Πανεπιστήμιο Πειραιώς

Τμήμα Ψηφιακών Συστημάτων

Πτυχιακή Εργασία

«Τεχνικές Εξόρυξης Γνώσης και Μηχανικής Μάθησης σε Οικονομικά Δεδομένα και Εφαρμογές σε Συναλλαγές Πιστωτικών Καρτών»

Φοιτητής

Κολιόπουλος Μιχαήλ

ΜΕ1714

Επιβλέπων Καθηγητής

Φιλιππάκης Μιχαήλ

Αθήνα 24/1/2020

ΠΕΡΙΕΧΟΜΕΝΑ

1. Εισαγωγή

1.1 Το Πρόβλημα της Ανίχνευσης Απάτης

Η έννοια της απάτης είναι το ίδιο παλιά όσο και η ανθρωπότητα και μπορεί να λάβει απεριόριστη ποικιλία μορφών. Συμπληρωματικά, η ανακάλυψη νέων τεχνολογιών παρέχει καινούριους τρόπους στους εγκληματίες για ναπραγματοποιήσουν απατηλές ενέργεις. Ενώ οι οικονομικές απώλειες επηρεάζουν τόσο τους επιχειρηματίες και τις τράπεζες, όσο και τους επιμέρους καταναλωτές. Για παράδειγμα, η Ευρωπαική Κεντρική Τράπεζα στην ετήσια αναφορά της το 2012 (European Central Bank. Reportoncardfraudavailable:https://www.ecb.europa.eu/paym/intro/mip online/2018/html/1809_fifth_report_on_card_fraud.en.html.) υπογραμμίζει οτι 1 Ευρό για κάθε 2,635 Ευρό που δαπανόνται σε συναλλαγές μέσω πιστωτικών καρτών, χάνεται λόγω απάτης. Ενώ η συνολική αξία της απάτης με πιστωτικές κάρτες ανέρχεται στα 1,33 δισεκατομμύρια Ευρό για το 2018.

Η ανίχνευση απάτης είναι, δοθέντος ενός συνόλου συναλλαγών, η διαδικασία αναγνώρισης του εάν μια επίσημη συναλλαγή ανήκει στην κατηγορία των απατηλών ή των κανονικών (αυθεντικών) συναλλαγών. Συμπληρωματικά, ένα σύστημα ανίχνευσης απάτης δεν θα πρέπει μόνο να εντοπίζει επιτυχώς τις απατηλές συναλλαγές, αλλά θα πρέπει να είναι αποδοτικό όσον αφορά το κόστος, με την έννοια ότι το κόστος που επενδύεται στην προσπάθεια ανίχνευσης της απάτης θα πρέπει να είναι μικρότερο απο τις οικονομικές απώλειες λόγω της απάτης.

Τυπικά οι συναλλαγές πρώτα φιλτράρονται με τον έλεγχο κάποιων βασικών συνθηκών και στη συνέχεια περνάνε στο προβλεπτικό μοντέλο το οποίο αξιολογεί τη συναλλαγή βάση χαμηλού ή υψηλού ρίσκου απάτης και παράγει συναγερμούς για τις περιπτώσεις υψηλού ρίσκου. Οι ερευνητές ελέγχουν τους συναγερμούς αυτούς και παρέχουν Feedback για κάθε συναγερμό, δηλαδή αν πρόκειται για True Positive (Απάτη) ή False Positive (κανονική). Τα Feedbackπου δίνουν οι ερευνητές μπορούν να χρησιμοποιηθούν για να βελτιώσουν το μοντέλο. Ένα προβλεπτικό μοντέλο μπορεί να φτιαχτεί στη βάση εξειδικευμένων κανόνων μαζί με τη γνώση απο τους ειδικούς του χώρου, αλλά αυτό θα απιτεί χειροκίνητη ρύθμιση και ανθρώπινη επίβλεψη.

Εναλλακτικά, μέσω της Μηχανικής Μάθησης μπορούμε αποδοτικά να ανακαλύψουμε απατηλά μοτίβα και να προβλέψουμε απατηλές συναλλαγές. Τέτοιες τεχνικές Μηχανικής Μάθησης στηρίζονται στη δημιουργία ενός προβλεπτικού μοντέλου στη βάση ενός συνόλου παραδειγμάτων. Στον χώρο της ανίχνευσης απάτης, η χρήση Μηχανικής Μάθησης είναι ιδιαίτερα ελκυστική για πολλούς λόγους: Πρώτον, επιτρέπει την ανακάλυψη μοτίβων σε σύνολα δεδομένων υψηλών διαστάσεων. Δεύτερον, επειδή οι απατηλές συναλλαγές συνήθως φανερώνουν υψηλέ συσχέτιση μεταξύ του χώρου και του χρόνου που αυτές έλαβαν μέρος. Τρίτον, τεχνικές Μηχανικής Μάθησης μπορούν να χρησιμοποιηθούν για την αναγνώριση και μοντελοποίηση ήδη υπάρχουσων στρατηγικών απάτης, καθώς και να αναγνωρίσουν νέες στρατηγικές που σχετίζονται με ασυνήθιστη συμπεριφορά απο τους κατόχους καρτών.

1.2 Ανίχνευση Απάτης Πιστωτικών Καρτών

Η ανίχνευση απάτης πιστωτικών καρτών στηρίζεται στην ανάλυση καταγεγραμμένων συναλλαγών. Τα αυτοματοποιημένα συστήματα είναι αναγκαία καθώς δεν είναι πάντα δυνατό ή εύκολο για έναν ανθρώπινο αναλυτή να αναγνωρίσει μοτίβα απάτης στα δεδομένα των συναλλαγών τα οποία χαρακτηρίζονται απο έναν μεγάλο αριθμό δειγμάτων, πολλές διαστάσεις και συνεχόμενες ανανεώσεις. Με τη Μηχανική Μάθηση αφήνουμε τους Ηλεκτρονικούς Υπολογιστές να ανακαλύψουν απατηλά μοτίβα. Βέβαια, η προσέγγιση αυτή χαρακτηρίζεται απο πλεονεκτήματα και μειονεκτήματα. Για παράδειγμα, οι αλγόριθμοι Μηχανικής Μάθησης μπορούν i) να εκπαιδευτούν σε περίπλοκες συνθέσεις απάτης, ii) να καταναλώσουν μεγάλες ποσότητες δεδομένων, iii) να μοντελοποιήσουν περίπλοκες κατανομές, iv) να προβλέψουν νέους τύπους απάτης και v) να προσαρμοστούν σε μια αλλαγή κατανομών λόγω της εξέλιξης της απάτης. Απο την άλλη μεριά, έχουν κάποια μειονεκτήματα όπως: i) απιατούν μεγάλα σε όγκο και ποιοτικά σύνολα δειγμάτων και ii) κάποια μοντέλα αποτελούν μαύρο κουτί, δηλαδή δεν είναι εύκολα ή και καθόλου ερμηνεύσιμα απο τον άνθρωπο.

Η κατασκευή ενός τέτοια συστήματος ανίχνευσης απάτης βασισμένο στη Μηχανική Μάθηση είναι ιδιαίτερα προκλητική για τους ακόλουθους λόγους:

1. Οι απάτες αποτελούν ένα πολύ μικρό κομμάτι των καθημερινών συναλλαγών.
2. Οι κατανομές των απατηλών συναλλαγών σεξελίσσονται στο βάθος του χρόνου λόγω της εποχικότητας και των νέων στρατηγικών απάτης.
3. Η αληθινή φύση της πλειοψηφίας των συναλλαγών γίνεται τυπικά γνωστή αρκετές ημέρες αφότου η συναλλαγή έλαβε μέρος.

Η πρώτη πρόκληση είναι γνωστή ώς πρόβλημα ανισσορροπίας, αφού η κατανομή των συναλλαγών γέρνει ισχυρά προς την αυθεντική κλάση. Συμπληρωματικά, οι κατανομές των αυθεντικών και απατηλών δειγμάτων δεν βρίσκονται μόνο σε ανισσορροπία αλλά και επικαλύπτονται (overlapping distributions). Οι περισσότεροι αλγόριθμοι Μηχανικής Μάθησης δεν είναι σχεδιασμένοι να διαχειρίζονται ταυτόχρονα κατανομές που είναι και ανισσόρροπες αλλά και επικαλυπτόμενες.

Η αλλαγή στις απατηλές ενέργειες και η συμπεριφορά των καταναλωτών είναι οι βασικοί λόγοι για τη μή στασιμότητα των συναλλαγών. Τα σύνολα δεδομένων θα πρέπει να ανανεώνονται συνέχεια ενώ, η Τρίτη πρόκληση αφορά το γεγονός ότι είναι αδύνατο να ελεγχθούν όλες οι συναλλαγές. Το κόστος της ανθρώπινης εργασίας περιορίχει αισθητά τον αριθμό των συναγερμών που μορούν να ελεγχθούν απο τους ερευνητές.

2. Εισαγωγή στη Μηχανική Μάθηση & στις Τεχνικές Εξόρυξης Γνώσης

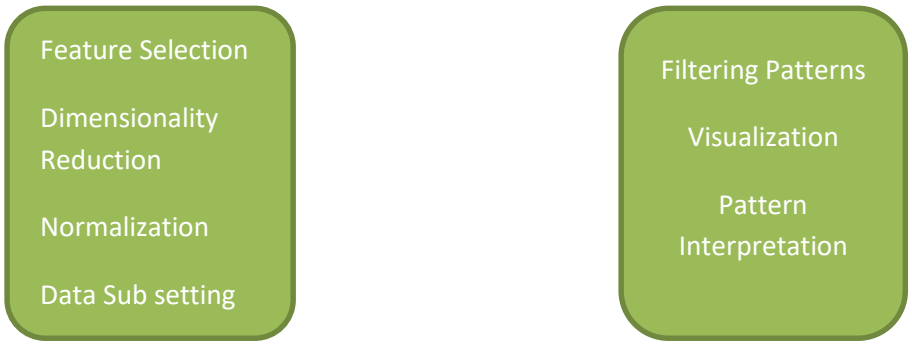
Η σύγκλιση της επιστήμης της πληροφορικής και των τηλεπικοινωνιών δημιούργησε μια κοινωνία που τρέφεται από την πληροφορία. Παρόλα αυτά, οι περισσότερες πληροφορίες βρίσκονται ακόμα στην ωμή μορφή τους: τα δεδομένα. Εάν τα δεδομένα χαρακτηρίζονται ως καταγεγραμμένα γεγονότα, τότε η πληροφορία αποτελεί το σύνολο προτύπων ή προσδοκιών που βρίσκονται μέσα στα δεδομένα. Υπάρχει

ένας τεράστιος όγκος πληροφορίας κλειδωμένος εντός των βάσεων δεδομένων, πληροφορία η οποία ενδεχομένως να είναι σημαντική αλλά παραμένει ανεξερεύνητη. Το Data Mining είναι η εξαγωγή της υπονοούμενης, προηγουμένως άγνωστης πληροφορίας από τα δεδομένα, ενώ το Machine Learningπαρέχει την τεχνική βάση για αυτό.

Η ταχεία ανάπτυξη της συλλογής δεδομένων και αποθήκευσης έχουν δώσει τη δυνατότητα σε οργανισμούς να παράγουν τεράστιες ποσότητες δεδομένων όμως, η εξαγωγή χρήσιμης πληροφορίας από τα δεδομένα αυτά παραμένει εξαιρετικά προκλητική. Συχνά, παραδοσιακά μέσα ανάλυσης δεδομένων και τεχνικές δεν μπορούν να χρησιμοποιηθούν λόγω του μεγάλου όγκου των δεδομένων. Άλλες φορές, η μη παραδοσιακή φύση των δεδομένων σημαίνει ότι παραδοσιακές προσεγγίσεις δεν μπορούν να χρησιμοποιηθούν ακόμα και αν το σύνολο των δεδομένων είναι μικρό. Σε άλλες περιπτώσεις, οι ερωτήσεις που πρέπει να απαντηθούν, δεν είναι εφικτό κάνοντας χρήση υπαρχόντων τεχνικών ανάλυσης και έτσι, καινούριες μέθοδοι πρέπει να δημιουργηθούν.

Το Data Mining είναι μια τεχνολογία που συνδυάζει τις παραδοσιακές μεθόδους ανάλυσης δεδομένων με εκλεπτυσμένους αλγόριθμους, για την επεξεργασία μεγάλου όγκου δεδομένων. Έχει δημιουργήσει, επίσης, συναρπαστικές ευκαιρίες για την εξερεύνηση και την ανάλυση νέων τύπων δεδομένων και την ανάλυση παλιότερων τύπων με νέους τρόπους. Για παράδειγμα, τεχνικές Data Mining μπορούν να χρησιμοποιηθούν για την υποστήριξη μεγάλου εύρους εφαρμογών επιχειρηματικής ευφυίας όπως η ανεύρεση των πελατειακών προφίλ, το στοχευμένο marketing και η ανίχνευση απάτης. Μπορεί επίσης να βοηθήσει τους εμπόρους να δώσουν σημαντικές απαντήσεις σε σημαντικές ερωτήσεις όπως « Ποιοι είναι οι πιο επικερδείς πελάτες?» και «Ποιες είναι οι προοπτικές εσόδων για την εταιρία την επόμενη οικονομική χρονιά».

Το Data Mining λοιπόν, είναι η διαδικασία της αυτοματοποιημένης ανακάλυψης χρήσιμης πληροφορίας σε μεγάλα αποθετήρια δεδομένων. Τεχνικές Data Mining αναπτύσσονται για να διατρέξουν μεγάλες βάσεις δεδομένων έτσι ώστε να βρουν νέα και χρήσιμα πρότυπα πληροφορίας που σε άλλη περίπτωση θα παρέμεναν άγνωστα. Παρέχουν επίσης δυνατότητες πρόβλεψης του αποτελέσματος μιας μελλοντικής παρατήρησης. Το Data Mining αποτελεί αναπόσπαστο κομμάτι της ανακάλυψης γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases–KDD) που αποτελεί τη συνολική διαδικασία της μετατροπής ωμών δεδομένων σε χρήσιμες πληροφορίες.



Τα εισαγόμενα δεδομένα μπορούν να αποθηκευτούννα αποθηκευτούν σε μια πληθώρα format και μπορούν να 'κατοικήσουν' σε ένα κεντρικό αποθετήριο δεδομένων ή να διανέμονται σε πολλαπλές τοποθεσίες. Σκοπός της προ επεξεργασίας είναι να μεταμορφώσει τα ωμά δεδομένα σε κατάλληλο formatγια ανάλυση που ακολουθεί. Τα βήματα που εμπεριέχονται στη διαδικασία της προ επεξεργασίας των δεδομένων περιλαμβάνουν τη συνένωση δεδομένων από διαφορετικές πηγές, τον καθαρισμό τους και την αφαίρεση του θορύβου και των διπλά εγγεγραμμένων παρατηρήσεων, καθώς και την επιλογή αρχείων και χαρακτηριστικών που είναι σχετικά με το συγκεκριμένο Data Mining έργο. Εξαιτίας των πολλών τρόπων συλλογής και αποθήκευσης των δεδομένων, η διαδικασία της προ επεξεργασίας αποτελεί το πιο κοπιώδες και χρονοβόρο βήμα στη συνολική διαδικασία ανακάλυψης γνώσης.

Από την άλλη πλευρά, η ενσωμάτωση των αποτελεσμάτων του Data Mining στα συστήματα υποστήριξης αποφάσεων χρειάζεται και μια μεταγενέστερη επεξεργασία (Post Processing) έτσι ώστε να εξασφαλιστεί ότι μόνο έγκυρα και χρήσιμα αποτελέσματα ενσωματώνονται σε ένα σύστημα υποστήριξης αποφάσεων. Ένα παράδειγμα μεταγενέστερης επεξεργασίας είναι η οπτικοποίηση (Visualization) που επιτρέπει στους αναλυτές να εξερευνήσουν τα δεδομένα και τα αποτελέσματα του Data Mining από μια ποικιλία οπτικών. Στατιστικά μέτρα και μέθοδοι ελέγχου υποθέσεων μπορούν επίσης να εφαρμοστούν κατά τη μεταγενέστερη επεξεργασία έτσι ώστε να ελαχιστοποιηθούν τα νόθα/ πλαστά αποτελέσματα που μπορεί να προκύψουν από το Data Mining

2.1 Οι προκλήσεις που έδωσαν ώθηση στην ανάπτυξη του Data Mining

Οι παραδοσιακές τεχνικές ανάλυσης, συνήθως συναντούσαν πρακτικές δυσκολίες στην αντιμετώπιση των προκλήσεων που παρουσίαζαν τα νέα σύνολα δεδομένων. Παρακάτω ακολουθούν κάποιες από τις πιο συνηθισμένες προκλήσεις.

2.1.1 Επεκτασιμότητα (Scalability)

Εξαιτίας της προόδου στη δημιουργία και συλλογή δεδομένων, τα σύνολα δεδομένων με μέγεθος της τάξης των Gigabytes, terra bytes ή και Petabytes γίνονται όλο και πιο συνηθισμένα. Εάν οι αλγόριθμοι εξόρυξης γνώσης είναι σε θέση να χειριστούν τα δεδομένα αυτά τότε πρέπει να είναι επεκτάσιμοι. Πολλοί τέτοιοι αλγόριθμοι κάνουν χρήση ειδικών στρατηγικών έρευνας προκειμένου να διαχειριστούν

τέτοια εκθετικά προβλήματα. Η επεκτασιμότητα μπορεί επίσης να απαιτεί την ενσωμάτωση νέων δομών δεδομένων προκειμένου να υπάρχει πρόσβαση σε ξεχωριστά αρχεία με έναν πιο αποδοτικό τρόπο.

2.1.2 Υψηλές Διαστάσεις (High Dimensionality)

Είναι, πλέον, συχνό φαινόμενο να συναντάς σύνολα δεδομένων υψηλών διαστάσεων με εκατοντάδες ή χιλιάδες χαρακτηριστικά (attributes) σε αντίθεση με τα σύνολα δεδομένων του παρελθόντος. Οι παραδοσιακές τεχνικές ανάλυσης οι οποίες είχαν αναπτυχθεί για σύνολα μικρών διαστάσεων συνήθως δεν δουλεύουν με ικανοποιητικά αποτελέσματα. Συμπληρωματικά, για κάποιους αλγόριθμους ανάλυσης, η υπολογιστική πολυπλοκότητα αυξάνεται κατακόρυφα με την αύξηση των διαστάσεων.

2.1.3 Ετερογενή και πολύπλοκα δεδομένα

Οι παραδοσιακές τεχνικές ανάλυσης δεδομένων συχνά αντιμετώπιζαν σύνολα δεδομένων που περιείχαν χαρακτηριστικά (attributes) του ίδιου τύπου είτε συνεχή είτε κατηγορικά. Όσο ο ρόλος του Data Mining στις επιχειρήσεις, στις επιστήμες, στη φαρμακοβιομηχανία και σε άλλα πεδία αυξήθηκε, τόσο αυξήθηκε και η ανάγκη για τεχνικές που μπορούν να χειριστούν ετερογενή χαρακτηριστικά. Οι τεχνικές που αναπτύχθηκαν για την εξόρυξη γνώσης τέτοιων συνόλων δεδομένων πρέπει να λαμβάνουν υπόψη τις σχέσεις μεταξύ των δεδομένων, όπως τη χρονική και χωρική συσχέτιση, τη συνδεσιμότητα των γράφων και τις σχέσεις γονέα-παιδιού μεταξύ των στοιχείων ενός ημιδομημένου κειμένου ή ενός XMLεγγράφου.

2.1.4 Ιδιοκτησία των δεδομένων και κατανομή

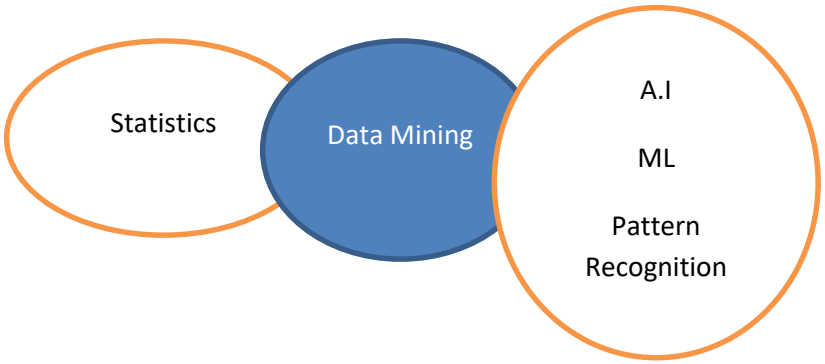
Πολλές φορές, τα δεδομένα που είναι απαραίτητα για μια ανάλυση δεν είναι αποθηκευμένα σε μια περιοχή ούτε ανήκουν μόνο σε έναν οργανισμό. Αντ’ αυτού, βρίσκονται γεωγραφικά κατανεμημένα σε πολλές περιοχές και ανήκουν σε πολλές οντότητες. Αυτό απαιτεί την ανάπτυξη κατανεμημένων τεχνικών Data Mining. Ανάμεσα στις σημαντικές προκλήσεις που αντιμετωπίζουν οι κατανεμημένοι αλγόριθμοι για Data Mining είναι (1) πως θα μειωθεί η ποσότητα της επικοινωνίας που απαιτείται για την πραγματοποίηση των κατανεμημένων υπολογισμών, (2) πως θα παγιωθούν αποτελεσματικά τα αποτελέσματα του Data Mining που προέρχονται από πολλές πηγές και (3) πως θα διευθετηθούν τα θέματα ασφάλειας των δεδομένων.

2.1.5 Μη παραδοσιακή ανάλυση

Η παραδοσιακή στατιστική προσέγγιση στηρίζεται στο παράδειγμα της υπόθεσης-ελέγχου. Με άλλα λόγια, αρχικά προτείνεται μια υπόθεση, στη συνέχεια σχεδιάζεται ένα πείραμα όπου συλλέγονται τα δεδομένα και κατόπιν τα δεδομένα αναλύονται βάσει της συγκεκριμένης υπόθεσης. Δυστυχώς, η διαδικασία αυτή είναι εξαιρετικά χρονοβόρα και κοστοβόρα. Οι παρούσες εργασίες ανάλυσης δεδομένων συχνά απαιτούν τη δημιουργία και αξιολόγηση χιλιάδων τέτοιων υποθέσεων, και κατά συνέπεια, η παραγωγή τεχνικών Data Mining έχει εμπνευστεί από την επιθυμία να αυτοματοποιηθεί η διαδικασία της δημιουργίας υποθέσεων και αξιολόγησης. Συμπληρωματικά, τα σύνολα δεδομένων που αναλύονται δεν αποτελούν το αποτέλεσμα ενός προσεκτικά σχεδιασμένου πειράματος, ενώ αρκετά συχνά εμπεριέχουν μη παραδοσιακούς τύπους δεδομένων και κατανομών.

2.2 Η προέλευση του Data Mining

Ερχόμενοι κοντά με στόχο να αντιμετωπίσουν τις προκλήσεις που αναπτύχθηκαν στο προηγούμενο κεφάλαιο, ερευνητές από διάφορα πεδία άρχισαν να εστιάζουν στην ανάπτυξη πιο αποδοτικών και επεκτάσιμων εργαλείων τα οποία θα μπορούσαν να χειριστούν ποικίλους τύπους δεδομένων. Η δουλειά αυτή, που είχε σαν αποκορύφωμα την ανάπτυξη του πεδίου του Data Mining βασίστηκε σε μεθοδολογίες και αλγόριθμους του παρελθόντος. Πιο συγκεκριμένα, το Data Mining βασίστηκε σε ιδέες όπως (1) η δειγματοληψία, η εκτίμηση και ο έλεγχος υποθέσεων από τη στατιστική και (2) τους αλγόριθμους έρευνας, τις τεχνικές μοντελοποίησης και θεωρίες μάθησης από την τεχνητή νοημοσύνη, την αναγνώριση προτύπων και τη μηχανική μάθηση. Συμπληρωματικά, σημαντικό ρόλο έπαιξαν και άλλες περιοχές. Πιο συγκεκριμένα, τα συστήματα βάσεων δεδομένων χρειάζονται για να παρέχουν υποστήριξη για πιο αποδοτική αποθήκευση, ευρετηρίαση και αναζήτηση πληροφορίας. Ενώ τεχνικές υψηλής υπολογιστικής απόδοσης είναι συχνά σημαντικές προκειμένου για τη διαχείριση των τεράστιων σε όγκο συνόλων δεδομένων.



2.3 Οι εργασίες του Data Mining

Σε γενικές γραμμές, οι εργασίες που αφορούν το Data Mining χωρίζονται σε δυο μεγάλες κατηγορίες:

Προβλεπτικές εργασίες

Σκοπός των συγκεκριμένων εργασιών είναι να προβλέψουν την τιμή ενός χαρακτηριστικού βάσει των τιμών των άλλων χαρακτηριστικών. Το χαρακτηριστικό πάνω στο οποίο πρόκειται να γίνει πρόβλεψη χαρακτηρίζεται συνήθως ως η στοχευμένη ή εξαρτώμενη μεταβλητή ενώ τα υπόλοιπα χαρακτηριστικά (attributes) που χρησιμοποιούνται για την πρόβλεψη χαρακτηρίζονται ως επεξηγηματικές ή ανεξάρτητες μεταβλητές.

Περιγραφικές εργασίες

Εδώ, σκοπός είναι να προκύψουν πρότυπα τα οποία υποβόσκουν στα δεδομένα. Οι περιγραφικές εργασίες είναι συνήθως εξερευνητικές από τη φύση τους και συχνά απαιτούν τεχνικές μεταγενέστερης επεξεργασίας προκειμένου να επικυρωθούν και να ερμηνεύσουν τα αποτελέσματα.

Στη συνέχεια θα ακολουθήσει μια σύντομη ανάλυση των τεσσάρων κομβικών μεθόδων Data Mining που ανήκουν είτε στην προβλεπτική είτε στην περιγραφική κατηγορία.

2.3.1 Προβλεπτική Μοντελοποίηση

Η προβλεπτική μοντελοποίηση αναφέρεται στη δημιουργία ενός μοντέλου για τη στοχευμένη μεταβλητή, σαν συνάρτηση των επεξηγηματικών μεταβλητών. Υπάρχουν δυο τύποι προβλεπτικής Μοντελοποίησης (Predictive Modelling): Η Κατηγοριοποίηση (Classification), η οποία χρησιμοποιείται για διακριτές μεταβλητές και η Παλινδρόμηση (Regression) που χρησιμοποιείται για συνεχόμενες στοχευμένες μεταβλητές. Για παράδειγμα, η πρόβλεψη του αν ένας χρήστης θα πραγματοποιήσει μια αγορά σε ένα διαδικτυακό βιβλιοπωλείο, είναι εργασία κατηγοριοποίησης καθώς η στοχευμένη μεταβλητή παίρνει δυαδικές (binary) αξίες. Από την άλλη, η πρόβλεψη της τιμής μιας μετοχής αποτελεί εργασία Παλινδρόμησης καθώς εδώ, η στοχευμένη μεταβλητή παίρνει συνεχόμενες αξίες. Ο στόχος και των δυο εργασιών είναι η εκπαίδευση ενός μοντέλου το οποίο ελαχιστοποιεί το σφάλμα μεταξύ των προβλεπόμενων και τωνπραγματικών αξιών της στοχευμένης μεταβλητής.

Παραδείγματα προβλεπτικής μοντελοποίησης αποτελούν η αναγνώριση του αν ένας πελάτης θα ανταποκριθεί θετικά σε μια διαφημιστική καμπάνια, η πρόβλεψη διαταράξεων στο γήινο οικοσύστημα ή η επιτυχής αναγνώριση του αν ένας ασθενής πάσχει από μια συγκεκριμένη ασθένεια βάσει προηγούμενων ιατρικών εξετάσεων

2.3.2 Ανάλυση Σχέσεων (Association Analysis)

Η ανάλυση σχέσεων χρησιμοποιείται για την ανακάλυψη προτύπων (patterns) τα οποία περιγράφουν ισχυρά συσχετισμένες σχέσεις μεταξύ των δεδομένων. Τα ανακαλυφθέντα πρότυπα εκφράζονται συνήθως με τη μορφή κανόνων επιπτώσεων ή υποσυνόλων χαρακτηριστικών. Εξαιτίας του εκθετικού μεγέθους του χώρου έρευνας, στόχος της ανάλυσης σχέσεων είναι η εξαγωγή των πιο ενδιαφερόντων προτύπων με τον πιο αποδοτικό τρόπο. Χρήσιμες εφαρμογές της ανάλυσης σχέσεων μπορούνα να αποτελούν η αναγνώριση ιστοσελίδων που οι χρήστες επισκέπτονται μαζί ή η κατανόηση των σχέσεων μεταξύ των διαφορετικών στοιχείων του κλιματικού συστήματος της γης.

2.3.3 Ανίχνευση Ανωμαλιών (Anomaly Detection)

Είναι η εργασία που αναγνωρίζει παρατηρήσεις (observations) των οποίων τα χαρακτηριστικά διαφέρουν σημαντικά από το υπόλοιπο σύνολο δεδομένων. Οι παρατηρήσεις αυτές είναι γνωστές ως ανωμαλίες ή outliers. Στόχος των συγκεκριμένων αλγόριθμων είναι η ανακάλυψη πραγματικών ανωμαλιών και η αποφυγή ψευδός καταχωρημένων κανονικών παρατηρήσεων ως ανώμαλες. Με άλλα λόγια, ένας καλός ανιχνευτής ανωμαλίας πρέπει να έχει έναν υψηλό ρυθμό ανίχνευσης και έναν χαμηλό ρυθμό ψευδούς συναγερμού. Εφαρμογές της ανίχνευσης ανωμαλίας περιλαμβάνουν την ανίχνευση απάτης, τις εισβολές σε δίκτυα, ασυνήθιστα πρότυπα ασθένειας και διαταράξεις οικοσυστημάτων.

2.3.4 Μηχανική Μάθηση (Machine Learning)

Η Μηχανική Μάθηση έχει γίνει πλέον ένας ιδιαίτερα ευρύς όρος που καλύπτει πολλές διαφορετικές περιοχές, από την ταξινόμηση (Classification) μέχρι τη Συσταδοποίηση (Clustering). Έτσι, δεν μπορεί να

δοθεί μια σαφής ετοιμολογία του όρου. Παρόλα αυτά, υπάρχουν αρκετές ομοιότητες με τις οποίες σχεδόν όλοι οι αλγόριθμοι Μηχανική Μάθησης λειτουργούν:

- Όλες οι διεργασίες πραγματοποιούνται με τη χρήση Ηλεκτρονικών Υπολογιστών αξιοποιώντας την ισχύ τους κάνοντας υπολογισμούς οι οποίοι θα ήταν αδύνατο να πραγματοποιηθούν με το χέρι.
- Όλες οι διεργασίες παίρνουν δεδομένα σαν είσοδο.
- Θεωρούν τα σημεία των δεδομένων ως δείγματα μιας κατανομής πιθανότητας του πραγματικού κόσμου.
- Τα δεδομένα είναι πινακοειδή. Υπάρχει μια σειρά για κάθε σημείο και μια στήλη για κάθε χαρακτηριστικό. Τα χαρακτηριστικά δε, είναι αριθμητικά, δυαδικά (binary) ή κατηγορικά.

Όλοι οι αλγόριθμοι είναι σχεδιασμένοι να διαχειρίζονται μόνο πινακοειδή δεδομένα. Τα πινακοειδή δεδομένα χρησιμοποιούνται για πληθώρα μαθηματικών αναλύσεων αφού οι σειρές ενός πίνακα με nσειρές και dστήλες μπορούν να θεωρηθούν ως τοποθεσίες σε ένα χώρο dδιαστάσεων.

Στις περισσότερες εφαρμογές Μηχανική Μάθησης τα σημεία των δεδομένων θεωρούνται ότι έχουν παρθεί από μια υποκείμενη κατανομή και ο στόχος είναι να βρεθούν πρότυπα (Patterns) στα δείγματα τα οποία θα δώσουν στοιχεία για την γενική κατανομή ή θα δώσουν τη δυνατότητα σε επεξεργασία και άλλων δειγμάτων από αυτή.

2.3.4.1 Ιστορικό Πλαίσιο

Η Μηχανική Μάθηση γεννήθηκε εν μέρει από τις αρχικές αποτυχίες στο χώρο της Τεχνητής Νοημοσύνης (Artificial Intelligence). Για μεγάλο χρονικό διάστημα, η ανθρώπινη σκέψη ήταν επικεντρωμένη στη ιδέα ότι θα μπορούσαν να κάνουν τους Ηλεκτρονικούς Υπολογιστές να σκεφτούν, και ήταν γενικότερα αποδεκτό ότι αυτό ήταν ζήτημα χρόνου. Εν τέλει, η Τεχνητή Νοημοσύνη απέτυχε (τουλάχιστον αναφορικά με τη δημοσιότητα που είχε δημιουργήσει) καθώς η τεχνολογία βρίσκεται και πάλι αρκετά μακριά στην προσπάθειά της να μιμηθεί την ανθρώπινη νοημοσύνη, μερικώς επειδή το ανθρώπινο μυαλό είναι εξαιρετικά πιο περίπλοκο από μια απλή λογική μηχανή.

Η κατεύθυνση μετατοπίστηκε από τη δημιουργία πραγματικής τεχνητής νοημοσύνης προς τη χρήση των Ηλεκτρονικών Υπολογιστών σε εργασίες που παραδοσιακά αναλαμβάνουν οι άνθρωποι. Αυτό περιλαμβάνει πράγματα όπως του να αναγνωρίσει, για παράδειγμα, εάν μια φωτογραφία περιέχει ένα συγκεκριμένο αντικείμενο ή αν ένα e-mail είναι κακόβουλο ή αν υπάρχει κάποιο ενδιαφέρον γεγονός σε μια χρονοσειρά. Η Μηχανική Μάθηση δομήθηκε στο να χρησιμοποιεί τους Ηλεκτρονικούς Υπολογιστές σαν υποκατάστατα της ανθρώπινης κρίσης σε πολύ συγκεκριμένες και περιορισμένες περιπτώσεις. Βέβαια, οι τεχνικές που αναπτύχθηκαν βρίσκουν εφαρμογή σε πολλές περιοχές, ακόμα και σε εκείνες που η ανθρώπινη κρίση δεν εμπλεκόταν ποτέ, και έτσι η Μηχανική Μάθηση ωρίμασε σε ένα δεδομένο εργαλείο στο ΄χώρο της επιστήμης των δεδομένων.

2.3.4.2 Μηχανική Μάθηση με Επιτήρηση και Μηχανική Μάθηση χωρίς επιτήρηση [Supervised Versus Unsupervised Machine Learning]

Υπάρχουν δυο βασικές κατηγορίες Μηχανική Μάθηση, αυτή που γίνεται με επιτήρηση και αυτή που γίνεται χωρίς επιτήρηση. Στην επιτηρούμενη Μηχανική Μάθηση, τα δεδομένα εκπαίδευσης αποτελούνται από κάποια σημεία και μια στοχευμένη αξία (μεταβλητή) που συνδέεται με αυτά. Ο στόχος

των αλγορίθμων είναι με κάποιο τρόπο να εκτιμήσουν τη στοχευμένη (targeted) μεταβλητή. Για παράδειγμα, μπορεί να υπάρχουν δεδομένα από αρκετούς ασθενείς τα οποία να δείχνουν την ανάλυση του αίματος τους και στη συνέχεια να σχετίζονται με μορφές καρκίνου. Εάν θελήσουμε να χρησιμοποιήσουμε δείγματα αίματος μελλοντικών ασθενών έτσι ώστε να εκτιμηθεί το ρίσκο εμφάνισης καρκίνου, αυτό αποτελεί ένα πρόβλημα Μηχανική Μάθηση με επιτήρηση.

Στη Μηχανική Μάθηση χωρίς επιτήρηση, υπάρχουν ωμά δεδομένα χωρίς κάποιο συγκεκριμένο στόχο πρόβλεψης. Οι αλγόριθμοι Μη Επιτηρούμενης Μηχανική Μάθησης χρησιμοποιούνται για την ανακάλυψη προτύπων σε δεδομένα γενικώς και αορίστως αποκωδικοποιώντας την επικείμενη δομή των δεδομένων. Οι αλγόριθμοι συσταδοποίησης, για παράδειγμα, χρησιμοποιούνται για να διασπάσουν ένα σύνολο δεδομένων σε συστάδες, δηλαδή ομάδες που έχουν μεταξύ τους κοινά χαρακτηριστικά και αυτό αποτελεί ένα σύνηθες παράδειγμα Μηχανική Μάθησης χωρίς επιτήρηση.

Η Μηχανική Μάθηση με επιτήρηση είναι πιο συνήθης σε πραγματικές εφαρμογές. Παρόλα αυτά οι αλγόριθμοι Μηχανική Μάθηση χωρίς επιτήρηση χρησιμοποιούνται συχνά σαν ένα προπαρασκευαστικό βήμα για την εξαγωγή χρήσιμης πληροφορίας από ένα σημείο δεδομένων και τελικά η πληροφορία αυτή θα χρησιμοποιηθεί για την επιτηρούμενη μάθηση.

2.3.4.3 Δεδομένα εκπαίδευσης (Train Data), Δεδομένα Ελέγχου (Test Data) και ο κίνδυνος του Over fitting

Μακράν ο μεγαλύτερος πονοκέφαλος σε ένα πρόβλημα Μηχανική Μάθησης είναι το Over fitting. Αυτό πρακτικά σημαίνει ότι τα αποτελέσματα είναι εξαιρετικά για τα δεδομένα τα οποία έχουν χρησιμοποιηθεί για την εκπαίδευση, αλλά δεν είναι σε θέση να γενικεύσουν για άλλα μελλοντικά δεδομένα. Για να γίνει καλύτερα κατανοητό, στο σημείο αυτό, θα μπορούσε να δοθεί ένα ακραίο παράδειγμα του Over fitting. Αν υποθέσουμε ότι ένα σύνολο δεδομένων ιατρικών ασθενών περιέχει τα ονόματά τους, και ο αλγόριθμος ταξινόμησης που χρησιμοποιήθηκε απλά είναι σε θέση να θυμάται το όνομα οποιουδήποτε πάσχει από καρκίνο και να κάνει προβλέψεις στηριζόμενος σε αυτό. Τότε, ο συγκεκριμένος αλγόριθμος θα είναι σε θέση να δίνει τέλειες προβλέψεις για οποιοδήποτε ασθενή του συνόλου εκπαίδευσης αλλά, είναι τελείως άχρηστος σε προβλέψεις μελλοντικών-νέων ασθενών.

Η λύση είναι να χρησιμοποιείς μέρος του συνόλου δεδομένων για την εκπαίδευση και το υπόλοιπο για την εκτίμηση της απόδοσης και τον έλεγχο. Αυτό μπορεί να γίνει με πολλούς τρόπους:

- Στο πιο απλό επίπεδο, το σύνολο δεδομένων χωρίζεται τυχαία σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο ελέγχου. Η τυχαιότητα είναι εξαιρετικά σημαντική έτσι ώστε να αποφευχθεί οποιαδήποτε μη σκόπιμη προκατάληψη. Η απλή αυτή προσέγγιση λειτουργεί αρκετά καλά στην πράξη.
- Μια λίγο πιο περίπλοκη μέθοδος που λειτουργεί συγκεκριμένα σε τεχνικές Μηχανική Μάθησης με επιτήρηση είναι το K-fold Cross Validation. Ο στόχος εδώ δεν είναι να μετρηθεί η απόδοση ενός ταξινομητή αλλά μιας οικογένειας ταξινομητών. Πιο συγκεκριμένα το k – fold Cross Validation πραγματοποιείται σύμφωνα με τα ακόλουθα βήματα :

Χωρίζεται το αρχικό σύνολο εκπαίδευσηςσε k τυχαία ισοδύναμα υποσύνολα. Κάθε υποσύνολοκαλείται fold (πτυχή).

Για i = 1 έως i = k :

A. Το fold I χρησιμοποιείται για την επικύρωση και τα υπόλοιπα k-1 foldsγια την εκπαίδευση.
B. Το μοντέλο μηχανικής μάθησης εκπαιδεύεται χρησιμοποιώντας το Cross Validation σύνολο εκπαίδευσης και υπολογίζεται η ακρίβειά του επικυρώνοντας τα αποτελέσματα της πρόβλεψης με το σύνολο επικύρωσης.
C. Τέλος, εκτιμάται η ακρίβεια του μοντέλου παίρνοντας τη μέση τιμή των αποτελεσμάτων των k περιπτώσεων του Cross Validation.

Με τη μέθοδο του k–fold Cross Validationόλα τα παραδείγματα (καταχωρήσεις) του αρχικού συνόλου εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση αλλά και για την επικύρωση του μοντέλου.

Συμπληρωματικά, κάθε παράδειγμα χρησιμοποιείται για την επικύρωση μόνο μία φορά, ενώ η τιμή του kδηλαδή του αριθμού των foldsσυνηθίζεται να παίρνει τιμή ίση με το 10, χωρίς όμως αυτό να αποτελεί κανόνα.

Γενικότερα, μικρότερες τιμές του kαποδίδουν μοντέλα τα οποία είναι φθηνότερα σε υπολογιστικές απαιτήσεις, μικρότερη διακύμανση αλλά και μεγαλύτερη προκατάληψη, ενώ μεγαλύτερες τιμές του kαποδίδουν μοντέλα τα οποία είναι πιο ακριβά σε υπολογιστικές απαιτήσεις, με μεγαλύτερη διακύμανση και μικρότερη προκατάληψη.

- Μια άλλη μέθοδος είναι ο διαχωρισμός του συνόλου δεδομένων σε ένα σύνολο εκπαίδευσης, σε ένα σύνολο ελέγχου και σε ένα σύνολο επικύρωσης. Το σύνολο επικύρωσης (Validation) δεν χρησιμοποιείται πάρα μόνο στο τέλος της όλης διαδικασίας προκειμένου να ελεγχθούν οι υποθέσεις που έγιναν και η απόδοση του μοντέλου. Αυτό γίνεται προκειμένου να αποφευχθεί έστω και η πάρα μικρή στατιστική προκατάληψη.

Αν υποθέσουμε ότι κάποιος είχε μόνο σύνολα εκπαίδευσης και ελέγχου και αρκετά μοντέλα Μηχανικής Μάθησης να διαλέξει, σε αυτή την περίπτωση θα επέλεγε αυτό που είχε την καλύτερη απόδοση όταν εκπαιδευόταν με το ένα σύνολο και ελεγχόταν με το άλλο. Αλλά αυτή είναι μια αδύναμη μορφή εκπαίδευσης στο σύνολο δεδομένων ελέγχου γιατί το σύνολο των δεδομένων που χρησιμοποιήθηκαν για τον έλεγχο επηρεάζουν την επιλογή του μοντέλου. Έτσι εισαγάγετε η έννοια του συνόλου δεδομένων επικύρωσης (validation data) και μπορεί να πραγματοποιηθεί ένας πραγματικός έλεγχος της απόδοσης του μοντέλου χωρίς προκατάληψη.

- Μια τελευταία προσέγγιση στο συγκεκριμένο θέμα αποτελεί η ακόλουθη η οποία βρίσκει χρήση σε εφαρμογές του πραγματικού κόσμου.

Πολλές φορές, δημιουργείται μια κατάσταση όπου ένα μοντέλο επαναεκπαιδεύεται περιοδικά, έστω κάθε εβδομάδα, ενσωματώνοντας νέα δεδομένα που αποκτήθηκαν την προηγούμενή εβδομάδα. Στις περιπτώσεις αυτές, έχει νόημα να εκπαιδευτεί το μοντέλο με όλα τα δεδομένα της εβδομάδας N και των προηγούμενων εβδομάδων και να ελεγχθεί με τα νέα δεδομένα της εβδομάδας N+1.

3. Εισαγωγή στην Κατηγοριοποίηση (Classification)

Το Data Mining περιλαμβάνει την εξαγωγή πληροφορίας από ένα σύνολο δεδομένων και τη μεταμόρφωσή της σε μια δομή που μπορεί να γίνει κατανοητή. Είναι η υπολογιστική διαδικασία της ανακάλυψης προτύπων σε μεγάλα σύνολα δεδομένων κάνοντας χρήση μεθόδων στο σταυροδρόμι της τεχνητής νοημοσύνης, της μηχανικής μάθησης, της στατιστικής και των συστημάτων βάσεων δεδομένων. Το Data Mining επιτρέπει τη λεπτομερή εξέταση μέσα από τον χαοτικό και επαναλαμβανόμενο ΄θόρυβο΄ των δεδομένων. Βοηθάει επίσης στην κατανόηση της σχετικής πληροφορίας και στην καλή χρήση αυτής, προκειμένου να εκτιμηθούν τα πιθανά αποτελέσματα. Έτσι, μέσω του Data Mining επιταχύνεται ο ρυθμός λήψης ορθών και ενημερωμένων αποφάσεων.

Υπάρχουν έξι κατηγορίες στο Data Miningκαι πιο συγκεκριμένα το Anomaly Detection, το Association Rule Learning, το Clustering, το Classificationκαι το Regression. Το Classification είναι μια λειτουργία του Data Mining η οποία αναθέτει αντικείμενα μιας συλλογής, σε κατηγορίες ή κλάσεις. Στοχεύει στην πρόβλεψη της υπό εξέταση κλάσης για κάθε περίπτωση του συνόλου δεδομένων. Για παράδειγμα, ένα μοντέλο ταξινόμησης μπορεί να βοηθήσει στην αναγνώριση αιτήσεων τραπεζικού δανείου ως ασφαλείς ή ριψοκίνδυνες. Οι διάφορες τεχνικές ταξινόμησης που χρησιμοποιούνται στο χώρο του Data Mining είναι τα Δένδρα Απόφασης (Decision Trees Induction), η μέθοδος δημιουργίας κανόνων (rule–based method), η μάθηση βασισμένη στη μνήμη (memory–based learning), τα Μπαγιεσιανά Δίκτυα (Bayesian Networks), τα Νευρωνικά Δίκτυα (Neural Networks) και τα Support Vector Machines.

Υπάρχουν δύο μεγάλες περιπτώσεις χρήσης για έναν ταξινομητή. Η πρώτη είναι η προφανής, έχουμε αντικείμενα που θέλουμε να ταξινομηθούν. Αυτό συμβαίνει συχνά στην παραγωγή όταν, για παράδειγμα, ένας Ηλεκτρονικός Υπολογιστής πρέπει να αποφασίσει για το ποια διαφήμιση να δείξει σε έναν χρήστη. Επίσης, συμβαίνει και όταν οι Ηλεκτρονικοί Υπολογιστές δεν παίρνουν αποφάσεις αυτόνομα αλλά επισημαίνουν πράγματα προκειμένου ένας άνθρωπος να λάβει τις ανάλογες αποφάσεις. Η άλλη χρήση των ταξινομητών είναι να δώσουν πληροφορίες που υπόκεινται στα δεδομένα. Σε περιπτώσεις σαν αυτή, σκοπός είναι η επεξεργασία των αποτελεσμάτων του ταξινομητή προκειμένου να γίνει η εξαγωγή των πληροφοριών για τυχόν πρότυπα που κρύβονται μέσα στα δεδομένα.

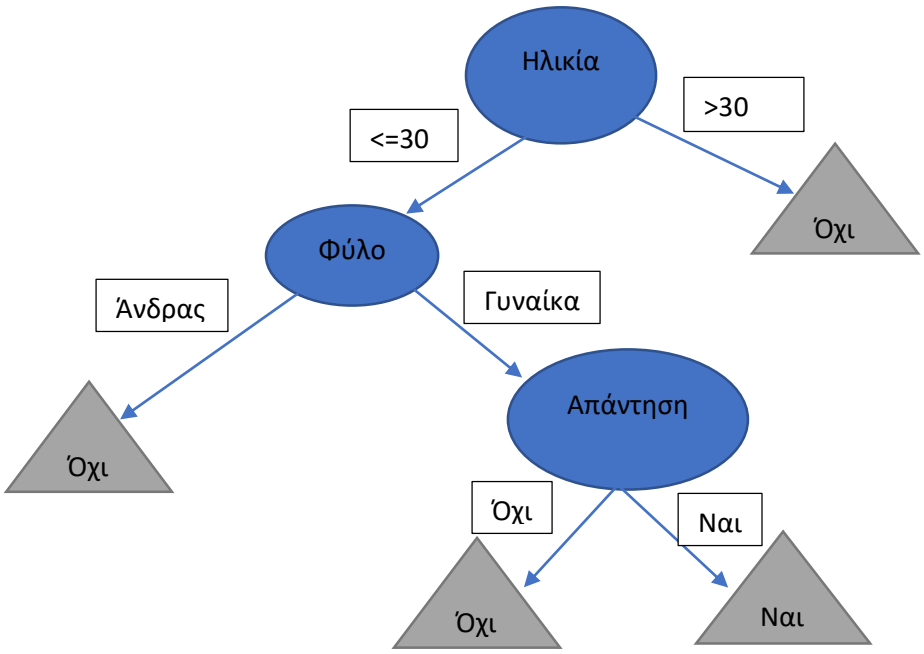
3.1 Δέντρα Απόφασης (Decision Trees)

Ένα Δέντρο Απόφασης είναι ένας κατηγοριοποιητής που εκφράζεται ως το αναδρομικό χώρισμα του χώρου των παραδειγμάτων. Το Δέντρο Απόφασης αποτελείται από κόμβος που δημιουργούν ένα ‘ριζωμένο δέντρο’ υπό την έννοια ότι το συγκεκριμένο δέντρο είναι ένα ‘κατευθυνόμενο δέντρο’ με έναν κόμβο να αποτελεί τη ρίζα, η οποία δεν έχει εισερχόμενες ακμές. Όλοι οι άλλοι κόμβοι έχουν ακριβώς μια εισερχόμενη ακμή. Συμπληρωματικά, οι κόμβοι αυτοί καλούνται ως ‘φύλλα’, κόμβοι απόφασης ή εσωτερικοί κόμβοι. Σε ένα Δέντρο Απόφασης κάθε εσωτερικός κόμβος διαμερίζει το πεδίο των παραδειγμάτων σε δυο ή περισσότερα υπό πεδία σύμφωνα με μια συγκεκριμένη λειτουργία των αξιών των χαρακτηριστικών εισόδου.

Στην πιο απλή και πιο συχνή περίπτωση, κάθε έλεγχος εξετάζει ένα μόνο χαρακτηριστικό έτσι ώστε ο χώρος των παραδειγμάτων να διαμερίζετε σύμφωνα με την αξία του χαρακτηριστικού. Στην περίπτωση

των αριθμητικών χαρακτηριστικών, η προηγούμενη προϋπόθεση αναφέρεται σε συγκεκριμένο εύρος τιμών. Κάθε φύλλο ανατίθεται σε μια κλάση η οποία αντιπροσωπεύει την πιο κατάλληλη στοχευμένη αξία. Εναλλακτικά, ένα φύλλο μπορεί να αντιπροσωπεύει ένα διάνυσμα πιθανότητας το οποίο καταδεικνύει την πιθανότητα η στοχευμένη μεταβλητή να έχει μια συγκεκριμένη αξία. Τα παραδείγματα κατηγοριοποιούνται ανάλογα με την διαδρομή τους από τη ρίζα του δέντρου προς το φύλλο, σύμφωνα με τα αποτελέσματα των ελέγχων κατά τη διαδρομή.

Το ακόλουθο σχήμα περιγράφει ένα Δέντρο Απόφασης το οποίο δείχνει εάν ένας μελλοντικός πελάτης θα απαντήσει θετικά στην άμεση αλληλογραφία (direct mailing). Οι εσωτερικοί κόμβοι παρουσιάζονται ως κύκλοι, ενώ τα φύλλα ως τρίγωνα. Ο συγκεκριμένος κατηγοριοποιητής ενσωματώνει κατηγορικά και αριθμητικά χαρακτηριστικά. Δοθέντος του συγκεκριμένου κατηγοριοποιητή ένας αναλυτής μπορεί να προβλέψει την αντίδραση ενός πιθανού πελάτη και να κατανοήσει τα συμπεριφορικά χαρακτηριστικά ολόκληρου του πληθυσμού των πιθανών πελατών αναφορικά με τις αντιδράσεις τους στην άμεση αλληλογραφία. Κάθε κόμβος σημειώνεται με το χαρακτηριστικό το οποίο ελέγχει ενώ τα κλαδιά του με τις αντίστοιχες αξίες. Στην περίπτωση των αριθμητικών χαρακτηριστικών, τα Δέντρα Απόφασης μπορούν να ερμηνευτούν γεωμετρικά σαν μια συλλογή υπερεπιπέδων, κάθε ένα ορθογώνιο σε έναν από τους άξονες.



Σε γενικές γραμμές προτιμάται η χρήση λιγότερο πολύπλοκων Δέντρων Απόφασης καθώς αυτά θεωρούνται περισσότερο ερμηνεύσιμα. Συμπληρωματικά, η πολυπλοκότητα ενός Δέντρου Απόφασης παίζει σημαντικό ρόλο στην ακρίβειά του. Η πολυπλοκότητα ενός δέντρου σαφέστατα ελέγχεται από τα κριτήρια παύσης (stopping criteria) και την εκάστοτε μέθοδο κλαδέματος (pruning method) που χρησιμοποιείται. Συνήθως η πολυπλοκότητα ενός Δέντρου Απόφασης μετριέται με μια από τις ακόλουθες μετρικές:

- Συνολικός αριθμός κόμβων.
- Συνολικός αριθμός φύλλων.
- Το βάθος του δέντρου.
- Ο αριθμός των χαρακτηριστικών που χρησιμοποιήθηκαν.

3.1.2 Το αλγοριθμικό πλαίσιο των Δέντρων Απόφασης

Υπάρχουν αρκετοί αλγόριθμοι που κατασκευάζουν αυτόματα ένα Δέντρο Απόφασης δοθέντος ενός συνόλου δεδομένων. Συνήθως, ο στόχος είναι να βρεθεί το άριστο Δέντρο Απόφασης ελαχιστοποιώντας το σφάλμα γενίκευσης. Βέβαια, για το σαυγκεκριμένο θέμα μπορούν να ακολουθηθούν και άλλες λειτουργίες όπως, για παράδειγμα, η ελαχιστοποίηση του αριθμού των κόμβων ή η ελαχιστοποίηση του μέσου βάθους του δέντρου.

Η επαγωγή ενός άριστου Δέντρου Απόφασης από ένα σύνολο δεδομένων θεωρείται δύσκολο έργο και στην πράξη χρησιμοποιούνται ευρετικές μέθοδοι για την επίλυση του προβλήματος. Σε γενικές γραμμές, οι μέθοδοι αυτοί χωρίζονται σε δυο ομάδες: την ‘από πάνω προς τα κάτω’ και την ‘από κάτω προς τα επάνω’ με μια σαφέστατη προτίμηση της βιβλιογραφίας στην πρώτη. Υπάρχουν αρκετοί αλγόριθμοι κατασκευής Δέντρων Απόφασης που χρησιμοποιούν την ‘από πάνω προς τα κάτω’ ευρετική (heuristic) μέθοδο κατασκευής όπως ο ID3 (Quinlan, 1986), ο C4.5 (Quinlan, 1993) και ο CART (Breiman et al., 1984). Κάποιοι από αυτούς αποτελούνται από δυο φάσεις: την φάση της ανάπτυξης και την φάση του κλαδέματος, ενώ κάποιοι πραγματοποιούν μόνο τη φάση της ανάπτυξης.

Γενικά, ένας αλγόριθμος Δέντρου Απόφασης ακολουθεί τα παρακάτω βήματα:

- Δοθέντος ενός συνόλου δεδομένων εκπαίδευσης Χ, βρίσκει το μοναδικό εκείνο χαρακτηριστικό που διαχωρίζει καλύτερα τα δεδομένα σε κλάσεις.
- Υπάρχουν αρκετοί τρόποι που ποσοτικοποιούν το κατά πόσο καλός είναι ο διαχωρισμός. Οι πιο κοινοί είναι το “information gain” και το “gini index”.

Πιο συγκεκριμένα, οι αλγόριθμοι Δένδρων Απόφασης χρησιμοποιούν το Information Gain για τον διαχωρισμό των κόμβων. Το Gini Index ή η εντροπία αποτελούν το κριτήριο για τον υπολογισμό του Information Gain. Το Gini Index και η Εντροπία δίνουν το μέγεθος της ακαθαρσίας (impurity) ενός κόμβου. Ένας κόμβος που έχει πολλαπλές κλάσεις θεωρείται ακάθαρτος, ενώ ένας κόμβος που έχει μια και μοναδική κλάση είναι καθαρός. Η Εντροπία πρακτικά δίνει το μέγεθος της αταξίας (disorder). Εάν υπάρχουν πολλαπλές κλάσεις σε έναν κόμβο, τότε υπάρχει αταξία στον κόμβο.

Το Information Gain είναι η Εντροπία του κόμβου γονέα μείον το άθροισμα των σταθμισμένων εντροπιών των κόμβων παιδιών. Το βάρος ενός κόμβου παιδιού είναι ο αριθμός των δειγμάτων του κόμβου προς τα δείγματα όλων των κόμβων παιδιών. Ομοίως, το Information Gain υπολογίζεται και από το Gini Index.

$$Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$$

$$Entropy = \sum_{i=1}^{n} -p(c_i)\log_2(p(ci))$$

όπου $p(c_i)$ είναι η πιθανότητα/ποσοστό της κλάσης (c_i) σε έναν κόμβο.

- Το μοναδικό χαρακτηριστικό που διαχωρίζει καλύτερα τα δεδομένα σε κλάσεις γίνεται η ρίζα του Δένδρου και στη συνέχεια διαχωρίζεται (partition) το σύνολο δεδομένων σύμφωνα με αυτόν τον κόμβο.
- Αναδρομικά εκπαιδεύεται κάθε κόμβος ¨παιδί¨ στο δικό του χώρισμα (partition) δεδομένων.
- Η αναδρομή σταματάει όταν, είτε όλα τα σημεία δεδομένων στο χώρισμα έχουν την ίδια κλάση είτε όταν η αναδρομή έχει φτάσει σε ένα προκαθορισμένο μέγιστο βάθος. Σε αυτό το σημείο τα αποτελέσματα που αποθηκεύτηκαν στον κόμβο θα είναι οι κατανομές των κλάσεων.

3.1.3 Μονομεταβλητά Κριτήρια Διαμερισμού

Στις περισσότερες περιπτώσεις, οι διακεκριμένες λειτουργίες διαμερισμού είναι μονομεταβλητές. Μονομεταβλητή σημαίνει ότι ένας εσωτερικός κόμβος διαμερίζει σύμφωνα με την αξία ενός χαρακτηριστικού. Συνεπώς, ο αλγόριθμος ψάχνει για το καλύτερο χαρακτηριστικό βάσει του οποίου θα κάνει τον διαμερισμό (splitting). Υπάρχουν πολλά μονομεταβλητά κριτήρια τα οποία μπορούν να χαρακτηριστούν με πολλούς τρόπους όπως:

- Σύμφωνα με την προέλευση του μέτρου (measure): Θεωρία Πληροφοριών, εξάρτηση και απόσταση.
- Σύμφωνα με τη δομή του μέτρου: κριτήρια βασισμένα στην ακαθαρσία (impurity), κριτήρια βασισμένα στην κανονικοποιημένη ακαθαρσία και δυαδικά κριτήρια.

Παρακάτω θα ακολουθήσει ανάλυση των πιο δημοφιλών κριτηρίων που συναντά κανείς στη βιβλιογραφία:

3.1.3.1 Κριτήρια βασισμένα στην ακαθαρσία (impurity-based criteria)

Δεδομένης μιας τυχαίας μεταβλητής X με K διακριτές αξίες, κατανεμημένης σύμφωνα με το $P = (p_1, p_2, ..., p_k)$, το μέτρο της ακαθαρσίας είναι μια συνάρτηση $\varphi : [0,1]^K \to R$ που ικανοποιεί τις ακόλουθες συνθήκες:

- $\Phi(\mathrm{P}) \geq 0$
- $\Phi(\mathrm{P})$ ελάχιστο αν $\exists i$ έτσι ώστε $p_i = 1$
- $\Phi(\mathrm{P})$ μέγιστο αν $\forall_i, 1 \leq i \leq k, p_i = 1/k$
- $\Phi(\mathrm{P})$ συμμετρικό με τα στοιχεία του P
- $\Phi(\mathrm{P})$ ομαλό σε όλο του το εύρος

Να σημειωθεί ότι όταν το διάνυσμα πιθανότητας έχει συνιστώσα τη μονάδα, τότε η μεταβλητή χαρακτηρίζεται ως αγνή. Από την άλλη, εάν όλες οι συνιστώσες είναι ίσες, τότε το επίπεδο της ακαθαρσίας φτάνει το μέγιστο.

Δοθέντος ενός συνόλου εκπαίδευσης S, το διάνυσμα πιθανότητας της στοχευμένης μεταβλητής y βρίσκεται από τον ακόλουθο τύπο:

$$P_y(S) = \left(\frac{|\sigma_{y=c_1} S|}{|S|}, \ldots, \frac{|\sigma_{y=c_{|dom(y)|}} S|}{|S|} \right)$$

Το πόσο καλά έγινε ο διαμερισμός λόγω της διακριτού χαρακτηριστικού a_i ορίζεται ως τη μείωση της ακαθαρσίας της στοχευμένης μεταβλητής αφότου αυτή έχει διαχωριστεί σύμφωνα με τις αξίες $v_{i,j} \in dom(a_i)$:

$$\Delta\Phi(a_i, S) = \phi(P_y(S)) - \sum_{j=1}^{|dom(a_i)|} \frac{|\sigma_{a_i} = \upsilon_{i,j} S|}{|S|} \cdot \phi(P_y(\sigma_{a_i} = \upsilon_{i,j} S))$$

3.1.3.2 Κέρδος Πληροφορίας (information Gain)

Το κέρδος της πληροφορίας χρησιμοποιεί το μέτρο της εντροπίας ως μέτρο της ακαθαρσίας

$$InformationGain(a_i, S) =$$
$$Entropy(y, S) - \sum_{u_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i} = u_{i,j} S|}{|S|} \cdot Entropy(y, \sigma_{a_i} = u_{i,j} S)$$

Όπου:

$$Entropy(y, S) = \sum_{c_j \in dom(y)} -\frac{|\sigma_y = c_j S|}{|S|} \cdot \log_2 \frac{|\sigma_y = c_j S|}{|S|}$$

3.1.3.3 Δείκτης Gini (Gini Index)

Ο δείκτης Gini είναι ένα κριτήριο βασισμένο στην ακαθαρσία που μετράει τις αποκλίσεις μεταξύ των κατανομών πιθανότητας για τις αξίες της στοχευμένης μεταβλητής.

$$Gini(y,S) = 1 - \sum_{c_j \in dom(y)} \left(\frac{|\sigma_y = c_j S|}{|S|} \right)^2$$

Συνεπώς το κριτήριο αξιολόγησης για την επιλογή του χαρακτηριστικου a_i , ορίζεται ως:

$$GiniGain(a_i,S) = Gini(y,s) - \sum_{u_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i} = u_{i,j} S|}{|S|} \cdot Gini(y, \sigma_{a_i} = u_{i,j} S)$$

3.1.3.4 Likelihood-Ratio Chi-Squared Statistics

Το Likelihood ratio ορίζεται ως:

$$G^2(a_i,S) = 2 \cdot \ln(2) \cdot |S| \cdot InformationGain(a_i,S)$$

Το likelihood-ratio είναι χρήσιμο για τη μέτρηση της στατιστικής σημασίας του κριτηρίου του κέρδους πληροφορίας. Η μηδενική υπόθεση (Ho) είναι ότι το χαρακτηριστικό εισόδου και το στοχευμένο χαρακτηριστικό είναι υπο όρους ανεξάρτητα. Εάν η υπόθεση κρατήσει, τότε το στατιστικό τέστ κατανέμεται ως x^2 με βαθμούς ελευθερίας ίσους με :

$$(dom(a_i) - 1) \cdot (dom(y) - 1).$$

3.1.3.5 Κριτήριο DKM

Το κριτήριο αυτό είναι βασισμένο στην ακαθαρσία και σχεδιάστηκε για χαρακτηριστικά δυαδικών κλάσεων. Ορίζεται ως:

$$DKM(y,S) = 2 \cdot \sqrt{\left(\frac{|\sigma_y = c_1 S|}{|S|} \right)} \cdot \sqrt{\frac{|\sigma_y = c_2 S|}{|S|}}$$

3.1.3.6 Κριτήρια βασισμένα στην κανονικοποιημένη ακαθαρσία (Normalized impurity-based criteria)

Το κριτήριο που αναλύθηκε πιο πάνω έχει μια προκατάληψη υπέρ χαρακτηριστικών εισόδου με πολλές αξίες σε σχέση με χαρακτηριστικά εισόδου με λίγες αξίες, γεγονός που πολλές φορές οδηγεί το δέντρο στο να κάνει φτωχές γενικεύσεις. Για τον λόγο αυτό, είναι χρήσιμο να κανονικοποιούνται τα μέτρα της ακαθαρσίας όπως περιγράφεται παρακάτω.

3.1.3.7 Αναλογία Κέρδους (Gain Ratio)

Το Gain ratio κανονικοποιεί το κέρδος της πληροφορίας με τον εξής τρόπο:

$$GainRatio(a_i, S) = \frac{InformationGain(a_i, S)}{Entropy(a_i, S)}$$

Καταρχάς, το κέρδος της πληροφορίας υπολογίζεται για όλα τα χαρακτηριστικά. Σαν συνέπεια, λαμβάνονται υπόψη μόνο εκείνα τα χαρακτηριστικά τα οποία απέδωσαν τουλάχιστον το ίδιο καλά με το μέσο κέρδος πληροφορίας και στη συνέχεια επιλέγεται εκείνο το χαρακτηριστικό που είχε την καλύτερη αναλογία κέρδους.

3.1.3.8 Μέτρο της απόστασης (Distance Measure)

Το μέτρο της απόστασης, όπως και η αναλογία κέρδους, κανονικοποιεί το μέτρο της ακαθαρσίας αλλά με έναν διαφορετικό τρόπο:

$$\frac{\Delta\Phi(a_i, S)}{-\sum_{u_{i,j} \in dom(a_i)} \sum_{c_k \in dom(y)} \frac{|\sigma_{a_i} = u_{i,j} AND y = c_k S|}{|S|} \cdot \log_2 \frac{|\sigma_{a_i} = u_{i,j} AND y = c_k S|}{|S|}}$$

Σε γενικές γραμμές η επιλογή του εκάστοτε κριτηρίου διαμερισμού δεν παίζει ιδιαίτερο ρόλο στην απόδοση του Δέντρου Απόφασης. Κάθε κριτήριο είναι ανώτερο σε κάποιες περιπτώσεις και κατώτερο σε άλλες.

3.1.4 Πολυμεταβλητά Κριτήρια

Στα πολυμεταβλητά κριτήρια διαμερισμού, αρκετά χαρακτηριστικά μπορούν να συμμετέχουν στον έλεγχο διαμερισμού ενός συγκεκριμένου κόμβου. Φυσικά, το να βρει κάποιος το καλύτερο πολυμεταβλητό κριτήριο είναι αρκετά πιο πολύπλοκο σε σχέση με το να βρει ένα μονομεταβλητό κριτήριο διαμερισμού. Συμπληρωματικά, ενώ αυτός ο τύπος κριτηρίου μπορεί να αυξήσει δραματικά την απόδοση του δέντρου, στην πράξη τα κριτήρια αυτά είναι πολύ λιγότερο δημοφιλή σε σχέση με τα μονομεταβλητά.

Τα περισσότερα πολυμεταβλητά κριτήρια στηρίζονται στον γραμμικό συνδυασμό των χαρακτηριστικών εισόδου. Η εύρεση του καλύτερου γραμμικού συνδυασμού μπορεί να γίνει με ένα άπληστο ψάξιμο (greedy search), με γραμμικό προγραμματισμό ή με γραμμική διακριτή ανάλυση.

3.1.5 Κριτήρια Παύσης

Η φάση της ανάπτυξης του Δέντρου Απόφασης συνεχίζεται μέχρι να ενεργοποιηθεί ένα κριτήριο παύσης. Οι παρακάτω καταστάσεις αποτελούν συνήθεις κανόνες παύσης:

- Όλα τα παραδείγματα του συνόλου εκπαίδευσης (training set) ανήκουν σε μια μόνο αξία.
- Το μέγιστο βάθος του δέντρου έχει επιτευχθεί.
- Ο αριθμός των περιπτώσεων σε ένα κόμβο απόφασης είναι μικρότερος από τον ελάχιστο αριθμό των περιπτώσεων στους κόμβους γονείς.
- Εάν ένας κόμβος είχε διαμεριστεί και ο αριθμός των περιπτώσεων σε έναν ή περισσότερους κόμβους παιδιά ήταν μικρότερος από τον ελάχιστο αριθμό των περιπτώσεων των κόμβων παιδιών.
- Το καλύτερο κριτήριο διαμερισμού δεν είναι μεγαλύτερο από ένα συγκεκριμένο κατώφλι.

3.1.6 Μέθοδοι Κλαδέματος (Pruning Methods)

Χρησιμοποιώντας αυστηρά κριτήρια παύσης προκαλείται η τάση δημιουργίας μικρών και under-fitted Δέντρων Απόφασης. Under fitting καλείται η περίπτωση όπου το μοντέλο ‘δεν έχει μάθει αρκετά’ από το σύνολο δεδομένων εκπαίδευσης, με αποτέλεσμα το χαμηλό επίπεδο γενίκευσης και τις αναξιόπιστες προβλέψεις.

Από την άλλη μεριά, χρησιμοποιώντας χαλαρά κριτήρια παύσης οδηγείται κανείς στη δημιουργία μεγάλων Δέντρων Απόφασης τα οποία είναι over-fitted στο σύνολο των δεδομένων. Το overfitting είναι η περίπτωση όπου η γενίκευση του μοντέλου είναι αναξιόπιστη αφού αυτό ‘έχει μάθει πάρα πολλά’ από το σύνολο δεδομένων εκπαίδευσης.

Οι μέθοδοι κλαδέματος δημιουργήθηκαν προκειμένου να δώσουν απάντηση σε αυτό το δίλλημα. Σύμφνα με τη μεθοδολογία, επιλέγεται να χρησιμοποιηθεί ένα χαλαρό κριτήριο παύσης το οποίο επιτρέπει στο Δέντρο Απόφασης να γίνει overfit στο σύνολο δεομένων εκπαίδευσης. Στη συνέχεια το over-fitted δέντρο κόβεται σε μικρότερα αφαιρώντας υπό κλαδιά τα οποία δεν συνεισφέρουν στην ακρίβεια της γενίκευσης. Είναι πλέον κατανοητό ότι χρησιμοποιώντας τεχνικές κλαδέματος οδηγούμαστε στη βελτίωση της απόδοσης της γενίκευσης του Δέντρου Απόφασης, ειδικά όταν υπάρχει αισθητός θόρυβος στα δεδομένα.

Υπάρχουν αρκετές τεχνικές κλαδέματος. Οι περισσότερες από αυτές πραγματοποιούν ένα πέρασμα στους κόμβους ‘από πάνω προς τα κάτω’ ή ‘από κάτω προς τα επάνω’. Ένας κόμβος κλαδεύεται όταν η επιχείρηση αυτή βελτιώνει συγκεκριμένα κριτήρια. Η ανάλυση των πιο δημοφιλών τεχνικών ακολουθεί παρακάτω.

3.1.6.1 Κλάδεμα πολυπλοκότητας κόστους (Cost complexity pruning)

Το συγκεκριμένο κλάδεμα λειτουργεί σε δυο στάδια. Στο πρώτο στάδιο χτίζεται μια ακολουθία δέντρων $T_o, T_1, \ldots, T_k$ όπου το T_0 είναι το αρχικό δέντρο πριν το κλάδεμα και το T_k είναι το δέντρο ρίζα. Στο

δεύτερο στάδιο, ένα από τα δέντρα επιλέγεται ως το κλαδεμένο δέντρο, βάσει του σφάλματος γενίκευσης. Το δέντρο T_{i+1} αποκτιέται αντικαθιστώντας ένα ή περισσότερα από τα υπο δέντρα στο προηγηθείς δέντρο T_i με κατάλληλα φύλλα. Τα υπό δέντρα που κλαδεύονται είναι εκείνα τα οποία έχουν αποκτήσει την χαμηλότερη αύξηση στον ρυθμό σφάλματος ανά κλαδεμένο φύλλο:

$$a = \frac{\varepsilon(pruned(T,t),S) - \varepsilon(T,S)}{|leaves(T)| - |leaves(pruned(T,t))|}$$

Όπου το $\varepsilon(T,S)$ υποδηλώνει τον ρυθμό σφάλματος του δέντρου Τ από το δείγμα S και το $|leaves(T)|$ υποδηλώνει τον αριθμό των φύλλων του Τ.

Στο δεύτερο στάδιο, εκτιμάται το σφάλμα γενίκευσης κάθε κλαδεμένου δέντρου $T_o, T_1, ..., T_k$ και επιλέγεται το καλύτερα κλαδεμένο.

3.1.6.2 Κλάδεμα μειωμένου σφάλματος (Reduced error pruning)

Η συγκεκριμένη μέθοδος αποτελεί μια αρκετά απλή διαδικασία και προτάθηκε αρχικά από τον Quinlan (1987). Ενώ διανύει τους εσωτερικούς κόμβους από κάτω προς τα επάνω, η διαδικασία ελέγχει κάθε εσωτερικό κόμβο εάν το κατά πόσο η αντικατάσταση με την πιο συνήθη κλάση δεν μειώνει την ακρίβεια του δέντρου. Εάν ισχύει αυτό τότε ο κόμβος κλαδεύεται, Η διαδικασία αυτή συνεχίζεται μέχρι το σημείο όπου κάποιο περαιτέρω κλάδεμα μειώνει την ακρίβεια.

3.1.6.3 Κλάδεμα ελάχιστου σφάλματος (Minimum error pruning)

Το συγκεκριμένο κλάδεμα πραγματοποιεί μια διασταύρωση από κάτω προς τα επάνω στους εσωτερικούς κόμβους. Σε κάθε κόμβο συγκρίνει την l-πιθανότητα της εκτίμησης του ρυθμού σφάλματος με ή χωρίς κλάδεμα. Η l-πιθανότητα εκτίμησης του ρυθμού σφάλματος αποτελεί διόρθωση της απλής εκτίμησης πιθανότητας χρησιμοποιώντας συχνότητες. Εάν το S_t υποδηλώνει τα παραδείγματα τα οποία έχουν φτάσει σε ένα φύλλο t, τότε ο αναμενόμενος ρυθμός σφάλματος στο συγκεκριμένο φύλλο είναι:

$$\varepsilon'(t) = 1 - \max_{c_i \in dom(y)} \frac{|\sigma_y = c_i S| + l \cdot p_{apr}(y = c_i)}{|S_t| + l}$$

Όπου $p_{apr}(y = c_i)$ είναι η εκ των προτέρων πιθανότητα του y να πάρει την αξία c_i, ενώ το l υποδηλώνει το βάρος.

Ο ρυθμός σφάλματος ενός εσωτερικού κόμβου είναι ο σταθμισμένος μέσος του ρυθμού σφάλματος των λαδιών του. Το βάρος καθορίζεται σύμφωνα με την αναλογία των παραδειγμάτων κατά μήκος του κλαδιού. Ο υπολογισμός αυτός πραγματοποιείται αναδρομικά μέχρι τα φύλλα. Εάν ένας εσωτερικός κόμβος κλαδευτεί, τότε γίνεται φύλλο και ο ρυθμός σφάλματος υπολογίζεται απευθείας χρησιμοποιώντας την προηγούμενη εξίσωση.

Σε αυτό το σημείο κρίνεται σκόπιμο να επισημανθεί ότι και εδώ δεν υπάρχει κάποια μέθοδος η οποία να δίνει καλύτερα αποτελέσματα από τις άλλες καθώς κάποιες οδηγούν σε υπερ-κλάδεμα (over-pruning),

δηλαδή στη δημιουργία μικρότερων Δέντρων Απόφασης με μικρότερη ακρίβεια, και κάποιές άλλες σε υπο-κλάδεμα (under-pruning)

3.1.7 Αλγόριθμοι δημιουργίας Δέντρων Απόφασης

3.1.7.1 ID3 (Quinlan, 1986)

Ο συγκεκριμένος αλγόριθμος θεωρείται ως ένας πολύ απλός αλγόριθμος δημιουργίας Δέντρων Απόφασης. Ο ID3 χρησιμοποιεί το κέρδος της πληροφορίας ως κριτήριο διαμερισμού. Η ανάπτυξη του Δέντρου σταματάει όταν όλα τα παραδείγματα ανήκουν σε μια αξία της στοχευμένης μεταβλητής (target feature) ή όταν το καλύτερο κέρδος πληροφορίας δεν είναι μεγαλύτερο του μηδενός. Ο ID3 δεν εφαρμόζει τεχνικές κλαδέματος ούτε διαχειρίζεται αριθμητικά χαρακτηριστικά ή χαμένες αξίες.

3.1.7.2 C4.5 (Quinlan, 1993)

Ο C4.5 είναι μια μετεξέλιξη του ID3. Χρησιμοποιεί την αναλογία κέρδους (gain ratio) σαν κριτήριο διαμερισμού. Ο διαμερισμός σταματάει όταν ο αριθμός των παραδειγμάτων που πρόκειται να δια μεριστούν έχει πέσει κάτω από ένα συγκεκριμένο κατώφλι. Μετά τη φάση της ανάπτυξης γίνεται χρήση κλαδέματος βασισμένο στο σφάλμα (error-based pruning). Ο C4.5 μπορεί να διαχειριστεί αριθμητικά δεδομένα καθώς και χαμένες αξίες.

3.1.7.3 CART (Breiman et al., 1984)

Ο CART σημαίνει Δέντρα Κατηγοριοποίησης και Παλινδρόμησης (Classification and Regression Trees). Χαρακτηρίζεται από το γεγονός ότι κατασκευάζει δυαδικά δέντρα όπου κάθε εσωτερικός κόμβος έχει ακριβώς δυο εξερχόμενες ακμές. Οι διαμερισμοί επιλέγονται χρησιμοποιώντας το κριτήριο Twoing και το τελικό δέντρο κλαδεύεται με τη μέθοδο πολυπλοκότητας κόστους. Ένα σημαντικό γνώρισμα του αλγόριθμου CART είναι ότι μπορεί να παράξει δέντρα παλινδρόμησης. Τα δέντρα αυτά είναι δέντρα που τα φύλλα τους προβλέπουν ένα πραγματικό νούμερο και όχι μια κλάση. Στην περίπτωση της παλινδρόμησης, ο CART πραγματοποιεί διαμερισμούς που ελαχιστοποιούν το τετραγωνικό σφάλμα της πρόβλεψης. Η πρόβλεψη σε κάθε φύλλο στηρίζεται στον σταθμισμένο μέσο του κόμβου.

3.1.7.4 CHAID (Kass,1980)

Ο αλγόριθμος CHAID (Chi square-Automatic-Interaction-Detection) αρχικά σχεδιάστηκε για να διαχειρίζεται μόνο nominal χαρακτηριστικά. Για κάθε χαρακτηριστικό εισόδου Ai, ο CHAID ψάχνει το ζευγάρι αξιών στο Vi το οποίο να έχει την μικρότερη σημαντική διαφορά από το στοχευμένο χαρακτηριστικό. Η σημαντική διαφορά μετριέται από την αξία (ρ) που προέρχεται από ένα στατιστικό τέστ. Το στατιστικό τέστ που χρησιμοποιείται εξαρτάται από τον τύπο του στοχευμένου χαρακτηριστικού. Εάν το στοχευμένο χαρακτηριστικό είναι συνεχές, τότε πραγματοποιείται ένα f-test. Εάν είναι nominal τότε πραγματοποιείται ένα Pearson chi-squared test. Ένα είναι ordinal τότε πραγματοποιείται ένα likelihood ratio test. Για κάθε επιλεγμένο ζευγάρι, ο CHAID ελέγχει εάν η αξία (ρ) είναι μεγαλύτερη από ένα συγκεκριμένο κατώφλι. Εάν η απάντηση είναι θετική, τότε συγχωνεύει τις αξίες και αναζητεί ένα άλλο ζευγάρι για να συγχωνεύσει. Η διαδικασία επαναλαμβάνεται μέχρι να μην μπορούν να βρεθούν άλλα σημαντικά ζευγάρια.

Στη συνέχεια επιλέγεταιτο καλύτερο χαρακτηριστικό εισόδου για τον διαμερισμό του εκάστοτε κόμβου, έτσι ώστε κάθε κόμβος παιδί να αποτελείται από μια ομάδα ομογενών αξιών του χαρακτηριστικού που έχει επιλεγεί. Η διαδικασία αυτή σταματάει όταν μια από τις ακόλουθες καταστάσεις έχει πραγματοποιηθεί:

1. Το μέγιστο βάθος του δέντρου έχει επιτευχθεί.
2. Όταν ένας κόμβος δεν μπορεί να διαμεριστεί περαιτέρω.
3. Όταν ο ελάχιστος αριθμός περιπτώσεων σε ένα κόμβο, για να είναι κόμβος παιδί, έχει επιτευχθεί.

Τελειώνοντας, ο CHAID δεν πραγματοποιεί κλάδεμα αλλά έχει τη δυνατότητα διαχείρισης χαμένων αξιών.

3.1.7.5 QUEST (Quick, Unbiased, Efficient, Statistical Tree) (Loh & Shih, 1997)

Ο αλγόριθμος QUEST υποστηρίζει μονομεταβλητούς διαμερισμούς και διαμερισμούς γραμμικών συνδυασμών. Για κάθε διαμερισμό, η συσχέτιση μεταξύ κάθε χαρακτηριστικού εισόδου και του στοχευμένου χαρακτηριστικού γίνεται με τη χρήση του ANOVA-Test ή του Levene-Test ή του Pearson Chi-Squared, ανάλογα με τον τύπο του στοχευμένου χαρακτηριστικού. Εάν το στοχευμένο χαρακτηριστικό είναι multinomial, τότε πραγματοποιείται συσταδοποίηση 2-μέσων προκειμένου να παραχθούν δυο υπερ κλάσσεις. Το χαρακτηριστικό που λαμβάνει τον υψηλότερο βαθμό συσχέτισης με το στοχευμένο χαρακτηριστικό επιλέγεται για τον διαμερισμό. Quadratic Discriminant Analysis πραγματοποιείται έτσι ώστε να βρεθεί το βέλτιστο σημείο διαμερισμού του χαρακτηριστικού εισόδου. Ο QUEST έχει μηδαμινή προκατάληψη και δημιουργεί δυαδικά Δέντρα Απόφασης. Για το κλάδεμα των δέντρων χρησιμοποιείται Ten-fold Cross Validation.

3.1.8 Πλεονεκτήματα και Μειονεκτήματα των Δέντρων Απόφασης

Στη βιβλιογραφία συναντάει κανείς πολλά πλεονεκτήματα της χρήσης ενός Δέντρου Απόφασης σαν εργαλείο κατηγοριοποίησης:

1. Τα Δέντρα Απόφασης είναι εύκολο να εξηγηθούν και όταν συμπτυχθούν μπορεί κανείς εύκολα να τα ακολουθήσει και να τα κατανοήσει ακόμα και αν δεν είναι επαγγελματίας του χώρου. Συμπληρωματικά, τα Δέντρα Απόφασης μπορούν να μετατραπούν σε ένα σύνολο κανόνων, γεγονός που κάνει την κατανόησή τους ακόμα πιο εύκολη.

2. Τα Δέντρα Απόφασης μπορούν να χειριστούν κατηγορικά αλλά και αριθμητικά χαρακτηριστικά εισόδου.

3. Η παρουσίαση ενός Δέντρου Απόφασης είναι αρκετά πλούσια για να εκπροσωπήσει οποιονσήποτε κατηγοριοποιητή διακριτής αξίας.

4. Τα Δέντρα Απόφασης είναι ικανά να χειριστούν σύνολα δεδομένων τα οποία μπορεί να έχουν λάθη.

5. Τα Δέντρα Απόφασης είναι ικανά να χειριστούν σύνολα δεδομένων τα οποία μπορεί να έχουν χαμένες αξίες.

6. Τα Δέντρα Απόφασης θεωρούνται ως μη παραμετρικές μέθοδοι. Αυτό σημαίνει ότι δεν κάνουν υποθέσεις για την χωρική κατανομή ή την δομή του κατηγοριοποιητή.

Από την άλλη πλευρά, τα Δέντρα Απόφασης έχουν μειονεκτήματα όπως:

1. Οι περισσότεροι αλγόριθμοι απαιτούν το στοχευμένο χαρακτηριστικό να έχει διακριτές αξίες.

2. Επειδή τα Δέντρα Απόφασης χρησιμοποιούν τη μέθοδο του 'διαίρει και βασίλευε', έχουν την τάση να αποδίδουν καλά μόνο όταν υπάρχουν λίγα σχετικά χαρακτηριστικά και να μην αποδίδουν καλά όταν υπάρχουν περίπλοκες αλληλεπιδράσεις μεταξύ των χαρακτηριστικών.

3. Ο 'άπληστος' χαρακτήρας των Δέντρων Απόφασης οδηγεί και σε ένα ακόμα μειονέκτημα. Αυτό της υπερευαισθησίας στο σύνολο δεδομένων εκπαίδευσης σε μη σχετικά χαρακτηριστικά και στον θόρυβο.

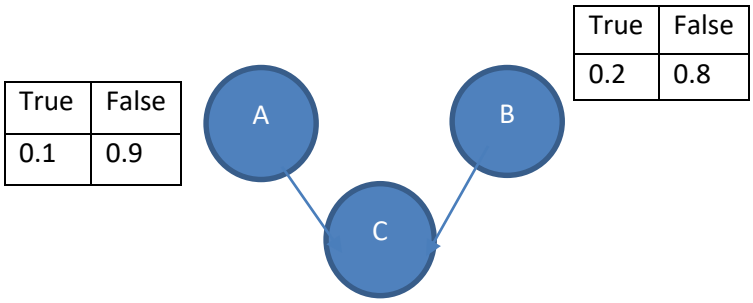
3.2 Δίκτυα Bayes (Bayesian Networks)

Τα Δίκτυα Bayes ανήκουν σε μια γενική ομάδα μοντέλων που λέγονται Γραφικά Μοντέλα Πιθανοφάνειας (Probabilistic Graphical Models) τα οποία δημιουργήθηκαν από τον συνδυασμό της Θεωρίας των Γράφων και της Θεωρίας Πιθανοφάνειας. Η επιτυχία τους έχει να κάνει με το γεγονός ότι είναι σε θέση να διαχειριστούν περίπλοκα μοντέλα πιθανοφάνειας αποσυνθέτοντας τα σε μικρότερα υπαγόμενα μέρη. Ένα γραφικό μοντέλο πιθανοφάνειας χαρακτηρίζεται από έναν γράφο όπου οι κόμβοι αντιπροσωπεύουν τις στοχαστικές μεταβλητές και τα τόξα αντιπροσωπεύουν τις εξαρτήσεις μεταξύ των μεταβλητών αυτών. Τα τόξα χαρακτηρίζονται από την κατανομή πιθανότητας που διέπει την αλληλεπίδραση μεταξύ των μεταβλητών αυτών.

Ένα γραφικό μοντέλο πιθανοφάνειας χαρακτηρίζεται ως Δίκτυο Bayes όταν ο γράφος που συνδέει τις μεταβλητές είναι κατευθυνόμενος και απεριοδικός (Directed Acyclic Graph – DAG). Ο γράφος αυτός αντιπροσωπεύει τις υποθέσεις για την υπό όρους ανεξαρτησία που χρησιμοποιήθηκαν για να πραγματοποιήσουν την κοινή κατανομή πιθανότητας των μεταβλητών του δικτύου κάνοντας έτσι την διαδικασία εκμάθησης από μεγάλα σύνολα δεδομένων υπαγόμενη.

Ένα Δίκτυο Bayes επαγόμενο από ένα σύνολο δεδομένων μπορεί να χρησιμοποιηθεί για να διερευνήσει τις μακρινές σχέσεις μεταξύ των μεταβλητών, να κάνει προβλέψεις και να επεξηγήσει, υπολογίζοντας την κατανομή της υπό όρους πιθανότητας μιας μεταβλητής, δοθέντος των αξιών άλλων μεταβλητών.

Ένα Δίκτυο Bayes αποτελείται από δυο μέρη: από ένα DAG και από την κατανομή της πιθανότητας. Οι κόμβοι του DAG παριστάνουν στοχαστικές μεταβλητές, ενώ τα τόξα παριστάνουν κατευθυνόμενες εξαρτήσεις μεταξύ των μεταβλητών, που ποσοτικοποιούνται από υπό όρους κατανομές πιθανοτήτων. Σαν παράδειγμα μπορεί να θεωρηθεί το ακόλουθο σενάριο όπου δυο μεταβλητές ελέγχουν την αξία μιας τρίτης. Χαρακτηρίζουμε αυτές τις μεταβλητές με τα γράμματα A, B και C, και θεωρούμε ότι καθεμία παίρνει δυο αξίες 'True' και 'False'. Το Δίκτυο Bayes περιγράφει την εξάρτηση των τριων μεταβλητών με έναν DAG όπου τα δυο τόξα που στρέφονται προς τον κόμβο C αντιπροσωπεύουν την κοινή δράση των μεταβλητών A, B.



A	B	True	False
False	False	0.3	0.7
False	True	0.6	0.4
True	False	0.7	0.3
True	True	0.9	0.1

Η απουσία οποιουδήποτε κατευθυνόμενου τόξου μεταξύ των Α και Β περιγράφει την οριακή ανεξαρτησία των δυο μεταβλητών, οι οποίες γίνονται εξαρτημένες όταν τεθούν οι όροι στο φαινότυπο. Ακολουθώντας την κατεύθυνση των τόξων, καλούμε τον κόμβο C παιδί των Α και Β, οι οποίοι τώρα γίνονται γονείς. Το Δίκτυο Bayes του παραδείγματος μας επιτρέπει να αποδομήσουμε την κοινή κατανομή πιθανότητας των τριών μεταβλητών που θα αποτελούνται από $2^3 - 1 = 7$ παραμέτρους, σε τρεις κατανομές πιθανοτήτων, μια υπό όρους κατανομή για την μεταβλητή C δοθέντος των γονέων, και δυο οριακές κατανομές για τις μεταβλητές γονείς Α και Β.

Οι πιθανότητες αυτές προσδιορίζονται από 1+1+4=6 παραμέτρους. Η αποδόμηση αποτελεί έναν από τους βασικούς παράγοντες που προσδίδουν ευκολία στην κατανόηση του συστήματος από τον άνθρωπο και στην αποθήκευση της συγκεκριμένης κατανομής η οποία μεγαλώνει εκθετικά με τον αριθμό των μεταβλητών του πεδίου. Ο δεύτερος παράγοντας κλειδί είναι η χρήση της υπό όρους ανεξαρτησίας μεταξύ των μεταβλητών του δικτύου που διασπάνε τη συνολική κατανομή τους σε επί μέρους ενωμένες ενότητες.

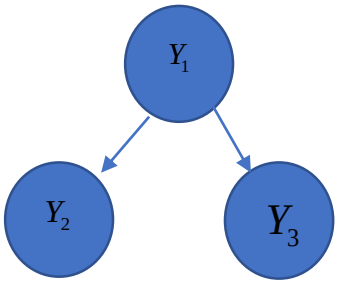
Εάν υποθέσουμε ότι έχουμε τρεις μεταβλητές τις Y_1, Y_2, Y_3, τότε η Y_1 και η Y_2 είναι ανεξάρτητες, δεδομένης της μεταβλητής Y_3, εάν η υπό όρους κατανομή της Y_1, δεδομένων των Y_2, Y_3, αποτελεί μόνο μια λειτουργία της Y_3:

$p(y_1 \mid y_2, y_3) = p(y_1 \mid y_3)$, όπου το $p(y \mid x)$ υποδηλώνει την υπό όρους πιθανότητα/ πυκνότητα του Y όταν $Y = X$.

Με κεφαλαία γράμματα υποδηλώνονται οι τυχαίες μεταβλητές και με μικρά γράμματα οι αξίες τους. Επίσης με τον τύπο $(Y_1 \perp Y_2 \mid Y_3)$ υποδηλώνεται η υπό όρους ανεξαρτησία των μεταβλητών Y_1 και Y_2 δεδομένης της μεταβλητής Y_3.

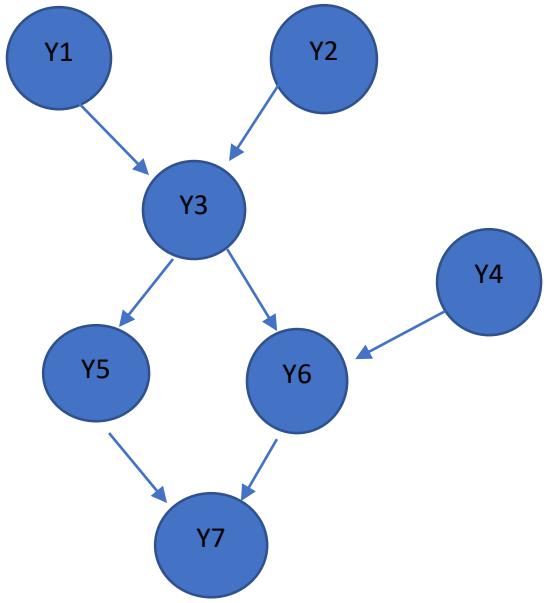
Οι έννοιες της υπό όρους και της οριακής ανεξαρτησίας είναι ουσιαστικά διαφορετικές μεταξύ τους. Για παράδειγμα, δυο μεταβλητές μπορούν να είναι οριακά ανεξάρτητες μεταξύ τους αλλά να γίνουν εξαρτημένες όταν διέπονται από μια Τρίτη μεταβλητή. Ο DAG του σχήματος 1 δείχνει αυτή την ιδιότητα: οι δυο μεταβλητές γονείς είναι οριακά ανεξάρτητες αλλά γίνονται εξαρτημένες όταν διέπονται από το κοινό τους παιδί. Γνωστή συνέπεια αυτού αποτελεί το Παράδοξο του Simpson: δυο μεταβλητές είναι ανεξάρτητες αλλά όταν παρατηρείται μια κοινή μεταβλητή παιδί, τότε γίνονται εξαρτημένες.

Αντιστρόφως, δυο μεταβλητές που είναι οριακά εξαρτημένες μπορεί να γίνουν υπό όρους ανεξάρτητες συστήνοντας μια Τρίτη μεταβλητή. Η κατάσταση αυτή απεικονίζεται από τον DAG του παρακάτω σχήματος που δείχνει δυο κόμβους παιδιά (Y_1 και Y_2) με έναν κοινό γονέα τον κόμβο Y_3.



Στην περίπτωση αυτή, οι δυο κόμβοι παιδιά είναι ανεξάρτητοι, δεδομένου του κοινού γονέα. Αλλά μπορεί να γίνουν εξαρτημένοι όταν περιθωριοποιηθεί ο γονέας.

Η συνολική λίστα των οριακών και υπό όρους ανεξαρτησιών που παρουσιάζεται από τον DAG συνοψίζεται από τις τοπικές και τις καθολικές ιδιότητες του Markov (Markov Properties) που παρουσιάζονται στο ακόλουθο σχήμα χρησιμοποιώντας ένα δίκτυο επτά μεταβλητών.



Local Markov Property

$Y \perp ND(Y) \mid P_a(Y)$
$Y_5 \perp Y_1, Y_2 \mid Y_3$
$Y_6 \perp Y_1, Y_2 \mid Y_3, Y_4$
$Y_7 \perp Y_1, Y_2, Y_3, Y_4 \mid Y_5, Y_6$

$ND(Y)$: Οι μη απόγονοι του Y, είναι όλοι οι κόμβοι όπου μπορούν να φτάσουνε το Y κατά μήκος μιας διαδρομής.

$P_a(Y)$: Υποδηλώνει τους γονείς του Y.

Global Markov Property

$Y \perp Y \setminus MB(Y) \mid MB(Y)$
$Y_1 \perp Y_4, Y_5, Y_6, Y_7 \mid Y_2, Y_3$
$Y_2 \perp Y_4, Y_5, Y_6, Y_7 \mid Y_1, Y_3$
$Y_3 \perp Y_7 \mid Y_1, Y_2, Y_4, Y_5, Y_6$

$MB(Y)$: Το Markov Blanket του Y δοθέντος από τους γονείς του Y, τα παιδιά του Y και τους γονείς των παιδιών του Y.

Το local markov property δηλώνει ότι κάθε κόμβος είναι ανεξάρτητος από τον μη απόγονό του δεδομένων των κόμβων γονέων και οδηγεί σε μια άμεση παραγοντοποίηση της κοινής κατανομής των μεταβλητών του δικτύου στο προϊόν της υπό όρους κατανομής κάθε μεταβλητής Y_i δεδομένων των γονέων της $P_a(Y_i)$.

Κατά συνέπεια, η κοινή πιθανότητα (ή πυκνότητα) του δικτυού U των μεταβλητών, μπορεί να γραφτεί ως:

$$P(y_1,...,y_v) = \prod_i \rho(y_i \mid p_a(y_i))$$

Με την αποδόμηση αυτή, η συνολική κατανομή διασπάται σε επιμέρους συγγενικές ενότητες και το δίκτυο συνοψίζει όλες τις σημαντικές εξαρτήσεις χωρίς καμία απώλεια πληροφορίας. Για παράδειγμα, εάν υποθέσουμε ότι όλες οι μεταβλητές στο προηγούμενο δίκτυο είναι κατηγορικές τότε η κοινή-συλλογική πιθανότητα $p(y_1,...,y_7)$ μπορεί να γραφτεί ως το αποτέλεσμα επτά εξαρτημένων κατανομών:

$$p(y_1)\cdot p(y_2)\cdot p(y_3 \mid y_1, y_2)\cdot p(y_4)\cdot p(y_5 \mid y_3)\cdot p(y_6 \mid y_3, y_4)\cdot p(y_7 \mid y_3, y_6)$$

Από την άλλη μεριά, το global markov property συνοψίζει όλες τις υπό όρους ανεξαρτησίες που ενσωματώνονται στον DAG μέσω της αναγνώρισης του markov blanket κάθε κόμβου.

3.2.1 Απλοϊκός Κατηγοριοποιητής Bayes

Ο Απλοϊκός Κατηγοριοποιητής Bayes (AKB) υπολογίζει την υπό όρους πιθανότητα της κλάσης θεωρώντας ότι τα χαρακτηριστικά είναι υπό όρους ανεξάρτητα, δεδομένης της κλάσης y . Η υπόθεση της υπό όρους ανεξαρτησίας απεικονίζεται ως εξής:

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y)$$

Όπου κάθε σύνολο χαρακτηριστικών $X = \{x_1, x_2, ...x_d\}$ αποτελείται από d χαρακτηριστικά.

Λόγω της υπόθεσης της ανεξαρτησίας, αντί να υπολογίζει την πιθανότητα της κλάσης για κάθε συνδυασμό του X , ο αλγόριθμος εκτιμά μόνο την πιθανότητα του κάθε X_i , δεδομένου του Y . Αυτή η προσέγγιση κρίνεται ιδιαιτέρως αποτελεσματική καθώς δεν απαιτείται ένα μεγάλο σύνολο δεδομένων εκπαίδευσης προκειμένου να υπολογιστεί η πιθανότητα.

Για να κατηγοριοποιηθεί ένα παράδειγμα, ο AKB υπολογίζει την μεταγενέστερη πιθανότητα για κάθε κλάση του Y :

$$P(Y \mid X) = \frac{P(Y)\cdot \prod_{i=1}^{d} P(X_i \mid Y)}{P(X)}$$

Αφού το $P(X)$ είναι σταθερό για κάθε Y , αρκεί να επιλεχτεί η κλάση εκείνη που μεγιστοποιεί τον όρο του αριθμητή $P(Y)\cdot \prod_{i=1}^{d} P(X_i \mid Y)$.

3.2.1.1 Υπολογισμός των πιθανοτήτων για τα χαρακτηριστικά

Για ένα κατηγορικό χαρακτηριστικό X_i, η πιθανότητα $P(X_i = x_i \mid Y = y)$ υπολογίζεται σύμφωνα με το κλάσμα των παραδειγμάτων εκπαίδευσης της κλάσης y, που παίρνουν την αξία ενός συγκεκριμένου χαρακτηριστικού x_i.

Όταν τα χαρακτηριστικά είναι συνεχή υπάρχουν δυο τρόποι υπολογισμού:

1) Μπορεί να ακολουθηθεί η διαδικασία του Discretize σε κάθε συνεχές χαρακτηριστικό και στη συνέχεια να αντικατασταθεί η αξία του χαρακτηριστικού με το αντίστοιχο διακριτό διάστημα. Η προσέγγιση αυτή μετατρέπει τα συνεχή χαρακτηριστικά σε ordinal χαρακτηριστικά. Η πιθανότητα $P(x_i \mid Y = y)$ υπολογίζεται από το κλάσμα των χαρακτηριστικών εκπαίδευσης που ανήκουν στην κλάση y και τα οποία πέφτουν εντός του αντίστοιχου διαστήματος x_i. Το σφάλμα της εκτίμησης προκύπτει ανάλογα με τη διαδικασία του discretize που ακολουθήθηκε και ανάλογα με τον αριθμό των διακριτών διαστημάτων. Εάν ο αριθμός των διαστημάτων είναι πολύ μεγάλος, τότε θα υπάρχουν πολύ λίγα παραδείγματα εκπαίδευσης σε κάθε διάστημα έτσι ώστε να δώσουν μια αξιόπιστη εκτίμηση για το $P(x_i \mid y)$. Από την άλλη πλευρά, αν ο αριθμός των διαστημάτων είναι πολύ μικρός, τότε κάποια διαστήματα ενδέχεται να συναθροίσουν παραδείγματα διαφορετικών κλάσεων γεγονός που μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα.

2) Μπορεί να υποτεθεί μια συγκεκριμένη κατανομή της πιθανότητας για τη συνεχή μεταβλητή και στη συνέχεια να εκτιμηθούν οι παράμετροι της κατανομής χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης. Η Γκαουσσιανή κατανομή επιλέγεται συνήθως να αντιπροσωπεύσει την πιθανότητα της κλάσης για συνεχή χαρακτηριστικά. Η κατανομή χαρακτηρίζεται από δυο παραμέτρους, τον μέσο μ, και τη διακύμανση σ^2. Για κάθε κλάση y_i, η πιθανότητα για το χαρακτηριστικό x_i να ανήκει στην κλάση προκύπτει από:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\Pi} \cdot \sigma_{ij}} \cdot \exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Η παράμετρος σ_{ij} υπολογίζεται από τον μέσο των $X_i(\overline{x})$ για όλα τα παραδείγματα εκπαίδευσης που ανήκουν στην κλάση y_j. Ομοίως, το σ_{ij}^2 μπορεί να εκτιμηθεί από τη διακύμανση S^2 των ίδιων παραδειγμάτων εκπαίδευσης.

3.2.1.2 Πλεονεκτήματα και Μειονεκτήματα των Απλοϊκών Κατηγοριοποιητών Bayes.

Οι ΑΚΒ σε γενικές γραμμές διέπονται από τα ακόλουθα χαρακτηριστικά:

1) Λειτουργούν σχετικά καλά ακόμα και αν το σύνολο εκπαίδευσης είναι μικρό.

2) Είναι καλοί στο να διαχειρίζονται απομονωμένα σημεία θορύβου καθώς τα σημεία αυτά υπολογίζονται κατά μέσο όρο όταν εκτιμώνται οι υπό όρους πιθανότητες των δεδομένων. Επίσης

έχουν τη δυνατότητα να διαχειρίζονται χαμένες αξίες αγνοώντας αυτές κατά τη φάση κατασκευής του μοντέλου και της κατηγοριοποίησης.

3) Είναι καλοί στο να διαχειρίζονται χαρακτηριστικά άσχετα. Εάν το x_i είναι ένα άσχετο χαρακτηριστικό, τότε το $P(x_i \mid y)$ κατανέμεται ενιαία. Η πιθανότητα της κλάσης για το x_i δεν επιφέρει καμία επίπτωση στο συνολικό υπολογισμό της μεταγενέστερης πιθανότητας.

4) Τα στενά συσχετισμένα χαρακτηριστικά μειώνουν αρκετά την απόδοση του ΑΚΒ καθώς η υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών πλέον δεν βρίσκει εφαρμογή για τα συγκεκριμένα χαρακτηριστικά.

3.3 Επαγωγή κανόνα (Rule Induction)

Η Επαγωγή Κανόνα αποτελεί μια από τις σημαντικότερες τεχνικές της Μηχανικής Μάθησης και αφού οι κανονικότητες που βρίσκονται κρυμμένες μέσα στα δεδομένα εκφράζονται συνήθως με τη μορφή κανόνων, η Επαγωγή Κανόνα αποτελεί και ένα από τα θεμελιώδη εργαλεία του Data Mining.

Οι κανόνες εκφράζονται συνήθως με τη μορφή:

If(attribute-1,value-1) and (attribute-2,value-2) and …. and (attribute-n, value-n) then (decision, value)

Κάποια συστήματα επαγωγής κανόνων δημιουργούν πιο περίπλοκους κανόνες στους οποίους οι αξίες των χαρακτηριστικών μπορούν να εκφραστούν με την άρνηση (απαγωγή) κάποιων άλλων αξιών ή από κάποιο υποσύνολο αξιών των χαρακτηριστικών του πεδίου.

Τα δεδομένα από τα οποία επάγονται οι κανόνες παρουσιάζονται συνήθως με μορφή παρόμοια με πίνακα στον οποίο τα παραδείγματα είναι ετικέτες (ή ονόματα) για τις σειρές, ενώ οι μεταβλητές επιγράφονται ως χαρακτηριστικά και σαν απόφαση. Στο παρόν κεφάλαιο, η προσοχή θα στραφεί στο κομμάτι της επαγωγής κανόνων της εποπτευόμενης μάθησης (supervised learning), δηλαδή όλες οι περιπτώσεις είναι κατηγοριοποιημένες από πριν. Με άλλα λόγια, η αξία απόφασης έχει αποφασιστεί από πριν για κάθε περίπτωση. Τα χαρακτηριστικά είναι ανεξάρτητες μεταβλητές και η απόφαση είναι η εξαρτημένη μεταβλητή.

Ένα τέτοιο απλό παράδειγμα περιγράφεται από τον ακόλουθο πίνακα:

Πίνακας 1

	Θερμοκρασία	**Πονοκέφαλος**	**Αδυναμία**	**Ναυτία**	**Πυρετός**
1	Πολύ υψηλή	Ναι	Ναι	Όχι	Ναι
2	υψηλή	Ναι	Όχι	Ναι	Ναι
3	Κανονική	Όχι	Όχι	Όχι	Όχι

4	Κανονική	Ναι	Ναι	Ναι	Ναι
5	Υψηλή	Όχι	Ναι	Όχι	Ναι
6	Υψηλή	Όχι	Όχι	Όχι	Όχι
7	Κανονική	Όχι	Ναι	Όχι	Όχι

Όπου τα χαρακτηριστικά είναι: Θερμοκρασία, Πονοκέφαλος, Αδυναμία, Ναυτία και η αξία απόφασης είναι ο Πυρετός.

Το σύνολο όλων των περιπτώσεων που χαρακτηρίζονται από την ίδια αξία απόφασης καλούνται concept (έννοια). Για παράδειγμα, στον παραπάνω πίνακα το σύνολο των περιπτώσεων (1,2,3,4,5) είναι ένα concept για όλες εκείνες τις περιπτώσεις που η αξία απόφασης είναι Πυρετός.

Σε αυτό το σημείο κρίνεται σκόπιμο να επισημανθεί ότι το σύνολο των δεδομένων εισόδου μπορεί να επηρεάζεται από κάποιο σφάλμα. Για παράδειγμα στο ακόλουθο πίνακα η περίπτωση νούμερο 7 έχει αξία για την μεταβλητή Αδυναμία ίση με 42.5, ένα προφανές σφάλμα αφού η συγκεκριμένη μεταβλητή παίρνει μόνο συμβολικές αξίες ίσες με το Ναι ή το Όχι. Τέτοια σφάλματα στα δεδομένα πρέπει να διορθωθούν πριν την Επαγωγή Κανόνα.

Πίνακας 2

	Θερμοκρασία	**Πονοκέφαλος**	**Αδυναμία**	**Ναυτία**	**Πυρετός**
1	Πολύ υψηλή	Ναι	Ναι	Όχι	Ναι
2	υψηλή	Ναι	Όχι	Ναι	Ναι
3	Κανονική	Όχι	Όχι	Όχι	Όχι
4	Κανονική	Ναι	Ναι	Ναι	Ναι
5	Υψηλή	Όχι	Ναι	Όχι	Ναι
6	Υψηλή	Όχι	Όχι	Όχι	Όχι
7	Κανονική	Όχι	42.5	Όχι	Όχι

Ένα άλλο πρόβλημα δημιουργείται από την ύπαρξη αριθμητικών χαρακτηριστικών. Για παράδειγμα, η Θερμοκρασία μπορεί να παριστάνεται από πραγματικά νούμερα όπως στον ακόλουθο πίνακα.

Πίνακας 3

	Θερμοκρασία	**Πονοκέφαλος**	**Αδυναμία**	**Ναυτία**	**Πυρετός**
1	39.8	Ναι	Ναι	Όχι	Ναι
2	36.6	Ναι	Όχι	Ναι	Ναι
3	38.9	Όχι	Όχι	Όχι	Όχι
4	40.3	Ναι	Ναι	Ναι	Ναι
5	37.7	Όχι	Ναι	Όχι	Ναι
6	39.9	Όχι	Όχι	Όχι	Όχι
7	37.9	Όχι	Ναι	Όχι	Όχι

Είναι δεδομένο ότι οι αριθμητικές αξίες πρέπει να μετασχηματιστούν σε συμβολικές αξίες πριν ή κατά τη διάρκεια της Επαγωγής Κανόνα. Η διαδικασία μετατροπής αριθμητικών αξιών σε συμβολικές ονομάζεται διακριτοποίηση (discretization) ή κβαντισμός (quantization).

Συμπληρωματικά, τα δεδομένα εισόδου μπορεί να είναι ατελή, δηλαδή να υπάρχουν χαμένες αξίες στα χαρακτηριστικά (βλέπε Πίνακα 4) ή να υπάρχει ασυνέπεια στα δεδομένα (βλέπε Πίνακα 5), δηλαδή να υπάρχουν αντιφατικά παραδείγματα όπου να έχουν τις ίδιες αξίες χαρακτηριστικών αλλά να παρουσιάζουν διαφορετικές αξίες απόφασης, όπως φαίνεται για τις περιπτώσεις 7 και 8 του Πίνακα 5.

Πίνακας 4

	Θερμοκρασία	**Πονοκέφαλος**	**Αδυναμία**	**Ναυτία**	**Πυρετός**
1	39.8	Ναι	Ναι	Όχι	Ναι
2	36.6	Ναι	Όχι	Ναι	Ναι
3	?	Όχι	Όχι	Όχι	Όχι
4	40.3	Ναι	Ναι	Ναι	Ναι
5	37.7	Όχι	?	Όχι	Ναι
6	39.9	?	Όχι	Όχι	Όχι
7	37.9	Όχι	Ναι	Όχι	Όχι

Πίνακας 5

	Θερμοκρασία	**Πονοκέφαλος**	**Αδυναμία**	**Ναυτία**	**Πυρετός**
1	39.8	Ναι	Ναι	Όχι	Ναι
2	36.6	Ναι	Όχι	Ναι	Ναι
3	38.9	Όχι	Όχι	Όχι	Όχι
4	40.3	Ναι	Ναι	Ναι	Ναι
5	37.7	Όχι	Ναι	Όχι	Ναι
6	39.9	Όχι	Όχι	Όχι	Όχι
7	37.9	Όχι	Ναι	Όχι	Ναι

3.3.1 Τύποι Κανόνων

Μια περίπτωση χ καλύπτεται από έναν κανόνα r όταν και μόνο όταν κάθε κατάσταση (condition) – ζεύγος χαρακτηριστικό-αξία – του r ικανοποιείται από την αντίστοιχη αξία του χαρακτηριστικού χ. Το concept C χαρακτηρίζεται από τη δεξιά πλευρά του κανόνα r. Λέμε ότι ένα concept C έχει καλυφθεί τελείως από ένα σύνολο κανόνων R όταν και μόνο όταν για κάθε περίπτωση χ του C υπάρχει ένας κανόνας r από το R, έτσι ώστε το r να καλύπτει το χ. Ένα σύνολο κανόνων R είναι ολοκληρωμένο όταν και μόνο όταν κάθε concept του συνόλου δεδομένων καλύπτεται πλήρως από το R.

Ένας κανόνας r είναι συνεπής με το σύνολο των δεδομένων όταν και μόνο όταν για κάθε περίπτωση χ που καλύπτεται από το r, το χ είναι μέλος του concept C που υποδηλώνεται από το r. Τέλος, ένα σύνολο κανόνων R είναι συνεπές όταν και μόνο όταν για κάθε κανόνα του R, το R είναι συνεπές με το σύνολο δεδομένων. Για παράδειγμα, η περίπτωση 1 του Πίνακα 1 καλύπτεται από τον ακόλουθο κανόνα:

$$(\Pi ονοκέφαλος, \Nαι) \rightarrow (\Pi υρετός, \Nαι)$$

Ο κανόνας r υποδηλώνει το concept (1,2,4,5). Συμπληρωματικά, το concept (1,2,4,5) δεν καλύπτεται πλήρως από ένα σύνολο κανόνων αποτελούμενο από το r, αφού το r καλύπτει μόνο τις περιπτώσεις 1,2 και 4, αλλά ο κανόνας r είναι συνεπής με το σύνολο δεδομένων του Πίνακα 1.

Από την άλλη, ο νόμος:

$$(\Pi ονοκέφαλος, Όχι) \rightarrow (\Pi υρετός, Όχι)$$

Καλύπτει πλήρως το concept (3,6,7) του Πίνακα 1, αν και ο συγκεκριμένος κανόνας δεν είναι συνεπής. Ο παραπάνω κανόνας καλύπτει τις περιπτώσεις 3,5,6 και 7. Καθένας από τους ακόλουθους κανόνες:

$$(\Pi ονοκέφαλος, \Nαι) \& (\Α δυναμία, \Nαι) \rightarrow (\Pi υρετός, \Nαι)$$

Και

$$(\Θ ερμοκρασία, \Υ ψηλή) \& (\Pi ονοκέφαλος, \Nαι) \rightarrow (\Pi υρετός, \Nαι)$$

Είναι συνεπής με το σύνολο δεδομένων του Πίνακα 1 αλλά το concept (1,2,4,5) δεν καλύπτεται πλήρως από το σύνολο κανόνων που αποτελείται από τους δυο παραπάνω κανόνες αφού η περίπτωση 5 δεν καλύπτεται από κανένα κανόνα. Ο πρώτος κανόνας καλύπτει τις περιπτώσεις 1 και 4, ενώ ο δεύτερος καλύπτει την περίπτωση 2.

Το πιο σύνηθες έργο της τεχνικής της Επαγωγής Κανόνα είναι η δημιουργία ενός συνόλου κανόνων R το οποίο είναι συνεπές και ολοκληρωμένο. Ένα τέτοιο σύνολο κανόνων R καλείται διακριτικό (discriminant). Ένα τέτοιο σύνολο για τα δεδομένα του Πίνακα 1 είναι το ακόλουθο το οποίο απαρτίζεται από 4 κανόνες και είναι συνεπές και ολοκληρωμένο.

$$(\Pi ονοκέφαλος, \Nαι) \rightarrow (\Pi υρετός, \Nαι)$$
$$(\Θ ερμοκρασία, \Υ ψηλή) \& (\Α δυναμία, \Nαι) \rightarrow (\Pi υρετός, \Nαι)$$
$$(\Θ ερμοκρασία, \Κ ανονική) \& (\Pi ονοκέφαλος, \Οχι) \rightarrow (\Pi υρετός, \Οχι)$$
$$(\Pi ονοκέφαλος, \Οχι) \& (\Α δυναμία, \Οχι) \rightarrow (\Pi υρετός, \Οχι)$$

Υπάρχουν πολλοί άλλοι τύποι κανόνων που χρησιμοποιούνται. Για παράδειγμα, κάποια συστήματα Επαγωγής Κανόνα δημιουργούν κανόνες που αποτελούνται από ισχυρούς κανόνες, δηλαδή σύνολα κανόνων όπου κάθε κανόνας καλύπτει πολλές περιπτώσεις. Ένα άλλο έργο είναι η επαγωγή προσεταιριστικών κανόνων, όπου και στις δυο πλευρές του κανόνα, αριστερά και δεξιά, οι εμπλεκόμενες μεταβλητές είναι χαρακτηριστικά. Ένα τέτοιο παράδειγμα αποτελεί ο ακόλουθος κανόνας:

$$(\Ν αυτία, \Nαι) \rightarrow (\Pi ονοκέφαλος, \Nαι)$$

3.3.2 Αλγόριθμοι Επαγωγής Κανόνων

Στο παρόν κεφάλαιο θα ακολουθήσει η ανάλυση ενός αλγόριθμου Επαγωγής Κανόνα. Σε γενικές γραμμές οι αλγόριθμοι Επαγωγής Κανόνα μπορούν να χαρακτηριστούν ως Καθολικοί (Global) ή Τοπικοί (Local). Στους καθολικούς αλγόριθμους Επαγωγής Κανόνα ο χώρος αναζήτησης είναι το σύνολο όλων των αξιών των χαρακτηριστικών, ενώ στους τοπικούς αλγόριθμους Επαγωγής Κανόνα ο χώρος αναζήτησης είναι το σύνολο των ζευγαριών χαρακτηριστικά-αξίες. Οι αλγόριθμος που ακολουθεί παράγει διακριτικά σύνολα κανόνων, και είναι καθολικός. Επίσης, στην ανάλυση θεωρείται ότι τα δεδομένα εισόδου δεν πλήττονται από σφάλματα, οι αριθμητικές αξίες έχουν ήδη διακριτοποιηθεί, τα δεδομένα είναι συνεπή και δεν υπάρχουν χαμένες αξίες.

3.3.2.1 Αλγόριθμος LEM1

Ο αλγόριθμος LEM1, είναι μέρος του συστήματος εξόρυξης γνώσης Lers (Learning from Examples using Rough Sets) και στηρίζεται σε κάποια ακατέργαστα σύνολα ορισμών (Pawlak,1991)(Pawlak et. Al,1995). Εάν υποτεθεί ότι το B είναι ένα μή-άδειο υποσύνολο του συνόλου όλων των χαρακτηριστικών A. Το U υποδηλώνει το σύνολο όλων των περιπτώσεων. Η αναντικατάστατη σχέση (indiscernibility relation) $IND(B)$ αναφορικά με το U ορίζεται για κάθε $x, y \in U$ με το $(x, y) \in IND(B)$ όταν και μόνο όταν και για τα δυο χ και y οι αξίες όλων των χαρακτηριστικών από το B είναι πανομοιότυπες.

Η αναντικατάστατη σχέση $IND(B)$ είναι σχέση ισοδυναμίας. Οι ισοδύναμες κλάσεις του $IND(B)$ καλούνται στοιχειώδη σύνολα του B. Για παράδειγμα, για τον Πίνακα 1 και B = {Θερμοκρασία, Πονοκέφαλος}, τα στοιχειώδη σύνολα του $IND(B)$ είναι τα {1}, {2}, {3,7}, {4}, {5,6}. Η οικογένεια όλων των στοιχειωδών συνόλων του B θα επισημαίνεται με B*, για παράδειγμα από τον Πίνακα 1:

{Θερμοκρασία, Πονοκέφαλος}* = { {1}, {2}, {3,7}, {4}, {5,6} }.

Για μια απόφαση d, λέμε ότι το {d} εξαρτάται από το B όταν και μόνο όταν το $B^* \le \{d\}^*$. Μια καθολική κάλυψη του {d} είναι ένα υποσύνολο B του A έτσι ώστε το {d} να εξαρτάται από το B και το B είναι ελάχιστο στο A. Έτσι, οι καθολικές καλύψεις του {d} υπολογίζονται με τη σύγκριση χωρισμάτων (partitions) του B* με το {d}*.

Στη βάση της καθολικής κάλυψης, οι κανόνες υπολογίζονται χρησιμοποιώντας την τεχνική της απόρριψης συνθηκών (Michalski,1983). Για έναν κανόνα της μορφής :

$$C_1 \& C_2 \& \ldots \& C_n \rightarrow D$$

Απόρριψη συνθηκών σημαίνει το σκανάρισμα της λίστας όλων των συνθηκών, από τα αριστερά προς τα δεξιά, με την προσπάθεια της απόρριψης κάθε συνθήκης, ελέγχοντας σύμφωνα με τον Πίνακα απόφασης όπου ο απλοποιημένος κανόνας δεν καταπατά τη συνοχή της διακριτικής περιγραφής. Για παράδειγμα, για τον Πίνακα 1 η διαδικασία γίνεται ως εξής:

$$\{\Theta\varepsilon\rho\mu οκρασία, \Pi ονοκέφαλος, \Alpha δυναμία, \Nαυτία\}^* = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\}\}$$

$$\{\Pi υρετός\}^* \{\{1,2,4,5\},\{3,6,7\}\}$$

και

$$\{\Theta\varepsilon\rho\mu οκρασία, \Pi ονοκέφαλος, \Alpha δυναμία, \Nαυτια\}^* \le \{\Pi υρετός\}^*$$

Κατόπιν, πρέπει να ελεγχτεί το κατά πόσο ισχύει ότι:

$$\{\Pi ονοκέφαλος, \Alpha δυναμία, \Nαυτία\}^* \le \{\Pi υρετός\}^*$$

Αυτή η συνθήκη είναι λανθασμένη αφού:

$$\{\Pi ονοκέφαλος, \Alpha δυναμία, \Nαυτία\}^* = \{\{1\},\{2\},\{3,6\},\{4\},\{5,7\}\}$$

Στη συνέχεια υπολογίζεται:

$$\{\Theta\varepsilon\rho\mu οκρασία, \Alpha δυναμία, \Nαυτία\}^* = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\}\}$$

Παρατηρείται ότι:

$$\{\Theta\varepsilon\rho\mu οκρασία, \Alpha δυναμία, \Nαυτία\}^* \le \{\Pi υρετός\}^*$$

Το επόμενο χώρισμα του συνόλου που υπολογίζεται είναι το:

$$\{\Theta\varepsilon\rho\mu οκρασία, \Nαυτία\}^* = \{\{1\},\{2\},\{3,7\},\{4\},\{5,6\}\}$$

Και

$$\{\Theta\varepsilon\rho\mu οκρασία, \Nαυτία\}^* \ge \{\Pi υρετός\}^*.$$

Το τελευταίο βήμα είναι να υπολογιστεί το:

$$\{\Theta\varepsilon\rho\mu οκρασία, \Alpha δυναμία\} = \{\{1\},\{2,6\},\{3\},\{4,7\},\{5\}\}$$

Αφού $\{\Theta\varepsilon\rho\mu οκρασία, \Alpha δυναμία\}^* \ge \{\Pi υρετός\}^*$, έτσι η συνολική κάλυψη είναι $\{\Theta\varepsilon\rho\mu οκρασία, \Alpha δυναμία, \Nαυτία\}$.

Η πρώτη περίπτωση του Πίνακα 1 αφήνει να εννοηθεί ο ακόλουθος πρωταρχικός κανόνας:

$$(\Theta\varepsilon\rho\mu οκρασία, \Pi ολύ\_\Upsilon ψηλή) \& (\Alpha δυναμία, \Nαι) \& (\Nαυτία, \Oχι) \to (\Pi υρετός, \Nαι)$$

Ο παραπάνω κανόνας καλύπτει μόνο την 1η περίπτωση. Η 1η συνθήκη

$(\Theta\varepsilon\rho\mu οκρασία, \Pi ολύ\_υψηλή)$ δεν μπορεί να απορριφθεί αφού ο νόμος:

$$(\Alpha δυναμία, \Nαι) \& (\Nαυτία, \Oχι) \to (\Pi υρετός, \Nαι)$$

Καλύπτει τις περιπτώσεις 1 και 7 που ανήκουν σε διαφορετικά concept. Παρόλα αυτά, μια προσπάθεια να απορριφθεί η επόμενη συνθήκη, (Αδυναμία, Ναι), κρίνεται επιτυχής αφού ο κανόνας:

$\{\Theta\varepsilon\rho\mu οκρασία, \Pi ολύ\_υψηλή\} \& \{\Nαυτία, \Οχι\} \to \{\Π υρετος, \Ναι\}$ καλύπτει μόνο την περίπτωση 1.

Η επόμενη πιθανότητα, να απορριφθεί η τελευταία συνθήκη (Αδυναμία, Ναι) είναι και αυτή επιτυχής, αφού ο κανόνας που δημιουργείται

$(\Theta\varepsilon\rho\mu οκρασία, \Π ολυ\_υψηλή) \to (\Π υρετός, \Ναι)$ καλύπτει μόνο την περίπτωση 1.

Με έναν παρόμοιο τρόπο δημιουργούνται και οι υπόλοιποι κανόνες. Το τελικό σύνολο κανόνων που παράγεται από τον αλγόριθμο LEM1, είναι:

$(\Theta\varepsilon\rho\mu οκρασία, \Π ολύ\_υψηλή) \to (\Π υρετός, \Ναι)$

$(\Ναυτία, \Ναι) \to (\Π υρετός, \Ναι)$

$(\Theta\varepsilon\rho\mu οκρασία, \Υ ψηλή) \& (\Αδυναμία, \Ναι) \to (\Π υρετός, \Ναι)$

$(\Αδυναμία, \Οχι) \& (\Ναυτία, \Οχι) \to (\Π υρετός, \Οχι)$

$(\Theta\varepsilon\rho\mu οκρασία, \Κανονική) \& (\Ναυτία, \Οχι) \to (\Π υρετός, \Οχι)$

3.3.3 Συστήματα Κατηγοριοποίησης

Τα σύνολα κανόνων που δημιουργούνται από σύνολα δεδομένων, χρησιμοποιούνται κυρίως για την κατηγοριοποίηση νέων περιπτώσεων που δεν έχουμε ξαναδεί. Στο συγκεκριμένο κεφάλαιο θα γίνει μια ανάλυση ενός συστήματος κατηγοριοποίησης που σχετίζεται με το LERS. Στο συγκεκριμένο σύστημα, η απόφαση του σε ποιο concept ανήκει μια περίπτωση γίνεται στη βάση τριών παραγόντων:

- Δύναμη (strength)
- Εξειδίκευση (specificity)
- Υποστήριξη (support)

Οι παράγοντες αυτοί χαρακτηρίζονται ως εξής:

- Δύναμη είναι ο συνολικός αριθμός περιπτώσεων που έχουν κατηγοριοποιηθεί σωστά με τον κανόνα κατά τη διάρκεια της εκπαίδευσης.
- Εξειδίκευση είναι ο συνολικός αριθμός των ζευγαριών χαρακτηριστικό-αξία στην αριστερή πλευρά του κανόνα. Οι κανόνες που ταιριάζουν με έναν μεγαλύτερο αριθμό ζευγαριών χαρακτηριστικού-αξίας, θεωρούνται πιο εξειδικευμένοι.
- Υποστήριξη είναι το άθροισμα όλων των προϊόντων της Δύναμης και της Εξειδίκευσης για όλους τους κανόνες που ταιριάζουν- αντιστοιχίζονται και αναφέρονται στο ίδιο concept. Το concept C για το οποίο η Υποστήριξη, δηλαδή η ακόλουθη έκφραση:

$$\sum_{matching\_rules\_r\_describing\_C} strength(r) * specificity(r)$$

Παίρνει τη μεγαλύτερη τιμή, είναι ο νικητής και η περίπτωση κατηγοριοποιείται γίνοντας μέρος του C.

Στο σύστημα κατηγοριοποίησης LERS, αν το πλήρες αντιστοίχισμα είναι αδύνατο, τότε αναγνωρίζονται όλοι οι μερικώς αντιστοιχισμένοι κανόνες. Αυτοί είναι κανόνες με τουλάχιστον ένα ζεύγος χαρακτηριστικού-αξίας που ταιριάζει στο αντίστοιχο ζεύγος χαρακτηριστικού-αξίας της περίπτωσης. Για οποιοδήποτε μερικώς αντιστοιχισμένο κανόνα r, ο επιπλέον παράγοντας που καλείται Παράγοντας

Αντιστοίχισης (Matching Factor) r , υπολογίζεται. Ο Παράγοντας Αντιστοίχισης r είναι η αναλογία του αριθμού των αντιστοιχισμένων ζευγών χαρακτηριστικού-αξίας του r με μια περίπτωση, προς το συνολικό αριθμό των ζευγαριών χαρακτηριστικού-αξίας του r.

Κατά τη μερική αντιστοίχιση, το Concept C για το οποίο η ακόλουθη έκφραση είναι μεγαλύτερη, είναι ο νικητής και η περίπτωση κατηγοριοποιείται ως μέρος του C.

$$\sum_{partially\_matching\_rules\_r\_describing\_C} Matching - Factor(r) * Strength(r) * Specificity(r)$$

3.3.4 Επικύρωση

Το πιο σημαντικό κριτήριο απόδοσης στις μεθόδους Επαγωγής Κανόνα είναι ο ρυθμός σφάλματος (error rate). Εάν ο αριθμός των περιπτώσεων είναι μικρότερος του 100, η μέθοδος 'άφησε-ένα-έξω' (leave-one-out) χρησιμοποιείται για να εκτιμήσει το ρυθμό σφάλματος του συνόλου των κανόνων. Στη μέθοδο leaving-one-out, ο αριθμός των πειραμάτων εκπαίδευσης και ελέγχου είναι ίσος με τον αριθμό των περιπτώσεων του συνόλου δεδομένων. Κατά τη διάρκεια του i πειράματος, η i περίπτωση αφαιρείται από το σύνολο των δεδομένων, ένα σύνολο κανόνων δημιουργείται από το σύστημα Επαγωγής Κανόνα των υπόλοιπων περιπτώσεων και η κατηγοριοποίηση της παραληφθείσας περίπτωσης, με τους νόμους που έχουν επαχθεί καταγράφεται. Το ποσοστό σφάλματος υπολογίζεται ως εξής:

$$\frac{\text{Αριθμός\_λανθασμένων\_κατηγοριοποιήσεων}}{\text{Αριθμός\_Περιπτώσεων}}$$

Από την άλλη μεριά, εάν ο αριθμός των περιπτώσεων του συνόλου δεδομένων είναι μεγαλύτερος ή ίσος με 100, χρησιμοποιείται ten-fold-cross-validation, τεχνική επικύρωσης που έχει αναπτυχθεί σε προηγούμενο κεφάλαιο.

3.4 Νευρωνικά Δίκτυα (Neural Networks)

Τα Νευρωνικά Δίκτυα ή τα Τεχνητά Νευρωνικά Δίκτυα αποτελούν ένα σημαντικό εργαλείο για την ποσοτική μοντελοποίηση (Quantitative Modelling). Είναι αρκετά δημοφιλή στο χώρο του Data Mining καθώς έχουν χρησιμοποιηθεί επιτυχώς στο παρελθόν για την επίλυση πληθώρας προβλημάτων σε όλα τα πεδία όπως επιχειρήσεις, βιομηχανία και επιστήμη. Σήμερα, θεωρούνται σαν ένα από τα κλασικά και πιο διαδεδομένα εργαλεία στο χώρο του Data Mining και χρησιμοποιούνται για πολλές εργασίες όπως της κατηγοριοποίησης μοτίβων, της ανάλυσης χρονοσειρών, της πρόβλεψης και της συσταδοποίησης.

Τα Νευρωνικά Δίκτυα είναι υπολογιστικά μοντέλα για την επεξεργασία της πληροφορίας και είναι ιδιαιτέρως χρήσιμα για την αναγνώριση θεμελιωδών σχέσεων μεταξύ των μεταβλητών ενός συνόλου δεδομένων. Δημιουργήθηκαν μέσω της έρευνας στον τομέα της Τεχνητής Νοημοσύνης και πιο συγκεκριμένα στην προσπάθεια μίμησης του τρόπου που μαθαίνουν τα βιολογικά Νευρωνικά Δίκτυα και ιδίως του ανθρώπινου εγκεφάλου, ο οποίος περιέχει πάνω από 10^{11} διασυνδεδεμένους νευρώνες.

Παρόλο που τα Τεχνητά Νευρωνικά Δίκτυα που θα αναλυθούν στο παρόν κεφάλαιο είναι εξαιρετικά πιο απλά και πιο μικρά σε σχέση με τα βιολογικά συστήματα αναφορικά με το μέγεθος, την ικανότητα και την ισχύ, μοιράζονται δυο πολύ σημαντικά χαρακτηριστικά: την παράλληλη επεξεργασία της πληροφορίας και τη μάθηση και γενίκευση μέσω εμπειρίας.

Τα Νευρωνικά Δίκτυα οφείλουν τη δημοτικότητά τους στην ισχυρή τους ικανότητα στην αναγνώριση προτύπων-μοτίβων. Ενώ αρκετά σημαντικά χαρακτηριστικά που διαθέτουν τα καθιστούν κατάλληλα και χρήσιμα στο χώρο του Data Mining. Καταρχάς, σε αντίθεση με τις παραδοσιακές μεθόδους στηριζόμενες σε μοντέλα, τα Νευρωνικά Δίκτυα δεν απαιτούν εκ των προτέρων υποθέσεις για τη διαδικασία δημιουργίας των δεδομένων ούτε συγκεκριμένες δομές μοντέλων. Αντ' αυτού, η διαδικασία μοντελοποίησης είναι εξαιρετικά προσαρμοστική και το μοντέλο προσδιορίζεται κυρίως από τα χαρακτηριστικά ή μοτίβα τα οποία το δίκτυο αναγνώρισε από τα δεδομένα κατά τη διαδικασία της εκμάθησης. Αυτή η οδηγούμενη από τα δεδομένα προσέγγιση είναι ιδανική για την αντιμετώπιση πραγματικών προβλημάτων στο χώρο του Data Mining όπου τα δεδομένα είναι αρκετά αλλά τα μοτίβα που έχουν σημασία ή οι δομές των δεδομένων δεν έχουν ανακαλυφθεί ακόμα και είναι αδύνατο να προκαθοριστούν.

Κατά δεύτερον, η μαθηματική ιδιότητα της καθολικής προσέγγισης είναι εξαιρετικά ισχυρή καθώς υποδεικνύει ότι τα Νευρωνικά Δίκτυα είναι πολύ πιο γενικά και ευέλικτα στη μοντελοποίηση σε σχέση με τις παραδοσιακές μορφές. Αφού πολλές από τις εγασίες του Data Mining όπως η αναγνώριση μοτίβων, η κατηγοριοποίηση και η πρόβλεψη μπορούν να αντιμπετωπιστούν ως προσεγγιστικά προβλήματα, όπου ακριβής αναγνώριση των βαθύτερων δομών και σχέσεων που κρύβονται στα δεδομένα κρίνεται ως ιδιαιτέρως σημαντική.

Τρίτον, τα Νευρωνικά Δίκτυα είναι μη γραμμικά μοντέλα. Καθώς τα δεδομένα του πραγματικού κόσμου και οι σχέσεις που διέπουν αυτά είναι μη γραμμικές, τα παραδοσιακά γραμμικά εργαλεία μπορεί να πάσχουν από σημαντικές προκαταλήψεις, γεγονός που καθιστά τα Νευρωνικά Δίκτυα με τη μη γραμμική και μη παραμετρική φύση τους ιδανικά εργαλεία για τη μοντελοποίηση πολύπλοκων προβλημάτων.

Τέλος, τα Νευρωνικά Δίκτυα είναι σε θέση να επιλύσουν προβλήματα που έχουν μη ακριβή μοτίβα ή δεδομένα και εμπεριέχουν θορυβώδη πληροφορία με έναν μεγάλο αριθμό μεταβλητών. Αυτή η ανοχή τους στο σφάλμα τα καθιστά ελκυστικά για την επίλυση πραγματικών προβλημάτων όπου συνήθως τα δεδομένα είναι 'βρώμικα' και δεν ακολουθούν ξεκάθαρες δομές πιθανοτήτων, που τυπικά απαιτούνται από τα παραδοσιακά στατιστικά μοντέλα.

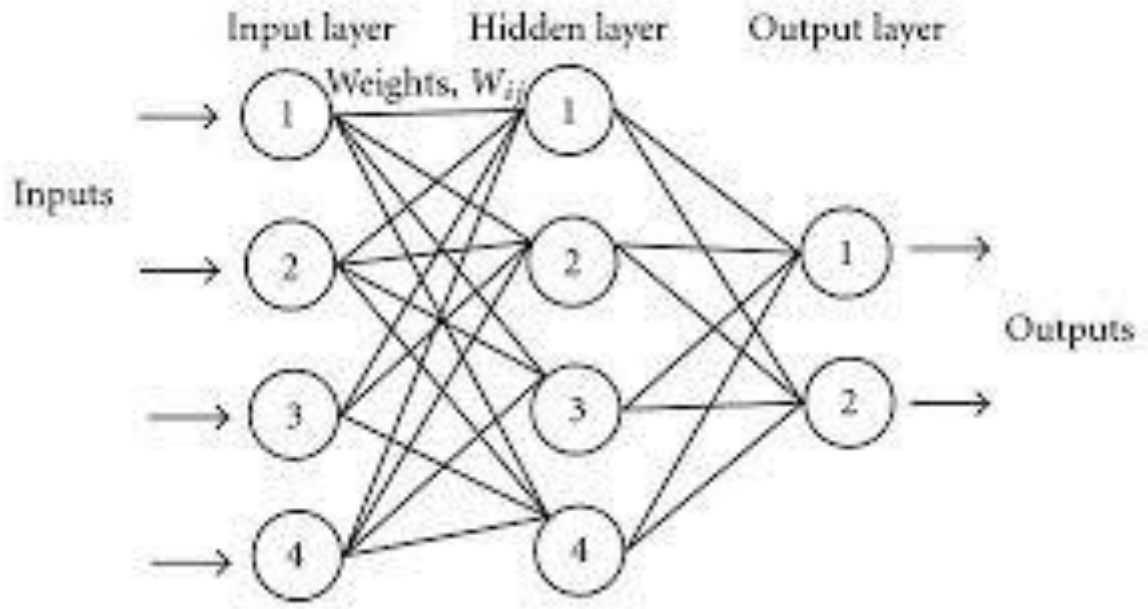
3.4.1 Μοντέλα Νευρωνικών Δικτύων – Feed-Forward Neural Networks

Στο κεφάλαιο που ακολουθεί θα γίνει μια ανάλυση του τρόπου λειτουργίας και της δομής του Νευρωνικού Δικτύου Feed-Forward. Τα multilayer feed-forward Νευρωνικά Δίκτυα καλούνται επίσης και Multi-layer Perceptorns (MLP), είναι τα πιο μελετημένα και χρησιμοποιημένα Νευρωνικά Δίκτυα στην πράξη. Σύμφωνα με τους Wong, Bodnovich και Selvi (1997) περίπου το 95% των εφαρμογών Νευρωνικών Δικτύων που συναντώνται στη βιβλιογραφία χρησιμοποιούν αυτόν τον τύπο Νευρωνικού Δικτύου, το Mylti-Layer Perceptron. Τα συγκεκριμένα Νευρωνικά Δίκτυα είναι ιδανικά για τη μοντελοποίηση σχέσεων μεταξύ ζεύγων μεταβλητών εισόδου και εξόδου. Με άλλα λόγια, είναι κατάλληλα για λειτουργικά προβλήματα χαρτογράφησης όπου θέλουμε να γνωρίζουμε πως ένας αριθμός μεταβλητών

εισόδου επιδρά στις μεταβλητές εξόδου. Αφού οι περισσότερες εργασίες πρόβλεψης και κατηγοριοποίησης μπορούν να αντιμετωπιστούν ως λειτουργικά προβλήματα χαρτογράφησης (functional mapping problems), τα δίκτυα MLP είναι ιδιαιτέρως ελκυστικά για το Data Mining.

3.4.1.1 Δομή του Μοντέλου

Ένα MLP είναι ένα δίκτυο αποτελούμενο από έναν αριθμό υψηλά διασυνδεδεμένων απλών υπολογιστικών μονάδων που καλούνται νευρώνες, κόμβοι ή κύτταρα, τα οποία οργανώνονται σε στρώσεις-στρώματα. Κάθε νευρώνας πραγματοποιεί μια απλή εργασία επεξεργασίας της πληροφορίας με το να μετατρέπει τις εισερχόμενες εισόδους σε επεξεργασμένες εξόδους. Μέσω των τόξων σύνδεσης μεταξύ των νευρώνων, η γνώση μπορεί να παραχθεί και να αποθηκευτεί σαν βάρη των τόξων, σε σχέση με το πόσο ισχυρή είναι η σχέση μεταξύ διαφορετικών κόμβων. Παρόλο που κάθε νευρώνας υλοποιεί τη λειτουργία του με αργό ρυθμό και ατελώς, όντας σε συνεργασία το Νευρωνικό Δίκτυο είναι ικανό να εκτελεί ποικιλία εργασιών με αποδοτικό τρόπο και να επιτυγχάνει θαυμαστά αποτελέσματα.



Το παραπάνω σχήμα απεικονίζει την αρχιτεκτονική ενός τέτοιου Fedd-froward Νευρωνικού Δικτύου τριών-στρωμάτων το οποίο αποτελείται από νευρώνες (κύκλους) οργανωμένους σε τρία στρώματα: στρώμα εισόδου, κρυφό στρώμα και στρώμα εξόδου.

Οι νευρώνες στους κόμβους εισόδου ανταποκρίνονται στις ανεξάρτητες ή προβλεπτικές μεταβλητές οι οποίες πιστεύεται ότι είναι χρήσιμες για την πρόβλεψη των εξαρτημένων μεταβλητών που ανταποκρίνονται στους νευρώνες εξόδου. Έτσι, οι νευρώνες του στρώματος εισόδου είναι παθητικοί. Δεν επεξεργάζονται την πληροφορία αλλά χρησιμοποιούνται απλώς για να δεχθούν τα μοτίβα των

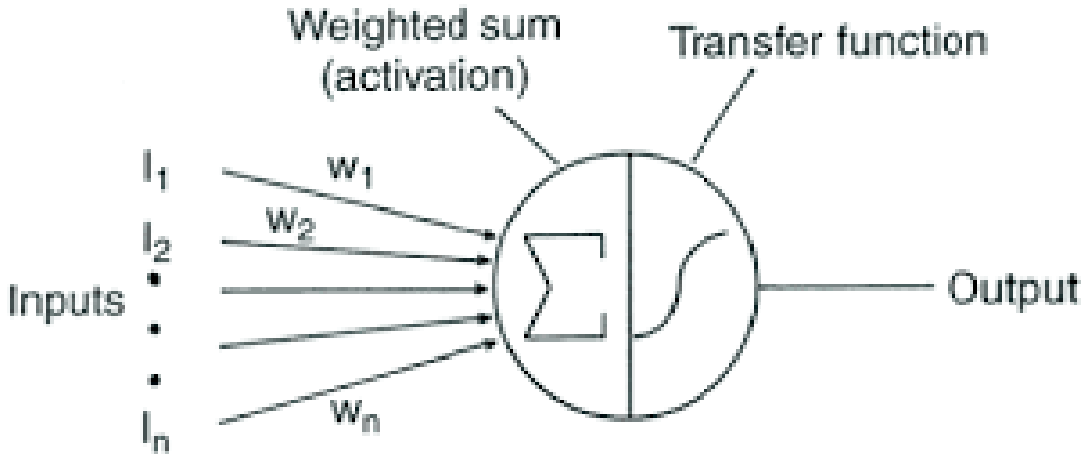
δεδομένων και να τα περάσουν στους νευρώνες του επόμενου στρώματος (layer). Οι νευρώνες του κρυφού στρώματος συνδέονται με τους νευρώνες εισόδου και εξόδου και αποτελούν συστατικό κλειδί για την εκμάθηση των μοτίβων που κρύβονται στα δεδομένα και για τη χαρτογράφηση των σχέσεων μεταξύ των μεταβλητών εισόδου και εξόδου.

Να αναφερθεί εδώ ότι, παρόλο που είναι εφικτό να υπάρχουν περισσότερα του ενός κρυφά στρώματα σε Νευρωνικά Δίκτυα πολλαπλών στρωμάτων, στην πράξη οι περισσότερες εφαρμογές χρησιμοποιούν μόνο ένα κρυφό στρώμα.

Με μη γραμμικές λειτουργίες μεταφοράς, οι κρυφοί νευρώνες μπορούν να επεξεργαστούν την πολύπλοκη πληροφορία που λαμβάνουν από τους νευρώνες εισόδου και στη συνέχεια να μεταφέρουν την επεξεργασμένη πληροφορία στο στρώμα εξόδου για περαιτέρω επεξεργασία και παραγωγή εξόδων. Στα feed-forward Νευρωνικά Δίκτυα, η ροή της πληροφορίας είναι μονοκατευθυνόμενη από την είσοδο, στο κρυφό στρώμα και στη συνέχεια στο στρώμα εξόδου, χωρίς να υπάρχει κάποια ανατροφοδότηση από την έξοδο. Έτσι ένα τέτοιο δίκτυο χαρακτηρίζεται από την αρχιτεκτονική του η οποία προσδιορίζεται από τον αριθμό των στρωμάτων, τον αριθμό των κόμβων σε κάθε στρώμα, τη λειτουργία μεταφοράς (transfer function) που χρησιμοποιείται σε κάθε στρώμα καθώς και από τον τρόπο που οι κόμβοι σε κάθε στρώμα ενώνονται με κόμβους γειτονικών στρωμάτων.

Παρόλο που η μερική συνδεσιμότητα μεταξύ κόμβων εισόδου και εξόδου είναι εφικτή, το πιο σύνηθες Νευρωνικό Δίκτυο είναι το πλήρες διασυνδεδεμένο υπο την έννοια ότι κάθε κόμβος του κάθε στρώματος είναι πλήρως διασυνδεδεμένος μόνο με όλους τους κόμβους των γειτονικών στρωμάτων.

Για να γίνει κατανοητός ο τρόπος λειτουργίας του παραπάνω σχήματος, πρέπει πρώτα να κατανοηθεί ο τρόπος με τον οποίο οι νευρώνες του κρυφού στρώματος και του στρώματος εξόδου επεξεργάζονται την πληροφορία. Το ακόλουθο σχήμα παρέχει έναν μηχανισμό που δείχνει πως ένας νευρώνας επεξεργάζεται την πληροφορία που προέρχεται από διάφορες εισόδους και την μετατρέπει, αυτή την πληροφορία, σε έξοδο.



Κάθε νευρώνας επεξεργάζεται την πληροφορία σε δυο βήματα. Στο πρώτο βήμα, οι είσοδοι (Xi) συνδυάζονται μαζί στη δημιουργία ενός σταθμισμένου αθροίσματος των εισόδων και των βαρών (Wi)

των διασυνδεδεμένων συνδέσμων. Το δεύτερο βήμα πραγματοποιεί μια μεταμόρφωση η οποία μετατρέπει το άθροισμα αυτό σε μια έξοδο μέσω της λειτουργίας μεταφοράς. Με άλλα λόγια, ο νευρώνας του σχήματος πραγματοποιεί τις ακόλουθες λειτουργίες:

$$OUT_n = f(\sum_i w_i x_i)$$

Όπου το OUT n είναι η μορφή της εξόδου του συγκεκριμένου νευρώνα και η f είναι η συνάρτηση μεταφοράς. Σε γενικές γραμμές, η συνάρτηση μεταφοράς είναι μια δεσμευμένη μη μειούμενη συνάρτηση. Παρόλο που υπάρχουν πολλές πιθανές επιλογές για συναρτήσεις μεταφοράς, μόνο μετικές από αυτές χρησιμοποιούνται στην πράξη, όπως:

- Η σιγμοειδής (λογιστική) συνάρτηση: $f(x) = (1+\exp(-x))^{-1}$
- Η υπερβολική εφαπτομένη συνάρτηση: $f(x) = \dfrac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$
- Η ημιτονοειδής συνάρτηση: $f(x) = \sin(x)$
- Η συνιμητονειδής συνάρτηση: $f(x) = \cos(x)$
- Η γραμμική ή ταυτοτική συνάρτηση: $f(x) = x$

Μεταξύ αυτών, η λογιστική συνάρτηση αποτελεί την πιο δημοφιλή επιλογή ειδικά για κόμβους του κρυφού στρώματος, εξαιτίας του γεγονότος ότι είναι απλή, έχει έναν αριθμό από καλά χαρακτηριστικά και φέρει καλύτερη ομοιότητα με τους πραγματικούς νευρώνες.

Τώρα, αν υποτεθεί ότι το $x = (x_1, x_2, ..., x_d)$ είναι ένα διάνυσμα d προβλεπτικών μεταβλητών, το $y = (y_1, y_2, ..., y_m)$ είναι m-διαστάσεων διάνυσμα εξόδου του δικτύου και w_1, w_2 οι μήτρες των βαρών των διασυνδεδεμένων τόξων από την είσοδο, στο κρυφό στρώμα και στην έξοδο αντίστοιχα. Τότε ένα Νευρωνικό Δίκτυο τριών στρωμάτων μπορεί να γραφτεί σαν ένα μη γραμμικό μοντέλο της μορφής:

$$y = f_2(w_2 f_1(w_1 x)),$$

Όπου τα f_1, f_2 είναι οι συναρτήσεις μεταφοράς για τους κρυφούς κόμβους και τους κόμβους εξόδου αντίστοιχα.

Πολλά δίκτυα περιλαμβάνουν επίσης κόμβους προκατάληψης (bias) οι οποίοι είναι σταθερές που προστίθενται στους κρυφούς ή τους κόμβους εξόδου για να ενισχύσουν την ευελιξία της μοντελοποίησης του Νευρωνικού Δικτύου. Η έννοια της προκατάληψης εδώ, λειτουργεί σαν την έννοια της αναχαίτισης (intercept) στη γραμμική παλινδρόμηση.

Σε προβλήματα κατηγοριοποίησης όπου οι επιθυμητές έξοδοι είναι δυαδικές ή κατηγορικές, η λογιστική συνάρτηση χρησιμοποιείται συχνά για τον περιορισμό του εύρους των εξόδων του δικτύου. Από την άλλη μεριά, για λόγους πρόβλεψης, αφού οι μεταβλητές εξόδου είναι σε γενικές γραμμές συνεχείς, η γραμμική συνάρτηση μεταφοράς αποτελεί καλύτερη επιλογή για τους κόμβους εξόδου. Τελικά, η εξίσωση μπορεί να έχει πολλά και διαφορετικά χαρακτηριστικά εξαρτώμενα από το είδος του προβλήματος, τη συνάρτηση μεταφοράς και τον αριθμό των κόμβων που χρησιμοποιήθηκαν στα στρώματα εισόδου, εξόδου και στο κρυφό.,

Για παράδειγμα, η δομή του Νευρωνικού δικτύου για ένα γενικό πολυμεταβλητό πρόβλημα πρόβλεψης με χρήση της λογιστικής συνάρτησης στους κρυφούς κόμβους και της ταυτοτικής συνάρτησης για τον κόμβο εξόδου μπορεί να εκφραστεί ως:

$$y_t = w_{10} + \sum_{j=1}^{q} w_{1j} f(\sum_{i=1}^{p} w_{ij} x_{it} + w_{oj})$$

Όπου το y_t είναι η παρατήρηση της μεταβλητής πρόβλεψης και $\{x_{it}, i = 1, 2, ..., p\}$ είναι οι p μεταβλητές πρόβλεψης για το χρόνο t, το p είναι επίσης ο αριθμός των κόμβων εισόδου, το q είναι ο αριθμός των κρυφών κόμβων και $\{w_{ij}, i = 0, 1, ..., p \,\&\, j = 1, 2, ..., q\}$ είναι τα βάρη από την είσοδο προς τους κρυφούς κόμβους, ενώ τα w_{10} και w_{0j} είναι οι όροι προκατάληψης και f είναι η λογιστική συνάρτηση.

3.4.1.2 Η εκπαίδευση του Δικτύου

Τα βάρη των τόξων (arc weights) αποτελούν τις παραμέτρους σε ένα μοντέλο Νευρωνικού Δικτύου. Όπως και σε ένα στατιστικό μοντέλο, αυτές οι παράμετροι χρειάζεται να εκτιμηθούν πριν το δίκτυο χρησιμοποιηθεί μελλοντικά. Η εκπαίδευση του Νευρωνικού Δικτύου αναφέρεται στη διαδικασία με την οποία τα βάρη αυτά καθορίζονται και ουσιαστικά αυτός είναι και ο τρόπος με τον οποίο το δίκτυο μαθαίνει. Η εκπαίδευση του δικτύου για προβλήματα κατηγοριοποίησης και πρόβλεψης πραγματοποιείται μέσω εποπτευόμενης μάθησης όπου γνωστές έξοδοι και οι συσχετιζόμενες με αυτές είσοδοι παρουσιάζονται στο δίκτυο.

Παρακάτω θα ακολουθήσει η βασική διαδικασία εκπαίδευσης ενός Νευρωνικού Δικτύου. Καταρχάς, το δίκτυο 'ταΐζεται' με παραδείγματα εκπαίδευσης τα οποία αποτελούνται από ένα σετ μοτίβων εισόδου και από τις επιθυμητές εξόδους αυτών. Στη συνέχεια, για κάθε μοτίβο εκπαίδευσης, σταθμίζονται και αθροίζονται οι αξίες εισόδου σε κάθε κρυφό κόμβο και το σταθμισμένο άθροισμα μετά μεταδίδεται μέσω μιας κατάλληλης συνάρτησης μεταφοράς στην αξία εξόδου του κρυφού κόμβου, όπου και γίνεται είσοδος για τους κόμβους εξόδου. Μετά, οι αξίες εξόδου που έχει παράγει το δίκτυο υπολογίζονται και συγκρίνονται με τις επιθυμητές αξίες προκειμένου να διευκρινιστεί το κατά πόσο κοντά είναι οι αξλιες του δικτύου με τις επιθυμητές. Η διαδικασία αυτή, τυπικά επαναλαμβάνεται πολλές φορές μέχρι οι αξίες εξόδου του δικτύου να είναι όσο το δυνατόν πιο κοντά στις επιθυμητές.

Για τη διευκόλυνση της εκπαίδευσης χρησιμοποιούνται κάποια συνολικά μέτρα σφάλματος όπως το μέσο τετραγωνικό σφάλμα (MSE) ή το άθροισμα των τετραγωνικών σφαλμάτων (SSE). Για παράδειγμα, στο συγκεκριμένο Νευρωνικό Δίκτυο το MSE μπορεί να καθοριστεί ως:

$$MSE = \frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{j=1}^{N}(d_{mj} - y_{mj})^2$$

Όπου το d_{mj}, y_{mj} εκπροσωπούν την επιθυμητή αξία και την έξοδο του δικτύου στον M-οστό κόμβο για το j-οστό μοτίβο εκπαίδευσης αντίστοιχα, το M είναι ο αριθμός των κόμβων εξόδου και N ο αριθμός των μοτίβων εκπαίδευσης. Ο στόχος της εκπαίδευσης είναι η εύρεση του κατάλληλου σετ βαρών τα οποία ελαχιστοποιούν την αντικειμενική συνάρτηση. Έτσι, η εκπαίδευση του δικτύου αποτελεί ουσιαστικά ένα

αβίαστο, μη γραμμικό πρόβλημα βελτιστοποίησης όπου χρειάζονται αριθμητικές μέθοδοι για την επίλυσή του.

Η πιο σημαντική και δημοφιλής μέθοδος εκπαίδευσης είναι ο αλγόριθμος της πίσω-διάδοσης (back propagation) ο οποίος ουσιαστικά αποτελεί μια μέθοδο της πιο απότομης κλίσης (gradient steepest method). Ο τρόπος λειτουργίας της μεθόδου αυτής είναι η εύρεση της καλύτερης κατεύθυνσης σε έναν πολύ-διάστατο χώρο σφάλματος και η μετακίνηση ή αλλαγή των βαρών που ελαχιστοποιούν περισσότερο την αντικειμενική συνάρτηση. Αυτό απαιτεί την μερική εξαγωγή(παραγώγιση) της αντικειμενικής συνάρτησης με σεβασμό σε κάθε βάρος που υπολογίζεται, γιατί η μερική παραγώγιση αντιπροσωπεύει το ρυθμό αλλαγής της αντικειμενικής συνάρτησης. Ο εκσυγχρονισμός των βαρών θα ακολουθεί τον παρακάτω κανόνα:

$$w_{ij}^{new} = w_{ij}^{old} + \Delta w_{ij}$$

$$\Delta w_{ij} = -n\frac{\partial \mathrm{E}}{\partial w_{ij}}$$

Όπου το Δw_{ij} αποτελεί τη βαθμίδα της αντικειμενικής συνάρτησης E με σεβασμό στο βάρος w_{ij}, ενώ το n καλείται ρυθμός εκμάθησης και πρακτικά ελέγχει το μέγεθος της κλίσης. Ο αλγόριθμος απαιτεί μια επαναλαμβανόμενη διαδικασία και υπάρχουν δυο εκδοχές του τρόπου εκσυγχρονισμού των βαρών: η batch κα η on-line. Στην batch εκδοχή, τα βάρη εκσυγχρονίζονται αφού προηγουμένως όλα τα μοτίβα εκπαίδευσης έχουν αξιολογηθεί, ενώ στην on-line εκδοχή, τα βάρη εκσυγχρονίζονται διαρκώς μετά από κάθε αναγνώριση μοτίβου. Τα βασικά βήματα της batch εκδοχής εκπαίδευσης συνοψίζονται ως εξής:

1. Δίνονται στα βάρη μικρές τυχαίες αξίες που προέρχονται από μια ενιαία κατανομή.

2. Επιλέγεται μοτίβο και διαδίδεται προς τα εμπρός προκειμένου να αποκτηθούν μερικές εξαγωγές του σφάλματος με σεβασμό στα βάρη.

3. Αθροίζονται όλα τα μονά μοτίβα.

4. Εκσυγχρονίζονται τα βάρη.

5. Επαναλαμβάνονται τα βήματα 2-5 για κάθε μοτίβο μέχρι όλα τα μοτίβα να έχουν περάσει από μέσα.

Να σημειωθεί εδώ ότι κάθε πέρασμα όλων των μοτίβων καλείται εποχή (epoch). Σε γενικές γραμμές, η εκσυγχρόνιση-ανανέωση των βαρών μειώνει το ολικό σφάλμα κατά πολύ λίγο οπότε αρκετές εποχές χρειάζονται προκειμένου αυτό να ελαχιστοποιηθεί.

3.4.2 Γνωστά θέματα με τη μοντελοποίηση ενός Νευρωνικού Δικτύου

Η ανάπτυξη ενός μοντέλου Νευρωνικού Δικτύου για μια Data Mining εφαρμογή δεν είναι εύκολο έργο. Παρόλο που υπάρχουν πολλά software πακέτα τα οποία διευκολύνουν τους χρήστες, είναι κριτικής σημασίας οι χρήστες να κατανοήσουν πλήρως πολλά σημαντικά ζητήματα γύρω από τη διαδικασία μοντελοποίησης, αφού αυτή είναι ένας συνδυασμός επιστήμης και τέχνης. Ένα σημαντικό θέμα είναι η

καλή κατανόηση της εκπαίδευσης και της γεωίκευσης που υπάρχει έμφυτη σε όλες τις εφαρμογές ενός τέτοιου δικτύου. Τα ζητήματα της εκπαίδευσης και της γενίκευσης μπορούν να κατανοηθούν μέσω των εννοιών της προκατάληψης του Μοντέλου και της διακύμανσης.

Η προκατάληψη του μοντέλου μετράει το συστηματικό σφάλμα ενός μοντέλου στο να μαθαίνει τις σχέσεις που υποβόσκουν μεταξύ των μεταβλητών ή των παρατηρήσεων. Η διακύμανση του μοντ΄λου σχετίζεται με τη σταθερότητα ενός μοντέλου που έχει δημιουργηθεί από διαφορετικά δείγματα δεδομένων και συνεπώς παρέχει πληροφορίες για την ικανότητα γενίκευσης του μοντέλου. Ένα προκαθορισμένο ή παραμετρικό μοντέλο το οποί είναι λίγο εξαρτώμενο από τα δεδομένα, μπορεί να αντιπροσωπεύει λάθος την πραγματική λειτουργική σχέση και έτσι να οδηγήσει σε μια μεγάλη προκατάληψη. Από την άλλη πλευρά, ένα ευέλικτο μοντέλο, οδηγούμενο από τα δεδομένα μπορεί να είναι υπερβολικά εξαρτημένο από το συγκεκριμένο σύνολο δεδομένων με αποτέλεσμα να παρουσιάζει μεγάλη διακύμανση. Η προκατάληψη και η διακύμανση είναι δυο πολύ σημαντικές έννοιες που επηρεάζουν έντονα τη χρησιμότητα του μοντέλου.

Ένα μοντέλο το οποίο είναι λιγότερο εξαρτώμενο από τα δεδομένα έχει την τάση να παρουσιάζει μικρότερη διακύμανση αλλά υψηλή προκατάληψη στην περίπτωση που το προκαθορισμένο μοντέλο είναι λανθασμένο. Από την άλλη, ένα μοντέλο που εφαρμόζει καλά στα δεδομένα έχει την τάση να έχει μικρή προκατάληψη αλλά μεγάλη διακύμανση όταν εφαρμόζεται σε νέα σύνολα δεδομένων. Συνεπώς ένα καλό προβλεπτικό μοντέλο πρέπει να βρίσκεται σε μια ισορροπία μεταξύ προκατάληψης και διακύμανσης.

Ένα άλλο σημαντικό ζήτημα είναι οι επιλογές αναφορικά με τον σχεδιασμό και την αρχιτεκτονική του Νευρωνικού Δικτύου. Όχι μόνο δεν υπάρχουν πολλοί τρόποι που μπορεί κανείς να χτίσει ένα Νευρωνικό Δίκτυο και πολλές επιλογές κατά τον σχεδιασμό του, αλλά υπάρχουν και πληθώρα παραμέτρων που πρέπει να εκτιμηθούν και να πειραματιστεί κανείς μέχρι να φτάσει στο επιθυμητό αποτέλεσμα.

Στην όλη δυσκολία προσθέτει και το γεγονός ότι δεν υπάρχουν παγιωμένες κατευθύνσεις για τη δημιουργία. Υπάρχουν πολλοί κανόνες διαθέσιμοι, οι οποίοι όμως δεν μπορούν να χρησιμοποιηθούν τυφλά σε οποιαδήποτε περίπτωση. Αναφορικά με την επιλογή της αρχιτεκτονικής υυπάρχουν πολλές αποφάσεις που πρέπει να παρθούν. Πρώτον, το μέγεθος του στρώματος εξόδου καθορίζεται συνήθως από τη φύση του προβλήματος. Για παράδειγμα, στα περισσότερα προβλήματα ανάλυσης χρονοσειρών, ένας κόμβος εξόδου χρησιμοποιείται συνήθως για την πρόβλεψη. Από την άλλη μεριά, για προβλήματα κατηγοριοποίησης, ο αριθμός των κόμβων εξόδου καθορίζεται από τον αριθμό των ομάδων στις οποίες επιθυμούμε να κατηγοριοποιήσουμε τα αντικείμενα.

Ο αριθμός των κόμβων εισόδου είναι ίσως η πιο σημαντική παράμετρος για ένα αποτελεσματικό Νευρωνικό Δίκτυο. Σε προβλήματα κατηγοριοποίησης, το νούνερο αυτό αντιστοιχεί στον αριθμό των μεταβλητών που θεωρούνται σημαντικές για την πρόβλεψη. Για πολυμεταβλητά προβλήματα χρονοσειρών είναι το νούμερο των παλαιών καθυστερημένων παρατηρήσεων. Ο καθορισμός του κατάλληλου συνόλου μεταβλητών εισόδου είναι ζωτικής σημασίας για ένα Νευρωνικό Δίκτυο για να κατανοηθούν οι βασικές σχέσεις που μπορεί να χρησιμοποιηθούν για μια επιτυχημένη πρόβλεψη.

Τώρα, όσο αφορά την επιλογή μοντέλου, αυτή γίνεται συνήθως μ΄σω της διαδικασίας του cross-validation. Τα δεδομένα χωρίζονται σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο επικύρωσης. Οι παράμετροι του δικτύου εκτιμώνται με το σύνολο εκπαίδευσης ενω η απόδοση του μοντέλου παρακολουθείται και εκτιμάται από το σύνολο επικύρωσης. Κατόπιν επιλέγεται ως καλύτερο εκείνο το

μοντέλο που έχει την καλύτερη απόδοση. Μετά τη διαδικασία μοντελοποίησης, το επιλεχθέν μοντέλο πρέπει να αξιολογηθεί σε ένα καινούριο σύνολο δεδομένων. Συμπληρωματικά, αφού τα Νευρωνικά Δίκτυα συχνά χρησιμοποιούνται ως μη γραμμικές εναλλακτικές λύσεις έναντι των παραδοσιακών στατιστικών μοντέλων, η απόδοσή τους πρέπει να συγκριθεί και με εκείνες των στατιστικών μεθόδων. Σε αυτό το σημείο κρίνεται σκόπιμο να επισημανθεί ότι αν έχει χρησιμοποιηθεί η μέθοδος του cross-validation για την επιλογή μοντέλου και τον πειραματισμό τότε, η απόδοση στο σύνολο επικύρωσης δεν πρέπει να ιδωθεί ότι απεικονίζει την πραγματική απόδοση του μοντέλου.

3.4.3 Εφαρμογές των Νευρωνικών Δικτύων στο Data Mining

Τα Νευρωνικά Δίκτυα έχουν χρησιμοποιηθεί εκτενώς στο χώρο του Data Mining για πληθώρα προβλημάτων σε ένα μεγάλο αριθμό πεδίων από τις επιχειρήσεις, τη μηχανολογία, τη βιομηχανία, την ιατρική και τις επιστήμες γενικότερα. Σε γενικές γραμμές τα Νευρωνικά Δίκτυα είναι πολύ καλά στο να λύνουν προβλήματα του Data Mining όπως η κατηγοριοποίηση, η πρόβλεψη και η Συσταδοποίηση.

Η κατηγοριοποίηση αποτελεί ένα από τα πιο συχνά αντιμετωπιζόμενα προβλήματα. Ένα πρόβλημα κατηγοριοποίησης λαμβάνει χώρα όταν ένα αντικείμενο χρειάζεται να ανατεθεί σε ένα προκαθορισμένο γκρουπ ή κλάση βάση του αριθμού των παρατηρούμενων χαρακτηριστικών που σχετίζονται με αυτό. Πολλά προβλήματα στο χώρο των επιχειρήσεων, της βιομηχανίας και της ιατρικής μπορούν να αντιμετωπιστούν ως προβλήματα κατηγοριοποίησης. Τέτοια παραδείγματα περιλαμβάνουν την αναγνώριση απάτης, τις ιατρικές διαγνώσεις, τον έλεχο της ποιότητας, αναγνώριση προτύπων χειρόγραφων και αναγνώριση φωνής. Τα feed-forward Νευρωνικά Δίκτυα πολλαπλών επιπέδων είναι αυτά που χρησιμοποιούνται κυρίως για τις εργασίες αυτές.

Η πρόβλεψη είναι κεντρική για τον αποτελεσματικό σχεδιασμό και τις λειτουργίες σε όλες τις επιχειρήσεις και τους οργανισμούς καθώς και στον χώρο των κρατικών υπηρεσιών. Η ικανότητα της ακριβής πρόβλεψης του μέλλοντος είναι θεμελιώδης για πολλές αποφάσεις και διαδικασίες στα οικονομικά, στη διαφήμιση και στην παραγωγή. Η βελτίωση της ακρίβειας της πρόβλεψης θα μπορούσε να έχει ως αποτέλεσμα σημαντικές αυξήσεις κερδών σε μια επιχείρηση. Η πρόβλεψη μπορεί να γίνει με δυο προσεγγίσεις: με αιτιώδη συνάφεια και με ανάλυση χρονοσειρών, όπου και οι δυο προσεγγίσεις είναι κατάλληλες για ένα Νευρωνικό Δίκτυο. Επιτυχείς εφαρμογές περιλαμβάνουν την πρόβλεψη πωλήσεων, τον αριθμό επιβατών, το μερίδιο της αγοράς, τον ρυθμό ανταλλαγής, τη μελλοντική τοποθέτηση των τιμών, την ανταπόδοση των μετοχών, τη ζήτηση ηλεκτρικού ρεύματος, τις περιβαλλοντικές αλλαγές και την κίνηση στους δρόμους.

Η συσταδοποίηση περιλαμβάνει την κατηγοριοποίηση ή την κατάτμηση παρατηρήσεων σε ομάδες ή συστάδες, έτσι ώστε κάθε συστάδα να είναι όσο το δυνατόν πιο ομογενής. Σε αντίθεση με τα προβλήματα κατηγοριοποίησης, οι ομάδες δεν είναι γνωστές από πριν ούτε προκαθορίζονται. Η συσταδοποίηση μπορεί να απλοποιήσει ένα πολύπλοκο και μεγάλο σύνολο δεδομένων σε μικρότερες ομάδες βάση της φυσικής δομής των δεδομένων. Η καλύτερη κατανόηση τηων δεδομένω και συνεπώς οι καλύτερες αποφάσεις αποτελούν μεγάλα οφέλη της συσταδοποίησης. Τα Νευρωνικά Δίκτυα και στο πεδίο αυτό αποτελούν ένα κατάλληλο εργαλείο. Εφαρμογές τέτοιας χρήσης αποτελούν μεταξύ άλλων, η κατάτμηση της αγοράς, η στόχευση πελατών, η αξιολόγηση πιστοληπτικής ικανότητας και η ανάκτηση εγγράφων.

Τα Νευρωνικά δίκτυα έχουν ήδη καταφέρει να σημειώσουν σημαντική πρόοδο και επιτυχία στο χώρο του Data Mining. Παρόλα αυτά, κρίνετε σκόπιμο να υπογραμμιστεί ότι έχουν και αρκετούς περιορισμούς και ότι δεν αποτελούν πανάκεια για κάθε πρόβλημα. Η χρήση τους απαιτεί εξαιρετική κατανόηση των δεδομένων, συνετή στρατηγική μοντελοποίησης και μεγάλη προσοχή στα ζητήματα μοντελοποίησης που μπορεί να εγερθούν.

3.5 Συσταδοποίηση (Clustering)

Η συσταδοποίηση και η κατηγοριοποίηση αποτελούν θεμελιώδεις εργασίες στο χώρο του DataMining. Η κατηγοριοποίηση χρησιμοποιείται συνήθως ως μέθοδος μάθησης με εποπτεία ενώ η συσταδοποίηση σαν μέθοδος μάθησης χωρίς εποπτεία, αν και υπάρχουν κάποιες τεχνικές συσταδοποίησης που χρησιμοποιούνται και για τα δυο. Ο στόχος της συσταδοποίησης είναι περιγραφικός ενώ της κατηγοριοποίησης προβλεπτικός. Αφού ο στόχος της συσταδοποίησης είναι η ανακάλυψη νέων ομάδων κατηγοριών, τα νέα γκρουπ αποτελούν πηγή ενδιαφέροντος από μόνα τους και η αξιολόγησή τους είναι εσωτερική (intrinsic). Από την άλλη πλευρά, στις εργασίες κατηγοριοποίησης ένα σημαντικό μέρος της αξιολόγησης είναι εξωτερικό (extrinsic) αφού οι ομάδες πρέπει να αντικατοπτρίζουν χαρακτηριστικά κάποιων ήδη υπάρχουσων κλάσεων.

Η συσταδοποίηση ομαδοποιεί παραδείγματα δεδομένων σε υποσύνολα με τέτοιο τρόπο έτσι ώστε παραδείγματα που ομαδοποιούνται μαζί να έχουν κοινές ομοιότητες, ενώ παραδείγματα που ομαδοποιούνται ξεχωριστά να διαφέρουν. Έτσι, τα παραδείγματα οργανώνονται σε μια αποτελεσματική παρουσίαση η οποία εκφράζει τον πληθυσμό από τον οποίο έχει γίνει η δειγματοληψία.

Τυπικά, η δομή της συσταδοποίησης παρουσιάζεται σαν ένα σετ από υποσύνολα: $C = c_1, \ldots, c_k$ του S, έτσι ώστε να ισχύει:

$$S = U_{i=1}^{k} C_i$$
$$C_i \cap C_j = \varnothing \quad \text{για κάθε } i \neq j$$

Συνεπώς, κάθε παράδειγμα του S ανήκει σε ακριβώς ένα και μόνο ένα υποσύνολο.

3.5.1 Μέτρα απόστασης

Αφού η συσταδοποίηση είναι η ομαδοποίηση ομοίων χαρακτηριστικών ή αντικειμένων, χρειάζεται κάποιο μέτρο το οποίο να καθορίζει αν τα δυο αντικείμενα είναι όμοια ή ανόμοια. Υπάρχουν δυο τύποι μέτρων που χρησιμοποιούνται για να εκτιμήσουν αυτή τη σχέση: τα μέτρα απόστασης και τα μέτρα ομοιότητας.

Πολλές μέθοδοι συσταδοποίησης χρησιμοποιούν μέτρα απόστασης προκειμένου να καθορίσουν την ομοιότητα ή την ανομοιότητα μεταξύ ενός ζεύγους αντικειμένων. Έτσι, είναι χρήσιμο να δηλώνεται η απόσταση μεταξύ δυο παραδειγμάτων x_i, x_j σαν $d(x_i, x_j)$. Ένα έγκυρο μέτρο απόστασης θα πρέπει να είναι συμμετρικό και να λαμβάνει την ελάχιστη αξία του (συνήθως το μηδέν) στην περίπτωση

πανομοιότυπων διανυσμάτων. Συμπληρωματικά, το μέτρο της απόστασης καλείται μετρικό εάν ικανοποιεί τις ακόλουθες ιδιότητες:

1. Τριγωνική ανισότητα: $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k) \forall x_i, x_j, x_k \in S$

2. $d(x_i, x_j) = 0 \Rightarrow x_i = x_j, \forall x_i, x_j \in S$

Παρακάτω θα ακολουθήσει μια ανάλυση των σημαντικότερων και πιο δημοφιλών μέτρων απόστασης.

3.5.1.1 Μέτρο απόστασης Minkowski για αριθμητικά χαρακτηριστικά

Δοθέντος δυο p-διάστατων παραδειγμάτων $x_j = (x_{j1}, x_{j2}, ...x_{jp}), x_i = (x_{i1,} x_{i2}, ..., x_{ip})$, η απόσταση μεταξύ τους μπορεί να υπολογιστεί χρησιμοποιώντας τη μετρική Minkowski ως εξής:

$$d(x_i, x_j) = (| x_{i1} - x_{j1} |^g + | x_{i2} - x_{j2} |^g + ... + | x_{ip} - x_{jp} |^g)^{1/g}$$

Η πιο συνήθης Ευκλείδεια απόσταση μεταξύ δυο αντικειμένων επιτυγχάνεται όταν g=2.

Το μέτρο της μονάδας που χρησιμοποιείται μπορεί να επηρεάσει την ανάλυση της συσταδοποίησης. Για να αποφευχθεί αυτή η εξάρτηση στην επιλογή μέτρων (measurement units), τα δεδομένα θα πρέπει να τυποποιηθούν-κανονικοποιηθούν. Τα κανονικοποιημένα μέτρα προσπαθούν να δώσουν σε όλες τις μεταβλητές ένα ίσο βάρος. Παρόλα αυτά, εάν σε κάθε μεταβλητή ανατίθεται ένα βάρος σύμφωνα με τη σημαντικότητά της, η σταθμισμένη απόσταση υπολογίζεται ως:

$$d(x_i, x_j) = (w_1 | x_{i1} - x_{j1} |^g + w_2 | x_{i2} - x_{j2} |^g + ... + w_p | x_{ip} - x_{jp} |^g)^{1/g}$$

όπου $w_i \in [0, \infty)$.

3.5.1.2 Μέτρο απόστασης για δυαδικά χαρακτηριστικά

Στην περίπτωση των δυαδικών χαρακτηριστικών, η απόσταση μεταξύ των αντικειμένων μπορεί να υπολογιστεί βάση ενός πίνακα ενδεχομένων (contingencytable). Ένα δυαδικό χαρακτηριστικό είναι συμμετρικό αν και οι δυο καταστάσεις του είναι εξίσου πολύτιμες. Στην περίπτωση αυτή, χρησιμοποιώντας έναν απλό συντελεστή αντιστοίχισης μπορεί να εκτιμηθεί η ανομοιότητα μεταξύ δυο αντικειμένων:

$$d(x_i, x_j) = \frac{r + s}{q + r + s + t}$$

Όπου:

- **q** είναι ο αριθμός των χαρακτηριστικών που ισούται με τη μονάδα και για τα δυο αντικείμενα.
- **t** είναι ο αριθμός των χαρακτηριστικών που ισούται με το μηδέν και για τα δυο αντικείμενα.
- **s** και rείναι ο αριθμός των χαρακτηριστικών που είναι άνισα και για τα δυο αντικείμενα.

Ένα δυαδικό χαρακτηριστικό είναι ασύμμετρο όταν και οι δυο καταστάσεις του δεν είναι εξίσου σημαντικές. Σε αυτή την περίπτωση, αγνοούνται οι (t) αρνητικές αντιστοιχίσεις στον παρανομαστή. Η σχέση που ακολουθεί λέγεται συντελεστής Jaccard:

$$d(x_i, x_j) = \frac{r+s}{q+r+s}$$

3.5.1.3 Μέτρο απόστασης για ονομαστικά (nominal) χαρακτηριστικά

Όταν τα χαρακτηριστικά είναι ονομαστικά, χρησιμοποιούνται δυο προσεγγίσεις:

1. Απλή αντιστοίχιση $d(x_i, x_j) = \frac{q-m}{p}$, όπου ρείναι ο συνολικός αριθμός των χαρακτηριστικών και m είναι ο αριθμός των αντιστοιχίσεων.

2. Δημιουργείται ένα δυαδικό χαρακτηριστικό για κάθε κατάσταση του λάθε ονομαστικού χαρακτηριστικού και υπολογίζεται η ανομοιότητα όπως πριν.

3.5.1.4 Μέτρο απόστασης για τακτικά (ordinal) χαρακτηριστικά

Αυτό που αλλάζει στην προκειμένη περίπτωση είναι ότι η αλληλουχία των αξιών έχει νόημα. Στην περίπτωση αυτή, τα χαρακτηριστικά μπορούν να αντιμετωπιστούν σαν αριθμητικά αφού προηγουμένως αντιστοιχιστεί το εύρος των αξιών τους σε [0,1]. Τέτοιά αντιστοίχιση γίνεται ως εξής:

$$Z_{i,n} = \frac{r_{i,n} - 1}{M_n - 1}$$

Όπου:

- Το $Z_{i,n}$ είναι η κανονικοποιημένη αξία του χαρακτηριστικού a_n του αντικειμένου i.
- Το $r_{i,n}$ είναι η αξία πριν την κανονικοποίηση.
- Το M_n είναι το άνω όριο του πεδίου του χαρακτηριστικού a_n .

3.5.1.5 Μέτρο απόστασης για ανάμεικτους τύπους χαρακτηριστικών

Στις περιπτώσεις αυτές το μέτρο της απόστασης μπορεί να υπολογιστεί συνδυάζοντας τις μεθόδους που προηγήθηκαν. Για παράδειγμα, όταν υπολογίζεται η απόσταση μεταξύ δυο παραδειγμάτων iκαι jχρησιμοποιώντας μια μετρική όπως η Ευκλείδεια απόσταση, κάποιος μπορεί να υπολογίσει τη διαφορά μεταξύ ονομαστικών (nominal) και δυαδικών (binary) χαρακτηριστικών σαν μηδέν και ένα, και τη διαφορά μεταξύ αριθμητικών χαρακτηριστικών σαν τη διαφορά μεταξύ των κανονικοποιημένων αξιών τους. Το τετράγωνο της κάθε διαφοράς θα προστίθεται στη συνολική απόσταση οπότε και θα έχουμε:

$$d(x_i, x_j) = \frac{\sum_{n=1}^{p} \delta_{ij}^{(n)} \cdot d_{ij}^{(n)}}{\sum_{n=1}^{p} \delta_{ij}^{(n)}}$$

Όπου ο δείκτης $\delta_{ij}^{(n)}$ =0 αν λείπει κάποια αξία.

Τώρα, η συμμετοχή του nστην απόσταση μεταξύ των δυο αντικειμένων $d^{(n)}(x_i, x_j)$ υπολογίζεται ως εξής:

- Εάν το χαρακτηριστικό είναι δυαδικό ή κατηγορικό,

$$d^{(n)}(x_i, x_j) = o \text{ αν } x_{in} = x_{jn}, \text{ αλλιώς } d^{(n)}(x_i, x_j) = 1.$$

- Εάν το χαρακτηριστικό είναι συνεχές

$$d_{ij}^{(n)} = \frac{| x_{in} - x_{jn} |}{\max_n \cdot x_{nm} - \min_n \cdot x_{nn}},$$

Όπου το hδιατρέχει όλα τα μη αγνοούμενα αντικείμενα για το n.

- Εάν το χαρακτηριστικό είναι τακτικό, οι κανονικοποιημένες αξίες του χαρακτηριστικού υπολογίζονται πρώτα και στη συνέχεια το $Z_{i,n}$ αντιμετωπίζεται σαν συνεχόμενη αξία.

3.5.2 Συναρτήσεις ομοιότητας

Μια διαφορετική προσέγγιση σε σχέση με αυτή της απόστασης είναι η λειτουργία-συνάρτηση ομοιότητας $S(x_i, x_j)$ που συγκρίνει δυο διανύσματα x_i, x_j . Η συνάρτηση αυτή πρέπει να είναι συμμετρική, δηλαδή να ισχύει $S(x_i, x_j) = S(x_j, x_i)$, και να έχει μεγάλη τιμή όταν τα x_i, x_j είναι σχετικά όμοια. Παρακάτω αναλύονται οι διαφορετικές συναρτήσεις ομοιότητας που χρησιμοποιούνται.

3.5.2.1 Μέτρο συνημιτόνου

Όταν η γωνία μεταξύ των δυο διανυσμάτων αποτελεί αξιόλογο μέτρο για την ομοιότητά τους, το κανονικοποιημένο εσωτερικό προϊόν μπορεί να είναι ένα κατάλληλο μέτρο ομοιότητας:

$$S(x_i, x_j) = \frac{x_i^T \cdot x_j}{\| x_i \| \cdot \| x_j \|}$$

3.5.2.2 Μέτρο συσχετισμού Pearson (Pearson correlation measure)

Ο κανονικοποιημένος συσχετισμός Pearson χαρακτηρίζεται ως:

$$S(x_i, x_j) = \frac{(x_i - \bar{x}_i)^T \cdot (x_j - \bar{x}_j)}{\| x_i - \bar{x}_i \| \cdot \| x_j - \bar{x}_j \|}$$

Όπου το $\bar{x}_i$ υποδηλώνει τη μέση αξία του χ σε όλες τις διαστάσεις.

3.5.2.3 Επεκταμένο μέτρο Jaccard (Extended Jaccard Measure)

Προτάθηκε από τους Strehlκαι Gnosh (2000) και χαρακτηρίζεται ως

$$S(x_i, x_j) = \frac{x_i^T x_j}{\| x_i \|^2 + \| x_j \|^2 - x_i^T x_j}$$

3.5.2.4 Μέτρο συσχέτισης Dice (Dice Coefficient Measure)

Είναι παρόμοιο με το μέτρο Jaccard και χαρακτηρίζεται ως:

$$S(x_i, x_j) = \frac{2x_i^T x_j}{\| x_i \|^2 + \| x_j \|^2}$$

3.5.3 Κριτήρια αξιολόγησης

Η αξιολόγηση του αν μια συσταδοποίηση είναι καλή ή όχι αποτελεί ένα προβληματικό και αμφιλεγόμενο ζήτημα. Χαρακτηριστικό παράδειγμα αποτελεί ο Bohner που ήταν ο πρώτος, το 1964, που δήλωσε πως δεν υπάρχει κάποιος καθολικός ορισμός για το τί σημαίνει καλή συσταδοποίηση. Η αξιολόγηση παραμένει κυρίως αντικείμενο αυτού που την κάνει. Παρόλα αυτά, υπάρχουν αρκετά κριτήρια αξιολόγησης στη βιβλιογραφία. Αυτά τα κριτήρια, συνήθως, χωρίζονται σε δυο κατηγορίες: τα εσωτερικά και τα εξωτερικά.

3.5.4 Εσωτερικά κριτήρια αξιολόγησης

Οι εσωτερικές μετρικές ποιότητας (internalqualitymetrics) μετράνε συνήθως το κατά πόσο συμπαγείς είναι οι συστάδες, χρησιμοποιώντας κάποιο μέτρο ομοιότητας. Συνήθως μετράνε την εσωτερική ομογένεια των συστάδων, την εσωτερική διαχωριστικότητα ή ένα συνδυασμό και των δυο. Δεν χρησιμοποιείται καμία άλλη εξωτερική πληροφορία πέραν των ίδιων των δεδομένων.

3.5.4.1 Άθροισμα του τετραγωνικού σφάλματος (Sum of Squared Error)

ΤοSSEείναι το πιο απλό και δημοφιλές κριτήριο για τη συσταδοποίηση. Υπολογίζεται από τον τύπο:

$$SSE = \sum_{k=1}^{k} \sum_{\forall x_i \in c_k} \| x_i - \mu_k \|^2$$

Όπου: c_k είναι το σύνολο των παραδειγμάτων στη συστάδα K και μ_κ το διανυσματικό μέσο της συστάδας.

Ενώ τα στελέχη του μ_κ υπολογίζονται από τον τύπο:

$$\mu_{\kappa,j} = \frac{1}{N_k} \sum_{\forall x_i \in c_k} x_{i,j}$$

Όπου: $N_k = | C_k |$ είναι ο αριθμός των παραδειγμάτων που ανήκουν στη συστάδα K.

Οι μέθοδοι που ελαχιστοποιούν το SSE συχνά καλούνται χωρίσματα ελάχιστης διακύμανσης, αφού μέσω μιας απλής αλγεβρικής μετατροπής το κριτήριο SSEμπορεί να γραφτεί ως:

$$SSE = \frac{1}{2} \sum_{k=1}^{k} N_k \bar{S}_k$$

Όπου:

$$\bar{S}_k = \frac{1}{N^2} \sum_{x_i, x_j \in c_k} \| x_i - x_j \|^2$$

Το κριτήριο SSE είναι κατάλληλο για περιπτώσεις όπου οι συστάδες δημιουργούν συμπαγή ‘σύννεφα’ τα οποία είναι καλά διαχωρισμένα το ένα από το άλλο.

3.5.4.2 Άλλα κριτήρια ελάχιστης διακύμανσης

Επιπρόσθετα κριτήρια ελάχιστης διακύμανσης μπορούν να παραχθούν απλά αντικαθιστώντας την αξία του S_k με εκφράσεις όπως:

$$\bar{S}_k = \frac{1}{N^2} \sum_{x_i, x_j \in c_k} S(x_i, x_j)$$

Ή

$$\bar{S}_k = \min_{x_i, x_j \in c_k} S(x_i, x_j)$$

3.5.4.3 Κριτήριο Condorcet

Μια άλλη κατάλληλη προσέγγιση αποτελεί η εφαρμογή της λύσης του Condorcet (1785) στο πρόβλημα κατάταξης. Στην περίπτωση αυτή το κριτήριο υπολογίζεται ως:

$$\sum_{c_i \in c} \sum_{x_j, x_k \in c_i} S(x_j, x_k) + \sum_{c_i \in c} \sum_{x_j \in c_i} d(x_j, x_k)$$

Όπου: $S(x_j, x_k)$ είναι το μέτρο της ομοιότητας και $d(x_j, x_k)$ είναι το μέτρο της απόστασης μεταξύ των διανυσμάτων x_j, x_k.

3.5.4.4 C-κριτήριο

Το C-κριτήριο (Fortier&Solomon,1996) αποτελεί επέκταση του κριτηρίου Condorcetκαι χαρακτηρίζεται ως:

$$\sum_{\substack{c_i \in c \\ x_j \neq x_k}} \sum_{x_j, x_k \in c_i} (S(x_j, x_k) - \gamma) + \sum_{c_i \in c} \sum_{\substack{x_j \in c_i \\ x_k \notin c_i}} (\gamma - S(x_j, x_k))$$

Όπου γ είναι η αξία κατώφλι.

3.5.5 Εξωτερικά κριτήρια αξιολόγησης

Τα εξωτερικά κριτήρια μπορούν να είναι χρήσιμα για την εξέταση του κατά πόσο η δομή των συστάδων ταιριάζει με κάποια προκαθορισμένη κατηγοριοποίηση των παραδειγμάτων. Παρακάτω ακολουθούν κάποια από τα πιο δημοφιλή τέτοια κριτήρια που συναντά κανείς στη βιβλιογραφία.

3.5.5.1 Μέτρο βασισμένο στην αμοιβαία πληροφορία (Mutual Information Based Measure)

Το συγκεκριμένο κριτήριο μπορεί να χρησιμοποιηθεί σαν εξωτερικό μέτρο Συσταδοποίησης (Strehletal.,2000). Το μέτρο για M παραδείγματα που έχουν συσταδοποιηθεί χρησιμοποιώντας $c = \{c_1, \ldots, c_y\}$ και αναφερόμενα στο στοχευμένο χαρακτηριστικό y όπου το πεδίο του είναι το $dom(y) = \{c_1, \ldots, c_k\}$ χαρακτηρίζεται ως:

$$c = \frac{2}{m} \sum_{l=1}^{g} \sum_{k=1}^{k} m_{l,h} \log_{g \cdot k} \left(\frac{m_{l,h} \cdot m}{m_{.,l} \cdot m_l} \right)$$

Όπου:

- Το $m_{l,h}$ υποδηλώνει τον αριθμό των παραδειγμάτων της συστάδας c_l και ανήκουν στην κλάση c_h.
- Το $m_{.,h}$ υποδηλώνει τον αριθμό των παραδειγμάτων της κλάσης c_h.
- Το $m_{l,.}$ υποδηλώνει τον αριθμό των παραδειγμάτων της κλάσης c_l.

3.5.5.2 Μέτρο ανάκλησης ακρίβειας (Precision-Recall Measure)

Η συστάδα μπορεί να αντιμετωπιστεί σαν αποτέλεσμα ερωτήσεων (queries) για μια συγκεκριμένη κλάση. Η ακρίβεια είναι το κλάσμα των ορθά ανακτημένων παραδειγμάτων, ενώ η ανάκληση είναι το κλάσμα των ορθά ανακτημένων παραδειγμάτων από όλες τις κλάσεις που ταιριάζουν. Ένα συνδυαζόμενο f-μέτρο μπορεί να είναι χρήσιμο για την αξιολόγηση της τεχνικής Συσταδοποίησης (Larsen and Aone, 1999).

3.5.5.3 Δείκτης Rand (Rand Index)

Ο δείκτης Rand (Rand, 1971) είναι ένα απλό κριτήριο που συγκρίνει μια επαγόμενη δομή Συσταδοποίησης (c_1) με μια δοσμένη δομή Συσταδοποίησης (c_2). Εάν το aείναι ο αριθμός των ζευγαριών παραδειγμάτων που ανατίθενται στην ίδια συστάδα c_1 και στην ίδια συστάδα c_2, το bείναι ο αριθμός των ζευγαριών παραδειγμάτων που βρίσκονται στην ίδια συστάδα c_1 αλλά όχι στην ίδια συστάδα c_2, cείναι ο αριθμός των ζευγαριών παραδειγμάτων που βρίσκονται στην ίδια συστάδα c_2 αλλά όχι στη c_1 και do αριθμός των ζευγαριών παραδειγμάτων που ανατίθενται σε διαφορετικές συστάδες c_1 και c_2. Τότε οι ποσότητες aκαι dμπορούν να μεταφραστούν σαν ‘συμφωνίες’, ενώ τα bκαι csαν ‘διαφωνίες’ και ο δείκτης Rand προκύπτει ως:

$$Rand = \frac{a+d}{a+b+c+d}$$

Ο δείκτης αυτός κυμαίνεται μεταξύ των 0 και 1. Όταν τα δυο χωρίσματα συμφωνούν απόλυτα ο δείκτης ισούται με 1.

3.5.6 Μέθοδοι Συσταδοποίησης - Αλγοριθμική Ανάλυση

Στο παρόν κεφάλαιο θα αναλυθούν οι πιο δημοφιλείς τεχνικές Συσταδοποίησης. Ο βασικός λόγος της ύπαρξης τόσο πολλών τεχνικών οφείλεται στο γεγονός ότι η έννοια της συστάδας δεν είναι ακριβώς καθορισμένη (Estirill – Castro, 2000). Συμπερασματικά, πολλές τεχνικές Συσταδοποίησης έχουν αναπτυχθεί, όπου κάθε μια χρησιμοποιεί και μια διαφορετική αρχή επαγωγής. Για παράδειγμα, οι Farley and Raftery (1998) προτείνουν τον διαχωρισμό των μεθόδων Συσταδοποίησης σε δυο ομάδες: ιεραρχικές και διαμερισμού.

3.5.6.1 Ιεραρχικές Μέθοδοι Συσταδοποίησης (Hierarchical Methods)

Οι μέθοδοι αυτοί κατασκευάζουν συστάδες χωρίζοντας αναδρομικά τα παραδείγματα με τρόπο διαμόρφωσης είτε από ΄κάτω προς τα πάνω’ είτε από ‘πάνω προς τα κάτω’. Οι μέθοδοι αυτοί μπορούν να χωριστούν με τη σειρά τους με τους ακόλουθους τρόπους:

3.5.6.2 Συσσωρευτική ΙεραρχικήΣυσταδοποίηση (Agglomerative Hierarchical Clustering)

Κάθε αντικείμενο, αρχικά, αντιπροσωπεύει μια συστάδα από μόνο του. Στη συνέχεια οι συστάδες ενώνονται διαδοχικά μέχρι να επιτευχθεί η επιθυμητή δομή συστάδας.

3.5.6.3 Διαιρετική ΙεραρχικήΣυσταδοποίηση (Divisive Hierarchical Clustering)

Όλα τα αντικείμενα, αρχικά, ανήκουν στην ίδια συστάδα. Μετά η συστάδα διαιρείται σε επιμέρους συστάδες, οι οποίες διαδοχικά διαιρούνται σε δικές τους υπό-συστάδες. Η διαδικασία συνεχίζεται μέχρι να επιτευχθεί η επιθυμητή δομή συστάδας.

Το αποτέλεσμα των ιεραρχικών μεθόδων είναι ένα δενδρόγραμμα το οποίο αντιπροσωπεύει την εμφωλευμένη ομαδοποίηση των αντικειμένων και τα επίπεδα ομοιότητας στα οποία οι ομάδες αυτές αλλάζουν. Η Συσταδοποίηση των αντικειμένων επιτυγχάνεται κόβοντας το δενδρόγραμμα στο επιθυμητό επίπεδο ομοιότητας.

Η συνένωση ή ο διαχωρισμός των συστάδων γίνεται σύμφωνα με κάποιο μέτρο ομοιότητας, το οποίο επιλέγεται με σκοπό τη βελτιστοποίηση κα΄ποιου κριτηρίου.

3.5.6.4 Πλεονεκτήματα και Μειονεκτήματα

Σε γενικές γραμμές, οι ιεραρχικές μέθοδοι έχουν τα εξής πλεονεκτήματα:

- Ευελιξία στον τρόπο Συσταδοποίησης παρέχοντας καλούς διαχωρισμούς.
- Παροχή πολλαπλών διαχωρισμών. Οι ιεραρχικές μέθοδοι δεν παράγουν μόνο ένα χώρισμα αλλά πολλά εμφωλευμένα, γεγονός που επιτρέπει στους χρήστες να επιλέγουν διαφορετικά χωρίσματα, ανάλογα με το επιθυμητό επίπεδο ομοιότητας.

Από την άλλη, τα βασικά πλεονεκτήματα είναι:

- Αδυναμία να κλιμακώσουν καλά. Οι ιεραρχικές μέθοδοι είναι χρονοβόρες και κοστοβόρες.
- Οι ιεραρχικές μέθοδοι δεν μπορούν να ξεκάνουν ότι έχει ήδη γίνει, εννοώντας ότι δεν διαθέτουν κάποια δυνατότητα επιστροφής σε προηγούμενη κατάσταση.

3.5.7 Μέθοδοι Διαμερισμού (Partitioning Methods)

Οι μέθοδοι διαμερισμού ανατοποθετούν τα παραδείγματα μεταφέροντας τα από τη μια συστάδα στην άλλη, βάση ενός αρχικού διαμερισμού. Τέτοιες μέθοδοι τυπικά απαιτούν ο αρχικός αριθμός των συστάδων να ορίζεται προηγουμένως από τον χρήστη. Προκειμένου να επιτευχθεί καθολική

βελτιστοποίηση χρειάζεται μια διαδικασία εξαντλητικής απαρίθμησης όλων των πιθανών διαμερισμών. Επειδή αυτό δεν είναι εφικτό, χρησιμοποιούνται διάφορες ‘άπληστες’ ευρετικές μέθοδοι (heuristics).

3.5.8 Αλγόριθμοι Ελαχιστοποίησης Σφάλματος (Error Minimization Algorithms)

Αυτοί οι αλγόριθμοι τείνουν να λειτουργούν καλά με απομονωμένες και συμπαγείς συστάδες και αποτελούν μια από τις πιο δημοφιλείς μεθόδους. Η βασική ιδέα είναι να βρεθεί μια δομή Συσταδοποίησης που ελαχιστοποιεί το σφάλμα βάση ενός κριτηρίου το οποίο μετράει την απόσταση κάθε παραδείγματος με την αντιπροσωπευτική του αξία. Το πιο γνωστό κριτήριο είναι το άθροισμα του τετραγωνικού σφάλματος (Sum of squared error–SSE), το οποίο μετράει την ολική Ευκλείδεια απόσταση των παραδειγμάτων στις αντιπροσωπευτικές τους αξίες. Το SSEμπορεί να βελτιστοποιηθεί καθολικά απαριθμώντας όλα τα χωρίσματα, το οποίο είναι πολύ χρονοβόρο, ή με το να δοθεί μια προσεγγιστική λύση χρησιμοποιώντας ευρετικές μεθόδους. Η τελευταία επιλογή είναι και η πιο κοινή εναλλακτική.

Ο πιο απλός και περισσότερο χρησιμοποιήσιμος αλγόριθμος αυτής της κατηγορίας είναι ο K-Means. Ο αλγόριθμος αυτός διαμερίζει τα δεδομένα σε K συστάδες $(c_1, c_2, ..., c_k)$ οι οποίες αντιπροσωπεύονται από τα κέντρα των μέσων τους. Το κέντρο κάθε συστάδας υπολογίζεται ως το μέσο όλων των παραδειγμάτων που ανήκουν στη συγκεκριμένη συστάδα. Ο αλγόριθμος ξεκινάει με ένα αρχικό σετ κέντρων συστάδων, επιλεγμένα στην τύχη ή σύμφωνα με κάποια ευρετική μέθοδο. Σε κάθε επανάληψη, κάθε παράδειγμα ανατίθεται στην πλησιέστερη συστάδα σύμφωνα με την Ευκλείδεια απόσταση μεταξύ των δυο. Κατόπιν, υπολογίζεται ξανά το κέντρο της συστάδας.

Το κέντρο κάθε συστάδας υπολογίζεται ως το μέσο όλων των παραδειγμάτων που ανήκουν σε αυτή τη συστάδα:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

Όπου το N_k είναι ο αριθμός των παραδειγμάτων που ανήκουν στη συστάδα K και το μ_κ είναι ο μέσος της συστάδας K.

Ο αλγόριθμος K-Means μπορεί να ιδωθεί σαν μια Gradient-descentδιαδικασία, η οποία ξεκινάει με έναν αρχικό αριθμό K κέντρων-συστάδων και διαδοχικά τις ανανεώνει εως την ελαχιστοποίηση του σφάλματος της λειτουργίας. Απόδειξη της πεπερασμένης σύγκλισης του K-Means παρέχεται στο (Selim&Ismail, 1984). Η πολυπλοκότητα T επαναλήψεων του αλγόριθμου σε ένα δείγμα mπαραδειγμάτων, όπου κάθε ένα χαρακτηρίζεται από N χαρακτηριστικά δίνεται από το:

$$\mathrm{O}(\mathrm{T} * \mathrm{K} * m * N)$$

Η γραμμική πολυπλοκότητα αυτή, είναι ένας από τους λόγους της δημοτικότητας των αλγορίθμων τύπου K-Means. Ακόμα και αν ο αριθμός των παραδειγμάτων είναι σημαντικά μεγάλος, ο αλγόριθμος παραμένει υπολογιστικά ελκυστικός. Έτσι έχει σαφέστατο πλεονέκτημα έναντι άλλων τεχνικών Συσταδοποίησης που έχουν μη-γραμμική πολυπλοκότητα.

Άλλοι λόγοι για τη δημοτικότητα του αλγορίθμου είναι η ευκολία της μετάφρασής του, η απλότητα εφαρμογής, η ταχύτητα σύγκλισης και η προσαρμοστικότητα στα δεδομένα. Η μεγάλη αδυνμάι του

αλγόριθμου είναι η επιλογή του αρχικού διαμερισμού. Ο αλγόριθμος είναι πολύ ευαίσθητος στην επιλογή αυτή, η οποία μπορεί να κάνει τη διαφορά μεταξύ καθολικού και τοπικού ελάχιστου. Όντας ένας τυπικός αλγόριθμος διαμερισμού, ο K-Means λειτουργεί καλά μόνο σε σύνολα δεδομένων που έχουν ισοτροπικές συστάδες και δεν είναι τόσο ευέλικτος όσο άλλοι αλγόριθμοι.

Συμπληρωματικά, ο αλγόριθμος είναι ευαίσθητος στον θόρυβο και σε υπερβολικά υψηλά (outliers), είναι εφαρμόσιμος μόνο όταν ο μέσος έχει χαρακτηριστεί και απαιτεί προκαταβολικά τον αριθμό των συστάδων, το οποίο δεν είναι εύκολο όταν δεν υπάρχει πρότερη γνώση.

Η χρήση του K-Means συχνά περιορίζεται μόνο σε αριθμητικά χαρακτηριστικά. Ο Haung (1998) παρουσίασε τον αλγόριθμο K-prototypeso οποίος βασίζεται στον K-Means αλλά αφαιρεί τους αριθμητικούς περιορισμούς ενώ διατηρε΄την αποτελεσματικότητα. Ο αλγόριθμος συσταδοποιεί αντικείμενα με αριθμητικά και κατηγορικά χαρακτηριστικά με τρόπο παρόμοιο του K-Means. Το μέτρο της ομοιότητας για τα αριθμητικά χαρακτηριστικά είναι το τετράγωνο της Ευκλείδειας απόστασης ενώ για τα κατηγορικά χαρακτηριστικά είναι ο αριθμός των αναντιστοιχιών μεταξύ των αντικειμένων και των πρωτότυπων συστάδων.

Ένας άλλος αλγόριθμος διαμερισμού που προσπαθεί να ελαχιστοποιήσει το SSEείναι ο K-Medoids ή PAM (Kaufmann & Rousseeuw, 1987). Ο αλγόριθμος αυτός είναι όμοιος με τον K-Means. Διαφέρει κυρίως στην παρουσίαση των διαφορετικών συστάδων. Κάθε συστάδα αντιπροσωπεύεται από το πιο κεντρικό αντικείμενο, σε αντίθεση με το μέσο της συστάδας. Η μέθοδος K-medoids είναι πιο εύρωστη από τον K-Means στην παρουσία θορύβου και υπερβολικά υψηλών σημείων (outliers) γιατί ένα medoid επηρεάζεται αισθητά λιγότερο από υπερβολικά υψηλά και άλλες υπερβολικές αξίες σε σχέση με τον μέσο. Όμως είναι αρκετά πιο κοστοβόρος σε σχέση με τον K-Means, ενώ και εδώ απαιτείται από τον χρήστη να καθορίσει τον αριθμό των συστάδων.

3.5.9 Graph-Theoretic Συσταδοποίηση

Οι Graph-Theoretic μέθοδοι παράγουν συστάδες μέσω γράφων. Οι ακμές του γράφου συνδέουν τα παραδείγματα που εκφράζονται ως κόμβος. Ένας πολύ γνωστός τέτοιος αλγόριθμος βασίζεται στο Minimal Spanning Tree(Zahn, 1971). Ασυνεπείς ακμές, είναι οι ακμές όπου το βάρος τους είναι συγκριτικά μεγαλύτερο από τον μέσο όρο των γειτονικών μηκών των ακμών.

4. Data Mining για Χρηματοοικονομικές εφαρμογές

Η πρόβλεψη της κίνησης των μετοχών του χρηματιστηρίου, ο ρυθμός ανταλλαγής συναλλάγματος, οι τραπεζικές χρεοκοπίες, η κατανόηση και η διαχείριση του χρηματοοικονομικού ρίσκου, η αξιολόγηση της πιστοληπτικής ικανότητας, η δανειακή διαχείριση, το profiling των πελατών μιας τράπεζας και αναλύσεις

ξεπλύματος χρήματος είναι όλες οι θεμελιώδεις εργασίες του Data Mining στον χώρο των χρηματοοικονομικών.

Η πρόβλεψη των τιμών του χρηματιστηρίου (Stock Market Forecasting) περιλαμβάνει την αποκάλυψη των τάσεων της αγοράς, τον σχεδιασμό επενδυτικών στρατηγικών, την αναγνώριση της καλύτερης στιγμής να αγοραστεί μια μετοχή και ποιες μετοχές θα αγοραστούν. Τα οικονομικά ιδρύματα παράγουν τεράστια σε όγκο σύνολα δεδομένων τα οποία χτίζουν τη βάση για την προσέγγιση αυτών των πολύπλοκων και δυναμικών προβλημάτων με τη χρήση εργαλείων Data Mining.

Σχεδόν κάθε υπολογιστική μέθοδος έχει εξερευνηθεί και χρησιμοποιηθεί για την χρηματοοικονομική μοντελοποίηση. Κάποιες πρόσφατες μελέτες περιλαμβάνουν: την προσομοίωση Monte-Carlo στην τιμολόγηση δικαιωμάτων προαίρεσης (Huang et al., 2004), την προσέγγιση πεπερασμένων διαφορών στα παράγωγα του ρυθμού επιτοκίων και τον μετασχηματισμό Fourier για την παράγωγη τιμολόγηση. Νέες εξελίξεις ενισχύουν τις παραδοσιακές τεχνικές αναλύσεις των καμπυλών του χρηματιστήριού (Murphy, 1999). Τεχνικές αναλύσεις που βελτιώνουν αισθητά στην αναγνώριση των σημείων αγοράς ή πώλησης μιας μετοχής.

Η εξόρυξη δεδομένων σαν μια διαδικασία αναγνώρισης χρήσιμων προτύπων και συσχετισμών κατέχει τη δική της θέση στο πεδίο της χρηματοοικονομικής μοντελοποίησης. Παρόμοια με τις άλλες υπολογιστικές μεθόδους, σχεδόν κάθε μέθοδος και τεχνική εξόρυξης δεδομένων έχει χρησιμοποιηθεί για τη χρηματοοικονομική μοντελοποίηση. Μια τέτοια λίστα περιλαμβάνει ποικιλία γραμμικών και μη-γραμμικών μοντέλων, Νευρωνικά Δίκτυα πολλαπλών στρώσεων, K-Means και ιεραρχική συσταδοποίηση, Κ-κοντινοί γείτονες, ανάλυση μέσω δέντρων απόφασης, παλινδρόμηση, ARIMA και μπαγιεσιανή μάθηση. Ενώ έχουν χρησιμοποιηθεί και αρκετές τεχνικές αξιολόγησης όπως το bootstrapping.

Η απλοϊκή προσέγγιση της εξόρυξης δεδομένων στον χρηματοοικονομικό χώρο υποθέτει ότι κάποιος μπορεί να χρησιμοποιήσει ένα εγχειρίδιο οδηγιών στο πως να επιτύχει το καλύτερο αποτέλεσμα, ενώ κάποιες δημοσιεύσεις υποστηρίζουν ακόμα αυτή την αδικαιολόγητη πεποίθηση. Στην πραγματικότητα, η μόνη ρεαλιστική προσέγγιση που φαίνεται να φέρνει επιτυχημένα αποτελέσματα είναι η παροχή συγκρίσεων μεταξύ διαφορετικών μεθόδων δείχνοντας τις δυνάμεις και τις αδυναμίες της καθεμίας σχετικά με τα χαρακτηριστικά του προβλήματος αφήνοντας στην κρίση του χρήστη την επιλογή της καταλληλότερης μεθόδου που ταιριάζει καλύτερα στο πρόβλημα. Στην ουσία, αυτό σημαίνει πως χρειάζεται μια ξεκάθαρη κατανόηση της εξόρυξης δεδομένων γενικότερα, και στον χρηματοοικονομικό χώρο αυτό αποτελεί περισσότερο τέχνη παρά επιστήμη.

Ευτυχώς σήμερα, υπάρχει ένας αυξανόμενος αριθμός βιβλίων που συζητά αυτά τα θέματα και τις μεθόδους. Για παράδειγμα, η κατανόηση της δύναμης των κανόνων if-then επί των δέντρων απόφασης μπορεί δραστικά να αλλάξει και να βελτιώσει την προσέγγιση του Data Mining στα χρηματοοικονομικά. Συγκριτικά με άλλα πεδία όπως η Γεωλογία ή η Φαρμακευτική όπου το τέστ της πρόβλεψης είναι ακριβό και δύσκολο, ακόμα και επικίνδυνο, μια οικονομική πρόβλεψη από την άλλη μεριά είναι εύκολο να γίνει χωρίς να εμπεριέχει κάποιο κίνδυνο ή ρίσκο.

Μέθοδοι μάθησης βάσει χαρακτηριστικών όπως τα Νευρωνικά Δίκτυα, η μέθοδος των κοντινότερων γειτόνων και τα δέντρα απόφασης κυριαρχούν σε οικονομικές εφαρμογές του Data Mining. Τέτοιες μέθοδοι είναι σχετικά απλές, αποδοτικές και μπορούν να χειριστούν θορυβώδη δεδομένα. Παρόλα αυτά, οι μέθοδοι αυτές έχουν δυο σημαντικά προβλήματα: περιορισμένη ικανότητα παρουσίασης ιστορικής

γνώσης και έλλειψη περίπλοκων σχέσεων. Σχεσιακές (relational) τεχνικές εξόρυξης δεδομένων που περιλαμβάνουν το Inductive Logic Programming επιχειρούν να υπερνικήσουν τους περιορισμούς αυτούς. Στο παρελθόν, αυτές οι μέθοδοι ήταν σχετικά μη αποδοτικές υπολογιστικά (Thulasiram, 1999) και είχαν περιορισμένες ικανότητες στην εφαρμογή αριθμητικών δεδομένων. Πλέον, οι μέθοδοι αυτοί έχουν ενισχυθεί και στις δυο προαναφερθείσες πτυχές και είναι η κατάλληλη περίοδος για την εφαρμογή τους κυρίως σε αναλύσεις τύπου πιθανολογικής σχεσιακής λογικής (probabilistic relational reasoning).

Διάφορες δημοσιεύσεις έχουν εκτιμήσει τη χρήση μεθόδων εξόρυξης γνώσης όπως υβριδικές αρχιτεκτονικές Νευρωνικών Δικτύων με γενετικούς αλγόριθμους, θεωρία του χάους και ασαφή λογική (Fuzzy logic) στη χρηματοοικονομία. Σημαντικές εκτιμήσεις τοποθετούν τη διαχείριση χαρτοφυλακίων της τάξης των 5 και 10 δισεκατομμυρίων στα χέρια εμπορικών μοντέλων Νευρωνικών Δικτύων. Ενώ το ποσό αυτό μεγαλώνει σταδιακά όσο οι εταιρίες πειραματίζονται και απαιτούν μεγαλύτερη εμπιστοσύνη με τα Νευρωνικά Δίκτυα και τις ανάλογες τεχνικές μεθόδους.

4.1 Ιδιαιτερότητες του Data Mining στον χώρο της χρηματοοικονομίας

Οι ιδιαιτερότητες του Data Mining στον χρηματοοικονομικό χώρο προέρχονται από τις ανάγκες:

- Της πρόβλεψης χρονοσειρών πολλών διαστάσεων με υψηλά επίπεδα θορύβου.
- Της δημιουργίας συγκεκριμένων κριτηρίων απόδοσης.
- Της πραγματοποίησης συντονισμένης πολύ-αναλυτικής πρόβλεψης (multi resolution forecasting) για λεπτά, ημέρες, εβδομάδες, μήνες και έτη.
- Της ενσωμάτωσης της ροής γραπτών σημάτων σαν είσοδο για τα προβλεπτικά μοντέλα.
- Της ικανότητας επεξήγησης της πρόβλεψης και των προβλεπτικών μοντέλων.
- Της δυνατότητας αποκόμισης κέρδους από πολύ απλά μοτίβα με περιορισμένο χρόνο ζωής.
- Της ενσωμάτωσης των επιπτώσεων των παικτών της αγοράς στην ικανότητα της αγοράς.

Η παρούσα θεωρία για την αγορά (efficient market theory) αποτρέπει την προσπάθεια ανακάλυψης μακροχρόνιων σταθερών κανόνων για τις συναλλαγές ή κανονικοτήτων με σημαντικό κέρδος. Η θεωρία στηρίζεται στην ιδέα ότι εάν υπάρχουν τέτοιες κανονικότητες θα είχαν ανακαλυφθεί και χρησιμοποιηθεί από παίκτες της αγοράς. Αυτό θα έκανε αυτούς τους κανόνες λιγότερο επικερδείς και κάποια στιγμή άχρηστους ή ακόμα και καταστροφικούς.

Η θεωρία της αποδοτικότητας της αγοράς δεν αποκλείει την ύπαρξη κρυμμένων βραχυχρόνιων τοπικών κανονικοτήτων. Αυτές οι κανονικότητες όμως δεν μπορούν να δουλεύουν για πάντα, ενώ χρειάζεται συχνή διόρθωση. Έχει δειχτεί ότι τα χρηματοοικονομικά δεδομένα δεν είναι τυχαία και ότι η υπόθεση για την αποδοτική αγορά είναι απλά ένα υποσύνολο μιας μεγαλύτερης χαοτικής υπόθεσης για την αγορά (Drake & Kim, 1997). Η υπόθεση αυτή δεν αποκλείει την ύπαρξη επιτυχημένων βραχυχρόνιων προβλεπτικών μοντέλων σε χαοτικές χρονοσειρές (Casdagli & Eubank, 1992).

Το Data Mining δεν προσπαθεί να δεχθεί ούτε να απορρίψει την θεωρία της απόδοσης της αγοράς. Το Data Mining δημιουργεί εργαλεία, τα οποία μπορούν να είναι χρήσιμα στην ανακάλυψη μικρών βραχυπρόθεσμων δεσμευμένων μοτίβων και τάσεων σε ένα μεγάλο εύρος του χρηματοοικονομικού χώρου. Αυτό σημαίνει ότι η συνεχόμενη επανεκπαίδευση πρέπει να αποτελεί ένα μόνιμο κομμάτι

του Data Mining στα χρηματοοικονομικά και όσοι ισχυρίζονται ότι υπάρχει μια χρυσή λύση πρέπει να αντιμετωπίζονται με καχυποψία.

4.1.1 Ανάλυση Χρονοσειρών

Ένα χρονικό σύνολο δεδομένων Τ που καλείται χρονοσειρά, μοντελοποιείται στην προσπάθεια ανακάλυψης των βασικών της χαρακτηριστικών όπως η μακροχρόνια τάση (long term trend), L(T), η κυκλική διακύμανση (cyclic variation), C(T), η εποχική διακύμανση (seasonal variation), S(T), και οι ακανόνιστες κινήσεις (irregular movements), I(T). Εάν υποτεθεί ότι το Τ είναι μια χρονοσειρά όπως το καθημερινό κλείσιμο της τιμής μιας μετοχής, ή ο δείκτης SP500, από τη στιγμή 0 στην παρούσα στιγμή Κ, τότε η επόμενη αξία της χρονοσειράς T(k+n) μοντελοποιείται από τον τύπο: T(k+n)=L(T)+C(T)+S(T)+I(T).

Παραδοσιακά, τα κλασικά μοντέλα ARIMA καταλαμβάνουν τον χώρο της εύρεσης παραμέτρων ή λειτουργιών για τέτοια μοντέλα. Τα μοντέλα ARIMA είναι αρκετά καλά ανεπτυγμένα αλλά είναι δύσκολο να χρησιμοποιηθούν για μη-στατικές στοχαστικές διαδικασίες. Πιθανές μέθοδοι Data Mining μπορούν να χρησιμοποιηθούν για την κατασκευή μοντέλων που υπερνικούν τους περιορισμούς των ARIMA. Το πλεονέκτημα αυτού του μοντέλου τεσσάρων συστατικών έναντι των μοντέλων 'μαύρων κουτιών' όπως τα Νευρωνικά Δίκτυα είναι ότι μπορούν να επεξηγηθούν.

4.1.2 Επιλογή Δεδομένων και ο Ορίζοντας Πρόβλεψης

Το Data Mining στα χρηματοοικονομικά αντιμετωπίζει τις ίδιες προκλήσεις όπως και γενικά κατά τη διαδικασία επιλογής δεδομένων για την κατασκευή μοντέλων. Στα χρηματοοικονομικά, η επιλογή αυτή είναι στενά συνδεδεμένη με την επιλογή της στοχευμένης μεταβλητής(target variable).

Υπάρχουν αρκετές επιλογές για τη στοχευμένη μεταβλητή y:

$y: \mathrm{y} = \mathrm{T}(\mathrm{k}+1), \mathrm{y} = \mathrm{T}(\mathrm{k}+2), \ldots, \mathrm{y} = \mathrm{T}(\mathrm{k}+\mathrm{n})$, όπου $\mathrm{y} = \mathrm{T}(\mathrm{k}+1)$ αντιπροσωπεύει την πρόβλεψη για την επόμενη χρονική κίνηση και το $\mathrm{y} = \mathrm{T}(\mathrm{k}+\mathrm{n})$ αντιπροσωπεύει την πρόβλεψη για n κινήσεις μπροστά. Η επιλογή του συνόλου δεδομένων Τ και του μεγέθους του για έναν συγκεκριμένο επιθυμητό ορίζοντα πρόβλεψης n αποτελεί σημαντική πρόκληση.

Για σταθερές στοχαστικές διαδικασίες, η απάντηση είναι γνωστή, ένα καλύτερο μοντέλο μπορεί να φτιαχτεί για μεγαλύτερη διάρκεια εκπαίδευσης. Αυτό όμως δεν ισχύει για χρηματοοικονομικές χρονοσειρές όπως ο δείκτης SP500. Εδώ η μεγαλύτερη διάρκεια εκπαίδευσης μπορεί να δημιουργήσει πολλά και αντιφατικά μοτίβα κέρδους που αντικατοπτρίζουν μόνο συγκεκριμένες περιόδους της αγοράς. Από την άλλη, μοντέλα που χτίζονται εκπαιδευόμενα σε μικρές χρονικές περιόδους υποφέρουν από overfitting και δεν είναι εφαρμόσιμα λοταν η αγορά κινείται σε κρίσιμα σημεία. Επίσης, στον χρηματοοικονομικό τομέα οι μακροχρόνιες αποδόσεις μπορεί να προβλέπονται καλύτερα σε σχέση με τις βραχυχρόνιες ανάλογα πάντα με το σύνολο δεδομένων εκπαίδευσης και τις παραμέτρους του μοντέλου.

Παραδοσιακά στο Data Mining είναι τυπικό να υποθέτει κανείς ότι η ποιότητα του μοντέλου δεν εξαρτάται από τη συχνότητα της χρήσης του. Σε μια χρηματοοικονομική εφαρμογή η συχνότητα χρήσης είναι μια από τις παραμέτρους που επηρεάζουν την ποιότητα του μοντέλου. Αυτό συμβαίνει επειδή στα χρηματοοικονομικά το κριτήριο της ποιότητας του μοντέλου δεν περιορίζεται στην ακρίβεια της πρόβλεψης, αλλά οδηγείται και από το κατά πόσο επικερδές είναι αυτό.

4.1.3 Μέτρα Επιτυχίας

Παραδοσιακά η ποιότητα των χρηματοοικονομικών Data Mining προβλεπτικών μοντέλων μετριέται με την τυπική απόκλιση μεταξύ των προβλεπόμενων και πραγματικών αξιών στα σύνολα εκπαίδευσης και ελέγχου. Η προσέγγιση αυτή δουλεύει καλά σε πολλά πεδία αλλά θα πρέπει να την ξανά επισκεφτεί κανείς όσον αφορά τις χρηματιστηριακές συναλλαγές. Δυο μοντέλα μπορεί να έχουν την ίδια τυπική απόκλιση αλλά να δίνουν διαφορετικά αποτελέσματα. Ένα μικρό R^2 δεν είναι ρακετό να κρίνει ότι το προβλεπτικό μοντέλο θα προβλέπει σωστά την αλλαγή κατεύθυνσης μια μετοχής.

Καταλληλότερα μέτρα επιτυχίας στο χρηματοοικονομικό Data Mining είναι μέτρα όπως το Average Monthly Excess Return (AMER) και το Potential Trading Profits (PTP) (Greenstone & Oyer, 2000):

$AMER_j = R_{ij} - \beta_\iota R_{500\,j} - (\sum_{j=1}^{12}(R_{ij} - \beta_i R_{500\,j})/12)$, όπου:

- Το R_{ij} είναι η μέση απόδοση για τον δείκτη SP500 στη βιομηχανία i και για το μήνα j.
- Το $R_{500\,j}$ είναι η μέση απόδοση για τον δείκτη SP500 για το μήνα j.
- Οι αξίες β_i τροποποιούν το AMER για την ευαισθησία του δείκτη στη συνολική αγορά.

Τώρα όσον αφορά το δεύτερο μέτρο απόδοσης το Potential Trading Profits: $PTP_{ij} = ij - R500_j$

Και δείχνει το κέρδος συναλλαγών του επενδυτή έναντι της εναλλακτικής επένδυσης βασισμένης στον ευρύτερο δείκτη SP500.

4.1.4 Ποιότητα των Μοτίβων και Αξιολόγηση των Υποθέσεων

Ένα σημαντικό ζήτημα στο Data Mining γενικά αλλά και πιο συγκεκριμένα στον χρηματοοικονομικό χώρο είναι η αξιολόγηση της ποιότητας του μοτίβου P που έχει ανακαλυφθεί, μετρημένο από τη στατιστική σημαντικότητά του. Μια τυπική προσέγγιση υποθέτει τον έλεγχο της αρχικής υπόθεσης (null hypothesis) Η ότι το μοτίβο δεν είναι στατιστικά σημαντικό στο επίπεδο (α). Ένα χρήσιμο στατιστικό τεστ απαιτεί ότι οι παράμετροι του μοτίβου όπως οι μήνες του χρόνου, πρέπει να επιλέγονται τυχαία. Γεγονός που σε πολλές περιπτώσεις δεν ισχύει.

Οι Greenstone & Oyer διαφωνούν ότι σε πολλές περιπτώσεις οι παράμετροι δεν επιλέγονται τυχαία αλλά προέρχονται από το ‘data snooping’, δηλαδή τον έλεγχο του συνδυασμού των μερών της βιομηχανίας και τους μήνες απόδοσης και κατόπιν η παροχή αναφοράς μόνο για λίγους σημαντικούς συνδυασμούς. Αυτό σημαίνει ότι ένα εξαντλητικό τεστ θα χρειάζεται τον έλεγχο διαφορετικών αρχικών υποθέσεων όχι μόνο για έναν σημαντικό συνδυασμό αλλά για ολόκληρη την οικογένεια των συνδυασμών. Κάθε

συνδυασμός αφορά συγκεκριμένο μέρος της βιομηχανίας με το μήνα απόδοσης. Στο περιβάλλον αυτό, η απόδοση της οικογένειας των συνδυασμών ελέγχεται έναντι της συνολικής απόδοσης της αγοράς.

Αρκετές επιλογές ελέγχου είναι διαθέσιμες. Οι Sullival et al. (1998) προτείνοθν μια μέθοδο bootstrapping για την εκτίμηση της στατιστικής σημαντικότητας τέτοιων υποθέσεων, συνυπολογίζοντας τις επιπτώσεις του data snooping στους συναλλακτικούς κανόνες και τις ημερολογιακές ανωμαλίες. Οι Greenstone & Oyer (2000) προτείνουν μια απλή υπολογιστική μέθοδο που συνδυάζει τα αποτελέσματα ξεχωριστών t-test χρησιμοποιώντας την ανισότητα Bonferroni ότι δοθέντος ενός συνόλου γεγονότων $A_1, A_2, ..., A_n$ η πιθανότητα της ένωσής τους είναι μικρότερη ή ίση από το άθροισμα των πιθανοτήτων τους: $P(A_1 \& A_2 \& ... \& A_k) \leq \sum_{i=1:k} P(A_i)$.

Όπου το A_i υποδηλώνει την ψευδή απόρριψη της δήλωσης i από μια οικογένεια με κ δηλώσεις (statements).

Μια άλλη επιλογή θα ήταν ο έλεγχος του κατά πόσο οι καταστάσεις αυτές είναι από κοινού αληθείς χρησιμοποιώντας ένα παραδοσιακό f-test. Παρόλα αυτά, εάν απορριφθεί η αρχική υπόθεση για μια τέτοια κοινή κατάσταση τότε δεν μπορεί να αναγνωριστούν οι επικερδείς στρατηγικές συναλλαγής.

Η διαδοχική σημασιολογική πιθανοτική συλλογιστική (sequential semantic probabilistic reasoning) που χρησιμοποιεί το f-test απευθύνεται σε αυτό το ζήτημα (Kovalerchuk & Vityaev, 2000). Με τη μέθοδο αυτή μπόρεσαν να βρεθούν επικερδή και στατιστικά σημαντικά μοτίβα για τον δείκτη SP500. Η ιδέα για τη διαδοχική πιθανοτική συλλογιστική προέρχεται από την αρχή του Occam's Razor (νόμος της απλότητας) στην επιστήμη και τη φιλοσοφία. Ανεπίσημα στον χώρο των χρηματοοικονομικών συναλλαγών δημιουργήθηκε από traders ως εξής:

- Όταν υφίστανται δυο αντιμαχόμενες θεωρίες συναλλαγών που κάνουν ακριβώς τις ίδιες προβλέψεις, τότε η πιο απλή είναι η καλύτερη και πιο επικερδής.
- Όταν υφίστανται δυο συναλλακτικές ή επενδυτικές θεωρίες όπου και οι δυο εξηγούν τα παρατηρούμενα δεδομένα, πρέπει να επιλέγεται η πιο απλή μέχρι να υπάρξουν περισσότερες αποδείξεις.
- Η πιο απλή εξήγηση για την κίνηση της τιμής κάποιας μετοχής ή εμπορεύματος είναι πιο πιθανό να οδηγεί σε καλύτερη ακρίβεια σε σχέση με μια πιο πολύπλοκη.
- Εάν υπάρχουν δυο ισότιμα πιθανές λύσεις για ένα πρόβλημα χρηματιστηριακής συναλλαγής, τότε επιλέγεται η πιο απλή.
- Η εξήγηση της κίνησης μιας τιμής που κάνει τις λιγότερες υποθέσεις είναι πιο πιθανά σωστή.

4.2 Πτυχές της Μεθοδολογίας του Data Mining στα Χρηματοοικονομικά

Το Data Mining στα χρηματοοικονομικά τυπικά ακολουθεί το ίδιο γενικό σετ βημάτων όπως και σε οποιαδήποτε άλλη εργασία. Αυτά τα βήματα περιλαμβάνουν την κατανόηση του προβλήματος, τη συλλογή και επεξεργασία των δεδομένων, την κατασκευή ενός μοντέλου, την αξιολόγηση του μοντέλου και την ανάπτυξή του (deployment).

Ένα άλλο σημαντικό βήμα στην όλη διαδικασία είναι η προσθήκη κανόνων στηριγμένους στη γνώση των ειδικών του χώρου, ειδικά όταν υπάρχει έλλειψη ή και απουσία δεδομένων. Το λεγόμενο 'expert-mining'

αποτελεί μια πολύτιμη πρόσθετη πηγή κανονικοτήτων. Παρόλα αυτά, τέτοια συστήματα ανταποκρίνονται με πολύ αργό τρόπο στις αλλαγές της αγοράς.

4.2.1 Σχεσιακές Μεθοδολογίες και Μεθοδολογίες βασισμένες στα Χαρακτηριστικά

Αρκετές παράμετροι χαρακτηρίζουν τις μεθοδολογίες Data Mining για χρηματοοικονομικές προβλέψεις. Οι κατηγορίες των δεδομένων και οι μαθηματικοί αλγόριθμοι είναι μερικές από τις πιο σημαντικές. Ο πρώτος τύπος δεδομένων αντιπροσωπεύεται από τα χαρακτηριστικά των αντικειμένων, δηλαδή κάθε αντικείμενο χ δίνεται από ένα σετ αξιών $A_1(x), A_2(\mathrm{x}), \ldots \mathrm{A}_n(x)$. Η κοινή Data Mining μεθοδολογία προσλαμβάνει αυτόν τον τύπο και είναι γνωστή ως μεθοδολογία βασισμένη στα χαρακτηριστικά (attribute-based) και καλύπτει ένα μεγάλο εύρος στατιστικών μεθόδων.

Τα σχεσιακά δεδομένα (relational data) είναι ο δεύτερος τύπος, όπου τα αντικείμενα παριστάνονται από τις σχέσεις τους με τα άλλα αντικείμενα, για παράδειγμα X>Y, Y<Z, X>Z. Σε αυτό το παράδειγμα μπορεί να μην γνωρίζουμε ότι X=3, Y=1 και Z=2. Έτσι, τα χαρακτηριστικά των αντικειμένων δεν είναι γνωστά αλλά είναι γνωστές οι σχέσεις τους. Τα αντικείμενα μπορεί να έχουν διαφορετικά χαρακτηριστικά αλλά οι μεταξύ τους σχέσεις παραμένουν ίδιες.

Ένα άλλο χαρακτηριστικό των δεδομένων, σημαντικό για τη μεθοδολογία στην χρηματοοικονομική μοντελοποίηση, είναι το πραγματικό σετ χαρακτηριστικών που χρησιμοποιείται. Μια θεμελιώδης προσέγγιση ανάλυσης χρησιμοποιεί όλα τα διαθέσιμα χαρακτηριστικά αλλά μια προσέγγιση τεχνικής ανάλυσης στηρίζεται μόνο σε χρονοσειρές όπως η τιμή μιας μετοχής και στις παραμέτρους που εξάγονται από αυτή.

Οι πιο δημοφιλείς χρονοσειρές είναι η αξία του δείκτη στο κλείσιμο, η υψηλότερη αξία του δείκτη, η χαμηλότερη αξία του δείκτη και ο όγκος των συναλλαγών καθώς και οι καθυστερημένες αποδόσεις από τις χρονοσειρές ενδιαφέροντος.

Το επόμενο χαρακτηριστικό της συγκεκριμένης μεθοδολογίας είναι η μορφή των σχέσεων μεταξύ των αντικειμένων. Συχνά, είναι δύσκολο να δικαιολογηθεί μια λειτουργική μορφή εξαρχής. Η μεθοδολογία του σχεσιακού Data Mining στα χρηματοοικονομικά δεν υποθέτει κάποια λειτουργική μορφή για τις σχέσεις. Πρόθεσή της είναι η εκμάθηση συμβολικών σχέσεων σε αριθμητικά δεδομένα ή χρηματοοικονομικές χρονοσειρές.

4.2.2 Σχεσιακό Data Mining στα Χρηματοοικονομικά

Οι μέθοδοι δέντρων απόφασης είναι πολύ δημοφιλείς σε εφαρμογές Data Mining γενικά και πιο συγκεκριμένα στα χρηματοοικονομικά. Παρέχουν ένα σύνολο δομημένων κανόνων επεξηγήσιμων από τον άνθρωπο αλλά η ανακάλυψη μικρών δέντρων για περίπλοκα προβλήματα αποτελεί σημαντική πρόκληση στο χώρο των χρηματοοικονομικών. Συμπληρωματικά, οι κανόνες που εξάγονται αδυνατούν να συγκρίνουν δυο αξίες χαρακτηριστικών, γεγονός που συμβαίνει με τις σχεσιακές μεθόδους.

Φαίνεται ότι οι σχεσιακές μέθοδοι του Data Mining κερδίζουν συνεχώς έδαφος σε διάφορα επιστημονικά πεδία (Muggleton 2002). Πιο συγκεκριμένα, στα χρηματοοικονομικά, το Data Mining δεν ακολουθεί απλά αυτή την τάση αλλά ηγείται της εφαρμογής των σχεσιακών μεθόδων για πολυδιάστατες χρονοσειρές όπως είναι οι χρηματιστηριακές. Η άποψη αυτή έχει ενισχυθεί από πληθώρα δημοσιεύσεων που

τοποθετούν το σχεσιακό Data Mining να κινείται προς πιθανολογικούς κανόνες πρώτης-τάξης, έτσι ώστε να αποφευχθούν οι περιορισμοί των ντετερμινιστικών συστημάτων.

Οι σχεσιακές μέθοδοι στα χρηματοοικονομικά όπως η Machine Method for Discovering Regularities (MMDR) (Kovalerchuk & Vityaer, 2000) είναι εξοπλισμένες με πιθανολογικούς μηχανισμούς που είναι απαραίτητοι για χρονοσειρές με μεγάλα επίπεδα θορύβου. Πιο συγκεκριμένα, το MMDR ταιριάζει κατάλληλα σε χρηματοοικονομικές εφαρμογές λόγω της ικανότητας του να χειρίζεται σύνολα δεδομένων με υψηλά επίπεδα θορύβου. Σε υπολογιστικά πειράματα, στρατηγικές συναλλαγών στηριζόμενες στο MMDR συνεχώς ξεπερνούν σε απόδοση άλλες σρατηγικές στηριζόμενες σε διαφορετικές μεθόδους.

4.3 Μοντέλα Data Mining και Εφαρμογές στα Χρηματοοικονομικά

4.3.1 Portfolio Management & Νευρωνικά Δίκτυα

Το Νευρωνικό Δίκτυο που χρησιμοποιείται περισσότερο σε χρηματοοικονομικούς οργανισμούς είναι το Multi-layer Perceptor (MLP) με ένα κρυφό στρώμα κόμβων για πρόβλεψη χρονοσειρών. Η κορυφή των ερευνητικών δραστηριοτήτων στον χώρο της χρηματοοικονομίας αναφορικά με τη χρήση Νευρωνικών Δικτύων ήταν στα μέσα της δεκαετίας του ’90 και κυρίως κάλυπτε τα MLP και τα επαναλαμβανόμενα (recurrent) Νευρωνικά Δίκτυα.

Παρακάτω παρουσιάζονται τα τυπικά βήματα που ακολουθούνται στη διαδικασία του Portfolio Management χρησιμοποιώντας ένα Νευρωνικό Δίκτυο το οποίο προβλέπει τις αξίες απόδοσης.

1. Συγκέντρωση 30-40 ιστορικών θεμελιωδών και τεχνικών παραγόντων για τη μετοχή S_1, για 10-20 έτη.

2. Κατασκευή του Νευρωνικού Δικτύου NN_1 για την πρόβλεψη των αξιών απόδοσης της μετοχής S_1.

3. Επανάληψη των βημάτων 1 και 2 για κάθε μετοχή S_i που παρακολουθείται από τον επενδυτή. Έστω ότι έχουμε 1000 μετοχές και 1000 δίκτυα NN_i.

4. Πρόβλεψη της απόδοσης της μετοχής $S_i(t+k)$ για κάθε μετοχή i και για κ ημέρες μπροστά, υπολογίζοντας το: $NN_i(S_i(t)) = S(t+k)$.

5. Επιλογή των n υψηλότερων $S_i(t+k)$ αξιών των προβλεπόμενων αποδόσεων των μετοχών.

6. Υπολογισμός της συνολικής προβλεπόμενης απόδοσης των επιλεγμένων μετοχών, Τ και υπολογισμός του $S_i(t+k)/\mathrm{T}$. Επένδυση σε κάθε μετοχή ανάλογα με το $S_i(t+k)/\mathrm{T}$.

7. Επαναυπολογισμός των NN_i μοντέλων για κάθε μετοχή i και για κάθε κ ημέρες, προσθέτοντας νέα δεδομένα στο σύνολο εκπαίδευσης. Επανάληψη όλων των βημάτων για την επόμενη ρύθμιση του Porfolio.

Τα παραπάνω βήματα δείχνουν γιατί η χρήση Νευρωνικών Δικτύων έγινε τόσο δημοφιλής στον χρηματοοικονομικό χ.ωρο. Ενδεχομένως όλα τα βήματα να μπορούν να γίνουν αυτόματα, ακόμα και αυτό της πραγματικής επένδυσης. Ακόμα και θεσμικοί επενδυτές μπορεί να μην έχουν διαθέσιμες πηγές να αναλύσουν χειροκίνητα 1000 μετοχές και τα 1000 Νευρωνικά Δίκτυα τους κάθε εβδομάδα. Εάν οι επενδυτικές αποφάσεις λαμβάνονται πιο συχνά, έστω κάθε μέρα, τότε το κίνητρο χρήσης Νευρωνικών Δικτύων με την υψηλή προσαρμοστικότητά τους γίνεται ακόμα πιο προφανές.

4.3.2 Ερμηνεύσιμοι Κανόνες Συναλλαγών & Σχεσιακό Data Mining

Η λογική του Portfolio Management στηριζόμενη στην ανακάλυψη ερμηνεύσιμων κανόνων συναλλαγής είναι η ίδια όπως με τα Νευρωνικά Δίκτυα με την υποκατάσταση του δικτύου με τεχνικές ανακάλυψεις κανόνων. Ανάλογα τις τεχνικές ανακάλυψεις κανόνων, μπορούν να παραχθούν πολύ διαφορετικοί κανόνες. Παρακάτω παρουσιάζονται μερικές κατηγορίες κανόνων.

Οι κατηγορικοί κανόνες προβλέπουν ένα κατηγορικό χαρακτηριστικό, όπως αύξηση/μείωση ή αγορά/πώληση. Ένα τυπικό παράδειγμα μοναδικού κατηγορικού κανόνα δίνεται παρακάτω:

$$If\ \_S_i(t) < \text{Value1}\_\&\ \_S_i(t-2) < Value2\ \_\text{Then}\_\text{S}_i(t+1)\ \_\text{will}\_\text{increase}\ .$$

Στο παράδειγμα αυτό, $S_i(t)$ είναι η συνεχής μεταβλητή, για παράδειγμα η τιμή της μετοχής τη στιγμή t. Εάν το $S_i(t)$ είναι μια διακεκριμένη (discrete) μεταβλητή από όπου λαμβάνονται τα Value1 και Value2, τότε θα έχουμε m διακεκριμένες αξίες. Ο κανόνας αυτός καλείται μοναδικός γιατί συγκρίνει μια μόνο χαρακτηριστική αξία με μια σταθερά. Τέτοιοι κανόνες μπορούν να παραχθούν από εκπαιδευμένα Δέντρα Απόφασης. Δυστυχώς όμως αυτά παράγουν μόνο τέτοιους κανόνες.

Ο ακόλουθος κανόνας τεχνικής ανάλυσης είναι ένας σχεσιακός κατηγορικός κανόνας, γιατί για να καταλήξει σε ένα συμπέρασμα συγκρίνει τις αξίες δυο χαρακτηριστικών όπως την κίνηση των μέσων (moving averages) για 5 και 15 ημέρες (ME5 & ME10) και τα παράγωγα της κίνησης των μέσων για 10 και 30 ημέρες (DerivativeME10 & DerivativeME30).

$$If\ \_ME5(t) = ME15(\text{t})\ \_\&\ \_\text{DerivativeME}10(\text{t}) > 0\ \_\&\ \_\text{DerivativeME}\,30(\text{t}) > 0\ \_\text{then}\_\text{buy}\_\text{stock}\_\text{at}\_\text{moment}\_(\text{t}+1)$$

Ο κανόνας αυτός μπορεί να διαβαστεί ως: εάν η κίνηση των μέσων για 5 και 15 ημέρες είναι ίση και τα παράγωγα για 10 και 30 ημέρες είναι θετικά, τότε αγόρασε τη μετοχή την επόμενη ημέρα.

4.3.3 Ανακάλυψη Ξεπλύματος Χρήματος και Σχεσιακό Data Mining Βασισμένο σε Χαρακτηριστικά

Η εγκληματολογική λογιστική είναι ένα πεδίο που ασχολείται με τις πιθανές παράνομες ή απατηλές χρηματοοικονομικές συνδιαλλαγές. Μια από τις τρέχουσες εστιάσεις του πεδίου αυτού είναι η ανάλυση

των μηχανισμών χρηματοδότησης της τρομοκρατίας όπου καθαρά χρήματα και προϊόντα που προκύπτουν από ξέπλυμα χρήματος χρησιμοποιούνται για πληθώρα δραστηριοτήτων που περιλαμβάνουν την απόκτηση και παραγωγή όπλων και των προδρόμων τους. Σε αντίθεση, οι παραδοσιακές παράνομες εργασίες και η διακίνηση ναρκωτικών κάνουν τα βρώμικα λεφτά να φαίνονται καθαρά.

Οι συγκεκριμένες εργασίες σε μια αυτοματοποιημένη εγκληματολογική λογιστική σχετιζόμενη με το Data Mining είναι η αναγνώριση κακοπροαίρετων και ασυνήθιστων ηλεκτρονικών συναλλαγών και η μείωση του αριθμού των 'λανθασμένα θετικών' (false positives) ύποπτών συνδιαλλαγών. Στην παρούσα, φτηνά, απλά συστήματα βασισμένα σε κανόνες, το προφίλ των πελατών, στατιστικές τεχνικές, Νευρωνικά Δίκτυα και γενετικοί αλγόριθμοι θεωρούνται ως τα κατάλληλα εργαλεία.

Για τις παραδοσιακά παράνομες επιχειρήσεις, υπάρχουν πολλοί δείκτες πιθανών παράνομων ή ασυνήθιστων συνδιαλλαγών. Αυτοί περιλαμβάνουν τη χρήση διαφόρων σχετικών ή και μη-σχετικών λογαριασμών πριν τα χρήματα μεταφερθούν off-shore, την έλλειψη ενός κατόχου του λογαριασμού, τραπεζικές συναλλαγές σε offshore ψεύτικες τράπεζες, μεταφορές συναλλάγματος σε νέα μέρη, συναλλαγές χωρίς αναγνωρίσιμους επαγγελματικούς λόγους.

Κάποιοι από αυτούς τους δείκτες μπορούν εύκολα να ενσωματωθούν σαν κόκκινες σημαίες σε κάποιο λογισμικό. Για κάποιους άλλους είναι πολύ δύσκολο καθώς παράγουν έναν πολύ μεγάλο αριθμό 'ψευδά θετικών' ύποπτων συναλλαγών. Το Data Mining μπορεί να βοηθήσει στην ανακάλυψη μοτίβων απατηλών δραστηριοτήτων που συνδέονται στενά με την τρομοκρατία, όπως συναλλαγές χωρίς κάποιον αναγνωρίσιμο επιχειρηματικό λόγο. Το πρόβλημα είναι ότι συχνά, τέτοιες συναλλαγές δεν αποκαλύπτουν την έλλειψη αναγνωρίσιμου λόγου. Έτσι, τεχνικές Data Mining μπορούν να ψάξουν για ύποπτα μοτίβα στη μορφή πιο περίπλοκων συνδυασμών των συναλλαγών και των αποδείξεων χρησιμοποιώντας προηγούμενη γνώση. Αυτό σημαίνει ότι τα δεδομένα εκπαίδευσης δεν δημιουργούνται μόνο από τις συναλλαγές μεμονωμένα αλλά από τον συνδυασμό δυο ή τριών συναλλαγών. Γεγονός που υπονοεί έναν τεράστιο όγκο αντικειμένων εκπαίδευσης. Το ποσοστό των ύποπτων καταχωρήσεων εντός του συνόλου των συναλλαγών είναι μικρό, αλλά το ποσοστό των ύποπτων συνδυασμών είναι μηδαμινό. Αυτό λοιπόν, αποτελεί τυπική εργασία της τεχνικής αναγνώρισης σπάνιων μοτίβων. Παραδοσιακές μέθοδοι και τεχνικές Data Mining δεν είναι κατάλληλα εξοπλισμένες στο να αντιμετωπίσουν τέτοια προβλήματα. Έτσι οι μέθοδοι του σχεσιακού Data Mining ανοίγουν νέες ευκαιρίες στο συγκεκριμένο πεδίο και περιγράφονται παρακάτω.

Ας θεωρήσουμε ένα σύνολο δεδομένων συναλλαγών που αποτελείται από τα ακόλουθα χαρακτηριστικά:

- Seller
- Buyer
- Item sold
- Item type
- Amount
- Cost
- Date
- Company name
- Type
- Company type

Θα σημειώσουμε κάθε καταχώρηση στο σύνολο δεδομένων σαν (<S>, <B>, <I>), όπου <S>, <B> και <I> είναι σύνολα χαρακτηριστικών για τους seller, buyer και item αντίστοιχα. Μπορεί να έχουμε δυο συνδεδεμένες καταχωρήσεις R1 = (<S1>, <B1>, <I1>) και R2 = (<S2>, <B2>, <I2>), έτσι ώστε ο πρώτος buyer B1 να είναι παράλληλα και ο seller S2, B1 = S2. Είναι επίσης πιθανό το item sold και στις δυο καταχωρήσεις να είναι το ίδιο I1 = I2.

Κατόπιν, φτιάχνουμε ένα καινούριο σύνολο δεδομένων από τις συνδεδεμένες καταχωρήσεις {<R1, R2>}. Οι μέθοδοι Data Mining θα εργαστούν πάνω σε αυτό το σύνολο δεδομένων για να ανακαλύψουν ύποπτες καταχωρήσεις ή ορισμούς φυσιολογικών και ύποπτων μοτίβων. Ακολουθεί μια λίστα με τέτοια μοτίβα:

- Ένα κανονικό μοτίβο (NP) – ένας κατασκευαστής αγοράζει ένα υλικό και πουλάει το αποτέλεσμα της παραγωγής (MBPSR)
- Ένα ύποπτο μοτίβο (SP) – ένας κατασκευαστής αγοράζει ένα υλικό και ξαναπουλάει το ίδιο υλικό (MBPSP)
- Ένα ύποπτο μοτίβο (SP) – μια εμπορική εταιρία αγοράζει ένα υλικό και ξαναπουλάει το ίδιο υλικό φθηνότερα (TBPSPC)
- Ένα κανονικό μοτίβο (NP) – μια εμπορική εταιρία αγοράζει ένα υλικό και πουλάει το αποτέλεσμα της παραγωγής (CBPSR)

Τώρα ένας αλγόριθμος Α, αναλύει τα ζεύγη καταχωρήσεων {<R1, R2>} με, έστω για παράδειγμα, 16 χαρακτηριστικά και μπορεί να συνδέσει ένα ζευγάρι (#5, #6) με ένα κανονικό μοτίβο MBPSR, A(#5, #6), ενώ ένα άλλο ζευγάρι (#1, #3) μπορεί να συνδεθεί με ένα ύποπτο μοτίβο A(#1, #3) = MBPSP.

Εάν έχουν δοθεί ορισμοί για τα ύποπτα μοτίβα, τότε η εύρεση ύποπτών καταχωρήσεων είναι απλά θέμα αποδοτικού υπολογιστικού ψάξιμου σε μια βάση δεδομένων. Αυτό δεν αποτελεί πρόκληση. Εκείνο που αποτελεί τεράστια πρόκληση είναι η αυτόματη παραγωγή περιγραφών για τα μοτίβα/ υποθέσεις. Το συγκεκριμένο θα μπορούσε να πραγματοποιηθεί και χειροκίνητα από τους χρήστες για κάποιο πολύ μικρό σύνολο δεδομένων, γεγονός που δεν υφίσταται εδώ αφού ο αριθμός των καταχωρήσεων, όπως και ο συνδυασμός τους, είναι πολύ μεγάλος.

Μια άλλη προσέγγιση βασισμένη στην ιδέα των αρνητικών μοτίβων μπορεί να βρει τέτοια σχέδια. Σύμφωνα με τη προσέγγιση αυτή, υψηλά πιθανά μοτίβα ανακαλύπτονται και στη συνέχεια αναιρούνται. Υποτίθεται πως ένα τέτοιο μοτίβο είναι κανονικό. Σε ποιο επίσημους όρους, η βασική υπόθεση (ΜΗ) της προσέγγισης αυτής είναι: εάν το Q είναι ένα υψηλά πιθανό μοτίβο (>0.9) τότε το Q αποτελεί ένα κανονικό μοτίβο και το μη (Q) μπορεί να αποτελέσει ένα ύποπτο – ασυνήθιστο μοτίβο.

Παρακάτω θα ακολουθήσει μια ανάλυση του τρόπου λειτουργίας ενός αλγορίθμου που στηρίζεται στην παραπάνω υπόθεση για να βρίσκει ύποπτα μοτίβα. Ο συγκεκριμένος σχεσιακός αλγόριθμος λέγεται MMRD (Machine Method for Discovering Regularities).

Ο αλγόριθμος αυτός που στηρίζεται στη βασική υπόθεση (ΜΗ) αποτελείται από 4 βήματα:

1. Ανακάλυψη μοτίβων, υπολογισμός της πιθανότητας για κάθε μοτίβο, επιλογή μοτίβων με πιθανότητες πάνω από ένα κατώφλι, έστω 0.9. Για να είναι σε θέση να υπολογίσει τις υπό όρους πιθανότητες των μοτίβων, αυτά θα πρέπει να έχουν τη μορφή ενός κανόνα: If A then B.

Τέτοια μοτίβα μπορούν να εξαχθούν με τη χρήση Δέντρων απόφασης για σχετικά απλούς κανόνες και με τη χρήση σχεσιακών αλγορίθμων για πιο πολύπλοκους κανόνες. Τα Νευρωνικά Δίκτυα και οι μέθοδοι παλινδρόμησης συνήθως δεν έχουν το if-κομμάτι. Παρόλα αυτά με μια επιπρόσθετη προσπάθεια τέτοιοι κανόνες μπορούν να εξαχθούν και από Νευρωνικά Δίκτυα ή εξισώσεις παλινδρόμησης.

2. Αναίρεση μοτίβων και υπολογισμός της πιθανότητας για κάθε αρνητικό μοτίβο.

3. Εύρεση καταχωρήσεων στη βάση δεδομένων που ικανοποιούν τα αρνητικά μοτίβα και ανάλυση των καταχωρήσεων εκείνων για πιθανό ψευδή συναγερμό (false alarm), καθώς οι καταχωρήσεις μπορεί να είναι κανονικές και όχι ύποπτες.

4. Αφαίρεση των καταχωρήσεων του 3^{ου} βήματος και παροχή λεπτομερής ανάλυσης για τις ύποπτες καταχωρήσεις.

4.4 Συμπεράσματα

Για να είναι επιτυχημένο ένα έργο Data Mining θα πρέπει να οδηγείται από τις ανάγκες της εφαρμογής και τα αποτελέσματα θα πρέπει να ελέγχονται γρήγορα. Οι χρηματοοικονομικές εφαρμογές παρέχουν ένα μοναδικό περιβάλλον όπου η αποδοτικότητα των μεθόδων μπορεί να ελεγχθεί επί τόπου, όχι μόνο χρησιμοποιώντας παραδοσιακά σύνολα εκπαίδευσης και ελέγχου αλλά κάνοντας πραγματικές προβλέψεις για την κίνηση των μετοχών και ελέγχοντας αυτές την ίδια μέρα. Αυτή η διαδικασία μπορεί να επαναλαμβάνεται καθημερινά για αρκετούς μήνες συλλέγοντας εκτιμήσεις ποιότητας.

Στο παρόν κεφάλαιο υπογραμμίστηκαν κάποια από τα προβλήματα του Data Mining στον χώρο της χρηματοοικονομίας, καθώς και συγκεκριμένες απαιτήσεις για τις μεθόδους που χρησιμοποιούνται. Οι μέθοδοι σχεσιακής ανάλυσης εξελίσσουν τις μεθόδους ανακάλυψης μοτίβων που διαχειρίζονται πολύπλοκα αριθμητικά και μη-αριθμητικά δεδομένα περιλαμβάνουν δομημένα αντικείμενα, κείμενο και δεδομένα μεγάλης ποικιλίας αναφορικά με τα διακριτά ή συνεχή χαρακτηριστικά τους. Όπως δείχτηκε και προηγουμένως, υπάρχουν σαφέστατα πλεονεκτήματα της χρήσης αυτών των μεθόδων στους τομείς της ανάλυσης χρηματιστηριακών συναλλαγών και εγκληματολογικής λογιστικής.

Επι του παρόντος, πειράματα εξόρυξης δεδομένων έχουν καταγραφεί με σημαντική επιτυχία στη βιβλιογραφία. Τυπικά αυτό γίνεται μέσω προσομοιώσεων συναλλαγών και μετέπειτα, συγκρίνοντας τα αποτελέσματα των προβλέψεων με τα αποτελέσματα άλλων μεθόδων. Δεν είναι λίγοι αυτοί που ισχυρίζονται ότι οι μέθοδοι Data Mining επιτυγχάνουν πολύ καλύτερα αποτελέσματα σε σχέση με τις στατιστικές μεθόδους (Huang et al., 2003).

Η μελλοντική κατεύθυνση είναι η δημιουργία πρακτικών συστημάτων υποστήριξης αποφάσεων (Decision Support Systems) που θα κάνουν πολύ πιο εύκολη τη λειτουργία του Data Mining για τις χρηματοοικονομικές εργασίες, όπου εκατοντάδες ή χιλιάδες μοντέλα όπως τα Νευρωνικά Δίκτυα και τα Δέντρα Απόφασης χρειάζεται να αναλύονται και να ρυθμίζονται καθημερινά λόγω της συνεχόμενης ροής δεδομένων. Πιο συγκεκριμένα, αναμένεται μια εκτενής αύξηση των υβριδικών μεθόδων που συνδυάζουν διαφορετικά μοντέλα και παρέχουν πολύ καλύτερη απόδοση συγκριτικά με τη χρήση

μεμονωμένων μοντέλων. Σε μια τέτοια προσέγγιση τα μεμονωμένα μοντέλα μπορούν να αντιμετωπιστούν ως εκπαιδευμένοι τεχνητοί εμπειρογνώμονες. Οπότε συνδυασμός αυτών μπορεί να αντιμετωπιστεί παρόμοια με ένα συμβούλιο αληθινών εμπειρογνωμόνων. Επιπροσθέτως, αυτοί οι τεχνητοί εμπειρογνώμονες, οδηγώντας τα αποτελέσματα σε εκπληκτικά επίπεδα.

Αναμένεται ότι στα επόμενα έτη το Data Mining στον χώρο της χρηματοοικονομίας θα δομηθεί σαν ένα διακριτό πεδίο που συνδυάζει τη γνώση από την χρηματοοικονομία και τη γνώση από το Data Mining, με τρόπο όμοιο με αυτόν της βιοπληροφορικής (bioinformatics) όπου η ενσωμάτωση συγκεκριμένων χαρακτηριστικών του πεδίου και το Data Mining έχουν φτάσει σε ένα πολύ ώριμο σημείο.

5. Το Σύνολο Δεδομένων

Το σύνολο δεδομένων συλλέχθηκε κατά τη διάρκεια μιας ερευνητικής συνεργασίας μεταξύ της Wordline και του Πανεπιστημίου των Βρυξελλών ULB.

Περιλαμβάνει συναλλαγές που πραγματοποιήθηκαν με πιστωτικές κάρτες το Σεπτέβριο του 2013 από ευρωπαίους κατόχους και ενδεχομένως να είναι το μοναδικό σύνολο δεδομένων που υπάρχει στο διαδίκτυο και αφορά πραγματικά, και όχι συνθετικά, δεδομένα. Πιο συγκεκριμένα, περιλαμβάνει συναλλαγές που πραγματοποιήθηκαν στη διάρκει δυο ημερών, όπου υπάρχουν 492 απατηλές συναλλαγές απο το συνολικό μέγεθος των 284.807 συναλλαγών. Το σύνολο δεδομένων είναι εξαιρετικά ανισόρροπο, με τη θετική κλάση (απάτη) να αντιστοιχεί μόνο στο 0,172% επί των συνολικών συναλλαγών.

Το σύνολο δεδομένω περιλαμβάνει μόνο αριθμητικές μεταβλητές, οι οποίες αποτελούν το αποτέλεσμα ενός PCA μετασχηματισμού. Δυστυχώς, για λόγους προσωπικού απορρήτου και εμπιστευτικότητας των στοιχείων, δεν είναι δυνατή η παροχή των αυθεντικών δεδομένων. Έτσι, τα χαρακτηριστικά V1,V2,…,V28 είναι τα βασικά στοιχεία (Principal Components) που αποκτήθηκαν απο την εφαρμογή του PCA. Τα μόνα χαρακτηριστικά που δεν μετασχηματίστηκαν απο τον PCA μετασχηματισμό είναι τα χαρακτηριστικά ‘Time’ και ‘Amount’. Το χαρακτηριστικό ‘Time’ περιλαμβάνει το χρόνο, σε δευτερόλεπτα, που πέρασε μεταξύ κάθε συναλλαγής και της πρώτης συναλλαγής του συνόλου των δεδομένων. Το χαρακτηριστικό ‘Amount’ είναι το χρηματικό ποσό της κάθε συναλλαγής. Τ¨ελος, το χαρακτηριστικό ‘Class’ είναι η στοχευμένη μεταβλητή και παίρνει την αξία 1 στην περίπτωση απάτης και 0 στην περίπτωση των αυθεντικών συναλλαγών.

5.1 Exploratory Data Analysis & PreProcess

Καταρχάς, ελέγχουμε εαν το σύνολο δεδομένων έχει χαμένες αξίες και παρατηρούμε πως δεν παρουσιάζονται χαμένες αξίες. Στη συνέχεια θα δούμε κάποια γενικά στατιστικά χαρακτηριστικά.

```
      Time                V1                   V2                   V3                   V4                   V5
 Min.   :     0   Min.   :-56.40751   Min.   :-72.71573   Min.   :-48.3256   Min.   :-5.68317   Min.   :-113.74331
 1st Qu.: 54202   1st Qu.: -0.92037   1st Qu.: -0.59855   1st Qu.: -0.8904   1st Qu.:-0.84864   1st Qu.:  -0.69160
 Median : 84692   Median :  0.01811   Median :  0.06549   Median :  0.1799   Median :-0.01985   Median :  -0.05434
 Mean   : 94814   Mean   :  0.00000   Mean   :  0.00000   Mean   :  0.0000   Mean   : 0.00000   Mean   :   0.00000
 3rd Qu.:139321   3rd Qu.:  1.31564   3rd Qu.:  0.80372   3rd Qu.:  1.0272   3rd Qu.: 0.74334   3rd Qu.:   0.61193
 Max.   :172792   Max.   :  2.45493   Max.   : 22.05773   Max.   :  9.3826   Max.   :16.87534   Max.   :  34.80167
       V6                  V7                   V8                   V9                  V10                  V11
 Min.   :-26.1605   Min.   :-43.5572   Min.   :-73.21672   Min.   :-13.43407   Min.   :-24.58826   Min.   :-4.79747
 1st Qu.: -0.7683   1st Qu.: -0.5541   1st Qu.: -0.20863   1st Qu.: -0.64310   1st Qu.: -0.53543   1st Qu.:-0.76249
 Median : -0.2742   Median :  0.0401   Median :  0.02236   Median : -0.05143   Median : -0.09292   Median :-0.03276
 Mean   :  0.0000   Mean   :  0.0000   Mean   :  0.00000   Mean   :  0.00000   Mean   :  0.00000   Mean   : 0.00000
 3rd Qu.:  0.3986   3rd Qu.:  0.5704   3rd Qu.:  0.32735   3rd Qu.:  0.59714   3rd Qu.:  0.45392   3rd Qu.: 0.73959
 Max.   : 73.3016   Max.   :120.5895   Max.   : 20.00721   Max.   : 15.59500   Max.   : 23.74514   Max.   :12.01891
       V12                 V13                  V14                  V15                  V16                  V17
 Min.   :-18.6837   Min.   :-5.79188   Min.   :-19.2143   Min.   :-4.49894   Min.   :-14.12985   Min.   :-25.16280
 1st Qu.: -0.4056   1st Qu.:-0.64854   1st Qu.: -0.4256   1st Qu.:-0.58288   1st Qu.: -0.46804   1st Qu.: -0.48375
 Median :  0.1400   Median :-0.01357   Median :  0.0506   Median : 0.04807   Median :  0.06641   Median : -0.06568
 Mean   :  0.0000   Mean   : 0.00000   Mean   :  0.0000   Mean   : 0.00000   Mean   :  0.00000   Mean   :  0.00000
 3rd Qu.:  0.6182   3rd Qu.: 0.66251   3rd Qu.:  0.4931   3rd Qu.: 0.64882   3rd Qu.:  0.52330   3rd Qu.:  0.39968
 Max.   :  7.8484   Max.   : 7.12688   Max.   : 10.5268   Max.   : 8.87774   Max.   : 17.31511   Max.   :  9.25353
       V18                 V19                  V20                  V21                  V22                  V23
 Min.   :-9.498746   Min.   :-7.213527   Min.   :-54.49772   Min.   :-34.83038   Min.   :-10.933144   Min.   :-44.80774
 1st Qu.:-0.498850   1st Qu.:-0.456299   1st Qu.: -0.21172   1st Qu.: -0.22839   1st Qu.: -0.542350   1st Qu.: -0.16185
 Median :-0.003636   Median : 0.003735   Median : -0.06248   Median : -0.02945   Median :  0.006782   Median : -0.01119
 Mean   : 0.000000   Mean   : 0.000000   Mean   :  0.00000   Mean   :  0.00000   Mean   :  0.000000   Mean   :  0.00000
 3rd Qu.: 0.500807   3rd Qu.: 0.458949   3rd Qu.:  0.13304   3rd Qu.:  0.18638   3rd Qu.:  0.528554   3rd Qu.:  0.14764
 Max.   : 5.041069   Max.   : 5.591971   Max.   : 39.42090   Max.   : 27.20284   Max.   : 10.503090   Max.   : 22.52841
       V24                 V25                  V26                  V27                  V28                Amount
 Min.   :-2.83663   Min.   :-10.29540   Min.   :-2.60455   Min.   :-22.565679   Min.   :-15.43008   Min.   :    0.00
 1st Qu.:-0.35459   1st Qu.: -0.31715   1st Qu.:-0.32698   1st Qu.: -0.070840   1st Qu.: -0.05296   1st Qu.:    5.60
 Median : 0.04098   Median :  0.01659   Median :-0.05214   Median :  0.001342   Median :  0.01124   Median :   22.00
 Mean   : 0.00000   Mean   :  0.00000   Mean   : 0.00000   Mean   :  0.000000   Mean   :  0.00000   Mean   :   88.35
 3rd Qu.: 0.43953   3rd Qu.:  0.35072   3rd Qu.: 0.24095   3rd Qu.:  0.091045   3rd Qu.:  0.07828   3rd Qu.:   77.17
 Max.   : 4.58455   Max.   :  7.51959   Max.   : 3.51735   Max.   : 31.612198   Max.   : 33.84781   Max.   :25691.16
    Class
Normal:284315
Fraud :   492
```

Όλες οι καλυμένες μεταβλητές είναι ήδη PCA μετασχηματισμένες, το οποίο σημαίνει ότι είναι ήδη κανονικοποιημένες και έχουν μέση τιμή το 0. Το μόνο που θα κάνουμε είναι να κανονικοποιήσουμε και τα χαρακτηριστικά ‘Time’ και ‘Amount’, ενώ θα μετατρέψουμε τη μεταβλητή ‘Class’ σε factor.

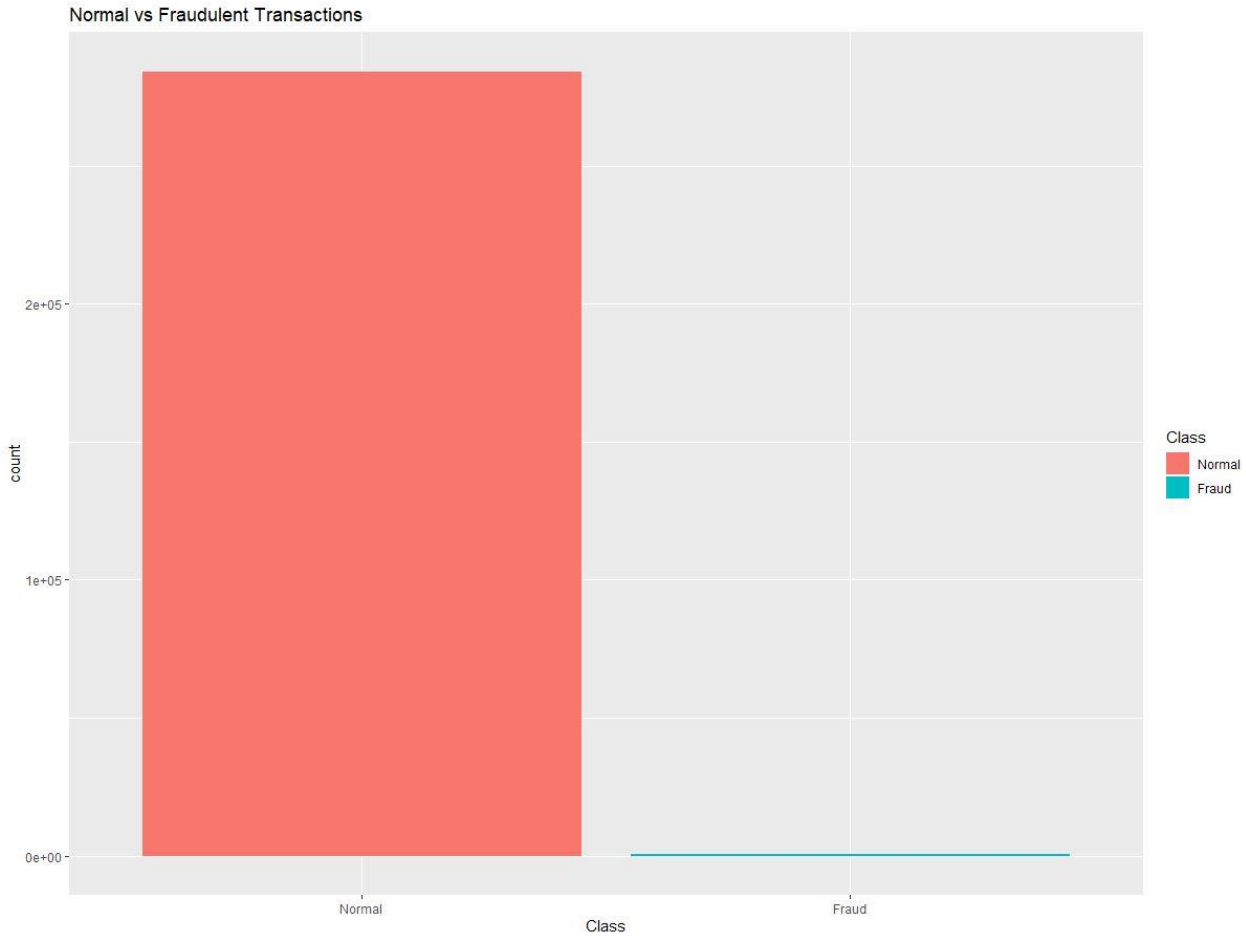
Ας δούμε την ποσοστιαία κατανομή των δυο κλάσεων.

```
           0           1
0.998272514 0.001727486
```

Όντως η θετική κλάση αντιστοιχεί μόνο στο 0,172%, ενω η άλλη κλάση στο 99,827%.

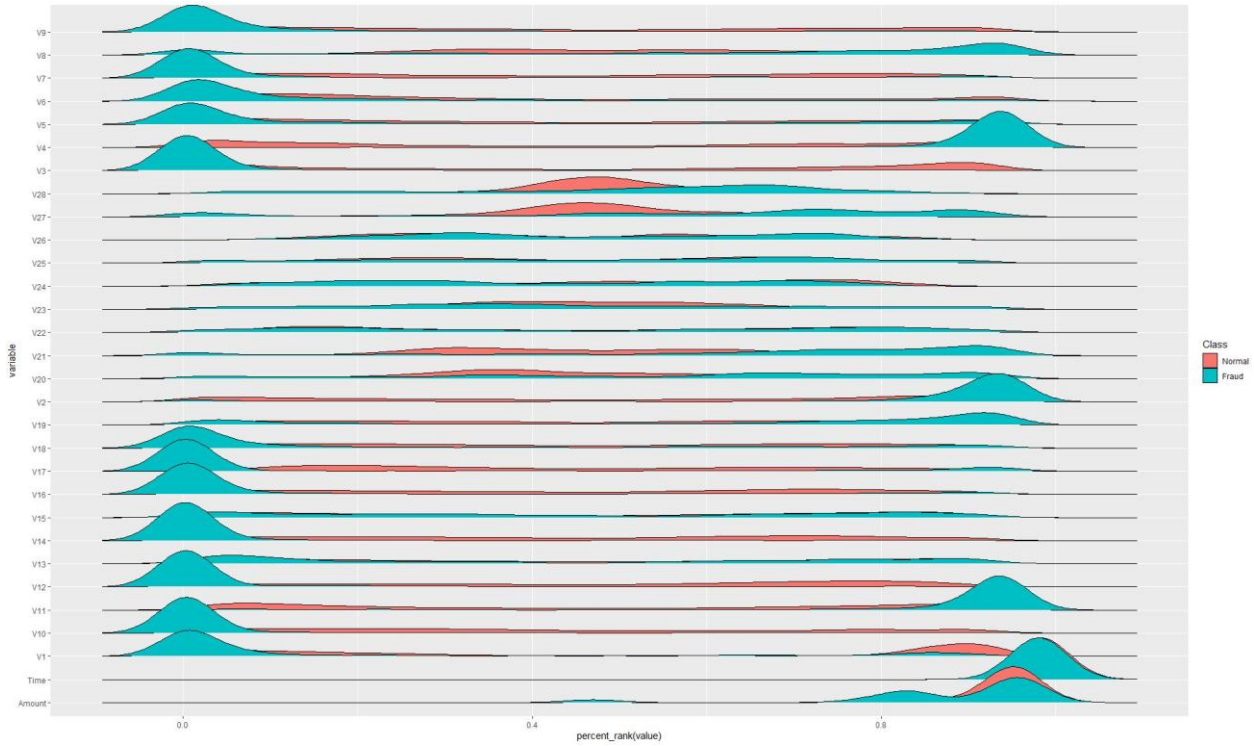
```
Normal  Fraud
284315    492
```

Ας το δούμε και σχηματικά.



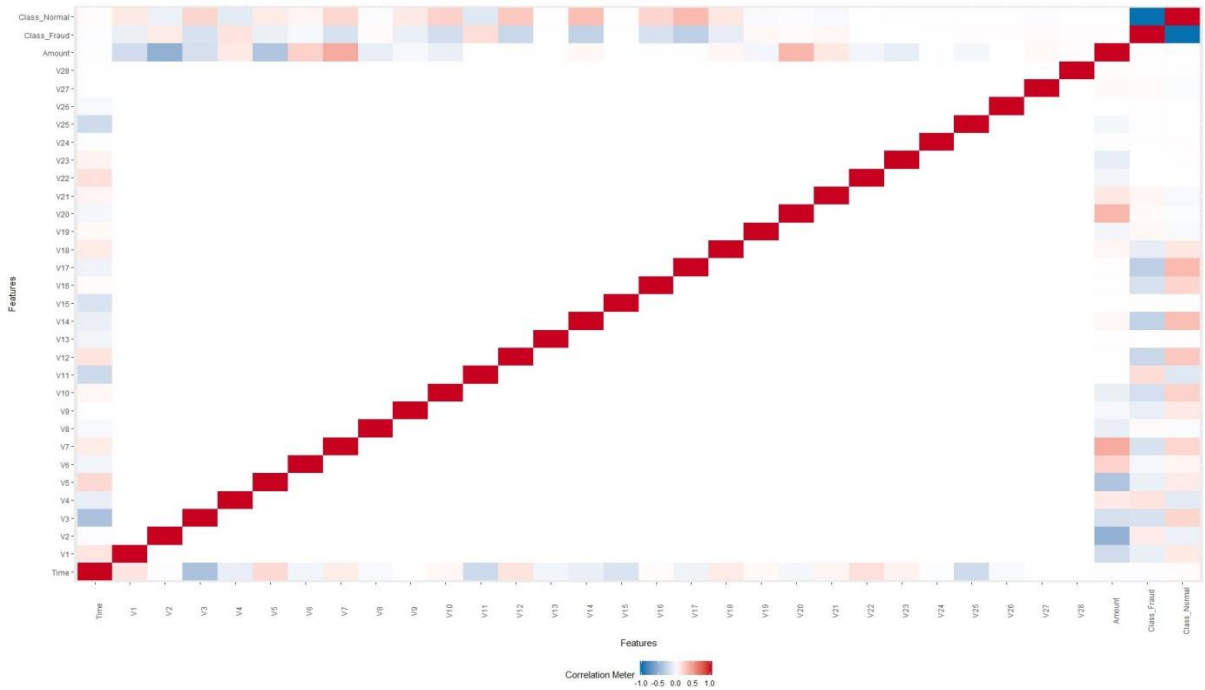
Έτσι, παρατηρούμε και πάλι το μέγεθος της ανισορροπίας των δυο κλάσεων αφού έχουμε 492 απατηλές συναλλαγές και 284.315 κανονικές.

Στη συνέχεια θα οπτικοποιήσουμε τις κατανομές των απατηλών και των κανονικών συναλλαγών.



Παρατηρούμε ότι οι κατανομές των μεταβλητών για τις απατηλές συναλλαγές είναι αρκετά διαφορετικές απο τις κανονικές, εκτός απο τη μεταβλητή 'Time' όπου παρατηρείται σχετικά η ίδια κατανομή.

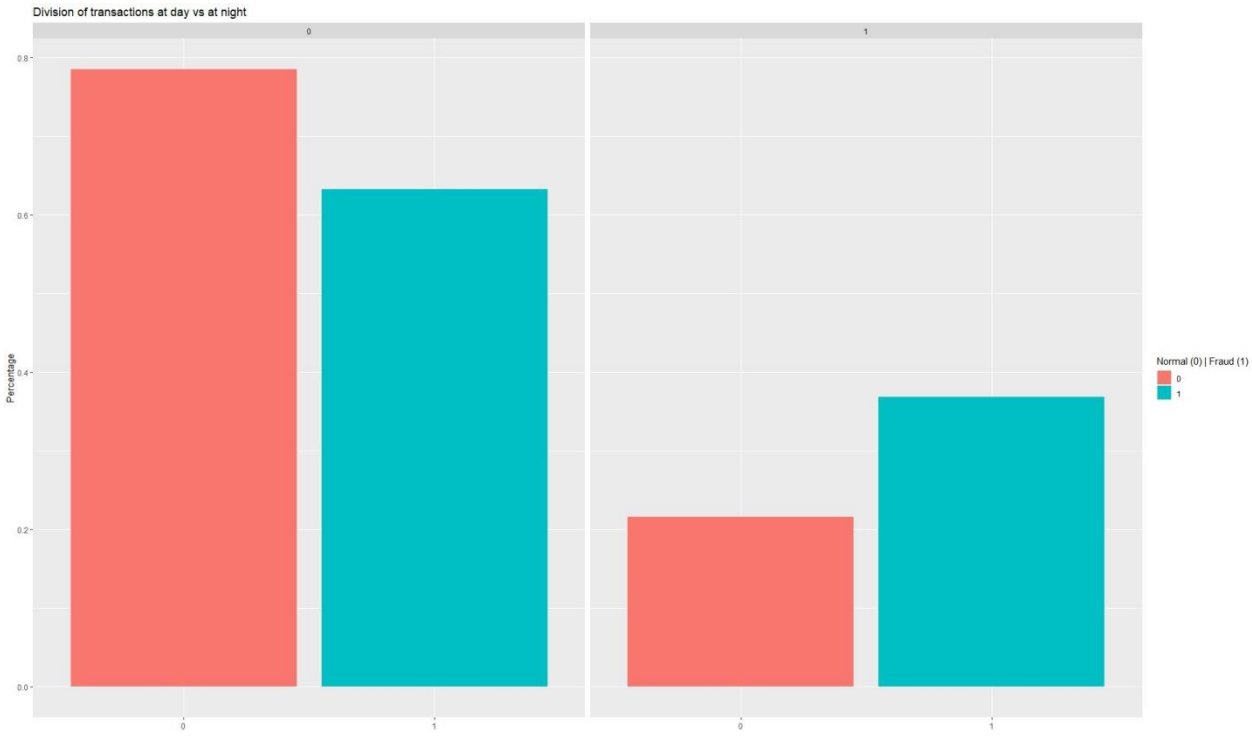
Στη συνέχεια θα κάνουμε ένα διάγραμμα συσχέτισης (correlation plot) για να δούμε ποιές και κατα πόσο μεταβλητές συσχετίζονται μεταξύ τους.



Βλέπουμε ότι καμία απο τις V1,V2,…,V28 μεταβλητές δεν έχει συσχέτιση με τις άλλες. Η μεταβλητή ‘Time’ παρουσιάζει κάποιες συσχετίσεις με κάποιες απο τις V μεταβλητές αλλά καμία με τη μεταβλητή ‘Amount’. Η μεταβλητή ‘Amount’ παρουσιάζει και αυτή κάποιες συσχετίσεις με τις V μεταβλητές αλλά όχι με την ‘Time’. Τέλος, η μεταβλητή που θα μας απασχολήσει, η ‘Class’, παρουσιάζει κάποιες αρνητικές και θετικές συσχετίσεις με τις μεταβλητές V, αλλά δεν φαίνεται να έχει κάποια έντονη συσχέτιση με τις μεταβλητές ‘Time’ και ‘Amount’.

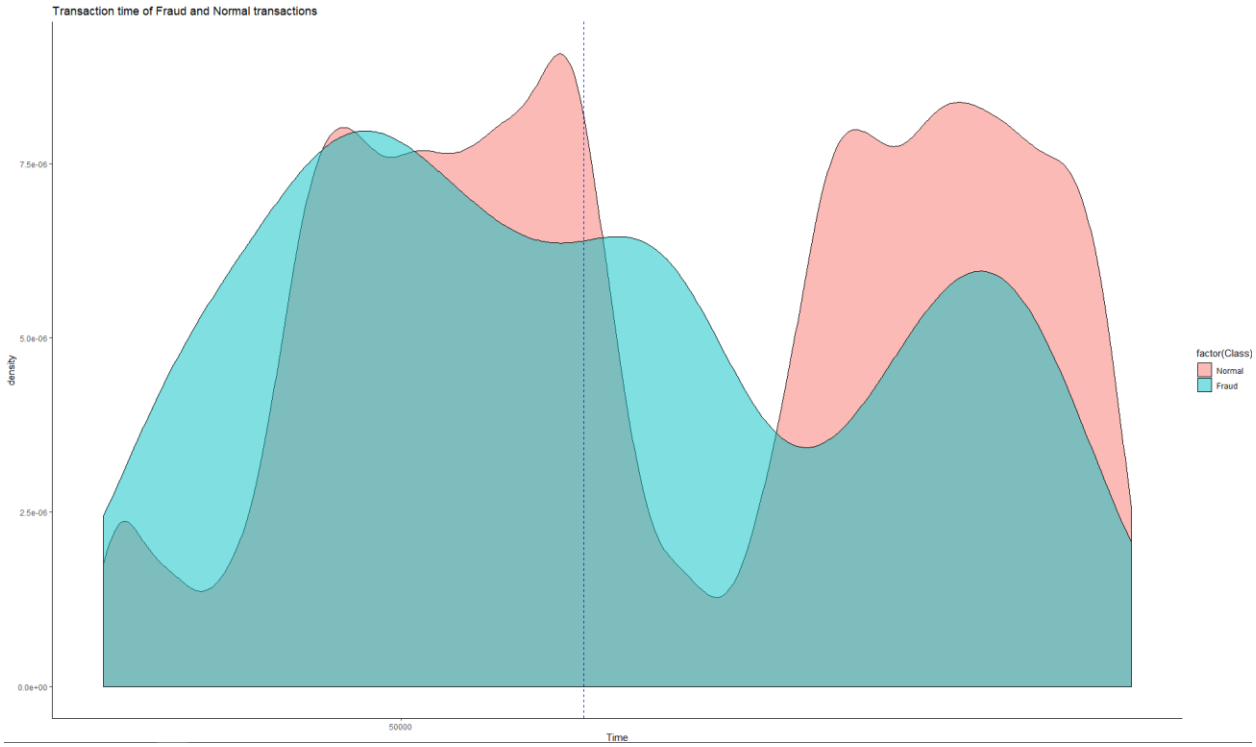
Στη συνέχεια, επείδη όπως παρατηρήσαμε προηγουμένως για την μεταβλητή ‘Time’ οι δυο κλάσεις μας παρουσιάζουν την ίδια κατανομή, θα συνεχίσουμε την διερεύνηση προσπαθώντας να διαπιστώσουμε εαν υπάρχει κάποιο μοτίβο των συναλλαγών (κανονικές και απατηλές) με τη στιγμή της ημέρας, δηλαδή αν παρατηρείται, για παράδειγμα, μεγαλύτερη συχνότητα απατηλών συναλλαγών τη νύχτα ή την ημέρα.

Προκειμένου να γίνει η διερεύνηση αυτή, το αρχικό σύνολο δεδομένων θα χωριστεί σε δυο υποσύνολα ανάλογα με την ετικέτα της εκάστοτε κλάσης, ενω η μεταβλητή ‘Time’ θα μετασχηματιστεί ¨ωστε να απεικονίζει την ώρα της ημέρας.



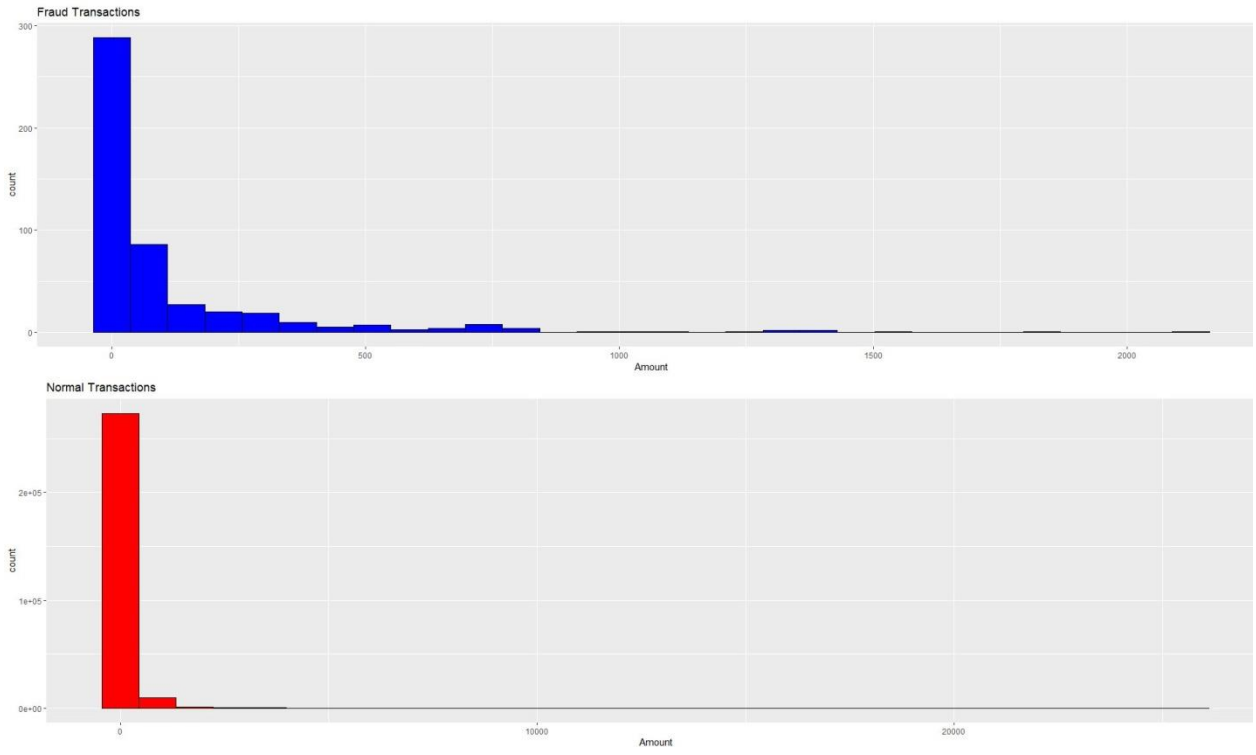
Αυτό που παρατηρείται απο το επάνω διάγραμμα είναι ότι οι κανονικές συναλλαγές είναι πολύ πιο πιθανό να συμβούν κατά τη διάρκεια της ημέρας (γεγονός που κρίνεται φυσιολογικό), ενώ οι απατηλές συναλλαγές είναι πολύ πιο πιθανό να εντοπιστούν κατά τη διάρκεια της νύχτας, συγκριτικά με τις κανονικές. Παρόλα αυτά, οι περισσότερες απατηλές συναλλαγές λαμβάνουν και αυτές μέρος κατά τη διάρκεια της ημέρας.

Συνεχίζοντας, θα δούμε την κατανομή των δυο κλάσεων εντός του χρονικού πλαισίου.



Μια παρατήρηση που θα μπορούσε να γίνει εδώ είναι ότι, οι κανονικές συναλλαγές φαίνεται να ακολουθούν μια κυκλική περιοδικότητα στο βάθος του χρόνου.

Προχωρώντας, θα γίνει μια διερεύνηση της σχέσης των δυο κλάσεων με το μέγεθος των συναλλαγών.



Και εδώ φαίνεται να μην υπάρχει κάποια σαφέστατη ένδειξη διαφορετικότητας μεταξύ των δυο κλάσεων. Σε γενικές γραμμές τα οικονομικά μεγέθη των συναλλαγών και για τις δυο κλάσεις είναι μικρά και κινούνται στα ίδια επίπεδα. Μοναδική εξαίρεση αποτελούν κάποιες περιορισμένες απατηλές συναλλαγές που φαίνεται να αντιστοιχούν σε μεγαλύτερα μεγέθη.

Συνεχίζοντας με την διαδικασία της προεπεξεργασίας πριν την μοντελοποίηση, δεν θα χρειαστεί να γίνουν αρκετά πράγματα καθώς το σύνολο δεδομένων έχει ήδη δεχτεί κάποια προεπεξεργασία λόγω της μεθόδου PCA που ακολουθήθηκε προκειμένου τα δεδομένα να γίνουν ανώνυμα για τον λόγο του προσωπικού απορρήτου που προαναφέρθηκε. Συμπληρωματικά, το σύνολο δεδομένων δεν έχει χαμένες αξίες οπότε το μόνο που θα γίνει είναι μια min-max κανονικοποίηση (normalization) για τις μεταβλητές 'Time' και 'Amount'.

Κατόπιν, για τις ανάγκες της μοντελοποίησης, το αρχικό σύνολο δεδομένων θα χωριστεί σε δύο επιμέρους υποσύνολα, το σύνολο εκπαίδευσης (Training Data set) και το σύνολο ελέγχου (Testing data set). Κάθε μοντέλο θα εκπαιδευτεί πάνω στο σύνολο εκπαίδευσης και ο έλεγχος της απόδοσής του θα γίνει στο σύνολο ελέγχου. Το σύνολο εκπαίδευσης θα αποτελεί το 70% του αρχικού, ενω το σύνολο ελέγχου το 30%. Σε αυτό το σημείο κρίνεται σκόπιμο να αναφερθεί ότι λόγω της ιδαιτερότητας του αρχικού συνόλου δεδομένων (ισχυρή ανισορροπία μεταξύ των δυο κλάσεων), δεν είναι δόκιμο να πραγματοποιηθεί τυχαία δειγματοληψία προκειμένου να κατασκευαστούν τα δυο επιμέρους υποσύνολα καθώς υπα΄ρχει ο κίνδυνος τα μοντέλα να εκπαιδευτούν σχεδόν αποκλειστικά πάνω στα δείγματα της ισχυρά πλειοψηφούσας κλάσης.

Για τον λόγο αυτό, η δειγματοληψία, προκειμένου να παραχθούν τα σύνολα ελέγχου και εκπαίδευσης απο το αρχικό σε ποσοστά 30% και 70% αντίστοιχα, θα γίνει με τη μέθοδο του Stratified Sampling. Η μέθοδος αυτή θα μας επιτρέψει να κάνουμε δειγματοληψία και να δημιουργήσουμε τα δυο υποσύνολα χωρίς να επηρεαστούν οι κατανομές των δυο κλάσεων σε κάθε υποσύνολο. Με άλλα λόγια και στα δυο υποσύνολα η θετική κλάση της απάτης θα αποτελεί το 0,17%.

6. Το πρόβλημα με την ανισορροπία των δεδομένων

Όταν οι κλάσεις του συνόλου δεδομένων βρίσκονται σε ισχυρή ανισορροπία μεταξύ τους, οι συνηθισμένοι αλγόριθμοι Μηχανικής Μάθησης που λειτουργούν μεγιστοποιώντας το ολικό accuracy, τείνουν να ταξινομούν όλες τις παρατηρήσεις σαν παραδείγματα της πλειοψηφικής κλάσης. Αυτό μεταφράζεται σε πολύ χαμηλό accuracy για την μειοψηφική κλάση (χαμηλό recall), η οποία είναι τυπικά και η κλάση ενδιαφέροντος. Η υποβάθμιση της απόδοσης του κατηγοριοποιητή δεν σχετίζεται μόνο με την ισχυρή υπο εκπροσώπηση της μειοψηφικής κλάσης αλλά και απο το γεγονός ότι οι δυο κλάσεις επικαλύπτονται (overlapping distributions). [80] [84]

Δυο δημοφιλείς τρόποι που συναντώνται στη βιβλιογραφία (Blagus & Lusa, 2013) (Dubey et al.,2014) (Zhang , 2014) (Ali et al., 2015) για την αντιμετώπιση του προβλήματος της ανισορροπίας του συνόλου δεδομένων είναι το Resampling και η δημιουργία νέων συνθετικών παραδειγμάτων για την κλάση που υπο εκπροσωπείται. Μια ανάλυση των συγκεκριμένων μεθόδων θα ακολουθήσει παρακάτω.

6.1 Random Under-Sampling

Εδώ το σύνολο δεδομένων εξισορροπείται μειώνοντας το μέγεθος των παραδειγμάτων της πλειοψηφικής κλάσης. Έτσι, κρατώντας όλα τα δείγματα της μειψηφικής κλάσης και τυχαία επιλέγοντας έναν ίσο αριθμό δειγμάτων απο την πλειοψηφική κλάση, παράγεται ένα νέο ισορροπημένο σύνολο δεδομένων που μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση.

6.2 Random Over-Sampling

Με τη μέθοδο αυτή γίνεται τυχαία δειγματοληψία (με αντικατάσταση) της μειοψηφικής κλάσης και παράγονται νέα δείγματα μέχρι η κλάση αυτή να έχει το ίδιο μέγεθος με την πλειοψηφική.

6.3 ROSE (Random Over-Sampling Examples)

Η τεχνική αυτή βασίζεται στη μέθοδο bootstrapκαι παράγει συνθετικά παραδείγματα της μειοψηφικής κλάσης μέχρι το σύνολο δεδομένων να εξισορροπηθεί. Τα συνθετικά παραδείγματα δημιουργούνται με την εκτίμηση της υπο όρους πυκνότητας (conditionaldensity estimate) των δυο κλάσεων.

6.4 SMOTE (Synthetic Minority Over-Sampling Technique)

Η τεχνική αυτή δουλεύει δημιουργώντας συνθετικά παραδείγματα της μειοψηφικής κλάσης μέχρι το σύνολο δεδομένων να εξισορροπηθεί. Ο αλγόριθμος διαλέγει δυο ή περισσότερα παραδείγματα (χρησιμοποιώντας ένα μέτρο απόστασης) και εισαγάγει ένα νέο παράδειγμα για ένα χαρακτηριστικό τη φορά με ένα τυχαίο μέγεθος, σύμφωνα με τη διαφορά των γειτονικών του παραδειγμάτων.

7. Μετρικές αξιολόγησης

Η πιο συχνά χρησιμοποιήσιμη μετρική προκειμένου να αξιολογηθεί η απόδοση ενός κατηγοριοποιητή είναι το Accuracy (ακρίβεια). Τυπικά, το Accuracy ενός προβλεπτικού μοντέλου είναι καλό όταν επιτυγχάνεται Accuracy μεγαλύτερο του 90%. Σε γενικές γραμμές το Accuracy υπολογίζεται απο το πηλίκο των σωστών προβλέψεων προς τον συνολικό αριθμό των προβλέψεων. Η συγκεκριμένη μετρική λειτουργεί καλά, παρόλα αυτά, σε σύνολα δεδομένων με ισχυρή ανισορροπία μεταξύ των κλάσεων, όπως είναι και αυτό της εργασίας, το Accuracy δεν μπορεί να χρησιμοποιηθεί, αφού όταν μια κλάση υποεκπροσωπείται αισθητά, ο κατηγοριοποιητής θα προβλέπει πάντοτε την πιο κοινά εμφανιζόμενη κλάση με ποσοστά επιτυχίας άνω του 99%, γεγονός το οποίο αποτελεί ψευδαίσθηση.

Στο συγκεκριμένο πρόβλημα λοιπόν, κρίνεται αναγκαίο να χρησιμοποιηθούν άλλες μετρικές αξιολόγησης. Στη βιβλιογραφία (Jeni et al., 2013) (Fatourechi et al., 2008) (Jaray and Anbari, 2017) ¨εχουν χρησιμοποιηθεί σε αντίστοιχα προβλήματα και προτείνονται το Kappa Coefficient και το Matthews

Correlation Coefficient. Μετρικές που θα χρησιμοποιηθούν και στην παρούσα εργασία για την αξιολόγηση.

7.1 Kappa Coefficient

Το Kappa Coefficient είναι ουσιαστικά μια μετρική που συγκρίνει την παρατηρήσιμη ακρίβεια (Observed Accuracy) με την αναμενώμενη ακρίβεια (Expected Accuracy). Σε γενικές γραμμές, θεωρείται ότι παρέχει πολύ πιο ποιοτική ακρίβεια στην αξιολόγηση καθώς λαμβάνει υπόψη την πιθανότητα η παρατηρήσιμη ακρίβεια και η αναμενώμενη να έρχονται σε συμφωνία κατά τύχη.

Η μετρική Kappa λαμβάνει τιμές μεταξύ 0 και 1. Η τιμή 1 αντιπροσωπεύει μια τέλεια συμφωνία μεταξύ της παρατηρήσιμης και της αναμενώμενης ακρίβειας, ενώ η τιμή 0 υποδηλώνει ότι δεν υπάρχει καμία συμφωνία.

Ο υπολογισμός της προκύπτει απο τον ακόλουθο τύπο.

$$k = \frac{N\sum_{i=1}^{n} m_{i,i} - \sum_{i=1}^{n}(G_i G_i)}{N^2 - \sum_{i=1}^{n}(G_i G_i)}$$

Όπου:

- i , είναι ο αριθμός των κλάσεων.
- N, ο συνολικός αριθμός των προβλέψεων συγκριτικά με τις πραγματικές τιμές.
- $m_{i,i}$, ο αριθμός των αξιών που ανήκουν στην κλάση i και όντως έχουν κατηγοριοποιηθεί στην κλάση i (δηλαδή οι αξίες που βρίσκονται στη διαγώνια γραμμή μιας μήτρας σύγχισης).
- c_i , είναι ο ολικός αριθμός των προβλεπόμενων αξιών που ανήκουν στην κλάση i .
- G_i , είναι ο αριθμός των πραγματικών αξιών που ανήκουν στην κλάση i .

7.2 Matthews Correlation Coefficient

Η μετρική MCC χρησιμοποιείται σαν ένα μέτρο ποιότητας της κατηγοριοποίησης για δυαδικά προβλήματα, δηλαδή προβλήματα δυο κλάσεων. Ο συντελεστής λαμβάνει υπόψη τα TP,FP,TN και FN και γενικά θεωρείται σαν μια πολύ ισορροπημένη μετρική η οποία μπορεί να χρησιμοποιηθεί και όταν οι δυο κλάσεις έχουν διαφορετικά μεγέθη, όπως και στο πρόβλημα που μας απασχολεί.

Το MCC είναι ουσιαστικά ένας συντελεστής συσχέτισης μεταξύ των προβλεπόμενων και των παρατηρούμενων αξιών της κατηγοριοποίησης. Λαμβάνει τιμές μεταξύ του -1 και +1. Ένας συντελεστής με τιμή το +1 αντιπροσωπεύει μια τέλεια πρόβλεψη. Αν η τιμή του είναι 0, αυτό σημαίνει ότι η πρόβλεψη

του κατηγοριοποιητή δεν είναι καλύτερη απο μια τυχαία πρόβλεψη. Ενώ, όταν η τιμή του είναι -1 υπάρχει ολική διαφωνία μεταξύ της τιμής πρόβλεψης και της πραγματικής.

Ενώ δεν υπάρχει κάποιος τέλειος τρόπος να περιγράψει κανείς τα αποτελέσματα μιας μήτρας σύγχισης με ένα και μόνο νούμερο, η μετρική MCC θεωρείται σαν ένας απο τους καλύτερους.

Μπορεί να υπολογιστεί απευθείας απο τη μήτρα σύγχισης:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Ή εναλλακτικά απο την ακόλουθη σχέση:

$$MCC = \frac{TP/n - \bar{S} \cdot \bar{P}}{\sqrt{\bar{S} \cdot \bar{P}(1 - \bar{S})(1 - \bar{P})}}$$

Όπου:

- $n = TN + TP + FN + FP$
- $\bar{S} = \frac{TP + FN}{n}$
- $\bar{P} = \frac{TP + FP}{n}$

8. Μεθοδολογία

Προκειμένου να ελεγχθούν οι τέσσερις τεχνικές Resampling, το σύνολο δεδομένων χωρίστηκε σε 2 κομμάτια σε ποσοστό αναλογίας 70% για το σύνολο εκπαίδευσης και 30% για το σύνολο ελέγχου. Κατά τη διάρκεια της μοντελοποίησης εφαρμόστηκε κάθε μια απο τις τέσσερις τεχνικές μόνο στο σύνολο εκπαίδευσης. Κατόπιν φτιάχτηκαν συνολικά τέσσερις κατηγοριοποιητές: LDA, Random Forest, SVM και XGBoost. Για κάθε κατηγοριοποιητή εκπαιδεύτηκαν πέντε διαφορετικά μοντέλα. Το πρώτο πάνω στο ανεπεξέργαστο σύνολο εκπαίδευσης και τα υπόλοιπα πάνω στα Resampled σύνολα εκπαίδευσης, ένα για κάθε αντίστοιχη τεχνική: Under-Sampling, Over-Sampling, ROSE και SMOTE. Όλα τα μοντέλα εκπαιδεύτηκαν με 5-fold cross validation και στη συνέχεια ελέγχθηκε η απόδοσή τους πάνω στο ανεπεξέργαστο σύνολο ελέγχου.

9. Αλγόριθμοι και Αποτελέσματα

9.1 Linear Discriminant Analysis

Ο αλγόριθμος αυτός έχει τις ρίζες του σε μια προσέγγιση που αναπτύχθηκε απο τον διάσημο στατιστικολόγο R.A.Fisher. Αυτός ενδιαφερόταν να βρει μια γραμμική προβολή για τα δεδομένα, η οποία να μεγιστοποιεί τη διακύμανση μεταξύ των κλάσεων σύμφωνα με τη διακύμανση των δεδομένων για την ίδια κλάση.

Στην περίπτωση των δυο κλάσεων, αναζητούμε ένα διάνυσμα προβολής α που μπορεί να χρησιμοποιηθεί για τον υπολογισμό των βαθμωτών προβολών $y = a \cdot x$ για τα διανύσματα εισόδου χ . Αυτό αποκτάται υπολογίζοντας τους μέσους της κάθε κλάσης μ_1 και μ_2 . Στη συνέχεια υπολογίζεται η διάσπαρτη μήτρα (scatter matrix) μεταξύ των κλάσεων $S_B = (\mu_2 - \mu_1) \cdot (\mu_2 - \mu_1)^{\mathrm{T}}$. Κατόπιν υπολογίζεται η διάσπαρτη μήτρα εσωτερικά των κλάσεων

$S_w = \sum_{i:C_i=1} (x_i - \mu_1) \cdot (x_i - \mu_1)^{\mathrm{T}} + \sum_{i:C_i=2} (x_i - \mu_2) \cdot (x_i - \mu_2)^{\mathrm{T}}$. Τέλος, μέσω της μεγιστοποίησης του πηλίκου Rayleigh, προκύπτει το ζητούμενο $J(\alpha) = \frac{\alpha^{\mathrm{T}} S_{B^a}}{\alpha^{\mathrm{T}} S_{W^a}}$.

Αποτελέσματα LDA

LDA	ACCURACY	MCC	KAPPA
Original	0.9995	0.8484	0.8459
Downsampling	0.9769	0.2249	0.1082
Oversampling	0.9881	0.2930	0.1838
ROSE	0.9968	0.5114	0.4562
SMOTE	0.9871	0.2824	0.1721

9.2 Random Forest

Ο αλγόριθμος Random Forest αποτελεί μια μέθοδο όπου λαμβάνεται ο μέσος όρος των αποτελεσμάτων πολλαπλών Δ¨εντρων Απόφασης, τα οποία έχουν εκπαιδευτεί σε διαφορετικά μέρη του συνόλου εκπαίδευσης, και έχει ως στόχο τη μείωση της διακύμανσης. Αυτό έρχεται με αντίτιμο μια μικρή αύξηση της προκατάληψης (bias) και μια μικρή απώλεια στην ερμηνεία των αποτελεσμάτων που προκύπτουν απο το μοντέλο. Παρόλα αυτά, τα Random Forest προσδίδουν μια αρκετά μεγάλη αύξηση στην απόδοση του τελικού μοντέλου.

Ο αλγόριθμος εκπαίδευσης για τα Random Forest εφαρμόζει την γενική τεχνική του Bootstrap Aggregation σε δέντρα μάθησης. Δοθέντος ενός συνόλου εκπαίδευσης $\mathrm{X} = \chi_1, ..., \chi_n$ με στοχευμένη μεταβλητή την $\mathrm{Y} = y_1, ..., y_n$, η επαναλαμβανόμενη διαδικασία (bagging) (B φορές) επιλέγει ένα τυχαίο

δείγμα με αντικατάσταση απο το σύνολο εκπαίδευσης και εφαρμόζειτα δέντρα πάνω σε αυτά τα δείγματα.

Έτσι, για $b = 1,...,B$, ο αλγόριθμος λειτουργεί ως εξής:

1. Γίνεται δειγματοληψία με αντικατάσταση για n παραδείγματα εκπαίδευσης απο τα X και Y, τα X_b, Y_b.
2. Εκπαιδεύεται ένα δέντρο κατηγοριοποίησης ή παλινδρόμησης f_b, πάνω στα X_b, Y_b.

Μετά την εκπαίδευση, οι προβλέψεις σε άγνωστα δείγματα X', παράγονται απο τον μέσο όρο των προβλέψεων όλων των μεμονωμένων δέντρων στο X': $\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(X')$ ή λαμβάνοντας την ψήφο της πλειοψηφίας στην περίπτωση της κατηγοριοποίησης.

Η διαδικασία του bootstrapping πδηγεί σε καλύτερη απόδοση του τελικού μοντέλου γιατί μειώνει τη διακύμανση του μοντέλου χωρίς να αυξάνει σημαντικά την προκατάληψη. Αυτό σημαίνει ότι ενώ οι προβλέψεις ενός μεμονωμένου μοντέλου είναι αρκετά ευαίσθητες στον θόρυβο του συνόλου εκπαίδευσης, ο μέσος όρος των αποτελεσμάτων πολλών δέντρων απόφασης δεν είναι, αρκεί τα δέντρα να μην συσχετίζονται μεταξύ τους. Γεγονός που εδώ συμβαίνει λόγω της μεθόδου bootstrapping αφού διαφορετικά δέντρα βλέπουν διαφορετικά σημεία του συνόλου εκπαίδευσης.

Αποτελέσματα Random Forest

Random Forest	ACCURACY	MCC	KAPPA
Original	0.9996	0.8622	0.8605
Downsampling	0.9843	0.2644	0.1494
Oversampling	0.9995	0.8484	0.8459
ROSE	0.9978	0.5848	0.5502
SMOTE	0.5495	0.1063	0.0942

9.3 Support Vector Machine

Σκοπός του αλγόριθμου SVM είναι η εύρεση ενός υπε-επιπέδου σε ένα χώρο N-διαστάσεων (όπου N, ο αριθμός των χαρακτηριστικών) όπου διακεκριμένα κατηγοριοποιούνται τα σημεία των δεδομένων. Προκειμένου να διαχωριστούν τα δείγματα των δυο κλάσεων, υπάρχουν πολλά πιθανά υπερ-επίπεδα που μπορούν να επιλεγούν. Το ζητούμενο είναι να βρεθεί εκείνο το υπερ-επίπεδο που έχει το μέγιστο περιθώριο, δηλαδή την μέγιστη απόσταση μεταξύ των παραδειγμάτων και των δυο κλάσεων. Μεγιστοποιώντας την απόσταση του περιθωρίου (margin distance) παρέχεται πρόσθετη ενίσχυση (reinforcement) έτσι ώστε τα μελλοντικά παραδείγματα να μπορούν να κατηγοριοποιηθούν με μεγαλύτερη αυτοπεποίθηση.

Τα υπερ-επίπεδα είναι όρια απόφασης που βοηθούν την κατηγοριοποίηση των παραδειγμάτων. Τα παραδείγματα που πέφτουν σε κάποια πλευρά του υπερ-επιπέδου αποδίδονται σε διαφορετικές κλάσεις. Επίσης, η διάσταση του υπερ-επιπέδου εξαρτάται απο τον αριθμό των χαρακτηριστικών. Εαν ο αριθμός των χαρακτηριστικών είναι 2, τότε το υπερ-επίπεδο θα είναι μια γραμμή. Εαν ο αριθμός των χαρακτηριστικών είναι 3, τότε το υπερ-επίπεδο γίνεται ένα δισδιάστατο επίπεδο.

Τα διανύσματα στήριξης (support vectors) είναι παραδείγματα που βρίσκονται πιο κοντά στο υπερ-επίπεδο και επηρεάζουν τη θέση και τον προσανατολισμό του υπερ-επιπέδου. Χρησιμοποιώντας αυτά τα διανύσματα στήριξης μεγιστοποιούμε το περιθώριο (margin) του κατηγοριοποιητή. Η συνάρτηση κόστους (loss function) που βοηθάει στη μεγιστοποίηση του περιθωρίου είναι το hinge loss:

$$c(x, y, f(x)) = \begin{Bmatrix} 0, \_if\ \_y * f(x) \geq 1 \\ 1 - y * f(x), \_else \end{Bmatrix}$$

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

Το κόστος είναι μηδέν εάν η προβλεπόμενη αξία είναι ίδια με την πραγματική. Εαν δεν είναι, υπολογίζεται η αξία κόστους. Στη συνέχεια εισάγετε μια παράμετρος κανονικοποίησης. Σκοπός αυτής της παραμέτρου είναι να ισορροπήσει το περιθώριο μεγιστοποίησης και κόστους. Μετά την εισαγωγή της παραμέτρου η συνάρτηση κόστους γίνεται:

$$\min_w \lambda \|w\|^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Στη συνέχεια ο αλγόριθμος παίρνει τα μερικά παράγωγα (partial derivatives) με σεβασμό στα βάρη, προκειμένου να βρει τα gradients. Βρίσκοντας τα gradients ανανεώνονται τα βάρη:

$$\frac{\delta}{\delta w_\kappa} \lambda \|w\|^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_\kappa} (1 - y_i \langle x_i, w \rangle)_+ = \begin{Bmatrix} o, \_if\ \_y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, \_else \end{Bmatrix}$$

Όταν δεν υπάρχει λανθασμένη κατηγοριοποίηση, απλά ανανεώνεται το gradient απο την παράμετρο κανονικοποίησης $w = w - a(2\lambda w)$.

Όταν υπάρχει λανθασμένη κατηγοριοποίηση, συνυπολογίζεται και το κόστος μαζί με την παράμετρο κανονικοποίησης προκειμένου να ανανεωθεί το gradient $w = w + a(y_i x_i - 2\lambda w)$.

Αποτελέσματα SVM

SVM	ACCURACY	MCC	KAPPA
Original	0.9991	0.7038	0.6940
Downsampling	0.9538	0.1544	0.0543
Oversampling	0.9920	0.3431	0.2470
ROSE	0.9940	0.3975	0.3093
SMOTE	0.9725	0.2011	0.0898

9.4 XGBoost

Ο XGBoost είναι ένας κατηγοριοποιητής ensemble, όπως και ο Random Forest. Κατασκευάζει και αυτός ένα σύνολο δέντρων απόφασης χρησιμοποιώντας την δομή του Gradient Boosting. Το Gradient Boosting αποτελεί μια ειδική περίπτωση του boosting ¨οπου τα σφάλματα ελαχιστοποιούνται με τη χρήση του Gradient Descent αλγορίθμου. Ο αλγόριθμος αυτός χρησιμοποιεί μια συνάρτηση κόστους που πρακτικά επισημαίνει πόσο καλό είναι το μοντέλο στις προβλέψεις για ένα σθγκεκριμένο σετ παραμέτρων. Η συνάρτηση κόστους έχει τη δική της καμπύλη και τα δικά της gradients. Η κλίση της καμπύλης δείχνει πως πρέπει να ανανεωθούν οι παράμετροι προκειμένου να επιτευχθεί μεγαλύτερη ακρίβεια στο μοντέλο.

Υπάρχουν δυο παράμετροι στη συνάρτηση κόστους που μπορούν να ρυθμιστούν: το βάρος (m) και η προκατάληψη (b). Εφόσον πρέπει να εκιμηθεί η επίδραση της κάθεμιας παραμέτρου στην τελική πρόβλεψη, θα πρέπει να χρησιμοποιηθούν τα μερικά παράγωγα, τα οποία υπολογίζονται σύμφωνα με την κάθε παράμετρο και αποθηκεύονται σε ένα gradient.

Δοθέντος της συνάρτησης κόστους λοιπόν:

$$f(m,b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - mx_i + b))^2$$

Το gradient υπολογίζεται ως:

$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

Στη συνέχεια, για να λύσουμε ως προς το gradient ο αλγόριθμος ξαναδιατρέχει τα παραδείγματα χρησιμοποιώντας τις καινούριες τιμές για τα m,b και υπολογίζονται τα μερικά παράγωγα. Το καινούριο

gradient δείχνει την κλίση της συνάρτησης κόστους, σύμφωνα με τις υπάρχουσες παραμέτρους, και την κατεύθυνση στην οποία πρέπει να κινηθούμε για να ανανεωθούν οι παράμετροι. Το μέγεθος της ανανέωσης ελέγχεται απο το ρυθμό εκμάθησης.

Αποτελέσματα XGBoost

XGBoost	ACCURACY	MCC	KAPPA
Original	0.9996	0.8730	0.8682
Downsampling	0.9625	0.1769	0.0686
Oversampling	0.9984	0.6539	0.6308
ROSE	0.9986	0.6797	0.6660
SMOTE	0.9787	0.2344	0.1168

10. Ανάλυση αποτελεσμάτων

Τα αποτελέσματα των μοντελοποιήσεων μετά την εφαρμογή των τεσσάρων διαφορετικών τεχνικών Resampling σε κάθε αλγόριθμο δείχνουν να υπολείπονται αισθητά σε σχέση με την αρχική μοντελοποίηση χωρίς Resampling. Παρατηρούμε ότι το Accuracy όλων των μοντέλων είναι εξαιρετικό (> 97%), παρόλα αυτά οι μετρικές MCC και Kappa δίνουν μια τελείως διαφορετική εικόνα. Τώρα από τους τέσσερις αρχικούς κατηγοριοποιητές, καλύτερα αποτελέσματα δίνει ο XGBoost, ακολουθούν οι Random Forest και LDA με αρκετά όμοιες αποδόσεις, ενώ ο SVM υπολείπεται αρκετά.

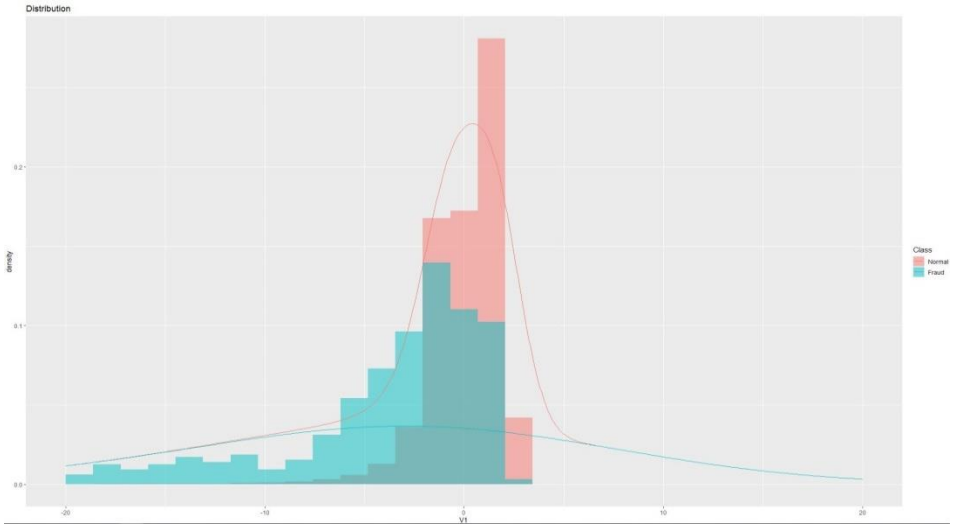
	LDA	Random Forest	SVM	XGBoost
MCC	0.8484	0.8622	0.7038	0.8730
KAPPA	0.8459	0.8605	0.6940	0.8682

Μια έρευνα στην περιορισμένη βιβλιογραφία για το ζήτημα της αισθητής μείωσης της απόδοσης των κατηγοριοποιητών μετά την εφαρμογή των τεχνικών Resampling, φανερώνει ότι η ισχυρή ανισορροπία μεταξύ των δυο κλάσεων του συνόλου δεδομένων δεν είναι ο μόνος παράγοντας που καθιστά δύσκολο το έργο της κατηγοριοποίησης των κλάσεων. Ένας άλλος παράγοντας με, ενδεχωμένως, μεγαλύτερη επιρροή είναι η περίπτωση που οι κλάσεις ενδιαφέροντος παρουσιάζουν επικαλυπτόμενες κατανομές (overlapping distributions).

Μάλιστα, ο Pratti (Pratti et al., 2004) έδειξε ότι η ανισορροπία των κλάσεων απο μόνη της δεν φαίνεται να αποτελεί σημαντικό πρόβλημα. Οι περισσότερες μελέτες (Domingos, 1999) (Japkowicz and Stephen, 2002)(Pratti et al., 2004)(Batista et al., 2004) προτείνουν μεθόδους που φαίνεται να λειτουργούν καλά κάτω απο συγκεκριμένες συνθήκες, παρόλα αυτά δεν φαίνεται να υπάρχουν εμπειρικές αποδείξεις ότι κάποια τεχνική υπερτερεί σε έναν γενικό βαθμό έναντι των άλλων. Σε γενικές γραμμές, δεν υπάρχει κάποια 'καλύτερη' μέθοδος, βέβαια σε κάποιες περιπτώσεις κάποιες τεχνικές είναι καλύτερες απο τις άλλες (no-free-lunch Theorem).

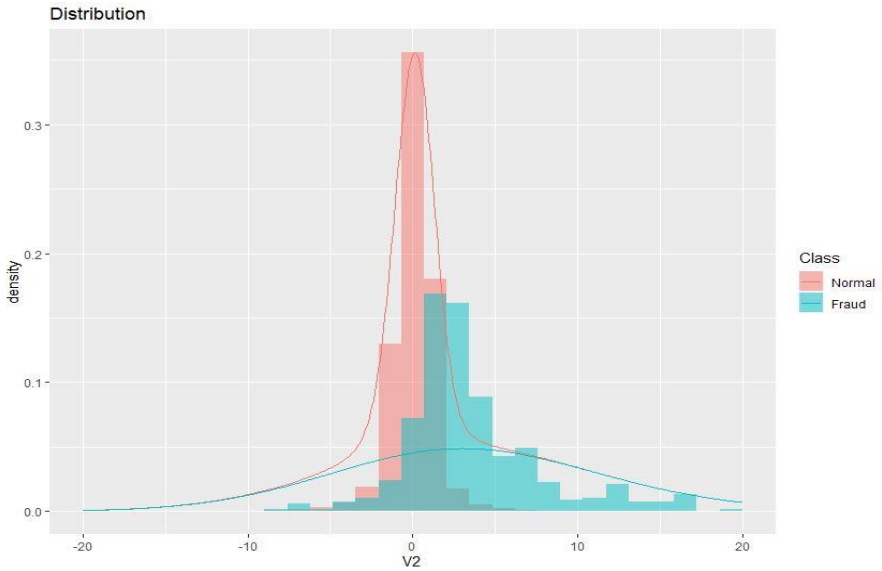
Στο κεφάλαιο αυτό, θα εξεταστεί λεπτομερώς το αν υφίσταται το γεγονός της ύπαρξης επικαλυπτόμενων κατανομών μεταξύ των δυο κλάσεων ‘Fraud’ και ‘Normal’. Ο έλεγχος αυτός θα πραγματοποιηθεί μέσω ιστογραμμάτων για κάθε ένα χαρακτηριστικο V1,V2,…,V28 σε σχέση με τις δυο κλάσεις.

V1



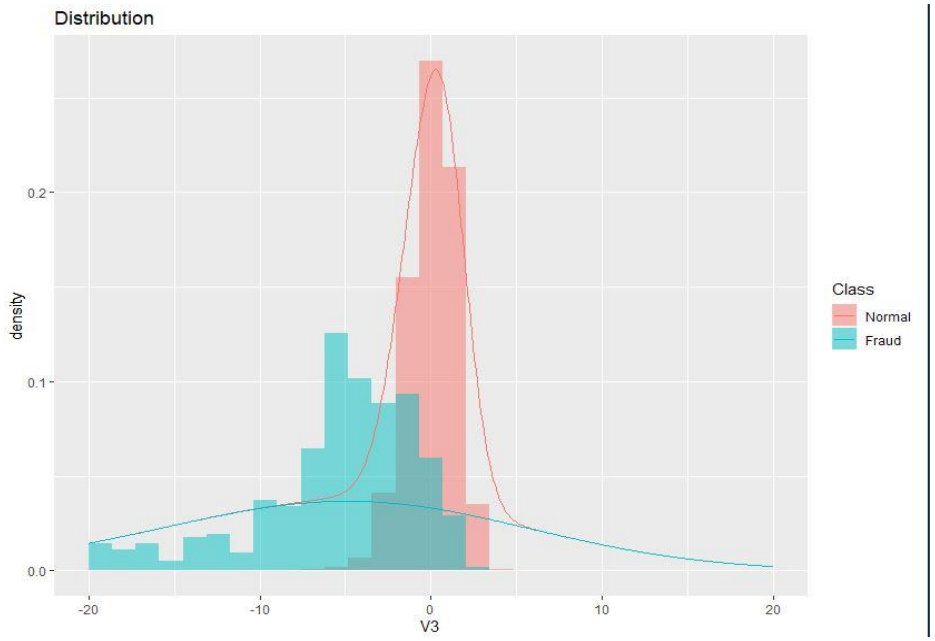
Παρατηρείται κάποια επικάλυψη της κατανομής της V1 για τις δυο κλάσεις.

V2



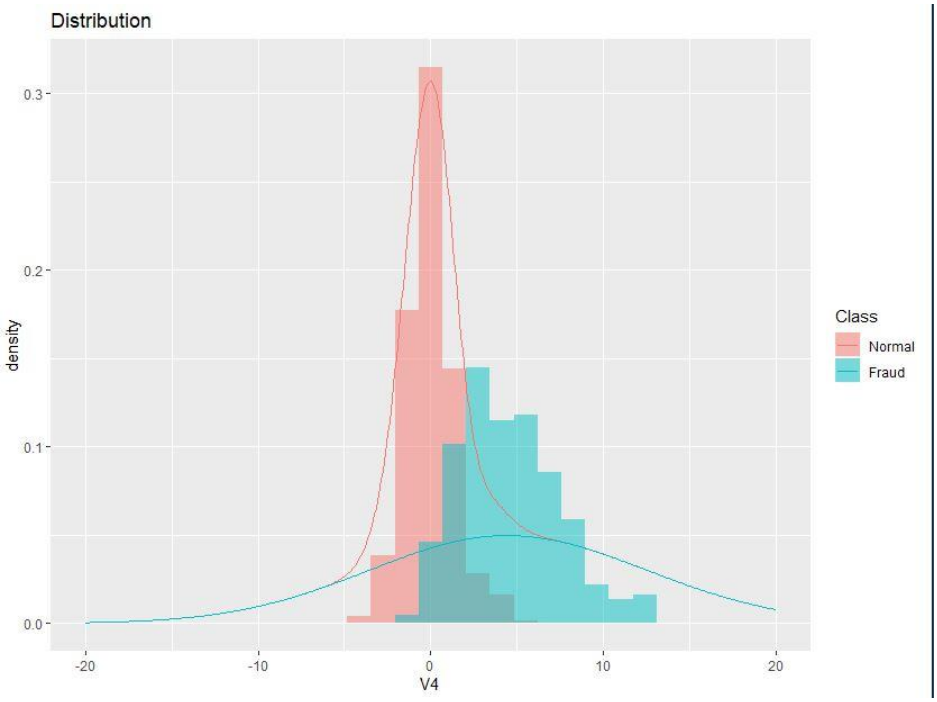
Παρατηρείται αρκετή επικάλυψη της κατανομής της V2 για τις δυο κλάσεις.

V3



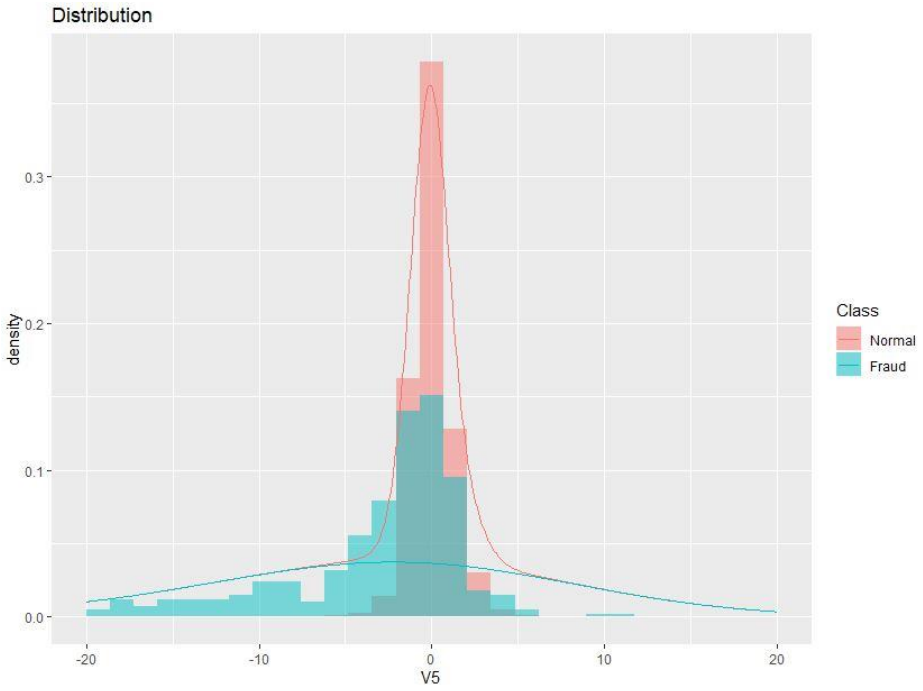
Παρατηρείται κάποια επικάλυψη της κατανομής της V3 για τις δυο κλάσεις.

V4



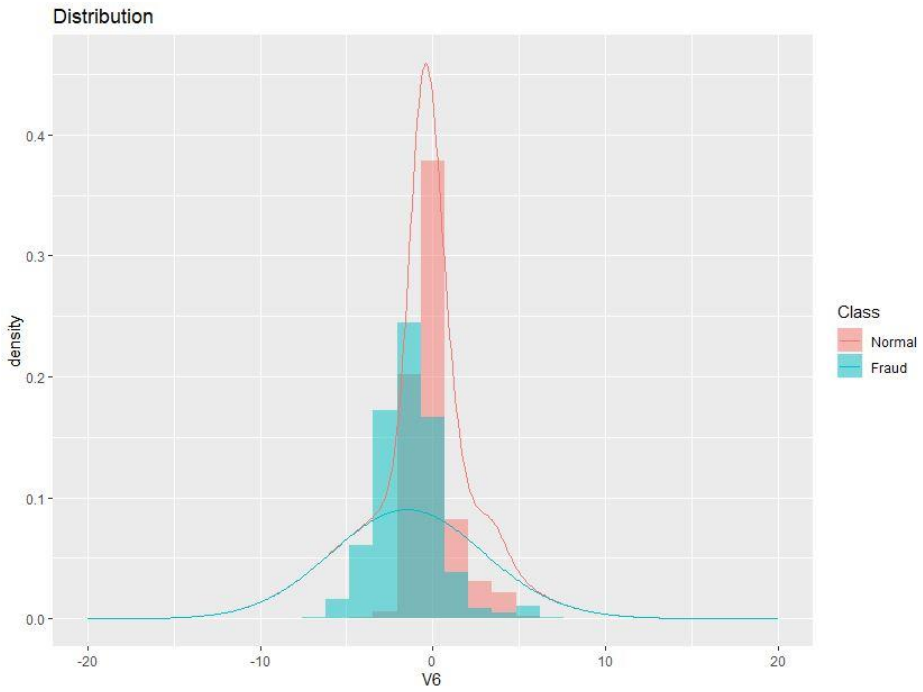
Η κατανομή δείχνει να μοιάζει αρκετά, ενω υπάρχει επικάλυψη.

V5



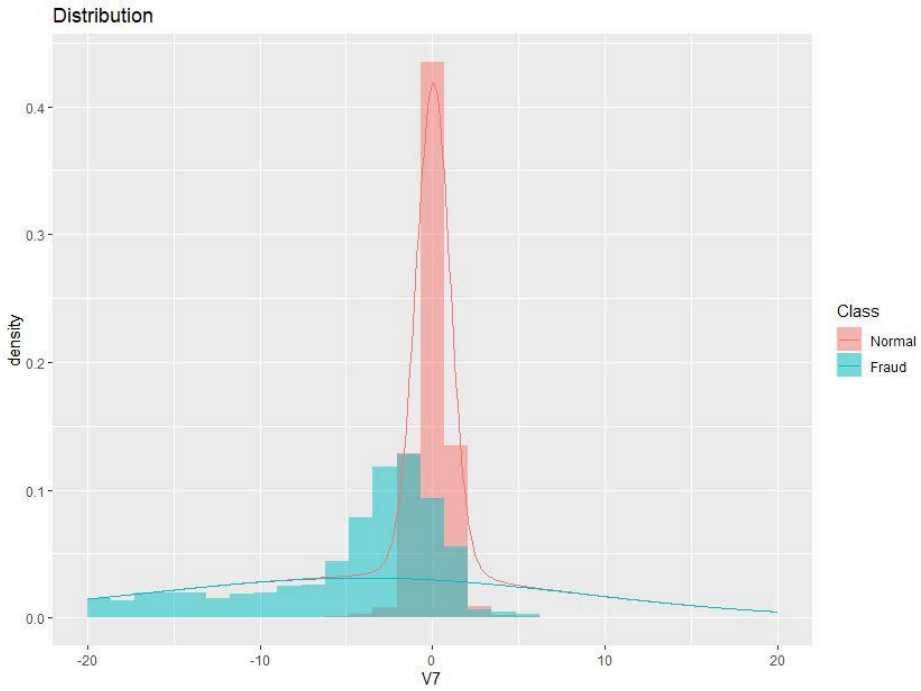
Η κατανομή δείχνει να μοιάζει αρκετά, ενω υπάρχει μεγάλη επικάλυψη.

V6



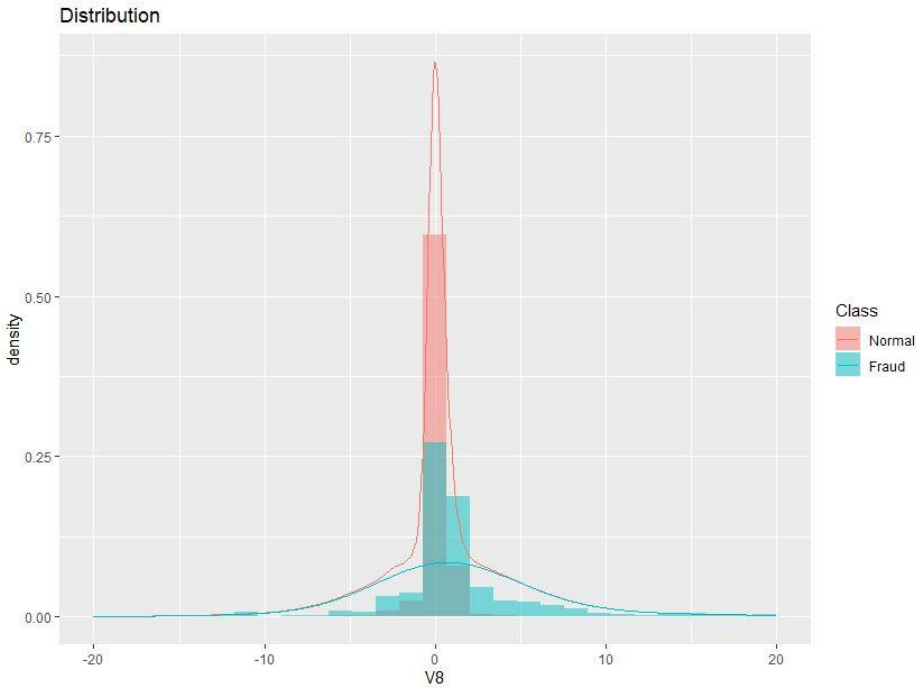
Η κατανομή δείχνει να μοιάζει αρκετά, ενω υπάρχει μεγάλη επικάλυψη.

V7



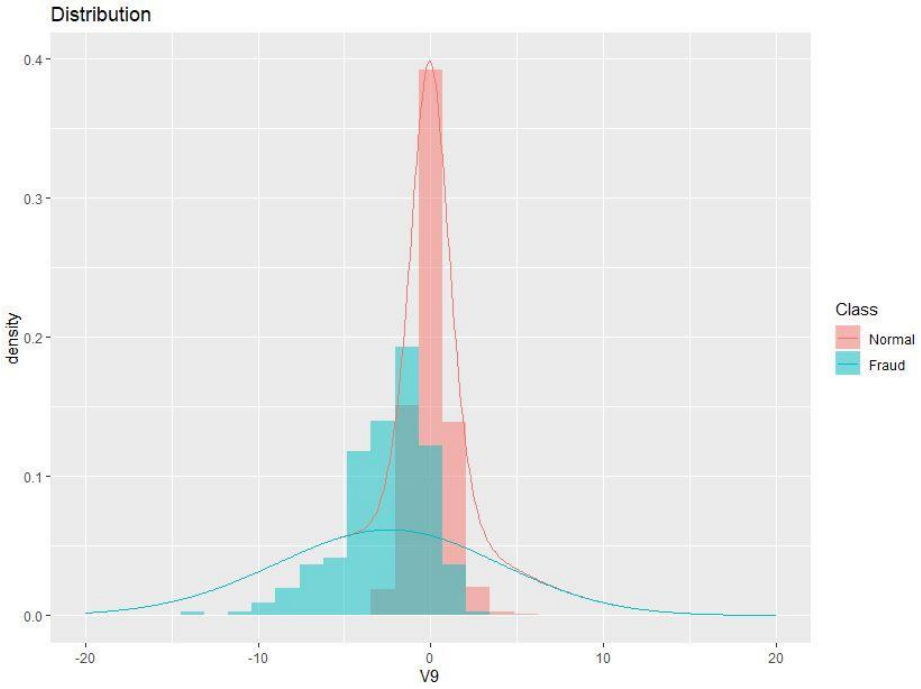
Η κατανομή δείχνει να μοιάζει αρκετά, ενω υπάρχει μεγάλη επικάλυψη.

V8



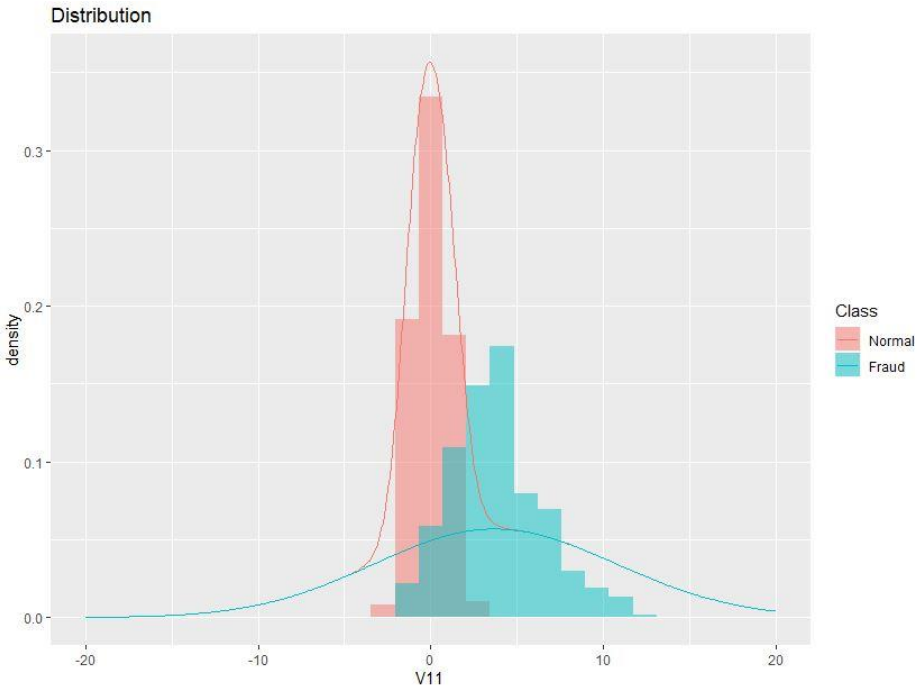
Η κατανομή είναι αρκετά όμοια, ενω φαίνεται σε συγκεκριμένα σημεία να υπάρχει σχεδόν πλήρης επικάλυψη.

V9



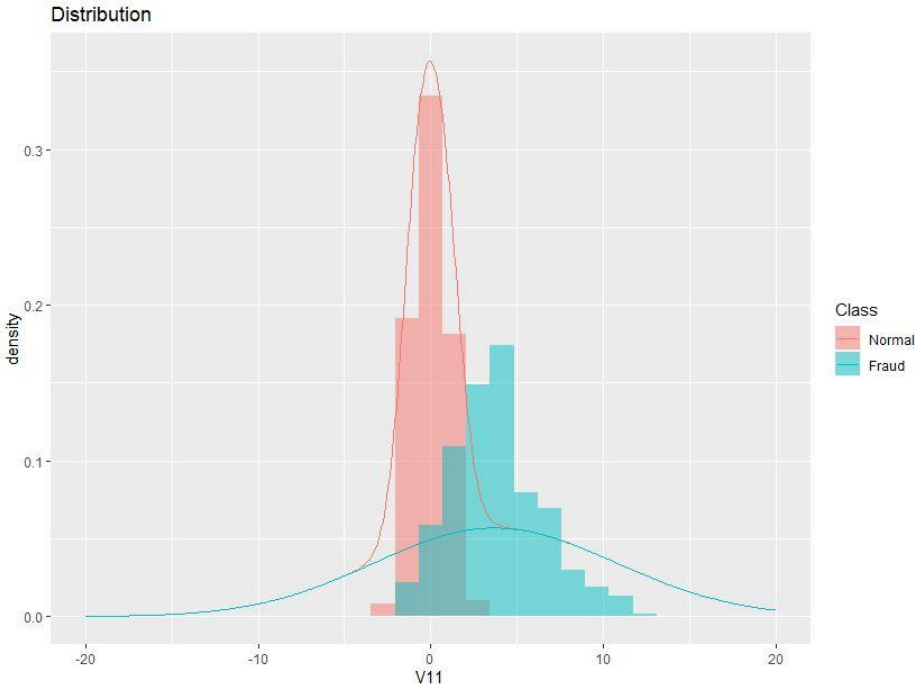
Η κατανομή είναι αρκετά όμοια, ενω φαίνεται να υπάρχει αρκετή επικάλυψη.

V10



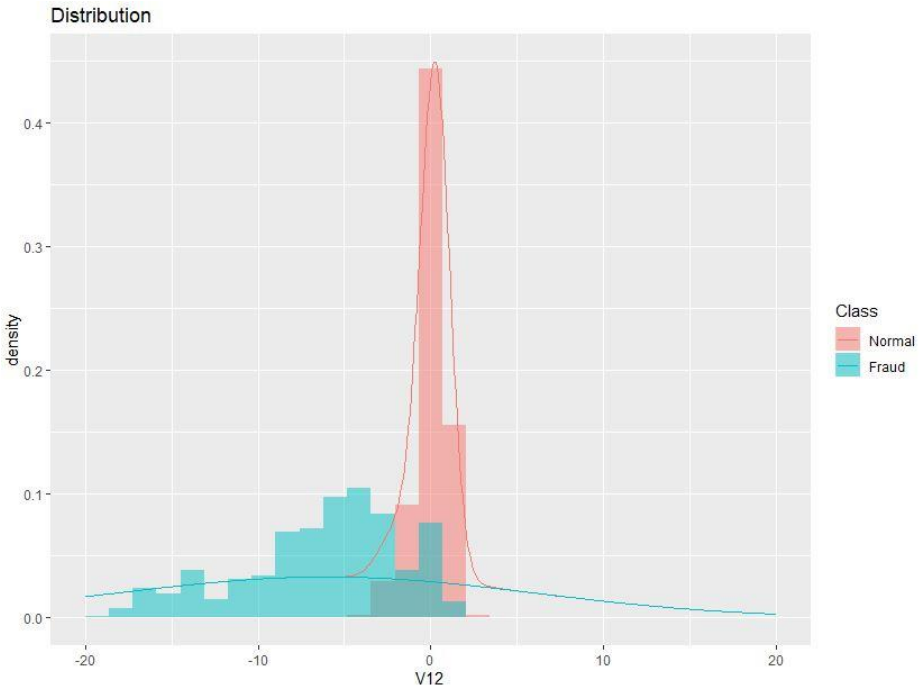
Η κατανομή είναι αρκετά όμοια, ενω φαίνεται να υπάρχει κάποια επικάλυψη.

V11



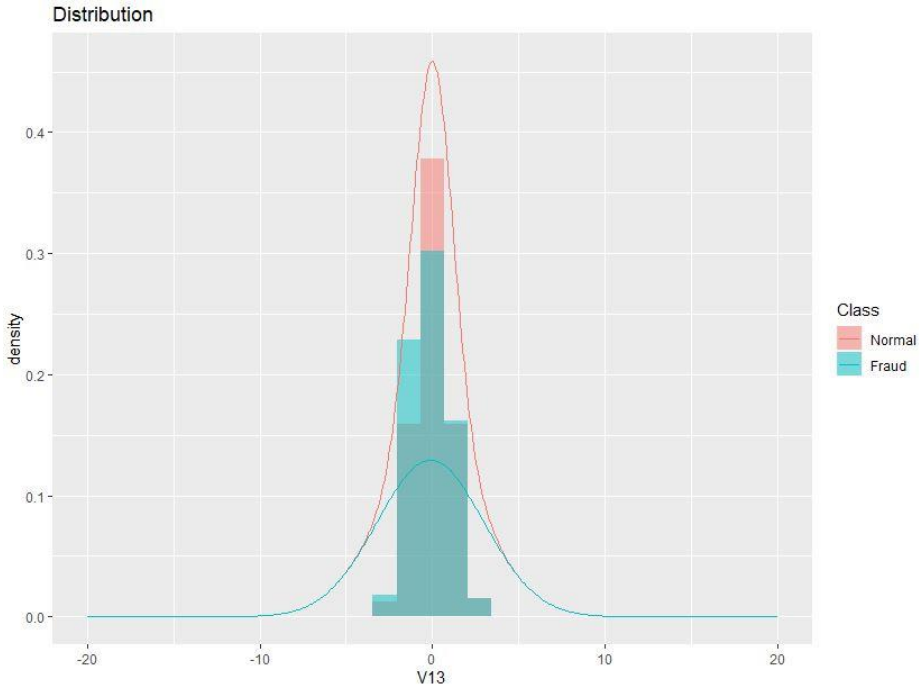
Η κατανομή φαίνεται διαφορετική εν μέρει, ενω υπάρχει επικάλυψη.

V12



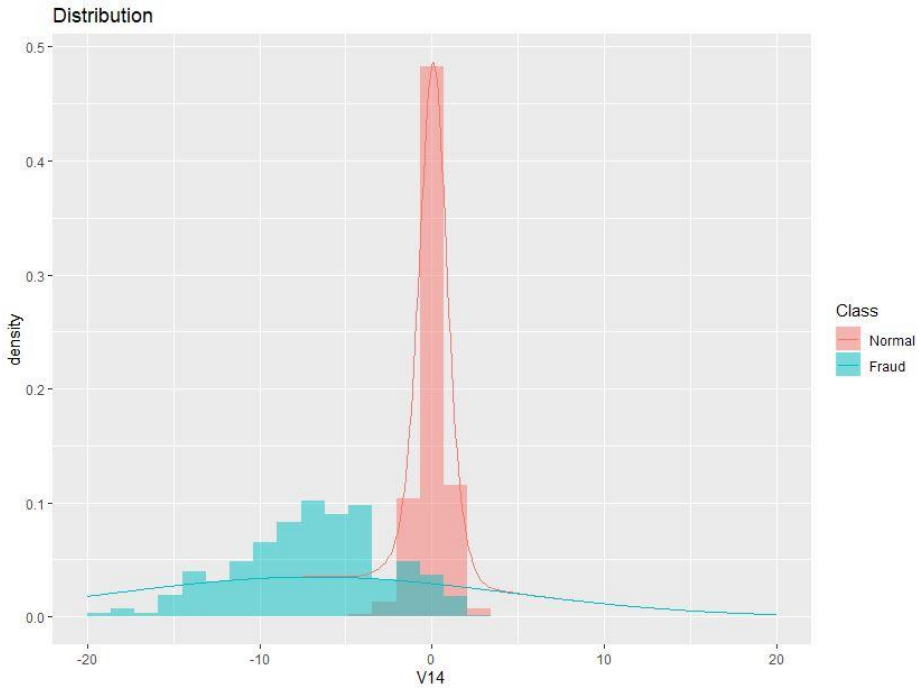
Η κατανομή μοιάζει, ενω υπάρχει επικάλυψη.

V13



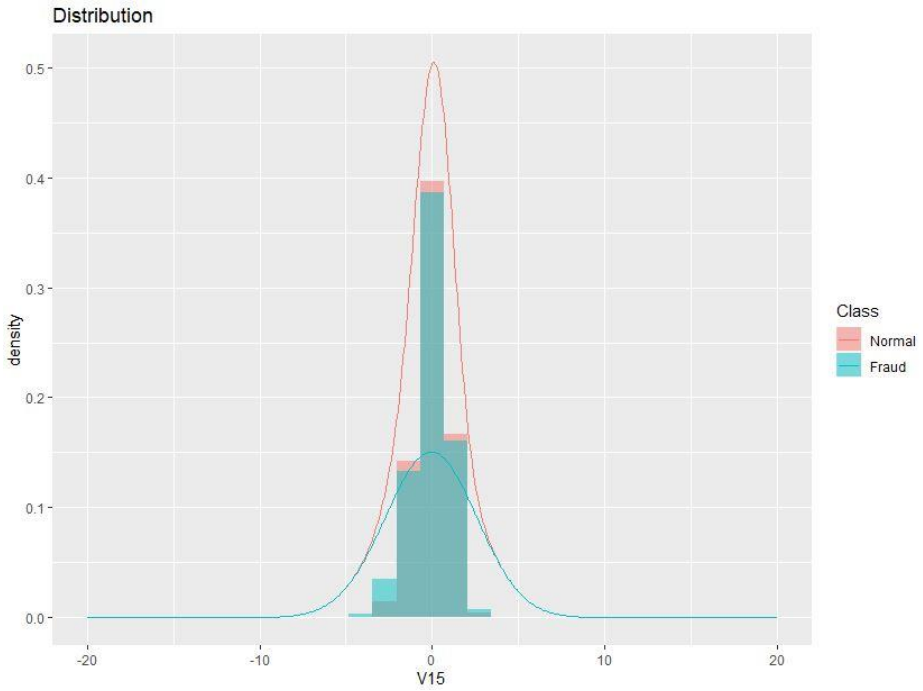
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V14



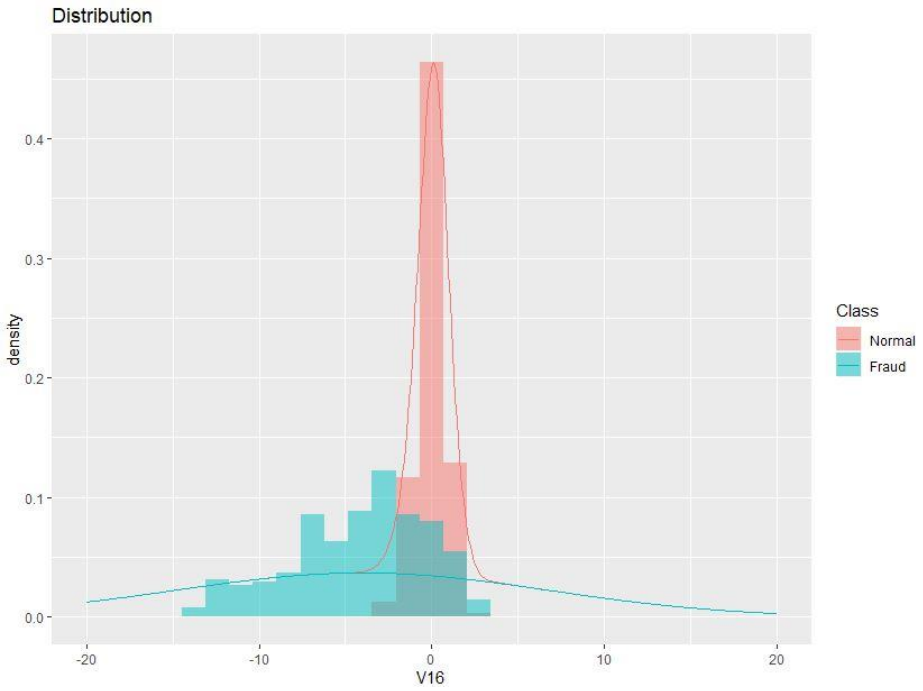
Εδώ οι κατανομές είναι διαφορετικές για κάθε κλάση, ενω δεν φαίνεται να παρουσιάζεται ιδιαίτερη επικάλυψη.

V15



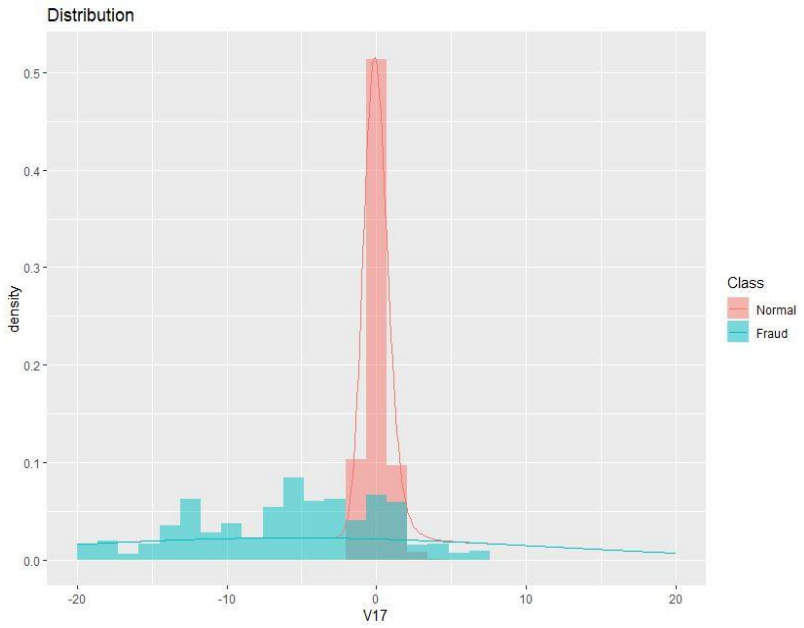
Η κατανομή είναι σχεδόν ίδια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V16



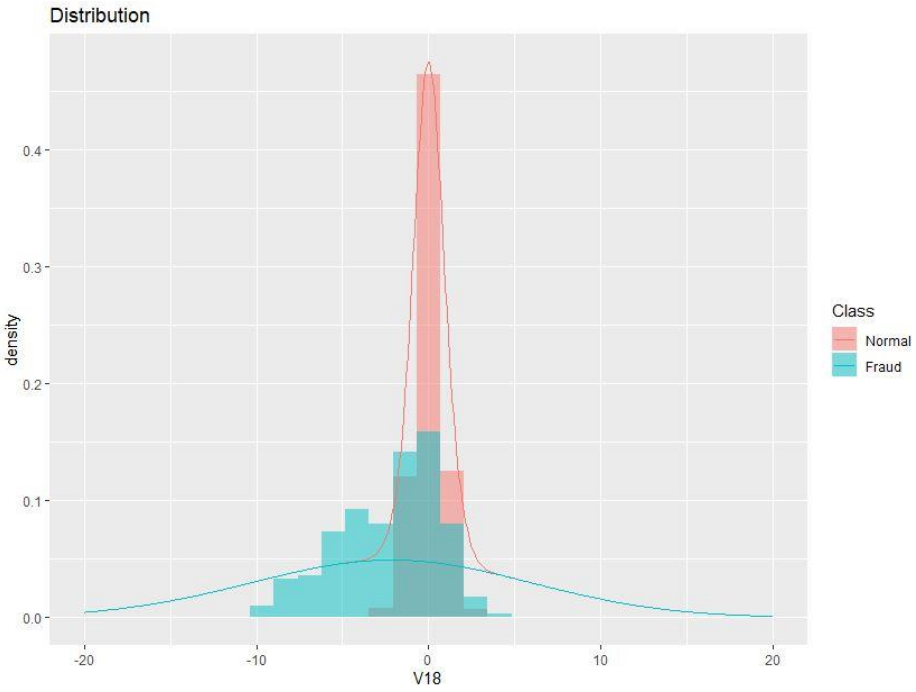
Εδώ οι κατανομές είναι διαφορετικές για κάθε κλάση, ενω δεν φαίνεται να παρουσιάζεται ιδιαίτερη επικάλυψη.

V17



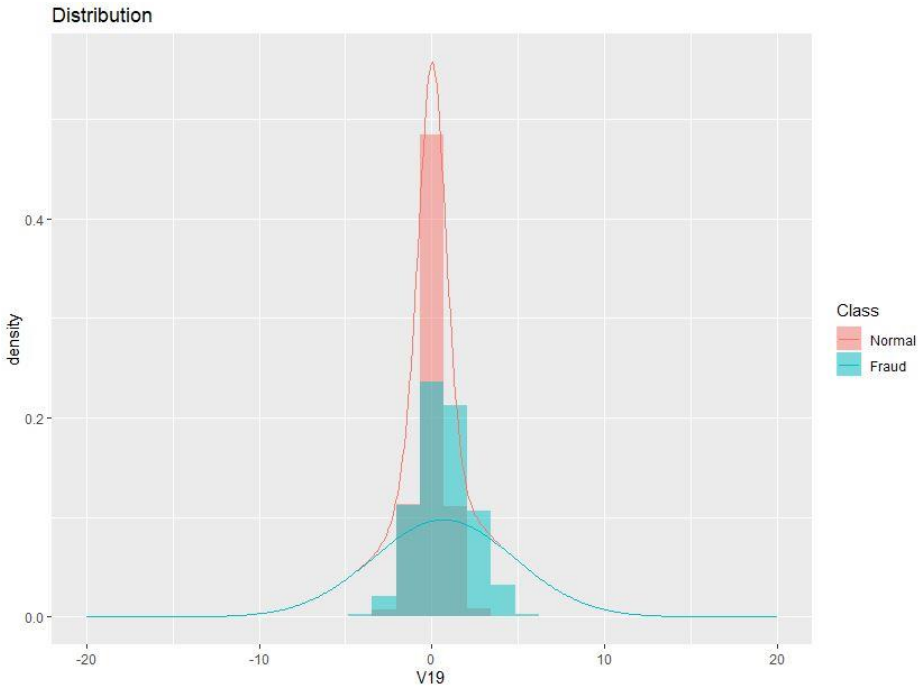
Και εδώ οι κατανομές είναι διαφορετικές για κάθε κλάση, ενω δεν φαίνεται να παρουσιάζεται ιδιαίτερη επικάλυψη.

V18



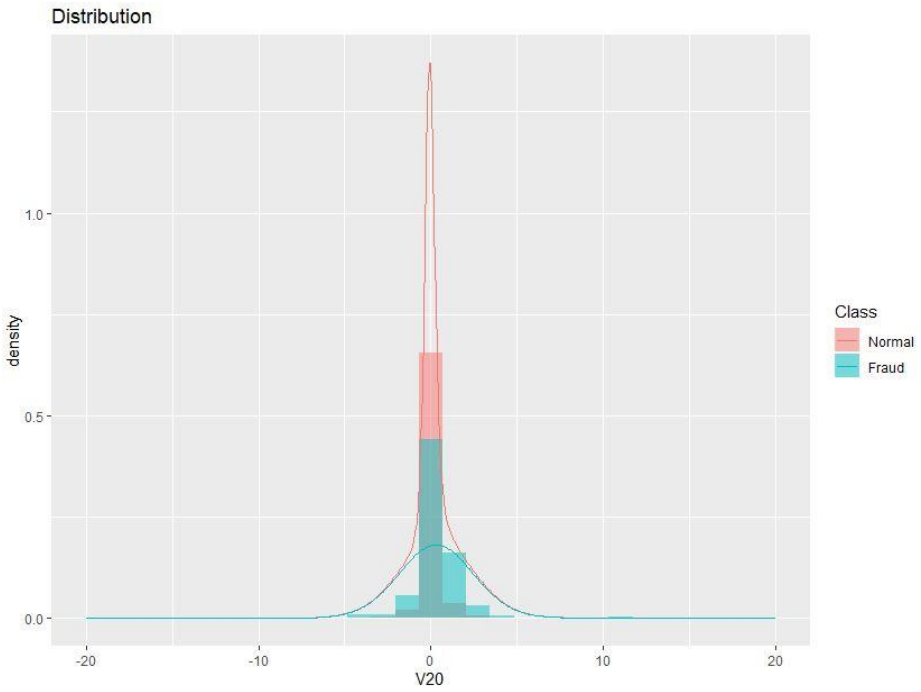
Και εδώ οι κατανομές είναι διαφορετικές για κάθε κλάση, ενω δεν φαίνεται να παρουσιάζεται ιδιαίτερη επικάλυψη.

V19



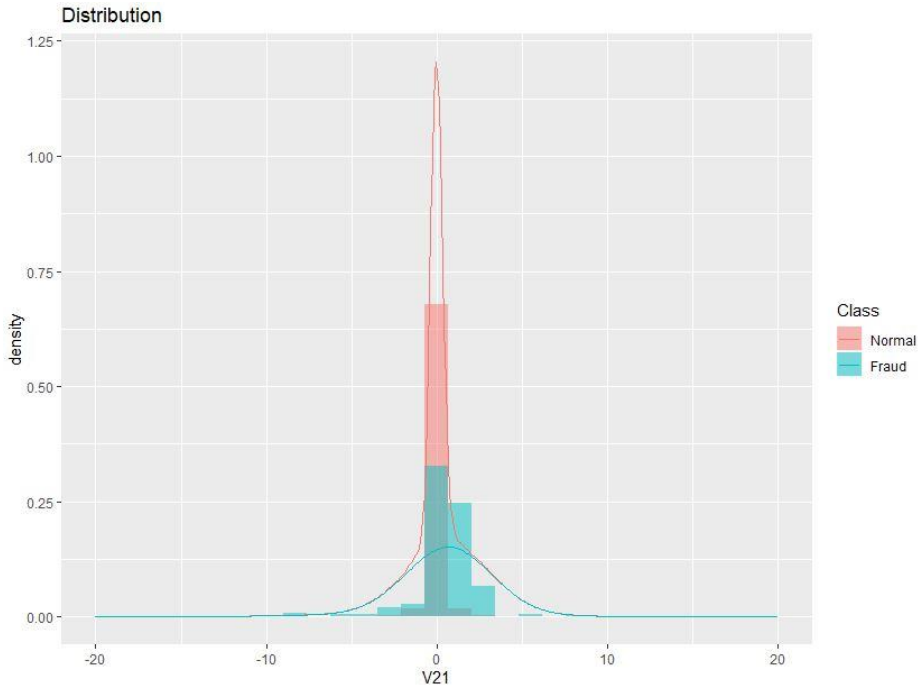
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V20



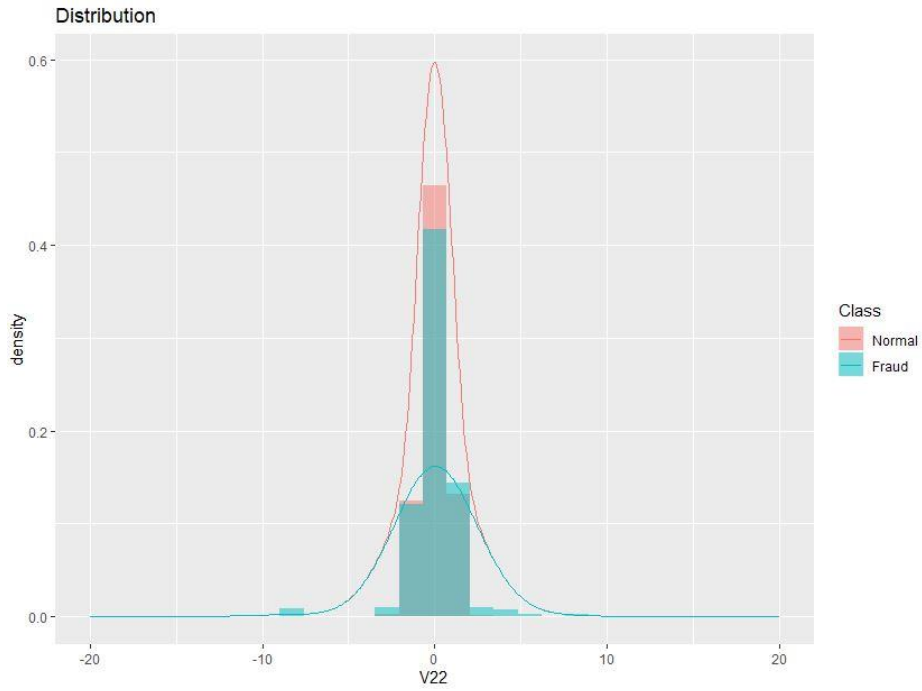
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V21



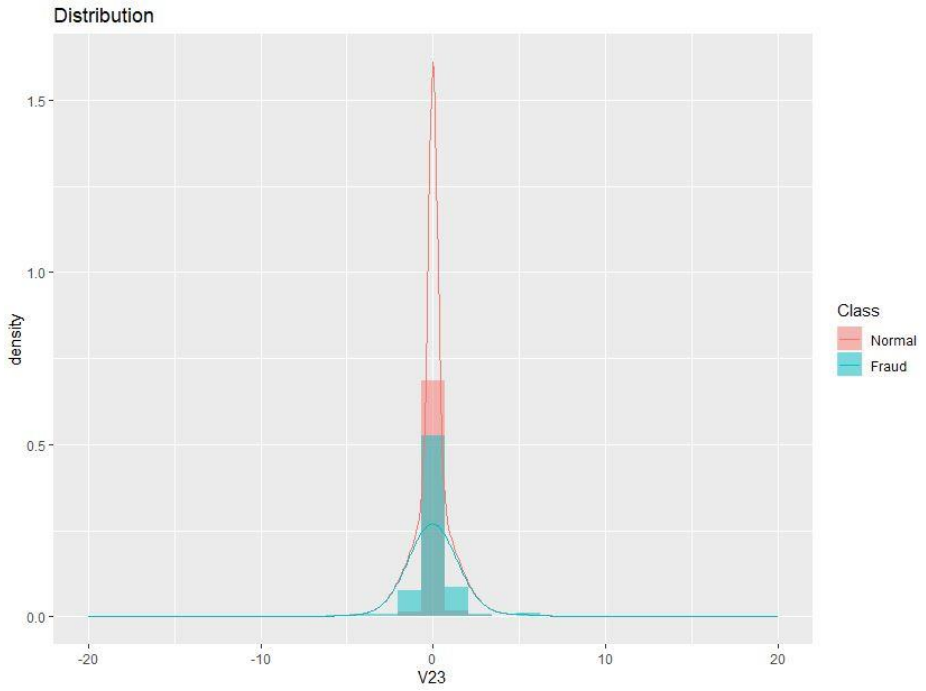
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V22



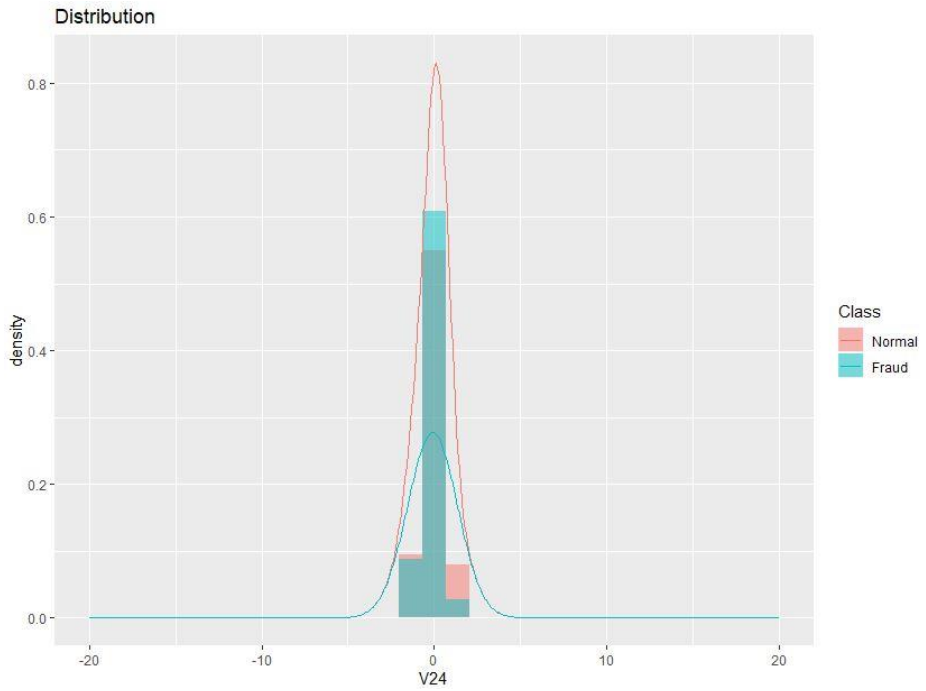
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V23



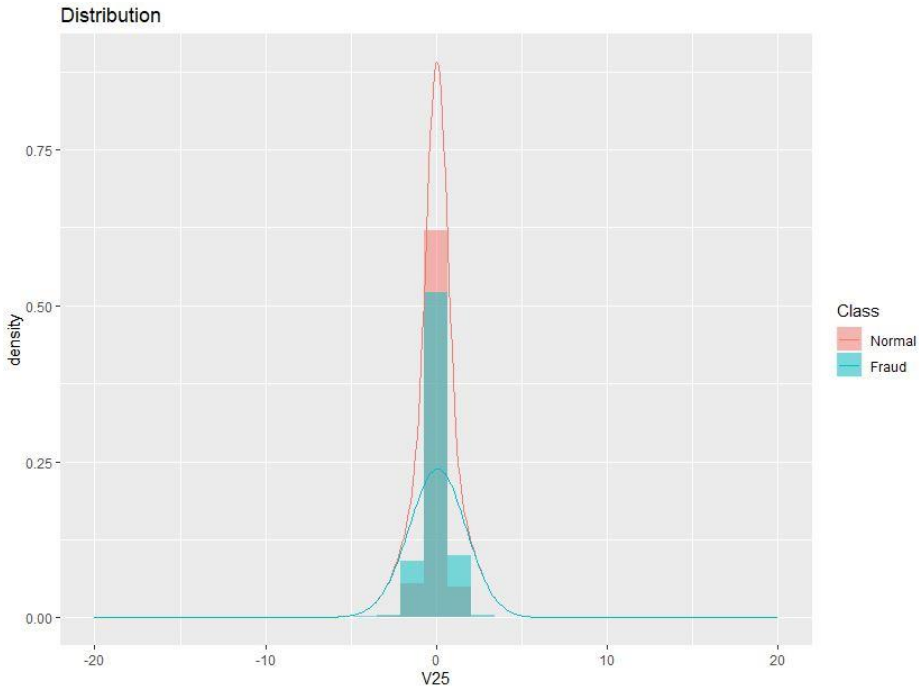
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V24



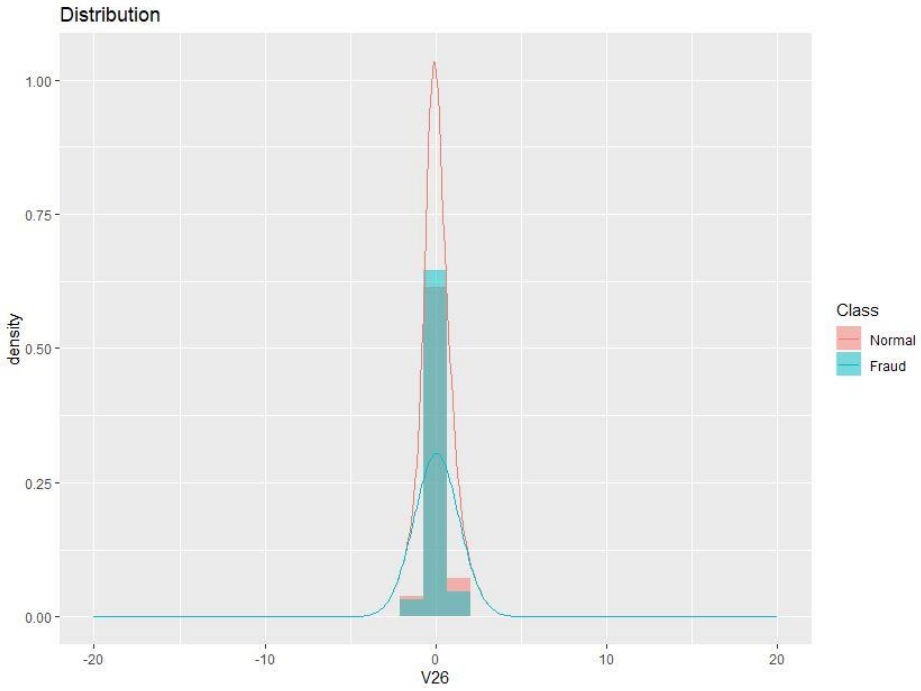
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V25



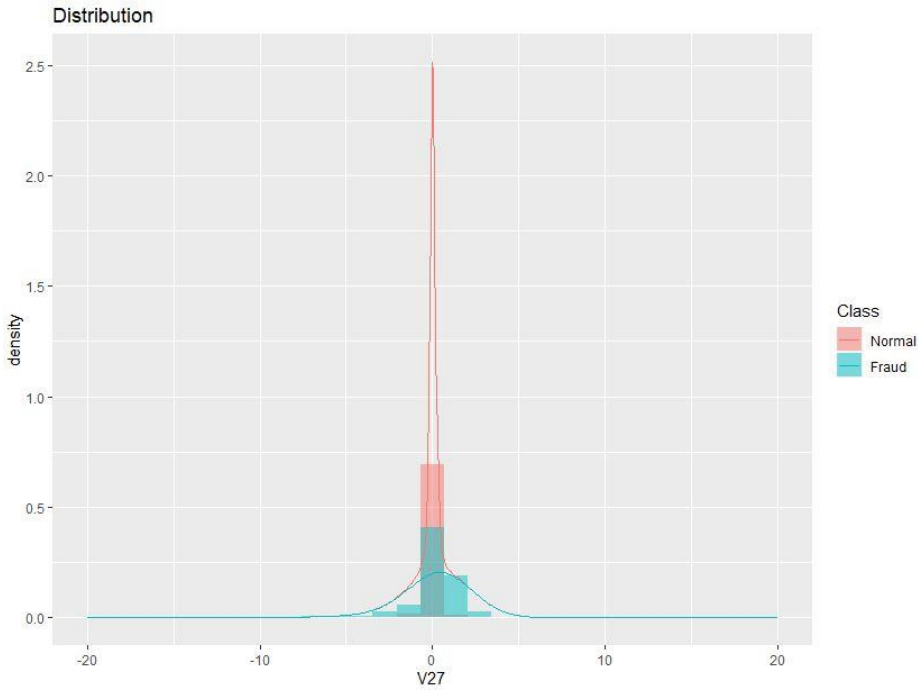
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V26



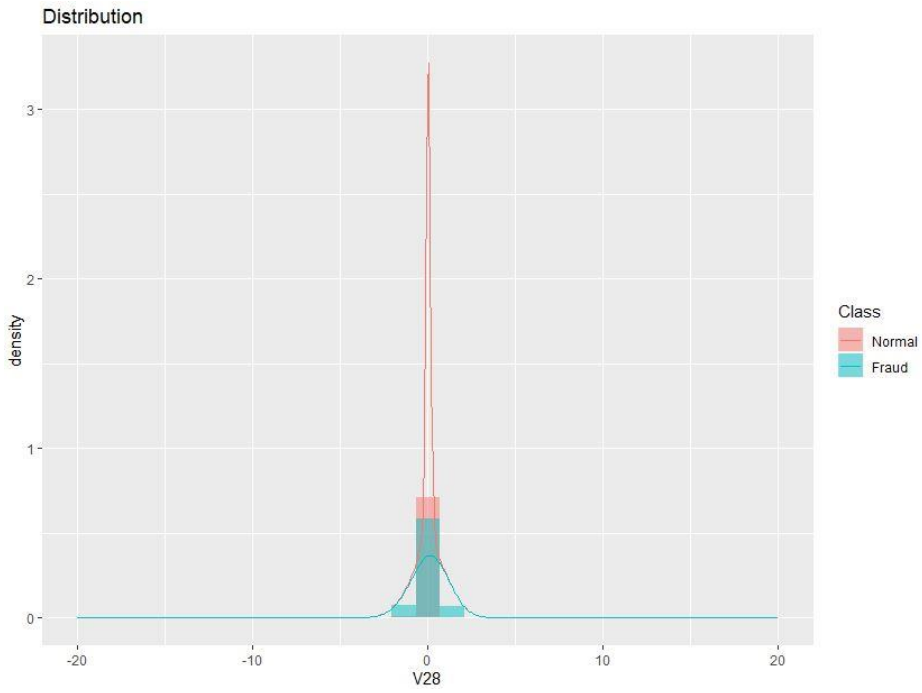
Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V27



Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

V28



Η κατανομή είναι εξαιρετικά όμοια, ενω φαίνεται να υπάρχει απόλυτη επικάλυψη.

11. Συμπεράσματα

Στην παρούσα εργασία έγινε μια βιβλιογραφική μελέτη πάνω στη Μηχανική Μάθηση και τις δίαφορες τεχνικές εξόρυξης γνώσης. Αναλύθηκαν κάποιοι απο τους πιο δημοφιλείς αλγόριθμους κατηγοριοποίησης, ενώ έγινε εκτενής ανάλυση των ιδιαιτεροτήτων και των ορθών πρακτικών της εφαρμογής της Μηχανικής Μάθησης στον χρηματοοικονομικό χώρο.

Μετέπειτα ορίστηκε το πρόβλημα της ανίχνευσης απάτης σε συναλλαγές πιστωτικών καρτών, ως πρόβλημα κατηγοριοποίησης σε σύνολο δεδομένων όπου παρουσιάζεται ισχυρή ανισορροπία στην εκπροσώπηση των δυο κλάσεων εντός του συνόλου. Κατασκευάστηκαν συνολικά 20 διαφορετικά μοντέλα κατηγοριοποίησης με την εφαρμογή των τεσσάρων πιο δημοφιλών τεχνικών χειρισμού ανισόροπων συνόλων δεδομένων. Τα μή ποιοτικά αποτελέσματα της απόδοσης των κατηγοριοποιητών ανέδειξαν ένα επιπρόσθετο πρόβλημα που σε πολλές περιπτώσεις, σύμφωνα με τη βιβλιογραφία, δείχνει να παίζει σημαντικότερο ρόλο απο αυτό της ανισορροπίας των κλάσεων στην προσπάθεια κατηγοριοποίησης, το πρόβλημα των επικαλυπτόμενων κατανομών. Γεγονός που υποδεικνύει πως σε γενικές γραμμές, δεν υπάρχει κάποια ‘καλύτερη’ μέθοδος ή αλγόριθμος, απλά σε κάποιες περιπτώσεις κάποιες τεχνικές λειτουργούν καλύτερα απο άλλες.

Βιβλιογραφία

1. *Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, 1993.*
2. *Rastogi, R., and Shim, K., PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning, 2000.*
3. *Arbel, R. and Rokach, L., Classifier evaluation under limited resources, Pattern Recognition Letters, 2006.*
4. *P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results, 1996.*
5. *N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers, 1997.*
6. *Chang C.-C. and Lin C.-J. Training support vector classifiers: Theory and algorithms. Neural Computation, 2001.*
7. *Cristianini N. and Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge Univ. Press, 2000.*
8. *Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies. Learning from Imbalanced Data Sets, 2000.*
9. *Stefanowski J. Algorithms of Decision Rule Induction in Data Mining. Poznan University of Technology Press, Poznan, Poland, 2001.*
10. *Arbel, R. and Rokach, L., Classifier evaluation under limited resources, Pattern Recognition Letters, 2006.*
11. *Huang, Z., Extensions to the k-means algorithm for clustering large data sets with categorical values,1998.*
12. *Rokach, L., Maimon, O., Data Mining with Decision Trees: Theory and Applications, 2008.*
13. *Michael H Cahill, Diane Lambert, Jost C Pinheiro, and Don X Sun. Detecting fraud in the real world, 2004.*
14. *Krishna M Gopinathan, Louis S Biafore, William M Ferguson, Michael A Lazarus, Anu K Pathria, and Allen Jost. Fraud detection using predictive modeling, 1998.*

15. *F. Provost, T. Fawcett, et al. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions, 1997.*
16. *Ryan N Lichtenwalter and Nitesh V Chawla. Adaptive methods for classification in arbitrarily imbalanced and drifting data streams, 2010.*
17. *G. Batista, A. Carvalho, and M. Monard. Applying one-sided selection to unbalanced datasets, 2000.*
18. *O. Maimon, L. Rokach. The data mining and knowledge discovery handbook, 2005.*
19. *Bart Baesens, Veronique Van Vlasselaer, and Wouter Verbeke. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. John Wiley & Sons, 2015.*
20. *Linda Delamaire, HAH Abdou, and John Pointon. Credit card fraud and detection techniques: a review, 2009.*
21. *R. Dubey, J. Zhou and Y. Wang, Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study, 2014.*
22. *H. Zhang and M. Li, RWO-Sampling: A random walk over-sampling approach to imbalanced data classification, 2014.*
23. *A. Jeni, J. Cohn and F. De la Tore, Facing Imbalanced Data--Recommendations for the Use of Performance Metrics, 2013.*
24. *M. Fatourechi, K. Ward and G. Mason, Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets, 2008.*
25. J. Zaki and W. Meira, Εξόρυξη και ανάλυση δεδομένων. Μετάφραση: Β. Μεγαλοοικονόμου και Χ. Μακρής, 2017.
26. *I. Written and E. Frank, Data Mining: Practical Machine Learning Tools & Techniques, 2000.*
27. *P. Tan, M. Steinbach and V. Kumar, Introduction to Datamining, 2006.*