ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΕΩΝ ΣΕ ΕΠΙΧΕΙΡΗΣΕΙΣ ΛΙΑΝΙΚΗΣ ΠΩΛΗΣΗΣ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ: ΧΑΛΙΚΙΑ ΕΛΕΝΗ ΜΑΡΙΑ & ΓΕΡΜΕΝΗ ΚΩΝΣΤΑΝΤΙΝΑ & ΚΑΤΣΙΟΥ ΜΑΡΙΑ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ: ΦΩΤΕΙΝΟΠΟΥΛΟΣ Μ.

ΠΑΤΡΑ 2016

ΠΡΟΛΟΓΟΣ

Η παρούσα εργασία έχει σκοπό την παρουσίαση των τεχνικών εξόρυξης δεδομένων και να παρουσιαστεί επίσης μια πρακτική εφαρμογή.

Κατά την εκπόνηση της εργασίας δεν αντιμετωπίσαμε κάποιο πρόβλημα, καθώς υπήρχε επαρκής διαθέσιμη βιβλιογραφία, ειδικά αγγλόφωνη. Οι χρήσεις του συγκεκριμένου πεδίου είναι πολλές και με συνεχώς αυξανόμενο ενδιαφέρον. Οι προκλήσεις όμως είναι εξίσου πολλές ιδιαίτερα στην περίπτωση που θέλουμε να επεξεργαστούμε Big Data.

Η εργασία γράφηκε με βάση τις προδιαγραφές που θέτει το Τεχνολογικό Εκπαιδευτικό Ίδρυμα Πατρών. Η γραμματοσειρά που χρησιμοποιούμε είναι Times New Roman 12 και χρησιμοποιούμε διάστιχο 1,5.

ΠΕΡΙΛΗΨΗ

Αρχικά θα γίνει μια εισαγωγή στις βασικές έννοιες της εξόρυξης δεδομένων. Στο πρώτο υποκεφάλαιο του πρώτου κεφαλαίου θα εξηγήσουμε τι είναι η εξόρυξη δεδομένων, θα αναλύσουμε τις χρήσεις της και τις βασικές διαδικασίες που ακολουθούνται. Ακολούθως θα μιλήσουμε πιο αναλυτικά για τη διαδικασία εξόρυξης δεδομένων, και θα αναλύσουμε για τα βασικά μοντέλα που χρησιμοποιούνται. Παρουσιάζονται οι κανόνες κατηγοριοποίησης και οι κανόνες συσχέτισης.

Θα αναλυθούν τα μέτρα αξιολόγησης κανόνων και το λογισμικό Weka. Στην πορεία θα μιλήσουμε πιο αναλυτικά για τις χρήσεις του data mining και ιδιαίτερα για τις χρήσεις από τις επιχειρήσεις.

Στα δύο τελευταία μέρη του πρώτου κεφαλαίου θα αναλύσουμε τη διαδικασία που προηγείται της εξόρυξης δεδομένων, την διαμόρφωση του dataset και την προετοιμασία πριν την εισαγωγή τους σε οποιοδήποτε ή σε συγκεκριμένο αλγόριθμο.

Στο δεύτερο κεφάλαιο θα αναλύσουμε την περίπτωση εξόρυξης δεδομένων σε καλάθι προϊόντων, στα αγγλικά market basket analysis. Θα παρουσιαστεί ο Apriori αλγόριθμος και θα δείξουμε τον τρόπο δημιουργίας Κανόνων Συσχέτισης από Frequent Itemsets.

Στο τρίτο κεφάλαιο θα γίνει αναλυτικά πρακτική εφαρμογή, όπου θα φαίνονται μέσω screenshots όλα τα βήματα της διαδικασίας.

Τέλος, θα παρουσιαστούν τα συμπεράσματα που εξαγάγαμε από την πρακτική εφαρμογή και η βιβλιογραφία που χρησιμοποιήθηκε.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ

ΚΕΦΑΛΑΙΟ 1: ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ (DATA MINING)

Εισαγωγή

Την τελευταία δεκαετία με την αυξημένη χρήση της τεχνολογίας και ειδικά του διαδικτύου συσσωρεύεται με αυξανόμενους ρυθμούς ένας τεράστιος όγκος δεδομένων. Τα δεδομένα αυτά που παλιότερα θα είχαν χαθεί τώρα με τη χρήση της τεχνολογίας είναι πολύ εύκολο να αποθηκευτούν τοπικά στους υπολογιστές μας και να γεμίσουν τους σκληρούς δίσκους των μεγάλων data centers. Παράγονται από μηχανήματα που αλληλεπιδρούν μεταξύ τους και καταγράφουν τις αποφάσεις μας, τις και τις προτιμήσεις μας. Από τις αγορές μας στο super market, τις οικονομικές μας συναλλαγές, τη μουσική που ακούμε και τις ταινίες που βλέπουμε τα μηχανήματα αυτά μονοπωλούν τις αισθήσεις και το μυαλό μας. Αυτή η συσσώρευση δεδομένων έρχεται σε αντιδιαστολή με την ικανότητά μας να τα επεξεργαζόμαστε και να βγάζουμε νόημα από αυτά. Μάλιστα όσο πιο πολύ αυξάνει ο όγκος τους τόσο πιο δύσκολο είναι να τα διαχειριστούμε. Κατά την διαδικασία εξόρυξης δεδομένων τα δεδομένα αποθηκεύονται ηλεκτρονικά και η διερεύνηση τους είναι αυτοματοποιημένη από υπολογιστή. Αν και αυτό δεν είναι μια ιδιαίτερα καινούρια ιδέα η αυτοματοποίηση της εύρεσης, επεξεργασίας επαλήθευσης και παρουσίασης μοτίβων μέσα στα δεδομένα είναι. Ιδιαίτερα εάν λάβουμε υπόψη μας ότι τα δεδομένα διπλασιάζονται κάθε είκοσι μήνες. Για παράδειγμα εάν θέλουμε να λύσουμε το πρόβλημα της αφοσίωσης των πελατών ενός τμήματος της αγοράς μπορούμε να βγάλουμε συμπεράσματα μελετώντας τα δεδομένα. Μία βάση δεδομένων με τις επιλογές των αγοραστών και μία βάση με τα profile τους θα μας δώσει την απάντηση. Χαρακτηριστικά αγοραστών μπορούν να συνδεθούν με τις αγοραστικές τους συνήθειες. Στην συνέχεια, από αυτά τα χαρακτηριστικά μπορούν να απομονωθούν τα διάφορα είδη των αγοραστών. Αυτοί που πιθανών να μείνουν πιστοί στο προϊόν που αγοράζουν και αυτοί που δεν θα μείνουν. Αφού απομονώσουμε τα είδη των αγοραστών που μας ενδιαφέρουν μπορούμε να κατασκευάσουμε τις στρατηγικές με τι οποίες θα αντιμετωπίσουμε αυτούς τους πελάτες.

1.1 Ορισμοί, χρήσεις, διαδικασία

Εξόρυξη δεδομένων είναι η επίλυση προβλημάτων αναλύοντας δεδομένα που υπάρχουν ήδη στις βάσεις δεδομένων. 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

Τα δεδομένα μπορούν να είναι οποιοδήποτε αντικείμενο ή στοιχείο μπορεί να επεξεργαστεί υπολογιστικό σύστημα. Μπορεί να είναι αριθμός, κείμενα. Στο διαδίκτυο υπάρχουν βάσεις που συγκεντρώνουν τα δεδομένα από διάφορες δραστηριότητες και με διαφορετικές κωδικοποιήσεις. Μπορεί να υπάρχουν δεδομένα με βάση το σκοπό χρήσης όπως τα παρακάτω:

- Δεδομένα πωλήσεων, προτιμήσεων σε πωλήσεις, λογιστικής και οικονομικής φύσεως κτλ που ονομάζονται επιχειρησιακά

- 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

Η πληροφορία είναι οι συσχετίσεις που μπορούμε να εξάγουμε με επεξεργασία των δεδομένων. Μπορούμε να αναγνωρίσουμε:

- Μοτίβα (patterns)
- Συσχετίσεις (associations)
- Συνάφειες (relationships)

Πιο συγκεκριμένα από ανάλυση δεδομένων μιας επιχειρήσεις μπορούμε να αναγνωρίσουμε κάποια συσχέτιση μεταξύ των προϊόντων που πωλούνται και των χαρακτηριστικών του καταναλωτή που τα επέλεξε. Για παράδειγμα το προϊόν Α έχει πωλήσεις όταν οι περισσότεροι πελάτες είναι γυναίκες.[1]

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.

[1] 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.

1.2 Διαδικασία εξόρυξης δεδομένων

Τα κύρια στάδια της εξόρυξης δεδομένων είναι τα εξής:

Προ-επεξεργασία

Πριν μπορέσουν να χρησιμοποιηθούν αλγόριθμοι εξόρυξης δεδομένων, πρέπει να δημιουργηθεί το σύνολο δεδομένων. Αυτό πρέπει να είναι αρκετά μεγάλο ώστε να περιέχει τα πρότυπα τα οποία εμφανίζονται στα δεδομένα, και συνάμα αρκετά συνοπτικό ώστε τα πρότυπα αυτά να εξορυχτούν μέσα σε ένα αποδεκτό χρονικό όριο. Η προ-επεξεργασία είναι απαραίτητη για την ανάλυση πολύ-παραγοντικών συνόλων δεδομένων πριν την εξόρυξη δεδομένων. Το σύνολο δεδομένων στη συνέχεια καθαρίζεται. Καθαρισμός Δεδομένων είναι η αφαίρεση των παρατηρήσεων που περιέχουν θόρυβο και εκείνων που έχουν ελλιπή δεδομένα.

1.2.1 Μοντέλα και διεργασίες

Μπορούμε να διακρίνουμε δυο τύπους μοντέλων εξόρυξης δεδομένων:

- Προβλεπτικά μοντέλα
- Περιγραφικά μοντέλα

Προβλεπτικό μοντέλο

Τα προβλεπτικά μοντέλα προσπαθούν να παράγουν προβλέψεις με βάση τις πληροφορίες που εξήχθησαν από την επεξεργασία των δεδομένων. Από το μοντέλο που δημιουργήθηκε με βάση τα ιστορικά δεδομένα παράγονται νέες τιμές θεωρητικών δεδομένων. Παράδειγμα είναι η πρόβλεψη του καιρού. Τα δεδομένα που

συλλέγονται από ανιχνευτές θερμοκρασίας, ατμοσφαιρικής πίεσης και υγρασίας κτλ, εισάγονται σε ένα μοντέλο πρόβλεψης του καιρού και γίνονται προβλέψεις για τις μελλοντικές τιμές των δεδομένων αυτών. Τα παραπάνω δεδομένα θα πρέπει να επεξεργάζονται και σε συνάρτηση με το χρόνο. Οι εργασίες εξόρυξης που γίνονται για τα προβλεπτικά μοντέλα είναι συνήθως οι παρακάτω:

- Κατηγοριοποίηση
- Παλινδρόμηση
- Ανάλυση χρονοσειρών
- Πρόβλεψη[2]

Κατηγοριοποίηση (Classification)

Κατά τη διαδικασία της κατηγοριοποίησης τα δεδομένα συγκεντρώνονται σε ομάδες ή τάξεις. Οι τάξεις έχουν προκαθοριστεί από πριν και μετά εισάγονται τα δεδομένα. Η διαδικασία της κατηγοριοποίησης είναι αντίστοιχη της αναγνώρισης μοτίβων, καθώς ένα δεδομένο εισάγεται σε μια ομάδα με βάση κάποια κοινά χαρακτηριστικά.

Παλινδρόμηση (Regression)

Κατά την παλινδρόμηση προσπαθούμε να ταιριάξει τα δεδομένα με κάποια ήδη γνωστή συνάρτηση, ώστε να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή πρόβλεψης και περιλαμβάνει την εκμάθηση της συνάρτησης που κάνει αυτή την απεικόνιση. Θα πρέπει βέβαια να βρεθεί μια συνάρτηση που να μπορεί να ταιριάξει με τα δεδομένα έστω προσεγγιστικά. Ακολούθως γίνεται και υπολογισμός του σφάλματος με βάση τη διαφορά πραγματικών και αναμενόμενων αποτελεσμάτων.

Ανάλυση Χρονοσειρών (Time Series Analysis)

Γίνεται συσχέτιση των παραπάνω δεδομένων και επεξεργασιών με το χρόνο. Κατόπιν παράγεται η χρονοσειρά και το διάγραμμα χρονοσειράς. Οι χρόνοι που λαμβάνονται

[2] Πολλές φορές κάποια από τα παραπάνω βήματα μπορούν να συνδυαστούν μεταξύ τους για το καλύτερο δυνατό αποτέλεσμα. Για παράδειγμα, τα βήματα του καθαρισμού και της ενσωμάτωσης των δεδομένων, μπορούν να υλοποιηθούν μαζί με στόχο την δημιουργία μια αποθήκης δεδομένων.

είναι τις περισσότερες φορές ίδιοι ώστε να μπορεί να γίνει σύγκριση και συσχέτιση των χρονοσειρών των γνωρισμάτων. Η ανάλυση των χρονοσειρών περιλαμβάνει τρεις διαφορετικές μεθόδους:
Αρχικά γίνεται σύγκριση των χρονοσειρών ώστε να εξαχθούν συμπεράσματα για την πιθανή ομοιότητα τους, μετά γίνεται ανάλυση της ίδιας της χρονοσειράς για την εύρεση των πιθανών κανόνων που διέπουν τη συμπεριφορά της και τέλος παράγεται η πρόβλεψη από τα διαγράμματα χρονοσειράς.

Πρόβλεψη (Prediction)
Σε αυτή την περίπτωση αναφέρεται σε τιμές μελλοντικής και όχι της τρέχουσας κατάστασης. Η πρόβλεψη σε αυτή την περίπτωση χρησιμοποιείται για ένα μεγάλο εύρος κλάδων, όπως τεχνητή μάθηση, πρόβλεψη καιρού και γενικά ακραίων καιρικών φαινομένων. Υπάρχουν και άλλες προσεγγίσεις από αυτές που αναφέραμε παραπάνω.

Περιγραφικό Μοντέλο (Descriptive Model)
Το περιγραφικό μοντέλο έχει διαφορετικό τρόπο προσέγγισης. Προσπαθεί να αναλύσει τα δεδομένα και όχι να προβλέψει. Η ανάλυση γίνεται μέσω μιας προσπάθειας να βρεθούν συσχετίσεις μεταξύ των δεδομένων. Για παράδειγμα σε μια εμπορική εταιρία χρησιμοποιώντας δεδομένα πώλησης αναγνωρίζεις τις συσχετίσεις με δημογραφικά χαρακτηριστικά των κατοίκων και αντίστοιχα προσαρμόζει τα προσφερόμενα προϊόντα. Στην αρχή γίνεται ομαδοποίηση των δεδομένων με βάση το συγκεκριμένο χαρακτηριστικό, και χρησιμοποιούνται για διαφήμιση βάση του χαρακτηριστικού αυτού.

Οι διεργασίες εξόρυξης δεδομένων για τα περιγραφικά μοντέλα είναι οι παρακάτω:

- Ομαδοποίηση (Clustering)
- Παρουσίαση Συνόψεων (Summarization)
- Κανόνες Συσχετίσεων (Association Rules)
- Ανακάλυψη Ακολουθιών

Ομαδοποίηση (Clustering)

Μετά την εύρεση κάποιας ομοιότητας στα δεδομένα αυτά χωρίζονται σε κατηγορίες. Αυτή η διαδικασία είναι διαφορετική της κατηγοριοποίησης γιατί οι ομάδες δεν είναι προκαθορισμένες, αλλά πηγάζουν από τα ίδια τα δεδομένα. Ονομάζεται επίσης μη εποπτευόμενη μάθηση. Σχετικότερα δεδομένα κατατάσσονται σε ίδιες ομάδες. Ένα γνωστό είδος ομαδοποίησης είναι η κατάτμηση, που από πολλούς βέβαια θεωρούνται και συνώνυμα.

Παρουσίαση Συνόψεων (Summarization)

Κατά τη διαδικασία αυτή, που ονομάζεται και χαρακτηρισμός ή γενίκευση, λαμβάνουμε πληροφορίες από γενικά χαρακτηριστικά των δεδομένων όπως ο μέσος όρος ή ο σταθμικός μέσος, ή ανακτώντας τμήματα των δεδομένων. Τα δεδομένα παρουσιάζονται ως υποσύνολα. Με αυτό τον τρόπο κάθε ομάδα δεδομένων αποκτά κάποια γενικά αναγνωριστικά χαρακτηριστικά.

Κανόνες Συσχετίσεων (Association Rules)

Οι κανόνες συσχετίσεων είναι μοντέλα που αναγνωρίζουν συσχετίσεις ανάμεσα στα δεδομένα. Υπάρχουν διάφοροι τρόποι για να γίνει αυτό, αλλά ο συχνότερος είναι μέσω στοιχειοσυνόλων. Για παράδειγμα ένας κανόνας συσχέτισης μπορεί να υπάρξει στην επανάληψη ενός δεδομένου πάνω από κάποιες φορές.

Ονομάζεται επίσης ανάλυση συγγένειας ή συσχέτιση. Έτσι αναγνωρίζουμε τα συχνά εμφανιζόμενα στοισειοσύνολα και βρίσκουμε τη συσχέτιση, αν υπάρχει, στην εμφάνιση τους. Οι συσχετίσεις ωστόσο πρέπει να ελέγχονται καθώς υπάρχει η πιθανότητα να είναι τυχαίες, και να μην αντιπροσωπεύουν καμία αληθινή σχέση μεταξύ των δεδομένων.

Ανακάλυψη Ακολουθιών

Η ανακάλυψη ακολουθιών είναι πανόμοια διαδικασία με τους κανόνες συσχετίσεων μόνο που στην περίπτωση αυτή γίνεται και σε συνάρτηση με το χρόνο. Παρακάτω θα

αναλύσουμε περεταίρω τις κατηγορίες εκπαίδευσης. Θα αναφέρουμε επίσης τις κατηγορίες εκπαίδευσης Εντοπισμός ανωμαλιών, Παλινδρόμηση και Περίληψη.

1.2.2 Κανόνες κατηγοριοποίησης

Πιο αναλυτικά η διαδικασία της κατηγοριοποίησης γίνεται μέσω της εξέτασης ενός χαρακτηριστικού ενός συνόλου δεδομένων και η τοποθέτηση του σε ένα σύνολο κλάσεων ή ομάδων. Μετά από αυτή τη διαδικασία που γίνεται κυρίως για ταξινόμηση των δεδομένων ώστε να υπάρχει ευκολότερη διαχείριση, παράγονται μοντέλα που χρησιμοποιούνται για περεταίρω ταξινόμηση των δεδομένων.

Η διαδικασία αυτή έχει δύο βήματα:

- Την εκμάθηση
- Την κατηγοριοποίηση

Κατά τη διαδικασία της εκμάθησης δημιουργούμε ένα μοντέλο από ήδη προκαθορισμένα στοιχεία. Οι μεταβλητές αυτές ονομάζονται training data ή δεδομένα εκπαίδευσης και αναλύονται από έναν αλγόριθμο ώστε να δημιουργηθεί το μοντέλο. Αποτελεί είδος εποπτευόμενης μάθησης γιατί τα στοιχεία που εισήχθησαν για τη δημιουργία του μοντέλου κατηγοριοποίησης είναι προκαθορισμένα από μας.

Κατά τη διαδικασία της κατηγοριοποίησης αξιολογούμε το μοντέλο που δημιουργήσαμε με την προηγούμενη διαδικασία. Υπάρχουν αρκετοί τρόποι να παράγουμε κανόνες κατηγοριοποίησης εκ των οποίων οι σημαντικότεροι είναι:

- Τα δέντρα απόφασης (decision trees)
- Οι κανόνες κατηγοριοποίησης (classification rules)
- Αλγόριθμοι Ακολουθιακής Κάλυψης (Sequential Covering Algorithms)
- Η μέθοδος των k-Κοντινότερων Γειτόνων (k-Nearest Neighbor – kNN)
- Τα Νευρωνικά Δίκτυα (Neural Networks)
- Οι Απλοϊκοί Μπευζιανοί Ταξινομητές (Naïve Bayesian Classifiers)
- Τα Μπευζιανά Δίκτυα (Bayesian Networks)
- Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

- Γενετικοί Αλγόριθμοι (Genetic Algorithms)
- Ασαφής Κατηγοριοποίηση

Τα δέντρα απόφασης

Η λογική των δέντρων αποφάσεων βασίζεται στη λογική της διαίρεσης των δεδομένων με συγκεκριμένα χαρακτηριστικά μέσα από κάποιες μεταβλητές. Το δέντρο αποφάσεων περιέχει κόμβους, κλαδιά και φύλλα. Έχει κατασκευαστεί με βάση ήδη κατηγοριοποιημένα δεδομένα, μετά από διαδικασία εκπαίδευσης. Κάθε κλάση του δέντρου πριν τον κόμβο αντιστοιχεί σε ένα συγκεκριμένο χαρακτηριστικό και επιλέγεται με βάση την τιμή του κάθε στοιχείο στο χαρακτηριστικό αυτό. Τα φύλλα αντιστοιχούν στις προκαθορισμένες κλάσεις. Άρα σε κάθε κόμβο γίνεται δοκιμή και αν το δεδομένο πληροί τις προϋποθέσεις συνεχίζει μέχρι να καταλήξει σε μία κατηγορία φύλλο. Από τη ρίζα του δέντρου μέχρι τον πρώτο κόμβο, σε κάθε κόμβο εξετάζονται τα γνωρίσματα διαδοχικά με το αν ικανοποιούν τα χαρακτηριστικά που απαιτεί ο κόμβος. Οι πιο γνωστοί αλγόριθμοι για την κατασκευή δέντρων απόφασης είναι οι: ID3, C4.5, SPRINT, SLIQ, CART, RainForest0.

income range	*life insurance promotion*	*credit card insurance*	*sex*	*age*
40-50 **χιλ**	OXI	OXI	Άντρας	45
30-40 **χιλ**	NAI	OXI	Γυναίκα	40
40-50 **χιλ**	OXI	OXI	Άντρας	42
30-40 **χιλ**	NAI	NAI	Άντρας	43
50-60 **χιλ**	NAI	OXI	Γυναίκα	38
20-30 **χιλ**	OXI	OXI	Γυναίκα	55
30-40 **χιλ**	NAI	NAI	Άντρας	35
20-30 **χιλ**	OXI	OXI	Άντρας	27
30-40 **χιλ**	OXI	OXI	Άντρας	43
30-40 **χιλ**	NAI	OXI	Γυναίκα	41
40-50 **χιλ**	NAI	OXI	Γυναίκα	43
20-30 **χιλ**	NAI	OXI	Άντρας	29
50-60 **χιλ**	NAI	OXI	Γυναίκα	39
40-50 **χιλ**	OXI	OXI	Άντρας	55
20-30 **χιλ**	NAI	NAI	Γυναίκα	19

Πίνακας 1.1 Παράδειγμα δέντρου απόφασης (Παπαδόπουλος, 2012)

Οι φάσεις δημιουργίας ενός δέντρου αποφάσεων είναι δύο: η φάση της οικοδόμησης και η φάση του κλαδέματος. Κατά τη φάση της οικοδόμησης όλα τα δεδομένα κάθε κατηγορίας τοποθετούνται στην κλάση τους μέσω συνεχών διακλαδώσεων και κατά τη φάση του κλαδέματος διορθώνεται η ακρίβεια.

Τα κύρια πλεονεκτήματα των δέντρων απόφασης είναι η ευκολία στη χρήση και η αποτελεσματικότητα τους και μπορούν να χρησιμοποιηθούν με ακρίβεια για πολλές βάσεις δεδομένων. Μπορούμε να κατασκευάσουμε επίσης δέντρα για επεξεργασία δεδομένων με πολλά διαφορετικά και ειδικά χαρακτηριστικά.

Τα κύρια μειονεκτήματα τους είναι το του υπερταιριάσματος, η δυσκολία χρήσης όταν υπάρχουν ελλιπή δεδομένα.

Τα δεδομένα κάθε προβλήματος μπορούν να αναπαρασταθούν με γραφική μέθοδο μέσω του δέντρου απόφασης. Πρακτικώς, είναι μια γραφική παράσταση που παρουσιάζονται όλοι οι παράγοντες επιρροής και οι δυνατές επιλογές ενός προβλήματος, αλλά και περισσότερων του ενός προβλήματος ανά περίπτωση

Με τα δέντρα αποφάσεων μπορεί ο υπεύθυνος να έχει μια καλύτερη εικόνα των περιβλημάτων και τι προηγείται και έπεται αυτών. Συνήθως κάθε απόφαση ενός προβλήματος ακολουθείται από συγκεκριμένες εναλλακτικές λύσεις για τις οποίες πρέπει να είμαστε ενήμεροι κατά τη λήψη αποφάσεων. Είναι ιδιαίτερα χρήσιμα στην περίπτωση που λαμβάνουμε κάποια απόφαση υπό συνθήκες αβεβαιότητας και στην περίπτωση που λαμβάνουμε κάποια απόφαση υπό συγκεκριμένο χρόνο. Κάθε δέντρο απόφασης μπορεί να χρησιμοποιηθεί και από αλγόριθμους τεχνητής νοημοσύνης και να παράγεται αυτόματα.

Υπάρχουν δυο πιθανές πορείες για τη δημιουργία δέντρων απόφασης:

Η πορεία προς τα εμπρός, στην οποία συμμετέχει και ο λήπτης της απόφασης. Στην πορεία αυτή φαίνεται η διάρθρωση και οι διακλαδώσεις του προβλήματος. Σε κάθε

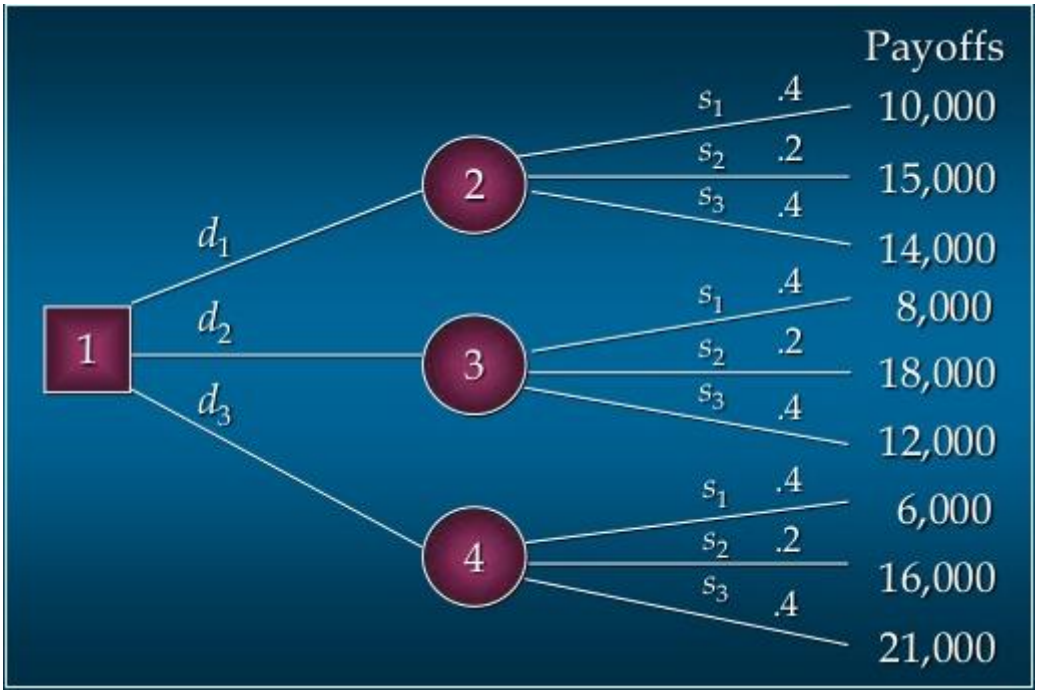
διακλάδωση απαιτείται μια απόφαση, και θα πρέπει να γίνει εκτενής υπολογισμός των χρηματικών απολαβών σε κάθε βήμα της πορείας, για να γίνει ορθή επιλογή. Επίσης, θα πρέπει να γίνεται καταγραφή και των πιθανοτήτων εμφάνισης των παραγόντων επιρροής.

Η προς τα πίσω πορεία, η οποία προσφέρεται για ανάλυση του προβλήματος και ανάγνωση των δυσκολιών λύσης του. Αυτή η διαδικασία είναι σχετικώς αντιστροφή της παραπάνω, αλλά δεν είναι απαραίτητη η παρουσία του λήπτη κατά τη διαδικασία αυτή.

Παρακάτω ακολουθεί παράδειγμα δέντρου απόφασης όπου διαφαίνονται τα στοιχεία της μήτρας αποφάσεων:

Κάθε δέντρο αποφάσεων αποτελείται από κόμβους και κλαδιά. Κάθε κόμβος αντιπροσωπεύει ένα συγκεκριμένο πρόβλημα και κάθε κλαδί μια συγκεκριμένη επιλογή. Επίσης κάθε κλαδί μπορεί αν αντιπροσωπεύει ένας παράγοντας επιλογής. Αν ο κόμβος αντιπροσωπεύει μια επιλογή d1, d2, d3 κτλ. τότε ονομάζεται κόμβος απόφασης. Αν αντιπροσωπεύει παράγοντα επιρροής ονομάζεται κόμβος παραγόντων επιρροής.

Αν έχει την περίπτωση κόμβων απόφασης τότε ο υπεύθυνος καλείται να διαλέξει μια οδό που αντιστοιχεί σε ένα κλαδί που με την σειρά του αντιπροσωπεύει την βέλτιστη επιλογή. Σε αυτή την περίπτωση ο ρόλος του υπεύθυνου είναι ενεργητικός. Αν ο κόμβος αντιπροσωπεύει παράγοντα επιρροής τότε ο υπεύθυνος δεν μπορεί να επιλέξει, και η επιλογή κλαδιού καθορίζεται από τις πιθανότητες εμφάνισης του συγκεκριμένου παράγοντα. Σε αυτή την περίπτωση ο ρόλος του υπευθύνου είναι παθητικός. Για την επιλογή της βέλτιστης απόφασης σε κάθε κόμβο ο υπεύθυνος μπορεί να χρησιμοποιήσει ένα από τα παραπάνω κριτήρια(www.palisade.com).



Σχήμα 1.1 Δέντρο απόφασης με τρεις αποφάσεις και τρεις παράγοντες επιρροής (Παπαδόπουλος, 2010).

Μειονεκτήματα της χρήσης δέντρων απόφασης

Όπως παραπάνω αναφέραμε, η χρησιμότητα των δέντρων αποφάσεων είναι μεγάλη και μπορεί να οδηγήσει τον λήπτη αποφάσεων στην αντιμετώπιση κάποιου προβλήματος απόφασης αλλά και επιλογής της ευνοϊκότερης λύσης. Ωστόσο, τα δέντρα αποφάσεων έχουν δεχθεί αρκετές κριτικές και έχουν συχνά αποδοκιμαστεί από τους αναλυτές όσον αφορά την χρησιμότητα τους. Κάποιοι αναλυτές τα χαρακτήρισαν μικρής αξίας ή και δύσκολης εφαρμογής σε πραγματικές καταστάσεις. Προσπαθώντας να συγκεντρώσουμε τα κυριότερα μειονεκτήματα της χρήσης δέντρων αποφάσεων, καταλήξαμε στα εξής:[3]

[3] Η έκβαση της εξέτασης καθορίζει το κλαδί που θα ακολουθηθεί στην συνέχεια, καθώς και τον επόμενο κόμβο. Η κλάση στην οποία θα ταξινομηθεί το νέο αντικείμενο αντιστοιχεί σε ένα από τα φύλλα του δέντρου απόφασης.

- Δυσκολία εκτίμησης των πιθανοτήτων που αφορούν τα διάφορα πιθανά σενάρια για το μέλλον, αφού η λήψη απόφασης χαρακτηρίζεται από το στοιχείο της αβεβαιότητας
- Δυσκολία αναπαράστασης του προβλήματος ώστε να απεικονίζει και να ανταποκρίνεται στις πραγματικές συνθήκες που αναπτύσσεται
- Πολυπλοκότητα δόμησης προβλημάτων με πολλούς παράγοντες
- Πολλές εναλλακτικές αποφάσεις (Κόλλια, 2012)

Κανόνες Κατηγοριοποίησης

Αν το δέντρο απόφασης είναι πολύ μεγάλο και δύσχρηστο, τότε μπορούμε να το αντικαταστήσουμε με ένα σύνολο κανόνων κατηγοριοποίησης. Η μετατροπή γίνεται αν κάθε μονοπάτι αντικατασταθεί με έναν κανόνα κατηγοριοποίησης από τη ρίζα ως το φύλλο. Η διαδικασία κατηγοριοποίησης γίνεται με IF THEN κανόνες. Παράδειγμα κανόνα:

IF ΔΙΑΡΚΕΙΑ (ΕΤΗ) < 1 THEN ΠΑΡΑΜΕΝΕΙ

Με αυτό τον τρόπο παριστάνουμε τις απαραίτητες συνθήκες για να εισαχθεί ένα στοιχείο στον κανόνα στο αριστερό μέρος, ενώ στο δεξί μέρος αντιστοιχεί το φύλλο.

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.

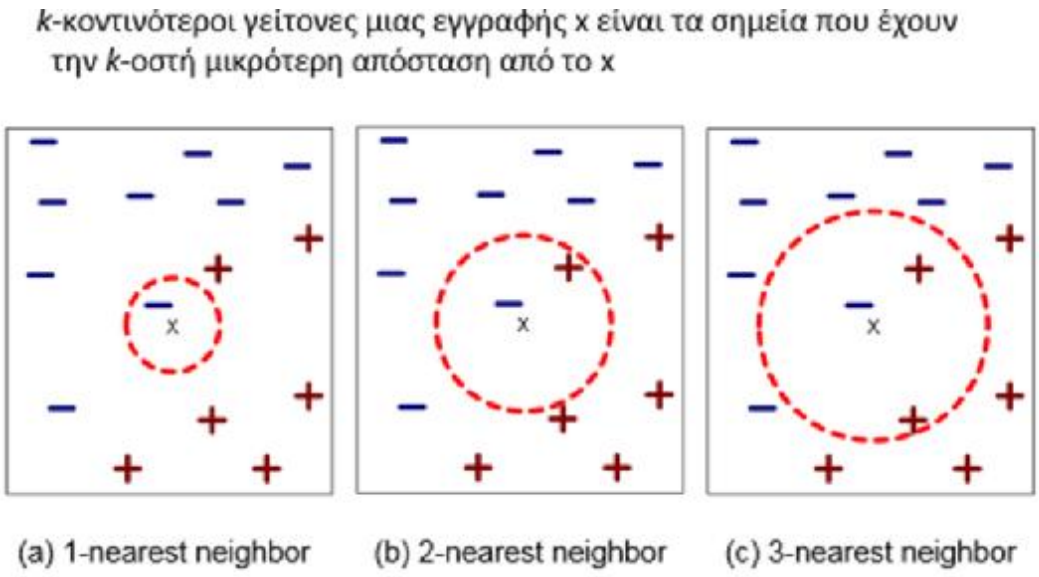
Μέθοδος των k-Κοντινότερων Γειτόνων

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.

Το πόσο κοντά βρίσκεται το στοιχείο σε μια πλειάδα εκπαίδευσης το καθορίζουμε με βάση μια απόσταση όπως η Ευκλείδεια:

1 11 12 1 (, ,...,) n X x x x = και 2 21 22 2 (, ,...,) n X x x x = , είναι 25 ( ) ( )2 1 2 1 2 1 , n i i i dist X X x x = = − ∑

Άρα όταν μια πλειάδα είναι πιο κοντά σε μια ήδη δοσμένη κλάση τότε της ανατίθεται η κλάση αυτή. Για την ταξινόμηση των k-κοντινότερων γειτόνων, στην άγνωστη πλειάδα ανατίθεται η πιο κοινή κλάση μεταξύ των k κοντινότερων γειτόνων της. Όταν k = 1, στην άγνωστη πλειάδα ανατίθεται η κλάση της πλειάδας εκπαίδευσης που βρίσκεται πλησιέστερα σε αυτή στο χώρο προτύπου.



Σχήμα 1.2 Αναπαράσταση της μεθόδου των k-Κοντινότερων Γειτόνων (Εικόνα από τον ιστότοπο www.mines.humanoriented.com)

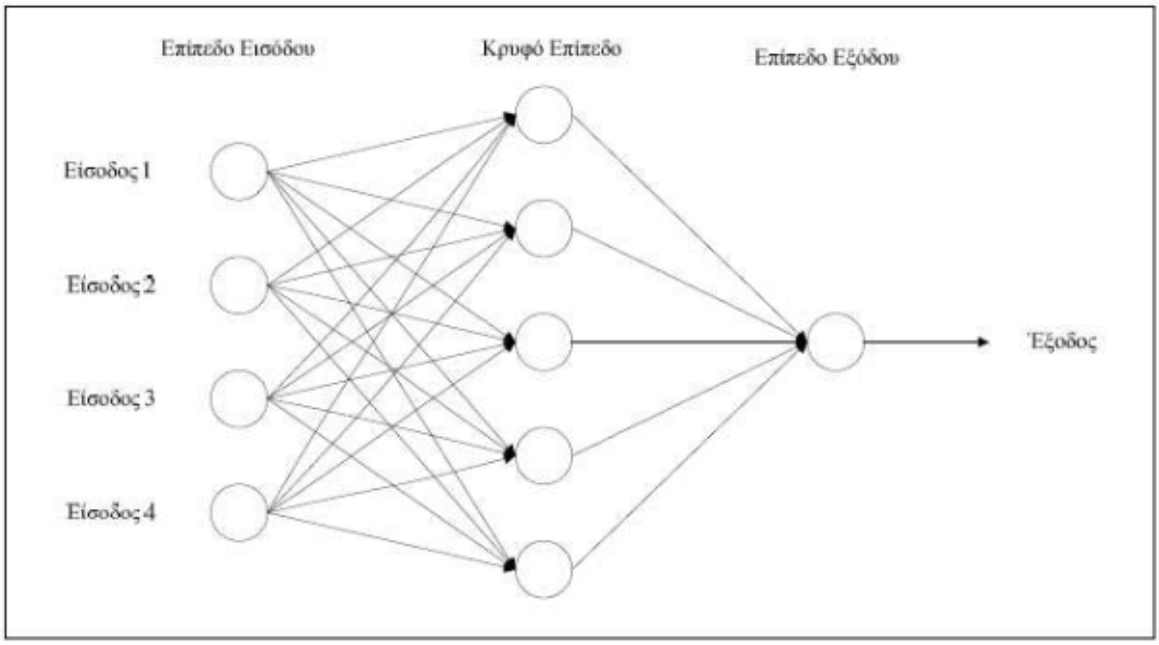
Η μέθοδος αυτή δεν θεωρείται ιδιαίτερα ακριβής στην περίπτωση που υπάρχουν πολλά διαφορετικά χαρακτηριστικά, γιατί κάθε χαρακτηριστικό έχει ίσο βάρος, ωστόσο μπορούν να γίνουν τροποποιήσεις και να υπάρξει διαφοροποίηση στη σημαντικότητα κάποιου χαρακτηριστικού.

Νευρωνικά Δίκτυα

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 (κρυφά ή μη) 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.



Σχήμα 1.3 Αναπαράσταση ενός νευρωνικού δικτύου (Σκούρα, σημειώσεις)

Οι ταξινομητές αυτοί έχουν τη δυνατότητα εύρεσης της κλάσης μιας πλειάδας και της πιθανότητας μια πλειάδα να ανήκει σε μια συγκεκριμένη κλάση. Η λειτουργία τους βασίζεται στο θεώρημα του Bayes.

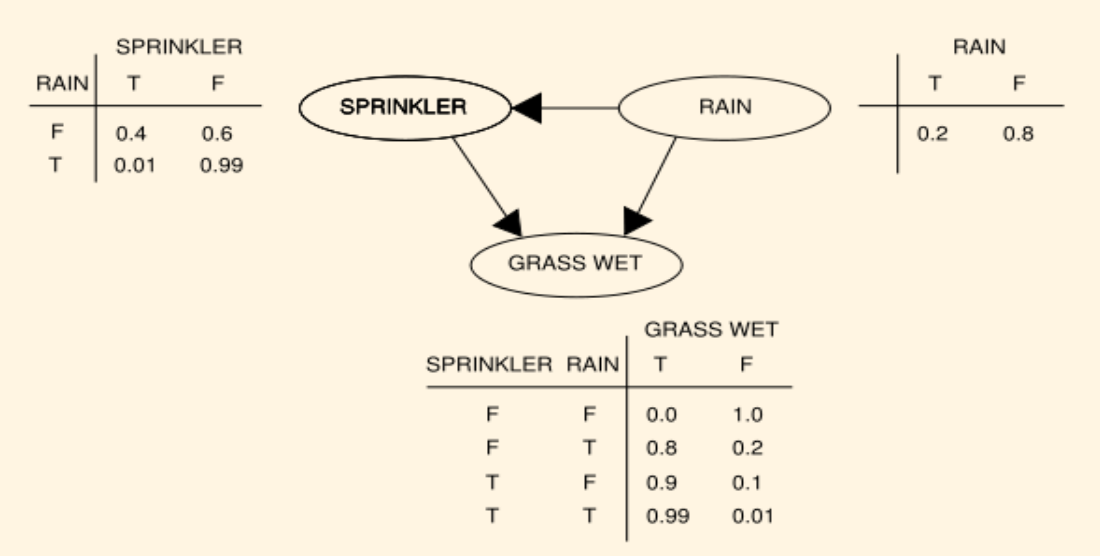
Υπάρχουν οι απλοϊκοί Μπεϋζιανοί ταξινομητές και τα Μπεϋζιανά δίκτυα. Οι απλοϊκοί έχουν ικανοποιητικότατες επιδόσεις και ταχύτατα, σε βαθμό συγκρίσιμο με τα δέντρα απόφασης. Μπορούν επίσης να εφαρμοστούν σε μεγάλες βάσεις δεδομένων.

Κατά τη μέθοδο αυτή θεωρούμε τα χαρακτηριστικά των στοιχείων ως ανεξάρτητα δεδομένα. Όσο πιο ανεξάρτητα είναι τα δεδομένα τόσο πιο ακριβείς είναι οι ταξινομητές αυτοί. Στην πραγματικότητα όμως τις περισσότερες φορές υπάρχει εξάρτηση μεταξύ των δεδομένων.

Οι απλοϊκοί Μπεϋζανοί ταξινομητές χρειάζονται μόνο ένα μικρό αριθμό δεδομένων ώστε να αρχίσει η εκπαίδευση τους για την ταξινόμηση των δεδομένων.

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 1.4 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.

Μηχανές Διανυσμάτων Υποστήριξης

Είναι μια νέα μέθοδος κατά την οποία τα δεδομένα απεικονίζονται γραμμικά σε μια διάσταση, μέσω ενός αλγορίθμου και κατόπιν μετασχηματίζονται σε μεγαλύτερη διάσταση. Τα δεδομένα μπορεί να είναι γραμμικά ή μη. Στη νέα μεγαλύτερη διάσταση χρησιμοποιείται αυτό που καλούμε υπερεπιφάνεια, ώστε να διαχωριστούν τα δεδομένα στις κλάσεις τους. Η υπερεπιφάνεια αυτή υπολογίζεται μέσω διανυσμάτων υποστήριξης.

Τα πλεονεκτήματα τους είναι η υψηλή ακρίβεια, το ότι μπορούν να μοντελοποιήσουν είτε γραμμικά είτε μη δεδομένα. Το κύριο μειονέκτημα τους είναι ο μεγάλος χρόνος εκπαίδευσης. Μπορούν τέλος να χρησιμοποιηθούν είτε για πρόβλεψη, είτε για ταξινόμηση.

Γενετικοί Αλγόριθμοι

Οι γενετικοί αλγόριθμοι είναι εμπνευσμένη 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.

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.

1.2.3 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. Όπως ήδη αναφέραμε 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.

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.[4]

1.3 Μέτρα Αξιολόγησης Κανόνων

Οι κανόνες που παρήχθησαν με τις παραπάνω διεργασίες 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.

[4] Οι κανόνες συσχέτισης χαρακτηρίζονται από το κατώφλι στήριξης (support threshold), που αναγνωρίζει τα στοιχεία των βάσεων δεδομένων που εμφανίζονται συχνά σε αυτά, καθώς και το κατώφλι εμπιστοσύνης (confidence threshold).

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.

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

1.4 Το λογισμικό Weka

Γενικά

Η εμπειρία δείχνει ότι κανένα σχήμα εκπαίδευσης μηχανής δεν είναι σωστό για όλες τα προβλήματα. Το λογισμικό Weka αποτελεί μια συλλογή από τους καλύτερους αλγόριθμους εκπαίδευσης και εργαλεία προ-επεξεργασίας. Έχει σχεδιαστεί ώστε να εφαρμόζει γρήγορα το κάθε αλγόριθμο στο στα σύνολα δεδομένων με ευέλικτους τρόπους. Παρέχει εκτενή υποστήριξη για την συνολική διαδικασία της πειραματικής εξόρυξης δεδομένων, περιέχοντας προεργασία των δεδομένων εισόδου, επαληθεύοντας στατιστικά τα σχήματα εκπαίδευσης και αστικοποιώντας τα σχήματα εκπαίδευσης και τα αποτελέσματα της μάθησης. Το λογισμικό παρέχεται μαζί με ένα απλό γραφικό περιβάλλον που το κάνει εύκολο στη χρήση και ονομάζεται **Explorer**. Το Weka έχει αναπτυχθεί από το Πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Είναι γραμμένο σε Java και διανέμεται σε με άδεια GNU.

Γραφικό περιβάλλον Weka

Ο Explorer δίνει πρόσβαση σε όλες τις λειτουργίες του λογισμικού. Μπορεί εύκολα να διαβαστεί ένα σύνολο δεδομένων από ένα **αρχείο ARFF** και να δημιουργηθεί ένα δέντρο αποφάσεων από αυτό. Το **Knowledge flow** επιτρέπει το σχεδιασμό της διαχείρισης για Streamed data processing. Ο Explorer διατηρεί όλη τη διαδικασία στην μνήμη και όταν ανοίγει ένα σύνολο δεδομένων αυτόματα το φορτώνει και αυτό στη μνήμη. Για το λόγο αυτό μπορεί να χρησιμοποιηθεί μόνο για μικρά σύνολα δεδομένων. Στο Knowledge flow ο χρήστης δημιουργεί μόνος του την σύνθεση της επεξεργασίας και τα δεδομένα φορτώνονται στη μνήμη ανάλογα με τη σύνθεση. Το τρίτο γραφικό περιβάλλον είναι το **Experimenter** και είναι σχεδιασμένο να παράγει μια απάντηση σε μια πρακτική ερώτηση που εφαρμόζεται σε τεχνικές κατηγοριοποίησης και παλινδρόμησης: Ποιοι μέθοδοι και τιμές παραμέτρων είναι καλύτερες για το συγκεκριμένο πρόβλημα. Αυτή η απάντηση μπορεί δοθεί και χρησιμοποιώντας τον Explorer και το Knowledge flow. Αλλά το Experimenter

επιτρέπει την αυτοματοποίηση της διαδικασίας κάνοντας εύκολη την χρήση φίλτρων και κατηγοριών με διαφορετικές τιμές παραμέτρων πάνω στο σύνολο των δεδομένων.

1.5 Χρήσεις της εξόρυξης δεδομένων

α) Η περίπτωση "DiapersandBeers" κατά την οποία παρατηρήθηκε ότι οι πελάτες που αγοράζουν πάνες αγοράζουν και μπύρες. Τότε τοποθετήθηκαν οι μπύρες κοντά με τις πάνες ώστε οι πελάτες να μπορούν εύκολα να αγοράζουν και τα δύο. Ακόμη, στο ενδιάμεσο πρόσθεσαν και τα πατατάκια οπότε αύξησαν και τις πωλήσεις από πατατάκια.

β) Οι τράπεζες κατασκευάζουν δέντρα αποφάσεων με χρήση ιστορικών στοιχείων για να υπολογίσουν το ρίσκο των δανείων και να μπορούν να παίρνουν πιο ασφαλείς αποφάσεις όταν δίνουν δάνεια σε πελάτες.

γ) Λήψη ιατρικών αποφάσεων. Είναι ένας σχετικά νέος τομέας όπου συντελείτε έρευνα. Τα αποτελέσματα των εξετάσεων κατηγοριοποιούνται σε Bayesian δέντρα αποφάσεων στα οποία βγαίνουν τα πιθανά αποτελέσματα.

δ) Στην γενετική στόχος είναι να βρεθούν το πως οι αλλαγές στην αλυσίδα DNA κάποιου ανθρώπου αλλάζουν το ρίσκο ανάπτυξης κάποιων ασθενειών όπως του καρκίνου. Η τεχνική που χρησιμοποιείται για την μελέτη αυτή ονομάζεται multifactor dimensionality reduction.

ε) Στην περιοχή της ηλεκτρολογίας, data mining τεχνικές έχουν χρησιμοποιηθεί ευρέως για επίβλεψη της κατάστασης ηλεκτρολογικού εξοπλισμού υψηλής τάσης. Ο σκοπός της επίβλεψης είναι η λήψη πολύτιμης πληροφορίας που αφορά την κατάσταση της μόνωσης του εξοπλισμού.

στ) Τέλος, το data mining εφαρμόζεται στην εκπαιδευτική έρευνα (educational research), όπου χρησιμοποιείται προς μελέτη των παραγόντων που οδηγούν τους μαθητές/φοιτητές σε δραστηριότητες, οι οποίες μειώνουν την μάθηση.

ζ) Μεγάλο ενδιαφέρον παρουσιάζει μια data mining εφαρμογή, την οποία χρησιμοποιεί ο Εθνικός Σύνδεσμος Καλαθόσφαιρας της Αμερικής (National Basketball Association - NBA) και η οποία χρησιμοποιεί στατιστικά δεδομένα και εικόνες καταγραμμένες από καλαθοσφαιρικούς αγώνες για να αναλύσει τις κινήσεις των παιχτών, βοηθώντας τους προπονητές στην επιλογή κατάλληλων παιχτών και στρατηγικών. Για παράδειγμα, η κατάλληλη ανάλυση στατιστικών δεδομένων κάποιου παιχνιδιού το 1995, έδειξε ότι όταν συγκεκριμένος παίχτης έπαιξε σε αμυντική θέση, τότε ένας άλλος παίχτης επιχείρησε τέσσερις βολές με 100% επιτυχία. Αυτό ήταν σημαντικό, επειδή το μοτίβο αυτό διαφοροποιείται κατά πολύ από τον μέσο όρο επιτυχημένων βολών της ομάδας στο συγκεκριμένο παιχνίδι, ο οποίος ήταν μόλις 49.3%.

1.6 Προεπεξεργασία της εξόρυξης δεδομένων

Προ επεξεργασία δεδομένων

Πριν τα δεδομένα εισαχθούν σε αλγόριθμο πρέπει να γίνουν κάποιες συγκεκριμένες διαδικασίες, όπως να καθαριστούν, να απορριφθούν κάποια να ερευνηθούν και να οριστεί η χρησιμότητα τους. Αν δεν γίνει σωστά αυτή η διαδικασία είναι προφανές ότι όσο ανεπτυγμένο και να είναι το σύστημα πρόβλεψης τα τελικά αποτελέσματα θα είναι λάθος.

Αρχικά θα πρέπει να ελεγχθούν τα δεδομένα με τρόπο όπου οι υπερβολικές τιμές , και οι τιμές που απουσιάζουν να μην δημιουργήσουν πρόβλημα στο σύστημα. Μια τιμή που λείπει είναι προτιμότερο να αντικατασταθεί παρά τα τη σβήσουν εντελώς. Θα πρέπει επίσης να κριθεί η αξιοπιστία και η σημαντικότητα των δεδομένων που

έχουμε. Αν υπάρχουν πολλές λανθασμένες τιμές ή τιμές που λείπουν τότε είναι προτιμότερο να αντικατασταθούν με μια τιμή που δεν θα επηρεάσει σημαντικά την πρόβλεψη, και η τιμή αυτή είναι περισσότερο κοντά στο μέσον όρο των υπολοίπων τιμών.

Για το λόγο αυτό χρησιμοποιούνται συγκεκριμένοι δείκτες που μπορούν να μειώσουν ικανοποιητικά το θόρυβο, δηλαδή την απόκλιση ή έλλειψη τιμών, και απορρίπτουν τις τιμές που θα αλλοιώσουν την πρόβλεψή μας. Οι δείκτες αυτοί μπορούν κατά μία έννοια να κανονικοποιήσουν τα δεδομένα του δείγματος βοηθώντας τον αλγόριθμο να καταλήξει ευκολότερα σε πρόβλεψη αυξάνοντας την ποιότητα των δεδομένων που εισάγονται.

Υπάρχουν και άλλες τεχνικές που χρησιμοποιούνται για την αναγνώριση κάποιων χαρακτηριστικών από το μοντέλο όσον αφορά τα δεδομένα και σε αυτήν την περίπτωση χρησιμοποιούνται τεχνικές όπως:

- Ανάλυση κύριων συνιστωσών (Principal component analysis)
- Ανάλυση ευαισθησίας (Sensitivity analysis)
- Ευρετικές (Heuristic) τεχνικές

Επιλογή μοντέλου και αλγορίθμων

Υπάρχουν πολλά μοντέλα και αλγόριθμοι που μπορούν να χρησιμοποιηθούν, αλλά όπως αναφέραμε σε προηγούμενο υποκεφάλαιο τα σημαντικότερα είναι τα Νευρωνικά δίκτυα και οι γραμμικές μέθοδοι. Επίσης σημαντικά είναι και τα δέντρα αποφάσεων. Τα πάρα πάνω χρησιμοποιούνται συχνά και σε ανάλυση χρηματοοικονομικών μεγεθών, καθώς είναι ευκολότερη η ποσοτικοποίηση των δεδομένων, που συνήθως είναι χρηματικές μονάδες.

Στις περισσότερες περιπτώσεις δεν γίνεται χρήση μόνο μιας τεχνικής αλλά δημιουργούνται υβριδικές τεχνικές για καλύτερη απόδοση. Κάθε τεχνική έχει δυνατά και αδύναμα σημεία, και αν χρησιμοποιηθούν συνδυαστικά μπορούμε να επιτύχουμε καλύτερα αποτελέσματα. Σε άλλες περιπτώσεις πιο εξειδικευμένων αντικειμένων είναι δυνατόν να χρησιμοποιηθεί μόνο ένα μοντέλο που θα έχει τη δυνατότητα να συνδυάζει μεγάλο αριθμό συναρτήσεων. Συνεπώς, η χρήση του κάθε μοντέλου είναι ειδική και εξαρτάται από το αντικείμενο ή την προσέγγιση που θέλουμε να κάνουμε. Τέτοια μοντέλα είναι μη παραμετρικά καθώς δεν χρειάζεται να υπάρχει άμεση σχέση μεταξύ των τιμών των παραμέτρων ενός μοντέλου με δεδομένα. Μπορούμε να συνοψίζουμε τα πλεονεκτήματα των παραμετρικών μοντέλων στα παρακάτω:

- Δίνεται η δυνατότητα να μοντελοποιούνται πολύπλοκες συναρτήσεις ή μεγάλος αριθμός συναρτήσεων
- Χρησιμοποιούν μεγάλο αριθμό μεταβλητών στο μοντέλο, γεγονός που ανεβάζει τις δυνατότητες επεξεργασίας πολυπαραγοντικών ζητημάτων

Ως κύριο μειονέκτημα αυτών των μοντέλων είναι η πολυπλοκότητα τους, και η δυσκολία στην ερμηνεία των αποτελεσμάτων.

Το κάθε μοντέλο πριν θεωρήσουμε ότι είναι κατάλληλο για την επεξεργασία των δεδομένων μας πρέπει να κριθεί και να αξιολογηθούν τα αποτελέσματα του. Κάθε μοντέλο εξετάζει με διαφορετικό τρόπο τους παράγοντες και παρέχει διαφορετική σημαντικότητα σε καθέναν. Συνεπώς, θα πρέπει να αιτιολογείται από το μοντέλο και να παρέχονται πληροφορίες και αποτελέσματα όσο το δυνατόν πιο έγκυρα.

Για να μπορέσουμε όμως να κρίνουμε πιο μοντέλο είναι κατάλληλο για την επεξεργασία των δεδομένων που θέλουμε, και σε πιο βαθμό είναι αξιόπιστες οι πληροφορίες που παρέχει έχουν αναπτυχθεί κάποιες συγκεκριμένες παράμετροί αξιολόγησης. Οι στρατηγικές αυτές έχουν κάποιους βασικούς παράγοντες με τους οποίους κρίνουμε. Αυτοί είναι:

Ακρίβεια (Accuracy). Όπου είναι το ποσοστό σωστών αποτελεσμάτων

Τετραγωνικό σφάλμα (Square error), που είναι χαρακτηριστικό της διακύμανσης των αποτελεσμάτων

Αξιοπιστία (Reliability). Όπου είναι ενδεικτική της αξιοπιστίας του μοντέλου στην πρόβλεψη. Παράδειγμα τέτοιο μπορεί να είναι ο βαθμός επανάληψης μιας πρόβλεψης με παρόμοια δεδομένα

Η προετοιμασία των δεδομένων για όλα τα εργαλεία εξόρυξης δεδομένων

Σε όλα τα μοντέλα και αλγορίθμους είπαμε ότι απαιτείται μια συγκεκριμένη προεργασία, που είναι ίσως ο σημαντικότερος παράγοντας σε πολλές περιπτώσεις. Πολλές φορές η προετοιμασία απαιτεί περισσότερο χρόνο από την ίδια την εξόρυξη δεδομένων. Μπορούμε να κατηγοριοποιήσουμε τα σημαντικότερα βήματα της προετοιμασίας στα παρακάτω:

- Έλεγχος και διόρθωση των δεδομένων. Μπορεί αν υπάρχουν λάθη στα σημεία στίξης, τιμές υπερβολικά μεγάλες λόγω έλλειψης τελειών ή κομμάτων κτλ.
- Αφαίρεση των πεδίων που δεν χρειάζονται στην ανάλυση μας. Τα δεδομένα μπορεί να περιέχουν πληροφορίες που δεν ενδιαφέρουν την ανάλυση μας. Οι τιμές αυτές αν εισαχθούν στον αλγόριθμο θα μας δώσει με βεβαιότητα λανθασμένα αποτελέσματα. Η αφαίρεση αυτών των δεδομένων μειώνει την πολυπλοκότητα του αλγόριθμου και τον αποτρέπει από λάθη όπως η εύρεση μοτίβων κτλ
- Η σωστή αναγραφή των κωδικών και η αντικατάσταση των λέξεων, ώστε να είναι εφικτή η σωστή ταξινόμηση
- Η κατηγοριοποίηση των δεδομένων. Σε πολλές περιπτώσεις χρειάζεται σωστή κατηγοριοποίηση, όπως τα δεδομένα αγοράς προϊόντων από διαφορετικές βάσεις δεδομένων.
- Η κατηγοριοποίηση των δεδομένων με βάση το πεδίο. Σε πολλές περιπτώσεις τα δεδομένα μπορούν να χρησιμοποιηθούν σε παραπάνω από μια χρήση. Τέτοια μπορεί να είναι οικονομικά δεδομένα που να χρησιμοποιηθούν για την εύρεση πολλών διαφορετικών οικονομικών παραμέτρων.

- Ο έλεγχος για μη φυσιολογικά στοιχεία και τιμές. Τέτοια λάθη μπορούν να προέλθουν βέβαια και από παράβλεψη κάποιας τελείας ή κόμματος αλλά σε πολλές περιπτώσεις μπορεί να είναι αι λανθασμένη μέτρηση. Σε αυτή την περίπτωση η χρήση κάποιου indicator προτείνεται. Σε άλλες περιπτώσεις ωστόσο, που υπάρχει μεγάλη διακύμανση δεδομένων καλό θα ήταν να συμπεριλαμβάνονται. Υπάρχουν είδη αντικειμένων που προβλέπουν ομοιογένεια μεταξύ των τιμών και άλλα datasets με τεράστιες αποκλίσεις. Θα πρέπει να μαρκαριστούν οι τιμές αυτές και να ελεγχθούν.
- Έλεγχος για έλλειψη κάποιας τιμής, και να γίνει η απαραίτητη αντικατάσταση.
- Η σωστή μεταχείριση σε μηδενικές τιμές. Πολιό αλγόριθμοι παρουσιάζουν πρόβλημα μεταχείρισης πολλών μηδενικών τιμών, και σε πολλές περιπτώσεις προτείνεται η αντικατάσταση του με την τιμή -1. Αυτό βέβαια ορίζει να μη υπάρχουν πολλές τιμές μικρού μεγέθους, της ίδιας ταράξεις μεγέθους με την τιμή 1. Δηλαδή σε τιμές κοντά στο 1000 θεωρείται αμελητέα η αλλαγή των τιμών από 0 σε -1.
- Η σωστή ταξινόμηση είναι απαραίτητα για την καλύτερη λειτουργία του αλγόριθμου. Για παράδειγμα μπορεί να μας ενδιαφέρει μόνο ένα προϊόν είδους τροφίμων και να θέλουμε να ελέγξουμε την αύξηση ή μείωση κατανάλωση του για να μπορέσει κάποια εταιρία να προμηθευτεί σωστό απόθεμα για συγκεκριμένο χρόνο. Αυτή η διαδικασία απαιτεί εξαιρετικά καλή ταξινόμηση.

Προετοιμασία για χρήση από συγκεκριμένο εργαλείο εξόρυξης δεδομένων

Όλα εργαλεία εξόρυξης δεδομένων απαιτούν συγκεκριμένη προεργασία πριν τα δεδομένα εισαχθούν στον αλγόριθμο. Χρειάζονται δηλαδή συγκεκριμένες μετατροπές, που υπερβαίνουν τις προηγούμενες διαδικασίες. Τέτοιες μετατροπές μπορεί να είναι:

- Μπορεί να χρειαστεί κατηγοριοποίηση των δεδομένων για τρεις διακριτές διαδικασίες. Η πρώτη ομάδα δεδομένων θα χρησιμοποιηθεί για να ρυθμιστούν οι παράμετροι, και η σημαντικότητες με βάση τις οποίες θα γίνουν οιο προβλέψεις από το εργαλείο. Η δεύτερη ομάδα δεδομένων εκλεχθεί τη χρήση των παραμέτρων ώστε να εξακριβωθεί η σωστή λειτουργία του μοντέλου. Η Τρίτη ομάδα δεδομένων χρησιμοποιείται για να την αξιολόγηση των αποτελεσμάτων του μοντέλου. Αυτά τα εργαλεία μπορεί αν είναι κατά σειρά ένα εργαλείο συσταδοποίησης (clustering tool), ένα εργαλείο νευρωνικών δικτύων (neural network tool) ή ένα εργαλείο δέντρου αποφάσεων (decision tree tool).
- Υπάρχει σε πολλές περιπτώσεις ανάγκη για ρύθμιση υπό συγκεκριμένες παραμέτρους που λειτουργούν ως στόχοι. Όταν αναφερόμαστε σε επιχειρήσεις ο στόχος συνήθως είναι το κέρδος, ή η κατανάλωση από συγκεκριμένες ομάδες πελατών. Για να γίνει η πρόβλεψη με βάση τη συγκεκριμένη ομάδα θα πρέπει να έχει προηγηθεί προεργασία όπου παράμετροι θα θεωρούνται τα κέρδη ή η ικανοποίηση των πελατών.
- Η κατηγοριοποίηση των δεδομένων σε κλίμακες. Πολλά εργαλεία εξόρυξης απαιτούν αυτή την προεργασία, όπως τα δέντρα αποφάσεων.
- Στα Νευρωνικά δίκτυα χρειάζεται να γίνει μια προεργασία ώστε ν γίνει εξομάλυνση των τιμών μεταξύ του μηδέν και του ένα. Στο συγκεκριμένο εργαλείο όλες οι τιμές είναι μεταξύ του μηδέν και του ένα.
- Κάποια εργαλεία απαιτούν την είσοδο μόνο αριθμητικών τιμών, και συνεπώς όλα τα δεδομένα θα πρέπει να μετατραπούν σε αριθμητικές τιμές.

1.7 Το μέλλον του data mining και τα big data

Τα λεγόμενα μεγάλα δεδομένα φαίνεται να είναι το μέλλον της εξόρυξης δεδομένων. Με την έννοια μεγάλα δεδομένα ή big data εννοούμε τα υψηλού όγκου, υψηλής ταχύτητας ή υψηλής ποικιλίας στοιχεία που απαιτούν αποδοτικές και καινοτόμες μορφές επεξεργασίας πληροφοριών. Το 2012 ξοδεύτηκαν σε επενδύσεις σε εξόρυξη δεδομένων σε μεγάλα δεδομένα ενενήντα έξι δισεκατομμύριά δολάρια. Αυτό βέβαια

δεν πρέπει να μας φαίνεται τόσο υπερβολικό καθώς στα μεγάλα δεδομένα συγκαταλέγονται και τα δεδομένα από τα μέσα κοινωνικής δικτύωσης, που είναι εξαιρετικά χρήσιμα για επιχειρήσεις, αλλά ανανεώνονται με υπερβολικά μεγάλη ταχύτητα.

Στα μεγάλα δεδομένα μπορούμε να συμπεριλάβουμε και βίντεο ή φωτογραφίες που δημοσιεύονται στα κοινωνικά δίκτυα, γεγονός που μας δείχνει πόσο υψηλής επεξεργαστικής ισχύος θα ρέπει να είναι το σύστημα επεξεργασίας τους.

Όλες οι μεγάλες εταιρίες τείνουν να δείχνουν μεγάλο ενδιαφέρον για τα μεγάλα δεδομένα, και ιδιαίτερα οι κολοσσοί του διαδικτύου. Φυσικά η πρώτη που ενδιαφέρθηκε να τα αξιοποιήσει είναι η Google η οποία είχε και πρόσβαση. Πολλές άλλες μεγάλες επιχειρήσεις έχουν κάνει σημαντικές επενδύσεις όπως η εταιρία τηλεφωνίας IOVOX.

Ωστόσο, πολλοί θεωρούν ότι η αξιοποίηση του τεράστιου όγκου δεδομένων που υπάρχει στο διαδίκτυο με τρόπο ικανοποιητικό αργεί πολύ. Προσπάθειες βέβαιά έχουν γίνει ήδη και αξιοποιούνται εμπορικά, όπως η άντληση πληροφοριών από τα μέσα κοινωνικής δικτύωσης για τις προτιμήσεις των πελατών ή τις μελλοντικές αγοραστικές τάσεις.

Είναι προφανές ότι όλες οι εταιρίες που επιθυμούν να γνωρίζουν τις επιθυμίες των καταναλωτών μια συγκεκριμένη χρονική περίοδο θα ήθελαν να έχουν συστήματα επεξεργασίας μεγάλων δεδομένων. Αυτό που συμβαίνει όμως είναι ότι οι περισσότερες επιχειρήσεις που προσπαθούν να εξάγουν αποτελέσματα από τα big data καταλήγουν σε αποτυχία. Πιο συγκεκριμένα, δυο στις τρεις επιχειρήσεις επιτυγχάνουν κατά την Hewlett Packard.

ΚΕΦΑΛΑΙΟ 2: ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

Πριν αναπτύξουμε τις βασικές ιδέες του κεφαλαίου αυτού αξίζει να αναφερθούμε σε 2-3 βασικά στοιχεία που χρησιμοποιεί.

Στοιχειοσύνολα:

Με δεδομένο ένα σύνολο από στοιχεία $I=\{I_1,I_2,\ldots,I_m\}$ και μία βάση δεδομένων από συναλλαγές $D=\{t_1,t_2,\ldots,t_n\}$ όπου $t_i=\{I_{i1},I_{i2},\ldots,I_{ik}\}$ και $I_{ij}\in I$, ένας **κανόνας συσχέτισης** είναι ένα επαγωγικό συμπέρασμα της μορφής $X\Longrightarrow Y$, όπου $X,Y\subset I$ είναι σύνολα στοιχείων που ονομάζονται στοιχειοσύνολα και $X\cup Y=\emptyset$

Συχνά Στοιχειοσύνολα:

Ένα συχνό στοιχειοσύνολο είναι ένα στοιχειοσύνολο του οποίου ο αριθμός των εμφανίσεων είναι πάνω από ένα κατώφλι, s. Χρησιμοποιούμε το Συμβολισμό L για να δηλώσουμε το σύνολο που αποτελείται από όλα τα συχνά στοιχειοσύνολα και το l για να δηλώσουμε ένα συγκεκριμένο συχνό στοιχειοσύνολο.

Υποστήριξη:

Για έναν κανόνα συσχέτισης $X\Longrightarrow Y$ υποστήριξη (support - s) είναι το ποσοστό των συναλλαγών στη βάση δεδομένων που περιέχουν το $X\cup Y$

Εμπιστοσύνη:

Για έναν κανόνα συσχέτισης $X\Longrightarrow Y$ εμπιστοσύνη (confidence - c) είναι το κλάσμα του αριθμού των συναλλαγών που περιέχουν το $X\cup Y$ προς τον αριθμό των συναλλαγών που περιέχουν το X.

Αλγόριθμος Apriori

Ο αλγόριθμος Apriori είναι ένας από τους πιο γνωστούς αλγορίθμους για την εύρεση κανόνων συσχέτισης και ο οποίος χρησιμοποιεί την ιδιότητα συχνών

στοιχειοσυνόλων, δηλαδή οποιοδήποτε υποσύνολο ενός συχνού στοιχειοσυνόλου πρέπει να είναι συχνό. Τα συχνά στοιχειοσύνολα ονομάζονται κλειστά προς κάτω επειδή εάν ένα στοιχειοσύνολο ικανοποιεί τις απαιτήσεις της ελάχιστης υποστήριξης, το ίδιο συμβαίνει και για όλα τα υποσύνολά του. Αντίστροφα τώρα εάν ένα στοιχειοσύνολο δεν είναι συχνό, δε δημιουργούμε κανένα υπερσύνολο του, ως υποψήφιο, επειδή και αυτό αποκλείεται να είναι συχνό.

Συνεπώς η βασική ιδέα του Apriori είναι η δημιουργία υποψήφιων στοιχειοσυνόλων ενός συγκεκριμένου μεγέθους και στη συνέχεια η σάρωση της βάσης δεδομένων για να μετρήσουμε και να δούμε αν αυτά είναι συχνά.

Κατά τη διάρκεια του περάσματος, καταμετρούνται τα υποψήφια στοιχειοσύνολα μεγέθους i, C_i ενώ μόνο εκείνοι οι υποψήφιοι που είναι συχνοί χρησιμοποιούνται για τη δημιουργία υποψηφίων για το επόμενο πέρασμα. Αυτό σημαίνει ότι το L_i χρησιμοποιείται για τη δημιουργία του C_{i+1}. Ένα στοιχειοσύνολο θεωρείται ως υποψήφιο μόνο όταν όλα του τα υποσύνολα είναι επίσης συχνά, Για τη δημιουργία υποψηφίων μεγέθους i+1 γίνονται συνενώσεις συχνών στοιχειοσυνόλων που βρίσκονται σε προηγούμενο πέρασμα.

Στο WEKA

Ο αλγόριθμος Apriori χρησιμοποιεί την ίδια καθορισμένη ελάχιστη τιμή για το confidence, η οποία δίνεται από τη minMetric παράμετρο. Το επίπεδο υποστήριξης εκφράζεται ως ποσοστό (μεταξύ 0 και 1) του συνολικού αριθμού των περιπτώσεων και ξεκινά με μια συγκεκριμένη τιμή (upper Bound Min Support).
Σε κάθε επανάληψη η υποστήριξη μειώνεται κατά μια σταθερή ποσότητα (δέλτα). Αυτή η διαδικασία γίνεται μέχρι να δημιουργηθεί ένας συγκεκριμένος αριθμός κανόνων(numRules) ή η υποστήριξη να φθάσει σε ένα ορισμένο ελάχιστο επίπεδο(lowerBoundMinSupport). Το number of cycles performed δείχνει ότι ο αλγόριθμος έτρεξε συγκεκριμένο αριθμό για να δημιουργήσει αυτούς του κανόνες.

2.1. Market Basket Analysis

Η τεχνική Market Basket Analysis βασίζεται στην εξέταση της πιθανότητας συσχέτιση μεταξύ ενός προϊόντος με κάποιο άλλο προϊόν ή σύνολο προϊόντων. Πιο αναλυτικά, αν ένας πελάτης αγοράσεις ένα προϊόν ή σύνολο προϊόντων Α, καλούμαστε να εξετάσουμε το πόσο πιθανόν είναι να αγοράσει ένα προϊόν ή σύνολο προϊόντων Β.[5]

Το προϊόν ή σύνολο προϊόντων που αγοράζει ένας πελάτης το ονομάζουμε itemset. Το κάθε itemset μπορεί να είναι ένα ή πολλά προϊόντα, που θα εξετάσουμε τη συσχέτιση μεταξύ τους. Όταν δύο προϊόντα Α και Β υπάγονται σε κάποια συσχέτιση τότε μπορεί η σχέση τους να παρασταθεί με την παρακάτω μορφή:

IF { προϊόν Α } THEN { προϊόν Β }

Η παραπάνω συσχέτιση μπορεί να επεκταθεί και σε παραπάνω από ένα προϊόν τη φορά, και να παριστάνει ένα σύνολο προϊόντων. Τότε θα μπορούσε ως παράδειγμα να παρασταθεί ως εξής:

IF { *γάλα, ψωμί* } THEN { *βούτυρο, δημητριακά* }

Τα παραπάνω δεδομένα μπορούν πλέον να εισαχθούν σε στατιστικές συναρτήσεις και να εξαχθούν χρήσιμα δεδομένα. Οι μεταβλητές αυτής της περίπτωσης ονομάζονται support και confidence.

[5] Τα εργαλεία και οι τεχνικές της εξόρυξης δεδομένων βοηθούν στο να αναλύσουμε τα οικονομικά δεδομένα και είναι τέτοια η συμβολή τους έτσι ώστε για παράδειγμα, τα οικονομικά ινστιτούτα να αναγνωρίζουν τις απάτες από παραποιημένα δεδομένα από τις διάφορες βάσεις δεδομένων και από το ιστορικό συναλλαγών που έγιναν από τους πελάτες.

Πιο αναλυτικά και πριν προχωρήσουμε θα παραθέσουμε τον σύντομο ορισμό της τεχνικής Market Basket Analysis. Στα ελληνικά θα μπορούσαμε να την ονομάσουμε ανάλυση καλαθιού αγορών, και είναι αυτό που περιγράφει το ίδιο το όνομα της. Τα αντικείμενα που εξετάζουμε κάθε φορά επιλέγονται από ένα σύνολο αντικειμένων και εξετάζεται η συσχέτιση τους. Τα αντικείμενα αυτά με τη σειρά τους επιλέχθηκαν από ένα μεγαλύτερο σύνολο αντικειμένων, που είναι τα διαθέσιμα αγαθά μιας επιχείρησης. Οι τρόποι συμβολισμού μαρτυρούν και την συσχέτιση που εξετάζουμε μεταξύ των αντικειμένων, και αν θέλουμε για παράδειγμα να εξετάσουμε ότι το αντικείμενο Α υπάρχει στο καλάθι αγορών αν υπάρχει και το αντικείμενο Β, τότε μπορούμε να φτιάξουμε τον κανόνα Α→Β, που υποδηλώνει την παραπάνω σχέση.

Όπου το αντικείμενο Α το ονομάζουμε προηγηθέν και το αντικείμενο Β το ονομάζουμε συνεπακόλουθο. Με αγγλική ορολογία το Α ονομάζεται antecedent item και το Β consequent. Μετά τη δημιουργία κανόνων γίνεται η εξέταση του βαθμού αλήθειας που περιέχουν. Μπορεί ένα κανόνας να είναι αληθής χωρίς όμως να είναι και τα υποσύνολά του αληθή απαραίτητα. Αυτή η σχέση μπορεί να παρασταθεί με τα A AND NOT B αν για παράδειγμα ισχύει ο κανόνας Α και όχι ο Β. Αναλυτικότερα για τις έννοιες support και confidence:

Ως support ενός κανόνα ονομάζουμε το ποσοστό αλήθειας που εμπεριέχει. Για παράδειγμα αν ο κανόνας Α→Β ισχύει κατά Χ%, τότε το Χ είναι το support του συγκεκριμένου κανόνα. Ως confidence ορίζουμε το ποσοστό συσχέτισης μεταξύ δύο προϊόντων ή υποσυνόλων. Δηλαδή αν τουλάχιστον Χ% είναι το ποσοστό εμφάνισης του προϊόντος Β σε καλάθια που περιέχουν το προϊόν Α, το confidence του κανόνα αυτού το ονομάζουμε Χ. Στη συγκεκριμένη τεχνική το confidence είναι μάλλον ο σημαντικότερος δείκτης, καθώς μας πληροφορεί σε πιο βαθμό μπορεί μια υπόθεση είναι αληθής ώστε να έχει αξία για τη λήψη κάποιας απόφασης. Δεν είναι ανάγκη να είναι 100% αληθής, αλλά να εμφανίζεται κάποια συσχέτιση σε κάποιο αξιόλογο ποσοστό. Ως παράδειγμα, αν εμφανιστεί κάποια συσχέτιση μεταξύ δυο προϊόντων ή δυο ειδών προϊόντων η επιχείρηση μπορεί να αλλάξει τη διαμόρφωση των χώρων πωλήσεις ώστε να ωθήσει τον καταναλωτή στη αγορά και των δύο προϊόντων.

Η συγκεκριμένη τεχνική, του να κρίνουμε δηλαδή πιθανή συσχέτιση μέσω του confidence, μπορεί να οδήγησε και σε λάθος συμπεράσματα. ένα προϊόν μπορεί να είναι εξαιρετικά δημοφιλές ούτως ή άλλως ανεξάρτητα από τα άλλα προϊόντα που βρίσκονται στη λίστα αγορών. Μπορεί ακόμα δύο προϊόντα να εμφανίζουν συνεχώς παρουσία ταυτόχρονα και να μην υπάρχει σχέση αιτίας που βρίσκονται ταυτόχρονα στο ίδιο καλάθι.

Για παράδειγμα έστω ότι υπάρχει confidence 80% για ένα προϊόν Β σε σχέση με το Α, παρόμοιου κανόνα Α→Β. Αυτό σημαίνει ότι κάθε φορά που εμφανίζεται το προϊόν Α στο καλάθι ακολουθεί το ποιόν Β στο 80% των περιπτώσεων.

Μπορεί όμως αν εξετάζουμε τη συσχέτιση με ένα όχι τόσο δημοφιλές συνεπακόλουθο προϊόν Β να μην εμφανίζεται ποτέ, ή σπάνια με κάποια εξάρτιση από το Α. Από την άλλη μπορεί να εμφανίζεται πάντα ως τρομερά δημοφιλές προϊόν χωρίς συσχέτιση.

Το support από την άλλη είναι ένας χρήσιμος δείκτης που παρέχει πληροφορίες για τη σημαντικότητα κάποιου κανόνα, και της προσοχής που του αρμόζει. Πρακτικώς μα δείχνει το ποσοστό αληθών περιπτώσεων ενός κανόνα.

2.2 Δημιουργία Κανόνων Συσχέτισης από Frequent Itemsets

Τα itemsets μπορούμε να τα προμηθευτούμε από πίνακες συναλλαγών ή μέσω βάσης δεδομένων. Από τα itemsets μπορούμε να παράγουμε κανόνες συσχέτισης και μετά να εξετάσουμε κατά πόσο ισχύουν. Για να θεωρηθεί ένας κανόνας επαρκής και να αξίζει προσοχής θα πρέπει να ικανοποιεί μια συγκεκριμένη τιμή support και μια συγκεκριμένη τιμή confidence. Οι τιμές αυτές ονομάζονται minimum support και minimum confidence, και συνολικά οι κανόνες ονομάζονται strong association rules.

Όπου οι κανόνες συσχέτισης γίνονται με χρήση της παρακάτω εξίσωσης:

confidence (A→B) = P (B|A) = support count (A ∪ B) / support count(A)

Πριν συνεχίσουμε όμως πρέπει να υπενθυμίσουμε τους παρακάτω σημαντικούς κανόνες:

Ως confidence ονομάζουμε τον ποσοτικό δείκτη που προσδιορίζει σε ένα itemset σε ποιο ποσοστό εμφανίζεται το συνεπακόλουθο Β σε σχέση με το προηγηθέν προϊόν Α. Ως support ορίζουμε το δείκτη που προσδιορίζει σε ποιο βαθμό ισχύει ένας υποθετικός κανόνας συσχέτισης ενός προϊόντος ή σύνολο προϊόντων Α με ένα συνεπακόλουθο προϊόν Β.

Ο κανόνας της μορφής $A \rightarrow B$ μας δείχνει την πιθανότητα να εμφανιστεί στο καλάθι αγορών το προϊόν Β, εφόσον ήδη έχει αγοραστεί το προϊόν Α. Η παραπάνω εξίσωση δεν είναι τίποτα παραπάνω από μια εξίσωση δεσμευμένης πιθανότητας μεταξύ δυο ενδεχομένων Α και Β. Η ένωση $A \cup B$ είναι το ενδεχόμενο να υπάρχουν στο καλάθι αγορών και το ενδεχόμενο Α και το ενδεχόμενο Β και ονομάζεται support count.

Για κάθε frequent itemset l, δημιούργησε όλα τα μη κενά υποσύνολα του l.

Για κάθε μη κενό υποσύνολο s του l, παρήγαγε τον κανόνα:

$s \rightarrow (l - s)$ (2)

εάν support_count(l)/support_count(s) ≥ min_conf

όπου min_conf είναι το κατώφλι minimum confidence. Εφόσον οι κανόνες έχουν δημιουργηθεί από frequent itemsets, τότε εξορισμού ικανοποιούν το minimum support κατώφλι. Με τον πιο πάνω περιορισμό, ικανοποιούν και το minimum confidence, οπότε έχουμε την παραγωγή strong association rules, οι οποίοι ικανοποιούν και το minimum support και το minimum confidence.

Παράδειγμα εφαρμογής του αλγορίθμου

Ως παράδειγμα θα χρησιμοποιήσουμε θεωρητική εταιρία και ως itemsets θα θεωρήσουμε τα προϊόντα L1, L2 και L3. Το παράδειγμα έχει αντληθεί από την μεταπτυχιακή εργασία του Μεπούρη (2008):

Θα παράξουμε κανόνες συσχέτισης με βάση τα frequent itemsets αυτά. Θα χρησιμοποιήσουμε το frequent itemset l = {I1, I2, I5}. Τα μη κενά υποσύνολα του l είναι:

{I1, I2}, {I2, I5}, {I1,I5}, {I1}, {I2} και {I5}

Οι κανόνες συσχέτισης που προκύπτουν σύμφωνα με την εξίσωση 2, είναι:

Για s={I1, I2} έχουμε: {I1, I2} → I5

confidence = support_count(l) / support_count(s) = 2 / 4 = 50%

Από τον πίνακα L3 προκύπτει ότι support_count(l) = support_count(I1, I2, I5) = 2 και από τον πίνακα L2 προκύπτει ότι support_count(s) = support_count(I1, I2) = 4. Η ίδια διαδικασία θα γίνει και για τα υπόλοιπα μη κενά υποσύνολα του l:

Για s={I1, I5} έχουμε: {I1, I5} → I2

confidence = support_count(l) / support_count(s) = 2 / 2 = 100%

Για s={I2, I5} έχουμε: {I2, I5} → I1

confidence = support_count(l) / support_count(s) = 2 / 2 = 100%

Για s=I1 έχουμε: I1 → {I2, I5}

confidence = support_count(l) / support_count(s) = 2 / 6 = 33%

Για s=I2 έχουμε: I2 → { I1, I5}

confidence = support_count(l) / support_count(s) = 2 / 7 = 29%

Για s=I5 έχουμε: I5 → { I1, I2}

confidence = support_count(l) / support_count(s) = 2 / 2 = 100%

Αν για παράδειγμα το κατώφλι minimum confidence είναι 70%, τότε μόνο ο δεύτερος, τρίτος και τελευταίος κανόνας το ικανοποιούν, αφού έχουν confidence support πάνω από το κατώφλι, άρα μόνο αυτοί χαρακτηρίζονται ως strong association rules. Έτσι, το αποτέλεσμα της εφαρμογής του Apriori αλγορίθμου στο παράδειγμά μας είναι οι εξής κανόνες συσχέτισης:

{I1, I5} → I2

Δεδομένης της αγοράς των προϊόντων I1 και I5, υπάρχει πιθανότητα 1 (confidence=100%) να έχει αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για το προϊόν I2.

{I2, I5} → I1

Δεδομένης της αγοράς των προϊόντων I2 και I5, υπάρχει πιθανότητα 1 (confidence=100%) να έχει αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για το προϊόν I1.

I5 → { I1, I2}

Δεδομένης της αγοράς του προϊόντος I5, υπάρχει πιθανότητα 1 (confidence=100%) να έχουν αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για τα προϊόντα I1 και I2.

Στα επόμενα κεφάλαια, παρουσιάζονται τα συστήματα από τα οποία αποτελείται η εφαρμογή. Πιο συγκεκριμένα, στο κεφάλαιο 6 αναλύεται το Σύστημα Εξαγωγής Κανόνων Συσχέτισης, ενώ στο κεφάλαιο 7 το Σύστημα Ανίχνευσης και Επεξεργασίας. Το κεφάλαιο 8 περιγράφει τις δοκιμές της εφαρμογής και τέλος στο κεφάλαιο 9 συζητάμε τα αποτελέσματα, καθώς και τη μελλοντική εργασία.

2.3 Apriori αλγόριθμος

Ο αλγόριθμος Apriori έχει προταθεί από τους R. Agrawal R. Srikant το 1994 [9]. Ο αλγόριθμος χρησιμοποιείται για ανόρυξη συχνών συνόλων αντικειμένων (itemsets) για εξόρυξη κανόνων συσχέτισης. Ο αλγόριθμος έχει πάρει το όνομα του από την προγενέστερη γνώση (prior knowledge) των χαρακτηριστικών των συχνών συνόλων αντικειμένων, που χρησιμοποιεί. Ο Apriori υιοθετεί την τεχνική αναζήτηση, level-wise, η οποία είναι μια επαναλαμβανόμενη τεχνική που χρησιμοποιεί τα k-itemsets για να κτίσει τα (k+1)-itemsets. Στην αρχή, ο αλγόριθμος βρίσκει τα συχνά εμφανιζόμενα 1-itemsets (το σύνολο αντικειμένων με 1 μόνο χαρακτηριστικό). Ο αλγόριθμος αναζητά και συναθροίζει τον αριθμό που εμφανίζεται κάθε αντικείμενο – χαρακτηριστικό στη βάση δεδομένων, και μετά συλλέγει τα αντικείμενα που ικανοποιούν το ελάχιστο support, στο σύνολο L1. Κατόπιν, χρησιμοποιώντας το σύνολο L1, χτίζεται το σύνολο L2 το οποίο περιλαμβάνει όλα τα συχνά σύνολα αντικειμένων με 2 χαρακτηριστικά (2-itemsets), το οποίο κι αυτό χρησιμοποιείται για να χτιστεί το L3, και ούτω κάθε εξής, μέχρι που να μην μπορεί βρεθεί άλλο σύνολο με kitemesets, δηλαδή το Lk να είναι κενό. Για να βρεθεί κάθε Lk απαιτείται μία αναζήτηση της βάσης δεδομένων. Για την δημιουργία κάθε επιπέδου με τα συχνά σύνολα αντικειμένων, χρησιμοποιείται η ιδιότητα Apriori (Apriori Property) η οποία μειώνει τον χώρο αναζήτησης και έτσι βελτιώνεται σημαντικά η αποδοτικότητα του αλγορίθμου. Η ιδιότητα Apriori αναφέρει ότι: 15 όλα τα μη κενά υποσύνολα των συχνών συνόλων αντικειμένων πρέπει να είναι επίσης συχνά. Η ιδιότητα Apriori βασίζεται στο ότι: εάν ένα σύνολο αντικειμένων I δεν ικανοποιεί το ελάχιστο όριο support (min_sup), τότε το I δεν είναι συχνό, P(I) < min_sup. Εάν το αντικείμενο A προστίθεται στο σύνολο αντικείμενων I, τότε το καινούργιο σύνολο I (IUA) δεν μπορεί να εμφανίζεται πιο συχνά από το I. Επομένως, ούτε το σύνολο IUA είναι συχνό, επειδή P(IUA) < min_sup. Η ιδιότητα Apriori χρησιμοποιείται για την παραγωγή του Lk από το Lk-1, για k≥2, και ακολουθείται μια διαδικασία δύο βημάτων, που αποτελείται από την διαδικασία ένωσης (join) και κλαδέματος (prune):

Διαδικασία Ένωσης: Για να βρεθεί το σύνολο Lk , παράγετε ένα σύνολο από υποψήφια σύνολα με k αντικείμενα (k-itemsets) από την ένωση του συνόλου Lk-1 με τον εαυτό του. Το σύνολο με τα υποψήφια σύνολα αντικειμένων καλείται Ck. Εάν το li είναι μέλος του Lk-1, τότε το li[j] αναφέρεται στο αντικείμενο j του συνόλου αντικειμένων li. Ο Apriori θεωρεί ότι τα αντικείμενα στα σύνολα είναι ταξινομημένα σε αλφαβητική σειρά. Για κάποιο σύνολο αντικειμένων li με (k-1) αντικείμενα, τα αντικείμενα είναι ταξινομημένα σε li[1]< li[2] < li[3]<...< li[k-1]. Όταν η ένωση Lk-1 Lk-1 εκτελείται, τα μέλη του Lk-1 μπορούν να ενωθούν εάν τα πρώτα (k-2) αντικείμενα είναι τα ίδια. Για παράδειγμα το l1 και l2 itemsets που ανήκουν στο σύνολο Lk-1 μπορούν να ενωθούν εάν (l1[1] = l2[1]) (l1[2] = l2[2]) … (l1[k-2] = l2[k-2]) (l1[k-1] < l2[k-1]). Ο έλεγχος (l1[k-1] < l2[k-2]) γίνεται για να εξασφαλιστεί ότι δεν θα παραχθεί κανένα αντίγραφο του ίδιου itemset στο Ck. Το αποτέλεσμα της ένωσης των l1 και l2 itemsets είναι l1[1], l1[2], l1[3],..., l1[k-1], l2[k-1]. Διαδικασία Κλαδέματος (prune): Κάποια από τα σύνολα αντικειμένων που ανήκουν στο Ck, μπορεί να είναι συχνά εμφανιζόμενα κι άλλα όχι, όμως όλα τα συχνά εμφανιζόμενα σύνολα k αντικειμένων (k-itemsets) συμπεριλαμβάνονται στο Ck. Θα πρέπει να γίνει μία αναζήτηση στη βάση δεδομένων για να μετρηθεί ο αριθμός όπου κάθε υποψήφιο σύνολο στο Ck, εμφανίζεται στη βάση δεδομένων. Όλα τα σύνολα αντικειμένων που περιλαμβάνονται στο Ck, και εμφανίζονται στη βάση δεδομένων όχι λιγότερό αριθμό από το ελάχιστο support, 16 τότε αυτό το σύνολο αντικειμένων προστίθεται στο Lk. Αυτό γίνεται, όπως αναφέρει η Apriori ιδιότητα, οποιοδήποτε (k-1)-itemset σύνολο αντικειμένων δεν είναι συχνό τότε δεν μπορεί να είναι υποσύνολο κάποιου k-itemset σύνολο αντικειμένων. Έτσι επειδή το σύνολο Ck, μπορεί να γίνει αρκετά μεγάλο, τα σύνολα αντικειμένων που δεν είναι συχνά αφαιρούνται.

Περιγραφή Ψευδοκώδικα Αλγόριθμου Apriori

Στην Εικόνα 3 παρουσιάζεται ο ψευδοκώδικας του αλγόριθμου Apriori και οι σχετικές διαδικασίες:

1. Καταρχάς ο αλγόριθμος δέχεται μια βάση δεδομένων με δοσοληψίες. Η βάση δεδομένων δοσοληψιών αποτελείται από ένα αρχείο, και κάθε εγγραφή του αρχείου

αντιπροσωπεύει μία δοσοληψία. Η δοσοληψία συνήθως περιλαμβάνει ένα μοναδικό αριθμό ταυτότητας και μία λίστα από αντικείμενα (items) – χαρακτηριστικά όπου συνθέτουν την δοσοληψία.

2. Στο πρώτο βήμα βρίσκονται όλα τα συχνά σύνολα αντικειμένων με 1 χαρακτηριστικό και φυλάγονται στο σύνολο L1

3. Στο βήμα 3 είναι η διαδικασία όπου το σύνολο υποψηφίων Ck παράγεται από την ένωση του Lk-1 με τον εαυτό του. Η διαδικασία apriori_gen παράγει τα υποψήφια σύνολα αντικειμένων και μετά χρησιμοποιεί την ιδιότητα Apriori για να αφαιρέσει τα αυτά που δεν είναι συχνά.

4. Στο βήμα 4 – 10 γίνεται μία αναζήτηση στη βάση δεδομένων, για να βρεθεί ο αριθμός που τα σύνολα αντικειμένων εμφανίζονται στην βάση. Και στο βήμα 9 βρίσκει τα υποψήφια σύνολα αντικειμένων που έχουν μεγαλύτερο από το ελάχιστο support και τα προσθέτει στο σύνολο Lk. 17

5. Στο τελικό βήμα 11 γίνεται μια ένωση όλων των συχνών συνόλων αντικειμένων των συνόλων Lk στο L. Έτσι μετά μια διαδικασία για εξαγωγή κανόνων συσχέτισης μπορεί να χρησιμοποιήσει το σύνολο L.

Αλγόριθμος: Apriori. Εύρεση των συχνών συνόλων αντικειμένων (itemsets) χρησιμοποιώντας την επαναλαμβανόμενη τεχνική level-wise βασισμένη στα παραγωγή υποψηφίων.

Είσοδος:
- D, βάση δεδομένων με δοσοληψίες
- min-sup, το ελάχιστο αριθμός support.

Έξοδος: L, σύνολο με όλα τα συχνά σύνολα αντικειμένων που ανήκουν στο D

Μέθοδος:

```
(1) L1 =find_frequent_1-itemsets(D);
(2) for(k = 2;Lk-1 ≠ 0;k++){
(3)     Ck = apriori_gen(Lk-1);
(4)     for each transaction t∈D { //scan D for counts
(5)             Ct = subset(Ck,t);  //get the subsets of t that are candidates
(6)             for each candidate c ∈Ct
(7)                     c.count++;
(8)     }
(9)     Lk = {c ∈ Ck| c.count ≥ min-sup}
(10) }
(11) return L = UkLk;

procedure apriori_gen(Lk-1 :frequent (k-1)-itemsets)
(1) for each itemset l1∈Lk-1
(2)     for each itemset l2 ∈ Lk-1
(3)             if((l1[1]=l2[1]) ^ (l1[2 =l2[2]) ^ … ^ ( l1 [k-2]= l2 [k-2]) ^
                   (l1 [k-1]]< l2[k-1])) then {
(4)                     c = l1 ⋈ l2; //join step: generate candidates
(5)                     if has_infrequent_subset(c, Lk-1) then
(6)                             delete c;  //prune step: remove unfruitful candidate
(7)                     else add c to Ck;
(8)             }
(9) return Ck;

procedure has_infrequent_subset(c: candidate k-itemset;
                                Lk-1: frequent (k - 1)-itemsets); //use prior knowledge
(1) for each (k - 1)-subset s of c
(2)     if s ∈ Lk-1 then
(3)             return TRUE;
(4)     return FALSE;
```

Εικόνα 3: Ψευδοκώδικας Αλγόριθμου Apriori

Σχήμα 2.1 Ψευδοκώδικας Αλγορίθμου apriori

Παράδειγμα 1

Έχουμε το παρακάτω σύνολο δεδομένων, όπου κάθε γραμμή είναι μία συναλλαγή και κάθε κελί ένα ξεχωριστό στοιχείο της συναλλαγής.

Άλφα	Βήτα	Έψιλον
Άλφα	Βήτα	Θήτα
Άλφα	Βήτα	Έψιλον
Άλφα	Βήτα	Θήτα

Πίνακας 2.1

Οι κανόνες συσχέτισης που προκύπτουν από το σύνολο δεδομένων είναι οι εξής:

1. 100% των συνόλων με Άλφα περιέχουν και Βήτα

2. 50% των συνόλων με Άλφα, Βήτα περιέχουν Έψιλον

3. 50% των συνόλων με Άλφα, Βήτα περιέχουν Θήτα

Παράδειγμα 2

Ας υποθέσουμε ότι ένα μεγάλο supermarket κρατάει τα δεδομένα από την αποθήκη για κάθε αντικείμενο. Κάθε αντικείμενο όπως το βούτυρο ή το ψωμί ξεχωρίζει από ένα μοναδικό αριθμό. Το Supermarket έχει μια βάση δεδομένων με συναλλαγές όπου κάθε συναλλαγή είναι ένα σύνολο από μοναδικούς αριθμούς όπου αγοράστηκαν μαζί. Οι συναλλαγές στη βάση είναι αποτελούνται από τα παρακάτω σύνολα:

Σύνολο αντικειμένων
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}

{2,4}

Πίνακας 2.2

Θα χρησιμοποιήσουμε Apriori για να προσδιορίσουμε τα πιο συχνά αντικείμενα της βάσης,. Για να γίνει αυτό, καθορίζουμε ότι ένα αντικείμενο είναι συχνό εάν εμφανίζεται σε τουλάχιστον 3 συναλλαγές. Ο αριθμός 3 είναι το όριο στήριξης. Το πρώτο βήμα είναι του αλγόριθμου Apriori είναι να μετρήσουμε τον αριθμό των εμφανίσεων του κάθε αντικειμένου ξεχωριστά, ο οποίος ονομάζεται υποστήριξη. Αυτό γίνεται σαρώνοντας την βάση για πρώτη φορά. Παίρνουμε τα παρακάτω αποτελέσματα:

Αντικείμενο	**Υποστήριξη**
{1}	3
{2}	6
{3}	4
{4}	5

Πίνακας 2.3

Όλα τα σύνολα με μέγεθος 1 έχουν υποστήριξη τουλάχιστον 3, άρα είναι όλα συχνά. Το επόμενο βήμα είναι να παραχθεί μια λίστα με όλα τα ζευγάρια των συχνών αντικειμένων.

Αντικείμενο	**Υποστήριξη**
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

Πίνακας 2.4

Τα ζευγάρια {1,2}, {2,3}, {2,4}, και {3,4} είναι ή μεγαλύτερα από την ελάχιστη υποστήριξη του 3, άρα είναι συχνά. Τα ζευγάρια {1,3} και {1,4} δεν είναι. Τώρα, επειδή {1,3} και {1,4} δεν είναι συχνά, κάθε σύνολο που τα περιέχει δεν είναι και αυτό συχνό. Έτσι μειώνουμε τα σύνολα. Ψάχνοντας για σύνολα τριών στην βάση και ενώ έχουνε αφαιρέσει τα σύνολα που περιέχουν τα δύο παραπάνω μένουν τα εξής:

Αντικείμενο	**Υποστήριξη**
{2,3,4}	2

Πίνακας 2.5

Στο παράδειγμά μας, δεν υπάρχουν συχνά σύνολα τριών αντικειμένων-το {2,3,4} είναι κάτω από το μικρότερο όριο και τα υπόλοιπα έχουν αποκλειστεί επειδή έχουν σύνολα δύο αντικειμένων που είναι ήδη κάτω από το όριο.

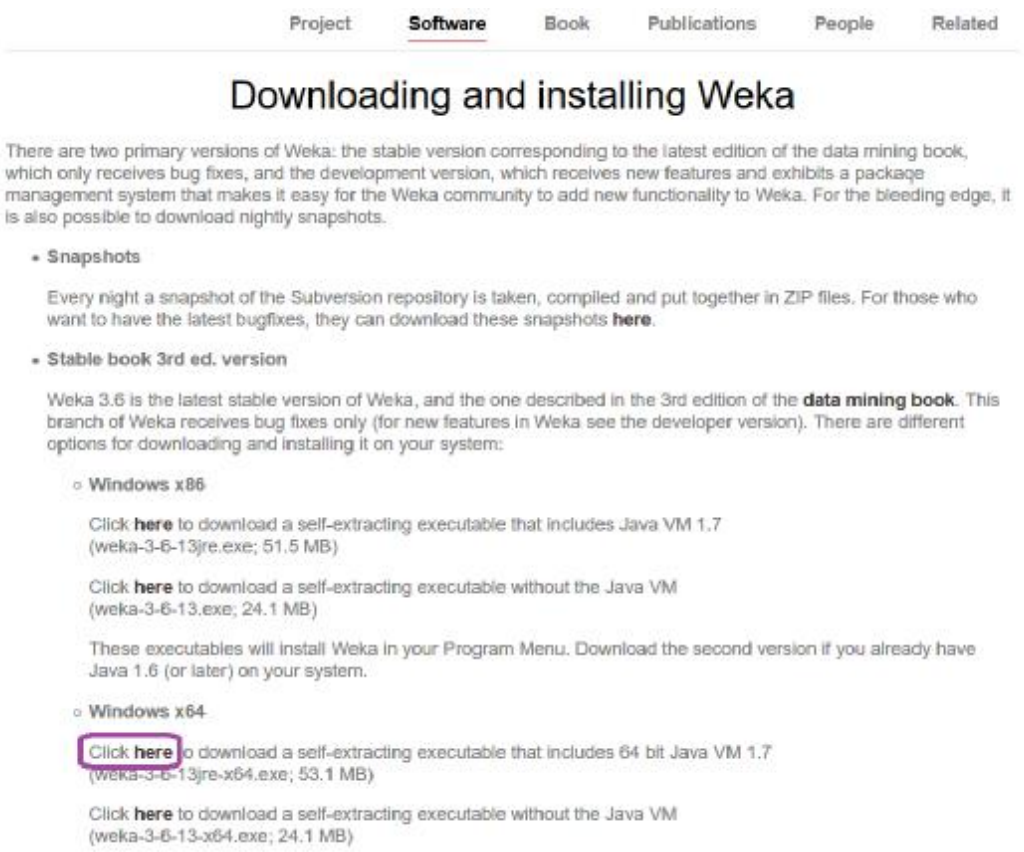
Περιορισμοί

Ο αλγόριθμος Apriori ενώ είναι σημαντικός για την συνεισφορά του στην έρευνα πάσχει από ένα σημαντικό αριθμό ανεπαρκειών και συμβιβασμών που ήταν το έναυσμα για να δημιουργηθούν άλλοι βελτιωμένοι αλγόριθμοι.

[1] .M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “Advances in Knowledge Discovery and Data Mini ng”, AAAI Press/MIT Press, 1996.

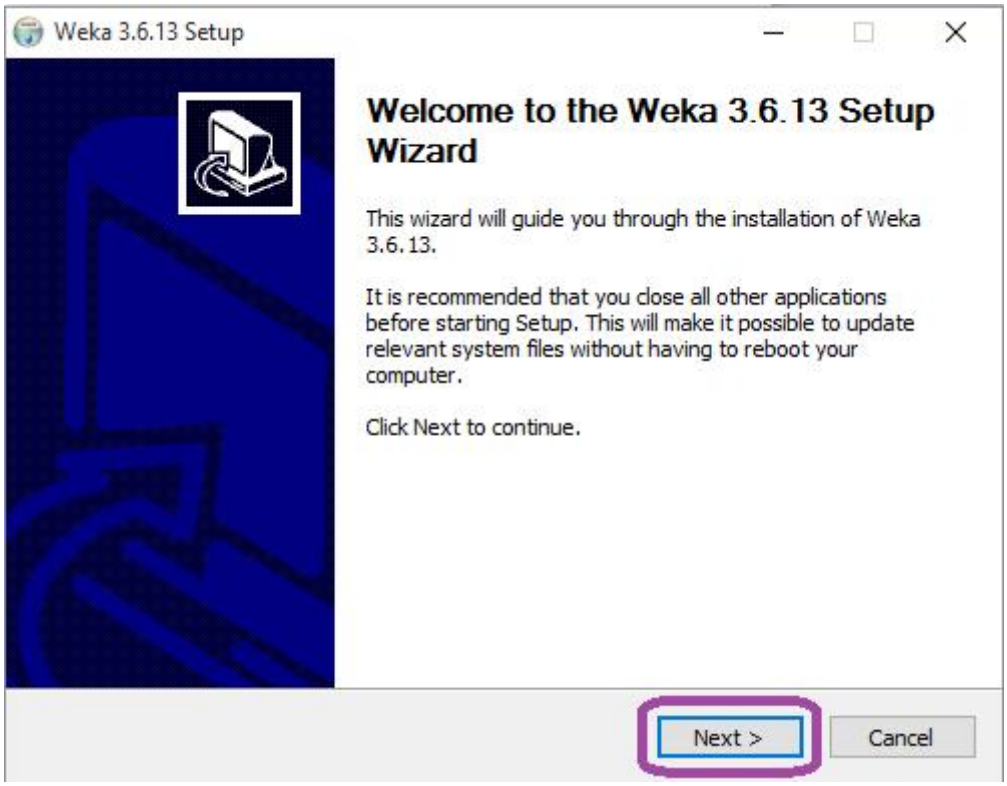
Εγκατάσταση weka

Για την εγκατάσταση του wekaπλοηγούμαστε στον ιστότοπο: www.cs.waikato.ac.nz και στη σελίδα του Softwareόπου επιλέγουμε την έκδοση του λογισμικού για την έκδοση του λειτουργικού μας.



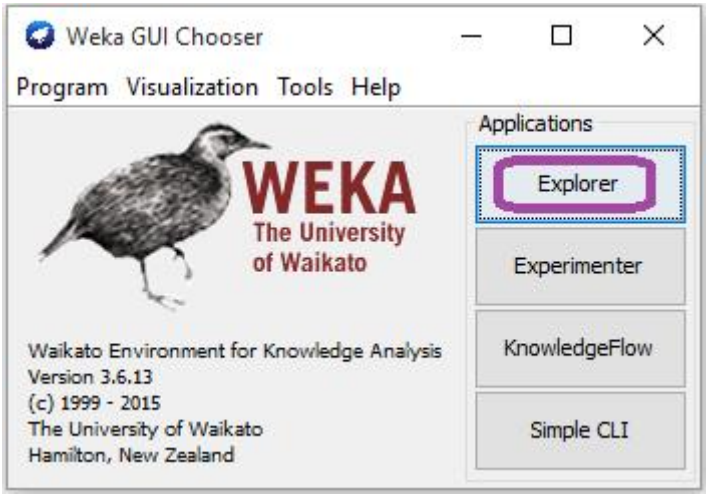
Σχήμα 3.1 Download Weka

Στη συνέχεια εκτελούμε το παρεχόμενο αρχείο και ακολουθούμε τα βήματα της εγκατάστασης.



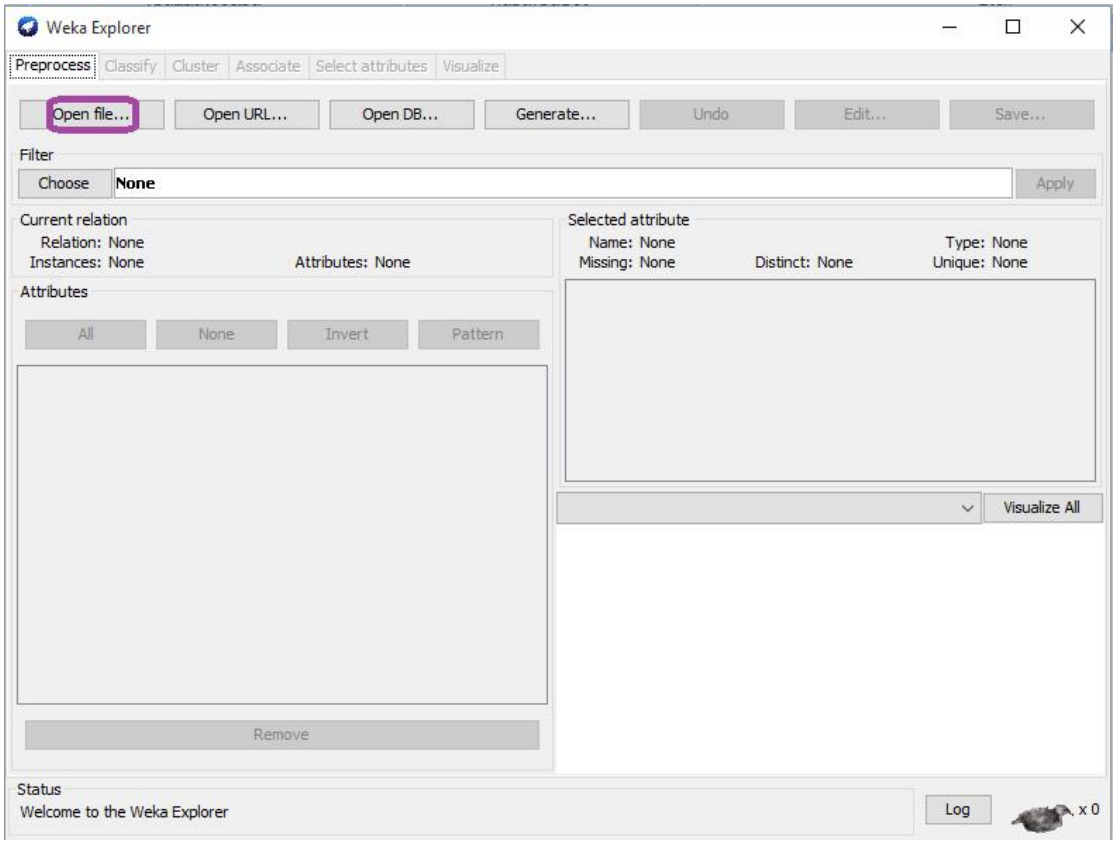
Σχήμα 3.2 Install Weka

Αφού τελειώσει η εγκατάσταση, εκκινούμε το λογισμικό Weka και επιλέγουμε Explorer.



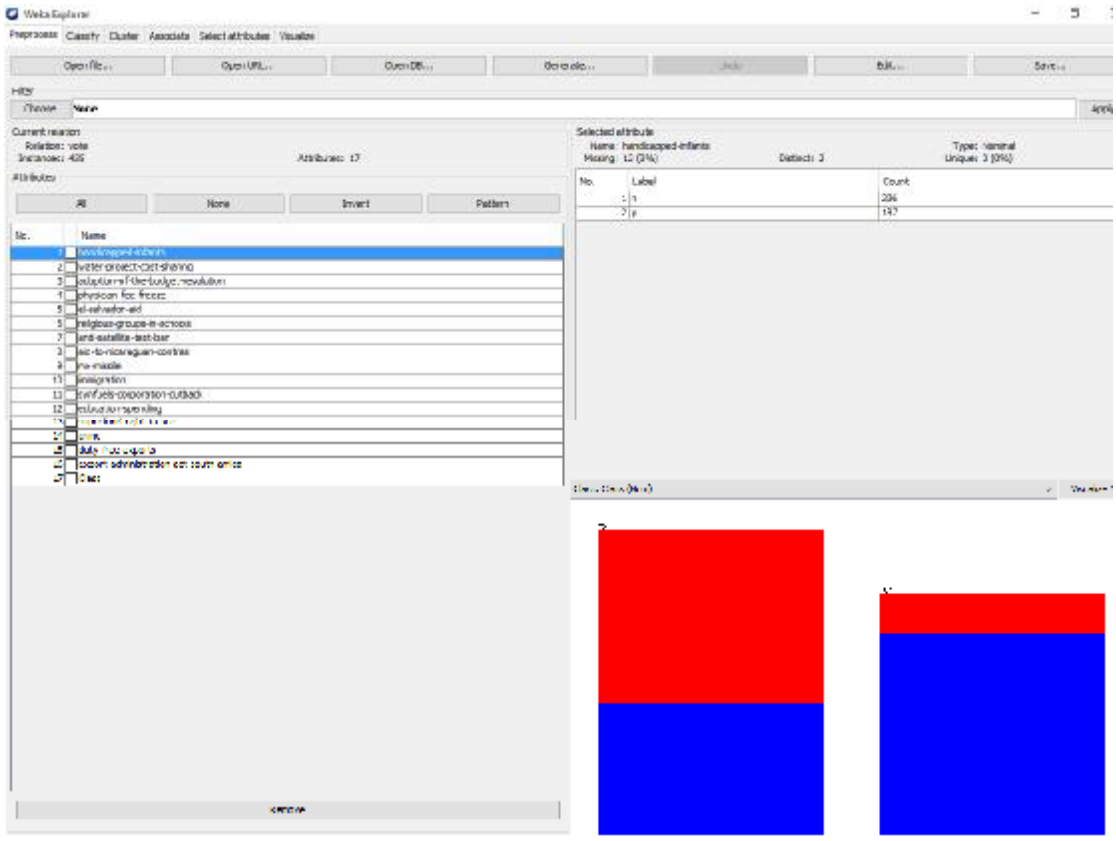
Σχήμα 3.3 Πρώτη οθόνη

Στη συνέχεια επιλέγουμε openfile και το αρχείο votes.arffπου μας ενδιαφέρει.



Σχήμα 3.4 Weka explorer

Ανοίγοντας το datasetεμφανίζεται η παρακάτω οθόνη.



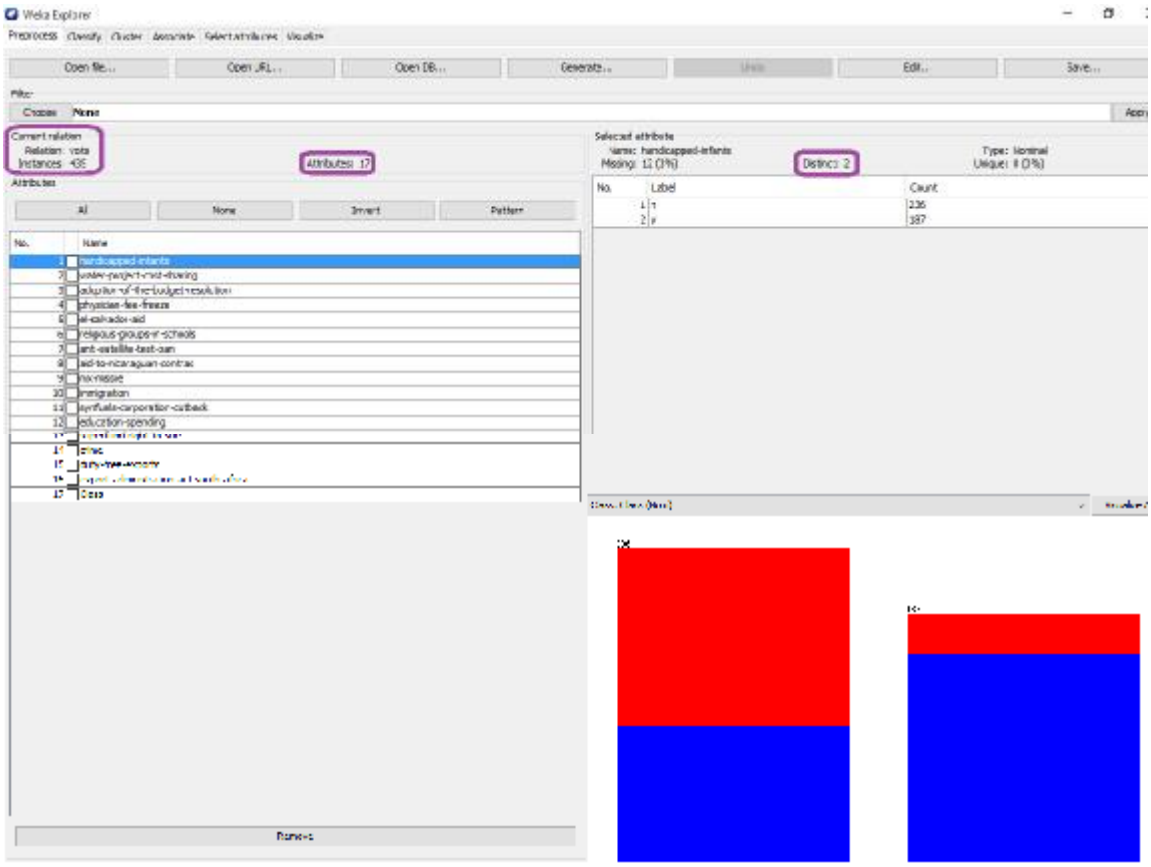
Σχήμα 3.5 Votesdataset

Περιγραφή των χαρακτηριστικών του συνόλου δεδομένων.

Το αρχείο δεδομένων vote.arff παρέχει πληροφορίες σχετικά με κάποια ψηφοφορία και την κατηγοριοποίηση των ψήφων σε 2 μεγάλες ομάδες (Δημοκράτες και Ρεπουμπλικάνους) όπως φαίνεται από το χαρακτηριστικό class.

Όλα τα χαρακτηριστικά του συνόλου votes είναι τύπου nominal (ονομαστικά) και οι τιμές που λαμβάνουν είναι της μορφής yes και no (y/n) (εκτός από το χαρακτηριστικό Class). Επιπρόσθετα όλα τα χαρακτηριστικά (εκτός από το χαρακτηριστικό Class) έχουν missing values (ελλιπείς τιμές) όπως φαίνεται στην εικόνα παρακάτω

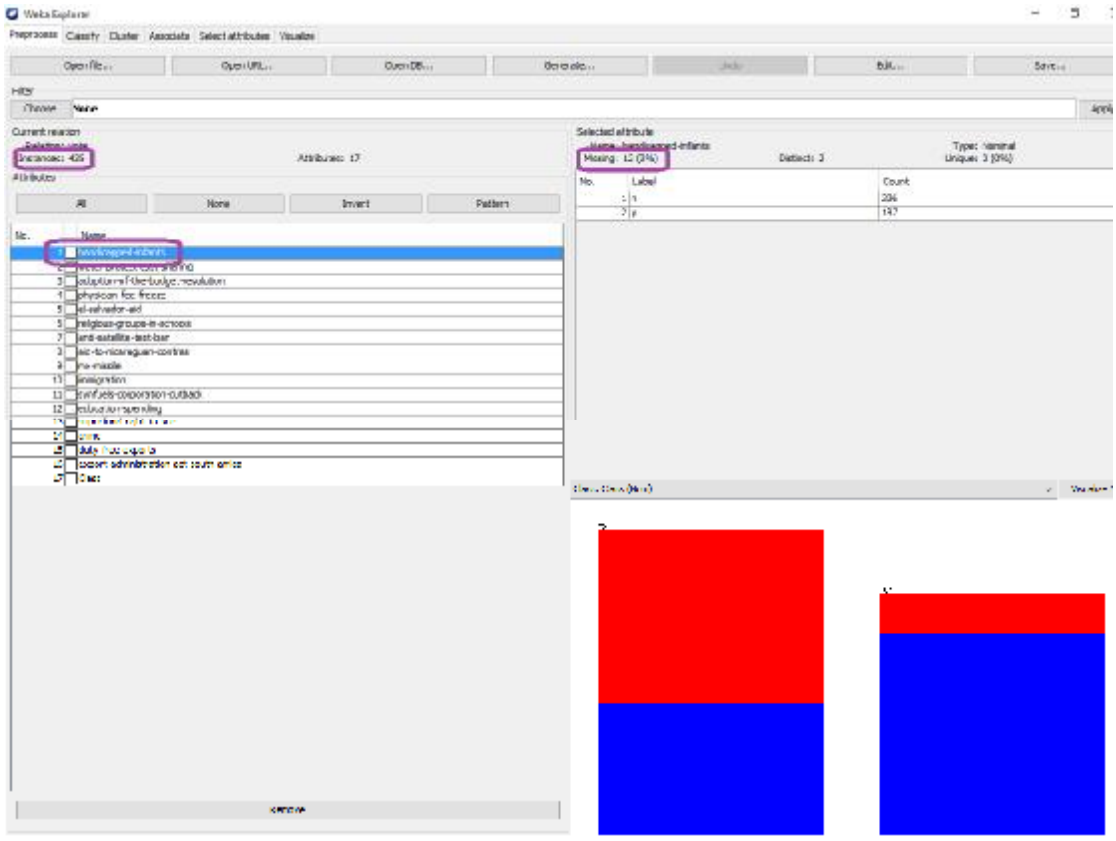
Πιο συγκεκριμένα το αρχείο των δεδομένων περιλαμβάνει 17 χαρακτηριστικά και 435 εγγραφές. Τα 17 χαρακτηριστικά του συνόλου δεδομένων, όπως μπορούμε να δούμε και από την καρτέλα Preprocess του weka, είναι:



Σχήμα 3.6 Χαρακτηριστικά, Ομάδες, Αριθμός ψήφων

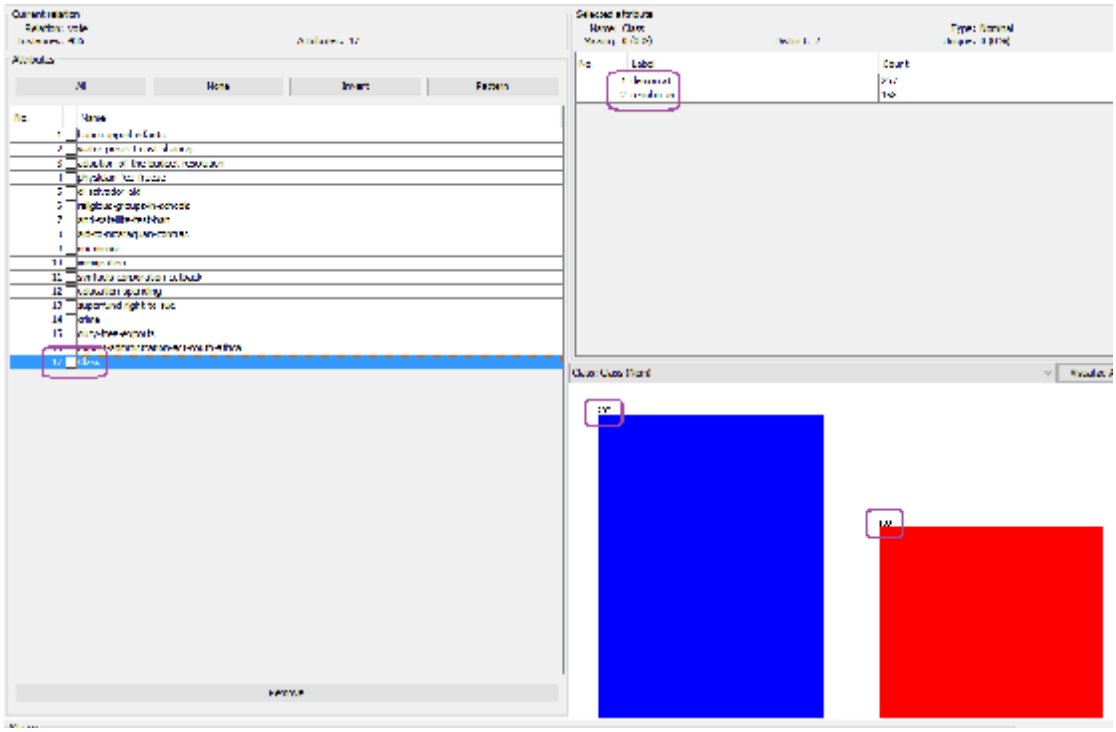
No.		Name
1	☐	handicapped-infants
2	☐	water-project-cost-sharing
3	☐	adoption-of-the-budget-resolution
4	☐	physician-fee-freeze
5	☐	el-salvador-aid
6	☐	religious-groups-in-schools
7	☐	anti-satellite-test-ban
8	☐	aid-to-nicaraguan-contras
9	☐	mx-missile
10	☐	immigration
11	☐	synfuels-corporation-cutback
12	☐	education-spending
13	☐	superfund-right-to-sue
14	☐	crime
15	☐	duty-free-exports
16	☐	export-administration-act-south-africa
17	☐	Class

Σχήμα 3.7 Data set attributes



Σχήμα 3.8 Dataset - missingvalues

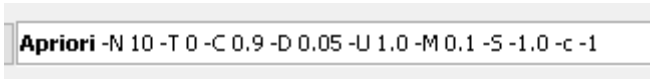
Το χαρακτηριστικό classδείχνει τις δύο κατηγορίες οι οποίες συννετείχαν στην ψηφοφορία.



Σχήμα 3.9 Datasetclassattribute

Εφαρμογή Apriori **αλγόριθμου και αποτελέσματα**

Αφού εισάγουμε το σύνολο δεδομένων στο weka. Εκτελούμε τον Apriori αλγόριθμο. Αυτό γίνεται από την καρτέλα (tab) associate. Επιλέγουμε Associator και επιλέγουμε αλγόριθμο Apriori Χρησιμοποιώντας τις παραμέτρους που φαίνονται παρακάτω.



Σχήμα 3.10 Παράμετροι του Apriori αλγόριθμου για το DatasetVotes

Ο αλγόριθμος μας δίνει τα παρακάτω αποτελέσματα παράγοντας δέκα κανόνες.

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

Best rules found:

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219    conf:(1)
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198    conf:(1)
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210    conf:(1)
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201    conf:(1)
5. physician-fee-freeze=n 247 ==> Class=democrat 245    conf:(0.99)
6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197    conf:(0.99)
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204    conf:(0.98)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198    conf:(0.98)
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197    conf:(0.97)
0. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210    conf:(0.96)
```

Σχήμα 3.11 Αποτελέσματα Apriori *αλγορίθμου στο* VotesDataset

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι το minimum support (ελάχιστη υποστήριξη) είναι s=0.45 ενώ το confidence (ελάχιστη εμπιστοσύνη) είναι c=0.9 σε 196 **εγγραφές**. Ο αλγόριθμος διενεργεί 4 περάσματα όπου σε κάθε πέρασμα περιλαμβάνει διαφορετικά στοιχειοσύνολα. Πιο συγκεκριμένα:

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

Best rules found:

 1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219    conf:(1)
 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198    conf:(1)
 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210    conf:(1)
 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201    conf:(1)
 5. physician-fee-freeze=n 247 ==> Class=democrat 245    conf:(0.99)
 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197    conf:(0.99)
 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204    conf:(0.98)
 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198    conf:(0.98)
 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197    conf:(0.97)
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210    conf:(0.96)
```

Σχήμα 3.12 Στοιχειοσύνολα VotesDataset

Ανάλυση των κανόνων

Από την εξομείωση του αλγορίθμου στο Σύνολο τιμών καταγράψαμε του δέκα κανόνες (bestrules) που παρήχθησανκαθώς και τις τιμές υποστήριξης και εμπιστοσύνης για κάθε ένα κανόνα. Παρακάτω παραθέτουμε τους κανόνες όπως παρήχθησαν από το πρόγραμμα και δίνουμε μία φυσική ερμηνεία στον κάθε ένα.

1ος Κανόνας

adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1)

Ερμηνεία:

«Εάν οι ψηφοφόροιυιοθετήσουν την ανάλυση του προϋπολογισμού και δε συμφωνούν με το πάγωμα του τέλους των ιατρών τότε είναι Δημοκρατικοίκατά 100%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=219/435=0.50, ενώ η εμπιστοσύνη είναι c=1 (100%).

Συνεπώς ο συγκεκριμένος κανόνας θα είναι πάντα αληθής.

2ος Κανόνας

adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 conf:(1)

Ερμηνεία:

«Εάν οι ψηφοφόροιυιοθετήσουντην ανάλυση του προϋπολογισμού, δε συμφωνούν με το πάγωμα του τέλους των ιατρών και ενισχύσουν τους NicaraguanContras τότε είναι Δημοκρατικοί κατά 100%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=198/435=0.46, ενώ η εμπιστοσύνη είναι c=1 (100%).

Συνεπώς ο συγκεκριμένος κανόνας θα είναι πάντα αληθής.

3ος Κανόνας

physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1)

Ερμηνεία:

«Εάν οι ψηφοφόροι δε συμφωνούν με το πάγωμα του τέλους των ιατρών και ενισχύσουν τους NicaraguanContras τότε είναι Δημοκρατικοί κατά 100%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=211/435=0.48, ενώ η εμπιστοσύνη είναι

c=1 (100%).

Συνεπώς ο συγκεκριμένος κανόνας θα είναι πάντα αληθής.

4ος Κανόνας

physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1)

Ερμηνεία:

«Εάν οι ψηφοφόροι δε συμφωνούν με το πάγωμα του τέλους των ιατρών και δεν υποστηρίξουν δαπάνες για εκπαίδευση τότε είναι Δημοκρατικοί κατά 100%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=202/435=0.46, ενώ η εμπιστοσύνη είναι c=1 (100%).

Συνεπώς ο συγκεκριμένος κανόνας θα είναι πάντα αληθής.

5ος Κανόνας

physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99)

Ερμηνεία:

«Εάν οι ψηφοφόροι δε συμφωνούν με το πάγωμα του τέλους των ιατρών τότε είναι Δημοκρατικοί κατά 99%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=247/435=0.57, ενώ η εμπιστοσύνη είναι c=0.99 (99%).

Συνεπώς ο συγκεκριμένος κανόνας δε θα είναι πάντα αληθής.

6ος Κανόνας

el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf:(0.99)

Ερμηνεία:

«Εάν οι ψηφοφόροι δεν ενισχύσουν τον ElSalvador και είναι Δημοκρατικοί τότε θα ενισχύσουν τους Nicaraguan Contras κατά 99%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=200/435=0.46, ενώ η εμπιστοσύνη είναι c=0.99 (99%).

Συνεπώς ο συγκεκριμένος κανόνας δε θα είναι πάντα αληθής.

7ος Κανόνας

el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf:(0.98)

Ερμηνεία:

«Εάν οι ψηφοφόροι δεν ενισχύσουν τον ElSalvador τότε θα ενισχύσουν τους Nicaraguan Contras κατά 98%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=204/435=0.48, ενώ η εμπιστοσύνη είναι c=0.98 (98%).

Συνεπώς ο συγκεκριμένος κανόνας δε θα είναι πάντα αληθής.

8ος Κανόνας

adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198 conf:(0.98)

Ερμηνεία:

«Εάν οι ψηφοφόροι υιοθετήσουν την ανάλυση του προϋπολογισμού, ενισχύσουν τους Nicaraguan Contras και είναι Δημοκρατικοί τότε θα συμφωνήσουν με το πάγωμα του τέλους των ιατρών κατά 98%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=198/435=0.47, ενώ η εμπιστοσύνη είναι c=0.98 (98%).

Συνεπώς ο συγκεκριμένος κανόνας δε θα είναι πάντα αληθής.

9ος Κανόνας

el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf:(0.97)

Ερμηνεία:

«Εάν οι ψηφοφόροι δεν ενισχύσουν τον ElSalvador και ενισχύσουν τους Nicaraguan Contras τότε είναι Δημοκρατικοί κατά 97%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=197/435=0.47, ενώ η εμπιστοσύνη είναι c=0.97 (97%).

Συνεπώς ο συγκεκριμένος κανόνας δε θα είναι πάντα αληθής.

10ος**Κανόνας**

aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210 conf:(0.96)

Ερμηνεία:

«Εάν οι ψηφοφόροι ενισχύσουν τους Nicaraguan Contras και είναι Δημοκρατικοί τότε δε θα συμφωνήσουν με το πάγωμα του τέλους των γιατρών κατά 96%»

Τιμές:

Η υποστήριξη του συγκεκριμένου κανόνα είναι s=218/435=0.50, ενώ η εμπιστοσύνη είναι c=0.96 (96%).

Συνεπώς ο συγκεκριμένος κανόνας δε θα είναι πάντα αληθής.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Από τους παραπάνω 10 κανόνες μόνο οι 4 πρώτοι βρέθηκαν να είναι πάντοτε αληθείς και αυτό διότι έχουν εμπιστοσύνη c=1 (100%).Συνοπτικά λοιπόν, οι κανόνες είναι οι εξής:

1ος Κανόνας

adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1)

2ος Κανόνας

adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 conf:(1)

3ος Κανόνας

physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1)

4ος Κανόνας

physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1)

ΒΙΒΛΙΟΓΡΑΦΙΑ

Data Mining. Practical machine learning Tools and Techniques. Ian H. Witten. Eibe Frank

Wikipedia. Εξόρυξη δεδομένων.

https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD

Web data mining. Exploring hyperlinks contents and usage data. Bing Liu

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.

Μιχάλης Βαζιργιάννης, Μαρία Χαλκίδη, Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, Εκδ. Gutenberg.

Sotiris Kotsiantis, Dimitris Kanellopoulos, Association Rules Mining: A Recent Overview,2006.

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

Σαλατάς Ι. (2011). Υλοποίηση και εφαρμογή Τεχνητών Νευρωνικών Δικτύων για την πρόβλεψη χρονοσειρών συναλλαγματικών ισοτιμιών, Ελληνικό Ανοιχτό Πανεπιστήμιο.

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

Κωνσταντίνος Δ. (2007). Τεχνητά Νευρωνικά Δίκτυα., Εκδόσεις Κλειδάριθμος, Αθήνα

Κωτσόπουλος, Δ. (2012). Προηγμένες μέθοδοι ταξινόμησης για την πρόβλεψη και την ανίχνευση μοτίβου σε δεδομένα ωοπαραγωγής. Θεσσαλονίκη: ΑΠΘ

Παπαδόπουλος, Χ. (2010). Predicting the Choice of Contraceptive Method using Classification. Πανεπιστήμιο Μακεδονίας

Σκούρα, σημειώσεις. http://slideplayer.gr/slide/2435319/