ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αλγόριθμοι Data Mining

Της φοιτήτριας

Καλέμου Δήμητρα

Αρ. Μητρώου: 05/2761

Επιβλέπων καθηγητής

Κ. Δημήτρης Δέρβος

Θεσσαλονίκη 2011

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αλγόριθμοι Data Mining

Της φοιτήτριας

Καλέμου Δήμητρα

Αρ. Μητρώου: 05/2761

Επιβλέπων καθηγητής

Κ. Δημήτρης Δέρβος

Θεσσαλονίκη 2011

ΠΡΟΛΟΓΟΣ

Η ανάγκη επεξεργασίας τεράστιου όγκου δεδομένων έχει οδηγήσει στην ανάγκη αποτελεσματικού χειρισμού και αποθήκευσης τους. Οι βάσεις δεδομένων που χρησιμοποιούνται για το σκοπό αυτό αναπτύσσονται συνεχώς και παρέχουν νέες λειτουργίες. Η επιστήμη των υπολογιστών έρχεται με τη σειρά της να συμβάλλει στο σκοπό αυτό παρέχοντας την υπολογιστική ισχύ και τη δυνατότητα προγραμματισμού των διαδικασιών επεξεργασίας δεδομένων. Ωστόσο, υπάρχει πλέον η ανάγκη όχι μόνο αποτελεσματικού χειρισμού των δεδομένων αλλά και δυνατότητα χρήσης τους για την εύρεση στοιχείων που δείχνουν συμπεριφορές ή και μελλοντικές τάσεις με βάση τα δεδομένα αυτά. Στο σημείο αυτό, η τεχνολογία της εξόρυξης δεδομένων παρέχει τις τεχνικές και τους αλγορίθμους της για την εύρεση προτύπων στα δεδομένα και την εξαγωγή χρήσιμων συμπερασμάτων. Στην παρούσα εργασία περιγράφονται οι τεχνικές και οι αλγόριθμοι της εξόρυξης δεδομένων.

ΠΕΡΙΛΗΨΗ

Με τον όρο εξόρυξη δεδομένων γίνεται αναφορά στη διαδικασία της ανάλυσης δεδομένων από διαφορετικές οπτικές γωνίες και συνοψίζοντας τα σε χρήσιμες πληροφορίες. Η εξόρυξη δεδομένων αποτελεί έναν αναπτυσσόμενο κλάδο της επιστήμης των υπολογιστών και εφαρμόζεται σε βάσεις δεδομένων. Μπορεί να χρησιμοποιηθεί σε οποιαδήποτε επιχείρηση ή οργανισμό που χρησιμοποιεί μεγάλες ποσότητες δεδομένων σε βάσεις δεδομένων. Το λογισμικό εξόρυξης είναι μια σειρά αναλυτικών εργαλείων για την ανάλυση των δεδομένων. Για τη διαδικασία της εξόρυξης χρησιμοποιείται ένα πλήθος τεχνικών και εργαλείων. Αρχικά, υπάρχουν οι ήδη γνωστές και καθιερωμένες τεχνικές της στατιστικής, όπως η κατάταξη, η συσταδοποίηση και η παλινδρόμηση, με πλήθος αλγορίθμων που εφαρμόζουν τις τεχνικές αυτές. Υπάρχουν επίσης και εργαλεία τα οποία γνωρίζουν ανάπτυξη τα τελευταία κυρίως χρόνια και είναι οι συνδυαστικοί κανόνες, η σειριακή ανάλυση και τα δένδρα αποφάσεων. Επίσης, λόγω της αύξησης της υπολογιστικής ισχύος των συστημάτων υπολογιστών, υπάρχει πλέον η δυνατότητα χρήσης πολυεπεξεργαστικών συστημάτων, δίνοντας έτσι τη δυνατότητα ανάπτυξης κατανεμημένων και παράλληλων αλγορίθμων που χρησιμοποιούνται στην εξόρυξη δεδομένων. Υπάρχουν πλήθος άλλων τεχνικών που χρησιμοποιούνται και επίσης η επιστήμη της εξόρυξης δεδομένων βρίσκει ευρεία εφαρμογή σε πολλούς επιχειρηματικούς και ερευνητικούς κλάδους και γενικά οπουδήποτε γίνεται διαχείριση μεγάλου όγκου δεδομένων και χρειάζεται να εξαχθούν συμπεράσματα από τα δεδομένα αυτά.

ABSTRACT

The term data mining refers to the process of analyzing data from different perspectives and summarizing them into useful information. Data mining is a growing branch of computer science and is applied to databases. It can be used in any business or organization that uses large data amounts in databases. The software is a series of analytical tools for data analysis. The mining process uses a number of techniques and tools. Initially, there are the lready known and established statistical techniques such as classification, clustering and regression, with many algorithms implementing these techniques. There are also tools that met increased, growth especially in recent years, and these are association rules, sequential analysis and decision trees. Also, due to the increased computing power of computer systems, it is now possible to use multiprocessor systems, thus enabling the development of distributed and parallel algorithms that are used in data mining. There are many other techniques used in this area and moreover, the data mining science is widely applied in many business and research sectors and generally everywhere that there is the need for managing large volumes of data and extract conclusions from these data.

ΕΥΧΑΡΙΣΤΙΕΣ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ ΚΑΙ ΠΙΝΑΚΩΝ

ΕΙΣΑΓΩΓΗ

Η εξόρυξη δεδομένων (data mining) είναι ουσιαστικά η εφαρμογή τεχνικών για την εξαγωγή χρήσιμων πληροφοριών και κανόνων από μεγάλα σύνολα δεδομένων. Πρόκειται για έναν συνεχώς εξελισσόμενο τομέα της επιστήμης της Πληροφορικής και συνδυάζει τις επιστήμες της στατιστικής, των βάσεων δεδομένων και της μηχανικής μάθησης, με ολοένα και αυξανόμενη εφαρμογή σε πολλούς τομείς. Λαμβάνοντας υπόψη το γεγονός ότι ο όγκος των πληροφοριών που χρησιμοποιούνται και αναλύονται σε οποιοδήποτε επιστημονικό και εμπορικό τομέα είναι πλέον τόσο μεγάλος ώστε η διαχείριση του είναι δυνατή μόνο με τη χρήση υπολογιστικών συστημάτων, η ανάγκη χρήσης τεχνικών διαχείρισης δεδομένων είναι απαραίτητη και παρέχει πολλά οφέλη.

Στο πλαίσιο αυτό, ο σκοπός της εργασίας είναι η παρουσίαση των βασικότερων κατηγοριών τεχνικών και αλγορίθμων εξόρυξης δεδομένων, καθώς και των προϊόντων λογισμικού που κάνουν εφαρμογή των τεχνικών και αλγορίθμων αυτών ώστε να εξάγουν χρήσιμες πληροφορίες. Οι στόχοι της εργασίας είναι αφενός η εύρεση των κατηγοριών των τεχνικών που χρησιμοποιούνται στην εξόρυξη δεδομένων και αφετέρου η εύρεση και παρουσίαση των βασικών αλγορίθμων της κάθε κατηγορίας και των χαρακτηριστικών τους. Στη συνέχεια, στόχος είναι η παρουσίαση των προϊόντων λογισμικού τα οποία εφαρμόζουν τις τεχνικές αυτές μέσα σε μια προσπάθεια εκτίμησης του εύρους εφαρμογών που βρίσκει η εξόρυξη δεδομένων.

Το περιεχόμενο των κεφαλαίων είναι ως εξής:

Στο πρώτο κεφάλαιο περιγράφονται όλες εκείνες οι έννοιες που έχουν να κάνουν με την εξόρυξη δεδομένων. Παρουσιάζονται τα χαρακτηριστικά της, οι χρήσεις και εφαρμογές της, τα ζητήματα που προκύπτουν, γίνεται μια κατηγοριοποίηση των βασικότερων τεχνικών εξόρυξης δεδομένων και παρουσιάζονται τα βασικά χαρακτηριστικά των δεδομένων που καθορίζουν και την επιλογή της αντίστοιχης τεχνικής.

Στο δεύτερο κεφάλαιο περιγράφονται οι βασικές τεχνικές στατιστικής, που είναι η κατάταξη και η παλινδρόμηση.

Στο τρίτο κεφάλαιο περιγράφονται μερικές ακόμα τεχνικές στατιστικής, που αφορούν την τμηματοποίηση και είναι η συσταδοποίηση, η δημογραφική συσταδοποίηση, οι αυξητικοί αλγόριθμοι και η γειτνίαση.

Στο τέταρτο κεφάλαιο περιγράφονται οι συνδυαστικοί κανόνες και οι αλγόριθμοι που εφαρμόζουν την τεχνική αυτή.

Στο πέμπτο κεφάλαιο περιγράφονται τα σειριακά πρότυπα και η διαδικασία με την οποία γίνεται εξόρυξη με σειριακούς αλγορίθμους.

Στο έκτο κεφάλαιο περιγράφεται πως η επιβλεπόμενη μάθηση εφαρμόζεται στην εξόρυξη δεδομένων.

Στο έβδομο κεφάλαιο περιγράφονται τα δένδρα απόφασης και οι αλγόριθμοι που χρησιμοποιούν την τεχνική αυτή για να κάνουν εξόρυξη δεδομένων σε κατάλληλα δεδομένα.

Στο όγδοο κεφάλαιο περιγράφεται η έννοια των παράλληλων και κατανεμημένων συστημάτων καθώς και των αλγορίθμων που χρησιμοποιούν τα συστήματα αυτά για λόγους εξόρυξης δεδομένων.

Στο ένατο κεφάλαιο παρουσιάζονται συνοπτικά για λόγους πληρότητας ορισμένες ακόμα τεχνικές που εφαρμόζονται στην εξόρυξη δεδομένων καθώς και εφαρμογές και προϊόντα λογισμικού που χρησιμοποιούν τεχνικές εξόρυξης ώστε να εξάγουν χρήσιμα συμπεράσματα.

Στο δέκατο κεφάλαιο βρίσκονται τα συμπεράσματα.

1. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΙΚΕΣ

1.1. Εισαγωγή

Η εξόρυξη δεδομένων, με άλλα λόγια η εξαγωγή κρυφών πληροφοριών πρόβλεψης από μεγάλες βάσεις δεδομένων, είναι μια ισχυρή νέα τεχνολογία με μεγάλες δυνατότητες, με σκοπό να βοηθήσει τις εταιρείες να επικεντρωθούν στις πιο σημαντικές πληροφορίες στις αποθήκες δεδομένων τους. Τα εργαλεία εξόρυξης δεδομένων προβλέπουν τις μελλοντικές τάσεις και συμπεριφορές, που επιτρέπουν στις επιχειρήσεις να πάρουν βασισμένες στη γνώση αποφάσεις. Οι αυτοματοποιημένες, προοπτικές αναλύσεις που προσφέρονται από την εξόρυξη δεδομένων κινούνται πέρα από τις αναλύσεις των γεγονότων του παρελθόντος που παρέχονται από αναδρομικά εργαλεία, τυπικά σε συστήματα υποστήριξης αποφάσεων. Τα εργαλεία εξόρυξης δεδομένων μπορούν να απαντήσουν σε ερωτήσεις επιχειρήσεων που παραδοσιακά ήταν πολύ χρονοβόρες να επιλυθούν. Κάνουν ξεκάθαρες τις βάσεις δεδομένων για τα κρυμμένα μοτίβα, βρίσκοντας πληροφορίες πρόβλεψης που οι εμπειρογνώμονες μπορεί να παραλείψουν επειδή βρίσκονται έξω από τις προσδοκίες τους (Thearling, 2010).. Στο κεφάλαιο αυτό παρουσιάζεται η έννοια της εξόρυξης δεδομένων και των πιο σχετικών σε αυτή εννοιών, καθώς και μια παρουσίαση των τεχνικών που χρησιμοποιεί.

1.2. Εξόρυξη Δεδομένων

1.2.1. Θεμελιώδεις Γνώσεις στην Εξόρυξη Δεδομένων

Η **εξόρυξη δεδομένων (data mining)**, γνωστή και ως **ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases, KDD)**, είναι η πρακτική της αυτόματης αναζήτησης μεγάλων αποθηκών δεδομένων για πρότυπα. Για να γίνει αυτό, η εξόρυξη δεδομένων χρησιμοποιεί υπολογιστικές τεχνικές από τη Στατιστική και την Αναγνώριση προτύπων (WordIQ.com, 2010).

Η διαδικασία KDD μπορεί να χωριστεί σε τέσσερα στάδια. Αρχικά, τα ακατέργαστα δεδομένα περνούν από το αρχικό στάδιο το οποίο είναι η **επιλογή των δεδομένων**, κατά το οποίο προσδιορίζεται το σύνολο δεδομένων και τα γνωρίσματα εκείνα που ενδιαφέρουν με βάση το στόχο. Το δεύτερο στάδιο είναι ο **καθαρισμός των δεδομένων** κατά το οποίο απομακρύνεται ο θόρυβος και οι προς εξαίρεση τιμές, μετασχηματίζονται οι τιμές των πεδίων σε κοινές μονάδες μέτρησης, δημιουργούνται νέα πεδία συνδυάζοντας τα ήδη υπάρχοντα και τοποθετούνται τα δεδομένα στο σχεσιακό σχήμα που θα χρησιμοποιηθεί στην είσοδο της επεξεργασίας της εξόρυξης δεδομένων. Στο στάδιο αυτό μπορεί να περιλαμβάνεται και η αποκανονικοποίηση των σχετικών πινάκων. Στη συνέχεια ακολουθεί το στάδιο της **εξόρυξης δεδομένων**, στο οποίο γίνεται η εξαγωγή των πραγματικών πρότυπων σχημάτων. Το τέταρτο στάδιο είναι το στάδιο της **αξιολόγησης**, τα δεδομένα παρουσιάζονται σε μια μορφή η οποία έχει διαμορφωθεί έτσι ώστε να είναι κατανοητή στον τελικό χρήστη, όπως για παράδειγμα με παραστατικό, οπτικό τρόπο. Τα αποτελέσματα οποιουδήποτε βήματος μπορούν να οδηγήσουν πίσω στην επεξεργασία κάποιου προηγούμενου σταδίου ώστε να επαναληφθεί η διαδικασία χρησιμοποιώντας τη νέα γνώση που αποκτήθηκε (Ramakrishnan R. et al., 2002).

1.2.1.1. Ορισμοί

Δυο ορισμοί που έχουν δοθεί στην εξόρυξη δεδομένων είναι οι ακόλουθοι: *"Η μη τετριμμένη εξαγωγή εσωτερικών, προηγουμένως άγνωστων, και ενδεχομένως χρήσιμων πληροφοριών από τα δεδομένα"* και *"Η επιστήμη της εξαγωγής χρήσιμων πληροφοριών από μεγάλα σύνολα δεδομένων ή βάσεις δεδομένων"*. Αν και συνήθως χρησιμοποιείται σε σχέση με την ανάλυση των δεδομένων, η εξόρυξη δεδομένων, όπως και η τεχνητή νοημοσύνη, είναι ένας γενικός όρος και χρησιμοποιείται με ποικίλο νόημα σε ένα ευρύ φάσμα πλαισίων (WordIQ.com,2010).

Σε γενικές γραμμές, η εξόρυξη δεδομένων (μερικές φορές αποκαλούμενη και ανακάλυψη γνώσης ή δεδομένων), είναι η διαδικασία της ανάλυσης δεδομένων από διαφορετικές οπτικές γωνίες και συνοψίζοντας τα σε χρήσιμες πληροφορίες - πληροφορίες που μπορούν να χρησιμοποιηθούν για την αύξηση των εσόδων, για περικοπές δαπανών, ή και τα δύο. Το λογισμικό εξόρυξης είναι μια σειρά

αναλυτικών εργαλείων για την ανάλυση των δεδομένων. Επιτρέπει στους χρήστες να αναλύουν δεδομένα από πολλές διαφορετικές διαστάσεις ή γωνίες, να ταξινομήσουν, και να συνοψίσουν τις σχέσεις που εντοπίστηκαν. Τεχνικά, η εξόρυξη δεδομένων είναι η διαδικασία της εύρεσης συσχετισμών ή μοτίβων ανάμεσα σε δεκάδες τομείς σε μεγάλες σχεσιακές βάσεις δεδομένων. Αν και η εξόρυξη δεδομένων είναι ένας σχετικά νέος όρος, η τεχνολογία δεν είναι. Οι εταιρείες έχουν χρησιμοποιήσει ισχυρούς υπολογιστές για να αναλύσουν όγκους δεδομένων σαρωτή σούπερ μάρκετ και να αναλύσουν εκθέσεις έρευνας αγοράς για χρόνια. Ωστόσο, οι συνεχείς καινοτομίες στην επεξεργαστική ισχύ του υπολογιστή, στην αποθήκευση σε δίσκους και στο λογισμικό στατιστικής αυξάνουν δραματικά την ακρίβεια της ανάλυσης, ενώ μειώνουν το κόστος (Palace, 1996).

1.2.1.2. Αποθήκες δεδομένων

Δραματικές εξελίξεις στη συλλογή δεδομένων, στην επεξεργαστική ισχύ, στη μετάδοση δεδομένων, καθώς και στις δυνατότητες αποθήκευσης επιτρέπουν στους οργανισμούς να ενσωματώσουν διάφορες βάσεις δεδομένων τους σε **αποθήκες δεδομένων (data warehouses)**. Η αποθήκη δεδομένων ορίζεται ως μια διαδικασία κεντρικής διαχείρισης των δεδομένων και ανάκτησης. Η αποθήκευση δεδομένων, όπως και η εξόρυξη δεδομένων, είναι ένας σχετικά νέος όρος, αν η ίδια η έννοια υπάρχει εδώ και χρόνια. Η αποθήκευση δεδομένων αποτελεί μια ιδεατή διατήρηση κεντρικού αποθετηρίου όλων των οργανωτικών στοιχείων. Η συγκέντρωση των δεδομένων είναι απαραίτητη για να μεγιστοποιήσει την πρόσβαση των χρηστών και την ανάλυση. Οι δραματικές τεχνολογικές εξελίξεις κάνουν αυτό το όραμα πραγματικότητα για πολλές εταιρείες. Οι εξίσου δραματικές εξελίξεις στο λογισμικό ανάλυσης δεδομένων επιτρέπουν στους χρήστες να έχουν πρόσβαση σε αυτά τα δεδομένα ελεύθερα. Το λογισμικό ανάλυσης δεδομένων είναι αυτό που υποστηρίζει η εξόρυξη δεδομένων (Palace, 1996).

1.2.1.3. Στοιχεία και τύποι ερωτημάτων

Η εξόρυξη δεδομένων αποτελείται από πέντε βασικά στοιχεία (Palace, 1996):

- Εξαγωγή, μετασχηματισμός και φόρτωση δεδομένων συναλλαγών στο σύστημα αποθήκευσης δεδομένων.

- Η αποθήκευση και διαχείριση των δεδομένων σε ένα πολυδιάστατο σύστημα βάσεων δεδομένων.
- Παροχή πρόσβασης σε δεδομένα σε επιχειρηματικούς αναλυτές και ειδικούς τεχνολογίας πληροφοριών.
- Ανάλυση των δεδομένων από λογισμικό εφαρμογών.
- Παρουσίαση των δεδομένων σε μια χρήσιμη μορφή, όπως σε έναν γράφο ή έναν πίνακα.

Ενώ η τεχνολογία πληροφοριών μεγάλης κλίμακας έχει εξελίξει τις ξεχωριστές συναλλαγές και τα συστήματα ανάλυσης δεδομένων, η εξόρυξη δεδομένων αποτελεί το συνδετικό κρίκο μεταξύ των δύο. Το λογισμικό εξόρυξης αναλύει τις σχέσεις και τα πρότυπα στα αποθηκευμένα δεδομένα συναλλαγών που βασίζονται σε αόριστα ερωτήματα χρήστη. Διάφοροι τύποι αναλυτικού λογισμικού είναι διαθέσιμα: στατιστική, μηχανική μάθηση, και νευρωνικά δίκτυα. Γενικά, ζητείται οποιοσδήποτε από τις τέσσερις τύπους σχέσεων (Palace, 1996):

- Κλάσεις (Classes): αποθηκευμένα δεδομένα χρησιμοποιούνται για να εντοπίσουν δεδομένα σε προκαθορισμένες ομάδες. Για παράδειγμα, μια αλυσίδα εστιατορίων θα μπορούσε να αναζητήσει στοιχεία αγοράς πελάτη για να καθορίσει πότε οι πελάτες επισκέπτονται και τι συνήθως παίρνουν. Οι πληροφορίες αυτές θα μπορούσαν να χρησιμοποιηθούν για την αύξηση της κίνησης με την χρήση σπεσιαλιτέ ημέρας.
- Συστάδες (Clusters)ι: Στοιχεία ομαδοποιούνται ανάλογα με λογικές σχέσεις ή προτιμήσεις καταναλωτών. Για παράδειγμα, τα δεδομένα μπορούν να εξορυχθούν να προσδιοριστεί το τμήμα της αγοράς ή τις σχέσεις των καταναλωτών.
- Συσχετίσεις (Associations): Τα δεδομένα μπορούν να εξορύσσονται για τον εντοπισμό των ενώσεων. Το παράδειγμα μπύρα-πάνα είναι ένα παράδειγμα συσχετιστικής εξόρυξης.
- Σειριακά Μοντέλα: τα δεδομένα εξορύσσονται για την πρόβλεψη προτύπων συμπεριφοράς και τάσεων. Για παράδειγμα, ένας εξωτερικός πωλητής

λιανικής πώλησης εξοπλισμού θα μπορούσε να προβλέψει την πιθανότητα αγοράς ενός σακιδίου με βάση την αγορά του καταναλωτή σε υπνόσακους και παπούτσια πεζοπορίας.

Σήμερα, οι εφαρμογές εξόρυξης δεδομένων είναι διαθέσιμες σε όλα τα μεγέθη συστημάτων για mainframe, client/server, καθώς και πλατφόρμες PC. Οι τιμές συστημάτων κυμαίνονται από αρκετές χιλιάδες δολάρια για τις μικρότερες εφαρμογές μέχρι και 1 εκατομμύριο δολάρια ανά terabyte για τις μεγαλύτερες. Οι εφαρμογές μεγάλων επιχειρήσεων, γενικά, κυμαίνονται σε μέγεθος από 10 gigabytes μέχρι πάνω από 11 terabytes. Η NCR έχει την ικανότητα να παραδώσει εφαρμογές που υπερβαίνουν τα 100 terabytes. Υπάρχουν δύο κρίσιμοι τεχνολογικοί οδηγοί (Palace, 1996):

- Μέγεθος της βάσης δεδομένων: όσα περισσότερα δεδομένα δέχονται επεξεργασία, τόσο πιο δυνατό σύστημα απαιτείται.
- Περιπλοκότητα ερωτήματος: όσο πιο περίπλοκα τα ερωτήματα και όσο μεγαλύτερος ο αριθμός των ερωτημάτων τα οποία δέχονται επεξεργασία από το σύστημα, τόσο πιο δυνατό σύστημα απαιτείται.

1.2.2. Χρήσεις

Η εξόρυξη δεδομένων χρησιμοποιείται κυρίως σήμερα από τις εταιρείες με ισχυρή εστίαση καταναλωτών - λιανικό εμπόριο, οικονομικά, επικοινωνία, και οι εμπορικές οργανώσεις. Δίνει τη δυνατότητα στις εταιρείες να καθορίζουν τις σχέσεις μεταξύ των «εσωτερικών» παραγόντων, όπως η τιμή, η τοποθέτηση του προϊόντος, ή τα προσόντα του προσωπικού, και των «εξωτερικών» παραγόντων όπως οι οικονομικοί δείκτες, ο ανταγωνισμός, και τα δημογραφικά στοιχεία των πελατών. Αυτό τους δίνει τη δυνατότητα να διαπιστώσουν την επίδραση επί των πωλήσεων, την ικανοποίηση των πελατών, και τα εταιρικά κέρδη. Τέλος, τους δίνει τη δυνατότητα να αναζητήσουν συνοπτικές πληροφορίες για να δουν αναλυτικά δεδομένα συναλλαγών. Με την εξόρυξη δεδομένων, ένας πωλητής θα μπορούσε να χρησιμοποιήσει τα αρχεία σημείου πώλησης αγορών και οι πελάτες να δέχονται ατομικές προωθητικές ενέργειες με βάση το ιστορικό αγοράς ενός ατόμου. Με την εξόρυξη δημογραφικών στοιχείων από την παρατήρηση ή τις

κάρτες εγγύησης, ο λιανοπωλητής θα μπορούσε να αναπτύξει τα προϊόντα και τις προσφορές σε συγκεκριμένες ομάδες πελατών (Palace, 1996).

Η τεχνολογία αποθήκευσης και διαχείρισης σε σχεσιακές βάσεις δεδομένων είναι επαρκής για πολλές εφαρμογές εξόρυξης δεδομένων με λιγότερο από 50 gigabytes. Ωστόσο, αυτή η υποδομή θα πρέπει να ενισχυθεί σημαντικά για την υποστήριξη μεγαλύτερων εφαρμογών. Ορισμένοι πωλητές έχουν προσθέσει εκτεταμένες δυνατότητες ευρετηρίου για τη βελτίωση της απόδοσης ερωτήματος. Άλλοι χρησιμοποιούν νέες αρχιτεκτονικές υλικού, όπως Μαζικούς Παράλληλους Επεξεργαστές (Massively Parallel Processors (MPP)), προκειμένου να επιτευχθούν βελτιώσεις μεγαλύτερης τάξης μεγέθους στο χρόνο ερωτήματος. Για παράδειγμα, τα συστήματα MPP από την NCR συνδέουν εκατοντάδες υψηλής ταχύτητας επεξεργαστές Pentium για την επίτευξη επιπέδων απόδοσης που υπερβαίνουν αυτές από μεγαλύτερους υπερυπολογιστές (Palace, 1996).

1.2.2.1. Οπτικοποίηση δεδομένων

Η **οπτικοποίηση δεδομένων (data visualization)** δίνει τη δυνατότητα στον αναλυτή να έχει μια βαθύτερη, πιο εσωτερική κατανόηση των δεδομένων και έτσι να εργαστεί σωστά στην εξόρυξη των δεδομένων. Η εξόρυξη δεδομένων όπως παρουσιάστηκε προηγουμένως, επιτρέπει στον αναλυτή να εστιάσει σε συγκεκριμένα πρότυπα και τάσεις και να τα εξερευνήσει σε βάθος χρησιμοποιώντας την οπτικοποίηση. Η οπτικοποίηση δεδομένων από μόνη της μπορεί να μπλοκάρει από τον όγκο των δεδομένων σε μια βάση δεδομένων αλλά με τη χρήση της εξόρυξης δεδομένων βοηθείται η εξερεύνηση (Pujari, 2001).

Η οπτικοποίηση δεδομένων βοηθά τους χρήστες να εξετάσουν μεγάλους όγκους δεδομένων και να ανιχνεύσουν οπτικά τα πρότυπα. Οι οπτικές απεικονίσεις δεδομένων όπως χάρτες, διαγράμματα και άλλες γραφικές αναπαραστάσεις δίνουν τη δυνατότητα της συμπαγούς παρουσίασης των δεδομένων στους χρήστες. Μια απλή γραφική οθόνη μπορεί να κωδικοποιήσει τόση πληροφορία όση μπορεί ένας μεγάλος αριθμός οθονών κειμένου (Pujari, 2001).

1.2.3. Εφαρμογές

Η εξόρυξη δεδομένων χρησιμοποιεί όπως είδαμε ένα σχετικά μεγάλο ποσό επιχειρησιακής υπολογιστικής ισχύος σε μία μεγάλη βάση δεδομένων για να προσδιορίσει κανονικότητες και συνδέσεις μεταξύ των σημείων δεδομένων. Οι αλγόριθμοι που χρησιμοποιούν τεχνικές από τη στατιστική, τη μηχανική μάθηση και την αναγνώριση προτύπων χρησιμοποιούνται για την αναζήτηση μεγάλων βάσεων δεδομένων αυτόματα. Όπως και ο όρος τεχνητή νοημοσύνη, η εξόρυξη δεδομένων είναι ένας γενικός όρος που μπορεί να εφαρμοστεί σε έναν αριθμό διαφορετικών δραστηριοτήτων. Στον κόσμο των επιχειρήσεων, η εξόρυξη δεδομένων χρησιμοποιείται πιο συχνά για να καθορίσει την κατεύθυνση των τάσεων και να προβλέψει το μέλλον. Χρησιμοποιείται για την κατασκευή μοντέλων και συστημάτων υποστήριξης αποφάσεων που δίνουν τις πληροφορίες που μπορούν να χρησιμοποιηθούν. Η εξόρυξη δεδομένων έχει ένα ρόλο πρώτης γραμμής στη μάχη κατά της τρομοκρατίας. Υποτίθεται ότι χρησιμοποιήθηκε για να καθοριστεί ο ηγέτης των επιθέσεων της 11/9 (Anissimov, 2011).

Οι ειδικοί εξόρυξης δεδομένων είναι στατιστικολόγοι που χρησιμοποιούν τεχνικές με ονόματα όπως μοντέλο πλησιέστερου γείτονα, συσταδοποίηση k-means, μέθοδος holdout, επικύρωση k-fold cross και ούτω καθεξής. Οι τεχνικές παλινδρόμησης που χρησιμοποιούνται αφαιρούν άσχετο μοτίβα, αφήνοντας μόνο χρήσιμες πληροφορίες. Ο όρος Bayesian που παρατηρείται συχνά στον κλάδο, αναφέρεται σε μια κατηγορία τεχνικών που προβλέπουν την πιθανότητα μελλοντικών γεγονότων με το συνδυασμό εκ των προτέρων πιθανοτήτων και των πιθανοτήτων που βασίζονται στα υπό όρους γεγονότα (Anissimov, 2011).

Το φιλτράρισμα ανεπιθύμητων μηνυμάτων είναι αναμφισβήτητα μια μορφή εξόρυξης δεδομένων. Τα δέντρα απόφασης χρησιμοποιούνται για το φιλτράρισμα μεγάλων ποσοτήτων δεδομένων. Σε ένα δέντρο αποφάσεων, όλα τα δεδομένα που περνάνε μέσα από ένα κόμβο εισόδου, βρίσκουν ένα φίλτρο που διαχωρίζει τα δεδομένα σε ροές, ανάλογα με τα χαρακτηριστικά τους. Για παράδειγμα, τα δεδομένα σχετικά με τη συμπεριφορά των καταναλωτών είναι πιθανό να φιλτράρονται με βάση δημογραφικούς παράγοντες (Anissimov, 2011).

Η εξόρυξη δεδομένων, δεν ασχολείται κατά κύριο λόγο με φανταχτερά γραφικά και τεχνικές απεικόνισης, αλλά είναι και αυτό τρόπος να παρουσιάσει ό, τι

έχει βρεθεί. Είναι γνωστό ότι ο άνθρωπος μπορεί να απορροφήσει περισσότερες στατιστικές πληροφορίες οπτικά παρά λεκτικά και αυτό το σχήμα για την παρουσίαση μπορεί να είναι πολύ πειστικό και ισχυρό εάν χρησιμοποιηθεί στο σωστό πλαίσιο. Καθώς ο πολιτισμός μας γίνεται όλο και πιο κορεσμένος από δεδομένα και οι αισθητήρες διανέμονται μαζικά σε τοπικά περιβάλλοντα, θα ανακαλυφθούν εκ παραδρομής τα πράγματα που έμειναν απαρατήρητα κατά το πρώτο πέρασμα. Η εξόρυξη δεδομένων θα επιτρέψει τη διόρθωση αυτών των λαθών και την ανακάλυψη νέων ιδεών που βασίζονται σε δεδομένα του παρελθόντος (Anissimov, 2011).

1.2.4. Προβλήματα και ζητήματα της εξόρυξης δεδομένων

Όταν χρησιμοποιείται στα τεχνικά πλαίσια αποθήκευσης δεδομένων και ανάλυσης, η εξόρυξη δεδομένων είναι ουδέτερη. Ωστόσο, μερικές φορές έχει μια πιο υποτιμητική χρήση που συνεπάγεται την επιβολή προτύπων (και ιδίως αιτιώδεις σχέσεις), σχετικά με τα δεδομένα εκεί που δεν υπάρχουν. Αυτή η επιβολή άσχετης αντιστοιχίας, παραπλανητικής αντιστοιχίας ή κοινότοπης αντιστοιχίας χαρακτηριστικού σωστά επικρίθηκε ως «εκβάθυνση δεδομένων» στη στατιστική βιβλιογραφία. Χρησιμοποιούμενα σε αυτήν την τελευταία έννοια, η εκβάθυνση δεδομένων βυθοκόρηση συνεπάγεται σάρωση των δεδομένων για κάθε σχέση που, στη συνέχεια, όταν βρεθεί να καταλήξει σε μια ενδιαφέρουσα εξήγηση. (Αυτό επίσης αναφέρεται ως «υπερβολικό ταίριασμα του δείγματος» (overfitting)) Το πρόβλημα είναι ότι τα μεγάλα σύνολα δεδομένων πάντοτε τυχαίνει να έχουν κάποιες συναρπαστικές σχέσεις σχετικές με αυτά τα δεδομένα. Επομένως, οποιαδήποτε συμπεράσματα είναι πιθανό να είναι πολύ ύποπτα. Παρόλα αυτά, απαιτούνται πάντα κάποιες διερευνητικές εργασίες δεδομένων σε κάθε εφαρμογή στατιστικής ανάλυσης για να παρθεί μια αίσθηση για τα δεδομένα, έτσι μερικές φορές η γραμμή μεταξύ καλής στατιστικής πρακτικής και εκβάθυνσης δεδομένων είναι λιγότερο από ξεκάθαρη (WordIQ.com, 2010).

Ένας πιο σημαντικός κίνδυνος είναι να βρεθούν συσχετίσεις που δεν υπάρχουν πραγματικά. Οι αναλυτές επενδύσεων φαίνεται να είναι ιδιαίτερα ευάλωτοι σε αυτό. Οι περισσότερες προσπάθειες εξόρυξης δεδομένων επικεντρώθηκαν στην ανάπτυξη ενός πολύ λεπτομερούς μοντέλου κάποιων μεγάλων συνόλων δεδομένων. Στο *Data Mining For Very Busy People* οι

ερευνητές στο West Virginia University και το University of British Columbia συζήτησαν μια εναλλακτική μέθοδο που περιλαμβάνει την εύρεση των ελάχιστων διαφορών μεταξύ των στοιχείων σε ένα σύνολο δεδομένων, με στόχο την ανάπτυξη απλούστερων μοντέλων που αντιπροσωπεύουν τα σχετικά δεδομένα (WordIQ.com, 2010).

Ένα από τα βασικά ζητήματα που προκύπτουν από την τεχνολογία εξόρυξης δεδομένων δεν είναι σχετικά με τις επιχειρήσεις ή τεχνολογικά, αλλά και κοινωνικά. Είναι το ζήτημα της ιδιωτικής ζωής του ατόμου. Η εξόρυξη δεδομένων καθιστά δυνατή την ανάλυση ρουτίνας επιχειρηματικών συναλλαγών και να βρει μια σημαντική ποσότητα πληροφοριών σχετικά με τα άτομα όσον αφορά τις αγοραστικές συνήθειες και προτιμήσεις (Palace, 1996). Επιπλέον, εάν για παράδειγμα ένας εργοδότης έχει πρόσβαση σε ιατρικά αρχεία, μπορεί να αποκλείσει άτομα με διαβήτη ή που έχουν υποστεί καρδιακή προσβολή. Ο αποκλεισμός των εν λόγω υπάλληλων θα μειώσει το κόστος για την ασφάλιση, αλλά δημιουργεί ηθικά και νομικά προβλήματα. Τα κυβερνητικά ή εμπορικά δεδομένα εξόρυξης για λόγους εθνικής ασφάλειας ή για σκοπούς επιβολής του νόμου, δημιουργούν επίσης ανησυχίες προστασία της ιδιωτικής ζωής (WordIQ.com, 2010).

Ένα άλλο ζήτημα είναι αυτό της ακεραιότητας των δεδομένων. Σαφώς, η ανάλυση των δεδομένων μπορεί να είναι μόνο τόσο καλή όσο τα δεδομένα που επεξεργάζεται. Μια βασική πρόκληση είναι η ενσωμάτωση της εφαρμογής με αντικρουόμενα ή πλεονάζοντα δεδομένα από διαφορετικές πηγές. Για παράδειγμα, μια τράπεζα μπορεί να διατηρεί την πίστωση των λογαριασμών καρτών σε πολλές διαφορετικές βάσεις δεδομένων. Οι διευθύνσεις (ή ακόμα και τα ονόματα) ενός κατόχου κάρτας μπορεί να είναι διαφορετικά σε κάθε κράτος. Το λογισμικό πρέπει να μεταφράσει τα δεδομένα από το ένα σύστημα στο άλλο και να επιλέξει τη διεύθυνση πιο αποθηκεύτηκε πρόσφατα (Palace, 1996).

Ένα πολυσυζητημένο τεχνικό ζήτημα είναι κατά πόσον είναι προτιμότερο να δημιουργηθεί μια σχεσιακή δομή βάσης δεδομένων ή μια πολυδιάστατη. Σε μια σχεσιακή δομή, τα δεδομένα αποθηκεύονται σε πίνακες, που επιτρέπουν τα ερωτήματα ad hoc. Σε μια πολυδιάστατη δομή, από την άλλη πλευρά, σύνολα κύβων είναι διατεταγμένα σε συστοιχίες, με υποσύνολα που δημιουργήθηκαν

σύμφωνα με την κατηγορία. Ενώ οι πολυδιάστατες δομές θα διευκολύνουν την πολυδιάστατη εξόρυξη δεδομένων, οι σχεσιακές δομές μέχρι σήμερα έχουν καλύτερη απόδοση σε περιβάλλοντα client/server. Ταυτόχρονα, με την έκρηξη του Διαδικτύου, ο κόσμος γίνεται ένα μεγάλο περιβάλλον client/server (Palace, 1996).

Τέλος, υπάρχει το θέμα του κόστους. Παρά το γεγονός ότι το κόστος του υλικού του συστήματος μειώθηκε δραματικά κατά την τελευταία πενταετία, η εξόρυξη δεδομένων και οι αποθήκες δεδομένων τείνουν να είναι αυτο-ενισχυόμενες. Όσο πιο μεγάλα είναι τα ερωτήματα εξόρυξης δεδομένων, τόσο μεγαλύτερη είναι η χρησιμότητα των πληροφοριών που βρίσκονται από τα δεδομένα και τόσο μεγαλύτερη είναι η πίεση για την αύξηση του όγκου των δεδομένων που συλλέγονται και διατηρούνται, γεγονός που αυξάνει την πίεση για ταχύτερα, ισχυρότερα ερωτήματα εξόρυξης δεδομένων. Αυτό αυξάνει την πίεση για τα μεγαλύτερα, πιο γρήγορα συστήματα, τα οποία είναι πιο ακριβά (Palace, 1996).

Υπάρχουν πολλές νόμιμες χρήσεις της εξόρυξης δεδομένων. Για παράδειγμα, μια βάση δεδομένων με τα συνταγογραφούμενα φάρμακα που λαμβάνονται από μια ομάδα ανθρώπων θα μπορούσε να χρησιμοποιηθεί για να βρείτε συνδυασμούς φαρμάκων με αρνητικές αντιδράσεις. Δεδομένου ότι ο συνδυασμός μπορεί να εμφανιστούν σε μόνο 1 στα 1000 άτομα, μια μεμονωμένη περίπτωση μπορεί να μην είναι εμφανής. Ένα σχέδιο στον τομέα των φαρμακείων θα μπορούσε να μειώσει τον αριθμό των παρενεργειών του φαρμάκου και, ενδεχομένως, να σώσει ζωές. Δυστυχώς, υπάρχει επίσης μια τεράστια δυνατότητα για κατάχρηση μιας τέτοιας βάσης δεδομένων. Βασικά, η εξόρυξη δεδομένων παρέχει πληροφορίες που δεν θα είναι διαθέσιμες με άλλο τρόπο. Θα πρέπει να ερμηνευτεί σωστά ώστε να είναι χρήσιμη. Όταν τα δεδομένα που συλλέχθηκαν περιλαμβάνουν πρόσωπα, υπάρχουν πολλά ερωτήματα σχετικά με προστασία της ιδιωτικής ζωής, της νομιμότητας, και της δεοντολογίας (WordIQ.com, 2010).

1.3. Στάδια

Η διαδικασία εξόρυξης δεδομένων αποτελείται από τρία στάδια: (1) της αρχικής έρευνας, (2) της δημιουργίας μοντέλου ή τον προσδιορισμό μοτίβου με την επικύρωση/επαλήθευση, και (3) την ανάπτυξη (δηλαδή, την εφαρμογή του

μοντέλου σε νέα δεδομένα, ώστε να δημιουργηθούν προβλέψεις) (StatSoft.com, Data Mining Techniques):

- **Στάδιο 1: Εξερεύνηση.** Το στάδιο αυτό ξεκινά συνήθως με την προετοιμασία των δεδομένων που μπορούν να αφορούν τον καθαρισμό των δεδομένων, τους μετασχηματισμούς δεδομένων, επιλέγοντας υποσύνολα των εγγραφών και - στην περίπτωση των συνόλων δεδομένων με μεγάλο αριθμό μεταβλητών («πεδίων») – την εκτέλεση ορισμένων προκαταρκτικών εργασιών επιλογής χαρακτηριστικών για να συμπληρωθεί ο αριθμός των μεταβλητών σε ένα διαχειρίσιμο φάσμα (ανάλογα με τις στατιστικές μεθόδους που εξετάζονται). Στη συνέχεια, ανάλογα με τη φύση του αναλυτικού προβλήματος, το πρώτο στάδιο της διαδικασίας εξόρυξης δεδομένων μπορεί να συνεπάγεται οτιδήποτε μεταξύ μιας απλής επιλογής ή απλής πρόβλεψης της πρόθεσης για ένα μοντέλο παλινδρόμησης, που θα επεξεργαστεί διερευνητικές αναλύσεις χρησιμοποιώντας μια ευρεία ποικιλία γραφικών και στατιστικών μεθόδων (π.χ. Διερευνητική Ανάλυση Δεδομένων (EDA)), προκειμένου να εντοπιστούν οι καταλληλότερες μεταβλητές και να προσδιοριστεί η πολυπλοκότητας και/ή ο γενικός χαρακτήρας των μοντέλων που μπορούν να ληφθούν υπόψη στην επόμενη φάση.

- Στάδιο 2: Δημιουργία μοντέλου και επικύρωση. Αυτό το στάδιο περιλαμβάνει την εξέταση των διαφόρων μοντέλων και την επιλογή της καλύτερης βάσης για την έξυπνη απόδοσή τους (δηλαδή, η εξήγηση της εν λόγω μεταβλητότητας την παραγωγή σταθερών αποτελεσμάτων σε όλα τα δείγματα). Αυτό μπορεί να ακούγεται σαν μια απλή λειτουργία, αλλά στην πραγματικότητα, αυτό σημαίνει μερικές φορές μια πολύ περίπλοκη διαδικασία. Υπάρχουν διάφορες τεχνικές που αναπτύχθηκαν για την επίτευξη αυτού του στόχου - πολλές από τις οποίες βασίζονται στη λεγόμενη «ανταγωνιστική αξιολόγηση των μοντέλων», δηλαδή, την εφαρμογή διάφορων μοντέλων για τη ρύθμιση των ίδιων δεδομένων και, στη συνέχεια, τη σύγκριση των επιδόσεων τους ώστε να επιλεγεί το καλύτερο. Οι τεχνικές αυτές - οι οποίες θεωρούνται συχνά ο πυρήνας της πρόβλεψης εξόρυξης δεδομένων - περιλαμβάνουν: Bagging (Λεπτομέρειες,

Μέσος Όρος), Ενίσχυση, Στοίβαξη (Στοιβαγμένες Γενικεύσεις), και Μετα-Μάθηση.

- Στάδιο 3: Ανάπτυξη. Αυτό το τελικό στάδιο περιλαμβάνει τη χρησιμοποίηση του μοντέλου που επιλέγεται ως το καλύτερο κατά το προηγούμενο στάδιο και η εφαρμογή των νέων δεδομένων, ώστε να δημιουργηθούν οι προβλέψεις ή εκτιμήσεις των αναμενόμενων αποτελεσμάτων.

Η έννοια της εξόρυξης δεδομένων γίνεται όλο και περισσότερο δημοφιλής ως εργαλείο ενημέρωσης για τη διαχείριση των επιχειρήσεων όπου αναμένεται να αποκαλύψει τις δομές της γνώσης που μπορούν να καθοδηγήσουν τις αποφάσεις και σε συνθήκες περιορισμένου δικαίου. Πρόσφατα, υπήρξε αύξηση του ενδιαφέροντος για την ανάπτυξη νέων αναλυτικών τεχνικών που έχουν σχεδιαστεί ειδικά για την αντιμετώπιση θεμάτων που αφορούν την επιχειρησιακή εξόρυξη δεδομένων (π.χ., Δέντρα Ταξινόμησης), αλλά η εξόρυξη δεδομένων εξακολουθεί να γίνεται με βάση τις εννοιολογικές αρχές της στατιστικής, συμπεριλαμβανομένων των παραδοσιακών Διερευνητική Ανάλυση Δεδομένων (ΕΟΑ) και τη μοντελοποίηση και μοιράζεται μαζί τους τόσο ορισμένα στοιχεία των γενικών προσεγγίσεων και ειδικές τεχνικές. Ωστόσο, μια σημαντική γενική διαφορά στην εστίαση και το σκοπό μεταξύ ης εξόρυξης δεδομένων και της παραδοσιακής Διερευνητικής Ανάλυσης Δεδομένων είναι ότι η εξόρυξη δεδομένων είναι περισσότερο προσανατολισμένη προς τις εφαρμογές παρά το βασικό χαρακτήρα των υποκείμενων φαινομένων. Με άλλα λόγια, η εξόρυξη δεδομένων ασχολείται σχετικά λιγότερο με τον εντοπισμό των ειδικών σχέσεων μεταξύ των εμπλεκόμενων μεταβλητών. Για παράδειγμα, η αποκάλυψη της φύσης των σχετικών λειτουργιών ή των συγκεκριμένων τύπων διαδραστικών, εξαρτήσεων πολλών μεταβλητών μεταξύ των μεταβλητών δεν είναι ο κύριος στόχος της εξόρυξης δεδομένων. Αντίθετα, η εστίαση είναι στην παραγωγή μιας λύσης που μπορεί να παράγει χρήσιμες προβλέψεις. Επομένως, η εξόρυξη δεδομένων δέχεται, μεταξύ άλλων, μια προσέγγιση «μαύρου κουτιού» στην εξερεύνηση δεδομένων ή γνώσης και χρησιμοποιεί όχι μόνο τις παραδοσιακές τεχνικές Διερευνητικής Ανάλυσης Δεδομένων αλλά και τεχνικές όπως νευρωνικά δίκτυα που μπορούν να δημιουργήσουν βάσιμες προβλέψεις, αλλά δεν είναι σε θέση να προσδιορίσουν τον ιδιαίτερο χαρακτήρα των αμοιβαίων σχέσεων μεταξύ των μεταβλητών στις οποίες στηρίζονται οι προβλέψεις. Η εξόρυξη δεδομένων

θεωρείται συχνά ως «ένα μείγμα από Στατιστική, Τεχνητή Νοημοσύνη και αναζήτηση σε βάσεις δεδομένων», η οποία μέχρι πολύ πρόσφατα δεν ήταν αναγνωρισμένη ευρέως ως ένα πεδίο ενδιαφέροντος για τους στατιστικολόγους, και μάλιστα θεωρείται από ορισμένους ως «μια βρώμικη λέξη στη Στατιστική». Λόγω της εφαρμοσμένης σημασίας του, ωστόσο, το πεδίο αναδύεται ως μια ταχύτατα αναπτυσσόμενη και μεγάλη περιοχή (επίσης στη Στατιστική), όπου γίνονται σημαντικές θεωρητικές εξελίξεις (StatSoft.com, Data Mining Techniques).

1.4. Τεχνικές Εξόρυξης Δεδομένων

Οι τεχνικές εξόρυξης δεδομένων είναι το αποτέλεσμα μιας μακράς διαδικασίας έρευνας και ανάπτυξης προϊόντων. Η εξέλιξη αυτή άρχισε όταν τα επιχειρηματικά δεδομένα για πρώτη φορά αποθηκεύτηκαν σε ηλεκτρονικούς υπολογιστές, συνεχίστηκε με βελτιώσεις στην πρόσβαση στα δεδομένα, και πιο πρόσφατα, στις τεχνολογίες παραγωγής που επιτρέπουν στους χρήστες να πλοηγούνται με τα στοιχεία τους σε πραγματικό χρόνο. Η εξόρυξη δεδομένων οδηγεί αυτήν την εξελικτική διαδικασία πέραν της αναδρομικής πρόσβασης στα δεδομένα και την πλοήγηση στην μελλοντική και προληπτική παροχή πληροφοριών. Η εξόρυξη δεδομένων εφαρμόζεται στην επιχειρηματική κοινότητα, διότι υποστηρίζεται από τρεις τεχνολογίες που είναι τώρα αρκετά ώριμες (Thearling, 2010):

- Μαζική συλλογή δεδομένων
- Ισχυροί πολυεπεξεργαστές υπολογιστών
- Αλγόριθμοι εξόρυξης δεδομένων

Οι εμπορικές βάσεις δεδομένων αυξάνονται με πρωτοφανείς ρυθμούς. Οι αλγόριθμοι εξόρυξης δεδομένων ενσωματώνουν τεχνικές που υπάρχουν εδώ και αρκετά χρόνια, αλλά μόνο πρόσφατα έχουν εφαρμοστεί ως ώριμα, αξιόπιστα, κατανοητά εργαλεία που ξεπερνούν σταθερά τις μεγαλύτερες στατιστικές μεθόδους. Στην εξέλιξη από τα δεδομένα των επιχειρήσεων στην πληροφόρηση των επιχειρήσεων, κάθε νέο βήμα έχει οικοδομηθεί με βάση το προηγούμενο. Για παράδειγμα, η δυναμική πρόσβαση στα δεδομένα είναι ζωτικής σημασίας για την αναζήτηση δεδομένων σε εφαρμογές πλοήγησης, καθώς και η δυνατότητα

αποθήκευσης μεγάλων βάσεων δεδομένων είναι κρίσιμη για την εξόρυξη δεδομένων. Από την πλευρά του χρήστη, τα τέσσερα βήματα που αναφέρονται στον Πίνακα 1 ήταν επαναστατικά, επειδή επέτρεψαν νέα ερωτήματα για την επιχείρηση να απαντηθούν με ακρίβεια και ταχύτητα (Thearling, 2010)..

Πίνακας 1: Εξελικτικά βήματα που οδήγησαν στην εξόρυξη δεδομένων (Thearling, 2010).

Εξελικτικό Βήμα	**Επιχειρηματική Ερώτηση**	**Τεχνολογίες**	**Παροχείς Προϊόντων**	**Χαρακτηριστικά**
Συλλογή Δεδομένων (1960)	«Ποιό ήταν το συνολικό εισόδημα τα τελευταία 5 χρόνια;»	Υπολογιστές, κασέτες, δίσκοι	IBM, CDC	Αναδρομική, στατική παράδοση δεδομένων
Πρόσβαση Στα Δεδομένα (1980)	«Ποιες ήταν οι μοναδιαίες πωλήσεις στην Νέα Αγγλία τον τελευταίο Μάρτιο;»	Σχεσιακές βάσεις δεδομένων (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Αναδρομική, δυναμική παράδοση δεδομένων σε επίπεδο εγγραφής
Αποθήκες Δεδομένων & Υποστήριξη Αποφάσεων (1990)	«Ποιες ήταν οι μοναδιαίες πωλήσεις στην Νέα Αγγλία τον τελευταίο Μάρτιο; Αναζήτηση στη Βοστώνη»	On-line αναλυτική επεξεργασία (OLAP), πολυδιάστατες βάσεις δεδομένων, αποθήκες δεδομένων	Pilot, Comshare, Arbor, Cognos, Microstrategy	Αναδρομική, δυναμική παράδοση δεδομένων σε πολλαπλά επίπεδα
Εξόρυξη Δεδομένων (Αναδυόμενη Σήμερα)	«Τι είναι πιθανό να συμβεί στις μοναδιαίες πωλήσεις της Βοστώνης τον	Εξελιγμένοι αλγόριθμοι, πολυεπεξεργαστικοί υπολογιστές, μαζικές βάσεις	Pilot, Lockheed, IBM, SGI, διάφορες νέες επιχειρήσεις	Αναδρομική, προορατική παράδοση πληροφοριών

	επόμενο μήνα;»	δεδομένων		

Τα βασικά συστατικά τεχνολογίας εξόρυξης δεδομένων βρίσκονται υπό ανάπτυξη εδώ και δεκαετίες, σε τομείς έρευνας όπως οι στατιστική, η τεχνητή νοημοσύνη και η μηχανική μάθησης. Σήμερα, η ωριμότητα των τεχνικών αυτών, σε συνδυασμό με υψηλής απόδοσης σχεσιακές βάσεις δεδομένων και ευρείες προσπάθειες ολοκλήρωσης δεδομένων, έχουν καταστήσει αυτές τις τεχνολογίες πρακτικές για τα τρέχοντα περιβάλλοντα αποθηκών δεδομένων (Thearling, 2010).

Οι **αλγόριθμοι εξόρυξης δεδομένων (data mining algorithms)** είναι προγραμματισμένα ερωτήματα και προγράμματα που χρησιμοποιούνται για την αναγνώριση προτύπων και τάσεις σε σύνολα δεδομένων. Η κύρια χρήση της εξόρυξης δεδομένων είναι όπως είδαμε να καθορίσει τις ανάγκες και τις προτιμήσεις των πελατών, με βάση την πραγματική δραστηριότητά τους. Αν και οι πληροφορίες βασίζονται σε επιδόσεις κατά το παρελθόν, μπορεί να είναι ένας άριστος δείκτης της συμπεριφοράς των πελατών και των τάσεων (Francois).

Στη συνέχεια θα γίνει μια περιγραφή των βασικών κατηγοριών τεχνικών εξόρυξης δεδομένων και αλγορίθμων εξόρυξης δεδομένων που θα περιγραφούν αναλυτικά στα επόμενα κεφάλαια.

1.4.1. Κλασικές Τεχνικές: Στατιστική, Γειτνίαση και Συσταδοποίηση

1.4.1.1. Στατιστική

Με τον αυστηρό ορισμό η **Στατιστική (Statistics)** ή στατιστική τεχνική δεν είναι εξόρυξη δεδομένων. Ο όρος αυτός είχε χρησιμοποιηθεί πολύ πριν επινοηθεί ο όρος εξόρυξη δεδομένων για να εφαρμόζεται στις εφαρμογές των επιχειρήσεων. Ωστόσο, οι στατιστικές τεχνικές οδηγούνται από τα δεδομένα και χρησιμοποιούνται για να ανακαλύψουν πρότυπα και να κατασκευαστούν μοντέλα πρόβλεψης. Από τη σκοπιά των χρηστών, υπάρχει μια συνειδητή επιλογή όταν επιλύεται ένα πρόβλημα «εξόρυξης δεδομένων» ως προς το αν θέλει να χρησιμοποιήσει στατιστικές μεθόδους ή άλλες τεχνικές εξόρυξης δεδομένων. Για το λόγο αυτό,

είναι σημαντικό να υπάρχει κάποια ιδέα για το πώς λειτουργούν οι στατιστικές τεχνικές εργασιών και πώς μπορούν να εφαρμοστούν (Berson et al., 2010).

Οι τεχνικές που χρησιμοποιούνται στην εξόρυξη δεδομένων, όταν ολοκληρωθούν με επιτυχία, επιτυγχάνουν για τους ίδιους λόγους για τους οποίους οι στατιστικές τεχνικές γίνονται δεκτές (π.χ. καθαρά στοιχεία, καλά καθορισμένος στόχος πρόβλεψης και καλή επικύρωση). Ως επί το πλείστον είναι οι τεχνικές που χρησιμοποιούνται στις ίδιες θέσεις για τους ίδιους τύπους προβλημάτων (πρόβλεψη, ανακάλυψη ταξινόμησης). Στην πραγματικότητα, μερικές από τις τεχνικές που κλασικά ορίζονται ως «εξόρυξη δεδομένων», όπως οι CART και CHAID, προέκυψαν από στατιστικολόγους. Υπάρχουν αρκετοί λόγοι όμως για τους οποίους οι τεχνικές αυτές δεν επαρκούν. Ο πρώτος είναι ότι οι κλασικές τεχνικές εξόρυξης δεδομένων, όπως η CART και οι τεχνικές πλησιέστερου γείτονα τείνουν να είναι πιο ανθεκτικές σε μαζικά πραγματικά δεδομένα, αλλά και πιο ισχυρές κατά τη χρήση τους από λιγότερο έμπειρους χρήστες. Αλλά αυτός δεν είναι ο μόνος λόγος. Ο άλλος λόγος είναι ότι ο χρόνος είναι σωστός, λόγω της χρήσης των ηλεκτρονικών υπολογιστών για κλειστού βρόχου επιχειρηματικά δεδομένα αποθήκευσης και παραγωγής, υπάρχουν πλέον μεγάλες ποσότητες δεδομένων διαθέσιμες στους χρήστες. Εάν δεν υπήρχαν δεδομένα δεν θα υπήρχε ενδιαφέρον για την εξόρυξη τους. Ομοίως, το γεγονός ότι το υλικό του υπολογιστή έχει αναβαθμιστεί κατά αρκετές τάξεις μεγέθους στην περιοχή της αποθήκευσης και επεξεργασίας δεδομένων, κάνει μερικές από τις πιο ισχυρές τεχνικές εξόρυξης δεδομένων εφικτές σήμερα. Η ουσία όμως, από ακαδημαϊκή άποψη, τουλάχιστον, είναι ότι υπάρχει ελάχιστη πρακτική διαφορά ανάμεσα σε μια στατιστική τεχνική και την κλασική τεχνική εξόρυξης δεδομένων (Berson et al., 2010).

1.4.1.2. Πλησιέστερος γείτονας

Η συσταδοποίηση που περιγράφεται στην επόμενη υποενότητα και το πρόβλεψη **Πλησιέστερου Γείτονα (Nearest Neighbor)** είναι από τις παλαιότερες τεχνικές που χρησιμοποιούνται στην εξόρυξη δεδομένων. Οι περισσότεροι άνθρωποι έχουν μια διαίσθηση ότι καταλαβαίνουν τι είναι η ομαδοποίηση - δηλαδή ότι είναι η ενέργεια κατά την οποία οι εγγραφές ομαδοποιούνται ή συγκεντρώνονται. Ο πλησιέστερος γείτονας είναι μια τεχνική πρόβλεψη που είναι αρκετά παρόμοια με την συσταδοποίηση - η ουσία της είναι ότι για να προβλεφθεί

ποια είναι μια τιμή πρόβλεψης είναι σε μια εγγραφή γίνεται αναζήτηση για εγγραφές με παρόμοιες τιμές πρόβλεψης στην βάση δεδομένων ιστορικού και χρησιμοποιείται η τιμή πρόβλεψης από την εγγραφή που είναι «πλησιέστερη» στην αταξινόμητη εγγραφή (Berson et al., 2010).

1.4.1.3. Συσταδοποίηδη

Συσταδοποίηση (clustering) είναι η μέθοδος με την οποία οι εγγραφές ομαδοποιούνται. Συνήθως αυτό γίνεται για να δώσει στον τελικό χρήστη μια υψηλού επιπέδου όψη του τι συμβαίνει στη βάση δεδομένων. Η συσταδοποίηση μερικές φορές χρησιμοποιείται με την έννοια της κατάτμησης (Berson et al., 2010).

1.4.2. Τεχνικές Νέας Γενιάς: Δένδρα και Κανόνες

Οι τεχνικές εξόρυξης δεδομένων σε αυτή την ενότητα αντιπροσωπεύουν τις πιο συχνά χρησιμοποιούμενες τεχνικές που έχουν αναπτυχθεί τις τελευταίες δύο δεκαετίες έρευνας. Αντιπροσωπεύουν επίσης τη συντριπτική πλειοψηφία των τεχνικών που αναφέρονται στο δημοφιλή Τύπο. Αυτές οι τεχνικές μπορούν να χρησιμοποιηθούν είτε για την ανακάλυψη νέων στοιχείων μέσα σε μεγάλες βάσεις δεδομένων ή για την κατασκευή μοντέλων πρόβλεψης. Αν και οι παλαιότερες τεχνικές δέντρου αποφάσεων, όπως η CHAID, χρησιμοποιούνται σήμερα ευρέως. η χρήση τν νέων τεχνικών, όπως η CART κερδίζει ευρύτερη αποδοχή (Berson et al., 2010).

1.4.2.1. Δένδρα Αποφάσεων

Ένα **δέντρο απόφασης (decision tree)** είναι ένα μοντέλο πρόβλεψης που, όπως υποδηλώνει το όνομά του, μπορεί να θεωρηθεί ως ένα δέντρο. Συγκεκριμένα κάθε κλάδος του δέντρου είναι ένα ερώτημα ταξινόμησης και τα φύλλα του δέντρου είναι τα διαμερίσματα του συνόλου δεδομένων με την ταξινόμηση τους (Berson et al., 2010).

1.4.2.2. Επαγωγή κανόνων

Η **επαγωγή κανόνων (rule induction)** είναι μια από τις σημαντικότερες μορφές εξόρυξης δεδομένων και ίσως η πιο κοινή μορφή ανακάλυψης γνώσης στην συστήματα μη επιβλεπόμενης μάθησης. Επίσης, είναι ίσως η μορφή της εξόρυξης δεδομένων που μοιάζει περισσότερο με τη διαδικασία με την οποία σκέφτονται οι περισσότεροι άνθρωποι όταν αναφέρονται στην εξόρυξη

δεδομένων, δηλαδή η «εξόρυξη» για το χρυσό μέσα σε μια τεράστια βάση δεδομένων. Ο χρυσός σε αυτή την περίπτωση είναι ένας κανόνας που έχει ενδιαφέρον, που λέει κάτι για τη βάση δεδομένων που δεν είναι ήδη γνωστό και πιθανόν να μην ήταν προφανές (Berson et al., 2010).

1.4.3. Επιλογή Τεχνικής

Δεν υπάρχει συγκεκριμένος κανόνας που υπαγορεύει πότε πρέπει να επιλεγεί μια ιδιαίτερη τεχνική σε αντίθεση με κάποια άλλη. Μερικές φορές αυτές οι αποφάσεις γίνονται σχετικά αυθαίρετα με βάση τη διαθεσιμότητα των αναλυτών εξόρυξης δεδομένων που είναι πιο έμπειροι σε μια τεχνική αντί κάποιας άλλης. Ακόμα και η επιλογή κάποιας κλασικής τεχνικής αντί για ορισμένες από τις νεότερες τεχνικές εξαρτάται από τη διαθεσιμότητα καλών εργαλείων και καλών αναλυτών. Είναι σαφές ότι ένα από τα δυσκολότερα πράγματα που γίνονται όταν αποφασίζεται η εφαρμογή ενός συστήματος εξόρυξης δεδομένων είναι να καθοριστεί ποια τεχνική θα χρησιμοποιηθεί πότε. Πότε χρησιμεύουν τα νευρωνικά δίκτυα και πότε τα δέντρα απόφασης; Πότε η εξόρυξη δεδομένων ενδείκνυται σε αντιδιαστολή με την απλή εργασία με σχεσιακές βάσεις δεδομένων και εκθέσεις; Πότε χρησιμοποιείται μόνο OLAP και πότε μια πολυδιάστατη βάση δεδομένων είναι κατάλληλη (Berson et al., 2010);

Ορισμένα από τα κριτήρια που είναι σημαντικά για τον καθορισμό της τεχνικής που θα χρησιμοποιηθεί καθορίζονται από τη μέθοδο δοκιμής και σφάλματος. Υπάρχουν σαφείς διαφορές στα είδη των προβλημάτων στα οποία είναι πλέον κατάλληλη η κάθε τεχνική, αλλά η πραγματικότητα των δεδομένων του πραγματικού κόσμου και ο δυναμικός τρόπος με τον οποίο οι αγορές, οι πελάτες και ως εκ τούτου, τα δεδομένα που τους εκπροσωπούν διαμορφώνονται σημαίνει ότι τα δεδομένα αλλάζουν συνεχώς. Η δυναμική αυτή σημαίνει ότι δεν υπάρχει πλέον νόημα να δημιουργηθεί το «τέλειο» μοντέλο για τα ιστορικά δεδομένα εφόσον ό,τι ήταν γνωστό κατά το παρελθόν δεν μπορεί να προβλέψει ικανοποιητικά το μέλλον, γιατί το μέλλον βρίσκεται σε αντίθεση με ό,τι έχει γίνει στο παρελθόν (Berson et al., 2010).

Κατά κάποιο τρόπο αυτή η κατάσταση είναι ανάλογη με το επιχειρηματικό πρόσωπο που περιμένει να λάβει όλες τις πληροφορίες πριν λάβει την απόφασή του. Προσπαθεί να χρησιμοποιήσει διάφορα σενάρια, διαφορετικούς τύπους και

έρευνα για νέες πηγές πληροφοριών. Αλλά αυτό είναι ένα έργο που δεν πρόκειται ποτέ να ολοκληρωθεί, τουλάχιστον εν μέρει, επειδή η επιχείρηση, η οικονομία, ακόμη και ο κόσμος αλλάζει με απρόβλεπτους και ακόμη και χαοτικούς τρόπους που δεν θα μπορούσαν ποτέ να προβλεφθούν ικανοποιητικά. Είναι καλύτερο να λάβουν ένα σταθερό μοντέλο που ίσως είναι υποδεέστερο σε σύγκριση με αυτό που μερικά από τα καλύτερα εργαλεία εξόρυξης δεδομένων θα μπορούσαν να προσφέρουν με μεγάλη ανάλυση και να το εκτελέσουν σήμερα και όχι αύριο, όταν μπορεί να είναι πολύ αργά (Berson et al., 2010).

1.5. Δεδομένα και κριτήρια επιλογής αλγορίθμων

Οι τύποι δεδομένων στην εξόρυξη δεδομένων είναι (Microsoft | Technet, 2008): Κείμενο (Text), Μακρύς (Long), Boolean, Διπλός (Double), Ημερομηνία (Date).

Η παρακάτω λίστα περιγράφει τους τύπους περιεχομένου που χρησιμοποιούνται στην εξόρυξη δεδομένων, και προσδιορίζει τους τύπους δεδομένων που υποστηρίζουν κάθε τύπο (Microsoft MSDN, Content Types (Data Mining), 2011).

Διακριτός (Discrete): Διακριτός σημαίνει ότι η στήλη περιέχει έναν πεπερασμένο αριθμό τιμών χωρίς συνέχεια μεταξύ των τιμών. Οι τιμές σε μια στήλη διακριτών χαρακτηριστικών δεν μπορεί να συνεπάγονται ταξινόμηση, ακόμα κι αν οι τιμές είναι αριθμητικές. Επιπλέον, ακόμη και αν οι τιμές που χρησιμοποιούνται για τη διακριτή στήλη είναι αριθμητικές, δεν μπορούν να υπολογιστούν κλασματικές τιμές. Ο τύπος περιεχομένου Discrete υποστηρίζεται από όλους τους τύπους δεδομένων εξόρυξης δεδομένων.

Συνεχής (Continuous): Συνεχής σημαίνει ότι η στήλη περιέχει τιμές που αντιστοιχούν σε αριθμητικά δεδομένα σε κλίμακα που επιτρέπουν τη λήψη ενδιάμεσων τιμών. Σε αντίθεση με μια διακριτή στήλη, που αντιστοιχεί σε πεπερασμένα, μετρήσιμα δεδομένα, η συνεχής στήλη αντιπροσωπεύει επεκτάσιμες μετρήσεις και είναι δυνατό για τα δεδομένα να περιέχουν έναν άπειρο αριθμό κλασματικών τιμών. Όταν μια στήλη περιέχει συνεχή αριθμητικά δεδομένα, και είναι γνωστή η κατανομή των δεδομένων, γίνεται να βελτιωθεί η ακρίβεια των

αναλύσεων, με την ένδειξη της αναμενόμενης κατανομής των τιμών. Κατά συνέπεια, η ρύθμιση ισχύει για όλα τα μοντέλα που βασίζονται στη δομή, Ο συνεχής τύπος περιεχομένου υποστηρίζεται από τους τύπους δεδομένων: Date, Double και Long.

Διακριτοποιημένος (Discretized): Διακριτοποίηση είναι η διαδικασία ανάθεσης τιμών ενός συνεχούς συνόλου δεδομένων σε κάδους, ώστε να υπάρχει ένας περιορισμένος αριθμός πιθανών τιμών. Μόνο αριθμητικά δεδομένα μπορούν να διακριτοποιηθούν. Έτσι, ο τύπος περιεχομένου discretized δείχνει ότι η στήλη περιέχει αξίες που αντιπροσωπεύουν ομάδες, ή κουβάδες, των αξιών που προέρχονται από μια συνεχή στήλη. Οι κάδοι αντιμετωπίζονται ως ταξινομημένοι και ως διακριτές τιμές. Ο τύπος περιεχομένου discretized υποστηρίζονται από τους τύπους δεδομένων: Date, Double, Long και Text.

Κλειδί (Key): Ο τύπος περιεχομένου κλειδιού σημαίνει ότι η στήλη προσδιορίζει μοναδικά μια σειρά. Σε έναν πίνακα περιπτώσεων, συνήθως η στήλη κλειδιών είναι αριθμητική ή αναγνωριστικό κείμενο. Ο τύπος περιεχομένου κλειδιού ορίζεται για να δείξει ότι η στήλη δεν θα πρέπει να χρησιμοποιείται για την ανάλυση, μόνο για την παρακολούθηση των εγγραφών. Οι ένθετοι πίνακες έχουν επίσης κλειδιά, αλλά η χρήση του κλειδιού ένθετου πίνακα είναι λίγο διαφορετική. Οι τιμές στο ένθετο κλειδί πίνακα πρέπει να είναι μοναδικές για κάθε υπόθεση, αλλά μπορεί να υπάρχουν διπλότυπα σε ολόκληρο το σύνολο των περιπτώσεων. Αυτός ο τύπος περιεχομένου υποστηρίζεται από τους τύπους δεδομένων: Date, Double, Long και Text.

Ακολουθία Κλειδιού (Key Sequence): Ο τύπος περιεχομένου ακολουθίας κλειδιού μπορεί να χρησιμοποιηθεί μόνο σε μοντέλα ακολουθίας συσταδοποίησης. Όταν ρυθμίζεται ο τύπος περιεχομένου ως ακολουθία κλειδιού, δείχνει ότι η στήλη περιέχει τις τιμές που αντιπροσωπεύουν μια ακολουθία γεγονότων. Οι τιμές είναι διατεταγμένες, αλλά δεν πρέπει να είναι σε ίση απόσταση μεταξύ τους. Αυτός ο τύπος περιεχομένου υποστηρίζεται από τους τύπους δεδομένων: Double, Long, Text και Date.

Χρόνος Κλειδιού (Key Time): Ο τύπος περιεχομένου χρόνου κλειδιού μπορεί να χρησιμοποιηθεί μόνο σε μοντέλα χρονικών σειρών. Όταν ρυθμίζεται ο τύπος περιεχομένου ως χρόνο κλειδιού, δείχνει ότι οι τιμές ταξινομούνται και

αντιπροσωπεύουν μια χρονική κλίμακα. Αυτός ο τύπος περιεχομένου που υποστηρίζεται από τους τύπους δεδομένων: Double, Long και Date.

Πίνακας (Table): Ο τύπος περιεχομένου πίνακα δείχνει ότι η στήλη περιέχει δεδομένα ενός άλλου πίνακα, με μία ή περισσότερες στήλες και μία ή περισσότερες σειρές. Για οποιαδήποτε ιδιαίτερη γραμμή στον πίνακα περίπτωσης, αυτή η στήλη μπορεί να περιέχει πολλαπλές τιμές, όλες σχετικές με την αρχική εγγραφή. Ο τύπος δεδομένων της στήλης αυτής είναι πάντα Table.

Κυκλικός (Cyclical): Ο κυκλικός τύπο περιεχομένου σημαίνει ότι η στήλη περιέχει τιμές που αντιπροσωπεύουν ένα κυκλικά διατεταγμένο σύνολο. Οι κυκλικές στήλες θεωρούνται τόσο διατεταγμένες όσο και διακριτές από πλευράς τύπου περιεχομένου. Αυτός ο τύπος περιεχομένου υποστηρίζεται από όλους τους τύπους δεδομένων εξόρυξης δεδομένων σε Υπηρεσίες ανάλυσης. Ωστόσο, οι περισσότεροι αλγόριθμοι αντιμετωπίζουν τις κυκλικές τιμές ως διακριτές τιμές και δεν επιτελούν ειδική επεξεργασία.

Διατεταγμένος (Ordered): Ο διατεταγμένος τύπος περιεχομένου υποδεικνύει επίσης ότι η στήλη περιέχει τιμές που ορίζουν μια ακολουθία ή εντολή. Ωστόσο, σε αυτό τον τύπο περιεχομένου οι τιμές που χρησιμοποιούνται για την παραγγελία δεν συνεπάγονται καμία σχέση απόστασης ή μεγέθους του συνόλου. Οι στήλες διατεταγμένων χαρακτηριστικών θεωρούνται ότι είναι διακριτές από την άποψη του τύπου περιεχομένου. Αυτός ο τύπος περιεχομένου υποστηρίζεται από όλους τους τύπους δεδομένων εξόρυξης δεδομένων σε Υπηρεσίες ανάλυσης. Ωστόσο, οι περισσότεροι αλγόριθμοι αντιμετωπίζουν τις διατεταγμένες τιμές ως διακριτές τιμές και δεν επιτελούν ειδική επεξεργασία.

Ταξινομημένος (Classified): Εκτός από τους προηγούμενους τύπους περιεχομένου που είναι σε κοινή χρήση για όλα τα μοντέλα, για ορισμένους τύπους δεδομένων μπορούν να χρησιμοποιηθούν διατεταγμένες στήλες για τον καθορισμό του τύπου περιεχομένου.

Το πρώτο βήμα για τη δημιουργία ενός παραγωγικού προγράμματος εξόρυξης δεδομένων είναι η συλλογή δεδομένων. Το κλειδί εδώ είναι ο εντοπισμός των κρίσιμων στοιχείων και την τοποθέτηση τους σε μια αποθήκη δεδομένων. Το επόμενο βήμα είναι η επιλογή ενός ή περισσοτέρων αλγορίθμων εξόρυξης

δεδομένων για το πρόβλημα. Στην αρχή, είναι πιθανώς μια καλή ιδέα ο πειραματισμός με διάφορες τεχνικές. Η επιλογή του αλγορίθμου θα εξαρτηθεί από τα δεδομένα που έχουν συγκεντρωθεί, το πρόβλημα που προσπαθεί να λύσει και τα εργαλεία πληροφορικής που είναι διαθέσιμα (Chapple, 2011).

Η επιλογή του καλύτερου αλγορίθμου που θα χρησιμοποιηθεί για μια συγκεκριμένη εργασία μπορεί να είναι μια πρόκληση. Ενώ μπορούν να χρησιμοποιηθούν διαφορετικοί αλγόριθμοι για την ίδια εργασία, κάθε αλγόριθμος παράγει ένα διαφορετικό αποτέλεσμα, και μερικοί αλγόριθμοι μπορούν να παράγουν περισσότερους από έναν τύπους αποτελέσματος. Επίσης, δεν χρειάζεται να χρησιμοποιηθούν οι αλγόριθμοι ανεξάρτητα. Σε μια ενιαία λύση εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν κάποιοι αλγόριθμοι για να εξερευνήσουν τα δεδομένα, και στη συνέχεια να χρησιμοποιηθούν άλλοι αλγόριθμοι για να προβλέψουν ένα συγκεκριμένο αποτέλεσμα με βάση αυτά τα δεδομένα. Για παράδειγμα, μπορεί να χρησιμοποιηθεί ένας αλγόριθμος, ο οποίος αναγνωρίζει τα πρότυπα, για να σπάσει τα δεδομένα σε ομάδες που είναι περισσότερο ή λιγότερο ομοιογενείς και, στη συνέχεια να χρησιμοποιήσει τα αποτελέσματα για να δημιουργήσει ένα καλύτερο μοντέλο δέντρου απόφασης. Μπορούν επίσης να χρησιμοποιηθούν πολλοί αλγόριθμοι μέσα σε μία λύση για την εκτέλεση ξεχωριστών εργασιών, για παράδειγμα με τη χρήση ενός αλγορίθμου δέντρου οπισθοδρόμησης για την απόκτηση χρηματοοικονομικών πληροφοριών πρόβλεψης, καθώς και ένα αλγόριθμο βασισμένο σε κανόνες για να εκτελέσει μια ανάλυση αγοράς. τα μοντέλα εξόρυξης μπορούν να προβλέψουν τις τιμές, να παράγουν περιλήψεις των δεδομένων και να βρουν κρυμμένες συσχετίσεις (Microsoft MSDN, Data Mining Algorithms (Analysis Services - Data Mining), 2011).

Στη συνέχεια αναφέρονται ενδεικτικά ορισμένα παραδείγματα χρήσης αλγορίθμων ανάλογα με το είδος των δεδομένων και περισσότερες πληροφορίες θα παρέχονται στα επόμενα κεφάλαια κατά την περιγραφή των αλγορίθμων.

Η παλινδρόμηση είναι η παλαιότερη και πιο γνωστή στατιστική τεχνική που χρησιμοποιεί η κοινότητα εξόρυξης δεδομένων. Ο κυριότερος περιορισμός αυτής της τεχνικής είναι ότι λειτουργεί μόνο καλά με συνεχή ποσοτικά δεδομένα (όπως βάρος, ταχύτητα ή ηλικία). Εάν επεξεργάζεται κατηγορικά δεδομένων, που να μην

είναι σημαντικά (όπως το χρώμα, το όνομα ή το φύλο), καλύτερα να χρησιμοποιηθεί κάποια άλλη τεχνική. Στην εργασία με κατηγορικά δεδομένα ή μείγμα συνεχών αριθμητικών και κατηγορικών δεδομένων η ταξινόμηση είναι καλή (Chapple, 2011). Ο αλγόριθμος ID3 που είναι αλγόριθμος δένδρου απόφασης είναι καλός για στιγμιότυπα που έχουν τη μορφή ζεύγους χαρακτηριστικού-τιμής και η συνάρτηση στόχος έχει διακριτές τιμές εξόδου (Verma et al.). Η επαγωγή κανόνων τυπικά χρησιμοποιείται σε βάσεις δεδομένων με πεδία είτε με πολλές διαφορετικές τιμές ή πολλές στήλες με δυαδικά πεδία (Berson et al., 2010).

Οι ακραίες τιμές (outliers) είναι σπάνια, ασυνήθιστα, ή απλά σπάνια γεγονότα που παρουσιάζουν ενδιαφέρον για την εξόρυξη δεδομένων σε πολλές περιπτώσεις, συμπεριλαμβανομένων της απάτης στον φόρο εισοδήματος, της ασφάλισης και των ηλεκτρονικών τραπεζικών συναλλαγών, καθώς και για το εμπόριο. Οι αναλύσεις που επικεντρώνονται στην ανακάλυψη αυτών των αντικειμένων δεδομένων ως αναλύσεις ακραίας τιμής. Η έννοια της ακραίας τιμής είναι (Williams, Outlier Analysis, 2010):

«παρατήρηση που αποκλίνει τόσο πολύ από τις άλλες παρατηρήσεις ώστε προκαλεί υποψίες ότι δημιουργήθηκε από έναν διαφορετικό μηχανισμό».

Οι αλγόριθμοι ανίχνευσης ακραίων τιμών συχνά εμπίπτουν σε μία από τις κατηγορίες μεθόδων εξ αποστάσεως, μεθόδων πυκνότητας, μεθόδων προβολής και μεθόδων κατανομής (Williams, Outlier Analysis, 2010).

1.6. Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκε η έννοια της εξόρυξης δεδομένων και τα βασικά χαρακτηριστικά της, οι εφαρμογές και οι χρήσεις της, τα διάφορα ζητήματα που προκύπτουν, τις βασικές κατηγορίες τεχνικών που χρησιμοποιούνται από τους αλγορίθμους της και τους τύπους δεδομένων που καθορίζουν ποιος αλγόριθμος θα επιλεγεί. Στα υπόλοιπα κεφάλαια θα γίνει μια πιο αναλυτική περιγραφή των τεχνικών εξόρυξης δεδομένων και των αλγορίθμων, ξεκινώντας με το κεφάλαιο 2 που περιλαμβάνει τις βασικές στατιστικές μεθόδους που είναι η κατάταξη και η παλινδρόμηση.

2. ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ (STATISTICS)

2.1. Εισαγωγή

Στο κεφάλαιο αυτό περιγράφονται οι στατιστικές μέθοδοι οι οποίες όπως περιγράφηκε σε προηγούμενο κεφάλαιο αποτελούν τις πρώτες τεχνικές που χρησιμοποιήθηκαν στην εξόρυξη δεδομένων και είναι οι μέθοδοι της κατάταξης και της παλινδρόμησης.

2.2. Κατάταξη (Classification)

Ως **κατάταξη (classification)** ορίζεται η οργάνωση διαφόρων αντικειμένων σε αμοιβαία αποκλειόμενες αλλά σχετιζόμενες τάξεις (BusinessDictionary.com, classification). Στην ενότητα αυτή θα περιγραφούν οι βασικότεροι αλγόριθμοι κατάταξης που χρησιμοποιούνται στην εξόρυξη δεδομένων.

2.2.1. Naive Bayes

Λαμβάνοντας υπόψη ένα σύνολο αντικειμένων, καθένα από τα οποία ανήκει σε μια γνωστή κατηγορία, και καθένα από τα οποία έχει ένα γνωστό διάνυσμα μεταβλητών, στόχος είναι να κατασκευαστεί ένας κανόνας που θα επιτρέψει την ανάθεση μελλοντικών αντικειμένων σε μια κατηγορία, δοθέντων μόνο των διανυσμάτων των μεταβλητών που περιγράφουν τα αντικείμενα. Προβλήματα του είδους αυτού, που ονομάζεται προβλήματα επιβλεπόμενης ταξινόμησης, είναι πανταχού παρόντα και έχουν αναπτυχθεί πολλές μέθοδοι για την κατασκευή αυτών των κανόνων. Μια πολύ σημαντική τεχνική είναι η μέθοδος **αφελής Bayes (naive Bayes)** που αποκαλείται επίσης simple Bayes και independence Bayes. Η μέθοδος αυτή είναι σημαντική για πολλούς λόγους. Είναι πολύ εύκολο να κατασκευαστεί, δεν χρειάζεται κανένα περίπλοκο επαναληπτικό

σύστημα υπολογισμού των παραμέτρων. Αυτό σημαίνει ότι μπορεί εύκολα να εφαρμοστεί σε τεράστια σύνολα δεδομένων. Είναι εύκολο να ερμηνευθεί, έτσι ώστε οι χρήστες που είναι μη εκπαιδευμένοι στην τεχνολογία κατάταξης να μπορούν να κατανοήσουν την ταξινόμηση που κάνει. Τέλος, όλο αυτό γίνεται συχνά εκπληκτικά καλά: μπορεί να μην είναι ο καλύτερος δυνατός ταξινομητής σε κάθε συγκεκριμένη εφαρμογή, αλλά μπορεί συνήθως να είναι βάσιμος για να είναι ουσιαστικός και λειτουργεί αρκετά καλά (Wu et al., 2008).

Στη συγκεκριμένη περίπτωση για απλούστερη εξήγηση θεωρούνται ότι υπάρχουν μόνο δύο κατηγορίες, οι οποίες συμβολίζονται με i=0,1. Στόχος είναι να χρησιμοποιηθεί το αρχικό σύνολο αντικειμένων με γνωστές τις κατηγορίες στις οποίες ανήκουν (το σύνολο εκπαίδευσης) και να κατασκευαστεί ένα αποτέλεσμα τέτοιο ώστε τα μεγαλύτερα αποτελέσματα να συνδέονται με της τάξης 1 αντικείμενα (για παράδειγμα) και τα μικρότερα αποτελέσματα με της τάξης 0 αντικείμενα. Η κατάταξη στη συνέχεια επιτυγχάνεται με τη σύγκριση αυτού του σκορ με ένα όριο, t. Εάν οριστεί το P(i|x) ότι είναι η πιθανότητα ότι ένα αντικείμενο με το διάνυσμα μέτρησης $x=(x_1,...,x_p)$ ανήκει στην κατηγορία i, τότε κάθε μονότονη συνάρτηση του P(i|x) θα κάνει κατάλληλη βαθμολογία. Ειδικότερα, ο δείκτης P(1|x)/P(0|x) θα ήταν κατάλληλος. Η στοιχειώδης πιθανότητα λέει ότι υπάρχει η δυνατότητα αποσύνθεσης του P (i|x), ανάλογο με την f(x|θ)P(θ), όπου f (x|i) είναι η δεσμευμένη κατανομή του x για την τάξη αντικειμένων i, και P(i) είναι η πιθανότητα ότι ένα αντικείμενο θα ανήκει στην κατηγορία i, αν δεν είναι γνωστό τίποτα περισσότερο γι 'αυτό (η «πριν» πιθανότητα της κατηγορίας i). Αυτό σημαίνει ότι ο λόγος γίνεται (Wu et al., 2008):

$$\frac{P(1|x)}{P(0|x)} = \frac{f(x|1)P(1)}{f(x|0)P(0)} \qquad (2.1)$$

Για να χρησιμοποιηθούν αυτά για την παραγωγή κατατάξεων, πρέπει να εκτιμηθεί η f(x|θ) και η P(i). Εάν το σύνολο κατάρτισης ήταν ένα τυχαίο δείγμα από τον συνολικό πληθυσμό, το P(i) να μπορεί να εκτιμηθεί άμεσα από το ποσοστό των αντικειμένων κατηγορίας i στο σύνολο εκπαίδευσης. Για την εκτίμηση της f(x|θ), η μέθοδος naive Bayes προϋποθέτει ότι τα στοιχεία του x είναι ανεξάρτητα, $f(x|i) = \prod_{j=1}^{p} f(x_j|i)$ και στη συνέχεια εκτιμά καθεμιάς από τις κατανομές απλής μεταβλητής f (x_j|i), j=(1,...,p),;i=0,1, χωριστά Έτσι το p διαστάσεων πρόβλημα

πολλών μεταβλητών έχει μειωθεί σε p προβλήματα εκτίμησης απλής μεταβλητής. Η μονοπαραγοντική εκτίμηση είναι γνωστή, απλή και απαιτεί μικρότερα σύνολα εκπαίδευσης για την απόκτηση ακριβών εκτιμήσεων Αυτό είναι ένα από τα ιδιαίτερα, και μάλιστα μοναδικά χαρακτηριστικά της μεθόδου naive Bayes: η εκτίμηση είναι απλή, πολύ γρήγορη και δεν απαιτεί πολύπλοκα συστήματα επαναληπτικής εκτίμησης (Wu et al., 2008).

Αν η οριακές κατανομές $f(x_j|i)$ είναι διακριτές, με κάθε x_j να λαμβάνει λίγες μόνο τιμές, τότε η εκτίμηση $f(x_j|i)$ είναι εκτιμητής ιστογράμματος πολυωνυμικού τύπου, μετρώντας απλώς το ποσοστό των αντικειμένων κατηγορίας i τα οποία εμπίπτουν σε κάθε κελί. Αν οι $f(x_j|i)$ είναι συνεχείς, τότε μια κοινή στρατηγική είναι η τμηματοποίηση καθεμιάς σε ένα μικρό αριθμό διαστημάτων και πάλι η χρήση πολυωνυμικών εκτιμητών, αλλά πιο περίτεχνες εκδόσεις βασίζονται σε συνεχείς εκτιμήσεις (π.χ. εκτιμήσεις πυρήνα) (Wu et al., 2008).

Δοθείσας της υπόθεσης ανεξαρτησίας, η αναλογία γίνεται:

$$\frac{P(1|x)}{P(0|x)} = \frac{\prod_{j=1}^{p} f(x_j|1)P(1)}{\prod_{j=1}^{p} f(x_j|0)P(0)} = \frac{P(1)}{P(0)} \prod_{j=1}^{p} \frac{f(x_j|1)}{f(x_j|0)} \quad (2.2)$$

Τώρα, ανακαλώντας ότι ο στόχος ήταν απλά η παραγωγή ενός σκορ μονοτονικά σχετιζόμενου με την $P(i|x)$, η παραπάνω σχέση λογαριθμίζεται (η log είναι μονοτονικά αυξάνουσα συνάρτηση). Αυτό δίνει ένα εναλλακτικό σκορ:

$$\ln \frac{P(1|x)}{P(0|x)} = \ln \frac{P(1)}{P(0)} + \ln \sum_{j=1}^{p} \frac{f(x_j|1)}{f(x_j|0)} \quad (2.3)$$

Αν οριστεί $w_j=\ln(f(x_j|1)/f(x_j|0))$ και μια σταθερά $k=\ln(P(1)/P(0))$ η προηγούμενη σχέση παίρνει τη μορφή ενός απλού αθροίσματος, άρα ο ταξινομητής έχει μια απλή δομή:

$$\ln \frac{P(1|x)}{P(0|x)} = k + \sum_{j=1}^{p} w_j \quad (2.4)$$

Η υπόθεση ανεξαρτησίας του x_j σε κάθε κατηγορία υπονοεί ότι το μοντέλο naive Bayes μπορεί να φαίνεται υπερβολικά περιοριστικό. Στην πραγματικότητα,

εντούτοις, διάφοροι παράγοντες μπορούν να έρθουν στο προσκήνιο το οποίο σημαίνει ότι η υπόθεση δεν είναι τόσο επιζήμια όσο φαίνεται. Πρώτον, έχει συμβεί πολλές φορές μια προηγούμενη φάση επιλογής μεταβλητής, στην οποία οι υψηλού βαθμού συσχέτισης μεταβλητές έχουν εξαλειφθεί με την αιτιολογία ότι είναι πιθανό να συμβάλουν με παρόμοιο τρόπο στο διαχωρισμό μεταξύ των τάξεων. Αυτό σημαίνει ότι οι σχέσεις μεταξύ των υπόλοιπων μεταβλητών θα μπορούσαν κάλλιστα να προσεγγιστούν από την ανεξαρτησία. Δεύτερον, αν υποτεθεί ότι η αλληλεπίδραση είναι μηδέν, προβλέπει σιωπηρό βήμα κανονικοποίησης, με αποτέλεσμα να μειωθεί η διακύμανση του μοντέλου και οδηγεί σε πιο ακριβείς ταξινομήσεις. Τρίτον, σε ορισμένες περιπτώσεις, όταν οι μεταβλητές συσχετίζονται, η βέλτιστη επιφάνεια απόφασης συμπίπτει με εκείνα που παράγονται με την παραδοχή της ανεξαρτησίας, έτσι η υπόθεση δεν είναι καθόλου επιζήμια για την απόδοση. Τέταρτον, φυσικά, η επιφάνεια απόφασης που παράγεται από το μοντέλο naive Bayes μπορεί στην πραγματικότητα να έχει μια πολύπλοκη γραμμική μορφή: η επιφάνεια είναι γραμμική στο w_j αλλά εξαιρετικά γραμμική στις αρχικές μεταβλητές x_j, έτσι ώστε να μπορούν να χωρέσουν αρκετά ανεπτυγμένες επιφάνειες (Wu et al., 2008).

Παράδειγμα

Ένα παράδειγμα που εμφανίζεται στην βιβλιογραφία για την κατανόηση των αλγορίθμων κατηγοριοποίησης, είναι αυτό της πρόβλεψης του «καλού καιρού για τένις». Τα δεδομένα εκπαίδευσης που θα χρησιμοποιήσουμε για αυτό το παράδειγμα παρουσιάζονται στον πίνακα 6. Ας δούμε σε ποια κατηγορία (ναι / όχι) θα κατηγοριοποιηθεί η πλειάδα <Ηλιόλουστος, Μικρή, Υψηλή, Ισχυρός>

Πίνακας 10: Δεδομένα εκπαίδευσης παραδείγματος

Ημέρα	Καιρός	Θερμοκρασία	Υγρασία	Άνεμος	Τένις
1	Ηλιόλουστος	Μεγάλη	Υψηλή	Αδύναμος	Όχι
2	Ηλιόλουστος	Μεγάλη	Υψηλή	Ισχυρός	Όχι
3	Συννεφιασμένος	Μεγάλη	Υψηλή	Αδύναμος	Ναι
4	Βροχερός	Μεσαία	Υψηλή	Αδύναμος	Ναι
5	Βροχερός	Μικρή	Κανονική	Αδύναμος	Ναι
6	Βροχερός	Μικρή	Κανονική	Ισχυρός	Όχι
7	Συννεφιασμένος	Μικρή	Κανονική	Ισχυρός	Ναι
8	Ηλιόλουστος	Μεσαία	Υψηλή	Αδύναμος	Όχι
9	Ηλιόλουστος	Μικρή	Κανονική	Αδύναμος	Ναι

10	Βροχερός	Μεσαία	Κανονική	Αδύναμος	Ναι

Καιρός: P(Ηλιόλουστος | όχι) = 3/5, P(Ηλιόλουστος | ναι) = 2/9,
P(Συννεφιασμένος | όχι) = 0/5, P(Συννεφιασμένος| ναι)=4/9,
P(Βροχερός | όχι) = 2/5, P(Βροχερός | ναι) = 3/9

Θερμοκρασία: P(Μεγάλη | όχι) = 2/5, P(Μεγάλη | ναι) = 2/9,
P(Μεσαία | όχι) = 2/5, P(Μεσαία | ναι) = 4/9,
P(Μικρή | όχι) = 1/5, P(Μικρή | ναι) = 3/9

Υγρασία: P(Υψηλή | όχι) = 4/5, P(Υψηλή | ναι) = 3/9 ,
P(Κανονική | όχι) = 1/5, P(Κανονική | ναι) = 6/9

Άνεμος: P(Αδύναμος | όχι) = 3/5 P(Αδύναμος | ναι) = 7/9
P(Ισχυρός | όχι) = 2/5 P(Ισχυρός | ναι) = 2/9

Κατηγορία ΟΧΙ:

P(όχι) * P(Ηλιόλουστος | όχι) * P(Μικρή | όχι) * P(Υψηλή | όχι) * P(Ισχυρός | όχι) = 5/14 * 3/5 * 1/5 * 4/5 * 2/5 =120/8750

Κατηγορία ΝΑΙ:

P(ναι) * P(Ηλιόλουστος | ναι) * P(Μικρή | ναι) * P(Υψηλή | ναι) * P(Ισχυρός | ναι) = 9/14 * 2/9 * 3/9 * 3/9 * 2/9 = 324/91854

Άρα κατηγοριοποιούμε την νέα πλειάδα στην κατηγορία όχι

2.2.2. Μηχανές Διανυσματικής Υποστήριξης

Στις σημερινές εφαρμογές μηχανικής μάθησης, οι μηχανές διανυσματικής υποστήριξης (support vector machines, SVM) θεωρούνται πολύ σημαντικές και προσφέρουν μία από τις πιο ισχυρές και ακριβείς μεθόδους μεταξύ όλων των γνωστών αλγορίθμων. Έχουν καλή θεωρητική θεμελίωση, απαιτούν μόνο μια δωδεκάδα παραδείγματα για την εκπαίδευση και επηρεάζονται από τον αριθμό των διαστάσεων. Επιπλέον, οι αποτελεσματικές μέθοδοι για την εκπαίδευση των SVM αναπτύσσονται με ταχείς ρυθμούς (Wu et al., 2008).

Σε μια εργασία μάθησης δύο κατηγοριών, ο στόχος της SVM είναι να βρει την καλύτερη λειτουργία ταξινόμησης για τη διάκριση μεταξύ των μελών των δύο κατηγοριών στα δεδομένα εκπαίδευσης. Η μετρική της έννοιας της «καλύτερης» συνάρτησης κατάταξης μπορεί να πραγματοποιηθεί γεωμετρικά. Για γραμμικά διαχωρίσιμα σύνολα δεδομένων, μια γραμμική συνάρτηση κατάταξης αντιστοιχεί σε ένα διαχωριστικό υπερεπίπεδο f(x) που διέρχεται από το μέσον των δύο τάξεων, χωρίζοντας τις. Όταν καθοριστεί αυτή η συνάρτηση, το νέο στιγμιότυπο δεδομένων x_n μπορεί να ταξινομηθεί με την απλή δοκιμή του πρόσημου της συνάρτησης f (x_n), το x_n ανήκει στην κατηγορία θετικό αν f (x_n)> 0 (Wu et al., 2008).

Επειδή υπάρχουν πολλά τέτοια γραμμικά υπερεπίπεδα, η SVM επιπλέον εγγυάται ότι η καλύτερη αυτή συνάρτηση βρίσκεται από τη μεγιστοποίηση του περιθωρίου μεταξύ των δύο κατηγοριών. Διαισθητικά, το περιθώριο ορίζεται ως η ποσότητα του χώρου, ή ο διαχωρισμός μεταξύ των δύο κλάσεων, όπως ορίζεται από το υπερεπίπεδο. Γεωμετρικά, το περιθώριο αντιστοιχεί στην ελάχιστη απόσταση μεταξύ των πλησιέστερων σημείων δεδομένων σε ένα σημείο του υπερεπιπέδου. Έχοντας αυτό τον γεωμετρικό ορισμό υπάρχει η δυνατότητα της εξερεύνησης πώς να μεγιστοποιηθεί το περιθώριο, έτσι ώστε ακόμα κι αν υπάρχει ένας άπειρος αριθμός υπερεπιπέδων, λίγες μόνο να χαρακτηρίζονται ως η λύση για την SVM (Wu et al., 2008).

Ο λόγος για τον οποίο η SVM επιμένει στην εύρεση των μέγιστων οριακών υπερεπιπέδων είναι ότι προσφέρει την καλύτερη δυνατότητα γενίκευσης. Επιτρέπει όχι μόνο την καλύτερη απόδοση κατάταξης (π.χ., την ακρίβεια) στα δεδομένα εκπαίδευσης, αλλά και αφήνει μεγάλα περιθώρια για την ορθή κατάταξη

των μελλοντικών στοιχείων. Για να εξασφαλιστεί ότι τα μέγιστα οριακά υπερεπίπεδα έχουν βρεθεί, ένας ταξινομητής SVM επιχειρεί να μεγιστοποιήσει την ακόλουθη συνάρτηση σε σχέση με τα $\vec{w}$ και b:

$$L_p = \frac{1}{2}\left\|\vec{w}\right\| - \sum_{i=1}^{t} a_i y_i \left(\vec{w} \bullet \overrightarrow{x_i} + \vec{b}\right) + \sum_{i=1}^{t} a_i \qquad (2.5)$$

όπου t είναι ο αριθμός των παραδειγμάτων εκπαίδευσης, και α_i, i=1,...,t, είναι μη-αρνητικοί αριθμοί έτσι ώστε τα παράγωγα του L_P σε σχέση με α_i είναι μηδέν. Τα α_i είναι οι πολλαπλασιαστές Lagrange και το L_P ονομάζεται Lagrangian. Σε αυτή την εξίσωση, τα διανύσματα $\vec{w}$ και η σταθερά b καθορίζουν το υπερεπίπεδο (Wu et al., 2008).

2.3. Παλινδρόμηση (Regression)

Ένα μοντέλο **παλινδρόμησης (regression)** προβλέπει την αξία ενός αριθμητικού πεδίου δεδομένων, που είναι το πεδίο-στόχος, σε μια δεδομένη εγγραφή δεδομένων από τις γνωστές τιμές των άλλων πεδίων δεδομένων της ίδιας εγγραφής. Οι γνωστές τιμές των άλλων πεδίων δεδομένων ονομάζονται πεδία δεδομένων εισόδου ή επεξηγηματικά πεδίων δεδομένων. Μπορούν να είναι αριθμητικά ή κατηγορικά. Η προβλεπόμενη τιμή, δεν μπορεί να ταυτίζεται με οποιαδήποτε τιμή περιέχονται στα δεδομένα που χρησιμοποιούνται για την κατασκευή του μοντέλου. Ο αλγόριθμος παλινδρόμησης εξόρυξης μπορεί να χρησιμοποιηθεί για τη δημιουργία μοντέλων παλινδρόμησης. Μπορεί επίσης να ελέγξει αυτά τα μοντέλα. Ένα μοντέλο παλινδρόμησης δημιουργείται και εκπαιδεύεται με βάση γνωστά σύνολα δεδομένων εγγραφών δεδομένων των οποίων οι τιμές των πεδίων στόχων είναι γνωστές. Το εκπαιδευμένο μοντέλο μπορεί να εφαρμοστεί σε γνωστά ή άγνωστα δεδομένα. Σε άγνωστα δεδομένα, είναι γνωστές οι τιμές των πεδίων εισόδου, ωστόσο, η τιμή του πεδίου στόχου δεν είναι γνωστή (IBM, Regression, 2011).

Η ανάλυση παλινδρόμησης αποσκοπεί να καθορίσει τις τιμές των παραμέτρων για μια λειτουργία που κάνει τη συνάρτηση να ταιριάζει καλύτερα σε ένα σύνολο παρατηρήσεων δεδομένων. Η ακόλουθη εξίσωση εκφράζει αυτές τις σχέσεις με σύμβολα. Δείχνει ότι η παλινδρόμηση είναι η διαδικασία του

υπολογισμού της αξίας του συνεχούς στόχου (y) ως μια συνάρτηση (F) ενός ή περισσοτέρων προβλέψεων ($x_1,x_2,...,x_n$), μια σειρά από παραμέτρους ($\theta_1,\theta_2,...,\theta_n$), και ένα μέτρο του σφάλματος (e) (Oracle, 2011):

$$y = F\left(x,\theta\right)+e \tag{2.6}$$

Οι προβλέψεις μπορούν να νοηθούν ως ανεξάρτητες μεταβλητές και ο στόχο τους ως εξαρτημένη μεταβλητή. Το σφάλμα, που ονομάζεται επίσης το υπόλοιπο, είναι η διαφορά μεταξύ των αναμενόμενων και προβλεπόμενων τιμών της εξαρτημένης μεταβλητής. Οι παράμετροι παλινδρόμησης είναι επίσης γνωστοί και ως συντελεστές παλινδρόμησης. Η διαδικασία εκπαίδευσης ενός μοντέλου παλινδρόμησης περιλαμβάνει την εξεύρεση των τιμών των παραμέτρων που ελαχιστοποιούν το μέτρο του σφάλματος, για παράδειγμα, το άθροισμα των τετραγώνων των σφαλμάτων. Υπάρχουν διάφορες οικογένειες συναρτήσεων παλινδρόμησης και διαφορετικοί τρόποι μέτρησης του σφάλματος (Oracle, 2011):

- Γραμμική Παλινδρόμηση (Linear Regression): μια γραμμική τεχνική παλινδρόμησης μπορεί να χρησιμοποιηθεί εάν η σχέση μεταξύ της πρόβλεψης και του στόχου μπορεί να προσεγγιστεί με μια ευθεία γραμμή.

- Πολυπαραγοντική Γραμμική Παλινδρόμηση (Multivariate Linear Regression): Ο όρος πολυπαραγοντική γραμμική παλινδρόμηση αναφέρεται σε γραμμική παλινδρόμηση με δύο ή περισσότερες προβλέψεις ($x_1,x_2,...,x_n$). Όταν χρησιμοποιούνται πολλαπλές πρόβλεψης, η γραμμή παλινδρόμησης δεν μπορεί να απεικονιστεί σε δισδιάστατο χώρο. Ωστόσο, η γραμμή μπορεί να υπολογιστεί απλά με την επέκταση της εξίσωσης για την γραμμική παλινδρόμηση απλής πρόγνωσης ώστε να συμπεριλάβει τις παραμέτρους για κάθε πρόβλεψη.

- Μη Γραμμική Παλινδρόμηση (Nonlinear Regression): Συχνά η σχέση μεταξύ των x και y δεν μπορεί να προσεγγιστεί με μια ευθεία γραμμή. Σε αυτή την περίπτωση, μια μη γραμμική τεχνική παλινδρόμησης μπορεί να χρησιμοποιηθεί. Εναλλακτικά, τα δεδομένα θα μπορούσαν να είναι προεπεξεργασμένα για να κάνουν τη σχέση γραμμική.

- Πολυπαραγοντική Μη Γραμμική Παλινδρόμηση (Multivariate Nonlinear Regression): Ο όρος πολυπαραγοντική μη γραμμική παλινδρόμηση αναφέρεται σε γραμμική παλινδρόμηση με δύο ή περισσότερες προβλέψεις (x_1,x_2,...,x_n). Όταν χρησιμοποιούνται πολλαπλές πρόβλεψεις, η γραμμική σχέση δεν μπορεί να απεικονιστεί σε δισδιάστατο χώρο.

2.4. Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκαν οι βασικές μέθοδοι στατιστικής που χρησιμοποιούνται στην εξόρυξη δεδομένων. Στη συνέχεια παρουσιάζονται οι τεχνικές τμηματοποίησης.

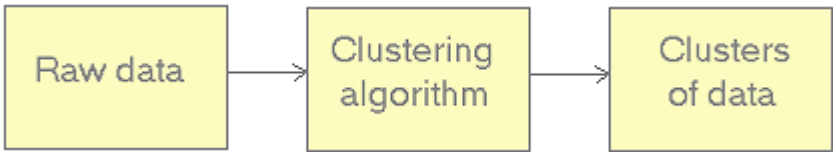
3. ΤΜΗΜΑΤΟΠΟΙΗΣΗ (SEGMENTATION)

3.1. Εισαγωγή

Στο κεφάλαιο αυτό περιγράφονται οι πιο συνηθισμένες τεχνικές τμηματοποίησης στην εξόρυξη δεδομένων και οι αλγόριθμοι που τις χρησιμοποιούν. Γίνεται αναφορά στη συσταδοποίηση, στη δημογραφική συσταδοποίηση, στους αυξητικούς αλγορίθμους και στη μέθοδο του k πλησιέστερου γείτονα.

3.2. Συσταδοποίηση (Clustering)

Η **συσταδοποίηση (clustering)** είναι ένα εργαλείο για την ανάλυση δεδομένων, το οποίο λύνει προβλήματα ταξινόμησης. Σκοπός του είναι να διανείμει περιπτώσεις (ανθρώπους, αντικείμενα, γεγονότα κ.λπ.) σε ομάδες, έτσι ώστε ο βαθμός της ενώσεως να είναι ισχυρός μεταξύ των μελών της ίδιας ομάδας και αδύναμος μεταξύ μελών διαφορετικών ομάδων. Με τον τρόπο αυτό κάθε ομάδα περιγράφει, όσον αφορά τα δεδομένα που συλλέγονται, την κλάση στην οποία ανήκουν τα μέλη της. Η συσταδοποίηση είναι εργαλείο ανακάλυψης. Είναι δυνατόν να αποκαλύψει τις ενώσεις και τη δομή των δεδομένων που, αν και προηγουμένως δεν προέκυπταν, παρόλα αυτά είναι λογικά και χρήσιμα όταν που βρέθηκαν. Τα αποτελέσματα της ανάλυσης συστάδων μπορούν να συμβάλουν στον καθορισμό ενός επίσημου συστήματος ταξινόμησης, όπως είναι η ταξινομία για σχετικά ζώα, έντομα ή φυτά, ή να προτείνουν στατιστικά μοντέλα για την περιγραφή του πληθυσμού, ή να υποδείξουν τους κανόνες για την ανάθεση νέων περιπτώσεων στις κατηγορίες για αναγνώριση και διαγνωστικούς σκοπούς, ή να παρέχουν τα μέτρα του ορισμού, το μέγεθος και την αλλαγή σε ό, τι στο παρελθόν ήταν μόνο γενικές έννοιες, ή να βρουν υποδείγματα για να αντιπροσωπεύσουν τις κατηγορίες (B&M Services). Η διαδικασία συσταδοποίησης παρουσιάζεται στο Σχήμα 1.



Σχήμα 1: Η διαδικασία συσταδοποίησης (B&M Services).

3.2.1. k-μέσων (k-means)

Ο αλγόριθμος **k-μέσων (k-means)** είναι μια απλή επαναληπτική μέθοδος για τη δημιουργία διαμερισμάτων σε έναν δεδομένο σύνολο δεδομένων σε σε καθορισμένο από το χρήστη πλήθος ομάδων, k. Ο αλγόριθμος λειτουργεί σε μία σειρά διανυσμάτων διάστασης d, D = {x_i|i=1,...,N}, όπου $x_i \in \Re^d$ σημαίνει το i-οστό σημείο δεδομένων. Ο αλγόριθμος ξεκινά επιλέγοντας k σημεία στο $\Re^d$ ως τους αρχικούς k αντιπροσώπους συστάδων ή «centroids». Οι τεχνικές για την επιλογή

αυτών των αρχικών φυτρών περιλαμβάνουν τυχαία δειγματοληψία από το σύνολο των δεδομένων, τον καθορισμό τους ως τη λύση της συσταδοποίησης σε ένα μικρό υποσύνολο των δεδομένων ή δημιουργώντας διαταραχές στην καθολική μέση τιμή των δεδομένων k φορές. Στη συνέχεια, ο αλγόριθμος επαναλαμβάνεται μεταξύ των δύο βημάτων μέχρι τη σύγκλιση (Wu et al., 2008):

- Βήμα 1: Εκχώρηση Δεδομένων (Data Assignment). Κάθε σημείο δεδομένων έχει ανατεθεί στο πλησιέστερο centroid του, με δεσμούς αυθαίρετα σπασμένους. Αυτό οδηγεί σε διαμερισμό των δεδομένων.

- Βήμα 2: Μεταφορά των «μέσων» (Relocation of “means”). Κάθε αντιπρόσωπος συστάδας μεταφέρεται προς το κέντρο (μέσος όρος) όλων των σημείων δεδομένων που του έχουν ανατεθεί. Εάν τα σημεία δεδομένων έρχονται με μέτρο πιθανότητας (βάρη), τότε η μετεγκατάσταση γίνεται προς τις αναμενόμενες τιμές (σταθμισμένος μέσος όρος) των διαμερισμάτων στοιχείων.

Ο αλγόριθμος συγκλίνει όταν οι αναθέσεις (και συνεπώς οι τιμές c_j) δεν αλλάζουν πλέον. Κάθε επανάληψη χρειάζεται N × k συγκρίσεις, οι οποίες προσδιορίζουν την χρονική πολυπλοκότητα μιας επανάληψης. Ο αριθμός των επαναλήψεων που απαιτούνται για τη σύγκλιση ποικίλλει και μπορεί να εξαρτηθεί από το N, αλλά ως πρώτη περικοπή, ο αλγόριθμος μπορεί να θεωρηθεί γραμμικός στο μέγεθος του συνόλου δεδομένων (Wu et al., 2008).

3.2.2. EM

Τα πεπερασμένα μεικτά χρησιμοποιούνται όλο και περισσότερο για να μοντελοποιήσουν τις κατανομές σε μεγάλη ποικιλία τυχαίων φαινομένων και να συγκεντρώσουν σύνολα δεδομένων. Τα μεικτά αυτά μοντέλα μπορούν να προσαρμοστούν ως προς τη μέγιστη πιθανότητα μέσω του αλγορίθμου EM (Expectation–Maximization) (Wu et al., 2008).

Έστω το p-διάστατο διάνυσμα ($y=(y_1,...,y_p)^T$) περιέχει τις τιμές των μεταβλητών p και μετράται σε κάθε μία από τις n (ανεξάρτητες) οντότητες προς συσταδοποίηση και έστω y_j η τιμή του y που αντιστοιχεί στην j-οστή οντότητα

(j=1,...,n). Με την μεικτή προσέγγιση στη συσταδοποίηση, τα $y_1,...,y_n$ υποτίθεται ότι είναι ένα παρατηρηθέν τυχαίο δείγμα από μείγμα ενός πεπερασμένου αριθμού, ας πούμε g, των ομάδων σε κάποιες άγνωστες αναλογίες $\pi_1,...,\pi_g$. Η πυκνότητα μείγματος y_j εκφράζεται ως:

$$f\left(y_i;\psi\right)=\sum_{i=1}^{g}\pi_i\, f_i\left(y_j;\theta_i\right)\left(j=1,...,n\right) \tag{3.1}$$

όπου οι αναλογίες μείξης $\pi_1,...,\pi_g$ αθροίζονται σε ένα και η υπό όρους πυκνότητα της ομάδας $f_i(y_j;\theta_i)$ καθορίζεται μέχρι ένα διάνυσμα θ_i αγνώστων παραμέτρων (i=1,...,g). Το διάνυσμα όλων των αγνώστων παραμέτρων δίνεται από:

$$\psi=\left(\pi_1,...,\pi_{g-1},\theta_1^T,...,\theta_g^T\right)^T \tag{3.2}$$

όπου ο εκθέτης T συμβολίζει αντιστροφή διανύσματος. Χρησιμοποιώντας μια εκτίμηση του ψ, η προσέγγιση αυτή δίνει μια πιθανοτική συσταδοποίηση των δεδομένων σε g συστάδες υπό όρους εκτιμήσεων των εκ των υστέρων πιθανοτήτων των συστατικών μελών,

$$\tau_i\left(y_j,\psi\right)=\frac{\pi_i\, f_i\left(y_j;\theta_i\right)}{f\left(y_j;\psi\right)} \tag{3.3}$$

όπου $\tau_i(y_j;\psi)$ είναι η εκ των υστέρων πιθανότητα ότι το yj (στην πραγματικότητα η οντότητα με παρατήρηση y_j) ανήκει στο i-οστό συστατικό του μείγματος (i=1,...,g; j=1,...,n). Το διάνυσμα παραμέτρου ψ μπορεί να υπολογιστεί με μέγιστη πιθανότητα. Η εκτίμηση μέγιστης πιθανότητας (maximum likelihood estimate, MLE) του ψ, δίνεται από μια κατάλληλη ρίζα της εξίσωσης πιθανότητας,

$$\partial \log L\left(\psi\right)/\partial\psi=0 \tag{3.4}$$

όπου

$$\log L\left(\psi\right)=\sum_{j=1}^{n}\log\, f\left(y_j;\psi\right) \tag{3.5}$$

είναι η συνάρτηση λογαριθμικής πιθανότητας του ψ. Οι λύσεις της παραπάνω συνάρτησης ανταποκρίνονται σε τοπικούς μεγιστοποιητές και μπορούν να υπολογιστούν με τον αλγόριθμο EM (Wu et al., 2008).

Για τη μοντελοποίηση συνεχών δεδομένων, οι υπό συνθήκη πυκνότητες συστατικών συνήθως υπολογίζονται να ανήκουν στην ίδια οικογένεια παραμέτρων, για παράδειγμα, την κανονική. Στην περίπτωση αυτή,

$$f_i\left(y_j;\theta_i\right)=\phi\left(y_j;\mu_i,\Sigma_i\right) \tag{3.6}$$

η p-διάστατη κανονική κατανομή πολλών μεταβλητών (Wu et al., 2008).

3.2.3. DBSCAN

Ο αλγόριθμος DBSCAN βασίζεται σε μια πυκνότητα που βασίζεται στην έννοια των συστάδων. Οι συστάδες προσδιορίζονται από την εξέταση της πυκνότητας των σημείων. Περιοχές με υψηλή πυκνότητα σημείων απεικονίζουν την ύπαρξη συστάδων ενώ περιοχές με χαμηλή πυκνότητα σημείων δείχνουν συστάδες θορύβου ή συστάδες ακραίων τιμών. Ο αλγόριθμος αυτός είναι ιδιαίτερα κατάλληλος για να ασχοληθεί με μεγάλα σύνολα δεδομένων, με θόρυβο, και είναι σε θέση να προσδιορίσει συστάδες με διαφορετικά μεγέθη και σχήματα. Η βασική ιδέα του αλγορίθμου DBSCAN είναι ότι, για κάθε σημείο μιας συστάδας, η γειτονιά μιας δεδομένης ακτίνας πρέπει να περιέχει τουλάχιστον έναν ελάχιστο αριθμό σημείων, δηλαδή, η πυκνότητα στη γειτονιά πρέπει να υπερβεί κάποιο προκαθορισμένο όριο. Ο αλγόριθμος αυτός χρειάζεται τρεις παραμέτρους εισόδου:

- k, το μέγεθος λίστας γειτονιάς
- Eps, η ακτίνα που οριοθετεί την περιοχή ενός σημείου (Eps-γειτονιά);
- MinPts, ο ελάχιστος αριθμός σημείων που πρέπει να υπάρχει στην Eps-γειτονιά.

Η διαδικασία συσταδοποίησης βασίζεται στην ταξινόμηση των σημείων στο σύνολο δεδομένων ως βασικά σημεία, οριακά σημεία και σημεία θορύβου, καθώς και για τη χρήση των σχέσεων μεταξύ της πυκνότητας σημείων για να διαμορφώσει τις συστάδες (Moreira et al., 2005).

Αλγόριθμος

1.Χαρακτήρισε κάθε σημείο ως κεντρικό , συνοριακό ή θόρυβο.

2. Αγνόησε όλα τα άλλα σημεία θορύβου.

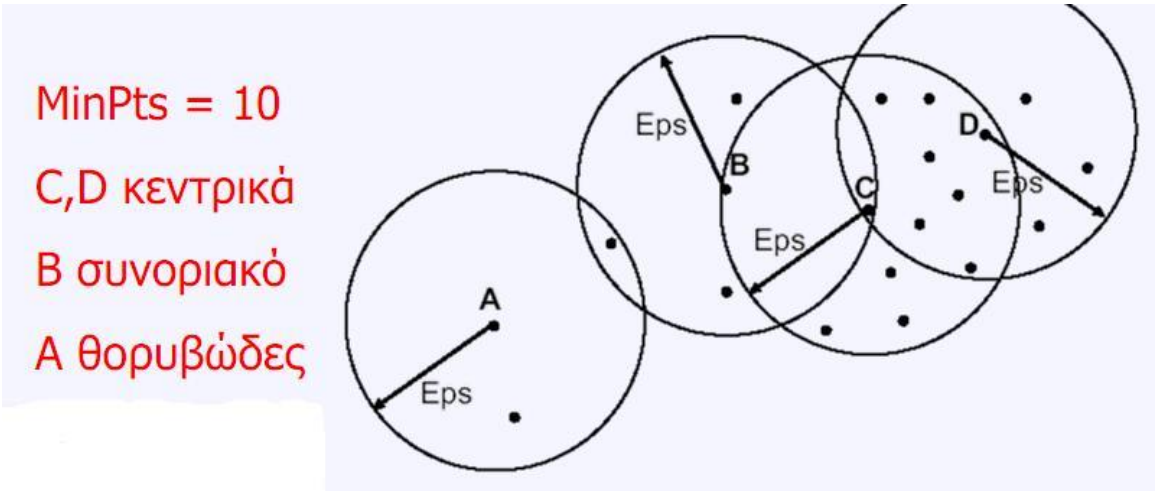
3.Δημιούργισε ένα γράφο μια κορυφή για κάθε σημείο.

4.Τοποθέτησε μια ακμή μεταξύ όλων των κεντρικών σημείων που είναι σε απόσταση ως Eps μεταξύ τους.

5. Θέσε κάθε ομάδα συνδεδεμένων βασικών σημείων ως μια διαφορετική συστάδα.

6.Ανάθεσε κάθε συνοριακό σημείο σε μια από τις συστάδες των συσχετιζόμενων βασικών σημείων

Παράδειγμα



Για κάθε σημείο βρίσκουμε το Κ πλησιέστερο προς αυτό , καθώς και τη μεταξύ τους απόσταση.

Ταξινομούμε τα σημεία ως προς την απόσταση τους ως προς το κ πλησιέστερο τους

Καθορίζουμε τις τιμές των Eps και MinPts έτσι ώστε να διαχωρίζονται τα σημεία που ανήκουν σε ομάδες από τα θορυβώδη σημεία.

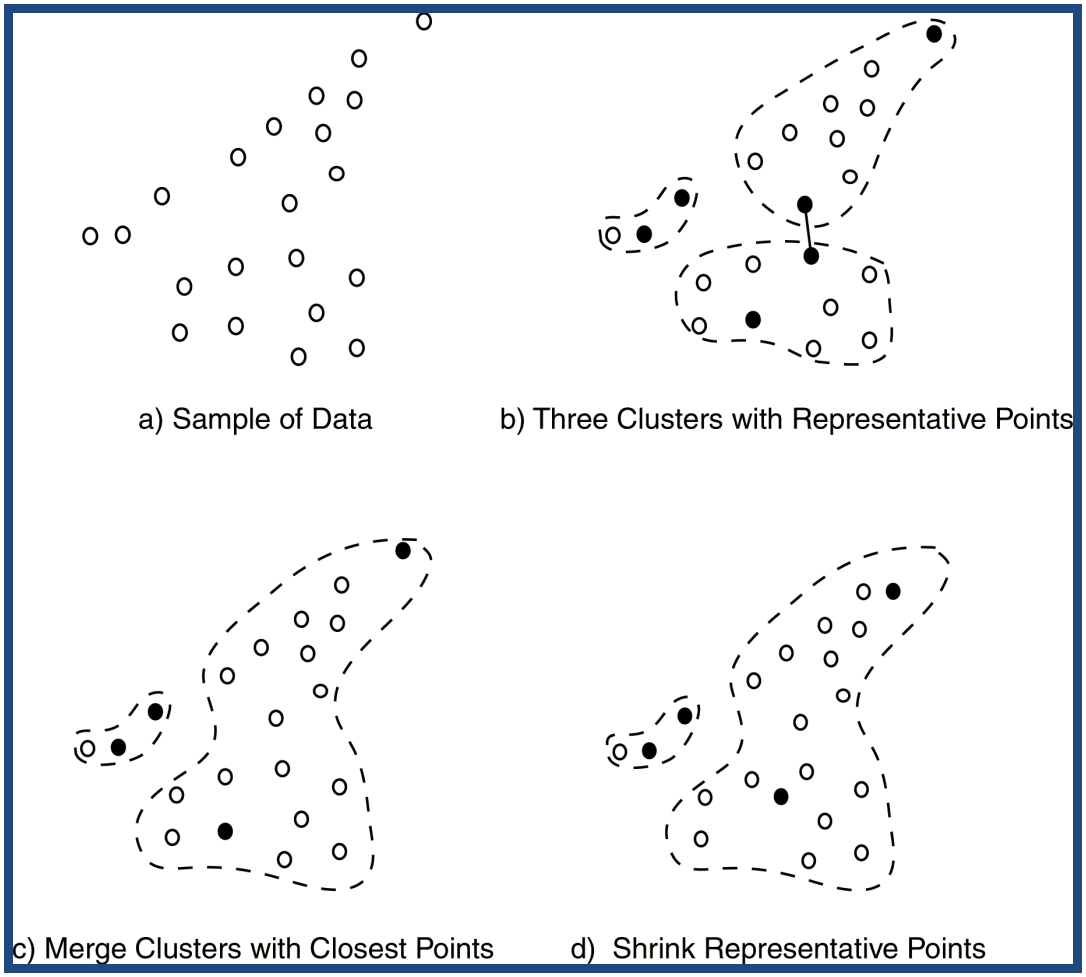
3.2.4. Αλγόριθμος CURE (Clustering Using Representatives)

Ο αλγόριθμος αρχίζει λαμβάνοντας κάθε σημείο εισόδου σαν ξεχωριστή συστάδα και σε κάθε βήμα που ακολουθεί συγχωνεύει τα πλησιέστερα ζευγάρια συστάδων. Για να υπολογιστεί η απόσταση μεταξύ των συστάδων, αποθηκεύονται για κάθε συστάδα c *αντιπρόσωποι* (*representatives).* Οι αντιπρόσωποι αυτοί καθορίζονται επιλέγοντας αρχικά τα πιο διάσπαρτα σημεία μέσα σε μία συστάδα και στη συνέχεια μετακινούμε τα σημεία προς το μέσο της

συστάδας κατά ένα ποσοστό α. Η απόσταση μεταξύ των συστάδων είναι η απόσταση μεταξύ των πιο κοντινών αντιπροσώπων δύο συστάδων. Έτσι μόνο τα σημεία αντιπρόσωποι μίας συστάδας χρησιμοποιούνται για να υπολογίσουμε την απόσταση της από μία άλλη συστάδα.

```
Input:
   D = {t_1, t_2, ..., t_n}    //Set of elements.
   k      // Desired number of clusters.
Output:
   Q      //Heap containing one entry for each cluster.
CURE Algorithm:
   T = build(D);              // Put each point in Tree
   Q = heapify(D);            // Initially build heap with one entry per item;
   repeat
      u = min(Q);
      delete(Q, u.close);
      w = merge(u, v);
      delete(T, u);
      delete(T, v);
      insert(T, w);
      for each x ∈ Q do
         x.close = find closest cluster to x;
         if x is closest to w then
            w.close = x;
      insert(Q, w);
   until number of nodes in Q is k;
```

Παράδειγμα



3.3. Δημογραφική Συσταδοποίηση (Demographic clustering)

Η **δημογραφική συσταδοποίηση (demographic clustering)** βασίζεται σε κατανομές. Παρέχει γρήγορη και φυσική ομαδοποίηση πολύ μεγάλων βάσεων δεδομένων. Οι συστάδες χαρακτηρίζονται από τις κατανομές τιμών των μελών τους. Καθορίζει αυτόματα τον αριθμό των ομάδων που πρόκειται να παραχθούν. Συνήθως, τα δημογραφικά στοιχεία περιέχουν πολλές κατηγορικές μεταβλητές. Η συνάρτηση εξόρυξης λειτουργεί καλά με σύνολα δεδομένων που αποτελούνται από αυτόν τον τύπο των μεταβλητών. Μπορούν επίσης να χρησιμοποιηθούν αριθμητικές μεταβλητές. Οι δημογραφικοί αλγόριθμοι αντιμετωπίζουν τις αριθμητικές μεταβλητές με την ανάθεση ομοιοτήτων, σύμφωνα με την αριθμητική διαφορά των τιμών (IBM, Demographic clustering, 2011).

Η δημογραφική συσταδοποίηση είναι μια επαναληπτική διαδικασία κατά την εισαγωγή δεδομένων. Κάθε εγγραφή εισόδου διαβάζεται σε διαδοχή. Η ομοιότητα της κάθε εγγραφής με καθεμιά από τις ήδη υπάρχουσες συστάδες υπολογίζεται. Αν η μεγαλύτερη υπολογισμένη ομοιότητα είναι πάνω από ένα συγκεκριμένο όριο,

η εγγραφή προστίθεται στη σχετική συστάδα. Τα χαρακτηριστικά αυτής της συστάδας αλλάζουν ανάλογα. Εάν η υπολογιζόμενη ομοιότητα δεν είναι πάνω από το κατώτατο όριο, ή αν δεν υπάρχει συστάδα (που είναι η αρχική περίπτωση) δημιουργείται μια νέα συστάδα που περιέχει μόνο την εγγραφή. Μπορεί να καθοριστεί ο μέγιστος αριθμός των ομάδων, καθώς και το όριο ομοιότητας (IBM, Demographic clustering, 2011).

Η δημογραφική συσταδοποίηση χρησιμοποιεί το στατιστικό κριτήριο Condorcet για να διαχειριστεί την ανάθεση των εγγραφών σε συστάδες και την δημιουργία νέων συστάδων. Το κριτήριο Condorcet αξιολογεί την ομοιογένεια κάθε ανακαλυφθείσας συστάδας (κατά πόσο τα στοιχεία που περιέχει είναι παρόμοια) και πόσο ετερογενείς είναι οι συστάδες μεταξύ τους. Η επαναληπτική διαδικασία της ανακάλυψης συστάδων σταματά μετά από δύο ή περισσότερα περάσματα πάνω από τα δεδομένα εισόδου, εφόσον η βελτίωση του αποτελέσματος της συσταδοποίησης σύμφωνα με το κριτήριο Condorcet δεν μπορεί να δικαιολογήσει ένα νέο πέρασμα (IBM, Demographic clustering, 2011).

3.4. Αυξητικοί Αλγόριθμοι (Incremental Algorithms)

3.4.1. BIRCH

Ο αλγόριθμος BIRCH (Balanced Iterative Reducing And Clustering Using Hierarchies) χρησιμοποιεί μια ιεραρχική δομή δεδομένων που ονομάζεται δέντρο Χαρακτηριστικού Συσταδοποίησης (Clustering Feature, CF) για τη διαμέριση των σημείων εισερχόμενων δεδομένων με προοδευτικό και δυναμικό τρόπο. Ο BIRCH εκτελεί τέσσερις διαφορετικές φάσεις κατά τη διάρκεια κάθε διαδικασίας συσταδοποίησης (Oberst, 2009):

1. Γραμμική σάρωση όλων των σημείων δεδομένων και εισαγωγή στο δέντρο CF, όπως περιγράφηκε προηγουμένως.

2. Συμπύκνωση του δέντρου CF σε επιθυμητό μέγεθος, ανάλογα με τον αλγόριθμο συσταδοποίησης που χρησιμοποιείται στο βήμα τρία. Αυτό μπορεί να περιλαμβάνει την αφαίρεση των ακραίων τιμών και την περαιτέρω συγχώνευση των συστάδων.

3. Χρήση αλγορίθμου καθολικής συσταδοποίησης χρησιμοποιώντας τα φύλλα του δέντρου CF ως είσοδο. Τα Χαρακτηριστικά Συσταδοποίησης επιτρέπουν την αποτελεσματική μέτρηση των αποστάσεων.

4. Προαιρετικά τελειοποίηση της εξόδου του βήματος τρία. Όλες οι ομάδες πλέον αποθηκεύονται στη μνήμη. Αν χρειάζεται τα πραγματικά σημεία δεδομένων μπορούν να συσχετιστούν με τις παραγόμενες συστάδες με την ανάγνωση όλων των σημείων από το δίσκο ξανά.

3.5. k Πλησιέστερος Γείτονας (kNN)

Ο αλγόριθμος k-πλησιέστερου γείτονα (k Nearest Neighbor, kNN) αποτελεί μέρος της μηχανικής μάθησης που έχει χρησιμοποιηθεί σε πολλές εφαρμογές στον τομέα της εξόρυξης δεδομένων, της στατιστικής αναγνώρισης προτύπων και πολλών άλλων. Ο kNN είναι μια μέθοδος για την ταξινόμηση των αντικειμένων με βάση τα πλησιέστερα παραδείγματα εκπαίδευσης στο χώρο των χαρακτηριστικών. Ένα αντικείμενο κατατάσσεται από την πλειοψηφία των γειτόνων του. Το k είναι πάντα ένας θετικός ακέραιος. Οι γείτονες λαμβάνονται από ένα σύνολο αντικειμένων για τα οποία η ορθή κατάταξη είναι γνωστή. Είναι σύνηθες να χρησιμοποιείται η Ευκλείδεια απόσταση, αν και άλλα μέτρα απόστασης, όπως η απόσταση Manhattean θα μπορούσαν εξαρχής να χρησιμοποιηθούν αντί αυτού. Ο αλγόριθμος για το πώς να υπολογιστούν οι k-πλησιέστεροι γείτονες έχει ως εξής (The Code Project, 2009):

1. Καθορισμός της παραμέτρου k=αριθμός των κοντινότερων γειτόνων εκ των προτέρων. Η τιμή αυτή καθορίζεται από το χρήστη.

2. Υπολογισμός της απόστασης μεταξύ του στιγμιότυπου ερωτήματος και όλων των δειγμάτων εκπαίδευσης. Μπορεί να χρησιμοποιηθεί οποιοσδήποτε αλγόριθμος απόστασης.

3. Ταξινόμηση των αποστάσεων για όλα τα δείγματα εκπαίδευσης και προσδιορισμός του πλησιέστερου γείτονα με βάση την k ελάχιστη απόσταση.

4. Δεδομένου ότι πρόκειται για επιβλεπόμενη μάθηση, χρήση όλων των κατηγοριών των δεδομένων εκπαίδευσης για την ταξινομημένο τιμή που εμπίπτει στο k.

5. Χρήση της πλειοψηφίας των κοντινότερους γειτόνων ως η τιμή πρόβλεψης.

Παράδειγμα

Χρησιμοποιώντας το δείγμα δεδομένων του πίνακα 9 και την κατηγοριοποίηση 1 σαν τιμή εξόδου του συνόλου εκπαίδευσης, κατηγοριοποιούμε την πλειάδα <Pat, Θ, 1.6>. Μόνο το ύψος χρησιμοποιείται για τον υπολογισμό της απόστασης αφού αυτό είναι το μόνο αριθμητικό χαρακτηριστικό. Είτε χρησιμοποιήσουμε το ευκλείδειο είτε το Manhattan μέτρο απόστασης θα έχουμε τα ίδια αποτελέσματα στην απόδοση. Άρα, η απόσταση είναι απλά η απόλυτη τιμή της διαφοράς των τιμών. Ας υποθέσουμε ότι δίνεται *K*=5. Έτσι στην συνέχεια, υπολογίζοντας της αποστάσεις, οι *K* κοντινότεροι γείτονες στην πλειάδα εισόδου είναι οι πλειάδες {<Kristina, Θ, 1.6 >, <Kathy, Θ, 1.6>, <Stephanie, Θ, 1.7>, <Dave, Α, 1.7>, <Wynette, Θ, 1.75>}. Από αυτά τα πέντε στοιχεία, τέσσερα είναι κατηγοριοποιημένα στην κατηγορία των κοντών και ένα στην κατηγορία των μέτριων. Έτσι ο KNN θα κατηγοριοποιήσει τον Pat στους κοντούς

3.6. Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκαν οι τεχνικές τμηματοποίησης που εφαρμόζονται στην εξόρυξη δεδομένων για να οργανώσουν τα δεδομένα σε ομάδες με συγκεκριμένα χαρακτηριστικά. στη συνέχεια ακολουθούν οι νέες τεχνικές που εφαρμόζονται στην εξόρυξη δεδομένων, ξεκινώντας με την εξόρυξη κανόνων και τους συνδυαστικούς κανόνες.

4. ΣΥΝΔΥΑΣΤΙΚΟΙ ΚΑΝΟΝΕΣ (ASSOCIATION RULES)

4.1. Εισαγωγή

Η **εξόρυξη συνδυαστικών κανόνων (association rule mining)** βρίσκει ενδιαφέρουσες ενώσεις ή/και σχέσεις αντιστοιχίας ανάμεσα σε μεγάλα σύνολα των στοιχείων δεδομένων. Οι συνδυαστικοί κανόνες δείχνουν συνθήκες απόδοσης τιμής που εμφανίζονται συχνά μαζί σε ένα συγκεκριμένο σύνολο δεδομένων. Οι συνδυαστικοί κανόνες παρέχουν πληροφορίες αυτού του είδους, με τη μορφή δηλώσεων «εάν-τότε» (if-then). Οι κανόνες αυτοί υπολογίζονται με βάση τα δεδομένα και, σε αντίθεση με τους κανόνες if-then της λογικής, οι συνδυαστικοί

κανόνες είναι πιθανολογικοί στη φύση. Εκτός από το προγενέστερο (το «αν» μέρος) και το επακόλουθο (το «τότε» μέρος), ένας συνδυαστικός κανόνας έχει δύο αριθμούς που εκφράζουν το βαθμό αβεβαιότητας σχετικά με τον κανόνα. Στην ανάλυση σύνδεσης το προγενέστερο και το επακόλουθο είναι σύνολα αντικειμένων (που ονομάζονται **στοιχειοσύνολα** ή **itemsets**) που είναι ασύνδετα (δεν έχουν οποιαδήποτε κοινά στοιχεία). Ο πρώτος αριθμός καλείται η **υποστήριξη (support)** για τον κανόνα. Η υποστήριξη είναι απλά ο αριθμός των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία των προγενέστερων και συνακόλουθων μερών του κανόνα. (Η υποστήριξη μερικές φορές εκφράζεται ως ποσοστό του συνολικού αριθμού των εγγραφών στη βάση δεδομένων.) Ο άλλος αριθμός είναι γνωστός ως η εμπιστοσύνη του κανόνα. Η εμπιστοσύνη είναι ο λόγος του αριθμού των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία του συνακόλουθου καθώς και του προγενέστερου (δηλαδή, η υποστήριξη) προς τον αριθμό των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία του προγενέστερου (XLMiner).

4.2. Μερική συσταδοποίηση με συνδυαστικούς κανόνες

Στην **μερική συσταδοποίηση (partial classification)**, ο κύριος σκοπός δεν είναι η πρόβλεψη ή οι μελλοντικές τιμές, αλλά κυρίως η ανακάλυψη χαρακτηριστικών ορισμένων τάξεων. Ο στόχος είναι η εκμάθηση κανόνων που είναι μεμονωμένα ακριβείς, αλλά μπορεί να μην καλύπτουν όλα τα παραδείγματα που ανήκουν σε μια κλάση. Επιπρόσθετα, τα παραδείγματα αυτά μπορούν να αλληλοεπικαλύπτονται. Στην περίπτωση αυτή επομένως χρησιμοποιούνται συνδυαστικοί κανόνες, σε πεδία όπου οι κλασικοί ταξινομητές μπορεί να είναι μη αποτελεσματικοί λόγω των ακόλουθων συνθηκών (Kamal et al):

1. Μεγάλος αριθμός χαρακτηριστικών.
2. Οι περισσότερες τιμές ενός χαρακτηριστικού λείπουν.
3. Η κατανομή είναι πολύ «λοξή» και ο χρήστης θέλει να κατανοήσει ορισμένες τάξεις χαμηλής συχνότητας.
4. Τα χαρακτηριστικά πρέπει να μοντελοποιηθούν με βάση τα άλλα χαρακτηριστικά.

5. Μεγάλος αριθμός παραδειγμάτων εκπαίδευσης.

4.3. Apriori

Ο Apriori είναι ένας αλγόριθμος εξόρυξης συνδυαστικών κανόνων και έχει αναπτυχθεί για εξόρυξη κανόνων σε μεγάλες βάσεις δεδομένων συναλλαγών. Το πρόβλημα της εξόρυξης συνδυαστικών κανόνων έχει οριστεί ότι αποτελείται από δύο μέρη (Stonebraker et al., 1998):

- Εύρεση όλων τους συνδυασμούς αντικειμένων που έχουν την υποστήριξη συναλλαγών πάνω από την ελάχιστη υποστήριξη. Οι συνδυασμοί αυτοί ονομάζονται **συχνά στοιχειοσύνολα (frequent itemsets)**.

- Χρήση των συχνών στοιχειοσυνόλων ώστε να παράγουν τους επιθυμητούς κανόνες. Η γενική ιδέα είναι ότι αν, π.χ. ABCD και AB είναι συχνά στοιχειοσύνολα, τότε μπορεί να καθοριστεί εάν ο κανόνας AB CD ισχύει με υπολογισμό του λόγου r=support(ABCD)/support(AB). Ο κανόνας ισχύει μόνο αν r>=ελάχιστη εμπιστοσύνη. Ο κανόνας θα έχει την ελάχιστη υποστήριξη, διότι το ABCD είναι συχνό.

Στον Πίνακα 2 παρουσιάζεται ο ψευδοκώδικας του αλγορίθμου Apriori (Stonebraker et al., 1998).

Πίνακας 2: Ο αλγόριθμος Apriori (Stonebraker et al., 1998).

```
Συνάρτηση AprioriAlg()
begin
L1 := {συχνά 1-στοιχειοσύνολα};
for ( k := 2; Lk-1 0; k++ ) do {
   Ck= apriori-gen(Lk-1) ; // νέοι υποψήφιοι
για όλες τις συναλλαγές  t στο σύνολο δεδομένων do {
  για όλους τους υποψήφιους  c Ck που περιέχονται στο t do
          c:count++
  }
  Lk = { c Ck | c:count >= min-support}
}
Απάντηση := k Lk
end
```

Ο αλγόριθμος κάνει πολλαπλά περάσματα στη βάση δεδομένων. Στο πρώτο πέρασμα, ο αλγόριθμος μετράει μόνο τις εμφανίσεις περιστατικών για να καθοριστούν τα συχνά 1-στοιχειοσύνολα (στοιχειοσύνολα με 1 στοιχείο). Ένα μεταγενέστερο πέρασμα, k, αποτελείται από δύο φάσεις. Πρώτον, τα συχνά στοιχειοσύνολα L_{k-1} (το σύνολο όλων των συχνών (k-1)-στοιχειοσυνόλων), που βρέθηκαν στο (k-1) πέρασμα χρησιμοποιούνται για να παραγάγουν τα στοιχειοσύνολα υποψηφίων C_k, χρησιμοποιώντας τη συνάρτηση Apriori-gen(). Η λειτουργία αυτή ενώνει αρχικά το L_{k-1} με το L_{k-1}, την προϋπόθεση ένωσης ότι τα λεξικογραφικά ταξινομημένα πρώτα k-2 στοιχεία είναι τα ίδια. Στη συνέχεια, διαγράφει όλα τα στοιχειοσύνολα από το ενωμένο αποτέλεσμα που έχουν κάποιο (k-1)-υποσύνολο που δεν είναι στο L_{k-1} δίνοντας C_k. Ο αλγόριθμος σαρώνει τη βάση δεδομένων. Για κάθε συναλλαγή, καθορίζει ποιες από τις υποψήφιες στο C_k περιλαμβάνονται στη συναλλαγή χρησιμοποιώντας μια δομή δεδομένων δένδρου κατακερματισμού και αυξάνει τον αριθμού των υποψηφίων. Στο τέλος του περάσματος, το C_k εξετάζεται ως προς το ποιες από τις υποψήφιες είναι συχνές,

δίνοντας L_k. Ο αλγόριθμος τερματίζει όταν το L_k αδειάσει (Stonebraker et al., 1998).

4.4. Άλλοι αλγόριθμοι

Στη ενότητα αυτή παρουσιάζονται ορισμένοι αλγόριθμοι που αποτελούν βελτιώσεις ή επεκτάσεις του αλγορίθμου Apriori (dataminingarticles.com, 2011):

- AIS: Ο αλγόριθμος αυτός χρησιμοποιεί δημιουργία υποψηφίων για την ανίχνευση συχνών στοιχειοσυνόλων και ήταν ο πρώτος αλγόριθμος που εισήγαγε το πρόβλημα της δημιουργίας συνδυαστικών κανόνων.

- SETM: Ο αλγόριθμος αυτός επίσης χρησιμοποιεί δημιουργία υποψηφίων στις συναλλαγές ανάγνωσης από τη βάση δεδομένων, αλλά δημιουργήθηκε κυρίως για υπολογισμούς SQL και διαχωρίζει τη δημιουργία υποψηφίων από τη μέτρηση.

- DHP (Direct Hashing and Pruning): Ο αλγόριθμος αυτός χρησιμοποιεί μια τεχνική κατακερματισμού για τη δημιουργία υποψηφίων στοιχειοσυνόλων και αποτελεσματικές τεχνικές κλαδέματος για να μειώσει το μέγεθος της βάσης δεδομένων συναλλαγών.

- Partition: Ο αλγόριθμος αυτός είναι θεμελιωδώς διαφορετικός από τους προηγούμενους αλγορίθμους, καθώς διαβάζει τη βάση δεδομένων τουλάχιστον δυο φορές μέχρι να δημιουργήσει όλους τους σημαντικούς συνδυαστικούς κανόνες.

- ECLAT (Echivalence Class Clustering and Bottom-up Lattice Traversal): Ο αλγόριθμος αυτός είναι ο πρώτος αλγόριθμος που χρησιμοποιεί μια δομή κάθετων δεδομένων (ανεστραμμένη) και είναι πολύ αποτελεσματικός για μεγάλα στοιχειοσύνολα αλλά λιγότερο αποτελεσματικός για μικρότερα στοιχειοσύνολα.

- FP-GROWTH: Ο αλγόριθμος αυτός αντιμετωπίζει τα δυο μεγάλα μειονεκτήματα του αλγορίθμου Apriori, υιοθετώντας μια στρατηγική διαίρει-και-βασίλευε και ένα δένδρο συχνών προτύπων.

- TREE-PROJECTION: Ο αλγόριθμος αυτός έχει μια καινοτομία, τη χρήση ενός λεξικογραφικού δένδρου που απαιτεί σημαντικά λιγότερη μνήμη από ένα δένδρο κατακερματισμού.

- PASCAL: Ο αλγόριθμος αυτός αποτελεί μια βελτιστοποίηση του αλγορίθμου Apriori, βρίσκει και συχνά και κλειστά σύνολα δέκα φορές γρηγορότερα από τον Apriori αλλά είναι μόνο πρακτικός όταν το μήκος του προτύπου είναι μικρό.

- H-MINE: Ο αλγόριθμος αυτός εισάγει την έννοια της δομής δεδομένων υπερσυνδέσεων (H-struct) και τη χρησιμοποιεί για να προσαρμόσει δυναμικά συνδέσμους στην διεργασία εξόρυξης.

- RELIM (Recursive Elimination): Ο αλγόριθμος αυτός μοιάζει πολύ με τον FP-growth και τον H-mine.

- MAX-MINER: Ο αλγόριθμος αυτός εξάγει μόνο τα μέγιστα συχνά στοιχειοσύνολα, δηλαδή αυτά που δεν έχουν υπερσύνολο που είναι συχνό.

- DepthProject: Ο αλγόριθμος αυτός βρίσκει συχνά στοιχειοσύνολα χρησιμοποιώντας αναζήτηση κατά βάθος σε ένα λεξικογραφικό δένδρο στοιχειοσυνόλων.

- MAFIA: Ο αλγόριθμος αυτός χρησιμοποιείται για εξόρυξη μέγιστων συχνών στοιχειοσυνόλων από μια βάση δεδομένων συναλλαγών και είναι ιδιαίτερα αποτελεσματικός όταν τα δεδομένα είναι πολύ μεγάλα.

- GenMax: Ο αλγόριθμος αυτός βασίζεται σε αντίστροφη αναζήτηση για εξόρυξη μέγιστων συχνών στοιχειοσυνόλων και χρησιμοποιεί προοδευτική εστίαση και τεχνικές ελέγχου μεγιστοποίησης και αύξησης της ταχύτητας.

- CLOSE: Ο αλγόριθμος αυτός εξορύσσει άμεσα συχνά στοιχειοσύνολα με αναζήτηση bottom-up για το μικρότερο συχνό στοιχειοσύνολο και στη συνέχεια συγκρίνει την υποστήριξη του κάθε συνόλου με τα υποσύνολά του στο προηγούμενο επίπεδο. Μια παραλλαγή του αλγορίθμου, αποκαλούμενη A-CLOSE, δεν καθορίζει όλα τα συχνά στοιχειοσύνολα, μειώνοντας το υπολογιστικό κόστος του.

- CLOSET: Ο αλγόριθμος αυτός εξορύσσει συχνά στοιχειοσύνολα με τη χρήση δομής FP-tree και μιας αναδρομικής στρατηγικής διαίρει-και-βασίλευε.
- CHARM: Ο αλγόριθμος αυτός εξορύσσει άμεσα συχνά στοιχειοσύνολα και ταυτόχρονα εξερευνεί το χώρο των στοιχειοσυνόλων και το χώρο των συναλλαγών, χρησιμοποιώντας μια υβριδική μέθοδο αναζήτησης που παρέχει μεγαλύτερη ταχύτητα αναγνώρισης των συχνών στοιχειοσυνόλων.

4.5. Επίλογος

Στο κεφάλαιο αυτό περιγράφηκαν οι συνδυαστικοί κανόνες και η εφαρμογή τους στην εξόρυξη δεδομένων. Στη συνέχεια ακολουθεί η σειριακή ανάλυση δεδομένων.

5. ΣΕΙΡΙΑΚΗ ΑΝΑΛΥΣΗ (SEQUENTIAL ANALYSIS)

5.1. Εισαγωγή

Στο κεφάλαιο αυτό περιγράφονται οι βασικοί αλγόριθμοι εξόρυξης δεδομένων που χρησιμοποιούνται στην σειριακή ανάλυση.

Σε πολλά επιχειρηματικά και επιστημονικά πεδία η ύπαρξη γεγονότων τα οποία συμβαίνουν σε μια **ακολουθία (sequence)** παρουσιάζει ενδιαφέρον. Η μελέτη τακτικών προτύπων στις βάσεις δεδομένων συνάντησε μεγάλη πρόοδο με την ανάλυση σειρών εγγραφών συναλλαγών. Είναι επίσης γνωστή στη βιβλιογραφία ως **ανάλυση καλαθιού δεδομένων (basket data analysis)** και αποτελείται από την ανακάλυψη αντικειμένων που συχνά αποκτώνται μαζί (Filho et al., 2004).

5.2. Αλγόριθμοι για την εύρεση σειριακών προτύπων

Οι αλγόριθμοι για την **εξόρυξη σειριακών προτύπων (sequential pattern mining)** διευθετούν το πρόβλημα της ανακάλυψης των υπαρχόντων μεγίστων ακολουθιών σε μια δοθείσα βάση δεδομένων. Οι αλγόριθμοι για το θέμα αυτό είναι σχετικοί όταν τα δεδομένα που πρόκειται να εξορυχθούν έχουν κάποια σειριακή φύση, π.χ. όταν κάθε δεδομένο είναι μια ταξινομημένη σειρά στοιχείων. Ο στόχος αυτής της μεθόδου εξόρυξης δεδομένων είναι να ανακαλύψει όλες τις συχνές ακολουθίες (frequent sequences) στοιχειοσυνόλων σε ένα σύνολο δεδομένων. Συγκεκριμένα, ένα στοιχειοσύνολο (itemset) είναι ένα μη κενό υποσύνολο των στοιχείων από ένα σύνολο C, τη συλλογή αντικειμένων, που αποκαλούνται αντικείμενα (items). Με τον τρόπο αυτό, ένα στοιχειοσύνολο αντιπροσωπεύει το σύνολο αντικειμένων που συμβαίνουν μαζί. Το στοιχειοσύνολο που αποτελείται από τα αντικείμενα α και b συμβολίζεται με (αb). Μια ακολουθία (sequence) είναι μια ταξινομημένη λίστα στοιχείων. Μια ακολουθία είναι μέγιστη αν δεν περιέχεται σε καμία άλλη ακολουθία. Μια ακολουθία με k αντικείμενα ονομάζεται k-ακολουθία. Ο αριθμός των αντικειμένων (στοιχειοσυνόλων) σε μια ακολουθία s είναι το μήκος της ακολουθίας και συμβολίζεται με |s|. Το i-οστό στοιχειοσύνολο στην ακολουθία συμβολίζεται με s_i και το σύνολο των ακολουθιών που λαμβάνονται υπόψη

θεωρείται ως η βάση δεδομένων (DB) και ο αριθμός των ακολουθιών θεωρείται το μέγεθος της βάσης δεδομένων (|DB|) (Antunes et al, 2004).

Στη συνέχεια περιγράφονται ορισμένοι αλγόριθμοι για την εύρεση σειριακών προτύπων.

AprioriAll: Ο αλγόριθμος AprioriAll δίνεται στον Πίνακα 3. Σε κάθε πέρασμα, χρησιμοποιούνται οι μεγάλες ακολουθίες από το προηγούμενο πέρασμα για τη δημιουργία των υποψηφίων ακολουθιών και στη συνέχεια μετράται η υποστήριξη τους, κάνοντας ένα πέρασμα πάνω από τη βάση δεδομένων. Στο τέλος του περάσματος, η υποστήριξη των υποψηφίων χρησιμοποιείται για τον προσδιορισμό των μεγάλων ακολουθιών. Στο πρώτο πέρασμα, η έξοδος της φάσης litemset χρησιμοποιείται για να αρχικοποιηθεί το σύνολο των μεγάλων 1-ακολουθιών. Οι υποψήφιοι αποθηκεύονται σε δένδρο κατακερματισμού (hash-tree) για να τη γρήγορη εύρεση όλων των υποψηφίων που περιέχονται σε μια ακολουθία πελάτη. Η συνάρτηση apriori-generate (Πίνακας 4) παίρνει ως όρισμα το L_{k-1}, το σύνολο όλων των μεγάλων (k-1)-ακολουθιών (Agrawal et al., 2004).

Πίνακας 3: Αλγόριθμος AprioriAll (Agrawal et al., 2004).

L_1 = μεγάλες 1-ακολουθίες; // Αποτέλεσμα φάσης litemset

for (k = 2; L_{k-1} 0; k++) **do**

begin

C_k = Νέοι Υποψήφιοι δημιουργημένοι από L_{k-1} (βλέπε παρακάτω)

for each ακολουθία πελάτη *c* στη βάση δεδομένων **do**

Αύξηση της μέτρησης όλων των υποψηφίων στο C_k που περιέχονται στο *c*.

L_k = Υποψήφιοι στο C_k με ελάχιστη υποστήριξη.

end

Απάντηση = Μέγιστες ακολουθίες στο $\bigcup_k L_k$;

Πίνακας 4: Η συνάρτηση apriori-generate (Agrawal et al., 2004).

insert into C_k

select $p.itemset_1$, ..., $p.itemset_{k-1}$, $q.itemset_{k-1}$

from L_{k-1} p, L_{k-1} q

where $p.itemset_1 = q.itemset_1$, . . .,

$p.itemset_{k-2} = q.itemset_{k-2}$;

AprioriSome: Ο αλγόριθμος AprioriSome επιχειρεί να μην μετρήσει τις μη μέγιστες ακολουθίες υπολογίζοντας τις μέγιστες ακολουθίες στην αρχή. Κατά την εκτέλεση του αλγορίθμου γίνονται δυο περάσματα. Στο πέρασμα προς τα εμπρός, μετρώνται μόνο μεγάλες ακολουθίες με ορισμένα μήκη. Η συνάρτηση στη συνέχεια παίρνει ως παράμετρο το μήκος των ακολουθιών που υπολογίζεται κατά το τελευταίο πέρασμα και επιστρέφει το μήκος των ακολουθιών που πρέπει να υπολογιστεί στο επόμενο πέρασμα. Έτσι, η συνάρτηση αυτή καθορίζει ακριβώς ποιες ακολουθίες μετριούνται και ισορροπεί την αντίστροφη σχέση μεταξύ του χρόνου που χάνεται στο μέτρημα των μη μέγιστων ακολουθιών έναντι της καταμέτρησης επεκτάσεων των μικρών υποψηφίων ακολουθιών (Agrawal et al., 2004).

Αλγόριθμος AprioriSome(Forward Phase)

L_1 = {large 1-sequences}

$C_1 = L_1$

last = 1

for (k = 2; $C_{k-1} \neq \{\}$ **and** $L_{last} \neq \{\}$; k++) **do**

begin

if (L_{k-1} known) **then**

C_k = New candidates generated from L_{k-1}

else

C_k = New candidates generated from C_{k-1}

if (k==next(last)) **then begin** // (next k to count?)

foreach customer-sequence c in the database **do**

Increment the count of all candidates in C_k

that are contained in c.

L_k = Candidates in C_k with minimum support.

last = k;

end

end

Αλγόριθμος AprioriSome(Backward Phase)

for (k--; k>=1; k--) **do**

if (L_k not found in forward phase) **then begin**

Delete all sequences in C_k contained in

some L_i i>k;

foreach customer-sequence c in D_T **do**

Increment the count of all candidates in C_k

that are contained in c

L_k = Candidates in C_k with minimum support

end

else // l_k already known

Delete all sequences in C_k contained in

some L_i i>k;

Answer = $U_k L_k$ //(Maximal Phase not Needed)

Notation: D_T; Transformed database

AprioriTid: Ο αλγόριθμος AprioriTid δημιουργεί μια νέα βάση δεδομένων C_k σε κάθε βήμα k, με τα αντικείμενα που είναι <TID, όλα τα υποσύνολα μεγέθους k> και σαρώνει αυτά παρά την πραγματική βάση δεδομένων στον επόμενο γύρο. Τα πλεονεκτήματα του αλγορίθμου είναι πως γίνονται λιγότερες αναζητήσεις στη βάση σε κάθε στάδιο και κάθε σειρά μπορεί να γίνει μικρότερη με μεγάλο k, ένα παίγνιο του τύπου (n επιλέγω k). Το βασικό μειονέκτημα είναι πως οι δημιουργούμενες βάσεις δεδομένων μπορεί να είναι μεγάλες (Stonebraker, 1998).

AprioriHybrid: Πρόκειται για μια υβριδική προσέγγιση: αρχικά χρησιμοποιείται ο Apriori για λίγο, στη συνέχεια γίνεται μετάβαση στον AprioriTid όταν η παραγόμενη βάση δεδομένων μπορεί να χωρέσει στη μνήμη χρησιμοποιώντας το C_k και τις μετρήσεις για κάθε υποψήφιο σύνολο (Stonebraker, 1998).

5.3. Επίλογος

Στο κεφάλαιο αυτό περιγράφηκαν οι αλγόριθμοι για την εύρεση σειριακών προτύπων σε ακολουθίες. Στη συνέχεια θα περιγραφεί πως η καθοδηγούμενη μάθηση εφαρμόζεται στην εξόρυξη δεδομένων.

6. ΚΑΘΟΔΗΓΟΥΜΕΝΗ ΕΚΜΑΘΗΣΗ (MACHINE LEARNING)

6.1. Εισαγωγή

Ως **καθοδηγούμενη ή μηχανική εκμάθηση (machine learning)** ορίζεται η ικανότητα μιας μηχανής να βελτιώσει τις επιδόσεις της με τη χρήση ενός λογισμικού που χρησιμοποιεί τεχνικές τεχνητής νοημοσύνης για να μιμηθεί τους τρόπους με τους οποίους μαθαίνουν οι άνθρωποι, όπως η επανάληψη και η εμπειρία (BusinessDictionary.com, machine learning). Στο κεφάλαιο αυτό παρουσιάζεται ο τρόπος με τον οποίο συνδέεται η μηχανική εκμάθηση με την εξόρυξη δεδομένων.

6.2. AdaBoost

Ο αλγόριθμος ADABoost (Adaptive Boosting) είναι ένα μετα-αλγόριθμος που χρησιμοποιείται για τη βελτίωση των αποτελεσμάτων ταξινόμησης. Η ιδέα είναι να γίνει η συνεργασία πολλών αδύναμων ταξινομητών για την ενίσχυση των αποτελεσμάτων. Η προσαρμοστικότητα σημαίνει στην περίπτωση αυτή ότι η

ανίχνευση λανθασμένης κατάταξης κάνει τον αλγόριθμο να εργάζεται περισσότερο σε αυτό (αλλάζοντας τα κέρδη και ρυθμίζοντας τον αλγόριθμο ώστε να προσπαθήσει περισσότερο εκεί όπου απέτυχε. Ο AdaBoost είναι ευαίσθητος σε θορυβώδη δεδομένα ή ακραίες τιμές (WebContentMining.com, 2010).

Η **ενίσχυση (boosting)** χρησιμοποιεί ένα αδύναμο αλγόριθμο μάθησης (ο οποίος προσδιορίζεται ως ο μαθητής). Ο αλγόριθμος απεικονίζεται στο Σχήμα 2. Γίνεται η υπόθεση ότι το σύνολο δεδομένων αποτελείται από N οντότητες που περιγράφονται χρησιμοποιώντας M μεταβλητές (γραμμές 1 και 2 του μετα-κώδικα). Η M-οστή μεταβλητή (δηλαδή, η τελευταία μεταβλητή της κάθε παρατήρησης) θεωρείται ότι είναι η ταξινόμηση της παρατήρησης. Τα δεδομένα εκπαίδευσης (μια μήτρα $\mathrm{N}\times(\mathrm{M}-1)$) συμβολίζονται με x (γραμμή 3) και η τάξη που συνδέεται με κάθε παρατήρηση στα δεδομένα εκπαίδευσης (ένα διάνυσμα μήκους M) ως y (γραμμή 4). Χωρίς απώλεια η τάξη μπορεί να περιοριστεί να είναι είτε 1 (ίσως εκπροσωπεί το *ναι*) ή -1 (που αντιπροσωπεύει το *όχι*). Αυτό θα απλοποιήσει τα μαθηματικά. Σε κάθε παρατήρηση στα δεδομένα εκπαίδευσης αποδίδεται αρχικά το ίδιο βάρος: $w_i = \frac{1}{N}$ (γραμμή 5). Ο αδύναμος μαθητής θα χρειαστεί να χρησιμοποιήσει τα βάρη που συνδέονται με κάθε παρατήρηση. Αυτό μπορεί να γίνει είτε απευθείας από το μαθητή (π.χ., το *rpart* παίρνει μια επιλογή για να καθορίσει το Roption[]weights είτε δημιουργεί ένα τροποποιημένο σύνολο δεδομένων με δειγματοληψία του αρχικού συνόλου δεδομένων με βάση τα βάρη. Το πρώτο μοντέλο, M_1 δημιουργείται με την εφαρμογή του ασθενούς μαθητή στα δεδομένα με τα βάρη w (γραμμή 7). Το M_1, προβλέποντας ή 1 ή -1, στη συνέχεια χρησιμοποιείται για να προσδιορίσει το σύνολο των δεικτών ή εσφαλμένα ταξινομημένων οντοτήτων (δηλ. όπου $M_1(x_p) \neq y_p$), που συμβολίζεται ως *ms* (γραμμή 8). Για ένα απόλυτα ακριβές μοντέλο θα είχαμε $M_1(x_i) \neq y_i$. Φυσικά το μοντέλο αναμένεται να είναι μόνο ελαφρώς καλύτερο από τυχαίο, έτσι το *ms* είναι απίθανο να είναι κενό. Ένα σχετικό σφάλμα ε_1 για το M_1 υπολογίζεται ως το σχετικό άθροισμα των βαρών των εσφαλμένα ταξινομημένων οντοτήτων (γραμμή 9). Αυτό χρησιμοποιείται για τον υπολογισμό του a_1 (γραμμή 10), που χρησιμοποιείται, με τη σειρά του, για να προσαρμόσει τα βάρη (γραμμή 11). Όλα

τα βάρη θα μπορούσαν είτε να μειωθούν ή να αυξηθούν ανάλογα με το αν το μοντέλο ταξινομεί ορθά την αντίστοιχη παρατήρηση. Ωστόσο, αυτό μπορεί να απλοποιηθεί με την αύξηση μόνο των βαρών των εσφαλμένα ταξινομημένων οντοτήτων. Οι οντότητες αυτές γίνονται έτσι πιο σημαντικές. Ο αλγόριθμος εκμάθησης εφαρμόζεται στη συνέχεια στα νέα σταθμισμένα δεδομένα με τον μαθητή να αναμένεται να δώσει μεγαλύτερη έμφαση στις δύσκολες οντότητες καθώς δομεί το επόμενο μοντέλο, M_2. Τα βάρη, στη συνέχεια, τροποποιούνται εκ νέου χρησιμοποιώντας τα λάθη από το M_2. Η δόμηση μοντέλου και η τροποποίηση βαρών στη συνέχεια επαναλαμβάνονται έως ότου το νέο μοντέλο να μην είναι καλύτερο από το τυχαίο (δηλαδή, το σφάλμα να είναι 50% ή περισσότερο: $e_i \geq 0.5$), ή είναι τέλειο (δηλαδή, το ποσοστό σφάλματος είναι 0% και το *ms* είναι άδειο), ή ίσως μετά από έναν ορισμένο αριθμό επαναλήψεων (Williams, AdaBoost Algorithm, 2010).

AdaBoost($data, learner$):

1 $N \leftarrow nrow(data)$
2 $M \leftarrow ncol(data)$
3 $x \leftarrow data[, 1 : M-1]$
4 $y \leftarrow data[, M]$
5 **for** $i \leftarrow 1$ **to** N: $w_i = \frac{1}{N}$

6 Repeat $i \leftarrow 1$, $i \leftarrow i+1$:
7 $\mathcal{M}_i \leftarrow learner(data, w)$
8 $ms = \{p | \mathcal{M}_i(x_p) \neq y_p\}$
9 $\epsilon_i = \frac{\sum_{i \in ms} w_j}{\sum_{j=1}^{n} w_j}$
10 $\alpha_i = \log((1-\epsilon_i)/\epsilon_i)$
11 **for** $j \in ms$: $w_j = w_j \times e^{\alpha_i}$
12 **for** $i \leftarrow 1$ **to** N: $w_i = \frac{w_i}{\sum_{j=1}^{n} w_j}$
13 **until** $\epsilon_i >= 0.5$ or $ms = \emptyset$

14 Return$[\mathcal{M}(x) = \mathbf{sign}(\sum_{j=1}^{T} \alpha_j \mathcal{M}_j(x))]$

Σχήμα 2: Ο αλγόριθμος AdaBoost (Williams, AdaBoost Algorithm, 2010).

Το τελικό μοντέλο M (γραμμή 14) συνδυάζει τα άλλα μοντέλα που χρησιμοποιούν ένα σταθμισμένο άθροισμα των εξόδων των άλλων μοντέλων. Τα βάρη, a_j, αντανακλούν την ακρίβεια καθενός από τα μοντέλα. Όσον αφορά την κατασκευή των μοντέλων, τα βάρη των παρατηρήσεων πολλαπλασιάζονται με e^{a_i}. Αν το μοντέλο που μόλις δομήθηκε (M_i) είναι αρκετά ακριβές (e κοντά στο 0), τότε λιγότερες οντότητες έχουν καταταχθεί λανθασμένα, και τα βάρη τους αυξάνονται σημαντικά (π.χ., για ένα λάθος 5% το βάρος πολλαπλασιάζεται επί 19). Για ανακριβή μοντέλα (καθώς το e προσεγγίζει το 0,5) ο πολλαπλασιαστής για τα βάρη προσεγγίζει το 1. Σημειώστε ότι αν $e_i = 0.5$ ο πολλαπλασιαστής είναι 1, και, επομένως, δεν επέρχεται μεταβολή στα βάρη των οντοτήτων. Φυσικά, η δόμηση ενός μοντέλου για το ίδιο σύνολο δεδομένων με τα ίδια βάρη θα δομήσει το ίδιο μοντέλο, ως εκ τούτου τα κριτήρια για τη συνέχιση της δόμησης ενός μοντέλο ελέγχει αν $e < 0.5$(Williams, AdaBoost Algorithm, 2010).

6.3. Επίλογος

Στο κεφάλαιο αυτό περιγράφηκε η εφαρμογή της μηχανικής μάθησης στην εξόρυξη αλγορίθμων. Στο επόμενο κεφάλαιο θα γίνει η περιγραφή της τεχνικής των δένδρων απόφασης και των αλγορίθμων που χρησιμοποιούνται.

7. ΔΕΝΔΡΑ ΑΠΟΦΑΣΕΩΝ (DECISION TREES)

7.1. Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται οι πιο σημαντικοί αλγόριθμοι που χρησιμοποιούνται στην εξόρυξη δεδομένων και εφαρμόζουν τεχνικές δένδρων αποφάσεων. Στο σημείο αυτό παρατίθεται ξανά ο ορισμός του δένδρου απόφασης. Ένα **δέντρο απόφασης (decision tree)** είναι ένα μοντέλο πρόβλεψης που, όπως υποδηλώνει το όνομά του, μπορεί να θεωρηθεί ως ένα δέντρο. Συγκεκριμένα κάθε κλάδος του δέντρου είναι ένα ερώτημα ταξινόμησης και τα φύλλα του δέντρου είναι τα διαμερίσματα του συνόλου δεδομένων με την ταξινόμηση τους (Berson et al., 2010).

7.2. ID3

Ο αλγόριθμος **ID3 (Induction of Decision Trees)**, είναι ένα σύστημα επιβλεπόμενης μάθησης το οποίο κατασκευάζει κανόνες κατάταξης με τη μορφή ενός δέντρου απόφασης. Παίρνει ένα σύνολο αντικειμένων, το σύνολο

εκπαίδευσης, ως πρώτη ύλη, και δομεί το δέντρο αποφάσεων διαμερίζοντας το σύνολο εκπαίδευσης. Τα χαρακτηριστικά επιλέγονται για να χωρίσουν το σύνολο και δημιουργείται ένα δένδρο για κάθε υποσύνολο, έως ότου όλα τα μέλη των υποσυνόλων ανήκουν στην ίδια κατηγορία. Μια **ευρετική (heuristic)** συνάρτηση χρησιμοποιείται για να επιλέξετε το καλύτερο χαρακτηριστικό διαχωρισμού με βάση το σύνολο εκπαίδευσης. Ο ID3 είναι ένας **άπληστος (greedy)** αλγόριθμος, και, επομένως, μια κακή επιλογή χαρακτηριστικού, μπορεί να επηρεάσει το τελικό αποτέλεσμα. Η αρχική έκδοση του ID3 χειρίζεται μόνο ένα μικρό αριθμό διακριτών τιμών, αλλά αργότερα τροποποιήθηκε ώστε να χειριστεί τόσο διατεταγμένα όσο και συνεχή χαρακτηριστικά. Άλλες παραλλαγές του αλγορίθμου ID3 περιλαμβάνουν την ικανότητα χειρισμού του θορύβου. Χρησιμοποιώντας μια δοκιμή, μόνο τα χαρακτηριστικά των οποίων η μη σχετικότητα μπορεί να απορριφθεί χρησιμοποιούνται για τον διαχωρισμό του συνόλου. Εάν τα χαρακτηριστικά λείπουν κατά τη διάρκεια της κατάταξης, εξερευνούνται όλα τα πιθανά κλαδιά και επιλέγεται η πιο πιθανή κατάταξη (Aasheim et al., 1996).

Ο ID3 αναπτύχθηκε από τον Quinlan και εν κατακλείδι, είναι ένας αλγόριθμος που έχει υψηλή ακρίβεια ταξινόμησης ακόμα και σε θορυβώδη σύνολα δεδομένων. Διαθέτει ένα γρήγορο στάδιο μάθησης, και χαμηλή πολυπλοκότητα χρόνου. Ο ID3 πρέπει να διαθέτει όλο το σύνολο εκπαίδευσης, αλλά υπάρχουν και διαφοροποιήσεις όσον αφορά την αυξητική εκμάθηση. Το δένδρο απόφασης που προκύπτει από τον ID3 δεν είναι πολύ απλό για τους ανθρώπους όταν χρησιμοποιούνται μεγάλες ποσότητες δεδομένων (Aasheim et al., 1996).

Ένα θεμελιώδες στοιχείο οποιουδήποτε αλγορίθμου κατασκευάζει ένα δέντρο απόφασης από ένα σύνολο δεδομένων είναι η μέθοδος με την οποία επιλέγει χαρακτηριστικά σε κάθε κόμβο του δένδρου. Μερικά χαρακτηριστικά διασπούν τα δεδομένα περισσότερο καθαρά από τα άλλα. Αυτό σημαίνει ότι οι τιμές τους αντιστοιχούν με μεγαλύτερη συνέπεια στις περιπτώσεις που έχουν ιδιαίτερη αξία στο χαρακτηριστικού στόχου (το ένα που θέλουμε να προβλέψουμε) από εκείνες ενός άλλου χαρακτηριστικού. Ως εκ τούτου, θα μπορούσαμε να πούμε ότι τέτοιες ιδιότητες έχουν κάποια υποκείμενη σχέση με το χαρακτηριστικό στόχο. Αλλά πώς μπορεί αυτό να ποσοτικοποιηθούν με κάποιο τρόπο; Ουσιαστικά, θα θέλαμε κάποια μέτρο που μας επιτρέπει να συγκρίνετε τις ιδιότητες μεταξύ τους

και στη συνέχεια να μπορεί να αποφασίσει να βάλει αυτούς που χωρίζουν τα δεδομένα πιο καθαρά πιο ψηλά στο δέντρο (decisiontrees.net).

Ένα μέτρο που χρησιμοποιείται από την Θεωρία Πληροφορίας στον αλγόριθμο ID3 και πολλούς άλλους που χρησιμοποιούνται στην κατασκευή δένδρων αποφάσεων είναι η **Εντροπία (Entropy)**. Ανεπίσημα, η εντροπία ενός συνόλου δεδομένων μπορεί να θεωρηθεί πως είναι διαταραγμένη. Έχει αποδειχθεί ότι η εντροπία σχετίζεται με τις πληροφορίες, υπό την έννοια ότι όσο μεγαλύτερη είναι η εντροπία, ή αβεβαιότητα, από κάποια δεδομένα, τότε απαιτούνται περισσότερες πληροφορίες προκειμένου να περιγράψουν πλήρως τα δεδομένα. Στην δόμηση ενός δένδρου αποφάσεων, στόχος είναι να μειωθεί η εντροπία του συνόλου δεδομένων, μέχρι να γίνουν δικτυακοί κόμβοι στο σημείο που το υποσύνολο που μένει είναι καθαρό, ή έχει μηδενική εντροπία και αντιπροσωπεύει όλες τις περιπτώσεις σε μία τάξη (όλες οι περιπτώσεις έχουν την ίδια τιμή για το χαρακτηριστικό-στόχο). Η εντροπία ενός συνόλου δεδομένων, S, σε σχέση με ένα γνώρισμα, στην περίπτωση αυτή το χαρακτηριστικό-στόχο, γίνεται με τον ακόλουθο υπολογισμό (decisiontrees.net):

$$Entropy(s) = -\sum_{i=1}^{c} p_i \log_2 p_i \qquad (7.1)$$

όπου P_i είναι η αναλογία των στιγμιότυπων στο σύνολο δεδομένων που παίρνουν την i-οστή τιμή στο χαρακτηριστικό στόχο, που έχει C διαφορετικές τιμές. Αυτά τα μέτρα πιθανότητας δίνουν μια ένδειξη για το πόση αβεβαιότητα υπάρχει σχετικά με τα δεδομένα και χρησιμοποιείται ένα μέτρο $\log_2$ καθώς αυτό αντιπροσωπεύει το πόσα bits θα πρέπει να χρησιμοποιηθούν για να προσδιορίσουν αν η κλάση (τιμή του χαρακτηριστικού στόχου) είναι ένα τυχαίο παράδειγμα. Στη συνέχεια υπάρχει ανάγκη ύπαρξης ενός ποσοτικού τρόπου για εμφάνιση του αποτελέσματος της διαίρεσης του συνόλου δεδομένων, χρησιμοποιώντας ένα συγκεκριμένο χαρακτηριστικό (το οποία αποτελεί μέρος της διαδικασίας οικοδόμησης δέντρο). Μπορεί να χρησιμοποιηθεί ένα μέτρο που ονομάζεται **Κέρδος Πληροφορίας (Information Gain)**, το οποίο υπολογίζει την μείωση της εντροπίας (κέρδος σε πληροφορίες) που θα έχει ως αποτέλεσμα την κατάτμηση των δεδομένων για μια ιδιότητα, A (decisiontrees.net):

$$Gain(S,A) = Entropy(S) - \sum_{v \in A} \frac{|S_v|}{|S|} Entropy(S_v) \quad (7.2)$$

όπου v είναι μια τιμή του A, $|S_v|$ είναι το υποσύνολο των στιγμιότυπων του S όπου το A παίρνει την τιμή v και |S| είναι ο αριθμός των στιγμιότυπων.

Στον Πίνακα 5 παρουσιάζεται ο ψευδοκώδικας του βασικού αλγορίθμου ID3 (decisiontrees.net):

Πίνακας 5: Ο αλγόριθμος ID3 (decisiontrees.net).

Είσοδος: Ένα σύνολο δεδομένων, S

Έξοδος: Ένα δένδρο απόφασης

Αν όλα τα στιγμιότυπα έχουν την ίδια αξία για το χαρακτηριστικό στόχο, τότε ένα δέντρο απόφασης είναι απλά αυτή η τιμή (δεν είναι πραγματικά ένα δέντρο – απλά ένας κορμός δένδρου).

Αλλιώς

1. Υπολογίζονται οι τιμές Gain για όλες τις ιδιότητες και επιλέγεται μια ιδιότητα με την υψηλότερη αξία και να δημιουργήσει ένα κόμβο για το γνώρισμα αυτό.
2. Γίνεται ένας κλάδος από αυτόν τον κόμβο για κάθε τιμή του χαρακτηριστικού
3. Εκχωρούνται όλες οι δυνατές τιμές του χαρακτηριστικού στα κλαδιά.
4. Ακολουθείται κάθε κλάδος μέσω της κατάτμησης του συνόλου δεδομένων ώστε να είναι μόνο στιγμιότυπα όπου η αξία του κλάδου είναι παρούσα και στη συνέχεια επιστροφή στο 1.

Παράδειγμα

Πίνακας 9 : Πίνακας δεδομένων του παραδείγματος κατηγοριοποίησης ατόμων στις κατηγορίες "κοντός", "ψηλός", "μέτριος"

Όνομα	Φύλο	Ύψος (μ)	Κατηγοριοποίηση 1	Κατηγοριοποίηση 2
Kristina	Θ	1,6	Κοντός	Μέτριος
Jim	Α	2	Ψηλός	Μέτριος
Maggie	Θ	1,9	Μέτριος	Ψηλός
Martha	Θ	1,88	Μέτριος	Ψηλός
Stephanie	Θ	1,7	Κοντός	Μέτριος
Bob	Α	1,85	Μέτριος	Μέτριος
Kathy	Θ	1,6	Κοντός	Μέτριος
Dave	Α	1,7	Κοντός	Μέτριος
Worth	Α	2,2	Ψηλός	Ψηλός
Steven	Α	2,1	Ψηλός	Ψηλός
Debbie	Θ	1,8	Μέτριος	Μέτριος
Todd	Α	1,95	Μέτριος	Μέτριος
Kim	Θ	1,9	Μέτριος	Ψηλός
Amy	Θ	1,8	Μέτριος	Μέτριος
Wynette	Θ	1,75	Μέτριος	Μέτριος

Τα δεδομένα εκπαίδευσης του πίνακα 3 (με την κατηγοριοποίηση 1) δείχνουν ότι 4/15 είναι κοντοί, 8/15 είναι μέτριοι και 3/15 είναι ψηλοί. Έτσι η εντροπία του αρχικού συνόλου είναι:

$$4/15 \log(15/4) + 8/15 \log(15/8) + 3/15 \log(15/3) = 0.4384$$

Επιλέγοντας το χαρακτηριστικό «φύλο» ως χαρακτηριστικό διάσπασης έχουμε 9 πλειάδες που είναι Θ και 6 που είναι Α. Η εντροπία του υποσυνόλου που είναι Θ είναι:

$$3/9 \log(9/3) + 6/9 \log(9/6) = 0.2764$$

Η εντροπία για τις πλειάδες Α είναι:

$$1/6 \log (6/1) + 2/6 \log(6/2) + 3/6 \log(6/3) = 0.4392$$

Ο ID3 πρέπει να καθορίσει ποιο είναι το κέρδος πληροφορίας χρησιμοποιώντας αυτή την διάσπαση. Για να γίνει αυτό υπολογίζει το σταθμισμένο άθροισμα των 2 παραπάνω εντροπιών.

$$((9/15)\ 0.2764) + ((6/15)\ 0.4392) = 0.34152$$

Άρα το κέρδος πληροφορίας χρησιμοποιώντας το χαρακτηριστικό «φύλο» είναι

$$0.4384 - 0.34152 = 0.09688$$

Τώρα κάνουμε την ίδια διαδικασία για το χαρακτηριστικό «ύψος». Έχουμε 2 άτομα που έχουν ύψος 1.6, 2 άτομα που έχουν ύψος 1.7, 1 άτομο με ύψος 1.75, 2 άτομα με ύψος 1.8, 1 άτομο με ύψος 1.85, 1 άτομο με ύψος 1.88, 2 άτομα με ύψος 1.9, 1 άτομο με ύψος 1.95, 1 άτομο με ύψος 2, 1 άτομο με ύψος 2.1 και 1 άτομο με ύψος 2.2. Όπως καταλαβαίνουμε θα είναι καλύτερο το να χωρίσουμε τα δεδομένα σε διαστήματα. Ο διαχωρισμός αυτός πρέπει να γίνει από έναν ειδικό στο πεδίο του προβλήματος. Τα διαστήματα που προκύπτουν από έναν διαχωρισμό θα μπορούσαν να είναι τα εξής:

$$(0,1.6],\ (1.6,\ 1.7],\ (1.7,\ 1.8],\ (1.8,\ 1.9],\ (1.9,\ 2],\ (2,\ \infty)$$

Υπάρχουν 2 πλειάδες στο πρώτο διάστημα με εντροπία (2/2(0)+0+0) = 0, 2 στο διάστημα (1.6,1.7] με εντροπία (0+4/4(0)+0)=0, 3 στο (1.7,1.8] με εντροπία (0 +3/3(0)+0) = 0, 4 στο (1.8,1.9] με εντροπία (0+4/4(0)+0)=0, 2 στο (1.9,2] με εντροπία (0 + 1/2(0.301) + 1/2(0.301) = 0.301 και 2 πλειάδες στο (2, ∞) με εντροπία (0 + 0 + 2/2(0)) = 0. όλες οι καταστάσεις είναι εντελώς διατεταγμένες και επομένως έχουν εντροπία 0, εκτός αυτή που αντιστοιχεί στο διάστημα (1.9, 2]. Το κέρδος με την χρήστη του χαρακτηριστικού «ύψος» είναι:

$$0.4384 – 2/15(0.301) = 0.3983$$

καταλαβαίνουμε ότι η διάσπαση χρησιμοποιώντας το χαρακτηριστικό «ύψος» είναι καλύτερη από την διάσπαση βάσει το χαρακτηριστικό «φύλο» αφού αυτή έχει μεγαλύτερο κέρδος.

7.3. C4.5

Ο C4.5 είναι ένας αλγόριθμος που χρησιμοποιείται για να παράγει ένα δέντρο απόφασης και αναπτύχθηκε από τον Ross Quinlan. Ο C4.5 είναι μια επέκταση του προγενέστερου ID3 αλγορίθμου. Τα δέντρα αποφάσεων που προκύπτουν από τον C4.5 μπορούν να χρησιμοποιηθούν για την ταξινόμηση, και για το λόγο αυτό, ο C4.5 αναφέρεται συχνά ως στατιστικός ταξινομητής (Classle.net, 2009).

Ο C4.5 δομεί δέντρα απόφασης από ένα σύνολο δεδομένων εκπαίδευσης κατά τον ίδιο τρόπο όπως ο ID3, χρησιμοποιώντας την έννοια της εντροπίας των πληροφοριών. Τα δεδομένα εκπαίδευσης είναι ένα σύνολο $S=s_1,s_2,...$ των δειγμάτων που έχουν ήδη ταξινομηθεί. Κάθε δείγμα $s_i=x_1,x_2,...$ είναι ένα διάνυσμα όπου $x_1,x_2,...$ αντιπροσωπεύουν γνωρίσματα ή χαρακτηριστικά του δείγματος. Τα δεδομένα εκπαίδευσης αυξάνονται με ένα διάνυσμα $C=c_1,c_2,...$ όπου $C_1,C_2,...$ αντιπροσωπεύει την κλάση στην οποία αντιστοιχεί κάθε δείγμα. Σε κάθε κόμβο του δένδρου, ο C4.5 επιλέγει ένα χαρακτηριστικό των στοιχείων που χωρίζει πλέον αποτελεσματικά το σύνολο των δειγμάτων σε υποσύνολα εμπλουτισμένα σε μια τάξη ή σε μια άλλη. Το κριτήριο του είναι το κανονικοποιημένο κέρδος πληροφοριών (διαφορά στην εντροπία) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Το χαρακτηριστικό με το υψηλότερο κέρδος κανονικοποιημένης πληροφορίας επιλέγεται να λάβει την απόφαση. Ο αλγόριθμος C4.5 τότε κάνει αναδρομή στους μικρότερους υποκαταλόγους (Classle.net, 2009).

Ο αλγόριθμος έχει μερικές βασικές περιπτώσεις (Classle.net, 2009).

- Όλα τα δείγματα της λίστας ανήκουν στην ίδια τάξης. Όταν συμβαίνει αυτό, απλώς δημιουργείται ένας κόμβο για το δέντρο απόφασης λέγοντας να γίνει επιλογή αυτής της τάξης.
- Κανένα από τα χαρακτηριστικά δεν παρέχει κέρδος πληροφοριών. Στην περίπτωση αυτή, ο C4.5 δημιουργεί έναν κόμβο απόφασης πιο ψηλά στο δέντρο με την αναμενόμενη τιμή της τάξης.
- Αντιμετωπίζεται στιγμιότυπο της απαρατήρητης τάξης. Και πάλι, ο C4.5 δημιουργεί έναν κόμβο απόφασης πιο ψηλά στο δέντρο με την αναμενόμενη τιμή.

Σε ψευδοκώδικα, ο αλγόριθμος παρουσιάζεται στον Πίνακα 6 (Classle.net, 2009):

Πίνακας 6: Ο αλγόριθμος C4.5 (Classle.net, 2009).

Έλεγχος για βασικές περιπτώσεις

Για κάθε γνώρισμα a

Εύρεση του κανονικοποιημένου κέρδους πληροφορίας διαχωρίζοντας στο a

Έστω a_{best} το γνώρισμα με το υψηλότερο κανονικοποιημένο κέρδους πληροφορίας

Δημιουργία ενός κόμβου απόφασης που διαχωρίζει στο a_{best}

Αναδρομή στις υπολίστες που αποκτήθηκαν διαχωρίζοντας στο a_{best} και πρόσθεση των κόμβων αυτών ως παιδιά του κόμβου

Ο C4.5 έκανε έναν αριθμό βελτιώσεων στον ID3. Μερικές από αυτές είναι (Classle.net, 2009):

- Χειρισμός συνεχών και διακριτών χαρακτηριστικών - για να χειριστεί τα συνεχή χαρακτηριστικά, ο C4.5 δημιουργεί ένα όριο και, στη συνέχεια, χωρίζει τον κατάλογο σε εκείνα των οποίων η τιμή γνωρίσματος υπερβαίνει το όριο και εκείνα που είναι μικρότερη ή ίση με αυτό.
- Χειρισμός δεδομένων εκπαίδευσης με τιμές γνωρισμάτων που λείπουν. Ο C4.5 επιτρέπει τις τιμές παραμέτρων να επισημανθούν ως «;» για όσα

λείπουν. Οι αγνοούμενες τιμές γνωρισμάτων απλά δεν χρησιμοποιούνται στους υπολογισμούς κέρδους και εντροπίας.

- Χειρισμός γνωρισμάτων με διαφορετικό κόστος.
- Κλάδεμα δένδρων μετά τη δημιουργία. Ο C4.5 πηγαίνει πίσω στο δέντρο μια φορά αφού δημιουργηθεί και προσπαθεί για την απομάκρυνση των κλαδιών που δεν βοηθούν με την αντικατάστασή τους με κόμβους φύλλα.

7.4. CART

Ο CART (Classification and Regression Trees) είναι ένας αλγόριθμος εξερεύνησης δεδομένων και πρόβλεψης που αναπτύχθηκε από τους Leo Breiman, Jerome Friedman, Richard Olshen και Charles Stone. Στη δόμηση του δέντρου CART κάθε προγνωστικό μπορεί να διαβαστεί με βάση το πόσο καλά χωρίζει τις εγγραφές με διαφορετικές προβλέψεις. Για παράδειγμα, ένα μέτρο που χρησιμοποιείται για να καθορίσει αν ένα δεδομένη σημείο διαχωρισμού ένα προγνωστικό παράγοντα είναι καλύτερο από ένα άλλο είναι η μετρική της εντροπίας (Berson et al., 2010).

Το δέντρο απόφασης CART είναι μια δυαδική επαναληπτική διαδικασία διαχωρισμού με δυνατότητα επεξεργασίας συνεχών και ονομαστικών χαρακτηριστικών, τόσο ως στόχοι όσο και ως μέθοδοι πρόβλεψης. Τα δεδομένα χρησιμοποιούνται στην ακατέργαστη μορφή τους. Δομούνται δένδρα σε ένα μέγιστο μέγεθος χωρίς τη χρήση ενός κανόνα διακοπής και στη συνέχεια κλαδεύονται στη ρίζα μέσω περικοπής κόστους πολυπλοκότητας. Η επόμενη διάσπαση για κλάδεμα είναι αυτή που συμβάλλει λιγότερο στις συνολικές επιδόσεις του δέντρου στα δεδομένα εκπαίδευσης (και περισσότερες από μία διασπάσεις μπορούν να αφαιρεθούν σε μια στιγμή). Η διαδικασία παράγει τα δέντρα που είναι αναλλοίωτα σε οποιαδήποτε σειρά, διατηρώντας τη μετατροπή της πρόβλεψης χαρακτηριστικών. Ο μηχανισμός CART έχει ως στόχο να παράγει όχι ένα, αλλά μια ακολουθία ένθετων κλαδεμένων δέντρων, τα οποία είναι υποψήφια βέλτιστα δέντρα. Το «σωστού μεγέθους» δέντρο αναγνωρίζεται από την αξιολόγηση της πρόβλεψης της απόδοσης κάθε δέντρου στην ακολουθία κλαδέματος (Wu et al., 2008).

Στη συνέχεια γίνεται μια περιγραφή του αλγορίθμου CART. Η βασική ιδέα είναι όπως περιγράφηκε προηγουμένως να γίνει μια επιλογή διάσπασης ανάμεσα σε όλες τις δυνατές διασπάσεις σε κάθε κόμβο έτσι ώστε οι προκύπτοντες κόμβοι απόγονοι να είναι οι βέλτιστοι. Στον αλγόριθμο αυτό θεωρούνται μόνο μονοπαραγοντικές διασπάσεις, δηλαδή κάθε διάσπαση εξαρτάται από την τιμή μόνο μιας μεταβλητής πρόβλεψης. Όλες οι δυνατές διασπάσεις αποτελούνται από πιθανές διασπάσεις της κάθε πρόβλεψης. Αν X είναι μια ονομαστική κατηγορική μεταβλητή I κατηγοριών, υπάρχουν 2^{I-1}-1 πιθανές διασπάσεις για την πρόβλεψη. Αν X είναι μια μεταβλητή κατηγορική τάξης ή συνεχής με K διαφορετικές τιμές, υπάρχουν K-1 διαφορετικές διασπάσεις στο X. Ένα δένδρο δομείται ξεκινώντας από τον κόμβο ρίζας χρησιμοποιώντας επαναλαμβανόμενα τα ακόλουθα βήματα σε κάθε κόμβο όπως παρουσιάζονται στον Πίνακα 7 (SPSS.com):

Πίνακας 7:Ο αλγόριθμος CART (SPSS.com).

1. Εύρεση καλύτερης διάσπασης της κάθε πρόβλεψης. Για κάθε πρόβλεψη συνεχή και τάξης, ταξινομούνται οι τιμές από τη μικρότερη στη μεγαλύτερη. Για κάθε ταξινομημένη πρόβλεψη, γίνεται διάσχιση σε κάθε τιμή από την κορυφή για εξέταση του κάθε υποψήφιου σημείου διάσπασης (δηλαδή v, αν $x \le v$, η περίπτωση πηγαίνει στον αριστερό κόμβο παιδί, αλλιώς πηγαίνει στον δεξιό) για να καθοριστεί το καλύτερο. Το καλύτερο σημείο διάσπασης είναι εκείνο που μεγιστοποιεί το κριτήριο διαχωρισμού περισσότερο όταν ο κόμβος διασπάται κατά μήκος του. Για κάθε ονομαστική πρόβλεψη, γίνεται εξέταση κάθε πιθανού υποσυνόλου κατηγοριών (δηλαδή A, αν $x \in A$, η περίπτωση πηγαίνει στον αριστερό κόμβο παιδί, αλλιώς πηγαίνει στον δεξιό) για να βρεθεί η καλύτερη διάσπαση.
2. Εύρεση της καλύτερης διάσπασης του κόμβου. Ανάμεσα στις καλύτερες διασπάσεις που βρέθηκαν στο βήμα 1, επιλογή εκείνης που μεγιστοποιεί το κριτήριο διαχωρισμού.
3. Διάσπαση του κόμβου χρησιμοποιώντας την καλύτερη διάσπαση που βρέθηκε στο βήμα 2 αν δεν ικανοποιούνται οι κανόνες σταματήματος.

Ένα από τα μεγάλα πλεονεκτήματα του CART είναι ότι ο αλγόριθμος έχει την επικύρωση του μοντέλου και την ανακάλυψη του βέλτιστο γενικού μοντέλου βαθιά δομημένη μέσα στον αλγόριθμο. Ο CART το πετυχαίνει αυτό με την δόμηση ενός πολύ σύνθετου δέντρου και στη συνέχεια με το κλάδεμα πίσω στο βέλτιστο γενικό δέντρο με βάση τα αποτελέσματα της πολλαπλής επικύρωσης ή της επικύρωσης του συνόλου δοκιμής. Το δέντρο κλαδεύεται πίσω με βάση τις επιδόσεις των διαφόρων κλαδεμένων εκδόσεων του δέντρου στο σύνολο δεδομένων δοκιμής. Το πιο σύνθετο δέντρο σπάνια είναι το καλύτερο για τα δεδομένα καθώς έχει ταιριάξει υπερβολικά στα δεδομένα εκπαίδευσης. Με τη χρήση επικύρωσης το δέντρο που είναι πιθανότερο να είναι το καλύτερο μπορεί να επιλεγεί (Berson et al., 2010).

Ο αλγόριθμος CART είναι σχετικά ισχυρός όσον αφορά τα ελλείποντα στοιχεία. Εάν η τιμή για μια συγκεκριμένη πρόγνωση λείπει σε μια συγκεκριμένη εγγραφή η εγγραφή αυτή δεν θα χρησιμοποιηθεί για την παρασκευή του προσδιορισμού της βέλτιστης διάσπασης όταν το δέντρο είναι υπό κατασκευή. Στην πραγματικότητα ο CART αξιοποιεί όσο περισσότερες πληροφορίες έχει σε ετοιμότητα, προκειμένου να λάβει την απόφαση για την επιλογή της καλύτερης δυνατής διάσπασης. Όταν ο CART χρησιμοποιείται για την πρόβλεψη σχετικά με τα νέα δεδομένα, οι τιμές που λείπουν μπορούν να αντιμετωπιστούν μέσω υποκατάστατων. Τα υποκατάστατα χωρίζουν τις τιμές και τους προγνωστικούς παράγοντες που μιμούνται την πραγματική διάσπαση στο δέντρο και μπορούν να χρησιμοποιηθούν όταν τα δεδομένα για την προτιμώμενη πρόβλεψης λείπουν (Berson et al., 2010).

7.5. CHAID

Μια εξίσου δημοφιλής τεχνολογία δέντρου απόφασης όπως ο CART είναι ο αλγόριθμος CHAID ή Chi-Square Automatic Interaction Detector. Ο CHAID είναι παρόμοιος με τον CART υπό την έννοια ότι δημιουργεί ένα δέντρο απόφασης αλλά διαφέρει στον τρόπο που επιλέγει τις διασπάσεις. Αντί της μέτρησης εντροπίας ή Gini για την βέλτιστη επιλογή χωρίζει με βάση την δοκιμή chi square που χρησιμοποιείται για πίνακες συνάφειας για να καθορίσει ποια κατηγορική πρόβλεψη απέχει περισσότερο από την ανεξαρτησία με τις τιμές πρόβλεψης.

Επειδή ο CHAID βασίζεται στους πινάκες συνάφειας για να σχηματίσει τη δοκιμή σημαντικότητας για κάθε πρόβλεψη όλες οι προβλέψεις πρέπει είτε να είναι κατηγορικές ή να εξαναγκάζονται σε κατηγορική μορφή μέσω της μεθόδου binning. Μέσω του binning μπορούν να υπάρξουν καταστροφικές συνέπειες στις πραγματικές επιδόσεις σε ακρίβεια του CART και ο CHAID έχουν αποδειχθεί να είναι συγκρίσιμος σε πραγματικού κόσμου άμεσα μοντέλα απόκρισης μάρκετινγκ (Berson et al., 2010).

Στη συνέχεια γίνεται μια περιγραφή του απλού αλγορίθμου CHAID. Ο αλγόριθμος αυτός δέχεται μόνο προβλέψεις ονομαστικές ή κατηγορικές τάξης. Όταν οι προβλέψεις είναι συνεχείς, μετατρέπονται σε προβλέψεις τάξης πριν χρησιμοποιηθούν στον αλγόριθμο, όπως περιγράφεται στον Πίνακα 8 (SPSS.com):

Πίνακας 8:Ο αλγόριθμος CHAID (SPSS.com).

Συγχώνευση

Για κάθε μεταβλητή πρόβλεψης X. γίνεται συγχώνευσης μη σημαντικών κατηγοριών. Κάθε τελική κατηγορία του X θα έχει ως αποτέλεσμα ένα κόμβο παιδί αν το X χρησιμοποιείται για να διασπάσει τον κόμβο. Το βήμα συγχώνευσης υπολογίζει επίσης την προσαρμοσμένη τιμή p που θα χρησιμοποιηθεί στο βήμα διάσπασης.

1. Αν το X έχει 1 κατηγορία μόνο, γίνεται διακοπή και η προσαρμοσμένη τιμή p γίνεται 1.
2. Αν το X έχει 2 κατηγορίες, μεταφορά στο βήμα 8.
3. Αλλιώς, εύρεση του επιτρεπόμενου ζεύγους κατηγοριών του X (ένα επιτρεπόμενο ζεύγος κατηγοριών για πρόβλεψη τάξης είναι δύο προσκείμενες κατηγορίες και για ονομαστική πρόβλεψη είναι δυο οποιεσδήποτε κατηγορίες) που είναι το λιγότερο σημαντικά διαφορετικά (δηλαδή πιο όμοια). Το ποιό όμοιο ζεύγος είναι το ζεύγος του οποίου η στατιστική δοκιμή δίνει τη μέγιστη τιμή p όσον αφορά την εξαρτημένη μεταβλητή Y.
4. Για το ζεύγος με τη μεγαλύτερη τιμή p, έλεγχος αν η τιμή p είναι

μεγαλύτερη από ένα καθορισμένο από το χρήστη επίπεδο άλφα α_{merge}(alpha_merge). Αν είναι, το ζεύγος αυτό συγχωνεύεται σε μια απλή κατηγορία συστατικών. Τότε δημιουργείται ένα νέο σύνολο κατηγοριών X. Αν όχι, μεταφορά στο βήμα 7.

5. (Προαιρετικό) Αν η νέα σχηματισμένη κατηγορία συστατικών αποτελείται από τρεις ή περισσότερες πρωτότυπες κατηγορίες, εύρεση της βέλτιστης δυαδικής διάσπασης μέσα στην κατηγορία συστατικών όπου η τιμή p είναι η μικρότερη. Η δυαδική αυτή διάσπαση γίνεται αν η τιμή p δεν είναι μεγαλύτερη από μια τιμή άλφα $\alpha_{split\text{-}merge}$(alpha_spli-merge).
6. Μεταφορά στο βήμα 2.
7. (Προαιρετικό) Κάθε κατηγορία που έχει πολύ λίγες παρατηρήσεις (σε σύγκριση με ένα καθορισμένο από το χρήστη ελάχιστο μέγεθος τμήματος) συγχωνεύεται με την πιο παραπλήσια κατηγορία καθώς μετράται από τη μέγιστη των τιμών p.
8. Η προσαρμοσμένη τιμή p υπολογίζεται για τις συγχωνευμένες κατηγορίες εφαρμόζοντας τις προσαρμογές Bonferroni.

Διάσπαση

Η «καλύτερη» διάσπαση για κάθε πρόβλεψη βρίσκεται στο βήμα συγχώνευσης. Το βήμα διάσπασης επιλέγει ποια πρόβλεψη θα χρησιμοποιηθεί για να διασπάσει καλύτερα τον κόμβο. η επιλογή επιτυγχάνεται συγκρίνοντας την προσαρμοσμένη τιμή p που σχετίζεται με κάθε πρόβλεψη. Η προσαρμοσμένη τιμή p υπολογίζεται στο βήμα συγχώνευσης.

1. Επιλογή της πρόβλεψης που έχει την μικρότερη προσαρμοσμένη τιμή p (δηλαδή την πιο σημαντική).
2. Αν η προσαρμοσμένη τιμή p είναι μικρότερη ή ίση από ένα καθορισμένο από το χρήστη α_{split}(alpha_split), διάσπαση του κόμβου χρησιμοποιώντας την πρόβλεψη. Αλλιώς, δεν γίνεται διάσπαση και ο κόμβος θεωρείται ως ένας τερματικός κόμβος.

Τερματισμός

Το βήμα τερματισμού ελέγχει αν η διαδικασία δόμησης του δένδρου θα έπρεπε να τερματιστεί σύμφωνα με τους ακόλουθους κανόνες τερματισμού.

1. Αν ένας κόμβος γίνεται καθαρός. Όλες οι περιπτώσεις σε έναν κόμβο έχουν ίδιες τιμές στην εξαρτημένη μεταβλητή και ο κόμβος δεν διασπάται.

2. Αν όλες οι περιπτώσεις σε έναν κόμβο έχουν ίδιες τιμές για κάθε πρόβλεψη, ο κόμβος δεν διασπάται.

3. Αν το τρέχον βάθος δένδρου φτάνει την καθορισμένη από το χρήστη μέγιστη οριακή τιμή βάθους δένδρου, ο κόμβος δεν διασπάται.

4. Αν η διάσπαση ενός κόμβου σε έναν κόμβο παιδί του οποίου το μέγεθος κόμβου είναι λιγότερο από την καθορισμένη από το χρήστη ελάχιστη τιμή μεγέθους κόμβου παιδιού, οι κόμβοι παιδιά που έχουν πολύ λίγες περιπτώσεις (σε σύγκριση με αυτό το ελάχιστο) θα συγχωνευτούν με τον πιο όμοιο κόμβο παιδί όπως μετράται από τη μέγιστη από τις τιμές p. Παρόλα αυτά, αν ο προκύπτων αριθμός κόμβων παιδιών είναι 1, ο κόμβος δεν διασπάται.

7.6. Επίλογος

Στο κεφάλαιο αυτό περιγράφηκαν οι βασικότεροι αλγόριθμοι δένδρων απόφασης. Στη συνέχεια θα γίνει περιγραφή μιας άλλης κατηγορίας αλγορίθμων που αφορούν τα παράλληλα και κατανεμημένα συστήματα.

8. ΚΑΤΑΝΕΜΗΜΕΝΟΙ ΚΑΙ ΠΑΡΑΛΛΗΛΟΙ ΑΛΓΟΡΙΘΜΟΙ

8.1. Εισαγωγή

Τα τελευταία χρόνια, υπάρχει ένα αυξανόμενο ενδιαφέρον για την έρευνα των **παράλληλων αλγορίθμων εξόρυξης δεδομένων (parallel data mining algorithms)**. Σε παράλληλο περιβάλλον, αξιοποιώντας τη μεγάλη συνολική κύρια μνήμη και την επεξεργαστική ισχύ των παράλληλων επεξεργαστών, οι παράλληλοι αλγόριθμοι μπορούν να έχουν και τον χρόνο εκτέλεσης και τα ζητήματα απαιτούμενης μνήμης σωστά διευθετημένα. Ωστόσο, δεν είναι ασήμαντη η παραλληλοποίηση των αλγορίθμων για την επίτευξη καλών επιδόσεων καθώς και την επεκτασιμότητα για τα μαζικά σύνολα δεδομένων. Πρώτον, είναι σημαντικό να σχεδιαστεί μια καλή οργάνωση δεδομένων και μια στρατηγική αποσύνθεσης, έτσι ώστε ο φόρτος εργασίας να είναι ομοιόμορφα κατανεμημένος μεταξύ όλων των διαδικασιών με ελάχιστη εξάρτηση από δεδομένα σε αυτούς. Δεύτερον, η ελαχιστοποίηση του συγχρονισμού ή/και γενικά η επικοινωνία είναι σημαντική, προκειμένου ο παράλληλος αλγόριθμος να κλιμακωθεί καλά καθώς ο αριθμός των διαδικασιών αυξάνεται. Η κατανομή του φόρτου εργασίας πρέπει επίσης να σχεδιαστεί προσεκτικά. Τέλος, το κόστος I/O δίσκου πρέπει να ελαχιστοποιηθεί (Li et al). Η **κατανεμημένη εξόρυξη δεδομένων (distributed data mining)**

ασχολείται με την εφαρμογή της κλασικής διαδικασίας εξόρυξης δεδομένων σε ένα κατανεμημένο υπολογιστικό περιβάλλον προσπαθώντας να χρησιμοποιήσει τους καλύτερους διαθέσιμους πόρους (δίκτυο επικοινωνίας, υπολογιστικές μονάδες και βάσεις δεδομένων). Η εξόρυξη δεδομένων γίνεται και τοπικά σε κάθε κατανεμημένη τοποθεσία και σε καθολικό επίπεδο όπου η τοπική γνώση διαχέεται για να ανακαλυφθεί η καθολική γνώση (Tsoumakas G., 2008).

8.2. FDM

Η κατανεμημένη εξόρυξη δεδομένων αλγόριθμοι βελτιστοποιεί την ανταλλαγή δεδομένων που απαιτούνται για την ανάπτυξη καθολικών μοντέλων που βασίζονται στις γνώσεις παράλληλης εξόρυξης απομακρυσμένων συνόλων δεδομένων. Ένας κατανεμημένος αλγόριθμος εξόρυξης δεδομένων, ο **FDM (Fast Distributed Mining of association rules)** παρουσιάζει τα ακόλουθα ιδιαίτερα χαρακτηριστικά. Η δημιουργία των υποψήφιων συνόλων είναι στο ίδιο πνεύμα με τον Apriori. Ωστόσο, ορισμένες σχέσεις μεταξύ των κατά τόπους μεγάλων συνόλων και σε καθολικό επίπεδο μεγάλων συνόλων πρέπει να διερευνηθούν για να δημιουργήσουν ένα μικρότερο σύνολο από υποψήφια σύνολα και έτσι να μειωθεί ο αριθμός των μηνυμάτων που θα περάσουν. Αφού τα υποψήφια σύνολα έχουν δημιουργηθεί, οι δύο τεχνικές κλαδέματος, το τοπικό κλάδεμα και το καθολικό κλάδεμα, αναπτύσσονται για να κλαδέψουν ορισμένα υποψήφια σύνολα σε κάθε επιμέρους τόπο. Ο αλγόριθμος FDM υλοποιείται με πλέγμα. Σε κάθε κόμβο του δικτύου, ο FDM βρίσκει την τοπική υποστήριξη, μετράει και κλαδεύει όλα τα σπάνια σύνολα στοιχείων. Μετά την ολοκλήρωση των τοπικών κλαδεμάτων, κάθε κόμβος στο δίκτυο μεταδίδει μηνύματα που περιέχουν όλα τα υπόλοιπα υποψήφια σύνολα για όλους τους άλλους κόμβους στο δίκτυο για να ζητήσουν την μέτρηση της υποστήριξή τους. Αποφασίζει τότε κατά πόσο τα μεγάλα στοιχειοσύνολα είναι καθολικά συχνά και δημιουργεί τα υποψήφια στοιχειοσύνολα από τα καθολικά συχνά στοιχειοσύνολα (Shakthi et al., 2008).

8.3. Άλλοι αλγόριθμοι

Στη ενότητα αυτή παρουσιάζονται ορισμένοι ακόμα παράλληλοι και κατανεμημένοι αλγόριθμοι που εφαρμόζονται στην εξόρυξη δεδομένων:

- Parallel HOP: Συσταδοποίηση Χωρικών Δεδομένων (Spatial Data): Ο HOP είναι ένας αλγόριθμος συσταδοποίησης βασισμένος στην πυκνότητα που αναγνωρίζει ομάδες σωματιδίων σε προσομοιώσεις N-σωμάτων (Li et al.).

- Count Distribution: Σε αυτό τον αλγόριθμο που βασίζεται σε γύρους, κάθε κόμβος πρώτα συμπληρώνει τους υποψήφιους για τα συχνά k-στοιχειοσύνολα των τοπικών του δεδομένων, ξεκινώντας με τα συχνά 1-στοιχειοσύνολα, δηλαδή τα αντικείμενα (Byrd et al.).

- Data Distribution: Ο αλγόριθμος Data Distribution μοιάζει με τον Count Distributionαλλά όλοι οι κόμβοι σε αυτόν υπολογίζουν ασύνδετα σύνολα υποψηφίων (Byrd et al.).

- DDM: Ο αλγόριθμος αυτός ανήκει στην ομάδα των αλγορίθμων Apriori. Εδώ, αφού οι τοπικές μετρήσεις συχνότητας υπολογίζονται για κάθε κόμβο, οι κόμβοι εκτελούν ένα κατανεμημένο πρωτόκολλο απόφασης σε κάθε γύρο, προκειμένου να καθοριστεί η δέσμη των καθολικά συχνών στοιχειοσυνόλων. Εξαιτίας αυτού του πρωτοκόλλου, ο DDM είναι πολύ αποδοτικός επικοινωνιακά. Επίσης, είναι πιο επεκτάσιμος και προσαρμόζεται καλύτερα στις ασυμμετρίες των δεδομένων σε σχέση με άλλους αλγορίθμους Apriori (Byrd et al.).

- D-Sampling: Ο αλγόριθμος αυτός είναι ένας συνδυασμός ενός κεντρικού αλγορίθμου δειγματοληψίας και του αλγορίθμου DDM. Ο D-Sampling προϋποθέτει ένα συγκεντρωτικό σύνολο δεδομένων και το διανέμει κατά την διάρκεια της εκτέλεσης (Byrd et al.).

- MLFPT (Multiple Local Frequent Pattern Tree): Ο αλγόριθμος αυτός υποθέτει μια αρχιτεκτονική κοινόχρηστης μνήμης. Δεν παράγει υποψήφιους για τα συχνά στοιχειοσύνολα αλλά αντίθετα χτίζει πολλαπλά συχνά δέντρα μοτίβων (FP-δέντρα) που χρειάζονται μόνο δύο πλήρεις σαρώσεις του

συνόλου δεδομένων. Έτσι, δεν μπορεί να χρησιμοποιηθεί για την εξισορρόπηση του φορτίου σε άλλους αλγορίθμους (Byrd et al.).

- ZigZag Algorithm: Αυτός ο αλγόριθμος βασίζεται στην παραδοχή ενός περιβάλλοντος όπου τα δεδομένα αρχικά κατανέμονται σε διαφορετικούς χώρους (όπως το δίκτυο δεδομένων για την ανίχνευση εισβολής) (Byrd et al.).

8.4. Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκαν ορισμένοι αλγόριθμοι που χρησιμοποιούνται σε κατανεμημένα και παράλληλα συστήματα. Στη συνέχεια ακολουθεί ένα συμπληρωματικό κεφάλαιο όπου περιγράφονται ορισμένες ακόμα τεχνικές εξόρυξης αλγορίθμων, εφαρμογές και προϊόντα λογισμικού.

9. ΆΛΛΕΣ ΤΕΧΝΙΚΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

9.1. Εισαγωγή

Στο κεφάλαιο αυτό θα γίνει μια συμπληρωματική περιγραφή των τεχνικών εξόρυξης δεδομένων που δεν αναφέρθηκαν στα προηγούμενα κεφάλαια και θα παρουσιαστούν οι εφαρμογές που βρίσκει η εξόρυξη δεδομένων καθώς και συγκεκριμένες εμπορικές εφαρμογές των τεχνικών εξόρυξης δεδομένων.

9.2. Άλλες τεχνικές και αλγόριθμοι

Μια σειρά από επιπλέον τεχνικές που χρησιμοποιούνται στην εξόρυξη δεδομένων είναι οι ακόλουθες:

- **Νευρωνικά Δίκτυα (Neural Networks)**

Στους συνηθέστερους αλγόριθμους εξόρυξης δεδομένων ανήκουν τα **νευρωνικά δίκτυα (neural networks)**, τα οποία έχουν ιδιαίτερο ενδιαφέρον, συμμετέχοντας στη διαμόρφωση σταδίων της τεχνολογίας εξόρυξης δεδομένων. Ωστόσο, τα νευρωνικά δίκτυα έχουν μειονεκτήματα που μπορούν να περιορίσουν την ευκολία χρήσης και εγκατάστασης, αλλά έχουν επίσης και κάποια σημαντικά πλεονεκτήματα. Στα πλεονεκτήματα τους ανήκουν τα πολύ ακριβή προβλεπτικά μοντέλα, τα οποία μπορούν να εφαρμοστούν σε ένα μεγάλο αριθμό διαφορετικών τύπων προβλημάτων. Με τον όρο «νευρωνικό δίκτυο» αναφερόμαστε σε ένα «τεχνητό νευρωνικό δίκτυο». Τα αληθινά νευρωνικά δίκτυα είναι τα βιολογικά συστήματα που ανιχνεύουν μορφές, κάνουν προβλέψεις και μαθαίνουν. Τα τεχνητά αυτά προγράμματα ηλεκτρονικών υπολογιστών εφαρμόζουν εξελιγμένη ανίχνευση πρότυπων και αλγόριθμους μηχανικής μάθησης σε έναν υπολογιστή για την κατασκευή μοντέλων πρόβλεψης από μεγάλες βάσεις. Τα νευρωνικά δίκτυα αναπτύχθηκαν από Τεχνητή Νοημοσύνη και όχι από την Στατιστική (Berson et al., 2010).

- **Γενετικοί Αλγόριθμοι (Generic Algorithms)**

Οι **γενετικοί αλγόριθμοι (genetic algorithms, GAs)** είναι μια τεχνική προγραμματισμού που μιμείται τη βιολογική εξέλιξη ως ένα πρόβλημα στρατηγικής επίλυσης. Λαμβάνοντας υπόψη ένα συγκεκριμένο πρόβλημα προς επίλυση, η είσοδος στον GA είναι ένα σύνολο πιθανών λύσεων στο πρόβλημα αυτό, που κωδικοποιούνται με κάποιο τρόπο, και μια μετρική που ονομάζεται συνάρτηση καταλληλότητας, που επιτρέπει σε κάθε υποψήφιο να αξιολογηθεί ποσοτικά. Οι υποψήφιοι μπορούν να είναι λύσεις που είναι ήδη γνωστό ότι λειτουργούν, με στόχο ο GA να τις βελτιώσει, αλλά τις περισσότερες φορές δημιουργούνται τυχαία. Ο GA τότε αξιολογεί κάθε υποψήφιο ανάλογα με τη λειτουργία ταιριάσματος. Σε μια ομάδα που δημιουργείται από τυχαίους υποψήφιους, βέβαια, οι περισσότεροι δεν θα λειτουργούν καθόλου, και αυτοί θα διαγραφούν. Ωστόσο, καθαρά από θέμα τύχης, λίγα μπορεί να κρατήσουν την υπόσχεση - μπορούν να εμφανίζουν δραστηριότητα, έστω και αδύναμη και ατελή δραστηριότητα, προς την επίλυση του προβλήματος (Marczyk, 2004). Η χρήση

τους στην εξόρυξη δεδομένων και στην αναγνώριση προτύπων είναι τεκμηριωμένη. Στο θέμα της αναγνώρισης προτύπων υπάρχουν δυο διαφορετικές προσεγγίσεις εφαρμογής τους. Η πρώτη αφορά την άμεση εφαρμογή τους ως ταξινομητή και η δεύτερη τη χρήση τους ως ένα εργαλείο βελτιστοποίησης για επαναφορά των παραμέτρων σε άλλους ταξινομητές (Minaei-Bidgoli, 2003).

- **Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό**

Η **εξόρυξη δεδομένων ιστού (web data mining)** περιλαμβάνει τη διαδικασία συλλογής και σύνοψης στοιχείων από τη δομή υπερ-συνδέσεων μιας τοποθεσίας Ιστού, το περιεχόμενο της σελίδας, ή τη σύνδεση χρήση τους, για να προσδιορίσει τα πρότυπα. Χρησιμοποιώντας την εξόρυξη δεδομένων, μια εταιρεία μπορεί να προσδιορίσει ένα δυνητικό ανταγωνιστή, τη βελτίωση της εξυπηρέτησης των πελατών, ή τις ανάγκες των πελατών-στόχων και των προσδοκιών. Μερικές κοινές τεχνικές εξόρυξης δεδομένων Ιστού περιλαμβάνουν την εξόρυξη περιεχομένου, την εξόρυξη χρήσης και την εξόρυξη δομής. Η εξόρυξη περιεχομένου εξετάζει το αντικείμενο μιας τοποθεσίας Ιστού. Η επεξεργασία φυσικής γλώσσας και η ανάκτηση πληροφοριών είναι δύο τεχνικές εξόρυξης δεδομένων που χρησιμοποιείται συχνά από τις εφαρμογές εξόρυξης δεδομένων Ιστού. Η εξόρυξη γνώσης χρήσης είναι συνήθως μια αυτοματοποιημένη διαδικασία κατά την οποία οι εξυπηρετητές Ιστού συγκεντρώνουν και να αναφέρουν τα πρότυπα πρόσβασης των χρηστών όσον αφορά την πρόσβαση σε server logs. Μια εταιρεία μπορεί, για παράδειγμα, να χρησιμοποιήσει ένα εργαλείο εξόρυξης δεδομένων χρήσης και να υποβάλει έκθεση σχετικά με την πρόσβαση στα αρχεία καταγραφής του διακομιστή και τα στοιχεία εγγραφής χρήστη, ώστε να δημιουργηθεί μια πιο αποτελεσματική δομή Ιστού. Η εξόρυξη δομής συχνά περιλαμβάνει την αποκάλυψη προτύπων υπερσυνδέσμων ή τις δομές εγγράφων σε μια ιστοσελίδα. Δύο γενικές τεχνικές εξόρυξης δεδομένων που μπορούν να χρησιμοποιηθούν από εφαρμογές εξόρυξης δεδομένων Ιστού η ανάλυση σύνδεσης εξόρυξης δεδομένων η παλινδρόμηση εξόρυξης δεδομένων (Delich).

9.3. Εφαρμογές και προϊόντα λογισμικού

Η εξόρυξη δεδομένων βρίσκει εφαρμογή σε πολλούς επιστημονικούς τομείς και στο χώρο των επιχειρήσεων και πολλές εφαρμογές λογισμικού χρησιμοποιούν τις τεχνικές της.

9.3.1. Εφαρμογές

Η τεχνολογία εξόρυξης δεδομένων χρησιμοποιείται συνήθως από οργανισμούς ή τμήματα επιχειρηματικής ευφυΐας, και από οικονομικούς αναλυτές. Ωστόσο, μπορεί να χρησιμοποιηθεί οπουδήποτε υπάρχει ανάγκη εξαγωγής χρήσιμης γνώσης από τεράστια σύνολα δεδομένων που συλλέγονται με τις σύγχρονες μεθόδους έρευνας και παρατήρησης (DataMining.gr).

9.3.2. Προϊόντα Λογισμικού

Δύο συστήματα συσταδοποίησης είναι το σύστημα PRIZM ™ της εταιρίας Claritas και Microvision ™ της εταιρίας Equifax. Οι εταιρείες αυτές έχουν ομαδοποιήσει του πληθυσμού από τις δημογραφικές πληροφορίες σε τμήματα τα οποία πιστεύουν ότι είναι χρήσιμα για άμεση εμπορική προώθηση και πωλήσεις. Για την κατασκευή αυτών ομάδων που χρησιμοποιούν στοιχεία όπως το εισόδημα, η ηλικία, το επάγγελμα, η στέγαση και η φυλή συλλεγμένα στο US Census. Στη συνέχεια αναθέτουν «ψευδώνυμα» εύκολα προς μνημόνευση για τις συστάδες (Berson et al., 2010).

Ορισμένα εργαλεία ελεύθερου λογισμικού που χρησιμοποιούν τεχνικές εξόρυξης δεδομένων είναι (DataMining.gr):

1 AlphaMiner: πλατφόρμα υλοποιημένη σε Java για εφαρμογές εξόρυξης δεδομένων. Αναπτύχθηκε στο πανεπιστήμιο του Hong Kong.

2 RapidMiner: εφαρμογή βασισμένη σε ελεύθερο λογισμικό για ανάλυση δεδομένων και εξόρυξη δεδομένων. Μπορεί να ενσωματωθεί σε άλλες εφαρμογές ή προϊόντα.

3 WEKA - Machine Learning Techniques: συλλογή εργαλείων και τεχνικών μηχανικής μάθησης για εφαρμογές εξόρυξης δεδομένων (και όχι μόνο), το

οποίο έχει αναπτυχθεί σε γλώσσα Java και ο κώδικας είναι ανοικτός στο κοινό.

Ορισμένα παραδείγματα των πιο γνωστών εταιρειών που χρησιμοποιούν την τεχνολογία της εξόρυξης δεδομένων είναι (DataMining.gr):

1. Oracle: Δημοφιλής και αξιόπιστη βάση δεδομένων, που υποστηρίζει, λειτουργίες εξόρυξης και ανάλυσης δεδομένων.
2. SPSS: Επιχειρηματικές λύσεις και λογισμικόγια ανάλυση δεδομένων.
3. Microsoft: Οι νεότερες εκδόσεις του SQL Server περιλαμβάνουν μια ειδική πλατφόρμα για επιχειρηματική νοημοσύνη με εξόρυξη δεδομένων.
4. IBM: Προσφέρει εργαλεία επιχειρηματικής ευφυΐας και εξόρυξης γνώσης για υψηλών απαιτήσεων αναλύσεις και προβλέψεις.

9.4. Επίλογος

Στο κεφάλαιο αυτό, που είναι συμπληρωματικό, περιγράφηκαν συνοπτικά μερικές ακόμα τεχνικές εξόρυξης δεδομένων και οι εφαρμογές που βρίσκει, αναφέροντας ορισμένα παραδείγματα λογισμικού που χρησιμοποιεί την τεχνολογία. Στο επόμενο κεφάλαιο βρίσκονται τα συμπεράσματα της εργασίας.

10. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός της εργασίας ήταν η παρουσίαση των βασικότερων κατηγοριών τεχνικών και αλγορίθμων εξόρυξης δεδομένων, καθώς και των προϊόντων λογισμικού που κάνουν εφαρμογή των τεχνικών και αλγορίθμων αυτών. Τα εργαλεία εξόρυξης δεδομένων προβλέπουν τις μελλοντικές τάσεις και συμπεριφορές, που επιτρέπουν στις επιχειρήσεις να πάρουν βασισμένες στη γνώση αποφάσεις.

Η εξόρυξη δεδομένων ή ανακάλυψη γνώσης σε βάσεις δεδομένων, είναι η πρακτική της αυτόματης αναζήτησης μεγάλων αποθηκών δεδομένων για πρότυπα. Η χρήση της στηρίζεται σε κάποια βάση δεδομένων ή σε αποθήκη

δεδομένων, η οποία ορίζεται ως μια διαδικασία κεντρικής διαχείρισης των δεδομένων και ανάκτησης.

Όταν χρησιμοποιείται στα τεχνικά πλαίσια αποθήκευσης δεδομένων και ανάλυσης, η εξόρυξη δεδομένων είναι ουδέτερη. Ωστόσο, μερικές φορές συνεπάγεται έναν αριθμό προβλημάτων, την επιβολή προτύπων σχετικά με τα δεδομένα εκεί που δεν υπάρχουν, ή προβλήματα ακεραιότητας των δεδομένων ή ζητήματα κόστους.

Η διαδικασία εξόρυξης δεδομένων Αποτελείται από 3 στάδια: Εξερεύνηση, Δημιουργία μοντέλου και επικύρωση, Ανάπτυξη.

Οι αλγόριθμοι εξόρυξης δεδομένων είναι προγραμματισμένα ερωτήματα και προγράμματα που χρησιμοποιούνται για την αναγνώριση προτύπων και τάσεις σε σύνολα δεδομένων. Δεν υπάρχει συγκεκριμένος κανόνας που υπαγορεύει πότε πρέπει να επιλεγεί μια ιδιαίτερη τεχνική σε αντίθεση με κάποια άλλη. Μερικές φορές αυτές οι αποφάσεις γίνονται σχετικά αυθαίρετα. Ορισμένα από τα κριτήρια που είναι σημαντικά για τον καθορισμό της τεχνικής που θα χρησιμοποιηθεί καθορίζονται από τη μέθοδο δοκιμής και σφάλματος. Ενώ μπορούν να χρησιμοποιηθούν διαφορετικοί αλγόριθμοι για την ίδια εργασία, κάθε αλγόριθμος παράγει ένα διαφορετικό αποτέλεσμα, και μερικοί αλγόριθμοι μπορούν να παράγουν περισσότερους από έναν τύπους αποτελέσματος. Επίσης, δεν χρειάζεται να χρησιμοποιηθούν οι αλγόριθμοι ανεξάρτητα.

Οι τύποι δεδομένων στην εξόρυξη δεδομένων είναι: Κείμενο (Text), Μακρύς (Long), Boolean, Διπλός (Double), Ημερομηνία (Date). Οι τύποι περιεχομένου είναι: Διακριτός (Discrete), Συνεχής (Continuous), Διακριτοποιημένος (Discretized), Κλειδί (Key), Ακολουθία Κλειδιού (Key Sequence), Χρόνος Κλειδιού (Key Time), Πίνακας (Table), Κυκλικός (Cyclical), Διατεταγμένος, Ordered), Ταξινομημένος (Classified).

Οι βασικές στατιστικές μέθοδοι είναι η κατάταξη (classification), με σημαντικότερους αλγορίθμους αφελή Bayes (naive Bayes) και τις μηχανές διανυσματικής υποστήριξης, και η παλινδρόμηση. Η ανάλυση παλινδρόμησης αποσκοπεί να καθορίσει τις τιμές των παραμέτρων για μια λειτουργία που κάνει τη συνάρτηση να ταιριάζει καλύτερα σε ένα σύνολο παρατηρήσεων δεδομένων.

Μια άλλη τεχνική που χρησιμοποιείται είναι η τμηματοποίηση, με βασικές μεθόδους τις: συσταδοποίηση, που διανείμει περιπτώσεις σε ομάδες, έτσι ώστε ο βαθμός της ενώσεως να είναι ισχυρός μεταξύ των μελών της ίδιας ομάδας και αδύναμος μεταξύ μελών διαφορετικών ομάδων, η δημογραφική συσταδοποίηση που βασίζεται σε κατανομές, οι αυξητικοί Αλγόριθμοι και ο αλγόριθμος k-πλησιέστερου γείτονα (k Nearest Neighbor, kNN) που αποτελεί μέρος της μηχανικής μάθησης.

Η εξόρυξη συνδυαστικών κανόνων βρίσκει ενδιαφέρουσες ενώσεις ή/και σχέσεις αντιστοιχίας ανάμεσα σε μεγάλα σύνολα των στοιχείων δεδομένων.

Σε πολλά επιχειρηματικά και επιστημονικά πεδία η ύπαρξη γεγονότων τα οποία συμβαίνουν σε μια ακολουθία παρουσιάζει ενδιαφέρον. Οι αλγόριθμοι για την εξόρυξη σειριακών προτύπων διευθετούν το πρόβλημα της ανακάλυψης των υπαρχόντων μεγίστων ακολουθιών σε μια δοθείσα βάση δεδομένων.

Ως καθοδηγούμενη ή μηχανική εκμάθηση ορίζεται η ικανότητα μιας μηχανής να βελτιώσει τις επιδόσεις της με τη χρήση ενός λογισμικού που χρησιμοποιεί τεχνικές τεχνητής νοημοσύνης για να μιμηθεί τους τρόπους με τους οποίους μαθαίνουν οι άνθρωποι, όπως η επανάληψη και η εμπειρία.

Ένα δέντρο απόφασης είναι ένα μοντέλο πρόβλεψης που, όπως υποδηλώνει το όνομά του, μπορεί να θεωρηθεί ως ένα δέντρο. Η τεχνική αυτή εφαρμόζεται στην εξόρυξη αλγορίθμων σε πολλές περιπτώσεις.

Σε παράλληλο περιβάλλον, αξιοποιώντας τη μεγάλη συνολική κύρια μνήμη και την επεξεργαστική ισχύ των παράλληλων επεξεργαστών, οι παράλληλοι αλγόριθμοι μπορούν να έχουν και τον χρόνο εκτέλεσης και τα ζητήματα απαιτούμενης μνήμης σωστά διευθετημένα. Επίσης, η κατανεμημένη εξόρυξη δεδομένων προσπαθεί να χρησιμοποιήσει τους καλύτερους διαθέσιμους πόρους.

Άλλες τεχνικές και αλγόριθμοι που χρησιμοποιούνται στην εξόρυξη δεδομένων είναι: νευρωνικά δίκτυα, γενετικοί αλγόριθμοι και εξόρυξη δεδομένων στον παγκόσμιο ιστό. Επίσης, τα πεδία εφαρμογής της εξόρυξης δεδομένων βρίσκονται κυρίως στις επιχειρήσεις αλλά και σε οποιοδήποτε φορέα η οργανισμό χρειάζεται να εξάγει συμπεράσματα από τα δεδομένα στις βάσεις δεδομένων του.

Συμπερασματικά, η εξόρυξη δεδομένων είναι ένας ραγδαία αναπτυσσόμενος επιστημονικός κλάδος που συνδυάζει τεχνικές από πολλούς άλλους κλάδους, όπως η στατιστική, η μηχανική μάθηση, η εξαγωγή κανόνων και άλλες, προσπαθώντας να αναλύσει μεγάλους αριθμούς δεδομένων οποιουδήποτε τύπου που βρίσκονται σε βάσεις ή αποθήκες δεδομένων. Σκοπός της είναι να βρει πρότυπα στα δεδομένα αυτά, τα οποία θα υποδεικνύουν κάποια συγκεκριμένη συμπεριφορά ή οργάνωση των δεδομένων με συγκεκριμένο τρόπο και θα επιτρέπει την εξαγωγή συμπερασμάτων και την πρόβλεψη μελλοντικών συμπεριφορών με βάση τα δεδομένα. Όπως γίνεται κατανοητό, αυτό αποτελεί ένα χαρακτηριστικό εξαιρετικά χρήσιμο ειδικά στις επιχειρήσεις αλλά και σε πολλούς άλλους τομείς. Κατά συνέπεια, η εξόρυξη δεδομένων αναμένεται να αναπτυχθεί ακόμα περισσότερο, παράλληλα με την ανάπτυξη των υπολογιστών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

DataMining.gr, http://www.datamining.gr/.

Ramakrishnan R., Gehrke J., Συστήματα Διαχείρισης Βάσεων Δεδομένων, Τόμος Β, Δεύτερη Έκδοση, Εκδόσεις Τζιόλα, Θεσσαλονίκη, 2002.

Tsoumakas G., Vlahavas I., Distributed Data Mining, Encyclopedia of Data Warehousing and Mining - 2nd Edition, John Wang (Ed.), Idea Group Reference, pp. 709-715, 2008.

ΞΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Aasheim Ø., Solheim H, Rough Sets as a Framework for Data Mining, Project report, Knowledge Systems Group, Faculty of Computer Systems and Telematics, The Norwegian University of Science and Technology, Trondheim, 1996, http://www.pvv.ntnu.no/~hgs/project/report/main.html.

Agrawal R., Ramakrishnan S., Mining Sequential Patterns, 2004, http://www-users.cs.umn.edu/~desikan/research/dataminingoverview.html.

Anissimov M., What is Data Mining?, wiseGEEK, 2011, http://www.wisegeek.com/what-is-data-mining.htm.

Antunes C. and Oliveira A. L. Sequential pattern mining algorithms: Trade-offs between speed and memory. In 2nd Workshop on Mining Graphs, Trees and Seq, Italy, 2004.

Arun K. Pujari, Data mining techniques, Universities Press (India) Private Limited, 2001: pp 48-49.

Berson A., Smith S, and Thearling K., An Overview of Data Mining Techniques, Excerpted from the book Building Data Mining Applications for CRM, 2010, http://www.thearling.com/text/dmtechniques/dmtechniques.htm.

BusinessDictionary.com, classification, http://www.businessdictionary.com/definition/classification.html.

BusinessDictionary.com, machine learning, http://www.businessdictionary.com/definition/machine-learning.html.

B&M Services, Clustering, http://www.bandmservices.com/Clustering/Clustering.htm.

Chapple M., Data Mining: An Introduction, About.com Guide, 2011, http://databases.about.com/od/datamining/a/datamining.htm.

Classle.net, C4.5 algorithm, 2009, http://www.classle.net/node/18681?q=node/27868.

dataminingarticles.com, Data Mining 101 - Part 5, Algorithms for Mining Frequent, Closed and Maximal Itemsets, 2011, http://www.dataminingarticles.com/association-rules-algorithms.html.

decisiontrees.net, Tutorial (4): ID3 Decision Trees tutorial > ID3 & Entropy, Decision Trees & Data Mining, http://www.decisiontrees.net/?q=node/27.

Delich C., What Is Web Data Mining?, wiseGEEK, http://www.wisegeek.com/what-is-web-data-mining.htm.

Filho J., Soibelman L. and Choo J., "SEQUENTIAL ANALYSIS OF REASONS FOR NON-COMPLETION OF ACTIVITIES: CASE STUDY AND FUTURE DIRECTIONS", 2004.

Francois C., What Are Data Mining Algorithms?, wiseGeek, http://www.wisegeek.com/what-are-data-mining-algorithms.htm.

IBM, Demographic clustering, Clustering, Data mining functions and algorithms, Data Mining, Data warehousing and analytics Data mining and text analysis, IBM DB2 Database for Linux, UNIX, and Windows Information Center, 2011, http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jsp.http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jspl

IBM, Regression, Data mining functions and algorithms, Data Mining, Data warehousing and analytics Data mining and text analysis, IBM DB2 Database for Linux, UNIX, and Windows Information Center, 2011, http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jsp.

Kamal A., Manganaris S., and Ramakrishnan S., Partial Classification Using Association Rules, in Proc. KDD, 1997, pp.115-118.

Li J., Liu Y., Liao W., Choudhary A., 1 Parallel Data Mining Algorithms for Association Rules and Clustering, http://users.eecs.northwestern.edu/~yingliu/papers/para_arm_cluster.pdf.

Marczyk A., Genetic Algorithms and Evolutionary Computation, The Talk Origins Archive, 2004, http://www.talkorigins.org/faqs/genalg/genalg.html.

Microsoft MSDN, Content Types (Data Mining), 2011, http://msdn.microsoft.com/en-us/library/ms174572.aspx.

Microsoft MSDN, Data Mining Algorithms (Analysis Services - Data Mining), 2011,.http://msdn.microsoft.com/en-us/library/ms175595.aspx.

Microsoft | Technet, Data Types (Data Mining), SQL Server 2005 Books Online, 2008, http://64.4.11.252/en-us/library/ms174796(SQL.90).aspx .

Minaei-Bidgoli, B. and Punch, W.F. Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System. In Proceedings of GECCO. 2003, 2252-2263.

Moreira, A., Santos, M., Carneiro, S. (2008). Density based clustering algorithms – DBSCAN and SNN, Version 1.0, 25.07.2005, http://ubicomp.algoritmi.uminho.pt/local/download/SNN&DBSCAN.pdf

Oberst J., 1 Efficient Data Clustering and How to Groom fast-Growing Trees, 2009, http://www.janoberst.com/_academics/2009_03_03_BIRCH-Efficient-Data-Clustering-Fast-Growing-Trees.pdf.

Oracle, 4 Regression, Oracle® Data Mining Concepts, 11g Release 1 (11.1), 2011, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/regress.htm .

Palace B., Data Mining, Technology Note prepared for Management 274A, Anderson Graduate School of Management at UCLA, 1996, http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/.

Sakthi U., Hemalatha R., and Bhuvaneswaran R. S., Parallel and Distributed Mining of Association Rule on Knowledge Grid, World Academy of Science, Engineering and Technology 42 2008.

Sharad Verma, Nikita Jain, Implementation of ID3 – Decision Tree Algorithm, Scribd.com, http://www.scribd.com/doc/22639832/ID3-Algorithm.

SPSS.com, Statistical Support – Algorithms, SPSS 14.0 Statistical Algorithms, http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/algorithms/

StatSoft.com, Electronic Statistics Textbook, Data Mining Techniques, http://www.statsoft.com/textbook/data-mining-techniques/.

Stonebraker M., Hellerstein J., Readings in Database Systems, Third Edition, Morgan Kaufmann Publishers, Inc, 1998.

The Code Project, K Nearest Neighbor Algorithm Implementation and Overview, 2009, http://www.codeproject.com/KB/recipes/K_nearest_neighbor.aspx.

Thearling K., An Introduction to Data Mining, Discovering hidden value in your data warehouse, Thearling.com , 2010, http://www.thearling.com/text/dmwhite/dmwhite.htm.

WebContentMining.com, ADABoost, 2010, http://webcontentmining.com/adaboost/.

Williams G., AdaBoost Algorithm, DATA MINING Desktop Survival Guide, 2010 http://www.togaware.com/datamining/survivor/AdaBoost_Algorithm.html.

Williams G., Outlier Analysis, DATA MINING Desktop Survival Guide, 2010, http://www.togaware.com/datamining/survivor/Outlier_Analysis.html.

WordIQ.com, KDD – Definition,2010, http://www.wordiq.com/definition/KDD.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A.F.M., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J., and Steinberg, D. Top 10 algorithms in data mining. In Proceedings of Knowl. Inf. Syst.. 2008: pp 1-37.

XLMiner, Association Rules, Online Help, Version 3, http://www.resample.com/xlminer/help/Assocrules/associationrules_intro.htm.

ΠΑΡΑΡΤΗΜΑ: ΚΩΔΙΚΕΣ ΑΛΓΟΡΙΘΜΩΝ ΣΕ JAVA

Πηγές όπου μπορούν να βρεθούν σε μορφή κώδικα οι παρακάτω αλγόριθμοι.

naive Bayes

(http://read.pudn.com/downloads95/sourcecode/book/383872/NaiveBayes.java__.htm**)**

k-means

(http://www.koders.com/java/fid427C06318D6557252A2AEBF2EBE8148EB26B421E.aspx)

ID3

(http://www.java2s.com/Open-Source/Java-Document/Science/weka/weka/classifiers/trees/Id3.java.htm**)**

κNN

(http://www.koders.com/java/fid42675E66D94C7C4B8425F76DA0D93E2BC217108A.aspx?s=IsNa#L23)

Apriori

(http://www.java2s.com/Open-Source/Java-Document/Science/weka/weka/associations/Apriori.java.htm)

ADABoost

(http://www.koders.com/kv.aspx?fid=B0F976B1445BCC6AA08C00D23DF6951156594164&s=297#L22**)**