ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Η ταξινόμηση ενός συνόλου εναλλακτικών αντικειμένων σε προκαθορισμένες ομοιογενείς κατηγορίες είναι ένα πρόβλημα ιδιαίτερου πρακτικού και ερευνητικού ενδιαφέροντος. Έχει αποδειχθεί ότι τα διάφορα χαρακτηριστικά των δεδομένων επιδρούν σημαντικά στην αποτελεσματικότητα των μεθόδων. Το αντικείμενο αυτής της έρευνας είναι η κατανόηση των δυνατοτήτων και των ορίων των διαφορετικών μεθόδων ταξινόμησης, καθώς και η επίδραση που έχουν τα εξαγόμενα δεδομένα στην αποτελεσματικότητα των μεθόδων. Για το σκοπό αυτό χρησιμοποιείται ένα συνθετικό σύνολο δεδομένων με προσεχτικά επιλεγμένα χαρακτηριστικά. Επιπρόσθετα, οι τεχνικές ταξινόμησης που χρησιμοποιούνται είναι τόσο από τον χώρο των στατιστικών και οικονομετρικών προσεγγίσεων, όσο και από τον χώρο των μη παραμετρικών προσεγγίσεων. Ο στόχος είναι η διερεύνηση της αποτελεσματικότητας των μεθόδων συναρτήσει των χαρακτηριστικών των εξεταζόμενων δεδομένων. Συγκεκριμένα, τα χαρακτηριστικά των δεδομένων επιλέγονται με βάσει τις δυνατότητες και τις αδυναμίες της κάθε μεθόδου, έτσι ώστε να εξετασθεί ο τρόπος με τον οποίο επηρεάζεται η αποτελεσματικότητα των διαφορετικών μεθόδων.

Στη συγκεκριμένη έρευνα γίνεται μια εισαγωγή στο πρόβλημα της ταξινόμησης και παρουσιάζονται οι κυριότερες τεχνικές ταξινόμησης. Στη συνέχεια αναλύονται τα δεδομένα που χρησιμοποιούνται και γίνεται μια αναλυτική παρουσίαση των αποτελεσμάτων. Τέλος, παρουσιάζονται τα γενικά συμπεράσματα που προκύπτουν καθώς και οι μελλοντικές επεκτάσεις που μπορεί να υπάρξουν.

ΚΕΦΑΛΑΙΟ 1

Μεθοδολογικό πλαίσιο εκτίμησης μεθόδων ταξινόμησης

1.1 Εισαγωγή

Τα συστήματα ταξινόμησης παίζουν ένα σημαντικό ρόλο στις διάφορες επιχειρηματικές αποφάσεις, κατατάσσοντας τις διαθέσιμες πληροφορίες σε κατηγορίες, οι οποίες βασίζονται σε διάφορα κριτήρια. Η αντιμετώπιση του προβλήματος της ταξινόμησης βάσει των διαθέσιμων μεθοδολογικών προσεγγίσεων συνίσταται στην ανάπτυξη ποσοτικών υποδειγμάτων, τα οποία υποστηρίζουν τη διαδικασία λήψης αποφάσεων στη βάση της προβληματικής της ταξινόμησης.

Η ταξινόμηση της πληροφορίας είναι ένα σημαντικό συστατικό για τις διάφορες επιχειρηματικές αποφάσεις. Πολλά θέματα αποφάσεων αποτελούν τμήματα ενός προβλήματος ταξινόμησης ή μπορούν εύκολα να μετατραπούν σε πρόβλημα ταξινόμησης (π.χ. προβλήματα πρόβλεψης, διάγνωσης, αναγνώρισης προτύπων κ.λ.π.). Η ταξινόμηση αποτέλεσε περισσότερο σημαντική με την έλευση του διαδικτύου. Το διαδίκτυο ως κανάλι επικοινωνίας και συναλλαγής, παρέχει τη δυνατότητα εφαρμογής νέων τεχνολογιών, όπως collaborative filtering and recommender systems, τα οποία διευκολύνουν το μάρκετινγκ και την εξυπηρέτηση της μαζικής πελατείας (Resnick and Varian, 1997). Ένας πρωτεύον σκοπός αυτών των συστημάτων είναι η ταξινόμηση της διαθέσιμης πληροφορίας βάσει κάποιων κριτηρίων.

Συγκεκριμένα, θεωρώντας ένα σύνολο εναλλακτικών δραστηριοτήτων που περιγράφονται από κάποια κριτήρια, υπάρχουν τέσσερις διαφορετικές αναλύσεις, για την υποστήριξη μιας απόφασης (Roy, 1985):

1. επιλογή της καλύτερης εναλλακτικής δραστηριότητας.
2. κατάταξη των εξεταζόμενων εναλλακτικών δραστηριοτήτων από τις καλύτερες προς τις χειρότερες βάσει των χαρακτηριστικών τους.
3. ταξινόμηση των εναλλακτικών δραστηριοτήτων σε προκαθορισμένες κατηγορίες.

4. περιγραφή των εναλλακτικών δραστηριοτήτων με στόχο τον εντοπισμό των βασικών τους χαρακτηριστικών και ιδιοτήτων.

Οι τρεις πρώτες κατηγορίες προβλημάτων οδηγούν σε ένα συγκεκριμένο αποτέλεσμα αξιολόγησης των εξεταζόμενων εναλλακτικών δραστηριοτήτων. Τα προβλήματα της επιλογής και της κατάταξης βασίζονται στην πραγματοποίηση σχετικών συγκρίσεων ανάμεσα στις εξεταζόμενες εναλλακτικές δραστηριότητες. Οι σχετικές συγκρίσεις που πραγματοποιούνται αφορούν τη σύγκριση όλων των εναλλακτικών δραστηριοτήτων μεταξύ τους. Κατά συνέπεια, το αποτέλεσμα της αξιολόγησης έχει και αυτό μια σχετική μορφή, δηλαδή επιλέγεται η εναλλακτική δραστηριότητα που είναι καλύτερη σε σχέση με τις υπόλοιπες ή κατατάσσονται οι εναλλακτικές από τις σχετικά καλύτερες προς τις σχετικά χειρότερες. Έτσι το αποτέλεσμα της αξιολόγησης δύναται να μεταβληθεί με τη μεταβολή του συνόλου των εξεταζόμενων εναλλακτικών δραστηριοτήτων.

Υπάρχει μια πληθώρα μεθόδων ταξινόμησης οι οποίες έχουν εφαρμοστεί στη χρηματοοικονομική διοίκηση και οικονομική πολιτική (Ζοπουνίδης, 1998, Ζοπουνίδης και Δούμπος, 1998), στην αναγνώριση προτύπων (pattern recognition, Ripley, 1996, Young and Fu, 1997), στη διαχείριση ανθρώπινου δυναμικού, στη διαχείριση παραγωγικών συστημάτων (Catelani and Fort, 2000), στο μάρκετινγκ (Dutka, 1995, Siskos et al., 1998), στην περιβαλλοντική και ενεργειακή διαχείριση (Diakoulaki et al., 1999), και στην ιατρική (Tsumoto, 1998). Ωστόσο, λίγες έρευνες έχουν γίνει για τον συστηματικό έλεγχο της απόδοσης των αλγορίθμων που χρησιμοποιούνται στην ταξινόμηση (Breese et al., 1998). Είναι φανερό από παλαιότερες μελέτες ότι υπάρχει μια μεγάλη διακύμανση στην απόδοση των αλγορίθμων ταξινόμησης, κάτω από διαφορετικά σενάρια (Dietterich et al., 1995, Meila and Heckerman, 2001).

Καθώς τα συστήματα ταξινόμησης αποτελούν αναπόσπαστο τμήμα των συστημάτων υποστήριξης αποφάσεων, η προσαρμογή στις διακυμάνσεις των χαρακτηριστικών των δεδομένων και στα δυναμικά επιχειρηματικά σενάρια γίνεται ολοένα και πιο σημαντική. Είναι, λοιπόν, επιτακτική ανάγκη η χρήση προσαρμοστικών συστημάτων ταξινόμησης, τα οποία χρησιμοποιούν κατάλληλες μεθόδους, αφού πρώτα αναλύσουν τα διαθέσιμα δεδομένα.

1.2 Αντικείμενο και δομή διατριβής

Έχει αποδειχθεί ότι τα διάφορα χαρακτηριστικά των δεδομένων επιδρούν σημαντικά στην αποτελεσματικότητα των μεθόδων. Ένα από τα κύρια μειονεκτήματα παλαιότερων ερευνών είναι ότι βασίζονται σε μη ελεγχόμενα δεδομένα (μεροληψίες). Το αντικείμενο αυτής της έρευνας είναι η κατανόηση των δυνατοτήτων και των ορίων των διαφορετικών μεθόδων ταξινόμησης, καθώς και η επίδραση που έχουν τα εξαγόμενα δεδομένα στην αποτελεσματικότητα των μεθόδων. Για το σκοπό αυτό θα χρησιμοποιηθεί ένα συνθετικό σύνολο δεδομένων με προσεχτικά επιλεγμένα χαρακτηριστικά. Επιπρόσθετα, οι τεχνικές ταξινόμησης που θα χρησιμοποιηθούν θα είναι τόσο από τον χώρο των στατιστικών και οικονομετρικών προσεγγίσεων, όσο και από τον χώρο των μη παραμετρικών προσεγγίσεων. Ο στόχος είναι η διερεύνηση της απολεσματικότητας των μεθόδων συναρτήσει των χαρακτηριστικών των εξεταζόμενων δεδομένων. Συγκεκριμένα, τα χαρακτηριστικά των δεδομένων θα επιλεγούν με βάσει τις δυνατότητες και τις αδυναμίες της κάθε μεθόδου, έτσι ώστε να εξετασθεί ο τρόπος με τον οποίο επηρεάζεται η αποτελεσματικότητα των διαφορετικών μεθόδων.

Οι παράγοντες που επιλέχθηκαν για την εξέταση της αποτελεσματικότητας των μεθόδων είναι η στατιστική κατανομή, την οποία ακολουθούν τα δεδομένα, η μορφή διάκρισης των κατηγοριών (γραμμική, μη γραμμική), το πλήθος των αντικειμένων στο δείγμα εκπαίδευσης και ο βαθμός συσχέτισης των κριτηρίων. Επιπρόσθετα, η αποτελεσματικότητα των μεθόδων ελέγχεται και για δύο διαφορετικούς τύπους δεδομένων, συνεχή και διακριτά. Τα παραπάνω εφαρμόζονται σε στατιστικές – οικονομετρικές και μη παραμετρικές προσεγγίσεις. Συγκεκριμένα, εξετάζονται εφτά τεχνικές ταξινόμησης, οι οποίες είναι η Γραμμική Διακριτική Ανάλυση (LDA), η Τετραγωνική Διακριτική Ανάλυση (QDA), το Λογιστικό Υπόδειγμα Πιθανότητας (LOGIT), ο Αλγόριθμός Πλησιέστερου Γείτονα (1NN), τα Πιθανολογικά Νευρωνικά Δίκτυα (PNN), οι Μηχανές Διανύσματος Υποστήριξης (SVM) και η μέθοδος UTADIS.

Τα αποτελέσματα της μελέτης αυτής μπορούν να βοηθήσουν στο σχεδιασμό συστημάτων ταξινόμησης. Επίσης, μπορεί να θέσει τις βάσεις για τον σχεδιασμό περισσότερο προσαρμοστικών συστημάτων ταξινόμησης.

Η διατριβή οργανώνεται με τον ακόλουθο τρόπο:

Στο κεφάλαιο 2 πραγματοποιείται μια εισαγωγή στο πρόβλημα της ταξινόμησης. Παρουσιάζονται οι βασικές έννοιες και ορισμοί του προβλήματος της ταξινόμησης, καθώς και ένα γενικό πλαίσιο των τεχνικών ανάπτυξης υποδειγμάτων ταξινόμησης.

Στο κεφάλαιο 3 παρουσιάζεται μια σύντομη ανασκόπηση των βασικότερων μεθοδολογικών προσεγγίσεων που έχουν προταθεί για την ανάπτυξη υποδειγμάτων ταξινόμησης. Η ανασκόπηση αυτή αφορά τόσο τις στατιστικές – οικονομετρικές προσεγγίσεις όσο και τις μη παραμετρικές προσεγγίσεις.

Στο κεφάλαιο 4 πραγματοποιείται μια ανάλυση των δεδομένων και των παραγόντων που εμπλέκονται στην εφαρμογή των τεχνικών ταξινόμησης. Επιπλέον, παρουσιάζεται μια σύντομη περιγραφή της διαδικασίας παραγωγής των δεδομένων.

Στο κεφάλαιο 5 παρουσιάζονται τα αποτελέσματα της ανάλυσης. Η ανάλυση των αποτελεσμάτων πραγματοποιείται με βάση τους παράγοντες που συνδυάζονται κατά την εφαρμογή των μεθόδων.

Τέλος, στο κεφάλαιο 6 παρουσιάζονται τα γενικά συμπεράσματα που προκύπτουν από την εφαρμογή των τεχνικών ταξινόμησης. Επιπλέον, προτείνονται διάφορες κατευθύνσεις για μελλοντικές έρευνες.

ΚΕΦΑΛΑΙΟ 2

Εισαγωγή στο πρόβλημα της ταξινόμησης

2.1 Ορισμός της ταξινόμησης

Η ταξινόμηση λαμβάνει χώρα σε ένα μεγάλο εύρος των ανθρώπινων δραστηριοτήτων. Ο όρος αυτός καλύπτει οποιοδήποτε πλαίσιο, μέσα στο οποίο κάποια απόφαση ή πρόβλεψη συντελείται στη βάση μιας διαθέσιμης πληροφορίας. Η διαδικασία ταξινόμησης είναι μια επίσημη μέθοδος για την επανάληψη τέτοιων κρίσεων σε καινούριες καταστάσεις.

Η ταξινόμηση έχει δύο διακριτές έννοιες. Η πρώτη αφορά ένα σύνολο δεδομένων με το οποίο πρέπει να καθοριστεί η ύπαρξη κατηγοριών ή ομάδων στα δεδομένα. Η δεύτερη αφορά την ύπαρξη συγκεκριμένων κατηγοριών, όπου σκοπός είναι να αναπτυχθεί κάποιος κανόνας με τον οποίο να ταξινομείται οποιαδήποτε νέα παρατήρηση σε κάποια από τις υπάρχουσες κατηγορίες. Η πρώτη περίπτωση αναφέρεται ως ομαδοποίηση (Clustering), ενώ η δεύτερη ως ταξινόμηση (discrimination).

Ένας λειτουργικός ορισμός της ταξινόμησης οφείλεται στον Mirkin (1996), ο οποίος καθορίζει το μηχανισμό αλλά και τη χρησιμότητα της:
Ταξινόμηση είναι η ιδεατή τοποθέτηση μαζί παρόμοιων αντικειμένων και ο διαχωρισμός των αντικειμένων τα οποία διαφέρουν, με απώτερο σκοπό:

1. Τη διαμόρφωση, οργάνωση και διατήρηση της γνώσης.
2. Την ανάλυση της δομής του φαινομένου που εξετάζεται.
3. Τη συσχέτιση των διαφόρων πλευρών του υπό εξέταση φαινομένου.

Ο όρος «διαμόρφωση της γνώσης» απαιτεί διευκρίνηση. Όπως έχει αναφερθεί παραπάνω, μια διαφορετική από την ταξινόμηση κατηγορία τεχνικών, χαρακτηριζόμενες από τον όρο ομαδοποίηση, αναλαμβάνει το διαχωρισμό των δεδομένων σε άγνωστες εκ των προτέρων ομάδες, γνωστού ή άγνωστου αριθμού, τοποθετώντας στην ίδια ομάδα αντικείμενα με όμοια ή παραπλήσια χαρακτηριστικά. Έτσι, στην ομαδοποίηση οι συγκρίσεις είναι σχετικές, και η σύνθεση του πληθυσμού των αντικειμένων επηρεάζει τόσο τις εξαγόμενες κατηγορίες, όσο και την

τοποθέτηση των αντικειμένων σε αυτές. Η κυριότερη διαφορά μεταξύ ταξινόμησης και ομαδοποίησης είναι γνωσιολογική. Στην ταξινόμηση, με a priori γνωστές τις κατηγορίες, η γνώση που παράγεται τοποθετείται σε μια υπάρχουσα δομή, της οποίας η ποιότητα παραμένει αμετάβλητη. Αντιθέτως, η ομαδοποίηση εξάγει την ίδια τη δομή της γνώσης, καθορίζοντας τις κατηγορίες των αντικειμένων βάσει εμπειρίας.

Στην αγγλική ορολογία χρησιμοποιούνται διάφοροι όροι για την αναφορά στο πρόβλημα της ταξινόμησης, οι συνηθέστεροι των οποίων είναι οι ακόλουθοι τρεις:

- Discrimination (διάκριση)
- Classification (ταξινόμηση)
- Sorting (διατεταγμένη ταξινόμηση)

Οι δύο πρώτοι όροι χρησιμοποιούνται κυρίως από στατιστικολόγους καθώς και από ερευνητές που δραστηριοποιούνται στο χώρο της τεχνητής νοημοσύνης (νευρωνικά δίκτυα, μηχανική μάθηση, κ.λ.π.). Αντίθετα, ο τρίτος όρος έχει εισαχθεί και χρησιμοποιείται κυρίως από ερευνητές του χώρου της πολυκριτήριας ανάλυσης αποφάσεων.

Είναι αναγκαίο να σημειωθεί, ότι παρ' όλες τις διαφορές μεταξύ τους, η ταξινόμηση και η ομαδοποίηση δεν είναι ανταγωνιστικές προσεγγίσεις. Αντιθέτως, μπορούν να χρησιμοποιηθούν διαδοχικά, με την ομαδοποίηση να εξάγει τη δομή της γνώσης που εμπεριέχει το δείγμα, και την ταξινόμηση να ακολουθεί, παρέχοντας ένα πρότυπο για την τοποθέτηση νέων αντικειμένων στη δομή αυτή.

2.1.1 Ζητήματα που αφορούν την ταξινόμηση

Υπάρχουν πολλά ζητήματα που αφορούν την ταξινόμηση. Μερικά από αυτά είναι τα ακόλουθα:

- Ακρίβεια. Η αξιοπιστία του κανόνα ταξινόμησης συνήθως, αντιπροσωπεύεται από την αναλογία των σωστών ταξινομήσεων. Παρόλα αυτά, είναι σημαντικό να ελέγχεται ο δείκτης σφαλμάτων.
- Ταχύτητα. Σε μερικές περιπτώσεις, η ταχύτητα του υποδείγματος ταξινόμησης αποτελεί ένα κύριο ζήτημα. Ένα υπόδειγμα ταξινόμησης, το οποίο είναι 90% ακριβές, είναι προτιμότερο από κάποιο, το οποίο

είναι 95% ακριβές, εάν το πρώτο είναι 100 φορές πιο γρήγορο στον έλεγχο.

- Σαφήνεια. Η διαδικασία της ταξινόμησης πρέπει να είναι εύκολα κατανοητή, διαφορετικά σφάλματα είναι πολύ εύκολο να συμβούν κατά την εφαρμογή.
- Χρόνος εκμάθησης. Σε ένα γρήγορα μεταβαλλόμενο περιβάλλον, είναι αναγκαία η γρήγορη εκμάθηση των κανόνων ταξινόμησης, ή οι τροποποιήσεις σε ένα υπάρχοντα κανόνα, σε πραγματικό χρόνο. Επίσης, «γρήγορα» μπορεί να υπονοείται ότι ένας μικρός αριθμός παρατηρήσεων είναι αρκετός για την εφαρμογή ενός συγκεκριμένου κανόνα ταξινόμησης.

2.1.2 Ορισμός των κατηγοριών

Ένα σημαντικό ζήτημα, το οποίο δεν κατανοείται σωστά σε πολλές μελέτες ταξινόμησης, είναι η φύση των κατηγοριών και ο τρόπος με τον οποίο ορίζονται. Διακρίνονται τρεις περιπτώσεις:

1. Οι κατηγορίες ανταποκρίνονται στα χαρακτηριστικά διαφορετικών πληθυσμών.
2. Οι κατηγορίες προκύπτουν από μια διαδικασία πρόβλεψης. Σε αυτή την περίπτωση, η κατηγορία είναι το αποτέλεσμα, το οποίο έχει προβλεφθεί από τη γνώση των κριτηρίων. Σε στατιστικούς όρους, η κατηγορία είναι μια τυχαία μεταβλητή.
3. Οι κατηγορίες έχουν προκαθοριστεί από τη διαίρεση του δείγματος (π.χ. από τα ίδια τα κριτήρια). Η κατηγορία θεωρείται ως μια συνάρτηση των κριτηρίων. Κατά συνέπεια, ένα αντικείμενο μπορεί να ταξινομηθεί εσφαλμένα, εάν κάποια κριτήρια βρίσκονται εκτός καθορισμένων ορίων, και σωστά διαφορετικά.

2.1.3 Ακρίβεια (accuracy)

Πρέπει να διευκρινιστεί, ότι η ακρίβεια που εκτιμάται στο δείγμα εκμάθησης και η ακρίβεια που εκτιμάται στο δείγμα ελέγχου, είναι συχνά εντελώς διαφορετικά. Πράγματι, δεν είναι ασυνήθιστο, ιδιαίτερα στις εφαρμογές της μηχανικής μάθησης, οι τεχνικές να παρουσιάζουν ικανοποιητική εφαρμογή στο δείγμα εκμάθησης, ενώ στο δείγμα ελέγχου να δίνουν απογοητευτικά αποτελέσματα. Συνήθως, αυτό που είναι σημαντικό, είναι η ακρίβεια που δίνει η ταξινόμηση ενός άγνωστου αντικειμένου. Μια γενικά αποδεκτή μέθοδος για την εκτίμηση του παραπάνω, είναι η χρήση των γνωστών δεδομένων, των οποίων οι κατηγορίες ταξινόμησης είναι γνωστές. Αρχικά, χρησιμοποιείται ένα δείγμα εκμάθησης για την εφαρμογή της διαδικασίας. Στη συνέχεια, αυτό ελέγχεται με το δείγμα ελέγχου και τα αποτελέσματα συγκρίνονται με ήδη υπάρχουσες γνωστές ταξινομήσεις. Η αναλογία που τα αποτελέσματα είναι σωστά στο δείγμα ελέγχου, είναι μια αμερόληπτη ακρίβεια, κατά την οποία επιβεβαιώνεται ότι το δείγμα ελέγχου έχει τυχαία επιλεγεί από τα υπόλοιπα δεδομένα.

2.2 Μαθηματική διατύπωση και ορισμοί

Οι παρακάτω ορισμοί περιγράφουν στοιχειώδεις έννοιες της ταξινόμησης. Οι συγκεκριμένες προέρχονται από τους Bauer και Kohavi (1999).

Παράδειγμα ή labeled instance ονομάζεται ένα ζεύγος (x,y), όπου $x \in X$, X ο χώρος των αντικειμένων ή παραδειγμάτων (instance space) ή το γνωστικό πεδίο (domain), $y \in Y$ και Y ο διακριτός χώρος των κατηγοριών (space of classes). Το x έχει n συνιστώσες και ονομάζεται διάνυσμα χαρακτηριστικών (vector of attributes), ενώ το y υποδεικνύει την κατηγορία στην οποία ανήκει το αντικείμενο.

Δείγμα (sample) ονομάζεται ένα σύνολο από αντικείμενα.

Υπόδειγμα ταξινόμησης (classifier) ή υπόθεση (hypothesis) είναι μια συνάρτηση $h : X \to Y$. Επιπλέον, ονομάζουμε αιτιοκρατική διαδικασία επαγωγής (deterministic inducer) μία αντιστοιχία $I : 2^x \to H$, όπου H ο χώρος των υποδειγμάτων ταξινόμησης και 2^x τα δυνατά υποσύνολα του χώρου X.

2.3 Γενικό πλαίσιο ανάπτυξης υποδειγμάτων ταξινόμησης

Η διαδικασία της ταξινόμησης απαιτεί την επιλογή ενός ικανοποιητικού υποδείγματος ταξινόμησης $h : X \to Y$ μέσω μιας διαδικασίας εκπαίδευσης (training). Η εκπαίδευση γίνεται πάντοτε με ένα δείγμα, που καλείται δείγμα εκμάθησης. Ως δείγμα εκμάθησης (training sample) ή σύνολο αναφοράς (reference set) ορίζεται το δείγμα των παρατηρήσεων που χρησιμοποιείται για την ανάπτυξη των υποδειγμάτων ταξινόμησης. Το υπόδειγμα ταξινόμησης h καλείται να ταξινομήσει άγνωστα αντικείμενα, δηλαδή αντικείμενα των οποίων η κατηγορία δεν είναι γνωστή. Από μαθηματικής απόψεως, θεωρείται ότι υπάρχει μια αιτιοκρατική συνάρτηση $h : X \to Y$, η οποία προσδιορίζει την ταξινόμηση κάθε αντικειμένου του συνόλου X σε μια κατηγορία του συνόλου Y.

Για την κατασκευή του υποδείγματος ταξινόμησης h, δεν αρκεί το δείγμα εκμάθησης. Πρέπει να καθοριστούν ο χώρος H από όπου θα επιλεγεί το υπόδειγμα ταξινόμησης, καθώς και ο ίδιος ο μηχανισμός αυτής της επιλογής. Οι δύο αυτές αποφάσεις ανάγονται στην επαγωγική διαδικασία I, όπου η διαδικασία επαγωγής είναι η συγκεκριμένη τεχνική ταξινόμησης που χρησιμοποιείται, ενώ ο χώρος H περιγράφει τη γενική μορφή των υποδειγμάτων ταξινόμησης που μπορεί να δώσει η τεχνική αυτή.

Η αναζήτηση του κατάλληλου υποδείγματος στο χώρο H γίνεται από τη διαδικασία επαγωγής μεταβάλλοντας τις παραμέτρους της γενικής μορφής υποδειγμάτων ταξινόμησης που διαθέτει. Γνώμονας κατά τη διαδικασία αυτή είναι συνήθως είτε η βελτιστοποίηση κάποιου μέτρου ποιότητας της ταξινόμησης στα αντικείμενα του δείγματος εκμάθησης, είτε κάποιος ευρετικός κανόνας (Δούμπος και Ζοπουνίδης, 2001).

Η διαδικασία επιλογής του υποδείγματος ταξινόμησης είναι εν γένει επαναληπτική. Αν διαπιστωθεί ότι το προτεινόμενο υπόδειγμα ταξινόμησης δεν παρέχει ικανοποιητικές εκτιμήσεις, ο αποφασίζων εκτελεί τη διαδικασία επαγωγής με ένα νέο δείγμα εκμάθησης. Αν όμως θεωρεί μια πιο ριζική αλλαγή, αναιρεί την ίδια τη διαδικασία επαγωγής, αλλάζει τεχνική ταξινόμησης, αναζητώντας υπόδειγμα ταξινόμησης διαφορετικής μορφής.

2.4 Ένα γενικό πλαίσιο για τα προβλήματα ταξινόμησης

Υπάρχουν τρία βασικά συστατικά σε ένα πρόβλημα ταξινόμησης**:**

1. Η σχετική συχνότητα με την οποία οι κατηγορίες εμφανίζονται σε ένα πληθυσμό η οποία εκφράζεται ως αρχική πιθανότητα κατανομής.
2. Τα κριτήρια διαχωρισμού των κατηγοριών.
3. Το κόστος που σχετίζεται με μια εσφαλμένη ταξινόμηση.

Η πλειοψηφία των τεχνικών συνδυάζουν τα συστατικά, παράγοντας ένα κανόνα ταξινόμησης ο οποίος προέρχεται από μία αρχική κατανομή και δεν μπορεί εύκολα να προσαρμοστεί σε μια μεταβολή της συχνότητας των κατηγοριών. Ωστόσο, στην θεωρία κάθε ένα από αυτά τα συστατικά μπορεί να μελετηθεί ξεχωριστά και στην συνέχεια να συνδυαστούν τα αποτελέσματα σε ένα κανόνα ταξινόμησης.

2.4.1 Αφελής κανόνας ταξινόμησης

Έστω ότι οι κατηγορίες ορίζονται ως A_i, i=1,…,q, τότε η πιθανότητα π_i της κατηγορίας A_i θα είναι:

$$\pi_i = p(A_i)$$

Είναι πάντα πιθανό να χρησιμοποιηθεί ο κανόνας των «μη-δεδομένων» (no-data rule): ταξινομείται κάθε νέα παρατήρηση στην κατηγορία A_k, ανεξάρτητα από τα χαρακτηριστικά της. Αυτός ο κανόνας αθέτησης μπορεί ακόμη να χρησιμοποιηθεί στην πράξη, εάν το κόστος συλλογής δεδομένων είναι υψηλό. Για παράδειγμα, οι τράπεζες παρέχουν πίστωση σε όλους τους καθιερωμένους πελάτες τους για χάρη των καλών πελατειακών σχέσεων (στην περίπτωση αυτή, το κόστος συλλογής δεδομένων είναι ο κίνδυνος της απώλειας πελατών). Ο κανόνας αθέτησης βασίζεται μόνο στη γνώση των αρχικών πιθανοτήτων ενώ ο κανόνας απόφασης που έχει τη μεγαλύτερη πιθανότητα επιτυχίας, κατανέμει κάθε νέα παρατήρηση στην κατηγορία με τη μεγαλύτερη συχνότητα. Ωστόσο, εάν κάποια σφάλματα ταξινόμησης είναι περισσότερο σημαντικά από κάποια άλλα, τότε υιοθετείται ο κανόνας του ελάχιστου προσδοκώμενου κόστους, και η κατηγορία k είναι αυτή με το ελάχιστο προσδοκώμενο κόστος.

2.4.2 Διαχωρισμός κατηγοριών

Έστω ότι παρατηρούνται δεδομένα x σε ένα δείγμα, και ότι η πιθανότητα κατανομής του x σε κάθε μία από τις κατηγορίες A_i είναι $P(x \mid A_i)$. Τότε για κάθε μία από τις κατηγορίες A_i , A_j ο δείκτης πιθανότητας $P(x \mid A_i) / P(x \mid A_j)$ παρέχει τη βέλτιστη θεωρητική μορφή διάκρισης των κατηγοριών, στη βάση του δεδομένου x. Η πλειοψηφία των τεχνικών ταξινόμησης μπορεί να θεωρηθεί ως μια άμεση ή έμμεση προσέγγιση του παραπάνω δείκτη πιθανοφάνειας.

2.4.3 Κόστος εσφαλμένης ταξινόμησης

Έστω ότι το κόστος εσφαλμένης ταξινόμησης της κατηγορίας A_i ως A_j είναι c(i,j). Οι αποφάσεις βασίζονται στην αρχή ότι το συνολικό κόστος εσφαλμένης ταξινόμησης θα πρέπει να ελαχιστοποιηθεί (για κάθε νέα παρατήρηση, αυτό σημαίνει την ελαχιστοποίηση του προσδοκώμενου κόστους εσφαλμένης ταξινόμησης).

Αρχικά, έστω ότι το προσδοκώμενο κόστος εφαρμογής του κανόνα αθέτησης είναι: *κατανομή όλων των νέων παρατηρήσεων στην κατηγορία A_d,* χρησιμοποιώντας το δείκτη d ως αναφορά για την κατηγορία απόφασης. Όταν η απόφαση A_d πραγματοποιείται για κάθε νέα παρατήρηση, τότε το κόστος c(i,d) συμβαίνει με πιθανότητα π_i. Οπότε το προσδοκώμενο κόστος C_d της απόφασης A_d είναι:

$$C_d = \sum_i \pi_i c(i,d)$$

Σύμφωνα με τον κανόνα του ελάχιστου κόστους του Bayes επιλέγεται εκείνη η κατηγορία με το ελάχιστο προσδοκώμενο κόστος. Στη σχέση μεταξύ των κανόνων του ελάχιστου σφάλματος και του ελάχιστου κόστους υποτίθεται ότι το κόστος εσφαλμένης ταξινόμησης είναι το ίδιο για όλα τα σφάλματα και μηδέν για τις σωστές ταξινομήσεις (π.χ. έστω c(i,j)=c για $i \neq j$ και c(i,j)=0 για i=j). Τότε το προσδοκώμενο κόστος είναι:

$$C_d = \sum_i \pi_i c(i,d) = \sum_{i \neq j} \pi_i c = c \sum_{i \neq j} \pi_i = c(1 - \pi_\delta)$$

όπου π_δ είναι η συνολική πιθανότητα σωστής ταξινόμησης κάθε νέας παρατήρησης στην κατηγορία A_d. Ο κανόνας του ελάχιστου κόστους εντοπίζει την κατηγορία με τη μεγαλύτερη αρχική πιθανότητα.

Το κόστος εσφαλμένης ταξινόμησης είναι δύσκολο να προσδιοριστεί στην πράξη. Ακόμη και σε περιπτώσεις που είναι εμφανείς μεγάλες ανισότητες στο μέγεθος των πιθανών ποινών ή ανταμοιβών για την πραγματοποίηση μιας σωστής ή λανθασμένης απόφασης, είναι συχνά δύσκολη η ποσοτικοποίηση του κόστους.

2.5 Ο κανόνας του Bayes

Στη συνέχεια παρουσιάζεται ο τρόπος με τον οποίο τα παραπάνω τρία συστατικά συνδυάζονται σε μια διαδικασία ταξινόμησης.

Όταν υπάρχει μια πληροφορία x για ένα αντικείμενο, οι πιθανότητες θεωρούνται ως δεσμευμένες και εκφράζονται συναρτήσει του x. Επιπλέον, ο κανόνας απόφασης με τη μικρότερη πιθανότητα σφάλματος εντοπίζει την κατηγορία με τη μεγαλύτερη συχνότητα, αλλά τώρα η σχετική πιθανότητα είναι η δεσμευμένη πιθανότητα $p(A_i \mid x)$ της κατηγορίας A_i δεδομένου του x (για τις πιθανότητες $p(A_i)=\pi_i$ και $p(A_i|x)$ συχνά χρησιμοποιούνται οι όροι prior (εκ των προτέρων) και posterior (εκ των υστέρων)).

Εάν επιθυμείται η χρήση ενός κανόνα ελάχιστου κόστους, θα πρέπει πρώτα να εκτιμηθούν τα προσδοκώμενα κόστη των διαφόρων αποφάσεων δοθέντος, της πληροφορίας x.

Για κάθε αντικείμενο x της κατηγορία A_d, το κόστος εσφαλμένης επιλογής c(i,d) για την κατηγορία A_i συμβαίνει με πιθανότητα $p(A_i|x)$. Όπως οι πιθανότητες $p(A_i \mid x)$ εξαρτώνται από το x, το ίδιο ισχύει και για τους κανόνες απόφασης. Οπότε, το προσδοκώμενο κόστος $C_d(x)$ επιλέγοντας την απόφαση A_d είναι:

$$C_d(x) = \sum_i p(A_i \mid x)\, c(i,d)$$

Στην ειδική περίπτωση όπου τα κόστη εσφαλμένων ταξινομήσεων είναι ίσα, ο κανόνας του ελάχιστου κόστους εντοπίζει την κατηγορία με τη μεγαλύτερη εκ των υστέρων πιθανότητα.

Σύμφωνα με τη θεωρία του Bayes οι δεσμευμένες πιθανότητες $p(A_i \mid x)$ των κατηγοριών εκτιμώνται εφόσον είναι γνωστές οι εκ των προτέρων πιθανότητες π_i, και στη συνέχεια εκτιμώνται οι δεσμευμένες πιθανότητες $P(x|A_i)$ για κάθε κατηγορία A_i. Κατά συνέπεια, για κάθε κατηγορία A_i θεωρείται ότι η πιθανότητα του δεδομένου x είναι $P(x|A_i)$. Η θεωρία του Bayes εκτιμά την εκ των υστέρων πιθανότητα $p(A_i \mid x)$ για κάθε κατηγορία A_i ως εξής:

$$p(A_i \mid x) = \pi_i P(x \mid A_i) / \sum_j \pi_j P(x \mid A_j)$$

Ο παρανομαστής είναι κοινός για όλες τις κατηγορίες, οπότε το $p(A_i \mid x)$ είναι ανάλογο του $\pi_i P(x|A_i)$. Η κατηγορία A_d με το ελάχιστο προσδοκώμενο κόστος (ελάχιστος κίνδυνος) είναι αυτή για την οποία ελαχιστοποιείται το ακόλουθο μέγεθος:

$$\sum_i \pi_i c(i,d) P(x \mid A_i)$$

Θεωρώντας ότι τα χαρακτηριστικά έχουν συνεχείς κατανομές, οι παραπάνω πιθανότητες γίνονται πυκνότητες πιθανότητα. Υποθέτοντας ότι οι παρατηρήσεις από τον πληθυσμό A_i έχουν συναρτήσεις πυκνότητας πιθανότητα $f_i(x) = f(x \mid A_i)$, τότε σύμφωνα με την θεωρία του Bayes, η πιθανότητα να ανήκει μια παρατήρηση x στην κατηγορία A_i είναι:

$$p(A_i \mid x) = \pi_i f_i(x) / \sum_j \pi_i f_i(x)$$

Στην συνέχεια, ο κανόνας ταξινόμησης τοποθετεί το x στην κατηγορία A_d με πιθανότητα:

$$p(A_d \mid x) = \max_i p(A_i \mid x)$$

Όπως και νωρίτερα, η κατηγορία A_d με το ελάχιστο προσδοκώμενο κόστος (ελάχιστος κίνδυνος) είναι αυτή για την οποία ελαχιστοποιείται το ακόλουθο μέγεθος:

$$\sum_i \pi_i c(i,d) f_i(x)$$

2.5.1 Ο κανόνας του Bayes όπως εφαρμόζεται στη στατιστική

Το $p(A_i \mid x)$ θα μπορούσε επίσης να προέρχεται από έναν εμπειρικό κανόνα, αντί για τον κανόνα του Bayes, αλλά αυτό θα απαιτούσε πολύ μεγάλο αριθμό δεδομένων. Ωστόσο, κατά κανόνα, συλλέγονται όλα τα δεδομένα που έχουν τα ίδια χαρακτηριστικά, από το δείγμα εκμάθησης, και στη συνέχεια εκτιμάται η πιθανότητα $p(A_i \mid x)$. Τέλος, ο κανόνας του ελάχιστου σφάλματος εντοπίζει την κατηγορία A_d με τη μεγαλύτερη πιθανότητα $P(A_d|x)$.

Οι αποφάσεις που βασίζονται στον κανόνα του Bayes θεωρούνται βέλτιστες επειδή ο κανόνας αυτός δίνει μικρότερο κάτω όριο για τους δείκτες των σφαλμάτων. Επίσης, ο κανόνας του Bayes παρέχει τη λογική βάση για όλους τους στατιστικούς αλγορίθμους, επειδή υποθέτει ότι η συνολική πληροφορία είναι γνωστή, σχετικά με τη στατιστική κατανομή σε κάθε κατηγορία. Οι στατιστικές διαδικασίες προσπαθούν να εκτιμήσουν την ελλιπή πληροφορία για την κατανομή, με ποικίλους τρόπους, αλλά υπάρχουν δύο κύριοι μέθοδοι: παραμετρικές και μη παραμετρικές. Οι παραμετρικές μέθοδοι υποθέτουν την φύση της κατανομής (συχνότερα υποτίθεται ότι οι κατανομές ακολουθούν την κανονική κατανομή) και το πρόβλημα περιορίζεται στην εκτίμηση των παραμέτρων των κατανομών (μέσες τιμές και διακυμάνσεις). Οι μη παραμετρικές μέθοδοι δεν κάνουν καμία υπόθεση για τις κατανομές που εμπλέκονται, και γι' αυτό θεωρούνται και περισσότερο ακριβείς.

ΚΕΦΑΛΑΙΟ 3

Μέθοδοι ταξινόμησης

3.1 Εισαγωγή

Η αυξημένη σημαντικότητα του προβλήματος της ταξινόμησης τόσο σε πρακτικό όσο και σε ερευνητικό επίπεδο, έχει ελκύσει το ενδιαφέρον πολλών ερευνητών από διάφορους επιστημονικούς χώρους. Ωστόσο, η ευρύτητα του προβλήματος της ταξινόμησης, καθιστά δύσκολη την πλήρη ανάλυση όλων των μεθοδολογικών προσεγγίσεων που έχουν κατά καιρούς αναπτυχθεί. Για το λόγο αυτό, η συγκεκριμένη ερευνητική διατριβή επικεντρώνεται στις ευρύτερα διαδεδομένες προσεγγίσεις, βάσει των ερευνητικών και πρακτικών τους εφαρμογών. Οι εξεταζόμενες προσεγγίσεις διακρίνονται σε δύο βασικές κατηγορίες:

- ✓ Στις στατιστικές και οικονομετρικές προσεγγίσεις, οι οποίες αποτελούν τον «παραδοσιακό» τρόπο αντιμετώπισης του προβλήματος της ταξινόμησης.
- ✓ Στις μη παραμετρικές προσεγγίσεις οι οποίες έχουν προταθεί κατά τις τελευταίες δύο δεκαετίες ως καινοτόμες και αποτελεσματικές τεχνικές ανάπτυξης υποδειγμάτων ταξινόμησης.

3.2 Στατιστικές και οικονομετρικές προσεγγίσεις

Οι στατιστικές και οικονομετρικές προσεγγίσεις παραμένουν, ακόμη και σήμερα, ιδιαίτερα διαδεδομένες, τόσο σε ερευνητικό όσο και σε πρακτικό επίπεδο. Το πλήθος των υπολογιστικών προγραμμάτων που είναι διαθέσιμα συμβάλλουν στην εύκολη εφαρμογή των προσεγγίσεων αυτών. Επιπρόσθετα, ιδιαίτερα διαδεδομένη είναι και η χρήση τους σε συγκριτικές έρευνες, οι οποίες στόχο έχουν την αξιολόγηση της αποτελεσματικότητας νέων τεχνικών ταξινόμησης που αναπτύσσονται. Προς την κατεύθυνση αυτή, οι στατιστικές-οικονομετρικές προσεγγίσεις αποτελούν σημείο αναφοράς βάσει του οποίου πραγματοποιούνται οι συγκρίσεις νέων τεχνικών ταξινόμησης.

3.2.1 Γραμμική Διακριτική Ανάλυση

Η γραμμική διακριτική ανάλυση (Fisher, 1936) είναι μια εμπειρική μέθοδος ταξινόμησης, η οποία χρησιμοποιεί ως δείγμα εκμάθησης ένα σύνολο εναλλακτικών δραστηριοτήτων, των οποίων η ταξινόμηση είναι γνωστή. Σκοπός της μεθόδου είναι η ανάπτυξη μιας σειράς διακριτικών συναρτήσεων οι οποίες μεγιστοποιούν τη διακύμανση μεταξύ των κατηγοριών σε σχέση με τη διακύμανση εντός των κατηγοριών.

Στην περίπτωση που υπάρχουν δύο κατηγορίες, έστω μ ο συνολικός μέσος και μ_1, μ_2 οι μέσες τιμές των διανυσμάτων των χαρακτηριστικών, και έστω α_1, ..., α_n ένα σύνολο συντελεστών. Τότε η γραμμική διάκριση μεταξύ των δύο κατηγοριών θα δίνεται από την ακόλουθη εξίσωση:

$$g(\mu) = \sum a_j \mu_j \qquad (3.1)$$

Η διακύμανση μεταξύ των κατηγοριών $(g(\mu_1) - g(\mu_2))$ πρέπει να είναι όσο το δυνατόν μεγαλύτερη και αυτό επιτυγχάνεται, εάν διαιρεθεί με την τυπική απόκλιση s_g:

$$\frac{g(\mu_1) - g(\mu_2)}{s_g} \qquad (3.2)$$

Αυτό το μέγεθος διάκρισης σχετίζεται με την εκτίμηση της εσφαλμένης ταξινόμησης, εάν θεωρηθεί ότι το g(μ) ακολουθεί την κανονική κατανομή. Η πιθανότητα Φ της εσφαλμένης ταξινόμησης μιας εναλλακτικής δραστηριότητας είναι:

$$\Phi\left(\frac{g(\mu_1) - g(\mu_2)}{2s_g} \right) \qquad (3.3)$$

Η γραμμική διακριτική ανάλυση είναι εύκολη στην εφαρμογή, όταν πρόκειται για δύο μόνο κατηγορίες ταξινόμησης. Τότε, το πρόβλημα είναι ισοδύναμο με ένα

πρόβλημα πολλαπλής παλινδρόμησης, όπου τα χαρακτηριστικά προβλέπουν την τιμή της κατηγορίας. Συγκεκριμένα, οι κατηγορίες θεωρούνται ως αριθμητικές μεταβλητές. Για παράδειγμα, αν στην κατηγορία A_1 δίνεται η τιμή 0 και στην κατηγορία A_2 η τιμή 1, χρησιμοποιώντας ένα πακέτο πολλαπλής παλινδρόμησης, εκτιμάται η τιμής της κατηγορίας. Εάν οι δύο κατηγορίες είναι ισοπίθανες, τότε η περιοχή που περιέχει τις κατηγορίες διχοτομείται. Διαφορετικά, η γραμμή διάκρισης είναι πλησιέστερη στην κατηγορία με την μικρότερη συχνότητα.

Η γραμμική διακριτική ανάλυση μπορεί να θεωρηθεί και ως μια μέθοδος μέγιστης πιθανοφάνειας. Ειδικότερα, τα διανύσματα των χαρακτηριστικών της κατηγορίας A_i θεωρούνται ότι είναι ανεξάρτητα και ότι ακολουθούν μια συγκεκριμένη κατανομή πιθανότητας με συνάρτηση πυκνότητας πιθανότητας f_i. Ένα νέο αντικείμενο με διάνυσμα χαρακτηριστικών x, ταξινομείται στην κατηγορία της οποίας η πιθανότητα $f_i(x)$ είναι μεγαλύτερη. Μια συχνή υπόθεση που πραγματοποιείται είναι ότι οι κατανομές είναι κανονικές με διαφορετικές μέσες τιμές αλλά με ίδιο πίνακα συνδιακύμανσης. Εάν η κατανομή είναι κανονική, τότε η εξίσωση που χρησιμοποιείται για την διάκριση των κατηγοριών είναι η εξής:

$$\frac{1}{\sqrt{|2\pi\Sigma|}}\exp(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)) \qquad (3.4)$$

όπου μ είναι ένα διάνυσμα n-διαστάσεων των μέσων τιμών της κατηγορίας, και Σ ένας n x n πίνακας συνδιακύμανσης (απαραίτητα θετικά ορισμένος). Στην περίπτωση που οι πιθανότητες των συναρτήσεων πυκνότητας είναι ίσες, τότε η διάκριση των δύο κατηγοριών πραγματοποιείται με την ακόλουθη εξίσωση:

$$x^{T}\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)^{\mathrm{T}}\Sigma^{-1}(\mu_1-\mu_2)=0 \qquad (3.5)$$

όπου μ_i δηλώνει τον πληθυσμό των μέσων της κατηγορίας A_i. Ωστόσο, στην ταξινόμηση, η ακριβής κατανομή δεν είναι συνήθως γνωστή, και είναι απαραίτητη η εκτίμηση των παραμέτρων των κατανομών.

Στην περίπτωση της ύπαρξης περισσότερων των δύο κατηγοριών, δεν είναι πλέον δυνατή η χρησιμοποίηση του ενός γραμμικού διακριτικού σκορ, για τον

διαχωρισμό των κατηγοριών. Μια απλή διαδικασία είναι η εκτίμηση ενός σκορ για την κάθε κατηγορία. Το σκορ αυτό υπολογίζεται από τη συνάρτηση πιθανότητας πυκνότητας, χωρίς όμως να περιέχει τους σταθερούς όρους..

Στην περίπτωση ύπαρξης περισσότερων των δύο κατηγοριών, έστω ότι η κοινή πιθανότητα της κατηγορίας A_i και του χαρακτηριστικού x είναι $\pi_i f_i(x)$, τότε η διάκριση μεταξύ των κατηγοριών πραγματοποιείται από την παρακάτω εξίσωση:

$$\log \pi_i + x^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i \qquad (3.6)$$

Ο συντελεστής β_i δοθέντος των συντελεστών του x είναι**:**

$$\beta_i = \Sigma^{-1} \mu_i \qquad (3.7)$$

Η αθροιστική σταθερά α_i εκτιμάται ως εξής:

$$a_i = \log \pi_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_t \qquad (3.8)$$

Οι εκτιμήσεις του συντελεστή και του σταθερού όρου δεν είναι μοναδικές, επειδή είναι δυνατή η ανάπτυξη μιας σειράς εναλλακτικών διακριτικών συναρτήσεων των οποίων οι συντελεστές β_i' και α_i' μπορούν να προκύψουν ως γραμμικοί μετασχηματισμοί των β_i και α_i.

3.2.2 Τετραγωνική Διακριτική Ανάλυση

Η τετραγωνική διακριτική ανάλυση είναι παρόμοια με τη γραμμική, μόνο που τα όρια των περιοχών διάκρισης των κατηγοριών είναι διαφορετικά, και η υπόθεση των ίσων πινάκων συνδιακύμανσης απορρίπτεται. Οι Clarke et al. (1979) θεωρούν ότι η τετραγωνική διακριτική ανάλυση είναι ευσταθής και ότι η μη κανονική κατανομή δεν μειώνει την ακρίβεια της μεθόδου. Ωστόσο, ο αριθμός των παραμέτρων που πρέπει να εκτιμηθεί αυξάνεται σε qn(n+1)/2 (q είναι ο αριθμός των κατηγοριών), και

πρέπει να ληφθεί υπόψη η διαφορά μεταξύ των διακυμάνσεων, ιδιαίτερα στα μικρά και μεσαίου μεγέθους σύνολα δεδομένων (Mark & Dunn, 1974). Περιστασιακά, οι διαφορές μεταξύ των συνδιακυμάνσεων είναι μικρής κλίμακας και μπορούν να γίνουν απλοποιήσεις (Kendall et al., 1983). Επιπλέον, η γραμμική διακριτική ανάλυση θεωρείται αποτελεσματική, εάν η απόκλιση από την ισότητα των συνδιακυμάνσεων είναι μικρή (Glbert, 1969).

Η τετραγωνική διακριτική συνάρτηση ορίζεται ως ο λογάριθμος της πυκνότητας πιθανότητας της κάθε κατηγορίας. Η εκτίμηση της τετραγωνικής διάκρισης πραγματοποιείται με την ακόλουθη εξίσωση:

$$\log \pi_i f_i(x) = \log \pi_i - \frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(x-\mu_t)^{\mathrm{T}} \Sigma_i^{-1}(x-\mu_i) \qquad (3.9)$$

Στην παραπάνω εξίσωση ο δείκτης i αναφέρεται στο δείγμα τιμών της κατηγορίας A_i.

Για κάθε κατηγορία εκτιμάται το σκορ διάκρισης και επιλέγεται αυτή με το μεγαλύτερο σκορ. Η εύρεση των τελικών δεσμευμένων πιθανοτήτων των κατηγοριών, πραγματοποιείται με την παρακάτω εξίσωση:

$$P(A_i \mid x) = \exp[\log(\pi_i) - \frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)] \qquad (3.10)$$

Το πιο συχνό πρόβλημα που παρουσιάζεται στην τετραγωνική διακριτική ανάλυση είναι όταν κάποιο χαρακτηριστικό έχει μηδενική διακύμανση σε μια κατηγορία, οπότε ο πίνακας συνδιακύμανσης δεν μπορεί να αντιστραφεί. Ένας τρόπος αποφυγής αυτού του προβλήματος, είναι η πρόσθεση μιας μικρής θετικής σταθεράς στους διαγώνιους όρους του πίνακα συνδιακύμανσης.

Τα κύρια προβλήματα της τετραγωνικής διάκρισης είναι ο μεγάλος αριθμός των παραμέτρων που πρέπει να εκτιμηθεί, καθώς και η παρουσία των μηδενικών ιδιοτιμών στους πίνακες συνδιακύμανσης. Για την επίλυση των προβλημάτων αυτών, ο Friedman (1989) πρότεινε μια συμβιβαστική λύση μεταξύ της γραμμικής και τετραγωνικής διάκρισης, μέσω της εκτίμησης δύο οικογενειών παραμέτρων.

Η μία παράμετρος ελέγχει την ομαλότητα του πίνακα συνδιακύμανσης ως εξής:

$$(1-\delta_i)\Sigma_i + \delta_i\Sigma \qquad (3.11)$$

όπου Σ_i είναι ο πίνακας συνδιακύμανσης της κατηγορίας i και Σ ο συνολικός πίνακας συνδιακύμανσης. Όταν το δ_i είναι μηδέν, τότε οι πίνακες συνδιακύμανσης είναι διαφορετικοί για την κάθε κατηγορία, ενώ όταν το δ_i είναι μονάδα, τότε είναι ίσοι για όλες τις κατηγορίες.

Η άλλη παράμετρος λ είναι μια σταθερά που προστίθεται στους διαγώνιους όρους του πίνακα συνδιακύμανσης, έτσι ώστε να είναι αναστρέψιμος. Όπως έχει αναφερθεί, οποιαδήποτε ιδιομορφία στον πίνακα συνδιακύμανσης μπορεί να προκαλέσει προβλήματα, καθώς πρέπει για την κάθε κατηγορία να εκτιμηθεί ο πίνακας συνδιακύμανσης. Ο κίνδυνος της ιδιομορφίας αυξάνει όταν οι κατηγορίες περιέχουν μικρού μεγέθους δείγματα.

3.2.3 Λογιστικό Υπόδειγμα Πιθανότητας (LOGIT)

Η μέθοδος αυτή βασίζεται σε μία αθροιστική συνάρτηση πιθανότητας και παρέχει την πιθανότητα ένα αντικείμενο να ανήκει σε μια από τις προκαθορισμένες κατηγορίες.

Κλασικά, προβλήματα δυαδικής ταξινόμησης χρησιμοποιούν το λογιστικό (logit) υπόδειγμα πιθανότητας. Το μοντέλο logit χρησιμοποιεί την αθροιστική λογιστική κατανομή. Ένα πλεονέκτημα του μοντέλου είναι ότι απαιτεί σημαντικά απλούστερες υπολογιστικές διαδικασίες βελτιστοποίησης για την εκτίμηση των παραμέτρων του, σε σχέση με το κανονικό υπόδειγμα πιθανότητας (PROBIT).

Σύμφωνα με το λογιστικό υπόδειγμα πιθανότητας (logit) και θεωρώντας δύο κατηγορίες, ορίζεται η παρακάτω εξίσωση:

$$\log\frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} = a + \beta' x \qquad (3.12)$$

όπου π_1 και π_2 είναι οι αρχικές πιθανότητες, f η λογιστική συνάρτηση, α και β είναι οι παράμετροι του μοντέλου που πρέπει να εκτιμηθούν. Η περίπτωση της κανονικής κατανομής με ίσους πίνακες συνδιακύμανσης αποτελεί μια ειδική περίπτωση, όπου οι

παράμετροι εκφράζονται συναρτήσει των πιθανοτήτων, των μέσων τιμών των κατηγοριών και του πίνακα συνδιακύμανσης. Ωστόσο, το λογιστικό μοντέλο καλύπτει και άλλες περιπτώσεις, όπως όταν τα αντικείμενα του δείγματος είναι ανεξάρτητα και παίρνουν τις τιμές 0 ή 1.

Στην πράξη, οι παράμετροι εκτιμώνται με την μέθοδο της μέγιστης πιθανότητας. Δεδομένου των τιμών των αντικειμένων x, οι δεσμευμένες πιθανότητες για τις κατηγορίες A_1 και A_2, είναι οι ακόλουθες:

$$P(A_1 \mid x) = \frac{e^{(a+\beta' x)}}{1+e^{(a+\beta' x)}}$$
$$P(A_2 \mid x) = \frac{1}{1+e^{(a+\beta' x)}} \qquad (3.13)$$

θεωρώντας ότι τα δείγματα εκμάθησης των δύο κατηγοριών είναι ανεξάρτητα, η δεσμευμένη πιθανότητα για τις παραμέτρους α και β, ορίζεται ως ακολούθως:

$$L(a,\beta) = \prod_{\{A_1\ δείγμα\}} P(A_1 \mid x) \quad * \prod_{\{A_2\ δείγμα\}} P(A_2 \mid x) \qquad (3.14)$$

Οι παράμετροι παίρνουν τις τιμές που μεγιστοποιούν την παραπάνω πιθανοφάνεια. Συνήθως, οι τιμές των παραμέτρων βρίσκονται μέσω επαναληπτικών μεθόδων, όπως προτείνουν και οι Cox (1966), Day & Kerridge (1967).

Στην περίπτωση που η ταξινόμηση αφορά περισσότερες των δύο κατηγοριών τότε το λογιστικό μοντέλο εφαρμόζεται υπό δύο μορφές: την πολλαπλή ονομαστική (multinomial) και τη διατεταγμένη (ordered).

Το διατεταγμένο λογιστικό υπόδειγμα οδηγεί στον υπολογισμό ενός διανύσματος συντελεστών β και ενός διανύσματος σταθερών όρων α, βάσει των οποίων η πιθανότητα P_{kj} να ανήκει το αντικείμενο x_j στην κατηγορία A_k υπολογίζεται με τον ακόλουθο τρόπο:

$$\begin{aligned} P_{1j} &= f(a_1 + \beta_j' x) \\ P_{2j} &= f(a_2 + \beta_j' x) - f(a_1 + \beta_j' x) \\ &\vdots \\ P_{kj} &= 1 - (P_{1j} + P_{2j} + \ldots + P_{k-1,j}) \end{aligned} \qquad (3.15)$$

Οι σταθεροί όροι ορίζονται έτσι ώστε: $\alpha_{k-1}>\alpha_{k-2}>\ldots>\alpha_2>0$ ($\alpha_1=0$). Ο υπολογισμός των παραμέτρων πραγματοποιείται μέσω τεχνικών μέγιστης πιθανοφάνειας, κατά παρόμοιο τρόπο με την περίπτωση των δύο κατηγοριών.

Σε αντίθεση με το διατεταγμένο λογιστικό υπόδειγμα, το ονομαστικό οδηγεί στην ανάπτυξη ενός συνόλου διανυσμάτων συντελεστών b_k και α_k για κάθε κατηγορία A_k (k=1,2,…,q). Οπότε η πιθανότητα P_{kj} να ανήκει το αντικείμενο x_j στην κατηγορία A_k, υπολογίζεται βάσει της σχέσης:

$$P_{kj} = \frac{e^{\beta_j' x_k + a_k}}{\sum_{1}^{q} e^{\beta_j' x_k + a_l}} \qquad (3.16)$$

Για λόγους κανονικοποίησης των παραμέτρων του υποδείγματος, τίθενται $\beta_1=0$ και $\alpha_1=0$, ενώ τα υπόλοιπα β_k και α_k υπολογίζονται μέσω τεχνικών μέγιστης πιθανοφάνειας.

Μία από τις πρώτες εφαρμογές του logit μοντέλου ήταν αυτή του Martin, 1977 για την πρόβλεψη της χρηματοοικονομικής αποτυχίας στο χώρο των τραπεζικών ιδρυμάτων.

3.3 Πολυκριτήριες προσεγγίσεις ταξινόμησης

Η πολυκριτήρια ανάλυση αποφάσεων αποτελεί έναν εξελιγμένο χώρο της επιχειρησιακής έρευνας. Βασικό ρόλο στην ανάπτυξη και διάδοση της πολυκριτήριας ανάλυσης αποτελεί η απλή διαπίστωση ότι η επίλυση πολύπλοκων και ιδιαίτερα σημαντικών προβλημάτων λήψης αποφάσεων δεν είναι δυνατό να πραγματοποιείται μέσω μιας μονοδιάστατης ανάλυσης. Η κύρια διαφορά της πολυκριτήριας ανάλυσης από άλλες εναλλακτικές προσεγγίσεις είναι η σύνθεση των παραμέτρων του προβλήματος υπό το πρίσμα της πολιτικής λήψης των αποφάσεων και του συστήματος προτιμήσεων και αξιών, το οποίο συνειδητά ή ασυνείδητα χρησιμοποιεί ο αποφασίζων.

3.3.1 Η μέθοδος UTADIS

Η μέθοδος UTADIS αποτελεί προσαρμογή της μεθόδου UTA (Jacquet – Lagreze και Siskos, 1982) στην περίπτωση όπου σκοπός δεν είναι η κατάταξη των εναλλακτικών δραστηριοτήτων, αλλά η ταξινόμηση τους σε προκαθορισμένες ομοιογενείς κατηγορίες. Οι κατηγορίες είναι διατεταγμένες από τις καλύτερες προς τις χειρότερες ως εξής:

$$A_1 \succ A_2 \succ ... \succ A_q$$

Ως A_1 συμβολίζεται η κατηγορία που αποτελείται από τις καλύτερες εναλλακτικές δραστηριότητες, ενώ στην τελευταία κατηγορία A_q ταξινομούνται οι χειρότερες εναλλακτικές δραστηριότητες. Προβλήματα ταξινόμησης στα οποία οι κατηγορίες ορίζονται κατά διατεταγμένο (ordinal) και όχι ονομαστικό (nominal) τρόπο εμφανίζονται ιδιαίτερα συχνά σε προβλήματα λήψης αποφάσεων. Χαρακτηριστικά αναφέρεται το πρόβλημα της αξιολόγησης αιτήσεων δανειοδότησης.

Σκοπός της μεθόδου είναι η ανάπτυξη ενός υποδείγματος σύνθεσης των κριτηρίων αξιολόγησης έτσι ώστε το αποτέλεσμα της σύνθεσης αυτής να αποδίδει υψηλά σκορ στις εναλλακτικές δραστηριότητες της κατηγορίας A_1 και σταδιακά χαμηλότερα σκορ στις δραστηριότητες που ανήκουν στις χαμηλότερες κατηγορίες.

Το υπόδειγμα σύνθεσης των κριτηρίων που χρησιμοποιείται στην UTADIS είναι το ακόλουθο:

$$U(g) = \sum_{i=1}^{n} p_i \, u_i(g_i) \qquad (3.17)$$

Όπου:

$g=(g_1,g_2, \ldots,g_n)$ είναι το διάνυσμα των n κριτηρίων αξιολόγησης

p_i είναι το βάρος του κριτηρίου g_i

$u_i(g_i)$ είναι η συνάρτηση μερικής χρησιμότητας του g_i.

Οι συναρτήσεις μερικών χρησιμοτήτων (marginal utility function) είναι μονότονες συναρτήσεις οριζόμενες στην κλίμακα του κάθε κριτηρίου αξιολόγησης.

Οι συναρτήσεις αυτές δύνανται να έχουν οποιαδήποτε μορφή, γραμμική ή μη γραμμική και ικανοποιούν τις ακόλουθες δύο βασικές συνθήκες:

$u_i(g_{i*})=0$
$u_i(g_i^*)=1$

Όπου, ως g_{i*} και g_i^* ορίζονται, αντίστοιχα, η λιγότερο και η περισσότερη προτιμητέα τιμή του κριτηρίου g_i.

Η αναγωγή των επιδόσεων των εναλλακτικών δραστηριοτήτων στα κριτήρια αξιολόγησης σε όρους χρησιμότητας, μέσω του ορισμού των κατάλληλων συναρτήσεων μερικών χρησιμοτήτων παρέχει τα ακόλουθα δύο βασικά πλεονεκτήματα:

1. Επιτρέπει τη μοντελοποίηση και αναπαράσταση στο υπό αναπτυσσόμενο υπόδειγμα της μη γραμμικής συμπεριφοράς του αποφασίζοντος κατά την αξιολόγηση των εναλλακτικών δραστηριοτήτων.
2. Επιτρέπει την αξιοποίηση ποιοτικών κριτηρίων αξιολόγησης χωρίς να απαιτείται η ποσοτικοποίηση τους μέσω του ορισμού μιας ποιοτικής κλίμακας.

Τα βασικά συστατικά στοιχεία του υποδείγματος ταξινόμησης που αναπτύσσεται μέσω της μεθόδου UTADIS περιλαμβάνουν τα βάρη των κριτηρίων αξιολόγησης και τη μορφή των μερικών συναρτήσεων χρησιμότητας. Τα δύο αυτά στοιχεία καθορίζουν τη μορφή της αναπτυσσόμενης προσθετικής συνάρτησης χρησιμότητας. Παράλληλα όμως, βασικό στοιχείο του αναπτυσσόμενου υποδείγματος ταξινόμησης, αποτελούν και τα όρια χρησιμότητας βάσει των οποίων λαμβάνεται η απόφαση για την ταξινόμηση των εναλλακτικών δραστηριοτήτων.

Ο καθορισμός αυτών των συστατικών στοιχείων του αναπτυσσόμενου υποδείγματος ταξινόμησης πραγματοποιείται στα γενικά πλαίσια που διέπουν την αναλυτική – συνθετική προσέγγιση. Η αναλυτική – συνθετική προσέγγιση αναφέρεται στην ανάλυση των ολικών προτιμήσεων του αποφασίζοντα ώστε να καθοριστεί το μοντέλο σύνθεσης των κριτηρίων βάσει του οποίου καταλήγει στις αποφάσεις που λαμβάνει. Το μοντέλο σύνθεσης των κριτηρίων που αναπτύσσεται είναι μια συνάρτηση χρησιμότητας η οποία συνήθως είναι προσθετικής μορφής. Η ανάπτυξη αυτής της συνάρτησης γίνεται χρησιμοποιώντας τεχνικές μονότονης παλινδρόμησης και γραμμικού προγραμματισμού. Πιο συγκεκριμένα, αρχικά

χρησιμοποιείται ένα σύνολο αναφοράς (reference set, το σύνολο αναφοράς είναι το αντίστοιχο του δείγματος εκμάθησης)αποτελούμενο από εναλλακτικές ενέργειες, οι οποίες αξιολογούνται από τον αποφασίζοντα (ταξινομούνται σε ομάδες) ανάλογα με τις προτιμήσεις του, τις εμπειρίες του και την πολιτική που ακολουθεί. Οι εναλλακτικές ενέργειες που περιλαμβάνονται στο σύνολο αναφοράς μπορούν να είναι είτε ένα υποσύνολο των εναλλακτικών ενεργειών που εξετάζονται, είτε εναλλακτικές ενέργειες που ήδη έχουν αξιολογηθεί από τον αποφασίζοντα. Στη συνέχεια, έχοντας ως δεδομένη την αξιολόγηση (ταξινόμηση σε ομάδες) των εναλλακτικών ενεργειών του συνόλου αναφοράς από τον αποφασίζοντα χρησιμοποιούνται τεχνικές γραμμικού προγραμματισμού ώστε να γίνει η ανάπτυξη της προσθετικής συνάρτησης χρησιμότητας ώστε να ελαχιστοποιηθούν οι διαφορές μεταξύ της ταξινόμησης του αποφασίζοντα και της ταξινόμησης που επιτυγχάνεται βάσει της προσθετικής συνάρτησης χρησιμότητας. Παρόμοια μεθοδολογία ακολουθείται και σε περιπτώσεις όπου το εξεταζόμενο πρόβλημα απαιτεί την κατάταξη των εναλλακτικών ενεργειών από τις καλύτερες προς τις χειρότερες. Λεπτομερής περιγραφή της μεθόδου μπορεί να βρεθεί στις εργασίες των Jacquet – Lagrèze (1995) και Doumpos and Zopounidis (2002).

3.3.2 Η μέθοδος M.H.DIS.

Η μέθοδος M.H.DIS (Zopounidis and Doumpos, 2000) αντιμετωπίζει το πρόβλημα της ταξινόμησης μέσω μιας ιεραρχικής διαδικασίας, η οποία βασίζεται στην πολυκριτήρια ανάλυση αποφάσεων και σε τεχνικές μαθηματικού προγραμματισμού. Δηλαδή σε κάθε στάδιο της ιεραρχικής διαδικασίας επιλύονται δύο προβλήματα γραμμικού προγραμματισμού και ένα πρόβλημα μικτού – ακεραίου προγραμματισμού. Σκοπός της επίλυσης των προβλημάτων είναι η επίτευξη των ακόλουθων δύο στόχων:

1. Ανάπτυξη δύο συναρτήσεων χρησιμότητας οι οποίες διαχωρίζουν τα αντικείμενα της κατηγορίας A_k από τα αντικείμενα χειρότερων κατηγοριών A_{k+1}, A_{k+2},…,A_q, έτσι ώστε να ελαχιστοποιηθεί το πλήθος των αντικειμένων που ταξινομούνται σε λάθος κατηγορία.

2. Τροποποίηση των παραπάνω συναρτήσεων χρησιμότητας έτσι ώστε να μεγιστοποιηθεί η «διαύγεια» της ταξινόμησης. Ο στόχος αυτός είναι παρόμοιος με τη μεγιστοποίηση της διακύμανσης μεταξύ των κατηγοριών στη γνωστή διακριτική ανάλυση.

Οι δύο αυτοί στόχοι επιτυγχάνονται μέσω μιας λεξικογραφικής διαδικασίας, κατά την οποία αρχικά πραγματοποιείται η ελαχιστοποίηση του πλήθους των αντικειμένων που ταξινομούνται σε λάθος κατηγορία και στη συνέχεια πραγματοποιείται η μεγιστοποίηση της διαύγειας της ταξινόμησης.

Συγκεκριμένα, η διαδικασία ταξινόμησης που εφαρμόζεται από την M.H.DIS προβαίνει σταδιακά στην ταξινόμηση των αντικειμένων, ξεκινώντας από την κατηγορία A_1. Κατά το πρώτο στάδιο της αξιολόγησης, τα αντικείμενα που βρίσκονται να ανήκουν στην κατηγορία A_1, (σωστά ή εσφαλμένα), αποκλείονται από περαιτέρω εξέταση. Το ζήτημα του δευτέρου σταδίου είναι να βρεθούν τα αντικείμενα που ανήκουν στην κατηγορία A_2. Ξανά, τα αντικείμενα που ταξινομούνται σε αυτή την κατηγορία αποκλείονται από περαιτέρω εξέταση και η ίδια διαδικασία επαναλαμβάνεται μέχρι να ταξινομηθούν όλα τα αντικείμενα στις προκαθορισμένες κατηγορίες. Ο αριθμός των σταδίων της ιεραρχικής διαδικασίας ταξινόμησης είναι q–1 (όπου q είναι ο αριθμός των κατηγοριών).

Η απόφαση που αφορά την ταξινόμηση των αντικειμένων βασίζεται στην ανάπτυξη δύο συναρτήσεων χρησιμότητας σε κάθε στάδιο k της προαναφερθείσας ιεραρχικής διαδικασίας ταξινόμησης. Οι συναρτήσεις αυτές έχουν την ακόλουθη μορφή:

$$U_k(g) = \sum_{i=1}^{n} h_{ki}\, u_{ki}(g_i) \quad \text{και} \quad U_{\sim k}(g) = \sum_{i=1}^{n} h_{\sim ki} u_{\sim ki}(g_i) \qquad (3.18)$$

Η πρώτη συνάρτηση χρησιμότητας $U_k(g)$ χαρακτηρίζει όλα τα αντικείμενα που ανήκουν στην κατηγορία A_k, ενώ η δεύτερη $U_{\sim k}(g)$ χαρακτηρίζει όλα τα αντικείμενα που ανήκουν σε κατώτερη (χειρότερη) κατηγορία από αυτή της A_k κατά το στάδιο k της ταξινόμησης. Οι $u_{ki}(g_i)$ και $u_{\sim ki}(g_i)$ είναι οι μερικές συναρτήσεις χρησιμότητας του κάθε κριτηρίου g_i, και ορίζονται μεταξύ των τιμών 0 και 1, ενώ τα βάρη των κριτηρίων h_{ki} και $h_{\sim ki}$ έχουν συνολικό άθροισμα ίσο με την μονάδα. Οι μερικές συναρτήσεις χρησιμότητας $u_{ki}(g_i)$ και $u_{\sim ki}(g_i)$ ορίζονται κατά παρόμοια με τις μερικές χρησιμότητες στη μέθοδο UTADIS.

Εάν η ολική χρησιμότητα του αντικειμένου σύμφωνα με την $U_k(g)$ είναι μεγαλύτερη από την συνολική χρησιμότητα που εκτιμήθηκε σύμφωνα με την $U_{\sim k}(g)$, τότε το αντικείμενο ταξινομείται στην κατηγορία A_k. Διαφορετικά, εάν η ολική χρησιμότητα του αντικειμένου σύμφωνα με την συνάρτηση χρησιμότητας $U_{\sim k}(g)$ είναι μεγαλύτερη από την ολική χρησιμότητα που εκτιμήθηκε με βάση τη $U_k(g)$, τότε το αντικείμενο δεν ταξινομείται στην κατηγορία A_k. Μία τέτοια περίπτωση επιδεικνύει ότι το αντικείμενο θα πρέπει να ταξινομηθεί σε μία από τις κατηγορίες A_{k+1}, $A_{k+2},\ldots,A_q$, και αυτό θα καθοριστεί κατά τη διάρκεια των υπόλοιπων σταδίων της ιεραρχικής διαδικασίας ταξινόμησης.

Η αντιμετώπιση του πρώτου στόχου της μεθόδου, αυτού της ελαχιστοποίησης των εσφαλμένων ταξινομήσεων, απαιτεί την ελαχιστοποίηση της ακόλουθης συνάρτησης:

$$EC = \sum_{i=1}^{N_k} I_{ki} + \sum_{i=1}^{N_{\sim k}} I_{\sim ki} \qquad (3.19)$$

όπου, I_{ki} και $I_{\sim ki}$ είναι δυαδικές 0-1 μεταβλητές οι οποίες αναπαριστούν τη σωστή ή την εσφαλμένη ταξινόμηση κάθε αντικειμένου που ανήκει στις κατηγορίες A_k και $\sim A_k$ αντίστοιχα (το 0 υποδεικνύει τη σωστή ταξινόμηση, ενώ το 1 υποδεικνύει εσφαλμένη ταξινόμηση). Ως N_k συμβολίζεται ο αριθμός των αντικειμένων που ανήκουν στην κατηγορία A_k, και αντίστοιχα ως $N_{\sim k}$ συμβολίζεται ο αριθμός των αντικειμένων που ανήκουν στο σύνολο των κατηγοριών $\sim A_k$. Η βασική δυσκολία που εντοπίζεται στη μεγιστοποίηση της παραπάνω συνάρτησης έγκειται στο γεγονός ότι αυτή μπορεί να επιτευχθεί μόνο μέσω διαμόρφωσης ενός προβλήματος μικτού – ακεραίου μαθηματικού προγραμματισμού. Η επίλυση ενός τέτοιου προβλήματος απαιτεί μια χρονοβόρα διαδικασία, ιδιαίτερα σε περιπτώσεις όπου το δείγμα των υπό εξέταση αντικειμένων είναι μεγάλο.

Στη μέθοδο M.H.DIS, η ανάπτυξη των προσθετικών συναρτήσεων χρησιμότητας πραγματοποιείται μέσω τεχνικών μαθηματικού προγραμματισμού. Όπως, έχει αναφερθεί και παραπάνω, σε κάθε στάδιο της ιεραρχικής διαδικασίας ταξινόμησης, επιλύονται δύο προβλήματα γραμμικού προγραμματισμού και ένα πρόβλημα ακέραιου προγραμματισμού για τον εντοπισμό της βέλτιστης ταξινόμησης. Αρχικά επιλύεται ένα πρόβλημα γραμμικού προγραμματισμού (LP1) για να ελαχιστοποιηθεί το μέγεθος των λανθασμένων ταξινομήσεων. Στη συνέχεια,

επιλύεται ένα πρόβλημα ακέραιου προγραμματισμού (MP1) για να ελαχιστοποιηθεί ο συνολικός αριθμός των λανθασμένων ταξινομήσεων μεταξύ αυτών που προκύπτουν μετά την επίλυση του προβλήματος LP1, διατηρώντας αμετάβλητο τον αριθμό των σωστών ταξινομήσεων. Τέλος, επιλύεται ένα δεύτερο πρόβλημα γραμμικού προγραμματισμού για να μεγιστοποιηθεί η διαύγεια των ταξινομήσεων που προκύπτουν μετά την επίλυση των LP1 και MP1. Λεπτομερής περιγραφή της μεθόδου μπορεί να βρεθεί στις εργασίες των Zopounidis and Doumpos (2000).

3.4 Μη παραμετρικές προσεγγίσεις

Οι στατιστικές ιδιότητες των εξεταζόμενων εναλλακτικών δραστηριοτήτων είναι συνήθως άγνωστες, καθώς πολλές φορές ο εντοπισμός του αντίστοιχου πληθυσμού είναι αδύνατος. Το γεγονός αυτό ώθησε πληθώρα ερευνητών στην ανάπτυξη μιας σειράς εναλλακτικών μη παραμετρικών προσεγγίσεων ταξινόμησης. Οι προσεγγίσεις αυτές δεν βασίζονται σε στατιστικές υποθέσεις και συνεπώς αναμένεται ότι μπορούν να προσαρμόζονται ικανοποιητικά, ανάλογα με τα χρησιμοποιούμενα σύνολα δεδομένων. Συνεπώς, τέτοιου είδους προσεγγίσεις παρέχουν αυξημένη ευελιξία στον αποφασίζοντα, χωρίς να είναι απαραίτητος ο εντοπισμός και η ανάλυση των στατιστικών ιδιοτήτων των δεδομένων που αφορούν το εξεταζόμενο πρόβλημα.

3.4.1 Δενδρική Ταξινόμηση και Παλινδρόμηση

Η δενδρική ταξινόμηση και παλινδρόμηση (Classification and Regression Trees-CART, Breiman et al., 1984, Yohannes and Webb, 1999) είναι μια μη παραμετρική προσέγγιση που αναπτύχθηκε για την ανάλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Σε κάθε περίπτωση το μοντέλο ταξινόμησης αναπαριστά τη μορφή ενός δένδρου αποφάσεων. Κύριος σκοπός της μεθόδου CART είναι να παράγει ένα ακριβές σύνολο από κανόνες ταξινόμησης βάσει των οποίων θα

προβλέπει σε ποια κατηγορία θα ανήκει κάθε μελλοντική παρατήρηση, σύμφωνα με τα αντίστοιχα χαρακτηριστικά της. Η δομή ενός κανόνα ταξινόμησης της μεθόδου CART επικεντρώνεται στους ορισμούς τριών κύριων παραγόντων:

1. του κανόνα διαχωρισμού του δείγματος παρατηρήσεων
2. των κριτηρίων αξιολόγησης της ποιότητας του διαχωρισμού
3. των κριτηρίων για την επιλογή του βέλτιστου δένδρου για ανάλυση

Τα βασικά βήματα για την δημιουργία ενός δένδρου ταξινόμησης είναι:

- Δημιουργία ενός δένδρου με μεγάλο αριθμό από κόμβους.
- Ένωση μερικών διακλαδώσεων για την παραγωγή μιας σειράς από μικρότερα δένδρα διαφορετικού μεγέθους.
- Επιλογή ενός βέλτιστου δένδρου μέσω της μέτρησης της ακρίβειας του δένδρου.

Για την ανάπτυξη ενός δένδρου ταξινόμησης, η μέθοδος CART χρησιμοποιεί μια πιθανοθεωρητική προσέγγιση η οποία μπορεί να υλοποιηθεί με τρεις τρόπους:

1. προσδιορισμός των a priori πιθανοτήτων των κατηγοριών από τα δεδομένα: $\pi_i=n_i/n$, όπου π_i η a priori πιθανότητα της κατηγορίας A_i, n ο αριθμός των αντικειμένων στο δείγμα, και n_i ο αριθμός των αντικειμένων της κατηγορίας A_i.
2. θεώρηση των a priori πιθανοτήτων των κατηγοριών ως ίσων.
3. προσδιορισμός των a priori πιθανοτήτων των κατηγοριών μέσω μιας υβριδικής προσέγγισης θεωρώντας τον μέσο όρο των δύο εκτιμήσεων που υπολογίζονται από τις προηγούμενες δύο προσεγγίσεις.

Η ανάπτυξη ενός δένδρου απαιτεί τον καθορισμό ενός συνόλου ερωτήσεων η απάντηση των οποίων οδηγεί στην ταξινόμηση των αντικειμένων, των κανόνων αξιολόγησης της ποιότητας των ερωτήσεων που αναπτύσσονται, και των κανόνων για τον προσδιορισμό της κατηγορίας σε κάθε τερματικό κόμβο το δένδρου.

Αρχικά, όλες οι παρατηρήσεις τοποθετούνται σε έναν αρχικό κόμβο, ο οποίος είναι ανομοιογενής καθώς περιέχει παρατηρήσεις από διάφορες κατηγορίες. Ο στόχος είναι η εύρεση εκείνων των κανόνων που θα διαχωρίσουν τις παρατηρήσεις δημιουργώντας νέους κόμβους σε κατώτερα επίπεδα του δένδρου, οι οποίοι θα είναι περισσότερο ομοιογενείς σε σχέση με τους προηγούμενους κόμβους.

Σε κάθε κόμβο t του δένδρου οι παρατηρήσεις του δείγματος διαχωρίζονται σε δύο επιμέρους κόμβους t_L και t_R στο αμέσως κατώτερο επίπεδο του δένδρου, ανάλογα με το εάν ικανοποιούν ή όχι ένα κανόνα της μορφής $x_{ij}<d_j$, όπου x_j είναι ένα χαρακτηριστικό και d_i είναι ένα όριο διαχωρισμού. Ειδικότερα, μια παρατήρηση i τοποθετείται στον κόμβο t_L εάν $x_{ij}<d_j$ διαφορετικά τοποθετείται στον κόμβο t_R. Ο βέλτιστος κανόνας διαχωρισμού καθορίζεται μεγιστοποιώντας τη μείωση της ανομοιογένειας που αποφέρει ο διαχωρισμός. Ένας διαχωρισμός θεωρείται ομοιογενής εάν δύο κόμβοι που δημιουργούνται από αυτόν περιλαμβάνουν παρατηρήσεις από διαφορετικές κατηγορίες. Εάν κάποιος κόμβος περιλαμβάνει παρατηρήσεις από διαφορετικές κατηγορίες, τότε ο διαχωρισμός θεωρείται ανομοιογενής. Βάσει αυτής της θεώρησης ως κριτήριο επιλογής του κατάλληλου διαχωρισμού θεωρείται η μεγιστοποίηση της ακόλουθης συνάρτησης:

$$\Delta_i(s,t) = i(t) - p_L[i(t_l)] - p_R[i(t_R)] \qquad (3.20)$$

Όπου s ο διαχωρισμός των παρατηρήσεων από τον κανόνα που αναπτύσσεται, p_L η αναλογία των περιπτώσεων του κόμβου t που καταλήγουν στον κόμβο t_L, p_R η αναλογία των περιπτώσεων του κόμβου t που καταλήγουν στον t_R, $i(t_L)$ η ομοιογένεια του κόμβου t_L, και $i(t_R)$ η ομοιογένεια του κόμβου t_R.

Αυτή η διαδικασία διαχωρισμού ξεκινά από τον αρχικό κόμβο του δένδρου και συνεχίζεται επαναληπτικά για κάθε νέο κόμβο που κατασκευάζεται. Εάν η διαδικασία εφαρμοστεί χωρίς κάποιο κριτήριο τερματισμού, τότε θα ολοκληρωθεί με την ανάπτυξη ενός μεγάλου και περίπλοκου δένδρου στο οποίο κάθε τελικός κόμβος θα περιέχει μόνο μια παρατήρηση του δείγματος εκμάθησης. Για να αποφευχθεί αυτό το φαινόμενο χρησιμοποιούνται τεχνικές μείωσης των διαστάσεων του δένδρου οι οποίοι υλοποιούνται είτε με την εισαγωγή κριτηρίων έγκαιρου τερματισμού της διαδικασίας ανάπτυξης του δένδρου, είτε με την περικοπή (pruning) του δένδρου μετά την πλήρη ανάπτυξη του.

Τα βασικά πλεονεκτήματα της μεθόδου CART είναι ότι δεν πραγματοποιείται καμία στατιστική υπόθεση όσον αφορά τα χαρακτηριστικά, είναι δυνατή η χρησιμοποίηση τόσο ποιοτικών όσο και ποσοτικών χαρακτηριστικών, τα αποτελέσματα της μεθόδου παραμένουν αμετάβλητα ανεξάρτητα από πιθανούς

μονότονους μετασχηματισμούς των δεδομένων και η κατανόηση του δένδρου ταξινόμησης είναι ιδιαίτερα εύκολη.

3.4.2 Μέθοδος Πλησιέστερου Γείτονα (K-nearest-neighbor)

Οι αλγόριθμοι πλησιέστερου γειτονα (Altman, 1968) διαφέρουν από τους υπόλοιπους καθώς δεν χρησιμοποιούν το δείγμα εκμάθησης για την ανάπτυξη μιας συνάρτησης ταξινόμησης, αλλά ως ένα σημείο αναφοράς προς το οποίο συγκρίνεται κάθε νέο αντικείμενο. Ειδικότερα για την ταξινόμηση ενός νέου αντικειμένου σε μια από τις προκαθορισμένες κατηγορίες, βρίσκονται τα αντικείμενα του δείγματος εκμάθησης που είναι πλησιέστερα στο εξεταζόμενο αντικείμενο. Βάσει των κατηγοριών στα οποία ανήκουν τα αντικείμενα αυτά και μέσω του κανόνα της πλειοψηφίας λαμβάνεται η απόφαση για την ταξινόμηση του αντικειμένου.

Έστω ότι από τα πρώτα k αντικείμενα, υπάρχουν k_m στην κατηγορία A_m (έτσι ώστε $\sum_{m=1}^{C} k_m = k$), και έστω ότι ο συνολικός αριθμός αντικειμένων στην κατηγορία A_m είναι n_m ($\sum_{m=1}^{C} n_m = n$). Τότε η δεσμευμένη πιθανότητα του x ως δεδομένο της κατηγορίας A_m είναι:

$$\hat{p}(x \mid A_m) = \frac{k_m}{n_m V} \qquad (3.21)$$

όπου V η συχνότητα εμφάνισης των αντικειμένων στην κατηγορία A_m.

Η αρχική πιθανότητα της κατηγορίας Α ορίζεται ως:

$$\hat{p}(A_m) = \frac{n_m}{n} \qquad (3.22)$$

Οπότε ο κανόνας απόφασης που ταξινομεί το x στην κατηγορία A_m είναι ο ακόλουθος:

$$\hat{p}(A_m \mid x) \geq \hat{p}(A_i \mid x) \quad \text{για όλα τα i} \qquad (3.23)$$

Κατά συνέπεια, ο κανόνας απόφασης τοποθετεί το x στην κατηγορία με τη μεγαλύτερη πιθανότητα μεταξύ των k πλησιέστερων γειτόνων. Εναλλακτικά το x μπορεί να ταξινομηθεί στην κατηγορία που έχει το πλησιέστερο μέσο διάνυσμα με αυτό του x (το μέσο διάνυσμα εκτιμάται συνολικά για τα k αντικείμενα). Επίσης, το x μπορεί να ταξινομηθεί στην περισσότερο ομοιογενή κατηγορία, δηλαδή σε αυτή για την οποία η απόσταση του k μέλους είναι η μικρότερη. Αυτό δεν απαιτεί επιπρόσθετους υπολογισμούς. Ο Dudani (1976) προτείνει έναν κανόνα απόστασης – βάρους (distance-weighted rule), σύμφωνα με τον οποίο, βάρη ορίζονται στους k γείτονες, δίνοντας μεγαλύτερο βάρος στον περισσότερο πλησιέστερο γείτονα. Ένα πρότυπο ορίζεται στην κατηγορία για την οποία τα βάρη των αντικειμένων μεταξύ των k γειτόνων αθροίζουν στη μεγαλύτερη τιμή.

Ο δείκτης εσφαλμένης ταξινόμησης στον κανόνα του πλησιέστερου γείτονα, ικανοποιεί την ακόλουθη συνθήκη (Cover and Hart, 1967):

$$e^* \leq e \leq e^*\left(2 - \frac{Ce^*}{C-1}\right) \qquad (3.24)$$

Όπου e^* είναι η πιθανότητα του Bayes και C ο αριθμός των κατηγοριών. Κατά συνέπεια, σε μεγάλα δείγματα ο δείκτης σφάλματος είναι οριακά μικρότερος από το διπλάσιο δείκτη σφάλματος του Bayes. Η ανισότητα αυτή μπορεί να αναστραφεί και να δώσει το εξής:

$$\frac{C-1}{C} - \sqrt{\frac{C-1}{C}}\sqrt{\frac{C-1}{C}} - e \leq e^* \leq e \qquad (3.25)$$

Η αριστερή ποσότητα, στην παραπάνω εξίσωση είναι μικρότερη από το όριο του δείκτη σφάλματος του Bayes. Επομένως, οποιαδήποτε ταξινόμηση πρέπει να έχει ένα δείκτη σφάλματος μεγαλύτερο από αυτό το όριο.

Για την επιτάχυνση της διαδικασίας εύρεσης του πλησιέστερου γείτονα, έχουν προταθεί πολλές προσεγγίσεις. Οι Fukunaka & Narendra (1975) χρησιμοποιούν έναν αλγόριθμο, σύμφωνα με τον οποίο ο χώρος των χαρακτηριστικών διαιρείται σε

διάφορες περιοχές. Οι περιοχές αυτές εξετάζονται μόνο αν υπάρχει πιθανότητα εύρεσης ενός πλησιέστερου γείτονα. Οι περιοχές αναλύονται σε υποσύνολα μέσω μιας ιεραρχικής διαδικασίας. Άλλοι τρόποι επιτάχυνσης της διαδικασίας είναι η χρήση ενός συγκεντρωτικού κανόνα πλησιέστερου γείτονα (condensed-nearest-neighbor, Hart, 1968), ενός μειωμένου κανόνα (reduced-nearest-neighbor, Gates, 1972) ή ενός επιμεληνένου κανόνα (edited-nearest-neighbor, Hand & Batchelor, 1978). Όλες οι παραπάνω μέθοδοι μειώνουν το δείγμα εκμάθησης κρατώντας εκείνες τις παρατηρήσεις που ταξινομούν σωστά όλα τα αντικείμενα.

3.4.3 Πιθανολογικά Νευρωνικά Δίκτυα

Τα πιθανολογικά νευρωνικά δίκτυα (Probabilistic neural networks PNN), αναπτύχθηκαν ως τεχνικές αξιολόγησης για προβλήματα ταξινόμησης (Parzen window method, Duda, 2001). Έχουν παρόμοια οργανωτική δομή με τα νευρωνικά δίκτυα και η μεθοδολογία ταξινόμησης συνδυάζει την υπολογιστική δύναμη και ευελιξία των τεχνιτών νευρωνικών δικτύων. Παράλληλα είναι απλά στην χρήση και σαφείς.

Ο σκοπός της μεθόδου είναι να καθοριστεί η κατηγορία A ($A=A_1, A_2, \ldots, A_n$) στην οποία θα ενταχθούν τα αντικείμενα ενός δείγματος εκμάθησης. Εάν οι συναρτήσεις πυκνότητας πιθανότητας $g_A(x)$ της κάθε κατηγορίας A είναι γνωστές, τότε σύμφωνα με το βέλτιστο κανόνα του Bayes, το αντικείμενο x ταξινομείται στην A_i εφόσον ισχύσει το ακόλουθο:

$$h_i c_i g_i(x) > h_j c_j g_j(x) \qquad (3.26)$$

για κάθε $i \neq j$. Ως h_i είναι η αρχική πιθανότητα να ανήκει το αντικείμενο στην κατηγορία A_i και c_i είναι το κόστος εσφαλμένης ταξινόμησης ενός αντικειμένου της κατηγορίας A_i.

Όταν οι συναρτήσεις πυκνότητας πιθανότητας (PDF) δεν είναι γνωστές για τις διάφορες κατηγορίες, τότε η πλειοψηφία των κλασικών στατιστικών τεχνικών ταξινόμησης προβαίνουν σε υποθέσεις για τη φύση των PDF. Ο Parzen (1962)

ανέπτυξε μια μη παραμετρική τεχνική για την εκτίμηση των PDF των διαφόρων κατηγοριών, για μονομεταβλητές κατανομές. Το παραπάνω όμως, επεκτάθηκε και για την περίπτωση των πολυμεταβλητών κατανομών, από τον Cacoullos (1966), ο οποίος απέδειξε ότι η εκτιμώμενη συνάρτηση πυκνότητας προσεγγίζει ασυμπτωτικά την πραγματική συνάρτηση πυκνότητας, καθώς το μέγεθος του δείγματος εκμάθησης αυξάνει. Εφόσον εκτιμηθούν οι PDF, εφαρμόζεται ο κανόνας του Bayes για την ταξινόμηση των αντικειμένων. Η παραπάνω διαδικασία είναι δύο βημάτων: εκτιμώνται τα $g_A(x)$ όλων των κατηγοριών και στη συνέχεια εφαρμόζεται ο κανόνας του Bayes.

Έστω $Y_{Ak}=y_1, y_2, \ldots,y_n$ (με k=1,…,n) ένα πολυμεταβλητό διάνυσμα ενός γνωστού συνόλου αντικειμένων της κατηγορίας A. Η εκτιμώμενη πυκνότητα πιθανότητα του x για μια δεδομένη κατηγορία A, $\hat{g}_A(x)$, καθορίζεται από τις συναρτήσεις βαρών W(x,y) ως προς τα $g_A(x)$:

$$\hat{g}_A(x) = \frac{1}{n}\sum_{j=1}^{n} W(x, y_j) \qquad (3.27)$$

Η μορφή του W(x,y) είναι τέτοια ώστε εκφράζει την αυξανόμενη επίδραση του y στο x καθώς η απόσταση d(x,y), μεταξύ του x και y μειώνεται. Η τελική πιθανότητα του x να ανήκει στην κατηγορία A, P[A|x], μπορεί να εκτιμηθεί μέσω του θεωρήματος του Bayes (Specht, 1990) ως εξής:

$$P[A \mid x] = \frac{h_A g_A(x)}{h_1 g_1(x) + h_2 g_2(x) + \ldots + h_n g_n(x)} \qquad (3.28)$$

το x ταξινομείται στην κατηγορία A_i εφόσον:

$$P[A_i|x] > P[A_j|x] \text{ για όλα τα } i \neq j \qquad (3.29)$$

Η επιλογή των βαρών W(x,y) είναι κρίσιμη για τον καθορισμό των $\hat{g}(x)$. Συνήθως τα W(x,y) επιλέγονται με τους ακόλουθους δύο τρόπους:

$$W(x,y) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

ή πιο απλά (3.30)

$$W(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

όπου σ είναι μια παράμετρος εξομάλυνσης που καθορίζει το εύρος των συναρτήσεων βαρών.

Τα PNN είναι δίκτυα παράλληλων μονάδων επεξεργασίας οι οποίες είναι οργανωμένες σε τρία επίπεδα. Το επίπεδο εισόδου (input layer) αποτελείται από μια σειρά n κόμβων, ένα για κάθε χαρακτηριστικό των εναλλακτικών δραστηριοτήτων. Οι είσοδοι του δικτύου συνδέονται πλήρως με τους m κόμβους του ενδιάμεσου επιπέδου (pattern layer), όπου m είναι ο αριθμός των εναλλακτικών δραστηριοτήτων του δείγματος εκμάθησης. Κάθε κόμβος k (k=1,2, …,m) του ενδιάμεσου επιπέδου συνδέεται με ένα διάνυσμα βάρους $w_k = (x_{k1}, x_{k2}, \dots, x_{kn})$. Η είσοδος x_i, ο ενδιάμεσος κόμβος k και το διάνυσμα βάρους w_k ορίζουν μια συνάρτηση η οποία δίνει το αποτέλεσμα του ενδιάμεσου κόμβου k.

Τα αποτελέσματα του ενδιάμεσου κόμβου μεταφέρονται στο τρίτο επίπεδο. Το επίπεδο αυτό αποτελείται από q κόμβους, όπου ο κάθε ένας αντιστοιχεί στις q προκαθορισμένες κατηγορίες $A_1, A_2, \dots, A_q$. Κάθε κόμβος του ενδιάμεσου επιπέδου συνδέεται μόνο με τον κόμβο του τρίτου επιπέδου (τον κόμβο ο οποίος αντιστοιχεί στην κατηγορία στην οποία ανήκει η εναλλακτική δραστηριότητα). Οι κόμβοι του τρίτου επιπέδου απλά αθροίζουν τα αποτελέσματα των κόμβων του ενδιάμεσου επιπέδου με τους οποίους συνδέονται. Κατά ασυνέπεια, το άθροισμα αυτό παρέχει q σκορ για κάθε χαρακτηριστικό x_i των εναλλακτικών δραστηριοτήτων, με αποτέλεσμα η εναλλακτική δραστηριότητα να ταξινομείται στην κατηγορία η οποία έχει παρόμοια χαρακτηριστικά με αυτή.

1.4.4 Μηχανές Διανύσματος Υποστήριξης

Οι μηχανές διανύσματος υποστήριξης (Support Vector Machines, SVM) αναπτύχθηκαν στα τέλη της δεκαετίας του ’70 (Vapnik, 1979), αλλά ιδιαίτερη ανάπτυξη γνωρίζουν τα τελευταία χρόνια όπου και θεωρούνται ως μια από τις σημαντικότερες μεθόδους για την ανάπτυξη μοντέλων ταξινόμησης. Κύριο χαρακτηριστικό τους αποτελεί το σημαντικό υπόβαθρο πάνω στο οποίο βασίζονται τα SVM, καθώς και η πληθώρα επιτυχημένων πρακτικών εφαρμογών.

Τα SVM διακρίνονται σε γραμμικά, που αφορούν διαχωρίσιμα δεδομένα (separable data) και σε μη γραμμικά, που αφορούν μη διαχωρίσιμα δεδομένα. Στόχος των SVM, στην πρώτη περίπτωση, είναι η ανάπτυξη ενός βέλτιστου υπερεπιπέδου της μορφής Zw-y για την ταξινόμηση των παρατηρήσεων, όπου ως Z συμβολίζεται ένας πίνακας nxm με τα στοιχεία του δείγματος εκμάθησης. Έστω ότι τα δεδομένα του δείγματος εκμάθησης είναι τα ακόλουθα: $\{x_i, y_i\}$, $i = 1,2,...,l$, $y_i \in \{-1,1\}$, $x_i \in R^d$. Συμβολίζοντας ως D ένα διαγώνιο πίνακα διαστάσεων nxn με την κύρια διαγώνιο να έχει τις τιμές +1 ή –1 (δύο κατηγορίες οι οποίες συμβολίζονται +1 και –1, στις οποίες ταξινομούνται τα δεδομένα) ανάλογα με την ταξινόμηση των δεδομένων του δείγματος εκμάθησης, και ως e το μοναδιαίο διάνυσμα nx1, ο εντοπισμός του βέλτιστου υπερεπιπέδου επιτυγχάνεται με την επίλυση του ακόλουθου τετραγωνικού προγράμματος (ως ν συμβολίζεται μια αυστηρά θετική σταθερά):

$$\begin{aligned} &\underset{w,y,d}{Min}\ ve'd + \frac{1}{2}w'w \\ &\upsilon\pi\acute{o}: \\ &D(Zw - ey) + d \geq e \\ &d \geq 0 \end{aligned} \qquad (3.31)$$

Ο τετραγωνικός όρος w΄w στην αντικειμενική συνάρτηση του προβλήματος (3.31) μεγιστοποιεί το περιθώριο μεταξύ των δύο υπερεπιπέδων Zw-y= +1 και Zw-y= -1, το οποίο ισούται με $2/\|w\|$. Εκτός της μεγιστοποίησης του περιθωρίου των κατηγοριών, το πρόβλημα (3.31) λαμβάνει υπόψη και το σφάλμα ταξινόμησης με τις μεταβλητές d (η σταθερά ν>0 αναπαριστά τη σχετική βαρύτητα που αποδίδεται στην

ελαχιστοποίηση των σφαλμάτων). Όταν όλες οι μεταβλητές d είναι ίσες με το μηδέν, τότε οι δύο κατηγορίες είναι αυστηρά γραμμικά διαχωρίσιμες. Τότε τα δύο επίπεδα καθορίζουν τα όρια των δύο κατηγοριών με ένα μη αρνητικό σφάλμα της μεταβλητής d:

$$\begin{aligned} Z_i'w + d_i &\geq y + 1, \quad \gamma\iota\alpha \quad D_{ii} = +1 \\ Z_i'w - d_i &\leq y - 1 \quad \gamma\iota\alpha \quad D_{ii} = -1 \end{aligned} \tag{3.32}$$

Με την επίλυση του προβλήματος (3.31) και τον προσδιορισμό των w και y που καθορίζουν το βέλτιστο υπερεπίπεδο, η ταξινόμηση κάθε αντικειμένου μπορεί να πραγματοποιηθεί ως εξής:

$$\text{Εάν} \quad Z_i'w - y \begin{cases} > 0, \quad \tau\acute{o}\tau\varepsilon \quad Z_i \in \{+1\} \\ < 0, \quad \tau\acute{o}\tau\varepsilon \quad Z_i \in \{-1\} \\ = 0, \quad \tau\acute{o}\tau\varepsilon \quad Z_i \in \{+1\} \acute{\eta} \; x \in \{-1\} \end{cases} \tag{3.33}$$

Το κύριο μειονέκτημα του προβλήματος βελτιστοποίησης (3.31) για τον προσδιορισμό του βέλτιστου μοντέλου ταξινόμησης αφορά τον αυξημένο υπολογιστικό φόρτο που απαιτεί η επίλυση του καθώς πρόκειται για ένα πρόβλημα τετραγωνικού προγραμματισμού. Για την αντιμετώπιση του παραπάνω προβλήματος, διαμορφώνεται η ακόλουθη συνάρτηση:

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} a_i y_i (Z_i w + b) + \sum_{i=1}^{l} a_i \tag{3.34}$$

όπου α_i οι θετικοί πολλαπλασιαστές του Lagrange.

Με βάσει τη συνάρτηση αυτή το δυϊκό πρόβλημα έχει την ακόλουθη μορφή:

$$L_D \equiv \sum_i a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j x_i x_j \tag{3.35}$$

υπό τους περιορισμούς:

$$0 \le a_i \le v$$
$$\sum_i a_i y_i = 0 \qquad (3.36)$$

Η λύση του οποίου είναι η εξής:

$$w=\sum^{Ns}_{i=1} \alpha_i y_i x_i \qquad (3.37)$$

όπου N_s ο αριθμός των διανυσμάτων υποστήριξης.

Στη δεύτερη περίπτωση των μη γραμμικών SVM (Boser et al., 1992), και γενικεύοντας τη μέθοδο, θεωρείται ότι τα αντικείμενα ορίζονται σε ένα ευκλείδειο υπερεπίπεδο **H** (Φ: $R^d \mapsto$ **H**). Επιπλέον, χρησιμοποιείται μια συνάρτηση πυρήνα της μορφής:

$$K(x_i, x_j)=\Phi(x_i)\,\Phi(x_j) \qquad (3.38)$$

Η εκτίμηση των μη γραμμικών SVM πραγματοποιείται με την ακόλουθη εξίσωση:

$$f(x) = \sum_{i=1}^{N_s} a_i y_i \Phi(s_i)\Phi(x) + b = \sum_{i=1}^{N_s} a_i y_i K(s_i, x) + b \qquad (3.39)$$

όπου s_i είναι τα διανύσματα υποστήριξης.

Ένα μειονέκτημα των SVM είναι ο μεγάλος χρόνος επίλυσης που απαιτείται όταν τα μεγέθη των δειγμάτων εκμάθησης και ελέγχου είναι πολύ μεγάλα. Δίνουν ικανοποιητικά αποτελέσματα, αλλά περιορίζεται η ταχύτητα επίλυσης του προβλήματος.

Τα SVM είναι μια νέα προσέγγιση επίλυσης προβλημάτων αναγνώρισης προτύπων, τα οποία έχουν άμεση σχέση με τη στατιστική θεωρία. Διαφέρουν από άλλες προσεγγίσεις, όπως τα νευρωνικά δίκτυα, γιατί τα SVM αναζητούν ένα ολικό ελάχιστο στο δείγμα εκμάθησης, και οι λύσεις ερμηνεύονται γεωμετρικά. Επιπλέον, στη γενικότερη περίπτωση, τα SVM εξαρτώνται κατά πολύ από την επιλογή του πυρήνα, έτσι ώστε η λύση να είναι αποδεκτή.

ΚΕΦΑΛΑΙΟ 4

Ανάλυση δεδομένων και εφαρμογή

4.1 Υποθέσεις στην εφαρμογή των μοντέλων

Η συγκεκριμένη διατριβή εστιάζεται στην εξέταση των χαρακτηριστικών των δεδομένων τα οποία μπορεί να επηρεάζουν την αποτελεσματικότητα των διαφορετικών επαγωγικών μεθόδων. Τα χαρακτηριστικά των δεδομένων επιλέχθηκαν έτσι ώστε να προσδιορίζουν την σταθερότητα και τις αδυναμίες της κάθε μεθόδου. Για την εξακρίβωση του παραπάνω, τα χαρακτηριστικά τα οποία επιλέχθηκαν είναι οι διάφορες κατανομές που μπορεί να ακολουθούν τα δεδομένα (κανονική, λογαριθμική – κανονική, μίξη κανονικής – λογαριθμικής), η μορφή διάκρισης των κατηγοριών (γραμμική, μη γραμμική), το πλήθος των αντικειμένων στο δείγμα εκμάθησης, και ο βαθμός συσχέτισης μεταξύ των κριτηρίων (ασυσχέτιστα, υψηλή συσχέτιση). Επιπρόσθετα, πραγματοποιήθηκε ακόμη μία διάκριση μεταξύ των δεδομένων που χρησιμοποιήθηκαν. Θεωρήθηκε σκόπιμο να εξεταστεί η αποτελεσματικότητα των μεθόδων όταν τα δεδομένα είναι συνεχή και όταν είναι διακριτά. Ο σκοπός του παραπάνω είναι να ελεγχθεί το μέγεθος επηρεασμού των χαρακτηριστικών των δεδομένων στην αποτελεσματικότητα των μεθόδων, όταν υπάρχουν δύο διαφορετικές περιπτώσεις δεδομένων. Δηλαδή, αυτό που προσδοκάται είναι να διερευνηθεί η αποτελεσματικότητα των μεθόδων συναρτήσει των χαρακτηριστικών των εξεταζόμενων δεδομένων. Συγκεκριμένα, αυτό που θα εξεταστεί είναι τα διάφορα ποσοστά εσφαλμένης ταξινόμησης των μεθόδων ως προς τον τύπο των δεδομένων (συνεχή, διακριτά) και ως προς τα διάφορα χαρακτηριστικά – παράγοντες. Στη συνέχεια παρουσιάζεται μια περισσότερο λεπτομερή περιγραφή των χαρακτηριστικών των δεδομένων για τις δύο διαφορετικές κατηγορίες δεδομένων που χρησιμοποιήθηκαν.

4.1.1 Συνεχή δεδομένα

Αρχικά, εξετάζονται τα συνεχή δεδομένα του δείγματος εκμάθησης για τους διάφορους παράγοντες. Επιπλέον, έχουν επιλεγεί πέντε κριτήρια για τα συνεχή δεδομένα. Ο πρώτος παράγοντας που λαμβάνεται υπόψη είναι η στατιστική κατανομή που ακολουθούν τα δεδομένα. Στη συγκεκριμένη περίπτωση θεωρούνται τρεις κατανομές, οι οποίες είναι οι ακόλουθες:

1. Κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1.
2. Λογαριθμική – Κανονική κατανομή με παραμέτρους θ=0.5 και σ=0.5.
3. Μίξη Κανονικής – Λογαριθμικής (θεωρούνται δύο κριτήρια για την κανονική κατανομή και τρία κριτήρια για την λογαριθμική).

Ο δεύτερος παράγοντας που λαμβάνεται υπόψη είναι η μορφή διάκρισης των κατηγοριών. Θεωρούνται δύο μορφές διάκρισης των κατηγοριών, οι οποίες είναι οι παρακάτω:

- Γραμμική διάκριση: η εξίσωση της γραμμικής διάκρισης είναι η εξής:

$$f_i = \sum_{j=1}^{5} a_j x_{ij}$$

όπου f_i είναι η βαθμολογία του αντικειμένου i, x_{ij} η επίδοση του αντικειμένου i στο κριτήριο j, α_j τυχαίοι διακριτικοί συντελεστές ομοιόμορφα κατανεμημένοι στο διάστημα [1, 10].

- Τετραγωνική διάκριση: η εξίσωση της τετραγωνικής διάκρισης είναι η εξής:

$$f_i = \sum_{j=1}^{5} a_j x_{ij} + \sum_{k=1}^{5} \sum_{j=1}^{5} a_{kj} x_{ik} x_{ij}$$

όπου α_{kj} είναι τυχαίοι συντελεστές, οι οποίοι κυμαίνονται στο διάστημα [0, 2] όταν k=j, ενώ για $k \neq j$ κυμαίνονται στο διάστημα [-2, 2].

Σε κάθε μία από τις παραπάνω περιπτώσεις ισχύει ο εξής κανόνας ταξινόμησης: εάν $f_i>0$ τότε το αντικείμενο i κατατάσσεται στην κατηγορία 1, διαφορετικά κατατάσσεται στην κατηγορία 2.

Στην παραπάνω διαδικασία έχει προστεθεί και ένας θόρυβος για να παραβιάζει τον κανόνα ταξινόμησης και να μην είναι τέλειος. Συγκεκριμένα, όταν

είναι διαθέσιμα όλα τα αποτελέσματα της ταξινόμησης, αυτά μεταβάλλονται τυχαία στο 10% των περιπτώσεων έτσι ώστε να παραβιαστεί ο κανόνας ταξινόμησης.

Ο τρίτος παράγοντας που λαμβάνεται υπόψη είναι το πλήθος των αντικειμένων του δείγματος εκμάθησης. Προηγούμενες έρευνες έχουν δείξει ότι το μέγεθος του δείγματος επηρεάζει την απόδοση των μεθόδων και την αξιοπιστία των εκτιμήσεων. Σε μερικές μεθόδους απαιτείται μεγάλο μέγεθος δείγματος έτσι ώστε να επιτευχθεί μέγιστη ακρίβεια στα αποτελέσματα, ενώ άλλες μέθοδοι απαιτούν σχετικά μικρό δείγμα εκμάθησης. Παρεμφερή με το πρόβλημα ενός μεροληπτικού δείγματος είναι και ο περιορισμός που μπορεί να υπάρξει στην διαθεσιμότητα και στις πηγές των δεδομένων. Στη συγκεκριμένη περίπτωση επιλέγεται το ακόλουθο πλήθος αντικειμένων:

1. 200
2. 500
3. 1000

Για την ακρίβεια, για το δείγμα εκμάθησης επιλέγονται τυχαία 200, 500 και 1000 αντικείμενα, ενώ για το δείγμα ελέγχου επιλέγονται 500 αντικείμενα, για κάθε συνδυασμό των παραγόντων.

Ο τέταρτος παράγοντας που λαμβάνεται υπόψη είναι ο βαθμός συσχέτισης μεταξύ των κριτηρίων. Θεωρούνται οι εξής δύο περιπτώσεις:

- ✓ Τα κριτήρια είναι ασυσχέτιστα μεταξύ τους.
- ✓ Ύπαρξη υψηλής συσχέτισης μεταξύ των κριτηρίων με:

$$x_i = \frac{1}{i}\sum_{k=1}^{i-1} x_k + \gamma \quad \text{με} \quad \gamma \sim N(0, 1)$$

4.1.2 Διακριτά δεδομένα

Όσον αφορά τα διακριτά δεδομένα, δεν υπάρχουν διαφορές στη χρήση των παραγόντων, με αυτούς που λαμβάνονται υπόψη στα συνεχή δεδομένα, με εξαίρεση τον πρώτο παράγοντα. Ουσιαστικά, και στην περίπτωση των διακριτών δεδομένων, η διαδικασία που ακολουθείται για την παραγωγή τους είναι όμοια με αυτή που πραγματοποιείται και στα συνεχή δεδομένα, καθώς και τα κριτήρια που χρησιμοποιούνται είναι επίσης πέντε.

Συγκεκριμένα, ο πρώτος παράγοντας αφορά το πλήθος των διακριτών επιπέδων που μπορεί να έχουν τα δεδομένα, σε αντίθεση με την στατιστική κατανομή

που ακολουθούν τα συνεχή δεδομένα, στην πρώτη περίπτωση. Το πλήθος των διακριτών επιπέδων που λαμβάνεται υπόψη και εξετάζεται για τα διακριτά δεδομένα, είναι τα ακόλουθα:

1. Διακριτές τιμές –1 και 1.
2. Διακριτές τιμές –1, 0, 1.
3. Μίξη των δύο και τριών επιπέδων.

Ο δεύτερος παράγοντας, που αφορά τη μορφή διάκρισης των κατηγοριών στις οποίες ταξινομούνται τα αντικείμενα του δείγματος εκμάθησης, παραμένει ο ίδιος, όπως και στην περίπτωση των συνεχών δεδομένων. Δηλαδή, υπάρχουν δύο μορφές διάκρισης των κατηγοριών, η γραμμική και η τετραγωνική, και ο θόρυβος εξακολουθεί να είναι ο ίδιος. Έχοντας τα αποτελέσματα της ταξινόμησης, αυτά μεταβάλλονται τυχαία στο 10% των περιπτώσεων ώστε ο κανόνας ταξινόμησης να είναι ατελής.

Όσον αφορά τον τρίτο παράγοντα, το πλήθος των αντικειμένων για το δείγμα εκμάθησης είναι 200, 500 και 1000 αντικείμενα, ενώ για το δείγμα ελέγχου είναι 500 αντικείμενα. Η μόνη διαφορά είναι ότι στην περίπτωση αυτή τα αντικείμενα είναι διακριτά.

Τέλος, ο παράγοντας 4 που αφορά το βαθμό συσχέτισης των κριτηρίων, παραμένει ο ίδιος και στην περίπτωση των διακριτών δεδομένων. Συγκεκριμένα, και στα διακριτά δεδομένα εξετάζονται οι περιπτώσεις των ασυσχέτιστων κριτηρίων και των κριτηρίων με υψηλή συσχέτιση.

4.2 Διαδικασία παραγωγής δεδομένων

Η διαδικασία παραγωγής των δεδομένων και η εφαρμογή των διάφορων επαγωγικών μεθόδων πραγματοποιείται με τη χρήση του λογισμικού MATLAB. Ο αριθμός των δεδομένων που παράγεται είναι ο συνολικός αριθμός ως προς τον τύπο δεδομένων (συνεχή, διακριτά), ως προς τους τέσσερις παράγοντες και ως προς τις μεθόδους. Η όλη διαδικασία πραγματοποιείται μέσω μιας προσομοίωσης Monte-Carlo. Για κάθε συνδυασμό παραγόντων συντελείται η παραγωγή των αντικειμένων, τα οποία στη συνέχεια ταξινομούνται σε δύο κατηγορίες, ανάλογα με τον τύπο ταξινόμησης (γραμμικό, τετραγωνικό). Η διαμόρφωση των δειγμάτων εκμάθησης και

ελέγχου γίνεται με τυχαία επιλογή των αντικειμένων. Συγκεκριμένα, για τη διαμόρφωση του δείγματος εκμάθησης επιλέγονται τυχαία 200, 500 και 1000 αντικείμενα, ενώ για το δείγμα ελέγχου επιλέγονται 500 αντικείμενα. Κάθε πείραμα επαναλαμβάνεται 30 φορές και τα αποτελέσματα που θα παρουσιαστούν παρακάτω είναι ο μέσος όρος των 30 επαναλήψεων. Η στατιστική ανάλυση των αποτελεσμάτων έγινε στο SPSS. Οι μέθοδοι ταξινόμησης που χρησιμοποιήθηκαν είναι συνολικά εφτά και είναι η Γραμμική Διακριτική Ανάλυση (LDA), η Τετραγωνική Διακριτική Ανάλυση (QDA), το Λογιστικό Υπόδειγμα Πιθανότητας (LOGIT), ο Αλγόριθμος Πλησιέστερου Γείτονα (1NN), τα Πιθανολογικά Νευρωνικά Δίκτυα (PNN), οι Μηχανές Διανύσματος Υποστήριξης (SVM), και η μέθοδος UTADIS. Στο επόμενο κεφάλαιο ακολουθεί η παρουσίαση των αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 5
Ανάλυση αποτελεσμάτων

5.1 Σύνοψη αποτελεσμάτων

Η ανάλυση των αποτελεσμάτων αφορά το ποσοστό των εσφαλμένων ταξινομήσεων του δείγματος ελέγχου. Η επεξεργασία των αποτελεσμάτων βασίζεται στην ανάλυση διασποράς ANOVA σε συνδυασμό με τον στατιστικό έλεγχο του Tukey ο οποίος επιτρέπει τη διαμόρφωση ομοιογενών ομάδων όπου κάθε μία περιλαμβάνει τα επίπεδα ενός παράγοντα για τα οποία δεν παρουσιάζουν στατιστικά σημαντικές διαφορές μεταξύ τους, ως προς το ποσοστό των εσφαλμένων ταξινομήσεων. Στους πίνακες που ακολουθούν, στις αντίστοιχες παρενθέσεις παρουσιάζεται η ομαδοποίηση των διαφόρων μορφών διακύμανσης σύμφωνα με το στατιστικό έλεγχο Tukey, σε επίπεδο σημαντικότητας 5%.

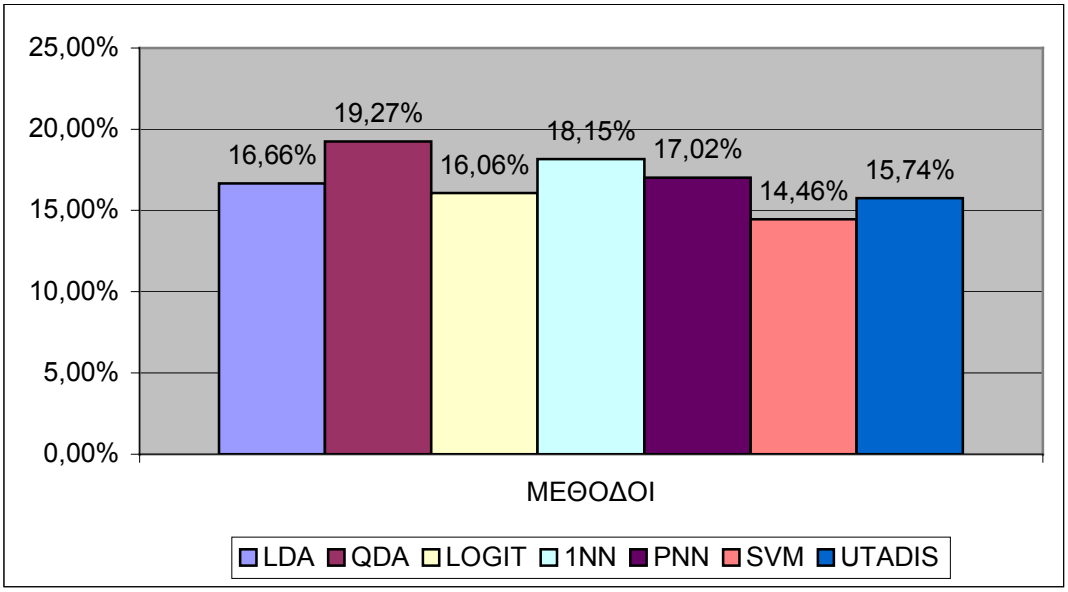
Το γενικό συμπέρασμα που προκύπτει εξετάζοντας μόνο τις μεθόδους είναι ότι το μικρότερο μέσο σφάλμα παρουσιάζουν οι μηχανές διανύσματος υποστήριξης και στη συνέχεια ακολουθούν η UTADIS, LOGIT και τα πιθανολογικά νευρωνικά δίκτυα. Το χειρότερο μέσο σφάλμα παρουσιάζει ο αλγόριθμος πλησιέστερου γείτονα με 22.40%. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα και αφορούν το μέσο σφάλμα των μεθόδων για το σύνολο των δεδομένων που εξετάζονται:

Πίνακας 5.1: Αποτελέσματα των μεθόδων.

ΜΈΘΟΔΟΙ	ΜΕΣΟ ΣΦΑΛΜΑ
LDA	15.20% (3)
QDA	16.53% (4)
LOGIT	14.63% (2)
1NN	22.40% (5)
PNN	14.67% (2)
SVM	13.70% (1)
UTADIS	14.25% (2)

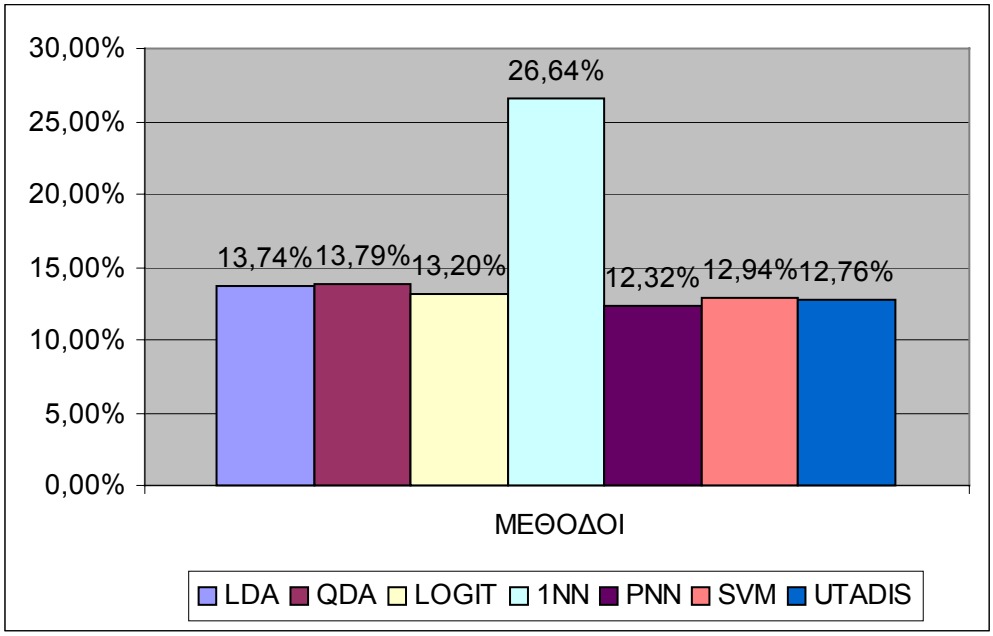
Εξετάζοντας τα ποσοστά σφαλμάτων ταξινόμησης για κάθε τύπο δεδομένων, στο σχήμα που ακολουθεί, προκύπτουν παρόμοια συμπεράσματα. Όσον αφορά την

εξέταση των μεθόδων ως προς τα συνεχή δεδομένα, το μεγαλύτερο ποσοστό εσφαλμένων ταξινομήσεων παρουσιάζει η τετραγωνική διακριτική ανάλυση, με σφάλμα 19,27% και το μικρότερο σφάλμα οι μηχανές διανύσματος υποστήριξης, με σφάλμα 14,46%. Τα παραπάνω αποτελέσματα παρουσιάζονται στο ακόλουθο σχήμα:



Σχήμα 5.1: Σύνοψη αποτελεσμάτων για συνεχή δεδομένα.

Εξετάζοντας τις μεθόδους ως προς τα διακριτά δεδομένα, παρατηρείται ότι τα ποσοστά σφαλμάτων όλων των μεθόδων κυμαίνονται από 12,32% (PNN) μέχρι 13,79% (QDA). Ωστόσο, το σφάλμα ταξινόμησης του αλγορίθμου πλησιέστερου γείτονα είναι πολύ μεγαλύτερο σε σχέση με τα υπόλοιπα και κυμαίνεται στο 26,64%. Στη συγκεκριμένη περίπτωση το μικρότερο σφάλμα, όπως αναφέρεται και παραπάνω, παρουσιάζουν τα πιθανολογικά νευρωνικά δίκτυα, με σφάλμα 12,32%. Τα αποτελέσματα των σφαλμάτων ταξινόμησης για τα διακριτά δεδομένα παρουσιάζονται στο σχήμα 5.2, το οποίο είναι το ακόλουθο:



Σχήμα 5.2: Σύνοψη αποτελεσμάτων για διακριτά δεδομένα.

Τα αποτελέσματα που αναλύθηκαν παραπάνω και παρουσιάζονται στα σχήματα 5.1 και 5.2 συνοψίζονται στον ακόλουθο πίνακα, όπου φαίνεται το μέσο σφάλμα των μεθόδων:

Πίνακας 5.2: Μέσο σφάλμα μεθόδων ανά τύπο δεδομένων.

Μέθοδοι	Συνεχή	Διακριτά
LDA	16.66% (3)	13.74% (3)
QDA	19.27% (5)	13.79% (3)
LOGIT	16.06% (2)	13.20% (2 ή 3)
1NN	18.15% (4)	26.64% (4)
PNN	17.02% (3)	12.32% (1)
SVM	14.46% (1)	12.94% (2)
UTADIS	15.74% (2)	12.76% (1 ή 2)

5.2 Αποτελέσματα ανά βαθμό συσχέτισης

Εξετάζοντας τις μεθόδους ως προς το βαθμό συσχέτισης των χαρακτηριστικών των δεδομένων, παρατηρείται ότι η τετραγωνική διακριτική ανάλυση παρουσιάζει ευστάθεια σε συσχετισμένα χαρακτηριστικά. Συγκεκριμένα, το

μέσο σφάλμα της μεθόδου για χαρακτηριστικά χαμηλής συσχέτισης είναι 16.54%, ενώ για υψηλής συσχέτισης 16.52%. Σε αντίθεση με την τετραγωνική διακριτική ανάλυση, τα πιθανολογικά νευρωνικά δίκτυα και ο αλγόριθμός πλησιέστερου γείτονα, παρουσιάζουν μεγάλη ευαισθησία σε χαρακτηριστικά χαμηλής και υψηλής συσχέτισης, αντίστοιχα. Ειδικότερα, το μέσο σφάλμα των PNN για χαρακτηριστικά χαμηλής συσχέτισης είναι 16.34%, ενώ για υψηλής συσχέτισης χαρακτηριστικά είναι 13.00%. Κάτι παρόμοιο συμβαίνει και με το 1NN, όπου το μέσο σφάλμα για χαμηλής και υψηλής συσχέτισης χαρακτηριστικά είναι 20.40% και 24.39%, αντίστοιχα. Ωστόσο, το λογιστικό υπόδειγμα πιθανότητας, τα πιθανολογικά νευρωνικά δίκτυα, οι μηχανές διανύσματος υποστήριξης και η UTADIS παρουσιάζουν βελτίωση καθώς η συσχέτιση των χαρακτηριστικών αυξάνει. Από τις παραπάνω μεθόδους, η UTADIS και οι SVM παρουσιάζουν μικρότερα ποσοστά σφαλμάτων σε σχέση με τις υπόλοιπες μεθόδους και η βελτίωση είναι της τάξης του 9.55% και 7.85%, αντίστοιχα. Παρόλα αυτά, οι SVM παρουσιάζουν το μικρότερο μέσο σφάλμα με τιμές για χαμηλή και υψηλή συσχέτιση χαρακτηριστικών 14.26% και 13.14%, αντίστοιχα. Τα αποτελέσματα φαίνονται αναλυτικά για όλες τις μεθόδους στο πίνακα που ακολουθεί:

Πίνακας 5.3: Μέσο σφάλμα ανά συσχέτιση χαρακτηριστικών.

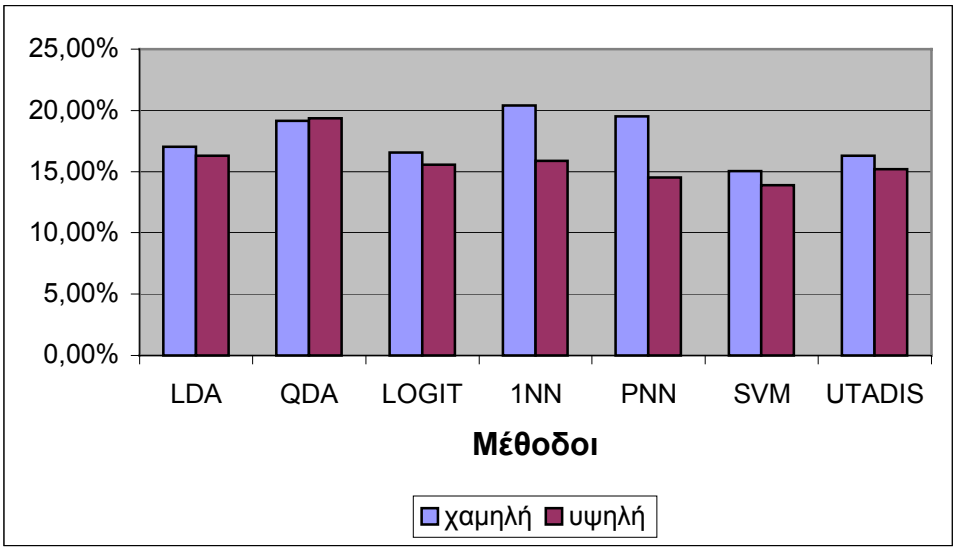
Μέθοδοι	Χαμηλή	Υψηλή
LDA	15.80% (3)	14.59% (3)
QDA	16.54% (4)	16.52% (4)
LOGIT	15.26% (2)	14.00% (2 ή 3)
1NN	20.40% (5)	24.39% (5)
PNN	16.34% (3 ή 4)	13.00% (1)
SVM	14.26% (1)	13.14% (1)
UTADIS	14.97% (2)	13.54% (1 ή 2)

Στη συνέχεια εξετάζεται η συσχέτιση των χαρακτηριστικών σε σχέση με τον τύπο δεδομένων (συνεχή, διακριτά). Το μέσο σφάλμα των μεθόδων ως προς τον τύπο δεδομένων και το βαθμό συσχέτισης, φαίνεται στον παρακάτω πίνακα:

Πίνακας 5.4: Μέσο σφάλμα ανά βαθμό συσχέτισης και ανά τύπο δεδομένων:

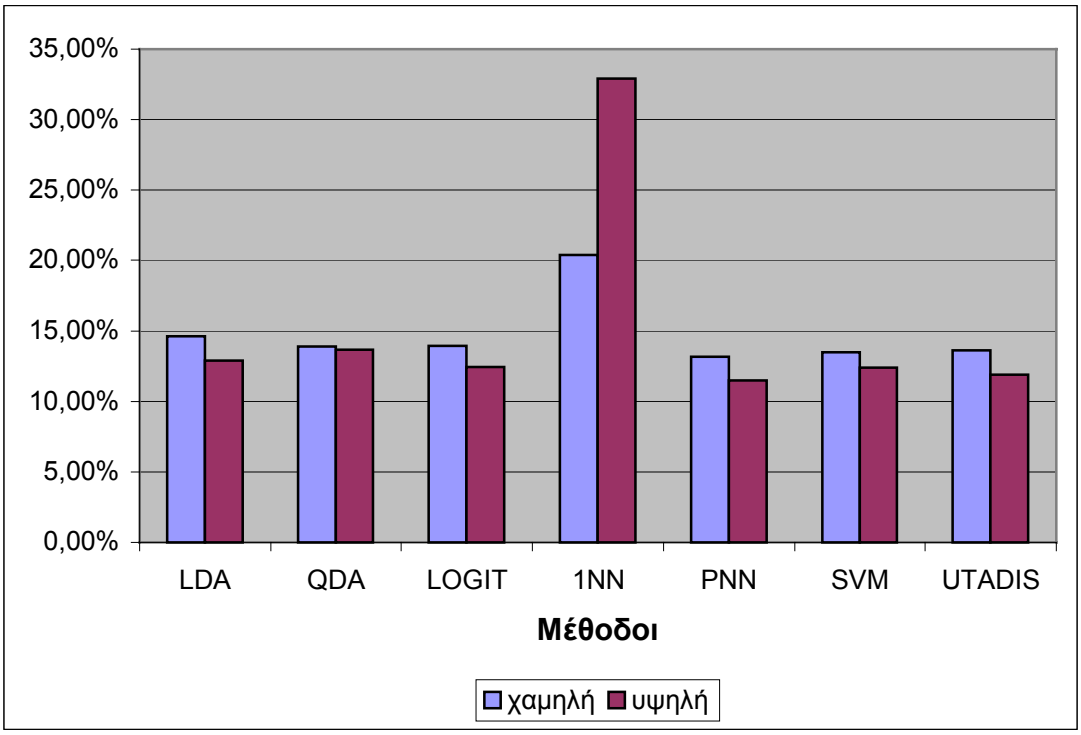
	Συνεχή		Διακριτά	
Μέθοδοι	Χαμηλή	Υψηλή	Χαμηλή	Υψηλή
LDA	17.01% (3)	16.30% (4)	14.60% (3)	12.88% (3)
QDA	19.16% (3)	19.37% (5)	13.91% (2 ή 3)	13.67% (4)
LOGIT	16.56% (2 ή 3)	15.55% (2 ή 3)	13.95% (2 ή 3)	12.45% (2 ή 3)
1NN	20.43% (5)	15.87% (3 ή 4)	20.38% (4)	32.90% (5)
PNN	19.53% (4)	14.50% (1)	13.15% (1)	11.49% (1)
SVM	15.05% (1)	13.87% (1)	13.47% (1 ή 2)	12.41% (2 ή 3)
UTADIS	16.29% (2)	15.19% (2)	13.64% (1 ή2)	11.88% (1 ή 2)

Σύμφωνα με τον παραπάνω πίνακα, και ελέγχοντας τις μεθόδους για συνεχή δεδομένα, παρατηρείται ότι η τετραγωνική διακριτική ανάλυση παρουσιάζει ευστάθεια σε συσχετισμένα χαρακτηριστικά. Συγκεκριμένα, το μέσο σφάλμα της QDA για χαμηλής και υψηλής συσχέτισης χαρακτηριστικά είναι 19.16% και 19.37%, αντίστοιχα. Αντίθετα, τα πιθανολογικά νευρωνικά δίκτυα και ο αλγόριθμος πλησιέστερου γείτονα παρουσιάζουν μεγάλη ευαισθησία σε χαρακτηριστικά χαμηλής συσχέτισης. Ωστόσο, οι δύο αυτές μέθοδοι παρουσιάζουν μεγάλη βελτίωση σε σχέση με τις υπόλοιπες μεθόδους, και η βελτίωση αυτή είναι τα τάξης του 25.75% για τα PNN, και 22.32% για το 1NN. Αναλυτικότερα, το μέσο σφάλμα των PNN για χαμηλής και υψηλής συσχέτισης χαρακτηριστικά είναι 19.53% και 14.50%, ενώ το μέσο σφάλμα του 1NN είναι 20.43% και 15.87%, αντίστοιχα. Από όλες τις παραπάνω μεθόδους, το μικρότερο μέσο σφάλμα παρουσιάζουν τα SVM με τιμές για χαμηλή και υψηλή συσχέτιση 15.05% και 13.87%, αντίστοιχα. Η εξέταση του ποσοστού σφάλματος ταξινόμησης, που αναλύθηκε παραπάνω φαίνεται στο παρακάτω σχήμα:



Σχήμα 5.3: Σφάλμα ταξινόμησης ανά βαθμό συσχέτισης, συνεχών δεδομένων

Εξετάζοντας το μέσο σφάλμα των μεθόδων ως προς το βαθμό συσχέτισης για τα διακριτά δεδομένα, παρατηρείται ότι και σε αυτή την περίπτωση η τετραγωνική διακριτική ανάλυση παρουσιάζει ευστάθεια σε συσχετισμένα χαρακτηριστικά. Το μέσο σφάλμα της QDA για χαμηλής και υψηλής συσχέτισης χαρακτηριστικών είναι 13.91% και 13.67%, αντίστοιχα. Γενικότερα, παρατηρώντας το μέσο σφάλμα όλων των μεθόδων συμπεραίνεται ότι παρουσιάζουν μια σχετική βελτίωση σε συσχετισμένα χαρακτηριστικά των δεδομένων με εξαίρεση αυτή του αλγορίθμου πλησιέστερου γείτονα, όπου το μέσο σφάλμα χαμηλής συσχέτισης χαρακτηριστικών βελτιώνεται αισθητά σε σχέση με το μέσο σφάλμα υψηλής συσχέτισης. Το μέσο σφάλμα του 1NN για χαμηλού και υψηλού βαθμού συσχέτισης χαρακτηριστικών είναι 20.38% και 32.90%, αντίστοιχα. Η ευαισθησία που παρουσιάζει στον υψηλό βαθμό συσχέτισης των χαρακτηριστικών, το 1NN, είναι ιδιαίτερα μεγάλη. Οι υπόλοιπες μέθοδοι παρουσιάζουν μια βελτίωση όσον αφορά τα χαρακτηριστικά με υψηλή συσχέτιση. Συγκεκριμένα, η γραμμική διακριτική ανάλυση παρουσιάζει μια βελτίωση της τάξης του 11.78%, τα πιθανολογικά νευρωνικά δίκτυα της τάξης του 12.62%, οι μηχανές διανύσματος υποστήριξης της τάξης του 7.86%, και τέλος η UTADIS της τάξης του 12.90%. Βέβαια, από όλες τις παραπάνω μεθόδους, το μικρότερο μέσο σφάλμα παρουσιάζουν τα PNN, με τιμές 13.15% και 11.49%, για χαρακτηριστικά χαμηλής και υψηλής συσχέτισης. Τα παραπάνω αποτελέσματα που προκύπτουν παρουσιάζονται στο παρακάτω σχήμα:



Σχήμα 5.4: Σφάλμα ταξινόμησης ανά βαθμό συσχέτισης, διακριτών δεδομένων.

5.3 Αποτελέσματα ανά μέγεθος δείγματος εκμάθησης

Όπως αναφέρθηκε στο τέταρτο κεφάλαιο, το πλήθος των αντικειμένων που επιλέγεται στο δείγμα εκμάθησης είναι 200, 500 και 1000 αντικείμενα, ενώ για το δείγμα ελέγχου επιλέγονται 500 αντικείμενα. Ανάλογα με το πλήθος των αντικειμένων που επιλέγεται κάθε φορά, επηρεάζεται η αποτελεσματικότητα των μεθόδων και η αξιοπιστία των αποτελεσμάτων. Μια γενική εικόνα που προκύπτει για το μέσο σφάλμα ανά πλήθους των αντικειμένων είναι η ακόλουθη:

Πίνακας 5.5: Μέσο σφάλμα σε σχέση με το πλήθος των αντικειμένων.

Πλήθος αντικείμενων	Μέσο σφάλμα
200	16.59% (2)
500	15.57% (1)
1000	15.57% (1)

Από τον παραπάνω πίνακα φαίνεται ότι το μέσο σφάλμα, όταν το πλήθος των αντικειμένων είναι 500 και 1000 είναι 15.57%. Αντίθετα, όταν το δείγμα εκμάθησης αποτελείται από 200 αντικείμενα, το μέσο σφάλμα είναι 16.59%. Συγκεκριμένα, από

τον παραπάνω πίνακα φαίνεται ότι με την αύξηση του πλήθους των αντικειμένων στο δείγμα εκμάθησης, βελτιώνεται το μέσο ποσοστό σφάλματος.

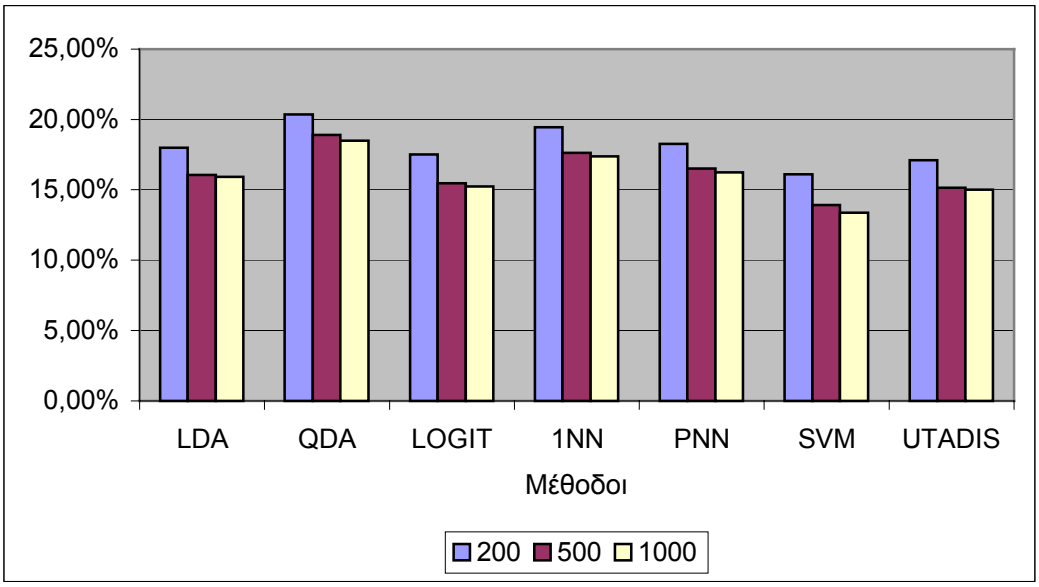
Εξετάζοντας το μέσο σφάλμα των μεθόδων ως προς το μέγεθος του δείγματος εκμάθησης, προκύπτουν τα ακόλουθα αποτελέσματα που παρουσιάζονται στον πίνακα 5.6:

Πίνακας 5.6: Μέσο σφάλμα των μεθόδων ανά μέγεθος δείγματος εκμάθησης.

Μέθοδοι	200	500	1000
LDA	16.17% (3)	14.83% (3)	14.59% (3)
QDA	17.87% (4)	16.20% (4)	15.52% (4)
LOGIT	15.66% (2 ή 3)	14.24% (2 ή 3)	13.99% (2 ή 3)
1NN	20.48% (5)	22.33% (5)	24.37% (5)
PNN	15.73% (2 ή 3)	14.29% (2 ή 3)	13.99% (2 ή 3)
SVM	14.95% (1)	13.31% (1)	12.84% (1)
UTADIS	15.26% (1 ή 2)	13.78% (1 ή 2)	13.72% (2)

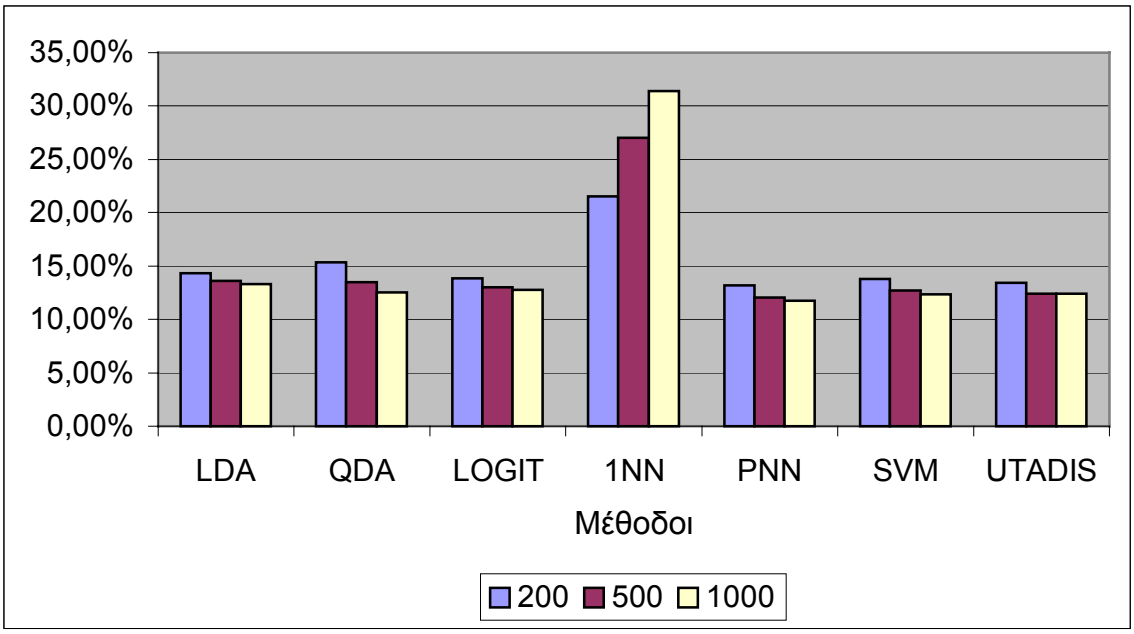
Από τον παραπάνω πίνακα προκύπτει ότι όλες οι μέθοδοι παρουσιάζουν βελτίωση καθώς το μέγεθος του δείγματος εκμάθησης αυξάνει, με εξαίρεση τον αλγόριθμο πλησιέστερου γείτονα, όπου σε αυτή την περίπτωση συμβαίνει ακριβώς το αντίθετο. Συγκεκριμένα, στον αλγόριθμο πλησιέστερου γείτονα, όσο αυξάνει το μέγεθος του δείγματος εκμάθησης τόσο αυξάνει το μέσο σφάλμα της μεθόδου. Ενώ, όταν το δείγμα αποτελείται από 200 αντικείμενα το μέσο σφάλμα είναι 20.48%, όταν το δείγμα αποτελείται από 1000 αντικείμενα το μέσο σφάλμα είναι 24.37%. Όλες οι υπόλοιπες μέθοδοι παρουσιάζουν βελτίωση καθώς το μέγεθος του δείγματος αυξάνει σε 1000 αντικείμενα. Συνολικά από όλες τις μεθόδους την καλύτερη βελτίωση παρουσιάζουν οι μηχανές διανύσματος υποστήριξης, όπου όταν το δείγμα αποτελείται από 1000 αντικείμενα, το μέσο σφάλμα είναι 12.84%. Επιπρόσθετα, μεγάλη βελτίωση παρουσιάζει και η UTADIS, όπου για μέγεθος δείγματος 200 αντικειμένων έχει μέσο σφάλμα 15.26%, ενώ για μέγεθος δείγματος 1000 αντικειμένων έχει μέσο σφάλμα 13.72%.

Τα προηγούμενα αποτελέσματα αφορούν και για τους δύο τύπους δεδομένων, δηλαδή όταν τα δεδομένα είναι συνεχή και όταν είναι διακριτά. Εξετάζοντας τη συμπεριφορά των μεθόδων μόνο στην περίπτωση των συνεχών δεδομένων, προκύπτει το ακόλουθο σχήμα:



Σχήμα 5.5: Σφάλμα ταξινόμησης ανά μέγεθος δείγματος συνεχών δεδομένων

Από το παραπάνω σχήμα είναι φανερό ότι όλες οι μέθοδοι παρουσιάζουν βελτίωση καθώς το μέγεθος του δείγματος εκμάθησης αυξάνει. Τα ποσοστά σφαλμάτων ταξινόμησης μειώνονται, με μικρότερο σφάλμα να παρουσιάζουν οι μηχανές διανύσματος υποστήριξης 13.37%, οι οποίες παρουσιάζουν και τη μεγαλύτερη βελτίωση σε σχέση με τις υπόλοιπες μεθόδους. Ο αλγόριθμος πλησιέστερου γείτονα παρουσιάζει βελτίωση, αν και το ποσοστό σφάλματος παραμένει το ίδιο, όταν το δείγμα αποτελείται από 500 και 1000 αντικείμενα, ίσο με 17.64%. Τα ποσοστά σφαλμάτων ταξινόμησης αφορούν την περίπτωση που τα δεδομένα είναι συνεχή. Όταν τα δεδομένα είναι διακριτά, υπάρχει μια μικρή διαφοροποίηση των αποτελεσμάτων. Τα σφάλματα ταξινόμησης που προκύπτουν στην περίπτωση των διακριτών δεδομένων, παρουσιάζονται στο ακόλουθο σχήμα:



Σχήμα 5.6:Σφάλμα ταξινόμησης ανά μέγεθος δείγματος διακριτών δεδομένων.

Από το παραπάνω σχήμα φαίνεται ότι όλες οι μέθοδοι παρουσιάζουν βελτίωση καθώς το μέγεθος του δείγματος εκμάθησης αυξάνει, με εξαίρεση τον αλγόριθμο πλησιέστερου γείτονα. Τη μεγαλύτερη βελτίωση παρουσιάζουν τα πιθανολογικά νευρωνικά δίκτυα, όπου για μέγεθος δείγματος 200 αντικειμένων έχουν μέσο σφάλμα 13.19% ενώ για μέγεθος δείγματος 1000 αντικειμένων έχουν 11.72%. Στην περίπτωση των διακριτών δεδομένων, ο αλγόριθμος πλησιέστερου γείτονα παρουσιάζει μεγάλο ποσοστό σφάλμα ταξινόμησης. Συγκεκριμένα, όταν το δείγμα εκμάθησης αποτελείται από 200 αντικείμενα, το σφάλμα ταξινόμησης είναι 21.49%, ενώ όταν το δείγμα αυξάνει σε 1000 αντικείμενα το σφάλμα ταξινόμησης που προκύπτει είναι 31.40%. Είναι εμφανές ότι το μέγεθος του δείγματος επηρεάζει την αποτελεσματικότητα του 1NN και την αξιοπιστία των αποτελεσμάτων. Ο 1NN παρουσιάζει την συγκεκριμένη ευαισθησία μόνο στην περίπτωση των διακριτών δεδομένων. Όπως έχει αναφερθεί παραπάνω, όταν τα δεδομένα είναι συνεχή παρουσιάζει κάποια σχετική βελτίωση καθώς το μέγεθος του δείγματος αυξάνει. Όπως, έχει παρουσιαστεί στον πίνακα 5.6, η γενική συμπεριφορά του 1NN στον παράγοντα του μεγέθους του δείγματος εκμάθησης είναι αρνητική, εφόσον επηρεάζεται η αποτελεσματικότητα της μεθόδου από το πλήθος των αντικειμένων. Ωστόσο, η αποτελεσματικότητα των υπολοίπων μεθόδων βελτιώνεται με την αύξηση του πλήθους των αντικειμένων στο δείγμα εκμάθησης, και αυτό φαίνεται από την

μείωση που παρουσιάζει το ποσοστό σφάλματος ταξινόμησης. Η βελτίωση αυτή είναι συνολική και για τους δύο τύπους δεδομένων, συνεχή και διακριτά.

5.4 Αποτελέσματα ανά μορφή ταξινόμησης

Εξετάζοντας συνολικά τις μεθόδους και για τους δύο τύπους δεδομένων στη μορφή διάκρισης των κατηγοριών (γραμμική, μη γραμμική), παρατηρείται ότι όλες οι μέθοδοι παρουσιάζουν ευαισθησία. Σε ορισμένες μέθοδοι δεν είναι τόσο μεγάλη η ευαισθησία που παρουσιάζουν και αυτό φαίνεται από τη μικρή διαφοροποίηση που υπάρχει στο μέσο σφάλμα. Το μέσο σφάλμα των μεθόδων χειροτερεύει στην περίπτωση της μη γραμμικής διάκρισης των κατηγοριών. Ακολουθεί ο πίνακας που παρουσιάζει το μέσο σφάλμα των μεθόδων ανάλογα με τη διάκριση των κατηγοριών:

Πίνακας 5.7: Μέσο σφάλμα ανά μορφή διάκρισης των κατηγοριών.

Μέθοδοι	Γραμμική	Μη γραμμική
LDA	13.12% (2)	17.28% (3)
QDA	15.87% (4)	17.19% (3)
LOGIT	12.46% (1)	16.79% (3)
1NN	22.13% (5)	22.66% (4)
PNN	13.79% (3)	15.55% (2)
SVM	13.08% (2)	14.31% (1)
UTADIS	12.45% (1)	16.06% (2)

Από τον παραπάνω πίνακα προκύπτει ότι οι μέθοδοι που παρουσιάζουν πολύ μεγάλη ευαισθησία στη μορφή διάκρισης είναι η UTADIS, η LOGIT και η γραμμική διακριτική ανάλυση. Συγκεκριμένα, η UTADIS στη γραμμική διάκριση έχει μέσο σφάλμα 12.45%, ενώ στην μη γραμμική το μέσο σφάλμα γίνεται 16.06%. Παρόμοια, στο λογιστικό υπόδειγμα το μέσο σφάλμα για την γραμμική και μη γραμμική διάκριση είναι 12.46% και 16.79%, αντίστοιχα. Ομοίως, στη γραμμική διακριτική ανάλυση το μέσο σφάλμα για την γραμμική και διακριτική ανάλυση είναι 13.12% και 17.28%, αντίστοιχα. Αυτές οι τρεις μέθοδοι παρουσιάζουν μεγαλύτερη ευαισθησία στη μορφή διάκρισης των κατηγοριών σε σύγκριση με τις υπόλοιπες, των οποίων η ευαισθησία είναι μικρότερη. Τα παραπάνω αποτελέσματα αφορούν και για τους δύο

τύπους δεδομένων, συνεχών και διακριτών. Το μέσο σφάλμα της μορφής διάκρισης των κατηγοριών για τα συνεχή και διακριτά δεδομένα παρουσιάζονται στον παρακάτω πίνακα:

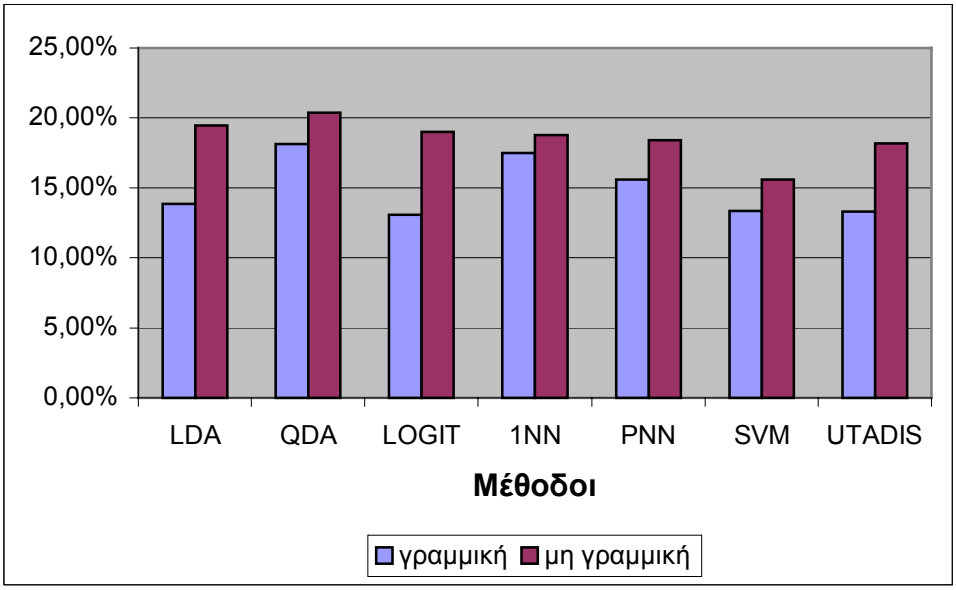
Πίνακας 5.8: Μέσο σφάλμα ανά μορφή διάκρισης και ανά τύπο δεδομένων.

	Συνεχή		Διακριτά	
Μέθοδοι	Γραμμική	Μη γραμμική	γραμμική	Μη γραμμική
LDA	13.85% (2)	19.46% (4)	12.38% (1 ή 2)	15.09% (3)
QDA	18.16% (5)	20.38% (5)	13.58% (3)	14.00% (2)
LOGIT	13.09% (1)	19.02% (3 ή 4)	11.83% (1)	14.56% (2 ή 3)
1NN	17.57% (4)	18.79% (2 ή 3 ή 4)	26.76% (4)	26.52% (4)
PNN	15.59% (3)	18.44% (2 ή 3)	11.98% (1)	12.66% (1)
SVM	13.33% (1)	15.60% (1)	12.84% (2 ή 3)	13.03% (1)
UTADIS	13.30% (1)	18.19% (2)	11.61% (1)	13.91% (2)

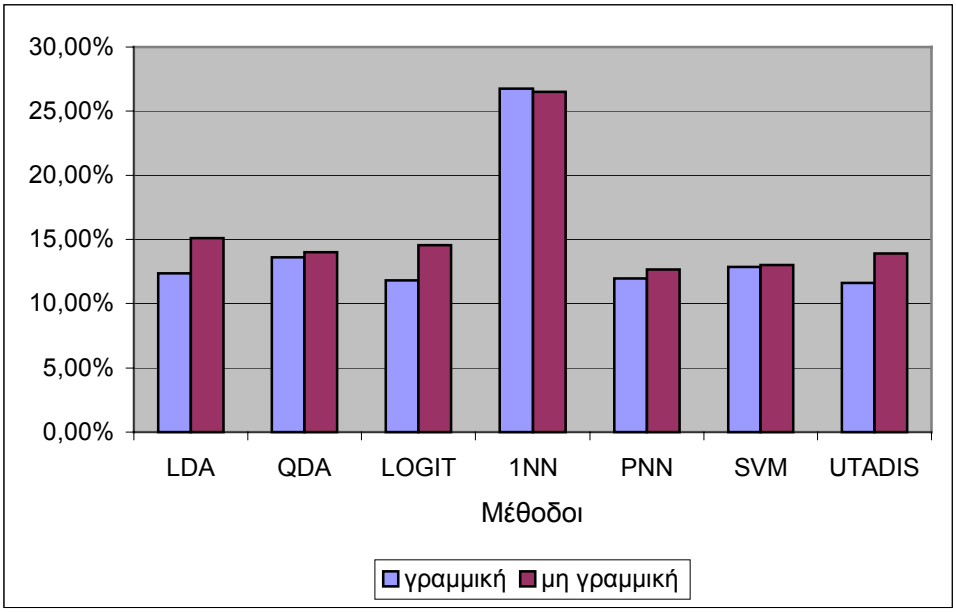
Από τον παραπάνω πίνακα προκύπτουν παρόμοια αποτελέσματα σε σύγκριση με αυτά του πίνακα 5.7. Όσον αφορά τα συνεχή δεδομένα, οι μέθοδοι που παρουσιάζουν τη μεγαλύτερη ευαισθησία στη μορφή διάκρισης των κατηγοριών είναι το λογιστικό υπόδειγμα πιθανότητας, η UTADIS, η γραμμική διακριτική ανάλυση και τα πιθανολογικά νευρωνικά δίκτυα, όπου το μέσο σφάλμα για τη γραμμική και μη γραμμική διάκριση είναι 13.09% και 19.02%, 13.30% και 18.19%, 13.85% και 19.46%, 15.59% και 18.44%, αντίστοιχα. Όσον αφορά τα διακριτά δεδομένα, η μέθοδος που παρουσιάζει τη μεγαλύτερη ευαισθησία στη μορφή διάκρισης των κατηγοριών, σε σύγκριση με τις υπόλοιπες, είναι η γραμμική διακριτική ανάλυση, όπου το μέσο σφάλμα για τη γραμμική και μη γραμμική διάκριση είναι 12.38% και 15.09%, αντίστοιχα. Αντίθετα, οι μηχανές διανύσματος υποστήριξης παρουσιάζουν κάποια ευστάθεια όταν τα δεδομένα είναι διακριτά. Συγκεκριμένα, το μέσο σφάλμα για τη γραμμική και μη γραμμική διάκριση είναι 12.84% και 13.03% αντίστοιχα.

Γενικότερα, όλες οι μέθοδοι παρουσιάζουν ευαισθησία στη μορφή διάκρισης των κατηγοριών, είτε εξετάζονται συνολικά και για τους δύο τύπους δεδομένων, είτε εξετάζονται χωριστά για τα συνεχή και διακριτά δεδομένα, με εξαίρεση τις μηχανές διανύσματος υποστήριξης όπου παρουσιάζουν μια σχετική ευστάθεια.

Από την εξέταση των ποσοστών σφαλμάτων ταξινόμησης των μεθόδων ανά τύπο δεδομένων και ανά μορφή διάκρισης των κατηγοριών προκύπτουν τα ακόλουθα δύο σχήματα:



Σχήμα 5.7: Σφάλμα ταξινόμησης ανά μορφή διάκρισης για συνεχή δεδομένα.



Σχήμα 5.8: Σφάλμα ταξινόμησης ανά μορφή διάκρισης για διακριτά δεδομένα.

Από τα παραπάνω δύο σχήματα είναι φανερή η ευαισθησία που παρουσιάζουν οι μέθοδοι στη μορφή διάκρισης των κατηγοριών. Η ευαισθησία είναι μεγαλύτερη στην

περίπτωση που τα δεδομένα είναι συνεχή και όχι τόσο στην περίπτωση των διακριτών δεδομένων.

5.5 Αποτελέσματα ανά στατιστική κατανομή και ανά μορφή διακριτών επιπέδων

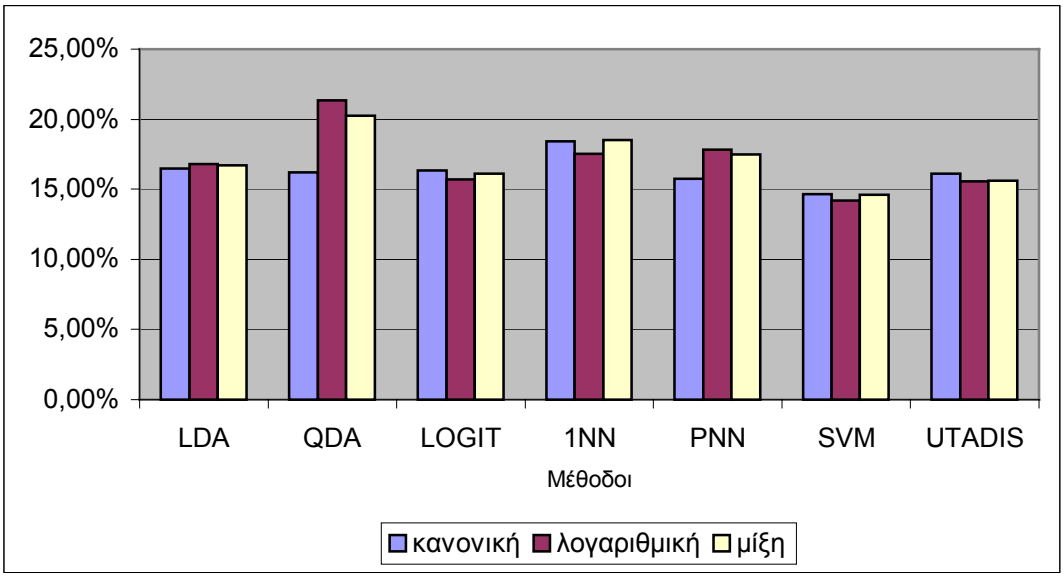
Ο πρώτος παράγοντας που αναλύθηκε στο κεφάλαιο 4 αφορούσε τη στατιστική κατανομή που ακολουθούν τα συνεχή δεδομένα και το πλήθος των διακριτών επιπέδων που διαθέτουν τα διακριτά δεδομένα. Η στατιστική κατανομή αφορούσε την κανονική, την λογαριθμική και τη μίξη κανονικής – λογαριθμικής, ενώ η μορφή των επιπέδων τις διακριτές τιμές –1 και 1, τις τιμές –1, 0, 1, και τη μίξη δύο και τριών επιπέδων.

Εξετάζοντας αρχικά τα συνεχή δεδομένα και το μέσο σφάλμα ανά στατιστική κατανομή, προκύπτει ο ακόλουθος πίνακας:

Πίνακας 5.9: Μέσο σφάλμα ανά στατιστική κατανομή.

Μέθοδοι	Κανονική	Λογαριθμική	Μίξη
LDA	16.48% (2)	16.79% (3)	16.70% (3)
QDA	16.20% (2)	21.33% (5)	20.27% (6)
LOGIT	16.34% (2)	15.71% (2)	16.12% (2 ή 3)
1NN	18.41% (3)	17.52% (3 ή 4)	18.52% (5)
PNN	15.76% (2)	17.82% (4)	17.47% (4)
SVM	14.64% (1)	14.17% (1)	14.58% (1)
UTADIS	16.10% (2)	15.55% (2)	15.58% (2)

Από τον παραπάνω πίνακα προκύπτει ότι όλες σχεδόν οι μέθοδοι παρουσιάζουν ευσταθή αποτελέσματα στους διάφορους τύπους κατανομών, με εξαίρεση την τετραγωνική διακριτική ανάλυση, η οποία παρουσιάζει ευαισθησία στις μη κανονικές κατανομές. Συγκεκριμένα, για κανονική κατανομή έχει μέσο σφάλμα 16.20%, ενώ όταν η κατανομή είναι λογαριθμική ή μίξη κανονικής – λογαριθμικής, το μέσο σφάλμα είναι 21.33% και 20.27%, αντίστοιχα. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στο ακόλουθο σχήμα που δείχνει τα ποσοστά σφάλματος ταξινόμησης:



Σχήμα 5.9: Σφάλμα ταξινόμησης ανά τύπο στατιστικής κατανομής.

Στο παραπάνω σχήμα είναι εμφανής η ευαισθησία που παρουσιάζει η τετραγωνική διακριτική ανάλυση σε μη κανονικές κατανομές. Κάποια ευαισθησία παρουσιάζουν και τα πιθανολογικά νευρωνικά δίκτυα, αλλά είναι σε μικρότερο βαθμό από αυτή της τετραγωνικής διακριτικής ανάλυσης. Αντίθετα οι μηχανές διανύσματος υποστήριξης και η UTADIS παρουσιάζουν ευσταθή αποτελέσματα για τους διάφορους τύπους κατανομών, σε σύγκριση με τις υπόλοιπες. Αναλυτικότερα, η UTADIS παρουσιάζει βελτίωση του μέσου σφάλματος στην περίπτωση της λογαριθμικής και της μίξης λογαριθμική – κανονικής κατανομής. Συγκεκριμένα, το μέσο σφάλμα της UTADIS για την κανονική κατανομή είναι 16.10%, ενώ για τη λογαριθμική και τη μίξη κανονικής – λογαριθμικής το μέσο σφάλμα βελτιώνεται σε 15.55% και 15.58%, αντίστοιχα.

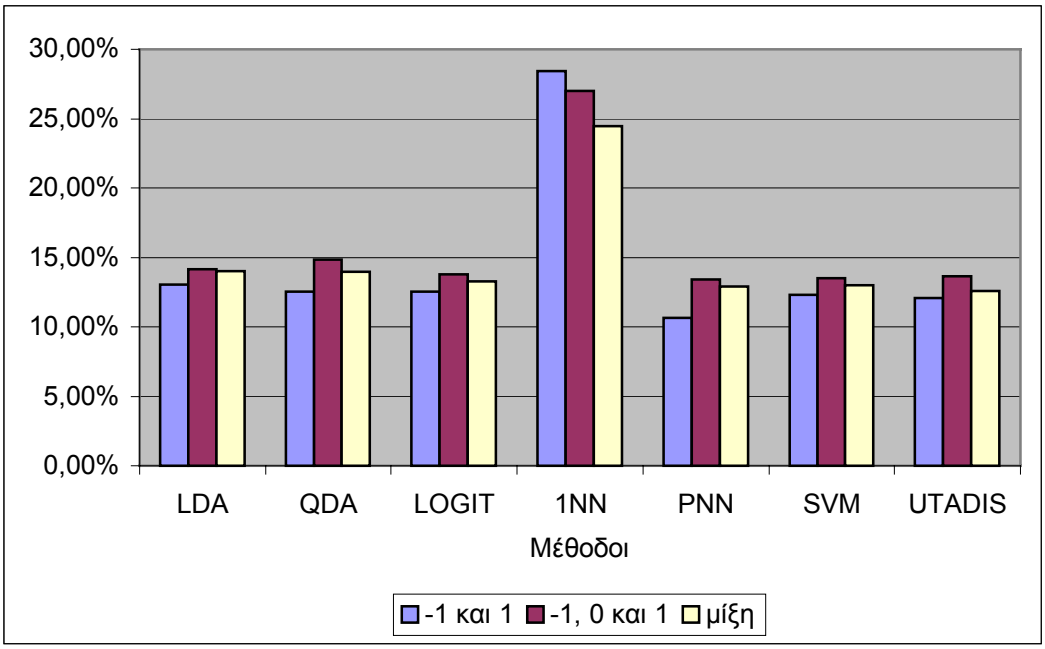
Αντίστοιχα αποτελέσματα προκύπτουν και από την εξέταση των μεθόδων ως προς τον τύπο των διακριτών επιπέδων που έχουν τα διακριτά δεδομένα, με τη διαφορά ότι στην περίπτωση αυτή η ευαισθησία που παρουσιάζουν οι μέθοδοι δεν είναι τόσο μεγάλη σε σύγκριση με αυτή, όταν τα δεδομένα είναι συνεχή. Το μέσο σφάλμα που προκύπτει ανά τύπο διακριτών επιπέδων, φαίνεται στον πίνακα που ακολουθεί:

Πίνακας 5.10: Μέσο σφάλμα ανά πλήθος διακριτών επιπέδων.

Μέθοδοι	-1 και1	1, 0 και 1	Μίξη
LDA	13.06% (2)	14.13% (1 ή 2)	14.02% (3)
QDA	12.53% (2)	14.86% (2)	13.98% (2 ή 3)
LOGIT	12.53% (2)	13.80% (1)	13.26% (1 ή 2 ή 3)
1NN	28.45% (3)	27.00% (3)	24.47% (4)
PNN	10.64% (1)	13.43% (1)	12.90% (1 ή 2)
SVM	12.32% (2)	13.49% (1)	12.99% (1 ή 2 ή 3)
UTADIS	12.07% (2)	13.64% (1)	12.58% (1)

Από τον παραπάνω πίνακα προκύπτει ότι οι μέθοδοι που δεν παρουσιάζουν ευσταθή αποτελέσματα στα διάφορα διακριτά επίπεδα είναι τα πιθανολογικά νευρωνικά δίκτυα, η τετραγωνική διακριτική ανάλυση και ο αλγόριθμος πλησιέστερου γείτονα. Οι δύο πρώτες μέθοδοι παρουσιάζουν ευαισθησία όταν ο τύπος των διακριτών επιπέδων είναι τρία και όταν πραγματοποιείται μίξη δύο και τριών επιπέδων. Αντίθετα, ο αλγόριθμος πλησιέστερου γείτονα παρουσιάζει βελτίωση στο μέσο σφάλμα όταν εφαρμόζεται στη μίξη δύο και τριών επιπέδων, με ποσοστό 24.47%, ενώ όταν εφαρμόζεται σε δύο και τρία επίπεδα, το μέσο σφάλμα που προκύπτει είναι 28.45% και 27.00%, αντίστοιχα. Οι υπόλοιπες μέθοδοι παρουσιάζουν γενικά ευσταθή αποτελέσματα στα διάφορα διακριτά επίπεδα, χωρίς ιδιαίτερες διαφοροποιήσεις στις τιμές του μέσου σφάλματος.

Τα παραπάνω αποτελέσματα που προέκυψαν από την εξέταση του ποσοστού σφάλματος ταξινόμησης των μεθόδων, παρουσιάζονται στο ακόλουθο σχήμα:



Σχήμα 5.10: Σφάλμα ταξινόμησης ανά πλήθος διακριτών επιπέδων.

Από το παραπάνω σχήμα είναι εμφανές ότι ο αλγόριθμος πλησιέστερου γείτονα παρουσιάζει ευαισθησία όταν ο τύπος του επιπέδου είναι δύο και όταν είναι τρία. Αντίθετα, στη μίξη των δύο και τριών επιπέδων το ποσοστό σφάλματος είναι μικρότερο. Τα πιθανολογικά νευρωνικά δίκτυα και η τετραγωνική διακριτική ανάλυση παρουσιάζουν ευαισθησία στα τρία διακριτά επίπεδα και στη μίξη δύο και τριών επιπέδων. Οι υπόλοιπες μέθοδοι παρουσιάζουν γενικά ευσταθή αποτελέσματα στις τρεις διαφορετικές περιπτώσεις διακριτών επιπέδων, με μεγαλύτερη ευστάθεια να παρουσιάζει η γραμμική διακριτική ανάλυση. Συγκεκριμένα, το μέσο σφάλμα της γραμμικής διακριτικής ανάλυσης για τις τρεις περιπτώσεις διακριτών επιπέδων είναι 13.06%, 14.13% και 14.02% αντίστοιχα.

5.6 Συμπεράσματα

Μια σύνοψη των αποτελεσματικότερων μεθόδων ταξινόμησης, εξετάζοντας όλους τους παράγοντες που χρησιμοποιήθηκαν κατά την εφαρμογή τους, φαίνεται στον ακόλουθο πίνακα:

Πίνακας 5.11: Σύνοψη των αποτελεσματικότερων μεθόδων ταξινόμησης βάσει των ποσοστών εσφαλμένων ταξινομήσεων.

		Τύπος δεδομένων	
		Συνεχή	**Διακριτά**
Συνολική αξιολόγηση		SVM	PNN, UTADIS
Βαθμός συσχέτισης	**Χαμηλός**	SVM	PNN, SVM, UTADIS
	Υψηλός	PNN, SVM	PNN, UTADIS
Μέγεθος δείγματος	**200**	SVM, UTADIS	SVM, UTADIS
	500	SVM, UTADIS	SVM, UTADIS
	1000	SVM	SVM
Μορφή ταξινόμησης	**Γραμμική**	SVM, UTADIS	PNN, UTADIS, LDA, LOGIT
	Μη γραμμική	SVM	PNN, SVM
Στατιστική κατανομή	**Κανονική**	SVM	
	Λογαριθμική	SVM	
	Μίξη	SVM	
Μορφή διακριτών επιπέδων	**-1 και 1**		PNN
	1, 0, και1		LOGIT, PNN,LDA SVM, UTADIS
	Μίξη		UTADIS, PNN

Σύμφωνα με τον παραπάνω πίνακα, εξετάζοντας τις τεχνικές ταξινόμησης ως προς τον τύπο των δεδομένων, καλύτερη αποτελεσματικότητα, χρησιμοποιώντας συνεχή δεδομένα, παρουσιάζουν οι μηχανές διανύσματος υποστήριξης με μέσο σφάλμα 14.46% και ακολουθεί η μέθοδος UTADIS με σφάλμα 15.74%. Αντίθετα, η τετραγωνική διακριτική ανάλυση παρουσιάζει το χειρότερο ποσοστό σφάλμα ταξινόμησης σε σχέση με τις υπόλοιπες. Όσον αφορά τη χρήση των διακριτών δεδομένων, καλύτερη αποτελεσματικότητα παρουσιάζουν τα πιθανολογικά νευρωνικά δίκτυα και η UTADIS, με μέσο σφάλμα 12.32% και 12.76% αντίστοιχα, ενώ τη χειρότερη ο αλγόριθμος πλησιέστερου γείτονα.

Ελέγχοντας τις μεθόδους ως προς το βαθμό συσχέτισης των χαρακτηριστικών και ως προς τον τύπο δεδομένων, προέκυψε ότι για τα συνεχή δεδομένα καλύτερη αποτελεσματικότητα παρουσιάζουν οι μηχανές διανύσματος υποστήριξης, ενώ για τα διακριτά δεδομένα καλύτερη αποτελεσματικότητα παρουσιάζουν τα πιθανολογικά νευρωνικά δίκτυα, τα SVM και η UTADIS. Η τετραγωνική διακριτική ανάλυση

παρουσιάζει ευστάθεια σε συσχετισμένα χαρακτηριστικά, ενώ τα πιθανολογικά νευρωνικά δίκτυα και ο αλγόριθμος πλησιέστερου γείτονα παρουσιάζουν μεγάλη ευαισθησία σε χαρακτηριστικά χαμηλής και υψηλής συσχέτισης, αντίστοιχα. Το παραπάνω ισχύει και για τους δύο τύπους δεδομένων.

Η εξέταση των μεθόδων ως προς το μέγεθος του δείγματος εκμάθησης έδειξε ότι γενικά όλες οι μέθοδοι παρουσιάζουν βελτίωση στην αύξηση του μεγέθους του δείγματος. Σε σύγκριση με τις υπόλοιπες μεθόδους μεγαλύτερη βελτίωση στην αύξηση του δείγματος παρουσιάζουν οι μηχανές διανύσματος υποστήριξης και η UTADIS, για δείγματα 200 και 500 αντικειμένων. Για δείγματα με μέγεθος 1000 αντικειμένων, καλύτερη αποτελεσματικότητα παρουσιάζουν οι μηχανές διανύσματος υποστήριξης. Εξαίρεση στον παράγοντα αυτό αποτελεί ο αλγόριθμος πλησιέστερου γείτονα, ο οποίος όσο αυξάνει το μέγεθος του δείγματος τόσο χειροτερεύει το μέσο σφάλμα. Γενικά, επηρεάζεται η αποτελεσματικότητα και η αξιοπιστία των αποτελεσμάτων του 1NN, σε μεγάλα δείγματα. Συγκεκριμένα, για μέγεθος δείγματος 200 αντικειμένων παρουσιάζει σφάλμα 20.48%, ενώ για μέγεθος δείγματος 1000 αντικειμένων παρουσιάζει σφάλμα 24.37%. Ελέγχοντας τις μεθόδους ανά τύπο δεδομένων και ανά μέγεθος δείγματος εκμάθησης, προκύπτουν ίδια συμπεράσματα.

Στον παράγοντα που αφορά τη μορφή διάκρισης των κατηγοριών προκύπτει ότι όλες οι τεχνικές ταξινόμησης παρουσιάζουν μεγάλη ευαισθησία. Εξετάζοντας τις μεθόδους ως προς τα συνεχή δεδομένα και ανά μορφή ταξινόμησης προκύπτει ότι στη γραμμική διάκριση καλύτερη αποτελεσματικότητα παρουσιάζουν οι μηχανές διανύσματος υποστήριξης και η UTADIS, ενώ στη μη γραμμική διάκριση καλύτερη αποτελεσματικότητα παρουσιάζουν τα SVM. Εξετάζοντας τις μεθόδους ως προς τα διακριτά δεδομένα και ανά μορφή ταξινόμησης, μεγαλύτερη αποτελεσματικότητα στη γραμμική διάκριση παρουσιάζουν τα πιθανολογικά νευρωνικά δίκτυα, η γραμμική διακριτική διάκριση το λογιστικό υπόδειγμα πιθανότητας και η UTADIS. Αντίθετα στη μη γραμμική διάκριση καλύτερη αποτελεσματικότητα παρουσιάζουν τα πιθανολογικά νευρωνικά δίκτυα και οι μηχανές διανύσματος υποστήριξης, σε σχέση με τις υπόλοιπες μεθόδους.

Ο τελευταίος παράγοντας αφορά τη στατιστική κατανομή που ακολουθούν τα συνεχή δεδομένα και τη μορφή των διακριτών επιπέδων που έχουν τα διακριτά δεδομένα. Όσον αφορά τα συνεχή δεδομένα και τη στατιστική κατανομή, προέκυψε ότι οι μηχανές διανύσματος υποστήριξης παρουσιάζουν καλύτερη αποτελεσματικότητα και στους τρεις τύπους κατανομών (κανονική, λογαριθμική, και

μίξη κανονικής – λογαριθμικής), σε σχέση με τις υπόλοιπες. Αντίθετα, η τετραγωνική διακριτική ανάλυση και τα πιθανολογικά νευρωνικά δίκτυα παρουσιάζουν ευαισθησία σε μη κανονικές κατανομές. Για τους διάφορους τύπους διακριτών επιπέδων, προέκυψε ότι τα πιθανολογικά νευρωνικά δίκτυα παρουσιάζουν καλύτερη αποτελεσματικότητα και στις τρεις μορφές διακριτών επιπέδων, ενώ η UTADIS παρουσιάζει καλή αποτελεσματικότητα όταν το πλήθος των διακριτών επιπέδων είναι -1, 0 και 1, και στη μίξη δύο και τριών επιπέδων. Επιπρόσθετα, όσον αφορά, την περίπτωση όπου τα διακριτά επίπεδα είναι τρία, καλύτερη αποτελεσματικότητα παρουσιάζουν το λογιστικό υπόδειγμα πιθανότητας, η γραμμική διακριτική ανάλυση και οι μηχανές διανύσματος υποστήριξης.

Μια σύνοψη όλων των διμερών συγκρίσεων των εξεταζόμενων μεθόδων ταξινόμησης σε όλους τους συνδυασμούς παραγόντων του πειραματικού σχεδιασμού παρουσιάζεται στον πίνακα 5.12:

Πίνακας 5.12: Διμερής σύγκριση των αποτελεσμάτων των προσεγγίσεων ταξινόμησης

	LDA	QDA	LOGIT	1NN	PNN	SVM	UTADIS
LDA	-	76,39%	1,39%	77,78%	31,94%	27,78%	2,78%
QDA	23,61%	-	20,83%	65,28%	23,61%	9,72%	16,67%
LOGIT	98,61%	79,17%	-	83,33%	41,67%	41,67%	31,34%
1NN	22,22%	34,72%	16,67%	-	11,11%	0,00%	12,50%
PNN	68,06%	76,39%	58,33%	88,89%	-	40,28%	52,78%
SVM	59,72%	90,28%	56,94%	100,00%	58,33%	-	54,17%
UTADIS	97,22%	83,33%	66,67%	87,50%	47,22%	45,83%	-

Τα αποτελέσματα του παραπάνω πίνακα δείχνουν ότι στην πλειοψηφία των περιπτώσεων οι μηχανές διανύσματος υποστήριξης υπερείχαν των υπολοίπων τεχνικών ταξινόμησης. Επιπλέον, τα πιθανολογικά νευρωνικά δίκτυα και η UTADIS ακολουθούν τις μηχανές διανύσματος υποστήριξης, παρουσιάζοντας τις αμέσως καλύτερες αποτελεσματικότητες. Αυτές οι δύο μέθοδοι βρίσκονται κοντά, όσον αφορά τα αποτελέσματα που προέκυψαν για τους διάφορους συνδυασμούς των παραγόντων. Η αποτελεσματικότητα των μεθόδων αυτών επιβεβαιώνεται και από την ανάλυση των αποτελεσμάτων του πειραματικού σχεδιασμού που πραγματοποιήθηκε. Τα ποσοστά σφαλμάτων ήταν χαμηλότερα στην πλειοψηφία των περιπτώσεων, ιδιαίτερα, όταν τα δεδομένα ήταν συνεχή.

Τέλος, στον πίνακα 5.13 παρουσιάζεται μια κατάταξη των μεθόδων, σύμφωνα με τα ποσοστά σφαλμάτων που προέκυψαν για όλους τους συνδυασμούς των παραγόντων που χρησιμοποιήθηκαν.

Πίνακας 5.13: Κατάταξη των μεθόδων σύμφωνα με τα ποσοστά σφαλμάτων

	$1^{η}$	$2^{η}$	$3^{η}$	$4^{η}$	$5^{η}$	$6^{η}$	$7^{η}$
LDA	0,00%	1,39%	13,89%	44,44%	63,89%	94,44%	100,00%
QDA	0,00%	9,72%	15,28%	22,22%	38,89%	73,61%	100,00%
LOGIT	19,44%	34,72%	59,72%	72,22%	93,06%	100,00%	100,00%
1NN	0,00%	1,39%	11,11%	19,44%	29,17%	37,50%	100,00%
PNN	29,17%	51,39%	62,50%	68,06%	80,56%	94,44%	100,00%
SVM	36,11%	51,39%	69,44%	83,33%	94,44%	100,00%	100,00%
UTADIS	19,44%	51,39%	68,06%	91,67%	100,00%	100,00%	100,00%

Ο πίνακας 5.13 παρουσιάζει την κατάταξη των μεθόδων, η οποία προέκυψε με τον ακόλουθο τρόπο: το κελί (i,j) του πίνακα παρουσιάζει το ποσοστό των περιπτώσεων που εξετάστηκαν στην ανάλυση (72 συνδυασμοί των πέντε παραγόντων) για τις οποίες η μέθοδος της γραμμής i κατατάσσεται το πολύ στη j θέση της κατάταξης των μεθόδων βάσει του σφάλματός τους. Για παράδειγμα, το κελί (1,2) δείχνει ότι μόνο στο 1,39% των περιπτώσεων (σε ένα συνδυασμό των πέντε εξεταζόμενων παραγόντων) η γραμμική διακριτική ανάλυση ήταν στις πρώτες δύο καλύτερες θέσεις. Από τον παραπάνω πίνακα προκύπτει ότι οι μηχανές διανύσματος υποστήριξης κατατάσσονται πρώτες στο 36.11% των περιπτώσεων (26 συνδυασμοί των πέντε εξεταζόμενων παραγόντων), ενώ δεν κατατάσσονται ποτέ στην τελευταία θέση. Η μέθοδος UTADIS ήταν στην πρώτη θέση στο 19.44% των περιπτώσεων (14 συνδυασμοί των πέντε εξεταζόμενων παραγόντων), ενώ δεν κατατάχθηκε ποτέ στις τελευταίες δύο θέσεις. Σε μεγάλο ποσοστό των περιπτώσεων, τα πιθανολογικά νευρωνικά δίκτυα ήταν στις δύο πρώτες θέσεις, και συγκεκριμένα κατατάχθηκαν στην πρώτη θέση στο 29.17% των περιπτώσεων, αλλά υπήρξε και ένα 5.56% των περιπτώσεων (4 συνδυασμοί των πέντε εξεταζόμενων παραγόντων) στις οποίες τα πιθανολογικά νευρωνικά δίκτυα κατατάχθηκαν στην τελευταία θέση της κατάταξης. Τέλος αξιοσημείωτο είναι το γεγονός ότι η γραμμική και η τετραγωνική διακριτική ανάλυση καθώς και ο αλγόριθμος πλησιέστερου γείτονα, δεν κατατάχθηκαν ποτέ στην πρώτη θέση.

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα και μελλοντικές επεκτάσεις

Το πρόβλημα της ταξινόμησης παρουσίαζε από ανέκαθεν αυξημένο ερευνητικό και πρακτικό ενδιαφέρον. Η διαπίστωση αυτή επιβεβαιώνεται από την πληθώρα των πρακτικών εφαρμογών που η προβληματική της ταξινόμησης παρουσιάζει και οι οποίες μπορούν να αντιμετωπιστούν πραγματοποιώντας συγκρίσεις με προκαθορισμένα δείγματα εκμάθησης και δείγματα ελέγχου.

Ο σκοπός της συγκεκριμένης έρευνας ήταν η διερεύνηση της αποτελεσματικότητας διαφόρων μεθόδων ταξινόμησης συναρτήσει των χαρακτηριστικών των εξεταζόμενων δεδομένων και των παραγόντων που τις επηρεάζουν.

Το ελεγχόμενο πείραμα που διεξήχθη με προσομοιωμένα δεδομένα έδειξε ότι οι τεχνικές ταξινόμησης είναι ευαίσθητες στις μεταβολές των χαρακτηριστικών των δεδομένων. Τα ποσοστά σφαλμάτων ταξινόμησης μπορεί να διαφοροποιούνται με την μεταβολή ενός και μόνο παράγοντα. Για παράδειγμα, με την αύξηση του μεγέθους του δείγματος εκμάθησης βελτιώνονται τα ποσοστά σφαλμάτων των μεθόδων ταξινόμησης. Γενικά, περισσότερες από μία μέθοδοι φαίνεται να είναι κατάλληλες με βάση τον τύπο των παραγόντων που χρησιμοποιείται στα δεδομένα. Τα διανύσματα υποστήριξης αποφάσεων και η UTADIS παρουσιάζουν την καλύτερη σχετική αποτελεσματικότητα με βάση τους περισσότερους παράγοντες. Τα αποτελέσματα που επιτεύχθηκαν είναι ιδιαίτερα ενθαρρυντικά και μπορούν να αποτελέσουν τη βάση για την επικέντρωση της περαιτέρω έρευνας στα επιπλέον χαρακτηριστικά των μεθόδων.

Η έρευνα έδειξε ότι καμία μέθοδος δεν υπερτερεί από τις άλλες σε όλους τους παράγοντες που χρησιμοποιήθηκαν. Για το λόγο αυτό, θα πρέπει να υπάρχει ένα σύστημα ταξινόμησης το οποίο θα εφαρμόζει την κατάλληλη μέθοδο στην κάθε περίπτωση. Το σύστημα ταξινόμησης θα πρέπει να επιλέγει ή να συνδυάζει μεθόδους ανάλογα με την παρουσία των διαφόρων παραγόντων που υπάρχουν και επηρεάζουν την αποτελεσματικότητα τους. Τα αποτελέσματα αυτής της έρευνας μπορούν να

χρησιμοποιηθούν για το σχεδιασμό τέτοιων συστημάτων ταξινόμησης. Δηλαδή, με βάση τα αποτελέσματα μπορούν να σχεδιαστούν συστήματα ταξινόμησης, τα οποία να εφαρμόζουν ένα σύνολο μεθόδων με βάση τα χαρακτηριστικά των δεδομένων, για ένα συγκεκριμένο πρόβλημα ταξινόμησης. Το παραπάνω θα αυξήσει την αποτελεσματικότητα και την προσαρμοστικότητα των μεθόδων ταξινόμησης.

Επιπρόσθετα, τα διάφορα χαρακτηριστικά των δεδομένων θα πρέπει να μελετηθούν περαιτέρω. Στη συγκεκριμένη περίπτωση επιλέχθηκαν τέσσερις παράγοντες, οι οποίοι θεωρήθηκαν ότι προσδιορίζουν τη σταθερότητα και τις αδυναμίες της κάθε μεθόδου. Θα πρέπει να γίνουν έρευνες που να εξετάζουν και άλλους παράγοντες καθώς και τις πιθανές αλληλεπιδράσεις μεταξύ τους. Για παράδειγμα στη μορφή διάκρισης των κατηγοριών, ο θόρυβος που επιλέχθηκε ήταν η τυχαία αλλαγή των αποτελεσμάτων στο 10% των περιπτώσεων. Ωστόσο, ο θόρυβος θα μπορούσε να επηρεάζει σε διαφορετικό βαθμό την αποτελεσματικότητα των μεθόδων ταξινόμησης. Απαιτούνται επιπρόσθετα πολύπλοκα πειράματα που να εξετάζουν τις αλληλεπιδράσεις των παραγόντων και την επίδραση των διαφόρων μη ελεγχόμενων δεδομένων (μεροληψιών) στα αποτελέσματα. Για παράδειγμα, η εξέταση της στατιστικής κατανομής των συνεχών δεδομένων περιλάμβανε τρεις περιπτώσεις, την κανονική, την λογαριθμική και τη μίξη κανονικής – λογαριθμικής κατανομής. Στην περίπτωση αυτή θα μπορούσε να εξεταστεί κάποια άλλη κατανομή, όπως η εκθετική. Πρέπει να κατανοηθεί η σταθερότητα και οι αδυναμίες της κάθε μεθόδου, και η πιθανότητα συνένωσης δύο ή περισσοτέρων αλγορίθμων ταξινόμησης για την επίλυση ενός προβλήματος. Ο σκοπός είναι η χρήση της σταθερότητας της μιας μεθόδου για να συμπληρώσει την αδυναμία μιας άλλης. Επιπλέον, απαιτείται περαιτέρω έρευνα και άλλων προσεγγίσεων κυρίως από τον χώρο της τεχνητής νοημοσύνης, όπως μηχανική μάθηση, ασαφής λογική, γενετικοί αλγόριθμοι, και άλλες. Είναι σημαντική η εξέταση αυτών των μεθόδων καθώς και η εξέταση της αποτελεσματικότητας τους στους διάφορους παράγοντες. Επιπρόσθετα, θα πρέπει να εξεταστεί και η περίπτωση των δεδομένων που περιέχουν ακραίες τιμές (outliers), έτσι ώστε να υπάρξει μια ολοκληρωμένη εικόνα για την αποτελεσματικότητα των μεθόδων ταξινόμησης. Τέλος, η αποτελεσματικότητα των μεθόδων θα πρέπει να διερευνηθεί και για τη μείξη ποιοτικών και ποσοτικών δεδομένων, τα οποία απαντώνται συχνά στην πράξη.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Altman, E.I. (1968). Financial Ratios, Discriminant analysis and the prediction of corpotate bankruptcy. Journal of Finance, 23, 589-609.
2. Bauer, E., and Kohani, R., (1999). An empirical comparison of voting classification algorithms: Bagging, boosting and variants, Machine Learning, 36, 105-139.
3. Boser, B.E., Guyon, I.M., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In fifth Annual Workshop on Computational Learning Theory, Pittsburgh, ACM.
4. Breese, J.S., Heckerman, D., and Kadie, C., (1998). Empirical analysis of predictive algorithms for collaborative filtering, Proceedings of the 14th Conference of Uncertainty in Artificial Intelligence, Madison, WI.
5. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., (1984). Classification and Regression Trees. Wadsworth and Brooks, Monterey, Ca.
6. Cacoullos, T. (1966). Estimation of Multivariate Density. Annals of the Institute of Statistical Mathematics, Tokyo, 18:2.
7. Catelani, M. and Forth, A. (2000). Fault diagnosis of electronic analog circuits using a radial basis function network classifier. Measurement, 28/3, 147-158.
8. Clarke, W.R., Lachenbruch, P.A., and Broffitt, B. (1979). How nonnormality affects the quadratic discriminant function. Comm. Statist. – Theory Meth., IT-16;41-46.
9. Cover, T.M., and Hart, P.E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13:21-27.
10. Cox, D.R. (1966). Some procedures associated with the logistic qualitative response curve. In David, F.N., editor, Research papers on statistics: Festschrift for J Neyman, pages 55-77. John Wiley, New York.
11. Day, N.E. and Kerridge, D.F., (1967). A general maximum likelihood discriminant. Biometrics, 23:3 13-324.

12. Diakoulaki, D., Zopounidis, C., Mavrotas, G. and Doumpos, M., (1999). The use of a preference disaggregation method in energy analysis and policy making. Energy – The International Journal, 24/2, 157-166.
13. Dietterich, T.G., Hild, H. and Bakiri, G., (1995). A comparison of ID3 and backpropagation for english text – to – speech mapping. Machine Learning 18, 51-80.
14. Δούμπος, Μ. και Ζοπουνίδης, Κ., (2001). Πολυκριτήριες Τεχνικές Ταξινόμησης. Αθήνα, Κλειδάριθμος.
15. Doumpos, M. and Zopounidis, C., (2002). Multicriteria Decision Aid Classification Methods. Kluwer Academic Plublishers, Dordrecht.
16. Duda, R.O., Hart, P.E. and Stork, D.G., (2001). Pattern Classification (2nd Edition). John Wiley, New York.
17. Dudani, S.A., (1976). The distance – weighted k – nearest neighbor rule. IEEE Transactions on Systems, Man and Cybernetics, 6(4):325-327.
18. Dutka, A., (1995). AMA Handbook of Customer Satisfaction: A Guide to Research, Planning and Implementation. NTC Publishing Group, Illinois.
19. Fisher, R.A., (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179-188.
20. Friedman, J., (1989). Regularized discriminant analysis. J. Amer. Statist. Assoc., 84:165-175.
21. Fukunaga, K. and Narendra, P.M., (1975). A branch and bound algorithm for computing k – nearest neighbor. IEEE Transactions on Computers, (24)7.
22. Gates, G.W., (1972). The reduced nearest neighbor rule. IEEE Transactions on Information Theory, IT-18:431.
23. Gilbert, E.S., (1969). The effect on unequal variance variance covariance matrices on fisher's liner discriminant function. Biometrics, 25:505-515.
24. Hand, D.J. and Batchelor, B.G., (1978). Experiments on the edited condensed nearest neighbor rule. Information Sciences, 14:171-180.
25. Hart, P.E., (1968). The condensed nearest neighbor rule. IEEE Transactions on Information Theory, 14:515-516.
26. Jacquet – Lagreze, E. (1995). An application of the UTA discriminant model for the evaluation of R&D projects, in: P.M. Pardalos, Y. Siskos and C. Zopounidis (eds.), Advances in Multicriteria Analysis, Kluwer Academic Publishers, Dordrecht, 203-211.

27. Jacquet – Lagreze, E. and Siskos, Y., (1982). Assessing a set of additive utility functions for multicriteria decision making: The UTA method. European Journal of Operational Research, 10, 151-164.
28. Kendall, M.G., Stuart, A. and Ord, J.K., (1983). The Advanced Theory of Statistics. Vol. 3, Design and Analysis and Time Series. Chapter 44. Griffin, London, 4th edition.
29. Marks, S. and Dunn, O.J., (1974). Discriminants functions when covariance matrices are unequal. J. Amer. Statist. Assoc., 69:555-559.
30. Martin, D., (1977). Early warning of bank failure: A logit regression approach. Journal of Banking and Finance, 1, 249-276.
31. Meila, M. and Heckerman, D., (2001). An experimental comparison of model – based clustering methods. Machine Learning 42, 9-29.
32. Mirkin, B., (1996). Mathematical Classification and Clustering, Kluwer Academic Publishers, Dordrecht.
33. Parzen, E., (1962). On estimation of a Probability Density Function and Mode. Annals of Mathematical Statistics 33.
34. Resnick, P. and Varian, H., (1997). Recommender Systems, Communication of the ACM 40(3), 56-58.
35. Ripley, B.D., (1996). Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.
36. Roy, B., (1985). Methodologie Multicritere d' Aide a la Decision. Economica, Paris.
37. Siskos, Y., Grigoroudis, E., Zopounidis, C. and Saurais, O. (1998). Measuring customer satisfaction using a survey based preference disaggregation model. Journal of Global Optimization, 12/2, 175-195.
38. Specht, D., (1990). Probabilistic Neural Networks. Neural Networks 3.
39. Tsumoto, S., (1998). Automated extraction of medical expert system rules from clinical databases on rough set theory. Information Sciences, 112, 67-84.
40. Vapnik, V., (1979). Estimation of Dependences Based on Empirical Data. Nauka, Moskow.
41. Yohannes, Y. and Webb, P., (1999). Classification and Regression Trees, CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity. Microcomputers in Policy Research 3, International Food Policy Research Institute.

42. Young, T.H. and Fu, K.S., (1997). Handbook of Pattern Recognition and Image Processing. Handbooks in Science and Technology, Academic Press.
43. Zopounidis, C., (1998). Operational Tools in the Management of Financial Risks. Kluwer Academic Publishers, Dordrecht.
44. Zopounidis, C. and Doumpos, M. (1997). Preference dissagregation methodology in segmentation problems: The case of financial distress, in: C. Zopounidis (ed.). New Operational Approaches for Finanacial Modelling, Springer-Verlag, Berlin-Heidelberg, 417-439.
45. Zopounidis, C. and Doumpos, M. (1998). Developing a multictiteria decision support system for financial classification problems: The FINCLAS system. Optimization Methods and Software, 8, 277-304.
46. Zopounidis, C. and Doumpos, M., (2000). Building additive utilities for multi – group hierarchical discrimination: The M.H.DIS. method. Optimization Methods and Software, 14/3, 219-240.