβεί όμως αυτό θα πρέπει να θυσιαστεί ένα μέρος της ακρίβειας των αποτελεσμάτων(Trnka, 2010).

ΚΕΦΑΛΑΙΟ 4: ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

4.1 Στατιστική

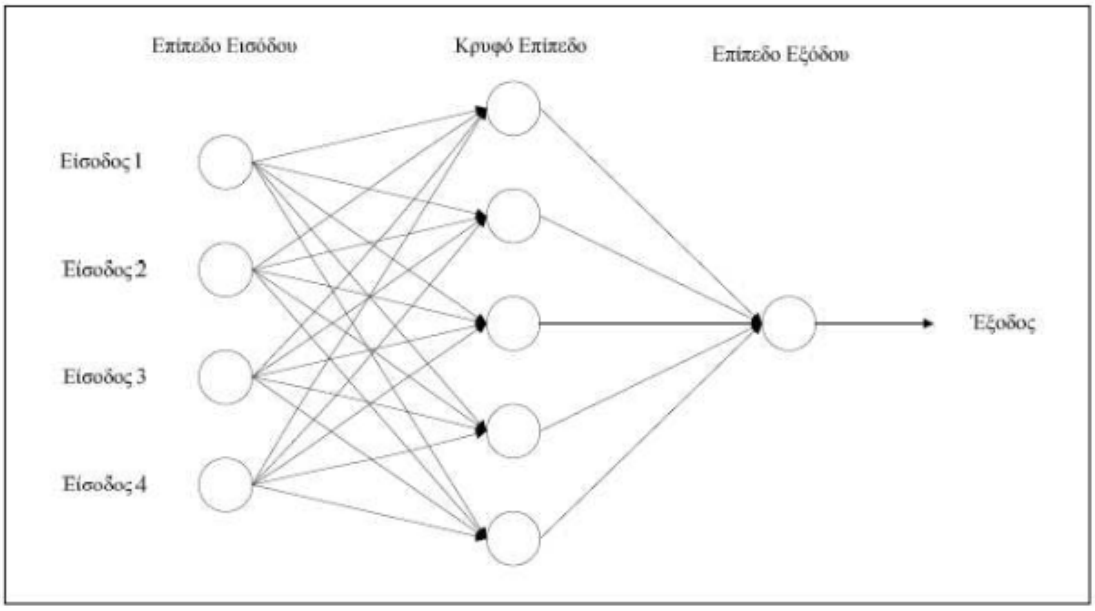
Το πρόβλημα της απόκτησης γνώσης από δεδομένα έχει αντιμετωπιστεί από τη στατιστική, πολύ πριν γίνουν τα πρώτα βήματα ανάπτυξης της τεχνητής νοημοσύνης. Για παράδειγμα, η ανάλυση συσχετίσεων εφαρμόζει στατιστικά εργαλεία για την ανάλυση της συσχέτισης μεταξύ δύο ή περισσοτέρων μεταβλητών. Η ανάλυση παραγόντων προσπαθεί να επισημάνει τις πιο σημαντικές μεταβλητές που περιγράφουν συγκεκριμένες ομάδες δεδομένων. Ορισμένες από τις δημοφιλείς τεχνικές που χρησιμοποιούνται για εργασίες ταξινόμησης είναι οι Iminary Immediate Linear Discr, οι Quadratic Discriminants, ο K-πλησιέστερος γείτονας, οι Μπεϋζιανοί ταξινομητές , η Logistic Regression και το CART. Ωστόσο, έχει γίνει εκτενής ανάλυση σε προηγούμενο κεφάλαιο.

4.2 Μηχανική μάθηση

Οι στατιστικές μέθοδοι υστερούν στην ενσωμάτωση υποκειμενικών, μη ποσοτικοποιήσιμων πληροφοριών στα μοντέλα τους. Πρέπει επίσης να υπολογιστούν και διάφορες κατανομές παραμέτρων και η ανεξαρτησία των χαρακτηριστικών. Διάφορες μελέτες κατέληξαν στο συμπέρασμα ότι η μηχανική μάθηση παράγει συγκρίσιμη (και συχνά καλύτερη) ακρίβεια πρόγνωσης. Η καλή επίδοσή της μηχανικής μάθησης σε σύγκριση με τις στατιστικές μεθόδους μπορεί να αποδοθεί στο γεγονός ότι είναι απαλλαγμένη από παραμετρικές και διαρθρωτικές υποθέσεις που αποτελούν τη βάση για τις στατιστικές μεθόδους. Μια άλλη αδυναμία των στατιστικών προσεγγίσεων για την ανάλυση δεδομένων είναι το πρόβλημα της ερμηνείας των αποτελεσμάτων(Hastie et al., 2001).

4.2.1 Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα είναι υπολογιστικά μοντέλα που αποτελούνται από πολλά μη γραμμικά στοιχεία επεξεργασίας που είναι διατεταγμένα σε ένα πρότυπο παρόμοιο με τα δίκτυα των βιολογικών νευρώνων. Ένα τυπικό νευρωνικό δίκτυο έχει μια τιμή ενεργοποίησης που συνδέεται με κάθε κόμβο και μια τιμή βάρους που συνδέεται με κάθε σύνδεση. Η λειτουργία ενεργοποίησης ρυθμίζει τη διάδοση δεδομένων μέσω συνδέσεων δικτύου. Το δίκτυο μπορεί επίσης να εκπαιδευτεί με παραδείγματα μέσω προσαρμογών βάρους σύνδεσης. Τα νευρωνικά δίκτυα είναι μια αναπαράσταση του τρόπου λειτουργίας του εγκεφάλου. Χρησιμοποιώντας ένα υπολογιστικό μοντέλο μαθαίνουν μέσω παραδειγμάτων, και με βάση αυτά αντιμετωπίζουν και λειτουργούν στις μελλοντικές εισόδους. Ένα τεχνητό νευρωνικό δίκτυο αποτελείται από ένα επίπεδο εισόδου και ένα ή πολλά επίπεδα εξόδου. Κάθε νευρωνικό δίκτυο μπορεί να θεωρηθεί ως πραγματικός νευρώνας που με βάση κάποιο ερέθισμα εκτελεί υπολογισμούς που με τη σειρά τους μεταφέρονται σε άλλο δίκτυο. Υπάρχει δυνατότητα σύνδεσης ενός νευρωνικού δικτύου με ένα άλλο, και η σύνδεση μπορεί να είναι ολική ή μερική(Hastie et al., 2001).



Σχήμα 4.1 Αναπαράσταση ενός νευρωνικού δικτύου

4.2.2 Γενετικοί Αλγόριθμοι

Οι γενετικοί αλγόριθμοι είναι αλγόριθμοι αναζήτησης που βασίζονται στη μηχανική της φυσικής επιλογής και της φυσικής γενετικής. Σκοπός είναι η επιβίωση των πιο κατάλληλων μεταξύ των δομών, σε δομημένα αλλά και τυχαία σύνολα. Σε κάθε γενιά, δημιουργείται μια νέα σειρά δομών με μέρη πιο ικανά της παλιάς γενιάς. Ένας απλός γενετικός αλγόριθμος που αποδίδει καλά αποτελέσματα, αποτελείται από τρεις φορείς: αναπαραγωγή, διασταύρωση και μετάλλαξη. Οι γενετικοί αλγόριθμοι διαφέρουν από τις πιο κανονικές διαδικασίες βελτιστοποίησης και αναζήτησης σε τέσσερα σημεία:

- Οι γενετικοί αλγόριθμοι λειτουργούν με την κωδικοποίηση του συνόλου παραμέτρων και όχι με την ίδια την παράμετρο.
- Οι γενετικοί αλγόριθμοι εξορύσσουν από έναν πληθυσμό τιμών, όχι μία τιμή.
- Οι γενετικοί αλγόριθμοι χρησιμοποιούν αντικειμενικές πληροφορίες λειτουργίας, όχι παράγωγα ή άλλες βοηθητικές γνώσεις.
- Οι γενετικοί αλγόριθμοι χρησιμοποιούν μεταβατικούς κανόνες πιθανότητας, όχι ντετερμινιστικούς κανόνες.

Οι γενετικοί αλγόριθμοι είναι εμπνευσμένοι από τη βιολογία, και τη θεωρία της εξέλιξης. Χρησιμοποιούν την λογική της φυσικής επιλογής. Αρχικά χρησιμοποιούνται κάποιοι κανόνες. Οι κανόνες αυτοί εφαρμόζονται στον πληθυσμό και αυτοί που έχουν τη μεγαλύτερη προσαρμοστικότητα στον πληθυσμό χρησιμοποιούνται για την παραγωγή καινούργιων. Η εκπαίδευση γίνεται σε έναν γνωστό πληθυσμό, με τη διαδικασία της ταξινόμησης. Ο κανόνας που έχει μεγαλύτερη ακρίβεια ταξινόμησης θεωρείται κατάλληλος.

Στους γενετικούς αλγόριθμους χρησιμοποιούνται οι λογικές επίσης της μετάλλαξης και της διασταύρωσης. Κατά τη διαδικασία της μετάλλαξης τυχαία ψηφία

αντιστρέφονται σε μια σειρά συμβόλων που αναπαριστούν τον κανόνα και στη διασταύρωση ανταλλάσσονται ψηφία από τη συμβολοσειρά ενός κανόνα σε έναν άλλο. Οι κανόνες αυτοί μπορούν ανά χρησιμοποιηθούν είτε για πρόβλεψη, είτε για ταξινόμηση. Στην εξόρυξη δεδομένων χρησιμοποιούνται και για να ελέγξουμε την καταλληλότητα των αλγορίθμων(Yanthy et al., 2009).

4.2.3 SVM

Τα SVM είναι οι μαθησιακές μηχανές που μπορούν να εκτελούν εργασίες δυαδικής ταξινόμησης και εκτίμησης παλινδρόμησης. Γίνονται όλο και πιο δημοφιλείς ως ένα νέο πρότυπο ταξινόμησης και μάθησης λόγω δύο σημαντικών παραγόντων. Πρώτον, αντίθετα με τις άλλες τεχνικές ταξινόμησης, τα SVM ελαχιστοποιούν το αναμενόμενο σφάλμα. Δεύτερον, τα SVM χρησιμοποιούν τη θεωρία δυαδικότητας του μαθηματικού προγραμματισμού για να δημιουργήσουν ένα σύστημα που χρησιμοποιεί αποδοτικές υπολογιστικές μεθόδους.

4.2.4 Δέντρα απόφασης

Ένα δέντρο απόφασης είναι ένα σχήμα ταξινόμησης που αποτελείται από ένα σύνολο κανόνων, όπου κάθε εσωτερικός κόμβος δηλώνει μια δοκιμή σε ένα χαρακτηριστικό, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα του τεστ και οι κόμβοι των φύλλων αντιπροσωπεύουν τις κλάσεις. Ο κορυφαίος κόμβος σε ένα δέντρο είναι ο κόμβος ρίζας(Yanthy et al., 2009).

4.2.5 Αλγόριθμοι Ακολουθιακής Κάλυψης

Η παραπάνω διαδικασία μπορεί να γίνει και εξ αρχής χωρίς την μετατροπή ενός δέντρου απόφασης. Αυτό επιτυγχάνεται με τους αλγορίθμους. ακολουθιακής

κάλυψης. Η διαδικασία γίνεται όταν ο αλγόριθμος παράγει από την αρχή τους κανόνες IF THEN μέσω εκπαίδευσης. Οι κανόνες εξάγονται ένας τη φορά, και από αυτό προέρχεται και το όνομα τους.

Οι πιο γνωστοί αλγόριθμοι ακολουθιακής κάλυψης είναι:

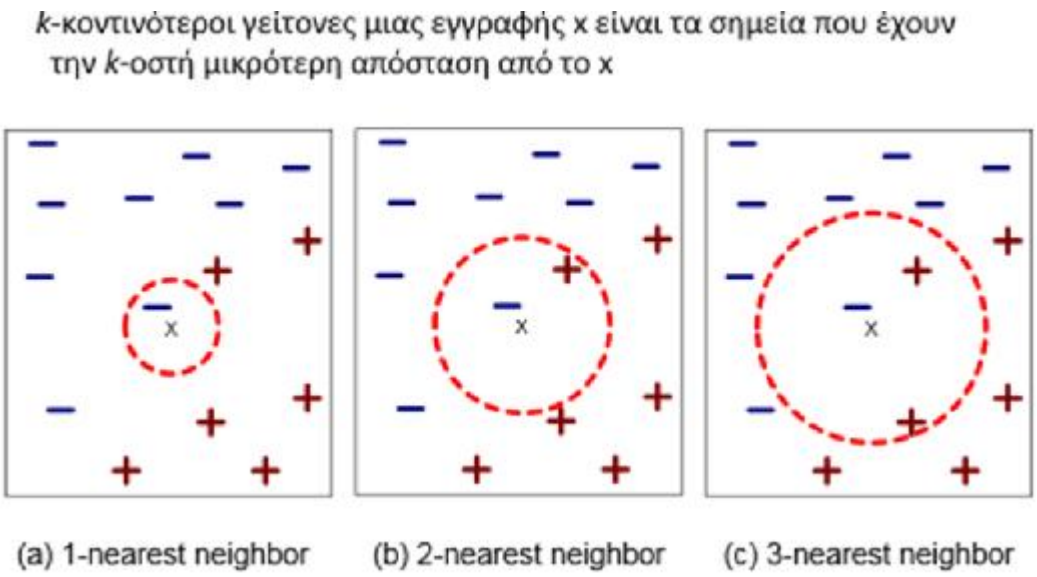
- AQ
- CN2
- RIPPER

Η λογική λειτουργίας τους είναι η εξής: Κάθε φορά που ένας κανόνας εφαρμόζεται απορρίπτονται τα στοιχεία που δεν πληρούν τον κανόνα με σκοπό να εξαχθεί μια κλάση. Κατόπιν αυτή η διαδικασία εφαρμόζεται από την αρχή. Παρατηρούμε τις διαφορές με τα δέντρα καθώς στα δέντρα εφαρμόζεται η εκμάθηση ταυτόχρονα.

Οι κανόνες που εξάγονται θα πρέπει να έχουν υψηλή ακρίβεια αλλά όχι απαραίτητα και υψηλή κάλυψη, καθώς μπορούν εφαρμόζονται εξ αρχής πολλές φορές, και να εφαρμόζονται περισσότεροι από ένας κανόνες για μια ομάδα. Η διαδικασία συνεχίζεται μέχρι να μην υπάρχουν άλλες πλειάδες εκπαίδευσης ή όταν οι κανόνες ικανοποιούν τον χρήστη, οπότε και τερματίζεται η διαδικασία(Yanthy et al., 2009).

4.2.6 Μέθοδος των k-Κοντινότερων Γειτόνων

Κατά τη μέθοδο αυτή μια ήδη γνωστή πλειάδα ελέγχου συγκρίνεται με άλλες πλειάδες εκπαίδευσης και με βάση το πόσο κοντά βρίσκονται στη γνωστή πλειάδα κατατάσσονται σε έναν ν-διάστατο χώρο. Αυτό εφαρμόζεται σε όλα τα δεδομένα και ταξινομούνται σε πλειάδες χώρου με βάση τους κ κοντινότερους γείτονες.



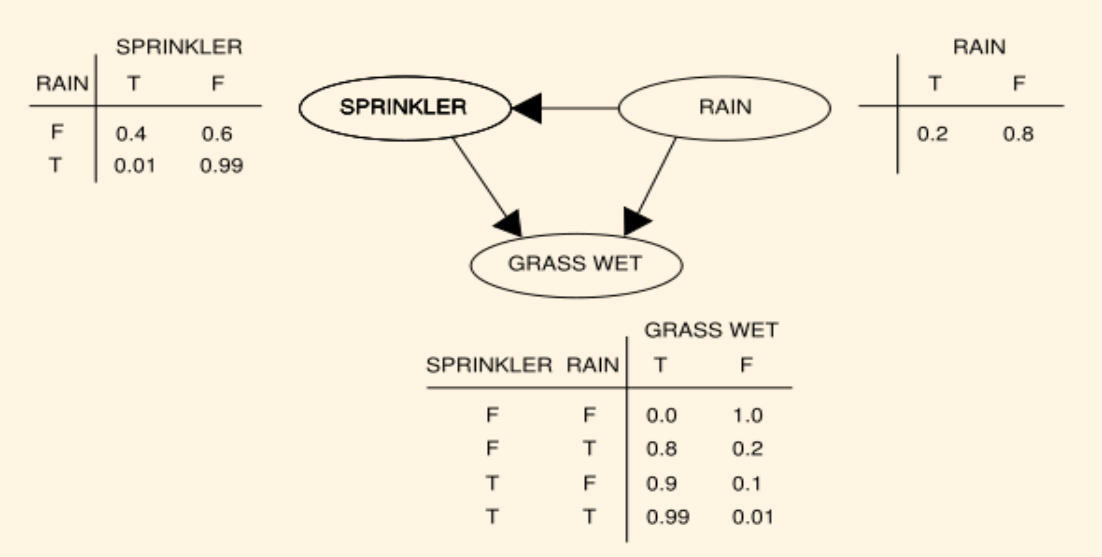
Σχήμα 4.2 Αναπαράσταση της μεθόδου των k-Κοντινότερων Γειτόνων (Εικόνα από τον ιστότοπο www.mines.humanoriented.com)

Η μέθοδος αυτή δεν θεωρείται ιδιαίτερα ακριβής στην περίπτωση που υπάρχουν πολλά διαφορετικά χαρακτηριστικά, γιατί κάθε χαρακτηριστικό έχει ίσο βάρος, ωστόσο μπορούν να γίνουν τροποποιήσεις και να υπάρξει διαφοροποίηση στη σημαντικότητα κάποιου χαρακτηριστικού(Weiss et al., 1991).

4.2.7 Μπεϋζιανά Δίκτυα

Οι συγκεκριμένοι ταξινομητές δεν κάνουν πρόβλεψη αλλά αποτίμηση πιθανότητας. Στόχος είναι το δείγμα να κατηγοριοποιηθεί σε κάποιες κλάσεις C1,C2,..,Cn χρησιμοποιώντας ένα μοντέλο πιθανότητας που ορίζεται σύμφωνα με τη θεωρία του Bayes. Είναι ένα γραφικό μοντέλο που κωδικοποιεί πιθανότητες σε ένα σύνολο μεταβλητών. Κάθε μεταβλητή σε ένα δίκτυο αναπαρίσταται με έναν κόμβο και κάθε κόμβος διαθέτει καταστάσεις ή διαφορετικά ένα σύνολο από πιθανές τιμές που αντιστοιχούν σε κάθε μεταβλητή. Οι κόμβοι συνδέονται μεταξύ τους με κατευθυνόμενα βέλη τα οποία δείχνουν την αλληλεξάρτηση των μεταβλητών, και

με ποια κατεύθυνση γίνεται αυτή η επιρροή. Κάθε βέλος αναπαριστά μια εξάρτηση πιθανότητας. Παράδειγμα αν το βέλος κατευθύνεται από έναν κόμβο Y σε έναν κόμβο Z, τότε ο Y είναι ένας γονέας ή άμεσος πρόγονος του Z, και ο Z είναι ένας απόγονος του Y.



Σχήμα 4.3 Ένα μπεϋζιανό δίκτυο, διαφαίνονται οι κόμβοι και η κατεύθυνση της πιθανοτικής εξάρτησης (Κωτσόπουλος, 2012).

Μπορεί να υπάρχουν παραπάνω από ένας κόμβοι εξόδου. Τα δίκτυα αυτά δεν έχουν προδιαγεγραμμένη λειτουργία, αλλά μπορούν να εφαρμοστούν διάφοροι αλγόριθμοι. Μπορεί να εξάγουμε κλάσεις ή πιθανότητα να ανήκει μια πλειάδα σε μια συγκεκριμένη κλάση(Weiss et al., 1991).

4.3 Ασαφής Κατηγοριοποίηση

Με εξαίρεση τους Μπεϋζιανούς ταξινομητές που μπορούν αναπαράγουν και πιθανότητα ένα στοιχείο να ανήκει σε μια κλάση, οι υπόλοιπες μέθοδοι ορίζουν αυστηρά ότι το δεδομένο ανήκει σε μια κλάση ή όχι. Η ασαφής κατηγοριοποίηση

εισάγει κάποιους βαθμούς αβεβαιότητας σε σχέση με το αν το δεδομένο είναι σε μια κλάση ή όχι.

Για την ταξινόμηση χρησιμοποιούνται επίσης ασαφείς κανόνες που μας δίνουν γενικό καθορισμό σε ποιες κλάσεις μπορεί να ανήκει μια ομάδα δεδομένων ή ένα δεδομένων. Για ένα δεδομένο xi και μια κλάση Ci προκύπτουν κανόνες της μορφής :

if {input is near xi} then class is Ci

Η έξοδος του συστήματος προκύπτει όταν ένα διάνυσμα ικανοποίησης μια συνθήκης ελέγχει τα αποτελέσματα. Έτσι καθορίζεται η καλύτερη δυνατή κλάση που μπορεί να ανήκει ένα δεδομένο(Yanthy et al., 2009).

ΚΕΦΑΛΑΙΟ 5: ΕΦΑΡΜΟΓΗ

5.1 Αλγόριθμος Apriori

Ο Apriori είναι ένας αλγόριθμος για την εξόρυξη και τη συσχέτιση στοιχείων με βάση συναλλαγές στις βάσεις δεδομένων συναλλαγών. Προχωράει εντοπίζοντας τα συχνότερα μεμονωμένα στοιχεία στη βάση δεδομένων και επεκτείνοντάς τα σε μεγαλύτερα και μεγαλύτερα σύνολα στοιχείων, αρκεί αυτά τα σύνολα στοιχείων να εμφανίζονται αρκετά συχνά στη βάση δεδομένων. Τα συνηθισμένα σύνολα στοιχείων που καθορίζονται από τον Apriori μπορούν να χρησιμοποιηθούν για τον καθορισμό των κανόνων συσχέτισης που υπογραμμίζουν τις γενικές τάσεις στη βάση δεδομένων, γεγονός έχει εφαρμογές σε τομείς όπως η ανάλυση καλαθιού αγοράς.

Ο αλγόριθμος Apriori προτάθηκε από την Agrawal και την Srikant το 1994. Ο Apriori έχει σχεδιαστεί για να λειτουργεί σε βάσεις δεδομένων που περιέχουν συναλλαγές (για παράδειγμα, συλλογές αντικειμένων που αγοράζονται από πελάτες ή στοιχεία ιστοσελίδων). Άλλοι αλγόριθμοι σχεδιάζονται για την εύρεση κανόνων συσχέτισης σε δεδομένα που δεν έχουν σχέση με συναλλαγές (Winepi και Minepi), ή δεν έχουν χρονολόγιο (αλληλούχιση DNA). Κάθε συναλλαγή θεωρείται ως ένα σύνολο στοιχείων. Δεδομένου ενός κατωφλίου C {\ displaystyle C} C, ο αλγόριθμος Apriori προσδιορίζει τα σύνολα στοιχείων που είναι υποσύνολα των συναλλαγών τουλάχιστον C {\ displaystyle C} C στη βάση δεδομένων.

Ο Apriori χρησιμοποιεί μια προσέγγιση "από κάτω προς τα πάνω", όπου συχνά υποσύνολα επεκτείνονται σε ένα στοιχείο κάθε φορά και ομάδες υποψηφίων συνόλων που δοκιμάζονται στα δεδομένα. Ο αλγόριθμος τερματίζεται όταν δεν βρεθούν άλλες επιτυχείς επεκτάσεις.

Ο Apriori χρησιμοποιεί δομή δέντρου Hash για να μετράει αποτελεσματικά τα υποψήφια στοιχεία. Δημιουργεί σύνολα υποψηφίων αντικειμένων τιμής k {\ displaystyle k} k από σύνολα θέσεων τιμής k - 1 {\ displaystyle k-1} k-1. Στη

συνέχεια, κατατάσσει το υποψήφιο σύνολο που περιέχει όλα τα συνηθισμένα σύνολα στοιχείων k {\ displaystyle k} k. Στη συνέχεια, ανιχνεύει τη βάση δεδομένων συναλλαγών για να καθορίσει συχνές σειρές στοιχείων μεταξύ των υποψηφίων στοιχείων.

Ο ψευδοκώδικας για τον αλγόριθμο δίνεται παρακάτω για μια βάση δεδομένων συναλλαγών T {\ displaystyle T} T, και ένα όριο υποστήριξης e {\ displaystyle \ epsilon} \ epsilon. Χρησιμοποιείται η συνήθης θεωρητική σημειογραφία, αν και σημειώστε ότι το T {\ displaystyleT} T είναι ένα πολλαπλό σύνολο. Το C k {{displaystyle C_ {k}} C_ {k} είναι το σύνολο υποψηφίων για το επίπεδο k {\ displaystyle k} k. Σε κάθε βήμα, ο αλγόριθμος θεωρείται ότι παράγει τα υποψήφια σύνολα από τα μεγάλα σύνολα στοιχείων του προηγούμενου επιπέδου, καλύπτοντας το προς τα κάτω. Ο αριθμός [c] αποκτά πρόσβαση σε ένα πεδίο της δομής δεδομένων που αντιπροσωπεύει το υποψήφιο σύνολο c {\ displaystyle c} c, το οποίο αρχικά θεωρείται μηδενικό.

5.2 WEKA

Το WEKA είναι μια σουίτα λογισμικού για μηχανική μάθηση και εξόρυξη δεδομένων. Αναπτύχθηκε στο πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και διατίθεται ως ελεύθερο λογισμικό. Η μεγάλη ποικιλία μεθόδων εξόρυξης δεδομένων που περιλαμβάνει, η συνεχής υποστήριξη και εξέλιξη του από μια διεθνή ομάδα προγραμματιστών, η ελεύθερη διανομή του πηγαίου κώδικα και η δυνατότητα εγκατάστασης του σε διαφορετικές πλατφόρμες υλικού και λογισμικού είναι ορισμένοι από τους παράγοντες που συμβάλλουν στην ευρύτερη αποδοχή και στη μεγάλη διάδοση του. Επίσης, η γραφική διεπαφή που διαθέτει επιτρέπει τη χρήση του από χρήστες, οι οποίοι δεν έχουν ικανότητες προγραμματισμού. Το παρόν κεφάλαιο αποτελεί μια σύντομη παρουσίαση του WEKA. Ειδικότερα, το κεφάλαιο αναφέρεται στο WEKA Explorer, το οποίο αποτελεί και τη δημοφιλέστερη διεπαφή. Με το WEKA Explorer ο χρήστης μπορεί να εκτελέσει εργασίες προεπεξεργασίας δεδομένων, κατηγοριοποίησης, ανάλυσης συστάδων, ανάλυσης κανόνων

συσχέτισης, επιλογής χαρακτηριστικών και οπτικοποίησης των δεδομένων. Σε ότι αφορά την προεπεξεργασία των δεδομένων, γίνεται αναφορά στις διάφορες πηγές δεδομένων και παρουσιάζονται τα αρχεία τύπου ARFF. Το γραφικό περιβάλλον του tab "Preprocess" επιτρέπει την εύκολη διερεύνηση της κατανομής τιμών στα διάφορα πεδία, τη διαγραφή πεδίων και την εκτέλεση διαφόρων αλγορίθμων προεπεξεργασίας, οι οποίοι εμφανίζονται υπό την ονομασία "filters". Παρέχονται εργαλεία για προσθήκη νέων υπολογιζόμενων πεδίων, για κανονικοποίηση και διακριτοποίηση αριθμητικών τιμών, για συγχώνευση ονομαστικών πεδίων, για δειγματοληψία, για μείωση διαστάσεων με Ανάλυση Κυρίων Συνιστωσών, για επιλογή χαρακτηριστικών κλπ.

Οι αλγόριθμοι και τα εργαλεία κατηγοριοποίησης που διαθέτει το WEKA είναι αξιοσημείωτοι. Παρέχονται υλοποιήσεις όλων των κύριων μεθόδων κατηγοριοποίησης, όπως Δένδρα Αποφάσεων, Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, Μπαϋεσιανοί κατηγοριοποιητές, Λογιστική Παλινδρόμηση, k-Πλησιέστεροι Γείτονες κλπ. Για κάθε μέθοδο υπάρχουν πολλές δυνατότητες παραμετροποίησης. Επίσης, διατίθενται πολλές παραλλαγές των βασικών μεθόδων, αλλά και εργαλεία για τη δημιουργία σύνθετων κατηγοριοποιητών bagging και boosting, κατηγοριοποιητών ευαίσθητων στο κόστος, κατηγοριοποιητών που χρησιμοποιούν ανάλυση συστάδων κλπ. Ο χρήστης μπορεί να επικυρώσει τα μοντέλα του εφαρμόζοντας τη μέθοδο cross validation, τη μέθοδο holdout ή χρησιμοποιώντας ένα ανεξάρτητο σύνολο δεδομένων. Για κάθε μοντέλο παρουσιάζονται αναλυτικά στοιχεία για τις επιδόσεις και τη δομή του (πχ τα βάρη των συνδέσεων ενός δικτύου Multilayer Perceptron). Το WEKA περιλαμβάνει αρκετούς αλγορίθμους Ανάλυσης Συστάδων, όπως τον k-Means, τη Συσσωρευτική Ιεραρχική ΑΣ και το DBSCAN. Κάθε αλγόριθμος μπορεί να παραμετροποιηθεί. Επίσης, υπάρχει δυνατότητα οπτικής αναπαράστασης της κατανομής των παρατηρήσεων στις συστάδες. Το tab "Associate" περιλαμβάνει αλγορίθμους για ανάλυση Κανόνων Συσχέτισης, μεταξύ των οποίων και τον βασικό αλγόριθμο Apriori. Υπάρχει η δυνατότητα εξόρυξης κανόνων συσχέτισης σε δεδομένα με πεδίο κλάσης. Οι κανόνες αυτοί θα έχουν στο δεξιό τμήμα τους μια τιμή κλάσης. Στο tab "Select attributes" ο χρήστης μπορεί να πειραματιστεί με

διάφορες μεθόδους επιλογής χαρακτηριστικών και να συνδυάσει μεθόδους αναζήτησης με μεθόδους αξιολόγησης χαρακτηριστικών. Τέλος, στο tab "Visualize" υπάρχει ένας πίνακας διαγραμμάτων διασποράς. Ο χρήστης, κάνοντας κλικ σε ένα διάγραμμα, μπορεί να το προβάλει σε ξεχωριστό παράθυρο.

5.3 Εφαρμογή

Το Weka είναι ένα λογισμικό που εφαρμόζει τεχνικές μηχανικής μάθησης με μια διαδικασία εύκολη, αποδοτική και διασκεδαστική. Πρόκειται για ένα εργαλείο GUI που σας επιτρέπει να φορτώσετε σύνολα δεδομένων, να εκτελέσετε αλγόριθμους και να σχεδιάσετε και να εκτελέσετε πειράματα με στατιστικά αρκετά ισχυρά αποτελέσματα για να δημοσιεύσετε. Το Weka συστήνεται και για αρχάριους στην εκμάθηση της συγκεκριμένης τεχνικής, διότι τους επιτρέπει να επικεντρωθούν στην εκμάθηση της διαδικασίας της εφαρμοσμένης μηχανικής μάθησης και όχι να αποξενωθούν από τα μαθηματικά και τον προγραμματισμό. Σε αυτήν την εργασία, θέλουμε να δείξουμε πόσο εύκολο είναι να φορτώσετε ένα σύνολο δεδομένων, να εκτελέσετε έναν προηγμένο αλγόριθμο ταξινόμησης και να εξετάσετε τα αποτελέσματα.

Τι ερευνούμε με αυτή την εφαρμογή;

Με αυτή την εφαρμογή γίνεται μια επίδειξη χρήσης του λογισμικού WEKA, για εξόρυξη δεδομένων και απόκτηση γνώσης από βάσεις δεδομένων από καλάθια αγορών. Συνεπώς, στο τέλος της εργασίας θα παρουσιαστούν κάποιοι κανόνες συσχέτισης μεταξύ προϊόντων, οι οποίοι μπορούν να χρησιμοποιηθούν με ποικίλες μεθόδους που αναφέραμε στα προηγούμενα κεφάλαια.

Που βρήκαμε τα δεδομένα;

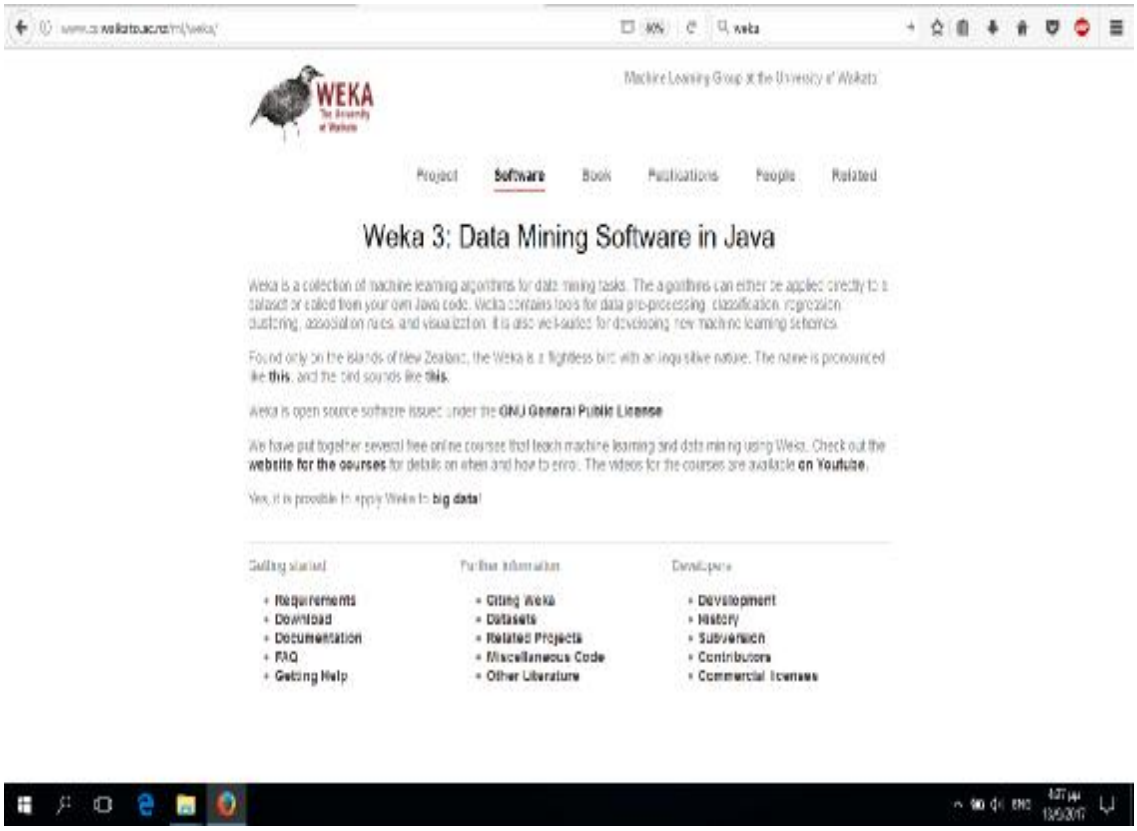
Τα δεδομένα που θα χρησιμοποιήσουμε είναι πραγματικά σύνολα που υπάρχουν άφθονα στο διαδίκτυο. Το σύνολο δεδομένων Iris Flower είναι ένα διάσημο σύνολο

δεδομένων από στατιστικά στοιχεία και δανείζεται σε μεγάλο βαθμό από ερευνητές στη μηχανική μάθηση.

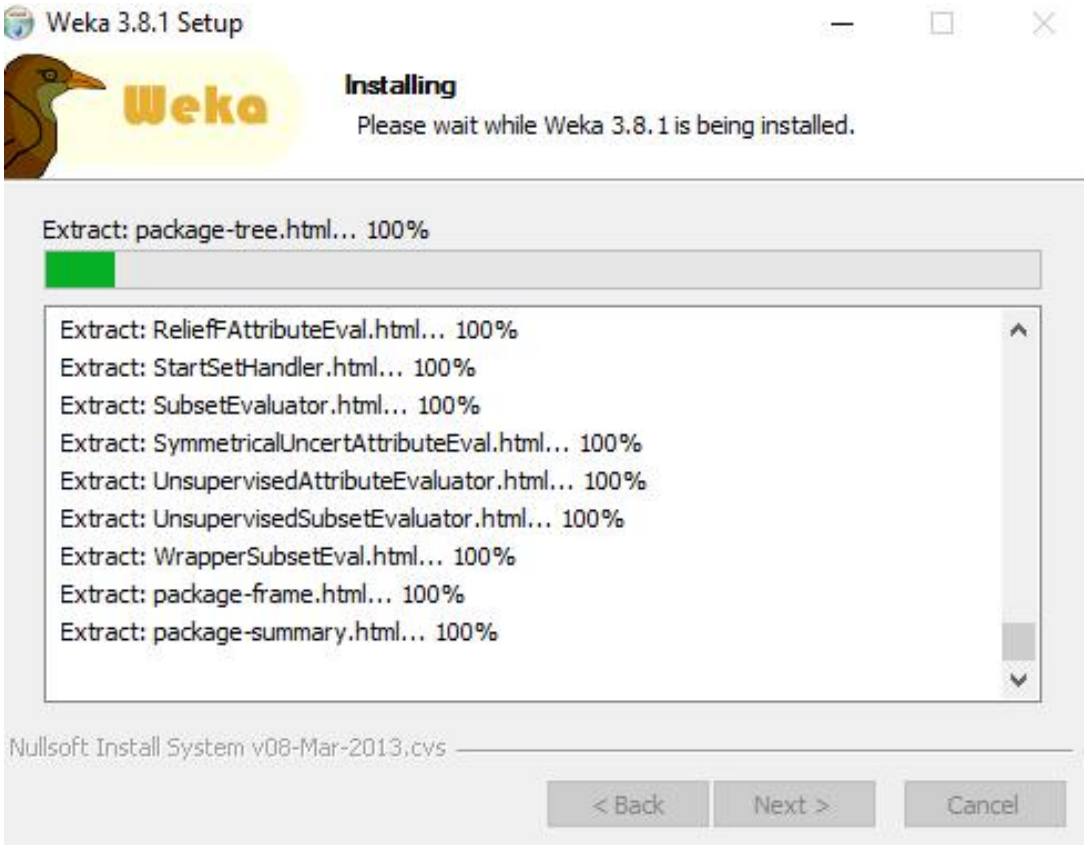
Το supermarket.arff είναι ένα σύνολο δεδομένων από POS. Τα δεδομένα είναι ονομαστικά και κάθε περίπτωση αντιπροσωπεύουν μια συναλλαγή ενός πελάτη σε ένα σούπερ μάρκετ, τα προϊόντα που αγοράστηκαν και τα εμπλεκόμενα τμήματα.

5.3.1. Κατέβασμα Weka

Επισκεφθείτε τη σελίδα λήψης Weka και εντοπίστε μια έκδοση του Weka που είναι κατάλληλη για τον υπολογιστή σας (Windows, Mac ή Linux).



Το Weka απαιτεί Java. Μπορεί να έχετε ήδη εγκαταστήσει Java και αν όχι, υπάρχουν εκδόσεις του Weka που αναφέρονται στη σελίδα λήψης (για Windows) που περιλαμβάνουν Java και το πρόγραμμα θα το εγκαταστήσει για σας.



5.3.2 Έναρξη Weka

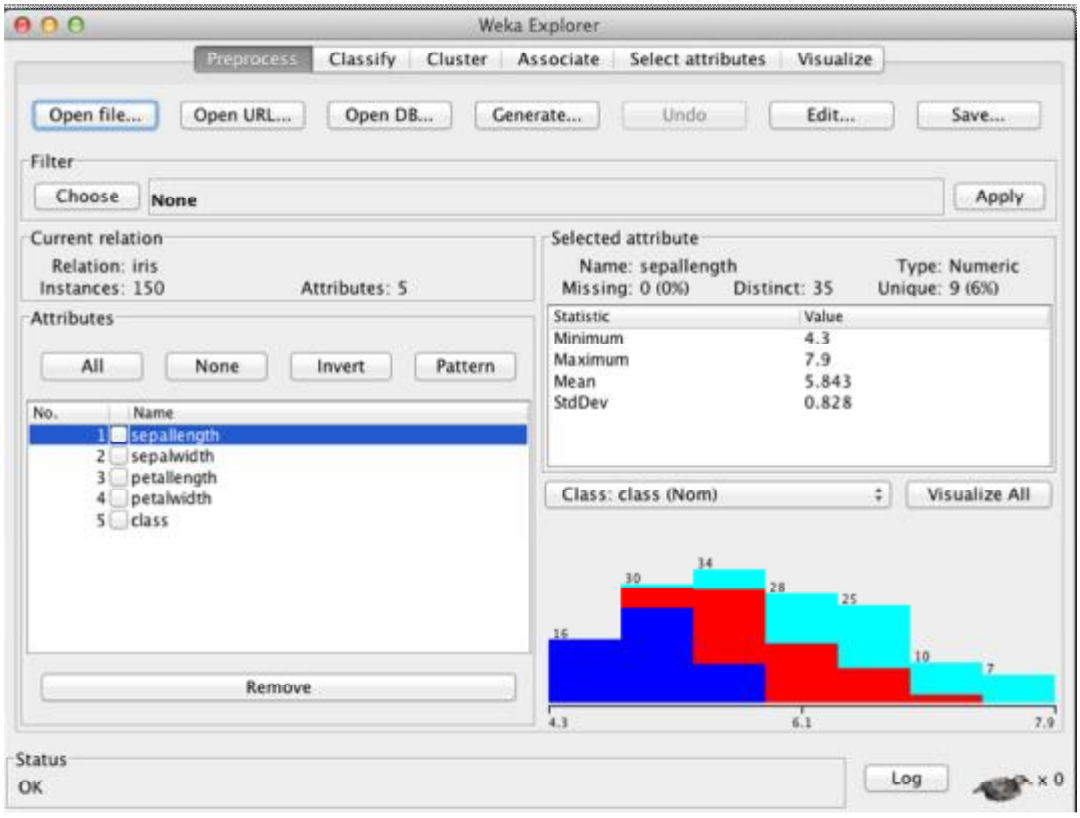
Ξεκινήστε το Weka. Αυτή η διαδικασία περιλαμβάνει την εύρεση του στο πρόγραμμα εκκίνησης ή το διπλό κλικ στο αρχείο weka.jar. Αυτό θα ξεκινήσει το Weka GUI.

Το Weka GUI σας επιτρέπει να επιλέξετε μία από τα Explorer, KnowledgeExplorer και το απλό CLI (περιβάλλον γραμμής εντολών).

Κάντε κλικ στο κουμπί «Explorer» για να ξεκινήσει το Weka Explorer.

Αυτό το GUI σας επιτρέπει να φορτώσετε σύνολα δεδομένων και να εκτελέσετε αλγόριθμους ταξινόμησης. Παρέχει επίσης άλλα χαρακτηριστικά, όπως το φιλτράρισμα δεδομένων, την ομαδοποίηση, την εξαγωγή κανόνων συσχέτισης, αλλά δεν θα χρησιμοποιήσουμε αυτά τα χαρακτηριστικά αυτήν τη στιγμή.

Ανοίγουμε το σύνολο δεδομένων / iris.arff



Κάνουμε κλικ στο κουμπί "Open file ..." για να ανοίξουμε ένα σύνολο δεδομένων και κάνουμε διπλό κλικ στον κατάλογο "data".

Το Weka παρέχει μια σειρά από μικρά σύνολα δεδομένων εκπαίδευσης που μπορούμε πριν την εφαρμογή να τρέξουμε για να δούμε το περιβάλλον και τον τρόπο λειτουργίας.

Επιλέγουμε το αρχείο «iris.arff» για να φορτώσουμε το σύνολο δεδομένων Iris.

Το σύνολο δεδομένων Iris Flower είναι ένα διάσημο σύνολο δεδομένων από στατιστικά στοιχεία και δανείζεται σε μεγάλο βαθμό από ερευνητές στη μηχανική μάθηση. Περιέχει 150 περιπτώσεις (σειρές) και 4 χαρακτηριστικά (στήλες) και ένα χαρακτηριστικό κατηγορίας για είδη λουλουδιών (setosa, versicolor, και virginica).

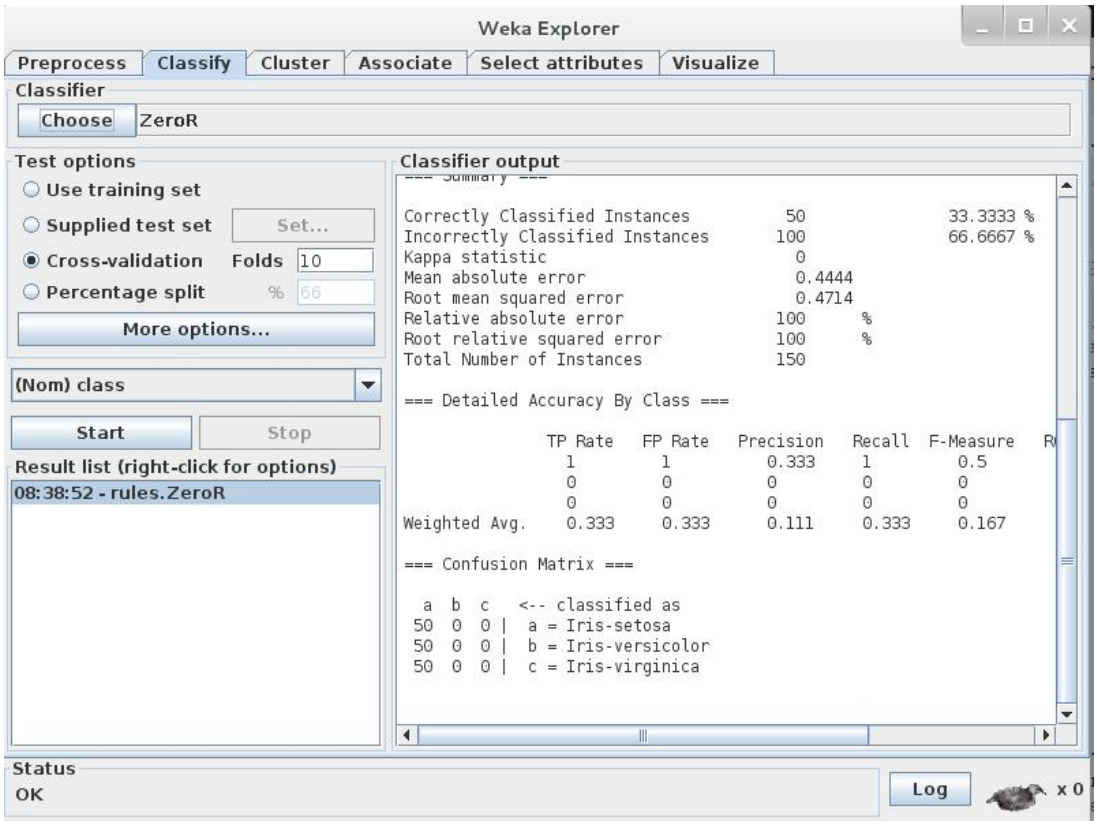
Επιλέγουμε και εκτελούμε έναν αλγόριθμο

Τώρα που έχουμε φορτώσει ένα σύνολο δεδομένων, ήρθε η ώρα να επιλέξουμε έναν αλγόριθμο μηχανικής μάθησης για να μοντελοποιήσουμε το πρόβλημα και να κάνουμε προβλέψεις.

Κάνουμε κλικ στην καρτέλα "Classify". Αυτή είναι η περιοχή για την εκτέλεση αλγορίθμων αφού ήδη έχει φορτωθεί σύνολο δεδομένων σε Weka.

Θα παρατηρήσετε ότι ο αλγόριθμος «ZeroR» είναι επιλεγμένος από προεπιλογή.

Κάνουμε κλικ στο κουμπί "Start" για να εκτελέσουμε αυτόν τον αλγόριθμο.



Το Weka δίνει αποτελέσματα από τον αλγόριθμο ZeroR στο σύνολο δεδομένων Iris.

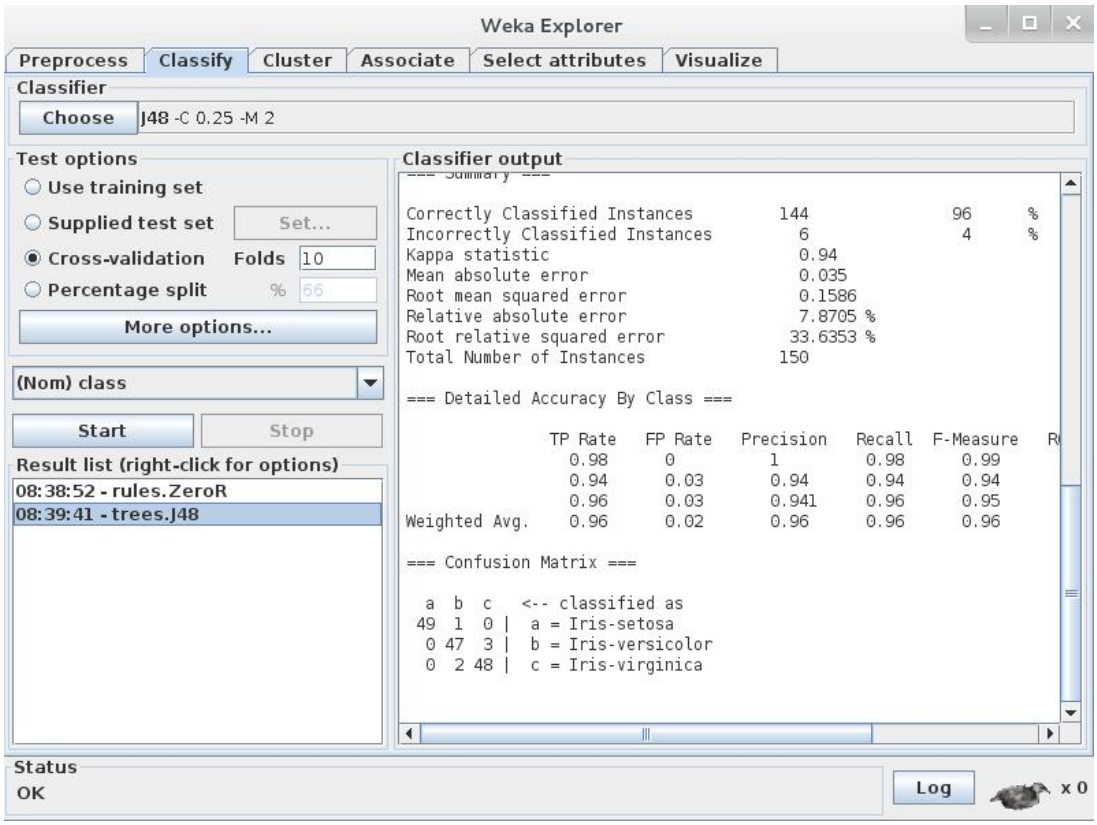
Ο αλγόριθμος ZeroR επιλέγει την πλειοψηφούσα τάξη στο σύνολο δεδομένων και χρησιμοποιεί αυτό το μέτρο για να κάνει όλες τις προβλέψεις. Αυτή είναι η βάση για το σύνολο δεδομένων και το μέτρο με το οποίο μπορούν να συγκριθούν όλοι οι αλγόριθμοι. Το αποτέλεσμα είναι 33%, όπως αναμενόταν (3 κατηγορίες, εκάστη εξ ίσου εκπροσωπούμενη, αποδίδοντας ένα από τα τρία σε κάθε πρόβλεψη οδηγεί σε 33% ακρίβεια ταξινόμησης).

Η επιλογές περνούν από επικύρωση δέκα συνολικά φορές. Αυτό σημαίνει ότι το σύνολο δεδομένων χωρίζεται σε 10 μέρη: τα πρώτα 9 χρησιμοποιούνται για την κατάρτιση του αλγορίθμου και ο δέκατος χρησιμοποιείται για την αξιολόγηση του αλγορίθμου. Αυτή η διαδικασία επαναλαμβάνεται, επιτρέποντας σε κάθε ένα από τα 10 μέρη του συνόλου δεδομένων που έχουν υποστεί διάκριση να λειτουργούν για την εφαρμογή δοκιμών.

Κάνουμε κλικ στο κουμπί "Choose" στην ενότητα "Classifier" και κλικ στον αλγόριθμο "J48".

Πρόκειται για μια εφαρμογή του αλγορίθμου C4.8 σε Java ("J" για Java, 48 για C4.8, εξ ου και το όνομα J48) και αποτελεί μια μικρή επέκταση στον περίφημο αλγόριθμο C4.5.

Κάνουμε κλικ στο κουμπί "Start" για να εκτελέσουμε τον αλγόριθμο.



5.3.3 Εξέταση αποτελεσμάτων

Αφού εκτελέσουμε τον αλγόριθμο J48, μπορούμε να σημειώσουμε τα αποτελέσματα στην ενότητα "Classifier output".

```
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   R
                 0.98      0          1         0.98      0.99
                 0.94      0.03       0.94      0.94      0.94
                 0.96      0.03       0.941     0.96      0.95
Weighted Avg.    0.96      0.02       0.96      0.96      0.96

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

Ο αλγόριθμος διεξήχθη με 10 φορές πολλαπλή επικύρωση: αυτό σημαίνει ότι του δόθηκε η ευκαιρία να κάνει μια πρόβλεψη για κάθε περίπτωση του συνόλου δεδομένων και το παρουσιαζόμενο αποτέλεσμα είναι μια σύνοψη αυτών των προβλέψεων.

Πρώτον, σημειώνουμε την ακρίβεια ταξινόμησης. Μπορείτε να δείτε ότι το μοντέλο πέτυχε ένα αποτέλεσμα 144/150 σωστό ή 96%, το οποίο φαίνεται πολύ καλύτερο από το βασικό επίπεδο του 33%.

5.3.4 Market Basket Analysis σε Weka

Το Data Mining λειτουργεί με βάση τη αρχή ότι οι αλγόριθμοι θα αναλύσουν τα δεδομένα και θα βρουν ενδιαφέροντα πρότυπα που θα μπορούσαμε να εκμεταλλευτούμε σε κάποια επιχείρηση.

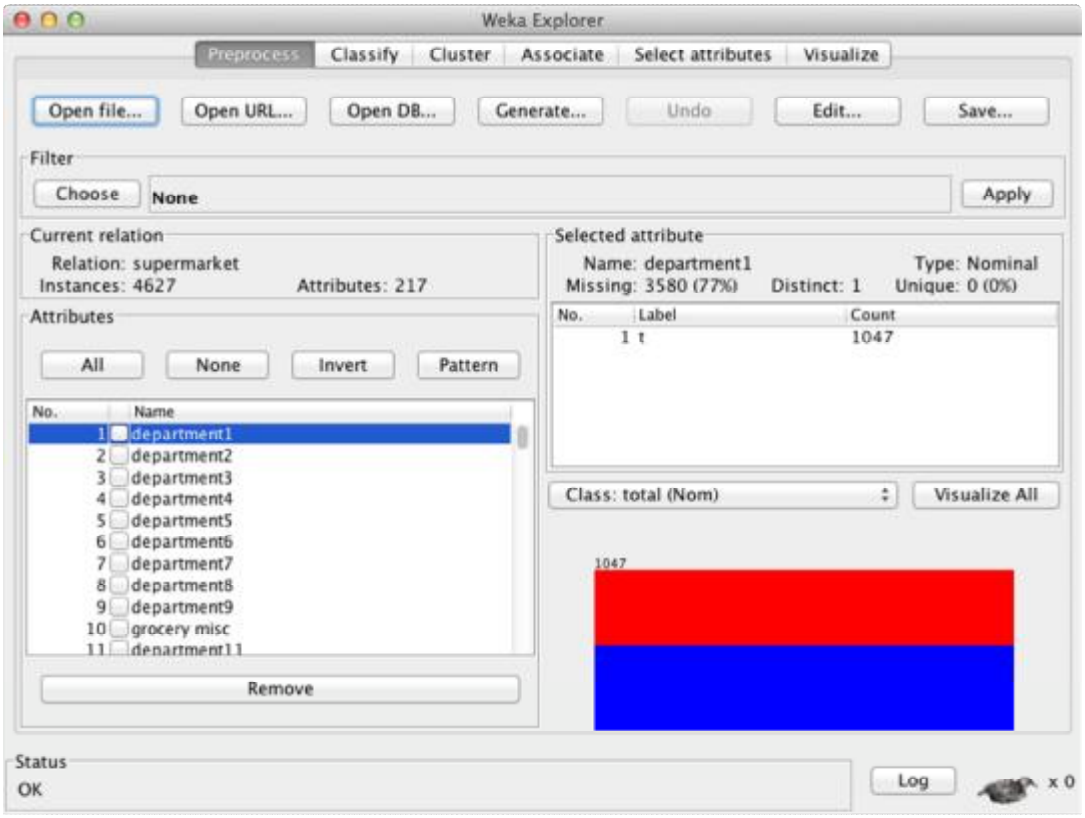
Ξεκινάμε από το Weka Explorer

Προηγουμένως, εξετάσαμε τη λειτουργία ενός ταξινομητή, το σχεδιασμό και τη διεξαγωγή ενός πειράματος, τον συντονισμό αλγορίθμων και τις μεθόδους επεξεργασίας ενός συνόλου δεδομένων.

Τοποθετούμε τα σύνολα δεδομένων

Το Weka εμπεριέχει μια σειρά από πραγματικά δεδομένα στον κατάλογο «data» της εφαρμογής Weka. Αυτό είναι πολύ βολικό, επειδή μπορούμε να εξερευνήσουμε και να πειραματιστούμε σε αυτά τα γνωστά προβλήματα και να μάθουμε για τις διάφορες μεθόδους του Weka.

Τοποθετούμε το σύνολο δεδομένων από καλάθι αγορών (supermarket.arff). Αυτό είναι ένα σύνολο δεδομένων από POS. Τα δεδομένα είναι ονομαστικά και κάθε περίπτωση αντιπροσωπεύουν μια συναλλαγή ενός πελάτη σε ένα σούπερ μάρκετ, τα προϊόντα που αγοράστηκαν και τα εμπλεκόμενα τμήματα. Δεν υπάρχουν πολλές πληροφορίες σχετικά με αυτό το σύνολο δεδομένων στο διαδίκτυο.



Τα δεδομένα περιέχουν 4.627 στοιχεία και 217 ιδιότητες. Κάθε ιδιότητα είναι σε δυαδικό σύστημα και έχει είτε μια τιμή ("t" για αληθινή) είτε καμία τιμή ("?" Για έλλειψη). Υπάρχει ένα χαρακτηριστικό ονομαστικής κλάσης που ονομάζεται "total" που υποδεικνύει αν η συναλλαγή ήταν μικρότερη από $ 100 (χαμηλή) ή μεγαλύτερη από 100 $ (υψηλή).

Δεν μας ενδιαφέρει να δημιουργήσουμε ένα πρότυπο πρόβλεψης για το σύνολο. Αντίθετα, μας ενδιαφέρει ποια στοιχεία αγοράστηκαν από κοινού. Ενδιαφερόμαστε για την εύρεση χρήσιμων μοτίβων σε αυτά τα δεδομένα τα οποία μπορεί να σχετίζονται ή να μην σχετίζονται.

Κανόνες συσχέτισης

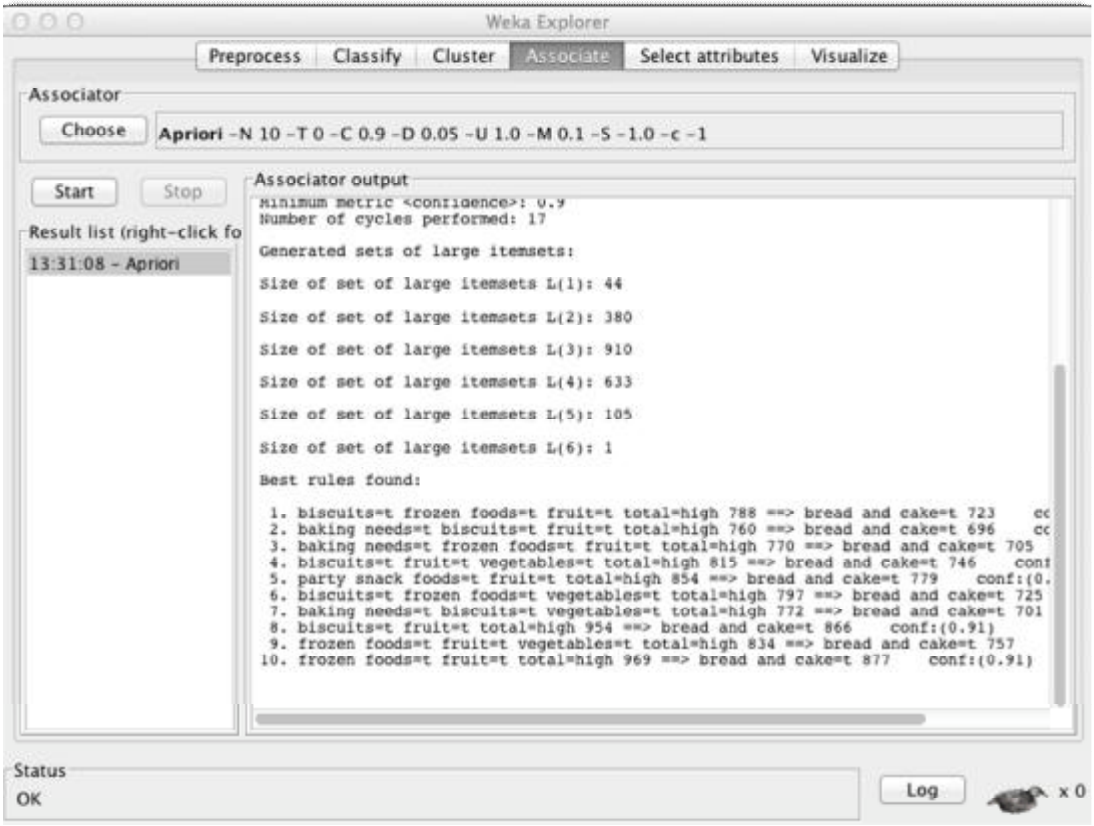
Εξετάζουμε ποιόν αλγόριθμο θα χρησιμοποιήσουμε. Θα χρησιμοποιήσουμε το αλγόριθμο "Apriori", ο οποίος θα είναι ήδη επιλεγμένος. Αυτή είναι η πιο γνωστή μέθοδος εκμάθησης κανόνων συσχέτισης, η πρώτη (Agrawal και Srikant το 1994) και είναι πολύ αποτελεσματική.

Κατ 'αρχήν ο αλγόριθμος είναι αρκετά απλός. Συγκροτεί σύνολα χαρακτηριστικών-τιμών (στοιχεία) που μεγιστοποιούν τον αριθμό των περιπτώσεων που μπορούν να εξηγηθούν (κάλυψη του συνόλου δεδομένων).

Κάνουμε κλικ στο κουμπί "Start" για να εκτελέσουμε τον Apriori στο σύνολο δεδομένων.

5.3.5 Εξαγωγή κανόνων

Η πραγματική εργασία για την εκμάθηση της συσχέτισης είναι η ερμηνεία των αποτελεσμάτων.



Αποτελέσματα του Apriori αλγόριθμου

Στο παράθυρο «exit associator», μπορείτε να δείτε ότι ο αλγόριθμος παρουσιάζει 10 κανόνες που αντλήθηκαν από το σύνολο δεδομένων για το σούπερ μάρκετ. Ο αλγόριθμος έχει ρυθμιστεί ώστε να σταματήσει στους 10 κανόνες από προεπιλογή, αλλά μπορείτε να κάνετε κλικ στο όνομα του αλγορίθμου και να ρυθμίσετε να βρείτε περισσότερους κανόνες, αν θέλετε, αλλάζοντας την τιμή «numRules».

Παρακάτω παρουσιάζονται οι κανόνες που εξήχθησαν:

biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 conf:(0.92)

baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 conf:(0.92)

baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 conf:(0.92)

biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 conf:(0.92)

party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 conf:(0.91)

biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 conf:(0.91)

baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 conf:(0.91)

biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 conf:(0.91)

frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 conf:(0.91)

frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 conf:(0.91)

Οι κανόνες παρουσιάζονται σε antecedent, δηλαδή ακολουθούμενη μορφή. Ο αριθμός που σχετίζεται με το προηγούμενο είναι η απόλυτη κάλυψη στο σύνολο δεδομένων (στην περίπτωση αυτή ένας αριθμός από ένα πιθανό σύνολο 4.627). Ο αριθμός δίπλα στον επακόλουθο είναι ο απόλυτος αριθμός περιπτώσεων που ταιριάζουν με το προηγούμενο και το επακόλουθο. Ο αριθμός στις αγκύλες στο

τέλος είναι η αξιοπιστία για τον κανόνα. Ως προεπιλογή είναι να μην χρησιμοποιηθεί κανένας κανόνας που να μην έχει αξιοπιστία μικρότερη του 91%.

Το ζήτημα με τους κανόνες ωστόσο είναι ότι παρότι μπορεί να υπάρχει κάποια στατιστική συσχέτιση δεν σημαίνει απαραίτητα ότι υπάρχει και σχέση αιτίας αιτιατού.

Εξήγηση των αποτελεσμάτων και συμπεριφορά του καταναλωτή

Έστω ότι η επιχείρηση η οποία καλείται να χρησιμοποιήσει τα παραπάνω δεδομένα είναι ένα σουπερμάρκετ. Τα δεδομένα από μόνα τους δεν μπορούν να παράγουν γνώση και η επιχείρηση πρέπει να βρει τους τρόπους να τα αξιοποιήσει εμπορικά. Συνεπώς, πως η επιχείρηση μπορεί να χρησιμοποιήσει τα παραπάνω αποτελέσματα για να προβλέψει τη συμπεριφορά του καταναλωτή; Με βάση λοιπόν τα παραπάνω αποτελέσματα παρατηρήσαμε ισχυρές συσχετίσεις σε πολλά από τα δεδομένα μας. Για παράδειγμα, η μπισκότων με ταυτόχρονη αγορά φρούτων συμπίπτει σε ποσοστό 86,6%. Η επιχείρηση μπορεί να χρησιμοποιήσει την παραπάνω πληροφορία με πολλούς τρόπους, μερικούς από τους οποίους παρουσιάζουμε παρακάτω:

- Μπορεί να αλλάξει τη θέση των προϊόντων, ώστε να είναι σε κοντινές θέσεις στο κατάστημα και στα ράφια, με σκοπό να αυξηθούν οι πωλήσεις του προϊόντος που πωλείται λιγότερο συχνά. Αν για παράδειγμα ένα συγκεκριμένο είδος μπισκότων έχει μειωμένη ζήτηση, τότε η μεταφορά του κοντά στην περιοχή των φρούτων αναμένεται να αυξήσει τις πωλήσεις του.
- Μπορεί η εταιρία να επιλέξει προϊόντα να οποία συνδυάζουν δύο ή περισσότερο προϊόντα που συσχετίζονται. Αν για παράδειγμα η αγορά γιαουρτιού συνδυάζεται με αγορά φρούτων η επιχείρηση μπορεί να αγοράσει προϊόν που συνδυάζει τα συσχετιζόμενα προϊόντα.
- Η επιχείρηση μπορεί να χρησιμοποιήσει κοινές προσφορές για συσχετιζόμενα προϊόντα. Οι προσφορές που θα γίνουν για κάποιο είδος μειωμένης ζήτησης μπορούν να συνδυαστούν με ένα άλλο είδος που θεωρούμε ότι θα αυξήσει τις πωλήσεις του.

- Μπορεί να γίνει κοινή διαφήμιση σε συσχετιζόμενα προϊόντα, όπως τα φρούτα και τα μπισκότα.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι πληροφορίες συλλέγονται σχεδόν παντού στην καθημερινότητά μας. Αυτό οδηγεί στην τεράστια αύξηση των διαθέσιμων δεδομένων. Η φυσική ανάλυση αυτών των τεράστιων ποσών πληροφοριών που είναι αποθηκευμένες στις σύγχρονες βάσεις δεδομένων είναι πολύ δύσκολη εώς αδύνατη. Η εξόρυξη δεδομένων παρέχει εργαλεία για την αποκάλυψη άγνωστων πληροφοριών σε μεγάλες βάσεις δεδομένων που υπάρχουν ήδη. Μια καλά γνωστή τεχνική εξόρυξης δεδομένων είναι η εξόρυξη κανόνων συσχέτισης.

Οι κανόνες συσχέτισης είναι πολύ αποτελεσματικοί στην αποκάλυψη όλων των ενδιαφερόντων σχέσεων σε μια σχετικά μεγάλη βάση δεδομένων με τεράστιο όγκο δεδομένων. Η μεγάλη ποσότητα των πληροφοριών που συλλέγονται μέσα από το σύνολο των κανόνων συσχέτισης μπορεί να χρησιμοποιηθεί όχι μόνο για την επεξήγηση των σχέσεων στη βάση δεδομένων, αλλά χρησιμοποιούνται επίσης για τη διαφοροποίηση μεταξύ των διαφόρων ειδών των κλάσεων σε μια βάση δεδομένων.

Η ανάλυση των υπαρχόντων αλγορίθμων δείχνει ότι η χρήση της ένωσης αλγορίθμων εξόρυξης κανόνα για την ανάλυση καλαθιού της αγοράς θα βοηθήσουν στην καλύτερη ταξινόμηση του τεράστιου όγκου των δεδομένων. Ο αλγόριθμος apriori μπορεί να τροποποιηθεί αποτελεσματικά για να μειώσει την πολυπλοκότητα του χρόνου και να βελτιώσει την ακρίβεια.

Η τεχνική MBA είναι χρήσιμη ως στρατηγικό εργαλείο για πωλητές και επιχειρήσεις που θα βοηθήσει να αυξήσουν την επιτυχία τους. Με τη χρήση της ανάλυσης καλαθιού αγοράς, οι κορυφαίοι λιανοπωλητές πετυχαίνουν αύξηση της ανταγωνιστικότητάς τους, εστιάζοντας απευθείας σε αγοραστικές συνήθειες του πελάτη, και στη συνέχεια, χρησιμοποιώντας αυτή τη γνώση για να προσαρμόσουν γρήγορα τις δραστηριότητές τους στις μεταβαλλόμενες ανάγκες των πελατών τους και των εμπορικών περιοχών.

Στο πρακτικό μέρος δείξαμε πως μπορεί να γίνει στην πράξη εξόρυξη δεδομένων και εξαγωγή κανόνων συσχέτισης. Το πρόγραμμα που χρησιμοποιήθηκε ήταν εύχρηστο και δεν παρουσιάστηκε κάποιο πρόβλημα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aczel, A.D. (1989). Complete Business Statistics, Irwin Series in Quantitative Analysis for Business, Irwin

Akaike, H. (1974). A new look at statistical model identification, IEEE Transactions on Automatic Control

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). Discrete Multivarate Analysis: Theory nad Practise, MIT Press.

Casella, G. and Berger, R.L. (2002). Statistical Inference, 2nd ed., Duxbury Advanced Series.

Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics, Chapman and Hall.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF)

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P. (1997). Statistical Themes and Lessons for Data Mining, Data Mining and Knowledge Discovery

Grabmeier, J. & Rudolph, A. (2002). Techniques of Cluster Algorithms in Data Mining, Data Mining and Knowledge Discovery

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On Clustering Validation Techniques, Journal of Intelligent Information Systems

Hand, D., Mannila, H. and Smyth, P. (2001). Principles of Data Mining, MIT Press, Cambridge

Hastie, T., Tibshirani, R. and Friedman, J. (2001). The elements of Statistical Learning: Data Mining, Inference and Prediction, Springer

R. Sumithra, S. Paul, “Using Distributed Apriori J Association Rule and Classical Apriori Mining Algorithms for Grid Based Knowledge Discovery,” International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1 - 5, 2010.

Ross, S. (1997). Simulation, 2nd ed., Academic Press.

Trnka, A.,’’Market Basket Analysis with Data Mining Methods”, International Conference on Networking and Information Technology (ICNIT). 446 - 450, 2010.

V. Bartik, “Association based Classification for Relational Data and its Use in Web Mining,” CIDM '09, IEEE Symposium on Computational Intelligence and Data Mining, Pp. 252 - 258, 2009.

W. Yanthy, T. Sekiya, K. Yamaguchi, “Mining Interesting Rules by Association and Classification Algorithms,” FCST '09. Fourth International Conference on Frontier of Computer Science and Technology, Pp. 177 - 182, 2009.

Weiss, S.I. and Kulikowski, C. (1991). Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems, Morgan Kaufmann

Xie Wen-xiu, Qi Heng-nian and Huang Mei-li, "Market Basket Analysis Based on Text Segmentation and Association Rule Mining", First International Conference on Networking and Distributed Computing (ICNDC), Pp. 309 - 313, 2010.

Y.J. Wang, Qin Xin, F. Coenen, “A Novel Rule Weighting Approach in Classification Association Rule Mining,” ICDM Workshops 2007, Seventh IEEE International Conference on Data Mining Workshops, Pp. 271 - 276, 2007.

Zhang, T., Ramakrishnan, R. and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases, In Proceedings of 1996 ACM-SIGMOD International Conference on Management of Data

Θεοδωρίδης Γ., Πελέκης Ν. (2011). Εξόρυξη Γνώσης από Δεδομένα - Συσταδοποίηση, Ομάδα Διαχείρισης Δεδομένων Πανεπιστήμιο Πειραιώς

Σταυλιώτης Ε. Γεράσιμος .(2009). Εξόρυξη Δεδομένων και Αναγνώριση προτύπων σε κατηγορικά δεδομένα μέσω συσταδοποίησης, Ελληνικό Στατιστικό Ινστιτούτο