ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη γνώσης σε σύγχρονες σχεσιακές βάσεις δεδομένων

Ευστάθιος Γ. Καλουτσίδης

Επιβλέπων: Δρ. Βασίλειος Ταμπακάς

Καθηγητής τμήματος Μηχανικών Πληροφορικής Τ.Ε.

Αντίρριο 2017

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Αντίρριο, 19-12-2017

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. ……..
2. ……..
3. ……..

Στους γονείς μου και
στην αδελφή μου.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλο το εκπαιδευτικό προσωπικό του τμήματος Μηχανικών Πληροφορικής καθώς και το διοικητικό και τεχνικό προσωπικό του τμήματος, για όλη την βοήθεια που έλαβα στα χρόνια φοίτησης μου.

Ιδιαίτερα θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή μου Βασίλειο Ταμπακά, πρόεδρο του Τμήματος Μηχανικών Πληροφορικής για όλη την βοήθεια, καθοδήγηση και υποστήριξη που έλαβα για την ολοκλήρωση αυτής της πτυχιακής εργασίας.

Επίσης αισθάνομαι την ανάγκη να ευχαριστήσω και τα άλλα δύο μέλη της τριμελούς εξεταστικής επιτροπής της παρούσας πτυχιακής εργασίας.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου Γεώργιο και Ελένη καθώς και την αδελφή μου Σοφία για την συμπαράσταση και υπομονή που έδειξαν προς το πρόσωπο μου όλα αυτά χρόνια.

Ευστάθιος Γ. Καλουτσίδης

Αντίρριο, 2017

Περίληψη

Με την ραγδαία ανάπτυξη του διαδικτύου και των τεχνολογιών της πληροφορικής, έχουμε φτάσει σε ένα σημείο όπου μπορούμε να έχουμε πρόσβαση στο διαδίκτυο σχεδόν σε όλους τους χώρους όπου παρευρισκόμαστε και με οποιοδήποτε μέσο, από προσωπικό υπολογιστή μέχρι και κινητό τηλέφωνο, κάτι που δεν θα φανταζόμασταν τόσο εύκολα μερικά χρόνια πριν. Αυτή η ραγδαία ανάπτυξη επιφέρει και έναν τεράστιο όγκο από δεδομένα, καθώς ο κάθε χρήστης πλοηγείται και δραστηριοποιείται στο διαδίκτυο. Όλα αυτά τα δεδομένα αποθηκεύονται συνήθως στις Βάσεις Δεδομένων. Μέσα από τις Βάσεις Δεδομένων μπορούμε να αντλήσουμε χρήσιμες πληροφορίες που μπορούν να μας δώσουν τη δυνατότητα να προσδιοριστούν οι θέσεις της αγοράς και να διευκολύνουν τη λήψη αποφάσεων των επιχειρήσεων. Η επιστήμη που ασχολείται με την διαδικασία άντλησης χρήσιμης πληροφορίας από σύνολα ή βάσεις δεδομένων μεγάλου μεγέθους ονομάζεται Εξόρυξη Γνώσης (Data Mining). Η τεχνολογία της Εξόρυξης Γνώσης χρησιμοποιείται συνήθως από οργανισμούς ή τμήματα επιχειρηματικής ευφυΐας και από οικονομικούς αναλυτές, αλλά πλέον η χρήση της επεκτείνεται συνεχώς και σε άλλες επιστήμες όπου γεννιέται η ανάγκη εξαγωγής χρήσιμης γνώσης. Για παράδειγμα: επιχείρηση, ιατρική, οικονομία, ανθρώπινα δικαιώματα, τηλεπικοινωνία κ.λπ. Η Εξόρυξη Γνώσης είναι κράμα της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής εκπαίδευσης και των βάσεων δεδομένων. Η παρούσα πτυχιακή εργασία έχει ως σκοπό να παρουσιάσει στον αναγνώστη τις Βάσεις Δεδομένων, τα Συστήματα Διαχείρισης Βάσεων Δεδομένων και τα εργαλεία που διαθέτουν για την εξόρυξη γνώσης, την Μηχανική Εκπαίδευση και τέλος την Εξόρυξη Γνώσης στην πράξη και τα αποτελέσματα αυτής.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Γραφημάτων

Κατάλογος Εικόνων

Κατάλογος Συντομογραφιών

ACID = Atomicity, Consistency, Isolation, and Durability

ANSI = American National Standards Institute

API = Application Programming Interface

ARFF = Attribute-Relation File Format

ASCII = American Standard Code for Information Interchange.

BI = Business Intelligence

BIDS = Business Intelligence Development Studio

BIML = Business Intelligence Markup Language

BPM = Business Process Management

CIA = Central Intelligence Agency

CLI = Call Level Interface

CLI = Command-Line Interface

CLOB = Character Large Object

CPU = Central Processing Unit

CSV = Comma Separated Values

DBMS = Database management system

DBWn = Database Writer process

DMX = Data Mining Extensions

GNU = GNU's Not Unix

GPL = General Public License

GUI = Graphical User Interface

HTML = Hypertext Markup Language

ID = Identity

ID3 = Iterative Dichotomiser 3

ISV = Independent Software Vendor

JDBC = Java Database Connectivity

JSON = javaScript Object Notation

LAMP = Linux Apache MySQL Perl/PHP/Python

LMT = Logistic Model Tree

MAP = Maximum a Posteriori

MDM = Master Data Management

MDS = Master Data Services

ML = Maximum Likelihood

MLP = Multilayer Perceptron

MVCC = Multiversion Concurrency Control

NN = Nearest Neighbour

OCI = Oracle Call Interface

ODBC = Open Database Connectivity

ODM = Oracle Data Mining

ODP.NET = Oracle Data Provider for .NET

OEM = Original Equipment Manufacturer

OLAP = Online Analytical Processing

OLTP = Online Transaction Processing

PDF = Portable Document Format

PHP = Hypertext Preprocessor

PL/SQL = Procedural Language/Structured Query Language

PMML = Predictive Model Markup Language

RDBMS = Relational database management system

ROLAP = Relational online analytical processing

RTF = Rich Text Format

SDK = Software Development Kit

SDL = Software Development Laboratories

SMO = Sequential Minimal Optimization

SQL = Structured Query Language

SSAS = SQL Server Analysis Services

SSDT = SQL Server Data Tools

SSIS = SQL Server Integration Services

SVM = Support Vector Machines

TDIDT = Top-Down Induction of Decision Trees

TDS = Tabular Data Stream

TXT = Text

UDB = Universal Database

URL = Uniform Resource Locator

WEKA = Waikato Environment for Knowledge Analysis

XLS = eXceL Spreadsheet

XML = Extensible Markup Language

YALE = Yet Another Learning Environment

ΒΔ = Βάση Δεδομένων

Η/Υ = Ηλεκτρονικοί Υπολογιστές

ΣΔΒΔ = Σύστημα Διαχείρισης Βάσης Δεδομένων

1 Εισαγωγικές Έννοιες

1.1 Εισαγωγή

Στις αρχές της δεκαετίας του 1960, ένας νέος όρος εμφανίστηκε στην βιβλιογραφία αλλά και στην πρακτική των ηλεκτρονικών υπολογιστών. Ο όρος ήταν «βάσεις δεδομένων» και αρχικά χρησιμοποιούταν από τους εργαζόμενους σε στρατιωτικές υπηρεσίες πληροφορικής, για να δηλώσουν συλλογές δεδομένων υπό τη διαχείριση μεγάλων υπολογιστικών συστημάτων διαμοιραζόμενου χρόνου. Η έννοια αυτή ήρθε να αντικαταστήσει τον ορό «ολοκληρωμένη επεξεργασία δεδομένων» (integrated data processing) που ήταν σε χρήση από την προηγούμενη δεκαετία. Σύντομα, ο όρος βάσεις δεδομένων διαδόθηκε σε όλο το εύρος των επιχειρήσεων και αρχικά υποδήλωνε τη συγκεντρωτική διαχείριση των δεδομένων ανεξάρτητων εφαρμογών. (Ταμπακάς 2011)

1.2 Τα Δεδομένα και οι Πληροφορίες

Στους υπολογιστές, οι έννοιες δεδομένα (data) και πληροφορία (information) πολλές φορές συγχέονται και χρησιμοποιούνται για τον ίδιο σκοπό. Αυτό είναι λάθος γιατί οι δυο έννοιες είναι διαφορετικές μεταξύ τους.

Τα δεδομένα είναι στοιχειά τυποποιημένα σε καθορισμένη μορφή και είναι κατάλληλα για επεξεργασία από ανθρώπους ή μηχανές. Τα δεδομένα που πρόκειται να αποθηκευτούν σε Η/Υ πρέπει να μετατραπούν σε κάποιον από τους γνωστούς κώδικες αναπαράστασης των υπολογιστών (π.χ. ASCII ή δυαδικός κώδικας). Με βάση τα παραπάνω το σύνολο των λέξεων {Νίκος, Γιώργος, Παντελής, Κώστας} και το σύνολο των αριθμών {7, 3, 10 ,9} είναι δεδομένα.

Η επεξεργασία των δεδομένων δημιουργεί την πληροφορία. Για την εξαγωγή της πληροφορίας είναι απαραίτητη η συγκεκριμένη γνώση των δεδομένων, π.χ. το πεδίο αναφοράς τους. Για παράδειγμα, αν γνωρίζουμε πως το σύνολο δεδομένων {Νίκος, Γιώργος, Παντελής, Κώστας} αναπαριστά ονόματα σπουδαστών και το σύνολο {7, 3, 10 ,9} τους αντίστοιχους βαθμούς τους σε ένα μάθημα, τότε η επεξεργασία τους μπορεί να δώσει πληροφορίες της μορφής:

«Ο Παντελής είναι ο καλύτερος όλων στο μάθημα»

«Ο Κώστας βαθμολογήθηκε με άριστα»

«Ο μέσος όρος των σπουδαστών στο μάθημα είναι 7,25»

(Ταμπακάς 2011)

1.3 Οι Βάσεις Δεδομένων

Ένα σημαντικό γνώρισμα των δεδομένων είναι πως μπορούν να κωδικοποιηθούν και να αποθηκευτούν στους Η/Υ. Βάση Δεδομένων (ΒΔ) είναι μια διαμοιραζόμενη συλλογή από λογικά σχετιζόμενα δεδομένα μαζί με την περιγραφή τους, που είναι σχεδιασμένα να ικανοποιούν τις πληροφοριακές ανάγκες ενός οργανισμού. Οι βάσεις δεδομένων επομένως, προσφέρουν την οργάνωση και αποθήκευση των δεδομένων στον Η/Υ, ώστε να είναι δυνατή η επεξεργασία τους και η εξαγωγή της επιθυμητής πληροφορίας. Η τεχνολογία των βάσεων δεδομένων βρίσκει σημαντικότατες εφαρμογές σε όλες τις περιοχές που χρησιμοποιούνται οι υπολογιστές όπως στις επιχειρήσεις, στην εκπαίδευση, στην διοίκηση, στην οικονομία, στην ιατρική και στα νομικά. (Ταμπακάς 2011)

1.3.1 Σύστημα διαχείρισης βάσης δεδομένων

Με τον όρο Σύστημα Διαχείρισης Βάσης Δεδομένων (ΣΔΒΔ) γνωστό ως Database Management system (DBMS) εννοείται είτε κάποιο λογισμικό μέσω του οποίου γίνεται η δημιουργία, η διαχείριση, η συντήρηση και η χρήση μιας ηλεκτρονικής βάσης δεδομένων, ανάλογα με τον τύπο βάσης δεδομένων που επιλέγεται ή ένα σύνολο αλληλοσυσχετιζόμενων προγραμμάτων που τρέχουν και διαχειρίζονται τα δεδομένα μιας τέτοιας βάσης. Το λογισμικό χρησιμοποιεί στερεότυπες (standard) μεθόδους καταλογοποίησης, ανάκτησης, και εκτέλεσης ερωτημάτων σχετικών με τα δεδομένα. Το σύστημα διαχείρισης οργανώνει τα εισερχόμενα δεδομένα με τρόπους χρησιμοποιήσιμους από εξωτερικούς χρήστες. (‘Σύστημα Διαχείρισης Βάσης Δεδομένων’ 2013)

Τα σύγχρονα συστήματα διαχείρισης βάσεων δεδομένων χειρίζονται και αποθηκεύουν πληροφορίες χρησιμοποιώντας το σχεσιακό (relational) μοντέλο διαχείρισης βάσεων δεδομένων.

2 Οι Σχεσιακές Βάσεις Δεδομένων

2.1 Εισαγωγή

Οι σχεσιακές βάσεις δεδομένων παρουσιάζουν ιδιαίτερα σημαντικό θεωρητικό αλλά και πρακτικό ενδιαφέρον. Το σχεσιακό μοντέλο χρησιμοποιεί τη μαθηματική τυποποίηση της σχέσης, η οποία χωρίς αυστηρότητα αντιστοιχεί στους γνωστούς μας πινάκες, ως βασικό εργαλείο λογικής οργάνωσης των δεδομένων. Έγινε αμέσως δεκτό εξαιτίας της απλότητας και της πρακτικής του αξίας. Παράλληλα προτάθηκε ένα σύνολο από πράξεις πάνω στο σχεσιακό μοντέλο που ονομάστηκε σχεσιακή άλγεβρα.

Το σχεσιακό μοντέλο, αν και αρχικά είχε να αντιμετωπίσει τον ανταγωνισμό των παλαιότερων λογικών μοντέλων (ιεραρχικό, δικτυωτό), γρήγορα χάραξε το δικό του δρόμο. Στην πράξη, έγινε γρήγορα εξαιρετικά δημοφιλές με αποτέλεσμα να χρησιμοποιείτε σήμερα από το μεγαλύτερο μέρος των εμπορικών Συστημάτων Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ). (Ταμπακάς 2011)

2.2 Σχεσιακή βάση δεδομένων

Με τον όρο σχεσιακή βάση δεδομένων εννοείται μία συλλογή δεδομένων οργανωμένη σε συσχετισμένους πίνακες που παρέχει ταυτόχρονα ένα μηχανισμό για ανάγνωση, εγγραφή, τροποποίηση ή και πιο πολύπλοκες διαδικασίες πάνω στα δεδομένα. Ο σκοπός μιας βάσης δεδομένων είναι η οργανωμένη αποθήκευση πληροφορίας και η δυνατότητα εξαγωγής της πληροφορίας αυτής, ιδίως σε πιο οργανωμένη μορφή, σύμφωνα με ερωτήματα που τίθενται στη σχεσιακή βάση δεδομένων. Τα δεδομένα είναι δυνατόν να αναδιοργανώνονται με πολλούς διαφορετικούς τρόπους, σε νοητούς πίνακες, χωρίς να είναι απαραίτητη η αναδιοργάνωση των φυσικών πινάκων που τα αποθηκεύουν. Τη σχεσιακή βάση δεδομένων επινόησε ο Έντγκαρ Κοντ το 1970.

Οι ερωτήσεις, είτε από το χρήστη είτε από λογισμικό, προς τη βάση δεδομένων, γίνονται συνήθως μέσω της διαδεδομένης διαλογικής γλώσσας SQL (Structured Query Language). Εκτελώντας ερωτήματα ο χρήστης (ή το λογισμικό που εκπροσωπεί το χρήστη) είναι δυνατόν, ανάλογα με τα δικαιώματά του, να δημιουργήσει, να

μεταβάλλει και να διαγράψει δεδομένα στη βάση, ή να ανασύρει πληροφορίες με σύνθετα κριτήρια αναζήτησης. (‘Σχεσιακή βάση δεδομένων’ 2013)

Κάθε πινάκας μοιράζεται τουλάχιστον ένα πεδίο με ένα άλλο πίνακα σε ‘ένα-προς-ένα’, ‘ένα-προς-πολλά’ ή ‘πολλά-προς-πολλά’ σχέση. Οι σχέσεις αυτές επιτρέπουν στον χρηστή της βάσης δεδομένων να έχει πρόσβαση στα δεδομένα με σχεδόν απεριόριστο αριθμό τρόπων, και να συνδυάζει τους πίνακες ως δομικά στοιχεία για τη δημιουργία σύνθετων και πολύ μεγάλων βάσεων δεδομένων. (businessdictionary n.d.)

Τα διάφορα συστήματα λογισμικού που χρησιμοποιούνται για τη διατήρηση σχεσιακών βάσεων δεδομένων είναι γνωστά ως συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων. (Relational Database Management System-RDBMS).
(‘Relational database’ 2016)

2.2.1 Σύστημα διαχείρισης Σχεσιακών Βάσεων Δεδομένων

Ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS) είναι ένα σύστημα διαχείρισης βάσεων δεδομένων (DBMS) που βασίζεται στο σχεσιακό μοντέλο.
(‘Relational database management system’ 2016) Το σχεσιακό μοντέλο έχει σχέσεις μεταξύ πινάκων χρησιμοποιώντας πρωτεύοντα κλειδιά, ξένα κλειδιά και δείκτες.
(Twinkle 2012) Τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων είναι μια συνηθισμένη επιλογή για την αποθήκευση των πληροφοριών σε νέες βάσεις δεδομένων που χρησιμοποιούνται για οικονομικά αρχεία, βιομηχανικές και λογιστικές πληροφορίες, δεδομένα προσωπικού, καθώς και άλλες εφαρμογές από το 1980.
Οι σχεσιακές βάσεις δεδομένων αντικατέστησαν τις ιεραρχικές βάσεις δεδομένων και τις βάσεις δεδομένων που χρησιμοποιούσαν το δικτυωτό μοντέλο, επειδή είναι ποιο εύκολες στην κατανόηση και στην λειτουργία τους. (‘Relational database management system’ 2016)

2.2.2 Σύστημα διαχείρισης βάσης δεδομένων VS Σύστημα διαχείρισης Σχεσιακών Βάσεων Δεδομένων

Στον παρακάτω πινάκα θα δούμε τις σημαντικότερες διαφορές μεταξύ DBMS και RDBMS.

#	DBMS	RDBMS
1	Δημιουργήθηκαν το 1960	Δημιουργήθηκαν το 1970
2	Κατά τη διάρκεια της εισαγωγής τους ακολούθησαν τις λειτουργίες του μοντέλου πλοήγησης (Navigational DBMS) για την αποθήκευση δεδομένων και την ανάκτηση τους.	Χρησιμοποιεί σχέσεις μεταξύ πινάκων χρησιμοποιώντας πρωτεύον κλειδί, ξένα κλειδιά και δείκτες.
3	Η ανάκτηση δεδομένων είναι πιο αργή για πολύπλοκα και μεγάλης ποσότητας δεδομένα.	Συγκριτικά ταχύτερο λόγω του σχεσιακού μοντέλου.
4	Χρησιμοποιείται για εφαρμογές που χρησιμοποιούν μικρή ποσότητα δεδομένων	Χρησιμοποιείται για πολύπλοκα δεδομένα και για μεγάλες ποσότητες δεδομένων.
5	Τα δεδομένα πλεονασμού είναι κοινά σε αυτό το μοντέλο	Κλειδιά και δείκτες χρησιμοποιούνται στους πίνακες για την αποφυγή πλεονασμών.
6	Παραδείγματα συστημάτων: dBase, Microsoft Access, LibreOffice Base, FoxPro.	Παραδείγματα συστημάτων: SQL Server, Oracle, MySQL, MariaDB, SQLite.

Πίνακας 2.1 DBMS vs RDBMS.

(Twinkle 2012)

2.2.3 Τα Συστήματα διαχείρισης Σχεσιακών Βάσεων Δεδομένων

Σύμφωνα με την εταιρεία ερευνών Gartner, οι πέντε κορυφαίες εμπορικές σχεσιακές βάσεις δεδομένων με βάση τα έσοδα το 2011 ήταν η Oracle (48,8%), η IBM (20,2%), η Microsoft (17,0%), η SAP συμπεριλαμβανομένων Sybase (4,6%), και Teradata (3,7%).

('Relational database management system' 2016)

Στην παρακάτω εικόνα βλέπουμε τα τοπ 15 συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων συμφώνα με την ιστοσελίδα DB-Engines, η οποία τα κατατάσσει συμφώνα με την δημοτικότητα τους. Η κατάταξη ενημερώνεται σε μηνιαία βάση.

116 systems in ranking, March 2016

Rank			DBMS	Database Model	Score		
Mar 2016	Feb 2016	Mar 2015			Mar 2016	Feb 2016	Mar 2015
1.	1.	1.	Oracle	Relational DBMS	1472.01	-4.13	+2.93
2.	2.	2.	MySQL	Relational DBMS	1347.71	+26.59	+86.62
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1136.49	-13.73	-28.31
4.	4.	4.	PostgreSQL	Relational DBMS	299.62	+10.97	+35.19
5.	5.	5.	DB2	Relational DBMS	187.94	-6.55	-10.91
6.	6.	6.	Microsoft Access	Relational DBMS	135.03	+1.95	-6.66
7.	7.	7.	SQLite	Relational DBMS	105.77	-1.01	+4.06
8.	8.	8.	SAP Adaptive Server	Relational DBMS	76.64	-3.39	-8.72
9.	9.	9.	Teradata	Relational DBMS	74.07	+0.69	+1.29
10.	10.	↑ 11.	Hive	Relational DBMS	50.51	-2.26	+11.18
11.	11.	↓ 10.	FileMaker	Relational DBMS	47.93	+0.90	-4.41
12.	12.	↑ 13.	SAP HANA	Relational DBMS	39.99	+1.91	+7.82
13.	13.	↓ 12.	Informix	Relational DBMS	31.87	-1.15	-5.95
14.	14.	14.	MariaDB	Relational DBMS	29.88	+1.11	+7.79
15.	15.	15.	Firebird	Relational DBMS	20.88	+0.76	-1.09

Εικόνα 2.1Τα ποιο δημοφιλή RDBMS για τον Μάρτιο του 2016.

(db-engines 2016)

3 Τα συστήματα διαχείρισης Σχεσιακών Βάσεων Δεδομένων

3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα αναφέρω και θα αναλύσω τα δημοφιλή, εμπορικά συστήματα διαχείρισης Σχεσιακών Βάσεων Δεδομένων, τα οποία έχουν κατακλίσει την αγορά σε παγκόσμιο επίπεδο και χρησιμοποιούνται ευρέως από επιχειρήσεις και οργανισμούς.

Στην εκτενέστερη ανάλυση αυτών των συστημάτων θα αναφερθούν οι ιδιότητες και οι λειτουργείς που υποστηρίζουν για την εξόρυξη δεδομένων (data mining) και για την επιχειρηματική ευφυΐα (Business Intelligence). Τι είναι όμως το data mining και το Business Intelligence?

Σε γενικές γραμμές, η εξόρυξη δεδομένων (data mining) (μερικές φορές ονομάζεται εξόρυξη γνώσης ή ανακάλυψη της γνώσης) είναι η διαδικασία της ανάλυσης των δεδομένων από διαφορετικές οπτικές γωνίες και συνοψίζοντας τα σε χρήσιμες πληροφορίες - πληροφορίες που μπορούν να χρησιμοποιηθούν για την αύξηση των εσόδων, την μείωση το κόστος, ή και τα δύο. Τεχνικά, η εξόρυξη δεδομένων είναι η διαδικασία για την εξεύρεση συσχετίσεων ή μοτίβων ανάμεσα σε δεκάδες πεδία σε μεγάλες σχεσιακές βάσεις δεδομένων. (Palace 1996)

Επιχειρηματική ευφυΐα (Business Intelligence-BI) είναι μια διαδικασία που βασίζεται στην τεχνολογία για την ανάλυση των δεδομένων και την υποβολή έτοιμων προς χρήση πληροφοριών για να βοηθήσει στελέχη επιχειρήσεων, διευθυντές επιχειρήσεων και άλλους τελικούς χρήστες να λαμβάνουν πιο σωστές επιχειρηματικές αποφάσεις. (Rouse 2016)

3.2 Oracle

Η Oracle Database (που συνήθως αναφέρεται ως Oracle RDBMS ή απλά ως Oracle) είναι ένα αντικείμενο-σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων που παράγεται και διατίθεται στο εμπόριο από την Oracle Corporation. Ο Larry Ellison μαζί με δύο φίλους του και πρώην συναδέλφους, Bob Miner και Ed Oates, ξεκίνησαν μια εταιρεία συμβούλων που ονομαζόταν Software Development Laboratories (SDL) το 1977. Η SDL ανέπτυξε την αρχική έκδοση του λογισμικού της Oracle. Το όνομα της

Oracle προέρχεται από την κωδική ονομασία ενός έργου που δούλευε πάνω σε αυτό ο Ellison και χρηματοδοτούνταν από την CIA, όταν ήταν υπάλληλος στην Ampex.

('Oracle Database' 2016)

Στον παρακάτω πινάκα διατίθενται μερικές χρήσιμες πληροφορίες για την Oracle Database.

Περιγραφή	Ευρέως χρησιμοποιούμενη
Μοντέλο Βάσης Δεδομένων	Relational DBMS
Αρχική έκδοση	1980
Τρέχουσα έκδοση	12 Έκδοση 1 (12.1.0.2), Ιούλιος 2014
Άδεια	Εμπορική
Γλώσσα Υλοποίησης	C και C ++
Λειτουργικά Συστήματα	AIX, HP-UX, Linux, OS X, Solaris, Windows, z/OS
XML support	Ναι
SQL	Ναι
APIs και άλλοι μέθοδοι πρόσβασης	ODP.NET, Oracle Call Interface (OCI), JDBC, ODBC
Υποστηριζόμενες γλώσσες προγραμματισμού	C, C#, C++, Clojure, Cobol, Eiffel, Erlang, Fortran, Groovy, Haskell, Java, JavaScript, Lisp, Objective C, OCaml, Perl, PHP, Python R, Ruby, Scala, Tcl, Visual Basic
Σχήμα δεδομένων	Ναι
Πληκτρολόγηση (Προκαθορίζει τύπους δεδομένων όπως float ή date)	Ναι
Ξένα κλειδιά	Ναι
Συγχρονισμός	Ναι
Ανθεκτικότητα	Ναι
Δυνατότητες στην μνήμη	Ναι

Πίνακας 3.1 Χρήσιμες πληροφορίες για Oracle Database.

(db-engines 2016)

Αυτήν την στιγμή βρισκόμαστε στην Oracle Database 12c όπου "c" δηλώνει το "cloud". (Seika 2013) Η βάση δεδομένων έχει λογικές δομές και φυσικές δομές. Επειδή οι φυσικές και λογικές δομές είναι ξεχωριστές, η φυσική αποθήκευση των δεδομένων μπορεί να διαχειριστεί χωρίς να επηρεάζεται η πρόσβαση σε λογικές δομές αποθήκευσης.

3.2.1 Επισκόπηση των φυσικών δομών στην βάση δεδομένων

Οι ακόλουθες ενότητες εξηγούν τις φυσικές δομές της βάσης δεδομένων μιας βάσης δεδομένων της Oracle, συμπεριλαμβανομένων αρχεία δεδομένων, redo αρχεία καταγραφής, και αρχεία ελέγχου.

Αρχεία δεδομένων

Κάθε βάση δεδομένων Oracle έχει ένα ή περισσότερα φυσικά αρχεία δεδομένων. Τα αρχεία δεδομένων περιλαμβάνουν όλα τα δεδομένα της βάσης δεδομένων. Τα δεδομένα των λογικών δομών δεδομένων, όπως πίνακες και ευρετήρια, είναι φυσικά αποθηκευμένα στα αρχεία δεδομένων που διατίθενται για μια βάση δεδομένων.

Τα χαρακτηριστικά των αρχείων δεδομένων είναι:

- Ένα αρχείο δεδομένων μπορεί να συνδέεται με μία μόνο βάση δεδομένων.
- Τα αρχεία δεδομένων έχουν ορισμένα χαρακτηριστικά σετ τα οποία, επεκτείνονται αυτομάτως όταν εξαντληθεί ο χώρος στην βάση δεδομένων.
- Ένα ή περισσότερα αρχεία δεδομένων σχηματίζουν μια λογική μονάδα αποθήκευσης δεδομένων που ονομάζεται tablespace,

Τα δεδομένα στα αρχεία δεδομένων “διαβάζονται”, όπως απαιτείται, κατά την κανονική λειτουργία της ΒΔ και αποθηκεύονται στην κρυφή μνήμη της Oracle. Για παράδειγμα, ας υποθέσουμε ότι ένας χρήστης θέλει να έχει πρόσβαση σε ορισμένα δεδομένα σε έναν πίνακα μιας βάσης δεδομένων. Εάν οι ζητούμενες πληροφορίες δεν είναι ήδη στη κρυφή μνήμη για την ΒΔ, τότε “διαβάζονται” από τα κατάλληλα αρχεία δεδομένων και αποθηκεύονται στη μνήμη. Τα τροποποιημένα ή νέα δεδομένα δεν είναι απαραίτητο να “γράφονται” σε ένα αρχείο δεδομένων αμέσως. Για να μειώσουμε το ποσό της πρόσβασης στο δίσκο και για να αυξήσουμε την απόδοση, τα δεδομένα συγκεντρώνονται στη μνήμα και “γράφονται” στα κατάλληλα αρχεία δεδομένων μονομιάς, όπως καθορίζεται από τη διαδικασία εγγραφής της βάσης δεδομένων (database writer process-DBWn), η οποία είναι διαδικασία που “τρέχει” στο παρασκήνιο.

Αρχεία ελέγχου

Κάθε βάση δεδομένων της Oracle έχει ένα αρχείο ελέγχου. Ένα αρχείο ελέγχου περιέχει καταχωρήσεις που προσδιορίζουν τη φυσική δομή της βάσης δεδομένων.

Για παράδειγμα, περιέχει τις ακόλουθες πληροφορίες:

- Το όνομα της ΒΔ
- Τα ονόματα και τις θέσεις των αρχείων δεδομένων και των redo αρχείων καταγραφής
- Χρονική σήμανση της δημιουργίας ΒΔ

Η Oracle μπορεί να πολλαπλασιάσει ένα αρχείο ελέγχου, δηλαδή, ταυτόχρονα να διατηρήσει μια σειρά από πανομοιότυπα αντίγραφα αρχείων ελέγχου, για την προστασία έναντι βλάβης του αρχείου ελέγχου. Κάθε φορά που ξεκάνει ένα νέο περιστατικό μιας ΒΔ της Oracle, το αρχείο ελέγχου προσδιορίζει τη ΒΔ και τα redo αρχεία καταγραφής έτσι ώστε να ανοίξει την λειτουργία της ΒΔ για να προχωρήσουμε. Εάν η φυσική εμφάνιση της ΒΔ μεταβάλλεται (για παράδειγμα, αν ένα νέο αρχείο δεδομένων ή ένα redo αρχείο καταγραφής δημιουργείται), τότε το αρχείο έλεγχου αυτόματα τροποποιείται από την Oracle για να αντικατοπτρίσει την αλλαγή. Ένα αρχείο ελέγχου χρησιμοποιείται επίσης στην ανάκτηση μιας βάσης δεδομένων (database recovery).

Redo αρχεία καταγραφής

Κάθε βάση δεδομένων της Oracle έχει ένα σετ από δύο ή περισσότερα redo αρχεία καταγραφής. Το σύνολο των redo αρχείων καταγραφής είναι συλλογικά γνωστά ως redo αρχεία για μια ΒΔ. Ένα redo αρχείο αποτελείται από redo εγγραφές που ονομάζονται και redo records. Η κύρια λειτουργία ενός redo αρχείου είναι να καταγράφει όλες τις αλλαγές που έγιναν στα δεδομένα. Εάν μια βλάβη εμποδίσει τροποποιημένα δεδομένα από το να είναι μόνιμα “γραμμένα” στα αρχεία δεδομένων, τότε οι αλλαγές μπορούν αν ληφθούν από τα redo αρχεία καταγραφής, έτσι ώστε να μην χαθεί πότε το έργο σας. Για την προστασία από μια αποτυχία που αφορά τα redo αρχεία καταγραφής, η Oracle επιτρέπει των πολλαπλασιασμό των redo αρχείων, έτσι ώστε δύο ή περισσότερα αντίγραφα των redo αρχείων μπορούν να διατηρηθούν σε διαφορετικούς δίσκους. Οι πληροφορίες σε ένα redo αρχείο καταγραφής χρησιμοποιούνται μονό για την ανάκτηση της βάσης δεδομένων από μια αποτυχία του

συστήματος ή των μέσων που αποτρέπουν τα δεδομένα της ΒΔ να "γραφτούν" στα αρχεία δεδομένων. Για παράδειγμα, αν μια απροσδόκητη διακοπή ρεύματος τερματίσει τη λειτουργία της ΒΔ, τότε τα δεδομένα στη μνήμη δεν μπορούν να "γραφτούν" στα αρχεία δεδομένων με αποτέλεσμα τα δεδομένα να χαθούν. Ωστόσο, τα χαμένα δεδομένα μπορούν να ανακτηθούν όταν ανοίξει η ΒΔ, μετά την αποκατάσταση της τροφοδοσίας. Εφαρμόζοντας τις πληροφορίες στα πιο πρόσφατα redo αρχεία καταγραφής στα αρχεία δεδομένων της ΒΔ, η Oracle επαναφέρει τη ΒΔ στο χρόνο κατά τον οποίο συνέβη η διακοπή ρεύματος. Η διαδικασία της εφαρμογής των redo αρχείων καταγραφής κατά την διάρκεια της διαδικασίας ανάκτησης ονομάζεται κύλισης προς τα εμπρός.

Αρχειοθέτηση στα αρχεία καταγραφής

Μπορείτε να ενεργοποιήσετε την αυτόματη αρχειοθέτηση των redo αρχείων καταγραφής. Η Oracle αρχειοθετεί αυτόματα τα αρχεία καταγραφής όταν η βάση δεδομένων είναι σε λειτουργία ARCHIVELOG.

Αρχεία παραμέτρων

Τα αρχεία παραμέτρων περιέχουν έναν κατάλογο των παραμέτρων διαμόρφωσης για ένα περιστατικό και για μια ΒΔ. Η Oracle συνιστά να δημιουργήσετε έναν διακομιστή για τα αρχεία παραμέτρων (spfile) ως ένα δυναμικό μέσο για τη διατήρηση των παραμέτρων αρχικοποίησης. Ένας διακομιστής αρχείων παραμέτρων σας επιτρέπει να αποθηκεύσετε και να διαχειριστείτε τις παραμέτρους αρχικοποίησης σας επίμονα σε ένα server-side disk file.

Αρχεία καταγραφής για ειδοποιήσεις και ανιχνεύσεις

Κάθε διακομιστής και διεργασία που γίνεται στο παρασκήνιο μπορούν να "γραφτούν" σε ένα σχετικό αρχείο ανίχνευσης. Όταν ένα εσωτερικό σφάλμα ανιχνεύεται από μια διαδικασία, αφήνει πληροφορίες σχετικά με το σφάλμα στο αρχείο ανίχνευσης του. Μερικές από τις πληροφορίες που καταγράφονται σε ένα αρχείο ανίχνευσης προορίζονται για το διαχειριστή της βάσης δεδομένων, ενώ άλλες πληροφορίες είναι για την Υπηρεσία Υποστήριξης της Oracle. Οι πληροφορίες από τα αρχεία ανίχνευσης χρησιμοποιούνται επίσης για να συντονίσουν εφαρμογές και περιστατικά. Το αρχείο ειδοποίησης ή καταγραφή ειδοποίησης είναι ένα ειδικό αρχείο ανίχνευσης. Το αρχείο

ειδοποίησης μιας βάσης δεδομένων είναι μια χρονολογική καταγραφή των μηνυμάτων και των σφαλμάτων.

Εφεδρικά αρχεία

Για να επαναφέρετε ένα αρχείο πρέπει να το αντικαταστήσετε με ένα εφεδρικό αντίγραφο ασφαλείας. Συνήθως, επαναφέρετε ένα αρχείο όταν υπάρχει αποτυχία των μέσων ενημέρωσης ή ένα σφάλμα του χρηστή προκάλεσε ζημιά ή διαγραφή του αρχικού αρχείου.

3.2.2 Επισκόπηση των λογικών δομών στην βάση δεδομένων

Οι λογικές δομές αποθήκευσης, συμπεριλαμβανομένων των μπλοκ δεδομένων, των extents καθώς και των segments, επιτρέπουν στην Oracle να έχει λεπτομερή έλεγχο της χρήσης χώρου στο δίσκο.

Tablespaces

Μια βάση δεδομένων είναι χωρισμένη σε λογικές μονάδες αποθήκευσης που ονομάζονται tablespaces, που συγκεντρώνουν συσχετιζόμενες λογικές δομές μεταξύ τους. Για παράδειγμα, τα tablespaces συνήθως συγκεντρώνουν όλα τα αντικείμενα εφαρμογής για την απλοποίηση ορισμένων διαχειριστικών λειτουργιών.

Ένα ή περισσότερα αρχεία δεδομένων δημιουργούνται ρητά για κάθε tablespace για να αποθηκεύσουν φυσικά τα δεδομένα όλων των λογικών δομών ενός tablespace. Το συνολικό μέγεθος των αρχείων δεδομένων σε ένα tablespace είναι η συνολική χωρητικότητα αποθήκευσης των tablespace. Κάθε βάση δεδομένων της Oracle περιέχει ένα SYSTEM tablespace και ένα SYSAUX tablespace. Η Oracle τα δημιουργεί αυτόματα όταν δημιουργείται μια βάση δεδομένων. Η προεπιλογή του συστήματος είναι να δημιουργήσει ένα smallfile tablespace, που είναι ο παραδοσιακός τύπος tablespace της Oracle. Το SYSTEM και το SYSAUX tablespaces δημιουργούνται ως smallfile tablespaces. Η Oracle σας επιτρέπει επίσης να δημιουργήσετε bigfile tablespaces. Αυτό επιτρέπει στην ΒΔ Oracle να περιέχει tablespaces που αποτελούνται από ενιαία μεγάλα αρχεία και όχι από πολυάριθμα μικρού μεγέθους. Επίσης επιτρέπει στην ΒΔ Oracle να αξιοποιήσει την ικανότητα των συστημάτων 64-bit για να δημιουργήσει και να διαχειριστεί εξαιρετικά μεγάλα αρχεία. Η συνέπεια αυτού είναι ότι η ΒΔ Oracle μπορεί τώρα να κλιμακωθεί μέχρι και σε 8 exabytes σε μέγεθος. Με την

Διαχείριση αρχείων της Oracle, τα bigfile tablespaces κάνουν τα αρχεία δεδομένων απολύτως ολοφάνερα για τους χρήστες. Με άλλα λόγια, μπορείτε να εκτελέσετε λειτουργίες σε tablespaces, παρά στα βαθύτερα στρώματα των αρχείων δεδομένων.

Online και Offline Tablespaces

Ένα tablespace μπορεί να είναι σε απευθείας σύνδεση (προσπελάσιμα) ή έκτος σύνδεσης (μη προσπελάσιμα). Ένα tablespace είναι γενικά σε απευθείας σύνδεση, έτσι ώστε οι χρήστες να μπορούν να έχουν πρόσβαση στις πληροφορίες του tablespace. Ωστόσο, μερικές φορές ένα tablespace μπορεί να είναι εκτός σύνδεσης έτσι ώστε ένα μέρος της βάσης δεδομένων να μην είναι διαθέσιμο, ενώ ταυτόχρονα επιτρέπει κανονική πρόσβαση στο υπόλοιπο της βάσης δεδομένων. Αυτό κάνει πολλά διοικητικά καθήκοντα πιο εύκολα να εκτελεστούν.

Μπλοκ δεδομένων της Oracle

Στο καλύτερο επίπεδο διακριτότητας, τα δεδομένα της βάσης δεδομένων Oracle αποθηκεύονται σε μπλοκ δεδομένων. Ένα μπλοκ δεδομένων αντιστοιχεί σε ένα συγκεκριμένο αριθμό bytes του φυσικού χώρου της βάσης δεδομένων στο δίσκο. Το τυπικό μέγεθος μπλοκ καθορίζεται από την παράμετρο προετοιμασίας DB_BLOCK_SIZE. Επιπλέον, μπορείτε να καθορίσετε έως και άλλα πέντε μεγέθη μπλοκ δεδομένων. Μια βάση δεδομένων χρησιμοποιεί και διαθέτει δωρεάν χώρο βάσης δεδομένων σε μπλοκ δεδομένων της Oracle.

Extents

Το επόμενο επίπεδο της λογικής βάσης δεδομένων χώρου είναι ένα extent. Ένα extent είναι ένας συγκεκριμένος αριθμός συνεχόμενων μπλοκ δεδομένων, που λαμβάνονται σε μια μόνο κατανομή, και χρησιμοποιείται για την αποθήκευση ενός ειδικού είδους πληροφορίας .

Segments

Πάνω από τα extents, στο επίπεδο της λογικής αποθήκευσης δεδομένων είναι τα Segments. Ένας Segment είναι ένα σύνολο από extents που διατίθενται για μια ορισμένη λογική δομή. Υπάρχουν διάφορα είδη segment τα οποία είναι: Data segment, Index segment, Temporary segment και Rollback segment. Η Oracle κατανέμει δυναμικά χώρο όταν τα υπάρχουσα extents ενός segment γίνουν πλήρες. Με άλλα λόγια, όταν τα extents ενός segment είναι πλήρη, η Oracle εκχωρεί ένα άλλο extent για

το συγκεκριμένο segment. Επειδή τα extents κατανέμονται ανάλογα με τις ανάγκες, τα extents ενός segment μπορεί να είναι ή μπορεί να μην είναι συνεχόμενα στο δίσκο. (Cyran et al. 2005)

3.2.3 Schema Αντικείμενα

Ένα χαρακτηριστικό ενός RDBMS είναι η ανεξαρτησία της φυσικής αποθήκευσης δεδομένων από τις λογικές δομές δεδομένων. Στην ΒΔ της Oracle, ένα schema είναι μια συλλογή από λογικές δομές δεδομένων, ή schema αντικειμένων. Ένας χρήστης της ΒΔ έχει ένα schema, το οποίο έχει το ίδιο όνομα με το όνομα χρήστη. Τα schema αντικειμένα είναι δομές που δημιουργούνται από τον χρήστη και αναφέρονται άμεσα στα δεδομένα της ΒΔ. Η ΒΔ υποστηρίζει πολλούς τύπους αντικειμένων schema, τα πιο σημαντικά είναι οι πίνακες και τα ευρετήρια. Ένα schema αντικείμενο είναι ένα είδος αντικειμένου μιας βάσης δεδομένων. Ορισμένα αντικείμενα βάσης δεδομένων, όπως τα προφίλ και οι ρόλοι, δεν ανήκουν στα αντικείμενα schema.

3.2.4 Πίνακες

Ένας πίνακας περιγράφει μια οντότητα όπως είναι οι εργαζόμενοι. Μπορείτε να καθορίσετε έναν πινάκα με ένα όνομα όπως εργαζόμενοι και το σύνολο των στηλών του πινάκα. Σε γενικές γραμμές, θα δώσετε σε κάθε στήλη ένα όνομα, έναν τύπο δεδομένων, και θα ορίσετε το πλάτος της στήλης κατά τη δημιουργία του πίνακα. Ένας πίνακας είναι ένα σύνολο γραμμών. Μια στήλη προσδιορίζει ένα χαρακτηριστικό της οντότητας που περιγράφεται από τον πίνακα, ενώ μια σειρά προσδιορίζει ένα περιστατικό της οντότητας. Για παράδειγμα, τα χαρακτηριστικά της οντότητας εργαζόμενοι αντιστοιχούν σε στήλες για ID εργαζομένων και επώνυμο. Ενώ μια σειρά προσδιορίζει ένα συγκεκριμένο εργαζόμενο. Μπορείτε προαιρετικά να ορίσετε έναν κανόνα, που ονομάζεται περιορισμός ακεραιότητας, για μια στήλη. Ένα παράδειγμα είναι ο NOT NULL περιορισμός ακεραιότητας. Αυτός ο περιορισμός αναγκάζει τη στήλη να περιέχει μια τιμή σε κάθε γραμμή.

3.2.5 Ευρετήρια

Ένα ευρετήριο είναι μια προαιρετική δομή δεδομένων που μπορείτε να δημιουργήσετε σε μία ή περισσότερες στήλες ενός πίνακα. Τα ευρετήρια μπορούν να αυξήσουν την απόδοση της ανάκτησης δεδομένων. Κατά την επεξεργασία μιας αίτησης, η βάση δεδομένων μπορεί να χρησιμοποιήσει τα διαθέσιμα ευρετήρια για να εντοπίσετε τις

ζητούμενες σειρές αποτελεσματικά. Τα ευρετήρια είναι χρήσιμα όταν οι εφαρμογές χρησιμοποιούν συχνά ερωτήματα σε μια συγκεκριμένη γραμμή ή σε σειρά γραμμών. Τα ευρετήρια είναι λογικά και φυσικά ανεξάρτητα από τα δεδομένα. Αυτό μας οδηγεί στην άνεση να διαγράφουμε, να δημιουργούμε ευρετήρια οποιαδήποτε στιγμή καθώς δεν δημιουργούν καμία επίδραση στους πίνακες ή σε άλλα ευρετήρια. Όλες οι εφαρμογές να συνεχίσουν να λειτουργούν χωρίς προβλήματα μετά την διαγραφή ενός ευρετηρίου.

3.2.6 Πρόσβαση στα δεδομένα

Μια γενική απαίτηση για τα DBMS είναι να συμμορφώνονται με αποδεκτά πρότυπα του κλάδου για μια γλωσσά προγραμματισμού, η οποία προσφέρει πρόσβαση στα δεδομένα.

Structured Query Language (SQL)

Η SQL βασίζεται στη δηλωτική γλώσσα και παρέχει μια διεπαφή για μια RDBMS, όπως η ΒΔ της Oracle. Σε αντίθεση με τις διαδικαστικές γλώσσες όπως η C, που περιγράφουν πώς πρέπει να γίνουν τα πράγματα, η SQL δεν είναι διαδικαστική γλωσσά προγραμματισμού και γι αυτόν τον λόγο περιγράφει τι πρέπει να γίνει. Η SQL είναι η τυπική γλώσσα για σχεσιακές βάσεις δεδομένων συμφώνα με το ANSI. Όλες οι εργασίες σχετικά με τα δεδομένα σε μια βάση δεδομένων της Oracle πραγματοποιούνται με τη χρήση SQL εντολών. Για παράδειγμα, μπορείτε να χρησιμοποιήσετε την SQL για τη δημιουργία πινάκων και ερωτημάτων και την τροποποίηση των δεδομένων στους πίνακες. Μια δήλωση στην SQL μπορεί να θεωρηθεί ως ένα πολύ απλό, αλλά ισχυρό, πρόγραμμα ηλεκτρονικού υπολογιστή ή εντολή. Οι χρήστες καθορίζουν το αποτέλεσμα που θέλουν (για παράδειγμα, τα ονόματα των υπαλλήλων), και όχι πώς θα το παράγουν. Μια δήλωση στην SQL είναι μια σειρά από SQL εντολές κειμένου, όπως η ακόλουθη:

```
SELECT first_name, last_name FROM employees;
```

Οι SQL εντολές σας επιτρέπουν να εκτελέσετε τις ακόλουθες εργασίες:

- Ερωτήσεις στα δεδομένα
- Εισαγωγή, ενημέρωση και διαγραφή γραμμών σε έναν πίνακα
- Δημιουργία, αντικατάσταση, αλλαγή και διαγραφή αντικειμένων

- Έλεγχος της πρόσβασης στη βάση δεδομένων και τα αντικείμενά της
- Εγγύηση συνέπειας και ακεραιότητας στην βάση δεδομένων

Η SQL ενοποιεί τις προηγούμενες εργασίες σε μία συνεπή γλώσσα. Η SQL της Oracle είναι μια εκτέλεση του προτύπου ANSI. Η SQL της Oracle υποστηρίζει πολλά χαρακτηριστικά που εκτείνονται πέρα από την τυπική SQL.

PL/SQL και Java

Η PL / SQL είναι μια διαδικαστική επέκταση της Oracle SQL. Η PL / SQL είναι ενσωματωμένη με την ΒΔ της Oracle, δίνοντας την δυνατότητα στον χρήστη να χρησιμοποιήσει όλες τις εντολές SQL απο την ΒΔ της Oracle, τις λειτουργίες και τους τύπους δεδομένων. Μπορείτε να χρησιμοποιήσετε την PL / SQL για τον έλεγχο της ροής του προγράμματος SQL, για να χρησιμοποιήσετε τις μεταβλητές, και για να γράφουν διαδικασίες χειρισμού σφαλμάτων. Ένα κύριο πλεονέκτημα της PL / SQL είναι η ικανότητα να αποθηκεύει λογική εφαρμογή στην ίδια την βάση δεδομένων. Μια PL / SQL διαδικασία ή λειτουργία είναι ένα schema αντικείμενο που αποτελείται από ένα σύνολο από δηλώσεις SQL και άλλες δομές PL / SQL οι οποίες, ομαδοποιούνται και αποθηκεύονται στη βάση δεδομένων και λειτουργούν ως μονάδα για να λύσουν ένα συγκεκριμένο πρόβλημα ή για να εκτελέσουν μια σειρά από σχετικές εργασίες. Το κύριο όφελος του server-side προγραμματισμού είναι ότι οι ενσωματωμένες λειτουργίες μπορούν να αναπτυχτούν οπουδήποτε. Η ΒΔ της Oracle μπορεί επίσης να αποθηκεύσει μονάδες προγραμμάτων που είναι γραμμένα σε Java. Τα προγράμματα που είναι γραμμένα σε Java δημοσιεύονται στην SQL και αποθηκεύονται στην ΒΔ για γενική χρήση. Υπάρχει η δυνατότητα να καλέσετε τα υφιστάμενα προγράμματα της PL / SQL από την Java και τα προγράμματα της Java από την PL / SQL.
(Ashdown & Kyte et al. 2015)

3.2.7 Oracle Advanced Analytics

Η Oracle Advanced Analytics 12c παραδίδει στη βάση δεδομένων παραλληλοποιημένες εκτελέσεις των αλγορίθμων εξόρυξης δεδομένων και τα ενσωματώνει στο πρόγραμμα ανοιχτού κώδικα R. Οι αναλυτές δεδομένων χρησιμοποιούν Oracle Data Miner GUI και R για να χτίσουν και να αξιολογήσουν

προγνωστικά μοντέλα καθώς και τα πλεονεκτήματα των πακέτων και γραφικών του προγράμματος R. Οι προγραμματιστές εφαρμογών αναπτύσσουν μοντέλα Oracle Advanced Analytics χρησιμοποιώντας τις λειτουργίες εξόρυξης δεδομένων της SQL καθώς και το πρόγραμμα R. Με την επιλογή Oracle Advanced Analytics, η Oracle επεκτείνει τη βάση δεδομένων της Oracle σε μια επεκτάσιμη πλατφόρμα ανάλυσης η όποια, κάνει εξόρυξη δεδομένων σε περισσότερα δεδομένα και τύπους δεδομένων, εξαλείφει την κίνηση δεδομένων, διατηρεί την ασφάλεια να προβλέψουμε τη συμπεριφορά των πελατών, ανιχνεύει πρότυπα, και παραδίδει πρακτικές πληροφορίες.

Η Oracle Big Data SQL προσθέτει νέες μεγάλες πηγές δεδομένων και η Oracle R Advanced Analytics για Hadoop παρέχει αλγόριθμους που τρέχουν σε Hadoop. Η Oracle Advanced Analytics, είναι ένας συνδυασμός της Oracle Data Mining και Oracle R Enterprise, προσφέρει προγνωστικά ανάλυσης, εξόρυξη δεδομένων, εξόρυξη κειμένου, στατιστική ανάλυση, προχωρημένους αριθμητικούς υπολογισμούς και διαδραστικά γραφικά στο εσωτερικό της βάσης δεδομένων. Φέρνει ισχυρούς υπολογισμούς στη βάση δεδομένων με αποτέλεσμα δραματικές βελτιώσεις στην ανακάλυψη πληροφοριών, την επεκτασιμότητα, την ασφάλεια και την εξοικονόμηση.

Προγνωστικά Analytics και πρακτικές πληροφορίες:

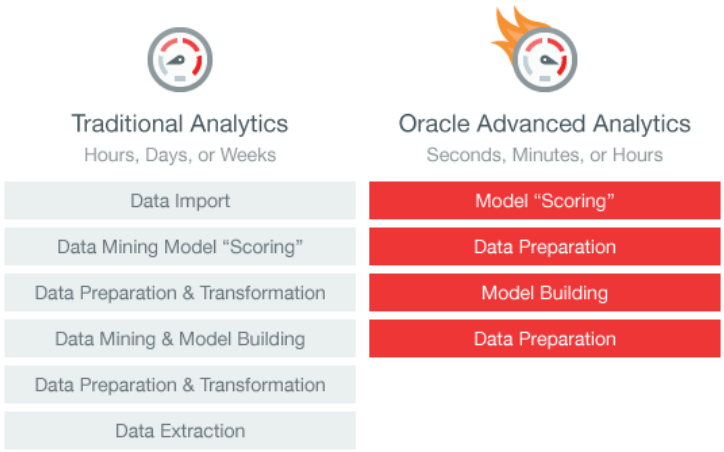
- Πρόβλεψη για την συμπεριφορά των πελατών και εντόπιση cross/up-sell ευκαιρίες
- Προσδιορισμός πότε κατά πάσα πιθανόν ένας πελάτης θα διακόψει την σχέση του με την εταιρία
- Ανάλυση των «καλαθίων της αγοράς» για την ανακάλυψη των ενώσεων, τα πρότυπα και τις σχέσεις μεταξύ των προϊόντων
- Εντοπισμός ανωμαλιών και καταπολέμηση της απάτης

Οι υψηλές επιδόσεις των αλγορίθμων εξόρυξης δεδομένων και οι στατιστικές λειτουργίες σε μια βάση δεδομένων είναι προσβάσιμες από την SQL και από το πρόγραμμα R. Η ενοποίηση με το πρόγραμμα ανοιχτού κώδικα R προσθέτει τη δυνατότητα στους χρήστες να γράφουν R scripts και να χρησιμοποιούν τα πακέτα R, ενώ ταυτόχρονα αξιοποιούν τα πλεονεκτήματα της βάσης δεδομένων.

Γραφικά περιβάλλοντα, το Oracle Data Miner είναι μια επέκταση της Oracle SQL Developer, παρέχει αναλυτές δεδομένων με ένα περιβάλλον ροή εργασίας για την εξερεύνηση των δεδομένων, δημιουργία μοντέλων και την ανάπτυξη αναλυτικών μεθοδολογιών. Μόλις ο αναλυτής δεδομένων είναι ικανοποιημένος με την αναλυτική μεθοδολογία του, το Oracle Data Miner μπορεί να δημιουργήσει SQL scripts για την άμεση ανάπτυξη της ΒΔ εξοικονομώντας χρόνο και χρήμα.

Βασικά πλεονεκτήματα της Oracle Advanced Analytics:

- Απόδοση και επεκτασιμότητα: διαχείριση όλων των δεδομένων, προετοιμασία και μετασχηματισμός δεδομένων και δημιουργία μοντέλων.
- Ο Γρηγορότερος τρόπος για να προσφέρεις εφαρμογές για επιχειρησιακές προβλέψεις και αναλύσεις, η βάση δεδομένων γίνεται η κλιμακωτή και ασφαλή πλατφόρμα για την παροχή προβλέψεων και ιδεών για BI dashboards και εφαρμογών.
- Χαμηλότερο συνολικό κόστος ιδιοκτησίας, δεν υπάρχει ανάγκη για ξεχωριστό servers για της αναλύσεις.



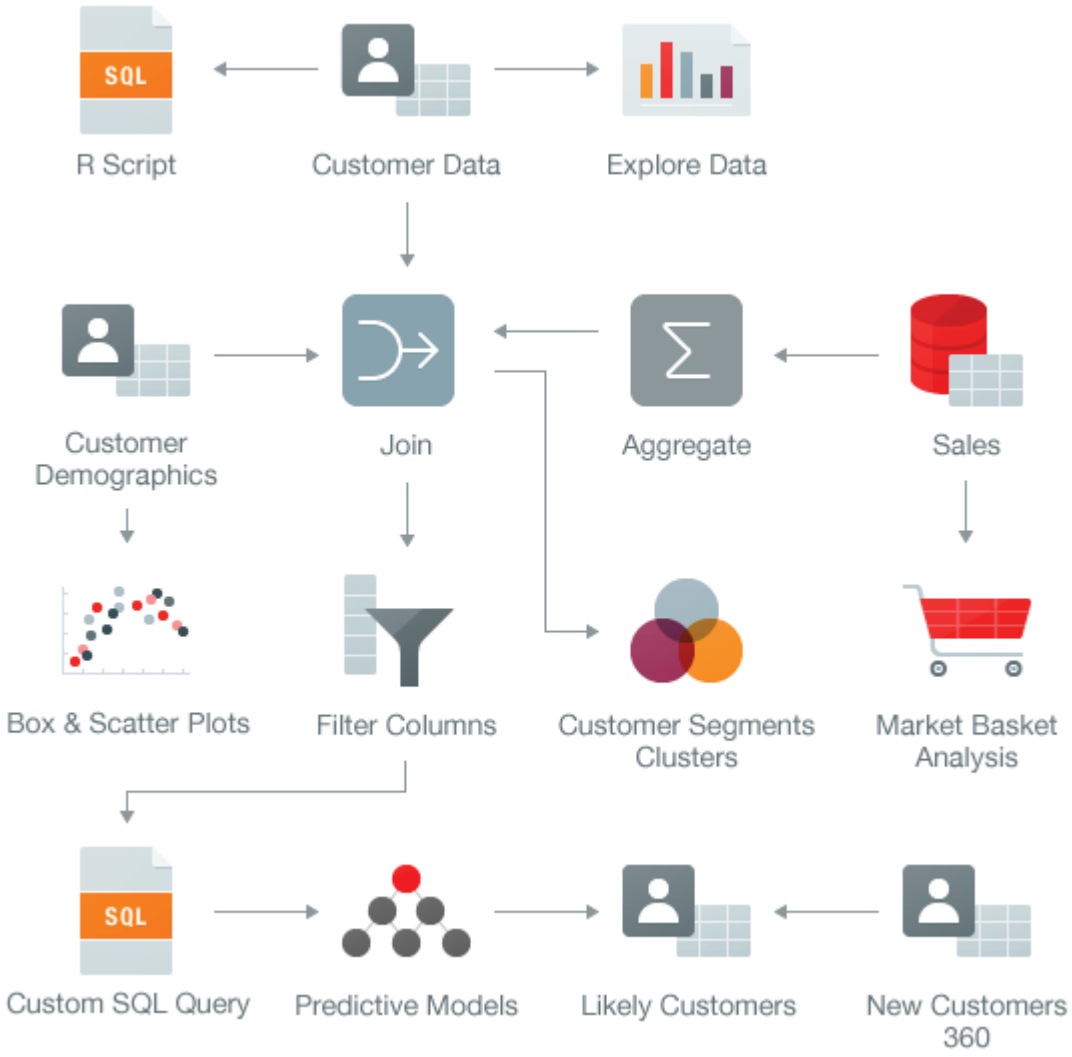
Εικόνα 3.1 Βασικές διαφορές Oracle Advanced Analytics από Συμβατικά Analytics.

(Oracle n.d.)

3.2.8 Oracle Data Mining

Το Oracle Data Mining (ODM), ένα συστατικό της Oracle Advanced Analytics, παρέχει ισχυρούς αλγορίθμους εξόρυξης δεδομένων που επιτρέπουν στους αναλυτές δεδομένων να ανακαλύψουν ιδέες, να κάνουν προβλέψεις, να αξιοποιούν τα δεδομένα τους και να κάνουν επενδύσεις. Με το ODM, μπορείτε να δημιουργήσετε και να εφαρμόσετε προβλεπτικά μοντέλα στο εσωτερικό της βάσης δεδομένων της Oracle τα οποία, θα σας βοηθήσουν να προβλέψετε τη συμπεριφορά των πελατών, να «στοχεύσετε» καλυτέρα τους πελάτες σας, να αναπτύξετε τα προφίλ των πελατών, να εντοπίσετε cross-selling ευκαιρίες, να ανιχνεύσετε ανωμαλίες και πιθανές απάτες. Οι αλγόριθμοι εκτελούνται ως λειτουργίες SQL και αξιοποιούν τα πλεονεκτήματα της βάσης δεδομένων της Oracle. Οι λειτουργίες εξόρυξης δεδομένων SQL μπορούν να εξορύξουν πίνακες δεδομένων και star schema δεδομένα συμπεριλαμβανομένων των δεδομένων συναλλαγών, συσσωματωμένων δεδομένων, αδόμητων δεδομένων δηλαδή τύπο δεδομένων CLOB και χωρικών δεδομένων. Η Oracle Advanced Analytics και οι SQL λειτουργίες εξόρυξης δεδομένων χρησιμοποιούν πλήρως το πλεονέκτημα από τον παραλληλισμό της βάσης δεδομένων για την κατασκευή μοντέλων και για την εφαρμογής τους και τιμά όλα τα δεδομένα και τα προνόμια των χρηστών και των συστημάτων ασφαλείας. Τα μοντέλα πρόβλεψης μπορούν να συμπεριληφθούν σε SQL ερωτήματα, BI dashboards και να ενσωματωθούν σε εφαρμογές πραγματικού χρόνου.

Oracle Data Miner GUI, μια επέκταση της Oracle SQL Developer, επιτρέπει αναλυτές δεδομένων, επιχειρηματικούς αναλυτές και επιστήμονες δεδομένων να εργαστούν άμεσα με τα δεδομένα στο εσωτερικό της βάσης δεδομένων χρησιμοποιώντας το γραφικό "drag and drop" στην ροή εργασιών και το συστατικό παλέτα. Στο Oracle Data Miner η ροή εργασιών συλλαμβάνει και αποθηκεύει σε ένα έγγραφο αναλυτικά τη μεθοδολογία του χρήστη και μπορεί να το μοιράσει σε άλλους για την αυτοματοποίηση των αναλυτικών μεθόδων. Το Oracle Data Miner μπορεί να δημιουργήσει SQL και PL / SQL scripts για την αυτοματοποίηση μοντέλων, το χρονοδιάγραμμα και την παράταξη σε ολόκληρη την επιχείρηση.



Εικόνα 3.2 Μερικές λειτουργίες του Oracle Data Miner.

Το Oracle Data Miner δημιουργεί μοντέλα πρόβλεψης που οι προγραμματιστές εφαρμογών μπορούν να τα ενσωματώσουν στις εφαρμογές για να αυτοματοποιήσουν την ανακάλυψη και τη διανομή των νέων προβλέψεων της επιχειρηματικής ευφυΐας, τα πρότυπα και τις ανακαλύψεις σε όλη την επιχείρηση.

(Oracle n.d.)

3.2.9 Business Intelligence

Η ανάλυση των πληροφοριών ενός οργανισμού είναι ένα βοήθημα για τη λήψη επιχειρηματικών αποφάσεων και είναι γνωστή ως επιχειρηματική ευφυΐα. Οι αναλυτικές εφαρμογές και η επιχειρηματική ευφυΐα κυριαρχούνται από τις ανακατατάξεις στις ιεραρχίες και από τις συγκρίσεις σε συνολικές τιμές μια μεταβλητής. Η Oracle Database παρέχει διάφορες τεχνολογίες για να υποστηρίξει τέτοιες λειτουργιές.

Analytic SQL, η Oracle Database εισήγαγε πολλές λειτουργίες SQL για την εκτέλεση των πράξεων ανάλυσης. Οι λειτουργιές αυτές περιλαμβάνουν κατατάξεις, κινητούς μέσους όρους, σωρευτικά αθροίσματα, αναλογία προς τις εκθέσεις και περίοδο-επί-περίοδο συγκρίσεις. Για παράδειγμα, η Oracle Database υποστηρίζει τις ακόλουθες μορφές Analytic SQL:

- SQL για aggregation: Μια aggregate function , όπως η COUNT επιστρέφει ένα αποτέλεσμα το οποίο, είναι μια γραμμή η οποία, βασίσετε σε μια ομάδα γραμμών. Η aggregation είναι θεμελιώδους σημασίας για της αποθήκες δεδομένων. Για τη βελτίωση των επιδόσεων με την aggregation σε αποθήκες δεδομένων, η βάση δεδομένων παρέχει επεκτάσεις GROUP BY για να κάνει την αναζήτηση και την υποβολή ερωτημάτων ευκολότερα και ταχύτερα.
- SQL για analysis: Μια analytic function , όπως η MAX ομαδοποιεί ομάδες γραμμών για να επιστρέψουν πολλαπλές γραμμές ως σύνολο αποτελεσμάτων. Η Oracle έχει προηγμένες δυνατότητες αναλυτικής επεξεργασίας SQL χρησιμοποιώντας μια οικογένεια των αναλυτικών λειτουργιών της SQL. Για παράδειγμα, αυτές οι analytic functions σας επιτρέπουν να υπολογίσετε βαθμολογίες, εκατοστημόρια κ.ά.
- SQL για modeling: Με την εντολή MODEL, μπορείτε να δημιουργήσετε έναν πολυδιάστατο πινάκα από τα αποτελέσματα των ερωτημάτων σας στον οποίο, μπορείτε να κάνετε όλες τις λειτουργιές της SQL και να βγάλετε νέα αποτελέσματα. Για παράδειγμα, μπορείτε να χωρίσετε τα δεδομένα σε μια προβολή των πωλήσεων ανά χώρα και να εκτελέσετε ένα μοντέλο υπολογισμού, όπως ορίζεται από πολλούς κανόνες, σε κάθε χώρα. Ένας κανόνας θα μπορούσε να υπολογίσει τις πωλήσεις ενός προϊόντος το 2008 ως το άθροισμα των πωλήσεων το 2006 και το 2007.

3.2.10 OLAP

Το Oracle online analytical processing (OLAP) παρέχει πολυδιάστατη αποθήκευση και ταχεία απόκριση χρόνου κατά την ανάλυση των δεδομένων σε πολλαπλές διαστάσεις. Το OLAP επιτρέπει στους αναλυτές να αποκτήσουν γρήγορα απαντήσεις σε σύνθετα, επαναληπτικά ερωτήματα ακόμα και κατά τη διάρκεια διακρατικών συνέδριων.

Το Oracle OLAP έχει τα ακόλουθα βασικά χαρακτηριστικά:

- Το Oracle OLAP είναι ενσωματωμένο στη βάση δεδομένων, έτσι ώστε να μπορείτε να χρησιμοποιήσετε το διοικητικό πρότυπο SQL, την δημιουργία ερωτημάτων, την αναζήτηση και τα εργαλεία αναφοράς.
- Ο κινητήρας OLAP τρέχει μέσα στον πυρήνα της Oracle Database.
- Τα τρισδιάστατα αντικείμενα αποθηκεύονται στην Oracle Database με την μητρική πολυδιάστατη μορφή τους.
- Οι κύβοι και τα άλλα τρισδιάστατα αντικείμενα ανήκουν στην πρώτη κατηγορία αντικειμένων δεδομένων και αντιπροσωπεύονται στο λεξικό δεδομένων της Oracle.
- Η ασφάλεια των δεδομένων διαχειρίζεται με τον κανονικό τρόπο, με τη χορήγηση και την ανάκληση των προνομίων της Oracle Database στους χρηστές της.

Το Oracle OLAP προσφέρει τη δύναμη της απλότητας: μία βάση δεδομένων, πρότυπο διοίκησης και ασφάλειας, πρότυπα διεπαφών και εργαλεία ανάπτυξης. (Ashdown & Kyte et al. 2015)

3.3 MySQL

Η MySQL είναι ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων ανοιχτού κώδικα, όπου τον Ιούλιο του 2013, ήταν η δεύτερη πιο διαδεδομένη RDBMS στον κόσμο, και το πιο ευρέως χρησιμοποιούμενο ανοιχτού κώδικα μοντέλο client-server RDBMS. Η MySQL πηρέ την ονομασία της από το όνομα My της κόρης του συν-ιδρυτή της Michael Widenius, η συντομογραφία SQL αναφέρεται στην Structured Query Language. Η MySQL έχει κάνει τον πηγαίο κώδικα της διαθέσιμο υπό τους όρους της GNU General Public License, καθώς και από μια ποικιλία από ιδιόκτητες συμφωνίες. Η MySQL ανήκε και χρηματοδοτούνταν από μια κερδοσκοπική σουηδική εταιρία την MySQL AB, πλέον ανήκει στην Oracle Corporation.

Η MySQL είναι μια δημοφιλής επιλογή βάσης δεδομένων για χρήση εφαρμογών web, και αποτελεί κύριο συστατικό της ευρέως χρησιμοποιούμενης πλατφόρμας ανοιχτού κώδικα LAMP stack. Το LAMP είναι ένα ακρωνύμιο για «Linux, Apache, MySQL, Perl / PHP / Python». Πολλά ελεύθερα λογισμικά ανοικτού κώδικα που απαιτούν ένα πλήρως εξοπλισμένο σύστημα διαχείρισης βάσεων δεδομένων συχνά χρησιμοποιούν την MySQL. Μερικές από τις εφαρμογές που χρησιμοποιούν τη βάση δεδομένων MySQL είναι οι εξής: TYPO3, MODx, Joomla, WordPress, phpBB, MyBB, Drupal κ.ά. ('MySQL' 2016) Η My SQL χρησιμοποιείτε επίσης και από διάσημες, μεγάλης κλίμακας επιχειρήσεις όπως: YouTube, PayPal, Google, facebook, Twitter, ebay, cisco, LinkedIn και Adobe. (mysql n.d.)

Στον παρακάτω πινάκα διατίθενται μερικές χρήσιμες πληροφορίες για την MySQL

Περιγραφή	Ευρέως χρησιμοποιούμενη open source RDBMS
Μοντέλο βάσης δεδομένων	Relational DBMS
Developer	Αρχικά από την MySQL AB έπειτα στην Sun Microsystems πλέον στην Oracle από το 2010.
Αρχική έκδοση	1995
Τρέχουσα έκδοση	5.7.11 από τον Φεβρουάριο του 2016
Άδεια	Ανοιχτού κώδικα (GPL version 2. Εμπορικές άδειες με εκτεταμένη λειτουργικότητα είναι διαθέσιμες)
Γλωσσά υλοποίησης	C και C++
Λειτουργικά συστήματα	FreeBSD, Linux, OS X, Solaris καιWindows
Σχήμα δεδομένων	Ναι
XML support	Ναι
SQL	Ναι
Πληκτρολόγηση (Προκαθορίζει τύπους δεδομένων όπως float ή date)	Ναι
Ξένα κλειδιά	Ναι
APIs και άλλοι μέθοδοι πρόσβασης	ADO.NET, JDBC και ODBC.
Υποστηριζόμενες γλώσσες προγραμματισμού	Ada, C, C#, C++, D, Eiffel, Erlang, Haskell, Java, Objective-C, OCaml, Perl, PHP, Python, Ruby, Scheme και Tcl
Συγχρονισμός	Ναι
Ανθεκτικότητα	Ναι
Δυνατότητες στην μνήμη	Ναι

Πίνακας 3.2 Χρήσιμες πληροφορίες για την MySQL.

(db-engines 2016)

3.3.1 Τα κύρια χαρακτηρίστηκα της MySQL

Σε αυτό το τμήμα θα περιγράψω μερικά από τα σημαντικότερα χαρακτηριστικά του λογισμικού βάσεων δεδομένων MySQL.

- Για φορητότητα, χρησιμοποιεί CMake από την MySQL 5.5 και σε νεότερες εκδόσεις, στις προηγούμενες εκδόσεις χρησιμοποιούσε το GNU Automake, Autoconf και Libtool.
- Έχει δοκιμαστεί με Purify (ένα εμπορικό ανιχνευτή διαρροής μνήμης), καθώς και με ένα εργαλείο GPLto Valgrind.
- Χρησιμοποιεί πολυεπίπεδο σχεδιασμό για τους σέρβερ της και ανεξάρτητα modules.
- Είναι σχεδιασμένη για να είναι πλήρως multi-threaded χρησιμοποιώντας kernel threads, μπορεί να χρησιμοποιήσει εύκολα πολλαπλές CPUs, εφόσον είναι διαθέσιμες.
- Χρησιμοποιεί πολύ γρήγορα B-tree πινάκες που βρίσκονται στον δίσκο (MyISAM) με την συμπίεση ευρετηρίων.
- Έχει σχεδιαστεί ώστε να είναι σχετικά εύκολο να προστεθούν άλλες μηχανές αποθήκευσης. Αυτό είναι χρήσιμο αν θέλετε να παρέχετε μια διεπαφή SQL για μια βάση δεδομένων in-house.
- Χρησιμοποιεί ένα αρκετά γρήγορο σύστημα κατανομής μνήμης το οποίο είναι thread-based.
- Εκτελεί πολύ γρήγορα joins χρησιμοποιώντας ένα βελτιστοποιημένο nested-loop join.
- Εφαρμόζει πίνακες κατακερματισμού στη μνήμη, οι οποίοι χρησιμοποιούνται ως προσωρινοί πίνακες.
- Υλοποιεί λειτουργίες SQL χρησιμοποιώντας μια ιδιαίτερα βελτιστοποιημένη βιβλιοθήκη κλάσης, η οποία θα πρέπει να είναι όσο το δυνατόν γρήγορη. Συνήθως δεν υπάρχει κατανομή μνήμης καθόλου μετά από ερωτήματα προετοιμασίας.
- Παρέχει τον διακομιστή ως ένα ξεχωριστό πρόγραμμα για χρήση σε ένα client / server δικτυωμένο περιβάλλον, και ως βιβλιοθήκη που μπορεί να ενσωματωθεί (συνδεθεί) σε αυτόνομες εφαρμογές. Τέτοιες εφαρμογές μπορούν να

χρησιμοποιηθούν μεμονωμένα ή σε περιβάλλοντα όπου η χρήση δικτύου δεν είναι διαθέσιμη.

Τύποι δεδομένων

- Πολλοί τύποι δεδομένων: signed/unsigned integers 1, 2, 3, 4, και 8 bytes μέγεθος, FLOAT, DOUBLE, CHAR, VARCHAR, BINARY, VARBINARY, TEXT, BLOB, DATE, TIME, DATETIME, TIMESTAMP, YEAR, SET, ENUM και OpenGIS spatial τύποι.
- Fixed-length και variable-length τύποι string.

Statements και Functions

- Πλήρης χειρισμός και υποστήριξη της εντολής SELECT και WHERE, των ερωτημάτων. Για παράδειγμα:

```
mysql> SELECT CONCAT(first_name, ' ', last_name)
    -> FROM citizen
    -> WHERE income/dependents > 10000 AND age > 30;
```

- Πλήρης υποστήριξη για SQL GROUP BY και ORDER BY. Υποστήριξη για τις functions (COUNT(), AVG(), STD(), SUM(), MAX(), MIN(), and GROUP_CONCAT()).
- Υποστήριξη για LEFT OUTER JOIN και RIGHT OUTER JOIN τόσο με το πρότυπο της SQL όσο και με την ODBC σύνταξη.
- Υποστήριξη για ψευδώνυμα σε πίνακες και στήλες, όπως απαιτείται από το πρότυπο της SQL.
- Υποστήριξη για τις εντολές DELETE, INSERT, REPLACE, και UPDATE για να επιστρέψουν τον αριθμό των γραμμών που άλλαξαν (επηρεάστηκαν), ή να επιστρέψουν τον αριθμό των γραμμών που ορίστηκαν με σημαία κατά τη σύνδεση με το διακομιστή.
- Συγκεκριμένη υποστήριξη της MySQL για την εντολή SHOW η οποία, ανακτά πληροφορίες σχετικά με την βάση δεδομένων, για μηχανές αποθήκευσης, πίνακες, και ευρετήρια. Η MySQL 5.0 προσθέτει υποστήριξη INFORMATION_SCHEMA στην βάση δεδομένων και εκτελείτε σύμφωνα με το πρότυπο της SQL.

- Παρέχει μια εντολή EXPLAIN για να δείξει πώς η βελτιστοποίηση επιλύει ένα ερώτημα.
- Παρέχει ανεξαρτησία των ονομάτων των functions από τα ονόματα των πινάκων ή των στηλών. Για παράδειγμα, το ABS είναι ένα έγκυρο όνομα μιας στήλης. Ο μόνος περιορισμός είναι ότι για την κλήση της συνάρτησης, δεν επιτρέπονται τα κενά διαστήματα μεταξύ του ονόματος της συνάρτησης και του "(" που ακολουθεί έπειτα.
- Μπορείτε να αναφέρεστε σε πίνακες από διαφορετικές βάσεις δεδομένων στην ίδια δήλωση.

Ασφάλεια

- Παρέχει ένα σύστημα για κωδικούς και δικαιώματα το οποίο, είναι πολύ ευέλικτο, ασφαλές και επιτρέπει την host-based επαλήθευση.
- Παρέχει ασφάλεια κωδικού πρόσβασης με κρυπτογράφηση όλης της κυκλοφορίας κωδικών πρόσβασης όταν συνδέεστε σε ένα διακομιστή.

Επεκτασιμότητα και Όρια

- Υποστήριξη για μεγάλες βάσεις δεδομένων. Χρησιμοποιούμε MySQL Server με βάσεις δεδομένων που περιέχουν 50 εκατομμύρια εγγραφές. Γνωρίζουμε επίσης χρήστες που χρησιμοποιούν MySQL Server με 200.000 πίνακες και περίπου 5 δισεκατομμύρια σειρές.
- Υποστήριξη έως και 64 ευρετήρια ανά πίνακα. Κάθε ευρετήριο μπορεί να αποτελείται από 1 έως 16 στήλες ή τμήματα στηλών. Το μέγιστο πλάτος του ευρετηρίου είναι 767 bytes για πίνακες InnoDB, ή 1000 bytes για πίνακες MyISAM. Ένα ευρετήριο μπορεί να χρησιμοποιεί ένα πρόθεμα μιας στήλης για τους τύπους CHAR, VARCHAR, BLOB ή TEXT.

Συνδεσιμότητα

- Οι πελάτες μπορούν να συνδεθούν με τον Server της MySQL, χρησιμοποιώντας διάφορα πρωτόκολλα:
    - Οι πελάτες μπορούν να συνδεθούν χρησιμοποιώντας υποδοχές TCP / IP σε οποιαδήποτε πλατφόρμα.

- Στα λειτουργικά συστήματα Windows, οι πελάτες μπορούν να συνδεθούν χρησιμοποιώντας named pipes εάν ο διακομιστής έχει ξεκινήσει με την επιλογή --enable-named-pipe option. Οι διακομιστές των Windows υποστηρίζουν επίσης συνδέσεις κοινής μνήμης, αν έχει ξεκινήσει ο διακομιστής με την επιλογή --shared-memory. Οι πελάτες μπορούν να συνδεθούν μέσω κοινής μνήμης με τη χρήση της επιλογής --protocol=memory.
- Στα λειτουργικά συστήματα Unix, οι πελάτες μπορούν να συνδεθούν χρησιμοποιώντας Unix domain socket αρχεία.

- APIs για C, C++, Eiffel, Java, Perl, PHP, Python, Ruby, και Tcl είναι διαθέσιμα, επιτρέποντας στα MySQL clients να γραφτούν σε πολλές γλώσσες.
- Η διεπαφή σύνδεσης / ODBC (MyODBC) παρέχει υποστήριξη MySQL για προγράμματα-clients που χρησιμοποιούν ODBC (Open Database Connectivity) συνδέσεις. Για παράδειγμα, μπορείτε να χρησιμοποιήσετε το MS Access για να συνδεθείτε με τον MySQL server σας. Clients μπορούν να τρέχουν σε Windows ή Unix. Όλες οι λειτουργίες ODBC 2.5 υποστηρίζονται, όπως και πολλές άλλες.
- Η διεπαφή σύνδεσης / J παρέχει υποστήριξη MySQL για τα προγράμματα-clients Java που χρησιμοποιούν JDBC συνδέσεις.
- Ο MySQL Connector / Net επιτρέπει στους προγραμματιστές να δημιουργούν εύκολα εφαρμογές .NET που απαιτούν ασφαλή, υψηλής απόδοσης συνδεσιμότητα δεδομένων με την MySQL. Υλοποιεί τις απαιτούμενες διασυνδέσεις ADO.NET και ενσωματώνει σε ADO.NET τα εργαλεία aware. Οι προγραμματιστές μπορούν να δημιουργήσουν εφαρμογές χρησιμοποιώντας την επιλογή τους από τις .NET γλώσσες. Ο MySQL Connector / Net είναι ένας πλήρες διαχειριζόμενος οδηγός του ADO.NET γραμμένος 100% σε C #.

Τοπική προσαρμογή

- Ο διακομιστής μπορεί να παρέχει μηνύματα λάθους στους πελάτες σε πολλές γλώσσες.
- Πλήρης υποστήριξη για πολλά διαφορετικά σύνολα χαρακτήρων, συμπεριλαμβανομένων latin1 (cp1252), γερμανικά, Big5, ujis, αρκετά σετ χαρακτήρων Unicode και πολλά άλλα. Για παράδειγμα, οι σκανδιναβικοί

χαρακτήρες “å”, “ä” και “ö” επιτρέπονται να χρησιμοποιηθούν στα ονόματα των πινάκων και στηλών.

- Όλα τα δεδομένα αποθηκεύονται στο επιλεγμένο σύνολο χαρακτήρων.
- Η ταξινόμηση και οι συγκρίσεις γίνονται σύμφωνα με το επιλεγμένο σύνολο χαρακτήρων και την συγκέντρωση τους (χρησιμοποιεί latin1 και σουηδική συγκέντρωση από προεπιλογή). Είναι δυνατόν να αλλάξει αυτή η ρύθμιση, όταν ξεκάνει ο διακομιστής της MySQL. Για να δείτε ένα παράδειγμα σε πολύ προχωρημένο στάδιο ταξινόμησης, δείτε τον Czech sorting code. Ο MySQL Server υποστηρίζει πολλά διαφορετικά σύνολα χαρακτήρων που μπορούν να καθοριστούν κατά τη μεταγλώττιση και κατά την εκτέλεση.
- Η ζώνη ώρας του server μπορεί να αλλάξει δυναμικά, και μεμονωμένοι πελάτες μπορούν να ορίσουν τη δική τους ζώνη ώρας.

Clients και Tools

- Η MySQL περιλαμβάνει διάφορα προγράμματα που χρησιμεύουν στον πελάτη. Αυτά περιλαμβάνουν τόσο προγράμματα γραμμής εντολών, όπως mysqldump και mysqladmin αλλά και γραφικά προγράμματα όπως το MySQL Workbench.
- Ο MySQL server έχει ενσωματωμένη υποστήριξη για SQL εντολές για: τον έλεγχο, την βελτιστοποίηση και την επισκευή πινάκων. Οι εντολές αυτές είναι διαθέσιμες από τη γραμμή εντολών μέσω του mysqlcheck client. Η MySQL περιλαμβάνει επίσης το myisamchk, ένα πολύ γρήγορο βοηθητικό πρόγραμμα γραμμής εντολών για την εκτέλεση των παραπάνω εντολέων σε πίνακες MyISAM.
- Τα προγράμματα της MySQL μπορούν να χρησιμοποιούν την δυνατότητα --help ή ?, για να αποκτήσουν αμέσως online βοήθεια.

(Oracle and/or its affiliates 1997, 2016,)

3.3.2 Data Mining και Business Intelligence στην MySQL

Η MySQL δεν έχει ενσωματωμένα εργαλεία για data mining και B.I. Για να έχουμε πρόσβαση σε λειτουργιές πάνω στο data mining και στο B.I θα πρέπει να χρησιμοποιήσουμε εξωτερικά προγράμματα, κάθε πρόγραμμα που υποστηρίζει ODBC-JDBC μπορεί να χρησιμοποιηθεί για την σύνδεση του με την MySQL.

Το Java Database Connectivity (JDBC) είναι ένα application programming interface (API) για τη γλώσσα προγραμματισμού Java, το οποίο καθορίζει πως ένας υπολογιστής μπορεί να έχει πρόσβαση σε μια βάση δεδομένων. Αποτελεί μέρος της πλατφόρμας Java Standard Edition, που παρέχεται από την Oracle Corporation. Παρέχει μεθόδους για την αναζήτηση και ενημέρωση δεδομένων σε μια βάση δεδομένων και είναι προσανατολισμένο προς τις σχεσιακές βάσεις δεδομένων. (‘Java Database Connectivity’ 2016)

Το Open Database Connectivity (ODBC) είναι ένα πρότυπο application programming interface (API) για την πρόσβαση σε συστήματα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ). Οι σχεδιαστές του ODBC είχαν ως στόχο να γίνει ανεξάρτητο από τα λειτουργικά συστήματα και τα συστήματα βάσεων δεδομένων. Μια εφαρμογή που μπορεί να χρησιμοποιήσει το ODBC αναφέρεται ως "ODBC συμβατή". Κάθε ODBC συμβατή εφαρμογή μπορεί να έχει πρόσβαση σε κάθε ΣΔΒΔ για το οποίο έχει εγκατασταθεί ο ανάλογος driver. Το ODBC αναπτύχθηκε αρχικά από τη Microsoft κατά τη διάρκεια της δεκαετίας του 1990, και έγινε η βάση για το Call Level Interface (CLI) που έχει τυποποιηθεί από την SQL Access Group στο Unix και στο mainframe. Το ODBC διατήρησε αρκετά χαρακτηριστικά που αφαιρέθηκαν ως μέρος της προσπάθειας του CLI. Το πλήρες ODBC αργότερα γύρισε πίσω σε εκείνες τις πλατφόρμες, και έγινε ένα de facto πρότυπο σημαντικά περισσότερο γνωστό από το CLI. Το CLI παραμένει παρόμοιο με το ODBC, και οι εφαρμογές μπορούν να μεταφερθούν από τη μια πλατφόρμα στην άλλη με λίγες αλλαγές. (‘Open Database Connectivity’ 2016)

Παρακάτω θα αναφέρω μερικά προγράμματα για data mining και business intelligence που μπορούν να συνδεθούν με την MySQL.

Μερικά από τα καλύτερα λογισμικά ανοιχτού κώδικα για την εξόρυξη δεδομένων και για BI τα οποία, διατίθενται δωρεάν είναι:

- Το Orange, ένα λογισμικό μηχανικής μάθησης και εξόρυξης δεδομένων. Διαθέτει φιλικό, ισχυρό, γρήγορο και ευέλικτο user interface χρησιμοποιώντας οπτικό προγραμματισμό για την διερευνητική ανάλυση των δεδομένων και την απεικόνιση τους. Επίσης χρησιμοποιεί Python bindings και βιβλιοθήκες για scripting. Ακόμη περιέχει πλήρες σύνολο στοιχείων: για την προεπεξεργασία των δεδομένων, με δυνατότητα βαθμολόγησης και φιλτραρίσματος, για την μοντελοποίηση, για την αξιολόγηση μοντέλου και για τις τεχνικές εξερεύνησης. Είναι γραμμένο σε C ++ και Python και η γραφική διεπαφή του χρήστη βασίζεται στην cross-platform Qt framework.
- Το RapidMiner που παλαιότερα ονομαζόταν YALE (Yet Another Learning Environment), είναι ένα περιβάλλον για πειράματα μηχανικής μάθησης και εξόρυξης δεδομένων, που χρησιμοποιείται για εργασίες εξόρυξης δεδομένων σε έρευνες αλλά και για έργα με πραγματικά δεδομένα. Καθιστά ικανά πειράματα που αποτελούνται από έναν τεράστιο αριθμό αυθαίρετων nestable μεταβλητών, να αναλύονται σε αρχεία XML και να δημιουργούνται με τη γραφική διεπαφή του χρήστη που διαθέτει το RapidMiner. Το RapidMiner παρέχει περισσότερες από 500 μεταβλητές για όλες τις κύριες διαδικασίες μηχανικής μάθησης και συνδυάζει επίσης την εκμάθηση σχεδίων και την αξιολόγηση των χαρακτηριστικών του προγράμματος Weka. Είναι διαθέσιμο ως αυτόνομο εργαλείο για την ανάλυση των δεδομένων και ως μηχανή εξόρυξης δεδομένων που μπορεί να ενσωματωθεί στα δικά σας προϊόντα.
- Το Weka (Waikato Environment for Knowledge Analysis) είναι ένα λογισμικό γραμμένο σε Java, είναι ένα πολύ γνωστό λογισμικό μηχανικής μάθησης που υποστηρίζει διάφορες τυπικές εργασίες εξόρυξης δεδομένων, συγκεκριμένα: προεπεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, οπισθοδρόμηση, οπτικοποίηση και επιλογή χαρακτηριστικών. Οι τεχνικές που χρησιμοποιεί το λογισμικό βασίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα ενιαίο επίπεδο αρχείο ή σε συγγενές αρχεία, όπου κάθε σημείο δεδομένων είναι χαρακτηρισμένο από ένα σταθερό αριθμό των χαρακτηριστικών των δεδομένων. Το Weka παρέχει πρόσβαση σε SQL βάσεις δεδομένων χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφεται από ένα ερώτημα. Η κυρία διεπαφή του χρηστή

είναι η Explorer αλλά η ίδια η λειτουργικότητα μπορεί να προσεγγιστεί από τη γραμμή εντολών ή μέσω της διεπαφής Flow.
(Auza 2010)

- Το JReport είναι ένα εμπορικό εργαλείο BI από την Jinfonet που υπερέχει στην ενσωμάτωση αναφορών και στην οπτικοποίηση των δεδομένων. Το χρησιμοποιούν μερικές από τις μεγαλύτερες επιχειρήσεις του κόσμου, τυπικά για την παροχή online υπηρεσιών πληροφόρησης στους πελάτες, αλλά χρησιμοποιείται επίσης σε πολλές επιχειρήσεις για την απευθείας παροχή πληροφοριών στους εργαζόμενους και στους συνεργάτες. Επί του παρόντος, περίπου τα δύο τρίτα της επιχείρησης Jinfonet προέρχονται από OEM (Original Equipment Manufacturer), ISVs (Independent Software Vendor) και άλλες τρίτες υπηρεσίες που παρέχουν ενσωματωμένη ευφυΐα σε μια ποικιλία υπηρεσιών και προϊόντων. Το JReport υποστηρίζει dashboards, διαγράμματα, αναφορές και συνδέεται με μια ευρεία ποικιλία των πηγών δεδομένων, συμπεριλαμβανομένων των Big Data, σχεσιακών βάσεων δεδομένων και δεδομένων που βασίζονται στο cloud.
- Το SpagoBI είναι ένα δωρεάν λογισμικό για BI, ουσιαστικά είναι μια πολύ μεγάλη συλλογή λογισμικών ανοιχτού κώδικα που ενώθηκαν μαζί για να δημιουργήσουν ένα ευρύ φάσμα δυνατοτήτων επιχειρηματικής ευφυΐας. Στην πραγματικότητα πηγαίνει πέρα από την παραδοσιακή έννοια της επιχειρηματικής ευφυΐας για να αγκαλιάσει τομείς όπως η εξόρυξη δεδομένων και BPM (Business Process Management). Υποστηρίζει 4 λογισμικά: JasperReport, BIRT, Accessible report, BO. Ακόμη το SpagoBI επιτρέπει να συνειδητοποιήσουμε δομημένες αναφορές, χρησιμοποιώντας δομημένη προβολή πληροφοριών (π.χ. λίστες, πίνακες, crosstabs, αναφορές) και να τις εξάγουμε χρησιμοποιώντας διάφορες μορφές αρχείων (π.χ. HTML, PDF, XLS, XML, TXT, CSV, RTF).
- Το Pentaho Community είναι ένα δωρεάν λογισμικό για BI, τα κυρία χαρακτηριστικά του είναι τα εργαλεία αναφοράς, η πλατφόρμα ενοποίησης δεδομένων, η πλατφόρμα ROLAP analytics και τα εργαλεία εξόρυξης δεδομένων. Με το Pentaho-Report-Designer μπορείτε να δημιουργήσετε αναφορές σε ένα γραφικό περιβάλλον. Οι αναφορές δημοσιεύονται συνήθως στην Pentaho-Platform, η οποία σας επιτρέπει να διαχειριστείτε, να τρέξετε και να προγραμματίσετε τις αναφορές που έχετε δημιουργήσει. Οι εσωτερικές

αναφορές εκτελούνται από την Pentaho Reporting Classic Engine. Η Pentaho Reporting περιλαμβάνει περισσότερες από δυο δωδεκάδες έργα λογισμικών που διευκολύνουν τη δημιουργία και τη δημοσίευση δεδομένων με γνώμονα τις αναφορές των επιχειρήσεων.

(Butler Analytics 2014)

3.4 Microsoft SQL server

Το Microsoft SQL Server είναι ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων, το οποίο αναπτύχθηκε από τη Microsoft. Ως εξυπηρετητής βάσης δεδομένων, είναι ένα λογισμικό με κύρια λειτουργιά του την αποθήκευση και ανάκτηση των δεδομένων, που μπορεί να ζητηθεί από άλλες εφαρμογές που τρέχουν στον ίδιο ηλεκτρονικό υπολογιστή ή από έναν άλλον ηλεκτρονικό υπολογιστή με τη χρήση διαδικτύου. Η Microsoft βγάζει στην αγορά τουλάχιστον μια ντουζίνα από διαφορετικές εκδόσεις του Microsoft SQL Server, που απευθύνονται σε διαφορετικό κοινό και για εργασίες που κυμαίνονται από μικρές εφαρμογές σε μεγάλες διαδικτυακές εφαρμογές με πολλούς ταυτόχρονους χρήστες.

('Microsoft SQL Server' 2016)

Στον παρακάτω πινάκα διατίθενται μερικές χρήσιμες πληροφορίες για τον Microsoft SQL Server

Περιγραφή	Microsofts relational DBMS
Μοντέλο Βάσης Δεδομένων	Relational DBMS
Αρχική έκδοση	1989
Τρέχουσα έκδοση	SQL Server 2014 τον Απρίλιο του 2014
Άδεια	Εμπορική (δωρεάν έκδοση είναι διαθέσιμη με περιορισμένες λειτουργιές)
Γλώσσα υλοποίησης	C++
Λειτουργικά συστήματα	Windows
XML support	Ναι
SQL	Ναι
Σχήμα δεδομένων	Ναι
APIs και άλλοι μέθοδοι πρόσβασης	OLE DB, Tabular Data Stream (TDS), ADO.NET, JDBC, ODBC
Υποστηριζόμενες γλώσσες προγραμματισμού	.Net, Java, PHP, Python, Ruby, Visual Basic
Συγχρονισμός	Ναι
Ανθεκτικότητα	Ναι
Δυνατότητες στην μνήμη	Ναι

Πληκτρολόγηση (Προκαθορίζει τύπους δεδομένων όπως float ή date)	Ναι
Ξένα κλειδιά	Ναι

Πίνακας 3.3 Χρήσιμες πληροφορίες για Microsoft SQL Server.

(db-engines 2016)

Η τρέχουσα ολοκληρωμένη έκδοση του Microsoft SQL Server είναι η SQL Server 2014 από τον Απρίλιο του 2014, πλέον όμως έχουμε και τον SQL Server 2016 Release Candidate 3 (RC3) από το επερχόμενο ολοκληρωμένο λογισμικό Microsoft SQL Server 2016.

3.4.1 Data Mining Tools

Ο Microsoft SQL Server Analysis Services παρέχει τα παρακάτω εργαλεία που μπορείτε να χρησιμοποιήσετε για να δημιουργήσετε λύσεις με την εξόρυξη δεδομένων:

- Ο Data Mining Wizard στον SQL Server Data Tools (SSDT) καθιστά εύκολο να δημιουργήσετε δομές εξόρυξης και μοντέλα εξόρυξης, χρησιμοποιώντας είτε σχεσιακές πηγές δεδομένων είτε πολυδιάστατα δεδομένα σε κύβους. Σε αυτόν τον οδηγό, μπορείτε να επιλέξετε τα δεδομένα που θέλετε να χρησιμοποιήσετε και στη συνέχεια να εφαρμόσετε συγκεκριμένες τεχνικές εξόρυξης δεδομένων, όπως η ομαδοποίηση, τα νευρωνικά δίκτυα, ή τη μοντελοποίηση χρονοσειρών.
- Model viewers παρέχονται στον SQL Server Management Studio αλλά και στον SQL Server Data Tools (SSDT), για την εξερεύνηση των μοντέλων εξόρυξης δεδομένων σας αφού πρώτα τα δημιουργήσετε.
- Το Prediction Query Builder παρέχετε στον SQL Server Management Studio αλλά και στον SQL Server Data Tools για να βοηθήσει στην δημιουργία ερωτημάτων πρόβλεψης. Μπορείτε επίσης να ελέγξετε την ακρίβεια των μοντέλων έναντι ρυθμιζόμενων διδομένων ή σε εξωτερικά δεδομένα ή να χρησιμοποιήσετε διασταυρούμενη επικύρωση για την αξιολόγηση της ποιότητας του συνόλου των δεδομένων σας.
- SQL Server Management Studio είναι το περιβάλλον όπου μπορείτε να διαχειριστείτε τις υπάρχουσες λύσεις της εξόρυξης δεδομένων που έχουν αναπτυχθεί σαν ένα περιστατικό των Analysis Services. Μπορείτε να επανεπεξεργαστείτε δομές και μοντέλα για να ενημερώσετε τα δεδομένα σε αυτά.

- SQL Server Integration Services περιλαμβάνει εργαλεία που μπορείτε να χρησιμοποιήσετε για να καθαρίσετε τα δεδομένα, για την αυτοματοποίηση εργασιών, όπως η δημιουργία προβλέψεων, η ενημέρωση μοντέλων και για την δημιουργία λύσεων στην εξόρυξη κείμενων.

Οι παρακάτω ενότητες παρέχουν περισσότερες πληροφορίες σχετικά με τα εργαλεία εξόρυξης δεδομένων στον SQL Server.

Data Mining Wizard

Χρησιμοποιήστε τον Data Mining Wizard για να ξεκινήσετε τη δημιουργία εξόρυξης δεδομένων. Ο Data Mining Wizard είναι γρήγορος και εύκολος στον χειρισμό του, θα σας καθοδηγεί στη διαδικασία της δημιουργίας μιας δομής εξόρυξης δεδομένων και σε ένα πρώτο σχετικό μοντέλο εξόρυξης δεδομένων. Επίσης περιλαμβάνει εργαλεία για την επιλογή αλγορίθμων και πηγών δεδομένων, καθώς και τον καθορισμό των δεδομένων σε περίπτωση που χρησιμοποιούνται για ανάλυση.

Data Mining Designer

Αφού έχετε δημιουργήσει δομές εξόρυξης δεδομένων και το μοντέλο εξόρυξης δεδομένων χρησιμοποιώντας τον Data Mining Wizard, μπορείτε να χρησιμοποιήσετε τον Data Mining Designer είτε από τον SQL Server Data Tools (SSDT) είτε από τον SQL Server Management Studio για να συνεργαστεί με τα υπάρχοντα μοντέλα και τις υπάρχουσες δομές.

Ο Data Mining Designer περιλαμβάνει εργαλεία για τις παρακάτω εργασίες:

- Τροποποίηση των ιδιοτήτων των δομών εξόρυξης δεδομένων, προσθήκη στηλών και δυνατότητα δημιουργίας προσωρινών ονομάτων στις στήλες (ψευδώνυμα).
- Προσθήκη νέων μοντέλων σε μια υπάρχουσα δομή, αντιγραφή μοντέλων, αλλαγή των ιδιοτήτων των μοντέλων ή των μεταδεδομένων ή ορισμός φίλτρων σε μοντέλα εξόρυξης δεδομένων.
- Περιήγηση στα πρότυπα και τους κανόνες μέσα στο μοντέλο, εξερεύνηση στα δέντρα απόφασης ή στα δέντρα σχέσεων. Λεπτομερή στατιστικά στοιχειά σχετικά με τα Custom viewers που παρέχονται για κάθε διαφορετικό μοντέλο

για την βοήθεια στην ανάλυση των δεδομένων και για την εξερεύνηση των προτύπων που αποκαλύπτονται από την εξόρυξη δεδομένων.

- Επικύρωση μοντέλων με την δημιουργία γραφημάτων lift ή αναλύοντας την καμπύλη κέρδους για τα μοντέλα. Συγκρίνει μοντέλα χρησιμοποιώντας ταξινομημένους πινάκες ή μπορεί να επικυρώσει ένα σύνολο δεδομένων και των μοντέλων του χρησιμοποιώντας διασταυρωμένη επικύρωση.
- Δημιουργία προβλέψεων και ερωτήματα περιεχομένου για την σύγκριση ανάμεσα σε υπάρχουσα μοντέλα εξόρυξης δεδομένων. Κατασκευή ερωτημάτων εφάπαξ ή δημιουργία ερωτημάτων για να παράγουν προβλέψεις για ολόκληρους πινάκες ή για εξωτερικά δεδομένα.

SQL Server Management Studio

Αφού δημιουργήσετε και να αναπτύξετε τα μοντέλα εξόρυξης δεδομένων σε ένα διακομιστή, μπορείτε να χρησιμοποιήσετε τον SQL Server Management Studio για τη διαχείριση της βάσης δεδομένων Analysis Services η οποία, φιλοξενεί τα αντικείμενα της εξόρυξης δεδομένων. Μπορείτε επίσης να συνεχίσετε να εκτελείτε εργασίες που χρησιμοποιούν το μοντέλο, όπως η διερεύνηση των μοντέλων, η επεξεργασία νέων δεδομένων και η δημιουργία προβλέψεων.

Το Management Studio περιλαμβάνει επίσης επεξεργαστές ερωτημάτων που μπορείτε να χρησιμοποιήσετε για να σχεδιάσετε και να εκτελέσετε Data Mining Extensions (DMX) ερωτήματα ή να ασχοληθείτε με αντικείμενα εξόρυξης δεδομένων χρησιμοποιώντας XMLA.

3.4.2 Integration Services Data Mining Tasks and Transformations

Ο SQL Server Integration Services παρέχει πολλά στοιχεία που υποστηρίζουν την εξόρυξη δεδομένων.

Μερικά εργαλεία Integration Services έχουν σχεδιαστεί για να βοηθήσουν την αυτοματοποίηση κοινών εργασιών εξόρυξης δεδομένων, συμπεριλαμβανομένης της πρόβλεψης, την δημιουργία μοντέλων και την επεξεργασία. Για παράδειγμα:

- Δημιουργία ενός πακέτου Integration Services που ενημερώνει αυτόματα το μοντέλο κάθε φορά που το σύνολο δεδομένων ενημερώνεται με νέους πελάτες.
- Εκτέλεση αυτοσχέδιας κατάτμησης ή αυτοσχέδιας δειγματοληψίας των αρχείων καταγραφής.
- Αυτόματη δημιουργία μοντέλων που περνάνε στις παραμέτρους.

Ωστόσο, μπορείτε επίσης να χρησιμοποιήσετε την εξόρυξη δεδομένων σε μια ροή εργασίας, ως είσοδος σε άλλες διαδικασίες. Για παράδειγμα:

- Χρησιμοποιήστε τις τιμές της πιθανότητας που δημιουργήθηκαν από το μοντέλο για την αξιολόγηση σκορ για εξόρυξη δεδομένων σε κείμενα ή άλλες εργασίες ταξινόμησης.
- Αυτόματη δημιουργία προβλέψεων βασισμένων στα προηγούμενα δεδομένα και χρήση των προηγούμενων τιμών αξιολόγησης για την εκτίμηση του κύρος των νέων δεδομένων.
- Χρησιμοποιήστε την λογιστική οπισθοδρόμηση σε τμήματα για εισερχομένους πελάτες από τον κίνδυνο.

(Microsoft 2016)

Business Intelligence Features

Τα χαρακτηριστικά του SQL Server που αποτελούν μέρος της πλατφόρμας Microsoft Business Intelligence είναι:Analysis Services, Integration Services, Master Data Services, Reporting Services και αρκετές εφαρμογές που χρησιμοποιούνται για την δημιουργία ή την εργασία πάνω σε αναλυτικά δεδομένα. Όλα τα χαρακτηριστικά του SQL Server, συμπεριλαμβανομένων των συστατικών BI εγκαθίστανται μέσο της εγκατάστασης του SQL Server. (Microsoft 2015)

Analysis Services

Analysis Services είναι ένα πρόγραμμα που συνδέεται στο διαδίκτυο για την ανάλυση των δεδομένων, χρησιμοποιείτε για την υποστήριξη στην λήψη αποφάσεων και στις αναλύσεις επιχειρηματικών δεδομένων. Με αυτό τον τρόπο παρέχει αναλυτικά στοιχεία για τις αναφορές των επιχειρήσεων και για εφαρμογές όπως Excel, Reporting Services reports και για άλλα τρίτα εργαλεία οπτικοποίησης δεδομένων.

(Microsoft 2016)

Integration Services

Η Microsoft Integration Services είναι μια πλατφόρμα για την δημιουργία ολοκληρωμένων δεδομένων σε επίπεδο επιτηρήσεων και για την δημιουργία λύσεων σε μετασχηματισμένα δεδομένα. Χρησιμοποιήστε Integration Services για την επίλυση σύνθετων επιχειρησιακών προβλημάτων με την αντιγραφή ή τη λήψη αρχείων, την αποστολή μηνυμάτων ηλεκτρονικού ταχυδρομείου για απάντηση στα περιστατικά, ενημέρωση των data warehouses, τον καθαρισμό των δεδομένων και την εξόρυξη δεδομένων, καθώς και τη διαχείριση των αντικειμένων του SQL Server και των δεδομένων. Τα πακέτα μπορούν να εργαστούν μόνα τους ή σε συνεννόηση με άλλα πακέτα για την αντιμετώπιση πολύπλοκων επιχειρηματικών αναγκών. Το Integration Services μπορεί να εξάγει και να μετατρέψει τα δεδομένα από μια μεγάλη ποικιλία πηγών, όπως αρχεία XML δεδομένων, flat αρχεία και από πηγές σχεσιακών δεδομένων και στη συνέχεια να φορτώσει τα δεδομένα σε έναν ή περισσότερους προορισμούς.

(Microsoft 2016)

Master Data Services

Master Data Services (MDS) είναι η λύση του SQL Server για τη διαχείριση master δεδομένων. Το Master data management (MDM) επιτρέπει στην επιχείρησή να ανακαλύψει και να ορίσει μη συναλλακτικούς καταλόγους δεδομένων, καθώς και να συντάσσει διατηρήσιμες και αξιόπιστες master λίστες. (Microsoft 2016)

3.4.3 Reporting Services

Ο SQL Server Reporting Services παρέχει: δημιουργία, ανάπτυξη, κινητή διαχείριση, αριθμημένες αναφορές για τον οργανισμό σας και τέλος ένα εύρη φάσμα από έτοιμα προς χρήση εργαλεία. Οι αναφορές μπορούν να είναι κινητές, αριθμημένες, διαδραστικές, σε μορφή πινάκων, με μορφή γραφικών, με μια ποικιλία δεδομένων, συμπεριλαμβανομένων σχεσιακών, πολυδιάστατων πηγών δεδομένων και σε πηγές δεδομένων που βασίζονται σε XML αρχεία. Οι αναφορές μπορούν να περιλαμβάνουν πλούσιες απεικονίσεις των δεδομένων, συμπεριλαμβανομένων διαγραμμάτων, χαρτών, γραφημάτων Sparkline, και KPIs. (Microsoft 2016)

3.4.4 Business Intelligence Development Studio

Το Business Intelligence Development Studio (BIDS) είναι το Microsoft Visual Studio 2008 με πρόσθετους τύπους εργασιών που είναι προσαρμοσμένοι ειδικά για τον τομέα της επιχειρηματικής ευφυΐας στον Microsoft SQL Server. Το BIDS είναι το κύριο περιβάλλον που θα χρησιμοποιήσετε για να αναπτύξετε επιχειρηματικές λύσεις που περιλαμβάνουν εργασίες σε: Analysis Services, Integration Services, και Reporting Services. Κάθε τύπος εργασιών παρέχει πρότυπα για τη δημιουργία των αντικειμένων που απαιτούνται για τις λύσεις στην επιχειρηματική ευφυΐα και προσφέρει μια ποικιλία από σχεδιαστές, εργαλεία και οδηγούς για να λειτουργήσουν πάνω στα αντικείμενα.

(Microsoft 2016)

Η λειτουργικότητα του BIDS μπορεί να αυξηθεί με το BIDS Helper, ένα Visual Studio πρόσθετο με χαρακτηριστικά που επεκτείνονται ώστε να ενισχυθεί η λειτουργικότητα της ανάπτυξης επιχειρηματικής ευφυΐας στον SQL Server 2005, 2008 και 2008 R2 BI Development Studio και στον SQL Server 2012 SQL Server Data Tools (SSDT). Το BIDS Helper φιλοξενείτε στην ιστοσελίδα CodePlex η όποια, φιλοξένει ανοιχτού κώδικα προγράμματα και είναι ένα έργο της Microsoft.

Η Business Intelligence Markup Language (Biml) μπορεί να χρησιμοποιηθεί στον BIDS για να δημιουργήσει λύσεις από την αρχή μέχρι το τέλος στον τομέα BI, μεταφράζοντας Biml μεταδιδόμενα σε SQL Server Integration Services (SSIS) και SQL Server Analysis Services (SSAS) στοιχειά για την πλατφόρμα Microsoft SQL Server. Το BIDS δεν υποστηρίζεται από το Visual Studio 2010 και μετά και έχει αντικατασταθεί από SQL Server Data Tools-Business Intelligence.

('Business Intelligence Development Studio' 2014)

Το BIDS περιλαμβάνεται στον SQL Server. Θα μπορέσετε να το εγκαταστήσετε αν επιλέξετε την επιλογή client tools κατά την εγκατάσταση του SQL Server. Μετά την εγκατάσταση, μπορείτε να το ξεκινήσετε από τον Microsoft SQL Server program group. (Barley n.d.)

3.5 PostgreSQL

Το PostgreSQL είναι ένα ισχυρό, ανοιχτού κώδικα αντικείμενο-σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων. Έχει περισσότερα από 15 χρόνια ενεργούς ανάπτυξης και μια αποδεδειγμένη αρχιτεκτονική που έχει κερδίσει μια ισχυρή φήμη για την αξιοπιστία, την ακεραιότητα των δεδομένων και την ορθότητα. Τρέχει σε όλα τα κύρια λειτουργικά συστήματα, συμπεριλαμβανομένων των: Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64) και Windows. Επίσης είναι πλήρως συμβατό με το ACID (Atomicity, Consistency, Isolation, Durability), έχει πλήρης υποστήριξη για τα ξένα κλειδιά, για joins, για views, για triggers και για αποθηκευμένες διαδικασίες (σε πολλές γλώσσες). Ακόμη περιλαμβάνει περισσότερο SQL:2008 τύπους δεδομένων συμπεριλαμβανομένων INTEGER, NUMERIC, BOOLEAN, CHAR, VARCHAR, DATE, INTERVAL και TIMESTAMP. Επιπλέον υποστηρίζει την αποθήκευση των μεγάλων δυαδικών αντικειμένων συμπεριλαμβανομένων εικόνων, ήχων ή βίντεο. Τέλος διαθέτει εγγενείς διεπαφές προγραμματισμού για C / C ++, Java, .Net, Perl, Python, Ruby, Tcl και ODBC.

Το PostgreSQL σαν μια βάση δεδομένων της επιχειρηματικής τάξης, διαθέτει εξελιγμένα χαρακτηριστικά όπως: το Multi-Version Concurrency Control (MVCC), την αποκατάσταση σε συγκεκριμένη χρονική στιγμή, tablespaces, ασύγχρονη αντιγραφή, ένθετες συναλλαγές (Savepoints), online/hot backups, ένα εξελιγμένο query planner/optimizer και καταγραφή για την ανοχή σφαλμάτων. Ακόμη υποστηρίζει διεθνή σύνολα χαρακτήρων, κωδικοποιήσεις χαρακτήρων multibyte, Unicode και είναι locale-aware για την ταξινόμηση, για την διάκριση πεζών-κεφαλαίων και την μορφοποίηση. Είναι εξαιρετικά επεκτάσιμο, τόσο στην καθαρή ποσότητα των δεδομένων που μπορεί να διαχειριστεί τόσο και στον αριθμό των ταυτόχρονων χρηστών που μπορεί να φιλοξενήσει.

Υπάρχουν ενεργά συστήματα PostgreSQL σε περιβάλλοντα παραγωγής που διαχειρίζονται πάνω από 4 terabytes δεδομένων. Ορισμένα γενικά όρια για το PostgreSQL περιλαμβάνονται στον παρακάτω πίνακα.

Όριο	Τιμή
Μέγιστο μέγεθος βάσης δεδομένων	Απεριόριστο
Μέγιστο μέγεθος πίνακα	32 TB
Μέγιστο μέγεθος σειράς	1.6 TB

Μέγιστο μέγεθος πεδίου	1 GB
Μέγιστες σειρές ανά πίνακα	Απεριόριστες
Μέγιστες στήλες ανά πίνακα	250 - 1600 ανάλογα με τον τύπο της στήλης
Μέγιστα ευρετήρια ανά πίνακα	Απεριόριστα

Πίνακας 3.4 Όρια τιμών PostgreSQL.

(PostgreSQL 2016)

Ακόμη στον παρακάτω πινάκα διατίθενται επιπλέον χρήσιμες πληροφορίες για το PostgreSQL.

Περιγραφή	Βασισμένο στο αντικείμενο-σχεσιακό ΣΔΒΔ Postgres (αναπτύχθηκε ως αντικειμενοστραφή ΣΔΒΔ, σταδιακά ενισχύθηκε με «πρότυπα» όπως η SQL)
Μοντέλο Βάσης Δεδομένων	Relational DBMS (με αντικειμενοστραφείς επεκτάσεις π.χ: ορίζονται από το χρήστη τύποι/λειτουργίες και η κληρονομικότητα. Χειρισμός των ζευγών κλειδιών/τιμών με την ενότητα hstore.)
Αρχική έκδοση	1989 (1989 Postgres, 1996 PostgreSQL)
Τρέχουσα έκδοση	9.5.2 Μάρτιος του 2016
Άδεια	Ανοιχτού κώδικα (BSD)
Γλώσσα υλοποίησης	C
Λειτουργικά συστήματα	FreeBSD, HP-UX, Linux, NetBSD, OpenBSD, OS X, Solaris, Unix και Windows
SQL	Ναι
Σχήμα δεδομένων	Ναι
APIs και άλλοι μέθοδοι πρόσβασης	native C library, streaming API for large objects, ADO.NET, JDBC και ODBC
Υποστηριζόμενες γλώσσες προγραμματισμού	.Net, C, C++, Java, Perl, Python και Tcl
Συγχρονισμός	Ναι
Ανθεκτικότητα	Ναι
Δυνατότητες στην μνήμη	Όχι
Πληκτρολόγηση (Προκαθορίζει τύπους δεδομένων όπως float ή date)	Ναι
Ξένα κλειδιά	Ναι

Πίνακας 3.5 Χρήσιμες πληροφορίες για PostgreSQL.

(db-engines 2016)

3.5.1 Εργαλεία για Data Mining και Business Intelligence

Όπως και η MySQL έτσι και το PostgreSQL δεν διαθέτει ενσωματωμένα εργαλεία για data mining και BI. Για να εκτελέσουμε εργασίες πάνω στο data mining και BI θα πρέπει να απευθυνθούμε σε τρίτα λογισμικά γι αυτόν τον λόγο υπάρχει η ενότητα Software Catalogue στην επιλογή Download στην επίσημη ιστοσελίδα postgresql.org στην οποία, υπάρχουν 13 λογισμικά τα οποία, μπορεί να βοηθήσουν τους χρήστες σε τέτοιες περιπτώσεις.

Τα λογισμικά αυτά είναι τα εξής: Chartio, check_postgres.pl, DaphnaBI, DbFacePHP, Orange, pgBadger, pgsnap, pg_top, PostgreSQL Dashboard, PostgreStats, PostgreStats Enterprise, Slemma και Ubiq. Προφανώς υπάρχει η «Σημείωση: Το PostgreSQL Global Development Group δεν εγκρίνει ή προτείνει οποιαδήποτε από τα λογισμικά που αναφέρονται και δεν μπορεί να εγγυηθεί για την ποιότητα ή την αξιοπιστία οποιονδήποτε εξ αυτών.» (PostgreSQL 2016) για την αποφυγή τυχών νομικών θεμάτων.

3.6 IBM DB2

Η IBM DB2 είναι ένα σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων με ενσωματωμένη υποστήριξη για μια σειρά από δυνατότητες σε NoSQL επίπεδο, όπως XML, graph store και Java Script Object Notation (JSON). Χρησιμοποιείται από οργανισμούς όλων των μεγεθών. Η DB2 παρέχει μια πλατφόρμα δεδομένων για λειτουργιές συναλλαγών και αναλύσεων. Επίσης προσφέρει τη συνεχή διαθεσιμότητα των δεδομένων για να κρατήσει τις ροές εργασίας των συναλλαγών και των αναλύσεων στην λειτουργική αποτελεσματικότητα τους. (Mullins 2015) Ιστορικά και σε αντίθεση με άλλες εταιρίες που δημιουργούσαν λογισμικά για βάσεις δεδομένων, η IBM παρήγαγε σε συγκεκριμένη πλατφόρμα ένα προϊόν DB2 για κάθε ένα από τα σημαντικά λειτουργικά συστήματα της. Ωστόσο, στη δεκαετία του 1990 η IBM άλλαξε τροχιά και παρήγαγε ένα DB2 "common server" προϊόν, σχεδιασμένο με μια κοινή βάση κώδικα για να τρέξει σε διαφορετικές πλατφόρμες. Σήμερα, υπάρχουν τρία κύρια προϊόντα στην οικογένεια DB2: DB2 για Linux, UNIX και Windows (ανεπίσημα γνωστό ως DB2 LUW), DB2 for z / OS (mainframe), και DB2 for i (πρώην OS / 400). Ένα τέταρτο προϊόν DB2 για VM / VSE είναι επίσης διαθέσιμο. (‘IBM DB2’ 2016)

Για την IBM DB2, η τρέχουσα έκδοση UDB (Universal Database) είναι η 10.5 με τις λειτουργιές του BLU Acceleration και με την κωδική ονομασία του ως «Kepler». Όλες οι εκδόσεις της DB2 μέχρι σήμερα αναφέρονται στον παρακάτω πινάκα:

Έκδοση	Κωδικό Όνομα
3.4	Cobweb
8.1, 8.2	Stinger
9.1	Viper
9.5	Viper 2
9.7	Cobra
9.8	Η Επιτροπή πρόσθεσε χαρακτηριστικά μόνο στο pureScale
10.1	Galileo
10.5	Kepler

Πίνακας 3.6 Εκδόσεις IBM DB2.

(TutorialsPoint 2016)

Επιπλέον στον παρακάτω πινάκα διατίθενται ακόμη μερικές πληροφορίες για την IBM DB2.

Περιγραφή	Κοινή σε περιβάλλοντα υποδοχής IBM, 2 διαφορετικές εκδόσεις για host και για Windows/Linux
Μοντέλο Βάσης Δεδομένων	Relational DBMS (Από την έκδοση 10.5 υποστηρίζει δεδομένα JSON/BSON και είναι συμβατή με την MongoDB)
Αρχική έκδοση	1983 host έκδοση
Τρέχουσα έκδοση	DB2 Data Server (10.5), Απρίλιος του 2013
Άδεια	Εμπορική (Δωρεάν έκδοση είναι διαθέσιμη)
Γλώσσα υλοποίησης	C και C++
Λειτουργικά συστήματα	Linux, Unix, Windows και z/OS
SQL	Ναι
Σχήμα δεδομένων	Ναι
APIs και άλλοι μέθοδοι πρόσβασης	JSON στυλ ερωτημάτων (συμβατό με MongoDB), XQuery, ADO.NET, JDBC και ODBC
Υποστηριζόμενες γλώσσες προγραμματισμού	C, C#, C++, Cobol, Fortran, Java, Perl, PHP, Python, Ruby και Visual Basic
Συγχρονισμός	Ναι
Ανθεκτικότητα	Ναι
Πληκτρολόγηση (Προκαθορίζει τύπους δεδομένων όπως float ή date)	Ναι
Ξένα κλειδιά	Ναι

Πίνακας 3.7 Χρήσιμες πληροφορίες για IBM DB2.

(db-engines 2016)

3.6.1 Data Mining

Η IBM DB2 Warehouse περιλαμβάνει λειτουργιές για την εξόρυξη και την ανάλυση των δεδομένων. Αυτές οι λειτουργιές παρέχουν ταχεία ενεργοποίηση της ανάλυσης της εξόρυξης δεδομένων, του ηλεκτρονικού εμπορίου ή των παραδοσιακών προγραμμάτων Online Transaction Processing (OLTP). Ενώ προηγουμένως αυτές οι λειτουργιές ήταν διαθέσιμες σαν ξεχωριστά προϊόντα πλέον είναι διαθέσιμα ως αναπόσπαστα τμήματα της IBM DB2 Warehouse.

Intelligent Miner

Στην DB2 Warehouse, ο Intelligent Miner αποτελεί ένα software development kit (SDK). Αυτό το SDK αποτελείται από επεκτάσεις της DB2 που περιλαμβάνουν SQL application program interface (API). Μπορείτε να ενσωματώσετε SQL API σε επιχειρηματικές εφαρμογές για την αξιοποίηση των λειτουργιών της εξόρυξης δεδομένων από την επιχειρησιακή εφαρμογή.

Τα SQL API αποτελούνται από τα εξής στρώματα που παρέχουν διαφορετική αναλυτικότητα και διαφορετικό βαθμό αφαίρεσης:

- Το API των διαδικασιών Easy Mining είναι προσανατολισμένο έργο. Μπορείτε να το χρησιμοποιήσετε για να εκτελέσετε συνήθεις εργασίες εξόρυξης δεδομένων.
- Το SQL / MM API συμμορφώνεται με το πρότυπο. Παρέχει ένα API που σας επιτρέπει να συνθέσετε εργασίες εξόρυξης δεδομένων που είναι «ραμμένες στο χέρι σας» για τις ιδιαίτερες απαιτήσεις σας. Μπορείτε να το χρησιμοποιήσετε σε SQL scripts ή από οποιαδήποτε εφαρμογή JDBC, CLI, ODBC, ή SQLJ.

Στην DB2, οι επαγγελματίες μπορούν να εκδίδουν διαδραστικά SQL δηλώσεις από το κέντρο εντολών ή από την γραμμή εντολών. Μπορούν επίσης να ξεκινήσουν τις λειτουργίες της εξόρυξης δεδομένων με μία από αυτές τις διεπαφές:

- Για να βοηθήσει τους προγραμματιστές εφαρμογών, η DB2 Warehouse Design Studio παρέχει plug-ins για την εξόρυξη δεδομένων. Αυτά τα plug-ins περιέχουν γραφικά wizards και editors που έχουν ενσωματωθεί στο περιβάλλον Eclipse. Με αυτά τα plug-ins, μπορείτε να δημιουργήσετε γραφικά μοντέλα εργασιών

στην εξόρυξη δεδομένων και να παράγετε SQL για να ενσωματώσετε την λειτουργικότητα του Intelligent Miner SQL στις επιχειρησιακές σας εφαρμογές.

- Για να βοηθήσει τους διαχειριστές βάσεων δεδομένων, ο Intelligent Miner παρέχει διοικητικές λειτουργίες στη διασύνδεση DB2 Warehouse Administration Console Web. Μπορείτε να ενεργοποιήσετε μια βάση δεδομένων για την εξόρυξη δεδομένων, να διαχειριστείτε τα μοντέλα εξόρυξης δεδομένων ή να διαχειριστείτε το μοντέλο cache του Intelligent Miner.

Predictive Model Markup Language (PMML) είναι μια γλώσσα σήμανσης για την εξόρυξη δεδομένων. Με βάση την XML, η PMML παρέχει ένα πρότυπο που επιτρέπει τα μοντέλα εξόρυξης δεδομένων να ανταλλάσσονται μεταξύ εφαρμογών διαφόρων προμηθευτών. Με την PMML, μπορείτε να ανταλλάξετε μοντέλα μεταξύ διαφορετικών εφαρμογών. Ο Intelligent Miner υποστηρίζει PMML μέχρι την έκδοση 3.2.

Ο Intelligent Miner Visualizer προσφέρει visualizers που βασίζονται σε Java και Flash ώστε να παρουσιάσουν αποτελέσματα μοντελοποίησης δεδομένων (μοντέλα PMML) για ανάλυση.
Οι ακόλουθοι visualizers είναι διαθέσιμοι:

- Associations Visualizer
- Sequences Visualizer
- Classification Visualizer
- Clustering Visualizer
- Regression Visualizer (μόνο για Intelligent Miner μοντελα)
- Time Series Visualizer

Μπορείτε να χρησιμοποιήσετε το Intelligent Miner Visualizer για να απεικονίσει PMML προσαρμοσμένα μοντέλα εξόρυξης δεδομένων. Οι εφαρμογές μπορούν να καλούν αυτούς του visualizers για να παρουσιάσουν τα αποτελέσματα των μοντέλων ή μπορείτε να αναπτύξετε τους visualizers ως applets σε ένα πρόγραμμα περιήγησης Web για μια εύκολη διάδοση τους.

3.6.2 IBM Cognos 10 Business Intelligence Reporting

Το IBM Cognos 10 Business Intelligence Reporting σας δίνει τη δυνατότητα να κάνετε έξυπνες επιχειρηματικές αποφάσεις παρέχοντας ένα ολοκληρωμένο σύνολο των δυνατοτήτων των αναφορών και την πρόσβαση στις πληροφορίες που χρειάζεστε. Δημιουργώντας αναφορές με το IBM Cognos10 Business Intelligence παρέχετε μια ενιαία, web-based, λύση για όλες τις συνιστώσες του κύκλου ζωής αναφορών. Το IBM Cognos 10 Business Intelligence Reporting έχει συνεργατικά χαρακτηριστικά αναφορών για να βοηθήσει τους συγγραφείς των αναφορών:

- Δημιουργία και τροποποίηση αναφορών με ευέλικτη διάταξη zone-based που προσαρμόζεται αυτόματα για να ταιριάζει διαφορετικά στοιχεία και αντικείμενα.

- Δημιουργία και εργασία με όλους τους τύπους αναφορών για να επεκτείνει τη βάση των δυνητικών χρηστών για κάθε αναφορά.

- Δημιουργία και ανάπτυξη μιας ενιαίας αναφοράς που μπορεί να εκτελεστεί σε πολλαπλές γλώσσες και μορφές εξόδου, όπως HTML, Adobe PDF, και Microsoft Excel.

- Δημιουργήστε θέματα αναφορών που περιλαμβάνουν τυποποιημένα αντικείμενα αναφορών, ερωτήματα αλλά και σχεδιαγράμματα.

(IBMn.d.)

4 Μηχανική Εκπαίδευση και Ταξινομητές

4.1 Εισαγωγή στη Μηχανική Εκπαίδευση

Η τεχνητή νοημοσύνη και ειδικότερα η μηχανική εκπαίδευση (machine learning) έχει αποκτήσει μεγάλο εύρος εφαρμογής τα τελευταία χρόνια. Κύριος στόχος είναι η αντιμετώπιση του προβλήματος της υπερ-πληροφόρησης (information overload), μέσω της ανάπτυξης συστημάτων τα οποία θα μπορούν αυτόματα να φιλτράρουν τον ολοένα και αυξανόμενο όγκο δεδομένων, αναζητώντας σχετική πληροφορία για τον τελικό χρήστη. Παρόλο που η διαδικασία εκπαίδευσης στους υπολογιστές απέχει αρκετά από τη διαδικασία εκπαίδευσης στους ανθρώπους, πληθώρα εφαρμογών έχουν επιτυχώς αναπτυχθεί τα τελευταία χρόνια οι οποίες χρησιμοποιούν τη μηχανική εκπαίδευση σε διάφορους τομείς όπως για παράδειγμα η ανακάλυψη γνώσης (knowledge discovery) από μεγάλο όγκο βάσεων δεδομένων κ.α.

Η μηχανική εκπαίδευση μπορεί να διακριθεί στην εκπαίδευση με επίβλεψη (supervised learning) και στη εκπαίδευση χωρίς επίβλεψη (unsupervised learning). Στη εκπαίδευση χωρίς επίβλεψη, δεν υπάρχει προκαθορισμένο σύνολο τιμών. Τα παραδείγματα εκπαίδευσης χωρίζονται σε, άγνωστες εκ των προτέρων, ομάδες με βάση τα χαρακτηριστικά τους. Χαρακτηριστικό παράδειγμα της εκπαίδευσης χωρίς επίβλεψη αποτελεί η εύρεση κανόνων συσχέτισης μεταξύ των τιμών των χαρακτηριστικών στα διανύσματα εκπαίδευσης.

Το μεγαλύτερο τμήμα της ερευνητικής δραστηριότητας στο χώρο της μηχανικής εκπαίδευσης αφορά την εκπαίδευση με επίβλεψη, τυπικό παράδειγμα της οποίας είναι τα προβλήματα ταξινόμησης (classification). Σε ένα πρόβλημα ταξινόμησης, κάθε παράδειγμα εκπαίδευσης αντιστοιχεί σε ένα διάνυσμα (x1, x2, . . . , xn, y), όπου x1, x2, . . . , xn είναι ένα σύνολο τιμών χαρακτηριστικών, ή αλλιώς γνωρισμάτων, και y είναι μια τιμή κλάσης η οποία περιγράφει ένα συγκεκριμένο γεγονός για μια θεματική περιοχή, ή αλλιώς, την έννοια στόχο. Στη συνέχεια της εργασίας θα ασχοληθούμε με προβλήματα επιβλεπόμενης μηχανικής εκπαίδευσης και δημιουργίας συνόλων ταξινομητών, αλλά πρώτα ας δώσουμε τον ορισμό του ταξινομητή.

Ορισμός 4.1. (Ταξινομητής). Το εκπαιδευμένο μοντέλο που προκύπτει από την εφαρμογή ενός αλγορίθμου ταξινόμησης σε ένα σύνολο διανυσμάτων χαρακτηριστικών καλείται και ταξινομητής (classifier).

Στην περίπτωση που ενδιαφερόμαστε για την καλύτερη δυνατή ακρίβεια ταξινόμησης, είναι δύσκολο (αν όχι αδύνατο) να βρεθεί ένας ταξινομητής που να αποδίδει αρκετά καλά σε όλες τις εργασίες. (Dietterich, 1857)

Έτσι γεννήθηκε το κίνητρο για το συνδυασμό ταξινομητών (ensemble of classifiers), το οποίο οφείλεται κυρίως στην παρατήρηση ότι δεν είναι δυνατό να βρεθεί ένας ταξινομητής που να είναι ο καλύτερος σε όλες τις θεματικές περιοχές, κάτι που είναι γνωστό και ως no free lunch theorem (Wolpert, 1992) ή conservation law of generalization performance. (Schaffer, 1994) Από την άλλη μεριά, ο συνδυασμός ταξινομητών μπορεί να οδηγήσει στην επίτευξη καλύτερων αποτελεσμάτων σε μετα-επίπεδο.

Ορισμός 4.2. (Συγκρότημα Ταξινομητών). Ένα συγκρότημα ταξινομητών (ensemble of classifiers) είναι μια συλλογή από ανεξάρτητους ταξινομητές, οι οποίοι συνεργάζονται με σκοπό την αύξηση της γενίκευση σε σχέση με αυτή που μπορεί να επιτυχθεί με ένα και μόνο ταξινομητή.

Από την αρχή της δεκαετίας ΄90, οι αλγόριθμοι οι οποίοι βασίζονται σε παρόμοιες ιδέες έχουν αναπτυχθεί σε πολλές αλλά σχετικές μορφές. Η προσέγγιση των συγκροτήματος ταξινομητών μπορεί να εκτιμηθεί σαν συνοθήλευμα δύο ειδών μεθόδων :

1. μια μέθοδος για τη δημιουργία ανεξάρτητων ταξινομητών.
2. μια μέθοδος για το συνδυασμό των αποτελεσμάτων των ταξινομητών που θα περιγράψουμε στη συνέχεια.

Βασικός σκοπός μας, είναι η διαδοχική δημιουργία μοντέλων ικανών να αλληλοσυμπληρώνονται, υπό την έννοια ότι το κάθε ένα αποδίδει τα μέγιστα σε ένα υποσύνολο του σώματος εκπαίδευσης, το οποίο τα υπόλοιπα δεν μπορούν να

αξιοποιήσουν αποτελεσματικά. (Κωτσιαντής, 2005) (Witten & Frank, 2005) Ένα σημαντικό ζήτημα για τη δημιουργία ταξινομητών είναι ότι πρέπει να είναι ακριβείς όμως και διαφορετικά, δηλαδή, να έχουν ασύνδετα λάθη. Η ακρίβεια νευρωνικών δικτύων σε ένα σύνολο εξαρτάται από τις αρχιτεκτονικές τους, ενώ η ποικιλομορφία τους εξαρτάται από το συσχετισμό των σφαλμάτων των νευρωνικών δικτύων.

Η βασική ιδέα που κρύβεται κάτω από αυτή την προσέγγιση είναι να βρεθούν οι τρόποι της προώθησης αντί της αδιαφορίας των πληροφοριών που περιλαμβάνονται στους ταξινομητές. Όπως είναι αναμενόμενο, η μέθοδος δουλεύει ιδιαίτερα καλά για ασταθείς αλγορίθμους εκπαίδευσης, όπως είναι τα νευρωνικά δίκτυα, οι οποίοι παράγουν ταξινομητές αρκετά διαφορετικούς με μικρές αλλαγές του συνόλου εκπαίδευσης. Στη βλιογραφία υπάρχει μια μεγάλη ποικιλία μοντέλων (αλγορίθμων μηχανικής εκπαίδευσης), οι οποίοι χωρίζονται σε κατηγορίες μερικές από τις οποίες παρουσιάζουμε συνοπτικά στη συνέχεια.

4.2 Μέθοδοι Μηχανικής Εκπαίδευσης

4.2.1 Δέντρα Απόφασης

Μια ευρέως χρησιμοποιούμενη μέθοδος μηχανικής εκπαίδευσης είναι και εκείνη που βασίζεται σε δέντρα απόφασης (decision trees), κατά την οποία επιχειρείται η προσέγγιση μιας κατηγορικής συνάρτησης στόχου, ακολουθώντας την τεχνική του διαίρει και βασίλευε (Divide and Conquer). Ο χώρος του προβλήματος χωρίζεται σε περιοχές από στιγμιότυπα που φέρουν την ίδια τιμή ως προς κάποιο χαρακτηριστικό, και η διαδικασία επαναλαμβάνεται αναδρομικά, αναπαριστώντας με τον τρόπο αυτό το παραγόμενο μοντέλο ως δέντρο απόφασης. Τα δέντρα απόφασης έχουν χρησιμοποιηθεί ευρέως για τρείς λόγους

1. Ο ταξινομητής στον οποίο καταλήγουν τα δέντρα αποφάσεων, όπως και στις μεθόδους εκπαίδευσης κανόνων, είναι η μορφή ταξινομητή πιο κοντά στην ανθρώπινη γλώσσα.
2. Ένας άλλος λόγος για την χρήση των μεθόδων αυτών είναι όταν η συνάρτηση στόχος ξέρουμε ότι είναι διάζευξη συζεύξεων καθώς τα δέντρα αποφάσεων αποτελούν τέτοιες εκφράσεις.
3. Τέλος ένα από τα βασικά πλεονεκτήματα των μεθόδων δέντρων αποφάσεων είναι ότι δεν επηρεάζονται από λάθη στις τιμές των στιγμιότυπων αλλά ούτε και από την έλλειψη των τιμών τους. Αυτό οφείλεται στο γεγονός ότι για την

κατασκευή των δέντρων αποφάσεων ασχολούμαστε με το σύνολο εκπαίδευσης και τα υποσύνολά του αντί να μας απασχολεί ξεχωριστά το κάθε στιγμιότυπο. Αυτό σημαίνει πως ένα λάθος σε μια τιμή ή η απουσία κάποιας τιμής δεν επηρεάζει ουσιαστικά την ανάπτυξη του δέντρου.

Η κατασκευή ενός δέντρου απόφασης από το σύνολο εκπαίδευσης είναι μια επαγωγική διαδικασία. Η περισσότερο χρησιμοποιούμενη κατηγορία επαγωγικής εκμάθησης για την κατασκευή δέντρων απόφασης είναι η επαγωγική κατασκευή δέντρων απόφασης από την κορυφή προς τα κάτω (Top Down Induction of Decision Trees TDIDT).

Κατά την εκπαίδευση, το δέντρο χτίζεται με την επαναλαμβανόμενη διάσπαση του δοσμένου συνόλου δεδομένων σύμφωνα με τις διάφορες ανεξάρτητες μεταβλητές. Η «σειρά» με την οποία χρησιμοποιούνται οι ανεξάρτητες μεταβλητές στη δόμηση του δέντρου εξαρτάται από τη δυνατότητα ταξινόμησης της κάθε ανεξάρτητης μεταβλητής. Υπάρχουν διάφοροι αλγόριθμοι για την επιλογή της σειράς, (Murthy, 1908) αλλά ο στόχος είναι πάντα ο ίδιος, δηλαδή, να επιλέξουμε την μεταβλητή εκείνη που διαχωρίζει καλύτερα τις τελικές κλάσεις. Ο αλγόριθμος σταματά όταν φτάσει σε κόμβο από τον οποίο δεν είναι δυνατό να ξεκινήσει μία νέα διάσπαση. Τότε ο κόμβος αυτός δεν έχει παιδιά και αποτελεί φύλλο του δέντρου. Ως προς τον τρόπο ανάπτυξής τους, τα δέντρα αποφάσεων διακρίνονται σε :

1. Δυαδικά: από κάθε κόμβο διακλαδίζονται δύο νέοι κόμβοι.
2. Σύνθετα: από κάθε κόμβο διακλαδίζονται δύο ή περισσότεροι νέοι κόμβοι.

Κάθε κόμβος και προφανώς και η ρίζα του δέντρου θέτουν ένα ερώτημα: «Το χαρακτηριστικό Α τι τιμή παίρνει ;» Μια εύλογη απορία είναι λοιπόν πιο χαρακτηριστικό θα χρησιμοποιηθεί σε κάθε κόμβο. Ο τρόπος επιλογής του χαρακτηριστικού συχνά είναι η μόνη διαφορά μεταξύ των αλγορίθμων. Το κυριότερο μέτρο της απόδοσης ενός χαρακτηριστικού είναι η εντροπία του. Οι αλγόριθμοι δέντρων αποφάσεων χειρίζονται σύνολα με συγκεκριμένα χαρακτηριστικά. Έτσι, αφού στο σύνολο όλων των στιγμιότυπων που χρησιμοποιούνται για την εκπαίδευση περιέχονται και θετικά και αρνητικά (αν οι κλάσεις στις οποίες χωρίζονται τα στιγμιότυπα είναι δύο), η φυσική σημασία της εντροπίας είναι αυτή ακριβώς η παρουσία και των δύο κλάσεων χωρίς κάποιο προφανή διαχωρισμό. Η τιμή της

εντροπίας εκφράζει το κατά πόσο είναι διαταραγμένο το σύνολο εκπαίδευσης. Η μείωση της εντροπίας εκφράζει την μείωση της διατάραξης, δηλαδή κάποια μορφή τάξης αρχίζει να επικρατεί στο σύνολο. Έτσι σκοπός των μεθόδων των δέντρων αποφάσεων είναι η μείωση της εντροπίας καθώς ακολουθούμε ένα μονοπάτι από τη ρίζα ως τα φύλλα ενός δέντρου. Στη συνέχεια θα δώσουμε τον ορισμό της εντροπίας.

Ορισμός 4.3. (Εντροπία). Έστω το σύνολο S των στιγμιότυπων πάνω στα οποία θα γίνει η εκπαίδευση. Η εντροπία ορίζεται σαν το αντίθετο της πιθανότητας εμφάνισης ενός θετικού στιγμιότυπου $\oplus$ επί τον δυαδικό λογάριθμο της πιθανότητας αυτής μείον την πιθανότητα να εμφανιστεί ένα αρνητικό στιγμιότυπο $\ominus$ επί τον δυαδικό λογάριθμο αυτής της πιθανότητας. Ο ορισμός της εντροπίας περιγράφεται από την έκφραση

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Προφανώς ο ορισμός αυτός αναφέρεται στην περίπτωση δύο κλάσεων κατάταξης των στιγμιότυπων. Εύκολα όμως επεκτείνεται ο ορισμός και για την περίπτωση ενός συνόλου όπου οι κλάσεις δεν είναι μόνο δύο. Έτσι ο ορισμός γίνεται :

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i$$

όπου το c είναι το πλήθος του συνόλου των κλάσεων κατάταξης των στιγμιότυπων.

ID3

Κατά καιρούς έχουν προταθεί διάφορα μέτρα επιλογής του κατάλληλου χαρακτηριστικού για κάποιον κόμβο. Τα περισσότερα από αυτά λαμβάνουν υπ' όψη τους την εντροπία του συνόλου εκπαίδευσης, αλλά αναλόγως και την εφαρμογή για την οποία κατασκευάστηκαν, επηρεάζονται και από άλλα χαρακτηριστικά του προβλήματος. Ένα πολύ διαδεδομένο μέτρο επιλογής είναι αυτό που χρησιμοποιεί και ο αλγόριθμος ID3, το κέρδος (Gain) από τη μείωση της εντροπίας που αποκτάμε εκχωρώντας το συγκεκριμένο χαρακτηριστικό στον κόμβο του δέντρου που αναπτύσσουμε.

Το κέρδος για ένα χαρακτηριστικό A ενός συνόλου S δίνεται από τον τύπο

$$Gain(S,A) = Entropy(S) - \sum_{u \in Values(A)} \frac{|S_u|}{|S|} Entropy(S_u)$$

$$\text{όπου } S_u = \{s \in S | A(s) = u\}$$

΄Οπως αναφέρθηκε νωρίτερα, σκοπός των αλγορίθμων είναι η μείωση της εντροπίας. Γι'αυτό το λόγο θεωρείται ως καλύτερη επιλογή, η επιλογή ενός χαρακτηριστικού το οποίο θα μειώνει όσο το δυνατόν περισσότερο την εντροπία. Αν υποτεθεί ότι ένα χαρακτηριστικό Α χωρίζει το σύνολο εκπαίδευσης σε πάρα πολλά υποσύνολα με πολύ μικρή εντροπία το καθένα. Αν χρησιμοποιούταν το κέρδος σαν μέτρο, το χαρακτηριστικό αυτό θα εκχωρούταν στην ρίζα και θα είχαμε ένα πάρα πολύ πλατύ δέντρο. Ένα τέτοιο δέντρο θα μπορούσε εύκολα να καταχωρήσει και με μεγάλη ακρίβεια οποιοδήποτε στιγμιότυπο του συνόλου εκπαίδευσης. Θα αποτύγχανε όμως σε ένα άλλο σύνολο πέραν του συνόλου εκπαίδευσης λόγω του υπερταιριάσματος (overfitting), καθώς το χαρακτηριστικό αυτό μπορεί να διαδραματίζει σημαντικό ρόλο για την κατάταξη του συνόλου εκπαίδευσης, όμως δεν μπορεί απαραίτητα να γενικευθεί και να καταχωρήσει ένα οποιοδήποτε στιγμιότυπο. Καταλήγουμε στο συμπέρασμα πως θα πρέπει να λάβουμε υπ' όψη μας και το κατά πόσο ένα χαρακτηριστικό χωρίζει σε πολλά υποσύνολα το αρχικό. ΄Ετσι ορίστηκε ένα καινούργιο μέγεθος, ο χωρισμός πληροφορίας που ορίζεται από τη σχέση

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Πρόκειται για ένα μέγεθος που εκφράζει το πόσο πολύ χωρίζεται ένα σύνολο από ένα χαρακτηριστικό. Προφανώς και δεν μπορεί από μόνο του να εκφράσει το συνολικό κέρδος που μας δίνει ένα χαρακτηριστικό σε αντίθεση με κάποιο άλλο. ΄Ετσι έχουμε το λόγο πληροφορίας που δίνεται από τον τύπο (Quinlan, 1986)

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

που συνδυάζει το κέρδος που προαναφέραμε με το χωρισμό της πληροφορίας.

C4.5

Ο πιο γνωστός αλγόριθμος για την κατασκευή δέντρων απόφασης που χρησιμοποιεί το Λόγο του Κέρδους Πληροφορίας είναι ο C4.5. (Quinlan, 1993) Μια από τις πιο πρόσφατες έρευνες που συγκρίνουν τα δέντρα απόφασης με άλλους αλγορίθμους εκπαίδευσης (Lim, Loh & Shih, 2000) έδειξε ότι ο C4.5 έχει έναν πολύ καλό συνδυασμό ακρίβειας και ταχύτητας εκπαίδευσης.

Η παρουσίαση του αλγορίθμου που ως τώρα αναπτύξαμε, προϋποθέτει τη χρήση κατηγορικών χαρακτηριστικών. Ο αλγόριθμος C4.5 ωστόσο και οι διάφορες επεκτάσεις του, έχουν τη δυνατότητα να διαχειριστούν και συνεχή χαρακτηριστικά, εφαρμόζοντας στην αρχή κάθε αναδρομικού βήματος μια διαδικασία μετατροπής τους σε ένα σύνολο διακριτών λογικών (Boolean) χαρακτηριστικών, γνωστή ως διακριτοποίηση (Discretization). Η διαδικασία αυτή ξεκινά με την διάταξη των στιγμιότυπων του υποσυνόλου του σώματος εκπαίδευσης που πρόκειται να χρησιμοποιηθούν στο τρέχον βήμα κατά αύξουσα σειρά, βάσει της τιμής που φέρουν για ένα συνεχές χαρακτηριστικό a_i. Με αυτό τον τρόπο εντοπίζονται όλα τα διαδοχικά στιγμιότυπα x_k, x_l τα οποία ανήκουν σε διαφορετική μεταξύ τους κλάση. Η τιμή του a_i στο δεύτερο κατά σειρά στιγμιότυπο κάθε τέτοιου ζεύγους ανατίθεται σε ένα κατώφλι (threshold), t_{ij}, με βάση το οποίο ορίζεται ένα λογικό χαρακτηριστικό αιθ. Το νέο χαρακτηριστικό παίρνει την τιμή 1 αν η τιμή του αρχικού a_i υπερβαίνει την τιμή του κατωφλίου t_{ij} , διαφορετικά η τιμή του είναι 0. Έτσι το χαρακτηριστικό a_i αντικαθίσταται από ένα σύνολο λογικών χαρακτηριστικών a_{i1}, a_{i2}, . . . , a_{im} με αντίστοιχα κατώφλια. Ο αλγόριθμος δημιουργίας του δέντρου σταματά όταν δεν είναι δυνατή η παραγωγή δύο νέων κλαδιών που το καθένα να περιέχει αριθμό εγγραφών ίσο ή μεγαλύτερο από έναν ελάχιστο αριθμό παρατηρήσεων σε κάθε κλαδί, τον οποίο έχει καθορίσει ο χρήστης στην αρχή της διαδικασίας.

M5

Ο M5 (Wang & Witten, 1997) αλγόριθμος είναι ένα υβρίδιο διασταύρωσης δέντρου απόφασης και γραμμικής παλινδρόμησης. Δουλεύει αρχικά σαν δέντρο απόφασης τμηματοποιώντας το διανυσματικό χώρο που ορίζουν οι ιδιότητες σε φύλλα, σε κάθε ένα από τα οποία εκπαιδεύεται ένα ξεχωριστό μοντέλο γραμμικής παλινδρόμησης, με το υποσύνολο των παραδειγμάτων του συνόλου εκπαίδευσης που αντιστοιχούν σε εκείνο το φύλλο.

4.3 Σύνολα Κανόνων

Η πιο κατανοητή μορφή ταξινομητή για τον παρατηρητή που μπορούν να εξάγουν οι αλγόριθμοι μηχανικής εκπαίδευσης είναι τα σύνολα κανόνων. Ουσιαστικά, οι κανόνες ταξινόμησης (classification rules) είναι δημοφιλείς ως εναλακτικό των δέντρων απόφασης επειδή η εκπαίδευση, η οποία καταλήγει σε λογικούς κανόνες μπορεί να γίνει και με την κατασκευή δέντρων αποφάσεων.

Οι κανόνες ταξινόμησης αντιπροσωπεύουν κάθε κατηγορία χρησιμοποιώντας την διαζευκτική κανονική μορφή (DNF). Ο στόχος είναι να κατασκευαστεί το μικρότερο σύνολο κανόνων που είναι σύμφωνο με τα στοιχεία κατάρτισης (δεδομένα εκπαίδευσης). Ο μεγάλος αριθμός κανόνων ταξινόμησης είναι συνήθως ένα σημάδι ότι ο αλγόριθμος εκπαίδευσης προσπαθεί να θυμηθεί το σύνολο κατάρτισης, αντί να ανακαλύπτει τις υποθέσεις που το κυβερνούν.

Αλγόριθμοι Ακολουθιακής Εκπαίδευσης

Βασική διαφορά των αλγορίθμων ακολουθιακής εκπαίδευσης από τους υπόλοιπους αλγορίθμους μηχανικής εκπαίδευσης είναι πως δεν προσπαθούν να κατασκευάσουν έναν ταξινομητή που να μπορεί να επαληθεύσει ολόκληρο το σύνολο εκπαίδευσης, αλλά ένα όσο πιο μεγάλο μέρος αυτού. Η λογική τους είναι να κατασκευάσουν έναν κανόνα που να μπορεί να επαληθεύει ένα μεγάλο σύνολο στιγμιοτύπων του συνόλου εκπαίδευσης. Αναζητείται λοιπόν, ένας κανόνας που θα έχει μεγάλη ακρίβεια αλλά όχι απαραίτητα μεγάλη κάλυψη. Δηλαδή θα πρέπει να μπορεί να έχει πολύ μεγάλη ακρίβεια πάνω στα στιγμιότυπα που περιγράφονται από χαρακτηριστικά που ο κανόνας αυτός προβλέπει, αλλά δεν είναι απαραίτητο να μπορεί να έχει εφαρμογή σε άλλα στιγμιότυπα. Στη συνέχεια αφαιρεί από το σύνολο εκπαίδευσης τα στιγμιότυπα τα οποία καλύπτει ο κανόνας που μόλις κατασκευάστηκε και επαναλαμβάνει τη διαδικασία ωσότου κατασκευάσει κανόνες που να καλύπτουν και το υπόλοιπο σύνολο εκπαίδευσης.

Ακολουθιακή Εναντίον Ταυτόχρονης Κάλυψης

Ο τρόπος επιλογής των χαρακτηριστικών συχνά είναι μια διαφορά μεταξύ των αλγορίθμων παραγωγής κανόνων ταξινόμησης. Σαφώς και για έναν αλγόριθμο ακολουθιακής εκπαίδευσης το κυριότερο μέτρο της απόδοσης ενός χαρακτηριστικού είναι η εντροπία του, αλλά όχι το μοναδικό, πληθώρα άλλων μέτρων έχουν χρησιμοποιηθεί. (Lavrac, 1999) Οι An και Cercone (An & Cercone, 2001) μετά από ένα σύνολο πειραμάτων απέδειξαν ότι δεν υπάρχει ένα μέτρο που να λειτουργεί καλύτερα

σε όλες τις περιπτώσεις. ΄Οπως αναφέραμε προηγουμένως οι αλγόριθμοι ακολουθιακής εκπαίδευσης εξάγουν άμεσα κανόνες για την ταξινόμηση στιγμιοτύπων, μαθαίνουν έναν έναν τους κανόνες σε αντίθεση με τους αλγόριθμους δέντρων αποφάσεων, οι οποίοι μαθαίνουν πολλούς κανόνες ταυτόχρονα. Γι' αυτό το λόγο οι τελευταίοι καλούνται και αλγόριθμοι ταυτόχρονης κάλυψης.

Το ερώτημα που εγείρεται είναι πιο είδος προτιμάται σε κάθε περίπτωση. Για την επιλογή του καταλληλότερου αλγορίθμου πρέπει να ληφθεί υπ' όψη το πλήθος των δεδομένων του συνόλου εκπαίδευσης. Αν αυτά είναι αρκετά τότε μπορεί κανείς να χρησιμοποιήσει αλγορίθμους ακολουθιακής κάλυψης. Αντιθέτως, αν υπάρχουν λίγα δεδομένα ο «διαμοιρασμός» των αποφάσεων μπορεί να αποδειχθεί πιο αποτελεσματικός. Οι αλγόριθμοι ακολουθιακής κάλυψης κάνουν αρκετά περισσότερους υπολογισμούς από αυτούς της ταυτόχρονης κάλυψης.

Επιπλέον, θα πρέπει να ληφθεί υπ' όψη αν έχει νόημα η επανεξέταση του ίδιου χαρακτηριστικού για διαφορετικούς κανόνες όπως γίνεται στους αλγόριθμους ταυτόχρονης κάλυψης σε αντίθεση με αυτούς της ακολουθιακής. ΄Οπως και στα δέντρα αποφάσεων έτσι και στους κανόνες ταξινόμησης υπάρχουν δύο κοινές προσεγγίσεις που οι αλγόριθμοι εκπαίδευσης χρησιμοποιούν για να αποφύγουν το υπερταίριασμα (Furnkranz, 1997) :

1. Τερματίζοντας τον αλγόριθμο κατάρτισης προτού φθάσει σε ένα σύνολο κανόνων το οποίο ταιριάζει απόλυτα στα δεδομένα εκπαίδευσης,
2. περικόπτοντας περιττούς κανόνες εκ των υστέρων.

Επίσης ο Bramer (Bramer,2002) παρατήρησε ότι η παρουσία θορύβου στο σύνολο εκπαίδευσης κάνει τους αλγόριθμους να παράγουν μεγαλύτερο πλήθος κανόνων ταξινόμησης από ότι χρειάζεται, γι΄ αυτό το λόγο πρότεινε μια μέθοδο περικοπής κανόνων (Jpruning) που βελτιώνει τα σημαντικά αποτελέσματα.

Περισσότερες πληροφορίες για τους κανόνες ταξινόμησης μπορούν να παρουσιάζονται από τον Furnkranz (Furnkranz, 1999), ο οποίος παρέχει μια άριστη ανασκόπηση της υπάρχουσας βιβλιογραφίας για τις μεθόδους παραγωγής κανόνων ταξινόμησης.

Ripper

Ο πιο γνωστός αλγόριθμος παραγωγής κανόνων ταξινόμησης (ακολουθιακής κάλυψης), τον οποίο χρησιμοποιούμε και στα πειράματα της εργασίας μας είναι ο Ripper. Ο

συγκεκριμένος αλγόριθμος διαμορφώνει τους κανόνες μέσω μιας διαδικασίας επαναλαμβανόμενης ανάπτυξης και περικοπής. Κατά τη διάρκεια της αυξανόμενης φάσης οι κανόνες γίνονται πιο περιοριστικοί προκειμένου να ταιριάζουν με τα δεδομένα εκπαίδευσης όσο το δυνατόν ερισσότερο. Κατά τη διάρκεια της φάσης περικοπής, οι κανόνες γίνονται λιγότερο περιοριστικοί για να αποφύγουν το υπερταίριασμα, το οποίο μπορεί να προκαλέσει την κακή απόδοση σε νέες περιπτώσεις (καινούρια στιγμιότυπα).

4.3.1 Εκπαίδευση κατά Bayes

Η συλλογιστική κατά Bayes (Bayesian reasoning) παρέχει μια πιθανοτική προσέγγιση στο πρόβλημα του επαγωγικού συμπερασμού. Στηρίζεται στην υπόθεση πως οι υπό μελέτη ποσότητες ακολουθούν πιθανοτικές κατανομές και πως οι βέλτιστες αποφάσεις μπορούν να παρθούν βάσει αυτών των κατανομών και των παρατηρούμενων δεδομένων.

Στα πλεονεκτήματα της συγκαταλέγεται η δυνατότητα συνδυασμού της προϋπάρχουσας γνώσης με τα παρατηρούμενα δεδομένα, η θεώρηση πιθανοτικών (μη ντετερμινιστικών) μοντέλων και η εκτίμηση της καταλληλότητας για κάθε μοντέλο, επιτρέποντας έτσι την εξέταση και εναλλακτικών μοντέλων πέραν του εκτιμώμενου βέλτιστου.

Στη μηχανική εκπαίδευση, συχνά μας ενδιαφέρει να βρούμε την καλύτερη υπόθεση σε ένα χώρο *H* με βάση τα γνωστά δεδομένα *D*. Ένας τρόπος να καθορίσουμε τι εννοούμε λέγοντας καλύτερη υπόθεση είναι να απαιτήσουμε την πιθανότερη υπόθεση με βάση τα δεδομένα *D* και την τυχόν προηγούμενη γνώση για τις πιθανότητες των υποθέσεων στο χώρο *H*. Το θεώρημα του Bayes, το οποίο είναι ο ακρογωνιαίος λίθος της ομώνυμης συλλογιστικής, παρέχει ένα άμεσο τρόπο υπολογισμού της πιθανότητας για μια υπόθεση *h*.

Θεώρημα 4.1. (Θεώρημα Bayes).

$$P(h\backslash D) = \frac{P(D\backslash h) \cdot P(h)}{P(D)}$$

- *P(h|D)* είναι η πιθανότητα να ισχύει η υπόθεση *h* με βάση τα παρατηρηθέντα δεδομένα *D* και καλείται εκ των υστέρων πιθανότητα (posterior probability) της *h*, γιατί εκφράζει την εμπιστοσύνη στην *h* αφού έχουμε δει τα δεδομένα *D*.

- *P(D|H)* είναι η πιθανότητα να παρατηρηθούν τα δεδομένα *D* σε κάποιο κόσμο που η υπόθεση *h* ισχύει και λέγεται πιθανοφάνεια (likelihood) των δεδομένων *D* δοθείσας της *h*.
- *P(h)* είναι η πιθανότητα να ισχύει η υπόθεση *h* πριν την παρατήρηση των δεδομένων και λέγεται εκ των προτέρων πιθανότητα (prior probability) της *h*. Εκφράζει την προηγούμενη γνώση που τυχόν έχουμε για την ισχύ της *h*.
- *P(D)* είναι η πιθανότητα να παρατηρηθούν τα δεδομένα *D* ανεξαρτήτως της υπόθεσης που ισχύει και λέγεται εκ των προτέρων πιθανότητα των δεδομένων *D*.

Σε πολλές περιπτώσεις, ο αλγόριθμος εκπαίδευσης θεωρεί ένα σύνολο υποψήφιων υποθέσεων *H* και αναζητεί την πιο πιθανή από αυτές δοθέντων των δεδομένων εκπαίδευσης. Μια τέτοια υπόθεση *h* ονομάζεται *υπόθεση μέγιστη εκ των υστέρων* (maximumum posteriori MAP). Ένας ευθύς τρόπος εύρεσης των υποθέσεων MAP είναι η εφαρμογή του θεωρήματος του Bayes για κάθε υπόθεση στο *H* και η επιλογή των μέγιστων από αυτές, δηλαδή:

$$h_{MAP} = \arg\max_{h \in H} P(h|D) = argnax \frac{P(D|h) \cdot P(h)}{P(D)} = argmaxP(D|h) \cdot P(h), \quad (4.1)$$

Στο τελευταίο βήμα, το *P(D)* παραλήφθηκε γιατί είναι σταθερά ως προς τις υποθέσεις. Μερικές φορές δεν έχουμε καμιά εκ των προτέρων γνώση για τις υποθέσεις η και δεν έχουμε λόγο να πιστεύουμε πως είναι ανισοπίθανες. Τότε μπορούμε να θεωρήσουμε πως και ο όρος *P(h)* είναι σταθερός για όλες τις υποθέσεις και να τον απαλείψουμε και αυτόν από τη σχέση (4.1). Έτσι, η υπόθεση MAP θα είναι αυτή που μεγιστοποιεί την πιθανοφάνεια *P(D|h)* και η οποία ονομάζεται υπόθεση μέγιστης πιθανοφάνειας (maximum likelihood ML):

$$h_{ML} = argmaxP(D|h), \qquad h \in H$$

4.3.2 Βέλτιστος Ταξινομητής Bayes

Στην πράξη, μας ενδιαφέρει συνήθως περισσότερο το ποια είναι η πιο πιθανή τιμή της συνάρτησης-στόχου ενός νέου στιγμιοτύπου δοθέντων των δεδομένων από το ποια είναι η πιο πιθανή υπόθεση δοθέντων των δεδομένων. Αν και μια απλή προσέγγιση είναι να θεωρήσουμε την τιμή της υπόθεσης MAP ως πιθανότερη τιμή, υπάρχει και

καλύτερη λύση. Αυτή προκύπτει αν λάβουμε υπόψη τις προβλέψεις όλων των υποθέσεων, υγισμένες κατά την εκ των υστέρων πιθανότητά τους. Έτσι, αν η συνάρτηση-στόχος παίρνει τιμές σε ένα πεπερασμένο σύνολο V , τότε η πιθανότητα $P(v_j|x,D)$ πως η σωστή τιμή για το στιγμιότυπο x είναι η v_j δίνεται από τη σχέση:

$$P(v_j|x,D) = \sum_{h \in H} P_h(v_j|x) \cdot P(h|D) \quad (4.2)$$

όπου $P_h(v_j|x)$ είναι η πιθανότητα να έχει το στιγμιότυπο ξ την τιμή v_j σύμφωνα με την υπόθεση h. Η σχέση (4.2), όπως φαίνεται, μπορεί να εφαρμοστεί και για μη ντετερμινιστικές υποθέσεις, δηλαδή υποθέσεις h που για ένα δεδομένο στιγμιότυπο x δεν ισχύουν απαραίτητα

$$P_h(v_j|x) = \begin{cases} 1 & v_j = v_x \\ 0 & διαφορετικά \end{cases} \quad για\ κάποιο\ v_x \in V \quad (4.3)$$

Η βέλτιστη απόφαση είναι η τιμή v_j για την οποία ο τύπος (4.2) μεγιστοποιείται :

$$v_{opt}(x,D) = argmax \sum_{h \in H} P_h(v_j|x) \cdot P(h|D), \quad v_j \in V \quad (4.4)$$

΄Ενα σύστημα που ταξινομεί τα στιγμιότυπα χρησιμοποιώντας την σχέση (4.4) καλείται βέλτιστος ταξινομητής Bayes (Bayes optimal classifier). Καμιά άλλη μέθοδος που θεωρεί τον ίδιο χώρο υποθέσεων, την ίδια a priori γνώση και τα ίδια δεδομένα δεν μπορεί να έχει καλύτερα αποτελέσματα κατά μέσο όρο. (Mitchell, 1997)

4.3.3 Αφελής Ταξινομητής Bayes

Δύο πρακτικά προβλήματα εμφανίζονται στη χρήση του βέλτιστου ταξινομητή Bayes. Το ένα είναι πως έχει γραμμική πολυπλοκότητα ως προς τον πληθάριθμο $|H|$ του χώρου των υποθέσεων, γεγονός που καθιστά την εφαρμογή του αδύνατη για απειροδιάστατους χώρους και μη αποδοτική για μεγάλους πεπερασμένους χώρους. Το άλλο είναι πως απαιτεί τη γνώση ή την εκτίμηση πάρα πολλών πιθανοτήτων: την πιθανοφάνεια $P(D|h)$ των δεδομένων D και την εκ των προτέρων πιθανότητα $P(h)$ για κάθε υπόθεση h. Μία μέθοδος κατά Bayes που αντιμετωπίζει σε μεγάλο βαθμό αυτές τις δυσκολίες είναι ο αφελής ταξινομητής Bayes (Lewis, 1998) (naive Bayes classifier NB για συντομία).

Ο NB εφαρμόζεται σε προβλήματα εκπαίδευσης όπου τα στιγμιότυπα αναπαρίστανται μέσω του μοντέλου του διανυσματικού χώρου, τα χαρακτηριστικά

παίρνουν διακριτές τιμές (αν κάποια είναι συνεχή, πρέπει να κβαντιστούν) και η συνάρτηση-στόχος παίρνει τιμές (ετικέτες - labels) σε ένα πεπερασμένο σύνολο V . Παρέχεται ένα σύνολο από διανύσματα εκπαίδευσης, βάσει του οποίου ο ταξινομητής πρέπει να προβλέψει την ετικέτα ενός νέου στιγμιοτύπου αναπαριστώμενου από το διάνυσμα $\langle \alpha_1, \alpha_{2,} \dots \dots \dots, \alpha_n \rangle$.

Η προσέγγιση κατά Bayes στην κατάταξη του νέου στιγμιοτύπου είναι η ανάθεση σε αυτό της πιο πιθανής τιμής v_{opt}, δεδομένων των τιμών των χαρακτηριστικών του, a_1, a_2, . . . a_n:

$$v_{opt} = argmaxP(V_j|a_1, a_2, \dots., a_n), u_j \in V$$

η οποία μέσω του θεωρήματος του Bayes εκφράζεται ως :

$$v_{opt} = argmax \frac{P(a_1, a_2, \dots \dots., a_n|v_j) \cdot P(v_j)}{P(a_1, a_2, \dots \dots., a_n)} = argmaxP(a_1, a_2, \dots..., a_n|v_j) \cdot P(v_j),$$

$$v_j \in V \quad (4.5)$$

Η εκτίμηση των πιθανοτήτων που εμφανίζονται στην εξίσωση (4.5) πρέπει να γίνει μέσω των δεδομένων εκπαίδευσης. Οι $P(v_j)$ μπορούν να εκτιμηθούν εύκολα ως η συχνότητα εμφάνισης κάθε ετικέτας v_j στα δεδομένα. Το ίδιο όμως δε μπορεί να γίνει για τις $P(a_1, a_2, . . . , a_n|v_j)$, δηλαδή τις πιθανότητες εμφάνισης κάθε δυνατού στιγμιοτύπου δεδομένης μιας ετικέτας, αφού για συνηθισμένα μεγέθη συνόλων εκπαίδευσης, τα περισσότερα στιγμιότυπα δε θα έχουν εμφανιστεί, και επομένως η συχνότητα εμφάνισής τους θα είναι μηδέν, που προφανώς δεν είναι αξιόπιστη εκτίμηση της πραγματικής πιθανότητας εμφάνισης τους.

Ο αφελής ταξινομητής Bayes βασίζεται στην απλουστευτική υπόθεση πως οι τιμές των χαρακτηριστικών είναι ανεξάρτητες δοθείσας της ετικέτας. Τότε, η πιθανότητα της κοινής εμφάνισης των a_1, a_2, . . . , a_n δεδομένης μιας ετικέτας είναι το γινόμενο των πιθανοτήτων εμφάνισης για καθένα από αυτά:

$$P(a_1, a_2, \dots \dots., a_n|v_j) = \prod_{i=1}^{n} P(a_i|v_j).$$

Αντικαθιστώντας αυτή την έκφραση στην εξίσωση (4.5) έχουμε την έκφραση του NB:

$$v_{NB} = argmax P(v_j) \cdot \prod_{i=1}^{n} P(a_i | v_j), \quad v_j \in V \quad (4.6)$$

Από την εξίσωση (4.6) φαίνεται πως το πλήθος των πιθανοτήτων $P(a_i|v_j)$ που πρέπει να εκτιμηθούν επιπλέον των $P(v_j)$ ισούται με το πλήθος των διαφορετικών τιμών των features επί το πλήθος των ετικετών, σημαντικά μικρότερο από αυτό που θα απαιτούνταν για όλες τις $P(a_1, a_2, \ldots, a_n|v_j)$, ακόμα κι αν οι εκτιμήσεις τους ήταν αξιόπιστες. Έτσι, ο NB στη φάση εκπαίδευσής του εκτιμά με βάση τα δεδομένα τις $P(v_j)$ και $P(a_i|v_j)$, το σύνολο των οποίων αποτελούν το μοντέλο ταξινόμησης που μαθαίνει, και στη φάση εξέτασης χρησιμοποιεί την εξίσωση (4.6) για να κατατάξει κάθε νέο στιγμιότυπο. Ένα ενδιαφέρον χαρακτηριστικό του είναι πως δεν ερευνά το χώρο υποθέσεων για την εντοπισμό της καλύτερης υπόθεσης, όπως κάνουν πολλοί αλγόριθμοι εκπαίδευσης, αλλά σχηματίζει άμεσα ένα μοντέλο, απλά μετρώντας τη συχνότητα των συνδυασμών των τιμών των features και των ετικετών μέσα στο σύνολο εκπαίδευσης.

Ο αφελής ταξινομητής Bayes, παρά την αρκετά δεσμευτική υπόθεση της υπό συνθήκη ανεξαρτησίας των χαρακτηριστικών, έχει να επιδείξει αναπάντεχα μεγάλη ακρίβεια και σε εφαρμογές που η υπόθεση της ανεξαρτησίας εμφανώς παραβιάζεται. Ένα ακόμα πλεονέκτημα του NB είναι η σχετική απλότητα των μοντέλων που κατασκευάζει, τα οποία μπορούν να γίνουν εύκολα κατανοητά από τον άνθρωπο, ιδιαίτερα μέσω οπτικοποίησης. (Becker, Kohavi & Sommerfield, 1997)

4.4 Μηχανές Διανυσμάτων Υποστήριξης

Θα ολοκληρώσουμε την παρουσίαση των αλγορίθμων εκπαίδευσης που χρησιμοποιήθηκαν στην παρούσα εργασία με την παρουσιάση των μηχανών διανυσμάτων υποστήριξης (support vector machines SVM), ένα είδος συγκερασμού γραμμικών μοντέλων και εκπαίδευσης βασισμένη σε στιγμιότυπα. Στόχος αυτής της κλάσης αλγορίθμων είναι η επιλογή ενός μικρού αριθμού στιγμιοτύπων εκπαίδευσης από κάθε κλάση, των διανυσμάτων υποστήριξης (support vectors), που συνορεύουν στο χώρο του προβλήματος με στιγμιότυπα άλλων κλάσεων. Τα επιλεγμένα στιγμιότυπα χρησιμοποιούνται για την κατασκευή μιας γραμμικής συνάρτησης διάκτισης (discriminant function), ικανής να τα διαχωρίσει όσο το δυνατόν περισσότερο.

Τα συστήματα ταξινόμησης που βασίζονται σε αυτόν τον αλγόριθμο αποτελούν σήμερα μία από τις δημοφιλέστερες προσεγγίσεις στο χώρο της κατηγοριοποίησης κειμένου,

λόγω της ευρωστίας, της αποτελεσματικότητας και της ταχύτητας που επιδεικνύουν, αλλά και της ικανότητάς τους να παράγουν μη γραμμικές επιφάνειες απόφασης, καθιστώντας έτσι υπολογιστικά εφικτή την επίλυση ενός μεγάλου αριθμού πρακτικών προβλημάτων εκπαίδευσης, τα οποία δεν μπορούν να αντιμετωπιστούν από γραμμικά μοντέλα. Στη συνέχεια θα αναπτύξουμε τα βασικά σημεία της θεωρίας των μηχανών διανυσμάτων υποστήριξης.

Υιοθετώντας το μοντέλο της διανυσματικής αναπαράστασης του χώρου ενός προβλήματος, θεωρούμε ένα σύνολο n διανυσμάτων εκπαίδευσης, διάστασης I + 1, έστω $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots\ldots, \overrightarrow{x_n}\}$, όπου $\overrightarrow{x_i} = \{a_0, a_1, \ldots\ldots, a_{I-1}, y_i\}$ με τα $a_0, a_1, \ldots, a_{I-1}$ να αποτελούν τα I χαρακτηριστικά του χώρου S και με $y_i \in \{-1, 1\}$ την κλάση στην οποία το i-στιγμιότυπο ανοίκει. Λόγω της γραμμικής διαχωρισιμότητας των κλάσεων που υποθέσαμε, μπορούμε να βρούμε ένα υπερεπίπεδο Π, το οποίο να διαχωρίζει κατά βέλτιστο τρόπο τα διανύσματα εκπαίδευσης, με την εξίσωση $\langle \vec{w}, \vec{x} \rangle + b = 0$, όπου $\vec{w}$ το κανονικό διάνυσμα του υπερεπίπεδου Π, $\langle .,. \rangle$ το εσωτερικό γινόμενο δύο διανυσμάτων και $|b|/\|\vec{w}\|$ η κατακόρυφη απόσταση της αρχής του συστήματος συντεταγμένων από το υπερεπίπεδο Π. Βάσει των παραπάνω, για το τυχαίο διάνυσμα εκπαίδευσης $\overrightarrow{x_i}$ θα ισχύουν οι ακόλουθες σχέσεις :

$$\langle \vec{w}, \overrightarrow{x_i} \rangle + b \geq +1, \alpha\nu\ y_i = 1 \quad (4.7)$$

$$\langle \vec{w}, \overrightarrow{x_i} \rangle + b \leq -1, \alpha\nu\ y_i = -1 \quad (4.8)$$

οι οποίες εκφράζονται ισοδύναμα ως εξής :

$$y_i(\langle \vec{w}, \vec{x} \rangle + b) - 1 \geq 0$$

Ας εστιάσουμε πλέον την προσοχή μας στα διανύσματα εκπαίδευσης που ικανοποιούν την ισότητα της σχέση (4.7). Αν τα θεωρήσουμε σαν σημεία του I-διάστατου χώρου S, τότε αυτά τα σημεία θα βρίσκονται στο υπερεπίπεδο Π_1: $\langle \vec{w}, \vec{x} \rangle + b = 1$, με $\vec{w}$ το κανονικό διάνυσμα του υπερεπίπεδου Π_1 και $|1\text{-}b|/\|\vec{w}\|$ την κατακόρυφη απόσταση της αρχής του συστήματος συντεταγμένων από το υπερεπίπεδο Π_1. Ομοίως, τα σημεία του S που ικανοποιούν την ισότητα της (4.8) θα βρίσκονται στο επίπεδο Π_2: $\langle \vec{w}, \vec{x} \rangle + b = -1$, με $\vec{w}$ το κανονικό του διάνυσμα και με $|\text{-}1\text{-}b|/\|\vec{w}\|$ την κατακόρυφη απόσταση της αρχής του συστήματος συντεταγμένων από το Π_2. Τα προαναφερθέντα διανύσματα εκπαίδευσης καλούνται διανύσματα υποστήριξης (support vectors), ενώ η απόσταση μεταξύ των δύο υπερεπιπέδων είναι ίση με $2/\|\vec{w}\|^2$ και μεγιστοποιείται όταν το $\|\vec{w}\|^2$ ελαχιστοποιηθεί. Θα πρέπει τέλος να σημειώσουμε ότι στην περίπτωση που εξετάζουμε, την περιοχή που ορίζεται από τα υπερεπίπεδα Π_1 και Π_2, και που ονομάζεται περιθώριο (margin), δεν αντιστοιχίζεται κανένα διάνυσμα εκπαίδευσης.

Αποδεικνύεται ότι η ελαχιστοποίηση της νόρμας του διανύσματος $\vec{w}$ μπορεί να επιτευχθεί μέσω της συνάρτησης :

$$F(\vec{a}) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} a_i \cdot a_j \langle \overrightarrow{x_i}, \overrightarrow{x_j} \rangle y_i y_j$$

όπου $\overrightarrow{x_i}$ είναι ένα διάνυσμα εκπαίδευσης και $\vec{a}$ ένα διάνυσμα πολλαπλασιαστών Lagrange με όλα του τα στοιχεία θετικά. Ένα διάνυσμα $\overrightarrow{x_j}$ ονομάζεται διάνυσμα υποστήριξης όταν η αντίστοιχη παράμετρος α_j είναι αυστηρά μεγαλύτερη του μηδενός. Έχοντας υπολογίσει τα διανύσματα υποστήριξης κατά την εκπαίδευση του αλγορίθμου, έστω r ο αριθμός τους, η ταξινόμηση του άγνωστου στιγμιότυπου $\vec{x}$ συνίσταται στον υπολογισμό της τιμής της συνάρτησης

$$f(\vec{x}) = sign(\langle \vec{w}, \vec{x} \rangle + b), \text{όπου } \vec{w} = \sum_{i=1}^{r} a_i y_i x_i$$

Εάν η υπόθεση της γραμμικής διαχωρισιμότητας των κλάσεων ισχύει, αναγκαζόμαστε να χαλαρώσουμε τους περιορισμούς των σχέσεων (4.7) και (4.8), επιτρέποντας σε κάποια διανύσματα εκπαίδευσης να βρίσκονται μεταξύ των υπερπεπιπέδων Π_1 και Π_2. Στην προκειμένη περίπτωση, η ποσότητα που θα πρέπει να ελαχιστοποιηθεί είναι

$$\|\vec{w}\|^2 + c\sum_{i=1}^{n} \xi_i$$

με τον περιορισμό ότι $y_i(\langle \vec{w}, \vec{x} \rangle + b) \geq 1 - \xi_i, με\ \xi_i \geq 0$.

Η παράμετρος ξ_i επιτρέπει στο αντίστοιχο διάνυσμα εκπαίδευσης να βρεθεί στην περιοχή του περιθωρίου εφόσον είναι μεγαλύτερη του μηδενός, ενώ η παράμετρος c, η οποία πρέπει να προσδιοριστεί από το χρήστη, εκφράζει την αυστηρότητα που αναμένεται να επιδείξει ο αλγόριθμος στην ανοχή στιγμιοτύπων στο περιθώριο, κατά την εύρεση του βέλτιστου υπερεπιπέδου. Όπως και στην περίπτωση των γραμμικά διαχωρίσιμων κλάσεων, μπορούμε να μεγιστοποιήσουμε τη συνάρτηση $F(\vec{a})$, υπό τον περιορισμό ότι $0 \leq a_i \leq c$ αντί του περιορισμού $a_i \geq 0$, ένα πρόβλημα, το οποίο επιλύεται μέσω των γενικευμένων τετραγωνικών προγραμματιστικών τεχνικών ή ακόμα και εξειδικευμένων στην περιοχή των SVMs.

Καθοριστική σημασία για την ικανότητα γενίκευσης του αλγορίθμου φέρει η επιλογή της παραμέτρου c, καθώς όσο μεγαλύτερη είναι η τιμή της, τόσο πιο αυστηρό είναι το επαγόμενο μοντέλο στον προσδιορισμό ενός υπερεπιπέδου ικανού να διαχωρίσει σωστά μεγάλες τιμές της παραμέτρου c επομένως καθιστούν πιθανή την εμφάνιση, σε σχετικά μικρό βαθμό του φαινομένου του υπερταιριάσματος, ιδιαίτερα όταν η διάσταση του χώρου είναι μεγάλη και τα διανύσματα εκπαίδευσης απομακρύνονται μεταξύ τους. Το γεγονός αυτό φαίνεται πως έρχεται σε αντίθεση τόσο με τη γραμμικότητα του μοντέλου, όσο και με τη φύση του αλγορίθμου, καθώς το φαινόμενο του υπερταιριάσματος μπορεί να παρατηρηθεί μόνο αν προστεθούν ή αφαιρεθούν στο

μοντέλο περιθωρίου. Στο σημείο αυτό θα πρέπει να τονίσουμε ότι το προαναφερθέν ενδεχόμενο θεωρείται σχετικά σπάνιο να παρατηρηθεί σε έναν ταξινομητή SVM, αφού τα διανύσματα υποστήριξης αποτελούν ένα πολύ μικρό ποσοστό των διανυσμάτων εκπαίδευσης, όχι όμως και εντελώς απίθανο.

΄Ενα κύριο πλεονέκτημα των SVMs είναι η ικανότητά τους να χειρίζονται πολύ μεγάλους χώρους χαρακτηριστικών, καθιστώντας το στάδιο της επιλογής χαρακτηστικών, που συνύθως προηγείται αυτού της εκπαίδευσης, περιττό. Επίσης είναι αξιοσημείωτη και η ανεκτικότητα που παρουσιάζουν όσον αφορά στο πλήθος των στιγμιοτύπων εκπαίδευσης, ιδιαίτερα όταν αυτό διαφέρει μεταξύ των δύο κλάσεων, καθώς τα SVMs δεν επιδιώκουν να ελαχιστοποιήσουν το σφάλμα των δεδομένων εκπαίδευσης, αλλά να τα διαχωρίσουν αποτελεσματικά σε ένα χώρο μεγάλης διάστασης. ΄Οσον αφορά τέλος στους χρόνους εκπαίδευσης και ελέγχου του αλγορίθμου, αυτοί αποδεικνύονται κάπως αυξημένοι, ιδιαίτερα όταν η διάσταση του χώρου είναι μεγάλη ή όταν η συνάρτηση διάκρισης δεν είναι γραμμική.

5 Εξόρυξη γνώσης χρησιμοποιώντας το εργαλείο Weka

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα σας παρουσιάσω τα αποτελέσματα από κάποια πειράματα που εκτέλεσα, χρησιμοποιώντας το εργαλείο Weka. Το εργαλείο Weka είναι μια συλλογή από αλγόριθμους μηχανικής μάθησης για την επίλυση πραγματικών προβλημάτων εξόρυξης δεδομένων. Περισσότερη και αναλυτικότερη ανάλυση περί του εργαλείου Weka βρίσκεται στο Παράρτημα Α. Επίσης θα γίνει σύγκριση των αποτελεσμάτων στους αλγορίθμους που χρησιμοποιήθηκαν ως προς τις οικογένειες ταξινομητών, στις οποίες ανήκουν.

5.2 Πειράματα

Στα πειράματα που εκτελεστήκαν, χρησιμοποιήθηκαν κάποια dataset με γενικά προβλήματα. Αυτά τα προβλήματα έτρεξαν στους εξής αλγορίθμους: BayesNet, Naive Bayes, Multi-layer Perceptron (MLP), Simple Logistic, Sequential Minimal Optimization (SMO), 1-Nearest Neighbors, 3-Nearest Neighbors, 10-Nearest Neighbors, Decision Table, JRip, PART, J48, Logistic Model Tree (LMT) και Random Forest. Γενικές πληροφορίες για το κάθε πρόβλημα παρατίθενται στον παρακάτω πινάκα.

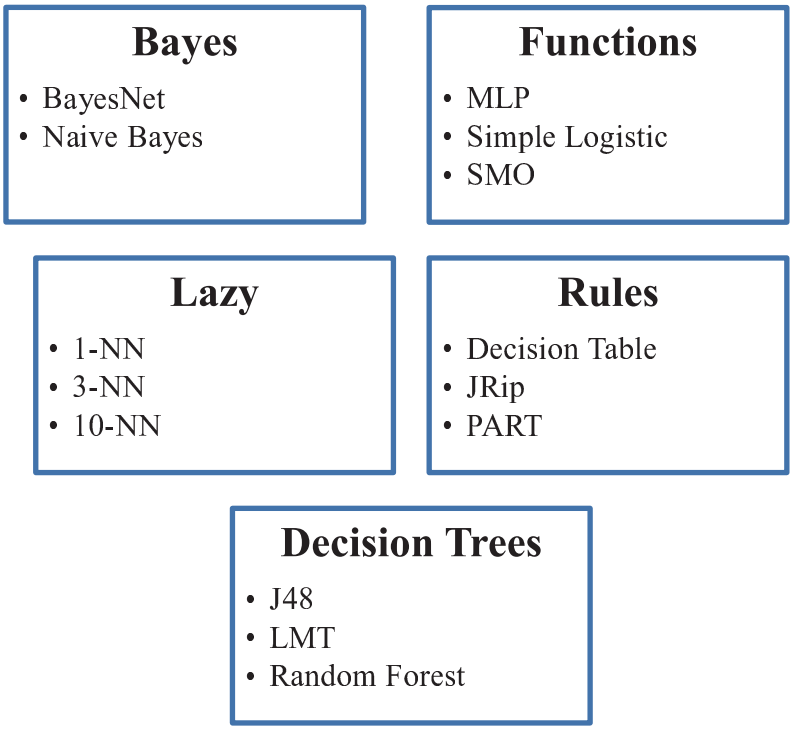
Dataset	Instances	Features	Classes
automobile	205	26	7
banana	5300	2	2
breast	286	9	2
bupa	345	6	2
cleveland	297	13	5
coil2000	9822	85	2
contraceptive	1473	9	3
crx	125	15	2
dermatology	366	33	6
german	1000	20	2
glass	214	9	7

haberman	306	3	2
heart	270	13	2
hepatitis	155	19	2
housevotes	435	16	2
iris	150	4	3
led7digit	500	7	10
magic	19020	10	2
mammographic	961	5	2
monk2	432	6	2
movement	360	90	15
mushroom	8124	22	2
page-blocks	5472	10	5
pendigits	10992	16	10
phoneme	5404	5	2
pima	768	8	2
ring	7400	20	2
satimage	6435	36	7
segment	2310	19	7
spambase	4597	55	2
splice	3190	60	3
texture	5500	40	11
thyroid	7200	21	3
tic-tac-toe	958	9	2
twonorm	7400	20	2
vehicle	846	18	4
wisconsin	683	9	2
wine	178	13	3
yeast	1484	8	10
zoo	101	17	7

Πίνακας 5.1 Πληροφορίες Προβλημάτων

5.3 Σύγκριση αποτελεσμάτων

Οι αλγόριθμοι που χρησιμοποιήσαμε χωρίζονται σε 5 οικογένειες ταξινομητών, όπως φαίνονται στον παρακάτω πινάκα. Η σύγκριση των αποτελεσμάτων θα γίνει σε δυο στάδια, το πρώτο στάδιο είναι η σύγκριση αλγορίθμων όσον αφορά το πλήθος νικών τους ανά οικογένεια ταξινομητών. Και το δεύτερο στάδιο είναι η σύγκριση της μέσης ακρίβειας όλων των αλγορίθμων μαζί.



Πίνακας 5.2 Οικογένειες Ταξινομητών

5.3.1 Σύγκριση αλγορίθμων ανά οικογένεια ταξινομητών

Η πρώτη οικογένεια ταξινομητών που θα συγκρίνουμε για το πλήθος νικών των αλγορίθμων στα σχετικά προβλήματα είναι οι αλγόριθμοι Bayes. Θα επακολουθήσουν στους σχετικούς πίνακες οι οικογένειες ταξινομητών Functions, Lazy, Rules & Decision Trees.

		Bayes	
Α/Α	**Πρόβλημα**	**Bayes Net**	**Naive Bayes**
1	appendicitis	84,91	86,73
2	australian	86,23	81,16
3	automobile	71,17	60,46
4	banana	71,81	72,15
5	breast	72,07	71,7

6	bupa	56,25	67,27
7	chess	87,83	87,8
8	cleveland	58,75	57,04
9	coil2000	82,86	82,34
10	contraceptive	50,92	51,4
11	crx	86,51	79,48
12	dermatology	97,54	97,27
13	ecoli	81,23	87,5
14	flare	74,3	74,4
15	german	75,5	74,9
16	glass	70,61	49,11
17	haberman	72,52	75,14
18	heart	83,44	84,13
19	hepatitis	83,21	85,75
20	housevotes	91,39	91,39
21	iris	92,67	96,67
22	led7digit	74,2	74
23	lymph	85,71	82,33
24	magic	77,68	76,12
25	mammographic	82,17	82,65
26	monk2	67,14	66,67
27	movement	61,39	65,56
28	nursery	90,37	90,37
29	page-blocks	93,62	93,99
30	pendigits	87,9	88,64
31	phoneme	76,96	78,24
32	pima	76,82	75,4
33	ring	97,5	97,8
34	saheart	70,79	69,92
35	satimage	81,43	82,13
36	segment	91,43	85,71
37	sonar	80,29	73,14
38	spambase	89,81	76,37

39	spect_heart	71,25	71,25
40	texture	79,45	80,31
41	thyroid	99,04	96,86
42	tic-tac-toe	69,41	69,62
43	titanic	77,6	77,83
44	twonorm	97,31	97,59
45	vehicle	59,81	60,76
46	vowel	97,57	97,17
47	wine	98,86	97,75
48	wisconsin	97,28	97,57
49	yeast	76,21	77,5
50	zoo	94,18	95,09
Μέση Τιμή		80,70	79,84

Πίνακας 5.3 BayesNet vs Naive Bayes

Από τον παραπάνω πίνακα 5.3 μπορούμε να διαπιστώσουμε με απλούς υπολογισμούς ότι, ο αλγόριθμος BayesNet σημειώνει 22 νίκες ανάμεσα σε αυτά τα 50 προβλήματα. Ο αλγόριθμος Naive Bayes σημειώνει 25 νίκες και σε 3 περιπτώσεις έχουμε ισοπαλία τον δυο αυτών αλγορίθμων.

		Functions		
A/A	**Πρόβλημα**	**MLP**	**Simple Logistic**	**SMO**
1	appendicitis	80,18	86,91	86,91
2	australian	85,22	84,93	84,93
3	automobile	47,21	76,83	69,96
4	banana	53,02	56,08	55,17
5	breast	71,01	75,16	69,63
6	bupa	57,98	66,92	58,28
7	chess	97,5	96,59	95,78
8	cleveland	62,35	60,35	60,03
9	coil2000	94,03	94,01	94,03

10	contraceptive	47,52	51,46	51,26
11	crx	86,82	86,06	86,2
12	dermatology	97,27	97,83	95,35
13	ecoli	62,79	87,18	84,22
14	flare	74,12	74,87	73,27
15	german	74,7	75,9	75,1
16	glass	42,51	64,03	56,13
17	haberman	73,53	73,83	73,53
18	heart	84,12	83,16	84,12
19	hepatitis	82,04	83,88	85,17
20	housevotes	95,29	96,56	96,99
21	iris	75,33	94	96
22	led7digit	57,8	74,2	73,8
23	lymph	79,71	83	86,43
24	magic	83,12	78,99	79,12
25	mammographic	81,81	82,77	82,29
26	monk2	67,14	67,14	67,14
27	movement	33,06	75,56	72,5
28	nursery	97,37	92,73	92,97
29	page-blocks	92,27	96,44	92,82
30	pendigits	93,46	95,45	97,96
31	phoneme	77,79	75,04	77,28
32	pima	75,78	77,35	77,34
33	ring	85,04	75,88	76,38
34	saheart	71,66	72,28	72,7
35	satimage	85,75	85,91	86,82
36	segment	90,52	95,11	93,07
37	sonar	75,98	76,95	75,95
38	spambase	83,85	92,59	90,41
39	spect_heart	68,75	65	70
40	texture	97,35	99,64	98,91
41	thyroid	93,74	95,78	93,79
42	tic-tac-toe	98,12	98,22	98,33

43	titanic	77,56	77,6	77,6
44	twonorm	97,61	97,84	97,8
45	vehicle	63,94	77,89	74,34
46	vowel	96,15	96,96	95,95
47	wine	97,19	98,86	98,3
48	wisconsin	96	96,43	96,71
49	yeast	76,08	75,68	74,46
50	zoo	73,36	94,18	96,18
Μέση Τιμή		78,25	82,76	81,99

Πίνακας 5.4 MLP vs Simple Logistic vs SMO

Από τον παραπάνω πίνακα 5.4 αντλούμε τις εξής πληροφορίες: ο αλγόριθμος MLP σημείωσε 9 καθαρές νίκες, με τον όρο καθαρές νίκες περιγραφώ τις νίκες στις οποίες, ο αλγόριθμος ήταν ο μόνος με το μεγαλύτερο αποτέλεσμα στο πρόβλημα χωρίς να υπάρχει ισοψηφία με κάποιον άλλον αλγόριθμο. Ο Simple Logistic σημείωσε 26 καθαρές νίκες και ο SMO 10. Υπάρχει 1 ισοπαλία μεταξύ όλων των αλγορίθμων, 2 ισοπαλίες μεταξύ MLP-SMO και 2 ισοπαλίες μεταξύ SL-SMO.

		Lazy		
A/A	**Πρόβλημα**	**1-NN**	**3-NN**	**10-NN**
1	appendicitis	80,09	85,09	86,91
2	australian	81,16	84,64	85,94
3	automobile	79,96	69,96	63,63
4	banana	87,25	88,4	89,68
5	breast	72,43	73,78	73,09
6	bupa	62,92	61,73	59,38
7	chess	96,34	96,4	94,74
8	cleveland	53,8	56,13	55,13
9	coil2000	90,71	92,96	94,03
10	contraceptive	43,59	44,67	48,13
11	crx	81,93	86,21	86,67

12	dermatology	94,54	95,92	95,64
13	ecoli	80,37	85,41	86,01
14	flare	74,12	74,31	74,03
15	german	72	73,3	74
16	glass	70,5	71,95	66,39
17	haberman	68,3	71,54	74,54
18	heart	76,22	81,11	82,46
19	hepatitis	80,63	81,25	82,63
20	housevotes	91,85	91,43	91
21	iris	95,33	95,33	96
22	led7digit	70,4	72,4	75,4
23	lymph	82,33	80,38	80,9
24	magic	81,06	83,23	83,62
25	mammographic	77,23	79,64	80,6
26	monk2	75,71	84,06	76,66
27	movement	86,67	79,72	65
28	nursery	97,68	97,68	97,68
29	page-blocks	96,09	96,14	95,16
30	pendigits	99,36	99,35	99,03
31	phoneme	90,49	88,9	86,23
32	pima	71,36	75,28	72,8
33	ring	75,15	72,09	67,64
34	saheart	64,7	69,47	69,02
35	satimage	90,46	90,96	90,15
36	segment	97,14	96,02	94,33
37	sonar	86,57	86,02	75,98
38	spambase	90,78	90,22	89,22
39	spect_heart	53,75	67,5	67,5
40	texture	99,05	98,76	97,95
41	thyroid	92,53	94,06	93,92
42	tic-tac-toe	98,96	98,96	98,85
43	titanic	79,06	79,06	78,69
44	twonorm	94,77	96,53	97,11

45	vehicle	69,27	70,92	69,15
46	vowel	100	99,69	98,08
47	wine	95,49	96,6	97,71
48	wisconsin	94,85	96,57	96,57
49	yeast	71,57	72,91	75,07
50	zoo	96,18	92,18	88,18
Μέση Τιμή		82,25	83,34	82,36

Πίνακας 5.5 1-NN vs 3-NN vs 10-NN

Ο πίνακας 5.5 μας παρέχει τις εξής πληροφορίες για τους Lazy αλγορίθμους: οι καθαρές νίκες του 1-NN είναι 14, 13 του 3-NN και 18 του 10-NN. Υπάρχουν 5 ισοπαλίες από τις οποίες, 2 είναι μεταξύ των 1-NN και 3-NN, άλλες 2 είναι μεταξύ των 3-NN και 10-NN και 1 ισοπαλία είναι μεταξύ και των τριών αλγορίθμων.

		Rules		
A/A	**Πρόβλημα**	**Decision Table**	**JRip**	**PART**
1	appendicitis	83,09	85	85,09
2	australian	83,48	85,8	85,36
3	automobile	64,29	78,71	78,67
4	banana	75,08	88,53	86,87
5	breast	73,47	70,95	71,33
6	bupa	57,72	64,64	63,81
7	chess	97,37	98,94	99,12
8	cleveland	56,09	54,46	51,17
9	coil2000	94,03	93,97	91,49
10	contraceptive	54,52	50,65	50,31
11	crx	84,97	86,05	85
12	dermatology	86,87	86,88	94,53
13	ecoli	76,77	80,64	83,6
14	flare	75,72	71,87	72,9
15	german	71	71,7	70,2

16	glass	68,2	68,66	68,14
17	haberman	71,24	72,83	69,58
18	heart	76,17	81,45	79,86
19	hepatitis	76,13	78	84,46
20	housevotes	95,69	96,99	94,4
21	iris	92,67	95,33	94
22	led7digit	68,4	72	74,2
23	lymph	77	77,76	76,24
24	magic	82,45	84,37	85,39
25	mammographic	83,61	83,49	82,77
26	monk2	67,14	73,86	91,19
27	movement	42,78	54,17	66,94
28	nursery	94,85	96,74	98,94
29	page-blocks	95,76	96,88	97,26
30	pendigits	75,78	96,39	96,82
31	phoneme	80,77	86,14	83,14
32	pima	74,48	74,48	74,74
33	ring	75,97	92,58	93,46
34	saheart	70,8	70,55	70,78
35	satimage	82,35	86,56	86,65
36	segment	87,66	95,41	96,23
37	sonar	69,21	73,07	80,31
38	spambase	90,31	92,39	93,59
39	spect_heart	61,25	62,5	66,25
40	texture	78,65	92,71	94,33
41	thyroid	99,07	99,49	99,54
42	tic-tac-toe	73,39	97,81	94,26
43	titanic	77,6	78,06	79,06
44	twonorm	76,77	90,43	92,01
45	vehicle	64,89	66,19	71,87
46	vowel	97,98	98,78	98,79
47	wine	86,96	93,27	91,63
48	wisconsin	93,99	96,14	94,85

49	yeast	74,6	75,13	76,28
50	zoo	86,27	87,27	92,18
Μέση Τιμή		78,11	82,13	83,19

Πίνακας 5.6 Decision Table vs JRip vs PART

Από τα αποτελέσματα του πίνακα 5.6, λαμβάνουμε την γνώση ότι ο αλγορίθμους Decision Table σημείωσε 7 νίκες, ο JRIP 16 νίκες και ο PART 27 νίκες. Να σημειωθεί ότι σε αυτήν την οικογένεια ταξινομητών δεν υπήρξαν ισοπαλίες για αυτά τα 50 προβλήματα. Αυτό γίνεται αντιληπτό και από το άθροισμα των νικών που σημείωσαν οι αλγόριθμοι, 7+16+27=50.

		Decision Trees		
A/A	**Πρόβλημα**	**J48**	**LMT**	**Random Forest**
1	appendicitis	85,09	86	87,91
2	australian	86,09	84,78	86,96
3	automobile	80,54	76,83	86,21
4	banana	89,04	89,28	88,87
5	breast	75,54	75,16	69,63
6	bupa	68,71	66,41	73,11
7	chess	99,41	99,59	99,25
8	cleveland	53,13	60,35	56,09
9	coil2000	93,95	94,01	92,86
10	contraceptive	54,72	55,13	52,34
11	crx	85,44	86,21	87,44
12	dermatology	94	97,83	95,9
13	ecoli	84,23	87,18	86,01
14	flare	74,68	74,87	74,4
15	german	70,5	75,9	76,4
16	glass	66,75	68,59	79,89
17	haberman	72,83	73,83,	68,97
18	heart	77,52	83,16	83,12
19	hepatitis	83,79	83,25	85,13
20	housevotes	96,12	96,56	96,56

21	iris	96	94	95,33
22	led7digit	72,8	74,2	71,2
23	lymph	76,95	83	83,67
24	magic	85,36	86,69	88,16
25	mammographic	83,98	85,06	79,64
26	monk2	90,08	87,08	95,14
27	movement	66,67	75,56	81,94
28	nursery	96,54	98,83	98,84
29	page-blocks	96,89	97,04	97,44
30	pendigits	96,56	98,54	99,13
31	phoneme	86,75	86,55	91,32
32	pima	74,35	77,48	77,08
33	ring	90,27	89,47	95,18
34	saheart	70,15	72,28	68,83
35	satimage	86,26	87,65	91,78
36	segment	96,93	96,19	97,88
37	sonar	71,17	77,9	81,24
38	spambase	92,98	93,74	95,5
39	spect_heart	71,25	66,25	63,75
40	texture	93,22	99,64	97,85
41	thyroid	99,65	99,5	99,64
42	tic-tac-toe	84,55	98,22	96,35
43	titanic	79,06	78,33	78,33
44	twonorm	84,68	97,84	96,65
45	vehicle	72,46	83,57	75,06
46	vowel	98,99	99,29	99,49
47	wine	93,86	98,86	98,3
48	wisconsin	94,13	96,43	96,86
49	yeast	75,2	77,43	77,03
50	zoo	92,18	94,18	93,18
Μέση Τιμή		83,24	85,55	93,18

Πίνακας 5.7 J48 vs LMT vs Random Forest

Η τελευταία οικογένεια ταξινομητών που θα συγκρίνουμε τους αλγορίθμους, ως προς τις νίκες που σημείωσαν στα σχετικά προβλήματα που εκτελέσαμε, είναι η οικογένεια των Decision Trees. Από τον πίνακα 5.7 αντλούμε τις πληροφορίες ότι, ο αλγόριθμος J48 σημείωσε 7 καθαρές νίκες, ο LMT σημείωσε 20 καθαρές νίκες και ο Random Forest 22 καθαρές νίκες. Υπήρξε όμως και 1 ισοπαλία μεταξύ των αλγορίθμων LMT-Random Forest.

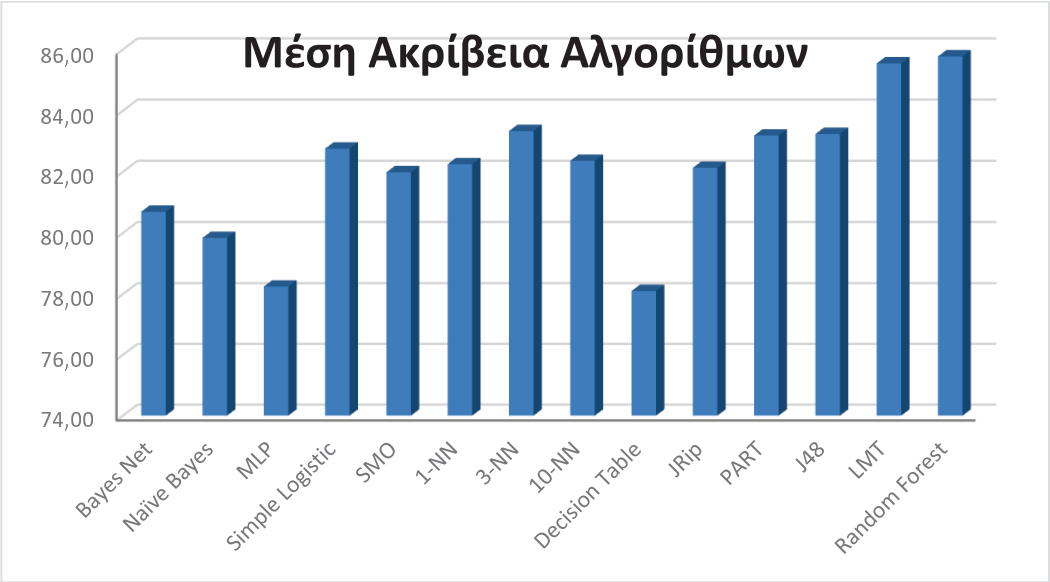
5.3.2 Σύγκριση όλων των αλγορίθμων

Όπως γνωρίζουμε ήδη οι αλγόριθμοι που χρησιμοποιήσαμε για να τρέξουμε τα 50 προβλήματα είναι οι εξής: BayesNet, Naive Bayes, MLP, Simple Logistic, SMO, 1-NN, 3-NN, 10-NN, Decision Table, JRip, PART, J48, LMT και Random Forest. Στην συνεχεία της εργασίας θα συγκρίνουμε όλους τους αλγορίθμους μαζί ως προς την μέση ακρίβεια που σημείωσαν σε αυτά τα 50 προβλήματα που χρησιμοποιήσαμε.

	Αλγόριθμοι													
	BayesNet	Naive Bayes	MLP	Simple Logistic	SMO	1-NN	3-NN	10-NN	Decision Table	JRip	PART	J48	LMT	Random Forest
Μέση Ακρίβεια	80,70	79,84	78,25	82,76	81,99	82,25	83,34	82,36	78,11	82,13	83,19	83,24	85,55	85,78

Πίνακας 5.8 Μέση Ακρίβεια Αλγορίθμων

Στον πίνακα 5.8 παρουσιάζονται οι μεσαίες ακρίβειες των αλγορίθμων που σημείωσαν στο πείραμα. Αναλύοντας λίγο τον πίνακα αυτόν, μπορούμε να διακρίνουμε ότι, η χαμηλότερη μέση ακρίβεια ανήκει στον αλγόριθμο Decision Table και η υψηλότερη μέση ακρίβεια ανήκει στον αλγόριθμο Random Forest.



Γράφημα 5.1 Μέση Ακρίβεια Αλγορίθμων

Με το παραπάνω γράφημα 5.1 μπορούμε να διακρίνουμε καλύτερα ότι οι αλγόριθμοι, οι οποίοι ανήκουν στην οικογένεια ταξινομητών Decision Trees (J48, LMT & Random Forest) έχουν υψηλή μέση ακρίβεια στο πείραμα μας και αυτό τους καθίστα πιο «δυνατούς» απέναντι στους υπόλοιπους αλγορίθμους.

5.4 Friedman aligned test

Η στατιστική σύγκριση των πολλαπλών αλγορίθμων σε πολλαπλά σύνολα δεδομένων είναι θεμελιώδης στην μηχανική μάθηση και συνήθως εκτελείτε μέσω στατιστικών τεστ. Ως εκ τούτου, χρησιμοποιήσαμε το μη παραμετρικό τεστ Friedman Aligned Ranking (Hodges & Lehmann, 1962) προκειμένου να αξιολογήσουμε την απόρριψη της υπόθεσης ότι όλοι οι ταξινομητές εκτελούνται εξίσου καλά σε ένα δεδομένο επίπεδο. Από την στιγμή που το τεστ είναι μη παραμετρικό, δεν απαιτεί ισομετρία στα μέτρα απέναντι σε διαφορετικά σύνολα δεδομένων, δεν υποθέτει την ομαλότητα των μέσων δειγματοληψίας και είναι ανθεκτικό στις αποκλίσεις.

Στον παρακάτω πίνακα παρατίθενται τα αποτελέσματα από το Friedman Aligned Test.

Ταξινομητής	Friedman Aligned Test
LMT	186.63
Random Forest	205.96
Simple Logistic	305.53
C4.5 (J48)	311.22
SMO	326.22
3NN	329.67
PART	344.58
10NN	358.59
Naive Bayes	376.44
Bayes Net	388.54
1NN	403.99
JRip	419.48
MLP	442.30
Decision Table	507.85

Πίνακας 5.9 Αποτελέσματα Friedman Aligned Test

Από τα αποτελέσματα του Friedman aligned test είναι αξιοσημείωτο να αναφέρουμε ότι, οι αλγόριθμοι LMT & Random Forest σημειώνουν τις χαμηλότερες τιμές και οι αλγόριθμοι MLP & Decision Table τις υψηλότερες τιμές. Στο Friedman aligned test όσο χαμηλότερη είναι η τιμή τόσο υψηλότερη είναι η απόδοση του αλγορίθμου και αντιθέτως όσο υψηλότερη είναι η τιμή τόσο χαμηλότερη είναι η απόδοση του αλγορίθμου. Έτσι καταλήγουμε στο συμπέρασμα ότι οι δυνατότεροι αλγόριθμοι που έτρεξαν το πείραμα μας είναι οι LMT & Random Forest οι οποίοι, ανήκουν στην οικογένεια ταξινομητών Decision Trees.

6 Συμπεράσματα

Τα τελευταία χρόνια με την εξάπλωση της χρήσης των υπολογιστών σε όλους τους τομείς της ζωής μας έχουν αυξηθεί σημαντικά οι δυνατότητές μας να παράγουμε και να συλλέγουμε πληροφορίες, γεγονός που οδήγησε στην συγκέντρωση μεγάλου όγκου πληροφορίας. Η αύξηση αυτή κάνει επιτακτική την ανάγκη εύρεσης νέων τεχνικών και εργαλείων που θα υποστηρίζουν την αυτόματη μετατροπή των υπό επεξεργασία πληροφοριών σε χρήσιμη γνώση. Η Μηχανική Μάθηση και η Εξόρυξη Γνώσης αποτελούν δύο από τους κυριότερους τομείς έρευνας προς αυτή την κατεύθυνση. (Κωτσιαντής, 2005) Ο σκοπός της παρούσας πτυχιακής εργασίας ήταν η παρουσίαση και η ανάλυση της Εξόρυξης Γνώσης σε σύγχρονες σχεσιακές βάσεις δεδομένων. Ο αναγνώστης είχε την ευκαιρία να γνωρίσει τα πιο δημοφιλή και σύγχρονα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων και να μάθει πια εργαλεία διαθέτει το καθένα από αυτά για την Εξόρυξη Γνώσης. Εφαρμόζοντας την Εξόρυξη Γνώσης με ένα πείραμα, το οποίο περιείχε 50 σύνολα δεδομένων με έναν περιορισμένο αριθμό αλγορίθμων, καταλήξαμε σε ένα συμπέρασμα. Οι αλγόριθμοι LMT & Random Forest, οι οποίοι ανήκουν στην οικογένεια ταξινομητών Decision Trees, ανταποκρίθηκαν καλυτέρα στο πείραμα μας σε σχέση με τους υπόλοιπους αλγόριθμους. Αυτό τους αναδεικνύει ως τους πιο αποδοτικούς. Με τη βοήθεια της Εξόρυξης Γνώσης έχουμε τη δυνατότητα να καταλήξουμε και σε άλλα πολύτιμα συμπεράσματα για τα δεδομένα μας ανάλογα με τον επιστημονικό κλάδο στον οποίο βρισκόμαστε και τον στόχο που έχουμε. Η βιοπληροφορική, η εγκληματολογία, η χρηματοοικονομική & τραπεζική, η εκπαίδευση, η υγεία κ.λπ., χρησιμοποιούν τεχνικές Εξόρυξης Γνώσης για να βελτιστοποιήσουν την εξαγωγή της δυνητικά ωφέλιμης γνώσης από τον μεγάλο όγκο δεδομένων. Όπως γίνεται αντιληπτό η Εξόρυξη Γνώσης διευρύνει τον ορίζοντα χρήσης της καθώς επίσης διευρύνεται η ανάγκη αξιοποίησης χρήσιμης γνώσης σε εμπορικούς και επιστημονικούς τομείς.

Παράρτημα Α

Λογισμικό Weka

Α.1 Παρουσίαση του Λογισμικού Weka

Το WEKA (Wekato Environment for knowledge Analysis) είναι μια συλλογή από αλγόριθμους μηχανικής μάθησης για την επίλυση πραγματικών προβλημάτων εξόρυξης δεδομένων. Δημιουργήθηκε από ερευνητές στο Πανεπιστήμιο Waikato στη Νέα Ζηλανδία και χρησιμοποιείται για έρευνα, εκπαίδευση και για εφαρμογές. Το Weka χρηματοδοτήθηκε από την NZ government το 1993 και είναι ένα λογισμικό ανοικτού κώδικα που εκδίδεται υπό την άδεια της GNU General Public License.

Εικόνα Α.1: Λογισμικό Weka

Είναι γραμμένο σε γλώσσα προγραμματισμού Java έτσι ώστε μπορεί να χρησιμοποιηθεί με όσο το δυνατόν περισσότερα λειτουργικά συστήματα και διατίθεται ελεύθερα (συμπεριλαμβανομένου του πηγαίου κώδικα). Παράλληλα διαθέτει γραφικό περιβάλλον εργασίας χρήστη (GUI) κάνοντας πιο εύκολη την πρόσβαση σε αυτό. Οι αλγόριθμοι είτε μπορούν να εφαρμοστούν κατευθείαν σε ένα σύνολο δεδομένων ή μπορεί να κληθούν από κάποιο κώδικα γραμμένο σε Java.

Το WEKA περιέχει εργαλεία για

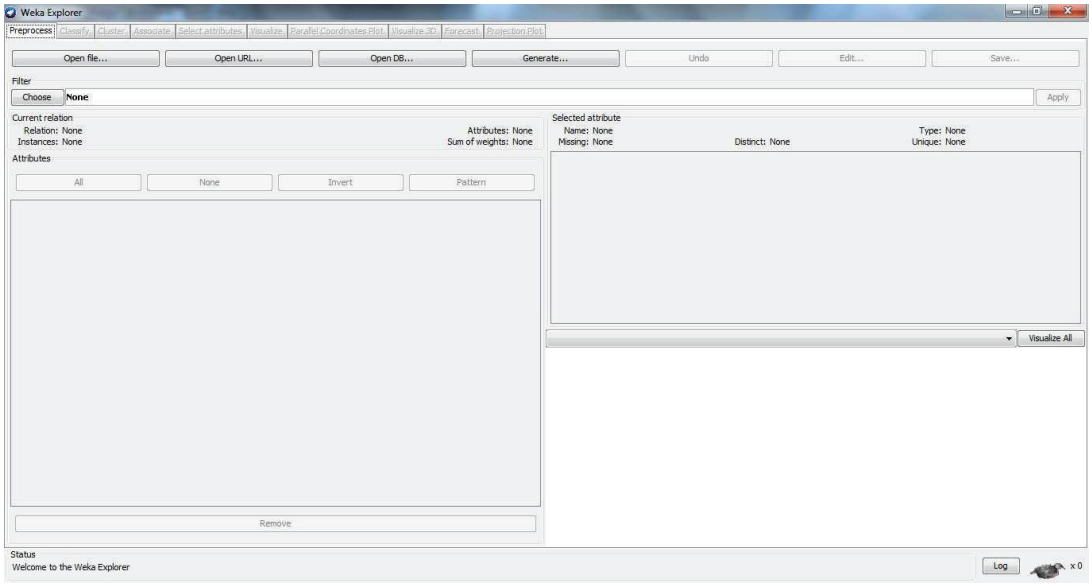
- Προ-επεξεργασία (pre process).
- Ταξινόμηση (classification).
- Παλινδρόμηση (regression).
- Ομαδοποίηση (clustering).
- Εύρεση κανόνων συσχέτισης (classification rules).
- Απεικόνιση των δεδομένων (visualization).

Επίσης, θεωρείται ιδιαίτερα κατάλληλο για την ανάπτυξη νέων συστημάτων μηχανικής μάθησης.

Το WEKA GUI Chooser window χρησιμοποιείται για να αρχίσει κάποιος τη χρήση του εργαλείου. Όπως παρουσιάζεται στην Εικόνα Α.1 υπάρχουν τέσσερα κουμπιά, τα οποία

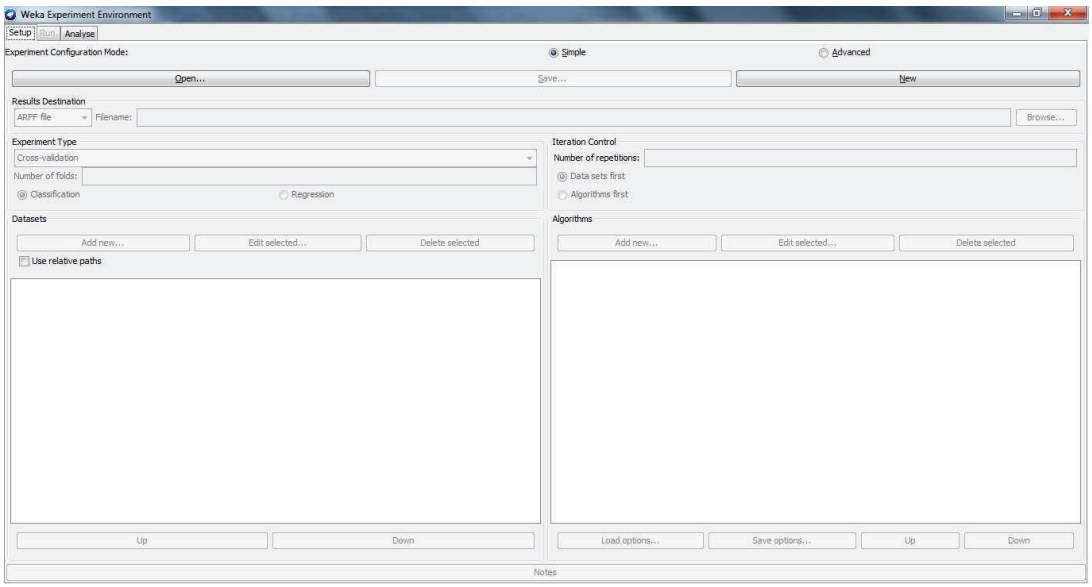
αναλύουμε συνοπτικά πιο κάτω.

- *Explorer*: Παρέχει γραφικό περιβάλλον για τις διεργασίες του Weka και τα συστατικά του μέρη, και χρησιμοποιείται περισσότερο για τη εξερεύνηση των δεδομένων.



Εικόνα Α.2: Explorer

- *Experimenter*: Επιτρέπει τη δημιουργία πειραμάτων και στατιστικών αναλύσεων των σχημάτων που παρέχονται.

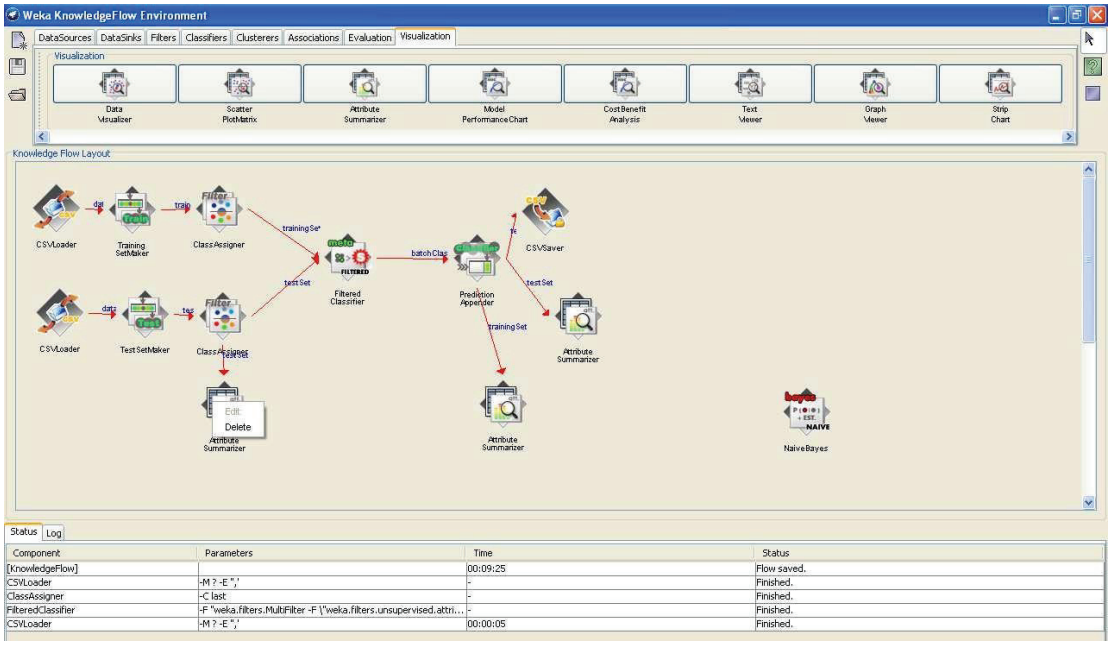


Εικόνα Α.3: Experimenter

- *Simple CLI* : Παρέχει γραμμή εντολών για τις διεργασίες του ωεκα και είναι περισσότερο για λειτουργικά συστήματα που δεν έχουν γραμμή εντολών.

Εικόνα A.4: Simple CLI

- *KnowledgeFlow*: Παρέχει τις ίδιες δυνατότητες με το Experiment αλλά με ένα περιβάλλον που επιτρέπει drag and drop.



Εικόνα A.5: KnowledgeFlow

A.2 Αρχεία ARFF

Ενα αρχείο ARFF (Attribute Relation File Format) είναι ένα αρχείο κειμένου ASCII, το οποίο περιγράφει έναν αριθμό παραδειγμάτων που έχουν κοινά ένα σύνολο από χαρακτηριστικά. Τα αρχεία ARFF έχουν δύο ξεχωριστά τμήματα. Το πρώτο τμήμα είναι η επικεφαλίδα (header), και το δεύτερο είναι το τμήμα των δεδομένων.

Η μορφή του αρχείου ARFF είναι η εξής :

@RELATION <relation name>
@ATTRIBUTE <όνομα-χαρακτηριστικού 1> <τύπος δεδομένου>
@ATTRIBUTE <όνομα-χαρακτηριστικού 2> <τύπος δεδομένου>
.
@ATTRIBUTE <όνομα-χαρακτηριστικού n> <τύπος δεδομένου>

@DATA
attr1_value, attr2_value, attr3_value, ..., attrn_value
attr1_value, attr2_value, attr3_value, ..., attrn_value
.
attr1 value, attr2 value, attr3 value, ..., attrn value

Η επικεφαλίδα του αρχείου ARFF περιέχει το όνομα της σχέσης, μια λίστα των μεταβλητών (οι στήλες στα δεδομένα), και τους τύπους τους. Το όνομα της σχέσης

ορίζεται ως η πρώτη γραμμή στο αρχείο arff. Η μορφή είναι :

@RELATION <relation-name>

όπου το <relation-name> είναι ένα string και αποτελεί το όνομα του συνόλου δεδομένων.

Ο ορισμός των χαρακτηριστικών έχει την μορφή μια ταξινομημένης σειράς δηλώσεων @ATTRIBUTE. Κάθε χαρακτηριστικό στο σετ δεδομένων έχει την δική του δήλωση @ATTRIBUTE, η οποία ορίζει μοναδικά το όνομα του χαρακτηριστικού και τον τύπο του. Η σειρά με την οποία ορίζονται τα χαρακτηριστικά δηλώνει τον αριθμό στήλης στο τμήμα δεδομένων του αρχείου. Ο τρόπος της δήλωσης @ATTRIBUTE είναι :

@ATTRIBUTE <όνομα-χαρακτηριστικού> <τύπος δεδομένου>

όπου το <όνομα-χαρακτηριστικού> είναι ένα string που πρέπει να ξεκινά με γράμμα. Ο <τύπος δεδομένου> μπορεί να είναι οποιοδήποτε από τους παρακάτω πέντε τύπους που αναγνωρίζει το WEKA:

- numeric.
- integer.
- real.
- string.
- <ονομαστικός-ορισμός>.

Ο μόνος τύπος που θα επεξηγηθεί είναι ο <ονομαστικός-ορισμός>, ο οποίος είναι ορισμός με βάση προκαθορισμένες τιμές, για παράδειγμα

@ATTRIBUTE attr {1,2,3,4}

ορίζει το χαρακτηριστικό attr, το οποίο μπορεί να πάρει μόνο τις τιμές που εμφανίζονται εντός των αγκυλών. Το τμήμα των δεδομένων του αρχείου ARFF, περιέχει την ραμμή ορισμού και τις γραμμές των δεδομένων.

Η δήλωση @DATA είναι μια μονή γραμμή που δηλώνει την αρχή των δεδομένων στο αρχείο. Κάθε παράδειγμα αντιπροσωπεύεται από μια μονή γραμμή, με μια αλλαγή γραμμής να δηλώνει το τέλος του. Οι τιμές των χαρακτηριστικών, όπως ορίστηκαν στην επικεφαλίδα, για κάθε παράδειγμα χωρίζονται με κόμμα.

Τα δεδομένα μπορούν να δοθούν και από μία ηλεκτρονική διεύθυνση ή από μία βάση δεδομένων. Επίσης, στο φάκελο C:\Program Files\Weka-3-5\data περιέχονται κάποια παραδείγματα τέτοιων αρχείων από πραγματικές εφαρμογές. Σε κάθε σύνολο δεδομένων μπορούν να εφαρμοστούν οι τεχνικές που παρουσιάσαμε στο Κεφάλαιο 4. Συγκεκριμένα το WEKA δίνει τη δυνατότητα στο χρήστη να εμφανιστούν γραφικά τα δεδομένα για κάθε ένα από τα γνωρίσματα ξεχωριστά καθώς και στατιστικές πληροφορίες για αυτές (π.χ. μέση τιμή, διασπορά κτλ). Εάν στο σύνολο δεδομένων δίνεται και κάποια κλάση στην οποία ταξινομούνται, τα δεδομένα που ανήκουν στην ίδια κλάση εμφανίζονται με το ίδιο χρώμα.

Ο χρήστης έχει τη δυνατότητα να χρησιμοποιήσει τις υλοποιήσεις των αλγορίθμων είτε από τη γραμμή εντολών είτε από το γραφικό περιβάλλον το οποίο προσφέρει το Weka, ενώ ο προγραμματιστής μπορεί να καλέσει τις υλοποιήσεις των αλγορίθμων από τα δικά του προγράμματα. Έτσι το Weka μπορεί να λειτουργήσει σαν μια βιβλιοθήκη υλοποιήσεων αλγορίθμων μηχανικής μάθησης, που μπορεί να χρησιμοποιηθεί για την δημιουργία νέων προγραμμάτων. Επίσης, καθώς παρέχει μια πλήρη βιβλιοθήκη με κώδικα για αξιολόγηση αποτελεσμάτων, μπορούν πολύ εύκολα να συγκριθούν νέες μέθοδοι με ήδη υπάρχουσες.

A.3 WEKA Explorer

Μόλις γίνει η εκκίνηση του Explorer παρατηρούμε πως στη κορυφή του παράθυρου υπάρχουν έξι καρτέλες, οι οποίες είναι

- Preprocess: Επιλέγει και τροποποιεί τα δεδομένα που χρησιμοποιούνται.
- Classify: Σχήματα εκπαίδευσης και ελέγχου που κατηγοριοποιούν ή εκτελούν παλινδρόμηση.
- Cluster: Δημιουργούνται συστάδες για δεδομένα.
- Associate: Δημιουργούνται κανόνες συσχέτισης για τα δεδομένα.
- Select attributes: Επιλέγει τις πιο κοινές μεταβλητές στα δεδομένα.
- Visualize: Προβάλει ένα διαδραστικό 2D διάγραμμα δεδομένων.

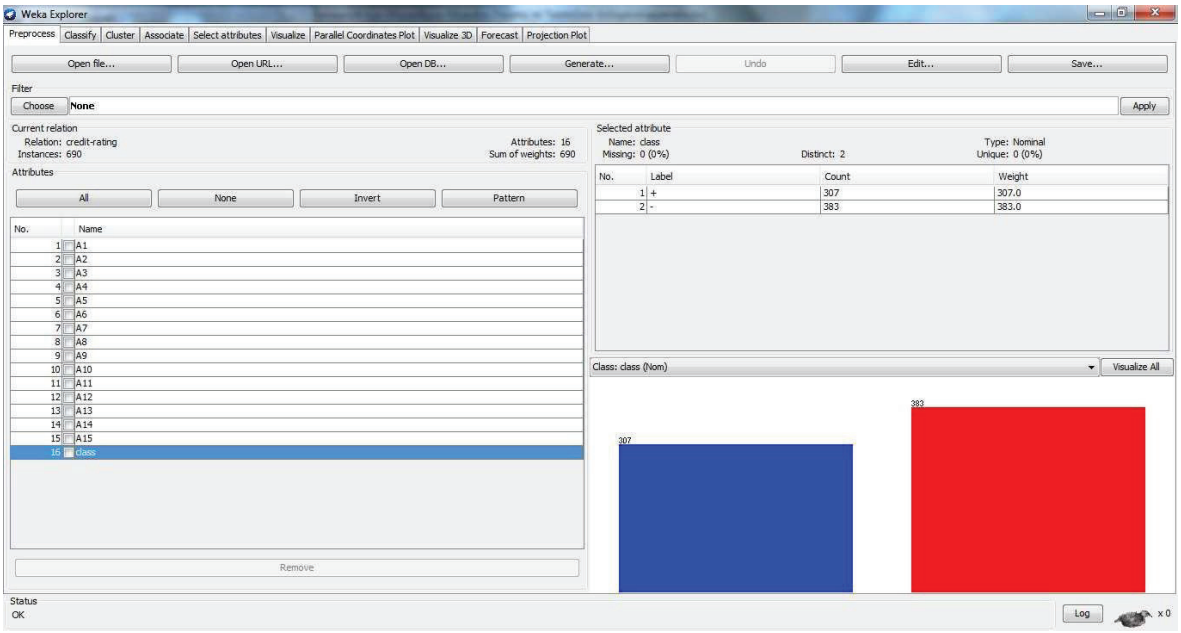
Σε όλες τις καρτέλες στο κάτω μέρος του παραθύρου εμφανίζεται το πεδίο Status, το οποίο δείχνει την κατάσταση που επικρατεί την εκάστοτε στιγμή. Με δεξί κλικ του ποντικιού σε εκείνο το σημείο, υπάρχει δυνατότητα να λάβουμε πληροφορίες για τη μνήμη του λογισμικού, καθώς και να δώσουμε εντολή ώστε να καθαριστούν κομμάτια μνήμης που χρησιμοποιούνται κάπου αλλού, ενώ δεν είναι απαραίτητο. Έτσι το πρόγραμμα εκτελείται ταχύτερα. Προφανώς, αυτές οι διεργασίες εκτελούνται στο παρασκήνιο.

Σε όλες τις καρτέλες υπάρχει επίσης το κουμπί Log. Επιλέγοντας το κουμπί ανοίγει ένα ξεχωριστό παράθυρο που περιέχει πληροφορίες κειμένου. Κάθε γραμμή του κειμένου αναφέρεται σε διαφορετική χρονική περίοδο που γράφτηκαν σχόλια μέσα στο αρχείο. Όσο εκτελείται το πρόγραμμα, το log κρατάει εγγραφές για το τι συμβαίνει.

A.3.1 Preprocess

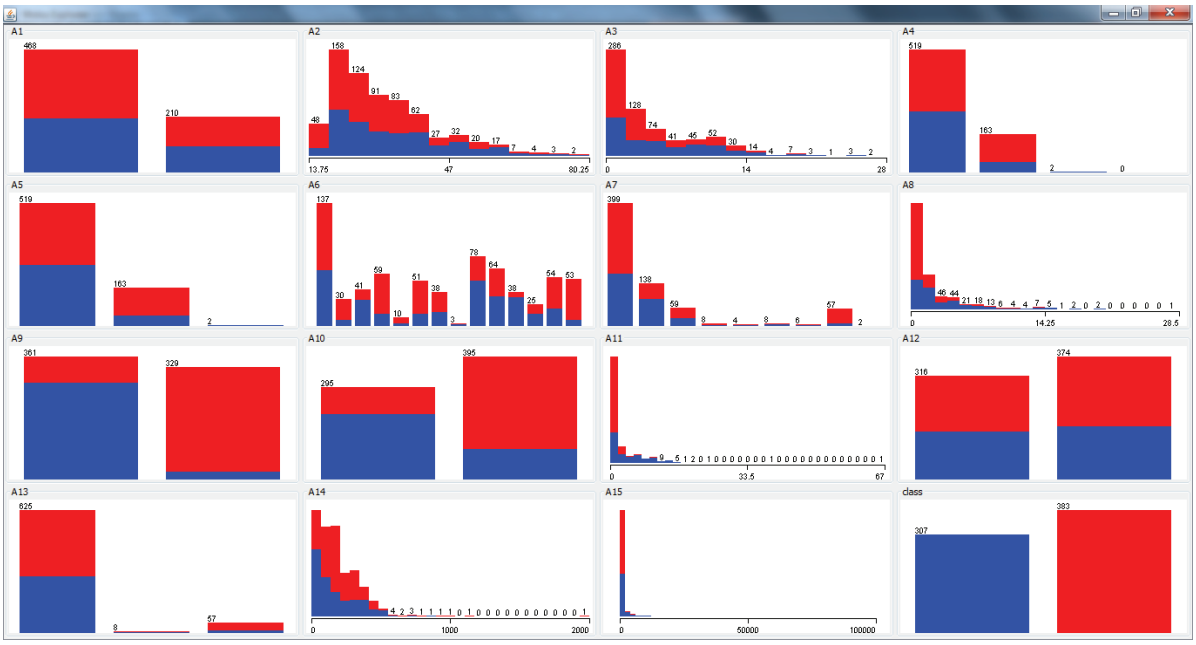
Τα πρώτα τρία κουμπιά που είναι ενεργά έχουν σχέση με το άνοιγμα του αρχείου, διεύθυνση URL ή βάσης δεδομένων. Χρησιμοποιώντας το κουμπί «Open file» μπορεί το λογισμικό να διαβάσει τα δεδομένα. Μόλις εισαχθούν τα δεδομένα, η καρτέλα «Preprocess» παρουσιάζει ένα πλήθος πληροφοριών όπως παρουσιάζεται στην Εικόνα Α6. Ο πίνακας «Current relation» έχει τρία πεδία.

- Relation: Παρουσιάζει το όνομα του αρχείου που περιέχει τα δεδομένα.
- Instances: Σημειώνει το πλήθος των εγγραφών που έχουμε στα δεδομένα.
- Attributes: Αναφέρει πόσα είναι τα γνωρίσματα.



Εικόνα Α.6: Καρτέλα Preprocess

Κάτω από το πίνακα που αναφέραμε υπάρχει ο πίνακας «Attributes». Σε αυτόν αναγράφονται όλες οι μεταβλητές που υπάρχουν και έχουμε τη δυνατότητα να επιλέξουμε συγκεκριμένες μεταβλητές και να τις διαγράψουμε. Κάθε φορά που γίνεται η επιλογή μίας μεταβλητής, αλλάζει ο πίνακας «selected attribute». Εκεί εμφανίζονται οι ιδιότητες κάθε επιλεγμένης μεταβλητής όπως το όνομα, ο τύπος, το πλήθος των ελλειπών εγγραφών, οι διαφορετικές τιμές και οι μοναδικές τιμές. Κάτω από τις παραπάνω ιδιότητες προβάλλονται περισσότερες πληροφορίες για τις μεταβλητές που εξαρτώνται από το τύπο δεδομένων. Για τις αριθμητικές τιμές δίνει τέσσερα στατιστικά στοιχεία (ελάχιστο, μέγιστο, μέσο όρο και τυπική απόκλιση). Αν είναι κατηγορηματικά δεδομένα εμφανίζει τον αριθμό των εγγραφών για κάθε κατηγορία. Κάτω από όλα αυτά τα στατιστικά στοιχεία υπάρχει ένα έγχρωμο ιστόγραμμα που περιγράφει τη συγκεκριμένη μεταβλητή. Υπάρχει και η δυνατότητα της ταυτόχρονης οπτικοποίησης των ιστογραμμάτων όλων των μεταβλητών με την επιλογή «Visualize All» (Εικόνα Α.7).



Εικόνα Α.7: Ιστογράμματα όλων των μεταβλητών

Α.3.1.1 Φίλτρα Προεπεξεργασίας Δεδομένων

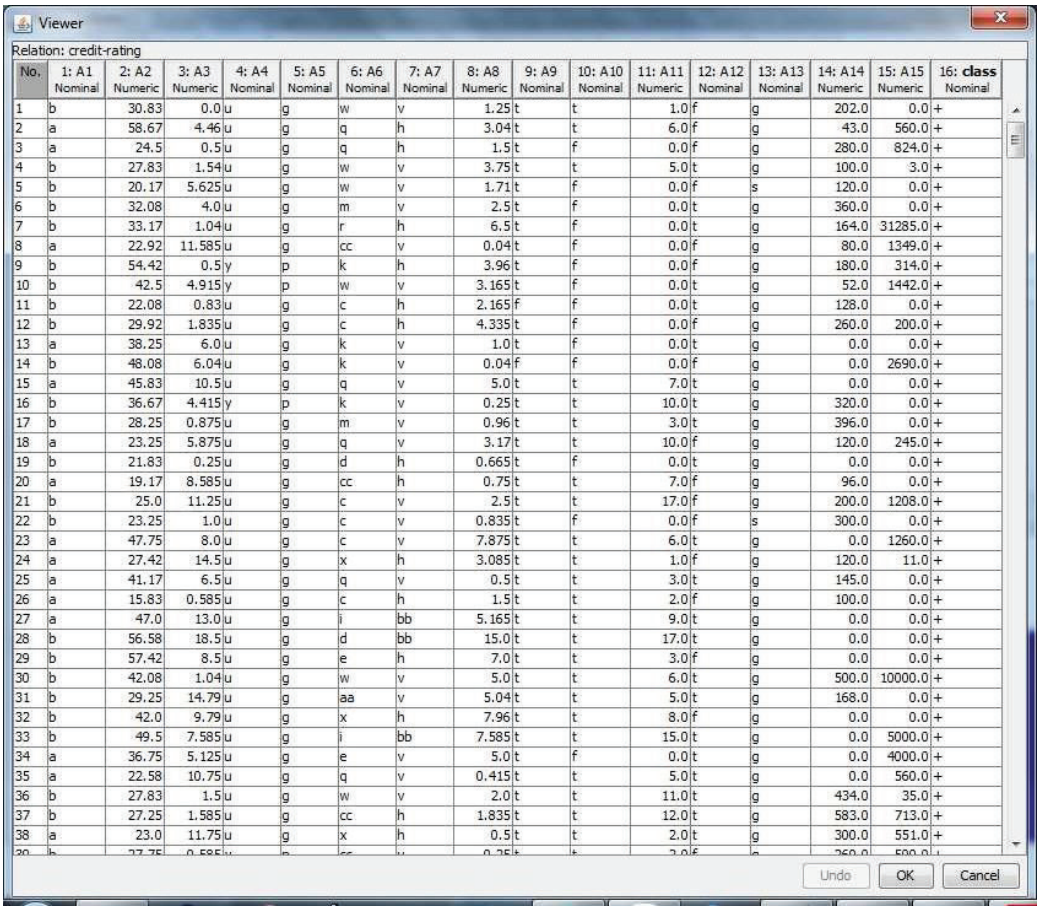
Το λογισμικό υποστηρίζει αρκετά φίλτρα που εφαρμόζονται στα δεδομένα ανάλογα με την όψη του προβλήματος που πρέπει να επιδιορθωθεί. Πιο συγκεκριμένα, αυτά τα φίλτρα μπορούν να χρησιμοποιηθούν για να μετασχηματίσουν τα δεδομένα (π.χ. να μετατρέψουν αριθμητικές τιμές σε αντίστοιχες διακριτές) και δύναται να διαγράψουν στιγμιότυπα και μεταβλητές σύμφωνα με συγκεκριμένα κριτήρια.

- Αν συγκεκριμένες τιμές πρέπει να αφαιρεθούν από τα δεδομένα, εφαρμόζεται το φίλτρο «Remove».

- Αν υπάρχουν αριθμητικά δεδομένα, που είναι πολύ μεγαλύτερα ή μικρότερα από το σύνολο δεδομένων, εφαρμόζεται το φίλτρο «NumericCleaner».

- Υπάρχει και ένα φίλτρο που αντικαθιστά τις ελλιπείς τιμές για κατηγορηματικά και αριθμητικά δεδομένα, το οποίο εφαρμόζεται με την εντολή «Apply».

Τέλος μπορεί να γίνει επεξεργασία των δεδομένων μέσα από το λογισμικό επιλέγοντας την εντολή «Edit». Το αρχείο παρουσιάζεται υπό τη μορφή πίνακα (Εικόνα Α.8) και έχει τη δυνατότητα ο χρήστης να αλλάξει τα δεδομένα καθώς και να συμπληρώσει τυχόν ελλειπείς τιμές.

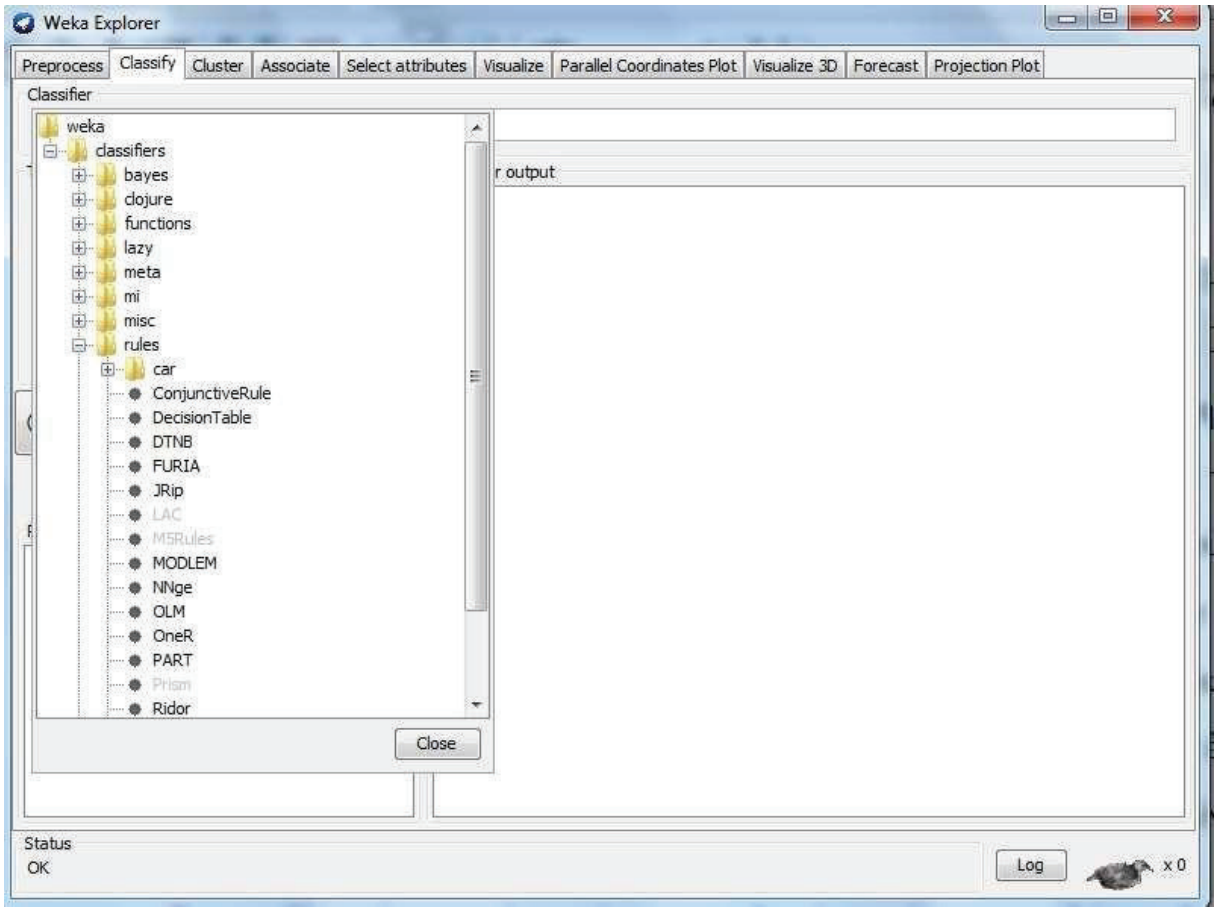


Εικόνα Α.8: Επεξεργασία των δεδομένων μέσα από το λογισμικό Weka

Α.3.2 Classify

Στην ενότητα ταξινόμηση υπάρχει ένα πλαίσιο «Classifier». Αυτό το πλαίσιο έχει ένα πεδίο κειμένου που προβάλει το όνομα του επιλεγόμενου αλγορίθμου ταξινόμησης και τις επιλογές του. Με το κουμπί «Choose» μπορούμε να διαλέξουμε τον αλγόριθμο που επιθυμούμε (Εικόνα Α.9).

Στο πεδίο «Test Option» υπάρχουν και οι επιλογές για τον επιλεγμένο αλγόριθμο. Η πρώτη επιλογή είναι η χρήση των δεδομένων εκπαίδευσης. Εδώ ο αλγόριθμος αξιολογείται από την επιτυχία της πρόβλεψης. Η δεύτερη επιλογή είναι η προμήθεια δεδομένων αξιολόγησης. Εδώ εκτιμάται ότι ο αλγόριθμος αξιολογείται με βάση ένα σύνολο δεδομένων που τα φορτώνουμε από ένα αρχείο. Η τρίτη επιλογή είναι η διασταυρωμένη επικύρωση, όπου πρέπει να δηλωθεί ο αριθμός των φακέλων που θα χρησιμοποιηθούν. Το ποσοστό διαχωρισμού, όπου δηλώνουμε ένα ποσοστό των δεδομένων για εκπαίδευση και το υπόλοιπο θα χρησιμοποιηθεί για αξιολόγηση.



Εικόνα Α.9: Επιλογή αλγορίθμου ταξινόμησης

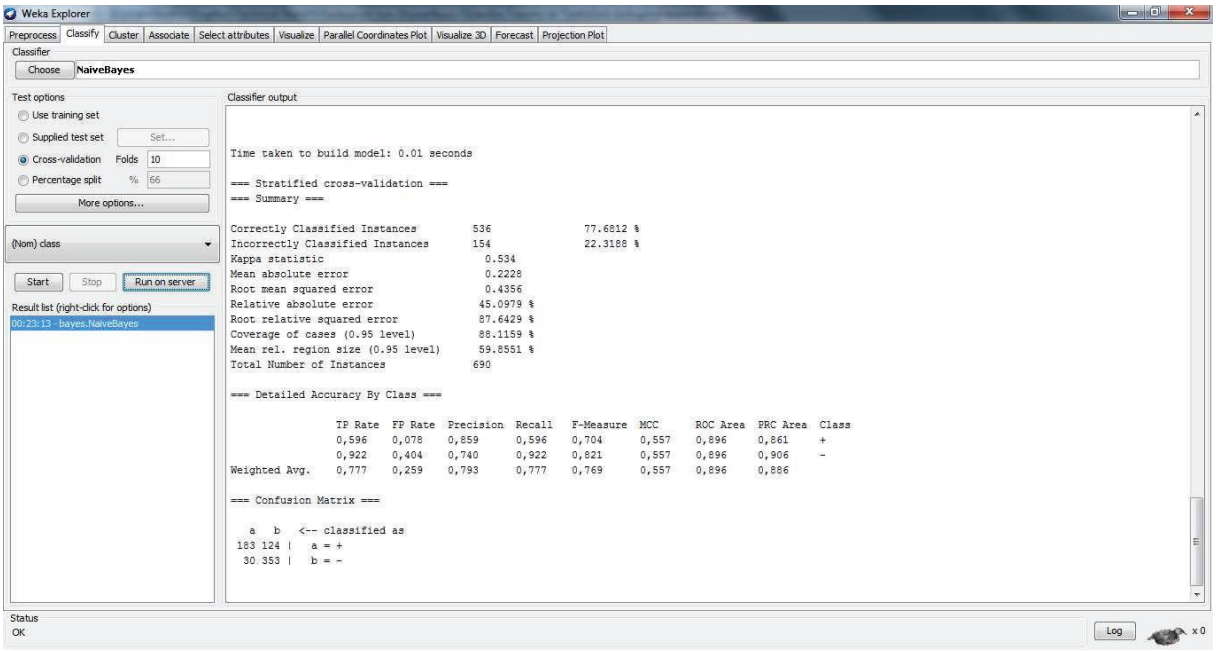
Στο πεδίο κάτω από αυτές τις επιλογές υπάρχει η κλάση, δηλαδή η μεταβλητή εξόδου. Το WEKA ταξινομεί χρησιμοποιώντας μόνο μία κλάση, η οποία είναι και ο στόχος της πρόβλεψης. Η προκαθορισμένη κλάση που είναι επιλεγμένη είναι η τελευταία από το σύνολο δεδομένων.

Αφού ολοκληρώσουμε τις επιλογές μας, με τη χρήση του κουμπιού «Start» γίνεται η επεξεργασία των δεδομένων όπου λαμβάνουμε και το αποτέλεσμα στο δεξί μέρος της οθόνης. Το αποτέλεσμα χωρίζεται σε ορισμένα κομμάτια:

- Run information: Υπάρχει μία λίστα από πληροφορίες σχετικά με τις επιλογές και τις ρυθμίσεις που κάναμε στα προηγούμενα βήματα.
- Classifier model (full training set): Μία περιγραφή του μοντέλου που χρησιμοποιήσαμε για τη ταξινόμηση των δεδομένων σε όλο το πλήθος των δεδομένων εκπαίδευσης.
- Evaluation on test split.
- Summary: Περιέχει τη ζητούμενη πληροφορία. Μία λίστα από στατιστική επεξεργασία και την εγκυρότητα του μοντέλου, που κρίνεται από το ποσοστό των δεδομένων αξιολόγησης που ταξινομήθηκαν σωστά με βάση τη πρόβλεψη.
- Detailed accuracy by class: Πιο λεπτομερής περιγραφή της πρόβλεψης.

- Confusion matrix: Παρουσιάζει πόσες περιπτώσεις έχουν ταξινομηθεί σε κάθε κλάση.

Μόλις γίνει η ταξινόμηση τα αποτελέσματα περιέχουν ορισμένες καταχωρήσεις. Με δεξί κλικ στη λίστα αποτελέσματος έχουμε κάποιες επιλογές. Η πιο χρήσιμη είναι η οπτικοποίηση του δέντρου απόφασης. Η συγκεκριμένη δημιουργεί σε γράφικη αναπαράσταση το μοντέλο ταξινόμησης.



Εικόνα Α.10: Εφαρμογή του αλγορίθμου Naive Bayes στο σύνολο δεδομένων

Οι ενότητες Cluster, Associate, Select attributes λειτουργούν με παρόμοιο τρόπο.

Βιβλιογραφία

Ταμπακάς, Β.Τ. 2011, Εισαγωγή στις Βάσεις Δεδομένων, Αθήνα.

‘Σύστημα διαχείρισης βάσης δεδομένων’ 2017, Wikipedia, wiki. Available at: https://el.wikipedia.org/wiki/Σύστημα_διαχείρισης_βάσης_δεδομένων. [23 February 2016].

‘Σχεσιακή βάση δεδομένων’ 2017, Wikipedia, wiki. Available at: https://el.wikipedia.org/wiki/Σχεσιακή_βάση_δεδομένων. [3 March 2016].

BusinessDictionary.com. 2017. What is relational database? definition and meaning - BusinessDictionary.com. Available at: http://www.businessdictionary.com/definition/relational-database.html. [5 March 2016].

‘Relational database’ 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/Relational_database. [5 March 2016].

‘Relational database management system’ 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/Relational_database_management_system. [5 March 2016].

Twinkle, 2017, What Is DBMS? What Is RDBMS? DBMS Vs RDBMS. Available from: http://www.mytecbits.com/microsoft/sql-server/what-is-dbms-what-is-rdbms. [11 March 2016].

DB-Engines 2017, DB-Engines Ranking of Relational DBMS. Available from: https://db-engines.com/en/ranking/relational+dbms. [11 March 2016].

Palace, 1996, Data Mining. Available from: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm. [13 March 2016].

Rouse, 2017, Business Intelligence (BI). Available from: http://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI. [13 March 2016].

‘Oracle Database’ 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/Oracle_Database. [13 March 2016].

DB-Engines 2017, Oracle System Properties. Available from: https://db-engines.com/en/system/Oracle. [13 March 2016].

Seika, 2013, Plug into the cloud. Available from: https://blogs.oracle.com/opnbenelux/entry/plug_into_the_cloud_with. [13 March 2016].

Cyran 2005, Oracle® Database. Available from: https://docs.oracle.com/cd/B19306_01/server.102/b14220/intro.htm#i62358. [14 March 2016].

Ashdown & Kyte 2017, Oracle® Database, Available from: https://docs.oracle.com/database/121/CNCPT/intro.htm#CNCPT88786. [17 March 2016].

Oracle n.d., Oracle Advanced Analytics. Available from: http://www.oracle.com/technetwork/database/options/advanced-analytics/overview/index.html. [24 March 2016].

Oracle n.d., Oracle Data Mining. Available from: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html. [24 March 2016].

'MySQL' 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/MySQL. [26 March 2016].

MySQL n.d., MySQL. Available from: https://www.mysql.com/. [26 March 2016].

DB-Engines 2017, MySQL System Properties. Available from: http://db-engines.com/en/system/MySQL. [26 March 2016].

Oracle and/or its affiliates 1997, 2017, MySQL 5.7 Reference Manual. Available from:

http://dev.mysql.com/doc/refman/5.7/en/features.html. [28 March 2016].

'Java Database Connectivity' 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/Java_Database_Connectivity. [10 April 2016].

'Open Database Connectivity' 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/Open_Database_Connectivity. [10 April 2016].

Auza, 2010, 5 of the Best Free and Open Source Data Mining Software. Available from: http://www.junauza.com/2010/11/free-data-mining-software.html. [12 April 2016].

Butler Analytics 2014, 10+ MySQL Reporting Tools. Available from: http://www.butleranalytics.com/10-mysql-reporting-tools/. [16 April 2016].

'Microsoft SQL Server' 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/Microsoft_SQL_Server#SQL_Server_2016. [18 April 2016].

DB-Engines 2017, Microsoft SQL Server System Properties. Available from: https://db-engines.com/en/system/Microsoft+SQL+Server. [18 April 2016].

Microsoft 2017, Data Mining Tools. Available from: https://msdn.microsoft.com/en-us/library/ms174467.aspx. [19 April 2016].

Microsoft 2016, Install SQL Server Business Intelligence Features. Available from: https://msdn.microsoft.com/en-us/library/hh231681.aspx. [26 April 2016].

Microsoft 2017, What is Analysis Services?. Available from: https://msdn.microsoft.com/en-us/library/bb522607.aspx. [26 April 2016].

Microsoft 2017, SQL Server Integration Services. Available from: https://msdn.microsoft.com/en-us/library/ms141026.aspx. [26 April 2016].

Microsoft 2017, Master Data Services Installation and Configuration. Available from: https://msdn.microsoft.com/en-us/library/ee633763.aspx. [26 April 2016].

Microsoft 2017, What is SQL Server Reporting Services (SSRS)?. Available from: https://msdn.microsoft.com/en-us/library/ms159106.aspx. [26 April 2016].

Microsoft 2016., Introducing Business Intelligence Development Studio. Available from: https://msdn.microsoft.com/en-us/library/ms173767.aspx. [28 April 2016].

‘Business Intelligence Development Studio’ 2016, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/Business_Intelligence_Development_Studio. [28 April 2016].

Barley n.d., Business Intelligence Development Studio (BIDS). Available from: https://www.mssqltips.com/sqlservertutorial/204/business-intelligence-development-studio-bids/. [28 April 2016].

PostgreSQL 2017, About. Available from: https://www.postgresql.org/about/. [2 May 2016].

DB-Engines 2017, PostgreSQL System Properties. Available from: https://db-engines.com/en/system/PostgreSQL. [2 May 2016].

PostgreSQL 2017, Software Catalogue - Reporting tools. Available from: https://www.postgresql.org/download/products/5-reporting-tools/. [4 May 2016].

Mullins, 2016, IBM DB2 relational DBMS overview. Available from: http://searchdatamanagement.techtarget.com/feature/IBM-DB2-relational-DBMS-overview. [6 May 2016].

‘IBM Db2’ 2017, Wikipedia, wiki. Available at: https://en.wikipedia.org/wiki/IBM_Db2. [6 May 2016].

Turorials Point 2017, DB2 – Introduction. Available from: http://www.tutorialspoint.com/db2/db2_introduction.htm. [6 May 2016].

DB-Engines 2017, DB2 System Properties. Available from: https://db-engines.com/en/system/DB2. [6 May 2016].

IBM n.d., IBM DB2 10.5 for Linux, Unix and Windows documentation. Available from: http://www.ibm.com/support/knowledgecenter/SSEPGG_10.5.0/com.ibm.db2.luw.kc.doc/welcome.html?lang=en. [7 May 2016].

Dietterich, T.G. 2001, "Ensemble methods in machine learning", Multiple Classifier Systems, volume 1857, 1–15.

Wolpert, D.W. 1992, "On the connection between in-sample testing and generalization error", Complex System, volume 6, 47–94.

Schaffer, C.S. 1994, "A conservation law of generalization performance", In the Proceedings of the 11th International Conference on Machine Learning (ICML), 259–265.

Κωτσιαντής, Σ.Β.Κ. 2005, Ομάδες ταξινομητών για την αύξηση της ακρίβειας των μεθόδων μηχανικής μάθησης και εξόρυξης γνώσης, Διδακτορική Διατριβή, Πανεπιστήμιο Πατρών.

I. Witten & E. Frank, 2005, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Elsevier, Morgan Kaufmann Publishers.

Murthy, 1908, "Automatic construction of decision trees from data: A multidisciplinary survey", Data Mining and Knowledge Discovery, volume 2, 345–389.

Quinlan, J.R.Q 1986, "Induction of decision trees", Machine Learning, volume 1, No. 1, 81–106.

Quinlan, J.R.Q 1993, C4.5: Programs for machine learning, Morgan Kaufmann, San Francisco.

TjenSien Lim, Wei Yin Loh, & Yu Shan Shih, 2000, "A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms", Machine Learning, volume 40, 203–228.

Y. Wang & I. Witten, 1997, "Induction of model trees for predicting continuous classes", In the Proceedings of the Poster Papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague, 128–137.

Lavrac, N.L. 1999, "Rule evaluation measures: A unifying view", Lecture Notes in Computer Science, volume 1634, 174–185.

A. An & N. Cercone, 2001, “Rule quality measures for rule induction systems: Description and evaluation”, Computational Intelligence, volume 17, 409–424.

Furnkranz, J.F. 1997, “Pruning algorithms for rule learning”, Machine Learning, volume 27, 139–171.

Bramer, M.B. 2002, “Using Jpruning to reduce overfitting of classification rules in noisy domains”, Lecture Notes in Computer Science, volume 2453, 433–442.

Furnkranz, J.F. 1999, “Separate-and-conquer rule learning”, Artificial Intelligence Review, volume 13, 3–54.

Mitchell, T.M.M. 1997, Machine Learning, McGrawHill.

Lewis, D.D.L. 1998, “Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval”, In the Proceedings of the ECML98, 10th European Conference on Machine Learning, Chemnitz, Germany, 4–15.

B. Becker, R. Kohavi & R. Sommerfield, 1997, “Visualizing the simple Bayesian classifier”, KDD97 Workshop on Issues in the Integration of Data Mining and Data Visualization.

J.L. Hodges & E.L. Lehmann, 1962, Rank methods for combination of independent experiments in analysis of variance. The Annals of Mathematical Statistics, 33(2):482-497.

Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Ελλάδας

Τμήμα Μηχανικών Πληροφορικής Τ.Ε.