ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ
Τομέας Οργάνωσης και Διοίκησης

'Γενετικοί Αλγόριθμοι στην Ανάπτυξη Μηχανών Διανυσμάτων Υποστήριξης: Μεθοδολογία και Εφαρμογές στην Εκτίμηση Πιστωτικού Κινδύνου'

Διατριβή που υπεβλήθη για την μερική ικανοποίηση των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης υπό

Άννα Σάτσιου

2005

Στην γιαγιά μου Άννα, που δεν είναι πια μαζί μας...

Η διατριβή της Άννας Σάτσιου, εγκρίνεται

Καθηγητής Κωνσταντίνος Ζοπουνίδης

Λέκτορας Μιχάλης Δούμπος

Αναπληρωτής Καθηγητής Νικόλαος Ματσατσίνης

Ευχαριστίες

Με την ολοκλήρωση της παρούσας μεταπτυχιακής διατριβής αισθάνομαι την ανάγκη να ευχαριστήσω πρώτα από όλα τον Θεό, που με αξίωσε να φέρω εις πέρας άλλον έναν από τους στόχους μου.

Ολόψυχα ευχαριστώ και τους γονείς μου Ιωάννη και Μαρία Σάτσιου, καθώς και τους αδερφούς μου Κώστα και Βαγγέλη Σάτσιο που με στηρίζουν πάντα με την αγάπη τους και μου δίνουν δύναμη. Είναι οι άνθρωποι στους οποίους οφείλω κάθε μου επιτυχία.

Ευχαριστώ μέσα από την καρδιά μου όλους μου τους φίλους και φίλες από το Πολυτεχνείο Κρήτης, αλλά και αυτούς που βρίσκονται μακριά για την ψυχολογική στήριξη και υπέροχη φιλία τους.

Ευχαριστώ πολύ και τον Καθηγητή κ. Κωνσταντίνο Ζοπουνίδη καθώς και τον Αναπληρωτή Καθηγητή κ. Νικόλαο Ματσατσίνη που υπήρξαν καθηγητές μου σε μεταπτυχιακά μαθήματα για τις πολύτιμες γνώσεις και βοήθεια που μου παρείχαν.

Τέλος, οι πιο θερμές ευχαριστίες ανήκουν στον Λέκτορα κ. Μιχάλη Δούμπο, για το ενδιαφέρον του, τις πολύτιμες γνώσεις και συμβουλές του, καθώς και την βοήθεια και συμπαράσταση του κατά την εκπόνηση αυτής της εργασίας.

Περιεχόμενα

Κεφάλαιο 1

'Εισαγωγή'

1.1 Εισαγωγικά

Οι Μηχανές Διανύσματος Υποστήριξης (Support Vector Machines, SVM) έχουν αναπτυχθεί τα τελευταία χρόνια ως το σημαντικότερο μεθοδολογικό εργαλείο στη θεωρία της στατιστικής θεωρίας μάθησης (statistical learning theory) σε προβλήματα αναγνώρισης προτύπων (pattern recognition) και ταξινόμησης (classification) με σημαντικές πρακτικές εφαρμογές σε διάφορα πεδία.

Οι SVM διακρίνονται σε γραμμικές και μη γραμμικές. Στην απλή γραμμική περίπτωση στόχος είναι η ανάπτυξη ενός υπερεπιπέδου (hyperplane) το οποίο διαχωρίζει τις προκαθορισμένες κατηγορίες αντικειμένων. Η ανάπτυξη του υπερεπιπέδου πραγματοποιείται στο χώρο των εισόδων του συστήματος (input space) χρησιμοποιώντας τεχνικές βελτιστοποίησης έτσι ώστε να περιοριστεί το σφάλμα εκπαίδευσης και να μεγιστοποιηθεί το περιθώριο μεταξύ των κατηγοριών. Η μεγιστοποίηση του περιθωρίου συνδέεται άμεσα με τη δυνατότητα γενίκευσης (generalization) του αναπτυσσόμενου υποδείγματος. Η μη γραμμική περίπτωση αποτελεί γενίκευση των γραμμικών SVM. Στην περίπτωση αυτή πραγματοποιείται μια μη γραμμική αναπαράσταση (non-linear mapping) των δεδομένων σε ένα χώρο μεγαλύτερων διαστάσεων. Στόχος της αναπαράστασης αυτής είναι να καταστήσει τις κατηγορίες γραμμικά διαχωρίσιμες στο νέο χώρο (feature space) που δημιουργείται, επιτρέποντας έτσι τη χρήση των ίδιων τεχνικών βελτιστοποίησης που χρησιμοποιούνται στην απλή γραμμική περίπτωση για την ανάπτυξη μη γραμμικών υποδειγμάτων.

Η επιτυχημένη χρήση των δυνατοτήτων που παρέχουν οι SVM καθορίζεται σε σημαντικό βαθμό από τη σωστή επιλογή του τρόπου με τον οποίο θα γίνει η προαναφερθείσα μη γραμμική αναπαράσταση των δεδομένων, την επιλογή των χαρακτηριστικών (παραγόντων) που θα ενσωματωθούν στην ανάλυση, καθώς και τον καθορισμό των διαφόρων επιμέρους τεχνικών παραμέτρων της διαδικασίας βελτιστοποίησης. Οι μέχρι σήμερα έρευνες έχουν επικεντρωθεί στο θέμα της επιλογής χαρακτηριστικών (feature selection), ενώ τα άλλα δύο θέματα δεν έχουν εξεταστεί.

Στόχος της προτεινόμενης ερευνητικής εργασίας είναι να αντιμετωπίσει τα παραπάνω σημαντικά θέματα σε ένα ολοκληρωμένο πλαίσιο συμβάλλοντας έτσι ουσιαστικά στην ανάπτυξη μιας κατάλληλης μεθοδολογίας για τη βέλτιστη επιλογή των προαναφερθέντων παραμέτρων. Η προτεινόμενη μεθοδολογία βασίζεται στη χρήση γενετικών αλγορίθμων, οι οποίοι έχουν βρει ευρεία χρήση σε θέματα επιλογής χαρακτηριστικών καθώς και στη βελτιστοποίηση της αρχιτεκτονικής νευρωνικών δικτύων. Στον αντίποδα, δεν έχει μέχρι σήμερα διερευνηθεί η αξιοποίηση των δυνατοτήτων που παρέχουν οι γενετικοί αλγόριθμοι στον καθορισμό όλων των παραμέτρων των SVM.

Παράλληλα με την ανάπτυξη του μεθοδολογικού πλαισίου για τη χρήση των γενετικών αλγορίθμων στην ανάπτυξη υποδειγμάτων SVM, πραγματοποιείται και εφαρμογή της μεθοδολογίας σε δεδομένα σχετικά με την ανάπτυξη συστημάτων εκτίμησης του πιστωτικού κινδύνου (credit scoring models). Η ανάπτυξη τέτοιων συστημάτων αποτελεί ένα σημαντικό πεδίο έρευνας στο χώρο της χρηματοοικονομικής επιστήμης. Στόχος των συστημάτων εκτίμησης του πιστωτικού κινδύνου είναι η βαθμολόγηση των πελατών (επιχειρήσεις / ιδιώτες) μιας επιχείρησης ανάλογα με την ικανότητά τους να ανταποκριθούν στις οικονομικές τους υποχρεώσεις και η ταξινόμησή τους σε αντίστοιχες σαφώς προκαθορισμένες κατηγορίες. Η πολυπλοκότητα του προβλήματος του πιστωτικού κινδύνου καθώς και η ιδιαίτερη σημασία του για τα χρηματοπιστωτικά ιδρύματα καθιστά αναγκαία τη χρήση υποδειγμάτων ταξινόμησης όπως αυτά που διερευνώνται στην προτεινόμενη εργασία. Σημειώνεται μάλιστα πως ήδη κορυφαίοι διεθνείς οργανισμοί βαθμολόγησης, όπως η Standard & Poor's, χρησιμοποιούν τις SVM ως κύριο μεθοδολογικό εργαλείο στην ανάπτυξη των συστημάτων βαθμολόγησης που

παρέχουν. Η εφαρμογή της προτεινόμενης μεθοδολογίας στο χώρο αυτό οδηγεί στην εξαγωγή χρήσιμων συμπερασμάτων σχετικά με την αποτελεσματικότητά της και την πραγματοποίηση συγκριτικών αναλύσεων με άλλες προσεγγίσεις.

1.2 Δομή της Εργασίας

Η εργασία δομείται ως εξής. Στο κεφάλαιο 2 περιγράφονται οι Μηχανές Διανύσματος Υποστήριξης. Αρχικά, στην παράγραφο 2.1 γίνεται αναφορά στα διάφορα συστήματα μάθησης και στην αναγνώριση προτύπων. Στην παράγραφο 2.2 εισάγονται οι Μηχανές Διανύσματος Υποστήριξης και περιγράφεται το βέλτιστο υπερεπίπεδο στην παράγραφο 2.2.1 και το γενικευμένο βέλτιστο υπερεπίπεδο στην παράγραφο 2.2.2. Το κεφάλαιο κλείνει με μια εκτενή αναφορά στην εκτίμηση της ικανότητας γενίκευσης των μοντέλων ταξινόμησης.

Στο κεφάλαιο 3 παρουσιάζονται οι γενετικοί αλγόριθμοι. Στην παράγραφο 3.1 δίνονται κάποια εισαγωγικά στοιχεία, ενώ στην παράγραφο 3.2 περιγράφεται ο απλός γενετικός αλγόριθμος και οι έξι βασικοί παράγοντές του. Η παράγραφος 3.3 αναλύει το μαθηματικό υπόβαθρο των γενετικών αλγορίθμων και τέλος στην παράγραφο 3.4 θίγεται το θέμα της επιλογής χαρακτηριστικών με τη βοήθεια των γενετικών αλγορίθμων.

Το κεφάλαιο 4 αναφέρεται στα συστήματα εκτίμησης πιστωτικού κινδύνου. Στην παράγραφο 4.1 γίνεται η εισαγωγή του θέματος, στην παράγραφο 4.2 δίνεται μια ιστορική αναδρομή και αναφορά στα συστήματα εκτίμησης πιστωτικού κινδύνου, ενώ στην παράγραφο 4.3 περιγράφεται η προτεινόμενη μεθοδολογία. Στην ενότητα 4.3.1 γίνεται η περιγραφή των δεδομένων και στην ενότητα 4.3.2 περιγράφονται και αναλύονται τα αποτελέσματα της προτεινόμενης μεθοδολογίας συγκριτικά και με άλλες μεθόδους, όπως την διακριτική ανάλυση, την βηματική διακριτική ανάλυση, την λογιστική παλινδρόμηση και την βηματική λογιστική παλινδρόμηση.

Τέλος, το κεφάλαιο 5 εμπεριέχει τα γενικά συμπεράσματα που προέκυψαν από την ανάλυση και τα αποτελέσματα της προτεινόμενης μεθοδολογίας.

Κεφάλαιο 2

'Μηχανές Διανυσμάτων Υποστήριξης για Ταξινόμηση'

2.1 Αναγνώριση Προτύπων στη Στατιστική Θεωρία Εκμάθησης

Η ανάπτυξη συστημάτων μάθησης (Machine Learning) διακρίνεται στην εποπτευμένη (supervised) και μη εποπτευμένη εκμάθηση (unsupervised learning). Επιπλέον, η εποπτευμένη εκμάθηση διακρίνεται στην διορθωτική εκμάθηση (corrective learning) και στην εκμάθηση ενίσχυσης (reinforcement learning) (π.χ. [RusNor95, Rojas96]). Στην διορθωτική εκμάθηση, ο μαθητής έχει πλήρη πρόσβαση στις λαμβανόμενες παρατηρήσεις (observations) και στα επιτευχθέντα αποτελέσματα κάθε δράσης. Αυτό το είδος εκμάθησης θα μπορούσε να παρομοιαστεί με το μάθημα ενός δασκάλου που παρέχει στους μαθητές τους τις ασκήσεις και τις σωστές απαντήσεις. Στην εκμάθηση ενίσχυσης, ο μαθητής τροφοδοτείται μόνο με μια αξιολόγηση της κάθε δράσης του και όχι με τα σωστά αποτελέσματα. Αυτό το είδος εκμάθησης είναι συγκρίσιμο με το μάθημα ενός δασκάλου που δεν λέει τις σωστές απαντήσεις, αλλά δίνει μόνο τους βαθμούς για την επίδοση των μαθητών. Στην μη εποπτευμένη εκμάθηση, ο μαθητής δεν έχει καθόλου πρόσβαση στα αποτελέσματα των ενεργειών του, παρά μόνο μπορεί να μάθει τις σχέσεις μεταξύ των παρατηρήσεων.

Στη Στατιστική Θεωρία Εκμάθησης (Statistical Learning Theory, βλ. π.χ. [SchSmo02]) το ενδιαφέρον επικεντρώνεται κυρίως στην διορθωτική εκμάθηση. Επομένως, ο μαθητής λαμβάνει ένα σύνολο υποδειγμάτων εκπαίδευσης $\varepsilon = \{(x_i, y_i) \setminus x_i \in \mathrm{X}, y_i \in \Upsilon, i = 1,...,n\}$ όπου X είναι ένα μη κενό σύνολο προτύπων (περιπτώσεις (cases), είσοδοι (inputs), περιστατικά (instances) ή παρατηρήσεις (observations)) και Υ είναι ένα μη κενό σύνολο στόχων (έξοδοι (outputs)). Στην

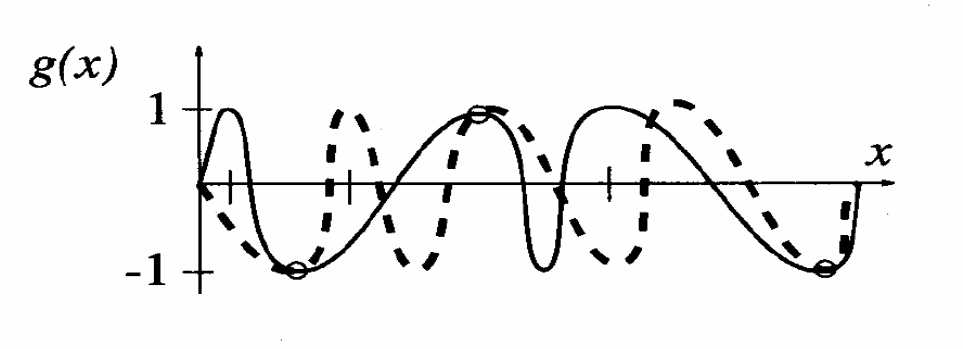
παρούσα εργασία το Υ θεωρείται σαν ένα πεπερασμένο σύνολο από επωνυμίες κλάσεων (class labels). Αυτό σημαίνει ότι ο μαθητής προσπαθεί να μάθει μία ταξινόμηση (classification) για τα πρότυπα εκπαίδευσης $x_1,...,x_n$. Η διαδικασία αυτή αναφέρεται συνήθως ως *αναγνώριση προτύπων (pattern recognition)*.

Η τυπική υπόθεση στη στατιστική θεωρία εκμάθησης είναι, ότι το σύνολο δεδομένων εκπαίδευσης ε παράγεται ανεξάρτητα από κάποια άγνωστη (αλλά συγκεκριμένη) κατανομή πιθανότητας *P(x,y)*. Ο στόχος είναι να κατασκευαστεί μια συνάρτηση αποφάσεων (decision function) ή αλλιώς μια υπόθεση (hypothesis) *f* βασισμένη στα δεδομένα εκπαίδευσης ε που θα ταξινομήσει σωστά όσο το δυνατό περισσότερα νέα δείγματα (*x,y*), τα οποία προέρχονται από την ίδια κατανομή *P(x,y)*, έτσι ώστε *f(x)= y*. Αυτό καλείται γενίκευση (generalization).

Στόχος λοιπόν μιας τέτοιας ανάλυσης είναι ο προσδιορισμός της συνάρτησης *f* από τα δείγματα εκπαίδευσης έτσι ώστε η υπόθεση που μαθαίνεται από αυτά τα δείγματα να μπορεί να γενικευτεί για οποιοδήποτε νέο δείγμα που προέρχεται από την ίδια κατανομή. Για αυτόν το λόγο πρέπει να χρησιμοποιηθεί μια αρχή επαγωγής (induction principle). Ωστόσο, το πρόβλημα είναι ότι η συνάρτηση *f* θα μπορούσε να επιλεχτεί από οποιαδήποτε κλάση συναρτήσεων, ενώ το μόνο που είναι γνωστό είναι το σύνολο ε. Θα μπορούσε συνεπώς εύκολα να επιλεγεί μια συνάρτηση *f* έτσι ώστε να ικανοποιεί απόλυτα όλα τα δεδομένα εκπαίδευσης (π.χ. σε ένα μονοδιάστατο πρόβλημα ταξινόμησης, μια πολυωνυμική συνάρτηση βαθμού n-1 δεδομένου ότι $x_i \neq x_j$ για $y_i \neq y_j$ για όλα τα $i,j \in \{1,...,n\}$). Αυτό σημαίνει ότι θα ισχύει $f(x_i) = y_i$ για ι = 1,...n, αλλά δεν εγγυάται ότι θα ισχύει *f* (*x*)= *y* για κάποιο νέο δείγμα (*x,y*)

Τα παραπάνω γίνονται πιο κατανοητά από το σχήμα 2.1, όπου παρουσιάζεται ένα μονοδιάστατο πρόβλημα ταξινόμησης με ένα σύνολο από τρία σημεία εκπαίδευσης που είναι σημειωμένα με κύκλους και σημεία ελέγχου που είναι σημειωμένα με κάθετες παύλες στον άξονα x. Ο στόχος του προβλήματος είναι η διάκριση της κλάσης -1 από την κλάση 1. Η ταξινόμηση πραγματοποιείται με ένα κατώφλι σε μία πραγματική συνάρτηση *g* έτσι ώστε *f(x)=sgn(g(x))*. Και οι δύο συναρτήσεις (διακεκομμένη και μη διακεκομμένη γραμμή) εξηγούν απόλυτα τα δεδομένα εκπαίδευσης, αλλά δίνουν αντίθετες προβλέψεις για τα δείγματα ελέγχου.

Επομένως, χωρίς επιπρόσθετη πληροφορία, δεν μπορεί να αποφασιστεί ποια συνάρτηση θα χρησιμοποιηθεί.



Σχήμα 2.1: Ένα μονοδιάστατο πρόβλημα ταξινόμησης (πηγή: [SchSmo02])

Προκειμένου να γίνουν τα παραπάνω μαθηματικά εμπεριστατωμένα, θεωρούμε μία μη αρνητική συνάρτηση απωλειών $\ell : \mathrm{X} \times \Upsilon \times \Upsilon \to [0, \infty]$ με την ιδιότητα $\ell(x, y, y) = 0$ για όλα τα $x \in \mathrm{X}, y \in \Upsilon$ η οποία δίνει κάποιο μέτρο για το λάθος που γίνεται προβλέποντας κάποιο δείγμα (x,y) (από την συνάρτηση κατανομής $P(x,y)$) με την βοήθεια της συνάρτησης f [SchSmo02]. Η απλούστερη επιλογή για μια συνάρτηση απωλειών θα ήταν

$$\ell(x, y, f(x)) = \begin{cases} 0 & \textit{εάν } y = f(x) \\ 1 & \textit{διαφορετικά} \end{cases} \qquad (2.1)$$

Μία σημαντική ιδιότητα της Στατιστικής Θεωρίας Εκμάθησης είναι ότι η ελαχιστοποίηση του μέσου λάθους εκπαίδευσης (average training error) ή αλλιώς εμπειρικού ρίσκου (empirical risk), [Vapnik98]

$$R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i, f(x_i)) \qquad (2.2)$$

δεν συνεπάγεται και ένα μικρό αναμενόμενο λάθος ή ρίσκο γενίκευσης (generalization error/risk) σε όλα τα δυνατά πρότυπα που λαμβάνονται από την κατανομή $P(x,y)$.

$$R[f] = \int_{x \times y} \ell(x, y, f(x)) dP(x, y) \qquad (2.3)$$

Σημειώνεται ότι η εξίσωση (2.3) δεν είναι πρακτικά υπολογίσιμη, εφόσον δεν είναι γνωστή η κατανομή *P(x,y)*. Ο όρος υπερπροσαρμογή (overfitting) αναφέρεται στην κατάσταση όπου το R_{emp} είναι πολύ μικρό έως και μηδενικό, ενώ το *R* είναι πολύ υψηλό. Αυτό σημαίνει ότι η συνάρτηση αποφάσεων *f(x)* καλύπτει πολύ καλά όλες τις ειδικές περιπτώσεις που παρέχονται από τα υποδείγματα εκπαίδευσης του συνόλου ε, αλλά δεν είναι αρκετά γενική για να καλύψει άλλα δείγματα. Επιπλέον θα πρέπει να αποφεύγεται μια συνάρτηση *f(x)* που να είναι τόσο ασαφής ώστε να μη μπορεί να εξηγήσει επαρκώς ούτε τα δείγματα εκπαίδευσης ούτε τα δείγματα ελέγχου, δηλαδή να εισάγει έναν υψηλό εμπειρικό κίνδυνο (υποπροσαρμογή (underfitting)). Για να το αποφύγει αυτό ο Vapnik [Vapnik95, Vapnik98] πρότεινε να περιοριστεί το σύνολο συναρτήσεων C από το οποίο επιλέγεται η συνάρτηση *f(x,y)* σε ένα σύνολο που να έχει μια κατάλληλη για τον όγκο των διαθέσιμων δεδομένων δυναμικότητα (capacity).

Μια από τις πιο γνωστές έννοιες δυναμικότητας είναι η αποκαλούμενη VC (Vapnik- Chervonenkis) διάσταση (βλ. [Vapnik98, SchSmo02, Vapnik95]). Η VC διάσταση ορίζεται ως ο μεγαλύτερος αριθμό *l* σημείων εκπαίδευσης ενός συνόλου που μια δεδομένη κλάση συναρτήσεων μπορεί να διαχωρίσει (shatter). Εάν δεν υπάρχει τέτοιος αριθμός *l*, τότε η VC διάσταση είναι άπειρη. Στην παρούσα εργασία θεωρείται ότι ένα σύνολο *l* σημείων διαχωρίζεται από μια κλάση συναρτήσεων C, εάν μπορούν να πραγματοποιηθούν όλοι οι πιθανοί διαχωρισμοί των σημείων, που προκύπτουν από οποιουσδήποτε πιθανούς συνδυασμούς επωνυμίας των προτύπων, με τη βοήθεια μιας συνάρτησης από το C. Για παράδειγμα εάν υπάρχουν δύο κλάσεις + 1 και -1, τότε υπάρχουν 2^3 πιθανοί τρόποι να επονομαστούν τρία πρότυπα εκπαίδευσης. Υποθέτοντας ότι όλα τα πρότυπα εκπαίδευσης βρίσκονται στο R^2 και είναι μη συγγραμικά, τότε κάθε πιθανός διαχωρισμός μπορεί να πραγματοποιηθεί με τη βοήθεια ενός υπερεπιπέδου. Αυτό σημαίνει ότι η κλάση των υπερεπιπέδων μπορεί να διαχωρίσει τρία σημεία, ωστόσο δεν μπορεί να διαχωρίσει τέσσερα, ανεξάρτητα από το πως είναι τοποθετημένα. Επομένως η VC διάσταση της κλάσης των υπερεπιπέδων στο R^2 είναι τρία. Μια κλάση συναρτήσεων είναι αφομοιώσιμη (learnable) μόνο, εάν η VC διάστασή της είναι πεπερασμένη.

Έστω ότι $h < n$ είναι η VC διάσταση της κλάσης των συναρτήσεων που μπορεί να υλοποιήσει η μαθητευόμενη μηχανή. Τότε για όλες τις συναρτήσεις της συγκεκριμένης κλάσης, ανεξάρτητα από την κατανομή *P(x,y)* και με μία πιθανότητα

τουλάχιστον *1-δ* για κάθε σχεδιασμό του δείγματος εκπαίδευσης ισχύει το παρακάτω όριο [Vapnik98,Vapnik95]:

$$R[f] \le R_{emp}[f] + \phi(h,n,\delta) \qquad (2.4)$$

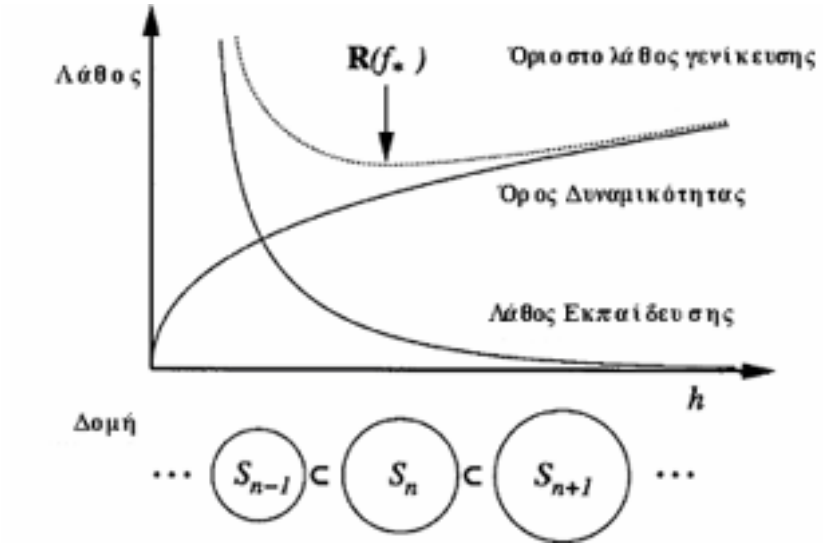
όπου ϕ είναι ένας όρος εμπιστοσύνης ή δυναμικότητας (confidence/ capacity term) που ορίζεται ως

$$\phi(h,n,\delta) = \sqrt{\frac{1}{n}(h(\ln\frac{2n}{h}+1)+\ln\frac{4}{\delta})} \qquad (2.5)$$

Για κάθε σύνολο πεπερασμένων δειγμάτων ε μπορεί πάντα να βρεθεί μια συνάρτηση *f* έτσι ώστε το λάθος εκπαίδευσης να είναι μηδενικό $R_{emp}[f] = 0$ (υπό τον όρο ότι τα δείγματα δεν είναι αντικρουόμενα). Το παραπάνω ισχύει ακόμη και εάν όλα τα πρότυπα εκπαίδευσης *x* και όλες οι επωνυμίες *y* είναι στατιστικά ανεξάρτητα μεταξύ τους, δηλαδή ισχύει *P(x,y)=P(x)P(y),* και τα *y* είναι ισοπίθανα. Σε αυτήν την περίπτωση, δεν μπορεί να γίνει μια καλή πρόβλεψη για την επωνυμία ενός προτύπου εκπαίδευσης. Ωστόσο, προκειμένου το λάθος εκπαίδευσης να είναι μηδενικό, είναι απαραίτητη η απαίτηση για μια μεγάλη VC διάσταση *h*. Δεδομένου ότι ο όρος εμπιστοσύνης (2.5) αυξάνεται μονότονα με τη VC διάσταση αυτό θα οδηγήσει σε έναν μεγάλο λάθος γενίκευσης *R*, δηλαδή σε υπερπροσαρμογή. Αυτό σημαίνει ότι η κλάση των συναρτήσεων, από την οποία λαμβάνεται το *f*, πρέπει να περιοριστεί έτσι ώστε η δυναμικότητά της (π.χ. VC διάσταση) να είναι αφενός όσο το δυνατόν μικρότερη (σε σχέση με το διαθέσιμο όγκο δεδομένων) για να αποδίδει καλά ως προς την γενίκευση (να αποφύγει υπερπροσαρμογή) και αφ' ετέρου αρκετά μεγάλη για να μοντελοποιήσει τις εξαρτήσεις που κρύβονται στα δεδομένα.

Ως εκ τούτου, ο Vapnik και Chervonenkis [Vapnik79, VapChe74] πρότειναν την ελαχιστοποίηση του δεξιού μέλους της ανισότητας (2.4), παρά απλά την ελαχιστοποίηση του εμπειρικού κινδύνου. Αυτό οδηγεί στην αρχή της Δομικής Ελαχιστοποίησης Κινδύνου (Structural Risk Minimization). Η κύρια ιδέα είναι να δημιουργηθεί μία ακολουθία από κλάσεις συναρτήσεων (ή δομές) $S_1 \subset S_2 \subset ... \subset S_h \subset ...$ αυξανόμενου μεγέθους (και επομένως, αυξανόμενης

δυναμικότητας) και να ελαχιστοποιηθεί το δεξιό μέλος της (2.4) με την επιλογή της κατάλληλης δομής. Κατά αυτόν τον τρόπο επιλέγεται μια συνάρτηση f_* που παρουσιάζει μικρό λάθος εκπαίδευσης και είναι στοιχείο μιας δομής που έχει χαμηλή δυναμικότητα h (σχήμα 2.2).



Σχήμα 2.2: Αρχή Δομικής Ελαχιστοποίησης Κινδύνου (πηγή: [SchSmo02])

Σημειώνεται ότι η εξίσωση (2.5) εξαρτάται επίσης από τον όγκο n των διαθέσιμων δεδομένων. Όσο μεγαλύτερη είναι η VC διάσταση h της κλάσης των συναρτήσεων C που λαμβάνεται υπόψη τόσο μεγαλύτερος είναι ο αριθμός των υποδειγμάτων εκπαίδευσης που χρειάζονται για να αποφευχθεί η υπερπροσαρμογή. Αφ' ετέρου, εάν έχουμε έναν απεριόριστο αριθμό υποδειγμάτων εκπαίδευσης, μπορεί να αποφευχθεί η υπερπροσαρμογή παίρνοντας απλά κάποια κλάση συναρτήσεων C με πεπερασμένη VC διάσταση.

Πρέπει να αναφερθεί ότι η προσέγγιση της Στατιστικής Θεωρίας Εκμάθησης που παρουσιάζεται εδώ δεν είναι η μοναδική στο πρόβλημα της αναγνώρισης προτύπων. Μια άλλη στατιστική προσέγγιση είναι για παράδειγμα η Bayesian εκμάθηση. Κατά την Bayesian άποψη η εκμάθηση δεν είναι τίποτα περισσότερο από ένα υποπρόβλημα του πιο θεμελιώδους προβλήματος των προβλέψεων [RusNor95, σελ. 588]. Η ιδέα είναι να χρησιμοποιηθούν οι υποθέσεις (συναρτήσεις απόφασης) ως μεσάζοντες μεταξύ των δεδομένων και της πρόβλεψης. Πρώτα υπολογίζεται η πιθανότητα κάθε υπόθεσης, λαμβάνοντας υπόψη τα δεδομένα. Κατόπιν οι προβλέψεις

γίνονται από τις υποθέσεις χρησιμοποιώντας τις μεταγενέστερες (posterior) πιθανότητές τους. Αυτό σημαίνει ότι οι προβλέψεις σταθμίζονται από την πιθανότητα της υπόθεσης.

Μία τελείως διαφορετική προσέγγιση προήλθε από τον χώρο της τεχνητής νοημοσύνης (Artificial Intelligence). Εδώ ο στόχος είναι να βρεθούν οι κανόνες (rules) πίσω από τα δεδομένα που μπορούν να αναπαρασταθούν με ένα συμβολικό τρόπο. Αυτό σημαίνει ότι η υπόθεση που 'μαθαίνεται' μπορεί να θεωρηθεί ως γνώση που μπορεί να κοινοποιηθεί στα ανθρώπινα όντα με έναν συμβολικό τρόπο. Ένα παράδειγμα αυτής της προσέγγισης είναι η μάθηση με τη βοήθεια των δέντρων απόφασης (decision trees) [RusNor95, σελ. 531]. Σε αντίθεση με αυτήν την προσέγγιση, οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) καθώς επίσης και τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) αναπαριστούν την υπόθεση με έναν 'υποσυμβολικό' τρόπο.

Επιπλέον, αντί κάποιος να προσπαθήσει να συμπεράνει αμέσως μια υπόθεση από όλα τα δεδομένα εκπαίδευσης με τη βοήθεια μιας αρχής επαγωγής, θα μπορούσε να ξεκινήσει με μια απλή υπόθεση και βαθμιαία να την βελτιώνει με κάθε νέο δείγμα εκπαίδευσης. Αυτή η προσέγγιση καλείται επαυξητική εκμάθηση (incremental learning) [RusNor95]. Ένα παράδειγμα μιας τέτοιας προσέγγισης είναι ο version-space αλγόριθμος εκμάθησης όπως περιγράφεται για παράδειγμα στο [RusNor95].

2.2 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs)

2.2.1 Βέλτιστο Υπερεπίπεδο

Στην παράγραφο 2.1 περιγράφτηκε η σημαντικότητα του ελέγχου της δυναμικότητας για οποιοδήποτε αλγόριθμο εκμάθησης. Επομένως το ζητούμενο είναι να βρεθεί μια κλάση συναρτήσεων των οποίων η δυναμικότητα να μπορεί να υπολογιστεί. Ο Vapnik et al. [VapLer63, VapChe74, VapChe79] θεώρησαν την κλάση των υπερεπιπέδων σε κάποιο διανυσματικό χώρο **H** με ένα ορισμένο εσωτερικό γινόμενο $\prec .,. \succ$:

$$\{\mathbf{x} \in \mathbf{H} \backslash \prec \mathbf{w}, \mathbf{x} \succ + b = 0\}, \mathbf{w} \in \mathbf{H}, b \in R \qquad (2.6)$$

Δεδομένου ότι οποιοδήποτε υπερεπίπεδο διαιρεί το χώρο σε δύο ημιδιαστήματα (και προκαλεί έτσι μία ταξινόμηση σε μια κλάση -1 και μια κλάση + 1) αυτό αντιστοιχεί στις συναρτήσεις απόφασης

$$f(\mathbf{x}) = \text{sgn}(\prec \mathbf{w}, \mathbf{x} \succ + b) \qquad (2.7)$$

Το περιθώριο (margin) ενός υπερεπιπέδου είναι η απόσταση μεταξύ του υπερεπιπέδου και του σημείου που είναι πλησιέστερα σε αυτό. Το σημαντικό είναι ότι μπορεί να αποδειχτεί ότι η δυναμικότητα (π.χ. η VC διάσταση) της κλάσης των υπερεπιπέδων με ένα δεδομένο περιθώριο μειώνεται όσο αυξάνεται το περιθώριο (π.χ. [SchSmo02, σελ. 142]). Εξετάζοντας την παράγραφο 2.1 διαφαίνεται ότι όσο μεγαλύτερο είναι το περιθώριο τόσο μικρότερος γίνεται ο δεξιός όρος της εξίσωσης (2.4). Αυτό είναι ακριβώς το επιζητούμενο. Με τη μεγιστοποίηση του περιθωρίου ελαχιστοποιούμε τον δομικό κίνδυνο. Ως εκ τούτου, ο Vapnik et al. πρότειναν να βρεθεί το αποκαλούμενο βέλτιστο υπερεπίπεδο που επιφέρει το μέγιστο περιθώριο με το διαχωρισμό δύο κλάσεων. Αυτό το βέλτιστο υπερεπίπεδο είναι μοναδικό και μπορεί να κατασκευαστεί με τον ακόλουθο τρόπο:

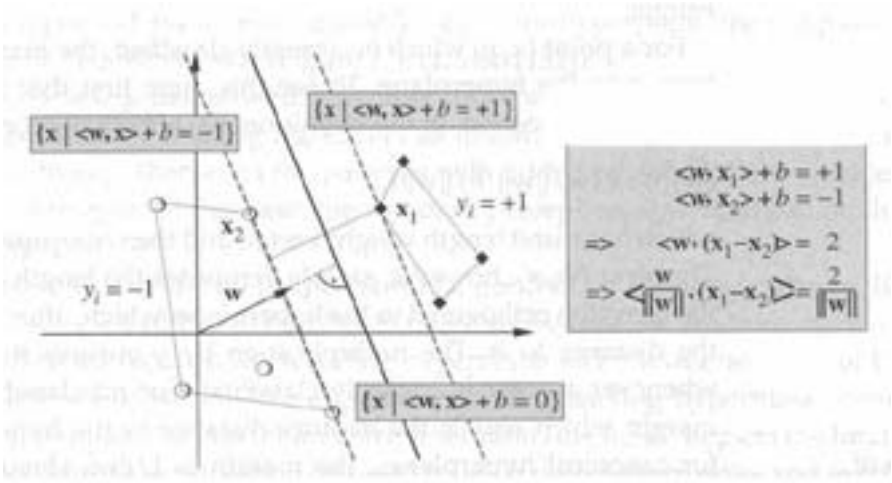
Έστω $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n) \in \mathbf{H} \times \{-1, 1\}$ είναι ένα σύνολο από υποδείγματα εκπαίδευσης. Τότε οποιοδήποτε υπερεπίπεδο της μορφής (2.6) μπορεί να κανονικοποιηθεί κατάλληλα έτσι ώστε

$$\min_{i=1,...,n} |\prec \mathbf{w}, \mathbf{x} \succ + b| = 1$$

που σημαίνει ότι το σημείο που βρίσκεται πλησιέστερα στο υπερεπίπεδο έχει μια απόσταση ίση με $1/\|\mathbf{w}\|$.

Στο σχήμα 2.3 παρουσιάζεται ένα δυαδικό πρόβλημα ταξινόμησης που είναι ο διαχωρισμός των σφαιρών από τα διαμάντια. Το βέλτιστο υπερεπίπεδο παρουσιάζεται με τη μη διακεκομμένη γραμμή. Δεδομένου ότι το πρόβλημα είναι γραμμικά διαχωρίσιμο, υπάρχει ένα διάνυσμα βάρους **w** και ένα κατώφλι b έτσι ώστε $y_i(\prec \mathbf{w}, \mathbf{x}_i \succ + b) > 0$, (i=1,...,n). Μεταβάλλοντας το **w** και b έτσι ώστε τα σημεία που είναι πιο κοντά στο υπερεπίπεδο να ικανοποιούν την $|\prec \mathbf{w}, \mathbf{x}_i \succ + b| = 1$, λαμβάνουμε

μια κανονικοποιημένη μορφή του υπερεπιπέδου που ικανοποιεί την $y_i(\prec \mathbf{w}, \mathbf{x}_i \succ +b) \geq 1$. Αυτό σημαίνει ότι το περιθώριο σε αυτήν την περίπτωση είναι ίσο με 1/||**w**||. Αυτό μπορεί να φανεί θεωρώντας δύο σημεία $\mathbf{x}_1, \mathbf{x}_2$ στις αντίθετες πλευρές του υπερεπιπέδου, οι οποίες ικανοποιούν ακριβώς την $|\prec \mathbf{w}, \mathbf{x}_i \succ +b| = 1$, και προβάλλοντας τα επάνω στο κανονικοποιημένο διάνυσμα του υπερεπιπέδου **w**/||**w**||.



Σχήμα 2.3 : Ένα δυαδικό πρόβλημα ταξινόμησης (πηγή: [SchSmo02])

Θεωρώντας αυτήν την κανονικοποίηση, το βέλτιστο υπερεπίπεδο μπορεί να κατασκευαστεί με την επίλυση του προβλήματος βελτιστοποίησης (βλ. [SchSmo02])

$$\min_{\mathbf{w}\in\mathbf{H}, b\in R} \tau(\mathbf{w}) = \frac{1}{2} \| \mathbf{w} \|^2 \qquad (2.8)$$

υπό τους περιορισμούς $y_i(\prec \mathbf{x}_i, \mathbf{w} \succ +b) \geq 1$ για κάθε i = 1,...,n

Αυτό είναι γνωστό ως το πρωτεύον πρόβλημα βελτιστοποίησης (primal optimization problem). Τα υπό περιορισμούς προβλήματα βελτιστοποίησης όπως αυτό, αποτελούν το αντικείμενο της θεωρίας βελτιστοποίησης. Η διαδικασία επίλυσης διευκολύνεται διατυπώνοντας το δυϊκό πρόβλημα :

$$\max_a W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j=1}^{n} a_i a_j y_i y_j \prec \mathbf{x}_i, \mathbf{x}_j \succ \qquad (2.9)$$

υπό τους περιορισμούς $a_i \geq 0$ για κάθε i=1,...,n, και $\sum_{i=1}^{n} a_i y_i = 0$

όπου $a = (a_1, ...a_n)$ είναι οι πολλαπλασιαστές Langrange. Η λύση του προβλήματος αυτού δίνει

$$\mathbf{w} = \sum_{i=1}^{n} a_i y_i \mathbf{x}_i \qquad (2.10)$$

οπότε η συνάρτηση ταξινόμησης έχει την ακόλουθη μορφή :

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{n} a_i y_i \prec \mathbf{x}, \mathbf{x}_i \succ + b) \qquad (2.11)$$

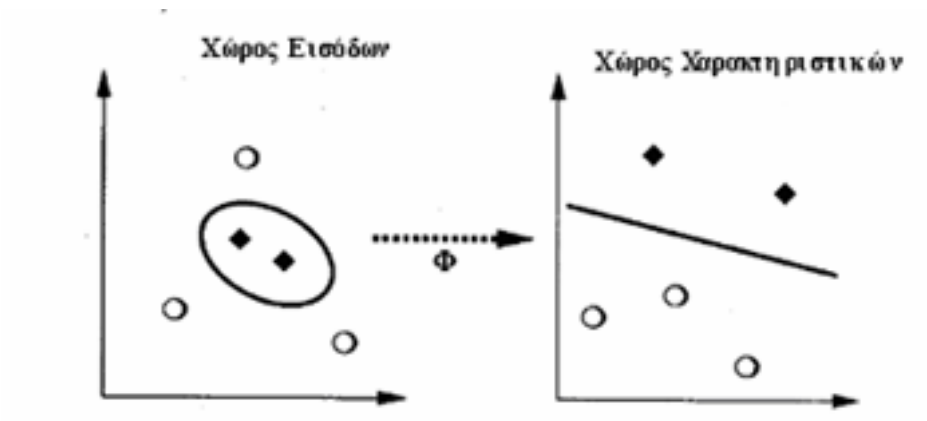
Προφανώς το **w** εξαρτάται μόνο από εκείνα τα διανύσματα εκπαίδευσης $\mathbf{x}_i$, για τα οποία το αντίστοιχα a_i είναι μη μηδενικά. Αυτά τα πρότυπα καλούνται διανύσματα υποστήριξης (support vectors). Όλα τα υπόλοιπα δείγματα δεν σχετίζονται με τη συνάρτηση απόφασης. Όλα τα διανύσματα υποστήριξης βρίσκονται ακριβώς στο περιθώριο, δεδομένου ότι μπορούν να ικανοποιήσουν μόνο την $y_i(\prec \mathbf{x}_i, \mathbf{w} \succ + b) = 1$. Η διαπίστωση αυτή συμβαδίζει με την διαίσθησή μας για το πρόβλημα: Εφόσον το υπερεπίπεδο καθορίζεται πλήρως από τα σημεία εκπαίδευσης που βρίσκονται πιο κοντά σε αυτό, η λύση δεν εξαρτάται από άλλα δείγματα.

2.2.2 Γενικευμένο Βέλτιστο Υπερεπίπεδο

Μέχρι τώρα η παρουσίαση εστίασε μόνο στην περίπτωση όπου τα δεδομένα εκπαίδευσης είναι πλήρως γραμμικά διαχωρίσιμα από ένα υπερεπίπεδο. Ωστόσο, υπάρχουν προβλήματα τα οποία δεν είναι γραμμικά διαχωρίσιμα (π.χ. το πρόβλημα XOR [Rojas96, σελ. 62]). Εάν θέλουμε και σε αυτήν την περίπτωση να χρησιμοποιήσουμε υπερεπίπεδα βέλτιστου περιθωρίου, τα οποία είναι ελκυστικά λόγω των διαθέσιμων ορίων δυναμικότητας, χρειάζεται μια μέθοδος για να αυξήσει την δυναμικότητα του συνόλου των συναρτήσεων απόφασης. Η ιδέα είναι να δημιουργηθεί μια αντιστοίχηση $\phi : \mathbf{X} \to \mathbf{H}$ από το σύνολο των εισόδων **X** στο χώρο του εσωτερικού γινομένου **H** (επίσης αποκαλούμενος ως χώρος χαρακτηριστικών (feature space)), έτσι ώστε τα δεδομένα που βρίσκονται στο **H** να είναι γραμμικά διαχωρίσιμα.

Για να επιτευχθεί ο παραπάνω στόχος, μπορεί να χρησιμοποιηθεί μια συνάρτηση $k : \mathbf{X} \times \mathbf{X} \to R$ έτσι ώστε $k(x, x') = \prec \phi(x), \phi(x') \succ$. Αυτή η συνάρτηση καλείται συνάρτηση πυρήνα (kernel function). Χρησιμοποιώντας αυτήν την συνάρτηση αποφεύγεται ο ρητός υπολογισμός της αντιστοίχησης φ. Αντί για αυτό, ορίζεται ένα εσωτερικό γινόμενο στον χώρο των χαρακτηριστικών με τη βοήθεια της συνάρτησης πυρήνα. Ο πυρήνας μπορεί να ερμηνευθεί ως μέτρο ομοιότητας μεταξύ των διανυσμάτων $\phi(x)$ και $\phi(x')$ στον χώρο **H**. Αυτό γίνεται πιο κατανοητό εάν θεωρήσουμε το κανονικό εσωτερικό γινόμενο μεταξύ δύο διανυσμάτων $\mathbf{x} = (x_1, ..., x_d)^{\mathrm{T}}, \mathbf{y} = (y_1, ..., y_d)^{\mathrm{T}} \in R^d$ $\prec \mathbf{x}, \mathbf{y} \succ = \sum_{i=1}^{d} x_i y_i$. Τότε το $\prec \mathbf{x}, \mathbf{y} \succ$ είναι ανάλογο προς το συνημίτονο της γωνίας μεταξύ του ***x*** και του ***y*** και έτσι αποτελεί έναν δείκτη της ομοιότητας μεταξύ των δύο διανυσμάτων. Η σκοπιμότητα των πυρήνων είναι να γενικευτεί αυτή η θεώρηση, πρώτα αντιστοιχώντας τα δεδομένα εκπαίδευσης σε ένα μεγαλύτερων διαστάσεων χώρο χαρακτηριστικών και έπειτα καθορίζοντας κάποιο εσωτερικό γινόμενο (και επομένως μέτρο ομοιότητας) σε εκείνο το χώρο.

Στο σχήμα 2.4 φαίνεται αυτή ακριβώς η θεώρηση των πυρήνων. Τα δεδομένα εκπαίδευσης αντιστοιχίζονται σε ένα μεγαλύτερων διαστάσεων χώρο χαρακτηριστικών (feature space), έτσι ώστε τα αντιστοιχούμενα δεδομένα να είναι γραμμικά διαχωρίσιμα με ένα μεγίστου περιθωρίου υπερεπίπεδο σε αυτό το χώρο. Αυτό προκαλεί ένα μη γραμμικό όριο απόφασης στον αρχικό χώρο εισόδων (input space). Με την χρήση μιας συνάρτησης πυρήνων, είναι δυνατό να υπολογιστεί το υπερεπίπεδο χωρίς να υπολογιστεί η αντιστοίχηση φ.



Σχήμα 2.4: Η θεώρηση των πυρήνων (πηγή:[SchSmo02])

Το ερώτημα που τίθεται είναι πώς αυτό μπορεί να γίνει στην πράξη. Πρέπει με κάποιο τρόπο να ελεγχθεί εάν κάποια συνάρτηση που έχει επιλεγεί αντιστοιχεί στην πραγματικότητα σε έναν πυρήνα και επομένως έμμεσα σε ένα χώρο χαρακτηριστικών ή όχι. Λόγω των ιδιοτήτων του εσωτερικού γινομένου οποιαδήποτε συνάρτηση *k* που είναι πυρήνας πρέπει να είναι συμμετρική. Κατά συνέπεια η μήτρα Gram ή μήτρα πυρήνων (Gram/kernel matrix) $\mathbf{K} = (k(x_i, x_j))_{ij}$ πρέπει να είναι επίσης συμμετρική. Επιπλέον μπορεί να αποδειχτεί [CrisSha00] ότι το *k* είναι μια συνάρτηση πυρήνα (δηλαδή ότι ικανοποιεί την $k(x,x') = \prec \phi(x), \phi(x') \succ$) εάν και μόνο εάν η μήτρα **K** είναι θετικά ορισμένη (δηλαδή έχει μη αρνητικά ιδιοδιανύσματα).

Οι πιο γνωστές συναρτήσεις πυρήνων είναι για παράδειγμα οι πυρήνες πολυώνυμων βαθμού g > 0 $k(x,x') = (\prec x, x' \succ + c)^g, c \in R$, ή οι ακτινωτές συναρτήσεις βάσης (RBF) με εύρος $\sigma > 0$ $k(x,x') = \exp(-\frac{\| x - x' \|^2}{2\sigma^2})$. Αυτές οι συναρτήσεις πυρήνων απαιτούν το X να είναι ένας διανυσματικός χώρος εσωτερικού γινομένου, αλλά υπάρχουν επίσης και πυρήνες για μη διανυσματικούς χώρους, παραδείγματος χάρη για στοιχειοσειρές (strings) (βλ. π.χ. [LesEskNob02]). Είναι επίσης δυνατό να δημιουργηθούν νέοι πυρήνες από τους ήδη υπάρχοντες. Εάν για παράδειγμα k_1 και k_2 είναι πυρήνες στο R^n, τότε και *ο* $k = k_1 + k_2$ είναι ένας πυρήνας. Αυτό μπορεί να φανεί εύκολα με την εξέταση των αντίστοιχων Gram μητρών $\mathbf{K}_1$, $\mathbf{K}_2$ και κάποιου διανύσματος $\mathbf{v} \in R^n$. Η μήτρα $\mathbf{K}_1$,+$\mathbf{K}_2$ είναι θετικά ορισμένη διότι $\mathbf{v}^{\mathbf{T}}(\mathbf{K}_1 + \mathbf{K}_2)\mathbf{v} = \mathbf{v}^{\mathbf{T}}\mathbf{K}_1\mathbf{v} + \mathbf{v}^{\mathbf{T}}\mathbf{K}_2\mathbf{v} \geq 0$.

Η έρευνα έχει κινηθεί και στην κατεύθυνση της εκτίμησης του πυρήνα απευθείας από τα δεδομένα εκπαίδευσης (π.χ. [CriShaEliKan02]). Αυτό γίνεται με τη βελτιστοποίηση της *ευθυγράμμισης (alignment)* $\frac{\prec \mathbf{K}, \mathbf{y}\mathbf{y}^{\mathbf{T}} \succ_F}{\sqrt{\prec \mathbf{K}, \mathbf{K} \succ_F \prec \mathbf{y}\mathbf{y}^{\mathbf{T}}, \mathbf{y}\mathbf{y}^{\mathbf{T}} \succ_F}}$ της

άγνωστης μήτρας πυρήνων **K** με τη μήτρα $\mathbf{yy}^{\mathbf{T}}$, όπου $\mathbf{y} \subset Y^n$ είναι το διάνυσμα με τις επωνυμίες των κλάσεων για τα δεδομένα εκπαίδευσης[1].

Επιστρέφοντας στα βέλτιστα υπερεπίπεδα περιθωρίου που εξετάστηκαν στην παράγραφο 2.2.1, η ιδέα είναι να αντικατασταθεί το εσωτερικό γινόμενο της εξίσωσης (2.9) από μια συνάρτηση πυρήνα που επιτρέπει ένα μη γραμμικό όριο απόφασης στον αρχικό χώρο εισόδων. Αυτό οδηγεί στην ακόλουθη διατύπωση του δυϊκού προβλήματος βελτιστοποίησης (βλ. [SchSmo02] :

$$\max_a W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j=1}^{n} a_i a_j y_i y_j k(x_i, x_j) \qquad (2.12)$$

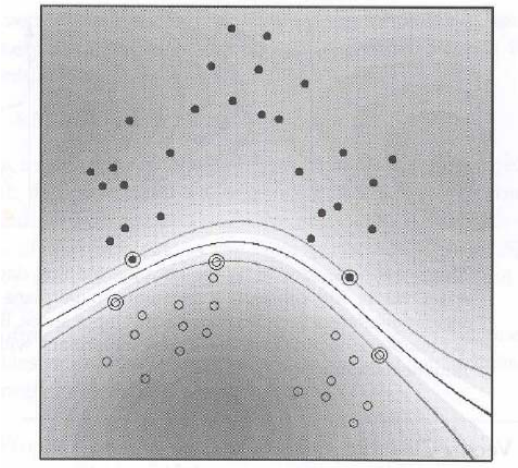
υπό τους περιορισμούς $a_i \geq 0$ για κάθε i=1,...n, και $\sum_{i=1}^{n} a_i y_i = 0$

που εισάγει μια συνάρτηση απόφασης

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{n} a_i y_i k(x, x_i) + b) \qquad (2.13)$$

Αυτή είναι η λεγόμενη αυστηρού περιθωρίου (hard margin) Μηχανή Διανύσματος Υποστήριξης. Σημειώνεται ότι το πρόβλημα βελτιστοποίησης πλέον διαμορφώνεται στον αρχικό χώρο εισόδων **X,** ο οποίος δεν είναι απαραίτητα ένας διανυσματικός χώρος.

[1] $\prec \mathbf{K}_1, \mathbf{K}_2 \succ_F$ είναι το Frobenius εσωτερικό γινόμενο που ορίζεται ως $\prec \mathbf{K}_1, \mathbf{K}_2 \succ_F = \sum_{i,j} (\mathbf{K}_1)_{ij} (\mathbf{K}_2^{\mathbf{T}})_{ij}$



Σχήμα 2.5: Παράδειγμα μίας αυστηρού περιθωρίου Μηχανής Διανύσματος Υποστήριξης (πηγή: [SchSmo02])

Παράδειγμα μίας αυστηρού περιθωρίου Μηχανής Διανύσματος Υποστήριξης που χρησιμοποιεί έναν πυρήνα RBF φαίνεται στο σχήμα 2.5. Οι κύκλοι και οι δίσκοι είναι δύο κλάσεις υποδειγμάτων εκπαίδευσης. Οι γραμμές δείχνουν το όριο απόφασης εντός του περιθωρίου. Τα διανύσματα υποστήριξης (που μαρκάρονται με τους πρόσθετους κύκλους) βρίσκονται ακριβώς πάνω στο περιθώριο.

Ένα ανοιχτό θέμα είναι τι γίνεται ωστόσο, όταν οι δύο κλάσεις δεν είναι απόλυτα διαχωρίσιμες εξαιτίας θορύβου στα δεδομένα. Σε αυτήν την περίπτωση δεν υπάρχει υπερεπίπεδο στον χώρο των χαρακτηριστικών που να διαχωρίζει τα δεδομένα. Ο Cortes και Vapnik [CorVap95] έλυσαν το πρόβλημα εισάγοντας μεταβλητές απόκλισης (slack variables) $\xi_i \geq 0, i = 1,...,n$ 'χαλαρώνοντας' έτσι τους περιορισμούς της εξίσωσης (2.8) απλά απαιτώντας

$$y_i(\prec \mathbf{x}_i, \mathbf{w} \succ + b) \geq 1 - \xi_i, i = 1,...,n \qquad (2.14)$$

Αυτό συχνά αναφέρεται ως χαλαρού περιθωρίου (soft margin) υπερεπίπεδο. Είναι φανερό ότι η παραπάνω σχέση θα μπορούσε να ισχύει πάντα, εάν τα ξ_i είναι αρκετά μεγάλα. Για αυτόν το λόγο η αντικειμενική συνάρτηση (2.8) τροποποιείται ως εξής

$$\min_{\mathbf{w}\in\mathbf{H}, b\in R} \tau(\mathbf{w}, \xi) = \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i \qquad (2.15)$$

όπου η σταθερά $C > 0$ καθορίζει την εξισορρόπηση (trade-off) μεταξύ της μεγιστοποίησης του περιθωρίου και την ελαχιστοποίηση του λάθους. Εισάγοντας έναν πυρήνα και χρησιμοποιώντας το δυϊκό πρόβλημα, το ζητούμενο είναι η μεγιστοποίηση της (2.12) υπό τους περιορισμούς

$$0 \le a_i \le C \text{ για κάθε i=1,...,n και } \sum_{i=1}^{n} a_i y_i = 0 \qquad (2.16)$$

Η μόνη διαφορά από την διαχωρίσιμη περίπτωση είναι το άνω όριο C στους πολλαπλασιαστές Langrange a_i. Με αυτόν τον τρόπο περιορίζεται η επιρροή μεμονωμένων προτύπων. Η λύση έχει την μορφή της (2.13). Αυτή είναι η αποκαλούμενη C-SVM που είναι και η πιο ευρέως χρησιμοποιούμενη Μηχανή Διανύσματος Υποστήριξης. Σημειώνεται ότι για $C \to \infty$ η λύση συγκλίνει σε αυτήν που μας δίνει μία αυστηρού περιθωρίου Μηχανή Διανύσματος Υποστήριξης.

Μια άλλη έκδοση είναι η αποκαλούμενη v-SVM (π.χ. [SchSmo02]). Αντί της παραμέτρου C, χρησιμοποιείται μια παράμετρος $v \in [0,1]$ η οποία μπορεί να αποδειχθεί [SchSmo02] ότι προσδιορίζει ένα κάτω όριο στο ποσοστό των περιπτώσεων που είναι διανύσματα υποστήριξης και ένα άνω όριο στο ποσοστό των περιπτώσεων με $\xi_i > 0$ (λάθη περιθωρίου (margin errors)). Τα πρότυπα που είναι λάθη περιθωρίου είναι είτε λάθος ταξινομημένα είτε βρίσκονται εντός του περιθωρίου. Το πρωτεύον πρόβλημα βελτιστοποίησης για τη v-SVM διατυπώνεται ως

$$\max_{\mathbf{w}\in R, \xi\in R^n, \rho, b\in R} \tau(\mathbf{w}, \xi, \rho) = \frac{1}{2} \| \mathbf{w} \|^2 - v\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i \qquad (2.17)$$

υπό τους περιορισμούς $y_i(\prec \mathbf{x}_i, \mathbf{w} \succ + b) \ge \rho - \xi_i$, και $\xi_i \ge 0$ για κάθε i=1,...,n, $\rho \ge 0$

όπου ρ είναι μια παράμετρος που ελέγχει το μέγεθος του περιθωρίου. Το δυϊκό πρόβλημα σε αυτήν την περίπτωση διατυπώνεται ως

$$\max_a W(a) = -\frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j k(x_i, x_j) \qquad (2.18)$$

$$\text{υπό τους περιορισμούς } 0 \le a_i \le \frac{1}{n},\ \sum_{i=1}^{n} a_i y_i = 0,\ \sum_{i=1}^{n} a_i \ge v$$

Η συνάρτηση απόφασης που προκύπτει έχει την ίδια μορφή όπως και στην περίπτωση της C-SVM. Μπορεί να αποδειχτεί [SchSmo02] ότι, εάν η v-SVM οδηγήσει σε $\rho>0$, η συνάρτηση απόφασης θα είναι ουσιαστικά η ίδια με αυτήν που προκύπτει από την C-SVM με $C=1/\rho$.

Το τελευταίο ανοιχτό ερώτημα είναι τι συμβαίνει όταν περισσότερες από δύο κλάσεις, έστω M πρέπει να διαχωριστούν. Η παρούσα εργασία δεν ασχολείται με τέτοιες περιπτώσεις, ωστόσο κοινές στρατηγικές σε αυτήν την περίπτωση παρουσιάζονται στο [SchSmo02, σελ. 211 - 213].

Τέλος, αξίζει να σημειωθεί ότι οι Μηχανές Διανύσματος Υποστήριξης δεν χρησιμοποιούνται μόνο στην αναγνώριση προτύπων, αλλά και στην εκτίμηση συναρτήσεων η οποία είναι γνωστή στην κλασσική στατιστική ως παλινδρόμηση (regression). Μια περιγραφή για το πως γενικεύεται ο αλγόριθμος των μηχανών διανύσματος υποστήριξης σε αυτές τις περιπτώσεις υπάρχει στο [SmoSch98].

2.2.3 Εκτίμηση της Ικανότητας Γενίκευσης

Στην παράγραφο 2.1 εξηγήθηκε πως η αρχή της δομικής ελαχιστοποίησης κινδύνου οδηγεί σε μια υπόθεση (συνάρτηση απόφασης) που επιτρέπει καλή γενίκευση και στην παράγραφο 2.2.1 επισημάνθηκε ότι η ιδέα πίσω από τις Μηχανές Διανύσματος Υποστήριξης είναι η ελαχιστοποίηση ενός ορίου στο ρίσκο γενίκευσης. Η ερώτηση που τίθεται είναι, πώς κάποιος μπορεί πραγματικά να μετρήσει την απόδοση γενίκευσης του μοντέλου ταξινόμησης που αναπτύσσεται. Όπως αναφέρθηκε στην παράγραφο 2.1, ένα ακριβές μέτρο της απόδοσης γενίκευσης δεν είναι δυνατό, δεδομένου ότι αυτό θα απαιτούσε τον υπολογισμό της εξίσωσης (2.3). Κατά συνέπεια μπορούν μόνο να χρησιμοποιηθούν στατιστικά εργαλεία για να υπολογιστεί η ικανότητα γενίκευσης. Πάλι το πρόβλημα είναι ότι έχουμε έναν πεπερασμένο αριθμό περιπτώσεων εκπαίδευσης από το οποίο θέλουμε να προβλέψουμε το λάθος που γίνεται σε οποιοδήποτε νέο σύνολο δεδομένων. Η μέτρηση του λάθους που λαμβάνεται από τα δεδομένα εκπαίδευσης θα υποτιμούσε συστηματικά το αληθινό λάθος, επειδή το μοντέλο έχει προσαρμοστεί στα δεδομένα εκπαίδευσης. Εάν είναι διαθέσιμος ένας πολύ μεγάλος αριθμός δεδομένων, μπορούν

απλά να χωριστούν τα στοιχεία σε ένα σύνολο που χρησιμοποιείται για την εκπαίδευση και σε ένα ανεξάρτητο σύνολο, που θα πρέπει να είναι όσο το δυνατό μεγαλύτερο, και θα χρησιμοποιείται για τον έλεγχο. Έτσι το μοντέλο ταξινόμησης κατασκευάζεται σύμφωνα με τα δεδομένα εκπαίδευσης, και καλείται έπειτα να προβλέψει τις εξόδους για τα δεδομένα του συνόλου ελέγχου[2].

Το κύριο πρόβλημα με την χρησιμοποίηση ενός πρόσθετου συνόλου ελέγχου είναι ότι προκειμένου να αποκτηθούν αξιόπιστα αποτελέσματα, χρειάζεται ένας πολύ μεγάλος όγκος δεδομένων στο σύνολο ελέγχου, διαφορετικά η αξιολόγηση μπορεί να είναι ευαίσθητη στα συγκεκριμένα δεδομένα που χρησιμοποιούνται στο στάδιο του ελέγχου. Κατά συνέπεια μια λύση στην περίπτωση που δεν είναι διαθέσιμα πολλά δεδομένα είναι να χρησιμοποιηθεί η τεχνική k-fold cross-validation [Rojas96]. Σύμφωνα με αυτήν την τεχνική το σύνολο δεδομένων ε διαιρείται σε k υποσύνολα , και η μέθοδος επαναλαμβάνεται k φορές. Κάθε φορά, ένα από τα k υποσύνολα χρησιμοποιείται για τον έλεγχο και τα υπόλοιπα k-1 υποσύνολα τίθενται μαζί και χρησιμοποιούνται για την εκπαίδευση. Κατόπιν η ακρίβεια ή το λάθος ταξινόμησης υπολογίζεται κατά μέσο όρο σε όλες τις k δοκιμές.

Το προφανές πλεονέκτημα αυτής της μεθόδου είναι ότι λειτουργεί ακόμα και εάν το ε είναι μικρό, επειδή δεν ενδιαφέρει πως θα χωριστούν τα δεδομένα. Ειδικά για πολύ μικρά σύνολα δεδομένων κάποιος μπορεί να επιλέξει την ακραία περίπτωση k = n. Αυτό σημαίνει ότι κάθε δεδομένο αντιμετωπίζεται ως ένα υποσύνολο. Σε κάθε δοκιμή n-1 δεδομένα χρησιμοποιούνται για την εκπαίδευση και 1 για τον έλεγχο. Αυτή η μέθοδος είναι γνωστή ως leave-one-out cross-validation. Πρόκειται για την λιγότερο 'προκατειλημμένη' εκτίμηση της ικανότητας γενίκευσης που μπορεί να δώσει η μέθοδος cross-validation. Ωστόσο είναι σαφές ότι είναι μια υπολογιστικά πολύ επίπονη μέθοδος, επειδή απαιτεί n επαναλήψεις για την εκπαίδευση και έλεγχο του μοντέλου ταξινόμησης.

[2] Οι ακόλουθοι όροι χρησιμοποιούνται συνήθως για να περιγράψουν τις ιδιότητες της ταξινόμησης ενός δείγματος (x,y) όσον αφορά μια κλάση C:

- (x,y) είναι αληθινά θετικό (T +): το πρότυπο x τέθηκε σωστά στην θεωρούμενη κλάση C.
- (x,y) είναι αληθινά αρνητικό (T -): το πρότυπο x δεν ανήκει στην κλάση C και ο ταξινομητής σωστά δεν τον έβαλε εκεί.
- (x,y) είναι ψεύτικα θετικό (F +): το πρότυπο x δεν ανήκει στην C, αλλά ο ταξινομητής το έβαλε εκεί.
- (x,y) είναι ψεύτικα αρνητικό (F -): το πρότυπο x στην πραγματικότητα ανήκει στην C, αλλά ο ταξινομητής δεν το έβαλε εκεί.

Μια άλλη λύση στο πρόβλημα των ελλιπών δεδομένων δίνει η μέθοδος bootstrapping . Η ιδέα πίσω από το bootstrap είναι η ακόλουθη [Rojas96]: Τα n σημεία εκπαίδευσης προέρχονται από μια άγνωστη κατανομή πιθανότητας P. Η υπόθεση του bootstrap είναι ότι αυτή η κατανομή μπορεί να προσεγγιστεί παίρνοντας k τυχαία υποδείγματα από το σύνολο ε με επανατοποθέτηση. Αυτό σημαίνει ότι σε ένα από αυτά τα τυχαία δείγματα κάθε σημείο εκπαίδευσης μπορεί να επιλεχτεί περισσότερες από μία φορά. Κάθε φορά το μοντέλο ταξινόμησης αναπτύσσεται σε ένα δείγμα bootstrap και ελέγχεται στις περιπτώσεις που δεν περιλαμβάνονται στο δείγμα αυτό. Κατόπιν η ακρίβεια ή το λάθος ταξινόμησης υπολογίζεται κατά μέσο από τις k επαναλήψεις της παραπάνω διαδικασίας. Όσο μεγαλύτερο είναι το k τόσο καλύτερη είναι η προσέγγιση της άγνωστης κατανομής P, και επομένως τόσο καλύτερη είναι και η εκτίμηση της απόδοσης γενίκευσης. Το πλεονέκτημα της μεθόδου bootstrap είναι ότι δεν χρειάζεται πολλά δεδομένα για να πετύχει αρκετά αξιόπιστα αποτελέσματα, ωστόσο το μειονέκτημα της είναι ότι είναι μια υπολογιστικά επίπονη μέθοδος.

Εκτός από αυτές τις γενικές προσεγγίσεις υπάρχουν ειδικά για τις Μηχανές Διανύσματος Υποστήριξης θεωρητικά όρια στο leave-one-out λάθος τα οποία παρουσιάζονται στους [Vapnik98, VapCha00, ChaVap00]

Το πιο γνωστό μέτρο της δυνατότητας γενίκευσης ενός μοντέλου ταξινόμησης είναι το ποσοστό ακρίβειας (accuracy rate) [UltschVL00] που προκύπτει από το ποσοστό των σωστών απαντήσεων ως προς τον αριθμό των δεδομένων στο σύνολο ελέγχου. Ομοίως κάποιος μπορεί να υπολογίσει το ποσοστό των λαθών σε όλα τα δεδομένα ελέγχου. Αυτό καλείται λάθος ταξινόμησης (classification error). Επιπλέον μπορεί να υπολογιστεί για κάθε κλάση η αναμενόμενη ακρίβεια ως ο αριθμός των σωστών απαντήσεων όσον αφορά την κλάση αυτή σε σχέση με όλα τα δεδομένα που ανήκουν στην κλάση (sensitivity)[3] ή συμπληρωματικά, ο αριθμός των δεδομένων που σωστά δεν τέθηκε σε μια ορισμένη κλάση σε σχέση με όλα τα δεδομένα που δεν

[3] $\text{Sensitivity} = \dfrac{\#\{T+\}}{\#\{T+\} + \#\{F-\}}$

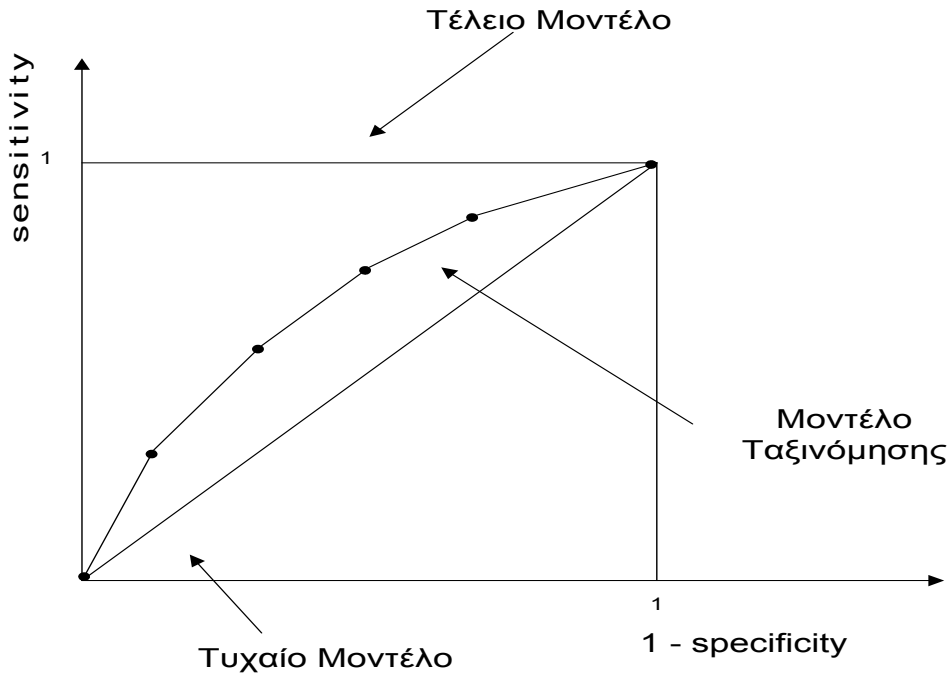
ανήκουν σε αυτήν την κλάση (specificity)[4]. Και οι δύο παραπάνω έννοιες αποτελούν ένα μέτρο για το πόσο καλά μια ορισμένη κλάση διαχωρίζεται από τις υπόλοιπες. Τέλος, για κάθε κλάση μπορεί κάποιος να υπολογίσει το ποσοστό των σωστών απαντήσεων όσον αφορά μια ορισμένη κλάση σε όλα τα δεδομένα που ταξινομήθηκαν σε αυτήν την κλάση είτε σωστά είτε λάθος (positive predictive value)[5].

Εκτός από τα παραπάνω μέτρα της δυνατότητας γενίκευσης των μοντέλων ταξινόμησης, υπάρχουν οι λεγόμενες καμπύλες ROC (Receiver Operating Characteristics) οι οποίες αποτυπώνουν γραφικά την αποτελεσματικότητα των μοντέλων ταξινόμησης. Ένας από τους πρώτους που υιοθέτησαν τις καμπύλες ROC στην εκμάθηση μηχανών ήταν ο Spackman [Spackman89], ο οποίος κατέδειξε την αξία των καμπυλών αυτών στην εκτίμηση και σύγκριση διαφόρων αλγορίθμων. Οι καμπύλες ROC χρησιμοποιούνται ευρέως πλέον στην εκμάθηση μηχανών, λόγω των ιδιοτήτων τους που τις καθιστάν ιδιαίτερα χρήσιμες σε περιπτώσεις που τα κόστη εσφαλμένων ταξινομήσεων δεν είναι γνωστά.

Ορισμένα μοντέλα ταξινόμησης, όπως τα νευρωνικά δίκτυα και οι Μηχανές Διανύσματος Υποστήριξης αποδίδουν ένα σκορ σε κάθε αντικείμενο, δηλαδή μια τιμή που αναπαριστά τον βαθμό κατά τον οποίο ένα αντικείμενο είναι μέρος μιας κλάσης. Έτσι τα αντικείμενα με σκορ που βρίσκονται πάνω από ένα κατώφλι C_{cut} ανήκουν στην μία κλάση και τα άλλα στην άλλη. Οι καμπύλες ROC κατασκευάζονται ως εξής : Για όλα τα κατώφλια C_{cut} που εμπεριέχονται στο εύρος των σκορ εκτίμησης, υπολογίζονται τα sensitivity και specificity, όπως περιγράφτηκαν παραπάνω. Η καμπύλη ROC είναι ένα γράφημα του sensitivity με το 1-specificity για όλες τις τιμές του C_{cut}. Στο σχήμα 2.6 φαίνονται ορισμένα παραδείγματα καμπύλων ROC.

[4] Specificity = $\frac{\#\{T-\}}{\#\{T-\}+\#\{F-\}}$

[5] positive predictive value = $\frac{\#\{T+\}}{\#\{T+\}+\#\{F+\}}$



Σχήμα 2.6 : Καμπύλες ROC για διάφορα μοντέλα

Η αποτελεσματικότητα ενός μοντέλου ταξινόμησης είναι καλύτερη όσο πιο απότομη είναι η ROC καμπύλη στο αριστερό τέλος και όσο πιο κοντά βρίσκεται στο σημείο (0,1). Ομοίως, το μοντέλο είναι καλύτερο όσο μεγαλύτερο είναι το εμβαδό κάτω από την καμπύλη (area under the curve – AUC). Το AUC μπορεί να ερμηνευτεί ως η μέση αποτελεσματικότητα των ελέγχων που αντιστοιχούν σε όλες τις δυνατές τιμές $C_{cut.}$ Το εμβαδό AUC είναι 0.5 για ένα τυχαίο μοντέλο χωρίς διακριτική ικανότητα και είναι 1 για ένα τέλειο μοντέλο. Είναι μεταξύ 0.5 και 1 για ένα οποιοδήποτε λογικό μοντέλο ταξινόμησης στην πράξη. Ένα άλλο μέτρο της ποιότητας ενός μοντέλου ταξινόμησης είναι ο δείκτης ακρίβειας (Accuracy Ratio – AR), ο οποίος προκύπτει από την σχέση AR = 2AUC – 1. Επομένως, όσο πλησιέστερα στην μονάδα είναι ο δείκτης ακρίβειας, τόσο καλύτερο θεωρείται ένα μοντέλο ταξινόμησης.

Κεφάλαιο 3

'Γενετικοί Αλγόριθμοι'

3.1 Εισαγωγικά

Οι γενετικοί αλγόριθμοι (Genetic Algorithms-GAs) [Whitley93, Goldberg98] αποτελούν μια από τις πιο πολλά υποσχόμενες εφαρμογές του εξελικτικού προγραμματισμού (Evolution Computing). Τα εξελικτικά συστήματα μελετήθηκαν στη δεκαετία του '50 και τη δεκαετία του '60 από διάφορους επιστήμονες υπολογιστών με την ιδέα ότι η 'εξέλιξη' θα μπορούσε να χρησιμοποιηθεί ως εργαλείο βελτιστοποίησης στα προβλήματα εφαρμοσμένης μηχανικής. Η ιδέα σε όλα αυτά τα συστήματα ήταν να εξελιχθεί ένας πληθυσμός από υποψήφιες λύσεις σε ένα δεδομένο πρόβλημα, που θα χρησιμοποιεί τελεστές εμπνεόμενους από τη φυσική γενετική μεταβλητότητα (natural genetic variation) και τη φυσική επιλογή (natural selection).

Οι γενετικοί αλγόριθμοι αναπτύχθηκαν από τον John Holland, [Holland75] και τους σπουδαστές και συναδέλφους του (π.χ. De Jong [DeJong75]) στο πανεπιστήμιο του Μίτσιγκαν στη δεκαετία του '60 και τη δεκαετία του '70, και οι περισσότερες εργασίες σε θεωρητικό επίπεδο για τους γενετικούς αλγορίθμους αναφέρονται σε αυτόν τον αρχικό (κανονικό) γενετικό αλγόριθμο. Ωστόσο υπήρξε μια αφθονία παραλλαγών του αλγορίθμου από άλλους ερευνητές που προσάρμοσαν τον αρχικό γενετικό αλγόριθμο στα προβλήματά τους. Σε αντίθεση με τις εξελικτικές στρατηγικές και τον εξελικτικό προγραμματισμό, ο αρχικός στόχος του Holland δεν ήταν να σχεδιάσει αλγορίθμους που θα επιλύουν συγκεκριμένα προβλήματα, αλλά κυρίως να μελετήσει τυπικά το φαινόμενο της προσαρμογής (adaptation) όπως αυτό

εμφανίζεται στη φύση και να αναπτύξει τρόπους με τους οποίους οι μηχανισμοί της φυσικής προσαρμογής θα μπορούσαν να εισαχθούν στα υπολογιστικά συστήματα.

Οι γενετικοί αλγόριθμοι κωδικοποιούν μια πιθανή λύση σε ένα συγκεκριμένο πρόβλημα σε μια δομή χρωμοσώματος (chromosome-like structure) και εφαρμόζουν τους τελεστές γενετικής διασταύρωσης (crossover) και μετάλλαξης (mutation) σε αυτή την δομή. Η βασική ιδέα είναι η αρχή της επιβίωσης του καταλληλότερου (survival of the fittest) , που σημαίνει ότι οι πιθανές λύσεις ανταγωνίζονται η μια με την άλλη με τέτοιο τρόπο ώστε τα χρωμοσώματα που αντιπροσωπεύουν τις καλύτερες λύσεις στο δεδομένο πρόβλημα να έχουν περισσότερες πιθανότητες για αναπαραγωγή.

Συνήθως οι γενετικοί αλγόριθμοι χρησιμοποιούνται στη βελτιστοποίηση συναρτήσεων. Οι παραδοσιακές τεχνικές βελτιστοποίησης όπως η κάθοδος κλίσης (gradient descent) έχουν τοπικό πεδίο. Τα βέλτιστα που επιδιώκουν αυτές οι τεχνικές είναι εκείνα τα σημεία που είναι τα καλύτερα σε μια τοπική μόνο περιοχή γύρω από το τρέχον σημείο. Αυτό δημιουργεί προβλήματα, εάν η συνάρτηση έχει πολλά τοπικά βέλτιστα ή, ακόμα χειρότερα, εάν η κλίση δεν είναι υπολογίσιμη. Οι απαριθμητικές μέθοδοι αναζήτησης (enumerative search methods) (π.χ. ο A*-αλγόριθμος [RusNor95]) από την άλλη, ισχύουν μόνο εάν το διάστημα αναζήτησης είναι διακριτό και όχι πάρα πολύ μεγάλο.

Το πλεονέκτημα των γενετικών αλγορίθμων έναντι των παραδοσιακών μεθόδων είναι ότι δεν χρειάζονται πολλές πληροφορίες για τη συνάρτηση που βελτιστοποιούν και επιπλέον αποδίδουν ακόμα κι αν το διάστημα αναζήτησης είναι πολύ μεγάλο. Οι γενετικοί αλγόριθμοι χρησιμοποιούν την τυχαία επιλογή ως εργαλείο για να οδηγήσουν μια ιδιαίτερα εξονυχιστική αναζήτηση μέσα από μια διακριτή κωδικοποίηση του διαστήματος αναζήτησης. Επομένως, η μετάβαση από μια κατάσταση στο διάστημα αναζήτησης σε μια άλλη είναι πιθανοκρατική και όχι αιτιοκρατική. Ένα άλλο χαρακτηριστικό των γενετικών αλγορίθμων είναι, ότι δεν ψάχνουν από ένα μόνο σημείο, αλλά από έναν πληθυσμό σημείων ταυτόχρονα. Αυτό σημαίνει ότι οι γενετικοί αλγόριθμοι εκτελούν μια σφαιρική διαδικασία αναζήτησης, εφόσον ερευνούν το διάστημα αναζήτησης από πολλά σημεία παράλληλα. Με αυτόν τον τρόπο μπορούν να αποφύγουν τα τοπικά βέλτιστα. Ωστόσο, πρέπει να αναφερθεί ότι οι γενετικοί αλγόριθμοι δεν είναι πάντα η καλύτερη επιλογή. Ανάλογα με το

πρόβλημα που αντιμετωπίζεται, μια παραδοσιακή κάθοδος κλίσης μπορεί να είναι μια πολύ γρηγορότερη μέθοδος από έναν σχετικά επίπονο υπολογιστικά γενετικό αλγόριθμο.

3.2 Ο Απλός Γενετικός Αλγόριθμος (Simple Genetic Algorithm)

Τα βήματα του απλού γενετικού αλγόριθμου όπως ονομάστηκε έτσι από τον Goldberg [Goldberg98] είναι:

(α) Ξεκινάει με έναν τυχαία παραγόμενο πληθυσμό από N l-bit χρωμοσώματα που αποτελούν τις υποψήφιες λύσεις για το εκάστοτε πρόβλημα.

(β) Υπολογίζει την καταλληλότητα (fitness) F_i κάθε χρωμοσώματος i στον πληθυσμό.

(γ) Επαναλαμβάνει τα ακόλουθα βήματα έως ότου δημιουργηθούν N απόγονοι.

i. Επιλέγει ένα ζευγάρι χρωμοσωμάτων από τον τρέχοντα πληθυσμό που θα αποτελέσει τους γονείς, με πιθανότητα επιλογής που είναι μια αύξουσα συνάρτηση της καταλληλότητας. Η επιλογή γίνεται "με αντικατάσταση", δηλαδή το ίδιο χρωμόσωμα μπορεί να επιλεχτεί περισσότερο από μία φορά για να γίνει γονέας.

ii. Με πιθανότητα p_c (πιθανότητα διασταύρωσης ή ποσοστό διασταύρωσης), διασταυρώνει το ζευγάρι σε ένα τυχαία επιλεγμένο σημείο για να διαμορφώσει δύο νέους απογόνους. Εάν δεν πραγματοποιείται διασταύρωση, οι δύο νέοι απόγονοι είναι τα ακριβές αντίγραφα των γονέων. (Σημειώνεται ότι στην παρούσα εργασία το ποσοστό διασταυρώσεων καθορίστηκε να είναι η πιθανότητα με την οποία δύο γονείς διασταυρώνονται σε ένα μόνο σημείο. Υπάρχουν ωστόσο παραλλαγές "πολυσημειακών διασταυρώσεων" του γενετικού αλγόριθμου στις οποίες το ποσοστό διασταυρώσεων για ένα ζευγάρι γονέων είναι ο αριθμός των σημείων στα οποία πραγματοποιείται μια διασταύρωση.)

iii. Μεταλλάσσει κάθε σημείο των δύο απογόνων με πιθανότητα p_m (πιθανότητα μετάλλαξης ή ποσοστό μετάλλαξης), και τοποθετεί τα προκύπτοντα χρωμοσώματα στο νέο πληθυσμό. Εάν το N είναι περιττός αριθμός, τότε μπορεί να απορριφθεί ένα μέλος του νέου πληθυσμού με τυχαίο τρόπο.

(δ) Αντικαθιστά τον τρέχοντα πληθυσμό με το νέο πληθυσμό που δημιουργήθηκε.

(ε) Επαναλαμβάνει τα βήματα (β)-(ε) μέχρι να ικανοποιηθεί η συνθήκη τερματισμού.

Όπως φαίνεται από τα παραπάνω, ο απλός γενετικός αλγόριθμος, όπως και κάθε άλλος απαιτεί τον καθορισμό έξι βασικών παραγόντων : (α) πως θα γίνει η αναπαράσταση της λύσης (solution representation), (β) ποια θα είναι η συνάρτηση επιλογής των χρωμοσωμάτων για αναπαραγωγή (selection function), (γ) ποιοι θα είναι οι γενετικοί τελεστές που θα πραγματοποιήσουν την αναπαραγωγή (genetic operators), (δ) πως θα επιλεχτεί ο αρχικός πληθυσμός (initial population), (ε) ποια θα είναι τα κριτήρια τερματισμού (termination criteria) και (στ) ποια θα είναι η συνάρτηση αξιολόγησης (evaluation function). Στη συνέχεια περιγράφονται αναλυτικά οι παραπάνω παράγοντες.

3.2.1 Αναπαράσταση της Λύσης

Για οποιοδήποτε γενετικό αλγόριθμο, η αναπαράσταση των χρωμοσωμάτων είναι αναγκαία για να περιγράψει κάθε πιθανή λύση στον υπό εξέταση πληθυσμό. Ο τρόπος της αναπαράστασης καθορίζει το πως θα δομηθεί το πρόβλημα στον γενετικό αλγόριθμο και επιπλέον καθορίζει τους γενετικούς τελεστές που θα χρησιμοποιηθούν. Κάθε χρωμόσωμα αποτελείται από μια ακολουθία γονιδίων από ένα ορισμένο αλφάβητο. Ένα αλφάβητο θα μπορούσε να αποτελείται από δυαδικά ψηφία (0 και 1), δεκαδικούς αριθμούς, ακέραιους, σύμβολα (π.χ Α, Β, Γ), πίνακες και πολλά άλλα.. Στον αρχικό σχεδιασμό του Holland το αλφάβητο περιορίστηκε στα δυαδικά ψηφία. Από τότε η αναπαράσταση του προβλήματος αποτέλεσε το αντικείμενο συστηματικής έρευνας.

Έχει αποδειχτεί ότι οι φυσικές αναπαραστάσεις παράγουν καλύτερες λύσεις [Michalewicz94]. Μία χρήσιμη αναπαράσταση ενός χρωμοσώματος για τη βελτιστοποίηση συναρτήσεων εμπεριέχει γονίδια ή μεταβλητές από ένα αλφάβητο με δεκαδικούς αριθμούς με τιμές μεταξύ του κατώτερου και ανώτερου ορίου στις μεταβλητές. Ο Michalewicz [Michalewicz94] έκανε εκτενή πειράματα για να συγκρίνει τους γενετικούς αλγορίθμους με πραγματικές τιμές και τους δυαδικούς γενετικούς αλγορίθμους και απέδειξε ότι ο πραγματικών τιμών γενετικός αλγόριθμος είναι πιο αποτελεσματικός αναφορικά με τον χρόνο επεξεργασίας που χρειάζεται. Απέδειξε επίσης, ότι η πραγματικών τιμών αναπαράσταση οδηγεί το πρόβλημα πιο κοντά σε εκείνη την αναπαράσταση του προβλήματος που προσφέρει υψηλότερη ακρίβεια και πιο αξιόπιστα αποτελέσματα.

3.2.2 Συνάρτηση Επιλογής

Η επιλογή των χρωμοσωμάτων που θα παράγουν τις διαδοχικές γενεές διαδραματίζει έναν εξαιρετικά σημαντικό ρόλο σε έναν γενετικό αλγόριθμο. Μια πιθανοκρατική επιλογή εκτελείται βασισμένη στη καταλληλότητα του κάθε χρωμοσώματος, έτσι ώστε τα καλύτερα χρωμοσώματα να έχουν μεγαλύτερη πιθανότητα επιλογής. Ένα χρωμόσωμα στον πληθυσμό μπορεί να επιλεχτεί περισσότερες από μία φορά αν και όλα τα χρωμοσώματα στον πληθυσμό είναι πιθανά για αναπαραγωγή στην επόμενη γενεά. Υπάρχουν διάφορα σχήματα για την διαδικασία επιλογής: επιλογή ρόδα ρουλέτας (roulette wheel selection) και οι προεκτάσεις αυτής της μεθόδου, τεχνικές βαθμονόμησης (scaling techniques), επιλογή τουρνουά (tournament), εκλεκτικά μοντέλα (elitist models), και μέθοδοι κατάταξης (ranking methods) [Goldberg98, Michalewicz94].

Μια κοινή προσέγγιση επιλογής ορίζει μια πιθανότητα επιλογής P_j για κάθε χρωμόσωμα j βασισμένη στην καταλληλότητα του κάθε χρωμοσώματος. Παράγεται μια σειρά από N τυχαίους αριθμούς καθένας από τους οποίους συγκρίνεται με την αθροιστική πιθανότητα, $C_i = \sum_{j=1}^{i} P_j$ του πληθυσμού. Έτσι, το κατάλληλο χρωμόσωμα i επιλέγεται για αναπαραγωγή, εάν $C_{i-1} < U(0,1) \le C_i$ και συνολικά επιλέγονται N χρωμοσώματα όσα αποτελούσαν και τον αρχικό πληθυσμό. Υπάρχουν διάφορες μέθοδοι που ορίζουν τις πιθανότητες των χρωμοσωμάτων: ρόδα ρουλέτας (roulette

wheel), γραμμική κατάταξη (linear ranking) και γεωμετρική κατάταξη (geometric ranking).

Η μέθοδος ρόδα ρουλέτας, που αναπτύχθηκε από τον Holland [Holland75], ήταν η πρώτη μέθοδος επιλογής. Η πιθανότητα P_i γιατί κάθε χρωμόσωμα ορίζεται ως:

$$P_i = \frac{F_i}{\sum_{j=1}^{N} F_j} \qquad (3.1)$$

όπου το F_i αντιστοιχεί στην καταλληλότητα του χρωμοσώματος *i*. Η χρησιμοποίηση της μεθόδου ρόδα ρουλέτας , ωστόσο, περιορίζει τον γενετικό αλγόριθμο μόνο στη μεγιστοποίηση, δεδομένου ότι η συνάρτηση αξιολόγησης πρέπει να αντιστοιχήσει τις λύσεις σε ένα πλήρως διαταγμένο σύνολο τιμών στο R^+. Γι' αυτόν τον λόγο προτάθηκαν προεκτάσεις της μεθόδου, όπως η παραθυροποίηση (windowing) και η βαθμονόμηση (scaling) που επιτρέπουν την ελαχιστοποίηση και την αρνητικότητα.

Οι μέθοδοι κατάταξης απαιτούν μόνο η συνάρτηση αξιολόγησης να αντιστοιχίζεται σε ένα μερικώς διαταγμένο σύνολο επιτρέποντας έτσι την ελαχιστοποίηση και την αρνητικότητα. Οι μέθοδοι κατάταξης ορίζουν μία πιθανότητα P_i βασισμένη στην σειρά κατάταξης της λύσης *i* σε σχέση με τις υπόλοιπες λύσεις. Η κανονικοποιημένη γεωμετρική κατάταξη [JoinesHouck94] ορίζει την P_i για κάθε χρωμόσωμα ως:

$$P_i = q'(1-q)^{r-1} \qquad (3.2)$$

όπου:

q = η πιθανότητα επιλογής του καλύτερου χρωμοσώματος
r = η κατάταξη του χρωμοσώματος, όταν το 1 είναι το καλύτερο
N = το μέγεθος του πληθυσμού

$$q' = \frac{q}{1-(1-q)^N}$$

Η μέθοδος επιλογής τουρνουά όπως και οι μέθοδοι κατάταξης, απαιτούν μόνο η συνάρτηση αξιολόγησης να αντιστοιχεί τις λύσεις σε ένα μερικώς διαταγμένο

σύνολο, ωστόσο δεν ορίζει τις πιθανότητες των χρωμοσωμάτων. Η μέθοδος επιλογής τουρνουά επιλέγει τυχαία *k* χρωμοσώματα, με αντικατάσταση, από τον πληθυσμό, και εισάγει το καλύτερο από αυτά τα *k* χρωμοσώματα στον νέο πληθυσμό. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου επιλεχτούν *N* χρωμοσώματα.

3.2.3 Γενετικοί Τελεστές

Οι γενετικοί τελεστές παρέχουν το βασικό μηχανισμό αναζήτησης του γενετικού αλγορίθμου. Οι τελεστές χρησιμοποιούνται για να δημιουργήσουν νέες λύσεις βασισμένες στις υπάρχουσες λύσεις στον πληθυσμό. Υπάρχουν δύο βασικοί τύποι τελεστών: διασταύρωση και μετάλλαξη. Η διασταύρωση παίρνει δύο χρωμοσώματα και παράγει δύο νέα, ενώ η μετάλλαξη αλλάζει μόνο ένα χρωμόσωμα για να παράγει μία μόνο νέα λύση. Η εφαρμογή των δύο αυτών βασικών τύπων τελεστών και τα παράγωγά τους εξαρτώνται από την αντιπροσώπευση του χρωμοσώματος (λύσης) που χρησιμοποιείται.

Έστω $\overline{X}, \overline{Y}$ δύο *l* – διάστατα διανύσματα που αναπαριστούν τα χρωμοσώματα – γονείς από τον πληθυσμό. Για δυαδικά $\overline{X}, \overline{Y}$, ορίζονται οι ακόλουθοι τελεστές: δυαδική μετάλλαξη (binary mutation) και απλή διασταύρωση (simple crossover).

Η δυαδική μετάλλαξη αντιστρέφει κάθε bit σε κάθε χρωμόσωμα στον πληθυσμό με πιθανότητα p_m σύμφωνα με την εξίσωση 3.3.

$$x_i' = \begin{cases} 1 - x_i, \text{εάν } U(0,1) < p_m \\ x_i, \text{ διαφορετικά} \end{cases} \qquad (3.3)$$

Η απλή διασταύρωση παράγει έναν τυχαίο αριθμό *r* από μια ομοιόμορφη κατανομή από το 1 μέχρι το *l* και δημιουργεί δύο νέα χρωμοσώματα ($\overline{X}', \overline{Y}'$) σύμφωνα με τις εξισώσεις 3.4 και 3.5.

$$x_i' = \begin{cases} x_i, \text{ εάν } i < r \\ y_i, \text{ διαφορετικά} \end{cases} \qquad (3.4)$$

$$y_i' = \begin{cases} y_i, & \text{εάν } i < r \\ x_i, & \text{διαφορετικά} \end{cases} \qquad (3.5)$$

Τελεστές για αναπαραστάσεις με πραγματικούς αριθμούς, όπως για παράδειγμα για ένα αλφάβητο από δεκαδικούς αριθμούς αναπτύχθηκαν από τον Michalewicz [Michalewicz94]. Για πραγματικούς $\overline{X}, \overline{Y}$ ορίζονται οι ακόλουθοι τελεστές: ομοιόμορφη μετάλλαξη (uniform mutation), μη-ομοιόμορφη μετάλλαξη (non-uniform mutation), πολλαπλή μη-ομοιόμορφη μετάλλαξη (multi-non-uniform mutation), μετάλλαξη ορίου (boundary mutation), απλή διασταύρωση (simple crossover), αριθμητική διασταύρωση (arithmetic crossover), και ευρετική διασταύρωση (heuristic crossover).

Έστω α_i και b_i είναι το κατώτερο και ψηλότερο όριο αντίστοιχα για κάθε μεταβλητή *i*. Στην ομοιόμορφη μετάλλαξη επιλέγεται τυχαία μια μεταβλητή, *j*, και εξισώνεται με έναν ομοιόμορφο τυχαίο αριθμό U(α_i, b_i) ως :

$$x_i' = \begin{cases} U(a_i, b_i), & \text{εάν } i = j \\ x_i, & \text{διαφορετικά} \end{cases} \qquad (3.6)$$

Στην μετάλλαξη ορίου επιλέγεται τυχαία μία μεταβλητή, *j,* και εξισώνεται είτε με το κατώτερο είτε με το ανώτερο όριο, όπου $r = \text{U}(0,1)$:

$$x_i' = \begin{cases} a_i, & \text{εάν } i = j, r < 0.5 \\ b_i, & \text{εάν } i = j, r \geq 0.5 \\ x_i, & \text{διαφορετικά} \end{cases} \qquad (3.7)$$

Στην μη-ομοιόμορφη μετάλλαξη επιλέγεται τυχαία μία μεταβλητή, *j*, και εξισώνεται με έναν μη-ομοιόμορφο τυχαίο αριθμό:

$$x_i' = \begin{cases} x_i + (b_i - x_i) f(G), & \text{εάν } r_1 < 0.5 \\ x_i - (x_i + a_i) f(G), & \text{εάν } r_1 \geq 0.5 \\ x_i, & \text{διαφορετικά} \end{cases} \qquad (3.8)$$

όπου

$$f(G) = (r_2(1 - \frac{G}{G_{\max}}))^b \qquad (3.9)$$

r_1, r_2 = ομοιόμορφος τυχαίος αριθμός μεταξύ (0,1),
G = η τρέχων γενεά,
$G_{\max}$ = ο μέγιστος αριθμός γενεών,
b = η παράμετρος διαμόρφωσης

Η πολλαπλή μη-ομοιόμορφη μετάλλαξη χρησιμοποιεί τον τελεστή της μη-ομοιόμορφης μετάλλαξης σε όλες τις μεταβλητές-γονίδια του χρωμοσώματος-γονέα $\overline{X}$.

Η πραγματικών τιμών απλή διασταύρωση είναι πανομοιότυπη με την δυαδική απλή διασταύρωση όπως περιγράφτηκε με τις εξισώσεις 3.4 και 3.5. Η αριθμητική διασταύρωση παράγει δύο γραμμικούς συνδυασμούς των γονέων, όπου r = U(0,1):

$$\overline{X}' = r\overline{X} + (1-r)\overline{Y} \qquad (3.10)$$

$$\overline{Y}' = (1-r)\overline{X} + r\overline{Y} \qquad (3.11)$$

Η ευρετική διασταύρωση παράγει μια γραμμική παρέκταση των δύο χρωμοσωμάτων. Πρόκειται για τον μοναδικό τελεστή που χρησιμοποιεί πληροφορίες για την καταλληλότητα των χρωμοσωμάτων. Ένα νέο χρωμόσωμα, $\overline{X}'$, παράγεται χρησιμοποιώντας την εξίσωση 3.12, όπου r = U(0,1) και το $\overline{X}$ είναι καλύτερο από το $\overline{Y}$ ως προς την καταλληλότητα. Εάν το $\overline{X}'$ είναι μια αδύνατη λύση (infeasible). (Η εφικτότητα (feasibility) δίνεται από τη εξίσωση 3.14), τότε παράγεται ένας νέος τυχαίος αριθμός r και δημιουργείται μια νέα λύση χρησιμοποιώντας την εξίσωση 3.12, διαφορετικά ο αλγόριθμος σταματάει. Για να εξασφαλιστεί η λήξη του αλγορίθμου, μετά κάποιο προκαθορισμένο αριθμό αποτυχιών, τα χρωμοσώματα-παιδιά εξισώνονται με τα χρωμοσώματα-γονείς και ο αλγόριθμος σταματάει.

$$\overline{X}' = \overline{X} + r(\overline{X} - \overline{Y}) \qquad (3.12)$$

$$\overline{Y}' = \overline{X} \qquad (3.13)$$

$$\text{εφικτότητα} = \begin{cases} 1, \text{ εάν } x_i' \geq a_i, x_i' \leq b_i \ \ \forall i \\ 0, \text{ διαφορετικά} \end{cases} \qquad (3.14)$$

3.2.4 Έναρξη, Τερματισμός, και Συναρτήσεις Αξιολόγησης

Ο γενετικός αλγόριθμος πρέπει να ξεκινήσει με έναν αρχικό πληθυσμό όπως υποδεικνύεται στο (α) βήμα του απλού γενετικού αλγορίθμου (παρ.3.2). Η πιο κοινή μέθοδος είναι να παραχθούν τυχαία οι λύσεις για ολόκληρο τον πληθυσμό. Ωστόσο, δεδομένου ότι ο γενετικός αλγόριθμος μπορεί επαναληπτικά να βελτιώσει τις υπάρχουσες λύσεις, ο αρχικός πληθυσμός μπορεί να τροφοδοτηθεί με μερικές εν δυνάμει καλές λύσεις (για παράδειγμα, λύσεις από άλλες τρέχουσες πρακτικές), ενώ ο υπόλοιπος πληθυσμός να παραχθεί τυχαία.

Ο γενετικός αλγόριθμος κινείται από γενεά σε γενεά επιλέγοντας και αναπαράγοντας τους γονείς έως ότου ικανοποιηθεί ένα κριτήριο τερματισμού. Το πιο διαδεδομένο κριτήριο τερματισμού είναι ένας συγκεκριμένος μέγιστος αριθμός από γενεές. Μια άλλη στρατηγική τερματισμού αφορά το κριτήριο σύγκλισης του πληθυσμού. Γενικά, οι γενετικοί αλγόριθμοι θα οδηγήσουν τα περισσότερα χρωμοσώματα από τον πληθυσμό να συγκλίνουν σε μία μόνο λύση. Όταν το άθροισμα των αποκλίσεων μεταξύ των χρωμοσωμάτων γίνει μικρότερο από ένα συγκεκριμένο κατώφλι, ο αλγόριθμος μπορεί να τερματιστεί. Ο αλγόριθμος μπορεί επίσης να τερματιστεί εφόσον δεν υπάρχει περαιτέρω βελτίωση της λύσης σε έναν ορισμένο αριθμό από γενεές. Εναλλακτικά, μία επιθυμητή τιμή για το μέτρο αξιολόγησης μπορεί να καθοριστεί βασισμένη σε κάποιο αυθαίρετα αποδεκτό κατώτατο όριο. Επιπλέον, μπορούν να χρησιμοποιηθούν διάφορες στρατηγικές η μια από κοινού με την άλλη.

Σε έναν γενετικό αλγόριθμο μπορούν να χρησιμοποιηθούν συναρτήσεις αξιολόγησης πολλών μορφών, υπό την ελάχιστη απαίτηση ότι η συνάρτηση μπορεί να αντιστοιχήσει τον πληθυσμό σε ένα μερικώς διαταγμένο σύνολο. Όπως ορίστηκε, η συνάρτηση αξιολόγησης είναι ανεξάρτητη από τον γενετικό αλγόριθμο.

3.3 Το Μαθηματικό Υπόβαθρο των Γενετικών Αλγορίθμων

Η παραδοσιακή θεωρία των γενετικών αλγορίθμων, η οποία μορφοποιήθηκε πρώτα από τον Holland [Holland75], υποθέτει ότι, σε ένα πολύ γενικό επίπεδο περιγραφής, οι γενετικοί αλγόριθμοι δουλεύουν ανακαλύπτοντας και επανασυνδυάζοντας καλές δομικές μονάδες (building blocks) των λύσεων με έναν παράλληλο τρόπο. Η ιδέα είναι ότι οι καλές λύσεις τείνουν να αποτελούνται από καλές δομικές μονάδες, δηλαδή συνδυασμούς από τιμές bits-γονίδια που προσδίδουν μεγαλύτερη καταλληλότητα στα χρωμοσώματα στα οποία βρίσκονται.

Ο Holland εισήγαγε την θεώρηση των σχημάτων (schemas or schemata) για να μορφοποιήσει την ιδέα των δομικών μονάδων. Ένα σχήμα είναι ένα σύνολο από σειρές bit που μπορεί να περιγραφτεί από ένα κώδικα με άσσους , μηδενικά και αστερίσκους, με τους αστερίσκους να αναπαριστούν σημεία αδιαφορίας. Για παράδειγμα, εάν ένα σχήμα H είναι το *11*0**, τότε το χρωμόσωμα A = 0111000 είναι ένα παράδειγμα του σχήματος H, εφόσον τα σταθερά σημεία του σχήματος στις θέσεις 2, 3 και 5 συμπίπτουν με τα σημεία που βρίσκονται στις αντίστοιχες θέσεις του Α. Σαν γενικός κανόνας, υπάρχουν $(k+1)^l$ σχήματα για αλφάβητα με k στοιχεία. Η τάξη ενός σχήματος (order) , που δηλώνεται ως $o(H)$, είναι απλά ο αριθμός των σταθερών θέσεων (σε ένα δυαδικό αλφάβητο είναι ο αριθμός των άσσων και μηδενικών) που υπάρχει σε έναν κώδικα. Για παράδειγμα, η τάξη του σχήματος 011*1** είναι τέσσερα. Το καθορισμένο μήκος ενός σχήματος (defining length), που ορίζεται ως $\delta(H)$, είναι απλά η απόσταση μεταξύ της πρώτης και της τελευταίας σταθερής θέσης. Παραδείγματος χάριν, το σχήμα 011*1** έχει καθορισμένο μήκος τέσσερα, ενώ το σχήμα 0****** έχει μηδενικό καθορισμένο μήκος.

3.3.1 Η Επιρροή της Αναπαραγωγής, Διασταύρωσης και Μετάλλαξης στα Σχήματα

Έχει αποδειχτεί ότι με την αναπαραγωγή τα σχήματα που έχουν πάνω από τον μέσο όρο καταλληλότητα αυξάνονται εκθετικά, ενώ αυτά που έχουν κάτω από τον μέσο όρο καταλληλότητα μειώνονται εκθετικά με τον χρόνο. Έστω, για παράδειγμα, ότι την χρονική στιγμή t, υπάρχουν m παραδείγματα ενός συγκεκριμένου σχήματος H που βρίσκονται στον πληθυσμό $A(t)$, δηλαδή $m = m(H, t)$. Κατά την αναπαραγωγή, ένα

χρωμόσωμα επιλέγεται για αναπαραγωγή με πιθανότητα (όταν η μέθοδος επιλογής είναι η ρουλέτα) $P_i = \frac{F_i}{\sum_{j=1}^{N} F_j}$ σύμφωνα με την καταλληλότητα του F_i. Μετά την επιλογή ενός μη επικαλυπτόμενου πληθυσμού μεγέθους *N* με αντικατάσταση από τον πληθυσμό *A(t)* , αναμένεται να υπάρχουν *m(H, t+1)* αντιπρόσωποι του σχήματος *H* στον πληθυσμό την χρονική στιγμή *t+1*

$$m(H, t+1) = m(H, t)NF(H)/\sum F_j \qquad (3.15)$$

όπου *F(H)* είναι η μέση καταλληλότητα των χρωμοσωμάτων που αναπαριστούν το σχήμα *H* την χρονική στιγμή *t*. Η εξίσωση 3.15 μπορεί να γραφτεί και ως

$$m(H, t+1) = m(H, t)F(H)/\bar{F} \qquad (3.16)$$

όπου $\bar{F} = \sum F_j /N$.

Από την εξίσωση 3.16 φαίνεται ότι ένα σχήμα μεγαλώνει (αποκτά περισσότερους αντιπροσώπους με τον χρόνο) όταν μεγαλώνει η μέση καταλληλότητα του σχήματος ως προς την μέση καταλληλότητα του πληθυσμού. Αν υποτεθεί ότι ένα συγκεκριμένο σχήμα *H* βρίσκεται πάνω από την μέση καταλληλότητα κατά $c\bar{F}$ με *c* μία σταθερά, τότε

$$m(H, t+1) = m(H, t)(\bar{F} + c\bar{F})/\bar{F} \qquad (3.17)$$

Ξεκινώντας από $t = 0$ και υποθέτοντας μία σταθερή τιμή του *c*, προκύπτει τελικά

$$m(H, t) = m(H, 0)(1+ c)^t \qquad (3.18)$$

που αποδεικνύει μαθηματικά ότι τα σχήματα με καταλληλότητα μεγαλύτερη από την μέση καταλληλότητα του πληθυσμού αυξάνονται εκθετικά με τον χρόνο. Αντίστοιχα αποδεικνύεται ότι τα σχήματα με καταλληλότητα μικρότερη από την μέση καταλληλότητα μειώνονται εκθετικά με τον χρόνο.

Στη συνέχεια περιγράφεται η επιρροή της διασταύρωσης στα σχήματα. Έστω δύο σχήματα, H_1 και H_2, με H_1 = *1****0 και H_2 = ***10**. Τότε είναι φανερό ότι το σχήμα H_1 είναι λιγότερο πιθανό να επιβιώσει από το H_2. Η πιθανότητα επιβίωσης στην απλή διασταύρωση είναι $p_s = 1 - \delta(H)/ l-1$ όπου l είναι ο αριθμός των bits των χρωμοσωμάτων. Εάν η διασταύρωση εκτελείται με τυχαία επιλογή, έστω με πιθανότητα p_c, σε ένα συγκεκριμένο ζευγάρωμα, η πιθανότητα επιβίωσης δίνεται από την $p_s \geq 1 - p_c \delta(H)/(l-1)$.

Θεωρώντας ανεξαρτησία μεταξύ των τελεστών αναπαραγωγής και διασταύρωσης, το συνδυασμένο αποτέλεσμα των τελεστών αυτών στα σχήματα δίνεται από την εξίσωση

$$m(H, t+1) \geq m(H, t) \frac{F(H)}{\bar{F}} [1 - p_c \frac{\delta(H)}{l-1}] \qquad (3.19)$$

Επομένως το αν το σχήμα μεγαλώσει ή εξασθενήσει στις επόμενες γενεές εξαρτάται από δύο στοιχεία: (α) εάν η καταλληλότητα του σχήματος είναι πάνω ή κάτω από την μέση καταλληλότητα του πληθυσμού, και (β) εάν το σχήμα έχει μικρό ή μεγάλο καθορισμένο μήκος.

Όσον αφορά την επιρροή της μετάλλαξης στα σχήματα, επισημαίνεται ότι προκειμένου ένα σχήμα H να επιζήσει κατά την μετάλλαξη, θα πρέπει να επιζήσουν όλες οι σταθερές θέσεις. Ένα απλό γονίδιο (bit του χρωμοσώματος) επιβιώνει με πιθανότητα (1- p_m). Ένα συγκεκριμένο σχήμα επιβιώνει όταν κάθε μία από τις $o(H)$ σταθερές θέσεις μέσα στο σχήμα επιβιώνει. Επομένως η πιθανότητα ένα σχήμα να επιβιώσει κατά την διαδικασία της μετάλλαξης είναι $(1 - p_m)^{o(H)}$. Όταν η πιθανότητα p_m είναι πολύ μικρή, $p_m << 1$, τότε η πιθανότητα επιβίωσης του σχήματος γίνεται 1- $o(H)\ p_m$.

Επομένως, ο αριθμός των αντιπροσώπων ενός σχήματος στην επόμενη γενεά μετά την αναπαραγωγή, διασταύρωση και μετάλλαξη δίνεται από την εξίσωση 3.20 (αγνοώντας τους πολύ μικρούς όρους)

$$m(H,t+1) \geq m(H,t)\frac{F(H)}{\bar{F}}[1-p_c\frac{\delta(H)}{l-1}-o(H)p_m] \qquad (3.20)$$

Η εξίσωση 3.20 αποτελεί την εξίσωση του βασικού θεωρήματος των γενετικών αλγορίθμων που είναι γνωστό ως Θεώρημα Σχημάτων (Schema Theorem) [Holland75]. Σύμφωνα με αυτό το θεώρημα ο αριθμός των χαμηλής τάξης σχημάτων με μικρό καθορισμένο μήκος και με καταλληλότητα πάνω από την μέση καταλληλότητα του πληθυσμού αυξάνεται εκθετικά με τον χρόνο. Αυτό σημαίνει ότι οι γενετικοί αλγόριθμοι πετυχαίνουν τον στόχο τους με τη δειγματοληψία και το συνδυασμό χαμηλής τάξης, μεγάλης καταλληλότητας σχημάτων με μικρό καθορισμένο μήκος για να δημιουργήσουν χρωμοσώματα μεγαλύτερης καταλληλότητας. Εφόσον ο αριθμός των σχημάτων με καταλληλότητα χαμηλότερη από την μέση καταλληλότητα του πληθυσμού μειώνεται εκθετικά με τον χρόνο, η μέση καταλληλότητα του πληθυσμού θα αυξάνεται με τον χρόνο και ο αλγόριθμος σταδιακά θα συγκλίνει στη βέλτιστη λύση.

3.4 Οι Γενετικοί Αλγόριθμοι στην Επιλογή Χαρακτηριστικών

Το πρόβλημα της επιλογής χαρακτηριστικών είναι ένα πολύ σημαντικό θέμα στην αναγνώριση προτύπων. Στην επιλογή χαρακτηριστικών ερευνείται ένα υποσύνολο των χαρακτηριστικών ενός συνόλου δεδομένων, έτσι ώστε ο αλγόριθμος που εφαρμόζεται στα δεδομένα που περιέχουν μόνο αυτά τα χαρακτηριστικά να παράγει ένα μοντέλο ταξινόμησης με τη μεγαλύτερη δυνατή ακρίβεια. [Kohavi97]. Υπάρχουν δύο κύριες προσεγγίσεις στην επίλυση αυτού του προβλήματος: η προσέγγιση του φίλτρου (filter approach) και η προσέγγιση περιβλήματος (wrapper method). Και οι δύο προσεγγίσεις διαφέρουν μόνο στον τρόπο με τον οποίο αποτιμούν ένα δεδομένο υποσύνολο χαρακτηριστικών.

Στην προσέγγιση φίλτρου η επιλογή χαρακτηριστικών γίνεται ως ένα βήμα προεπεξεργασίας στον αλγόριθμο εκπαίδευσης. Αυτό σημαίνει ότι ο αλγόριθμος επιλογής χαρακτηριστικών επιλέγει πρώτα τα x πιο σχετικά χαρακτηριστικά και μετά το μοντέλο ταξινόμησης εκπαιδεύεται με τα δεδομένα εισόδου διάστασης x. Όσον αφορά την 'σχετικότητα' των χαρακτηριστικών υπάρχουν αρκετοί ορισμοί στην βιβλιογραφία [Kohavi97, σελ. 4, σελ. 4-5], [BluLan97, σελ. 2-4]. Αυτό το μέτρο

σχετικότητας δεν σχετίζεται, ωστόσο, με την απόδοση του συστήματος μάθησης. Αντίθετα, η προσέγγιση περιβλήματος εκτελεί μία αναζήτηση στον χώρο όλων των δυνατών υποσυνόλων χαρακτηριστικών. Έπειτα, για κάθε υποσύνολο χαρακτηριστικών που θεωρεί, εκπαιδεύει το μοντέλο ταξινόμησης και αποτιμάει το υποσύνολο χαρακτηριστικών εκτιμώντας την ικανότητα γενίκευσης (για παράδειγμα, το αναμενόμενο ρίσκο) του συστήματος μάθησης. Η αναζήτηση μπορεί να ξεκινήσει είτε με ένα κενό υποσύνολο χαρακτηριστικών και σταδιακά να προστίθενται χαρακτηριστικά (forward selection), είτε με το πλήρες σύνολο χαρακτηριστικών και σταδιακά να αφαιρούνται μερικά από αυτά (backward selection) [Kohavi97, σελ.9]. Είναι φανερό, ωστόσο, ότι η προσέγγιση περιβλήματος είναι ιδιαίτερα επίπονη σε περιπτώσεις μεγάλων χώρων αναζήτησης.

Ένας τελείως διαφορετικός τρόπος δειγματοληψίας μεγάλων χώρων αναζήτησης δίνεται από τους γενετικούς αλγόριθμους [FerKadKit93, BriBroMar92, RicLan96, RayPunGooKuhJai00]. Εάν ο συνολικός αριθμός των χαρακτηριστικών είναι d, τότε ο χώρος αναζήτησης είναι μεγέθους 2^d, εφόσον κάθε χαρακτηριστικό μπορεί είτε να επιλεγεί (να πάρει την τιμή 1), είτε όχι (να πάρει την τιμή 0) και επομένως κάθε υποσύνολο του $\{0, 1\}^d$ είναι μια δυνατή λύση. Η τυπική προσέγγιση είναι να θεωρηθεί κάθε υποσύνολο του χώρου $\{0, 1\}^d$ όλων των δυνατών συνδυασμών χαρακτηριστικών ως ένα δυαδικό χρωμόσωμα και να υπολογιστεί η καταλληλότητα του βάσει της ακρίβειας που επιτυγχάνεται από το μοντέλο ταξινόμησης όταν εκπαιδεύεται στα επιλεγμένα χαρακτηριστικά, είτε σε ένα ανεξάρτητο σύνολο ελέγχου, είτε μέσω της τεχνικής k-fold cross validation.

Όσον αφορά τις Μηχανές Διανύσματος Υποστήριξης, ο γενετικός αλγόριθμος μπορεί να χρησιμοποιηθεί όχι μόνο για την επιλογή των χαρακτηριστικών που θα ληφθούν στην ανάλυση, αλλά και για την βελτιστοποίηση της παραμέτρου κανονικοποίησης C της εξίσωσης 2.15 όπως και για την επιλογή παραμέτρου του πυρήνα που χρησιμοποιείται. Τα τρία αυτά θέματα εξετάζονται από την παρούσα εργασία και η ολοκληρωμένη μεθοδολογία περιγράφεται στην παράγραφο 4.3.

Κεφάλαιο 4

'Εκτίμηση Πιστωτικού Κινδύνου'

4.1 Εισαγωγικά

Τα συστήματα εκτίμησης πιστωτικού κινδύνου (credit scoring models) είναι ουσιαστικά τεχνικές που βοηθάνε τους οργανισμούς-επιχειρήσεις να αποφασίσουν εάν θα δώσουν πίστωση στους αιτούντες, εκτιμώντας τον κίνδυνο που εμπεριέχει μια τέτοια απόφαση. Τα συστήματα αυτά έχουν εξελιχθεί εντυπωσιακά κατά τη διάρκεια των τελευταίων 20 ετών εξαιτίας κάποιων γεγονότων που καταστούν την μέτρηση του πιστωτικού κινδύνου αναγκαιότερη από ποτέ. Μεταξύ αυτών των γεγονότων είναι: (α) η παγκόσμια αύξηση στον αριθμό των πτωχεύσεων, (β) πιο ανταγωνιστικά περιθώρια στα δάνεια, (γ) μια μειωμένη αξία των πραγματικών ενεργητικών στοιχείων (και έτσι της επιβοηθητικής εγγύησης) σε πολλές αγορές και (δ) μια δραματική αύξηση των στοιχείων εκτός ισολογισμού (off-balance sheet instruments) με έμφυτη την έκθεση κινδύνου μη αποπληρωμής [McKinsey93], συμπεριλαμβανομένων των παραγώγων πιστωτικού κινδύνου.

Τα μοντέλα πιστωτικού κινδύνου βασίζονται σε στατιστικές τεχνικές και τεχνικές από τον χώρο της επιχειρησιακής έρευνας. Στα στατιστικά εργαλεία συγκαταλέγονται (α) η διακριτική ανάλυση (discriminant analysis), (β) η λογιστική παλινδρόμηση και (γ) τα δέντρα ταξινόμησης. Οι τεχνικές από τον χώρο της επιχειρησιακής έρευνας αφορούν διάφορες παραλλαγές του γραμμικού προγραμματισμού. Εκτός από τα παραπάνω μοντέλα, υπάρχουν και διάφορες μη παραμετρικές μεθοδολογίες μοντελοποίησης που προέρχονται από τον χώρο της τεχνητής νοημοσύνης. Ανάμεσα σε αυτές τις προσεγγίσεις είναι τα νευρωνικά δίκτυα

(neural networks), τα έμπειρα συστήματα (expert systems), οι γενετικοί αλγόριθμοι και οι μέθοδοι πλησιέστερου γείτονα (nearest neighbor methods).

Προκειμένου να χρησιμοποιηθούν οι παραπάνω μέθοδοι, λαμβάνεται αρχικά ένα δείγμα από αιτούντες, που μπορεί να ποικίλει από λίγες χιλιάδες μέχρι και εκατοντάδες χιλιάδες (αυτό δεν αποτελεί πρόβλημα σε μία βιομηχανία όπου οι εταιρίες συχνά διαθέτουν τα χαρτοφυλάκια δεκάδων εκατομμυρίων πελατών). Για κάθε πελάτη στο δείγμα, γίνεται χρήση κάποιων πληροφοριών που τον αφορούν. Εάν οι πελάτες είναι εταιρείες, οι πληροφορίες αντλούνται συνήθως από τον ισολογισμό των εταιριών καθώς και από την πιστωτική τους ιστορία, εάν οι πελάτες είναι ιδιώτες οι πληροφορίες αντλούνται από το δελτίο αίτησης που έχουν συμπληρώσει και αφορά πληροφορίες όπως την οικογενειακή, οικονομική και επαγγελματική τους κατάσταση. Επιπλέον λαμβάνεται υπόψη και η ιστορία της αξιοπιστίας τους σε ένα δεδομένο χρονικό διάστημα – για παράδειγμα δώδεκα, δεκαοχτώ ή εικοσιτέσσερις μήνες. Έπειτα εξετάζεται εάν αυτή η ιστορία είναι αποδεκτή, δηλαδή εάν ο πελάτης μπορεί να θεωρηθεί ως ασυνεπής. Συνήθως ένας πελάτης χαρακτηρίζεται ως ασυνεπής, εάν δεν έχει ανταποκριθεί σε τρεις συνεχόμενες μηνιαίες πληρωμές. Θα υπάρχει, ωστόσο ένας αριθμός πελατών για τους οποίους δεν θα είναι δυνατό να εξακριβωθεί εάν είναι συνεπείς ή ασυνεπείς, λόγω του ότι δεν υπήρξαν πελάτες για πολύ καιρό ή επειδή η ιστορία τους δεν είναι ξεκάθαρη. Αυτό το σύνολο πελατών συνηθίζεται να αφαιρείται από το δείγμα.

Ένα άλλο ανοιχτό ζήτημα είναι η αναλογία των συνεπών και ασυνεπών πελατών που θα πρέπει να υπάρχει στο δείγμα, εάν δηλαδή θα πρέπει να αντικατοπτρίζει τις αναλογίες στον πληθυσμό ή εάν θα ήταν προτιμότερο να υπάρχει ίσος αριθμό συνεπών και ασυνεπών πελατών στο δείγμα. Ο Henley [Henley95] περιγράφει μερικά από αυτά τα ζητήματα στο διδακτορικό του.

Η εκτίμηση πιστωτικού κινδύνου μπορεί τελικά να θεωρηθεί ως ένα πρόβλημα ταξινόμησης, όπου τα χαρακτηριστικά εισόδου είναι οι πληροφορίες - δεδομένα που λαμβάνονται για τους πελάτες, ενώ η έξοδος είναι ο διαχωρισμός των πελατών σε συνεπείς και ασυνεπείς. Το ζητούμενο είναι ο διαχωρισμός του συνόλου των πληροφοριών A σε δύο υποσύνολα: (α) στα δεδομένα για εκείνους τους πελάτες που αποδείχτηκαν ασυνεπείς, $\mathbf{x} \in A_B$ και (β) στα δεδομένα για εκείνους τους πελάτες

που αποδείχτηκαν συνεπείς, $\mathbf{x} \in A_G$. Ο κανόνας για τους καινούριους αιτούντες τότε θα είναι: αποδοχή της αίτησης τους εάν τα δεδομένα για αυτούς βρίσκονται στο σύνολο A_G, απόρριψη της αίτησης τους εάν τα δεδομένα για αυτούς βρίσκονται στο σύνολο A_B.

Είναι, επίσης, απαραίτητο να υπάρχει συνέπεια και συνέχεια σε αυτά τα σύνολα και επομένως δεν είναι δυνατό να ταξινομηθούν όλοι οι πελάτες στο δείγμα σωστά. Η τέλεια ταξινόμηση δεν είναι εφικτή, εφόσον μερικές φορές το ίδιο σύνολο δεδομένων αποδίδεται και σε έναν συνεπή πελάτη και σε έναν ασυνεπή. Ωστόσο, το ζητούμενο είναι ένας κανόνας που θα ταξινομεί λάθος όσο το δυνατό λιγότερους πελάτες και επιπλέον θα ικανοποιεί κάποιες λογικές απαιτήσεις συνέχειας.

4.2 Μέτρηση Πιστωτικού Κινδύνου

4.2.1 Έμπειρα Συστήματα και Υποκειμενική Ανάλυση

Πριν από είκοσι χρόνια, οι περισσότεροι οικονομικοί οργανισμοί βασίζονταν κυρίως στην υποκειμενική ανάλυση ή στα αποκαλούμενα έμπειρα συστήματα τραπεζών για να εκτιμούν τον πιστωτικό κίνδυνο στα εταιρικά δάνεια. Συγκεκριμένα, οι τράπεζες χρησιμοποιούσαν πληροφορίες για διάφορα χαρακτηριστικά των πελατών τους, όπως για παράδειγμα τον χαρακτήρα και την φήμη τους (Character), το κεφάλαιό που ζητάνε (Capital), την δυναμικότητά τους (Capacity) δηλαδή την ικανότητα τους να αποπληρώσουν τα χρέη τους και την μεταβλητότητα που υπάρχει στα κέρδη τους, τις εγγυήσεις τους (Collateral) και τις συνθήκες (Conditions) της αγοράς. Τα παραπάνω είναι γνωστά σαν τα 5Cs που οδηγούσαν σε μία υποκειμενική κρίση ενός ειδικού για το αν η τράπεζα θα δώσει τελικά πίστωση ή όχι στον πελάτη.

Σε μια δημοσίευση τους, οι Sommerville και Taffler, [SommerTaffler95] έδειξαν ότι (α) οι αποφάσεις που παίρνουν οι τράπεζες βασιζόμενες σε υποκειμενικές αξιολογήσεις, όπως παραπάνω, τείνουν να είναι σε μεγάλο βαθμό απαισιόδοξες ως προς τον πιστωτικό κίνδυνο των πελατών, και (β) τα πολυμεταβλητά συστήματα εκτίμησης κινδύνου που θα περιγραφτούν παρακάτω αποδίδουν καλύτερα από τα έμπειρα συστήματα. Γι αυτό άλλωστε τα τελευταία είκοσι χρόνια σημειώθηκε

στροφή των χρηματοοικονομικών ιδρυμάτων από τα έμπειρα συστήματα σε συστήματα που βασίζονται σε πιο αντικειμενικές αξιολογήσεις.

4.2.2 Βασισμένα σε Χρηματοοικονομικά Μεγέθη Συστήματα Εκτίμησης Πιστωτικού Κινδύνου

Στα μονομεταβλητά βασισμένα σε χρηματοοικονομικά μεγέθη συστήματα εκτίμησης πιστωτικού κινδύνου, το χρηματοοικονομικό ίδρυμα συγκρίνει διάφορους βασικούς χρηματοοικονομικούς δείκτες και συναφή μεγέθη των πιθανών πελατών με τα πρότυπα βιομηχανίας ή ομάδας. Κατά την χρησιμοποίηση πολυμεταβλητών μοντέλων, οι βασικές μεταβλητές συνδυάζονται και σταθμίζονται για να παραγάγουν είτε μία εκτίμηση του πιστωτικού κινδύνου είτε μια πιθανότητα αθέτησης της υποχρέωσης του πελάτη. Εάν η εκτίμηση του πιστωτικού κινδύνου, ή η πιθανότητα αθέτησης, έχει τιμή μεγαλύτερη από ένα κρίσιμο όριο, τότε ο πελάτης είτε απορρίπτεται είτε υποβάλλεται σε διεξοδικότερη έρευνα.

Όσον αφορά τα πολυμεταβλητά μοντέλα εκτίμησης πιστωτικού κινδύνου, υπάρχουν τουλάχιστον τέσσερις μεθοδολογικές προσεγγίσεις στην ανάπτυξη τους: (α) το γραμμικό μοντέλο πιθανότητας (linear probability model), (β) το λογιστικό μοντέλο (logit model), (γ) το probit μοντέλο, και (δ) το μοντέλο διακριτικής ανάλυσης (discriminant analysis model). Οι μεθοδολογίες που κυριαρχούσαν στις δημοσιεύσεις των περιοδικών, ήταν η διακριτική ανάλυση ακολουθούμενη από την λογιστική ανάλυση.

Ο Altman et al. [Altman77] ανέπτυξε το σήμερα πλέον ευρέως διαδεδομένο και χρησιμοποιημένο διακριτικό μοντέλο ZETA®. Πρόκειται για την πιο κοινή μορφή διακριτικής ανάλυσης η οποία επιδιώκει να βρει μια γραμμική συνάρτηση ορισμένων χρηματοοικονομικών μεταβλητών που διακρίνουν καλύτερα τις δύο κατηγορίες πελατών (συνεπείς και ασυνεπείς). Αυτή η μέθοδος απαιτεί την ανάλυση ενός συνόλου μεταβλητών που θα μεγιστοποιεί την μεταβλητότητα μεταξύ των μεταβλητών των δύο ομάδων ενώ θα ελαχιστοποιεί την μεταβλητότητα μεταξύ των μεταβλητών της ίδιας ομάδας. Παρόμοια, η λογιστική ανάλυση χρησιμοποιεί ένα σύνολο μεταβλητών για να προβλέψει την πιθανότητα ασυνέπειας ενός πελάτη,

υποθέτοντας ότι αυτή η πιθανότητα μοντελοποιείται μέσω μιας λογιστικής συνάρτησης που εξ' ορισμού περιορίζει τις πιθανότητες μεταξύ 0 και 1.

Ο Martin [Martin77] χρησιμοποίησε και τη λογιστική και τη διακριτική ανάλυση για να προβλέψει τις αποτυχίες τραπεζών στην περίοδο 1975-1976. Και τα δύο μοντέλα έδωσαν παρόμοιες ταξινομήσεις όσον αφορά την αναγνώριση των αποτυχιών και των μη αποτυχιών. Ο West [West85] χρησιμοποίησε το λογιστικό μοντέλο (μαζί με την παραγοντική ανάλυση) για να μετρήσει τις οικονομικές συνθήκες των χρηματοοικονομικών ιδρυμάτων και να τους προσδώσει μια πιθανότητα για το αν είναι προβληματικές τράπεζες. Το ενδιαφέρον αποτέλεσμα που προέκυψε είναι ότι οι παράγοντες που καθορίστηκαν από το λογιστικό μοντέλο είναι παρόμοιοι με τους συντελεστές βαθμολόγησης του CAMEL μοντέλου που χρησιμοποιείται από τους εξεταστές τραπεζών.

Ο Lawrence et al. [Lawrence92] χρησιμοποίησε το λογιστικό μοντέλο για να προβλέψει την πιθανότητα ασυνέπειας των στεγαστικών δανείων και έφτασε στο συμπέρασμα ότι η ιστορία πληρωμής είναι η πλέον σημαντικότερη μεταβλητή πρόβλεψης. Οι Smith και Lawrence [SmithLawrence95] χρησιμοποιούν ένα λογιστικό μοντέλο για να βρουν τις μεταβλητές που δίνουν την καλύτερη πρόβλεψη ενός δανείου που δεν θα αποπληρωθεί.

4.2.3 Σύγχρονα Μοντέλα Εκτίμησης Πιστωτικού Κινδύνου

Αν και τα πολυμεταβλητά χρηματοοικονομικά μοντέλα εκτίμησης πιστωτικού κινδύνου έχουν αποδεδειγμένα αποδώσει αρκετά καλά σε διαφορετικά χρονικά διαστήματα και σε πολλές διαφορετικές χώρες, υπόκεινται σε τουλάχιστον τρεις κριτικές. Πρώτα απ' όλα, επειδή είναι κυρίως βασισμένα σε χρηματοοικονομικά δεδομένα τα οποία μετριούνται σε διακριτά χρονικά διαστήματα, αποτυγχάνουν να συλλάβουν μικρότερες και πιο γρήγορες αλλαγές στις συνθήκες της αγοράς. Δεύτερον, ο κόσμος είναι εγγενώς μη γραμμικός, έτσι ώστε η γραμμική διακριτική ανάλυση και τα γραμμικά μοντέλα πιθανότητας να αποτυγχάνουν να κάνουν προβλέψεις με τόση ακρίβεια όση επιτυγχάνουν μέθοδοι που χαλαρώνουν την υπόθεση της γραμμικότητας μεταξύ των επεξηγηματικών μεταβλητών. Τέλος, τα μοντέλα πρόβλεψης πτώχευσης που περιγράφτηκαν στην παράγραφο 4.2.2, δεν

βασίζονται πάντα σε κάποιο θεωρητικό μοντέλο. Έτσι, προέκυψε ένα πλήθος νέων προσεγγίσεων, οι περισσότερες από τις οποίες είχαν διερευνητική φύση, και προτάθηκαν ως εναλλακτικές λύσεις των παραδοσιακών μοντέλων εκτίμησης πιστωτικού κινδύνου.

Μια κατηγορία μοντέλων πτώχευσης με ισχυρό θεωρητικό υπόβαθρο είναι τα λεγόμενα 'risk of ruin' μοντέλα. Στην απλούστερη εκδοχή , μία εταιρεία χρεοκοπεί όταν η αγοραστική αξία των ενεργητικών της στοιχείων (A) πέσει κάτω από τις υποχρεώσεις της σε εξωτερικούς πιστωτές (B). Μοντέλα αυτού του τύπου μπορούν να βρεθούν στους [Wilcox73], [Scott81], [SantomeroVinso77]. Όπως αναγνωρίστηκε από τον Scott, αυτά τα μοντέλα είναι από πολλές απόψεις παρόμοια με τα μοντέλα αποτίμησης δικαιωμάτων (Option Pricing Models - OPM) των Black και Scholes [BlackScholes73], καθώς επίσης και με εκείνα του Merton [Merton74] και των Hull και White [HullWhite95]. Στο μοντέλο των Black, Scholes και Merton, η πιθανότητα πτώχευσης μιας εταιρείας εξαρτάται αποφασιστικά από την αγοραία αξία του ενεργητικού της εταιρείας στην αρχή της περιόδου (A), σχετικά με το εξωτερικό χρέος της (B), καθώς επίσης και την μεταβλητότητα της αγοραίας αξίας των ενεργητικών στοιχείων της εταιρείας σ_A.

Οι ιδέες των risk of ruin και OPM μοντέλων έχουν κερδίσει την αυξανόμενη εμπιστοσύνη στην εμπορική περιοχή. Ένα χαρακτηριστικό παράδειγμα είναι το KMV [KMV93] μοντέλο και το μοντέλο του Kealhofer [Kealhofer96]. Στο KMV μοντέλο, οι κρίσιμες είσοδοι στην εκτίμηση της πιθανότητας ασυνέπειας είναι το A και το σ_A. Οι τιμές των παραμέτρων αυτών μπορούν να εκτιμηθούν για όλες τις επιχειρήσεις εφόσον υπάρχουν επαρκή δεδομένα για τις αποπληρωμές των δανείων. Βάσει των στοιχείων αυτών υπολογίζεται μια αναμενόμενη συχνότητα αθέτησης (expected default frequency - EDF) για κάθε εταιρεία, η οποία αναπαριστά την πιθανότητα σε κάποια μελλοντική περίοδο η αξία των ενεργητικών στοιχείων της εταιρείας να κυμανθεί σε επίπεδα χαμηλότερα των βραχυπρόθεσμων υποχρεώσεων της.

Ένα άλλο μοντέλο βασισμένο στην κεφαλαιαγορά είναι το μοντέλο ποσοστού θνησιμότητας (mortality rate model) του Altman [Altman88], [Altman89] και η προσέγγιση γήρανσης (aging approach) του Asquith et al. [Asquith89]. Αυτά τα μοντέλα ποσοστού θνησιμότητας - αθέτησης επιδιώκουν να αντλήσουν τις

πραγματικές πιθανότητες αθέτησης από προηγούμενα δεδομένα για αθετήσεις ομολόγων από τον βαθμό πίστωσης και τη χρονική τους διάρκεια. Όλοι οι οργανισμοί βαθμολόγησης έχουν υιοθετήσει και τροποποιήσει την προσέγγιση θνησιμότητας (για παράδειγμα η Moody's [Moody's90] και η Standard and Poor's [Standard and Poor's91]) και τώρα την χρησιμοποιούν ευρέως στις δομημένες οικονομικές αναλύσεις των εγγράφων τους. [McElraveyShah96].

Μοντέλα σαν τα παραπάνω έχουν τη δυνατότητα να επεκταθούν για την πιθανότητα της μη αποπληρωμής των δανείων, αλλά αυτό συχνά είναι δύσκολο λόγω της έλλειψης μεγάλων βάσεων για δεδομένα δανείων που βρίσκονται σε κατάσταση ασυνέπειας. Για παράδειγμα, οι McAllister και Mingo [McAllisterMingo94] εκτιμούν ότι για την ανάπτυξη αξιόπιστων εκτιμήσεων των πιθανοτήτων ασυνέπειας, ένα χρηματοοικονομικό ίδρυμα θα χρειαζόταν 20.000 με 30.000 περιπτώσεις ασυνέπειας στη βάση δεδομένων του. Πολύ λίγα χρηματοοικονομικά ιδρύματα διαθέτουν όμως τέτοιες βάσεις δεδομένων. Αυτό μπορεί να εξηγήσει διάφορες τρέχουσες πρωτοβουλίες στις ΗΠΑ, μεταξύ των μεγαλύτερων τραπεζών, προς την ανάπτυξη μιας κοινής εθνικής βάσης από δεδομένα ποσοστών θνησιμότητας - μη αποπληρωμής δανείων (ένα τρέχον πρόγραμμα της Robert Morris Associates στην Φιλαδέλφεια).

Μια νεώτερη προσέγγιση είναι η εφαρμογή της ανάλυσης νευρωνικών δικτύων στο πρόβλημα της ταξινόμησης πιστωτικού κινδύνου. Ουσιαστικά, η ανάλυση των νευρωνικών δικτύων είναι παρόμοια με μια μη γραμμική διακριτική ανάλυση, δεδομένου ότι δεν υποθέτουν ότι οι μεταβλητές που εισάγονται στη συνάρτηση πρόβλεψης πτώχευσης συσχετίζονται γραμμικά και ανεξάρτητα. Συγκεκριμένα, τα μοντέλα νευρωνικών δικτύων για τον πιστωτικό κίνδυνο ερευνούν ενδεχομένως κρυμμένες συσχετίσεις μεταξύ των μεταβλητών πρόβλεψης που εισάγονται έπειτα ως πρόσθετες επεξηγηματικές μεταβλητές στη μη γραμμική συνάρτηση πρόβλεψης πτώχευσης. Οι εφαρμογές των νευρωνικών δικτύων στην ανάλυση πρόβλεψης κινδύνου περιλαμβάνουν έρευνες που έχουν γίνει από τον Altman et al. [Altman94], τους Coats και Fant [CoatsFant93] και διάφορες μελέτες που συνοψίζονται σους Trippi και Turban [TrippiTurban96].

Μία άλλη μη παραμετρική στατιστική προσέγγιση στην εκτίμηση πιστωτικού κινδύνου είναι η λεγόμενη μέθοδος των πλησιέστερων γειτόνων (nearest neighbors).

Στη μεθοδολογία αυτή ένας υποψήφιος θα ταξινομηθεί σύμφωνα με το σε ποια ομάδα – συνεπών, ασυνεπών πελατών – ανήκει η πλειοψηφία των πλησιέστερων γειτόνων του. Ο εντοπισμός των πλησιέστερων γειτόνων πραγματοποιείται υπολογίζοντας την απόσταση του εξεταζόμενου πελάτη με το σύνολο των δεδομένων στο δείγμα εκμάθησης, συνήθως χρησιμοποιώντας την ευκλείδεια απόσταση. Η ανάλυση των Henley και Hand [HenleyHand96] υποθέτει ότι η ταξινόμηση είναι ευσταθής στην επιλογή των πλησιέστερων γειτόνων που θα θεωρηθούν, και το σύστημα έχει το πλεονέκτημα ότι καινούρια δεδομένα μπορούν να προστεθούν και επομένως μπορεί να αναβαθμιστεί το σύστημα χωρίς αλλαγή στον κώδικα.

4.3 Προτεινόμενη Μεθοδολογία για την Εκτίμηση Πιστωτικού Κινδύνου

Η προτεινόμενη μεθοδολογία βασίζεται στην χρήση γενετικών αλγορίθμων για την επιλογή (α) των χαρακτηριστικών – κριτηρίων στην εκτίμηση πιστωτικού κινδύνου, (β) του πυρήνα και (γ) της παραμέτρου του πυρήνα των Μηχανών Διανυσμάτων Υποστήριξης. Όπως περιγράφτηκε στην παράγραφο 3.4, κάθε υποσύνολο του χώρου $\{0, 1\}^d$ όλων των δυνατών συνδυασμών χαρακτηριστικών - κριτηρίων θεωρείται ως ένα δυαδικό χρωμόσωμα και υπολογίζεται η καταλληλότητα του βάσει της ακρίβειας που επιτυγχάνεται από το μοντέλο ταξινόμησης όταν εκπαιδεύεται στα επιλεγμένα χαρακτηριστικά.

Επιπλέον, ο γενετικός αλγόριθμος χρησιμοποιείται στην παρούσα εργασία για την βελτιστοποίηση της παραμέτρου κανονικοποίησης C της εξίσωσης 2.15 όπως και για την επιλογή παραμέτρου του πυρήνα που χρησιμοποιείται. Για παράδειγμα, όταν χρησιμοποιείται πολυωνυμικός πυρήνας, γίνεται η επιλογή του βαθμού του πολυωνύμου, ενώ όταν χρησιμοποιείται RBF πυρήνας, γίνεται η επιλογή της παραμέτρου σ. Εφόσον το χρωμόσωμα είναι δυαδικό, τότε και η παράμετρος C όπως και η παράμετρος του πυρήνα μπορούν να αναπαρασταθούν δυαδικά και να ενσωματωθούν στο χρωμόσωμα. Αυτό σημαίνει ότι ο γενετικός αλγόριθμος προσπαθεί να επιλέξει το βέλτιστο υποσύνολο χαρακτηριστικών ταυτόχρονα με τη βέλτιστη παράμετρο C και την βέλτιστη παράμετρο του πυρήνα. Αυτό είναι λογικό, καθώς η επιλογή της παραμέτρου C και της παραμέτρου του πυρήνα επηρεάζονται από το υποσύνολο των χαρακτηριστικών που θεωρείται και αντίστροφα.

Όσον αφορά την παράμετρο *C* μόνο με τρία δυαδικά ψηφία μπορούν να αναπαρασταθούν οι αριθμοί 0 έως 7. Εάν μετατοπίσουμε την αναπαράσταση αυτή κατά -3 και υψώσουμε το αποτέλεσμα σαν δύναμη του δέκα προκύπτουν οι εξής δυνατές τιμές για την παράμετρο *C*: 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000. Αντίστοιχα και η παράμετρος του πυρήνα μπορεί να αναπαρασταθεί με τρία δυαδικά ψηφία. Ο βαθμός του πολυωνύμου βρίσκεται μετατοπίζοντας τον αριθμό που προκύπτει κατά +2. Έτσι προκύπτουν πολυώνυμα βαθμού 2 έως 9. Η παράμετρος *σ* του RBF πυρήνα βρίσκεται μετατοπίζοντας τον αριθμό κατά – 6 και υψώνοντας το αποτέλεσμα που προκύπτει σαν δύναμη του δύο. Έτσι προκύπτουν οι δυνατές τιμές 2^{-6} , 2^{-5}, 2^{-4},...,2 για την παράμετρο *σ*. Οι παραπάνω παράμετροι μπορούν δυνητικά να πάρουν οποιεσδήποτε πραγματικές τιμές, ωστόσο ο στόχος είναι να ερευνηθεί ένα μεγάλο εύρος από λογικές τιμές των παραμέτρων και όχι ένας εξονυχιστικός έλεγχος όλων των δυνατών τιμών.

Επομένως, οι γενετικοί αλγόριθμοι στην παρούσα εργασία χρησιμοποιούν δυαδικά χρωμοσώματα που το αρχικό τους μέρος αποτελείται από τόσα γονίδια-bit όσα και τα συνολικά χαρακτηριστικά – κριτήρια στην εκτίμηση πιστωτικού κινδύνου (το 1 θα δηλώνει ότι το συγκεκριμένο κριτήριο λαμβάνεται υπόψη στην εκπαίδευση, ενώ το 0 ότι δεν λαμβάνεται), τα επόμενα τρία bit αναπαριστούν την τιμή της παραμέτρου *C* και τα τρία τελευταία την παράμετρο του πυρήνα. Η καταλληλότητα του κάθε χρωμοσώματος υπολογίζεται βάση της ακρίβειας που πετυχαίνεται από τις Μηχανές Διανύσματος Υποστήριξης ως μοντέλου ταξινόμησης όταν εκπαιδεύεται στα επιλεγμένα χαρακτηριστικά, την παράμετρο *C* και την παράμετρο του πυρήνα που προκύπτουν από το εν λόγω χρωμόσωμα.

Για τον υπολογισμό της ακρίβειας που επιτυγχάνεται από την πληροφορία ενός χρωμοσώματος προκειμένου να εκτιμηθεί η καταλληλότητα του, χρησιμοποιείται στην παρούσα εργασία η τεχνική k-fold cross-validation, όπως περιγράφτηκε στην παράγραφο 2.2.3. Συγκεκριμένα το σύνολο εκπαίδευσης διαιρείται σε 5 υποσύνολα , και η μέθοδος επαναλαμβάνεται 5 φορές. Κάθε φορά, ένα από τα 5 υποσύνολα χρησιμοποιείται για τον έλεγχο και τα υπόλοιπα 4 υποσύνολα τίθενται μαζί και χρησιμοποιούνται για την εκπαίδευση. Κατόπιν η ακρίβεια υπολογίζεται κατά μέσο όρο σε όλες τις 5 δοκιμές.

Τελικά, ο γενετικός αλγόριθμος συγκλίνει σε μία λύση-χρωμόσωμα που οδηγεί στην μέγιστη δυνατή ακρίβεια και δίνει την βέλτιστη τιμή της παραμέτρου *C*, την βέλτιστη τιμή του πυρήνα και ορίζει εκείνα τα χαρακτηριστικά - κριτήρια που είναι τα πλέον σχετικά και θα πρέπει να λαμβάνονται υπόψη κατά την εκτίμηση πιστωτικού κινδύνου.

4.3.1 Δεδομένα

Χρησιμοποιήθηκαν τρία σύνολα δεδομένων προκειμένου να εξεταστεί η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας. Το πρώτο σύνολο δεδομένων αφορά επιχειρήσεις από το χαρτοφυλάκιο χορηγήσεων μιας Ελληνικής Τράπεζας [DoumpoKosmBaouZop02]. Συνολικά εξετάζονται 1411 εταιρίες, εκ των οποίων οι 1000 χρησιμοποιούνται στο δείγμα εκπαίδευσης και οι 411 στο δείγμα ελέγχου.

Πίνακας 4.1 : Κριτήρια για το πρώτο σύνολο δεδομένων

Κριτήρια	**Ερμηνεία**
A1	Κέρδη προ τόκων και φόρων / Σύνολο Ενεργητικού
A2	Πωλήσεις / Σύνολο Ενεργητικού
A3	Μικτά Κέρδη / Σύνολο Ενεργητικού
A4	Συνολικές Υποχρεώσεις / Σύνολο Ενεργητικού
A5	Μακροπρόθεσμες Υποχρεώσεις / (Μακροπρόθ. Υποχρ. + Ίδια Κεφάλαια)
A6	Άμεση Ρευστότητα
A7	Βραχυπρόθεσμες Υποχρεώσεις / Ίδια Κεφάλαια
A8	Συνολικές Υποχρεώσεις / Κεφάλαιο Κίνησης

Βάσει των διαθέσιμων οικονομικών στοιχείων των εταιριών, εξετάζονται 8 κριτήρια ως επαρκή μέτρα του πιστωτικού κινδύνου εταιριών τα οποία παρουσιάζονται στον πίνακα 4.1. Η επιλογή αυτών των κριτηρίων έχει εκτελεσθεί με τη συνεργασία των ειδικών αναλυτών πιστωτικού κινδύνου από την εμπορική τράπεζα της Ελλάδας προκειμένου να εξεταστεί η πολιτική πιστωτικού κινδύνου της τράπεζας και η οικονομική ανάλυση που χρησιμοποιείται στην καθημερινή πρακτική από τους αναλυτές πιστωτικού κινδύνου. Πρέπει επίσης να παρατηρηθεί ότι σύμφωνα με τη διεθνή οικονομική βιβλιογραφία [Courtis78] τα επιλεγμένα κριτήρια καλύπτουν όλες

τις πτυχές της χρηματοοικονομικής απόδοσης των εταιριών, συμπεριλαμβανομένης της αποδοτικότητας, της φερεγγυότητας και της διευθυντικής απόδοσης.

Τα κριτήρια Α1, Α2 και Α3 συσχετίζεται με την αποδοτικότητα των εταιριών. Υψηλές τιμές αυτών των αναλογιών αντιστοιχούν σε κερδοφόρες εταιρίες. Κατά συνέπεια, όλες αυτές οι αναλογίες συσχετίζονται αρνητικά με τον πιστωτικό κίνδυνο. Το κριτήριο Α6 αφορά την ρευστότητα της εταιρίας. Οι εταιρίες που έχουν μεγάλη ρευστότητα μπορούν πιο εύκολα να αντεπεξέλθουν στις βραχυπρόθεσμες υποχρεώσεις τους στους πιστωτές τους. Κατά συνέπεια, και αυτό το κριτήριο συσχετίζεται αρνητικά με τον πιστωτικό κίνδυνο. Τέλος, τα κριτήρια Α4, Α5, Α7 και Α8 συσχετίζονται με τη φερεγγυότητα (οικονομική δύναμη) των εταιριών. Υψηλές τιμές δείχνουν υψηλό δανεισμό, δηλαδή ότι οι εταιρίες πρέπει να παράγουν περισσότερο εισόδημα για να εκπληρώσουν τις υποχρεώσεις τους και να ξεπληρώσουν τα χρέη τους. Συνεπώς αυτά τα κριτήρια συσχετίζονται θετικά με τον πιστωτικό κίνδυνο.

Το δεύτερο σύνολο δεδομένων προέρχεται από μια έρευνα του Τμήματος Στατιστικής και Οικονομετρίας του Πανεπιστημίου του Αμβούργου και αφορά δεδομένα για 1000 πελάτες εκ των οποίων οι 700 χρησιμοποιήθηκαν στο δείγμα εκπαίδευσης και οι 300 στο δείγμα ελέγχου. Τα δεδομένα αυτά αντλήθηκαν από μια βάση δεδομένων για την εκτίμηση πιστωτικού κινδύνου που ονομάζεται UCI Machine Learning Repository. Εξετάστηκαν 7 ποσοτικά και 13 ποιοτικά κριτήρια με τις δυνατές καταστάσεις τους, όπως φαίνονται στον πίνακα 4.2. Για κάθε ποιοτική μεταβλητή χρησιμοποιήθηκε μία κωδικοποίηση, προκειμένου να εφαρμοστεί ο γενετικός αλγόριθμος. Για κάθε ποιοτική μεταβλητή, δημιουργούνται τόσες στήλες όσες είναι οι δυνατές καταστάσεις της μείον 1, δηλαδή για παράδειγμα για την ποιοτική μεταβλητή Β1 δημιουργούνται τρεις στήλες. Εάν ισχύει η κατάσταση Β11, τότε η πρώτη στήλη συμπληρώνεται με έναν άσσο και οι άλλες δύο με μηδενικά. Η συνθήκη Β14 ικανοποιείται όταν και οι τρεις στήλες συμπληρωθούν με μηδέν. Κατά αντιστοιχία κωδικοποιούνται και οι υπόλοιπες ποιοτικές μεταβλητές. Όσον αφορά τις ποσοτικές μεταβλητές, αυτές κανονικοποιήθηκαν μεταξύ 0 και 1.

Πίνακας 4.2 : Κριτήρια για το δεύτερο σύνολο δεδομένων

B1	**Λογαριασμός Επιταγών :**
B11	...< 0 Γερμανικά μάρκα
B12	0 <=...< 200 Γερμανικά Μάρκα
B13	...>= 200 Γερμανικά Μάρκα
B14	Δεν υπάρχει Λογαριασμός Επιταγών
B2	**Διάρκεια σε μήνες**
B3	**Πιστωτική Ιστορία :**
B31	Δεν υπάρχει ή όλες οι υποχρεώσεις αποπληρώθηκαν δεόντως
B32	Όλες οι υποχρεώσεις στην συγκεκριμένη τράπεζα αποπληρώθηκαν δεόντως
B33	Οι υπάρχουσες υποχρεώσεις έχουν ξεπληρωθεί δεόντως μέχρι τώρα
B34	Καθυστέρηση στην εκπλήρωση υποχρεώσεων στο παρελθόν
B35	Υπάρχουσα πίστωση σε άλλες τράπεζες
B4	**Σκοπός :**
B41	Αγορά καινούριου αυτοκινήτου
B42	Αγορά χρησιμοποιημένου αυτοκινήτου
B43	Αγορά επίπλων / εξοπλισμού
B44	Αγορά ράδιο / τηλεόρασης
B45	Αγορά οικιακών συσκευών
B46	Επιδιορθώσεις
B47	Εκπαίδευση
B48	Επανεκπαίδευση
B49	Επιχειρήσεις
B50	Άλλα
B5	**Ποσό πίστωσης**
B6	**Λογαριασμός Καταθέσεων :**
B61	...< 100 Γερμανικά Μάρκα
B62	100 <= ... < 500 Γερμανικά Μάρκα
B63	500 <= ... < 1000 Γερμανικά Μάρκα
B64	... >= 1000 Γερμανικά Μάρκα
B65	Άγνωστος ή δεν υπάρχει Λογαριασμός Καταθέσεων
B7	**Χρονικό διάστημα που είναι εργαζόμενος :**
B71	Άνεργος
B72	... < 1 χρόνο
B73	1 <= ... < 4 χρόνια
B74	4 <= ... < 7 χρόνια
B75	... >= 7 χρόνια
B8	**Αξία δόσης σαν ποσοστό εισοδήματος**
B9	**Οικογενειακή Κατάσταση και φύλο :**
B91	Άντρας χωρισμένος

B92	Γυναίκα χωρισμένη ή παντρεμένη
B93	Άντρας ελεύθερος
B94	Άντρας παντρεμένος ή χήρος
B10	**Άλλοι χρεώστες ή εγγυητές :**
B101	Κανένας
B102	Χρεώστης
B103	Εγγυητής
B11	**Χρονικό διάστημα που είναι μόνιμος κάτοικος**
B12	**Ιδιοκτησία :**
B121	Ακίνητα
B122	Ασφάλεια Ζωής
B123	Αυτοκίνητο ή άλλα
B124	Άγνωστη ή δεν υπάρχει
B13	**Ηλικία σε χρόνια**
B14	**Άλλες Υποχρεώσεις :**
B141	Τράπεζα
B142	Καταστήματα
B143	Καμία
B15	**Κατοικία :**
B151	Ενοικίαση
B152	Ιδιοκτησία
B153	Δωρεάν
B16	**Πλήθος υποχρεώσεων στην συγκεκριμένη τράπεζα**
B17	**Δουλειά :**
B171	Άνεργος ή Ανειδίκευτος και όχι κάτοικος της περιοχής
B172	Ανειδίκευτος και κάτοικος της περιοχής
B173	Ειδικευμένος υπάλληλος
B174	Διοικητικός ή Ιδιώτης ή Υψηλά Ειδικευμένος υπάλληλος
B18	**Αριθμός ανθρώπων που μπορούν να εγγυηθούν για πελάτη**
B19	**Τηλέφωνο :**
B191	Δεν υπάρχει
B192	Υπάρχει στο όνομα του πελάτη
B20	**Ξένος εργάτης :**
B201	Ναι
B202	Όχι

Το τρίτο σύνολο δεδομένων αφορά 666 πελάτες που συμπλήρωσαν αιτήσεις για χορήγηση πιστωτικής κάρτας, εκ των οποίων οι 460 χρησιμοποιήθηκαν στο δείγμα εκπαίδευσης και οι 206 στο δείγμα ελέγχου. Ωστόσο, τα κριτήρια στα οποία υποβλήθηκαν οι πελάτες είναι εμπιστευτικά και δεν γνωστοποιούνται, παρά μόνο

μετατράπηκαν σε σύμβολα χωρίς νόημα. Είναι, παρ' όλα αυτά, γνωστό ότι υπάρχουν 5 ποσοτικά και 9 ποιοτικά κριτήρια με τις δυνατές καταστάσεις τους. Τα δεδομένα αυτά προέκυψαν όπως και για το δεύτερο σύνολο δεδομένων από την βάση UCI Machine Learning Repository. Σημειώνεται ότι και σε αυτό το σύνολο δεδομένων χρησιμοποιήθηκε η κωδικοποίηση των ποιοτικών μεταβλητών που περιγράφτηκε για το δεύτερο σύνολο δεδομένων, καθώς και η κανονικοποίηση των ποσοτικών μεταβλητών.

4.3.2 Περιγραφή και Ανάλυση Αποτελεσμάτων

4.3.2.1 Η Αναμενόμενη Ακρίβεια ως Συνάρτηση Καταλληλότητας

Αρχικά, η συνάρτηση καταλληλότητας που χρησιμοποιήθηκε στον γενετικό αλγόριθμο είναι η Αναμενόμενη Ακρίβεια (sensitivity), όπως ορίστηκε στην παράγραφο 2.2.3 και αποτελεί ένα μέτρο της ακρίβειας του μοντέλου ταξινόμησης. Δηλαδή, η καταλληλότητα του κάθε χρωμοσώματος καθορίζεται από το πόσο καλή Αναμενόμενη Ακρίβεια επιτυγχάνει η Μηχανή Διανύσματος Υποστήριξης για κάθε κλάση, λαμβάνοντας τα δεδομένα από το εν λόγω χρωμόσωμα.. Έτσι ένα χρωμόσωμα έχει μεγάλη καταλληλότητα, εάν ο αριθμός των σωστών ταξινομήσεων όσον αφορά την κάθε κλάση σε σχέση με όλα τα δεδομένα που ανήκουν στην συγκεκριμένη κλάση είναι μεγάλος. Στην παρούσα εφαρμογή στην μία κλάση ανήκουν οι πελάτες με μικρό πιστωτικό κίνδυνο και στην άλλη ανήκουν οι πελάτες με μεγάλο πιστωτικό κίνδυνο τους οποίους η τράπεζα θα πρέπει να απορρίψει.

Το μέγεθος του πληθυσμού, αρχικά ορίστηκε στα 50 χρωμοσώματα, ενώ τα χρωμοσώματα όπως αναφέρθηκε είναι δυαδικά και το μήκος τους καθορίζεται από τον αριθμό των κριτηρίων του κάθε αρχείου συν τρία bit για την παράμετρο C και άλλα τρία στην περίπτωση μη γραμμικού πυρήνα για τον καθορισμό της παραμέτρου του. Ο αλγόριθμος καθορίστηκε να τερματίζει εφόσον δεν υπάρχει βελτίωση της λύσης σε 20 συνεχόμενες γενεές. Κατόπιν εξετάστηκε πληθυσμός με μέγεθος 20 και συνθήκη τερματισμού 50. Γενικά, όταν ο πληθυσμός είναι μικρός, είναι προτιμότερο ο αλγόριθμος να τερματίζει εφόσον δεν υπάρχει βελτίωση μετά από πολλές συνεχόμενες γενεές, προκειμένου να υπάρχει σύγκλιση.

Σύμφωνα με την μελέτη του De Jong [DeJong75] για τους γενετικούς αλγορίθμους, η καλή απόδοση τους χρειάζεται την επιλογή μιας υψηλής πιθανότητας διασταύρωσης και μίας χαμηλής πιθανότητας μετάλλαξης. Γι' αυτόν τον λόγο στην παρούσα εργασία επιλέχτηκαν να χρησιμοποιηθούν σαν τελεστές η απλή διασταύρωση με $p_c = 0.6$ και η δυαδική μετάλλαξη με $p_m = 0.05$. Και οι δύο τελεστές περιγράφονται αναλυτικά στην παράγραφο 3.2.3. Στον πίνακα 4.3 παρατίθενται όλες οι τιμές των παραμέτρων που χρησιμοποιήθηκαν στο πρώτο μέρος της ανάλυσης.

Πίνακας 4.3 : Τιμές των παραμέτρων για το πρώτο μέρος της ανάλυσης

Μέγεθος Πληθυσμού (N)	50
Αριθμός γενεών χωρίς μεταβολή στην λύση (G_t)	20
Πιθανότητα διασταύρωσης (p_c)	0.6
Πιθανότητα μεταλλαγής (p_m)	0.05
Συνάρτηση Καταλληλότητας	Αναμενόμενη Ακρίβεια

Η μεθοδολογία εξετάστηκε για διάφορες συναρτήσεις επιλογής του γενετικού αλγορίθμου που έχουν ήδη περιγραφτεί στην παράγραφο 3.2.2. Συγκεκριμένα ερευνήθηκε η μέθοδος ρουλέτας, η γεωμετρική επιλογή με παράμετρο q ίση με 0.05, 0.15 και 0.3, και η μέθοδος τουρνουά με παράμετρο k ίση με 5, 10 και 15. Στον πίνακα 4.4 εμφανίζονται οι μέσοι όροι για τις τιμές της Αναμενόμενης Ακρίβειας για κάθε συνάρτηση επιλογής, που προκύπτουν από όλα τα αρχεία δεδομένων και για όλες τις περιπτώσεις των πυρήνων (γραμμικός, πολυώνυμο, RBF), όταν χρησιμοποιούνται οι τιμές των παραμέτρων του πίνακα 4.3. Στην τρίτη στήλη εμφανίζεται ο μέσος όρος των τιμών Αναμενόμενης Ακρίβειας και για τις δύο κλάσεις. Οι τιμές της Αναμενόμενης Ακρίβειας είναι αυτές που προκύπτουν εφαρμόζοντας την λύση – χρωμόσωμα από τον γενετικό αλγόριθμο στο δείγμα ελέγχου.

Στον ίδιο πίνακα φαίνονται και οι μέσοι όροι των χαρακτηριστικών – κριτηρίων που λήφθηκαν υπόψη από τον γενετικό αλγόριθμο ανάλογα με την συνάρτηση επιλογής που χρησιμοποιήθηκε. Όσο λιγότερα είναι αυτά τα χαρακτηριστικά, τόσο απλούστερη γίνεται η ανάλυση, εφόσον το ζητούμενο είναι να εκτιμηθεί ο πιστωτικός κίνδυνος των εταιριών με όσο το δυνατό λιγότερα στοιχεία. Επειδή δεν μπορούν να παρουσιαστούν όλα τα αποτελέσματα, οι μέσοι όροι των τιμών

εξυπηρετούν στο ότι δείχνουν ενδεικτικά την αποτελεσματικότητα της κάθε συνάρτησης επιλογής και ως προς την Αναμενόμενη Ακρίβεια και ως προς τον αριθμό των χαρακτηριστικών που επιλέχτηκαν.

Πίνακας 4.4 : Μέσοι όροι των τιμών Αναμενόμενης Ακρίβειας και αριθμών των κριτηρίων για κάθε συνάρτηση επιλογής

Συναρτήσεις Επιλογής	**Αναμενόμενη Ακρίβεια**		**Μέσος Όρος**	**Μέσος όρος κριτηρίων**
	Κλάση 1	**Κλάση 2**		
Ρόδα Ρουλέτας	0.8541	0.7748	0.8144	18.88
Γεωμ Επιλογή (0.05)	0.8507	0.7821	0.8164	17.55
Γεωμ. Επιλογή (0.15)	0.8545	0.7942	0.8243	16
Γεωμ. Επιλογή (0.3)	0.8412	0.7825	0.8118	15.55
Τουρνουά (5)	0.8661	0.7681	0.8171	17.44
Τουρνουά (10)	0.8656	0.7686	0.8171	16.44
Τουρνουά (15)	0.8538	0.7888	0.8213	16.77

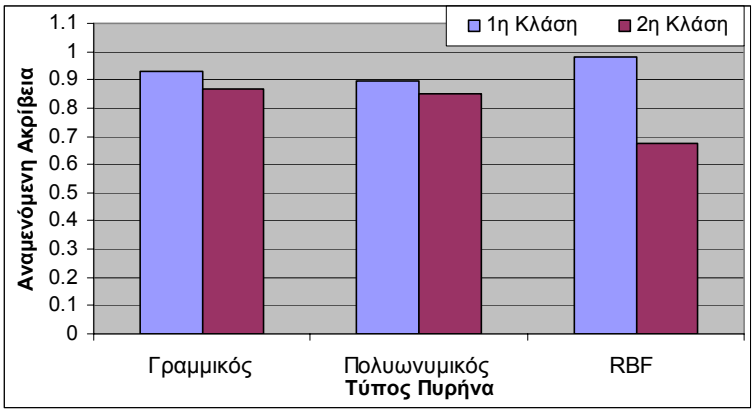
Όπως φαίνεται από τον πίνακα 4.4, όλες οι συναρτήσεις επιλογής αποδίδουν περίπου το ίδιο. Ωστόσο, η συνάρτηση επιλογής που υπερέχει έναντι των άλλων ως προς τις τιμές Αναμενόμενης Ακρίβειας για την κάθε κλάση και ως προς τον αριθμό των κριτηρίων που επιλέγει, είναι η Γεωμετρική Επιλογή με παράμετρο q ίση με 0.15, γι΄ αυτό και υιοθετείται στην συνέχεια της ανάλυσης ως συνάρτηση επιλογής.

Στη συνέχεια, στους πίνακες 4.5, 4.6 και 4.7 παρουσιάζονται τα αποτελέσματα για κάθε αρχείο δεδομένων. Συγκεκριμένα εμφανίζονται τα αποτελέσματα για κάθε τύπο πυρήνα, όταν χρησιμοποιείται η γεωμετρική επιλογή με παράμετρο q = 0.15 και οι τιμές των παραμέτρων που εμφανίζονται στον πίνακα 4.3. Στους ίδιους πίνακες παρατίθενται και τα αποτελέσματα για πληθυσμό N = 20 και συνθήκη τερματισμού G_t = 50. Επιπλέον, παρουσιάζονται και τα αποτελέσματα που προκύπτουν από την εφαρμογή των μεθόδων λογιστικής παλινδρόμησης, βηματικής λογιστικής παλινδρόμησης, διακριτικής ανάλυσης και βηματικής διακριτικής ανάλυσης. Στις βηματικές εκδοχές των μεθόδων γίνεται επιλογή ορισμένων χαρακτηριστικών, ενώ στις υπόλοιπες λαμβάνονται όλα τα χαρακτηριστικά στην επίλυση του προβλήματος.

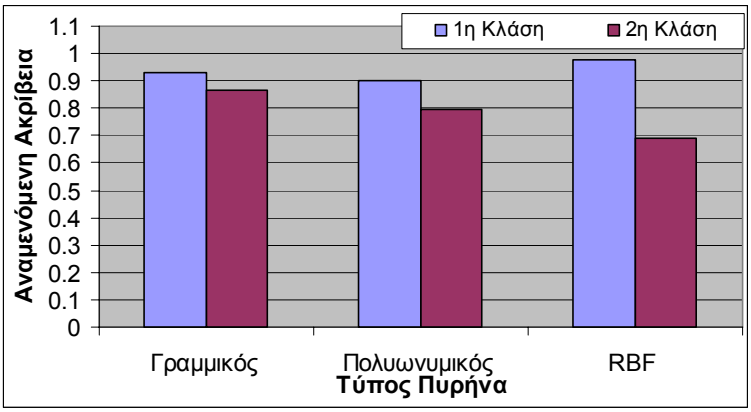
Πίνακας 4.5 : Αποτελέσματα για το πρώτο αρχείο δεδομένων

Τύπος Πυρήνα για $N = 50$, $G_t = 20$	**Αναμενόμενη Ακρίβεια**		**Μέσος Όρος**	**Αριθμός Κριτηρίων**
	Κλάση 1	**Κλάση 2**		
Γραμμικός	0.93003	0.86765	0.89884	5 / 8
Πολυωνυμικός	0.89796	0.85294	0.87545	6 / 8
RBF	0.98251	0.67647	0.82949	5 / 8
Τύπος Πυρήνα για $N = 20$, $G_t = 50$				
Γραμμικός	0.93003	0.86765	0.89884	5 / 8
Πολυωνυμικός	0.90087	0.79412	0.84749	6 / 8
RBF	0.97959	0.69118	0.83538	5 / 8
Στατιστικές Μέθοδοι				
Λογ.Παλινδρόμηση	0.968	0.691	0.8295	8 / 8
Βηματ.Λογ.Παλινδρόμηση	0.971	0.706	0.8385	4 / 8
Διακριτική Ανάλυση	0.956	0.647	0.8015	8 / 8
Βηματ.Διακρ.Ανάλυση	0.959	0.632	0.7955	6 / 8

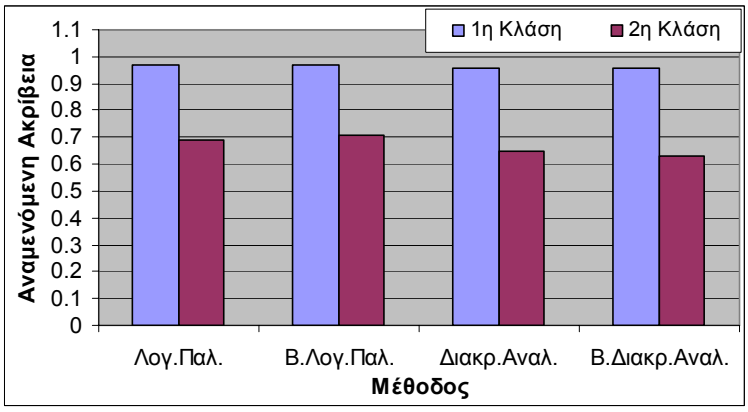
Όσον αφορά το πρώτο σύνολο δεδομένων, τα σχήματα 4.1, 4.2 παρουσιάζουν σχηματικά τις τιμές της Αναμενόμενης Ακρίβειας για κάθε τύπο πυρήνα και για κάθε κλάση πελατών, όταν N=20, G_t=50 και όταν N=50, G_t=20, αντίστοιχα. Στο σχήμα 4.2 εμφανίζονται οι τιμές της Αναμενόμενης Ακρίβειας για τις διάφορες στατιστικές μεθόδους.



Σχήμα 4.1 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση και πυρήνα, όταν N=50, G_t =20



Σχήμα 4.2 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση και πυρήνα, όταν N=20, G_t =50



Σχήμα 4.3 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση σε κάθε μέθοδο

Από τα αποτελέσματα του πίνακα 4.5 και τα γραφήματα 4.1, 4.2 και 4.3 προκύπτουν διάφορες χρήσιμες παρατηρήσεις για τις διάφορες μεθόδους όσον αφορά το πρώτο σύνολο δεδομένων. Παρατηρούμε ότι όταν χρησιμοποιείται ο γραμμικός πυρήνας, τα αποτελέσματα δεν διαφοροποιούνται με την αλλαγή του πληθυσμού και την συνθήκη τερματισμού του αλγορίθμου, κάτι όμως που δεν παρατηρείται στην περίπτωση του πολυωνύμου και του RBF πυρήνα.

Όσον αφορά τις στατιστικές μεθόδους, αυτές παρατηρούμε ότι επιτυγχάνουν μεγάλες τιμές Αναμενόμενης Ακρίβειας για την πρώτη κλάση πελατών, λίγο μεγαλύτερες από αυτές που επιτυγχάνει ο γενετικός αλγόριθμος, ωστόσο οι τιμές Αναμενόμενης Ακρίβειας για την δεύτερη κλάση πελατών είναι πολύ χαμηλότερες από αυτές του γενετικού αλγορίθμου, γι' αυτό και ο μέσος όρος και για τις δύο κλάσεις κυμαίνεται σε αρκετά μικρότερες τιμές από αυτές που προκύπτουν από την

προτεινόμενη μεθοδολογία. Το ζητούμενο είναι να υπάρχει μια ισορροπία των τιμών Αναμενόμενης Ακρίβειας για τις δύο κλάσεις πελατών, δηλαδή να βρίσκονται περίπου στα ίδια αρκετά υψηλά επίπεδα τιμών. Παρατηρούμε, επιπλέον, ότι γενικά η λογιστική παλινδρόμηση παρουσιάζει λίγο καλύτερα αποτελέσματα από την διακριτική ανάλυση, ενώ οι βηματικές εκδοχές των μεθόδων δεν διαφοροποιούνται σημαντικά από τις κανονικές.

Μια ισορροπία στις τιμές Αναμενόμενης Ακρίβειας μεταξύ των δύο κλάσεων παρατηρούμε ότι επιτυγχάνει ο γραμμικός πυρήνας με 93% Αναμενόμενη Ακρίβεια για την πρώτη κλάση και 86.7% για την δεύτερη κλάση. Τα χαρακτηριστικά – κριτήρια που λήφθηκαν υπόψη στην ανάλυση ως τα πλέον σημαντικότερα είναι τα Α1 έως Α4 και το Α6, δηλαδή όλα τα κριτήρια αποδοτικότητας και ρευστότητας και από τα κριτήρια φερεγγυότητας μόνο ο δείκτης Συνολικές Υποχρεώσεις / Σύνολο Ενεργητικού. Η παράμετρος κανονικοποίησης C που επιλέχτηκε από το χρωμόσωμα – λύση του γενετικού αλγορίθμου ορίστηκε στην τιμή 10.000. Συγκριτικά με τα αποτελέσματα όλων των μεθόδων, ο γενετικός αλγόριθμος με τον γραμμικό πυρήνα προσφέρει τα καλύτερα αποτελέσματα και όσον αφορά τις τιμές Αναμενόμενης Ακρίβειας και όσο αφορά τον αριθμό των χαρακτηριστικών που ελήφθησαν στην ανάλυση.

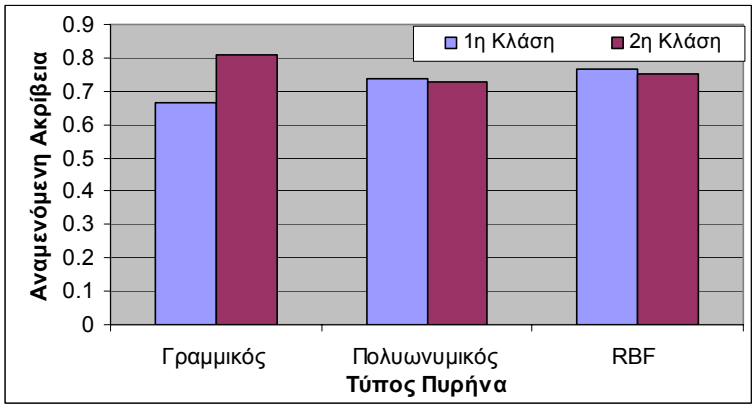
Σημειώνεται ότι τα χαρακτηριστικά που επιλέγονται από τους πυρήνες δεν αλλάζουν με την επιλογή του ζευγαριού N, G_t. Επιπλέον, τα χαρακτηριστικά που επιλέχτηκαν από τον RBF πυρήνα είναι τα ίδια με αυτά που επιλέχτηκαν από τον γραμμικό πυρήνα. Ωστόσο ο πυρήνας πολυωνύμου επέλεξε τα ίδια κριτήρια αποδοτικότητας και ρευστότητας, αλλά διαφορετικά κριτήρια φερεγγυότητας, συγκεκριμένα επέλεξε τα κριτήρια Α1-Α3, Α6-Α8.

Από τα αποτελέσματα του πίνακα 4.6 και τα γραφήματα 4.4, 4.5 και 4.6 προκύπτουν χρήσιμα συμπεράσματα για τις διάφορες μεθόδους όταν εφαρμόζονται στο δεύτερο σύνολο δεδομένων. Συγκεκριμένα, παρατηρείται ότι οι τιμές Αναμενόμενης Ακρίβειας που επιτυγχάνονται για κάθε τύπο πυρήνα δεν επηρεάζονται σημαντικά από το μέγεθος του πληθυσμού και την συνθήκη τερματισμού του αλγορίθμου. Ωστόσο, τα αποτελέσματα που προκύπτουν από τον

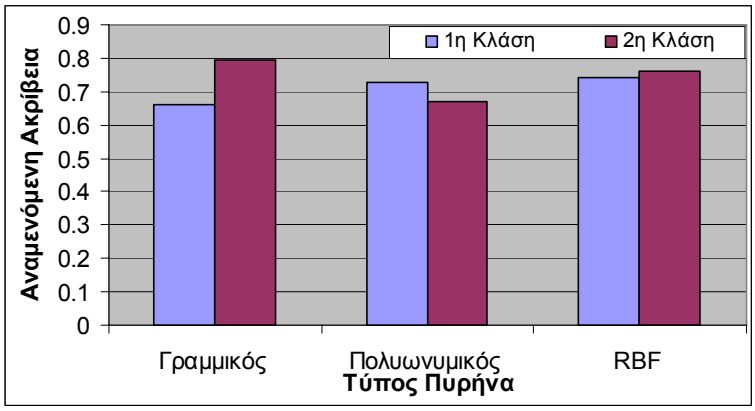
μεγαλύτερο πληθυσμό και την μικρότερη συνθήκη τερματισμού είναι ελαφρώς καλύτερα.

Πίνακας 4.6 : Αποτελέσματα για το δεύτερο αρχείο δεδομένων

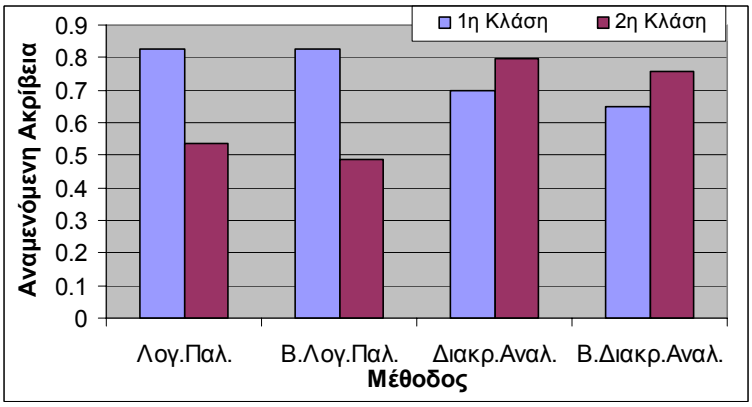
Τύπος Πυρήνα για $N = 50$, $G_t = 20$	**Αναμενόμενη Ακρίβεια**		**Μέσος Όρος**	**Αριθμός Κριτηρίων**
	Κλάση 1	**Κλάση2**		
Γραμμικός	0.66509	0.80682	0.73595	27 / 48
Πολυωνυμικός	0.73113	0.68182	0.70647	23 / 48
RBF	0.76415	0.75	0.75707	27 / 48
Τύπος Πυρήνα για $N = 20$, $G_t = 50$				
Γραμμικός	0.66038	0.79545	0.72791	24 / 48
Πολυωνυμικός	0.72642	0.67045	0.69843	23 / 48
RBF	0.74057	0.76136	0.75096	28 / 48
Στατιστική Μέθοδος				
Λογ.Παλινδρόμηση	0.825	0.534	0.6795	48 / 48
Βηματ.Λογ.Παλινδρόμηση	0.825	0.489	0.6570	15 / 48
Διακριτική Ανάλυση	0.698	0.795	0.7465	48 / 48
Βηματ.Διακρ.Ανάλυση	0.651	0.759	0.7050	16 / 48



Σχήμα 4.4 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση και πυρήνα, όταν N=50, G_t =20



Σχήμα 4.5 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση και πυρήνα, όταν N=20, G_t =50



Σχήμα 4.6 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση σε κάθε μέθοδο

Όσον αφορά τις στατιστικές μεθόδους, η λογιστική παλινδρόμηση και η βηματική λογιστική παλινδρόμηση δίνουν μεγάλες τιμές Αναμενόμενης Ακρίβειας στην πρώτη κλάση, αλλά πολύ μικρές στην δεύτερη κλάση. Η διακριτική ανάλυση και η βηματική διακριτική ανάλυση δίνουν σχετικά χαμηλές τιμές και στις δύο κλάσεις πελατών, με λίγο καλύτερες τιμές στην δεύτερη κλάση. Για άλλη μια φορά, οι τιμές Αναμενόμενης Ακρίβειας δεν επηρεάζονται σημαντικά από την χρήση της βηματικής εκδοχής της μεθόδου έναντι της κανονικής, αν και οι βηματικές εκδοχές φαίνεται να υστερούν λίγο έναντι των απλών μεθόδων.

Τελικά, παρατηρούμε ότι τις καλύτερες τιμές Αναμενόμενης Ακρίβειας και για τις δύο κλάσεις από όλες τις μεθόδους τις δίνει ο πυρήνας RBF για N = 50 και G_t = 20. Συγκεκριμένα, επιτυγχάνει τιμές Αναμενόμενης Ακρίβειας 76.4% για την πρώτη κλάση και 75% για την δεύτερη κλάση πελατών. Τα χαρακτηριστικά –

κριτήρια που δίνονται από το χρωμόσωμα – λύση του γενετικού αλγορίθμου είναι 27 από τα 48 συνολικά που εξετάζονται από τον αλγόριθμο. Συγκεκριμένα, θεωρούνται τα κριτήρια B11 – B13, B2, B31 – B33, B41, B43, B45 – B49, B61, B62, B64, B91 – B93, B121, B122, B142, B151, B152, B172, B191, B201. Η τιμή της παραμέτρου *C* που προκύπτει από το ίδιο χρωμόσωμα - λύση είναι 10, ενώ η παράμετρος *σ* του RBF πυρήνα είναι 2^{-6}.

Παρατηρούμε ότι δεν λαμβάνονται καθόλου στην ανάλυση κριτήρια, όπως το χρονικό διάστημα που κάποιος είναι εργαζόμενος, η αξία δόσης ως ποσοστό εισοδήματος, εάν υπάρχουν άλλοι χρεώστες ή εγγυητές, το χρονικό διάστημα μόνιμης κατοικίας στην περιοχή, η ηλικία, το πλήθος των υποχρεώσεων στην συγκεκριμένη τράπεζα και ο αριθμός των ατόμων που μπορούν να εγγυηθούν για τον πελάτη. Αντίθετα, η ανάλυση επικεντρώνεται κυρίως στον λογαριασμό επιταγών και καταθέσεων, στην πιστωτική ιστορία, στην διάρκεια σε μήνες, στο φύλο και οικογενειακή κατάσταση, στον σκοπό, στο αν υπάρχουν ακίνητα ή ασφάλεια ζωής, εάν υπάρχουν υποχρεώσεις σε καταστήματα, εάν υπάρχει δικιά του κατοικία ή νοικιάζει, εάν είναι ανειδίκευτος, εάν δεν έχει τηλέφωνο και αν είναι ξένος εργάτης.

Σημειώνεται ότι οι άλλοι δύο πυρήνες, δηλαδή ο γραμμικός και ο πυρήνας πολυωνύμου επιλέγουν 15 και 12 κοινά χαρακτηριστικά αντίστοιχα με τον πυρήνα RBF. Τα χαρακτηριστικά που είναι κοινά και στους τρεις πυρήνες είναι 9 και είναι τα B11, B12, B2, B46, B47, B61, B121, B191 και B201.

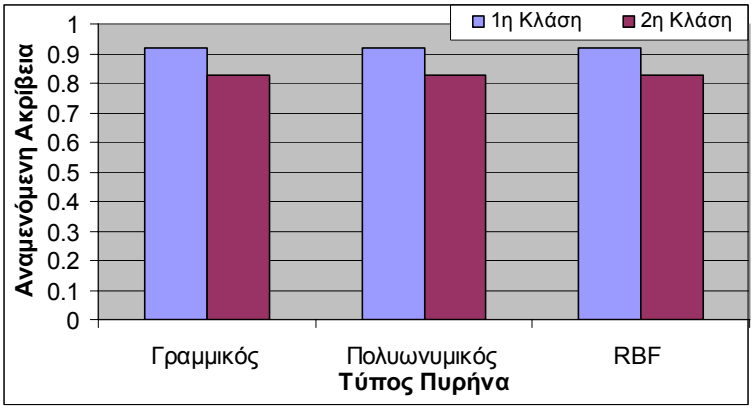
Τέλος από τον πίνακα 4.7 και τα σχήματα 4.7, 4.8 και 4.9 εξάγονται ενδιαφέρουσες πληροφορίες για τις διάφορες μεθόδους όταν εφαρμόζονται στο τρίτο σύνολο δεδομένων. Για αυτό το αρχείο δεδομένων, παρατηρείται ότι όλες οι μέθοδοι αποδίδουν καλά και περίπου το ίδιο. Όσον αφορά τις στατιστικές μεθόδους η διακριτική ανάλυση αποδίδει καλύτερα από την λογιστική παλινδρόμηση. Η διακριτική ανάλυση που θεωρεί όλα τα χαρακτηριστικά δίνει τις μεγαλύτερες τιμές Αναμενόμενης Ακρίβειας, 93% για την πρώτη κλάση και 82.5% για την δεύτερη. Τις ίδιες τιμές δίνουν και οι Μηχανές Διανύσματος Υποστήριξης με γραμμικό πυρήνα, θεωρώντας 21 από τα 40 κριτήρια και παράμετρο *C* ίση με 1, όταν N=20 και G_t=50. Εφόσον, με λιγότερα κριτήρια επιτυγχάνονται οι ίδιες τιμές Αναμενόμενης Ακρίβειας

για τις δύο κλάσεις πελατών, είναι προτιμότερη η προτεινόμενη μεθοδολογία της παρούσας εργασίας.

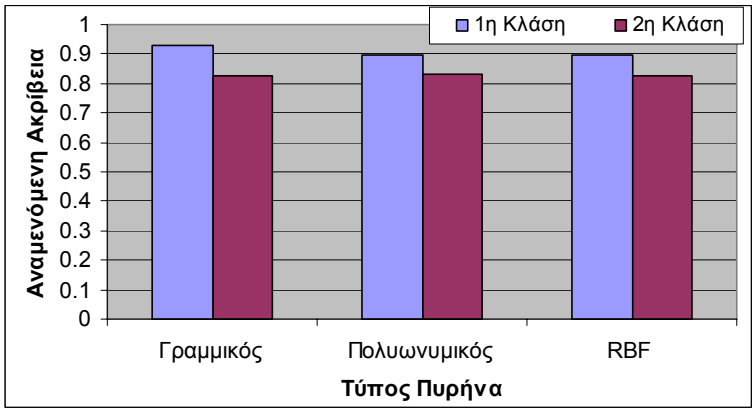
Για άλλη μια φορά, ο μεγαλύτερος πληθυσμός δίνει ελαφρώς καλύτερα αποτελέσματα με εξαίρεση τον γραμμικό πυρήνα. Τα χαρακτηριστικά που λήφθηκαν στην ανάλυση από τους άλλους πυρήνες είναι περίπου τα ίδια. Συγκεκριμένα, τα κριτήρια που επιλέχτηκαν και από τους τρεις πυρήνες είναι 9.

Πίνακας 4.7 : Αποτελέσματα για το τρίτο αρχείο δεδομένων

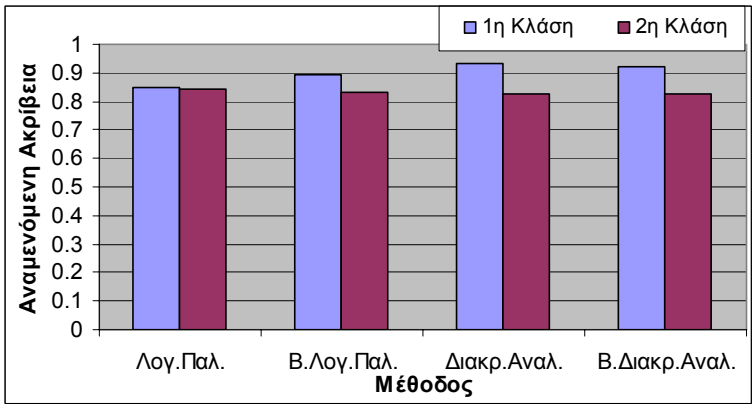
Τύπος Πυρήνα για $N = 50$, $G_t = 20$	**Αναμενόμενη Ακρίβεια**		**Μέσος Όρος**	**Αριθμός Κριτηρίων**
	Κλάση 1	**Κλάση 2**		
Γραμμικός	0.9186	0.825	0.8718	17 / 40
Πολυωνυμικός	0.9186	0.825	0.8718	22 / 40
RBF	0.9186	0.825	0.8718	22 / 40
Τύπος Πυρήνα για $N = 20$, $G_t = 50$				
Γραμμικός	0.93023	0.825	0.8776	21 / 40
Πολυωνυμικός	0.89535	0.8333	0.8643	17 / 40
RBF	0.89535	0.825	0.8601	23 / 40
Στατιστική Μέθοδος				
Λογ.Παλινδρόμηση	0.849	0.842	0.845	40 / 40
Βηματ.Λογ.Παλινδρόμηση	0.895	0.833	0.864	7 / 40
Διακριτική Ανάλυση	0.930	0.825	0.877	40 / 40
Βηματ.Διακρ.Ανάλυση	0.919	0.825	0.872	8 / 40



Σχήμα 4.7 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση και πυρήνα, όταν N=50, G_t =20



Σχήμα 4.8 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση και πυρήνα, όταν N=20, G_t =50



Σχήμα 4.9 : Τιμές Αναμενόμενης Ακρίβειας για κάθε κλάση σε κάθε μέθοδο

4.3.2.2 Ο Δείκτης Ακρίβειας ως Συνάρτηση Καταλληλότητας

Στο δεύτερο μέρος της ανάλυσης ως συνάρτηση καταλληλότητας επιλέχτηκε ο Δείκτης Ακρίβειας, όπως περιγράφτηκε στην παράγραφο 2.2.3. Οι υπόλοιπες παράμετροι του γενετικού αλγορίθμου ορίζονται όπως και στο πρώτο κομμάτι της ανάλυσης και παρουσιάζονται στον πίνακα 4.8.

Κατά αντιστοιχία με το πρώτο μέρος της ανάλυσης, η μεθοδολογία εξετάστηκε για διάφορες συναρτήσεις επιλογής του γενετικού αλγορίθμου όπως τη μέθοδο ρουλέτας, τη γεωμετρική επιλογή με παράμετρο q ίση με 0.05, 0.15 και 0.3, και τη μέθοδο τουρνουά με παράμετρο k ίση με 5, 10 και 15. Στον πίνακα 4.9 εμφανίζονται οι μέσοι όροι των Δεικτών Ακρίβειας και των χαρακτηριστικών – κριτηρίων για κάθε συνάρτηση επιλογής, που προκύπτουν από όλα τα αρχεία δεδομένων και για όλες τις

περιπτώσεις των πυρήνων (γραμμικός, πολυώνυμο, RBF), όταν χρησιμοποιούνται οι τιμές των παραμέτρων του πίνακα 4.8. Οι τιμές των Δεικτών Ακρίβειας είναι αυτές που προκύπτουν εφαρμόζοντας την λύση – χρωμόσωμα από τον γενετικό αλγόριθμο στο δείγμα ελέγχου.

Πίνακας 4.8 : Τιμές των παραμέτρων για το δεύτερο μέρος της ανάλυσης

Μέγεθος Πληθυσμού (N)	50
Αριθμός γενεών χωρίς μεταβολή στην λύση (G_t)	20
Πιθανότητα διασταύρωσης (p_c)	0.6
Πιθανότητα μεταλλαγής (p_m)	0.05
Συνάρτηση Καταλληλότητας	Δείκτης Ακρίβειας

Πίνακας 4.9 : Μέσοι όροι των Δεικτών Ακρίβειας και αριθμών των κριτηρίων για κάθε συνάρτηση επιλογής

Συναρτήσεις Επιλογής	**Μέσος όρος των Δεικτών Ακρίβειας**	**Μέσος όρος των κριτηρίων**
Ρόδα Ρουλέτας	0.7649	16
Γεωμ Επιλογή (0.05)	0.7672	16
Γεωμ. Επιλογή (0.15)	0.7730	15.77
Γεωμ. Επιλογή (0.3)	0.7638	15.88
Τουρνουά (5)	0.7720	15.55
Τουρνουά (10)	0.7671	14.44
Τουρνουά (15)	0.7622	15.11

Όπως φαίνεται από τον πίνακα 4.9, τα αποτελέσματα που προκύπτουν για κάθε συνάρτηση επιλογής δεν διαφοροποιούνται σημαντικά, ωστόσο και σε αυτή την περίπτωση η Γεωμετρική Επιλογή με παράμετρο q = 0.15 φαίνεται να αποδίδει καλύτερα από τις άλλες συναρτήσεις επιλογής, γι' αυτό και υιοθετείται και σε αυτό το μέρος της ανάλυσης.

Στη συνέχεια, για κάθε σύνολο δεδομένων παρατίθενται στους πίνακες 4.10, 4.11 και 4.12 οι Δείκτες Ακρίβειας και τα επιλεγμένα χαρακτηριστικά για κάθε τύπο

πυρήνα, όταν N=50, G_t=20 και όταν N=20, G_t=50. Από τα αποτελέσματα αυτά προκύπτει ότι γενικά η μεταβολή στο μέγεθος του πληθυσμού και της συνθήκης τερματισμού δεν διαφοροποιούν σημαντικά τα αποτελέσματα και επιπλέον το ζευγάρι τιμών N=50, G_t=20 δίνει λίγο καλύτερες τιμές. Τα παραπάνω συμπεράσματα ισχύουν και για τα τρία αρχεία δεδομένων, γι' αυτό και στο υπόλοιπο μέρος της ανάλυσης θα εξετάζεται μόνο η περίπτωση N=50, G_t=20.

Όσον αφορά το πρώτο αρχείο δεδομένων, προκειμένου να γίνει σύγκριση της προτεινόμενης μεθοδολογίας με τις στατιστικές μεθόδους, στον πίνακα 4.13 εμφανίζονται οι τιμές AUC που προκύπτουν από τις καμπύλες ROC στα σχήματα 4.10 και 4.11. Στο σχήμα 4.10 παρουσιάζονται οι καμπύλες ROC για τον κάθε τύπο πυρήνα όταν γίνεται επιλογή χαρακτηριστικών και όταν δεν γίνεται επιλογή χαρακτηριστικών, δηλαδή όταν λαμβάνονται όλα τα χαρακτηριστικά στην ανάλυση. Στο σχήμα 4.11 παρουσιάζονται οι καμπύλες ROC για την κάθε στατιστική μέθοδο (Λογιστική Παλινδρόμηση, Βηματική Λογιστική Παλινδρόμηση, Διακριτική Ανάλυση και Βηματική Διακριτική Ανάλυση). Παρ΄ ότι οι τιμές AUC για κάθε τύπο πυρήνα, όταν γίνεται επιλογή χαρακτηριστικών μπορούν να προκύψουν από τους Λόγους Ακρίβειας (AR) του πίνακα 4.10 από την σχέση AR = 2AUC-1, παρατίθενται και αυτές στον πίνακα 4.13 για να μπορεί να γίνει σύγκριση των διαφόρων μεθόδων

Πίνακας 4.10 : Αποτελέσματα για το 1ο Αρχείο Δεδομένων

Τύπος Πυρήνα για N = 50, G_t = 20	**Δείκτης Ακρίβειας**	**Αριθμός Κριτηρίων**
Γραμμικός	0.91803	6 / 8
Πολυωνυμικός	0.87747	3 / 8
RBF	0.90525	6 / 8
Τύπος Πυρήνα για N = 20, G_t = 50		
Γραμμικός	0.91802	6 / 8
Πολυωνυμικός	0.87738	3 / 8
RBF	0.90233	7 / 8

Πίνακας 4.11 : Αποτελέσματα για το 2ο Αρχείο Δεδομένων

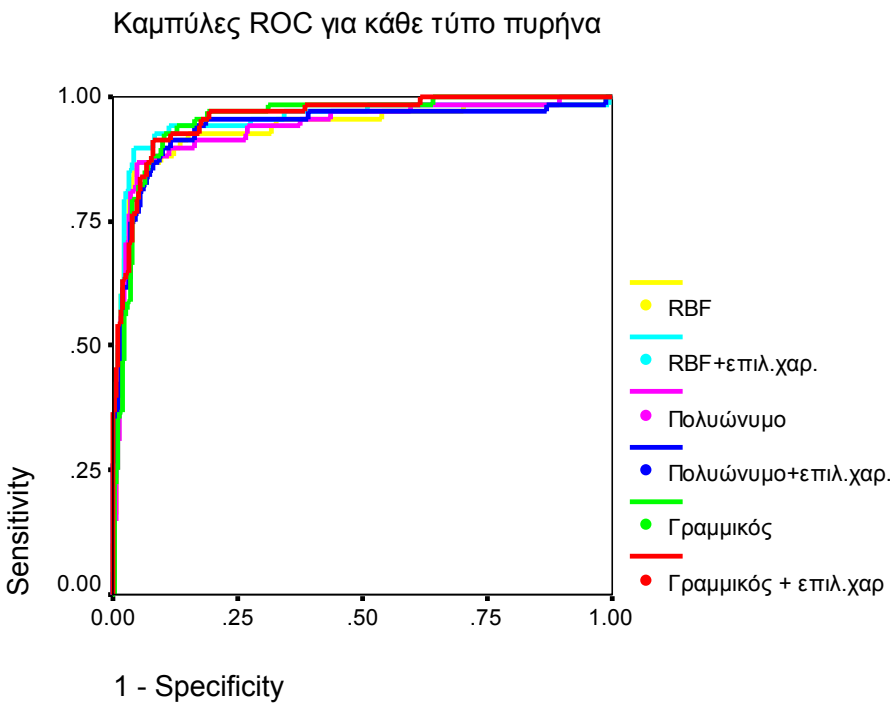
Τύπος Πυρήνα για $N = 50$, $G_t = 20$	**Δείκτης Ακρίβειας**	**Αριθμός Κριτηρίων**
Γραμμικός	0.58212	24 / 48
Πολυωνυμικός	0.52412	24 / 48
RBF	0.55843	27 / 48
Τύπος Πυρήνα για $N = 20$, $G_t = 50$		
Γραμμικός	0.55467	19 / 48
Πολυωνυμικός	0.47792	25 / 48
RBF	0.55403	24 / 48

Πίνακας 4.12 : Αποτελέσματα για το 3ο Αρχείο Δεδομένων

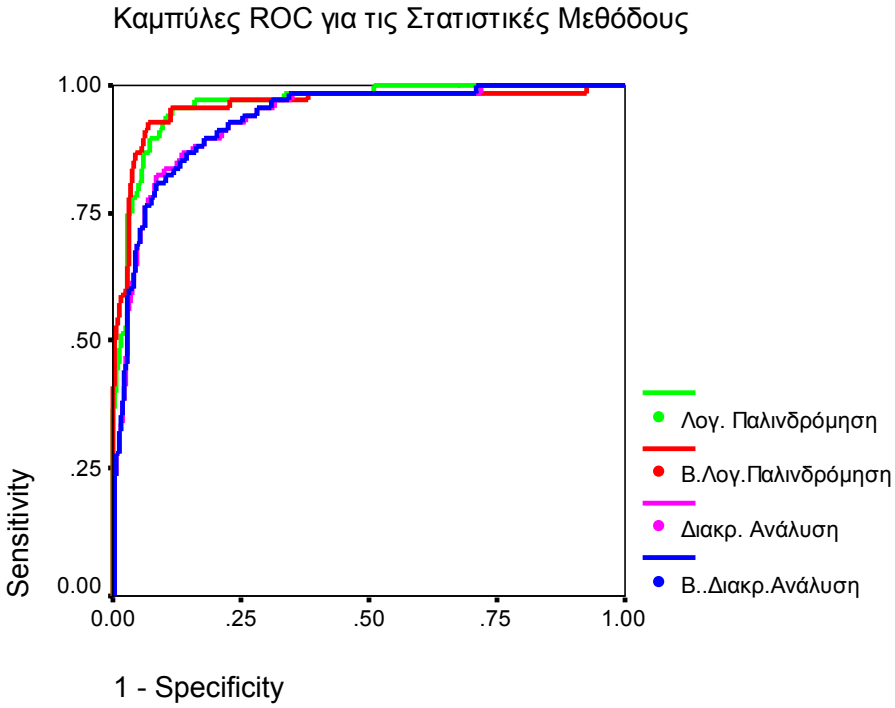
Τύπος Πυρήνα για $N = 50$, $G_t = 20$	**Δείκτης Ακρίβειας**	**Αριθμός Κριτηρίων**
Γραμμικός	0.85988	15 / 40
Πολυωνυμικός	0.87665	19 / 40
RBF	0.85514	18 / 40
Τύπος Πυρήνα για $N = 20$, $G_t = 50$		
Γραμμικός	0.85640	16 / 48
Πολυωνυμικός	0.85514	15 / 48
RBF	0.84089	21 / 40

Πίνακας 4.13 : Τιμές AUC και Αριθμός Κριτηρίων για κάθε μέθοδο

Τύπος Πυρήνα	**AUC**	**Αριθμός Κριτηρίων**
Γραμμικός	0.955	8 / 8
Γραμμικός + επιλ.χαρακτ.	0.959	6 / 8
Πολυωνυμικός	0.938	8/ 8
Πολυωνυμικός + επιλ.χαρ.	0.939	3 / 8
RBF	0.938	8 / 8
RBF + επιλ.χαρακτ.	0.953	6 / 8
Στατιστική Μέθοδος		
Λογ.Παλινδρόμηση	0.963	8 / 8
Βηματ.Λογ.Παλινδρόμηση	0.960	4 / 8
Διακριτική Ανάλυση	0.934	8 / 8
Βηματ.Διακρ.Ανάλυση	0.934	6 / 8



Σχήμα 4.10 : Καμπύλες ROC για κάθε τύπο πυρήνα



Σχήμα 4.11 : Καμπύλες ROC για τις Στατιστικές Μεθόδους

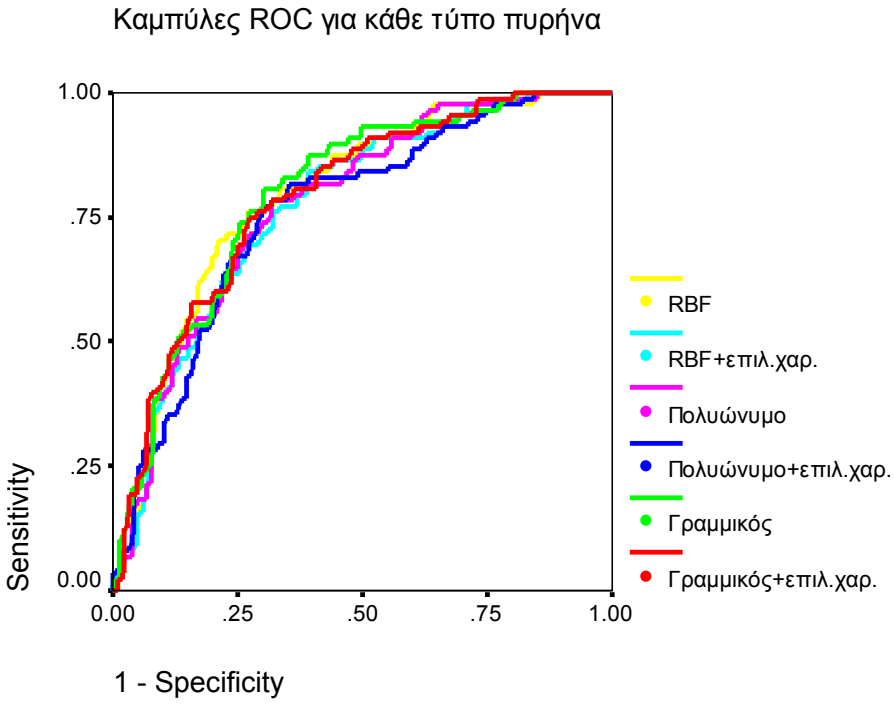
Από τον πίνακα 4.13 και τις αντίστοιχες καμπύλες ROC (σχήματα 4.10 – 4.12) παρατηρούμε ότι όσο αφορά την προτεινόμενη μεθοδολογία τα καλύτερα αποτελέσματα παρουσιάζει ο γραμμικός πυρήνας όταν γίνεται επιλογή χαρακτηριστικών. Συγκεκριμένα, δίνει τιμή AUC ίση με 95,9% και χρησιμοποιούνται στην ανάλυση 6 από τα 8 κριτήρια που είναι τα A1 έως A4, A6 και A8, δηλαδή όλα τα κριτήρια αποδοτικότητας και ρευστότητας και από τα κριτήρια φερεγγυότητας ο δείκτης Συνολικές Υποχρεώσεις / Σύνολο Ενεργητικού και ο δείκτης Συνολικές Υποχρεώσεις / Κεφάλαιο Κίνησης. Η τιμή της παραμέτρου C που προέκυψε από την λύση – χρωμόσωμα είναι 10000.

Ο γραμμικός πυρήνας υπερέχει της Διακριτικής Ανάλυσης, όχι όμως και της Λογιστικής Παλινδρόμησης που παρουσιάζει λίγο μεγαλύτερη τιμή AUC ίση με 96.3% όταν θεωρεί όλα τα χαρακτηριστικά στην ανάλυση. Οι υπόλοιποι πυρήνες παρουσιάζουν επίσης μεγάλες τιμές AUC και χρησιμοποιούν στην ανάλυση τρία κοινά κριτήρια, τα A1, A3 και A4. Σημειώνεται επίσης ότι η προτεινόμενη μεθοδολογία παρουσιάζει λίγο καλύτερα αποτελέσματα όταν γίνεται η επιλογή κριτηρίων από όταν δεν γίνεται.

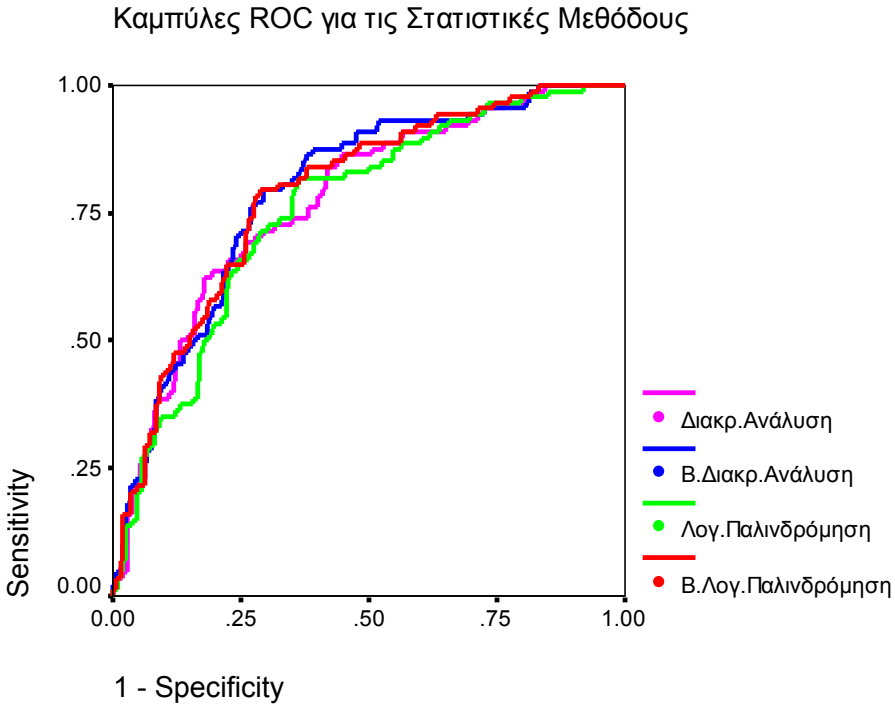
Όσον αφορά το δεύτερο αρχείο δεδομένων στον πίνακα 4.14 εμφανίζονται οι τιμές AUC που προκύπτουν από τις καμπύλες ROC στα σχήματα 4.12 και 4.13.

Πίνακας 4.14 : Τιμές AUC και Αριθμός Κριτηρίων για κάθε μέθοδο

Τύπος Πυρήνα	**AUC**	**Αριθμός Κριτηρίων**
Γραμμικός	0.797	48 / 48
Γραμμικός + επιλ.χαρακτ.	0.791	24 / 48
Πολυωνυμικός	0.777	48/ 48
Πολυωνυμικός + επιλ.χαρ.	0.762	24 / 48
RBF	0.794	48 / 48
RBF + επιλ.χαρακτ.	0.779	27 / 48
Στατιστική Μέθοδος		
Λογ.Παλινδρόμηση	0.757	48 / 48
Βηματ.Λογ.Παλινδρόμηση	0.788	15 / 48
Διακριτική Ανάλυση	0.773	48 / 48
Βηματ.Διακρ.Ανάλυση	0.792	16 / 48



Σχήμα 4.12 : Καμπύλες ROC για κάθε τύπο πυρήνα



Σχήμα 4.13 : Καμπύλες ROC για τις Στατιστικές Μεθόδους

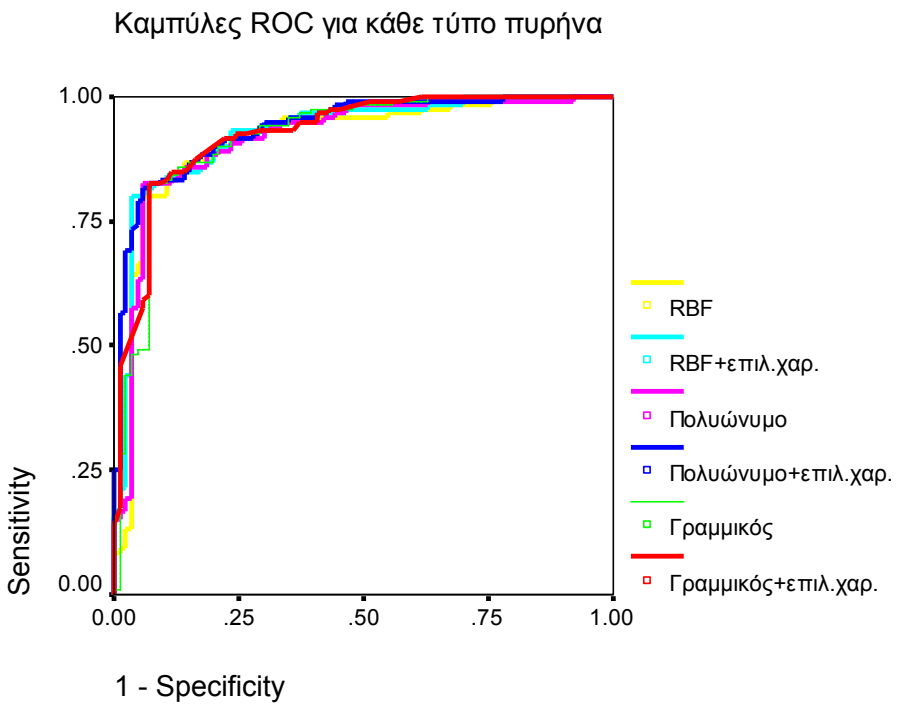
Σχετικά με το δεύτερο αρχείο δεδομένων, παρατηρείται από τον πίνακα 4.14 και τα σχήματα 4.12 και 4.13 ότι την μεγαλύτερη τιμή AUC από όλες τις μεθόδους δίνει ο γραμμικός πυρήνας που θεωρεί όλα τα χαρακτηριστικά στην ανάλυση. Συγκεκριμένα η τιμή AUC είναι 79,7% και η τιμή της παραμέτρου C προέκυψε από την λύση – χρωμόσωμα ίση με 10000. Παρόμοιες τιμές AUC δίνει και ο RBF πυρήνας που θεωρεί όλα τα χαρακτηριστικά με τιμή AUC ίση με 79,4%, καθώς και η βηματική Διακριτική Ανάλυση με τιμή AUC ίση με 79.2%.

Για το τρίτο αρχείο δεδομένων τα αποτελέσματα φαίνονται στον πίνακα 4.15 και τα σχήματα 4.14 και 4.15.

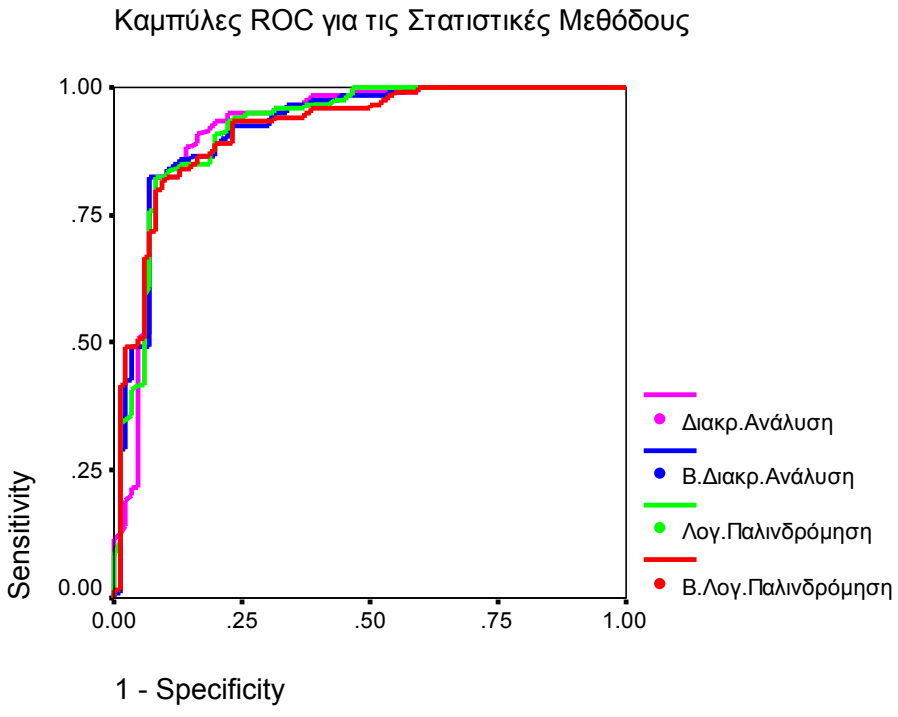
Πίνακας 4.15 : Τιμές AUC και Αριθμός Κριτηρίων για κάθε μέθοδο

Τύπος Πυρήνα	**AUC**	**Αριθμός Κριτηρίων**
Γραμμικός	0.919	40/ 40
Γραμμικός + επιλ.χαρακτ.	0.929	15 / 40
Πολυωνυμικός	0.917	40/ 40
Πολυωνυμικός + επιλ.χαρ.	0.938	19 / 40
RBF	0.913	40 / 40
RBF + επιλ.χαρακτ.	0.928	18 / 40
Στατιστική Μέθοδος		
Λογ.Παλινδρόμηση	0.923	40 / 40
Βηματ.Λογ.Παλινδρόμηση	0.919	7 / 40
Διακριτική Ανάλυση	0.922	40 / 40
Βηματ.Διακρ.Ανάλυση	0.921	8 / 40

Για αυτό το αρχείο δεδομένων, η προτεινόμενη μεθοδολογία δίνει γενικότερα καλύτερα αποτελέσματα από τις Στατιστικές Μεθόδους. Την μεγαλύτερη τιμή AUC δίνει ο πυρήνας πολυωνύμου με βαθμό πολυωνύμου ίσο με 8 και παράμετρο C ίση με 0.001 και θεωρώντας τα 19 από τα 40 κριτήρια στην ανάλυση. Ακολουθεί ο γραμμικός πυρήνας με AUC ίσο με 92.9%, θεωρώντας 15 κριτήρια στην ανάλυση και παράμετρο C ίση με 10000. Τέλος, ο πυρήνας RBF με παράμετρο σ ίση με 0.2 και παράμετρο C ίση με 1 και θεωρώντας 18 κριτήρια δίνει μια τιμή AUC ίση με 92.8%. Σημειώνεται επίσης ότι και οι τρεις πυρήνες επέλεξαν 9 κοινά χαρακτηριστικά.



Σχήμα 4.14 : Καμπύλες ROC για κάθε τύπο πυρήνα



Σχήμα 4.15 : Καμπύλες ROC για τις Στατιστικές Μεθόδους

Κεφάλαιο 5

'Συμπεράσματα'

5.1 Σύνοψη της Εργασίας

Η παρούσα εργασία ασχολήθηκε με ένα από τα σημαντικότερα μεθοδολογικά εργαλεία στη θεωρία της στατιστικής θεωρίας μάθησης σε προβλήματα ταξινόμησης, τις Μηχανές Διανύσματος Υποστήριξης (Support Vector Machines, SVM). Στόχος της εργασίας ήταν η ανάπτυξη μιας κατάλληλης και ολοκληρωμένης μεθοδολογίας για το πρόβλημα της ταξινόμησης.

Προκειμένου να χρησιμοποιηθούν επιτυχημένα οι δυνατότητες που παρέχουν οι SVM χρειάζεται η σωστή επιλογή των χαρακτηριστικών (κριτηρίων) που θα ενσωματωθούν στην ανάλυση, ο καθορισμός της τεχνικής παραμέτρου της διαδικασίας βελτιστοποίησης C και ο τρόπος με τον οποίο γίνεται η μη γραμμική αναπαράσταση των δεδομένων, δηλαδή η επιλογή του πυρήνα και της αντίστοιχης παραμέτρου του.

Η παρούσα εργασία αντιμετώπισε τα παραπάνω προβλήματα με τη χρήση των γενετικών αλγορίθμων. Συγκεκριμένα, χρησιμοποιήθηκαν γενετικοί αλγόριθμοι με δυαδικά χρωμοσώματα που το αρχικό τους μέρος αποτελείται από τόσα γονίδια-bit όσα και τα συνολικά χαρακτηριστικά (το 1 θα δηλώνει ότι το συγκεκριμένο χαρακτηριστικό επιλέγεται στην ανάλυση , ενώ το 0 ότι δεν επιλέγεται), τα επόμενα τρία bit αναπαριστούν την τιμή της παραμέτρου C και τα τρία τελευταία την παράμετρο του πυρήνα. Η καταλληλότητα του κάθε χρωμοσώματος υπολογίστηκε βάσει της ακρίβειας που πετυχαίνεται από τις Μηχανές Διανύσματος Υποστήριξης ως

μοντέλου ταξινόμησης όταν εκπαιδεύεται στα επιλεγμένα χαρακτηριστικά, την παράμετρο *C* και την παράμετρο του πυρήνα που προκύπτουν από το εν λόγω χρωμόσωμα. Τελικά, ο γενετικός αλγόριθμος συγκλίνει σε μία λύση-χρωμόσωμα που οδηγεί στην μέγιστη δυνατή ακρίβεια και δίνει την βέλτιστη τιμή της παραμέτρου *C*, την βέλτιστη τιμή του πυρήνα και ορίζει εκείνα τα χαρακτηριστικά που είναι τα πλέον σχετικά και θα πρέπει να λαμβάνονται υπόψη κατά την ταξινόμηση.

Η προτεινόμενη μεθοδολογία εφαρμόστηκε και εξετάστηκε σε τρία σύνολα δεδομένων σχετικά με την ανάπτυξη συστημάτων εκτίμησης πιστωτικού κινδύνου, προκειμένου να βαθμολογήσει τους πελάτες (επιχειρήσεις / ιδιώτες) μιας επιχείρησης ανάλογα με την ικανότητά τους να ανταποκριθούν στις οικονομικές τους υποχρεώσεις και έτσι να τους κατατάξει σε δύο κατηγορίες πελατών, στους συνεπείς και στους ασυνεπείς ως προς της υποχρεώσεις τους. Επιπλέον, εξετάστηκε η αποτελεσματικότητά της συγκριτικά με κάποιες άλλες στατιστικές μεθόδους, όπως η διακριτική ανάλυση και η λογιστική παλινδρόμηση.

Η ανάλυση των αποτελεσμάτων χωρίστηκε σε δύο κύρια μέρη, σε αυτά που προέκυψαν όταν μελετήθηκε (α) η Αναμενόμενη Ακρίβεια και (β) ο Δείκτης Ακρίβειας, ως μέτρο της ικανότητας γενίκευσης του Μοντέλου Ταξινόμησης και επομένως ως συνάρτηση καταλληλότητας του γενετικού αλγορίθμου. Ωστόσο, και τα δύο μέρη της ανάλυσης οδηγούν στα ίδια βασικά συμπεράσματα.

Αρχικά, μελετήθηκαν διάφορες συναρτήσεις επιλογής του γενετικού αλγορίθμου, όπως η μέθοδος ρουλέτας, η γεωμετρική επιλογή και η μέθοδος τουρνουά. Παρατηρήθηκε ότι και για στα δύο μέρη της ανάλυσης, οι διάφορες συναρτήσεις επιλογής δεν παρουσίασαν σημαντικές διαφοροποιήσεις μεταξύ τους ως προς τα αποτελέσματα. Ωστόσο, ένα μικρό προβάδισμα έναντι των άλλων παρουσίασε η γεωμετρική επιλογή γι΄ αυτό και υιοθετήθηκε και στα δύο μέρη της ανάλυσης.

Επιπλέον, και στα δύο μέρη της ανάλυσης εξετάστηκαν δύο ζευγάρια τιμών για το μέγεθος του πληθυσμού και τη συνθήκη τερματισμού. Παρατηρήθηκε ότι γενικά το ζευγάρι με το μεγαλύτερο μέγεθος στον πληθυσμό αποδίδει τις περισσότερες φορές λίγο καλύτερα από το άλλο ζευγάρι, για οποιοδήποτε επιλογή πυρήνα και συνάρτηση καταλληλότητας γι’ αυτό και χρησιμοποιήθηκε περισσότερο στις αναλύσεις.

Όσον αφορά την επιλογή του πυρήνα, θα μπορούσε να σημειωθεί ότι ο γραμμικός πυρήνας δίνει τις περισσότερες φορές τα καλύτερα αποτελέσματα, ακολουθούμενος από τον πυρήνα RBF και τέλος τον πυρήνα πολυωνύμου. Ειδικά, όταν εξετάστηκε ο Δείκτης Ακρίβειας ως συνάρτηση καταλληλότητας, ο γραμμικός πυρήνας υπερείχε έναντι των άλλων για όλα τα αρχεία δεδομένων. Το συμπέρασμα που προέκυψε σχετικά με τα χαρακτηριστικά που επιλέγονταν από κάθε πυρήνα είναι ότι και οι τρεις πυρήνες επιλέγουν κάποια κοινά χαρακτηριστικά που μπορούν να θεωρηθούν και ως τα πλέον σημαντικά στην ανάλυση.

Στο δεύτερο μέρος της ανάλυσης όπου μελετήθηκαν οι πυρήνες όχι μόνο όταν γίνεται επιλογή χαρακτηριστικών , αλλά και όταν δεν γίνεται, παρατηρήθηκε ότι για τα δύο μικρότερα αρχεία η επιλογή χαρακτηριστικών βελτίωνε τα αποτελέσματα, όχι όμως και για το δεύτερο αρχείο δεδομένων.

Συγκριτικά με τις στατιστικές μεθόδους, παρατηρείται ότι τις περισσότερες φορές η προτεινόμενη μεθοδολογία υπερέχει έναντι αυτών ή βρίσκεται περίπου στα ίδια επίπεδα τιμών. Συνολικά λοιπόν διαπιστώνεται ότι η μεθοδολογία που παρουσιάστηκε στα πλαίσια αυτής της μεταπτυχιακής διατριβής αποδίδει πολύ καλά και ωθεί στην περαιτέρω διερεύνηση και βελτίωσή της. Για παράδειγμα, θα μπορούσαν να εξεταστούν διαφορετικές παράμετροι για τους γενετικούς τελεστές του γενετικού αλγορίθμου ή ακόμα και να ερευνηθούν νέοι τελεστές διασταύρωσης και μετάλλαξης και νέες συναρτήσεις πυρήνων για τις Μηχανές Διανύσματος Υποστήριξης.

‘Βιβλιογραφία’

[Altman77] E. I. Altman, R. Haldeman, , P. Narayanan, Zeta analysis: A new model to identify bankruptcy risk of corporations. Journal of Banking and Finance, pp. 29-54, 1977

[Altman88] E. I. Altman, Default Risk, Mortality Rates, and the Performance of Corporate Bonds. Research Foundation, Institute of Chartered Financial Analysts, Charlottesville, VA, 1988

[Altman89] E. I. Altman, Measuring corporate bond mortality and performance. Journal of Finance, September, pp. 909-922, 1989

[Altman94] E. I. Altman, G. Marco, F. Varetto, Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (The Italian Experience), Journal of Banking and Finance, pp. 505-529, 1994

[Asquith89] P. Asquith, D.W. Mullins Jr., E. D. Wolff , Original issue high yield bonds: Aging analysis of defaults, exchanges and calls. Journal of Finance, pp. 923-953, 1989

[BlackScholes73] F. Black, M. Scholes, The pricing of options and corporate liabilities. Journal of Political Economy, pp. 637-659, 1973

[BluLan97] A. L. Blum, P. Langley, Selection of Relevant Features and Examples in Machine Learning, In: Artificial Intelligence, Vol. 97:12, S. 245-271, 1997

[BriBroMar92] F. Brill, D. Brown, W. Martin, Fast genetic selection of features for neural network classifiers, In: IEEE Transactions on Neural Networks, 3(2): 324 - 328, 1992

[ChaVap00] O. Chapelle, V. Vapnik, Model selection for support vector machines, In: Sara A. Solla, Todd K. Leen, Klaus-Robert Muller (editors), Advances in Neural Information Processing Systems 12, MIT Press, Cambridge, MA, 2000

[CoatsFant93] P. Coats, L. Fant, Recognizing financial distress patterns using a neural network tool. Financial Management, pp. 142-155, 1993

[CorVap95] C. Cortes, V. Vapnik, Support vector networks, In: Machine Learning, 20:273-297, 1995

[Courtis78] J. K. Courtis, Modelling a financial ratios categoric framework. Journal of Business Finance and Accounting 5 (4), pp. 371–386, 1978

[CrisSha00] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000

[CrisShaEliKan02] N. Cristianini, J. Shawe-Taylor, A. Elisseff, J. Kandola, On Kernel-Target Alignment, In: T. G. Dietterich, S. Becker, Z. Ghahramani (editors), Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA, 2002

[DeJong75] K. DeJong, An Analysis of the Behavious of a Class of Genetic Adaptive Systems, PhD Dissertion, Department of Computer and Communication Sciences, University of Michigan, Ann Arbor, 1975

[DoumpoKosmBaouZop02] M. Doumpos, K. Kosmidou, G. Baourakis, C. Zopounidis, Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis, European Journal of Operational Research 138. pp. 392–412, 2002

[FerKadKit93] F. J. Ferri, V. Kadirkamanathan, J. Kittler, Feature Subset Search using Genetic Algorithms, In: IEE/IEEE Workshop on Natural Algorithms in Signal Processing, Essex, 1993

[Goldberg98] D. Goldberg, Genetic Algorithms in Search, Optimization in Machine Learning, Addison Wesley, Reading, 1998

[Henley95] W. E. Henley, Statistical aspects of credit scoring. PhD thesis, Open University, 1995

[HenleyHand96] W. E. Henley, D. J. Hand, A k-NN classifier for assessing consumer credit risk. The Statistician 65, pp. 77–95, 1996

[Holland75] J. H. Holland, Adaption in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, 1975

[HullWhite95] J. Hull, A. White, The impact of default risk on the prices of options and other derivative securities. Journal of Banking and Finance, pp. 299-322, 1995

[JoinesHouck94] J. Joines, C. Houck, On the use of non-stationary penalty functions to solve constrained optimization problems with genetic algorithms, In: IEEE International Symposium Evolutionary Computation, Orlando, Fl, pp. 579-584, 1994

[Kealhofer96] S. Kealhofer, Measuring Default Risk in Portfolios of Derivatives. Mimeo KMV Corporation, San Francisco, CA, 1996

[KMV93] KMV Corporation, Credit Monitor Overview, San Francisco, Ca, USA, 1993

[Kohavi97] R. Kohavi, G. John, Wrappers for Feature Subset Selection, In: Artificial Intelligence, Vol. 97:12, pp. 273 – 324, 1997

[Lawrence92] E. L. Lawrence, S. Smith, M. Rhoades, An analysis of default risk in mobile home credit. Journal of Banking and Finance, pp. 299-312, 1992

[LesEskNob02] C. Leslie, E. Eskin, W. Noble, The spectrum kernel: A string kernel for SVM protein classification, In: Proc. Pacific Symposium on Bio-computing, pp. 564 - 575, 2002

[Martin77] D. Martin, Early warning of bank failure: A logit regression approach. Journal of Banking and Finance, pp 249-276, 1977

[McAllisterMingo94] P. McAllister, J. J. Mingo, Commercial loan risk management, credit-scoring and pricing: The need for a new shared data base. Journal of Commercial Bank Lending, pp. 6-20, 1994

[McElraveyShah96] J. N. McElravey, V. Shah, Rating Cash Flow Collateralized Bond Obligations. Special Report, Asset Backed Securities, Duff and Phelps Credit Rating Co., Chicago, IL, USA, 1996

[McKinsey93] McKinsey, Special report on ``The new world of financial services'' The McKinsey Quarterly, Number 2, 1993

[Merton74] R. Merton, On the pricing of corporate debt. Journal of Finance, pp. 449-470, 1974

[Michalewicz94] Z. Michalewicz, Genetic Algorithms and Data Structures, Evolution Programs. AI Series. Springer-Verlag, New York, 1994

[Moody's90] Moody's Special Report, Corporate Bond Defaults and Default Rates, 1970-1989, April, 1990

[RayPunGooKuhJai00] M. Raymer, W. Punch, E. Goodman, L. Kuhn, A. Jain, Dimensionality Reduction Using Genetic Algorithms, In: IEEE Transactions on Evolutionary computing, 2000

[RicLan96] M. Richeldi, P. Lanzi, A Tool for Performing effective feature selection by investigating the deep structure of the data, In: Proceedings of the International Conference on Tools with Artificial Intelligence, pp. 102 - 105, 1996

[Rojas96] R. Rojas, Neural Networks - A Systematic Introduction, Springer, Berlin, 1996

[RusNor95] S.J. Russel, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, Englewood Cliffs, New Jersey 07632, 1995

[SantomeroVinso77] A. Santomero, J. Vinso, Estimating the probability of failure for firms in the banking system. Journal of Banking and Finance, pp. 185-206, 1977

[SchSmo02] B. Scholkopf, A. J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002

[Scott81] J. Scott, The probability of bankruptcy: A comparison of empirical predictions and theoretical models. Journal of Banking and Finance. September, pp. 317-344, 1981

[SmithLawrence95] L. D. Smith, E. Lawrence, Forecasting losses on a liquidating long-term loan portfolio. Journal of Banking and Finance, pp. 959-985, 1995

[SmoSch98] A. J. Smola, B. Scholkopf: A Tutorial on Support Vector Regression, NeuroCOLT2 NC2-TR-1998-030, 1998

[SommerTaffler95] R. A. Sommerville, R. J. Taffler, Banker judgement versus formal forecasting models: The case of country risk assessment. Journal of Banking and Finance, pp. 281-297, 1995

[Spackman89] K. A. Spackman, Signal detection theory: Valuable tools for evaluating inductive learning, In: Proceedings of the Sixth International Workshop on Machine Learning, pp. 160-163 San Mateo, CA. Morgan Kaufman, 1989

[Standard and Poor's91] Standard and Poor's, Corporate Bond Default Study. Credit Week, September 16, 1991

[TrippiTurban96] R. Trippi, E. Turban, Neural Networks in Finance and Investing, revised ed. Irwin, Homewood, IL, 1996

[UltschVL00] A. Ultsch, Knowledge Discovery, Vorlesung an der Universitat Marburg, Sommersemester 2000

[Vapnic79] V. Vapnik, Estimation of Dependencies Based on Empirical Data [in Russian], Nauka, Moscow, 1979 (English translation: Springer Verlag, New York, 1982)

[Vapnic95] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, NY, 1995

[Vapnic98] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, New York, 1998

[VapCha00] V. Vapnik, O. Chapelle, Bounds on error expectation for support vector machines, In: Neural Computation, 12 (9), 2000

[VapChe74] V. Vapnik, A. Chervonenkis, Ordered risk minimization, In: Automation and Remote Control, 35:1226-1235, 1403-1412, 1974

[VapChe79] V. Vapnik, A. Chervonenkis, Theory of Pattern Recognition, Nauka, Moscow, 1974

[West85] R. C. West, A factor-analytic approach to bank condition. Journal of Banking and Finance, pp. 253-266, 1985

[Whitley93] D. Whitley, A Genetic Algorithm Tutorial, Department of Computer Science, Colorado State University, 1993

[Wilcox73] J. W. Wilcox, A Prediction of business failure using accounting data. Journal of Accounting Research, Vol. 2, 1973