ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Δ.Π.Μ.Σ. ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Μεταπτυχιακή Διατριβή

Μηχανική Μάθηση και Μέτρα Πληροφορίας στην πρόβλεψη Χρηματοπιστωτικής Φερεγγυότητας

(Machine Learning and Information Theory Measures in financial solvency prediction)

ΜΗΤΡΟΠΟΥΛΟΥ ΑΙΚΑΤΕΡΙΝΗ

Α.Μ.: 09416017

Επιβλέπων: Κουκουβίνος Χρήστος
Καθηγητής Ε.Μ.Π.

Αθήνα, 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Μηχανική Μάθηση και Μέτρα Πληροφορίας στην πρόβλεψη Χρηματοπιστωτικής Φερεγγυότητας

ΜΗΤΡΟΠΟΥΛΟΥ ΑΙΚΑΤΕΡΙΝΗ

Αθήνα, 2018

Στον άνθρωπο που με έφερε στην ζωή.

Σε ευχαριστώ.

Πρόλογος

Η Επιστήμη των Δεδομένων αποτελεί σχετικά νέο διεπιστημονικό πεδίο, αντικείμενο του οποίου είναι η εξαγωγή γνώσης από αδόμητα ή δομημένα δεδομένα. Αποτελεί τη συνέχεια επιστημών όπως η Στατιστική, η Μηχανική Μάθηση, η Θεωρία Πληροφοριών και η Εξόρυξη Δεδομένων. Η εμφάνιση της Επιστήμης των Δεδομένων έγινε σταδιακά, ξεκινώντας από τα τέλη της δεκαετίας του 1980, μια εποχή όπου μεσουρανούσαν τα σχεσιακά συστήματα βάσεων, τα οποία εξυπηρετούσαν τις ανάγκες αποθήκευσης, διαχείρισης και οργάνωσης επιχειρήσεων και οργανισμών. Ωστόσο, η εξέλιξη της τεχνολογίας τις δύο τελευταίες δεκαετίες, με ορόσημο την ραγδαία εξάπλωση του Internet, διευκόλυνε σημαντικά την διάδοση της πληροφορίας, οδηγώντας αντίστοιχα σε αλματώδη αύξηση των απαιτήσεων διαχείρισης μεγάλου όγκου δεδομένων.

Σήμερα, το πλήθος και η πολυδιαστατικότητα των διαθέσιμων δεδομένων αυξάνεται εκθετικά, οδηγώντας στην καθιέρωση καινοτόμων τεχνικών κατανεμημένης διαχείρισης και ανάλυσης, προς αντικατάσταση των παραδοσιακών σχεσιακών μεθόδων, οι οποίες σταδιακά απαξιώνονται. Τα τεράστια οφέλη που προκύπτουν από την αξιοποίηση των συγκεντρωμένων δεδομένων για την εξόρυξη κρυφού γνωσιακού πλούτου, έχουν μόλις αρχίσει να γίνονται αντιληπτά, χάρη στην εφαρμογή ευφυών αλγορίθμων, που προσφέρονται από το νέο αυτό επιστημονικό πεδίο. Εμφανώς πιο επιδέξιες λύσεις σε σύγκριση με αυτές που είχαν προταθεί από την εφαρμογή καθαρά αιτιοκρατικών προσεγγίσεων γίνονται διαθέσιμες για την επίλυση καθημερινών προβλημάτων. Παράλληλα, πολύπλοκες διαδικασίες φαίνεται να εκτελούνται πλήρως αυτοματοποιημένα και με μεγάλη ακρίβεια, με τη βοήθεια υπολογιστικών συστημάτων τα οποία επιδεικνύουν χαρακτηριστικά που σχετίζονται άμεσα με την ανθρώπινη νοημοσύνη.

Οι προαναφερθείσες - φαινομενικά αδιανόητες - τεχνολογίες είναι σήμερα εφικτές γιατί, στον πυρήνα της, η Επιστήμη Δεδομένων αφορά μια από τις πιο διαχρονικές ανθρώπινες ανάγκες, την ανάγκη για πρόβλεψη. Από την αρχή του κόσμου, ο άνθρωπος στηριζόταν

σε διάφορες μεθόδους για να προβλέψει τις συνέπειες των δράσεών του, από την ερμηνεία των φυσικών φαινομένων και τους προφήτες, ως τον σχηματισμό στοιχειοθετημένων εικασιών βασισμένων σε προηγούμενη γνώση. Είναι γεγονός ότι οι επιστημονικές προβλέψεις είναι πιο αξιόπιστες, ωστόσο περιορίζονται από τα δεδομένα που μπορούμε να παρατηρήσουμε και να μοντελοποιήσουμε συστηματικά. Η Επιστήμη των Δεδομένων έρχεται να διευρύνει το υπάρχον πεδίο, επιτρέποντας στους υπολογιστές να χρησιμοποιούν παρελθοντικά δεδομένα για να προβλέψουν μελλοντικές συμπεριφορές, αποτελέσματα και τάσεις, χωρίς να έχουν προγραμματιστεί ρητά για αυτό τον σκοπό.

Στη σύγχρονη εποχή, είναι φανερό ότι η πρόβλεψη ορισμένων καθημερινών φαινομένων μπορεί να γίνει εύκολα ακόμα και υποσυνείδητα, χάρη την ανθρώπινη νόηση. Αντίθετα, άλλα φαινόμενα είναι τόσο πολύπλοκα, που ίσως παραμείνουν για πάντα ανεξερεύνητα, παρά τη συνδυασμένη χρήση ανθρώπινης και τεχνητής νοημοσύνης. Η Επιστήμη των Δεδομένων μελετά τις ενδιάμεσες περιπτώσεις, επιχειρώντας να δώσει λύσεις εκεί που οι συμβατικές μέθοδοι αποτυγχάνουν. Η πορεία της εξέλιξής της είναι τόσο ιλιγγιώδης και απρόβλεπτη που μπορεί να παρομοιαστεί με ένα αδιάκοπα επιταχυνόμενο τραίνο με άγνωστο προορισμό. Όσο για εμάς, που θέλουμε να μελετήσουμε τις πραγματικές συνεισφορές της στην ανθρώπινη ζωή, δεν έχουμε παρά να γίνουμε επιβάτες του για να μάθουμε που θα μπορεί να μας οδηγήσει.

Περιεχόμενα

Περιεχόμενα Σχημάτων

Περιεχόμενα Πινάκων

Περίληψη

Αντικείμενο της παρούσας Μεταπτυχιακής Διατριβής αποτελεί η ανάδειξη της πρακτικής χρησιμότητας των μεθόδων επιλογής χαρακτηριστικών με χρήση μέτρων πληροφορίας, στη διαδικασία εξόρυξης δεδομένων. Αυτό επιτυγχάνεται μέσω της παρουσίασης και συγκριτικής αξιολόγησης μεθόδων μηχανικής μάθησης, σε συνδυασμό με τη μέθοδο επιλογής χαρακτηριστικών mRMR σε ένα πραγματικό σύνολο δεδομένων από τον τραπεζοοικονομικό τομέα. Συγκεκριμένα, προτείνεται μια ολοκληρωμένη μεθοδολογία ανάλυσης για την πρόβλεψη της χρηματοπιστωτικής φερεγγυότητας των πελατών μιας τράπεζας, δεδομένων των ιστορικών οικονομικών τους στοιχείων. Προκειμένου να γίνει η ορθή ταξινόμηση κάθε πελάτη σε κατηγορίες (ικανότητα ή αδυναμία πληρωμής), η προτεινόμενη μεθοδολογία περιλαμβάνει την υλοποίηση συγκεκριμένων βημάτων, με κυριότερα την προεπεξεργασία των διαθέσιμων δεδομένων, την επιλογή των βέλτιστων χαρακτηριστικών που θα χρησιμοποιηθούν κατά τη διαδικασία πρόβλεψης, την εκπαίδευση των αντίστοιχων μοντέλων μηχανικής μάθησης και την αξιολόγησή τους, ώστε να αναδειχθεί εκείνο που επιτυγχάνει την μεγαλύτερη απόδοση. Παράλληλα, με την προαναφερθείσα πρακτική εφαρμογή, στα πλαίσια της διατριβής γίνεται αναλυτική περιγραφή του απαιτούμενου θεωρητικού και μαθηματικού υποβάθρου των μεθόδων που χρησιμοποιούνται. Μια σύνοψη του περιεχομένου των Κεφαλαίων που περιλαμβάνονται στην εργασία, έχει ως εξής:

Στο Κεφάλαιο 1 γίνεται εισαγωγή στην επιστήμη της Εξόρυξης Δεδομένων, παρατίθενται οι απαραίτητες θεωρητικές έννοιες και παρουσιάζονται τα βασικά στάδια που συνθέτουν τη διαδικασία εξόρυξης γνώσης. Ιδιαίτερη έμφαση δίνεται στο κομμάτι της μηχανικής μάθησης, όπου περιγράφεται η κατηγοριοποίηση σε μεθόδους επιβλεπόμενης, μη-επιβλεπόμενης και ενισχυτικής μάθησης, ενώ δίνονται συνοπτικές παρουσιάσεις μεθόδων κάθε κατηγορίας. Στο ίδιο κεφάλαιο, παρατίθενται παραδείγματα των κυριότερων εφαρμογών εξόρυξης γνώσης σε διάφορους τομείς της ανθρώπινης δραστηριότητας, καθώς και οι σημαντικότερες προκλήσεις που εμφανίζονται κατά την χρήση τους.

Το Κεφάλαιο 2 επικεντρώνεται στην περιγραφή των μεθόδων επιβλεπόμενης μάθησης που θα χρησιμοποιηθούν στο Κεφάλαιο 5 για την επίλυση ενός προβλήματος δυαδικής ταξινόμησης. Στο πλαίσιο αυτό, γίνεται περιγραφή του μαθηματικού υποβάθρου και περιγραφή της αλγοριθμικής διαδικασίας με χρήση ψευδοκώδικα του πολυεπίπεδου δικτύου Perceptron (MLP), του αλγορίθμου Τυχαίων Δασών (Random Forest), του αλγορίθμου k-πλησιέστερων γειτόνων (kNN) και των Μηχανών Διανυσμάτων Υποστήριξης (SVM), ενώ γίνεται εκτενής αναφορά στα πλεονεκτήματα, τα μειονεκτήματα και τις ιδιαιτερότητες της κάθε μεθόδου.

Το Κεφάλαιο 3 αποτελεί μια σύντομη εισαγωγή στην Επιστήμη της Πληροφορίας (Information Theory όπου αναλύονται οι βασικότερες έννοιες που την αφορούν, με έμφαση στα μέτρα πληροφορίας. Παράλληλα, γίνεται αναφορά στο απαραίτητο μαθηματικό υπόβαθρο κάθε μέτρου, ώστε να γίνει κατανοητή τόσο η χρησιμότητά τους όσο και αλληλοσυσχέτιση που εμφανίζουν. Στο πλαίσιο αυτό, παρουσιάζονται οι ορισμοί βασικών μέτρων πληροφορίας, όπως η εντροπία, η αμοιβαία πληροφορία, η υπό συνθήκη πληροφορία, κτλ.) Το Κεφάλαιο λειτουργεί ως προαπαιτούμενο για την κατανόηση της λειτουργίας της μεθόδου επιλογής χαρακτηριστικών, η οποία ακολουθεί στο Κεφάλαιο 4.

Στο Κεφάλαιο 4 γίνεται λεπτομερής αναφορά στις διαφορετικές κατηγορίες μεθόδων επιλογής χαρακτηριστικών (filter, wrapper, embedded) παρουσιάζοντας τα βασικά χαρακτηριστικά, πλεονεκτήματα και μειονεκτήματα της καθεμιάς. Το Κεφάλαιο εμβαθύνει στην περιγραφή της μεθόδου επιλογής χαρακτηριστικών mRMR, μιας από τις πιο πετυχημένες filter μεθόδους, η οποία βασίζεται στο μέτρο της αμοιβαίας πληροφορίας.

Το Κεφάλαιο 5, το οποίο αποτελεί το κύριο πρακτικό κομμάτι της μεταπτυχιακής διατριβής, αφορά τη χρήση των μεθόδων μηχανικής μάθησης που περιγράφηκαν στο Κεφάλαιο 2 σε συνδυασμό με τη μέθοδο επιλογής χαρακτηριστικών που αναλύθηκε στο Κεφάλαιο 4 για την επίλυση ενός πραγματικού προβλήματος δυαδικής ταξινόμησης. Ξεκινά με την εισαγωγή του αναγνώστη στο χρησιμοποιούμενο σύνολο δεδομένων, και συνεχίζει με την διεξοδική ανάλυση της ακολουθούμενης στρατηγικής αντιμετώπισης του προβλήματος, τόσο σε επίπεδο μεθοδολογίας με την περιγραφή των σχετικών βημάτων, όσο και σε αλγοριθμικό επίπεδο με την τμηματική επεξήγηση του αντίστοιχου

κώδικα σε γλώσσα προγραμματισμού R.

Το Κεφάλαιο 6 συγκεντρώνει τα αποτελέσματα και τις παρατηρήσεις που προέκυψαν από την μοντελοποιητική διαδικασία του προηγούμενου κεφαλαίου, με σκοπό την συγκριτική αξιολόγηση των χρησιμοποιούμενων μεθόδων και την τελική επιλογή της βέλτιστης, βάσει συγκεκριμένων κριτηρίων, τα οποία περιγράφονται διεξοδικά. Η ανάλυση των σχετικών αποτελεσμάτων αποτελεί πηγή εξαγωγής χρήσιμων συμπερασμάτων που αφορούν το σύνολο της εργασίας, καθώς και προτάσεων για συνέχιση της έρευνας.

Abstract

The subject of this Postgraduate Dissertation is the emergence of the usefulness of information-driven feature selection methods in the data mining process. This is accomplished by presenting and comparatively evaluating the performance of several machine learning methods, in conjunction with the mRMR feature selection method, applied to a realistic dataset related to the banking and finance fields. More specifically, a comprehensive analytical methodology is proposed for the prediction of financial solvency of a bank's clients, given their historical financial data. In order to successfully classify each client into categories (creditworthiness/default) the proposed methodology incorporates the implementation of specific steps, the most important being the preprocessing of the available data, the selection of the optimal features that are going to be exploited during prediction, the training of the relevant machine learning models, as well as their evaluation, so as to conclude to the one with the greatest performance. In parallel to the aforementioned application, a thorough description of the implemented methods' relevant theoretical and mathematical background is provided, as part of the dissertation. A summary of the contents of each Chapter contained in this Postgraduate Dissertation is as follows:

Chapter 1 introduces the Data Mining science, as well as its necessary theoretical concepts and presents the basic stages that make up the data mining process. Particular emphasis is placed on the machine learning part, by elaborating on its categorization in supervised, unsupervised and reinforcement learning, while providing summarizing presentations of methods belonging to each category. In the same chapter, examples of the main knowledge discovery applications in various areas of human activity are presented, as well as the most important challenges arising during its application to real-world problems.

Chapter 2 focuses on the description of the specific supervised learning methods that are going to be implemented in Chapter 5 for the solution of a binary classification problem.

In this context, the explanation of the mathematical background and the algorithmic procedure of the following methods is given: Multi-Layer Perceptron (MLP), Random Forest (RF), k-nearest neighbours (kNN) and Support Vector Machines (SVM). Extensive reference to the advantages, disadvantages and unique characteristics of each method is also made.

In Chapter 3, the basic concepts related to the development of Information Theory are formulated and analyzed, with emphasis on information measures. Furthermore, reference is made to the necessary mathematical background of every measure, in order to understand both their usefulness and correlation between them. In this sense, the definitions of the main information measures like entropy, mutual information, conditional information, etc. are presented. This Chapter works as a prerequisite for understanding the functionality of the feature selection method which follows in the next Chapter.

Chapter 4 provides a detailed overview of the different categories of feature selection methods (filter, wrapper, embedded), showing the key features, pros and cons of each. The chapter delves deeper into the description of the mRMR feature selection method, one of the most successful filter methods, which is based on the measure of mutual information.

Chapter 5, which forms the main practical part of this postgraduate dissertation, involves the use of the machine learning methods described in Chapter 2, in conjunction with the feature selection method discussed in Chapter 4, to solve a real binary classification problem. It begins with the introduction of the reader to the dataset in use, and continues with a thorough analysis of the proposed strategic approach to the problem, both at the methodology level with the description of the relevant steps and at the algorithmic level with the explanation of the corresponding code, developed in R programming language.

Chapter 6 summarizes the results and observations that emerged from the modeling procedures of the previous chapter, in order to evaluate the methods used and finally select the optimal, based on specific criteria which are described in detail. The analysis of the relevant results is a source of useful conclusions regarding the whole work, as well as of suggestions for future research.

Ευχαριστίες

Θα ήθελα πρωτίστως να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Χρήστο Κουκουβίνο που μου έδωσε την ευκαιρία να προσανατολίσω τη μελέτη μου προς το συγκεκριμένο επιστημονικό πεδίο. Επιπλέον, θα ήθελα να ευχαριστήσω την υποψήφια διδάκτορα Λάππα Αγγελική για την συμβολή της στην περάτωση της εργασίας, καθώς υπήρξε διαθέσιμη και συνεργάσιμη. Θα ήθελα επίσης να ευχαριστήσω ιδιαιτέρως τον υποψήφιο διδάκτορα Παρασκευά Σπύρο, για την πολύτιμη βοήθειά του, την ακούραστη προθυμία του και τον συμβουλευτικό του ρόλο.

Οι ευχαριστίες επεκτείνονται σε δύο ακόμα άτομα. Στον σύντροφό μου Δημήτρη Παπαδόπουλο, ο οποίος είναι στο πλευρό μου με πίστη και επιμονή όλα αυτά τα χρόνια και με οπλίζει με ψυχική δύναμη και πνευματική ηρεμία και στον φίλο μου και καθηγητή κ. Αναστάσιο Μαυραγάνη, ο οποίος ήταν επίσης δίπλα μου καθ'όλη την διάρκεια των σπουδών μου, προσφέροντάς μου διαρκή καθοδήγηση.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου. Την μητέρα μου που φρόντισε να μου παρέχει όλα τα απαραίτητα ώστε να ολοκληρώσω και να εξελίξω τις σπουδές μου και μαζί με τον πατέρα μου, ο οποίος δεν είναι εν ζωή, φρόντισαν να μου καλλιεργήσουν κατά τα παιδικά μου χρόνια όλα τα απαραίτητα χαρακτηριστικά ώστε να γίνω ολοκληρωμένος άνθρωπος, με αυτοπεποίθηση, εσωτερική γαλήνη και διάθεση για συνεχόμενη εξέλιξη.

Μητροπούλου Αικατερίνη

Εθνικό Μετσόβιο Πολυτεχνείο.
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Αθήνα, 2018

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή στην Εξόρυξη δεδομένων

1.1 Εισαγωγή

Από τα πρώτα στάδια της ύπαρξής του ο άνθρωπος βρίσκεται σε μια συνεχόμενη προσπάθεια για αποτελεσματική και ταχεία επικοινωνία για τη μετάδοση διαφόρων πληροφοριών που αφορούν την καθημερινότητά του. Στη σήμερον ημέρα, η σύγκλιση της προόδου υπολογιστικών συστημάτων έχει επιφέρει την ραγδαία ανάπτυξη των μέσων καταγραφής, αποθήκευσης, επεξεργασίας και μετάδοσης δεδομένων, καθώς επίσης και των συναφών τεχνολογιών επικοινωνίας. Σε μια κοινωνία που παράγει συνεχώς νέα πληροφορία και ο τεράστιος όγκος δεδομένων που παράγεται συσσωρεύεται σε βάσεις και αποθήκες δεδομένων (data warehouses), δημιουργήθηκε με την σειρά της η ανάγκη για την αξιοποίηση τους με αποδοτικό τρόπο.

Η ραγδαία αυτή τεχνολογική εξέλιξη στη συλλογή και αποθήκευση δεδομένων τις τελευταίες δεκαετίες έχει επιτρέψει στους οργανισμούς να συγκεντρώνουν τεράστια ποσά δεδομένων. Ωστόσο, η εξαγωγή χρήσιμων πληροφοριών από αυτά έχει αποδειχθεί εξαιρετικά δύσκολο εγχείρημα. Συχνά, τα παραδοσιακά εργαλεία και οι μέχρι πρότινος χρησιμοποιούμενες τεχνικές εξαγωγής πληροφοριών δεν μπορούν να χρησιμοποιηθούν, εξαιτίας του τεράστιου όγκου που χαρακτηρίζει τα αποθηκευμένα δεδομένα. Μερικές φορές, η ημι-δομημένη φύση των συσσωρευμένων δεδομένων συνεπάγεται ότι οι υπάρχουσες προσεγγίσεις δεν δύνανται να εφαρμοστούν ακόμη και αν ο όγκος τους είναι σχετικά μικρός. Σε άλλες περιπτώσεις πάλι, οι υφιστάμενες τεχνικές ανάλυσης δεν είναι καν σχεδιασμένες ώστε να δίνουν απάντηση στα ζητούμενα ερωτήματα, απαιτώντας κατά συνέπεια την ανάπτυξη νέων μεθόδων. Το περίφημο ρητό του John

Naisbitt "Είμαστε πνιγμένοι στα δεδομένα, αλλά λιμοκτονούμε για γνώση" συνοψίζει εύστοχα αυτή την αναγκαιότητα. Εκτιμάται ότι πάνω από 6 zettabytes ($6 \cdot 10^{21}$ bytes) πληροφοριών παράγονται κάθε χρόνο σε παγκόσμιο επίπεδο. Προκειμένου να αποκαλυφθούν τα ενδιαφέροντα πρότυπα που κρύβουν αυτά τα δεδομένα, χρειάζονται αποτελεσματικές και υπολογιστικά επεκτάσιμες μέθοδοι εξαγωγής γνώσης. Η γνώση, άλλωστε, είναι πολύτιμη μόνο όταν μπορεί να χρησιμοποιηθεί αποτελεσματικά για τη βελτίωση της διαδικασίας λήψης αποφάσεων.

Είναι φανερό λοιπόν ότι ο τεράστιος αυτός όγκος δεδομένων που προκύπτει καθημερινά, δεν μπορεί να αξιοποιηθεί έτσι όπως είναι, ακατέργαστος. Το ερώτημα που προκύπτει είναι εάν υπάρχει τρόπος να διαχειριστούμε τις βάσεις δεδομένων που καθημερινά ανανεώνονται με επιπλέον πληροφορία. Όλες αυτές οι απαιτήσεις προκάλεσαν το ενδιαφέρον και οδήγησαν στην επιστήμη της Εξόρυξης Δεδομένων (Data Mining).

Η επιστήμη της Εξόρυξης Δεδομένων συνδυάζει παραδοσιακές μεθόδους ανάλυσης με εξελιγμένους αλγορίθμους για την επεξεργασία μεγάλων όγκων δεδομένων. Στοχεύοντας στην εξαγωγή χρήσιμων μοτίβων και συμπερασμάτων, αξιοποιεί πληθώρα μεθόδων από διαφορετικά τεχνολογικά πεδία, συνδυάζοντας την επιστήμη της στατιστικής και της πληροφορικής, μεθόδους που βασίζονται στην Τεχνητή Νοημοσύνη (Artificial Intelligence) και την Μηχανική Μάθηση (Machine Learning) (Tan et al., 2006). Αποτελεί αναπόσπαστο μέρος της ανακάλυψης γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases - KDD), η οποία αποτελεί τη συνολική διαδικασία μετατροπής πρωτογενών και ακατέργαστων δεδομένων σε χρήσιμες πληροφορίες.

Έχει ενδιαφέρον να αναφέρουμε ότι υπάρχουν αντικρουόμενες απόψεις όσον αφορά τον σαφή ορισμό της Εξόρυξης Δεδομένων. Ο πιο περιεκτικός, ίσως, που προέρχεται από τους Hand et al., 2001 παρατίθεται παρακάτω.

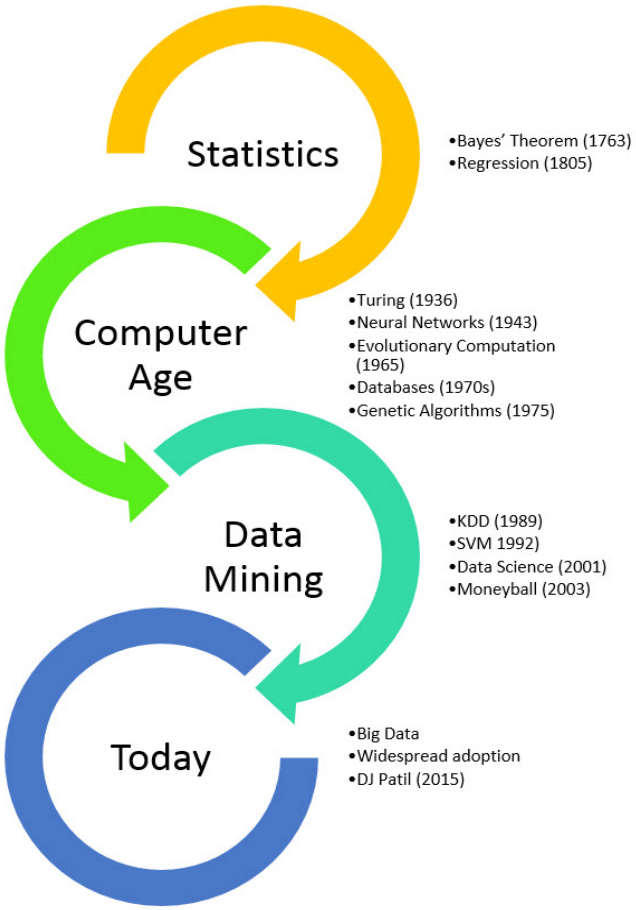
"Εξόρυξη Δεδομένων (Data Mining) είναι η ανάλυση – συνήθως τεράστιων – παρατηρούμενων συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων."

Η έννοια “σχέσεις” που περιέχει ο ορισμός αναφέρεται στα μοντέλα (models) ή πρότυπα (patters) που χαρακτηρίζουν τα δεδομένα. Βασικός στόχος της Εξόρυξης Δεδομένων είναι η περιγραφή και η πρόβλεψη των προτύπων. Τα πρότυπα αυτά, τα οποία συναντάμε σε διάφορες μορφές, όπως συσχετίσεις, ανωμαλίες, συστάδες, κλάσεις κ.λπ., αποτελούν δομές ή περιστατικά, που εμφανίζονται στα δεδομένα και έχουν κάποια ιδιαίτερη σημασία από στατιστικής πλευράς. Η αναγνώρισή τους γίνεται μέσω γραμμικών εξισώσεων, κανόνων, διάκρισης σε συστάδες, απόδοσης γραφημάτων και δομών σε μορφή δέντρου, καθώς και επαναλαμβανόμενων προτύπων σε μορφή χρονοσειρών.

Τα μοντέλα για την Εξόρυξη Δεδομένων χωρίζονται σε δύο κατηγορίες (Fayyad et al. - a, 1996):

- Μοντέλα Πρόβλεψης (Predictive Models). Χρησιμοποιούν μερικές μεταβλητές για να προβλέψουν άγνωστες ή μελλοντικές μεταβλητές. Για παράδειγμα μπορεί να εισάγουμε ορισμένα στοιχεία για ένα πελάτη και να προσπαθήσουμε να προβλέψουμε αν αυτός θα κηρύξει πτώχευση, βασισμένοι σε ήδη υπάρχοντα στοιχεία. Είναι φανερό ότι τα συγκεκριμένα μοντέλα απαιτούν πολλά στοιχεία.
- Περιγραφικά Μοντέλα (Descriptive Models). Προσπαθούν να ανακαλύψουν, ήδη υπάρχουσες τάσεις και κανόνες, που δεν είναι ορατοί χωρίς την Εξόρυξη Δεδομένων. Για παράδειγμα, αν κάποιος αγοράσει έναν νέο φορητό υπολογιστή, πιθανόν να προβεί και στην αγορά μιας τσάντας μεταφοράς γι' αυτόν.

1.2 Ιστορική Αναδρομή



ΣΧΗΜΑ 1.1 Ορόσημα στην Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων είναι παντού πλέον γύρω μας, αλλά η ιστορία της ξεκινάει πολλά χρόνια πριν, για να φτάσει σε αυτό που αποκαλούμε σήμερα “Big Data”. Παρακάτω παρουσιάζονται τα πλέον σημαντικότερα ορόσημα και οι "πρωτιές" στην ιστορία της εξόρυξης δεδομένων καθώς και το πώς εξελίσσεται και συνδυάζεται με την επιστήμη των δεδομένων και τα Big Data (Ramzan et al., 2014).

Ξεκινώντας από το 1763, το έγγραφο του Thomas Bayes που δημοσιεύεται μεταθάνατον με τίτλο Θεώρημα του Bayes (Bayes' theorem), αφορά ένα θεώρημα για τη συσχέτιση της εκ των υστέρων πιθανότητας (posterior probability) με την εκ των προτέρων πιθανότητα (prior probability). Αποδεικνύεται θεμελιώδους σημασίας για την εξόρυξη δεδομένων και την επιστήμη των πιθανοτήτων, καθώς επιτρέπει την κατανόηση σύνθετων πραγματικοτήτων που βασίζονται σε εκτιμώμενες πιθανότητες. Περίπου 40 χρόνια μετά (1805) ο Adrien-Marie Legendre και ο Carl Friedrich Gauss εφαρμόζουν παλινδρόμηση για να καθορίσουν τις τροχιές των σωμάτων γύρω από τον Ήλιο (κομήτες και πλανήτες). Ο στόχος της ανάλυσης παλινδρόμησης είναι να εκτιμηθούν οι σχέσεις μεταξύ των μεταβλητών και η συγκεκριμένη μέθοδος που χρησιμοποιείται σε αυτή την περίπτωση είναι η μέθοδος των ελάχιστων τετραγώνων. Η παλινδρόμηση πλέον είναι ένα από τα βασικά εργαλεία στην εξόρυξη δεδομένων.

Κάνοντας ένα μεγάλο βήμα αρκετά χρόνια μπροστά στο 1936 συναντάμε την αυγή της

εποχής των υπολογιστών, η οποία καθιστά δυνατή τη συλλογή και επεξεργασία μεγάλων ποσοτήτων δεδομένων. Σε ένα paper του 1937, με τίτλο “On Computable Numbers”, ο Alan Turing παρουσίασε την ιδέα μιας παγκόσμιας μηχανής ικανής να εκτελεί υπολογισμούς όπως οι σύγχρονοι υπολογιστές. Ο σύγχρονος υπολογιστής βασίζεται στις ιδέες που πρωτοεισήγαγε ο Turing. Λίγο αργότερα, το 1943 Ο Warren McCulloch και ο Walter Pitts ήταν οι πρώτοι που δημιούργησαν ένα εννοιολογικό μοντέλο ενός νευρωνικού δικτύου. Σε ένα paper με τίτλο “A logical calculus of the ideas immanent in nervous activity”, περιγράφουν την ιδέα ενός νευρώνα μέσα σε ένα δίκτυο. Κάθε ένας από αυτούς τους νευρώνες μπορεί να κάνει 3 πράγματα: να λαμβάνει εισόδους, να επεξεργάζεται αυτές τις εισόδους και εν τέλει να παράγει αποτελέσματα. Το 1965 η επιστήμη της εξόρυξης δεδομένων έχει αρχίσει να εδραιώνεται, και η πρώτη εταιρία για εφαρμογές εξελικτικού προγραμματισμού ιδρύεται, από τον Lawrence J. Fogel με εμπορική επωνυμία Decision Science, Inc.. Ήταν η πρώτη εταιρεία στα χρονικά που είχε σαν καθορισμένο στόχο την εφαρμογή του εξελικτικού υπολογισμού (evolutionary computation) για την επίλυση προβλημάτων του πραγματικού κόσμου.

Μπαίνοντας στην δεκαετία του '70 πλέον, έχοντας στην διάθεσή μας εξελιγμένα συστήματα διαχείρισης βάσεων δεδομένων, κατέστη δυνατή η αναζήτηση σε terabytes και petabytes δεδομένων. Επιπλέον, οι αποθήκες δεδομένων επιτρέπουν στους χρήστες να μετακινηθούν από τον κλασσικό τρόπο σκέψης που είναι προσανατολισμένος στις συναλλαγές σε έναν πιο αναλυτικό τρόπο αντιμετώπισης των δεδομένων. Ωστόσο, η εξόρυξη σύνθετων πληροφοριών από αυτές τις αποθήκες δεδομένων αξιοποιώντας πολυδιάστατα μοντέλα είναι ακόμα πολύ περιορισμένη. Στα μέσα της δεκαετίας, το 1975, Ο John Henry Holland έγραψε το “Adaptation in Natural and Artificial Systems”, ένα πρωτοποριακό βιβλίο για τους γενετικούς αλγόριθμους. Είναι το βιβλίο που ουσιαστικά ξεκίνησε αυτόν τον τομέα έρευνας, παρουσιάζοντας τα θεωρητικά θεμέλια του αντικειμένου και διερευνώντας τις πρακτικές εφαρμογές του.

Προχωρώντας άλλη μια δεκαετία, το '80, η HNC κατοχυρώνει εμπορικά τη φράση “database mining”, συνδέοντας άρρηκτα τον εαυτό της με την εξόρυξη βάσεων δεδομένων. Το εμπορικό αυτό σήμα προορίζεται να κατοχυρώσει ένα προϊόν που ονομάζεται DataBase Mining Workstation. Ένα εργαλείο γενικής χρήσης που χρησίμευε για την δημιουργία μοντέλων νευρωνικών δικτύων και τώρα δεν είναι πλέον διαθέσιμο.

Είναι επίσης κατά τη διάρκεια αυτής της περιόδου που έγινε επιτρεπτό για τους εξελιγμένους αλγορίθμους να μπορούν να "μαθαίνουν" πιθανές σχέσεις από τα δεδομένα, που θα επιτρέψουν στους εμπειρογνώμονες του αντικειμένου να διασαφηνίσουν τι μπορεί να σημαίνουν οι σχέσεις αυτές. Στο τέλος της δεκαετίας (1989) ο όρος "ανακάλυψη γνώσεων σε βάσεις δεδομένων" (“Knowledge Discovery in Databases” - KDD) δημιουργείται από τον Gregory Piatetsky-Shapiro. Είναι επίσης, η ίδια στιγμή που συνδιοργανώνει το πρώτο workshop που φέρει επίσης το όνομα KDD (Piatetsky-Shapiro, 1991).

Ακόμη μια δεκαετία μπροστά, το '90, ο όρος "εξόρυξη δεδομένων" εμφανίζεται στην κοινότητα των βάσεων δεδομένων. Οι εταιρείες λιανικής και η χρηματοοικονομική κοινότητα χρησιμοποιούν την εξόρυξη δεδομένων για να αναλύσουν δεδομένα και να αναγνωρίσουν τάσεις για την αύξηση της πελατειακής τους βάσης, να προβλέψουν διακυμάνσεις των επιτοκίων, των τιμών των μετοχών και της ζήτησης των πελατών. Το '92 οι Bernhard E. Boser, Isabelle M. Guyon και Vladimir N. Vapnik προτείνουν μια βελτίωση στην αρχική έκδοση του αλγορίθμου “Μηχανή Διανυσμάτων Υποστήριξης” (Support Vector Machine – SVM) που επιτρέπει τη δημιουργία μη γραμμικών ταξινομητών. Το '93 ο Gregory Piatetsky-Shapiro ξεκινάει το ενημερωτικό δελτίο με τίτλο “Knowledge Discovery Nuggets” (KDnuggets). Αρχικά σχεδιάστηκε για τη διασύνδεση των ερευνητών που παρακολούθησαν το εργαστήριο KDD. Ωστόσο, το KDnuggets.com φαίνεται να έχει ένα πολύ ευρύτερο κοινό πια.

Μπαίνουμε στην νέα χιλιετία. Αν και ο όρος επιστήμη των δεδομένων υπάρχει από τη δεκαετία του 1960, δεν είναι μέχρι το 2001 που ο William S. Cleveland την παρουσιάζει για πρώτη φορά ως ανεξάρτητο επιστημονικό κλάδο. Όσον αφορά τις Build Data Science Teams, ο DJ Patil και ο Jeff Hammerbacher χρησιμοποιούν τον όρο για να περιγράψουν τους ρόλους τους στο LinkedIn και στο Facebook. Το 2003 δημοσιεύεται το Moneyball, από τον Michael Lewis, και αλλάζει τον τρόπο με τον οποίο λειτουργούν πολλά μεγάλα γραφεία εξυπηρέτησης πελατών. Το Oakland Athletics χρησιμοποιεί μια στατιστική προσέγγιση που βασίζεται σε δεδομένα για να διακρίνει σε κλάσεις ποιότητας τους παίκτες που ήταν υποτιμημένοι και φθηνότεροι να αποκτηθούν. Με αυτόν τον τρόπο, συγκεντρώνουν με επιτυχία μια ομάδα παιχτών που τους έφερε στα play-off του 2002 και του 2003 με μόλις το 1/3 του μέσου προϋπολογισμού της ομάδας.

Φτάνουμε στο παρόν. Τον Φεβρουάριο του 2015, ο DJ Patil γίνεται ο πρώτος επικεφαλής Data Scientist στο Λευκό Οίκο. Η επιστήμη της εξόρυξης δεδομένων αποτελεί πλέον αναπόσπαστο κομμάτι της σημερινής κοινωνίας. Είναι ευρέως διαδεδομένη στις επιχειρήσεις, την επιστήμη, τη μηχανική καθώς και την ιατρική, αναφέροντας μόνο μερικούς από τους κλάδους στους οποίους έχει εισχωρήσει δίνοντας λύσεις σε ποικίλα προβλήματα. Η εξόρυξη δεδομένων από συναλλαγές με πιστωτικές κάρτες, κινήσεις των χρηματιστηριακών αγορών, δεδομένα εθνικής ασφάλειας, δεδομένα αλληλουχιών γονιδιωμάτων και κλινικές δοκιμές είναι μόνο η κορυφή του παγόβουνου που ονομάζεται “εφαρμογές εξόρυξης δεδομένων” (data mining applications). Όροι όπως το Big Data είναι πλέον συνηθισμένοι, με τη συλλογή των δεδομένων να γίνεται φθηνότερη και το πλήθος των συσκευών που είναι σε θέση να συλλέγουν δεδομένα να πολλαπλασιάζεται συνεχώς.

1.3 Βασικές Έννοιες

Όταν επεξεργαζόμαστε μια τεράστια βάση δεδομένων, είναι πολύ πιθανό να ανακαλύψουμε την ύπαρξη “κρυμμένης γνώσης”. Δηλαδή, μπορεί να εντοπίσουμε συσχετίσεις, αλληλεξάρτηση ή ομαδοποιήσεις μεταξύ των δεδομένων, πράγματα τα οποία μπορεί να μην είναι άμεσα εμφανή από τα δεδομένα. Αυτό το είδος γνώσης θεωρείται ότι ναι μεν δεν είναι εκ των προτέρων διαθέσιμο, αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Υπό αυτές τις συνθήκες, κρίνεται απαραίτητη η αυτοματοποιημένη ανάκτηση γνώσης από τα δεδομένα, χωρίς να απαιτείται η συνεχόμενη ενασχόληση κάποιου ανθρώπου, η οποία θα υποστηρίζεται κατά κύριο λόγο από την εφαρμογή αλγορίθμων. Στόχος είναι η ανακάλυψη αυτής της κρυμμένης γνώσης. Αυτήν την ανάγκη έρχεται να καλύψει η επιστήμη της εξόρυξης δεδομένων, η οποία αποτελεί και τον πυρήνα της διαδικασίας ανακάλυψης της γνώσης από βάσεις δεδομένων (KDD).

Η διαδικασία KDD αναφέρεται στη διεργασία εξόρυξης γνώσης από μεγάλες αποθήκες δεδομένων. Ο όρος εξόρυξη δεδομένων χρησιμοποιείται πολλές φορές ως συνώνυμο της

KDD, αλλά αποτελεί αναφορά στις πραγματικές τεχνικές (εργαλεία) που χρησιμοποιούνται για την ανάλυση και εξαγωγή της γνώσης από διάφορα σύνολα δεδομένων. Για να είναι σαφής η διαφορά μεταξύ διαδικασίας και εργαλείων, ο όρος KDD χρησιμοποιείται για την περιγραφή ολόκληρης της διαδικασίας ανακάλυψης γνώσης από ένα σύνολο δεδομένων, ενώ ο όρος εξόρυξη δεδομένων αναφέρεται μεμονωμένα στις τεχνικές που χρησιμοποιούνται για την ανακάλυψη γνώσης (Zaïane, 1999). Οι τεχνικές αυτές βασίζονται κατά κόρον στην επιστήμη της στατιστικής και της πληροφορικής, και πιο συγκεκριμένα στη χρήση μεθόδων που βασίζονται στην τεχνητή νοημοσύνη και την μηχανική μάθηση.

Ένας άλλος όρος που χρησιμοποιείται συχνά αντί της εξόρυξης δεδομένων είναι η "εξόρυξη γνώσης". Θεωρείται, όμως, ότι ο όρος αυτός δε δίνει έμφαση στην ανάλυση και εξαγωγή προτύπων. Ο όρος εξόρυξη δεδομένων αντιπροσωπεύει καλύτερα τη διαδικασία εύρεσης δομών γνώσης που περιγράφουν με ακρίβεια σύνολα πρωτογενών δεδομένων. Οι δομές αυτές αναδεικνύουν κρυμμένη γνώση (συνάφειες / κανόνες) που δεν είναι άμεσα ορατή και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων.

Η εξόρυξη δεδομένων τυπικά ασχολείται με δεδομένα που έχουν συλλεχθεί για κάποιον άλλο σκοπό εκτός της ανάλυσης τους στα πλαίσια του data mining. Οι αντικειμενικοί στόχοι της εξόρυξης δεδομένων δεν παίζουν κανένα ρόλο στην στρατηγική που ακολουθείται για τη συλλογή των δεδομένων. Αυτό είναι ένα σημείο στο οποίο η εξόρυξη δεδομένων διαφοροποιείται από τις συνηθισμένες στατιστικές αναλύσεις οι οποίες συχνά συλλέγουν δεδομένα χρησιμοποιώντας αποτελεσματικές στρατηγικές για να απαντήσουν σε συγκεκριμένα ερωτήματα. Επίσης, οι κλασικές στατιστικές αναλύσεις χρησιμοποιούν μικρά σύνολα δεδομένων, σε αντίθεση με το data mining που χρησιμοποιεί σύνολα τα οποία εντάσσονται στην κατηγορία των big data.

Για τη διεξαγωγή αποτελεσματικής εξόρυξης δεδομένων, πρέπει πρώτα να εξεταστεί τι είδους χαρακτηριστικά αναμένεται να έχει ένα εφαρμοσμένο σύστημα ανακάλυψης γνώσης και τι είδους προκλήσεις μπορεί να αντιμετωπιστούν στην ανάπτυξη των τεχνικών εξόρυξης δεδομένων (Chen et al., 1997).

i. **Χειρισμός διαφορετικών τύπων δεδομένων.** Πλέον υπάρχουν πολλά είδη δεδομένων και βάσεων δεδομένων που χρησιμοποιούνται στις διαφορετικές

εφαρμογές. Γι'αυτό το λόγο μπορεί κανείς να υποθέσει ότι ένα σύστημα ανακάλυψης γνώσης θα πρέπει να είναι σε θέση να εκτελέσει αποτελεσματική εξόρυξη δεδομένων σε διαφορετικά είδη δεδομένων. Δεδομένου ότι οι περισσότερες διαθέσιμες βάσεις δεδομένων είναι σχεσιακές, είναι σημαντικό ένα σύστημα εξόρυξης δεδομένων να μπορεί να εκτελεί αποτελεσματική και αποδοτική ανακάλυψη γνώσης σε σχεσιακά δεδομένα. Επιπλέον, πολλές βάσεις δεδομένων περιέχουν σύνθετους τύπους δεδομένων, όπως δομημένα δεδομένα και πολύπλοκα αντικείμενα δεδομένων, δεδομένων πολυμέσων, χωρικά και χρονικά δεδομένα, δεδομένα συναλλαγών κλπ. Ένα ισχυρό σύστημα θα πρέπει να είναι σε θέση να εκτελέσει αποτελεσματική εξόρυξη δεδομένων για τέτοιου είδους πολύπλοκους τύπους δεδομένων. Ωστόσο, η ποικιλία των τύπων δεδομένων και διαφορετικών στόχων της εξόρυξης δεδομένων καθιστούν μη ρεαλιστικό να αναμένουμε ένα ενιαίο σύστημα εξόρυξης δεδομένων το οποίο θα μπορεί να χειριστεί όλα αυτά τα είδη των δεδομένων. Καλό θα ήταν να διαμορφωθούν εξειδικευμένα συστήματα για την εξόρυξη γνώσης πάνω σε όλους αυτούς τους σύνθετους τύπους δεδομένων.

ii. **Χρησιμότητα, βεβαιότητα και εκφραστικότητα των αποτελεσμάτων της εξόρυξης δεδομένων.** Η ανακάλυψη γνώσης θα πρέπει να απεικονίζει με ακρίβεια το περιεχόμενο της βάσης δεδομένων και είναι χρήσιμη για ορισμένες εφαρμογές. Η ακρίβεια των αποτελεσμάτων θα πρέπει να εκφράζεται μέσω κάποιων μέτρων αβεβαιότητας, προσεγγιστικά ή ποσοτικά. Θόρυβος και ακραίες τιμές θα πρέπει να αντιμετωπίζονται αποτελεσματικά από τα συστήματα εξόρυξης δεδομένων. Αυτό παρακινεί μια συστηματική μελέτη πάνω στην μέτρηση της ποιότητας της εξορυγμένης γνώσης, η οποία θα χρησιμοποιεί στατιστικά ή αναλυτικά μοντέλα, μοντέλα προσομοίωσης, καθώς και τα εργαλεία αυτών.

iii. **Αποδοτικότητα και εξελιξιμότητα αλγορίθμων εξόρυξης δεδομένων.** Για να είναι αποδοτική η εξόρυξη γνώσης από μεγάλα σύνολα δεδομένων, οι αλγόριθμοι ανακάλυψης γνώσης πρέπει να είναι κατάλληλα προσαρμοσμένοι σε αυτά ώστε να είναι αποτελεσματικοί. Δηλαδή, ο χρόνος λειτουργίας ενός αλγορίθμου εξόρυξης δεδομένων πρέπει να είναι προβλέψιμος και αποδεκτός σε μεγάλες βάσεις δεδομένων. Πρακτικά αυτό σημαίνει ότι αλγόριθμοι με εκθετική ή πολυωνυμική πολυπλοκότητα δεν θεωρούνται αποδοτικοί στη χρήση.

iv. **Έκφραση των αποτελεσμάτων της εξόρυξης δεδομένων με διαφορετικούς τρόπους.** Διαφορετικά είδη γνώσης μπορεί να ανακαλυφθούν από ένα μεγάλο όγκο δεδομένων. Επίσης, μπορεί κανείς να θέλει να εξετάσει την αποκτηθείσα γνώση από διαφορετική άποψη και να την παρουσιάσει σε διαφορετικές μορφές. Επιπλέον θα ήταν χρήσιμο να εκφραστούν τόσο τα αιτήματα εξόρυξης δεδομένων όσο και η εξορυγμένη γνώση σε γλώσσες υψηλού επιπέδου ή μέσω γραφικών διεπαφών των χρηστών, έτσι ώστε το έργο εξόρυξης δεδομένων να μπορεί να εφαρμόζεται από μη ειδικούς και η εξορυγμένη γνώση να είναι κατανοητή και άμεσα χρησιμοποιήσιμη από όλους. Αυτό απαιτεί επίσης τα συστήματα ανακάλυψης γνώσης να υιοθετήσουν εκφραστικές τεχνικές αναπαράστασης γνώσης.

v. **Εξόρυξη πληροφοριών από διαφορετικές πηγές δεδομένων.** Το φαινόμενο ύπαρξης πλήθους διαφόρων πηγών δεδομένων που οφείλεται στην ευρέως διαθέσιμη σύνδεση υπολογιστών στο διαδίκτυο, οδηγεί στην δημιουργία μεγάλων κατανεμημένων και ετερογενών βάσεων δεδομένων. Επιπλέον, το τεράστιο μέγεθος των βάσεων δεδομένων, η υψηλή κατανομή των δεδομένων και η υπολογιστική πολυπλοκότητα ορισμένων μεθόδων εξόρυξης δεδομένων προωθούν την ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων εξόρυξης δεδομένων.

vi. **Διαλογική εξόρυξη γνώσης σε πολλαπλά εννοιολογικά επίπεδα.** Δεδομένου ότι είναι δύσκολο να προβλεφθεί τι ακριβώς θα μπορούσε να ανακαλυφθεί από μια βάση δεδομένων, θα μπορούσε να καθοριστεί μια σειρά ερωτήσεων της εξόρυξης δεδομένων, προκειμένου να διαμορφωθεί η εστίαση στα δεδομένα, να δημιουργηθεί ένα λεπτομερέστερο επίπεδο εξόρυξης γνώσης και να παρατηρηθούν τα αποτελέσματα της εξόρυξης δεδομένων σε πολλαπλά επίπεδα και από διαφορετικές πτυχές.

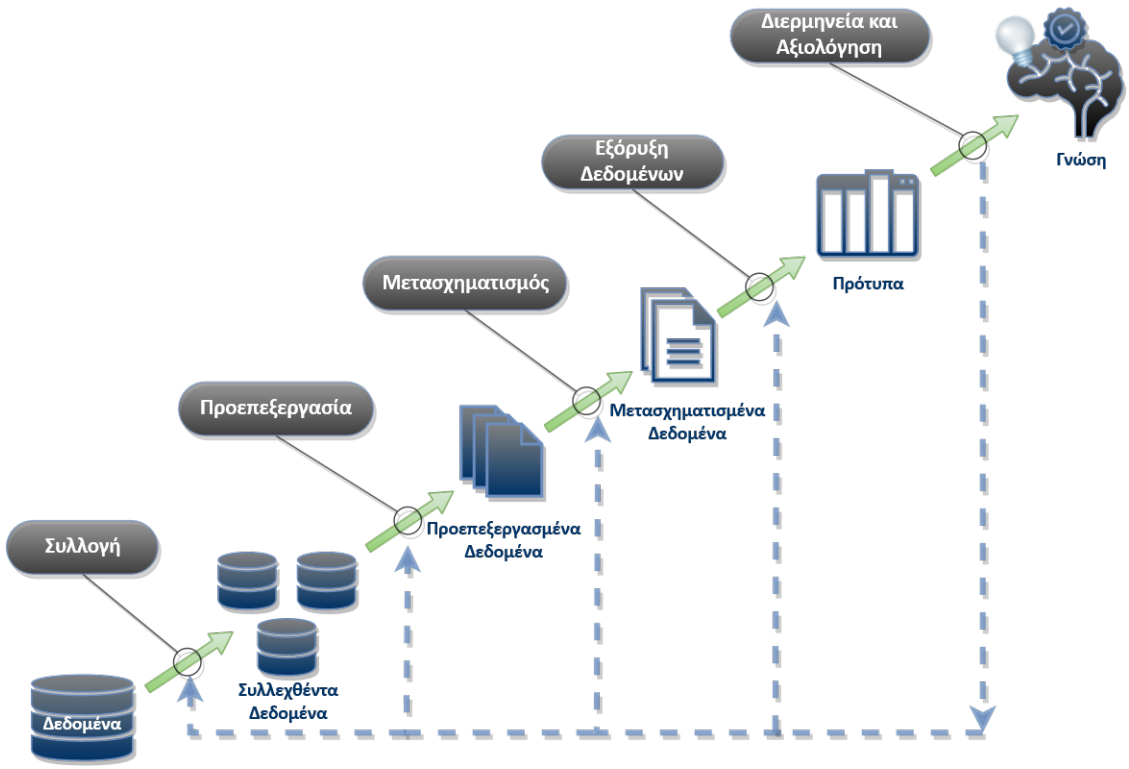
vii. **Προστασία της ιδιωτικότητας και της ασφάλειας των δεδομένων.** Όταν τα δεδομένα μπορούν να παρατηρηθούν από πολλές διαφορετικές οπτικές γωνίες, απειλείται η προστασία και η αποκλειστικότητα τους. Είναι σημαντικό να μελετηθεί πότε μπορεί να οδηγηθούμε σε παραβίαση της ιδιωτικότητας μέσω της KDD και τι αντίμετρα μπορούν να αναπτυχθούν για την αποφυγή αποκάλυψης ευαίσθητων πληροφοριών.

Σημειώνεται ότι κάποιοι από αυτές τις απαιτήσεις μπορεί να προκαλέσουν αντικρουόμενους

στόχους στην ανάπτυξη του data mining. Είναι όμως σημαντικό να παρουσιαστεί η γενική εικόνα των απαιτήσεων όσον αφορά την αποτελεσματική εξόρυξη δεδομένων.

1.4 Στάδια Εξόρυξης Γνώσης από τα Δεδομένα

Όπως προαναφέρθηκε, η επιστήμη της Εξόρυξης Δεδομένων (Data Mining) ερευνά τη σειρά από τεχνικές που είναι απαραίτητες για την αποδοτική διαχείριση των δεδομένων και βασίζονται σε ανάπτυξη αλγορίθμων. Είναι ουσιαστικά μία ημι-αυτοματοποιημένη διαδικασία, σκοπός της οποίας είναι να αναλύσει έναν μεγάλο όγκο δεδομένων που αφορούν ένα συγκεκριμένο πρόβλημα, συνήθως είτε εμπορικού είτε επιστημονικού ενδιαφέροντος, για την παραγωγή προτύπων (patterns), καθώς και την περιγραφή και την πρόβλεψη αυτών.



ΣΧΗΜΑ 1.2 Στάδια Εξόρυξης Δεδομένων

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων - ΑΓΒΔ (Knowledge Discovery in Databases - KDD) είναι μια συγκροτημένη μεθοδολογία και αποτελείται από συγκεκριμένα στάδια (Fayyad et al. - b, 1996). Πρόκειται για την αποκάλυψη ή παραγωγή χρήσιμης, λειτουργικής και κατανοητής γνώσης μέσα από την ανάλυση των δεδομένων, η οποία μέχρι στιγμής δεν υπήρχε. Ο όρος αναφέρεται σε ολόκληρη τη διαδικασία που πρέπει να ακολουθήσει κάποιος, από τη συλλογή δεδομένων μέχρι την αξιοποίηση των αποτελεσμάτων σε πιο πρακτικό επίπεδο. Πρόκειται ουσιαστικά για μια επαναληπτική διαδικασία, αφού είναι πιθανό σε πολλές περιπτώσεις να απαιτηθεί η επιστροφή σε κάποιο προηγούμενο βήμα, προκειμένου να βελτιστοποιηθεί το αποτέλεσμα.

Τα βασικά στάδια της "ΑΓΒΔ" είναι:

1. **Συλλογή Δεδομένων (Data Collection):** Σε αυτό το στάδιο λαμβάνεται ολόκληρο το δείγμα των δεδομένων προκειμένου να αποκτήσουμε μία βασική ιδέα για την σημασία και τη χρησιμότητά τους, ενώ τίθενται οι στόχοι και οι προσδοκίες που θα προκύψουν από την ανάλυσή τους. Ουσιαστικά αφορά τεχνικές αυτόματης και μη αυτόματης συλλογής δεδομένων.
2. **Προεπεξεργασία Δεδομένων (Preprocessing):** Αποτελεί εξαιρετικά σημαντικό στάδιο για τη σωστή εξόρυξη δεδομένων καθώς αφορά την προετοιμασία του συνόλου που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου (training set). Σε αυτό το στάδιο τα δεδομένα καθαρίζονται από τον θόρυβο, ενώ αντιμετωπίζονται προβλήματα ελλειπουσών τιμών, τυπογραφικά λάθη κτλ.
3. **Μετασχηματισμός Δεδομένων (Transformation):** Αφορά τεχνικές οι οποίες καθορίζουν τη βέλτιστη δομή και μέγεθος του συνόλου των χαρακτηριστικών. Το σύνολο εκπαίδευσης πρέπει να είναι αρκετά μεγάλο έτσι ώστε να συμπεριλαμβάνει όλους τους λανθάνοντες συσχετισμούς αλλά ταυτόχρονα αρκετά μικρό έτσι ώστε να εξάγει πληροφορίες σε ένα λογικό χρονικό πλαίσιο. Με άλλα λόγια πρέπει να είναι αντιπροσωπευτικό του συνόλου των δεδομένων.
4. **Εξόρυξη Δεδομένων (Data Mining):** Αποτελεί το βασικό στάδιο επεξεργασίας και αφορά την εφαρμογή πληθώρας διαφορετικών αλγοριθμικών προσεγγίσεων για την εξαγωγή συμπερασμάτων από δεδομένα, όπως τη χρήση τεχνικών παλινδρόμησης, ταξινόμησης, συσταδοποίησης, κ.α. Οι αλγόριθμοι χρησιμοποιούνται για την δημιουργία ενός προγνωστικού μοντέλου με τη βοήθεια ενός συνόλου εκπαίδευσης

(training set). Συνήθως, η ακρίβεια πρόγνωσης του μοντέλου θα αξιολογηθεί με την βοήθεια ενός συνόλου επικύρωσης (validation set), το οποίο αποτελείται από δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση.

5. **Διερμηνεία και Αξιολόγηση (Interpretation/Evaluation):** Σε αυτό το στάδιο τα αποτελέσματα επαληθεύονται, συνήθως με τη χρήση συνόλου επικύρωσης (validation set) ώστε να επιβεβαιωθεί η ορθή εκπαίδευση του μοντέλου και στη συνέχεια παρουσιάζονται αναλυτικά με χρήση τεχνικών απεικόνισης.

Στη συνέχεια γίνεται αναλυτικότερη περιγραφή των επιμέρους σταδίων.

1.4.1 Συλλογή Δεδομένων

Το πρώτο βήμα της "ΑΓΒΔ" είναι η συλλογή και η αποθήκευση των δεδομένων. Η λανθασμένη καταγραφή των δεδομένων που θα χρησιμοποιηθούν, ή η λανθασμένη εισαγωγή αυτών στις κατάλληλες ηλεκτρονικές βάσεις δεδομένων και η άγνοια της διαχείρισης των βάσεων δεδομένων μειώνουν σημαντικά την αξιοπιστία της ανάλυσης τους, οδηγώντας σε μη έγκυρα αποτελέσματα.

Η συλλογή των δεδομένων συνήθως γίνεται είτε αυτόματα, π.χ. με χρήση αισθητήρων, είτε μη αυτόματα, π.χ. με χρήση ερωτηματολογίων. Όσον αφορά την πρώτη κατηγορία, στην καθημερινότητα χρησιμοποιούνται πολλοί αισθητήρες για τη μέτρηση φυσικών δεδομένων και τη μετατροπή τους σε αναγνώσιμα ψηφιακά σήματα. Αυτά τα δεδομένα ύστερα μεταφέρονται σε ένα σημείο συλλογής δεδομένων μέσω ενσύρματων ή ασύρματων δικτύων για περαιτέρω επεξεργασία και αποθήκευση. Άλλη μία μέθοδος είναι η απόκτησή τους μέσω του διαδικτύου. Προς το παρόν, το διαδίκτυο στην απόκτηση δεδομένων αξιοποιείται μέσω του web crawling. Ένας web crawler, στη ουσία, προβαίνει στην καταγραφή των δεδομένων περιήγησης του χρήστη, δηλαδή στις διευθύνσεις που αυτός επισκέπτεται και στο περιεχόμενό τους, ενώ η διαδικασία επαναλαμβάνεται μέχρι να διακοπεί η χρήση του διαδικτύου. Όσον αφορά την δεύτερη κατηγορία, ένα ερωτηματολόγιο μπορεί να υλοποιηθεί είτε μέσω ταχυδρομείου, είτε μέσω τηλεφώνου, είτε με προσωπική συνέντευξη, είτε μέσω διαδικτύου, είτε με άμεση

παράδοση και παραλαβή.

Δυσλειτουργία στους αισθητήρες ή αδυναμία απάντησης κάποιας ερώτησης στα ερωτηματολόγια μπορεί να οδηγήσει σε θορυβώδη ή ελλιπή δεδομένα. Τα συγκεκριμένα προβλήματα, που ενδεχομένως να προκύψουν κατά τη συλλογή δεδομένων, αναλαμβάνει να τα αντιμετωπίσει το επόμενο στάδιο.

1.4.2 Προεπεξεργασία Δεδομένων

Το δεύτερο και πιο σημαντικό στάδιο της "ΑΓΒΔ" είναι η προεπεξεργασία του συνόλου δεδομένων, δηλαδή το σύνολο των εργασιών προετοιμασίας τους, οι οποίες εκτελούνται πριν την καθαυτό εξόρυξη γνώσης. Οι εργασίες αυτές γίνονται με στόχο τον καθαρισμό των δεδομένων, δηλαδή την τακτοποίηση δεδομένων που έχουν αλληλοσυγκρουόμενες πληροφορίες, ασυνέπειες ως προς την κωδικοποίηση, την ονοματοδοσία πεδίων και τις μονάδες μέτρησης, καθώς και χαμένες τιμές και θόρυβο, τυχαία δηλαδή κυμαινόμενα δεδομένα χωρίς ουσιαστικό περιεχόμενο.

Η προεπεξεργασία των δεδομένων περιλαμβάνει τον καθαρισμό τους, αλλά δεν περιορίζεται μόνο σε αυτόν. Ειδικές απαιτήσεις των διαφόρων μεθόδων επεξεργασίας συχνά επιβάλλουν και την περαιτέρω μετατροπή των δεδομένων πριν αυτά χρησιμοποιηθούν. Ορισμένες μέθοδοι δεν μπορούν να χειριστούν συνεχείς αριθμητικές τιμές, αλλά χρειάζονται διακριτές τιμές. Άλλες μέθοδοι που μπορούν να χειριστούν συνεχόμενες αριθμητικές τιμές, αντιμετωπίζουν προβλήματα εάν ορισμένες μεταβλητές περιέχουν πολύ μεγάλες τιμές, ενώ άλλες μεταβλητές περιέχουν πολύ μικρές τιμές.

Έτσι, οι δύο πιο συνηθισμένες εργασίες μετασχηματισμού δεδομένων είναι η διακριτοποίηση και η κανονικοποίηση τους. Ο όρος διακριτοποίηση αναφέρεται στον μετασχηματισμό πεδίων με συνεχείς τιμές σε πεδία με διακριτές τιμές, ενώ ο όρος κανονικοποίηση αναφέρεται στη μετατροπή ενός πεδίου με αριθμητικές τιμές σε νέες αριθμητικές τιμές, οι οποίες είναι πιο "κατάλληλες" διότι έχουν μικρότερη σχετική απόσταση μεταξύ τους.

Αξίζει να σημειωθεί ότι η προεπεξεργασία των δεδομένων δεν είναι μικρό κομμάτι της

“ΑΓΒΔ” και δεν πρέπει να αντιμετωπίζεται ως κάτι προαιρετικό. Μπορεί να απαιτήσει έως και το 60% της συνολικής προσπάθειας και αυτό διότι, αν τα δεδομένα δεν είναι “καθαρά” και στην κατάλληλη μορφή, δεν έχει νόημα να μιλάμε για ποιότητα αποτελεσμάτων.

1.4.3 Μετασχηματισμός Δεδομένων

Ο μετασχηματισμός των δεδομένων αποτελεί το τρίτο στάδιο της “ΑΓΒΔ”. Ουσιαστικά, πρόκειται για τη μετατροπή των δεδομένων κάτω από ένα κοινό πλαίσιο, για επεξεργασία. Χρησιμοποιείται κυρίως για την μείωση του όγκου των χαρακτηριστικών ή για τη δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα.

Το κύριο θέμα που εμπίπτει στον μετασχηματισμό των δεδομένων είναι η μείωση του όγκου τους, καθώς δεδομένα μεγάλου όγκου αυξάνουν την πολυπλοκότητα του προβλήματος και μπορούν να προκαλέσουν προβλήματα στις μεθόδους επεξεργασίας και μεγάλες καθυστερήσεις στη διεξαγωγή των αναλύσεων. Αξίζει να σημειωθεί ότι η αύξηση της πολυπλοκότητας και του συνακόλουθου υπολογιστικού κόστους δεν επιφέρει πάντοτε βελτίωση των αποτελεσμάτων και μείωση του λάθους. Από ένα σημείο και μετά η αύξηση της πολυπλοκότητας έχει μηδενική επίπτωση στη μείωση του σφάλματος. Σε κάθε περίπτωση, η μείωση του όγκου δεν είναι μια τετριμμένη εργασία, καθώς τα αποτελέσματα της ανάλυσης των μειωμένων δεδομένων πρέπει να είναι τα ίδια ή περίπου τα ίδια με τα αποτελέσματα της ανάλυσης του συνόλου των δεδομένων.

Μια ειδική περίπτωση μείωσης του όγκου είναι η επιλογή σημαντικών χαρακτηριστικών (feature selection). Τα διαθέσιμα δεδομένα περιλαμβάνουν πολλά χαρακτηριστικά (αλλιώς αναφέρονται ως διαστάσεις ή γνωρίσματα). Ωστόσο, πολλές φορές για μια συγκεκριμένη εργασία εξόρυξης, δεν είναι χρήσιμα όλα τα διαθέσιμα χαρακτηριστικά, ή ακόμα κάποια χαρακτηριστικά ενώ είναι χρήσιμα μπορεί να σχετίζονται μεταξύ τους, οπότε δεν χρειάζονται όλα ταυτόχρονα. Σε τέτοιες περιπτώσεις, με στόχο την επίτευξη καλύτερων αποτελεσμάτων χρησιμοποιούνται διάφορες τεχνικές για την επιλογή εκείνου του υποσυνόλου των χαρακτηριστικών, το οποίο είναι το πλέον κατάλληλο για τη

συγκεκριμένη εργασία εξόρυξης γνώσης

1.4.4. Εξόρυξη Δεδομένων

Σε αυτό το στάδιο της "ΑΓΒΔ" εφαρμόζεται κάποιος αλγόριθμος για την παραγωγή ενός μοντέλου. Έχοντας πλέον καθαρίσει τα δεδομένα από πιθανούς θορύβους, μετατρέψει σε κατάλληλες μορφές και επιλέξει το βέλτιστο υποσύνολο, είναι έτοιμα να χρησιμοποιηθούν από κάποιον αλγόριθμο, ώστε να δημιουργηθεί κάποιο μοντέλο, συνήθως κατηγοριοποίησης ή πρόβλεψης. Σκοπός είναι να χρησιμοποιήσουμε το μοντέλο αυτό, το οποίο δημιουργήθηκε με βάση κάποια γνωστά δεδομένα, έτσι ώστε να μπορεί να μας δώσει απάντηση για την τιμή ενός χαρακτηριστικού-μεταβλητής στόχου για νέα, άγνωστα δεδομένα.

Το στάδιο αυτό περιλαμβάνει την δοκιμή διαφόρων μοντέλων και την επιλογή του πιο κατάλληλου, με κριτήριο την απόδοση των προβλέψεών του. Ίσως ακούγεται σαν μια απλή διαδικασία, αλλά συνήθως πρόκειται για μια πολύπλοκη και δύσκολη διαδικασία. Έχει αναπτυχθεί πλήθος τεχνικών για να επιτύχουν αυτό τον σκοπό, με τις περισσότερες από αυτές να βασίζονται στην αξιολόγηση ανταγωνισμού των μοντέλων, δηλαδή στην εφαρμογή διαφορετικών μοντέλων στο ίδιο σύνολο δεδομένων και στη συνέχεια στην σύγκριση των αποδόσεων τους για την επιλογή του πιο κατάλληλου.

Ύστερα από την επιλογή του βέλτιστου μοντέλου, γίνεται χρήση αυτού και εφαρμογή σε νέα δεδομένα ώστε να παράγει προβλέψεις για αυτά ή να εκτιμήσει το αποτέλεσμά τους. Τα πιο συνηθισμένα εργαλεία εξόρυξης είναι τα:

- ✔ Τεχνητά Νευρωνικά Δίκτυα – Artificial Neural Networks
- ✔ Δέντρα Αποφάσεων – Decision Trees
- ✔ Επαγωγή Κανόνων – Rule Induction
- ✔ Γενετικοί Αλγόριθμοι – Genetic Algorithms
- ✔ Μέθοδος Γειτνίασης – Nearest Neighbor Method
- ✔ Μέθοδοι Πυρήνα – Kernel Methods

1.4.5 Διερμηνεία και Αξιολόγηση

Στο τελευταίο στάδιο της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων γίνεται η διερμηνεία και η αξιολόγηση των αποτελεσμάτων που παρήχθησαν από την όλη διαδικασία. Η φύση των σύνθετων δεδομένων έφερε την ανάγκη για τρόπους απεικόνισής τους, ώστε να γίνουν καλύτερα αντιληπτές οι σχέσεις και οι δομές που κρύβονται μέσα στα αυτά. Η οπτικοποίηση των δεδομένων (Data Visualization) αποτελεί ουσιαστικά τη διαδικασία μετατροπής των αριθμητικών δεδομένων σε εικόνα με σκοπό την πιο εύκολη διερμηνεία τους.

Έχουν προταθεί κατά καιρούς διάφοροι τρόποι ταξινόμησης των μεθόδων οπτικοποίησης πολυδιάστατων δεδομένων, με επικρατέστερη την ταξινόμηση στις παρακάτω πέντε κατηγορίες (Keim και Kriegel, 1996):

1. Γεωμετρικές Τεχνικές – Geometric Techniques
2. Τεχνικές Εικονογραφημάτων ή Εικονογραφικές Τεχνικές - Icon-Based Techniques
3. Ιεραρχικές Τεχνικές – Hierarchical Techniques
4. Τεχνικές Εικονοστοιχείων – Pixel-Oriented Techniques
5. Τεχνικές Γραφημάτων – Graph- Based Techniques

Η κατηγοριοποίηση αυτή βασίζεται στον τρόπο με τον οποίο παρουσιάζονται τα δεδομένα. Η επιλογή της κατάλληλης μεθόδου έχει να κάνει με την φύση των δεδομένων, τα ερωτήματα που καλούμαστε να απαντήσουμε κατά τη μελέτη τους αλλά και με τα διαθέσιμα εργαλεία που έχουμε στα χέρια μας κάθε φορά.

1.5 Κατηγοριοποίηση μεθόδων Μηχανικής Μάθησης

Εστιάζοντας στο στάδιο της εξόρυξης γνώσης από τα δεδομένα, αυτό υλοποιείται κυρίως με χρήση τεχνικών Μηχανικής Μάθησης. Η Μηχανική Μάθηση αποτελεί ραγδαία αναπτυσσόμενο κλάδο που χρησιμοποιείται κατά κόρον για την εξαγωγή πολύτιμων

συμπερασμάτων από τα δεδομένα. Εν γένει, συνίσταται σε τρεις τρόπους μάθησης, ανάλογους με τους τρόπους με τους οποίους μαθαίνει ο άνθρωπος, και συγκεκριμένα στην επιβλεπόμενη μάθηση (supervised learning), στην μη επιβλεπόμενη μάθηση (unsupervised learning) και στην ενισχυτική μάθηση (reinforcement learning) (Κύρκος, 2015):

Η επιβλεπόμενη μάθηση αφορά την κατασκευή μιας συνάρτησης-μοντέλου που αντιστοιχεί δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα ταξινόμησης, πρόγνωσης και αξιολόγησης. Η λειτουργία των αλγορίθμων αυτών απαιτεί την ύπαρξη ενός συνόλου εκπαίδευσης με ετικέτες (labeled dataset), το οποίο περιλαμβάνει ζεύγη εισόδου-εξόδου που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου.

Η μη επιβλεπόμενη μάθηση στοχεύει στην κατασκευή ενός μοντέλου για κάποιο σύνολο παρατηρήσεων-εισόδων, χωρίς ωστόσο να γνωρίζει τις εξόδους τους. Το σύνολο εκπαίδευσης στην περίπτωση αυτή λοιπόν, δεν περιλαμβάνει ετικέτες-εξόδους παρά μόνον παρατηρήσεις-εισόδους, από τις οποίες το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι. Χρησιμοποιείται σε προβλήματα ανάλυσης συσχετίσεων, ομαδοποίησης - συσταδοποίησης, ανίχνευσης ανωμαλιών κλπ.

Η ενισχυτική μάθηση είναι εμπνευσμένη από τα αντίστοιχα ανάλογα της μάθησης με επιβράβευση και τιμωρία που συναντώνται ως μοντέλα μάθησης των έμβιων όντων. Σκοπός του συστήματος μάθησης είναι να μεγιστοποιήσει μια συνάρτηση του αριθμητικού σήματος ενίσχυσης (ανταμοιβή), για παράδειγμα την αναμενόμενη τιμή του σήματος ενίσχυσης στο επόμενο βήμα. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού, όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

1.5.1 Μέθοδοι επιβλεπόμενης μάθησης

Η εκμάθηση με επίβλεψη είναι μια λειτουργία των μηχανών εκμάθησης στην περίπτωση της οποίας κάθε δείγμα μας αποτελείται από ένα ζευγάρι που περιέχει την επεξηγηματική μεταβλητή x και μια "ετικέτα" (label), τιμής απόκρισης, την μεταβλητή y. Ο αλγόριθμος εκμάθησης με επίβλεψη αναλύει ένα γνωστό σύνολο δεδομένων που καλείται σύνολο εκπαίδευσης και παράγει μια συνάρτηση f η οποία μπορεί να χρησιμοποιηθεί για να χαρτογραφηθούν (ταξινομηθούν) νέα άγνωστα σύνολα δεδομένα. Με άλλα λόγια η f καθορίζει την κλάση στην οποία θα αντιστοιχιστούν τα νέα άγνωστα δεδομένων (Mitchell, 1997). Για να γίνει πιο κατανοητή η διαδικασία παραθέτονται κάποιοι ορισμοί.

Πρόβλημα Ταξινόμησης στον τομέα στης μηχανικής μάθησης είναι το πρόβλημα κατά το οποίο ζητείται ο αυτοματοποιημένος προσδιορισμός της κλάσης (κατηγορίας) που ανήκει μια νέα παρατήρηση π.χ ένα e-mail είναι spam ή όχι.

Σύνολο εκπαίδευσης είναι το σύνολο το οποίο αποτελείται από γνωστά ζευγάρια δεδομένων (x_{ij}, y_i), όπου τα $x_{ij}=(x_{i1}, x_{i2}, \ldots x_{in})$ είναι οι επεξηγηματικές μεταβλητές (τα χαρακτηριστικά των δεδομένων) και τα y_i είναι οι κλάσεις στις οποίες ανήκουν τα αντίστοιχα x_{ij} (η κατηγορία που αντιστοιχεί στο κάθε x_{ij}). Το σύνολο εκπαίδευσης δίνεται σαν είσοδος στον αλγόριθμο εκμάθησης έτσι ώστε ο αλγόριθμος να εκπαιδευτεί πάνω σε αυτό και να εξάγει μοτίβα ή κανόνες αντιστοίχισης των x_{ij} με τα y_i.

Σύνολο επικύρωσης είναι ένα σύνολο που αποτελείται και πάλι από γνωστά ζευγάρια δεδομένων (x_{ij}, y_i). Το σύνολο αυτό χρησιμοποιείται αφού γίνει η εκπαίδευση του αλγορίθμου για να γίνει έλεγχος της απόδοσης του.

Τα βήματα που ακολουθούνται κατά την εκμάθηση με επίβλεψη είναι τα ακόλουθα:

1. Καθορισμός του είδους των δειγμάτων που είναι διαθέσιμα.
2. Διαχωρισμός του συνόλου εκπαίδευσης και του συνόλου επικύρωσης
3. Επιλογή του βασικού αλγορίθμου μηχανικής μάθησης (π.χ. decision trees)
4. Εκπαίδευση του αλγορίθμου πάνω στο σύνολο εκπαίδευσης
5. Έλεγχος της ακρίβειας του μοντέλου πάνω στο σύνολο επικύρωσης

Ο αλγόριθμος εκμάθησης με επίβλεψη λειτουργεί ως εξής:

i. Τροφοδότηση του αλγορίθμου με ένα σύνολο εκπαίδευσης S που περιέχει m ζευγάρια παραδειγμάτων: $S=\left\{\left(x_{1j},y_1\right),\left(x_{2j},y_2\right),\ldots\ldots,\left(x_{mj},y_m\right)\right\}$.

ii. Εύρεση μιας προσεγγιστικής συνάρτησης $h:X\rightarrow Y$ από τον αλγόριθμο, όπου X είναι ο χώρος εισόδου και Y ο χώρος εξόδου. Η h καλείται συνάρτηση πρόβλεψης.

Οι μέθοδοι για εκμάθηση με επίβλεψη μπορούν να αντιμετωπίσουν προβλήματα ταξινόμησης και παλινδρόμησης και συνήθως αυτές οι μέθοδοι είναι γρήγορες και ακριβείς.

ΣΧΗΜΑ 1.3 Διαδικασία εκμάθησης με επίβλεψη

Παραδείγματα αλγορίθμων εκμάθησης με επίβλεψη είναι τα δέντρα αποφάσεων (decision trees προαιρετικά με χρήση τεχνικών bagging ή boosting), τυχαία δάση (random forest), μέθοδος k πλησιέστερων γειτόνων (k nearest neighbor - kNN), γραμμική παλινδρόμηση, λογιστική παλινδρόμηση, "Μηχανές Διανυσμάτων Υποστήριξης" (Support Vector Machines – SVM).

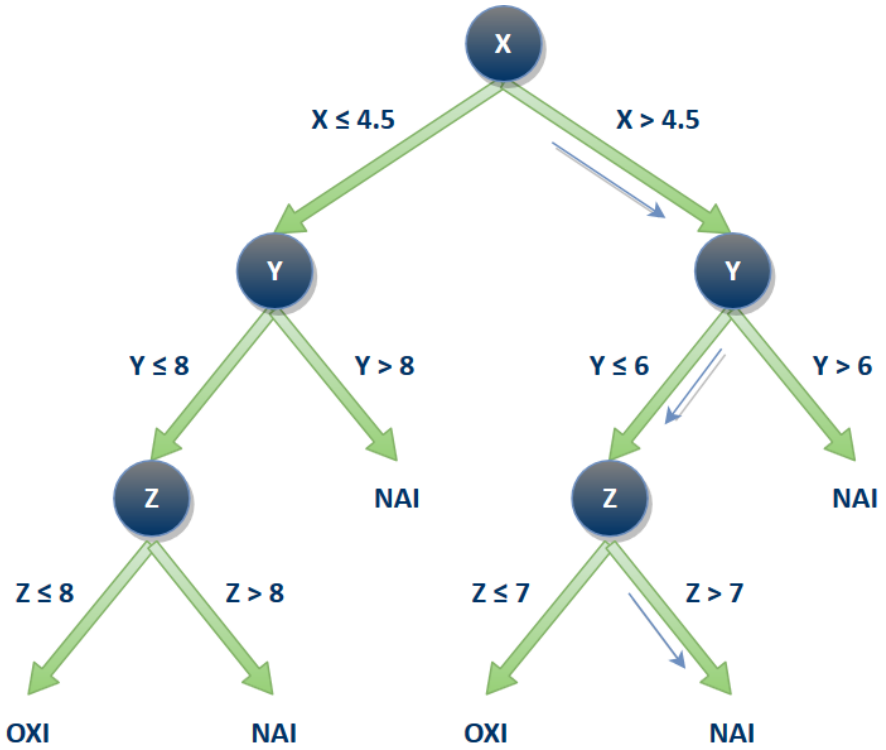
Δέντρα Απόφασης

Τα δέντρα απόφασης είναι συναρτήσεις ταξινόμησης οι οποίες αναπαρίστανται ως δέντρα και είναι μια από τις πιο βασικές μεθόδους κατηγοριοποίησης και πρόβλεψης (Quinlan, 1986). Τα δέντρα απόφασης είναι δέντρα με τις ακόλουθες ιδιότητες:

- Οι κόμβοι του δέντρου αναπαριστούν τα χαρακτηριστικά και ο κόμβος που βρίσκεται στο υψηλότερο επίπεδο ονομάζεται ρίζα του δέντρου.

- Οι κόμβοι του δέντρου συνδέονται μεταξύ τους με κλαδιά τα οποία αναπαριστούν τις δυνατές συνδέσεις του κόμβου “πατέρα” με τους κόμβους “παιδιά”.
- Οι κόμβοι του δέντρου από τους οποίους δεν ξεκινά κάποιο κλαδί λέγονται φύλλα και παίρνουν το όνομα κάποιας κλάσης.

Μια νέα παρατήρηση κατηγοριοποιείται ξεκινώντας από την ρίζα του δέντρου, η οποία αναπαριστά ένα από τα χαρακτηριστικά, αφού ελέγξουμε ποια τιμή έχει το αντίστοιχο χαρακτηριστικό της νέας παρατήρησης προχωράμε στο κατάλληλο κλαδί. Συνεχίζουμε επαναληπτικά την ίδια διαδικασία για κάθε επόμενο κόμβο στον οποίο καταλήγει το κάθε κλαδί που ακολουθήσαμε. Η παραπάνω διαδικασία σταματάει όταν φτάσουμε σε κάποιο φύλλο, το οποίο μας υποδεικνύει και την κλάση του παραδείγματος.



ΣΧΗΜΑ 1.4 Σχηματική αναπαράσταση δέντρου απόφασης

Για την κατασκευή ενός δέντρου απόφασης ακολουθούνται τα παρακάτω βήματα:

1. Δοθέντος ενός συνόλου εκπαίδευσης, αρχίζουμε με έναν κόμβο ο οποίος περιέχει όλα τα παραδείγματα του συνόλου εκπαίδευσης.
2. Ακολουθεί η διάσπαση του κόμβου με βάση μια συνθήκη σε κάποιο από τα

χαρακτηριστικά.

3. Το βήμα (2) εκτελείται επαναληπτικά για κάθε κόμβο μέχρις ότου οι εγγραφές ενός τελικού κόμβου να ανήκουν σε μία μόνο κλάση.
4. Αφού κατασκευαστεί τα δέντρο μπορούν να χρησιμοποιηθούν αν χρειαστεί κάποιες τεχνικές βελτιστοποίησης.

1.5.2 Μέθοδοι μη επιβλεπόμενης μάθησης

Στις μηχανές εκμάθησης η εκμάθηση χωρίς επίβλεψη αναφέρεται στα προβλήματα όπου τα δεδομένα είναι μη κατηγοριοποιημένα, δηλαδή δεν έχουν “ετικέτα” ή τιμή απόκρισης Y (unlabeled data). Έτσι η εκμάθηση χωρίς επίβλεψη προσπαθεί να βρει κρυφά μοτίβα ή τυχαίες συσχετίσεις μεταξύ των δεδομένων. Αυτή η περίπτωση δεν είναι καθόλου σπάνια και τότε δίνεται σαν είσοδος στον αλγόριθμο μόνο οι μεταβλητές $\{x_{1j}, x_{2j}, \dots, x_{nj}\}$. Οι είσοδοι αυτοί για παράδειγμα μπορεί να αντιστοιχούν στα pixels μιας φωτογραφίας, στα αντικείμενα ενός καλαθιού αγορών κ.α.. Αυτό έχει ως αποτέλεσμα να μην μπορεί να χρησιμοποιηθεί κάποια συνάρτηση σφάλματος ή ανταμοιβής κατά την εκπαίδευση, με βάση την οποία θα αξιολογηθεί η ενδεχόμενη λύση. Αυτός ακριβώς είναι και ο λόγος που η εκμάθηση χωρίς επίβλεψη συνδέεται άμεσα με μια πολύ διαδεδομένη τεχνική, την εκτίμηση πυκνότητας πιθανότητας (probability density estimation) μιας μη παρατηρήσιμης συνάρτησης βάσει ενός συνόλου στοιχείων παρατήρησης.

Απλά παραδείγματα Μη-Επιβλεπόμενης Μάθησης είναι η Συσταδοποίηση (Clustering), η Μοντελοποίηση Θεμάτων (Topic Modeling) και η Μείωση Διαστάσεων (Dimensionality Reduction).

Συσταδοποίηση K-Means

Η συσταδοποίηση (Clustering) γενικότερα είναι μια τεχνική μη επιβλεπόμενης μάθησης η οποία αποσκοπεί στην ομαδοποίηση αντικειμένων μεταξύ τους με βάση κάποιο δείκτη ομοιότητας έτσι ώστε τα αντικείμενα που παρουσιάζουν τη μεγαλύτερη ομοιότητα να

βρίσκονται στην ίδια ομάδα (συστάδα). Οι αλγόριθμοι συσταδοποίησης χωρίζονται σε δύο χαρακτηριστικές κατηγορίες: αλγόριθμοι διαχωρισμού (partitioning) και συσσωρευτικοί (agglomerative) αλγόριθμοι, οι οποίοι εκπροσωπούνται από το μοντέλο K-Means και την Ιεραρχική Συσταδοποίηση αντίστοιχα.

Η συσταδοποίηση K-Means είναι μια μέθοδος κβάντωσης διανυσμάτων, προερχόμενη από την επεξεργασία σημάτων, και αρκετά διαδεδομένη για ανάλυση συστάδων στην εξόρυξη δεδομένων (Yadav et al., 2013). Στοχεύει στο διαχωρισμό n παρατηρήσεων σε k συστάδες, των οποίων ο αριθμός είναι προκαθορισμένος. Η κύρια ιδέα είναι ο καθορισμός k κέντρων, ενός για κάθε μία εκ των συστάδων. Έτσι κάθε παρατήρηση θεωρείται ότι ανήκει σε εκείνη τη συστάδα με τον πλησιέστερο μέσο.

Αυτά τα κέντρα πρέπει να τοποθετηθούν με όσο το δυνατόν πιο βέλτιστο τρόπο διότι διαφορετικές τοποθεσίες επιφέρουν διαφορετικά αποτελέσματα. Για αυτό το λόγο η καλύτερη επιλογή είναι να τοποθετηθούν όσο μακριά γίνεται μεταξύ τους. Ο υπολογισμός της απόστασης γίνεται με χρήση της Ευκλείδειας απόστασης. Το επόμενο βήμα είναι κάθε σημείο που ανήκει στο σύνολο δεδομένων να αντιστοιχιστεί στο κοντινότερο κέντρο. Όταν δεν εκκρεμεί πλέον κανένα σημείο προς αντιστοίχηση, το πρώτο βήμα του αλγορίθμου το οποίο αποτελεί την πρωταρχική ομαδοποίηση ολοκληρώνεται.

Έπειτα πρέπει να ξαναϋπολογιστούν τα k νέα κέντρα ως βαρύκεντρα των συστάδων που προέκυψαν από το προηγούμενο βήμα. Εφόσον λοιπόν υπολογιστούν αυτά τα k νέα κέντρα των συστάδων, πρέπει να γίνει ένα νέο ταίριασμα μεταξύ των σημείων δεδομένων και των πλησιέστερων νέων κέντρων των συστάδων. Αυτό το βήμα γίνεται επαναληπτικά και η τοποθεσία των κέντρων των k συστάδων συνεχώς μεταβάλλεται μέχρις ότου αυτά να συγκλίνουν στα βέλτιστα σημεία. Δηλαδή μέχρι να σταματήσουν να βελτιώνονται σημαντικά οπότε και να μετακινούνται.

Η αξιολόγηση της βελτιστοποίησης ουσιαστικά δεν είναι τίποτα άλλο από την ελαχιστοποίηση μιας αντικειμενικής συνάρτησης, γνωστής και ως συνάρτησης τετραγωνικού σφάλματος η οποία δίνεται από τη σχέση:

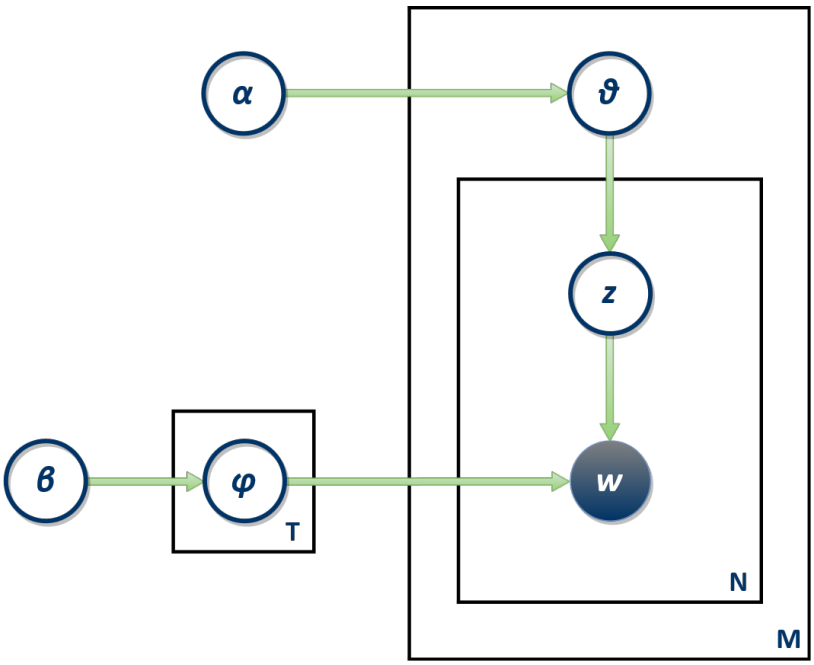
$$J(v)=\sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(\|x_i-v_j\|\right)^2, \qquad 1.1$$

όπου το $\|x_i - v_j\|$ είναι η Ευκλείδεια απόσταση μεταξύ των x_i που συμβολίζουν τα σημεία των δεδομένων της συστάδας και v_j όπου συμβολίζει το κέντρο αυτής, το c_i συμβολίζει τον αριθμό των σημείων δεδομένων στην i-οστή συστάδα και τέλος το c συμβολίζει το πλήθος των κέντρων συστάδων.

Μοντελοποίηση Θεμάτων με Λανθάνουσα Κατανομή Dirichlet (LDA)

Η Λανθάνουσα Κατανομή Dirichlet (Latent Dirichlet Allocation - LDA) είναι ένα παραγωγικό πιθανοτικό μοντέλο για συλλογές διακριτών δεδομένων όπως κείμενα (Blei et al., 2003). Τα δεδομένα μοντελοποιούνται ως μια κατανομή Θεμάτων (Topics), και κάθε Θέμα με τη σειρά του ως κατανομή Λέξεων (Words).

Στο παρακάτω σχήμα διαφαίνεται η γραφική αναπαράσταση του μοντέλου LDA. Οι σκιασμένοι κόμβοι είναι οι παρατηρούμενες μεταβλητές και οι μη-σκιασμένοι είναι οι λανθάνουσες (latent) μεταβλητές. Τα βέλη αναπαριστούν τις εξαρτήσεις και τα ορθογώνια αναπαριστούν τις διαδικασίες επαναλαμβανόμενης δειγματοληψίας. Οι τιμές M, N, και T είναι αντίστοιχα: το πλήθος των Κειμένων στη συλλογή, το πλήθος των Λέξεων σε κάθε κείμενο και το πλήθος των Θεμάτων.



ΣΧΗΜΑ 1.7 Σχηματική αναπαράσταση Λανθάνουσας Κατανομής Dirichlet

Η τιμή z αντιστοιχεί στο Θέμα από το οποίο εξάγεται μια συγκεκριμένη λέξη w. Οι ανά-Κείμενο πολυωνυμικές κατανομές Θεμάτων δίνονται από το θ, ενώ το φ δίνει τις ανά-Θέμα πολυωνυμικές κατανομές λέξεων. Προγενέστερες (prior) κατανομές Dirichlet τοποθετούνται πάνω από αυτές τις κατανομές. Η Dirichlet είναι μια πολυμεταβλητή κατανομή. Αφού ο LDA ακολουθεί ιδέες όπως το ότι κάθε κείμενο μπορεί να αποτελείται από πολλαπλά θέματα και ότι κάθε θέμα μπορεί να αποτελείται από πολλαπλές λέξεις, δημιουργούνται ανάγκες μοντελοποίησης αυτών των συσχετίσεων, και αυτές καλύπτονται από τις κατανομές Dirichlet που είναι στη φύση τους να αντιμετωπίζουν τέτοιες πολλαπλότητες. Αυτές οι Dirichlet κατανομές παραμετροποιούνται από τα α (Alpha) και το β (Beta) αντίστοιχα και δεν φαίνονται από τα δεδομένα, για αυτό και τις αποκαλούμε λανθάνουσες (latent) ή κρυφές (hidden). Όσο πιο χαμηλές οι τιμές των α και β, τόσο πιο λίγα τα θέματα ανά κείμενο και οι λέξεις ανά θέμα αντίστοιχα.

1.5.3 Μέθοδοι ενισχυτικής μάθησης

Ως Ενισχυτική Μάθηση (reinforcement learning) αναφερόμαστε στο πρόβλημα που αντιμετωπίζει ένας πράκτορας ο οποίος διαμορφώνει τη συμπεριφορά του εντός ενός δυναμικού περιβάλλοντος διαμέσου αλληλεπιδράσεων της μορφής προσπάθεια–και–λάθος (trial and error).

Το μοντέλο αλληλεπίδρασης που μελετάται έχει ως εξής: σε κάθε διακριτό σημείο στο χρόνο ο πράκτορας μελετά την τρέχουσα κατάσταση (state) του περιβάλλοντος (environment) και επιλέγει την πραγματοποίηση μιας ενέργειας (action). Σαν αποτέλεσμα, το περιβάλλον μεταβαίνει σε μια νέα κατάσταση και ο πράκτορας λαμβάνει μια τιμή ανταμοιβής (reward). Αυτό το σήμα αποτελεί ένα μέτρο της ποιότητας των ενεργειών του πράκτορα, όπως καθορίζονται από το περιβάλλον.

Οι τεχνικές μάθησης που βασίζονται σε ένα τέτοιο μοντέλο είναι πολύ ελκυστικές. Κι αυτό διότι αν οι πράκτορες υπόκεινται σε ενισχυτική μάθηση, οι σχεδιαστές καλούνται μόνο να παράγουν την τιμή της ανταμοιβής.

Παραδείγματα μεθόδων ενισχυτικής μάθησης αποτελούν ο αλγόριθμος Q-Μάθησης και

η Μάθηση χρονικών διαφορών (Temporal Difference learning – TD learning).

Αλγόριθμος Q-Μάθησης

Στην ενισχυτική μάθηση Q (Vlachogiannis et al.,2004) η βέλτιστη συνάρτηση επιβράβευσης ορίζεται με χρήση της εξίσωσης Bellman, ως εξής:

$$Q^*(s,a) = E\left(r(s,a) + \gamma \cdot \max_{a'} Q^*(s',a')\right). \quad 1.2$$

Η εξίσωση αυτή αναπαριστά το αναμενόμενο άθροισμα των επιβραβεύσεων που λαμβάνεται ξεκινώντας από μια αρχική κατάσταση (s), εκτελώντας τη δράση (a) και επιλέγοντας βέλτιστες δράσεις (a') στις επόμενες αναζητήσεις. Αυτό γίνεται σε περιορισμένο ή θεωρητικά άπειρο χρονικά ορίζοντα μέχρι να επιτευχθεί η βέλτιστη τιμή της συνάρτησης $Q(Q^*(s,a))$. Η εκπτωτική παράμετρος γ $(0 \le \gamma \le 1)$ χρησιμοποιείται για να μειωθεί εκθετικά το βάρος των επιβραβεύσεων που λαμβάνεται κατά τις επόμενες αναζητήσεις. Εφόσον έχουμε τη βέλτιστη τιμή $Q^*(s,a)$ είναι εύκολο να καθοριστεί η βέλτιστη δράση a* με τη χρήση της βέλτιστης πολιτικής. Ένας απλός τρόπος είναι να ερευνηθούν όλες οι δυνατές δράσεις (a) για μία δεδομένη κατάσταση (s) και να επιλεγεί εκείνη με τη μεγαλύτερη τιμή:

$$a^* = \arg\max_a Q^*(s,a). \quad 1.3$$

Η συνάρτηση Q (μνήμη-Q) συνήθως αποθηκεύεται σε κάποιον πίνακα με διαστάσεις τον αριθμό των καταστάσεων και τον αριθμό των δράσεων. Λαμβάνοντας αρχικά αυθαίρετες τιμές μπορεί, μέσω επαναλήψεων, να προσεγγιστεί η βέλτιστη συνάρτηση Q σύμφωνα με τα προκαθορισμένα κριτήρια.

Στη συνέχεια η εγγραφή στον πίνακα που απεικονίζει την απόδοση της δράσης (a) πάνω στην κατάσταση (s) γίνεται σύμφωνα με την παρακάτω επαναληπτική εξίσωση:

$$Q(s,a) = (1-a) \cdot Q(s,a) + a \cdot \left(r + \gamma \cdot \max_{a'} Q(s',a')\right). \quad 1.4$$

Είναι σημαντικό να σημειωθεί ότι η νέα τιμή της μνήμης $Q(s,a)$ βασίζεται τόσο στην τρέχουσα τιμή της $Q(s,a)$, (μνήμη-Q) όσο και στις τιμές (άμεσες επιβραβεύσεις) των δράσεων που λαμβάνονται από τις επόμενες δράσεις. Έτσι, η παράμετρος a $(0 \le a \le 1)$

αναπαριστά το ποσοστό της νέας γνώσης που εναποτίθεται στη μνήμη επηρεάζοντας έτσι τον αριθμό των επαναλήψεων μάθησης. Η παράμετρος $(1-a)$ εκφράζει το συνολικό ποσοστό των τιμών Q που παραμένουν ως "μνήμη" στη συνάρτηση Q.

1.6 Εφαρμογές και Προκλήσεις

Τα παραδείγματα εξόρυξης γνώσης από δεδομένα ποικίλουν ανάλογα με τον τομέα στον οποίο εφαρμόζονται. Στη σημερινή εποχή όπου τα δεδομένα υπάρχουν σχεδόν παντού και τις περισσότερες φορές βρίσκονται σε ηλεκτρονική μορφή, η σωστή ανάλυση τους οδηγεί πάντα στην ανάδειξη και οργάνωση της πληροφορίας, η γνώση της οποίας είναι ο σημαντικότερος παράγοντας για την εύρεση μιας στρατηγικής και την ορθολογική λήψη αποφάσεων. Ο χρηματοοικονομικός τομέας, ο τομέας των τηλεπικοινωνιών, της υγείας και της εκπαίδευσης, ο δημόσιος τομέας, ο τομέας των τηλεπικοινωνιών καθώς επίσης και αυτός της βιομηχανίας και της έρευνας, αποτελούν ίσως το μεγαλύτερο δείγμα εφαρμογών των τεχνολογιών εξόρυξης γνώσης από δεδομένα (Wang, 2005), (Yu και Guo, 2016), (Han et al., 2011).

Επιχειρήσεις. Μια εφαρμογή των μηχανών εκμάθησης είναι η ανάλυση του καλαθιού αγορών των καταναλωτών (basket analysis), για παράδειγμα στα πλαίσια μιας αλυσίδας πολυκαταστημάτων. Σκοπός είναι η εύρεση συσχέτισης μεταξύ των προϊόντων που αγοράζουν οι πελάτες. Για παράδειγμα αν ένας πελάτης που αγοράζει το Χ προϊόν συνήθως αγοράζει και το Υ, ενώ ένας άλλος πελάτης που αγοράζει το Χ προϊόν δεν συνηθίζει να αγοράζει και το Υ προϊόν, τότε ο δεύτερος πελάτης είναι υποψήφιος αγοραστής του Υ προϊόντος. Στόχος είναι η εύρεση αυτών των προϊόντων και η τοποθέτηση τους σε κοντινές αποστάσεις, διαφημίσεις κ.α ώστε να γίνει μεγιστοποίηση των πωλήσεων τους (cross selling). Με την ίδια λογική θα ήταν χρήσιμο αν για κάποιο χρήστη μιας συγκεκριμένης ιστοσελίδας μπορούσε να γίνει πρόβλεψη του συνδέσμου (link) που θα πιθανώς θα ήθελε να επισκεφθεί ώστε να φορτωθούν από πριν τα δεδομένα του συνδέσμου για γρηγορότερη πρόσβαση.

Τραπεζικές Εφαρμογές. Ένα χρηματοπιστωτικό ίδρυμα, για παράδειγμα μια τράπεζα,

προσφέρει δάνεια σε κάποιο πελάτη τα οποία πρέπει να επιστραφούν πίσω με τόκο, συνήθως σε δόσεις. Είναι σημαντικό για την τράπεζα να βρίσκεται σε θέση να προβλέψει εκ των προτέρων τον κίνδυνο που υπάρχει ο πελάτης να μην πληρώσει όλο το ποσό του δανείου. Με αυτήν την εκ των προτέρων γνώση η τράπεζα εξασφαλίζει το κέρδος της. Ο κίνδυνος αυτός υπολογίζεται βάσει του ποσού του δανείου και πληροφοριών σχετικά με τον πελάτη (επάγγελμα, εισόδημα, αποταμιεύσεις, ηλικία, εξασφαλίσεις κ.α) καθώς και από προηγούμενη εμπειρία σχετικά με εξοφλήσεις δανείων γενικότερα. Σκοπός των μηχανών εκμάθησης είναι η κατασκευή ενός ισχυρού κανόνα που μοντελοποιεί τα δεδομένα του παρελθόντος και τα χαρακτηριστικά του πελάτη και είναι σε θέση να προβλέψει τον κίνδυνο για ένα νέο δάνειο.

Αναγνώριση χειρόγραφων κειμένων. Σε πολλές περιπτώσεις δημιουργείται η ανάγκη να αναγνωριστεί τι είναι γραμμένο πάνω σε μια επιταγή ή σε μια επιστολή ή γενικά σε κάποιο χειρόγραφο έγγραφο. Η διαφοροποίηση που δημιουργείται από τον ανθρώπινο παράγοντα στο συγκεκριμένο πρόβλημα είναι πολύ μεγάλη, για παράδειγμα ο τρόπος γραφής ενός τυπικού γράμματος όπως το “Ε” διαφέρει σημαντικά από άτομο σε άτομο. Σκοπός των μηχανών εκμάθησης είναι η κατασκευή ενός κανόνα που θα αναγνωρίζει πιο γράμμα αντιστοιχεί σε κάθε σύμβολο του εκάστοτε γραφικού χαρακτήρα, δηλαδή στο παράδειγμα με το γράμμα “Ε” αναζητούνται όλα αυτά τα χαρακτηριστικά που κάνουν όλα τα “Ε” (από διαφορετικούς τρόπους γραφής) να είναι όμοια, ώστε τελικά να μετατραπεί το χειρόγραφο κείμενο σε ηλεκτρονικό έγγραφο.

Αναγνώριση προσώπου. Μια διαδικασία ταυτοποίησης στοιχείων, είναι πιθανόν να γίνει μέσω αναζήτησης εικόνας προσώπου. Στόχος είναι η εικόνα που εισάγεται σε μια μηχανή εκπαίδευσης να αντιστοιχιστεί σε κάποια από αυτές που υπάρχουν στην βάση δεδομένων, όταν πρόκειται για το ίδιο πρόσωπο. Η μηχανή εκμάθησης εκπαιδεύεται στο να ανιχνεύει συγκεκριμένα χαρακτηριστικά μέσω μοτίβων προσώπου σε εικόνες με σκοπό να εντοπίσει αυτά τα χαρακτηριστικά σε κάθε εικόνα από την βάση δεδομένων και να τα συγκρίνει με την εικόνα που είναι προς διερεύνηση.

Ιατρικές διαγνώσεις. Στην περίπτωση των ιατρικών διαγνώσεων, οι πληροφορίες που δίνονται ως είσοδοι σε έναν αλγόριθμο εκμάθησης είναι τα στοιχεία που είναι διαθέσιμα για κάποιον ασθενή όπως η ηλικία του, το βάρος του, το φύλο του, το ιατρικό του ιστορικό, ίσως οι συνήθειές του και το επάγγελμά του, και τέλος τα συμπτώματα που

παρουσιάζει. Οι μηχανές εκμάθησης αναλαμβάνουν να εντοπίσουν την πιο πιθανή ασθένεια από την οποία μπορεί να πάσχει ο συγκεκριμένος ασθενής, συνδέοντάς τον με μια ασθένεια από την βάση δεδομένων, μοντελοποιώντας τα χαρακτηριστικά του και εξάγοντας διάφορες νόρμες.

Οικονομία. Οι μηχανές εκμάθησης χρησιμοποιούνται και στον κλάδο της οικονομίας για την εύρεση πιθανών ακραίων τιμών σε σύνολα δεδομένων. Πρακτικά βρίσκουν τιμές οι οποίες δεν υπακούν στον κανόνα που κατασκευάστηκε μέσα από την εκμάθηση τους. Οι τιμές αυτές ύστερα θεωρούνται ανωμαλίες που χρήζουν προσοχής. Για παράδειγμα μπορεί να πρόκειται για κάποια απάτη ή παρανομία όπως ξέπλυμα μαύρου χρήματος κ.α.

Μετεωρολογία. Μια πληθώρα συστημάτων έχει υιοθετηθεί, τα οποία έχουν εκπαιδευτεί κατάλληλα ώστε να είναι δυνατή η ανάλυση κλιματολογικών συνθηκών και η εξαγωγή προβλέψεων μέσα από αυτήν σχετικά με κλιματολογικές αλλαγές και φαινόμενα όπως οι κυκλώνες, οι καταιγίδες, οι καύσωνες και άλλα πολλά.

Τηλεπικοινωνίες. Η ανάλυση των στοιχείων λειτουργίας των δικτύων είναι ένα επιπλέον πεδίο εφαρμογής μεθόδων Εξόρυξης Δεδομένων. Τα δεδομένα αυτά παράγονται με αυτόματο τρόπο από διάφορα υποσυστήματα του δικτύου και ο όγκος τους είναι τέτοιος, που καθιστά αδύνατη την επεξεργασία τους χωρίς τη χρήση εξελιγμένων τεχνικών. Μια συγκεκριμένη βλάβη στο δίκτυο μπορεί να προκαλέσει πολλά και διαφορετικά μηνύματα σφάλματος. Με την εφαρμογή τεχνικών κατηγοριοποίησης και ανάλυσης αλληλουχίας, μέσα από μια μηχανή εκμάθησης, μπορεί να γίνει συσχετισμός των μηνυμάτων και εντοπισμός της βλάβης. Τα στοιχεία λειτουργίας του δικτύου μπορούν να χρησιμοποιηθούν και για την ανάλυση της ποιότητας των υπηρεσιών.

Λογιστική – Ελεγκτική. Δύο μεγάλα προβλήματα της Ελεγκτικής, στα οποία βρίσκουν εφαρμογή οι τεχνικές Εξόρυξης Δεδομένων και ειδικότερα οι τεχνικές κατηγοριοποίησης, είναι η πρόβλεψη χρεοκοπίας και ο εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων. Η πρόβλεψη χρεοκοπίας είναι ένα από τα σημαντικότερα προβλήματα λήψης αποφάσεων στον τομέα των επιχειρήσεων. Στόχος των μηχανών εκμάθησης είναι η κατασκευή ενός μοντέλου ικανού να προβλέπει έγκαιρα τις περιπτώσεις χρεοκοπίας. Αυτό βασίζεται στην υπόθεση ότι η χρεοκοπία είναι ένα φαινόμενο που εξελίσσεται στη διάρκεια του χρόνου και όχι ένα στιγμιαίο συμβάν, και έτσι ένας αλγόριθμος μπορεί να

εκπαιδευτεί πάνω στα οικονομικά δεδομένα που αφορούν τις επιχειρήσεις και να χτίσει συσχετίσεις.

Κυβερνοασφάλεια. Μέθοδοι εξόρυξης γνώσης μπορούν να χρησιμοποιηθούν για τη δημιουργία φίλτρων διαχωρισμού της κακόβουλης κυκλοφορίας από την συνήθη και για τον χαρακτηρισμό της δικτυακής κίνησης (Papadopoulos, 2017). Με τη βοήθεια τεχνικών μηχανικής μάθησης καθίσταται δυνατή η διαχείριση, καταγραφή και αποθήκευση δεδομένων για την ανίχνευση ανωμαλιών σε σχεδόν πραγματικό χρόνο. Σχετικά λογισμικά μπορούν να αναλύουν ταυτόχρονα δισεκατομμύρια καταγραφές, με σκοπό τον εντοπισμό άγνωστων και κρυφών απειλών, προσφέροντας νέες δυνατότητες δικτυακής ορατότητας και προβλέποντας τις πιο πιθανές απειλές δικτυακής ασφάλειας.

Όπως αναφέρθηκε σε αυτό το κεφάλαιο, μετά την απόκτηση των δεδομένων, ακολουθεί η χρήση των διαφόρων τεχνικών μηχανικής μάθησης, και συγκεκριμένα η εξαγωγή των χαρακτηριστικών και η ανάλυσή τους με σκοπό τη δημιουργία του επαγωγικού μοντέλου. Ωστόσο, η διαδικασία αυτή συνοδεύεται από τις δικές της ιδιαιτερότητες (Aliferis et al., 2006). Για παράδειγμα, εάν το πλήθος των παρατηρήσεων είναι υπερβολικά μεγάλο, μπορεί να δημιουργηθεί ένα εξαιρετικά πολύπλοκο μοντέλο που να ενσωματώνει εκατοντάδες παράγοντες, ακολουθώντας τη λογική ότι οι περισσότεροι από αυτούς είναι πραγματικά σημαντικοί. Αντιστρόφως, όταν υπάρχει ένας μικρός αριθμός παρατηρήσεων η ανάλυση είναι συνήθως απλούστερη, δίνοντας έμφαση στην αξιόπιστη ανίχνευση μόνο των κύριων επιδράσεων. Οι συνηθέστερες προκλήσεις που συναντώνται κατά την εξόρυξη γνώσης από τα λεγόμενα δεδομένα υψηλής διάστασης (high dimensional data) συνοψίζονται παρακάτω:

1. Κατάρα της διαστατικότητας (curse of dimensionality): Ο όρος αυτός, ο οποίος επινοήθηκε από τον Bellman, αφορά το ελάχιστο απαιτούμενο μέγεθος του δείγματος που απαιτείται για την ορθή εκτίμηση ενός μοντέλου. Συγκεκριμένα, αναφέρει πως ο αριθμός των δειγμάτων που απαιτούνται για την εκτίμηση μιας λειτουργίας ή κατάστασης με ένα δεδομένο επίπεδο ακρίβειας, αυξάνεται εκθετικά με τον αριθμό των μεταβλητών-διαστάσεων που την χαρακτηρίζουν. Αυτό σημαίνει ότι όσο μεγαλύτερη είναι η υποκείμενη διάσταση ενός προβλήματος, τόσο εκθετικά αυξημένο πρέπει να είναι και το πλήθος των δεδομένων με τα οποία πρέπει να τροφοδοτηθεί ο αλγόριθμος μηχανικής μάθησης.

2. Ψευδώς θετικά/αρνητικά αποτελέσματα (false positives/negatives): Υπάρχουν πολλοί παράγοντες που καθορίζουν την ακρίβεια ενός μοντέλου μηχανικής μάθησης, με σημαντικότερους την ύπαρξη αντιπροσωπευτικού συνόλου εκπαίδευσης, την έλλειψη θορύβου, τον χρόνο εκπαίδευσης κ.α. Ωστόσο, σε ένα πολύπλοκο σύνολο δεδομένων, υπάρχει πάντα η περίπτωση λανθασμένων προβλέψεων από τον αλγόριθμο μηχανικής μάθησης, είτε πρόκειται για λάθη ταξινόμησης είτε για λανθασμένες ομαδοποιήσεις παρατηρήσεων στην ίδια συστάδα. Ανάλογα με τη σύμβαση που ακολουθήθηκε κατά τον ορισμό του προβλήματος έχουμε την περίπτωση ψευδώς θετικής πρόβλεψης (πχ. Όταν ένα άτομο είναι υγιές, αλλά ο αλγόριθμος εκτιμά λανθασμένα ότι είναι ασθενής) και την περίπτωση ψευδώς αρνητικής πρόβλεψης (πχ. Όταν ένα άτομο είναι ασθενής, αλλά ο αλγόριθμος εκτιμά λανθασμένα ότι είναι υγιής).

3. Υπερπροσαρμογή (overfitting): Ένα από τα βασικά προβλήματα που μπορεί να εμφανιστεί κατά την εκπαίδευση ενός μοντέλου είναι αυτό της υπερβολικής εκπαίδευσης. Στην περίπτωση αυτή, το συνολικό σφάλμα για το σύνολο εκπαίδευσης γίνεται πολύ μικρό, αλλά παραμένει υπερβολικά για το σύνολο επικύρωσης, με αποτέλεσμα το μοντέλο να μην γενικεύει καλά. Αυτό συμβαίνει συνήθως όταν ένα στατιστικό μοντέλο περιγράφει το τυχαίο σφάλμα ή τον θόρυβο αντί της επικείμενης σχέσης μεταξύ των παρατηρήσεων και οδηγεί σε κακή προγνωστική απόδοση, αφού μπορεί να διογκωθούν οι μικρές διακυμάνσεις που υπάρχουν στα δεδομένα. Σε δεδομένα υψηλής διάστασης, όπου ο αριθμός των παρατηρήσεων είναι μικρός και ο αριθμός των μεταβλητών εισόδου είναι μεγάλος, το φαινόμενο δημιουργίας μοντέλων με χαμηλή δυνατότητα γενίκευσης είναι αρκετά συχνό.

ΚΕΦΑΛΑΙΟ 2

Ταξινόμηση με χρήση Μεθόδων Μηχανικής Μάθησης

2.1 Εισαγωγή

Η χρήση τεχνικών επιβλεπόμενης μάθησης για ταξινόμηση παρατηρήσεων σε κλάσεις αποτελεί πλέον με διαφορά την πιο διαδεδομένη εφαρμογή μεθόδων μηχανικής μάθησης σε τεχνικό, ερευνητικό και βιομηχανικό επίπεδο. Η εκμάθηση ενός μοντέλου από δεδομένα έτσι ώστε αυτό να είναι σε θέση να κατηγοριοποιεί επιτυχώς νέες παρατηρήσεις, αποτελεί μονάχα την προφανή λειτουργία των μεθόδων αυτών, αποτελώντας εξάλλου βασική δυνατότητα και των συστημάτων βασιζόμενων σε κανόνες. Η ιδιαιτερότητα των ταξινομητών επιβλεπόμενης μάθησης ωστόσο, έγκειται στην πρόσθετη δυνατότητά τους να δημιουργούν γενικεύσεις κατά τη διάρκεια της εκπαίδευσής τους, μαθαίνοντας τα σύνορα μεταξύ των ορίων απόφασης, ανεξαρτήτως της πολυπλοκότητας και του πλήθους των μεταβλητών που ανήκουν στο σύνολο δεδομένων (Duda et al., 2000), (Theodoridis και Koutroumbas, 2008).

Η επεκτασιμότητα, η απλότητα στη χρήση καθώς και η ευκολία αυτοματοποιημένης εκπαίδευσής τους, καθιστούν τις μεθόδους επιβλεπόμενης μάθησης ως τις πλέον εύρωστες λύσεις ταξινόμησης για συστήματα που απαιτούν ανθεκτικότητα σε μελλοντικές αλλαγές, σε αντίθεση με τις παραδοσιακές rule-based προσεγγίσεις, όπου κάθε αλλαγή θα πρέπει να προστίθεται χειροκίνητα στο σύνολο κανόνων.

Στο συγκεκριμένο κεφάλαιο περιγράφονται τέσσερις από τις πλέον ευρέως χρησιμοποιούμενες μεθόδους ταξινόμησης, οι οποίες θα υλοποιηθούν προγραμματιστικά στο Κεφάλαιο 5 για την επίλυση ενός πραγματικού προβλήματος

ταξινόμησης. Οι μέθοδοι που θα περιγραφούν είναι: το πολυεπίπεδο δίκτυο Perceptron (Multi Layer Perceptron - MLP), ο αλγόριθμος Random Forest, ο αλγόριθμος k-πλησιέστερων γειτόνων και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM). Στις ακόλουθες ενότητες γίνεται σύντομη περιγραφή του μαθηματικού υποβάθρου και παρουσίαση της αλγοριθμικής διαδικασίας για κάθε μια από αυτές, ενώ παρατίθεται παράλληλα τα βασικά πλεονεκτήματα και μειονεκτήματά τους καθώς και οι ιδιαιτερότητες στην εφαρμογή τους.

2.2 Τεχνητά Νευρωνικά Δίκτυα

2.2.1 Εισαγωγή

Τα τεχνητά νευρωνικά δίκτυα είναι απλοποιημένα μοντέλα του κεντρικού συστήματος του ανθρώπου. Σκοπός τους είναι να μιμούνται τη λειτουργία των βιολογικών νευρώνων του εγκεφάλου και την δομή των βιολογικών νευρωνικών δικτύων. Αποτελούνται από διασυνδεδεμένα υπολογιστικά στοιχεία που έχουν την ικανότητα να ανταποκρίνονται σε ερεθίσματα που δέχονται στη είσοδο και να μαθαίνουν να προσαρμόζονται στο περιβάλλον τους.

Τα συνήθη τεχνητά νευρωνικά δίκτυα χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων τέτοια ώστε να διατηρούν μόνο τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στη νευρολογία. Τα τεχνητά νευρωνικά μοντέλα έχουν δηλαδή ελάχιστη σχέση με τα βιολογικά νευρωνικά συστήματα. Ωστόσο οι λεπτομέρειες δεν έχουν ιδιαίτερη σημασία στην κατανόηση της ευφυούς συμπεριφοράς των βιολογικών νευρωνικών συστημάτων. Ακόμη και αυτά τα απλά μοντέλα μπορούν να δημιουργήσουν ενδιαφέροντα δίκτυα αρκεί να πληρούν δύο βασικά χαρακτηριστικά:

✔ οι νευρώνες να έχουν ρυθμιζόμενες παραμέτρους ώστε να διευκολύνεται η

διαδικασία της μάθησης – ιδιότητα γνωστή ως πλαστικότητα των νευρώνων

- ✔ το δίκτυο να αποτελείται από πολλούς νευρώνες ώστε να επιτυγχάνεται παραλληλισμός της επεξεργασίας και κατανομή της πληροφορίας.

Τα τεχνητά νευρωνικά δίκτυα μοιάζουν με τον εγκέφαλο στα εξής σημεία:

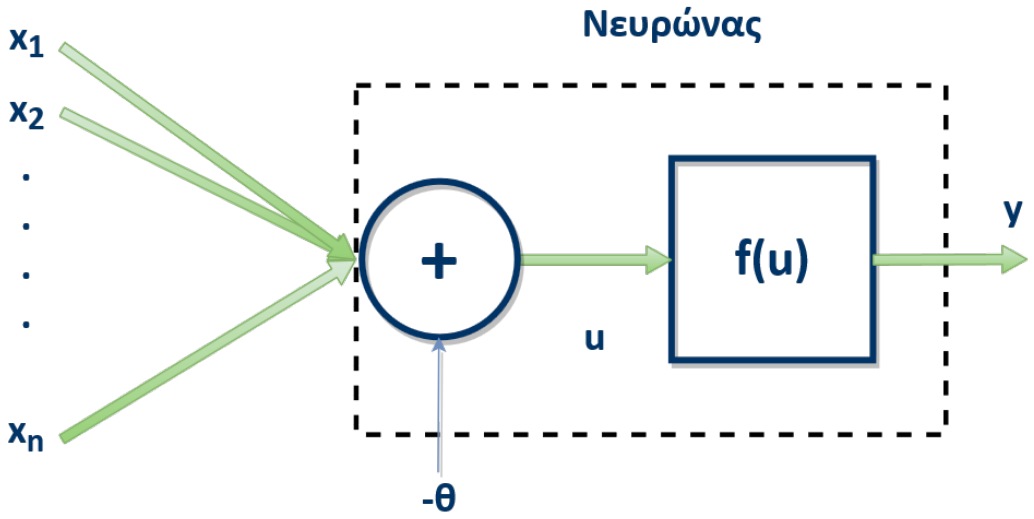
- ➢ η γνώση αποκτάται από το δίκτυο μέσα από μια διαδικασία μάθησης – εκπαίδευσης
- ➢ η γνώση αποθηκεύεται στις δυνάμεις σύνδεσης των νευρώνων, οι οποίες είναι τα συναπτικά (synaptic) βάρη.

Ένα τεχνητό νευρωνικό δίκτυο οφείλει την υπολογιστική του ισχύ πρώτον στην παράλληλη, κατανεμημένη δομή του και δεύτερον στην ικανότητά του να μαθαίνει και να γενικεύει. Η δεύτερη ιδιότητα αναφέρεται στην παραγωγή λογικών εξόδων για εισόδους τις οποίες δεν έχει συναντήσει κατά την διάρκεια της εκπαίδευσής του. Αυτές οι δύο ιδιότητες δίνουν στα νευρωνικά δίκτυα τη δυνατότητα να βρίσκουν καλές προσεγγιστικές λύσεις σε πολύπλοκα προβλήματα, τα οποία είναι μη επιδεκτικά σε λύσεις.

Το αντικείμενο των τεχνητών νευρωνικών δικτύων είναι η ανάπτυξη κατάλληλων αλγορίθμων εκπαίδευσης και ανάκλησης της πληροφορίας που αυτά περιέχουν έτσι ώστε να προσομοιάζονται ευφυείς διαδικασίες. Για να γίνει αυτό, πρέπει να οριστεί το κατάλληλο περιβάλλον εκπαίδευσης, π.χ. αν το δίκτυο εκπαιδεύεται με επίβλεψη ή χωρίς. Η θεμελιώδης μονάδα επεξεργασίας της πληροφορίας στα τεχνητά νευρωνικά δίκτυα είναι ο νευρώνας (neuron) ή κόμβος (node), ο οποίος αποτελείται από εισόδους και εξόδους. Υλοποιεί τοπικά έναν υπολογισμό με βάση τις εισόδους που δέχεται και μεταδίδει το αποτέλεσμα σε άλλες μονάδες επεξεργασίας με τις οποίες συνδέεται. Το σύστημα λειτουργεί παράλληλα και πολλές μονάδες έχουν δυνατότητα να πραγματοποιούν ταυτόχρονα τους υπολογισμούς τους. Κάθε σύνδεση μεταξύ μονάδων χαρακτηρίζεται από μια τιμή βάρους. Οι τιμές των βαρών των συνδέσμων αποτελούν τη γνώση που είναι αποθηκευμένη στο τεχνητό νευρωνικό δίκτυο και χαρακτηρίζουν τη λειτουργία του. Συνήθως ένα τεχνητό νευρωνικό δίκτυο αναπτύσσει μια συλλογική λειτουργικότητα μέσω μιας μορφής εκπαίδευσης, Στο δίκτυο υπάρχουν και οι κρυφές μονάδες, οι οποίες δεν είναι ορατές στον εξωτερικό κόσμο και οι είσοδοι τους καθώς και οι έξοδοι τους βρίσκονται εντός του δικτύου.

2.2.2 Το μοντέλο του Perceptron

Το μοντέλο του νευρώνα Perceptron αποτελεί το βασικότερο είδος τεχνητού νευρωνικού δικτύου που εφευρέθηκε το 1957 στο Αεροναυτικό Εργαστήριο του Κορνέλλ (Cornell Aeronautical Laboratory) από τον Φρανκ Ρόζενμπλαττ (F. Rosenblatt, 1961). Μπορεί να χαρακτηριστεί ως ένα απλό είδος ενός εμπροσθοτροφοδοτούμενου (feed-forward) νευρωνικού δικτύου και αποτελεί την βάση για πιο περίπλοκα δίκτυα που αναπτύχθηκαν αργότερα. Ο νευρώνας Perceptron χαρακτηρίζεται από ορισμένο αριθμό εισόδων και από μία μοναδική έξοδο, όπως φαίνεται στο σχήμα (Σχήμα 2.1).



ΣΧΗΜΑ 2.1 Perceptron

Στο μοντέλο του Perceptron, οι μοναδικές συνδέσεις που υπάρχουν είναι αυτές μεταξύ των εισόδων $x_1,\dots,x_n$, και του νευρώνα. Κάθε σύνδεση του νευρώνα με το εισερχόμενο σήμα $x_1,\dots,x_n$, χαρακτηρίζεται από το αντίστοιχο συναπτικό βάρος w_i, που αναπαριστά την επίδραση του σήματος στον νευρώνα. Επομένως, η ποσότητα από την οποία εξαρτάται η ενεργοποίηση του νευρώνα είναι το γινόμενο $x_i \cdot w_i$ και όχι η τιμή του βάρους w ή του σήματος x ξεχωριστά. Ο νευρώνας ακολουθεί το μοντέλο McCulloch-Pitts και περιγράφεται από τις ακόλουθες εξισώσεις:

$$u = \sum_{i=1}^{n} w_i \cdot x_i - \theta. \qquad 2.1$$

Η έξοδος του νευρώνα είναι η $y=f(u)$, η οποία γράφεται εναλλακτικά ως εξής:

$$y=f\left(\sum_{i=1}^{n} w_i \cdot x_i - \theta\right), \qquad 2.2$$

όπου u είναι η διέγερση του νευρώνα, f η συνάρτηση ενεργοποίησης που απεικονίζει το διάνυσμα εισόδου $x=[x_1,\dots,x_n]^T$ στην έξοδο y και τέλος θ το κατώφλι ενεργοποίησης πάνω από το οποίο ενεργοποιείται ο νευρώνας.

Η διέγερση του νευρώνα είναι θετική εάν το άθροισμα $\sum_{i=1}^{n} w_i \cdot x_i$ ξεπεράσει αυτό το όριο. Έτσι διακρίνουμε τις παρακάτω τρεις περιπτώσεις:

- $u>0$, αν $\sum_{i=1}^{n} w_i \cdot x_i > \theta$
- $u=0$, αν $\sum_{i=1}^{n} w_i \cdot x_i = \theta$
- $u<0$, αν $\sum_{i=1}^{n} w_i \cdot x_i < \theta$.

Η συνάρτηση ενεργοποίησης f, αποτελεί μη-γραμμική συνάρτηση, η οποία τροφοδοτείται από τη διέγερση u και δίνει την έξοδο του νευρώνα σε μία από τις ακόλουθες μορφές:

1. $f(u)=\begin{cases} 1, & \text{αν } u>0 \\ 0, & \text{αν } u\leq 0 \end{cases}$

2. $f(u)=\begin{cases} +1, & \text{αν } u>0 \\ -1, & \text{αν } u\leq 0. \end{cases}$

Στην πρώτη περίπτωση η έξοδος $y=f(u)$ είναι δυαδική, λαμβάνοντας τιμή $y=1$ εάν ο νευρώνας ενεργοποιηθεί ή $y=0$ εάν ο νευρώνας παραμείνει αδρανής. Στη δεύτερη περίπτωση, η έξοδος $y=f(u)$ είναι λαμβάνει τιμή $y=1$ εάν ο νευρώνας ενεργοποιηθεί ή $y=-1$ εάν η διέγερση u δεν ενεργοποιήσει τον νευρώνα. Η ενεργητικότητα του Perceptron εξαρτάται από τρεις παραμέτρους: τα βάρη των συνάψεων, τις τιμές εισόδου και την τιμή του κατωφλιού. Πολλές φορές, το κατώφλι ενεργοποίησης θ μπορεί να θεωρηθεί ως ενα επιπλέον συναπτικό βάρος $w_0=\theta$, το οποίο αντιστοιχεί στην έξοδο

$x_0 = -1$ και είναι ξεχωριστό από τα υπόλοιπα βάρη αλλά δρα με τον ίδιο τρόπο. Στην περίπτωση αυτή, η διέγερση εκφράζεται με τη σχέση:

$$u = \sum_{i=1}^{n} w_i \cdot x_i - \theta = \sum_{i=1}^{n} w_i \cdot x_i - w_0 \cdot x_0 = \sum_{i=0}^{n} w_i \cdot x_i, \qquad 2.3$$

Η παραπάνω εξίσωση αποτελεί το εσωτερικό γινόμενο των διανυσμάτων w και x ($u = w^T \cdot x$). Με την πρόσθεση μιας επιπλέον εισόδου x_0 με σταθερή τιμή -1, το διάνυσμα εισόδου γίνεται $x = [x_0, \ldots, x_n]^T$ και το αντίστοιχο διάνυσμα βαρών γίνεται $w = [w_1, \ldots, w_n]^T$. Το βάρος $w_0 = -\theta$ θεωρείται ότι αντιστοιχεί στην σταθερή είσοδο $x_0 = +1$ και ονομάζεται πόλωση (bias ή b). Ο όρος $w_0 = b$ δεν έχει καμία φυσική σημασία, αλλά αντιμετωπίζεται ως ένα εξωτερικό ερέθισμα που προστίθεται στο υπόλοιπο άθροισμα για να αποδώσει το σωστό δυναμικό ενεργοποίησης του νευρώνα.

Η εκπαίδευση του Perceptron αφορά την αυτόματη εκμάθηση των παραμέτρων του ώστε να επιτευχθεί ο επιθυμητός στόχος, δηλαδή η εύρεση της διαχωριστικής γραμμής μεταξύ δύο κλάσεων C_1 και C_2. Ουσιαστικά αναζητείται ένα διάνυσμα βαρών w τέτοιο ώστε να ικανοποιείται μία από τις παρακάτω ανισότητες:

i. $w^T \cdot x > 0$ για κάθε πρότυπο εισόδου που ανήκει στην κλάση C_1

ii. $w^T \cdot x < 0$ για κάθε πρότυπο εισόδου που ανήκει στην κλάση C_2

Εφόσον έχουμε μάθηση με επίβλεψη, η εκπαίδευση γίνεται με επιθυμητή έξοδος δίνεται για κάθε πρότυπο εκπαίδευσης (εισόδου) x χρησιμοποιώντας κάποιον επαναληπτικό αλγόριθμο. Ο κλασσικός κανόνας εκπαίδευσης είναι αυτός της σταθερής αύξησης (fixed increment rule) όπου τα πρότυπα παρουσιάζονται με κυκλική σειρά στο μοντέλο και, όταν τελειώσουν επαναλαμβάνονται από την αρχή. Ένας κύκλος εμφάνισης όλων των προτύπων (εισόδων) x του συνόλου εκπαίδευσης ονομάζεται εποχή (epoch). Με βάση τον παραπάνω κανόνα, το διάνυσμα των βαρών μεταβάλλεται μόνο όταν υπάρχει σφάλμα ταξινόμησης, δηλαδή όταν ο στόχος για το συγκεκριμένο πρότυπο εισόδου διαφέρει από την παραγόμενη έξοδο του νευρωνικού δικτύου.

Ο αλγόριθμος σύγκλισης για Perceptron n εισόδων με έναν υπολογιστικό νευρώνα, όπου το διάνυσμα εισόδων είναι το $x(n) = [x_0 = +1, x_1(n), \ldots, x_n(n)]^T$ και το διάνυσμα βαρών

$w(n)=[w_0=b, w_1(n), \ldots, w_n(n)]^T$ ακολουθεί παρακάτω:

Algorithm Perceptron

Input: $x(n)=[+1, x_1(n), \ldots, x_n(n)]^T$, $w(n)=[b, w_1(n), \ldots, w_n(n)]^T$ datasets of observations X and weights (εισαγωγή των δειγμάτων x και των βαρών w)

Output: $w(n)$, b final weights and bias b (τελικά βάρη και η πόλωση)

$w(0)=0$ (ανάθεση των βαρών w ίσα με 0)

For $k=1$ **to** $k=n$ (έλεγχος συνθήκης τερματισμού)

$x(n)=x_k(n)$ (εφαρμογή του διανύσματος k)

$u(n)=w^T(n)\cdot x(n)$ (υπολογισμός της διέγερσης του νευρώνα)

$y(n)=-sgn(u(n))=sgn[w^T(n)\cdot x(n)]$ (υπολογισμός της πραγματικής εξόδου του νευρώνα

όπου $sgn(u)=\begin{cases}+1 & \alpha\nu\ u>0\\ -1 & \alpha\nu\ u\le 0\end{cases}$)

$w(n+1)=w(n)+\eta\cdot[d(n)-y(n)]\cdot x(n)$ (προσαρμογή του διανύσματος των βαρών όπου

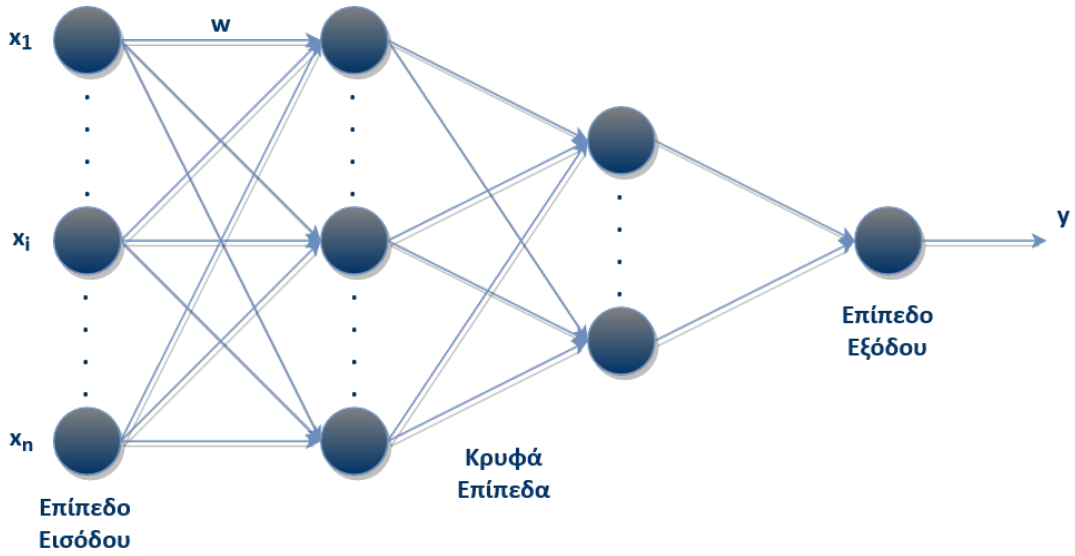
$\eta\in[0,1]$ είναι μια παράμετρος μάθησης και $d(n)=\begin{cases}+1 & \alpha\nu\ X(n)\in C_1\\ -1 & \alpha\nu\ X(n)\in C_2\end{cases}$)

end for

2.2.3 Πολυεπίδεδο δίκτυο Perceptron – Multi Layer Perceptron (MLP)

Το δίκτυο Perceptron πολλών επιπέδων (MLP) αποτελεί γενίκευση του απλού Perceptron (Bishop, 1995). Αποτελείται από ένα σύνολο κόμβων που αποτελούν το επίπεδο εισόδου (πηγαίοι κόμβοι), ένα ή περισσότερα κρυφά επίπεδα υπολογιστικών κόμβων (hidden layers) και ένα επίπεδο υπολογιστικών κόμβων εξόδου. Βασικό χαρακτηριστικό του πολυεπίπεδου Perceptron είναι ότι οι νευρώνες οποιουδήποτε στρώματος l τροφοδοτούν αποκλειστικά τους νευρώνες του επόμενου στρώματος l+1 και

τροφοδοτούνται αποκλειστικά από τους νευρώνες του προηγούμενου στρώματος $l-1$. Το σήμα εισόδου κινείται επομένως από τα πίσω προς τα μπροστά επίπεδα. Παρακάτω απεικονίζεται η αρχιτεκτονική ενός απλού MLP με 2 κρυφά επίπεδα:



ΣΧΗΜΑ 2.2 Δίκτυο MLP 2 κρυφών επιπέδων

Τα δίκτυα MLP αποτελούν μέθοδο ταξινόμησης επιβλεπόμενης μάθησης, επομένως απαιτούν την ύπαρξη συνόλου εκπαίδευσης για τη σωστή εκτίμηση του διανύσματος βαρών w. Για το σκοπό αυτό χρησιμοποιείται ο αλγόριθμος οπίσθιας διάδοσης σφάλματος (back-propagation), ο οποίος βασίζεται στον κανόνα μάθησης με διόρθωση σφάλματος και αποτελεί γενίκευση του αλγορίθμου Ελαχίστων Μέσων Τετραγώνων (Least Mean Square Algorithm). Η λειτουργία της οπίσθιας διάδοσης σφάλματος περιγράφεται αναλυτικά στην Υποενότητα 2.2.4. Ένα δίκτυο MLP έχει τα τρία παρακάτω χαρακτηριστικά:

(1) Το μοντέλο κάθε νευρώνα στο δίκτυο περιλαμβάνει μια μη-γραμμική συνάρτηση ενεργοποίησης στην έξοδό του. Είναι σημαντικό να τονιστεί ότι η μη γραμμικότητα αυτή πρέπει να είναι ομαλή, δηλαδή παντού διαφορίσιμη. Μια κοινώς χρησιμοποιούμενη μορφή μη γραμμικότητας που ικανοποιεί αυτή την απαίτηση είναι η σιγμοειδής (sigmoid) ή λογιστική συνάρτηση, που ορίζεται από την παρακάτω σχέση:

$$y_j = \frac{1}{\left(1+e^{(-u_j)}\right)}, \qquad 2.4$$

όπου u_j είναι η διέγερση του νευρώνα j και y_j η έξοδος του νευρώνα.

Η παρουσία μη-γραμμικότητας είναι σημαντική καθώς σε διαφορετική περίπτωση η σχέση εισόδου-εξόδου του δικτύου θα εκφυλιζόταν σε αυτή του μονοεπίπεδου Perceptron. Επιπλέον, η χρήση της συνάρτησης αυτής έχει το πλεονέκτημα ότι προσομοιάζει τη βιολογική φάση της ανάσχεσης (refractory phase) που συναντάται στους φυσικούς νευρώνες.

(2) Το δίκτυο MLP περιλαμβάνει ένα ή περισσότερα επίπεδα κρυφών νευρώνων που δεν αποτελούν ούτε μέρος της εισόδου, ούτε μέρος της εξόδου του δικτύου, αλλά ενδιάμεσο στάδιο. Οι κρυφοί νευρώνες επιτρέπουν στο δίκτυο να εκπαιδεύεται και να εκτελεί περίπλοκες εργασίες, εξάγοντας προοδευτικά τα βάρη των χαρακτηριστικών εκείνων που έχουν τη μεγαλύτερη σημασία για την επιτυχή ταξινόμηση των προτύπων στις σωστές κλάσεις.

(3) Το δίκτυο παρουσιάζει υψηλό βαθμό συνδεσιμότητας που καθορίζεται από τις συνάψεις μεταξύ των νευρώνων του δικτύου. Μια αλλαγή στη συνδεσιμότητα του δικτύου επιφέρει συνολική αλλαγή στον πληθυσμό των συνάψεων ή στις τιμές των βαρών τους.

Στο Σχήμα 2.2 παρουσιάστηκε η αρχιτεκτονική ενός πλήρους διασυνδεδεμένου MLP δύο κρυφών επιπέδων. Ένα δίκτυο καλείται πλήρως διασυνδεδεμένο (fully connected) όταν κάθε νευρώνας του είναι συνδεδεμένος με όλους τους νευρώνες του προηγούμενου επιπέδου. Σε δίκτυα αυτής της μορφής το σήμα μεταδίδεται προοδευτικά από αριστερά προς τα δεξιά και από επίπεδο σε επίπεδο. Στην περίπτωση που κάποιες από τις συνδέσεις δεν υπάρχουν, τότε το δίκτυο καλείται μερικώς διασυνδεδεμένο (partially connected).

Η μετάδοση σημάτων στα δίκτυα MLP αφορά τα σήματα συναρτήσεων (function signals) και τα σήματα σφάλματος (error signals), τα οποία έχουν αντίθετη φορά μεταξύ τους. Συγκεκριμένα, τα σήματα συναρτήσεων αποτελούν ερεθίσματα που ξεκινούν από τους κόμβους εισόδου του δικτύου και διαδίδονται προς τα εμπρός, καταλήγοντας στους νευρώνες εξόδου. Σε κάθε νευρώνα του MLP από τον οποίο περνάει το σήμα, αυτό υπολογίζεται ως συνάρτηση όλων των εισερχόμενων σημάτων και των αντίστοιχων βαρών των συνάψεων που καταλήγουν στο συγκεκριμένο νευρώνα. Αντίθετα, τα σήματα

σφάλματος ξεκινούν από τους νευρώνες εξόδου και διαδίδονται προς τα πίσω από επίπεδο σε επίπεδο, υπολογιζόμενα από μια συνάρτηση που εξαρτάται από την διαφορά επιθυμητής εξόδου με την παραγόμενη έξοδο από το δίκτυο.

Οι νευρώνες στα ενδιάμεσα στάδια του δικτύου MLP καθώς και οι νευρώνες εξόδου εκτελούν τους εξής υπολογισμούς:

i. Υπολογίζουν το σήμα συνάρτησης που εμφανίζεται στην έξοδο του νευρώνα ως μια συνεχή μη-γραμμική συνάρτηση των σημάτων που εισέρχονται σε αυτόν καθώς και των αντίστοιχων βαρών των συνάψεων που σχετίζονται με τον νευρώνα.

ii. Υπολογίζουν τη στιγμιαία προσέγγιση του διανύσματος κλίσης (δηλαδή της κλίσης της επιφάνειας σφάλματος ως προς τα βάρη που σχετίζονται με τις συνάψεις του νευρώνα) με χρήση της οπίσθιας διάδοσης σφάλματος που θα περιγραφεί στην επόμενη υποενότητα.

2.2.4 Αλγόριθμος οπίσθιας διάδοσης σφάλματος (error back propagation)

Όπως αναφέρθηκε παραπάνω, η εκπαίδευση των δικτύων MLP γίνεται με τον αλγόριθμο οπίσθιας διάδοσης σφάλματος (error back propagation), ο οποίος μπορεί να θεωρηθεί σαν μία γενίκευση του αλγορίθμου Ελαχίστων Μέσων Τετραγώνων - EMT (Least Mean Squares). Η διαδικασία εκπαίδευσης στον back propagation περιλαμβάνει υπολογισμούς που υλοποιούνται σε δύο περάσματα μέσω των επιπέδων του δικτύου: ένα κατά την ευθεία φορά και ένα κατά την αντίστροφη φορά (από την έξοδο προς την είσοδο). Κατά την εφαρμογή του αλγόριθμου, η εκπαίδευση επιτυγχάνεται με την παρουσίαση στο δίκτυο ενός συνόλου παραδειγμάτων εκπαίδευσης, ενώ με την ολοκλήρωση εμφάνισης όλων των προτύπων εκπαίδευσης του συνόλου (δειγμάτων), ολοκληρώνεται μια εποχή (epoch). Η εκπαίδευση γίνεται προκειμένου να επιτευχθεί η κατάλληλη σύγκλιση των βαρών κάθε διανύσματος εκπαίδευσης βάσει του μέσου τετραγωνικού σφάλματος. Η διαδικασία τερματίζεται όταν το μέσο τετραγωνικό σφάλμα μειωθεί κάτω από μια προκαθορισμένη αποδεκτή τιμή. Η αλγοριθμική διαδικασία φαίνεται παρακάτω:

Algorithm Back-Propagation

Input: $y^0 = p$ dataset of initial observations p (εισαγωγή των αρχικών δειγμάτων p)

Output: $w(n)$, b final weights and bias b (τελικά βάρη και η πόλωση)

$w(0)$ (αρχικοποίηση των βαρών w π.χ. με τυχαίο τρόπο)

For $a = 1$ **to** $a = K$ (έλεγχος συνθήκης τερματισμού epoch)

For $c = 1$ **to** $c = N$ (έλεγχος συνθήκης τερματισμού προτύπου εκπαίδευσης)

$y^0 = p_c$ (τροφοδότηση με νέο πρότυπο εκπαίδευσης)

$y^{m+1} = f^{m+1}\left(W^{m+1} + y^m + b^{m+1}\right),\ m = 1, \ldots, M$ (υπολογισμός διεγέρσεων κάθε σταδίου – forward pass)

$s^m = -2 \cdot F^m(n^m) \cdot \left(W^{m+1}\right)^T \cdot s^{m+1},\ m = 0, 1, \ldots, M$ (ανάστροφος υπολογισμός των σημάτων σφάλματος – backward pass)

$W^m(k+1) = W^m(k) - a \cdot s^m \cdot \left(y^{m-1}\right)^T$ (υπολογισμός των βαρών)

$b^m(k+1) = b^m(k) - a \cdot s^m$ (υπολογισμός των πολώσεων)

end for

end for

Η εκπαίδευση με τη χρήση back propagation για ένα δεδομένο σύνολο εκπαίδευσης διακρίνεται σε δύο κατηγορίες: σε εκπαίδευση ανά ομάδα (batch training) και σε εκπαίδευση ανά πρότυπο (online training). Στην πρώτη περίπτωση, το δίκτυο τροφοδοτείται με όλα τα πρότυπα εκπαίδευσης και στη συνέχεια γίνεται η προσαρμογή των βαρών, τα οποία ενημερώνονται βάσει του μέσου τετραγωνικού σφάλματος που προκύπτει έπειτα από την παρουσίαση όλου του συνόλου εκπαίδευσης. Έτσι, η αναπροσαρμογή των βαρών γίνεται μία φορά μετά το τέλος κάθε εποχής με στόχο την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος για κάθε είσοδο, κάτι που απαιτεί μεγάλες ικανότητες αποθήκευσης και επεξεργασίας. Στην online εκπαίδευση αντίθετα, η μεταβολή των βαρών του δικτύου γίνεται για κάθε νέο πρότυπο που παρουσιάζεται, εκτελώντας ουσιαστικά τόσο το ευθύ όσο και το αντίστροφο πέρασμα που περιγράφηκε προηγουμένως. Στην περίπτωση αυτή, μια εποχή ολοκληρώνεται με την παρουσίαση στο

δίκτυο του τελευταίου προτύπου του συνόλου εκπαίδευσης.

Στην πράξη προτιμάται το online training, ιδιαίτερα όταν η εκπαίδευση αφορά μεγάλα σύνολα δεδομένων που παρουσιάζουν προκλήσεις τόσο σε επίπεδο υπολογιστικών πόρων όσο και στον χρόνο εκπαίδευσης. Ακόμη, χάρη στην εκπαίδευση ανά πρότυπο ενισχύεται ο στοχαστικός χαρακτήρας της διαδικασίας και μειώνεται ο κίνδυνος παγίδευσης του αλγορίθμου σε τοπικά ακρότατα. Αξίζει, ωστόσο, να σημειωθεί ότι η εκπαίδευση ανά ομάδα υπολογίζει με μεγαλύτερη ακρίβεια το διάνυσμα κλίσης της επιφάνειας σφάλματος.

Μια ακόμη δυνατή διαφοροποίηση στην εκτέλεση της οπίσθιας διάδοσης σφάλματος αφορά την σειρά τροφοδότησης των προτύπων εκπαίδευσης στο δίκτυο. Ο σειριακός τρόπος τροφοδότησης συνήθως συνοδεύεται από χαμηλό σφάλμα εκπαίδευσης, αλλά εγκυμονεί τον κίνδυνο της υπερπροσαρμογής του δικτύου στο σύνολο εκπαίδευσης, οδηγώντας τελικά σε υψηλότερο σφάλμα ταξινόμησης. Για τον λόγο αυτό, συνήθως προτιμάται η τυχαία σειρά τροφοδότησης των προτύπων εκπαίδευσης, η οποία μπορεί μεν να χαρακτηρίζεται από μικρές ταλαντώσεις του σφάλματος εκπαίδευσης, ωστόσο ενισχύει το στοχαστικό χαρακτήρα της μάθησης και συνήθως οδηγεί σε χαμηλότερο σφάλμα ταξινόμησης.

2.3 Random Forest

2.3.1 Εισαγωγή

Ο αλγόριθμος Random Forest είναι μια μέθοδος εκμάθησης συνόλων για κατηγοριοποίηση και λειτουργεί με το να κατασκευάζει ένα πλήθος από δέντρα απόφασης σε δείγμα του συνόλου των δεδομένων κατά τη φάση εκπαίδευσης του μοντέλου και ύστερα να συνυπολογίζει όλα τα δέντρα για να καθορίσει την τελική έξοδο (Breiman, 2001). Αποτελεί στην πραγματικότητα μία τροποποίηση της μεθόδου bagging,

η οποία παίρνει πολλά αμερόληπτα μοντέλα με θόρυβο και βρίσκει την μέση τιμή αυτών, μειώνοντας την διακύμανση της εξόδου. Ο Random Forest προσθέτει σε αυτήν την αρχική ιδέα μεταφέροντας κατά μια έννοια την μεθοδολογία του bootstrap, εκτός από τα δείγματα (instances) και στα χαρακτηριστικά (attributes) των δεδομένων.

Οι ιδιότητες που έχουν τα δέντρα απόφασης, τα καθιστούν ιδανικούς υποψηφίους για μεθόδους bagging διότι μπορούν να συλλάβουν πολύπλοκες αλληλεπιδράσεις μεταξύ των δεδομένων, ενώ αν μεγαλώσουν αρκετά βαθιά, έχουν σχετικά χαμηλή μεροληψία. Επιπρόσθετα, λόγω του θορύβου που έχουν, ο μέσος όρος τους αποτελεί έναν καλό δείκτη της πραγματικής εξόδου. Κάθε δέντρο που παράγεται θεωρείται ότι είναι πανομοιότυπα κατανεμημένο (identically distributed), δηλαδή είναι ανεξάρτητο από τα άλλα δέντρα και παρέχει την ίδια κατανομή πιθανότητας ως προς την τελική έξοδο. Με αυτό τον τρόπο, ο μέσος όρος όλων των δέντρων παρέχει την ίδια μεροληψία με αυτή που παρέχει ένα δέντρο από μόνο του, οπότε πετυχαίνουμε βελτίωση μέσω μείωσης της διακύμανσης

Τα βασικότερα πλεονέκτηματά αυτού του αλγορίθμου είναι ότι έχει βέλτιστη ακρίβεια μεταξύ των υπαρχόντων αλγορίθμων, η ταχύτητά του είναι πολύ καλή ακόμα και σε πολύ μεγάλα σύνολα δεδομένων εκπαίδευσης και τέλος μπορεί να χειριστεί αποδοτικά πάρα πολύ μεγάλο αριθμό χαρακτηριστικών (ακόμα και χιλιάδες εγγραφές). Ακόμη μερικά πλεονεκτήματα αυτού του αλγορίθμου είναι ότι δίνει μια εκτίμηση για το ποια χαρακτηριστικά είναι τα πιο σημαντικά στην κατηγοριοποίηση, δεν χρειάζεται την χρήση διαφορετικού συνόλου δεδομένων για τον έλεγχο ακρίβειας (δεν είναι δηλαδή απαραίτητο το cross--validation), καθώς η εκτίμηση του λάθους γενίκευσης γίνεται από τον ίδιο τον αλγόριθμο κατά την εκτέλεσή του. Τέλος να σημειωθεί ότι μπορεί να χειριστεί αποδοτικά ελλιπή δεδομένα και δεν παρουσιάζει φαινόμενα υπερεκπαίδευσης.

2.3.2 Δέντρα Απόφασης

Από τους δημοφιλέστερους αλγόριθμους μηχανικής μάθησης με επίβλεψη είναι τα δέντρα απόφασης (decision trees). Είναι μια μέθοδος κατασκευής και προσέγγισης

συναρτήσεων με έξοδο διακριτές τιμές. Το παραγόμενο αποτέλεσμα αυτού του αλγορίθμου είναι μια δενδροειδής μορφή, με κόμβους και κλαδιά, η οποία περιγράφει τα δεδομένα και τους κανόνες που χρησιμοποιήθηκαν για την κατασκευή του δέντρου (decision rules).

Κάθε εσωτερικός κόμβος του δέντρου, αντιστοιχεί σε μια συνθήκη ελέγχου της τιμής ενός χαρακτηριστικού (attribute) των περιπτώσεων (instances) που χρησιμοποιήθηκαν στην κατασκευή του δέντρου. Κάθε κλαδί που φεύγει από έναν κόμβο πατέρα προς τα κάτω (επόμενο επίπεδο), αντιστοιχεί σε διαφορετική πιθανή τιμή ή εύρος τιμών με κοινή κλάση του χαρακτηριστικού που ελέγχθηκε στον συγκεκριμένο κόμβο. Προχωρώντας στο παρακάτω επίπεδο, τα δεδομένα διαχωρίζονται ως προς αυτό το χαρακτηριστικό, και αυτή η διαδικασία συνεχίζεται μέχρις ότου ο αλγόριθμος καταλήξει στους εξωτερικούς κόμβους του δέντρου (φύλλα), οι οποίοι αντιστοιχούν στις τιμές εξόδου (classes) των δεδομένων. Προκειμένου να χαρακτηριστεί κάποιος κόμβος του δέντρου ως φύλλο, θα πρέπει όλα τα δεδομένα που ανήκουν σε αυτόν να ανήκουν και στην ίδια κατηγορία (κλάση) μεταξύ τους, και τότε η τιμή της κλάσης αυτής θα είναι και η τιμή του αντίστοιχου κόμβου.

Το σημαντικότερο πλεονέκτημα των δέντρων απόφασης είναι η ευκολία που προσφέρουν στην φυσική ερμηνεία των αποτελεσμάτων που εξάγονται από αυτά. Οι δεντρικές δομές είναι πολύ απλές στην χρήση τους, στην κατανόησή τους από τον χρήστη, αλλά και στην κατασκευή τους, καθώς το υπολογιστικό τους κόστος είναι της τάξης του $O(\log(n))$ στην περίπτωση των δυαδικών δέντρων, τα οποία είναι και τα πιο συνηθισμένα. Αντίστοιχο είναι το κόστος το οποίο έχουμε και στις περιπτώσεις ανεύρεσης και ταξινόμησης δεδομένων που βρίσκονται σε μορφή δέντρων.

Όπως είναι φανερό, το βασικό σημείο της διαδικασίας είναι η εύρεση του χαρακτηριστικού που θα χρησιμοποιηθεί για την κατασκευή του εκάστοτε κόμβου και τον διαχωρισμό τον δεδομένων στο επόμενο επίπεδο. Οι μέθοδοι που εφαρμόζονται είναι στατιστικές και υπολογίζουν μεταβλητές οι οποίες μας δίνουν ένα μέτρο του “βέλτιστου” χαρακτηριστικού.

Η επιλογή του βέλτιστου χαρακτηριστικού γίνεται σε κάθε βήμα που πρέπει να υπολογιστεί (το οποίο αντιστοιχεί και σε έναν νέο κόμβο, ένα επίπεδο πιο κάτω), και σε αυτήν συμμετέχουν κάθε φορά μόνο τα δεδομένα που ανήκουν σε αυτό τον κόμβο που

είναι τώρα ο αλγόριθμος (σύμφωνα με όλους τους προηγούμενους διαχωρισμούς) και όχι ολόκληρο το αρχικό σύνολο δεδομένων.

Η μέθοδος εύρεσης του βέλτιστου χαρακτηριστικού χρησιμοποιεί μια έννοια από την θεωρία πληροφορίας. Αυτή είναι η έννοια της εντροπίας η οποία, όπως θα εξηγηθεί και στο επόμενο κεφάλαιο, αποτελεί ένα μέτρο της συνεισφοράς του συγκεκριμένου χαρακτηριστικού στην κλάση που ανήκουν τα δείγματα (instances). Η εξίσωση αυτής είναι η παρακάτω:

$$H = -\sum_{i=1}^{n} p_i \cdot \log(p_i). \qquad 2.5$$

Βέβαια η εντροπία δεν δίνει καμιά πληροφορία για το κατά πόσο θα μεταβληθεί η εντροπία στα εναπομείναντα δεδομένα, εφόσον επιλεγεί το συγκεκριμένο χαρακτηριστικό. Γι' αυτό τον λόγο εισάγεται η έννοια του "Κέρδους Πληροφορίας" (Information Gain), το οποίο δείχνει κατά πόσο θα μειωθεί η εντροπία των δεδομένων εκπαίδευσης, στην περίπτωση που επιλεχθεί το επόμενο χαρακτηριστικό A_i, από το σύνολο των χαρακτηριστικών $A_i, \ldots, A_n$. Η εξίσωση του κέρδους είναι η ακόλουθη:

$$G(S, A_i) = H - \sum_{u \in A_i} \frac{|S_u|}{|S|} \cdot H(S_u), \qquad 2.6$$

όπου S είναι το σύνολο εκπαίδευσης και S_u το σύνολο των δεδομένων με τιμή u στο χαρακτηριστικό A_i.

Στην ουσία το κέρδος πληροφορίας δίνει την μείωση της εντροπίας στο σύνολο των δεδομένων, αν επιλεγεί το χαρακτηριστικό A_i ως μεταβλητή διαχωρισμού. Μείωση της εντροπίας οδηγεί στην αύξηση της πυκνότητας της πληροφορίας και συνεπώς και σε περισσότερο "συμπαγή" δεδομένα. Στην πράξη, ο δεύτερος όρος της προηγούμενης σχέσης, δίνει την εντροπία των δεδομένων μετά τον διαχωρισμό που θα γίνει με βάση το συγκεκριμένο χαρακτηριστικό. Συνεπώς, σε κάθε κόμβο υπολογίζεται το κέρδος πληροφορίας για κάθε υποψήφιο χαρακτηριστικό, και επιλέγεται αυτό που δίνει το μέγιστο κέρδος κάθε φορά.

Η κατασκευή του δέντρου γίνεται με την χρήση αναδρομικών μεθόδων, από πάνω προς τα κάτω (top down) και αποτελείται από τα εξής στάδια:

Algorithm Decision Tree

Input: S dataset of classified observations (εισαγωγή των δεδομένων με τις κλάσεις τους)

Output: Decision Tree structure (δομή δέντρου απόφασης)

$A \leftarrow \max(G(S, A_i))$ (εύρεση του χαρακτηριστικού που μεγιστοποιεί το κέρδος της πληροφορίας)

$Root \leftarrow A$ (ορισμός αυτού του χαρακτηριστικού ως ρίζα του δέντρου)

RECURSIVE STEP: For $i = 1$ **to** $i = u_{A_i}$ (έλεγχος συνθήκης κάθε πιθανής τιμής ή εύρους τιμών του χαρακτηριστικού A_i)

$Attributes \leftarrow Attributes \setminus A$ (αφαίρεση από το σύνολο των διαθέσιμων χαρακτηριστικών)

ADD Branch (εισαγωγή κλαδιού)

SPLIT subsets (διαχωρισμός δεδομένων σε υποσύνολα)

If $subset = \emptyset$ **then** (έλεγχος συνθήκης για κενό υποσύνολο)

ADD Leaf (εισαγωγή φύλλου με ετικέτα την πιο συνήθη κλάση)

Else

$A \leftarrow \max(G(S, A_i))$, $A_i \in Attributes$ (εύρεση του νέου χαρακτηριστικού που μεγιστοποιεί το κέρδος της πληροφορίας)

ADD Node (εισαγωγή κόμβου)

GoTo RECURSIVE STEP

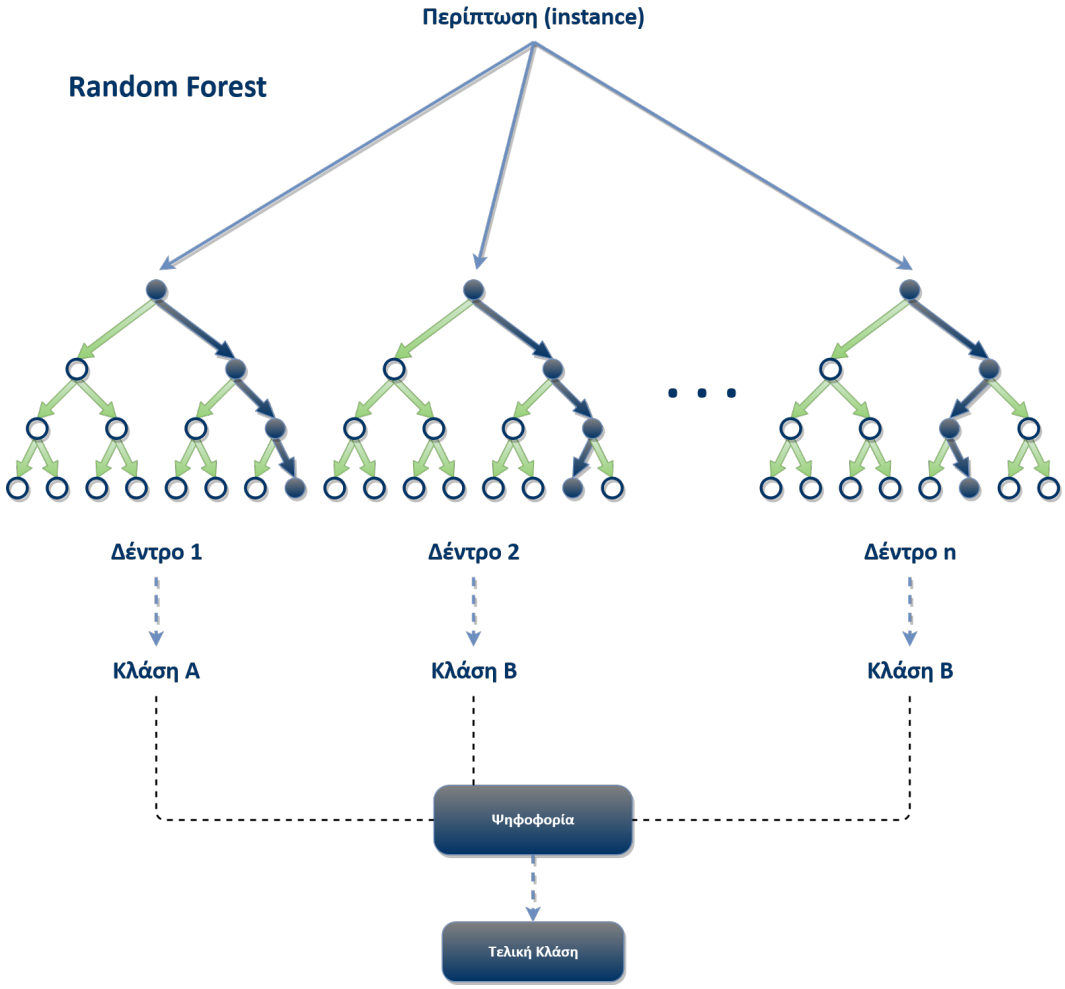
end if

end for

2.3.3 Random Forest

Η καινοτομία του αλγορίθμου Random Forest σε σχέση με τα απλά δέντρα απόφασης, έγκειται στη χρήση της μεθόδου Bagging (Bootstrap Aggregating) (Efron και Gong, 1983).

Χρησιμοποιώντας την μέθοδο bootstrapping, καθίσταται δυνατή η δημιουργία m νέων υποσυνόλων εκπαίδευσης από ένα αρχικό σύνολο δεδομένων, σε κάθε ένα από τα οποία εφαρμόζεται ο επιθυμητός αλγόριθμος, και η εξάγονται m διαφορετικά αποτελέσματα (δέντρα). Η γραφική αναπαράσταση του αλγορίθμου φαίνεται στο σχήμα που ακολουθεί (Σχήμα 2.3).



ΣΧΗΜΑ 2.3 Αλγόριθμος Random Forest

Ο αλγόριθμος που χρησιμοποιείται για την εκπαίδευση ονομάζεται βασικός ταξινομητής (base classifier) και στην περίπτωση του Random Forest είναι το Decision Tree. Αυτά τα αποτελέσματα στο επόμενο βήμα συγκρίνονται μεταξύ τους έτσι ώστε να επιλεχθεί το καταλληλότερο. Στην περίπτωση των προβλημάτων ταξινόμησης, στα αποτελέσματα χρησιμοποιείται η μέθοδος της ψηφοφορίας (voting), ενώ αν το πρόβλημα είναι πρόβλημα παλινδρόμησης (regression) τότε υπολογίζεται συνήθως ο μέσος όρος. Γενικά

μπορούν να χρησιμοποιηθούν διάφορες στατιστικές μέθοδοι για να συγκριθούν τα αποτελέσματα. Το ποια θα εφαρμοστεί εξαρτάται από την φύση του προβλήματος που εξετάζεται. Η μέθοδος αυτή ονομάστηκε Bagging από τα αρχικά των λέξεων Bootstrap Aggregating και δίνει καλύτερα αποτελέσματα από ότι θα έδινε η εφαρμογή μόνο του βασικού ταξινομητή στο αρχικό σύνολο δεδομένων. Ο αλγόριθμος του Random Forest ακολουθεί παρακάτω:

Algorithm Random Forest

Input: `S` dataset of classified observations, `B` number of trees (εισαγωγή των δεδομένων με τις κλάσεις τους και του πλήθους των δέντρων)

Output: Random Forest Model

For `i=1` **to** `i=B` (έλεγχος συνθήκης κάθε υποσυνόλου i)

`Feature Bagging` (επιλογή bootstrap δείγματος με N δεδομένα)

Do`Algorithm Decision Tree` (υλοποίηση του αλγορίθμου δέντρου απόφασης για το υποσύνολο χαρακτηριστικών)

end for

`CONSTRUCT ensemble model` (δημιουργία μοντέλου με χρήση voting)

2.4 Αλγόριθμος k-πλησιέστερων γειτόνων (k-nearest neighbours, kNN)

2.4.1 Εισαγωγή

Η μέθοδος k-πλησιέστερων γειτόνων αποτελεί μια ιδιαίτερα διαδεδομένη και ευρέως χρησιμοποιούμενη τεχνική ταξινόμησης που βασίζεται στη χρήση μέτρων τα οποία είναι βασισμένα στην απόσταση (Fix και Hodges, 1951) (Cover και Hart, 1967). Ως μέθοδος

επιβλεπόμενης μάθησης, προϋποθέτει ότι το σύνολο εκπαίδευσης δεν περιλαμβάνει μόνο τα δεδομένα εισόδου (χαρακτηριστικά) αλλά επίσης και την επιθυμητή κατηγοριοποίηση (κλάση) για κάθε στοιχείο. Για την κατηγοριοποίηση κάθε νέου στοιχείου σε κάποια κλάση, απαιτείται ο υπολογισμός της απόστασής του από κάθε στοιχείο του συνόλου εκπαίδευσης, λαμβάνοντας τελικά υπόψη μόνο τις k κοντινότερες εκχωρήσεις.

Η μέθοδος k-πλησιέστερων γειτόνων ανήκει στην οικογένεια ταξινομητών βασισμένων σε παραδείγματα (Instance Based Classifiers, IBC), στους οποίους η μάθηση βασίζεται στην αναλογία και όχι στην παραγωγή κάποιου μοντέλου γενίκευσης (όπως στην περίπτωση των νευρωνικών δικτύων ή των δέντρων απόφασης). Έτσι, στην περίπτωση του kNN δεν υπάρχει κάποιο στάδιο εκπαίδευσης και δεν παράγεται κάποιο μοντέλο, μέχρι να χρειαστεί να κατηγοριοποιηθεί μια νέα παρατήρηση. Για τον λόγο αυτό, οι κατηγοριοποιητές IBC καλούνται και “οκνηροί ταξινομητές” (lazy classifiers). Εφόσον για να κατηγοριοποιηθεί μια νέα παρατήρηση πρέπει να συγκριθεί με γνωστές παρατηρήσεις του συνόλου εκπαίδευσης, αυτό απαιτεί την αποθήκευση όλων ή τουλάχιστον ενός μέρους των δεδομένων εκπαίδευσης. Αντιθέτως, σε άλλες τεχνικές όπως οι SVM μπορούν να απορριφθούν όλες οι παρατηρήσεις εκπαίδευσης που δεν είναι διανύσματα υποστήριξης.

Στα πλεονεκτήματα της μεθόδου kNN συγκαταλέγεται η αποτελεσματικότητα ανίχνευσης σύνθετων εξαρτήσεων μεταξύ των μεταβλητών, η απλότητα στην υλοποίηση και χρήση και οι γενικά υψηλές επιδόσεις ταξινόμησης. Ωστόσο, το γεγονός ότι απαιτούνται πολλές συγκρίσεις μεταξύ παρατηρήσεων απαιτεί αντίστοιχα πολύ αποτελεσματικές τεχνικές καταλογοποίησης (indexing), ειδάλλως η κατηγοριοποίηση νέων παρατηρήσεων διαρκεί για περισσότερο χρόνο, ειδικά στις περιπτώσεις όπου ο αριθμός των εν δυνάμει γειτόνων είναι μεγάλος. Ακόμη, τα αποτελέσματα ταξινόμησης παρουσιάζουν ευαισθησία στα τοπικά χαρακτηριστικά των δεδομένων, στην ύπαρξη μη σημαντικών μεταβλητών εισόδου και στο πλήθος των γειτόνων, αυξάνοντας τον κίνδυνο υπερπροσαρμογής.

2.4.2 Μέτρα απόστασης

Όπως αναφέρθηκε, η βασική ιδέα την μεθόδου k-πλησιέστερων γειτόνων είναι ότι κάθε στοιχείο του συνόλου δεδομένων που απεικονίζεται στην ίδια κατηγορία θεωρείται ότι είναι πιο κοντά σε στοιχεία της ίδιας κατηγορίας από όσο είναι σε στοιχεία τα οποία ανήκουν σε άλλες κατηγορίες. Για το σκοπό αυτό, μπορούν να χρησιμοποιηθούν μέτρα ομοιότητας (ή απόστασης) ώστε να ποσοτικοποιηθεί η ομοιότητα ανάμεσα στα διαφορετικά στοιχεία του συνόλου δεδομένων. Το κάθε στοιχείο μπορεί να απεικονιστεί ως πλειάδα, δηλαδή ως σημείο ενός ν-διάστατου χώρου, όπου ν είναι ο αριθμός των χαρακτηριστικών που συνθέτουν τα δεδομένα εισόδου. Τα συνηθέστερα είδη αποστάσεων που χρησιμοποιούνται σαν μέτρα ομοιότητας ανάμεσα σε πλειάδες ενός συνόλου δεδομένων είναι τα εξής:

- Ευκλείδεια Απόσταση

$$d_{ij}=\sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2} \qquad 2.7$$

- Σταθμισμένη Ευκλείδεια Απόσταση

$$d_{ij}=\sqrt{\sum_{k=1}^{p}w_k^2\cdot\left(x_{ik}-x_{jk}\right)^2} \qquad 2.8$$

- Απόσταση Manhattan

$$d_{ij}=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right| \qquad 2.9$$

- Απόσταση Minkowski

$$d_{ij}=\left(\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{\lambda}\right)^{1/\lambda}. \qquad 2.10$$

Η Ευκλείδεια απόσταση αποτελεί το πλέον διαδεδομένο μέτρο απόστασης για την μέθοδο των k-πλησιέστερων γειτόνων. Ωστόσο ο υπολογισμός της ομοιότητας βάσει αυτής της απόστασης χαρακτηρίζεται από την αδυναμία ότι οι μεταβλητές που έχουν μεγάλο εύρος τιμών επηρεάζουν περισσότερο το αποτέλεσμα της μεθόδου από τις

μεταβλητές που έχουν μικρό εύρος τιμών. Αυτή η ιδιαιτερότητα εμποδίζει την αμεροληψία των αποτελεσμάτων της μεθόδου, καθώς οι μεταβλητές με μεγάλο εύρος τιμών αποκτούν μεγαλύτερο βάρος (ασκούν μεγαλύτερη επιρροή). Το πρόβλημα αυτό της μεροληψίας αντιμετωπίζεται με την κανονικοποίηση των αριθμητικών τιμών, π.χ. διαιρώντας τις τιμές του εκάστοτε γνωρίσματος με την περιοχή τιμών του αντίστοιχου γνωρίσματος. Άλλες λιγότερο γνωστές αποστάσεις που έχουν προταθεί ως μέτρα ομοιότητας είναι οι αποστάσεις Chebychev και Canberra, ο συντελεστής Czekanowski, τα μέτρα Dice και Jaccard καθώς και αυτοσχέδια μέτρα ομοιότητας που έχουν εφαρμοστεί κατά περίπτωση σε εξειδικευμένα προβλήματα.

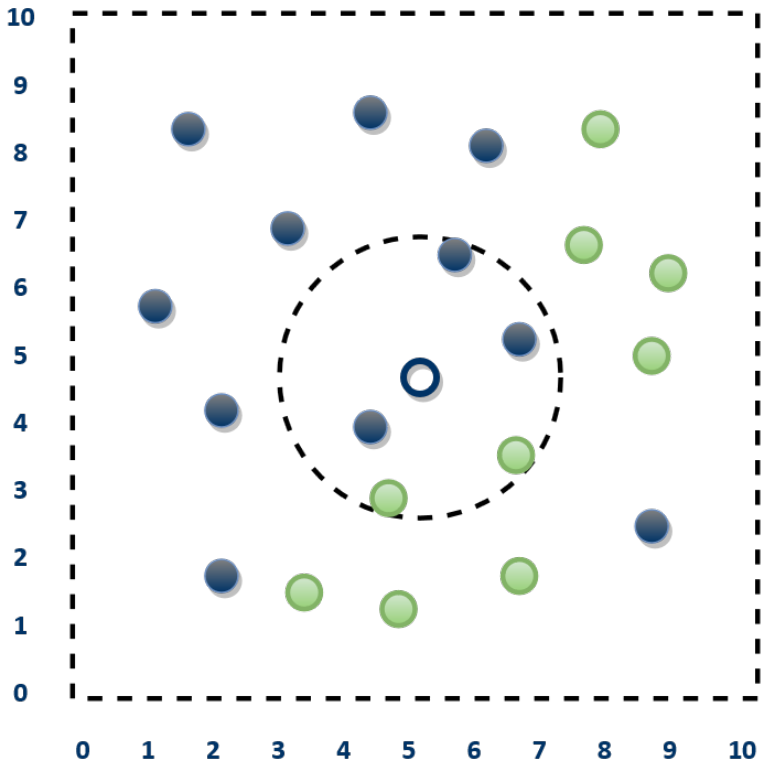
Ακόμη, ο υπολογισμός της ομοιότητας με βάση την Ευκλείδεια απόσταση προϋποθέτει την ισότιμη συμμετοχή όλων των γνωρισμάτων, κάτι που δεν ισχύει στη γενική περίπτωση. Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα, έχει προταθεί η εισαγωγή βαρών για κάθε διάσταση, ανάλογα με τις ανάγκες του προβλήματος. Μια εναλλακτική πρόταση είναι η απόφαση ταξινόμησης για ένα σημείο να λαμβάνεται με βάση την σταθμισμένη συμμετοχή των k πλησιέστερων γειτόνων, ώστε τα κοντινότερα σημεία στη νέα παρατήρηση να συνεισφέρουν περισσότερο, με χρήση κάποιου συντελεστή βαρύτητας.

2.4.3 Περιγραφή της μεθόδου k-πλησιέστερων γειτόνων

Η ταξινόμηση με χρήση της μεθόδου kNN βασίζεται στην εύρεση και στην εξέταση των πλησιέστερων γειτόνων κάθε νέας παρατήρησης. Ο αριθμός k, ο οποίος αντιστοιχεί στο πλήθος των πλησιέστερων γειτόνων που χρησιμοποιούνται για την επίτευξη της κατηγοριοποίησης ενός νέου σημείου με τη μεγαλύτερη δυνατή ακρίβεια, θεωρείται γνωστός εκ των προτέρων και σταθερός καθ'όλη τη διάρκεια των υπολογισμών. Η αλγοριθμική διαδικασία επιλογής της κατάλληλης κλάσης για μια νέα παρατήρηση μπορεί για λόγους απλότητας να εξηγηθεί για ένα πρόβλημα ταξινόμησης όπου κάθε στοιχείο (πλειάδα) του συνόλου εκπαίδευσης αποτελείται από δύο χαρακτηριστικά και χαρακτηρίζεται από μια κλάση. Σε αυτό το υποθετικό παράδειγμα, μια παρατήρηση X

μπορεί να θεωρηθεί ως ένα σημείο στον δισδιάστατο χώρο που απέχει κάποια απόσταση από μια άλλη παρατήρηση Y. Η απόσταση $d(X,Y)$ μπορεί να υπολογιστεί ως η Ευκλείδεια απόσταση σύμφωνα με την Εξίσωση 2.7 που αναφέρθηκε παραπάνω.

Ο αλγόριθμος αρχικά αναζητά μέσα στον δισδιάστατο χώρο τα k σημεία-παρατηρήσεις που βρίσκονται πλησιέστερα στην προς ταξινόμηση παρατήρηση και ύστερα την εκχωρεί στην κλάση που πλειοψηφεί μεταξύ των k πλησιέστερων γειτόνων. Για $k=1$, η νέα παρατήρηση εκχωρείται στην κλάση της πιο όμοιας παρατήρησης του συνόλου εκπαίδευσης, για $k=3$ στην κλάση στην οποία ανήκουν τουλάχιστον 2 από τις 3 πιο όμοιες παρατηρήσεις, κ.ο.κ. Ο ίδιος αλγόριθμος ισχύει κατ' αντιστοιχία για παρατηρήσεις με n αριθμητικές διαστάσεις (χαρακτηριστικά) και για k πλησιέστερους γείτονες. Σε αυτή την περίπτωση οι παρατηρήσεις θεωρούνται σημεία στον n-διάστατο χώρο και η Ευκλείδεια απόσταση υπολογίζεται αναλόγως. Η γραφική αναπαράσταση του αλγορίθμου k-πλησιέστερων γειτόνων φαίνεται στο σχήμα που ακολουθεί (Σχήμα 2.4).



ΣΧΗΜΑ 2.4 Αλγόριθμος k-πλησιέστερων γειτόνων για k=5

Η αλγοριθμική διαδικασία περιγράφεται στη συνέχεια:

Algorithm k Nearest Neighbours

Input: S dataset of classified observations, K number of nearest neighbours, t instance to be classified (εισαγωγή των δεδομένων με τις κλάσεις τους, του πλήθους των πλησιέστερων γειτόνων και της πλειάδας προς κατηγοριοποίηση)

Output: c Class of instance (κλάση της πλειάδας)

$I = \emptyset$ (ορισμός του συνόλου με τους k πλησιέστερους γείτονες ως το κενό)

For $i = 1$ **to** $i = m$ (έλεγχος συνθήκης κάθε σημείου - παρατήρηση i)

COMPUTE DISTANCE $d(x_i, t)$ (υπολογισμός απόστασης, συνήθως Ευκλείδειας $d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$)

end for

COMPUTE set I (υπολογισμός του συνόλου με τους k πλησιέστερους γείτονες ως αυτού με τις k μικρότερες αποστάσεις $d(x_i, t)$)

Return $\arg\max_i c_i$ (υπολογισμός της κλάσης στην οποία μπήκαν οι περισσότεροι k πλησιέστεροι γείτονες)

2.5 Μηχανές Διανυσμάτων Υποστήριξης

2.5.1 Εισαγωγή

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM) ανήκουν στην κατηγορία των επιβλεπόμενων μεθόδων μηχανικής μάθησης και λειτουργούν αναλύοντας τα δεδομένα και αναγνωρίζοντας μοτίβα με σκοπό την ταξινόμησή τους σε κλάσεις (Vapnick και Cortes, 1995). Η γενική ιδέα των μηχανών διανυσματικής υποστήριξης είναι ο υπολογισμός ενός υπερεπιπέδου διαχωρισμού, μεγιστοποιώντας

έτσι τα περιθώρια (αποστάσεις) μεταξύ των κλάσεων των δεδομένων. Οι ταξινομητές αυτού του είδους ανήκουν στην ευρύτερη περιοχή των μεθόδων πυρήνα (kernel methods), μιας κατηγορίας αλγορίθμων που η εξάρτησή τους από τα δεδομένα συνοψίζεται αποκλειστικά και μόνο στον υπολογισμό του εσωτερικού γινομένου. Αυτό μπορεί να αντικατασταθεί από μια συνάρτηση πυρήνα που χρησιμοποιείται για τον υπολογισμό του εσωτερικού γινομένου σε χώρο υψηλότερης διάστασης. Η ιδιότητα αυτή - αν και φαινομενικά αυξάνει την πολυπλοκότητα του αλγορίθμου - αποτελεί το σημαντικότερο χαρακτηριστικό του SVM, καθώς επιτρέπει την χρήση του σε δεδομένα τα οποία δεν έχουν προφανή αναπαράσταση σε ένα διανυσματικό χώρο συγκεκριμένης διάστασης (μη γραμμικά διαχωρίσιμα προβλήματα). Στον αντίποδα, η σωστή χρήση των SVM προϋποθέτει καλή κατανόηση του τρόπου λειτουργίας τους, προεπεξεργασία των μεταβλητών εισόδου, ενώ απαιτείται και η κατάλληλη επιλογή του πυρήνα και η ρύθμιση (finetuning) των αντίστοιχων παραμέτρων, προκειμένου να επιτευχθεί αποδεκτή απόδοση.

2.5.2 Γραμμικώς Διαχωρίσιμα Προβλήματα

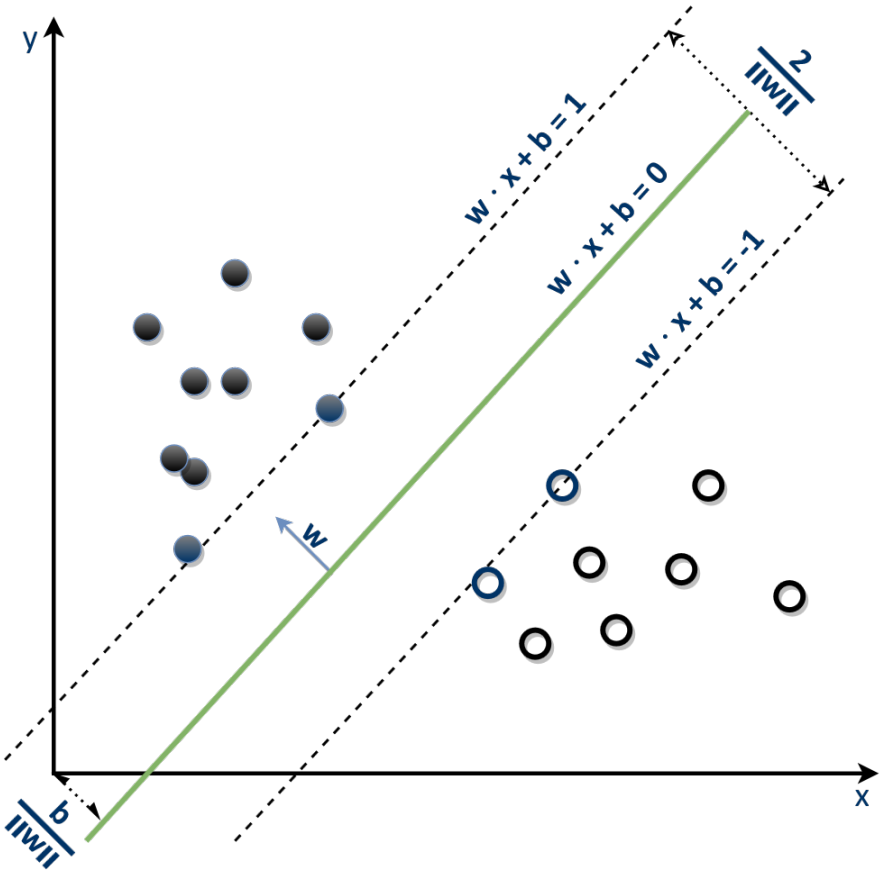
Οι Μηχανές Διανυσματικής Υποστήριξης μπορούν να χρησιμοποιηθούν για την επίλυση ενός προβλήματος δύο κλάσεων, μεγιστοποιώντας την απόσταση μεταξύ των πλησιέστερων σημείων της κάθε κλάσης. Το αποτέλεσμα είναι η εύρεση του (μοναδικού) διαχωριστικού υπερεπιπέδου που μεγιστοποιεί το περιθώριο στο σύνολο εκπαίδευσης, επιτυγχάνοντας ουσιαστικά καλύτερη απόδοση ταξινόμησης.

Έστω ένα σύνολο από N ζευγάρια δειγμάτων (ζεύγη εισόδου - εξόδου) του συνόλου εκπαίδευσης: $\{x_i, y_i\}$, $i=1,\dots,N$ όπου $y_i \in \{-1,+1\}$ και $x_i \in R^p$. Αναζητούμε μια συνάρτηση της μορφής $f(x)$ που θα υπολογίζει το y στο x:

$$f(x) = b + w^{\tau} x = b + \sum_{i=1}^{p} w_i - x_i, \qquad 2.11$$

όπου το $b + w^{\tau} x = 0$ είναι το υπερεπίπεδο μας (εδώ ευθεία) και το w είναι διάνυσμα

κάθετο στο υπερεπίπεδο (Σχήμα 2.5).



ΣΧΗΜΑ 2.5 Σχηματική αναπαράσταση μηχανής διανυσμάτων υποστήριξης

Τα διανύσματα υποστήριξης (support vectors) είναι τα σημεία που βρίσκονται πλησιέστερα στο υπερεπίπεδο.

Στόχος της μεθόδου είναι να επιλέξει τέτοια b και w έτσι ώστε τα δεδομένα να μπορούν να περιγραφούν από τις παρακάτω ανισότητες:

$$\begin{aligned} b+w^{\tau}x_i \le -1, &\quad \text{όταν } y_i=-1 \\ b+w^{\tau}x_i \ge +1, &\quad \text{όταν } y_i=+1 \end{aligned}. \qquad 2.12$$

Το παραπάνω ισοδύναμα γράφεται:

$$y_i\left(b+w^{\tau}x_i\right) \ge +1 \Rightarrow y_i\left(b+w^{\tau}x_i\right)-1 \ge 0. \qquad 2.13$$

Ενώ τα διανύσματα υποστήριξης μπορούν να περιγραφούν από δυο υπερεπίπεδα (στις περίπτωσή μας, ευθείες) τα:

$$\begin{matrix} b+w^{\tau}x_i=-1 \\ b+w^{\tau}x_i=+1 \end{matrix}. \qquad 2.14$$

Βάσει του παραπάνω, η απόσταση μεταξύ των δύο υπερεπιπέδων αποδεικνύεται ότι είναι ίση με $\frac{2}{\|w\|}$. Στόχος του αλγορίθμου είναι η μεγιστοποίηση αυτής της απόστασης, οδηγώντας στο πρόβλημα μεγιστοποίησης:

$$\max \frac{2}{\|w\|}, \ \text{με περιορισμό:}\ y_i\left(b+w^{\tau}x_i\right)-1\geq 0, \qquad 2.15$$

ή ισοδύναμα:

$$\min \frac{\|w\|}{2}, \ \text{με περιορισμό:}\ y_i\left(b+w^{\tau}x_i\right)-1\geq 0. \qquad 2.16$$

Όταν οι κλάσεις του προβλήματος είναι τέλεια διαχωρισμένες, η λύση αυτή λειτουργεί αποτελεσματικά (hard-margin SVM). Σε αρκετά προβλήματα του πραγματικού κόσμου, ωστόσο, οι κλάσεις των δεδομένων δεν είναι γραμμικά διαχωρίσιμες καθώς μερικές παρατηρήσεις βρίσκονται στην απέναντι μεριά του διαχωριστικού υπερεπιπέδου. Αυτό το πρόβλημα αντιμετωπίζεται με την εισαγωγή “χαλαρών μεταβλητών” για τα σημεία που βρίσκονται στην λάθος πλευρά του περιθωρίου ώστε να επιτυγχάνεται μερική χαλάρωση των περιορισμών μας (soft-margin SVM). Η διαδικασία υπολογισμού είναι παρόμοια με αυτήν της προηγούμενης περίπτωσης, με την επιπλέον εισαγωγή μιας παραμέτρου κόστους C, η οποία μπορεί να ρυθμιστεί έτσι ώστε το “μαλακό περιθώριο” να περιλαμβάνει έναν συγκεκριμένο αριθμό παρατηρήσεων, λαμβάνοντας ωστόσο υπόψιν ότι η χρήση του τεχνάσματος αυτού οδηγεί σε αντίστοιχη μείωση της απόδοσης ταξινόμησης.

Ενδεικτικά παρατίθεται σε μορφή ψευδοκώδικα ο αλγόριθμος hard-margin SVM για την επίλυση προβλημάτων ταξινόμησης:

Algorithm Support Vector Machine

Input: S dataset of classified observations (εισαγωγή των δεδομένων με τις κλάσεις τους)

Output: Weight vector w, dual solution α^*, margin γ^* and function f implementing the decision rule represented by the hyperplane (διάνυσμα βαρών, λύση διπλού πολλαπλασιαστή Lagrange, περιθώριο, συνάρτηση υπολογισμού)

Solve optimization problem $W(\alpha) = -\sum_{i,j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \kappa \cdot (x_i, x_j)$, subject to

$\sum_{i=1}^{l} y_i \cdot \alpha_i = 0, \sum_{i=1}^{l} \alpha_i = 1, 0 \leq \alpha_i, \; i = 1, \ldots, l$ (επίλυση του προβλήματος βελτιστοποίησης)

$\gamma^* = \sqrt{-W(\alpha^*)}$ (υπολογισμός του Υ που αντιστοιχεί στην βέλτιστη λύση)

Find all i **such that** $0 < \alpha_i^*$ (εύρεση i τέτοιου ώστε $0 < \alpha_i^*$)

For all i (έλεγχος συνθήκης κάθε i)

$b = y_i \cdot (\gamma^*)^2 - \sum_{j=1}^{l} \alpha_j^* \cdot y_j \cdot \kappa \cdot (x_j, x_i)$ (υπολογισμός του συντελεστή b)

end for

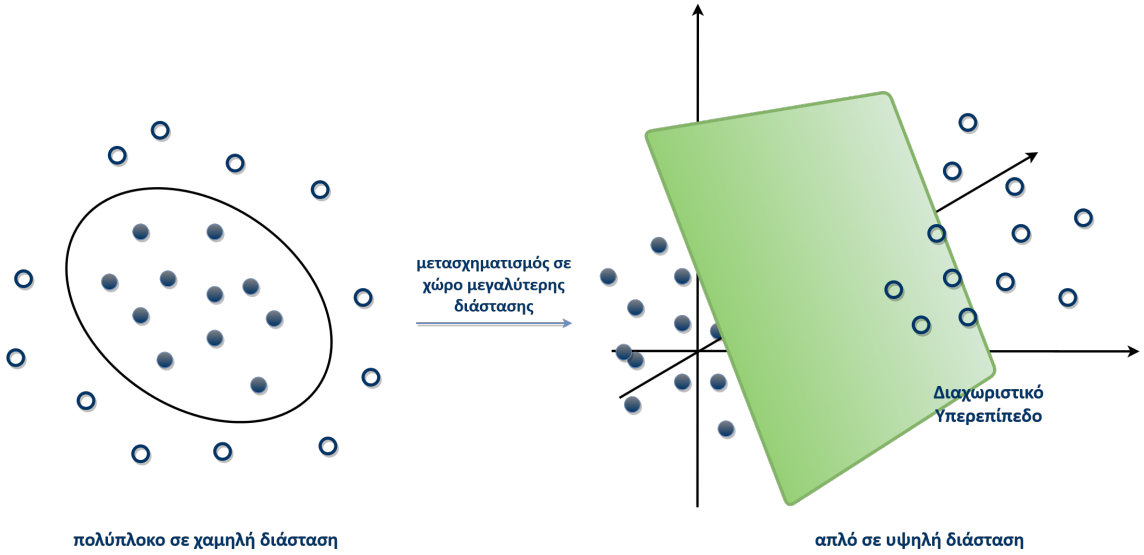
$f(\cdot) = \operatorname{sgn}\left(\sum_{j=1}^{l} \alpha_j^* \cdot y_j \cdot \kappa \cdot (x_j, \cdot) + b \right)$ (ταξινόμηση στοιχείων σε κλάση)

$W = \sum_{j=1}^{l} y_j \cdot \alpha_j^* \cdot \varphi(x_j)$ (υπολογισμός διανύσματος βαρών)

2.5.3 Μη Γραμμικώς Διαχωρίσιμα Προβλήματα

Παραπάνω παρουσιάστηκε η λειτουργία του SVM για γραμμικώς διαχωρίσιμα προβλήματα, και συγκεκριμένα για περιπτώσεις στις οποίες τα δεδομένα εκπαίδευσης είναι δυνατόν να διαχωριστούν γραμμικά από ένα βέλτιστο υπερεπίπεδο χωρίς λάθη (hard-margin SVM) ή έστω με κάποια λάθος ταξινομημένα δεδομένα (soft-margin SVM).

Οι Μηχανές Διανυσματικής Υποστήριξης μπορούν να επεκτείνουν τη χρήση τους και στην περίπτωση κλάσεων που δεν μπορούν καθόλου να διαχωριστούν γραμμικά. Σε αυτήν την περίπτωση, οι υπάρχουσες συντεταγμένες των σημείων αντιστοιχίζονται στο χώρο και συγκεκριμένα σε ένα χώρο μεγαλύτερης διάστασης, χρησιμοποιώντας μη γραμμικές συναρτήσεις. Ο χώρος στον οποίο αντιστοιχίζεται κάθε σημείο είναι χώρος υψηλού αριθμού διαστάσεων με το βασικό χαρακτηριστικό ότι δύο κλάσεις μπορούν να διαχωριστούν με ένα γραμμικό ταξινομητή όπως στην Υποενότητα 2.5.2 (Σχήμα 2.6).



ΣΧΗΜΑ 2.6 Μετασχηματισμός δεδομένων σε υψηλότερη διάσταση, όπου οι κλάσεις είναι γραμμικώς διαχωρίσιμες

Η διαδικασία μετασχηματισμού σε χώρο υψηλότερης διάστασης γίνεται με τη βοήθεια της λεγόμενης συνάρτησης πυρήνα, ενώ η διαδικασία καλείται kernel trick (κόλπο του πυρήνα). Στο σημείο αυτό είναι χρήσιμο να δοθεί ο ορισμός ενός εκ των πιο δημοφιλών πυρήνων, του Gaussian kernel, ο οποίος ορίζεται από την ακόλουθη σχέση:

$$k(x, x') = e^{(-\gamma \|x - x'\|^2)}, \qquad 2.17$$

όπου η k συνάρτηση πυρήνα, η οποία μετασχηματίζει ένα σημείο x σε x′ στο νέο χώρο και γ>0, μια παράμετρος η οποία ελέγχει το πλάτος της Gaussian κατανομής. Καθώς αυξάνει η τιμή του γ, αυξάνει η καμπυλότητα του ορίου απόφασης. Γενικά, η τιμή της παραμέτρου γ και ο βαθμός του πυρήνα (διάσταση πυρήνα) επηρεάζουν την ικανότητα

του SVM στην ταξινόμηση δεδομένων. Συγκεκριμένα πυρήνες μεγαλύτερης διάστασης επιτρέπουν ένα περισσότερο ευέλικτο όριο απόφασης με μικρότερη γραμμικότητα. Αντίστοιχα, αύξηση της παραμέτρου γ μειώνει την γραμμικότητα του ορίου απόφασης, ωστόσο αν δοθούν πολύ υψηλές τιμές μπορεί να οδηγηθούμε σε υπερεκπαίδευση.

Συνοψίζοντας, οι Μηχανές Διανυσματικής Υποστήριξης θεωρούνται ως μια από τις πιο αποτελεσματικές μεθόδους ταξινόμησης και μοντελοποίησης δεδομένων. Όπως προαναφέρθηκε ωστόσο, είναι σημαντική η γνώση του εκάστοτε προβλήματος προκειμένου να καταστεί δυνατή η βέλτιστη ρύθμιση των προαναφερθεισών παραμέτρων. Ακόμη, παρά τις υψηλές της επιδόσεις σε προβλήματα δύο κλάσεων (διαχωρίσιμα ή μη), η μέθοδος δεν παρουσιάζει παρόμοια επίδοση σε προβλήματα με περισσότερες από δύο κλάσεις, καθώς χρησιμοποιεί προσεγγιστικές μεθόδους για μείωση της πολυπλοκότητας που επιδρούν αρνητικά στην ακρίβεια της ταξινόμησης.

Στο σημείο αυτό ολοκληρώνεται η περιγραφή του μαθηματικού υποβάθρου και της αλγοριθμικής διαδικασίας των μεθόδων μηχανικής μάθησης που υλοποιήθηκαν προγραμματιστικά στα πλαίσια της μεταπτυχιακής διατριβής για την επίλυση ενός προβλήματος δυαδικής ταξινόμησης. Ο σχετικός κώδικας και τα παραγόμενα αποτελέσματα που προέκυψαν περιγράφονται αναλυτικά στα Κεφάλαια 5 και 6 αντίστοιχα.

ΚΕΦΑΛΑΙΟ 3

Εισαγωγή στην Θεωρία Πληροφορίας

3.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο αναπτύχθηκε το μαθηματικό και αλγοριθμικό υπόβαθρο των μεθόδων ταξινόμησης που θα εφαρμοστούν στα πλαίσια της παρούσας μεταπτυχιακής διατριβής. Ωστόσο, όπως παρατηρήθηκε κατά την μελέτη των σταδίων Εξόρυξης Δεδομένων στο Κεφάλαιο 1, η διαδικασία του μετασχηματισμού δεδομένων προηγείται της χρήσης τεχνικών μηχανικής μάθησης και είναι απαραίτητη προκείμενου να εξασφαλιστεί η αποδοτική εξόρυξη γνώσης από τα δεδομένα. Μια από τις βασικότερες μεθόδους μετασχηματισμού, που αποτελεί παράλληλα και σημαντικό πεδίο έρευνας, είναι η επιλογή χαρακτηριστικών (feature selection) με χρήση μέτρων πληροφορίας. Επομένως, σε αυτό το κεφάλαιο κρίθηκε σκόπιμη η παράθεση του βασικού μαθηματικού υποβάθρου της Θεωρίας της Πληροφορίας, στην οποία οφείλεται η ψηφιακή επανάσταση που οδήγησε στην ανάπτυξη και την εδραίωση νέων μέσων επικοινωνίας – μεταξύ των οποίων και το Internet. Η Θεωρία Πληροφορίας χρησιμοποιείται ευρέως για την επίλυση γρίφων σε γνωστικούς τομείς τόσο διαφορετικούς μεταξύ τους όπως η Πληροφορική, η Γενετική Μηχανική, τα Νευρωνικά Συστήματα, η Γλωσσολογία, η Φωνητική, η Ψυχολογία και τα Οικονομικά (Guiasu, 1977). Στο παρόν κεφάλαιο περιγράφονται οι βασικές έννοιες που διέπουν το συγκεκριμένο επιστημονικό πεδίο, και των οποίων η κατανόηση αποτελεί προαπαιτούμενο για την ανάλυση της μεθόδου επιλογής χαρακτηριστικών που περιγράφεται στο επόμενο κεφάλαιο.

Η Θεωρία Πληροφορίας είναι το τμήμα των εφαρμοσμένων μαθηματικών που ασχολείται με την έννοια της πληροφορίας, την ποσοτικοποίησή της, τα μέτρα και τις

εφαρμογές της. Πιο συγκεκριμένα, στα θέματα που απασχολούν τη Θεωρία Πληροφορίας συγκαταλέγονται η ποσοτικοποίηση της αβεβαιότητας στην πρόβλεψη της πληροφορίας (ή εντροπία) και οι μονάδες μέτρησης αυτής, η ροή της πληροφορίας, η κωδικοποίησή της, η συλλογή και η ανάκτησή της καθώς και η κατασκευή συστημάτων επεξεργασίας και μετάδοσής της. Τα όρια της Θεωρίας Πληροφοριών είναι αρκετά ασαφή. Η θεωρία αυτή επικαλύπτει κατά ένα μεγάλο μέρος τη Θεωρία Επικοινωνίας, όμως είναι περισσότερο προσανατολισμένη στην επεξεργασία και τη μετάδοση των πληροφοριών και λιγότερο στις λεπτομερείς λειτουργίες των συσκευών που χρησιμοποιούνται στα δίκτυα επικοινωνίας.

Η βασική έννοια της Θεωρίας της Πληροφορίας, είναι η ίδια η έννοια της πληροφορίας. Η λέξη πληροφορία αναφέρεται σαν μια αλληλουχία συμβόλων, που είτε καταγράφονται είτε μεταδίδονται και μπορεί να ερμηνευτεί ως ένα μήνυμα, το οποίο μεταφέρει κάποια γνώση για κάποιο αντικείμενο ή κάτι καινούργιο σχετικά με αυτό. Η πληροφορία για ένα γεγονός έχει άμεση σχέση με τη πιθανότητα του να συμβεί.

Αυτή η σύνδεση στην πραγματικότητα είναι πολύ λογική αφού η πληροφορία συνδέεται με την πιθανότητα μέσω της έννοιας της αβεβαιότητας. Όσο μικρότερη είναι η πιθανότητα P να συμβεί ένα γεγονός, τόση περισσότερη ποσότητα πληροφορίας συνοδεύει την πραγματοποίηση του. Και αντίστροφα, αν η πιθανότητα πραγματοποίησης ενός γεγονότος είναι μεγάλη, τότε η πληροφορία που μεταφέρει το γεγονός αυτό είναι μικρή. Για παράδειγμα, αν κάποιος πει πως “Θα περάσει ένα μπλε αυτοκίνητο τώρα.”, το μήνυμα αυτό, έχει μεγάλη πληροφορία, γιατί είναι ένα αβέβαιο γεγονός. Αν όμως πει κάποιος πως “Θα περάσει ένα μπλε αυτοκίνητο μέσα στην μέρα.”, τότε το κείμενο αυτό έχει πολύ μικρή πληροφορία. Γιατί στο μήνυμα αυτό η πιθανότητα να περάσει κάποια στιγμή ένα μπλε αυτοκίνητο είναι πολύ μεγάλη, ίσως αγγίζει και το 100%.

Για τη γενική περίπτωση, ορίζεται η μέση πληροφορία $H(A)$ που μπορεί να φέρει ένα πιθανοθεωρητικό πείραμα A, σε ένα δειγματικό χώρο X, να ισούται με (Shannon, 1948):

$$H(A) = -\sum_{i=1}^{n} p_i \cdot \log(p_i), \qquad 3.1$$

όπου με p_i συμβολίζεται η πιθανότητα του ενδεχομένου $x_i \in X$.

Ο λόγος που επιλέχθηκε η λογαριθμική συνάρτηση στον ορισμό της ποσότητας πληροφορίας, είναι ότι πληροί τη σχέση:

$$f(x^y) = y \cdot f(x), \qquad 3.2$$

και έτσι ανταποκρίνεται στη διαίσθησή μας ότι η ποσότητα πληροφορίας ενός μηνύματος αποτελούμενου από κ σύμβολα θα πρέπει να είναι κ φορές μεγαλύτερη από αυτή ενός μηνύματος που αποτελείται από 1 σύμβολο.

Η έννοια της πληροφορίας, βέβαια, είναι πολύ ευρεία για να καλυφθεί πλήρως από έναν και μόνο ορισμό. Για αυτό ακριβώς, αναπτύχθηκαν συγκεκριμένα μέτρα τα οποία είναι υπεύθυνα για την μέτρηση αυτής. Σε αυτό το κεφάλαιο εισάγονται οι περισσότεροι από τους βασικούς ορισμούς που απαιτούνται για την ανάπτυξη της Θεωρίας Πληροφοριών. Αφού διατυπωθούν, θα αναλυθούν εκτενέστερα οι μεταξύ τους σχέσεις και ερμηνείες.

Για κάθε κατανομή πιθανότητας θα οριστεί μια ποσότητα που ονομάζεται εντροπία. Η εντροπία αποτιμά τον μέσο όρο πληροφορίας που φέρει μια τυχαία μεταβλητή, και έχει πολλές ιδιότητες που συμφωνούν με όσα θα αναμέναμε διαισθητικά από ένα μέτρο πληροφορίας. Επεκτείνοντας αυτή την έννοια, θα οριστεί και η έννοια της αμοιβαίας πληροφορίας, η οποία είναι ένα μέτρο της ποσότητας πληροφορίας που περιέχει μια τυχαία μεταβλητή σχετικά με κάποια άλλη. Υπό αυτό το πρίσμα, η εντροπία είναι η αυτοπληροφορία μιας τυχαίας μεταβλητής. Η αμοιβαία πληροφορία είναι ειδική περίπτωση μιας γενικότερης ποσότητας που ονομάζεται σχετική εντροπία, η οποία είναι ένα μέτρο της απόστασης μεταξύ δύο κατανομών πιθανότητας. Θα οριστούν και μερικές άλλες σχετικές έννοιες. Όλες αυτές οι ποσότητες σχετίζονται στενά μεταξύ τους, και έχουν ορισμένες απλές κοινές ιδιότητες, μερικές από τις οποίες θα αποδειχτούν σε αυτό το κεφάλαιο.

3.2 Ορισμοί

Τα κυριότερα μέτρα πληροφορίας είναι τα παρακάτω (Κουκουβίνος, 2003) (Ζορκάδης, 2002):

1. Η **εντροπία** (entropy), η οποία μετράει την μέση πληροφορία που φέρει μια τυχαία μεταβλητή X.
2. Η **σχετική εντροπία** (relative entropy), η οποία μετράει την ομοιότητα των X και Y.
3. Η **κοινή εντροπία** (joint entropy), η οποία μετράει τη συνολική πληροφορία των X και Y.
4. Η **δεσμευμένη** ή **υπό συνθήκη εντροπία** (conditional entropy), η οποία μετράει την πληροφορία του X, όταν η Y είναι γνωστή και αντιστρόφως.
5. Η **αμοιβαία πληροφορία** ή **διαπληροφορία** (mutual information), η οποία μετρά την μείωση της αβεβαιότητας για το X, όταν είναι γνωστή η Y μεταβλητή.
6. Η **υπο συνθήκη αμοιβαία πληροφορία** (conditional mutual information), η οποία μετρά την αναμενόμενη αμοιβαία πληροφορία μεταξύ δύο μεταβλητών X, Y όταν είναι γνωστή μια τρίτη μεταβλητή Z.
7. Η **υπο συνθήκη σχετική εντροπία** (conditional relative entropy), η οποία μετράει το σταθμισμένο άθροισμα των σχετικών εντροπιών των δεσμευμένων κατανομών των X , Y για τις διάφορες τιμές του Y.

Πιο κάτω παρουσιάζονται αναλυτικά κάθε μια από αυτές τις έννοιες.

3.2.1 Εντροπία

Η πρώτη και ίσως και η δυσκολότερη και πιο αφηρημένη έννοια που ορίστηκε είναι αυτή της εντροπίας. Η λέξη εντροπία είναι σύνθετη και προέρχεται από τις λέξεις “εν” και “τροπή” και ουσιαστικά σημαίνει εσωτερική αλλαγή ή αλλαγή εντός ενός συστήματος. Ενώ ο όρος συναντάται για πρώτη φορά στο επιστημονικό πεδίο της Θερμοδυναμικής, το 1948 ο Shannon, μέσω της γνωστής, πλέον εργασίας του “A Mathematical Theory of Communication”, καταφέρνει να ποσοτικοποιήσει την πληροφορία και αποφασίζει να ονομάσει την ποσότητα που μετρούσε την πληροφορία εντροπία.

Η εντροπία στη Θεωρία Πληροφοριών είναι στην ουσία το μέτρο αβεβαιότητας που

διακατέχει το σύστημα. Ο Shannon είδε πως όσος λιγότερος θόρυβος παράγεται σε ένα μοντέλο επικοινωνίας, πομπού και δέκτη, τόση περισσότερη πληροφορία μεταδίδει. Και αντιστρόφως, όσο αυξάνεται η αταξία (θόρυβος) ενός συστήματος τόσο λιγότερη πληροφορία μεταδίδει αυτό. Από αυτό απορρέει το συμπέρασμα ότι η πληροφορία του συστήματος αποτελεί μέτρο της εσωτερικής του τάξης (δηλ. αντιστρόφως ανάλογη με την αταξία). Αλλά η εντροπία είναι το μέτρο της αταξίας ενός συστήματος, άρα η πληροφορία είναι αντιστρόφως ανάλογη της εντροπίας. Αυτός είναι βασικά και ο λόγος που συχνά αναφέρεται η πληροφορία A σαν η αρνητική εντροπία H, δηλαδή ισχύει ότι $A=-H$. Άρα προκύπτει ο εξής ορισμός για την εντροπία:

Αν X είναι μια διακριτή τυχαία μεταβλητή με δειγματικό χώρο $X=\{x_1, x_2, \ldots, x_n\}$ και συνάρτηση μάζας πιθανότητας p_i, με $p_i>0$ και $\sum_{i=1}^{n} p_i=1$, τότε η μέση ποσότητα πληροφορίας (ή μέση πληροφορία) της X, $H(X)$, δίνεται από τη σχέση:

$$H_b(X)=-\sum_{i=1}^{n} p_i \cdot \log_b(p_i). \qquad 3.3$$

Η μέση πληροφορία ονομάζεται διαφορετικά και εντροπία (όπως αναφέρθηκε και προηγουμένως).

Η βάση b του λογαρίθμου, συνήθως λαμβάνει τη τιμή 2, και η εντροπία σε αυτήν την περίπτωση μετριέται σε bits, ενώ όταν η βάση b του λογαρίθμου ισούται με e, τότε η εντροπία μετριέται σε nats. Επιπλέον, ισχύει η σύμβαση ότι $0 \cdot \log(0)=0$, η οποία αποδεικνύεται και από τον ορισμό της συνέχειας, από τον οποίο ισχύει ότι η ποσότητα $x \cdot \log(x) \to 0$ καθώς το $x \to 0$. Σημειώνεται ότι η εντροπία ορίζεται συναρτήσει της κατανομής της τυχαίας μεταβλητής X. Δεν εξαρτάται όμως από τις πραγματικές τιμές που παίρνει η μεταβλητή X, αλλά μόνο από τις πιθανότητες που έχουν οι τιμές αυτές.

Όπως μπορεί να συνάγει κάποιος και από τον ορισμό της εντροπίας, η ποσότητα πληροφορίας (ή το πληροφορικό περιεχόμενο) ενός γεγονότος x_i της τυχαίας μεταβλητής X είναι ίση με την αρνητική ποσότητα του λογαρίθμου της πιθανότητας εμφάνισής του p_i, δηλαδή ίση με $(-\log(p_i))$. Επομένως, η ποσότητα πληροφορίας ενός γεγονότος x_i είναι αντιστρόφως ανάλογη της πιθανότητας εμφάνισής του.

Οι ιδιότητες της μέσης (ποσότητας) πληροφορίας, που παράλληλα έχουν τεθεί και ως απαιτήσεις κατά τον ορισμό της, δηλαδή κατά την αναζήτηση από τον Shannon και άλλους ερευνητές της κατάλληλης συνάρτησης, διακρίνονται στις πέντε ακόλουθες:

i. Το ποσό της πληροφορίας σε ένα γεγονός x εξαρτάται μόνον από την πιθανότητά του. Αυτή είναι μία φυσική απαίτηση, μιας και όσο πιο απίθανο είναι ένα γεγονός να πραγματοποιηθεί, τόσο περισσότερη πληροφορία περιέχει.

ii. Η μέση πληροφορία $H(X)$ είναι συνεχής ως προς το p.

iii. Η εντροπία είναι προσθετική. Η ιδιότητα αυτή αναφέρεται στην περίπτωση κατά την οποία δύο τυχαίες μεταβλητές X και Y, οι οποίες είναι ανεξάρτητες μεταξύ τους, συνδυάζονται. Τότε, για τη συνδυασμένη ποσότητα πληροφορίας ισχύει $H(X,Y)=H(X)+H(Y)$.

iv. Η εντροπία $H(X)$ παίρνει τη μέγιστη τιμή της όταν όλα τα ενδεχόμενα είναι ισοπίθανα. Τότε, η αβεβαιότητα είναι η μέγιστη δυνατή και, κατά συνέπεια, η επιλογή ενός μηνύματος προσφέρει τη μέγιστη δυνατή μέση πληροφορία. Αντίθετα, η $H(X)$, γίνεται ελάχιστη, όταν ένα ενδεχόμενο έχει πιθανότητα ίση με 1.

v. Η μέση πληροφορία $H(X)$ είναι συμμετρική, δηλαδή η διάταξη των πιθανοτήτων δεν την επηρεάζει. Έτσι, διαφορετικές τυχαίες μεταβλητές με κατανομές πιθανοτήτων που προέρχονται από μεταθέσεις της ίδιας κατανομής πιθανοτήτων έχουν ίση εντροπία. Σε ορισμένες περιπτώσεις, ακόμα και διαφορετικές κατανομές πιθανοτήτων οδηγούν στην ίδια μέση ποσότητα πληροφορίας.

Ένας σχετικός ορισμός με αυτό της εντροπίας, είναι αυτός της προσδοκίας (expectation). Η προσδοκία (αναμενόμενη τιμή) συμβολίζεται με E, και μετρά τη προσδοκώμενη τιμή της τυχαίας μεταβλητής $g(X)$, όταν η $X\sim p(x)$. Η τιμή της δίνεται από την σχέση:

$$E_p g(x)=\sum_{x\in X} g(x)\cdot p(x), \qquad 3.4$$

και απλούστερα γράφεται και ως $E_g(X)$, όταν η συνάρτηση μάζας πιθανότητας εννοείται από τα συμφραζόμενα. Ιδιαίτερο ενδιαφέρον, παρουσιάζει επίσης η αυτοαναφορική αναμενόμενη τιμή της $g(X)$ ως προς την $p(x)$ όταν αυτές οι δύο συνδέονται με τη σχέση $g(X)=\log\left(\frac{1}{p(X)}\right)$. Σε αυτήν την περίπτωση η εντροπία της X μπορεί επίσης να ερμηνευτεί

ως η αναμενόμενη τιμή της τυχαίας μεταβλητής $\log(\frac{1}{p(X)})$, όπου η X λαμβάνεται σύμφωνα με τη συνάρτηση μάζας πιθανότητας $p(x)$. Συνεπώς,

$$H(X)=E_p \log(\frac{1}{p(X)}). \qquad 3.5$$

Για τον ορισμό της εντροπίας πρέπει να καθοριστούν και οι δύο ιδιότητες που είναι ικανές και αναγκαίες συνθήκες για να να είναι ορθός.

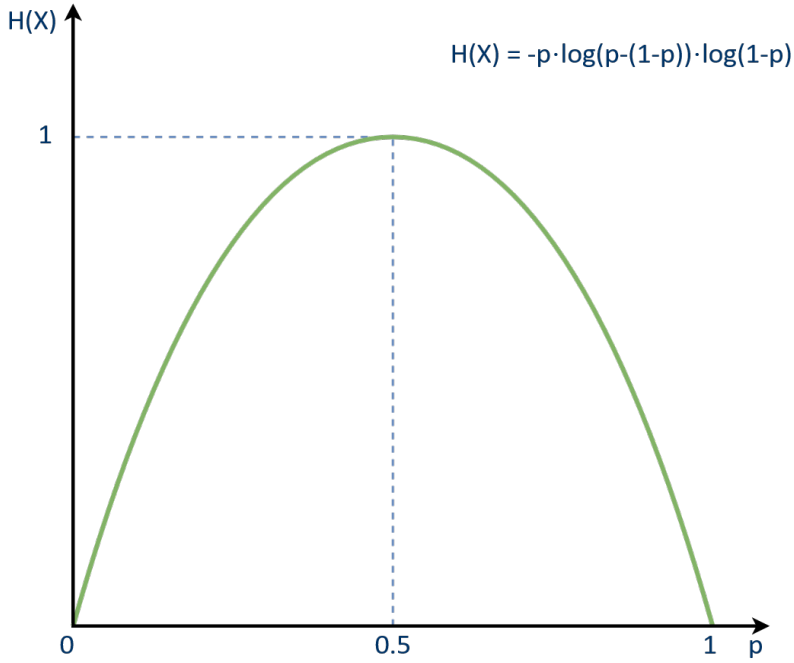
- Η εντροπία παίρνει πάντα θετικές τιμές, δηλαδή $H(X)\geq 0$.
- Για την εντροπία πρέπει να ισχύει ότι $H_b(X)=\log_b(a)\cdot H_a(X)$.

Η πρώτη ιδιότητα προκύπτει άμεσα από τους ορισμούς της πιθανότητας και του λογαρίθμου, βάση των οποίων απαιτούνται ότι $0\leq p(x)\leq 1$ και $\log(\frac{1}{p(x)})\geq 0$, ενώ η δεύτερη ιδιότητα αποδεικνύεται άμεσα από τη γνωστή σχέση που ισχύει για τους λογαρίθμους, $\log_b(p)=\log_b(a)\cdot\log_a(p)$ η οποία στην ουσία περιγράφει ότι είναι επιτρεπτό να αλλάξουμε τη βάση του λογαρίθμου για την εντροπία, απλά πολλαπλασιάζοντας τη σχέση με τον κατάλληλο συντελεστή.

Στην απλούστερη περίπτωση μιας διακριτής τυχαίας μεταβλητής X, η οποία έχει μόνο δύο ενδεχόμενα, (π.χ. εκπομπή ενός από δύο δυνατά μηνύματα) και οι πιθανότητες αυτών ισούται με p και $(1-p)$, αντίστοιχα, η εξίσωση της εντροπίας θα είναι:

$$H(X)=-p\cdot\log(p)-(1-p)\cdot\log(1-p). \qquad 3.6$$

Στο Σχήμα 3.1 που ακολουθεί φαίνεται η γραφική παράσταση της συμπεριφοράς της μέσης ποσότητας πληροφορίας ως συνάρτηση της πιθανότητας p. (Η μονάδα μέτρησης της μέσης ποσότητας πληροφορίας είναι το bit, δηλαδή ο λογάριθμος είναι με βάση το 2.)



ΣΧΗΜΑ 3.1 Η μέση ποσότητα πληροφορίας ως συνάρτηση της p

Είναι φανερό από τη γραφική παράσταση (Σχήμα 3.1) ότι η μέση πληροφορία παίρνει τη μέγιστη τιμή της, που ισούται με ένα, όταν τα δύο γεγονότα είναι ισοπίθανα, δηλαδή έχουν την ίδια πιθανότητα, $p=\frac{1}{2}$, να συμβούν. Από την άλλη πλευρά, αν $p=1$ ή $p=0$, τότε η εντροπία είναι 0, αφού το τελικό αποτέλεσμα (η έκβαση του πειράματος) είναι βέβαιο.

Σε ένα άλλο λίγο πιο πολύπλοκο παράδειγμα όπου, έστω ότι έχουμε έναν αγώνα αυτοκινήτων. Σε αυτόν συμμετέχουν οκτώ αυτοκίνητα με αντίστοιχες πιθανότητες νίκης για το καθένα από αυτά $(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64})$. Τότε η εντροπία του αγώνα θα υπολογιζόταν ως εξής:

$$H(X)=-\frac{1}{2}\cdot\log(\frac{1}{2})-\frac{1}{4}\cdot\log(\frac{1}{4})-\frac{1}{8}\cdot\log(\frac{1}{8})-\frac{1}{16}\cdot\log(\frac{1}{16})-\frac{1}{64}\cdot\log(\frac{1}{64})$$
$$-\frac{1}{64}\cdot\log(\frac{1}{64})-\frac{1}{64}\cdot\log(\frac{1}{64})-\frac{1}{64}\cdot\log(\frac{1}{64})=2\,\text{bits}.$$

3.2.2 Σχετική Εντροπία

Η σχετική εντροπία ή αλλιώς απόκλιση κατά Kullback–Leibler, είναι ένα μέτρο της απόστασης μεταξύ δύο κατανομών. Στη στατιστική, εμφανίζεται ως αναμενόμενος λογάριθμος του λόγου πιθανοφανειών. Η σχετική εντροπία $D(p||q)$ είναι ένα μέτρο του πόσο άστοχο είναι να θεωρηθεί ότι η κατανομή που περιγράφει την τυχαία μεταβλητή είναι η q όταν η πραγματική της κατανομή είναι η p. Για παράδειγμα, αν γνωρίζαμε την πραγματική κατανομή p της τυχαίας μεταβλητής, θα μπορούσαμε να κατασκευάσουμε έναν κώδικα με μέσο μήκος περιγραφής $H(p)$. Αν αντ' αυτού χρησιμοποιούσαμε τον κώδικα για μια κατανομή q, θα χρειαζόμασταν κατά μέσο όρο $H(p)+D(p||q)$ για να περιγράψουμε την τυχαία μεταβλητή.

Συγκεκριμένα, η σχετική εντροπία $D(p||q)$, μεταξύ δύο συναρτήσεων μάζας πιθανότητας $p(x)$ και $q(x)$ ισούται με:

$$\begin{aligned} D(p||q) &= \sum_{x\in X} p(x)\cdot \log(q(x)) + H(X) \\ &= \sum_{x\in X} p(x)\cdot \log\left(\frac{p(x)}{q(x)}\right) \\ &= E_p \log\left(\frac{p(X)}{q(X)}\right). \end{aligned} \tag{3.7}$$

Η εντροπία Shannon είναι μια τυχαία ειδική περίπτωση της σχετικής εντροπίας. Πράγματι η εντροπία Shannon, μιας τυχαίας μεταβλητής, είναι η σχετική εντροπία ως προς μια κατάσταση που είναι γνωστή με απόλυτη βεβαιότητα (μεταβλητή Y), δηλαδή $H(X)=H(X|Y)$, όπου $P(Y=y)=1$, για κάποια τιμή του Y.

Η σχετική εντροπία είναι πάντα μη αρνητική και ισούται με μηδέν αν και μόνο αν ισχύει ότι $p=q$. Παρόλο που συγκαταλέγεται στη σχετική λίστα των μέτρων πληροφορίας, η σχετική εντροπία, δεν είναι ένα πραγματικό μέτρο. Αυτό γιατί δεν είναι συμμετρική, δηλαδή η διαφορά του p από το q δεν ισούται με τη διαφορά του q από το p και επιπλέον δεν ικανοποιεί την τριγωνική ανισότητα. Είναι χρήσιμο κάποιος να την αντιληφθεί καλύτερα σαν μια "απόσταση" μεταξύ δύο κατανομών παρόλο που δεν είναι μια πραγματική απόσταση.

3.2.3 Κοινή Εντροπία

Πολλές φορές έχει ενδιαφέρον να εξεταστεί η ποσότητα πληροφορίας ενός συνδυασμού δύο τυχαίων μεταβλητών, δηλαδή ενός πειράματος το οποίο αποτελείται από δύο υποπειράματα. Έστω ένα τυχαίο πείραμα (X,Y) έχει ως δυνατά αποτελέσματα όλους τους δυνατούς συνδυασμούς των αποτελεσμάτων των δύο υποπειραμάτων του $X=\{x_1,x_2,\dots,x_n\}$ και $Y=\{y_1,y_2,\dots,y_m\}$, επομένως έχει το δειγματοχώρο:

$$(X,Y)=\{(x_1,y_1),(x_1,y_2),\dots,(x_1,y_m),\dots,(x_n,y_1),(x_n,y_2),\dots,(x_n,y_m)\}. \quad 3.8$$

Η κατανομή πιθανοτήτων του δίνεται από:

$$P=\{p(x_1,y_1),p(x_1,y_2),\dots,p(x_1,y_m),\dots,p(x_n,y_1),p(x_n,y_2),\dots,p(x_n,y_m)\}. \quad 3.9$$

Αν (X,Y) είναι ένα τυχαίο πείραμα με δισδιάστατο δειγματοχώρο και κατανομή πιθανοτήτων αυτή που αναφέρεται παραπάνω, τότε η συνδυασμένη πληροφορία του $H(X,Y)$ ορίζεται ως η μέση τιμή:

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_i)\cdot\log(p(x_i,y_i)), \quad 3.10$$

το οποίο μπορεί να γραφτεί και με την αναμενόμενη τιμή Ε ως:

$$H(X,Y)=-E_p\log(p(X,Y)). \quad 3.11$$

Όταν οι κατανομές είναι ανεξάρτητες μεταξύ τους, τότε ισχύει ότι:

$$H(X,Y)=H(X)+H(Y). \quad 3.12$$

Αφού οι δύο τυχαίες μεταβλητές είναι ανεξάρτητες, τότε ισχύει επίσης για την πιθανότητα της τομής τους ότι $p_{ij}=p(x_i)\cdot p(y_j)$ και έτσι προκύπτει:

$$\begin{aligned}H(X,Y)&=-\sum_{i=1}^{n}\sum_{j=1}^{m}p_{ij}\cdot\log(p_{ij})\\&=-\sum_{i=1}^{n}\sum_{j=1}^{m}p_i\cdot p_j\cdot\log(p_i\cdot p_j)\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i=1}^{n} p_i \sum_{j=1}^{m} (\log(p_i) + \log(p_j)) \\
&= -\sum_{i=1}^{n} p_i \sum_{j=1}^{m} (p_j \cdot \log(p_i)) - \sum_{i=1}^{n} p_i \sum_{j=1}^{m} (p_j \cdot \log(p_j)) \\
&= -\sum_{i=1}^{n} p_i \cdot \log(p_i) \sum_{j=1}^{m} p_j - \sum_{i=1}^{n} p_i \sum_{j=1}^{m} (p_j \cdot \log(p_j)) \\
&= -\sum_{i=1}^{n} p_i \cdot \log(p_i) - \sum_{j=1}^{m} (p_j \cdot \log(p_j)) \\
&= H(X) + H(Y). \qquad 3.13
\end{aligned}
$$

Ο ορισμός της μέσης ποσότητας πληροφορίας $H(X,Y)$ μπορεί να επεκταθεί και για περισσότερες από δύο διαστάσεις, δηλαδή $H(X_{1,}\ldots,X_n)$. Σε κάθε περίπτωση λαμβάνονται υπόψη όλοι οι δυνατοί συνδυασμοί αποτελεσμάτων και, εφόσον είναι γνωστές οι πιθανότητες αυτών, μπορεί να υπολογιστεί η συνδυασμένη ποσότητα πληροφορίας. Άρα για n μεταβλητές, έστω $X_1,\ldots,X_n$, η κοινή εντροπία ισούται με:

$$
\begin{aligned}
H(X_{1,}\ldots,X_n) &= -\sum_{X_1} \ldots \sum_{X_n} p(x_{1,}\ldots,x_n) \cdot \log(p(x_1,\ldots,x_n)) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1},\ldots,X_1). \qquad 3.14
\end{aligned}
$$

Αυτό αποδεικνύεται ως εξής:

$$
\begin{aligned}
H(X_1,X_2,\ldots X_n) &= -\sum_{x_1,x_2,\ldots x_n} p(x_1,x_2,\ldots,x_n) \cdot \log(p(x_1,x_2,\ldots,x_n)) \\
&= -\sum_{x_1,x_2,\ldots x_n} p(x_1,x_2,\ldots,x_n) \cdot \log\left(\prod_{i=1}^{n} p(x_i|x_{i-1},\ldots,x_1)\right) \\
&= -\sum_{x_1,x_2,\ldots x_n} \sum_{i=1}^{n} p(x_1,x_2,\ldots,x_n) \cdot \log(p(x_i|x_{i-1},\ldots,x_1)) \\
&= -\sum_{i=1}^{n} \sum_{x_1,x_2,\ldots x_n} p(x_1,x_2,\ldots,x_n) \cdot \log(p(x_i|x_{i-1},\ldots,x_1)) \\
&= -\sum_{i=1}^{n} \sum_{x_1,x_2,\ldots x_i} p(x_1,x_2,\ldots,x_i) \cdot \log(p(x_i|x_{i-1},\ldots,x_1)) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1},\ldots,X_1), \qquad 3.15
\end{aligned}
$$

όπου με $p(x_1,\ldots,x_n)$ συμβολίζεται η ποσότητα $p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1},\ldots,x_1)$. Όσον

αφορά το δεύτερο μέρος του τύπου εφαρμόζεται κατ' επανάληψη ο κανόνας για το ανάπτυγμα της εντροπίας για δύο μεταβλητές, οπότε προκύπτει ότι:

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1) \tag{3.16}$$

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \end{aligned} \tag{3.17}$$

$$\vdots$$

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &= H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_{n-1}, \dots, X_1) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots X_1). \end{aligned} \tag{3.18}$$

3.2.4 Δεσμευμένη Εντροπία

Αρκετές φορές, αναλόγως το πρόβλημα, μπορεί να έχει ενδιαφέρον να υπολογιστεί η ποσότητα πληροφορίας μιας τυχαίας μεταβλητής, X, όταν είναι γνωστό το αποτέλεσμα μιας δεύτερης τυχαίας μεταβλητής, Y, θεωρώντας πάντα δεδομένο ότι για τις δύο μεταβλητές ισχύει $(X, Y) \sim p(x, y)$. Η ποσότητα αυτή, καλείται είτε δεσμευμένη ή υπό συνθήκη ποσότητα πληροφορίας της μεταβλητής X ως προς την μεταβλητή Y και συμβολίζεται με $H(X|Y)$. Με άλλα λόγια η υπό συνθήκη εντροπία μετρά την αβεβαιότητα της τυχαίας μεταβλητής X όταν είναι γνωστή η Y μεταβλητή. Η αναλυτική της μορφή είναι:

$$\begin{aligned} H(X|Y) &= \sum_{y \in Y} p(y) \cdot H(X|Y = y) \\ &= -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \cdot \log(p(x|y)) \\ &= -\sum_{y \in Y} \sum_{x \in X} p(y, x) \cdot \log(p(x|y)) \\ &= -\sum_{y \in Y, x \in X} p(y, x) \cdot \log(p(x|y)) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{y\in Y, x\in X} p(y,x)\cdot \log\left(\frac{p(y,x)}{p(y)}\right) \\
&= \sum_{y\in Y, x\in X} p(y,x)\cdot \log\left(\frac{p(y)}{p(y,x)}\right).
\end{aligned}
\tag{3.19}
$$

Εναλλακτικά γράφεται και ως:

$$
\begin{aligned}
H(X|Y) &= \sum_{y\in Y} p(y)\cdot H(X|Y=y) \\
&= -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\cdot \log\left(p(x|y)\right) \\
&= -\sum_{y\in Y}\sum_{x\in X} p(y,x)\cdot \log\left(p(x|y)\right) \\
&= -E_p \log\left(p(X|Y)\right).
\end{aligned}
\tag{3.20}
$$

Στην περίπτωση όπου, η μία τυχαία μεταβλητή, X, καθορίζεται πλήρως μέσω της άλλης τυχαίας μεταβλητής, Y, ισχύει ότι η υπό συνθήκη ποσότητα πληροφορίας της πρώτης μεταβλητής, X, ως προς την δεύτερη μεταβλητή, Y, ισούται με μηδέν, δηλαδή, $H(X|Y)=0$. Αντίστοιχα, στην περίπτωση που οι δυο τυχαίες μεταβλητές, X, Y είναι εντελώς ανεξάρτητες μεταξύ τους, ισχύει για την δεσμευμένη ποσότητα πληροφορίας ότι $H(X|Y)=H(X)$, δηλαδή είναι ίση με την εντροπία της τυχαίας μεταβλητής X. Αυτές οι δύο ισότητες αποδεικνύονται σχεδόν άμεσα και από τον τύπο της δεσμευμένης εντροπίας, αλλά είναι και διαισθητικά ορθές με βάση αυτό που θα περίμενε κανείς από τον ορισμό της.

Αναφέρεται σε αυτό το σημείο, ότι η φυσικότητα του ορισμού της από κοινού εντροπίας και της δεσμευμένης εντροπίας αποκαλύπτεται από το γεγονός ότι η εντροπία ενός ζεύγους τυχαίων μεταβλητών, έστω X,Y, είναι η εντροπία της μιας μεταβλητής, Y, προστιθέμενη με τη δεσμευμένη εντροπία αυτής της μεταβλητής, Y, ως προς την μεταβλητή X. Δηλαδή:

$$
H(Y,X) = H(Y) + H(X|Y), \tag{3.21}
$$

και λόγω συμμετρίας ισχύει επίσης ότι:

$$
H(X,Y) = H(X) + H(Y|X). \tag{3.22}
$$

Αυτό αποδεικνύεται ακολούθως:

$$\begin{aligned} H(Y,X) &= -\sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(y,x)) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(y)\cdot p(x|y)) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(y)) - \sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(p(x|y)) \\ &= -\sum_{y\in Y} p(y)\cdot\log(p(y)) - \sum_{x\in X}\sum_{y\in Y} p(y,x)\cdot\log(x|y) \\ &= H(Y)+H(X|Y). \end{aligned} \quad 3.23$$

Ισοδύναμα, μπορεί να διατυπωθεί:

$$\log(p(Y,X)) = \log(p(Y)) + \log(p(X|Y)). \quad 3.24$$

3.2.5 Αμοιβαία Πληροφορία

Η αμοιβαία πληροφορία $I(X;Y)$ είναι ένα μέγεθος που μετράει την ποσότητα της πληροφορίας που μια τυχαία μεταβλητή περιέχει για μια άλλη τυχαία μεταβλητή. Με άλλα λόγια, υπολογίζει σε ποιον βαθμό μπορεί η γνώση που σημειώνεται για την δεύτερη μεταβλητή να μειώσει την αβεβαιότητα που υπάρχει για την πρώτη μεταβλητή. Το μέτρο αυτό ουσιαστικά, βασίζεται στην αμοιβαία εξάρτηση που υπάρχει μεταξύ των δύο μεταβλητών. Πιο συγκεκριμένα, έστω ότι δίνονται δύο τυχαίες μεταβλητές X, Y, για τις οποίες η από κοινού σ.μ.π. $p(x,y)$ είναι γνωστή, καθώς επίσης είναι γνωστές και οι αντίστοιχες περιθώριες κατανομές πιθανότητας των X, Y, $p(x)$ και $p(y)$. Τότε η ποσότητα της αμοιβαίας πληροφορίας $I(X;Y)$ αυτών των δύο μεταβλητών θα είναι ίση με τη σχετική εντροπία μεταξύ της από κοινού σ.μ.π. $p(x,y)$ και των περιθώριων κατανομών πιθανότητας $p(x)$ και $p(y)$, δηλαδή:

$$\begin{aligned} I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log\left(\frac{p(x|y)}{p(x)}\right) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(p(x)) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(p(x|y)) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x \in X} p(x) \cdot \log(p(x)) - (-\sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \log(p(x|y))) \\
&= H(X) - H(X|Y). \qquad 3.25
\end{aligned}
$$

Λόγω συμμετρίας έπεται επίσης ότι:

$$I(X;Y) = H(Y) - H(Y|X), \qquad 3.26$$

Άρα, όταν είναι γνωστή η πρώτη μεταβλητή X, δίνει για την άγνωστη δεύτερη μεταβλητή Y τόση πληροφορία, όση δίνει αντίστροφα και η δεύτερη μεταβλητή Y όταν είναι γνωστή, για την πρώτη άγνωστη μεταβλητή X. Ένας διαφορετικός τρόπος αποτύπωσης της αμοιβαίας πληροφορίας είναι ο ακόλουθος:

$$
\begin{aligned}
I(X;Y) &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \log\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right) \\
&= D(p(x,y) \| p(x) \cdot p(y)) \\
&= E_{p(X,Y)} \log\left(\frac{p(X,Y)}{p(X) \cdot p(Y)}\right). \qquad 3.27
\end{aligned}
$$

Όπως αποδείχτηκε σε παραπάνω ενότητα (Ενότητα 3.2.4), για την κοινή εντροπία δύο τυχαίων μεταβλητών X, Y ισχύει ότι $H(X,Y) = H(X) + H(Y|X)$, επομένως μετά από πράξεις προκύπτει και η εξής σχέση:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \qquad 3.28$$

Τέλος μπορεί κάποιος να παρατηρήσει ότι:

$$I(X;X) = H(X) - H(X|X) = H(X). \qquad 3.29$$

Άρα συμπεραίνεται ότι η αμοιβαία πληροφορία που έχει οποιαδήποτε τυχαία μεταβλητή με τον ίδιο της τον εαυτό, ισούται στην πραγματικότητα με την ίδια την εντροπία της τυχαίας μεταβλητής. Αυτός είναι άλλωστε και ο λόγος, για τον οποίο μερικές φορές η εντροπία μιας τυχαίας μεταβλητής αποκαλείται και αυτοπληροφορία. Μπορεί επίσης να γίνει εμφανές, πως η αμοιβαία ποσότητα πληροφορίας που έχουν δύο ανεξάρτητες τυχαίες μεταβλητές X, Y είναι ίση με μηδέν, δηλαδή $I(X;Y) = 0$. Αυτό είναι ένα συμπέρασμα που προκύπτει άμεσα από το γεγονός ότι όταν δύο μεταβλητές X, Y είναι ανεξάρτητες μεταξύ τους, τότε για την από κοινού σ.μ.π. ισχύει η εξίσωση

$p(x,y)=p(x)\cdot p(y)$ και άρα για τον λογάριθμο προκύπτει ότι $\log(\frac{p(x,y)}{p(x)\cdot p(y)})=\log(1)=0$,

οπότε τελικά για την αμοιβαία πληροφορία ισχύει:

$$\begin{aligned} I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(\frac{p(x,y)}{p(x)\cdot p(y)}) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log(1) \\ &= 0. \end{aligned} \qquad 3.30$$

Από την άλλη πλευρά, στην περίπτωση που η X είναι μια ντετερμινιστική συνάρτηση της Y (ή το αντίθετο), δηλαδή $H(X|Y)=0$, τότε όλη την πληροφορία την οποία μεταφέρει το X, θα την μοιράζεται με το Y. Με λίγα λόγια γνωρίζοντας το X, θα μπορεί κάποιος να καθορίσει πλήρως το Y (και το αντίστροφο). Σε αυτή την περίπτωση, η αμοιβαία πληροφορία των δύο μεταβλητών, θα είναι ίση με την αβεβαιότητα που περιέχει μόνη της η μεταβλητή Y (ή η μεταβλητή X αντίστοιχα). Επομένως θα είναι ίση με την εντροπία της μεταβλητής Y ή αντίστοιχα την εντροπία της μεταβλητής X, $I(X;Y)=H(X)=H(Y)$.

Ακόμη, σε περίπτωση που οι δύο τυχαίες μεταβλητές X,Y δεν είναι διακριτές, αλλά συνεχείς τότε η $I(X;Y)$ δίνεται από την σχέση:

$$I(X;Y)=\iint_{Y,X} p(x,y)\cdot\log(\frac{p(x,y)}{p(x)\cdot p(y)})dxdy, \qquad 3.31$$

όπου $p(x,y)$ είναι η σ.π.π. των δύο μεταβλητών (X,Y) και $p(x)$, $p(y)$ είναι οι περιθώριες κατανομές αυτών αντίστοιχα.

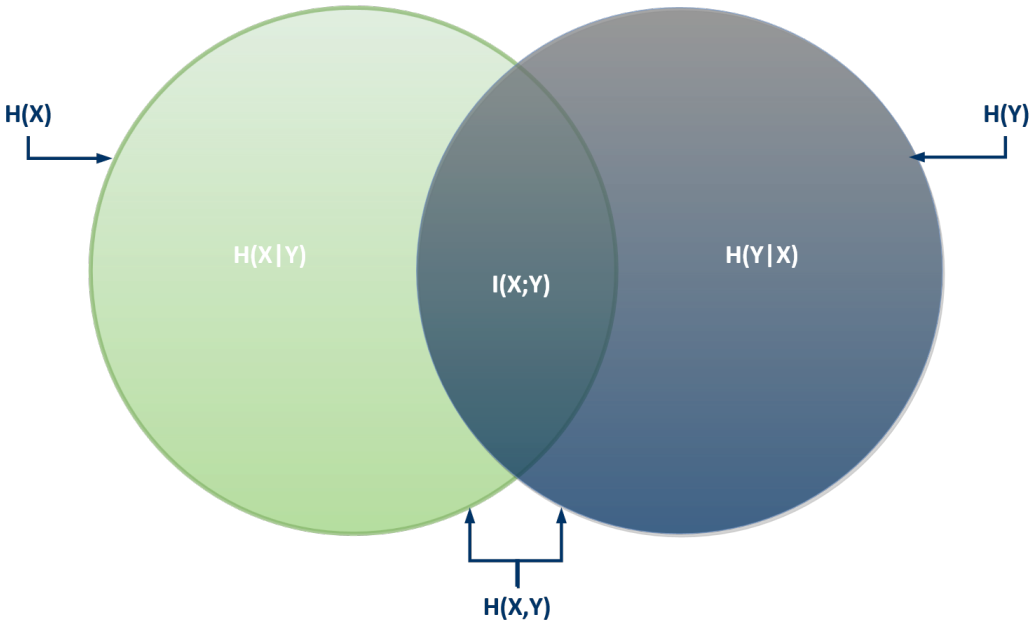
Τέλος δύο αναγκαίες και ικανές συνθήκες για την αμοιβαία πληροφορία είναι οι ακόλουθες:

- Η αμοιβαία πληροφορία είναι πάντοτε θετική, δηλαδή $I(X;Y)\geq 0$.
- Η αμοιβαία πληροφορία είναι πάντα συμμετρική, δηλαδή $I(X;Y)=I(Y;X)$.

Η δεύτερη ιδιότητα αναφέρθηκε και παραπάνω.

Οι σχέσεις που διέπουν όλα αυτά τα μέτρα που περιγράφηκαν και αναλύθηκαν παραπάνω ($H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, $H(Y|X)$, $I(X;Y)$) μπορούν να αναπαρασταθούν με την βοήθεια ενός διαγράμματος Venn (Σχήμα 3.2), και παράλληλα να γίνουν και καλύτερα κατανοητές. Σε μία τέτοια αναπαράσταση η αμοιβαία πληροφορία $I(X;Y)$ των

μεταβλητών X, Y αντιστοιχεί στην τομή της πληροφορίας που περιέχει η X με την πληροφορία που περιέχει η Y.



ΣΧΗΜΑ 3.2 Σχέσεις μεταξύ δύο μέτρων ποσότητας πληροφορίας

Η αμοιβαία πληροφορία ικανοποιεί τον παρακάτω κανόνα αλυσίδας:

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i-1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1). \qquad 3.32$$

Η απόδειξη του κανόνα είναι:

$$\begin{aligned} I(X_1, X_2, \dots, X_n; Y) &= H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n | Y) \\ &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1, Y) \\ &= \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \dots X_{i-1}). \end{aligned} \qquad 3.33$$

3.2.6 Υπό Συνθήκη Αμοιβαία Πληροφορία

Η υπό συνθήκη αμοιβαία πληροφορία είναι επίσης ένα από τα πιο βασικά μέτρα στην θεωρία πληροφοριών και την συμβολίζεται με $I_z(X;Y)$. Η ποσότητα αυτή μπορεί να ερμηνευτεί ως η μείωση της αβεβαιότητας της τυχαίας μεταβλητής X λόγω της γνώσης που υπάρχει για τη τυχαία μεταβλητή Y, με δεδομένο ότι έχει ήδη παρατηρηθεί και μια τρίτη τυχαία μεταβλητή Z. Γενικά ορίζεται ως:

$$\begin{aligned} I_z(X;Y) &= I(X;Y|Z) = E_z(I(X;Y|Z)) \\ &= \sum_{z\in Z} p_Z(z) \sum_{y\in Y}\sum_{x\in X} p_{X,Y|Z}(x,y|z)\cdot \log\left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)\cdot p_{Y|Z}(y|z)}\right) \\ &= \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} p_{X,Y,Z}(x,y,z)\cdot \log\left(\frac{p_Z(z)\cdot p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)\cdot p_{Y,Z}(y,z)}\right). \end{aligned} \qquad 3.34$$

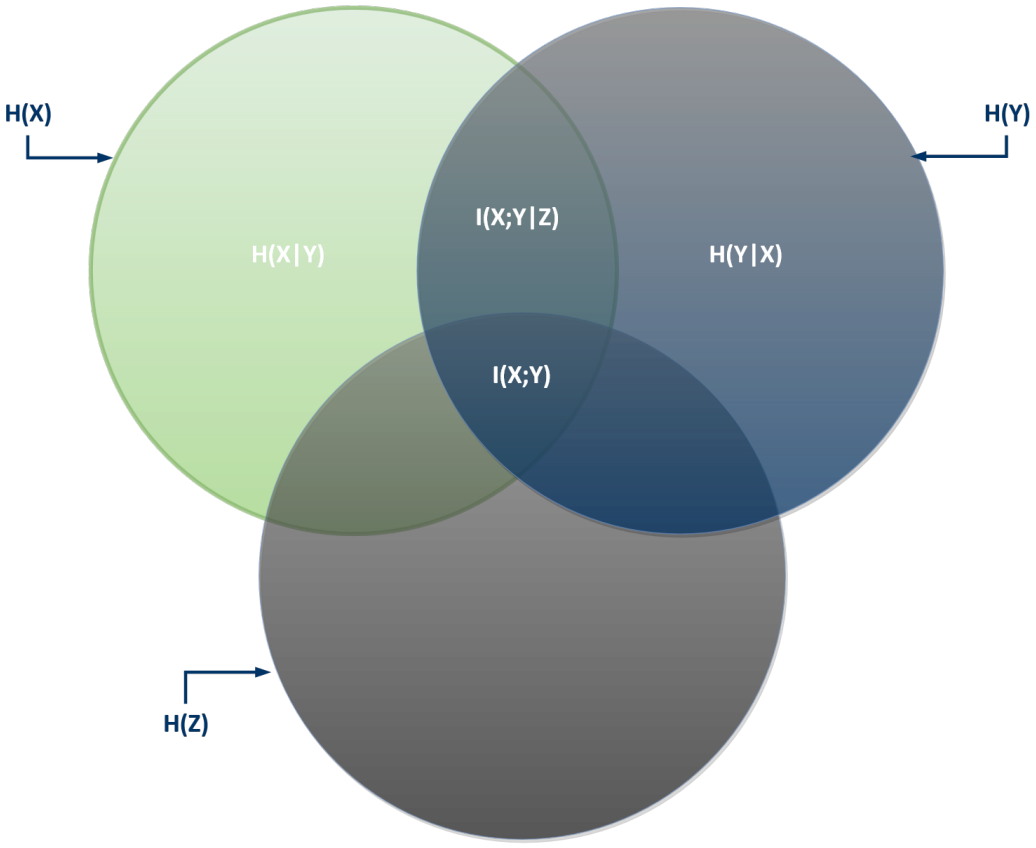
Εναλλακτικά, μπορεί να εκφραστεί μέσω της προσδοκίας ως:

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} p_{X,Y,Z}(x,y,z)\cdot \log\left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)\cdot p_{Y|Z}(y|z)}\right) \\ &= E_{p(x,y,z)} \log\left(\frac{p(X,Y|Z)}{p(X|Z)\cdot p(Y|Z)}\right). \end{aligned} \qquad 3.35$$

Τέλος, εύκολα αποδεικνύεται ότι μπορεί να περιγραφεί και με την παρακάτω μορφή:

$$I(X;Y|Z) = H(X,Z)+H(Y,Z)-H(X,Y,Z)-H(Z) = H(X|Z)-H(X|Y,Z). \qquad 3.36$$

Κατ'αντιστοιχία με την περίπτωση δύο μεταβλητών X, Y, και στην περίπτωση ύπαρξης τριών τυχαίων μεταβλητών X, Y, Z, οι σχέσεις που συνδέουν τα ποσότητες $H(X)$, $H(Y)$, $H(Z)$, $H(X|Y)$, $H(Y|X)$, $I(X;Y)$ και $I(X;Y|Z)$ μπορούν να αναπαρασταθούν μέσω ενός διαγράμματος Venn (Σχήμα 3.3), επεκτείνοντας ουσιαστικά το διάγραμμα του Σχήματος 3.2, που αναφέρεται σε δύο μεταβλητές, σε τρεις. Αντίστοιχα, σε αυτό το διάγραμμα η υπό συνθήκη αμοιβαία πληροφορία $I_z(X;Y)$ αναπαριστάται από την τομή της πληροφορίας που περιέχει μόνο η μεταβλητή X με την πληροφορία που περιέχει η μεταβλητή Y, μείον την τομή της πληροφορίας που περιέχουν και οι τρεις μεταβλητές X, Y, Z μαζί.



ΣΧΗΜΑ 3.3 Σχέσεις μεταξύ τριών μέτρων ποσότητας πληροφορίας

Για την υπό συνθήκη αμοιβαία πληροφορία, οι μεταβλητές X, Y, Z δεν είναι αναγκαίο να αντιπροσωπεύουν αποκλειστικά επιμέρους τυχαίες μεταβλητές αλλά θα μπορούσαν επίσης να αντιπροσωπεύουν την από κοινού κατανομή κάθε συνδυασμού τυχαίων μεταβλητών, οι οποίες όμως να ορίζονται στο ίδιο χώρο πιθανοτήτων. Με άλλα λόγια θα μπορούσε να ισχύει $I(X_1, X_2; Y_1, Y_2 | Z_1; Z_2)$.

3.2.7 Υπό Συνθήκη Σχετική Εντροπία

Έστω οι δύο συναρτήσεις $p(x,y)$ και $q(x,y)$, οι οποίες εκφράζουν τις από κοινού συναρτήσεις μάζας πιθανότητας των δύο μεταβλητών. Τότε η ποσότητα της δεσμευμένης σχετικής εντροπίας $D(p(y|x)||q(y|x))$ ισούται με το μέσο όρο των σχετικών εντροπιών μεταξύ των δύο δεσμευμένων συναρτήσεων μάζας πιθανότητας $p(y|x)$ και $q(y|x)$

υπολογισμένο ως προς τη συνάρτηση μάζας πιθανότητας $p(x)$. Συγκεκριμένα,

$$D(p(y|x)||q(y|x)) = \sum_x p(x) \sum_y p(y|x) \cdot \log\left(\frac{p(y|x)}{q(y|x)}\right) = E_{p(x,y)} \log\left(\frac{p(Y|X)}{q(Y|X)}\right). \quad 3.37$$

Ο παραπάνω συμβολισμός για την από κοινού σχετική εντροπία δεν είναι πλήρης, διότι δεν αναφέρει την κατανομή $p(x)$ της δεσμεύουσας τυχαίας μεταβλητής. Συνήθως όμως εννοείται από τα συμφραζόμενα.

Η σχετική εντροπία, μεταξύ δύο από κοινού κατανομών ενός ζεύγους τυχαίων μεταβλητών μπορεί να εκφραστεί ως άθροισμα μιας σχετικής εντροπίας και μιας δεσμευμένης σχετικής εντροπίας. Αυτός θεωρείται άλλωστε και ο κανόνας αλυσίδας για την σχετική εντροπία. Η εξίσωση είναι:

$$D(p(y,x)||q(y,x)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \quad 3.38$$

Και η αντίστοιχη απόδειξη:

$$\begin{aligned} D(p(y,x)||q(y,x)) &= \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(x,y)}{q(x,y)}\right) \\ &= \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(x) \cdot p(y|x)}{q(x) \cdot q(y|x)}\right) \\ &= \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(x)}{q(x)}\right) + \sum_x \sum_y p(x,y) \cdot \log\left(\frac{p(y|x)}{q(y|x)}\right) \\ &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \end{aligned} \quad 3.39$$

ΚΕΦΑΛΑΙΟ 4

Επιλογή χαρακτηριστικών με χρήση μέτρων πληροφορίας

4.1 Εισαγωγή

Με την απόκτηση της πληροφορίας να γίνεται όλο και πιο εύκολη και το μέγεθος των βάσεων δεδομένων συνεχώς να αυξάνεται, η ευκολία να μελετηθεί ένα πρόβλημα, να εξαχθούν συμπεράσματα και τελικά να πραγματοποιηθεί η επίλυσή του μειώνεται δραματικά. Ο λόγος είναι ότι μεγάλο μέρος των δεδομένων που υπάρχουν διαθέσιμα για το εκάστοτε πρόβλημα, μπορεί να αφορά άχρηστες πληροφορίες, με την έννοια ότι μπορεί να είναι ασύνδετες με το υπό μελέτη πρόβλημα.

Σε ένα σχετικό και με το αντικείμενο την παρούσας εργασίας παράδειγμα που θα βοηθήσει τον αναγνώστη να κατανοήσει το πρόβλημα, έστω ότι ζητούμενο είναι να ταξινομηθούν οι πελάτες μιας τράπεζας – χρήστες πιστωτικών καρτών σε δύο κατηγορίες: επίφοβοι προς πτώχευση, οικονομικά φερέγγυοι. Είναι φανερό ότι για έναν αλγόριθμο ταξινόμησης, η συμμετοχή χαρακτηριστικών όπως το ονοματεπώνυμο του πελάτη ή ο τόπος καταγωγής του κ.α. στο σύνολο εκπαίδευσης, δε θα συμβάλει στην βελτίωση της ακρίβειας του μοντέλου. Αντίθετα η συμμετοχή χαρακτηριστικών όπως το μηνιαίο εισόδημα ή η έγκυρη καταβολή των δόσεων σε αυτό, αναμένεται να αποτελέσουν καθοριστικούς παράγοντες πρόβλεψης. Όπως γίνεται εμφανές, μερικά από τα διαθέσιμα χαρακτηριστικά μπορεί να αποδειχτούν πολύ χρήσιμα, έως και καθοριστικής σημασίας, ενώ άλλα εντελώς ασυσχέτιστα με το πρόβλημα, που ακόμα και αν παραληφθούν δεν θα παρατηρηθεί καμία βελτίωση στην απόδοση του αλγορίθμου (αντίθετα σε ακραίες περιπτώσεις μπορεί να παρατηρηθεί και μείωση της απόδοσης λόγω εισαγωγής θορύβου).

Σε πολλά προβλήματα, λοιπόν, του πραγματικού κόσμου όπως σε εφαρμογές αναγνώρισης προτύπων, στατιστικής ανάλυσης προσδιορισμού συναρτήσεων από πεπερασμένα σύνολα δεδομένων κ.ά., όπου προκύπτει η αναγκαιότητα διαχείρισης και επεξεργασίας τεράστιας ποσότητας πληροφορίας αποδοτικά και γρήγορα, αυτό συνεπάγεται τον προσδιορισμό των μεταβλητών που περιγράφουν με επάρκεια το πρόβλημα με όσο το δυνατόν βέλτιστο τρόπο, προτού κανείς προβεί σε οποιουδήποτε είδους εργασία πάνω στον ακατέργαστο όγκο μετρήσεων που ενδεχομένως είναι διαθέσιμος (Guyon και Elisseeff, 2003).

Οι τεχνικές που συνθέτουν το πεδίο ελάττωσης διαστάσεων μπορούν να καταταχθούν σε διάφορες κατηγορίες αναλόγως ποιο κριτήριο λαμβάνεται υπόψιν. Ωστόσο, συνηθίζεται να ταξινομούνται σύμφωνα με τη γενική λειτουργία που αυτές επιτελούν, διότι είναι πιο χρήσιμο. Σύμφωνα με αυτό το κριτήριο, διακρίνονται στις τεχνικές εξαγωγής χαρακτηριστικών (feature extraction), των οποίων ο στόχος είναι η δημιουργία νέων χαρακτηριστικών μέσω μετασχηματισμών του αρχικού χώρου των ακατέργαστων δεδομένων, και σε αυτές της επιλογής χαρακτηριστικών (feature selection) που επιλέγουν τα πιο αντιπροσωπευτικά χαρακτηριστικά από τα ήδη υπάρχοντα στον αρχικό χώρο ακατέργαστων δεδομένων.

Η κύρια ιδέα της επιλογής χαρακτηριστικών είναι η επιλογή ενός υποσυνόλου μεταβλητών, εξαλείφοντας εκείνα τα στοιχεία με μικρή ή περιττή πληροφορία. Με αυτόν τον τρόπο μειώνεται η διάσταση των δεδομένων, αφού τα περιττά διανύσματα δεν συμμετέχουν στις περαιτέρω διαδικασίες. Παρακάτω ακολουθεί ένας ορισμός για αυτήν την διαδικασία:

"Δεδομένου ενός D-διάστατου συνόλου N δειγμάτων $\{x_i\}_{i=1}^{N}$ *με άγνωστη κατανομή πιθανότητας, στόχος είναι να προσδιοριστεί μια νέα d-διάστατη απεικόνιση ώστε να διατηρείται η βασική δομή των δεδομένων με το ελάχιστο δυνατό σφάλμα αναπαράστασης."*

Συνοπτικά η σημασία των διαδικασιών επιλογής χαρακτηριστικών αναλύεται στις ακόλουθες διαστάσεις (Kira and Rendell, 1992):

i. Πιθανή μείωση του θορύβου στα δεδομένα, η οποία οφείλεται στην ύπαρξη χαρακτηριστικών που δεν παρέχουν αξιόπιστη πληροφορία.

ii. Περιορισμός του υπολογιστικού φόρτου που απαιτείται για την υλοποίηση της ανάλυσης και την ανάπτυξη βέλτιστων υποδειγμάτων.

iii. Απλοποίηση των αναπτυσσόμενων υποδειγμάτων, καθώς υποδείγματα που εξετάζουν περιορισμένη πληροφορία έχουν πιο απλή μορφή και συνεπώς μπορούν να ερμηνευτούν πιο εύκολα.

iv. Μείωση του χρόνου και του κόστους της χρήσης των υποδειγμάτων, καθώς περιορίζεται η ποσότητα της πληροφορίας που πρέπει να είναι διαθέσιμη για τη χρήση τους.

Σε αυτό το σημείο πρέπει να αναφερθεί ότι η δυσκολία του προβλήματος της επιλογής ενός υποσυνόλου χαρακτηριστικών οφείλεται σε δύο κυρίως λόγους. Πρώτον, το πλήθος των δυνατών υποσυνόλων που θα μπορούσαν να επιλεγούν αυξάνεται εκθετικά σε σχέση με τον αριθμό των χαρακτηριστικών του αρχικού συνόλου. Αν υποθέσουμε ότι δίνεται ένα κριτήριο αξιολόγησης υποσυνόλων, η εύρεση του καλύτερου υποσυνόλου ως προς αυτό δεν είναι υπολογιστικά εφικτή από κάποιο μέγεθος αρχικού συνόλου και έπειτα. Δεύτερον, η ποιότητα ενός υποσυνόλου εξαρτάται από πολλούς παράγοντες και έτσι δεν μπορεί να οριστεί εύκολα ένα αντικειμενικό κριτήριο αξιολόγησης. Με άλλα λόγια δεν υπάρχει τρόπος να αποτιμηθεί με ακρίβεια η ποιότητα ενός υποσυνόλου, αντίθετα στην πράξη με χρήση ευρετικών (heuristic) τεχνικών επιλέγεται τελικά ένα υποσύνολο που αναμένεται ότι θα οδηγήσει σε καλή απόδοση τον αλγόριθμο μάθησης που θα το χρησιμοποιήσει.

Μια τυπική διαδικασία επιλογής χαρακτηριστικών, αποτελείται από δύο φάσεις. Η πρώτη είναι η επιλογή των χαρακτηριστικών και η δεύτερη η αξιολόγησή τους και περιλαμβάνει τα ακόλουθα βήματα: Αρχικά, δημιουργία ενός υποψηφίου σετ που περιέχει ένα υποσύνολο από τα αρχικά χαρακτηριστικά μέσω ορισμένων στρατηγικών. Ακολούθως, αξιολόγηση του υποψήφιου συνόλου και εκτίμηση της χρησιμότητας των χαρακτηριστικών στο σύνολο αυτό. Με βάση αυτή την αξιολόγηση, ορισμένα χαρακτηριστικά στο υποψήφιο σύνολο μπορεί να απορριφθούν, ενώ κάποια άλλα μπορεί να προστεθούν. Τέλος, χρησιμοποιούνται ορισμένα κριτήρια διακοπής,

προκειμένου να καθοριστεί εάν το τρέχον σύνολο των επιλεγμένων χαρακτηριστικών, είναι αρκετά καλό ή όχι.

Οι πιο γνωστές μέθοδοι επιλογής υποσυνόλου χαρακτηριστικών είναι οι μέθοδοι φίλτρων (filter), οι μέθοδοι περιτυλίγματος (wrapper), και οι ενσωματωμένες μέθοδοι (embedded) (Yang και Pedersen, 1997), (Μητροπούλου, 2016). Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στην πρώτη κατηγορία εφαρμόζονται πριν τη χρήση κάποιας τεχνικής ταξινόμησης και συνεπώς δεν επηρεάζονται από αυτή. Ουσιαστικά, οι αλγόριθμοι αυτής της κατηγορίας λειτουργούν ως φίλτρα για την απαλοιφή των μη σχετικών ή πλεοναστικών χαρακτηριστικών. Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στη δεύτερη κατηγορία χρησιμοποιούν τη μέθοδο ταξινόμησης ως μέρος της διαδικασίας (John et al., 1994). Ειδικότερα, βασιζόμενοι σε εμπρόσθιες, ανάστροφες ή τυχαίες διαδικασίες, οι αλγόριθμοι της κατηγορίας αυτής χρησιμοποιούν τη απόδοση της μεθόδου ταξινόμησης για την αξιολόγηση της αποτελεσματικότητας του συνόλου των χαρακτηριστικών που επιλέγονται. Τέλος, στην τρίτη κατηγορία περιλαμβάνονται αλγόριθμοι και τεχνικές, η εφαρμογή των οποίων είναι άμεσα συνδεδεμένη με μια συγκεκριμένη τεχνική ταξινόμησης.

4.2 Κατηγορίες Μεθόδων Επιλογής Χαρακτηριστικών

Ξεκινώντας από την κατηγορία των φίλτρων, σε αυτήν ανήκουν όσοι αλγόριθμοι βασίζονται στην έννοια της συνάφειας μεταξύ χαρακτηριστικών και κλάσης, και όχι σε κάποιο ταξινομητή προκειμένου να εκτιμήσουν την ποιότητα ενός υποσυνόλου χαρακτηριστικών (Kakoyan, 2010). Πρακτικά με την χρήση στατιστικών μέτρων προσπαθούν να εντοπίσουν συναφή χαρακτηριστικά. Με άλλα λόγια αξιολογούν τη σχετικότητα των χαρακτηριστικών ερευνώντας μόνο τις ιδιότητες των στοιχείων. Η πιο κοινή διαδικασία είναι βασισμένη στον υπολογισμό της σχετικότητας χαρακτηριστικών γνωρισμάτων. Τα χαρακτηριστικά γνωρίσματα με τη χαμηλή σχετικότητα αφαιρούνται. Τα υπόλοιπα χαρακτηριστικά γνωρίσματα χρησιμοποιούνται ως είσοδος στον αλγόριθμο ταξινόμησης.

Οι τεχνικές φίλτρων έχουν το πλεονέκτημα ότι είναι απλές και γρήγορες, είναι εφικτές από άποψη υπολογιστικής πολυπλοκότητας που εφαρμόζεται στα δεδομένα υψηλής διάστασης (όπως δεδομένα DNA, κείμενο κτλ) και επίσης είναι ανεξάρτητες από τον αλγόριθμο ταξινόμησης. Το τελευταίο πλεονέκτημα είναι πολύ σημαντικό, δεδομένου ότι κάποιος μπορεί να εφαρμόσει μια μέθοδο φίλτρων για να παράξει ένα βέλτιστο υποσύνολο χαρακτηριστικών γνωρισμάτων που μπορεί να αξιολογηθεί χρησιμοποιώντας διαφορετικούς ταξινομητές (classifiers).

Το κύριο μειονέκτημα των αλγορίθμων αυτών είναι ότι αγνοούν την αλληλεπίδραση που πιθανόν υπάρχει μεταξύ του συνόλου των χαρακτηριστικών που επιλέγεται και της τεχνικής ταξινόμησης που χρησιμοποιείται για την ανάπτυξη του υποδείγματος ταξινόμησης, αφού κατά την διαδικασία της επιλογής των χαρακτηριστικών δεν υπάρχει οποιαδήποτε αλληλεπίδραση με τον ταξινομητή. Αυτό σημαίνει ότι κάθε χαρακτηριστικό γνώρισμα ελέγχεται χωριστά και κατά συνέπεια οι εξαρτήσεις χαρακτηριστικών γνωρισμάτων αγνοούνται. Το γεγονός αυτό μπορεί να οδηγήσει στη χαμηλότερη απόδοση ταξινόμησης σε σύγκριση με άλλες τεχνικές επιλογής χαρακτηριστικών γνωρισμάτων. Σε μια προσπάθεια να ελεγχθούν οι εξαρτήσεις χαρακτηριστικών γνωρισμάτων πολλών μεταβλητών οι τεχνικές φίλτρων έχουν βελτιωθεί, στοχεύοντας στην ενσωμάτωση των εξαρτήσεων χαρακτηριστικών γνωρισμάτων μέχρι ενός ορισμένου βαθμού.

Οι μέθοδοι φίλτρου διακρίνονται σε δύο βασικές κατηγορίες, τις μονοπαραγοντικές μεθόδους (univariate) και τις πολυπαραγοντικές μεθόδους (multivariate). Οι μέθοδοι της πρώτης κατηγορίας αρχικά αξιολογούν μεμονωμένα κάθε χαρακτηριστικό, με βάση τη συσχέτιση του με τις κλάσεις. Όσο μεγαλύτερη συσχέτιση υπάρχει, τόσο πιο χρήσιμο θεωρείται το χαρακτηριστικό. Ύστερα επιλέγονται τα k πιο συσχετισμένα χαρακτηριστικά, όπου το k καθορίζεται ανάλογα με την περίσταση. Η κυριότερη αδυναμία αυτών των μεθόδων, είναι η εμφάνιση φαινομένων πλεονασμού, δηλαδή περιπτώσεις όπου επιλέγονται περιττά χαρακτηριστικά, με την έννοια ότι είναι όμοια μεταξύ τους και έτσι ο συνδυασμός τους δεν προσφέρει πολύ περισσότερη πληροφορία για την κατηγορία από αυτή που θα προσέφερε κάθε χαρακτηριστικό από μόνο του. Αυτό συμβαίνει κατά κύριο λόγο επειδή κάθε χαρακτηριστικό αξιολογείται ξεχωριστά, χωρίς να λαμβάνονται υπόψη τα άλλα που έχουν ήδη επιλεγεί.

Μερικά από τα κριτήρια που έχουν χρησιμοποιηθεί για τη μέτρηση της συσχέτισης είναι το κριτήριο του Fischer το οποίο μπορεί να χρησιμοποιηθεί σε προβλήματα δύο κατηγοριών, το F-test που μπορεί να χρησιμοποιηθεί για προβλήματα με K κατηγορίες και τέλος η μέτρηση της αμοιβαίας τους πληροφορίας, η οποία μπορεί να ανιχνεύσει και τις γραμμικές εξαρτήσεις μεταξύ των μεταβλητών.

Οι μέθοδοι της δεύτερης κατηγορίας, σε αντίθεση με αυτές της πρώτης, αξιολογούν τα χαρακτηριστικά λαμβάνοντας υπόψιν την παρουσία και των άλλων χαρακτηριστικών, προσπαθώντας έτσι να αποφύγουν την επιλογή περιττών χαρακτηριστικών και τα φαινόμενα πλεονασμού που έχει η πρώτη κατηγορία. Στην πράξη αυτό που κάνουν είναι να φτιάχνουν ένα βέλτιστο υποσύνολο επιλέγοντας χαρακτηριστικά που έχουν μεγάλη συσχέτιση σε σχέση με την κλάση (όπως και οι μονοπαραγοντικές μέθοδοι) ενώ παράλληλα ελέγχουν τα χαρακτηριστικά που επιλέγονται να είναι όσο το δυνατόν πιο ανόμοια μεταξύ τους. Τα υποσύνολα δηλαδή αξιολογούνται με βάση την περιεχόμενη πληροφορία, και τη σχετική ανεξαρτησία που έχουν μεταξύ τους.

Συνεχίζοντας με την κατηγορία των μεθόδων περιτυλίγματος, αυτές ενσωματώνουν τους ταξινομητές (classifiers) μέσα στην αναζήτηση υποσυνόλων χαρακτηριστικών γνωρισμάτων (Van Dijck και Van Hulle, 2006). Στην κατηγορία αυτών των μεθόδων εντάσσονται όλοι οι αλγόριθμοι επιλογής χαρακτηριστικών που χρησιμοποιούν την ακρίβεια ταξινόμησης ως κριτήριο αξιολόγησης των διάφορων υποσυνόλων των χαρακτηριστικών. Επιπλέον, σε αυτές τις μεθόδους τα χαρακτηριστικά γνωρίσματα συνήθως αξιολογούνται σε ομάδες και όχι χωριστά. Όλα τα ήδη διαθέσιμα χαρακτηριστικά γνωρίσματα χρησιμοποιούνται για να παραγάγουν τα υποσύνολα χαρακτηριστικών γνωρισμάτων που αξιολογούνται. Για να βρεθεί το διάστημα όλων των υποσυνόλων χαρακτηριστικών γνωρισμάτων, ένας αλγόριθμος αναζήτησης είναι συνέχεια “τυλιγμένος” (wrapped) γύρω από το πρότυπο ταξινόμησης. Εντούτοις, καθώς το διάστημα του υποσυνόλου χαρακτηριστικών γνωρισμάτων αυξάνεται εκθετικά με τον αριθμό χαρακτηριστικών γνωρισμάτων, χρησιμοποιούνται ευρετικές (heuristic) μέθοδοι αναζήτησης για να καθοδηγήσουν την αναζήτηση ενός βέλτιστου υποσυνόλου.

Η διαδικασία που ακολουθεί μια τυπική μέθοδος περιτυλίγματος είναι η εξής: Αρχικά χωρίζει τα δεδομένα εκπαίδευσης σε δύο νέα σύνολα, το σύνολο εκπαίδευσης (training) και το σύνολο επικύρωσης (validation). Ακολούθως, διαγράφονται όσα χαρακτηριστικά

δεν ανήκουν στο υποψήφιο προς επιλογή υποσύνολο. Στη συνέχεια, ο ταξινομητής, εκπαιδεύεται με το τροποποιημένο σύνολο εκπαίδευσης και βάση αυτής της εκπαίδευσης κατατάσσει τα στοιχεία που ανήκουν στο τροποποιημένο σύνολο επικύρωσης σε μια σειρά. Η ακρίβεια με την οποία τα δεδομένα αυτά ταξινομούνται, είναι το κριτήριο αξιολόγησης των μεθόδων περιτυλίγματος για ένα οποιοδήποτε υποψήφιο σύνολο χαρακτηριστικών.

Βέβαια, η αξιολόγηση με βάση την απόδοση του ταξινομητή, η οποία απαιτεί την κατασκευή του ταξινομητή για κάθε ξεχωριστό υποσύνολο χαρακτηριστικών που εξετάζεται, έχει ως αρνητικό επακόλουθο το αυξημένο υπολογιστικό κόστος σε σχέση με τις πιο εξελιγμένες ενσωματωμένες μεθόδους ή τα φίλτρα. Το αυξημένο υπολογιστικό κόστος είναι ίσως και το πιο βασικό μειονέκτημα των μεθόδων περιτυλίγματος. Η αποτίμηση κάθε υποψήφιου υποσυνόλου, που συνεπάγεται την εκπαίδευση του ταξινομητή και ύστερα τη μέτρηση της απόδοσης του στο σύνολο επικύρωσης, είναι συνήθως χρονοβόρα διαδικασία και κάνει το υπολογιστικό κόστος ακόμα πιο υψηλό.

Ένα ακόμα από τα πιο βασικά μειονεκτήματα αυτών των μεθόδων, εμφανίζεται στην περίπτωση που το διαθέσιμο σύνολο δεδομένων είναι μικρό. Τότε δεν υπάρχει η δυνατότητα να σχηματιστεί μεγάλο σύνολο επικύρωσης γιατί το σύνολο εκπαίδευσης γίνεται υπερβολικά μικρό και αντίστροφα. Και στις δύο περιπτώσεις υπάρχει πρόβλημα, πιο συγκεκριμένα, στη περίπτωση που το σύνολο εκπαίδευσης είναι μικρό, δεν μπορεί να γίνει καλή εκπαίδευση του ταξινομητή, ενώ στη περίπτωση που το σύνολο επικύρωσης είναι μικρό, δεν μπορεί να γίνει αξιόπιστη εκτίμηση όσο αφορά την ακρίβεια στη ταξινόμηση. Σε αυτήν την περίπτωση χρησιμοποιούνται τεχνικές επαναληπτικής δειγματοληψίας (resampling techniques), όπως το cross-validation (Stone, 1974) και το bootstrap (Efron και Gong, 1983) χάρη στις οποίες αποφεύγεται και η υπερπροσαρμογή (over-fitting) του συνόλου εκπαίδευσης, το οποίο είναι ένα πολύ συχνό φαινόμενο. Στην τεχνική cross-validation τα παραδείγματα εκπαίδευσης χωρίζονται σε k ξένα υποσύνολα (υποσύνολα παραδειγμάτων). Ο ταξινομητής εκπαιδεύεται στα παραδείγματα των k-1 υποσυνόλων, τα οποία παίζουν τον ρόλο του συνόλου εκπαίδευσης, ενώ το ένα υποσύνολο που περισσεύει παίζει το ρόλο του συνόλου επικύρωσης. Η ίδια διαδικασία επαναλαμβάνεται k φορές έτσι ώστε κάθε υποσύνολο να παίξει το ρόλο του συνόλου επικύρωσης ακριβώς μία φορά. Ο μέσος όρος της ακρίβειας ταξινόμησης στα k

διαφορετικά σύνολα επικύρωσης είναι το κριτήριο αξιολόγησης. Σε περίπτωση που χρησιμοποιείται κάποια τεχνική επαναληπτικής δειγματοληψίας όπως το cross-validation, το ήδη μεγάλο υπολογιστικό κόστος που υπήρχε, αυξάνεται περαιτέρω σε σημαντικό βαθμό.

Ένα ακόμη μειονέκτημα είναι ότι λόγω του μεγάλου όγκου των υποσυνόλων που εξετάζονται αυξάνεται κατά πολύ την πιθανότητα να βρεθεί τελικά κατά τύχη ένα υποσύνολο που δίνει πολύ καλή απόδοση στο σύνολο επικύρωσης, χωρίς όμως να έχει καλή ικανότητα γενίκευσης, ενώ την ίδια στιγμή είναι πολύ πιθανό άλλα υποσύνολα με σημαντικά μικρότερη ακρίβεια ταξινόμησης στο σύνολο επικύρωσης να επιτυγχάνουν καλύτερη ικανότητα γενίκευσης. Το πρόβλημα αυτό είναι γενικά γνωστό ως το πρόβλημα πολλαπλών συγκρίσεων και γίνεται ακόμα πιο έντονο όταν το σύνολο με τα διαθέσιμα δεδομένα εκπαίδευσης είναι πολύ μικρό.

Το ισχυρότερο επιχείρημα υπέρ της χρήσης των μεθοδολογιών περιτυλίγματος είναι ότι λαμβάνουν υπόψη την επαγωγική μεροληψία (inductive bias) του ταξινομητή. Κάθε ταξινομητής έχει τα δικά του ιδιαίτερα χαρακτηριστικά και τον δικό του τρόπο που απεικονίζει την είσοδο που δέχεται, σε έξοδο. Η επαγωγική μεροληψία είναι το σύνολο όλων αυτών των υποθέσεων που κάνει ο ταξινομητής, στη περίπτωση που δεν υπάρχουν επαρκή στοιχεία, έτσι ώστε να μπορέσει να κατατάξει δεδομένα στη σωστή κατηγορία. Σε αυτή τη περίπτωση, εφόσον δεν υπάρχουν παρόμοια παραδείγματα στα δεδομένα εκπαίδευσης και ο ταξινομητής δεν είναι βέβαιος, το πρόβλημα δεν μπορεί να λυθεί πλήρως. Αυτό σημαίνει ότι το καλύτερο υποσύνολο χαρακτηριστικών για έναν ταξινομητή τύπου Α δεν είναι απαραίτητα το καλύτερο υποσύνολο για έναν ταξινομητή τύπου Β. Η ακρίβεια ταξινόμησης είναι το πιο αξιόπιστο κριτήριο για να ελεγχθεί αν ένα υποσύνολο χαρακτηριστικών δουλεύει καλά σε συνδυασμό με έναν ταξινομητή.

Ακόμα ένα πλεονέκτημα τους είναι ότι θεωρητικά η χρήση μεθοδολογιών περιτυλίγματος δίνει τη δυνατότητα ανακάλυψης αλληλεπιδράσεων μεταξύ χαρακτηριστικών κι αυτό γιατί τα χαρακτηριστικά δεν αξιολογούνται μεμονωμένα αλλά ως μέρη ενός υποσυνόλου. Φυσικά αν υπάρχουν χαρακτηριστικά που αλληλεπιδρούν, η ανακάλυψη τους εξαρτάται από το αν θα τύχει να βρεθούν στο ίδιο υποψήφιο υποσύνολο ώστε να αξιολογηθούν ως ομάδα.

Τέλος, όσον αφορά την κατηγορία των ενσωματωμένων μεθόδων, σε αυτήν συναντά

κάποιος μεθόδους παρόμοιας περίπου φιλοσοφίας με τις μεθόδους περιτυλίγματος. Εδώ ανήκουν οι μέθοδοι και οι τεχνικές η εφαρμογή των οποίων είναι άμεσα συνδεδεμένη με μια συγκεκριμένη τεχνική ταξινόμησης. Ουσιαστικά έχουν σχεδιαστεί με στόχο να δουλεύουν αποκλειστικά και μόνο σε συνεργασία με ένα ταξινομητή συγκεκριμένου τύπου και ποτέ μόνες τους. Σε αντίθεση με τις μεθόδους περιτυλίγματος που απλώς χρησιμοποιούν την έξοδο ενός ταξινομητή για αξιολόγηση, οι μέθοδοι αυτές επιλέγουν χαρακτηριστικά με βάση το πως επηρεάζεται κάποια συνάρτηση κόστους που εμπλέκεται στη διαδικασία εκπαίδευσης του ταξινομητή. Ταυτόχρονα, δηλαδή, με την ανάπτυξη των υποδειγμάτων ταξινόμησης οι τεχνικές αυτές ενσωματώνουν στη δομή τους κατάλληλες διαδικασίες επιλογής των χαρακτηριστικών που συμμετέχουν στο τελικό υπόδειγμα.

Από αυτήν την ενσωμάτωση της διαδικασίας της επιλογής χαρακτηριστικών στη διαδικασία της εκπαίδευσης προκύπτουν διάφορα πλεονεκτήματα σε σχέση με τις μεθόδους περιτυλίγματος. Το πρώτο και ίσως κυριότερο είναι ότι αυτή η κατηγορία μεθόδων έχει μεγάλο κέρδος σε υπολογιστικό κόστος συγκριτικά με την προηγούμενη. Άλλο ένα πλεονέκτημα είναι ότι, οι ενσωματωμένες μέθοδοι καταφέρνουν να κάνουν καλύτερη χρήση των διαθέσιμων δεδομένων (δεδομένα εκπαίδευσης) αφού δεν υπάρχει η ανάγκη αυτά να χωριστούν σε σύνολα εκπαίδευσης και επικύρωσης σε αντίθεση με τις μεθόδους περιτυλίγματος.

Όσον αφορά τις μεθόδους φίλτρων, και πάλι οι μέθοδοι αυτής της κατηγορίας υπερτερούν διότι έχουν το πλεονέκτημα, να λαμβάνουν υπόψη τους την επαγωγική μεροληψία, όπως και οι μέθοδοι περιτυλίγματος.

4.3 Μέθοδος επιλογής χαρακτηριστικών mRMR

4.3.1 Περιγραφή μεθόδου mRMR

Μια από τις πιο δημοφιλείς μεθόδους επιλογής χαρακτηριστικών, στην κατηγορία των

φίλτρων, που χρησιμοποιείται και βασίζεται πάνω στο μέτρο της αμοιβαίας πληροφορίας, είναι η μέθοδος mRMR (minimal redundancy-maximal relevance), δηλαδή η μέθοδος της μέγιστης συνάφειας και του ελάχιστου πλεονασμού (Peng et al., 2005)

Η μέθοδος αυτή, δημιουργήθηκε, προκειμένου να καλύψει το πρόβλημα που υπήρχε όσο αφορά μια άλλη μέθοδο, αυτή του κριτηρίου της μέγιστης εξάρτησης (Max-Dependency), που και αυτή με τη σειρά της βασίζεται στη αμοιβαία πληροφορία. Η μέθοδος αυτή εμφάνιζε ορισμένες δυσκολίες στην άμεση εφαρμογή της. Αυτά ακριβώς τα προβλήματα, ήρθε να καλύψει η mRMR, η οποία είναι ισοδύναμη με αυτή, ίσως και αποδοτικότερη, κάτι που αποδεικνύεται και πειραματικά μέσω της σύγκρισης των δύο διαδικασιών.

Όπως είναι γνωστό, ο εντοπισμός των σημαντικότερων χαρακτηριστικών από ένα σύνολο δεδομένων, είναι αναγκαίος στη προσπάθεια για ελαχιστοποίηση του σφάλματος στη ταξινόμηση. Δοθέντος ενός προβλήματος, για το οποίο θα πρέπει να οριστούν τα "βέλτιστα" χαρακτηριστικά του, είναι απαραίτητο να οριστεί ένας αλγόριθμος, που να έχει τη δυνατότητα να επιλέξει το καλύτερο υποσύνολο. Η συνθήκη των βέλτιστων χαρακτηριστικών, είναι συχνά ταυτόσημη με το ελάχιστο σφάλμα στη ταξινόμηση.

Σε καταστάσεις, στις οποίες, οι ταξινομητές δεν είναι συγκεκριμένοι, προκειμένου να επιτευχθεί το ελάχιστο σφάλμα, είναι αναγκαίο να υπάρχει η μέγιστη δυνατή εξάρτηση της κλάσης c (target class), με την κατανομή των δεδομένων στο υποσύνολο με τα m χαρακτηριστικά, ή με άλλα λόγια τον υπόχωρο R^m, από τα συνολικά M χαρακτηριστικά (δηλαδή τον χώρο R^m).

Η πιο δημοφιλής προσέγγιση, προκειμένου να επιτευχθεί η μέγιστη εξάρτηση (Max-Dependency) είναι μέσω της μεθόδου επιλογής χαρακτηριστικών με τη μέγιστη συνάφεια (Max-Relevance) σε σχέση με την κλάση c. Η συνάφεια συνήθως ορίζεται με όρους συσχέτισης ή αμοιβαίας πληροφορίας, από τις οποίες η δεύτερη είναι και η πιο δημοφιλής που χρησιμοποιείται για να καθοριστεί η εξάρτηση των μεταβλητών.

Στο κριτήριο της μέγιστης συνάφειας, επιλέγονται τα m χαρακτηριστικά x_i, τα οποία έχουν τη μεγαλύτερη αμοιβαία πληροφορία, $I(x_i;c)$, με τη επιλεγμένη κλάση c, κάτι που συνεπάγεται ότι θα έχουν και τη μεγαλύτερη εξάρτηση με τη κλάση αυτή. Συχνά παρατηρείται το φαινόμενο, κατά τη διαδοχική αναζήτηση να επιλέγονται τα m

κορυφαία χαρακτηριστικά, αυτά δηλαδή που έχουν την μέγιστη αμοιβαία πληροφορία $I(x_i;c)$, ανεξάρτητα το ένα από το άλλο, ως τα m χαρακτηριστικά που βελτιστοποιούν τη ταξινόμηση. Αυτό δεν είναι βέλτιστο, διότι στη μέθοδο της επιλογής χαρακτηριστικών, είναι ευρέως αποδεκτό, ότι οι συνδυασμοί μεμονωμένων καλών χαρακτηριστικών δεν οδηγούν απαραιτήτως σε μια καλή ταξινόμηση - “the m best features are not the best m features”.

Έχει γίνει πλήθος ερευνών για τη μελέτη συγκεκριμένων τρόπων, οι οποίες χρησιμοποιούν είτε άμεσα είτε έμμεσα μέσα, προκειμένου να περιοριστεί η περίσσεια στα χαρακτηριστικά που επιλέγονται και να επιλεχθούν τελικά εκείνα τα χαρακτηριστικά με τον ελάχιστον πλεονασμό (min-Redundancy). Ο περιορισμός των χαρακτηριστικών ως προς την μεταξύ τους εξάρτηση δεν επιφέρει κανένα απολύτως πρόβλημα, αντιθέτως, είναι δυνατό κατά τη διαδοχική αναζήτηση, η από κοινού εξάρτηση των χαρακτηριστικών με την κλάση c να μεγιστοποιείται και αντίστοιχα ο πλεονασμός μεταξύ των χαρακτηριστικών να μειώνεται.

Παρακάτω θα ακολουθήσει καταρχήν η παρουσίαση του κριτηρίου της μέγιστης εξάρτησης και της μεθόδου mRMR, καθώς και μια θεωρητική ανάλυση των δύο. Επιπλέον, σκοπός είναι να δειχθεί πως είναι εφικτό, η μέθοδος mRMR, να συνδυαστεί και με άλλες μεθόδους επιλογής χαρακτηριστικών σχηματίζοντας έναν αλγόριθμο δύο σταδίων, ο οποίος να δίνει ένα συμπαγές υποσύνολο που αποτελείται από τα καλύτερα χαρακτηριστικά με πολύ χαμηλό υπολογιστικό κόστος.

Χρησιμοποιώντας όρους από τη θεωρία πληροφορίας, το ιδανικό υποσύνολο χαρακτηριστικών που θα επιλεγεί πρέπει να είναι αυτό που έχει τη μέγιστη αμοιβαία πληροφορία $I(x_i;c)$ με την κλάση c που μας ενδιαφέρει, δηλαδή αυτό από το οποίο η κατηγορία c να έχει τη μεγαλύτερη εξάρτηση. Το υποσύνολο χαρακτηριστικών που πρέπει να επιλεγεί είναι αυτό που μεγιστοποιεί την ποσότητα:

$$\max D(S,c),\ D=I(\{x_i,i=1,\dots,m\};c). \qquad 4.1$$

Προφανώς όταν ισχύει ότι $m=1$, η λύση ταυτίζεται με το χαρακτηριστικό που μεγιστοποιεί την αμοιβαία πληροφορία $I(x_j;c)\ (1\le j\le M)$, ενώ όταν $m>1$ η αναζήτηση γίνεται προσθέτοντας ένα χαρακτηριστικό κάθε φορά. Δεδομένου δηλαδή ενός συνόλου με $m-1$ χαρακτηριστικά, το m-οστό χαρακτηριστικό, μπορεί να οριστεί σαν αυτό το

οποίο προσφέρει τη μεγαλύτερη αύξηση της $I(S;c)$ η οποία ορίζεται ως εξής:

$$\begin{aligned} I(S_m;c) &= \int\int p(S_m,c)\cdot\log\left(\frac{p(S_m,c)}{p(S_m)\cdot p(c)}\right)\ dS_m\,dc \\ &= \int\int p(S_{m-1},x_m,c)\cdot\log\left(\frac{p(S_{m-1},x_m,c)}{p(S_{m-1},x_m)\cdot p(c)}\right)\ dS_{m-1}\,dx_m\,dc \\ &= \int\ldots\int p(x_1,\ldots,x_m,c)\cdot\log\left(\frac{p(x_1,\ldots,x_m,c)}{p(x_1,\ldots,x_m)\cdot p(c)}\right)\ dx_1\ldots dx_m\,dc, \end{aligned} \qquad 4.2$$

όπου $S(m)=\{x_1,x_2,\ldots,x_m\}$ είναι το σύνολο των χαρακτηριστικών που επιλέγεται, c είναι η κλάση (target class), $p(S_m)$ είναι η από κοινού κατανομή των m χαρακτηριστικών $x_1,x_2,\ldots,x_m$ και τέλος $p(S_m,c)$ η από κοινού κατανομή των m χαρακτηριστικών $x_1,x_2,\ldots,x_m$ και του c μαζί.

Παρά τη μεγάλη θεωρητική αξία που έχει το κριτήριο της μέγιστης εξάρτησης, συνήθως, είναι δύσκολο να προσδιοριστούν με ακρίβεια οι ποσότητες $p(S_m)$ και $p(S_m,c)$ και επομένως η αμοιβαία πληροφορία $I(S_m;c)$. Αυτό οφείλεται στις εξής δυσκολίες που παρουσιάζονται σε χώρους μεγάλων διαστάσεων. Καταρχήν, η εκτίμηση της πυκνότητας πολλών μεταβλητών, συχνά απαιτεί και τον υπολογισμό του αντιστρόφου πίνακα συνδιασποράς κάτι που αρκετές φορές μπορεί να μην είναι εφικτό.

Ένα επιπλέον πρόβλημα είναι το γεγονός ότι ο αριθμός των δειγμάτων που υπάρχει τις περισσότερες φορές είναι σχετικά μικρός. Αυτό συμβαίνει γιατί ο αριθμός παραδειγμάτων που απαιτούνται για την εκτίμηση των από κοινού κατανομών $p(S_m,c)$ και $p(S_m)$ αυξάνεται εκθετικά ως προς των αριθμό των χαρακτηριστικών S_m. Για παράδειγμα, ας υποτεθεί ότι ζητείται η εκτίμηση της κατανομής $p(x_k)$ και ότι τα $(x_1,x_2,\ldots,x_m)$ είναι διακριτά χαρακτηριστικά με δυαδικό πεδίο ορισμού. Υπάρχουν 2^m διαφορετικές τιμές που μπορεί να πάρει το x_k, χρειάζεται επομένως ο υπολογισμός 2^m πιθανοτήτων. Ο αριθμός αυτός θα είναι πολύ μεγαλύτερος από το πλήθος των διαθέσιμων παραδειγμάτων εκτός και αν το m είναι πολύ μικρό.

Τέλος, ένα άλλο πρόβλημα που παρουσιάζεται σε αυτή τη μέθοδο είναι ο μεγάλος χρόνος που χρειάζεται προκειμένου να γίνουν οι διάφοροι υπολογισμοί, με αποτέλεσμα να την καθιστά ιδιαίτερη αργή.

Παρόλο που το κριτήριο της μέγιστης εξάρτησης είναι καλό στις περιπτώσεις όπου απαιτείται η επιλογή λίγων χαρακτηριστικών από ένα μεγάλο σύνολο, δεν είναι κατάλληλο για τις περιπτώσεις όπου χρειάζεται να επιτευχθεί ταξινόμηση με υψηλή ακρίβεια. Για αυτούς ακριβώς τους λόγους μια εναλλακτική πρόταση, είναι ο συνδυασμός του κριτηρίου της μέγιστης συνάφειας (max relevance) και του ελάχιστου πλεονασμού (min redundancy), προκειμένου να κατασκευαστεί μια νέα μέθοδος πιο αποτελεσματική, η mRMR.

Εφόσον η αμοιβαία πληροφορία $I(S_m;c)$ μεταξύ των χαρακτηριστικών και της κλάσης, δεν μπορεί να υπολογιστεί, πρέπει να χρησιμοποιηθεί κάποια άλλη προσέγγιση ώστε να βρεθεί και να επιλεγεί το υποσύνολο από το οποίο η κλάση να έχει τη μεγαλύτερη εξάρτηση. Η μονοπαραγοντική (univariate) προσέγγιση στο πρόβλημα αυτό είναι να υπολογιστεί η εξάρτηση της κλάσης από κάθε χαρακτηριστικό ξεχωριστά και εν συνεχεία να επιλεγούν τα m χαρακτηριστικά από τα οποία υπάρχει η μεγαλύτερη εξάρτηση. Δηλαδή στόχος είναι να βρεθεί ένα υποσύνολο S_p μέσω της σχέσης

$$S_p = \arg\max \{ \sum_{x_i \in S_m} I(x_i;c) \}. \qquad 4.3$$

Το βασικό πρόβλημα της μονοπαραγοντικής αυτής προσέγγισης είναι ότι επιλέγεται μεγάλος αριθμός περιττών χαρακτηριστικών. Αν ένα χαρακτηριστικό επιλέγεται γιατί έχει υψηλή αμοιβαία πληροφορία με την κλάση, τότε χαρακτηριστικά πολύ όμοια με αυτό θα έχουν επίσης υψηλή αμοιβαία πληροφορία με την κλάση και θα επιλεγούν και αυτά. Όμως ένα σύνολο που αποτελείται από χαρακτηριστικά που έχουν πολλές ομοιότητες μεταξύ τους προσφέρει λίγη μόνο περισσότερη πληροφορία για την κλάση από αυτήν που προσφέρουν μερικά μόνο χαρακτηριστικά του συνόλου. Με άλλα λόγια η διακριτική δύναμη της κλάσης δεν θα άλλαζε και πολύ αν μερικά από αυτά τα χαρακτηριστικά αφαιρεθούν από το σύνολο επιλεγμένων χαρακτηριστικών. Ένα ακραίο αλλά όχι απίθανο παράδειγμα είναι η επιλογή δύο χαρακτηριστικών που έχουν ίδιες μεταξύ τους τιμές σε κάθε παράδειγμα του συνόλου εκπαίδευσης.

Ίσως να είναι χρήσιμη η επιλογή ενός εκ των δύο αλλά η γνώση του δεύτερου δεν προσφέρει κανένα κέρδος σε πληροφορία. Ένα καλύτερο υποσύνολο χαρακτηριστικών μπορεί να προκύψει αν επιλέγονται χαρακτηριστικά που έχουν μεν μεγάλη συνάφεια με την κλάση, αλλά την ίδια στιγμή είναι μεταξύ τους όσο το δυνατόν περισσότερο ανόμοια.

Δυστυχώς, δεν υπάρχει κάποιος προφανής τρόπος για το πώς μπορεί να μετρηθεί ο βαθμός στον οποίο τα χαρακτηριστικά του υποσυνόλου S είναι μεταξύ τους όμοια. Σαν εναλλακτική λύση χρησιμοποιείται η μέση τιμή της ομοιότητας μεταξύ όλων των πιθανών ζευγών από χαρακτηριστικά του S, η οποία μετριέται με την μέση τιμή της αμοιβαίας πληροφορίας που έχουν αυτά τα ζεύγη. Η ποσότητα αυτή θα αναφέρεται στο εξής ως περιττή πληροφορία (redundancy) του S, και σε αυτήν την ιδέα βασίζεται ουσιαστικά η μέθοδος mRMR.

Όπως προαναφέρθηκε η μέθοδος mRMR, δεν βασίζεται στην επιλογή χαρακτηριστικών με βάση την ανεξαρτησία που υπάρχει μεταξύ τους. Αντιθέτως προσπαθεί να επιλέξει χαρακτηριστικά τα οποία ελαχιστοποιούν τον πλεονασμό και παράλληλα μεγιστοποιούν τη συνάφεια που υπάρχει μεταξύ αυτών και της κλάσης. Άλλωστε στη πράξη, για πραγματικά δεδομένα, συνήθως ένα σύνολο από χαρακτηριστικά τελείως ανεξάρτητα μεταξύ τους δεν οδηγεί και σε τόσο καλά αποτελέσματα. Αντίθετα, είναι δυνατόν, χαρακτηριστικά με από κοινού επίδραση να οδηγούν σε πολύ καλά αποτελέσματα. Επιπλέον, είναι δυνατόν να μειωθεί άμεσα ο πλεονασμός των χαρακτηριστικών, απλά υπολογίζοντας την αμοιβαία πληροφορία μεταξύ των συνεχών χαρακτηριστικών μεταβλητών.

Συγκεκριμένα, η μέθοδος mRMR εφαρμόζει μια στοιχειώδη επιλογή πρώτης τάξης, προκειμένου να δημιουργήσει ένα υποψήφιο σύνολο χαρακτηριστικών, κάτι που διευκολύνει ουσιαστικά την εφαρμογή σε αυτό το σύνολο των μεθόδων wrapper, προκειμένου να καταλήξουν σε συμπαγή υποσύνολα χαρακτηριστικών με μια βέλτιστη ακρίβεια στη ταξινόμηση. Ο mRMR είναι ιδιαίτερα χρήσιμος, είτε στην περίπτωση που έχουμε να κάνουμε με χαρακτηριστικά μεγάλου μεγέθους, είτε όταν έχουμε να αντιμετωπίσουμε προβλήματα επιλογής μεταβλητών στα οποία υπάρχουν χιλιάδες υποψήφια χαρακτηριστικά.

Η μέθοδος mRMR θέτει δύο συνθήκες οι οποίες πρέπει να ικανοποιούνται από ένα υποσύνολο χαρακτηριστικών:

1. Τα χαρακτηριστικά του υποσυνόλου πρέπει να έχουν όσο το δυνατόν μεγαλύτερη συνάφεια με την κατηγορία (max-relevance).
2. Τα χαρακτηριστικά του υποσυνόλου πρέπει να είναι όσο το δυνατόν ανόμοια μεταξύ

τους (min-redundancy).

Η καλύτερη προσέγγιση της $D(S,c)$ γίνεται μέσω του κριτηρίου της μέγιστης συνάφειας, το οποίο βασίζεται στον υπολογισμό της μέσης τιμής όλων των αμοιβαίων πληροφοριών, μεταξύ των χαρακτηριστικών x_i και της κλάσης c. Δηλαδή η μέγιστη συνάφεια $\max D(S,c)$ υπολογίζεται από την εξίσωση:

$$\max D(S,c),\ D(S,c)=\frac{1}{|S|}\cdot\sum_{x_i\in S_m} I(x_i;c), \qquad 4.4$$

όπου $S=\{x_1,x_2,\ldots,x_m\}$ είναι ένα υποσύνολο χαρακτηριστικών και c η κλάση.

Όσον αφορά την ελαχιστοποίηση των όμοιων χαρακτηριστικών, προκειμένου να επιτευχθεί ο ελάχιστος πλεονασμός, εκτιμάται η ομοιότητα μεταξύ δύο χαρακτηριστικών x_i και x_j χρησιμοποιώντας την αμοιβαία πληροφορία $I(x_i,x_j)$ και ελαχιστοποιώντας την ποσότητα $R(S)$ που δίνεται από τη σχέση:

$$\min R(S),\ R(S)=\frac{1}{|S|^2}\cdot\sum_{x_i,x_j\in S} I(x_i,x_j). \qquad 4.5$$

Η μέθοδος mRMR είναι στην ουσία ο συνδυασμός των δύο πιο πάνω κριτηρίων. Στόχος της είναι η εύρεση ενός υποσυνόλου χαρακτηριστικών το οποίο μεγιστοποιεί τον τελεστή $\Phi(D,R)$ που συνδυάζει την ελαχιστοποίηση της περιττής πληροφορίας $R(S)$ και ταυτόχρονα την μεγιστοποίηση της συνάφειας $D(S)$. Κατά κανόνα η αύξηση της συνάφειας συνοδεύεται από αύξηση της περιττής πληροφορίας και έτσι δεν υπάρχει μοναδικό υποσύνολο που να υπερέχει έναντι των άλλων με βάση και τα δύο κριτήρια. Ο τελεστής $\Phi(D,R)$ ορίζεται μέσω της ακόλουθης σχέσης:

$$\max \Phi(D,R),\ \Phi(D,R)=D(S,c)-R(S). \qquad 4.6$$

Στην πράξη, προκειμένου η διαδικασία να γίνει υπολογιστικά εφικτή, το υποσύνολο σχηματίζεται χρησιμοποιώντας μία προς τα εμπρός μέθοδο άπληστης αναζήτησης (forward selection). Αρχικά επιλέγεται το πιο συναφές χαρακτηριστικό (αυτό με τη μεγαλύτερη αμοιβαία πληροφορία με την κλάση). Στη συνέχεια, σε κάθε επανάληψη επιλέγεται ένα χαρακτηριστικό που έχει μεγάλη συνάφεια με την κλάση και ταυτόχρονα μικρή ομοιότητα με τα ήδη επιλεγμένα χαρακτηριστικά.

Τροποποιώντας κατάλληλα την πιο πάνω εξίσωση δημιουργείται μία συνάρτηση που

αντί να αξιολογεί υποσύνολα χαρακτηριστικών, αξιολογεί τα χαρακτηριστικά μεμονωμένα. Έστω ότι στην m-οστή επανάληψη έχουν επιλεχθεί ήδη τα πρώτα m-1 χαρακτηριστικά και έχει σχηματιστεί το υποσύνολο S_{m-1}. Επιλέγεται ένα επιπλέον χαρακτηριστικό, το m-οστό, από το σύνολο $X-S_{m-1}$ το οποίο έχει μεγάλη συνάφεια με την κλάση και ταυτόχρονα μικρή ομοιότητα με τα ήδη m-1 επιλεγμένα χαρακτηριστικά, μεγιστοποιώντας επί της ουσίας την συνάρτηση $\Phi(D,R)$. Τότε αυτό μπορεί να βρεθεί μεγιστοποιώντας την ακόλουθη τροποποιημένη συνάρτηση:

$$\max_{x_j \in X-S_{m-1}} \left[I(x_j;c) - \frac{1}{m-1} \cdot \sum_{x_i \in S_{m-1}} I(x_j;x_i) \right], \qquad 4.7$$

του οποίου η υπολογιστική πολυπλοκότητα είναι $O(|S| \cdot M)$.

Ένας ενδεικτικός αλγόριθμος που να συνοψίζει τη μέθοδο mRMR, όπου στόχος είναι η δημιουργία ενός συνόλου S που να αποτελείται από m βέλτιστα χαρακτηριστικά ακολουθεί παρακάτω:

Algorithm mRMR

Input: data set of observations X and the corresponding labels Y (εισαγωγή των δειγμάτων X με όλα τα χαρακτηριστικά και της αντίστοιχης κλάσης Y)

Output: final subset of feature set S (τελικό υποσύνολο χαρακτηριστικών)

$S \leftarrow \{\ \}$ (ανάθεση του υποσυνόλου επιλεγμένων χαρακτηριστικών ως το κενό)

$X^* \leftarrow \max\{I(X;c)\}$ (εύρεση του χαρακτηριστικού με τη μέγιστη αμοιβαία πληροφορία με την κλάση c)

$S \leftarrow S \cup X^*$ (εισαγωγή αυτού στο υποσύνολο επιλεγμένων χαρακτηριστικών)

For $m \geq 2$ **to** $m = M$ (έλεγχος συνθήκης τερματισμού)

$X^* = \max_{x_j \in X-S_{m-1}} \left[I(x_j;c) - \frac{1}{m-1} \cdot \sum_{x_i \in S_{m-1}} I(x_j;x_i) \right]$ (εύρεση του νέου χαρακτηριστικού με την μέγιστη τροποποιημένη συνάρτηση)

$S \leftarrow S \cup X^*$ (εισαγωγή του νέου χαρακτηριστικού στο υποσύνολο επιλεγμένων χαρακτηριστικών)

end for

4.3.2 Ισοδυναμία mRMR και κριτηρίου μέγιστης εξάρτησης

Όταν επιλέγεται ένα χαρακτηριστικό κάθε φορά, δηλαδή η επιλογή είναι διαδοχική πρώτης τάξεως (first order), τότε η μέθοδος mRMR είναι ισοδύναμη με το κριτήριο της μέγιστης εξάρτησης.

Πράγματι, έστω ότι έχει ήδη επιλεγεί το σύνολο S_{m-1}, δηλαδή το σύνολο των $m-1$ χαρακτηριστικών που είναι πιο σημαντικά και αναζητά κάποιος το m-οστό χαρακτηριστικό από το σύνολο $X-S_{m-1}$. Εφόσον, η εξάρτηση D αντιπροσωπεύεται από την αμοιβαία πληροφορία, δηλαδή $D=I(S_m;c)$, όπου $S_m=\{S_{m-1},x_m\}$ είναι μια πολυμεταβλητή, από τον ορισμό της αμοιβαίας πληροφορίας θα ισχύει:

$$\begin{aligned} I(S_m;c) &= H(c)+H(S_m)-H(S_m,c) \\ &= H(c)+H(S_{m-1},x_m)-H(S_{m-1},x_m,c), \end{aligned} \tag{4.8}$$

όπου $H(\cdot)$ είναι η εντροπία των αντίστοιχων μεταβλητών.

Επιπρόσθετα ορίζεται η ποσότητα $J(S_m)=J(x_1,x_2,\ldots,x_m)$ να είναι η παρακάτω:

$$J(x_1,\ldots,x_m)=\int\ldots\int p(x_1,\ldots,x_m)\cdot\log\frac{p(x_1,\ldots,x_m)}{p(x_1)\ldots p(x_m)}dx_1\ldots dx_m. \tag{4.9}$$

Παρομοίως ορίζεται και η ποσότητα $J(S_m,c)=J(x_1,\ldots,x_m,c)$ ως:

$$J(x_1,\ldots,x_m,c)=\int\ldots\int p(x_1,\ldots,x_m,c)\cdot\log\frac{p(x_1,\ldots,x_m,c)}{p(x_1)\ldots p(x_m)\cdot p(x_m)}dx_1\ldots dx_m dc. \tag{4.10}$$

Από τις δύο παραπάνω ποσότητες προκύπτει άμεσα ότι:

$$H(S_{m-1},x_m)=H(S_m)=\sum_{i=1}^{m}H(x_i)-J(S_m), \tag{4.11}$$

$$H(S_{m-1},x_m,c)=H(S_m,c)=H(c)+\sum_{i=1}^{m}H(x_i)-J(S_m,c). \tag{4.12}$$

Αντικαθιστώντας, τους δύο τελευταίους όρους, με τους αντίστοιχους στην εξίσωση της αμοιβαίας πληροφορίας (4.8) προκύπτει ότι:

$$\begin{aligned} I(S_m;c) &= J(S_m,c)-J(S_m) \\ &= J(S_{m-1},x_m,c)-J(S_{m-1},x_m). \end{aligned} \qquad 4.13$$

Προφανώς η μέγιστη εξάρτηση, είναι ισοδύναμη με την πιο πάνω εξίσωση, μεγιστοποιώντας το πρώτο όρο και ελαχιστοποιώντας συγχρόνως τον δεύτερο.

4.3.3 Υπολογισμός κατώτατου ορίου του δεύτερου όρου

Θεωρώντας ότι ισχύει η ανισότητα $\log(z)\le z-1$, η σχέση (4.9) μπορεί να γραφτεί ως:

$$\begin{aligned} -J(x_1,\ldots,x_m) &= \int\ldots\int p(x_1,\ldots,x_m)\cdot\log\frac{p(x_1)\ldots p(x_m)}{p(x_1,\ldots,x_m)}dx_1\ldots dx_m \\ &\le \int\ldots\int p(x_1,\ldots,x_m)\cdot\left[\frac{p(x_1)\ldots p(x_m)}{p(x_1,\ldots,x_m)}-1\right]dx_1\ldots dx_m \\ &= \int\ldots\int p(x_1)\ldots p(x_m)dx_1\ldots dx_m-\int\ldots\int p(x_1,\ldots,x_m)dx_1\ldots dx_m \\ &= 1-1=0. \end{aligned} \qquad 4.14$$

Είναι προφανές ότι η ελάχιστη τιμή επιτυγχάνεται όταν όλες οι μεταβλητές είναι ανεξάρτητες μεταξύ τους, όπου θα ισχύει ότι $p(x_1,\ldots,x_m)=p(x_1)\ldots p(x_m)$. Αφού έχουν ήδη επιλεγεί m-1 χαρακτηριστικά, αυτή η συνθήκη ανεξαρτησίας σημαίνει ότι η αμοιβαία πληροφορία μεταξύ του x_m και οποιουδήποτε χαρακτηριστικού x_i, $(i=1,\ldots,m-1)$ θα είναι η ελάχιστη. Αυτό είναι και το κριτήριο του ελάχιστου πλεονασμού.

4.3.4 Εύρεση άνω ορίου του πρώτου όρου

Για τον όρο $J(x_1,\ldots,x_m)$ ισχύει η παρακάτω ανισότητα:

$$
\begin{aligned}
J(x_1,\dots,x_m) &= \int \dots \int p(x_1,\dots,x_m)\cdot \log \frac{p(x_1,\dots,x_m)}{p(x_1)\dots p(x_m)}\, dx_1 \dots dx_m \\
&= \int \dots \int p(x_1,\dots,x_m)\cdot \frac{p(x_1|x_2,\dots,x_m)p(x_2|x_3,\dots,x_m)\dots p(x_{m-1}|x_m)p(x_m)}{p(x_1)\dots p(x_{m-1})p(x_m)}\, dx_1 \dots dx_m \\
&= \sum_{i=1}^{m-1} H(x_i) - H(x_1|x_2,\dots,x_m) - H(x_2|x_3,\dots,x_m) - H(x_{m-1}|x_m) \\
&\le \sum_{i=1}^{m-1} H(x_i). \qquad 4.15
\end{aligned}
$$

Η παραπάνω σχέση μπορεί να αναπτυχθεί ισοδύναμα ως:

$$
J(x_1,\dots,x_m) \le \min\left\{ \sum_{i=2}^{m} H(x_i), \sum_{i=1, i\neq 2}^{m} H(x_i), \dots, \sum_{i=1, i\neq m-1}^{m} H(x_i), \sum_{i=1}^{m-1} H(x_i) \right\}. \qquad 4.16
$$

Είναι αρκετά εύκολο να διαπιστώσει κάποιος ότι το μέγιστο για τον πρώτο όρο της εξίσωσης (4.13), $J(S_{m-1},\dots,x_m,c)$, επιτυγχάνεται όταν όλες οι μεταβλητές είναι στο μέγιστο εξαρτημένες μεταξύ τους. Αυτό σημαίνει ότι όταν το σύνολο S_{m-1} είναι καθορισμένο, η x_m θα έχει τη μέγιστη δυνατή εξάρτηση με την κατηγορία c. Αυτό είναι και το κριτήριο της μέγιστης συνάφειας (max-relevance).

4.3.5 Διαφορές των δύο μεθόδων

Παρόλο που μαθηματικά μπορεί κάποιος να καταλήξει μέσω της μιας μεθόδου στη άλλη, εντούτοις αυτές παρουσιάζουν κάποιες διαφορές. Καταρχήν, κάποιος μπορεί να οδηγηθεί στη μέγιστη εξάρτηση μέσω της μεγιστοποίησης του $J(S_m,c)$. Η διαφορά μεταξύ της μεθόδου mRMR και του κριτηρίου της μέγιστης εξάρτησης είναι ότι ενώ στη μέγιστη εξάρτηση η συσχέτιση μεταξύ των κατανεμημένων δεδομένων στο υποσύνολο R^m και της κλάσης c έπαιζε πολύ σημαντικό ρόλο, κάτι τέτοιο δεν συμβαίνει στη μέθοδο mRMR, κάτι που φαίνεται συγκρίνοντας την εξίσωση (4.10) με τις εξισώσεις (4.1) και (4.2). Επιπλέον, δεν μπορεί να παραγνωριστεί το γεγονός ότι, ενώ στη μέγιστη εξάρτηση υπολογίζονται οι ποσότητες $p(x_1,\dots,x_m)$ και $p(x_1,\dots,x_m,c)$, η mRMR αποφεύγει κάτι τέτοιο υπολογίζοντας μόνο τις ποσότητες $p(x_i,x_j)$ και $p(x_i,c)$, κάτι το οποίο είναι

προφανώς πολύ πιο εύκολο να γίνει και επιπλέον δίνει πολύ πιο ακριβή αποτελέσματα.

4.3.6 Αλγόριθμοι επιλογής χαρακτηριστικών δύο σταδίων

Όπως έχει αναφερθεί, ένα από τα κυριότερα προβλήματα που υπάρχει σε ένα πρόβλημα επιλογής του καλύτερου υποσυνόλου έχει να κάνει με τον καθορισμό του βέλτιστου αριθμού των χαρακτηριστικών που θα πρέπει να περιέχει το τελικό υποσύνολο.

Η μέθοδος mRMR που μόλις παρουσιάστηκε είναι μια μέθοδος στοιχειώδους πρώτης επιλογής χαρακτηριστικών, χωρίς την δυνατότητα να αφαιρεί περιττά χαρακτηριστικά από αυτά που έχει ήδη επιλέξει. Αυτό όμως που μπορεί να κάνει είναι να συνδυάζεται μαζί με άλλες μεθόδους επιλογής χαρακτηριστικών, οι οποίες σε δεύτερο στάδιο θα μπορούσαν να κάνουν αυτή τη δουλειά.

Συγκεκριμένα, στο πρώτο στάδιο με τη χρήση της μεθόδου mRMR, καθορίζεται ένα υποψήφιο σύνολο χαρακτηριστικών. Με τη βοήθεια του μοντέλου επαλήθευσης cross-validation, υπολογίζεται το σφάλμα στη ταξινόμηση και στη συνέχεια απομονώνεται μια σταθερή περιοχή Ω στην οποία το σφάλμα είναι μικρό. Ο βέλτιστος αριθμός των χαρακτηριστικών, που θα έχει το υποσύνολο θα εξαρτάται από αυτή τη περιοχή Ω. Συγκεκριμένα ακολουθούνται τα παρακάτω βήματα:

i. Με τη χρήση της μεθόδου mRMR, επιλέγονται n διαδοχικά χαρακτηριστικά από ένα σύνολο δεδομένων X. Αυτή η επιλογή, οδηγεί σε m διαδοχικά υποσύνολα χαρακτηριστικών, έστω τα $S_1 \subset S_2 \subset ... \subset S_{m-1} \subset S_m$.

ii. Γίνεται σύγκριση όλων των υποσυνόλων χαρακτηριστικών που επιλέχθηκαν μεταξύ τους, με σκοπό τη δημιουργία ενός νέου συνόλου το οποίο θα αποτελείται από αυτά τα υποσύνολα, στα οποία το σφάλμα είναι μικρό (δηλ. έχει μικρή μέση τιμή και μικρή διασπορά). Έστω ότι το νέο σύνολο υποσυνόλων το οποίο ονομάζεται Ω, θα αποτελείται από k υποσύνολα.

iii. Από τα υποσύνολα που ανήκουν στο Ω, υπολογίζεται αυτό με το μικρότερο σφάλμα ταξινόμησης e_k. Το υποψήφιο σύνολο χαρακτηριστικών που αναζητείται, προκειμένου να θεωρείται βέλτιστο, θα πρέπει να έχει μέγεθος ίσο με το υποσύνολο

k για το οποίο αντιστοιχεί το μικρότερο σφάλμα.

Υπάρχουν αρκετά εξεζητημένες μέθοδοι, οι οποίες θα μπορούσαν να συνδυαστούν με τη μέθοδο mRMR σε δεύτερη φάση, προκειμένου, να δημιουργήσουν συμπαγή σύνολα χαρακτηριστικών, αφαιρώντας χαρακτηριστικά από τα ήδη επιλεγμένα. Συγκεκριμένα, με τη χρήση της διαδικασίας mRMR σε πρώτο στάδιο, δημιουργείται ένα μικρό σύνολο από υποψήφια χαρακτηριστικά, για τα οποία στη συνέχεια εφαρμόζονται σε αυτά άλλες μέθοδοι, οι οποίες είναι συνήθως wrapper, προκειμένου να βελτιωθεί το αποτέλεσμα. Αναφέρεται ότι δύο από τους κύριους αλγορίθμους που μπορούν να χρησιμοποιηθούν, είναι οι wrapper μέθοδοι της προς τα πίσω επιλογής (backward elimination) και της προς τα εμπρός επιλογής (forward selection).

Ο λόγος που χρησιμοποιούνται wrapper μέθοδοι σε δεύτερο στάδιο έχει να κάνει με την ίδια την λειτουργία της mRMR μεθόδου. Είναι πολλές φορές πιθανόν τα διάφορα χαρακτηριστικά που επιλέγονται από τη μέθοδο mRMR, να μην οδηγούν σε τόσο καλά αποτελέσματα, δηλαδή να μην καταφέρνουν να μειώσουν σημαντικά το σφάλμα στη ταξινόμηση ή ακόμη και να μην μπορούν να το περιορίσουν καθόλου. Υπάρχουν ακόμη και περιπτώσεις όπου η επιλογή ενός χαρακτηριστικού να οδηγεί σε αύξηση του σφάλματος, δηλαδή να υπάρχουν διακυμάνσεις στο σφάλμα όσο αυξάνεται ο αριθμός των χαρακτηριστικών που επιλέγεται. Αυτό μπορεί να οφείλεται σε πολλούς λόγους. Μια πιθανή αιτία, είναι το γεγονός ότι κάποια από τα πρόσθετα χαρακτηριστικά είναι “noisy”, δηλαδή παράγουν θόρυβο. Μια δεύτερη πιθανή αιτία, αφορά τη μέθοδο “cross-validation” που χρησιμοποιείται και η οποία είναι δυνατό να ευθύνεται και αυτή για τις διακυμάνσεις που υπάρχουν πάνω στη καμπύλη του σφάλματος. Τέλος μια ακόμα σημαντική πιθανή αιτία, είναι το γεγονός ότι η μέθοδος mRMR χρησιμοποιεί ως κριτήριο επιλογής χαρακτηριστικών την εξίσωση $\Phi = D - R$, η οποία από τους όρους που επιλέγονται με τη μέγιστη συνάφεια, αφαιρεί τον πλεονασμό τους. Αυτό μπορεί να οδηγήσει σε φαινόμενα κατά τα οποία ένα χαρακτηριστικό το οποίο να είναι να ουσιαστικά περιττό, να παρουσιάζει ταυτόχρονα μεγάλο ενδιαφέρον όσο αφορά τη συνάφεια του με αποτέλεσμα να επιλέγεται τελικά ανάμεσα στα κορυφαία χαρακτηριστικά. Ακριβώς λόγω του πιο πάνω φαινομένου, για να περιοριστεί το σφάλμα στη ταξινόμηση, χρησιμοποιούνται σε 2^{η} φάση και άλλες μέθοδοι επιλογής χαρακτηριστικών, προκειμένου να αφαιρέσουν όσα χαρακτηριστικά επιλέχθηκαν σε

πρώτη φάση και τα οποία δεν χρειάζονται είτε γιατί είναι περιττά είτε γιατί δεν προσφέρουν κάτι ουσιαστικό στη μείωση του σφάλματος.

ΚΕΦΑΛΑΙΟ 5

Εφαρμογή εξόρυξης γνώσης σε πραγματικά δεδομένα

5.1 Εισαγωγή

Η παρούσα ενότητα αποτελεί το κύριο τεχνικό κομμάτι της διατριβής, καθώς εγκολπώνει την πρακτική εφαρμογή των μεθόδων εξόρυξης δεδομένων που περιγράφηκαν στο Κεφάλαιο 2 καθώς και της μεθόδου επιλογής χαρακτηριστικών που αναλύθηκε στο Κεφάλαιο 4 για την επίλυση ενός πραγματικού προβλήματος δυαδικής ταξινόμησης. Συγκεκριμένα, γίνεται χρήση ενός ελεύθερα διαθέσιμου στο Διαδίκτυο συνόλου δεδομένων από τον τραπεζοοικονομικό τομέα, για την εκπαίδευση και τη συγκριτική αξιολόγηση διαφόρων μοντέλων μηχανικής μάθησης, εκ των οποίων καλούμαστε να επιλέξουμε το βέλτιστο, με κύριο γνώμονα την ακρίβεια ταξινόμησης και δευτερεύοντες την πολυπλοκότητα και τον χρόνο εκτέλεσης των παραγόμενων αλγορίθμων. Ο συνδυασμός των μοντέλων αυτών με μεθόδους επιλογής χαρακτηριστικών οδηγούμενων από την πληροφορία, αποτελεί το καινοτόμο χαρακτηριστικό της εργασίας, καθώς στοχεύει στην επίτευξη ίσης ή ακόμη και μεγαλύτερης ακρίβειας ταξινόμησης, με χρήση ενός υποσυνόλου των διαθέσιμων χαρακτηριστικών που περιλαμβάνονται στο σύνολο δεδομένων.

Στις ακόλουθες υποενότητες γίνεται περιγραφή του συνόλου δεδομένων και των επιμέρους χαρακτηριστικών του, ενώ δίνεται έμφαση στην περιγραφή της μεθοδολογίας ανάλυσης, η οποία ενσωματώνει όλα τα στάδια της εξόρυξης δεδομένων που περιγράφηκαν στο Κεφάλαιο 1, και συγκεκριμένα το κομμάτι συλλογής, της προεπεξεργασίας, του μετασχηματισμού, της ανάλυσης με χρήση μεθόδων μηχανικής μάθησης και τελικά της αξιολόγησης και οπτικοποίησης των αποτελεσμάτων. Στην

ανάλυση έχει περιληφθεί επίσης η χρήση προς τα εμπρός επιλογής χαρακτηριστικών, ώστε να μελετηθεί η επίδρασή τους στην ακρίβεια ταξινόμησης. Η τελευταία υποενότητα είναι αφιερωμένη στην αλγοριθμική διαδικασία, όπου σχολιάζονται αναλυτικά τα προαναφερθέντα βήματα σε επίπεδο κώδικα τα οποία οδήγησαν στην επιλογή του βέλτιστου μοντέλου.

5.2 Περιγραφή του συνόλου δεδομένων

Στα πλαίσια της μεταπτυχιακής διατριβής έγινε χρήση του συνόλου δεδομένων “default of credit card clients” το οποίο διατίθεται ελεύθερα στο Διαδίκτυο από το University of California, Irvine (UCI)[1]. Το dataset αφορά την περίπτωση αδυναμίας αποπληρωμής των δόσεων πιστωτικών καρτών από πελάτες τραπεζών της Ταϊβάν.

Την προηγούμενη δεκαετία, το τραπεζικό σύστημα της χώρας σημαδεύτηκε από μια πρωτόγνωρη οικονομική κρίση, η οποία αποδόθηκε στην υπερ-έκδοση πιστωτικών καρτών σε αιτούντες που δεν πληρούσαν τα ανάλογα οικονομικά κριτήρια, γεγονός που οδήγησε σε μαζικές αδυναμίες πληρωμής. Ταυτόχρονα, οι περισσότεροι χρήστες καρτών παρασύρθηκαν σε μαζικές αγορές, χωρίς να νοιάζονται για την ενδεχόμενη αδυναμία εκπλήρωσης των πιστωτικών τους υποχρεώσεων, προκαλώντας σημαντικό πλήγμα τόσο στις τράπεζες όσο και στην καταναλωτική κοινότητα. Η παντελής αδυναμία πρόβλεψης του οικονομικού αυτού αδιεξόδου από το χρηματοπιστωτικό σύστημα της χώρας συνετέλεσε στην ανάδειξη της Επιστήμης των Δεδομένων ως ένα πολύτιμο εργαλείο ανάλυσης, πρόβλεψης και μείωσης της αβεβαιότητας σε ένα πλήρως πραγματικό πρόβλημα, αυτό της διαχείρισης οικονομικού κινδύνου (Yeh και Lien, 2009).

Το συγκεκριμένο σύνολο δεδομένων είναι πολυμεταβλητό (Multivariate) και περιέχει ακέραιες, πραγματικές τιμές στα χαρακτηριστικά του. Η εξαρτημένη μεταβλητή του συνόλου (default.payment.next.month) παίρνει δύο πιθανές τιμές: 1 και 0 οι οποίες κατά σειρά αντιστοιχούν στην αδυναμία πληρωμής τον επόμενο μήνα από τους πελάτες ή μη. Το πλήθος των παρατηρήσεων (instances) που έχουν καταχωρηθεί στο προς μελέτη

1 https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

σύνολο είναι 30.000, ενώ το πλήθος των χαρακτηριστικών που έχουν καταγραφεί για κάθε μια από αυτές είναι 24, συμπεριλαμβανομένης της μεταβλητής απόκρισης. Τέλος αξίζει να σημειωθεί ότι στην περιγραφή του συγκεκριμένου dataset αναφέρεται ότι δεν υπάρχουν ελλείπουσες τιμές (Πίνακας 5.1).

ΠΙΝΑΚΑΣ 5.1 Χαρακτηριστικά Συνόλου Δεδομένων

Χαρακτηριστικά Συνόλου :	Πολυμεταβλητό	**Πλήθος Δειγμάτων :**	30.000
Ιδιότητες Χαρακτηριστικών :	Ακέραιοι, Πραγματικοί	**Πλήθος Χαρακτηριστικών :**	24
Σχετικές εργασίες :	Ταξινόμηση	**Ελλείπουσες τιμές :**	Καμία
Πεδίο :	Τομέας Επιχειρήσεων		

Η γραμμογράφηση των 23 επεξηγηματικών μεταβλητών του συνόλου με όλες τις πιθανές τιμές που μπορούν αυτές να έχουν παρατίθεται αναλυτικά παρακάτω:

- **X1:** Ποσό της δεδομένης πίστωσης (σε New Taiwan dollar): Περιλαμβάνει τόσο την ατομική καταναλωτική πίστωση όσο και την συμπληρωματική πίστωση της οικογένειάς του.
- **X2:** Φύλο καταναλωτή (1 = αρσενικό, 2 = θηλυκό).
- **X3:** Μορφωτικό επίπεδο καταναλωτή (0: άγνωστο, 1 = μεταπτυχιακή εκπαίδευση, 2 = πανεπιστημιακή εκπαίδευση, 3 = τριτοβάθμια εκπαίδευση, 4 = άλλα, 5: άγνωστο, 6: άγνωστο).
- **X4:** Οικογενειακή κατάσταση (0: άγνωστο, 1 = παντρεμένος, 2 = ελεύθερος, 3 = άλλο).
- **X5:** Ηλικία (σε έτη).
- **X6 – X11:** Ιστορικό προηγούμενης πληρωμής. Αναφέρεται στα προηγούμενα μηνιαία αρχεία πληρωμών (από τον Απρίλιο έως τον Σεπτέμβριο του 2005) ως εξής: X6 = κατάσταση αποπληρωμής τον Σεπτέμβριο του 2005, X7 = κατάσταση αποπληρωμής τον Αύγουστο του 2005, ... , X11 = κατάσταση αποπληρωμής τον Απρίλιο του 2005. Η κλίμακα μέτρησης για την κατάσταση αποπληρωμής είναι: -1 = πλήρης καταβολή, 1 = καθυστέρηση πληρωμής για ένα μήνα, ... , 9 = καθυστέρηση πληρωμής για εννέα μήνες και άνω.
- **X12 – X17:** Ποσό λογαριασμού (σε New Taiwan dollar): X12 = ποσό λογαριασμού τον Σεπτέμβριο του 2005, ... , X17 = ποσό λογαριασμού τον Απρίλιο του 2005.

- **X18 – X23:** Ποσό προηγούμενης πληρωμής (New Taiwan dollar): X18 = ποσό που καταβλήθηκε τον Σεπτέμβριο του 2005, … , X23 = ποσό που καταβλήθηκε τον Απρίλιο του 2005.

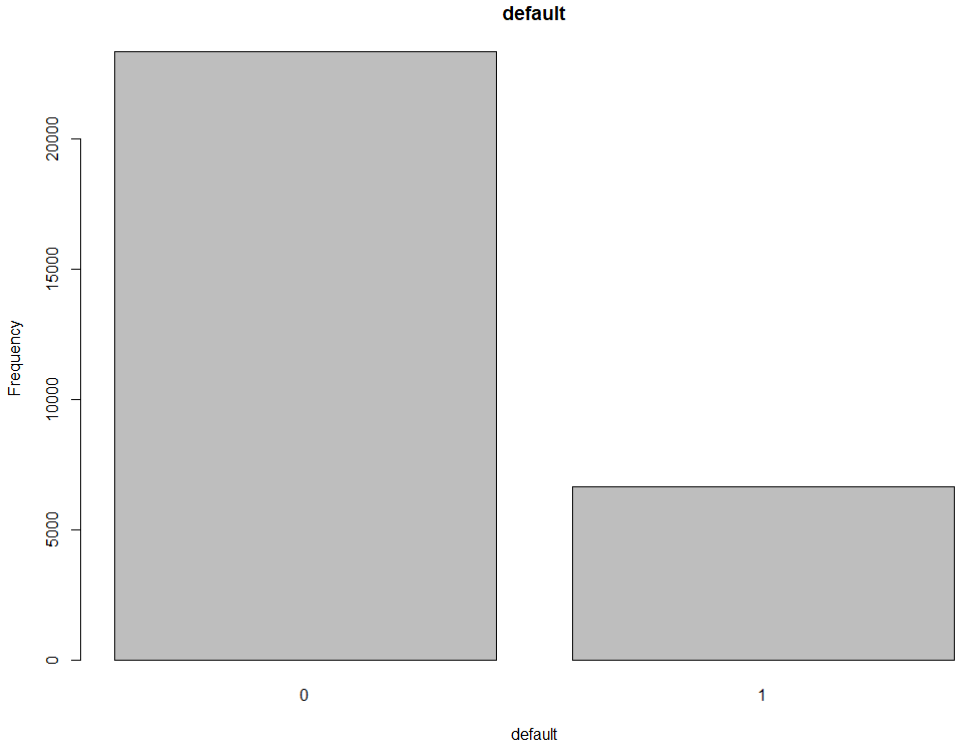
Επιπλέον παρακάτω παρατίθενται και οι πρώτες δέκα γραμμές και από τις 24 μεταβλητές αυτού του συνόλου όπως παράχθηκαν από την R.

```
'data.frame':  30000 obs. of  24 variables:
 $ default  : int  1 1 0 0 0 0 0 0 0 0 ...
 $ LIMIT_BAL: int  20000 120000 90000 50000 50000 50000 500000 100000
140000 20000 ...
 $ SEX      : int  2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION: int  2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE : int  1 2 2 1 1 2 2 2 1 2 ...
 $ AGE      : int  24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0    : int  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2    : int  2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3    : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4    : int  -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5    : int  -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6    : int  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1: int  3913 2682 29239 46990 8617 64400 367965 11876 11285
0 ...
 $ BILL_AMT2: int  3102 1725 14027 48233 5670 57069 412023 380 14096
0 ...
 $ BILL_AMT3: int  689 2682 13559 49291 35835 57608 445007 601 12108
0 ...
 $ BILL_AMT4: int  0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
 $ BILL_AMT5: int  0 3455 14948 28959 19146 19619 483003 -159 11793 13007
...
 $ BILL_AMT6: int  0 3261 15549 29547 19131 20024 473944 567 3719
13912 ...
 $ PAY_AMT1 : int  0 0 1518 2000 2000 2500 55000 380 3329 0 ...
 $ PAY_AMT2 : int  689 1000 1500 2019 36681 1815 40000 601 0 0 ...
 $ PAY_AMT3 : int  0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4 : int  0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
 $ PAY_AMT5 : int  0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
 $ PAY_AMT6 : int  0 2000 5000 1000 679 800 13770 1542 1000 0 ....
```

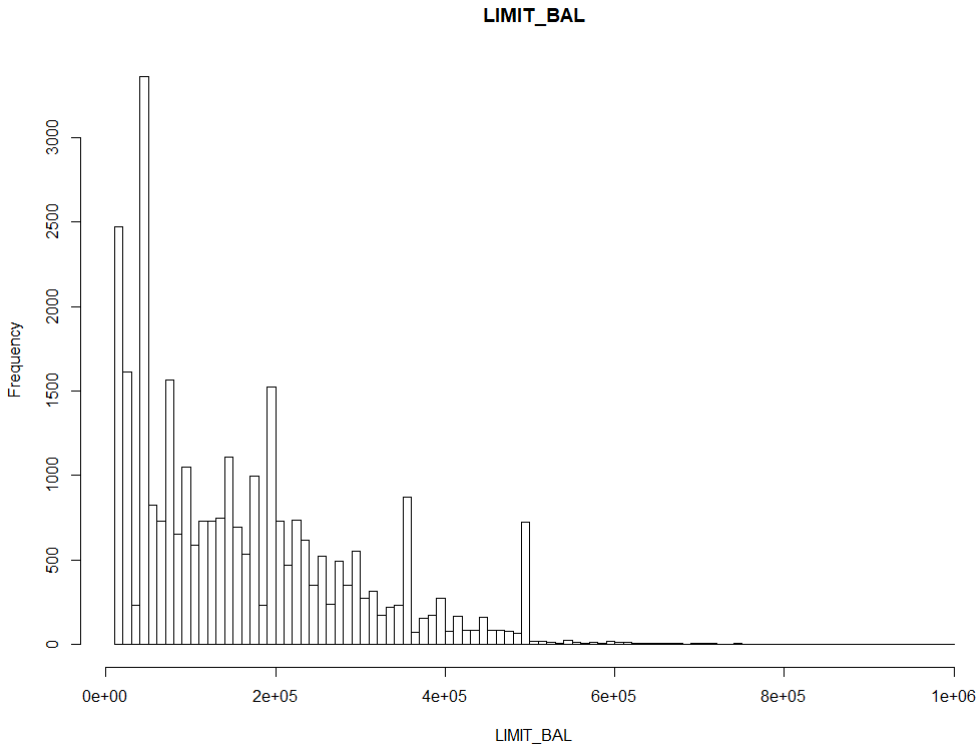
Ο αναγνώστης μπορεί να δει και αναλυτικότερα ένα δείγμα από το υπό μελέτη σύνολο και σε μορφή πίνακα στο Παράρτημα.

Παρότι στο συγκεκριμένο σύνολο δεδομένων δεν υπάρχουν ελλείπουσες τιμές όπως αναφέρει και η περιγραφή που το συνοδεύει, ακολούθησε και οπτικός έλεγχος. Αναφορικά με τις διπλότυπες παρατηρήσεις, αυτές υπήρχαν σε πολύ μικρό ποσοστό και αφαιρέθηκαν κατά την προεπεξεργασία των δεδομένων.

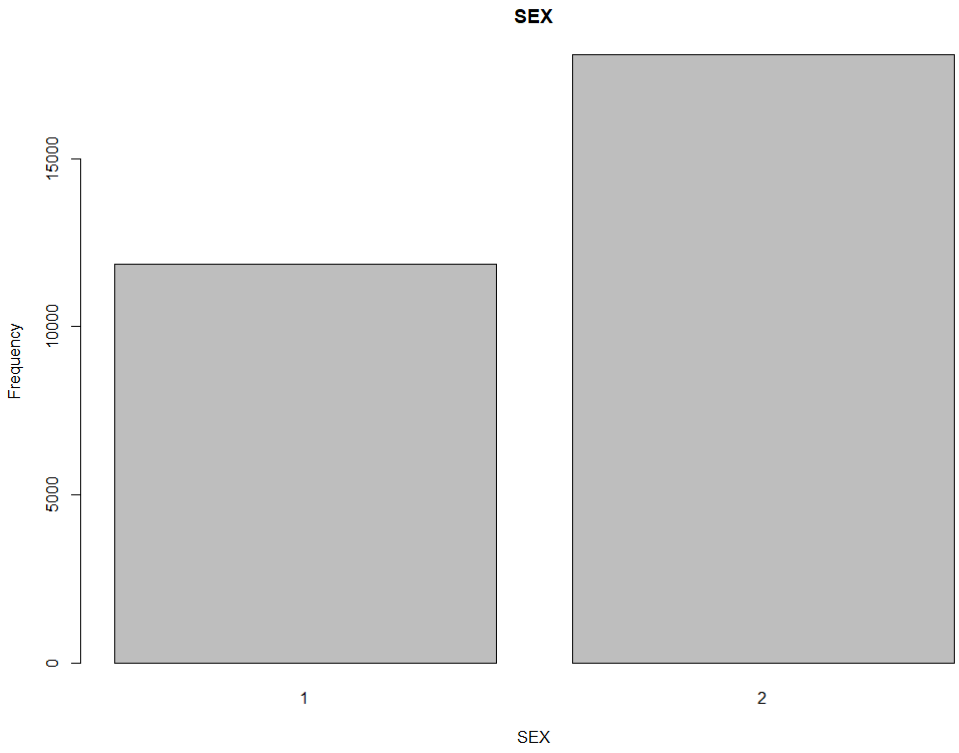
Ακολουθούν τα διαγράμματα του τμήματος “D. Exploratory Data Analysis” του κώδικα για κάθε μια από τις 24 μεταβλητές. Ακόμη θα παρουσιαστούν και οι αντίστοιχες συνόψεις των 5 αριθμών (five-number summary) για όσες από αυτές τις μεταβλητές έχει νόημα, εναλλακτικά θα παρουσιαστούν οι αντίστοιχες συχνότητες αυτών.



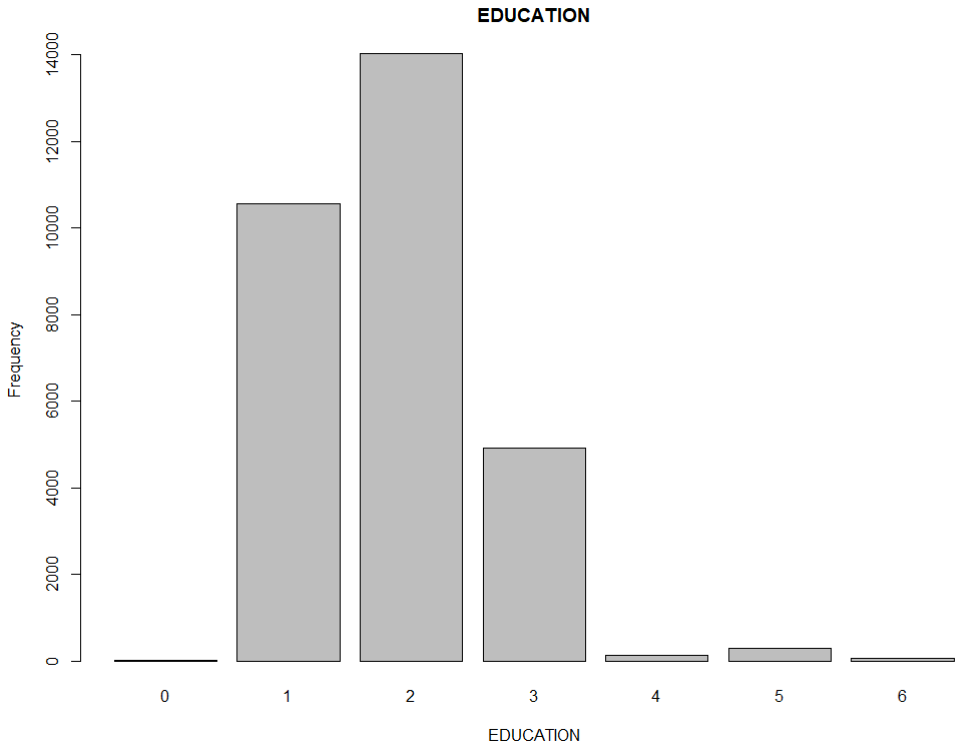
ΣΧΗΜΑ 5.1 Ιστόγραμμα συχνότητας εμφάνισης της εξαρτημένης μεταβλητής default, όπου 0: ικανότητα πληρωμής και 1: αδυναμία πληρωμής



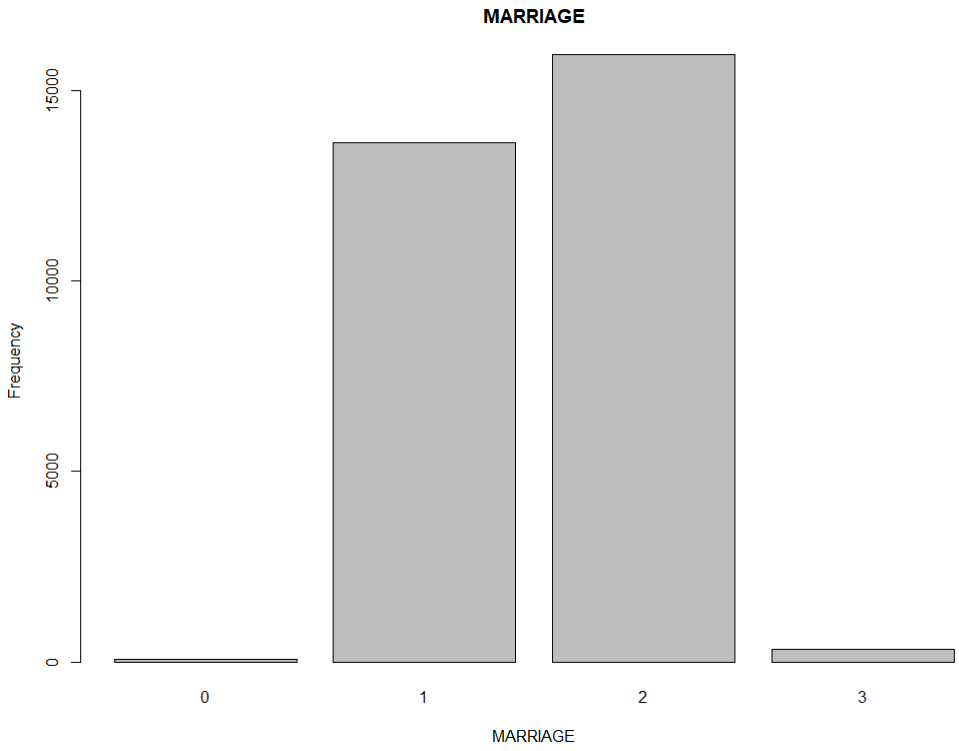
ΣΧΗΜΑ 5.2 Ιστόγραμμα ποσού δεδομένης πίστωσης



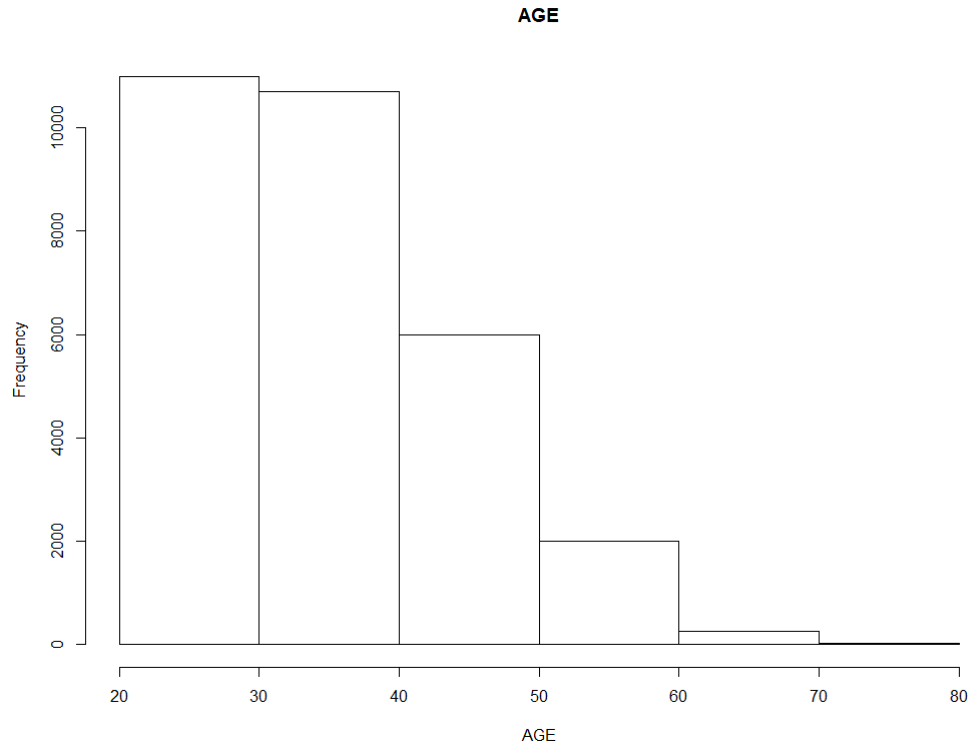
ΣΧΗΜΑ 5.3 Ιστόγραμμα συχνότητας εμφάνισης φύλου, όπου 1: αρσενικό και 2: θηλυκό



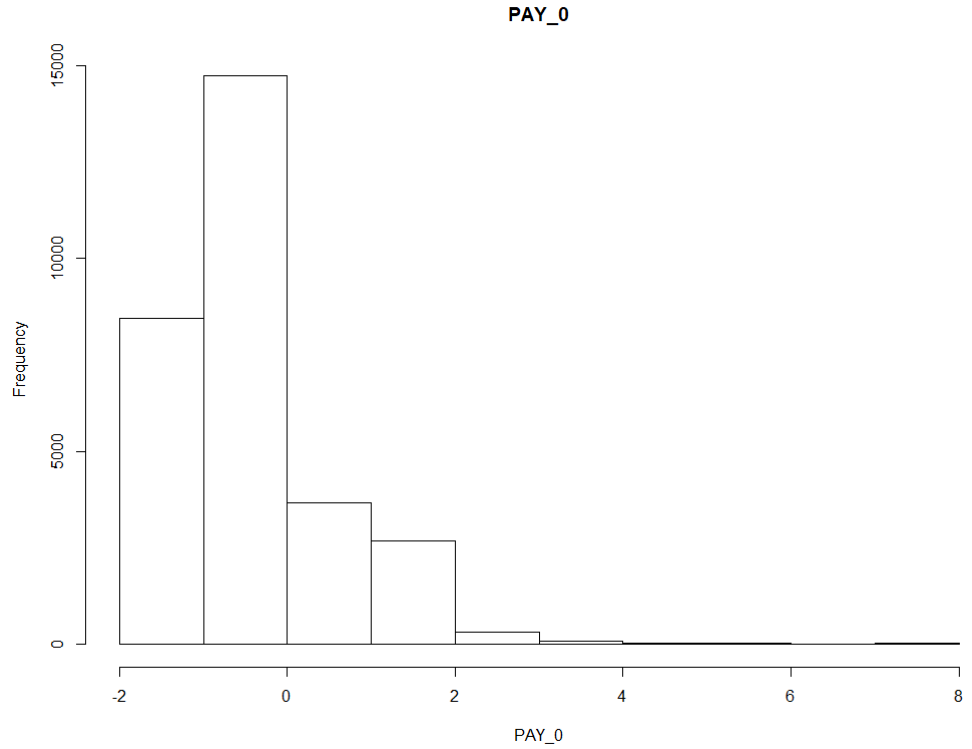
ΣΧΗΜΑ 5.4 Ιστόγραμμα συχνότητας εμφάνισης μορφωτικού επιπέδου, όπου 1: μεταπτυχιακή, 2: πανεπιστημιακή, 3: τριτοβάθμια, 4: άλλο, 5-6: άγνωστο



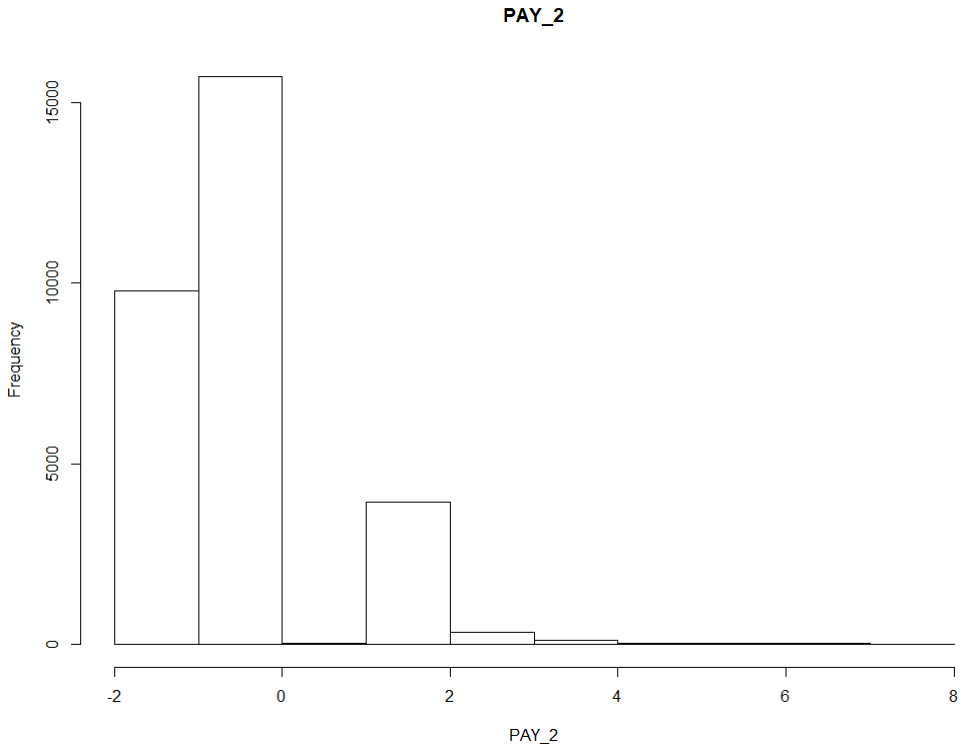
ΣΧΗΜΑ 5.5 Ιστόγραμμα συχνότητας εμφάνισης οικογενειακής κατάστασης, όπου 0: άγνωστο, 1: παντρεμένος, 2: ελεύθερος, 3: άλλο



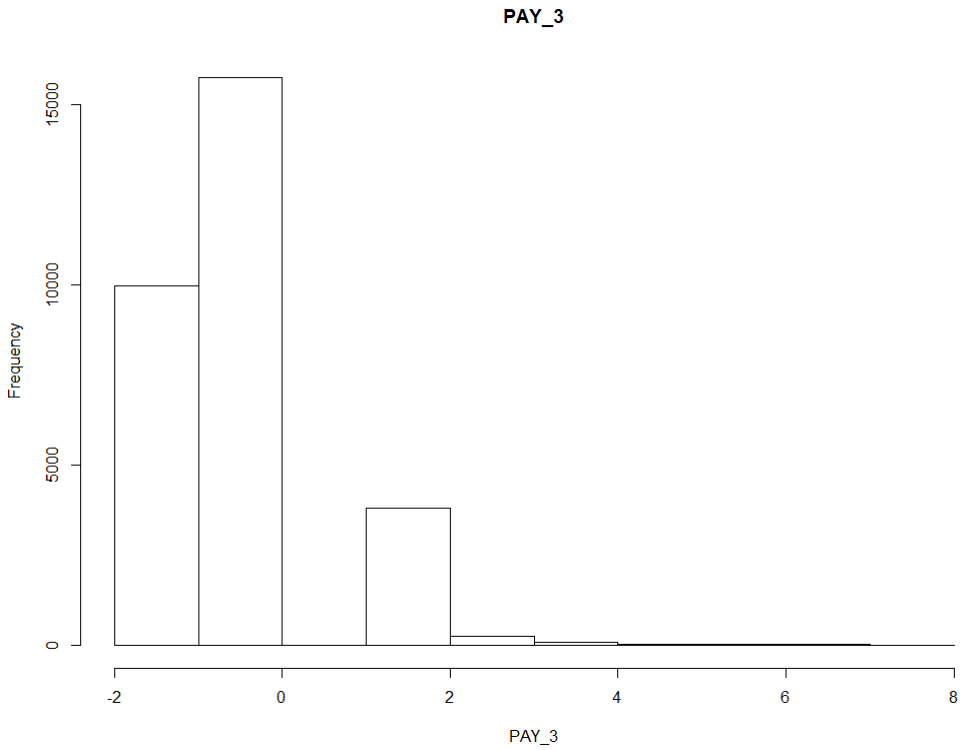
ΣΧΗΜΑ 5.6 Ιστόγραμμα ηλικίας



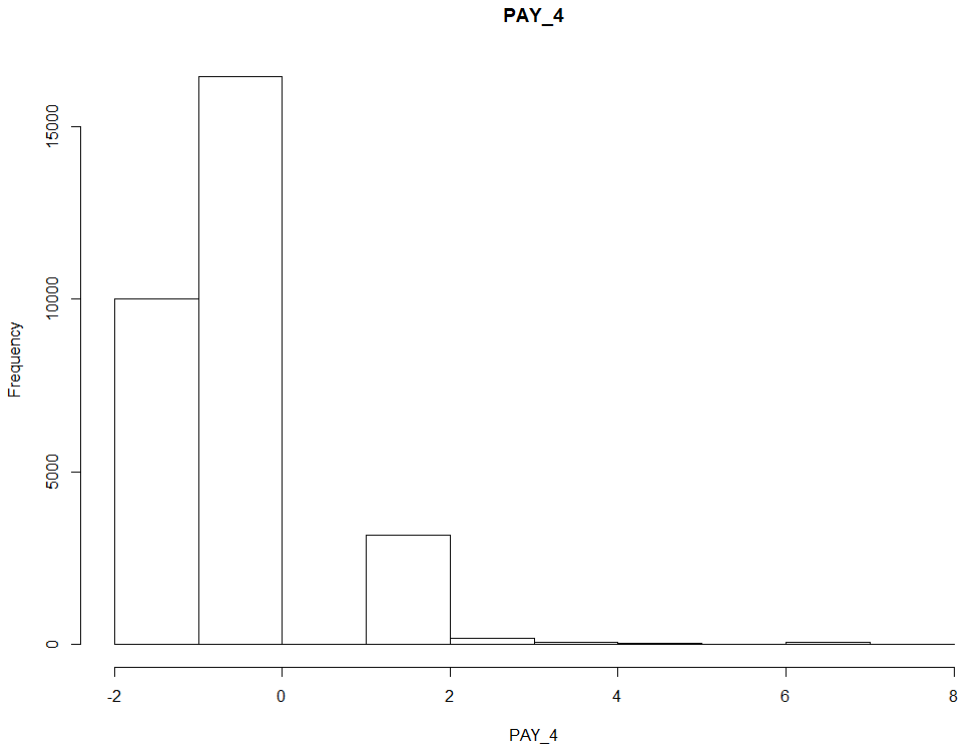
ΣΧΗΜΑ 5.7 Ιστόγραμμα ιστορικού προηγούμενης πληρωμής, για τον μήνα Σεπτέμβριο του 2005



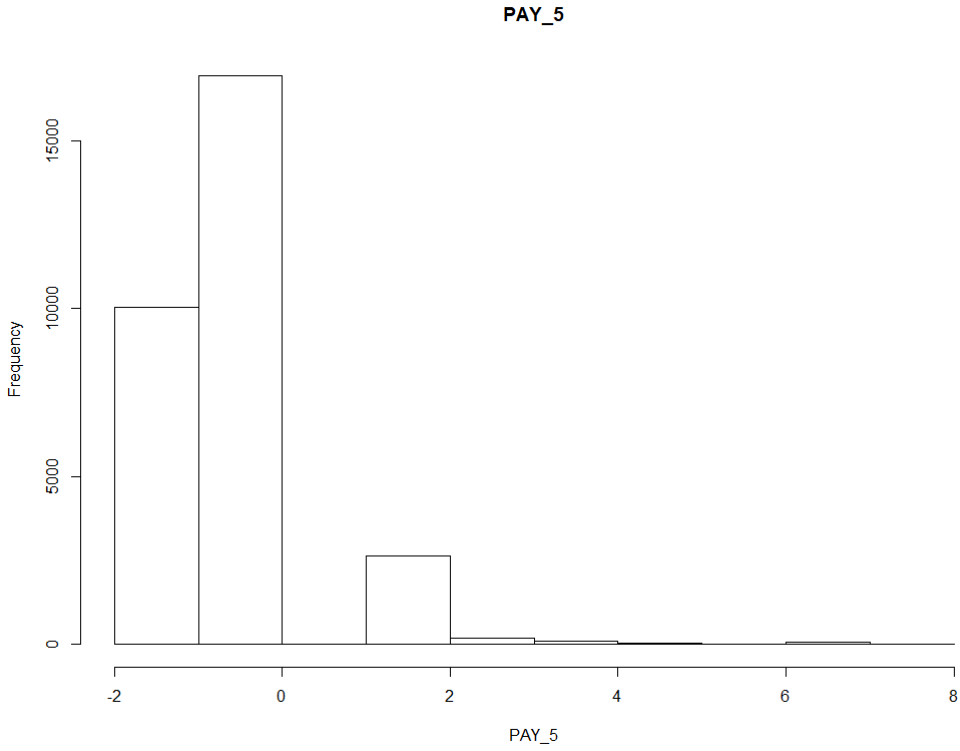
ΣΧΗΜΑ 5.8 Ιστόγραμμα ιστορικού προηγούμενης πληρωμής, για τον μήνα Αύγουστο του 2005



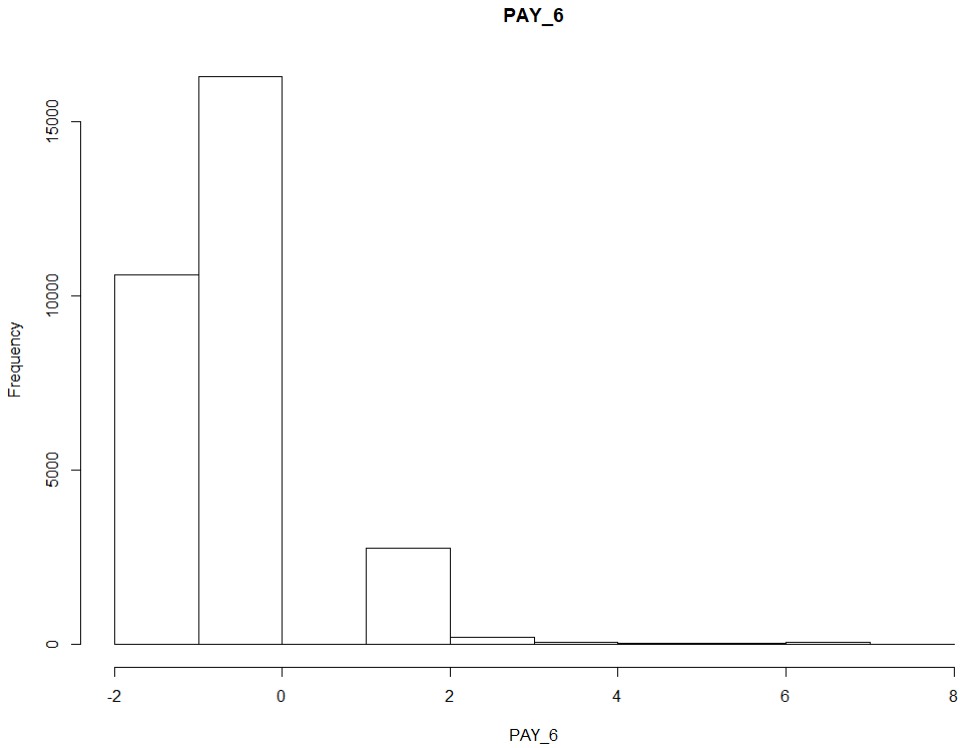
ΣΧΗΜΑ 5.9 Ιστόγραμμα ιστορικού προηγούμενης πληρωμής, για τον μήνα Ιούλιο του 2005



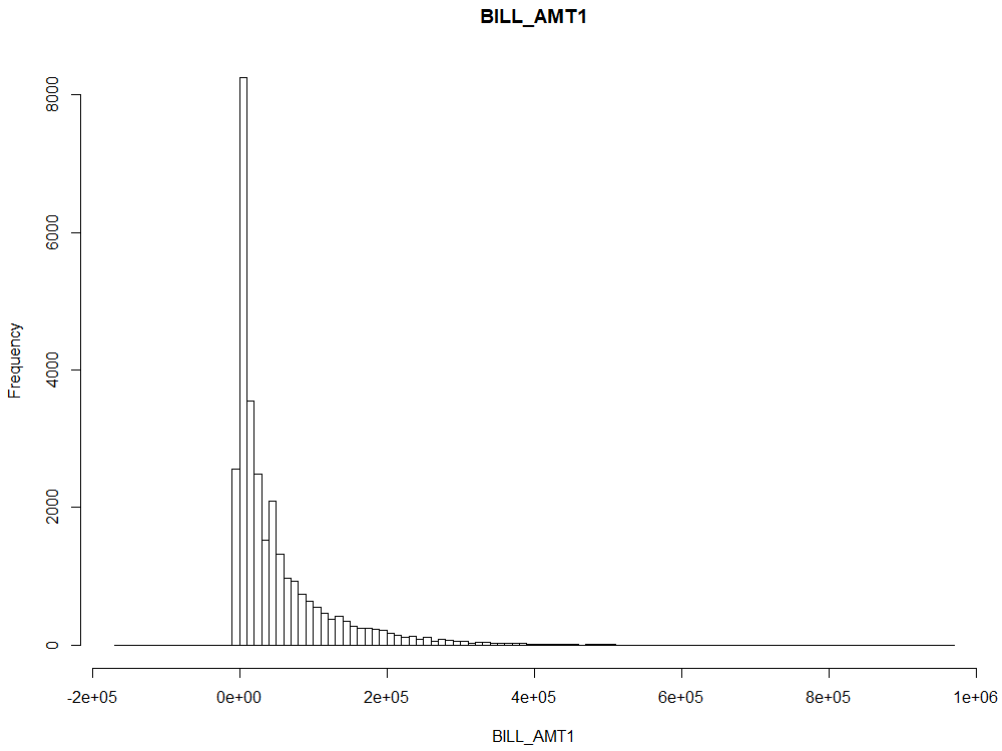
ΣΧΗΜΑ 5.10 Ιστόγραμμα ιστορικού προηγούμενης πληρωμής, για τον μήνα Ιούνιο του 2005



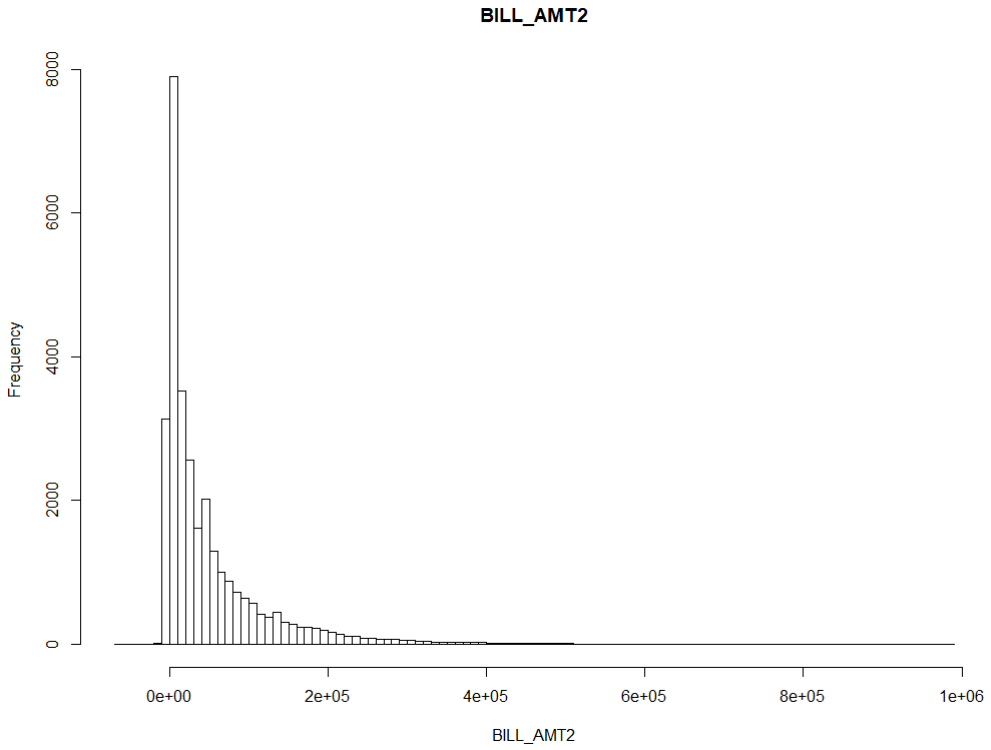
ΣΧΗΜΑ 5.11 Ιστόγραμμα ιστορικού προηγούμενης πληρωμής, για τον μήνα Μάιο του 2005



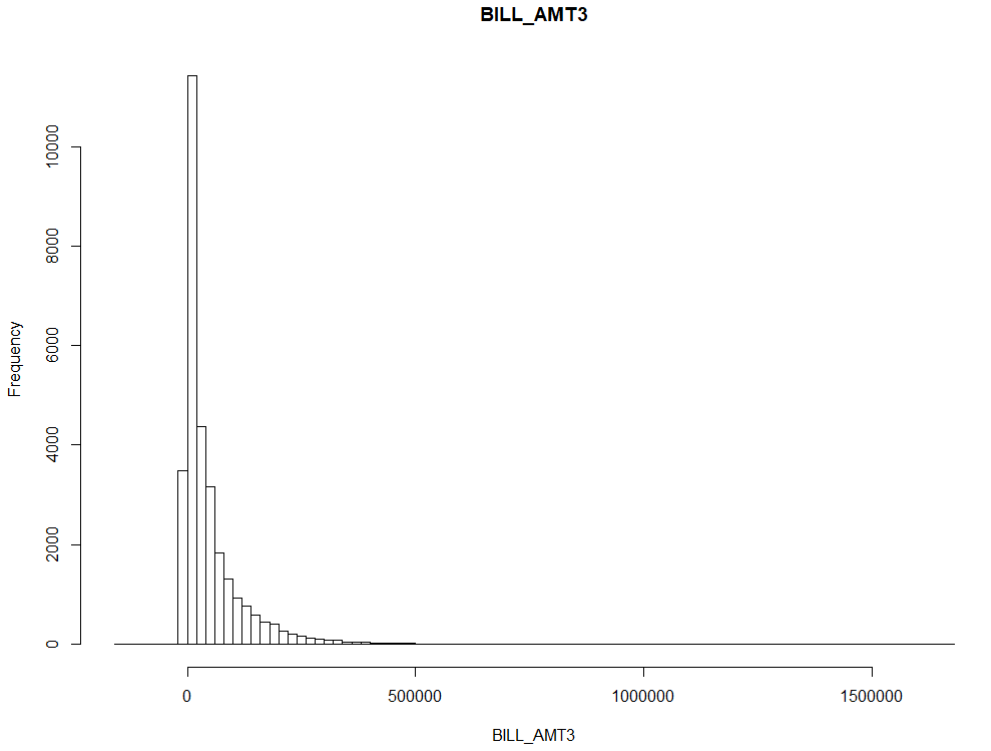
ΣΧΗΜΑ 5.12 Ιστόγραμμα ιστορικού προηγούμενης πληρωμής, για τον μήνα Απρίλιο του 2005



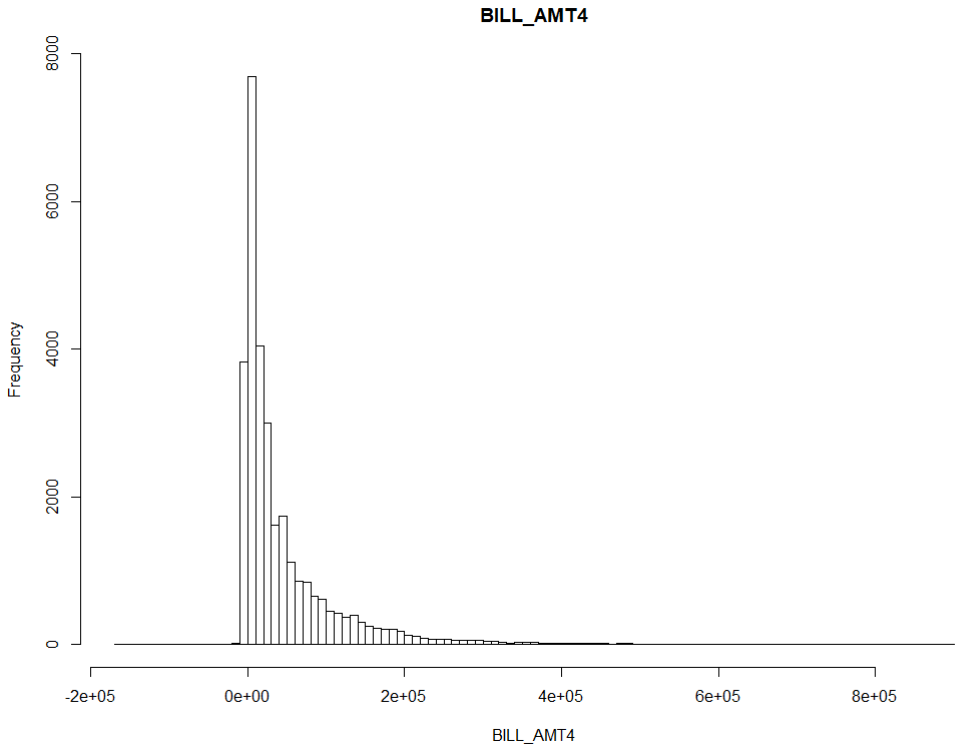
ΣΧΗΜΑ 5.13 Ιστόγραμμα ποσού λογαριασμού, για τον μήνα Σεπτέμβριο του 2005



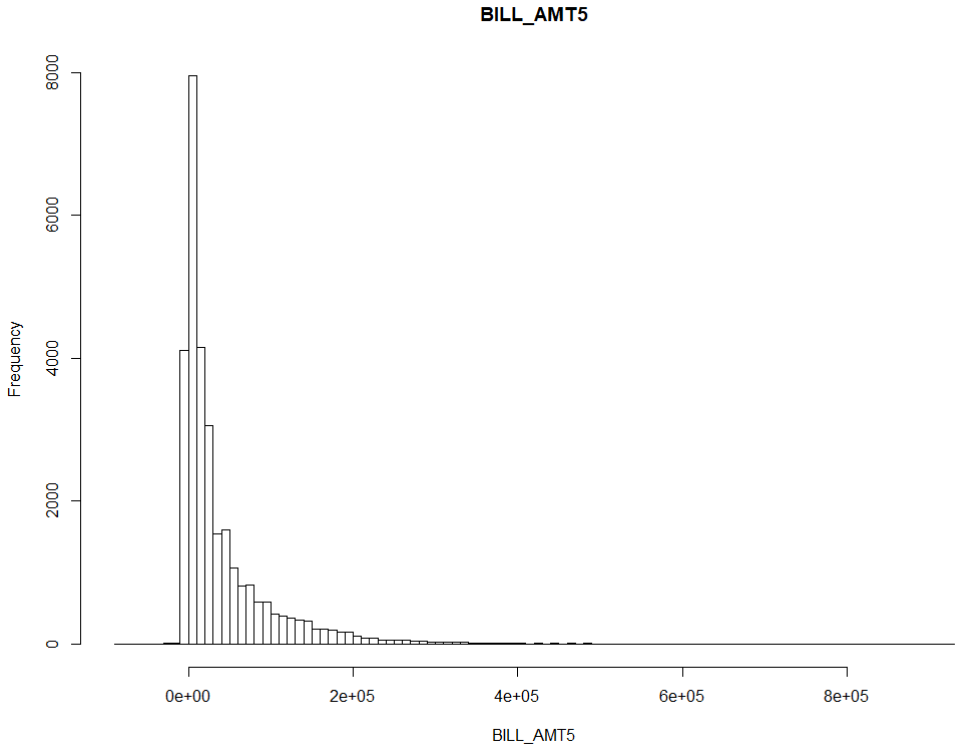
ΣΧΗΜΑ 5.14 Ιστόγραμμα ποσού λογαριασμού, για τον μήνα Αύγουστο του 2005



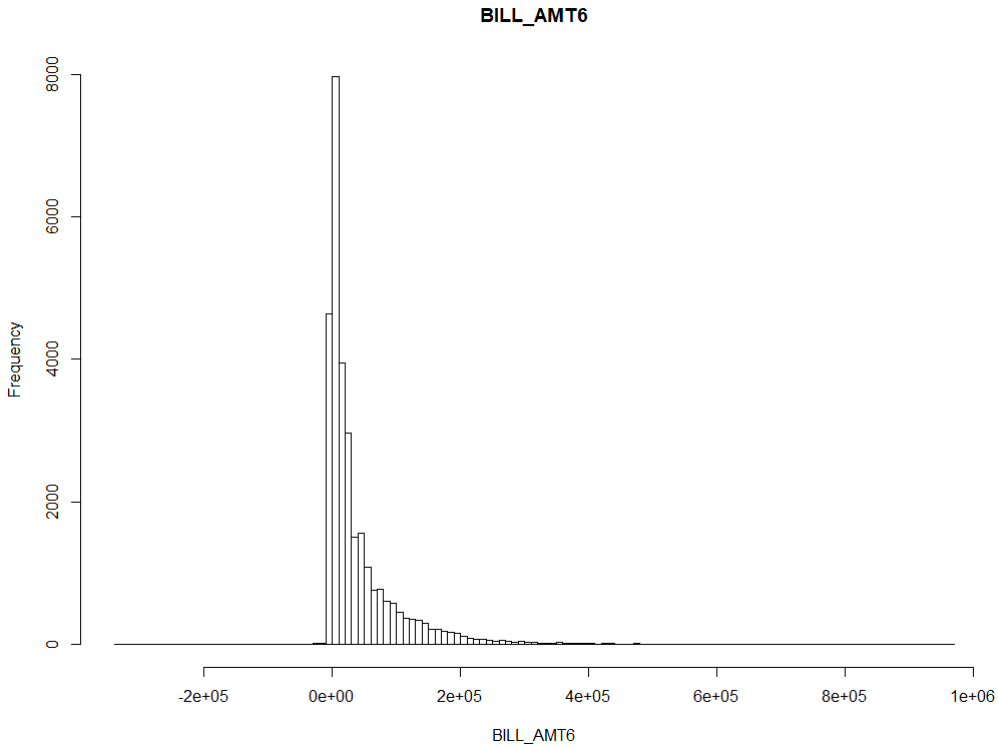
ΣΧΗΜΑ 5.15 Ιστόγραμμα ποσού λογαριασμού, για τον μήνα Ιούλιο του 2005



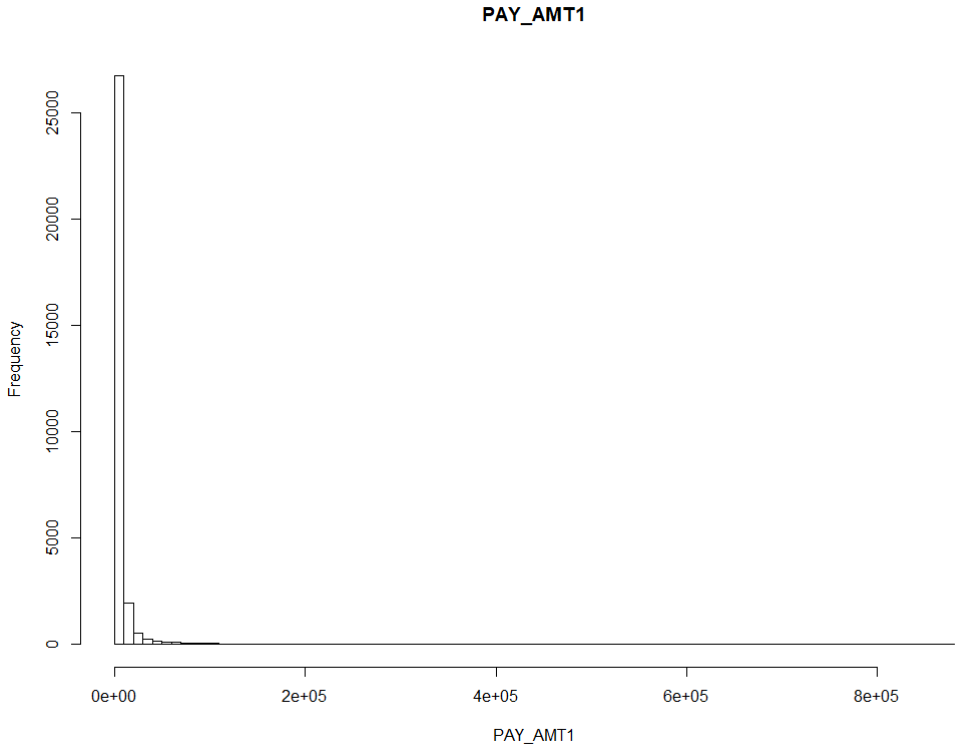
ΣΧΗΜΑ 5.16 Ιστόγραμμα ποσού λογαριασμού, για τον μήνα Ιούνιο του 2005



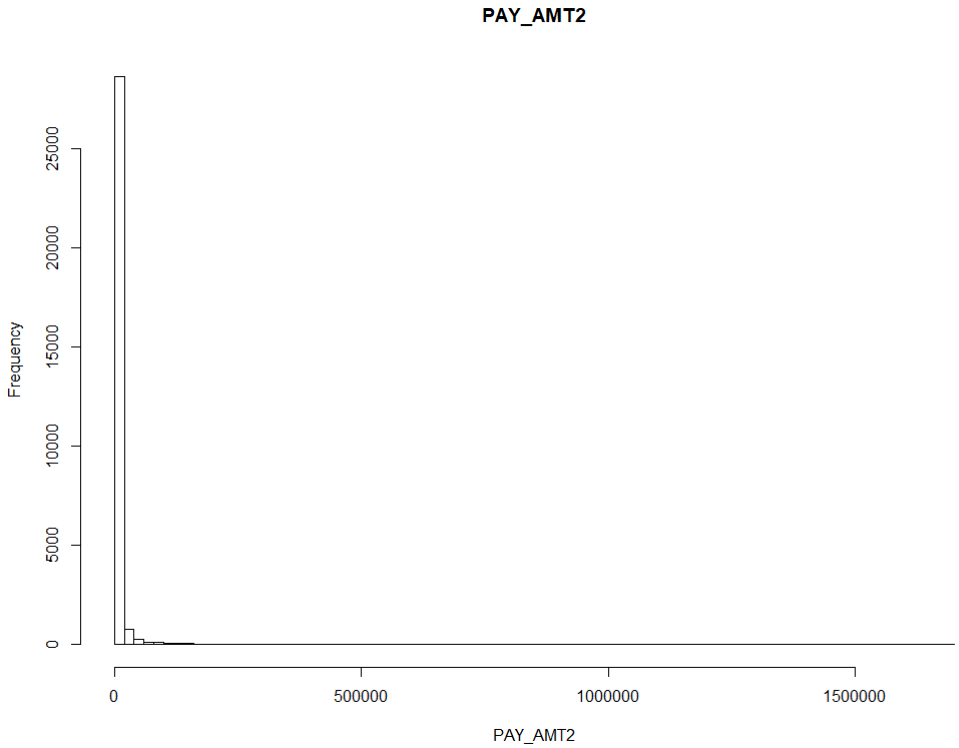
ΣΧΗΜΑ 5.17 Ιστόγραμμα ποσού λογαριασμού, για τον μήνα Μάιο του 2005



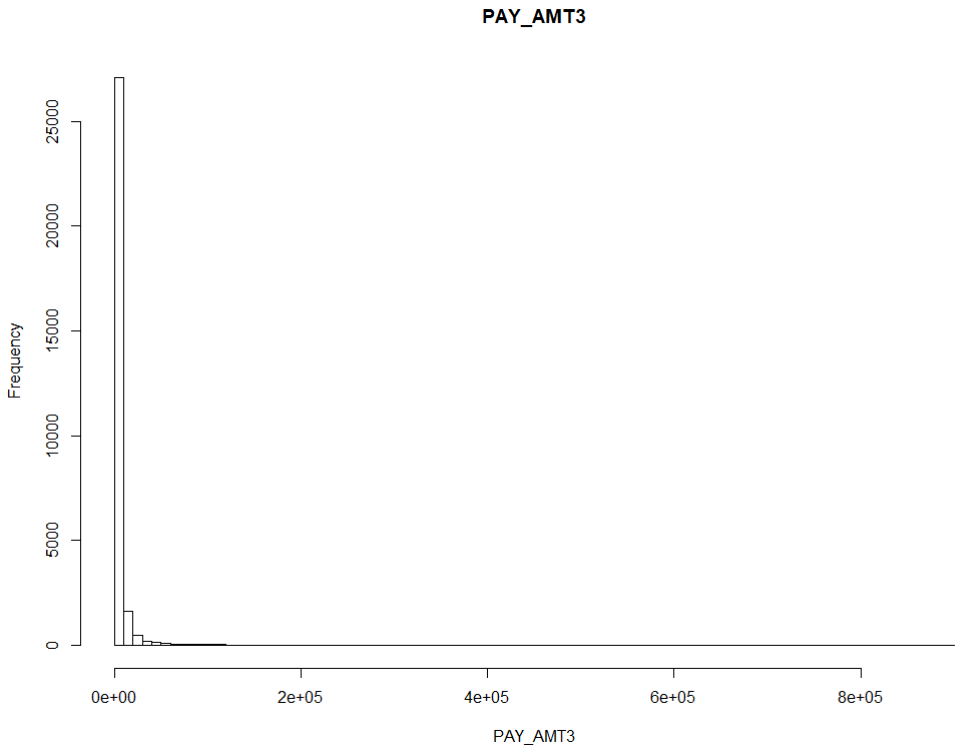
ΣΧΗΜΑ 5.18 Ιστόγραμμα ποσού λογαριασμού, για τον μήνα Απρίλιο του 2005



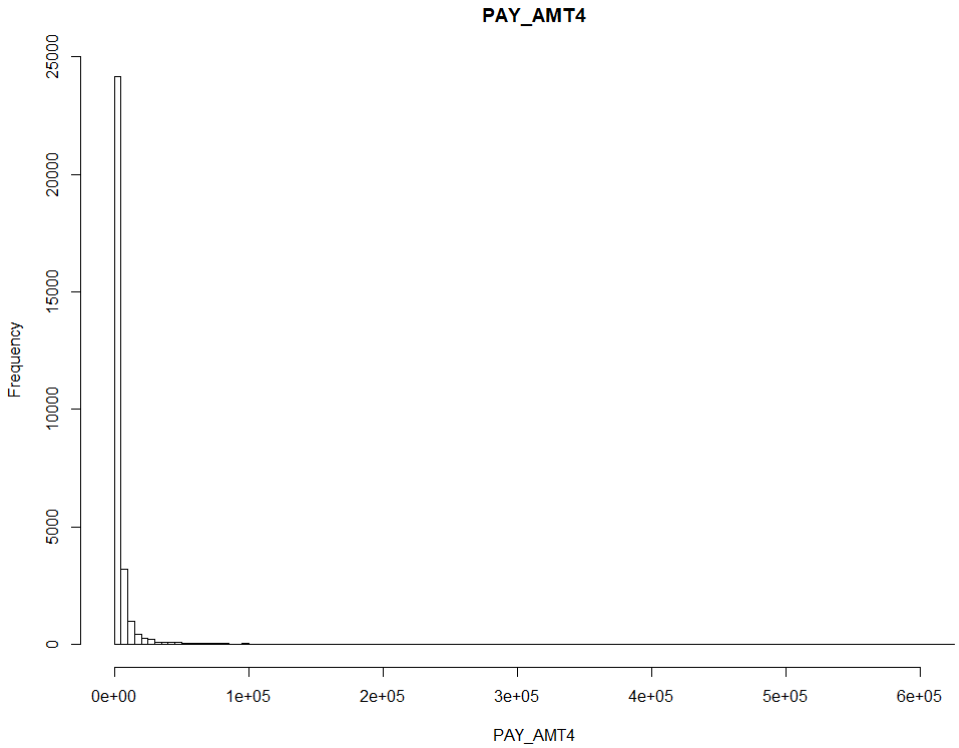
ΣΧΗΜΑ 5.19 Ιστόγραμμα ποσού προηγούμενης πληρωμής, για τον μήνα Σεπτέμβριο του 2005



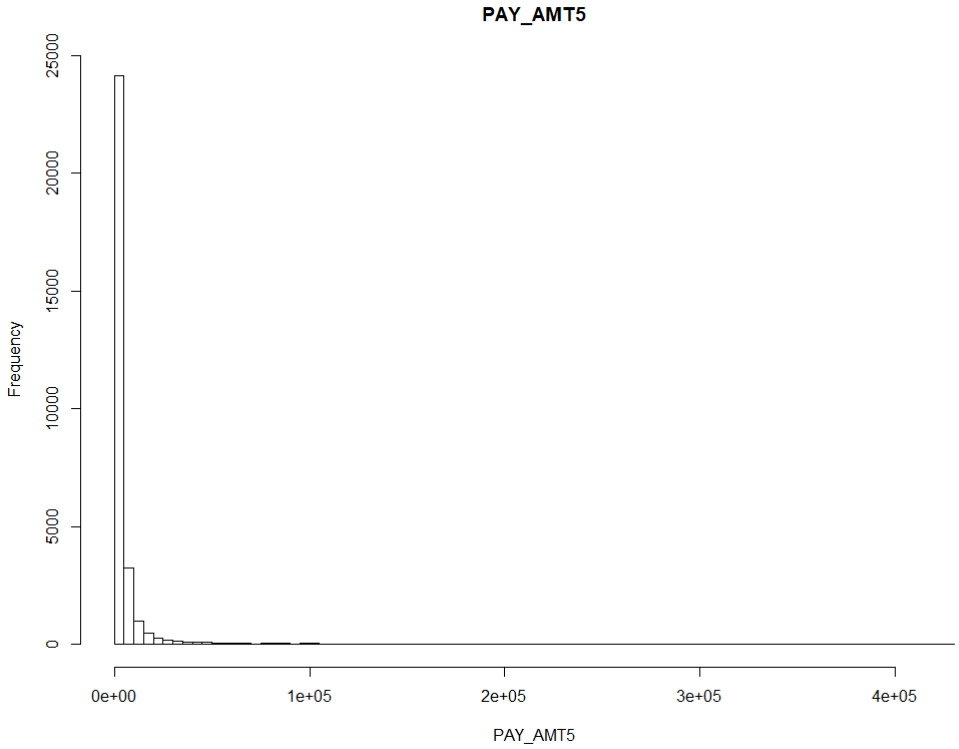
ΣΧΗΜΑ 5.20 Ιστόγραμμα ποσού προηγούμενης πληρωμής, για τον μήνα Αύγουστο του 2005



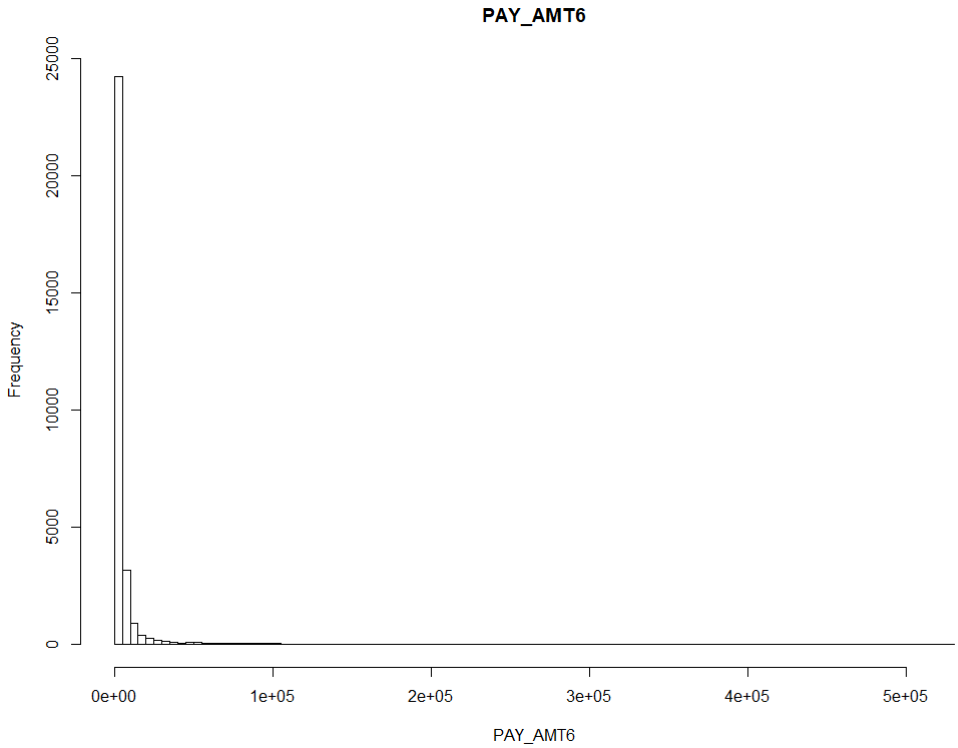
ΣΧΗΜΑ 5.21 Ιστόγραμμα ποσού προηγούμενης πληρωμής, για τον μήνα Ιούλιο του 2005



ΣΧΗΜΑ 5.22 Ιστόγραμμα ποσού προηγούμενης πληρωμής, για τον μήνα Ιούνιο του 2005



ΣΧΗΜΑ 5.23 Ιστόγραμμα ποσού προηγούμενης πληρωμής, για τον μήνα Μάιο του 2005



ΣΧΗΜΑ 5.24 Ιστόγραμμα ποσού προηγούμενης πληρωμής, για τον μήνα Απρίλιο του 2005

Συχνότητες εμφάνισης της εξαρτημένης μεταβλητής default

```
table(data$default)
```

```
    0     1
23335  6630
```

Σύνοψη των 5 αριθμών ποσού δεδομένης πίστωσης

```
summary(data$LIMIT_BAL)
```

```
 Min.  1st Qu.  Median    Mean  3rd Qu.     Max.
10000    50000  140000  167442   240000  1000000
```

Συχνότητες εμφάνισης φύλου

```
table(data$SEX)
```

```
    0     1
11874 18091
```

Συχνότητες εμφάνισης μορφωτικού επιπέδου

```
table(data$EDUCATION)
```

```
 0     1     2     3    4    5   6
14 10563 14019  4915  123  280  51
```

Συχνότητες εμφάνισης οικογενειακής κατάστασης

```
table(data$MARRIAGE)
```

```
 0     1     2    3
54 13643 15945  323
```

Σύνοψη των 5 αριθμών ηλικίας

```
summary(data$AGE)
```

```
 Min. 1stQu. Median   Mean 3rdQu.   Max.
21.00   28.00   34.00 35.49   41.00  79.00
```

Συνόψεις των 5 αριθμών ιστορικού προηγουμένων πληρωμών

```
lapply( data[,7:12],summary)
```

```
$`PAY_0`
    Min.   1stQu.   Median      Mean   3rdQu.    Max.
-2.00000 -1.00000  0.00000 -0.01675  0.00000 8.00000

$PAY_2
   Min.   1stQu.  Median     Mean  3rdQu.   Max.
-2.0000 -1.0000  0.0000 -0.1319  0.0000 8.0000

$PAY_3
   Min.   1stQu.  Median     Mean  3rdQu.   Max.
-2.0000 -1.0000  0.0000 -0.1644  0.0000 8.0000

$PAY_4
   Min.   1stQu.  Median     Mean  3rdQu.   Max.
-2.0000 -1.0000  0.0000 -0.2189  0.0000 8.0000

$PAY_5
   Min.   1stQu.  Median     Mean  3rdQu.   Max.
-2.0000 -1.0000  0.0000 -0.2645  0.0000 8.0000

$PAY_6
   Min.   1stQu.  Median     Mean  3rdQu.   Max.
-2.0000 -1.0000  0.0000 -0.2894  0.0000 8.0000
```

Συνόψεις των 5 αριθμών ποσών λογαριασμών

```
lapply(data[,13:18],summary)
```

```
$`BILL_AMT1`
    Min.  1stQu.  Median   Mean  3rdQu.    Max.
-165580     3595   22438  51283   67260  964511
$BILL_AMT2
  Min.  1stQu.  Median   Mean  3rdQu.    Max.
-69777    3010   21295  49236   64109  983931
$BILL_AMT3
   Min.  1stQu.  Median   Mean  3rdQu.     Max.
-157264    2711   20135  47068   60201  1664089
$BILL_AMT4
   Min.  1stQu.  Median   Mean  3rdQu.    Max.
-170000    2360   19081  43313   54601  891586
$BILL_AMT5
  Min.  1stQu.  Median   Mean  3rdQu.    Max.
-81334    1787   18130  40358   50247  927171
$BILL_AMT6
   Min.  1stQu.  Median   Mean  3rdQu.    Max.
-339603    1262   17124  38917   49252  961664
```

Συνόψεις των 5 αριθμών ποσών προηγουμένων πληρωμών

```
lapply(data[,19:24],summary)
```

```
$`PAY_AMT1`
Min.  1stQu.  Median  Mean  3rdQu.    Max.
   0    1000    2102  5670    5008  873552
$PAY_AMT2
Min.  1stQu.  Median  Mean  3rdQu.     Max.
   0     850    2010  5928    5000  1684259
$PAY_AMT3
Min.  1stQu.  Median  Mean  3rdQu.    Max.
   0     390    1804  5232    4512  896040
$PAY_AMT4
Min.  1stQu.  Median  Mean  3rdQu.    Max.
   0     300    1500  4832    4016  621000
$PAY_AMT5
Min.  1stQu.  Median  Mean  3rdQu.    Max.
   0     261    1500  4805    4042  426529
$PAY_AMT6
Min.  1stQu.  Median  Mean  3rdQu.    Max.
   0     131    1500  5222    4000  528666
```

5.3 Μεθοδολογία και στρατηγική ανάλυσης

Όπως αναφέρθηκε παραπάνω, βασική ιδέα της ανάλυσης που περιγράφεται στην παρούσα ενότητα είναι η πρόβλεψη με χρήση μοντέλων μηχανικής μάθησης της αδυναμίας πληρωμής χρηστών πιστωτικών καρτών βάσει των ιστορικών χαρακτηριστικών τους, ενσωματώνοντας information-driven μεθόδους επιλογής χαρακτηριστικών (mRMR) στο στάδιο της προεπεξεργασίας. Τα χαρακτηριστικά (features) που συνθέτουν το χρησιμοποιούμενο σύνολο δεδομένων περιγράφηκαν εκτενώς στην προηγούμενη ενότητα. Ακολουθεί, επομένως, η περιγραφική ανάπτυξη της μεθοδολογίας εξόρυξης γνώσης που πραγματοποιήθηκε, η οποία θα αποτυπωθεί στην τελευταία υποενότητα του παρόντος κεφαλαίου με τη μορφή κώδικα και σχετικών σχολίων.

Στάδια ανάλυσης:

1. Η διαδικασία εξόρυξης δεδομένων για την επίλυση του προβλήματος ξεκινά με τη συλλογή του σχετικού συνόλου δεδομένων και την εγκατάσταση των απαιτούμενων βιβλιοθηκών επεξεργασίας, μοντελοποίησης και απεικόνισης αποτελεσμάτων.
2. Περνώντας στο στάδιο της προεπεξεργασίας, ακολουθεί οπτικός έλεγχος του dataset και μετατροπή των ανεξάρτητων μεταβλητών του σε κατάλληλη προς ανάλυση μορφή. Αυτό το βήμα συνοδεύεται από την αφαίρεση τυχόν διπλότυπων εγγραφών.
3. Το σύνολο δεδομένων είναι πλέον έτοιμο να υποβληθεί σε διαδικασία επιλογής χαρακτηριστικών, όπου γίνεται χρήση δύο παραλλαγών του mRMR για συλλογική μάθηση (ensemble learning) με χρήση εκτιμητών συσχέτισης Pearson και Spearman αντίστοιχα και εξάγονται τα αντίστοιχα φίλτρα επιλογής χαρακτηριστικών που θα εφαρμοστούν στο σύνολο δεδομένων σε επόμενο στάδιο (Mukaka, 2012).
4. Ακολουθεί το στάδιο της διερευνητικής ανάλυσης δεδομένων (Exploratory Data Analysis - EDA), όπου με χρήση μεθόδων περιγραφικής στατιστικής δίνεται η δυνατότητα στον αναγνώστη να κατανοήσει καλύτερα τα χαρακτηριστικά του χρησιμοποιούμενου dataset. Τα διαγράμματα και οι διάφοροι δείκτες που παρουσιάστηκαν στην Υποενότητα 5.2 προήλθαν από αυτό το στάδιο.

5. Ένα μικρό μέρος του συνόλου δεδομένων αποκόπτεται από το αρχικό, ώστε να χρησιμοποιηθεί για την αξιολόγηση των τελικών μοντέλων πρόβλεψης. Το μεγαλύτερο μέρος χρησιμοποιείται για την εκπαίδευση και τον έλεγχο της ακρίβειας πρόβλεψης των μοντέλων μηχανικής μάθησης, ακολουθώντας τη στρατηγική k-πλής διασταυρωμένης επικύρωσης (k-fold cross-validation) (Stone, 1974). Σύμφωνα με τη μέθοδο αυτή, τα δεδομένα χωρίζονται τυχαία σε k ισοπληθή υποσύνολα, εκ των οποίων τα k-1 χρησιμοποιούνται για την εκπαίδευση των μοντέλων και το τελευταίο για τον υπολογισμό του σφάλματος ταξινόμησης. Η διαδικασία επαναλαμβάνεται k φορές, με διαφορετικό σύνολο επαλήθευσης κάθε φορά, και τελικά τα k εκπαιδευμένα μοντέλα χρησιμοποιούνται για την τελική πρόβλεψη (π.χ. πλειοψηφία ή μέσος όρος).

6. Εφαρμογή των φίλτρων επιλογής χαρακτηριστικών mRMR με εκτιμητές συσχέτισης Pearson και Spearman, και δημιουργία αντίστοιχων συνόλων δεδομένων που περιέχουν μόνο 10 από τα 23 χαρακτηριστικά του αρχικού συνόλου δεδομένων.

7. Εκκίνηση της διαδικασίας μοντελοποίησης με χρήση των αλγορίθμων που περιγράφηκαν εκτενώς στο Κεφάλαιο 2 και συγκεκριμένα των ακόλουθων: Μηχανές Διανυσματικής Υποστήριξης (SVM), k-Πλησιέστεροι Γείτονες (kNN), Πολυεπίπεδο Νευρωνικό Δίκτυο (MLP-nnet) και Τυχαία Δάση (RF). Σημειώνεται ότι η πλειονότητα των μεθόδων υποβάλλεται σε διαδικασία παραμετροποίησης, ώστε να προκύψει το βέλτιστο δυνατό μοντέλο κάθε μοντελοποιητικής οικογένειας. Με την ολοκλήρωση της διαδικασίας, συγκρίνονται τα αποτελέσματα ταξινόμησης για το σύνολο των μοντέλων, ώστε να επιλεγούν οι καλύτεροι ταξινομητές που θα υποβληθούν στο επόμενο στάδιο. Τα σχετικά αποτελέσματα που προέκυψαν σε αυτό το στάδιο αλλά και στα επόμενα, έχουν συγκεντρωθεί μαζί με τις αντίστοιχες ερμηνείες τους στο Κεφάλαιο 6.

8. Τα δύο καλύτερα - ως προς την ακρίβεια ταξινόμησης - μοντέλα υποβάλλονται σε προς τα εμπρός επιλογή χαρακτηριστικών με χρήση mRMR, προκειμένου να διαπιστωθεί η επίδραση της αύξησης των μεταβλητών απόφασης στη διαδικασία πρόβλεψης. Η επαναληπτική αυτή διεργασία ολοκληρώνεται με την παραγωγή των αντίστοιχων αποτελεσμάτων και διαγραμμάτων.

9. Το τελευταίο στάδιο της μεθοδολογίας ανάλυσης περιλαμβάνει την αξιολόγηση των

εξαχθέντων μοντέλων με χρήση του εναπομείναντος υποσυνόλου που είχε αφαιρεθεί από το σύνολο εκπαίδευσης, ώστε να διαπιστωθεί εάν τα προκύπτοντα μοντέλα έχουν τις αναμενόμενες επιδόσεις πρόβλεψης σε ένα “άγνωστο” για αυτά σύνολο. Η διαδικασία ολοκληρώνεται με την παράθεση των σχετικών αποτελεσμάτων.

Στην ακόλουθη υποενότητα γίνεται αντιστοίχιση των παραπάνω σταδίων με συγκεκριμένα κομμάτια της αλγοριθμικής διαδικασίας που υλοποιήθηκε στα πλαίσια της μεταπτυχιακής διατριβής.

5.4 Περιγραφή αλγοριθμικής διαδικασίας

Σε αυτή την υποενότητα αναλύεται τμηματικά ο κώδικας που παράχθηκε στα πλαίσια της παρούσας μεταπτυχιακής διατριβής σε γλώσσα προγραμματισμού R v.3.5.1, ο οποίος αναπτύχθηκε και εκτελέστηκε σε περιβάλλον RStudio v.1.1.383 σε φορητό υπολογιστή με επεξεργαστή i7-3635QM και μνήμη 8GB DDR3. Ο συνολικός χρόνος εκτέλεσης του κώδικα ήταν 17.85 ώρες, ενώ σημειώνεται ότι με την αξιοποίηση βιβλιοθηκών βασικών αλγορίθμων γραμμικής άλγεβρας για διανυσματικές πράξεις (OpenBLAS), μπορεί να επισπευσθεί σημαντικά ο χρόνος εκτέλεσης.

Ακολουθεί η παράθεση των επιμέρους τμημάτων του κώδικα που αφορούν καθένα από τα διακριτά στάδια ανάλυσης που περιγράφηκαν στην Υποενότητα 5.3, συνοδευόμενα από λεπτομέρειες σχετικά με τις επιμέρους εντολές που χρησιμοποιήθηκαν, ενώ - όπου κρίνεται απαραίτητο - παρατίθεται και το αποτέλεσμα των εντολών αυτών. Σημειώνεται ότι το σύνολο των αποτελεσμάτων που αφορούν τις επιδόσεις πρόβλεψης των μοντέλων μηχανικής μάθησης συγκεντρώνεται στο Κεφάλαιο 6. Το σύνολο του κώδικα παρατίθεται και στο Παράρτημα για καλύτερη αναγνωσιμότητα.

```
---
title: "R Notebook"
output:
  word_document: default
  html_notebook: default
---


### A. Install and load necessary libraries, set working directory and read data.
```{r}

rm(list=ls(all=TRUE))
start_time <- Sys.time()
#install.packages("rstudioapi") # run this if it's your first time using it to
install
#install.packages("caret")      # run this if it's your first time using it to
install
#install.packages("plyr")       # run this if it's your first time using it to
install
#install.packages("reshape2")   # run this if it's your first time using it to
install
#install.packages("class")      # run this if it's your first time using it to
install
#install.packages("ggplot2")    # run this if it's your first time using it to
install

# Set the working directory to the relevant one:
setwd(dirname("C:/Users/Katerina/Desktop/Kate/diatrivi_master/R_Codes/"))

# Load necessary libraries
library(rstudioapi)
library(caret)
library(plyr)
library(reshape2)
library(class)
library(ggplot2)

data <- read.csv(file=
"C:/Users/Katerina/Desktop/Kate/diatrivi_master/R_Codes/CreditCard.csv",
header=TRUE, sep=",", dec=".")
# make sure to change the filepath in chunk H. as well.
```
```

Η αλγοριθμική διαδικασία ξεκινά με καθαρισμό του workspace, εγκατάσταση των απαραίτητων βιβλιοθηκών (σε περίπτωση που λείπουν από το σύστημα) και φόρτωσή τους. Καθορίζεται επίσης ο φάκελος εργασίας στον οποίο θα αποθηκεύονται τα outputs του κώδικα (πχ. εκπαιδευμένα μοντέλα), γίνεται ανάγνωση του συνόλου δεδομένων “CreditCard.csv” και εισαγωγή του στο workspace ως data.frame με την ονομασία “data”.

```
### B. Data Quality & Sanity checks

```{r}

# Reassuring correct type of variable acknowledgement
str(data)

# default is the dependent variable and should be a factor type
class(data$default)
data$default <- as.factor(data$default)

# limit_bal is an indepedent variable and should be a numerical type
class(data$LIMIT_BAL)
data$LIMIT_BAL <- as.numeric(data$LIMIT_BAL)

# sex (aka gender) is an independent variable and should be a factor type
class(data$SEX)
data$SEX <- as.factor(data$SEX)

# education is an independent variable and should be a factor type
class(data$EDUCATION)
data$EDUCATION <- as.factor(data$EDUCATION)

# marriage is an indepedent variable and should be a factor type
class(data$MARRIAGE)
data$MARRIAGE <- as.factor(data$MARRIAGE)

# age is an independent variable and should be a numeric type
class(data$AGE)
data$AGE <- as.numeric(data$AGE)

# pay_0 to pay_6 are independent variables and should be a numeric type
data[,7:12]<-lapply(data[,7:12],as.numeric)

# bill_amt1 to bill_amt6 are independent variables and should be numeric
data[,13:18]<-lapply(data[,13:18],as.numeric)

# pay_amt1 to pay_amt6 are independent variables and should be numeric
data[,19:24]<-lapply(data[,19:24],as.numeric)

################################################################

# Ensure no duplicate records exist
data.u <- unique(data)

ifelse(length(nrow(data.u))==nrow(data),print("no duplicates
found"),print("duplicates found - keeping unique records"))

# Remove duplicates
data <- data.u
rm(data.u)

```
```

Η παραπάνω σειρά εντολών υλοποιεί τον προαπαιτούμενο μετασχηματισμό δεδομένων, ώστε αυτά να αποκτήσουν την κατάλληλη μορφή πριν αξιοποιηθούν από τα μοντέλα εξόρυξης γνώσης. Δεδομένου ότι το χρησιμοποιούμενο dataset αποτελείται αποκλειστικά από αριθμητικά δεδομένα, κρίνεται ποια από αυτά αντιπροσωπεύουν ποιοτικά χαρακτηριστικά (πχ. φύλο, εκπαίδευση) και άρα θα πρέπει να

μετασχηματιστούν σε κατηγορικές μεταβλητές, και ποια αφορούν ποσοτικά χαρακτηριστικά (πχ. υπόλοιπο λογαριασμού, μήνες καθυστέρησης) και θα πρέπει να αποδοθούν ως αριθμητικές μεταβλητές. Τέλος, γίνεται αφαίρεση των διπλότυπων εγγραφών, προκειμένου να αποφευχθούν φαινόμενα υπερεκπαίδευσης. Ακολουθεί το αποτέλεσμα της παραπάνω διαδικασίας μετασχηματισμού μετά τις αλλαγές (πρώτες δέκα γραμμές του συνόλου όπως παράχθηκαν από την R).

```
'data.frame':  29965 obs. of  24 variables:
 $ default  : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 1 ...
 $ LIMIT_BAL: num  20000 120000 90000 50000 50000 50000 500000 100000
140000 20000 ...
 $ SEX      : Factor w/ 2 levels "1","2": 2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION: Factor w/ 7 levels "0","1","2","3",..: 3 3 3 3 3 2 2 3 4
4 ...
 $ MARRIAGE : Factor w/ 4 levels "0","1","2","3": 2 3 3 2 2 3 3 3 2 3 ...
 $ AGE      : num  24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0    : num  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2    : num  2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3    : num  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4    : num  -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5    : num  -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6    : num  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1: num  3913 2682 29239 46990 8617 ...
 $ BILL_AMT2: num  3102 1725 14027 48233 5670 ...
 $ BILL_AMT3: num  689 2682 13559 49291 35835 ...
 $ BILL_AMT4: num  0 3272 14331 28314 20940 ...
 $ BILL_AMT5: num  0 3455 14948 28959 19146 ...
 $ BILL_AMT6: num  0 3261 15549 29547 19131 ...
 $ PAY_AMT1 : num  0 0 1518 2000 2000 ...
 $ PAY_AMT2 : num  689 1000 1500 2019 36681 ...
 $ PAY_AMT3 : num  0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4 : num  0 1000 1000 1100 9000 ...
 $ PAY_AMT5 : num  0 0 1000 1069 689 ...
 $ PAY_AMT6 : num  0 2000 5000 1000 679 ...
```

```
### C. Feature selection based on mRMRe
```

```
```{r}

#install.packages("mRMRe")      # run this if it's your first time using it to
install
library(mRMRe)

# Necessary conversions for mRMR feature selection
data.num  <- sapply(data, as.numeric)
data.num  <- data.frame(data.num)
mrmr.data <- mRMR.data(data.num)
rm(data.num)

# Select number of important features
nf = 10

# mRMR with pearson coef.
classic1 <-mRMR.classic("mRMRe.Filter", data=mrmr.data, target_indices = 1,
feature_count = nf, method = "bootstrap", continuous_estimator = "pearson")

# mRMR with spearman coef.
classic2 <-mRMR.classic("mRMRe.Filter", data=mrmr.data, target_indices = 1,
feature_count = nf, method = "bootstrap", continuous_estimator = "spearman")

```
```

Σε αυτό το στάδιο υλοποιείται η επιλογή χαρακτηριστικών με χρήση της μεθόδου mRMR. Συγκεκριμένα, γίνεται φόρτωση της αντίστοιχης βιβλιοθήκης καθώς και οι απαιτούμενες μετατροπές δεδομένων σε αριθμητικές μεταβλητές, προκειμένου να χρησιμοποιηθούν ως είσοδοι στην διαδικασία επιλογής. Αφού οριστεί ο επιθυμητός αριθμός χαρακτηριστικών (nf=10) που θέλουμε να προκύψουν από την εκτέλεση, γίνεται εφαρμογή του αλγορίθμου για δύο περιπτώσεις: τη δημιουργία του φίλτρου “classic1” που αφορά επιλογή χαρακτηριστικών mRMR με χρήση εκτιμητή συσχέτισης Pearson και του φίλτρου “classic2” που αφορά επιλογή χαρακτηριστικών mRMR με χρήση εκτιμητή συσχέτισης Spearman. Τα 10 σημαντικότερα χαρακτηριστικά που επιλέχθηκαν από τις παραπάνω μεθόδους φαίνονται κατά σειρά παρακάτω:

```
> solutions(classic1)
```

```
$`1`

      [,1]
 [1,]    7
 [2,]    5
 [3,]   19
 [4,]    3
 [5,]   23
 [6,]    4
```

```
 [7,]  24
 [8,]  11
 [9,]  22
[10,]  21
```

```
> solutions(classic2)
```

```
$`1`

      [,1]
 [1,]    7
 [2,]   21
 [3,]    3
 [4,]    5
 [5,]    4
 [6,]    2
 [7,]    6
 [8,]   20
 [9,]   12
[10,]   19
```

D. Exploratory Data Analysis

```r
# Data structure summary
str(data)

# default is a factor type ~ plot (Sxima 5.1)
table(data$default)
plot(data$default,main=colnames(data[1]),xlab = colnames(data[1]),ylab =
"Frequency")

# limit_bal is a numeric type ~ histogram (Sxima 5.2)
summary(data$LIMIT_BAL)
hist(data$LIMIT_BAL,breaks=100,main=colnames(data[2]),xlab = colnames(data[2]))

# sex is a factor type ~ plot (Sxima 5.3)
table(data$SEX)
plot(data$SEX,main=colnames(data[3]),xlab = colnames(data[3]),ylab = "Frequency")

# education is a factor type ~ plot (Sxima 5.4)
table(data$EDUCATION)
plot(data$EDUCATION,main=colnames(data[4]),xlab = colnames(data[4]),ylab =
"Frequency")

# marriage is a factor type ~ plot (Sxima 5.5)
table(data$MARRIAGE)
plot(data$MARRIAGE,main=colnames(data[5]),xlab = colnames(data[5]),ylab =
"Frequency")

# age is a numerical type ~ histogram (Sxima 5.6)
summary(data$AGE)
hist(data$AGE,breaks=5,main=colnames(data[6]),xlab = colnames(data[6]))

# pay_0 to pay_6 are numerical types ~ histograms (Sximata 5.7 - 5.12)
lapply( data[,7:12],summary)
for (i in 7:12) {
```

```
  hist(data[,i],breaks=10,main=colnames(data)[i],xlab = colnames(data[i]))
}

# bill_amt1 to bill_amt6 are numerical types ~ histograms (Sximata 5.13 - 5.18)
lapply(data[,13:18],summary)
for (i in 13:18) {
  hist(data[,i],breaks=100,main=colnames(data)[i],xlab = colnames(data[i]))
}

# pay_amt1 to pay_amt6 are numerical types ~ histograms (Sxima 5.19 - 5.24)
lapply(data[,19:24],summary)
for (i in 19:24) {
  hist(data[,i],breaks=100,main=colnames(data)[i],xlab = colnames(data[i]))
}

```

Το παραπάνω κομμάτι κώδικα αφορά την οπτικοποίηση του εύρους τιμών των επιμέρους χαρακτηριστικών που συνθέτουν το σύνολο δεδομένων, με χρήση μεθόδων διερευνητικής ανάλυσης δεδομένων, στοχεύοντας στην καλύτερη κατανόηση του προβλήματος. Τα παραχθέντα διαγράμματα, οι συχνότητες εμφάνισης και οι συνόψεις των 5 αριθμών παρουσιάστηκαν στην Υποενότητα 5.2 (Σχήματα 5.1 – 5.24).

E. Split data to CV-evaluation subsets, set resampling strategy

```{r}

library(caret)

portion       <- createDataPartition(data$default, p = 0.8, list=FALSE)
data.cv       <- data[portion,]  # 80% of the initial dataset on which to perform
model selection
data.evaluate <- data[-portion,] # 20% of the initial dataset on which to
evaluate model accuracy

# Keep safe initial dataset
data.initial <- data
# Rename data.cv portion into data and remove data.cv
data <- data.cv
rm(data.cv)

# Repeated cross-validation strategy:
# Split available data into 5 equal parts, train in 4 and evaluate in the 5th

set.seed(123)
fitControl <- trainControl(method="repeatedcv",
                           number=10,
                           repeats=1,
                           classProbs  = FALSE,
                           verboseIter = FALSE,
                           savePredictions = TRUE,
                           allowParallel = TRUE)

rm(portion)

```

Σε αυτό το σημείο, γίνεται διαχωρισμός του συνόλου δεδομένων, σε υποσύνολο που θα χρησιμοποιηθεί για την εκπαίδευση και τον έλεγχο ακρίβειας πρόβλεψης των μοντέλων και σε υποσύνολο επικύρωσης. Συγκεκριμένα, δημιουργώντας μια διαμέριση (portion, p=0.8) του dataset λαμβάνουμε δύο επιμέρους υποσύνολα, το “data.cv” το οποίο θα χρησιμοποιηθεί για την επιλογή του βέλτιστου μοντέλου με χρήση cross-validation, και το “data.evaluate” βάσει του οποίου θα αξιολογηθεί η ορθότητα της επιλογής μας. Ακολούθως, αφού το αρχικό dataset μετονομαστεί σε “data.initial” για να κρατηθεί ασφαλές στο workspace, εργαζόμαστε με το “data.cv”, το οποίο μετονομάζεται σε “data”. Τέλος, ορίζονται οι παράμετροι της στρατηγικής επαναδειγματοληψίας του cross-validation μέσω της μεθόδου fitcontrol, η οποία εφαρμόζει 10-fold cross validation. Αυτό σημαίνει ότι το χρησιμοποιούμενο σύνολο δεδομένων θα χωριστεί σε 10 μέρη ώστε κάθε μοντέλο να εκπαιδεύεται διαδοχικά σε 9 από αυτά και να ελέγχεται στο 10ο.

```
### F. Preprocessing features and creation of different dataset versions.
```{r}

set.seed(123)

# Rename necessary 0 and 1 to X.0 and X.1 respectively
data$default <- revalue(data$default , c("1"="X.1", "0"="X.0"))

# mRMR pearson subset
data.p <- data[,c(1,unname(unlist(classic1@filters)))]

# mRMR spearman subset
data.s <- data[,c(1,unname(unlist(classic2@filters)))]

```
```

Στο τελευταίο βήμα προεπεξεργασίας πριν την έναρξη της μοντελοποίησης, γίνεται μετασχηματισμός των τιμών της μεταβλητής απόκρισης “default” (X.0 , X.1 αντί 0,1), ώστε να διασφαλιστεί η ορθή επεξεργασία τους από το πακέτο caret. Τέλος, γίνεται εφαρμογή των φίλτρων “classic1” και “classic2” στα δεδομένα, ώστε να προκύψουν δύο αντίστοιχα datasets: το “data.p” το οποίο περιέχει τα 10 βέλτιστα features βάσει της επιλογής χαρακτηριστικών mRMR με εκτιμητή συσχέτισης Pearson, και το “data.s” που περιέχει τα 10 βέλτιστα features βάσει της επιλογής χαρακτηριστικών mRMR με εκτιμητή συσχέτισης Spearman. Τα δύο αυτά datasets, μαζί με το “data” που περιλαμβάνει και τα 23 features θα χρησιμοποιηθούν στο επόμενο στάδιο της μοντελοποίησης.

G. Modeling (using svm, knn, nnet, rf)

```{r}

# Support vector machines (all features , mRMR with Pearson coef. , mRMR with Spearman coef.)

set.seed(123)

svm <- train(default ~ .,
             data = data,
             method = "svmRadial",
             metric = "Accuracy",
             trControl = fitControl)

svm.p <- train(default ~ .,
             data = data.p,
             method = "svmRadial",
             metric = "Accuracy",
             trControl = fitControl)

svm.s <- train(default ~ .,
             data = data.s,
             method = "svmRadial",
             metric = "Accuracy",
             trControl = fitControl)

# k-Nearest Neighbours (all features , mRMR with Pearson coef. , mRMR with Spearman coef.)

knn <-train(default ~ .,
             data = data,
             method = "knn",
             metric = "Accuracy",
             trControl = fitControl)

knn.p <-train(default ~ .,
             data = data.p,
             method = "knn",
             metric = "Accuracy",
             trControl = fitControl)

knn.s <-train(default ~ .,
             data = data.s,
             method = "knn",
             metric = "Accuracy",
             trControl = fitControl)

# Neural Networks (all features , mRMR with Pearson coef. , mRMR with Spearman coef.)

set.seed(123)

nnGrid=expand.grid(size=c(2),decay=c(0.02))

nn <- train(default ~ .,
             data = data,
             method = 'nnet',
             tuneGrid = nnGrid,
             metric = "Accuracy",
             trControl = fitControl)

nn.p <- train(default ~ .,
             data = data.p,
             method = 'nnet',
             tuneGrid = nnGrid,
             metric = "Accuracy",
             trControl = fitControl)
```

```
nn.s <- train(default ~ .,
              data = data.s,
              method = 'nnet',
              tuneGrid = nnGrid,
              metric = "Accuracy",
              trControl = fitControl)

# Random Forests (all features , mRMR with Pearson coef. , mRMR with Spearman
coef.)

set.seed(123)

tunegrid <- expand.grid(.mtry=c(4,5,6))

rf <- train(default ~.,
                  data    = data,
                  method = "rf",
                  metric = "Accuracy",
                  tuneGrid = tunegrid,
                  trControl=fitControl,
                  ntree=100)

rf.p <- train(default ~.,
                  data    = data.p,
                  method = "rf",
                  metric = "Accuracy",
                  tuneGrid = tunegrid,
                  trControl=fitControl,
                  ntree=100)

rf.s <- train(default ~.,
                  data    = data.s,
                  method = "rf",
                  metric = "Accuracy",
                  tuneGrid = tunegrid,
                  trControl=fitControl,
                  ntree=100)


# Cumulative cross-validation accuracy results of classifiers

results <- resamples(list(
                           SVM        =svm,
                           SVM.p      =svm.p,
                           SVM.s      =svm.s,
                           KNN        =knn,
                           KNN.p      =knn.p,
                           KNN.s      =knn.s,
                           NN         = nn,
                           NN.p       = nn.p,
                           NN.s       = nn.s,
                           RF         = rf,
                           RF.p       = rf.p,
                           RF.s       = rf.s
                           ))

# Saving trained models
saveRDS(svm, file = 'svm' )
saveRDS(svm.p, file = 'svm.p' )
saveRDS(svm.s, file = 'svm.s' )
saveRDS(knn, file = 'knn' )
saveRDS(knn.p, file = 'knn.p' )
saveRDS(knn.s, file = 'knn.s' )
saveRDS(nn, file = 'nn' )
saveRDS(nn.p, file = 'nn.p' )
saveRDS(nn.s, file = 'nn.s' )
saveRDS(rf, file = 'rf' )
```

```
saveRDS(rf.p, file = 'rf.p' )
saveRDS(rf.s, file = 'rf.s' )

# Print cross-validation accuracy results of classifiers
summary(results)
dotplot(results)
bwplot(results)

```

Το συγκεκριμένο στάδιο του αλγορίθμου περιέχει το σύνολο εντολών για τη δημιουργία των ταξινομητών που βασίζονται στις μοντελοποιητικές οικογένειες SVM, kNN, MLP-nnet και RF, όπως αυτές διατίθενται από το πακέτο caret. Συγκεκριμένα, καθεμιά από τις τέσσερεις μοντελοποιητικές οικογένειες καλείται να εκπαιδεύσει τα μοντέλα της βάσει τριών διαφορετικών συνόλων δεδομένων, του “data” που περιέχει και τα 23 αρχικά χαρακτηριστικά, του “data.p” που περιέχει τα 10 βέλτιστα βάσει του mRMR Pearson και του “data.s” που περιέχει τα 10 βέλτιστα βάσει του mRMR Spearman. Επομένως αναμένεται να προκύψουν συνολικά 12 εκπαιδευμένα μοντέλα. Σημειώνεται ότι όλα τα μοντέλα υποβλήθηκαν σε cross-validation με χρήση του fitcontrol, ενώ ως μέτρο ακρίβειας της απόδοσης του μοντέλου ορίστηκε η ακρίβεια ταξινόμησης (accuracy), η οποία ορίζεται ως το κλάσμα των σωστών ταξινομήσεων προς το άθροισμα σωστών και λανθασμένων ταξινομήσεων. Ανάλογα με τον αλγόριθμο εκπαίδευσης απαιτούνταν μερικές περαιτέρω παραμετροποιήσεις κατά περίπτωση, οι οποίες αναφέρονται συνοπτικά παρακάτω:

- ✔ **SVM:** χρήση μη-γραμμικού Gaussian πυρήνα svmRadial
- ✔ **kNN:** προκαθορισμένη παραμετροποίηση που περιλαμβάνει επιλογή της βέλτιστης υποπερίπτωσης για k=5, 7, 9 πλησιέστερους γείτονες
- ✔ **nnet:** καθορισμός του size=2 για τον αριθμό των hidden layers, και του decay=0.02 για την απόσβεση των βαρών.
- ✔ **RF:** καθορισμός mtry=4, 5, 6 ως του αριθμού μεταβλητών που επιλέγονται τυχαία ως υποψήφιες σε κάθε split του κάθε δέντρου, καθορισμός συνολικού αριθμού δέντρων ntree=100)

Μετά την ολοκλήρωση της διαδικασίας cross-validation, ακολουθεί η παρουσίαση των αποτελεσμάτων, η αποθήκευση των μοντέλων για μελλοντική χρήση καθώς και η παραγωγή των σχετικών διαγραμμάτων (boxplots) που παρατίθενται στο Κεφάλαιο 6.

Από την παραπάνω ανάλυση, την καλύτερη απόδοση ταξινόμησης σημείωσαν οι μοντελοποιητικές οικογένειες SVM και RF όπως θα παρουσιαστεί στα αποτελέσματα του Κεφαλαίου 6.

```
### H. Comparative modeling with forward feature selection (mRMR Pearson) between two best classifiers, with respect to CV accuracy
```

```{r}

# Fetching the initial preprocessed dataset
data <- data.initial

set.seed(123)

# Set tunegrid for rf
tunegrid <- expand.grid(.mtry=c(4,5,6))

# Initialise accuracy results vectors
TotalEvalAccRF  = c()
TotalEvalAccSVM = c()

portionF       <- createDataPartition(data$default, p = 0.8, list=FALSE)
data.cvF       <- data[portionF,]  # 80% of the initial dataset on which to perform model selection
data.evaluateF <- data[-portionF,] # 20% of the initial dataset on which to evaluate model accuracy

# Keep safe initial dataset
data.initial <- data

# Rename data.cv portion into data and remove data.cvF
data <- data.cvF #
rm(data.cvF)

# Repeated cross-validation strategy:
# Split available data into 5 equal parts, train in 4 and evaluate in the 5th

set.seed(123)
fitControl <- trainControl(method="repeatedcv",
                           number=10,
                           repeats=1,
                           classProbs  = FALSE,
                           verboseIter = FALSE,
                           savePredictions = TRUE,
                           allowParallel = TRUE)

rm(portionF)

# Rename necessary 0 and 1 to X.0 and X.1 respectively
data$default <- revalue(data$default , c("1"="X.1", "0"="X.0"))

# Forward selection (mRMR Pearson) from 1 to 23 features
for (i in 1:(ncol(data)-1)) {

    classic1F <-mRMR.classic("mRMRe.Filter", data=mrmr.data, target_indices = 1, feature_count = i, method = "bootstrap", continuous_estimator = "pearson")

    set.seed(123)

    # mRMR pearson subset
    data.pF <- data[,c(1,unname(unlist(classic1F@filters)))]
```

```
    # Support vector machine

    svm.pF <- train(default ~ .,
                    data = data.pF,
                    method = "svmRadial",
                    metric = "Accuracy",
                    trControl = fitControl)

    # Random Forest

    rf.pF <- train(default ~.,
                     data   = data.pF,
                     method = "rf",
                     metric = "Accuracy",
                     tuneGrid = tunegrid,
                     trControl=fitControl,
                     ntree=100)


    # Saving trained models, needed to compute cv accuracy later
     saveRDS(svm.pF, file = paste('svmp', i, sep="") )
     saveRDS(rf.pF, file = paste('rfp', i, sep="") )

    # Transform evaluation subset to include only mRMR selected features
    data.p.evaluateF  <- data.evaluateF[,c(1,unname(unlist(classic1F@filters)))]

    # Compute Random Forest evaluation accuracy for mRMR selected features
    preds <- predict(rf.pF, newdata =
data.frame( data.p.evaluateF[colnames(data.p.evaluateF)[2:(i+1)]]))
    TotalEvalAccRF[i]<-(table(preds,data.p.evaluateF$default)[1,1]+
table(preds,data.p.evaluateF$default)[2,2])/(nrow(data.p.evaluateF))

    # Compute SVM evaluation accuracy for mRMR selected features
    preds <- predict(svm.pF, newdata =
data.frame( data.p.evaluateF[colnames(data.p.evaluateF)[2:(i+1)]]))
    TotalEvalAccSVM[i]<-(table(preds,data.p.evaluateF$default)[1,1]+
table(preds,data.p.evaluateF$default)[2,2])/(nrow(data.p.evaluateF))

}

# Print Random Forest and SVM CV accuracy for all mRMR selected subsets
modelnamesSVMp <- paste0("svmp", 1:i)
modelsSVMp <- lapply(modelnamesSVMp, function(x) readRDS(x))
resultsSVMp <- resamples( modelsSVMp ,  modelNames = modelnamesSVMp )
summary(resultsSVMp)

modelnamesRFp <- paste0("rfp", 1:i)
modelsRFp <- lapply(modelnamesRFp, function(x) readRDS(x))
resultsRFp <- resamples( modelsRFp ,  modelNames = modelnamesRFp )
summary(resultsRFp)

# Respective plots for CV accuracy
bwplot(resultsSVMp, main='CV Accuracy SVMp')
bwplotSVMp <- bwplot(resultsSVMp, main='CV Accuracy SVMp')
dotplotSVMp <-dotplot(resultsSVMp, main='CV Accuracy SVMp')
#bwplot(resultsSVMp, main='CV Accuracy SVMp', scales = list(relation = "free"),
#xlim = list(c(0.8, 0.85), c(0.25, 0.4)))

bwplot(resultsRFp, main='CV Accuracy RFp')
bwplotRFp <- bwplot(resultsRFp, main='CV Accuracy RFp')
dotplotRFp <- dotplot(resultsRFp, main='CV Accuracy RFp')
#bwplot(resultsRFp, main='CV Accuracy SVMp', scales = list(relation = "free"),
#xlim = list(c(0.8, 0.85), c(0.25, 0.4)))

# Print Random Forest and SVM evaluation accuracy for all mRMR selected subsets
cat('Total RF Evaluation Accuracy:',TotalEvalAccRF)
cat('Total SVM Evaluation Accuracy:',TotalEvalAccSVM)
```

```
# Respective plots for evaluation accuracy
plot(TotalEvalAccRF , main='Evaluation Accuracy RF' , xlab = 'Number of
features', ylab = 'Accuracy', type ='o')
axis(1, at = seq(1, (ncol(data)-1), by = 1))
grid( NA , 15 , lwd = 2 )

plot(TotalEvalAccSVM , main='Evaluation Accuracy SVM' , xlab = 'Number of
features', ylab = 'Accuracy', type ='o')
axis(1, at = seq(1, (ncol(data)-1), by = 1))
grid( NA , 15 , lwd = 2 )

```

Η παραπάνω σειρά εντολών αποτελεί αυτόνομο τμήμα κώδικα και υλοποιεί την προς-τα-εμπρός επιλογή χαρακτηριστικών για τους δύο καλύτερους ταξινομητές (SVM, RF) με χρήση της μεθόδου mRMR με εκτιμητή συσχέτισης Pearson. Σκοπός της συγκριτικής αυτής ανάλυσης είναι η υλοποίηση μιας feature selection wrapper μεθόδου, ώστε να ελεγχθεί η επίδραση της αύξησης των χαρακτηριστικών κατά την μοντελοποιητική διαδικασία, ως προς την ακρίβεια ταξινόμησης των παραγόμενων μοντέλων. Για το λόγο αυτό, γίνεται χρήση του αρχικού dataset που περιλαμβάνει και τα 23 features, και επαναλαμβάνονται όλες οι επιμέρους διαδικασίες που προηγούνται της μοντελοποίησης και που περιγράφηκαν παραπάνω.

Στη συνέχεια, αντί της κλασικής διαδικασίας εκπαίδευσης μοντέλων με σταθερό αριθμό χαρακτηριστικών, εκκινείται μια επαναληπτική διαδικασία για i από 1 έως 23, όπου σε κάθε επανάληψη εκτελούνται τα εξής βήματα:

i. Χρησιμοποιείται η μέθοδος mRMR Pearson για την δημιουργία φίλτρου επιλογής i βέλτιστων χαρακτηριστικών

ii. Το φίλτρο εφαρμόζεται στο αρχικό dataset, ώστε να προκύψει το νέο dataset “data.pF” που περιλαμβάνει i βέλτιστα χαρακτηριστικά βάσει του mRMR.

iii. Οι αλγόριθμοι SVM και RF υπόκεινται σε cross-validation με χρήση του “data.pF” και εξάγονται τα αντίστοιχα μοντέλα ταξινόμησης “svm.pF”, “rf.pF” για i χαρακτηριστικά. Ακολούθως, τα μοντέλα αποθηκεύονται.

iv. Το φίλτρο επιλογής χαρακτηριστικών εφαρμόζεται στο σύνολο επικύρωσης “data.evaluateF”, ώστε να προκύψει το νέο σύνολο “data.p.evaluateF” με i χαρακτηριστικά που θα χρησιμοποιηθεί για την αξιολόγηση της απόδοσης των μοντέλων σε “άγνωστα” δεδομένα.

v. Γίνεται υπολογισμός της ακρίβειας ταξινόμησης των μοντέλων “svm.pF”, “rf.pF” i χαρακτηριστικών, στο σύνολο επικύρωσης.

Μετά την ολοκλήρωση της επαναληπτικής διαδικασίας, τα 23 αποθηκευμένα μοντέλα SVM και τα 23 αντίστοιχα RF χρησιμοποιούνται για την παραγωγή συγκεντρωτικών διαγραμμάτων από τα οποία καθίσταται δυνατή η ποιοτική σύγκριση της απόδοσης (ακρίβειας ταξινόμησης) κάθε μοντέλου, τόσο στο υποσύνολο cross-validation, όσο και στο σύνολο επικύρωσης. Τα διαγράμματα αυτά θα παρουσιαστούν αναλυτικά στο Κεφάλαιο 6.

```
### I. Evaluation
```

```{r}

# Transform evaluation subset to include only mRMR Pearson selected features
data.p.evaluate  <- data.evaluate[,c(1,unname(unlist(classic1@filters)))]

# Transform evaluation subset to include only mRMR Spearman selected features
data.s.evaluate  <- data.evaluate[,c(1,unname(unlist(classic2@filters)))]

# SVM evaluation results
preds <- predict(svm, data.evaluate[,2:24])
table(preds,data.evaluate$default)
SVM_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))

# SVM.p evaluation results
preds <- predict(svm.p, data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
SVM.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# SVM.s evaluation results
preds <- predict(svm.s, data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
SVM.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# KNN evaluation results
preds <- predict(knn, data.evaluate[,2:24])
table(preds,data.evaluate$default)
KNN_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))

# KNN.p evaluation results
preds <- predict(knn.p, data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
KNN.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# KNN.s evaluation results
preds <- predict(knn.s, data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
KNN.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# NN evaluation results
```

```
preds <- predict(nn, data.evaluate[,2:24])
table(preds,data.evaluate$default)
NN_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))

# NN.p evaluation results
preds <- predict(nn.p, data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
NN.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# NN.s evaluation results
preds <- predict(nn.s, data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
NN.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# RF evaluation results
preds <- predict(rf, newdata = data.evaluate[,2:24])
table(preds,data.evaluate$default)
RF_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))

# RF.p evaluation results
preds <- predict(rf.p, newdata = data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
RF.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# RF.s evaluation results
preds <- predict(rf.s, newdata = data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
RF.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# Cumulative Evaluation results
Model    <- c( 'SVM' , 'SVM.p' , 'SVM.s' , 'KNN' , 'KNN.p' , 'KNN.s' , 'NN' ,
'NN.p' , 'NN.s' , 'RF' , 'RF.p' , 'RF.s' )
Accuracy <- c( SVM_Accuracy, SVM.p_Accuracy, SVM.s_Accuracy, KNN_Accuracy,
KNN.p_Accuracy, KNN.s_Accuracy, NN_Accuracy, NN.p_Accuracy, NN.s_Accuracy,
RF_Accuracy, RF.p_Accuracy, RF.s_Accuracy )

Evaluation_Results <- data.frame(Model,Accuracy)
Evaluation_Results

# Plot for cumulative Evaluation results
barplot(Evaluation_Results$Accuracy, names.arg = Evaluation_Results$Model,
main='Cumulative Evaluation Accuracy' , xlab = 'Classifier' , ylab = 'Accuracy',
ylim=range(0.72,0.85), xpd = FALSE )
grid(nx=NA, ny=NULL)

end_time <- Sys.time()

end_time - start_time

```

Το τελευταίο κομμάτι της αλγοριθμικής διαδικασίας είναι αφιερωμένο στον έλεγχο της απόδοσης των μοντέλων που παρήχθησαν στην αρχική διαδικασία μοντελοποίησης “G. Modeling (using svm, knn, nnet, rf)“, δηλαδή στο σύνολο μοντέλων που προέκυψαν με επιλογή 10 βέλτιστων χαρακτηριστικών, χωρίς να συγχέεται με το αμέσως προηγούμενο

αυτόνομο στάδιο. Σκοπός αυτού είναι η αξιολόγηση της απόδοσης ταξινόμησης του εκάστοτε μοντέλου σε “άγνωστο” σύνολο δεδομένων, καθώς η διαδικασία εφαρμόζεται στο dataset “data.evaluate” του τμήματος “E. Split data to CV-evaluation subsets, set resampling strategy”. Στο πλαίσιο αυτό, γίνεται υπολογισμός του μέτρου ακρίβειας ταξινόμησης (accuracy) για καθένα από τα 12 παραχθέντα μοντέλα ως προς το σύνολο επικύρωσης και παράγεται το αντίστοιχο συγκεντρωτικό διάγραμμα αξιολόγησης που παρατίθεται στο Κεφάλαιο 6.

Το παρόν κεφάλαιο επικεντρώθηκε στην παρουσίαση του τρόπου αντιμετώπισης ενός πραγματικού προβλήματος δυαδικής ταξινόμησης, περιλαμβάνοντας την περιγραφή του συνόλου δεδομένων, την καταγραφή σε high-level της μεθοδολογίας ανάλυσης με απώτερο σκοπό την ανάδειξη του βέλτιστου μοντέλου ταξινόμησης και τέλος, την αποτύπωση των επιμέρους βημάτων σε επίπεδο κώδικα. Τα αποτελέσματα που προέκυψαν από τα στάδια της αλγοριθμικής διαδικασίας έχουν σκοπίμως συμπεριληφθεί μαζικά στο Κεφάλαιο 6, ώστε να αποτελέσουν πηγή εξαγωγής χρήσιμων συμπερασμάτων για την αποδοτικότητα της συνολικής διαδικασίας καθώς και για να πυροδοτήσουν τυχόν ιδέες για περαιτέρω βελτιώσεις, προσθήκες και μελλοντική εξέλιξη του παρουσιασθείσας στρατηγικής ανάλυσης.

ΚΕΦΑΛΑΙΟ 6

Συγκριτική αξιολόγηση μεθόδων, Αποτελέσματα και Συμπεράσματα

6.1 Εισαγωγή

Το παρόν κεφάλαιο αποτελεί το καταληκτικό τμήμα της παρούσας εργασίας και συγκεντρώνει τα αποτελέσματα της αλγοριθμικής διαδικασίας, η οποία περιγράφηκε αναλυτικά στο προηγούμενο κεφάλαιο (Κεφάλαιο 5). Βάσει των αποτελεσμάτων που θα παρουσιαστούν παρακάτω, στις Υποενότητες 6.2 έως 6.4 εξάγονται χρήσιμες παρατηρήσεις σχετικά με την απόδοση της μεθόδου επιλογής χαρακτηριστικών mRMR και των τεχνικών μηχανικής μάθησης, οι οποίες αποτέλεσαν αντικείμενο συγκριτικής αξιολόγησης για την εξαγωγή του βέλτιστου μοντέλου δυαδικής ταξινόμησης, όσον αφορά την ακρίβεια ταξινόμησης και άλλους παράγοντες που αναλύονται στη συνέχεια. Επιπλέον, η Υποενότητα 6.5 συνοψίζει σε μορφή συμπερασμάτων τα βασικά ευρήματα της συνολικής μελέτης, ενώ στην τελευταία υποενότητα προτείνονται ιδέες για μελλοντική συνέχιση της έρευνας.

6.2 Περιγραφή αποτελεσμάτων μοντελοποίησης με χρήση cross-validation

Σε αυτή την υποενότητα αναλύονται τα αποτελέσματα που προέκυψαν από την μοντελοποιητική διαδικασία με χρήση cross-validation, ακολουθώντας την μεθοδολογία

που περιγράφηκε εκτενώς στην Υποενότητα 5.3. Βάσει της διαδικασίας αυτής προέκυψαν πολύ ενδιαφέροντα συγκεντρωτικά αποτελέσματα, τα οποία επέτρεψαν την επιλογή των καλύτερων ταξινομητών για χρήση τους σε περαιτέρω δοκιμές.

Τα αποτελέσματα που παρουσιάζονται ακολούθως αφορούν την ακρίβεια/πιστότητα ταξινόμησης (accuracy) και τον συντελεστή kappa (Cohen's kappa) (Sasikala et al., 2017). Υπενθυμίζεται ότι η ακρίβεια ταξινόμησης αποτελεί το πλέον σύνηθες μέτρο εκτίμησης της δυνατότητας πρόβλεψης των μοντέλων και υπολογίζεται ως το κλάσμα σωστών ταξινομήσεων προς το σύνολο όλων των ταξινομήσεων. Ο δείκτης kappa, υπολογίζεται από την R ως δευτερεύον μέτρο εκτίμησης και συνδέει το επίπεδο της παρατηρηθείσας συμφωνίας με το επίπεδο της τυχαίας συμφωνίας, εκτιμώντας τη μεταβλητότητα που υπάρχει σε κάθε ταξινομητή. Περιλαμβάνεται στα αποτελέσματα για λόγους πληρότητας, καθώς η σύγκριση των ταξινομητών ως προς την απόδοση θα περιοριστεί στην εξέταση της ακρίβειας ταξινόμησης του καθενός.

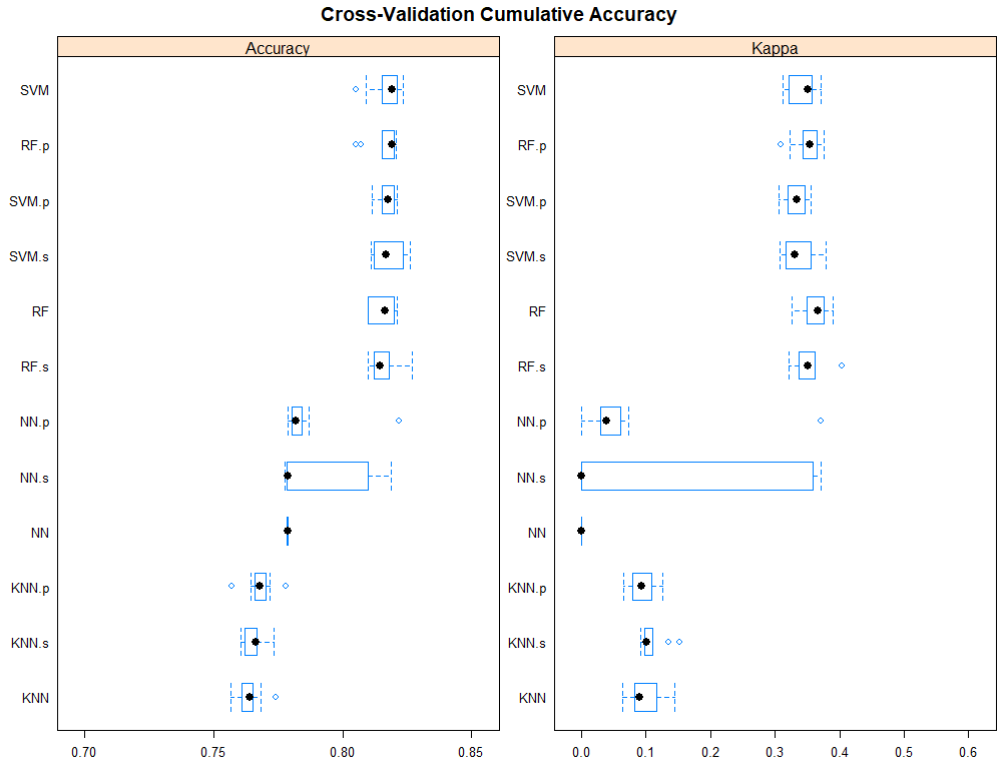
Ακολουθούν τα συγκεντρωτικά αποτελέσματα ακρίβειας των μοντελοποιητικών οικογενειών SVM, RF, NN (nnet) και kNN οι οποίες υποβλήθηκαν σε σύγκριση λαμβάνοντας υπόψη 3 διαφορετικές υποπεριπτώσεις (χωρίς επιλογή χαρακτηριστικών, με επιλογή χαρακτηριστικών mRMR και εκτιμητή συσχέτισης Pearson, με επιλογή χαρακτηριστικών mRMR και εκτιμητή συσχέτισης Spearman). Παρατίθενται λοιπόν τα αποτελέσματα 12 μοντέλων ταξινόμησης, τόσο ως συνόψεις 5 αριθμών όπως εξήχθησαν από το workspace της R, όσο και ως σχηματική αναπαράσταση σε μορφή boxplot (Σχήμα 6.1):

```
Accuracy
```

	Min.	1stQu.	Median	Mean	3rdQu.	Max.	NA's
SVM	0.8051731	0.8157071	0.8191865	0.8171619	0.8207804	0.8235294	0
SVM.p	0.8115096	0.8154328	0.8175183	0.8170785	0.8197184	0.8211009	0
SVM.s	0.8110138	0.8127606	0.8168163	0.8179540	0.8229775	0.8260325	0
KNN	0.7567793	0.7611349	0.7641785	0.7638073	0.7651168	0.7743012	0
KNN.p	0.7570952	0.7659818	0.7678831	0.7678534	0.7698883	0.7780559	0
KNN.s	0.7605340	0.7630127	0.7664233	0.7659771	0.7668648	0.7734668	0
NN	0.7784731	0.7785655	0.7788441	0.7787419	0.7788903	0.7788903	0
NN.p	0.7787980	0.7806862	0.7819016	0.7859170	0.7840235	0.8218607	0
NN.s	0.7776387	0.7785655	0.7788903	0.7892135	0.8020442	0.8188648	0
RF	0.8097622	0.8106561	0.8164755	0.8156184	0.8199270	0.8210263	0
RF.p	0.8051731	0.8154794	0.8191113	0.8163274	0.8198311	0.8206839	0
RF.s	0.8097622	0.8125977	0.8143899	0.8157852	0.8181060	0.8268669	0

Kappa

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
SVM	0.31270737	0.32764212	0.35039565	0.34465530	0.35765276	0.3709704	0
SVM.p	0.30663937	0.32268805	0.33386611	0.33292288	0.34427483	0.3556021	0
SVM.s	0.30733947	0.31828513	0.33025986	0.33646681	0.35215506	0.3796803	0
KNN	0.06424369	0.08292181	0.09012064	0.09866079	0.11418437	0.1440873	0
KNN.p	0.06515175	0.08060248	0.09388803	0.09504636	0.10718406	0.1267305	0
KNN.s	0.09267893	0.09889830	0.10200630	0.10952320	0.10956600	0.1528839	0
NN	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.0000000	0
NN.p	0.00000000	0.03162179	0.03896099	0.06938627	0.05676438	0.3710668	0
NN.s	0.00000000	0.00000000	0.00000000	0.12226402	0.30121955	0.3712281	0
RF	0.32669286	0.35257976	0.36622905	0.36349493	0.37510403	0.3893581	0
RF.p	0.30874286	0.34578058	0.35381460	0.34997756	0.36335102	0.3755027	0
RF.s	0.32110151	0.33892332	0.35099640	0.35285312	0.36211504	0.4045654	0



ΣΧΗΜΑ 6.1 Αποτελέσματα συγκριτικής αξιολόγησης όλων των ταξινομητών με χρήση 10-fold cross validation

Όπως γίνεται εμφανές από τα παραπάνω αποτελέσματα, οι μοντελοποιητικές οικογένειες Μηχανών Διανυσμάτων Υποστήριξης (SVM) και Τυχαίων Δασών (RF) παρουσίασαν καλύτερη ακρίβεια ταξινόμησης από τα μοντέλα βασισμένα σε Πολυεπίπεδο Δικτυο Perceptron (MLP/ nnet) και μέθοδο k-Πλησιέστερων Γειτόνων (kNN). Το πιο ενδιαφέρον συμπέρασμα, ωστόσο, έγκειται στο γεγονός ότι τα μοντέλα SVM και RF στα οποία προηγήθηκε η μέθοδος επιλογής χαρακτηριστικών mRMR (SVM.p, SVM.s,

RF.p, RF.s) για εξαγωγή των 10 βέλτιστων χαρακτηριστικών, εμφάνισαν ακρίβεια περίπου ίση με τα αντίστοιχα μοντέλα που αξιοποίησαν και τα 23 χαρακτηριστικά. Η παρατήρηση αυτή δημιούργησε την ανάγκη για περαιτέρω έλεγχο της επίδρασης των χαρακτηριστικών στην απόδοση πρόβλεψης, οδηγώντας στην συγκριτική ανάλυση με χρήση forward feature selection που περιγράφηκε στην Υποενότητα 5.3 και τα αποτελέσματα της οποίας αναλύονται στη συνέχεια.

6.3 Περιγραφή αποτελεσμάτων συγκριτικής αξιολόγησης των δύο καλύτερων ταξινομητών

Στο σημείο αυτό περιγράφονται τα αποτελέσματα της διαδικασίας προς-τα-εμπρός επιλογής χαρακτηριστικών για τους δύο καλύτερους ταξινομητές (SVM, RF) με χρήση της μεθόδου mRMR με εκτιμητή συσχέτισης Pearson, η οποία υλοποιήθηκε στο επόμενο στάδιο “H. Comparative modeling with forward feature selection (mRMR Pearson) between two best classifiers, with respect to CV accuracy” του κώδικα. Υπενθυμίζεται ότι στόχος της διαδικασίας ήταν ο έλεγχος της επίδρασης της αύξησης των χαρακτηριστικών κατά τη μοντελοποιητική διαδικασία, γι' αυτό το λόγο τα αποτελέσματα που παρουσιάζονται αφορούν την απόδοση των αλγορίθμων με 1 έως και 23 διαθέσιμα χαρακτηριστικά. Συγκεκριμένα, στην Υποενότητα 6.3.1 παρατίθενται και αναλύονται τα αποτελέσματα συγκριτικής αξιολόγησης των δύο μοντελοποιητικών οικογενειών από τη χρήση 10-fold cross validation, ενώ στην Υποενότητα 6.3.2 η απόδοση των ίδιων εκπαιδευμένων μοντέλων συγκρίνεται στο υποσύνολο επικύρωσης.

6.3.1 Ανάλυση αποτελεσμάτων συγκριτικής αξιολόγησης με χρήση cross-validation

Αρχικά παρατίθενται τα αποτελέσματα του ταξινομητή SVM με επιλογή χαρακτηριστικών mRMR και εκτιμητή συσχέτισης Pearson για 1 έως 23 χαρακτηριστικά που προέκυψαν

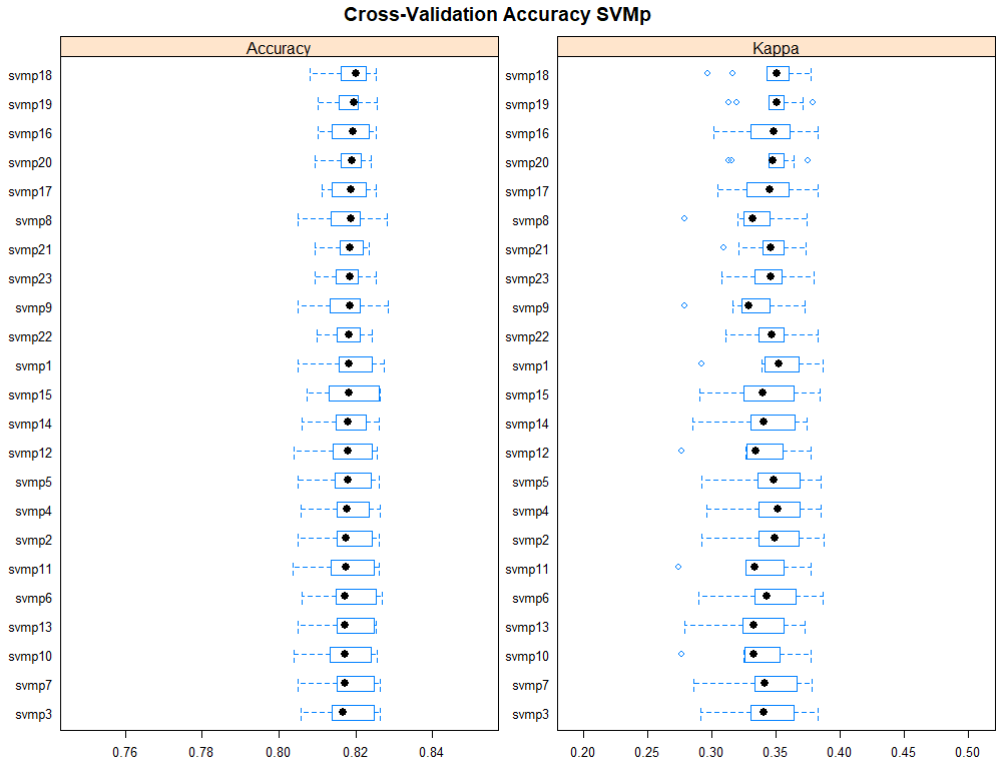
από τη χρήση cross-validation, πρώτα ως συνόψεις 5 αριθμών καθώς και ως γραφική απεικόνιση μέσω boxplot διαγραμμάτων στην συνέχεια:

```
Accuracy

            Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
svmp1   0.8048374 0.8159623 0.8181818 0.8189142 0.8240509 0.8272841    0
svmp2   0.8048374 0.8152513 0.8175182 0.8184137 0.8240509 0.8260325    0
svmp3   0.8056714 0.8142080 0.8166839 0.8180382 0.8243638 0.8264497    0
svmp4   0.8056714 0.8152513 0.8177265 0.8184971 0.8235294 0.8264497    0
svmp5   0.8048374 0.8149383 0.8179351 0.8184554 0.8239466 0.8260325    0
svmp6   0.8060884 0.8151005 0.8173095 0.8187473 0.8249896 0.8268669    0
svmp7   0.8048374 0.8153092 0.8171011 0.8183719 0.8245724 0.8264497    0
svmp8   0.8048374 0.8136795 0.8187693 0.8178296 0.8209220 0.8281185    0
svmp9   0.8048374 0.8132430 0.8185232 0.8176210 0.8207322 0.8285357    0
svmp10  0.8040033 0.8134711 0.8173095 0.8174959 0.8232165 0.8256154    0
svmp11  0.8035863 0.8138881 0.8175181 0.8176211 0.8235294 0.8260325    0
svmp12  0.8040033 0.8142009 0.8179352 0.8177880 0.8233208 0.8256154    0
svmp13  0.8048374 0.8152049 0.8173095 0.8177044 0.8236337 0.8251982    0
svmp14  0.8060884 0.8152048 0.8179353 0.8182884 0.8223821 0.8260325    0
svmp15  0.8073394 0.8138490 0.8181437 0.8185803 0.8247810 0.8264497    0
svmp16  0.8102585 0.8139535 0.8193953 0.8184968 0.8234251 0.8251982    0
svmp17  0.8110926 0.8141620 0.8187695 0.8183716 0.8224864 0.8251982    0
svmp18  0.8081735 0.8163905 0.8199833 0.8182880 0.8222965 0.8251982    0
svmp19  0.8101002 0.8160777 0.8196040 0.8182464 0.8206091 0.8256154    0
svmp20  0.8092654 0.8163905 0.8189784 0.8179543 0.8213392 0.8239466    0
svmp21  0.8094245 0.8160392 0.8185613 0.8177458 0.8216521 0.8235294    0
svmp22  0.8098415 0.8152049 0.8183527 0.8177458 0.8209220 0.8243638    0
svmp23  0.8094245 0.8148920 0.8185612 0.8176624 0.8204005 0.8251982    0


Kappa

            Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
svmp1   0.2917799 0.3429191 0.3523735 0.3522329 0.3670960 0.3868417    0
svmp2   0.2917799 0.3379320 0.3488224 0.3494742 0.3670960 0.3878077    0
svmp3   0.2912210 0.3306407 0.3404271 0.3434050 0.3634330 0.3827045    0
svmp4   0.2959936 0.3373380 0.3513108 0.3498907 0.3670668 0.3848715    0
svmp5   0.2917799 0.3367476 0.3481602 0.3483844 0.3669781 0.3848715    0
svmp6   0.2897327 0.3339346 0.3426489 0.3459745 0.3649881 0.3867264    0
svmp7   0.2857581 0.3339026 0.3416343 0.3446990 0.3661003 0.3777989    0
svmp8   0.2783954 0.3255125 0.3322685 0.3343139 0.3443287 0.3745601    0
svmp9   0.2783954 0.3239056 0.3290842 0.3329219 0.3440480 0.3725401    0
svmp10  0.2765546 0.3256327 0.3326769 0.3359945 0.3523156 0.3777378    0
svmp11  0.2743931 0.3270647 0.3331473 0.3362796 0.3533364 0.3777378    0
svmp12  0.2765546 0.3278106 0.3340972 0.3366668 0.3528414 0.3777378    0
svmp13  0.2783954 0.3247911 0.3323860 0.3345862 0.3532991 0.3726148    0
svmp14  0.2848641 0.3309478 0.3405052 0.3413656 0.3619741 0.3740126    0
svmp15  0.2900861 0.3267500 0.3398930 0.3417158 0.3606488 0.3847414    0
svmp16  0.3014404 0.3307801 0.3485426 0.3452662 0.3608933 0.3828501    0
svmp17  0.3045110 0.3289933 0.3448914 0.3440944 0.3595968 0.3828501    0
svmp18  0.2967715 0.3432552 0.3509311 0.3465439 0.3602549 0.3770832    0
svmp19  0.3131898 0.3444522 0.3506699 0.3482629 0.3555237 0.3790565    0
svmp20  0.3131898 0.3450797 0.3476076 0.3466262 0.3561907 0.3751066    0
svmp21  0.3090087 0.3409740 0.3456672 0.3453861 0.3564320 0.3731268    0
svmp22  0.3111010 0.3377685 0.3471499 0.3455644 0.3544034 0.3829459    0
svmp23  0.3078425 0.3353926 0.3460798 0.3447195 0.3531542 0.3799595    0
```



ΣΧΗΜΑ 6.2 Αποτελέσματα συγκριτικής αξιολόγησης μεθόδου SVM με επιλογή χαρακτηριστικών mRMR Pearson για 1 έως 23 χαρακτηριστικά και με χρήση cross-validation

Ακολουθούν τα αντίστοιχα αποτελέσματα για το ταξινομητή RF με επιλογή χαρακτηριστικών mRMR και εκτιμητή συσχέτισης Pearson για 1 έως 23 χαρακτηριστικά:

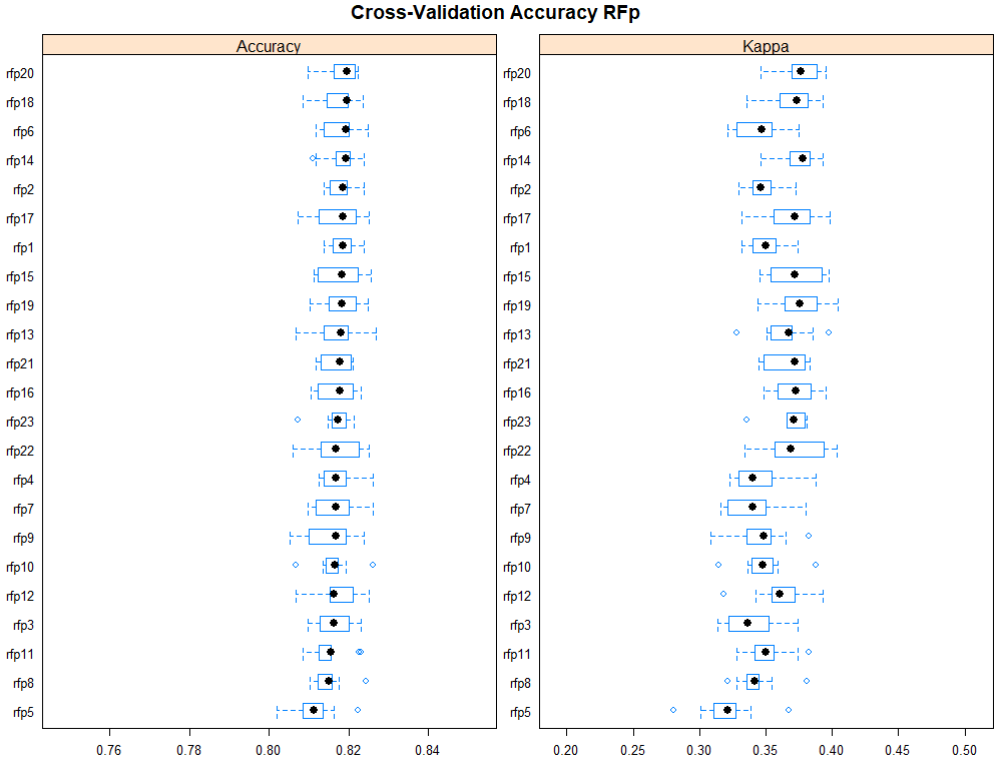
Accuracy

	Min.	1stQu.	Median	Mean	3rdQu.	Max.	NA's
rfp1	0.8139341	0.8163905	0.8185611	0.8186636	0.8203255	0.8240200	0
rfp2	0.8139341	0.8157840	0.8186940	0.8184549	0.8195286	0.8239466	0
rfp3	0.8098415	0.8128454	0.8163987	0.8161607	0.8195851	0.8231860	0
rfp4	0.8126825	0.8141620	0.8168545	0.8175372	0.8191967	0.8261051	0
rfp5	0.8019183	0.8088235	0.8112226	0.8114053	0.8134974	0.8223520	0
rfp6	0.8118481	0.8141620	0.8193577	0.8178293	0.8201731	0.8248540	0
rfp7	0.8098415	0.8127414	0.8168544	0.8167447	0.8198601	0.8261051	0
rfp8	0.8102585	0.8127019	0.8151471	0.8152427	0.8158774	0.8244370	0
rfp9	0.8051731	0.8116587	0.8168542	0.8156599	0.8191489	0.8240200	0
rfp10	0.8068419	0.8145329	0.8165693	0.8162854	0.8173287	0.8261051	0
rfp11	0.8085106	0.8131454	0.8156028	0.8155348	0.8156797	0.8231860	0
rfp12	0.8068419	0.8154328	0.8164372	0.8170365	0.8205608	0.8252711	0
rfp13	0.8068419	0.8140190	0.8181060	0.8173699	0.8197456	0.8269391	0
rfp14	0.8110926	0.8171289	0.8193575	0.8183299	0.8202481	0.8238731	0
rfp15	0.8113523	0.8133670	0.8183525	0.8180375	0.8218119	0.8255426	0
rfp16	0.8105966	0.8123892	0.8178974	0.8169943	0.8210636	0.8231860	0
rfp17	0.8072591	0.8130803	0.8185613	0.8172449	0.8218421	0.8252711	0
rfp18	0.8085106	0.8154215	0.8195661	0.8177037	0.8198499	0.8236030	0
rfp19	0.8102585	0.8156493	0.8183146	0.8180792	0.8210937	0.8248540	0
rfp20	0.8098415	0.8166267	0.8196038	0.8180795	0.8212907	0.8223520	0
rfp21	0.8118481	0.8139990	0.8178974	0.8172867	0.8206465	0.8211009	0

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
rfp22	0.8060884	0.8130412	0.8168920	0.8169114	0.8220880	0.8252711	0
rfp23	0.8072591	0.8161710	0.8173478	0.8169531	0.8191967	0.8214435	0

Kappa

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
rfp1	0.3315488	0.3416093	0.3496463	0.3511021	0.3561500	0.3740569	0
rfp2	0.3293061	0.3405411	0.3462829	0.3490960	0.3531694	0.3727629	0
rfp3	0.3135977	0.3235300	0.3366143	0.3390969	0.3521226	0.3741932	0
rfp4	0.3230913	0.3306437	0.3404879	0.3461383	0.3542677	0.3880474	0
rfp5	0.2805803	0.3107413	0.3210785	0.3202167	0.3263754	0.3670777	0
rfp6	0.3212177	0.3292403	0.3467721	0.3446067	0.3542890	0.3749573	0
rfp7	0.3157086	0.3252162	0.3405049	0.3394914	0.3486728	0.3799360	0
rfp8	0.3211820	0.3362318	0.3415143	0.3429851	0.3442766	0.3811655	0
rfp9	0.3087429	0.3372549	0.3486848	0.3457811	0.3533757	0.3822251	0
rfp10	0.3146637	0.3397459	0.3475767	0.3481538	0.3548283	0.3880474	0
rfp11	0.3283480	0.3433001	0.3502909	0.3518394	0.3555906	0.3823200	0
rfp12	0.3180420	0.3545223	0.3602991	0.3598953	0.3707915	0.3930522	0
rfp13	0.3279799	0.3550011	0.3676245	0.3649459	0.3697596	0.3978746	0
rfp14	0.3464554	0.3691569	0.3776243	0.3742623	0.3831889	0.3931963	0
rfp15	0.3455353	0.3568789	0.3718559	0.3727166	0.3902039	0.3976366	0
rfp16	0.3486405	0.3603928	0.3727981	0.3724224	0.3840795	0.3950850	0
rfp17	0.3321348	0.3562643	0.3717475	0.3683033	0.3827510	0.3979113	0
rfp18	0.3359361	0.3628172	0.3731235	0.3701556	0.3811556	0.3931351	0
rfp19	0.3440801	0.3660811	0.3758557	0.3741896	0.3854890	0.4045779	0
rfp20	0.3463136	0.3698794	0.3760835	0.3745113	0.3868751	0.3951475	0
rfp21	0.3443356	0.3531321	0.3717853	0.3673735	0.3791654	0.3833202	0
rfp22	0.3339438	0.3567432	0.3685664	0.3705846	0.3901167	0.4036405	0
rfp23	0.3353503	0.3664983	0.3712555	0.3693105	0.3785047	0.3812851	0



ΣΧΗΜΑ 6.3 Αποτελέσματα συγκριτικής αξιολόγησης μεθόδου RF με επιλογή χαρακτηριστικών mRMR Pearson για 1 έως 23 χαρακτηριστικά και με χρήση cross-validation

Από τα παραπάνω αποτελέσματα, μπορεί να συμπεράνει κανείς ότι η ακρίβεια ταξινόμησης δε μεταβάλλεται σημαντικά από το πλήθος των χαρακτηριστικών που αξιοποιούνται από τις δύο μεθόδους. Συγκεκριμένα, εξάγονται οι ακόλουθες παρατηρήσεις:

- ✔ Η διαφοροποίηση που εμφανίζεται στην ακρίβεια ταξινόμησης κατά την αύξηση του αριθμού των χαρακτηριστικών είναι αμελητέα (της τάξεως του 3ου δεκαδικού ψηφίου).
- ✔ Αξιοσημείωτο είναι το γεγονός ότι η ακρίβεια ταξινόμησης για χρήση του συνόλου και των 23 χαρακτηριστικών τόσο για τον SVM όσο και για τον RF είναι μικρότερη από την ακρίβεια ταξινόμησης για ένα σημαντικά μικρότερο υποσύνολο χαρακτηριστικών (πχ. 8 χαρακτηριστικά για τον SVM και 6 χαρακτηριστικά για τον RF).
- ✔ Ακόμα και με τη χρήση ενός μόνο χαρακτηριστικού, οι δυνατότητες πρόβλεψης των δύο μοντελοποιητικών οικογενειών είναι ιδιαίτερα ικανοποιητικές, καθώς προσεγγίζουν την ακρίβεια ταξινόμησης των βέλτιστων ταξινομητών. Το γεγονός αυτό προφανώς σχετίζεται άμεσα με τις ιδιότητες των χαρακτηριστικών του χρησιμοποιούμενου dataset και κυρίως με την ανεξάρτητη μεταβλητή PAY_0, η οποία επιλέχθηκε ως το βέλτιστο χαρακτηριστικό από τη μέθοδο mRMR.
- ✔ Από τα παραπάνω, είναι ασφαλές να συμπεράνει κανείς ότι η ικανότητα πληρωμής (εξαρτημένη μεταβλητή default) είναι άρρηκτα συνδεδεμένη με το ιστορικό προηγούμενης πληρωμής για τον μήνα Σεπτέμβριο του 2005 (ανεξάρτητη μεταβλητή PAY_0).

6.3.2 Ανάλυση αποτελεσμάτων συγκριτικής αξιολόγησης στο σύνολο επικύρωσης

Στη συνέχεια παρατίθενται τα αποτελέσματα των δύο μοντελοποιητικών οικογενειών SVM και RF για 1 έως 23 χαρακτηριστικά, στο υποσύνολο επικύρωσης. Σκοπός της διαδικασίας είναι η αξιολόγηση της απόδοσης μοντέλων που εκπαιδεύτηκαν με χρήση

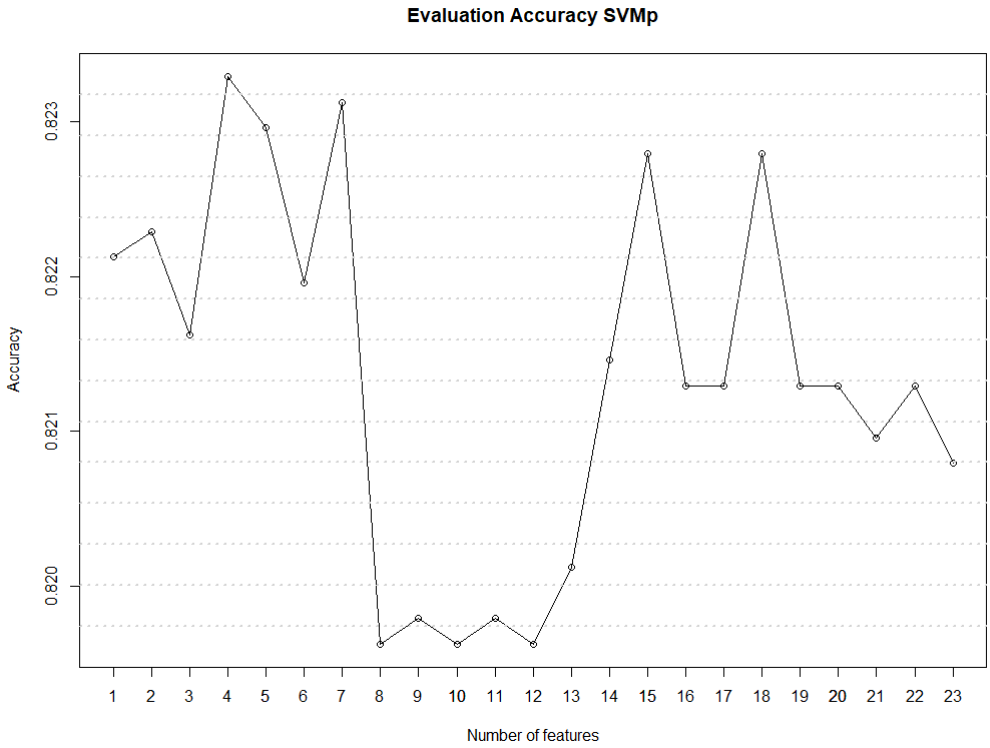
cross-validation, αναθέτοντάς τους την ταξινόμηση σε κάποιο "άγνωστο" για αυτά σύνολο δεδομένων, ώστε να διαπιστωθεί εάν αυτά θα είχαν τις αναμενόμενες επιδόσεις πρόβλεψης σε μελλοντικές περιπτώσεις αδυναμίας πληρωμής. Σημειώνεται ότι μέτρο απόδοσης σε αυτήν την περίπτωση είναι η (μοναδική) τιμή accuracy, καθώς η αξιολόγηση των μοντέλων γίνεται σε ένα σύνολο επικύρωσης σε αντίθεση με την τεχνική cross-validation.

Στον επόμενο πίνακα παρουσιάζεται η επίδραση της προσθήκης χαρακτηριστικών στην ακρίβεια ταξινόμησης του αλγορίθμου SVM (Πίνακας 6.1):

ΠΙΝΑΚΑΣ 6.1 Συγκεντρωτικά αποτελέσματα απόδοσης SVM με επιλογή χαρακτηριστικών mRMR Pearson για 1 έως 23 χαρακτηριστικά στο σύνολο επικύρωσης

Number of best features	SVM Accuracy
1	0.8221258
2	0.8222927
3	0.8216252
4	0.8232938
5	0.8229601
6	0.821959
7	0.823127
8	0.8196229
9	0.8197898
10	0.8196229
11	0.8197898
12	0.8196229
13	0.8201235
14	0.8214584
15	0.8227933
16	0.8212915
17	0.8212915
18	0.8227933
19	0.8212915
20	0.8212915
21	0.8209578
22	0.8212915
23	0.8207909

Η παραπάνω διακύμανση στην απόδοση (ακρίβεια ταξινόμησης) αποτυπώνεται επίσης στο ακόλουθο διάγραμμα (Σχήμα 6.4):



ΣΧΗΜΑ 6.4 Διάγραμμα απόδοσης SVM με επιλογή χαρακτηριστικών mRMR Pearson για 1 έως 23 χαρακτηριστικά στο σύνολο επικύρωσης

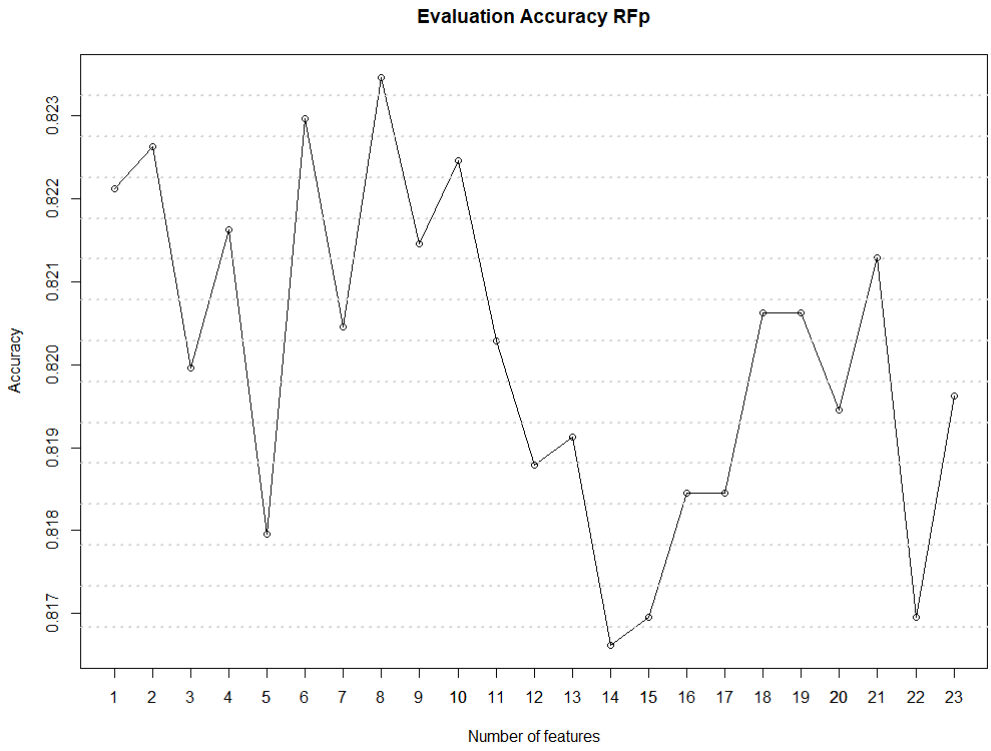
Στον επόμενο πίνακα παρουσιάζεται η αντίστοιχη επίδραση της προσθήκης χαρακτηριστικών στην ακρίβεια ταξινόμησης του αλγορίθμου RF (Πίνακας 6.2):

ΠΙΝΑΚΑΣ 6.2 Συγκεντρωτικά αποτελέσματα απόδοσης RF με επιλογή χαρακτηριστικών mRMR Pearson για 1 έως 23 χαρακτηριστικά στο σύνολο επικύρωσης

Number of best features	Random Forest Accuracy
1	0.8221258
2	0.8226264
3	0.8199566
4	0.821625
5	0.8179543
6	0.8229601
7	0.8204572
8	0.8234607
9	0.8214584
10	0.8224595
11	0.8202903
12	0.8187886
13	0.8191223

Number of best features	Random Forest Accuracy
14	0.8166194
15	0.8169531
16	0.8184549
17	0.8184549
18	0.8206241
19	0.8206241
20	0.819456
21	0.8212915
22	0.8169531
23	0.8189554

Η παραπάνω διακύμανση στην απόδοση (ακρίβεια ταξινόμησης) αποτυπώνεται επίσης στο ακόλουθο διάγραμμα (Σχήμα 6.5):



ΣΧΗΜΑ 6.5 Διάγραμμα απόδοσης RF με επιλογή χαρακτηριστικών mRMR Pearson για 1 έως 23 χαρακτηριστικά στο σύνολο επικύρωσης

Από την ανάλυση των παραπάνω αποτελεσμάτων για τις δύο μοντελοποιητικές οικογένειες, παρατηρείται συνολικά πολύ μικρή διακύμανση στην ακρίβεια ταξινόμησης

στο σύνολο επικύρωσης. Πιο συγκεκριμένα εξάγονται τα παρακάτω συμπεράσματα:

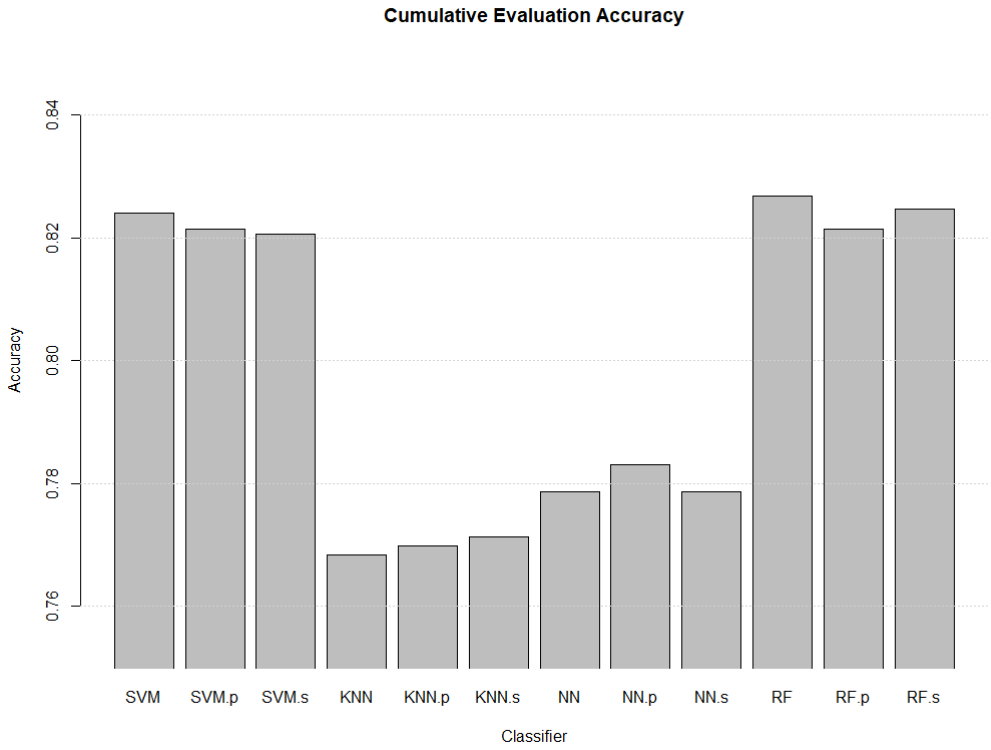
- ✔ Επιβεβαιώνεται η αρχική παρατήρηση (Υποενότητα 6.3.1) ότι η εκπαίδευση ταξινομητών με χρήση ενός μικρού υποσυνόλου των διαθέσιμων χαρακτηριστικών μπορεί να οδηγήσει ακόμη και σε μεγαλύτερη ακρίβεια ταξινόμησης, σε σχέση με την εκπαίδευση που γίνεται με χρήση του συνολικού αριθμού χαρακτηριστικών.
- ✔ Τα καλύτερα αποτελέσματα ταξινόμησης εμφανίζονται με χρήση μόλις 4 βέλτιστων χαρακτηριστικών για τον SVM και 8 βέλτιστων χαρακτηριστικών για τον RF, γεγονός που επισημαίνει την αναγκαιότητα χρήσης μεθόδων επιλογής χαρακτηριστικών (όπως η mRMR) με στόχο τη βελτιστοποίηση της μοντελοποιητικής διαδικασίας.
- ✔ Η μικρή μείωση της ακρίβειας ταξινόμησης που παρατηρείται για τον SVM κατά τη χρήση 8 έως 13 βέλτιστων χαρακτηριστικών καθώς και η ανάλογη μείωση για τον RF κατά τη χρήση 12 έως 17 χαρακτηριστικών αποδίδεται τόσο στις ιδιαιτερότητες εκπαίδευσης της κάθε μεθόδου, όσο και στο ενδεχόμενο ύπαρξης θορύβου που δυσχεραίνει την ορθή ταξινόμηση των δειγμάτων στις πραγματικές τους κλάσεις.

6.4 Περιγραφή αποτελεσμάτων στο υποσύνολο επικύρωσης

Στην παρούσα υποενότητα αναλύονται τα αποτελέσματα ακρίβειας ταξινόμησης από τη χρήση των 12 μοντέλων σε “άγνωστο” για αυτά σύνολο επικύρωσης. Τα αντίστοιχα αποτελέσματα εκπαίδευσης αυτών των μοντέλων με χρήση cross-validation παρουσιάστηκαν αναλυτικά στην Υποενότητα 6.2. Κατά αντιστοιχία με την προηγούμενη υποενότητα τα αποτελέσματα που προέκυψαν παρατίθενται αρχικά σε μορφή πίνακα (Πίνακας 6.3) και στην συνέχεια αποτυπώνονται και σε συγκεντρωτικό ραβδόγραμμα (Σχήμα 6.6):

ΠΙΝΑΚΑΣ 6.3 Συγκεντρωτικά αποτελέσματα απόδοσης ταξινομητών στο σύνολο επικύρωσης

Model	Accuracy
SVM	0.8241281
SVM.p	0.8214584
SVM.s	0.8206241
KNN	0.7683965
KNN.p	0.7698982
KNN.s	0.7714000
NN	0.7787419
NN.p	0.7830803
NN.s	0.7787419
RF	0.8267979
RF.p	0.8214584
RF.s	0.8247956



ΣΧΗΜΑ 6.6 Ραβδόγραμμα συγκεντρωτικών αποτελεσμάτων απόδοσης ταξινομητών στο σύνολο επικύρωσης

Τα παραπάνω αποτελέσματα φαίνεται να συμφωνούν με τις αρχικές αξιολογήσεις των μοντέλων βάσει της χρήσης cross-validation (Υποενότητα 6.2). Συγκεκριμένα, από την ανάλυση των αποτελεσμάτων παρατηρούμε τα εξής:

- ✔ Οι μοντελοποιητικές οικογένειες SVM και RF επιβεβαιώνουν την αρχική μας επιλογή ως οι βέλτιστοι ταξινομητές, καθώς παρουσιάζουν την μεγαλύτερη ακρίβεια ταξινόμησης στο σύνολο επικύρωσης. Αντίστοιχα, τα μοντέλα που βασίζονται στις μεθόδους kNN και NN(nnet) εμφάνισαν μικρότερη ακρίβεια, όπως αναμενόταν από τα αποτελέσματα της Υποενότητας 6.2.
- ✔ Η απόδοση των ταξινομητών όσον αφορά την ακρίβεια ταξινόμησης φαίνεται να επηρεάζεται ελάχιστα από τον αριθμό των διαθέσιμων χαρακτηριστικών. Συγκεκριμένα, οι μέθοδοι στις οποίες έχει προηγηθεί επιλογή 10 βέλτιστων χαρακτηριστικών με χρήση mRMR (κατάληξη .s ή .p) έχουν αμελητέες διαφορές στην ακρίβεια σε σχέση με τις αντίστοιχές τους που χρησιμοποιούν και τα 23 χαρακτηριστικά.
- ✔ Η απόδοση των "νικητήριων μοντέλων" (SVM, RF) μπορεί να βελτιωθεί περαιτέρω, όπως φάνηκε στην Υποενότητα 6.3, καθώς σε συνδυασμό με τη μέθοδο επιλογής χαρακτηριστικών mRMR κατέστη δυνατή η εκπαίδευση μοντέλων με μεγαλύτερη ακρίβεια ταξινόμησης, χρησιμοποιώντας μονοψήφιο αριθμό χαρακτηριστικών (βλ. Υποενότητα 6.3.2).

Κλείνοντας το κομμάτι παρουσίασης των αποτελεσμάτων, αξίζει να σημειωθεί ότι η παραπάνω συγκριτική αξιολόγηση επικεντρώθηκε μονάχα στον έλεγχο της ακρίβειας ταξινόμησης για την εξαγωγή της βέλτιστης μεθόδου. Ωστόσο, η ακρίβεια ταξινόμησης δεν αποτελεί το μοναδικό κριτήριο επιλογής μιας μεθόδου εξόρυξης δεδομένων, ιδιαίτερα σε περιπτώσεις που πρέπει να δοθεί έμφαση στην ταχύτητα παραγωγής αποτελεσμάτων, στο υπολογιστικό κόστος και στην πολυπλοκότητα των εξαχθέντων μοντέλων (Yao et al., 2017). Κρίνεται λοιπόν σκόπιμο, να διευκρινιστούν τα εξής:

- ✔ Όσον αφορά καθαρά την ακρίβεια ταξινόμησης, είναι εμφανές ότι η μοντελοποιητική οικογένεια SVM έχει ελάχιστα καλύτερη απόδοση από την RF, είτε με χρήση μεθόδων επιλογής χαρακτηριστικών, είτε χωρίς.
- ✔ Όσον αφορά το υπολογιστικό κόστος που σχετίζεται άμεσα με την πολυπλοκότητα των δύο μεθόδων, αξίζει να σημειωθεί ότι ο χρόνος εκπαίδευσης των μοντέλων RF ήταν τάξεις μεγέθους μικρότερος από τον αντίστοιχο χρόνο εκπαίδευσης μοντέλων SVM, καθιστώντας την χρήση μοντέλων που βασίζονται στην οικογένεια SVM

ιδιαίτερα χρονοβόρα για μεγάλα σύνολα δεδομένων. Ιδιαίτερα σε περιπτώσεις στις οποίες ο μεγάλος αριθμός των διαθέσιμων χαρακτηριστικών οδηγεί στην αύξηση των υπολογιστικών και αποθηκευτικών απαιτήσεων, η εκπαίδευση μοντέλων SVM χωρίς να προηγηθεί μέθοδος feature selection, κρίνεται εξαιρετικά ασύμφορη.

- ✔ Σταθμίζοντας τη σημασία των δύο παραπάνω κριτηρίων, καταλήγουμε στην επιλογή της μεθόδου RF με χρήση επιλογής χαρακτηριστικών mRMR, ως το βέλτιστο μοντέλο για την επίλυση του συγκεκριμένου προβλήματος δυαδικής ταξινόμησης.

6.5 Σύνοψη και τελικά συμπεράσματα

Στην παρούσα υποενότητα επιχειρείται να δοθεί μια συνοπτική παρουσίαση των κομβικών σημείων της παρούσας μεταπτυχιακής διατριβής καθώς και να εξαχθούν χρήσιμα συμπεράσματα αναφορικά με την μεθοδολογία που ακολουθήθηκε, την υπολογιστική διαδικασία και τα παραχθέντα αποτελέσματα. Αρχικά, κρίνεται σκόπιμο να γίνει μια σύντομη ανασκόπηση του περιεχομένου των προηγούμενων κεφαλαίων:

Στο Κεφάλαιο 1 έγινε εισαγωγή στην επιστήμη της Εξόρυξης Δεδομένων, παρατέθηκαν οι απαιτούμενες βασικές έννοιες και δόθηκαν παραδείγματα εφαρμογής των μεθόδων της σε πλήθος τομέων της ανθρώπινης δραστηριότητας. Παρουσιάστηκαν τα 5 βασικά στάδια που συνθέτουν τη διαδικασία εξόρυξης γνώσης, δίνοντας ιδιαίτερη έμφαση στο κομμάτι της μηχανικής μάθησης, όπου περιγράφηκε η κατηγοριοποίηση σε μεθόδους επιβλεπόμενης, μη-επιβλεπόμενης και ενισχυτικής μάθησης, ενώ δόθηκαν παραδείγματα για την κάθε κατηγορία.

Το Κεφάλαιο 2 εμβάθυνε στις μεθόδους επιβλεπόμενης μάθησης που χρησιμοποιήθηκαν στο Κεφάλαιο 5 για την επίλυση ενός προβλήματος δυαδικής ταξινόμησης. Συγκεκριμένα, έγινε περιγραφή του μαθηματικού υποβάθρου και περιγραφή της αλγοριθμικής διαδικασίας με χρήση ψευδοκώδικα του πολυεπίπεδου δικτύου Perceptron (MLP), του αλγορίθμου Τυχαίων Δασών (Random Forest), του αλγορίθμου k-πλησιέστερων γειτόνων (kNN) και των Μηχανών Διανυσμάτων Υποστήριξης (SVM), ενώ

έγινε αναφορά στα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου.

Το Κεφάλαιο 3 αποτέλεσε μια σύντομη εισαγωγή στην Επιστήμη της Πληροφορίας (Information Theory), ως προαπαιτούμενο για την κατανόηση της λειτουργίας των μεθόδων επιλογής χαρακτηριστικών η οποία ακολούθησε στο Κεφάλαιο 4. Στο πλαίσιο αυτό, παρουσιάστηκαν οι ορισμοί στα βασικά μέτρα θεωρίας της πληροφορίας, όπως η εντροπία, η αμοιβαία πληροφορία, η υπό συνθήκη πληροφορία κτλ.)

Στο Κεφάλαιο 4 έγινε αναφορά στις διαφορετικές κατηγορίες μεθόδων επιλογής χαρακτηριστικών (filter, wrapper, embedded) παρουσιάζοντας τα βασικά χαρακτηριστικά, πλεονεκτήματα και μειονεκτήματα της καθεμιάς. Ακολούθησε η αναλυτική περιγραφή της μεθόδου επιλογής χαρακτηριστικών mRMR, μιας από τις πιο πετυχημένες filter μεθόδους, η οποία βασίζεται στο μέτρο της αμοιβαίας πληροφορίας.

Το Κεφάλαιο 5 αποτέλεσε το κεντρικό πρακτικό κομμάτι της μεταπτυχιακής διατριβής, ενσωματώνοντας τη χρήση των μεθόδων εξόρυξης δεδομένων που περιγράφηκαν στο Κεφάλαιο 2 και της μεθόδου επιλογής χαρακτηριστικών που αναλύθηκε στο Κεφάλαιο 4 για την επίλυση ενός πραγματικού προβλήματος δυαδικής ταξινόμησης. Σε αυτό το πλαίσιο, έγινε εισαγωγή του αναγνώστη στο χρησιμοποιούμενο σύνολο δεδομένων, ενώ αναλύθηκε διεξοδικά η στρατηγική ανάλυσης τόσο σε επίπεδο μεθοδολογίας με την περιγραφή των σχετικών βημάτων, όσο και σε αλγοριθμικό επίπεδο με την τμηματική επεξήγηση του αντίστοιχου κώδικα σε γλώσσα προγραμματισμού R.

Το Κεφάλαιο 6 συγκέντρωσε τα αποτελέσματα που προέκυψαν από την μοντελοποιητική διαδικασία του προηγούμενου κεφαλαίου, ώστε να καταστεί δυνατή η συγκριτική αξιολόγηση των χρησιμοποιούμενων μεθόδων και τελικά να επιλεγεί εκείνη με την βέλτιστη απόδοση, λαμβάνοντας υπόψη κυρίως την ακρίβεια ταξινόμησης (accuracy) και δευτερευόντως το υπολογιστικό κόστος. Η ανάλυση των σχετικών αποτελεσμάτων αποτέλεσε πηγή εξαγωγής χρήσιμων παρατηρήσεων που αναφέρονται στις αντίστοιχες υποενότητες του παρόντος κεφαλαίου.

Ακολουθούν τα συνολικά συμπεράσματα που εξήχθησαν κατά τη διαδικασία εκπόνησης της μεταπτυχιακής διατριβής και την πρακτική εφαρμογή μεθόδων ανάλυσης δεδομένων για την επίλυση ενός ρεαλιστικού προβλήματος:

- ✔ Η Εξόρυξη Δεδομένων αποτελεί αναδυόμενο επιστημονικό πεδίο με τεράστιο εύρος

δυνατοτήτων, καθώς τα πλεονεκτήματα από τη χρήση της αντικατοπτρίζονται ήδη σε πληθώρα τεχνολογικών, κοινωνικών και οικονομικών εφαρμογών. Τα στάδια που απαιτούνται για την εξόρυξη γνώσης από τα δεδομένα αποτελούν σαφείς κατευθυντήριες γραμμές και πρέπει να τηρούνται απαρέγκλιτα, ώστε να καταστεί δυνατή η επίτευξη ουσιαστικών αποτελεσμάτων. Συγκεκριμένα για τα στάδια προ-επεξεργασίας και μετασχηματισμού διαπιστώνεται ότι απαιτούν ιδιαίτερη προσοχή, καθώς αφορούν τις ενέργειες προετοιμασίας του συνόλου δεδομένων πριν την έναρξη της μοντελοποίησης και επομένως επηρεάζουν σε μεγάλο βαθμό το τελικό αποτέλεσμα.

- ✔ Η ορθή επιλογή μεθόδων μηχανικής μάθησης καθώς και η επιτυχής τους εφαρμογή για την επίλυση ενός προβλήματος απαιτούν τόσο την καλή εξοικείωση με το μαθηματικό υπόβαθρο και τις αντίστοιχες παραμέτρους που τις χαρακτηρίζουν, όσο και τη βαθύτερη γνώση του κλάδου στον οποίο ανήκει το εκάστοτε πρόβλημα (domain knowledge). Αυτό κατέστη ιδιαίτερα σαφές κατά την επίλυση του προβλήματος δυαδικής ταξινόμησης στο Κεφάλαιο 5, από την ανάγκη παραμετροποίησης των αλγοριθμικών μεθόδων για αύξηση της ακρίβειας ταξινόμησης και από την οπτικοποίηση του χώρου χαρακτηριστικών με χρήση EDA για καλύτερη κατανόηση του συνόλου δεδομένων, αντίστοιχα.
- ✔ Η χρήση μεθόδων επιλογής χαρακτηριστικών πριν τη διαδικασία μοντελοποίησης αποδεικνύεται ιδιαίτερα χρήσιμη, καθώς μπορεί να οδηγήσει σε εξαιρετικά αποτελέσματα, εφόσον παραμετροποιηθεί σωστά και συνδυαστεί με κατάλληλες τεχνικές μηχανικής μάθησης. Στα πλαίσια της εργασίας, διαπιστώθηκε η αποτελεσματικότητα της information-driven μεθόδου επιλογής χαρακτηριστικών mRMR στην ουσιαστική μείωση της πολυπλοκότητας του τελικού μοντέλου, χωρίς καμία έκπτωση στην απόδοση (ακρίβεια ταξινόμησης).
- ✔ Ειδικότερα η συνδυαστική χρήση της μεθόδου επιλογής χαρακτηριστικών mRMR με την ensemble μέθοδο μηχανικής μάθησης Random Forest οδήγησε στην εξαγωγή ενός εύρωστου, γρήγορου στην εκπαίδευση και σχετικά χαμηλής πολυπλοκότητας μοντέλου, με ακρίβεια ταξινόμησης αντίστοιχη άλλων περισσότερο κοστοβόρων υπολογιστικά μοντέλων (SVM, MLP). Για τον λόγο αυτό, αναδείχθηκε ο “τελικός νικητής” του προβλήματος ταξινόμησης.

- ✔ Από την συνολική διερεύνηση και μεθοδολογία που ακολουθήθηκε στο Κεφάλαιο 5, εξάγεται το τελικό συμπέρασμα ότι η έξυπνη χρήση συνδυαστικών προσεγγίσεων (πχ. feature selection με ensemble machine learning) μπορεί να αποφέρει ίσα ή και καλύτερα αποτελέσματα από αυτά που προκύπτουν με την απευθείας χρήση state-of-the-art τεχνικών, όπως τα νευρωνικά δίκτυα. Όπως προαναφέρθηκε, απαιτείται επαρκής γνώση του προβλήματος καθώς και σύγκριση τεχνικών και κριτηρίων για την επιλογή αυτών που ταιριάζουν καλύτερα στις ιδιαιτερότητες του εκάστοτε προβλήματος, όπως ορίζει και το “No free lunch theorem” (Lattimore και Hutter, 2013).

6.6 Προτάσεις για μελλοντική έρευνα

Στην τελευταία υποενότητα της εργασίας προτείνονται κατευθύνσεις για μελλοντική έρευνα, επικεντρώνοντας στην αξιοποίηση των παραπάνω συμπερασμάτων και στη διεύρυνση της πρακτικής εφαρμογής των μεθόδων που χρησιμοποιήθηκαν.

Όσον αφορά το κομμάτι των μεθόδων μηχανικής μάθησης, προτείνεται τόσο η περαιτέρω παραμετροποίηση των υλοποιηθέντων μοντέλων, όσο και η κατασκευή νέων, προερχόμενων από άλλες μοντελοποιητικές οικογένειες. Συγκεκριμένα, ανατρέχοντας στην Υποενότητα 5.4 ο αναγνώστης θα διαπιστώσει ότι, ενώ τα παραχθέντα μοντέλα μηχανικής μάθησης υποβλήθηκαν σε διαδικασία παραμετροποίησης, δεν εξαντλήθηκε σε καμία περίπτωση το εύρος των δυνατών συνδυασμών που μπορεί να οδηγήσει σε βελτιωμένη ακρίβεια ταξινόμησης σε σχέση με την επιτευχθείσα. Εξάλλου, στόχος της παρούσας μεταπτυχιακής διατριβής δεν ήταν η βελτιστοποίηση αλγορίθμων μηχανικής μάθησης μέσω της εξονυχιστικής παραμετροποίησής τους, αλλά η πρακτική εφαρμογή τους σε συνδυασμό με μεθόδους επιλογής χαρακτηριστικών για την διαπίστωση της χρησιμότητας των δεύτερων. Ακόμη, προτείνεται η μελέτη της συμπεριφοράς άλλων αλγορίθμων, δίνοντας έμφαση στις tree-based ensemble μεθόδους όπως οι XGBoost, LightGBM, AdaBoost, ExtraTrees κ.α.

Αναφορικά με τις μεθόδους επιλογής χαρακτηριστικών, προτείνεται η ακολούθηση

παρόμοιας στρατηγικής, δηλαδή η περαιτέρω διερεύνηση των δυνατοτήτων της μεθόδου mRMR, καθώς και η εισαγωγή και σύγκρισή της με νέες μεθόδους επιλογής χαρακτηριστικών, βασισμένων στην Θεωρία Πληροφοριών, όπως η μέθοδος Joint Mutual Information (JMI) και οι παραλλαγές της (JMIM, NJMIM), η Double Input Symmetrical Relevance (DISR) και η Information Gain (IG). Ενδιαφέρουσα θα ήταν επίσης και η χρήση της δημοφιλούς filter μεθόδου Principal Components Analysis (PCA), των wrapper Recursive Feature Elimination (RFE) και Genetic Algorithms (GA) και της embedded Least Absolute Shrinkage and Selection Operator (LASSO).

Τέλος, προτείνεται η αναδιαμόρφωση της αλγοριθμικής δομής μέσω της προσθήκης αυτοματοποιήσεων και επιπλέον δυνατοτήτων που θα οδηγήσουν σε μια πλήρη μεθοδολογία σύγκρισης των διαθέσιμων μεθόδων. Για παράδειγμα, κρίνεται χρήσιμη η επέκταση της επαναληπτικής διαδικασίας προς-τα-εμπρός επιλογής χαρακτηριστικών (η οποία υλοποιήθηκε μόνο για τους δύο καλύτερους ταξινομητές), ώστε αυτή να γίνεται για όλα τα διαθέσιμα μοντέλα και η χρήση επιπλέον μέτρων αποτίμησης μοντέλου, πέραν της ακρίβειας ταξινόμησης, όπως η ανάκληση (recall), η ακρίβεια (precision) και η καμπύλη ROC (Receiver Operating Characteristic Curve).

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Aliferis, C. F., Statnikov, A., & Tsamardinos, I. (2006). Challenges in the Analysis of Mass-Throughput Data: A Technical Commentary from the Statistical Machine Learning Perspective. *Cancer Informatics, 2*, 133–162.

[2] Boser, B., Guyon, I., Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory (COLT '92). ACM*, New York, NY, USA, 144-152.

[3] Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press, Inc., New York, NY, USA.

[4] Blei, D., Ng, A., Jordan, M., Lafferty, J. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022.

[5] Breiman, L. (2001), Random Forests, *Mach. Learn.* 45, 1, pp.5-32.

[6] Chen, M., Han, J., Yu, P. (1997). Data mining: An overview from a database perspective. Knowledge and Data Engineering, *IEEE Transactions*, 8., pp.866 - 883.

[7] Cover, T., Hart, P. (1967) Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* 13, p.21–27.

[8] Duda, R., Hart, P., Stork D. (2000). *Pattern Classification* (2nd Edition). Wiley-Interscience, New York, NY, USA.

[9] Efron, B., Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, *The American Statistician*, vol. 37, Issue 1, pp. 36-48.

[10] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth P. and Uthurusamy, Ft. -a. (1996). *Advances in Knowledge Discovery and Data Mining*, (AKDDM), AAAI/MIT Press.

[11] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. -b. (1996). From Data Mining to

Knowledge Discovery in Databases, *AI Magazine*, 37-54.

[12] Fix, E., Hodges, J.-L. (1951) Discriminatory analysis, nonparametric discrimination: Consistency properties, Technical Report 4, U.S. Air Force, School of Aviation Medicine, Randolph Field, TX.

[13] Guiasu, S. (1977). *Information Theory with Applications*, McGraw-Hill Inc.

[14] Guyon, I., Elisseeff, A. (2003). An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 3, 1157-1182.

[15] Han, J., Kamber, M., Pei L. (2011). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA.

[16] Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining*, MIT Press, Cambridge, Massachusetts.

[17] Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* Oxford, England: U Michigan Press.

[18] John, G. H., Kohavi, R., Pfleger, K. (1994). Irrelevant feature and the subset selection problem, *ICML'94 Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, p.p. 121-129.

[19] Kakoyan, M. (2010). Επιλογή χαρακτηριστικών (feature selection) από βάσεις δεδομένων με φίλτρο αμοιβαίας πληροφορίας (mutual information filter), ΤΕΙ Σερρών, Διπλωματική Εργασία.

[20] Keim, D., Kriegel H.-P. (1996). Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Trans. on Knowl. and Data Eng.* 8, 6 pp.923-938.

[21] Kira, K., Rendell, L. (1992). A Practical Approach to Feature Selection, ML92, *Proceedings of the ninth international workshop on Machine learning* Pages, p.p. 249-256.

[22] Lattimore, T., Hutter, M., (2013). No free lunch versus Occam's razor in supervised learning, In *Algorithmic Probability and Friends*. Bayesian Prediction and Artificial Intelligence, pp.223-235. Springer, Berlin, Heidelberg.

[23] McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus of the ideas immanent in

nervous activity. Bulletin of Mathematical Biophysics, 5:115-133.

[24] Mitchell, T.-M. (1997), *Machine Learning*, McGraw-Hill, New York.

[25] Mukaka, M. (2012). A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal : The Journal of Medical Association of Malawi*, 24(3), pp.69–71.

[26] Papadopoulos, D. (2017), Design and development of a cognitive data analytics engine for network security, implementing Big Data technologies & machine learning techniques, Master Thesis, School of Production Engineering and Management, Technical University of Crete.

[27] Peng, H., Long, F., Ding, C. (2005). Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min Redundancy, *IEEE Trans.Pattern Analysis and Machine Intelligence*, vol. 27, p.p. 1226-1238.

[28] Piatetsky-Shapiro, G. (1991), Knowledge discovery in real databases: A report on the IJCAI-89 Workshop, *AI Mag*. 11, 5, pp.68-70.

[29] Quinlan, J.-R. 1986. Induction of Decision Trees. *Mach. Learn.* 1, 1, pp.81-106.

[30] Ramzan, M., Ahmad, M. (2014) Evolution of data mining: An overview, *Conference on IT in Business, Industry and Government (CSIBIG)*, Indore, pp.1-4.

[31] Rosenblatt, F. (1961). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books, Washington DC.

[32] Sasikala, B., Biju V., Prashanth C. (2017). Kappa and accuracy evaluations of machine learning classifiers, *2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, Bangalore,pp.20-23.

[33] Shannon, C. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal* 27, vol. 3, p.p. 379-423.

[34] Stone, M. (1974) Cross-validation and multinomial prediction, *Biometrika* 61, vol. 3, p.p. 509-515.

[35] Tan, P.-N., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education.

[36] Theodoridis, S., Koutroumbas, K., (2008). *Pattern Recognition*, Fourth Edition, Academic Press, Inc., Orlando, FL, USA.

[37] Turing, A. M. (1937), On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42: 230-265.

[38] Van Dijck, G., Van Hulle, M. M. (2006). Speeding up the wrapper feature subset selection in regression by mutual information relevance and redundancy analysis, Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) *ICANN 2006*, LNCS, Springer, vol. 4131, p.p. 31-40.

[39] Vapnik, V., Cortes, C. (1995). Support-Vector Networks. *Mach. Learn*, 20, 3, pp.273-297.

[40] Vlachogiannis, G., Hatziargyriou N. (2004) *Reinforcement Learning (RL) to Optimal Reconfiguration of Radial Distribution System (RDS). In: Vouros G.A., Panayiotopoulos T. (eds) Methods and Applications of Artificial Intelligence*. SETN 2004, vol 3025. Springer, Berlin, Heidelberg.

[41] Wang, L. (2005). *Support Vector Machines: Theory and Applications, Studies in fuzziness and soft computing*, Springer, vol.177.

[42] Yadav J., Sharma, M. (2013). A Review of K - mean Algorithm, *International Journal of Engineering Trends and Technology (IJETT)*. vol.4(7), pp.2972-2976.

[43] Yang, Y., Pedersen, J. (1997). A comparative study on feature selection in text categorization. ICML.

[44] Yao, Y., Xiao. Z., Wang, B., Viswanath, B., Zheng H., Zhao B. (2017). Complexity vs. performance: empirical analysis of machine learning as a service. In *Proceedings of the 2017 Internet Measurement Conference (IMC '17)*. ACM, New York, NY, USA, pp.384-397.

[45] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), pp.2473-2480.

[46] Yu, S., Guo, S., (2016). *Big data concepts, theories, and applications. Edited by Yu,*

Shui and Guo, Song, Springer International Publishing, Cham, Switzerland.

[47] Zaiane, O. (1999). *Principles of Knowledge Discovery in Databases*, Department of Computing Science, University of Alberta.

[48] Ζορκάδης, Β. (2002). Θεωρία Πληροφορίας και Κωδικοποίησης, Τόμος Α', Ελληνικό Ανοικτό Πανεπιστήμιο.

[49] Κουκουβίνος, Χ. (2003). *Θεωρία Πληροφοριών και Κωδίκων*. Πανεπιστημιακές Εκδόσεις ΕΜΠ.

[50] Κύρκος, Ε. (2015). *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

[51] Μητροπούλου, Α. (2016), Επιλογή χαρακτηριστικών για ταξινόμηση με τη βοήθεια μέτρων πληροφορίας, Διπλωματική Εργασία, ΣΕΜΦΕ, ΕΜΠ.

ΠΑΡΑΡΤΗΜΑ

Π.1 Αλγόριθμος πρόβλεψης χρηματοπιστωτικής φερεγγυότητας

```
---
title: "R Notebook"
output:
  word_document: default
  html_notebook: default
---


### A. Install and load necessary libraries, set working directory and read data.

```{r}

rm(list=ls(all=TRUE))
start_time <- Sys.time()
#install.packages("rstudioapi") # run this if it's your first time using it to
install
#install.packages("caret")      # run this if it's your first time using it to
install
#install.packages("plyr")       # run this if it's your first time using it to
install
#install.packages("reshape2")   # run this if it's your first time using it to
install
#install.packages("class")      # run this if it's your first time using it to
install
#install.packages("ggplot2")    # run this if it's your first time using it to
install

# Set the working directory to the relevant one:
setwd(dirname("C:/Users/Katerina/Desktop/Kate/diatrivi_master/R_Codes/"))

# Load necessary libraries
library(rstudioapi)
library(caret)
library(plyr)
library(reshape2)
library(class)
library(ggplot2)

data <- read.csv(file=
"C:/Users/Katerina/Desktop/Kate/diatrivi_master/R_Codes/CreditCard.csv",
header=TRUE, sep=",", dec=".")
# make sure to change the filepath in chunk H. as well.
```
```

B. Data Quality & Sanity checks

```{r}

# Reassuring correct type of variable acknowledgement
str(data)

# default is the dependent variable and should be a factor type
class(data$default)
data$default <- as.factor(data$default)

# limit_bal is an indepedent variable and should be a numerical type
class(data$LIMIT_BAL)
data$LIMIT_BAL <- as.numeric(data$LIMIT_BAL)

# sex (aka gender) is an independent variable and should be a factor type
class(data$SEX)
data$SEX <- as.factor(data$SEX)

# education is an independent variable and should be a factor type
class(data$EDUCATION)
data$EDUCATION <- as.factor(data$EDUCATION)

# marriage is an indepedent variable and should be a factor type
class(data$MARRIAGE)
data$MARRIAGE <- as.factor(data$MARRIAGE)

# age is an independent variable and should be a numeric type
class(data$AGE)
data$AGE <- as.numeric(data$AGE)

# pay_0 to pay_6 are independent variables and should be a numeric type
data[,7:12]<-lapply(data[,7:12],as.numeric)

# bill_amt1 to bill_amt6 are independent variables and should be numeric
data[,13:18]<-lapply(data[,13:18],as.numeric)

# pay_amt1 to pay_amt6 are independent variables and should be numeric
data[,19:24]<-lapply(data[,19:24],as.numeric)

##################################################################

# Ensure no duplicate records exist
data.u <- unique(data)

ifelse(length(nrow(data.u))==nrow(data),print("no duplicates
found"),print("duplicates found - keeping unique records"))

# Remove duplicates
data <- data.u
rm(data.u)
```

C. Feature selection based on mRMRe

```{r}

#install.packages("mRMRe")      # run this if it's your first time using it to
install
library(mRMRe)

# Necessary conversions for mRMR feature selection
data.num  <- sapply(data, as.numeric)
data.num  <- data.frame(data.num)
mrmr.data <- mRMR.data(data.num)
rm(data.num)
```

```
# Select number of important features
nf = 10

# mRMR with pearson coef.
classic1 <-mRMR.classic("mRMRe.Filter", data=mrmr.data, target_indices = 1,
feature_count = nf, method = "bootstrap", continuous_estimator = "pearson")

# mRMR with spearman coef.
classic2 <-mRMR.classic("mRMRe.Filter", data=mrmr.data, target_indices = 1,
feature_count = nf, method = "bootstrap", continuous_estimator = "spearman")
```

D. Exploratory Data Analysis

```{r}

# Data structure summary
str(data)

# default is a factor type ~ plot (Sxima 5.1)
table(data$default)
plot(data$default,main=colnames(data[1]),xlab = colnames(data[1]),ylab =
"Frequency")

# limit_bal is a numeric type ~ histogram (Sxima 5.2)
summary(data$LIMIT_BAL)
hist(data$LIMIT_BAL,breaks=100,main=colnames(data[2]),xlab = colnames(data[2]))

# sex is a factor type ~ plot (Sxima 5.3)
table(data$SEX)
plot(data$SEX,main=colnames(data[3]),xlab = colnames(data[3]),ylab = "Frequency")

# education is a factor type ~ plot (Sxima 5.4)
table(data$EDUCATION)
plot(data$EDUCATION,main=colnames(data[4]),xlab = colnames(data[4]),ylab =
"Frequency")

# marriage is a factor type ~ plot (Sxima 5.5)
table(data$MARRIAGE)
plot(data$MARRIAGE,main=colnames(data[5]),xlab = colnames(data[5]),ylab =
"Frequency")

# age is a numerical type ~ histogram (Sxima 5.6)
summary(data$AGE)
hist(data$AGE,breaks=5,main=colnames(data[6]),xlab = colnames(data[6]))

# pay_0 to pay_6 are numerical types ~ histograms (Sximata 5.7 - 5.12)
lapply( data[,7:12],summary)
for (i in 7:12) {
  hist(data[,i],breaks=10,main=colnames(data)[i],xlab = colnames(data[i]))
}

# bill_amt1 to bill_amt6 are numerical types ~ histograms (Sximata 5.13 - 5.18)
lapply(data[,13:18],summary)
for (i in 13:18) {
  hist(data[,i],breaks=100,main=colnames(data)[i],xlab = colnames(data[i]))
}

# pay_amt1 to pay_amt6 are numerical types ~ histograms (Sxima 5.19 - 5.24)
lapply(data[,19:24],summary)
for (i in 19:24) {
  hist(data[,i],breaks=100,main=colnames(data)[i],xlab = colnames(data[i]))
}
```

E. Split data to CV-evaluation subsets, set resampling strategy

```{r}

library(caret)

portion       <- createDataPartition(data$default, p = 0.8, list=FALSE)
data.cv       <- data[portion,]  # 80% of the initial dataset on which to perform
model selection
data.evaluate <- data[-portion,] # 20% of the initial dataset on which to
evaluate model accuracy

# Keep safe initial dataset
data.initial <- data
# Rename data.cv portion into data and remove data.cv
data <- data.cv
rm(data.cv)

# Repeated cross-validation strategy:
# Split available data into 5 equal parts, train in 4 and evaluate in the 5th

set.seed(123)
fitControl <- trainControl(method="repeatedcv",
                           number=10,
                           repeats=1,
                           classProbs  = FALSE,
                           verboseIter = FALSE,
                           savePredictions = TRUE,
                           allowParallel = TRUE)

rm(portion)
```


F. Preprocessing features and creation of different dataset versions.

```{r}

set.seed(123)

# Rename necessary 0 and 1 to X.0 and X.1 respectively
data$default <- revalue(data$default , c("1"="X.1", "0"="X.0"))

# mRMR pearson subset
data.p <- data[,c(1,unname(unlist(classic1@filters)))]

# mRMR spearman subset
data.s <- data[,c(1,unname(unlist(classic2@filters)))]
```


G. Modeling (using svm, knn, nnet, rf)


```{r}

# Support vector machines (all features , mRMR with Pearson coef. , mRMR with
Spearman coef.)

set.seed(123)

svm <- train(default ~ .,
             data = data,
             method = "svmRadial",
             metric = "Accuracy",
             trControl = fitControl)
```

```
svm.p <- train(default ~ .,
               data = data.p,
               method = "svmRadial",
               metric = "Accuracy",
               trControl = fitControl)

svm.s <- train(default ~ .,
               data = data.s,
               method = "svmRadial",
               metric = "Accuracy",
               trControl = fitControl)

# k-Nearest Neighbours (all features , mRMR with Pearson coef. , mRMR with
Spearman coef.)

knn <-train(default ~ .,
               data = data,
               method = "knn",
               metric = "Accuracy",
               trControl = fitControl)

knn.p <-train(default ~ .,
               data = data.p,
               method = "knn",
               metric = "Accuracy",
               trControl = fitControl)

knn.s <-train(default ~ .,
               data = data.s,
               method = "knn",
               metric = "Accuracy",
               trControl = fitControl)

# Neural Networks (all features , mRMR with Pearson coef. , mRMR with Spearman
coef.)

set.seed(123)

nnGrid=expand.grid(size=c(2),decay=c(0.02))

nn <- train(default ~ .,
               data = data,
               method = 'nnet',
               tuneGrid = nnGrid,
               metric = "Accuracy",
               trControl = fitControl)

nn.p <- train(default ~ .,
               data = data.p,
               method = 'nnet',
               tuneGrid = nnGrid,
               metric = "Accuracy",
               trControl = fitControl)

nn.s <- train(default ~ .,
               data = data.s,
               method = 'nnet',
               tuneGrid = nnGrid,
               metric = "Accuracy",
               trControl = fitControl)

# Random Forests (all features , mRMR with Pearson coef. , mRMR with Spearman
coef.)

set.seed(123)

tunegrid <- expand.grid(.mtry=c(4,5,6))
```

```
rf <- train(default ~.,
                  data   = data,
                  method = "rf",
                  metric = "Accuracy",
                  tuneGrid = tunegrid,
                  trControl=fitControl,
                  ntree=100)

rf.p <- train(default ~.,
                  data   = data.p,
                  method = "rf",
                  metric = "Accuracy",
                  tuneGrid = tunegrid,
                  trControl=fitControl,
                  ntree=100)

rf.s <- train(default ~.,
                  data   = data.s,
                  method = "rf",
                  metric = "Accuracy",
                  tuneGrid = tunegrid,
                  trControl=fitControl,
                  ntree=100)


# Cumulative cross-validation accuracy results of classifiers

results <- resamples(list(
                          SVM        =svm,
                          SVM.p      =svm.p,
                          SVM.s      =svm.s,
                          KNN        =knn,
                          KNN.p      =knn.p,
                          KNN.s      =knn.s,
                          NN         = nn,
                          NN.p       = nn.p,
                          NN.s       = nn.s,
                          RF         = rf,
                          RF.p       = rf.p,
                          RF.s       = rf.s
                          ))

# Saving trained models
saveRDS(svm, file = 'svm' )
saveRDS(svm.p, file = 'svm.p' )
saveRDS(svm.s, file = 'svm.s' )
saveRDS(knn, file = 'knn' )
saveRDS(knn.p, file = 'knn.p' )
saveRDS(knn.s, file = 'knn.s' )
saveRDS(nn, file = 'nn' )
saveRDS(nn.p, file = 'nn.p' )
saveRDS(nn.s, file = 'nn.s' )
saveRDS(rf, file = 'rf' )
saveRDS(rf.p, file = 'rf.p' )
saveRDS(rf.s, file = 'rf.s' )

# Print cross-validation accuracy results of classifiers
summary(results)
dotplot(results)
bwplot(results)
```

H. Comparative modeling with forward feature selection (mRMR Pearson) between two best classifiers, with respect to CV accuracy

```{r}
```

```
# Fetching the initial preprocessed dataset
data <- data.initial

set.seed(123)

# Set tunegrid for rf
tunegrid <- expand.grid(.mtry=c(4,5,6))

# Initialise accuracy results vectors
TotalEvalAccRF  = c()
TotalEvalAccSVM = c()

portionF       <- createDataPartition(data$default, p = 0.8, list=FALSE)
data.cvF       <- data[portionF,]  # 80% of the initial dataset on which to
perform model selection
data.evaluateF <- data[-portionF,] # 20% of the initial dataset on which to
evaluate model accuracy

# Keep safe initial dataset
data.initial <- data

# Rename data.cv portion into data and remove data.cvF
data <- data.cvF #
rm(data.cvF)

# Repeated cross-validation strategy:
# Split available data into 5 equal parts, train in 4 and evaluate in the 5th

set.seed(123)
fitControl <- trainControl(method="repeatedcv",
                           number=10,
                           repeats=1,
                           classProbs  = FALSE,
                           verboseIter = FALSE,
                           savePredictions = TRUE,
                           allowParallel = TRUE)

rm(portionF)

# Rename necessary 0 and 1 to X.0 and X.1 respectively
data$default <- revalue(data$default , c("1"="X.1", "0"="X.0"))

# Forward selection (mRMR Pearson) from 1 to 23 features
for (i in 1:(ncol(data)-1)) {

    classic1F <-mRMR.classic("mRMRe.Filter", data=mrmr.data, target_indices = 1,
feature_count = i, method = "bootstrap", continuous_estimator = "pearson")

    set.seed(123)

    # mRMR pearson subset
    data.pF <- data[,c(1,unname(unlist(classic1F@filters)))]

    # Support vector machine

    svm.pF <- train(default ~ .,
                    data = data.pF,
                    method = "svmRadial",
                    metric = "Accuracy",
                    trControl = fitControl)

    # Random Forest

    rf.pF <- train(default ~.,
                     data   = data.pF,
                     method = "rf",
                     metric = "Accuracy",
```

```
                    tuneGrid = tunegrid,
                    trControl=fitControl,
                    ntree=100)


    # Saving trained models, needed to compute cv accuracy later
     saveRDS(svm.pF, file = paste('svmp', i, sep="") )
     saveRDS(rf.pF, file = paste('rfp', i, sep="") )

    # Transform evaluation subset to include only mRMR selected features
    data.p.evaluateF  <- data.evaluateF[,c(1,unname(unlist(classic1F@filters)))]

    # Compute Random Forest evaluation accuracy for mRMR selected features
    preds <- predict(rf.pF, newdata =
data.frame( data.p.evaluateF[colnames(data.p.evaluateF)[2:(i+1)]]))
    TotalEvalAccRF[i]<-(table(preds,data.p.evaluateF$default)[1,1]+
table(preds,data.p.evaluateF$default)[2,2])/(nrow(data.p.evaluateF))

    # Compute SVM evaluation accuracy for mRMR selected features
    preds <- predict(svm.pF, newdata =
data.frame( data.p.evaluateF[colnames(data.p.evaluateF)[2:(i+1)]]))
    TotalEvalAccSVM[i]<-(table(preds,data.p.evaluateF$default)[1,1]+
table(preds,data.p.evaluateF$default)[2,2])/(nrow(data.p.evaluateF))

}

# Print Random Forest and SVM CV accuracy for all mRMR selected subsets
modelnamesSVMp <- paste0("svmp", 1:i)
modelsSVMp <- lapply(modelnamesSVMp, function(x) readRDS(x))
resultsSVMp <- resamples( modelsSVMp ,  modelNames = modelnamesSVMp )
summary(resultsSVMp)

modelnamesRFp <- paste0("rfp", 1:i)
modelsRFp <- lapply(modelnamesRFp, function(x) readRDS(x))
resultsRFp <- resamples( modelsRFp ,  modelNames = modelnamesRFp )
summary(resultsRFp)

# Respective plots for CV accuracy
bwplot(resultsSVMp, main='CV Accuracy SVMp')
bwplotSVMp <- bwplot(resultsSVMp, main='CV Accuracy SVMp')
dotplotSVMp <-dotplot(resultsSVMp, main='CV Accuracy SVMp')
#bwplot(resultsSVMp, main='CV Accuracy SVMp', scales = list(relation = "free"),
#xlim = list(c(0.8, 0.85), c(0.25, 0.4)))

bwplot(resultsRFp, main='CV Accuracy RFp')
bwplotRFp <- bwplot(resultsRFp, main='CV Accuracy RFp')
dotplotRFp <- dotplot(resultsRFp, main='CV Accuracy RFp')
#bwplot(resultsRFp, main='CV Accuracy SVMp', scales = list(relation = "free"),
#xlim = list(c(0.8, 0.85), c(0.25, 0.4)))

# Print Random Forest and SVM evaluation accuracy for all mRMR selected subsets
cat('Total RF Evaluation Accuracy:',TotalEvalAccRF)
cat('Total SVM Evaluation Accuracy:',TotalEvalAccSVM)

# Respective plots for evaluation accuracy
plot(TotalEvalAccRF , main='Evaluation Accuracy RF' , xlab = 'Number of
features', ylab = 'Accuracy', type ='o')
axis(1, at = seq(1, (ncol(data)-1), by = 1))
grid( NA , 15 , lwd = 2 )

plot(TotalEvalAccSVM , main='Evaluation Accuracy SVM' , xlab = 'Number of
features', ylab = 'Accuracy', type ='o')
axis(1, at = seq(1, (ncol(data)-1), by = 1))
grid( NA , 15 , lwd = 2 )
```

I. Evaluation

```{r}

# Transform evaluation subset to include only mRMR Pearson selected features
data.p.evaluate  <- data.evaluate[,c(1,unname(unlist(classic1@filters)))]

# Transform evaluation subset to include only mRMR Spearman selected features
data.s.evaluate  <- data.evaluate[,c(1,unname(unlist(classic2@filters)))]

# SVM evaluation results
preds <- predict(svm, data.evaluate[,2:24])
table(preds,data.evaluate$default)
SVM_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))

# SVM.p evaluation results
preds <- predict(svm.p, data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
SVM.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# SVM.s evaluation results
preds <- predict(svm.s, data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
SVM.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# KNN evaluation results
preds <- predict(knn, data.evaluate[,2:24])
table(preds,data.evaluate$default)
KNN_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))

# KNN.p evaluation results
preds <- predict(knn.p, data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
KNN.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# KNN.s evaluation results
preds <- predict(knn.s, data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
KNN.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# NN evaluation results
preds <- predict(nn, data.evaluate[,2:24])
table(preds,data.evaluate$default)
NN_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))

# NN.p evaluation results
preds <- predict(nn.p, data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
NN.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# NN.s evaluation results
preds <- predict(nn.s, data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
NN.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# RF evaluation results
preds <- predict(rf, newdata = data.evaluate[,2:24])
table(preds,data.evaluate$default)
RF_Accuracy = (table(preds,data.evaluate$default)[1,1]+
table(preds,data.evaluate$default)[2,2])/(nrow(data.evaluate))
```

```
# RF.p evaluation results
preds <- predict(rf.p, newdata = data.p.evaluate[,2:11])
table(preds,data.p.evaluate$default)
RF.p_Accuracy = (table(preds,data.p.evaluate$default)[1,1]+
table(preds,data.p.evaluate$default)[2,2])/(nrow(data.p.evaluate))

# RF.s evaluation results
preds <- predict(rf.s, newdata = data.s.evaluate[,2:11])
table(preds,data.s.evaluate$default)
RF.s_Accuracy = (table(preds,data.s.evaluate$default)[1,1]+
table(preds,data.s.evaluate$default)[2,2])/(nrow(data.s.evaluate))

# Cumulative Evaluation results
Model    <- c( 'SVM' , 'SVM.p' , 'SVM.s' , 'KNN' , 'KNN.p' , 'KNN.s' , 'NN' ,
'NN.p' , 'NN.s' , 'RF' , 'RF.p' , 'RF.s' )
Accuracy <- c( SVM_Accuracy, SVM.p_Accuracy, SVM.s_Accuracy, KNN_Accuracy,
KNN.p_Accuracy, KNN.s_Accuracy, NN_Accuracy, NN.p_Accuracy, NN.s_Accuracy,
RF_Accuracy, RF.p_Accuracy, RF.s_Accuracy )

Evaluation_Results <- data.frame(Model,Accuracy)
Evaluation_Results

# Plot for cumulative Evaluation results
barplot(Evaluation_Results$Accuracy, names.arg = Evaluation_Results$Model,
main='Cumulative Evaluation Accuracy' , xlab = 'Classifier' , ylab = 'Accuracy',
ylim=range(0.72,0.85), xpd = FALSE )
grid(nx=NA, ny=NULL)

end_time <- Sys.time()

end_time - start_time
```

Π.2 Δείγμα υπό μελέτη συνόλου δεδομένων

default	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0
1	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000
0	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000
0	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000
0	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679
0	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800
0	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770
0	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542
0	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000
0	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0
0	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66
0	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640
0	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0
1	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0
0	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000
0	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100
1	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0
0	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10385	10000	75940	20000	195599	50000
0	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0