Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Εξόρυξη δεδομένων σε Ανοιχτά Διασυνδεδεμένα Δεδομένα

Χαραλαμπόπουλος Αθανάσιος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Εφαρμοσμένης Στατιστικής

Πειραιάς
Δεκέμβριος 2015

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμό. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής Ν. Πελέκης (Επιβλέπων)
- Καθηγητής Ι.Θεοδωρίδης
- Επίκουρος Καθηγητής Ελ.Κοφίδης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα

U N I V E R S I T Y OF P I R A E U S

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Data Mining in Open Linked Data

By

Charalampopoulos Athanasios

MSc Dissertation

Submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
December 2015

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ.Νίκο Πελέκη που μου έδωσε τη δυνατότητα να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα και για την συνεργασία που είχαμε όλο αυτό το διάστημα. Επίσης, θα ήθελα να ευχαριστήσω θερμά την οικογένεια μου για την αμέριστη συμπαράσταση και υποστήριξη τους κατά την διάρκεια των σπουδών μου.

Περίληψη

Ο σκοπός της παρούσας διπλωματικής είναι διττός. Από τη μια πλευρά, σκοπός μας είναι να εκμεταλλευτούμε τη δυνατότητα των ανοιχτών διασυνδεδεμένων δεδομένων και τις πληροφορίες που αυτά παρέχουν. Από την άλλη η εφαρμογή τεχνικών εξόρυξης δεδομένων σε διασυνδεδεμένα δεδομένα με σκοπό την ανακάλυψη γνώσης η οποία κρύβεται σε αυτά.

Η ολοένα και αυξανόμενη χρήση του διαδικτύου, έχει καταστήσει τον Παγκόσμιο Ιστό, τη μεγαλύτερη αποθήκη δεδομένων και πληροφοριών. Η συνεισφορά των ανοιχτών διασυνδεδεμένων δεδομένων είναι η δημοσίευση και διασύνδεση δομημένων πληροφοριών στον Ιστό, έτσι ώστε αυτές να γίνονται κατανοητές από τις υπολογιστικές μηχανές μέσω του Σημασιολογικού Ιστού. Τα δεδομένα αυτά αναπαρίστανται μέσω του σχήματος RDF (Resource Description Framework) και της SPARQL γλώσσας για την δυνατότητα αναζητήσεων σε RDF δεδομένα στον Σημασιολογικό Ιστό.

Η ραγδαία εξέλιξη των διασυνδεδεμένων δεδομένων και η χρησιμότητα τους, ώθησε κυβερνήσεις, δημόσιους φορείς, μουσεία, εγκυκλοπαίδειες, βιβλιοθήκες κ.α. να συμμετέχουν στο εγχείρημα αυτό. Παράδειγμα αποτελεί το DBpedia, ένα project για την διασύνδεση και επαναχρησιμοποίηση δομημένης πληροφορίας από την Wikipedia κάτω από τις αρχές των ανοιχτών διασυνδεδεμένων δεδομένων. Μέσω ερωτημάτων SPARQL στην DBpedia, εξάγαμε πληροφορίες για 2000 ταινίες με σκοπό να εφαρμόσουμε τεχνικές εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων με τη χρήση αλγορίθμων οι οποίοι βασίζονται στη στατιστική και μηχανική μάθηση, μας δίνουν την δυνατότητα να αναλύσουμε και να επεξεργαστούμε μεγάλες βάσεις δεδομένων με σκοπό να εξάγουμε χρήσιμες πληροφορίες από αυτά. Συγκεκριμένα, εφαρμόσαμε τεχνικές κατηγοριοποίησης με σκοπό την ταξινόμηση μιας ταινίας σε “καλή” ή “κακή”, βασιζόμενοι στα χαρακτηριστικά των ταινιών τα οποία συλλέξαμε. Στη συνέχεια, μελετήσαμε, παραμετροποιήσαμε κατάλληλα και αξιολογήσαμε αλγορίθμους κατηγοριοποίησης.

Abstract

The purpose of the present thesis is twofold. On the one hand, our goal is to take advantage of the potential of the linked open data and of the information that they could provide. On the other hand the application of data mining techniques on linked data with a view to discover the hidden knowledge in them.

The ever increasing use of Internet, has without doubt converted the World Wide Web into the largest data and information storage. The contribution of the linked open data is the link and the publication of structured information on the Web, so that they can be understood by the computational engines via the Semantic Web. These data are represented by the RDF schema (Resource Description Framework) and SPARQL language for the searchable data in RDF in the Semantic Web.

The rapid development of the linked data and their usefulness, has urged governments, public institutions, museums, encyclopedias, libraries etc. to participate in this endeavor. One example is the DBpedia, a project about linking and reusing structured information through Wikipedia under the principles of the linked open data. Through SPARQL queries on DBpedia, we extracted information about 2000 films in order to apply data mining techniques.

Data mining using algorithms which are based on statistical and machine learning enable us to analyze and process large databases in order to extract useful information from them. More specifically, we implemented categorization techniques to classify a film as "good" or "bad" based on the film characteristics that we collected. Then, we studied, customized and appropriately evaluated classification algorithms.

Έχω διαβάσει και κατανοήσει τους κανόνες του ΠΜΣ που περιέχονται στον Οδηγό Συγγραφής ΔΕ και ιδιαίτερα όσα συνιστούν λογοκλοπή. Δηλώνω ότι η παρούσα διπλωματική εργασία αποτελεί προϊόν αποκλειστικά δικής μου προσπάθειας, υπό την καθοδήγηση του επιβλέποντος καθηγητή, ενώ για όλες τις πηγές που χρησιμοποιήθηκαν περιλαμβάνονται οι αντίστοιχες αναφορές.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Πινάκων

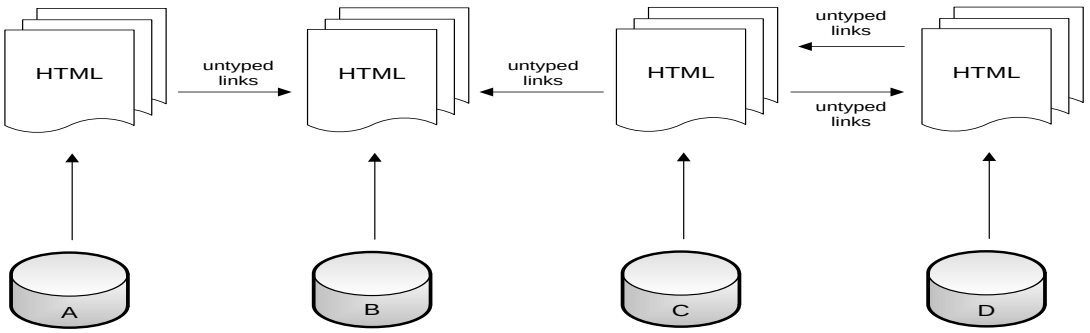
Κατάλογος Σχημάτων

ΚΕΦΑΛΑΙΟ 1

Σημασιολογικός Ιστός

1.1 Εισαγωγή

Η ραγδαία αύξηση της χρήσης του Παγκόσμιου Ιστού, τον έχει μετατρέψει στη μεγαλύτερη πηγή δημοσίευσης πληροφοριών. Κάθε στιγμή ένα τεράστιο πλήθος πληροφοριών, δημοσιεύονται σε ιστοσελίδες. Οι πληροφορίες αυτές είναι ετερογενής, διασκορπισμένες και αλληλοκαλυπτόμενες. Αυτό έχει σαν αποτέλεσμα να δυσχεραίνεται η αναζήτηση τους ενώ ταυτόχρονα ένα μεγάλος όγκος πληροφοριών παραμένει ανεκμετάλλευτος. Βασικό κίνητρο είναι η εκμετάλλευση του τεράστιου όγκου πληροφοριών παράγοντας γνώση η οποία βρίσκεται μέσα σε αυτές. Το βασικότερο αντικείμενο πληροφορίας στο διαδίκτυο είναι η ιστοσελίδα. Η πρόσβαση στις ιστοσελίδες γίνεται από τους χρήστες μέσω φυλλομετρητών (browser). Στο διαδίκτυο (internet), τα μη επεξεργασμένα δεδομένα υπάρχουν αποθηκευμένα σε διαφορετικές βάσεις δεδομένων και τοποθετούνται στο διαδίκτυο μέσα από τις σελίδες HTML. Ο Παγκόσμιος Ιστός αποτελείται από έγγραφα και τους συνδέσμους μεταξύ αυτών. Τα έγγραφα αυτά είναι ημίδομημένα και η σημασιολογία του περιεχομένου των σελίδων και των συνδέσμων δεν είναι φανερή. Έτσι, ο Ιστός μπορεί να γίνει αντιληπτός από τους ανθρώπους αλλά όχι από τις μηχανές. Στο παρακάτω σχήμα, παρουσιάζεται η μορφή του Ιστού όπως τον γνωρίζουμε μέχρι σήμερα, ο οποίος ονομάζεται και Ιστός των εγγράφων [A1]



Σχήμα 1.1: Ο Ιστός των Εγγράφων

Αυτή είναι και η κύρια διαφορά των συνδεδεμένων δεδομένων. Η πληροφορία να μπορεί να γίνεται αντιληπτή από τις μηχανές. Το 2006 ο Tim Berners-Lee παρουσίασε ένα σημείωμα [A2] όπου εμφανίζει τους κανόνες που χρειάζεται να διέπουν τα δεδομένα ώστε ο Παγκόσμιος Ιστός να μετατραπεί από τον Ιστό των Εγγράφων στον Ιστό των Συνδεδεμένων Δεδομένων :

1. Πρώτα όλοι οι πόροι χρειάζεται να αντιστοιχηθούν σε ένα URI όπου θα τους αντιπροσωπεύει. Ο κανόνας αυτός προϋποθέτει ότι όλα τα έγγραφα στον Ιστό, όπως ένα πρόσωπο ή ένα αντικείμενο έχει την δυνατότητα να ταυτοποιηθεί από ένα συγκεκριμένο URI (Uniform Resource Identifier – Καθολικό Αναγνωριστικό Πόρου) και με την σειρά του να μπει στον Ιστό των Δεδομένων. Επίσης, το URL (Uniform Resource Locator – Καθολικός Εντοπιστής Πόρου) που συναντάμε στον γνωστό Ιστό είναι μία συγκεκριμένη κατηγορία URI που χαρακτηρίζει ένα έγγραφο.

2. Επίσης, όλα τα URIs χρειάζεται να είναι HTTP URIs, με σκοπό ο κόσμος να έχει την δυνατότητα να πάρει δεδομένα για τους πόρους που αυτά αντιπροσωπεύουν λειτουργώντας το πρωτόκολλο HTTP. Ακόμα ο δεύτερος κανόνας παρουσιάζει ότι χρειάζεται να λειτουργούν HTTP URIs, με σκοπό τα URIs να είναι dereferenceable μέσω του πρωτοκόλλου HTTP. Dereferenceable URI ονομάζεται ένα URI όπου αν χρησιμοποιηθεί από ένα άτομο (π.χ. μέσω ενός browser), του παρέχει πληροφορίες που έχουν να κάνουν με το αντικείμενο που ταυτοποιεί.

3. Στην περίπτωση όπου κάποιος αναζητά ένα URI πρέπει να του παρέχονται σωστές πληροφορίες, βάσει των προτύπων RDF, RDF Schema, SPARQL. Ο τρίτος κανόνας συμπληρώνει τον δεύτερο και οδηγεί τους παρόχους συνδεδεμένων δεδομένων να δώσουν τις πληροφορίες που έχουν να κάνουν με τα δεδομένα τους. Το 2009, ο Berners-Lee τροποποίησε τον κανόνα αυτό, αναφέροντας ότι είναι απαραίτητη η χρήση των RDF.

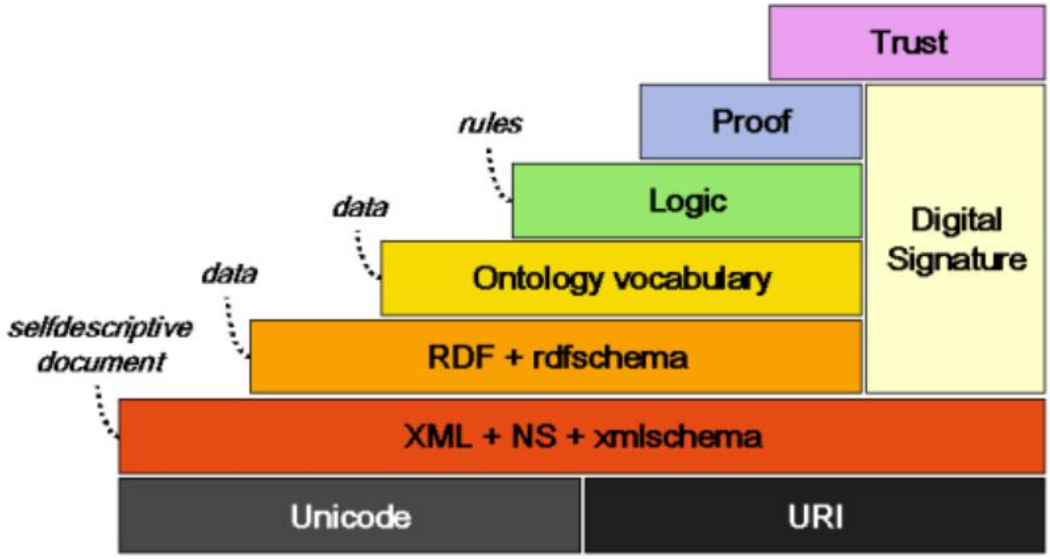
4. Με σκοπό την σωστή περιγραφή των URIs πρέπει να περιλαμβάνονται συνδέσεις σε άλλα URIs, με στόχο το άτομο να έχει την δυνατότητα να βρει ακόμα περισσότερες πληροφορίες. Ο τέταρτος κανόνας παρουσιάζει τις συνδέσεις ανάμεσα στα αντικείμενα. Όπως στον γνωστό Ιστό, μία σελίδα χρειάζεται να έχει συνδέσεις με άλλες σελίδες που περιέχουν παρόμοιες πληροφορίες, έτσι και στον Ιστό των

Συνδεδεμένων Δεδομένων ένας πόρος πρέπει να ενώνεται με άλλους πόρους του Ιστού με σκοπό να παρέχεται η δυνατότητα σε ένα άτομο να περιηγηθεί και να μπορέσει να βρει ακόμα περισσότερες πληροφορίες.

1.2 Δομή του Σημασιολογικού Ιστού

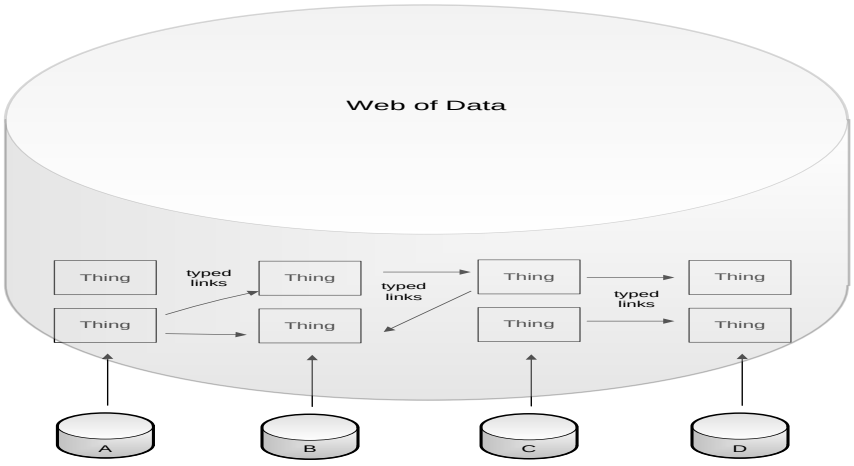
Ο όρος σημασιολογικός ιστός σχεδιάστηκε αρχικά από τον Tim-Berners Lee, εφευρέτη του παγκόσμιου ιστού. Αυτός ο όρος χρησιμοποιείται για να περιγράψει ένα διασυνδεδεμένο "ιστό δεδομένων" ο οποίος μπορεί να υλοποιηθεί μέσω υπολογιστικών μηχανών. Επί του παρόντος, η πρόσβαση στα δεδομένα του ιστού γίνεται μέσω ιστοσελίδων, δηλαδή HTML αρχείων τα οποία είναι συνδεδεμένα με υπερσυνδέσμους. Βασική ιδέα του ιστού είναι να παρέχει το πλαίσιο το οποίο θα επιτρέπει τα δεδομένα να επεξεργάζονται και να επαναχρησιμοποιούνται.

Ο Σημασιολογικός Ιστός αποτελείται από ένα σύνολο επιπέδων. Στο σχήμα 1.2 που ακολουθεί, φαίνεται ότι αποτελείται από 7 στρωματοποιημένα επίπεδα [A6]. Στο πρώτο επίπεδο, ο Σημασιολογικός Ιστός βασίζεται στα URIs και στα Unicode για την ονομασία και τον εντοπισμό των αντικειμένων στο Σημασιολογικό Ιστό. Το δεύτερο επίπεδο, βασίζεται στην XML, μια γλώσσα η οποία επιτρέπει τη χρησιμοποίηση δομημένων εγγράφων με ένα ορισμένο από το χρήστη λεξιλόγιο ώστε να ενσωματωθούν οι έννοιες του Σημασιολογικού Ιστού. Στο επόμενο επίπεδο, το RDF και το rdfschema συμβάλλουν στην γραφή δηλώσεων για τα αντικείμενα και στην οργάνωση τους σε ιεραρχίες, όπως οι κλάσεις και οι ιδιότητες. Το επίπεδο Οντολογίας (Ontology) υποστηρίζει την χρήση και την εξέλιξη των λεξιλογίων αφού έτσι ορίζονται οι σχέσεις μεταξύ των διάφορων εννοιών. Τα 3 τελευταία επίπεδα βρίσκονται σε ερευνητικό στάδιο. Το επίπεδο της λογικής, αναφέρεται σε κανόνες και συμπεράσματα, βασισμένα στις οντολογίες που θα έχουν δομηθεί και τα οποία θα οδηγούν τις μηχανές σε "λογικές" αποφάσεις. Στο επίπεδο απόδειξης, οι κανόνες αυτοί θα εκτελούνται και μαζί με το τελευταίο επίπεδο εμπιστοσύνης θα αποδεικνύεται ο βαθμός στον οποίο οι πληροφορίες αυτές είναι αξιόπιστες.



Σχήμα 1.2: Στρωματική απεικόνιση του Σημασιολογικού Ιστού

Η διαφορά του Σημασιολογικού Ιστού σε σχέση με τον κλασικό ιστό, είναι ότι προσδίδει δομή στο νόημα του περιεχομένου των ιστοσελίδων επιτρέποντας σε πράκτορες λογισμικού (software agents) να πλοηγούνται στις ιστοσελίδες και να πραγματοποιούν εργασίες. Έτσι, ο Ιστός αλλάζει σε μία παγκόσμια βάση δεδομένων, η οποία αποτελείται από τα ίδια τα δεδομένα. Στο επόμενο σχήμα εμφανίζεται η μορφή του Σημασιολογικού Ιστού [A1] :



Σχήμα 1.3: Ο Ιστός των Δεδομένων

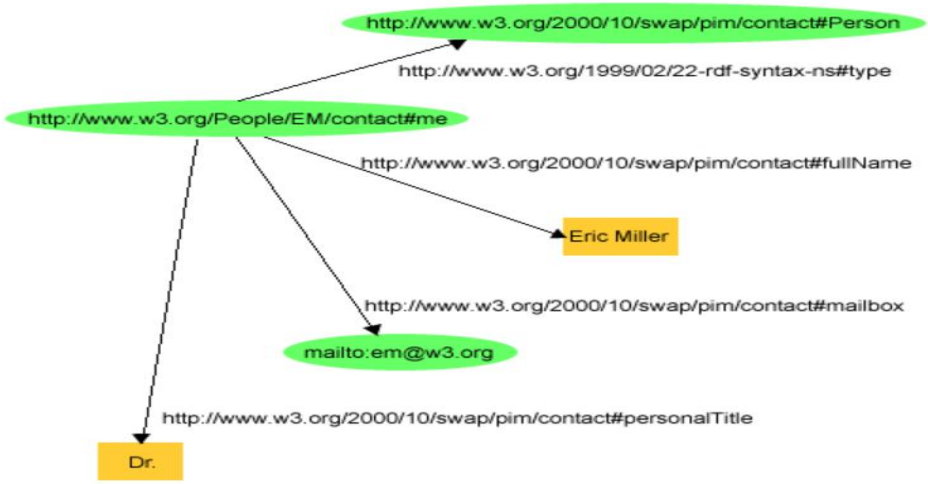
1.3 RDF

Βάση του σημασιολογικού ιστού, αποτελεί το RDF (Resource Description Framework), το οποίο είναι ουσιαστικά μια γλώσσα του W3C (World Wide Web) για την αναπαράσταση

πόρων στον ιστό. Αποτελεί επί της ουσίας έναν τρόπο μέσω του οποίου παρέχεται η δυνατότητα σε υπολογιστές να διαβάσουν και να κατανοήσουν πληροφορίες του ιστού. Η βασική μονάδα του RDF προτύπου είναι η πρόταση-statement. Ένα RDF statement παρουσιάζεται ως τριάδα (Σχήμα 1.4). Η τριάδα αποτελείται από τρία μέρη: το υποκείμενο, το κατηγόρημα και το αντικείμενο. Το υποκείμενο μπορεί να είναι πόρος (resource) που περιγράφεται από ένα IRI ή ανώνυμος – κενός κόμβος (blank node). Το κατηγόρημα είναι η ιδιότητα μέσω της οποίας συνδέεται το υποκείμενο με το αντικείμενο και περιγράφεται επίσης από ένα μοναδικό IRI. Το αντικείμενο αποτελεί την τιμή της ιδιότητας για το συγκεκριμένο υποκείμενο και μπορεί να είναι πόρος με συγκεκριμένο IRI, κενός κόμβος ή Literal. Τα Literals είναι τιμές διαφόρων τύπων δεδομένων όπως string, int, double, date κλπ και δεν περιγράφονται με IRI.

Σχήμα 1.4: RDF Statements

Πολλές RDF προτάσεις δημιουργούν έναν RDF γράφο στον οποίο οι κόμβοι αναπαριστούν τους πόρους (υποκείμενα, αντικείμενα) και οι ακμές τις μεταξύ τους σχέσεις (κατηγορήματα).



Σχήμα 1.5: Παράδειγμα αναπαράστασης RDF Statements σε RDF Graph.

1.4 RDF Schema

Το RDF Schema (RDFS) είναι ένα πρότυπο του W3C, το οποίο αρχικά εκδόθηκε το 1998 και με την τελική του μορφή το 2004 με στόχο να προσθέσει επιπλέον σημασιολογία σε RDF προτάσεις. Συγκεκριμένα ορίζει ένα λεξιλόγιο που επιτρέπει την περιγραφή ομάδων από συναφείς πόρους καθώς και των μεταξύ τους σχέσεων. Οι πόροι μπορούν να οργανωθούν σε κλάσεις και όταν γίνει αυτό οι πόροι αναφέρονται ως στιγμιότυπα της κλάσης. Οι κλάσεις μπορούν να οργανωθούν και να διαβαθμιστούν ιεραρχικά. Για την περιγραφή των ιδιοτήτων που χαρακτηρίζουν κλάσεις ή αντικείμενα/πόρους και την οργάνωσή τους σε μια ιεραρχία ορίζονται οι ιδιότητες (properties – subproperties) του RDFS. Το RDFS εισάγει έναν πιο σαφή διαχωρισμό μεταξύ των δύο κύριων τύπων των στοιχείων που βρίσκονται σε μια οντολογία το t-box και το a-box. Το T-box είναι η πληροφορία που περιγράφει τις έννοιες του τομέα και τις σχέσεις. Το a-box απαρτίζεται από τα απτά άτομα που έχουν επιβεβαιωθεί εντός των δεδομένων. Το λεξιλόγιο του RDFS, οι κλάσεις και οι ιδιότητες περιγράφονται από IRIs. Το λεξιλόγιο του RDFS αποτελείται από προκαθορισμένους πόρους των οποίων το URI έχει σαν πρόθεμα http://www.w3.org/2000/01/rdf-schema#. Ακολουθούν οι κλάσεις των RDF Schema [A8]

- `rdfs:Resource`, η κλάση των πόρων
- `rdfs:Class`, η κλάση όλων των κλάσεων
- `rdfs:Literal`, η κλάση των λεκτικών (χαρακτήρες και αριθμοί)
- `rdfs:Datatype`, η κλάση τύπου των δεδομένων (υποκατηγορία της `Literal`)
- `rdfs:langString`, η κλάση της γλώσσας των χαρακτήρων
- `rdfs:HTML`, η κλάση των HTML λεκτικών
- `rdfs:XMLLiteral`, η κλάση των XML λεκτικών
- `rdf:Property`, η κλάση των RDF ιδιοτήτων.

1.5 OWL

Η OWL αποτελεί μια γλώσσα του σημασιολογικού ιστού με αυξημένες βασικές έννοιες που ενισχύουν την εκφραστικότητα των RDFS. Ο στόχος της OWL είναι ίδιος με εκείνο των RDFS, τον ορισμό οντολογιών οι οποίες περιλαμβάνουν τις κλάσεις, τις ιδιότητες και τις

σχέσεις για συγκεκριμένες εφαρμογές. Ωστόσο η OWL μπορεί να σχηματίσει πιο σύνθετες σχέσεις και ως εκ τούτου μπορούν να φτιαχτούν εφαρμογές με ενισχυμένες συλλογιστικές ικανότητες. Το λεξιλόγιο της OWL αποτελείται από προκαθορισμένους πόρους των οποίων το URI έχει σαν πρόθεμα http://www.w3.org/2002/07/owl# (συναντάται και ως owl:). Η OWL, για παράδειγμα, χρησιμοποιεί δικό της ορισμό κλάσεων (owl: Classes) και χωρίζει τις ιδιότητες σε αυτές που έχουν ως αντικείμενο πόρους (owl: ObjectProperty) και σε αυτές που έχουν κάποιο τύπο δεδομένων – literal (owl: DatatypeProperty). Οι ιδιότητες μπορούν να οριστούν σαν μεταβατικές (transitive), συμμετρικές (symmetric), αντίστροφες (inverse) κ.λπ.

1.6 SPARQL

Η SPARQL είναι μια γλώσσα αναζήτησης RDF και ένα πρωτόκολλο για την πρόσβαση σε RDF δεδομένα στον σημασιολογικό ιστό. Η SPARQL 1.0 έγινε ένα πρότυπο W3C το 2008 και αναθεωρήθηκε σε SPARQL 1.1 το 2013.

Η SPARQL 1.0 παρείχε 4 μορφές ερωτημάτων:

- SELECT – Εξάγει μη επεξεργασμένα δεδομένα για ένα δεδομένο ερώτημα.
- CONSTRUCT – Εξάγει μια έγκυρη δομή RDF για ένα δεδομένο ερώτημα.
- ASK – Επιστρέφει αποτέλεσμα με τη μορφή True/False σε ένα δεδομένο ερώτημα.
- DESCRIBE – Εξάγει RDF δεδομένα από μια πηγή, ωστόσο το περιεχόμενο του γραφήματος είναι αποτέλεσμα του επεξεργαστή του ερωτήματος και όχι το πραγματικό ερώτημα. Αυτό χρησιμοποιείται όταν ο εκτελεστής του ερωτήματος δεν γνωρίζει πολλά για την γραφική παράσταση των δεδομένων και χρειάζεται περισσότερες πληροφορίες.

Η SPARQL 1.1 περιλαμβάνει τη δυνατότητα τροποποίησης των RDF δεδομένων μέσα στην πηγή προέλευσης. Αυτό περιλαμβάνει την προσθήκη/αφαίρεση των τριάδων και τη δημιουργία/διαγραφή των γραφημάτων.

- INSERT DATA – Προσθέτει RDF τριάδες σε ένα RDF γράφημα. Δημιουργεί ένα γράφημα αν δεν υπάρχει ήδη.
- DELETE DATA – Αφαιρεί RDF τριάδες από ένα RDF γράφημα.

1.7 Πόροι και Διαπραγμάτευση Περιεχομένου

Αρχικά οι πόροι όπου υπάρχουν στον Ιστό των Δεδομένων διαχωρίζονται σε πληροφοριακούς (informational resources) και μη-πληροφοριακούς (non-informational resources). Πληροφοριακοί λέγονται οι πόροι, όπου τα κύρια χαρακτηριστικά τους έχουν την δυνατότητα να μεταφέρονται μέσω μηνυμάτων. Η πλειοψηφία των πόρων που συναντάμε στον Ιστό, όπως για παράδειγμα έγγραφα, εικόνες και άλλα αρχεία μέσων, είναι πληροφοριακοί πόροι. Οι πληροφοριακοί πόροι έχουν μία ή περισσότερες αναπαραστάσεις, όπου έχουν την δυνατότητα να χρησιμοποιηθούν λειτουργώντας το πρωτόκολλο HTTP. Αυτές οι αναπαραστάσεις μπορούν να μεταδοθούν μέσα σε κάποιο μήνυμα ικανοποιώντας την έννοια των πληροφοριακών πόρων. Οι υπόλοιποι πόροι λέγονται μη-πληροφοριακοί. Ουσιαστικά στην συγκεκριμένη κατηγορία πόρων βρίσκονται οι ιδέες, τα φυσικά αντικείμενα και ότι δεν υπάρχει στο χώρο πληροφορίας του Ιστού. Ακόμα οι πόροι αυτοί, σε σχέση με τους πληροφοριακούς, δεν έχουν κάποια συγκεκριμένη αναπαράσταση.

Ο μηχανισμός του πρωτοκόλλου HTTP είναι η διαπραγμάτευση περιεχομένου και έχει να κάνει με την αναπαράσταση διαφορετικών μορφών του ίδιου εγγράφου, σχετικά με αυτό που αναζητά ο καθένας. Στην περίπτωση που κάποιος ανατρέξει στο URI ενός πόρου, χρειάζεται να λάβει ως απάντηση πληροφορίες για τον συγκεκριμένο πόρο. Η συγκεκριμένη διαδικασία διαφέρει για τους πληροφοριακούς και τους μη-πληροφοριακούς πόρους [A10].

Στην περίπτωση των πληροφοριακών πόρων δεν συναντώνται δυσκολίες. Όταν ανατρέχουμε στο URI ενός πληροφοριακού πόρου, ο εξυπηρετητής του URI αυτού υλοποιεί μία καινούρια αναπαράσταση του πόρου και την επιστρέφει ως απάντηση. Επίσης, η συγκεκριμένη διαδικασία δεν έχει διάφορα από την επίσκεψη κάποιου σε μία σελίδα κανονικού Ιστού, με την αντιστοιχία ότι το URI του ζητούμενου πληροφοριακού πόρου είναι το URL της σελίδας. Έτσι, το dereferencing του URI ενός πληροφοριακού πόρου έχει την δυνατότητα να γίνει είτε άμεσα με την επιστροφή του εγγράφου που αντιστοιχεί στο ζητούμενο URI, είτε έμμεσα με την επιστροφή ενός άλλου εγγράφου που έχει καλύτερη αναπαράσταση ή με την ανακατεύθυνση σε ένα άλλο έγγραφο που ανταποκρίνεται καλύτερα στις απαιτήσεις του.

Τέλος ένας μη-πληροφοριακός πόρος είναι περισσότερο από ένα απλό έγγραφο, οπότε δεν έχει την δυνατότητα να σχετίζεται με το URL ενός εγγράφου. Όμως έχοντας το URI ενός μη-

πληροφοριακού πόρου πρέπει να έχουμε την δυνατότητα να ανακτήσουμε πληροφορίες για τον πόρο αυτό. Επομένως, προκύπτει το ερώτημα πως μπορούμε να ανακτήσουμε τα έγγραφα που περιγράφουν έναν μη-πληροφοριακό πόρο, έχοντας μόνο το URI του πόρου. Αυτό θα μπορούσε να πραγματοποιηθεί από μηχανές αναζήτησης όπως αυτές που γνωρίζουμε, όπου θα είχαμε την δυνατότητα να αναζητήσουμε για την περιγραφή ενός πόρου δίνοντας σαν είσοδο ένα URI. Επομένως για την αναζήτηση μη πληροφοριακών πόρων έχουμε την δυνατότητα να χρησιμοποιήσουμε τον ίδιο τον Ιστό.

ΚΕΦΑΛΑΙΟ 2

ΑΝΟΙΧΤΑ ΔΙΑΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ

2.1 Εισαγωγή

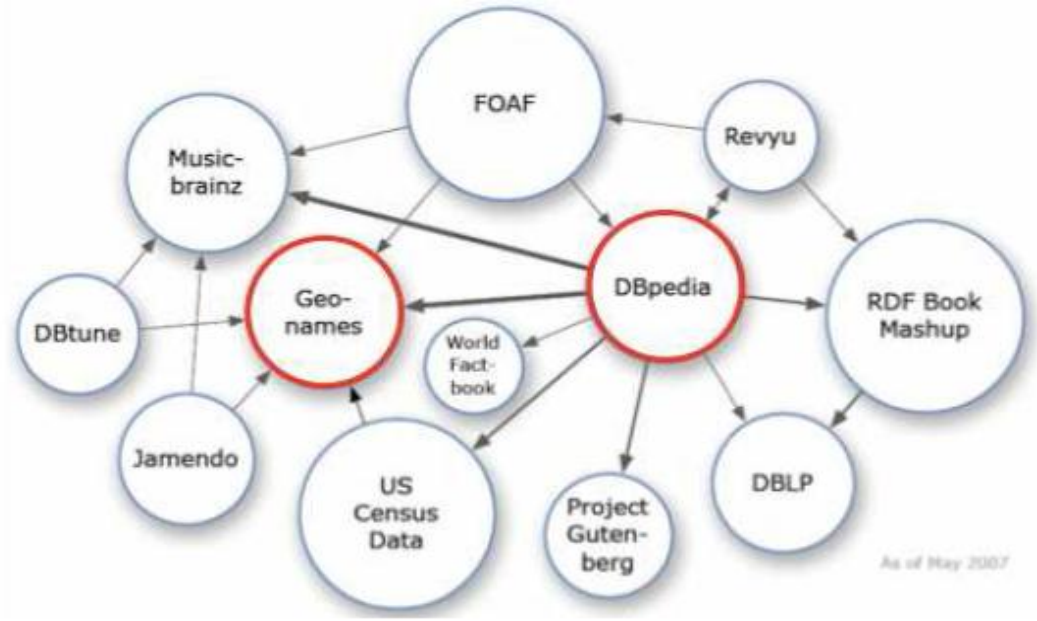
Ο Σημασιολογικός Ιστός δεν αφορά απλά την δημοσίευση δεδομένων στο διαδίκτυο. Αποσκοπεί σε κάτι πιο σημαντικό. Κύριος στόχος, είναι η δημιουργία συνδέσεων μεταξύ των δεδομένων έτσι ώστε όταν ένας άνθρωπος ή μια μηχανή θέλει, να μπορεί να εξερευνήσει το σύνολο των διαθέσιμων δεδομένων. Ανοιχτά διασυνδεδεμένα δεδομένα είναι τα συνδεδεμένα δεδομένα τα οποία δημοσιεύονται κάτω από μια ανοικτή άδεια ώστε να διευκολύνεται η ελεύθερη και δωρεάν επαναχρησιμοποίηση τους. Τον Φεβρουάριο του 2007, ξεκίνησε μια κοινοτική προσπάθεια για τη θεμελίωση και την εφαρμογή των κανόνων των Ανοιχτών Συνδεδεμένων Δεδομένων, μέσω του Linked Open Data project, που πραγματοποιήθηκε από τους Chris Bizer και Richard Cyganiak και είχε την υποστήριξη του W3C. Στόχος του project, είναι η επέκταση του Ιστού χρησιμοποιώντας τα ήδη διαθέσιμα δεδομένα κάτω από ανοιχτές άδειες μετατρέποντας τα σε RDF δεδομένα. Τον Οκτώβριο του 2007, μέσα σε λίγους μόνο μήνες τα συνδεδεμένα σύνολα δεδομένων αποτελούνταν από περισσότερες από 2 δισεκατομμύρια τριάδες, οι οποίες ήταν διασυνδεδεμένες με σχεδόν 2 εκατομμύρια συνδέσμους μεταξύ τους. Τον Σεπτέμβριο του 2011 και με την εξέλιξη του project ήταν διαθέσιμες 31 δισεκατομμύρια τριάδες και 504 εκατομμύρια σύνδεσμοι αντίστοιχα.

Αρχικά, το project στηρίχτηκε στην πλειοψηφία του από ερευνητές, πανεπιστημιακά ερευνητικά κέντρα και εταιρίες. Τα περισσότερα διασυνδεδεμένα σύνολα δεδομένων προέρχονταν στα πρώτα του βήματα από το:

- DBpedia
- Geonames
- Musicbrainz
- DBtune
- DBLP

- Revyu
- US census data
- RDF Book Mashup

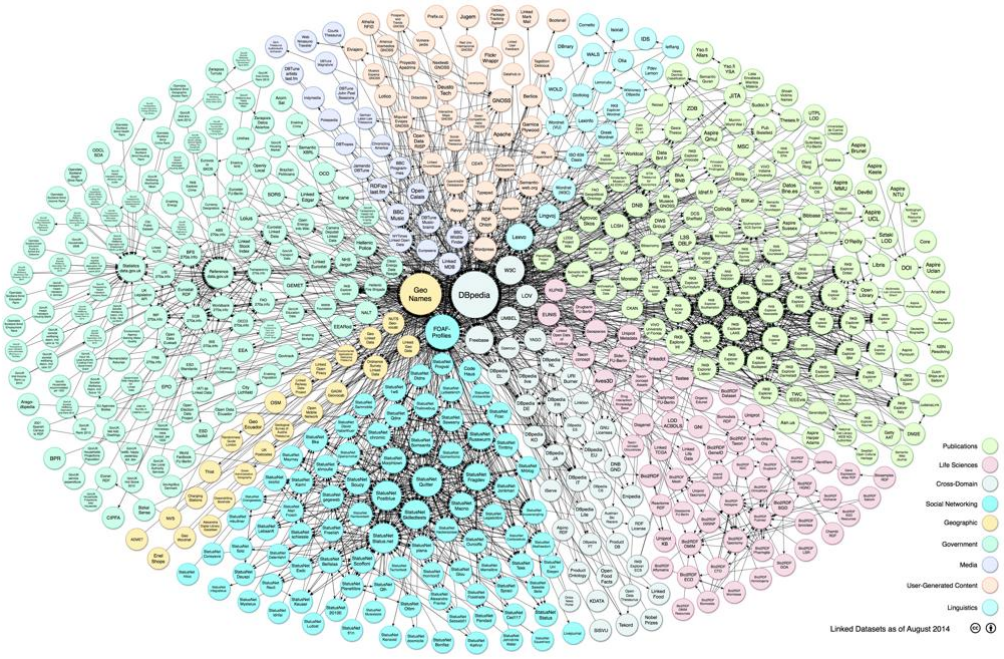
Όλα αυτά τα δεδομένα ενώνονταν με 120.00 RDF συνδέσμους όπου συνέδεαν μέσω των URI τα αντικείμενα από ένα σύνολο δεδομένων σε ένα άλλο σύνολο. Χρησιμοποιώντας τους συγκεκριμένος συνδέσμους, παρέχεται η δυνατότητα να πλοηγηθεί κάποιος από τη DBpedia στη βάση δεδομένων απογραφής των ΗΠΑ, από μια ταινία της DBpedia σε αξιολογήσεις του αντίστοιχου βιβλίου που δίνονται από το RDF Book Mashup ή από μια πόλη στη DBpedia σε περεταίρω πληροφορίες για αυτήν, στη γεωγραφική βάση δεδομένων του Geonames.



Σχήμα 2.1: Διάγραμμα-Σύννεφο Linking Open Data project, Μάιος 2007

Από τον Μάιο του 2007 και έκτοτε, το project εξελίχθηκε με γρήγορους ρυθμούς έχοντας και την βοήθεια μεγάλων οργανισμών, όπως το BBC, το Thomson Reuters και τη βιβλιοθήκη του Κογκρέσου [A11]. Η ταχεία αυτή εξέλιξη, υποστηρίζεται και από την φύση του project, αφού οποιοσδήποτε μπορεί να συμμετάσχει δημοσιεύοντας ένα σύνολο διασυνδεδεμένων δεδομένων αρκεί να πληρούν τους κανόνες που έχουν οριστεί. Η απήχηση και εξέλιξη των Ανοιχτών Διασυνδεδεμένων Δεδομένων απεικονίζεται στο σχήμα που ακολουθεί όπου παρατηρούμε το διάγραμμα-σύννεφο των Linked Data τον Αύγουστο του 2015. Από αυτό είναι φανερό το πόσο μεγάλο είναι το μέγεθος του Ιστού Δεδομένων αλλά και το πλήθος των

φορέων που συμμετέχουν σε αυτό. Όλοι οι κόμβοι στο διάγραμμα-σύννεφο παρουσιάζουν ένα διαφορετικό μέρος δεδομένων που δημοσιεύεται ως Συνδεδεμένα Δεδομένα [A15].



Σχήμα 2.2: Διάγραμμα-σύννεφο Linking Open Data, Αύγουστος 2014

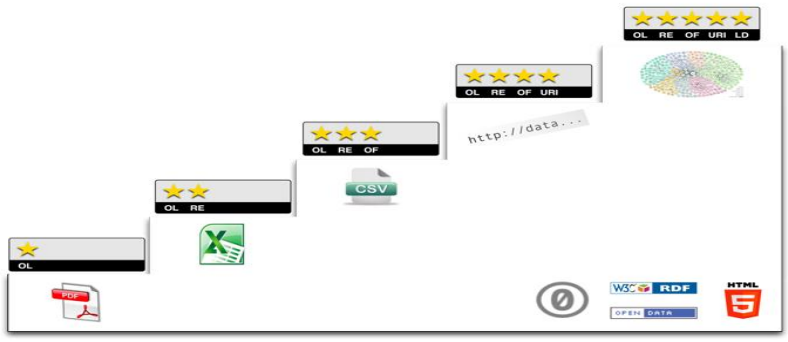
Τα βέλη στο Σχήμα 2.2 υποδηλώνουν την ύπαρξη συνδέσεων ανάμεσα στα δύο σύνολα δεδομένων. Τα αμφίδρομα βέλη φανερώνουν εξερχόμενους συνδέσμους και από τα δυο σύνολα δεδομένων και τα έντονα βέλη φανερώνουν την ύπαρξη περισσότερων συνδέσεων ανάμεσα σε δύο σύνολα δεδομένων. Όπως φαίνεται και από το σχήμα, το κέντρο του σύννεφου αποτελούν το DBpedia, GEOnames FOAFprofiles τα οποία περιέχουν και τους περισσότερους συνδέσμους με τα υπόλοιπα σύννεφα. Τα δεδομένα αυτά καλύπτουν ένα ευρύ φάσμα διαφορετικών τομέων από τα οποία το 18.05% Media, 21% Geographic, 9.47% Publications, 18.05% Government, 4.04% Cross Domain, 8.19% Life Sciences, 51.28% Social Networking. Έτσι λοιπόν, το περιεχόμενο του σύννεφου περιέχει πληροφορίες από γεωγραφικές τοποθεσίες, δημογραφικά χωρών και πόλεων, βιβλία, ταινίες, ιστορικά γεγονότα, στατιστικά δεδομένα, ιατρικά, φαρμακευτικά κ.α.

Ένα ζήτημα το οποίο προκύπτει από τη φύση των Ανοιχτών Διασυνδεδεμένων Δεδομένων είναι η κοινή χρήση οντολογιών και URI όταν αναφερόμαστε σε ίδια αντικείμενα μεταξύ του συνόλου των δεδομένων. Για παράδειγμα, όταν δεδομένα προέρχονται από διαφορετικούς εκδότες και η μια βάση χρησιμοποιεί τον όρο author ενώ η άλλη βάση τον όρο creator και

ενώ στην ουσία και οι δυο αναφέρονται στον συγγραφέα του βιβλίου. Η εισαγωγή λεξιλογίων συμβάλλει σε μεγάλο βαθμό στην ενοποίηση των δεδομένων. Για να γίνει αυτός ο διαχωρισμός είναι απαραίτητη η χρήση λεξιλογίων όπου θα δηλώνεται ότι αυτές οι δυο έννοιες είναι ταυτόσημες. Έως τώρα, ο αριθμός των ομάδων που χρησιμοποιούν το ίδιο URI για να αναφερθούν σε αντικείμενα είναι σχετικά μικρός σε σχέση με τις διαθέσιμες βάσεις δεδομένων. Από αυτό συμπεραίνουμε ότι αν θέλουμε να συνδέσουμε όλα τα σύνολα δεδομένων, είναι απαραίτητο να βρούμε τρόπους για να αποφασίζεται αν δύο αντικείμενα είναι ίδια [A12]. Αυτή τη στιγμή υπάρχουν διαθέσιμα λεξιλόγια στα ήδη υπάρχοντα συνδεδεμένα σύνολα με την χρήση κοινά συμφωνημένων λεξιλογίων όπως foaf, dc, dbpedia για την διαχείριση και ενσωμάτωση μεγάλων ανοιχτών συνόλων δεδομένων από βιβλιοθήκες, μουσεία, εφημερίδες κ.α.

2.2 Αξιολόγηση Linked Open Data

Όπως είδαμε και παραπάνω αρχικά υιοθετήθηκαν οι 4 κανόνες για τον τρόπο δημοσίευσης και διασύνδεσης δεδομένων τα οποία θα γίνονται αντιληπτά από τον άνθρωπο αλλά και από τις μηχανές με τη βοήθεια οντολογιών. Οι 4 αυτοί κανόνες αποτελούν γενικές κατευθυντήριες γραμμές αλλά κινούνται περισσότερο σε τεχνικές λεπτομέρειες. Λίγα χρόνια αργότερα, το 2010 ο Tim-Berners Lee πρόσθεσε την αξιολόγηση 5 αστέρων για τα ανοιχτά διασυνδεδεμένα δεδομένα. Στόχος δεν είναι να υποδείξει τον τρόπο με τον οποίο δημιουργούνται "καλά" ανοιχτά συνδεδεμένα δεδομένα αλλά να ενθαρρύνει κυρίως τις κυβερνήσεις αλλά και τους ιδιοκτήτες δεδομένων να δημοσιεύσουν τα δεδομένα τους σύμφωνα με τις αρχές των συνδεδεμένων δεδομένων. Σύμφωνα με την αξιολόγηση αυτή, πρότεινε το σχέδιο ανάπτυξης 5 αστέρων (5 star deployment scheme), μια βαθμωτή κλίμακα η οποία αντικατοπτρίζει την ποιότητα των δεδομένων.



Σχήμα 2.3: Απεικόνιση της αξιολόγησης των 5 αστέρων

Σύμφωνα με την αξιολόγηση αυτή [A17] έχουμε σε κάθε επίπεδο

☆ Διαθέστε τα δεδομένα σας (σε οποιαδήποτε μορφή) στον Παγκόσμιο Ιστό κάτω από μια ανοικτή άδεια

☆☆ Διαθέστε τα δεδομένα σας ως δομημένα δεδομένα (π.χ., Excel αντί για σάρωση εικόνας ενός πίνακα)

☆☆☆ Χρησιμοποιήστε μη ιδιόκτητες μορφές non-proprietary (π.χ. CSV αντί για Excel)

☆☆☆☆Χρησιμοποιήστε URIs (Uniform Resource Identifiers) για να ταυτοποιήσετε τα αντικείμενα σας, ώστε ο κόσμος να μπορεί να δείχνει τα δεδομένα σας

☆☆☆☆☆ Συνδέστε τα δεδομένα σας σε άλλα δεδομένα ώστε να παρέχετε το σημασιολογικό τους πλαίσιο.

Κάθε βήμα της απόκτησης ενός αστεριού συνοδεύεται από τα αντίστοιχα οφέλη αλλά και κόστη από την πλευρά του καταναλωτή (as consumer) και του εκδότη (as publisher).

Από την απόκτηση του 1^{ου} αστεριού ως καταναλωτής ωφελείται αφού μπορεί να δει, να τα εκτυπώσει, να τα αποθηκεύσει (σε ένα σκληρό δίσκο ή σε ένα USB), να τα τροποποιήσει, να τα εισάγει σε οποιοδήποτε άλλο σύστημα ακόμα και να τα μοιραστεί με όποιον επιθυμεί. Από την πλευρά του εκδότη, είναι απλό να δημοσιευτούν και δεν χρειάζεται να εξηγεί επανειλημμένως σε αυτούς που χρησιμοποιούν τα δεδομένα. Στο επίπεδο αυτό, τα δεδομένα μας είναι εγκλωβισμένα σε ένα έγγραφο και δεν μπορούν να εξαχθούν από αυτό παρά μόνο με τη βοήθεια ενός ειδικού προγράμματος ανίχνευσης. Είναι σημαντικό, τα δεδομένα να είναι διαθέσιμα στον Ιστό κάτω από μια άδεια όπως την PDDL ή την CCO.

Από την απόκτηση του 2^{ου} αστεριού ως καταναλωτής, διατηρεί τα οφέλη που έχει από την από την αξιολόγηση του 1^{ου} αστεριού και επιπλέον μπορεί απευθείας να επεξεργαστεί τα δεδομένα με το κατάλληλο λογισμικό ή να προχωρήσει σε υπολογισμούς ή οπτικοποίηση καθώς και να τα εξάγει σε μια άλλη δομημένη μορφή. Ως εκδότης παραμένει απλό να τα δημοσιεύσει. Σε αυτό το επίπεδο τα δεδομένα βρίσκονται εντός του Ιστού. Τα στοιχεία δεδομένων έχουν ένα URI και μπορούν να διαμοιρασθούν στον Ιστό. Τα δεδομένα αναπαρίστανται μέσω RDF.

Από την απόκτηση του 3ου αστεριού ως καταναλωτής διατηρεί τα οφέλη και επιπρόσθετα μπορεί να χειριστεί τα δεδομένα με οποιοδήποτε τρόπο επιθυμεί ανεξάρτητα από τις δυνατότητες του λογισμικού που χρησιμοποιεί. Από την πλευρά του εκδότη μπορεί να παραμένει απλή η δημοσίευση τους όμως μπορεί να χρειάζονται μετατροπείς ή πρόσθετα για να εξάγει τα δεδομένα από την ιδιόκτητη μορφή τους. Σε αυτό το επίπεδο, τα δεδομένα όχι απλά είναι διαθέσιμα στον Ιστό αλλά ο καθένας τώρα μπορεί να τα χρησιμοποιήσει εύκολα. Όμως, τα δεδομένα ακόμα βρίσκονται πάνω στον Ιστό αλλά δεν είναι δεδομένα μέσα στον Ιστό.

Από την απόκτηση του 4ου αστεριού μπορεί πλέον να συνδέσει τα δεδομένα του με οποιοδήποτε άλλο μέρος στο διαδίκτυο ή τοπικά. Μπορεί ακόμα να τα ορίσει στον σελιδοδείκτη, να επαναχρησιμοποιήσει τμήματα των δεδομένων καθώς και να χρησιμοποιήσει ήδη υπάρχοντα εργαλεία κι βιβλιοθήκες ακόμα και αν αυτά μπορούν να καταλάβουν ένα μέρος του σχεδίου (pattern) που έχει χρησιμοποιηθεί από τον εκδότη. Επίσης μπορεί με ασφάλεια να συνδυάσει τα δεδομένα με άλλα δεδομένα καθώς τα URIs είναι ένα παγκόσμιο σχήμα. Το κόστος για την απόκτηση του 4ου αστεριού είναι ότι τώρα χρειάζεται περισσότερη προσπάθεια η κατανόηση της δομής ενός RDF γραφήματος από ότι αν ήταν σε μορφή πίνακα (Excel/CSV) ή δέντρου (XML/JSON). Ως εκδότης μπορεί να έχει λεπτομερή έλεγχο των στοιχείων των δεδομένων και μπορεί να βελτιώσει την προσβασιμότητα τους ενώ παράλληλα και άλλοι εκδότες μπορούν να συνδεθούν με τα δεδομένα του. Όμως θα πρέπει να αφιερώσει κάποιο χρόνο ώστε να διαμερίσει και να συνθέσει τα δεδομένα. Ακόμη θα πρέπει να εκχωρήσει URIs στα στοιχεία των δεδομένων και να σκεφτεί το πώς αυτά θα αναπαρασταθούν ή να βρει είτε τα υπάρχοντα πρότυπα για την επαναχρησιμοποίηση ή να δημιουργήσει νέα. Στο επίπεδο αυτό, τα δεδομένα μας πια είναι μέσα στον Ιστό. Στα στοιχεία αντιστοιχεί ένα URI και μπορούν να διαμοιρασθούν στον Ιστό.

Από την απόκτηση του 5ου αστεριού μπορεί να ανακαλύψει ακόμη περισσότερα συναφή δεδομένα και μπορεί άμεσα να μάθει το σχήμα των δεδομένων του. Θα πρέπει να αντιμετωπίσει "σπασμένες" συνδέσεις δεδομένων (π.χ. λάθος ιστοσελίδας 404 errors page not found). Παρουσιάζοντας δεδομένα από τυχαία link εμπεριέχεται ο κίνδυνος να αφήσουν κάποιον να περιλαμβάνει περιεχόμενο από οποιοδήποτε διαδικτυακό τόπο στις σελίδες του. Ως εκδότης, τα δεδομένα τώρα είναι ανιχνεύσιμα ενώ ταυτόχρονα αυξάνεται και η αξία των δεδομένων του. Ωστόσο θα πρέπει να βρει πόρους ώστε να συνδέσει τα δεδομένα του με

άλλα δεδομένα στον Ιστό ενώ θα χρειαστεί να διορθώσει και τυχόν εσφαλμένες συνδέσεις. Στο επίπεδο αυτό, τα δεδομένα μας είναι μέσα στον Ιστό αλλά ταυτόχρονα είναι και συνδεδεμένα με άλλα δεδομένα.

Τόσο ως καταναλωτής όσο και ως εκδότης, έχοντας δεδομένα σύμφωνα με την αξιολόγηση των 5 αστέρων, επωφελούνται από το φαινόμενο του δικτύου. Ως φαινόμενο του δικτύου ορίζεται εκείνο το χαρακτηριστικό γνώρισμα κατά το οποίο η αξία ενός αντικειμένου εξαρτάται από τον αριθμό των ατόμων που το χρησιμοποιούν. Η χρηστική αξία των δεδομένων εξαρτάται έμεσα από τα πόσα άτομα μπορούν να χρησιμοποιήσουν τα δεδομένα μας. Αν το άτομο απλά κατέχει τα δεδομένα για τον εαυτό του, τότε η αξία τους είναι μηδενική. Με την διασύνδεση των δεδομένων του και όσο περισσότερα άτομα τα χρησιμοποιούν τόσο αυξάνει η αξία τους.

2.3 DBpedia

Η προσφορά της κοινότητα Dbpedia είναι ουσιαστικά να εξάγει πληροφορίες που βρίσκονται στη Wikipedia και να τις καθιστά ευρέως διαθέσιμες μέσω οργανωμένων προτύπων του σημασιολογικού ιστού και των διασυνδεδεμένων δεδομένων. Το σύνολο δεδομένων της DBpedia αποτελείται από RDF τριάδες που έχουν εξαχθεί από τα infoboxes που είναι τοποθετημένα στη δεξιά πλευρά των άρθρων της Wikipedia. Η Wikipedia είναι το έβδομο πιο δημοφιλές website, η πρώτη σε αναζητήσεις εγκυκλοπαίδεια και ίσως από τα καλύτερα παραδείγματα δημιουργίας συνεργατικών δεδομένων. Ωστόσο λόγο της δομή των άρθρων της δεν προσφέρει ιδιαίτερες δυνατότητες αναζήτησης και διερεύνησης. Ένα παράδειγμα αποτελεί η δυσκολία που παρουσιάζεται στην εύρεση όλων των ποταμών που εκβάλλουν στην μεσόγειο ή όλων των ελλήνων ποιητών του 20^{ου} αιώνα. Στόχος του project Dbpedia είναι να προσφέρει αυτές τις δυνατότητες αναζήτησης και διερεύνησης στο ευρύ κοινό, εξάγοντας δομημένες πληροφορίες από τη Wikipedia οι οποίες θα απαντούν σε σύνθετα ερωτήματα όπως αυτά του προηγούμενου παραδείγματος.

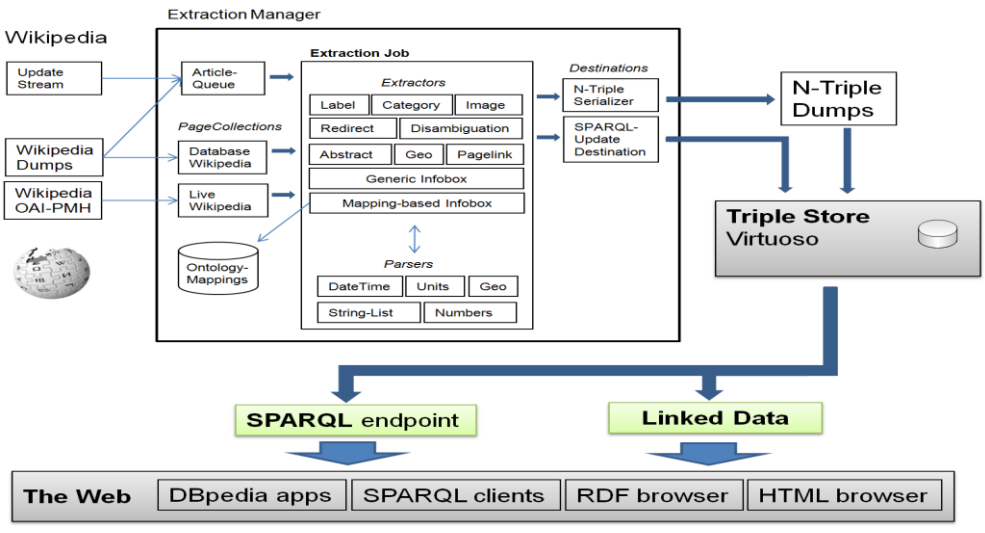
2.3.1 DBpedia project

Το project DBpedia ξεκίνησε το 2006 και έκτοτε έχει κεντρίσει το ενδιαφέρον ερευνών και πρακτικών. Έχει αποδειχθεί ένας πολύ σημαντικός παράγοντας για την επιτυχία των ανοιχτών διασυνδεδεμένων δεδομένων καθώς χρησιμεύει ως διασυνδετικός κόμβος για άλλα σύνολα δεδομένων. Για την ερευνητική κοινότητα προσφέρει μια πλατφόρμα δοκιμών για

πραγματικά δεδομένα που καλύπτουν διάφορους τομείς και πάνω από 100 εκδόσεις γλώσσας. Πλήθος εφαρμογών, αλγορίθμων και εργαλείων έχουν εφαρμοστεί ή κατασκευαστεί γύρω από τη DBpedia. Εξαιτίας της συνεχούς ανάπτυξης της Wikipedia αλλά και των βελτιώσεων που έχουν γίνει της DBpedia εξαγόμενα δεδομένα παρέχουν μια αυξανόμενη προστιθέμενη αξία στην απόκτηση δεδομένων, στην επαναχρησιμοποίηση και στις εργασίες ενσωμάτωσης μέσα στους οργανισμούς.

Τα άρθρα της Wikipedia αποτελούνται κυρίως από ελεύθερο κείμενο, ωστόσο περιέχουν επίσης διάφορους τύπους δομημένων πληροφοριών στην μορφή των wiki επισημάνσεων. Αυτές οι πληροφορίες περιλαμβάνουν πρότυπα infobox, πληροφορίες κατηγοριοποίησης, εικόνες, γεω-συντεταγμένες, συνδέσμους για εξωτερικές ιστοσελίδες, σελίδες αποσαφήνισης, ανακατευθύνσεις ανάμεσα σε σελίδες και σε συνδέσμους ανάμεσα σε όλες τις διαθέσιμες γλώσσες της Wikipedia. Το project DBpedia εξάγει αυτές τις δομημένες πληροφορίες και τις μετατρέπει σε μια πλούσια βάση δεδομένων. Το σχήμα 2.4 δίνει τη συνολική εικόνα του πλαισίου εξαγωγής δεδομένων. Τα βασικά στοιχεία του πλαισίου είναι:

- ✓ Page collections: τα οποία αντλούν τοπικές ή απομακρυσμένες πηγές από τα Wiki άρθρα.
- ✓ Destinations: αποθηκεύουν ή σειριοποιούν τις εξαγμένες RDF τριάδες.
- ✓ Extractors: μετατρέπουν ένα συγκεκριμένο τύπο wiki σήμανσης σε τριάδα.
- ✓ Parsers: στηρίζουν τους extractors με τον καθορισμό των τύπων των δεδομένων, μετατρέποντας τιμές μεταξύ μονάδων και διαχωριστικών σημάνσεων σε λίστες



Σχήμα 2.4: Αρχιτεκτονική του πλαισίου εξαγωγής δεδομένων της DBpedia

Το πλαίσιο αποτελείται από 11 extractors που επεξεργάζονται τους ακόλουθους τύπους περιεχομένων του Wikipedia:

- ✓ Labels: πρόκειται για τους τίτλους των άρθρων Wiki.
- ✓ Abstracts: μικρές ή μεγαλύτερες περιλήψεις (πρώτες σειρές, παράγραφος πριν από τα περιεχόμενα) μέσα από τα άρθρα Wiki.
- ✓ Interlanguage links: σύνδεσμοι οι οποίοι συνδέουν άρθρα που αφορούν το ίδιο θέμα σε διαφορετικές γλώσσες.
- ✓ Image: η πρώτη εικόνα μιας σελίδας Wikipedia.
- ✓ Redirects: συνώνυμες ή εναλλακτικές εκφράσεις των URIs της DBpedia (εκτός από συνώνυμα περιλαμβάνονται ακρώνυμα, συχνά ορθογραφικά λάθη κτλ)
- ✓ Disambiguation: δηλαδή όρους οι οποίοι μπορούν να περιγραφούν ή να συνδέονται εννοιολογικά με πολλά διαφορετικά URIs της DBpedia και γι'αυτό το λόγο υπάρχει μια ασάφεια ως προς το ποιο είναι το καταλληλότερο URI για να τους περιγράψει.
- ✓ Pagelinks: όλοι οι σύνδεσμοι μεταξύ των άρθρων από το Wikipedia.
- ✓ Homepages: οι σύνδεσμοι προς την αρχική σελίδα φορέων όπως είναι οργανισμοί ή εταιρίες.
- ✓ Geo-coordinates: όλες οι γεω-συντεταγμένες.

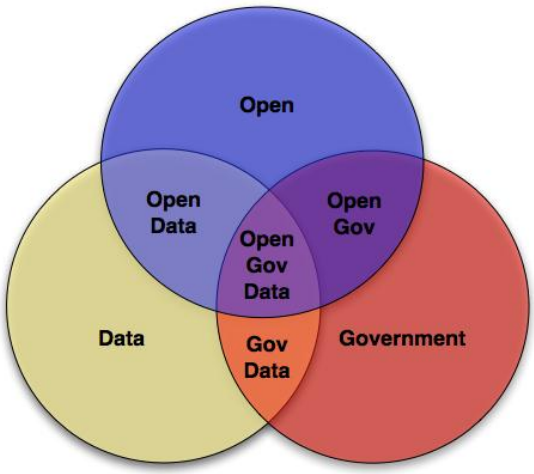
2.3.2 Η βάση δεδομένων της DBpedia

Η βάση δεδομένων της DBpedia αποτελείται πάνω από 274 εκατομμύρια RDF τριάδες, οι οποίες έχουν εξαχθεί από 35 διαφορετικές εκδόσεις της Wikipedia. Η βάση δεδομένων περιγράφει πάνω από 2,6 εκατομμύρια φορείς, περιέχει ακόμα labels και abstracts σε 30 διαφορετικές γλώσσες, 609 χιλιάδες συνδέσμους σε εικόνες και πάνω από 3,15 εκατομμύρια συνδέσμους σε εξωτερικές ιστοσελίδες. Η DBpedia αναγνωρίζει αγγλικούς τίτλους άρθρων. Οι πληροφορίες που βρίσκονται σε διαφορετικές γλώσσες αντιστοιχίζονται μέσω αμφίδρομης αξιολόγησης των διαγλωσσικών (interlingual) συνδέσμων μεταξύ των άρθρων Wikipedia. Οι πόροι συγκροτούν ένα URI μέσω του προτύπου http://dbpedia.org/resource/Name, όπου το Name έχει ληφθεί από τη διεύθυνση URL της πηγής Wikipedia, η οποία έχει τη μορφή: http://en.wikipedia.org/wiki/Name. Τα URIs της DBpedia καλύπτουν ένα μεγάλο εύρος εγκυκλοπαιδικών τίτλων. Υπάρχουν σαφείς πολιτικές σε ισχύ για τη διαχείρισή τους. Τα δεδομένα εντός της DBpedia συνδέονται με άλλα εξωτερικά σύνολα δεδομένων από τη βάση των ανοιχτών διασυνδεδεμένων δεδομένων.

Η DBpedia προσφέρει ένα τερματικό σημείο SPARQL (http://dbpedia.org/sparql), στο οποίο οι χρήστες-εφαρμογές μπορούν να αποστείλουν ερωτήματα για την ανάκτηση δεδομένων. Αυτός ο τύπος πρόσβασης καθίσταται περισσότερο κατάλληλος όταν ο κατασκευαστής της εφαρμογής που τρέχει τα ερωτήματα έχει πρότερη γνώση του τι πληροφορία θέλει να ανακτήσει. Για να προστατευθεί από υπερφόρτωση το τερματικό σημείο SPARQL υπάρχει ένα όριο στις πληροφορίες που μπορεί να ανακτηθούν για κάθε δεδομένο ερώτημα.

2.4 Ανοιχτά κυβερνητικά δεδομένα

Με τον όρο Κυβερνητικά Δεδομένα χαρακτηρίζονται οι πληροφορίες του Δημοσίου Τομέα (Public Sector Information - PSI) ενώ με τον όρο Ανοιχτά Κυβερνητικά Δεδομένα χαρακτηρίζεται η δημοσίευση των Κυβερνητικών Δεδομένων σε ανοιχτές, πρωτογενείς (raw) μορφές και με διαδικασίες που τις κάνουν προσιτές και άμεσα διαθέσιμες σε όλους και δίνουν την δυνατότητα επαναχρησιμοποίησης. Επομένως, Ανοιχτά Κυβερνητικά Δεδομένα (Open Government Data) είναι τα ανοιχτά δεδομένα που παράγονται από τις κυβερνήσεις ή από φορείς υπό την επίβλεψη των κυβερνήσεων, τα οποία μπορούν να χρησιμοποιηθούν, να επαναχρησιμοποιηθούν και να αναδιανεμηθούν από τον καθένα [A19].



Σχήμα 2.5: Απεικόνιση των Ανοιχτών Κυβερνητικών Δεδομένων

Για παράδειγμα, εμπεριέχονται επιλεγμένα αποθηκευμένα δεδομένα της Eurostat όπως επίσης και πολλές πληροφορίες του δημόσιου τομέα της βρετανικής κυβέρνησης. Η επιτυχής υλοποίηση των ελευθέρων κυβερνητικών δεδομένων στα ευρωπαϊκά κράτη δεν μπορεί να πραγματοποιηθεί αν απλά αντιγραφούν ξένα πρότυπα εκμοντερνισμού κρατών και διοίκησης. Αυτό συμβαίνει διότι εγείρονται συγκεκριμένοι περιορισμοί από τις αντιλήψεις, τις παραδόσεις και τις κουλτούρες του κάθε κράτους για την πρόσβαση του κοινού σε τέτοιου είδους πληροφορίες αλλά και από ζητήματα διαφάνειας. Ως εκ τούτου, ο κάθε φορέας διακυβέρνησης θα πρέπει να δημιουργεί τις δικές του ιδέες συμπληρώνοντας την έννοια της ανοιχτής πρόσβασης στα κυβερνητικά δεδομένα.

2.5 Αλλαγή μοντέλου για τα ανοιχτά κυβερνητικά δεδομένα

Είναι σημαντικό να υπάρξει μια διαμόρφωση της γνώμης των πολιτικών και διοικητικών φορέων καθώς χρειάζεται μια ρεαλιστική προσέγγιση στην διαχείριση των αποθηκευμένων δεδομένων. Τα κράτη και οι διοικήσεις βρίσκονται σε μια αλλαγή προτύπου, την εποχή των ανοιχτών κυβερνητικών δεδομένων. Η ελεύθερη πρόσβαση στα δεδομένα μπορεί να χρησιμοποιηθεί ως εργαλείο για το άνοιγμα και την επιρροή στις δομές, στις οργανωτικές αλυσίδες και τη διαδικασία λήψης αποφάσεων. Για την επίλυση των προβλημάτων που προκύπτουν θα πρέπει να υπάρξουν στο υπάρχων μοντέλο αλλαγές σε τρία διαφορετικά επίπεδα.

Στο πρώτο επίπεδο, επηρεάζεται η έννοια των δημόσιων και κρυφών δεδομένων. Μέχρι τώρα όλα τα δεδομένα είναι κρυφά εκτός από αυτά που είναι μαρκαρισμένα ως δημόσια. Στην αλλαγή αυτού του προτύπου όλα τα δεδομένα θα πρέπει να είναι δημόσια εκτός από αυτά που θα είναι μαρκαρισμένα ως κρυφά.

Όσον αφορά το δεύτερο επίπεδο επηρεάζεται το εύρος, ο τύπος και η χρονική στιγμή της δημοσίευσης δεδομένων. Το εύρος και η χρονική στιγμή, σύμφωνα με το παλιό πρότυπο καθορίζονταν από τη δημόσια αρχή. Στο σύγχρονο πρότυπο δεν υπάρχει προστασία απορρήτου των φακέλων και δημοσιεύονται ολόκληρα.

Η τρίτη αλλαγή αφορά την προστασία της χρήσης αυτών των δεδομένων. Σύμφωνα με το παλιό πρότυπο τα δεδομένα προορίζονταν για ιδιωτική χρήση καθώς οτιδήποτε περαιτέρω απαιτεί έγκριση κατόπιν αίτησης. Στο καινούριο πρότυπο τα δεδομένα αυτά είναι ελεύθερα για οποιαδήποτε χρήση ακόμα και διαφημιστική.

Είναι σαφές, ότι οι κυβερνήσεις έχουν δύο κύριες επιλογές στην κατεύθυνση της δημοσίευσης δεδομένων του Δημοσίου στο Διαδίκτυο. Η πρώτη είναι το να κατέχει η ίδια η κυβέρνηση υπηρεσίες πληροφόρησης κατευθείαν στους πολίτες, όπου λειτουργούν τα δημόσια δεδομένα, και που είναι πολύ συνηθισμένο σήμερα. Για παράδειγμα στο http://geodata.gov.gr οι πολίτες μπορούν να βρουν διαθέσιμα τα δρομολόγια της αστικής συγκοινωνίας για την Αθήνα. Τα δεδομένα περιλαμβάνουν τις στάσεις και τις διαδρομές για λεωφορεία, τρόλεϊ, μετρό, τραμ, ηλεκτρικό σιδηρόδρομο και προαστιακό. Αυτή είναι μια κατευθείαν υπηρεσία πληροφόρησης του Δημόσιου Τομέα. Από την άλλη, έχει τη δυνατότητα να δημοσιοποιεί στο Διαδίκτυο απλά πρωτογενή δεδομένα (raw data) με σκοπό να γίνει πιο εύκολη η προσβασιμότητα και η χρησιμότητα των κυβερνητικών δεδομένων. Οι τρεις κυριότεροι λόγοι για τους οποίους πρέπει τα κυβερνητικά δεδομένα να είναι ανοιχτά συνοψίζονται παρακάτω [A21]

1. Διαφάνεια

Σε μια καλά δομημένη και δημοκρατική κοινωνία οι πολίτες πρέπει να γνωρίζουν τι κάνει η κυβέρνηση τους. Για να γίνει αυτό, θα πρέπει να έχουν πρόσβαση στα κυβερνητικά στοιχεία ελεύθερα και να μπορούν να μοιράζονται αυτές τις πληροφορίες με άλλους πολίτες. Η διαφάνεια δεν αφορά μόνο την πρόσβαση σε αυτά αλλά και την ανταλλαγή και επαναχρησιμοποίηση τους ώστε να είναι σε θέση οι πολίτες να κατανοήσουν το υλικό. Για να επιτευχθεί η κατανόηση των δεδομένων, αρκετές φορές πρέπει να αναλυθούν ή να οπτικοποιηθούν και αυτό προϋποθέτει ότι τα στοιχεία αυτά είναι ανοικτά έτσι ώστε να μπορούν ελεύθερα να χρησιμοποιηθούν και να επαναχρησιμοποιηθούν.

2. Απελευθέρωση Κοινωνικής και Εμπορικής αξίας

Στην ψηφιακή εποχή, τα δεδομένα αποτελούν το βασικό κλειδί για τις κοινωνικές και εμπορικές δραστηριότητες. Από την αναζήτηση ενός καταστήματος των ΕΛΤΑ έως την κατασκευή μιας διαδικτυακής μηχανής αναζήτησης (building search engine) απαιτείται πρόσβαση σε δεδομένα, τα οποία δημιουργούνται ή διατηρούνται από τις κυβερνήσεις. Με το άνοιγμα των δεδομένων, η κυβέρνηση μπορεί να βοηθήσει στη δημιουργία καινοτόμων επιχειρήσεων και υπηρεσιών που θα παρέχουν κοινωνική και εμπορική αξία.

3 Συμμετοχική διακυβέρνηση

Με το άνοιγμα των δεδομένων, οι πολίτες έχουν τη δυνατότητα να ενημερώνονται και να συμμετέχουν άμεσα στη διαδικασία λήψης αποφάσεων. Αυτό είναι κάτι περισσότερο από ότι

η διαφάνεια, αυτό μπορεί να επιφέρει μια πλήρη "read/write" society, όχι μόνο για να γνωρίζει ο πολίτης τι συμβαίνει στη διαδικασία της διακυβέρνησης, αλλά να είναι σε θέση να συμβάλλει σε αυτή.

Υπάρχουν πολλά είδη ανοιχτών κυβερνητικών δεδομένων τα οποία μπορούν να φανούν χρήσιμα όπως

Πολιτιστικά δεδομένα σχετικά με έργα τέχνης ή πολιτιστικά έργα όπως τίτλους και συντάκτες τα οποία βρίσκονται σε βιβλιοθήκες πινακοθήκες ή μουσεία.

Επιστημονικά δεδομένα που παράγονται ως μέρος της επιστημονικής έρευνας από την αστρονομία έως και τη ζωολογία.

Οικονομικά δεδομένα όπως οι λογαριασμοί του Δημοσίου, δαπάνες ή έσοδα , και πληροφορίες σχετικά με τις χρηματοπιστωτικές αγορές (μετοχές, ομόλογα κ.α.)

Στατιστικά στοιχεία που παράγονται από τις αρμόδιες στατιστικές αρχές όπως η απογραφή του πληθυσμού ή ο υπολογισμός σημαντικών κοινωνικοοικονομικών δεικτών όπως ο πληθωρισμός, δείκτες ανεργίας ή φτώχειας ενός πληθυσμού.

Μετεωρολογικά διαφόρου τύπου δεδομένα τα οποία σχετίζονται με τον καιρό και τα οποία μπορούν να συμβάλλουν στην πρόβλεψη/ανάλυση του καιρού/της κλιματικής αλλαγής.

Περιβαλλοντικά δεδομένα τα οποία σχετίζονται με το φυσικό περιβάλλον, όπως το επίπεδο ρύπων σε αστικά κέντρα ή τα επίπεδα μόλυνσης ποταμών και θαλασσών.

Μεταφορών στοιχεία όπως χρονοδιαγράμματα, δρομολόγια δημοσίων συγκοινωνιών, ακόμα και ανάλυση τροχαίων ατυχημάτων σε πλήθος και αριθμό παθόντων.

Στη συνέχεια, και συγκεκριμένα το 2007, το Open Government Working Group όρισε έναν αριθμό κανόνων όπου χρειάζεται να διέπουν τη δημοσίευση κυβερνητικών δεδομένων έτσι ώστε να θεωρούνται ανοιχτά. Σύμφωνα με αυτούς τους κανόνες τα δεδομένα χρειάζεται να είναι [A22]:

1. Ολοκληρωμένα. Όλα τα δημόσια δεδομένα γίνονται διαθέσιμα. Τα δημόσια δεδομένα είναι δεδομένα που δεν υπόκεινται σε νόμιμη ιδιωτικότητα (privacy), ασφάλεια ή περιορισμούς.

2. Πρωτογενή. Τα δεδομένα συλλέγονται από την πηγή τους, στην αρχική τους μορφή και όχι σε συγκεντρωτικές ή τροποποιημένες μορφές.

3. Έγκαιρα. Τα δεδομένα γίνονται διαθέσιμα όσο το συντομότερο απαιτείται για τη διατήρηση της αξία τους.

4. Προσβάσιμα. Τα δεδομένα διατίθενται στο ευρύτερο φάσμα χρηστών για το ευρύτερο φάσμα σκοπών.

5. Επεξεργάσιμα από μηχανές. Τα δεδομένα είναι λογικά δομημένα για να επιτρέπουν την αυτόματη επεξεργασία τους.

6. Χωρίς διακρίσεις. Τα δεδομένα είναι διαθέσιμα σε όλους, χωρίς να απαιτείται εγγραφή του χρήστη σε αυτά.

7. Σε μη-ιδιόκτητες μορφές (non-proprietary formats). Τα δεδομένα διατίθενται σε μορφή της οποίας κανένας φορέας δεν έχει τον αποκλειστικό έλεγχο.

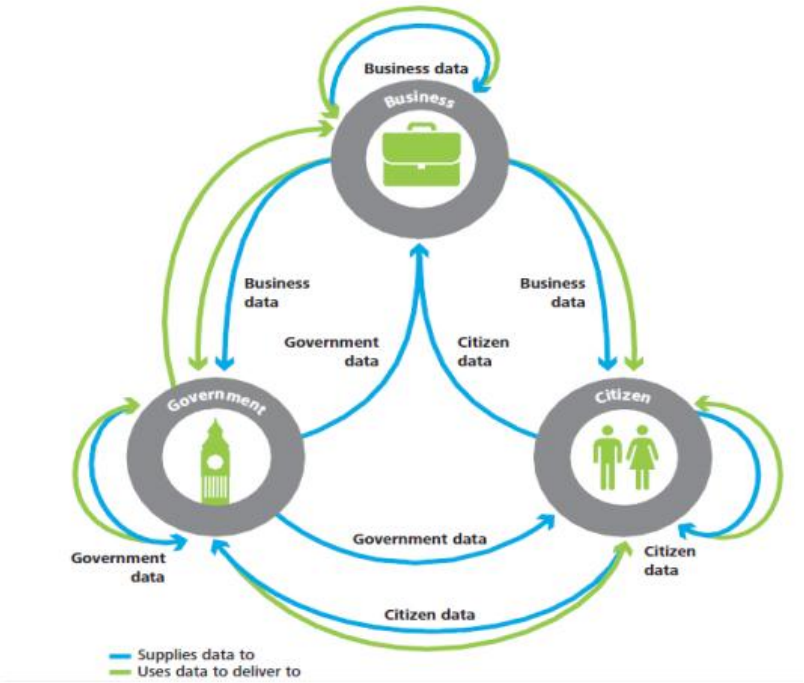
8. Χωρίς άδεια. Τα δεδομένα δεν υπόκεινται σε πνευματικά δικαιώματα, πατέντας, εμπορικού σήματος ή κανονισμούς εμπορίου. Εύλογη ιδιωτικότητα, ασφάλεια και περιορισμοί δικαιωμάτων επιτρέπονται.

Συμπερασματικά, μια ταξινόμηση των υπαρχόντων δεδομένων στην πολιτική και τη διοίκηση μπορεί να περιλαμβάνει αξιόλογα ευρήματα που πιθανώς θα επηρεάσουν τη διαδικασία λήψης αποφάσεων. Επί του παρόντος, χρήση και επεξεργασία μπορεί να γίνει μόνο σε ανοιχτές και ιδιόκτητες φόρμες δεδομένων ωστόσο όλες οι φόρμες δεδομένων δεν δημοσιεύονται και κανένας ενδιαφερόμενος δεν μπορεί να επηρεάσει τις προδιαγραφές που τηρούνται για τα δεδομένα. Σε ολόκληρο τον κόσμο, υπάρχουν διάφορα συστήματα παροχής αδειών που χρησιμοποιούν διαβαθμισμένα επίπεδα πρόσβασης στην διανομή και επεξεργασία δεδομένων. Η πρόσβαση στα πολύ αποθηκευμένα μπορεί να χορηγηθεί χωρίς τέλη και άλλα εμπόδια. Ωστόσο, μπορούν να υπάρξουν επιχειρηματικά μοντέλα που θα χρησιμοποιούν συνδρομές, έξοδα συναλλαγής, προμήθειες και χρεώσεις όγκου δεδομένων. Εάν υπάρξουν αποθηκευμένα δεδομένα τα οποία έχουν συλλεχθεί από τη δημόσια τάξη και δεν έχουν δηλωθεί ως δεδομένα δημοσίου τομέα, συχνά οικειοποιούνται από κράτη και τοπικές αρχές. Εναλλακτικά, μια εταιρεία ή μια ένωση θα μπορούσε να πάρει άδεια για να βελτιώσει τα αποθηκευμένα

δεδομένα και να τα πουλήσει αποκομίζοντας κέρδος. Οι πάροχοι υπηρεσιών μπορούν να δραστηριοποιηθούν στη συλλογή την ομαδοποίηση και την τελική επεξεργασία και βελτίωση των δεδομένων. Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν τόσο για δημόσια όσο και ιδιωτική χρήση.

2.6 Ανοιχτά Κυβερνητικά Δεδομένα και επιχειρήσεις

Η αξιοποίηση των διασυνδεδεμένων δεδομένων έχει δημιουργήσει μια νέα επιχειρηματική δραστηριότητα. Υπάρχουν εταιρίες οι οποίες στηρίζουν την δραστηριότητα τους στην αξιοποίηση αυτών των δεδομένων. Σε αυτό συνέβαλε και η οδηγία της Ευρωπαικής Ένωσης το 2003, η οποία αφορούσε το κόστος της χρησιμοποίησης τους. Μέχρι τότε οποιοσδήποτε χρειαζόταν να επαναχρησιμοποιήσει δημόσια πληροφορία έπρεπε να πληρώσει το αντίστοιχο κόστος, το οποίο τις περισσότερες φορές ήταν ασύμφορο. Με την οδηγία αυτή, τα δεδομένα τα οποία συλλέγονταν από τις δημόσιες αρχές θα διετίθεντο ελεύθερα σε όποιον ήθελε να τα χρησιμοποιήσει. Η οδηγία αυτή βασίστηκε στο σκεπτικό ότι η χρηματοδότηση είχε προκύψει από τους φορολογούμενους Ευρωπαίους πολίτες. Αυτό έδωσε το έναυσμα της ενασχόλησης των επιχειρήσεων αναφορικά με τη ροή ανοιχτών κυβερνητικών δεδομένων μεταξύ κυβέρνησης, πολιτών και επιχειρήσεων. Οι επιχειρήσεις και οι κυβερνήσεις χρησιμοποιούν τα δεδομένα με σκοπό να παρέχουν υπηρεσίες οι οποίες θα είναι χρήσιμες και στους τρεις αυτούς παράγοντες. Τα κυβερνητικά δεδομένα μαζεύονται, συλλέγονται από το Δημόσιο Τομέα με περιορισμό σε θέματα που αφορούν είτε την εθνική ασφάλεια είτε ευαίσθητα ή ιδιωτικά δεδομένα. Τα δεδομένα των επιχειρήσεων παράγονται/συλλέγονται από τον Ιδιωτικό Τομέα και δημοσιεύονται ελεύθερα και ανοιχτά. Περιορισμοί στα δεδομένα που δημοσιοποιούνται τίθενται ανάλογα από την επιχείρηση. Τέλος, οι πολίτες μπορούν να διαθέσουν ιδιωτικά ή μη ιδιωτικά δεδομένα στο διαδίκτυο. Στο σχήμα 2.6 που ακολουθεί φαίνεται πως οι τρεις αυτοί παράγοντες αλληλεπιδρούν μεταξύ τους και πως ο ένας προμηθεύει και εκμεταλλεύεται τα δεδομένα των άλλων.



Σχήμα 2.6: Απεικόνιση της αλληλεπίδρασης των 3 φορέων

Σαν αποτέλεσμα της δημοσιοποίησης οι κυβερνήσεις βελτιώνουν τις δημόσιες υπηρεσίες τους και έχουν την εποπτεία των δραστηριοτήτων τους σε ένα ευρύ φάσμα φορέων του δημόσιου τομέα. Με τη σειρά τους οι επιχειρήσεις επιτυγχάνουν την τόνωση της οικονομικής τους ανάπτυξης εκμεταλλευόμενες τα ανοιχτά δεδομένα, δημιουργούν ανταγωνισμό αναγκάζοντας και άλλες επιχειρήσεις να δημοσιοποιήσουν τα δεδομένα τους, βελτιώνουν τις επιχειρηματικές τους αποφάσεις και προσθέτουν αξία στα δεδομένα τους, Τέλος, οι πολίτες επωφελούνται από τα μέτρα κατά της διαφθοράς και υπέρ της διαφάνειας, διευκολύνουν τις αγορές τους, μειώνεται το κόστος για την πρόσβαση σε δεδομένα ανοιχτά ενώ ταυτόχρονα αυξάνεται η πρόσβαση του πολίτη στα δεδομένα και βελτιώνεται η ποιότητα των δεδομένων.

ΚΕΦΑΛΑΙΟ 3

Στατιστική Μάθηση

3.1 Ιστορική αναδρομή στατιστικής μάθησης

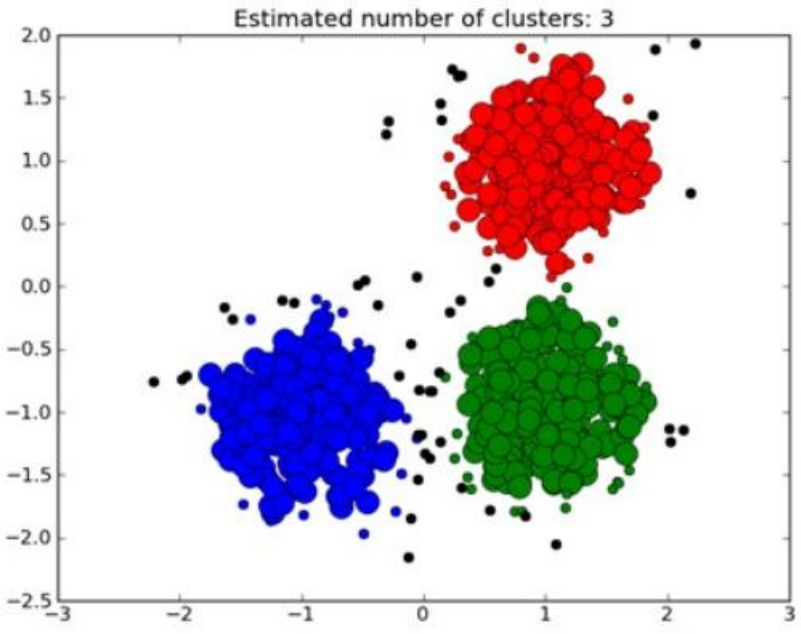
Η στατιστική μάθηση αναφέρεται σε μια ευρεία επιλογή εργαλείων τα οποία έχουν ως στόχο να "καταλάβουμε" τα δεδομένα. Η στατιστική μάθηση παρέχει τη θεωρητική βάση για τους περισσότερους αλγόριθμους που έχουν αναπτυχθεί στη μηχανική μάθηση. Ο όρος στατιστική μάθηση εισήχθη, το 1960 στη Ρωσία όπου απέκτησε μεγάλη δημοτικότητα την δεκαετία του 1990, με την ανάπτυξη των support vector machines. Μέχρι τις αρχές του 1990 αποτελούσε μια θεωρητική ανάλυση του προβλήματος της εκτίμησης για μια συγκεκριμένη συλλογή δεδομένων. Στα μέσα της δεκαετίας του '90, αναπτύχθηκαν οι αλγόριθμοι μάθησης support vector machines. Οι νέοι αυτοί αλγόριθμοι άλλαξαν τη στατιστική μάθηση από ένα εργαλείο θεωρητικής ανάλυσης σε ένα εργαλείο για τη δημιουργία πρακτικών αλγόριθμων με στόχο την επίλυση πραγματικών προβλημάτων σε διάφορους τομείς. Αν και σχετικά πρόσφατα αναπτύχθηκε η στατιστική μάθηση πολλές από τις έννοιες οι οποίες χρησιμοποιήθηκαν ήταν ήδη γνωστές από τις αρχές του 19ου αιώνα. Χαρακτηριστικό παράδειγμα, είναι η γραμμική παλινδρόμηση η οποία βασίζεται και είναι στην ουσία η εφαρμογή της θεωρίας των Gauss και Legendre αναφορικά με τη μέθοδο των ελαχίστων τετραγώνων. Η πρώτη επιτυχής εφαρμογή έγινε σε προβλήματα αστρονομίας για να ακολουθήσουν αργότερα εφαρμογές και σε άλλους τομείς όπως τα οικονομικά, το marketing, η ιατρική κ.α. Αργότερα, και μέσω της ιατρικής δημιουργήθηκε η ανάγκη να απαντήσουμε σε ερωτήματα όπως το αν θα επιζήσει ένας ασθενής ή όχι, στα οικονομικά αν θα ανέβει η όχι η μετοχή μιας εταιρίας, στα τραπεζικά η έγκριση ενός δανείου ή όχι κ.α. Έτσι το 1936 ο Fisher πρότεινε τη θεωρία της γραμμικής διακριτής ανάλυσης (linear discriminant analysis). Η θεωρία αυτή βρήκε απήχηση και λίγο αργότερα προτάθηκε μια διαφορετική προσέγγιση η λογιστική παλινδρόμηση. Έτσι το 1970 οι Nelder και Wedderburn εισήγαγαν τον όρο γενικευμένα γραμμικά μοντέλα (generalized linear model) για ολόκληρη την κατηγορία της στατιστικής μάθησης η οποία περιλαμβάνει τα γραμμικά και λογιστικά μοντέλα. Στη συνέχεια και έως το 1970 αναπτύχθηκαν και άλλες τεχνικές για την μάθηση από τα δεδομένα αλλά όλες αφορούσαν γραμμικές μεθόδους καθώς μέχρι την έως τότε ανάπτυξη των

υπολογιστών ήταν ανέφικτη η εφαρμογή μη γραμμικών μεθόδων. Τη δεκαετία του 1980 με την ανάπτυξη των ηλεκτρονικών υπολογιστών, η χρήση μη γραμμικών μεθόδων ήταν πια εφικτή υπό την έννοια του υπολογιστικού κόστους και έτσι οι Brieman, Friedman, Olshen και Stone παρουσίασαν την κατηγοριοποίηση (classification) μέσω της παλινδρόμηση δέντρων (regression trees). Αργότερα, στα τέλη του 1980 οι Hastie και Tibshirani εισήγαγαν τον όρο των γενικευμένων πρόσθετων μοντέλων (generalised additive models), μια μη γραμμική επέκταση των γραμμικών μοντέλων. Από εκείνη τη στιγμή και με την έκρηξη της μηχανικής μάθησης, η στατιστική μάθηση αναδείχθηκε σε ένα πολύτιμο εργαλείο. Την τελευταία δεκαπενταετία, με την ευρεία ανάπτυξη των υπολογιστών και την ταυτόχρονη μείωση του υπολογιστικού κόστους αλλά και την ολοένα και αυξανόμενη εφαρμογή των τεχνικών αυτών σε πρακτικά πεδία, δημιουργήθηκε πληθώρα στατιστικών πακέτων τα οποία διευκολύνουν σε μεγάλο βαθμό τη στατιστική ανάλυση αλλά και την ανάπτυξη νέων πολύπλοκων τεχνικών.

Ειδικότερα, η στατιστική μάθηση χωρίζεται σε δύο κύριες κατηγορίες, την εποπτευόμενη μάθηση (supervised learning) και τη μη εποπτευόμενη μάθηση (unsupervised learning). Η εποπτευόμενη μάθηση έχει ως στόχο τη δημιουργία ενός στατιστικού μοντέλου για την πρόβλεψη της κλάσης των δεδομένων εισόδου έχοντας στη διάθεση της όμως τα δεδομένα εξόδου (την κλάση στην οποία ανήκουν). Αντίθετα στη μη εποπτευόμενη μάθηση έχουμε μόνο τα δεδομένα εισόδου διαθέσιμα, δηλαδή δεν έχουμε εποπτεία στα δεδομένα μας, με σκοπό να δημιουργήσουμε ομάδες από τα δεδομένα εισόδου που μοιάζουν μεταξύ τους. Ανάμεσα στην εποπτευόμενη και την μη-εποπτευόμενη μάθηση οι οποίες είναι δύο πλήρως διαφορετικές μέθοδοι από την άποψη της ύπαρξης ή όχι των δεδομένων εξόδου, υπάρχει η ημι-εποπτευόμενη μάθηση (semi-supervised). Στην ημι-εποπτευόμενη μάθηση υπάρχει συνήθως ένας πολύ μικρός αριθμός μεταβλητών εξόδου για τις μεταβλητές εισόδου ενώ για τις υπόλοιπες δεν διατίθενται. Αυτά τα μοντέλα χρησιμοποιούν πρώτα τα δεδομένα για τα οποία υπάρχει η μεταβλητή εξόδου με σκοπό να κατασκευάσουν ένα αρχικό μοντέλο και στη συνέχεια ενσωματώνουν τα δεδομένα τα οποία δεν έχουν μεταβλητή εξόδου αντιστοιχίζοντας σε αυτά την μεταβλητή εξόδου την οποία έχει προβλέψει το μοντέλο εκπαίδευσης μέχρι εκείνο το σημείο. Στόχος της είναι να ανακαλύψει τις σχέσεις και τη δομή που έχουν τα δεδομένα σε αυτή τη περίπτωση.

3.2 Μη εποπτευόμενη μάθηση

Η μη εποπτευόμενη μάθηση έχει ως στόχο να ανακαλύψει ομάδες από τα δεδομένα βασιζόμενη στα χαρακτηριστικά τους. Η συσταδοποίηση ή αλλιώς ομαδοποίηση είναι η πιο σημαντική μέθοδος μη εποπτευόμενης μάθησης. Το πλήθος των ομάδων είτε είναι γνωστό από την αρχή είτε αποφασίζεται από τον αλγόριθμο μάθησης. Οι ομάδες ή αλλιώς συστάδες (clusters) αποτελούνται από ένα σύνολο των δεδομένων με βάση τη μεταξύ τους ομοιότητα. Το βασικότερο μέτρο βάσει του οποίου αναγνωρίζουμε πότε δυο παρατηρήσεις είναι όμοιες ή ανόμοιες μεταξύ τους, είναι η απόσταση. Παρατηρήσεις που μοιάζουν πολύ, θα πρέπει να έχουν μικρή απόσταση μεταξύ τους ενώ αντίστοιχα ανόμοιες παρατηρήσεις θα πρέπει να απέχουν αρκετά. Έως σήμερα έχουν προταθεί πολλά μέτρα απόστασης για τη δημιουργία συστάδων ανάλογα τα χαρακτηριστικά τα οποία πρόκειται να ομαδοποιήσουμε. Για συνεχή χαρακτηριστικά, τα κύρια μέτρα απόστασης είναι η Ευκλείδεια απόσταση, η απόσταση Manhattan, η απόσταση Minkowski, η απόσταση Chebyshev κ.α. Όταν τα χαρακτηριστικά είναι δίτιμα, δηλαδή μεταβλητές που μπορούν να πάρουν μόνο τις τιμές 0 ή 1, οι παραπάνω αποστάσεις μειονεκτούν. Για παράδειγμα, αν χρησιμοποιούσαμε την Ευκλείδεια απόσταση, αυτή θα ισούταν με 0 ή 1 και θα αντιμετώπιζε με τον ίδιο τρόπο είτε είχαμε συμφωνία στη παρουσία είτε στην απουσία του χαρακτηριστικού (Συνήθως με 1 συμβολίζεται η παρουσία του χαρακτηριστικού και με 0 η απουσία του). Σε αυτές τις περιπτώσεις, κατασκευάζουμε ένα πίνακα συνάφειας και μέσω αυτού δίνουμε βάρη στη παρουσία ή μη ενός χαρακτηριστικού. Τέτοιες αποστάσεις είναι η Simple matching απόσταση, η οποία δίνει βαρύτητα στις ασυμφωνίες, η Sokal and Sneath η οποία δίνει διπλάσσιο βάρος στις συμφωνίες κ.α. Όσο πιο μεγάλη είναι η τιμή αυτών των αποστάσεων τόσο πιο πολύ διαφέρουν οι παρατηρήσεις. Μια άλλη κατηγορία μέτρων που μπορούν να χρησιμοποιηθούν είναι τα μέτρα ομοιότητας (similarity measures), τα οποία για παρατηρήσεις που είναι όμοιες μεταξύ τους, δίνουν μεγάλες τιμές ενώ για παρατηρήσεις που είναι ανόμοιες δίνουν μικρές τιμές. Έτσι, σε αυτή τη περίπτωση, οι παρατηρήσεις που ανήκουν στην ίδια ομάδα, θα έχουν μεταξύ τους μεγάλες τιμές στο μέτρο ομοιότητας. Ένα μέτρο ομοιότητας για συνεχείς παρατηρήσεις, είναι ο γνωστός μας συντελεστή συσχέτισης. Για δίτιμες μεταβλητές, τα κύρια μέτρα ομοιότητας είναι ξανά το Simple matching αλλά αυτή τη φορά μιας και αναφερόμαστε σε μέτρα ομοιότητας, μετρά τις συμφωνίες και η Sokal and Sneath η οποία τώρα δίνει διπλάσσιο βάρος στις ασυμφωνίες.



Σχήμα 3.1: Γραφική απεικόνιση μιας συσταδοποίησης 3 ομάδων

Αφού ορίσαμε λοιπόν τα μέτρα απόστασης, τα οποία καθορίζουν την ομοιότητα δυο παρατηρήσεων, θα εξετάσουμε στη συνέχεια τις διαφορετικές προσεγγίσεις σύμφωνα με τις οποίες μπορούμε να ομαδοποιήσουμε τα δεδομένα μας. Οι μέθοδοι ομαδοποίησης χωρίζονται σε δυο κατηγορίες ανάλογα τον τρόπο με τον οποίο δημιουργούν τις διαφορετικές ομάδες. Αυτές είναι, οι ιεραρχικές και οι μη ιεραρχικές μέθοδοι. Όταν ο αριθμός των ομάδων είναι γνωστός εκ των προτέρων αναφερόμαστε στις μη ιεραρχικές ενώ όταν ο αριθμός των ομάδων είναι άγνωστος στις ιεραρχικές. Οι ιεραρχικοί μέθοδοι, χωρίζονται με τη σειρά τους, σε συσσωρευτικούς και διαιρετικούς. Οι συσσωρευτικές μέθοδοι, δημιουργούν τις ομάδες σταδιακά. Αρχικά, κάθε παρατήρηση αποτελεί και μια συστάδα. Οι συστάδες συγχωνεύονται επαναληπτικά σχηματίζοντας συνεχώς μεγαλύτερες συστάδες μέχρι να καταλήξουμε σε μια συστάδα η οποία θα αποτελείται από το σύνολο των παρατηρήσεων. Γενικά, ένας αλγόριθμος μιας συσσωρευτικής μεθόδου με τη μορφή βημάτων (Μ.Κούτρας-2013) λειτουργεί ως εξής:

1. Ξεκινάμε με n ομάδες (clusters) του ενός ατόμου η καθεμία, και με τον nxn πίνακα των αποστάσεων $D=[d_{ij}]$ (εναλλακτικά θα μπορούσε να χρησιμοποιηθεί ένας πίνακας μέτρων ομοιότητας)
2. Εντόπισε στον πίνακα D το ζεύγος των πλησιέστερων (πιο όμοιων) ομάδων έστω Q και R.
3. Συγχώνευσε τις ομάδες Q και R σε μια ομάδα, την P=(QR) μειώνοντας έτσι τον

αριθμό των ομάδων κατά ένα. Ανανέωσε τον πίνακα αποστάσεων D διαγράφοντας τις γραμμές και στήλες που αντιστοιχούσαν στις ομάδες Q και R και προσθέτοντας μια γραμμή και μια στήλη που περιέχει τις αποστάσεις της ομάδας P=(QR) από τις υπόλοιπες ομάδες.

4. Επανέλαβε τα βήματα 2 και 3 συνολικά n-1 φορές έτσι ώστε με τη λήξη του αλγορίθμου, όλα τα άτομα να αποτελούν μια μόνο ομάδα. Σε κάθε βήμα κατάγραψε τις λεπτομέρειες συνένωσης, την ταυτότητα των ομάδων και τα επίπεδα (distances ή similarities) στα οποία πραγματοποιούνται οι συγχωνεύσεις.

Οι πιο δημοφιλείς μέθοδοι των συσσωρευτικών αλγορίθμων είναι η μέθοδος του κοντινότερου γείτονα (Nearest Neighbor Method), της απλής συνένωσης (Single Linkage Method), των κέντρων βάρους (Centroid Method), η μέθοδος του Ward και αυτή των σταθμισμένων μέσων (Weighted Average Linkage Method). Οι διάφορες συγκρίσεις μεταξύ των συσσωρευτικών μεθόδων έχουν δείξει ότι καλύτερα αποτελέσματα δίνουν, η μέθοδος του Ward και η μέθοδος των σταθμισμένων μέσων ενώ αυτή με τα χειρότερα αποτελέσματα είναι του κοντινότερου γείτονα.

Οι διαιρετικές μέθοδοι λειτουργούν ακριβώς αντίθετα. Αρχικά όλες οι παρατηρήσεις ανήκουν σε μια συστάδα και στη συνέχεια οι μεγάλες συστάδες προοδευτικά διαιρούνται έως ότου στο τέλος κάθε παρατήρηση θα είναι από μόνη της μια συστάδα. Η κύρια ιδέα των διαιρετικών μεθόδων είναι να δημιουργούν υποομάδες των ήδη υπαρχόντων ομάδων, οι οποίες να είναι όλο και περισσότερο ανόμοιες μεταξύ τους. Ο τρόπος με τον οποίο αποφασίζεται ποια ομάδα θα διασπαστεί βασίζεται στις μεταβλητές που εξετάζονται μια κάθε φορά ή σε όλες τις μεταβλητές που εξετάζονται συγχρόνως. Από τους πιο γνωστούς διαιρετικούς αλγόριθμους είναι αυτός των Edwards και Cavalli-Sforza (1965), ο οποίος διαχωρίζει τις ομάδες βάσει του διαμερισμού ο οποίος ελαχιστοποιεί το άθροισμα των τετραγωνικών αποκλίσεων. Το βασικό μειονέκτημα των διαιρετικών μεθόδων είναι το υπολογιστικό κόστος αφού χρειάζονται πολύ περισσότεροι υπολογισμοί από ότι οι συσσωρευτικοί. Οι πιθανοί διαμερισμοί ενός συνόλου n αντικειμένων σε δυο ομάδες είναι $2^{n} - 1$ ενώ σε ένα συσσωρευτικό αλγόριθμο, οι πιθανές ενώσεις είναι $n(n-1)/2$.

Οι μη ιεραρχικοί μέθοδοι, όπως αναφέραμε και παραπάνω θεωρούν ότι ο αριθμός των ομάδων είναι από πριν γνωστός. Δοθέντος λοιπόν του αριθμού των συστάδων που θέλουμε

να δημιουργήσουμε ο αλγόριθμος αναζητά τις βέλτιστες διαμερίσεις των παρατηρήσεων. Ο αλγόριθμος ξεκινά με μια αρχική ομαδοποίηση των παρατηρήσεων σε k-ομάδες, όσες είναι ο αριθμός που θέλουμε να δημιουργήσουμε, στη συνέχεια μετακινούνται οι παρατηρήσεις μεταξύ των ομάδων μέχρις ότου να φτάσει σε ένα σημείο ισορροπίας, δηλαδή κανένα άτομο να μην αλλάζει ομάδα. Εναλλακτικά, ο αλγόριθμος μπορεί να επιλέξει k-αρχικά στοιχεία τα οποία λέγονται μητρικά και στη συνέχεια με βάση αυτά τα στοιχεία ταξινομούνται τα υπόλοιπα μέχρι να δημιουργηθούν οι επιθυμητές ομάδες. Η κύρια διαφορά των μεθόδων αυτών βρίσκεται στον τρόπο με τον οποίο γίνεται η ανανέωση των κέντρων των ομάδων. Ο πιο δημοφιλής αλγόριθμος των μη ιεραρχικών μεθόδων, είναι ο k-means όπως αυτός προτάθηκε από τον MacQueen το 1967 και υλοποιείται με βάση τα ακόλουθα βήματα

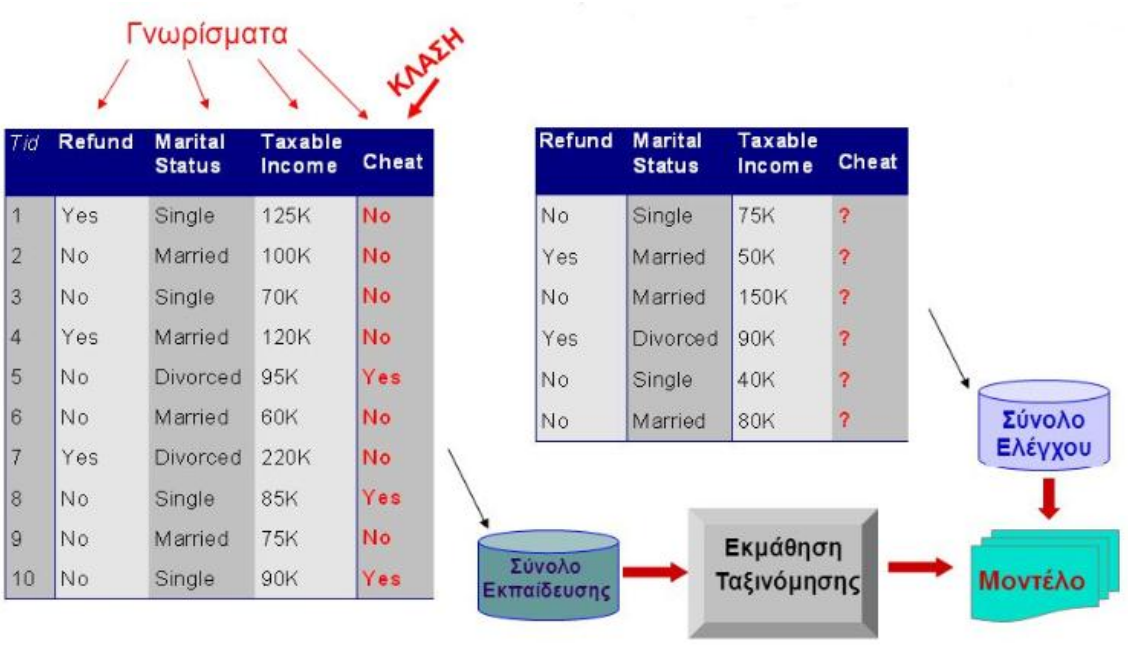
1. Όρισε ένα αρχικό σύνολο από k-μητρικά στοιχεία (centroid)
Κατέταξε τα υπόλοιπα στοιχεία του αρχικού συνόλου στην ομάδα της οποίας το κέντρο βρίσκεται πιο κοντά. Μετά από κάθε ανάθεση υπολόγισε ξανά το κέντρο βάρους της ομάδας
2. Όταν όλα τα στοιχεία έχουν ανατεθεί στις ομάδες, θέσε τα κέντρα βάρους ως μητρικά στοιχεία και επανέλαβε το βήμα 2 μέχρι να επιτύχουμε ισορροπία.

Με τον όρο "ισορροπία" εννοούμε την κατάσταση κατά την οποία τα στοιχεία δεν μετακινούνται μεταξύ των ομάδων. Ο αλγόριθμος k-means είναι αρκετά γρήγορος αφού συχνά τερματίζει μετά από ένα μικρό αριθμό επαναλήψεων ενώ ταυτόχρονα δεν έχει μεγάλο υπολογιστικό κόστος αφού δεν απαιτείται να κρατά στη μνήμη του πολλά δεδομένα. Ωστόσο είναι ευάλωτος και εξαρτάται σε μεγάλο βαθμό από τα αρχικά μητρικά στοιχεία. Η μη σωστή επιλογή αρχικών μητρικών στοιχείων μπορεί να οδηγήσει σε τοπικές βέλτιστες λύσεις. Για την αντιμετώπιση αυτού του προβλήματος, γίνονται αρκετές δοκιμές με διαφορετικά αρχικά μητρικά στοιχεία.

3.3 Εποπτευόμενη Μάθηση

Τα δεδομένα εισόδου ενός προβλήματος κατηγοριοποίησης ή ταξινόμησης (classification) αποτελούνται από ένα σύνολο εγγραφών. Κάθε εγγραφή χαρακτηρίζεται από ένα σύνολο χαρακτηριστικών (attributes) και από την τιμή ενός γνωρίσματος, που καλείται κατηγορία ή κλάση (class). Οι τιμές των χαρακτηριστικών της κάθε εγγραφής μπορεί να είναι συνεχείς ή

διακριτές ενώ η κλάση πρέπει οπωσδήποτε να παίρνει διακριτές τιμές. Ταξινόμηση, ονομάζεται η διαδικασία μάθησης μιας συνάρτησης f, η οποία πρέπει για κάθε εγγραφή εισόδου να απεικονίζει το σύνολο των χαρακτηριστικών της, στην τιμή της κάθε κλάσης. Η συνάρτηση αυτή ονομάζεται μοντέλο ταξινόμησης. Ένα τέτοιο μοντέλο αποτελεί στην ουσία ένα εργαλείο μέσω του οποίου οι εγγραφές διαχωρίζονται στις κατηγορίες τις οποίες ανήκουν. Ένα μοντέλο ταξινόμησης χρησιμοποιείται κυρίως για προγνωστικούς λόγους όταν εφαρμόζεται για να ταξινομήσει νέες εγγραφές, των οποίων οι τιμές των χαρακτηριστικών είναι γνωστές, αλλά δεν γνωρίζουμε την τιμή της κλάσης στην οποία ανήκει. Επειδή στο σύνολο δεδομένων εισόδου οι τιμές κλάσης των εγγραφών είναι γνωστές εκ των προτέρων, η ταξινόμηση ανήκει στην κατηγορία των εποπτευόμενων μεθόδων μάθησης και αυτή είναι και η κύρια διαφορά με τις μεθόδους μη εποπτευόμενης μάθησης που μελετήσαμε νωρίτερα. Η διαφορά της ταξινόμησης από το πρόβλημα της παλινδρόμησης (μέθοδος εποπτευόμενης μάθησης) έγκειται στο ότι η παλινδρόμηση απαιτεί οι τιμές της εξαρτημένης μεταβλητής να είναι συνεχείς και όχι διακριτές. Κάθε αλγόριθμος εφαρμόζει μια τεχνική μάθησης, για να αναγνωρίσει το μοντέλο που προσαρμόζεται καλύτερα στις συσχετίσεις μεταξύ των γνωρισμάτων και των τιμών κλάσης των εγγραφών εισόδου. Το μοντέλο που παράγεται από τον αλγόριθμο ταξινόμησης πρέπει να προσαρμόζεται καλά στα υπάρχοντα δεδομένα, αλλά και να μπορεί να προβλέπει με όσο το δυνατόν μεγαλύτερη ακρίβεια τις κλάσεις νέων εγγραφών.



Σχήμα 3.2: Κατασκευή ενός μοντέλου ταξινόμησης.

Η επίλυση ενός προβλήματος ταξινόμησης ακολουθεί τα επόμενα βήματα όπως αυτά φαίνονται και στο σχήμα 3.2. Καταρχήν, θα πρέπει να έχουμε ένα σύνολο εκπαίδευσης (training set) με εγγραφές των οποίων οι κλάσεις να είναι γνωστές. Το σύνολο αυτό χρησιμοποιείται για την εκπαίδευση του μοντέλου ταξινόμησης με βάση κάποιον αλγόριθμο. Μετέπειτα, το εκπαιδευμένο μοντέλο εφαρμόζεται για να ταξινομήσει τις εγγραφές ενός συνόλου δοκιμής (test set) στο οποίο οι τιμές της κλάσης είναι άγνωστες. Σε περιπτώσεις όπου δεν δίνεται σύνολο δοκιμής, χωρίζουμε το αρχικό μας σύνολο δεδομένων σε δυο σύνολα, σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο δοκιμής με τυχαίο τρόπο.

Στη συνέχεια της παρούσας διπλωματικής θα μελετήσουμε διεξοδικά διαφόρους αλγόριθμους ταξινόμησης. Οι αλγόριθμοι ταξινόμησης στους οποίους θα αναφερθούμε παρακάτω είναι τα δένδρα απόφασης, η λογιστική παλινδρόμηση, η διαχωριστική ανάλυση, ο KNN, τα τεχνητά νευρωνικά δίκτυα, τα support vector machines, ο απλός ταξινομητής Naive-Bayes και οι μέθοδοι bagging, random forest και boosting.

3.3.1 Λογιστική Παλινδρόμηση

Όταν η μεταβλητή απόκρισης είναι δίτιμη (binary) μπορούμε να χρησιμοποιήσουμε τη λογιστική παλινδρόμηση για την επίλυση ενός προβλήματος ταξινόμησης. Η λογιστική παλινδρόμηση χρησιμοποιεί ένα μετασχηματισμό της μέσης τιμής της μεταβλητής Y,

$$n_i = g(p_i) = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots\ldots + b_k X_{ik}$$

Όπου η g είναι μια συνάρτηση που απεικονίζει το διάστημα [0,1] στην πραγματική ευθεία έτσι ώστε

$$g^{-1}(n_i) = p_i\,, \qquad \epsilon\,[0,1]$$

Η συνάρτηση σύνδεσης που θα χρησιμοποιήσουμε είναι η Logit,

$$n_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

Μια μεταβλητή Y, λέμε ότι ακολουθεί τη λογιστική κατανομή με παραμέτρους μ και σ^2 όταν η αθροιστική της συνάρτηση γράφεται στη μορφή

$$P(Y \le y) = \frac{1}{1 + e^{-(y-\mu)/s}} \qquad -\infty < y < \infty$$

Θεωρώντας ότι η μέση τιμή θα ισούται με $\mu_i = b_0 + b_1 x_i$ και για s=1 παίρνουμε

$$p_i = P(Y_i = 1) = 1 - \frac{1}{1 + e^{b_0 + b_1 x_i}} = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}}$$

Αντίστοιχα για το 1-p_i θα έχουμε ότι

$$1 - p_i = P(Y_i = 0) = \frac{1}{1 + e^{b_0 + b_1 x_i}}$$

Διαιρώντας τις δύο τελευταίες σχέσεις κατά μέλη παίρνουμε ότι

$$\frac{p_i}{1 - p_i} = e^{b_0 + b_1 x_i}$$

Στη συνέχεια λογαριθμίζοντας προκύπτει η παρακάτω σχέση

$$\log\left(\frac{p_i}{1 - p_i}\right) = b_0 + b_1 x_i$$

Από την παραπάνω σχέση παρατηρούμε ότι η ερμηνευτική μεταβλητή συνδέεται με την πιθανότητα επιτυχίας μέσω της συνάρτησης logit. Η συνάρτηση logit αναφέρεται στο λογάριθμο της σχετικής πιθανότητας του ενδεχομένου που μας ενδιαφέρει, δηλαδή της επιτυχίας. Επίσης από την τελευταία σχέση παρατηρούμε ότι η αύξηση της μεταβλητής της x κατά μια μονάδα προκαλεί πολλαπλασιαστική αύξηση της σχετικής πιθανότητας επιτυχίας κατά $\exp(\widehat{b_1})$ δεδομένου ότι οι άλλες μεταβλητές παραμένουν σταθερές. Η συνάρτηση logit είναι συμμετρική για την πιθανότητα επιτυχίας και αποτυχίας με αποτέλεσμα να μην έχει σημασία ποιό από τα δύο αποτελέσματα της δίτιμης μεταβλητής ορίζουμε ως επιτυχία ή αποτυχία αντίστοιχα. Ο λόγος $\frac{p_i}{1-p_i}$ λέγεται odds και μπορεί να πάρει τιμές μεταξύ μηδέν και άπειρο. Τιμές της σχετικής πιθανότητας μεγαλύτερες της μονάδας υποδεικνύουν ότι το ενδεχόμενο στον αριθμητή είναι πιο πιθανό να συμβεί από αυτό στον παρονομαστή. Αξίζει σε αυτό το σημείο να τονίσουμε, ότι επειδή η σχέση μεταξύ p_i και x_i δεν είναι γραμμική, το b_1 δεν αντιστοιχει στην μεταταβολή του p_i δεδομένου ότι το x αυξήθηκε κατα μια μονάδα. Το πόσο θα αλλάξει η τιμή του p_i δεδομένου ότι η τιμή του x θα αυξηθεί κατά 1 μονάδα θα εξαρτάται και από την τρέχουσα τιμή του x. Ανεξάρτητα όμως από την τιμή του X, θετικό b_1 σημαίνει ότι θα έχουμε αύξηση της τιμής του p_i ενώ αρνητική τιμή του b_1 θα σημαίνει αντίστοιχα μείωση.

Η εκτίμηση των παραμέτρων b_0,b_1,... γίνεται χρησιμοποιώντας τη μέθοδο μέγιστης πιθανοφάνειας. Με τη μέθοδο της μεγίστης πιθανοφάνειας και θεωρώντας το δίανυσμα των παραμέτρων $b' = (b_0, b_1, \dots \dots, b_k)$, προκύπτει μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας

$$l(b') = \prod_{i:y_i=1} p(x_i) \prod_{i':y'_i=0} 1 - p(x_i) \text{ ως προς } b_i \text{ για } i=0,1,2,..,k$$

Στην ουσία προσπαθούμε να εκτιμήσουμε τα $\widehat{b_0}$, $\widehat{b_1}$,… έτσι ώστε οι εκτιμήσεις του μοντέλου να είναι κοντά στη μονάδα όταν έχουμε επιτυχία και κοντά στο μηδέν για την αποτυχία, έτσι ώστε να μπορούμε να ταξινομήσουμε τα δεδομένα μας με βάση τα χαρακτηριστικά τους.

Η λογιστική παλινδρόμηση έχει ένα βασικό πλεονέκτημα έναντι της κλασσικής παλινδρόμησης, ότι δεν απαιτεί η εξαρτημένη μεταβλητή Y να έχει σταθερή διακύμανση και ακόμα να ακολουθεί την Κανονική Κατανομή, με αποτέλεσμα να βρίσκει εφαρμογή σε μια μεγάλη ποικιλία προβλημάτων. Επίσης, οι εκτιμητές των παραμέτρων προκύπτουν με τη μέθοδο της μεγίστης πιθανοφάνειας και ως εκ τούτου έχουν μια σειρά από επιθυμητές ιδιότητες.

3.3.2. Linear Discriminant Analysis (LDA)

Ο κύριος στόχος και της Διαχωριστικής Ανάλυσης είναι να διαχωρίσει κάθε παρατήρηση στις k γνωστές ομάδες-κλάσεις. Αυτό πραγματοποιείται με τη χρήση ενός διαχωριστικού κανόνα, μέσω μιας διαχωριστικής συνάρτησης, που σκοπεύει να ταξινομήσει σωστά όσο το δυνατόν περισσότερες παρατηρήσεις. Η συνάρτηση αυτή στην ουσία είναι ένα υπερεπίπεδο το οποίο διαχωρίζει κατάλληλα τον χώρο των χαρακτηριστικών των παρατηρήσεων. Μια διαχωριστική συνάρτηση μπορεί να είναι γραμμική, τετραγωνική ή και πολυωνυμική. Στη περίπτωση όπου έχουμε δυο κλάσεις, η γραμμική συνάρτηση είναι ένα υπερεπίπεδο που διχοτομεί τις δυο κλάσεις. Συχνά χρησιμοποιείται η συνάρτηση Fischer για τον διαχωρισμό των κλάσεων. Σύμφωνα με αυτή τη συνάρτηση, τα δεδομένα εκπαίδευσης κάθε κατηγορίας ακολουθούν την Κανονική Κατανομή και μεγιστοποιώντας την Ευκλείδεια απόσταση ανάμεσα στις παρατηρήσεις των δυο κλάσεων εκτιμά ένα βέλτιστο όριο ταξινόμησης τους. Η συνάρτηση Fischer, δεν λαμβάνει υπόψη της το μέγεθος της κάθε κλάσης το οποίο μπορεί να διαφέρει σημαντικά. Για τον λόγο αυτό θα χρησιμοποιήσουμε τον κανόνα του Bayes. Όπως είδαμε παραπάνω, η λογιστική παλινδρόμηση περιλαμβάνει τον απευθείας υπολογισμό της

$\Pr(Y = k|X = x)$ χρησιμοποιώντας τη λογιστική συνάρτηση. Στην LDA χρησιμοποιούμε μια έμμεση προσέγγιση για την εκτίμηση αυτών των πιθανοτήτων. Με τη μέθοδο αυτή, μοντελοποιούμε την εκ των υστέρων κατανομή πιθανότητας για κάθε κλάση χρησιμοποιώντας τον κανόνα του Bayes. Έστω ότι π(k) συμβολίζει την εκ των προτέρων πιθανότητα ότι μια παρατήρηση προέρχεται από την k κλάση και έστω ότι $f_k(x)$ συμβολίζει την συνάρτηση πυκνότητας πιθανότητας μιας παρατήρησης να ανήκει στη κλάση k.

$$p_k(x) = f_k(x) = \Pr(X = x|Y = k)$$

Η $f_k(x)$ παίρνει μεγάλες τιμές όταν υπάρχει μεγάλη πιθανότητα να ανήκει στην k κλάση για X=x και μικρές τιμές όταν αυτό δεν είναι πιθανό. Τότε το Θ. Bayes αναφέρει ότι

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

Από αυτό φαίνεται, ότι αντί για τον απευθείας υπολογισμό του $p_k(x)$ μπορούμε να συνδυάσουμε τις εκτιμήσεις των π_k και $f_k(x)$ μέσω της τελευταίας σχέσης. Γενικά η εκτίμηση του π_k υπολογίζεται εύκολα αφού ισούται με την αναλογία των παρατηρήσεων των δεδομένων εκπαίδευσης που ανήκουν στην k κλάση.

Ωστόσο, η εκτίμηση του $f_k(x)$ είναι πιο δύσκολη εκτός αν υποθέσουμε κάποιες παραδοχές για τις πυκνότητες πιθανότητες. Αναφερόμαστε στο $p_k(x)$ ως την εκ των προτέρων πιθανότητα μια παρατήρηση X=x να ανήκει στην k κλάση, δηλαδή είναι η πιθανότητα ότι η παρατήρηση ανήκει στην κλάση κ δεδομένου του X. Θέλουμε να εκτιμήσουμε το $f_k(x)$ ώστε μέσω της τελευταίας σχέσης να εκτιμήσουμε το $p_k(x)$ και στη συνέχεια θα ταξινομήσουμε τη παρατήρηση στη κλάση με τη μεγαλύτερη πιθανότητα. Υποθέτουμε πρώτα ότι η $f_k(x)$ ακολουθεί Κανονική Κατανομή και ότι έχουμε ίσες διακυμάνσεις μεταξύ των κλάσεων. Με βάση αυτές τις παραδοχές και μέσω της συνάρτησης πυκνότητας πιθανότητας της Κανονικής κατανομής, τελικά, μια παρατήρηση ταξινομείται στη κλάση όπου η τιμή της $\delta_k(x)$ είναι μεγαλύτερη

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{{\mu_k}^2}{2\sigma^2} + \log(\pi_k)$$

όπου μ_k , σ^2, η μέση τιμή και η διακύμανση της k κλάσης και $\pi_k = n_k/n$, δηλαδή ο αριθμός των παρατηρήσεων που ανήκουν στην k-κλάση προς το σύνολο των παρατηρήσεων.

Όταν οι κατανομές των ανεξάρτητων μεταβλητών προσεγγίζουν την Κανονική Κατανομή τα αποτελέσματα της LDA είναι παρόμοια με αυτά της λογιστικής παλινδρόμησης. Όταν οι κλάσεις είναι καλά διαχωρισμένες, οι εκτιμήσεις των παραμέτρων με τη μέθοδο της λογιστικής παλινδρόμησης είναι ιδιαίτερα ασταθής ενώ με τη μέθοδο LDA δεν συμβαίνει αυτό. Επίσης, όταν το n είναι πολύ μικρό και η κατανομή των ανεξάρτητων μεταβλητών προσεγγίζουν τη Κανονική Κατανομή, το μοντέλο LDA είναι πιο σταθερό από αυτό της λογιστικής παλινδρόμησης.

3.3.3 Δένδρα απόφασης

Τα δέντρα απόφασης (decision tree) αποτελούν μια απλή αλλά πολύ σημαντική τεχνική κατηγοριοποίησης. Βασίζονται σε μια σειρά από ερωτήματα, τα οποία τίθενται σε κάθε εγγραφή για κάποιες από τις τιμές των χαρακτηριστικών της. Τα ερωτήματα αυτά και οι πιθανές απαντήσεις τους απεικονίζονται με τη μορφή ενός δένδρου. Ένα δένδρο αποτελείται από φύλλα, κόμβους και ακμές. Κάθε κόμβος αντιστοιχεί και σε ένα ερώτημα, ενώ κάθε ακμή που εξέρχεται από τον κόμβο αυτό αντιστοιχεί σε μια πιθανή απάντηση του ερωτήματος. Σε κάθε τελικό κόμβο, που καλείται και φύλλο, αντιστοιχεί μια τιμή της κλάσης. Η διαδικασία ταξινόμησης μιας εγγραφής αφού έχει κατασκευαστεί ένα δέντρο είναι απλή. Αρχίζοντας από τον πρώτο κόμβο, που καλείται και ρίζα, τίθεται ένα ερώτημα για κάποιο χαρακτηριστικό της. Ανάλογα με την απάντηση θα ακολουθήσουμε και την αντίστοιχη ακμή η οποία θα μας οδηγήσει σε ένα νέο κόμβο. Ακολουθώντας τους κόμβους και ανάλογα τα χαρακτηριστικά της νεο εισερχόμενης εγγραφής καταλήγουμε σε κάποιο φύλλο από το οποίο αντιστοιχούμε και την τιμή της κλάσης της εγγραφής. Ο αριθμός των πιθανών δέντρων απόφασης που μπορούν να προκύψουν από το ίδιο σύνολο χαρακτηριστικών είναι εκθετικός. Αυτό έχει σαν αποτέλεσμα να καθιστά την εύρεση της βέλτιστης λύσης μια υπολογιστικά μη πρακτική διαδικασία λόγω του εκθετικού χρόνου που απαιτείται. Ωστόσο, έχουν σχεδιαστεί αλγόριθμοι, οι οποίοι κατασκευάζουν αρκετά ακριβή δένδρα απόφασης με ανεκτό υπολογιστικό κόστος. Οι αλγόριθμοι αυτοί εφαρμόζουν μια άπληστη (greedy) στρατηγική, χτίζοντας το δέντρο απόφασης παίρνοντας μια σειρά από τοπικές βέλτιστες λύσεις. Οι άπληστοι αλγόριθμοι δεν δίνουν πάντα τη βέλτιστη λύση αλλά τη προσεγγίζουν. Με απλά λόγια, οι αλγόριθμοι αυτοί έχουν ως στόχο τη διάσπαση εγγραφών με βάση έναν έλεγχο γνωρίσματος βελτιστοποιώντας ένα κριτήριο. Η στρατηγική greedy προτιμά κόμβους με ομοιογενείς κατανομές κλάσεων, χρησιμοποιώντας μια μέτρηση της μη καθαρότητας του κόμβου, επιλέγει εκείνον ο οποίος έχει μικρό βαθμό μη καθαρότητας. Ο αλγόριθμος του

Hunt, αποτελεί τη βάση για πολλούς αλγόριθμους κατασκευής δέντρων οι οποίοι αναπτύχθηκαν αργότερα, όπως του ID3, του C4.5, του CART κ.α. Ο αλγόριθμος του Hunt κατασκευάζει ένα δέντρο αναδρομικά. Έστω λοιπόν, ότι έχουμε ένα σύνολο από εγγραφές εκπαίδευσης που βρίσκονται σε έναν κόμβο. Τα βασικά βήματα του αλγορίθμου συνοψίζονται παρακάτω

1. Άρχισε με έναν κόμβο ο οποίος θα αποτελεί τη ρίζα του δένδρου
2. Αν όλες οι εγγραφές που αντιστοιχούν στο κόμβο ανήκουν στην ίδια κλάση, θέσε τον κόμβο ως φύλλο και ανάθεσε του την τιμή της κλάσης αυτής.
3. Αν όλες οι εγγραφές που αντιστοιχούν στο κόμβο δεν ανήκουν στην ίδια κλάση, τότε αντιστοίχισε στον κόμβο ένα έλεγχο γνωρίσματος, με σκοπό τη διαμέριση των εγγραφών. Δημιούργησε μια ακμή για κάθε έλεγχο και διαμέρισε τις οντότητες στους κόμβους που δημιουργούνται αντίστοιχα
4. Επανέλαβε τα βήματα 2 και 3 αναδρομικά σε κάθε κόμβο που έχεις δημιουργήσει.

Ένα ζήτημα που αρκετές φορές μας απασχολεί είναι πως διακριτοποιούνται τα χαρακτηριστικά τα οποία έχουν συνεχές γνώρισμα. Η διακριτοποίηση συνεχών χαρακτηριστικών αποτελεί ένα σημαντικό ζήτημα της προεπεξεργασίας των δεδομένων και η ακρίβεια αρκετών αλγόριθμων ταξινόμησης εξαρτάται σε μεγάλο βαθμό από το πόσο κοντά στο βέλτιστο υλοποιείται η διαδικασία αυτή. Μια αποτελεσματική μέθοδος διακριτοποίησης συνεχών γνωρισμάτων είναι αυτή που χρησιμοποιεί το δείκτη GINI, την οποία χρησιμοποιεί ο αλγόριθμος CART, SLIQ και SPRINT. Όταν ένας κόμβος διασπάται σε k κόμβους, αυτό συνεπάγεται ότι οι εγγραφές του κόμβου έχουν χωριστεί σε k-υποσύνολα και η ποιότητα του διαχωρισμού υπολογίζεται ως

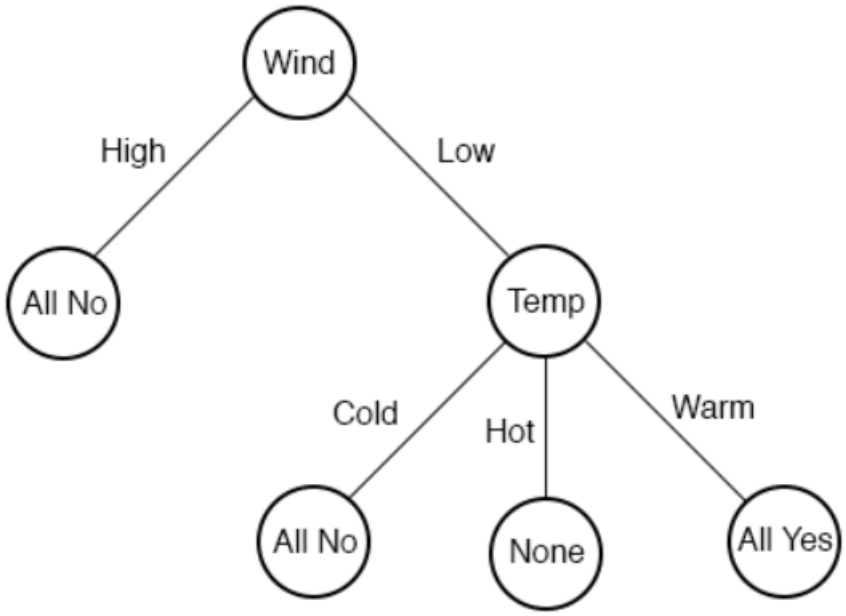
$$\mathrm{GINI(t)} = 1 - \sum_{j=1}^{c} [p(j|t)]^2$$

$p(j|t)$ = σχετική συχνότητα της κλάσης j στον κόμβο t

και c = ο αριθμός των κλάσεων

Ο δείκτης GINI παίρνει τιμές από μηδέν (όταν όλες οι τιμές ανήκουν σε μια κλάση) μέχρι μικρότερες της μονάδας. Τιμές κοντά στη μονάδα, σημαίνει ότι όλες οι εγγραφές είναι

ομοιόμορφα κατανεμημένες στις κλάσεις. Ταυτόχρονα αποτελεί και ένα μέτρο της διακύμανσης σε όλες τις τάξεις k.



Σχήμα 3.3: Δέντρο Απόφασης

Ένα δένδρο απόφασης αποτελείται από ένα σύνολο κανόνων, με κάθε κανόνα να παράγεται από ένα μονοπάτι του δένδρου, ξεκινώντας από τη ρίζα και καταλήγοντας σε κάθε φύλλο του δέντρου. Οι κανόνες που διέπουν το δέντρο της εικόνας 3.3 είναι οι ακόλουθοι:

WIND=High → Class=All No

WIND=Low → Temp=Warm → Class=All Yes

WIND=Low → Temp=Hot → Class=None

WIND=Low → Temp=Cold → Class=All No

Τα πλεονεκτήματα των δέντρων απόφασης είναι ότι μπορούν να εξηγηθούν με απλό τρόπο ακόμα και σε κάποιον που δεν έχει κάποια ιδιαίτερη σχέση με τις τεχνικές ταξινόμησης. Τα δέντρα απόφασης μπορούν εύκολα να χειριστούν συνεχείς μεταβλητές αφού δεν απαιτείται η δημιουργία ψευδομεταβλητών. Επίσης, μπορούν εύκολα να απεικονιστούν γραφικά κάνοντας την ερμηνεία τους αρκετά εύκολη. Ωστόσο, τα δέντρα απόφασης δεν έχουν ισάξια προβλεπτική ικανότητα όπως άλλες τεχνικές κατηγοριοποίησης.

3.3.4 Ταξινόμηση βάσει των στιγμιότυπων - K-Nearest Neighbors

Μια άλλη κατηγορία μοντέλων ταξινόμησης είναι αυτοί που βασίζονται στα στιγμιότυπα (instance-based classifiers). Οι ταξινομητές αυτοί χαρακτηρίζονται από την ιδιότητα να εκπαιδεύονται την ίδια στιγμή που ταξινομούν τις εγγραφές εισόδου. Το σύνολο εκπαίδευσης απλά αποθηκεύεται και όταν καλείται να ταξινομήσει μια νέα εγγραφή αυτή γίνεται βάσει των εγγραφών που έχουν αποθηκευτεί. Για τον λόγο αυτό οι μέθοδοι που βασίζονται στα στιγμιότυπα αναφέρονται και ως "τεμπέλικη" μάθηση (lazy learning). Συχνά, η διαδικασία ταξινόμησης είναι πιο χρονοβόρα από τη διαδικασία εκπαίδευσής τους. Η λογική στην οποία βασίζονται είναι ότι οι εγγραφές μπορούν να αναπαρασταθούν ως σημεία ενός χώρου, όπου κάθε μία από τις διαστάσεις αντικατοπτρίζει ένα από τα χαρακτηριστικά των εγγραφών. Η απόσταση δύο εγγραφών στο χώρο αντιστοιχεί σε έναν n-διάστατο Ευκλείδειο χώρο R^n, όπου n ο αριθμός των χαρακτηριστικών. Ως απόσταση ορίζεται μια συνάρτηση d, που πρέπει να ικανοποιεί τις παρακάτω ιδιότητες:

1. $d(x,y) \geq 0$ για κάθε x,y και αν $d(x,y)=0 \Leftrightarrow x=y$
2. $d(x,y) \leq d(x,z) + d(z,y)$ τριγωνική ανισότητα
3. $d(x,y)=d(y,x)$ συμμετρική ιδιότητα

Από ολόκληρο το σύνολο των εγγραφών δημιουργείται ένας συμμετρικός πίνακας που περιέχει τις αποστάσεις των στοιχείων x και y οι οποίες συμβολίζονται ως d(x,y). Οι x,y αποτελούν δυο ξεχωριστές εγγραφές του συνόλου. Η πιο συχνά χρησιμοποιούμενη μετρική απόσταση είναι η Ευκλείδεια. Η Ευκλείδεια απόσταση όμως εξαρτάται από την κλίμακα μέτρησης και όταν αυτό συμβαίνει μπορούμε να πάρουμε πολύ διαφορετικά αποτελέσματα. Μια μετρική η οποία λύνει το πρόβλημα αυτό είναι η απόσταση του Pearson αλλά και αυτή με τη σειρά της δεν λαμβάνει υπόψη της, τις συνδιακυμάνσεις μεταξύ των μεταβλητών. Μια τέτοια απόσταση είναι η απόσταση του Mahalanobis. Η απόσταση Manhattan, η Chebyshev και η απόσταση Minkowski είναι κάποια άλλα μέτρα απόστασης για συνεχή δεδομένα που χρησιμοποιούνται συχνά.

Μετρική	Τύπος
Pearson	$d_{i,j} = \sqrt{\sum_{r=1}^{p} (\frac{x_{ir} - x_{jr}}{s_r})^2}$
Chebyshev	$d_{i,j} = \max_{r=1,2,..p} \lvert x_{ir} - x_{jr} \rvert$
Manhattan	$d_{i,j} = \sum_{r=1}^{p} \lvert x_{ir} - x_{jr} \rvert$
Minkowski	$d_{i,j} = \left(\sum_{r=1}^{p} \lvert x_{ir} - x_{jr} \rvert^{\lambda} \right)^{1/\lambda}$

Πίνακας 3.1: Μετρικές αποστάσεις για συνεχείς μεταβλητές.

Οι αλγόριθμοι μάθησης οι οποίοι βασίζονται στα στιγμιότυπα είναι ο αλγόριθμος του κοντινότερου γείτονα (nearest neighbors), οι μηχανές kernel (kernel machines), της Rote μάθησης και της τοπικά σταθμισμένης παλινδρόμησης.

Ο πιο γνωστός αλγόριθμος βασιζόμενος στα στιγμιότυπα, είναι ο αλγόριθμος των k πλησιέστερων γειτόνων (k nearest neighbors) και ακολουθεί τα παρακάτω βήματα

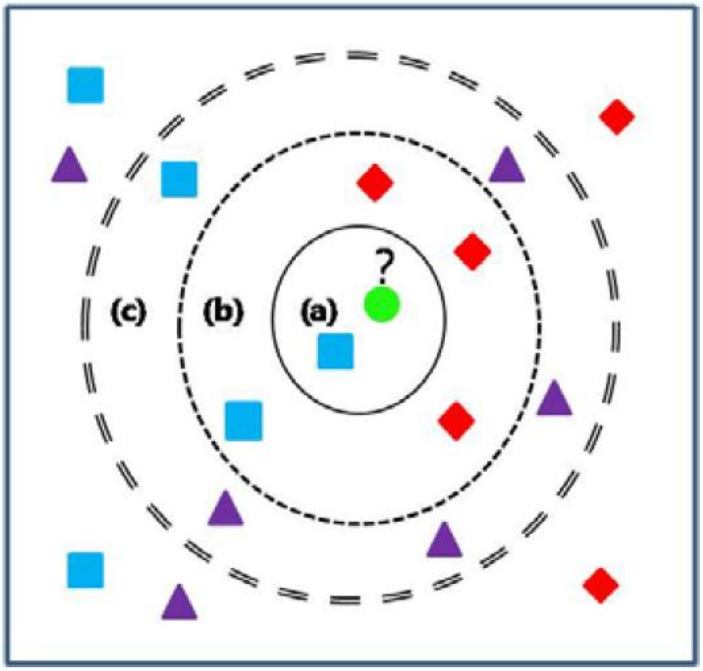
1. Για κάθε εγγραφή του συνόλου δοκιμής, βρες στο σύνολο εκπαίδευσης τις k πλησιέστερες εγγραφές της.
2. Ανάθεσέ της την πιο συχνά εμφανιζόμενη τιμή κλάσης.

Η μέθοδος του k-κοντινότερου γείτονα, είναι μια μέθοδος η οποία προσπαθεί να ταξινομήσει ένα νέο στοιχείο σε μια κλάση βασιζόμενο στην απόσταση εξετάζοντας τα k πλησιέστερα σημεία σε αυτό. Ένα νέο στοιχείο τοποθετείται στην κλάση που έχει την πλειοψηφία ανάμεσα στα κ κοντινότερα σημεία.

Δύο είναι οι παράμετροι οι οποίοι παίζουν σημαντικό ρόλο για τον τρόπο με τον οποίο θα γίνει η ταξινόμηση με τη μέθοδο του KNN. Η τιμή του k, δηλαδή ο αριθμός των πλησιέστερων σημείων για μια παρατήρηση x_0 και η απόσταση η οποία θα χρησιμοποιήσουμε για να βρούμε τις πλησιέστερες παρατηρήσεις. Ο αριθμός k συνήθως,

είναι ένας μικρός θετικός περιττός αριθμός ώστε να μην προκύπτει θέμα ισοψηφίας. Σε περίπτωση ισοψηφίας ο αλγόριθμος δεν θα μπορούσε να ταξινομήσει μια εγγραφή σε κάποια κλάση. Για μικρό αριθμό k το όριο απόφασης είναι αρκετά ευέλικτο με αποτέλεσμα να δημιουργεί μοτίβα τα οποία έχουν μικρή μεροληψία και μεγάλες αποκλίσεις, δηλαδή μεγάλες διακυμάνσεις στις προβλέψεις ενώ μεγάλος αριθμός του k έχει σαν αποτέλεσμα ο ταξινομητής να έχει μεγάλη μεροληψία και μικρές αποκλίσεις. Όσο αναφορά την απόσταση, η μετρική η οποία θα χρησιμοποιήσουμε εξαρτάται από τη φύση του προβλήματος και των δεδομένων. Επίσης, το σφάλμα στα δεδομένα εκπαίδευσης δεν σχετίζεται με το σφάλμα στα τεστ δεδομένα. Γενικά όσο πιο ευέλικτη γίνεται η μέθοδος ταξινόμησης, το σφάλμα στα δεδομένα εκπαίδευσης μπορεί να μειώνεται αλλά αυτό να μην συμβαίνει στα τεστ δεδομένα. Έχει παρατηρηθεί ότι στα τεστ δεδομένα, μέχρι ένα σημείο μειώνεται το σφάλμα αλλά από ένα σημείο και μετά, εκεί που η μέθοδος έχει αρχίζει να δείχνει μεγάλη μεροληψία, το σφάλμα αρχίζει σταδιακά να ανεβαίνει δημιουργώντας μια κυρτή γραμμή.

Στο σχήμα 3.4 που ακολουθεί δίνεται ένα παράδειγμα της προσέγγισης της μεθόδου του KNN και το πως αυτό επηρεάζεται από την επιλογή του k. Το σύνολο εκπαίδευσης αποτελείται από 16 παρατηρήσεις, 5 μπλε τετράγωνα, 5 κόκκινους ρόμβους και 6 μωβ τρίγωνα. Σκοπός μας είναι να ταξινομήσουμε τον πράσινο κύκλο σε μία από αυτές τις τάξεις ανάλογα με την τιμή του κ σε κάθε περίπτωση. Όπως φαίνεται στο σχήμα, για k=1, η μέθοδος του KNN θα εντοπίσει το κοντινότερο σημείο το οποίο είναι το μπλε τετράγωνο και είναι η περιοχή (α). Για k=5 έχουμε 2 μπλε τετράγωνα και 3 κόκκινους ρόμβους, περιοχή (β), με αποτέλεσμα οι συχνότητες να είναι 2/5 για την μπλε τάξη και 3/5 για την κόκκινη αντίστοιχα. Επομένως ο KNN αλγόριθμος θα προβλέψει ότι ανήκει στην κόκκινη τάξη. Αντίστοιχα για k=10 ο αλγόριθμος θα προβλέψει ότι ανήκει στην μωβ τάξη, περιοχή (c). Συμπεραίνουμε λοιπόν ότι η επιλογή του k παίζει καθοριστικό ρόλο στην ταξινόμηση ενός στοιχείου και μπορεί να επηρεάσει σε μεγάλο βαθμό την ποιότητα των αποτελεσμάτων. Όπως αναφέραμε και παραπάνω πρέπει η τιμή του κ να εξισορροπεί την μεροληψία με την διακύμανση του αλγορίθμου. Μια μέθοδος μέσω της οποίας μπορούμε να καθορίσουμε τον βέλτιστο αριθμό k είναι η μέθοδος Cross Validation η οποία είναι μια καθαρά υπολογιστική μέθοδος η οποία υπολογίζει για διάφορες τιμές του k την ακρίβεια του μοντέλου αφήνοντας κάθε φορά από τα δεδομένα εκπαίδευσης μια παρατήρηση την οποία και προσπαθεί να ταξινομήσει. Το k για το οποίο έχουμε την μεγαλύτερη ακρίβεια σωστών ταξινομήσεων αυτό και επιλέγεται.



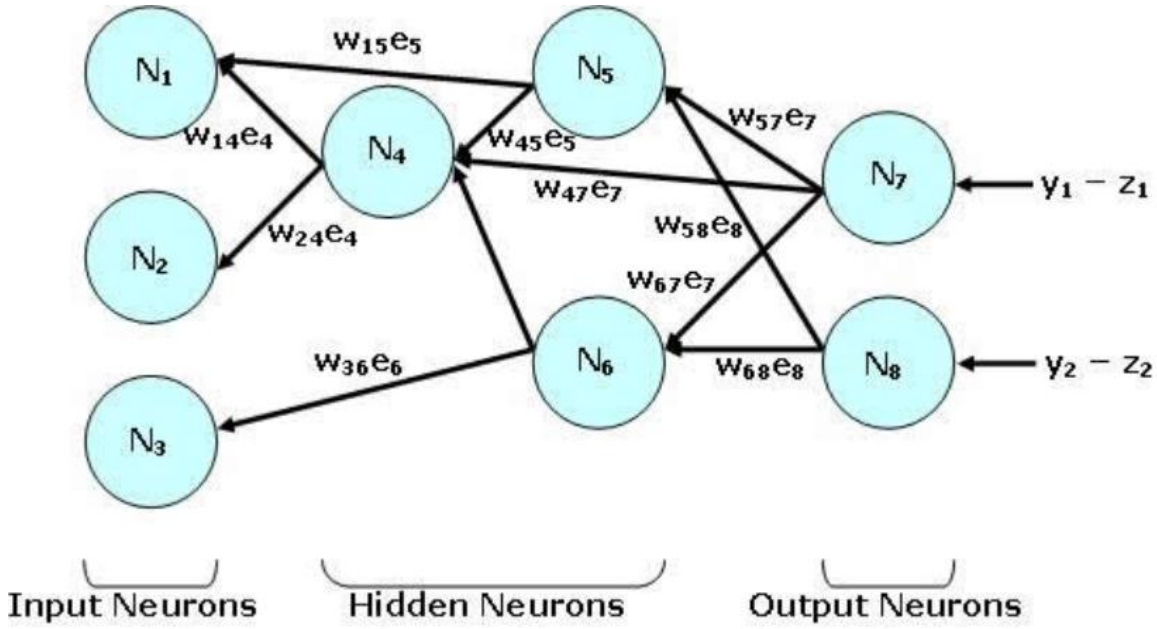
Σχήμα 3.4: Ταξινόμηση με τη μέθοδο k-κοντινότερου γείτονα

Συμπερασματικά μπορούμε να πούμε ότι ενώ η μέθοδος του KNN δεν βασίζεται σε κάποιο πιθανοθεωρητικό μοντέλο και ως εκ τούτου δεν λαμβάνει υπόψη του τη μεταβλητότητα και την πιθανή εξάρτηση μεταξύ των μεταβλητών. Παρόλα αυτά η μέθοδος αυτή δίνει αξιόπιστα αποτελέσματα και χρησιμοποιείται ευρέως σε διάφορα προβλήματα. Η αξιοπιστία του έγκειται στο ότι παράγει συχνά αποτελέσματα τα οποία είναι πολύ κοντά στη βέλτιστη ταξινόμηση με τη μέθοδο του Bayes η οποία βασίζεται σε πιθανοθεωρητικό μοντέλο.

3.3.5 Τεχνητά νευρωνικά δίκτυα

Το αντικείμενο των τεχνητών νευρωνικών δικτύων (artificial neural networks) έχει γνωρίσει ραγδαία εξέλιξη τις τελευταίες δεκαετίες. Αποτελούν ένα αυτόνομο πεδίο της επιστήμης που ανήκει στο ευρύτερο πλαίσιο της τεχνητής νοημοσύνης και των ευφυών συστημάτων. Τα τεχνητά νευρωνικά δίκτυα (ΤΝΔ) αποτελούν μια προσπάθεια προσέγγισης του ανθρώπινου εγκεφάλου. Οι εφαρμογές τους, καλύπτουν όλο το φάσμα των θετικών επιστημών και της μηχανικής από την αεροδιαστημική, τα αυτοκίνητα τις τραπεζικές εργασίες, τη ρομποτική ακόμη και μέχρι τα προβλήματα επιβλεπόμενης μάθησης στην αναγνώριση προτύπων. Το μοντέλο των νευρωνικών δικτύων που εξειδικεύεται στην ταξινόμηση καλείται perceptron. Το perceptron είναι από τα πιο απλά ΤΝΔ και χρησιμοποιείται για την ταξινόμηση γραμμικά διαχωριζόμενων συνόλων. Αποτελούνται από

ένα μόνο επίπεδο απλών νευρώνων, οι οποίοι λειτουργούν ως είσοδοι και έξοδοι του δικτύου ταυτόχρονα. Το perceptron ενός επιπέδου, αντιστοιχεί σε κάθε ένα χαρακτηριστικό βάρη $w_1, w_2, \ldots . w_n$, για κάθε εγγραφή εισόδου. Τα βάρη αυτά παίρνουν τιμές από -1 έως 1. Για κάθε εγγραφή εισόδου, υπολογίζεται ως έξοδος το άθροισμα $\sum_{k=1}^{n} w_k x_{i,k}$, της i-παρατήρησης για το k-χαρακτηριστικό. Με βάση τη τιμή αυτή και κάποιων τιμών κατωφλίου που έχουμε ορίσει, αποφασίζεται σε ποια κλάση ανήκει η εγγραφή. Η μέθοδος εκπαίδευσης γίνεται με την επαναληπτική εκτέλεση του αλγόριθμου στο σύνολο των εγγραφών εισόδου, μέχρι ο αλγόριθμος να ταξινομεί σωστά όλες τις εγγραφές εισόδου.



Σχήμα 3.5: Απεικόνιση ενός perceptron πολλών επιπέδων (back propagation algorithm).

Το απλό perceptron ταξινομεί σωστά μόνο γραμμικά διαχωρίσιμα σύνολα. Αν δεν είναι γραμμικά διαχωρίσιμα, το perceptron δεν θα καταφέρει ποτέ να ταξινομήσει σωστά όλες τις εγγραφές εισόδου. Στη περίπτωση αυτή χρησιμοποιούμε το perceptron πολλών επιπέδων. Ένα τέτοιο δίκτυο αποτελείται από ένα σύνολο κόμβων που αποτελούν το επίπεδο εισόδου, από ενδιάμεσα κρυφά επίπεδα υπολογιστικών κόμβων και από ένα επίπεδο υπολογιστικών κόμβων εξόδου. Στο perceptron πολλών επιπέδων, δεν υπάρχουν απευθείας συνδέσεις μεταξύ εισόδου και εξόδου. Όμως το δίκτυο είναι πλήρως διασυνδεδεμένο, δηλαδή ο νευρώνας του κάθε επιπέδου είναι πλήρως διασυνδεδεμένος με όλους τους νευρώνες του προηγούμενου επιπέδου. Το επίπεδο εισόδου στέλνει τα σήματα εισόδου σε όλους τους νευρώνες του κρυφού επιπέδου. Το κρυφό επίπεδο αποτελείται από μη γραμμικούς νευρώνες ενώ το

επίπεδο εξόδου από γραμμικούς και μη γραμμικούς νευρώνες. Κάθε εγγραφή θεωρείται και ένα perceptron. Εδώ θα αναφερθούμε στον αλγόριθμο μάθησης της πίσω διάδοσης (back propagation algorith). Τα βήματα ενός αλγόριθμου πίσω διάδοσης αποτελούνται από ένα πέρασμα εμπρός και από ένα πέρασμα πίσω. Ένα διάνυσμα εισόδου , δηλαδή μια εγγραφή, εφαρμόζεται στους νευρώνες εισόδου. Η επίδραση του διαδίδεται μέσα στο δίκτυο, από επίπεδο σε επίπεδο, παράγοντας ένα σύνολο από εξόδους ως τιμή της πραγματικής απόκρισης του δικτύου. Κατά τη διάρκεια του εμπρός περάσματος, στα βάρη του δικτύου δίνονται σταθερές αυθαίρετες τιμές. Κατά τη διάρκεια του πίσω περάσματος, η πραγματική απόκριση του δικτύου αφαιρείται από την επιθυμητή απόκριση για την παραγωγή ενός σφάλματος το οποίο διαδίδεται προς τα πίσω στο δίκτυο. Τα βάρη μεταβάλλονται ώστε να μειωθεί το σφάλμα και η πραγματική απόκριση να πλησιάζει την επιθυμητή. Συνήθως η διαδικασία τερματισμού του αλγορίθμου ορίζεται όταν το σφάλμα πέφτει κάτω από ένα όριο.

3.3.6 Support vector machines

Οι Μηχανές Διανυσμάτων Υποστήριξης-ΜΔΥ (Support Vector Machines-SVMs) είναι μια μέθοδος μάθησης για δυαδικά προβλήματα ταξινόμησης υποθέτωντας ότι υπάρχει γραμμική διαχωριστικότητα των δεδομένων. Ο αρχικός αλγόριθμος των ΜΔΥ αναπτύχθηκε από τον Vlatimir Vapnik και Alexey Chervonenkis το 1963. Πολύ αργότερα το 1992 οι Boser, Guyon και Vapnik πρότειναν ένα τρόπο για τη δημιουργία μη γραμμικών ταξινομητών των ΜΔΥ, ώσπου το 1995 οι Cortes και Vapnik θεμελίωσαν τα ΜΔΥ όπως αυτά είναι γνωστά έως σήμερα. Αποτελούν την πιο πρόσφατα αναπτυγμένη τεχνική κατηγοριοποίησης με αξιοσημείωτα αποτελέσματα απόδοσης και ως εκ τούτου έχει μεγάλη απήχηση σε εκείνους που χρησιμοποιούν τεχνικές κατηγοριοποίησης. Σκοπός των ΜΔΥ είναι η επιλογή ενός βέλτιστου υπερεπιπέδου το οποίο θα διαχωρίζει τα σημεία των δυο κλάσεων βέλτιστα. Ο αλγόριθμος προβάλλει τις παρατηρήσεις σε ένα χώρο διαστάσεων και βρίσκει το υπερεπίπεδο το οποίο διαχωρίζει βέλτιστα τα σημεία. Μια νέα παρατήρηση ταξινομείται σύμφωνα με την πλευρά του υπερεπιπέδου στην οποία βρίσκεται. Τα διανύσματα αυτά τα οποία καθορίζουν το υπερεπίπεδο λέγονται διανύσματα υποστήριξης. Ο αλγόριθμος επιλέγει ένα μικρό αριθμό παρατηρήσεων από κάθε κλάση τα οποία ορίζουν το μέγιστο περιθώριο μεταξύ των δυο κλάσεων. Η εξίσωση της επιφάνειας σύμφωνα με την οποία διαχωρίζονται οι δυο κλάσεις δίνεται από τον τύπο

$$w^T x + b = 0$$

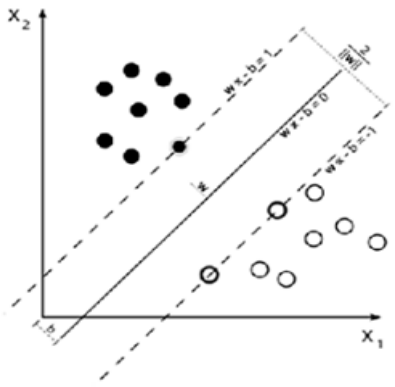
όπου το διάνυσμα w καθορίζει τη διεύθυνση και το b τη θέση στο χώρο του υπερεπιπέδου. Για ένα σύνολο δεδομένων εκπαίδευσης μπορούν να προκύψουν παραπάνω από ένα υπερεπίπεδα τα οποία να διαχωρίζουν τις δυο κλάσεις. Έτσι το πρόβλημα ανάγεται στην εύρεση του κατάλληλου διανύσματος w και του αριθμού b, ώστε

$$wx_i + b \geq 1$$

$$wx_i + b \leq -1, \text{ για κάθε διάνυσμα } x_i$$

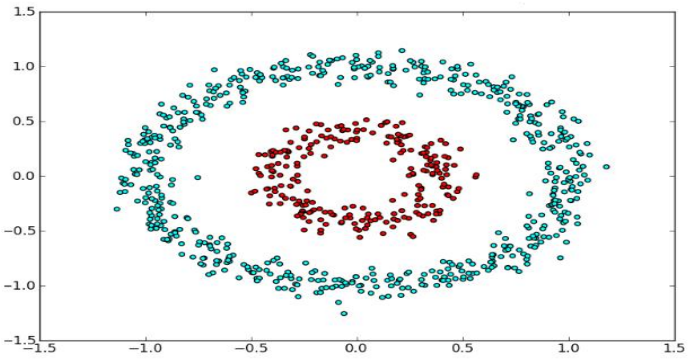
Αυτό έχει σαν αποτέλεσμα ο κανόνας ταξινόμησης να προκύπτει ανάλογα με το αν η ποσότητα $wx_i + b$ είναι μικρότερη ή μεγαλύτερη από το μηδέν, δηλαδή για c_i=1 ή -1. Άρα, $|\mathrm{w^T x + b}| = c_i(\mathrm{w^T x + b})$, όπου c_i=1 ή -1 ανάλογα τη θέση του σημείου στο χώρο. Το βέλτιστο υπερεπίπεδο θα έχει συνολική απόσταση μεταξύ των δυο κλάσεων ίση με $\frac{2}{||\mathrm{w}||}$. Επομένως, ο βέλτιστος διαχωρισμός προκύπτει από την επίλυση της ελαχιστοποίησης της νόρμας, $\frac{1}{2}\|w\|^2$ έτσι ώστε $c_i(wx_i + b) \geq 1$, για i=1,2…

Ο διαχωρισμός μεταξύ του υπερεπιπέδου και των πλησιέστερων σημείων ονομάζεται περιθώριο διαχωρισμού. Βέλτιστο υπερεπίπεδο θεωρείται αυτό που μεγιστοποιεί το περιθώριο διαχωρισμού. Ο λόγος για τον οποίο η μέθοδος επιβάλλει το μέγιστο περιθώριο διαχωρισμού αφορά την κατηγοριοποίηση των τεστ δεδομένων, αφού παρέχεται καλύτερη δυνατότητα ταξινόμησης σε παρατηρήσεις οι οποίες θα διαφέρουν από τα δεδομένα εκπαίδευσης. Τα σημεία τα οποία βρίσκονται οριακά στα δυο υπερεπίπεδα αποτελούν τα διανύσματα υποστήριξης που αντιστοιχούν στο κάθε υπερεπίπεδο.



Σχήμα 3.6: Απεικόνιση των support vectors.

Έτσι λοιπόν, η πολυπλοκότητα ταξινόμησης των ΜΔΥ είναι ανεξάρτητη του πλήθους των δεδομένων εισόδου και του πλήθους των χαρακτηριστικών τους. Η διαδικασία της εκπαίδευσης είναι χρονοβόρα αφού απαιτείται $O(n^2)$. Από την άλλη, η πολυπλοκότητα εκπαίδευσής του κυριαρχείται από το χρόνο επίλυσης δεν θεωρείται αποδοτικός για τεράστιο όγκο δεδομένων εισόδου. Η επίλυση του τετραγωνικού προγράμματος καταλήγει πάντα σε μια βέλτιστη λύση, με αποτέλεσμα να μην υπάρχει κίνδυνος σύγκλισής σε τοπικά βέλτιστα, όπως συμβαίνει με την εκπαίδευση άλλων αλγόριθμων κατηγοριοποίησης. Όπως αναφέραμε και προηγουμένως, όλα τα παραπάνω προϋποθέτουν ότι στα δεδομένα εκπαίδευσης υπάρχει η γραμμική διαχωριστικότητα των δεδομένων. Τι γίνεται όμως εάν αυτή η προϋπόθεση δεν ισχύει; Για παράδειγμα, τι αποτελέσματα θα είχαμε με τις ΜΔΥ αν είχαμε δεδομένα εκπαίδευσης όπως αυτά του σχήματος 3.7. Προφανώς, στην περίπτωση αυτή, οι ΜΔΥ δεν θα μπορούσαν να διαχωρίσουν τις δυο αυτές κλάσεις αφού η γραμμική διαχωριστικότητα δεν ικανοποιείται.



Σχήμα 3.7: SVM για μη γραμμικά διαχωρίσιμα δεδομένα

Το παραπάνω πρόβλημα μπορεί να αντιμετωπιστεί με τη βοήθεια μιας μεταβλητής ξ, η οποία θα επιτρέπει ένα μικρό περιθώριο για κάποιες λανθασμένες ταξινομήσεις. Ο πραγματικός, μη αρνητικός αριθμός ξ, χαλαρώνει τους περιορισμούς και αναγάγει το πρόβλημα τετραγωνικού προγραμματισμού στη μορφή

$$c_i(w^T x_i + b) \geq 1 - \xi_i$$

Αν η $\xi_i > 1$, τότε η παρατήρηση x_i ταξινομείται σε λάθος κλάση. Αυτό συνεπάγεται ότι το σύνολο των μεταβλητών ξ είναι το άνω όριο του συνόλου των δεδομένων που επιτρέπεται να ταξινομηθούν λάθος. Έτσι, η ποσοότητα η οποία χρειάζεται να ελαχιστοποιήσουμε για να βρούμε το διαχωριστικό επίπεδο μετατρέπεται σε

$$\frac{1}{2}||\mathrm{w}||^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{P}} \xi_{\iota}$$

Η σταθερά C, προσδιορίζει το βάρος του κόστους που εμείς θέλουμε να δώσουμε στις λάθος ταξινομήσεις. Έτσι, για μεγάλες τιμές του C δίνουμε μεγάλη βαρύτητα στις λάθος ταξινομήσεις ενώ για τιμές κοντά στο μηδέν όχι. Μια άλλη λύση είναι να απεικονιστούν οι εγγραφές σε ένα χώρο περισσότερων διαστάσεων (feature space) και να βρεθεί εκεί ένα υπερεπίπεδο διαμέρισης. Ο βέλτιστος διαχωρισμός μπορεί να επιτευχθεί σε μη γραμμικά υπερεπίπεδα. Ο νέος αυτός χώρος καλείται μετασχηματισμένος χώρος γνωρισμάτων. Ο μετασχηματισμός αυτός δε γίνεται κατευθείαν αλλά μέσω μιας συνάρτησης πυρήνα (kernel function), η οποία μετασχηματίζει το εσωτερικό γινόμενο όλων των δυνατών εγγραφών στο νέο χώρο. Στόχος είναι μέσω της συνάρτησης πυρήνα να τα προβάλουμε σε ένα χώρο μεγαλύτερης διάστασης ευελπιστώντας να πετύχουμε ένα γραμμικό διαχωρισμό με λιγότερα σφάλματα. Η κατάλληλη επιλογή της συνάρτησης πυρήνα καθορίζει σε σημαντικό βαθμό τα αποτελέσματα της ταξινόμησης. Έστω x_1 και x_2 τα διανύσματα γνωρισμάτων δυο οντοτήτων. Οι πιο σημαντικές συναρτήσεις πυρήνα είναι:

Πολυωνυμική βαθμού p: $K(x_1, x_2) = (x_1 x_2 + 1)^p$, p=1 μετατρέπεται σε γραμμική

Γκαουσιανή RBF: $K(x_1, x_2) = e^{-\|x_1 - x_2\|^2/2\sigma^2}$

Σιγμοειδής βαθμού p: $K(x_1, x_2) = \tan h(kx_1 x_2 - \delta)^p$

3.3.7 Απλός ταξινομητής Naive-Bayes

Στο σημείο αυτό θα περιγράψουμε τις γενικές σχέσεις που διέπουν ένα σύνολο εγγραφών, αξιοποιώντας τις συσχετίσεις και τις ανεξαρτησίες μεταξύ των χαρακτηριστικών των εγγραφών. Ο ταξινομητής Bayes είναι ένα πιθανοθεωρητικό μοντέλο και χρησιμοποιείται για να εκτιμήσουμε την πιθανότητα μια παρατήρηση να ανήκει σε μια από τις προκαθορισμένες κλάσεις. Για την εκτίμηση των πιθανοτήτων αυτών χρησιμοποιούμε το θεώρημα του Bayes σύμφωνα με το οποίο

$$\Pr[\mathrm{H}|\mathrm{E}] = \frac{\Pr[\mathrm{E}|\mathrm{H}]\Pr[\mathrm{H}]}{\Pr[\mathrm{E}]}$$

Pr[H] : εκ των προτέρων πιθανότητα του H, η πιθανότητα του γεγονότος χωρίς επίκληση μαρτυρίας

$\Pr[E|H]$: εκ των υστέρων πιθανότητατου Η, η πιθανότητα του γεγονότος με την επίκληση μαρτυρίας

Με τον όρο "μαρτυρία" εννοούμε την εγγραφή στα δεδομένα εκπαίδευσης ενώ με τον όρο "γεγονός" εννοούμε την κλάση της εγγραφής. Ο αλγόριθμος του Bayes προϋποθέτει δυο παραδοχές, Πρώτον, ότι τα χαρακτηριστικά είναι εξίσου σημαντικά. Δεύτερον, ότι είναι στατιστικά ανεξάρτητα μεταξύ τους δεδομένης της κλάσης των εγγραφών, δηλαδή δεδομένου του χαρακτηριστικού μιας παρατήρησης δεν μπορούμε να ισχυριστούμε τίποτα για ένα άλλο χαρακτηριστικό. Προϋποθέτοντας λοιπόν την υπό συνθήκη ανεξαρτησία η παραπάνω σχέση γίνεται

$$\Pr[H|E] = \frac{\Pr[E_1|H]\ \Pr[E_2|H] \ldots . \Pr[E_n|H]\ \Pr[H]}{\Pr[E]}$$

Έτσι λοιπόν, για τα δεδομένα εκπαίδευσης, εκτιμάμε την εκ των προτέρων πιθανότητα $\Pr[H]$ υπολογίζοντας απλά την αναλογία της κάθε κλάσης στα δεδομένα εκπαίδευσης. Η πιθανότητα $\Pr[E]$ είναι σταθερή για όλες τις κλάσεις. Τέλος, οι πιθανότητες $\Pr[E_1|H]$, $\Pr[E_2|H]$, ... ισούνται με την αναλογία του κάθε χαρακτηριστικού δεδομένου της κλάσης την οποία υπολογίζουμε. Αυτό πρέπει να γίνει για όλα τα χαρακτηριστικά και για όλες τις τιμές των χαρακτηριστικών. Μια νέα παρατήρηση, θα ανήκει στην κλάση στην οποία έχει την μεγαλύτερη εκ των υστέρων πιθανότητα. Η κατηγοριοποίηση του Naïve-Bayes αν και είναι ιδιαίτερα απλοική δίνει καλά αποτελέσματα ακόμη και αν στην πράξη δεν ισχύει κάποια από τις προυποθέσεις που απαιτούνται. Η χρήση του Naïve-Bayes για ταξινόμηση είναι εξαιρετικά αποδοτική από άποψη χρόνου, τόσο στο στάδιο της εκπαίδευσης του μοντέλου, όσο και σε αυτό της εφαρμογής του στο σύνολο δοκιμής. Ακόμα, χαρακτηρίζεται για την ανοχή του στο θόρυβο, τα outliers, και την απουσία δεδομένων. Από την άλλη, το Naïve-Bayes μοντέλο βασίζεται σε κάποιες απλουστευτικές παραδοχές για τις οποίες και κατακρίνεται. Το Naïve-Bayes μοντέλο δεν δίνει ακριβείς εκτιμήσεις κατανομών για μη παραμετρικά συνεχή χαρακτηριστικά και η ακρίβεια ταξινόμησής του μειώνεται όταν δεν πληρείται η ανεξαρτησία των χαρακτηριστικών. Στην πράξη έχει αποδειχτεί, πως ο Naïve-Bayes ταξινομητής είναι αρκετά ακριβής ακόμα και υπό συνθήκες χαλάρωσης των δύο βασικών παραδοχών του αρκεί, είτε τα συνεχή χαρακτηριστικά να μην αποκλίνουν σημαντικά από την κανονική κατανομή, είτε να μην είναι μεταξύ τους ισχυρά συσχετισμένα. Για το δεύτερο, αυτό συμβαίνει γιατί η κατηγοριοποίηση του Naïve-Bayes δεν χρειάζεται να

δίνει ακριβείς εκτιμήσεις πιθανοτήτων, αρκεί η μέγιστη πιθανότητα να αντιστοιχίζεται στη σωστή κλάση.

3.3.8 Bagging

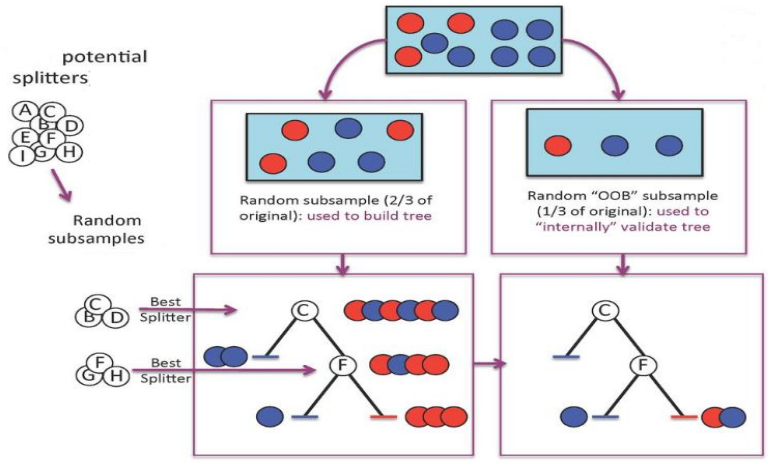
Η μέθοδος bagging ανήκει σε μια κατηγορία μεθόδων, όπως και οι μέθοδοι random forest και boosting που θα δούμε παρακάτω, οι οποίες εκμεταλλεύονται την αστάθεια που παρουσιάζουν ορισμένοι αλγόριθμοι μάθησης. Σε αυτές τις μεθόδους οι ταξινομητές προκύπτουν από την εκπαίδευση ενός αλγορίθμου σε υποσύνολα των δεδομένων εκπαίδευσης. Η μέθοδος bagging χρησιμοποιεί τα δέντρα απόφασης για την κατασκευή πιο ισχυρών μοντέλων πρόβλεψης και χρησιμοποιείται όταν το σύνολο των δεδομένων εκπαίδευσης είναι μικρό και όταν στα δεδομένα εκπαίδευσης υπάρχει μεγάλη μεταβλητότητα. Όπως αναφέραμε παραπάνω, τα δέντρα απόφασης έχουν ένα ισχυρό μειονέκτημα αφού πάσχουν από μεγάλη μεταβλητότητα. Αυτό έχει σαν αποτέλεσμα, ότι αν χωρίσουμε τα δεδομένα εκπαίδευσης σε δύο υποκατηγορίες και εφαρμόσουμε δέντρα απόφασης τα αποτελέσματα μπορεί να είναι πολύ διαφορετικά λόγο της υψηλής μεταβλητότητας. Αντίθετα αν είχαμε εφαρμόσει μια τεχνική με χαμηλή μεταβλητότητα σε δύο υποκατηγορίες τα αποτελέσματα θα ήταν παρόμοια. Η μέθοδος bagging έχει ως στόχο τη μείωση της μεταβλητότητας. Υποθέτουμε ότι έχουμε Y_1, Y_2,...., Y_n παρατηρήσεις με διακύμανση ίση με σ^2, τότε η διακύμανση του μέσου $\overline{Y}$ ισούται με $\frac{\sigma^2}{n}$, δηλαδή κατα μέσο όρο το σύνολο των παρατηρήσεων μειώνει τη διακύμανση, επομένως ένας απλος τρόπος για να μειώσουμε τη διακύμανση και ως εκ τούτου να αυξήσουμε την ακρίβεια της πρόβλεψης είναι να δημιουργήσουμε πολλά υποσύνολα δεδομένων εκπαίδευσης, όπου σε καθένα από αυτά θα εφαρμόσουμε τον αλγόριθμο ταξινόμησης. Κάθε υποσύνολο εκπαίδευσης παράγεται επιλέγοντας n-φορές ένα δείγμα από το αρχικό σύνολο. Η επιλογή γίνεται με επανατοποθέτηση σύμφωνα με την Ομοιόμορφη Κατανομή. Αυτό έχει σαν αποτέλεσμα κάποια από τα δεδομένα εκπαίδευσης να επιλέγονται παραπάνω από μια φορά και άλλα να μην επιλέγονται καθόλου. Κάθε υποσύνολο που έχουμε δημιουργήσει χρησιμοποιείται ως είσοδος στον αλγόριθμο. Επομένως έχουμε δημιουργήσει μια σειρά από διαφορετικά μοντέλα τα οποία θα χρησιμοποιήσουμε για την πρόβλεψη της κλάσης της εξαρτημένης μεταβλητής. Σε κάθε μοντέλο δίνεται ίση βαρύτητα. Στη συνέχεια καταγράφουμε την κλάση στην οποία έχει ταξινομηθεί σε κάθε ένα από τα υποσύνολα εκπαίδευσης και η τελική πρόβλεψη θα είναι αυτή που εμφανίζεται στα περισσότερα μοντέλα (μέθοδος πλειοψηφίας). Ο αριθμός των

επαναλαμβανόμενων υποσυνόλων που θα δημιουργήσουμε από τα δεδομένα εκπαίδευσης είναι μια σημαντική παράμετρος της μεθόδου bagging. Δεν συνεπάγεται ότι όσο μεγαλύτερος είναι ο αριθμός τόσο καλύτερη θα είναι η προσαρμογή του μοντέλου. Στην πράξη χρησιμοποιούμε ένα μεγάλο αριθμό τέτοιο ώστε να μειώνεται το σφάλμα. Υπάρχει ένας απλός τρόπος που επιβεβαιώνει τη μείωση του σφάλματος με τη μέθοδο αυτή χωρίς να χρειάζεται η εφαρμογή του μοντέλου σε τεστ δεδομένα. Κάθε υποσύνολο των δεδομένων εκπαίδευσης χρησιμοποιεί τα 2/3 των παρατηρήσεων. Το υπόλοιπο 1/3 αναφέρονται ως out-of-bag (OOB) και χρησιμοποιούνται για την πρόβλεψη της i-οστής παρατήρησης χρησιμοποιώντας κάθε φορά ένα από τα δέντρα στα οποία η παρατήρηση αυτή ήταν OOB. Έστω ότι έχουμε κατασκευάσει n-υποσύνολα δεδομένων εκπαίδευσης. Έτσι για την παρατήρηση i έχουμε n/3 διαθέσιμες προβλέψεις. Προκειμένου να έχουμε μια ενιαία πρόβλεψη για την i-παρατήρηση θα χρησιμοποιήσουμε την πλειοψηφία των προβλέψεων. Αυτό οδηγεί σε μια OOB πρόβλεψη για την i-παρατήρηση. Μια OOB πρόβλεψη μπορεί να εξαχθεί για καθεμία από τις n-παρατηρήσεις, για τις οποίες μπορεί να υπολογιστεί το OOB σφάλμα. Το OOB σφάλμα, είναι μια έγκυρη εκτίμηση σφάλματος για τη μέθοδο bagging αφού οι μεταβλητές απόκρισης που έχουν χρησιμοποιηθεί για την πρόβλεψη της i-παρατήρησης, χρησιμοποιούν τα δέντρα τα οποία δεν πήραν μέρος σε αυτή την παρατήρηση. Το μειονέκτημα της μεθόδου bagging είναι πως χάνουμε το βασικό πλεονέκτημα των δέντρων απόφασης που είναι η εύκολη ερμηνεία των αποτελεσμάτων. Ένα από τα πλεονεκτήματα της μεθόδου, είναι ότι δεν είναι υποκύπτει σε υπερεκπαίδευση του αλγορίθμου και ότι παράγει καλύτερα αποτελέσματα στην περίπτωση θορύβου. Με τη μέθοδο αυτή βελτιώνουμε την ακρίβεια του μοντέλου αλλά δεν μπορούμε να ερμηνεύσουμε εύκολα το μοντέλο. Αυτό οφείλεται στο ότι δεν μπορούμε να αναπαραστήσουμε τα αποτελέσματα οπτικά με ένα δέντρο αλλά ούτε και να διαπιστώσουμε ποιες από τις μεταβλητές ήταν σημαντικές για το μοντέλο. Συμπερασματικά, μπορούμε να πούμε ότι κερδίζουμε σε ακρίβεια αλλά χάνουμε σε επεξηγηματικότητα. Μπορούμε όμως να αποκτήσουμε μια γενική εικόνα της σημαντικότητας των επεξηγηματικών μεταβλητών με το δείκτη GINI, συγκρίνοντας το συνολικό ποσό που ο δείκτης GINI μειώνεται λόγω των διασπάσεων των δέντρων για ένα συγκεκριμένο χαρακτηριστικό.

3.3.9 Random Forest-Τυχαίο Δάσος

Η μέθοδος random forest αποτελεί μια βελτίωση της μεθόδου bagging. Και εδώ κατασκευάζουμε δέντρα απόφασης για υποσύνολα των δεδομένων εκπαίδευσης αλλά κάθε

φορά θεωρούμε ότι έχουμε μια διάσπαση του δέντρου. Σε κάθε διάσπαση ένα τυχαίο σύνολο από p-χαρακτηριστικά επιλέγονται από το αρχικό σύνολο των n-χαρακτηριστικών. Ο αριθμός των χαρακτηριστικών που επιλέγουμε είναι ίσος με $p \approx \sqrt{n}$, δηλαδή ίσος με τη τετραγωνική ρίζα του συνολικού αριθμού των χαρακτηριστικών. Στην ουσία ο αλγόριθμος δεν επιτρέπει σε κάθε διάσπαση του δέντρου να εξετάζεται το σύνολο των διαθέσιμων χαρακτηριστικών. Αυτό συμβαίνει για τον εξής λόγο. Αν υποθέσουμε ότι υπάρχει ένα χαρακτηριστικό το οποίο είναι πολύ ισχυρός προγνωστικός παράγοντας και ότι τα υπόλοιπα χαρακτηριστικά έχουν μέτρια προγνωστική σημασία. Τότε σε κάθε υποσύνολο, σχεδόν όλα τα δέντρα θα χρησιμοποιούν αυτό ως ρίζα του δέντρου. Συνεπώς όλα τα δέντρα σε κάθε υποσύνολο θα είναι παρόμοια μεταξύ τους με αποτέλεσμα οι προβλέψεις των δέντρων σε κάθε υποσύνολο να είναι ισχυρά συσχετισμένες μεταξύ τους κάτι το οποίο δεν θα οδηγήσει στη μείωση της διακύμανσης που είναι και ο κύριος σκοπός μας. Το πρόβλημα αυτό αντιμετωπίζεται με τη μέθοδο random forest αφού σε κάθε διάσπαση επιλέγεται ένα υποσύνολο των χαρακτηριστικών. Με αυτό το τρόπο, περίπου $(p - m)/p$ των δέντρων δεν θα εξετάσουν τον ισχυρό παράγοντα και έτσι τα υπόλοιπα χαρακτηριστικά θα μπορέσουν να αξιολογηθούν ως προς την συνεισφορά τους στο μοντέλο αυξάνοντας έτσι την αξιοπιστία του μοντέλου αφού με αυτό τον τρόπο αντιμετωπίζεται με αποτελεσματικό τρόπο η διακύμανση. Η κύρια διαφορά μεταξύ της μεθόδου bagging και random forest είναι η επιλογή των m-χαρακτηριστικών σε κάθε υποσύνολο. Για m=n τότε η μέθοδος random forest θα δώσει τα ίδια αποτελέσματα με τη μέθοδο bagging. Έτσι λοιπόν όταν στα δεδομένα υπάρχει ένας μεγάλος αριθμός συσχετισμένων χαρακτηριστικών είναι χρήσιμο να επιλέγουμε μια μικρή τιμή για το m.



Σχήμα 3.8: Απεικόνιση αλγορίθμου Random Forest

Στο σχήμα 3.8 απεικονίζεται ο αλγόριθμος του τυχαίου δάσους. Όπως φαίνεται από το σχήμα, τα δεδομένα μας ταξινομούνται σε 2 κλάσεις, την κόκκινη και την μπλε. Επίσης, κάθε παρατήρηση καθορίζεται από 9 χαρακτηριστικά, A,B,C,D,E,F,G,H,I. Όπως και με τη μέθοδο bagging, έτσι και εδώ, δημιουργούμε υποσύνολα των δεδομένων εκπαίδευσης χρησιμοποιώντας τα 2/3 για την κατασκευή του δέντρου απόφασης. Το υπόλοιπο 1/3 χρησιμοποιείται ως out-of-bag και από αυτό θα υπολογίσουμε την ακρίβεια του μοντέλου. Για τη ρίζα του δέντρου, επιλέγουμε τυχαία $3 = \sqrt{9}$, από τα διαθέσιμα χαρακτηριστικά, αυτά είναι τα B,C,D χαρακτηριστικά. Το C χαρακτηριστικό είναι αυτό που τελικά επιλέχθηκε για τη ρίζα του δέντρου. Για τη δημιουργία του επόμενου κόμβου, επιλέξαμε τυχαία πάλι 3 από τα διαθέσιμα χαρακτηριστικά, τα G,F,H και τελικά επιλέγει το F χαρακτηριστικό. Μετά την κατασκευή του δέντρου, ελέγχεται η ακρίβεια του ταξινομητή από το OOB σύνολο. Και σε αυτή τη μέθοδο, ισχύει ο κανόνας της πλειοψηφίας για την πρόβλεψη μιας νέας εγγραφής. Κάθε νέα εγγραφή ανατίθεται στη κλάση με τη μεγαλύτερη συχνότητα. Τα δέντρα με τη μέθοδο αυτή αναπτύσσονται πλήρως, στο μέγιστο μέγεθος τους, χωρίς τη τεχνική του κλαδέματος.

Τα πλεονεκτήματα των τυχαίων δασών είναι ότι μπορούν να αποδώσουν εξίσου καλά και σε δεδομένα υψηλών διαστάσεων όπως είναι οι εικόνες και τα κείμενα. Λόγω του μεγάλου πλήθους των δέντρων στο δάσος δεν αντιμετωπίζει το φαινόμενο της υπερεκπαίδευσης ενώ ταυτόχρονα παρουσιάζει ανεκτικότητα σε περίπτωση ελλιπών δεδομένων. Επίσης, η δημιουργία ενός δέντρου από τη ρίζα έως τα φύλλα του δέντρου πραγματοποιείται σε λογαριθμικό χρόνο ως προς το πλήθος των φύλλων του.

3.3.10 Boosting

Μια ακόμη μέθοδος η οποία βελτιώνει τη προβλεπτική ικανότητα των δέντρων απόφασης είναι η μέθοδος boosting. Στις προηγούμενες μεθόδους κατασκευάζαμε πολλαπλά υποσύνολα από τα αρχικά δεδομένα εκπαίδευσης και σε κάθε ένα από αυτά εφαρμόζαμε ένα δέντρο απόφασης. Κάθε δέντρο απόφασης εφαρμοζόταν σε ένα υποσύνολο ανεξάρτητα από τα άλλα δέντρα. Για τη δημιουργία του τελικού μοντέλου συνυπολογίζονταν όλα τα δέντρα απόφασης. Η μέθοδος boosting λειτουργεί σχεδόν με τον ίδιο τρόπο αλλά με τη διαφορά ότι χρησιμοποιεί τη γνώση που έχει προκύψει από τα προηγούμενα δέντρα. Δηλαδή τα δέντρα δεν εφαρμόζονται παράλληλα και ανεξάρτητα αλλά διαδοχικά και κάθε δέντρο εφαρμόζεται στο σύνολο των δεδομένων εκπαίδευσης. Αρχικά, ο αλγόριθμος ταξινομεί όλα τα δεδομένα

εκπαίδευσης στα οποία δίνει ίση βαρύτητα, η οποία ισούται με w=1/n (n ο αριθμός των δεδομένων εκπαίδευσης). Έτσι στο πρώτο βήμα απλά εφαρμόζεται ένας αδύναμος ταξινομητής στα αρχικά δεδομένα εκπαίδευσης. Το μοντέλο αυτό ταξινομεί όλες τις παρατηρήσεις των δεδομένων εκπαίδευσης σε κάποια κλάση. Στη συνέχεια τα βάρη αναπροσαρμόζονται. Για τις παρατηρήσεις που ταξινομήθηκαν σωστά τα βάρη μειώνονται ενώ για τις εσφαλμένες ταξινομήσεις τα βάρη αυξάνονται. Με αυτό το τρόπο ο ταξινομητής επικεντρώνεται στις εσφαλμένες ταξινομήσεις. Η διαδικασία αυτή επαναλαμβάνεται έως τη δημιουργία ενός προκαθορισμένου Β-αριθμού μοντέλων. Μοντέλα που το σφάλμα τους είναι μηδέν ή πολύ μεγάλα απορρίπτονται. Τα πρώτα γιατί θεωρούνται ότι πάσχουν από το φαινόμενο της υπερεκπαίδευσης και τα δεύτερα γιατί αδυνατούν να ταξινομήσουν τις εναπομείναντες παρατηρήσεις. Τα μοντέλα τα οποία δεν έχουν απορρριφτεί, χρησιμοποιούνται για την τελική ταξινόμηση. Για την ταξινόμηση των τεστ δεδομένων συμμετέχουν όλα τα μοντέλα που έχουν κατασκευαστεί αλλά το κάθε ένα έχει διαφορετική βαρύτητα ανάλογα με το σφάλμα του μοντέλου.

Η μέθοδος boosting καθορίζεται από 3 παραμέτρους:

Ο αριθμός των δέντρων Β. Η μέθοδος boosting, πάσχει από το φαινόμενο της υπερεκπαίδευσης όταν το Β είναι αρκετά μεγάλο αν και αυτό τείνει να συμβαίνει σταδιακά ή σχεδόν καθόλου. Για την επιλογή του Β χρησιμοποιούμε τη μέθοδο του cross validation.

Ο αριθμός συρρίκνωσης λ, ένας μικρός θετικός αριθμός ο οποίος ελέγχει τον ρυθμό με τον οποίο η μέθοδος εκπαιδεύεται. Λαμβάνει τιμές από 0.01 εως 0.001 και η σωστή επιλογή εξαρτάται από τα δεδομένα και τις τιμές εισόδου. Πολύ μικρή τιμή του λ απαιτεί πολύ μεγάλη τιμή του Β για καλύτερη προσαρμογή του μοντέλου.

Ο αριθμός δ των διασπάσεων που γίνονται σε κάθε δέντρο. Για δ=1 συνήθως λειτουργεί καλά, σε αυτή τη περίπτωση κάθε δέντρο αποτελείται από μόνο μια διάσπαση. Το δ καθορίζει το μέγεθος της αλληλεπίδρασης βάθους (interaction depth) και ελέγχει το βαθμό αλληλεπίδρασης μεταξύ των μεταβλητών στο τελικό μοντέλο, αφού οι δ-διασπάσεις μπορούν να περιέχουν το πολύ δ-χαρακτηριστικά.

Γενικά, οι μέθοδοι bagging, random forest και boosting βελτιώνουν την ακρίβεια ενός μοντέλου όταν χρησιμοποιούν ασταθείς αλγόριθμους όπως τα νευρωνικά δίκτυα και τα

δέντρα απόφασης. Για αλγόριθμους, όπως αυτός του k-πλησιέστερου γείτονα και του Naïve-Bayes οι οποίοι είναι ιδιαίτερα σταθεροί δεν παράγουν ικανοποιητικά αποτελέσματα.

ΚΕΦΑΛΑΙΟ 4

Υλοποίηση Αλγορίθμων Ταξινόμησης

4.1 Εισαγωγή-Περιγραφή Challenge

Στην παρούσα διπλωματική σκοπός μας είναι να εκμεταλλευτούμε τη δυνατότητα των ανοιχτών διασυνδεδεμένων δεδομένων και τις πληροφορίες που αυτά παρέχουν και την εφαρμογή μεθόδων κατηγοριοποίησης με σκοπό να προβλέψουμε αν μια ταινία είναι "καλή" ή "κακή". Οι δύο αυτές κλάσεις για τις ταινίες έχουν δημιουργηθεί βάσει των αξιολογήσεων των κριτικών που υποβλήθηκαν στην ιστοσελίδα Metacritic. Με βάση το σύστημα αξιολόγησης που χρησιμοποιεί η συγκεκριμένη ιστοσελίδα μια ταινία μπορεί να χαρακτηριστεί ως θετική, αρνητική ή μεικτή. Στη παρούσα μελέτη όμως μόνο δύο κλάσεις θα απαιτηθούν. Μια ταινία θα χαρακτηρίζεται ως "κακή" όταν η βαθμολογία της στην ιστοσελίδα είναι μικρότερη απο 40 ενώ θα χαρακτηρίζεται ως "καλή" όταν είναι μεγαλύτερη από 60. Σκοπός μας είναι να χρησιμοποιήσουμε μεθόδους ταξινόμησης ώστε να προβλέψουμε με όσο το δυνατόν μεγαλύτερη ακρίβεια την κατηγορία στην οποία ανήκει μια ταινία. Δύο ήταν οι κύριες προκλήσεις για την επίλυση αυτού του προβλήματος. Πρώτον, η συλλογή των χαρακτηριστικών των ταινιών από διαφορετικές πηγές δεδομένων . Δεύτερον, η εφαρμογή και η σύγκριση διαφορετικών μεθόδων ταξινόμησης ώστε να δημιουργήσουμε ένα αξιόπιστο μοντέλο πρόβλεψης.

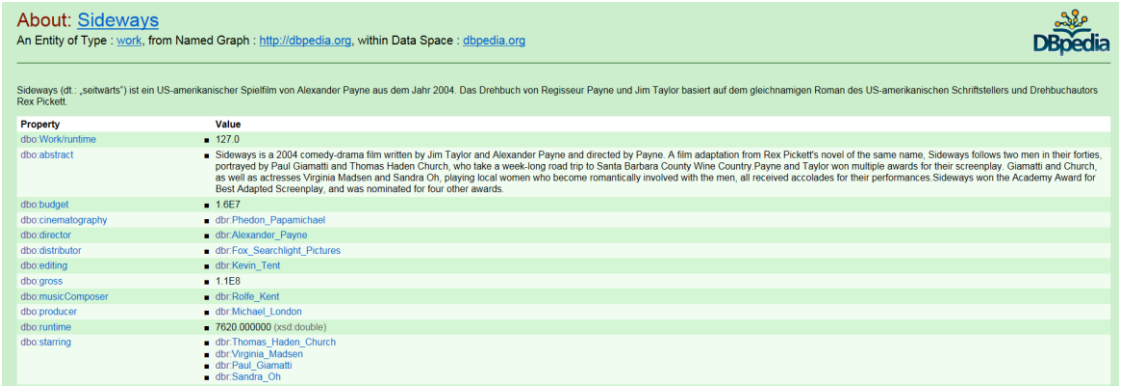
4.2 Στατιστικό πακέτο R

Η συλλογή των δεδομένων και οι τεχνικές κατηγοριοποίησης πραγματοποιήθηκαν χρησιμοποιώντας το στατιστικό πακέτο R (έκδοση 3.2.2), το οποίο είναι ένα ολοληρωμένο λογισμικό για ανάλυση δεδομένων και γραφημάτων. Η R είναι ένα υπολογιστικό πακέτο που βασίζεται στη γλώσσα προγραμματισμού S και πρόκειται για λογισμικού ανοιχτού κώδικα. Το λογισμικό R ξεκίνησε από τους **R**oss Ihaka και **R**obert Gentleman από τους οποίους πήρε και το όνομα του. Υποστηρίζει πολλές πλατφόρμες και λειτουργικά συστήματα όπως Linux, Mac OS και MS Windows. Από το 1997 αναπτύσσεται και συντηρείται από το R Development Core Team. Η R διατίθεται ελεύθερα στην ιστοσελίδα https://www.r-project.org/ και στηρίζεται στην ανάπτυξη προγραμμάτων μέσω πακέτων τα οποία

διατίθενται και πάλι ελεύθερα. Αυτή τη στιγμή πάνω από 2000 πακέτα διατίθενται τα οποία χρησιμοποιούνται για την επίλυση προβλημάτων σε διάφορους τομείς όπως τα χρηματοοικονομικά, τις κλινικές δοκιμές, τη μηχανική μάθηση κ.α. Αξίζει να αναφερθεί, ότι μέσω της R μπορούμε να συνδεθούμε και με άλλες γλώσσες προγραμματισμού όπως η C, C++, Java ή ακόμα και με άλλα ελεύθερα στατιστικά πακέτα όπως η Python. Επιπρόσθετα, δίνεται η δυνατότητα στους χρήστες να δημιουργήσουν το δικό τους πακέτο και στη συνέχεια να το καταστήσουν διαθέσιμο στους υπόλοιπους χρήστες. Τη τελευταία δεκαετία, πολλοί οργανισμοί αλλά και εταιρίες χρησιμοποιούν τις δυνατότητες που παρέχει η R σε ένα ευρύ φάσμα πεδίων, όπως η Google για την αξιολόγηση και τη βελτίωση των διαφημίσεων που προβάλλει, η Mozilla για την οπτικοποίηση της διαδικτυακής δραστηριότητας των χρηστών, η ANZ Bank για την ανάλυση του πιστωτικού κινδύνου.

4.3 Συλλογή Δεδομένων

Τα δεδομένα μας αποτελούνται από τη συλλογή 2000 ταινιών τα οποία περιέχουν το όνομα της ταινίας, την ημερομηνία που εκδόθηκε η ταινία, το URI από το DBpedia και την κατηγορία, ‘‘καλή’’ ή ‘‘κακή’’, στην οποία ανήκει η κάθε ταινία. Από το σύνολο των δεδομένων μας, οι 1600 ταινίες χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης και οι υπόλοιπες 400 ως τεστ δεδομένα. Στα αρχικά μας δεδομένα έχουμε διαθέσιμα τα URI από το DBpedia. Όπως είδαμε και σε προηγούμενο Κεφάλαιο, το DBpedia είναι ένα project για την εξαγωγή, διασύνδεση και επαναχρησιμοποίηση δομημένης πληροφορίας μέσω του Web από τη Wikipedia. Οι πληροφορίες είναι διαθέσιμες με τη χρήση του Σημασιολογικού Ιστού και των Διασυνδεδεμένων δεδομένων. Όπως φαίνεται και από το σχήμα 4.1. που ακολουθεί η δομή της αποτελείται από μια τριάδα (triples), η οποία αποτελείται από το υποκείμενο (subject), το κατηγόρημα (predicate) και το αντικείμενο (object). Το υποκείμενο μπορεί να είναι πόρος (resource), το κατηγόρημα αποτελεί την ιδιότητα που συνδέει το υποκείμενο με το αντικείμενο και το αντικείμενο αποτελεί την τιμή της ιδιότητας για το συγκεκριμένο υποκείμενο.



Σχήμα 4.1 DBpedia

Από το παραπάνω σχήμα, το υποκείμενο είναι η ταινία Sideways, το κατηγόρημα μπορεί να είναι ο παραγωγός της ταινίας (producer) και το αντικείμενο το όνομα του παραγωγού της ταινίας Michael London. Με ερωτήματα SPARQL και μέσω του στατιστικού πακέτου R και ειδικότερα κάνοντας χρήση της βιβλιοθήκης SPARQL πήραμε για την κάθε ταινία τα βασικά χαρακτηριστικά της όπως τον παραγωγό, το διανομέα (distributor), τη χώρα (country), τη λεζάντα (caption), τη διάρκεια της ταινίας (runtime), τον συγγραφέα (writer), το είδος της ταινίας (genre) και τους πρωταγωνιστές (starring) κ.α.

```
library(SPARQL)

# Live DBpedia endpoint
endpoint <- 'http://live.dbpedia.org/sparql'
options <- NULL

prefix <- c("db","http://dbpedia.org/resource/")

sparql_prefix <- "PREFIX dbp: <http://dbpedia.org/property/>
                  PREFIX dc: <http://purl.org/dc/terms/>
                  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
                  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
"
q <- paste(sparql_prefix,
           'SELECT ?actor ?movie ?director
            WHERE {
                ?m dc:subject <http://dbpedia.org/resource/Category:American_films> .
                ?m rdfs:label ?movie .
                FILTER(LANG(?movie) = "en")

                ?m dbp:starring ?a .
                ?a rdfs:label ?actor .
                FILTER(LANG(?actor) = "en")

                ?m dbp:director ?d .
                ?d rdfs:label ?director .
                FILTER(LANG(?director) = "en")
           }')

res <- SPARQL(url=endpoint,query=q,ns=prefix)$results
```

Σχήμα 4.2 Ερωτήματα μέσω SPARQL στην R

Στη συνέχεια και επειδή τα χαρακτηριστικά αυτά δεν διαχωρίζουν σε σημαντικό βαθμό την κλάση μιας ταινίας προχωρήσαμε στο επόμενο στάδιο συλλογής χαρακτηριστικών μέσω

άλλων διαθέσιμων δεδομένων. Για να γίνει αυτό χρησιμοποιήσαμε τους συνδέσμους sameAS από το DBpedia και μέσω του Freebase ώστε να κατορθώσουμε να ανακτήσουμε το IMDB ID της κάθε ταινίας. Το IMDB ID είναι ένας αναγνωριστικός μοναδικός αριθμός για την κάθε ταινία. Στη συνέχεια, έχοντας στη διάθεση μας το IMDB ID εφαρμόσαμε τη τεχνική web scraping και μέσω της ιστοσελίδας OMDBapi ανακτήσαμε κάποια επιπλέον χαρακτηριστικά για την κάθε ταινία. Το OMDBapi είναι μια δωρεάν υπηρεσία web με σκοπό τη παροχή πληροφοριών για τις διαθέσιμες ταινίες. Το Web API είναι μια διεπαφή που ένα υπολογιστικό σύστημα παρέχει, προκειμένου να επιτρέψει να γίνουν προς αυτό αιτήματα από άλλα προγράμματα για τη συλλογή δεδομένων. Στην ουσία, ένα Web API είναι ένα «κανάλι επικοινωνίας» μιας υπηρεσίας διαδικτύου με μία εξωτερική πηγή η οποία μέσω προγραμματιστικών εντολών, που ορίζονται από έναν οδηγό εντολών κάνοντας χρήση του API. Χρησιμοποιώντας τη βιβλιοθήκη της R omdbapi, έχουμε πολλές δυνατότητες για την συλλογή χαρακτηριστικών των ταινιών. Όπως φαίνεται στην εικόνα που ακολουθεί μπορούν να γίνουν ερωτήματα βάσει του τίτλου της ταινίας η οποία μας ενδιαφέρει. Στην περίπτωση όπου υπάρχουν περισσότερες από μια ταινίες με το ίδιο όνομα η αναζήτηση μπορεί να γίνει και μέσω του IMDB ID.

```
1
2  library(XML)
3  library(omdbapi)
4
5  search_by_title("Best Kept Secret")
6  # Title  Year     imdbID  Type
7  # (chr) (chr)      (chr) (chr)
8  # 1                  Best Kept Secret  2013 tt2433448 movie
9  # 2 Orgasmic Birth: The Best-Kept Secret  2008 tt1331111 movie
10 # 3                  Best Kept Secret  2006 tt0832340 movie
11 # 4           America's Best Kept Secret  1988 tt0393039 movie
12 # 5            Canada's Best Kept Secret  2011 tt2112122 movie
13 # 6               The Best Kept Secret  1986 tt3503162 movie
14 # 7            Best Kept Economic Secret  2003 tt2042461 movie
15 find_by_title("Best Kept Secret")
16 #>       Title: Winter Is Coming
17 #>        Year: 2011
18 #>       Rated: TV-MA
19 #>    Released: 2011-04-17
20 #>     Runtime: 62 min
21 #>       Genre: Adventure, Drama, Fantasy
22 #>    Director: Timothy Van Patten
23 #>      Writer: David Benioff (created by), D.B. Weiss (created by), George R.R.
24 #>              Martin ("A Song of Ice and Fire" by), David Benioff, D.B.
25 #>              Weiss
26 #>      Actors: Sean Bean, Mark Addy, Nikolaj Coster-Waldau, Michelle Fairley
27 #>        Plot: Jon Arryn, the Hand of the King, is dead. King Robert Baratheon plans
28 #>              to ask his oldest friend, Eddard Stark, to take Jon's
29 #>              place. Across the sea, Viserys Targaryen plans to wed his
30 #>              sister to a nomadic warlord in exchange for an army.
31 #>    Language: English
32 #>     Country: USA
33 #>      Awards: N/A
34 #>      Poster: http://ia.media-imdb.com/images/M/
35 #>              MV5BMTk5MDU3OTkzMF5BMl5BanBnXkFtZTcwOTc0ODg5NA@@._V1_SX300.jpg
36 #>  Metascore: N/A
37 #> imdbRating: 8.5
38 #>  imdbVotes: 12584
39 #>     imdbID: tt1480055
40 #>       Type: episode
```

Σχήμα 4.3 Ερωτήματα στο OMDBapi μέσω της R βάσει του τίτλου ταινίας

Στο σχήμα 4.4 που ακολουθεί, βλέπουμε κάποιους από τους τρόπους μέσω των οποίων μπορούμε είτε μέσω του τίτλου της ταινίας είτε μέσω του IMDB ID να ανακτήσουμε συγκεκριμένες πληροφορίες για μια ταινία. Συγκεκριμένα, στις γραμμές 64-79 πραγματοποιήσαμε ερωτήματα για την ανάκτηση περαιτέρω και πιο χρήσιμων

χαρακτηριστικών μέσω του OMDBapi των ταινιών, όπως τα κέρδη της ταινίας, την αξιολόγηση της στο IMDB, τον αριθμό των χρηστών που έχουν αξιολογήσει την ταινία.

```
get_genres(find_by_title("Best Kept Secret"))
#>  [1] "Documentary" "Drama"

get_writers(find_by_title("Best Kept Secret"))
#[1] "Francisco Bello" "Samantha Buck"   "Zeke Farrow"

get_directors(find_by_id("tt2433448"))
# [1] "Samantha Buck"

get_writers(find_by_id("tt2433448"))
# [1] "Francisco Bello" "Samantha Buck"   "Zeke Farrow"

get_genres(find_by_id("tt2433448"))
# [1] "Documentary" "Drama"

get_actors(find_by_id("tt2433448"))
# [1] "Alyce Barnhardt" "Carla Byrd"      "Johnson Green"   "Quran Key"

get_countries(find_by_id("tt2433448"))
# [1] "USA"


library(data.table)
movie<-fread("lod_data_movie_IMDB test.csv")
ID<-movie$`imdb URI`

l<-list()
i<-1
for(ind in ID){

  #ind<-"tt0068646"
  eval(parse(text=paste0("l[[",i,"]]","<-as.data.table(t(unlist(find_by_id('",'"',ind,'"',"'))))")))

  i<-i+1
}

library(plyr)
movie attributes <-rbind.fill(l)

```

Σχήμα 4.4 Ερωτήματα μέσω της βιβλιοθήκης omdbapi στην R βάσει του imdbid

4.4 Επιλογή μεταβλητών και καθαρισμός δεδομένων

Μέχρι τώρα ανακτήσαμε ένα σημαντικό πλήθος μεταβλητών για κάθε ταινία. Πριν προχωρήσουμε στην εφαρμογή των μεθόδων, θα αξιολογήσουμε τις μεταβλητές αυτές αλλά θα ελέγξουμε τις συσχετίσεις και τη πολυσυγραμμικότητα μεταξύ των χαρακτηριστικών που έχουμε ανακτήσει. Αρχικά, κάποιες από τις κατηγορικές μας μεταβλητές αφαιρέθηκαν από τα μοντέλα εξαιτίας του πλήθους των επιπέδων που είχαν. Για παράδειγμα, η μεταβλητή actor αποτελούνταν από 1187 επίπεδα (διαφορετικούς ηθοποιούς) ενώ η μεταβλητή director από 1246. Στη συνέχεια ελέγξαμε για ισχυρές συσχετίσεις μεταξύ των μεταβλητών. Μεταβλητές οι οποίες ήταν συσχετισμένες πάνω από 0.85 αφαιρέθηκαν από τα μοντέλα μας. Ισχυρά θετικά συσχετισμένες μεταβλητές ήταν το Tomatoes Rating, Tomatoes Meter, Tomatoes UserMeter. Ο έλεγχος των συσχετίσεων με τη μέθοδο του Pearson έγινε με τη συνάρτηση cor. Στη συνέχεια ελέγξαμε τη πολυσυγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών μας χρησιμοποιώντας την συνάρτηση alias της R. Η συνάρτηση αυτή, ανακαλύπτει τις πλήρως πολύσυγγραμικές μεταβλητές ώστε στη συνέχεια να μπορούμε να υπολογίσουμε το παράγοντα διόγκωσης διασποράς (Variance Inflation Factor-VIF). Από τον έλεγχο πολυσυγραμμικότητας αφαιρέθηκαν οι μεταβλητές Tomatoes Rotten και Language Final. Οι

συντελεστές προσδιορισμού εκφράζουν το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που εξηγείται από την ύπαρξη των ανεξάρτητων μεταβλητών στο μοντέλο και παίρνουν τιμές στο διάστημα [0,1]. Όταν παίρνουν την τιμή 1, δεν μπορεί να υπολογιστεί το VIF αφού $VIF_i = \frac{1}{1-R_i^2}$ όπου R_i^2 ο συντελεστής προσδιορισμού όταν η X_i χρησιμοποιείται ως εξαρτημένη μεταβλητή και οι υπόλοιπες X χρησιμοποιούνται ως ανεξάρτητες μεταβλητές. Όταν η τιμή του VIF είναι μεγαλύτερη από το 10 τότε υπάρχει ένδειξη για έντονη πολυσυγραμμικότητα στο μοντέλο οπότε και η αντίστοιχη μεταβλητή πρέπει να αφαιρείται. Όλες οι μεταβλητές μας έχουν VIF μικρότερο από 10, οπότε δεν αφαιρέσαμε κάποια μεταβλητή.

Σημαντικό ρόλο στην εκπαίδευση των αλγορίθμων που χρησιμοποιούνται είναι η ορθότητα των δεδομένων. Πρέπει να δίνεται ιδιαίτερη προσοχή σε πιθανά σφάλματα κάποιων τιμών των χαρακτηριστικών. Η μη ορθότητα κάποιων τιμών ονομάζεται θόρυβος (noise ή outlier). Η παρουσία θορύβου στα δεδομένα εκπαίδευσης, μπορεί να οδηγήσει σε αποπροσανατολισμό του αλγορίθμου και στην επιλογή μιας μη βέλτιστης λύσης. Για την ανίχνευση θορύβου στα δεδομένα, μπορούμε απλά να χρησιμοποιήσουμε διαγνωστικά διαγράμματα, όπως τα boxplots όπου εμφανίζονται πιθανά outliers. Επίσης, μπορούμε να εξετάσουμε την ύπαρξη μη λογικών τιμών στα δεδομένα μας. Για παράδειγμα, το imdbRating παίρνει τιμές από 0 έως 10 σύμφωνα με την αξιολόγηση μιας ταινίας στο IMDB. Επομένως, μια αρνητική τιμή του χαρακτηριστικού αυτού ή μια τιμή μεγαλύτερη του 10 θα αποτελούσε θόρυβο και θα έπρεπε να απομακρυνθεί. Επιπρόσθετα, η μεγάλη απόκλιση μεταξύ της μέσης τιμής και της παρατηρούμενης παρατήρησης αποτελεί ένδειξη θορύβου. Στην εικόνα που ακολουθεί, φαίνεται το summary των σημαντικότερων χαρακτηριστικών. Ένα ακόμα σημείο το οποίο ελέγξαμε είναι η μεταβλητότητα των χαρακτηριστικών. Η μεταβλητότητα είναι απαραίτητο χαρακτηριστικό καθώς μεταβλητές με σταθερές τιμές δεν προσφέρουν τίποτα στην εκπαίδευση του αλγορίθμου. Η ύπαρξη σταθερών τιμών ελέγχεται με τον υπολογισμό της τυπικής απόκλισης ή του συντελεστή συσχέτισης. Από τους διαγνωστικούς ελέγχους που πραγματοποιήσαμε δεν διαπιστώθηκε η ύπαρξη θορύβου στα δεδομένα μας.

Από τις μεταβλητές που έχουμε κρατήσει έως τώρα, οι μεταβλητές, Box Office, Tomatoes Reviews, Tomatoes userReviews, Tomatoes userRating και Tomatoes Fresh έχουν ελλιπή δεδομένα σε 120 παρατηρήσεις. Η απουσία τιμών κάποιων χαρακτηριστικών από τα δεδομένα ονομάζεται missing values. Μερικές από τις μεθόδους ταξινόμησης διαχειρίζονται

τα ελλιπή δεδομένα αγνοώντας τις παρατηρήσεις αυτές. Αυτό όμως θα είχε σαν αποτέλεσμα τη μείωση των δεδομένων εκπαίδευσης. Μια άλλη λύση είναι η αντικατάσταση των ελλιπών δεδομένων με την μέση τιμή του χαρακτηριστικού αυτού ή την επικρατούσα τιμή όταν πρόκειται για κατηγορική μεταβλητή. Μια τροποποίηση της προηγούμενης λύσης θα ήταν η αντικατάσταση με τη μέση ή την επικρατούσα τιμή (για κατηγορικές μεταβλητές) του χαρακτηριστικού σύμφωνα με την κλάση στην οποία ανήκει η παρατήρηση αυτή. Στην παρούσα μελέτη, για την αντικατάσταση των ελλιπών δεδομένων χρησιμοποιήσαμε μια άλλη μέθοδο όπως αυτή προτείνεται από τους J.Honaker, G.King και M.Blackwell στη βιβλιοθήκη Amelia της R. Σύμφωνα με αυτή, η αντικατάσταση των ελλιπών δεδομένων γίνεται χρησιμοποιώντας bootstrapping και EM αλγόριθμο παράγοντας πολλαπλά σύνολα δεδομένων. Αρχικά ο αλγόριθμος χωρίζει σε υποσύνολα τα δεδομένα με τη μέθοδο bootstrapped και στη συνέχεια εκτιμά τις ελλιπείς τιμές με τον EM αλγόριθμο. Ο αλγόριθμος υποθέτει ότι τα δεδομένα ακολουθούν την πολυμεταβλητή κανονική κατανομή (multivariate normal distribution). Ο αλγόριθμος EM εφαρμόζεται σε 2 βήματα. Το πρώτο βήμα λέγεται προσδοκία (expectation) και το δεύτερο μεγιστοποίηση (maximization). Στο πρώτο βήμα, η αναμενόμενη τιμή του λογαρίθμου της πιθανοφάνειας υπολογίζεται λαμβάνοντας υπόψη τα υπάρχοντα δεδομένα και στο δεύτερο βήμα η τιμή αυτή μεγιστοποιείται. Δηλαδή, επαναϋπολογίζεται ο λογάριθμος της πιθανοφάνειας έως ότου η πιθανοφάνεια που παράγεται σε δύο επαναλήψεις να μην διαφέρει σημαντικά. Τότε, λέμε ότι ο αλγόριθμος συγκλίνει. Στη συνέχεια, συνδυάζοντας τα αποτελέσματα των πολλαπλών συνόλων (μέση τιμή των εκτιμήσεων), δίνεται η τελική τιμή για το χαρακτηριστικό.

Τελικά, από τα χαρακτηριστικά που ανακτήσαμε αυτά τα οποία θα χρησιμοποιηθούν στους αλγορίθμους ταξινόμησης είναι τα παρακάτω

Year: χρονιά έκδοσης της ταινίας

Runtime: χρονική διάρκεια της ταινίας σε λεπτά

imdbRating: η βαθμολογία της ταινίας στο IMDB

imdbVotes: ο αριθμός των χρηστών που βαθμολόγησαν την ταινία στο IMDB

Oscars: ο αριθμός των Oscar

BoxOffice: τα κέρδη της ταινίας

Genre: το είδος της ταινίας

LanguageFinal: η γλώσσα της ταινίας

LanguageTotal: το σύνολο των γλωσσών που έχει μεταγλωτιστεί

Country: η χώρα προέλευσης της ταινίας

CountryUSA: ψευδομεταβλητή, αν η ταινία προέρχεται ή όχι από την Αμερική

CountryTOTAL: το σύνολο των χωρών που προβλήθηκε η ταινία

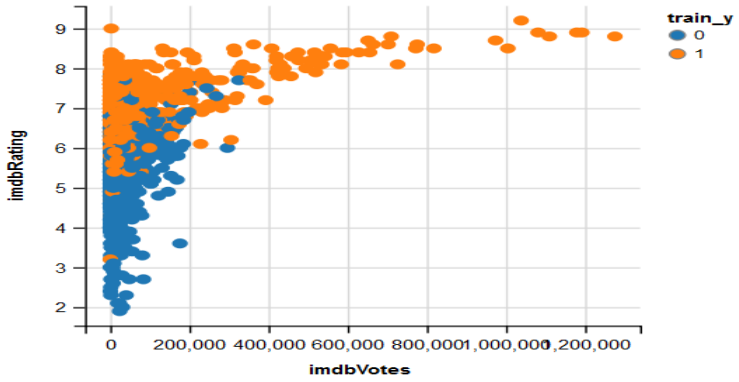
Tomatoes Reviews: ο αριθμός των χρηστών που έγραψαν κριτική για την ταινία

Tomatoes userReviews: ο αριθμός των αξιολογήσεων στο Tomatoes Rotten

Tomatoes userRating: η βαθμολογία των χρηστών που αξιολόγησαν την ταινία

Tomatoes Fresh: ο αριθμός των κριτικών με αξιολόγηση πάνω από 60%

Στο σχήμα 4.5 που ακολουθεί, παρατηρούμε ότι οι δυο μεταβλητές imdbRating και imdbVotes διαχωρίζουν σε μεγάλο βαθμό τις δύο κλάσεις της μεταβλητής απόκρισης. Αυτό μας δίνει τη διαίσθηση ότι τα αποτελέσματα των μοντέλων θα αποδώσουν σε ικανοποιητικό βαθμό.



Σχήμα 4.5 Απεικόνιση των κλάσεων της μεταβλητής απόκρισης βάσει των χαρακτηριστικών imdbRating και imdbVotes

4.5 Μέτρα Ακρίβειας

Εδώ θα αναφέρουμε στα μέτρα ακρίβειας με τα οποία θα αξιολογήσουμε τις μεθόδους που θα εφαρμόσουμε. Το σημαντικότερο μέτρο και σε αυτό στο οποίο θα βασιστούμε είναι η

συνολική ακρίβεια του μοντέλου (accuracy). Ωστόσο θα υπολογίσουμε και άλλα μέτρα τα οποία είναι ιδιαίτερα χρήσιμα όπως το Precision, και το Recall. Σε δίτιμα δεδομένα, όπως αυτά που μελετάμε, έχουμε τα θετικά δεδομένα τα οποία έχουν το χαρακτηριστικό που ζητάμε και τα αρνητικά από τα οποία απουσιάζει. Στα δεδομένα που αναλύσαμε θεωρούμε ότι η κακή ταινία είναι το επιθυμητό χαρακτηριστικό και συμβολίζεται με μηδέν. Έτσι, επιτυχής θετική πρόβλεψη (true positive prediction-TP) το μοντέλο θα έχει προβλέψει σωστά ένα αριθμό κακών ταινιών. Ακολουθεί ο πίνακας σύγχυσης (confusion matrix) ο οποίος μας βοηθά να υπολογίσουμε τα μέτρα ακριβείας

Test data\Prediction	0	1
0	True Positive	False Negative
1	False Positive	True Negative

Πίνακας 4.1 Confusion Matrix (πίνακας σύγχυσης)

Τα μέτρα ακριβείας που θα υπολογίσουμε ορίζονται ως εξής:

- Accuracy, είναι το ποσοστό των σωστών ταξινομήσεων, $Acc = \frac{TP+TN}{TP+FP+TN+FN}$
- Precision, το ποσοστό αυτό δηλώνει πόσες από τις εγγραφές που το μοντέλο έχει κατηγοριοποιήσει ως κακές είναι πραγματικά κακές, $Prec = \frac{TP}{TP+FP}$.
- Recall, το ποσοστό αυτό δηλώνει πόσες από τις κακές εγγραφές κατάφερε ο κατηγοριοποιητής να ταξινομήσει, $Rec = \frac{TP}{TP+FN}$.

4.6 Εφαρμογή των τεχνικών κατηγοριοποίησης με την R

Εφαρμογή της μεθόδου KNN

Για την εφαρμογή του αλγορίθμου KNN θα χρησιμοποιήσουμε τις συναρτήσεις knn και knn.cv από τη βιβλιοθήκη class. Για να αποφύγουμε τις ισοψηφίες θα χρησιμοποιήσουμε περιττό αριθμό πλησιέστερων γειτόνων. Η συνάρτηση knn χρησιμοποιεί την Ευκλείδεια απόσταση για την εύρεση των κοντινότερων γειτόνων και μας δίνει τη δυνατότητα να επιλέξουμε διάφορες τιμές για το k. Ταυτόχρονα ελέγχουμε την ακρίβεια του αλγορίθμου αφού δέχεται ως όρισμα τα τεστ δεδομένα. Στον πίνακα που ακολουθεί φαίνονται τα αποτελέσματα για τις διάφορες τιμές του k.

knn	k=1	k=3	k=5	k=7	k=11	k=13	k=15	k=19	k=23	k=35
Acc	56.3%	57.3%	52.5%	52.5%	52.3%	53.5%	53.8%	54.5%	54.0%	53.8%
Prec	59.0%	55.4%	48.7%	43.1%	44.6%	47.7%	45.6%	48.2%	44.6%	45.6%
Rec	54.8%	56.3%	51.4%	51.5%	51.2%	52.5%	53.0%	53.7%	53.4%	53.0%

Πίνακας 4.2 Αποτελέσματα της μεθόδου knn

Στο πίνακα που ακολουθεί θα δούμε τις εκτιμήσεις του μοντέλου χρησιμοποιώντας leave-one-out cross validation, με την συνάρτηση knn.cv για το σύνολο των δεδομένων

knn.cv	k=1	k=3	k=5	k=7	k=11	k=13	k=15	k=19	k=23	k=35
Acc	55.5%	56.7%	57.7%	57.7%	56.3%	56.2%	55.9%	56.5%	58.7%	58.9%
Prec	53.8%	54.9%	56.3%	56.4%	54.6%	54.6%	54.3%	54.8%	57.4%	57.3%
Rec	53.7%	55.5%	54.4%	52.8%	54.0%	52.3%	52.3%	54.2%	54.7%	57.2%

Πίνακας 4.3 Αποτελέσματα της μεθόδου knn με cross validation

Στους παραπάνω πίνακες παρατηρούμε ότι η απόδοση του μοντέλου δεν είναι ικανοποιητική. Αυτό οφείλεται σε ένα βαθμό, ότι η κλίμακα των χαρακτηριστικών είναι ανόμοια. Για να το αντιμετωπίσουμε αυτό κανονικοποιήσαμε τα δεδομένα μας σε μια κλίμακα από 0 έως 1 (βλέπε παράρτημα) και εφαρμόσαμε ξανά τη συνάρτηση του knn. Στο πίνακα ακολουθούν τα αποτελέσματα των κανονικοποιημένων δεδομένων

knn	k=1	k=3	k=5	k=7	k=11	k=13	k=15	k=19	k=23	k=35
Acc	86.8%	87.5%	88.0%	88.0%	88.3%	88.5%	88.3%	87.5%	88.3%	87.3%
Prec	80.0%	80.0%	81.0%	81.0%	81.0%	80.5%	79.5%	79.5%	80.5%	77.9%
Rec	91.8%	93.4%	93.5%	93.5%	94.0%	95.2%	95.7%	93.9%	94.6%	95.0%

Πίνακας 4.4 Αποτελέσματα της μεθόδου knn με κανονικοποίηση των δεδομένων

Με την κανονικοποίηση των χαρακτηριστικών πετύχαμε εμφανή βελτίωση της απόδοσης του αλγορίθμου. Από τον πίνακα 3 μπορούμε να δούμε ότι η συνολική απόδοση του αλγορίθμου μεγιστοποιείται για 13 πλησιέστερους γείτονες. Με το ποσοστό Recall να φτάνει το 95.2% αντιλαμβανόμαστε ότι το μοντέλο κάνει ελάχιστα λάθη τύπου false negative ενώ και το ποσοστό του Precision κυμαίνεται σε ικανοποιητικό επίπεδο ίσο με 80.5%. Επίσης, παρατηρούμε ότι αρχικά όσο αυξάνονται οι τιμές των κοντινότερων γειτόνων η απόδοση του αλγορίθμου βελτιώνεται, έως το k=13, από το σημείο αυτό και μετά όμως η απόδοση του μειώνεται σταδιακά.

Εφαρμογή των Support Vector Machines

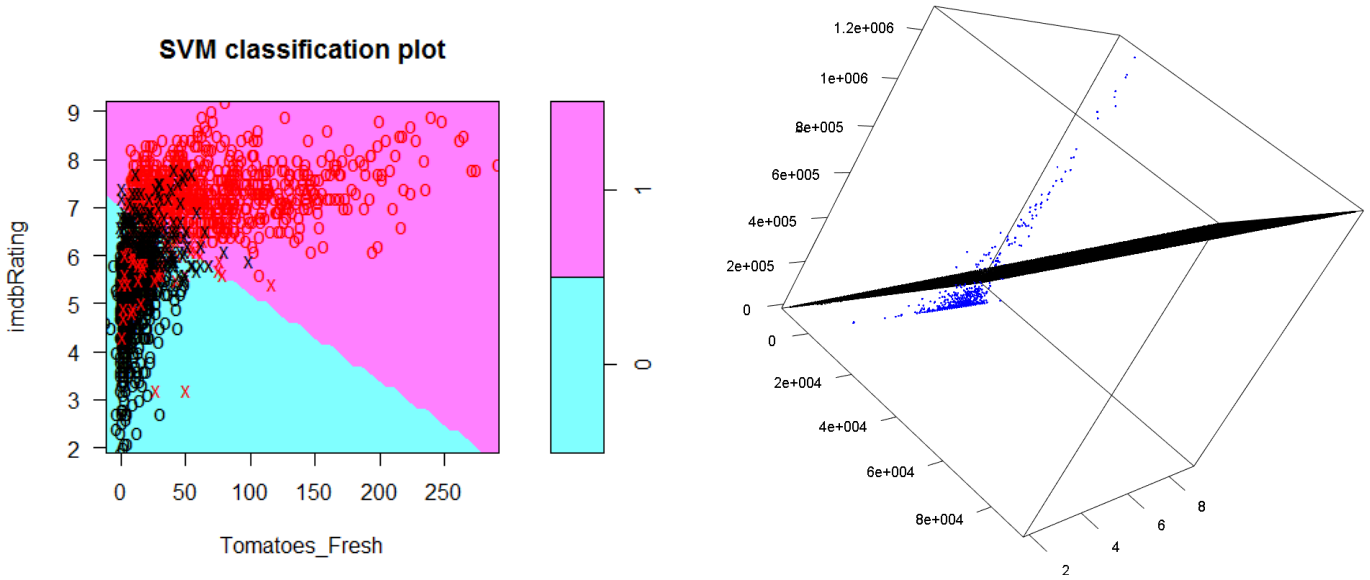
Η εφαρμογή των support vector machine πραγματοποιείται με τη συνάρτηση svm της βιβλιοθήκης e1071. Ένα από τα ορίσματα της συνάρτησης είναι το scale, το οποίο φέρνει

στην ίδια κλίμακα τόσο τις εξαρτημένες μεταβλητές όσο και την ανεξάρτητη μεταβλητή όταν αυτό απαιτείται. Επίσης, με το όρισμα Kernel μπορούμε να ορίσουμε τη συνάρτηση πυρήνα. Στον πίνακα που ακολουθεί μπορούμε να δούμε τα αποτελέσματα της μεθόδου για τις 4 συναρτήσεις πυρήνα. Η μέγιστη απόδοση της συνάρτησης επιτυγχάνεται με τη γραμμική συνάρτηση πυρήνα. Συμπεραίνουμε λοιπόν, ότι τα δεδομένα μας είναι γραμμικά διαχωρίσιμα. Μάλιστα επιτυγχάνεται ποσοστό συνολικής ακρίβειας ίσο με 98% και Precision ίσο με 98.4, το οποίο σημαίνει ότι αναγνωρίζει σε πολύ υψηλό ποσοστό τις "κακές" ταινίες των τεστ δεδομένων. Αυτό επιβεβαιώνεται και από τον πίνακα συνάφειας των αποτελεσμάτων, όπου 190 από τις 195 "κακές" ταινίες τις ταξινομεί σωστά.

svm	Linear	Polynomial	Radial	Sigmoid
Acc	98.0%	94.0%	95.8%	85.3%
Prec	98.4%	96.7%	97.8%	88.6%
Rec	97.4%	90.8%	93.3%	80.0%

Πίνακας 4.5 Αποτελέσματα της μεθόδου svm

Στα γραφήματα του σχήματος 4.6 που ακολουθούν, μπορούμε να δούμε κάποιες από τις γραφικές δυνατότητες της R. Απεικονίζεται ο διαχωρισμός που επιτυγχάνεται με τη χρήση των support vector machines στις 2 και στις 3 διαστάσεις αντίστοιχα.



Σχήμα 4.6 Απεικόνιση του διαχωρισμού με τη χρήση των svm

Εφαρμογή Λογιστικής Παλινδρόμησης

Η R εκτελεί τα γενικευμένα γραμμικά μοντέλα με τη συνάρτηση glm. Χρειάζεται ως όρισμα τα δεδομένα εκπαίδευσης, ο τύπος ο οποίος καθορίζει τη μορφή του μοντέλου και χωρίζει την επεξηγηματική μεταβλητή απο τις ερμηνευτικές μεταβλητές και μια συνάρτηση σύνδεσης (link function) με το όρισμα family. Στο μοντέλο μας, ορίσαμε ως συνάρτηση σύνδεσης τη Διωνυμική (binomial). Η glm διαχειρίζεται από μόνη της, τις κατηγορικές μεταβλητές κατασκευάζοντας ψευδομεταβλητές (dummy variables) για το κάθε επίπεδο της κατηγορικής μεταβλητής. Ακολουθούν τα αποτελέσματα της λογιστικής παλινδρόμησης.

glm	Binomial
Acc	96.5%
Prec	98.4%
Rec	94.4%

Πίνακας 4.6 Αποτελέσματα της μεθόδου glm

Με τη λογιστική παλινδρόμηση επιτυγχάνεται υψηλό ποσοστό της συνολικής ακρίβειας ίσο με 96.5%. Εκτελώντας την εντολή summary παίρνουμε κάποιες χρήσιμες πληροφορίες για το μοντέλο. Από τις τιμές z-value που μας δίνονται βλέπουμε ότι ισχυροί προγνωστικοί παράγοντες είναι τα Oscars, Tomatoes_Reviews, Tomatoes_Fresh, Tomatoes_userReviews. Αρκετές από τις μεταβλητές δεν είναι στατιστικά σημαντικές. Συνεπώς εφαρμόζουμε ένα νέο μοντέλο λογιστικής παλινδρόμησης με τις στατιστικά σημαντικές μεταβλητές μόνο και τα αποτελέσματα παρουσιάζονται στον επόμενο πίνακα.

glm	Binomial
Acc	98.2%
Prec	99.0%
Rec	97.4%

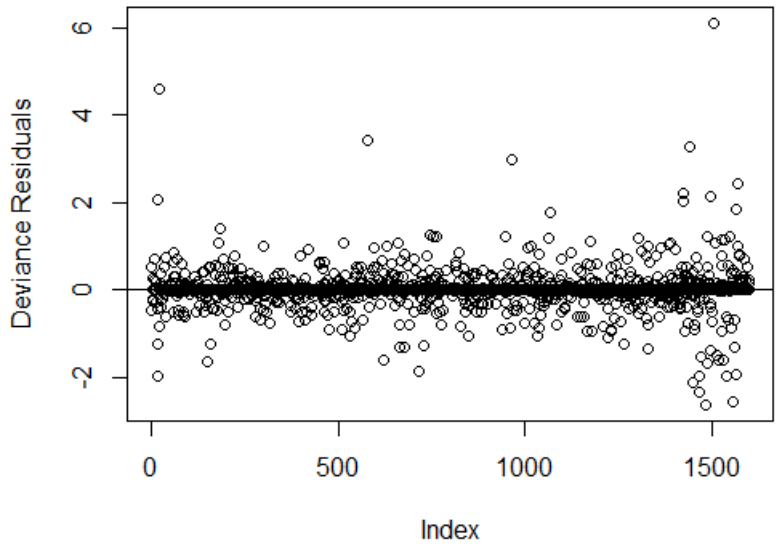
Πίνακας 4.7 Αποτελέσματα της μεθόδου glm με το βέλτιστο μοντέλο

Παρατηρούμε ότι η απόδοση του μοντέλου βελτιώθηκε στο 98.3% ενώ και το Recall ανέβηκε στο 97.4% από το 94.4%. Αυτό φαίνεται και από τον έλεγχο σύγκρισης των μοντέλων όπου έχει για μηδενική υπόθεση ότι το μοντέλο εφαρμόζει καλύτερα στα δεδομένα.

```
glm.fit_movie_1=glm(y~. , data=train_1 ,family =binomial )
glm.fit_movie_2=glm(y~. , data=train_2 ,family =binomial )

anova(glm.fit_movie_2,glm.fit_movie_1)
#  Analysis of Deviance Table
#
#  Model 1: y ~ Oscars + Tomatoes_Reviews + Tomatoes_Fresh + Tomatoes_userReviews
#  Model 2: y ~ Runtime + imdbRating + imdbVotes + Year + Oscars + Tomatoes_Reviews +
#    Tomatoes_Fresh + Tomatoes_userRating + Tomatoes_userReviews +
#    BoxOffice + Genre_A + Language_Final + Language_Total + CountryFinal +
#    Country_USA + Country_Total
#  Resid. Df Resid. Dev Df Deviance
#  1      1594     327.79
#  2      1522     245.78 72    82.01
```

Σχήμα 4.7 Έλεγχος σύγκρισης μοντέλων της glm

Ο έλεγχος X^2 δίνει p-value μεγαλύτερο από 0.05 (1-pchisq(3.053,2)=0.19). Έτσι, δεν απορρίπτεται η μηδενική υπόθεση και επομένως το νέο μοντέλο εφαρμόζει καλύτερα. Στο διάγραμμα που ακολουθεί, έχουν υπολογιστεί τα κατάλοιπα απόκλισης. Από το γράφημα φαίνεται ότι έχουμε κάποιες ακραίες τιμές που ξεπερνούν τη τιμή 4 όπως αυτή της παρατήρησης 1502 με τιμή ίση με 6.11. Η αρνητική τιμή των καταλοίπων υποδηλώνει ότι η εκτιμώμενη τιμή είναι μεγαλύτερη από τη παρατηρούμενη.



Σχήμα 4.8 Απεικόνιση των καταλοίπων της glm

Εφαρμογή Διαχωριστικής Ανάλυσης

Οι συναρτήσεις που χρειάζονται για την εφαρμογή της Διαχωριστικής Ανάλυσης βρίσκονται στη βιβλιοθήκη MASS. Με την εντολή lda και εφαρμόζοντας το μοντέλο παίρνουμε τις αρχικές αναλογίες (prior probabilities) της εξαρτημένης μεταβλητής και τις

μέσες τιμές κάθε ανεξάρτητης μεταβλητής για κάθε επίπεδο της εξαρτημένης μεταβλητής, όπως φαίνεται στη παρακάτω εικόνα.

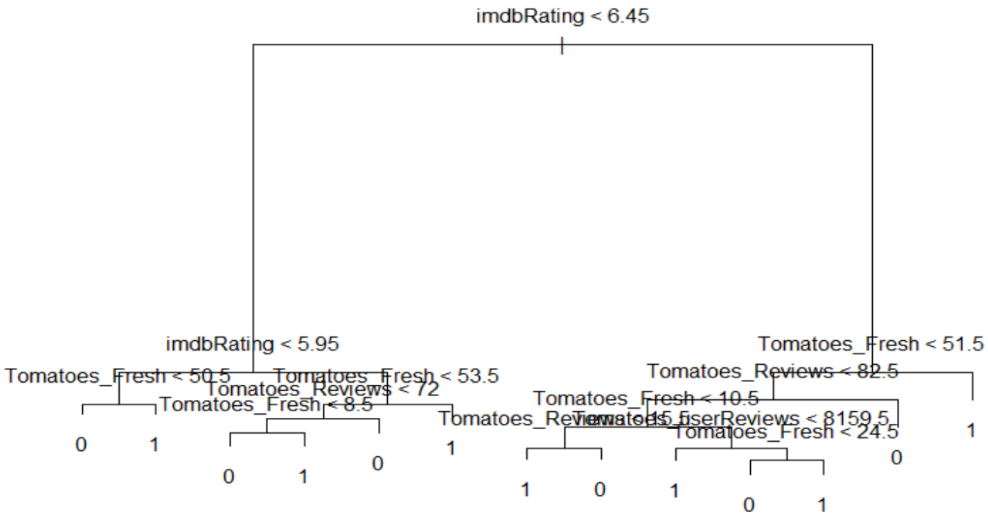
```
Call:
lda(y ~ ., data = train)

Prior probabilities of groups:
        0         1
0.4809256 0.5190744

Group means:
    Runtime imdbRating imdbVotes     Year     Oscars Tomatoes_Reviews Tomatoes_Fresh Tomatoes_userRating
0  98.52016   5.397789  30517.36 2004.085 0.00130039         72.08322       14.28609            2.967750
1 106.75542   7.184699  73970.25 2001.524 0.20602410         80.42048       65.86747            3.599277
  Tomatoes_userReviews BoxOffice Genre_AAdventure Genre_AAnimation Genre_ABiography Genre_AComedy
0             489644.4  31514219       0.04681404       0.02470741       0.01560468     0.3797139
1             450682.0  33140055       0.03132530       0.03975904       0.04819277     0.2385542
  Genre_ACrime Genre_ADocumentary Genre_ADrama Genre_AFamily Genre_AFantasy Genre_AHorror Genre_AMusical
0   0.05071521         0.01040312    0.1586476    0.00390117    0.001300390    0.07412224     0.00130039
1   0.06506024         0.19638554    0.2722892    0.00000000    0.001204819    0.01686747     0.00000000
```

Σχήμα 4.9 Output της συνάρτησης lda στην R

Γενικότερα, στα αποτελέσματα της διαχωριστικής ανάλυσης, όταν οι διαχωριστές είναι περισσότεροι από δύο τότε ταξινομούνται με σειρά, από αυτόν με το μεγαλύτερο ίχνος προς τον μικρότερο. Από τη τιμή αυτή, κάθε διαχωριστή συμπεραίνουμε και τη διαχωριστική ικανότητα του. Με τη μέθοδο αυτή επιτυγχάνετε ποσοστό συνολικής ακρίβειας ίσο με 94.8% ενώ το Recall και το Precision ισούνται με 96.3% και 92.8% αντίστοιχα.

Εφαρμογή δέντρων απόφασης

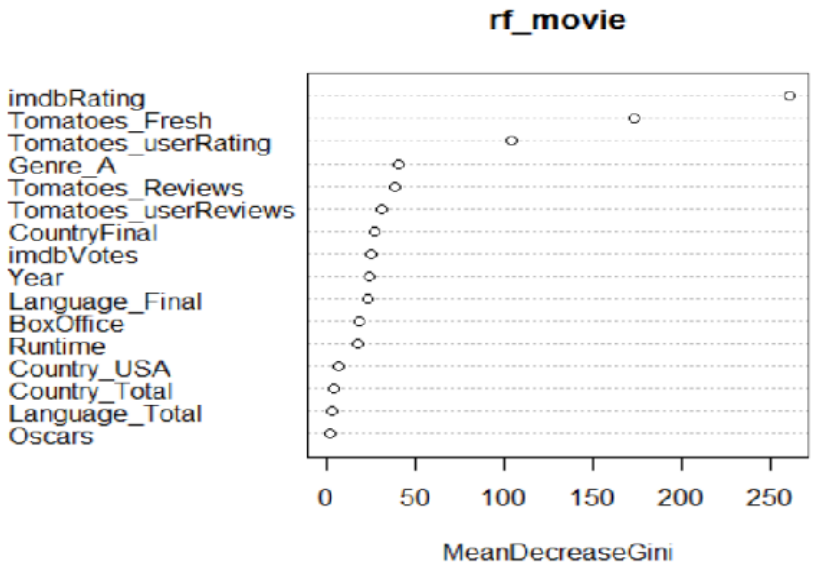
Η υλοποίηση των δέντρων απόφασης πραγματοποιήθηκε μέσω της συνάρτησης tree και της αντίστοιχης βιβλιοθήκης. Η συνάρτηση μέσω του ορίσματος na.action διαχειρίζεται τα ελλιπή δεδομένα ενώ μέσω του ορίσματος split μας επιτρέπει να καθορίσουμε τον τρόπο με τον οποίο γίνεται κάθε διάσπαση όπου μπορούμε να επιλέξουμε τη μέθοδο Gini ή την απόκλιση. Η διάσπαση του δέντρου η οποία μεγιστοποιεί τη μείωση του δείκτη Gini αυτή και επιλέγεται. Επίσης, η συνάρτηση αυτή έχει ένα μειονέκτημα καθώς δεν μπορεί να διαχειριστεί κατηγορικές μεταβλητές με περισσότερα από 32 επίπεδα. Ο τρόπος που χειρίζεται τις συνεχείς μεταβλητές είναι δημιουργώντας διαστήματα. Στην εικόνα που ακολουθεί, βλέπουμε πως η ρίζα του δέντρου είναι το imdbRating και η διάσπαση έγινε με βάση αν η ταινία έχει μικρότερη ή μεγαλύτερη βαθμολογία από 6.45. Για την δημιουργία των τερματικών κόμβων, οι διασπάσεις συνεχίζονται έως ότου οι τερματικοί κόμβοι είναι αρκετά μικροί ή αρκετά λίγοι για να διασπαστούν. Με τη μέθοδο των δέντρων απόφασης έχουμε σωστή ταξινόμηση ίση με 91.5% .



Σχήμα 4.10 Απεικόνιση του δέντρου απόφασης στην R

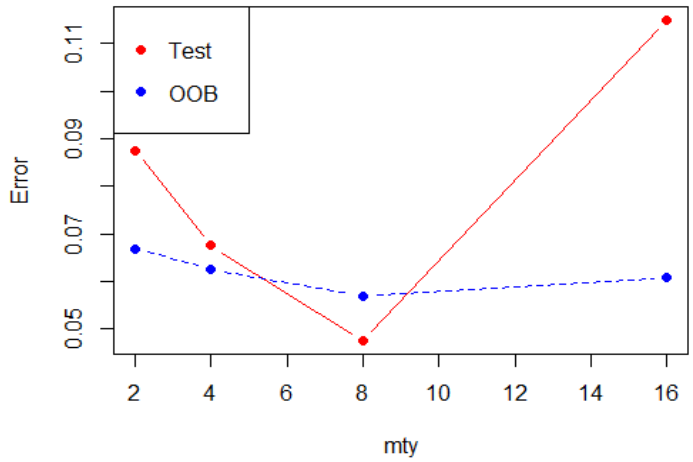
Εφαρμογή της μεθόδου random forest

Για τη μέθοδο random forest χρειαζόμαστε την βιβλιοθήκη και την συνάρτηση randomForest. Η συνάρτηση βασίζεται στον αρχικό αλγόριθμο των Breiman και Culter σε κώδικα Fortran. Με το όρισμα importance της συνάρτησης, μας δίνεται η δυνατότητα της αξιολόγησης των ανεξάρτητων μεταβλητών η οποία βασίζεται στη μέση μείωση του δείκτη Gini. Επίσης, με τα ορίσματα ntree μπορούμε να δηλώσουμε τον αριθμό των δέντρων που θα κατασκευαστούν στο δάσος και με το mtry τον αριθμό των μεταβλητών που επιλέγονται τυχαία σε κάθε διάσπαση ενός δέντρου.



Σχήμα 4.11 Συνεισφορά των ανεξάρτητων μεταβλητών στα random forest

Εφαρμόζοντας τον αλγόριθμο, από τα αποτελέσματα φαίνεται ότι οι σημαντικότερες μεταβλητές για τον διαχωρισμό της εξαρτημένης μεταβλητής και με σημαντική διαφορά από τις υπόλοιπες μεταβλητές είναι το imdbrating με τιμή ίση με 260.5. Ακολουθούν το Tomatoes Fresh με 172.7, το Tomatoes userRating με 103.4 και το Genre με 40.5. Η συνεισφορά όλων των μεταβλητών φαίνεται στο σχήμα 4.11. Επίσης, μας δίνεται το OOB error του μοντέλου το οποίο είναι ίσο με 5.75% και ο προεπιλεγμένος αριθμός των μεταβλητών. Επειδή στα τυχαία δάση, το σφάλμα ταξινόμησης υπολογίζεται από τα out-of-bag δεδομένα, ακολουθεί η γραφική παράσταση που απεικονίζει το πόσο μειώνεται ή αυξάνεται το σφάλμα όσο αυξάνεται ο αριθμός των μεταβλητών που επιλέγονται για τη διάσπαση ενός δέντρου του δάσους. Επίσης στο σχήμα 4.12 απεικονίζεται και η αντίστοιχη μεταβολή του σφάλματος στα τεστ δεδομένα. Παρατηρούμε, ότι η μείωση του OOB σφάλματος δεν συνεπάγεται και ταυτόχρονη μείωση του σφάλματος στα τεστ δεδομένα. Η μέθοδος random forest μας έδωσε 95% ποσοστό συνολικής ακρίβειας με ιδιαίτερα υψηλό Precision της τάξεως του 98.5%.



Σχήμα 4.12 Απεικόνιση OOB και Test σφάλματος με τη random forest

Εφαρμογή της μεθόδου boosting

Στη συνέχεια ακολουθεί η μέθοδος boosting. Για την εφαρμογή της μεθόδου χρειαζόμαστε τη συνάρτηση boosting της βιβλιοθήκης adabag. Η συνάρτηση έχει ως ορίσματα την συνάρτηση των μεταβλητών, τα δεδομένα εκπαίδευσης και με το όρισμα coeflearn τον τρόπο με τον οποίο υπολογίζονται τα βάρη. Μπορούμε να επιλέξουμε μεταξύ των μεθόδων Breiman, Freund, Zhu. Για τη μέθοδο Breiman που μας έδωσε και τα μεγαλύτερα ποσοστά ακρίβειας, ισούνται με alpha=1/2ln((1-err)/err). Η συνεισφορά των μεταβλητών στον διαχωρισμό της εξαρτημένης μεταβλητής, με τη μέθοδο boosting φαίνονται στο σχήμα 4.13

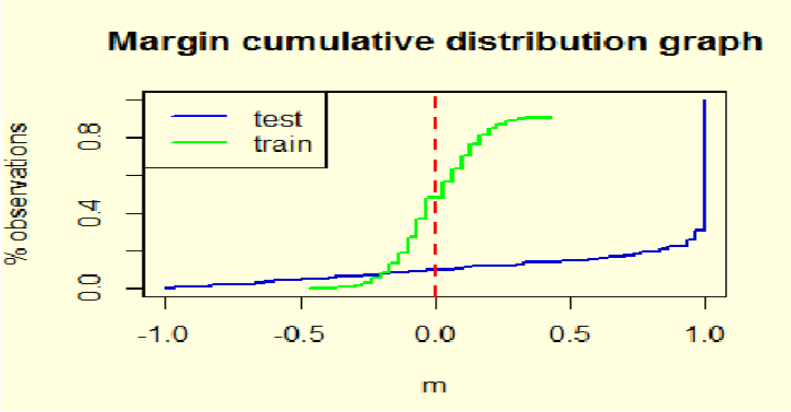
που ακολουθεί. Παρατηρούμε ότι οι σημαντικότερες μεταβλητές είναι Tomatoes Fresh και το Tomatoes Reviews ενώ τα Oscar και το Country USA έχουν ελάχιστη συνεισφορά. Με τη μέθοδο αυτή, το ποσοστό της συνολικής ακρίβειας του μοντέλου ισούται με 97.8%.

Variable	Importance
importance.Tomatoes_Fresh	21.29
importance.Tomatoes_Reviews	14.46
importance.Tomatoes_userReviews	8.76
importance.Year	7.83
importance.Genre_A	7.57
importance.BoxOffice	7.36
importance.imdbRating	6.93
importance.CountryFinal	6.92
importance.imdbVotes	6.60
importance.Runtime	4.56
importance.Tomatoes_userRating	4.21
importance.Language_Final	1.82
importance.Language_Total	0.89
importance.Country_Total	0.72
importance.Oscars	0.09
importance.Country_USA	0.00

Σχήμα 4.13 Συνεισφορά των ανεξάρτητων μεταβλητών με τη μέθοδο boosting

Εφαρμογή της μεθόδου Bagging

Για την μέθοδο Bagging, θα χρησιμοποιήσουμε τη συνάρτηση bagging της βιβλιοθήκης adabag. Η συνάρτηση αυτή χρησιμοποιεί ως ορίσματα την εξίσωση των μεταβλητών, τον αριθμό των δέντρων που κατασκευάζονται με προεπιλεγμένο αριθμό τα 100 (όρισμα mfinal), τα δεδομένα εκπαίδευσης και το όρισμα control το οποίο καθορίζει τον ελάχιστο αριθμό των παρατηρήσεων που πρέπει να υπάρχουν σε ένα κόμβο ώστε να πραγματοποιηθεί η διάσπαση ή τις μέγιστες διασπάσεις που θα επιτρέπονται σε έναν κόμβο (depth node). Το όρισμα control, συντάσσεται σύμφωνα με την συνάρτηση rpart. Από τα αποτελέσματα, μας δίνεται ένας πίνακας ο οποίος περιέχει τις ψήφους για κάθε παρατήρηση από τις ταξινομήσεις που πραγματοποιήθηκαν σε κάθε δέντρο για τις δυο κλάσεις. Επίσης, μας δίνεται και ο αντίστοιχος πίνακας για κάθε παρατήρηση, της πιθανότητας να ανατεθεί σε κάθεμια από τις δυο κλάσεις. Με την εντολή samples, μας παρέχεται η δυνατότητα να δούμε τα bootstrap δείγματα που χρησιμοποιήθηκαν σε κάθε επανάληψη του αλγορίθμου ενώ η σημαντικότητα των μεταβλητών υπολογίζεται και στη συνάρτηση αυτή με βάση τον δείκτη Gini.



Σχήμα 4.14 Απεικόνιση περιθωρίων με τη μέθοδο bagging

Το σχήμα 4.14 απεικονίζει το περιθώριο κάθε παρατήρησης των δεδομένων εκπαίδευσης. Στην ουσία το περιθώριο, προσδιορίζει το πόσο ασφαλής ήταν η ταξινόμηση μιας παρατήρησης στην κλάση στην οποία ανατέθηκε. Υπολογίζεται, ως η διαφορά της στήριξης της σωστής ανάθεσης από τη μέγιστη διαφορά της λανθασμένης. Από το παραπάνω γράφημα, παρατηρούμε ότι στα δεδομένα εκπαίδευσης το περιθώριο κυμαίνεται από 0.5 έως -0.5 με αρκετές από τις παρατηρήσεις να βρίσκονται γύρω από το μηδέν. Αντίθετα, στα τεστ δεδομένα, οι 312 από τις 400 παρατηρήσεις έχουν περιθώριο μεγαλύτερο ή ίσο του 0.90. Αυτό συνεπάγεται ότι οι αναθέσεις των παρατηρήσεων στα τεστ δεδομένα έχει γίνει με μεγαλύτερη ασφάλεια. Με τη μέθοδο αυτή επιτυγχάνεται ποσοστό συνολικής ακρίβειας 91.75%. Οι μεταβλητές οι οποίες διαχωρίζουν τα δεδομένα μας, με τη μέθοδο bagging είναι το imdbRating, Tomatoes Fresh και το Tomatoes Reviews.

Εφαρμογή της μεθόδου Naïve-Bayes

Η συνάρτηση naiveBayes της βιβλιοθήκης e1071 υπολογίζει την εκ των προτέρων πιθανότητα της κλάσης μιας κατηγορικής μεταβλητής δεδομένου των μεταβλητών χρησιμοποιώντας το Θεώρημα του Naïve-Bayes. Για τα χαρακτηριστικά με ελλειπούσες τιμές όπου δεν μπορεί να υπολογιστεί η πιθανότητα η συνάρτηση παραλείπει τις αντίστοιχες πρατηρήσεις. Όμως, με το όρισμα laplace, επιτυγχάνεται ένας έλεγχος εξομάλυνσης για τις πιθανότητες που ισούνται με μηδέν, προσθέτοντας μια τιμή στο μετρητή κάθε παρατήρησης του χαρακτηριστικού. Ως προεπιλογή έχει τη τιμή μηδέν, η οποία απενεργοποιεί την εξομάλυνση αυτή. Ακολουθούν τα αποτελέσματα της ταξινόμησης με τη μέθοδο του Naïve-Bayes από τα οποία συμπεραίνουμε ότι δεν επιτυγχάνει ιδιαίτερα υψηλά ποσοστά σωστής ταξινόμησης σε σύγκριση με τις μεθόδους που έχουμε μελετήσει έως τώρα.

Naïve-Bayes	Linear
Acc	86.3%
Prec	83.3%
Rec	89.7%

Πίνακας 4.8 Αποτελέσματα της μεθόδου Naïve-Bayes

4.7 Συμπερασματολογία

Παρουσιάστηκαν παραπάνω 9 μέθοδοι κατηγοριοποίησης από τις οποίες μπορούμε να ταξινομήσουμε με μεγάλη ασφάλεια μια ταινία στην κλάση την οποία ανήκει, αφού τα ποσοστά συνολικής ακρίβειας κυμαίνονται από 88.5% έως 98.3%. Από το σύνολο των μεθόδων, οι μέθοδοι της λογιστικής παλινδρόμησης, των support vector machine και η μέθοδος boosting δίνουν τα μεγαλύτερα ποσοστά στα μέτρα ακριβείας που υπολογίσαμε ενώ η μέθοδος του Naïve-Bayes και η μέθοδος των πλησιέστερων γειτόνων τα χαμηλότερα. Υψηλά ποσοστά ακρίβειας επιτυγχάνονται με τη διαχωριστική ανάλυση, τη μέθοδο bagging, τα δέντρα απόφασης και τα τυχαία δάση. Το μεγαλύτερο recall επιτυγχάνεται από τις μεθόδους της λογιστικής παλινδρόμησης και των support vector machine και είναι ίσο με 97.4%. Η λογιστική παλινδρόμηση και η μέθοδος boosting υπερισχύουν ελαφρά στο precision έναντι των support vector machine. Επομένως, με αυτές τις δυο μεθόδους έχουμε τα υψηλότερα ποσοστά εγγραφών που ταξινομήθηκαν ως κακές και είναι πραγματικά κακές μειώνοντας έτσι τον αριθμό των ανεπιτυχών θετικών προβλέψεων (false positive).

	GLM	SVM	KNN	LDA	Boosting	Bagging	Decision Trees	Random Forest	Naïve-Bayes
Acc	98.3%	98.0%	88.5%	94.8%	97.8%	90.5%	91.5%	95.0%	86.3%
Prec	99.0%	98.4%	80.5%	96.3%	98.9%	92.9%	92.1%	98.3%	83.3%
Rec	97.4%	97.4%	95.2%	92.8%	96.4%	87.2%	90.3%	91.3%	89.7%

Πίνακας 4.9 Συγκεντρωτικός πίνακας αποτελεσμάτων των μεθόδων ταξινόμησης

Επίσης, είδαμε ότι οι τεχνικές που βασίζονται σε έναν αλγόριθμο μάθησης και την δειγματοληψία των δεδομένων εκπαίδευσης βελτιώνουν τα ποσοστά ακρίβειας. Τα δέντρα απόφασης έχουν ποσοστό ακρίβειας 91.5% ενώ με τη μέθοδο boosting και random forest αυξάνεται σε 97.8% και 95% αντίστοιχα.

Βιβλιογραφία

Ελληνική Βιβλιογραφία

[Ε1] Γ.Θεοδωρίδης, Ν.Πελέκης (2013), Στατιστικές μέθοδοι Εξόρυξης Δεδομένων

[Ε2] Μ.Κούτρας (2013), Εφαρμοσμένη Πολυμεταβλητή Ανάλυση

[Ε3] Κωνσταντίνος Φωκιανός, Χαράλαμπος Χαραλάμπους (2010), Εισαγωγή στην R

[Ε4] Ι.Βλαχάβας, Π.Κεφαλάς, Ν.Βασιλειάδης, Φ.Κόκκορας, Η.Σακελλαρίου (2011), Τεχνητή Νοημοσύνη

Ξένη Βιβλιογραφία

[A1] C. Bizer, R. Cyganiak, T. Heath (2007), How to publish Linked Data on the Web

[A2] T.Berners-Lee (July 2006), Linked Data

[A3] T.Heath, M. Hausenblas, C. Bizer, R.Cyganiak, O. Hartig (November 2008), Presentation on How to Publish Linked Data on the Web

[A4] Florian Bauer, Martin Kaltenböck, Linked Open Data: The Essentials

[A5] I. Jacobs, N. Walsh (December 2004), Architecture of the World Wide Web

[A6] Marja-Riitta Koivunen and Eric Miller (November 2001) W3C Semantic Web Activity

[A7] Patel-Schneider, P.F Hayes, Horrocks (2006), OWL Semantics and Abstract Syntax

[A8] Dan Brickley, R.V. Guha (February 2014), RDF Schema 1.1

[A9] Patrick J. Hayes, Florida IHMC (February 2014), RDF 1.1 Semantics

[A10] R. Lewis (2007), Dereferencing HTTP URIs.

[A11] C. Bizer, T. Heath, T. Berners-Lee (2009), Linked Data – The Story So Far

[A12] Tony Segaran, Jeff Hammerbacher (August 2009), Beautiful Data

[A13] Tom Heath, Christian Bizer (February 2011), Linked Data: Evolving the Web into a Global Data Space

[A14] Grigoris Antoniou, Frank van Harmelen (2003), A semantic web primer

[A15] Richard Cyganiak (August 2014), The Linking Open Data Diagram

[A16] L. Sauermann, R. Cyganiak (December 2008), Cool URIs for the Semantic Web

[A17] T.Berners-Lee (2010), 5 stars Linked Open Data

[A18] David Wood (2011), Linking Government Data

[A19] Daniel Bennett (September 2009), Publishing Open Government Data

[A20] Josh Tauberer (February 2009), Open Government Data Standards and Setting Expectations

[A21] Tim Berners-Lee (June 2009), Putting Government Data Online

[A22] Carl Malamud, Tim O'Reilly (December 2007), 8 Open Government Data Principles

[A23] Sunlight Foundation (2010), Principles for Transparency in Government

[A24] Eric Prud, Andy Seaborne (January 2008) SPARQL Query Language for RDF

[A25] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas,Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, Christian Bizer (May 2015), DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

[A26] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann (May 2009), DBpedia - A Crystallization Point for the Web of Data

[A27] F. Wu, D. Weld (2008), Automatically Refining the Wikipedia Infobox Ontology

[A28] Andrezej Kobylinski, Andrezej Sobczak (2013), Perspectives in Business Informatics Research

[A29] Richard Hammel (2012) Open Data Driving Growth, ingenuity and innovation

[A30] Christopher M. Bishop (2006), Pattern Recognition and Machine Learning

[A31] S. B. Kotsiantis (July 2007), Supervised Machine Learning: A Review of Classification Techniques

[A32] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2014), Introduction to Statistical Learning

[A33] Kristína Machová, František Barčák, Peter Bednár (Vol. 3, No. 2, 2006), A Bagging Method using Decision Trees in the Role of Base Classifiers

[A34] Alfaro, Esteban; Gamez, Matias and Garcia, Noelia; with contributions from Li Guo (October 14, 2015) Package 'adabag'

[A35] Lior Rokach, Oded Maimon (2013), DECISION TREES

[A36] Neha Patel, Divakar Singh (February 2015), An Algorithm to Construct Decision Tree for Machine Learning based on Similarity Factor

[A37] Gorunescu, F (Springer, 2011), Data Mining: Concepts, Models, and Techniques

[A38] Jiawei Han, Michelini Kamber (2006), Data Mining Concept and Techniques

[A39] J. Zaki and W. Meira Jr. (2014), Fundamentals of Data Mining Algorithms

[A40] Brian Ripley (June 2015), Package 'tree'

[A41] Brian Ripley, William Venables (August 2015), Package 'class'

[A42] J. Furnkranz (2014), Instance-Based Learning

[A43] Leo Breiman, Adele Cutler, Andy Liaw, Matthew Wiener, (October 2015) Package 'randomForest'

[A44] Andy Liaw, Matthew Wiener (December 2002), Classification and Regression by randomForest

[A45] David Meyer,Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch,Chih-Chung Chang, Chih-Chen Lin (August 2015), Package 'e1071'

[A46] G.Cauwenberghs, T.Poggio (2000), Incremental and decremental support vector machine learning

[A47] Yanchang Zhao (April 2013), R and Data Mining: Examples and Case Studies

[A48] James Honaker, Gary King, Matthew Blackwell (February 2015), Package 'Amelia'

[A49] Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, David Firth (November 2015), Package 'MASS'

[A50] Jesse Davis, Mark Goadrich (2010), The Relationship Between Precision-Recall and ROC Curves

[A51] Sergio A. Alvarez, An exact analytical relation among recall, precision and classification accuracy in information retrieval

[A52] K. G. Clark, L. Feigenbaum, E. Torres (January 2008), SPARQL Protocol for RDF

[A53] M. Hausenblas, I. Herman, B. Adida (October 2008), RDFa - Bridging the Web of Documents and the Web of Data

[A54] Peter Flach (August 2012), The art and science of algorithmsthat make sense of data

Ιστοσελίδες

[I1] http://opengovernmentdata.org/

[I2] https://okfn.org/opendata/

[I3] https://okfn.org/

[I4] http://www.w3.org/

[I5] http://www.linkeddatatools.com/introducing-rdf-part-2

[I6] http://www.linkeddatatools.com/introducing-rdfs-owl

[I7] http://www.linkeddatatools.com/querying-semantic-data

[I8] http://www.w3.org/TR/2009/NOTE-egov-improving-20090512/#OGD

[I9] https://www.r-project.org/

[I10] http://geodata.gov.gr

[I11] http://wiki.dbpedia.org/