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

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Η Προγραμματιστική Διεπαφή (API) προϊόντων λογισμικού Data Mining

Της φοιτήτριας **Επιβλέπων καθηγητής**

Μπεϊλέρη Όλγα **Δέρβος Δ. Δημήτριος**

Αρ. Μητρώου: 03/2399

Θεσσαλονίκη 2011

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή, που σκοπό έχει την ανάλυση της προγραμματιστικής διεπαφής (API) των προϊόντων λογισμικού data mining, εξηγούμε κάποιους βασικούς όρους της εξόρυξης δεδομένων. Αναλύεται περετέρο τι είναι το data mining, η χρησιμότητα του και που εφαρμόζεται καθώς και παρουσιάζονται διάφορα λογισμικά προϊόντα που κάνουν εξόρυξη δεδομένων (εμπορικά και ανοιχτού κώδικα). Στη συνέχεια γίνεται λεπτομερής αναφορά των τριών δημοφιλέστερων εργαλείων data mining που χρησιμοποιούνται ευρέως τόσο στον επιχειρησιακό όσο και στον εκπαιδευτικό τομέα (Intelligent Miner, Microsoft Analysis Services, Weka).

Ακολουθεί ένα κεφάλαιο για την επεξήγηση του τι είναι API και όλες τις θεωρητικές και τεχνικές του λεπτομέρειες. Στο σημαντικότερο κεφάλαιο της εργασίας παρουσιάζονται αναλυτικά οι κυριότεροι αλγόριθμοι και μέθοδοι εξόρυξης δεδομένων. Το τελευταίο κεφάλαιο περιέχει παραδείγματα εφαρμογής του προγράμματος που υλοποιήθηκε σε Java από το API του Weka και τρέχει πάνω σε δεδομένα που χρησιμοποιούνται για εκπαιδευτικά πειράματα. Στο τέλος της εργασίας υπάρχει παράρτημα με το documentation του προγράμματος.

Κατάλογος περιεχομένων

Ευρετήριο πινάκων

Ευρετήριο σχημάτων

ΕΙΣΑΓΩΓΗ

Σκοπός της εργασίας αυτής είναι η έρευνα και η παρουσίαση αλγορίθμων και μεθόδων που χρησιμοποιούν τα API των προϊόντων λογισμικού για εξόρυξη δεδομένων. Επίσης η ανάπτυξη μιας εφαρμογής με την υλοποίηση του API ενός

εργαλείου Data Mining, του Weka. Τα κεφάλαια είναι αριθμημένα έτσι ώστε να γίνουν γνωστοί όλοι οι όροι και οι τεχνικές που χρησιμοποιούνται στην εξόρυξη δεδομένων και στη χρήση ενός API. Αυτό κάνει την εργασία πιο κατανοητή ακόμα και σε άτομα που δεν έχουν μεγάλη συνάφεια με το αντικείμενο χωρίς όμως να στερείται επιστημονικής γνώσης. Εκτός από την ανάλυση των πιο γνωστών αλγορίθμων εξόρυξης δεδομένων, έμφαση δίνεται και στην ανάπτυξη προγράμματος χρησιμοποιώντας το API του Weka. Οι δοκιμές γίνονται με 2 σετ δεδομένων. Το πρώτο είναι για τις κλιματολογικές συνθήκες και πως επηρεάζουν το εάν θα παίξουμε τέννις ή όχι και το άλλο είναι τα δεδομένα Iris που περιγράφονται αναλυτικότερα στο τελευταίο κεφάλαιο.

ΚΕΦΑΛΑΙΟ 1

Εξόρυξη Δεδομένων

ΕΙΣΑΓΩΓΗ

Η εξόρυξη δεδομένων (Data Mining) είναι ένας κλάδος της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης, μια διαδικασία εξόρυξης προτύπων

από δεδομένα.[1] Η εξόρυξη δεδομένων θεωρείται ως ένα όλο και πιο σημαντικό εργαλείο από σύγχρονες επιχειρήσεις ώστε να μετατρέψουν τα δεδομένα σε επιχειρηματική ευφυΐα δίνοντας ένα πληροφοριακό πλεονέκτημα. Χρησιμοποιείται σήμερα σε ένα ευρύ φάσμα χαρακτηριστικών πρακτικών(profiling practices), όπως το marketing , την επιτήρηση(surveillance), την ανίχνευση απάτης και την επιστημονική ανακάλυψη.

Σε γενικές γραμμές, εξόρυξη δεδομένων είναι η διαδικασία της ανάλυσης δεδομένων από διαφορετικές οπτικές γωνίες και συνοψίζοντας το σε χρήσιμες πληροφορίες - πληροφορίες που μπορούν να χρησιμοποιηθούν για την αύξηση των εσόδων, περικοπές δαπανών, ή και τα δύο. Το λογισμικό εξόρυξης είναι ένα από μια σειρά αναλυτικών εργαλείων για την ανάλυση των δεδομένων. Επιτρέπει στους χρήστες να αναλύουν δεδομένα από πολλές διαφορετικές διαστάσεις ή γωνίες και να ταξινομήσουν και να συνοψίσουν τις σχέσεις που εντοπίζονται. Τεχνικά, εξόρυξη δεδομένων είναι η διαδικασία της εύρεσης συσχετισμών ή μοτίβων ανάμεσα σε δεκάδες τομείς σε μεγάλες σχεσιακές βάσεις δεδομένων.

Αν και εξόρυξη δεδομένων είναι ένας σχετικά νέος όρος, η τεχνολογία αυτή δεν είναι. Οι εταιρείες έχουν χρησιμοποιήσει ισχυρούς υπολογιστές για να "κοσκινίσουν" όγκους δεδομένων από σαρωτή σούπερ μάρκετ και έχουν αναλύσει εκθέσεις έρευνας αγοράς για χρόνια. Ωστόσο, οι συνεχείς καινοτομίες στην επεξεργαστική ισχύ των υπολογιστών, την αποθήκευση σε δίσκους και το λογισμικό στατιστικής αυξάνουν δραματικά την ακρίβεια της ανάλυσης ενώ μειώνουν το κόστος.

Η εξόρυξη δεδομένων, είναι μια ισχυρή νέα τεχνολογία με μεγάλες δυνατότητες και έχει βοηθήσει τις εταιρείες να επικεντρωθούν στις πιο σημαντικές πληροφορίες από τις αποθήκες δεδομένων τους. Τα εργαλεία εξόρυξης δεδομένων προβλέπουν μελλοντικές τάσεις και συμπεριφορές, που επιτρέπει στις επιχειρήσεις να παίρνουν προληπτικές αποφάσεις βασιζόμενες στη γνώση. Οι αυτοματοποιημένες, μελλοντικές αναλύσεις που παρέχονται από την εξόρυξη δεδομένων κινούνται πέρα από τις αναλύσεις των γεγονότων του παρελθόντος που παρέχονται από αναδρομικά εργαλεία τυπικών συστημάτων υποστήριξης αποφάσεων. Τα εργαλεία εξόρυξης δεδομένων μπορούν να απαντήσουν σε ερωτήσεις των επιχειρήσεων που παραδοσιακά ήταν πολύ χρονοβόρες να επιλυθούν. Καθαρίζουν βάσεις δεδομένων για τα κρυμμένα μοτίβα και την εύρεση προβλέψιμων πληροφοριών που εμπειρογνώμονες μπορεί να χάσουν επειδή βρίσκεται έξω από τις προσδοκίες τους.

Οι περισσότερες εταιρείες έχουν ήδη συλλέξει και επεξεργαστεί τεράστιες ποσότητες δεδομένων. Οι τεχνικές εξόρυξης δεδομένων μπορούν να υλοποιηθούν γρήγορα σε ήδη υπάρχον λογισμικό και πλατφόρμες hardware για να ενισχύσει την

αξία των ήδη υπαρχόντων πόρων πληροφοριών και μπορούν να ενσωματωθούν με νέα προϊόντα και

συστήματα, όταν αρχίσουν να χρησιμοποιούνται. Όταν εφαρμόζονται σε υψηλών επιδόσεων client / server ή παράλληλης επεξεργασίας υπολογιστές, τα εργαλεία εξόρυξης δεδομένων μπορούν να αναλύσουν τεράστιες βάσεις δεδομένων για την παροχή απαντήσεων σε ερωτήσεις όπως, «Ποιοί πελάτες είναι πιο πιθανό να ανταποκριθούν στην επόμενη διαφημιστική αλληλογραφία μου, και γιατί;"

Οι πιο συχνά χρησιμοποιούμενες τεχνικές στην εξόρυξη δεδομένων είναι οι εξής:

* Τεχνητά νευρωνικά δίκτυα: Μη-γραμμική προβλεπτικά μοντέλα που μαθαίνουν μέσω της κατάρτισης και μοιάζουν στη δομή με βιολογικά νευρωνικά δίκτυα.

* Δέντρα Αποφάσεων: Δομές σε σχήμα δέντρων που αντιπροσωπεύουν σύνολα αποφάσεων. Οι αποφάσεις αυτές παράγουν κανόνες για την ταξινόμηση του συνόλου δεδομένων. Ειδικές μέθοδοι αποφάσεων περιλαμβάνουν Classification and Regression Trees (CART) και Chi Square Automatic Interaction Detection (CHAID).

* Γενετικοί αλγόριθμοι: τεχνικές βελτιστοποίησης που χρησιμοποιούν μεθόδους όπως το γενετικό συνδυασμό, τη μετάλλαξη και τη φυσική επιλογή ενός σχεδίου που βασίζεται στις έννοιες της εξέλιξης.

* Μέθοδος κοντινότερου γείτονα (Nearest neighbor method): Μια τεχνική που κατατάσσει κάθε εγγραφή σε ένα σύνολο δεδομένων βασιζόμενο σε ένα συνδυασμό των κατηγοριών της k εγγραφής (-ών) που μοιάζουν περισσότερο με αυτό σε ένα ιστορικό σύνολο δεδομένων (όπου k ³ 1). Μερικές φορές ονομάζεται και τεχνική του k-πλησιέστερου γείτονα.

* Κανόνας επαγωγής: Η εξαγωγή χρήσιμων if-then κανόνων από δεδομένα που βασίζονται σε στατιστική σημαντικότητα (Η στατιστική σημαντικότητα ενός αποτελέσματος είναι η πιθανότητα ότι η παρατηρηθείσα σχέση (π.χ., μεταξύ των μεταβλητών) ή της διαφοράς (π.χ., μεταξύ των μέσων) σε ένα δείγμα εμφανίστηκε κατά καθαρή τύχη και ότι στον πληθυσμό από τον οποίο το δείγμα προήλθε, καμία τέτοια σχέση ή διαφορά δεν υπάρχει.).

Πολλές από αυτές τις τεχνολογίες έχουν χρησιμοποιηθεί για περισσότερο από μια δεκαετία σε εξειδικευμένα εργαλεία ανάλυσης που λειτουργούν με σχετικά μικρού όγκου δεδομένα. Οι δυνατότητες αυτές εξελίσσεται πια για να ενσωματωθούν με τις αποθήκες δεδομένων βιομηχανικών προτύπων και πλατφόρμες OLAP.

Η εξόρυξη δεδομένων περιλαμβάνει συνήθως τέσσερις κατηγορίες εργασιών:

* Ομαδοποίηση (Clustering) - είναι η εργασία της ανακάλυψης ομάδων και δομών σε δεδομένα που είναι κατά κάποιο τρόπο "παρόμοια", χωρίς τη βοήθεια γνωστών δομών δεδομένων.

* Η κατάταξη (Classification) - είναι η εργασία της γενίκευσης γνωστής δομής για να ισχύει για τα νέα δεδομένα. Για παράδειγμα, ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου ενδέχεται να προσπαθήσει να χαρακτηρίσει ένα μήνυμα ηλεκτρονικού ταχυδρομείου ως νόμιμο ή spam. Κοινοί αλγόριθμοι περιλαμβάνουν δέντρα αποφάσεων, μέθοδος πλησιέστερου γείτονα, ταξινόμηση αφελούς Bayesian, νευρωνικά δίκτυα και μηχανές υποστήριξης διανυσμάτων.

* Υποχώρηση (Regression) - Προσπάθειες να βρεθεί μια λειτουργία που μοντελοποιεί τα δεδομένα λιγότερο λανθασμένα.

* Μάθηση συλλογικού κανόνα (Association rule learning)- Ψάχνει για τις σχέσεις μεταξύ των μεταβλητών. Για παράδειγμα, ένα σούπερ-μάρκετ μπορεί να συλλέξει στοιχεία για τις αγοραστικές συνήθειες των καταναλωτών. Χρησιμοποιώντας αυτό, το σούπερ μάρκετ μπορεί να διαπιστώσει ποια προϊόντα αγοράζονται συχνά μαζί και χρησιμοποιεί τις πληροφορίες αυτές για εμπορικούς σκοπούς. Αυτό μερικές φορές αναφέρεται ως ανάλυση καλαθιού αγοράς (market basket analysis).

1.1 Ιστορικό

Η χειροκίνητη εξόρυξη προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Πρώιμες μέθοδοι αναγνώρισης προτύπων σε δεδομένα περιλαμβάνουν το «Θεώρημα του Bayes (1700) και τη παλινδρομική ανάλυση (regression analysis) (1800). Η πολλαπλασιαστικά αυξανόμενη δύναμη της τεχνολογίας των υπολογιστών έχει αυξήσει στη συλλογή δεδομένων, την αποθήκευση και το χειρισμό τους. Καθώς τα σύνολα δεδομένων έχουν αυξηθεί σε μέγεθος και πολυπλοκότητα,η άμεση πρακτική ανάλυση δεδομένων έχει αυξηθεί όλο και περισσότερο με τις έμμεσες, αυτόματες επεξεργασίες δεδομένων. Αυτό έχει επιτευχθεί από άλλες ανακαλύψεις στην επιστήμη των υπολογιστών, όπως τα νευρωνικά δίκτυα ,το clustering, γενετικούς αλγόριθμους (1950), δένδρα αποφάσεων (1960) και την υποστήριξη διανυσματικών μηχανών (support vector machines) (1980). Η εξόρυξη δεδομένων είναι η διαδικασία της εφαρμογής των μεθόδων αυτών σε δεδομένα με σκοπό την αποκάλυψη κρυμμένων μοντέλων-patterns.[2]

Ένας κύριος λόγος για τη χρήση του data mining είναι να βοηθήσει στην ανάλυση των συλλογών παρατηρήσεων της συμπεριφοράς. Τα δεδομένα αυτά είναι ευάλωτα σε συγγραμμικότητα λόγω της άγνοιας των συσχετισμών. Ένα

αναπόφευκτο γεγονός της εξόρυξης δεδομένων είναι ότι το (υπο-) σύνολο των στοιχείων που αναλύθηκε δεν μπορεί να είναι αντιπροσωπευτικό του συνόλου ενός τομέα και ως εκ τούτου ενδέχεται να μην περιέχει παραδείγματα ορισμένων κρίσιμων σχέσεων και συμπεριφορών που υπάρχουν σε ορισμένα άλλα τμήματα του τομέα. Για την αντιμετώπιση αυτών των ειδών θεμάτων, η ανάλυση μπορεί να αυξηθεί με περαματικές βάσεις και άλλες προσεγγίσεις, όπως η επιλογή μοντέλων (Choice Modelling) που προορίζονται για ανθρώπινα δεδομένα. Σε αυτές τις περιπτώσεις, οι συσχετίσεις μπορούν είτε να ελέγχονται είτε και να καταργούνται, κατά τη διάρκεια της κατασκευής του πειραματικού σχεδιασμού (experimental design.) .

Υπήρξαν κάποιες προσπάθειες για τον καθορισμό προτύπων για την εξόρυξη δεδομένων, όπως για παράδειγμα το 1999 το European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) και του 2004 το πρότυπο Java Data Mining (JDM 1,0). Αυτά τα πρότυπα εξελίσσονται και νεότερες εκδόσεις τους είναι υπό ανάπτυξη. Ανεξάρτητα από αυτές τις προσπάθειες τυποποίησης, ελεύθερα διαθέσιμα ανοικτά συστήματα λογισμικού όπως το R Project, το Weka, το KNIME, το RapidMiner, το jHepWork και άλλα, έχουν καθορίσει άτυπα το πρότυπο για τον ορισμό των διαδικασιών data-mining. Αξίζει να σημειωθεί ότι όλα αυτά τα συστήματα είναι δυνατό να εισάγουν και να εξάγουν μοντέλα σε PMML (Predictive Model Markup Language) το οποίο παρέχει έναν πρότυπο τρόπο να εκπροσωπούν μοντέλα εξόρυξης δεδομένων, έτσι ώστε να μπορούν να μοιραστούν μεταξύ διαφόρων στατιστικών εφαρμογών. [3]Η PMML είναι μια γλώσσα που βασίζονται στο XML και αναπτύχθηκε από το Data Mining Group (DMG), μια ανεξάρτητη ομάδα που αποτελείται από πολλές εταιρείες εξόρυξης δεδομένων. Η έκδοση PMML 4.0 κυκλοφόρησε τον Ιούνιο του 2009.

1.2 Διαδικασία

1.2.1 Προ-επεξεργασία

Πριν οι αλγόριθμοι για την εξόρυξη δεδομένων χρησιμοποιηθούν, πρέπει να διαμορφωθεί ένα σύνολο δεδομένων. Αφού η εξόρυξη δεδομένων μπορεί να αποκαλύψει μόνο πρότυπα που υπάρχουν ήδη στα στοιχεία, το σύνολο δεδομένων πρέπει να είναι αρκετά μεγάλο ώστε να περιέχει αυτά τα πρότυπα παραμένοντας όμως συνοπτικό ώστε να εξορυχθεί σε ένα αποδεκτό χρονικό διάστημα. Μια συνήθης πηγή για τα δεδομένα είναι μια Datamart ή μια αποθήκη δεδομένων. Η προ-επεξεργασία είναι απαραίτητη για την ανάλυση των συνόλων δεδομένων με πολλές μεταβλητές πριν την ομαδοποίηση ή την εξόρυξη.

Το σύνολο στη συνέχεια καθαρίζεται. Ο καθαρισμός αφαιρεί τις παρατηρήσεις με θόρυβο και τα στοιχεία με ελλείψεις.

Τα καθαρά στοιχεία μειώνονται σε διανύσματα με χαρακτηριστικά γνωρίσματα, ένα διάνυσμα ανά παρατήρηση. Ένα διάνυσμα είναι μια συνοπτική έκδοση της αρχικής παρατήρησης δεδομένων. Για παράδειγμα, μια ασπρόμαυρη εικόνα ενός προσώπου που είναι 100px από 100px, περιέχει 10.000 κομμάτια των ανεπεξέργαστων δεδομένων. Αυτό θα μπορούσε να μετατραπεί σε ένα διάνυσμα γνωρισμάτων, εντοπίζοντας τα μάτια και το στόμα στην εικόνα. Κάτι τέτοιο θα μείωνε τα στοιχεία κάθε φορέα από 10.000 κομμάτια σε τρεις κωδικούς για τις θέσεις, μειώνοντας εντυπωσιακά το μέγεθος του συνόλου δεδομένων που πρέπει να εξορυχθεί, και ως εκ τούτου μειώνοντας την επεξεργαστική προσπάθεια. Το χαρακτηριστικό που θα επιλεγεί εξαρτάται από το ποιος είναι ο στόχος (οι), προφανώς, επιλέγοντας το "σωστό" χαρακτηριστικό έχει θεμελιώδη σημασία για την επιτυχή εξόρυξη δεδομένων.

Τα χαρακτηριστικά διανύσματα χωρίζονται σε δύο ομάδες, το «σύνολο κατάρτισης» και η «σύνολο δοκιμών". Το σύνολο εκπαίδευσης χρησιμοποιείται για να "εκπαιδεύσουμε" τον αλγόριθμο εξόρυξης δεδομένων, ενώ το σύνολο δοκιμής χρησιμοποιείται για την επαλήθευση της ακρίβειας των τυχόν μοντέλων που ανακαλύπτονται.

1.2.2 Επικύρωση αποτελεσμάτων

Το τελικό βήμα στην ανακάλυψη γνώσης από τα δεδομένα, είναι η επαλήθευση των μοντέλων που παράγονται από τους αλγορίθμων της εξόρυξης δεδομένων που εμφανίζονται στο ευρύτερο σύνολο δεδομένων. Δεν είναι όλα τα patterns που βρέθηκαν από τους αλγορίθμους εξόρυξης αναγκαστικά έγκυρα. Είναι σύνηθες για τους αλγορίθμους εξόρυξης να βρίσκουν πρότυπα στο σύνολο εκπαίδευσης, τα οποία δεν βρίσκονται στον γενικό σύνολο δεδομένων, αυτό ονομάζεται overfitting. Για να ξεπεραστεί αυτό, στην αξιολόγηση χρησιμοποιείται δοκιμαστικά στοιχεία στα οποία ο αλγόριθμος εξόρυξης δεν είχε εκπαιδευτεί. Τα διδαγμένα πρότυπα εφαρμόζονται σε αυτό το σύνολο δοκιμής και τα αποτελέσματα συγκρίνονται με το επιθυμητό αποτέλεσμα. Για παράδειγμα, ένας αλγόριθμος που προσπαθεί να διακρίνει ένα spam από ένα νόμιμο e-mail θα εκπαιδευτεί σε ήδη καταρτισμένο δείγμα. Αφού εκπαιδευτούν, τα διδαγμένα πρότυπα θα εφαρμοστούν στο σύνολο των e-mail στα οποία δεν είχαν εκπαιδευθεί, η ακρίβεια αυτών των μοντέλων μπορούν έπειτα να μετρηθούν από το πόσα e-mail κατατάσσονται σωστά. Ένας αριθμός στατιστικών μεθόδων μπορούν να χρησιμοποιηθούν για την αξιολόγηση του αλγορίθμου, όπως οι καμπύλες ROC.

Αν τα διδαγμένα πρότυπα δεν πληρούν τις επιθυμητές προδιαγραφές, τότε είναι απαραίτητο να επαναξιολογηθούν και να αλλάξει η προ- επεξεργασία και η εξόρυξη δεδομένων. Αν τα πρότυπα πληρούν τις επιθυμητές προδιαγραφές, στη συνέχεια, το τελικό βήμα είναι να ερμηνευτούν και να μετατραπούν σε γνώση.

1.3 Αξιοσημείωτες χρήσεις

1.3.1 Επιχειρήσεις

Η εξόρυξη δεδομένων μπορεί να συμβάλει σημαντικά σε εφαρμογές διαχείρισης πελατιακών σχέσεων. Εξελιγμένες μέθοδοι μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση των πόρων σε πολλές καμπάνιες, ώστε να μπορεί κανείς να προβλέψει σε ποια προσφορά ένα άτομο είναι πιο πιθανό να ανταποκριθεί από όλες τις πιθανές προσφορές. Επιχειρήσεις που χρησιμοποιούν εξόρυξη δεδομένων μπορεί να δουν κέρδος από την επένδυση, αλλά και να αναγνωρίσουν ότι ο αριθμός των μοντέλων πρόβλεψης μπορεί γρήγορα να γίνει πολύ μεγάλο. Η εξόρυξη δεδομένων μπορεί επίσης να είναι χρήσιμη για την υπηρεσία ανθρώπινων πόρων κατά τον προσδιορισμό των χαρακτηριστικών των πιο επιτυχημένων εργαζομένων τους. Ένα άλλο παράδειγμα της εξόρυξης δεδομένων, που συχνά αποκαλείται ανάλυση καλαθιού της αγοράς, σχετίζεται με τη χρήση της στον τομέα των λιανικών πωλήσεων. Η ανάλυση του καλαθιού της αγοράς έχει χρησιμοποιηθεί επίσης για να προσδιορίσει τα πρότυπα αγοράς των καταναλωτών. Το Data Mining είναι ένα εξαιρετικά αποτελεσματικό εργαλείο στον τομέα της βιομηχανίας εμπορίας καταλόγου.

1.3.2 Επιστήμη και τεχνολογία

Τα τελευταία χρόνια, η εξόρυξη δεδομένων έχει χρησιμοποιηθεί ευρέως στον τομέα της επιστήμης και της μηχανικής, όπως η βιοπληροφορική , η γενετική , η ιατρική , η εκπαίδευση και η μηχανική ηλεκτρικής ενέργειας.

Στον τομέα της μελέτης για την ανθρώπινη γενετική, ένας σημαντικός στόχος είναι να μάθουμε πώς οι αλλαγές στην αλληλουχία του DNA του κάθε ενός επηρεάζουν τον κίνδυνο ανάπτυξης κοινών ασθενειών όπως ο καρκίνος . Αυτό είναι πολύ σημαντικό διότι συμβάλει στη βελτίωση της διάγνωσης, της πρόληψης και της θεραπείας των ασθενειών.

Στον τομέα της μηχανικής ηλεκτρικής ενέργειας, οι τεχνικές εξόρυξης δεδομένων έχουν χρησιμοποιηθεί ευρέως για την παρακολούθηση της κατάστασης της υψηλής τάσης του ηλεκτρικού εξοπλισμού. Ο σκοπός της παρακολούθησης της κατάστασης είναι να αντληθούν πολύτιμες πληροφορίες για τη μονωτική κατάσταση της υγείας του εξοπλισμού.

Ένας άλλος τομέας εφαρμογής εξόρυξης δεδομένων στον τομέα της επιστήμης και μηχανικής είναι η εκπαιδευτική έρευνα, όπου η εξόρυξη δεδομένων έχει χρησιμοποιηθεί για τη μελέτη των παραγόντων που οδηγούν τους μαθητές να επιλέξουν να συμμετάσχουν σε συμπεριφορές που μειώνουν τη μάθηση τους [25]

και να κατανοήσουμε τους παράγοντες που επηρεάζουν τη διατήρηση των φοιτητών.

Άλλα παραδείγματα της εφαρμογής των τεχνικών εξόρυξης δεδομένων είναι τα βιοϊατρικά δεδομένα που διευκολύνουν τον τομέα οντολογίας , [27] την εξόρυξη δεδομένα των κλινικών μελετών, [28], την ανάλυση της κυκλοφορίας χρησιμοποιώντας SOM, [29] και τα λοιπά.

1.3.3 Χωροταξική εξόρυξη δεδομένων

Η Χωροταξική εξόρυξη δεδομένων είναι η εφαρμογή τεχνικών εξόρυξης γνώσης σε χωροταξικά δεδομένα. Η Χωροταξική εξόρυξη δεδομένων ακολουθεί τα ίδια καθήκοντα στην εξόρυξη δεδομένων, με τελικό στόχο να βρουν πρότυπα στη γεωγραφία. Μέχρι σήμερα, η εξόρυξη δεδομένων και των Geographic Information Systems (GIS) έχουν υπάρξει ως δύο ξεχωριστές τεχνολογίες, το καθένα με, τις δικές του μεθόδους και προσεγγίσεις για την ανάλυση δεδομένων και την οπτικοποίηση τους.

Η εξόρυξη δεδομένων, η οποία είναι εν μέρει αυτοματοποιημένη να αναζητά για κρυφά σχέδια σε μεγάλες βάσεις δεδομένων, και προσφέρει μεγάλα δυνατά οφέλη για εφαρμογή λήψης αποφάσεων που βασίζονται στο GIS. Πρόσφατα, το έργο της ενσωμάτωσης αυτών των δύο τεχνολογιών έχει καταστεί κρίσιμο, ιδιαίτερα όσον αφορά διάφορους δημόσιους και ιδιωτικούς οργανισμούς που διαθέτουν τεράστιες βάσεις δεδομένων με θεματικά και γεωγραφικά δεδομένα και αρχίζουν να συνειδητοποιούν το τεράστιο δυναμικό στις πληροφορίες που κρύβονται εκεί. Μεταξύ των οργανισμών αυτών είναι:

- γραφεία που απαιτούν ανάλυση και διάδοση των στατιστικών στοιχείων
- υπηρεσίες γεωαναφοράς δημόσιας υγείας που ψάχνουν για εξηγήσεις έξαρσης κρουσμάτων
- περιβαλλοντικοί οργανισμοί αξιολόγησης του αντίκτυπου των μεταβαλλόμενων προτύπων χρήσης γης για την κλιματική αλλαγή
- geo-μάρκετινγκ εταιρείες που κάνουν την κατηγοριοποίηση των πελατών με βάση τη χωροταξική τους θέση.

1.3.4 Προκλήσεις

Οι Γεω-Χωροταξικές αποθήκες δεδομένων τείνουν να είναι πολύ μεγάλες. Επιπλέον, οι υπάρχουσες σειρές δεδομένων GIS είναι συχνά κατακερματισμένες σε χαρακτηριστικό και ιδιότητα των εξαρτημάτων, που συμβατικά αρχειοθετούνται σε υβριδικά συστήματα διαχείρισης δεδομένων. Οι αλγοριθμικές απαιτήσεις διαφέρουν σημαντικά για σχεσιακή διαχείριση δεδομένων και τοπολογική

διαχείριση δεδομένων. [32] Σχετικό με αυτό είναι το εύρος και η ποικιλομορφία της γεωγραφικής μορφής των δεδομένων, που παρουσιάζει επίσης μοναδικές προκλήσεις. Η ψηφιακή επανάσταση γεωγραφικών δεδομένων είναι η δημιουργία νέων τύπων της μορφής των δεδομένων πέρα από τις παραδοσιακές "vector" και "raster" μορφές. Γεωγραφικές αποθήκες δεδομένων περιλαμβάνουν όλο και πιο άρρωστα-δομημένων δεδομένων, όπως εικόνες και γεωαναφορές multi-media. [33]

1.3.5 Εποπτείας

Προηγούμενη εξόρυξη δεδομένων χρησιμοποιείται για το τρομοκρατικό προγράμματα στο πλαίσιο της αμερικανικής κυβέρνησης και περιλαμβάνει την Total Information Awareness (TIA) το πρόγραμμα Secure Flight (παλαιότερα γνωστό ως Computer-Assisted επιβατών Prescreening System (CAPPS II)), Ανάλυση, Διάδοση, Οπτικοποίηση, Insight, Semantic Enhancement, και το Multi-state Anti-Terrorism Information Exchange. [41]

1.3.6 Pattern εξόρυξης

Το "Πρότυπο εξόρυξης" είναι μια τεχνική εξόρυξης δεδομένων που περιλαμβάνει διαπίστωση υπαρχόντων πρότυπων στα δεδομένα. Σε αυτό το πλαίσιο τα πρότυπα συχνά σημαίνουν κανόνες συσχέτισης. Το αρχικό κίνητρο για την αναζήτηση κανόνων συσχέτισης προέκυψε από την επιθυμία για την ανάλυση δεδομένων συναλλαγών στο σούπερ μάρκετ, δηλαδή, να εξετάσει τη συμπεριφορά των πελατών όσον αφορά τα αγορασθέντα προϊόντα.

Στο πλαίσιο της εξόρυξης πρότυπου ως εργαλείο για τον εντοπισμό τρομοκρατικών δραστηριοτήτων, το Εθνικό Συμβούλιο Έρευνας δίνει τον εξής ορισμό: "Το Πρότυπο εξόρυξης δεδομένων αναζητά πρότυπα (συμπεριλαμβανομένων των ανώμαλων πρότυπων δεδομένων) που θα μπορούσαν να συνδέονται με τρομοκρατικές δραστηριότητες - αυτά τα σχέδια θα μπορούσαν να θεωρηθούν μικρά σήματα σε ένα μεγάλο ωκεανό θορύβου. " [43] [44] [45] Το Pattern Mining περιλαμβάνει νέους τομείς, μια τέτοια είναι η Music Information Retrieval (MIR), όπου τα πρότυπα που φαίνονται στους διαχρονικούς και μη διαχρονικούς τομείς εισάγονται σε τεχνικές αναζήτησης και ανακάλυψης της κλασική γνώσης.

1.3.7 Απόρρητο και δεοντολογία

Μερικοί άνθρωποι πιστεύουν ότι η εξόρυξη δεδομένων είναι ηθικά ουδέτερη. [46] Είναι σημαντικό να σημειωθεί ότι ο όρος εξόρυξη δεδομένων δεν έχει ηθικές επιπτώσεις. Ο όρος συνδέεται συχνά με την εξόρυξη των πληροφοριών σε σχέση με τη συμπεριφορά των ανθρώπων. Ωστόσο, η εξόρυξη δεδομένων είναι μια

στατιστική τεχνική που εφαρμόζεται σε ένα σύνολο πληροφοριών, ή ένα σύνολο δεδομένων. Συνδυάζοντας αυτά τα σύνολα δεδομένων με τους ανθρώπους είναι μια ακραία στένωση των τύπων δεδομένων που είναι διαθέσιμα στην τεχνολογική κοινωνία του σήμερα.

Η εξόρυξη δεδομένων απαιτεί προετοιμασία των δεδομένων που μπορεί να αποκαλύψει τις πληροφορίες ή τα σχέδια που ενδέχεται να διακυβεύσει εμπιστευτικές και απόρρητες υποχρεώσεις. Ένας κοινός τρόπος για να συμβεί αυτό είναι μέσω άθροισης δεδομένων . Άθροιση δεδομένων είναι όταν τα δεδομένα είναι τα δεδουλευμένα, πιθανόν από διάφορες πηγές, και μαζί, ώστε να μπορούν να αναλυθούν.

Συνιστάται ότι ένα άτομο έχει ενημερωθεί για τα ακόλουθα πριν τα δεδομένα συλλεχθούν:

- το σκοπό της συλλογής δεδομένων και τα σχέδια εξόρυξης δεδομένων,
- το πώς τα δεδομένα θα χρησιμοποιηθούν,
- το οποίοι θα είναι σε θέση να εξορύξουν τα δεδομένα και τη χρήση τους,
- την ασφάλεια γύρω από την πρόσβαση στα δεδομένα, και, επιπλέον,
- τον τρόπο με τον οποίο συλλέγονται τα δεδομένα και το πως ενημερώνονται. [50]

1.3.8 Έρευνες Marketplace

Αρκετοί ερευνητές και οργανισμοί έχουν διεξάγει αξιολογήσεις των εργαλείων εξόρυξης δεδομένων και τις έρευνες των data miners. Αυτά προσδιορίζουν ορισμένα από τα πλεονεκτήματα και τις αδυναμίες των πακέτων λογισμικού. Επίσης, παρέχουν μια επισκόπηση των προτιμήσεων συμπεριφοράς, καθώς και τις απόψεις των data miners.

ΚΕΦΑΛΑΙΟ 2

Λογισμικό Εξόρυξης Δεδομένων

ΕΙΣΑΓΩΓΗ

Οι σημερινές επιχειρήσεις απαιτούν ανώτερα εργαλεία για να είναι ανταγωνιστικές σε μια συνεχώς εξελισσόμενη παγκόσμια αγορά. Το λογισμικό εξόρυξης παρέχει σημαντικές πληροφορίες για την επιχείρηση, από τον εντοπισμό των τάσεων, απομονώνοντας πληροφορίες ζωτικής σημασίας, καθώς και το ουσιαστικό περιεχόμενο του μεγάλου όγκου δεδομένων. Με αυτόματη και ακαριαία εξόρυξη μιας ποικιλίας πληροφοριών, ένα ανώτερο λογισμικό μπορεί να προσφέρει το είδος της σημαντικής λεπτομέρειας με το οποίο οι εταιρείες μπορούν να είναι αποτελεσματικά ανταγωνιστικές. Το λογισμικό εξόρυξης δεδομένων μπορεί να παρέχει το εύρος της νοημοσύνης που απαιτούν οι επιτυχημένες επιχειρήσεις και τη δυνατότητα για τη διάδοση των κρίσιμων πληροφοριών σε ολόκληρο τον οργανισμό. Στο κεφάλαιο αυτό παρατίθενται ορισμένες ευρέως γνωστές εφαρμογές εξόρυξης δεδομένων τόσο εμπορικής χρήσης όσο και open source.

2.1 SAS (Statistical Analysis System)

Το SAS (αρχικά Statistical Analysis System) είναι ένα ολοκληρωμένο σύστημα προϊόντων λογισμικού που παρέχονται από τη SAS Institute Inc.που επιτρέπει στους προγραμματιστές να πραγματοποιούν:

- εισαγωγή δεδομένων, ανάκτηση, διαχείριση και εξόρυξη
- σύνταξη εκθέσεων και γραφικών
- στατιστική ανάλυση
- επιχειρηματικό σχεδιασμό, πρόβλεψη, και υποστήριξη αποφάσεων
- επιχειρησιακής έρευνα και διαχείριση έργων
- βελτίωση της ποιότητας
- ανάπτυξη εφαρμογών
- αποθήκευση δεδομένων (απόσπασμα, μετατροπή, φορτίο)
- ανεξάρτητη πλατφόρμα και απομακρυσμένους υπολογιστές

Επιπλέον, το SAS έχει πολλές επιχειρηματικές λύσεις που επιτρέπουν λύσεις λογισμικού σε μεγάλη κλίμακα για τομείς όπως η διαχείριση IT, η διαχείριση ανθρώπινων πόρων, η οικονομική διαχείριση, η επιχειρηματική ευφυΐα, η διαχείριση πελατειακών σχέσεων και άλλα.

Η SAS οδηγείται από τα SAS προγράμματα, που καθορίζουν μια σειρά ενεργειών που πρέπει να εκτελεστούν σε δεδομένα που αποθηκεύονται ως πίνακες. Αν και υπάρχουν μη-προγραμματιστικές γραφικές διεπαφές στη SAS (όπως η SAS Enterprise Guide), τα εν λόγω GUI είναι τις περισσότερες φορές απλά ένα front-end που αυτοματοποιεί ή διευκολύνει την παραγωγή των προγραμμάτων SAS. Οι λειτουργίες του περιεχομένου της SAS, είναι προσβάσιμα μέσω διασυνδέσεων προγραμματιστικών εφαρμογών, με τη μορφή δηλώσεων και διαδικασιών.

Ένα πρόγραμμα SAS έχει τρία βασικά μέρη:

- το βήμα ΔΕΔΟΜΕΝΩΝ
- βήματα της διαδικασίας (ουσιαστικά, όλα όσα δεν περιέχονται στο βήμα ΔΕΔΟΜΕΝΩΝ)
- γλώσσα μακροεντολών

Οι Μηχανές Βιβλιοθήκης SAS και οι Υπηρεσίες απομακρυσμένης Βιβλιοθήκης επιτρέπουν την πρόσβαση σε δεδομένα που αποθηκεύονται σε εξωτερικές δομές δεδομένων και σε απομακρυσμένες πλατφόρμες υπολογιστών.

Το βήμα δεδομένων ενός προγράμματος SAS, όπως και άλλες γλώσσες προγραμματισμού τέταρτης γενιάς με γνώμονα τις βάσεις δεδομένων, όπως η SQL ή Focus, προϋποθέτουν μια προεπιλεγμένη δομή αρχείων, και αυτοματοποιεί τη διαδικασία αναγνώρισης των αρχείων του λειτουργικού συστήματος, το άνοιγμα του αρχείου εισόδου, διαβάζοντας την επόμενη εγγραφή, ανοίγοντας το αρχείο εξόδου, γράφοντας την επόμενη εγγραφή, καθώς και το κλείσιμο των αρχείων. Αυτό επιτρέπει στο χρήστη / προγραμματιστή να επικεντρωθεί στις λεπτομέρειες της εργασίας με τα δεδομένα μέσα σε κάθε εγγραφή, στην πραγματικότητα εργάζεται σχεδόν αποκλειστικά μέσα σε έναν σιωπηρό βρόχο προγράμματος που τρέχει για κάθε εγγραφή.

Όλα τα άλλα καθήκοντα επιτυγχάνονται με διαδικασίες οι οποίες εφαρμόζονται στο σύνολο των δεδομένων. Τυπικές εργασίες περιλαμβάνουν την εκτύπωση ή πραγματοποίηση στατιστικών αναλύσεων, και μπορεί απλώς να απαιτούν από το χρήστη / προγραμματιστή να προσδιορίσει το σύνολο δεδομένων. Οι διαδικασίες δεν περιορίζονται σε μία μόνο συμπεριφορά και κατά συνέπεια επιτρέπουν

εκτεταμένη προσαρμογή, που ελέγχεται από τις μίνι-γλώσσες που ορίζονται στο πλαίσιο των διαδικασιών. Η SAS διαθέτει επίσης μια εκτενή διαδικασία SQL, που επιτρέπει σε SQL προγραμματιστές να χρησιμοποιήσουν το σύστημα με λίγη επιπλέον γνώση.

Υπάρχουν επεκτάσεις μακροεντολών προγραμματισμού, που επιτρέπουν την ορθολογική οργάνωση των επαναλαμβανόμενων τμημάτων του προγράμματος. Η σωστή επιτακτική ανάγκη και οι διαδικαστικές δομές προγραμματισμού μπορούν να προσομοιωθούν με τη χρήση του μακροεντολών "ανοικτού κώδικα" ή της Interactive Matrix Language SAS/IML.

Ο μακρο-κώδικας σε ένα πρόγραμμα της SAS, εάν υπάρχει, υφίσταται προεπεξεργασία. Κατά το χρόνο εκτέλεσης, στο βήμα Δεδομένων τα στοιχεία συλλέγονται, γίνονται compile και οι διαδικασίες ερμηνεύονται και τρέχουν με τη σειρά που εμφανίζονται στο πρόγραμμα της SAS. Ένα πρόγραμμα SAS απαιτεί το λογισμικό της SAS για να τρέξει.

Σε σύγκριση με το γενικής χρήσης γλωσσών προγραμματισμού, αυτή η δομή επιτρέπει στο χρήστη / προγραμματιστή να επικεντρωθούν λιγότερο σχετικά με τις τεχνικές λεπτομέρειες των δεδομένων και το πως είναι αποθηκευμένα, και περισσότερο για τις πληροφορίες που περιέχονται στα δεδομένα. Αυτό θολώνει τα όρια μεταξύ χρήστη και προγραμματιστή, γίνεται ελκυστικό για τα άτομα στον τομέα των «επιχειρήσεων» ή στον τομέα της «έρευνας» και λιγότερο στον τομέα της «τεχνολογίας της πληροφορίας», δεδομένου ότι η SAS δεν επιβάλλει σε μια δομή μια διαρθρωμένη και επικεντρωμένη προσέγγιση στα δεδομένα και τη διαχείριση των υποδομών.

Το SAS τρέχει σε IBM mainframes, Unix, Linux, OpenVMS Alpha, και Microsoft Windows. Ο κώδικας είναι "σχεδόν" διαφανής και διακινείται μεταξύ αυτών των περιβαλλόντων. Οι παλαιότερες εκδόσεις υποστήριζαν PC-DOS, Apple Macintosh, VMS, VM / CMS, PrimeOS, Data General AOS και OS / 2.

2.2 KNIME

Το Konstanz Information Mine, είναι μια φιλική προς το χρήστη πλατφόρμα, συνεκτική open source ανάλυσης δεδομένων, αναφοράς και ολοκλήρωσης. Η KNIME ενσωματώνει διάφορα στοιχεία για τη μηχανική μάθηση και εξόρυξη δεδομένων μέσω του concept της των τμηματικών δεδομένων. Η γραφική διεπαφή χρήστη επιτρέπει την εύκολη και γρήγορη συναρμολόγηση των κόμβων για την προ-επεξεργασία δεδομένων (Εξαγωγή, Μετασχηματισμός, Φόρτωση), για την

μοντελοποίηση και ανάλυση των δεδομένων και την απεικόνιση. Η KNIME από το 2006 χρησιμοποιείται στη φαρμακευτική έρευνα, αλλά χρησιμοποιείται και σε άλλους τομείς, [1]όπως CRM ανάλυση των πελατιακών δεδομένων, business intelligence και χρηματοοικονομική ανάλυση των δεδομένων.
Η KNIME επιτρέπει στους χρήστες να δημιουργήσουν οπτικές ροών δεδομένων (ή αγωγούς), που επιλεκτικά εκτελούν ορισμένα ή όλα τα βήματα ανάλυσης, και αργότερα επιθεωρεί τα αποτελέσματα, τα μοντέλα, και τα διαδραστικά views. Το KNIME είναι γραμμένο σε Java και βασίζεται σε Eclipse και κάνει χρήση του μηχανισμού επέκτασής του για να προσθέσετε plugins που παρέχουν πρόσθετη λειτουργικότητα. Η έκδοση του πυρήνα περιλαμβάνει ήδη τις εκατοντάδες ενοτήτων για την ενσωμάτωση δεδομένων (file I /O, κόμβοι βάσεων δεδομένων που υποστηρίζει όλα τα διαδεδομένα συστήματα διαχείρισης βάσεων δεδομένων), μετασχηματισμοί δεδομένων (φίλτρο, μετατροπέας, συνδυαστική), καθώς και τις συνηθέστερα χρησιμοποιούμενες μεθόδους για την ανάλυση δεδομένων και οπτικοποίησης. Με τη δωρεάν επέκταση Report Designer,οι ροές εργασίας της KNIME μπορούν να χρησιμοποιηθούν ως σύνολα δεδομένων για να δημιουργηθεί πρότυπο αναφοράς που μπορεί να εξαχθεί σε έγγραφο μορφής όπως doc, ppt, xls, pdf και άλλα. Άλλες δυνατότητες της KNIME είναι οι εξής:

- Η βασική αρχιτεκτονική KNIME επιτρέπει την επεξεργασία μεγάλου όγκου δεδομένων που περιορίζεται μόνο από το διαθέσιμο χώρο στο σκληρό δίσκο (τα περισσότερα άλλα open source εργαλεία ανάλυσης δεδομένων εργάζονται στην κύρια μνήμη και ως εκ τούτου περιορίζονται από τη διαθέσιμη μνήμη RAM). Π.χ. Η KNIME επιτρέπει την ανάλυση των 300 εκατ. διευθύνσεις πελατών, 20 εκατομμύρια εικόνες των κυττάρων και 10 εκατ. μοριακών δομών.

- Πρόσθετα plugins επιτρέπουν την ενσωμάτωση των μεθόδων για Text Mining, Mining εικόνας, καθώς και ανάλυση χρονολογικών σειρών.Η KNIME ενσωματώνει διάφορα άλλα open-source-έργα, π.χ. αλγόριθμοι μηχανικής μάθησης από Weka, το στατιστικό πακέτο R, καθώς και LibSVM, JFreeChart, ImageJ,και το Chemistry Development Kit.

- Το KNIME είναι υλοποιημένο σε Java, αλλά επίσης επιτρέπει στουσ wrappers να καλούν άλλον κωδικό εκτός από την παροχή στους κόμβους που επιτρέπουν να τρέχει Java, Python, Perl και άλλων κομματιών κώδικα.

2.3 Orange

Το Orange είναι μια component-based σουίτα εξόρυξης δεδομένων και μηχανικής μάθησης λογισμικού, η οποία διαθέτει φιλικό αλλά ισχυρό και ευέλικτο οπτικό προγραμματισμό front-end για διερευνητική ανάλυση των δεδομένων και την απεικόνιση, και συνδέσεις και βιβλιοθήκες Python για δέσμες ενεργειών. Περιλαμβάνει πλήρη σειρά εξαρτημάτων για προ-επεξεργασία δεδομένων, δυνατότητα βαθμολόγησης και φιλτραρίσματος, μοντελοποίηση, αξιολόγηση μοντέλου, και τις τεχνικές εξερεύνησης. Υλοποιείται σε C + + (ταχύτητα) και Python (ευελιξία). Η γραφική του διεπαφή χρήστη βασίζεται στο πλαίσιο Qt cross-platform. Το Orange διανέμεται δωρεάν υπό την GPL. Διατηρείται και αναπτύσσεται στο Εργαστήριο Βιοπληροφορικής του Τμήματος Ηλεκτρονικών Υπολογιστών και Πληροφορικής, στο Πανεπιστήμιο της Ljubljana, στη Σλοβενία.

2.4 Γλώσσα προγραμματισμού R

Η R είναι μια γλώσσα προγραμματισμού και το περιβάλλον λογισμικού για στατιστικούς υπολογισμούς και γραφικά. Η γλώσσα R έχει γίνει το de facto πρότυπο μεταξύ των στατιστικολόγων για την ανάπτυξη στατιστικού λογισμικού [2] [3]και χρησιμοποιείται ευρέως για τη στατιστική ανάπτυξη λογισμικού και την ανάλυση δεδομένων[3].
Η R είναι μια υλοποίηση της γλώσσας προγραμματισμού S σε συνδυασμό με την λεξική σημασιολογία οριοθέτησης του πεδίου εμπνευσμένη από το Scheme. Η S δημιουργήθηκε από τον John Chambers, στα εργαστήρια της Bell. Η R δημιουργήθηκε από τον Ross Ihakaκαι Robert Gentleman[4], στο Πανεπιστήμιο του Όκλαντ της Νέας Ζηλανδίας, και τώρα αναπτύσσεται από την R Development Core Team, της οποίας ο Chambers είναι μέλος. Η R έχει ονομαστεί εν μέρει από τα πρώτα ονόματα των δύο πρώτων συντακτών R (Robert Gentleman και Ross Ihaka), και εν μέρει ως ένα παιχνίδι στο όνομα του S[5].
Η R είναι μέρος του έργου GNU[6][7]. Ο κώδικάς του είναι ελεύθερα διαθέσιμος υπό την GNU General Public License, και pre-compiled δυαδικές εκδόσεις παρέχονται για διάφορα λειτουργικά συστήματα. Η R χρησιμοποιεί διασύνδεση γραμμής εντολών, ωστόσο, αρκετές γραφικές διεπαφές χρήστη είναι διαθέσιμες για χρήση με τη R.

2.5 TANAGRA

Το TANAGRA είναι ένα δωρεάν λογισμικό εξόρυξης δεδομένων. Εξόρυξη δεδομένων,διερευνητική ανάλυση των δεδομένων, μηχανική μάθηση και αλγόριθμοι στατιστικής μάθησης είναι διαθέσιμες. Το TANAGRA μπορεί να χρησιμοποιηθεί σε πειραματισμούς για τις ακαδημαϊκές εκδόσεις ή μελέτες σε εφαρμογές πραγματικού κόσμου. Ο πηγαίος κώδικας του προγράμματος είναι δωρεάν και μπορείτε να το κατεβάσετε. Τα παράγωγα λογισμικά μπορεί να διανεμηθούν ελεύθερα, αλλά πρέπει να μην είναι επίσης εμπορικού χαρακτήρα. Το TANAGRA διανέμεται από τον Δεκέμβριο του 2003. Έχει γίνει compiled για Win32 πλατφόρμα, αλλά είναι δυνατόν να εκτελεστεί και σε άλλα συστήματα (π.χ. στο WINE με LINUX).

2.6 Carrot²

Το Carrot²[8] είναι μια open source μηχανή αναζήτησης αποτελεσμάτων ομαδοποίησης[9]. Μπορεί αυτόματα να ομαδοποιεί μικρές συλλογές εγγράφων, π.χ. αποτελέσματα αναζήτησης ή αποσπάσματα εγγράφων, σε θεματικές κατηγορίες. Εκτός από δύο εξειδικευμένα αποτελέσματα αναζήτησης αλγορίθμων ομαδοποίησης, το Carrot² προσφέρει έτοιμα προς χρήση στοιχεία για αποτελέσματα αναζήτησης από διάφορες πηγές. Το Carrot² είναι γραμμένο σε Java και διανέμεται υπό την άδεια της BSD.

2.7 RapidMiner

Το RapidMiner, πρώην YALE (Yet Another Learning Environment), είναι ένα περιβάλλον για τη μηχανική μάθηση, την εξόρυξη δεδομένων, την εξόρυξη κειμένου, την προγνωστική ανάλυση και των επιχειρηματικών analytics. Χρησιμοποιείται για την έρευνα, την εκπαίδευση, την κατάρτιση,την ταχεία εκπόνηση πρωτοτύπων, την ανάπτυξη εφαρμογών, και τις βιομηχανικές εφαρμογές. Σε μια δημοσκόπηση της KDnuggets, μιας εφημερίδας data-mining, το RapidMiner καταλαμβάνει τη δεύτερη θέση στην εξόρυξη δεδομένων / εργαλεία ανάλυσης που χρησιμοποιούνται για έργα που εκτελέστηκαν το 2009[10] και ήταν πρώτο το 2010[11]. Διανέμεται υπό την AGPL open source και έχει φιλοξενηθεί από το SourceForge το 2004.
Το έργο του RapidMiner ξεκίνησε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa, και Simon Fischer στην Artificial Intelligence Unit του Πανεπιστημίου του

Dortmund. Το 2006 ο Ingo Mierswa και Ralf Klinkenberg ίδρυσαν την εταιρία Rapid-I που είναι τώρα το κύριο αίτιο για την περαιτέρω ανάπτυξη του RapidMiner καθώς και περισσότεροι από 30 προγραμματιστές σε όλον τον κόσμο.

2.9 GATE - General Architecture for Text Engineering

Η Γενική Αρχιτεκτονική Μηχανικής Κειμένου ή GATE είναι μια σουίτα εργαλείων Java που αναπτύχθηκε αρχικά στο Πανεπιστήμιο του Σέφιλντ άαρχισε το 1995 και χρησιμοποιείται τώρα σε όλο τον κόσμο από μια μεγάλη κοινότητα επιστημόνων, εταιριών, εκπαιδευτικούς και μαθητές για όλα τα είδη των εργασιών επεξεργασίας φυσικής γλώσσας, συμπεριλαμβανομένης της εξόρυξης των πληροφοριών σε πολλές γλώσσες.
Η GATE περιλαμβάνει[12]:

- ένα IDE, GATE Developer: ένα ολοκληρωμένο περιβάλλον ανάπτυξης για συνιστώσες της φυσικής γλώσσας συνδιασμένο με ένα ευρέως χρησιμοποιούμενο σύστημα εξαγωγής πληροφορίας και μια πλήρη σειρά άλλων plugins.

- ένα WEb app, GATE Teamware: ένα συνεργατικό περιβάλλον σχολιασμού για το factory-style σημασιολογικό έργο σχολιασμού χτισμένο γύρω από μια μηχανή ροής εργασίας και μια βαριά-βελτιστοποιημένη backend υποδομή παροχής υπηρεσιών.

- ένα πλαίσιο, GATE Embedded: μια βιβλιοθήκη αντικειμένων βελτιστοποιημένη για συμπερίληψη σε διάφορες εφαρμογές που παρέχει πρόσβαση σε όλες τις υπηρεσίες που χρησιμοποιούνται από το GATE GATE Developer και περισσότερα.

- μια αρχιτεκτονική: μια υψηλού επιπέδου οργανωτική εικόνα γλώσσας σύνθεσης λογισμικού επεξεργασίας.

- μια διαδικασία για τη δημιουργία ισχυρών και διατηρήσιμων υπηρεσιών.

Υπό ανάπτυξη:

- ένα wiki / CMS [13]

- GATE Cloud, μια λύση για το cloud computing που φιλοξενεί μεγάλης κλίμακας επεξεργασία κειμένου.

Η GATE στοχεύει στην εξάλειψη της ανάγκης για την επίλυση κοινών προβλημάτων μηχανικής πριν γίνει η έρευνα στον τομέα, και την αναδιοργάνωσή τους, πριν την ανάπτυξη των αποτελεσμάτων της έρευνας σε εφαρμογές. Οι βασικές λειτουργίες της GATE φροντίσει τη μερίδα του λέοντος της μηχανικής:

- μοντελοποίηση και επιμονή των εξειδικευμένων δομών δεδομένων
- μέτρηση, αξιολόγηση και συγκριτική αξιολόγηση
- οπτικοποίηση και επεξεργασία των σχολιασμών, οντολογίες, ανάλυση δέντρων, κλπ.
- μια περιορισμένου εύρους γλώσσα μεταγωγής για ταχεία προτυποποίηση και αποτελεσματική εφαρμογή των ρηχών μεθόδων ανάλυσης
- εξόρυξη δειγμάτων κατάρτισης για μηχανική μάθηση
- pluggable εφαρμογές μηχανικής μάθησης (Weka, SVM Light, μια εσωτερική άνιση εφαρμογή περιθωρίων SVM[14] και άλλα.)

Πέρα από τις βασικές λειτουργίες, η GATE περιλαμβάνει στοιχεία για ποικίλες εργασίες επεξεργασίας φυσικής γλώσσας, π.χ. parsers, μορφολογία, τοποθέτηση πινακίδων, εργαλεία ανάκτησης πληροφοριών, στοιχείων εξαγωγής πληροφοριών για διάφορες γλώσσες, και πολλά άλλα. Έχει εφαρμοστεί ευρέως σε τομείς όπως η βιοπληροφορική[15] και άλλους. Το GATE Developer και Embedded παρέχονται με το σύστημα εξαγωγής πληροφορίας (Annie), η οποία έχει προσαρμοστεί και αξιολογηθεί σε πολύ μεγάλο βαθμό (πολλά βιομηχανικά συστήματα, συστημάτων έρευνας σε MUC, TREC, ACE, DUC, Pascal, NTCIR, κ.λπ.). Το ANNIE χρησιμοποιείται συχνά για τη δημιουργία RDF ή OWL (μεταδεδομένα) για αδόμητο περιεχόμενο (σημασιολογικής αποτύπωσης). Η GATE έχει συγκριθεί με το NLTK, την R και το RapidMiner[16]. Καθώς χρησιμοποιείται ευρέως από μόνη της, αποτελεί και τη βάση της σημασιολογικής πλατφόρμας KIM[17].

Η GATE κοινότητα και η έρευνα συμμετέχει σε πολλά ευρωπαϊκά ερευνητικά έργα, συμπεριλαμβανομένων των TAO, Sekt, NeOn, Media-Campaign, Musing, Service-Finde, LIRICS και KnowledgeWeb, καθώς και πολλά άλλα έργα.

2.10 jHepWork

Το jHepWork είναι ένα ελεύθερο πλαίσιο ανάλυσης δεδομένων για επιστήμονες, μηχανικούς και φοιτητές γραμμένο σε Java. Το πρόγραμμα είναι σχεδιασμένο για διαδραστικά επιστημονικά τμήματα σε 2D και 3D και περιέχει αριθμητικές επιστημονικές βιβλιοθήκες που εφαρμόζεται στην Java για μαθηματικές λειτουργίες, τυχαίους αριθμούς, στατιστικές αναλύσεις, καμπύλες και άλλους αλγορίθμους εξόρυξης δεδομένων. Το jHepWork βασίζεται σε μια υψηλού επιπέδου γλώσσα προγραμματισμού τη Jython (Python που εφαρμόζεται στην Java), αλλά η κωδικοποίηση Java μπορεί επίσης να χρησιμοποιηθεί για την κλήση αριθμητικών και γραφικών βιβλιοθηκών του jHepWork.
Το jHepWork είναι μια προσπάθεια να δημιουργίας ενός περιβάλλοντος ανάλυσης δεδομένων που χρησιμοποιούν ανοικτού κώδικα πακέτα με μια συνεπή διεπαφή χρήστη για να δημιουργήσει ένα ανταγωνιστικό εργαλείο στα εμπορικά προγράμματα. Η ιδέα πίσω από το σχέδιο είναι να ενσωματωθούν ανοικτού κώδικα μαθηματικά και αριθμητικά πακέτα λογισμικού με τύπους GUI των διεπαφών σε ένα συνεκτικό πρόγραμμα στο οποίο η κύρια διεπαφή χρήστη βασίζεται σε σύντομα ονόματα κλάσεων Java / Python. Αυτό ήταν απαραίτητο για να δημιουργηθεί ένα περιβάλλον ανάλυσης με τη χρήση Java scripting.
Το HepWork τρέχει σε οποιαδήποτε πλατφόρμα (Windows, Mac και Linux, κ.λπ.) όπου η Java μπορεί να εγκατασταθεί. Scripts και κώδικας Java (στην περίπτωση του προγραμματισμού με Java) μπορεί να λειτουργήσει είτε σε ένα πρόγραμμα επεξεργασίας γραφικών GUI του jHepWork είτε ως προγράμματα batch. Η γραφικές βιβλιοθήκες του jHepWork μπορούν να χρησιμοποιηθούν για τη δημιουργία applets. Όλα τα γραφήματα (ή "Καμβάδες") που χρησιμοποιούνται για την αναπαράσταση δεδομένων μπορούν να ενσωματωθούν σε Web browsers.
Το jHepWork μπορεί να χρησιμοποιηθεί παντού όπου είναι απαραίτητη μια ανάλυση μεγάλου όγκου αριθμητικών δεδομένων, εξόρυξη δεδομένων, στατιστική ανάλυση δεδομένων και μαθηματικών. Το πρόγραμμα μπορεί να χρησιμοποιηθεί στις φυσικές επιστήμες, τη μηχανική, τη μοντελοποίηση και ανάλυση των χρηματοοικονομικών αγορών. Το jHepWork θεωρείται μεταξύ των πέντε καλύτερων ελεύθερων λογισμικών ανοικτού κώδικα εξόρυξης δεδομένων[18]. Υπάρχουν πολλές άλλες κριτικές διαθέσιμες για το jHepWork[19][20]. Παρόλο που το πρόγραμμα ανήκει στην κατηγορία λογισμικού ανοικτού κώδικα, δεν είναι εντελώς δωρεάν για εμπορική χρήση.

ΕΠΙΛΟΓΟΣ

Όπως είδαμε, στην αγορά λογισμικού εξόρυξης δεδομένων υπάρχει μια πληθώρα ποικιλία εργαλείων για κάθε σκοπό, επιχείρηση, εκπαιδευτική κοινότητα ή απλό χρήστη τόσο σε εφαρμογές ανοιχτού κώδικα όσο και σε εμπορικές σουίτες εξόρυξης. Όλο και περισσότερα πανεπιστήμια ασχολούνται με την ανάπτυξη εργαλείων στον τομέα της έρευνας της μηχανικής μάθησης και της ανάκτησης πληροφοριών. Στη συνέχεια αναπτύσσονται τα τρία (3) πιο γνωστά και πολυεφαρμοσμένα περιβάλλοντα λογισμικού εξόρυξης δεδομένων για τα οποία θα γίνει λεπτομερείς ανάλυση και σύγκριση του τρόπου που επιτυγχάνουν την ανάλυση δεδομένων.

ΚΕΦΑΛΑΙΟ 3

Προγράμματα Εξόρυξης Δεδομένων (Intelligent Miner-Microsoft Analysis Services-Weka)

ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό γίνεται αναλυτική παρουσίαση των πιο γνωστών εφαρμογών εξόρυξης δεδομένων, του IBM Intelligent Miner, του Microsoft SQL Server Analysis Services και του Weka.

[1]Ο IBM Intelligent Miner είναι πρωτοπόρος στην παροχή βοήθειας για την αναγνώριση και να εξαγωγή υψηλής αξίας επιχειρηματικής ευφυΐας από τα δεδομένα. Η διαδικασία είναι μία ανακάλυψη. Οι εταιρείες έχουν το δικαίωμα να αναμοχλεύσουν πληροφορίες κρυμμένες μέσα σε δεδομένα για τις επιχειρήσεις και να ανακαλύψουν συσχετίσεις, πρότυπα και τάσεις, να ανιχνεύσουν αποκλίσεις, να ομαδοποιήσουν και να ταξινομήσουν πληροφορίες και να αναπτύξουν μοντέλα πρόβλεψης.

Το Microsoft SQL Server Analysis Services αποτελεί μέρος του Microsoft SQL Server , ένα σύστημα διαχείρισης βάσης δεδομένων. Η Microsoft έχει συμπεριλάβει μια σειρά υπηρεσιών στον SQL Server που σχετίζονται με την επιχειρηματική ευφυΐα και το data warehousing . Οι υπηρεσίες αυτές περιλαμβάνουν το Integration Services και το Analysis Services. Το Analysis Services περιλαμβάνει μια ομάδα OLAP και δυνατότητες εξόρυξης δεδομένων.

Το Weka είναι μια δημοφιλής σουίτα λογισμικού της μηχανικής μάθησης γραμμένο σε Java, που αναπτύχθηκε στο Πανεπιστήμιο του Waikato ,της Νέα Ζηλανδία . Το Weka είναι ελεύθερο λογισμικό διαθέσιμο υπό την GNU General Public License.

3.1 Intelligent Miner

3.1.1 Ιστορία

Ο βραβευμένος Intelligent Miner της IBM, κυκλοφόρησε το 1996. Επιτρέπει στους χρήστες να εξορύξουν δομημένα δεδομένα που αποθηκεύονται σε βάσεις δεδομένων ή συμβατικά επίπεδα αρχεία. Οι πελάτες και οι εταίροι έχουν επιτυχή ανάπτυξη αλγορίθμων εξόρυξης για την αντιμετώπιση τέτοιων επιχειρηματικών τομών όπως η ανάλυση αγοράς, η απάτη και η κατάχρηση, και η διαχείριση σχέσεων πελατών.

3.1.2 Ενδεδειγμένοι Πελάτες

Οι προσφορές Intelligent Miner προορίζονται για χρήση από τους αναλυτές δεδομένων και Τεχνολόγους επιχειρήσεων σε τομείς όπως το μάρκετινγκ, τη χρηματοδότηση,τη διαχείριση προϊόντων, καθώς και τη διαχείριση πελατειακών σχέσεων. Επιπλέον, οι τεχνολογίες εξόρυξης κείμενου έχουν εφαρμογή σε ένα ευρύ φάσμα χρηστών που τακτικά επανεξετάζουν ή ερευνούν έγγραφα - για παράδειγμα, οι δικηγόροι διπλωμάτων ευρεσιτεχνίας, εταιρικοί βιβλιοθηκάριοι, ομάδες δημόσιων σχέσεων, ερευνητές και φοιτητές.

3.1.3 Ποιος είναι ο Intelligent Miner

Ο Intelligent Miner της IBM είναι μια σουίτα στατιστικών, επεξεργαστικών, και λειτουργιών εξόρυξης που μπορούν να χρησιμοποιηθούν για την ανάλυση μεγάλων βάσεων δεδομένων. Παρέχει επίσης εργαλεία οπτικοποίησης για την προβολή και την ερμηνεία των αποτελεσμάτων της εξόρυξης. Το λογισμικό του Server τρέχει σε AIX, AS/400, OS/390, και Sun Solaris λειτουργικά συστήματα. Τα AIX, OS / 2, και Windows μπορούν να χρησιμοποιηθούν για τους πελάτες.

Μερικές από τις δυνατότητες που παρέχονται από τον πρόγραμμα Intelligent Miner είναι:

- Επέκταση συσχετίσεων, ταξινόμηση, ομαδοποίηση, και μεθόδους πρόβλεψης
- Νευρωνική πρόβλεψη
- Στατιστικές Λειτουργίες
- Εξαγωγή και εισαγωγή βάσεων εξόρυξης από διαφορετικά λειτουργικά συστήματα
- Αξιοποίηση του DB2 Parallel Edition and DB2 Universal Database Enterprise Extended Edition
- Επαναληπτικές ακολουθίες
- API για όλες τις πλατφόρμες server

Ο Intelligent Miner παρέχει ένα πλήρες γραφικό περιβάλλον εργασίας χρήστη με TaskGuides που οδηγούν μέσα από βήματα δημιουργίας διαφόρων αντικειμένων του Intelligent Miner. Γενική βοήθεια για κάθε TaskGuide παρέχει πρόσθετες πληροφορίες, παραδείγματα, και έγκυρες τιμές για τους ελέγχους σε κάθε σελίδα.

3.1.4 Data Mining με τον Intelligent Miner

Η εξόρυξη δεδομένων είναι η διαδικασία της ανακάλυψης έγκυρων, μέχρι τώρα άγνωστων, και τελικά κατανοητών πληροφοριών από μεγάλες αποθήκες δεδομένων. Μπορεί να χρησιμοποιηθεί για την εξαγωγή πληροφοριών για τον σχηματισμό ενός μοντέλου πρόβλεψης ή την κατάταξη ή να εντοπιστούν ομοιότητες μεταξύ των εγγραφών βάσης δεδομένων. Οι πληροφορίες που προκύπτουν μπορούν να βοηθήσουν στη λήψη πιο σωστών αποφάσεων.
Ο Intelligent Miner βοηθά τους οργανισμούς να εκτελούν εργασίες εξόρυξης δεδομένων. Για παράδειγμα, ένα κατάστημα λιανικής θα μπορούσε να χρησιμοποιήσει τον Intelligent Miner για να προσδιορίσει τις ομάδες πελατών που είναι πιο πιθανό να ανταποκριθούν σε νέα προϊόντα και υπηρεσίες ή να εντοπίσει νέες ευκαιρίες για cross selling. Μια ασφαλιστική εταιρεία μπορεί να χρησιμοποιήσει τον Intelligent Miner για την απομόνωση πιθανής απάτης.

3.1.5 Επισκόπηση του Intelligent Miner

Ο Intelligent Miner συνδέει τις λειτουργίες εξόρυξης και επεξεργασίας στο server με τα εργαλεία διοίκησης και οπτικοποίησης στον client. Τα περιεχόμενα του client περιλαμβάνουν ένα περιβάλλον χρήστη από την οποίο μπορεί να επικαλεστούν η εξόρυξη και οι λειτουργίες επεξεργασίας σε έναν server Intelligent Miner. Τα αποτελέσματα της διαδικασίας εξόρυξης μπορούν να επιστραφούν στον πελάτη, όπου μπορεί να τα απεικονίσει και να τα αναλύσει.
Τα περιεχόμενα του client είναι διαθέσιμα για AIX, OS / 2, Windows NT και Windows 95 λειτουργικά συστήματα. Τα περιεχόμενα του server είναι διαθέσιμα για AIX, OS/390, AS/400 και συστήματα Sun Solaris. Είναι επίσης διαθέσιμα για RS/6000 SP και εκμεταλλεύονται παράλληλα εξόρυξης σε πολλούς κόμβους επεξεργασίας. Μπορεί επίσης να υπάρχει ο client και ο server στο ίδιο μηχάνημα.

3.1.6 Συναρτήσεις εξόρυξης και στατιστικής

Τα αντικείμενα ρυθμίσεων εξόρυξης και στατιστικής είναι τα ίδια με εκείνα που αντιπροσωπεύουν αναλυτικές λειτουργίες που ερχόταν σε αντίθεση με τα δεδομένα. Και στις δυο περιπτώσεις, θα πρέπει να αναφερθεί ποια αντικείμενα δεδομένων θα χρησιμοποιηθούν.
Τα αντικείμενα ρυθμίσεων εξόρυξης και στατιστικής παράγουν ως αποτελέσματα αντικείμενο, όταν τρέχουν. Τα αντικείμενα αποτελέσματα μπορούν να προβληθούν με εργαλεία οπτικοποίησης. Επίσης μπορεί να αναφερθεί στις ρυθμίσεις για τις λειτουργίες αυτές ότι θέλουμε να δημιουργήσουμε δεδομένα εξόδου αντί αντικείμενο αποτελεσμάτων .

Ο Intelligent Miner έχει πολλά είδη λειτουργιών εξόρυξης και στατιστικής :

Πίνακας 1: Λειτουργίες Εξόρυξης και Στατιστικής

Mining	**Statistics**
Assosiations	Cross-correlation
Clustering – demographic	Correlation matrixes
Clustering – neural	Factor analysis
Sequential patterns	Linear regression
Time sequence	Principal component analysis
Classification – tree	Univariate curve fitting
Classification – neural	Bivariate statistics
Prediction – Radial-Basis-Function	
Prediction – neural	

3.1.7 Λειτουργίες επεξεργασίας

Οι λειτουργίες επεξεργασίας χρησιμοποιούνται για να κάνουν τα δεδομένα κατάλληλα για την εξόρυξη ή την ανάλυση. Οι ρυθμίσεις επεξεργασίας αντικειμένων ισχύουν μόνο για τους πίνακες της βάσης δεδομένων και τα Views, διότι, επωφελούμενοι από την επεξεργαστική ικανότητα του μηχανισμού διαχείρισης βάσεων δεδομένων.

Ο Intelligent Miner έχει πολλές λειτουργίες επεξεργασίας:

Πίνακας 2: Λειτουργίες Επεξεργασίας

Aggregate values	Filter fields
Calculate values	Filter records
Clean up input data or output data	Filter records using a value set
Convert to lowercase or uppercase	Get random sample
Copy records to file	Group records
Discard records with missing values	Join data sources
Discretization into quantiles	Map values
Discretization using ranges	Pivot fields to records
Encode missing values	Run SQL
Encode nonvalid values	

Τα αντικείμενα ρυθμίσεων επεξεργασίας διαβάζουν πάντα την είσοδο από μια βάση δεδομένων και δημιουργούν τα δεδομένα εξόδου σε μια βάση δεδομένων. Η μόνη εξαίρεση είναι η λειτουργία Copy Records to File, η οποία αντιγράφει τα δεδομένα σε ένα αρχείο. Όταν δημιουργείτε ένα αντικείμενο ρυθμίσεων επεξεργασίας ή
ενημερώνεται ένα υπάρχον, μπορείτε να χρησιμοποιήσετε τα αντικείμενα ρυθμίσεων επεξεργασίας για τον προσδιορισμό των δεδομένων εισόδου ή δεδομένων εξόδου. Με αυτό τον τρόπο το όνομα ενός πίνακα βάσης δεδομένων ή η προβολή αντιγράφεται στο αντικείμενο ρυθμίσεων επεξεργασίας. Οι επόμενες αλλαγές στις ρυθμίσεις των δεδομένων αντικειμένου δεν επηρεάζουν το αντικείμενο ρυθμίσεων επεξεργασίας.

3.1.8 Λειτουργίες

Το πώς τα αντικείμενα αποτελεσμάτων χρησιμοποιούνται με τον Intelligent Miner εξαρτάται από το ποια λειτουργία υλοποιείται. Ο Intelligent Miner παρέχει τις ακόλουθες λειτουργίες στο πλαίσιο εκτέλεσης της διαδικασίας εξόρυξης:

Εκπαίδευση

Στη λειτουργία της εκπαίδευσης, μια λειτουργία mining χτίζει ένα μοντέλο με βάση τα επιλεγμένα δεδομένα εισόδου.

Clustering
Στην ομαδοποίηση, οι λειτουργίες χτίζουν ένα μοντέλο στη βάση των επιλεγμένων δεδομένων εισόδου. Η λειτουργία ομαδοποίησης είναι παρόμοια με τη λειτουργία εκπαίδευσης για τους έξυπνους αλγόριθμους. Ομαδοποίηση προσφέρει τη δυνατότητα επιλογής από το χρήση στατιστικού υποβάθρου από τα δεδομένα εισόδου ή αποτέλεσμα των εισροών.

Δοκιμής
Στη δοκιμαστική λειτουργία, μια λειτουργία mining χρησιμοποιεί νέα ή τα ίδια δεδομένα με
γνωστά αποτελέσματα για την επαλήθευση ότι το μοντέλο που δημιουργήθηκε στον τομέα της εκπαιδευτικής λειτουργίας παράγει σταθερά αποτελέσματα. Τα αποτελέσματα των αντικειμένων που χρησιμοποιούνται για την είσοδο και δημιουργήθηκαν ως έξοδο.

Εφαρμογή
Στην κατάσταση εφαρμογής, μια συνάρτηση εξόρυξης χρησιμοποιεί ένα μοντέλο που δημιουργήθηκε κατά την εκπαίδευση για να προβλέψει συγκεκριμένο πεδίο για κάθε εγγραφή στα νέα δεδομένα εισόδου. Η μορφή των δεδομένων πρέπει να είναι πανομοιότυπη με εκείνη που χρησιμοποιείται για τη δημιουργία του μοντέλου.

3.2 Microsoft Analysis Services

3.2.1 Ιστορία

Το 1996 , η Microsoft ξεκίνησε την καριέρα της στην επιχείρηση Server OLAP με την εξαγορά της τεχνολογίας λογισμικού OLAP από το ισραηλινό Panorama Software[1]. Λίγο περισσότερο από δύο χρόνια αργότερα, το 1998, η Microsoft κυκλοφόρησε το OLAP Services, στο πλαίσιο του SQL Server 7. Το OLAP Services υποστήριξε MOLAP , ROLAP και HOLAP αρχιτεκτονικές, και το χρησιμοποιούσε το OLEDB για OLAP πελατειακή πρόσβαση API και MDX ως

query language.. Θα μπορούσε να δουλέψει με client-server mode ή λειτουργία χωρίς σύνδεση με τα τοπικά αρχεία κύβου[2].

Το 2000, η Microsoft κυκλοφόρησε Analysis Services 2000. Μετονομάστηκε από "OLAP Services" λόγω της συμπερίληψης των υπηρεσιών εξόρυξης δεδομένων. Το Analysis Services 2000 θεωρήθηκε μια εξελικτική κυκλοφορία, καθώς χτίστηκε στην ίδια αρχιτεκτονική με τις υπηρεσίες OLAP και, συνεπώς, ήταν συμβατό με αυτό. Στις σημαντικές βελτιώσεις περιλαμβάνονται η μεγαλύτερη ευελιξία στο σχεδιασμό διαστάσεων με την υποστήριξη των διαστάσεων παιδιών γονέα, η αλλαγή διαστάσεων, και οι εικονικές διαστάσεις. Ένα άλλο χαρακτηριστικό ήταν μια πολύ βελτιωμένη μηχανή υπολογισμού με υποστήριξη για μοναδιαίους φορείς, custom συλλογές, και υπολογισμούς κυττάρων. Άλλα χαρακτηριστικά ήταν η ασφάλεια διαστάσεων, η διακριτή μέτρηση, η συνδεσιμότητα μέσω HTTP, οι κύβοι συνόδου, τα επίπεδα ομαδοποίησης, και πολλά άλλα[3].

Το 2005 , η Microsoft κυκλοφόρησε την επόμενη γενιά των OLAP και της τεχνολογίας εξόρυξης δεδομένων Analysis Services 2005. Διατήρησε τη συμβατότητα προς τα πίσω με το επίπεδο API: αν και οι εφαρμογές που ήταν γραμμένες με OLE DB για OLAP και MDX συνέχισαν να εργάζονται, η αρχιτεκτονική του προϊόντος ήταν εντελώς διαφορετική. Η μεγάλη αλλαγή ήρθε με το υπόδειγμα της μορφής UDM - Unified Dimensional Model[4].

Το Microsoft SQL Server Analysis Services αποτελεί μέρος του Microsoft SQL Server , μια βάση δεδομένων του συστήματος διαχείρισης. Η Microsoft έχει συμπεριλάβει μια σειρά υπηρεσιών στον SQL Server που σχετίζονται με επιχειρηματική ευφυΐα και data warehousing. Οι υπηρεσίες αυτές περιλαμβάνουν Integration Services και υπηρεσίες ανάλυσης. Το Analysis Services περιλαμβάνει μια ομάδα OLAP και δυνατότητες εξόρυξης δεδομένων.

3.2.2 Λειτουργίες Αποθήκευσης

Η Microsoft Analysis Services υιοθετεί ουδέτερη θέση στην αντιπαράθεση MOLAP εναντίον ROLAP μεταξύ των OLAP προϊόντων. Επιτρέπει σε όλες τις γεύσεις της MOLAP , ROLAP και HOLAP να χρησιμοποιηθούν εντός του ίδιου μοντέλου.

Λειτουργίες τμηματικής αποθήκευσης

- MOLAP - Πολυδιάστατοι OLAP - Και τα δύο στοιχεία, πραγματικά δεδομένα και συναθροίσεις επεξεργάζονται, αποθηκεύονται, απαριθμίζονται χρησιμοποιώντας μια ειδική μορφή βελτιστοποιημένη για πολυδιάστατα δεδομένα.

- ROLAP - Relational OLAP - Και τα δύο στοιχεία, πραγματικά δεδομένα και συγκεντρώσεις παραμένουν στην σχεσιακή προέλευση δεδομένων, καταργώντας την ανάγκη για ειδική επεξεργασία.
- HOLAP - Hybrid OLAP - Αυτή η λειτουργία χρησιμοποιεί τη σχεσιακή προέλευση δεδομένων για να αποθηκεύουν τα πραγματικά δεδομένα, αλλά προ-επεξεργασία, ομαδοποίηση και ευρετηριοποίηση, αποθηκεύουν αυτά σε μια ειδική μορφή, βελτιστοποιημένη για πολυδιάστατα δεδομένα.

Λειτουργίες αποθήκευσης με Διαστάσεις

- MOLAP - ιδιότητες διαστάσεων και ιεραρχίες επεξεργάζοναι και αποθηκεύονται σε ειδική μορφή
- ROLAP - ιδιότητες διαστάσεων που δεν υποβάλλονται σε επεξεργασία και παραμένουν στην σχεσιακή προέλευση δεδομένων.

3.2.3 APIs και Object Models

Η Microsoft Analysis Services υποστηρίζει διαφορετικά σύνολα API και μοντέλα αντικειμένων για διάφορες εργασίες και σε διαφορετικά περιβάλλοντα προγραμματισμού.

Querying

- XML για την Ανάλυση - Το χαμηλότερο επίπεδο API. Μπορεί να χρησιμοποιηθεί από οποιαδήποτε πλατφόρμα και σε οποιαδήποτε γλώσσα που υποστηρίζουν HTTP και XML
- OLE DB για OLAP - Επέκταση OLEDB. Με COM βάση και είναι κατάλληλο για C / C + + προγράμματα για Windows πλατφόρμα.
- ADOMD - Επέκταση της ADO . Με COM που βασίζεται στον αυτοματισμό και είναι κατάλληλο για VB προγράμματα για Windows πλατφόρμα.
- ADOMD.NET - Επέκταση του ADO.NET. Με .NET βάση και είναι κατάλληλο για διαχειριζόμενο κώδικα προγραμμάτων για CLR πλατφόρμες.

Διοίκηση και Διαχείριση

- DSO - Για το 2000. Για COM που βασίζεται στον αυτοματισμό και είναι κατάλληλο για VB προγράμματα για Windows πλατφόρμα.
- AMO - Για το 2005. Για .NET βάση και είναι κατάλληλο για διαχειριζόμενο κώδικα προγραμμάτων για CLR πλατφόρμες.

3.2.4 Αλγόριθμος εξόρυξης δεδομένων (Analysis Services - Data Mining)

[5]Ο αλγόριθμος εξόρυξης δεδομένων είναι ο μηχανισμός που δημιουργεί ένα μοντέλο εξόρυξης δεδομένων. Για τη δημιουργία ενός πρότυπου, ένας αλγόριθμος αναλύει πρώτα ένα σύνολο δεδομένων και αναζητά συγκεκριμένα μοτίβα και τάσεις. Ο αλγόριθμος χρησιμοποιεί τα αποτελέσματα αυτής της ανάλυσης για να καθορίσει τις παραμέτρους του μοντέλου εξόρυξης. Αυτές οι παράμετροι, στη συνέχεια εφαρμόζονται σε ολόκληρο το σύνολο δεδομένων για την εξαγωγή patterns και λεπτομερών στατιστικών.

Το μοντέλο εξόρυξης που δημιουργεί ένας αλγόριθμος μπορεί να λάβει διάφορες μορφές, όπως:

- Ένα σύνολο κανόνων που περιγράφουν τον τρόπο που τα προϊόντα συγκεντρώνονται σε μια συναλλαγή.
- Ένα δέντρο απόφασης που προβλέπει αν ένας συγκεκριμένος πελάτης θα αγοράσει ένα προϊόν.
- Ένα μαθηματικό μοντέλο για προβλέψεις πωλήσεων.
- Μια σειρά από συμπλέγματα που περιγράφουν τον τρόπο που σχετίζονται οι περιπτώσεις σε ένα σύνολο δεδομένων.

Το Microsoft SQL Server Analysis Services παρέχει διάφορους αλγορίθμους για χρήση σε προβλήματα data mining. Αυτοί οι αλγόριθμοι είναι ένα υποσύνολο του συνόλου των αλγορίθμων που μπορούν να χρησιμοποιηθούν για την εξόρυξη δεδομένων. Μπορούν επίσης να χρησιμοποιηθούν τρίτοι αλγόριθμοι που συμμορφώνονται με τα OLE DB για τις προδιαγραφές του Data Mining.

3.2.5 Τύποι αλγορίθμων εξόρυξης δεδομένων

Το Analysis Services περιλαμβάνει τους ακόλουθους τύπους αλγόριθμων:

- αλγόριθμοι ταξινόμησης που προβλέπουν μία ή περισσότερες διακριτές μεταβλητές, με βάση τα άλλα χαρακτηριστικά στο σύνολο δεδομένων. Ένα παράδειγμα ενός τέτοιου αλγορίθμου κατάταξης είναι ο Αλγόριθμος Δένδρων Απόφασης της Microsoft.
- αλγόριθμους παλινδρόμησης που προβλέπουν μία ή περισσότερες συνεχείς μεταβλητές, όπως κέρδη ή ζημίες, βάσει άλλων χαρακτηριστικών

στο σύνολο δεδομένων. Ένα παράδειγμα ενός τέτοιου αλγορίθμου παλινδρόμησης είναι ο Αλγόριθμος Χρονοσειρών της Microsoft.

- αλγόριθμοι Τμηματοποίησης που χωρίζουν τα δεδομένα σε ομάδες, ή ομάδες, των στοιχείων που έχουν παρόμοιες ιδιότητες. Ένα παράδειγμα ενός τέτοιου αλγορίθμου κατάτμησης είναι ο Microsoft Clustering Algorithm.

- αλγόριθμοι συσχετίσεων που βρίσκουν συσχετίσεις μεταξύ των διαφορετικών ιδιοτήτων σε ένα σύνολο δεδομένων. Η πιο κοινή εφαρμογή αυτού του είδους του αλγορίθμου είναι για η δημιουργία κανόνων συσχέτισης, οι οποίοι μπορούν να χρησιμοποιηθούν σε μια ανάλυση καλαθιού αγοράς. Ένα παράδειγμα ενός τέτοιου αλγορίθμου συσχέτισης είναι ο Αλγόριθμος συσχέτισης της Microsoft.

- αλγόριθμοι ανάλυσης αλληλουχίας που συνοψίζουν συχνές ακολουθίες ή επεισόδια σε δεδομένα, όπως μια ροή Web μονοπατιού. Ένα παράδειγμα ενός τέτοιου αλγορίθμου ανάλυσης αλληλουχίας είναι ο αλγόριθμος αλληλουχίας της Microsoft.

3.2.6 Εφαρμόζοντας τους Αλγόριθμους

Επιλέγοντας τον καλύτερο αλγόριθμο που θα χρησιμοποιηθεί για μια συγκεκριμένη εργασία μπορεί να είναι μια πρόκληση. Ενώ μπορούν να χρησιμοποιηθούν διαφορετικοί αλγόριθμοι για την εκτέλεση της ίδιας εργασίας, κάθε αλγόριθμος παράγει ένα διαφορετικό αποτέλεσμα, και μερικοί αλγόριθμοι μπορούν να παράγουν περισσότερους από έναν τύπους αποτελέσματος. Για παράδειγμα, μπορεί να χρησιμοποιηθεί ο Αλγόριθμος Δένδρων Απόφασης της Microsoft όχι μόνο για την πρόβλεψη, αλλά και ως ένας τρόπος για να μειωθεί ο αριθμός των στηλών σε ένα σύνολο δεδομένων, επειδή το δέντρο αποφάσεων μπορεί να εντοπίσει τις στήλες που δεν επηρεάζουν το τελικό μοντέλο εξόρυξης.

Επίσης, δεν χρειάζεται να χρησιμοποιηθούν αλγόριθμοι ανεξάρτητα. Σε μια ενιαία λύση εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν κάποιοι αλγόριθμοι για την εξερεύνηση των δεδομένων, και στη συνέχεια να χρησιμοποιηθεί άλλος αλγόριθμος για να προβλέψει ένα συγκεκριμένο αποτέλεσμα με βάση αυτά τα δεδομένα. Για παράδειγμα, μπορεί να γίνει χρήση ενός αλγορίθμου, η οποία αναγνωρίζει τα πρότυπα, για να σπάσει τα δεδομένα σε ομάδες που είναι περισσότερο ή λιγότερο ομοιογενής και, στη συνέχεια να χρησιμοποιήσει τα αποτελέσματα για να δημιουργήσει ένα καλύτερο μοντέλο δέντρο απόφασης. Χρησιμοποιούνται πολλοί αλγόριθμοι μέσα σε μία λύση για την εκτέλεση ξεχωριστών εργασιών, για παράδειγμα με τη χρήση ενός αλγορίθμου δέντρου οπισθοδρόμησης για την πρόβλεψη χρηματοοικονομικών πληροφοριών, καθώς και ένα βασισμένο σε κανόνες αλγόριθμο για να εκτελέσει μια ανάλυση καλαθιού αγοράς.

Τα μοντέλα εξόρυξης μπορούν να προβλέψουν τιμές, περιλήψεις παραγωγής των δεδομένων, και να βρουν κρυμμένες συσχετίσεις. Ο ακόλουθος πίνακας παρέχει προτάσεις αλγόριθμων που χρησιμοποιούνται για συγκεκριμένες εργασίες.

Πίνακας 3: Αλγόριθμοι και Εργασίες

Καθήκον	Αλγόριθμοι Microsoft για χρήση
Πρόβλεψη διακριτού χαρακτηριστικού. Για παράδειγμα, η πρόβλεψη του κατά πόσον ο αποδέκτης μιας στοχευμένης αλληλογραφίας θα αγοράσει ένα προϊόν.	• Αλγόριθμος Δένδρων Απόφασεων Microsoft • Αλγόριθμος Naive Bayes Microsoft • Αλγόριθμος Microsoft Clustering • Αλγόριθμος Νευρωνικών Δικτύων Microsoft
Προβλέποντας ένα συνεχές χαρακτηριστικό. Για παράδειγμα, η πρόβλεψη των πωλήσεων του επόμενου έτους.	• Αλγόριθμος Δένδρων Απόφασεων Microsoft • Αλγόριθμος Χρονοσειρών Microsoft
Προβλέποντας μια ακολουθία. Για παράδειγμα, εκτελώντας μια ανάλυση συνδεσμοδιαδρομή για το Web site μιας εταιρείας.	• Αλγόριθμος Ακολουθίας Microsoft
Εύρεση ομάδων κοινών στοιχείων σε συναλλαγές. Για παράδειγμα, χρησιμοποιούν την ανάλυση καλαθιού αγοράς να προτείνουν επιπλέον προϊόντα σε έναν πελάτη για την αγορά.	• Αλγόριθμος σύνδεσης Microsoft • Αλγόριθμος Δένδρων Απόφασεων Microsoft
Εύρεση ομάδων παρόμοιων στοιχείων. Για παράδειγμα, τμηματοποιούν δημογραφικά δεδομένα σε ομάδες για να κατανοηθούν καλύτερα οι σχέσεις μεταξύ των χαρακτηριστικών.	• Αλγόριθμος Microsoft Clustering • Αλγόριθμος Ακολουθίας Microsoft

Επειδή κάθε μοντέλο επιστρέφει ένα διαφορετικό είδος αποτελέσματος, το Analysis Services παρέχει ένα ξεχωριστό πρόγραμμα προβολής για κάθε αλγόριθμο. Όταν εξετάζετε ένα μοντέλο εξόρυξης στο Analysis Services, το πρότυπο εμφανίζεται στην καρτέλα του Mining Model Viewer, του Data Mining Designer το οποίο χρησιμοποιεί το κατάλληλο πρόγραμμα προβολής για το μοντέλο.

3.2.6 Λεπτομέρειες Αλγόριθμων

Ο ακόλουθος πίνακας παρέχει τις συνδέσεις με τους τύπους των διαθέσιμων πληροφοριών για κάθε αλγόριθμο:

- Περιγραφή βασικού αλγόριθμου. Παρέχει μια βασική εξήγηση για το τι κάνει ο αλγόριθμος και πώς λειτουργεί, μαζί με ένα σενάριο εγκατάστασης, όταν ο αλγόριθμος μπορεί να είναι χρήσιμος.

- Τεχνική Ανάλυση. Λίστα αναφοράς με τις παραμέτρους που μπορείτε να ορίσετε για τον έλεγχο της συμπεριφοράς του αλγόριθμου και να προσαρμόσετε τα αποτελέσματα του μοντέλου. Παρέχει πρόσθετες τεχνικές λεπτομέρειες σχετικά με την εφαρμογή του αλγορίθμου, συμβουλές για την απόδοση, και απαιτούμενα στοιχεία.

- Αναζήτηση ενός μοντέλου. Παρέχει παραδείγματα από τα ερωτήματα που μπορείτε να χρησιμοποιήσετε με κάθε τύπο. Μπορείτε να ρωτήσετε ένα μοντέλο για να μάθετε περισσότερα σχετικά με τα πρότυπα του μοντέλου, ή να κάνετε προβλέψεις με βάση αυτά τα πρότυπα.

- Περιεχόμενο μοντέλου εξόρυξης. Περιγράφει το πώς οι πληροφορίες αποθηκεύονται σε μια κοινή δομή για όλα τα είδη μοντέλου, και εξηγεί τον τρόπο ερμηνείας των πληροφοριών. Αφού έχετε δημιουργήσει ένα πρότυπο, μπορείτε να εξερευνήσετε το μοντέλο με τη χρήση των viewers που προβλέπονται στο BI Development Studio, ή μπορείτε να γράψετε ερωτήματα για να επιστρέψει πληροφορίες απευθείας από το περιεχόμενο μοντέλου χρησιμοποιώντας DMX.

Πίνακας 4: Αλγόριθμοι Analysis Services Data Mining

Βασική Περιγραφή Αλγόριθμου	Τεχνική αναφορά	Επερωτήσεις	Περιεχόμενο Μοντέλου Εξόρυξης
Αλγόριθμος σύνδεσης Microsoft	Τεχνική αναφορά Αλγόριθμου σύνδεσης Microsoft	Επερωτήσεις ενός μοντέλου σύνδεσης (Υπηρεσίες Ανάλυσης - Data Mining)	Περιεχόμενο Μοντέλου Εξόρυξης για μοντέλα Σύνδεσης (Υπηρεσίες Ανάλυσης - Data Mining)
Αλγόριθμος Clustering Microsoft	Τεχνική αναφορά Αλγόριθμου Clustering Microsoft	Επερωτήσεις σε ένα μοντέλο Clustering (Υπηρεσίες Ανάλυσης - Εξόρυξη Δεδομένων)	Περιεχόμενο Μοντέλου Εξόρυξης για Μοντέλα Clustering (Υπηρεσίες Ανάλυσης - Data Mining)
Αλγόριθμος Δένδρων Απόφασης Microsoft	Τεχνική αναφορά Αλγόριθμου Δένδρων Απόφασης	Επερωτήσεις Μοντέλου Δένδρων απόφασης(Υπηρεσίες Ανάλυσης - Data Mining)	Περιεχόμενο Μοντέλου Εξόρυξης για Δένδρα Απόφασης(Υπηρεσίες Analysis - Data Mining)

Βασική Περιγραφή Αλγόριθμου	Τεχνική αναφορά	Επερωτήσεις	Περιεχόμενο Μοντέλου Εξόρυξης
	Microsoft		
Γραμμικό αλγόριθμος Παλινδρόμησης Microsoft	Τεχνική αναφορά Γραμμικού αλγόριθμου Παλινδρόμησης Microsoft	Επερωτήσεις σε μια γραμμική ανάλυση παλινδρόμησης (Υπηρεσίες Ανάλυσης - Data Mining)	Περιεχόμενο Μοντέλου Εξόρυξης για Γραμμικά Μοντέλα Παλινδρόμησης (Υπηρεσίες Ανάλυσης - Data Mining)
Logistic Αλγόριθμος Παλινδρόμησης Microsoft	Τεχνική αναφορά Logistic Αλγόριθμου Παλινδρόμησης Microsoft	Επερωτήσεις σε ένα υπόδειγμα της λογιστικής παλινδρόμησης (Υπηρεσίες Ανάλυσης - Data Mining)	Περιεχόμενο Μοντέλου Εξόρυξης για Υπόδειγμα λογιστικής παλινδρόμησης (Υπηρεσίες Ανάλυσης - Data Mining)
Αλγόριθμος Naive Bayes Microsoft	Τεχνική αναφορά Αλγόριθμου Naive Bayes Microsoft	Αναζήτηση σε ένα Naive Bayes μοντέλο (Υπηρεσίες Ανάλυσης - Εξόρυξη Δεδομένων)	Περιεχόμενο Μοντέλου Εξόρυξης για Naive Bayes Μοντέλα (Υπηρεσίες Ανάλυσης - Data Mining)
Αλγόριθμος Νευρωνικών Network Microsoft	Τεχνική αναφορά Αλγόριθμου Νευρωνικών Network Microsoft	Αναζήτηση σε ένα μοντέλο Νευρωνικού δικτύου (Ανάλυση-Data Mining Services)	Περιεχόμενο Μοντέλου Εξόρυξης νευρωνικών δικτύων (Υπηρεσίες Ανάλυσης - Data Mining)
Αλγορίθμος Ακολουθίας Microsoft	Τεχνική αναφορά Αλγόριθμου Ακολουθίας Microsoft	Αναζήτηση σε ένα μοντέλο ακολουθίας Clustering (Υπηρεσίες Ανάλυσης - Εξόρυξη Δεδομένων)	Περιεχόμενο Μοντέλου Εξόρυξης για Ακολουθία Clustering (Υπηρεσίες Ανάλυσης - Data Mining)
Αλγόριθμος Χρονοσειράς Microsoft	Τεχνική αναφορά Αλγόριθμου Χρονοσειράς Microsoft	Επερωτήσεις σε ένα μοντέλο χρονοσειρών (Υπηρεσίες Ανάλυσης - Εξόρυξη Δεδομένων)	Περιεχόμενο Μοντέλου Εξόρυξης για Χρονοσειρά (Υπηρεσίες Ανάλυσης - Data Mining)

3.3 Weka

3.3.1 Περιγραφή

Ο σταθμός εργασίας του Weka[1] περιέχει μια συλλογή από εργαλεία οπτικοποίησης και αλγόριθμους για την ανάλυση δεδομένων και μοντέλα προβλέψεων , σε συνδυασμό με γραφικές διεπαφές χρήστη για την εύκολη πρόσβαση σε αυτή τη λειτουργία. Η αρχική μη Java έκδοση του Weka ήταν TCL /

TK front-end για τη μοντελοποίηση αλγορίθμων που εφαρμόζονται σε άλλες γλώσσες προγραμματισμού, καθώς και τα προεπεξεργασμένα δεδομένα επιχειρήσεων κοινής ωφέλειας σε C, και ένα Makefile με βάση το σύστημα για την εκτέλεση πειραμάτων μηχανικής μάθησης. Αυτή η πρωτότυπη έκδοση είχε αρχικά σχεδιαστεί ως εργαλείο για την ανάλυση των δεδομένων από γεωργικό τομέα, [2] [3] , αλλά η πιο πρόσφατη πλήρως Java-based έκδοση (Weka 3), της οποίας η ανάπτυξη ξεκίνησε το 1997, χρησιμοποιείται σήμερα σε πολλές εφαρμογές διαφορετικών τομέων, ιδίως για εκπαιδευτικούς σκοπούς και την έρευνα. Πλεονεκτήματα του Weka περιλαμβάνουν:

- ελεύθερη διαθεσιμότητα υπό την GNU General Public License
- φορητότητα, δεδομένου ότι εφαρμόζεται πλήρως στη γλώσσα προγραμματισμού Java και έτσι τρέχει σε σχεδόν οποιαδήποτε μοντέρνα πλατφόρμα πληροφορικής
- περιέχει μια ολοκληρωμένη συλλογή προεπεξεργασίας δεδομένων και τεχνικές μοντελοποίησης
- ευκολία στη χρήση, λόγω της γραφικής διεπαφης του

Το Weka υποστηρίζει αρκετά πρότυπο εξόρυξης δεδομένων, πιο συγκεκριμένα, την προ-επεξεργασία δεδομένων, την ομαδοποίηση, την ταξινόμηση , την οπισθοδρόμηση, την απεικόνιση, και την επιλογή των χαρακτηριστικών γνωρισμάτων . Όλες οι τεχνικές του Weka βασιζόταν στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα ενιαίο flat αρχείο ή σχέση, όπου κάθε σημείο δεδομένων περιγράφεται με έναν σταθερό αριθμό χαρακτηριστικών. Το Weka παρέχει πρόσβαση σε SQL βάσεις δεδομένων που χρησιμοποιούν Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφουν από ένα ερώτημα βάσης δεδομένων. Δεν είναι σε θέση να κάνει πολυ-σχεσιακή εξόρυξη δεδομένων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογή από συνδεδεμένους πίνακες βάσεων δεδομένων σε ένα ενιαίο πίνακα που είναι κατάλληλος για τη μεταποίηση που χρησιμοποιεί το Weka [4] . Ένας άλλος σημαντικός τομέας που επί του παρόντος δεν καλύπτεται από τους αλγορίθμους που περιλαμβάνονται στη διανομή του Weka είναι η ακολουθιακής μοντελοποίηση.

Η κύρια διεπαφή χρήστη του Weka είναι το Explorer, αλλά ουσιαστικά η ίδια λειτουργικότητα μπορεί να προσεγγιστεί μέσω του component-based interface και από τη γραμμή εντολών . Υπάρχει επίσης ο πειραματιστής, το οποίο επιτρέπει τη συστηματική σύγκριση των επιδόσεων των πρόβλέψεων των αλγορίθμων του Weka για μια συλλογή συνόλου στοιχείων.

Η διεπαφή Explorer έχει πολλά panels που δίνουν πρόσβαση στα κύρια συστατικά του σταθμού εργασίας:

- Το panel της Προεργασίας έχει δυνατότητες για εισαγωγή δεδομένων από μια βάση, ένα CSV αρχείο, κ.λπ. και για προεπεξεργασία δεδομένων χρησιμοποιώντας το λεγόμενο αλγόριθμο φιλτραρίσματος. Τα φίλτρα αυτά μπορούν να χρησιμοποιηθούν για το μετασχηματισμό των στοιχείων και να καταστεί δυνατή η διαγραφή περιπτώσεων και ιδιότητες σύμφωνα με ειδικά κριτήρια.

- Η ομάδα Ταξινόμησης επιτρέπει στο χρήστη να εφαρμόσει αλγορίθμους ταξινόμησης και παλινδρόμησης προς το σύνολο δεδομένων που προκύπτει, την εκτίμηση της ακρίβειας του μοντέλου προβλέψεων, καθώς και για την οπτικοποίηση λανθασμένων προβλέψεων, ROC καμπύλες , κ.λπ., ή το ίδιο το μοντέλο.

- Το πάνελ Συσχέτισης παρέχει πρόσβαση σε εκπαιδευόμενους κανόνες συσχέτισης που επιχειρούν να προσδιορίσουν όλες τις σημαντικές αλληλεξαρτήσεις μεταξύ των χαρακτηριστικών των δεδομένων.

- Το Cluster δίνει πρόσβαση στις τεχνικές ομαδοποίηση του Weka. Υπάρχει επίσης μια εφαρμογή του αλγορίθμου μεγιστοποίησης προσδοκίας για την κανονική κατανομή .

- Η Επιλογή Χαρακτηριστικού ορίζει αλγορίθμους για τον εντοπισμό των πιο έξυπνων χαρακτηριστικά σε ένα σύνολο δεδομένων.

- Το panel Οπτικοποίησης δείχνει μια γραφική παράσταση πίνακα, όπου τα ατομικά διαγράμματα διασποράς μπορούν να επιλεγούν και να διευρυνθούν, και να αναλυθούν περαιτέρω με τη χρήση διαφόρων φορέων επιλογής.

3.3.2 Ιστορία

- Το 1993, το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία ξεκίνησε την ανάπτυξη της αρχικής έκδοσης του Weka (που ήταν μείγμα TCL / TK, C, και Makefiles).
- Το 1997, η απόφαση λήφθηκε και ξεκίνησε η ανάπλαση του Weka από το μηδέν σε Java, συμπεριλαμβανομένων υλοποιήσης αλγορίθμων μοντελοποίησης. [5]

- Το 2005, το Weka λαμβάνει το βραβείο Data Mining and Knowledge Discovery Service Award[6] [7]
- Το 2006, η Pentaho Corporation απέκτησε αποκλειστική άδεια χρήσης του Weka για την επιχειρηματική ευφυΐα που αποτελεί την εξόρυξη δεδομένων και προβλεπτική συνιστώσα analytics της σουίτας επιχειρηματικών πληροφοριών Pentaho.
- Όλων των εποχών κατάταξη για Sourceforge.net από 2009-06-11, 246 (με 1.566.318 downloads)

ΚΕΦΑΛΑΙΟ 4

Application programming interface – Προγραμματιστική Διεπαφή Εφαρμογών (API)

ΕΙΣΑΓΩΓΗ

Μια Προγραμματιστική Διεπαφή Εφαρμογών (API) είναι ένα συγκεκριμένο σύνολο κανόνων και προδιαγραφών που ακολουθούν τα προγράμματα λογισμικού για να επικοινωνούν μεταξύ τους. Χρησιμεύει ως μια διασύνδεση μεταξύ των διαφόρων προγραμμάτων λογισμικού και διευκολύνει την αλληλεπίδραση τους, παρόμοια με τον τρόπο που το περιβάλλον εργασίας χρήστη διευκολύνει την αλληλεπίδραση μεταξύ ανθρώπων και υπολογιστών.

Ένα API μπορεί να δημιουργηθεί για τις εφαρμογές , τις βιβλιοθήκες , τα λειτουργικά συστήματα , κλπ., ως ένας τρόπος προσδιορισμού των λεξιλογίων και των αιτημάτων συμβάσεων πόρων. Μπορεί να περιλαμβάνει προδιαγραφές για ρουτίνες , δομές δεδομένων , κλάσεις αντικειμένων , καθώς και πρωτόκολλα που χρησιμοποιούνται για την επικοινωνία μεταξύ του προγράμματος για τους καταναλωτές και το πρόγραμμα υλοποίησης του API[1][2].

4.1 Έννοια

Ένα API είναι μια αφαίρεση που περιγράφει μια διεπαφή για την αλληλεπίδραση με μια σειρά από λειτουργίες που χρησιμοποιούνται από συνιστώσες ενός συστήματος λογισμικού . Το λογισμικό που παρέχει τις λειτουργίες που περιγράφονται από ένα API λέγεται ότι είναι μια υλοποίηση του API.

Ένα API μπορεί να είναι:

- Γενικά, το σύνολο των API που είναι συνδυασμένη με τις βιβλιοθήκες μιας γλώσσας προγραμματισμού, π.χ. Standard Βιβλιοθήκη προτύπων σε C ++ ή Java API .
- συγκεκριμένα, ως στόχο να αντιμετωπίσει ένα συγκεκριμένο πρόβλημα, π.χ. το Google Maps API ή Java API για την XML Web Services .
- Εξαρτώμενο από τη γλώσσα, που σημαίνει ότι είναι διαθέσιμο μόνο με χρήση της σύνταξης και τα στοιχεία μιας συγκεκριμένης γλώσσας, η οποία καθιστά το API πιο βολικό στη χρήση.

- Ανεξάρτητο γλώσσας, γραπτό έτσι ώστε να μπορεί να κληθεί από διάφορες γλώσσες προγραμματισμού. Αυτό είναι ένα επιθυμητό χαρακτηριστικό για μια service-oriented API που δεν είναι συνδεδεμένο με μια συγκεκριμένη διαδικασία ή σύστημα και μπορεί να παρέχει κλήσεις απομακρυσμένης διαδικασίας ή των υπηρεσιών web . Για παράδειγμα, μια ιστοσελίδα που επιτρέπει στους χρήστες να εξετάσουν τοπικά εστιατόρια είναι σε θέση κατηγοριοποιήσουν τις κριτικές τους από τους χάρτες που λαμβάνονται από το Google Maps, επειδή τα Google Maps έχουν ένα API που διευκολύνει αυτή τη λειτουργικότητα. Το Google Maps API ελέγχει τα στοιχεία που μια ιστοσελίδα ενός τρίτου μπορεί να χρησιμοποιήσει και πώς να τα χρησιμοποιήσει.

Το API μπορεί να χρησιμοποιηθεί για να αναφερθεί σε ένα ολοκληρωμένο περιβάλλον εργασίας, μία μεμονωμένη λειτουργία ή ακόμη και ένα σύνολο από API που παρέχει ένας οργανισμός. Έτσι, το πεδίο εφαρμογής του συνήθως καθορίζεται από το πλαίσιο της χρήσης.

4.2 Αναλυτική επεξήγηση

Ένα API μπορεί να περιγράψει τον τρόπο με τον οποίο μια συγκεκριμένη εργασία εκτελείται.

Σε γλώσσες διαδικασίας , όπως τη γλώσσα C , η ενέργεια γίνεται συνήθως με τη μεσολάβηση μιας κλήσης συνάρτησης .

Για παράδειγμα: η math.h που περιλαμβάνει αρχείο για τη γλώσσα C , περιέχει τον ορισμό των πρωτοτύπων λειτουργίας των μαθηματικών λειτουργιών που είναι διαθέσιμες στη βιβλιοθήκη της γλώσσας C για μαθηματική επεξεργασία (συνήθως ονομάζεται libm). Αυτό το αρχείο περιγράφει πώς να χρησιμοποιείτε τις λειτουργίες που περιλαμβάνονται στη δεδομένη βιβλιοθήκη: το πρωτότυπο της συνάρτησης είναι μια υπογραφή που περιγράφει τον αριθμό και το είδος των παραμέτρων που πρέπει να περαστούν με τις λειτουργίες και τον τύπο της τιμής επιστροφής.

Η συμπεριφορά των λειτουργιών συνήθως περιγράφεται με περισσότερες λεπτομέρειες σε μια ανθρώπινα αναγνώσιμη μορφή στα τυπωμένα βιβλία ή σε ηλεκτρονική μορφή, όπως η σελίδες man : π.χ. για Unix συστήματα από την εντολή

```
man 3 sqrt
```

θα παρουσιάσει την υπογραφή της συνάρτησης sqrt με τη μορφή:

```
SYNOPSIS
            #include <math.h>
            double sqrt(double X);
            float  sqrtf(float X);
DESCRIPTION
       DESCRIPTION
       sqrt computes the positive square root of the argument. ...
RETURNS
       On success, the square root is returned. If X is real and
positive...
```

Αυτό σημαίνει ότι η συνάρτηση επιστρέφει την τετραγωνική ρίζα ενός θετικού αριθμού κινητής υποδιαστολής (single ή double ακρίβεια) ως έναν άλλον αριθμό κινητής υποδιαστολής. Εξ ου και το API στην περίπτωση αυτή μπορεί να ερμηνευθεί ως η συλλογή του συμπεριλαμβανόμενου αρχείου που χρησιμοποιείται από τη γλώσσα C και η αναγνώσιμη περιγραφή του παρέχεται από τις man pages.

4.3 Το API στις σύγχρονες γλώσσες

Οι περισσότερες από τις σύγχρονες γλώσσες προγραμματισμού παρέχουν τα έγγραφα που σχετίζονται με ένα API σε κάποια ψηφιακή μορφή που το καθιστά εύκολο να γνωμοδοτήσει σε έναν υπολογιστή. Π.χ.

Η perl έρχεται με το εργαλείο perldoc :

```
$ perldoc -f sqrt

       sqrt EXPR
       sqrt    #Return the square root of EXPR.  If EXPR is omitted,
returns
               #square root of $_.  Only works on non-negative operands,
unless
               #you've loaded the standard Math::Complex module.
```

Η python έρχεται με το εργαλείο pydoc :

```
$ pydoc math.sqrt
Help on built-in function sqrt in math:
math.sqrt = sqrt(...)
    sqrt(x)
  Return the square root of x.
```

Η ruby έρχεται με το εργαλείο ri:

```
$ ri Math::sqrt
---------------------------------------------------------- Math::sqrt
     Math.sqrt(numeric)    => float
------------------------------------------------------------------------

   Returns the non-negative square root of _numeric_.
```

Η Java έρχεται με την τεκμηρίωση οργανωμένη σε html σελίδες (JavaDoc μορφή), ενώ η Microsoft διανέμει την τεκμηρίωση API για τις γλώσσες της (Visual C + + , C # , Visual Basic , F # , κλπ. ..), ενσωματωμένα στο σύστημα βοήθειας του Visual Studio.

4.4 API σε αντικειμενοστραφείς γλώσσες

Σε αντικειμενοστραφείς γλώσσες, ένα API συνήθως περιλαμβάνει μια περιγραφή μιας σειράς ορισμών κλάσεων, με μια σειρά από συμπεριφορές που σχετίζονται με αυτές τις κλάσεις. Μια συμπεριφορά είναι το σύνολο των κανόνων για το πώς ένα αντικείμενο, που προέρχεται από αυτή την κλάση, θα δράσει σε μια συγκεκριμένη περίσταση. Αυτή η αφηρημένη έννοια, σχετίζεται με τις λειτουργίες που εκτίθενται πραγματικά, ή διατίθενται, από τις τάξεις που εφαρμόζονται με όρους των μεθόδων κλάσης.

Το API σε αυτή την περίπτωση μπορεί να θεωρηθεί ως το σύνολο όλων των μεθόδων που εκτίθενται στο κοινό από τις κλάσεις (συνήθως ονομάζεται η κλάση interface). Αυτό σημαίνει ότι το API ορίζει τις μεθόδους με τις οποίες κάποιος αλληλεπιδρά με/χειρίζεται τα αντικείμενα που προέρχονται από τους ορισμούς της κλάσης.

Γενικότερα, μπορεί κανείς να δει το API, όπως τη συλλογή όλων των ειδών των αντικειμένων που μπορεί κανείς να αντλήσει από τους ορισμούς της κλάσης και τις δυνατές συνδεόμενες συμπεριφορές τους. Και πάλι: η χρήση γίνεται με τη μεσολάβηση των public μεθόδων, αλλά σε αυτή την ερμηνεία, οι μέθοδοι θεωρούνται ως τεχνική λεπτομέρεια για το πώς εφαρμόζεται η συμπεριφορά.

Για παράδειγμα: μια κλάση που αντιπροσωπεύει ένα Stack (Στοίβα) μπορεί απλά να εκθέσει δημοσίως δύο μεθόδων push() (για να προσθέσει ένα νέο στοιχείο στη στοίβα), και pop() (για να εξαγάγει το τελευταίο στοιχείο, σε ιδανική θέση στην κορυφή της στοίβας).

Στην περίπτωση αυτή, το API μπορεί να ερμηνευθεί ως οι δύο μέθοδοι pop() και push() , ή, γενικότερα, ώς η ιδέα ότι μπορεί κανείς να χρησιμοποιήσει ένα αντικείμενο τύπου Stack που υλοποιεί τη συμπεριφορά μιας στοίβας: ένας σωρός που εκθέτει την κορυφή για την προσθήκη / αφαίρεση στοιχείων.

Η έννοια αυτή μπορεί να πραγματοποιηθεί μέχρι το σημείο όπου μια κλάση διασύνδεσης σε ένα API δεν έχει καμία μεθόδων σε όλα, αλλά μόνον συμπεριφορές που συνδέονται με αυτό. Για παράδειγμα, το API για τη γλώσσα Java και Lisp περιλαμβάνει το interface Serializable , το οποίο απαιτεί από κάθε κλάση που υλοποιεί να συμπεριφέρεται σε serialized μόδα. Αυτό δεν απαιτεί να έχει οποιαδήποτε δημόσια μέθοδο, αλλά απαιτεί κάθε κλάση που το υλοποιεί να έχει μια παράσταση που μπορεί να αποθηκευτεί (συνεχόμενο), ανά πάσα στιγμή (αυτό συνήθως ισχύει για οποιαδήποτε κλάση που περιέχει απλά δεδομένα και δεν συνδέεται με εξωτερικούς πόρους , όπως μια σύνδεση με ένα αρχείο, ένα απομακρυσμένο σύστημα, ή μια εξωτερική συσκευή).

Με αυτή την έννοια, σε αντικειμενοστραφείς γλώσσες, το API ορίζει ένα σύνολο συμπεριφορών αντικειμένου, πιθανώς με τη μεσολάβηση ενός συνόλου μεθόδων κλάσης.

Σε αυτές τις γλώσσες, το API ακόμα διανέμεται ως βιβλιοθήκη. Για παράδειγμα, οι βιβλιοθήκες της γλώσσας Java περιλαμβάνουν ένα σύνολο από API που παρέχονται με τη μορφή του JDK που χρησιμοποιούνται από τους προγραμματιστές για την κατασκευή νέων προγραμμάτων Java. Το JDK περιλαμβάνει την τεκμηρίωση του API στη σημειογραφία JavaDoc.

Η ποιότητα των εγγράφων που σχετίζονται με ένα API είναι συχνά ένας παράγοντας που καθορίζει την επιτυχία του όσον αφορά την ευκολία χρήσης.

4.5 API και πρωτόκολλα

Ένα API μπορεί επίσης να είναι μια υλοποίηση ενός πρωτοκόλλου .

Σε γενικές γραμμές η διαφορά μεταξύ ενός API και ενός πρωτοκόλλου είναι ότι το πρωτόκολλο ορίζει ένα πρότυπο τρόπο στα αιτήματα ανταλλαγής και απαντήσεων με βάση μια κοινή μεταφορά, ενώ ένα API παρέχει μια βιβλιοθήκη που πρέπει να χρησιμοποιείται άμεσα: ως εκ τούτου δεν μπορεί να υπάρξει μεταφορά (δεν υπάρχουν στοιχεία που μεταφέρονται φυσικά από κάποιο απομακρυσμένο μηχάνημα), αλλά μόνο απλή ανταλλαγή πληροφοριών μέσω προσκλήσεων λειτουργίας (τοπικά στο μηχάνημα όπου η επεξεργασία λαμβάνει χώρα).

Όταν ένα API εφαρμόζει ένα πρωτόκολλο μπορεί να βασίζεται σε μεθόδους proxy για απομακρυσμένες επικλήσεις που υπάρχουν κάτω από το πρωτόκολλο επικοινωνίας. Ο ρόλος του API μπορεί να είναι ακριβώς για να κρύψει τις λεπτομέρειες του πρωτοκόλλου μεταφοράς.

Τα πρωτόκολλα συνήθως μοιράζονται μεταξύ διαφορετικών τεχνολογιών (με βάση το σύστημα σε δεδομένη γλώσσα προγραμματισμού ηλεκτρονικών υπολογιστών σε ένα δεδομένο λειτουργικό σύστημα) και συνήθως επιτρέπουν στις διαφορετικές τεχνολογίες να ανταλλάσσουν πληροφορίες, που ενεργούν ως ένα επίπεδο αφαίρεσης / διαμεσολαβητή μεταξύ των δύο κόσμων. Ενώ το API είναι ειδικά για μια δεδομένη τεχνολογία: εξ ου και η API μιας συγκεκριμένης γλώσσας δεν μπορεί να χρησιμοποιηθεί σε άλλες γλώσσες, εκτός αν οι κλίσεις λειτουργίας περιέχουν ειδικές προσαρμοσμένες βιβλιοθήκες.

4.6 Web API

Όταν χρησιμοποιείται στο πλαίσιο της ανάπτυξης ιστοσελίδων , ένα API είναι συνήθως ένα καθορισμένο σύνολο του Hypertext Transfer Protocol (HTTP μηνύματα αιτήσεων), μαζί με έναν ορισμό της δομής των μηνυμάτων απόκρισης, η οποία είναι συνήθως σε Extensible Markup Language (XML) ή JavaScript Object Notation (JSON) μορφή. Ενώ το "Web API" είναι ουσιαστικά ένα συνώνυμο για διαδικτυακή υπηρεσία , η πρόσφατη τάση (τα λεγόμενα Web 2.0) έχει απομακρυνθεί από το Simple Object Access Protocol (SOAP) υπηρεσίες που βασίζονται σε πιο άμεσο Representational State Transfer (REST)[3] στυλ επικοινωνιών. Τα Web API επιτρέπουν το συνδυασμό πολλαπλών υπηρεσιών σε νέες εφαρμογές γνωστές ως mashups[4] .

Χρήση του API για κοινή χρήση περιεχόμενου

Η πρακτική της έκδοσης API επέτρεψε στις διαδικτυακές κοινότητες τη δημιουργία μιας ανοικτής αρχιτεκτονικής για τη διανομή περιεχομένου και δεδομένων μεταξύ των κοινοτήτων και των εφαρμογών. Με αυτόν τον τρόπο, το περιεχόμενο που δημιουργείται σε ένα μέρος μπορεί δυναμικά να δημοσιευτεί και να ενημερωθεί σε διάφορες τοποθεσίες στον παγκόσμιο ιστό.

1. Οι φωτογραφίες μπορούν να μοιραστούν από ιστοσελίδες όπως το Flickr και το Photobucket σε κοινωνικό δικτυακό τόπο όπως το Facebook και το MySpace .

2. Το περιεχόμενο μπορεί να ενσωματωθεί, για παράδειγμα ενσωματώνοντας μια παρουσίαση από SlideShare στο προφίλ του LinkedIn .
3. Το περιεχόμενο μπορεί να δημοσιευτεί δυναμικά . Κοινή χρήση ζωντανών σχολίων που έγιναν στο Twitter με έναν λογαριασμό στο Facebook, για παράδειγμα, όταν έχει ενεργοποιηθεί από API τους.
4. Το περιεχόμενο βίντεο μπορεί να ενσωματωθεί σε ιστοσελίδες που εξυπηρετούνται από ένα άλλο host.
5. Πληροφορίες χρήστη μπορούν να ανταλλάσσονται από διαδικτυακές κοινότητες σε εξωτερικές εφαρμογές, προσφέροντας νέα λειτουργικότητα με την διαδικτυακή κοινότητα να μοιράζεται τα δεδομένα του χρήστη μέσω ενός ανοικτού API. Ένα από τα καλύτερα παραδείγματα αυτού είναι το Facebook Application Platform . Μια άλλη είναι η πλατφόρμα Open Social[5].

4.7 Εφαρμογές

Το POSIX πρότυπο ορίζει ένα API που επιτρέπει σε ένα ευρύ φάσμα των κοινών λειτουργιών πληροφορικής να είναι γραμμένο με τρόπο τέτοιο ώστε να μπορεί να λειτουργεί σε πολλά διαφορετικά συστήματα (Mac OS X , και διάφορα Berkeley Software Distributions (BSDs) να εφαρμόσει αυτό το interface) ωστόσο, για να γίνει χρήση αυτής απαιτείται εκ νέου compiling για κάθε πλατφόρμα. Ένα συμβατό API, από την άλλη πλευρά, επιτρέπει στον compiled object-code να λειτουργήσει χωρίς καμία αλλαγή στο σύστημα εφαρμογής του εν λόγω API. Αυτό είναι προς όφελος τόσο των παρόχων λογισμικού (όπου μπορούν να διανέμουν το υπάρχον λογισμικό για τα νέα συστήματα που δεν παράγουν και διανέμουν αναβαθμίσεις) και των χρηστών (όπου μπορούν να εγκαταστήσουν ένα παλαιότερο λογισμικό σε νέα συστήματα χωρίς την αγορά αναβαθμίσεων), αν και αυτό απαιτεί, γενικώς, ότι διάφορες βιβλιοθήκες λογισμικού εφαρμόσουν τα απαραίτητα API.

Η Microsoft έχει επιδείξει ισχυρή δέσμευση για ένα συμβατό API, ιδίως στο πλαίσιο της βιβλιοθήκης Windows API (Win32), έτσι ώστε παλαιότερες εφαρμογές να μπορούν να τρέχουν σε νεότερες εκδόσεις των Windows χρησιμοποιώντας ένα εκτελέσιμο ειδικής ρύθμισης που ονομάζεται "Compatibility Mode"[6].

Η Apple Inc έχει δείξει λιγότερη ανησυχία, σπάζοντας τη συμβατότητα ή την εφαρμογή μιας API σε μια πιο αργή "κατάσταση εξομοίωσης", επιτρέπει μεγαλύτερη ελευθερία στην ανάπτυξη, στο κόστος του να κάνει τα παλιότερα λογισμικά απαρχαιωμένα.

Μεταξύ των Unix-like λειτουργικών συστημάτων, υπάρχουν πολλά σχετικά αλλά ασυμβίβαστα λειτουργικά συστήματα που εκτελούνται σε μια κοινή πλατφόρμα υλικού (κυρίως Intel 80386 -συμβατών συστημάτων). Έχουν γίνει αρκετές προσπάθειες για την τυποποίηση των API έτσι ώστε οι προμηθευτές λογισμικού να μπορούν να διανέμουν μια δυαδική εφαρμογή για όλα αυτά τα συστήματα, ωστόσο, μέχρι σήμερα, κανένα από αυτά δεν έχει γνωρίσει μεγάλη επιτυχία. ToLinux Standard Base προσπαθεί να το κάνει αυτό για τη Linux πλατφόρμα, ενώ πολλά από τα BSD Unixes, όπως FreeBSD , NetBSD , και OpenBSD , εφαρμόζουν διάφορα επίπεδα συμβατότητας API για την προς τα πίσω συμβατότητα (επιτρέποντας στα προγράμματα που γράφτηκαν για τις παλαιότερες εκδόσεις να τρέχουν στις νεώτερες διανομές του συστήματος) και cross-platform συμβατότητα (που επιτρέπει την εκτέλεση του ξένου κώδικα χωρίς μεταγλώττιση).

4.8 Πολιτική Κυκλοφορίας

Οι δύο επιλογές για την κυκλοφορία API είναι οι εξής:

1. Προστασία των πληροφοριών του API από το ευρύ κοινό. Για παράδειγμα, η Sony συνήθιζε να κάνει το επίσημο PlayStation 2 API διαθέσιμο μόνο σε αδειούχους προγραμματιστές PlayStation. Αυτό επέτρεψε στη Sony να ελέγχει οποίος έγραφε παιχνίδια για το PlayStation 2. Αυτό δίνει στις επιχειρήσεις προνόμιο έλεγχο της ποιότητας και μπορεί να τους παρέχει τη δυνατότητα εσόδων αδειοδότησης.

2. Κάνοντας τα API διαθέσιμα χωρίς περιορισμούς. Για παράδειγμα, η Microsoft κάνει το Microsoft Windows API κοινό, και η Apple απελευθερώνει API του Carbon και Cocoa, έτσι ώστε λογισμικό να μπορεί να γραφτεί για τις πλατφόρμες αυτές.

Ο συνδυασμός των δύο συμπεριφορών μπορεί να χρησιμοποιηθεί επίσης.

4.9 ABI

Ο σχετικός όρος application binary interface - διεπαφή δυαδικών εφαρμογών (ABI) είναι ένα χαμηλότερο επίπεδο, όσον αφορά τον ορισμό λεπτομερειών επιπέδου στη γλώσσα assembly. Για παράδειγμα, το Linux Standard Base είναι ένα ABI, ενώ το POSIX είναι ένα API.[7]

4.10 Σύνδεσμοι γλώσσας και γεννήτριες διεπαφών

Τα API τα οποία προορίζονται να χρησιμοποιηθούν από περισσότερους του ενός υψηλού επιπέδου γλώσσες προγραμματισμού συχνά παρέχουν, ή είναι αυξημένα, με εγκαταστάσεις που αυτόματα χαρτοποιούν το API για τα χαρακτηριστικών (συντακτικά ή σημασιολογικά) που είναι πιο φυσικό σε αυτές τις γλώσσες. Αυτό είναι γνωστό ως σύνδεσμος γλώσσας, και είναι ο ίδιος ένα API. Ο στόχος είναι να ενσωματώσει την περισσότερη από την απαιτούμενη λειτουργικότητα του API, αφήνοντας ένα "λεπτό" στρώμα κατάλληλο για την κάθε γλώσσα.

Παρακάτω παρατίθενται ορισμένα εργαλεία γεννητριών διασυνδέσεων που δεσμεύουν γλώσσες του API κατά τη μεταγλώττιση .

- Η opensource Swig γεννήτρια διεπαφής από πολλές γλώσσες σε πολλές γλώσσες (Τυπικά Compiled-> δέσμης ενεργειών)
- F2PY: Fortran σε Python γεννήτρια διεπαφής.

ΚΕΦΑΛΑΙΟ 5

Κύριοι αλγόριθμοι και εργαλεία εξόρυξης δεδομένων

ΕΙΣΑΓΩΓΗ

- Classifiers - Ταξινομητές (καλύπτει επιβλεπόμενη ταξινόμηση και παλινδρόμηση)
- Clusterers - Συστάδες (εκμάθηση χωρίς επίβλεψη)
- Associations - Συσχετίσεις
- Attribute Selection - Επιλογή βάση Ιδιότητας (αξιολογητές και τις μεθόδους αναζήτησης)
- Preprocessing Filters - Φίλτρα Προεπεξεργασίας (προεπεξεργασία δεδομένων με επίβλεψη και χωρίς)

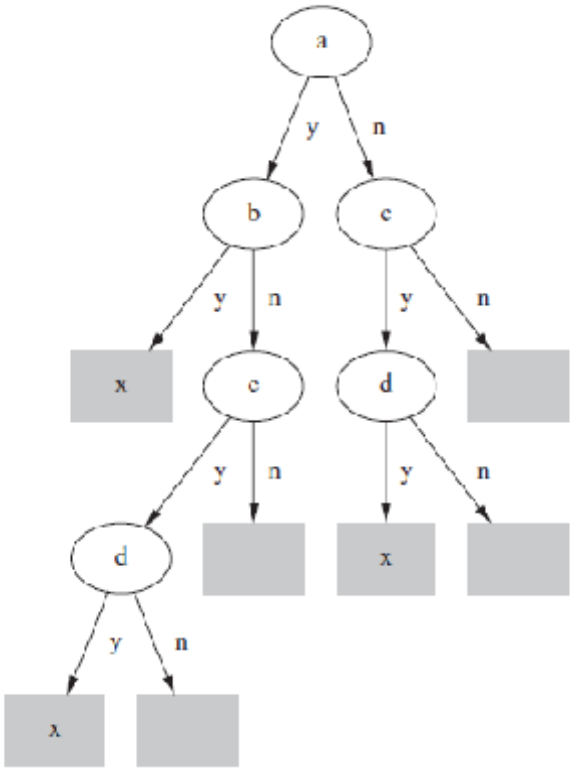
5.1 Classifiers - Ταξινομητές

[1]Οι Κανόνες ταξινόμησης είναι μια δημοφιλής εναλλακτική λύση για τα δέντρα αποφάσεων. Ο προηγούμενος, ή προϋπόθεση, ενός κανόνα είναι μια σειρά τεστ και η επακόλουθη, ή το συμπέρασμα, παρέχει την κλάση ή τις κλάσεις που ισχύουν για τις περιπτώσεις που καλύπτονται από τον εν λόγω κανόνα, ή δίνει ίσως μια κατανομή πιθανότητας για τις κλάσεις Σε γενικές γραμμές, οι προϋποθέσεις έχουν λογικά AND μαζί, και όλες οι δοκιμές επιτυγχάνουν αν ο κανόνας είναι σε εφαρμογή. Ωστόσο, σε ορισμένες διατυπώσεις του κανόνα οι προϋποθέσεις είναι γενικές λογικές εκφράσεις και όχι απλοί σύνδεσμοι. Συχνά σκεφτείτε τους επιμέρους κανόνες ως αποτελεσματικά λογικά OR μαζί: εάν ένα ισχύει, η κλάση (ή την πιθανότητα διανομής) που αναφέρεται στο συμπέρασμά της εφαρμόζεται στην περίπτωση. Ωστόσο, οι συγκρούσεις προκύπτουν όταν αρκετοί από τους κανόνες με διαφορετικές συμπεράσματα ισχύουν.
Είναι εύκολο να διαβάσει ένα σύνολο κανόνων απευθείας από ένα δέντρο απόφασης. Ένας κανόνας παράγεται για κάθε φύλλο. Ο προηγούμενος του κανόνα περιλαμβάνει έναν όρο για κάθε κόμβο σχετικά με τη διαδρομή από τη ρίζα προς το φύλλο, και η επακόλουθη του κανόνα είναι η κλάση που έχει ανατεθεί από το φύλλο. Αυτή η διαδικασία παράγει κανόνες που είναι σαφής στο ότι η σειρά με την οποία εκτελούνται είναι άνευ σημασίας. Ωστόσο,γενικά οι κανόνες που διαβάζονται απευθείας από ένα δέντρο αποφάσεων είναι πολύ πιο περίπλοκες

από ότι αναγκαίες, και οι κανόνες που προέρχονται από δέντρα συνήθως κλαδεύονται για να αφαιρούνται οι περιττές δοκιμές.

Επειδή τα δέντρα αποφάσεων δεν μπορούν να εκφράσουν εύκολα τη διάζευξη μεταξύ των σιωπηρών διαφορετικών κανόνων σε ένα σύνολο, η μετατροπή ενός γενικού σύνολου κανόνων σε ένα δέντρο δεν είναι τόσο απλό. Ένα καλό παράδειγμα είναι αυτό που συμβαίνει όταν οι κανόνες έχουν την ίδια δομή αλλά διαφορετικές ιδιότητες, όπως:

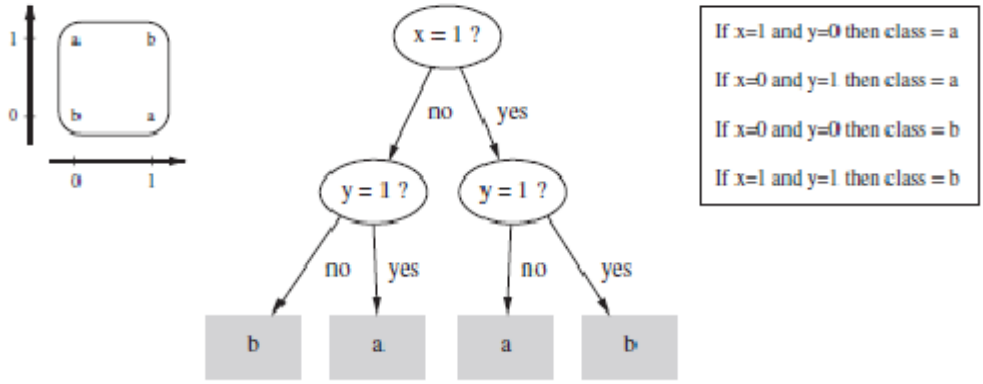
Αν a και b τότε x
Αν c και d τότε x



Σχήμα 1: Δέντρο απόφασης για απλή Διαχώριση

Στη συνέχεια, είναι αναγκαίο να σπάσει η συμμετρία και να επιλεγεί ένα ενιαίο τεστ για τον κόμβο ρίζα. Αν, για παράδειγμα, επιλεγεί, ο δεύτερος κανόνας πρέπει, στην πραγματικότητα, να επαναληφθεί δύο φορές στο δέντρο, όπως φαίνεται στο Σχήμα 1. Αυτό είναι γνωστό ως πρόβλημα αναπαραγωγής υποδέντρου.
Το πρόβλημα αναπαραγωγής υποδέντρου είναι αρκετά σημαντικό. Το Σχήμα 2

εμφανίζει μια λειτουργία αποκλειστικού or για την οποία η έξοδος είναι ένα, αν x = 1 or (ή) y = 1, αλλά όχι και τα δύο.
Για να γίνει αυτό σε ένα δέντρο, θα πρέπει να χωριστεί σε ένα χαρακτηριστικό πρώτα, με αποτέλεσμα να έχει δομή όπως αυτή που φαίνεται στο κέντρο. Αντίθετα, οι κανόνες μπορούν να απεικονίζουν με ακρίβεια, την πραγματική συμμετρία του προβλήματος σε σχέση με τις ιδιότητες.



Σχήμα 2: Το πρόβλημα του αποκλειστικού -or

Εάν ένα σύνολο κανόνων δίνει πολλαπλές ταξινομήσεις για ένα συγκεκριμένο παράδειγμα, μία λύση είναι ν μη δοθεί κανένα συμπέρασμα. Μια άλλη είναι να μετρηθεί πόσο συχνά κάθε κανόνας ανάβει στα δεδομένα εκπαίδευσης και να πάει με το πιο δημοφιλές. Οι στρατηγικές αυτές μπορούν να οδηγήσουν σε ριζικά διαφορετικά αποτελέσματα. Ένα διαφορετικό πρόβλημα εμφανίζεται όταν ένα παράδειγμα αντιμετωπιστεί και οι κανόνες αποτυγχάνουν να το κατατάξουν. Και πάλι, αυτό δεν μπορεί να συμβεί με δέντρα αποφάσεων, ή με κανόνες που διαβάζουν κατευθείαν από αυτά, αλλά μπορεί εύκολα να συμβεί με γενικά σύνολα κανόνων. Ένας τρόπος για την αντιμετώπιση αυτής της κατάστασης είναι να αποτύχει για να κατηγοριοποιήσει ένα τέτοιο παράδειγμα, άλλο είναι να επιλέξουν την πιο συχνά εμφανιζόμενη κλάση ως προεπιλογή.
Και πάλι, ριζικά διαφορετικά αποτελέσματα μπορεί να ληφθούν για τις στρατηγικές αυτές. Οι μεμονωμένοι κανόνες είναι απλοί, και τα σύνολα κανόνων φαίνονται απατηλά απλά, αλλά με δεδομένο απλά ένα σύνολο κανόνων που δεν έχει συμπληρωματικές πληροφορίες, δεν είναι σαφές πώς θα πρέπει να ερμηνευθούν.

Μια ιδιαίτερα απλή κατάσταση εμφανίζεται όταν οι κανόνες οδηγούν σε μια κατηγορία που
είναι Boolean (ας πούμε, ναι και όχι) και όταν μόνο κανόνες οι οποίοι οδηγούν σε ένα αποτέλεσμα (δηλαδή, ναι) εκφράζονται. Η υπόθεση είναι ότι αν μια συγκεκριμένη περίπτωση δεν είναι σε κλάση ναι, τότε πρέπει να είναι στην κλάση

όχι – μια μορφή υπόθεσης κλειστού κόσμου. Αν αυτή είναι η περίπτωση, τότε οι κανόνες δεν μπορούν συγκρούονται και δεν υπάρχει καμία ασάφεια ως προς την ερμηνεία του κανόνα:
οποιασδήποτε ερμηνευτική στρατηγική θα δώσει το ίδιο αποτέλεσμα. Ένα τέτοιο σύνολο κανόνων μπορεί να γραφτεί ως λογική έκφραση σε αυτό που ονομάζεται *διαζευκτική* κανονική μορφή: ότι είναι, ως μια διάζευξη (OR) των συνδετικών (AND) συνθηκών.
Είναι αυτή η απλή ειδική περίπτωση που σαγηνεύει τους ανθρώπους και υποθέτουν ότι οι κανόνες είναι πολύ εύκολοι ώστε να ασχοληθείς μαζί τους, γιατί εδώ κάθε κανόνας λειτουργεί ως ένα νέο, ανεξάρτητο κομμάτι των πληροφοριών που συμβάλλουν σε έναν απλό τρόπο για τη διάζευξη. Δυστυχώς, αυτό ισχύει μόνο για τα Boolean αποτελέσματα και απαιτεί την υπόθεση κλειστού κόσμου, και οι δύο αυτοί οι περιορισμοί είναι εξωπραγματικοί, στις περισσότερες πρακτικές καταστάσεις. Οι αλγόριθμοι της μηχανικής μάθησης που δημιουργούν τους κανόνες πάντοτε παράγουν διατεταγμένων συνόλων κανόνων σε καταστάσεις πολυ-ταξικές (multiclass), και αυτό θυσιάζει κάθε δυνατότητα της σπονδυλωτής κατασκευής, διότι η σειρά εκτέλεσης είναι κρίσιμη.

[2]

Πίνακας 5: Επισκόπηση του συνόλου των συστημάτων ταξινόμησης.

Bayes	**Λειτουργίες**	**Δέντρα**
• AODE • AODEsr • BayesianLogisticRegression • BayesNet • ComplementNaiveBayes • DMNBtext • HNB • NaiveBayes • NaiveBayesMultinomial • NaiveBayesMultinomialUpdateable • NaiveBayesSimple • NaiveBayesUpdateable • WAODE	• GaussianProcesses • IsotonicRegression • LeastMedSq • LibLINEAR • LibSVM • LinearRegression • Logistic • MultilayerPerceptron • PaceRegression • PLSClassifier • RBFNetwork • SimpleLinearRegression • SimpleLogistic • SGD • SMO • SMOreg • SPegasos	• ADTree • BFTree • DecisionStump • FT • ID3 • J48 • J48graft • LADTree • LMT • M5P • NBTree • RandomForest • RandomTree • REPTree • SimpleCart • UserClassifier

	• SVMreg • VotedPerceptron • Winnow	
Lazy • I.B.1 • IBK • KStar • LBR • LWL •	**Κανόνες(Rules)** • ConjunctiveRule • DecisionTable • DTNB • Furia • JRip • M5Rules • NNge • Öner • PART • Prism • Ridor • ZeroR	**MetaData** • AdaBoostM1 • AdditiveRegression • AttributeSelectedClassifier • Bagging • ClassificationViaClustering • ClassificationViaRegression • CostSensitiveClassifier • CVParameterSelection • Dagging • Decorate • END • EnsembleSelection • FilteredClassifier • Grading • GridSearch • LogitBoost • MetaCost • MultiBoostAB • MultiClassClassifier • MultiScheme • OneClassClassifier • OrdinalClassClassifier • RacedIncrementalLogitBoost • RandomCommittee • RandomSubSpace • RealAdaBoost • RegressionByDiscretization • RotationForest • Stacking • StackingC • ThresholdSelector • Vote • ClassBalancedND • DataNearBalancedND • ND

Multi-Instance • CitationKNN • MDD • MIBoost • MIDD • MIEMDD • MILR • MINND • MIOptimalBall • MISMO • MISVM • MIWrapper • SimpleMI • TLD • TLDSimple	**Διάφορα** • HyperPipes • InputMappedClassifier • MinMaxExtension • OLM • OSDL • SerializedClassifier • VFI	

5.1.1 Naive Bayes Classifier Εισαγωγική Επισκόπηση

[3]Η Naive Bayes Classifier τεχνική βασίζεται στο λεγόμενη Bayesian θεώρημα και ενδείκνυται ιδιαίτερα όταν ο χαρακτήρας των εισροών είναι υψηλός. Παρά την απλότητά του, οι Naive Bayes μπορεί να ξεπεράσουν συχνά πιο εξελιγμένες μεθόδους ταξινόμησης.

Για να αποδειχτεί η έννοια της ταξινόμησης Naive Bayes, θεωρούμε το παράδειγμα που εμφανίζεται στην παραπάνω εικόνα. Όπως αναφέρεται, τα αντικείμενα μπορούν να ταξινομηθούν είτε ως πράσινο είτε κόκκινο. Το καθήκον μας είναι να χαρακτηρίσουμε τις νέες υποθέσεις που προκύπτουν, δηλαδή, να αποφασιστεί σε ποια ετικέτα κατηγορίας ανήκουν, με βάση το τρέχον αντικείμενο εξόδου.

Δεδομένου ότι υπάρχουν δύο φορές περισσότερα ΠΡΑΣΙΝΟ απ' ότι ΚΟΚΚΙΝΟ, είναι λογικό να πιστεύουμε σε μια νέα υπόθεση (η οποία δεν έχει παρατηρηθεί ακόμη) ότι είναι δύο φορές πιο πιθανό να έχουν συμμετοχή τα ΠΡΑΣΙΝΟ αντί τα ΚΟΚΚΙΝΟ. Στη Bayesian ανάλυση, αυτή η πεποίθηση είναι γνωστή ως εκ των προτέρων πιθανότητα. Πρώιμες πιθανότητες βασίζονται στην προηγούμενη εμπειρία, στην περίπτωση αυτή το ποσοστό πράσινων και κόκκινων αντικείμενων,

και συχνά χρησιμοποιείται για να προβλέψει τα αποτελέσματα πριν πραγματικά συμβούν.

Έτσι, μπορούμε να γράψουμε:

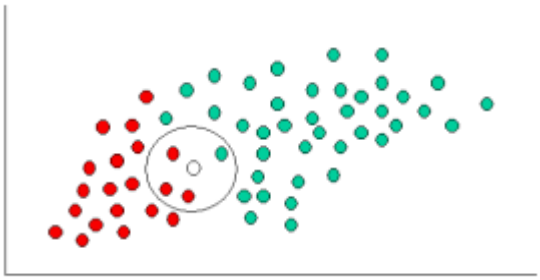
$$Prior\ probability\ for\ GREEN \propto \frac{Number\ of\ GREEN\ objects}{Total\ number\ of\ objects}$$

$$Prior\ probability\ for\ RED \propto \frac{Number\ of\ RED\ objects}{Total\ number\ of\ objects}$$

Δεδομένου ότι υπάρχει ένα σύνολο 60 αντικειμένων των οποίων οι 40 είναι πράσινα και 20 κόκκινα,οι πρώιμες πιθανότητές μας για την συμμετοχής της κατηγορία είναι οι εξής:

$$Prior\ probability\ for\ GREEN \propto \frac{40}{60}$$

$$Prior\ probability\ for\ RED \propto \frac{20}{60}$$



Αφού διατυπωθεί η πρώιμη πιθανότητα μας, είμαστε έτοιμοι να χαρακτηρίσουμε ένα νέο αντικείμενο (άσπρο κύκλο). Δεδομένου ότι τα αντικείμενα είναι καλά συγκεντρωμένα, είναι λογικό να υποθέσει κανείς ότι όσο πιο πολλά ΠΡΑΣΙΝΟ (ή κόκκινο) αντικείμενα βρίσκονται πλησίον του X, τόσο πιο πιθανό ότι τα νέα κρούσματα θα ανήκουν σε αυτό το ειδικό χρώμα. Για τη μέτρηση αυτής της πιθανότητας, θα σχεδιάσετε έναν κύκλο γύρω από X το οποίο περιλαμβάνει έναν αριθμό (που πρέπει να επιλεγεί εκ των προτέρων) σημείων, ανεξάρτητα από την ετικέτα κατηγορίας τους. Τότε θα υπολογιστεί ο αριθμός των σημείων στον κύκλο που ανήκουν σε κάθε ετικέτα κατηγορίας. Από αυτό, υπολογίζουμε την πιθανότητα:

$$Likelihood\ of\ X\ given\ GREEN\ \propto\ \frac{Number\ of\ GREEN\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ GREEN\ cases}$$

$$Likelihood\ of\ X\ given\ RED\ \propto\ \frac{Number\ of\ RED\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ RED\ cases}$$

Από το παραπάνω σχήμα, είναι σαφές ότι το ενδεχόμενο το Χ να δώσει πράσινο είναι μικρότερο από το ενδεχόμενο το Χ να δώσει κόκκινο, δεδομένου ότι ο κύκλος περιλαμβάνει 1 πράσινο αντικείμενο και 3 κόκκινα. Έτσι:

$$Probability\ of\ X\ given\ GREEN \propto \frac{1}{40}$$

$$Probability\ of\ X\ given\ RED \propto \frac{3}{20}$$

Αν και η προηγούμενες πιθανότητες δείχνουν ότι το Χ μπορεί να ανήκουν σε ΠΡΑΣΙΝΟ (δεδομένου ότι υπάρχουν διπλάσιες σε σύγκριση με το ΠΡΑΣΙΝΟ-ΚΟΚΚΙΝΟ) η πιθανότητα δηλώνει το αντίθετο. Ότι η ταξική ένταξη του Χ είναι ΚΟΚΚΙΝΟ (δεδομένου ότι υπάρχουν περισσότερα ΚΟΚΚΙΝΟ αντικείμενα που βρίσκονται κοντά στο Χ από τα ΠΡΑΣΙΝΟ). Στη Bayesian ανάλυση, η τελική κατάταξη παράγεται από συνδυασμό και των δύο πηγών πληροφοριών, δηλαδή, η πρώιμη και η πιθανότητα, να διαμορφώσει μια μετα-πιθανότητα χρησιμοποιώντας τον λεγόμενο κανόνα του Bayes.

$$Posterior\ probability\ of\ X\ being\ GREEN \propto$$
$$Prior\ probability\ of\ GREEN \times Likelihood\ of\ X\ given\ GREEN$$
$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$
$$Posterior\ probability\ of\ X\ being\ RED \propto$$
$$Prior\ probability\ of\ RED \times Likelihood\ of\ X\ given\ RED$$
$$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

Τέλος, ταξινομούμε το Χ ως ΚΟΚΚΙΝΟ καθώς η συμμετοχή της κλάσης του επιτυγχάνει τη μεγαλύτερη μετα-πιθανότητα.

5.2 Association rule - Κανόνες συσχέτισης

Στην εξόρυξη δεδομένων , οι κανόνες συσχέτισης είναι μια δημοφιλής μέθοδος που έχει ερευνηθεί καλά για να ανακαλύψετε ενδιαφέρουσες σχέσεις μεταξύ των μεταβλητών σε μεγάλες βάσεις δεδομένων. Ο [4]Piatetsky-Shapiro περιγράφει την ανάλυση και παρουσίαση των ισχυρών κανόνων που ανακαλύφθηκαν σε βάσεις δεδομένων, χρησιμοποιώντας διάφορα μέτρα ενδιαφέροντος. Με βάση την έννοια των ισχυρών κανόνων, ο [5]Agrawal εισήγαγε κανόνες συσχέτισης για να ανακαλύψει ισότητες μεταξύ προϊόντων σε μεγάλη κλίμακα δεδομένων από συναλλαγές που καταγράφονται από συστήματα σημείου πώλησης (point-of-sale, POS), στα σούπερ μάρκετ. Για παράδειγμα, ο κανόνας που βρέθηκε στα δεδομένα των πωλήσεων σε ένα σουπερμάρκετ, δείχνουν ότι εάν ένας πελάτης αγοράζει τα κρεμμύδια και τις πατάτες μαζί, αυτός ή αυτή είναι πιθανό να αγοράσει και μπέργκερ. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν ως βάση για τις αποφάσεις σχετικά με δραστηριότητες μάρκετινγκ, όπως, π.χ., η τιμολόγηση ή οι τοποθετήσεις προϊόντων. Εκτός από το παραπάνω παράδειγμα από την ανάλυση του καλαθιού αγοράς (market basket analysis), οι κανόνες συσχέτισης χρησιμοποιούνται σήμερα σε πολλούς τομείς, όπου συμπεριλαμβάνουν την εξόρυξη στο Διαδίκτυο, την ανίχνευσης εισβολής και τη βιοπληροφορική.

5.2.1 Ορισμός

Με βάση τον αρχικό ορισμό του [5]Agrawal, το πρόβλημα της εξόρυξης κανόνων συσχέτισης ορίζεται ως εξής: Έστω $I = \{i_1, i_2, \ldots, i_n\}$ είναι ένα σύνολο από *n* δυαδικά χαρακτηριστικά που ονομάζονται αντικείμενα. Έστω $D = \{t_1, t_2, \ldots, t_m\}$ είναι ένα σύνολο πράξεων που ονομάζεται βάση δεδομένων. Κάθε συναλλαγή στο D έχει ένα μοναδικό ID συναλλαγής και περιέχει ένα υποσύνολο των στοιχείων του *I*. Ένας κανόνας ορίζεται ως μια επίπτωση της μορφής $X \Rightarrow Y$ όπου $X, Y \subseteq I$ και $X \cap Y = \emptyset$. Τα σύνολα στοιχείων, Χ και Υ είναι τα αποκαλούμενα "προηγούμενο" (αριστερή πλευρά ή LHS) και "συνακόλουθο" (δεξιά πλευρά ή RHS) του κανόνα, αντίστοιχα.
Για να απεικονίσουμε τις έννοιες, χρησιμοποιούμε ένα μικρό παράδειγμα από το σούπερ μάρκετ. Το σύνολο των στοιχείων είναι *I*= {γάλα, ψωμί, βούτυρο, μπύρα} και μια μικρή βάση δεδομένων που περιέχει τα στοιχεία (1 για την παρουσία και 0

για την απουσία ενός στοιχείου σε μια συναλλαγή) όπως εμφανίζεται στον πίνακα δεξιά.
Ένας κανόνας για παράδειγμα για το σουπερμάρκετ θα μπορούσε να είναι $\{\text{butter}, \text{bread}\} \Rightarrow \{\text{milk}\}$ πράγμα που σημαίνει ότι όταν το βούτυρο και το ψωμί αγοράζονται, οι πελάτες αγοράζουν επίσης και γάλα.
Σημείωση: αυτό το παράδειγμα είναι εξαιρετικά μικρό. Σε πρακτικές εφαρμογές, οι κανόνες χρειάζονται την υποστήριξη πολλών εκατοντάδων συναλλαγών για να μπορεί να θεωρηθεί στατιστικά σημαντική, και τα σύνολα δεδομένων περιέχουν συχνά χιλιάδες ή εκατομμύρια συναλλαγές.

Πίνακας 6: Παράδειγμα βάσης δεδομένων με τα 4 στοιχεία και 5 συναλλαγές

ID Συναλλαγής	Γάλα	Ψωμί	Βούτηρο	Μπύρα
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

5.2.2 Χρήσιμες Αρχές

Για να επιλεγούν ενδιαφέροντες κανόνες από το σύνολο όλων των πιθανών κανόνων, μπορούν να χρησιμοποιηθούν περιορισμοί σχετικά με διάφορα αξιακά και σημασιολογικά μέτρα. Τα πιο γνωστά προβλήματα είναι τα ελάχιστα όρια για τη στήριξη και την εμπιστοσύνη.

- Η υποστήριξη supp(X) του στοιχειοσυνόλου X ορίζεται ως το ποσοστό των συναλλαγών στο σύνολο των δεδομένων που περιέχουν το στοιχειοσύνολο. Στη βάση δεδομένων του παραδειγματος, το στοιχειοσύνολο {γάλα, ψωμί, βούτυρο} έχει υποστήριξη 5.1 = 0.2 δεδομένου ότι εμφανίζεται σε ποσοστό 20% του συνόλου των συναλλαγών (1 στα 5 συναλλαγές).

- Η *εμπιστοσύνη* του κανόνα ορίζεται $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$.Για παράδειγμα, ο κανόνας $\{\text{milk}, \text{bread}\} \Rightarrow \{\text{butter}\}$ έχει την εμπιστοσύνη της τάξης του 0,2 / 0,4

= 0,5 στη βάση δεδομένων, πράγμα που σημαίνει ότι για το 50% των συναλλαγών που περιέχουν γάλα και ψωμί ο κανόνας είναι σωστός.

- Η εμπιστοσύνη μπορεί να ερμηνευθεί ως η εκτίμηση της πιθανότητας *P (Y | X),* η πιθανότητα της εύρεσης του κανόνα RHS στις συναλλαγές, υπό την προϋπόθεση ότι οι συναλλαγές αυτές περιέχουν επίσης τα LHS.[6]

- Ο *ανελκυστήρας (lift)* του κανόνα ορίζεται ως

$$\operatorname{lift}(X \Rightarrow Y) = \frac{\operatorname{supp}(X \cup Y)}{\operatorname{supp}(Y) \times \operatorname{supp}(X)}$$ ή ο λόγος της παρατηρούμενης στήριξης σε σχέση με την αναμενόμενη, αν X και Y ήταν ανεξάρτητες . Ο κανόνας $\{\text{milk}, \text{bread}\} \Rightarrow \{\text{butter}\}$ διαθέτει ανελκυστήρα $\frac{0.2}{0.4 \times 0.4} = 1.25.$

- Το φρόνημα *(conviction)* ενός κανόνα ορίζεται ως

$$\operatorname{conv}(X \Rightarrow Y) = \frac{1 - \operatorname{supp}(Y)}{1 - \operatorname{conf}(X \Rightarrow Y)}$$. Ο κανόνας $\{\text{milk}, \text{bread}\} \Rightarrow \{\text{butter}\}$ έχει ένα φρόνημα $\frac{1 - 0.4}{1 - .5} = 1.2$ και μπορεί να ερμηνευθεί ως ο λόγος της αναμενόμενης συχνότητας του X χωρίς να εμφανίζεται το Y (δηλαδή, η συχνότητα εσφαλμένης πρόβλεψης από τον κανόνα) αν X και Y ήταν ανεξάρτητες διαιρούμενες με την παρατηρούμενη συχνότητα των εσφαλμένων προβλέψεων. Σε αυτό το παράδειγμα, η αξία φρονήματος των 1,2 δείχνει ότι ο κανόνας $\{\text{milk}, \text{bread}\} \Rightarrow \{\text{butter}\}$ Θα ήταν λάθος 20% πιο συχνά (1,2 φορές πιο συχνά) εάν η σύνδεση μεταξύ X και Y ήταν καθαρά τυχαία επιλογή.

- Η ιδιότητα της *σαφήνειας* (που χαρακτηρίζεται από σαφή, ακριβή έκφραση με λίγες λέξεις) ενός περιορισμού. Ένας περιορισμός είναι σύντομος, αν είμαστε σε θέση να γράψουμε ρητά όλα τα σύνολα-θέσεις, που ικανοποιούν τον περιορισμό.
  Παράδειγμα: περιορισμός C = S. Τύπος = {μη φαγώσιμο}
  Τα προϊόντα που θα ικανοποιήσουν αυτόν τον περιορισμό είναι για τα πρώην. {Ακουστικά, παπούτσια, χαρτί υγείας}
  Χρηστικό Παράδειγμα: Αντί να χρησιμοποιήσουμε τον αλγόριθμο Apriori για να βρούμε τα συχνά σύνολα μπορούμε αντ' αυτού να δημιουργήσουμε όλα τα στοιχειοσύνολα και να τρέξουμε τον μετρητή υποστήριξης μόνο μία φορά.

5.2.3 Διαδικασία

Οι κανόνες συσχέτισης συνήθως απαιτείται να ικανοποιούν μια ελάχιστη υποστήριξη που καθορίζει ο χρήστης και μια ελάχιστη εμπιστοσύνη ταυτόχρονα. Η παραγωγή κανόνων διαιρείται συνήθως σε δύο χωριστά στάδια:

1. Πρώτον, η ελάχιστη στήριξη εφαρμόζεται για την εύρεση όλων των συχνών συνόλων στοιχείων σε μια βάση δεδομένων.
2. Δεύτερον, τα σύνολα αυτά και η ελάχιστη παροχή εμπιστοσύνης χρησιμοποιούνται για να διαμορφώσουν τους κανόνες.

Ενώ το δεύτερο βήμα είναι η ευθεία προς τα εμπρός, το πρώτο βήμα χρειάζεται περισσότερη προσοχή.

Η εύρεση όλων των συχνών συνόλων σε μια βάση δεδομένων είναι δύσκολη, δεδομένου ότι προϋποθέτει την αναζήτηση όλων των πιθανών συνόλων στοιχείων (συνδυασμός στοιχείων). Το σύνολο των πιθανών συνόλων είναι το δυναμοσύνολο πέραν του I και έχει μέγεθος $2^n - 1$ (με εξαίρεση το κενό σύνολο που δεν είναι έγκυρο στοιχειοσύνολο). Αν και το μέγεθος του δυναμοσυνόλου αυξάνεται εκθετικά με τον αριθμό των στοιχείων n στο I, αποτελεσματική αναζήτηση είναι δυνατή με τη χρήση της στήριξης του "κλείσιμο ιδιοκτησίας προς τα κάτω" [7](downward-closure property) (ονομάζεται επίσης κατά της μονοτονίας[8]), η οποία εγγυάται ότι, για ένα συχνό στοιχειοσύνολο, όλα τα υποσύνολα του είναι επίσης συχνά και, συνεπώς, για ένα σπάνιο στοιχειοσύνολο, όλα τα υπερσύνολά του πρέπει επίσης να είναι σπάνια. Εκμεταλλευόμενοι αυτή την ιδιότητα, αποδοτικοί αλγόριθμοι (π.χ., Apriori και Eclat) μπορούν να βρουν όλα τα συχνά σύνολα στοιχείων.

5.2.4 Ιστορία

Η έννοια των κανόνων συσχέτισης διαδόθηκε κυρίως το 1993 λόγω του άρθρου του Agrawal[5], το οποίο έχει αποκτήσει περισσότερες από 6000 αναφορές σύμφωνα με το Google Scholar, τον Μάρτιο του 2008, και έτσι είναι ένα από τα πιο αναφερόμενα άρθρα στον τομέα Εξόρηξης Δεδομένων. Ωστόσο, είναι πιθανό ότι αυτό που τώρα ονομάζεται "κανόνες συσχέτισης" είναι παρόμοιο με αυτό που φαίνεται στο [9]άρθρο του 1966 στο Guha, μια γενική μέθοδος εξόρυξης δεδομένων που αναπτύχθηκε από τον [10]Petr Hajek.

5.2.5 Εναλλακτικά μέτρα ενδιαφέροντος

Μαζί με την εμπιστοσύνη έχουν προταθεί και άλλα μέτρα ενδιαφέροντος για κανόνες συσχέτισης. Μερικά δημοφιλή μέτρα είναι τα εξής:

- Ολοκληρωτική εμπιστοσύνη.[11]
- Συλλογική δύναμη[12]
- Φρόνημα (Conviction)[13]

- Μόχλευση[14]
- Lift (που αρχικά λεγόταν συμφέρον)[13]

5.2.6 Αλγόριθμοι

Πολλοί αλγόριθμοι για την παραγωγή κανόνων συσχέτισης παρουσιάστηκαν στην πάροδο του χρόνου.

- Apriori
- FilteredAssociator
- FPGrowth
- GeneralizedSequentialPatterns
- HotSpot
- PredictiveApriori
- Tertius

Apriori αλγόριθμος

[15]Ο Apriori είναι ο πιο γνωστός αλγόριθμος για την εξόρυξη κανόνων συσχέτισης. Χρησιμοποιεί μια στρατηγική αναζήτησης για να μετράει την υποστήριξη των στοιχειοσυνόλων και χρησιμοποιεί μια υποψήφια συνάρτηση παραγωγής η οποία εκμεταλλεύεται την ιδιότητα του πτωτικού κλεισίματος της στήριξης. Ο αλγόριθμος Apriori έχει σχεδιαστεί για να λειτουργεί με βάσεις δεδομένων που περιέχουν συναλλαγές (για παράδειγμα, συλλογές προϊόντων που αγοράζονται από πελάτες, ή λεπτομέρειες της επισκεψιμότητας μιας ιστοσελίδας). Άλλοι αλγόριθμοι έχουν σχεδιαστεί για την εξεύρεση προτύπων στα δεδομένα που δεν έχουν συναλλαγές (Winepi και Minepi), ή που δεν έχουν χρονική στιγμή (αλληλουχία DNA).

Όπως συνηθίζεται στην εξόρυξη κανόνων συσχέτισης , δίνεται ένα σύνολο από στοιχειοσύνολα(για παράδειγμα, σύνολα λιανικών συναλλαγών, κάθε απαρίθμησης μεμονωμένων ειδών που αγοράστηκαν), ο αλγόριθμος προσπαθεί να βρει υποσύνολα που είναι κοινά σε τουλάχιστον ένα ελάχιστο αριθμό C του στοιχειοσυνόλου. Ο Apriori χρησιμοποιεί μια «από κάτω προς τα άνω» προσέγγιση, όπου συχνά τα υποσύνολα παρατείνονται ανά ένα στοιχείο κάθε φορά (ένα βήμα γνωστήό ως υποψήφια γενιά), καθώς και ομάδες υποψηφίων που δοκιμάζονται με βάση τα δεδομένα. Ο αλγόριθμος τερματίζει όταν δεν βρήσκεται καμία περαιτέρω επιτυχής επέκταση.

Ο Apriori χρησιμοποιεί breadth-first αναζήτηση και μια δομή δέντρο για να μετρήσει τα υποψήφια στοιχειοσύνολα αποτελεσματικά. Δημιουργεί υποψήφια

στοιχειοσύνολα μήκους k από στοιχειοσύνολα μήκους k-1. Στη συνέχεια, “κλαδεύει” τα υποψηφία που έχουν ένα σπάνιο υπο-μοτίβο. Σύμφωνα με το πτωτικό κλείσιμο λήμμα (downward closure lemma), το υποψήφιο σύνολο περιέχει όλα τα συχνά στοιχειοσύνολα μήκους -k. Στη συνέχεια, σαρώνει τη βάση δεδομένων της συναλλαγής για να καθορίσει το συχνό σύνολο στοιχείων μεταξύ των υποψηφίων.

Ο αλγόριθμος Apriori, ενώ είναι ιστορικά σημαντικός, πάσχει από διάφορες δυσλειτουργίες ή συμβιβασμούς, οι οποίες έχουν γεννήσει άλλους αλγορίθμους. Η υποψήφια γενιά δημιουργεί μεγάλους αριθμούς υποσυνόλων (ο αλγόριθμος προσπαθεί να φορτώσει το υποψήφιο σύνολο όσο το δυνατόν περισσότερο πριν από κάθε σάρωση).

Επιλογές

Ο παρακάτω πίνακας περιγράφει τις διαθέσιμες επιλογές για τον αλγόριθμο Apriori.[2]

Πίνακας 7: Επιλογές για τον αλγόριθμο Apriori

Επιλογή	Περιγραφή
Αυτοκίνητο (car)	Αν ενεργοποιηθεί, κανόνες συσχέτισης κλάσεως εξορύσσεται αντί των (γενικών) κανόνων.
classIndex	Δείκτης του χαρακτηριστικού κλάσης. Αν οριστεί σε -1, το τελευταίο χαρακτηριστικό λαμβάνεται ως χαρακτηριστικό κλάσης.
Δέλτα (delta)	Επαναληπτική μείωση στήριξης από αυτόν τον παράγοντα. Μειώνει τη στήριξη στο ελάχιστο ή μέχρι να δημιουργηθεί ο απαιτούμενος αριθμός κανόνων.
lowerBoundMinSupport	Κάτω φράγμα για την ελάχιστη υποστήριξη.
metricType	Καθορίζει τον τύπο μέτρησης ο οποίος ταξινομεί τους κανόνες. Η εμπιστοσύνη είναι το ποσοστό των παραδειγμάτων που καλύπτονται από την προϋπόθεση που καλύπτονται επίσης και από το αποτέλεσμα (οι κανόνες συσχέτισης κλάσεων μπορούν να εξορυχθούν μόνο με εμπιστοσύνη). Lift είναι η εμπιστοσύνη που διαιρείται με το ποσοστό σε όλα τα παραδείγματα που καλύπτονται από το αποτέλεσμα. Αυτό είναι ένα μέτρο της σημασίας της σύνδεσης, η οποία είναι ανεξάρτητη από τη στήριξη. Η μόχλευση είναι η αναλογία των επιπλέον παραδείγματα που καλύπτονται τόσο από την υπόθεση και κατά συνέπεια υψηλότερες από τις αναμενόμενες, αν το σκεπτικό και το αποτέλεσμα ήταν ανεξάρτητες μεταξύ τους. Ο συνολικός αριθμός των παραδειγμάτων που αυτό αντιπροσωπεύει παρουσιάζεται σε παρενθέσεις μετά την μόχλευση. Καταδίκη είναι ένα άλλο μέτρο της εξόδου από την ανεξαρτησία.

Επιλογή	Περιγραφή
	Καταδίκη δίνεται από
minMetric	Ελάχιστο σκορ μέτρησης. Λαμβάνονται υπόψιν μόνο οι κανόνες με σκορ μεγαλύτερο από την τιμή αυτή.
numRules	Αριθμός κανόνων που μπορούν να βρεθούν.
outputItemSets	Εάν είναι ενεργοποιημένο τα στοιχειοσύνολα είναι και αποτέλεσμα.
removeAllMissingCols	Καταργεί όλες τις στήλες με τις τιμές που λείπουν.
σημαντικότητας(significanceLevel)	Επίπεδο σημασίας. Έλεγχος σημαντικότητας (εμπιστοσύνη μετρικό μόνο).
upperBoundMinSupport	Ανώτατο όριο για την ελάχιστη υποστήριξη. Αρχίζει επαναληπτική μείωση για την ελάχιστη στήριξη από την τιμή αυτή.
πολύλογος (verbose)	Εάν είναι ενεργοποιημένο, ο αλγόριθμος θα εκτελεστεί σε verbose mode.

Αλγόριθμος FP (FPGrowth.)

Η FP- ανάπτυξη (συχνή ανάπτυξη μοτίβου[16]), χρησιμοποιεί μια εκτεταμένη δομή πρόθεμα-δέντρο (FP- tree) για την αποθήκευση των δεδομένων σε συμπιεσμένη μορφή. Η FP- ανάπτυξη υιοθετεί μια “διαίρει και βασίλευε” προσέγγιση για να αποσυνδέσει τις εργασίες εξόρυξης και τις βάσεις δεδομένων. Χρησιμοποιεί μια μέθοδο ανάπτυξης μοτίβου για να αποφευχθεί η δαπανηρή διαδικασία της παραγωγής υποψηφίων και των δοκιμών που χρησιμοποιούνται από τον αλγόριθμο Apriori.

Επιλογές

Ο παρακάτω πίνακας περιγράφει τις διαθέσιμες επιλογές για FPGrowth.[2]

Πίνακας 8: Επιλογές FPGrowth

Επιλογή	Περιγραφή
Δέλτα (delta)	Επαναληπτική μείωση στήριξης από αυτόν τον παράγοντα. Μειώνει τη στήριξη στο ελάχιστο ή μέχρι να δημιουργηθεί ο απαιτούμενος αριθμός κανόνων.
findAllRulesForSupportLevel	Βρείτε όλους τους κανόνες που πληρούν το κατώτατο όριο για την ελάχιστη στήριξη του και την ελάχιστη μετρική πίεση. Όσον αφορά αυτή την κατάσταση για να απενεργοποιήσει την επαναληπτική διαδικασία της μείωσης της υποστήριξης

Επιλογή	Περιγραφή
	για να βρείτε τον καθορισμένο αριθμό κανόνων.
lowerBoundMinSupport	Κάτω φράγμα για την ελάχιστη υποστήριξη.
maxNumberOfItems	Ο μέγιστος αριθμός στοιχείων που θα περιλαμβάνονται σε συχνές σημείο σύνολα. -1 Σημαίνει ότι δεν υπάρχει όριο.
metricType	Καθορίστε τον τύπο των μετρικών με τις οποίες να ταξινομήσει τους κανόνες. Η εμπιστοσύνη είναι το ποσοστό των παραδειγμάτων που καλύπτονται από την προϋπόθεση ότι καλύπτονται επίσης και από το αποτέλεσμα (Class κανόνες συσχέτισης μπορούν να εξορυχθούν με εμπιστοσύνη). Lift είναι η εμπιστοσύνη διαιρείται με το ποσοστό του σε όλα τα παραδείγματα που καλύπτονται από το αποτέλεσμα.Αυτό είναι ένα μέτρο της σημασίας της σύνδεσης, η οποία είναι ανεξάρτητη από τη στήριξη. Η μόχλευση είναι η αναλογία των επιπλέον παραδείγματα που καλύπτονται τόσο από την υπόθεση και κατά συνέπεια υψηλότερες από τις αναμενόμενες, αν το σκεπτικό και το αποτέλεσμα ήταν ανεξάρτητες μεταξύ τους. Ο συνολικός αριθμός των παραδειγμάτων που αυτό αντιπροσωπεύει παρουσιάζεται σε παρενθέσεις μετά την μόχλευση. Καταδίκη είναι ένα άλλο μέτρο της εξόδου από την ανεξαρτησία.
minMetric	Ελάχιστη μετρικούς σκορ. Σκεφτείτε μόνο τους κανόνες με σκορ μεγαλύτερο από την τιμή αυτή.
numRulesToFind	Ο αριθμός των κανόνων για την έξοδο
positiveIndex	Ρυθμίστε το δείκτη του δυαδικού αποτιμάται ιδιότητες που πρέπει να θεωρείται ο θετικός δείκτης. Δεν έχει κανένα αποτέλεσμα για το αραιό των δεδομένων (στην περίπτωση αυτή το πρώτο δείκτη (δηλαδή μη μηδενικών τιμών) είναι αντιμετωπίζονται πάντα ως θετικές. Επίσης δεν έχει καμία επίπτωση για μοναδιαίοι αποτιμώνται ιδιότητες (π.χ. κατά τη χρήση του Weka Apriori-style μορφοτύπου για τα δεδομένα καλάθι της αγοράς, η οποία χρησιμοποιεί διαθέσιμη τιμή ";" για να δείξει έλλειψη ενός στοιχείου.
upperBoundMinSupport	Ανώτατο όριο για την ελάχιστη υποστήριξη. Αρχίζει επαναληπτική μείωση για την ελάχιστη στήριξη από την τιμή αυτή.

FilteredAssociator

[2]Κατηγορία για τη λειτουργία ενός αυθαίρετου συνδέσμου (associator) σε δεδομένα που έχουν περάσει μέσα από ένα αυθαίρετο φίλτρο. Όπως και ο σύνδεσμος, η δομή του φίλτρου βασίζεται αποκλειστικά στα δεδομένα εκπαίδευσης και οι περιπτώσεις δοκιμών θα υποβληθούν σε επεξεργασία από το φίλτρο χωρίς να μεταβληθεί η δομή τους.

Επιλογές

Ο παρακάτω πίνακας περιγράφει τις διαθέσιμες επιλογές για τον αλγόριθμο FilteredAssociator.

Πίνακας 9: Επιλογές αλγόριθμου FilteredAssociator

Επιλογή	Περιγραφή
associator	Ο βασικός σύνδεσμος που θα χρησιμοποιηθεί.
classIndex	Δείκτης του χαρακτηριστικού κλάσης. Αν οριστεί σε -1, το τελευταίο χαρακτηριστικό λαμβάνεται ως χαρακτηριστικό κλάσης.
filter	Το φίλτρο που θα χρησιμοποιηθούν.

GeneralizedSequentialPatterns

[2]Η κλάση εφαρμόζει έναν αλγόριθμο GSP για την ανακάλυψη διαδοχικών προτύπων σε ένα διαδοχικό σύνολο δεδομένων.
Το χαρακτηριστικό που προσδιορίζει τις διαφορετικές αλληλουχίες δεδομένων που περιέχονται στο σύνολο, καθορίζεται από την αντίστοιχη επιλογή. Επιπλέον, το σύνολο των αποτελεσμάτων που παράγονται μπορούν να περιοριστούν, ορίζοντας ένα ή περισσότερα χαρακτηριστικά που πρέπει να περιέχονται σε κάθε στοιχείο / στοιχειοσύνολο της ακολουθίας.

Επιλογές

Ο παρακάτω πίνακας περιγράφει τις διαθέσιμες επιλογές για GeneralizedSequentialPatterns.

Πίνακας 10: Επιλογές αλγόριθμου GeneralizedSequentialPatterns

Επιλογή	Περιγραφή
dataSeqID	Ο αριθμός του χαρακτηριστικού που αντιπροσωπεύει το ID της σειράς

Επιλογή	Περιγραφή
	δεδομένων.
debug	Αν οριστεί σε true, ο αλγόριθμος μπορεί να παράγει επιπλέον πληροφορίες στην κονσόλα.
filterAttributes	Οι αριθμοί του χαρακτηριστικού (π.χ. "0, 1") που χρησιμοποιούνται για το αποτελεσματικό φιλτράρισμα, μόνο ακολουθίες που περιέχουν τα ειδικά χαρακτηριστικά σε καθένα από τα στοιχεία τους / στοιχειοσύνολα θα βγουν στην έξοδο. Το -1 τις εκτυπώνει όλες.
minSupport	Ελάχιστο όριο στήριξης.

HotSpot

[2]Το HotSpot μαθαίνει ένα σύνολο κανόνων (που παρουσιάζεται σε μία δομή που μοιάζει με δέντρο), που μεγιστοποιούν / ελαχιστοποιούν μια ενδιαφέρουσα μεταβλητή-στόχο. Με έναν ονομαστικό στόχο, θα μπορούσε κανείς να ψάξει για τμήματα δεδομένων, όπου υπάρχει μεγάλη πιθανότητα να συμβεί μια μειονότητας αξίας (δεδομένου του περιορισμού της ελάχιστης στήριξης). Για έναν αριθμητικό στόχο, θα μπορούσε να ενδιαφέρει η εξεύρεση τμημάτων όπου είναι υψηλότερα κατά μέσο όρο σε σχέση με το συνολικό σύνολο δεδομένων. Για παράδειγμα, σε ένα σενάριο ασφάλισης υγείας, βρίσκοντας ποιες ομάδες ασφάλισης έχουν το μεγαλύτερο ρίσκο (έχουν τον υψηλότερο δείκτη αξίωσης), ή ποιες ομάδες έχουν το υψηλότερο μέσο όρο πληρωμής ασφάλισης.

Επιλογές

Ο παρακάτω πίνακας περιγράφει τις διαθέσιμες επιλογές για το HotSpot.

Πίνακας 11: Επιλογές HotSpot

Επιλογή	Περιγραφή
debug	Πληροφορίες εξόδου αποσφαλμάτωσης (διπλός κανόνας αναζήτησης στατιστικών του hash table).
maxBranchingFactor	Μέγιστος παράγοντας διακλάδωσης. Ο μέγιστος αριθμός των παιδιών για το ενδεχόμενο επέκτασης σε κάθε κόμβο.
minImprovement	Ελάχιστη βελτίωση της τιμής-στόχου, προκειμένου να εξετάσει την προσθήκη μιας νέας διακλάδοσης.

Επιλογή	Περιγραφή
minimizeTarget	Ελαχιστοποίηση και όχι τη μεγιστοποίηση του στόχου.
support	Η ελάχιστη υποστήριξη. Οι τιμές μεταξύ 0 και 1 ερμηνεύονται ως ποσοστό του συνολικού πληθυσμού, τιμές > 1 ερμηνεύονται ωςένας απόλυτος αριθμός περιπτώσεων.
target	Το χαρακτηριστικό στόχος που ενδιφέρει.
targetIndex	Η αξία του στόχου (ονομαστικά χαρακτηριστικά μόνο (nominal attributes)) ενδιαφέροντος.

Προγνωστικός αλγόριθμος Apriori (PredictiveApriori)

[2]Η κλάση εφαρμόζει τον προγνωστικό αλγόριθμο Apriori για να εξορύξει κανόνες συσχέτισης. Αναζητεί με αυξανόμενο όριο στήριξης των κανόνων τα 'n' καλύτερα, στηριζόμενη σε μια διορθωμένη τιμή εμπιστοσύνης. Ένας κανόνας προστίθεται εάν:
η αναμενόμενη ακρίβεια του κανόνα αυτού είναι μεταξύ των 'n' καλύτερων και δεν θα αφομοιωθεί από έναν κανόνα με τουλάχιστον την ίδια αναμενόμενη πρόβλεψη ακριβείας.

Επιλογές

Ο παρακάτω πίνακας περιγράφει τις διαθέσιμες επιλογές για τον PredictiveApriori.

Πίνακας 12: Επιλογές PredictiveApriori

Επιλογή	Περιγραφή
car	Αν ενεργοποιηθεί οι κανόνες συσχέτισης της κλάσης εξορύσσεται αντί των (γενικών) κανόνων συσχέτισης.
classIndex	Δείκτης του χαρακτηριστικού κλάσης. Αν οριστεί σε -1, το τελευταίο χαρακτηριστικό θα λαμβάνεται ως χαρακτηριστικό κλάσης.
numRules	Αριθμός των κανόνων να βρεθούν.

Tertius αλγόριθμος

[18]Ο αλγόριθμος Tertius χτίζει κανόνες από τις τιμές των χαρακτηριστικών στα δεδομένα προς εκπαίδευση και τις κατατάσσει σύμφωνα με το πόσο αξιόπιστες είναι - δηλαδή, πόσες φορές ο κανόνας ισχύει. Ένας κανόνας αποτελείται από ένα σώμα και ένα κεφάλι. Το σώμα περιέχει τους όρους (Γνωστά ως προτάσεις) που απαιτούνται να κατέχει ο κανόνας και μπορεί να αποτελείται από οποιοδήποτε αριθμό προτάσεων. Η κεφαλή περιέχει το συμβάν που παρουσιάζεται όταν ο κανόνας ισχύει. Κατά τη διάρκεια της μάθησης, ο Tertius ξεκινά με έναν κενό κανόνα. Ο κανόνας στη συνέχεια καθαρίζεται με την προσθήκη ζευγαριών χαρακτηριστικό-τιμή με τη σειρά που εμφανίζονται στο σύνολο δεδομένων. Μόλις γίνει αυτό, ο αλγόριθμος υπολογίζει τον αριθμό των φορών που ο κανόνας ισχύει και τις φορές που ο κανόνας δίνει ψευδές θετικό. Ο αλγόριθμος Tertius προσφέρει μερικά χρήσιμα χαρακτηριστικά γνωρίσματα. Είναι σε θέση να αντιμετωπίσει τιμές που λείπουν, το οποίο είναι χρήσιμο στις περιπτώσεις όπου ορισμένες ιδιότητες έχουν παραλειφθεί στην καταχώρηση των δεδομένων. Ο Tertius έχει επίσης τη λειτουργικότητα να περιλαμβάνει αρνήσεις στις προτάσεις του, για παράδειγμα, θα μπορούσε να προβλέψει μια μελλοντική κατάσταση. Η λειτουργικότητα που το χωρίζει από τους άλλους αλγόριθμους είναι ότι μπορεί να ρυθμιστεί για τα χαρακτηριστικά κλάσεων και μόνο. Οι άλλοι αλγόριθμοι δεν μπορούν να το κάνουν αυτό. Ένα μειονέκτημα του Tertius είναι ο σχετικά μεγάλος χρόνος λειτουργίας του, το οποίο σε μεγάλο βαθμό εξαρτάται από τον αριθμό των προτάσεων στους κανόνες. Η αύξηση του αριθμού των προτάσεων αυξάνει το χρόνο εκτέλεσης εκθετικά.

Επιλογές

Ο παρακάτω πίνακας περιγράφει τις διαθέσιμες επιλογές για Tertius.[2]

Πίνακας 13: Επιλογές Tertius

Επιλογή	Περιγραφή
classIndex	Δείκτης χαρακτηριστικού κλάσης. Αν οριστεί σε 0, η κλάση θα είναι το τελευταίο χαρακτηριστικό.
classification	Βρήσκει μόνο κανόνες με την κλάση στο κεφάλι.
confirmationThreshold	Ελάχιστη επιβεβαίωση των κανόνων.
confirmationValues	Αριθμός των καλύτερων τιμών επιβεβαίωσης για να βρεθούν.
frequencyThreshold	Ελάχιστο ποσοστό των περιπτώσεων που πληρούν το κεφάλι και το σώμα των κανόνων
hornClauses	Βρίσκει κανόνες με μία πρόταση συμπέρασμα μόνο.
missingValues	Ορίζει τον τρόπο χειρισμού των τιμών που λείπουν. Οι τιμές που

Επιλογή	Περιγραφή
	λείπουν μπορούν να ρυθμιστούν ώστε να ταιριάζουν με οποιαδήποτε τιμή, ή να ταιριάζουν με τις τιμές και να είναι σημαντικές και ενδεχομένως να εμφανίζονται σε κανόνες.
negation	Καθορίστε τον τύπο της άρνησης που επιτρέπεται στον κανόνα. Η άρνηση μπορεί να επιτραπεί στο σώμα, στο κεφάλι, και στα δύο ή σε κανένα.
noiseThreshold	Μέγιστο ποσοστό του μετρητή ιδιοτήτων των κανόνων. Αν οριστεί σε 0, μόνο οι ικανοποιημένοι κανόνες θα δοθούν.
numberLiterals	Μέγιστος αριθμός των προτάσεων σε έναν κανόνα.
repeatLiterals	Επαναληπτικές ιδιότητες επιτρέπονται.
rocAnalysis	Επιστροφή των TP-rate και FP-rate για κάθε κανόνα που βρίσκει.
valuesOutput	Δίνει οπτική ανατροφοδότηση κατά τη διάρκεια της αναζήτησης. Οι τρέχουσες καλύτερες και τις χειρότερες τιμές μπορούν να εξαχθούν είτε στο stdout ή σε ξεχωριστό παράθυρο.

5.3 Ανάλυση κατά συστάδες – Clusterers

Ανάλυση κατά συστάδες ή ομαδοποίηση είναι η ανάθεση ενός συνόλου παρατηρήσεων σε υποσύνολα (που ονομάζονται clusters), έτσι ώστε οι παρατηρήσεις στο ίδιο σύμπλεγμα να είναι παρόμοιες υπό κάποια έννοια. Η ομαδοποίηση είναι μια μέθοδος εκμάθησης χωρίς επίβλεψη , και μια κοινή τεχνική για την στατιστική ανάλυση των δεδομένων που χρησιμοποιούνται σε πολλούς τομείς, συμπεριλαμβανομένης της μηχανικής μάθησης (machine learning), της εξόρυξης δεδομένων, της αναγνώρισης προτύπων, της ανάλυσης εικόνας, της ανάκτησης πληροφοριών, και της βιοπληροφορικής.

Εκτός από τον όρο ομαδοποίηση, υπάρχει μια σειρά από όρους με παρόμοια σημασία, όπως αυτόματη ταξινόμηση, αριθμητική ταξινόμηση, botryology και τυπολογική ανάλυση.

5.3.1 Τύποι ομαδοποίησης

Οι Ιεραρχικοί αλγόριθμοι βρίσκει διαδοχικές ομάδες από ομάδες που έχουν ήδη συσταθεί. Αυτοί οι αλγόριθμοι συνήθως είναι είτε συσσωρευτικοί (agglomerative)

("bottom-up") ή διχαστικοί ("top-down»). Οι Agglomerative αλγόριθμοι ξεκινούν με κάθε στοιχείο ως ξεχωριστό σύμπλεγμα και τα συγχωνεύει διαδοχικά σε μεγαλύτερες ομάδες. Οι Διχαστικοί αλγόριθμοι ξεκινούν με το σύνολο και το χωρίζουν σε διαδοχικά μικρότερες ομάδες.

Οι διαιρετικοί (Partitional) αλγόριθμοι συνήθως καθορίζουν όλα τα συμπλέγματα με τη μία, αλλά μπορεί επίσης να χρησιμοποιηθεί ως διχαστικός αλγόριθμοι στην ιεραρχική ομαδοποίηση.

Οι αλγόριθμοι ομαδοποίησης με βάση την πυκνότητα επινοήθηκαν για να ανακαλύψουν συστάδες με αυθαίρετο σχήμα. Σε αυτήν την προσέγγιση, μια συστάδα θεωρείται ως μια περιοχή στην οποία η πυκνότητα των δεδομένων αντικειμένων υπερβαίνει ένα όριο. Οι DBSCAN και OPTICS είναι δύο χαρακτηριστικοί αλγόριθμοι αυτού του είδους.

Η μέθοδος ομαδοποίησης Subspace ψάχνει για τις ομάδες που μπορούν μόνο να βρεθούν σε μια συγκεκριμένη προβολή των δεδομένων. Αυτές οι μέθοδοι μπορεί έτσι να αγνοήσουν άνευ σημασίας χαρακτηριστικά. Το γενικό πρόβλημα είναι επίσης γνωστό ως Συσχέτιση ομαδοποίηση , ενώ η ειδική περίπτωση του άξονα-παράλληλου Subspace είναι επίσης γνωστή ως διπλή ομαδοποίηση, συν-ομαδοποίηση ή biclustering : σε αυτές τις μεθόδους όχι μόνο τα αντικείμενα είναι συγκεντρωμένοι, αλλά και τα χαρακτηριστικά των αντικειμένων, δηλαδή, αν τα δεδομένα εκπροσωπούνται σε μια μήτρα δεδομένων , οι γραμμές και οι στήλες είναι οι συγκεντρωμένες ταυτόχρονα. Συνήθως δεν λειτουργούν με αυθαίρετους συνδυασμούς χαρακτηριστικών, όπως σε γενικές μεθόδους subspace. Αλλά αυτή η ειδική περίπτωση χρήζει προσοχής λόγω των πολλών εφαρμογών της στην βιοπληροφορική.

Πολλοί αλγόριθμοι ομαδοποίησης απαιτούν την προδιαγραφή του αριθμού των clusters για την παραγωγή στην εισροή του συνόλου δεδομένων, πριν από την εκτέλεση του αλγορίθμου.

5.3.2 Απόσταση μέτρο

Ένα σημαντικό βήμα στις περισσότερες ομαδοποιήσεις είναι η επιλογή ενός μέτρου απόστασης, το οποίο θα καθορίσει το πώς η *ομοιότητα* των δύο στοιχείων θα υπολογίζεται. Αυτό θα επηρεάσει τη μορφή των clusters, δεδομένου ότι ορισμένα στοιχεία μπορεί να είναι κοντά το ένα στο άλλο σύμφωνα με ένα μέτρα και πιο μακριά, σύμφωνα με ένα άλλο. Για παράδειγμα, σε έναν 2-διαστάσεων χώρο, η απόσταση μεταξύ του σημείου *(x* = 1, *y* = 0) και προσανατολισμό *(x* = 0, *y* = 0) είναι πάντα 1 σύμφωνα με τους συνήθεις κανόνες, αλλά η απόσταση μεταξύ των σημείο *(x* = 1, *y* = 1) και προσανατολισμό 2, $\sqrt{2}$ ή 1 αν πάρετε, αντίστοιχα, το 1-κανόνα, 2- κανόνα ή άπειρο- κανόνα απόσταση.

Συνήθεις μέθοδοι απόστασης:

- Η Ευκλείδεια απόσταση (ονομάζεται επίσης απόσταση σε ευθεία γραμμή ή 2-κανόνα απόσταση). Μια επανεξέταση της ανάλυσης συστάδων στην έρευνα της ψυχικής υγείας διαπίστωσε ότι το πιο κοινό μέτρο απόστασης σε δημοσιευμένες μελέτες σε αυτόν τον τομέα της έρευνας είναι η Ευκλείδεια απόσταση ή το τετράγωνο της Ευκλείδειας απόστασης.
- Η απόσταση Manhattan (γνωστή και ως κανόνας ταξί ή 1-κανόνα)
- Ο μέγιστος κανόνας (γνωστός και ως κανόνας άπειρο)
- Η απόσταση Mahalanobis διορθώνει τα δεδομένα για διάφορες κλίμακες και συσχετισμούς στις μεταβλητές.
- Η γωνία μεταξύ δύο διανυσμάτων μπορεί να χρησιμοποιηθεί ως μέτρο απόστασης όταν ομαδοποιούνται δεδομένα υψηλής διάστασης.
- Η απόσταση Hamming μέτρα τον ελάχιστο αριθμό των αντικαταστάσεων που απαιτούνται για την αλλαγή ένα μέλους σε άλλο.

Μια άλλη σημαντική διάκριση είναι το κατά πόσον η ομαδοποίηση χρησιμοποιεί συμμετρικές ή ασύμμετρες αποστάσεις. Πολλές από τις μεθόδους αποστάσεων που αναφέρονται παραπάνω έχουν την ιδιότητα ότι οι αποστάσεις είναι συμμετρικές (η απόσταση από το αντικείμενο A στο B είναι το ίδιο με την απόσταση από το B στην A). Σε άλλες εφαρμογές αυτό δεν συμβαίνει. (Μια αληθινή μέτρηση δίνει συμμετρική μέτρηση απόστασης.)

5.3.3 Ιεραρχική ομαδοποίηση

Η Ιεραρχική ομαδοποίηση δημιουργεί μια ιεραρχία από clusters, η οποία αναπαριστάται με μια δομή δέντρου που ονομάζεται δενδρόγραμμα. Η ρίζα του δένδρου αποτελείται από ένα ενιαίο σύμπλεγμα που περιέχει όλες τις παρατηρήσεις και τα φύλλα αντιστοιχούν σε επιμέρους παρατηρήσεις.

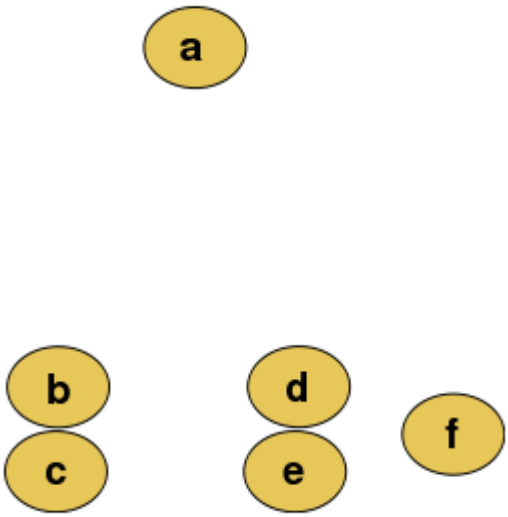
Οι αλγόριθμοι για ιεραρχική ομαδοποίηση είναι εν γένει είτε agglomerative, στην οποία ένας ξεκινάει από τα φύλλα και συγχωνεύει διαδοχικά clusters μαζί, ή διχαστική, κατά την οποία ξεκινά από τη ρίζα και αναδρομικά χωρίζει τις ομάδες.

Οποιαδήποτε έγκυρη μέτρηση μπορεί να χρησιμοποιηθεί ως μέτρο της ομοιότητας μεταξύ ζευγαριών παρατηρήσεων. Η επιλογή των ομάδων που είναι να συγχωνευθούν ή διαιρεθούν προσδιορίζεται από το κριτήριο σύνδεσης, το οποίο αποτελεί μέθοδο των αποστάσεων σε ζεύγη μεταξύ των παρατηρήσεων.

Κόβοντας το δέντρο σε ένα δεδομένο ύψος, θα δώσει ομαδοποίηση σε μια επιλεγμένη ακρίβεια. Στο παρακάτω παράδειγμα, το κόψιμο μετά τη δεύτερη σειρά θα αποφέρει συστάδες {a} {BC} {de} {f}. Κόβοντας μετά την τρίτη σειρά θα αποφέρει συστάδες {a} {BC} {def}, το οποίο είναι μια ομαδοποίηση πιο χονδροειδής, με ένα μικρότερο αριθμό μεγαλύτερων clusters.

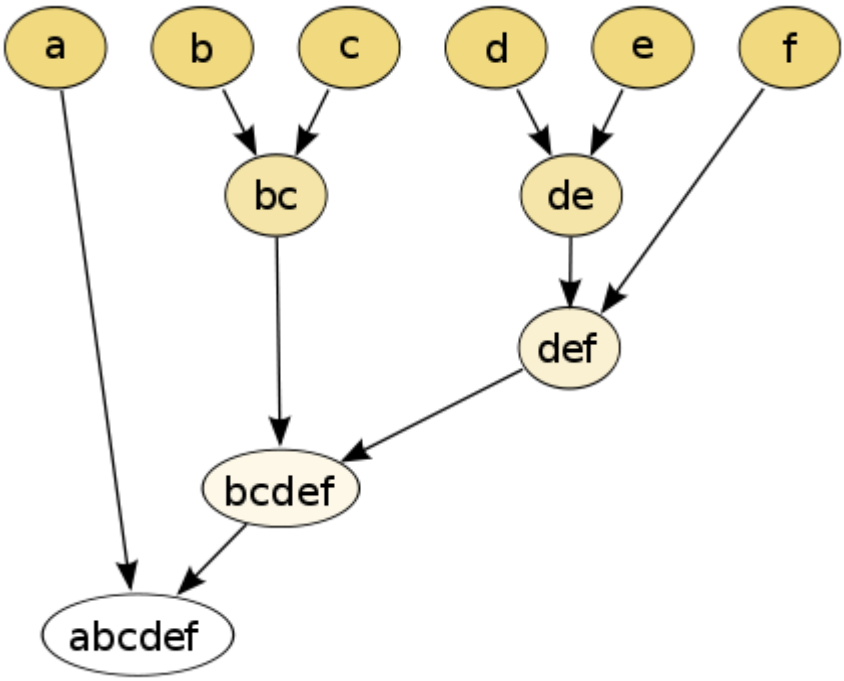
5.3.4 Agglomerative ιεραρχική ομαδοποίηση

Για παράδειγμα, ας υποθέσουμε ότι αυτά τα δεδομένα συγκεντρώνονται, και η ευκλείδεια απόσταση είναι το μέτρο απόστασης.



Σχήμα 3: Ανεπεξέργαστα δεδομένα

Η ιεραρχική ομαδοποίηση δενδρόγραμμα θα ήταν:



Σχήμα 4: Παραδοσιακή αναπαράσταση

Η μέθοδος αυτή βασίζεται στην ιεραρχία από τα επιμέρους στοιχεία με σταδιακή συγχώνευση clusters. Στο παράδειγμά μας, έχουμε έξι στοιχεία {a} {b} {c} {d} {e} και {f}. Το πρώτο βήμα είναι να καθοριστεί ποια στοιχεία να συγχωνευθούν σε ένα cluster. Συνήθως, θέλουμε να αναλάβουμε τα δύο πλησιέστερα στοιχεία, σύμφωνα με την επιλεγμένη απόσταση.

Προαιρετικά, μπορεί κανείς να κατασκευάσει επίσης έναν πίνακα αποστάσεων σε αυτό το στάδιο, όπου ο αριθμός της *i-οστής* γραμμής *j-οστής* στήλης είναι η απόσταση μεταξύ του *i-οστού* και *j-οστού* στοιχείου. Έτσι, όπως και η ομαδοποίηση προχωρά, οι γραμμές και στήλες συγχωνεύονται, όπως είναι και τα clusters και οι αποστάσεις ενημερώνονται. Αυτός είναι ένας κοινός τρόπος για την εφαρμογή αυτού του τύπου της ομαδοποίησης, και έχει το πλεονέκτημα της προσωρινής αποθήκευσης των αποστάσεων μεταξύ των clusters.

Ας υποθέσουμε ότι έχουμε συγχωνεύσει τα δύο πλησιέστερα στοιχεία *b* και *c* , τώρα έχουμε τις εξής ομάδες *{a}, {b, c}, {d}, {ε}* και *{f},* και θέλουν να τις συνενώσουμε περαιτέρω. Για να γίνει αυτό, πρέπει να πάρουμε την απόσταση μεταξύ {a} και {*b c*}, και ως εκ τούτου να καθορίσουμε την απόσταση μεταξύ των δύο ομάδων. Συνήθως η απόσταση μεταξύ δύο clusters $\mathcal{A}$ και $\mathcal{B}$ είναι ένα από τα ακόλουθα:

- Η μέγιστη απόσταση μεταξύ των στοιχείων της κάθε ομάδας (που ονομάζεται επίσης πλήρης συνδεδεμένη ομαδοποίηση):

$$\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \,\}.$$

- Η ελάχιστη απόσταση μεταξύ των στοιχείων της κάθε ομάδας (γνωστή επίσης ως μονή συνδεδεμένη ομαδοποίηση):

$$\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \,\}.$$

- Η μέση απόσταση μεταξύ των στοιχείων της κάθε ομάδας (που ονομάζεται επίσης κατά μέσο όρο συνδεδεμένη ομαδοποίηση):

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y).$$

- Το άθροισμα όλων των διακυμάνσεων εντός του cluster.
- Η αύξηση της μεταβλητότητας του cluster που συγχωνεύεται).
- Η πιθανότητα ότι τα υποψήφια clusters παράγονται από την ίδια συνάρτηση κατανομής (V-σύνδεση).

Κάθε οικισμός(agglomeration) εμφανίζεται σε μεγαλύτερη απόσταση μεταξύ των ομάδων από τον προηγούμενο οικισμό, και μπορεί να αποφασίσει να σταματήσει την ομαδοποίηση είτε όταν οι ομάδες είναι πάρα πολύ μακριά για να συγχωνευθούν (κριτήριο απόστασης) ή όταν υπάρχει ένας αρκετά μικρός αριθμός ομάδων (αριθμητικό κριτήριο).

5.3.5 Φασματική ομαδοποίηση

Λαμβάνοντας υπόψη ένα σύνολο δεδομένων σημείων A, ο πίνακας ομοιότητας μπορεί να οριστεί ως ένας πίνακας *S* όπου το *S i j* αντιπροσωπεύει το μέτρο της ομοιότητας μεταξύ των σημείων $i, j \in A$. Οι Φασματική τεχνικές ομαδοποίησης, κάνουν χρήση του φάσματος του πίνακα ομοιότητας των δεδομένων για να εκτελέσουν τη μείωση των διαστάσεων για την ομαδοποίηση σε λιγότερες διαστάσεις.

Μια τέτοια τεχνική είναι ο αλγόριθμος Κανονικοποιημένου Τεμαχίου του Shi-Malik, που χρησιμοποιούνται συνήθως για την τμηματοποίηση εικόνων. Χωρίζει σημεία σε δύο σύνολα (S 1, S 2) με βάση το ιδιοδιάνυσμα V που αντιστοιχεί στη δεύτερη μικρότερη ιδιοτιμή του πίνακα Laplac

$$L = I - D^{-1/2} S D^{-1/2}$$

του *S,* όπου *D* είναι ο διαγώνιος πίνακας

$$D_{ii} = \sum_{j} S_{ij}.$$

Αυτή η κατάτμηση μπορεί να γίνει με διάφορους τρόπους, όπως με τη λήψη του διάμεσου *m* από τα στοιχεία της V, και τοποθετώντας όλα τα σημεία των οποίων η συνιστώσα στη *V* είναι μεγαλύτερη από την m στην S_1, και τα υπόλοιπα στην S_2.Ο αλγόριθμος μπορεί να χρησιμοποιηθεί για ιεραρχική ομαδοποίηση με επανειλημμένη τμηματοποίηση των υποσυνόλων με αυτό τον τρόπο.

Ένας συναφής αλγόριθμος είναι ο αλγόριθμος Meila-Shi , ο οποίος παίρνει τα ιδιοδιανύσματα που αντιστοιχούν στις *k* μεγαλύτερες ιδιοτιμές του πίνακα $P = S D^{-1}$ για κάποια *k,* και στη συνέχεια επικαλείται ένα άλλο (π.χ. *K-μέσα)* για να ομαδοποιήσει σημεία από τα αντίστοιχα *k* στοιχεία τους σε αυτά τα ιδιοδιανύσματα.

5.3.6 Εφαρμογές

Βιολογία

Στη βιολογία η ομαδοποίηση έχει πολλές εφαρμογές

- Στην απεικόνιση, η ομαδοποίηση δεδομένων μπορεί να λάβει διαφορετική μορφή με βάση τη διάσταση των δεδομένων. Για παράδειγμα, η SOCR EM Μείγμα μοντέλου δραστηριότητας τμηματοποίησης και applet, δείχνει πώς να αποκτήσετε το σημείο, την περιοχή ή την κατάταξη χρησιμοποιώντας τις online υπολογιστικές βιβλιοθήκες SOCR .
- Στους τομείς της φυτικής και ζωικής οικολογίας , η ομαδοποίηση χρησιμοποιείται για να περιγράψει και να κάνει χωρικές και χρονικές συγκρίσεις των κοινοτήτων (των συνόλων) των οργανισμών σε ετερογενή περιβάλλοντα, χρησιμοποιείται επίσης σε συστηματική φυτών για την παραγωγή τεχνητών φυλογενέσεων ή ομάδων οργανισμών (ιδιωτών) στο είδος, το γένος ή το υψηλότερο επίπεδο από αυτά που μοιράζονται έναν αριθμό χαρακτηριστικών
- Σε θέματα υπολογιστικής βιολογίας και της βιοπληροφορικής :
  - Στις μεταγραφωμικές (transcriptomics), η ομαδοποίηση χρησιμοποιείται για να χτίσει ομάδες γονιδίων που σχετίζονται με πρότυπα έκφρασης (επίσης γνωστά ως συνεκφράζονται γονίδια). Συχνά, οι ομάδες αυτές περιλαμβάνουν λειτουργικά συνδεόμενες πρωτεΐνες, όπως τα ένζυμα για μια συγκεκριμένη διαδρομή , ή τα γονίδια που είναι συν-ρυθμιστές. Υψηλής απόδοσης πειράματα χρησιμοποιώντας ετικέτες εξέφρασης

ακολουθίας (ESTs) ή μικροσυστοιχίες DNA μπορούν να είναι ένα ισχυρό εργαλείο για ανάλυση γονιδιώματος , μια γενική όψη της γονιδιωματικής.
    - Στην ανάλυση ακολουθίας , η ομαδοποίηση χρησιμοποιείται για τη δημιουργία ομάδων από ομόλογες ακολουθίες σε οικογένειες γονιδίων . Αυτή είναι μια πολύ σημαντική έννοια στην βιοπληροφορική και την εξελικτική βιολογία εν γένει.
    - Σε υψηλής απόδοσης πλατφόρμες γονοτυπικής οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την αυτόματη εκχώρηση γονότυπων .
- Σε QSAR και μοριακές μελέτες μοντελοποίησης όπως επίσης και στη χημειοπληροφορική.

Ιατρική, Ψυχολογία και τις Νευροεπιστήμες

Στην ιατρική απεικόνιση , όπως η τομογραφία εκπομπής ποζιτρονίων (PET scans), η ανάλυση των clusters μπορεί να χρησιμοποιηθεί για τη διαφοροποίηση μεταξύ των διαφόρων τύπων ιστών και αίματος σε ένα τρισδιάστατο είδωλο. Σε αυτή την εφαρμογή, την πραγματική θέση δεν έχει σημασία, αλλά η ένταση voxel θεωρείται ως ένα διάνυσμα, με μια διάσταση για κάθε εικόνα που λήφθηκε στην πάροδο του χρόνου. Η τεχνική αυτή επιτρέπει, για παράδειγμα, η ακριβή μέτρηση του ρυθμού ενός ραδιενεργού ιχνηθέτη παραδίδεται στην περιοχή ενδιαφέροντος, χωρίς ξεχωριστή δειγματοληψία του αρτηριακού αίματος, μια παρεμβατική τεχνική που είναι πιο γνωστή σήμερα.

Έρευνα αγοράς

Η ανάλυση των clusters χρησιμοποιείται ευρέως στην έρευνα αγοράς , όταν ασχολείται με δεδομένων πολυμεταβλητών από έρευνες και πάνελ ελέγχου. Οι ερευνητές αγοράς χρησιμοποιούν την ανάλυση των clusters για να χωρίσουν το γενικό πληθυσμό των καταναλωτών σε τμήματα της αγοράς και να κατανοήσουν καλύτερα τις σχέσεις μεταξύ των διαφόρων ομάδων καταναλωτών / υποψήφιων καταναλωτών.

- Κατακερματισμός της αγορά και καθορισμός των αγορών-στόχων
- Τοποθέτησης προϊόντος
- Ανάπτυξη νέων προϊόντων
- Επιλογή αγορών δοκιμής

Η εκπαιδευτική έρευνα

Στην εκπαιδευτική ερευνητική ανάλυση, τα στοιχεία για την ομαδοποίηση μπορεί να είναι οι μαθητές, οι γονείς, το φύλο ή ο βαθμός των διαγωνισμάτων. Η

ομαδοποίηση είναι μια σημαντική μέθοδος για την κατανόηση και τη χρησιμότητα της διασποράς στην εκπαιδευτική έρευνα. [18]Η ανάλυση των clusters στην εκπαιδευτική έρευνα μπορεί να χρησιμοποιηθεί για την εξερεύνηση των δεδομένων, την επιβεβαίωση της διασποράς και τον ελέγχο της υπόθεσης. Η εξερεύνηση των δεδομένων χρησιμοποιείται όταν υπάρχουν λίγες πληροφορίες σχετικά με τα σχολεία ή τους μαθητές που θα συγκεντρωθούν. Στόχος της είναι η ανακάλυψη οποιασδήποτε ουσιαστικής ομάδας από μονάδες με βάση τη μέτρηση σε ένα σύνολο υπεύθυνων μεταβλητών. [19]Η επιβεβαίωση των Clusters χρησιμοποιείται για την επιβεβαίωση των προηγούμενων αναφορών αποτελεσμάτων των Clusters. [20]Οι έλεγχοι υποθέσεων χρησιμοποιούνται για τη διευθέτηση της δομής των clusters.

5.3.7 Άλλες εφαρμογές

Ανάλυση των κοινωνικών δικτύων

Στη μελέτη των κοινωνικών δικτύων , η ομαδοποίηση, μπορεί να χρησιμοποιηθεί για να αναγνωριστούν οι κοινότητες μέσα σε μεγάλες ομάδες ανθρώπων.

Λογισμικό εξέλιξης

Η ομαδοποίηση είναι χρήσιμη στην εξέλιξη του λογισμικού, δεδομένου ότι βοηθά στη μείωση των κληρονομικών ιδιοτήτων στον κώδικα μεταρρυθμίζοντας τις λειτουργίες που έχουν διασπαστεί. Πρόκειται για μια μορφή της αναδιάρθρωσης και, ως εκ τούτου είναι ένας τρόπος για άμεση προληπτική συντήρηση.

Η τμηματοποίηση εικόνων

Το clustering μπορεί να χρησιμοποιηθεί για τη διαίρεση μιας ψηφιακής εικόνας σε ξεχωριστές περιοχές για παραμεθόρια ανίχνευση ή για την αναγνώριση αντικειμένων.

Εξόρυξη δεδομένων

Πολλές εφαρμογές εξόρυξης δεδομένων αφορούν τμηματοποίηση δεδομένων σε αντίστοιχα υποσύνολα. Μια άλλη συνηθισμένη εφαρμογή είναι η κατανομή των εγγράφων, όπως των σελίδων του World Wide Web, σε είδη.

Ομαδοποίηση αποτελεσμάτων αναζήτησης

Κατά τη διαδικασία της ευφυούς ομαδοποίησης των αρχείων και ιστοσελίδων, η ομαδοποίηση, μπορεί να χρησιμοποιηθεί για να δημιουργηθεί ένα πιο σχετικό σύνολο αποτελεσμάτων αναζήτησης σε σύγκριση με τις κανονικές

μηχανές αναζήτησης όπως το Google . Υπάρχει σήμερα μια σειρά διαδικτυακών εργαλείων ομαδοποίησης, όπως το Clusty .

Βελτιστοποίηση χάρτη

Στο Flickr ο χάρτης των φωτογραφιών και οι άλλες σελίδες με χάρτες χρησιμοποιούν clustering για τη μείωση του αριθμού των δεικτών σε ένα χάρτη. Αυτό τις καθιστά πιο γρήγορες και μειώνει την οπτική ακαταστασία.

IMRT κατάτμηση

Το clustering μπορεί να χρησιμοποιείται για να διαιρέσει ένα χάρτη πυκνότητας ενέργειας σε ξεχωριστές περιοχές για τη μετατροπή σε παραδοτέους τομείς της MLC Ακτινοθεραπείας.

Ομαδοποίηση στοιχείων Shopping

Το clustering μπορεί να χρησιμοποιηθεί για την ομαδοποίηση όλων των στοιχείων που είναι διαθέσιμο στο Διαδίκτυο σε ένα σύνολο μοναδικών προϊόντων. Για παράδειγμα, όλα τα στοιχεία του eBay μπορούν να ομαδοποιηθούν σε μοναδικά προϊόντα.

Συστήματα Σύστασης

Τα συστήματα Recommender είναι σχεδιασμένα να συστήνουν νέα στοιχεία βάσει των προτιμήσεων του χρήστη. Χρησιμοποιούν μερικές φορές αλγόριθμους ομαδοποίησης για να προβλέψουν τις προτιμήσεις του χρήστη με βάση τις προτιμήσεις των άλλων χρηστών στο cluster του χρήστη.

Μαθηματική χημεία

Για να βρει τη δομική ομοιότητα, κλπ., για παράδειγμα, 3000 χημικές ενώσεις ήταν συγκεντρωμένες στο χώρο των 90 τοπολογικών δεικτών[21].

Κλιματολογία

Για να βρει καιρικές συνθήκες ή μοντέλα για το επίπεδο της θάλασσας ατμοσφαιρικής πίεσης[22].

Πετρελαϊκής Γεωλογίας

Το Cluster Analysis χρησιμοποιείται για την ανακατασκευή εκλιπόντων κάτω οπών βασικών δεδομένων ή καμπύλων που λείπουν ώστε να εκτιμηθούν οι ιδιότητες των ρεζερβουάρ.

Φυσική Γεωγραφία

Η ομαδοποίηση των χημικών ιδιοτήτων σε διαφορετικές τοποθεσίες-δείγματα.

Ανάλυση Εγκλημάτων

Η Cluster ανάλυση μπορεί να χρησιμοποιηθεί για να εντοπιστούν οι τομείς όπου υπάρχουν οι μεγαλύτερες επιπτώσεις συγκεκριμένων μορφών εγκληματικότητας. Με τον προσδιορισμό διακριτών τέτοιων περιοχών ή "hot spots" όπου ένα παρόμοιο έγκλημα συνέβη κατά τη διάρκεια μιας χρονικής περιόδου, είναι δυνατή η διαχείριση των πόρων επιβολής του νόμου πιο αποτελεσματικά.

Εξελικτικοί αλγόριθμοι

Η Ομαδοποίηση, μπορεί να χρησιμοποιηθεί για τον εντοπισμό διαφορετικών θέσεων μέσα στον πληθυσμό ενός εξελικτικού αλγορίθμου, έτσι ώστε η αναπαραγωγική δυνατότητα να μπορεί να διανεμηθούν πιο ομοιόμορφα μεταξύ των εξελισσόμενων ειδών ή των υπο-ειδών.

5.3.8 Αξιολόγηση της ομαδοποίησης

Η αξιολόγηση της ομαδοποίησης μερικές φορές αναφέρεται ως επικύρωση Cluster.
Υπήρξαν πολλές προτάσεις για ένα μέτρο ομοιότητας μεταξύ δύο clusterings. Ένα τέτοιο μέτρο μπορεί να χρησιμοποιηθεί για να συγκρίνουν πως διαφορετικοί αλγόριθμοι ομαδοποίησης δεδομένα εκτελούνται σε ένα σύνολο δεδομένων. Τα μέτρα αυτά συνήθως συνδέονται με τον τύπο του κριτηρίου που εξετάζονται κατά την αξιολόγηση της ποιότητας της μεθόδου ομαδοποίησης δεδομένων.

5.3.9 Εσωτερικό κριτήριο ποιότητας

Οι μέθοδοι της αξιολόγησης ομαδοποίησης που εφαρμόζουν στο εσωτερικό κριτήριο αναθέτει την καλύτερη βαθμολογία στον αλγόριθμο που παράγει clusters με υψηλή ομοιότητα μέσα σε μια συστάδα και χαμηλής ομοιότητας μεταξύ των ομάδων. Ένα μειονέκτημα της χρήσης του εσωτερικού κριτηρίου στην αξιολόγηση clusters είναι ότι οι υψηλές βαθμολογίες ενός εσωτερικού μέτρου δεν οδηγεί απαραίτητα σε αποτελεσματικές εφαρμογές ανάκτησης πληροφοριών [23].
Οι ακόλουθες μέθοδοι μπορούν να χρησιμοποιηθούν για την αξιολόγηση της ποιότητας των αλγορίθμων ομαδοποίησης με βάση το εσωτερικό κριτήριο:

- **Δείκτης Davies-Bouldin**

  Ο δείκτης Davies-Bouldin μπορεί να υπολογιστεί με τον ακόλουθο τύπο:

$$DB = \frac{1}{n} \sum_{i=1, i \neq j}^{n} \max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

όπου n είναι ο αριθμός των συνεργατικών σχηματισμών, cx είναι το κέντρο βάρους της διασποράς x, σx είναι η μέση απόσταση όλων των στοιχείων στο σύμπλεγμα x με κέντρο βάρους cx, και d (ci , cj) η απόσταση μεταξύ centroids ci και cj. Από τη στιγμή που οι αλγόριθμοι που παράγουν συσπειρώσεις με χαμηλή απόσταση ενδο-cluster (υψηλή ομοιότητα ενδο-cluster) και μεγάλες αποστάσεις μεταξύ των συνεργατικών σχηματισμών (χαμηλή ομοιότητα μεταξύ cluster) θα έχουν χαμηλό δείκτη Davies-Bouldin.Ο αλγόριθμος που παράγει μια συλλογή από clusters με το μικρότερο δείκτη Davies-Bouldin θεωρείται ο καλύτερος αλγόριθμος με βάση αυτά τα κριτήρια.

- **Δείκτης Dunn**

  Ο δείκτης Dunn στοχεύει στον εντοπισμό πυκνών και καλά διαχωρισμένων συστάδων. Ορίζεται ως ο λόγος μεταξύ της ελάχιστης απόστασης μεταξύ των συνεργατικών σχηματισμών σε μέγιστη απόσταση εντός του συμπλέγματος. Για κάθε διαμέρισμα συμπλέγματος, ο δείκτης Dunn μπορεί να υπολογιστεί με τον ακόλουθο τύπο[24]:

$$D = \min_{1\le i\le n}\left\{\min_{1\le j\le n, i\neq j}\left\{\frac{d(i,j)}{\max_{1\le k\le n} d'(k)}\right\}\right\}$$

  όπου d (i, j) αντιπροσωπεύει την απόσταση μεταξύ clusters i και j, και d' (k) τα μέτρα της ενδο-cluster απόστασης k συμπλέγματος. Η απόσταση μεταξύ των συστάδων d (i, j) ανάμεσα σε δύο ομάδες μπορεί να είναι οποιοσδήποτε αριθμός των εξ αποστάσεως μέτρων, όπως η απόσταση μεταξύ των centroids των clusters. Ομοίως, η απόσταση ενδο-cluster d' (k) μπορεί να μετρηθεί με διάφορους τρόπους, όπως είναι η μέγιστη απόσταση μεταξύ κάθε ζεύγους στοιχείων των clusters k. Δεδομένου ότι το εσωτερικό κριτήριο επιδιώξει συστάδες με υψηλή ομοιότητα ενδο-διασποράς και χαμηλής ομοιότητας μεταξύ των συστάδων, οι αλγόριθμοι που παράγουν ομάδες με υψηλό δείκτη Dunn είναι οι πιο επιθυμητοί.

5.3.10 Εξωτερικό κριτήριο ποιότητας

Οι μέθοδοι της αξιολόγησης ομαδοποίησης που εφαρμόζουν εξωτερικό κριτήριο, συγκρίνουν τα αποτελέσματα των αλγορίθμων κατά ορισμένων εξωτερικών σημείων αναφοράς. Τα εν λόγω σημεία αποτελούνται από ένα σύνολο προ-ταξινομημένων στοιχείων, και αυτά τα σύνολα συχνά δημιουργούνται από ανθρώπους εμπειρογνώμονες. Έτσι, το σύνολο των σημείων αναφοράς μπορεί να θεωρηθεί ως χρυσός κανόνας για την αξιολόγηση. Αυτοί οι τύποι των μεθόδων αξιολόγησης μετρούν πόσο κοντά είναι η ομαδοποίηση στις προκαθορισμένες κλάσεις αναφοράς. Ωστόσο, συζητήθηκε προσφάτως αν αυτό είναι επαρκές για

πραγματικά δεδομένα, ή μόνο για τα συνθετικά σύνολα δεδομένων με πραγματολογική αλήθεια εδάφους, δεδομένου ότι οι κλάσεις μπορούν να περιέχουν εσωτερική δομή, τα χαρακτηριστικά του παρόντος μπορεί να μην επιτρέπουν διαχωρισμό των ομάδων ή οι κλάσεις μπορεί να περιέχουν ανωμαλίες.

Ορισμένα από τα μέτρα της ποιότητας ενός αλγορίθμου διασποράς χρησιμοποιώντας εξωτερικό κριτήριο περιλαμβάνουν:

- Μέτρο Rand

  Ο δείκτης Rand υπολογίζει πόσο συναφή είναι τα clusters (που επέστρεψε ο αλγόριθμος) με το σημείο αναφοράς. Κάποιος μπορεί επίσης να δεί το δείκτη Rand ως μέτρο του ποσοστού των σωστών αποφάσεων που λαμβάνονται από τον αλγόριθμο. Μπορεί να υπολογιστεί χρησιμοποιώντας τον ακόλουθο τύπο:

  $$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

  όπου TP είναι ο αριθμός των αληθώς θετικών, TN είναι ο αριθμός των πραγματικών αρνητικών, FP είναι ο αριθμός των ψευδώς θετικών, και FN είναι ο αριθμός των ψευδώς αρνητικών. Ένα θέμα με το δείκτη Rand είναι ότι τα ψευδώς θετικά και ψευδώς αρνητικά σταθμίζονται εξίσου. Αυτό μπορεί να είναι ανεπιθύμητο χαρακτηριστικό για κάποιες εφαρμογές ομαδοποίησης. Η F-μέτρηση αντιμετωπίζει την ανησυχία αυτή.

- F-measure (F-μέτρηση)

  Η F-μέτρηση μπορεί να χρησιμοποιηθεί για την εξισορρόπηση της συμβολής των ψευδώς αρνητικών με ανάκληση στάθμισης μέσω μιας παραμέτρου $\beta \geq 0$. Η ακρίβεια και η ανάκληση ορίζονται ως εξής:

  $$P = \frac{TP}{TP + FP}$$

  $$R = \frac{TP}{TP + FN}$$

  όπου P είναι το ποσοστό ακρίβειας και R είναι το ποσοστό ανάκλησης. Μπορούμε να υπολογίσει το F-measure, χρησιμοποιώντας τον ακόλουθο τύπο[23]:

  $$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

Παρατηρήστε ότι όταν β = 0, F_0 = P. Με άλλα λόγια, η ανάκληση δεν επηρεάζει το F-measure όταν β = 0, και η αύξηση του β διαθέτει ένα αυξανόμενο ποσό του βάρους να ανακλάται στην τελική F-μέτρηση.

- Jaccard δείκτης

  Ο δείκτης Jaccard χρησιμοποιείται για να ποσοτικοποιηθεί η ομοιότητα μεταξύ δύο συνόλων δεδομένων. Ο δείκτης Jaccard παίρνει μια τιμή μεταξύ 0 και 1. Ένας δείκτης του 1 σημαίνει ότι τα δύο σύνολο δεδομένων είναι ταυτόσημα, και ο δείκτης 0 δηλώνει ότι τα σύνολα δεδομένων δεν έχουν κοινά στοιχεία. Ο δείκτης Jaccard ορίζεται από τον ακόλουθο τύπο:

  $$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

  Αυτό είναι απλά ο αριθμός των μοναδικών στοιχείων που είναι κοινά στα δύο σύνολα διαιρούμενο με το συνολικό αριθμό των μοναδικών στοιχείων και στα δύο σύνολα.

- Fowlkes-Mallows δείκτης
- Πίνακας σύγχυσης

  Ένας πίνακας σύγχυσης μπορεί να χρησιμοποιηθεί για να απεικονίσει γρήγορα τα αποτελέσματα του αλγορίθμου διαβάθμισης (ή clustering). Δείχνει πόσο διαφορετικά είναι ένα cluster από το χρυσό πρότυπο των clusters

5.3.11 Σχετικό κριτήριο ποιότητας

Οι μέθοδοι της αξιολόγησης ομαδοποίησης, οι οποίες ενσωματώνουν σχετικό κριτήριο αξιολογούν άμεσα τον αλγόριθμο συγκέντρωσης σε σχέση με τις ανάγκες των χρηστών. Για παράδειγμα, ένας αλγόριθμος μπορεί να εκτελείται αξιοθαύμαστα βασιζόμενος σε διάφορα εσωτερικά κριτήρια, αλλά ο αλγόριθμος μπορεί να είναι αδικαιολόγητα αργός. Εάν ο χρήστης επιδιώκει ταχεία ανταπόκριση της ομαδοποίησης, ένας πιο γρήγορος αλγόριθμος που εκτελείται ελαφρώς φτωχότερα κατά του εσωτερικού κριτηρίου μπορεί να είναι πιο επιθυμητός. Αυτή η μέθοδος αποτίμησης είναι πιο άμεση και απαιτεί προσεκτικά τον ορισμό των «ανάγκες του χρήστη". Η προσέγγιση αυτή μπορεί επίσης να είναι πιο ακριβής, ιδιαίτερα όταν πρόκειται για μεγάλες μελέτες και οφείλει να κατανοεί πλήρως τις προτιμήσεις του χρήστη[23].

5.3.12 Αλγόριθμοι

Τα τελευταία χρόνια σημαντική προσπάθεια καταβλήθηκε για τη βελτίωση των επιδόσεων των αλγόριθμων (Z. Huang, 1998). Ανάμεσα στα πιο δημοφιλή είναι CLARANS (Ng και Han, 1994), DBSCAN και BIRCH (Zhang et al., 1996).

[2]

- CLOPE
- Cobweb
- DBScan
- EM
- FarthestFirst
- FilteredClusterer
- HierarchicalClusterer
- MakeDensityBasedClusterer
- OPTICS
- sIB
- SimpleKMeans
- XMeans

5.4 Επιλογή του χαρακτηριστικού

[25]Στη μηχανική μάθησης και τις στατιστικές , η επιλογή χαρακτηριστικών, επίσης γνωστή και ως επιλογή μεταβλητής, ή μείωση των χαρακτηριστικών, επιλογή χαρακτηριστικού ή επιλογή υποσυνόλου μεταβλητών, είναι η τεχνική της επιλογής ενός υποσυνόλου των σχετικών στοιχείων για την οικοδόμηση ισχυρών μοντέλων μάθησης. Όταν εφαρμόζεται στον τομέα της βιολογίας, η τεχνική αυτή ονομάζεται επίσης διακριτική επιλογή γονίδιου ,η οποία ανιχνεύει γονίδια με επιρροή που βασίζονται στα πειράματα του τομέα του DNA microarray. Με την αφαίρεση των άσχετων και περιττών χαρακτηριστικών από τα στοιχεία, η επιλογή των χαρακτηριστικών συμβάλλει στη βελτίωση της απόδοσης των μαθησιακών μοντέλων από:

- Ανακούφιση από την επίδραση της κατάρα των διαστάσεων.
- Ενίσχυση της ικανότητας γενίκευσης.

- Επιτάχυνση της μαθησιακής διαδικασίας.
- Βελτίωση της ερμηνείας του μοντέλου.

Η επιλογή των χαρακτηριστικών βοηθά επίσης τους ανθρώπους να αποκτήσουν καλύτερη κατανόηση για τα δεδομένα τους, λέγοντάς τους ποια είναι τα σημαντικά χαρακτηριστικά γνωρίσματα και το πώς αυτά σχετίζονται μεταξύ τους.

5.4.1 Εισαγωγή

Οι αλγόριθμοι επιλογής των χαρακτηριστικών είναι οι πιο συνήθεις, αλλά υπάρχουν και πιο μεθοδικές προσεγγίσεις. Από θεωρητική σκοπιά, μπορεί να αποδειχθεί ότι η βέλτιστη επιλογή των χαρακτηριστικών γνωρισμάτων για τα προβλήματα της επιβλεπόμενης μάθησης απαιτεί μια εξαντλητική αναζήτηση όλων των δυνατών υποσυνόλων των χαρακτηριστικών του επιλεγμένου πληθυσμού. Σε περίπτωση που μεγάλος αριθμός χαρακτηριστικών είναι διαθέσιμος, αυτό είναι ανέφικτο. Για πρακτικούς επιβλεπόμενους αλγορίθμους μάθησης, η αναζήτηση γίνεται για ένα ικανοποιητικό σύνολο χαρακτηριστικών αντί ενός βέλτιστου συνόλου.

Οι αλγόριθμοι επιλογής των χαρακτηριστικών κατά κανόνα εμπίπτουν σε δύο κατηγορίες: την κατάταξη χαρακτηριστικού και την επιλογή υποσυνόλου. Η κατάταξη χαρακτηριστικού κατατάσσει τα χαρακτηριστικά γνωρίσματα από μια μέτρηση και εξαλείφει όλες τις λειτουργίες που δεν επιτυγχάνουν την κατάλληλη βαθμολογία. Η επιλογή υποσυνόλου αναζητεί το σύνολο των πιθανών χαρακτηριστικών για το βέλτιστο υποσύνολο.

Στις στατιστικές, η πιο δημοφιλής μορφή της επιλογής χαρακτηριστικού είναι η σταδιακή οπισθοδρόμηση . Είναι ένας άπληστος αλγόριθμος που προσθέτει το καλύτερο χαρακτηριστικό γνώρισμα (ή διαγράφει το χειρότερο χαρακτηριστικό) σε κάθε γύρο. Το κύριο θέμα του ελέγχου είναι να αποφασίσει πότε θα σταματήσει ο αλγόριθμος. Στη μηχανική μάθηση, αυτό γίνεται συνήθως με διασταυρωμένη επικύρωση(cross-validation) . Στις στατιστικές, κάποια κριτήρια έχουν βελτιστοποιηθεί. Αυτό οδηγεί στο εγγενές πρόβλημα της ένθεσης. Περισσότερες αξιόπιστες μέθοδοι έχουν διερευνηθεί, όπως κλάδοι και φραγές (branch and bound) και τμήματα γραμμικού δικτύου.

5.4.2 Επιλογή Υποσυνόλου

[26]Η Επιλογή Υποσυνόλου αξιολογεί ένα υποσύνολο των χαρακτηριστικών γνωρισμάτων ως μια ομάδα για την καταλληλότητα. Οι αλγόριθμοι επιλογής υποσυνόλου μπορούν να χωριστούν σε αλγορίθμους Wrappers, σε αλγορίθμους φίλτρα και σε ενσωματωμένους αλγορίθμους. Οι Wrappers χρησιμοποιούν έναν αλγόριθμο αναζήτησης, για αναζήτηση μέσω του χώρου των πιθανών χαρακτηριστικών και αξιολογούν κάθε υποσύνολο εκτελώντας ένα μοντέλο για το υποσύνολο. Οι αλγόριθμοι Wrappers μπορεί να είναι υπολογιστικά ακριβείς και έχουν έναν κίνδυνο για over fitting στο μοντέλο. Τα φίλτρα είναι παρόμοια με τους Wrappers για την προσέγγιση της αναζήτησης, αλλά αντί για την αξιολόγηση με βάση ένα μοντέλο, αξιολογείται ένα απλούστερο φίλτρο. Οι ενσωματωμένες τεχνικές είναι ενσωματωμένες ειδικά μέσα σε ένα μοντέλο.

Πολλές δημοφιλείς προσεγγίσεις αναζήτησης χρησιμοποιούν την άπληστη αναρρίχηση λόφων , η οποία αξιολογεί επαναληπτικά ένα υποψήφιο υποσύνολο χαρακτηριστικών, στη συνέχεια, τροποποιεί το υποσύνολο και αξιολογεί αν το νέο υποσύνολο αποτελεί βελτίωση του παλιού. Η αξιολόγηση των υποσυνόλων απαιτεί ένα μετρικό σύστημα βαθμολόγησης που θα βαθμολογεί ένα υποσύνολο χαρακτηριστικών γνωρισμάτων. Η εξαντλητική αναζήτηση είναι γενικά ανέφικτη, έτσι σε κάποιο φορέα ορίζεται το σημείο ακινητοποίησης, και το υποσύνολο των χαρακτηριστικών γνωρισμάτων με την υψηλότερη βαθμολογία μέχρι εκείνο το σημείο επιλέγεται ως το ικανοποιητικό υποσύνολο χαρακτηριστικών. Το κριτήριο διακοπής ποικίλλει ανάλογα με τον αλγόριθμο, πιθανά κριτήρια περιλαμβάνουν: ένα σκορ υποσυνόλου που υπερβαίνει ένα κατώτατο όριο, ένα ανώτατο όριο χρόνου εκτέλεσης ενός προγράμματος που έχει ξεπεραστεί, κλπ.

Οι προσεγγίσεις της αναζήτησης περιλαμβάνουν τους εξής αλγορίθμους:

- Εξαντλητικός
- Καλύτερος πρώτος
- Προσομοιωμένη πυράκτωση
- Γενετικός αλγόριθμος
- Άπληστη επιλογής προς τα εμπρός
- Άπληστη εξάλειψη προς τα πίσω

Δύο δημοφιλή φίλτρα μέτρησης για προβλήματα ταξινόμησης είναι η συσχέτιση και η αμοιβαία πληροφόρηση , αν και δεν είναι αληθινές μετρήσεις ή «μέτρα αποστάσεως» με τη μαθηματική έννοια του όρου, αφού αποτυγχάνουν να

υπακούσουν στην τριγωνική ανισότητα και επομένως δεν υπολογίζουν καμία πραγματική «απόσταση» - θα έπρεπε μάλλον να θεωρηθούν «βαθμοί/αποτελέσματα». Αυτά τα αποτελέσματα υπολογίζονται μεταξύ ενός υποψηφίου χαρακτηριστικού (ή του σύνολου στοιχείων) και της επιθυμητής κατηγορίας εξόδου.

Άλλα διαθέσιμα φίλτρα μετρήσεων περιλαμβάνουν:

- Κατηγορία διαχωρισμού
  - Πιθανότητα σφάλματος
  - Ενδο-ταξική απόσταση
  - Πιθανή απόσταση
  - Εντροπία
- Επιλογή χαρακτηριστικών με βάση τη συνέπεια
- Επιλογή χαρακτηριστικών με βάση τη συσχέτιση

5.4.3 Κριτήριο Βέλτιστου

Υπάρχει μια ποικιλία κριτηρίων βέλτιστου που μπορούν να χρησιμοποιηθούν για τον έλεγχο της επιλογής χαρακτηριστικών. Τα παλαιότερα είναι η στατιστική Cp Mallows και το κριτήριο πληροφοριών Akaike (AIC). Αυτά προσθέτουν μεταβλητές αν το t-statistic είναι μεγαλύτερη από $\sqrt{2}$.

Άλλα κριτήρια είναι το Bayesian κριτήριο πληροφοριών (BIC), το οποίο χρησιμοποιεί $\sqrt{\log n}$, ελάχιστο μήκος περιγραφής (MDL), το οποίο χρησιμοποιεί ασυμπτωτικά $\sqrt{\log n}$ αλλά ορισμένοι υποστηρίζουν ότι αυτό το ασύμπτωτο δεν υπολογίζεται σωστά, Bonnferroni / RIC που χρησιμοποιούν $\sqrt{2 \log p}$, και μια ποικιλία από νέα κριτήρια, που βασίζονται σε ψευδές ποσοστό ανακάλυψης (FDR), που χρησιμοποιούν κάτι κοντά σε $\sqrt{2 \log \frac{p}{q}}$.

5.4.4 Ελάχιστος πλεονασμός – Μέγιστη συνάφεια επιλογή χαρακτηριστικού

Τα χαρακτηριστικά μπορούν να επιλεγούν με πολλούς διαφορετικούς τρόπους. Ένας τρόπος είναι να επιλεγούν τα χαρακτηριστικά που συσχετίζονται ισχυρότερα με μεταβλητή ταξινόμησης. Αυτό καλείται μέγιστης σημασίας/συνάφειας επιλογή. Πολλοί ευρετικοί αλγορίθμου μπορούν να χρησιμοποιηθούν, όπως η διαδοχική προς τα εμπρός, προς τα πίσω, ή κυμαινόμενη επιλογή.

Από την άλλη πλευρά, τα χαρακτηριστικά μπορεί να επιλεγούν να είναι αμοιβαία μακριά το ένα από το άλλο, ενώ εξακολουθούν να έχουν "μεγάλη" συσχέτιση με τη μεταβλητή ταξινόμησης. Ο εν λόγω τρόπος, ο οποίος αποκαλείται επιλογή ελάχιστου πλεονασμού– μέγιστης συνάφειας (mRMR),και έχει βρεθεί να είναι πιο ισχυρός από την επιλογή μέγιστου ενδιαφέροντος.

Σαν ειδική περίπτωση, η "συσχέτιση" μπορεί να αντικατασταθεί από τη στατιστική εξάρτηση μεταξύ των μεταβλητών. Η αμοιβαία πληροφορία μπορεί να χρησιμοποιηθεί για την ποσοτικοποίηση της εξάρτησης. Σε αυτή την περίπτωση, αποδεικνύεται ότι η mRMR είναι μια προσέγγιση για την μεγιστοποίηση της εξάρτησης μεταξύ της διανομής των επιλεγμένων λειτουργιών και της μεταβλητής ταξινόμησης.

Ενσωματωμένες μέθοδοι που συνεπάγονται την επιλογής των χαρακτηριστικών

- Τυχαίο πολυωνυμικό logit (Random multinomial logit) (RMNL)
- Αραιά παλινδρόμησης, LASSO
- Δέντρο Αποφάσεων
- Μιμιδικοί αλγόριθμος
- Αυτόματα κωδικοποιημένα δίκτυα με συμφόρηση-layer
- Πολλές άλλες μέθοδοι μηχανικής μάθησης την εφαρμογή βημάτων pruning.

5.4.5 Λογισμικό για την επιλογή των χαρακτηριστικών γνωρισμάτων

Πολλά τυποποιημένο λογισμικά συστήματα ανάλυσης δεδομένων χρησιμοποιούνται συχνά για την επιλογή χαρακτηριστικών, όπως το MATLAB , SciLab , NumPy και η γλώσσα R . Υπάρχουν επίσης συστήματα λογισμικού, ειδικά προσαρμοσμένα για το έργο της επιλογής χαρακτηριστικών:

- Weka - διατίθενται ελεύθερα και είναι ανοιχτού κώδικα λογισμικό σε Java.
- Εργαλειοθήκη Επιλογής Χαρακτηριστικού 3 (Feature Selection Toolbox 3) - διατίθενται ελεύθερα και είναι ανοιχτού κώδικα λογισμικό σε C + +.
- RapidMiner - διατίθενται ελεύθερα και είναι ανοιχτού κώδικα λογισμικό.
- Orange - διατίθενται ελεύθερα και είναι ανοικτού κώδικα λογισμικό (ενότητα orngFSS).
- TOOLDIAG εργαλειοθήκη αναγνώρισης Pattern - ελεύθερα διαθέσιμη εργαλειοθήκη σε C .
- minimum redundancy feature selection tool - διατίθενται ελεύθερα σε C / Matlab κώδικες για την επιλογή των ελάχιστων περιττών χαρακτηριστικών.

- AC # Εφαρμογή της άπληστης επιλογής προς τα εμπρός υποσυνόλου χαρακτηριστικών για διάφορους ταξινομητές (π.χ., LibLinear, SVM-light).
- MCFS-ID (η Μόντε Κάρλο επιλογή χαρακτηριστικών και αλληλεξάρτηση Discovery) είναι ένα εργαλείο βασισμένο στη μέθοδο Monte Carlo για την επιλογή των χαρακτηριστικών γνωρισμάτων. Επιτρέπει, επίσης, την ανακάλυψη των αλληλεξαρτήσεων μεταξύ των σχετικών στοιχείων. Το MCFS-ID είναι ιδιαίτερα κατάλληλο για την ανάλυση των πολλών διαστάσεων, ασαφών συναλλαγών και βιολογικών δεδομένων.

5.4.6 Αλγόριθμοι

[2]

Πίνακας 14: Αλγόριθμοι Επιλογής Χαρακτηριστικού

Subset evaluators

- CfsSubsetEval
- ClassifierSubsetEval
- ConsistencySubsetEval
- CostSensitiveSubsetEval
- FilteredSubsetEval
- WrapperSubsetEval

Attribute evaluators

- ChiSquaredAttributeEval
- ClassifierAttributeEval
- CostSensitiveAttributeEval
- FilteredAttributeEval
- GainRatioAttributeEval
- InfoGainAttributeEval
- OneRAttributeEval
- ReliefFAttributeEval
- SVMAttributeEval
- SymmetricalUncertAttributeEval
- SymmetricalUncertAttributeSetEval

Attribute transformers

- PrincipalComponents (attribute transformer)
- LatentSemanticAnalysis

Search methods

- BestFirst
- ExhaustiveSearch
- FCBFSearch
- GeneticSearch
- GreedyStepwise
- LinearForwardSelection
- RaceSearch
- RandomSearch
- Ranker
- RankSearch
- ScatterSearchV1
- SubsetSizeForwardSelection
- TabuSearch

5.5 Φίλτρα Προ-Επεξεργασίας

[27]Η προ-επεξεργασία δεδομένων είναι ένα συχνά παραμελημένο αλλά σημαντικό βήμα στη διαδικασία εξόρυξης δεδομένων. Η φράση «Garbage in, Garbage Out» ισχύει ιδιαίτερα για τα έργα της εξόρυξης δεδομένων και της μηχανικής μάθησης. Οι μέθοδοι συλλογής δεδομένων συχνά ελέγχονται επιεικώς, με αποτέλεσμα εκτός εύρους τιμών (π.χ., εισόδημα: -100), αδύνατους συνδυασμούς δεδομένων (π.χ., φύλο: άνδρας, έγκυος: Ναι), τιμές που λείπουν , κ.λ.π. Αναλύοντας τα δεδομένα που δεν έχουν προσεκτικά ελεγχθεί μπορούν να παραχθούν παραπλανητικά αποτελέσματα. Έτσι, η παρουσίαση και η ποιότητα των δεδομένων είναι πρωταρχικά, πριν να εκτελεστεί μια ανάλυση.

Εάν υπάρχει πολύ άσχετη και περιττή πληροφορία ή γεμάτα θόρυβο και αναξιόπιστα δεδομένα, στη συνέχεια, η ανακάλυψη της γνώσης κατά τη φάση της εκπαίδευσης είναι πιο δύσκολη. Η προετοιμασία των δεδομένων και τα μέτρα φιλτραρίσματος μπορούν να πάρουν σημαντικό χρόνο επεξεργασίας. Η προετοιμασία των δεδομένων περιλαμβάνει καθαρισμό , τυποποίηση, μεταποίηση, εξαγωγή χαρακτηριστικών και την επιλογή, κ.λπ. Το προϊόν της προ-επεξεργασίας των δεδομένων είναι το τελικό σύνολο εκπαίδευσης.

Η προεπεξεργασία των δεδομένων μπορεί συχνά να έχει σημαντική επίπτωση σχετικά στις επιδόσεις γενίκευσης του επιβλεπόμενου ML αλγόριθμου. Η εξάλειψη των περιπτώσεων με θόρυβο είναι ένα από τα πιο δύσκολα προβλήματα στην επαγωγική ML. Συνήθως όσες έχουν αφαιρεθεί περιέχουν υπερβολικές περιπτώσεις που
έχουν πάρα πολλές μηδενικές τιμές ως χαρακτηριστικό. Τα χαρακτηριστικά που παρεκκλίνουν υπερβολικά αναφέρονται επίσης ως ακραίες τιμές.

Επιπλέον, κοινή προσέγγιση για την αντιμετώπιση της αδυναμίας της μάθησης από πολύ μεγάλα σύνολα δεδομένων είναι να επιλέγει ένα μόνο δείγμα από το μεγάλο σύνολο δεδομένων. Ο χειρισμός της έλλειψης δεδομένων είναι άλλο θέμα που συχνά αντιμετωπίζεται με τα βήματα προετοιμασίας των δεδομένων. Οι αλγόριθμοι της συμβολικής, λογικής μάθησης είναι σε θέση να κινήσουν τη διαδικασία σε συμβολικά, κατηγορικά δεδομένα μόνο. Ωστόσο, στον πραγματικό κόσμο τα προβλήματα αφορούν τόσο συμβολικά όσο και αριθμητικά στοιχεία. Γνωστό είναι το πρόβλημα ότι χαρακτηριστικά με πάρα πολλές τιμές είναι υπερτιμημένα στην διαδικασία επιλογής των πιο κατατοπιστικών χαρακτηριστικά,

τόσο για τη χρήση δέντρων απόφασης όσο και για την αποκρυπτογράφηση των κανόνων απόφασης.

Επιπλέον, στα δεδομένα του πραγματικού κόσμου, η αναπαράσταση των δεδομένων συχνά χρησιμοποιεί πάρα πολλά χαρακτηριστικά, αλλά μόνο λίγα από αυτά μπορούν να είναι σχετικά με την στοχευόμενο concept. Μπορεί να υπάρξουν απολύσεις, όπου ορισμένα χαρακτηριστικά να συσχετίζονται έτσι ώστε να μην είναι απαραίτητο να συμπεριληφθούν όλα τους στη μοντελοποίηση, και αλληλεξάρτηση, όπου δύο ή περισσότερα χαρακτηριστικά μεταξύ τους μεταφέρουν σημαντικές πληροφορίες που είναι ασαφές, αν κάποιο από αυτά περιλαμβάνεται στο ίδιο. Η επιλογή του υποσυνόλου των χαρακτηριστικών είναι η διαδικασία προσδιορισμού και απομάκρυνσης όσο πιο άσχετων και περιττών πληροφοριών είναι δυνατόν. Αυτό μειώνει τη διάσταση των δεδομένων και μπορεί να επιτρέψει στον αλγόριθμο μάθησης να λειτουργεί ταχύτερα και πιο αποτελεσματικά. Σε ορισμένες περιπτώσεις, η ακρίβεια σχετικά με τη μελλοντική ταξινόμηση μπορεί να βελτιωθεί, κατά τα άλλα, το αποτέλεσμα είναι πιο συμπαγές και εύκολο να ερμηνευθεί η παρουσίαση του στοχευόμενου concept. Επιπλέον, το πρόβλημα της αλληλεπίδρασης των χαρακτηριστικών μπορεί να αντιμετωπιστεί με την κατασκευή νέων χαρακτηριστικών από το βασικό σύνολο χαρακτηριστικών. Τα αλλοιωμένα χαρακτηριστικά που προκύπτουν από την κατασκευή χαρακτηριστικών μπορεί να παρέχουν την καλύτερη διακριτική ικανότητα από ότι το καλύτερο υποσύνολο των δεδομένων χαρακτηριστικών.

5.5.1 Επιλογή δείγματος και ανίχνευση ακραίων τιμών

Σε γενικές γραμμές, οι προσεγγίσεις επιλογής δείγματος διακρίνονται μεταξύ φίλτρου[28] και περιτυλίγματος (wrappers)[29]. Το φίλτρο αξιολόγησης θεωρεί σημαντική μόνο τη μείωση των δεδομένων, αλλά δεν λαμβάνει υπόψη τις δραστηριότητες. Σε αντίθεση, οι προσεγγίσεις wrappers υπογραμμίζουν ρητά την πτυχή της ML και αξιολογούν τα αποτελέσματα με τη χρήση του
ειδικού αλγόριθμου ML για την ενεργοποίηση επιλογής χαρακτηριστικού.

Ωστόσο, η επιλογή δείγματος δεν χρησιμοποιείται μόνο για το χειρισμό του θορύβου
αλλά και για την αντιμετώπιση της αδυναμίας της μάθησης από πολύ μεγάλα σύνολα δεδομένων. Η επιλογή δείγματος στην περίπτωση αυτή είναι ένα πρόβλημα βελτιστοποίησης που προσπαθεί να διατηρήσει την ποιότητα της εξόρυξης

καθώς ελαχιστοποιείται το μεγέθους του δείγματος. Μειώνει τα δεδομένα και επιτρέπει σε έναν αλγόριθμο μάθησης να λειτουργεί και να εργάζεται αποτελεσματικά με τεράστιο όγκο δεδομένων. Υπάρχει μια ποικιλία διαδικασιών για την επιλογή δείγματος από μια σειρά μεγάλων συνόλων δεδομένων.

Οι πιο γνωστές είναι[30]:

- Τυχαία δειγματοληψία που επιλέγει ένα υποσύνολο των περιπτώσεων τυχαία.
- Δειγματοληψία σε στρώματα που εφαρμόζεται όταν οι τιμές της τάξης δεν είναι ομοιόμορφα κατανεμημένες στα σύνολα εκπαίδευσης.

Τα δείγματα της μειονότητας των τάξεων επιλέγονται συχνότερα, προκειμένου να εξομαλυνθεί η κατανομή.

Η δειγματοληψία είναι καλώς αποδεκτή από την κοινότητα των στατιστικών, η οποία
παρατηρεί ότι «μια ισχυρά έντονη υπολογιστική διαδικασία που λειτουργεί σε ένα υποσύνολο των δεδομένων μπορεί στην πραγματικότητα να παρέχει ανώτερη ακρίβεια από μια λιγότερο προηγμένη η οποία χρησιμοποιεί μία ολόκληρη βάση δεδομένων. Στην πράξη, όσο η ποσότητα των δεδομένων αυξάνεται, ο ρυθμός αύξησης της ακρίβειας μικραίνει, με αποτέλεσμα τη γνώριμη καμπύλη μάθησης. Το εάν η δειγματοληψία είναι αποτελεσματική εξαρτάται από το πόσο δραματικά επιβραδύνεται ο ρυθμός αύξησης.

Τάξεις που περιλαμβάνουν λίγα δείγματα μπορεί να αγνοηθούν σε μεγάλο βαθμό από τους αλγορίθμους εκμάθησης διότι το κόστος των καλών επιδόσεις για την υπέρ-προβολή της κλάσης υπερκαλύπτει το κόστος μιας πιο φτωχής ενέργειας για τις μικρότερες κλάσεις. Ένας άλλος παράγοντας συμβάλλει στην προκατάληψη είναι το Over-fitting. Το Over-fitting εμφανίζεται όταν ένας αλγόριθμος εκμάθησης δημιουργεί μια υπόθεση που εκτελείται καλά στα δεδομένα εκπαίδευσης, αλλά δεν γενικεύεται καλά στα δεδομένων που δεν έχουν μελετηθεί. Αυτό μπορεί να συμβεί σε μια τάξη υπέρ-προβολής επειδή ο αλγόριθμος εκμάθησης δημιουργεί μια υπόθεση που μπορεί εύκολα να λειτουργήσει σε μικρό αριθμό δειγμάτων, αλλά εφαρμόζεται σε αυτά πολύ συγκεκριμένα. Ανισόρροπα σύνολα δεδομένων έχουν επιστήσει την προσοχή της κοινότητας της μηχανικής μάθησης. Κοινές λύσεις της επιλογής δειγμάτων περιλαμβάνουν:

- Αντιγραφή παραδειγμάτων εκπαίδευσης της τάξη που υπό-προβάλλεται. Αυτό είναι στην ουσία επανάληψη της δειγματοληψίας και αναφέρεται ως υπέρ-δειγματοληψία.

- Κατάργηση παραδειγμάτων εκπαίδευσης της τάξης που υπέρ-προβάλλεται. Αυτό αναφέρεται ως συρρίκνωση, για να δείξει ότι το συνολικό μέγεθος του συνόλου δεδομένων είναι μικρότερο μετά από αυτή την τεχνική εξισορρόπησης που έχει πραγματοποιηθεί.

5.5.2 Τιμές Ελλειπόντων χαρακτηριστικών

Τα ελλιπή δεδομένα αποτελούν αναπόφευκτο πρόβλημα στην αντιμετώπιση των περισσότερων από τις πηγές δεδομένων στον πραγματικό κόσμο. Σε γενικές γραμμές, υπάρχουν ορισμένοι σημαντικοί παράγοντες που πρέπει να λαμβάνονται υπόψη κατά την επεξεργασία τιμών με άγνωστο χαρακτηριστικό. Μία από τις πιο σημαντικές είναι η πηγή “unknownness” :(i) η τιμή λείπει, επειδή ξεχάστηκε ή χάθηκε (ii) ένα συγκεκριμένο χαρακτηριστικό δεν ισχύει για ένα δεδομένο δείγμα, για παράδειγμα, δεν υπάρχει για ένα συγκεκριμένο δείγμα? (iii) για μια δεδομένη παρατήρηση, ο σχεδιαστής του συνόλου εκπαίδευσης δεν ενδιαφέρεται για την αξία της σε ένα συγκεκριμένο χαρακτηριστικό.

Αναλογικά με την υπόθεση, ο εμπειρογνώμονας πρέπει να επιλέξει από έναν αριθμό μεθόδων για το χειρισμό των ελλειπόντων στοιχείων[31]:

- Μέθοδος Αγνόησης δειγμάτων με Αξίες Άγνωστου Χαρακτηριστικό: Αυτή η μέθοδος είναι η πιο απλή: απλά αγνοεί τις περιπτώσεις, οι οποίες έχουν τουλάχιστον μία άγνωστη τιμή χαρακτηριστικού.

- Η πιο συνηθισμένη Αξία χαρακτηριστικού : Η τιμή του χαρακτηριστικού που προκύπτει πιο συχνά επιλέγεται να είναι η τιμή για όλες τις άγνωστες τιμές του χαρακτηριστικού.

- Concept της πιο συνηθισμένης Αξίας χαρακτηριστικού : Αυτή τη φορά η αξία του χαρακτηριστικού, η οποία προκύπτει πιο συχνά μέσα στην ίδια κλάση επιλέγεται να είναι η τιμή για το σύνολο των άγνωστων τιμών του χαρακτηριστικού

- Μέση υποκατάσταση: Αναπληρώνει τη μέση τιμή ενός χαρακτηριστικού υπολογιζόμενο από τις διαθέσιμες περιπτώσεις ώστε να ταιριάζει με τις αξίες των ελλειπόντων στοιχείων στις υπόλοιπες περιπτώσεις. Μια πιο

έξυπνη λύση από ότι η χρήση του «γενικού» μέσου χαρακτηριστικού είναι να χρησιμοποιηθεί το μέσος χαρακτηριστικό για όλα τα δείγματα που ανήκουν στην ίδια κλάση για να ταιριάζουν στην τιμή που λείπει.

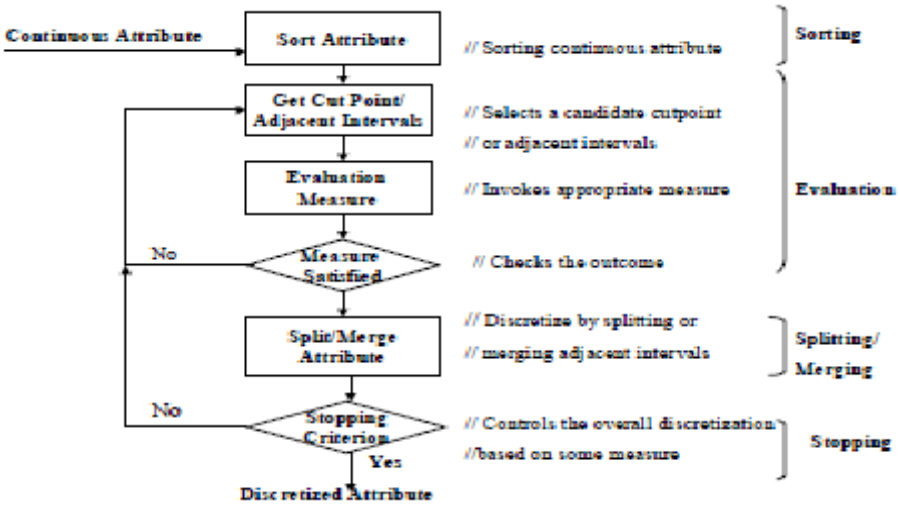
- Παλινδρόμηση ή μέθοδοι ταξινόμησης: Η ανάπτυξη
της παλινδρόμησης ή του μοντέλου ταξινόμησης βασίζεται σε πλήρεις υπόθεση
δεδομένων για ένα συγκεκριμένο χαρακτηριστικό, αντιμετωπίζοντας το ως αποτέλεσμα και χρησιμοποιώντας όλα τα άλλα συναφή χαρακτηριστικά ως προγνωστικά.

- Καταλογισμός Hot καταστρώματος: Εντοπίζει την πιο όμοια περίπτωση στην
περίπτωση με μια τιμή που λείπει και αντικαθιστά την πιο παρεμφερή αξία Υ της περίπτωσης αυτής για της αξία Υ της περίπτωσης που λείπει.

- Μέθοδος αντιμετώπισης των ελλειπόντων τιμών χαρακτηριστικού ως Ειδικής
Τιμής: Αντιμετωπίζει τον "άγνωστο" τον εαυτό της ως μια νέα τιμή για τα χαρακτηριστικά που περιέχουν τιμές που λείπουν.

5.5.3 Διακριτοποίηση

Η διακριτοποίηση μειώνει σημαντικά τον αριθμό των πιθανών τιμών του συνεχούς χαρακτηριστικού καθώς μεγάλος αριθμός των πιθανών τιμών χαρακτηριστικού συμβάλλει στην αργή και αναποτελεσματική διαδικασία της επαγωγικής ML. Το πρόβλημα της επιλογής του μεσοδιαστήματος των συνόρων/ορίων και της ορθής επιλογής του αριθμού των μεταβλητών για τη διακριτοποίηση ενός εύρους αριθμητικών τιμών παραμένει ένα ανοικτό πρόβλημα στον χειρισμό αριθμητικών χαρακτηριστικών . Η τυπική διαδικασία της διακριτοποίησης παρουσιάζεται στο σχήμα. 5.



Σχήμα 5: Διαδικασία Διακριτοποίησης

Γενικά, οι αλγόριθμοι διακριτοποίησης μπορούν να διαιρεθούν σε αλγορίθμους χωρίς επίβλεψη που διακριτοποιούν χαρακτηριστικά χωρίς να λαμβάνουν υπόψη το label της κλάσης και τους αλγορίθμους με επόπτευση που δακριτοποιούν λαμβάνοντας υπόψη το χαρακτηριστικό κλάσης. Η απλούστερη μέθοδος διακριτοποίησης είναι μια άμεση μέθοδος χωρίς επίβλεψη που ονομάζεται διακριτοποίηση ίσου μεγέθους. Υπολογίζει το μέγιστο και το ελάχιστο του χαρακτηριστικού που διακριτοποιείται και χωρίζει το εύρος παρατήρησης σε k ίσου μεγέθους διαστήματα. Η ίση συχνότητα είναι μια άλλη μέθοδος χωρίς επιτήρηση. Μετρά τον αριθμό των τιμών του χαρακτηριστικού που προσπαθούμε να διακριτοποιήσουμε και το χωρίζει σε χρονικά διαστήματα που περιέχουν τον ίδιο αριθμό δειγμάτων.

Οι περισσότερες μέθοδοι διακριτοποίησης διαιρούνται σε “από πάνω προς τα κάτω”(top-down) και“από κάτω προς τα πάνω” (bottom-up) μεθόδους. Οι top-down μέθοδοι ξεκινούν από το αρχικό διάστημα και το διαχωρίζουν αναδρομικά σε μικρότερα χρονικά διαστήματα. Οι Bottom-up μέθοδοι ξεκινούν από το σύνολο του διαστήματος μοναδικής τιμής και συγχωνεύει επαναληπτικά γειτονικά διαστήματα. Μερικές από αυτές τις μεθόδους απαιτούν παραμέτρους από το χρήστη για να τροποποιήσουν τη συμπεριφορά του κριτηρίου διακριτοποίησης ή να δημιουργήσουν ένα όριο για τον κανόνα διακοπής.
Στατικές μεθόδους, όπως το binning και αυτές που βασίζονται στην εντροπία, καθορίζουν τον αριθμό των κατατμήσεις για κάθε ανεξάρτητο χαρακτηριστικό από τα άλλα χαρακτηριστικά. Από την άλλη πλευρά,οι δυναμικές μεθόδους διεξάγουν

μια έρευνα μέσα στο χώρο των πιθανών k κατατμήσεων για όλες τα χαρακτηριστικά ταυτόχρονα, με αποτέλεσμα να την καταγραφή των αλληλεξαρτήσεων στη διακριτοποίηση χαρακτηριστικών. Ωστόσο, αναφέρεται ότι δεν υπάρχει σημαντική βελτίωση στη χρήση δυναμικής διακριτοποίησης σε σχέση με τις στατικές μεθόδους.

5.5.4 Κανονικοποίηση δεδομένων

Η Κανονικοποίηση είναι μια μεταμόρφωση "αποκλιμάκωσης" των χαρακτηριστικών. Μέσα σε ένα χαρακτηριστικό υπάρχει συχνά μεγάλη διαφορά μεταξύ της μέγιστης και ελάχιστης τιμής, π.χ. 0,01 και 1000. Όταν εφαρμόζεται η κανονικοποίηση τα μεγέθη των τιμών διαβαθμίζονται σε αισθητά χαμηλές τιμές. Αυτό είναι σημαντικό για πολλά νευρωνικά δίκτυα και αλγορίθμους k- Κοντινότερης Γειτονίας. Οι δύο πιο κοινές μέθοδοι για αυτό το πεδίο είναι:

- κανονικοποίηση min-max:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

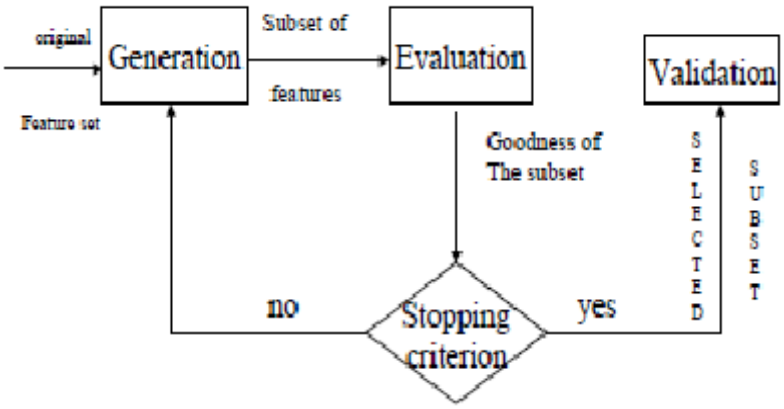
- κανονικοποίηση z-score:

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

όπου *v* είναι η παλιά τιμή του χαρακτηριστικού και *v′* το νέο.

5.5.5 Επιλογή Χαρακτηριστικού

Η επιλογή υποσυνόλου χαρακτηριστικών είναι η διαδικασία εντοπισμού και αφαίρεσης όσο άσχετων και περιττών χαρακτηριστικών, είναι δυνατον. Αυτό μειώνει η διάσταση των δεδομένων και δίνει τη δυνατότητα στους αλγόριθμους μάθησης να λειτουργούν ταχύτερα και πιο αποτελεσματικά. Σε γενικές γραμμές, τα χαρακτηριστικά αναφέρονται ως:
• Σχετικά: Αυτά τα χαρακτηριστικά έχουν επιρροή στο αποτέλεσμα και ο ρόλος τους δεν μπορεί να ληφθεί υπόψιν από τα υπόλοιπα.
• Άσχετα: Τα άσχετα χαρακτηριστικά ορίζονται ως εκείνα τα χαρακτηριστικά που δεν έχουν καμία επιρροή στο αποτέλεσμα, και των οποίων οι τιμές δημιουργούνται τυχαία για κάθε παράδειγμα.
• Εφεδρικά: Ένας πλεονασμός υπάρχει κάθε φορά που ένα χαρακτηριστικό μπορεί να αναλάβει το ρόλο του άλλου (ίσως και ο απλούστερος τρόπος για να το εφεδρικό μοντέλο).



Σχήμα 6: Επιλογή Υποσυνόλου Χαρακτηριστικών

Οι αλγόριθμοι επιλογής χαρακτηριστικών σε γενικές γραμμές έχουν δύο συστατικά στοιχεία: έναν αλγόριθμο επιλογής που δημιουργεί προτεινόμενα υποσύνολα των χαρακτηριστικών και προσπαθεί να βρει το βέλτιστο υποσύνολο, καθώς και έναν αλγόριθμο αξιολόγησης που καθορίζει πόσο «καλό» είναι το προτεινόμενο υποσύνολο χαρακτηριστικών, επιστρέφοντας κάποιο μέτρο της ποιότητας στον αλγόριθμο επιλογής. Ωστόσο, χωρίς κατάλληλο κριτήριο διακοπής η διαδικασία επιλογής χαρακτηριστικού μπορεί να τρέξει εξαντλητικά ή για πάντα μέσα από το χώρο των υποσυνόλων. Το κριτήριο διακοπής μπορεί να είναι: (i) κατά πόσο η προσθήκη (ή
διαγραφή) κάθε χαρακτηριστικού, δεν παράγει καλύτερο υποσύνολο και
(ii) κατά πόσο ένα βέλτιστο υποσύνολο επιτυγχάνεται σύμφωνα με ορισμένες λειτουργίες αξιολόγησης.

Διάφορες μέθοδοι επιλογής χαρακτηριστικών συγκεντρώνονται σε δύο ευρείες ομάδες (δηλαδή, το φίλτρο και το [wrappers] περιτύλιγμα) με βάση την εξάρτησή τους από
τον επαγωγικό αλγόριθμο που θα χρησιμοποιήσει τελικά το επιλεγμένο υποσύνολο. Οι μέθοδοι φίλτρου είναι ανεξάρτητες από τον επαγωγικό αλγόριθμο ως λειτουργία αξιολόγησης. Το φίλτρο αξιολόγησης χαρακτηριστικών μπορεί να χωριστεί σε τέσσερις κατηγορίες: της απόστασης, της πληροφόρησης, της εξάρτησης και της συνέπειας.

• Απόσταση: Για προβλήματα δύο κλάσεων, ένα χαρακτηριστικό γνώρισμα X προτιμάται σε ένα άλλο χαρακτηριστικό γνώρισμα Y αν το X προκαλεί μεγαλύτερη διαφορά μεταξύ των δύο κλάσεων υπό όρους πιθανοτήτων του Y.
• Πληροφορίες: Το X χαρακτηριστικό προτιμάται από το Y χαρακτηριστικό αν οι

πληροφορίες από το Χ χαρακτηριστικό είναι μεγαλύτερες από εκείνες του χαρακτηριστικού Υ.
• Εξάρτηση: Ο συντελεστής είναι ένα κλασικό μέτρο εξάρτησης και μπορεί να χρησιμοποιηθεί για να βρεθεί η αντιστοιχία μεταξύ ενός χαρακτηριστικού και μίας κλάσης. Εάν η συσχέτιση του Χ χαρακτηριστικού με την κλάση C είναι υψηλότερη από το συσχετισμό του χαρακτηριστικού Υ με τη C, τότε το Χ χαρακτηριστικό είναι προτιμότερο από το Υ.
• Συνέπεια: δύο δείγματα είναι σε σύγκρουση, εάν έχουν τις ίδιες τιμές για ένα υποσύνολο χαρακτηριστικών, αλλά διαφωνούν στην κλάση που εκπροσωπούν.

5.5.6 Κατασκευή Χαρακτηριστικών

Το πρόβλημα της αλληλεπίδρασης των χαρακτηριστικών μπορεί επίσης να αντιμετωπιστεί με την κατασκευή νέων χαρακτηριστικών από το βασικό σύνολο χαρακτηριστικών γνωρισμάτων. Αυτή η τεχνική ονομάζεται κατασκευή χαρακτηριστικών/μετασχηματισμός. Η νέες δυνατότητες που δημιουργούνται μπορεί να οδηγήσουν στη δημιουργία περισσότερων συνοπτικών και ακριβών ταξινομητών. Επιπλέον, η ανακάλυψη
ουσιαστικών χαρακτηριστικών συμβάλλει στην καλύτερη κατανόηση του παραγόμενου ταξινομητή, και την καλύτερη κατανόηση της έννοιας της μάθησης.

5.5.7 Επίλογος

Οι αλγόριθμοι της μηχανικής μάθησης αυτόματα βγάζουν εκχύλισμα γνώσης από πληροφορίες αναγνώσιμες από μηχανήματα. Δυστυχώς, η επιτυχία τους συνήθως εξαρτάται από την ποιότητα των δεδομένων πάνω στα οποία λειτουργούν. Εάν τα δεδομένα είναι ανεπαρκή, ή περιέχουν ξένες και άσχετες πληροφορίες, οι αλγόριθμοι μηχανικής μάθησης μπορούν να παράγουν λιγότερο ακριβή και λιγότερο κατανοητά αποτελέσματα, ή μπορεί να αποτύχουν να ανακαλύψει οτιδήποτε χρήσιμο. Έτσι, η προ-επεξεργασία δεδομένων είναι ένα σημαντικό βήμα στη μηχανική μαθησιακή διαδικασία. Το βήμα της προ-επεξεργασίας είναι απαραίτητο στην επίλυση πολλών ειδών προβλήματα περιλαμβανομένου των δεδομένων με θόρυβο, των πλεοναζόντων δεδομένων, των εκλειπόντων τιμών δεδομένων, κλπ. Όλα οι επαγωγικοί αλγόριθμοι μάθησης βασίζονται σε μεγάλο βαθμό στο προϊόν του σταδίου αυτού, που είναι το τελικό σύνολο εκπαίδευσης.

Με την επιλογή των συναφών περιπτώσεων, οι ειδικοί μπορούν συνήθως να αφαιρέσουν τα άσχετα καθώς και αυτά με θόρυβο ή/και τα πλεονάζοντα δεδομένα. Η υψηλή ποιότητα των δεδομένων θα οδηγήσει σε αποτελέσματα υψηλής ποιότητας και μειωμένου κόστος για την εξόρυξη δεδομένων. Επιπλέον, όταν ένα σύνολο δεδομένων είναι πολύ μεγάλο, μπορεί να μην είναι δυνατόν να τρέξετε ένα αλγόριθμο ML. Σε αυτή την περίπτωση, η Επιλογή δείγματος μειώνει τα δεδομένα και επιτρέπει σε έναν αλγόριθμο ML να λειτουργεί αποτελεσματικά με τεράστια δεδομένα.
Στις περισσότερες περιπτώσεις, τα δεδομένα που λείπουν θα πρέπει να προ-επεξεργαστούν, έτσι ώστε να επιτρέψει στο σύνολο δεδομένων υποβληθεί σε επεξεργασία από έναν επιβλεπόμενο αλγόριθμο ML. Επιπλέον, οι περισσότεροι από τους υπάρχοντες αλγορίθμους ML είναι σε θέση να εξάγουν γνώση από σύνολα δεδομένα που αποθηκεύουν διακριτά χαρακτηριστικά. Εάν τα χαρακτηριστικά είναι συνεχόμενα, οι αλγόριθμοι μπορούν να ενσωματώνονται με έναν αλγόριθμο διακριτοποίησης που τους μετατρέπει σε διακριτά χαρακτηριστικά. Μια σειρά μελετών συγκρίνοντας τα αποτελέσματα από τη χρήση διαφόρων τεχνικών διακριτοποίησης (σχετικά με τους κοινούς τομείς και αλγορίθμους ML) έχουν βρει ότι οι μέθοδοι που βασίζονται στην εντροπία είναι συνολικά οι ανώτεροι.

Η επιλογή υποσυνόλου είναι η διαδικασία εντοπισμού και αφαίρεσης όσο άσχετων και περιττών πληροφοριών είναι δυνατό. Τα wrappers χαρακτηριστικών επιτυγχάνουν συχνά καλύτερα αποτελέσματα από ότι τα φίλτρα και οφείλεται στο γεγονός ότι είναι συντονισμένα με ειδικές αλληλεπιδράσεις μεταξύ ενός αλγόριθμου επαγωγής και της κατάρτισης δεδομένων. Εντούτοις, είναι πολύ πιο αργή από ότι τα φίλτρα χαρακτηριστικών.
Επιπλέον, το πρόβλημα της αλληλεπίδρασης χαρακτηριστικών μπορεί να αντιμετωπιστεί με την κατασκευή νέων χαρακτηριστικών από το βασικό σύνολο χαρακτηριστικών (Κατασκευή χαρακτηριστικού). Σε γενικές γραμμές, τα μεταμορφωμένα χαρακτηριστικά που δημιουργούνται από την κατασκευή χαρακτηριστικών μπορεί να παρέχουν καλύτερη διακριτική ικανότητα από ότι το καλύτερο υποσύνολο των χαρακτηριστικών, αλλά αυτά τα νέα χαρακτηριστικά γνωρίσματα δεν μπορούν να έχουν μια ξεκάθαρη φυσική έννοια.
Θα ήταν ωραίο αν μια μονή ακολουθία αλγόριθμου προ-επεξεργασίας δεδομένων είχε τις καλύτερες επιδόσεις για κάθε σύνολο δεδομένων, αλλά αυτό δεν συμβαίνει.

[2]

Πίνακας 15: Αλγόριθμοι Φίλτρων Προεπεξεργασίας

Supervised instance-based

- Resample
- SpreadSubsample
- StratifiedRemoveFolds
- SMOTE

Supervised attribute-based

- AddClassification
- AttributeSelection
- ClassOrder
- Discretize
- NominalToBinary
- PLSFilter

Unsupervised instance-based

- Denormalize
- NonSparseToSparse
- Normalize (instance)
- Randomize
- RemoveFolds
- RemoveFrequentValues
- RemoveMisclassified
- RemovePercentage
- RemoveRange
- RemoveWithValues
- Resample (unsupervised)
- ReservoirSample
- SparseToNonSparse
- SubsetByExpression

Unsupervised attribute-based

- Add
- AddCluster
- AddExpression
- AddID
- AddNoise
- AddValues
- Center
- ChangeDateFormat
- ClassAssigner
- ClusterMembership
- Copy
- Discretize (unsupervised)
- EMImputation
- FirstOrder
- InterquartileRange
- KernelFilter
- MakeIndicator
- MathExpression
- MergeManyValues
- MergeTwoValues
- MILESFilter
- MultiInstanceToPropositional
- NominalToBinary (unsupervised)
- NominalToString
- Normalize (attribute)
- NumericCleaner
- NumericToBinary
- NumericToNominal
- NumericTransform
- Obfuscate
- PartitionedMultiFilter
- PKIDiscretize
- PrincipalComponents (filter)
- PropositionalToMultiInstance
- RandomProjection
- RandomSubset

- RELAGGS
- Remove
- RemoveByName
- RemoveType
- RemoveUseless
- RenameAttribute
- Reorder
- ReplaceMissingValues
- SortLabels
- Standardize
- StringToNominal
- StringToWordVector
- SwapValues
- TimeSeriesDelta
- TimeSeriesTranslate
- Wavelet

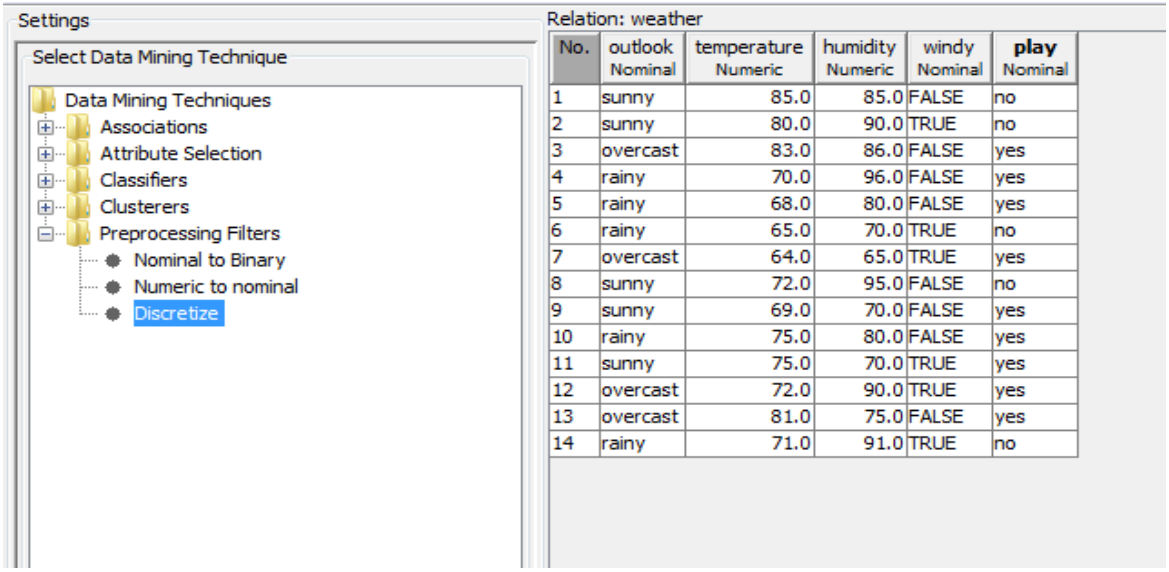
ΚΕΦΑΛΑΙΟ 6

Εξόρυξη δεδομένων με τη χρήση της custom εφαρμογής που δημιουργήθηκε με την υλοποίηση του Java Api του Weka.

ΕΙΣΑΓΩΓΗ

Για να δημιουργήσουμε την εφαρμογή μας στο weka έπρεπε να ερευνήσουμε το api του και να βρούμε τα αντικείμενα που υλοποιούν τις μεθόδους data mining που αναφέραμε κατά στα προηγούμενα κεφάλαια. Παρακάτω παρουσιάζουμε έναν πίνακα με όλα τα αντικείμενα τύπου Java. Μετά το βήμα αυτό θα προχωρήσουμε στην περαιτέρω ανάλυση μερικών από αυτών τα οποία χρησιμοποιήθηκαν και στην εφαρμογή μας.

6.1 Κανόνες συσχέτισης και Φίλτρα Προεπεξεργασίας

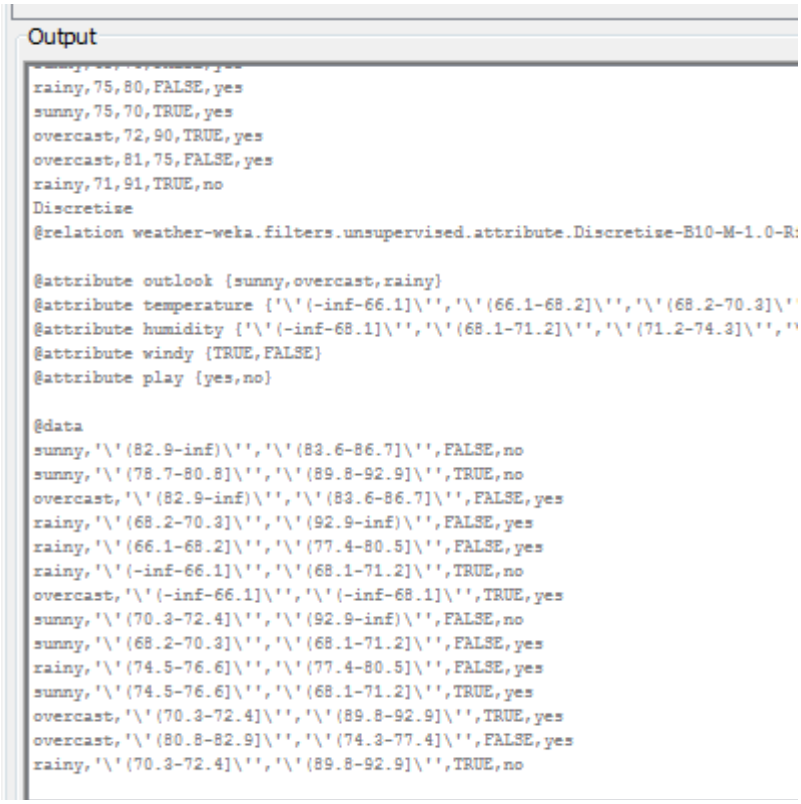


No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Σχήμα 7: Διακριτοποίηση

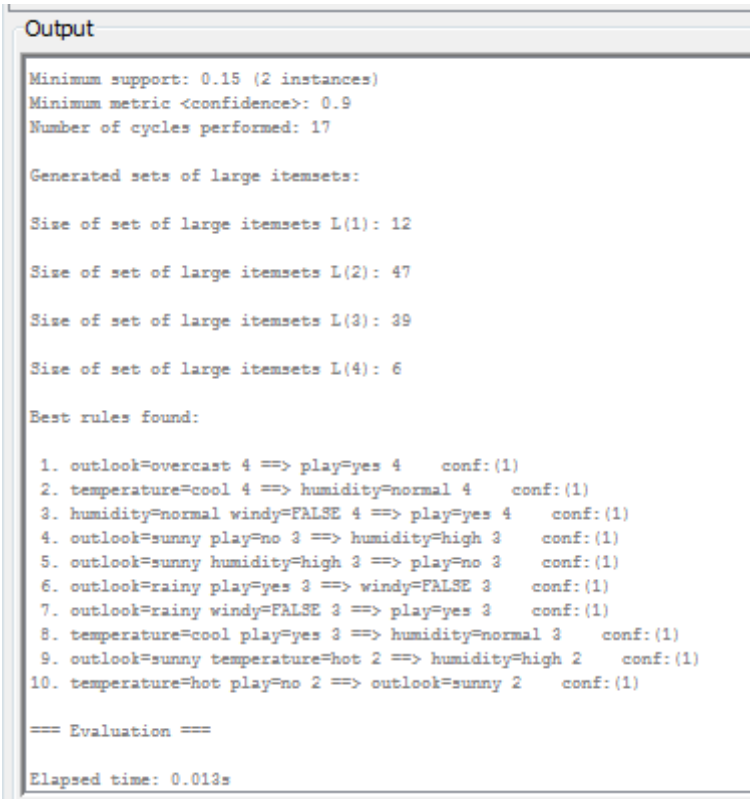
Το ζητούμενο στο παρακάτω πείραμα είναι να βγάλουμε κανόνες συσχέτισης οι οποίοι θα μας δείχνουν ποιες μέρες μπορούμε να παίξουμε ή όχι τέννις ανάλογα με τις κλιματικές συνθήκες που επικρατούν. Αφού τα δεδομένα μας περιέχουν τόσο αριθμητικά όσο και μη αριθμητικά στοιχεία και για να εφαρμόσουμε τον αλγόριθμο Apriori πρέπει να μην υπάρχουν αριθμητικά δεδομένα, θα χρειαστεί να κάνουμε προεπεξεργασία για την μετατροπή τους σε μη αριθμητικά. Εκτός αυτού

στα δεδομένα θα εφαρμόσουμε και άλλον αλγόριθμο προεπεξεργασίας για διακριτοποίηση των θερμοκρασιών ώστε να μειώσουμε τους εξαγόμενους κανόνες.



Σχήμα 8: Output διακριτοποίησης

Αποθηκεύουμε σε αρχείο .arff το output και φορτώνουμε τα προεπεξεργασμένα δεδομένα για να εξάγουμε τους κανόνες συσχέτισης. Οι κανόνες που προκύπτουν από τον αλγόριθμο Apriori είναι οι παρακάτω, από τους οποίους μας ενδιαφέρουν μόνο όσοι έχουν ως αποτέλεσμα yes ή no στο χαρακτηριστικό play που μας λέει αν μπορούμε να παίξουμε τέννις ή όχι. Το output με τους κανόνες το αποθηκεύουμε ως txt για περαιτέρω χρήση.



Σχήμα 9: Output Apriori

Για να εφαρμόσουμε τον αλγόριθμο Tertius ακολουθούμε την ίδια διαδικασία προεπεξεργασίας με τον Apriori και οι κανόνες που παίρνουμε είναι οι εξής:

Output

```
 1. /* 0,633754 0,071429 */ play = yes ==> humidity = normal or outlook = overcast
 2. /* 0,607625 0,000000 */ humidity = normal ==> temperature = cool or play = yes
 3. /* 0,607625 0,000000 */ temperature = cool ==> humidity = normal
 4. /* 0,594071 0,214286 */ humidity = normal ==> temperature = cool
 5. /* 0,590214 0,000000 */ humidity = high and outlook = sunny ==> play = no
 6. /* 0,555556 0,000000 */ play = no ==> windy = TRUE or outlook = sunny
 7. /* 0,486606 0,000000 */ play = no and outlook = sunny ==> humidity = high
 8. /* 0,486606 0,000000 */ humidity = normal ==> play = yes or outlook = rainy
 9. /* 0,469374 0,000000 */ outlook = overcast ==> play = yes
10. /* 0,469374 0,000000 */ windy = FALSE and outlook = overcast ==> temperature = hot
11. /* 0,469374 0,000000 */ outlook = overcast ==> temperature = hot or windy = TRUE
12. /* 0,469374 0,000000 */ temperature = hot and play = yes ==> outlook = overcast
13. /* 0,469374 0,000000 */ play = no ==> humidity = high or windy = TRUE
14. /* 0,469374 0,000000 */ temperature = hot ==> play = no or outlook = overcast
15. /* 0,469374 0,000000 */ temperature = hot ==> humidity = high or outlook = overcast
16. /* 0,469374 0,000000 */ humidity = high and play = no ==> temperature = mild or outlook = sunny
17. /* 0,469374 0,000000 */ temperature = mild and play = yes ==> windy = TRUE or outlook = rainy
18. /* 0,469374 0,000000 */ outlook = sunny ==> temperature = cool or windy = TRUE or play = no
19. /* 0,467119 0,357143 */ play = yes ==> outlook = overcast
20. /* 0,458333 0,071429 */ play = yes ==> windy = FALSE or outlook = overcast
21. /* 0,458333 0,071429 */ humidity = high and play = no ==> outlook = sunny
22. /* 0,439100 0,071429 */ play = no ==> humidity = high
23. /* 0,439100 0,071429 */ humidity = high ==> temperature = mild or play = no
24. /* 0,439100 0,071429 */ humidity = high ==> temperature = mild or outlook = sunny

Number of hypotheses considered: 1724
Number of hypotheses explored: 689
=== Evaluation ===
```

Σχήμα 10: Output Tertius

Όπως διαπιστώνουμε και με μια πρώτη ματιά οι κανόνες αυτοί είναι πολύ διαφορετικοί από εκείνους που πήραμε με τον Apriori. Οι κανόνες που παράγονται δεν είναι ακριβείς (δεν έχουμε πληροφορίες για την ακρίβεια αυτή) και δεν μας βοηθούν στην έρευνά μας.

6.2 Ταξινόμηση (Classifiers)

Iris Δεδομένα

Το πρόβλημα μας περιέχει δεδομένα μετρήσεων που αποτελούνται από 150 φυτά τριών ειδών της οικογένειας iris, το iris setosa, iris versicolor, και iris virginica. Τις μετρήσεις για τα δεδομένα τις έχει πάρει για πρώτη φορά ο Anderson Fisher το 1936. Κάθε είδος των προαναφερθέντων iris μπορεί να ταξινομηθεί ανάλογα με το μήκος και πλάτος των σέπαλων του, και το μήκος και πλάτος των πετάλων του.

Παρακάτω βλέπουμε ένα μικρό κομμάτι από τα δεδομένα.

Πλάτος πετάλων	Μήκος πετάλων	Πλάτος σέπαλων	Μήκος πετάλων	Είδος Iris
2	14	33	50	0 (setosa)
21	54	31	69	1 (versicolor)
13	40	25	55	2 (virginica)

Πίνακας 16: Δεδομένα Iris

Στα παραπάνω δεδομένα θα εφαρμόσουμε αλγορίθμους ταξινόμησης με σκοπό να ταξινομήσουμε το δείγμα μας σε ξεχωριστές κατηγορίες ανάλογα με το μέγεθος τον πετάλων και σέπαλων του.

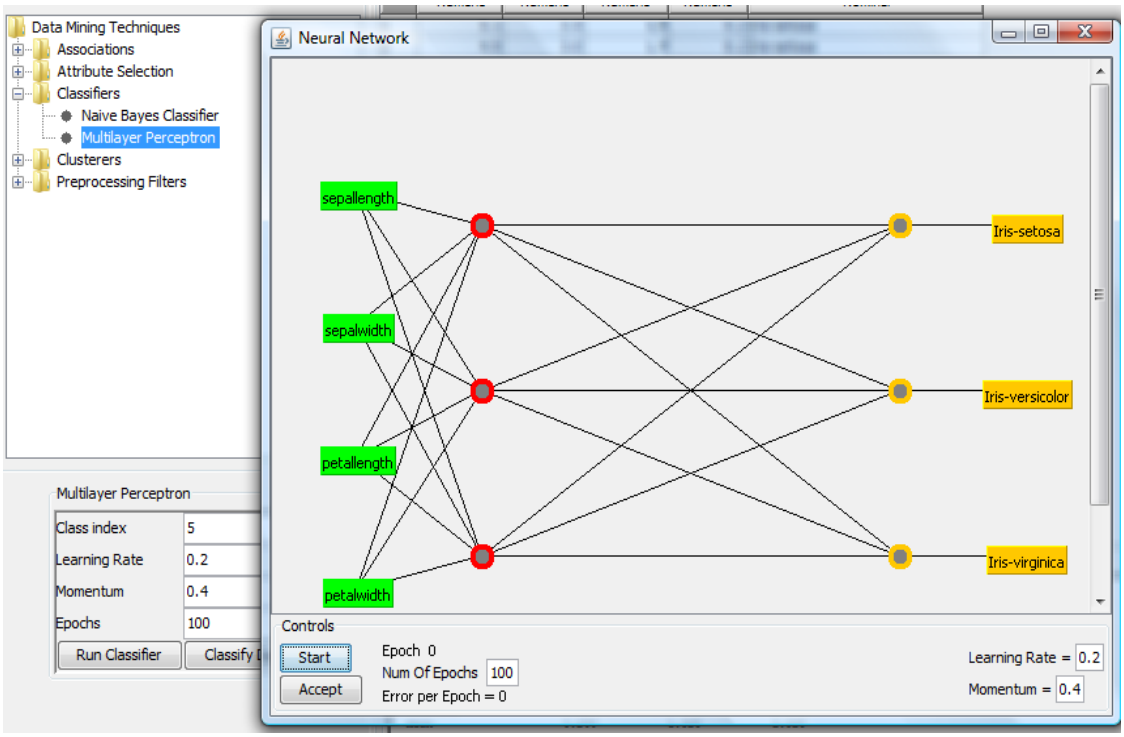
Output

```
  weight sum                50                50                50
  precision               0.01              0.01              0.01

sepalwidth
  mean                   3.418              2.77             2.974
  std. dev.             0.3772            0.3106            0.3193
  weight sum                50                50                50
  precision               0.01              0.01              0.01

petallength
  mean                   1.464              4.26             5.552
  std. dev.             0.1718            0.4652            0.5463
  weight sum                50                50                50
  precision               0.01              0.01              0.01

petalwidth
  mean                   0.244             1.326             2.026
  std. dev.             0.1061            0.1958            0.2719
  weight sum                50                50                50
  precision               0.01              0.01              0.01


=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

Σχήμα 11: Output Ταξινόμησης

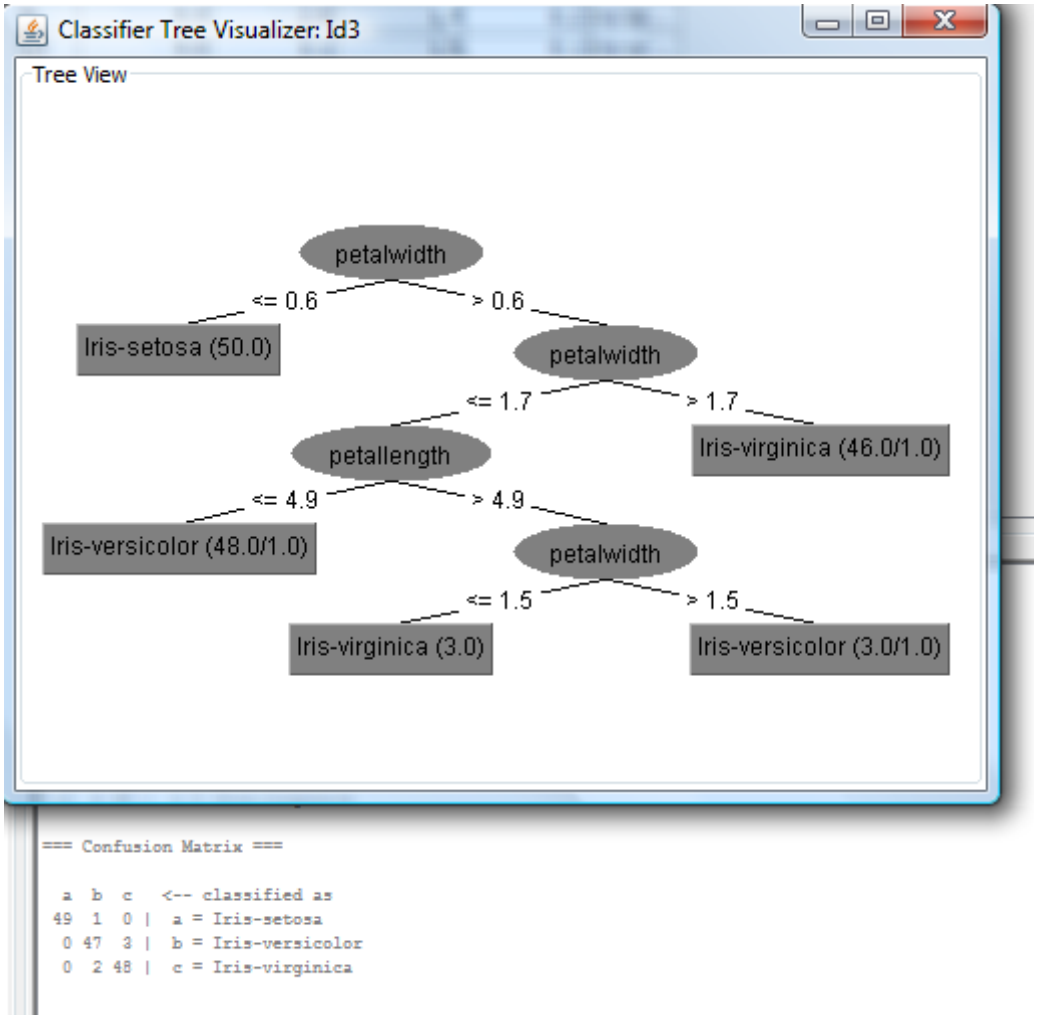
Όπως βλέπουμε στο confusion matrix, ο αλγόριθμος Naive Bayes ταξινόμησε τα iris σε 3 κατηγορίες.

Με τον αλγόριθμο Multilayer Perceptron δημιουργείται ένα νευρωνικό δίκτυο από τα δεδομένα που έχουμε, και παρουσιάζει ποια ανήκουν σε ποια κλάση.



Σχήμα 12: Multilayer Perceptron

Με το Start βάζουμε τα δεδομένα στη διαδικασία εκπαίδευσής τους, τα οποία θα τρέχουν για όσες “εποχές” ορίσουμε από τις παραμέτρους, καθώς και το learning rate για την προσαρμογή στα δεδομένα μετά από κάθε εποχή και το momentun για να ξεπερνά μικροπροβλήματα όπως το θόρυβο στα δεδομένα.
Ο επόμενος αλγόριθμος είναι το J48 Tree που υπερτερεί σε σχέση με τους άλλους καθώς βλέπουμε και με γραφικό τρόπο τα βήματα που ακολουθεί για το αποτέλεσμα που μας δίνει στην έξοδο στον confusion matrix.



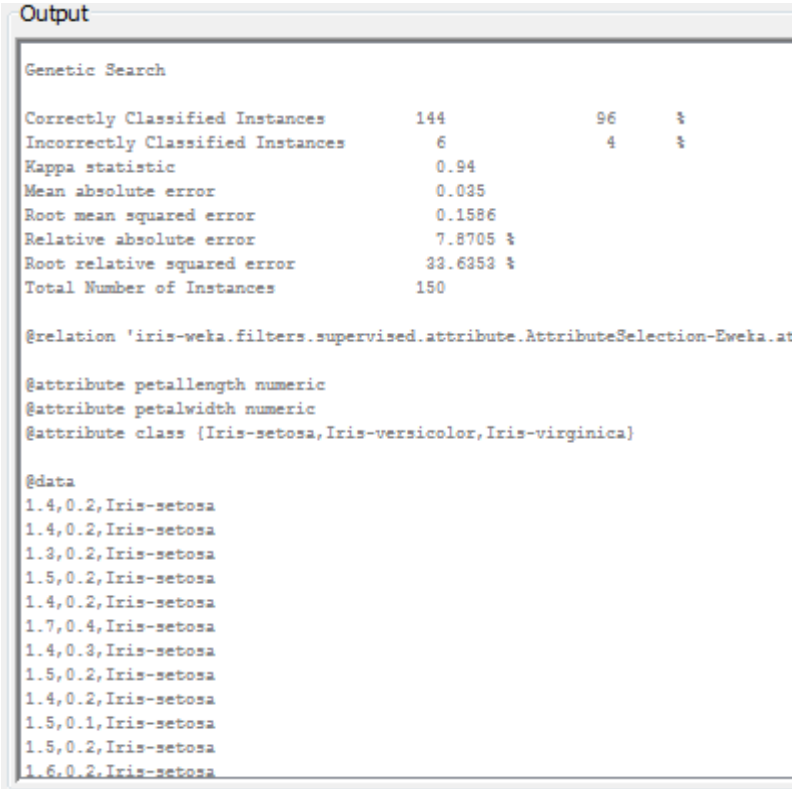
Σχήμα 13: J48 Tree

6.3 Επιλογή Χαρακτηριστικού (Attribute Selection)

Στη συνέχεια παρουσιάζουμε την εφαρμογή αλγορίθμων για επιλογή χαρακτηριστικού όπως ο Generic Search και ο Greedy Search.

Ο Generic Search χρησιμοποιεί έναν γενετικό αλγόριθμο και εφαρμόζει το πρόβλημα με τη θεωρία της εξέλιξης σε έναν πληθυσμό μεγέθους όσο το population size. Στη διαδικασία αυτή αφαιρούνται χαρακτηριστικά από το σύνολο δεδομένων που δε θεωρούνται σημαντικά. Κάθε άτομο στον πληθυσμό μπορεί να μεταλλαχτεί από το να αφαιρεί ένα στοιχείο από τα δεδομένα σε άλλο με την πιθανότητα που ορίζεται στο mutation rate. Τα άτομα στον πληθυσμό αναπαράγονται μεταξύ τους και έχουν πιθανότητα να αλλάξουν δεδομένα όσο δίνει

το crossover. Ο αλγόριθμος επαναλαμβάνεται για έναν αριθμό κύκλων όσο το generation no.



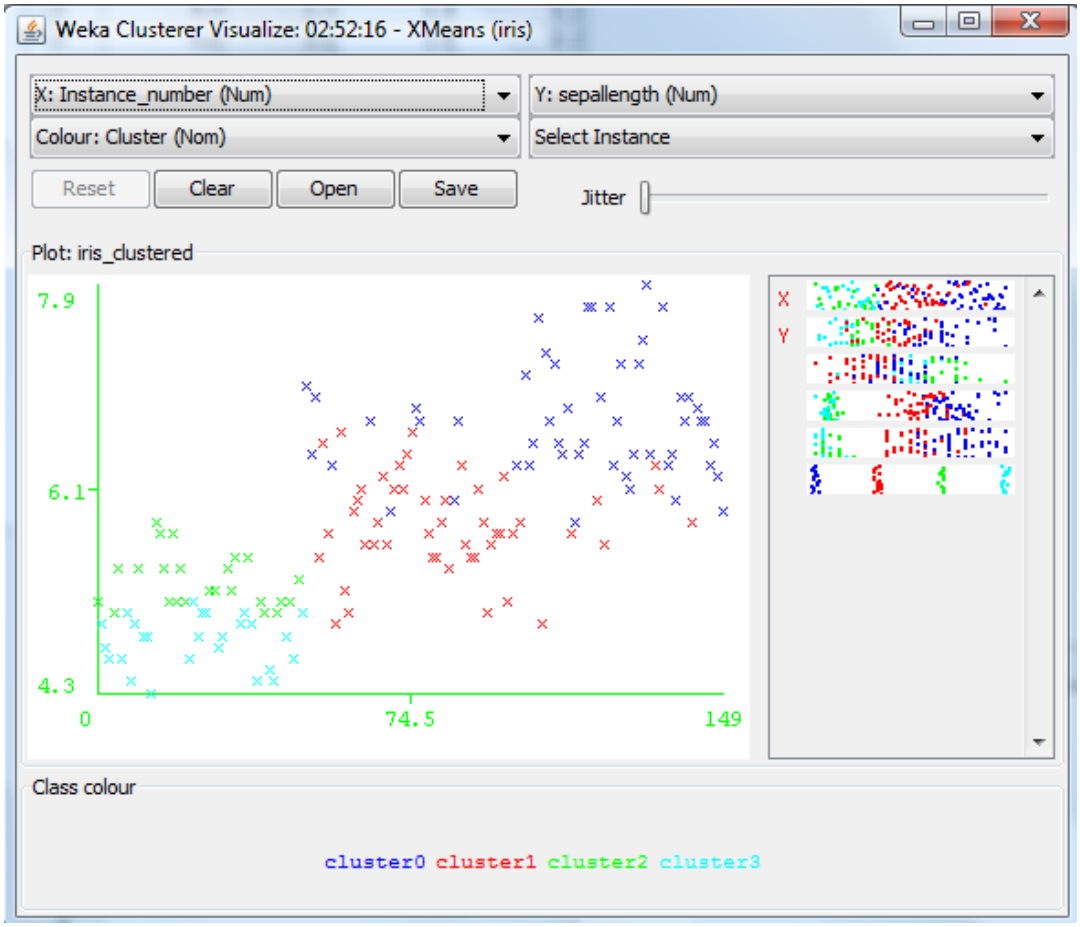
Σχήμα 14: Output Attribute Selection

Ο αλγόριθμος greedy όπως λέει και το όνομά του “άπληστος” ψάχνει όλες τις πιθανές προσθαφαιρέσεις χαρακτηριστικών. Αν και οι δυο αυτοί αλγόριθμοι μας δίνουν το ίδιο αποτέλεσμα, ο greedy μπορεί να εγγυηθεί πως η λύση που προσφέρει είναι η καλύτερη δυνατή, κάτι που ο generic δεν κάνει, όμως είναι εξαιρετικά αργός και χρονοβόρος σε σύνολα δεδομένων με πολλές στήλες/χαρακτηριστικά.

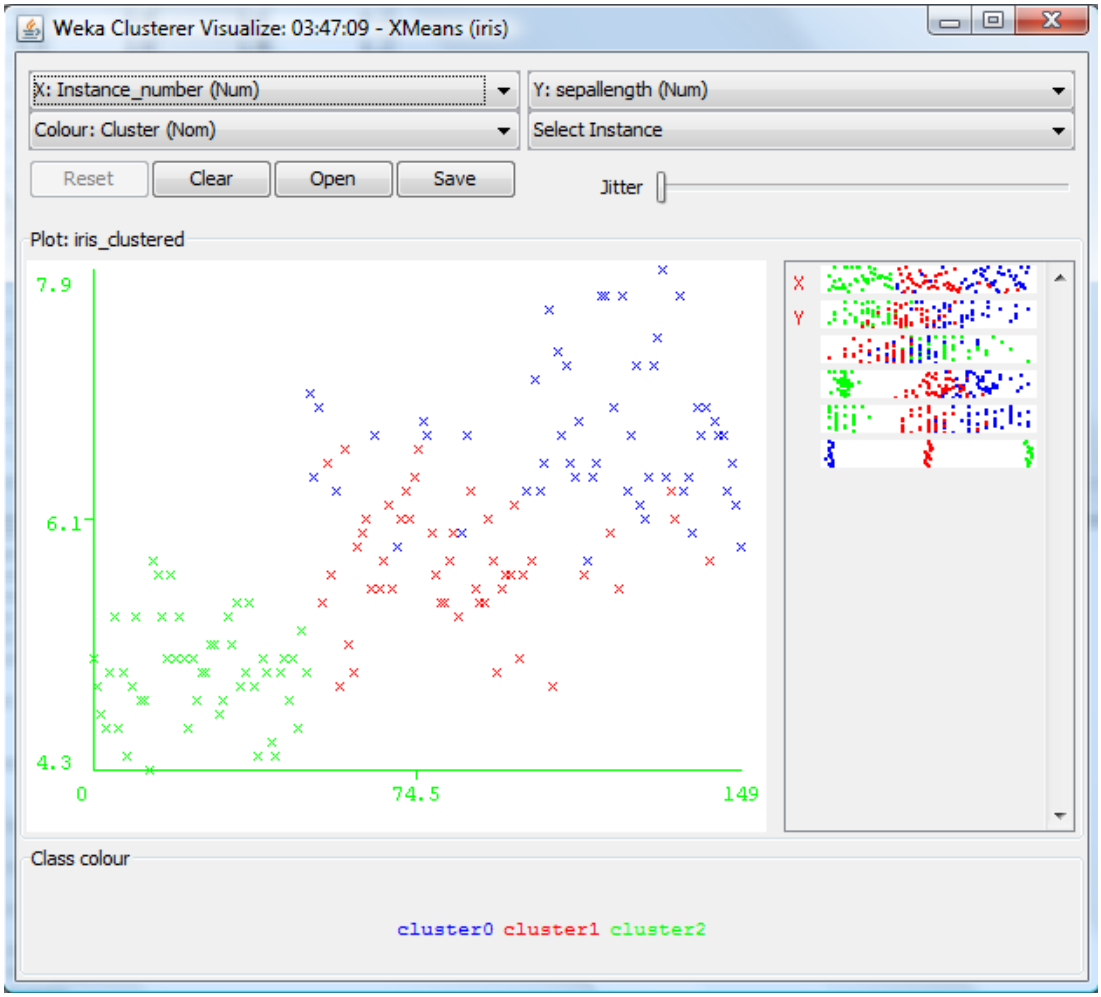
6.4 Ανάλυση κατά συστάδες (Clusters)

Τελευταία κατηγορία αλγόριθμων στη demo εφαρμογή μας είναι τα Clusters(ή Συστάδες) και υλοποιούμε δυο αλγορίθμους για ανάλυση κατά συστάδες ή αλλιώς ομαδοποίηση.
Τα δεδομένα στα οποία θα εφαρμόσουμε αυτούς τους αλγορίθμους δεν έχουν χαρακτηριστικό κλάσεις αφού αυτό είναι που ψάχνουμε να βρούμε. Αν λοιπόν

τρέξουμε το cluster για 4 κλάσεις (ενώ γνωρίζουμε ότι οι κλάσεις είναι 3) παρατηρούμε στο Cluster Visualizer ότι τα 2 πρώτα Clusters αντιστοιχούν στις πραγματικές κλάσεις ενώ το τρίτο μοιράζεται σε δυο μικρότερα. Το γεγονός αυτό υποδηλώνει ότι η κλάση αυτή περιέχει τα πιο ανομοιογενή δεδομένα και θα μπορούσε να μελετηθεί σε μικρότερες κατηγορίες. Αν το τρέξουμε όμως για 3 Clusters διαπιστώνουμε ότι ο διαχωρισμός γίνεται σχεδόν σωστά. Τα δεδομένα αυτά μπορούν να αποθηκευτούν σε αρχείο .arff για περαιτέρω ανάλυση. Επίσης υπάρχει η δυνατότητα αλλαγής των αξόνων του γραφήματος με κάποιο άλλο χαρακτηριστικό για να έχουμε καλύτερη αντίληψη των δεδομένων.



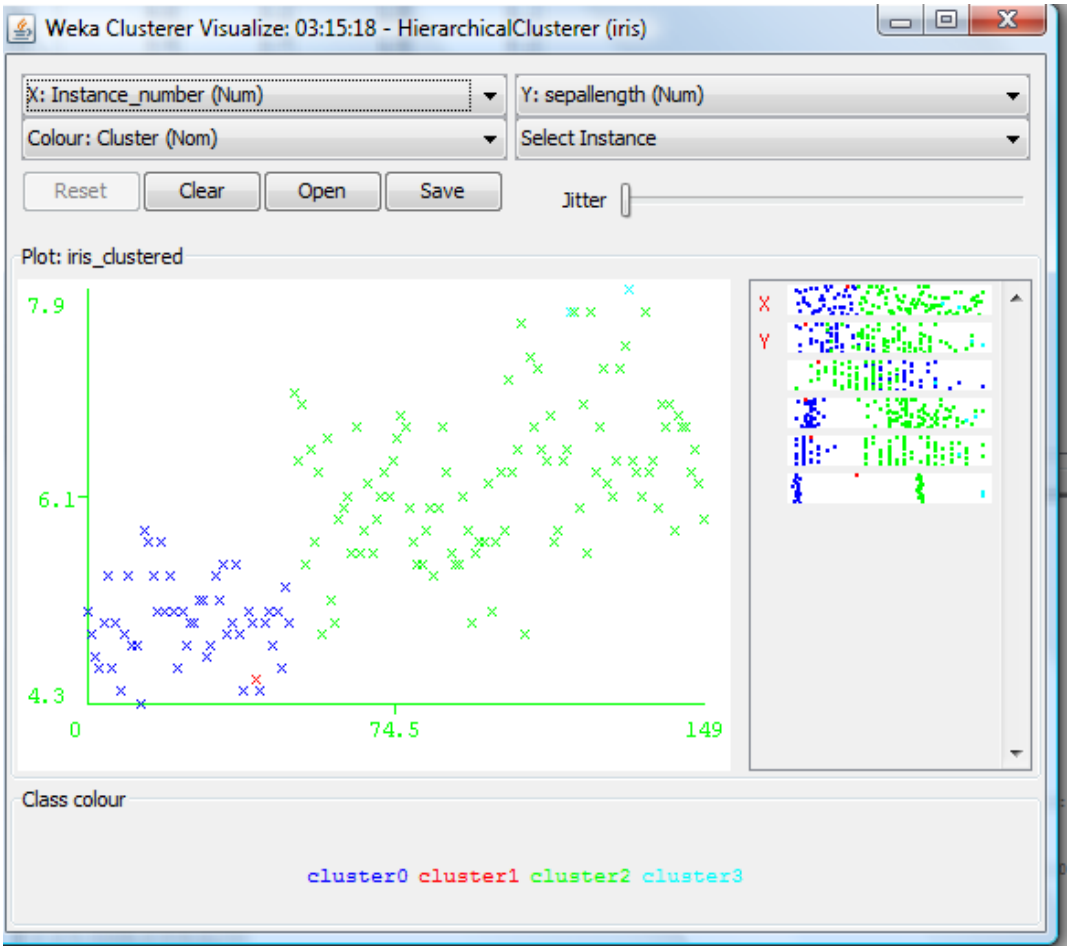
Σχήμα 15: XMeans για 4 κλάσεις



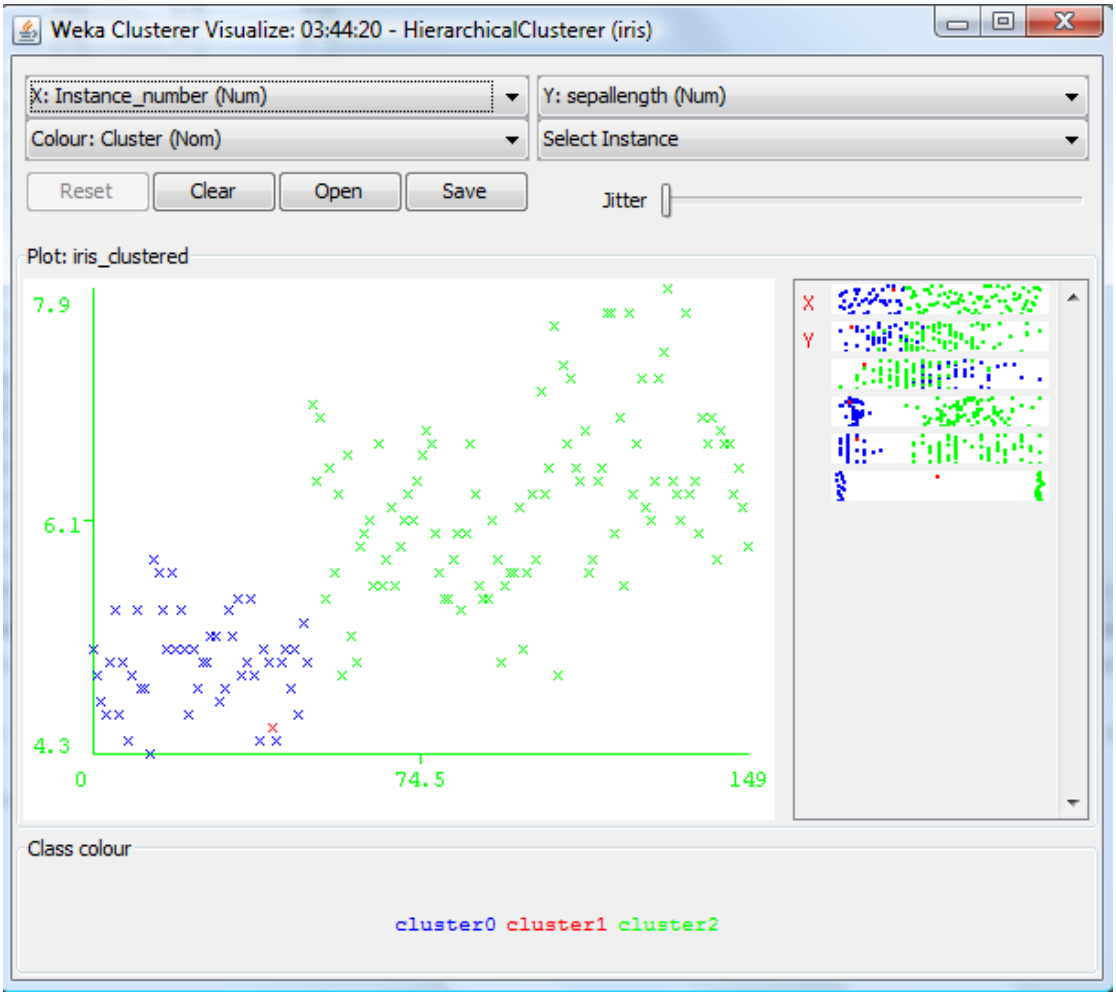
Σχήμα 16: XMeans για 3 κλάσεις

Αν τώρα εφαρμόσουμε το ίδιο πείραμα και με τον αλγόριθμο hierarchical clusterer παρατηρούμε ότι έχει χαθεί ένα ολόκληρο cluster, καθώς έχουν συγχωνευθεί δυο

μεταξύ τους και έχει ξεχωρίσει αυτό που εντοπίσαμε πιο πάνω ως ευδιάκριτο. Καταλαβαίνουμε ότι αυτό δε μας παρέχει χρήσιμα αποτελέσματα και έτσι με τη γνώση που έχουμε θέτουμε τον αριθμό των clusters 3. Παρατηρούμε και πάλι όμως ότι δυο clusters έχουν συγχωνευτεί. Από αυτό συμπεραίνουμε ότι η απόδοση του συγκεκριμένου αλγορίθμου δεν είναι ικανοποιητική και οφείλεται σε δυο βασικούς παράγοντες: ο πρώτος παράγοντας είναι ότι η υλοποίηση του αλγορίθμου αυτού δεν είναι η βέλτιστη στο Api του Weka και ο δεύτερος παράγοντας είναι, ότι δεν ταιριάζει στα συγκεκριμένα δεδομένα που επιλέξαμε να αναλύσουμε. Για πιο ασφαλή συμπεράσματα θα πρέπει να γίνουν πειράματα με αρκετά και διαφορετικά σύνολα δεδομένων και ίσως διαφορετικές υλοποιήσεις του αλγόριθμου αυτού.



Σχήμα 17: HierarchicalClusterer για 4 κλάσεις



Σχήμα 18: HierarchicalClusterer για 3 κλάσεις

ΕΠΙΛΟΓΟΣ

Αξιολογώντας την cystom αυτή εφαρμογή που χρησιμοποιεί το API του Weka μπορούμε να πούμε ότι έχουμε ένα εργαλείο data mining γρήγορο που όμως απαιτεί γνώση τεχνικών λεπτομερειών, κάτι πού ένας άπειρος χρήστη δεν κατέχει. Σε σύγκριση με τα αντίστοιχα API εργαλεία της IBM και Microsoft, μας παρέχει μια διαφάνεια λόγω του ότι είναι open source με αποτέλεσμα να καθιστά εύκολη την ανάπτυξη εφαρμογών. Τα δυο αυτά εργαλεία, ενώ παρέχουν ένα API συγκρίσιμο με του Weka, δεν είναι εύκολο να γίνει χρήση των API τους, Η IBM μας δίνει ένα API με λίγη τεκμηρίωση κ ο πιο εύκολος τρόπος πρόσβασης είναι η markup glossas pmll. Από την άλλη πλευρά η Microsoft κάνει διαθέσιμα τα εργαλεία της για εξόρυξη δεδομένων μέσω ερωτημάτων sql.

Αν και τα τρία εργαλεία που αναφέραμε υλοποιούν όλες τις μεθόδους που αναπτύξαμε προηγουμένως, ο λόγος που επιλέξαμε το API του Weka είναι η εύκολη χρήση του.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Κλείνοντας εδώ την εργασία θα πρέπει να πούμε ότι σαφώς δεν έχουμε καλύψει πλήρως και λεπτομερώς όλα όσα υπάρχουν για το αντικείμενο αυτό, όμως έχουμε προσπαθήσει να κάνουμε κατανοητά όσα το δυνατόν περισσότερα στοιχεία με έναν εύχρηστο τρόπο, παρουσιάζοντας παραδείγματα αλλά και πειράματα. Θα μπορούσαμε να είχαμε αναπτύξει μια πλήρης εφαρμογή εξόρυξης δεδομένων με περισσότερες μεθόδους και λειτουργίες όμως αυτό θα ξεπερνούσε τον καθαρά εκπαιδευτικό χαρακτήρα της εργασίας. Το ιδανικό σενάριο θα περιελάμβανε την υλοποίηση και άλλων APIs σε μια κοινή εφαρμογή, κάτι πού είναι δύσκολο, αν όχι αδύνατον, καθώς πολλά προϊόντα λογισμικού δεν έχουν ένα ανοιχτό API ή όσα έχουν δε σημαίνει πως αλληλεπιδρούν μεταξύ τους, λόγω διαφορετικής γλώσσας ή τεχνολογίας. Αυτό που πετυχαίνει η εργασία αυτή είναι να φωτίσει λίγο όσους ενδιαφέρονται για την εξόρυξη δεδομένων και την ανάπτυξη εφαρμογών για το σκοπό αυτο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Κεφάλαιο 1

1.Clifton, Christopher (2010). "Encyclopedia Britannica: Ορισμός του Data Mining»

2.^ Kantardzic, Mehmed (2003):. Data Mining Έννοιες, μοντέλα, μέθοδοι, και Αλγόριθμοι.

3.Alex Guazzelli, Wen-Ching Lin, Tridivesh Jena. PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics. CreateSpace, 2010

4. R. Baker."Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model". Workshop on Data Mining for User Modeling 2007 .

5.JF Superby, JP.Vandamme, N. Meskens."Determination of factors influencing the achievement of the first-year university students using data mining methods". Workshop on Educational Data Mining 2006 .

6.Xingquan Zhu, Ian Davidson (2007). Knowledge Discovery and Data Mining: Challenges and Realities Hershey, New York.

7.Yudong Chen, Yi Zhang, Jianming Hu, Xiang Li.Traffic Data Analysis Using Kernel PCA and Self-Organizing Map". Intelligent Vehicles Symposium, 2006 IEEE

8.Healey, R., 1991, Database Management Systems.In Maguire, D., Goodchild, MF, and Rhind, D., (eds.), Geographic Information Systems: Principles and Applications (London: Longman).

9.Câmara, AS and Raper, J., (eds.), 1999, Spatial Multimedia and Virtual Reality, (London: Taylor and Francis)

10.Secure Flight Program report , MSNBC.

11.grawal et al., Fast discovery of association rules , in Advances in knowledge discovery and data mining

12.National Research Council, Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment , Washington, DC: National Academies Press, 2008.

13.Stephen Haag et al.(2006). Management Information Systems for the information ageToronto: McGraw-Hill Ryerson.William Seltzer. The Promise and Pitfalls of Data Mining: 14.William Seltzer. The Promise and Pitfalls of Data Mining: Ethical Issues .
http://www.amstat.org/committees/ethics/linksdir/Jsm2005Seltzer.pdf

15.Think Before You Dig: Privacy Implications of Data Mining & Aggregation , NASCIO Research Brief, September 2004.

Κεφάλαιο 2

1.Abhishek Tiwaria and Arvind K.T. Sekhar: Workflow based framework for life science informatics, Computational Biology and Chemistry, Volume 31, Issues 5–6, Pages 305–319, Elsevier, October 2007.

2.Fox, John and Andersen, Robert (January 2005) (PDF). Using the R Statistical Computing Environment to Teach Social Statistics Courses. Department of Sociology, McMaster University. Retrieved 2006.

3.Vance, Ashlee (2009). "Data Analysts Captivated by R's Power". New York Times. Retrieved 2009. "R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca..."

4."Robert Gentleman's home page". Retrieved 2009

5.Kurt Hornik. The R FAQ: Why is R named R?. 2008

6."Free Software Foundation (FSF) Free Software Directory: GNU R". 2010.

7. "What is R?". 2009.

8.Carrot² project website

9. Carrot² search results clustering demo

10.Data Mining Tools Used Poll (May 2009)".

11."Data Mining / Analytic Tools Used Poll (May 2010)".

12. GATE Family page on the GATE website

13.GATE Wiki

14.Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Journal Of Natural Language Engineering 2009 (Y. Li, K. Bontcheva and H. Cunningham)

15."Combining Biological Databases and Text Mining to Support New Bioinformatics Applications", by René Witte and Christopher J.O. Baker (in "Lecture Notes in Computer Science, Springer Berlin, Volume 3513, 2005)

16."Open Source Text Analytics" web article by Seth Grimes

17."KIM – a semantic platform for information extraction and retrieval", by Popov et al (Natural Language Engineering (2004), 10:375-392)

18.TeachSource. 5 of the Best Free and Open Source Data Mining Software

19.SoftSea Ediror review

20.DreamCSS.COM. 8 useful open source information graphics software

Κεφάλαιο 3

Intelligent Miner

1.Peter Cabena, Hyun Hee Choi, Il Soo Kim, Shuichi Otsuka, Joerg Reinschmidt, Gary Saarenvirta. Intelligent Miner for Data Applications Guide

Microsoft Analysis Services

1."Microsoft Announces Acquisition Of Panorama Online Analytical Processing (OLAP) Technology".

2.MS SQL Server 7.0 OLAP Services".

3."SQL Server 2000 – Analysis Services".

4."SQL Server 2005 Analysis Services"

5.http://technet.microsoft.com/en-us/library/ms175595.aspx 2011

Weka

1.Ian H. Witten; Eibe Frank (2005). "Data Mining: Practical machine learning tools and techniques, 2nd Edition". Morgan Kaufmann, San Francisco.

2.G. Holmes; A. Donkin and I.H. Witten (1994). "Weka: A machine learning workbench". Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia.

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

4. P. Reutemann; B. Pfahringer and E. Frank (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.

5.an H. Witten; Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham (1999). "Weka: Practical Machine Learning Tools and Techniques with Java Implementations". Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems. pp. 192–196.

6.Gregory Piatetsky-Shapiro (2005-06-28). "KDnuggets news on SIGKDD Service Award 2005".

7."Overview of SIGKDD Service Award winners". 2005.

Κεφάλαιο 4

1."Definition of: API". PC Magazine. 1996.

2.Orenstein, David (2000-01-10). "QuickStudy: Application Programming Interface (API)". Computerworld. Retrieved

3. Benslimane, Djamal; Schahram Dustdar, and Amit Sheth (2008). "Services Mashups: The New Generation of Web Applications". IEEE Internet Computing, vol. 12, no. 5. Institute of Electrical and Electronics Engineers.

4.Niccolai, James (2008), "So What Is an Enterprise Mashup, Anyway?", PC World

5."Dynamic Community content via APIs". 2009

6.Microsoft (October 2001). "Run Older Programs On Windows XP" (in EN). Microsoft. pp. 4.

7.Stoughton, Nick (April 2005). "Update on Standards" (PDF). USENIX.

Κεφάλαιο 5

1. Ian H. Witten & Eibe Frank. Data Mining, Practical Machine Learning Tools and Techniques

2.http://wiki.pentaho.com/display/DATAMINING/Data+Mining+Algorithms+and+Tools+in+Weka

3.http://www.statsoft.com/textbook/naive-bayes-classifier/

4.Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & WJ Frawley, eds, 'Knowledge Discovery in Databases'

5.Agrawal; T. Imielinski; A. Swami: Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference 1993

6.Hipp Jochen, Güntzer Ulrich, και Nakhaeizadeh Gholamreza.Algorithms for association rule mining - A general survey and comparison.

7.Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" . Introduction to Data Mining .

8.Jian Pei, Jiawei Han, and Laks VS Lakshmanan. Mining frequent itemsets with convertible constraints.

9.Hajek P., Havel I., Chytil M.: The GUHA method of automatic hypotheses determination, Computing 1

10.Petr Hajek, Tomas Feglar, Jan Rauch, David Coufal. Database Support for Data Mining Applications

11.Edward R. Omiecinski. Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering, 15(1):57-69, Jan/Feb 2003.

12.C. C. Aggarwal and P. S. Yu. A new framework for itemset generation. In PODS 98, Symposium on Principles of Database Systems, pages 18-24, Seattle, WA, USA, 1998.

13.Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In SIGMOD 1997,

Proceedings ACM SIGMOD International Conference on Management of Data, pages 255-264, Tucson, Arizona, USA, May 1997.

14.Piatetsky-Shapiro, G., Discovery, analysis, and presentation of strong rules. Knowledge Discovery in Databases, 1991

15.Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

16. Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation. Data Mining and Knowledge Discovery

17.Application of the Weka Machine Learning Library to Hospital Ward Occupancy Problems Ian Harris1, J¨org Denzinger2 and Dean Yergens3 1Faculty of Medicine, University of Calgary, Canada 2Department of Computer Science, University of Calgary, Canada 3Faculty of Medicine, University of Manitoba, Canada Technical

18.Cluster Analysis: Basic Concepts and Algorithms

19.Finch, H. (2005). Comparison of distance measures in cluster analysis with dichotomous data. Journal of Data Science, 3

20.Huberty, C. J., Jordan, E. M., & Brandt, W. C. (2005). Cluster analysis in higher education research. In J. C. Smart (Ed.), Higher Education: Handbook of Theory and Research (Vol. 20, pp. 437-457). Great Britain: Springer.

21.Basak S.C., Magnuson V.R., Niemi C.J., Regal R.R. "Determing Structural Similarity of Chemicals Using Graph Theoretic Indices".

22.Huth R. et al. "Classifications of Atmospheric Circulation Patterns: Recent Advances and Applications". Ann. N.Y. Acad. Sci., 1146

23.Christopher D. Manning, Prabhakar Raghavan & Hinrich Schutze.Introduction to Information Retrieval. Cambridge University Press

24.Dunn, J. (1974). "Well separated clusters and optimal fuzzy partitions".Journal of Cybernetic

25.Tutorial Outlining Feature Selection Algorithms

26.Efficient Feature Subset Selection and Subset Size Optimization (Survey, 2010)

27.S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas. Data Preprocessing for Supervised Leaning

28.Marek Grochowski, Norbert Jankowski: Comparison of Instance Selection Algorithms II. Results and Comments. ICAISC 2004a:

29.Norbert Jankowski, Marek Grochowski: Comparison of Instances Selection Algorithms I. Algorithms Survey. ICAISC 2004b: 598-603.

30.J.R. Cano, F. Herrera, M. Lozano. Strategies for Scaling Up Evolutionary Instance Reduction Algorithms for Data Mining. In: L.C. Jain, A. Ghosh (Eds.) Evolutionary Computation in Data Mining, Springer, 2005, 21-39

31.Lakshminarayan K., S. Harp & T. Samad, Imputation of Missing Data in

Industrial Databases, Applied Intelligence 11, 259–275 (1999).

ΠΑΡΑΡΤΗΜΑ 1

Κώδικας

Associations

Arff Αρχεία

```
ArffLoader loader = new ArffLoader();
    loader.setFile(new File(dataFile));
    train = loader.getDataSet();
    train.setClassIndex(selectIndex-1);
```

Ένα .arff αρχείο το φορτώνουμε από το δίσκο με τον *ArffLoader loader.* Με το αντικείμεναο *ArffLoader loader*

καλούμε τι μέθοδο *set file* η οποία παίρνει ένα αρχείο File στην τοποθεσία του αρχείου μασς στο δίσκο. Για να ανακτίσουμε δεδομένα τύπου Instances από το arff, καλούμε τη μέθοδο *getData* του *loader.* Μια ακόμη σημαντική μέθοδος του *loader* είναι η *setClassIndex*

που μας επιτρέπει να ορίσουμε την κύρια κλάση των δεδομένων μας, όπου αυτό είναι απαραίτητο.

Apriori

```
Apriori ap=new Apriori();
    ap.buildAssociations(train);
    AssociatorEvaluation eval = new AssociatorEvaluation();
    eval.evaluate(ap, trainInstances);
```

Για να δημιουργήσουμε έναν *Associator* τύπου *Apriori* στο Weka πρώτα δημιουργήσαμε ένα αντικείμενο της κλάσης *Apriori.* Στη συνέχεια καλούμε τη μέθοδο *buildAssociations* που παίρνει τα δεδομένα *train* τύπου *Instances.*

Tertius

```
Tertius ta=new Tertius();
    ta.buildAssociations(train);
    AssociatorEvaluation eval = new AssociatorEvaluation();
    eval.evaluate(ta, trainInstances);
```

Έναν *Associator* τύπου *Tertius* μπορεί να δημιουργηθεί με τον ίδιο τρόπο όπως ο *Apriori.* Δημιουργούμε και εδώ ένα αντικείμενο της κλάσης *Tertius* όπως και στο *Apriori,* με το οποίο καλούμε η μέθοδο *buildAssociations.*

Atrribute selection

Greedy k Genetic Search

Για τη δημιουργία αντικειμένων τέτοιων αναζητήσεων που κάνουν αναζήτηση τον βέλτιστων στηλών στα δεδομένα, η διαδικασία είναι λίγο πιο πολύπλοκη απ' ότι προηγουμένως.

Αρχικά δημιουργούμε τα αντικείμενα που θα εκτελέσουν την αναζήτηση, τα οποία για κάθε περίπτωση είναι:

```
GeneticSearch gs=new GeneticSearch();
GreedyStepwise gs=new GreedyStepwise()
```

Ακόμα, χρείαζεται ένα αντικείμενο τύπου *classifier*

που θα παίξει το ρόλο του αξιολογητή, στην περίπτωση αυτή χρησιμοποιούμε ένα δέντρο τύπου J48

J48 base = new J48();

αυτό μαζί με το αντικείμενα gs που κάνει την αναζήτηση τα δηλώνουμε στο παρακάτω AtribureSelectionClassifier.

```
AttributeSelectedClassifier classifier = new AttributeSelectedClassifier();
        CfsSubsetEval eval = new CfsSubsetEval();
        classifier.setClassifier(base);
        classifier.setEvaluator(eval);
        classifier.setSearch(gs);
 meta, vasi me ton pio kato kodika
        Evaluation evaluation = new Evaluation(train);
        evaluation.crossValidateModel(classifier, train, 10, new Random(1));

        weka.filters.supervised.attribute.AttributeSelection filter = new
 weka.filters.supervised.attribute.AttributeSelection();

        filter.setEvaluator(eval);
        filter.setSearch(gs);
        filter.setInputFormat(train);

        Instances newData = Filter.useFilter(train, (Filter)filter);
```

Δημιουργούμε αξιολόγηση Evaluation η οποία με cross validation αξιολογεί τα δεδομένα μας. Ακολούθως δημιουργούμε ένα φίλτρο το οποίο μέσω της μεθόδου *useFilter* θα μας δώσει δεδομένα τύπου *Instances* όπου είναι τα επιλεγμένα χαρακτηριστικά. Στην εφαρμογή μας τα δείγματα αυτά τα αποθηκεύουμε σε αρχείο arff.

Classifiers

Για τους Classifiers δημιουργήσαμε τρία αντικείμενα, ένα δέντρο J48, ένα Multilayer Perceptron και έναν Naive Bayes Classifier.

```
J48 j48=new J48();
MultilayerPerceptron mlp=new MultilayerPerceptron();
NaiveBayesUpdateable nb = new NaiveBayesUpdateable();
```

Για να τρέξουμε τα δυο πρώτα καλούμε τη μέθοδο *buildClassifier*

που παίρνει δεδομένα τύπου *Instances* και στη συνέχεια την αξιολογούμε για το output με ένα αντικείμενο *evaluation.* Αυτά μπορούμε να τα δούμε παρακάτω για τον *Classifier* J48.

```
j48.buildClassifier(train);
        Evaluation eval = new Evaluation(train);
        Random rand = new Random(1);  // using seed = 1
        int folds = 10;
        eval.crossValidateModel(j48, train, folds, rand);
```

Για τον Naive Bayes Classifier καλούμε τη μέθοδο *buildClassifier* όπως και πριν αλλά με τη διαφορά ότι τώρα πρέπει να δείξουμε με έναν βρόγχο επανάληψης όλα τα instances ξεχωριστά.

```
nb.buildClassifier(train);
    Instance current;
    while ((current = loader.getNextInstance(train)) != null)
       nb.updateClassifier(current);
```

Η αξιολόγηση μέσου του αντικειμένου Evaluation παραμένει η ίδια.

Κάθε ένας από τους Classifiers παρέχει τη μέθοδο classifyInstance η οποία μασς δίνει τη δυνατότητα μετά την υλοποίηση του classifier να κάνουμε classify. Ως είσοδο η classifyInstance παίρνει αντικείμενα τύπου Instance (όχι Instances) που αντιπροσωπεύουν μια γραμμή δεδομένων.

Clusterers

Στην εφαρμογή μας υλοποιήσαμε δυο clusterers.

Έναν με βάση τον αλγόριθμο XMeans/KMeans και έναν hierarchical clusterer.

```
HierarchicalClusterer hc=new HierarchicalClusterer();

 XMeans xm=new XMeans();
```

Οι πιο αξιοσημείωτες μέθοδοι για τη δημιουργία ενός clusterer είναι η buildClusterer που παίρνει τα δεδομένα Intances και η setNumClusters που δηλώνει τον αριθμό των clusters που θα παραχθούν. Στη συνέχεια δημιουργούμε ένα αντικείμενο ClusterEvaluation που αξιολογεί τον cluster ώστε να κρίνουμε τα αποτελέσματά του. Αυτά φαίνονται πιο κάτω για τον αλγόριθμο hierarchical clusterer.

```
hc.setNumClusters(classNo);

        hc.buildClusterer(train);
        ClusterEvaluation eval = new ClusterEvaluation();
        eval.setClusterer(hc);
        eval.evaluateClusterer(train);
        System.out.println(eval.clusterResultsToString());
```

Preprocessing Filters

Για τους σκοπούς της εφαρμογής μας δημιουργήσαμε και τρία αντικείμενα που επεξεργάζονται τα δεδομένα. Αυτά είναι το *numeric to nominal,* το *nominal to binary,* και το *descritize.*

Discretize dv=new Discretize()

NominalToBinary nb=new NominalToBinary();

NumericToNominal nn=new NumericToNominal();

Η χρήση ενός φίλτρουθ προεπεξεργασίας στο weka είναι αρκετά απλή. Αρχικά ορίζουμε τον τύπο των δεδομένων στο φίλτρο μέσω της *setInputFormat*

που δέχεται Instances και έπειτα καλούμε τη μέθοδο useFilter που μας παρέχει το κάθε φίλτρο και παίρνουμε ως έξοδο ένα φιλτραρισμένο σύνολο Instances.

Όπως και στον παρακάτω κώδικα,

dv.setInputFormat(train);

Instances newData=dv.useFilter(train, dv);

System.out.println(newData);

μετά τη χρήση του useFilter

επιλέγουμε να αποθηκεύσουμε τα δεδομένα σε ένα αρχείο arff.

ΠΑΡΑΡΤΗΜΑ

Documentation της custom Εφαρμογής.

Associators:
AbstractAssociator
Apriori

AprioriItemSet
AssociatorEvaluation
CaRuleGeneration
CheckAssociator
FilteredAssociator
FPGrowth
FPGrowth.AssociationRule
FPGrowth.BinaryItem
GeneralizedSequentialPatterns
ItemSet
LabeledItemSet
PredictiveApriori
PriorEstimation
RuleGeneration
RuleItem
SingleAssociatorEnhancer
Tertius

Atribute selection
ASEvaluation
ASSearch
AttributeSelection
AttributeSetEvaluator
BestFirst
CfsSubsetEval
CheckAttributeSelection
ChiSquaredAttributeEval
ClassifierSubsetEval
ConsistencySubsetEval
CostSensitiveASEvaluation
CostSensitiveAttributeEval
CostSensitiveSubsetEval
ExhaustiveSearch
FilteredAttributeEval
FilteredSubsetEval
GainRatioAttributeEval
GeneticSearch
GreedyStepwise

HoldOutSubsetEvaluator
InfoGainAttributeEval
LatentSemanticAnalysis
LFSMethods
LinearForwardSelection
OneRAttributeEval
PrincipalComponents
RaceSearch
RandomSearch
Ranker
RankSearch
ReliefFAttributeEval
ScatterSearchV1
SubsetSizeForwardSelection
SVMAttributeEval
SymmetricalUncertAttributeEval
UnsupervisedAttributeEvaluator
UnsupervisedSubsetEvaluator
WrapperSubsetEval

Classifiers
BVDecompose
BVDecomposeSegCVSub
CheckClassifier
CheckSource
Classifier
CostMatrix
Evaluation
IteratedSingleClassifierEnhancer
MultipleClassifiersCombiner
RandomizableClassifier
RandomizableIteratedSingleClassifierEnhancer
RandomizableMultipleClassifiersCombiner
RandomizableSingleClassifierEnhancer
SingleClassifierEnhancer
AODE
AODEsr
BayesianLogisticRegression
BayesNet

ComplementNaiveBayes
DMNBtext
HNB
NaiveBayes
NaiveBayesMultinomial
NaiveBayesMultinomialUpdateable
NaiveBayesSimple
NaiveBayesUpdateable
WAODE
GaussianProcesses
IsotonicRegression
LeastMedSq
LibLINEAR
LibSVM
LinearRegression
Logistic
MultilayerPerceptron
PaceRegression
PLSClassifier
RBFNetwork
SimpleLinearRegression

SimpleLogistic
SMO
SMOreg
SPegasos
VotedPerceptron
Winnow
IB1
IBk
KStar
LBR
LWL
AdaBoostM1
AdditiveRegression
AttributeSelectedClassifier
Bagging
ClassificationViaClustering
ClassificationViaRegression
CostSensitiveClassifier
CVParameterSelection
Dagging
Decorate

END
FilteredClassifier
Grading
GridSearch
LogitBoost
MetaCost
MultiBoostAB
MultiClassClassifier
MultiScheme
OrdinalClassClassifier
RacedIncrementalLogitBoost
RandomCommittee
RandomSubSpace
RegressionByDiscretization
RotationForest
Stacking
StackingC
ThresholdSelector
Vote
ConjunctiveRule
DecisionTable

DecisionTableHashKey
DTNB
JRip
M5Rules
NNge
OneR
PART
Prism
Ridor
Rule
RuleStats
ZeroR
ADTree
BFTree
DecisionStump
FT
Id3
J48
J48graft
LADTree
LMT

M5P
NBTree
RandomForest
RandomTree
REPTree
SimpleCart
UserClassifier

Clusterers
AbstractClusterer
AbstractDensityBasedClusterer
CheckClusterer
CLOPE
ClusterEvaluation
Cobweb

DBScan
EM
FarthestFirst
FilteredClusterer
HierarchicalClusterer
MakeDensityBasedClusterer
OPTICS
RandomizableClusterer
RandomizableDensityBasedClusterer
RandomizableSingleClustererEnhancer
sIB
SimpleKMeans
SingleClustererEnhancer
XMeans

Preprocessing filters
AbstractTimeSeries
Add
AddCluster
AddExpression
AddID
AddNoise
AddValues
Center
ChangeDateFormat
ClassAssigner
ClusterMembership
Copy
Discretize
FirstOrder
InterquartileRange
KernelFilter
MakeIndicator
MathExpression
MergeTwoValues
MultiInstanceToPropositional

NominalToBinary
NominalToString
Normalize
NumericCleaner
NumericToBinary
NumericToNominal
NumericTransform
Obfuscate
PartitionedMultiFilter
PKIDiscretize
PotentialClassIgnorer
PrincipalComponents
PropositionalToMultiInstance
RandomProjection
RandomSubset
RELAGGS
Remove
RemoveType
RemoveUseless
Reorder
ReplaceMissingValues

Standardize
StringToNominal
StringToWordVector
SwapValues
TimeSeriesDelta
TimeSeriesTranslate
Wavelet