ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ
ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ & ΠΑΡΑΓΩΓΗΣ

ΜΕΛΕΤΗ ΛΟΓΙΣΜΙΚΟΥ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ WEKA - ΘΕΩΡΗΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

ΌΝΟΜΑ ΣΠΟΥΔΑΣΤΗ: ΙΑΚΩΒΑΚΗΣ ΕΥΘΥΜΙΟΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΔΡΟΣΟΣ ΧΡΗΣΤΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΙΓΑΛΕΩ 2019

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους όσοι συνέβαλαν στο να επιτευχθεί αυτή η εργασία για την πολύτιμη βοήθειά τους και πάνω απ'όλα την στίρηξη που μου δείξανε κυρίως δείχνοντας ανοχή στις ιδιαιτερότητες του χαρακτήρα μου καθώς και στις όποιες παράλογες απαιτήσεις μου ακόμα και τις πιο ακατάλληλες ώρες και στιγμές. Πιο συγκεκριμένα, τους πάντα παρόντες στις δύσκολες μάχες μέχρι στιγμής της ζωής μου, την οικογένειά μου. Χωρίς τη δική τους στίρηξη καμία προσπάθεια δε γίνεται πιο εύκολη σε τόσο μεγάλο βαθμό όσο έγινε για εμένα η επίτευξη αυτής της εργασίας. Στη συνέχεια ένα μεγάλο ευχαριστώ στον υπεύθυνο καθηγητή που με την καθοδήγησή του οι στόχοι γίνονταν ξεκάθαροι και αρκούσε κάθε φορά να με πάει ένα βήμα πιο μπροστά. Θα ήθελα επίσης να ευχαριστήσω κι εκείνη την ομάδα ανθρώπων που βρίσκεται εκεί έξω με σκοπό να κάνει πραγματικότητα το όνειρο της εργασίας και που με τη στάση τους ο καθένας φροντίζει να μου μαθαίνει αρχικά και έκτοτε να μου υπενθυμίζει ότι τίποτα δε γίνεται αν δεν το θες, αν δεν το επιδιώκεις και πως τίποτα στη ζωή μας δεν είναι πραγματικά εύκολο. Χρειάζεται πείσμα, επιμονή, όρεξη, κάποιες φορές να πηγαίνεις κόντρα στα θέλω σου και στο "εγώ" σου, να έχεις τα μάτι σου ανοιχτά, να διεκδικείς ευκαιρίες και με σωστή μέθοδο και τακτική όλα είναι ζήτημα χρόνου να συμβούν. Τελευταίο και καλύτερο θέλω να ευχαριστήσω αυτόν που κάθε πρωί στον καθρέφτη, με το μεγάλο του χαμόγελο, μου χαρίζει απίστευτη θετική ενέργεια παρά τις όποιες δυσκολίες και την όση κούραση επικρατούν.

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος ΙΑΚΩΒΑΚΗΣ ΕΥΘΥΜΙΟΣ,
του ΠΑΝΑΓΙΩΤΗ, με αριθμό μητρώου 36107 φοιτητής του Τμήματος **Μηχανικών Αυτοματισμού Τ.Ε.** του Α.Ε.Ι. Πειραιά Τ.Τ. πριν αναλάβω την εκπόνηση της Πτυχιακής Εργασίας μου, δηλώνω ότι ενημερώθηκα για τα παρακάτω:

«Η Πτυχιακή Εργασία (Π.Ε.) αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο του συγγραφέα, όσο και του Ιδρύματος και θα πρέπει να έχει μοναδικό χαρακτήρα και πρωτότυπο περιεχόμενο.

Απαγορεύεται αυστηρά οποιοδήποτε κομμάτι κειμένου της να εμφανίζεται αυτούσιο ή μεταφρασμένο από κάποια άλλη δημοσιευμένη πηγή. Κάθε τέτοια πράξη αποτελεί προϊόν λογοκλοπής και εγείρει θέμα Ηθικής Τάξης για τα πνευματικά δικαιώματα του άλλου συγγραφέα. Αποκλειστικός υπεύθυνος είναι ο συγγραφέας της Π.Ε., ο οποίος φέρει και την ευθύνη των συνεπειών, ποινικών και άλλων, αυτής της πράξης.

Πέραν των όποιων ποινικών ευθυνών του συγγραφέα σε περίπτωση που το Ίδρυμα του έχει απονείμει Πτυχίο, αυτό ανακαλείται με απόφαση της Συνέλευσης του Τμήματος. Η Συνέλευση του Τμήματος με νέα απόφασης της, μετά από αίτηση του ενδιαφερόμενου, του αναθέτει εκ νέου την εκπόνηση της Π.Ε. με άλλο θέμα και διαφορετικό επιβλέποντα καθηγητή. Η εκπόνηση της εν λόγω Π.Ε. πρέπει να ολοκληρωθεί εντός τουλάχιστον ενός ημερολογιακού 6μήνου από την ημερομηνία ανάθεσης της. Κατά τα λοιπά εφαρμόζονται τα προβλεπόμενα στο άρθρο 18, παρ. 5 του ισχύοντος Εσωτερικού Κανονισμού.»

Ο Δηλών

Ημερομηνία
20 Ιουνίου 2019

Περίληψη

Η παρούσα πτυχιακή εργασία αποτελεί μια προσπάθεια περιγραφής της διαδικασίας εξόρυξης δεδομένων και συγκεκριμένα του προγράμματος weka. Στο πρώτο κεφάλαιο γίνεται ανάλυση της διαδικασίας εξόρυξης δεδομένων (data mining), ενώ στο δεύτερο κεφάλαιο πραγματοποιείται ανάλυση της μελέτης των δεδομένων. Τέλος στο τρίτο και τέταρτο κεφάλαιο της εργασίας, παρουσιάζεται και αναλύεται το πρόγραμμα και το λογισμικό weka, προκειμένου να γίνει κατανοητή η χρήση του καθώς και για να εκτιμηθεί η χρηστικότητά του.

Λέξεις κλειδιά: μηχανική μάθηση, εξόρυξη δεδομένων, δεδομένα, weka

Abstract

This thesis is an attempt to describe the data mining process, namely the weka project. The first chapter analyzes the data mining process, while the second chapter analyzes the data study. Finally, in the third and fourth chapters of the paper, the program and the weka software are presented and analyzed to understand its use and to evaluate its usefulness.

Key words: mechanical learning, data mining, data, weka

Περιεχόμενα

Κεφάλαιο 4ο- Λογισμικό WEKA

Εισαγωγή

Η εξόρυξη δεδομένων έχει οριστεί ως η μη κερδοσκοπική εξαγωγή προηγουμένως άγνωστων και δυνητικά χρήσιμων πληροφοριών από βάσεις δεδομένων / αποθήκες δεδομένων. Χρησιμοποιεί μηχανικές μάθησης, τεχνικές στατιστικής και οπτικοποίησης για να ανακαλύψει και να παρουσιάσει τη γνώση σε μια μορφή, η οποία είναι εύκολα περιεκτική για τον άνθρωπο (Han, & Kamber, 2006).

Η εξόρυξη δεδομένων, η εξαγωγή κρυφών προγνωστικών πληροφοριών από μεγάλες βάσεις δεδομένων, είναι μια ισχυρή νέα τεχνολογία με μεγάλη δυνατότητα να βοηθήσει τους χρήστες να επικεντρωθούν στις πιο σημαντικές πληροφορίες στις αποθήκες τους. Τα εργαλεία εξόρυξης δεδομένων προβλέπουν τις μελλοντικές τάσεις και συμπεριφορές, επιτρέποντας στις επιχειρήσεις να κάνουν προληπτικές αποφάσεις που βασίζονται στη γνώση. Οι αυτοματοποιημένες, προοπτικές αναλύσεις που προσφέρονται από την εξόρυξη δεδομένων ξεπερνούν τις αναλύσεις προηγούμενων γεγονότων που παρέχονται από τα αναδρομικά εργαλεία που είναι τυπικά για συστήματα υποστήριξης αποφάσεων.

Τα εργαλεία εξόρυξης δεδομένων μπορούν να απαντήσουν σε επιχειρηματικές ερωτήσεις που παραδοσιακά ήταν πολύ χρονοβόρες για την επίλυση. Τρέχουν βάσεις δεδομένων για κρυμμένα μοτίβα, βρίσκοντας προγνωστικές πληροφορίες που οι εμπειρογνώμονες μπορεί να χάσουν επειδή βρίσκονται εκτός των προσδοκιών τους. Οι τεχνικές εξόρυξης δεδομένων μπορούν να υλοποιηθούν γρήγορα σε υπάρχουσες πλατφόρμες λογισμικού και υλικού για την ενίσχυση της αξίας των υφιστάμενων πόρων πληροφόρησης και μπορούν να ενσωματωθούν με νέα προϊόντα και συστήματα, καθώς αυτά μεταφέρονται στο διαδίκτυο (Marakas, 2005).

Το Weka (Waikato Environment for Knowledge Analysis) είναι ένα δημοφιλές λογισμικό μηχανικής μάθησης γραμμένο σε Java, που αναπτύχθηκε στο Πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Το Weka είναι ελεύθερο λογισμικό διαθέσιμο υπό την Γενική Άδεια Δημόσιας Χρήσης GNU. Ο πίνακας εργασίας Weka περιέχει μια συλλογή εργαλείων οπτικοποίησης και αλγορίθμων για ανάλυση

δεδομένων και πρόβλεψης μοντέλων, μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτή τη λειτουργικότητα.

Το Weka είναι μια συλλογή αλγορίθμων μηχανικής μάθησης για την επίλυση πραγματικών προβλημάτων εξόρυξης δεδομένων. Είναι γραμμένο σε Java και λειτουργεί σχεδόν σε οποιαδήποτε πλατφόρμα. Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων είτε να καλούνται από τον δικό σας κώδικα Java. Είναι μια συλλογή αλγορίθμων μηχανικής μάθησης για εργασίες εξόρυξης δεδομένων. Το Weka περιλαμβάνει εργαλεία για την προεπεξεργασία δεδομένων, την ταξινόμηση, την παλινδρόμηση, την ομαδοποίηση, τους κανόνες σύνδεσης και την οπτικοποίηση. Είναι επίσης κατάλληλο για την ανάπτυξη νέων μηχανισμών εκμάθησης μηχανών.

Κεφάλαιο 1^{ο} – Εξόρυξη Δεδομένων (Data Mining)

1.6 Θεωρία Εξόρυξης Δεδομένων

Η έρευνα σχετικά με την εξόρυξη δεδομένων (Data Mining) και η ανακάλυψη γνώσεων πάνω σε βάσεις δεδομένων, έχει επικεντρωθεί κυρίως στην ανάπτυξη ιδανικών αλγορίθμων, ώστε να εξυπηρετούν της διάφορες διεργασίες της εξόρυξης δεδομένων. Ορισμένα μέρη της ερευνητικής προσπάθειας έχουν προχωρήσει στη διερεύνηση της διεργασίας, των θεμάτων διεπαφής χρηστών, των θεμάτων βάσης δεδομένων ή της απεικόνισης (Fayyad, et al., 1996). Ωστόσο οι δημοσιεύσεις που αφορούν τα θεωρητικά θεμέλια της εξόρυξης δεδομένων είναι περιορισμένα.

Οι βάσεις δεδομένων υπήρχαν ήδη στη δεκαετία του 1960, αλλά ο τομέας θεωρήθηκε ότι είναι μια κοινή συνισταμένη διαφορετικών εφαρμογών, χωρίς σαφή δομή και χωρίς ενδιαφέροντα θεωρητικά ζητήματα. Το σχεσιακό μοντέλο του Codd ήταν ένα χρήσιμο και απλό πλαίσιο, για τον καθορισμό της δομής των δεδομένων και των λειτουργιών που πρέπει να εκτελεστούν σε αυτό. Η μαθηματική περιγραφή του σχεσιακού μοντέλου, επέτρεψε την ανάπτυξη προηγμένων μεθόδων βελτιστοποίησης ερωτημάτων και συναλλαγών, οι οποίες με τη σειρά τους κατέστησαν εφικτά αποτελεσματικά συστήματα διαχείρισης βάσεων δεδομένων γενικής χρήσης (Mannila, 2000). Το σχεσιακό μοντέλο είναι ένα σαφές παράδειγμα του τρόπου με τον οποίο η θεωρία στην επιστήμη των υπολογιστών, έχει μετατρέψει μια περιοχή από μια σειρά από μη συνδεδεμένες μεθόδους, σε ένα ενδιαφέρον και κατανοητό σύνολο.

Δεδομένου ότι η θεωρία είναι χρήσιμη, θα πρέπει να επισημανθεί ποιες πρέπει να είναι οι ιδιότητες που πρέπει να ικανοποιήσει ένα θεωρητικό πλαίσιο, ώστε να μπορεί να χαρακτηριστεί ως «θεωρία για την εξόρυξη δεδομένων». Ένα θεωρητικό πλαίσιο πρέπει να είναι απλό και εύκολο να εφαρμοστεί. Θα πρέπει επίσης να μπορεί να δώσει χρήσιμα αποτελέσματα που θα μπορούν να εφαρμοστούν στην ανάπτυξη αλγορίθμων και μεθόδων εξόρυξης δεδομένων (Mannila, 2000).

Μια απλή προσέγγιση στη θεωρία της εξόρυξης δεδομένων είναι η περιγραφή της, ως ένα σύνολο στατιστικών στοιχείων (ίσως σε μεγαλύτερα σύνολα δεδομένων από ό, τι στο παρελθόν) και έτσι η αναζήτηση ενός θεωρητικού πλαισίου για την εξόρυξη δεδομένων θα μπορούσε να περιοριστεί σε αυτό. Εντουτοις η θεωρία της εξόρυξης δεδομένων έχει χαρακτήρα «στατιστικής» (ως επιστήμη). Η εξόρυξη δεδομένων είναι προφανώς πολύ κοντά στα στατιστικά στοιχεία και οι ερευνητές εξόρυξης δεδομένων με υπόβαθρα πληροφορικής, συνήθως έχουν πολύ λίγη εκπαίδευση στα στατιστικά στοιχεία (Mannila, 2000).

Ωστόσο, μπορεί κανείς να υποστηρίξει ότι υπάρχουν σημαντικές διαφορές μεταξύ των περιοχών. Ο όγκος των δεδομένων συνιστά μια σημαντική διαφορά, όπως επίσης και ο αριθμός των μεταβλητών ή των χαρακτηριστικών, που έχει συχνά πολύ πιο σημαντικό αντίκτυπο στις εφαρμοστέες μεθόδους ανάλυσης. Για παράδειγμα, η εξόρυξη δεδομένων έχει αντιμετωπιστεί με προβλήματα σχετικά με το τι πρέπει να κάνει κάποιος σε περιπτώσεις όπου ο αριθμός των μεταβλητών είναι τόσο μεγάλος, ώστε η εξέταση όλων των ζευγών μεταβλητών, δεν είναι υπολογιστικά εφικτή. Συνολικά, το θέμα της υπολογιστικής σκοπιμότητας έχει πολύ σαφέστερο ρόλο στην εξόρυξη δεδομένων, από ότι στις στατιστικές (Boulicaut, Klemettinen, & Mannila, 1998). Μια άλλη διαφορά είναι ότι η εξόρυξη δεδομένων αποτελεί συνήθως δευτερεύουσα ανάλυση δεδομένων: τα δεδομένα έχουν συλλεχθεί για κάποιο άλλο σκοπό, ως απάντηση σε μια συγκεκριμένη ερώτηση αναλυτικών στοιχείων (Mannila, 2000).

Κάποιες άλλες διαφορές μεταξύ των περιοχών, θα μπορούσαν να επισημανθούν. Για παράδειγμα, η έμφαση στην ολοκλήρωση της βάσης δεδομένων στην απλότητα χρήσης και στην κατανόηση των αποτελεσμάτων, είναι χαρακτηριστική των μεθόδων εξόρυξης δεδομένων. Αρκεί να επισημανθεί ότι τουλάχιστον σήμερα, το θεωρητικό πλαίσιο των στατιστικών φαίνεται να είναι σχετικά απομακρυσμένο από την πραγματική εξέλιξη των μεθόδων εξόρυξης δεδομένων. Επίσης, η στατιστική θεωρία δεν φαίνεται να δίνει ιδιαίτερη προσοχή στο χαρακτήρα διεργασίας της εξόρυξης δεδομένων (Mannila, 2000).

Μια παρόμοια (αλλά πιο αδύναμη) περίπτωση μείωσης της εξόρυξης δεδομένων σε μια υπάρχουσα περιοχή έχει γίνει από την άποψη της μηχανικής μάθησης. Θα

μπορούσε κάποιος να πει, ότι στη εξόρυξη δεδομένων εφαρμόζεται μηχανική μάθηση, και έτσι η θεωρία της εξόρυξης δεδομένων είναι ίση με τη θεωρία της μηχανικής μάθησης. Και πάλι, η προσέγγιση αυτή αποτυγχάνει για δύο λόγους. Πρώτον, υπάρχουν σημαντικές διαφορές μεταξύ της μηχανικής μάθησης και των στατιστικών και, δεύτερον, οι θεωρητικές προσεγγίσεις μηχανικής μάθησης (όπως το μοντέλο PAC) δεν ανταποκρίνονται πραγματικά στις ειδικές απαιτήσεις που έχει να κάνει με τη θεωρία της εξόρυξης δεδομένων (Boulicaut, Klemettinen, & Mannila, 1998).

Μια πιθανή θεωρητική προσέγγιση για την εξόρυξη δεδομένων είναι η προβολή της ως ένας τρόπος να βρεθεί η υποκείμενη κοινή κατανομή των μεταβλητών στα δεδομένα. Τυπικά, κάποιος στοχεύει στην εύρεση μιας σύντομης και κατανοητής αναπαράστασης της κοινής διανομής, όπως για παράδειγμα ένα Μπαγεσιανό δίκτυο (Heckerman, 1997) ή ένα ιεραρχικό Μπαγεσιανό μοντέλο (Gelman, et al., 1995, Gilks, Richardson, & Spiegelhalter, 1996). Αυτή η προσέγγιση προφανώς συνδέεται στενά με την αναγωγική προσέγγιση της προβολής της εξόρυξης δεδομένων ως στατιστικών στοιχείων.

Τα πλεονεκτήματα της προσέγγισης, είναι ότι το υπόβαθρο είναι πολύ σταθερό και είναι εύκολο να τεθούν επίσημα ερωτήματα. Εργασίες όπως η ομαδοποίηση ή η ταξινόμηση ταιριάζουν εύκολα στην προσέγγιση αυτή. Αυτό που φαίνεται να λείπει, όπως και στις περισσότερες από τις προσεγγίσεις, είναι οι τρόποι που χρησιμοποιούνται, για να ληφθεί υπόψη ο επαναληπτικός και διαδραστικός χαρακτήρας της διαδικασίας εξόρυξης δεδομένων

Εν κατακλείδι επισημαίνεται, ότι μια καλή και επαρκής θεωρία για την εξόρυξη δεδομένων θα πρέπει να εξετάζει τη διαδικασία της εξόρυξης δεδομένων, να έχει πιθανολογικούς χαρακτήρες, να είναι σε θέση να περιγράψει διαφορετικά καθήκοντα εξόρυξης δεδομένων, όπως επίσης και να είναι σε θέση να επιτρέψει την παρουσία γνώσεων υποβάθρου κλπ. (Mannila, 2000).

1.2 Data mining – ορισμός

Η ανάπτυξη της Πληροφορικής έχει δημιουργήσει έναν μεγάλο αριθμό βάσεων δεδομένων, καθώς επίσης και μια πληθώρα δεδομένων, σε διάφορους τομείς. Η έρευνα σε διάφορες βάσεις και η τεχνολογία της πληροφορίας, έχουν με τη σειρά τους οδηγήσει σε μια προσέγγιση για την αποθήκευση και τον χειρισμό αυτών των πολύτιμων δεδομένων για περαιτέρω λήψη αποφάσεων (Ramageri, 2010).

Η εξόρυξη δεδομένων (Data Mining) είναι μια διαδικασία εξαγωγής χρήσιμων πληροφοριών και σχεδίων, μέσα από έναν τεράστιο όγκο δεδομένων. Καλείται επίσης ως διαδικασία αποκάλυψης γνώσης, εξόρυξη γνώσης από δεδομένα, εξαγωγή γνώσης ή ανάλυση δεδομένων / προτύπου. Η εξόρυξη δεδομένων είναι μια λογική διαδικασία που χρησιμοποιείται για την αναζήτηση μεγάλου όγκου πληροφοριών για την εύρεση χρήσιμων δεδομένων (Jiawei, & Micheline, 2006). Η εξόρυξη δεδομένων (Data Mining) αποτελεί στην ουσία, τη διαδικασία ανεύρεσης «μοτίβων», σε μεγάλα σύνολα δεδομένων, που περιλαμβάνουν μεθόδους σχετικά με τη διασταύρωση της μηχανικής μάθησης, των στατιστικών και των συστημάτων βάσεων δεδομένων. Ο όρος «εξόρυξη δεδομένων» είναι στην πραγματικότητα μια εσφαλμένη ονομασία, αφού ο στόχος είναι η εξαγωγή προτύπων και γνώσεων μέσα από μεγάλες ποσότητες δεδομένων και όχι η εξαγωγή (εξόρυξη) των ίδιων των δεδομένων (Jiawei, & Micheline, 2006)

Ο στόχος αυτής της τεχνικής είναι να βρεθούν μοτίβα που ήταν προηγουμένως άγνωστα. Μόλις εντοπιστούν αυτά τα μοτίβα (πρότυπα), μπορούν να χρησιμοποιηθούν περαιτέρω για να ληφθούν ορισμένες αποφάσεις με σκοπό την ανάπτυξη των επιχειρήσεων τους. Υπάρχουν τρία βήματα που ακολουθούνται κατά τη διαδικασία (Ramageri, 2010):

- Εξερεύνηση
- Ταυτοποίηση μοτίβου (προτύπου)
- Ανάπτυξη

Αναλυτικότερα τα βήματα περιγράφονται ως εξής:

Εξερεύνηση: Στο πρώτο βήμα της εξερεύνησης δεδομένων, τα δεδομένα καθαρίζονται και μετατρέπονται σε άλλη μορφή, ενώ καθορίζονται σημαντικές μεταβλητές και στη συνέχεια η φύση των δεδομένων με βάση το πρόβλημα (Ramageri, 2010).

Αναγνώριση μοτίβων: Μόλις εξερευνηθούν, επεξεργαστούν και καθοριστούν τα δεδομένα για τις συγκεκριμένες μεταβλητές, το δεύτερο βήμα είναι να διαμορφωθεί η αναγνώριση προτύπου. Στη συνέχεια, προσδιορίζονται και επιλέγονται τα πρότυπα εκείνα, που εξασφαλίζουν την καλύτερη πρόβλεψη (Ramageri, 2010).

Ανάπτυξη: Τα μοτίβα αναπτύσσονται για το επιθυμητό αποτέλεσμα (Ramageri, 2010).

1.3 Στόχοι εφαρμογής εξόρυξης δεδομένων

Ο στόχος της εξόρυξης δεδομένων (Data Mining) είναι να εντοπιστούν έγκυροι, νέοι, δυνητικά χρήσιμοι και κατανοητοί συσχετισμοί και πρότυπα, σε υπάρχοντα δεδομένα (Chung, & Grey 1999). Η εύρεση χρήσιμων μοτίβων στα δεδομένα, είναι γνωστή με διαφορετικά ονόματα (συμπεριλαμβανομένου της εξόρυξης δεδομένων) σε διάφορες κοινότητες (π.χ. εξαγωγή γνώσης, ανακάλυψη πληροφοριών, συλλογή πληροφοριών, και επεξεργασία προτύπων δεδομένων) (Fayyad, et al, 1996). Ο όρος «εξόρυξη δεδομένων» χρησιμοποιείται κυρίως από τους στατιστικολόγους, τους ερευνητές βάσεων δεδομένων, τα πληροφοριακά συστήματα MIS και τις επιχειρηματικές κοινότητες. Ο όρος «Discovery Knowledge in Databases» (KDD) χρησιμοποιείται γενικά για να αναφερθεί στη συνολική διαδικασία ανεύρεσης χρήσιμων γνώσεων από δεδομένα, που συνιστά ένα συγκεκριμένο βήμα σε αυτή τη διαδικασία. (Fayyad, et al., 1996· Peacock, 1998α· Han, & Kamber, 2000) Τα πρόσθετα βήματα στη διαδικασία KDD, όπως η προετοιμασία δεδομένων, η επιλογή δεδομένων, ο καθαρισμός των δεδομένων και η σωστή ερμηνεία των αποτελεσμάτων της διαδικασίας εξόρυξης δεδομένων, εξασφαλίζουν ότι οι χρήσιμες γνώσεις προέρχονται από τα δεδομένα.

Η εξόρυξη δεδομένων είναι μια επέκταση της παραδοσιακής ανάλυσης δεδομένων και των στατιστικών προσεγγίσεων, δεδομένου ότι ενσωματώνει αναλυτικές τεχνικές οι οποίες προέρχονται από μια σειρά επιστημονικών κλάδων που περιλαμβάνουν, αλλά δεν περιορίζονται για: α) αριθμητική ανάλυση, β) αντιστοίχηση προτύπων και περιοχές τεχνητής νοημοσύνης όπως μηχανική μάθηση, γ) νευρωνικά δίκτυα και γενετικοί αλγόριθμοι.

Ενώ πολλές εργασίες εξόρυξης δεδομένων ακολουθούν μια παραδοσιακή προσέγγιση υπολογισμού που βασίζεται στην υπόθεση, είναι συνηθισμένο να χρησιμοποιείται μια ευκαιριακή προσέγγιση που βασίζεται σε δεδομένα, η οποία ενθαρρύνει τους αλγορίθμους ανίχνευσης προτύπων να βρίσκουν χρήσιμες τάσεις, μοτίβα και σχέσεις. Ουσιαστικά, οι δύο τύποι προσεγγίσεων εξόρυξης δεδομένων διαφέρουν ως προς το α) αν επιδιώκουν να δημιουργήσουν μοντέλα ή, β) να βρουν μοτίβα.

Η πρώτη προσέγγιση, που αφορά τη δόμηση μοντέλων, είναι, εκτός από τα προβλήματα που είναι εγγενή από τα μεγάλα μεγέθη των συνόλων δεδομένων, παρόμοια με τις συμβατικές διερευνητικές στατιστικές μεθόδους. Ο στόχος είναι να παραχθεί μια συνολική σύνοψη ενός συνόλου δεδομένων για να προσδιοριστούν και να περιγραφούν τα κύρια χαρακτηριστικά του σχήματος της κατανομής (Hand, 1998). Παραδείγματα τέτοιων μοντέλων περιλαμβάνουν την κατάτμηση ανάλυσης συστάδων ενός συνόλου δεδομένων, ένα μοντέλο παλινδρόμησης για πρόβλεψη και ένας κανόνας ταξινόμησης που βασίζεται σε «δέντρα». Κατά τη δόμηση μοντέλων, γίνεται μερικές φορές διάκριση μεταξύ εμπειρικών και μηχανιστικών μοντέλων (Box and Hunter 1965· Cox 1990· Hand 1995). Το πρότερο (επίσης μερικές φορές αποκαλούμενο «επιχειρησιακό») επιδιώκει να μοντελοποιήσει σχέσεις χωρίς να τους στηρίξει σε οποιαδήποτε υποκείμενη θεωρία. Το τελευταίο (μερικές φορές αποκαλούμενο «ουσιαστικό» ή «φαινομενολογικό») βασίζεται σε κάποια θεωρία ή μηχανισμό για την υποκείμενη διαδικασία δημιουργίας δεδομένων. Η εξόρυξη δεδομένων, σχεδόν εξ ορισμού, ασχολείται κυρίως με τη λειτουργία.

Ο δεύτερος τύπος προσέγγισης εξόρυξης δεδομένων, η ανίχνευση προτύπων, επιδιώκει να εντοπίσει μικρές (αλλά ενδεχομένως σημαντικές) αναχωρήσεις από τον κανόνα, για να ανιχνεύσει ασυνήθιστα πρότυπα συμπεριφοράς. Παραδείγματα περιλαμβάνουν ασυνήθιστα μοντέλα δαπανών στη χρήση πιστωτικών καρτών (για

ανίχνευση απάτης), σποραδικές κυματομορφές στα ίχνη EEG και αντικείμενα με μοτίβα χαρακτηριστικών σε αντίθεση με άλλα. Γενικά, οι βάσεις δεδομένων των επιχειρήσεων, αποτελούν ένα μοναδικό πρόβλημα για την εξόρυξη προτύπων λόγω της πολυπλοκότητας τους. Η πολυπλοκότητα προκύπτει από ανωμαλίες όπως την ασυνέχεια, το θόρυβο, την αμφισημία και την ατέλεια (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Και ενώ οι περισσότεροι αλγόριθμοι εξόρυξης δεδομένων είναι σε θέση να διαχωρίσουν τα αποτελέσματα τέτοιων ασύνδετων χαρακτηριστικών για τον προσδιορισμό του πραγματικού προτύπου, η προγνωστική δύναμη των αλγορίθμων εξόρυξης μπορεί να μειωθεί, καθώς ο αριθμός αυτών των ανωμαλιών αυξάνεται (Rajagopalan, & Krovi, 2002).

1.4 Ιστορική αναδρομή εξόρυξης δεδομένων

Από τους αρχαίους χρόνους, οι άνθρωποι αναζητούσαν χρήσιμες πληροφορίες από τα δεδομένα με το χέρι. Ωστόσο, με τον ταχέως αυξανόμενο όγκο δεδομένων στη σύγχρονη εποχή, απαιτούνται πιο αυτόματες και αποτελεσματικές προσεγγίσεις εξόρυξης. Οι πρώιμες μέθοδοι όπως το θεώρημα του Bayes στις δεκαετίες του 1700 και η ανάλυση παλινδρόμησης κατά τις δεκαετίες του 1800, ήταν μερικές από τις πρώτες τεχνικές που χρησιμοποιήθηκαν για τον προσδιορισμό των μορφών στα δεδομένα.

Μετά από το 19^{ο} αιώνα, με την εξάπλωση, την έντονη παρουσία και τη συνεχώς αναπτυσσόμενη δύναμη της τεχνολογίας των υπολογιστών, η συλλογή δεδομένων και η αποθήκευση τους, αυξήθηκαν αξιοσημείωτα. Καθώς τα σύνολα δεδομένων είχαν αυξηθεί σε μέγεθος και πολυπλοκότητα, η άμεση ανάλυση δεδομένων εμφανιζόταν όλο και περισσότερο συνδεδεμένη με την αυτόματη επεξεργασία δεδομένων. Αυτό βοηθήθηκε από άλλες ανακαλύψεις στην επιστήμη των υπολογιστών, όπως τα νευρωνικά δίκτυα, την ομαδοποίηση, τους γενετικούς αλγόριθμους στη δεκαετία του 1950, αλλά και τα «δέντρα απόφασης» στη δεκαετία του 1960 και τα μηχανήματα φορέα υποστήριξης στη δεκαετία του '80. Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων σε σύνολα δεδομένων, με σκοπό την αποκάλυψη κρυφών μοτίβων (Kantardzic, 2003). Η τεχνολογία εξόρυξης δεδομένων, έχει

χρησιμοποιηθεί εδώ και πολλά χρόνια από πολλούς τομείς όπως επιχειρήσεις, επιστημονικές κοινότητες, καθώς και κυβερνήσεις. Χρησιμοποιήθηκε επίσης για τη διερεύνηση όγκων δεδομένων, όπως πληροφορίες για τους ταξιδιώτες αεροπορικών εταιρειών, δεδομένα πληθυσμού και στοιχεία μάρκετινγκ, για τη δημιουργία αναφορών έρευνας αγοράς, παρόλο που η αναφορά αυτή μερικές φορές δεν θεωρείται ως εξόρυξη δεδομένων.

Κατά τη δεκαετία του 1960, η τεχνολογία βάσεων δεδομένων και η τεχνολογία των πληροφοριών αναπτύχθηκαν σταδιακά από το βασικό σύστημα επεξεργασίας εγγράφων, σε ένα πιο περίπλοκο και πιο ισχυρό σύστημα βάσεων δεδομένων. Για παράδειγμα η ιεραρχική βάση δεδομένων και η βάση δεδομένων δικτύου, είναι τυπικά αντιπροσωπευτικά αυτής της εποχής, με ελάχιστη ανεξαρτησία και αφαίρεση δεδομένων. Κατά τη δεκαετία του 1970, εμφανίζονται σχεσιακές βάσεις δεδομένων, επιτρέποντας στους χρήστες πρόσβαση σε μια ευέλικτη γλώσσα και διεπαφή πρόσβασης δεδομένων, ενώ η τεχνολογία OLTP καθιστά την εφαρμογή τεχνολογίας σχεσιακής βάσης δεδομένων, δημοφιλή. Μέσα στη δεκαετία του '80, η άνοδος ενός ισχυρού συστήματος βάσεων δεδομένων, έρχεται να προτείνει πολλά προηγμένα μοντέλα δεδομένων,. Μετά το 2000, η δυνατότητα αποθήκευσης μεγάλων ποσοτήτων δεδομένων υπερβαίνει την ικανότητα ανάλυσης και κατανόησης του ανθρώπου, ενώ δεν υπάρχει κατάλληλο εργαλείο για να βοηθήσει στην εξαγωγή πληροφοριών και γνώσεων από τα δεδομένα. Η ύπαρξη συγκεκριμένων προτύπων και κανόνων μπορεί να βρεθεί μέσω εργαλείων εξόρυξης δεδομένων σε μεγάλο αριθμό δεδομένων, τα οποία μπορούν να παράσχουν τις απαραίτητες πληροφορίες για την εμπορική δραστηριότητα, την επιστημονική έρευνα και την ιατρική έρευνα και πολλούς άλλους τομείς (Weiping, & Yuming, 2013).

Η εξόρυξη δεδομένων περιλαμβάνει συνήθως τέσσερις κατηγορίες καθηκόντων (Usama, Piatetsky-Shapiro, & Smyth, 1996): 1) ταξινόμηση, (κατανομή των δεδομένων σε προκαθορισμένες ομάδες, 2) ομαδοποίηση, (ταξινόμηση σε ομάδες που δεν είναι προκαθορισμένες, οπότε ο αλγόριθμος θα προσπαθήσει να ομαδοποιήσει παρόμοια στοιχεία μαζί, 3) παλινδρόμηση, (εύρεση μιας συνάρτησης η οποία διαμορφώνει τα δεδομένα με το ελάχιστο σφάλμα, και, 4) σύνδεση κανόνα μάθησης και αναζήτησης σχέσεων μεταξύ μεταβλητών. Σύμφωνα με τον Han και τον Kamber (2001), οι λειτουργίες εξόρυξης δεδομένων περιλαμβάνουν τον χαρακτηρισμό

δεδομένων, τη διάκριση δεδομένων, την ανάλυση συσχέτισης, την ταξινόμηση, τη ομαδοποίηση και την ανάλυση εξέλιξης δεδομένων.

Ο χαρακτηρισμός των δεδομένων είναι μια σύνοψη των γενικών χαρακτηριστικών ή χαρακτηριστικών μιας κατηγορίας στόχων δεδομένων. Η διάκριση δεδομένων είναι μια σύγκριση των γενικών χαρακτηριστικών των αντικειμένων τάξης στόχου με τα γενικά χαρακτηριστικά των αντικειμένων από ένα σύνολο κατηγοριών αντίθεσης. Η ανάλυση της σύνδεσης είναι η ανακάλυψη κανόνων σύνδεσης που εμφανίζουν συνθήκες χαρακτηριστικού-τιμής που συμβαίνουν συχνά μαζί σε ένα συγκεκριμένο σύνολο δεδομένων. Η ταξινόμηση είναι η διαδικασία εύρεσης ενός συνόλου μοντέλων ή λειτουργιών που περιγράφουν και διακρίνουν τάξεις (κλάσεις) ή έννοιες δεδομένων, με σκοπό να είναι σε θέση να χρησιμοποιήσουν το μοντέλο για να προβλέψουν την κατηγορία αντικειμένων των οποίων η ετικέτα τάξεως είναι άγνωστη. Η ομαδοποίηση αναλύει αντικείμενα δεδομένων χωρίς να συμβουλευτεί ένα γνωστό μοντέλο τάξεως. Η ανάλυση εξέλιξης δεδομένων περιγράφει και μοντελοποιεί τάσεις για αντικείμενα των οποίων η συμπεριφορά μεταβάλλεται με το χρόνο (Han, & Kamber, 2001).

1.5 Εφαρμογές εξόρυξης δεδομένων

Η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί σε μια πληθώρα εφαρμογών και να αποτελέσει ένα πολύ χρήσιμο εργαλείο. Η εξόρυξη δεδομένων είναι μια σχετικά νέα τεχνολογία που δεν έχει ωριμάσει πλήρως. Παρ 'όλα αυτά, υπάρχουν αρκετές βιομηχανίες που την χρησιμοποιούν ήδη σε τακτική βάση. Ορισμένες από αυτές τις οργανώσεις περιλαμβάνουν καταστήματα λιανικής πώλησης, νοσοκομεία, τράπεζες και ασφαλιστικές εταιρείες. Πολλοί από αυτούς τους οργανισμούς συνδυάζουν την εξόρυξη δεδομένων με στατιστικά στοιχεία, αναγνώριση προτύπων και άλλα σημαντικά εργαλεία. Η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για την εύρεση σχεδίων και συνδέσεων που διαφορετικά θα ήταν δύσκολο να βρεθούν. Αυτή η τεχνολογία είναι δημοφιλής σε πολλές επιχειρήσεις επειδή τους επιτρέπει να μάθουν περισσότερα για τους πελάτες τους και να κάνουν έξυπνες αποφάσεις μάρκετινγκ. (Bharati, & Ramageri, 2010)

Η εξόρυξη δεδομένων έχει μεγάλες δυνατότητες βελτίωσης των συστημάτων υγείας. Χρησιμοποιεί δεδομένα και αναλύσεις για τον εντοπισμό βέλτιστων πρακτικών που βελτιώνουν τη φροντίδα και μειώνουν το κόστος (Antonie, Zaiane, & Coman, 2001). Οι ερευνητές χρησιμοποιούν προσεγγίσεις εξόρυξης δεδομένων όπως πολυδιάστατες βάσεις δεδομένων, μηχανική μάθηση, οπτικοποίηση δεδομένων και στατιστικές. Η εξόρυξη μπορεί να χρησιμοποιηθεί για την πρόβλεψη του όγκου των ασθενών σε κάθε κατηγορία. Αναπτύσσονται διαδικασίες που διασφαλίζουν ότι οι ασθενείς λαμβάνουν την κατάλληλη φροντίδα στο σωστό μέρος και την κατάλληλη στιγμή. Η εξόρυξη δεδομένων μπορεί επίσης να βοηθήσει τους ασφαλιστές υγειονομικής περίθαλψης να ανιχνεύσουν απάτες και καταχρήσεις.

Η ανάλυση του καλαθιού αγοράς είναι μια τεχνική μοντελοποίησης βασισμένη σε μια θεωρία ότι κάποιος αγοράζει μια συγκεκριμένη ομάδα αντικειμένων είναι πιθανότερο να αγοράσει μια άλλη ομάδα αντικειμένων. Αυτή η τεχνική μπορεί να επιτρέψει στον πωλητή να κατανοήσει την αγοραστική συμπεριφορά ενός αγοραστή. Αυτές οι πληροφορίες μπορεί να τον βοηθήσουν να γνωρίζει τις ανάγκες του αγοραστή και να αλλάζει ανάλογα τη διάταξη του καταστήματος. Χρησιμοποιώντας τη διαφορική ανάλυση σύγκρισης των αποτελεσμάτων μεταξύ των διαφόρων καταστημάτων, μπορεί να γίνει διάκριση μεταξύ πελατών σε διαφορετικές δημογραφικές ομάδες.

Υπάρχει ένας νέος αναδυόμενος τομέας που ονομάζεται Εκπαιδευτικό Data Mining (Educational Data Mining) και περιλαμβάνει ανησυχίες για την ανάπτυξη μεθόδων που ανακαλύπτουν τη γνώση από τα δεδομένα που προέρχονται από τα εκπαιδευτικά περιβάλλοντα. Οι στόχοι της εφαρμογής αυτής, προσδιορίζονται ως πρόβλεψη της μελλοντικής μαθησιακής συμπεριφοράς των μαθητών, μελετώντας τα αποτελέσματα της εκπαιδευτικής υποστήριξης και προωθώντας τις επιστημονικές γνώσεις σχετικά με τη μάθηση. Η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί από ένα ίδρυμα για να λάβει ακριβείς αποφάσεις και επίσης να προβλέψει τα αποτελέσματα του κάθε μαθητή. Με τα αποτελέσματα αυτά, το ίδρυμα μπορεί να επικεντρωθεί σε αυτά που πρέπει να διδάξουν αλλά και τον τρόπο διδασκαλίας. Τα μαθησιακά πρότυπα των μαθητών μπορούν να συλλεχθούν και να χρησιμοποιηθούν για την ανάπτυξη τεχνικών για τη διδασκαλία τους.

Η γνώση είναι το καλύτερο στοιχείο που θα διαθέτει μια επιχείρηση παραγωγής. Τα εργαλεία εξόρυξης δεδομένων μπορούν να είναι πολύ χρήσιμα για να ανακαλύψουν τα πρότυπα σε περίπλοκη διαδικασία παραγωγής (Rukshan, Menik, & Chandrika, 2009). Η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί σε σχεδιασμό σε επίπεδο συστήματος για την εξαγωγή των σχέσεων μεταξύ της αρχιτεκτονικής του προϊόντος, του χαρτοφυλακίου προϊόντος και των αναγκών των πελατών. Μπορεί επίσης να χρησιμοποιηθεί για την πρόβλεψη του χρόνου ανάπτυξης του προϊόντος, του κόστους και των εξαρτήσεων μεταξύ άλλων καθηκόντων.

Η Διαχείριση Σχέσεων Πελατών (Customer Relationship Management), αφορά αποκλειστικά την απόκτηση και διατήρηση πελατών, τη βελτίωση της εμπιστοσύνης των πελατών και την εφαρμογή στρατηγικών που εστιάζουν στους πελάτες. Για να διατηρηθεί μια σωστή σχέση με έναν πελάτη, μια επιχείρηση χρειάζεται να συλλέξει δεδομένα και να αναλύσει τις πληροφορίες. Σε αυτό το σημείο, η εξόρυξη δεδομένων μπορεί να παίξει έναν πολύ σημαντικό ρόλο (Rukshan, Menik, & Chandrika, 2009). Με τις τεχνολογίες εξόρυξης δεδομένων τα συλλεχθέντα δεδομένα μπορούν να χρησιμοποιηθούν για ανάλυση. Αντί να υπάρχει σύγχηση σχετικά με του που πρέπει μια επιχείρηση να επικεντρωθεί για να διατηρήσει τον πελάτη, με την εξόρυξη δεδομένων μπορούν να έχουν έτοιμα, φιλτραρισμένα αποτελέσματα.

Επιπλέον, δισεκατομμύρια δολάρια έχουν χαθεί από διάφορες απάτες στον επιχειρησιακό τομέα. Οι παραδοσιακές μέθοδοι ανίχνευσης απάτης είναι χρονοβόρες και πολύπλοκες. Η τεχνολογία εξόρυξης δεδομένων βοηθά στην παροχή ουσιαστικών προτύπων και τη μετατροπή των δεδομένων σε πληροφορίες. Οποιεσδήποτε πληροφορίες είναι έγκυρες και χρήσιμες, αποτελούν σημαντική γνώση. Ένα τέλειο σύστημα ανίχνευσης απάτης πρέπει να προστατεύει τις πληροφορίες όλων των χρηστών. Μια εποπτευόμενη μέθοδος περιλαμβάνει συλλογή αρχείων δειγμάτων. Αυτά τα αρχεία είναι ταξινομημένα ως επικίνδυνα ή μη επικίνδυνα. Ένα μοντέλο κατασκευάζεται χρησιμοποιώντας αυτά τα δεδομένα και ο αλγόριθμος γίνεται για να προσδιοριστεί εάν κάποιο αρχείο είναι παράνομο ή όχι.

Οποιαδήποτε ενέργεια που θα θέσει σε κίνδυνο την ακεραιότητα και την εμπιστευτικότητα ενός πόρου αποτελεί μια εισβολή. Τα αμυντικά μέτρα για την αποφυγή διείσδυσης περιλαμβάνουν έλεγχο ταυτότητας χρήστη, αποφυγή σφαλμάτων

προγραμματισμού και προστασία πληροφοριών. Η εξόρυξη δεδομένων μπορεί να βοηθήσει στη βελτίωση της ανίχνευσης εισβολών προσθέτοντας ένα επίπεδο εστίασης στην ανίχνευση ανωμαλιών. Βοηθά έναν αναλυτή να διακρίνει μια δραστηριότητα από την κοινή καθημερινή δραστηριότητα δικτύου. Η εξόρυξη δεδομένων βοηθά επίσης στην εξαγωγή δεδομένων που είναι πιο συναφή με το πρόβλημα (Bharati, & Ramageri, 2010).

Επίσης, οι αρχές επιβολής του νόμου μπορούν να χρησιμοποιήσουν τεχνικές εξόρυξης για να διερευνήσουν εγκλήματα, να παρακολουθήσουν την επικοινωνία των ύποπτων τρομοκρατών. Αυτό το αρχείο περιλαμβάνει εξόρυξη κειμένων επίσης. Αυτή η διαδικασία επιδιώκει να βρει εύχρηστα πρότυπα σε δεδομένα τα οποία είναι συνήθως τη μορφή ενός αδόμητου κειμένου. Το δεδομένα που συλλέγονται από προηγούμενες έρευνες συγκρίνονται μεταξύ τους και δημιουργούν ένα μοντέλο ανίχνευσης ψεύδους. Με αυτό το μοντέλο οι διαδικασίες μπορούν να αναπτυχθούν ανάλογα με τις ανάγκες κάθε φορά.

Η παραδοσιακή έρευνα αγοράς μπορεί να βοηθήσει στον διαχωρισμό πελατών, αλλά η εξόρυξη δεδομένων προχωράει πιο βαθιά και αυξάνει την αποτελεσματικότητα της αγοράς. Η εξόρυξη δεδομένων βοηθά στην ευθυγράμμιση των πελατών σε ένα ξεχωριστό τμήμα και μπορεί να προσαρμόσει τις ανάγκες σύμφωνα με τους πελάτες. Η αγορά αφορά πάντα την διατήρηση των πελατών (Bharati, & Ramageri, 2010). Η εξόρυξη δεδομένων επιτρέπει την εύρεση ενός τμήματος πελατών βασισμένου σε κάποια συγκεκριμένη ιδιαιτερότητα. Με αυτόν τον τρόπο η κάθε επιχείρηση έχει τη δυνατότητα να προσφέρει ειδικές προσφορές και να αυξήσει την ικανοποίησή των πελατών αυτών.

Με την ηλεκτρονικό τραπεζικό σύστημα το οποίο πλέον επικρατεί, παράγεται ένας τεράστιος όγκος δεδομένων με τις νέες συναλλαγές. Η εξόρυξη δεδομένων μπορεί να συμβάλει στην επίλυση επιχειρηματικών προβλημάτων στον τραπεζικό και χρηματοπιστωτικό τομέα, βρίσκοντας πρότυπα, αιτίες και συσχετισμούς στις επιχειρηματικές πληροφορίες και στις τιμές της αγοράς, οι οποίες δεν είναι άμεσα εμφανείς στους διαχειριστές, επειδή τα δεδομένα όγκου είναι πολύ μεγάλα ή δημιουργούνται πολύ γρήγορα (Rukshan, Menik, & Chandrika, 2009). Οι

διαχειριστές μπορούν να βρουν αυτές τις πληροφορίες για καλύτερη κατάτμηση, στόχευση, απόκτηση και διατήρηση ενός κερδοφόρου πελάτη.

Η εταιρική εποπτεία είναι η παρακολούθηση της συμπεριφοράς ενός ατόμου ή μιας ομάδας από μια εταιρεία. Τα στοιχεία που συλλέγονται χρησιμοποιούνται συχνότερα για σκοπούς μάρκετινγκ ή πωλούνται σε άλλες εταιρείες, αλλά επίσης μοιράζονται τακτικά με κυβερνητικές υπηρεσίες. Μπορεί να χρησιμοποιηθεί από την επιχείρηση για να προσαρμόσει τα προϊόντα της, στις επιθυμίες των πελατών της. Τα δεδομένα μπορούν να χρησιμοποιηθούν για σκοπούς άμεσου μάρκετινγκ, όπως στοχευμένες διαφημίσεις στο Google και στο Yahoo, όπου οι διαφημίσεις στοχεύουν στον χρήστη της μηχανής αναζήτησης, αναλύοντας το ιστορικό αναζήτησης και τα μηνύματα ηλεκτρονικού ταχυδρομείου.

Κεφάλαιο 2^{ο}- Μελέτη Δεδομένων

2.1 Περιγραφή προτύπων και μοντέλων

Υπάρχουν δύο βασικοί τύποι μοντέλων εξόρυξης δεδομένων. Αυτά είναι: Προγνωστικά και περιγραφικά. Το περιγραφικό μοντέλο αναγνωρίζει τα σχέδια ή τις σχέσεις στα δεδομένα και ανακαλύπτει τις ιδιότητες των δεδομένων που μελετήθηκαν. Για παράδειγμα, Clustering, Summarization, Rule Association, Sequence discovery κλπ. Η ομαδοποίηση είναι σαν την ταξινόμηση, ωστόσο οι ομάδες δεν είναι προκαθορισμένες, αλλά είναι και πάλι καλά καθορισμένες μόνο από τα δεδομένα. Αναφέρεται επίσης ως ακαδημαϊκή μάθηση ή υποδιαίρεση. Είναι ο τοίχος ή ο χωρισμός των δεδομένων σε συλλογές ή ομάδες. Τα clusters είναι καλά καθορισμένα με την εκμάθηση της απόδοσης των δεδομένων από τους ειδικούς τομέα. Ο όρος διαίρεση χρησιμοποιείται σε πολύ συγκεκριμένο πλαίσιο.

Οι προγνωστικές αναλύσεις έχουν οριστεί από τον Delen & Demirkan (2013), ώστε να έχουν προτυποποίηση δεδομένων ως προϋπόθεση κατά τη λήψη αξιόπιστων προβλέψεων για το μέλλον χρησιμοποιώντας επιχειρηματικές προβλέψεις και προσομοιώσεις. Οι διαφορετικές μελέτες των Lechevalier, Narayanan, & Rachuri (2014) ορίζουν τα προγνωστικά ως εργαλείο που «χρησιμοποιεί τις στατιστικές τεχνικές, τη μηχανική μάθηση και την εξόρυξη δεδομένων να ανακαλυφθούν γεγονότα για να πραγματοποιηθούν προβλέψεις για άγνωστα μελλοντικά γεγονότα ", στη διερεύνηση ενός πλαισίου ειδικά για το συγκεκριμένο τομέα. Το μοντέλο πρόβλεψης κάνει πρόβλεψη για τιμές μη αναγνωρισμένων δεδομένων χρησιμοποιώντας τις προσδιορισμένες τιμές.

Περιγραφικά μοντέλα

Η σύνοψη είναι η διαδικασία παροχής των πληροφοριών ανακεφαλαίωσης από τα δεδομένα. Ο κανόνας σύνδεσης ανακαλύπτει τη σύνδεση μεταξύ των διαφορετικών χαρακτηριστικών και είναι μια διαδικασία δύο βημάτων: Η εύρεση όλων των συχνών

αντικειμένων θέτει ένα ισχυρό κανόνα δημιουργίας δεσμών από τα συνηθισμένα σύνολα στοιχείων. Σύμφωνα με τους Mortenson, Doherty & Robinson (2014), οι περιγραφικές αναλύσεις ανασυντάσσουν και μετασχηματίζουν τα δεδομένα σε εκφραστικές πληροφορίες για την αναφορά και τη φροντίδα ενός ατόμου, αλλά επίσης επιτρέπουν την εμπεριστατωμένη εξέταση για να απαντήσουν σε ερωτήματα όπως "τι συνέβη;" και " (SAP) (2014) περιγράφει επίσης τα περιγραφικά στοιχεία ανάλυσης ως εφαρμογές πίνακα ελέγχου που υποστηρίζουν την υλοποίηση της ανάπτυξης στις πωλήσεις και τη διαχείριση των διαδικασιών, επιτρέποντας την παρακολούθηση σε πραγματικό χρόνο.

Η συνοπτική παρουσίαση μπορεί να παρατηρηθεί ως συμπιέζοντας ένα δεδομένο σύνολο συναλλαγών σε ένα μικρότερο σύνολο σχεδίων, ενώ συγχρόνως θυμίζει τις ανώτατες πιθανές πληροφορίες. Η σύνοψη είναι μια κοινή και έγκυρη αλλά συχνά χρονοβόρα μέθοδος για την εξέταση μεγάλων συνόλων δεδομένων. Για παράδειγμα, ας υποθέσουμε ότι κάποιος θέλει να εξετάσει δεδομένα απογραφής για να εκτιμήσει τη σχέση μεταξύ του επιπέδου εκπαίδευσης και του μισθού σε μια χώρα. Μια πολύ πυκνή περίληψη της απογραφής μπορεί να παρατηρηθεί με τον σχεδιασμό του μέσου μισθού ανά επίπεδο εκπαίδευσης. Αυτή η περίληψη θα είναι επαρκής για κάποιες αποφάσεις, αλλά άλλοι ίσως χρειαστούν περισσότερο χρόνο για να πραγματοποιήσουν μια πιο υγιή κατανόηση των δεδομένων.

Τα δεδομένα ακολουθίας είναι εναλλακτικό εργαλείο που χρησιμοποιείται ως περιγραφικό μοντέλο. Οι ακολουθίες κυνηγιού αγορών καταναλωτών, τα δεδομένα ιατρικής περίθαλψης και τα δεδομένα που σχετίζονται με φυσικές καταστροφές, τα δεδομένα των επιστημών και των μηχανικών διαδικασιών, τα δεδομένα των αποθεμάτων και των αγορών, τα σχέδια κατοχής τηλεφώνου, οι ακολουθίες εκτέλεσης πακέτων και τα δεδομένα συναρμολογήσεων είναι μερικές περιπτώσεις δεδομένων αλληλουχίας. Τα δεδομένα της ακολουθίας συνδέουν κανονικά τα δεδομένα αλλά δεν προβλέπουν στο μέλλον, αν και μπορεί να ληφθεί απόφαση μετά την έκφρασή της.

Η ανάλυση συμπλέγματος είναι ένας άλλος τύπος Περιγραφικού μοντέλου που συγκεντρώνει αντικείμενα (παρατηρήσεις, γεγονότα) με βάση τις πληροφορίες που βρέθηκαν στα δεδομένα που περιγράφουν τα αντικείμενα ή τις σχέσεις τους. Ο

στόχος είναι ότι τα αντικείμενα μιας ομάδας θα είναι παρόμοια (ή σχετίζονται) με ένα άλλο και διαφορετικά από (ή μη σχετιζόμενα με) τα αντικείμενα σε άλλες ομάδες.

Προγνωστικό Μοντέλο

Οι μέθοδοι χρονοσειράς είναι επίσης μέρη Προγνωστικών μοντέλων, χρησιμοποιώντας μεθόδους όπως κινητούς μέσους όρους, εκθετική εξομάλυνση, αυτορυθμιζόμενα μοντέλα, γραμμική, μη γραμμική και λογιστική παλινδρόμηση(Souza, 2014). Η βάση δεδομένων αποτελείται από ακολουθίες τιμών ή συμβάντων που λαμβάνονται επαναλαμβανόμενες μετρήσεις χρόνου. Οι τιμές τυπικά μετρούνται σε ίσα χρονικά διαστήματα (π.χ. ωριαία, ημερήσια, εβδομαδιαία). Οι βάσεις δεδομένων χρονολογικών σειρών είναι δημοφιλείς σε πολλές εφαρμογές, όπως η ανάλυση χρηματιστηριακών αγορών, η πρόβλεψη οικονομικών και πωλήσεων, η ανάλυση του προϋπολογισμού, οι μελέτες χρησιμότητας, οι μελέτες αποθεμάτων, οι προβολές απόδοσης, οι προβολές φόρτου εργασίας, ο έλεγχος της διαδικασίας και του ελέγχου ποιότητας, η παρατήρηση φυσικών φαινομένων, επιστημονικά και μηχανολογικά πειράματα και ιατρικές θεραπείες. Μια βάση δεδομένων χρονοσειρών είναι επίσης μια βάση δεδομένων αλληλουχίας. Ωστόσο, μια βάση δεδομένων αλληλουχίας είναι οποιαδήποτε βάση δεδομένων που αποτελείται από ακολουθίες παραγγελιών, με ή χωρίς συγκεκριμένες αντιλήψεις χρόνου. Για παράδειγμα, οι ακολουθίες μετατόπισης ιστοσελίδων και οι ακολουθίες συναλλαγών αγορών πελατών είναι δεδομένα ακολουθίας, αλλά μπορεί να μην είναι δεδομένα χρονοσειράς.

2.2 Τεχνικές μελέτης δεδομένων για τη δημιουργία προτύπων

Δύο τεχνικές μοντελοποίησης δεδομένων που σχετίζονται με ένα περιβάλλον αποθήκευσης δεδομένων είναι η μοντελοποίηση ER (Entity-Relationship) και η πολυδιάστατη μοντελοποίηση. Η ER modeling παράγει ένα μοντέλο δεδομένων της συγκεκριμένης περιοχής ενδιαφέροντος, χρησιμοποιώντας δύο βασικές έννοιες: οντότητες και τις σχέσεις μεταξύ αυτών των οντοτήτων. Το μοντέλο ER είναι ένα εργαλείο αφαίρεσης επειδή μπορεί να χρησιμοποιηθεί για να κατανοήσει και να

απλοποιήσει τις διφορούμενες σχέσεις δεδομένων στον κόσμο των επιχειρήσεων και τα περίπλοκα περιβάλλοντα συστημάτων.

Η πολυδιάστατη μοντελοποίηση χρησιμοποιεί τρεις βασικές έννοιες: τα μέτρα, τα γεγονότα και τις διαστάσεις. Η πολυδιάστατη μοντελοποίηση είναι ισχυρή στην εκπροσώπηση των απαιτήσεων του επιχειρηματικού χρήστη στο πλαίσιο των πινάκων βάσης δεδομένων. Και η ER και η πολυδιάστατη μοντελοποίηση μπορούν να χρησιμοποιηθούν για να δημιουργήσουν ένα αφηρημένο μοντέλο του ειδικού αντικειμένου (Patel, & Patel, 2012).

Η έννοια ER (Entity-Relationship) χρησιμοποιείται εκτεταμένα για το σχεδιασμό βάσης δεδομένων σε περιβάλλον σχεσιακών βάσεων δεδομένων, η οποία τονίζει τις λειτουργίες της ημέρας σήμερα. Η πολυδιάστατη μοντελοποίηση δεδομένων (MD), από την άλλη πλευρά, είναι κρίσιμη για τον σχεδιασμό της αποθήκευσης δεδομένων, η οποία στοχεύει στην υποστήριξη της διαχείρισης των αποφάσεων. Υποστηρίζει τη λήψη αποφάσεων επιτρέποντας στους χρήστες να δουλεύουν για πιο λεπτομερείς πληροφορίες. Όταν σχεδιάζουμε ένα μοντέλο MD ανεξάρτητα από το αν αυτό περιλαμβάνει την αναγνώριση ενός γεγονότος, διαστάσεων και χαρακτηριστικών μέτρησης (Patel, & Patel, 2012).

2.3 Είδη δεδομένων

Όταν κάποιος επιχειρεί να δημιουργήσει ένα μοντέλο εξόρυξης ή μια δομή εξόρυξης, πρέπει να ορίσει τους τύπους δεδομένων για κάθε μια από τις στήλες της δομής εξόρυξης. Ο τύπος δεδομένων αναφέρει στον μηχανισμό ανάλυσης αν τα δεδομένα στην πηγή δεδομένων είναι αριθμητικά ή κείμενα και πώς πρέπει να επεξεργάζονται τα δεδομένα. Για παράδειγμα, αν τα δεδομένα προέλευσης περιέχουν αριθμητικά δεδομένα, μπορούν να καθορίσουν αν οι αριθμοί θα αντιμετωπίζονται ως ακέραιοι αριθμοί ή χρησιμοποιώντας δεκαδικά ψηφία (Guyer, & Rabeler, 2018).

Εάν κάποιος θέλει να δημιουργήσει απευθείας το μοντέλο εξόρυξης χρησιμοποιώντας τις Επεκτάσεις Εξόρυξης Δεδομένων (DMX), μπορεί να καθορίσει τον τύπο

δεδομένων για κάθε στήλη, με τον τρόπο που καθορίζεται το μοντέλο. Οι υπηρεσίες ανάλυσης θα δημιουργήσουν την αντίστοιχη δομή εξόρυξης με τους καθορισμένους τύπους δεδομένων ταυτόχρονα. Εάν δημιουργήσει το μοντέλο εξόρυξης ή τη δομή εξόρυξης χρησιμοποιώντας έναν οδηγό, οι Υπηρεσίες ανάλυσης θα προτείνουν έναν τύπο δεδομένων ή μπορεί να επιλέξει έναν τύπο δεδομένων από μια λίστα.

Εάν επιθυμεί να αλλάξει τον τύπο δεδομένων μιας στήλης, πρέπει πάντα να επανεπεξεργαστεί τη δομή εξόρυξης που βασίζεται σε αυτή τη δομή. Μερικές φορές, αν αλλάξει τον τύπο δεδομένων, αυτή η στήλη δεν μπορεί πλέον να χρησιμοποιηθεί σε ένα συγκεκριμένο μοντέλο. Σε αυτήν την περίπτωση, οι υπηρεσίες ανάλυσης είτε θα εμφανίσουν ένα σφάλμα κατά την επανεπεξεργασία του μοντέλου, είτε θα επεξεργαστούν το μοντέλο αλλά θα αφήσουν εκτός αυτής τη συγκεκριμένη στήλη (Guyer, & Rabeler, 2018).

Με την αυξανόμενη χρήση εφαρμογών βάσεων δεδομένων, η εξόρυξη ενδιαφέρουσας πληροφορίας από τεράστιες βάσεις δεδομένων προκαλεί μεγάλη ανησυχία και έχουν προταθεί διάφοροι αλγόριθμοι εξόρυξης τα τελευταία χρόνια. Όπως γνωρίζουμε, τα δεδομένα που επεξεργάζονται στην εξόρυξη δεδομένων μπορούν να ληφθούν από πολλές πηγές στις οποίες μπορούν να χρησιμοποιηθούν διαφορετικοί τύποι δεδομένων. Ωστόσο, κανένας αλγόριθμος δεν μπορεί να εφαρμοστεί σε όλες τις εφαρμογές λόγω της δυσκολίας προσαρμογής των τύπων δεδομένων στον αλγόριθμο. Η επιλογή ενός κατάλληλου αλγόριθμου εξόρυξης δεδομένων βασίζεται όχι μόνο στον στόχο της εφαρμογής, αλλά και στην προσαρμοστικότητα των δεδομένων. Επομένως, ο μετασχηματισμός του μη προσαρμοσμένου τύπου δεδομένων σε έναν στόχο είναι επίσης σημαντικός για την εξόρυξη δεδομένων, αλλά η εργασία είναι συχνά κουραστική ή πολύπλοκη, καθώς υπάρχουν πολλοί τύποι δεδομένων στον πραγματικό κόσμο. Η συγχώνευση παρόμοιων τύπων δεδομένων ενός συγκεκριμένου επιλεγμένου αλγόριθμου εξόρυξης σε έναν γενικευμένο τύπο δεδομένων φαίνεται να είναι μια καλή προσέγγιση για τη μείωση της πολυπλοκότητας του μετασχηματισμού (Mon-Fong, Shian-Shyong, Shan-Yi, 1999).

2.4 Αποτελέσματα αλγορίθμων εξόρυξης δεδομένων

Ένας αλγόριθμος εξόρυξης δεδομένων είναι ένα σύνολο ευρετικών και υπολογισμών που δημιουργούν ένα μοντέλο εξόρυξης δεδομένων από δεδομένα (Microsoft, 2016). Μπορεί να είναι μια πρόκληση για κάποιον να επιλέξει τον κατάλληλο ή τον καλύτερα προσαρμοσμένο αλγόριθμο που θα εφαρμοστεί για την επίλυση ενός συγκεκριμένου προβλήματος. Παρόλο που κάποιος μπορεί να χρησιμοποιήσει διαφορετικούς αλγόριθμους για να εκτελέσει τις ίδιες εργασίες, κάθε αλγόριθμος αποφέρει ένα διαφορετικό σύνολο αποτελεσμάτων και ορισμένοι αλγόριθμοι μπορούν ακόμη να παράγουν περισσότερους από έναν τύπους αποτελεσμάτων. Μερικοί αλγόριθμοι μπορούν να εκτελέσουν μια διαδικασία ταξινόμησης, δηλαδή, μπορούν να προβλέψουν μία ή περισσότερες διακριτές μεταβλητές, με βάση τα άλλα χαρακτηριστικά του συνόλου δεδομένων.

Μερικοί αλγόριθμοι εκτελούν σκοπούς παλινδρόμησης και μπορούν να προβλέψουν περισσότερες ή συνεχείς μεταβλητές βάσει των άλλων χαρακτηριστικών στο σύνολο δεδομένων. Όπως επισήμανε η Microsoft (2016), ορισμένοι αλγόριθμοι μπορούν να εκτελούν διαχωρισμό, δηλαδή διαιρούν τα δεδομένα σε ομάδες που έχουν παρόμοιες ιδιότητες. Ενώ ορισμένοι αλγόριθμοι μπορούν να είναι συνειρμικοί με την εύρεση συσχετισμών μεταξύ διαφορετικών χαρακτηριστικών σε ένα σετ, μερικοί μπορούν να χρησιμοποιηθούν για διαδικασίες ανάλυσης αλληλουχίας, δηλαδή μπορούν να χρησιμοποιηθούν για να συνοψίσουν μια αλληλουχία ή επεισόδια δεδομένων, όπως μια ροή διαδρομής ιστού (Microsoft, 2016). Ωστόσο, όλοι οι προαναφερθέντες τύποι αλγορίθμων μπορούν να ταξινομηθούν σε δύο μεγάλες κατηγορίες: Εποπτευόμενοι εκμάθησης και μη εποπτευόμενοι αλγόριθμοι εκμάθησης.

Οι αλγόριθμοι εποπτευόμενης μάθησης είναι αυτοί για τους οποίους οι τιμές χαρακτηριστικών κλάσης για το σύνολο δεδομένων είναι γνωστές πριν από την εκτέλεση του αλγορίθμου. Αυτά τα δεδομένα ονομάζονται δεδομένα εκπαίδευσης (Gundecha, & Liu, 2012). Οι περιπτώσεις σε αυτό το σετ είναι πλειάδες στη μορφή (x, y) όπου το x είναι ένα διάνυσμα και το y είναι το χαρακτηριστικό κλάσης, συνήθως κλιμακωτό. Η εποπτευόμενη μάθηση χτίζει ένα μοντέλο που χαρτογραφεί το x στο y. Ο στόχος είναι να βρεθεί μια χαρτογράφηση m (.) τέτοια ώστε m (x) = y.

Επίσης, παρέχεται το μη επισημασμένο σύνολο δεδομένων ή σύνολο δεδομένων δοκιμών, όπου οι περιπτώσεις είναι σε μορφή (x,?) και οι y είναι άγνωστες. Δεδομένου ότι το m (.) που αντλήθηκε από τα δεδομένα εκπαίδευσης και το x μιας μη επισημασμένης παρουσίας, μπορεί να υπολογιστεί το m (x) που οδηγεί στην πρόβλεψη της ετικέτας για την περίπτωση που δεν έχει επισημανθεί (Zafarani, Abbasi, & Liu, 2014).

Κανονικά, όταν συζητιέται η μη εποπτευόμενη μάθηση, οι περισσότεροι ερευνητές επικεντρώνονται στην ομαδοποίηση (Gundecha, & Liu, 2012). Κατά την ομαδοποίηση, τα δεδομένα συχνά δεν έχουν επισημανθεί. Έτσι, η ετικέτα για κάθε περίπτωση δεν είναι γνωστή στον αλγόριθμο ομαδοποίησης. Αυτή είναι η κύρια διαφορά μεταξύ της εποπτευόμενης και της μη εποπτευόμενης μάθησης. Οποιοσδήποτε αλγόριθμος ομαδοποίησης απαιτεί μέτρηση απόστασης. Οι περιπτώσεις τοποθετούνται σε διαφορετικά σμήνη με βάση την απόσταση τους από άλλες περιπτώσεις (Han, & Kamber, 2010). Το πιο δημοφιλές μέτρο απόστασης για συνεχή χαρακτηριστικά είναι η ευκλείδεια απόσταση: d (X, Y) = √ (x1-y1) 2+ (x2-y2) 2 + ... + (xn-yn)

Κεφάλαιο 3^{ο}- Περιγραφή WEKA

3.1 Πρόγραμμα WEKA – περιγραφή

Το Weka περιέχει μια συλλογή εργαλείων απεικόνισης και αλγορίθμων για την ανάλυση δεδομένων και τη μέθοδο πρόβλεψης , μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτές τις λειτουργίες (Witten, Frank, & Hall. 2011). Η αρχική μη Weba έκδοση του Weka ήταν ένας Tcl / Tk front-end σε αλγόριθμους μοντελοποίησης (κυρίως τρίτου μέρους) που εφαρμόστηκαν σε άλλες γλώσσες προγραμματισμού, καθώς και βοηθητικά προγράμματα προεπεξεργασίας δεδομένων στο C και ένα σύστημα βασισμένο σε Makefile για λειτουργία μηχανής μάθησης. Αυτή η αρχική έκδοση σχεδιάστηκε πρωτίστως ως εργαλείο για την ανάλυση δεδομένων από αγροτικούς τομείς, αλλά η πιο πρόσφατη πλήρως Java (Weka 3), για την οποία ξεκίνησε η ανάπτυξη το 1997, χρησιμοποιείται πλέον σε πολλούς διαφορετικούς τομείς εφαρμογής, ιδίως για εκπαιδευτικούς σκοπούς και έρευνα (Garner, et al., 1995). Τα πλεονεκτήματα του Weka περιλαμβάνουν:

- Δωρεάν διαθεσιμότητα βάσει της Γενικής Δημόσιας Άδειας GNU .
- Φορητότητα, δεδομένου ότι υλοποιείται πλήρως στη γλώσσα προγραμματισμού Java και επομένως λειτουργεί σχεδόν σε οποιαδήποτε σύγχρονη πλατφόρμα υπολογιστών.
- Μια ολοκληρωμένη συλλογή τεχνικών προεπεξεργασίας και μοντελοποίησης δεδομένων.
- Ευκολία στη χρήση λόγω των γραφικών διεπαφών χρήστη.

Ο Weka υποστηρίζει αρκετά καθιερωμένα καθήκοντα εξόρυξης δεδομένων, πιο συγκεκριμένα, προεπεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, παλινδρόμηση, οπτικοποίηση και επιλογή χαρακτηριστικών. Όλες οι τεχνικές του Weka βασίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα ενιαίο αρχείο ή σχέση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό χαρακτηριστικών (κανονικά, αριθμητικά ή ονομαστικά χαρακτηριστικά, αλλά υποστηρίζονται επίσης μερικοί άλλοι τύποι χαρακτηριστικών).

Το Weka παρέχει πρόσβαση σε βάσεις δεδομένων SQL χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφεται από ένα ερώτημα βάσης δεδομένων. Το Weka παρέχει πρόσβαση σε βαθιά μάθηση με Deeplearning4j. Δεν είναι ικανό για εξόρυξη δεδομένων πολλαπλών σχέσεων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής συνδεδεμένων πινάκων βάσης δεδομένων σε έναν ενιαίο πίνακα που είναι κατάλληλος για επεξεργασία με χρήση του Weka (Reutemann, Pfahringer, & Frank, 2004). Ένας άλλος σημαντικός τομέας, που επί του παρόντος δεν καλύπτεται από τους αλγόριθμους που περιλαμβάνονται στην κατανομή Weka, είναι η μοντελοποίηση ακολουθιών.

3.2 Ιστορική αναδρομή

Το σχέδιο WEKA χρηματοδοτήθηκε από τη διοίκηση της Νέας Ζηλανδίας από το 1993 μέχρι πρόσφατα. Η αρχική αίτηση χρηματοδότησης κατατέθηκε στα τέλη του 1992 και δήλωσε τους στόχους του έργου ως εξής: «Το πρόγραμμα στοχεύει στην κατασκευή ενός υπερσύγχρονου εξοπλισμού για την ανάπτυξη τεχνικών μηχανικής μάθησης και διερεύνησης της εφαρμογής τους σε βασικούς τομείς της οικονομίας της Νέας Ζηλανδίας.

Ειδικότερα, είναι σε θέση να δημιουργήσει έναν τομέα εργασίας για την εκμάθηση μηχανών, θα καθορίσει τους παράγοντες που συμβάλλουν στην επιτυχή εφαρμογή του στις γεωργικές βιομηχανίες και θα αναπτύξει νέες μεθόδους μηχανικής μάθησης και τρόπους αξιολόγησης της αποτελεσματικότητάς τους. Το μεγαλύτερο μέρος της υλοποίησης πραγματοποιήθηκε στο C, με μερικές αξιολογήσεις που γράφτηκαν στο Prolog, και η διεπαφή χρήστη που παράχθηκε χρησιμοποιώντας το TCL / TK. Κατά τη διάρκεια αυτής της περιόδου, το ακρωνύμιο WEKA σχεδιάστηκε και δημιουργήθηκε η μορφή αρχείου συσχετίσεων αρχείων (ARFF) από το σύστημα (Altintas, et al., 2004).

Η πρώτη απελευθέρωση του WEKA ήταν εσωτερική και συνέβη το 1994. Το λογισμικό ήταν σε βήτα στάδιο. Η πρώτη δημοσίευση (έκδοση 2.1) έγινε τον Οκτώβριο του 1996. Τον Ιούλιο του 1997 κυκλοφόρησε το WEKA 2.2. Περιλάμβανε οχτώ αλγορίθμους εκμάθησης (οι υλοποιήσεις των οποίων παρασχέθηκαν από τους αρχικούς συγγραφείς) που ενσωματώθηκαν σε wrappers που χρησιμοποιούσαν το WEKA με βάση την προεπεξεργασία των δεδομένων σε ετικέτες που γράφτηκαν στο C. Τα WEKA 2.2 επίσης διέθεταν μια εγκατάσταση βασισμένη σε Unix Make files , για τη διαμόρφωση και τη διεξαγωγή πειραμάτων μεγάλης κλίμακας βάσει αυτών των αλγορίθμων.

Μέχρι τώρα έγινε ολοένα και πιο δύσκολο να διατηρηθεί το λογισμικό. Παράγοντες όπως οι αλλαγές στις υποστηρικτικές βιβλιοθήκες, η διαχείριση των εξαρτήσεων και η πολυπλοκότητα της διαμόρφωσης καθιστούσαν τη δουλειά του δημιουργού και την εμπειρία εγκατάστασης απογοητευτική για τους χρήστες. Σχετικά με αυτό το διάστημα αποφασίστηκε να ξαναγραφεί το σύστημα εξ ολοκλήρου στην Java, συμπεριλαμβανομένων των υλοποιήσεων των αλγορίθμων μάθησης. Αυτό ήταν μια κάπως ριζοσπαστική απόφαση, δεδομένου ότι η Java ήταν δυο ετών. Επιπλέον, η απόδοση χρόνου εκτέλεσης της Java κατέστησε μια αμφισβητήσιμη επιλογή για την υλοποίηση υπολογιστικών εντατικών αλγορίθμων μηχανικής μάθησης. Εντούτοις, αποφασίστηκε ότι τα πλεονεκτήματα όπως "Write Once, Run Anywhere" και η απλή συσκευασία και διανομή αντισταθμίζουν αυτά τα κενά και θα διευκόλυναν την ευρύτερη αποδοχή του λογισμικού.

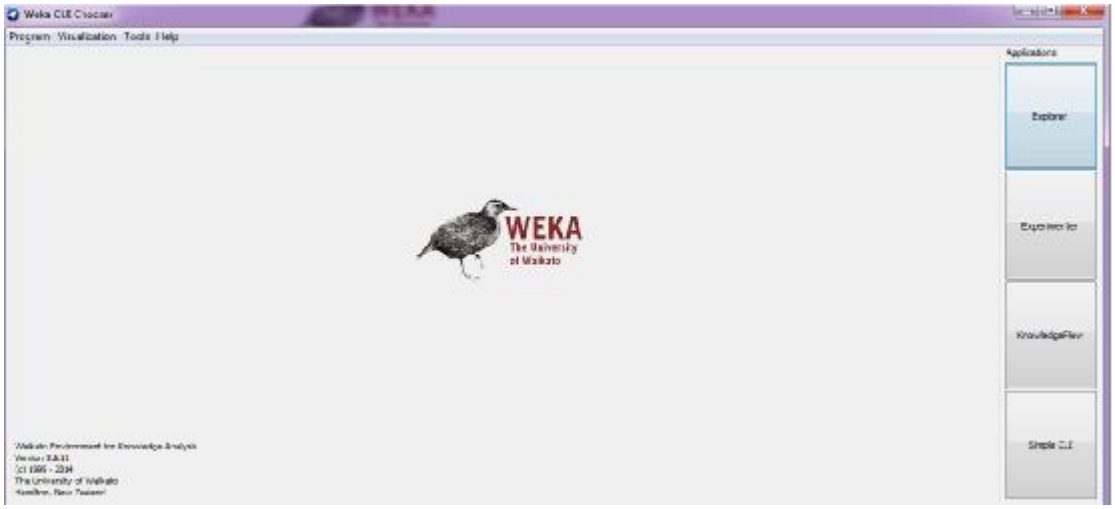
Το Μάιο του 1998 πραγματοποιήθηκε η τελική απελευθέρωση του συστήματος TCL / TK (WEKA 2.3) και στα μέσα του 1999 κυκλοφόρησε το 100% Java WEKA 3.0. Αυτή η μη γραφική έκδοση του WEKA συνοδεύει την πρώτη έκδοση του miningbook από τους Witten και Frank (Witten, & Frank, 2000). Τον Νοέμβριο του 2003, κυκλοφόρησε μια σταθερή έκδοση του WEKA (3.4) εν αναμονή της δημοσίευσης της δεύτερης έκδοσης του βιβλίου (Witten, & Frank, 2005). Στο χρονικό διάστημα μεταξύ 3,0 και 3,4, αναπτύχθηκαν οι τρεις κύριες γραφικές επεξηγήσεις χρήστη.

Το 2005, η ομάδα ανάπτυξης του WEKA έλαβε το βραβείο SIGKDDData Mining and Discovery Service. Το βραβείο αναγνώρισε τη μακροζωία και την ευρεία

υιοθέτηση του WEKA. Το 2006, η Pentaho Corporation έγινε ο κύριος χορηγός του λογισμικού και την υιοθέτησε για να διαμορφώσει το στοιχείο εξόρυξης δεδομένων και προβλεπτικό συστατικό της επιχειρηματικής ευφυΐας. Η Pentaho είναι πλέον ενεργός συνεισφέρων στη βάση κώδικα και ο πρώτος συγγραφέας είναι επί του παρόντος ο επικεφαλής του λογισμικού. Από αυτό το γράψιμο, το WEKA 3.6 (που κυκλοφόρησε τον Δεκέμβριο του 2008) είναι η τελευταία έκδοση του WEKA, το οποίο, δίνοντας την ομοιόμορφη αρίθμηση της έκφρασης cheme, θεωρείται ότι είναι χαρακτηριστικό σταθερό.

3.3 Επιλογές WEKA

Το WEKA μπορεί να διατεθεί για οποιοδήποτε λογισμικό Windows, Mac, ή Linux ενώ υπάρχει επίσης μια επιλογή η οποία εμπεριέχει το Java Virtual Machine. Το Java Virtual Machine επιτρέπει στον χρήση την γραφή ενός κώδικα τον οποίο και στη συνέχεια μπορεί να ενσωματώσει στο WEKA, προκειμένου να εξυπηρετήσει δικές του ανάγκες. Όταν το πρόγραμμα κατέβει και εγκατασταθεί είναι πιθανό να ζητηθεί να πραγματοποιηθεί αναβάθμιση λογισμικού προκειμένου να γίνει χρήση java. Παρακάτω ακολουθεί το περιβάλλον της αρχικής οθόνης του WEKA (Λαζάρου, & Χατζηδάκης, 2015).



Εικόνα 3.3.1: Αρχική οθόνη WEKA
Πηγή: Λαζάρου, & Χατζηδάκης, 2015

Στο αρχικό μενού του WEKA θα εμφανιστούν οι εξής επιλογές (Λαζάρου, & Χατζηδάκης, 2015):

Explorer: Το Explorer συνιστά το γραφικό εκείνο περιβάλλον που χρησιμοποιείται για την επεξεργασία δεδομένων τα οποία δεν έχουν επεξεργαστεί.

Experimenter: Το Experimenter αποτελεί ένα περιβάλλον όπου πραγματοποιούνται εργασίες που έχουν να κάνουν με τη στατιστική.

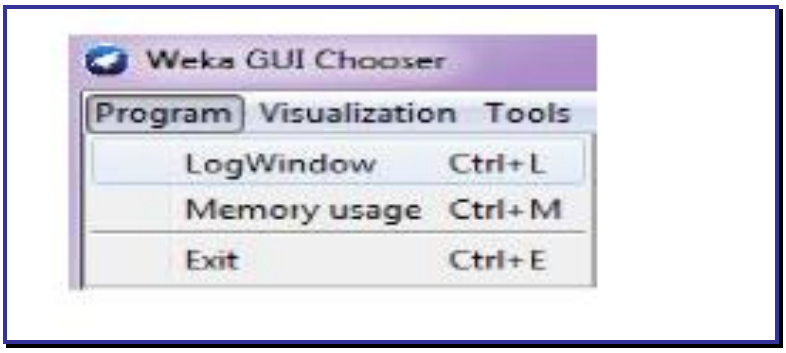
Knowledge Flow: Το Knowledge Flow έχει ίδια χρήση με αυτή του Explorer. Ωστόσο με το Knowledge Flow ο χρήστης μπορεί να πραγματοποιήσει drag & drop να μεταφέρει με το ποντίκι δηλαδή, δεδομένα από άλλα αρχεία στο αρχείο που χρησιμοποιεί. Επίσης το Knowledge Flow υποστηρίζει την incremental learning η οποία είναι μέθοδος εκμάθησης που συνδέει διάφορες τεχνικές. Με αυτόν τον τρόπο ο χρήστης αντλεί γνώσεις από παλαιότερες εργασίες.

Simple CLI: Με το Simple CLI ο χρήστης μπορεί να χρησιμοποιήσει μέσα από μια σειρά εντολών το γραφικό περιβάλλον. Το περιβάλλον της τελευταίας επιλογής έχει συγκεκριμένες εντολές που παρέχουν τα ανάλογα αποτελέσματα. Για παράδειγμα η εντολή :java weka.classifiers.trees.J48-t temp.arff θα ανοίξει το αρχείο μας temp.arff και θα εφαρμόσει τον αλγόριθμο J48 που έχει να κάνει με τα δέντρα αποφάσεων. Επίσης θα πρέπει ο χρήστης να γνωρίζει ότι και για την αποθήκευση σε αυτήν την επιλογή θα πρέπει να την πραγματοποιήσει πάλι με τον ίδιο τρόπο.

3.4 Περιγραφή Menu WEKA

Το Menu του WEKA διαμορφώνεται ως εξής: Αρχικά στην οθόνη υπάρχει η επιλογή Program όπου υπάρχουν οι επιλογές LogWindow, Memory usage και Exit. Επιλέγοντας το LogWindow ανοίγει ένα καινούριο παράθυρο. Το παράθυρο αυτό καταγράφει πρόσθετες διαδικασίες. Η επιλογή Memory usage, αναφέρεται στο

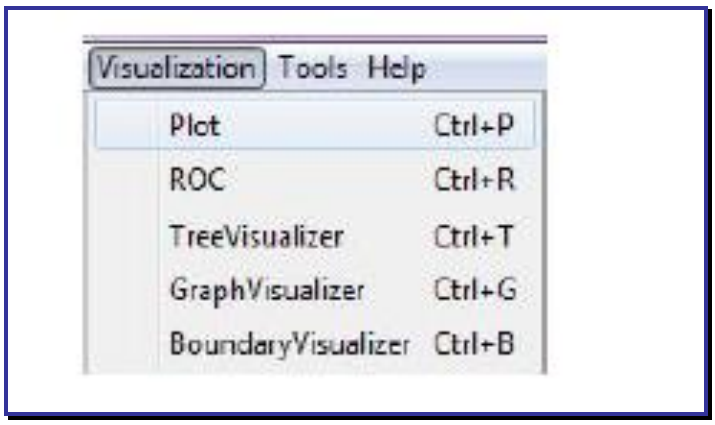
ποσοστό μνήμης που χρησιμοποιείται από το πρόγραμμα. Τέλος με την επιλογή Exit, ο χρήστης κλείνει το πρόγραμμα του.



Εικόνα 3.4.1: Menu WEKA

Πηγή: Λαζάρου, & Χατζηδάκης, 2015

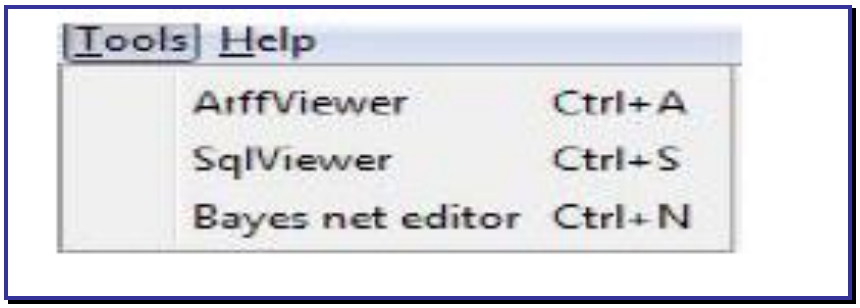
Δίπλα από την επιλογή Program υπάρχει η επιλογή Visualization. Η Visualization παραθέτει μια σειρά νέων επιλογών όπως: Plot, ROC, TreeVisualizer, GraphVisualizer, BoundaryVisualizer. Η επιλογή Plot χρησιμοποιεί το αρχείο του χρήστη για να σχεδιάσει δυσδιάστατα μοντέλα. Η ROC επιτρέπει την επιλογή την καμπύλης που έχει αποθηκευτεί πιο πρόσφατα. Η TreeVisualizer χρησιμοποιείται από τον χρήστη για να απεικονίσει διάφορα γραφήματα. Η GraphVisualizer βοηθάει στην απεικόνιση XML, DOT ή BIF που έχουν να κάνουν με γραφήματα Bayesian. Τέλος η BoundaryVisualizer δίνει τη δυνατότητα στον χρήστη να απεικονίσει την ταξινόμηση των δεδομένων στα όρια που ο ίδιος έχει θέσει.



Εικόνα 3.4.2: Menu WEKA

Πηγή: Λαζάρου, & Χατζηδάκης, 2015

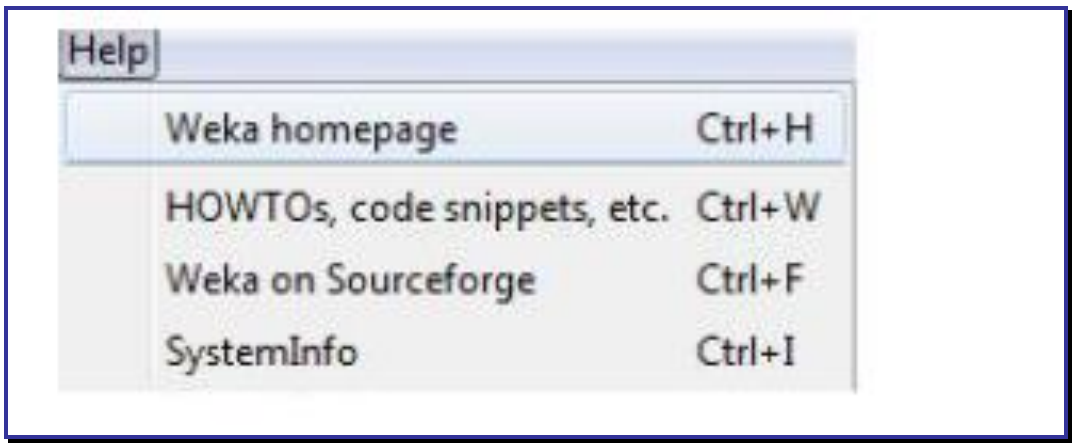
Στην επιλογή Tools υπάρχουν οι επιλογές: ArffViewer, SqlViewer και Bayes net editor. Η επιλογή ArffViewer δίνει τη δυνατότητα στον χρήστη να βλέπει τα αρχεία .arff με τη μορφή ενός υπολογιστικού φύλλου. Με τη SqlViewer ο χρήστης μπορεί να αναπαραστήσει ένα φύλλο εργασίας sql ώστε να το επεξεργαστεί μέσω JDBC (Συνδετικότητα Βάσης Δεδομένων JAVA). Τέλος με τη Bayes net editor ο χρήστης μπορεί να επεξεργάζεται, να απεικονίζει και να μαθαίνει μέσα από την χρήση δικτύων Bayes.



Εικόνα 3.4.3: Menu WEKA

Πηγή: Λαζάρου, & Χατζηδάκης, 2015

Στην επιλογή Help εμφανίζονται οι επιλογές: Weka homepage, HOWTOs, code snippets,etc, Weka on Sourceforge και SystemInfo. Η επιλογή Weka homepage μεταφέρει τον χρήστη στη σελίδα του προγράμματος. Η HOWTOs, code snippets,etc στην ουσία αποτελεί έναν οδηγό του προγράμματος το οποίο μέσα από την παράθεση παραδειγμάτων, βοηθάει τον χρήστη να εξοικειωθεί με τη λειτουργία του προγράμματος. Η επιλογή Weka on Sourceforge συνιστά την ιστοσελιδά του Weka στο Sourceforge. Τέλος η SystemInfo δίνει στο χρήστη χρήσιμες πληροφορίες για την λειτουργία του συστήματος.



Εικόνα 3.4.4: Menu WEKA

Πηγή: Λαζάρου, & Χατζηδάκης, 2015

3.5 Υποστηριζόμενα Αρχεία WEKA

Το WEKA υποστηρίζει αρχεία .arff. Ένα αρχείο arff (Μορφή αρχείου συσχετίσεων) είναι ένα αρχείο κειμένου ASCII που περιγράφει μια λίστα εμφανίσεων που μοιράζονται ένα σύνολο χαρακτηριστικών. Τα αρχεία arff αναπτύχθηκαν από το Μηχανικό Πρόγραμμα Μάθησης στο Τμήμα Πληροφορικής του Πανεπιστημίου Waikato για χρήση με το λογισμικό εκμάθησης μηχανών Weka.

Τα αρχεία ARFF έχουν δύο διαφορετικές ενότητες. Η πρώτη ενότητα είναι οι πληροφορίες κεφαλίδας , οι οποίες ακολουθούν τις πληροφορίες δεδομένων . Η επικεφαλίδα του αρχείου ARFF περιέχει το όνομα της σχέσης, μια λίστα με τα χαρακτηριστικά (τις στήλες στα δεδομένα) και τους τύπους τους. Μια κεφαλίδα παράδειγμα στο τυπικό σύνολο δεδομένων IRIS μοιάζει με αυτό:

```
% 1. Τίτλος: Βάση δεδομένων φυτών Iris
   %
   % 2. Πηγές:
   % (α) Δημιουργός: RA Fisher
   % (β) Δωρητής: Michael Marshall (MARSHALL_PLU@io.arc.nasa.gov)
   % (γ) Ημερομηνία: Ιούλιος, 1988
   %
   @RELATION ίριδα

   @ATTRIBUTE sepallength NUMERIC
   @ ATTRIBUTE sepalwidth NUMERIC
   @ ATTRIBUTE petallength NUMERIC
   @ ATTRIBUTE πενταπλάσια NUMERIC
```

```
@ ATTRIBUTE τάξη {Iris-setosa, Iris-versicolor, Iris-virginica}
```

Τα δεδομένα του αρχείου ARFF μοιάζουν με τα ακόλουθα:

```
@ΔΕΔΟΜΕΝΑ
5.1,3.5,1.4,0.2, Iris-setosa
4,9,3,0,1,4,0,2, Iris-setosa
4.7,3.2,1.3,0.2, Iris-setosa
4.6,3,1,1,5,0,2, Iris-setosa
5.0,3.6,1.4,0.2, Iris-setosa
5,4,3,9,1,7,0,4, Iris-setosa
4.6,3.4,1.4,0.3, Iris-setosa
5,0,3,4,1,5,0,2, Iris-setosa
4.4,2.9,1.4,0.2, Iris-setosa
4,9,3,1,1,5,0,1, Iris-setosa
```

Οι γραμμές που αρχίζουν με% είναι σχόλια. Οι δηλώσεις @RELATION , @ATTRIBUTE και @DATA δεν είναι ευαίσθητες στην περίπτωση.

Παραδείγματα

Αρκετά γνωστά σύνολα δεδομένων μάθησης μηχανής διανέμονται με τον Weka στον κατάλογο $ WEKAHOME / δεδομένων ως αρχεία ARFF.

Το τμήμα κεφαλίδας ARFF

Το τμήμα Κεφαλίδας ARFF του αρχείου περιέχει τις δηλώσεις σχέσης και τις δηλώσεις χαρακτηριστικών.

Η Δήλωση Συναλλαγών

Το όνομα σχέσης ορίζεται ως η πρώτη γραμμή στο αρχείο ARFF. Η μορφή είναι:

```
@ σχέση <όνομα_αναφοράς>
```

όπου <όνομα_αναφοράς> είναι μια συμβολοσειρά. Η συμβολοσειρά πρέπει να αναφέρεται όταν το όνομα περιλαμβάνει κενά.

Οι δηλώσεις @attribute

Οι δηλώσεις χαρακτηριστικών έχουν τη μορφή μιας σειράς **παραγγελιών** @attribute. Κάθε ιδιότητα στο σύνολο δεδομένων έχει τη δική του δήλωση @attribute που ορίζει με μοναδικό τρόπο το όνομα αυτού του χαρακτηριστικού και τον τύπο δεδομένων του. Η σειρά που δηλώνονται τα χαρακτηριστικά υποδεικνύει τη θέση της στήλης στην ενότητα δεδομένων του αρχείου. Για παράδειγμα, αν ένα χαρακτηριστικό είναι το τρίτο που δηλώνεται τότε ο Weka αναμένει ότι όλες οι τιμές των χαρακτηριστικών θα βρεθούν στην τρίτη στήλη που οριοθετείται με κόμμα.

Η μορφή της δήλωσης @attribute είναι:

```
@attribute <attribute-name> <datatype>
```

όπου το <*attribute-name*> πρέπει να ξεκινά με έναν αλφαβητικό χαρακτήρα. Εάν πρέπει να συμπεριληφθούν τα κενά στο όνομα, πρέπει να αναφερθεί ολόκληρο το όνομα.

Το <*datatype*> μπορεί να είναι οποιοσδήποτε από τους τέσσερις τύπους που υποστηρίζει ο Weka:

- αριθμητικός
- <nominal-specification>
- σειρά
- ημερομηνία [<ημερομηνία-μορφή>]

όπου οι <ονομαστικές προδιαγραφές> και <ημερομηνία-μορφή> καθορίζονται παρακάτω. Οι αριθμοί λέξεων-κλειδιών , η συμβολοσειρά και η ημερομηνία δεν είναι ευαίσθητες στην περίπτωση.

Αριθμητικά χαρακτηριστικά

Τα αριθμητικά χαρακτηριστικά μπορούν να είναι πραγματικοί ή ακέραιοι αριθμοί.

Ονομαστικά χαρακτηριστικά

Οι ονομαστικές τιμές ορίζονται με την παράθεση <ονομαστικής προδιαγραφής> των πιθανών τιμών: {<ονομαστικής ονομασίας1>, <ονομαστικής ονομασίας2>, <ονομαστικής ονομασίας3>, ...}

Για παράδειγμα, η τιμή κλάσης του συνόλου δεδομένων Iris μπορεί να οριστεί ως εξής:

```
@ ATTRIBUTE τάξη {Iris-setosa, Iris-versicolor, Iris-virginica}
```

Πρέπει να αναφέρονται τιμές που περιέχουν κενά.

Χαρακτηριστικά στοιχειοσειράς

Τα χαρακτηριστικά στοιχειοσειράς μας επιτρέπουν να δημιουργούμε χαρακτηριστικά που περιέχουν αυθαίρετες τιμές κειμένου. Αυτό είναι πολύ χρήσιμο σε εφαρμογές εξόρυξης κειμένου, καθώς μπορούμε να δημιουργούμε σύνολα δεδομένων με χαρακτηριστικά στοιχειοσειράς και στη συνέχεια να γράφουμε φίλτρα Weka για να χειριστούμε συμβολοσειρές (όπως το StringToWordVectorFilter). Τα χαρακτηριστικά των συμβολοσειρών δηλώνονται ως εξής:

```
@ ATTRIBUTE σειρά LCC
```

Χαρακτηριστικά ημερομηνίας

Οι δηλώσεις χαρακτηριστικών ημερομηνίας έχουν τη μορφή:

```
@attribute <όνομα> ημερομηνία [<ημερομηνία-μορφή>]
```

όπου <name> είναι το όνομα για το χαρακτηριστικό και <date-format> είναι μια προαιρετική συμβολοσειρά που καθορίζει πώς πρέπει να αναλύονται και να εκτυπώνονται οι τιμές ημερομηνίας (αυτό είναι το ίδιο σχήμα που χρησιμοποιείται από το SimpleDateFormat). Η προεπιλεγμένη συμβολοσειρά μορφής δέχεται τη μορφή συνδυασμένης ημερομηνίας και ώρας ISO-8601: "yyyy-MM-dd'T'HH: mm: ss".

Οι ημερομηνίες πρέπει να καθορίζονται στην ενότητα δεδομένων ως τις αντίστοιχες αναπαραστάσεις συμβολοσειρών της ημερομηνίας / ώρας

Αρχείο δεδομένων ARFF

Η ενότητα Δεδομένα ARFF του αρχείου περιέχει τη γραμμή δήλωσης δεδομένων και τις πραγματικές γραμμές παρουσίας.

Η Δήλωση @data

Η δήλωση @data είναι μία γραμμή που δηλώνει την αρχή του τμήματος δεδομένων στο αρχείο. Η μορφή είναι:

```
@δεδομένα
```

Τα δεδομένα παρουσίας

Κάθε παράσταση αντιπροσωπεύεται σε μία γραμμή, με την επιστροφή του οχήματος να υποδηλώνει το τέλος του στιγμιότυπου.

Οι τιμές των χαρακτηριστικών για κάθε περίπτωση οριοθετούνται με κόμματα. Πρέπει να εμφανίζονται με τη σειρά που δηλώθηκαν στο τμήμα κεφαλίδας (δηλ. Τα δεδομένα που αντιστοιχούν στη nth **@** declaration είναι πάντα το nth πεδίο του χαρακτηριστικού).

Οι τιμές που λείπουν αντιπροσωπεύονται από ένα μόνο ερωτηματικό, όπως:

```
@δεδομένα
4.4, α, 1.5, δ, Iris-setosa
```

Οι τιμές των συμβολοσειρών και των ονομαστικών ιδιοτήτων είναι ευαίσθητες σε πεζά και πρέπει να αναγράφονται όσες περιέχουν χώρο, ως εξής:

```
@ LCCvsLCSH

@attribute LCC string
@attribute LCSH συμβολοσειρά
```

```
@δεδομένα
AG5, «Εγκυκλοπαίδειες και λεξικά, 20ος αιώνας».
AS262, «Επιστήμη - Σοβιετική Ένωση - Ιστορία».
AE5, "Εγκυκλοπαίδειες και λεξικά".
AS281, «Αστρονομία, Ασύρο-Βαβυλωνιανή, Φεγγάρι - Φάσεις».
AS281, «Αστρονομία, Ασσύρο-Βαβυλωνιανή, Σελήνη - Πίνακες».
```

Οι ημερομηνίες πρέπει να καθορίζονται στην ενότητα δεδομένων χρησιμοποιώντας τη συμβολοσειρά που ορίζεται στη δήλωση χαρακτηριστικών. Για παράδειγμα:

```
@RELATION Timestamps

@ ATTRIBUTE χρονική σφραγίδα DATE "εεεε-ΜΜ-ηη ΗΗ: mm: ss"

@ΔΕΔΟΜΕΝΑ
"2001-04-03 12:12:12"
"2001-05-03 12:59:55"
```

Αρχεία αραιών αρχείων ARFF

Τα αρχεία αραιών αρχείων ARFF είναι πολύ παρόμοια με τα αρχεία ARFF, αλλά τα δεδομένα με την τιμή 0 δεν εκπροσωπούνται ρητά.

Τα αρχεία αραιών αρχείων ARFF έχουν την ίδια κεφαλίδα (δηλαδή ετικέτες @relation και @attribute), αλλά η ενότητα δεδομένων είναι διαφορετική. Αντί να αναπαριστάμε κάθε τιμή με τη σειρά, όπως αυτό:

```
@δεδομένα
0, X, 0, Y, "κλάση Α"
0, 0, W, 0, "κλάση Β"
```

Τα μη μηδενικά χαρακτηριστικά προσδιορίζονται ρητά από τον αριθμό του χαρακτηριστικού και την τιμή που δηλώνεται, όπως παρακάτω:

```
@δεδομένα
{1 X, 3 Y, 4 "κλάση Α"}
{2 W, 4 "κατηγορίας Β"}
```

Κάθε παράσταση περιβάλλεται από αγκύλες και το σχήμα για κάθε καταχώρηση είναι: <index> <space> <value> όπου index είναι ο δείκτης χαρακτηριστικών (ξεκινώντας από 0).

Σημειώστε ότι οι παραλειπόμενες τιμές σε μια αραιή παρουσία είναι **0** , δεν είναι "λείπουν" τιμές! Εάν μια τιμή είναι άγνωστη, πρέπει να την εκπροσωπείτε ρητά με ένα ερωτηματικό (?).

Προειδοποίηση: Υπάρχει ένα γνωστό πρόβλημα που αποθηκεύει τα αντικείμενα SparseInstance από σύνολα δεδομένων που έχουν χαρακτηριστικά στοιχειοσειράς. Στο Weka, οι τιμές των συμβολοσειρών και των ονομαστικών δεδομένων αποθηκεύονται ως αριθμοί. αυτοί οι αριθμοί λειτουργούν ως ευρετήρια σε μια σειρά από πιθανές τιμές χαρακτηριστικών (αυτό είναι πολύ αποτελεσματικό). Ωστόσο, η πρώτη τιμή συμβολοσειράς έχει εκχωρηθεί δείκτης 0: αυτό σημαίνει ότι, εσωτερικά, αυτή η τιμή αποθηκεύεται ως 0. Όταν γράφεται ένα SparseInstance, οι συμβολοσειρές με εσωτερική τιμή 0 δεν εξάγονται, έτσι η τιμή συμβολοσειράς τους έχει χαθεί (και όταν το αρχείο arff διαβάζεται ξανά, η προεπιλεγμένη τιμή 0 είναι το ευρετήριο μιας διαφορετικής τιμής συμβολοσειράς, οπότε η τιμή του χαρακτηριστικού φαίνεται να αλλάζει).

3.6 Εντολές σε περιβάλλον WEKA

Στο πρόγραμμα WEKA υπάρχει μια σειρά εντολών που βοηθούν τον χρήστη να κατανοήσει τη λειτουργία του προγράμματος και να εκτελέσει τις ενέργειες που επιθυμεί με τρόπο εύκολο και γρήγορο (Bouckaert, et al., 2008). Οι ακόλουθες εντολές είναι διαθέσιμες στο απλό CLI:

- java <classname> [<args>]

Επικαλείται μια κλάση java με τα συγκεκριμένα επιχειρήματα (αν υπάρχουν)

- break

Σταματά την τρέχουσα ροή , π.χ. έναν τρέχον ταξινομητή, με φιλικό τρόπο

- kill

Σταματά την τρέχουσα ροή κατά τρόπο μη εχθρικό

- cls

Καθαρίζει την περιοχή εξόδου

- exit

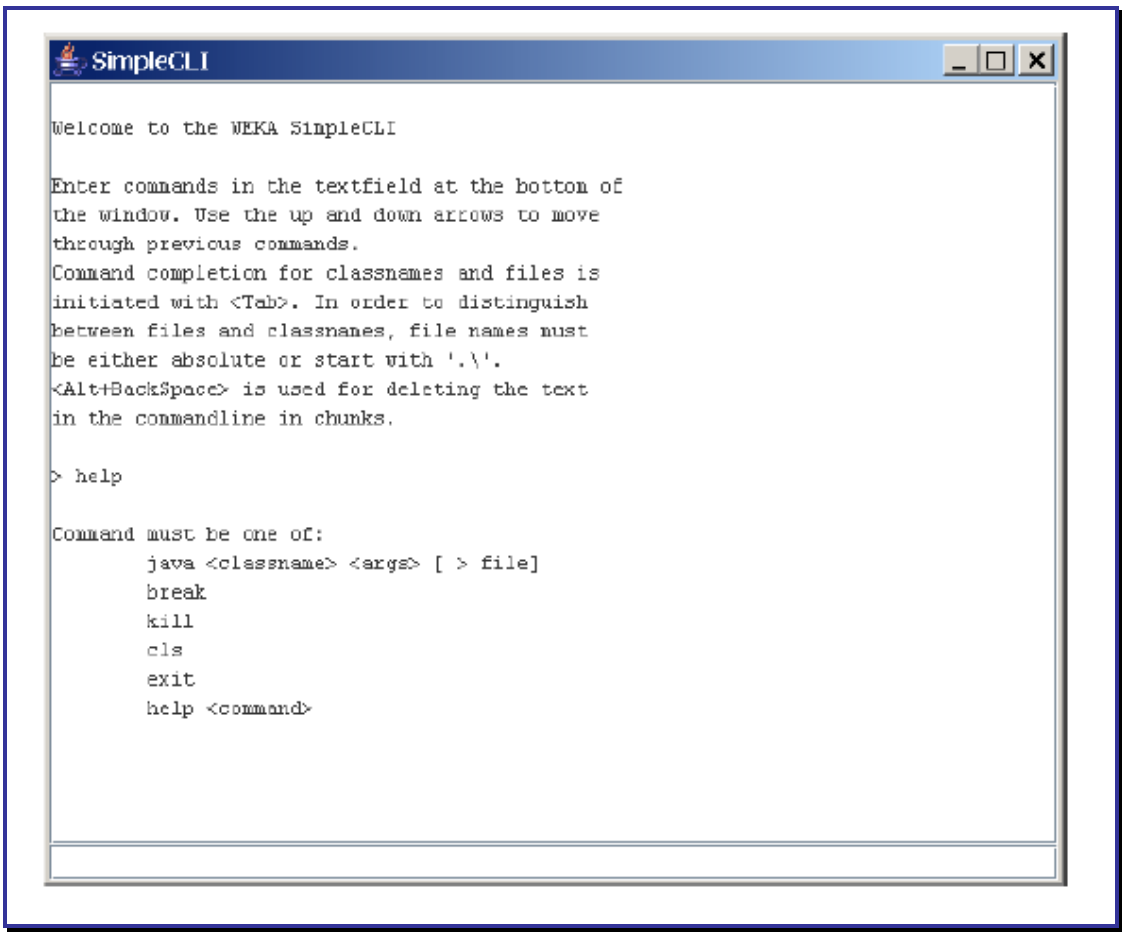
Εξέρχεται από το απλό CLI

- help [<command>]

Παρέχει μια επισκόπηση των διαθέσιμων εντολών αν δεν υπάρχει καθορισμένη εντολή, διαφορετικά περισσότερη βοήθεια για την καθορισμένη εντολή

Απλό CLI

Το απλό CLI παρέχει πλήρη πρόσβαση σε όλες τις κατηγορίες του Weka, δηλ. Ταξινομητές, φίλτρα, ομάδες, κλπ., Αλλά χωρίς την ταλαιπωρία του CLASSPATH (διευκολύνει εκείνο με το οποίο ξεκίνησε η Weka). Προσφέρει ένα απλό κέλυφος Weka με ξεχωριστή γραμμή εντολών και έξοδο (Bouckaert, et al., 2008)



Εικόνα 3.6.1: Εντολές WEKA

Πηγή: (Bouckaert, et al., 2008)

Κεφάλαιο 4ο- Λογισμικό WEKA

4.1 Επεκτάσεις στο λογισμικό WEKA

Το Programming Interface (API) της WEKA επιτρέπει την εύκολη ενσωμάτωση σε άλλες εφαρμογές Java καθώς και την επέκτασή της με νέα χαρακτηριστικά που μπορεί να είναι είτε πρόσθετοι αλγόριθμοι μηχανικής μάθησης και εργαλεία για την οπτικοποίηση δεδομένων, είτε ακόμα και επεκτάσεις του γραφικού περιβάλλοντος χρήστη (GUI) για να υποστηριχθούν διαφορετικές ροές εργασίας, όπως για παράδειγμα το περιβάλλον χρόνου ανάλυσης και πρόβλεψης (Hall, et al., 2009).

Εκτός από τον πηγαίο κώδικα WEKA, για την υλοποίηση οποιασδήποτε επέκτασης, μπορεί κανείς να χρησιμοποιήσει διάφορα εξωτερικά εργαλεία προγραμματισμού και βιβλιοθήκες τρίτων που περιγράφονται παρακάτω. Ο σκοπός αυτών των εργαλείων είναι η αυτοματοποίηση πολλών κοινών εργασιών προγραμματισμού, όπως η δοκιμή μονάδων ή η διαχείριση του πηγαίου κώδικα.

Για τη δοκιμή μονάδων, το WEKA χρησιμοποιεί τη βιβλιοθήκη JUnit . Το JUnit είναι ένα πλαίσιο που χρησιμοποιείται για τη δημιουργία αυτοματοποιημένων περιπτώσεων δοκιμών και διανέμεται δωρεάν σύμφωνα με την Άδεια Κοινής Δημόσιας Άδειας Χρήσης v 1.0.

Ο πηγαίος κώδικας του WEKA είναι επίσης διαθέσιμος μέσω αποθετηρίου λογισμικού που βασίζεται στο Apache Subversion . Η εφαρμογή Subversion του Apache επιτρέπει στους προγραμματιστές να ελέγχουν τις τροποποιήσεις στον πηγαίο κώδικα σε διάφορα στάδια του κύκλου ανάπτυξης λογισμικού, οδηγώντας σε μια πιο αποτελεσματική συνεργασία μεταξύ αυτών που εμπλέκονται στην ανάπτυξη και συνεπώς στην αύξηση της παραγωγικότητας, ειδικά στην περίπτωση ανάπτυξης μιας μεγάλης εφαρμογής ανοιχτού κώδικα, η οποία περιλαμβάνει μεγάλο αριθμό γεωγραφικά κατανεμημένων προγραμματιστών. Το Apache Subversion διανέμεται δωρεάν υπό την Κοινή Δημόσια Άδεια v 1.0.

Η διαδικασία δημιουργίας ενός νέου πακέτου επέκτασης WEKA απαιτεί το εργαλείο Apache Ant build, το οποίο επίσης διανέμεται δωρεάν υπό την έκδοση Apache License 2.0 (Salatas, 2011).

Τα εργαλεία που περιγράφονται παραπάνω είναι συνήθως ενσωματωμένα σε άλλα εργαλεία προγραμματισμού (δηλαδή Κώδικα επεξεργαστές / debuggers) σε ένα ενιαίο IDE. Τα δύο πιο δημοφιλή IDE για ανάπτυξη Java είναι πιθανώς τα εξής:

- Το Netbeans, το οποίο διανέμεται δωρεάν υπό διπλή άδεια: Common Development and DistributionLicense (CDDL) v1.0 και GNU GPL v2.
- Η Eclipse, η οποία επίσης διανέμεται δωρεάν υπό την Eclipse Public License (EPL) v. 1.0.

Και οι δύο αυτές IDE μπορούν να ρυθμιστούν για την ανάπτυξη των επεκτάσεων του WEKA (Salatas, 2011).

4.2 Τα δίκτυα RBF και MLP στο WEKA

Ένα δίκτυο (MLP) είναι μια κατηγορία τροφοδοτούμενου τεχνητού νευρικού δικτύου. Ένα MLP αποτελείται από τουλάχιστον τρία στρώματα κόμβων: ένα στρώμα εισόδου, ένα κρυμμένο στρώμα και ένα στρώμα εξόδου. Εκτός από τους κόμβους εισόδου, κάθε κόμβος είναι ένας νευρώνας που χρησιμοποιεί μια μη γραμμική λειτουργία ενεργοποίησης. Το MLP χρησιμοποιεί μια εποπτευόμενη τεχνική εκμάθησης που ονομάζεται backpropagation για εκπαίδευση (Rosenblatt, 1961) Τα πολλαπλά στρώματα και η μη γραμμική ενεργοποίησή του διακρίνουν το MLP από ένα γραμμικό perceptron . Μπορεί να διακρίνει δεδομένα που δεν είναι γραμμικά διαχωρίσιμα (Rumelhart, Hinton, & Williams, 1986).

Εάν ένα perceptron πολλαπλών στρώσεων έχει μια γραμμική συνάρτηση ενεργοποίησης σε όλους τους νευρώνες, δηλαδή μια γραμμική συνάρτηση που χαρτογραφεί τις σταθμισμένες εισροές στην έξοδο κάθε νευρώνα, τότε η γραμμική άλγεβρα δείχνει ότι οποιοσδήποτε αριθμός στρωμάτων μπορεί να μειωθεί σε ένα

στρώμα εισόδου- εξόδου. Σε MLPs ορισμένοι νευρώνες χρησιμοποιούν μια μη γραμμική λειτουργία ενεργοποίησης που αναπτύχθηκε για να μοντελοποιήσει τη συχνότητα των δυνατοτήτων δράσης ή την πυροδότηση των βιολογικών νευρώνων. Τα MLP είναι χρήσιμα στην έρευνα για την ικανότητά τους να επιλύουν προβλήματα στοχαστικά, τα οποία συχνά επιτρέπουν προσεγγιστικές λύσεις για εξαιρετικά περίπλοκα προβλήματα όπως η προσέγγιση της φυσικής κατάστασης.

Τα MLPs είναι καθολικές προσεγγίσεις λειτουργιών όπως έδειξε το θεώρημα του Cybenko, έτσι ώστε να μπορούν να χρησιμοποιηθούν για τη δημιουργία μαθηματικών μοντέλων με ανάλυση παλινδρόμησης (Rumelhart, Hinton, & Williams, 1986). Καθώς η ταξινόμηση είναι μια ιδιαίτερη περίπτωση παλινδρόμησης όταν η μεταβλητή απόκρισης είναι κατηγορηματική , τα MLPs ταξινομούν αλγορίθμους.

MLPs ήταν μια δημοφιλής λύση μηχανικής μάθησης στη δεκαετία του 1980, βρίσκοντας εφαρμογές σε διάφορους τομείς, όπως η αναγνώριση φωνής , η αναγνώριση εικόνας και αυτόματη μετάφραση λογισμικού, αλλά στη συνέχεια αντιμετωπίζει έντονο ανταγωνισμό από πολύ πιο απλή (και των σχετικών) Μηχανές διανυσμάτων υποστήριξης. Το ενδιαφέρον για δίκτυα backpropagation επέστρεψε λόγω των επιτυχιών της βαθιάς μάθησης.

Στον τομέα της μαθηματικής μοντελοποίησης, ένα δίκτυο λειτουργιών ακτινικής βάσης είναι ένα τεχνητό νευρωνικό δίκτυο που χρησιμοποιεί λειτουργίες ακτινικής βάσης σαν λειτουργίες ενεργοποίησης. Η έξοδος του δικτύου είναι ένας γραμμικός συνδυασμός λειτουργιών ακτινικής βάσης των εισόδων και παραμέτρων νευρώνων. Τα δίκτυα λειτουργίας ακτινικής βάσης έχουν πολλές χρήσεις, συμπεριλαμβανομένης της προσέγγισης των λειτουργιών, της πρόβλεψης χρονοσειρών, της ταξινόμησης και του ελέγχου του συστήματος . Διατυπώθηκαν για πρώτη φορά σε ένα έγγραφο του 1988 από τους Broomhead και Lowe, και οι δύο ερευνητές στο Royal Signals και το Radar Establishment (Broomhead, & Lowe, 1988).

Τα δίκτυα βάσης ακτινωτής βάσης (RBF) έχουν συνήθως τρία επίπεδα: ένα στρώμα εισόδου, ένα κρυμμένο στρώμα με μη γραμμική λειτουργία ενεργοποίησης RBF και γραμμικό επίπεδο στρώματος εξόδου.

4.3 Διεπαφές χρήστη

Το πρόγραμμα WEKA στοχεύει στην παροχή ολοκληρωμένης συλλογής αλγορίθμων μηχανικής μάθησης και εργαλείων προεπεξεργασίας δεδομένων σε ερευνητές και επαγγελματίες. Επιτρέπει στους χρήστες να δοκιμάσουν γρήγορα και να συγκρίνουν διαφορετικές μεθόδους εκμάθησης μηχανών σε νέα σύνολα δεδομένων. Η αρθρωτή, επεκτάσιμη αρχιτεκτονική της επιτρέπει να δημιουργηθούν εξελιγμένες διεργασίες εξόρυξης δεδομένων από την ευρεία συλλογή αλγορίθμων βασικής μάθησης και εργαλείων που παρέχονται. Η επέκταση του Toolkit είναι εύκολη χάρη σε ένα απλό API, μηχανισμούς plugin και εγκαταστάσεις που αυτοματοποιούν την ενσωμάτωση νέων αλγορίθμων μάθησης με τις γραφικές διεπαφές χρήστη του WEKA. Ο πίνακας εργασίας περιλαμβάνει αλγόριθμους για παλινδρόμηση, ταξινόμηση, ομαδοποίηση, εξόρυξη κανόνα σύνδεσης και επιλογή χαρακτηριστικών. Η προκαταρκτική εξερεύνηση δεδομένων καλύπτεται καλά από τις εγκαταστάσεις οπτικοποίησης δεδομένων και από πολλά εργαλεία προεπεξεργασίας. Αυτά, όταν συνδυάζονται με τη στατιστική αξιολόγηση των μαθησιακών σχημάτων και την οπτικοποίηση των αποτελεσμάτων της μάθησης, υποστηρίζουν μοντέλα διεργασιών εξόρυξης δεδομένων όπως το CRISP-DM (Shearer, 2000).

Το WEKA διαθέτει πολλές διασυνδέσεις γραφικών που επιτρέπουν την εύκολη πρόσβαση στις βασικές λειτουργίες. Η κύρια γραφική διεπαφή χρήστη είναι ο "Explorer". Διαθέτει μια διεπαφή με βάση τα πάνελ, όπου διαφορετικές ομάδες αντιστοιχούν σε διαφορετικές εργασίες εξόρυξης δεδομένων. Στην πρώτη ομάδα, που ονομάζεται "Προεπεξεργασία", μπορεί να φορτωθούν και να μετατραπούν δεδομένα χρησιμοποιώντας τα εργαλεία προεπεξεργασίας δεδομένων WEKA, που ονομάζονται "φίλτρα". Τα δεδομένα μπορούν να φορτωθούν από διάφορες πηγές, συμπεριλαμβανομένων των αρχείων, των διευθύνσεων URL και των βάσεων δεδομένων. Οι υποστηριζόμενες μορφές αρχείων περιλαμβάνουν τη μορφή ARFF της WEKA, μορφή CSV, μορφή LibSVM και μορφή C4.5. Είναι επίσης δυνατό να δημιουργηθούν δεδομένα χρησιμοποιώντας μια τεχνητή πηγή δεδομένων και να

επεξεργαστούν δεδομένα με μη αυτόματο τρόπο χρησιμοποιώντας ένα πρόγραμμα επεξεργασίας δεδομένων (Khoussainov, Zuo, & Kushmerick, 2004).

Το δεύτερο πάνελ στον Explorer παρέχει πρόσβαση στους αλγόριθμους ταξινόμησης και παλινδρόμησης του WEKA. Ο αντίστοιχος πίνακας ονομάζεται "Ταξινόμηση", επειδή οι τεχνικές παλινδρόμησης θεωρούνται προβλέπτες των "συνεχών τάξεων". Από προεπιλογή, ο πίνακας εκτελεί μια διασταυρούμενη επικύρωση για έναν επιλεγμένο αλγόριθμο εκμάθησης στο σύνολο δεδομένων που έχει καταρτιστεί στον πίνακα Preprocess για την εκτίμηση της πρόβλεψης απόδοσης. Παρουσιάζει επίσης μια γραπτή αναπαράσταση του μοντέλου που κατασκευάστηκε από το πλήρες σύνολο δεδομένων. Ωστόσο, άλλοι τρόποι αξιολόγησης, π.χ. με βάση ένα ξεχωριστό σύνολο δοκιμών, υποστηρίζονται επίσης. Εάν είναι εφαρμόσιμο, ο πίνακας παρέχει επίσης πρόσβαση σε γραφικές παραστάσεις μοντέλων, π.χ. δέντρα αποφάσεων (Khoussainov, Zuo, & Kushmerick, 2004). Επιπλέον, μπορεί να απεικονίσει σφάλματα πρόβλεψης σε διαγράμματα διασποράς και επίσης επιτρέπει την αξιολόγηση μέσω καμπυλών ROC και άλλων "καμπυλών κατωφλίου". Τα μοντέλα μπορούν επίσης να αποθηκευτούν και να φορτωθούν σε αυτόν τον πίνακα.

Μαζί με τους εποπτευόμενους αλγορίθμους, το WEKA υποστηρίζει επίσης την εφαρμογή αλγορίθμων χωρίς επίβλεψη, δηλαδή αλγορίθμων ομαδοποίησης και μεθόδων εξόρυξης κανόνα σύνδεσης. Αυτά είναι προσβάσιμα στον Explorer μέσω του τρίτου και τέταρτου πίνακα αντίστοιχα. Ο πίνακας "Συστάδες" επιτρέπει στους χρήστες να εκτελούν έναν αλγόριθμο ομαδοποίησης των δεδομένων που έχουν φορτωθεί στον πίνακα Preprocess. Παρέχει απλές στατιστικές για την αξιολόγηση της απόδοσης της ομαδοποίησης: απόδοση που βασίζεται σε πιθανότητες για τους αλγόριθμους στατιστικής ομαδοποίησης και σύγκριση με την "αληθινή" ένταξη του συμπλέγματος, αν αυτό καθορίζεται σε ένα από τα χαρακτηριστικά των δεδομένων. Εάν είναι εφικτό, είναι επίσης δυνατή η απεικόνιση της δομής ομαδοποίησης και τα μοντέλα μπορούν να αποθηκευτούν επίμονα εάν είναι απαραίτητο.

Η υποστήριξη του WEKA για τα καθήκοντα συγκέντρωσης δεν είναι τόσο εκτεταμένη όσο η υποστήριξή της στην ταξινόμηση και την παλινδρόμηση, αλλά έχει περισσότερες τεχνικές για ομαδοποίηση από ό, τι για την εξόρυξη κανόνα σύνδεσης, η οποία έως τώρα έχει παραμεληθεί κάπως. Παρ 'όλα αυτά, περιέχει μια εφαρμογή

του πιο γνωστού αλγορίθμου σε αυτόν τον τομέα, καθώς και μερικές άλλες. Αυτές οι μέθοδοι μπορούν να αποκτήσουν πρόσβαση μέσω του πίνακα "Συνεργάτες" στον Explorer (Witten, Frank, & Hall, 2011).

Ίσως ένα από τα πιο σημαντικά καθήκοντα στην πρακτική εξόρυξη δεδομένων είναι το καθήκον του προσδιορισμού ποια χαρακτηριστικά στα δεδομένα είναι τα πιο προγνωστικά. Για το σκοπό αυτό, ο Explorer του WEKA έχει μια ειδική ομάδα για την επιλογή χαρακτηριστικών, "Επιλογή χαρακτηριστικών", η οποία παρέχει πρόσβαση σε μια μεγάλη ποικιλία αλγορίθμων και κριτηρίων αξιολόγησης για τον εντοπισμό των σημαντικότερων χαρακτηριστικών σε ένα σύνολο δεδομένων. Λόγω του γεγονότος ότι είναι δυνατό να συνδυαστούν διαφορετικές μέθοδοι αναζήτησης με διαφορετικά κριτήρια αξιολόγησης, είναι δυνατό να διαμορφωθεί ένα ευρύ φάσμα πιθανών υποψήφιων τεχνικών. Η ανθεκτικότητα του επιλεγμένου συνόλου χαρακτηριστικών μπορεί να επικυρωθεί μέσω προσέγγισης που βασίζεται σε διασταυρούμενη επικύρωση.

Σημειώστε ότι ο πίνακας επιλογής χαρακτηριστικών έχει σχεδιαστεί κυρίως για διερευνητική ανάλυση δεδομένων. Ο "FilteredClassifier" του WEKA (προσβάσιμος από τον πίνακα "Ταξινόμηση") πρέπει να χρησιμοποιηθεί για την εφαρμογή τεχνικών επιλογής χαρακτηριστικών σε συνδυασμό με έναν υποκείμενο αλγόριθμο ταξινόμησης ή παλινδρόμησης, προκειμένου να αποφευχθεί η εισαγωγή αισιόδοξων προκαταλήψεων στις εκτιμήσεις απόδοσης που αποκτήθηκαν. Αυτή η προειδοποίηση ισχύει επίσης για μερικά από τα εργαλεία προεπεξεργασίας - πιο συγκεκριμένα, τα εποπτευόμενα εργαλεία που είναι διαθέσιμα από τον πίνακα Preprocess.

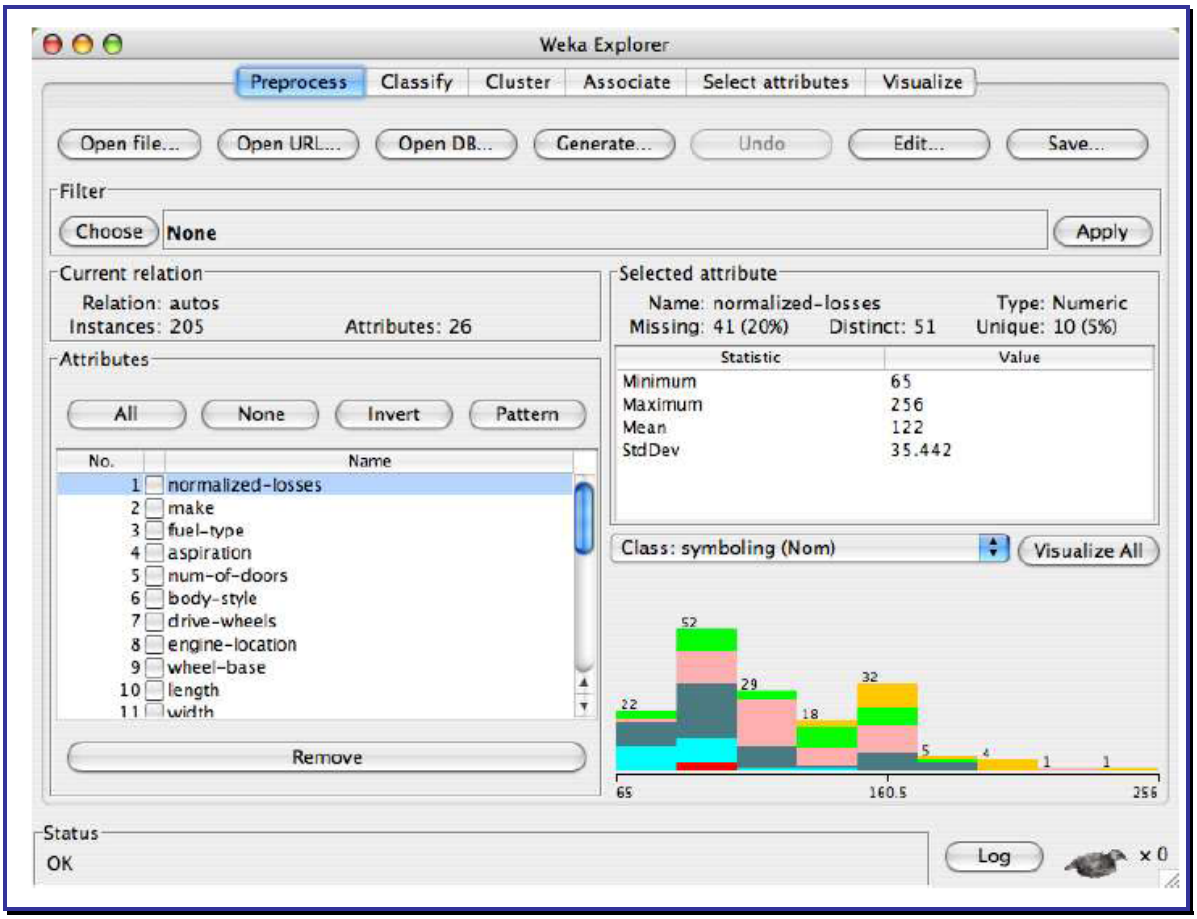
Σε πολλές πρακτικές εφαρμογές, η οπτικοποίηση δεδομένων παρέχει σημαντικές πληροφορίες. Αυτά μπορεί ακόμη να επιτρέψουν την αποφυγή περαιτέρω ανάλυσης χρησιμοποιώντας μηχανικές μάθησης και αλγόριθμους εξόρυξης δεδομένων (Witten, Frank, & Hall, 2011). Αλλά ακόμα κι αν αυτό δεν συμβαίνει, μπορούν να ενημερώσουν τη διαδικασία επιλογής ενός κατάλληλου αλγορίθμου για το πρόβλημα στο χέρι. Ο τελευταίος πίνακας στον Explorer, ο οποίος ονομάζεται "Οπτικοποίηση", παρέχει μια χρωματική γραφική παράσταση στο διάγραμμα σκέδασης, μαζί με την επιλογή της διάτρησης, επιλέγοντας μεμονωμένα οικόπεδα σε αυτόν τον πίνακα και επιλέγοντας τμήματα των δεδομένων για απεικόνιση. Είναι επίσης δυνατή η λήψη

πληροφοριών σχετικά με μεμονωμένα datapoints και η τυχαία διαταραχή των δεδομένων με επιλεγμένη ποσότητα για την αποκάλυψη συγκεχυμένων δεδομένων.

Ο Explorer έχει σχεδιαστεί για επεξεργασία δεδομένων βασισμένη σε παρτίδες: τα δεδομένα εκπαίδευσης φορτώνονται στη μνήμη στο σύνολό της και στη συνέχεια επεξεργάζονται. Αυτό μπορεί να μην είναι κατάλληλο για προβλήματα που αφορούν μεγάλα σύνολα δεδομένων. Ωστόσο, το WEKA έχει υλοποιήσεις κάποιων αλγορίθμων που επιτρέπουν τη δημιουργία στοιχειωδών μοντέλων, τα οποία μπορούν να εφαρμοστούν σε διαδοχική λειτουργία από μια διεπαφή γραμμής εντολών. Η αυξητική φύση αυτών των αλγορίθμων αγνοείται στον Explorer, αλλά μπορεί να εκμεταλλευτεί χρησιμοποιώντας μια πιο πρόσφατη προσθήκη στο σύνολο των γραφικών διεπαφών χρήστη του WEKA, δηλαδή τη λεγόμενη "Knowledge Flow/ροή γνώσης" που παρουσιάζεται στο σχήμα 4.3.2. Περισσότερες λειτουργίες που μπορούν να αντιμετωπιστούν με τον Explorer μπορεί επίσης να αντιμετωπιστεί από τη ροή γνώσης. Ωστόσο, εκτός από την κατάρτιση με βάση την παρτίδα, το μοντέλο ροής δεδομένων επιτρέπει κλιμακωτές ενημερώσεις με κόμβους επεξεργασίας που μπορούν να φορτώσουν και να προεπεξεργαστούν μεμονωμένες περιπτώσεις πριν τους τροφοδοτήσουν σε κατάλληλους αλγόριθμους αυξητικής μάθησης. Παρέχει επίσης κόμβους για απεικόνιση και αξιολόγηση. Μόλις ρυθμιστεί η ρύθμιση των διασυνδεδεμένων κόμβων επεξεργασίας, μπορεί να αποθηκευτεί για μετέπειτα επαναχρησιμοποίηση (Witten, Frank, & Hall, 2011).

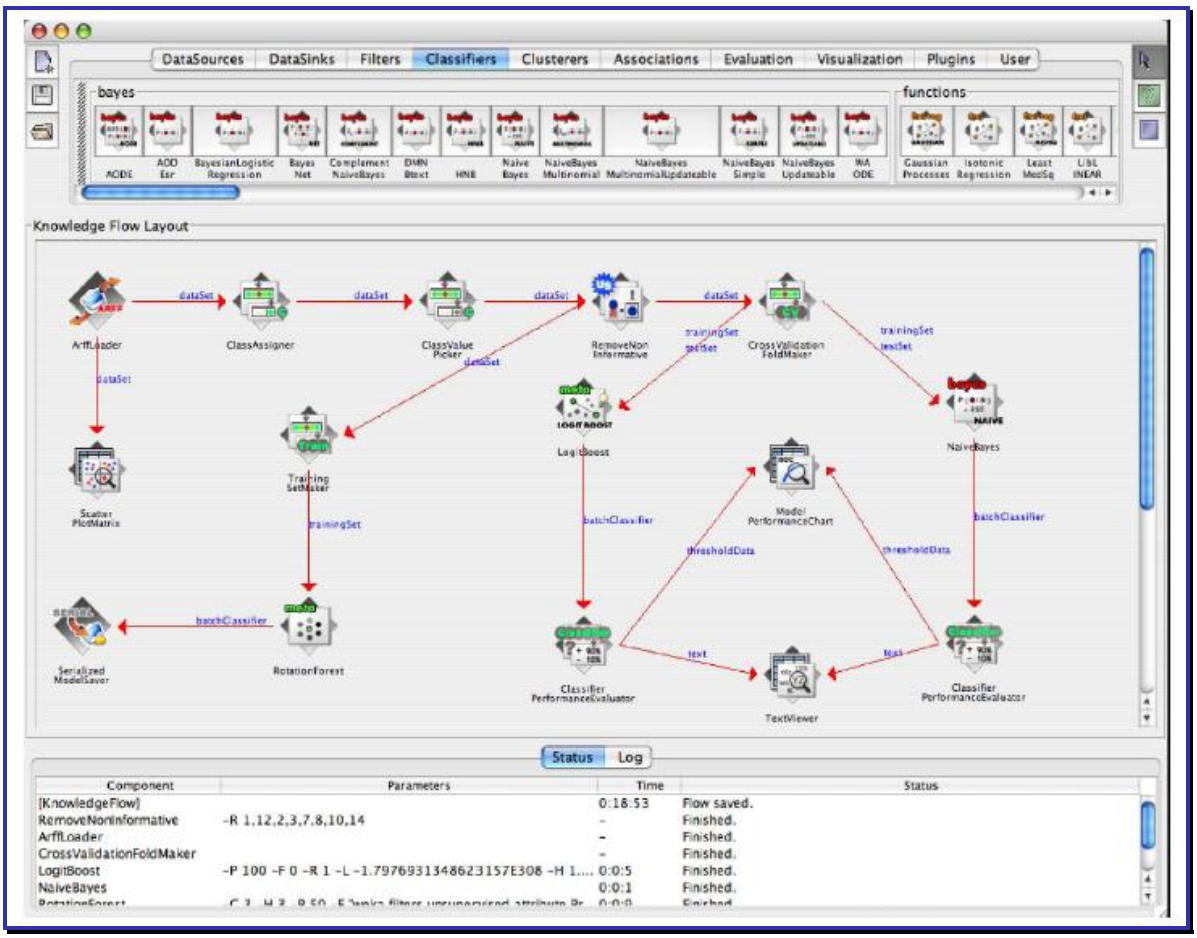
Το τρίτο κύριο γραφικό περιβάλλον χρήστη στο WEKA είναι το "Experimenter" . Αυτή η διεπαφή έχει σχεδιαστεί για να διευκολύνει την πειραματική σύγκριση της πρόβλεψης των αλγορίθμων με βάση τα πολλά διαφορετικά κριτήρια αξιολόγησης που είναι διαθέσιμα στο WEKA. Τα πειράματα μπορούν να περιλαμβάνουν πολλούς αλγόριθμους που εκτελούνται σε πολλαπλά σύνολα δεδομένων. για παράδειγμα, χρησιμοποιώντας επαναλαμβανόμενη διασταυρωμένη επικύρωση. Τα πειράματα μπορούν επίσης να διανεμηθούν σε διάφορους υπολογιστικούς κόμβους σε ένα δίκτυο για τη μείωση του υπολογιστικού φορτίου για μεμονωμένους κόμβους. Μόλις δημιουργηθεί ένα πείραμα, μπορεί να αποθηκευτεί είτε σε μορφή XML είτε σε δυαδική μορφή, ώστε να μπορεί να ξαναεπισκεφθεί εάν είναι απαραίτητο. Τα παραμετροποιημένα και αποθηκευμένα πειράματα μπορούν επίσης να εκτελεστούν από τη γραμμή εντολών. Σε σύγκριση με τις άλλες διεπαφές χρήστη του WEKA, το

Experimenter ίσως χρησιμοποιείται λιγότερο συχνά από επαγγελματίες εξόρυξης δεδομένων. Ωστόσο, αφού πραγματοποιηθεί προκαταρκτικός πειραματισμός στον Explorer, είναι συχνά πολύ πιο εύκολο να εντοπιστεί ένας κατάλληλος αλγόριθμος για ένα συγκεκριμένο σύνολο δεδομένων ή συλλογή συνόλων δεδομένων, χρησιμοποιώντας αυτή την εναλλακτική διεπαφή (Witten, Frank, & Hall, 2011).



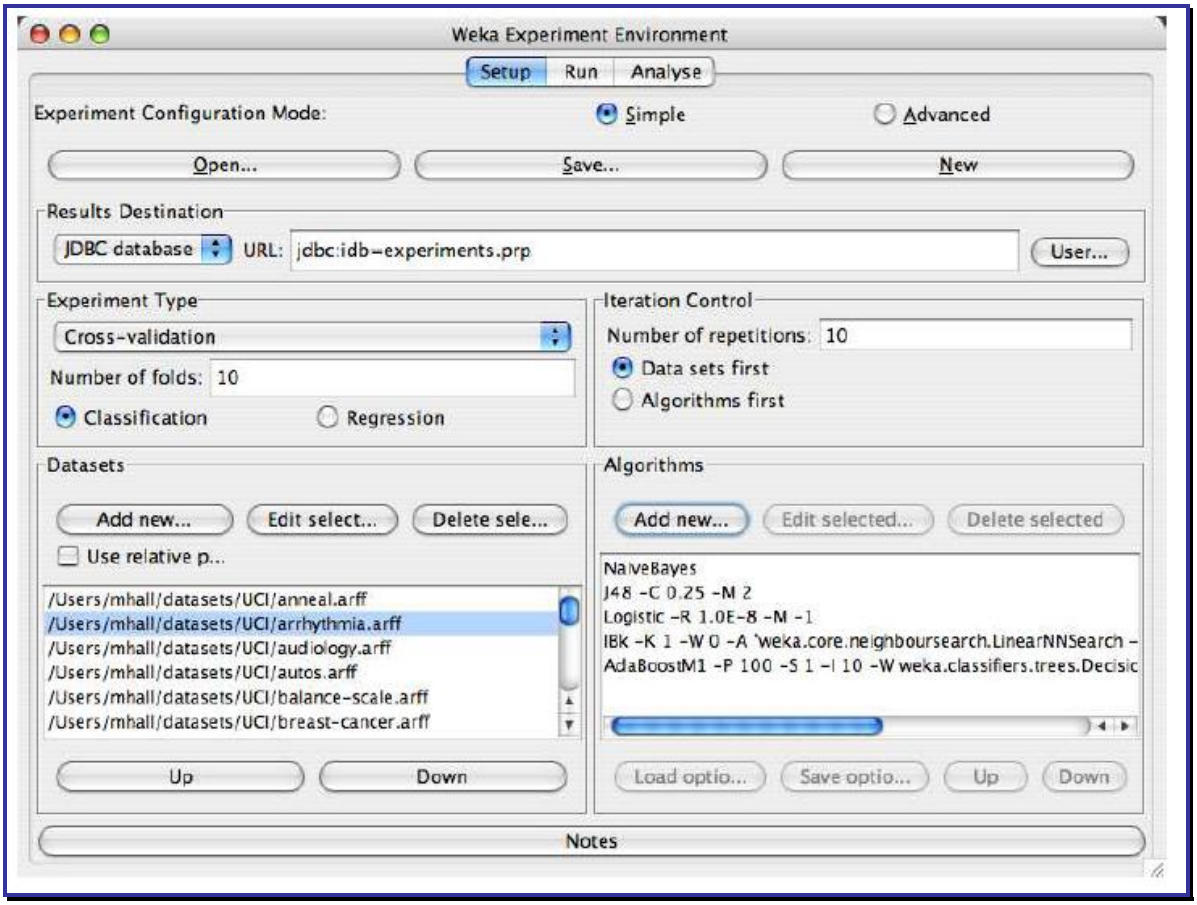
Εικόνα 4.3.1: Η διεπαφή χρήστη WEKA Explorer

Πηγή: https://www.kdd.org/exploration_files/p2V11n1.pdf



Εικόνα 4.3.2: Η διεπαφή χρήστη WEKA Knowledge Flow

Πηγή: https://www.kdd.org/exploration_files/p2V11n1.pdf



Εικόνα 4.3.3: Η διεπαφή χρήστη WEKA Experimenter

Πηγή: https://www.kdd.org/exploration_files/p2V11n1.pdf

4.4 Πλεονεκτήματα χρήσης WEKA

Το Weka είναι μια συλλογή αλγορίθμων μηχανικής μάθησης για την επίλυση πραγματικών προβλημάτων εξόρυξης δεδομένων. Είναι γραμμένο σε Java και λειτουργεί σχεδόν σε οποιαδήποτε πλατφόρμα. Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων είτε να καλούνται από τον δικό σας κώδικα Java (Sudhir, & Kodge, 2013). Η αρχική μη Weba έκδοση του Weka ήταν TCL / TK front-end σε αλγόριθμους μοντελοποίησης (κυρίως τρίτων) που εφαρμόστηκαν σε άλλες γλώσσες προγραμματισμού, καθώς και βοηθητικά προγράμματα προεπεξεργασίας δεδομένων στο C και ένα σύστημα βασισμένο σε Makefile για πειράματα εκμάθησης μηχανικής μάθησης. Αυτή η πρωτότυπη έκδοση

σχεδιάστηκε κυρίως ως εργαλείο για την ανάλυση δεδομένων από αγροτικούς τομείς, αλλά η πιο πρόσφατη εκδοχή βασισμένη στην Java (Weka 3), για την οποία ξεκίνησε η ανάπτυξη το 1997, χρησιμοποιείται τώρα σε πολλούς διαφορετικούς τομείς εφαρμογής, τους σκοπούς και την έρευνα. Τα πλεονεκτήματα του Weka περιλαμβάνουν:

1. Ελεύθερη διαθεσιμότητα βάσει της γενικής δημόσιας άδειας GNU.
2. Φορητότητα, δεδομένου ότι υλοποιείται πλήρως στη γλώσσα προγραμματισμού Java και επομένως λειτουργεί σχεδόν σε οποιαδήποτε σύγχρονη πλατφόρμα υπολογιστών.
3. Μια ολοκληρωμένη συλλογή τεχνικών προεπεξεργασίας και μοντελοποίησης δεδομένων.
4. Ευκολία στη χρήση λόγω των γραφικών διεπαφών χρήστη

Το Weka υποστηρίζει διάφορες τυπικές εργασίες εξόρυξης δεδομένων, ειδικότερα, προεπεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, παλινδρόμηση, οπτικοποίηση και επιλογή χαρακτηριστικών (Sunita, & Lobo, 2011). Όλες οι τεχνικές του Weka βασίζονται στην παραδοχή ότι τα δεδομένα είναι διαθέσιμα ως ένα ενιαίο επίπεδο αρχείο ή σχέση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό χαρακτηριστικών (κανονικά, αριθμητικά ή ονομαστικά χαρακτηριστικά, αλλά υποστηρίζονται επίσης ορισμένοι άλλοι τύποι χαρακτηριστικών).

Η Weka παρέχει πρόσβαση σε βάσεις δεδομένων SQL χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφεται από ένα ερώτημα βάσης δεδομένων. Δεν είναι ικανή για εξόρυξη πολλών σχεσιακών δεδομένων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής συνδεδεμένων πινάκων βάσης δεδομένων σε έναν ενιαίο πίνακα που είναι κατάλληλος για επεξεργασία με χρήση του Weka. Ένας άλλος σημαντικός τομέας που επί του παρόντος δεν καλύπτεται από τους αλγόριθμους που περιλαμβάνονται στην κατανομή Weka είναι η μοντελοποίηση ακολουθιών (Sudhir, & Kodge, 2013).

4.5 Μειονεκτήματα χρήσης WEKA

Πιθανόν το πιο σημαντικό μειονέκτημα των συστημάτων εξόρυξης δεδομένων είναι ότι δεν εφαρμόζουν τις νεότερες τεχνικές. Για παράδειγμα, το MLP που εφαρμόζεται έχει έναν πολύ βασικό αλγόριθμο κατάρτισης και το SVM χρησιμοποιεί μόνο πολυωνυμικούς πυρήνες και δεν υποστηρίζει την αριθμητική εκτίμηση. Επομένως, θα χρειαστεί να συνδυαστεί το WEKA με κάποια από τα άλλα εργαλεία όπως το Netlab ή το SVM_torch. Ένα άλλο σημαντικό μειονέκτημα προκύπτει από το γεγονός ότι το λογισμικό είναι δωρεάν: η τεκμηρίωση για το GUI είναι αρκετά περιορισμένη. Το βιβλίο εξόρυξης δεδομένων Witten και Frank είναι περισσότερο ή λιγότερο μια περίληψη των λειτουργιών του προγράμματος και το κεφάλαιο 8 είναι ένα σεμινάριο για αυτό. Δεν περιγράφει όμως τίποτα για το GUI. Καθώς το λογισμικό αυξάνεται συνεχώς, η τεκμηρίωση δεν είναι ενημερωμένη με τα πάντα (τα πιο ενημερωμένα και πλήρεις πληροφορίες σχετικά με τις επιλογές αλγορίθμου μπορούν να ληφθούν χρησιμοποιώντας την επιλογή -h στη διεπαφή γραμμής εντολών) (Sudhir, & Kodge, 2013).

Ένα τρίτο πιθανό πρόβλημα είναι η κλιμάκωση. Για δύσκολες εργασίες σε μεγάλα σύνολα δεδομένων, ο χρόνος εκτέλεσης μπορεί να είναι αρκετά μεγάλος και η java μερικές φορές δίνει ένα σφάλμα OutOfMemory. Αυτό το πρόβλημα μπορεί να μειωθεί χρησιμοποιώντας την επιλογή «-mx x» όταν καλείται η java, όπου «x» είναι μέγεθος μνήμης (π.χ. '50m'). Για τα μεγάλα σύνολα δεδομένων θα είναι πάντοτε απαραίτητο να μειωθεί το μέγεθος ώστε να είναι σε θέση να εργαστεί εντός λογικών προθεσμιών. Ένα τέταρτο πρόβλημα είναι ότι το GUI δεν εφαρμόζει όλες τις πιθανές επιλογές. Τα πράγματα που θα μπορούσαν να είναι πολύ χρήσιμα, όπως η βαθμολόγηση ενός συνόλου δοκιμών, δεν παρέχονται στο GUI, αλλά μπορούν να καλούνται από τη διεπαφή γραμμής εντολών. Επομένως, μερικές φορές θα χρειαστεί να γίνει μετάβαση μεταξύ GUI και γραμμής εντολών. Τέλος, οι τεχνικές προετοιμασίας δεδομένων και απεικόνισης που προσφέρονται ενδέχεται να μην είναι αρκετές. Οι περισσότερες από αυτές είναι πολύ χρήσιμες, αλλά στις περισσότερες εργασίες εξόρυξης δεδομένων η καλή γνώση των δεδομένων είναι απαραίτητη (Sudhir, & Kodge, 2013).

Συμπεράσματα

Η τεχνολογία μας επιτρέπει πλέον να συλλαμβάνουμε και να αποθηκεύουμε τεράστιες ποσότητες δεδομένων. Η εύρεση προτύπων, τάσεων και ανωμαλιών σε αυτά τα σύνολα δεδομένων και η σύνοψη αυτών με απλά ποσοτικά μοντέλα είναι μία από τις μεγάλες προκλήσεις της πληροφορίας που μετατρέπει την ηλικία στις πληροφορίες και μετατρέπει τις πληροφορίες στη γνώση. Υπήρξε εκπληκτική πρόοδος στην εξόρυξη δεδομένων και την εκμάθηση μηχανών. Η σύνθεση των στατιστικών, η μηχανική μάθηση, η θεωρία των πληροφοριών και η πληροφορική έχουν δημιουργήσει μια σταθερή επιστήμη, με μια σταθερή μαθηματική βάση και με πολύ ισχυρά εργαλεία.

Η σύγκλιση της πληροφορικής και της επικοινωνίας έχει δημιουργήσει μια κοινωνία που τροφοδοτεί με πληροφορίες. Ωστόσο, οι περισσότερες πληροφορίες είναι στην ακατέργαστη μορφή τους: δεδομένα. Εάν τα δεδομένα χαρακτηρίζονται ως καταγεγραμμένα γεγονότα, τότε η πληροφορία είναι το σύνολο των προτύπων ή των προσδοκιών που υποκινούν τα δεδομένα. Υπάρχει ένα τεράστιο ποσό πληροφοριών κλειδωμένο σε βάσεις δεδομένων-πληροφορίες που είναι δυνητικά σημαντικές αλλά δεν έχουν ακόμη ανακαλυφθεί ή αρθρωθεί.

Η εξόρυξη δεδομένων είναι η εξαγωγή των προηγουμένως άγνωστων και δυνητικά χρήσιμων πληροφοριών από δεδομένα. Η ιδέα είναι να οικοδομηθούν προγράμματα ηλεκτρονικών υπολογιστών που αδειάζουν αυτόματα τις βάσεις δεδομένων, αναζητώντας κανονικότητες ή σχέδια. Τα ισχυρά μοτίβα, αν βρεθούν, θα γενικευθούν και πιθανόν να κάνουν ακριβείς προβλέψεις για τα μελλοντικά δεδομένα. Φυσικά, είναι πιθανό να υπάρξουν προβλήματα. Πολλά μοτίβα μπορεί να είναι αδιάφορα, ενώ άλλα να είναι ψεύτικα, εξαρτώμενα από τυχαίες συμπτώσεις στο συγκεκριμένο σύνολο δεδομένων που χρησιμοποιείται.

Επιπλέον, τα πραγματικά δεδομένα μπορεί να είναι ατελή: ορισμένα τμήματα ίσως έχουν αλλοιωθεί και ορισμένα ίσως παραλείπονται. Οτιδήποτε ανακαλυφθεί μπορεί να είναι ανακριβές: Ωστόσο θα υπάρξουν εξαιρέσεις σε κάθε κανόνα και περιπτώσεις που δεν καλύπτονται από κανόνα. Οι αλγόριθμοι πρέπει να είναι αρκετά ισχυροί για

να αντιμετωπίσουν τα ατελείωτα δεδομένα και να εξάγουν κανονικοποιήσεις που είναι ανακριβείς αλλά χρήσιμες.
Το Weka περιέχει μια συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για την ανάλυση δεδομένων και την προγνωστική μοντελοποίηση, μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτές τις λειτουργίες. Το Weka υποστηρίζει διάφορες βασικές διεργασίες εξόρυξης δεδομένων· πιο συγκεκριμένα, προεπεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, παλινδρόμηση, απεικόνιση, και την δυνατότητα επιλογής.

Το Weka έχει τρία βασικά πλεονεκτήματα έναντι των περισσότερων άλλων δεδομένων miningsoftware. Πρώτον, είναι ανοιχτό πρόγραμμα, που όχι μόνο σημαίνει ότι μπορεί να αποκτηθεί δωρεάν, αλλά -όπως είναι σημαντικότερο- είναι διατηρήσιμο χωρίς να εξαρτάται από τη οπιαδήποτε δέσμευση, ενός συγκεκριμένου ιδρύματος ή εταιρείας. Δεύτερον, παρέχει έναν πλούτο αλγορίθμων μηχανικής μάθησης που μπορούν να αναπτυχθούν σε ένα δεδομένο πρόβλημα. Τρίτον, υλοποιείται πλήρως στην Java και λειτουργεί σχεδόν σε οποιαδήποτε πλατφόρμα - ακόμη και σε Personal Digital Assistant.

Το κύριο μειονέκτημα είναι ότι το μεγαλύτερο μέρος της λειτουργικότητας εφαρμόζεται μόνο αν όλα τα δεδομένα διατηρούνται στην κύρια μνήμη. Συμπεριλαμβάνονται μερικοί αλγόριθμοι που είναι σε θέση να επεξεργάζονται δεδομένα σταδιακά ή σε παρτίδες. Ένα δεύτερο μειονέκτημα είναι η πλευρά flp της φορητότητας: μια υλοποίηση Java είναι γενικά κάπως πιο χαλαρή από ό, τι ένα ισοδύναμο σε C / C ++. Όλες οι τεχνικές του Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα απλό αρχείο ή συσχέτιση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό των χαρακτηριστικών.

Βιβλιογραφία

Heckerman, D. (1997). Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery 1, 1, 79-119.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). Bayesian Data Analysis, Chapman & Hall

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (1996). Markov Chain Monte Carlo in Practice, Chapman & Hall,

Boulicaut, J.F., Klemettinen, M., & Mannila, H. (1998). Querying inductive databases: a case study on the MINE RULE operator. 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)}, Nantes, France, September 23-26, p. 194-202.

Mannila, H. (2000). Theoretical Frameworks for Data Mining. SIGKDD Explorations, 1(2), 30-32

Fayyad, U.M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), AAAI Press, p. 1-34.

Han, Jiawei; Kamber, Micheline (2001). Data mining: concepts and techniques. Morgan Kaufmann. p. 5. ISBN 978-1-55860-489-6. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long

Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

Ramageri, B. M. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering, 1(4), 301-305

Patel, J., & Patel, A. (2012). DATA MODELING TECHNIQUES FOR DATA WAREHOUSE. 2. 240-246.

Jiang, Mon-Fong & Tseng, Shian-Shyong & Liao, Shan-Yi. (1999). Data types generalization for data mining algorithms. 3. 928 - 933 vol.3. 10.1109/ICSMC.1999.823352.

Guyer, G., & Rabeler, C. (2018). Data Types (Data Mining). Ανακτήθηκε από https://docs.microsoft.com/

Microsoft. Data Mining Algorithms. Analysi Services-Data Mining. 2016.

Gundecha, P., & Liu, H. (2012). Mining Social Media: A Brief Introduction. Tutorials in operations research,

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social Media Mining, Cambridge University Press.

Han, J., & Kamber, M. (2010). Data Mining Concepts and Techniques. Presentation Slides of Prof Anita Wasilewska.

Witten, I.H., Frank, E., & Hall, M.A. (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco.

Garner, S.R., Cunningham, S.J., Holmes, G., Nevill-Manning, C.G.,Witten, I.H. (1995). "Applying a machine learning workbench: Experience with agricultural databases" (PDF). Proc Machine Learning in Practice Workshop, Machine Learning Conference, Tahoe City, CA, USA. pp. 14–21. Retrieved 2007-06-25.

Reutemann, P., Pfahringer, B., & Frank, E. (2004). "Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners". 17th Australian Joint Conference on Artificial Intelligence (AI2004). Springer-Verlag. Retrieved 2007-06-25.

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludscher, B., & Mock, S. (2004). Kepler: An extensible system for design and execution of scientific workflows. In In SS- DBM, pages 21–23

Witten, I.H., & Frank, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.

Witten, I.H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 2 edition,

Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2008). WEKA manual for version 3-6-0.

Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, 2nd ed., Morgan Kaufmann publishers, SanFrancisco, 2006

Marakas, G.M. (2005). Modern Data Warehousing, Mining, and Visualization, Pearson Education, New Delhi

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4)

Khoussainov, R., Zuo, X., & Kushmerick, N. (2004). Gridenabled Weka: A toolkit for machine learning on the grid. ERCIM News, 59

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Burlington, MA: Morgan Kaufmann Publishers.

Salatas, J. (2011). Extending WEKA. Ανακτήθηκε από https://jsalatas.ictpro.gr/extending-weka/

Sunita B. A., Lobo L. M. R. J. (2011). Data Mining in Educational System using Weka, International Conference on Emerging Technology Trends (ICETT), Proceedings published by International Journal of Computer Applications® (IJCA) Number 3, pp-20-25.

Sudhir, B. J., & Kodge B. G. (2013). Census Data Mining and Data Analysis using WEKA. International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, 2009, Volume 11, Issue 1, pp.10-18

Bharati, M., & Ramageri. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering. 1.

Rukshan, A., Menik, T., & Chandrika, F. (2009). Data Mining Applications: Promise and Challenges. 10.5772/6449.

Antonie, M.L., Zaıane, O. R., & Coman, A. (2001) Application of data mining techniques for medical image classification, Second Intl. Workshop on Multimedia Data Mining (MDM/KDD'2001), pp. 94-101, San Francisco, USA.