ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ
ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Διπλωματική εργασία με θέμα:

ΣΥΓΚΡΙΤΙΚΗ ΠΕΙΡΑΜΑΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕΘΟΔΩΝ ΤΑΞΙΝΌΜΗΣΗΣ ΣΕ ΔΙΑΚΡΙΤΑ ΔΕΔΟΜΕΝΑ

ΣΦΑΤΚΙΔΗΣ ΙΩΑΝΝΗΣ

Επιβλέπων Καθηγητής:

Δούμπος Μιχαήλ

Χανιά
Ιανουάριος 2004

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1ο

Εισαγωγή

1.1 Είδη των προβλημάτων λήψης αποφάσεων και η έννοια της ταξινόμησης

Στην καθημερινή πρακτική επιχειρήσεων και οργανισμών, η διαδικασία λήψης αποφάσεων αποτελεί τη βάση όλων των λειτουργικών και αναπτυξιακών τους δραστηριοτήτων. Ένας τρόπος προσέγγισης των προβλημάτων λήψης αποφάσεων, βασίζεται στην προβληματική της ταξινόμησης. Απώτερος στόχος της προσέγγισης αυτής, είναι ο διαχωρισμός του συνόλου των εφικτών λύσεων ή των εναλλακτικών τρόπων δράσης σε κατάλληλα προκαθορισμένες ομοιογενείς κατηγορίες.

Μια από τις πλέον γενικές κατηγοριοποιήσεις των προβλημάτων λήψης αποφάσεων που θα μπορούσαν να πραγματοποιηθούν, συνίσταται στη διάκριση των ακόλουθων δύο μεγάλων κατηγοριών (Σχήμα 1.1):

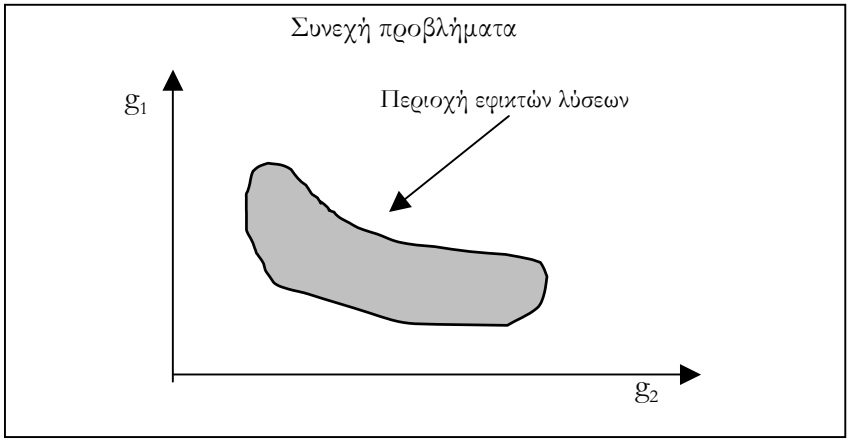
- **Διακριτά προβλήματα:** στην κατηγορία αυτή εντάσσονται προβλήματα τα οποία αφορούν την ανάλυση ενός πεπερασμένου συνόλου συγκεκριμένων εναλλακτικών δραστηριοτήτων. Ως εναλλακτικές δραστηριότητες θεωρούνται τα αντικείμενα τα οποία αναλύονται και κάθε εναλλακτική δραστηριότητα περιγράφεται από ένα σύνολο χαρακτηριστικών.

- **Συνεχή προβλήματα:** στην κατηγορία αυτή εντάσσονται προβλήματα, στα οποία δεν είναι δυνατός ο άμεσος και σαφής καθορισμός ενός πεπερασμένου συνόλου εναλλακτικών δραστηριοτήτων. Αντίθετα, είναι δυνατή η οριοθέτηση του χώρου μέσα στον οποίο βρίσκονται οι δυνατές λύσεις του προβλήματος και κάθε μια από τις οποίες ουσιαστικά αντιστοιχεί σε έναν εναλλακτικό τρόπο δράσης.

Διακριτά προβλήματα

m εναλλακτικές δραστηριότητες
n κριτήρια/χαρακτηριστικά

		Κριτήρια			
		g_1	g_2	….	g_n
Εναλλακτικές Δραστηριότητες	x_1	g_{11}	g_{12}	….	g_{1n}
	x_2	g_{21}	g_{22}	….	g_{2n}
	.	.	.	….	.
	.	.	.	….	.
	.	.	.	….	.
	x_m	g_{m1}	g_{m2}	….	g_{mn}



Σχήμα 1.1: Συνεχή και διακριτά προβλήματα ταξινόμησης

Τα προβλήματα της πρώτης κατηγορίας (διακριτά) μπορούν να ενταχθούν σε επιμέρους κατηγορίες ανάλογα με το επιδιωκόμενο αποτέλεσμα της ανάλυσης. Ο Roy (1985) θεώρησε τις ακόλουθες κατηγορίες διακριτών προβλημάτων:

α) Επιλογή (choice) της καλύτερης εναλλακτικής.

β) Κατάταξη (ranking) των εξεταζόμενων εναλλακτικών από τις καλύτερες στις χειρότερες βάσει των χαρακτηριστικών τους.

γ) Ταξινόμηση (sorting, discrimination ή classification) των εναλλακτικών δραστηριοτήτων σε προκαθορισμένες κατηγορίες.

δ) Περιγραφή των εναλλακτικών δραστηριοτήτων για τον εντοπισμό των βασικών τους ιδιοτήτων (description).

Τα προβλήματα της επιλογής και της κατάταξης βασίζονται στην πραγματοποίηση σχετικών συγκρίσεων (relative comparisons) ανάμεσα στις εξεταζόμενες εναλλακτικές δραστηριότητες. 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.

Αντίθετα, το πρόβλημα της ταξινόμησης βασίζεται στην πραγματοποίηση απόλυτων συγκρίσεων (absolute comparisons). Κάθε εναλλακτική δραστηριότητα εντάσσεται σε μια εκ των προκαθορισμένων κατηγοριών βάσει ενός συγκεκριμένου κανόνα, ο οποίος συνήθως αναφέρεται στη σύγκριση με συγκεκριμένα πρότυπα τα οποία διαχωρίζουν τις κατηγορίες. Η εφαρμογή του κανόνα ταξινόμησης, και κατά συνέπεια το αποτέλεσμα της αξιολόγησης, δεν επηρεάζεται από το σύνολο των εξεταζόμενων δραστηριοτήτων.

Όπως προαναφέρφηκε, το πρόβλημα της ταξινόμησης αναφέρεται στην ένταξη ορισμένων προκαθορισμένων εναλλακτικών δραστηριοτήτων ή αντικειμένων σε κατηγορίες. Ένας περισσότερο αυστηρός ορισμός δόθηκε από τον Mirkin (1998), ο οποίος όρισε την ταξινόμηση ως εξής:

«Ταξινόμηση είναι η ρεαλιστική ή ιδεατή τοποθέτηση μαζί παρόμοιων αντικειμένων και ο διαχωρισμός των αντικειμένων τα οποία διαφέρουν, με απώτερο σκοπό:

(α) τη διαμόρφωση, οργάνωση και διατήρηση της γνώσης,
(β) την ανάλυση της δομής του φαινομένου που εξετάζεται,
(γ) τη συσχέτιση των διαφόρων πλευρών του υπό εξέταση φαινομένου.»

Σε αυτό το σημείο είναι χρήσιμο να τονιστεί και η διαφορά του προβλήματος της ταξινόμησης από το παρόμοιο πρόβλημα της ομαδοποίησης (clustering). Η κύρια διαφορά των δύο αυτών προβλημάτων συνίσταται στο γεγονός ότι ενώ στην

προβληματική της ταξινόμησης επιδιώκεται η τοποθέτηση ενός συνόλου εναλλακτικών τρόπων δράσης-αντικειμένων σε προκαθορισμένες ομοιογενείς κατηγορίες (groups, classes), στην ομαδοποίηση επιδιώκεται η οργάνωση ενός συνόλου εναλλακτικών τρόπων δράσης-αντικειμένων σε ομοιογενείς ομάδες (clusters), χωρίς όμως να είναι γνωστός ο αριθμός των ομάδων και η έννοια τους.

1.2 Πεδία εφαρμογής του προβλήματος της ταξινόμησης

Το πρόβλημα της ταξινόμησης δεν βρίσκει εφαρμογή μόνο στην επιστημονική έρευνα λόγω της πολυπλοκότητας που παρουσιάζει αλλά και σε πολλούς άλλους τομείς που παρουσιάζονται παρακάτω.

1. *Στην χρηματοοικονομική διοίκηση* και πιο συγκεκριμένα όσον αφορά την πρόβλεψη της πτώχευσης επιχειρήσεων, εκτίμηση του πιστωτικού κινδύνου επιχειρήσεων και καταναλωτών, επιλογή και διαχείριση χαρτοφυλακίων επενδύσεων και χρεογράφων, αξιολόγηση δανειοληπτικής ικανότητας χωρών, στις εξαγορές και συγχωνεύσεις επιχειρήσεων και στον κίνδυνο χώρας (Zopounidis (1998), Doumpos και Zopounidis (1998)).
2. *Ιατρική:* Πραγματοποίηση ιατρικών διαγνώσεων ταξινομώντας τους ασθενείς σε κατηγορίες με βάση τα συμπτώματα που παρουσιάζουν (Tsumoto (1998), Belacel(2000)).
3. *Αναγνώριση προτύπων:* Διερεύνηση των χαρακτηριστικών φυσικών προσώπων ή αντικειμένων και ταξινόμησή τους σε ανάλογες κατηγορίες. Χαρακτηριστικά παραδείγματα της αναγνώρισης βασικών ανθρώπινων χαρακτηριστικών είναι η αναγνώριση φωνής, δακτυλικών αποτυπωμάτων και οι εφαρμογές τους στην ασφάλεια καίριων συστημάτων (Ripley (1996), Young και Fu (1997), Nieddu και Patrizi (2000)).
4. *Διαχείριση ανθρωπίνου δυναμικού:* Αξιολόγηση του ανθρώπινου δυναμικού βάσει των προσόντων του, με απώτερο σκοπό τον προσδιορισμό της κατάλληλης θέσης (Rulon et al. (1967) και Gochet et al. (1997)).

5. *Διαχείριση Τεχνικών Συστημάτων και Τεχνική Διάγνωση:* Παρακολούθηση της λειτουργίας πολύπλοκων συστημάτων παραγωγής για την έγκαιρη διάγνωση πιθανών βλαβών (Catelani και Ford (2000), Shen et al. (2000)).

6. *Μάρκετινγκ:* Μέτρηση της ικανοποίησης πελατών, μελέτη των επιμέρους χαρακτηριστικών διαφορετικών κατηγοριών καταναλωτών, ανάπτυξη κατάλληλων πολιτικών για την διείσδυση προϊόντων στην αγορά, κ.ά. (Dutka (1995), Siskos et al. (1998)).

7. *Περιβαλλοντική και ενεργειακή διαχείριση, οικολογία:* Ανάλυση και έγκαιρη διάγνωση των περιβαλλοντικών επιπτώσεων διαφόρων ενεργειακών πολιτικών, διερεύνηση της αποτελεσματικότητας ενεργειακών πολιτικών σε κρατικό επίπεδο (Diakoulaki et al., 1999).

1.3 Σκοπός και δομή της εργασίας

Η επιτυχής πρακτική εφαρμογή οποιασδήποτε μεθοδολογικής προσέγγισης για την αντιμετώπιση ενός προβλήματος λήψης αποφάσεων είναι συνάρτηση όχι μόνο της ποιότητας και της ποσότητας της παρεχόμενης υποστήριξης, αλλά και της αξιοπιστίας της χρησιμοποιούμενης μεθοδολογικής προσέγγισης. Ο σκοπός της έρευνας που παρουσιάζεται σε αυτή την εργασία είναι η σύγκριση διαφόρων τεχνικών ταξινόμησης με συγκεκριμένα χαρακτηριστικά και η κατάταξη τους ως προς την αποτελεσματικότητα-αξιοπιστία τους.

Η εργασία οργανώνεται σε τέσσερα κεφάλαια καλύπτοντας τα εξής θέματα:

Στο δεύτερο κεφάλαιο παρουσιάζονται οι βασικές έννοιες του προβλήματος της ταξινόμησης. Γίνεται αναφορά στις μεθοδολογικές προσεγγίσεις που έχουν προταθεί για την ανάπτυξη υποδειγμάτων ταξινόμησης τόσο με τις «παραδοσιακές» στατιστικές προσεγγίσεις, όσο και με τα νέα τεχνολογικά εργαλεία τα οποία έχουν αναπτυχθεί κατά τη διάρκεια των τελευταίων δύο δεκαετιών.

Στο τρίτο κεφάλαιο πραγματοποιείται μια συγκριτική έρευνα διαφόρων προσεγγίσεων ταξινόμησης, με σκοπό την κατάταξή τους ως προς την αποτελεσματικότητά τους. Παρουσιάζεται μια γενική περιγραφή των εξεταζόμενων τεχνικών ταξινόμησης και των παραγόντων που επιδρούν στην αποτελεσματικότητά τους, όπως επίσης και η διαδικασία υλοποίησης της συγκριτικής έρευνας και η ανάλυση των αποτελεσμάτων που προκύπτουν.

Τέλος, στο τέταρτο κεφαλαίο παρουσιάζονται τα βασικά συμπεράσματα που επιτεύχθηκαν από την έρευνα που πραγματοποιήθηκε και προτείνονται μελλοντικές ερευνητικές κατευθύνσεις, οι οποίες θα συμβάλλουν στην καλύτερη αντιμετώπιση του προβλήματος της ταξινόμησης, αλλά και στην περαιτέρω διερεύνηση των ιδιαιτεροτήτων, ομοιοτήτων και διαφορών που χαρακτηρίζουν τις εξεταζόμενες μεθόδους ταξινόμησης.

ΚΕΦΑΛΑΙΟ 2ο

Μεθοδολογίες ανάπτυξης υποδειγμάτων ταξινόμησης

2.1 Το πρόβλημα της ταξινόμησης

Όπως έχει προαναφερθεί στο προηγούμενο κεφάλαιο το πρόβλημα της ταξινόμησης βασίζεται στην πραγματοποίηση απόλυτων συγκρίσεων με στόχο την ένταξη ορισμένων προκαθορισμένων εναλλακτικών δραστηριοτήτων ή αντικειμένων σε κατηγορίες.

Οι συνηθέστεροι όροι που χρησιμοποιούνται στην αγγλική βιβλιογραφία για την αναφορά στο πρόβλημα της ταξινόμησης είναι:

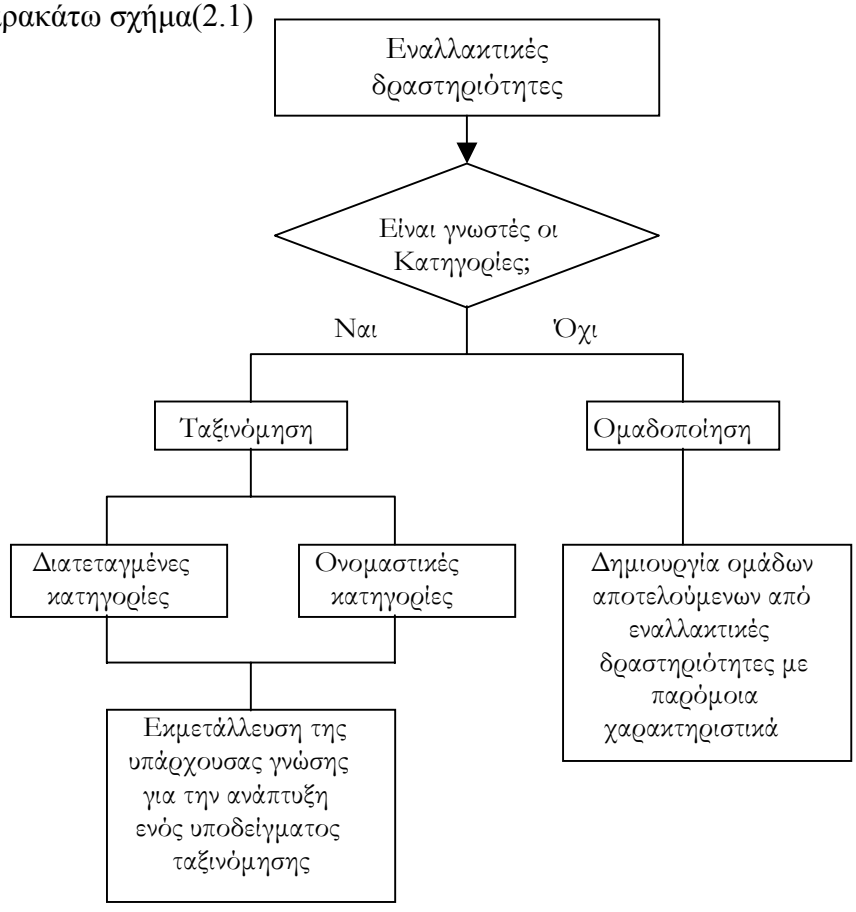
- *Discrimination (διάκριση)*
- *Classification (ταξινόμηση)*
- *Sorting (διατεταγμένη ταξινόμηση)*

Οι όροι *discrimination* και *classification* αναφέρονται σε προβλήματα όπου οι κατηγορίες στις οποίες θα ενταχθούν οι εξεταζόμενες εναλλακτικές δραστηριότητες, ορίζονται κατά ονομαστικό τρόπο (nominal categories). Αυτό σημαίνει ότι οι εναλλακτικές δραστηριότητες που εντάσσονται σε μια συγκεκριμένη κατηγορία παρουσιάζουν διαφορετικά χαρακτηριστικά σε σχέση με τις εναλλακτικές δραστηριότητες των άλλων κατηγοριών, χωρίς όμως αυτό να συνεπάγεται ότι οι

εναλλακτικές δραστηριότητες της μιας κατηγορίας είναι κατά οποιονδήποτε τρόπο καλύτερες ή χειρότερες από τις εναλλακτικές δραστηριότητες των άλλων κατηγοριών. Το πλέον διαδεδομένο πρόβλημα αυτής της μορφής είναι το πρόβλημα αναγνώρισης προτύπων (pattern recognition) με πληθώρα πρακτικών εφαρμογών σε εφαρμογές αναγνώρισης γραμμάτων, φυσικών προσώπων, αντικειμένων, κλπ.

Αντίθετα, ο όρος sorting αναφέρεται σε προβλήματα όπου οι οριζόμενες κατηγορίες, στις οποίες θα πραγματοποιηθεί η ταξινόμηση των εναλλακτικών δραστηριοτήτων, είναι διατεταγμένες από τις καλύτερες στις χειρότερες. Χαρακτηριστικό είναι το παράδειγμα της αξιολόγησης των επιδόσεων και της βιωσιμότητας των επιχειρήσεων, κατατάσσοντας αυτές σε κατηγορίες όπως για παράδειγμα στην κατηγορία των υγιών και δυναμικών, στην κατηγορία των μέτριων και στην κατηγορία των προβληματικών εταιριών. Μια τέτοια ταξινόμηση προφανώς θεωρεί ότι οποιαδήποτε επιχείρηση που ταξινομείται ως υγιής είναι καλύτερη από οποιαδήποτε επιχείρηση που ταξινομείται ως μέτρια ή προβληματική.

Όλες οι παραπάνω παρατηρήσεις για το πρόβλημα της ταξινόμησης συνοψίζονται στο παρακάτω σχήμα(2.1)



Σχήμα 2.1:Τα προβλήματα της ταξινόμησης και της ομαδοποίησης

Υπενθυμίζεται ότι η ομαδοποίηση είναι η οργάνωση ενός συνόλου εναλλακτικών τρόπων δράσης-αντικειμένων σε ομοιογενείς ομάδες, χωρίς όμως να είναι γνωστός ο αριθμός των ομάδων και η έννοια τους ενώ ταξινόμηση είναι η τοποθέτηση ενός συνόλου εναλλακτικών τρόπων δράσης-αντικειμένων σε προκαθορισμένες ομοιογενείς κατηγορίες.

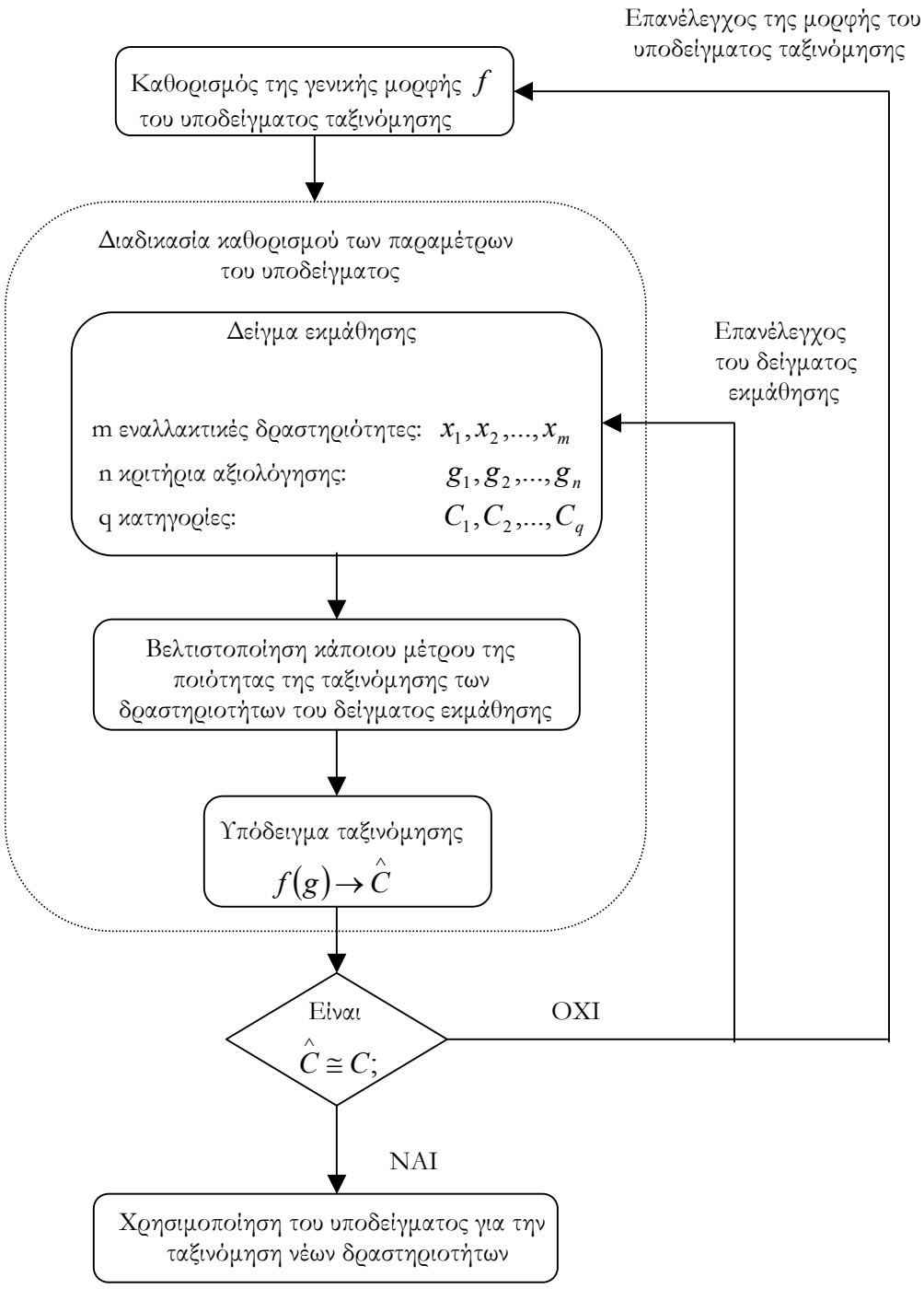
Κλείνοντας την ενότητα αυτή ανακεφαλαιώνουμε τις βασικές έννοιες του προβλήματος της ταξινόμησης.

- Εναλλακτικοί (alternative) τρόποι δράσης: Τα αντικείμενα που εξετάζονται στο πρόβλημα. Π.χ. επιχειρήσεις, χώρες, μετοχές
- Χαρακτηριστικά (attributes): Παράγοντες που περιγράφουν επαρκώς τους εναλλακτικούς τρόπους δράσεις.
- Κριτήρια αξιολόγησης (evaluation criteria): Χαρακτηριστικά που ενσωματώνουν την έννοια της προτίμησης.
- Μοντέλο-Υπόδειγμα ταξινόμησης: Η σύνθεση των χαρακτηριστικών και των κριτηρίων για την ταξινόμηση των εναλλακτικών.

2.2 Ανάπτυξη υποδειγμάτων ταξινόμησης

Οι μεθοδολογικές προσεγγίσεις για την ανάπτυξη υποδειγμάτων ταξινόμησης ακολουθούν την γενική φιλοσοφία της παλινδρόμησης, προσπαθώντας να αξιοποιήσουν τη διαθέσιμη γνώση και πληροφορία που απορρέει από το γεγονός ότι οι κατηγορίες είναι προκαθορισμένες. Το γενικό περίγραμμα της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης φαίνεται στο σχήμα 2.2.

Στην κλασσική στατιστική παλινδρόμηση, στόχος είναι ο εντοπισμός της συναρτησιακής σχέσης που συνδέει μια εξαρτημένη μεταβλητή Y με ένα διάνυσμα μεταβλητών X βάσει της ανάλυσης ενός συνόλου δεδομένων παρατηρήσεων (Y,X). Κατά ανάλογο τρόπο αντιμετωπίζεται και το πρόβλημα της ταξινόμησης. Η ιδιαιτερότητα όμως του προβλήματος της ταξινόμησης σε σχέση με αυτό της στατιστικής παλινδρόμησης έγκειται στο γεγονός ότι η εξαρτημένη μεταβλητή δεν είναι συνεχής, αλλά αφορά ένα περιορισμένο σύνολο διακριτών επιπέδων καθένα από τα οποία αντιστοιχεί σε μια κατηγορία.



Σχήμα 2.2:Γενικό περίγραμμα της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης

Σε αυτό το σημείο θα πρέπει να εξηγηθούν κάποιοι συμβολισμοί ώστε να γίνει κατανοητή η ανάπτυξη των υποδειγμάτων ταξινόμησης. Έτσι εφεξής ως C θα συμβολίζεται η εξαρτημένη μεταβλητή που υποδηλώνει την ταξινόμηση των εναλλακτικών δραστηριοτήτων σε ένα σύνολο κατηγοριών q, ενώ ως C_1, C_2, …. ,C_q τα επιμέρους επίπεδα (κατηγορίες) που λαμβάνει το C. Το δείγμα των παρατηρήσεων που χρησιμοποιείται για την ανάπτυξη των υποδειγμάτων ταξινόμησης ονομάζεται δείγμα

εκμάθησης (training sample) ή σύνολο αναφοράς (reference set) και περιλαμβάνει ζεύγη της μορφής (C,X).

Η επίλυση του προβλήματος της ταξινόμησης συνίσταται στην ανάπτυξη ενός υποδείγματος της μορφής $f(X) \to C^*$ ώστε να ελαχιστοποιηθεί ένα μέτρο των διαφορών που εντοπίζονται μεταξύ της εκτιμώμενης ταξινόμησης C* και της δεδομένης ταξινόμησης C. Μετά την ολοκλήρωση της διαδικασίας ανάπτυξης του υποδείγματος ταξινόμησης, και εφόσον κριθεί ικανοποιητικό όσον αφορά την εκτίμηση των παραμέτρων του και την αποτελεσματικότητα του στο δείγμα εκμάθησης μπορεί πλέον αυτό να χρησιμοποιηθεί για την ταξινόμηση οποιονδήποτε άλλων νέων εναλλακτικών δραστηριοτήτων, οι οποίες δεν συμπεριλαμβάνονται στο δείγμα εκμάθησης. Η χρησιμότητα της παραπάνω διαδικασίας βασίζεται στην εκμετάλλευση της υπάρχουσας γνώσης από το δείγμα εκμάθησης, με σκοπό την μοντελοποίηση και αναπαράστασή της σε ένα υπόδειγμα ταξινόμησης, το οποίο θα διαθέτει την απαραίτητη ικανότητα γενίκευσης.

Τα αναπτυσσόμενα υποδείγματα ταξινόμησης είναι μια συνάρτηση, η οποία συνδυάζει όλα τα επιμέρους χαρακτηριστικά των εναλλακτικών δραστηριοτήτων σε έναν ολικό ποσοτικό δείκτη βάσει του οποίου λαμβάνονται οι αποφάσεις για την ταξινόμηση των εναλλακτικών δραστηριοτήτων. Ο δείκτης αυτός συνήθως αναπαριστά την πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα σε μια κατηγορία ή την αξία της κάθε εναλλακτικής σε μια κανονικοποιημένη ή μη κανονικοποιημένη κλίμακα.

2.3 Μεθοδολογικές προσεγγίσεις για την ανάπτυξη υποδειγμάτων ταξινόμησης

Η αυξημένη σημαντικότητα του προβλήματος της ταξινόμησης τόσο σε πρακτικό όσο και σε ερευνητικό επίπεδο, έχει ελκύσει το ενδιαφέρον πολλών ερευνητών από διαφορετικούς επιστημονικούς χώρους. Βέβαια, η ευρύτητα του προβλήματος της ταξινόμησης, καθιστά ιδιαίτερα δύσκολη την πλήρη κάλυψη όλων των μεθοδολογικών προσεγγίσεων που έχουν κατά καιρούς αναπτυχθεί. Για το λόγο αυτό η παρουσίαση των βασικότερων μεθοδολογικών προσεγγίσεων οι οποίες έχουν προταθεί για την ανάπτυξη υποδειγμάτων ταξινόμησης επικεντρώνεται στις ευρύτερα διαδεδομένες

προσεγγίσεις, βάσει των ερευνητικών και πρακτικών τους εφαρμογών. Οι προσεγγίσεις αυτές διακρίνονται σε δύο βασικές κατηγορίες:

α) Στις στατιστικές και οικονομετρικές προσεγγίσεις

β) Στις μη παραμετρικές προσεγγίσεις.

2.3.1 Στατιστικές και οικονομετρικές προσεγγίσεις

Οι στατιστικές και οικονομετρικές προσεγγίσεις, αποτελούν τον «παραδοσιακό» τρόπο αντιμετώπισης του προβλήματος της ταξινόμησης. Οι σχετικές τεχνικές που έχουν αναπτυχθεί περιλαμβάνουν τόσο μονοδιάστατες όσο και πολυδιάστατες στατιστικές μεθόδους. Οι πρώτες αναφέρονται στην ανάπτυξη και εφαρμογή στατιστικών ελέγχων περιγραφικού χαρακτήρα. Όσο για τις πολυδιάστατες στατιστικές μεθόδους οι βάσεις τέθηκαν ουσιαστικά από τον Fisher το 1936. Στην ερευνητική του εργασία ο Fisher ανέπτυξε την πρώτη πολυδιάστατη μέθοδο ταξινόμησης, τη γραμμική διακριτική ανάλυση (linear discriminant analysis), η οποία για πολλές δεκαετίες υπήρξε η πλέον διαδεδομένη μεθοδολογία για την ανάπτυξη υποδειγμάτων ταξινόμησης. Αργότερα ο Smith (1947) επέκτεινε την εργασία του Fisher αναπτύσσοντας την τετραγωνική διακριτική ανάλυση (quadratic discriminant analysis). Τόσο η γραμμική διακριτική ανάλυση (LDA) όσο και η τετραγωνική διακριτική ανάλυση (QDA) αναπτύσσονται περαιτέρω σε παρακάτω κεφάλαιο. Μια άλλη πολυδιάστατη μέθοδος ταξινόμησης αυτής της κατηγορίας είναι το λογιστικό (η οποία επίσης αναπτύσσεται περαιτέρω σε παρακάτω κεφάλαιο) και το κανονικό υπόδειγμα πιθανότητας(logit και probit analysis, αντίστοιχα).

2.3.2 Μη παραμετρικές προσεγγίσεις

Οι μη παραμετρικές προσεγγίσεις έχουν προταθεί κατά τις τελευταίες δύο δεκαετίες ως καινοτόμες και αποτελεσματικές τεχνικές ανάπτυξης υποδειγμάτων ταξινόμησης. Οι προσεγγίσεις αυτές δεν βασίζονται σε στατιστικές υποθέσεις και συνεπώς αναμένεται ότι μπορούν να προσαρμόζονται ικανοποιητικά, ανάλογα με τα χρησιμοποιούμενα σύνολα δεδομένων, είτε ως γραμμικά υποδείγματα ταξινόμησης είτε ως μη γραμμικά υποδείγματα. Συνεπώς, τέτοιου είδους προσεγγίσεις παρέχουν αυξημένη ευελιξία στον αποφασίζοντα, απαλλάσσοντάς τον από την ανάγκη εντοπισμού και ανάλυσης των στατιστικών ιδιοτήτων των δεδομένων που αφορούν το εξεταζόμενο πρόβλημα. Οι

σημαντικότερες από τις μη παραμετρικές προσεγγίσεις είναι τα νευρωνικά δίκτυα (neural networks) τα οποία παρουσιάζονται περαιτέρω σε επόμενο κεφάλαιο, η μηχανική μάθηση (machine learning), η θεωρία της ασαφούς λογικής (fuzzy set theory) η οποία αναπτύχθηκε από τον Zadeh (1965) και η θεωρία των προσεγγιστικών συνόλων (rough set theory) η οποία αναπτύχθηκε από τον Pawlak (1982).

ΚΕΦΑΛΑΙΟ 3ο

Συγκριτική έρευνα υποδειγμάτων ταξινόμησης

3.1 Σκοπός της έρευνας

Είναι προφανές, ότι ο κάθε αποφασίζων, ενός προβλήματος λήψης αποφάσεων, προσβλέπει στην χρήση του πλέον αποτελεσματικού τρόπου αντιμετώπισης του εξεταζομένου προβλήματος μεταξύ των εναλλακτικών εργαλείων που διαθέτει. Συνεπώς, η επιτυχής πρακτική εφαρμογή οποιασδήποτε μεθοδολογικής προσέγγισης για την αντιμετώπιση ενός προβλήματος λήψης αποφάσεων είναι συνάρτηση όχι μόνο της ποιότητας και της ποσότητας της παρεχόμενης υποστήριξης, αλλά και της αξιοπιστίας της χρησιμοποιούμενης μεθοδολογικής προσέγγισης. Σκοπός της έρευνας που παρουσιάζεται στο κεφαλαίο αυτό, είναι η σύγκριση διαφόρων τεχνικών ταξινόμησης με συγκεκριμένα χαρακτηριστικά, τα οποία παρουσιάζονται στη συνέχεια του κεφαλαίου, και η κατάταξη τους ως προς την αποτελεσματικότητα-αξιοπιστία τους.

3.2 Εξεταζόμενες τεχνικές ταξινόμησης

Σε κάθε συγκριτική ανάλυση μεταξύ εναλλακτικών μεθοδολογικών προσεγγίσεων ενός προβλήματος, οι εξεταζόμενες προσεγγίσεις θα πρέπει να επιλέγονται έτσι ώστε να είναι αντιπροσωπευτικές όλων των διαθέσιμων επιλογών και ευρείας αποδοχής στον χώρο του προβλήματος. Στην παρούσα ενότητα πραγματοποιείται μια σύγκριση πέντε τεχνικών ταξινόμησης, με συγκεκριμένα χαρακτηριστικά, τα οποία αναφέρονται σε επόμενη ενότητα, ως προς την αποτελεσματικότητα τους υπό διαφορετικές συνθήκες, όπως π.χ. διαφορετικό επίπεδο θορύβου ή με διαφορετική συσχέτιση των εξεταζομένων παραγόντων κ.α. Οι προς σύγκριση τεχνικές ταξινόμησης που θα απασχολήσουν τη συγκεκριμένη έρευνα είναι οι παρακάτω:

1. Η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis).
2. Η τετραγωνική διακριτική ανάλυση (Quadratic Discriminant Analysis).
3. Το λογιστικό υπόδειγμα πιθανότητας (Logistic Regression).
4. Οι μηχανές διανύσματος υποστήριξης (Support vector machines).
5. Τα πιθανοτικά νευρωνικά δίκτυα (Probabilistic Neural Networks).

Στις ενότητες που ακολουθούν παρουσιάζονται τα βασικά χαρακτηριστικά των προαναφερθέντων τεχνικών ταξινόμησης.

3.2.1 Γραμμική διακριτική ανάλυση

Η γραμμική διακριτική ανάλυση (LDA) αποτέλεσε την πρώτη πολυδιάστατη μέθοδο ταξινόμησης και αναπτύχθηκε αρχικά από τον Fisher (1936). Σκοπός της μεθόδου είναι η ανάπτυξη μιας σειράς διακριτικών συναρτήσεων οι οποίες μεγιστοποιούν τη διακύμανση μεταξύ των κατηγοριών σε σχέση με την διακύμανση εντός των κατηγοριών, χρησιμοποιώντας ως δείγμα εκμάθησης ένα σύνολο εναλλακτικών δραστηριοτήτων η ταξινόμηση των οποίων είναι γνωστή. Στην γενική περίπτωση όπου η ταξινόμηση πραγματοποιείται σε q κατηγορίες, αναπτύσσονται q-1 γραμμικές συναρτήσεις της μορφής:

$$\mathbf{Z_{kl}=a_{kl}+b_{kl1}g_1+b_{kl2}g_2+\ldots.+b_{kln}g_n} \qquad (3.1)$$

όπου $g1$, $g2$, …, gn είναι τα χαρακτηριστικά που περιγράφουν τις εναλλακτικές δραστηριότητες $x_1,x_2,\ldots x_m$, a_{kl} είναι μια σταθερά, και $b_{kl1,}$ $b_{kl2,\ldots}$ b_{kln} είναι οι συντελεστές των χαρακτηριστικών στη διακριτική συνάρτηση. Οι δείκτες k και l

αναφέρονται σε ένα ζεύγος κατηγοριών οι οποίες συμβολίζονται αντίστοιχα ως C_k και C_l.

Ο υπολογισμός του σταθερού όρου a_{kl} και του διανύσματος b_{kl} βασίζεται στην υπόθεση ότι οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών είναι ίσοι και ότι οι επιδόσεις των εναλλακτικών δραστηριοτήτων στα εξεταζόμενα χαρακτηριστικά ακολουθούν την πολυμεταβλητή κανονική κατανομή. Βάσει των υποθέσεων αυτών οι υπολογισμοί των a_{kl} και b_{kl} πραγματοποιούνται ως εξής:

$$\mathbf{a_{kl}= -[\mu_k+\mu_l]^1 b_{kl}/2} \tag{3.2}$$

$$\mathbf{b_{kl}= \Sigma^{-1} [\mu_k-\mu_l]} \tag{3.3}$$

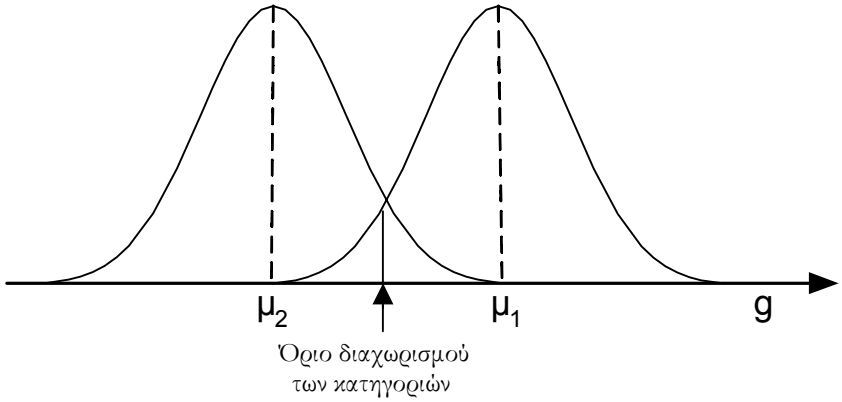
Όπου μ_k είναι το διάνυσμα των μέσων τιμών των χαρακτηριστικών για τις εναλλακτικές δραστηριότητες της κατηγορίας C_k και C_l. Ως $\boldsymbol{\Sigma}$ συμβολίζεται ο πίνακας διακύμανσης-συνδιακύμανσης μεταξύ των κατηγοριών (within groups variance-covariance matrix). Ως m συμβολίζεται το πλήθος των εναλλακτικών δραστηριοτήτων του δείγματος εκμάθησης, ως $\mathbf{g}_j=(g_{j1}, g_{j2}, \ldots, g_{jn})$ το διάνυσμα της περιγραφής της εναλλακτικής δραστηριότητας x_j βάσει των χαρακτηριστικών **g**, και ως q το πλήθος των κατηγοριών.

Η ταξινόμηση κάθε εναλλακτικής δραστηριότητας x_j σε μια εκ των προκαθορισμένων κατηγοριών πραγματοποιείται βάσει των σκορ διάκρισης της δραστηριότητας όπως αυτά υπολογίζονται από την κάθε συνάρτηση. Πιο συγκεκριμένα, μια εναλλακτική δραστηριότητα x_j θα ταξινομηθεί στην κατηγορία C_k εάν για όλες τις άλλες κατηγορίες C_l ισχύει:

$$Z_{kl}\left(g_j\right) \geq \ln \frac{K(k \mid l)\pi_\lambda}{K(l \mid k)\pi_k} \tag{3.4}$$

Ως $Z_{kl}\left(g_j\right)$ συμβολίζεται το σκορ διάκρισης (discriminant score) που αποδίδεται στην εναλλακτική δραστηριότητα x_j από τη διακριτική συνάρτηση Z_{kl}, ως K(k|l) συμβολίζεται το κόστος της εσφαλμένης ταξινόμησης μιας εναλλακτικής δραστηριότητας, η οποία ενώ ανήκει στην κατηγορία C_l εντάσσεται στην κατηγορία C_k, ενώ τέλος ως π_k συμβολίζεται η εκ των προτέρων πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα στην κατηγορία C_k . Θεωρώντας τα κόστη εσφαλμένων ταξινομήσεων ίσα όπως και τις εκ των προτέρων πιθανότητες, αυτός ο γραμμικός

κανόνας ταξινόμησης μπορεί να αποδοθεί γραφικά μέσω του Σχήματος 3.1, για την απλή περίπτωση της διάκρισης μεταξύ δύο κατηγοριών.



Σχήμα 3.1:Σχηματική απεικόνιση του κανόνα ταξινόμησης της γραμμικής διακριτικής ανάλυσης

3.2.2 Τετραγωνική διακριτική ανάλυση

Η τετραγωνική διακριτική ανάλυση αναπτύχθηκε από τον Smith (1947). Χρησιμοποιείται, αντί της γραμμικής διακριτικής ανάλυσης στην περίπτωση όπου οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών δεν είναι ίσοι. Η μορφή της τετραγωνικής συνάρτησης που αναπτύσσεται για κάθε ζεύγος κατηγοριών C_k και C_l είναι η ακόλουθη:

$$Z_{kl} = a_{kl} + \sum_{i=1}^{n} b_{kli} g_i + \sum_{i=1}^{n} \sum_{h=1}^{n} c_{klih} g_i g_h \tag{3.5}$$

Οι συντελεστές και ο σταθερός όρος της παραπάνω συνάρτησης δίνονται από τους παρακάτω τύπους:

$$b_{kl} = -2[\mu_k' \Sigma_k^{-1} - \mu_l' \Sigma_l^{-1}] \tag{3.6}$$

$$c_{kl} = \Sigma_k^{-1} - \Sigma_l^{-1} \tag{3.7}$$

$$a_{kl} = \mu_k' \Sigma_k^{-1} \mu_k - \mu_l' \Sigma_l^{-1} \mu_l - \ln |\Sigma_l \Sigma_k^{-1}| \tag{3.8}$$

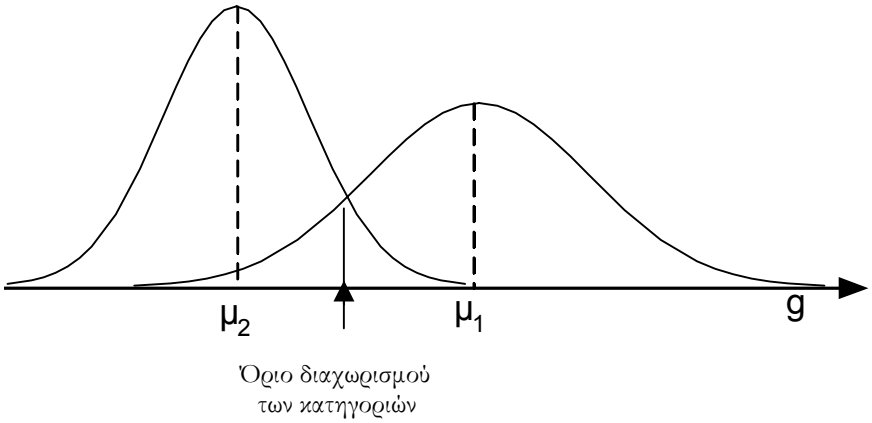
Οι πίνακες διακύμανσης-συνδιακύμανσης Σ_k και Σ_l υπολογίζονται από τη σχέση:

$$\Sigma_k = \frac{\sum_{\forall x_j \in c_k} [g_j - \mu_k][g_j - \mu_k]'}{m_k - 1} \tag{3.9}$$

Ο κανόνας ταξινόμησης της τετραγωνική διακριτική ανάλυση διαμορφώνεται ως εξής (Σχήμα 3.2): η εναλλακτική x_j θα ενταχθεί στην κατηγορία C_k εάν και μόνο εάν για όλες τις άλλες κατηγορίες C_l ισχύει:

$$Z_{kl}(g_j) \geq -2\ln\frac{K(k\mid l)\pi_l}{K(l\mid k)\pi_k} \tag{3.10}$$

όπου $Z_{kl}(g_j)$ το σκορ διάκρισης μιας εναλλακτικής δραστηριότητας x_j.



Σχήμα 3.2: Σχηματική απεικόνιση του κανόνα ταξινόμησης της τετραγωνικής διακριτικής ανάλυσης

3.2.3 Το λογιστικό υπόδειγμα πιθανότητας

Το λογιστικό υπόδειγμα πιθανότητας γνώρισε ιδιαίτερη διάδοση μετά τη δεκαετία του 1970 και τις εργασίες του βραβευμένου με Νόμπελ Οικονομίας, Daniel McFadden (1974, 1980) για την ανάπτυξη της θεωρίας της διακριτής επιλογής (discrete choice). Η θεωρία αυτή αποτέλεσε την αναγκαία θεωρητική βάση για την κατανόηση των βασικών εννοιών της εν λόγω προσέγγισης.

Το λογιστικό υπόδειγμα πιθανότητας βασίζεται στην ανάπτυξη μιας μη γραμμικής συνάρτησης βάσει της οποίας υπολογίζεται η πιθανότητα των εναλλακτικών δραστηριοτήτων να ανήκουν σε κάθε μια από τις υπό εξέταση κατηγορίες. Έτσι, στην περίπτωση της ταξινόμησης σε δύο κατηγορίες, η πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα x_j στην κατηγορία C_2 δίνεται από την παρακάτω σχέση:

$$P_j = F(a+bg_j) = \frac{1}{1+e^{-a-bg_j}} \tag{3.11}$$

Ο υπολογισμός του σταθερού όρου α και του διανύσματος b το οποίο περιέχει τους συντελεστές των χαρακτηριστικών, πραγματοποιείται χρησιμοποιώντας τεχνικές

μέγιστης πιθανοφάνειας, και πιο συγκεκριμένα μεγιστοποιώντας την ακόλουθη συνάρτηση:

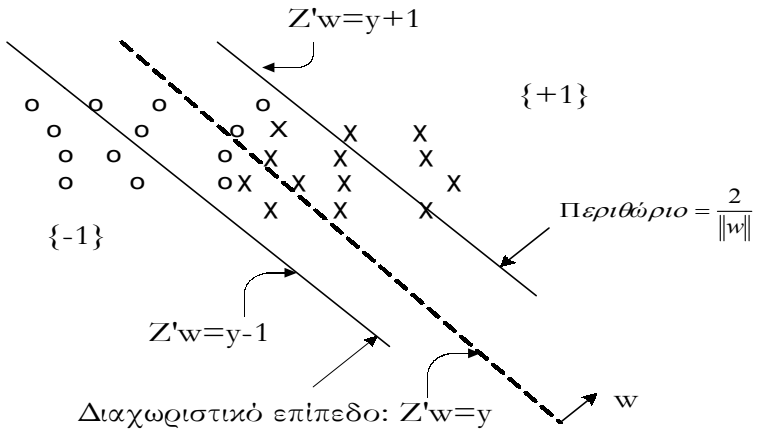
$$\ln L = \sum_{\forall x_j \in c_2} \ln\left(P_j\right) + \sum_{\forall x_j \in c_1} \ln\left(1 - P_j\right) \tag{3.12}$$

Από τη μορφή της συνάρτησης γίνεται σαφές ότι η εκτίμηση των παραμέτρων ανάγεται σε ένα πρόβλημα μη γραμμικής βελτιστοποίησης, η επίλυση του οποίου είναι πολλές φορές ιδιαίτερα δύσκολη. Επίσης πρέπει να τονιστεί ότι η σημαντικότητα των επιμέρους χαρακτηριστικών στην πραγματοποίηση της ταξινόμησης είναι δυνατόν να εκτιμηθεί μέσω γνωστών στατιστικών ελέγχων όπως το *t*-τεστ.

Η ταξινόμηση των εναλλακτικών δραστηριοτήτων πραγματοποιείται βάσει των πιθανοτήτων που υπολογίζονται μέσω του λογιστικού υποδείγματος πιθανότητας. Πιο συγκεκριμένα, κάθε εναλλακτική δραστηριότητα ταξινομείται στη κατηγορία όπου η αντίστοιχη πιθανότητα είναι μεγαλύτερη. Έτσι εάν η πιθανότητα που υπολογίζεται από τη σχέση (3.11), να ανήκει μια εναλλακτική δραστηριότητα στην κατηγορία *C*2 είναι μεγαλύτερη από 0,5, τότε η εναλλακτική δραστηριότητα εντάσσεται στην κατηγορία *C*2, διαφορετικά εντάσσεται στην κατηγορία *C*1.

3.2.4 Μηχανές διανύσματος υποστήριξης

Οι μηχανές διανύσματος υποστήριξης (SVM), έχουν αναπτυχθεί τελευταία από τον Vapnik (1995). Η λογική των μηχανών διανύσματος υποστήριξης παρουσιάζεται συνοπτικά στο σχήμα 3.3 όπου απεικονίζεται ένα πρόβλημα ταξινόμησης m αντικειμένων οι οποίες περιγράφονται βάσει n χαρακτηριστικών, σε δύο κατηγορίες οι οποίες συμβολίζονται ως +1 και -1.



Σχήμα 3.3: Γραφική απεικόνιση των μηχανών διανύσματος υποστήριξης

Στόχος των SVM, στην απλή γραμμική περίπτωση, είναι η ανάπτυξη του βέλτιστου υπερεπιπέδου της μορφής Zw-y για την ταξινόμηση των παρατηρήσεων. Ο εντοπισμός του βέλτιστου υπερεπιπέδου επιτυγχάνεται με την επίλυση του ακόλουθου τετραγωνικού προγράμματος:

$$\left.\begin{array}{l} \underset{w,y,d}{Min} \quad ve'd + \dfrac{1}{2}w'w \\ \upsilon\pi\acute{o}: \\ \qquad D(Zw - ey) + d \geq e \\ \qquad d \geq 0 \end{array}\right\} \tag{3.13}$$

Όπου:

Z: Συμβολίζεται ένας πίνακας διαστάσεων mxn με τα στοιχεία των παρατηρήσεων του δείγματος εκμάθησης.

D: Είναι ένας διαγώνιος πίνακας διαστάσεων mxm με την κύρια διαγώνιο να έχει τιμές +1 ή -1 ανάλογα με την ταξινόμηση των παρατηρήσεων του δείγματος εκμάθησης.

e: Το μοναδιαίο διάνυσμα διαστάσεων mx1.

v: Συμβολίζεται μια αυστηρά θετική σταθερά.

Ο τετραγωνικός όρος $w'w$ στην αντικειμενική συνάρτηση του προβλήματος (3.13) μεγιστοποιεί το περιθώριο μεταξύ των δυο υπερεπιπέδων $Zw - y = +1$ και $Zw - y = -1$, το οποίο ισούται με $2/\|w\|$. Εκτός της μεγιστοποίησης του περιθωρίου των κατηγοριών, το πρόβλημα (3.13) 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 διαχωρισμένες και το επίπεδο $Z'w = y + 1$ περικλείει όλα τα αντικείμενα της κατηγορίας +1, ενώ το επίπεδο $Z'w = y - 1$ περικλείει όλα τα αντικείμενα της κατηγορίας -1. Εάν οι κατηγορίες είναι γραμμικά διαχωρίσιμες (σχήμα 3.3), τα δύο επίπεδα καθορίζουν τα όρια των δύο κατηγοριών με ένα μη αρνητικό σφάλμα της μεταβλητής d. Έτσι λοιπόν με την επίλυση του προβλήματος (3.13) και τον προσδιορισμό των w και y που καθορίζουν το βέλτιστο υπερεπίπεδο, η ταξινόμηση κάθε αντικειμένου μπορεί εύκολα να πραγματοποιηθεί ως εξής:

$$\text{Εάν} \quad Z_i'w - y \begin{cases} > 0, & \tau\acute{o}\tau\varepsilon \quad x \in \{+1\}, \\ < 0, & \tau\acute{o}\tau\varepsilon \quad x \in \{-1\}, \\ = 0, & \tau\acute{o}\tau\varepsilon \quad x \in \{+1\} \quad \acute{\eta} \quad x \in \{-1\} \end{cases} \tag{3.14}$$

Το κύριο μειονέκτημα του προβλήματος βελτιστοποίησης (3.13) για τον προσδιορισμό του βέλτιστου μοντέλου ταξινόμησης αφορά τον αυξημένο υπολογιστικό φόρτο που απαιτεί η επίλυσή του καθώς πρόκειται για ένα πρόβλημα τετραγωνικού προγραμματισμού. Για την αντιμετώπιση του προβλήματος αυτού έχουν προταθεί διάφορες μεθοδολογίες για την 'επιτάχυνση' της διαδικασίας εκπαίδευσης των SVM. Μια από τις πλέον διαδεδομένες μεθοδολογίες είναι ο αλγόριθμος SVM-light (Joackins, 1998) ο οποίος χρησιμοποιήθηκε και στην παρούσα εργασία.

Όλα βέβαια τα παραπάνω, ισχύουν στην απλή γραμμική περίπτωση όπως προαναφέρθηκε. Στην μη γραμμική περίπτωση τα δεδομένα αναπαριστώνται σε ένα άλλο χώρο υψηλότερων διαστάσεων *H*, χρησιμοποιώντας μια συνάρτηση Φ ώστε:

$$\Phi : R^n \mapsto H$$

Έτσι επιτυγχάνεται ο αλγόριθμος εκπαίδευσης να εξαρτάται μόνο από τα δεδομένα που βρίσκονται στο χώρο *H*, δηλαδή μόνο από συναρτήσεις της μορφής $\Phi(x_i)\cdot\Phi(x_j)$. Στην περίπτωση όμως που ο χώρος *H* είναι εξαιρετικά μεγάλης διάστασης δεν είναι δυνατό να χρησιμοποιηθούν συναρτήσεις της μορφής Φ. Για την αντιμετώπιση αυτού του προβλήματος εισάγεται ένας πυρήνας (Kernel) *K*, της μορφής $K(x_i, x_j)$αντί των $\Phi(x_i)\cdot\Phi(x_j)$ και έτσι είναι δυνατή η εκπαίδευση μη γραμμικών SVM. Οι πυρήνες που έχουν αναπτυχθεί για την αντιμετώπιση προβλημάτων ταξινόμησης είναι διαφόρων τύπων. Στην παρούσα εργασία αυτός που χρησιμοποιήθηκε είναι ο radial basis function (RBF) ο οποίος είναι της εξής μορφής:

$$K(x, y) = e^{-\|x-y\|^2/2\sigma^2}$$

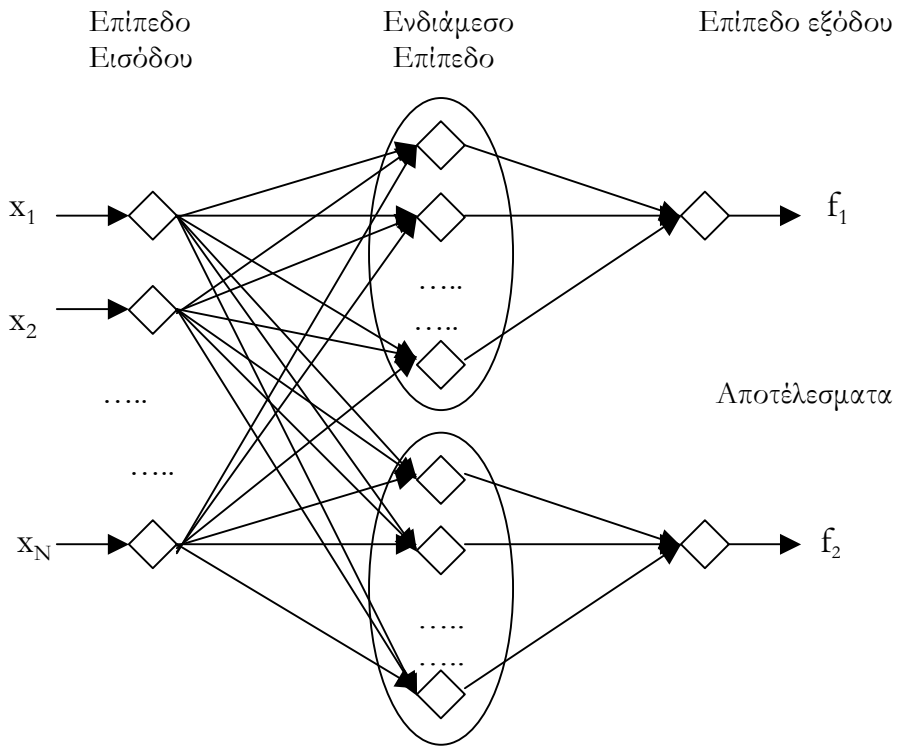
3.2.5 Πιθανοτικά νευρωνικά δίκτυα

Τα πιθανοτικά νευρωνικά δίκτυα (PNN) είναι μια κατηγορία νευρικών δικτύων, τα οποία συνδυάζουν μερικές από τις καλύτερες ιδιότητες της στατιστικής αναγνώρισης προτύπων (Pattern Recognition), και των συμβατικών νευρικών δικτύων. Τα πιθανοτικά νευρωνικά δίκτυα εισήχθησαν από τον Donald Specht στο τέλος της δεκαετίας του 1980.

Κάθε PNN είναι ένα δίκτυο παράλληλων μονάδων επεξεργασίας οι οποίες είναι οργανωμένες σε μια σειρά επιπέδων (layers). Το σχήμα (3.4) δείχνει την τυπική αρχιτεκτονική ενός PNN για προβλήματα ταξινόμησης σε δύο κατηγορίες, αλλά μπορεί

να αναχθεί και σε πολλαπλές κατηγορίες ταξινόμησης ανάλογα με τις απαιτήσεις του εκάστοτε προβλήματος. Η αρχιτεκτονική δομή ενός PNN αποτελείται από τα εξής επίπεδα όπως φαίνεται και στο σχήμα (3.4).

1) Ένα επίπεδο εισόδου (input layer) αποτελούμενο από μια σειρά κόμβων, έναν για κάθε είσοδο.
2) Το επίπεδο εξόδου (output layer) το οποίο αποτελείται από τόσους κόμβους, όσο και το πλήθος των κατηγοριών.
3) Μια σειρά ενδιάμεσων επιπέδων (hidden layers) χωρισμένα σε ομάδες. Κάθε ομάδα αντιστοιχεί και σε μια κατηγορία.



Σχήμα 3.4: Τυπική αρχιτεκτονική ενός πιθανοτικού νευρωνικού δικτύου

Όπως φαίνεται στο παραπάνω σχήμα (3.4) το επίπεδο εισόδου αποτελείται από m κόμβους, έναν για κάθε μία είσοδο, που αναπαριστάται με ένα χαρακτηριστικό διάνυσμα. Οι ενδιάμεσοι κόμβοι είναι χωρισμένοι σε ομάδες, μία ομάδα για κάθε κατηγορία. Στην περίπτωση των δύο κατηγοριών, υπάρχει μια ομάδα P με διανύσματα $\{x(p): p = 1,2,...P\}$ η οποία αντιστοιχεί στα αντικείμενα εκπαίδευσης της κατηγορίας 1

και μια ομάδα Q με διανύσματα $\{y(q): q = 1,2,...Q\}$ η οποία αντιστοιχεί στα αντικείμενα εκπαίδευσης της κατηγορίας 2. Στο ενδιάμεσο επίπεδο άρα υπάρχουν P κόμβοι για την κατηγορία 1 και Q κόμβοι για την κατηγορία 2. Οι συναρτήσεις Gauss για τις κατηγορίες 1, 2 δίνονται από τις αντίστοιχες σχέσεις:

$$g_1(x) = [1/\sqrt{(2\pi\sigma^2)}]\exp\{-\| x - x^p \|^2 /(2\sigma^2)\} \quad (3.15)$$

$$g_2(y) = [1/\sqrt{(2\pi\sigma^2)}]\exp\{-\| y - y^q \|^2 /(2\sigma^2)\} \quad (3.16)$$

Η τιμή του σ μπορεί να είναι η μέση τιμή της απόστασης των διανυσμάτων που ανήκουν στην ίδια ομάδα ή η μέση τιμή της απόστασης των διανυσμάτων της μιας ομάδας από τα κοντινότερα διανύσματα της άλλης ομάδας. Τέλος γίνεται η άθροιση των τιμών που λαμβάνονται από τις εξισώσεις (3.15) και (3.16) για κάθε ομάδα σύμφωνα με τις σχέσεις:

$$f_1(x) = [1/\sqrt{(2\pi\sigma^2)^P}](1/P)\sum_{(p=1,P)}\exp\{-\| x - x^{(p)} \|^2 /(2\sigma^2)\} \quad (3.17)$$

$$f_2(y) = [1/\sqrt{(2\pi\sigma^2)^Q}](1/Q)\sum_{(q=1,Q)}\exp\{-\| y - y^{(q)} \|^2 /(2\sigma^2)\} \quad (3.18)$$

Σε κάθε διάνυσμα εισόδου εφαρμόζονται και οι δύο παραπάνω σχέσεις (3.17), (3.18) και η μέγιστη τιμή από τις f_1 και f_2 καθορίζει την ταξινόμηση.

3.3 Πειραματικός σχεδιασμός

3.3.1 Εξεταζόμενοι παράγοντες

Η σύγκριση όλων των προαναφερθέντων μεθόδων βασίζεται στην πραγματοποίηση μιας εκτεταμένης προσομοίωσης Monte Carlo. Η προσομοίωση αυτή επιτρέπει την πραγματοποίηση της σύγκρισης σε δεδομένα τα οποία διαθέτουν συγκεκριμένες στατιστικές ιδιότητες, συμβάλλοντας με τον τρόπο αυτό στην εξαγωγή των αντίστοιχων συμπερασμάτων όσον αφορά την αποτελεσματικότητα των εξεταζόμενων μεθοδολογιών στην αντιμετώπιση του προβλήματος της ταξινόμησης. Τα δεδομένα της παρούσας συγκριτικής έρευνας θεωρούνται ότι μπορούν να αξιολογηθούν σε μια ποιοτική κλίμακα. Η αποτελεσματικότητα των εξεταζόμενων μεθόδων ελέγχεται βάσει των παρακάτω παραγόντων:

1. Πλήθος διακριτών επιπέδων.
2. Τρόπος ταξινόμησης των δεδομένων.
3. Πλήθος των αντικειμένων εκπαίδευσης
4. Συσχέτιση των χαρακτηριστικών (κριτηρίων αξιολόγησης).
5. Επίπεδο θορύβου που εισάγεται στα δεδομένα.

Ο Πίνακας 3.1 παρουσιάζει τον τρόπο με τον οποίο χρησιμοποιούνται οι παραπάνω παράγοντες στον πειραματικό σχεδιασμό που πραγματοποιείται με σκοπό την αξιολόγηση της αποτελεσματικότητας των εξεταζόμενων μεθόδων ταξινόμησης.

Πίνακας 3.1: Εξεταζόμενοι παράγοντες στον πειραματικό σχεδιασμό για την σύγκριση των μεθόδων ταξινόμησης.

Παράγοντες		**Επίπεδα**
Π_1	Τεχνικές ταξινόμησης	1. Πιθανοτικά νευρωνικά δίκτυα (PNN) 2. Γραμμική διακριτική ανάλυση (LDA) 3. Λογιστικό υπόδειγμα πιθανότητας (LOGIT) 4. Τετραγωνική διακριτική ανάλυση (QDA) 5. Μηχανές διανύσματος υποστήριξης (SVM)
Π_2	Πλήθος διακριτών επιπέδων	1. {-1,1} 2. {-1,0,1}
Π_3	Ταξινόμηση των δεδομένων	1. Γραμμική ταξινόμηση 2. Μη γραμμική ταξινόμηση
Π_4	Πλήθος αντικειμένων εκπαίδευσης	1. 500 αντικείμενα, 5 κριτήρια 2. 1000 αντικείμενα, 5 κριτήρια
Π_5	Συσχέτιση χαρακτηριστικών	1. Χαμηλή συσχέτιση 2. Υψηλή συσχέτιση
Π_6	Επίπεδο θορύβου	1. 10% 2. 20%

Ο παράγοντας Π_2 είναι ο πρώτος από τους παράγοντες που καθορίζουν τη μορφή των εξεταζομένων δεδομένων. Ο παράγοντας αυτός αναφέρεται στο πλήθος των διακριτών επιπέδων των εναλλακτικών δραστηριοτήτων. Για τον παράγοντα αυτό εξετάζονται δύο περιπτώσεις. Στην πρώτη περίπτωση εξετάζεται τα αντικείμενα προς αξιολόγηση, να είναι δυαδικά και παίρνουν τις τιμές –1 και 1 ενώ στη δεύτερη

περίπτωση εξετάζονται οι περιπτώσεις που τα αντικείμενα μπορούν τα αξιολογηθούν σε μια ποιοτική κλίμακα τριών βαθμίδων και παίρνουν τιμές –1,0 και 1.

Ο επόμενος παράγοντας Π_3 καθορίζει την ταξινόμηση των δεδομένων. Στον παρόντα πειραματικό σχεδιασμό εξετάζονται δύο είδη ταξινόμησης των δεδομένων. Στην πρώτη περίπτωση η ταξινόμηση είναι γραμμική και στη δεύτερη περίπτωση είναι μη γραμμική.

Ο επόμενος παράγοντας Π_4 καθορίζει το πλήθος των αντικειμένων εκμάθησης. Στον παρόντα πειραματικό σχεδιασμό εξετάζονται δύο επίπεδα για τον παράγοντα αυτό, σύμφωνα με τα οποία το δείγμα εκμάθησης περιλαμβάνει 500 ή 1000 αντικείμενα, τα οποία περιγράφονται από ένα σύνολο πέντε χαρακτηριστικών (κριτηρίων αξιολόγησης). Καθώς το δείγμα εκμάθησης αυξάνεται προστίθεται και νέα πληροφορία, αλλά παράλληλα αυξάνεται και η πολυπλοκότητα του προβλήματος ταξινόμησης. Συνεπώς, η εξέταση του παράγοντα αυτού περιλαμβάνει την αξιολόγηση της αποτελεσματικότητας των παραπάνω προσεγγίσεων σε περιπτώσεις που η παρεχόμενη πληροφορία είναι περιορισμένη και άρα και η πολυπλοκότητα, αλλά και τις περιπτώσεις που υπάρχει αυξημένη πληροφορία και άρα και αυξημένη πολυπλοκότητα.

Ο επόμενος παράγοντας Π_5 καθορίζει την συσχέτιση των χαρακτηριστικών (κριτηρίων αξιολόγησης). Για τον παράγοντα αυτό εξετάζονται δύο περιπτώσεις. Η πρώτη περίπτωση είναι να υπάρχει χαμηλή συσχέτιση μεταξύ των κριτηρίων αξιολόγησης και η δεύτερη να υπάρχει υψηλή συσχέτιση.

Ο τελευταίος παράγοντας Π_5 σχετίζεται με το επίπεδο του θορύβου που εισάγεται στα δεδομένα του παρόντος πειραματικού σχεδιασμού. Για τον παράγοντα αυτό εξετάζονται δύο περιπτώσεις. Στην πρώτη περίπτωση εισάγεται στην ταξινόμηση των δεδομένων του πειραματικού σχεδιασμού ένα επίπεδο θορύβου της τάξης των 10% και στην δεύτερη περίπτωση εισάγεται ένα μεγαλύτερο ποσοστό θορύβου της τάξης του 20%.

3.3.2 Διαδικασία παραγωγής των δεδομένων

Για κάθε συνδυασμό των κατηγοριών Π_2 έως και Π_6, η παραπάνω διαδικασία χρησιμοποιείται για την παραγωγή δύο συνόλων δεδομένων. Το πρώτο χρησιμοποιείται ως δείγμα εκμάθησης και το δεύτερο ως δείγμα ελέγχου. Οι εναλλακτικές δραστηριότητες που παράγονται τόσο στο δείγμα ελέγχου όσο και στο δείγμα

εκμάθησης είναι 5000. Από τα 5000 αυτά αντικείμενα επιλέγονται κάποια τυχαία ώστε να συμφωνούν με τους παράγοντες Π_2 και Π_5 (βλ. Πίνακα 3.1). Το μέγεθος του δείγματος εκμάθησης καθορίζεται από τον παράγοντα Π_4 και περιλαμβάνει 500 ή 1000 εναλλακτικές δραστηριότητες, ενώ το δείγμα ελέγχου αποτελείται σε κάθε περίπτωση από 500 εναλλακτικές δραστηριότητες. Επίσης, τόσο στο δείγμα ελέγχου όσο και στο δείγμα εκμάθησης, ο αριθμός των εναλλακτικών δραστηριοτήτων των δύο κατηγοριών είναι ίδιος.

Ο παραπάνω πειραματικός έλεγχος επαναλαμβάνεται 30 φορές για κάθε συνδυασμό των παραγόντων Π_2 έως και Π_6 (32 δυνατοί συνδυασμοί). Συνολικά ελέγχονται 960 (32 συνδυασμοί x 30 επαναλήψεις) διαφορετικά δείγματα εκμάθησης, στα οποία αντιστοιχούν ισάριθμα δείγματα ελέγχου. Σε κάθε ένα από αυτά τα σύνολα δεδομένων εφαρμόζονται οι πέντε τεχνικές ταξινόμησης που περιλαμβάνονται στον πίνακα Π_1 (βλ. Πίνακα 3.1).

Η περίπτωση της παραγωγής δεδομένων από τον παράγοντα Π_3, που όπως προαναφέρθηκε καθορίζει την ταξινόμηση των δεδομένων, ακολουθεί τους εξής κανόνες:

α) Όσον αφορά την πρώτη περίπτωση, η ταξινόμηση των δεδομένων είναι γραμμική και εκφράζεται από μια εξίσωση της μορφής: $Z = a_1x_2 + a_2x_2 + ... + a_5x_5$

Οι συντελεστές $\alpha_1,\alpha_2,\ldots,\alpha_5$ της παραπάνω εξίσωσης, είναι τυχαίοι αριθμοί και ακολουθούν την ομοιόμορφη κατανομή στο διάστημα [1,10].

β) Όσον αφορά την δεύτερη περίπτωση η ταξινόμηση των δεδομένων είναι μη γραμμική και εκφράζεται από μια εξίσωση της μορφής: $Z' = Z + \sum_{i=1}^{5}\sum_{j=1}^{5} a_{ij}x_i x_j$. Οι συντελεστές α_{ij} της παραπάνω εξίσωσης, είναι τυχαίοι αριθμοί και ακολουθούν την ομοιόμορφη κατανομή στο διάστημα [0,1].

3.4 Ανάλυση των αποτελεσμάτων

Η υλοποίηση του παρόντος πειραματικού σχεδιασμού πραγματοποιήθηκε στο Matlab και η στατιστική ανάλυση και επεξεργασία των αποτελεσμάτων, που ακολουθεί, πραγματοποιήθηκε στο στατιστικό πακέτο SPSS. Πιο συγκεκριμένα, η επεξεργασία των αποτελεσμάτων βασίζεται στην ανάλυση διασποράς (ANOVA) και τον στατιστικό έλεγχο του Tukey ο οποίος επιτρέπει τη διαμόρφωση ομοιογενών ομάδων όπου κάθε

μια περιλαμβάνει τα επίπεδα ενός παράγοντα για τα οποία δεν παρουσιάζονται στατιστικά σημαντικές διαφορές μεταξύ τους, ως προς το ποσοστό των εσφαλμένων ταξινομήσεων. Στις αντίστοιχες παρενθέσεις, δίπλα από το ποσοστό εσφαλμένων ταξινομήσεων ,εφεξής , παρουσιάζεται η ομαδοποίηση των διαφόρων μορφών διακύμανσης σύμφωνα με τον στατιστικό έλεγχο Tukey, σε επίπεδο σημαντικότητας 5%.

Τα αποτελέσματα που εξετάζονται για κάθε μέθοδο ταξινόμησης αφορούν μόνο το ποσοστό των εσφαλμένων ταξινομήσεων του δείγματος ελέγχου, καθώς αυτά αντικατοπτρίζουν πληρέστερα την πραγματική ικανότητα των μεθόδων να αντιμετωπίσουν με επιτυχία το πρόβλημα της ταξινόμησης. Έτσι, η αξιολόγηση των αποτελεσμάτων του συγκεκριμένου πειραματικού σχεδιασμού, επιτρέπει την εξαγωγή αμερόληπτων εκτιμήσεων όσον αφορά την αναμενόμενη αποτελεσματικότητα των εξεταζόμενων μεθόδων στην αντιμετώπιση των προβλημάτων ταξινόμησης που συναντώνται στην πράξη.

Ο πίνακας 3.2 παρουσιάζει τα αποτελέσματα της ανάλυσης διασποράς των πέντε εξεταζόμενων παραγόντων, στα αποτελέσματα ταξινόμησης του δείγματος ελέγχου.

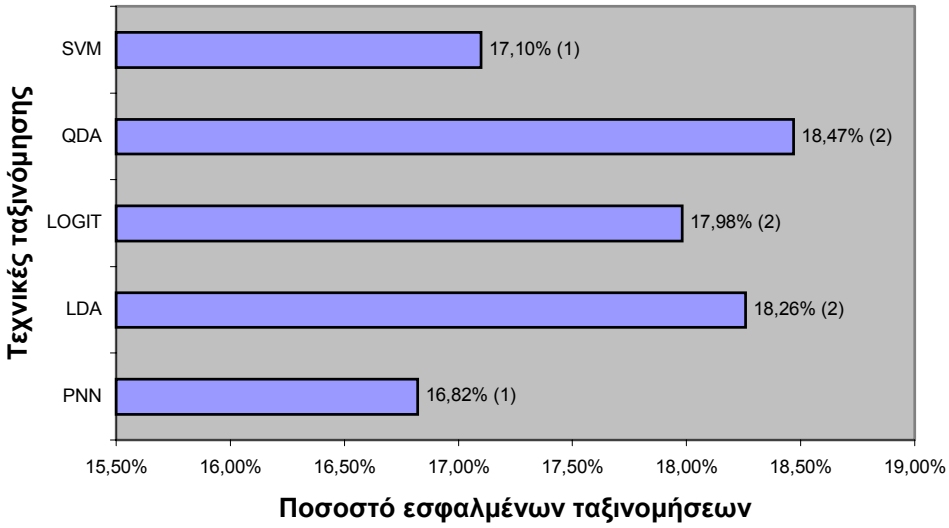
Πίνακας 3.2: Ανάλυση διασποράς των αποτελεσμάτων του πειραματικού σχεδιασμού στο δείγμα ελέγχου

	Βαθμοί ελευθερίας	Μέσα τετραγώνα	F	ω^2
Π_4	1	13,10242	26588,04	77,8%
Π_2	1	0,459386	932,2078	2,7%
Π_3	1	0,219376	445,167	1,3%
Π_1	4	0,050576	102,6303	1,2%
$\Pi_1 \times \Pi_3$	4	0,033027	67,01918	0,8%
$\Pi_1 \times \Pi_5$	4	0,014495	29,41468	0,3%
$\Pi_1 \times \Pi_3 \times \Pi_5$	4	0,011728	23,79971	0,3%
$\Pi_1 \times \Pi_2$	4	0,010686	21,68592	0,2%
$\Pi_1 \times \Pi_2 \times \Pi_5$	4	0,009103	18,47303	0,2%
$\Pi_1 \times \Pi_6$	4	0,004517	9,166624	0,1%
$\Pi_1 \times \Pi_2 \times \Pi_3$	4	0,002870	5,825281	0,1%

Καθένας από τους παραπάνω παράγοντες επιδρά σημαντικά στην αποτελεσματικότητα με την οποία μπορεί να αντιμετωπιστεί το πρόβλημα ταξινόμησης όσον αφορά τη δυνατότητα γενίκευσης των αναπτυσσόμενων υποδειγμάτων στο δείγμα ελέγχου. Η πειραματική ανάλυση που θα πραγματοποιηθεί βασίζεται μόνο στις επιδράσεις και αλληλεπιδράσεις που παρουσιάζονται στον παραπάνω πίνακα, έτσι ώστε

να διευκολυνθεί η διαδικασία ανάλυσης των αποτελεσμάτων του πειραματικού σχεδιασμού.

Στο σχήμα 3.5 παρουσιάζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό εσφαλμένων ταξινομήσεων όλων των εξεταζομένων τεχνικών ταξινόμησης στο δείγμα ελέγχου.



Σχήμα 3.5: Μέσο ποσοστό εσφαλμένων ταξινομήσεων των εξεταζόμενων τεχνικών ταξινόμησης στο δείγμα ελέγχου

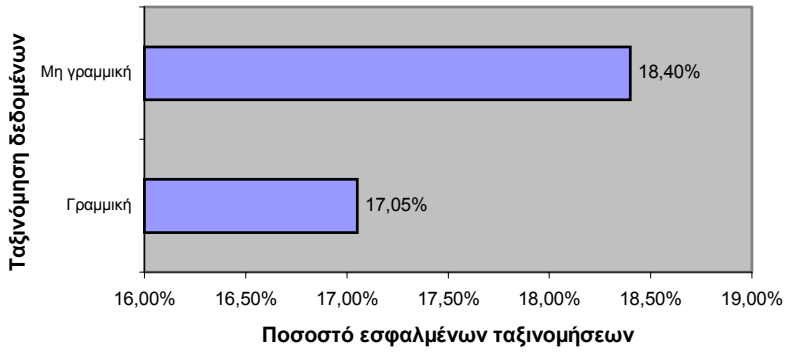
Τα αποτελέσματα δείχνουν ότι η μέθοδος PNN αποδίδει το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων. Με λίγο μεγαλύτερο σφάλμα ταξινόμησης ακολουθεί η μέθοδος SVM και έπονται η LOGIT και η LDA. Την χαμηλότερη αποτελεσματικότητα παρουσιάζει η μέθοδος QDA. Τα αποτελέσματα αυτά αποδίδουν τη γενική εικόνα της αποτελεσματικότητας των εξεταζομένων τεχνικών ταξινόμησης, αλλά δεν συμβάλουν στον εντοπισμό του τρόπου με τον οποίο επιδρούν οι επιμέρους παράμετροι του πειραματικού σχεδιασμού στην αποτελεσματικότητα των τεχνικών ταξινόμησης. Για τον λόγο αυτό κατά την διάρκεια της πειραματικής ανάλυσης θα συνδυαστούν κάποιοι παράγοντες, η αλληλεπίδραση των οποίων είναι σημαντική στην αποτελεσματικότητα του προβλήματος ταξινόμησης.

Μια από τις αλληλεπιδράσεις αυτές είναι η αλληλεπίδραση μεταξύ των εξεταζομένων προσεγγίσεων και της ταξινόμησης των δεδομένων ($\Pi_1 X \Pi_3$). Τα σχετικά αποτελέσματα παρουσιάζονται στον πίνακα 3.3. Τα αποτελέσματα αυτά δείχνουν ότι τόσο στην περίπτωση που η ταξινόμηση των δεδομένων είναι γραμμική όσο και στην

περίπτωση που είναι μη γραμμική, η μέθοδος PNN αποδίδει το μικρότερο μέσο ποσοστό εσφαλμένων ταξινομήσεων σε σύγκριση με τις άλλες μεθόδους ταξινόμησης. Αναλυτικότερα, στην περίπτωση οπού η ταξινόμηση των δεδομένων είναι γραμμική, η τεχνική PNN παρουσιάζει το μικρότερο μέσο ποσοστό εσφαλμένων ταξινομήσεων, ακολουθούμενη κατά σειρά τις LOGIT, SVM και LDA ενώ τη μικρότερη αποτελεσματικότητα με σημαντική διαφορά από τις υπόλοιπες τεχνικές ταξινόμησης παρουσιάζει η QDA. Στην περίπτωση όπου η ταξινόμηση των δεδομένων είναι μη γραμμική, η αποτελεσματικότητα της QDA βελτιώνεται σημαντικά, υποσκελίζοντας τόσο την LDA που παρουσιάζει τη μικρότερη αποτελεσματικότητα στην συγκεκριμένη περίπτωση όσο και την LOGIT. Παρόλο όμως τη σημαντική βελτίωση που παρουσιάζει, οι διαφορές μεταξύ της τετραγωνικής διακριτικής ανάλυσης και των τεχνικών SVM και PNN παραμένουν στατιστικά σημαντικές. Σε κάθε περίπτωση πάντως, το μέσο ποσοστό εσφαλμένων ταξινομήσεων, για όλες τις μεθόδους, είναι σημαντικά μειωμένο στην περίπτωση της γραμμικής ταξινόμησης των δεδομένων, όπως φαίνεται και στο σχήμα 3.6.

Πίνακας 3.3: Μέσο ποσοστό εσφαλμένων ταξινομήσεων των εξεταζομένων προσεγγίσεων ταξινόμησης στο δείγμα ελέγχου συναρτήσει της ταξινόμησης των δεδομένων.

	Γραμμική	**Μη γραμμική**
PNN	**16,47% (1)**	**17,17% (1)**
LDA	**16,93% (1)**	**19,59% (2)**
LOGIT	**16,68% (1)**	**19,28% (2)**
QDA	**18,26% (2)**	**18,68% (2)**
SVM	**16,91% (1)**	**17,30% (1)**

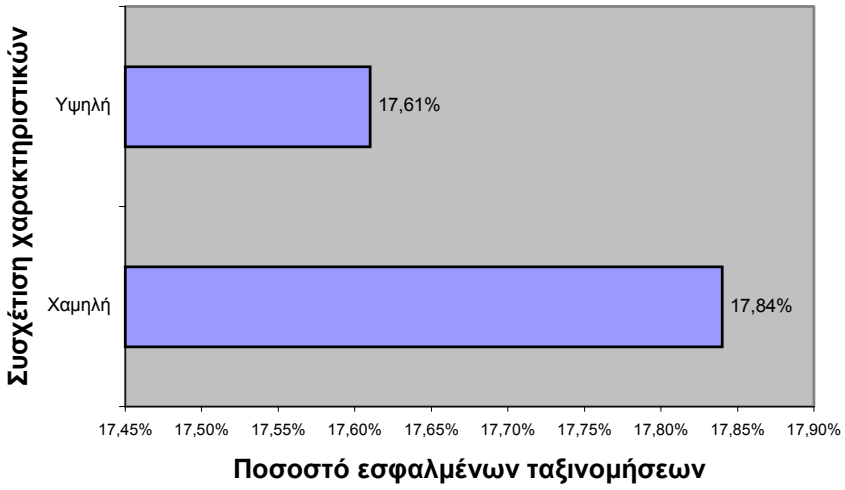


Σχήμα 3.6: Μέσο ποσοστό εσφαλμένων ταξινομήσεων συναρτήσει της μορφής της ταξινόμησης των δεδομένων

Μια άλλη αλληλεπίδραση που θα εξεταστεί είναι αυτή μεταξύ των εξεταζομένων προσεγγίσεων και της συσχέτισης που παρουσιάζουν τα χαρακτηριστικά ($\Pi_1 X \Pi_5$). Τα σχετικά αποτελέσματα παρουσιάζονται στον πίνακα 3.4. Τα αποτελέσματα αυτά δείχνουν ότι όσον αφορά την περίπτωση που η συσχέτιση των χαρακτηριστικών είναι χαμηλή, επιβεβαιώνεται για άλλη μια φορά η αποτελεσματικότητα της τεχνικής PNN, αφού παρουσιάζει το μικρότερο μέσο ποσοστό εσφαλμένων ταξινομήσεων, ακολουθούμενη από την τεχνική SVM ενώ με σημαντική διαφορά ακολουθούν κατά σειρά οι QDA, LOGIT και LDA. Στην περίπτωση όπου η συσχέτιση των χαρακτηριστικών είναι υψηλή η κατάταξη για τις δύο πιο αποτελεσματικές τεχνικές ταξινόμησης δεν αλλάζει, ενώ, η αποτελεσματικότητα της QDA μειώνεται κατά ένα μεγάλο ποσοστό και κατατάσσεται τελευταία. Για τις περισσότερες τεχνικές ταξινόμησης, το μέσο ποσοστό εσφαλμένων ταξινομήσεων, μειώνεται όσο υψηλότερος είναι ο βαθμός συσχέτισης (σχήμα 3.7).

Πίνακας 3.4: Μέσο ποσοστό εσφαλμένων ταξινομήσεων των εξεταζομένων προσεγγί-σεων ταξινόμησης στο δείγμα ελέγχου συναρτήσει της συσχέτισης των χαρακτηριστικών

	Χαμηλή συσχέτιση	**Υψηλή συσχέτιση**
PNN	**16,74% (1)**	**16,91% (1)**
LDA	**18,77% (3)**	**17,75% (1)**
LOGIT	**18,43% (3)**	**17,53% (1)**
QDA	**18,04% (2,3)**	**18,90% (2)**
SVM	**17,24% (1,2)**	**16,97% (1)**

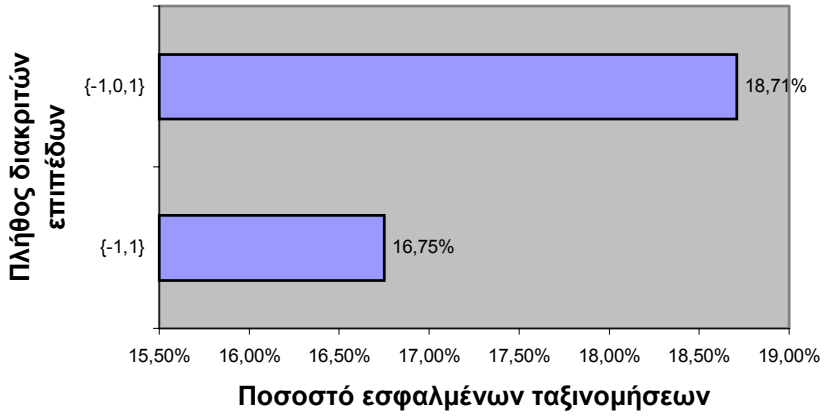


Σχήμα 3.7: Μέσο ποσοστό εσφαλμένων ταξινομήσεων συναρτήσει της συσχέτισης των χαρακτηριστικών

Μια άλλη αλληλεπίδραση δευτέρου βαθμού που θα εξεταστεί είναι αυτή μεταξύ των εξεταζομένων προσεγγίσεων και του πλήθους των διακριτών επιπέδων ($Π_1ΧΠ_2$).Τα σχετικά αποτελέσματα παρουσιάζονται στον πίνακα 3.5. Για άλλη μια φορά επιβεβαιώνεται η αποτελεσματικότητα της τεχνικής PNN, τόσο στην περίπτωση που το πλήθος των διακριτών επιπέδων των χαρακτηριστικών είναι δυαδικό{-1,1}, όσο και στην περίπτωση όπου υπάρχει μια ποιοτική κλίμακα (τριών βαθμίδων) αξιολόγησης των χαρακτηριστικών. Αξιοσημείωτη είναι επίσης η σημαντική μείωση των εσφαλμένων ταξινομήσεων, για όλες τις τεχνικές ταξινόμησης, που παρατηρείται στην περίπτωση όπου το πλήθος των διακριτών επιπέδων των χαρακτηριστικών είναι δυαδικό (σχήμα 3.8). Επίσης για άλλη μια φορά παρατηρείται, μια σημαντική αύξηση του ποσοστού των εσφαλμένων ταξινομήσεων για την τεχνική QDA, στην περίπτωση όπου υπάρχει μια ποιοτική κλίμακα αξιολόγησης των χαρακτηριστικών σε τρία επίπεδα.

Πίνακας 3.5: Μέσο ποσοστό εσφαλμένων ταξινομήσεων των εξεταζομένων προσεγγίσεων ταξινόμησης στο δείγμα ελέγχου συναρτήσει του πλήθους των διακριτών επιπέδων

	{-1,1}	**{-1,0,1}**
PNN	**15,51% (1)**	**18,13% (1)**
LDA	**17,67% (2)**	**18,85% (1,2)**
LOGIT	**17,32% (2)**	**18,64% (1)**
QDA	**17,26% (2)**	**19,67% (2)**
SVM	**15,97% (1)**	**18,24% (1)**

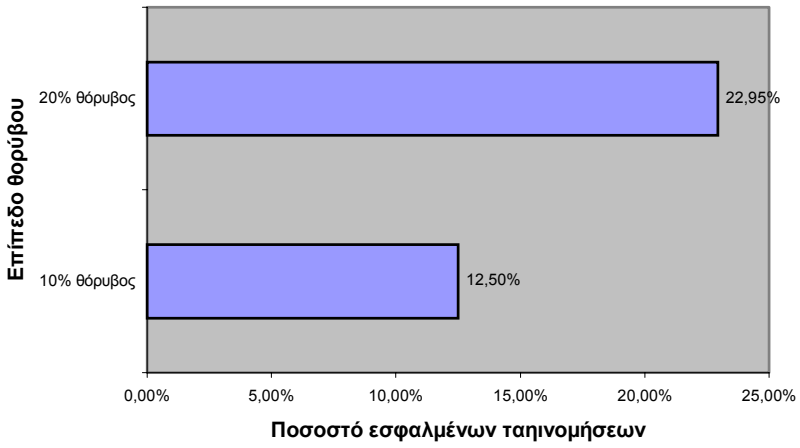


Σχήμα 3.8: Μέσο ποσοστό εσφαλμένων ταξινομήσεων συναρτήσει του πλήθους των διακριτών επιπέδων

Η τελευταία αλληλεπίδραση δευτέρου βαθμού η οποία παρουσιάζεται, αφορά την αλληλεπίδραση μεταξύ των προσεγγίσεων ταξινόμησης και το επίπεδο του θορύβου που εισάγεται στα δεδομένα ($\Pi_1 X \Pi_6$).Τα σχετικά αποτελέσματα παρουσιάζονται στον πίνακα 3.6. Τα αποτελέσματα του πίνακα 3.6 δείχνουν ότι η αύξηση του θορύβου στα δεδομένα οδήγησε σε σημαντική μείωση της αποτελεσματικότητας όλων των μεθόδων ταξινόμησης, όπως άλλωστε ήταν αναμενόμενο (σχήμα 3.9). Για ακόμα μια φορά είναι φανερή η αποτελεσματικότητα της PNN και στις δύο περιπτώσεις, αλλά και της SVM που ακολουθεί με μικρή διαφορά. Παρατηρείται επίσης μια σημαντική αύξηση του ποσοστού των εσφαλμένων ταξινομήσεων για την τεχνική QDA, στην περίπτωση όπου υπάρχει επίπεδο θορύβου 20% σε σύγκριση με το αντίστοιχο ποσοστό που παρουσιάζει η QDA για επίπεδο θορύβου 10%.

Πίνακας 3.6: Μέσο ποσοστό εσφαλμένων ταξινομήσεων των εξεταζομένων προσεγγίσεων ταξινόμησης στο δείγμα ελέγχου συναρτήσει του θορύβου στα δεδομένα

	Θόρυβος 10%	**Θόρυβος 20%**
PNN	**11,67% (1)**	**21,98% (1)**
LDA	**13,28% (2)**	**23,23% (3)**
LOGIT	**12,88% (2)**	**23,08% (3)**
QDA	**12,96% (2)**	**23,98% (4)**
SVM	**11,73% (1)**	**22,48% (2)**



Σχήμα 3.8: Μέσο ποσοστό εσφαλμένων ταξινομήσεων συναρτήσει του επιπέδου του θορύβου

Εκτός των παραπάνω αλληλεπιδράσεων δευτέρου βαθμού, θα εξεταστούν και τρεις αλληλεπιδράσεις τρίτου βαθμού που βοηθούν να εξηγηθούν τα αποτελέσματα του παρόντος πειραματικού σχεδιασμού. Η πρώτη από τις αλληλεπιδράσεις αυτές αφορά

την αλληλεπίδραση των παραγόντων που αφορούν τις εξεταζόμενες προσεγγίσεις ταξινόμησης, την ταξινόμηση των δεδομένων και τη συσχέτιση των χαρακτηριστικών ($Π_1ΧΠ_3ΧΠ_5$). Τα αντίστοιχα αποτελέσματα συνοψίζονται στον πίνακα 3.7 και επιτρέπουν την πληρέστερη ανάλυση των συμπερασμάτων που παρουσιάστηκαν προηγουμένως όσον αφορά την επίδραση της ταξινόμησης των δεδομένων και της συσχέτισης των χαρακτηριστικών στην αποτελεσματικότητα των εξεταζόμενων προσεγγίσεων ταξινόμησης. Αναλυτικότερα, η ταυτόχρονη εξέταση των δύο αυτών παραγόντων δείχνει ότι όταν η συσχέτιση των δεδομένων είναι χαμηλή τότε η PNN αποδίδει το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων, ενώ όταν η συσχέτιση των δεδομένων είναι υψηλή τότε η PNN αποδίδει το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων μόνο στην περίπτωση που και η ταξινόμηση είναι γραμμική. Η QDA παρουσιάζει το μεγαλύτερο ποσοστό εσφαλμένων ταξινομήσεων εκτός της περίπτωσης που η ταξινόμηση των δεδομένων είναι μη γραμμική και η συσχέτιση χαμηλή όπου το μεγαλύτερο ποσοστό εσφαλμένων ταξινομήσεων παρουσιάζει η LDA. Συνεπώς επιβεβαιώνεται για άλλη μια φορά η αποτελεσματικότητα της PNN και πιο συγκεκριμένα αποδίδει το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων στην περίπτωση που η ταξινόμηση των δεδομένων είναι γραμμική και η συσχέτιση των χαρακτηριστικών χαμηλή. Επίσης παρατηρείται ότι το ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου που παρουσιάζει η τεχνική QDA δεν επηρεάζεται σημαντικά συναρτήσει της συσχέτισης των χαρακτηριστικών και της ταξινόμησης των δεδομένων.

Πίνακας 3.7: Μέσο ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου συναρτήσει της ταξινόμησης των δεδομένων και της συσχέτισης των χαρακτηριστικών

Χαμηλή συσχέτιση των χαρακτηριστικών		
	Γραμμική ταξινόμηση	Μη γραμμική ταξινόμηση
PNN	16,24% **(1)**	17,23% **(1)**
LDA	16,65% **(1)**	20,89% **(2)**
LOGIT	16,37% **(1)**	20,49% **(2)**
QDA	17,63% **(1)**	18,46% **(1)**
SVM	16,96% **(1)**	17,51% **(1)**

Υψηλή συσχέτιση των χαρακτηριστικών		
	Γραμμική ταξινόμηση	Μη γραμμική ταξινόμηση
PNN	16,71% **(1)**	17,11% **(1)**
LDA	17,22% **(1)**	18,28% **(1,2)**
LOGIT	16,98% **(1)**	18,08% **(1,2)**
QDA	18,89% **(2)**	18,90% **(2)**
SVM	16,86% **(1)**	17,08% **(1)**

Η δεύτερη σημαντική αλληλεπίδραση τρίτου βαθμού αφορά των συνδυασμό των παραγόντων που αναφέρονται στις εξεταζόμενες προσεγγίσεις ταξινόμησης, στο πλήθος των διακριτών επιπέδων και στη συσχέτιση των χαρακτηριστικών ($Π_1ΧΠ_2ΧΠ_5$). Τα αντίστοιχα αποτελέσματα συνοψίζονται στον πίνακα 3.8. Και πάλι η PNN παρουσιάζει το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων εκτός της περίπτωσης όπου η συσχέτιση των χαρακτηριστικών είναι υψηλή και το πλήθος των διακριτών επιπέδων εκφράζεται σε ποιοτική κλίμακα τριών βαθμίδων στην οποία υπερτερεί η τεχνική SVM. Επίσης παρατηρείται μια μεγάλη αύξηση του ποσοστού των εσφαλμένων ταξινομήσεων των μεθόδων PNN και SVM στην περίπτωση όπου η συσχέτιση των χαρακτηριστικών είναι χαμηλή και το πλήθος των διακριτών επιπέδων εκφράζεται σε ποιοτική κλίμακα, σε σχέση με το αντίστοιχο ποσοστό όπου το πλήθος των κατηγοριών είναι δυαδικής μορφής. Αντίθετα οι LOGIT και LDA αυξάνουν ελάχιστα το ποσοστό των εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου για την αντίστοιχη περίπτωση.

Πίνακας 3.8: Μέσο ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου συναρτήσει του πλήθους των διακριτών επιπέδων και του ποσοστού συσχέτισης των χαρακτηριστικών

Χαμηλή συσχέτιση των χαρακτηριστικών		
	{-1,0}	{-1,0,1}
PNN	15,34% **(1)**	18,13% **(1)**
LDA	18,56% **(3)**	18,98% **(1)**
LOGIT	18,11% **(3)**	18,75% **(1)**
QDA	17,20% **(2,3)**	18,88% **(1)**
SVM	15,81% **(1,2)**	18,66% **(1)**
Υψηλή συσχέτιση των χαρακτηριστικών		
	{-1,1}	{-1,0,1}
PNN	15,68% **(1)**	18,14% **(1)**
LDA	16,79% **(1,2)**	18,71% **(1)**
LOGIT	16,54% **(1,2)**	18,52% **(1)**
QDA	17,33% **(2)**	20,47% **(2)**
SVM	16,13% **(1,2)**	17,82% **(1)**

Η τελευταία αλληλεπίδραση τρίτου βαθμού που θα εξεταστεί αφορά των συνδυασμό των παραγόντων που αναφέρονται στις εξεταζόμενες προσεγγίσεις ταξινόμησης, στο πλήθος των διακριτών επιπέδων και στην ταξινόμηση των δεδομένων ($\Pi_1 X \Pi_2 X \Pi_3$). Τα αντίστοιχα αποτελέσματα συνοψίζονται στον πίνακα 3.9. Η PNN παρουσιάζει το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου και σε αυτή την περίπτωση, με τη διαφορά ότι και άλλες μέθοδοι ταξινόμησης αποδίδουν εξίσου αποτελεσματικά. Πιο συγκεκριμένα στην περίπτωση όπου η ταξινόμηση των δεδομένων είναι γραμμική και τα δεδομένα είναι ποιοτικής μορφής τριών βαθμίδων, τόσο η LOGIT όσο και η LDA παρουσιάζουν εξίσου καλή αποτελεσματικότητα. Όσον αφορά τη μέθοδο QDA για άλλη μια φορά έχει το μεγαλύτερο ποσοστό εσφαλμένων ταξινομήσεων εκτός της περίπτωσης όπου η ταξινόμηση των δεδομένων είναι μη

γραμμική και τα δεδομένα είναι δυαδικής μορφής, στην οποία το μεγαλύτερο σφάλμα παρουσιάζει η LDA.

Πίνακας 3.8: Μέσο ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου συναρτήσει του πλήθους των διακριτών επιπέδων και της ταξινόμησης των δεδομένων

Γραμμική ταξινόμηση δεδομένων		
	{-1,0}	{-1,0,1}
PNN	15,22% **(1)**	17,73% **(1)**
LDA	16,12% **(1,2)**	17,75% **(1)**
LOGIT	15,84% **(1,2)**	17,51% **(1)**
QDA	17,19% **(2)**	19,31% **(2)**
SVM	15,90% **(1,2)**	17,93% **(1,2)**
Μη γραμμική ταξινόμηση δεδομένων		
	{-1,1}	{-1,0,1}
PNN	15,80% **(1)**	18,54% **(1)**
LDA	19,23% **(3)**	19,94% **(1,2)**
LOGIT	18,81% **(3)**	19,76% **(1,2)**
QDA	17,34% **(2)**	20,02% **(2)**
SVM	16,04% **(1,2)**	18,55% **(1)**

3.5 Συμπεράσματα

Ο πειραματικός σχεδιασμός στο κεφάλαιο αυτό, συνέβαλε στην εξέταση της σχετικής αποτελεσματικότητας των υποδειγμάτων ταξινόμησης που εξετάστηκαν, σε ένα ευρύ φάσμα συνθηκών σχετικά με τη μορφή και τις ιδιότητες των εξεταζομένων δεδομένων. Η παραπάνω έρευνα παρέχει τις απαραίτητες βάσεις για την πλήρη κατανόηση των προσεγγίσεων ταξινόμησης, των ιδιαιτεροτήτων και χαρακτηριστικών της κάθε προσέγγισης και του τρόπου με τον οποίο μπορούν οι διάφοροι παράγοντες να επιδρούν στην αποτελεσματικότητα της κάθε μεθόδου.

Τα βασικά συμπεράσματα της έρευνας που παρουσιάστηκε συνοψίζονται στα ακόλουθα στοιχεία.

1. Τα πιθανοτικά νευρωνικά δίκτυα μπορούν να θεωρηθούν γενικά ως μια αποτελεσματική μέθοδος ταξινόμησης. Σε σύγκριση μάλιστα με τις τεχνικές ταξινόμησης που παρουσιάστηκαν στον παραπάνω πειραματικό σχεδιασμό αποδίδει σαφώς το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου, ανεξαρτήτως των παραγόντων που επιδρούν και καθορίζουν την ταξινόμηση, και κατά συνέπεια μπορεί να θεωρηθεί ως η αποτελεσματικότερη τεχνική ταξινόμησης. Μια μόνο μικρή αδυναμία της μεθόδου παρατηρείται ως προς το ποσοστό των εσφαλμένων ταξινομήσεων που αποδίδει, όταν το πλήθος των διακριτών επιπέδων εκφράζεται από μια κλίμακα ποιοτικής μορφής τριών επιπέδων και ταυτόχρονα η συσχέτιση που παρουσιάζουν τα δεδομένα είναι υψηλή. Στην περίπτωση αυτή αποτελεσματικότερη μέθοδος είναι η SVM, χωρίς όμως αυτό να σημαίνει ότι η PNN δεν χαρακτηρίζεται αξιόπιστη και σε αυτή την περίπτωση.

2. Όσον αφορά, της μηχανές διανύσματος υποστήριξης (SVM), μπορεί να θεωρηθεί ως εξίσου αποτελεσματική μέθοδος ταξινόμησης με τα πιθανοτικά νευρωνικά δίκτυα. Ειδικά σε κάποιες περιπτώσεις, όπως αυτή που αναφέρθηκε στην προηγούμενη επισήμανση, η SVM αποδίδει μικρότερο ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου από ότι αποδίδει στην αντίστοιχη περίπτωση η PNN. Η αδυναμία της SVM εντοπίστηκε, σύμφωνα με την παραπάνω έρευνα, στην περίπτωση όπου το επίπεδο του θορύβου στα δεδομένα του πειραματικού σχεδιασμού μεγαλώνει. Στην περίπτωση αυτή η PNN αποδίδει σημαντικά μικρότερο ποσοστό εσφαλμένων ταξινομήσεων από ότι αποδίδει η SVM, χωρίς όμως να υπονοείται ότι σε μια τέτοια κατάσταση το αποτέλεσμα που θα παρουσιάσει θα θεωρείται αναξιόπιστο.

3. Τέλος, όσον αφορά τα υπόλοιπα υποδείγματα ταξινόμησης, η τετραγωνική διακριτική ανάλυση στους περισσότερους συνδυασμούς των κριτηρίων ταξινόμησης αποδίδει το μεγαλύτερο ποσοστό εσφαλμένων ταξινομήσεων, συγκρινόμενο πάντα με τα αντίστοιχα ποσοστά των υπολοίπων προσεγγίσεων ταξινόμησης που παρουσιάστηκαν στον παρόντα πειραματικό σχεδιασμό. Μια παρατήρηση που αφορά την μέθοδο αυτή και ίσως είναι χρήσιμη σε αντίστοιχες περιπτώσεις, είναι ότι το ποσοστό εσφαλμένων ταξινομήσεων στην περίπτωση

όπου η συσχέτιση των χαρακτηριστικών είναι υψηλή δεν επηρεάζεται από την ταξινόμηση των δεδομένων (βλ. Πίνακα 3.7). Όσον αφορά τη γραμμική διακριτική ανάλυση (LDA) και το λογιστικό υπόδειγμα πιθανότητας (LOGIT), το ποσοστό των εσφαλμένων ταξινομήσεων που αποδίδουν, στις περισσότερες των περιπτώσεων, έχει μεγάλες διακυμάνσεις εξαρτώμενο από τις ιδιαιτερότητες των κριτηρίων αξιολόγησης και δεν μπορεί να εξαχθεί ένα ασφαλές συμπέρασμα. Κάτω από κάποια συγκεκριμένα γνωρίσματα των κριτηρίων ταξινόμησης και με το σωστό συνδυασμό τους, το αποτέλεσμα της ταξινόμησης που παρουσιάζουν αυτές οι δυο τεχνικές ταξινόμησης μπορεί να θεωρηθεί αξιόπιστο.

ΚΕΦΑΛΑΙΟ 4ο

Συμπεράσματα και μελλοντικές κατευθύνσεις

Το πρόβλημα της ταξινόμησης παρουσίαζε ανέκαθεν αυξημένο ερευνητικό και πρακτικό ενδιαφέρον. Οι έρευνες για την αντιμετώπιση των προβλημάτων ταξινόμησης επικεντρώνονται στην εφαρμογή κατάλληλων τεχνικών για την ανάπτυξη υποδειγμάτων ταξινόμησης, τα οποία συνθέτουν όλες τις παραμέτρους του εκάστοτε εξεταζόμενου προβλήματος και παρουσιάζουν με σαφή τρόπο τόσο την κατηγορία στην οποία εντάσσονται, όσο και την επίδραση τους στην αξιολόγηση των εναλλακτικών δραστηριοτήτων και στις διαφοροποιήσεις που παρατηρούνται μεταξύ των κατηγοριών.

Η ανάπτυξη των υποδειγμάτων ταξινόμησης βασίστηκε σε στατιστικές προσεγγίσεις. Οι προσεγγίσεις αυτές συνέβαλαν στην κατανόηση της προβληματικής της ταξινόμησης, των χαρακτηριστικών και των ιδιαιτεροτήτων που παρουσιάζει, καθώς επίσης και των κανόνων που διέπουν τα αναπτυσσόμενα υποδείγματα ταξινόμησης. Οι συχνά, όμως, περιοριστικές υποθέσεις που διέπουν κάθε στατιστική ανάλυση, οδήγησε στην ανάπτυξη άλλων εναλλακτικών προσεγγίσεων, αξιοποιώντας

τα επιτεύγματα άλλων επιστημονικών πεδίων. Δίνεται έτσι η δυνατότητα σε κάθε αποφασίζων, με την ανάπτυξη των σύγχρονων τεχνικών ταξινόμησης, να αναλύσει ακόμα και τα πιο πολύπλοκα προβλήματα λήψης αποφάσεων τα οποία βασίζονται στην προβληματική της ταξινόμησης.

Η παρούσα διπλωματική εργασία επικεντρώθηκε στην παρουσίαση πέντε τεχνικών ταξινόμησης, τόσο στατιστικών όσο και μη παραμετρικών, και στην σύγκρισή τους ως προς την αποτελεσματικότητα τους. Στόχος της παραπάνω έρευνας και κατ' επέκταση και παρόμοιων ερευνών με αυτή, είναι να παρουσιαστεί το κατά πόσο τα υποδείγματα ταξινόμησης είναι ικανά να παρέχουν συνεχή και αξιόπιστη υποστήριξη προς τον αποφασίζοντα. Άρα η έρευνα ως προς την αποτελεσματικότητα των τεχνικών ταξινόμησης παρουσιάζει σημαντικές προοπτικές περαιτέρω έρευνας.

Οι κύριες από τις προοπτικές αυτές εντοπίζονται στη διερεύνηση με άλλες εναλλακτικές μεθοδολογίες, κυρίως από τον χώρο των πολυκριτήριων τεχνικών. Κύριος στόχος είναι η διερεύνηση εκείνων των συνθηκών υπό τις οποίες οι διάφορες προσεγγίσεις ταξινόμησης θα παρουσιάζουν παρόμοια αποτελέσματα. Απαραίτητος είναι και ο προσδιορισμός εκείνων των παραγόντων που θα συμβάλλουν στην υψηλότερη αποτελεσματικότητα των τεχνικών ταξινόμησης.

Βιβλιογραφία

[1] Belacel, N. (2000), "Multicriteria Assignment Method PROAFTN: Methodology and Medical Applications," European Journal of Operational research, 125, 175-183.

[2] Burges, C.J.C. (1998), "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery 2(2),121-167

[3] Catelani, M. and Ford, A. (2000), "Fault Diagnosis of Electronic Analog Circuits using a Radial Basis Function Network Classifier," Measurement 28(3), 147-158.

[4] 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.

[5] Δούμπος, Μ. και Ζοπουνίδης, Κ. (2001), «Πολυκριτήριες Τεχνικές Ταξινόμησεις: Θεωρία και Εφαρμογές», Εκδόσεις 'Κλειδάριθμος'.

[6] Doumpos, M. and Zopounidis, C. (1998), 'The Use of the Preference Disaggregation Analysis in the Assessment of Financial Risks," Fuzzy Economic Review 3(1),39-57.

[7] Dutka, A. (1995), "AMA Handbook of Customer Satisfaction: A Guide to Research, Planning and Implementation," NTC Publishing Group, Illinois.

[8] Fisher, R.A. (1936), "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 7, 179-188.

[9] Fung, G. and Mangasarian, O.L. (2001), "Proximal Support Vector Machine Classifiers," Data Mining Institute Technical Report 01-02, Association for Computing Machinery, New York,77-86.

[10] Gochet, W. Stam, A.; Srinivasan, V. and Chen, S. (1997), "Multigroup Discriminant Analysis using Linear Programming," Operations Research 45(2), 213-225.

[11] McFadden, D. (1974), "Conditional logit analysis in qualitative choice behavior", in: P. Zarembka (ed.), Frontiers in Econometrics, Academic Press, New York.

[12] McFadden, D. (1980), "Structural discrete probability models derived from the theories of choice", in: C.F. Manski and D. McFadden (eds.), Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, Mass.

[13] Mirkin, B. (1996), "Mathematical Classification and Clustering," Kluwer Academic Publishers, Dordrecht.

[14] Nieddu, L. and Patrizi, G. (2000), "Formal Methods in Pattern Recognition: A review," European Journal of Operational Research 120, 459-495.

[15] Pawlak, Z. (1982), "Rough sets", International Journal of Information and Computer Sciences,11,31-356.

[16] Ripley, B.D. (1996), "Pattern Recognition and Neural Networks," Cambridge University Press, Cambridge.

[17] Roy, B. (1985), Méthodologie Multicritère d'Aide à la Décision, Economica, Paris.

[18] Rulon, P.J. Tiedeman, D.V., Tatsuoka, M.M., and Langmuir, C.R. (1967), "Multivariate Statistics for Personnel Classification," Wiley, New York.

[19] 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.

[20] Shen, L., F.E.H., Qu, L. and Shen, Y. (2000), "Fault Diagnosis using Rough Sets Theory," Computer in Industry 43, 61-72.

[21] Smith, C. (1947), "Some examples of discrimination", Annals of Eugenics,13,272-282.

[22] Specht,D.F., (1988), "Probabilistic Neural Networks for Classification, mapping or Associative Memory " Proc. IEEE Int. Conf. Neural Networks,1,525-532.

[23] Specht,D.F., (1990), "Probabilistic Neural Networks" 1(3), 109-118.

[24] T. Joackims. Making large-scale SVM learning practical. In B. Scholkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods-Support Vector Learning,* pages 169-184. MIT Press, 1998.

[25] Tsumoto, S. (1998), "Automated Extraction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory," Information Sciences 112, 67-84.

[26] Vapnik, V.N. (1995), "The Nature of Statistical Learning Theory," Springer-Verlang, New York, USA.

[27] Young, T.Y., and Fu, K.S. (1997), "Handbook of Pattern Recognition and Image Processing," Hanbooks in Science and Technology, Academic Press, New York.

[28] Zadeh, L.A. (1965), "Fuzzy sets", Information and Control,8,338-353.

[29] Zopounidis, C. (1998), "Operational Tools in the Management of Financial Risks," Kluwer Academic Publishers, Dordrecht.