Π Ο Λ Υ Τ Ε Χ Ν Ε Ι Ο   Κ Ρ Η Τ Η Σ
Τ Μ Η Μ Α   Η Λ Ε Κ Τ Ρ Ο Ν Ι Κ Ω Ν   Μ Η Χ Α Ν Ι Κ Ω Ν   Κ Α Ι
Μ Η Χ Α Ν Ι Κ Ω Ν   Υ Π Ο Λ Ο Γ Ι Σ Τ Ω Ν

Τ Ο Μ Ε Α Σ   Τ Η Λ Ε Π Ι Κ Ο Ι Ν Ω Ν Ι Ω Ν

Εκτίμηση και Αξιολόγηση Ταξινομητών στη Βιοϊατρική

Διατριβή που υπεβλήθη για τη μερική ικανοποίηση των
απαιτήσεων για την απόκτηση μεταπτυχιακού
διπλώματος στον Τομέα Τηλεπικοινωνιών

Σαλάππα Αθηνά

Χανιά 2009

Η διατριβή της Σαλάππα Αθηνά εγκρίνεται από τους κ.κ.

Ζερβάκης Μιχάλης	Καθηγητής
Μπάλας Κωνσταντίνος	Αναπληρωτής Καθηγητής
Σταυρακάκης Γεώργιος	Καθηγητής

Π Ε Ρ Ι Ε Χ Ο Μ Ε Ν Α

Κ Α Τ Α Λ Ο Γ Ο Σ   Σ Χ Η Μ Α Τ Ω Ν

Κ Α Τ Α Λ Ο Γ Ο Σ   Π Ι Ν Α Κ Ω Ν

Ε Υ Χ Α Ρ Ι Σ Τ Ι Ε Σ

Με το πέρας της παρούσας ερευνητικής εργασίας θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Ζερβάκη Μιχαήλ τόσο για τη σημαντική βοήθεια που μου παρείχε στη διεκπεραίωση της εργασίας, όσο και για τις γνώσεις που μου προσέφερε κατά τη διάρκεια του μεταπτυχιακού μου.

Επίσης, θα ήθελα να ευχαριστώ τον Επίκουρο Καθηγητή του Τμήματος Μηχανικών Παραγωγής και Διοίκησης κ. Μιχάλη Δούμπο για τη βοήθεια που μου προσέφερε σε διάφορες δυσκολίες που αντιμετώπισα κατά την εκπόνηση αυτής.

Τέλος, ευχαριστώ την οικογένειά μου για την ηθική υποστήριξή της όλα αυτά τα χρόνια που λείπω από κοντά της και στην οποία αφιερώνεται και η παρούσα ερευνητική μεταπτυχιακή διατριβή.

Σαλάππα Αθηνά
Χανιά
Φεβρουάριος 2009

Κεφάλαιο 1°

ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΣΥΓΚΡΙΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

1.1 Εισαγωγή

Η ταξινόμηση ενός συνόλου εναλλακτικών παρατηρήσεων ή αντικειμένων σε προκαθορισμένες ομοιογενείς κατηγορίες είναι ένα πρόβλημα ιδιαίτερου πρακτικού και ερευνητικού ενδιαφέροντος. Αυτού του είδους τα προβλήματα αναφέρονται ως προβλήματα απλής ταξινόμησης (classification) ή διατεταγμένης ταξινόμησης (sorting), ανάλογα με το εάν οι κατηγορίες ταξινόμησης ορίζονται ονομαστικά (nominal) ή είναι διατεταγμένες (ordinal). Η προβληματική της ταξινόμησης παρέχει το πλαίσιο μιας εναλλακτικής θεώρησης των προβλημάτων σε σύγκριση με άλλες προβληματικές, όπως η ομαδοποίηση (clustering), η επιλογή της καλύτερης εναλλακτικής (choice), η κατάταξη των εξεταζόμενων εναλλακτικών από τις καλύτερες στις χειρότερες βάσει των χαρακτηριστικών τους (ranking) και η περιγραφή των εναλλακτικών δραστηριοτήτων για τον εντοπισμό των βασικών τους ιδιοτήτων (description). Για τη μελέτη των προβλημάτων ταξινόμησης έχουν αναπτυχθεί κατά καιρούς διάφορες μεθοδολογίες από τους χώρους της στατιστικής και της οικονομετρίας, της τεχνικής νοημοσύνης και της επιχειρησιακής έρευνας. Η ανάπτυξη και χρήση ποσοτικών τεχνικών ταξινόμησης κρίνεται απαραίτητη τόσο για την καλύτερη αντιμετώπιση του εξεταζόμενου προβλήματος, όσο και για την σημαντική μείωση του χρόνου και του κόστους που απαιτούνται για την αντιμετώπισή του. Γενικά, όπως αναφέρεται και παρακάτω, η αντιμετώπιση του προβλήματος της ταξινόμησης βάσει των διαθέσιμων μεθοδολογικών προσεγγίσεων συνίσταται στην ανάπτυξη ποσοτικών υποδειγμάτων, τα οποία υποστηρίζουν την διαδικασία επίλυσης προβλημάτων στη βάση της προβληματικής της ταξινόμησης.

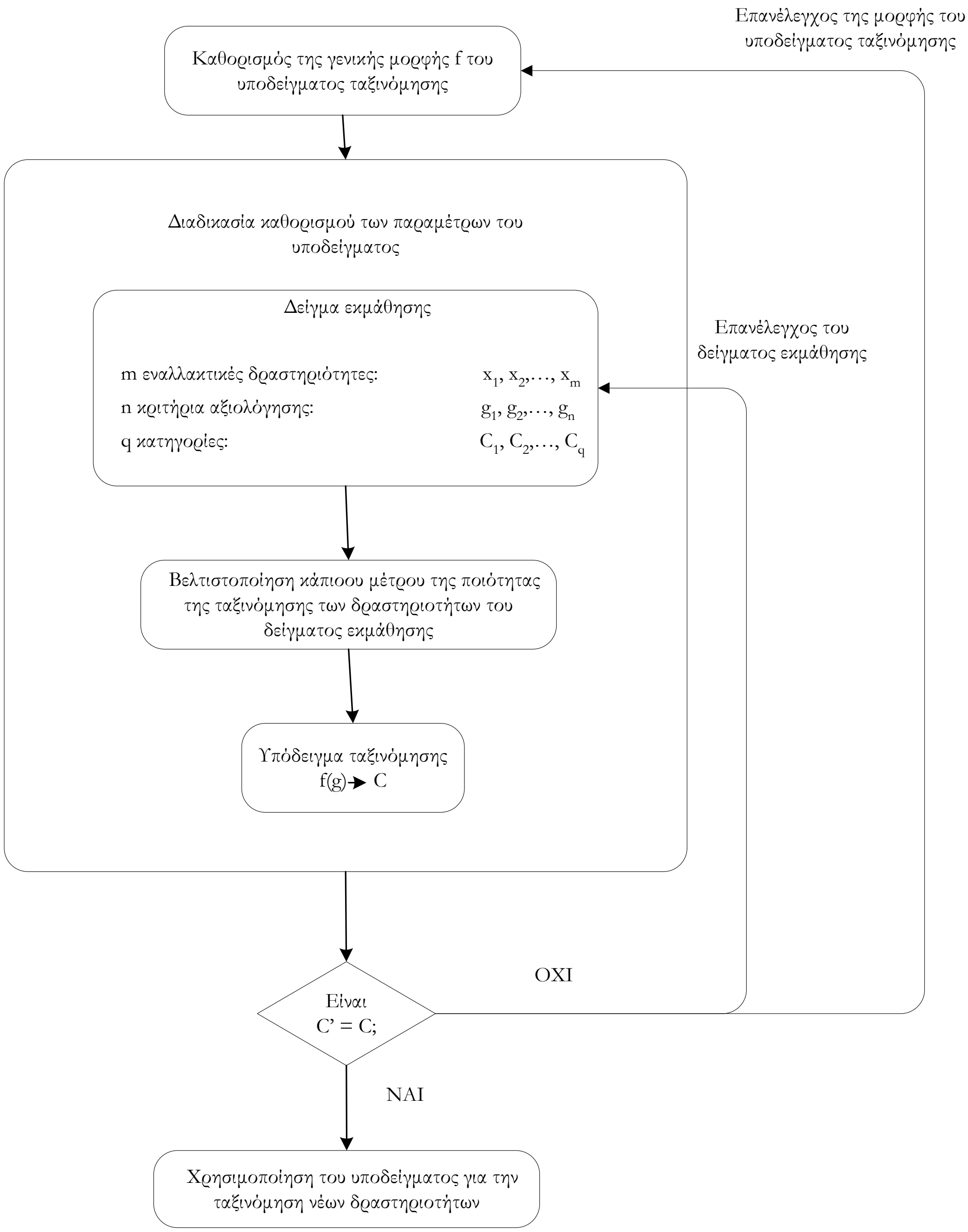
1.2 Ανάπτυξη υποδειγμάτων ταξινόμησης

Οι μεθοδολογικές προσεγγίσεις για την ανάπτυξη υποδειγμάτων ταξινόμησης ακολουθούν την γενική φιλοσοφία της παλινδρόμησης, προσπαθώντας να αξιοποιήσουν τη διαθέσιμη γνώση και πληροφορία που απορρέει από το γεγονός ότι οι κατηγορίες είναι προκαθορισμένες.

Στην ανάλυση παλινδρόμησης στόχος είναι ο εντοπισμός της συναρτησιακής σχέσης που συνδέει μια εξαρτημένη μεταβλητή Υ με ένα διάνυσμα μεταβλητών Χ βάσει της ανάλυσης ενός συνόλου δεδομένων παρατηρήσεων (Χ, Υ). Κατά ανάλογο τρόπο αντιμετωπίζεται και το πρόβλημα της ταξινόμησης με την μόνη διαφορά ότι η εξαρτημένη μεταβλητή δεν είναι συνεχής, αλλά αφορά ένα περιορισμένο σύνολο διακριτών επιπέδων καθένα από τα οποία αντιστοιχεί σε μια κατηγορία. Το δείγμα των παρατηρήσεων που χρησιμοποιείται για την ανάπτυξη των υποδειγμάτων ταξινόμησης ονομάζεται δείγμα εκμάθησης και περιλαμβάνει ζεύγη της μορφής (Χ, C), όπου ως C συμβολίζεται η εξαρτημένη μεταβλητή που υποδηλώνει την ταξινόμηση των εναλλακτικών δραστηριοτήτων σε ένα σύνολο κατηγοριών q. Η επίλυση του προβλήματος της ταξινόμησης συνίσταται στην ανάπτυξη ενός υποδείγματος της μορφής $f(X) \rightarrow C$ το οποίο ελαχιστοποιεί ένα μέτρο των διαφορών που εντοπίζονται μεταξύ της εκτιμώμενης ταξινόμησης $\hat{C}$ και της δεδομένης ταξινόμησης C. Εφόσον, ολοκληρωθεί η ανάπτυξη του υποδείγματος ταξινόμησης μπορεί πλέον αυτό να χρησιμοποιηθεί για την ταξινόμηση οποιονδήποτε άλλων εναλλακτικών, οι οποίες δεν συμπεριλαμβάνονται στο δείγμα εκμάθησης. Η χρησιμότητα της παραπάνω διαδικασίας βασίζεται στην εκμετάλλευση της υπάρχουσας γνώσης από το δείγμα εκμάθησης, με σκοπό την μοντελοποίηση και αναπαράστασή της σε ένα υπόδειγμα ταξινόμησης, το οποίο θα διαθέτει την απαραίτητη ικανότητα γενίκευσης.

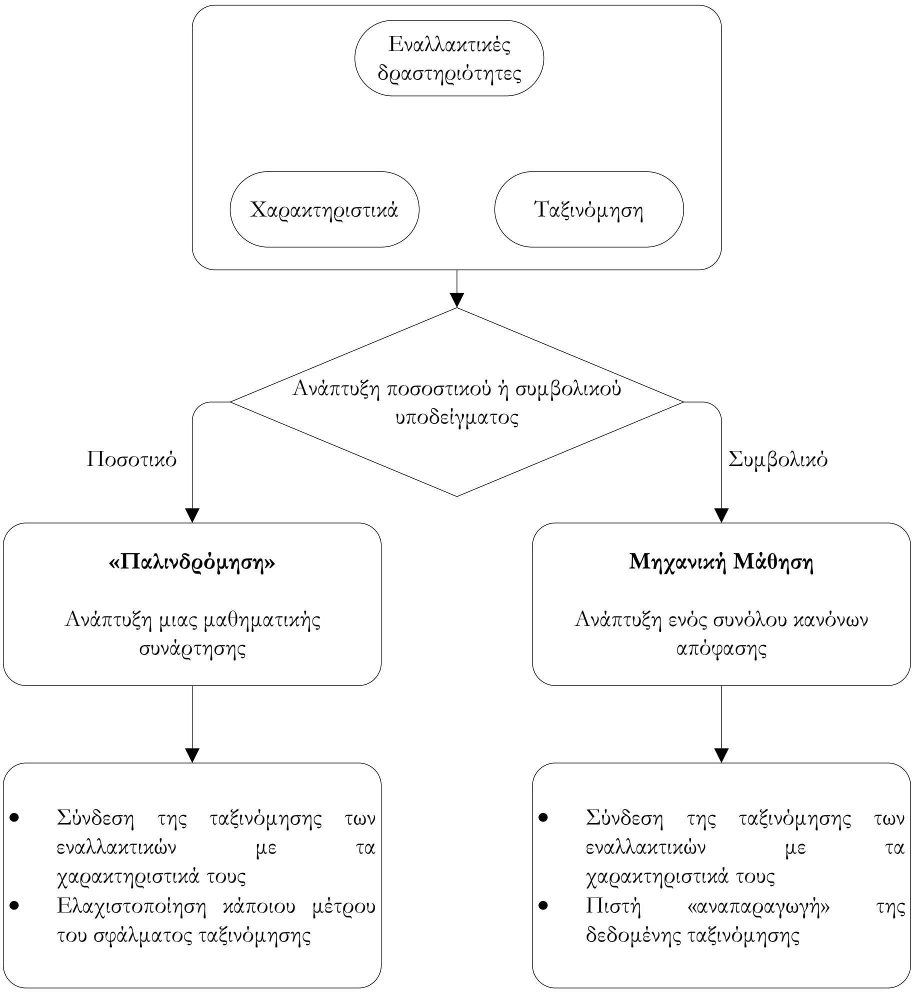
Γενικά, τα αναπτυσσόμενα υποδείγματα ταξινόμησης είναι μια συνάρτηση, η οποία συνδυάζει όλα τα επιμέρους χαρακτηριστικά των εναλλακτικών δραστηριοτήτων σε έναν ολικό ποσοτικό δείκτη βάσει του οποίου λαμβάνονται οι αποφάσεις για την ταξινόμηση των εναλλακτικών δραστηριοτήτων. Ο δείκτης αυτός μπορεί να αναπαριστά την πιθανότητα να ανήκει μια εναλλακτική σε μια κατηγορία ή κάποια αυθαίρετη «βαθμολογία» η οποία σε συνδυασμό με κατάλληλους κανόνες οδηγεί στην ταξινόμηση των εναλλακτικών.

Ένα γενικό περίγραμμα της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης φαίνεται παρακάτω (σχήμα 1.1):



Σχήμα 1. 1: Γενικό περίγραμμα της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης

Η παραπάνω διαδικασία ακολουθείται, τουλάχιστον στη γενική της φιλοσοφία, από τη συντριπτική πλειοψηφία των μεθόδων ταξινόμησης. Μεταξύ όμως των διαφόρων μεθόδων ταξινόμησης υπάρχουν ουσιαστικές διαφορές τόσο όσον αφορά τη μορφή των υποδειγμάτων ταξινόμησης που αναπτύσσονται, όσο καις τη διαδικασία καθορισμού των παραμέτρων τους. Ένας γενικός διαχωρισμός των μεθόδων ταξινόμησης ως προς τη μορφή των υποδειγμάτων που αναπτύσσουν παρουσιάζεται στο σχήμα 1.2:



Σχήμα 1. 2: Μορφές υποδειγμάτων ταξινόμησης

Τα αναπτυσσόμενα υποδείγματα ταξινόμησης, στη συνηθέστερη μορφή τους, είναι μια συνάρτηση η οποία συνδυάζει όλα τα επιμέρους χαρακτηριστικά των εναλλακτικών δραστηριοτήτων σε έναν ολικό ποσοτικό δείκτη βάσει του οποίου λαμβάνονται οι αποφάσεις για την ταξινόμησης των εναλλακτικών δραστηριοτήτων. Ο δείκτης αυτός συνήθως αναπαριστά την πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα σε μια κατηγορία ή την αξία της κάθε εναλλακτικής δραστηριότητας σε μια κανονικοποιημένη ή μη κανονικοποιημένη κλίμακα. Καθώς η αναπτυσσόμενη συνάρτηση χαρακτηρίζει όλες τις εναλλακτικές δραστηριότητες, είναι δυνατό να μην είναι σε θέση να αναπαράγει πλήρως τη δεδομένη ταξινόμηση των εναλλακτικών δραστηριοτήτων του δείγματος εκμάθησης. Για το λόγο αυτό η ανάπτυξή της βασίζεται στην ελαχιστοποίηση κάποιου μέτρου του σφάλματος ταξινόμησης. Τα μοντέλα αυτά αναφέρονται ως ποσοτικά με τη λογική ότι το αποτέλεσμά τους αφορά τον υπολογισμό ενός ποσοτικού δείκτη για την πραγματοποίηση της επιθυμητής ταξινόμησης.

Η δεύτερη βασική μορφή υποδειγμάτων ταξινόμησης είναι συμβολικής μορφής. Τεχνικές που ακολουθούν αυτή τη μεθοδολογική κατεύθυνση οδηγούν στην ανάπτυξη ενός συνόλου κανόνων της μορφής «ΕΑΝ συνθήκες ΤΟΤΕ αποτέλεσμα ταξινόμησης». Το τμήμα των συνθηκών του κάθε κανόνα αναφέρεται στα χαρακτηριστικά των εναλλακτικών δραστηριοτήτων, ουσιαστικά δηλαδή στις συνθήκες που πρέπει να πληροί η κάθε εναλλακτική δραστηριότητα προκειμένου να ενταχθεί στην κατηγορία που υποδεικνύει το συμπέρασμα του κανόνα. Σε πολλές περιπτώσεις το αποτέλεσμα ταξινόμησης (συμπέρασμα κανόνων) συνδυάζεται και με έναν αριθμητικό συντελεστή, ο οποίος αναπαριστά τη βεβαιότητα ή την ισχύ του αποτελέσματος. Ο επιστημονικός χώρος ο οποίος ασχολείται με την ανάπτυξη και εφαρμογή μεθοδολογιών για την ανάπτυξη τέτοιων υποδειγμάτων ταξινόμησης είναι γνωστός ως μηχανική μάθηση (machine learning). Οι κανόνες που αναπτύσσονται έχουν γενικά τη δυνατότητα απόλυτης προσαρμογής στα δεδομένα του δείγματος εκμάθησης, καθώς είναι δυνατό να αναπτυχθεί ένας κανόνας για κάθε εναλλακτική δραστηριότητα χωρίς βέβαια να συνεπάγεται και υψηλότερη ικανότητα γενίκευσης. Για το λόγο αυτό συχνά χρησιμοποιούνται τεχνικές επιλογής ενός περιορισμένου συνόλου αποτελούμενου από γενικούς κανόνες οι οποίοι είναι σε θέση να καλύψουν πολλές περιπτώσεις (pruning).

1.2.1 Το πρόβλημα της ταξινόμησης

Το πρόβλημα της ταξινόμησης βασίζεται στην πραγματοποίηση απόλυτων συγκρίσεων (absolute comparisons). Κάθε εναλλακτική δραστηριότητα εντάσσεται σε μια εκ των προκαθορισμένων κατηγοριών βάσει ενός συγκεκριμένου κανόνα, ο οποίος συνήθως αναφέρεται στη σύγκριση με συγκεκριμένα πρότυπα τα οποία διαχωρίζουν τις κατηγορίες. Η εφαρμογή του κανόνα ταξινόμησης, καθώς επίσης και το αποτέλεσμα της αξιολόγησής της, δεν επηρεάζεται από το σύνολο των εξεταζόμενων δραστηριοτήτων. Για παράδειγμα, μια εκτίμηση της μορφής «το προϊόν Χ είναι ακατάλληλο» είναι απόλυτη, καθώς δεν επηρεάζεται με κανέναν τρόπο από το σύνολο των ομοειδών προϊόντων του Χ. Βέβαια μερικές φορές, τέτοιου είδους εκτιμήσεις δεν είναι πάντα απόλυτες, καθώς καθορίζονται μέσα στο γενικό πλαίσιο που χαρακτηρίζει το περιβάλλον του προβλήματος ή της απόφασης. Για παράδειγμα, μια επιχείρηση μπορεί να πληροί τις προϋποθέσεις χρηματοδότησης από ένα χρηματοπιστωτικό ίδρυμα υπό συγκεκριμένες συνθήκες του γενικότερου οικονομικού και επιχειρηματικού περιβάλλοντος. Καθώς όμως οι συνθήκες αυτές εξελίσσονται δυναμικά, μεταβάλλονται ανάλογα και οι προϋποθέσεις χρηματοδότησης, με αποτέλεσμα η ίδια επιχείρηση να κριθεί ακατάλληλη προς χρηματοδότηση. Γενικά, παρά τις όποιες μεταβολές που είναι δυνατόν να πραγματοποιηθούν στο χρησιμοποιούμενο κανόνα ταξινόμησης, ο καθορισμός του πραγματοποιείται πάντα ανεξάρτητα από το σύνολο των εξεταζόμενων εναλλακτικών δραστηριοτήτων. Αυτή είναι και η βασικότερη διαφορά του προβλήματος της ταξινόμησης από άλλες προβληματικές (επιλογή, κατάταξη).

Ένας γενικός ορισμός της ταξινόμησης όπως προτάθηκε από τον Mirkin (1996) είναι ο εξής:

«Ταξινόμηση είναι η ρεαλιστική ή ιδεατή τοποθέτηση μαζί παρόμοιων αντικειμένων και ο διαχωρισμός των αντικειμένων τα οποία διαφέρουν, με απώτερο σκοπό: (α) τη διαμόρφωση, οργάνωση και διατήρηση της γνώσης, (β) την ανάλυση της δομής του φαινομένου που εξετάζεται, (γ) τη συσχέτιση των διαφόρων πλευρών του υπό εξέταση φαινομένου.»

Οι συνηθέστεροι όροι που χρησιμοποιούνται για την αναφορά στο πρόβλημα της ταξινόμησης είναι οι εξής:

- Ταξινόμηση (Classification)
- Διατεταγμένη ταξινόμηση (Sorting)

Ο πρώτος όρος αναφέρεται σε προβλήματα όπου οι κατηγορίες στις οποίες θα ενταχθούν οι εξεταζόμενες εναλλακτικές δραστηριότητες, ορίζονται κατά ονομαστικό τρόπο. Χαρακτηριστικά παραδείγματα αυτής της κατηγορίας είναι τα προβλήματα αναγνώρισης προτύπων (pattern recognition) τα οποία παρουσιάζουν σημαντικές εφαρμογές σε τεχνολογίες αναγνώρισης φωνής, γραφής, ακόμα και στην ιατρική. Αντίθετα, ο δεύτερος όρος αναφέρεται σε προβλήματα όπου οι οριζόμενες κατηγορίες, στις οποίες θα πραγματοποιηθεί η ταξινόμηση των εναλλακτικών δραστηριοτήτων, είναι διατεταγμένες από τις καλύτερες στις χειρότερες. Τέτοια προβλήματα απαντώνται συχνά στην πράξη σε χώρους, όπως στη χρηματοοικονομική διοίκηση, το μάρκετινγκ, τη διοίκηση ανθρωπίνου δυναμικού, στη περιβαλλοντική διαχείριση, κ.ά.

1.2.2 Σπουδαιότητα της διαδικασίας ταξινόμησης

Η σημασία του προβλήματος ταξινόμησης δεν περιορίζεται μόνο στην πολυπλοκότητα που παρουσιάζει ως ένα επιστημονικό πεδίο έρευνας, αλλά επεκτείνεται και σε πρακτικό επίπεδο. Χαρακτηριστικές είναι οι παρακάτω πρακτικές εφαρμογές:

- *Ιατρική:* Πραγματοποίηση ιατρικών διαγνώσεων ταξινομώντας τους ασθενείς σε κατηγορίες με βάση τα συμπτώματα που παρουσιάζουν (Tsumoto (1998), Belacel(2000)).

- *Αναγνώριση προτύπων:* Διερεύνηση των χαρακτηριστικών φυσικών προσώπων ή αντικειμένων και ταξινόμησή τους σε ανάλογες κατηγορίες. Χαρακτηριστικά παραδείγματα της αναγνώρισης βασικών ανθρώπινων χαρακτηριστικών είναι η αναγνώριση φωνής, δακτυλικών αποτυπωμάτων και οι εφαρμογές τους στην ασφάλεια καίριων συστημάτων (Ripley (1996), Young και Fu (1997), Nieddu και Patrizi (2000)).

- *Διαχείριση ανθρωπίνου δυναμικού:* Αξιολόγηση του ανθρώπινου δυναμικού βάσει των προσόντων του, με απώτερο σκοπό τον προσδιορισμό της κατάλληλης θέσης (Rulon et al. (1967) και Gochet et al. (1997)).

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

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

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

- *Χρηματοοικονομική Διοίκηση και Οικονομική Θεωρία:* Πρόβλεψη της πτώχευσης επιχειρήσεων, εκτίμηση του πιστωτικού κινδύνου επιχειρήσεων και καταναλωτών, επιλογή και διαχείριση χαρτοφυλακίων επενδύσεων, αξιολόγηση δανειοληπτικής ικανότητας χωρών (Zopounidis (1998), Doumpos και Zopounidis (1998)).

Τα παραπάνω προβλήματα πιστοποιούν την σπουδαιότητα του προβλήματος ταξινόμησης και της ανάπτυξης των αντίστοιχων αποτελεσματικών υποδειγμάτων.

1.3 Προτεινόμενη μεθοδολογική προσέγγιση και στόχοι της έρευνας

Για την αντιμετώπιση προβλημάτων ταξινόμησης, στη διεθνή βιβλιογραφία έχουν προταθεί διάφοροι αλγόριθμοι και μεθοδολογίες, που έχουν εφαρμοστεί με επιτυχία στη Βιοϊατρική. Παρά όμως τη σημαντική έρευνα που έχει πραγματοποιηθεί στον χώρο αυτό δεν υπάρχει μια ολοκληρωμένη έρευνα που να αναλύει θέματα όπως, ο συνδυασμός αλγορίθμων επιλογής χαρακτηριστικών σε βιοϊατρικά δεδομένα με μεγάλο πλήθος χαρακτηριστικών, η επιλογή κριτηρίων μέτρησης της διαφορετικότητας των υποδειγμάτων ταξινόμησης και τεχνικών συνδυασμού των υποδειγμάτων και τέλος, η επιμέρους εκτίμηση της αποτελεσματικότητας των προτεινόμενων μεθοδολογιών και αλγορίθμων, καθώς και η ανάλυση αυτής με τη χρήση ευρέως διαδεδομένων και γνωστών στατιστικών μεθόδων. Μια τέτοια ολοκληρωμένη έρευνα μπορεί να εξάγει χρήσιμα αποτελέσματα για τις «αλληλεπιδράσεις» μεταξύ των μεθοδολογιών και αλγορίθμων που χρησιμοποιούνται και να εξετάσει όλες τις πτυχές και τις διαδικασίες ελέγχου μέτρησης της αποτελεσματικότητας στην αντιμετώπιση προβλημάτων ταξινόμησης σε βιοϊατρικά δεδομένα.

Βασικός σκοπός της συγκεκριμένης διατριβής είναι η διεξαγωγή μιας πλήρης ανάλυσης της επίδοσης και αποτελεσματικότητας του συνδυασμού ταξινομητών στην αντιμετώπιση προβλημάτων ταξινόμησης. Η όλη πειραματική ανάλυση ξεκινάει με την χρήση αλγορίθμων επιλογής χαρακτηριστικών, οι οποίοι θα συσχετιστούν με σκοπό την επιλογή του πλέον κατάλληλου συνόλου χαρακτηριστικών, τα οποία θα περιγράφουν με ικανοποιητικό βαθμό την ταξινόμηση ενός αντικειμένου. Εν συνεχεία, επιλέγονται οι εξεταζόμενες τεχνικές ταξινόμησης, οι οποίες είναι γνωστές στατιστικές τεχνικές, αλλά και διαδεδομένες μη παραμετρικές προσεγγίσεις, οι οποίες θα εξεταστούν, είτε μεμονωμένα είτε σε συνδυασμό, με σκοπό την εκτίμηση της αποτελεσματικότητας ταξινόμησης. Για να συνδυαστούν οι εξεταζόμενες τεχνικές και να έχει νόημα ο συνδυασμός αυτών πρέπει αρχικά να εντοπιστεί εκείνο το κριτήριο με το οποίο θα μετρηθεί η διαφορετικότητα (συσχέτιση) των τεχνικών ταξινόμησης. Με αυτόν τον τρόπο θα εξεταστεί το πώς ένα υπόδειγμα αλληλεπιδρά με ένα άλλο και θα επισημανθούν οι βασικές διαφορές, ως προς τη λειτουργία και τη δομή τους. Τέτοιου είδους κριτήρια συντελούν στο εντοπισμό του πλέον κατάλληλου και αποτελεσματικού συνδυασμού υποδειγμάτων ταξινόμησης, χωρίς να απαιτείται η πραγματοποίηση όλων των δυνατών συνδυασμών αυτών για την εκτίμηση εκείνου του συνδυασμού που θα αποδώσει υψηλότερη ακρίβεια ταξινόμησης. Επομένως, με την χρήση ενός τέτοιου κριτηρίου επιλέγεται ο αριθμός των ταξινομητών που θα συνδυαστούν και χρησιμοποιείται τελικά μια μέθοδος συνδυασμού, ανάλογη με τη πληροφορία που έχουν δώσει οι διάφορες τεχνικές ταξινόμησης (εκτιμώμενες ταξινομήσεις ή πιθανότητες ταξινόμησης). Κύριος στόχος είναι να μετρηθεί η ακρίβεια ταξινόμησης, η ικανότητα των υποδειγμάτων στην αντιμετώπιση τέτοιων προβλημάτων (F Score), το εμβαδόν κάτω από την καμπύλη ROC και η απόδοση των διαστημάτων εμπιστοσύνης με μεθόδους διωνυμικής κατανομής ως εκτιμητές της ακρίβειας ταξινόμησης.

Επισημαίνεται ότι σε αντίστοιχες έρευνες από τη διεθνή βιβλιογραφία, η εκτίμηση της αποτελεσματικότητας στην αντιμετώπιση προβλημάτων ταξινόμησης περιορίζονταν στην εκτίμηση και τον υπολογισμό της ακρίβειας ταξινόμησης. Στην συγκεκριμένη διατριβή θα αποδειχθεί ότι μια πιο ολοκληρωμένη έρευνα για τη μέτρηση της αποτελεσματικότητας ταξινόμησης απαιτεί τη συμβολή απλών στατιστικών μεθόδων ελέγχου, όχι μόνο για τη συσχέτιση των αλγορίθμων επιλογής χαρακτηριστικών, αλλά και για την εκτίμηση της εκτιμώμενης ακρίβειας ταξινόμησης, ως το πλέον βασικό και αξιόπιστο μέτρο εκτίμησης και αξιολόγησης των προβλημάτων ταξινόμησης. Συνεπώς, οι αποδόσεις των υποδειγμάτων ταξινόμησης μεμονωμένα, αλλά και σε συνδυασμό εκτιμήθηκαν με τη μέτρηση της ακρίβειας σωστών ταξινομήσεων και τη χρήση των διαστημάτων εμπιστοσύνης για την εκτίμηση των αποκλίσεων στην εκτιμώμενη τιμή της ακρίβειας.

Η όλη διαδικασία ελέγχου που χρησιμοποιήθηκε στη συγκεκριμένη διατριβή δεν έχει εφαρμοστεί ολοκληρωτικά σε βιοϊατρικά δεδομένα και κατά αυτό τον τρόπο. Εν μέρει έχουν πραγματοποιηθεί κάποιες αναλύσεις, αλλά όχι στην ίδια διάσταση μεθοδολογιών και αλγορίθμων ταξινόμησης και με την συγκεκριμένη βηματική αλληλουχία μεθόδων, τεχνικών, διαδικασιών και μέτρων εκτίμησης και αξιολόγησης προβλημάτων ταξινόμησης.

1.4 Δομή της εργασίας

Η υπόλοιπη ανάλυση που πραγματοποιείται στην παρούσα εργασία οργανώνεται σε πέντε κεφάλαια ως εξής:

Στο δεύτερο κεφάλαιο που ακολουθεί γίνεται μια συνοπτική εισαγωγή της έννοιας της μηχανικής μάθησης (machine learning) και των κατηγοριών αυτής. Επίσης, πραγματοποιείται μια γενική περιγραφή της προετοιμασίας των δεδομένων και των μέτρων εκτίμησης των αλγορίθμων μηχανικής μάθησης, όπως το σφάλμα πρόβλεψης, η μέση απόλυτη απόκλιση, το μέσο τετραγωνικό σφάλμα, το τυπικό σφάλμα διαστήματος εμπιστοσύνης, ο πίνακας σύγχυσης (confusion matrix), το μέτρο F. Τέλος, πραγματοποιείται μια συνοπτική αναφορά στον τρόπο κατασκευής της καμπύλης ROC, ως διαδικασία αξιολόγησης υποδειγμάτων ταξινόμησης.

Στο τρίτο κεφάλαιο γίνεται μια ιστορική αναδρομή στις πιο διαδεδομένες τεχνικές ταξινόμησης και μια συνοπτική παρουσίαση των αλγορίθμων επιλογής χαρακτηριστικών σύμφωνα με τη στρατηγική διερεύνησης των χαρακτηριστικών που ακολουθούν, τη διαδικασία επιλογής των χαρακτηριστικών και τα κριτήρια αξιολόγησης της ποιότητας των επιλεγμένων χαρακτηριστικών. Επίσης, πραγματοποιείται μια γενική περιγραφή των εξεταζόμενων τεχνικών ταξινόμησης (κλασσική πιθανοθεωρητική θεωρία του Bayes, αλγόριθμος πλησιέστερου γείτονα, δέντρα λογιστικής παλινδρόμησης, μηχανές διανύσματος υποστήριξης και ακτινωτό λειτουργικό δίκτυο). Τέλος, γίνεται μια συνοπτική αναφορά στη διαδικασία ελέγχου cross-validation, ως μια τεχνική επαναληπτικής δειγματοληψίας για τον έλεγχο της αποτελεσματικότητας των αναπτυσσόμενων υποδειγμάτων ταξινόμησης.

Στο τέταρτο κεφάλαιο πραγματοποιείται μια γενική περιγραφή της έννοιας της επαγωγικής στατιστικής και της εκτιμητικής. Αναλύονται οι όροι στης σημειακής εκτίμησης και της εκτίμησης διαστήματος και δίνεται συνοπτικά μια θεωρητική προσέγγιση των διαστημάτων εμπιστοσύνης. Τέλος, εξηγείται ο τρόπος υπολογισμού των διαστημάτων εμπιστοσύνης για παραμέτρους γνωστών κατανομών (κανονική και διωνυμική κατανομή).

Στο πέμπτο κεφάλαιο γίνεται αναφορά στην έννοια του συνδυασμού υποδειγμάτων ταξινόμησης (fusion classifier) και καταγράφονται διάφοροι τρόποι μέτρησης της διαφορετικότητας (diversity) των υποδειγμάτων ταξινόμησης. Επίσης, γίνεται αναφορά σε διάφορες τεχνικές συνδυασμού υποδειγμάτων ταξινόμησης, ως μέσο αντιμετώπισης και βελτίωσης του προβλήματος της ταξινόμησης.

Στο έκτο κεφάλαιο πραγματοποιείται ένας εκτενής πειραματικός σχεδιασμός με σκοπό της διερεύνηση της αποτελεσματικότητας της απόδοσης των εξεταζόμενων μεθοδολογικών προσεγγίσεων ταξινόμησης και συνδυασμού αυτών για την αντιμετώπιση του προβλήματος ταξινόμησης με τη εκτίμηση και υπολογισμό της ακρίβειας ταξινόμησης, του μέτρου F, της καμπύλης ROC και των διαστημάτων εμπιστοσύνης. Επίσης, πραγματοποιείται ανάλυση των αποτελεσμάτων και τονίζονται τα βασικά συμπεράσματα του πειραματικού σχεδιασμού που πραγματοποιήθηκε.

Τέλος, στο έβδομο κεφάλαιο παρουσιάζονται τα βασικά συμπεράσματα που επιτεύχθηκαν από την έρευνα που πραγματοποιήθηκε και προτείνονται μελλοντικές ερευνητικές κατευθύνσεις, οι οποίες θα συμβάλλουν στην καλύτερη αντιμετώπιση του προβλήματος της ταξινόμησης, αλλά και στην περαιτέρω διερεύνηση των ιδιαιτεροτήτων, ομοιοτήτων και διαφορών που χαρακτηρίζουν τις εξεταζόμενες τεχνικές ταξινόμησης και το συνδυασμό (fusion) αυτών.

Κ ε φ ά λ α ι ο 2°

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

2.1 Εισαγωγή

Η μηχανική μάθηση (machine learning) είναι ένα υποπεδίο της τεχνητής νοημοσύνης που ασχολείται με τη σχεδίαση αυτόματων διαδικασιών ικανών να μαθαίνουν από παραδείγματα (Καλαπανίδας, 2003). Η εκμάθηση από παραδείγματα αφορά στην παραγωγή μιας λογικής περιγραφής των αναγκαίων και ικανών συνθηκών που αντιστοιχούν σε μια ομάδα αντικειμένων. Μία σημαντική φροντίδα είναι η εύρεση των αναγκαίων συμβιβασμών ανάμεσα στην πολυπλοκότητα των κανόνων και στην προσαρμογή των δεδομένων, έτσι ώστε να αποφευχθεί η υπερβολική προσαρμογή (over fitting) και να καταστεί δυνατή η ερμηνεία των αποτελεσμάτων. Μεταξύ των διαφόρων μορφών συστημάτων μάθησης (Kodratoff και Michalski, 1990), η επαγωγική μάθηση μέσω παραδειγμάτων (inductive learning) έχει γνωρίσει τη μεγαλύτερη διάδοση.

Σε αντίθεση με όλες τις προαναφερθείσες προσεγγίσεις ταξινόμησης (Δούμπος και Ζοπουνίδης, 2001), η χρησιμοποίηση της επαγωγικής μάθησης μέσω παραδειγμάτων εισάγει μια εντελώς διαφορετική αντιμετώπιση του προβλήματος, ιδιαίτερα όσον αφορά τη μορφή των υποδειγμάτων ταξινόμησης που αναπτύσσονται. Αναλυτικότερα, κατά την αντιμετώπιση προβλημάτων λήψης αποφάσεων όπου απαιτείται η ταξινόμηση των εξεταζόμενων εναλλακτικών δραστηριοτήτων, σκοπός της επαγωγικής μάθησης είναι η εξαγωγή γνώσης από ένα σύνολο παραδειγμάτων των αποφάσεων που λαμβάνονται και η αναπαράσταση/ οργάνωση της εξαγχθείσας γνώσης σε ένα σύνολο κανόνων αποφάσεων της μορφής:

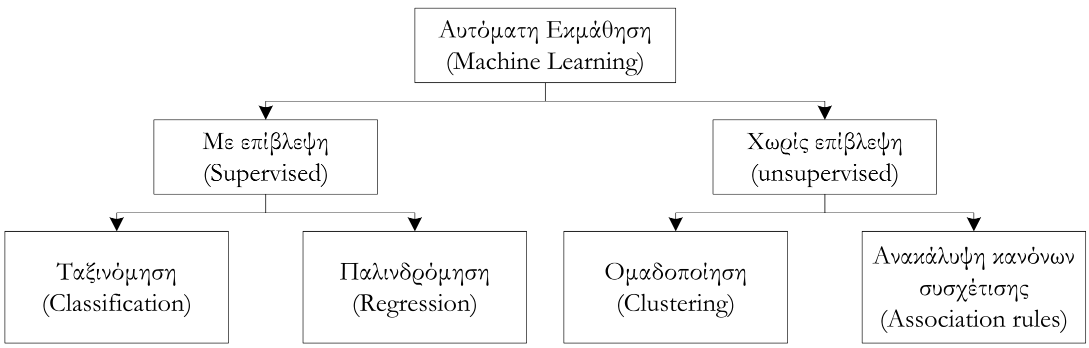
ΕΑΝ *στοιχειώδεις συνθήκες* ΤΟΤΕ *συμπέρασμα*

Το πρώτο τμήμα του κανόνα περιλαμβάνει τις απαραίτητες και αναγκαίες συνθήκες που πρέπει να πληροί μια εναλλακτική δραστηριότητα προκειμένου να ταξινομηθεί στην κατηγορία που υποδεικνύει το συμπέρασμα του κανόνα. Οι στοιχειώδεις συνθήκες συνδέονται με το λογικό τελεστή «ΚΑΙ».

Μία από τις πλέον γνωστές τεχνικές στον χώρο της μηχανικής μάθησης για την ανάπτυξη ενός συνόλου τέτοιων κανόνων απόφασης είναι ο αλγόριθμος C4.5, ο οποίος αναπτύχθηκε από τον Quinlan (1993), ως εξέλιξη του γνωστού αλγορίθμου ID3 (Quinlan, 1983, 1986). Βασικά πλεονεκτήματα του αλγορίθμου C4.5 έναντι του προκατόχου του ID3 αποτελούν η δυνατότητα διαχείρισης ποσοτικών κριτηρίων, η δυνατότητα διαχείρισης δεδομένων με ελλιπή στοιχεία και η αποφυγή της μεγάλης προσαρμογής στα δεδομένα του δείγματος εκμάθησης (over fitting).

2.2 Κατηγορίες μηχανικής μάθησης

Η μηχανική μάθηση περιλαμβάνει μία ιεραρχία μεθοδολογιών, όπως αυτές απεικονίζονται στο σχήμα 2.1.



Σχήμα 2. 1: Οι κατηγορίες της μηχανικής μάθησης

Στο πρώτο επίπεδο χωρίζεται σε δύο κλάδους: την εκμάθηση με επίβλεψη και την εκμάθηση χωρίς επίβλεψη. Στην πρώτη περίπτωση σε κάθε παράδειγμα είναι γνωστή η τιμή ή η κατηγορία στην οποία ανήκει μία παρατήρηση y, ενώ στη δεύτερη η κατηγορία είναι άγνωστη και η μηχανική μάθηση επιχειρεί να την προσεγγίσει ομαδοποιώντας τα παραδείγματα. Η πρόβλεψη της αυριανής μέσης θερμοκρασίας με δεδομένες μετρήσεις της μέσης ημερήσιας θερμοκρασίας είναι για παράδειγμα ένα πρόβλημα με επίβλεψη, γιατί στα παραδείγματα του παρελθόντος υπάρχει ακριβής τιμή της μετρούμενης θερμοκρασίας. Αντίθετα, το πρόβλημα της εύρεσης της κατηγορίας ενός νέου προϊόντος σε σχέση με τα υπόλοιπα προϊόντα σε ένα πολυκατάστημα με βάση τα χαρακτηριστικά του και τις πωλήσεις του είναι ένα συνηθισμένο παράδειγμα προβλήματος χωρίς επίβλεψη.

Με μαθηματικούς όρους, έστω ένα δείγμα δεδομένων εκπαίδευσης (X, Y). Στα προβλήματα με επίβλεψη αυτό αποτελείται από n ζευγάρια παρατηρήσεων $(x_1, y_1),\ldots,(x_n, y_n)$, όπου $x_i \in R^p$ και $y_i \in R$ με $i = 1,\ldots,n$. Στα προβλήματα χωρίς πρόβλεψη δεν υπάρχει η μεταβλητή Y και το σύνολο (X, Y) αποτελείται από n παρατηρήσεις $x_1,\ldots,x_n$, όπου $x_i \in R^p$. Η μεταβλητή Y ονομάζεται εξαρτημένη μεταβλητή και οι μεταβλητές X_i ονομάζονται ανεξάρτητες μεταβλητές ή αλλιώς ιδιότητες του προβλήματος.

Η εκμάθηση με επίβλεψη αφορά σε δύο περαιτέρω κατηγορίες προβλημάτων, την παλινδρόμηση και την ταξινόμηση. Στην πρώτη κατηγορία η μεταβλητή y είναι μία συνεχής μεταβλητή, ενώ στη δεύτερη μία διακριτή κατηγορική μεταβλητή. Το παράδειγμα της θερμοκρασίας είναι πρόβλημα παλινδρόμησης όταν η θερμοκρασία μετράται σαν πραγματική συνεχής τιμή, ενώ είναι πρόβλημα ταξινόμησης αν μετράται σαν θερμοκρασιακή ζώνη (π.χ. κρύο/ ζέστη).

Τα προβλήματα χωρίς επίβλεψη χωρίζονται σε ομαδοποίηση και ανακάλυψη κανόνων συσχέτισης. Η μεθοδολογία της ομαδοποίησης επιχειρεί να ξεχωρίσει τα δεδομένα σε άτυπες ομάδες, με βάση τις αποστάσεις ανάμεσα στις ιδιότητες των παρατηρήσεων. Το παράδειγμα της κατάταξης ενός νέου προϊόντος σε ομάδα που αναφέρθηκε προηγουμένως, ανήκει στην πρώτη κατηγορία. Η ανακάλυψη κανόνων συσχέτισης από την άλλη διερευνά τα δεδομένα για να ανακαλύψει κανόνες που διέπουν τις ιδιότητες των παρατηρήσεων. Αν για παράδειγμα υπήρχε ένα σύνολο δεδομένων που περιελάμβανε τις αγορές προϊόντων σε ένα κατάστημα από κάθε πελάτη, τότε ένας αλγόριθμος ανακάλυψης κανόνων συσχέτισης θα συμπέρανε κανόνες του τύπου «το 30% των πελατών που αγόρασαν ψυγεία, αγόρασαν και ηλεκτρική σκούπα» ή «το 68% των πελατών που αγόρασαν σοκολατάκια ΚΑΙ αγόρασαν επίσης σαμπάνια, πλήρωσαν με πιστωτική κάρτα».

Η πλειοψηφία των προβλημάτων ανήκει στην κατηγορία των προβλημάτων εκμάθησης με επίβλεψη.

2.3 Η προετοιμασία των δεδομένων στην μηχανική μάθηση

Μία σημαντική υπόθεση για την επιτυχία μιας εφαρμογής μηχανικής μάθησης είναι η σωστή εκμετάλλευση των παρεχόμενων δεδομένων. Όλοι οι αλγόριθμοι της μηχανικής μάθησης απαιτούν από αυτά να είναι εντοπισμένα σε ένα αρχείο εισόδου ακολουθώντας συγκεκριμένη τυποποίηση. Το ζητούμενο είναι οι παρατηρήσεις να αντιπροσωπεύονται από τις γραμμές του αρχείου και τα χαρακτηριστικά από τις στήλες. Πολλές δυσκολίες προκύπτουν προς την κατεύθυνση της ενοποίησης των δεδομένων. Αυτές μπορεί να υπάρχουν σε διάφορες μορφές και σε διαφορετικά αποθηκευτικά μέσα. Μπορεί να είναι κατανεμημένες ή συγκεντρωμένες, σε αρχεία κειμένου, σε βάσεις δεδομένων, σε λογιστικά φύλλα, ταξινομημένες ή αταξινόμητες, χρονικά διατεταγμένες ή χωρικά διεσπαρμένες.

2.4 Μέτρα εκτίμησης των αλγορίθμων της μηχανικής μάθησης

Ο έλεγχος της ποιότητας της προσαρμογής των δεδομένων στους παραγόμενους κανόνες γίνεται με τον υπολογισμό και την εκτίμηση του σφάλματος. Υπάρχουν πολλά μέτρα εκτίμησης του σφάλματος στη βιβλιογραφία. Μερικά από αυτά παρουσιάζονται παρακάτω τα πιο διαδομένα και ευρείας χρήσης γενικότερα. Τα πρώτα μέτρα αναφέρονται σε προβλήματα παλινδρόμησης και ταξινόμησης, ενώ τα τελευταία μόνο σε προβλήματα τύπου ταξινόμησης.

2.4.1 Σφάλμα πρόβλεψης

Το σφάλμα πρόβλεψης e_i για την παρατήρηση I είναι η διαφορά ανάμεσα στην πραγματική μετρήσιμη τιμή της μεταβλητής y' και στην εκτιμώμενη από τον αλγόριθμο τιμή προσέγγισης y:

$$e_i = y' - y$$

Το αντίστοιχο εκατοστιαίο σφάλμα πρόβλεψης είναι ο λόγος του σφάλματος προς την πραγματική τιμή:

$$p_{e_i} = \frac{e_i}{y}$$

2.4.2 Μέση απόλυτη απόκλιση

Η μέση απόλυτη απόκλιση (mean absolute deviation – MAD) για ένα σύνολο δεδομένων είναι η μέση τιμή για την απόλυτη τιμή του σφάλματος πρόβλεψης για όλες τις παρατηρήσεις του συνόλου δεδομένων:

$$MAD = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

Όπου n το πλήθος των παρατηρήσεων στο σύνολο των δεδομένων.

2.4.3 Μέσο απόλυτο εκατοστιαίο σφάλμα

Το μέσο απόλυτο εκατοστιαίο σφάλμα (mean absolute percentage error – MAPE) για ένα σύνολο δεδομένων είναι η μέση τιμή του απόλυτου εκατοστιαίου σφάλματος πρόβλεψης για όλες τις παρατηρήσεις του συνόλου δεδομένων:

$$MAPE = \frac{\sum_{i=1}^{n} |p_{e_i}|}{n}$$

2.4.4 Μέσο τετραγωνισμένο σφάλμα

Το μέσο τετραγωνισμένο σφάλμα (mean squared error – MSE) είναι η μέση τιμή του σφάλματος πρόβλεψης για όλες τις παρατηρήσεις του συνόλου δεδομένων:

$$MSE = \frac{\sum_{i=1}^{n} |e_i^{\,2}|}{n}$$

2.4.5 Μέσο τετραγωνικό σφάλμα

Το μέσο τετραγωνικό σφάλμα είναι η ρίζα του αντίστοιχου μέσου τετραγωνισμένου σφάλματος MSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left| e_i^{\,2} \right|}{n}}$$

2.4.6 Τυπικό σφάλμα ενός διαστήματος εμπιστοσύνης

Στα προβλήματα ταξινόμησης όπου η εξαρτημένη μεταβλητή Υ είναι κατηγορική, το σφάλμα πρόβλεψης είναι πάντα ακέραιος αριθμός.

Σε αυτά τα προβλήματα υπολογίζεται το ποσοστό σφάλματος e' ως ο λόγος του αριθμού των παρατηρήσεων που ταξινομήθηκαν εσφαλμένα από τον αλγόριθμο προς το πλήθος αυτών:

$$e' = \frac{l_{\{y \neq y'\}}}{n}$$

Αν e το πραγματικό ποσοστό σφάλματος του υποδείγματος, τότε το διάστημα εμπιστοσύνης 1-α είναι:

$$e' - z_{\alpha/2} SE \leq e \leq e' + z_{\alpha/2} SE$$

Όπου SE (Standard Error) το τυπικό σφάλμα του διαστήματος εμπιστοσύνης.

2.4.7 Ειδικά μέτρα εκτίμησης προβλημάτων ταξινόμησης

Ειδικά στα προβλήματα ταξινόμησης όπου η μεταβλητή εξόδου μπορεί να είναι ποιοτική, να δέχεται δηλαδή διακριτές τιμές, κατασκευάζεται ένας πίνακας σύγχυσης (confusion matrix) για να υποδείξει τα σφάλματα ταξινόμησης του εκάστοτε αλγορίθμου ανά τιμή της εξαρτημένης μεταβλητής. Αν δηλαδή Υ είναι η πραγματική μεταβλητή εξόδου με δυνατές τιμές στο σύνολο $\{y_1, y_2, ..., y_k\}$ και Y' η εκτίμηση αυτής που κατασκευάζει το υπόδειγμα ταξινόμησης με δυνατές τιμές στο σύνολο $\{y_1', y_2', ..., y_k'\}$, n ο αριθμός των διακριτών τιμών της Υ και της Y', τότε ο confusion matrix είναι η μήτρα:

$$\begin{vmatrix} sum(y_1' \mid Y = y_1) & \dots & sum(y_k' \mid Y = y_1) \\ \dots & \dots & \dots \\ sum(y_1' \mid Y = y_k) & \dots & sum(y_k' \mid Y = y_k) \end{vmatrix}$$

Οι πραγματικά θετικές παρατηρήσεις (True Positive ή Sensitivity) για την κλάση $i \in [1,\ldots,k]$ ορίζονται ως ο αριθμός των παρατηρήσεων που ταξινομήθηκαν σωστά από τον αλγόριθμο ταξινόμησης, δηλαδή:

$$TP = sum(y_i' \mid Y = y_i)$$

Οι εσφαλμένα θετικές παρατηρήσεις (False Positive ή 1-Spesificity) για την κλάση $i \in [1,\ldots,k]$ είναι ο αριθμός των παρατηρήσεων που εκτιμήθηκαν ότι ανήκουν στην κλάση i, ενώ δεν ανήκουν σε αυτή:

$$FP = sum(y_i' \mid Y \neq y_i)$$

Η ακρίβεια (Precision) για την κλάση $i \in [1,\ldots,k]$ ορίζεται ως ο λόγος των πραγματικά θετικών ως προς το άθροισμα πραγματικά θετικών και εσφαλμένα θετικών παρατηρήσεων:

$$\mathrm{Pr}\,ecision = \frac{sum(y_i' \mid Y = y_i)}{sum(y_i' \mid Y = y_i) + sum(y_i' \mid Y \neq y_i)}$$

Η επανάκληση (Recall ή Accuracy Ratio) για την κλάση $i \in [1,\ldots,k]$ ορίζεται ως ο λόγος των πραγματικά θετικών ως προς τον αριθμό των συνολικά θετικών παρατηρήσεων:

$$\mathrm{Re}\,call = \frac{sum(y_i' \mid Y = y_i)}{sum(y_i')}$$

Γενικά υπάρχει μία συμφωνία ανάμεσα στις υψηλές τιμές του μέτρου ακρίβειας και στις υψηλές τιμές του μέτρου επανάκλησης.

Το μέτρο F είναι ο συνδυασμός της ακρίβειας και της επανάκλησης. Ορίζεται ως ο αρμονικός μέσος αυτών των δύο και θεωρείται ένα καλό συγκεντρωτικό μέτρο για την ικανότητα ταξινόμησης της κλάσης i:

$$F = \frac{2 \cdot \mathrm{Recall} \cdot \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}$$

2.5 Καμπύλη ROC (Receiver Operating Characteristic Curve)

Η ανάπτυξη των υποδειγμάτων ταξινόμησης έχει ένα και μοναδικό σκοπό, την όσο το δυνατόν καλύτερη και ορθότερη ταξινόμηση των εναλλακτικών στις διάφορες κατηγορίες. Στην ταξινόμηση όμως αυτή, πάντα υπεισέρχεται κάποιο σφάλμα, διότι δεν υπάρχει ή έστω δεν έχει βρεθεί ακόμα το ιδανικό υπόδειγμα το οποίο θα πραγματοποιεί μια 100% σωστή ταξινόμηση. Πάντα υπάρχουν κάποιες εναλλακτικές οι οποίες θα καταχωρηθούν σε λάθος κατηγορίες, από αυτές στις οποίες πραγματικά ανήκουν. Για το λόγο αυτό έχουν αναπτυχθεί διάφορες διαδικασίες ελέγχου με τα οποία μπορεί να μετρηθεί η αποτελεσματικότητα των μεθόδων ταξινόμησης και υπό αυτή την έννοια να εξαχθούν συμπεράσματα για την αποτελεσματικότητά τους.

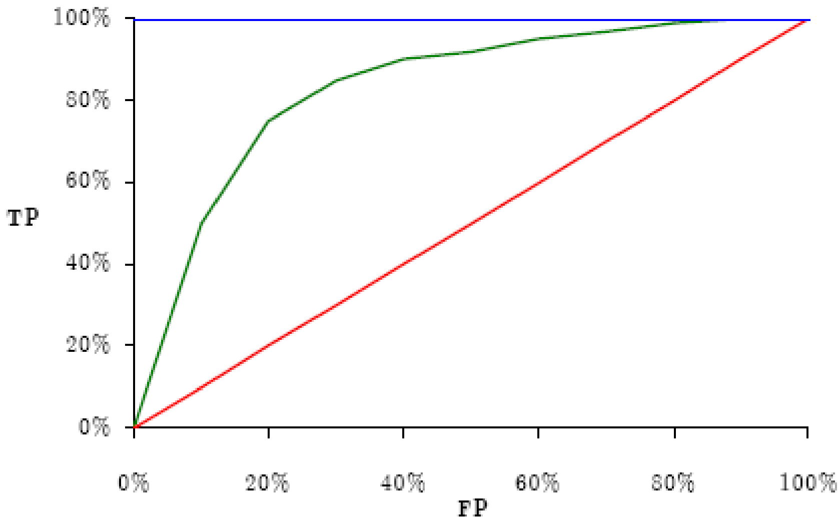
Ο πιο απλός τρόπος για την αξιολόγηση της αποτελεσματικότητας των υποδειγμάτων ταξινόμησης, είναι ο υπολογισμός του σφάλματος ταξινόμησης (Error Rate or Misclassification Rate). Το σφάλμα ταξινόμησης ορίζεται πολύ απλά ως το πλήθος των εναλλακτικών οι οποίες ταξινομήθηκαν λανθασμένα από το σύνολο των εναλλακτικών. Όμως, παρά το ότι αποτελεί το πιο διαδεδομένο μέτρο για την αξιολόγηση της αποτελεσματικότητας των υποδειγμάτων ταξινόμησης, παρουσιάζει ορισμένα προβλήματα κατά την εφαρμογή του. Αυτά οφείλονται στο γεγονός ότι, λόγω έλλειψης πληθώρας δεδομένων, τα τελευταία χρησιμοποιούνται τόσο για την κατασκευή του μοντέλου, δηλαδή τον υπολογισμό του σφάλματος ταξινόμησης, όσο και για την αξιολόγηση των υποδειγμάτων ταξινόμησης. Το γεγονός αυτό οδήγησε στην ανάπτυξη πιο επιδέξιων μεθόδων για τον υπολογισμό του σφάλματος ταξινόμησης, όπως είναι οι μέθοδοι jacknife και bootstrap. Αναλυτική αναφορά για αυτές γίνεται από τον Hand (1997). Μπορεί να ειπωθεί ότι παρότι το σφάλμα είναι το πιο διαδεδομένο κριτήριο αποτίμησης των υποδειγμάτων ταξινόμησης, αυτό δε σημαίνει ότι είναι και τα πιο κατάλληλο κριτήριο.

Μία ιδιαίτερα διαδεδομένη διαδικασία ελέγχου αξιολόγησης υποδειγμάτων ταξινόμησης βασίζεται στην κατασκευή του διαγράμματος ROC (Receiving Operating Characteristic, Fawcett, 2003). Η κατασκευή του διαγράμματος ROC υλοποιείται στην περίπτωση δύο κατηγοριών, αλλά είναι δυνατή η επέκταση της ανάλυσης και στην περίπτωση περισσότερων κατηγοριών. Δεδομένων λοιπόν, δύο κατηγοριών C_1 και C_2, το διάγραμμα ROC είναι ένα διάγραμμα δύο αξόνων (Σχήμα 2.2). Ο οριζόντιος άξονας δείχνει το ποσοστό των εναλλακτικών που ταξινομήθηκαν από τον αποφασίζοντα στην κατηγορία C_2 και το υπόδειγμα τις ταξινομεί στην κατηγορία C_1 (False Positives, FP), ενώ ο κάθετος άξονας δείχνει το ποσοστό των εναλλακτικών που ταξινομήθηκαν από τον αποφασίζοντα στην κατηγορία C_1 και το υπόδειγμα τις ταξινομεί επίσης στην κατηγορία C_1 (True Positives, TP). Υπάρχουν διάφορες ονομασίες, ανάλογα και με το πρόβλημα που αντιμετωπίζεται κάθε φορά, για τους δύο άξονες του διαγράμματος ROC. Έτσι, τα TP και FP μπορεί να ονομάζονται αντίστοιχα, Sensitivity και 1-Spesificity (κυρίως στο χώρο της επιδημιολογίας), False Negative Rate και True Negative Rate, Positive Predicted Value και Negative Predicted Value, κ.ά.

Για την κατασκευή του διαγράμματος θεωρείται ότι το εξεταζόμενο υπόδειγμα V αποδίδει μια βαθμολογία $V(x_i)$ σε κάθε εναλλακτική x_i. Βάσει αυτής της βαθμολογίας η ταξινόμηση των εναλλακτικών πραγματοποιείται εισάγοντας ένα σημείο διαχωρισμού t, έτσι ώστε:

$$V(x_i) \geq t \Rightarrow x_i \in C_1$$
$$V(x_i) < t \Rightarrow x_i \in C_2$$

Το διάγραμμα ROC καταγράφει τα αποτελέσματα του υποδείγματος συναρτήσει του σημείου διαχωρισμού t. Για τις διαφορετικές τιμές του σημείου διαχωρισμού καταγράφεται η συμπεριφορά του υποδείγματος στα μεγέθη TP και FP, τα οποία στη συνέχεια τοποθετούνται σε ένα γράφημα δύο αξόνων διαμορφώνοντας έτσι το διάγραμμα ROC του υποδείγματος. Σε ένα εξαιρετικό (άριστο) υπόδειγμα οι βαθμολογίες των εναλλακτικών της κατηγορίας C_1 είναι υψηλότερες από τις βαθμολογίες των εναλλακτικών της κατηγορίας C_2 (μπλε γραμμή στο σχήμα 2.2). Δηλαδή, μια ιδανική καμπύλη ROC θα ακολουθούσε τους δύο κάθετους άξονες, και θα ταξινομούσε το 100% των θετικών περιπτώσεων (positive values) στην ίδια κατηγορία και το 0% των αρνητικών περιπτώσεων (negative values) ως θετικές για κάποια συγκεκριμένη τιμή του ορίου διαχωρισμού t. Αντίθετα, σε ένα ανεπαρκές υπόδειγμα δε θα παρατηρείται καμία σχέση μεταξύ των βαθμολογιών των εναλλακτικών και της ταξινόμησής τους, οπότε η καμπύλη ROC στην περίπτωση αυτή είναι μια ευθεία γραμμή που ξεκινά από την αρχή των αξόνων (κόκκινη γραμμή στο σχήμα 2.2) και έχει κλίση 45^o. Δηλαδή είναι η ευθεία TP=FP (κόκκινη γραμμή στο σχήμα 2.2). Ένα ικανοποιητικό υπόδειγμα βρίσκεται στο ενδιάμεσο των δύο περιπτώσεων (πράσινη γραμμή στο σχήμα 2.2).



Σχήμα 2. 2: Γενική περιγραφή της καμπύλης ROC

Η αποτελεσματικότητα του υποδείγματος βάσει του διαγράμματος ROC εξετάζεται από το εμβαδόν της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC). Προφανώς, το εμβαδόν αυτό για ένα άριστο υπόδειγμα είναι ίσο με τη μονάδα, ενώ για ένα ανεπαρκές υπόδειγμα είναι μικρότερο από 0.5. Στη γενικότερη και πιο συνήθη περίπτωση, το εμβαδόν που περικλείεται κάτω από την καμπύλη ROC κυμαίνεται μεταξύ 0.5 και 1. Συνάγεται ότι υψηλότερες τιμές του AUC υποδεικνύουν υψηλότερη αποτελεσματικότητα, για το υπόδειγμα ταξινόμησης. Στην περίπτωση όπου AUC<0.5, τότε το υπόδειγμα είναι τελείως ανεπαρκές και χρειάζεται να γίνουν κάποιες αναθεωρήσεις στο μαθηματικό υπόβαθρο του υποδείγματος, ώστε η καμπύλη ROC που θα προκύψει να βρίσκεται πάνω από την ευθεία TP=FP ούτως ώστε να προκύψει AUC>0.5.

Για το υπολογισμό του AUC, έστω ότι εξετάζεται ένα υπόδειγμα το οποίο παρέχει b διακριτές βαθμολογίες $s_1 < s_2 < \ldots < s_b$ στις υπό εξέταση εναλλακτικές. Η πιθανότητα μια εναλλακτική της κατηγορίας C_1 να λάβει το πολύ τη βαθμολογία s_i συμβολίζεται ως $p_1^{\ i}$ (προφανώς $p_1^{\ b} = 1$). Η πιθανότητα αυτή ουσιαστικά αντιστοιχεί στο ποσοστό των εναλλακτικών που ταξινομήθηκαν από τον αποφασίζοντα στην κατηγορία C_1 και λαμβάνουν από το υπόδειγμα βαθμολογία μικρότερη από ή ίση με s_i. Αντίστοιχα, ως $p_2^{\ i}$ συμβολίζεται το ποσοστό (πιθανότητα) των εναλλακτικών της κατηγορία C_2 που λαμβάνουν βαθμολογία μικρότερη από ή ίση με s_i (προφανώς $p_2^{\ b} = 1$). Βάσει αυτών των στοιχείων μπορεί να υπολογιστεί το εμβαδόν κάτω από τη καμπύλη ROC του υποδείγματος ως εξής:

$$AUC = \sum_{i=1}^{b} \frac{1}{2}\left(p_1^{\ i} - p_1^{\ i-1}\right)\left(p_2^{\ i} + p_2^{\ i-1}\right)$$

Υπολογιζόμενο κατά τον τρόπο αυτό, το εμβαδόν της περιοχής κάτω από την καμπύλη ROC αναπαριστά την πιθανότητα μια εναλλακτική της κατηγορίας C_1 να λάβει υψηλότερη βαθμολογία από μια εναλλακτική της κατηγορία C_2. Όσο υψηλότερη είναι αυτή η πιθανότητα τόσο καλύτερο θεωρείται το μοντέλο ταξινόμησης.

Ένας εναλλακτικός τρόπος για την μέτρηση της αποτελεσματικότητας των υποδειγμάτων ταξινόμησης, ο οποίος όμως βασίζεται σε πολύ μεγάλο ποσοστό στον υπολογισμό του AUC όπως αναφέρθηκε παραπάνω, αποτελεί ο δείκτης ακρίβειας (Accuracy Ratio). Ο δείκτης αυτός ονομάζεται εναλλακτικά και συντελεστής Gini (Gini Coefficient), ενώ στη διεθνή βιβλιογραφία κατά καιρούς έχει λάβει και πολλές άλλες ονομασίες. Ο δείκτης ακρίβειας ισούται δύο φορές με το εμβαδόν που περιέχεται ανάμεσα στην καμπύλη ROC και στη διαγώνιο ευθεία TP=FP. Δηλαδή, ορίζεται ως Accuracy Ratio=2(AUC-0.5), όπου το AUC υπολογίζεται σύμφωνα με τα παραπάνω, ενώ το 0.5 είναι το εμβαδόν που περικλείεται κάτω από τη διαγώνιο του 1ου τεταρτημορίου, δηλαδή την ευθεία TP=FP, όπως προαναφέρθηκε.

Γενικότερα, είναι προφανές και στην περίπτωση του AUC, ότι ο δείκτης ακρίβειας για ένα άριστο υπόδειγμα είναι ίσος με τη μονάδα, αφού τότε AUC=1 και Accuracy Ratio=1, ενώ για ένα ανεπαρκές υπόδειγμα ο δείκτης ακρίβειας είναι ίσος με το μηδέν, αφού τότε AUC=0.5 και Accuracy Ratio=0. Επομένως, συμπεραίνεται ότι ο δείκτης ακρίβειας για ένα ικανοποιητικό υπόδειγμα, κυμαίνεται μεταξύ 0 και 1, ενώ όσο μεγαλύτερη είναι η τιμή του δείκτη ακρίβειας, τόσο καλύτερο είναι το υπόδειγμα ταξινόμησης.

Στην πειραματική μελέτη που πραγματοποιήθηκε για την παρούσα διατριβή και η οποία αναλύεται στο 6ο Κεφάλαιο, η αξιολόγηση της αποτελεσματικότητας των υποδειγμάτων ταξινόμησης έγινε με την χρήση του δείκτη ακρίβειας.

Κεφάλαιο 3ο

ΕΚΤΙΜΗΣΗ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΠΡΟΒΛΗΜΑΤΩΝ ΤΑΞΙΝΟΜΗΣΗΣ

3.1 Εισαγωγή – Ιστορική Αναδρομή

Το πρόβλημα της ταξινόμησης γενικά φορά την τοποθέτηση μιας σειράς εναλλακτικών λύσεων (αντικειμένων ή παρατηρήσεων) τα οποία περιγράφονται βάσει ενός συνόλου κριτηρίων (χαρακτηριστικά) σε προκαθορισμένες ομοιογενείς κατηγορίες. Οι ομάδες αυτές δύνανται να είναι προδιατεταγμένες από τις καλύτερες προς τις χειρότερες, χωρίς βεβαία αυτό να είναι απαραίτητο.

Το 1936, ο Fisher ήταν ο πρώτος ο οποίος ανέπτυξε ένα ολοκληρωμένο πλαίσιο για την αντιμετώπιση προβλημάτων ταξινόμησης λαμβάνοντας υπόψη την πολυδιάστατη φύση τους. Η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis-LDA) που πρότεινε ο Fisher χρησιμοποιήθηκε για δεκαετίες ως η κύρια τεχνική ταξινόμησης και χρησιμοποιείται ακόμα και σήμερα τουλάχιστον ως σημείο αναφοράς για την πραγματοποίηση συγκρίσεων με τις νέες μεθοδολογίες που διαρκώς αναπτύσσονται. Το 1947, ο Smith ανέπτυξε την τετραγωνική διακριτική ανάλυση (Quadratic Discriminant Analysis-QDA) με σκοπό να υπερκεράσει την υπόθεση ότι οι πίνακες διασποράς – συνδιασποράς μεταξύ των ομάδων είναι ίσοι, που περιόριζε την χρήση της LDA. Αργότερα αναπτύχθηκαν και άλλες πολυμεταβλητές στατιστικές μέθοδοι για την αντιμετώπιση προβλημάτων ταξινόμησης. Μεταξύ αυτών το λογιστικό (logit) και το κανονικό (probit) υπόδειγμα πιθανότητας είναι οι πιο διαδεδομένες τεχνικές, οι οποίες υπόκεινται στη βασική περιοριστική υπόθεση των LDA και QDA ότι οι μεταβλητές είναι κανονικά κατανεμημένες. Το 1977, ο Eisenbeis παρουσίασε μια αναλυτική κριτική σχετική με τα προβλήματα της διακριτικής ανάλυσης.

3.2 Άλλες τεχνικές ταξινόμησης

Οι συνεχείς εξελίξεις και σε άλλους επιστημονικούς κλάδους εκτός της στατιστικής, όπως στην επιχειρησιακή έρευνα και την τεχνητή νοημοσύνη, γρήγορα οδήγησαν τους ερευνητές στην αξιοποίηση των δυνατοτήτων που προσφέρουν οι κλάδοι αυτοί στην ανάπτυξη νέων τεχνικών ταξινόμησης. Μεταξύ των προσπαθειών οι οποίες έχουν πραγματοποιηθεί, μπορούν να αναφερθούν η χρήση των νευρωνικών δικτύων (neural networks), της μηχανικής μάθησης (machine learning), της θεωρίας των ασαφών συνόλων (fuzzy sets), της θεωρίας των προσεγγιστικών συνόλων (rough sets), καθώς και της πολυκριτήριας ανάλυσης (Multicriteria Decision Aid-MCDA). Βασικό χαρακτηριστικό και πλεονέκτημα της πολυκριτήριας ανάλυσης αποφάσεων αποτελεί το γεγονός ότι παρέχει ένα ευρύ σύνολο μη παραμετρικών μεθοδολογιών ταξινόμησης οι οποίες δεν υπόκεινται σε στατιστικές μεθόδους και περιορισμούς. Παρακάτω θα γίνει μια μικρή αναφορά στα πλεονεκτήματα και μειονεκτήματα μερικών προαναφερθέντων νέων τεχνικών ταξινόμησης.

Η θεωρία των προσεγγιστικών συνόλων (Rough sets theory) αποτελεί ένα σημαντικό εργαλείο για τη μελέτη προβλημάτων ταξινόμησης. Βασίζεται σε παραδείγματα προηγούμενων αποφάσεων με σκοπό την ανάπτυξη ενός συνόλου κανόνων απόφασης που να αντιπροσωπεύει τις προτιμήσεις του αποφασίζοντα με ένα κατανοητό τρόπο. Επιπλέον τα προσεγγιστικά σύνολα είναι ικανά να χειρίζονται ποσοτικά και ποιοτικά κριτήρια, όπως και μη μονότονες προτιμήσεις. Από την άλλη όμως αδυνατούν να λάβουν υπόψη τους συνεχείς μεταβλητές. Οι συνεχείς μεταβλητές μετατρέπονται σε διακριτές μέσω μιας διαδικασίας διακριτοποίησης, η οποία πρέπει να επιλεχθεί και να εφαρμοστεί έτσι, ώστε να αντιπροσωπεύει τις προτιμήσεις του αποφασίζοντα.

Η φιλοσοφία των έμπειρων συστημάτων έχει ελκύσει στο παρελθόν το ενδιαφέρον πολλών ερευνητών στο χώρο της διοίκησης. Τα βασικότερα πλεονεκτήματα τους αφορούν την ικανότητά τους να παρέχουν εκτιμήσεις χρησιμοποιώντας έναν κατανοητό τρόπο που βασίζεται στη γνώση ειδικών, όπως και στην ικανότητά τους να εξηγούν τις εκτιμήσεις τους. Παρά τα παραπάνω πλεονεκτήματα και τον αρχικό ενθουσιασμό γύρω από τα έμπειρα συστήματα, η εφαρμογή τους επέφερε κάποιους αρχικούς περιορισμούς. Τα μεγαλύτερα μειονεκτήματα που παρουσιάζονται κατά την εφαρμογή των έμπειρων συστημάτων στα προβλήματα ταξινόμησης αφορούν το μεγάλο χρόνο που απαιτείται για την άντληση γνώσης από τους ειδικούς, την απαιτούμενη ακρίβεια κατά την μοντελοποίηση των μεταβλητών απόφασης και την έλλειψη ευελιξίας σε αλλαγές του αποφασίζοντα ή του περιβάλλοντος της απόφασης (Tafti και Nikbakht, 1993 & Klein και Methlie, 1995).

Τα νευρωνικά δίκτυα έχουν κεντρίσει το ενδιαφέρον των ερευνητών την τελευταία κυρίως δεκαετία. Τα νευρωνικά δίκτυα προσομοιώνουν τη δομή του ανθρώπινου εγκεφάλου για να παρέχουν εκτιμήσεις. Όταν το νευρωνικό δίκτυο έχει εκπαιδευτεί, μπορεί να χρησιμοποιηθεί εύκολα και σε άλλες περιπτώσεις. Εκμεταλλευόμενα την παράλληλη επεξεργασία διαφόρων μονάδων (νευρώνες), τα νευρωνικά δίκτυα έχουν τη δυνατότητα να δώσουν εκτιμήσεις σε πραγματικό χρόνο, αλλά το πιο σημαντικό είναι ότι μπορούν να προσαρμοστούν σε πιθανές αλλαγές του αποφασίζοντα. Η μέχρι σήμερα κριτική των νευρωνικών δικτύων έχει επικεντρωθεί σε τρία σημεία: (i) ο καθορισμός των δομικών παραμέτρων (π.χ. αριθμός επιπέδων, αριθμός των μονάδων επεξεργασίας σε κάθε επίπεδο, κ. ά.) είναι περισσότερο υποκειμενικός και βασίζεται συνήθως σε ένα μεγάλο αριθμό πειραματικών εφαρμογών, (ii) η λειτουργία τους, η οποία έχει χαρακτηριστεί ως «μαύρο κουτί», δεν παρέχει στον αποφασίζοντα πληροφορίες σχετικά με την τεκμηρίωση των αποτελεσμάτων, και (iii) συνήθως απαιτείται ένα σημαντικά μεγάλο χρονικό διάστημα για την εκπαίδευση του νευρωνικού δικτύου (Altman et al., 1994).

Οι πολυκριτήριες μέθοδοι αποφάσεων (Multicriteria Decision Aid Methods-MCDA) κατέχουν σημαντική θέση στη μελέτη προβλημάτων ταξινόμησης. Οι πολυκριτήριες μέθοδοι δεν υπόκεινται σε στατιστικούς περιορισμούς, αλληλεπιδρούν με τον αποφασίζοντα κατά την ανάλυση των προβλημάτων, μπορούν να χειριστούν κριτήρια ποιοτικής φύσεως και προσαρμόζονται εύκολα παίρνοντας υπόψη τη δυναμική φύση του συστήματος και τις προτιμήσεις του αποφασίζοντα.

Οι πολυκριτήριες προσεγγίσεις υλοποιούνται στα πολυκριτήρια συστήματα υποστήριξης αποφάσεων (Multicriteria Decision Support Systems-MCDSSs), τα οποία αποτελούν ένα σημαντικό εργαλείο για την ταξινόμηση. Τα πολυκριτήρια συστήματα υποστήριξης αποφάσεων παρέχουν όλα τα απαραίτητα μέσα για την εφαρμογή διαφόρων τεχνικών που έχουν σκοπό την υποστήριξη των ατόμων στη λήψη των αποφάσεων σε πραγματικό χρόνο. Η αλληλεπιδραστική δομή και λειτουργία των πολυκριτήριων συστημάτων υποστήριξης αποφάσεων επιτρέπει το συνδυασμό τεχνικών διαχείρισης βάσεων δεδομένων με τις δυνατότητες των πολυκριτήριων μεθόδων, γεγονός που τα καθιστά ευέλικτα σε αλλαγές του περιβάλλοντος λήψης αποφάσεων, καθώς και στην πολιτική και στις προτιμήσεις του εκάστοτε αποφασίζοντα. Η λειτουργία τέτοιων μεθόδων παρουσιάζεται στη σχετική εργασία των Zopounidis and Doumpos (1998).

Μια διαφορετική προσέγγιση στην αντιμετώπιση του προβλήματος της βελτιστοποίησης πολλαπλών στόχων γίνεται μέσω του προγραμματισμού στόχων. Σε αντίθεση με τον πολυκριτήριο γραμμικό προγραμματισμό, οι στόχοι εκφράζονται ως προκαθορισμένες από τον αποφασίζοντα τιμές οι οποίες θα πρέπει να επιτευχθούν. Στη συνέχεια, με την βοήθεια τεχνικών γραμμικού προγραμματισμού βελτιστοποιούνται (ελαχιστοποιούνται ή μεγιστοποιούνται) οι αποκλίσεις από τις τιμές αυτές. Ο αποφασίζων είναι σε θέση να καθορίσει ορισμένα βάρη τα οποία αναπαριστούν τις προτεραιότητες που δίνουν στην επίτευξη των προκαθορισμένων στόχων. Το γεγονός αυτό παρέχει ευελιξία στην ανάπτυξη ικανοποιητικών λύσεων ανάλογα με την πολιτική και τις προτιμήσεις του αποφασίζοντα. Με τη χρήση του μοντέλου προγραμματισμού στόχων επιτυγχάνεται η ανάλυση των ταυτόχρονων μεταβολών που λαμβάνουν χώρα στις παραμέτρους του μοντέλου. Η ανάλυση αυτή είναι ιδιαίτερα χρήσιμη ιδίως όταν πρόκειται για παραμέτρους που αναπαριστούν αντικρουόμενους στόχους. Έτσι δίνεται η δυνατότητα να τροποποιηθούν αλληλεπιδραστικά οι παράμετροι των στόχων προκειμένου να βελτιστοποιηθούν οι αρχικοί στόχοι. Παρακάτω θα γίνει μια εκτενέστερη ανάλυση των τεχνικών προγραμματισμού στόχων και των μεθόδων που τη διέπουν.

3.3 Προτεινόμενες τεχνικές ταξινόμησης

Οι τεχνικές που χρησιμοποιήθηκαν στην παρούσα έρευνα είναι αντιπροσωπευτικές των διαθέσιμων προσεγγίσεων. Ειδικότερα, εξετάζονται τόσο γνωστές στατιστικές τεχνικές, που αποτελούν τον παραδοσιακό τρόπο ανάπτυξης υποδειγμάτων ταξινόμησης, όσο και διαδεδομένες μη παραμετρικές προσεγγίσεις, οι οποίες έχουν εξελιχθεί ραγδαία τις τελευταίες δύο δεκαετίες ως αποτελεσματικά εργαλεία για την ανάπτυξη υποδειγμάτων ταξινόμησης. Παρακάτω παρουσιάζονται αναλυτικά οι εξεταζόμενες τεχνικές.

3.3.1 Κλασική Πιθανοθεωρητική Προσέγγιση του Naïve Bayes (NaïveBayes)

Από τους σημαντικότερους στόχους της πιθανοθεωρητικής προσέγγισης στη μάθηση είναι η εύρεση της πιο πιθανής υπόθεσης του χώρου υποθέσεων H, δεδομένου ενός σώματος εκπαίδευσης D και της όποιας γνώσης ενδεχομένως διατίθεται για τις πιθανότητες των διαφόρων υποθέσεων $h \in H$. Η πιθανότητα ισχύος μιας υπόθεσης h δεδομένου ενός συνόλου στιγμιότυπων D δίδεται από το νόμο Bayes:

$$\Pr(h/D) = \frac{\Pr(h)\Pr(D/h)}{\Pr(D)}$$

Όπου:

- $\Pr(h)$: η εκ των προτέρων πιθανότητα ισχύος της h, χωρίς να προηγηθεί παρατήρηση των δεδομένων του D.

- $\Pr(D/h)$: η δεσμευμένη πιθανότητα που εκφράζει το ενδεχόμενο παρατήρησης των δεδομένων του D, ισχυούσης της h (πιθανοφάνεια-likelihood).

- $\Pr(D)$: η εκ των προτέρων πιθανότητα παρατήρησης των δεδομένων του D. Ο συγκεκριμένος όρος απλοποιείται και δε συμμετέχει στους υπολογισμούς.

- $\Pr(h/D)$: η ζητούμενη εκ των προτέρων πιθανότητα ισχύος της h δεδομένης της παρατήρησης των δεδομένων του D.

Η αναζήτηση επομένως της πιο πιθανής υπόθεσης h δεδομένου του D ανάγεται στην εύρεση της υπόθεσης εκείνης με τη μεγαλύτερη εκ των υστέρων πιθανότητα (maximum a-posteriori ή MAP hypothesis). Ορίζουμε την υπόθεση αυτή ως:

$$h_{MAP} = \arg\max_{h \in H} \{\Pr(h/D)\} = \arg\max_{h \in H} \left\{ \frac{\Pr(h)\Pr(D/h)}{pr(h)} \right\} = \arg\max_{h \in H} \{\Pr(h)\Pr(D/h)\}$$

Ο παραπάνω αλγόριθμος συναντάτε στη βιβλιογραφία με το όνομα Βέλτιστος ταξινομητής Bayes, καθώς αποδεικνύεται θεωρητικά (Theodoridis and Koutroumbas, 1999) πως είναι σε θέση να υποδείξει το άνω φράγμα των επιδόσεων ενός συστήματος ταξινόμησης για ένα συγκεκριμένο πρόβλημα.

Για το πρόβλημα της ταξινόμησης, χρησιμοποιώντας διανυσματική αναπαράσταση των δεδομένων, έστω:

- C: τυχαία μεταβλητή που δείχνει την κλάση μιας παρατήρησης (εναλλακτικής).
- X: διάνυσμα τυχαίων μεταβλητών που δείχνει τις τιμές των παρατηρούμενων χαρακτηριστικών.
- c: μια συγκεκριμένη ετικέτα κλάσης.
- x: ένα συγκεκριμένο παρατηρούμενο διάνυσμα.

Έχοντας μια παρατήρηση δοκιμής x για ταξινόμηση, χρησιμοποιείται ο νόμος του Bayes για να υπολογιστεί η εκ των υστέρων πιθανότητα για κάθε κλάση δεδομένου του διανύσματος x, και επιλέγεται η μεγαλύτερη από αυτές:

$$\begin{aligned}&\arg\max_{\forall c}\{p(C=c \mid X=x)\}=\arg\max_{\forall c}\left\{\frac{p(C=c)p(X=x \mid C=c)}{p(X=x)}\right\}=\\&=\arg\max_{\forall c}\{p(C=c)p(X=x \mid C=c)\}\end{aligned}$$

Εδώ X=x αναπαριστά το γεγονός $X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots X_k = x_k$, οπότε:

$$\arg\max_{\forall c}\{p(C=c \mid X=x)\}=\arg\max_{\forall c}\left\{p(C=c)p\left(\bigcap_i X_i = x_i \mid C=c\right)\right\}$$

Όπου οι πιθανότητες στο δεξί μέλος της ισότητας εκτιμώνται από τις ταξινομημένες παρατηρήσεις του δείγματος εκπαίδευσης.

Χαρακτηριστικός είναι ο Naïve Bayes ταξινομητής, ο οποίος προσφέρει μια απλή πιθανοθεωρητική προσέγγιση στα προβλήματα μάθησης με επίβλεψη, όπου στόχος είναι η επακριβής πρόβλεψη της κατηγορίας-κλάσης των παρατηρήσεων δοκιμής χρησιμοποιώντας ταξινομημένες παρατηρήσεις εκπαίδευσης που περιλαμβάνουν την πληροφορία της κλάσης που ανήκουν. Βασίζεται σε δύο σημαντικές εκλαϊκευτικές υποθέσεις. Συγκεκριμένα, υποθέτει ότι κάθε χαρακτηριστικό των παρατηρήσεων είναι στοχαστικά ανεξάρτητα των υπολοίπων, δεδομένης κλάσης και ότι δεν υπάρχουν άλλα κρυφά χαρακτηριστικά που να επηρεάζουν τη διαδικασία της πρόβλεψης. Έτσι, η πιθανότητα της παραπάνω σχέσης μετατρέπεται σε γινόμενο πιθανοτήτων. Οπότε:

$$\arg\max_{\forall c}\{p(C=c \mid X=x)\} = \arg\max_{\forall c}\left\{p(C=c)\prod_i p(X_i = x_i \mid C=c)\right\}$$

Ο παράγοντας $p(C=c)$ υπολογίζεται βάσει της συχνότητας εμφάνισης της κλάσης c στις παρατηρήσεις του δείγματος εκπαίδευσης. Οι δεσμευμένες πιθανότητες $p(X_i = x_i \mid C=c)$ υπολογίζονται ανάλογα με το αν το χαρακτηριστικό X_i είναι διακριτό ή συνεχές. Για τα διακριτά χαρακτηριστικά των διανυσμάτων, εκείνα δηλαδή που παίρνουν διακριτές τιμές, η πιθανότητα αυτή είναι ένας πραγματικός αριθμός μεταξύ 0 και 1 που αντιπροσωπεύει την πιθανότητα το εκάστοτε χαρακτηριστικό X_i να πάρει την τιμή x_i δεδομένης της κλάσης c. Για τα συνεχή χαρακτηριστικά, θεωρείται ότι οι τιμές ακολουθούν μια πιθανοτική κατανομή (ξεχωριστή για κάθε χαρακτηριστικό), η οποία προσεγγίζεται από τα δείγματα εκπαίδευσης. Η συνήθης θεώρηση είναι οι τιμές των χαρακτηριστικών να είναι κανονικά κατανεμημένες. Οπότε για συνεχή χαρακτηριστικά προκύπτει:

$$p(X_i = x_i \mid C=c) = g(x_i; \mu_{i,c}, \sigma_{i,c})$$

Όπου $g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ η συνάρτηση πυκνότητας πιθανότητας μια κανονικής (Gaussian) κατανομής.

3.3.2 Δέντρα Λογιστικής παλινδρόμησης – Logistic Model Trees (LMT)

Τα δέντρα αποφάσεων δίνουν υψηλές αποδόσεις ταξινόμησης, αλλά παρουσιάζουν προβλήματα όταν υπάρχουν πολλοί κόμβοι αποφάσεων, καθώς είναι δύσκολο να προβλεφθεί μια σωστή ταξινόμηση του δείγματος. Εντούτοις, αποδεικνύεται ότι η μέθοδος boosting μπορεί, επίσης, να χρησιμοποιηθεί για την κατασκευή βέλτιστων δέντρων αποφάσεων. Παραδείγματος χάριν, ο αλγόριθμος LogitBoost έχει χρησιμοποιηθεί για να δημιουργήσει δέντρα με την εφαρμογή μοντέλων λογιστικής παλινδρόμησης. Αυτά είναι γνωστά ως Logistic Model Trees (LMT) και λειτουργούν με τον ίδιο τρόπο όπως τα δέντρα λογιστικής παλινδρόμησης.

Ο αλγόριθμος LogitBoost είναι μέρος της μεθόδου λογιστικής παλινδρόμησης. Θεωρείται ότι κάθε επανάληψη του αλγορίθμου προσαρμόζει μια απλή συνάρτηση παλινδρόμησης σε όλα τα χαρακτηριστικά του δείγματος, αναζητώντας εκείνη της συνάρτηση που θα σημειώσει το μικρότερο σφάλμα ταξινόμησης και προσθέτοντας την στο μοντέλο ταξινόμησης. Το αποτέλεσμα του αλγορίθμου LogitBoost είναι ένα μοντέλο λογιστικής παλινδρόμησης πολλαπλών συναρτήσεων και μέγιστης πιθανότητας. Ένας κατάλληλος αριθμός επαναλήψεων boosting μπορεί να προσδιοριστεί με τον υπολογισμό της αναμενόμενης απόδοσης για έναν δεδομένο αριθμό επαναλήψεων με cross-validation και η διαδικασία τερματίζεται όταν η εκτιμώμενη απόδοση παύει να αυξάνεται. Μια απλή επέκταση αυτού του αλγορίθμου οδηγεί στη μέθοδο Logistic Model Trees (LMT, Landwehr, Hall and Frank).

Συγκεκριμένα, η μέθοδος LMT (Yohannes and Webb, 1999) είναι μια μη παραμετρική προσέγγιση που αναπτύχθηκε για την ανάλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Σε κάθε περίπτωση το μοντέλο ταξινόμησης ή παλινδρόμησης που αναπτύσσεται αναπαρίσταται με τη μορφή ενός δέντρου αποφάσεων. Στην περίπτωση της ταξινόμησης κύριος σκοπός της μεθόδου είναι να παράγει ένα ακριβές σύνολο από κανόνες ταξινόμησης βάσει των οποίων θα προβλέπεται σε ποια κατηγορία ανήκει κάθε ένα αντικείμενο, σύμφωνα με τα αντίστοιχα χαρακτηριστικά του. Η δομή ενός κανόνα ταξινόμησης της μεθόδου επικεντρώνεται στους ορισμούς τριών κύριων παραγόντων: (α) του κανόνα διαχωρισμού του δείγματος των αντικειμένων, (β) των κριτηρίων αξιολόγησης της ποιότητας του διαχωρισμού, (γ) των κριτηρίων για την επιλογή του βέλτιστου δέντρου για ανάλυση. Τα βασικά βήματα για την δημιουργία ενός δέντρου ταξινόμησης είναι: (α) δημιουργία ενός δέντρου με μεγάλο αριθμό κόμβων, (β) ένωση μερικών διακλαδώσεων για την παραγωγή μιας σειράς από μικρότερα δέντρα διαφορετικού μεγέθους, (γ) επιλογή ενός βέλτιστου δέντρου μέσω της μέτρησης της ακρίβειας του δέντρου.

Για την ανάπτυξη ενός δέντρου ταξινόμησης χρησιμοποιείται η ίδια λογική με τα δέντρα ταξινόμησης με τη μόνη διαφορά ότι δεν απαιτείται ο προσδιορισμός a priori πιθανοτήτων των κατηγοριών και η εκτίμηση σφάλματος ταξινόμησης. Η ανάπτυξη ενός δέντρου παλινδρόμησης απαιτεί τον καθορισμό τριών στοιχείων: (α) ενός συνόλου ερωτήσεων η απάντηση των οποίων οδηγεί στην ταξινόμηση των αντικειμένων, (β) των κανόνων αξιολόγησης της ποιότητας των ερωτήσεων που αναπτύσσονται, και (γ) των κανόνων για το προσδιορισμό της κατηγορίας σε κάθε τερματικό κόμβο του δέντρου. Σε ένα δέντρο παλινδρόμησης, η εξαρτημένη μεταβλητή είναι διακριτή ή συνεχής. Το κριτήριο αξιολόγησης της ποιότητας του διαχωρισμού είναι το μέσο άθροισμα των τετραγώνων της εξαρτημένης μεταβλητής και ο κανόνας διαχωρισμού εκτιμάται από την μείωση του σταθμισμένου αθροίσματος των τετραγώνων.

Αρχικά, όλα τα αντικείμενα τοποθετούνται σε έναν αρχικό κόμβο, ο οποίος είναι ανομοιογενής καθώς περιέχει αντικείμενα από διάφορες κατηγορίες. Ο κύριος στόχος είναι η εύρεση εκείνων των κανόνων που θα διαχωρίσουν τα αντικείμενα δημιουργώντας νέους κόμβους σε κατώτερα επίπεδα του δέντρου, οι οποίοι θα είναι περισσότερο ομοιογενείς σε σχέση με τους προηγούμενους κόμβους. Η μέθοδος CART (Classification and Regression Trees, Breimani et al., 1984) χρησιμοποιείται για την εκτίμηση όλων των πιθανών διαχωρισμών των αντικειμένων. Σε κάθε κόμβο t του δέντρου τα αντικείμενα του δείγματος διαχωρίζονται σε δύο επιμέρους κόμβους t_L και t_R στο αμέσως κατώτερο επίπεδο του δέντρου, ανάλογα με τον εάν ικανοποιούν ή όχι έναν κανόνα (ερώτηση) της μορφής $x_{ij} \leq d_j$, όπου x_j είναι ένα χαρακτηριστικό και d_j είναι ένα όριο διαχωρισμού. Ειδικότερα, μια παρατήρηση i τοποθετείται στον κόμβο t_L εάν $x_{ij} \leq d_j$, διαφορετικά τοποθετείται στον κόμβο t_R. Ο βέλτιστος κανόνας διαχωρισμού καθορίζεται μεγιστοποιώντας τη μείωση της ανομοιογένειας (impurity) που αποφέρει ο διαχωρισμός. Ένας διαχωρισμός θεωρείται ομοιογενής εάν οι δύο κόμβοι που δημιουργούνται από αυτόν περιλαμβάνουν (ο καθένας) αντικείμενα από διαφορετικές κατηγορίες. Εάν κάποιος κόμβος περιλαμβάνει αντικείμενα από διαφορετικές κατηγορίες, τότε ο διαχωρισμός θεωρείται ως ανομοιογενής. Βάσει αυτής της θεώρησης ως κριτήριο επιλογής του κατάλληλου διαχωρισμού (επιλογή χαρακτηριστικού x_j και τιμής ορίου d_j) θεωρείται η μεγιστοποίηση της ακόλουθης συνάρτησης:

$$\varphi(s,t) = SS(t) - SS(t_L) - SS(t_R),$$

Όπου s ο διαχωρισμός των αντικειμένων από τον κανόνα που αναπτύσσεται, $SS(t_L)$ και $SS(t_R)$, το άθροισμα των τετραγώνων των ανεξάρτητων μεταβλητών που καταλήγουν αντίστοιχα, στον αριστερό και το δεξιό κόμβο και χαρακτηρίζουν την ομοιογένεια διαχωρισμού του κάθε κόμβου.

Ο υπολογισμός του όρου $SS(t)$ εξαρτάται από τους κανόνες της ποιότητας διαχωρισμού των αντικειμένων και υπάρχουν δύο κριτήρια που συντελούν στη μείωση της ανομοιογένειας που αποφέρει ο διαχωρισμός. Τα κριτήρια αυτά είναι η συνάρτηση ελαχίστων τετραγώνων (LS-Least Squares) και η συνάρτηση ελάχιστης τυπικής απόκλισης (LAD-Least Absolute Deviation).

Δεδομένου της χρήσης του κριτηρίου των ελαχίστων τετραγώνων, ο όρος $SS(t)$ υπολογίζεται από τη σχέση $SS(t) = \sum \left(y_{i(t)} - \bar{y}_{(t)} \right)^2$, για $i = 1,2,...,N_t$, όπου $y_{i(t)}$ οι ατομικές τιμές της ανεξάρτητες μεταβλητής στον κόμβο t και $\bar{y}_{(t)}$ η μέση τιμή της ανεξάρτητης μεταβλητής στον κόμβο t.

Ένας εναλλακτικός υπολογισμός του όρου $SS(t)$ είναι με τη χρήση της σταθμισμένης διακύμανσης στον αριστερό και δεξιό κόμβο, όπου τα βάρη είναι οι αναλογίες των περιπτώσεων που καταλήγουν στον αριστερό κόμβο (t_L) και στο δεξί κόμβο (t_R). Εάν θεωρεί ότι $p(t) = N_t / N$ είναι η αναλογία των περιπτώσεων που καταλήγουν στον κόμβο t και $s^2(t)$ η διακύμανση της ανεξάρτητης μεταβλητής στον κόμβο t, τότε η διακύμανση ορίζεται ως $s^2(t) = \frac{1}{N_{(t)}} \sum_{i=1}^{N_t} \left[y_i - \bar{y}_{(t)} \right]^2$ και ως κριτήριο επιλογής του κατάλληλου διαχωρισμού θεωρείται η μεγιστοποίηση της ακόλουθης συνάρτησης:

$$\varphi(s,t) = s^2(t) - \left[p_L s^2(t_L) + p_R s^2(t_R) \right]$$

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

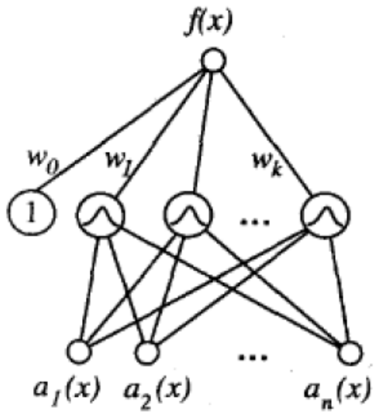
3.3.3 Ακτινωτό Λειτουργικό Δίκτυο – Radial Basis Function Network (RBFNetwork)

Ένας τρόπος προσέγγισης μιας συνάρτησης, ο οποίος είναι αρκετά κοντά στη μέθοδο του σταθμισμένου γείτονα με την απόσταση αλλά και με τα Νευρωνικά Δίκτυα, είναι η μάθηση με τη βοήθεια συναρτήσεων ακτινικής βάσης (Radial Basis Functions Networks).

Σε αυτή τη μέθοδο, η συνάρτηση μάθησης είναι η $f(x) = w_0 + \sum_{u=1}^{k} w_u K_u(d(x_u, x))$. Όπου x_u είναι τα παραδείγματα εκπαίδευσης και $K_u(d(x_u, x))$ είναι μία φθίνουσα συνάρτηση, όσο το d μεγαλώνει. Η συνάρτηση K συνήθως επιλέγεται να είναι μια Gaussian συνάρτηση με κέντρο x_u και διακύμανση $\sigma_u{}^2$.

$$K_u(d(x_u, x)) = e^{\frac{1}{2\sigma_k{}^2} d^2(x_u, x)}$$

Η μάθηση με δίκτυα συναρτήσεων ακτινικής βάσης μοιάζει πολύ με τη μάθηση ενός Νευρωνικού Δικτύου δύο επιπέδων, όπου το πρώτο επίπεδο υπολογίζει τις τιμές της συνάρτησης $K_u(d(x_u, x))$ και το δεύτερο επίπεδο υπολογίζει ένα γραμμικό συνδυασμό των τιμών του πρώτου επιπέδου (σχήμα 3.1). Έτσι, το αποτέλεσμα είναι μια ολοκληρωτική προσέγγιση του γραμμικού συνδυασμού των κοντινότερων συναρτήσεων πυρήνα (Mitchell (1997), Alpaydin (2004), Bishop (2006)).

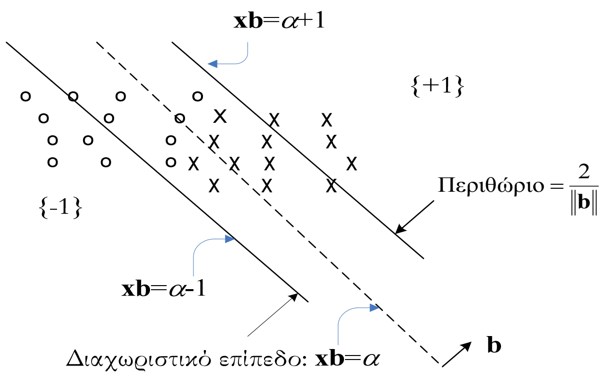


Σχήμα 3. 1: Αρχιτεκτονική δομή της μεθόδου ταξινόμησης RBFNetwork

3.3.4 Μηχανές Διανύσματος Υποστήριξης – Support Vector Machines (SVM)

Οι μηχανές διανύσματος υποστήριξης (Support Vector Machines – SVM, Vapnik, 2000 και Burges, 1998) έχουν αναπτυχθεί τα τελευταία χρόνια ως μια από τις σημαντικότερες μεθόδους για την ανάπτυξη μοντέλων ταξινόμησης. Κύριο χαρακτηριστικό τους αποτελεί το σημαντικό θεωρητικό υπόβαθρο πάνω στο οποίο βασίζονται οι SVM, καθώς και η πληθώρα επιτυχημένων πρακτικών εφαρμογών.

Η λογική των SVM παρουσιάζεται συνοπτικά στο σχήμα 3.2 όπου απεικονίζεται ένα πρόβλημα ταξινόμησης n αντικειμένων οι οποίες περιγράφονται βάσει m χαρακτηριστικών, σε δύο κατηγορίες οι οποίες συμβολίζονται ως +1 και -1.



Σχήμα 3. 2: Γραφική απεικόνιση των SVM

Στόχος των SVM, στην απλή γραμμική περίπτωση, είναι η ανάπτυξη του βέλτιστου υπερεπιπέδου της μορφής $\mathbf{xb}-\alpha$ για την ταξινόμηση των αντικειμένων, όπου ως $\mathbf{x}$ συμβολίζεται ένας πίνακας διαστάσεων nxm με τα στοιχεία των αντικειμένων του δείγματος εκμάθησης. Συμβολίζοντας ως $\mathbf{D}$ ένα διαγώνιο πίνακα διαστάσεων nxn με την κύρια διαγώνιο να έχει τιμές +1 ή -1 ανάλογα με την ταξινόμηση των αντικειμένων του δείγματος εκμάθησης, και ως $\mathbf{e}$ το μοναδιαίο διάνυσμα διαστάσεων nx1, ο εντοπισμός του βέλτιστου υπερεπιπέδου επιτυγχάνεται με την επίλυση του ακόλουθου τετραγωνικού προγράμματος (ως ν συμβολίζεται μια αυστηρά θετική σταθερά):

$$\left.\begin{array}{l} \min\limits_{\mathbf{b},\alpha,\mathbf{d}} \ \nu \mathbf{e}^T \mathbf{d} + \frac{1}{2}\mathbf{b}^T \mathbf{b} \\ \text{υπό}: \\ \qquad \mathbf{D}\left(\mathbf{x}\mathbf{b} - \mathbf{e}\alpha\right) + \mathbf{d} \geq \mathbf{e} \\ \qquad \mathbf{d} \geq 0 \end{array}\right\}$$

Ο τετραγωνικός όρος $\mathbf{b}^T\mathbf{b}$ στην αντικειμενική συνάρτηση του προβλήματος μεγιστοποιεί το περιθώριο μεταξύ των δυο υπερεπιπέδων $\mathbf{xb} - \alpha = +1$ και $\mathbf{xb} - \alpha = -1$, το οποίο ισούται με $2/\|\mathbf{b}\|$. Εκτός της μεγιστοποίησης του περιθωρίου των κατηγοριών, το πρόβλημα (4.6) λαμβάνει υπόψη και το σφάλμα ταξινόμησης με τις μεταβλητές του διανύσματος $\mathbf{d}$ (η σταθερά $\nu > 0$ αναπαριστά τη σχετική βαρύτητα που αποδίδεται στην ελαχιστοποίηση των σφαλμάτων). Όταν όλες οι μεταβλητές του διανύσματος $\mathbf{d}$ είναι ίσες με το μηδέν, τότε οι δύο κατηγορίες είναι αυστηρά γραμμικά διαχωρισμένες και το επίπεδο $\mathbf{xb} = \alpha + 1$ περικλείει όλα τα αντικείμενα της κατηγορίας +1, ενώ το επίπεδο $\mathbf{xb} = \alpha - 1$ περικλείει όλα τα αντικείμενα της κατηγορίας -1.

Με την επίλυση του παραπάνω προβλήματος και τον προσδιορισμό των $\mathbf{b}$ και α που καθορίζουν το βέλτιστο υπερεπίπεδο, η ταξινόμηση κάθε αντικειμένου μπορεί εύκολα να πραγματοποιηθεί ως εξής:

$$\text{Εάν } x_i b - a \begin{cases} > 0, \text{τότε } x_i \in \{+1\} \\ < 0, \text{τότε } x_i \in \{-1\} \\ = 0, \text{τότε } x_i \in \{+1\} \text{ ή } x_i \in \{-1\} \end{cases}$$

Το κύριο μειονέκτημα του προβλήματος βελτιστοποίησης για τον προσδιορισμό του βέλτιστου μοντέλου ταξινόμησης αφορά τον αυξημένο υπολογιστικό φόρτο που απαιτεί η επίλυσή του, καθώς πρόκειται για ένα πρόβλημα τετραγωνικού προγραμματισμού. Για την αντιμετώπιση του προβλήματος αυτού οι Fung και Mangasarian (2001) πρότειναν μια εναλλακτική διατύπωση του προβλήματος ως εξής (ως $\mathbf{A}$ συμβολίζεται ο πίνακας διαστάσεων nxm με τα στοιχεία του δείγματος εκπαίδευσης):

$$\min_{b,a,d} \nu \frac{1}{2}\|d\|^2 + \frac{1}{2}\left(b^T b + a^2\right)$$
$$\text{υπό}:$$
$$D\left(Ab - ea\right) + d \geq e$$

Η αντικειμενική συνάρτηση του νέου προβλήματος αφορά την ελαχιστοποίηση της νόρμας δεύτερης τάξης του διανύσματος $\mathbf{d}$ (σε αντίθεση με την νόρμα πρώτης τάξεως που χρησιμοποιείται στο πρόβλημα). Επιπλέον, η μεγιστοποίηση του περιθωρίου πραγματοποιείται τόσο σε σχέση με τη διεύθυνση $\mathbf{b}$ του διαχωριστικού υπερεπιπέδου, όσο και σε σχέση με τη σχετική του θέση α ως προς την αρχή των αξόνων. Στην εναλλακτική αυτή διατύπωση δεν απαιτείται ο περιορισμός μη αρνητικότητας του $\mathbf{d}$, καθώς εάν υπάρχει κάποιο d_i αρνητικό τότε η αντικειμενική συνάρτηση μπορεί να μειωθεί θέτοντας $d_i = 0$, κάτι που δεν παραβιάζει τον περιορισμό ανισότητας.

Το παραπάνω πρόβλημα απλοποιείται ακόμα περισσότερο εάν ο περιορισμός ανισότητας μετατραπεί σε περιορισμό ισότητας ως εξής::

$$\min_{b,a,d} \nu \frac{1}{2}\|d\|^2 + \frac{1}{2}\left(b^T b + a^2\right)$$
$$\text{υπό}:$$
$$D\left(Ab - ea\right) + d = e$$

Σε αυτή τη μορφή, η ανάπτυξη του βέλτιστου υπερεπιπέδου ταξινόμησης ανάγεται σε ένα πρόβλημα βελτιστοποίησης υπό περιορισμούς ισότητας, το οποίο μπορεί να επιλυθεί εύκολα χρησιμοποιώντας γνωστές τεχνικές βελτιστοποίησης (πολλαπλασιαστές Lagrange).

Στη μη γραμμική περίπτωση, τα δεδομένα αναπαριστώνται σε ένα άλλο χώρο υψηλότερων διαστάσεων H, χρησιμοποιώντας μια συνάρτησης Φ, τέτοια ώστε $\Phi: R^n \rightarrow H$. Έτσι ο αλγόριθμος εκπαίδευσης εξαρτάται μόνο από τα δεδομένα που βρίσκονται στον χώρο H, δηλαδή από τις συναρτήσεις $\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$. Στην περίπτωση όμως, που ο χώρος H είναι εξαιρετικά μεγάλης διάστασης ο προσδιορισμός της συνάρτησης αντιστοίχησης Φ και ο υπολογισμός των εσωτερικών γινομένων $\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$ απαιτεί αυξημένο υπολογιστικό φόρτο. Για την αντιμετώπιση του προβλήματος αυτού εισάγεται μια συνάρτηση πυρήνα (kernel function) K τέτοια ώστε $K(\mathbf{x}_i,\, \mathbf{x}_j) = \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$. Η γραμμική περίπτωση που αναπτύχθηκε προηγούμενα αντιστοιχεί στο γραμμικό πυρήνα $K(\mathbf{x}_i,\ \mathbf{x}_j) = \mathbf{x}_i\mathbf{x}_j^T$, ενώ οι πιο δημοφιλείς μη γραμμικοί πυρήνες είναι ο πολυωνυμικός πυρήνας $K(\mathbf{x}_i,\ \mathbf{x}_j) = (\mathbf{x}_i\mathbf{x}_j + 1)^p$ και ο εκθετικός πυρήνας $K(\mathbf{x}_i,\ \mathbf{x}_j) = \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / 2\sigma^2\right)$. Στην παρούσα έρευνα χρησιμοποιείται ο εκθετικός πυρήνας.

Με την εισαγωγή της συνάρτησης πυρήνα, το παραπάνω πρόβλημα βελτιστοποίησης διατυπώνεται ως εξής:

$$\begin{aligned}&\min_{u,a,d} \nu \frac{1}{2}\|d\|^2 + \frac{1}{2}\left(u^T u + a^2\right)\\ &\text{υπό}:\\ &D\left(K\left(A, A\right)Du - ea\right) + d = e\end{aligned}$$

Με την επίλυση του παραπάνω μη γραμμικού προβλήματος το βέλτιστο υπερεπίπεδο διαχωρισμού των αντικειμένων διατυπώνεται ως εξής:

$$f\left(\mathbf{x}\right) = K(\mathbf{x}, \mathbf{A})\mathbf{Du} - \alpha$$

3.3.5 Ο Αλγόριθμος του Πλησιέστερου Γείτονα – k-Nearest Neighbor Classifier (IBk)

Αντικειμενικός σκοπός των αλγορίθμων των πλησιέστερων γειτόνων (Nearest Neighbors, NN) είναι η εκτίμηση της υπό συνθήκη πιθανότητας ένα αντικείμενο $\mathbf{x}_i$ να ανήκει σε μία κατηγορία. Ο υπολογισμός αυτής της πιθανότητας πραγματοποιείται βάσει του πλήθους των αντικειμένων του δείγματος εκμάθησης, τα οποία ανήκουν στην εξεταζόμενη κατηγορία και βρίσκονται στον γειτονικό χώρο του $\mathbf{x}_i$. Ο προσδιορισμός των γειτονικών αντικειμένων του $\mathbf{x}_i$ μπορεί εύκολα πραγματοποιηθεί προσδιορίζοντας την απόσταση κάθε αντικειμένου του δείγματος εκμάθησης από το $\mathbf{x}_i$, χρησιμοποιώντας για παράδειγμα την Ευκλείδεια απόσταση, και θεωρώντας ότι όλες οι μεταβλητές (χαρακτηριστικά) έχουν την ίδια σπουδαιότητα. Ταυτόχρονα, θα πρέπει να καθοριστεί και το εύρος του γειτονικού χώρου του

$\mathbf{x}_i$ με τον καθορισμό μιας παραμέτρου K η οποία προσδιορίζει το πλήθος των γειτονικών αντικειμένων που θα εξεταστούν.

Δεδομένων των K γειτονικών αντικειμένων του $\mathbf{x}_i$, η ταξινόμησή του μπορεί εύκολα να πραγματοποιηθεί εξετάζοντας την κατηγορία στην οποία ανήκουν τα γειτονικά του στοιχεία. Ειδικότερα, μέσω του απλού κανόνα της πλειοψηφίας, αποφασίζεται η ταξινόμηση του $\mathbf{x}_i$ στην κατηγορία στην οποία ανήκει η πλειοψηφία των Κ πλησιέστερων γειτόνων του.

Στην παρούσα έρευνα χρησιμοποιείται ο αλγόριθμος του πλησιέστερου γείτονα, θεωρώντας $K = 1$. Περαιτέρω ανάλυση των ιδιοτήτων, των χαρακτηριστικών και των υπολογιστικών μεθόδων των αλγορίθμων των πλησιέστερων γειτόνων δίδονται στο βιβλίο του Hand (1997).

3.4 Αλγόριθμοι Επιλογής Χαρακτηριστικών

3.4.1 Σπουδαιότητα της διαδικασίας επιλογής χαρακτηριστικών

Οι έρευνες σε θέματα επιλογής χαρακτηριστικών επικεντρώνονται στην ανάπτυξη νέων μεθοδολογικών προσεγγίσεων οι οποίες επιτρέπουν την αποτελεσματικότερη αξιοποίηση ενός συνόλου δεδομένων (δείγμα εκμάθησης) με στόχο την ανάπτυξη αξιόπιστων υποδειγμάτων ταξινόμησης. Ανεξαρτήτως όμως της χρησιμοποιούμενης μεθοδολογίας η επιτυχία κάθε υποδείγματος ταξινόμησης εξαρτάται άμεσα από την ποιότητα των δεδομένων που αναλύονται. Η έννοια της ποιότητας μπορεί να αναλυθεί σε δύο διαστάσεις:

1. Την επάρκεια των δεδομένων: Η ανάπτυξη ενός αξιόπιστου υποδείγματος ταξινόμησης πρέπει να βασίζεται σε δεδομένα τα οποία είναι αντιπροσωπευτικά του προβλήματος που αναλύεται. Σε αντίθετη περίπτωση είναι προφανές ότι το αναπτυσσόμενο υπόδειγμα δεν μπορεί να είναι ρεαλιστικό και να έχει την απαραίτητη δυνατότητα γενίκευσης. Η εξακρίβωση όμως της επάρκειας των δεδομένων είναι μια ιδιαίτερα δύσκολη διαδικασία, δεδομένου ότι προκειμένου να διαπιστωθεί η επάρκεια των δεδομένων θα πρέπει να υπάρχει συνολική γνώση του εξεταζόμενου πεδίου και όλων των πιθανών περιπτώσεων που μπορούν να εμφανιστούν.

2. Τον μη πλεονασμό των δεδομένων: Η ανάλυση δεδομένων που δεν διαθέτουν την ιδιότητα του μη πλεονασμού εγκυμονεί δύο κινδύνους: (α) την ενσωμάτωση θορύβου στα δεδομένα, και (β) την ανάλυση συσχετισμένων ή περιττών πληροφοριών. Η

περίπτωση του θορύβου προφανώς έχει αρνητικές συνέπειες στην ανάπτυξη ενός αποτελεσματικού υποδείγματος, καθώς οδηγεί σε στρέβλωση των δεδομένων με την ενσωμάτωση μη ρεαλιστικών περιπτώσεων. Αντίστοιχα, η ανάλυση συσχετισμένων ή περιττών πληροφοριών μπορεί επίσης να έχει αρνητική επίδραση στην ανάπτυξη του υποδείγματος, ενώ παράλληλα αυξάνει τον υπολογιστικό φόρτο που απαιτείται για την πραγματοποίηση της ανάλυσης.

Το δεύτερο από τα παραπάνω δύο θέματα έχει εξελιχθεί σε ένα από τα βασικά πεδία έρευνας στο χώρο της ταξινόμησης. Οι σχετικές έρευνες έχουν δώσει κύρια έμφαση στο πρόβλημα της επιλογής χαρακτηριστικών (feature selection). Δεδομένου ενός συνόλου χαρακτηριστικών $X = \{x_1, \ldots, x_m\}$ τα οποία περιγράφουν τα αντικείμενα, το πρόβλημα της επιλογής χαρακτηριστικών αναφέρεται στην επιλογή των κατάλληλων χαρακτηριστικών ο συνδυασμός των οποίων σε ένα υπόδειγμα μεγιστοποιεί την αναμενόμενη αποτελεσματικότητα του υποδείγματος. Η σημασία των διαδικασιών επιλογής χαρακτηριστικών αναλύεται στις ακόλουθες διαστάσεις (Kira and Rendell, 1992):

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

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

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

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

Στη διεθνή βιβλιογραφία έχουν προταθεί διάφορες μεθοδολογίες στο πρόβλημα της επιλογής χαρακτηριστικών. Πολλές από τις μεθοδολογίες αυτές είναι άρρηκτα συνδεδεμένες με συγκεκριμένες τεχνικές ταξινόμησης, ενώ άλλες είναι γενικοί αλγόριθμοι οι οποίοι μπορούν να εφαρμοστούν ανεξαρτήτως του τρόπου που υλοποιείται η ανάπτυξη ενός υποδείγματος. Στις μέχρι σήμερα έρευνες, η αποτελεσματικότητα των μεθοδολογιών αυτών έχει ελεγχθεί για περιορισμένο αριθμό τεχνικών ταξινόμησης και σε περιορισμένα σύνολα δεδομένων. Τα στοιχεία αυτά καθιστούν δύσκολη την εξαγωγή ασφαλών συμπερασμάτων σχετικά με την πραγματική αποτελεσματικότητα των μεθοδολογιών επιλογής χαρακτηριστικών.

3.4.2 Κατηγορίες Αλγορίθμων Επιλογής Χαρακτηριστικών

Οι αλγόριθμοι επιλογής χαρακτηριστικών (Feature Selection Algorithm – FSA) μπορούν να χωριστούν σε τρεις μεγάλες κατηγορίες σύμφωνα με τον τρόπο λειτουργίας τους: (1) ενσωματωμένες διαδικασίες (embedded scheme), (2) διαδικασίες filter, και (3) διαδικασίες wrapper.

Στις ενσωματωμένες διαδικασίες (**embedded scheme**) περιλαμβάνονται αλγόριθμοι και τεχνικές η εφαρμογή των οποίων είναι άμεσα συνδεδεμένη με μια συγκεκριμένη τεχνική ταξινόμησης. Σε αυτή την κατηγορία ανήκουν παραδοσιακοί αλγόριθμοι μηχανικής μάθησης, όπως οι μέθοδοι ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), CART (Breiman et al., 1984), και τα προσεγγιστικά σύνολα (rough sets, Pawlak, 1982). Οι τεχνικές αυτές χρησιμοποιούνται ευρέως για την ανάπτυξη δέντρων ή κανόνων απόφασης σε προβλήματα ταξινόμησης. Ταυτόχρονα με την ανάπτυξη των υποδειγμάτων ταξινόμησης οι τεχνικές αυτές ενσωματώνουν στη δομή τους κατάλληλες διαδικασίες επιλογής των χαρακτηριστικών που συμμετέχουν στο τελικό υπόδειγμα. Στην ίδια κατηγορία ενσωματώνονται και διαδικασίες ελέγχου της πολυπλοκότητας των υποδειγμάτων ταξινόμησης, οι οποίες χρησιμοποιούνται ευρέως σε τεχνικές, όπως τα νευρωνικά δίκτυα (weight decay, Moody, 1992) και οι μηχανές διανύσματος υποστήριξης (Vapnik, 1998).

Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στις **διαδικασίες filter** αυτή εφαρμόζονται πριν τη χρησιμοποίηση κάποιας τεχνικής ταξινόμησης και συνεπώς, δεν επηρεάζονται από αυτή. Ουσιαστικά, οι αλγόριθμοι αυτής της κατηγορίας λειτουργούν ως φίλτρα για την απαλοιφή των μη σχετικών ή πλεοναστικών χαρακτηριστικών. Το κύριο μειονέκτημα των αλγορίθμων αυτών είναι ότι αγνοούν την αλληλεπίδραση που πιθανόν υπάρχει μεταξύ του συνόλου των χαρακτηριστικών που επιλέγεται και της τεχνικής ταξινόμησης που χρησιμοποιείται για την ανάπτυξη του υποδείγματος ταξινόμησης.

Ένας πρώτος αλγόριθμος της διαδικασίας filter είναι ο FOCUS (Almuallim and Dietterich, 1991, 1994), που ακολουθεί μια εκθετική εμπρόσθια στρατηγική επιλογής χαρακτηριστικών και χρησιμοποιεί τη συνέπεια ως μέτρο αξιολόγησης των χαρακτηριστικών που επιλέγονται.

Ο αλγόριθμος FOCUS αναζητεί ένα μικρού μεγέθους σύνολο χαρακτηριστικών το οποίο επαρκεί για την περιγραφή της ταξινόμησης των αντικειμένων. Έχει όμως διαπιστωθεί ότι η έμφαση που αποδίδει ο αλγόριθμος στην επιλογή του μικρότερου δυνατού αριθμού χαρακτηριστικών συχνά οδηγεί σε επιλογή λανθασμένων χαρακτηριστικών. Παράλληλα, η εφαρμογή του αλγορίθμου απαιτεί μεγάλο υπολογιστικό φόρτο, λόγω της εξαντλητικής διερεύνησης των χαρακτηριστικών που πραγματοποιείται.

Ένας άλλος, επίσης, αλγόριθμος της διαδικασίας filter είναι ο RELIEF (Kira and Rendell, 1992), ο οποίος ακολουθεί μια στρατηγική επιλογής χαρακτηριστικών βασιζόμενος σε τυχαία δειγματοληψία των αντικειμένων από ένα σύνολο δεδομένων και χρησιμοποιεί την εσωτερική απόκλιση ως κριτήριο επιλογής. Ο αλγόριθμος ορίζει για κάθε χαρακτηριστικό ένα αριθμητικό βάρος που αντιπροσωπεύει την σημασία του χαρακτηριστικού στην ταξινόμηση των αντικειμένων. Κύριος στόχος του αλγορίθμου είναι ο εντοπισμός όλων των χρήσιμων χαρακτηριστικών, χωρίς όμως να δίνεται έμφαση στον εντοπισμό της περιττής πληροφορίας. Στην αρχική του μορφή ο αλγόριθμος αναπτύχθηκε για προβλήματα ταξινόμησης σε δύο κατηγορίες, αλλά αργότερα παρουσιάστηκαν επεκτάσεις για προβλήματα πολλαπλών κατηγοριών (Kononenko, 1994).

Τα δέντρα απόφασης και συναφείς τεχνικές, εκτός από μια τεχνική για την ανάπτυξη υποδειγμάτων ταξινόμησης μπορούν να χρησιμοποιηθούν και ως αλγόριθμοι επιλογής χαρακτηριστικών. Ο Cardie (1993) παρουσίασε μια τέτοια προσέγγιση χρησιμοποιώντας ένα δέντρο ταξινόμησης ως μέσο επιλογής χαρακτηριστικών για την ανάπτυξη ενός υποδείγματος ταξινόμησης με τον αλγόριθμο του πλησιέστερου γείτονα. Στην περίπτωση αυτή τα δέντρα απόφασης (ή άλλες συναφείς τεχνικές) λειτουργούν ως φίλτρα επιλογής χαρακτηριστικών τα οποία στη συνέχεια αποτελούν την είσοδο σε κάποια τεχνική ταξινόμησης. Βασικό μειονέκτημα αυτής της μεθόδου αποτελεί το γεγονός ότι τα χαρακτηριστικά που μπορούν να θεωρηθούν χρήσιμα σύμφωνα με ένα δέντρο απόφασης πιθανόν να θεωρηθούν περιττά από τη μέθοδο ταξινόμησης που χρησιμοποιείται στη συνέχεια.

Άλλες μεθοδολογίες που βασίζονται σε διαδικασίες τύπου filter για την επιλογή χαρακτηριστικών περιλαμβάνουν τους αλγορίθμους Las Vegas (Liu and Setiono, 1996a, 1998b), sequential forward generation (Pudil et al., 1994) και sequential backward generation (Choubey et al., 1996), ενώ ενδιαφέρον παρουσιάζουν και τεχνικές που βασίζονται σε διαδικασίες κλάδου και φράγματος (Narendra and Fukunaga, 1977, Dash and Liu, 1998) και δυναμικού προγραμματισμού (Chang, 1973).

Αλγόριθμοι επιλογής χαρακτηριστικών **διαδικασίας τύπου wrapper** χρησιμοποιούν τη μέθοδο ταξινόμησης ως μέρος της διαδικασίας (John et al., 1994). Ειδικότερα, βασιζόμενοι σε εμπρόσθιες, ανάστροφες ή τυχαίες διαδικασίες, οι αλγόριθμοι της κατηγορίας αυτής χρησιμοποιούν τη μέθοδο ταξινόμησης για την αξιολόγηση της αποτελεσματικότητας του συνόλου των χαρακτηριστικών που επιλέγονται. Για την επίτευξη αξιόπιστων αποτελεσμάτων, χρησιμοποιούνται τεχνικές επαναληπτικής δειγματοληψίας (resampling techniques), όπως το cross-validation (Stone, 1974) και το bootstrap (Efron, 1983). Το βασικό πλεονέκτημα της προσέγγισης αυτής είναι ότι λαμβάνει υπόψη τις αλληλεπιδράσεις μεταξύ της διαδικασίας επιλογής χαρακτηριστικών και της συγκεκριμένης τεχνικής που χρησιμοποιείται για την ανάπτυξη του υποδείγματος. Στον αντίποδα, κύριο μειονέκτημα αποτελεί ο αυξημένος υπολογιστικός φόρτος που απαιτείται για την υλοποίηση τέτοιων προσεγγίσεων, δεδομένου ότι αυτός καθορίζεται σημαντικά τόσο από την τεχνική ταξινόμησης που χρησιμοποιείται, όσο και από τη διαδικασία επαναληπτικής δειγματοληψίας (Blum and Langley, 1997).

Η λειτουργία ενός FSA μπορεί να περιγραφθεί βάσει των παρακάτω ιδιοτήτων (Blum and Langley, 1997, Doak, 1992, Liu and Motoda, 1998): (1) τη στρατηγική διερεύνησης των χαρακτηριστικών, (2) τη διαδικασία επιλογής των χαρακτηριστικών, και (3) το κριτήριο αξιολόγησης της ποιότητας των χαρακτηριστικών.

3.4.3 Στρατηγικές διερεύνησης FSAs

Κάθε αλγόριθμος επιλογής χαρακτηριστικών ακολουθεί μια συγκεκριμένη στρατηγική προκειμένου να διερευνήσει το σύνολο των χαρακτηριστικών. Η στρατηγική διερεύνησης στοχεύει στον προσδιορισμό κατάλληλων συντελεστών στάθμισης $w_1, \ldots, w_m$ των χαρακτηριστικών ανάλογα με την αναμενόμενη συμβολή τους στην ανάπτυξη ενός αξιόπιστου υποδείγματος ταξινόμησης. Οι συντελεστές στάθμισης μπορεί να είναι πραγματικοί αριθμοί στο διάστημα [0, 1] αναπαριστώντας τη σημαντικότητα του κάθε χαρακτηριστικού ή να έχουν δυαδική μορφή {0, 1} ανάλογα με το εάν ένα χαρακτηριστικό επιλέγεται ($w_j = 1$) ή όχι ($w_j = 0$).

Γενικά, η διερεύνηση ενός βέλτιστου διανύσματος βαρών μπορεί να επιτευχθεί μέσω τριών βασικών στρατηγικών (Molina et al., 2002): (1) εκθετική, (2) σειριακή, και (3) τυχαία:

Οι εκθετικές στρατηγικές διερεύνησης βασίζονται στην εξαντλητική διερεύνηση του συνόλου των πιθανών λύσεων και απαιτούν υπολογιστικό φόρτο $O(2^m)$. Οι στρατηγικές σειριακής διερεύνησης βασίζονται στον προσδιορισμό ενός συνόλου k πιθανών επόμενων λύσεων βάσει της τρέχουσας λύσης. Παρά τον αυξημένο υπολογιστικό φόρτο, η υλοποίηση τέτοιων στρατηγικών εγγυάται την εύρεση μιας ολικά βέλτιστης λύσης, χωρίς όμως αυτό να σημαίνει ότι ο εντοπισμός της ολικά βέλτιστης λύσης απαιτεί μια εξαντλητική διερεύνηση. Χαρακτηριστικό παράδειγμα αλγορίθμων που εφαρμόζουν αυτή τη στρατηγική είναι μέθοδοι τύπου κλάδου και φράγματος (Narendra and Fukunaga, 1977).

Οι στρατηγικές σειριακής διερεύνησης βασίζονται στον προσδιορισμό ενός συνόλου k πιθανών επόμενων λύσεων βάσει της τρέχουσας λύσης. Στρατηγικές της μορφής αυτής, συγκρινόμενες με τις εκθετικές στρατηγικές διερεύνησης, μειώνουν τον υπολογιστικό φόρτο, καθώς έχουν πολυωνυμική πολυπλοκότητα $O\left(m^{k+1}\right)$, αλλά είναι πιθανό να μην οδηγήσουν σε μια ολικά βέλτιστη λύση, δεδομένου ότι αυτή μπορεί να βρίσκεται σε μέρος του χώρου των λύσεων το οποίο δεν εξετάζεται κατά τη σειριακή διερεύνηση.

Τέλος, οι στρατηγικές τυχαίας διερεύνησης (Liu and Motoda, 1998) βασίζονται στη χρήση της τυχαιότητας ως μέσο αποφυγής τοπικά βέλτιστων λύσεων. Κύριο χαρακτηριστικό των στρατηγικών αυτών είναι ότι μπορούν να οδηγήσουν στον εντοπισμό περισσοτέρων του ενός διαφορετικών υποσυνόλων χαρακτηριστικών ή σε μέρος του χώρου λύσεων με περισσότερο αναξιόπιστα χαρακτηριστικά.

3.4.4 Διαδικασίες επιλογής χαρακτηριστικών

Οι διαδικασίες επιλογής των χαρακτηριστικών χωρίζονται σε πέντε κατηγορίες (Koller and Sahami, 1996): εμπρόσθιες (forward), ανάστροφες (backward), σύνθετες (compound), σταθμισμένες (weighting) και τυχαίες (random). Όλες αυτές οι διαδικασίες προσδιορίζουν με κάποιο τρόπο τα βάρη w_t των χαρακτηριστικών, έτσι ώστε να ενημερώσουν την τρέχουσα λύση.

Οι εμπρόσθιες διαδικασίες ξεκινούν από το κενό σύνολο, στο οποίο σταδιακά προστίθενται εκείνα τα χαρακτηριστικά που μεγιστοποιούν ένα κριτήριο αξιολόγησης της ποιότητας κάθε υποσυνόλου χαρακτηριστικών. Σε αντίθεση με μια εμπρόσθια διαδικασία, η ανάστροφη προσέγγιση ξεκινά από το σύνολο των χαρακτηριστικών και διερευνά την σταδιακή απαλοιφή χαρακτηριστικών ώστε να μεγιστοποιηθεί ένα κριτήριο αξιολόγησης. Οι σύνθετες διαδικασίες αποτελούν συνδυασμό εμπρόσθιων και ανάστροφων βημάτων επιτρέποντας, τόσο την προσθήκη χαρακτηριστικών στο υποσύνολο που έχει επιλεγεί, όσο και την αφαίρεση χαρακτηριστικών από αυτό, γεγονός που επιτρέπει την καλύτερη διερεύνηση των αλληλεπιδράσεων μεταξύ των χαρακτηριστικών. Στις διαδικασίες στάθμισης, η επιλογή των χαρακτηριστικών πραγματοποιείται έμμεσα αποδίδοντάς τους βάρη που αντικατοπτρίζουν τη συνεισφορά τους στο τελικό υπόδειγμα ταξινόμησης. Συνεπώς, η ανάλυση στην περίπτωση αυτή δεν οδηγεί στην απαλοιφή χαρακτηριστικών. Ο υπολογισμός των συντελεστών στάθμισης πραγματοποιείται, έτσι ώστε να μεγιστοποιηθεί η ποιότητα της ταξινόμησης που επιτυγχάνεται. Τέλος, η τυχαία επιλογή χαρακτηριστικών βασίζεται στη δυνατότητα διαμόρφωσης ενός οποιουδήποτε υποσυνόλου χαρακτηριστικών από ένα δεδομένο υποσύνολο επιλεγμένων χαρακτηριστικών.

3.4.5 Κριτήρια αξιολόγησης FSAs

Κάθε στρατηγική διερεύνησης και διαδικασία επιλογής χαρακτηριστικών βασίζεται σε κάποιο κριτήριο αξιολόγησης για τη μέτρηση της ποιότητας του κάθε υποσυνόλου χαρακτηριστικών που επιλέγεται.

Το πλέον δημοφιλές κριτήριο αφορά τη μεγιστοποίηση της αναμενόμενης ακρίβειας ταξινόμησης (Devijver and Kittler, 1982). Ένα εναλλακτικό κριτήριο αφορά την επιλογή χαρακτηριστικών, ώστε να μεγιστοποιούνται οι διαφορές στις κατανομές πυκνότητας πιθανότητας των κατηγοριών (Devijver and Kittler, 1982, Ben-Bassat, 1982). Ανάλογη πληροφορία με το παραπάνω κριτήριο παρέχουν και προσεγγίσεις οι οποίες βασίζονται στη μεγιστοποίηση της απόστασης μεταξύ των κατηγοριών (interclass distance), όπου η έννοια της απόστασης συνήθως, αφορά την Ευκλείδεια απόσταση μεταξύ των αντικειμένων διαφορετικών κατηγοριών. Ένα ακόμα διαδεδομένο κριτήριο αφορά την επιλογή χαρακτηριστικών βάσει των αλληλεξαρτήσεων τους, ώστε να αποφευχθεί η επιλογή εξαρτημένων χαρακτηριστικών. Ο πλέον απλοϊκός τρόπος μέτρησης των αλληλεξαρτήσεων των χαρακτηριστικών είναι η χρήση του συντελεστή συσχέτισης, αλλά είναι δυνατή και η εφαρμογή εναλλακτικών προσεγγίσεων (Hall, 1999).

Δύο ακόμα διαδεδομένα κριτήρια επιλογής χαρακτηριστικών αφορούν τον προσδιορισμό της εντροπίας στην ταξινόμηση των αντικειμένων και της συνέπειας των χαρακτηριστικών. Η έννοια της εντροπίας αναφέρεται στην ανομοιογένεια που παρατηρείται στην ταξινόμηση των αντικειμένων χρησιμοποιώντας ένα σύνολο χαρακτηριστικών (η σωστή ταξινόμηση των αντικειμένων δεν εμπεριέχει καμία ανομοιογένεια και συνεπώς, η εντροπία είναι μηδέν). Χαρακτηριστικά που συμβάλουν στη μείωση της εντροπίας θεωρείται ότι αποδίδουν αυξημένη πληροφορία για τη σωστή ταξινόμηση των αντικειμένων. Κριτήρια αυτής της κατηγορίας (εντροπία του Shannon, information gain ratio, κ.ά.) χρησιμοποιούνται ιδιαίτερα συχνά στην ανάπτυξη υποδειγμάτων ταξινόμησης υπό τη μορφή δέντρων ή κανόνων απόφασης (Breiman et al., 1984, Quinlan, 1993). Αντίστοιχα, η έννοια της συνέπειας αναφέρεται στη δυνατότητα διάκρισης των αντικειμένων διαφορετικών κατηγοριών με βάση ένα δεδομένο υποσύνολο χαρακτηριστικών (Almuallim and Dietterich, 1994). Ένα υποσύνολο χαρακτηριστικών θεωρείται συνεπές, όταν δεν υπάρχουν αντικείμενα διαφορετικών κατηγοριών με την ίδια περιγραφή στα χαρακτηριστικά του συνόλου.

Τέλος, ένα υπόδειγμα ταξινόμησης μπορεί να χρησιμοποιηθεί ως κριτήριο αξιολόγησης για τη μέτρηση της ποιότητας του υποσυνόλου των χαρακτηριστικών που επιλέγεται, ελαχιστοποιώντας το σφάλμα ταξινόμησης ή εναλλακτικά κριτήρια, όπως το εμβαδόν κάτω από την καμπύλη ROC (Receiver Operating Characteristic, Fawcett, 2003).

3.5 Διαδικασία ελέγχου Cross-Validation

Η διαδικασία ελέγχου cross-validation (Stone, 1974), μαζί με τη διαδικασία bootstrap, είναι μια διαδεδομένη διαδικασία στην κατηγορία των τεχνικών επαναληπτικής δειγματοληψίας (resampling techniques) η χρήση τη οποίας έχει ως στόχο την αξιολόγηση του βαθμού γενίκευσης των αποτελεσμάτων των υποδειγμάτων ταξινόμησης. Οι σημαντικότεροι λόγοι για τους οποίους είναι σκόπιμο να μετρηθεί ο βαθμός γενίκευσης των αποτελεσμάτων είναι πρώτον, για τον έλεγχο της αποτελεσματικότητας υποδειγμάτων ταξινόμησης και παλινδρόμησης βάσει ενός συνόλου δεδομένων (σύνολο αναφοράς) και δεύτερον, τη σύγκριση των διαφόρων μεθόδων ταξινόμησης μεταξύ τους, με σκοπό την επιλογή της πλέον κατάλληλης.

Οι τεχνικές αυτές αποσκοπούν την εξαγωγή ασφαλών και αξιόπιστων εκτιμήσεων για την αποτελεσματικότητα των εξεταζόμενων υποδειγμάτων ταξινόμησης χρησιμοποιώντας ένα κοινό δείγμα τόσο για την ανάπτυξη των υποδειγμάτων όσο και για τον έλεγχο τους. Γενικότερα, οι τεχνικές αυτές χρησιμοποιούνται για την αντιμετώπιση προβλημάτων που υπάρχουν τόσο στη συλλογή αντιπροσωπευτικών δεδομένων κατά τη φάση του ελέγχου, καθώς επίσης και τη μείωση του χρόνου και κόστους για τη συλλογή αυτών, όσο στην αξιολόγηση των νέων αντικειμένων (εναλλακτικών), ώστε να είναι δυνατός ο έλεγχος της αξιοπιστίας των υποδειγμάτων ταξινόμησης.

Η εφαρμογή του εκάστοτε υποδείγματος ταξινόμησης, συνοδεύεται από τη δημιουργία του δείγματος εκμάθησης και του δείγματος ελέγχου. Το δείγμα εκμάθησης χρησιμοποιείται για την ανάπτυξη ενός κατάλληλου μοντέλου ταξινόμησης, ανάλογα με τις αρχές τη κάθε μεθόδου, ενώ το δείγμα ελέγχου χρησιμοποιείται στον έλεγχο της αποτελεσματικότητας του μοντέλου. Με τον τρόπο όμως αυτό, πραγματοποιείται ένας και μόνο έλεγχος της αποτελεσματικότητας κάθε φορά, γεγονός που εμπεριέχει κίνδυνο τα αποτελέσματα που λαμβάνονται να είναι τυχαία, δηλαδή να μην είναι αντιπροσωπευτικά της μεθοδολογία που χρησιμοποιείται. Το πρόβλημα αυτό επιλύεται με την χρήση της επαναληπτικής διαδικασίας ελέγχου cross-validation.

Η διαδικασία πραγματοποιείται επαναληπτικά σε T στάδια. Η πλέον διαδεδομένη επιλογή για το πλήθος T των επαναλήψεων είναι 10 και τότε η διαδικασία ονομάζεται 10-fold cross-validation. Ανάλογα με το πλήθος των T επαναλήψεων και δεδομένου ενός συνόλου αναφοράς A αποτελούμενου από n αντικείμενα, η διαδικασία υλοποιείται ως εξής:

1. Διάσπαση του συνόλου αναφοράς A κατά τυχαίο τρόπο σε T αλληλοαποκλειόμενα υποσύνολα $A_1, A_2, \ldots, A_T$ μεγέθους n/T.

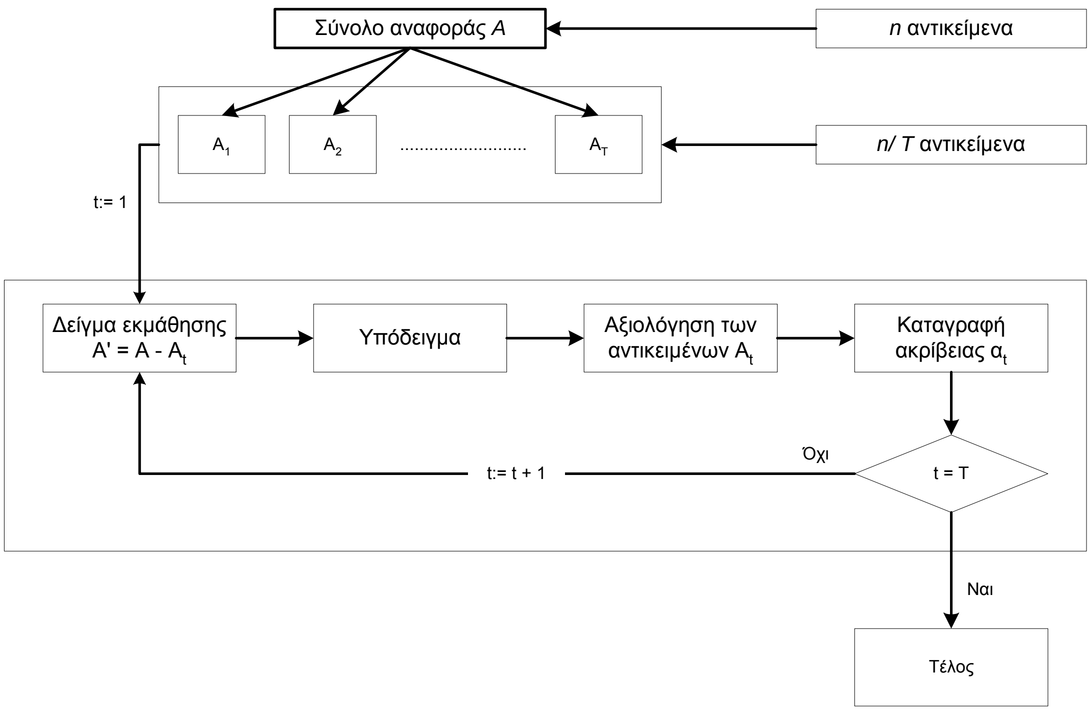
2. Για την πρώτη επανάληψη τίθεται $t=1$.

3. Για την τρέχουσα επανάληψη t επιλέγεται το σύνολο $A-A_t$ για την ανάπτυξη του υποδείγματος ταξινόμησης.

4. Για το υπόδειγμα που αναπτύχθηκε καταγράφεται η αποτελεσματικότητά του a_t συγκρινόμενη με ένα προεπιλεγμένο δείκτη a. Η αποτελεσματικότητα του υποδείγματος ταξινόμησης προσδιορίζεται βάσει των αντικειμένων του συνόλου A_t.

5. Εάν $t < T$ τότε τίθεται $t = t + 1$ και η διαδικασία επαναλαμβάνεται από το στάδιο 3, αλλιώς η διαδικασία τερματίζεται.

Από την παραπάνω επαναληπτική διαδικασία υπολογίζεται η αναμενόμενη αποτελεσματικότητα $E(a) = \frac{1}{T}\sum_{t=1}^{T} a_t$ ως ο μέσος όρος της αποτελεσματικότητας των T επιμέρους υποδειγμάτων ταξινόμησης που αναπτύχθηκαν και διατυπώνεται μια ολοκληρωμένη άποψη για την αποτελεσματικότητα της μεθόδου που χρησιμοποιείται για την ανάπτυξη ενός κατάλληλου και αξιόπιστου υποδείγματος.



Σχήμα 3. 3: Σχηματική απεικόνιση της διαδικασίας ελέγχου k-fold cross validation

Κεφάλαιο 4°

ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

4.1 Εισαγωγή

Στην παρούσα διατριβή, η συσχέτιση των επιλεγμένων αλγορίθμων επιλογής χαρακτηριστικών σε συνδυασμό με την εφαρμογή διαφόρων υποδειγμάτων ταξινόμησης ελέγχεται μέσω του υπολογισμού της ακρίβειας ταξινόμησης των δεδομένων. Το υπόδειγμα ταξινόμησης που λειτουργεί στα δεδομένα με το επιλεγμένο σύνολο χαρακτηριστικών μπορεί να παρουσιαστεί ως εκτιμητής της ακρίβειας ταξινόμησης. Το αποτέλεσμα του εκτιμητή σε νέα δεδομένα εκτιμάται μέσω της διαδικασίας ελέγχου cross-validation. Επειδή, όμως, το αποτέλεσμα στην αντιμετώπιση του προβλήματος ταξινόμησης διαφέρει από ένα σύνολο δεδομένων σε ένα άλλο, χρειάζεται η παρουσίαση της δυνατότητας εκτίμησης με τη βοήθεια ενός διαστήματος εμπιστοσύνης μέσα στο οποίο κυμαίνεται η ακρίβεια του ταξινομητή σε τυχαία άγνωστα δεδομένα. Η ακρίβεια ταξινόμησης, καθώς και το διάστημα εμπιστοσύνης αυτής αναλύονται στο 6° Κεφάλαιο της συγκεκριμένης διατριβής. Προτού, πραγματοποιηθεί η ανάλυση των διαστημάτων εμπιστοσύνης της ακρίβειας ταξινόμησης των εξεταζόμενων τεχνικών (Κεφάλαιο 3) και του συνδυασμού αυτών (Κεφάλαιο 5), θα ακολουθήσει μια αναφορά στη θεωρητική προσέγγιση και εκτίμηση των διαστημάτων εμπιστοσύνης και θα προταθεί ο τρόπος υπολογισμού αυτών.

4.2 Επαγωγική Στατιστική

Η **Επαγωγική Στατιστική** είναι μέρος της Στατιστικής, το οποίο παρέχει το θεωρητικό υπόβαθρο για την εξαγωγή συμπερασμάτων για όλο τον πληθυσμό από τις πληροφορίες που προέρχονται από το δείγμα. Στην Επαγωγική Στατιστική μελετούνται δείγματα προκειμένου τα συμπεράσματα που βγαίνουν από αυτά, να γενικευτούν στον αρχικό πληθυσμό από τον οποίο έχουν επιλεγεί. Η θεωρία της στηρίζεται στη θεωρία των Πιθανοτήτων. Για την αποτελεσματική χρησιμοποίηση της Επαγωγικής Στατιστικής αρκεί η γνώση της μεθοδολογίας, των βασικών προϋποθέσεων που πρέπει να πληρούνται για τη χρήση κάποιας τεχνικής και η κατανόηση του τρόπου ανάγνωσης και παρουσίασης των αποτελεσμάτων (Αποστολάκης, Καστανιά και Πιερράκου, 2003).

Ο πρώτος τρόπος επαγωγής είναι η εκτίμηση κάποιων άγνωστων τιμών του πληθυσμού που καλούνται **παράμετροι**, από το δείγμα. Αρχικά υπολογίζονται τα περιγραφικά στατιστικά μέτρα του δείγματος, δηλαδή η μέση τιμή, η διασπορά, κ.λ.π. και στη συνέχεια, προσδιορίζεται κάποια κατάλληλη συνάρτηση για την εκτίμηση μιας παραμέτρου θ. η συνάρτηση αυτή ονομάζεται **εκτιμήτρια συνάρτηση** της παραμέτρου θ του πληθυσμού και συμβολίζεται με $\hat{\theta}$.

Η εκτιμήτρια συνάρτηση είναι:

- Συνάρτηση των παρατηρήσεων του δείγματος, δηλαδή δεν περιέχει άγνωστες παραμέτρους.
- Τυχαία μεταβλητή.
- Ακολουθεί κανονική κατανομή.

Αφού προσδιοριστεί η κατανομή της εκτιμήτριας, στη συνέχεια με βάση τη θεωρία της εκτιμητικής, που είναι αποτέλεσμα του συνδυασμού της Πιθανοθεωρίας και της Δειγματοληψίας, εκτιμώνται οι παράμετροι του πληθυσμού. Συνήθως χρησιμοποιούνται δύο ειδών εκτιμήσεις:

1. **Μονότιμες ή Σημειακές Εκτιμήσεις** (Point Estimations), οι οποίες παρέχουν μια μόνο τιμή για κάθε μία από τις άγνωστες παραμέτρους. Δηλαδή, σημειακή είναι η εκτίμηση, όταν προσδιορίζει συγκεκριμένη τιμή για την παράμετρο του πληθυσμού.

2. **Εκτιμήσεις Διαστημάτων** (Interval Estimations), με τις οποίες καθορίζεται ένα διάστημα τιμών μέσα στο οποίο μπορούμε να πούμε ότι περιλαμβάνεται η άγνωστη

παράμετρος με ορισμένη πιθανότητα. Η πιθανότητα για την εκτίμηση του διαστήματος καθορίζεται από ένα συντελεστή που ονομάζεται **συντελεστής εμπιστοσύνης** (confidence coefficient).

4.2.1 Σημειακή Εκτίμηση

Σημειακή λέγεται η εκτίμηση όταν προσδιορίζει συγκεκριμένη τιμή για την παράμετρο του πληθυσμού. Η τιμή αυτή είναι ένας απλός αριθμός που υπολογίζεται από το δείγμα και που είναι αρκετά κοντά στην άγνωστη παράμετρο του πληθυσμού. Συνήθως υπάρχουν πολλές εκτιμήτριες συναρτήσεις για μια παράμετρο. Το σημαντικό είναι να επιλεχθεί εκείνη η εκτιμήτρια συνάρτηση Τ, η οποία προσεγγίζει όσο το δυνατόν καλύτερα την άγνωστη παράμετρο. Μερικά από τα χρησιμοποιούμενα κριτήρια για την επιλογή της εκτιμήτριας συνάρτησης είναι τα εξής:

- **Αμεροληψία:** Όταν η μέση τιμή της εκτιμήτριας ισούται με την παράμετρο. Συμβολικά $E(T) = \mu$.

- **Ελάχιστο μέσο τετραγωνικό σφάλμα:** Όταν η εκτιμήτρια δίνει την ελάχιστη μέση τετραγωνική απόκλιση από την πραγματική τιμή. Συμβολικά $\min E(T-\theta)^2$.

- **Συνέπεια:** Όταν μεγαλώνοντας απεριόριστα το δείγμα η μέση τιμή της εκτιμήτριας συνάρτησης πλησιάζει όλο και περισσότερο την τιμή της παραμέτρου.

- **Επάρκεια:** Όταν από τις παρατηρήσεις του δείγματος αντλείται επιπλέον πληροφορία για τη συγκεκριμένη παράμετρο, πέραν της γνώσης που υπάρχει μέσω της εκτιμήτριας Τ.

Για παράδειγμα σε μια τυχαία δειγματοληψία, η μέση τιμή του δείγματος είναι γνωστό από τη θεωρία ότι είναι αμερόληπτη εκτιμήτρια της πληθυσμιακής μέσης τιμής. Επομένως, όταν υπολογίζεται η μέση τιμή στο δείγμα, μπορεί να θεωρηθεί ότι αυτή είναι η εκτιμώμενη μέση τιμή στον πληθυσμό. Στην ουσία προτείνεται μια τιμή που είναι αρκετά κοντά στην παράμετρο του πληθυσμού.

4.2.2 Εκτίμηση Διαστήματος

Η σημειακή όμως εκτίμηση, αν και δίνει μια τιμή (ένα σημείο) που πρέπει να είναι κοντά στην υπό εκτίμηση συνάρτηση, δε δίνει καμία ιδέα για την ακρίβεια ή το σφάλμα εκτίμησης. Για να υπάρχει μεγαλύτερη πιθανότητα προσέγγισης των πραγματικών παραμέτρων του πληθυσμού, εκτιμάται ένα διάστημα μέσα στο οποίο θα έχει μεγάλη πιθανότητα να βρίσκεται η παράμετρος του πληθυσμού.

Αν με θ συμβολίζεται η παράμετρος του πληθυσμού, υπάρχει **διάστημα εμπιστοσύνης** (Confidence Interval) (L,U), τέτοιο ώστε η πιθανότητα η παράμετρος να βρίσκεται μέσα σε αυτό το διάστημα ισούται με (1-α), όπου α είναι το **επίπεδο σημαντικότητας**. Το επίπεδο σημαντικότητας α είναι μια μικρή πιθανότητα η οποία επιλέγεται αυθαίρετα και εκφράζει τη πιθανότητα σφάλματος. Συμβολικά ισχύει: $P(L < \theta < U) = 1 - \alpha$, όπου L, U τα άκρα του διαστήματος, τα οποία ονομάζονται **κάτω και άνω όρια**, αντίστοιχα.

4.3 Θεωρητική προσέγγιση των διαστημάτων εμπιστοσύνης

Τα διαστήματα εμπιστοσύνης για μία παράμετρο χρησιμοποιούνται όταν η παράμετρος αυτή δεν μπορεί να θεωρηθεί ως μια σημειακή εκτίμηση. Αντί αυτού μπορεί με μια σχετική βεβαιότητα να ειπωθεί ότι η παράμετρος αυτή βρίσκεται μέσα σε καθορισμένα όρια ανοχής. Αυτό οδηγεί στην ακόλουθη έννοια:

Μια εκτίμηση διαστήματος μιας παραμέτρου θ είναι ένα διάστημα (θ_1, θ_2) τα ακραία σημεία του οποίου είναι συναρτήσεις $\theta_1 = g_1(T)$ και $\theta_2 = g_2(T)$ του διαστήματος παρατήρησης T. Το αντίστοιχο τυχαίο διάστημα (θ_1, θ_2) είναι μια εκτιμήτρια διαστήματος της θ. Θα θεωρείται ότι το (θ_1, θ_2) είναι ένα διάστημα εμπιστοσύνης γ της θ, εάν: $P(\theta_1 < \theta < \theta_2) = \gamma$.

Η σταθερά γ είναι ο συντελεστής εμπιστοσύνης της εκτίμησης και η διαφορά 1-γ=α είναι το επίπεδο σημαντικότητας ή εμπιστοσύνης. Έτσι ο γ είναι ένα υποκειμενικό μέτρο εμπιστοσύνης, ότι η άγνωστη θ βρίσκεται στο διάστημα (θ_1, θ_2). Εάν το γ είναι κοντά στο 1, τότε βεβαιώνεται ότι η άγνωστη παράμετρος βρίσκεται στην πραγματικότητα μέσα στο διάστημα (θ_1, θ_2) και η εκτίμηση της παραμέτρου είναι σωστή σε ποσοστό 100γ% των περιπτώσεων.

Ο αντικειμενικός σκοπός της εκτίμησης διαστήματος είναι ο καθορισμός των συναρτήσεων $\theta_1 = g_1(T)$ και $\theta_2 = g_2(T)$, έτσι ώστε να ελαχιστοποιείται το μήκος $\theta_2 - \theta_1$ του διαστήματος (θ_1, θ_2) για συγκεκριμένες τιμές του γ ή α. Όσο αυξάνεται ο συντελεστής γ, αυξάνονται οι πιθανότητες να βρίσκεται μέσα στο διάστημα εμπιστοσύνης η πραγματική τιμή της εκτιμώμενης παραμέτρου και συνεπώς αυξάνεται και το μήκος του διαστήματος εμπιστοσύνης. Για ένα υποθετικό διάστημα εμπιστοσύνης απείρου μήκους είναι απολύτως βέβαιο ότι η πραγματική τιμή της παραμέτρου είναι μέσα σε αυτό το διάστημα, δηλαδή γ=1. οι πιθανότητες μειώνονται όταν μικραίνει το διάστημα εμπιστοσύνης.

Μια γενική μέθοδος κατασκευής διαστήματος εμπιστοσύνης συντελεστού 1-α είναι η ακόλουθη:

Έστω τυχαίο διάνυσμα $X_1, X_2, ..., X_n$:

1. Υπολογίζεται μια στατιστική συνάρτηση Τ($X_1, X_2, ..., X_n$) της οποίας η κατανομή να εξαρτάται από το θ. Συνήθως ως Τ επιλέγεται μια εκτιμήτρια του θ ή του $g(\theta)$.

2. Κατασκευάζεται συνάρτηση $Y = h(T, g(\theta))$ η κατανομή της οποίας να μην εξαρτάται από το θ.

3. Υπολογίζονται δύο σταθερές θ_1, θ_2 έτσι, ώστε να ισχύει: $P(\theta_1 < \theta < \theta_2) = 1 - a$

4. Εφόσον έχουν εκτιμηθεί τα θ_1, θ_2, επιλύεται η ανισότητα $\theta_1 \leq Y = h(T, g(\theta)) \leq \theta_2$ ως προς $g(\theta)$, έτσι προκύπτει μια νέα ανισότητα της μορφής $L = L(X_1, X_2, ..., X_n) \leq g(\theta) \leq U = U(X_1, X_2, ..., X_n)$

Το παραπάνω ενδεχόμενο θα έχει και αυτό πιθανότητα 1-α και επομένως, το διάστημα (L,U) θα είναι ένα διάστημα εμπιστοσύνης για το $g(\theta)$ συντελεστού 1-α. Τα θ_1, θ_2 συνήθως επιλέγονται έτσι ώστε $P(Y > \theta_2) = P(Y < \theta_1) = \dfrac{a}{2}$. Δηλαδή, το θ_2 είναι το άνω $\dfrac{a}{2}$ όριο της κατανομής της Υ, ενώ το θ_1 είναι το άνω $\left(1 - \dfrac{a}{2}\right)$ όριο της ίδιας κατανομής.

4.4 Υπολογισμός διαστημάτων εμπιστοσύνης για παραμέτρους γνωστών κατανομών

4.4.1 Κανονική κατανομή

4.4.1.1 Διάστημα εμπιστοσύνης για το μέσο κανονικής κατανομής όταν σ^2 γνωστό

Για τυχαία μεταβλητή X, η οποία ακολουθεί κανονική κατανομή με άγνωστη μέση τιμή μ και γνωστή διασπορά σ^2, τότε η μέση τιμή $\overline{X}$ της μεταβλητής X θα ακολουθεί κανονική κατανομή $N\left(\mu, \dfrac{\sigma^2}{n}\right)$ και η μεταβλητή $Z = \dfrac{X - \mu}{\sigma/\sqrt{n}}$, ακολουθεί την τυπική κανονική κατανομή N(0,1). Συνεπώς, το 100(1-α)% διάστημα εμπιστοσύνης για τη μέση τιμή μπορεί να υπολογιστεί χρησιμοποιώντας των πίνακα τιμών της τυπικής κανονικής κατανομής.

$$1 - \alpha = P\left(-z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2} \right)$$

$$\Rightarrow 1 - \alpha = P\left(-\frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \overline{X} - \mu \le \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

$$\Rightarrow 1 - \alpha = P\left(-\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le -\mu \le -\overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

$$\Rightarrow 1 - \alpha = P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \mu \le \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

Άρα το 100(1-α)% διάστημα εμπιστοσύνης για τη μέση τιμή θα είναι:

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \quad \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right]$$

4.4.1.2 Διάστημα εμπιστοσύνης για το μέσο κανονικής κατανομής όταν σ2 άγνωστο

Εφόσον η διακύμανση του πληθυσμού είναι άγνωστη, εκτιμάται με την τιμή της δειγματικής διακύμανσης

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

Όταν η κατανομή πληθυσμού μπορεί να υποτεθεί κανονική και το δείγμα είναι μεγάλο ($n \geq 30$) τότε η τυχαία μεταβλητή $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ ακολουθεί την τυπική κανονική κατανομή. Συνεπώς, το 100(1-α)% διάστημα εμπιστοσύνης για τη μέση τιμή μπορεί να υπολογιστεί όπως καις την πρώτη περίπτωση με τη μόνη διαφορά ότι θα χρησιμοποιηθεί η δειγματική τυπική απόκλιση αντί της τυπικής απόκλισης του πληθυσμού. Επομένως, αν $\overline{X}$ και s^2 είναι η τιμή του μέσου και της διακύμανσης, αντίστοιχα, σε ορισμένο τυχαίο δείγμα μεγέθους n, τότε θα διάστημα εμπιστοσύνης θα είναι:

$$\left[\overline{X} - \frac{s}{\sqrt{n}} z_{1-\alpha/2}, \quad \overline{X} + \frac{s}{\sqrt{n}} z_{1-\alpha/2}\right]$$

Και θα περιέχει την μ με επίπεδο εμπιστοσύνης 1-α.

Σε περίπτωση που το δείγμα είναι μικρό (n<30) τότε η κατανομή πληθυσμού ακολουθεί κατανομή Student αντί κανονικής κατανομής και συμβολίζεται με t_{n-1} (κατανομή t με n-1 βαθμούς ελευθερίας). Επομένως, το διάστημα εμπιστοσύνης θα είναι:

$$\left[\overline{X} - \frac{S}{\sqrt{n}} t_{n-1;\alpha/2}, \quad \overline{X} + \frac{S}{\sqrt{n}} t_{n-1;\alpha/2}\right]$$

4.4.1.3 Διάστημα εμπιστοσύνης για τη διασπορά κανονικής κατανομής

Αν s^2 η διακύμανση τυχαίου δείγματος μεγέθους n από μία κανονική τυχαία μεταβλητή με διακύμανση σ^2, τότε η τυχαία μεταβλητή $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ ακολουθεί την κανονική κατανομή χ^2 με $\nu = n-1$ βαθμούς ελευθερίας. Συνεπώς, το 100(1-α)% διάστημα εμπιστοσύνης για τη διακύμανση μπορεί να υπολογιστεί χρησιμοποιώντας τον πίνακα τιμών της χι-τετράγωνο κατανομής:

Έχουμε:

$$1-\alpha = P\left(\chi^2_{\nu,1-a/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\nu,a/2} \right)$$

$$\Rightarrow 1-\alpha = P\left(\frac{(n-1)s^2}{\chi^2_{\nu,a/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\nu,1-a/2}} \right)$$

Επομένως, αν σε ορισμένο δείγμα n στοιχείων υπολογιστεί η διακύμανση s^2, τότε εκτιμάται ότι το διάστημα που θα περιέχει τη διακύμανση πληθυσμού σ^2 με επίπεδο εμπιστοσύνης (1-α) είναι το:

$$\left[\frac{(n-1)s^2}{\chi^2_{\nu,a/2}}, \frac{(n-1)s^2}{\chi^2_{\nu,1-a/2}} \right]$$

4.4.1.4 Διάστημα εμπιστοσύνης για τη διαφορά των μέσων δύο ανεξάρτητων κανονικών πληθυσμών

Έστω δύο ανεξάρτητα δείγματα $X_1, X_2, ..., X_n$ και $Y_1, Y_2, ..., Y_n$ που ακολουθούν κανονική κατανομή $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, αντίστοιχα. Τα διαστήματα εμπιστοσύνης συντελεστού 1-α για τη διαφορά των μέσων μ_1-μ_2, υπολογίζονται με τη βοήθεια των παρακάτω σχέσεων, ανάλογα με το εάν οι διασπορές είναι γνωστές ή όχι και το μέγεθος του δείγματος.

$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$	σ_1^2, σ_2^2 γνωστά
$\overline{X} - \overline{Y} \pm S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\ t_{n_1+n_2-2;\alpha/2}$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$ άγνωστο $n_1, n_2 < 30$, $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$
$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\ \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]^*$	σ_1^2, σ_2^2 άγνωστα $n_1, n_2 \ge 30$,

4.4.1.5 Διάστημα εμπιστοσύνης για το λόγο των διασπορών δύο ανεξάρτητων κανονικών πληθυσμών

Έστω δύο ανεξάρτητα δείγματα $X_1, X_2, ..., X_n$ και $Y_1, Y_2, ..., Y_n$ που ακολουθούν κανονική κατανομή $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, αντίστοιχα. Τα διαστήματα εμπιστοσύνης συντελεστού 1-α για τη διαφορά των μέσων $\frac{\sigma_2^2}{\sigma_1^2}$, υπολογίζονται με τη βοήθεια της παρακάτω σχέσης, δεδομένου ότι οι μέσες τιμές των κατανομών (μ_1 και μ_2) είναι άγνωστες.

$\left[\frac{S_1^2}{S_2^2} F_{n_2-1,n_1-1;1-\alpha/2}, \ \frac{S_1^2}{S_2^2} F_{n_2-1,n_1-1;\alpha/2}\right]$	$\frac{1}{F_{n_2-1,n_1-1;\alpha/2}} = F_{n_1-1,n_2-1;1-\alpha/2}$

Στην περίπτωση που οι μέσες τιμές των κατανομών (μ_1 και μ_2) είναι γνωστές τότε τα διαστήματα εμπιστοσύνης υπολογίζονται από την παρακάτω σχέση:

$$\left[\frac{n_1 \sum_{i=1}^{n_2} (Y_i - \mu_2)^2}{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2} F_{n_1,n_2\left(1-\frac{a}{2}\right)}, \ \frac{n_1 \sum_{i=1}^{n_2} (Y_i - \mu_2)^2}{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2} F_{n_1,n_2\left(\frac{a}{2}\right)}\right]$$

4.4.2 Διωνυμική Κατανομή

Σε αυτή την κατηγορία παρουσιάζονται η μέθοδος Wilson Score, όπως επίσης και η μέθοδος Wald οι οποίες και θα χρησιμοποιηθούν για τον υπολογισμό των διαστημάτων εμπιστοσύνης στην πειραματική ανάλυση του Κεφαλαίου 6. Αυτό που θα πρέπει να τονιστεί είναι ότι επειδή στα προβλήματα ταξινόμησης εκτιμάται η ακρίβεια ταξινόμησης, δηλαδή η αναλογία σωστών ταξινομήσεων μιας σειράς πειραμάτων τύχης, για αυτό και επιλέγεται ο υπολογισμός διαστημάτων εμπιστοσύνης διωνυμικής κατανομής. Η μέθοδος Clopper-Pearson αποκλείεται από την πειραματική ανάλυση καθώς παρέχει αξιόπιστα διαστήματα εμπιστοσύνης μόνο όταν χρησιμοποιείται μικρό μέγεθος δείγματος.

4.4.2.1 Wald Method

Αν η αναλογία στον πληθυσμό ισούται με p, όπου p το ποσοστό επιτυχίας σε n ανεξάρτητες δοκιμές Bernoulli, τότε η κατανομή της δειγματικής αναλογίας $\hat{p}$ σε τυχαίο δείγμα μεγέθους n, όταν το μέγεθος του δείγματος είναι μεγάλο, $(n \geq 30)$ προσεγγίζει την κατανομή $N\left(p, p(1-p)/n\right)$ και η μεταβλητή $Z = \dfrac{\hat{p} - p}{\sqrt{\hat{p}\left(1-\hat{p}\right)/n}}$ ακολουθεί την τυπική κανονική κατανομή N(0,1). Συνεπώς, το 100(1-α)% διάστημα εμπιστοσύνης για την αναλογία μπορεί να υπολογιστεί χρησιμοποιώντας τον πίνακα τιμών της τυπικής κανονικής κατανομής.

Έχουμε:

$$1-a = P\left(-z_{1-a/2} \leq \frac{\hat{p} - p}{\sqrt{\hat{p}\left(1-\hat{p}\right)/n}} \leq z_{1-a/2} \right) \Rightarrow$$

$$1-a = P\left(-\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2} \leq \hat{p} - p \leq \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2} \right) \Rightarrow$$

$$1-a = P\left(-\hat{p} - \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2} \leq -p \leq -\hat{p} + \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2} \right) \Rightarrow$$

$$1-a = P\left(\hat{p} - \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2} \leq p \leq \hat{p} + \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2} \right)$$

Άρα το 100(1-α)% διάστημα εμπιστοσύνης για τη μέση τιμή θα είναι:

$$\left[\hat{p} - \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2}, \hat{p} + \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} z_{1-a/2} \right]$$

Όπου $\hat{p}$ η αναλογία του στατιστικού δείγματος, n το μέγεθος του δείγματος και $z_{1-a/2}$ προκύπτει από την κανονική κατανομή (Wald and Wolfowitz, 1939).

Η μέθοδος αυτή είναι απλή όσον αφορά τους υπολογισμούς της και δίνει πάντα ικανοποιητικά αποτελέσματα όταν το μέγεθος του δείγματος είναι μεγαλύτερο από 100. Στην περίπτωση που το δείγμα είναι μικρού μεγέθους, η μέθοδος δίνει στενά διαστήματα εμπιστοσύνης. Η μέση κάλυψή της είναι κατά προσέγγιση 60% για επίπεδο εμπιστοσύνης 95%. Αυτό δημιουργεί σε μία εσφαλμένη αίσθηση ακρίβειας σε αυτούς που την χρησιμοποιούν και θα επιθυμούσαν η ονομαστική κάλυψη να συμπίπτει και στην πράξη με την πραγματική. Επίσης, έχει αποδειχθεί ότι η μέθοδος είναι κατάλληλη όταν np>5 και n(1-p)>5 (Brown et. al., 2001).

Επίσης, είναι γνωστό ότι η μέθοδος Wald, μερικές φορές, παρουσιάζει μη ικανοποιητικά αποτελέσματα, ακόμα και εάν το δείγμα είναι αρκετά μεγάλου μεγέθους. Χαρακτηριστικά παραδείγματα, αυτού του συμπεράσματος αποτελούν οι έρευνες των Ghosh (1979) και Blyth και Still (1983). Η διαδικασία εκτίμησης του διαστήματος εμπιστοσύνης είναι αρκετά συντηρητική και αυτό οφείλεται στη διωνυμική κατανομή. Με τον όρο συντηρητική διαδικασία εννοείται ότι η εμπειρική τιμή του συντελεστή εμπιστοσύνης είναι μεγαλύτερη από ονομαστικό επίπεδο 1-α.

4.4.2.2 Wilson Score Method

Η μέθοδος αυτή είναι μια βελτίωση της Wald Method και αναπτύχθηκε από τον Wilson το 1927. Δεν είναι πολύ συντηρητική και παρέχει κάλυψη 95% για επίπεδο εμπιστοσύνης 95%. Σημειώνει ικανοποιητικές τιμές ακόμα για μικρό μέγεθος δείγματος, αλλά παρουσιάζει μεγάλη υπολογιστική πολυπλοκότητα για συγκεκριμένες περιπτώσεις. Τα διαστήματα εμπιστοσύνης αυτής της μεθόδου υπολογίζονται από την παρακάτω σχέση:

$$\frac{\hat{p}+\frac{1}{2n}z_{1-a/2}^{2} \pm z_{1-a/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}+\frac{z_{1-a/2}^{2}}{4n^{2}}}}{1+\frac{1}{n}z_{1-a/2}^{2}}$$

4.4.2.3 Clopper-Pearson Method

Η μέθοδος (Clopper and Pearson, 1934) αυτή παρέχει πιο αξιόπιστα διαστήματα εμπιστοσύνης όταν χρησιμοποιείται μικρό μέγεθος δείγματος. Στην πραγματικότητα, παράγει πολύ συντηρητικά διαστήματα με κάλυψη 99% για επίπεδο εμπιστοσύνης 95%. Αυτό σημαίνει ότι για μικρό μέγεθος δείγματος, τα διαστήματα που προκύπτουν από τη μέθοδο αυτή είναι πολύ ευρεία, ενώ αυτά που προκύπτουν από τη εφαρμογή της μεθόδου Wald είναι πολύ στενά. Τα διαστήματα εμπιστοσύνης αυτής της μεθόδου υπολογίζονται από την παρακάτω σχέση:

$$\left[1+\frac{n-x+1}{xF_{2x,2(n-x+1),1-a/2}}\right]^{-1} < p < \left[1+\frac{n-x}{(x+1)F_{2(x+1),2(n-x),a/2}}\right]^{-1}$$

Επειδή η μέθοδος αυτή χρησιμοποιεί την αθροιστική συνάρτηση της διωνυμικής κατανομής, το διάστημα εμπιστοσύνης μπορεί να γραφεί και ως εξής:

$$\{\theta \mid P[Bin(n;\theta) \le X] \ge a/2\} \cap \{\theta \mid P[Bin(n;\theta) \ge X] \ge a/2\}$$

Όπου X ο αριθμός των επιτυχιών, $Bin(n;\theta)$ τυχαία μεταβλητή που ακολουθεί διωνυμική κατανομή με n δοκιμές και πιθανότητα επιτυχίας θ.

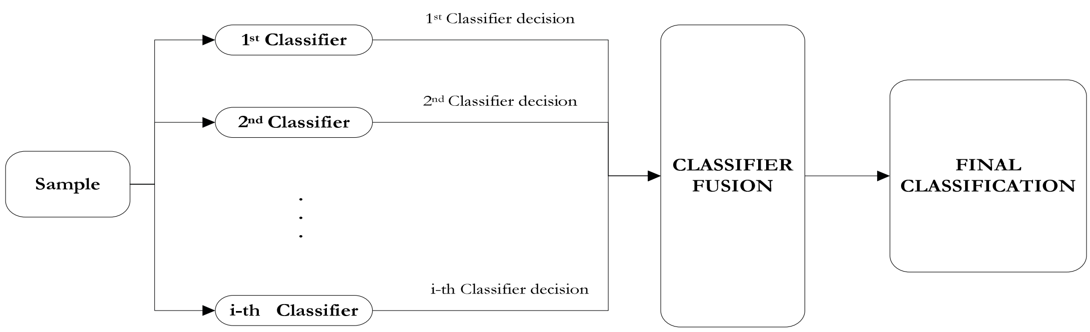
Η μέθοδος αυτή δίνει αρκετά ικανοποιητικά διαστήματα εμπιστοσύνης, καθώς ο υπολογισμός τους βασίζεται στη διωνυμική κατανομή και όχι στην θεώρηση ότι σε μεγάλα μεγέθη δειγμάτων n η κατανομή γίνεται κανονική.

Κεφάλαιο 5ο

ΣΥΝΔΥΑΣΜΟΣ ΤΑΞΙΝΟΜΗΤΩΝ (CLASSIFIER FUSION)

5.1 Εισαγωγή

Τα υποδείγματα ταξινόμησης (Κεφάλαιο 3), ή όπως αλλιώς αναφέρονται ως ταξινομητές, μπορούν να λειτουργήσουν μεμονωμένα. Αυτό σημαίνει ότι μπορούν να εκπαιδευτούν με ένα σύνολο δεδομένων εκμάθησης και εν συνεχεία, να δεχθούν νέα σύνολα δεδομένων ελέγχου, με τα οποία θα εκτιμηθεί η αποτελεσματικότητα των υποδειγμάτων στη διαδικασία της ταξινόμησης των δεδομένων. Πλέον είναι απαραίτητος ο συνδυασμός διαφόρων ταξινομητών, καθώς παρουσιάζουν καλύτερη απόδοση από ότι να αντιμετωπίσουν το πρόβλημα της ταξινόμησης μεμονωμένα (Ruta and Gabryst, 2005). Στο σχήμα 5.1 παρουσιάζεται σχηματικά η δομή Fusion, έτσι ώστε να κατανοηθεί ο όρος Classifier Fusion.



Σχήμα 5. 1: Σχηματική απεικόνιση δομής Classifier Fusion

Είναι αυτονόητο ότι ο συνδυασμός των ταξινομητών που λειτουργούν κάτω από την ίδια λογική μπορεί να μην αποφέρει ουσιώδεις βελτιώσεις στην απόδοση του συνολικού μοντέλου, αλλά από την άλλη να συμβάλλει αισθητά στην αύξηση της πολυπλοκότητας του συστήματος. Στην περίπτωση, όμως, που τα υποδείγματα ταξινόμησης προέρχονται από διαφορετικά πεδία εφαρμογής, ο συνδυασμός αυτών μπορεί να συντελέσει στην αύξηση της αποτελεσματικότητας ταξινόμησης. Γενικότερα, επικρατεί η αντίληψη ότι οι βέλτιστοι συνδυασμοί τεχνικών ταξινόμησης θα πρέπει να έχουν ταυτόχρονα ικανοποιητικές μεμονωμένες αποδόσεις και επαρκή διαφορετικότητα (diversity).

Προκειμένου να εντοπιστεί εκείνος ο συνδυασμός των ταξινομητών που θα αποφέρει τις βέλτιστες αποδόσεις, απαιτείται ο υπολογισμός όλων των δυνατών συνδυασμών. Αυτό βέβαια, θεωρείται πρακτικά δύσκολο, αλλά όχι αδύνατον και συνοδεύεται από υψηλό υπολογιστικό φόρτο, στη περίπτωση δε που ο αριθμός των υπό εξέταση ταξινομητών είναι ιδιαίτερα μεγάλος. Για να μειωθεί λοιπόν, ο υπολογιστικός φόρτος και να επιτευχθεί ο όσο το δυνατόν πιο αξιόπιστος και αποδοτικότερος συνδυασμός ταξινομητών, προτείνεται ο υπολογισμός της διαφορετικότητας (Kuncheva and Whitaker, 2003 and Shipp and Kuncheva, 2002) όλων των πιθανών συνδυασμών και εν συνεχεία, η εφαρμογή των fusion μεθόδων που αποδίδουν συνδυασμό με τη μεγαλύτερη διαφορετικότητα (Kuncheva, Whitaker and Shipp, 2000). Γενικά, συνδυάζοντας τα υποδείγματα ταξινόμησης επιτυγχάνονται πολύ μεγαλύτερες αποδόσεις στην αποτελεσματικότητα της ταξινόμησης, ανεξαρτήτου του πόσο αποδοτικός θεωρείται ότι είναι μεμονωμένα ένας ταξινομητής.

Στη συγκεκριμένη μεταπτυχιακή διατριβή θα μελετηθεί η επιλογή εκείνων των ταξινομητών (Κεφάλαιο 3), που θα αποφέρει τη βέλτιστη απόδοση στην προβληματική της ταξινόμησης. Η επιτυχία του συνδυασμού εξαρτάται κυρίως από το πόσο διαφορετικά (Duin, 2002) είναι τα υποδείγματα ταξινόμησης που πρόκειται να χρησιμοποιηθούν κατά τη διαδικασία Fusion, καθώς και από τη σωστή επιλογή της μεθόδου που θα χρησιμοποιηθεί για το συνδυασμό των επιλεγμένων ταξινομητών. Γενικότερα, υπάρχουν διάφορες προσεγγίσεις με τις οποίες μπορούν να εντοπιστούν ταξινομητές με μεγάλη διαφορετικότητα (diversity). Στην παρούσα διατριβή η προσέγγιση η οποία θα χρησιμοποιηθεί, είναι ο εντοπισμός εκείνων των ταξινομητών από το σύνολο των εξεταζόμενων τεχνικών ταξινόμησης, που όταν συνδυαστούν παρουσιάζουν τη μεγαλύτερη διαφορετικότητα (diversity). Με αυτό τον τρόπο επιτυγχάνεται η μείωση του υπολογιστικού φόρτου, καθώς δεν χρησιμοποιείται τεχνική εξαντλητικής μέτρησης της απόδοσης των ταξινομητών που επιλέχθηκαν τυχαία για κάθε πιθανό συνδυασμό τους.

Γενικότερα, υπάρχουν πολλά κριτήρια εκτίμησης της διαφορετικότητας των υποδειγμάτων ταξινόμησης, τα οποία θα παρουσιαστούν και θα αναλυθούν στη συνέχεια του κεφαλαίου. Προτού όμως, γίνει αυτό πρέπει να γίνουν κάποιες υποθέσεις.

Έστω, λοιπόν, $D=\{D_1,D_2,...,D_L\}$ το σύνολο L ταξινομητών, $C=\{C_1,C_2,...,C_q\}$ το σύνολο από q κατηγορίες (κλάσεις) και τα δείγματα αποτελούνται από n χαρακτηριστικά. Κάθε υπόδειγμα ταξινόμησης δέχεται σαν είσοδο ένα διάνυσμα $x\in\Re^n$, το οποίο δεν αντιστοιχεί απλά σε μία κατηγορία από το σύνολο C, αλλά δίνει για κάθε κλάση την πιθανότητα το δείγμα να ανήκει σε αυτή. Έτσι η έξοδος που προκύπτει για το δείγμα $x\in\Re^n$ για κάθε ταξινομητή ξεχωριστά αναπαριστάται υπό μορφή q-διάστατου διανύσματος $D_i(x)=\begin{bmatrix} d_{i,1}(x) \\ ... \\ d_{i,q}(x)\end{bmatrix}$, όπου $i=1,...,L$ και $d_{i,j}(x)$ η πιθανότητα το δείγμα x να ανήκει στην κατηγορία j, σύμφωνα με τον i ταξινομητή, δεδομένου ότι το διάνυσμα x προέρχεται από την κλάση C_j. Τα $d_{i,j}(x)$ παίρνουν τιμές στο διάστημα [0, 1]. Έτσι, όταν τα $d_{i,j}(x)$ παίρνουν οποιαδήποτε συνεχή τιμή στο διάστημα [0, 1], τότε θεωρούνται ως soft labels, ενώ όταν παίρνουν διακριτές τιμές, 0 ή 1 (ανήκει ή δεν ανήκει σε μία κλάση), τότε θεωρούνται ως hard labels. Ο συνδυασμός των ταξινομητών στοχεύει στην εύρεση της κλάσης που ανήκει το διάνυσμα x, βάσει των εξόδων των L ταξινομητών $D_1(x),D_2(x),...,D_L(x)$. Για το λόγο αυτό υπολογίζεται ένα διάνυσμα $\mu_D(x)=\begin{bmatrix} \mu_D{}^1(x) \\ ... \\ \mu_D{}^q(x)\end{bmatrix}$, το οποίο περιέχει τις q τελικές πιθανότητες για κάθε κατηγορία, χρησιμοποιώντας τα soft labels για το x. Γενικότερα, τα soft labels δίνουν περισσότερη πληροφορία σχετική με την έξοδο του ταξινομητή και επιπλέον, μπορούν να μετατραπούν σε hard labels με την χρήση ενός ορίου (κατωφλιού).

Στην παρούσα διατριβή θα χρησιμοποιηθούν soft labels για το συνδυασμό των ταξινομητών, καθώς οι πιθανότητες που δίνει ο κάθε ταξινομητής για κάθε κατηγορία ταξινόμησης, συμβαδίζουν με τη θεωρητική προσέγγιση και ανάλυση της καμπύλης ROC (Κεφάλαιο 2), αλλά και με τη θεωρητική προσέγγιση της διωνυμικής κατανομής, στην οποία ανήκουν και τα εξεταζόμενα διαστήματα εμπιστοσύνης. Όσον αφορά την καμπύλη ROC (Κεφάλαιο 2) η τιμή του κατωφλιού που θα διαχωρίζει και τις δύο κατηγορίες είναι 0.5.

5.2 Εκτίμηση της διαφορετικότητας των υποδειγμάτων ταξινόμησης

Ένα σημαντικό στοιχείο για να επιτευχθεί βέλτιστη απόδοση, όταν πρόκειται να συνδυαστούν διαφορετικού είδους μεμονωμένοι ταξινομητές, είναι η εκτίμηση της διαφορετικότητας μεταξύ των διαφόρων υποδειγμάτων ταξινόμησης. Στη διεθνή βιβλιογραφία προτείνονται δύο (2) τρόποι υπολογισμού της διαφορετικότητας των ταξινομητών. Στην πρώτη κατηγορία ανήκουν εκείνα τα κριτήρια διαφορετικότητας που εφαρμόζονται σε ζεύγη υποδειγμάτων ταξινόμησης (pairwise diversity measures) και στη δεύτερη κατηγορία ανήκουν εκείνα τα κριτήρια διαφορετικότητας που εφαρμόζονται σε όλο το σύνολο των ταξινομητών (non-pairwise diversity measures)(Kuncheva and Whitaker., 2003 και Shipp and Kuncheva, 2002).

Στην παρούσα διατριβή επιλέχθηκε να χρησιμοποιηθεί ένα pairwise κριτήριο διαφορετικότητας και συγκεκριμένα, το Q Statistic, το οποίο και θα αναλυθεί παρακάτω. Επειδή, όμως σύμφωνα με το Κεφάλαιο 3 πρέπει να εξεταστεί η διαφορετικότητα πέντε (5) τεχνικών ταξινόμησης, τότε για την εύρεση της διαφορετικότητας περισσότερων των δύο ταξινομητών, υπολογίζεται με τα pairwise κριτήρια η διαφορετικότητα για όλα τα πιθανά ζεύγη των ταξινομητών και στη συνέχεια, χρησιμοποιείται ο μέσος όρος αυτών.

5.2.1 Κριτήρια απόδοσης για ζεύγη υποδειγμάτων ταξινόμησης (Pairwise Diversity Measures)

Δεδομένου δύο (2) ταξινομητών, D_i και D_k, οι οποίοι ταξινόμησαν το ίδιο σύνολο δεδομένων, υπάρχουν οι εξής περιπτώσεις:

Οι δύο (2) ταξινομητές να έχουν ταξινομήσει σωστά τα δεδομένα του δείγματος.

Οι δύο (2) ταξινομητές να έχουν ταξινομήσει λάθος τα δεδομένα του δείγματος.

Ο D_i να έχει ταξινομήσει σωστά κάποια δεδομένα, ενώ ο D_k, τα ίδια δεδομένα, να τα έχει ταξινομήσει λάθος και αντίστροφα.

Οι περιπτώσεις αυτές αναλύονται βάσει πιθανοτήτων και καταγράφονται στον πίνακα 5.1.

Πίνακας 5. 1: Κατανομή ταξινομήσεων μεταξύ των δύο ταξινομητών

	D_k (σωστή ταξινόμηση)	D_k (λάθος ταξινόμηση)
D_i (σωστή ταξινόμηση)	A	B
D_i (λάθος ταξινόμηση)	C	D

Όπως είναι λογικό αυτό που ισχύει είναι ότι $A+B+C+D=1$. Τα A, B, C, D είναι ακρίβειες ταξινόμησης, που προκύπτουν από τον συνδυασμό των ταξινομητών και χρησιμοποιούνται προκειμένου να μελετηθούν τα παρακάτω κριτήρια απόδοσης της διαδικασίας Fusion. Σε περίπτωση που υπάρχει μεγάλος αριθμός L ταξινομητών δημιουργούνται $\frac{L(L-1)}{2}$ τιμές για κάθε κριτήριο. Η τελική τιμή του κριτηρίου προκύπτει από τον μέσο όρο των παραπάνω τιμών.

5.2.1.1 Q-Statistic

Για δύο (2) ταξινομητές D_i και D_k, το Q-Statistic (Kuncheva et al.) υπολογίζεται ως το κλάσμα:

$$Q_{i,k} = \frac{AD - BC}{AD + BC}$$

Η τιμή του Q παίρνει τιμές μεταξύ -1 και 1. Θετικές τιμές του Q αντιστοιχούν σε καλές αποδόσεις των ταξινομητών, καθώς τα υποδείγματα ταξινομούν σωστά τα ίδια δείγματα. Το αντίθετο συμβαίνει για ταξινομητές που σημειώνουν αρνητικές τιμές του Q. Όσο μεγαλύτερη είναι η τιμή του Q, τόσο μικρότερη είναι η διαφοροποίηση μεταξύ των ταξινομητών. Για στατιστικά ανεξάρτητους ταξινομητές ισχύει ότι $Q_{i,k} = 0$. Για αριθμό ταξινομητών L υπολογίζεται ο μέσος όρος των τιμών Q από όλα τα ζεύγη ταξινομητών. Συγκεκριμένα, η τιμή του Q για L αριθμό ταξινομητών υπολογίζεται από την παρακάτω σχέση:

$$Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k}$$

5.2.1.2 Συντελεστής συσχέτισης ρ

Για δύο (2) ταξινομητές D_i και D_k, ο συντελεστής συσχέτισης ρ υπολογίζεται ως το κλάσμα:

$$\rho_{i,k} = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Όσο μεγαλύτερη είναι η τιμή του συντελεστή συσχέτισης, τόσο μικρότερη είναι η διαφοροποίηση μεταξύ των ταξινομητών. Για στατιστικά ανεξάρτητους και ασυσχέτιστους ταξινομητές ισχύει ότι $\rho_{i,k} = 0$. Για οποιοσδήποτε δύο ταξινομητές, το Q και το ρ έχουν το ίδιο πρόσημο και μπορεί να αποδειχτεί ότι $|\rho| \le |Q|$.

5.2.1.3 Κριτήριο διαφωνίας

Για δύο (2) ταξινομητές D_i και D_k, το κριτήριο διαφωνίας (disagreement measure) υπολογίζεται από το κλάσμα $DIS_{i,k} = \dfrac{B+C}{A+B+C+D}$ και εκφράζει την πιθανότητα οι δύο ταξινομητές να μην συμφωνούν στις ταξινομήσεις τους για τα ίδια δεδομένα. Όσο μεγαλύτερη είναι η τιμή του κριτηρίου διαφωνίας, τόσο μικρότερη είναι η συσχέτιση μεταξύ των δύο υποδειγμάτων ταξινόμησης, δηλαδή τόσο μεγαλύτερη είναι η διαφοροποίηση μεταξύ των ταξινομητών. Το κριτήριο αυτό χρησιμοποιήθηκε από το Skalak (1996) για να χαρακτηρίσει τη διαφοροποίηση ανάμεσα σε ένα κύριο ταξινομητή και ένα συμπληρωματικό. Επίσης, σημαντική ήταν και η συμβολή του σε μελέτες του Ho (1998).

5.2.1.4 Κριτήριο συμφωνίας

Για δύο (2) ταξινομητές D_i και D_k, το κριτήριο συμφωνίας (agreement measure) υπολογίζεται από το κλάσμα $DAG_{i,k} = \dfrac{A+D}{A+B+C+D}$ και εκφράζει την πιθανότητα οι δύο ταξινομητές να συμφωνούν στις ταξινομήσεις τους για τα ίδια δεδομένα. Όσο μεγαλύτερη είναι η τιμή του κριτηρίου συμφωνίας, τόσο μεγαλύτερη είναι η συσχέτιση μεταξύ των δύο υποδειγμάτων ταξινόμησης. Το πρόβλημα με το κριτήριο συμφωνίας είναι ότι δεν είναι ιδιαίτερα αντιπροσωπευτικό στην εκτίμηση της διαφορετικότητας των υποδειγμάτων ταξινόμησης, καθώς εάν συνδυαστούν δύο (2) πολύ καλοί ταξινομητές είναι αναμενόμενο να ταξινομήσουν σωστά τα περισσότερα δείγματα.

5.2.1.5 Κριτήριο Διπλού Σφάλματος

Για δύο (2) ταξινομητές D_i και D_k, το κριτήριο διπλού σφάλματος (Double-Fault Measure (DF)) υπολογίζεται ως το κλάσμα $DF_{i,k} = \dfrac{D}{A+B+C+D}$ και εκφράζει την πιθανότητα και οι δύο ταξινομητές να έχουν λάθος ταξινομήσεις για τα ίδια δεδομένα. Όσο μεγαλύτερη είναι η τιμή του κριτηρίου διπλού σφάλματος, τόσο μεγαλύτερη είναι η συσχέτιση μεταξύ των δύο υποδειγμάτων ταξινόμησης, δηλαδή τόσο μικρότερη είναι η διαφοροποίηση μεταξύ των ταξινομητών. Αυτό το κριτήριο χρησιμοποιήθηκε από τους Giacinto και Roli (2000) για να συμπληρωθεί ένας πίνακας διαφοροποίησης ταξινομητών ανά ζεύγη και συνεπώς, να επιλεχθούν εκείνοι οι ταξινομητές που είναι λιγότερο σχετικοί.

5.2.1.6 Kappa Statistic

Για δύο (2) ταξινομητές D_i και D_k, το κριτήριο Kappa Statistic υπολογίζεται ως το κλάσμα:

$$K_{i,k} = \frac{2(AD - BC)}{(A+B)(B+D) + (A+C)(C+D)}$$

Το κριτήριο αυτό εκφράζει το βαθμό συμφωνίας των υποδειγμάτων στις ταξινομήσεις των δεδομένων του δείγματος. Όταν η τιμή του κριτηρίου είναι μικρότερη του 0.4, τότε υπάρχει μικρή συμφωνία μεταξύ των ταξινομητών, όταν είναι ανάμεσα στο 0.4 και στο 0.75 υπάρχει μερική συμφωνία, ενώ για τιμές του κριτηρίου μεγαλύτερες του 0.75 υπάρχει μεγάλη συμφωνία μεταξύ των ταξινομητών και επομένως, υψηλή απόδοση της διαδικασίας Fusion.

5.2.2 Κριτήρια απόδοσης για συνδυασμό περισσότερων υποδειγμάτων ταξινόμησης

Δεδομένου ότι υπάρχουν L υποδείγματα ταξινόμησης και επιλέγεται να μην συνδυαστούν ανά ζεύγη, αλλά στο σύνολο τους, αυτό συνεπάγεται την χρήση των παρακάτω κριτηρίων. Θεωρείται ότι N ο αριθμός των παρατηρήσεων του δείγματος δεδομένων και y_{ji} η απόφαση του ταξινομητή i για το αντικείμενο j. Όταν $y_{ji} = 1$ τότε έγινε σωστή ταξινόμηση του αντικειμένου από το υπόδειγμα i, ενώ εάν $y_{ji} = 0$ τότε έγινε λάθος ταξινόμηση του αντικειμένου από το υπόδειγμα i.

5.2.2.1 Εντροπία

Για τους ταξινομητές L, η τιμή του κριτηρίου της εντροπίας υπολογίζεται από τη σχέση:

$$E = \frac{1}{N}\sum_{J=1}^{N}\frac{1}{\left(L-\frac{L}{2}-1\right)}\min\left(\sum_{I=1}^{N} y_{j,i}, N-\sum_{I=1}^{N} y_{j,i}\right)$$

Η τιμή του κριτηρίου παίρνει τιμές μεταξύ 0 και 1. Όταν η τιμή του κριτηρίου είναι 0 τότε υπάρχει μεγάλη συσχέτιση μεταξύ των υποδειγμάτων ταξινόμησης, ενώ όταν η τιμή του κριτηρίου είναι 1 τότε υπάρχει μεγάλη διαφοροποίηση μεταξύ των υποδειγμάτων. Το κριτήριο της εντροπίας παίρνει την τιμή 1, όταν ένα συγκεκριμένο αντικείμενο x_j των δεδομένων του δείγματος, οι μισοί ταξινομητές προβλέψουν σωστά και οι άλλοι μισοί όχι. Επιπλέον, το κριτήριο εντροπίας παίρνει μικρότερη τιμή όταν όλοι οι ταξινομητές δίνουν ακριβώς τις ίδιες ταξινομήσεις, που σημαίνει ότι υπάρχουν μικρές διαφοροποιήσεις μεταξύ των ταξινομητών.

5.2.2.2 Διακύμανση Kohavi-Wolpert

Για τους ταξινομητές L, η τιμή του κριτηρίου της διακύμανσης Kohavi-Wolpert υπολογίζεται από τη σχέση:

$$kw = \frac{1}{NL^2}\sum_{J=1}^{N} l(z_j)(L - l(z_j))$$

Όπου $l(z_j)$ ο αριθμός των υποδειγμάτων ταξινόμησης που πρόβλεψαν σωστά την κλάση του z_j.

5.3 Τεχνικές συνδυασμού υποδειγμάτων ταξινόμησης

Όπως αναφέρθηκε και στα προηγούμενα, υπάρχουν διάφορες μέθοδοι που μπορεί να ακολουθήσει κανείς για να συνδυάσει τα αποτελέσματα των υποδειγμάτων ταξινόμησης, με σκοπό την βελτιστοποίηση της απόδοσης του μοντέλου ταξινόμησης. Δεν είναι πάντοτε ξεκάθαρο ποια μέθοδος θα επιφέρει την βέλτιστη απόδοση, γι' αυτό συνήθως οι ταξινομητές συνδυάζονται με αρκετές μεθόδους και εν συνεχεία, επιλέγεται η βέλτιστη από πλευράς απόδοσης. Υπάρχουν δύο βασικές κατηγορίες στις οποίες χωρίζονται οι μέθοδοι συνδυασμού των ταξινομητών, ως προς το είδος της εισόδου που δέχονται. Στην πρώτη κατηγορία ανήκουν οι μέθοδοι που δέχονται hard labels ως εισόδους, ενώ στην δεύτερη κατηγορία οι μέθοδοι απαιτούν soft labels (Shipp and Kuncheva, 2002).

Στην παρούσα διατριβή επιλέχθηκε να χρησιμοποιηθεί μια soft label μέθοδος και συγκεκριμένα, η Average Rule Method, δεδομένου ότι χρησιμοποιούνται ως δεδομένα εισόδου οι πιθανότητες που δίνει ο κάθε ταξινομητής για κάθε κατηγορία ταξινόμησης.

5.3.1 Hard Labeling Μέθοδοι Συνδυασμού Ταξινομητών

Οι μέθοδοι αυτοί έχουν εκπαιδευτεί ως μεμονωμένοι ταξινομητές, χρησιμοποιούν τα hard labeling των δειγμάτων που έχουν δημιουργηθεί και εξάγουν την τελική απόφαση για κάθε δείγμα, η οποία είναι και αυτή σε hard labeling.

5.3.1.1 Majority Vote

Πρόκειται για μια hard label μέθοδο, με την οποία η τελικά απόφαση για την κατηγορία στην οποία ανήκει το x βασίζεται στην έξοδο που εμφανίζει η πλειοψηφία των ταξινομητών. Για περιττό αριθμό ταξινομητών, η μέθοδος αυτή δεν παρουσιάζει κανένα ιδιαίτερο πρόβλημα, στην αντίθετη περίπτωση και όταν υπάρχει ισοψηφία επιλέγεται τυχαία η κατηγορία ταξινόμησης (Ruta and Gabryst, 2005).

5.3.1.2 Behavior – Knowledge Space (BKS)

Έστω $s_1,...,s_L$ οι αντιστοιχίες hard labels των ταξινομητών $D = \{D_1, D_2,..., D_L\}$, όσον αφορά την ταξινόμηση ενός διανύσματος x. Το s μπορεί να θεωρηθεί ως μια L-διάστατη τυχαία μεταβλητή και η εκτίμηση των πιθανοτήτων γίνεται ως εξής $\mu_i(x) = P(\omega_i \mid s)$. Κάθε πιθανός συνδυασμός των class labels μπορεί να θεωρηθεί ως ένας δείκτης σε ένα κελί ενός BKS πίνακα. Ο πίνακας σχηματίζεται χρησιμοποιώντας ένα labeled σύνολο από δεδομένα x. Κάθε στοιχείο $x_j \in X$ τοποθετείται σε ένα κελί του πίνακα το οποίο υποδεικνύεται από το $D_1(x_j), D_2(x_j),..., D_L(x_j)$. Κατόπιν, γίνεται καταμέτρηση του πλήθους των στοιχείων που υπάρχουν σε κάθε κελί και η πιο αντιπροσωπευτική κλάση επιλέγεται για το συγκεκριμένο κελί. Στην περίπτωση ισοψηφίας η κλάση επιλέγεται τυχαία. Μετά την κατασκευή του πίνακα το διάνυσμα x αντιστοιχίζεται στην κλάση που αντιπροσωπεύει το κελί που υποδεικνύεται από το $D_1(x), D_2(x),..., D_L(x)$.

5.3.1.3 Wernecke's Method (WER)

Η μέθοδος αυτή είναι παρόμοια με τη BKS, η διαφορά έγκειται στο γεγονός ότι για την κατασκευή του πίνακα με τη μέθοδο Wernecke χρησιμοποιείται το 95% των διαστημάτων εμπιστοσύνης των συχνοτήτων σε κάθε κελί. Στην περίπτωση επικάλυψης μεταξύ των διαστημάτων από τους L ταξινομητές, επιλέγεται ο λιγότερο λάθος ταξινομητής, προκειμένου να ταξινομηθεί το x. Για αυτό υπολογίζονται L πιθανότητες, μία για κάθε ταξινομητή. Ο ταξινομητής με την μικρότερη πιθανότητα προσδιορίσει το label του κελιού.

5.3.2 Soft Labeling Μέθοδοι Συνδυασμού Ταξινομητών

Οι μέθοδοι που ακολουθούν χρησιμοποιούν ως εισόδους τα soft labels των μεμονωμένων ταξινομητών. Έστω, λοιπόν, $D = \{D_1, D_2, ..., D_L\}$ το σύνολο L ταξινομητών, $C = \{C_1, C_2, ..., C_q\}$ το σύνολο από q κατηγορίες (κλάσεις) και $\mu_j(x)$ η πιθανότητα αντιστοίχησης του x στην κατηγορία C_q από το συνδυασμό των ταξινομητών, οι soft labeling μέθοδοι λειτουργούν ως εξής:

5.3.2.1 Minimum, Maximum, Average, Product Rule

- **Minimum Rule:** Προκειμένου να βρεθεί η κατηγορία του x εφαρμόζεται ο κανόνας $\mu_j(x) = \min(d_{1,j}(x), ..., d_{L,j}(x)), j = 1, ..., q$. Συγκεκριμένα, το $\mu_j(x)$ βρίσκεται από την ελάχιστη τιμή όλων των εξόδων των ταξινομητών. Ανάλογα με το πιο $\mu_j(x)$ έχει τη μεγαλύτερη τιμή, το διάνυσμα x ανήκει στην αντίστοιχη j κατηγορία.

- **Maximum Rule:** Προκειμένου να βρεθεί η κατηγορία του x εφαρμόζεται ο κανόνας $\mu_j(x) = \max(d_{1,j}(x), ..., d_{L,j}(x)), j = 1, ..., q$. Συγκεκριμένα, το $\mu_j(x)$ βρίσκεται από την μέγιστη τιμή όλων των εξόδων των ταξινομητών. Ανάλογα με το πιο $\mu_j(x)$ έχει τη μεγαλύτερη τιμή, το διάνυσμα x ανήκει στην αντίστοιχη j κατηγορία.

- **Average Rule:** Προκειμένου να βρεθεί η κατηγορία του x εφαρμόζεται ο κανόνας $\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(x), j = 1, ..., q$. Συγκεκριμένα, το $\mu_j(x)$ βρίσκεται από την μέση τιμή όλων των εξόδων των ταξινομητών. Ανάλογα με το πιο $\mu_j(x)$ έχει τη μεγαλύτερη τιμή, το διάνυσμα x ανήκει στην αντίστοιχη j κατηγορία.

- **Product Rule:** Προκειμένου να βρεθεί η κατηγορία του x εφαρμόζεται ο κανόνας $\mu_j(x) = \prod_{i=1}^{L} d_{i,j}(x), j = 1, ..., q$. Συγκεκριμένα, το $\mu_j(x)$ βρίσκεται από το γινόμενο όλων των εξόδων των ταξινομητών. Ανάλογα με το πιο $\mu_j(x)$ έχει τη μεγαλύτερη τιμή, το διάνυσμα x ανήκει στην αντίστοιχη j κατηγορία.

5.3.2.2 Decision Template

Το Decision Template (DT) για κάθε κλάση, υπολογίζονται από τη μέση τιμή των Decision Profiles (DP) όλων των δειγμάτων που ανήκουν σε αυτή την κλάση.

Το Decision Profile (DP) είναι πίνακας διαστάσεων Mxq, όπου οι στήλες αναπαραστούν για κάθε κλάση, την πιθανότητα που δίνει ο κάθε ταξινομητής, το δείγμα x να ανήκει σε αυτήν, ενώ οι γραμμές του πίνακα αφορούν την έξοδο του κάθε ταξινομητή. Είναι αυτονόητο ότι το άθροισμα της κάθε γραμμής ισούται με την μονάδα. Το Decision Profile (DP) έχει την εξής μορφή:

$$DP(x) = \begin{bmatrix} d_{1,1}(x) & \dots & d_{1,j}(x) & \dots & d_{1,q}(x) \\ \dots & & \dots & & \dots \\ d_{i,1}(x) & \dots & d_{i,j}(x) & \dots & d_{i,q}(x) \\ \dots & & \dots & & \dots \\ d_{L,1}(x) & \dots & d_{L,j}(x) & \dots & d_{L,q}(x) \end{bmatrix}$$

Επομένως, το Decision Template (DT) για κάθε κλάση υπολογίζεται από τη σχέση:

$$DT_j = \frac{1}{N_j} \sum_{\substack{z_i \in C_i \\ z_i \in Z}} DP(z_i),\ j = 1,2,\dots,q$$

Το Decision Templates και το Decision Profile είναι πίνακες ίδιων διαστάσεων και αφού κατασκευαστεί το Decision Profile κάθε δείγματος συγκρίνεται με το Decision Template της κάθε κλάσης με τη χρήση κάποιου κριτηρίου ομοιότητας.

Τα ευρέως διαδεδομένα κριτήρια ομοιότητας που χρησιμοποιούνται είναι η ευκλείδεια απόσταση και η συμμετρική διαφορά. Επομένως, η πιθανότητα που προκύπτει από το συνδυασμό των ταξινομητών, να ανήκει το δείγμα x στην κλάση j είναι:

- **Ευκλείδεια απόσταση:** $\mu_j(x) = 1 - \frac{1}{Lq} \sum_{i=1}^{L} \sum_{k=1}^{q} \left(DT_j(i,k) - d_{i,k}(x)\right)^2$

- **Συμμετρική απόσταση:**

$$\mu_j(x) = 1 - \frac{1}{Lq}\sum_{i=1}^{L}\sum_{k=1}^{q}\max\{\min\{(DT_j(i,k),(1-d_{i,k}(x)))\},\min\{((1-DT_j(i,k))-d_{i,k}(x))\}\}$$

Όπου το $DT_j(i,k)$ είναι το (i,k) στοιχείο του DT_j.

Κ ε φ ά λ α ι ο 6°

ΠΕΙΡΑΜΑΤΙΚΗ ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ

6.1 Σκοπός της έρευνας

Στόχος της προτεινόμενης έρευνας είναι να πραγματοποιήσει μια συνολική και διεξοδική ανάλυση της αποτελεσματικότητας του συνδυασμού τεχνικών ταξινόμησης (Classifier Fusion) καλύπτοντας ένα ευρύ φάσμα μεθόδων και δεδομένων. Η ανάλυση αυτή θα συμβάλλει ουσιαστικά:

1. Στον εντοπισμό των αλληλεπιδράσεων που πιθανόν να υπάρχουν στο συνδυασμό των τεχνικών ανάπτυξης υποδειγμάτων ταξινόμησης.

2. Στην εκτίμηση των ορίων ανοχής στην αποτελεσματικότητα των υποδειγμάτων ταξινόμησης, καθώς και του συνδυασμού αυτών, με τη βοήθεια των διαστημάτων εμπιστοσύνης.

3. Στην ανάλυση ευαισθησίας των τεχνικών ταξινόμησης με τη βοήθεια της καμπύλης ROC (Receiver Operating Characteristic, Fawcett, 2003).

Η έννοια της αποτελεσματικότητας αναφέρεται στην αναμενόμενη ακρίβεια των αναπτυσσόμενων υποδειγμάτων ταξινόμησης. Με την υλοποίηση της έρευνας αυτής θα εξαχθούν χρήσιμα συμπεράσματα τα οποία αφορούν όχι μόνο τις χρησιμοποιούμενες τεχνικές ταξινόμησης, αλλά και τις συνθήκες υπό τις οποίες ο συνδυασμός διαφόρων τεχνικών ταξινόμησης μπορεί να συμβάλλει στην καλύτερη κατανόηση της λειτουργίας των εξεταζόμενων τεχνικών ταξινόμησης και στην εξαγωγή χρήσιμων συμπερασμάτων ως προς τις πρακτικές τους εφαρμογές. Επίσης, ο συνδυασμός των διαφόρων τεχνικών θα αποτελέσει χρήσιμο εργαλείο για την αντιμετώπιση προβλημάτων ταξινόμησης στα οποία εντοπίζονται προβλήματα ως προς τις επιδόσεις των παρατηρήσεων στα κριτήρια αξιολόγησης.

6.2 Εξεταζόμενες τεχνικές ταξινόμησης

Όπως προαναφέρθηκε, ένα ακόμα καινοτόμο στοιχείο της προτεινόμενης έρευνας είναι η εξέταση ενός ευρύ συνόλου τεχνικών ταξινόμησης. Οι τεχνικές που χρησιμοποιούνται, επιλέγονται με τέτοιο τρόπο, έτσι ώστε να είναι αντιπροσωπευτικές όλων των διαθέσιμων προσεγγίσεων. Ειδικότερα, εξετάζονται τόσο γνωστές στατιστικές τεχνικές, όσο και μη παραμετρικές προσεγγίσεις από τους χώρους της μηχανικής μάθησης και της επιχειρησιακής έρευνας. Οι στατιστικές προσεγγίσεις αποτελούν τον παραδοσιακό τρόπο ανάπτυξης υποδειγμάτων ταξινόμησης. Οι πλέον διαδεδομένες στατιστικές τεχνικές περιλαμβάνουν η κλασική πιθανοθεωρητική θεωρία του Bayes (Naïve Bayes). Παρά τα προβλήματα που παρουσιάζει η χρήση των τεχνικών αυτών, οι στατιστικές τεχνικές παραμένουν ιδιαίτερα διαδεδομένες και χρησιμοποιούνται συχνά ως σημείο αναφοράς για νέες τεχνικές ταξινόμησης που αναπτύσσονται. Ταυτόχρονα, εξετάζονται και μη παραμετρικές τεχνικές οι οποίες έχουν εξελιχθεί ραγδαία τις τελευταίες δύο δεκαετίες ως αποτελεσματικά εργαλεία για την ανάπτυξη υποδειγμάτων ταξινόμησης. Οι τεχνικές αυτές παρέχουν αυξημένη ευελιξία στον αποφασίζοντα, καθώς δεν βασίζονται σε στατιστικές υποθέσεις και συνεπώς, αναμένεται ότι μπορούν να προσαρμόζονται ικανοποιητικά ανάλογα με τα χρησιμοποιούμενα σύνολα δεδομένων, είτε ως γραμμικά υποδείγματα είτε ως μη γραμμικά. Χαρακτηριστικές μη παραμετρικές τεχνικές, που σε συνδυασμό με τις προαναφερθείσες στατιστικές μεθόδους θα χρησιμοποιηθούν στη συγκεκριμένη πειραματική ανάλυση, είναι ο αλγόριθμος του πλησιέστερου γείτονα (nearest neighbors algorithm, Duda et al., 2001), τα δέντρα λογιστικής παλινδρόμησης (logistic model trees), οι μηχανές διανύσματος υποστήριξης (support vector machines, Vapnik, 1998) και τα ακτινωτό λειτουργικό δίκτυο (radial basis function network).

6.3 Σχεδιασμός του πειράματος

6.3.1 Εισαγωγή

Η υλοποίηση της ανάλυσης πραγματοποιείται σε δύο φάσεις. Αρχικά, γίνεται ένας εκτεταμένος πειραματικός σχεδιασμός ώστε να διερευνηθεί μεμονωμένα η αποτελεσματικότητα των εξεταζόμενων τεχνικών ταξινόμησης σε πραγματικά δεδομένα. Η ανάλυση αυτή βοηθάει στην εξαγωγή χρήσιμων συμπερασμάτων όσον αφορά την ικανότητα των διαφόρων τεχνικών να ελαχιστοποιήσουν το σφάλμα ταξινόμησης. Σε δεύτερη φάση, γίνεται συνδυασμός (Fusion) των εξεταζόμενων τεχνικών ταξινόμησης και εξετάζεται η συνολική ικανότητα αντιμετώπισης προβλημάτων ταξινόμησης και η επίδραση του συνόλου των εξεταζόμενων ταξινομητών μεμονωμένα σε κάθε υπόδειγμα ταξινόμησης. Η πραγματοποίηση της ανάλυσης αυτής συμβάλλει στη διαμόρφωση μιας ολοκληρωμένης εικόνας για την αποτελεσματικότητα των διαφόρων τεχνικών ταξινόμησης, των δυνατοτήτων και τις υπάρχουσες αλληλεπιδράσεις από το συνδυασμό των διαφόρων τεχνικών ανάπτυξης υποδειγμάτων ταξινόμησης. Για την ανάλυση αυτή χρησιμοποιήθηκαν δεδομένα διαφορετικού μεγέθους παρατηρήσεων, αλλά ίδιων κριτηρίων αξιολόγησης, έτσι ώστε να ελεγχθεί η συμπεριφορά και ο υπολογιστικός φόρτος των εξεταζόμενων τεχνικών ταξινόμησης.

6.3.2 Διαδικασία παραγωγής των δεδομένων και υλοποίηση της πειραματικής ανάλυσης

Η υλοποίηση της ανάλυσης πραγματοποιείται σε πραγματικά δεδομένα:

1. Από τη βάση δεδομένων UCI Machine Learning Repository (Blake et al., 1998) τα οποία χρησιμοποιούνται ευρέως σε έρευνες σχετικές με θέματα ταξινόμησης. Το σύνολο των δεδομένων εκπαίδευσης (train set) που χρησιμοποιήθηκε βρίσκεται στη βάση δεδομένων με το όνομα Pima Indians Diabetes, αποτελείται από 770 αντικείμενα και 8 χαρακτηριστικά και αναλύεται μέσω της διαδικασίας ελέγχου 10-fold cross validation. Αντίστοιχα τα δεδομένα ελέγχου ανεξαρτήτου συνόλου (independent test set) είναι διαστάσεων 600x8, τα οποία και θα χρησιμοποιηθούν στη διαδικασία Fusion.

2. Πραγματικά δεδομένα εκπαίδευσης (train set) που σχετίζονται με τη λευχαιμία και περιέχουν 38 ασθενείς (αντικείμενα) και 7129 γονίδια (χαρακτηριστικά) για κάθε ασθενή (δεδομένα εκπαίδευσης). Αντίστοιχα, τα δεδομένα ελέγχου ανεξαρτήτου συνόλου

(independent test set) είναι διαστάσεων 34x7129, τα οποία και θα χρησιμοποιηθούν στη διαδικασία Fusion.

3. Πραγματικά δεδομένα εκπαίδευσης (train set) που σχετίζονται με τον καρκίνο του μαστού και περιέχουν 78 ασθενείς (αντικείμενα) και 20000 γονίδια (χαρακτηριστικά) για κάθε ασθενή (δεδομένα ελέγχου). Αντίστοιχα, τα δεδομένα ελέγχου ανεξαρτήτου συνόλου (independent test set) είναι διαστάσεων 19x20000, τα οποία και θα χρησιμοποιηθούν στη διαδικασία Fusion.

Η υλοποίηση του πειραματικού σχεδιασμού και η ανάπτυξη των εξεταζόμενων μοντέλων πραγματοποιήθηκε με τη χρήση του λογισμικού WEKA 3.5.8 (Kirkby, Frank and Reutemann, 2008), που έχει αναπτυχθεί από το Πανεπιστήμιο Waikato της Νέας Ζηλανδίας και διατίθεται στο διαδίκτυο δωρεάν για εκπαιδευτικούς σκοπούς. Το πρόγραμμα περιλαμβάνει ένα μεγάλο αριθμό μεθόδων και δυνατοτήτων ανάπτυξης μοντέλων, χρησιμοποιώντας ένα εύχρηστο και λειτουργικό περιβάλλον εργασίας. Το λογισμικό Weka βρίσκεται στο διαδίκτυο: http://www.cs.waikato.ac.nz/ml/weka/.

Η επεξεργασία και διαμόρφωση των πραγματικών δεδομένων στο πλαίσιο των απαιτήσεων εφαρμογής ειδικού λογισμικού πραγματοποιήθηκε στο περιβάλλον του Matlab 7.0. Η στατιστική ανάλυση και επεξεργασία των αποτελεσμάτων πραγματοποιήθηκε σε Excel. Η διαδικασία ταξινόμησης του αρχικού δείγματος συμπεριλαμβανομένου και του 10-fold cross validation, επαναλαμβάνεται 10 φορές. Συνεπώς, δημιουργούνται 10 (cross validation) x 10 (replications) διαφορετικά δείγματα εκπαίδευσης και αντίστοιχος αριθμός δειγμάτων ελέγχου ανεξαρτήτου συνόλου.

Το σχήμα 6.1 απεικονίζει τη διαδικασία υλοποίησης του πειραματικού σχεδιασμού και τον τρόπο εξαγωγής των αποτελεσμάτων. Αναλυτικά, η μεθοδολογία που ακολουθήθηκε κατά τη φάση της πειραματικής ανάλυσης είναι η εξής:

A. Χρήση διαδικασιών Επιλογής Χαρακτηριστικών (FS – Feature Selection), με σκοπό το ποσοστό μείωσης της εξεταζόμενης πληροφορίας και τον εντοπισμό του πλέον κατάλληλου και χρήσιμου συνόλου χαρακτηριστικών, τα οποία θα περιγράφουν με ικανοποιητικό βαθμό την ταξινόμηση ενός αντικειμένου. Γενικότερα, η ανάπτυξη και εφαρμογή των FSAs σε τεχνικές ταξινόμησης κρίνεται απαραίτητη, τόσο για την καλύτερη αντιμετώπιση του εξεταζόμενου προβλήματος, όσο και για τη σημαντική μείωση του χρόνου και του κόστους που απαιτούνται για την αντιμετώπισή του.

Στα πλαίσια της συγκεκριμένης ερευνητικής διατριβής εξετάζονται αλγόριθμοι επιλογής χαρακτηριστικών που ανήκουν σε διαδικασίες filter, σύμφωνα με τον τρόπο λειτουργίας τους (Κεφάλαιο 3). Συγκεκριμένα, όπως και προαναφέρθηκε, οι αλγόριθμοι αυτοί εφαρμόζονται πριν την χρήση κάποιας τεχνικής ταξινόμησης και συνεπώς, δεν επηρεάζονται από αυτή. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι εξής:

1. **Genetic Search:** Ερευνά την αναζήτηση του πλέον κατάλληλου συνόλου χαρακτηριστικών χρησιμοποιώντας έναν απλό γενετικό αλγόριθμο, όπως αυτός περιγράφθηκε από τον Goldberg (1989).

2. **Greedy Stepwise:** Ερευνά την αναζήτηση του πλέον κατάλληλου συνόλου χαρακτηριστικών εφαρμόζοντας μία εξαντλητική εμπρόσθια ή ανάστροφη (greedy forward and backward) αναζήτηση. Η διαδικασία αναζήτησης χαρακτηριστικών ξεκινάει με ένα υποσύνολο στο οποίο περιέχεται κανένα ή όλα τα χαρακτηριστικά και σταματάει αντίστοιχα, την πρόσθεση ή την αφαίρεση χαρακτηριστικών από το υποσύνολο όταν τα εναπομείναντα χαρακτηριστικά οδηγούν σε μείωση του κριτηρίου αξιολόγησης της ποιότητας των επιλεγμένων χαρακτηριστικών ή μη αισθητή βελτίωση σε μια ακολουθία συνεχιζόμενων επαναλήψεων του αλγορίθμου.

3. **Ranker:** Κατάταξη των χαρακτηριστικών βάσει της ποιότητας τους, χρησιμοποιώντας διάφορα κριτήρια αξιολόγησης, όπως είναι τα ReliefF, information Gain Ratio, Εντροπία, κ.ά.

Από την εφαρμογή των τριών παραπάνω αλγορίθμων επιλέγονται εκείνα τα χαρακτηριστικά που παρουσιάζουν μεγαλύτερη συχνότητα εμφάνισης. Έτσι διαμορφώνεται το τελικό σύνολο δεδομένων εκπαίδευσης (train set) και δεδομένων ελέγχου (independent test set), το οποίο και θα χρησιμοποιηθεί στο επόμενο βήμα κατά την χρήση των εξεταζόμενων τεχνικών ταξινόμησης. Στον πίνακα 6.1, φαίνονται οι διαστάσεις των πραγματικών δεδομένων και οι διαστάσεις αυτών μετά την εφαρμογή των παραπάνω αλγορίθμων επιλογής χαρακτηριστικών.

Πίνακας 6. 1: Εφαρμογή FSAs για τη μείωση των χαρακτηριστικών των δεδομένων

	Εφαρμογή FSAs για τη μείωση των χαρακτηριστικών των δεδομένων					
Πραγματικά δεδομένα	Pima Indians Diabetes (ασθενείς x χαρακτηριστικά)		Leukemia (ασθενείς x χαρακτηριστικά)		**Breast Cancer** (ασθενείς x χαρακτηριστικά)	
	Αρχικό σύνολο δεδομένων	Μειωμένο σύνολο δεδομένων	Αρχικό σύνολο δεδομένων	Μειωμένο σύνολο δεδομένων	Αρχικό σύνολο δεδομένων	Μειωμένο σύνολο δεδομένων
Train Set	770x8	770x3	38x7129	38x715	78x20000	78x605
Independent Test Set	600x8	600x3	34x7129	34x715	19x20000	19x605

B. Τα μειωμένα δεδομένα εκπαίδευσης (train set) εισάγονται στις εξεταζόμενες τεχνικές ταξινόμησης και εξετάζονται τα αποτελέσματα που προκύπτουν όσον αφορά το πρόβλημα ταξινόμησης. Κατά την εκτίμηση της αποτελεσματικότητας της ταξινόμησης των αναπτυσσόμενων τεχνικών ταξινόμησης εξετάζονται:

1. Η ακρίβεια ταξινόμησης.
2. Το μέτρο F με τη βοήθεια του Confusion Matrix.
3. Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC).
4. Τα διαστήματα εμπιστοσύνης (Confidence Interval) της ακρίβειας ταξινόμησης με τη χρήση των μεθόδων Wald Method και Wilson Score Method για διάφορα επίπεδα σημαντικότητας.

C. Για κάθε τεχνική ταξινόμησης και για όλα τα τρεξίματα (runs) των δεδομένων εκπαίδευσης χρησιμοποιούνται οι εκτιμώμενες ταξινομήσεις των δειγμάτων και οι πιθανότητες που δίνει ο κάθε ταξινομητής, το δείγμα να ανήκει σε μία εκ των δύο κατηγοριών, προκειμένου να υπολογιστεί το κριτήριο διαφορετικότητας Q Statistic των τεχνικών ταξινόμησης και να επιλεχθεί ο καλύτερος συνδυασμός ταξινομητών (Κεφάλαιο 5).

Όπως φαίνεται αναλυτικά και στο σχήμα 6.1 αρχικά υπολογίζεται το Q Statistic για όλα τα ζεύγη των εξεταζόμενων τεχνικών ταξινόμησης (10 πιθανοί συνδυασμοί), προκειμένου να εκτιμηθεί η διαφορετικότητα των τεχνικών ταξινόμησης. Για να υπολογιστεί το Q Statistic για κάθε ζεύγος ταξινομητών πρέπει αρχικά να εκτιμηθεί, όπως φαίνεται και στο κεφάλαιο 5, η συμφωνία ταξινομήσεων του ζεύγους των υποδειγμάτων ταξινόμησης. Συγκεκριμένα, το Q Statistic υπολογίζεται από το κλάσμα $Q_{i,k} = \frac{AD - BC}{AD + BC}$, όπου A το πλήθος των περιπτώσεων που ταξινόμησαν σωστά και οι δύο ταξινομητές, D το πλήθος των περιπτώσεων που ταξινόμησαν λάθος και οι δύο ταξινομητές, B το πλήθος των περιπτώσεων που ταξινόμησε σωστά ο πρώτος ταξινομητής και λάθος ο δεύτερος και C το πλήθος των περιπτώσεων που ταξινόμησε λάθος ο πρώτος ταξινομητής και σωστά ο δεύτερος.

Προκειμένου να ελεγχθεί η διαφορετικότητα περισσότερων από δύο ταξινομητών, έστω L ταξινομητών, υπολογίζεται ο μέσος όρος των Q Statistic από τα ζεύγη των ταξινομήσεων. Επομένως, προκειμένου να εκτιμηθεί η διαφορετικότητα και των υπολοίπων πιθανών συνδυασμών των τεχνικών ταξινόμησης (16 επιπλέον πιθανοί συνδυασμοί ταξινομητών), υπολογίζεται το Q Statistic από τον τύπο $Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k}$. Οι τιμές του Q Statistic για όλους του πιθανούς συνδυασμούς των ταξινομητών, καθώς και η κατάταξη των συνδυασμών από την καλύτερη απόδοση στην χειρότερη με βάσει το Q Statistic φαίνεται στον πίνακα 6.2.

Πίνακας 6. 2: Κατάταξη των πιθανών συνδυασμών των τεχνικών ταξινόμησης με βάσει το Q Statistic

FUSION FOR PIMA INDIAN DIABETES	**Q STATISTIC**	**FUSION FOR BREAST CANCER**	**Q STATISTIC**	**FUSION FOR LEUKEMIA**	**Q STATISTIC**
NB-SVM	0.998	NB-RBFN	0.986	NB-RBFN	1.000
NB-LMT-SVM	0.998	LMT-SVM	0.919	NB-SVM	1.000
NB-LMT	0.998	NB-RBFN-SVM	0.911	NB-IBK	1.000
LMT-SVM	0.998	RBFN-SVM	0.829	SVM-IBK	1.000
NB-LMT-RBFN-SVM	0.988	LMT-RBFN-SVM	0.819	NB-SVM-IBK	1.000
NB-RBFN-SVM	0.986	NB-LMT-RBFN-SVM	0.807	NB-RBFN-SVM-IBK	0.998
NB-LMT-RBFN	0.985	NB-LMT-RBFN	0.775	NB-RBFN-SVM	0.997
NB-RBFN	0.983	NB-LMT-SVM	0.771	RBFN-SVM	0.997
LMT-RBFN-SVM	0.983	NB-SVM	0.767	NB-RBFN-IBK	0.996
RBFN-SVM	0.977	SVM-IBK	0.734	RBFN-SVM-IBK	0.995
LMT-RBFN	0.975	LMT-SVM-IBK	0.730	LMT-SVM	0.994
NB-LMT-RBFN-SVM-IBK	0.893	NB-RBFN-SVM-IBK	0.718	NB-LMT-SVM	0.993
NB-LMT-SVM-IBK	0.872	NB-LMT-RBFN-SVM-IBK	0.710	NB-LMT-RBFN-SVM	0.991
NB-RBFN-SVM-IBK	0.870	LMT-RBFN	0.710	RBFN-IBK	0.989
NB-LMT-RBFN-IBK	0.869	LMT-RBFN-SVM-IBK	0.704	LMT-RBFN-SVM	0.988
LMT-RBFN-SVM-IBK	0.868	RBFN-SVM-IBK	0.686	NB-LMT-RBFN-SVM-IBK	0.987
NB-RBFN-IBK	0.832	NB-LMT-SVM-IBK	0.680	NB-LMT-SVM-IBK	0.986
NB-SVM-IBK	0.831	NB-SVM-IBK	0.666	NB-LMT-RBFN	0.985
NB-LMT-IBK	0.830	NB-RBFN-IBK	0.660	NB-LMT	0.984
LMT-SVM-IBK	0.830	NB-LMT-RBFN-IBK	0.642	LMT-RBFN-SVM-IBK	0.981
RBFN-SVM-IBK	0.830	NB-LMT	0.629	NB-LMT-RBFN-IBK	0.980
LMT-RBFN-IBK	0.828	LMT-RBFN-IBK	0.580	LMT-SVM-IBK	0.977
RBFN-IBK	0.764	NB-LMT-IBK	0.555	NB-LMT-IBK	0.973
NB-IBK	0.748	LMT-IBK	0.537	LMT-RBFN	0.972
SVM-IBK	0.748	NB-IBK	0.498	LMT-RBFN-IBK	0.966
LMT-IBK	0.745	RBFN-IBK	0.494	LMT-IBK	0.936

Με βάση τις παραπάνω κατατάξεις για κάθε εξεταζόμενο δείγμα εκπαίδευσης διαπιστώνεται ότι ο καλύτερος συνδυασμός ταξινομητών θα ήταν Naïve Bayes με RBFNetwork (για τα δεδομένα Breast Cancer και Leukemia) και Naïve Bayes με SVM (για τα δεδομένα Pima Indians Diabetes). Σύμφωνα με τη διεθνή βιβλιογραφία, δε συνίσταται ο συνδυασμός ζυγού αριθμού ταξινομητών, ειδικά όταν πρόκειται για ταξινομήσεις δύο (2) κατηγοριών, καθώς στην περίπτωση που οι μισοί ταξινομητές έχουν ταξινομήσει σωστά και οι άλλοι μισοί λάθος, επιλέγεται τυχαία η εκτιμώμενη ταξινόμηση. Για το σκοπό αυτό θα πραγματοποιηθεί συνδυασμός μονού αριθμού ταξινομητών.

Επομένως, με βάσει τις αποδόσεις του κριτηρίου Q θα συνδυαστούν οι εξής ταξινομητές, όπως και φαίνεται στο σχήμα 6.1: α) Naïve Bayes – RBFNetwork – SVM, β) Naïve Bayes - LMT – RBFNetwork, γ) Naïve Bayes – LMT – SVM, δ) LMT – RBFNetwork – SVM και ε) ALL CLASSIFIERS. Από το συνδυασμό των ταξινομητών εξαιρείται η τεχνική ταξινόμησης IBk, καθώς ο συνδυασμός αυτής με τις υπόλοιπες τεχνικές δεν αποδίδει ικανοποιητικά αποτελέσματα. Οι παραπάνω πέντε (5) συνδυασμοί ταξινομητών χρησιμοποιούνται προκειμένου να εφαρμοστούν στα δεδομένα ελέγχου.

D. Στη φάση αυτή εφαρμόζονται τα δεδομένα ελέγχου μεμονωμένα στις εξεταζόμενες τεχνικές ταξινόμησης και εξετάζονται τα αποτελέσματα που προκύπτουν όσον αφορά το πρόβλημα ταξινόμησης. Κατά την εκτίμηση της αποτελεσματικότητας της ταξινόμησης των αναπτυσσόμενων τεχνικών ταξινόμησης εξετάζονται:

1. Η ακρίβεια ταξινόμησης.
2. Το μέτρο F με τη βοήθεια του Confusion Matrix.
3. Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC).
4. Τα διαστήματα εμπιστοσύνης της ακρίβειας ταξινόμησης με τη χρήση των μεθόδων Wald Method και Wilson Score Method με επίπεδο σημαντικότητας 5%, καθώς από την εξέταση των διαστημάτων εμπιστοσύνης στα δεδομένα εκπαίδευσης παρατηρήθηκε ότι για μικρότερο επίπεδο σημαντικότητας το εύρος των διαστημάτων εμπιστοσύνης ήταν αρκετά μεγάλο, ενώ για επίπεδο εμπιστοσύνης 5% ή 10% το εύρος των διαστημάτων εμπιστοσύνης μειώνονταν. Ανάμεσα στις δύο (2)

τελευταίες τιμές του επιπέδου εμπιστοσύνης επιλέχθηκε το 5% γιατί πρώτον, παρουσιάζει, όπως και θα διαπιστωθεί στην παρακάτω ανάλυση, μικρές διαφοροποιήσεις εν συγκρίσει με το 10% στο εύρος των διαστημάτων και δεύτερον, είναι ευρέως χρησιμοποιούμενο στη διεθνή βιβλιογραφία.

Τα διαστήματα εμπιστοσύνης υπολογίζονται από τη σχέση:

- Wilson Score Method : $\dfrac{\hat{p}+\dfrac{1}{2n}z_{1-a/2}^{2}\pm z_{1-a/2}\sqrt{\dfrac{\hat{p}\left(1-\hat{p}\right)}{n}+\dfrac{z_{1-a/2}^{2}}{4n^{2}}}}{1+\dfrac{1}{n}z_{1-a/2}^{2}}$

- Wald Method: $\left[\hat{p}-\sqrt{\dfrac{\hat{p}\left(1-\hat{p}\right)}{n}}z_{1-a/2},\hat{p}+\sqrt{\dfrac{\hat{p}\left(1-\hat{p}\right)}{n}}z_{1-a/2}\right]$

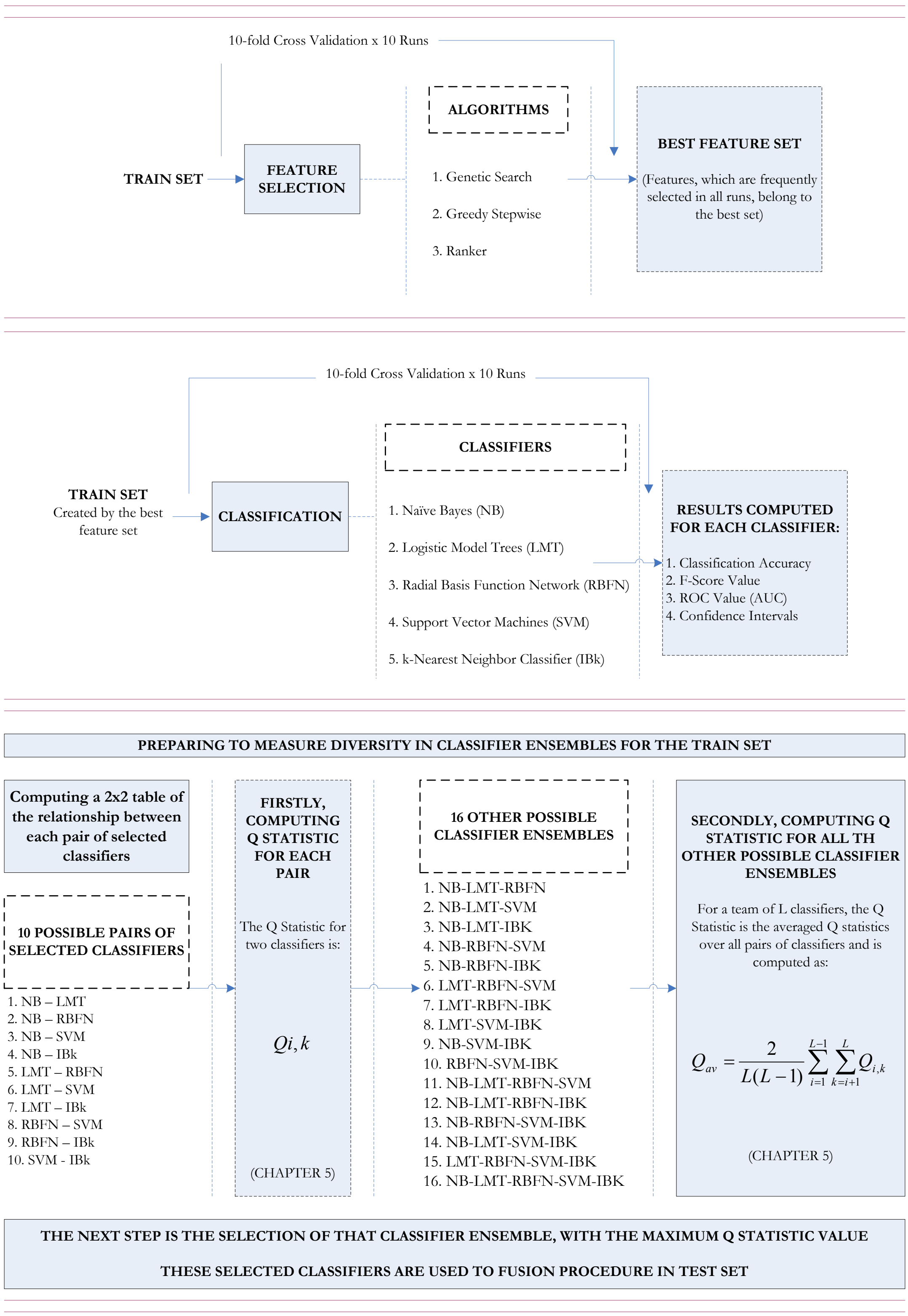
Όπου p η εκτιμώμενη ακρίβεια ταξινόμησης.

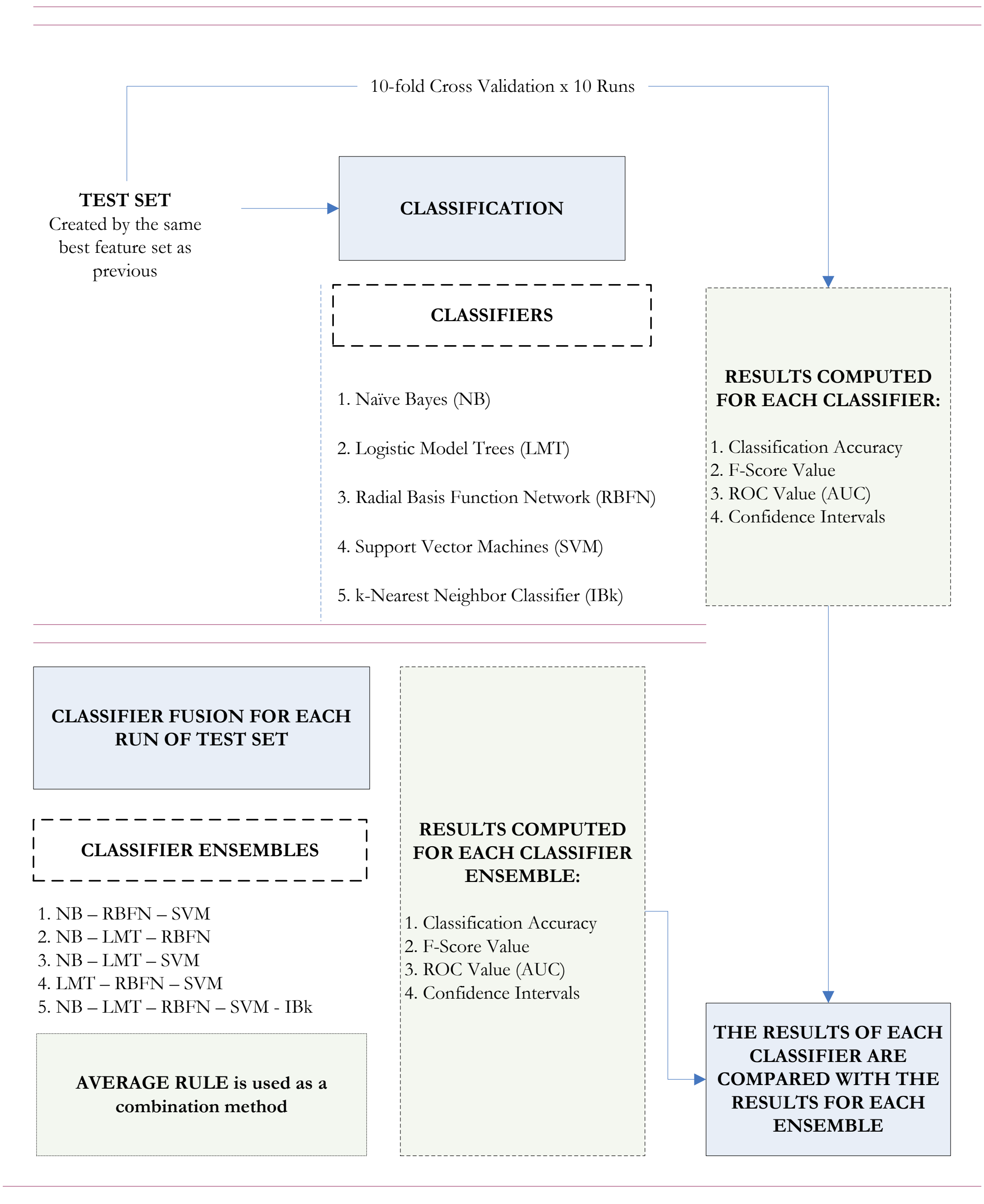
Ε. Οι πέντε (5) συνδυασμοί ταξινομητών που επιλέχθηκαν παραπάνω με το κριτήριο απόδοσης Q, πραγματοποιούνται με το Average Rule (Κεφάλαιο 5) και αφού υπολογιστούν οι εκτιμώμενες ταξινομήσεις των δειγμάτων ελέγχου και οι πιθανότητες που δίνει ο κάθε συνδυασμός το δείγμα να ανήκει σε μία εκ των δύο κατηγοριών, εκτιμώνται:

1. Η ακρίβεια ταξινόμησης.
2. Το μέτρο F με τη βοήθεια του Confusion Matrix.
3. Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC).
4. Τα διαστήματα εμπιστοσύνης (Confidence Interval) της ακρίβειας ταξινόμησης με τη χρήση των μεθόδων Wald Method και Wilson Score Method με επίπεδο σημαντικότητας 5%.

Όσον αφορά τις πιθανότητες που δίνει ο κάθε ταξινομητής, η τιμή του κατωφλιού που διαχωρίζει και τις δύο κατηγορίες είναι 0.5, γεγονός που συμβαδίζει και με την καμπύλη ROC (Κεφάλαιο 2), αλλά και με τη θεωρητική προσέγγιση της διωνυμικής κατανομής, στην οποία ανήκουν και τα εξεταζόμενα διαστήματα εμπιστοσύνης. Για το λόγο αυτό επιλέχθηκε να γίνει Fusion των ταξινομητών με τη βοήθεια του Average Rule, καθώς ο συγκεκριμένος κανόνας έχει να κάνει με τις πιθανότητες ταξινόμησης στις προκαθορισμένες κατηγορίες. Το Average Rule, όπως αναλύεται και στο Κεφάλαιο 5, υπολογίζεται από τη σχέση $\mu_j(X) = \frac{1}{M}\sum_{i=1}^{M} d_{i,j}(X),\ j = 1,...,q$.

F. Τέλος, συγκρίνονται τα αποτελέσματα των δειγμάτων ελέγχου που προκύπτουν μεμονωμένα από τους εξεταζόμενους ταξινομητές και από τους διάφορους συνδυασμούς των τεχνικών ταξινόμησης.


10-fold Cross Validation x 10 Runs
TRAIN SET
FEATURE SELECTION
ALGORITHMS
1. Genetic Search
2. Greedy Stepwise
3. Ranker
BEST FEATURE SET
(Features, which are frequently selected in all runs, belong to the best set)
10-fold Cross Validation x 10 Runs
TRAIN SET
Created by the best feature set
CLASSIFICATION
CLASSIFIERS
1. Naïve Bayes (NB)
2. Logistic Model Trees (LMT)
3. Radial Basis Function Network (RBFN)
4. Support Vector Machines (SVM)
5. k-Nearest Neighbor Classifier (IBk)
RESULTS COMPUTED FOR EACH CLASSIFIER:
1. Classification Accuracy
2. F-Score Value
3. ROC Value (AUC)
4. Confidence Intervals
PREPARING TO MEASURE DIVERSITY IN CLASSIFIER ENSEMBLES FOR THE TRAIN SET
Computing a 2x2 table of the relationship between each pair of selected classifiers
10 POSSIBLE PAIRS OF SELECTED CLASSIFIERS
1. NB – LMT
2. NB – RBFN
3. NB – SVM
4. NB – IBk
5. LMT – RBFN
6. LMT – SVM
7. LMT – IBk
8. RBFN – SVM
9. RBFN – IBk
10. SVM - IBk
FIRSTLY, COMPUTING Q STATISTIC FOR EACH PAIR
The Q Statistic for two classifiers is:
Qi,k
(CHAPTER 5)
16 OTHER POSSIBLE CLASSIFIER ENSEMBLES
1. NB-LMT-RBFN
2. NB-LMT-SVM
3. NB-LMT-IBK
4. NB-RBFN-SVM
5. NB-RBFN-IBK
6. LMT-RBFN-SVM
7. LMT-RBFN-IBK
8. LMT-SVM-IBK
9. NB-SVM-IBK
10. RBFN-SVM-IBK
11. NB-LMT-RBFN-SVM
12. NB-LMT-RBFN-IBK
13. NB-RBFN-SVM-IBK
14. NB-LMT-SVM-IBK
15. LMT-RBFN-SVM-IBK
16. NB-LMT-RBFN-SVM-IBK
SECONDLY, COMPUTING Q STATISTIC FOR ALL TH OTHER POSSIBLE CLASSIFIER ENSEMBLES
For a team of L classifiers, the Q Statistic is the averaged Q statistics over all pairs of classifiers and is computed as:
$Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k}$
(CHAPTER 5)
THE NEXT STEP IS THE SELECTION OF THAT CLASSIFIER ENSEMBLE, WITH THE MAXIMUM Q STATISTIC VALUE
THESE SELECTED CLASSIFIERS ARE USED TO FUSION PROCEDURE IN TEST SET



Σχήμα 6. 1: Ακολουθία βημάτων για την υλοποίηση της πειραματικής ανάλυσης

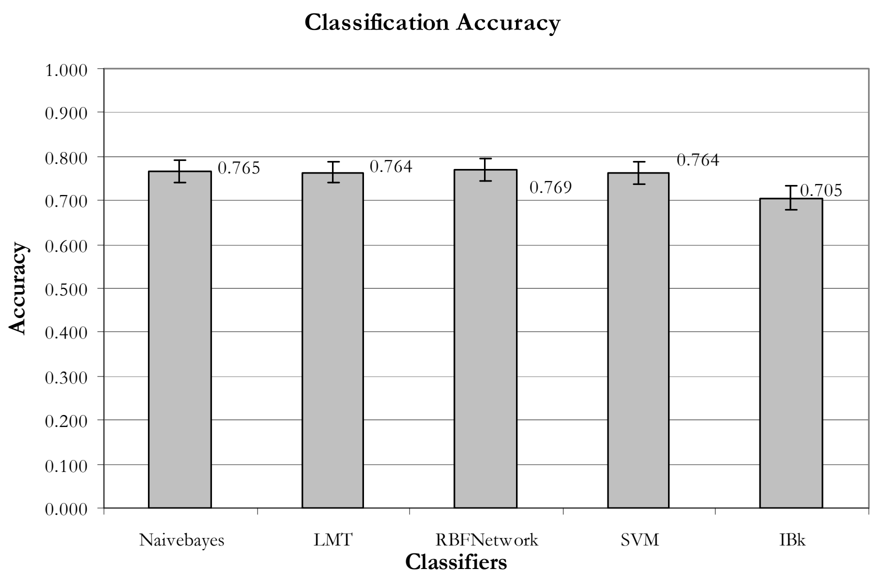
6.3 Ανάλυση των αποτελεσμάτων

Η ανάλυση των αποτελεσμάτων που επιτυγχάνεται για κάθε μέθοδο ταξινόμησης αφορά το ποσοστό των σωστών ταξινομήσεων των δειγμάτων, καθώς αυτά εξηγούν πληρέστερα την ικανότητα των αντίστοιχων μεθόδων στην αντιμετώπιση προβλημάτων ταξινόμησης.

6.3.1 Δεδομένα Pima Indian Diabetes

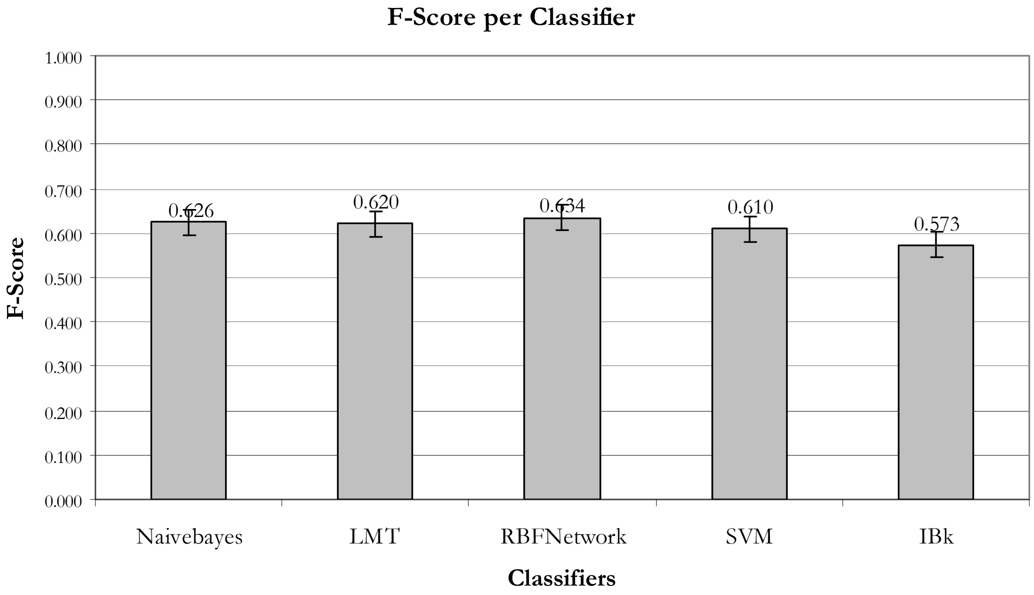
6.4.1.1 Ανάλυση αποτελεσμάτων μεμονωμένα για κάθε τεχνική ταξινόμησης (Δεδομένα: 770x3)

Στο σχήμα 6.2 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων όλων των εξεταζόμενων τεχνικών στο δείγμα εκπαίδευσης. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζει η προσέγγιση RBFNetwork (76.9%) έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Σημαντικά μεγαλύτερο σφάλμα ταξινόμησης από την RBFNetwork (76.9%) παρουσιάζει η προσέγγιση IBk (70.5%). Οι υπόλοιπες τρεις προσεγγίσεις Naïve Bayes (76.5%), LMT (76.4%) και SVM (76.4%) παρουσιάζουν μικρές διαφοροποιήσεις όσον αφορά την ακρίβεια ταξινόμησης. Επομένως, μπορεί να θεωρηθεί ότι οι εν λόγω τεχνικές ταξινόμησης αποδίδουν παρόμοια αποτελέσματα.



Σχήμα 6. 2: Ποσοστό ακρίβειας ταξινόμησης για τα δεδομένα Pima Indian Diabetes

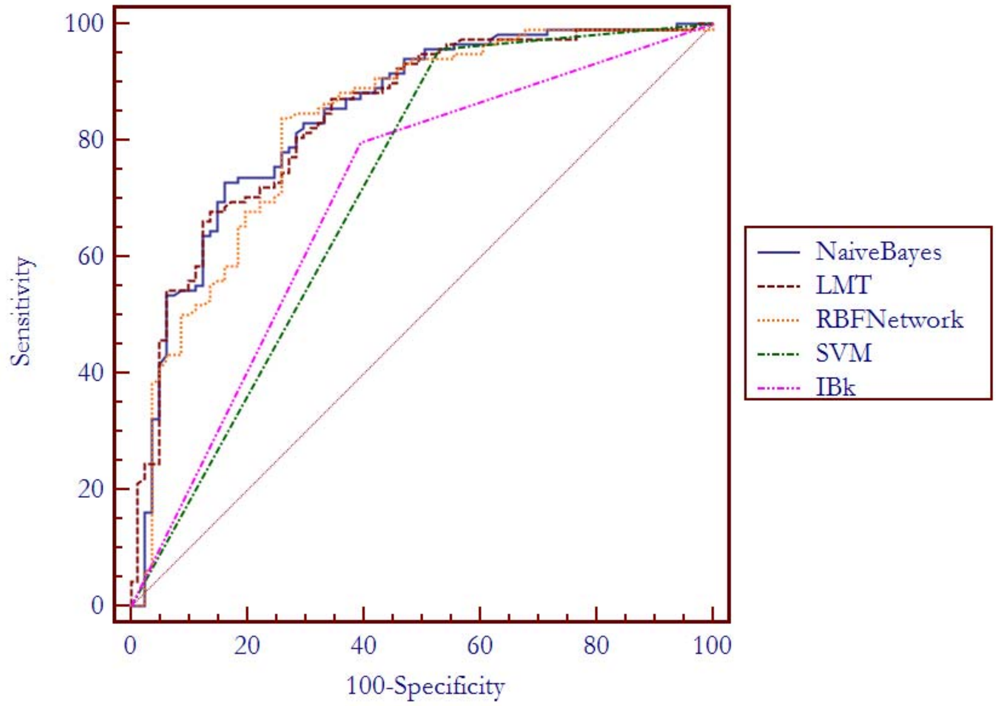
Στο σχήμα 6.3 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με την ικανότητα ταξινόμησης όλων των εξεταζόμενων τεχνικών στο δείγμα εκπαίδευσης, όπως εκτιμάται με τη βοήθεια του μέτρου F. Τα αποτελέσματα δείχνουν την υψηλή ικανότητα ταξινόμησης που παρουσιάζει η προσέγγιση RBFNetwork (63.4%) έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Σημαντικά τη μικρότερη τιμή του μέτρου παρουσιάζει η προσέγγιση IBk (57.3%). Οι υπόλοιπες τρεις προσεγγίσεις με αυστηρή σειρά, Naïve Bayes (62.6%), LMT (62%) και SVM (61%), παρουσιάζουν μία μέση απόδοση στην ικανότητα ταξινόμησης.



Σχήμα 6. 3: Εκτίμηση F-Score για τα δεδομένα Pima Indian Diabetes

Στο σχήμα 6.4 εξετάζεται η αποτελεσματικότητα των υποδειγμάτων ταξινόμησης με τη βοήθεια του εμβαδού της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC). Προφανώς, το εμβαδόν αυτό για ένα άριστο υπόδειγμα είναι ίσο με τη μονάδα, ενώ για ένα ανεπαρκές υπόδειγμα είναι μικρότερο από 0.5. Όσον αφορά τις εξεταζόμενες τεχνικές ταξινόμησης η μέθοδος IBk δίνει την χαμηλότερη τιμή εμβαδού (AUC=0.701), γεγονός που αποδεικνύει την αδυναμία της μεθόδου να αποδώσει υψηλά ποσοστά ακρίβειας. Ακολουθεί η τεχνική SVM (AUC=0.713) με μέτρια απόδοση ακρίβειας. Σαφώς υψηλότερες ακρίβειες ταξινόμησης σημειώνουν οι τεχνικές Naïve Bayes (AUC=0.843) και LMT

(AUC=0.842), με μικρές διαφοροποιήσεις και ακολουθεί η RBFNetwork (AUC=0.828) με εξίσου ικανοποιητική απόδοση.



Σχήμα 6. 4: Σχηματική απεικόνιση καμπύλης ROC
για τα δεδομένα Pima Indian Diabetes

Πίνακας 6. 3: Αποτελέσματα καμπύλης ROC για τα δεδομένα Pima Indian Diabetes

Classifiers	**AUC**	**95% CI**
NaiveBayes	0.843	0.785 to 0.890
LMT	0.842	0.784 to 0.890
RBFNetwork	0.828	0.768 to 0.877
SVM	0.713	0.645 to 0.775
IBk	0.701	0.632 to 0.763

Στο σύνολο των σχημάτων 6.5 εξετάζονται τα διαστήματα εμπιστοσύνης της ακρίβειας ταξινόμησης των εξεταζόμενων υποδειγμάτων ταξινόμησης. Για τον υπολογισμό των διαστημάτων εμπιστοσύνης, όπως και προαναφέρθηκε, χρησιμοποιήθηκαν δύο μέθοδοι εκτίμησης αυτών (Wald Method και Wilson Score Method) και υπολογίστηκαν τα διαστήματα εμπιστοσύνης για διάφορους συντελεστές εμπιστοσύνης. Συγκεκριμένα, στα σχήματα απεικονίζονται τα διαστήματα εμπιστοσύνης της ακρίβειας των ταξινομητών για συντελεστή εμπιστοσύνης 99.5% (α=0.005, 1η δεκάδα κάθε διαγράμματος), 97.5% (α=0.025, 2η δεκάδα κάθε διαγράμματος), 95% (α=0.05, 3η δεκάδα κάθε διαγράμματος) και 90% (α=0.1, 4η και τελευταία δεκάδα κάθε διαγράμματος). Από τα παρακάτω σχήματα προκύπτουν οι εξής διαπιστώσεις:

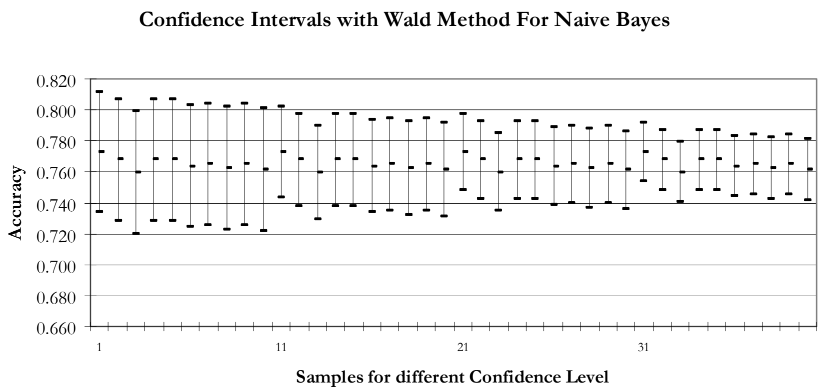
1. Οι δύο εξεταζόμενες μέθοδοι εκτίμησης των διαστημάτων εμπιστοσύνης, Wald Method και Wilson Score Method, παρουσιάζουν παρόμοιες εκτιμήσεις με μηδαμινές διαφοροποιήσεις μεταξύ τους.

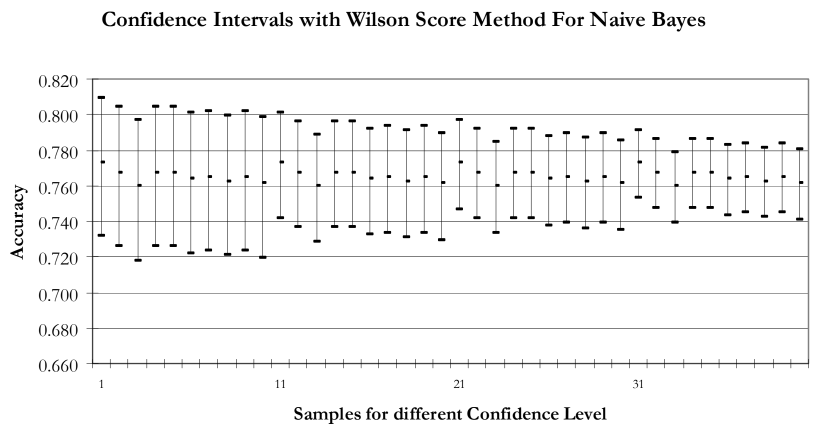
2. Όσο μικραίνει ο συντελεστής εμπιστοσύνης τόσο μικραίνει και το εύρος των διαστημάτων εμπιστοσύνης. Συγκεκριμένα, για συντελεστή εμπιστοσύνης 99.5% εκτιμάται το εύρος του διαστήματος περίπου 0.08, για 97.5% περίπου 0.06, για 95% περίπου 0.05, ενώ για 90% περίπου 0.04. Συνεπώς, ο συντελεστής εμπιστοσύνης επηρεάζει τα άκρα του διαστήματος εμπιστοσύνης και όσο αυξάνεται ο συντελεστής, αυξάνονται και οι πιθανότητες να βρίσκεται μέσα στο διάστημα εμπιστοσύνης η πραγματική τιμή της εκτιμώμενης παραμέτρου.

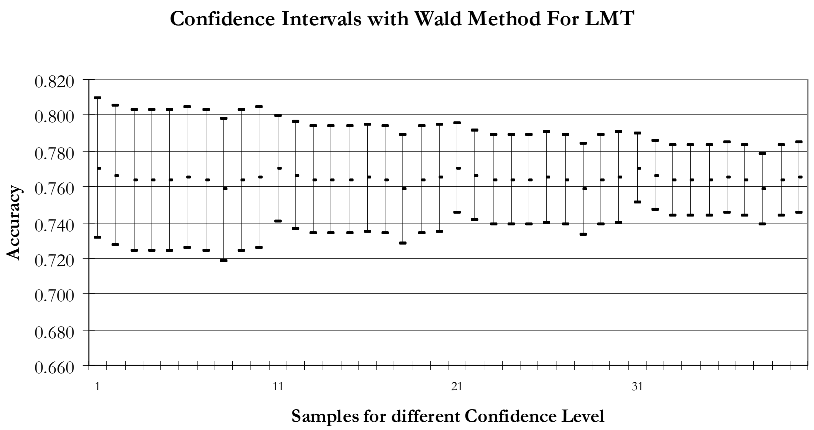
3. Όσον αφορά τις χρησιμοποιούμενες τεχνικές ταξινόμησης και για συντελεστή εμπιστοσύνης 95% με τη χρήση της μεθόδου Wald, η πραγματική διαφορά στην εκτίμηση της μέγιστης από την ελάχιστη τιμή ακρίβειας, που μπορεί να αποδώσει το κάθε υπόδειγμα ταξινόμησης για τα πραγματικά δεδομένα Pima Indian Diabetes, βρίσκεται στο διάστημα, κατά προσέγγιση μέση τιμής, [0.740, 0.790] για την τεχνική Naïve Bayes, [0.739, 0.789] για την LMT, [0.744, 0.794] για την τεχνική RBFNetwork, [0.739, 0.789] για την SVM και [0.678, 0.732] για την τεχνική IBk.

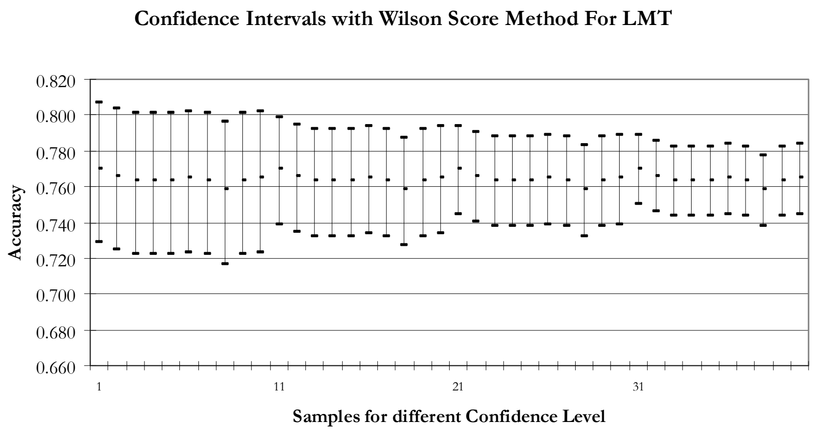
4. Δε σημειώνεται σταθερή μεταβολή των ακρών των διαστημάτων εμπιστοσύνης για όλες τις τεχνικές ταξινόμησης.

Και από την εκτίμηση των διαστημάτων εμπιστοσύνης συμπεραίνεται ότι η τεχνική IBk δεν είναι ιδιαίτερα αξιόπιστη ως προς την αντιμετώπιση προβλημάτων ταξινόμησης. Σαφώς καλύτερα αποτελέσματα αποδίδουν οι υπόλοιπες τεχνικές, με την RBFNetwork να υπερτερεί έναντι των τεχνικών Naïve Bayes, LMT και SVM με μικρή διαφορά.

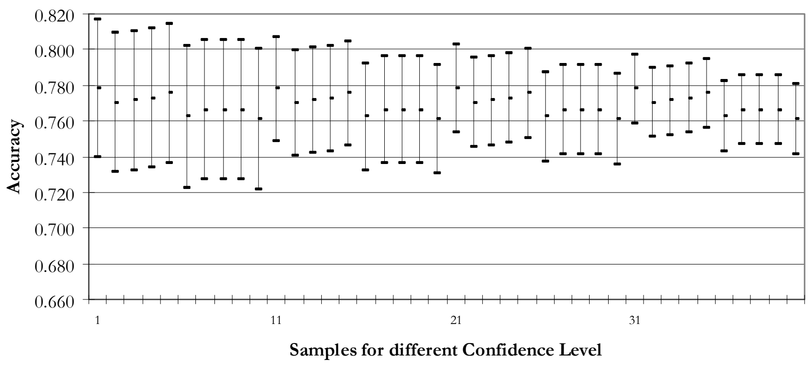

Confidence Intervals with Wald Method For Naive Bayes
0.820
0.800
0.780
0.760
0.740
0.720
0.700
0.680
0.660
Accuracy
1
11
21
31
Samples for different Confidence Level


Confidence Intervals with Wilson Score Method For Naive Bayes
0.820
0.800
0.780
0.760
0.740
0.720
0.700
0.680
0.660
Accuracy
1
11
21
31
Samples for different Confidence Level

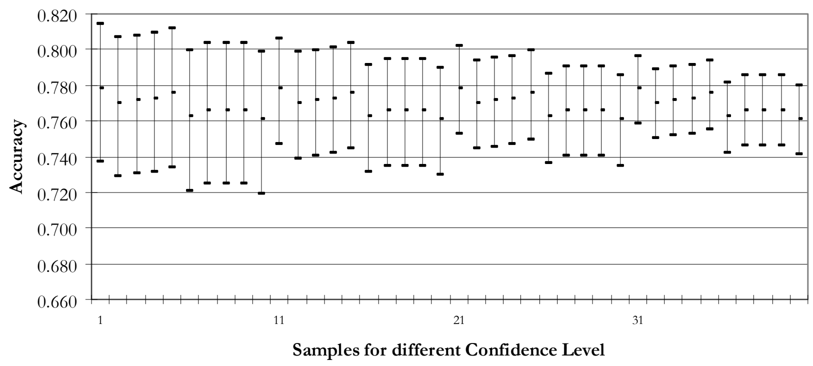

Confidence Intervals with Wald Method For LMT
0.820
0.800
0.780
0.760
0.740
0.720
0.700
0.680
0.660
Accuracy
1
11
21
31
Samples for different Confidence Level


Confidence Intervals with Wilson Score Method For LMT
0.820
0.800
0.780
0.760
0.740
0.720
0.700
0.680
0.660
Accuracy
1
11
21
31
Samples for different Confidence Level

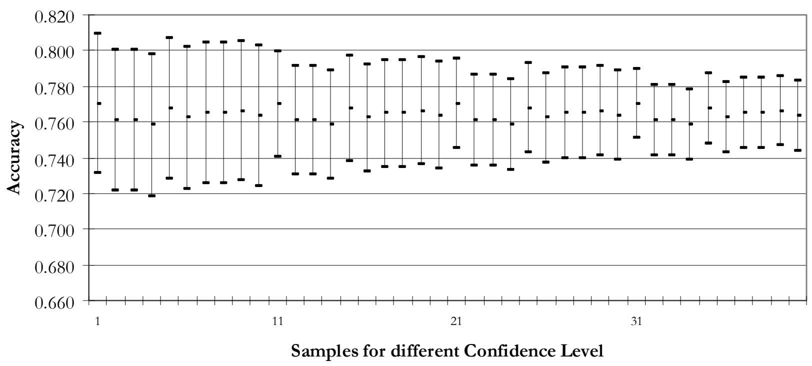
Confidence Intervals with Wald Method For RBFNetwork



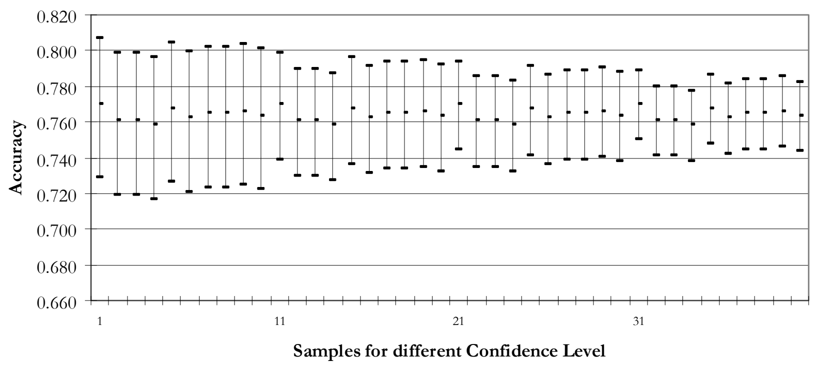
Confidence Intervals with Wilson Score Method For RBFNetwork

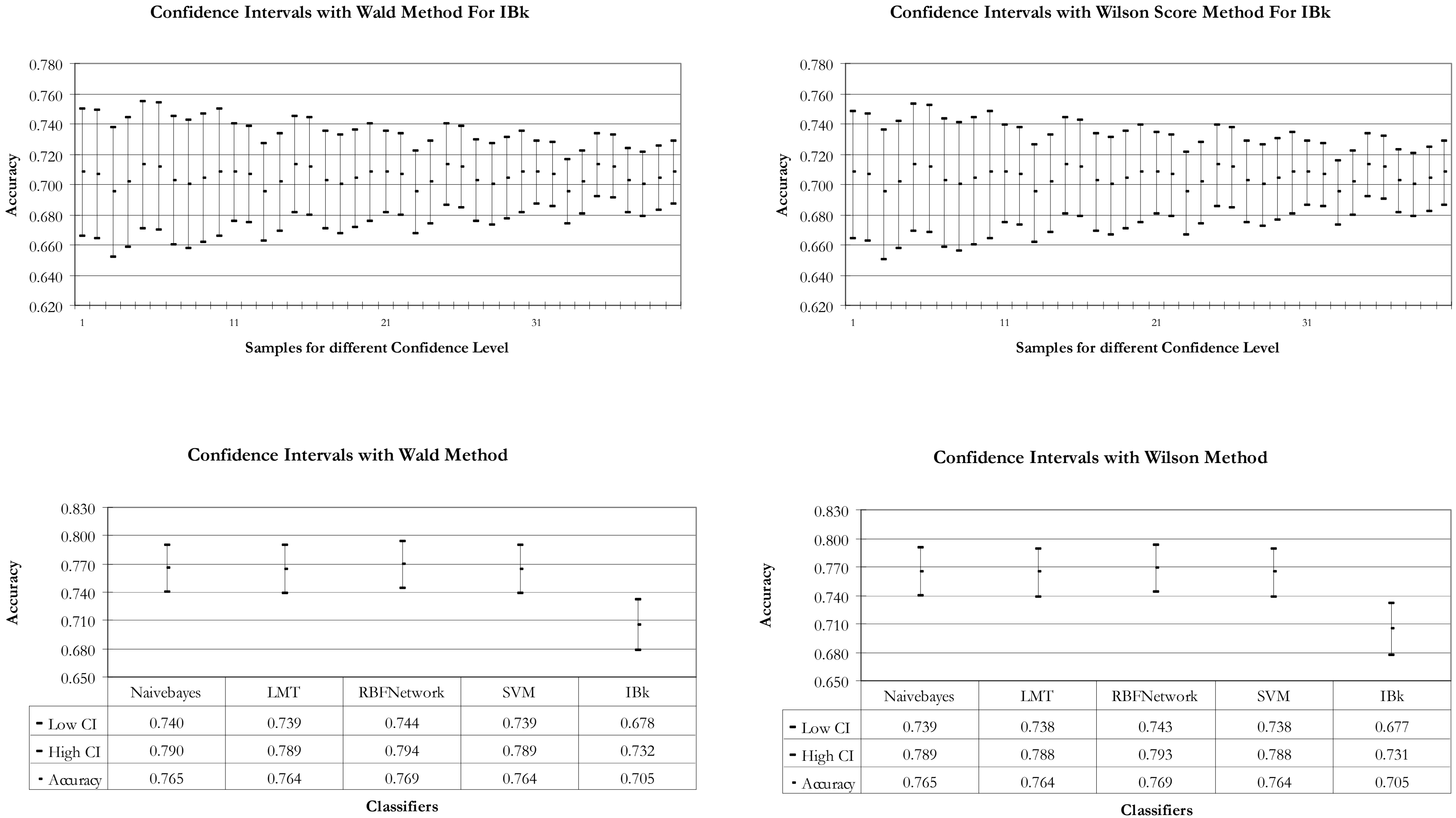


Confidence Intervals with Wald Method For SVM



Confidence Intervals with Wilson Score Method For SVM





Confidence Intervals with Wald Method

	Naivebayes	LMT	RBFNetwork	SVM	IBk
Low CI	0.740	0.739	0.744	0.739	0.678
High CI	0.790	0.789	0.794	0.789	0.732
Accuracy	0.765	0.764	0.769	0.764	0.705

Confidence Intervals with Wilson Method

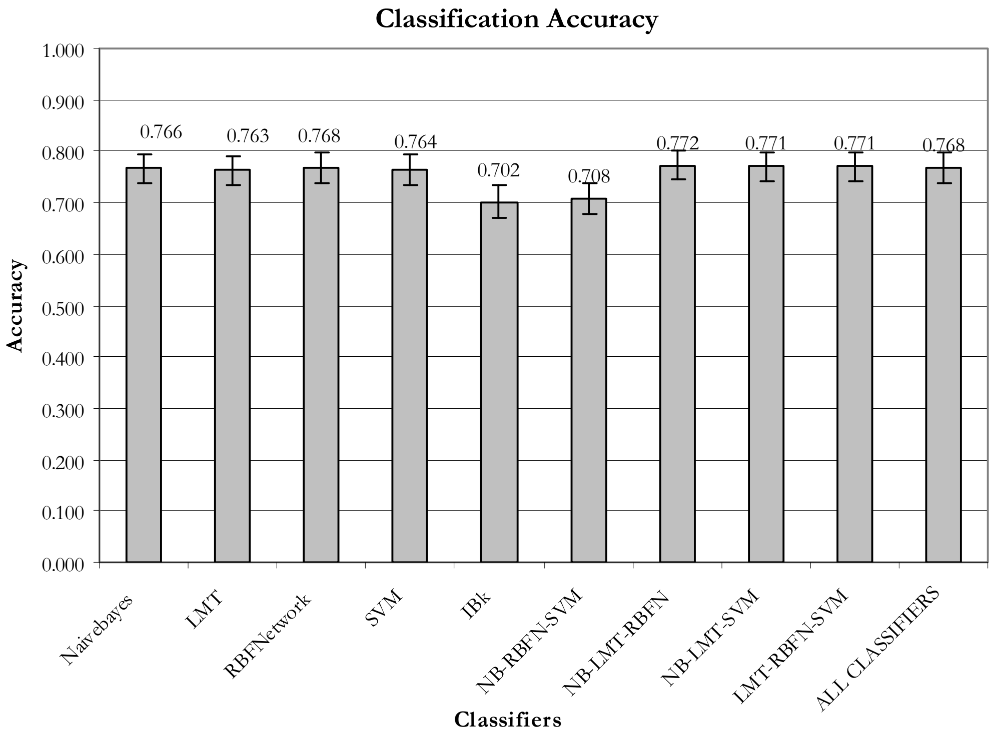
	Naivebayes	LMT	RBFNetwork	SVM	IBk
Low CI	0.739	0.738	0.743	0.738	0.677
High CI	0.789	0.788	0.793	0.788	0.731
Accuracy	0.765	0.764	0.769	0.764	0.705

Σχήμα 6. 5: Εκτίμηση των διαστημάτων εμπιστοσύνης σύμφωνα με τις μεθόδους Wald και Wilson Score για τα δεδομένα Pima Indian Diabetes

6.4.1.2 Ανάλυση αποτελεσμάτων της διαδικασία Fusion για κάθε τεχνική ταξινόμησης (Δεδομένα: 600x3)

Στο σχήμα 6.6 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων των μεμονωμένα εξεταζόμενων τεχνικών στο δείγμα ελέγχου (independent test set) εν συγκρίσει με τους πέντε (5) επιλεγμένους συνδυασμούς ταξινομητών. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζουν οι τέσσερις (4) από τους πέντε (5) συνδυασμούς ταξινομητών έναντι της μεμονωμένα εφαρμογής των υποδειγμάτων ταξινόμησης, στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Συγκεκριμένα, υψηλή ακρίβεια ταξινόμησης σημειώνεται όταν γίνουν οι συνδυασμοί NaiveBayes – LMT – RBFNetwork (77.2%), NaiveBayes – LMT – SVM ή LMT – RBFNetwork – SVM (77.1%). Ακολουθεί ο συνδυασμός όλων των ταξινομητών (76.8%) με μηδαμινή διαφοροποίηση εν συγκρίσει με την εφαρμογή μεμονωμένων τεχνικών ταξινόμησης.

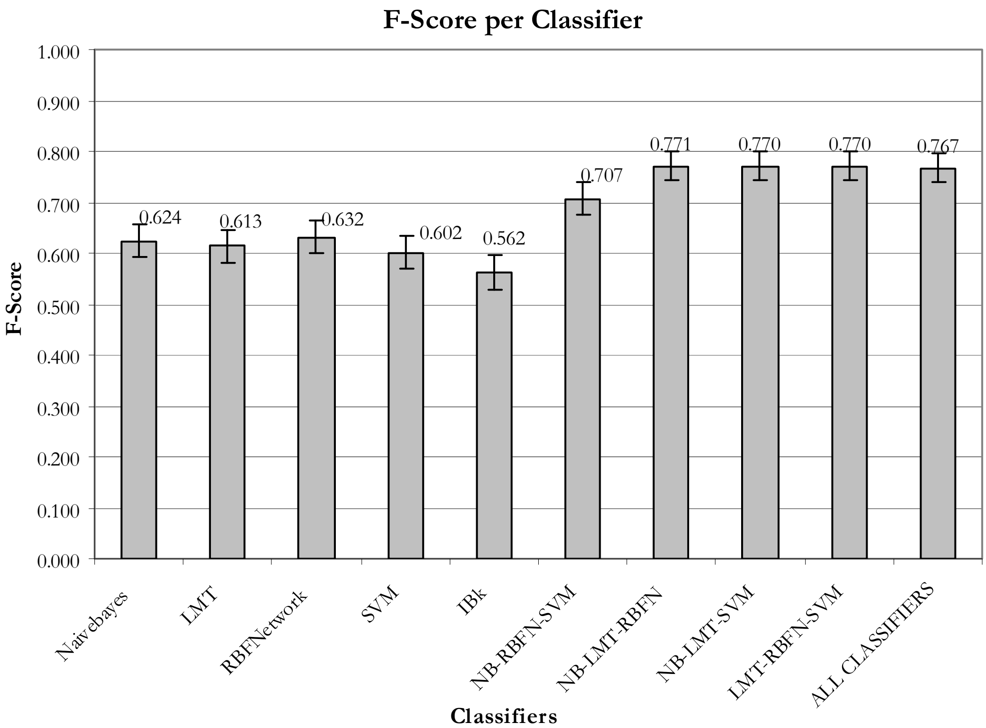
Το γεγονός ότι συνδυάστηκαν όλοι οι ταξινομητές, ενώ η αποτελεσματικότητα ταξινόμησης δεν βελτιώθηκε αισθητά, οφείλεται στη συμμετοχή της μεθόδου IBk στη διαδικασία Fusion, η οποία εμποδίζει τη βελτίωση και αντιμετώπιση του προβλήματος ταξινόμησης. Αυτό άλλωστε φαίνεται και από το σχήμα 6.6, σύμφωνα με το οποίο την χειρότερη ακρίβεια ταξινόμησης σημειώνει η τεχνική IBk με ποσοστό ακρίβειας 70.2%. Αυτό άλλωστε επιβεβαιώνεται και από το σχήμα 6.5 και την εφαρμογή της μεθόδου IBk στα δεδομένα εκπαίδευσης. Σε παρόμοια επίπεδα ακρίβειας ταξινόμησης κυμαίνεται και ο συνδυασμός των υποδειγμάτων ταξινόμησης NaïveBayes – RBFNetwork – SVM με ποσοστό σωστών ταξινομήσεων 70.8%. Από όλους τους συνδυασμούς των ταξινομητών συμπεραίνεται ότι μόνο αυτός σημειώνει το μικρό ποσοστό αποτελεσματικότητας ταξινόμησης. Φαίνεται ότι η τεχνική ταξινόμησης LMT είναι καθοριστικός παράγοντας στη βελτίωση της ακρίβειας, αφού όλοι οι άλλοι συνδυασμοί εμπεριέχουν την τεχνική LMT. Συνεπώς, η εν λόγω τεχνική συνδυάζεται πολύ καλά με οποιαδήποτε από τις άλλες τεχνικές εκτός από την IBk. Ικανοποιητικά ποσοστά και μικρές διαφοροποιήσεις στην ακρίβεια ταξινόμησης παρουσιάζουν οι τεχνικές ταξινόμησης NaïveBayes (76.6%), LMT (76.3%), RBFNetwork (76.8%) και SVM (76.4%)



Σχήμα 6. 6: Ποσοστό ακρίβειας ταξινόμησης για τα δεδομένα Pima Indian Diabetes (Fusion)

Στο σχήμα 6.7 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων των μεμονωμένα εξεταζόμενων τεχνικών στο δείγμα ελέγχου (independent test set) εν συγκρίσει με τους πέντε (5) επιλεγμένους συνδυασμούς ταξινομητών. Η παραπάνω εκτίμηση πραγματοποιείται με τη βοήθεια του μέτρου F. Τα αποτελέσματα δείχνουν την υψηλή ικανότητα ταξινόμησης που παρουσιάζουν οι συνδυασμοί των ταξινομητών έναντι της εφαρμογής μεμονωμένα της κάθε τεχνικής ταξινόμησης. Όπως και προηγουμένως, στους συνδυασμούς των ταξινομητών που ενσωματώνεται και η τεχνική LMT σημειώνεται υψηλή αποτελεσματικότητα ταξινόμησης. Συγκεκριμένα, σημειώθηκαν τα εξής ποσοστά, NaiveBayes – LMT – RBFNetwork (77.1%), NaiveBayes – LMT – SVM ή LMT – RBFNetwork – SVM (77%). Παρατηρείται, επίσης, ότι οι τεχνικές NaiveBayes και RBFNetwork επιδρούν με το ίδιο ακριβώς τρόπο όταν συνδυάζονται με τις τεχνικές LMT και SVM.

Όπως και στο σχήμα 6.6, και εδώ παρατηρείται ότι η τεχνική IBk αδυνατεί να αποδώσει υψηλές ακρίβειες ταξινόμησης, για αυτό και επιδρά αρνητικά όταν συνδυάζεται με τις υπόλοιπες τεχνικές ταξινόμησης. Συγκεκριμένα, ο συνδυασμός όλων των ταξινομητών σημείωσε F Score 76.7%. Όσον αφορά, την ικανότητα ταξινόμησης μεμονωμένα του κάθε υποδείγματος ταξινόμησης, καλύτερα αποτελέσματα σημειώνει η προσέγγιση RBFNetwork (63.2%) έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Τη μικρότερη τιμή του μέτρου παρουσιάζει η προσέγγιση IBk (56.2%). Οι υπόλοιπες τρεις προσεγγίσεις με αυστηρή σειρά, Naïve Bayes (62.4%), LMT (61.3%) και SVM (60.2%), παρουσιάζουν μία μέση απόδοση στην ικανότητα ταξινόμησης.



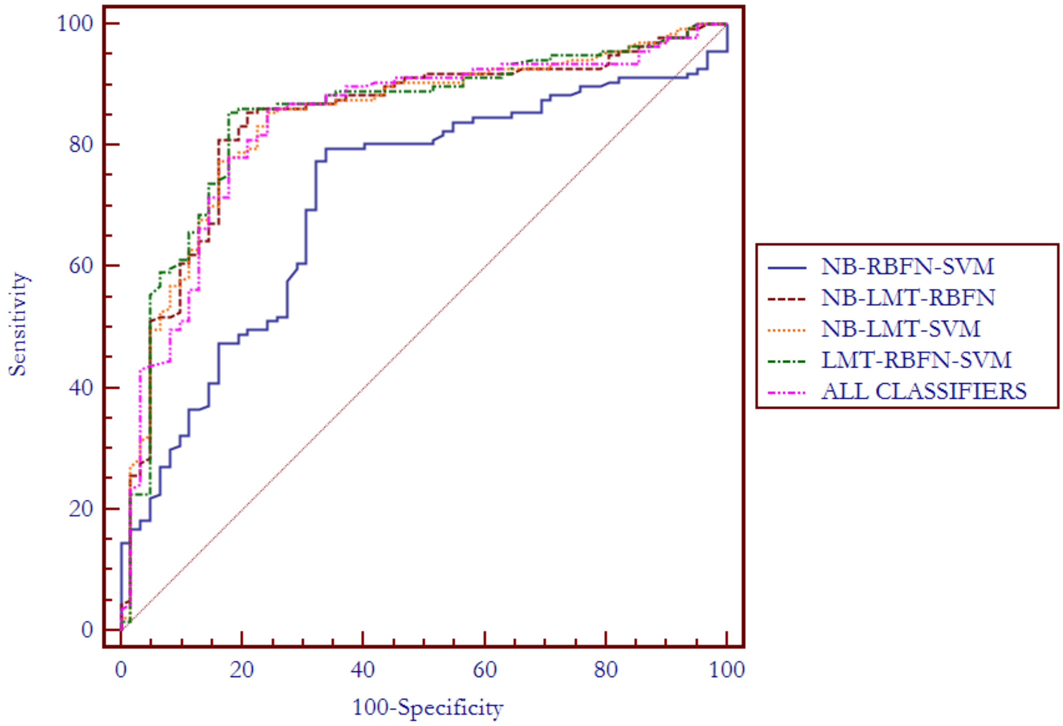
Σχήμα 6. 7: Εκτίμηση F-Score για τα δεδομένα Pima Indian Diabetes (Fusion)

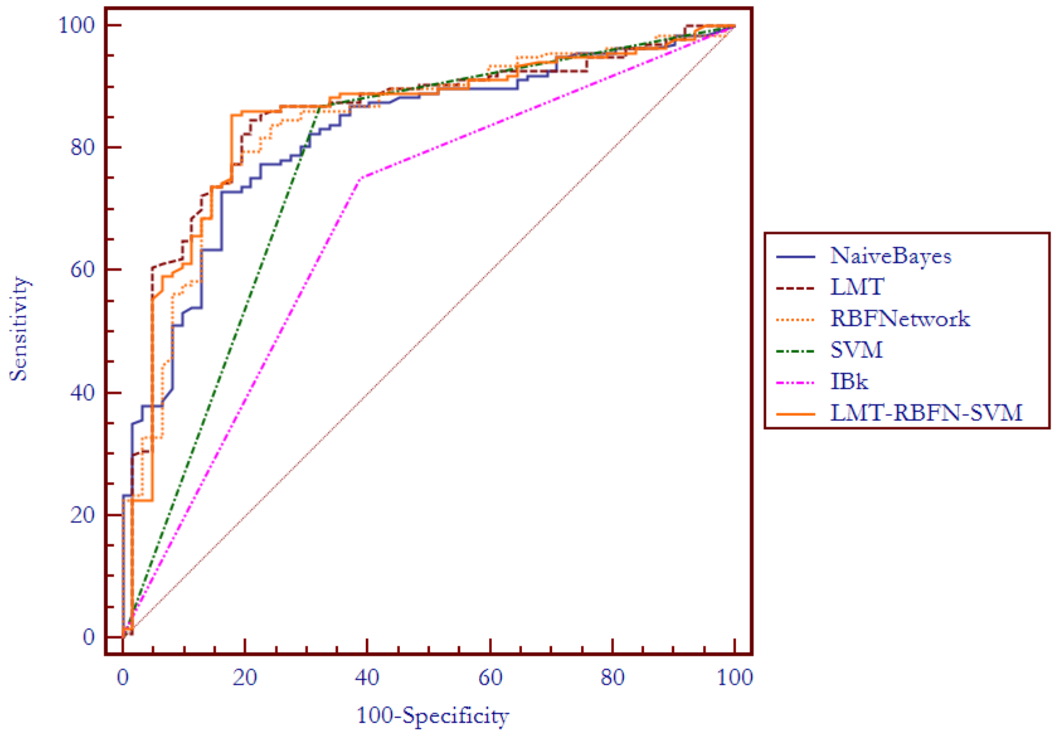
Στα σχήματα 6.8 εξετάζεται η αποτελεσματικότητα των υποδειγμάτων ταξινόμησης με τη βοήθεια του εμβαδού της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC). Προφανώς, το εμβαδόν αυτό για ένα άριστο υπόδειγμα είναι ίσο με τη μονάδα, ενώ για ένα ανεπαρκές υπόδειγμα είναι μικρότερο από 0.5. Στο πρώτο παρουσιάζονται σχηματικά οι καμπύλες ROC για διάφορους συνδυασμούς των τεχνικών ταξινόμησης, όπως αυτοί επιλέχθηκαν με την χρήση του κριτηρίου Q Statistic και συνδυάστηκαν με τη βοήθεια της τεχνικής συνδυασμού ταξινομητών Average Rule (Κεφάλαιο 5). Στο δεύτερο σχήμα παρουσιάζονται σχηματικά οι καμπύλες ROC των διαφόρων τεχνικών ταξινόμησης μεμονωμένα εν συγκρίσει με τον συνδυασμό ταξινομητών, με βάσει το πρώτο σχήμα, που σημείωσε την υψηλότερη τιμή για το εμβαδόν κάτω από την καμπύλη ROC.

Από τα σχήματα 6.8 και τον πίνακα 6.4 παρατηρείται ευκρινώς ότι ο συνδυασμός NaïveBayes – RBFNetwork – SVM δεν αποδίδει τόσα καλά αποτελέσματα όσο οι υπόλοιποι συνδυασμοί. Την καλύτερη απόδοση με υψηλότερη τιμή εμβαδού σημειώνει ο συνδυασμός των ταξινομητών LMT – RBFNetwork – SVM (AUC=0.848). Ακολουθούν οι συνδυασμοί NaiveBayes – LMT – SVM ή NaiveBayes – LMT – RBFNetwork (AUC=0.842), όλοι οι ταξινομητές μαζί με AUC=0.837 και τέλος, NaiveBayes – RBFNetwork – SVM με AUC=0.713.

Την καλύτερη απόδοση ταξινόμησης του συνδυασμού των ταξινομητών NaïveBayes – RBFNetwork – SVM έρχεται να «κοντράρει», η τεχνική LMT η οποία σημειώνει ελάχιστα υψηλότερη τιμή εμβαδού (AUC=0.851) και η τεχνική RBFNetwork με ελάχιστα χαμηλότερη τιμή εμβαδού (AUC=0.841).

Όσον αφορά τις υπόλοιπες εξεταζόμενες τεχνικές ταξινόμησης η μέθοδος IBk δίνει την χαμηλότερη τιμή εμβαδού (AUC=0.684), γεγονός που αποδεικνύει την αδυναμία της μεθόδου να αποδώσει υψηλά ποσοστά ακρίβειας. Ακολουθεί η τεχνική SVM (AUC=0.773) με μέτρια απόδοση ακρίβειας. Σαφώς υψηλότερη ακρίβεια ταξινόμησης σημειώνει η τεχνική Naïve Bayes (AUC=0.823).





Σχήμα 6. 8: Σχηματική απεικόνιση καμπύλης ROC για τα δεδομένα Pima Indian Diabetes (Fusion)

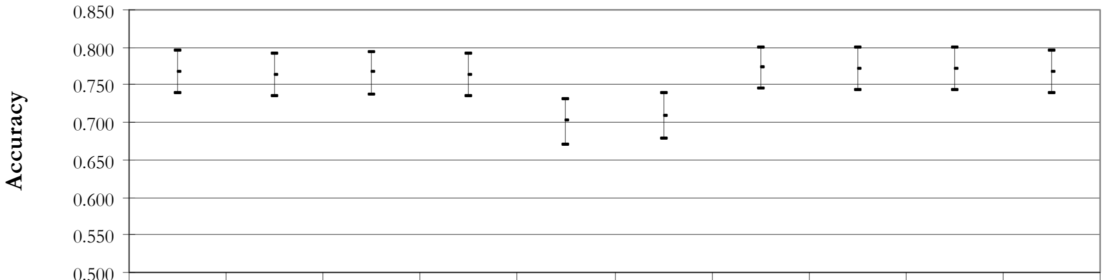
Πίνακας 6. 4: Αποτελέσματα καμπύλης ROC για τα δεδομένα Pima Indian Diabetes (Fusion)

CLASSIFIERS ARE COMPARED WITH THE BEST CLASSIFIER FUSION	AUC	95% CI
NaiveBayes	0.823	0.763 to 0.874
LMT	0.851	0.794 to 0.897
RBFNetwork	0.841	0.782 to 0.888
SVM	0.773	0.708 to 0.829
IBk	0.684	0.614 to 0.748
NB_RBFN_SVM	0.713	0.645 to 0.775
NB_LMT_RBFN	0.842	0.784 to 0.890
NB_LMT_SVM	0.842	0.783 to 0.889
LMT_RBFN_SVM	0.848	0.790 to 0.894
ALL_CLASSIFIERS	0.837	0.779 to 0.886

Στο σύνολο των σχημάτων 6.9 εξετάζονται τα διαστήματα εμπιστοσύνης της ακρίβειας ταξινόμησης των εξεταζόμενων υποδειγμάτων ταξινόμησης. Για τον υπολογισμό των διαστημάτων εμπιστοσύνης, όπως και προαναφέρθηκε, χρησιμοποιήθηκαν δύο μέθοδοι εκτίμησης αυτών (Wald Method και Wilson Score Method) και υπολογίστηκαν τα διαστήματα εμπιστοσύνης για συντελεστή εμπιστοσύνης 95%. Από τα παρακάτω σχήματα προκύπτουν ότι οι δύο εξεταζόμενες μέθοδοι εκτίμησης των διαστημάτων εμπιστοσύνης, Wald Method και Wilson Score Method, παρουσιάζουν παρόμοιες εκτιμήσεις, με την Wald Method να σημειώνει ελάχιστα μικρότερου εύρους διαστήματα εμπιστοσύνης.

Όλες οι τεχνικές ταξινόμησης, καθώς και οι συνδυασμοί αυτών παρουσιάζουν του ιδίου εύρους διαστήματα εμπιστοσύνης. Από όλες τις τεχνικές μόνο η μέθοδος IBk σημειώνει τις μικρότερες επιδόσεις ταξινόμησης, καθώς η υψηλότερη τιμή του διαστήματος εμπιστοσύνης της είναι η χαμηλότερη τιμή των διαστημάτων εμπιστοσύνης των υπολοίπων τεχνικών. Το ίδιο ισχύει και για τον συνδυασμό NaïveBayes – RBFNetwork – SVM ο οποίος δεν αποδίδει ικανοποιητικά αποτελέσματα ταξινόμησης. Προφανώς, όπως και διαπιστώθηκε παραπάνω, μπορεί οι 2 τεχνικές ταξινόμησης NaïveBayes και RBFNetwork να συμπεριφέρονται κατά τον ίδιο τρόπο όταν συνδυάζονται με τις τεχνικές LMT – SVM, όταν όμως συνδυάζονται μεταξύ τους δεν αποδίδουν ικανοποιητικά αποτελέσματα.

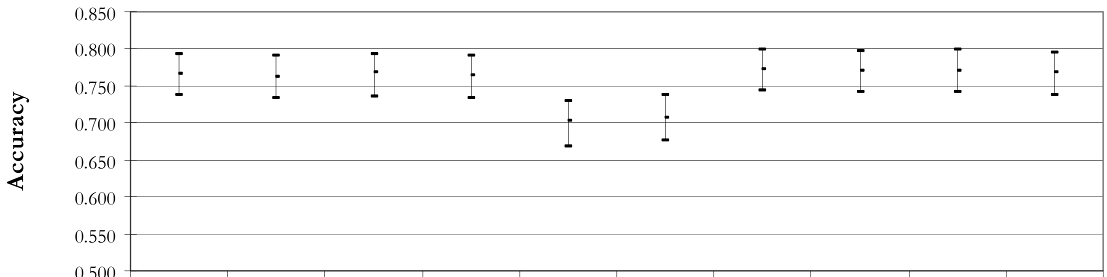
Confidence Intervals with Wald Method



	Naivebayes	LMT	RBFNetwork	SVM	IBk	NB-RBFN-SVM	NB-LMT-RBFN	NB-LMT-SVM	LMT-RBFN-SVM	ALL CLASSIFIERS
Low CI	0.738	0.734	0.737	0.735	0.670	0.677	0.744	0.742	0.743	0.739
High CI	0.795	0.791	0.794	0.792	0.731	0.738	0.800	0.799	0.799	0.796
Accuracy	0.766	0.763	0.768	0.764	0.702	0.708	0.772	0.771	0.771	0.768

Classifiers

Confidence Intervals with Wilson Score Method



	Naivebayes	LMT	RBFNetwork	SVM	IBk	NB-RBFN-SVM	NB-LMT-RBFN	NB-LMT-SVM	LMT-RBFN-SVM	ALL CLASSIFIERS
Low CI	0.737	0.733	0.736	0.734	0.669	0.676	0.743	0.741	0.742	0.738
High CI	0.793	0.790	0.792	0.791	0.730	0.737	0.799	0.797	0.798	0.795
Accuracy	0.766	0.763	0.768	0.764	0.702	0.708	0.772	0.771	0.771	0.768

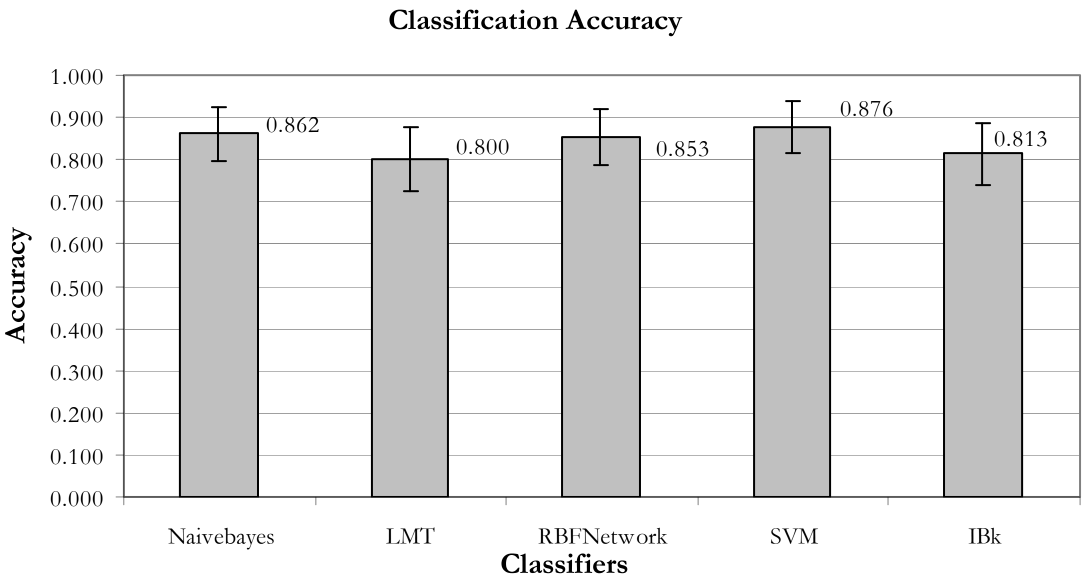
Classifiers

Σχήμα 6. 9: Εκτίμηση των διαστημάτων εμπιστοσύνης σύμφωνα με τις μεθόδους Wald και Wilson Score για επίπεδο σημαντικότητας 5% για τα δεδομένα Pima Indian Diabetes (Fusion)

6.3.2 Δεδομένα Breast Cancer

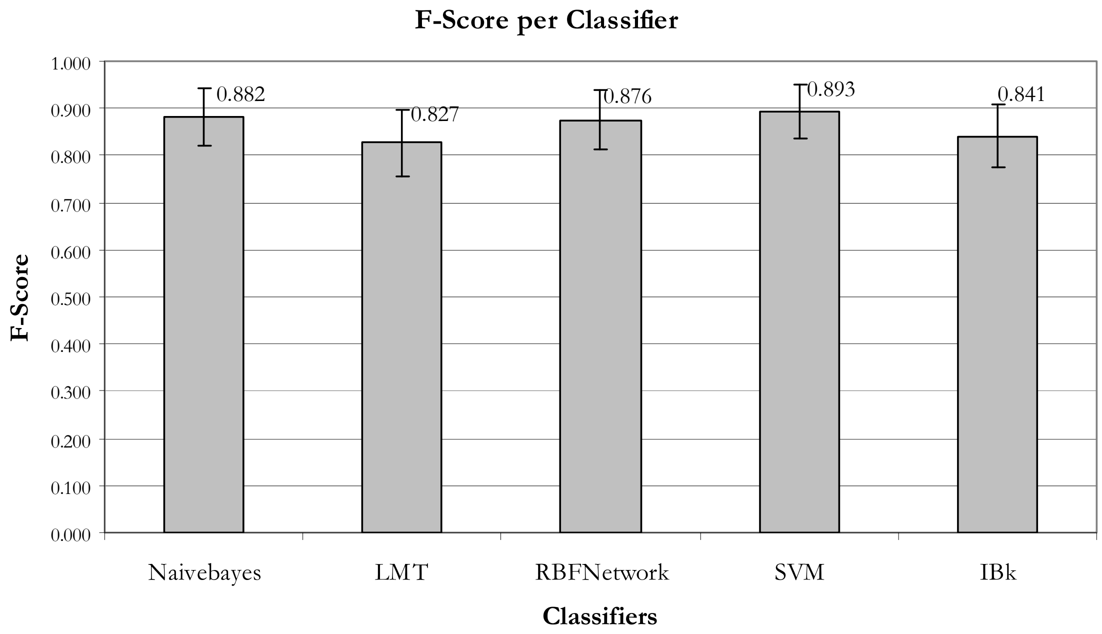
6.3.2.1 Ανάλυση αποτελεσμάτων μεμονωμένα για κάθε τεχνική ταξινόμησης (Δεδομένα: 78x605)

Στο σχήμα 6.10 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων όλων των εξεταζόμενων τεχνικών στο δείγμα εκπαίδευσης. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζει η προσέγγιση SVM (87.6%) έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Μεγαλύτερο σφάλμα ταξινόμησης από όλες τις εξεταζόμενες τεχνικές παρουσιάζει η προσέγγιση LMT (80%). Οι υπόλοιπες τρεις προσεγγίσεις, με αυστηρή σειρά ακρίβειας σημειώνουν τα εξής ποσοστά, Naïve Bayes (86.2%), RBFNetwork (85.3%) και IBk (81.3%).



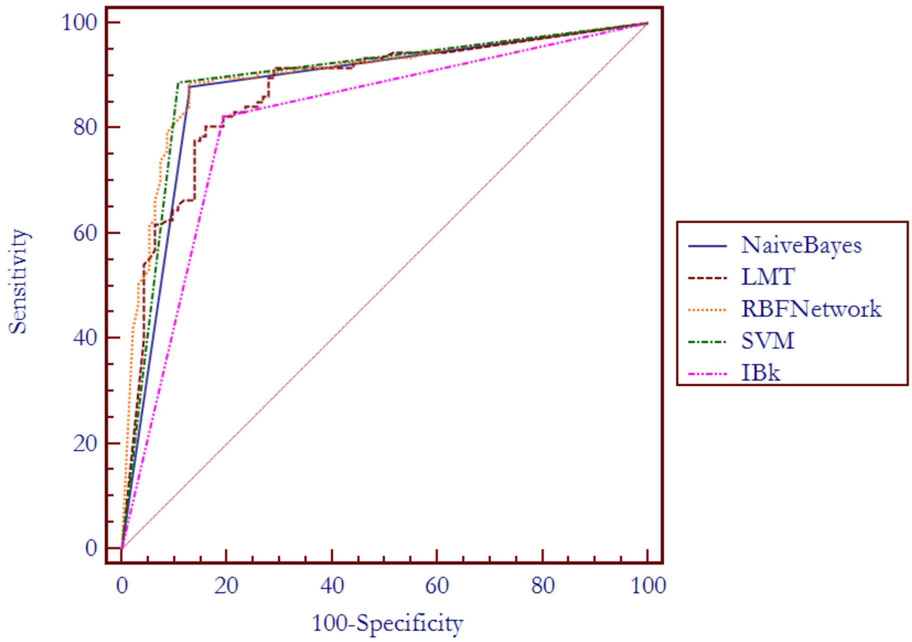
Σχήμα 6. 10: Ποσοστό ακρίβειας ταξινόμησης για τα δεδομένα Breast Cancer

Στο σχήμα 6.11 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με την ικανότητα ταξινόμησης των εξεταζόμενων τεχνικών στο δείγμα εκπαίδευσης, όπως εκτιμάται με τη βοήθεια του μέτρου F. Τα αποτελέσματα δείχνουν την υψηλή ικανότητα ταξινόμησης που παρουσιάζει η προσέγγιση SVM (89.3%) έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Σημαντικά τη μικρότερη τιμή του μέτρου παρουσιάζει η προσέγγιση LMT (82.7%). Οι υπόλοιπες τρεις προσεγγίσεις με αυστηρή σειρά, Naïve Bayes (88.2%), RBFNetwork (87.6%) και IBk (84.1%), παρουσιάζουν μία μέση απόδοση στην ικανότητα ταξινόμησης.



Σχήμα 6. 11: Εκτίμηση F-Score για τα δεδομένα Breast Cancer

Στο σχήμα 6.12 εξετάζεται η αποτελεσματικότητα των υποδειγμάτων ταξινόμησης με τη βοήθεια του εμβαδού της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC). Προφανώς, το εμβαδόν αυτό για ένα άριστο υπόδειγμα είναι ίσο με τη μονάδα, ενώ για ένα ανεπαρκές υπόδειγμα είναι μικρότερο από 0.5. Όσον αφορά τις εξεταζόμενες τεχνικές ταξινόμησης, οι εκτιμήσεις του εμβαδού της καμπύλης ROC, όπως φαίνεται στον πίνακα 6.5, έδωσαν ικανοποιητικά αποτελέσματα για όλες τις τεχνικές ταξινόμησης. Γενικότερα, με αυστηρή σειρά, οι τιμές που σημειώθηκαν για κάθε ταξινομητή είναι RBFNetwork (AUC=0.899) και SVM (AUC=0.890) με μικρές διαφοροποιήσεις, Naïve Bayes (AUC=0.874) και LMT (AUC=0.869) με επίσης, μικρές διαφοροποιήσεις και τέλος, IBk (AUC=0.814) με μικρή τιμή αλλά εξίσου καλή απόδοση.



Σχήμα 6. 12: Εμβαδόν κάτω από την καμπύλη ROC για τα δεδομένα Breast Cancer

Πίνακας 6. 5: Αποτελέσματα καμπύλης ROC για τα δεδομένα Breast Cancer

Classifiers	AUC	95% CI
NaiveBayes	0.874	0.820 to 0.917
LMT	0.869	0.814 to 0.912
RBFNetwork	0.899	0.848 to 0.937
SVM	0.890	0.838 to 0.930
IBk	0.814	0.754 to 0.866

Στο σύνολο των σχημάτων 6.13 εξετάζονται τα διαστήματα εμπιστοσύνης της ακρίβειας ταξινόμησης των εξεταζόμενων υποδειγμάτων ταξινόμησης. Για τον υπολογισμό των διαστημάτων εμπιστοσύνης, όπως και προαναφέρθηκε, χρησιμοποιήθηκαν δύο μέθοδοι εκτίμησης αυτών (Wald Method και Wilson Score Method) και υπολογίστηκαν τα διαστήματα εμπιστοσύνης για διάφορους συντελεστές εμπιστοσύνης. Συγκεκριμένα, στα σχήματα απεικονίζονται τα διαστήματα εμπιστοσύνης της ακρίβειας των ταξινομητών για συντελεστή εμπιστοσύνης 99.5% (α=0.005, 1η δεκάδα κάθε διαγράμματος), 97.5% (α=0.025, 2η δεκάδα κάθε διαγράμματος), 95% (α=0.05, 3η δεκάδα κάθε διαγράμματος) και 90% (α=0.1, 4η και τελευταία δεκάδα κάθε διαγράμματος). Από τα παρακάτω σχήματα προκύπτουν οι εξής διαπιστώσεις:

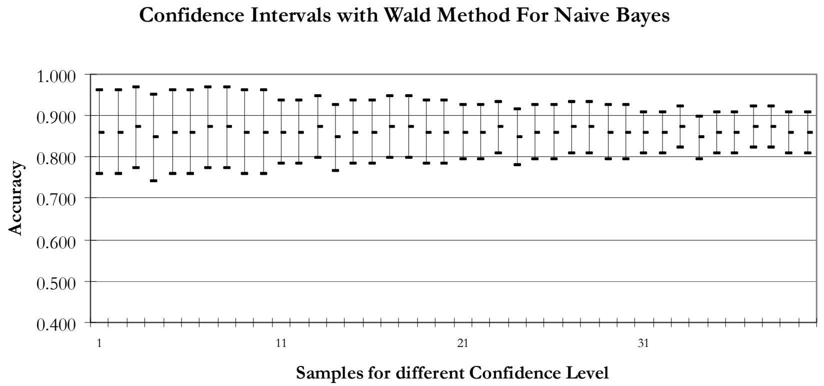
1. Οι δύο εξεταζόμενες μέθοδοι εκτίμησης των διαστημάτων εμπιστοσύνης, Wald Method και Wilson Score Method, παρουσιάζουν μικρές διαφοροποιήσεις στην εκτίμηση των διαστημάτων εμπιστοσύνης, που αποδεικνύουν όμως την προτίμηση της μεθόδου Wald. Η μέθοδος Wald δίνει πιο υψηλό εύρος τιμών ακρίβειας από ότι η μέθοδο Wilson Score, όπως και παρατηρείται από τα παρακάτω διαγράμματα.

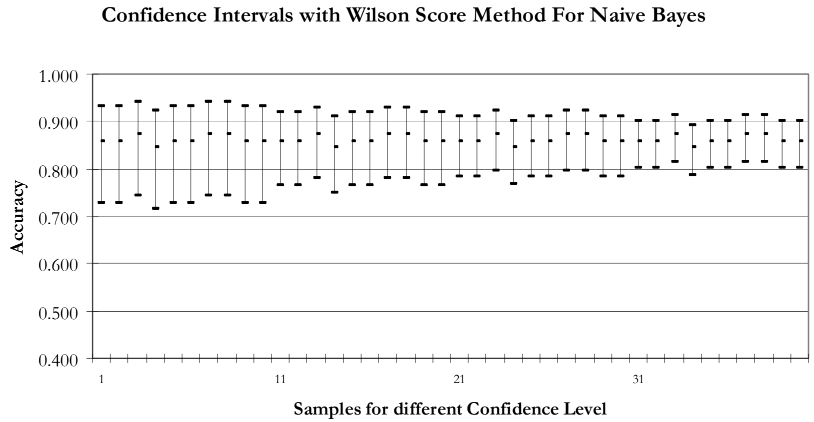
2. Όσο μικραίνει ο συντελεστής εμπιστοσύνης τόσο μικραίνει και το εύρος των διαστημάτων εμπιστοσύνης. Συγκεκριμένα, για συντελεστή εμπιστοσύνης 99.5% εκτιμάται το εύρος του διαστήματος περίπου 0.22, για 97.5% περίπου 0.18, για 95% περίπου 0.14, ενώ για 90% περίπου 0.1. Συνεπώς, ο συντελεστής εμπιστοσύνης επηρεάζει τα άκρα του διαστήματος εμπιστοσύνης και όσο αυξάνεται ο συντελεστής, αυξάνονται και οι πιθανότητες να βρίσκεται μέσα στο διάστημα εμπιστοσύνης η πραγματική τιμή της εκτιμώμενης παραμέτρου.

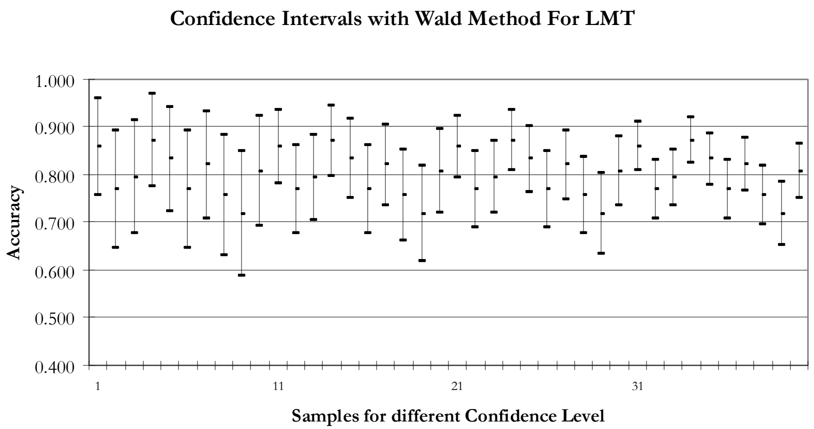
3. Όσον αφορά τις χρησιμοποιούμενες τεχνικές ταξινόμησης και για συντελεστή εμπιστοσύνης 95% με τη χρήση της μεθόδου Wald, η πραγματική διαφορά στην εκτίμηση της μέγιστης από την ελάχιστη τιμή ακρίβειας, που μπορεί να αποδώσει το κάθε υπόδειγμα ταξινόμησης για τα πραγματικά δεδομένα Breast Cancer, βρίσκεται στο διάστημα, κατά προσέγγιση μέση τιμής, [0.797, 0.926] για την τεχνική Naïve Bayes, [0.726, 0.874] για την LMT, [0.787, 0.918] για την τεχνική RBFNetwork, [0.814, 0.937] για την SVM και [0.740, 0.885] για την τεχνική IBk.

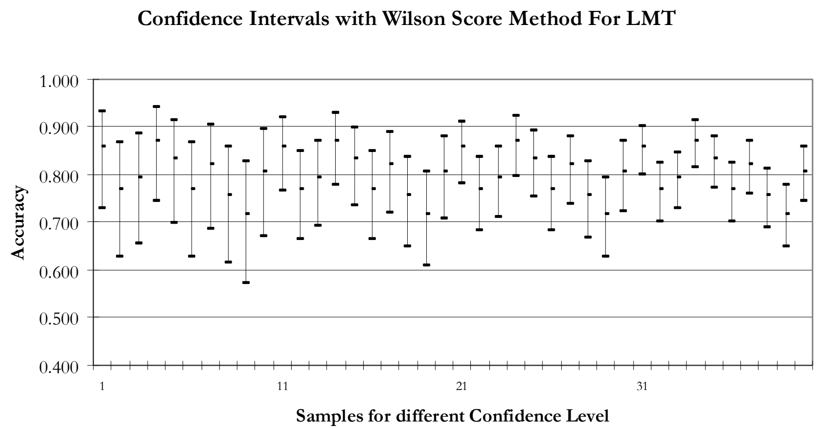
4. Η τεχνική ταξινόμησης LMT παρουσιάζει αισθητά διαφορετικές εκτιμήσεις διαστημάτων εμπιστοσύνης ως προς την ακρίβεια ταξινόμησης για διαφορετικά δείγματα, ίδιου μεγέθους και από το ίδιο σύνολο δεδομένων. Αυτό αποδεικνύει την αδυναμία την μεθόδου στην επίλυση προβλημάτων ταξινόμησης.

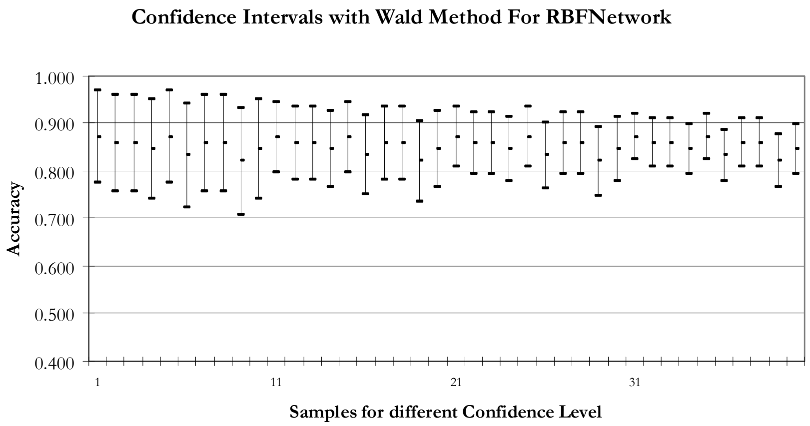
Έτσι λοιπόν, και από την εκτίμηση των διαστημάτων εμπιστοσύνης συμπεραίνεται ότι η τεχνική LMT δεν είναι ιδιαίτερα αξιόπιστη ως προς την αντιμετώπιση προβλημάτων ταξινόμησης. Σαφώς καλύτερα αποτελέσματα αποδίδουν οι υπόλοιπες τεχνικές, με την SVM να υπερτερεί έναντι των τεχνικών Naïve Bayes και RBFNetwork με μικρή διαφορά.

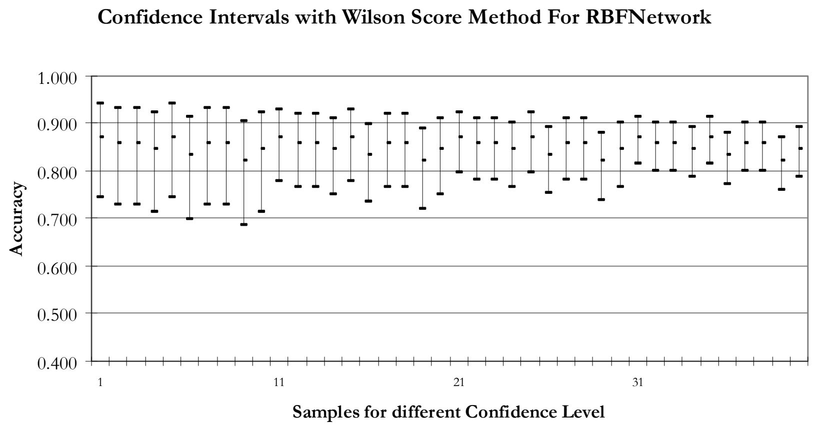

Confidence Intervals with Wald Method For Naive Bayes
Accuracy
1.000
0.900
0.800
0.700
0.600
0.500
0.400
1
11
21
31
Samples for different Confidence Level

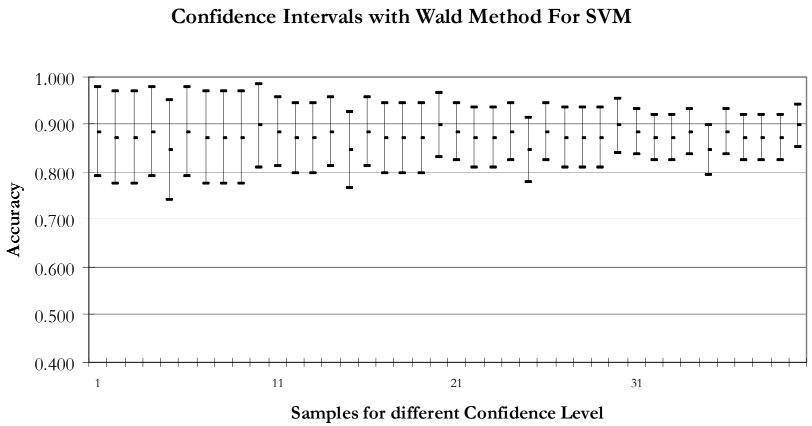

Confidence Intervals with Wilson Score Method For Naive Bayes
Accuracy
1.000
0.900
0.800
0.700
0.600
0.500
0.400
1
11
21
31
Samples for different Confidence Level

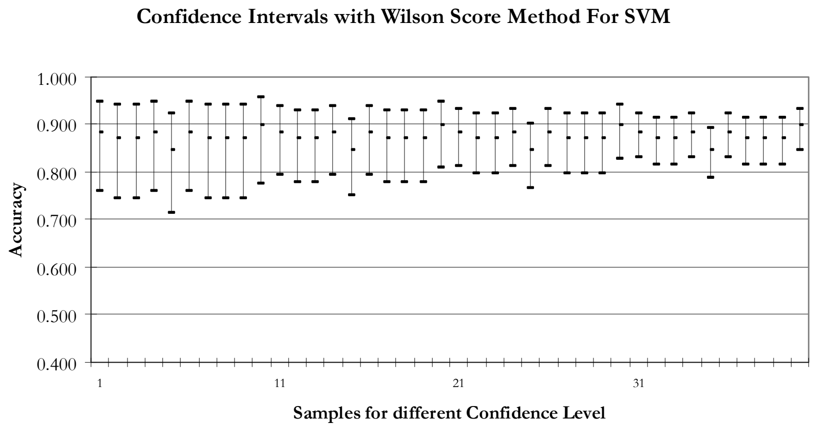

Confidence Intervals with Wald Method For LMT
Accuracy
1.000
0.900
0.800
0.700
0.600
0.500
0.400
1
11
21
31
Samples for different Confidence Level

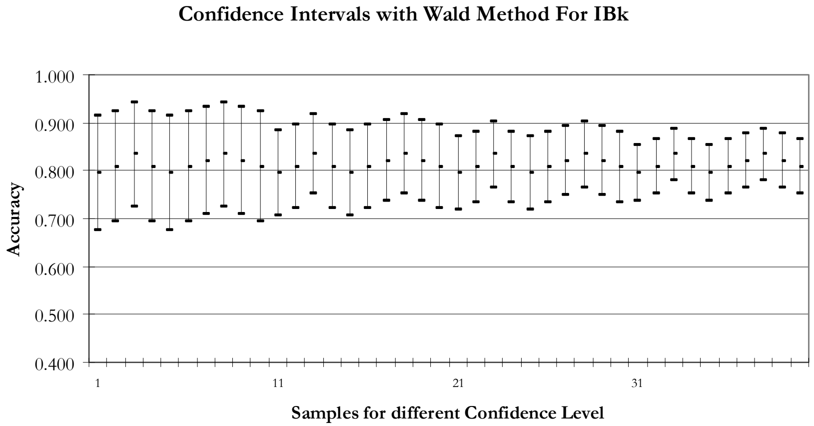

Confidence Intervals with Wilson Score Method For LMT
Accuracy
1.000
0.900
0.800
0.700
0.600
0.500
0.400
1
11
21
31
Samples for different Confidence Level

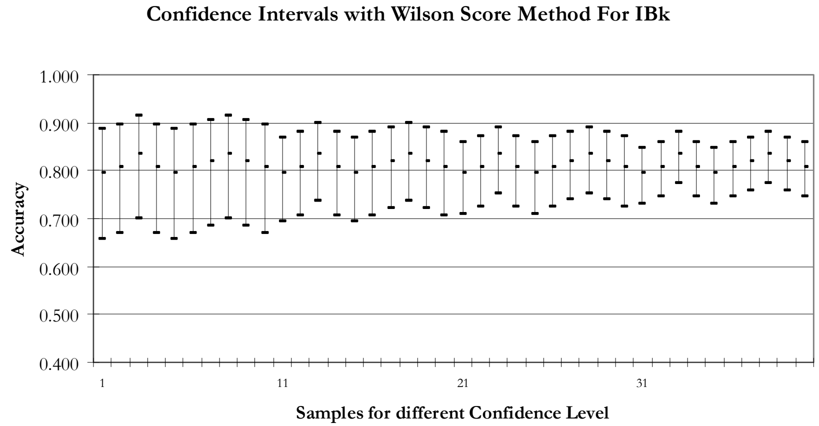

Confidence Intervals with Wald Method For RBFNetwork
1.000
0.900
0.800
0.700
0.600
0.500
0.400
Accuracy
1
11
21
31
Samples for different Confidence Level

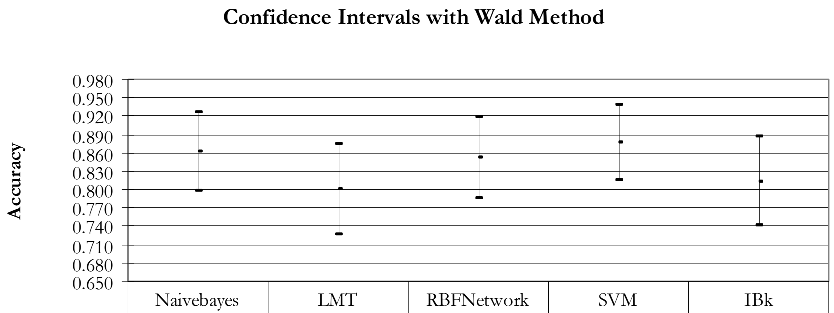

Confidence Intervals with Wilson Score Method For RBFNetwork
1.000
0.900
0.800
0.700
0.600
0.500
0.400
Accuracy
1
11
21
31
Samples for different Confidence Level


Confidence Intervals with Wald Method For SVM
1.000
0.900
0.800
0.700
0.600
0.500
0.400
Accuracy
1
11
21
31
Samples for different Confidence Level


Confidence Intervals with Wilson Score Method For SVM
1.000
0.900
0.800
0.700
0.600
0.500
0.400
Accuracy
1
11
21
31
Samples for different Confidence Level

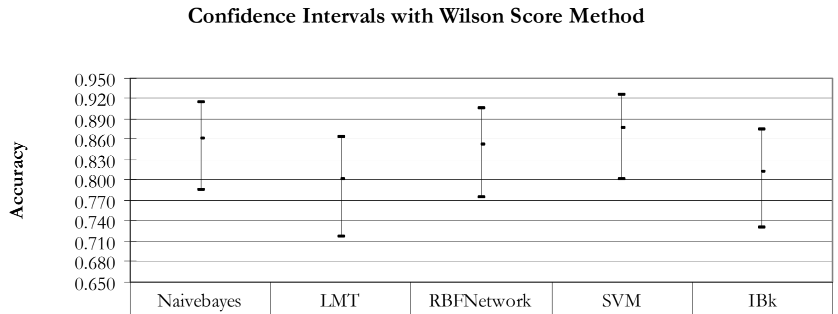






	Naivebayes	LMT	RBFNetwork	SVM	IBk
Low CI	0.797	0.726	0.787	0.814	0.740
High CI	0.926	0.874	0.918	0.937	0.885
Accuracy	0.862	0.800	0.853	0.876	0.813

Classifiers



	Naivebayes	LMT	RBFNetwork	SVM	IBk
Low CI	0.785	0.717	0.775	0.801	0.730
High CI	0.914	0.863	0.907	0.925	0.874
Accuracy	0.862	0.800	0.853	0.876	0.813

Classifiers

Σχήμα 6. 13: Εκτίμηση των διαστημάτων εμπιστοσύνης σύμφωνα με τις μεθόδους Wald και Wilson Score για τα δεδομένα Breast Cancer

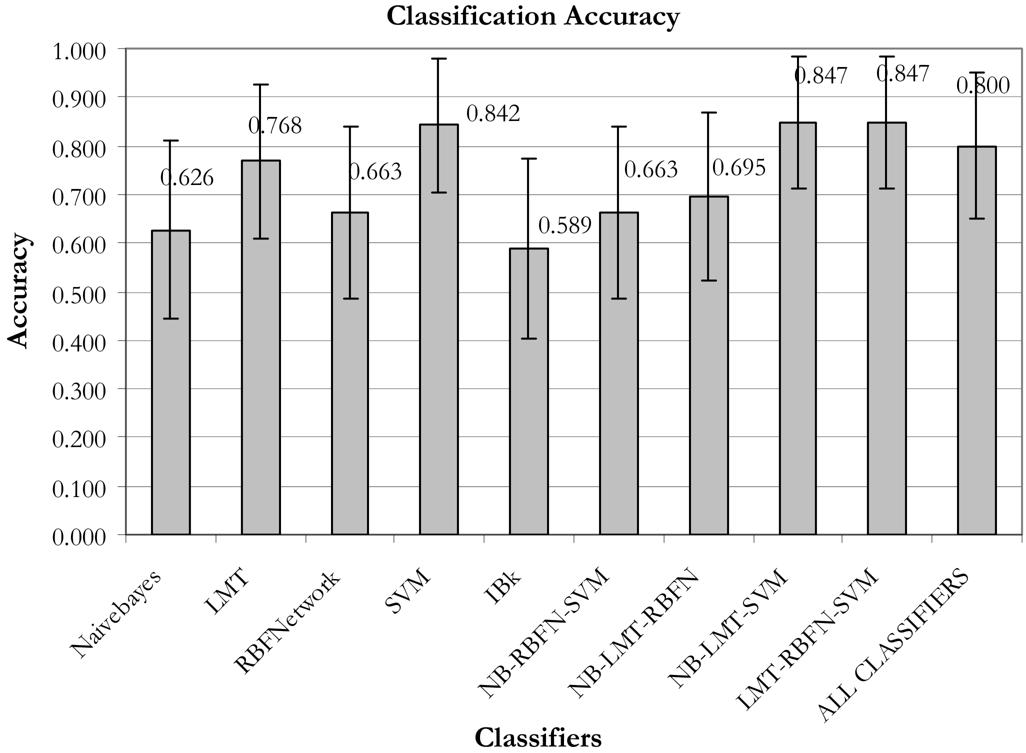
6.3.2.2 Ανάλυση αποτελεσμάτων της διαδικασία Fusion για κάθε τεχνική ταξινόμησης (Δεδομένα: 19x605)

Στο σχήμα 6.14 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων των μεμονωμένα εξεταζόμενων τεχνικών στο δείγμα ελέγχου (independent test set) εν συγκρίσει με τους πέντε (5) επιλεγμένους συνδυασμούς ταξινομητών. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζουν οι δύο (2) από τους πέντε (5) συνδυασμούς ταξινομητών έναντι της μεμονωμένα εφαρμογής των υποδειγμάτων ταξινόμησης, στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Συγκεκριμένα, υψηλή ακρίβεια ταξινόμησης σημειώνεται όταν γίνουν οι συνδυασμοί NaiveBayes – LMT – SVM ή LMT – RBFNetwork – SVM (84.7%). Ακολουθεί η τεχνική ταξινόμησης SVM με ποσοστό ακρίβειας ταξινόμησης της τάξεως 84.2% και αμέσως μετά, ο συνδυασμός όλων των ταξινομητών (80%). Επίσης, καλή απόδοση σημειώνει και η τεχνική ταξινόμησης LMT με ποσοστό σωστών ταξινομήσεων 76.8%.

Οι υπόλοιπες εκτιμήσεις όσον αφορά την ακρίβεια ταξινόμησης, με φθίνουσα σειρά, είναι οι εξής: NaiveBayes – LMT – RBFNetwork (69.5%), RBFNetwork ή NaiveBayes – RBFNetwork – SVM (66.3%) και NaiveBayes (62.6%). Όπως και αναμενόταν, η τεχνική ταξινόμησης IBk σημείωσε τη χειρότερη επίδοση με ποσοστό σωστών ταξινομήσεων 58.9%.

Διαπιστώνεται ότι, εν συγκρίσει με τα δεδομένα Pima Indians Diabetes, τα αποτελέσματα και οι εκτιμήσεις μεταβλήθηκαν σημαντικά. Αυτό ίσως οφείλεται στη σημαντική διαφορά των προς ταξινόμηση παρατηρήσεων. Συγκεκριμένα, στα δεδομένα Pima Indians Diabetes ταξινομήθηκαν 600 ασθενείς. Σαφώς μια μεμονωμένη τεχνική ταξινόμησης μπορεί να παρουσιάσει δυσκολίες στην ταξινόμηση μεγάλου αριθμού παρατηρήσεων. Επομένως, η εφαρμογή συνδυασμού ταξινομητών μπορεί να καταστεί αναγκαία, προκειμένου να βελτιωθούν τα αποτελέσματα στην αποτελεσματικότητα της ταξινόμησης. Αντιθέτως, τα δεδομένα Breast Cancer είχαν αισθητά πολύ μικρό αριθμό παρατηρήσεων προς ταξινόμηση γεγονός που κάνει δυνατή την εκτίμηση υψηλών τιμών ακρίβειας κατά την εφαρμογή μεμονωμένων ταξινομητών, αλλά δεν δικαιολογεί τις χαμηλές τιμές ακρίβειας στο συνδυασμό των ταξινομητών.

Ο συνδυασμός των ταξινομητών NaiveBayes – LMT – SVM ή LMT – RBFNetwork – SVM σημείωσε υψηλά αποτελέσματα ταξινόμησης και στα δύο σύνολα δεδομένων (Pima Indians Diabetes και Breast Cancer) ανεξαρτήτου αριθμού παρατηρήσεων. Επομένως, μπορεί να θεωρηθεί ότι οι συγκεκριμένοι συνδυασμοί πετυχαίνουν πάντα υψηλές αποδόσεις και δεν επηρεάζονται από το πλήθος των παρατηρήσεων του δείγματος.

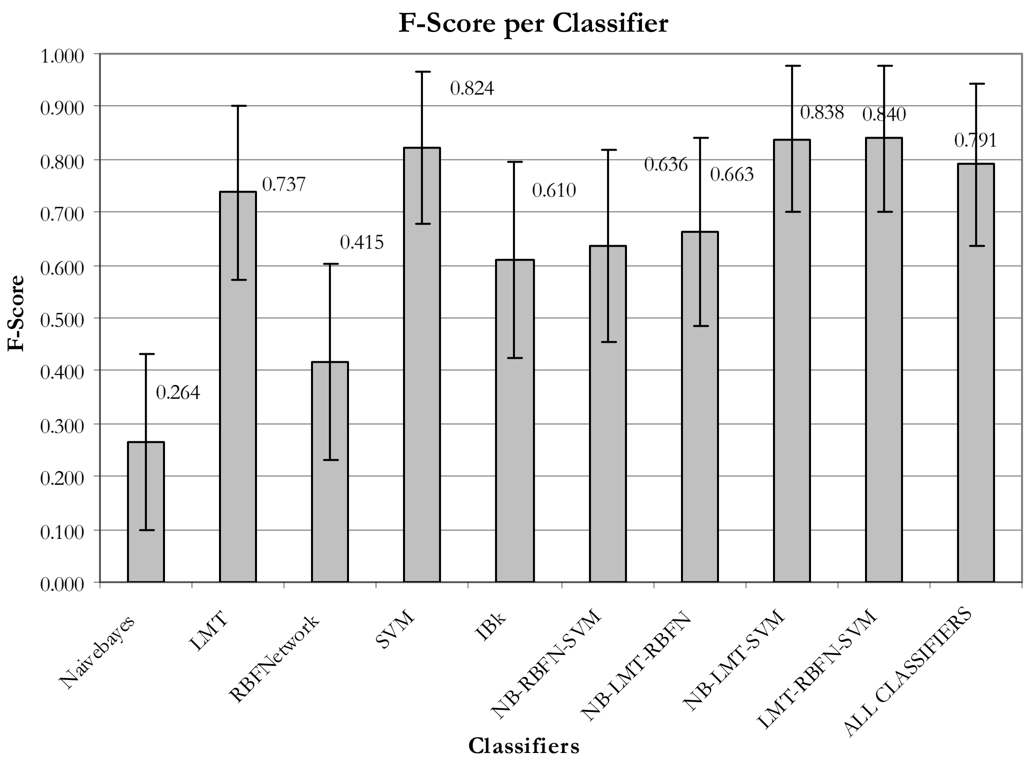


Σχήμα 6. 14: Ποσοστό ακρίβειας ταξινόμησης για τα δεδομένα Breast Cancer (Fusion)

Στο σχήμα 6.15 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων των μεμονωμένα εξεταζόμενων τεχνικών στο δείγμα ελέγχου (independent test set) εν συγκρίσει με τους πέντε (5) επιλεγμένους συνδυασμούς ταξινομητών. Η παραπάνω εκτίμηση πραγματοποιείται με τη βοήθεια του μέτρου F. Τα αποτελέσματα δείχνουν την υψηλή ικανότητα ταξινόμησης που παρουσιάζουν οι δύο (2) συνδυασμοί ταξινομητών έναντι της εφαρμογής μεμονωμένα της κάθε τεχνικής ταξινόμησης. Όπως και προηγουμένως, οι συνδυασμοί των ταξινομητών NaiveBayes – LMT – SVM ή LMT – RBFNetwork – SVM σημειώνουν υψηλή ικανότητα αντιμετώπισης του προβλήματος της ταξινόμησης. Συγκεκριμένα, για το κριτήριο F σημείωσαν 83.8% και 84%, αντίστοιχα. Ακολουθούν η τεχνική ταξινόμησης SVM με ποσοστό 82.4%, ο συνδυασμός όλων των ταξινομητών με ποσοστό 79.1 και η τεχνική ταξινόμησης LMT με ποσοστό 73.7%. Όπως διαπιστώνεται, οι εν λόγω τεχνικές, SVM και LMT, είτε εφαρμοστούν μεμονωμένα είτε συνδυαστούν αποδίδουν υψηλές αποδόσεις.

Όλες οι υπόλοιπες τεχνικές ταξινόμησης παρουσιάζουν χαμηλή ικανότητα ταξινόμησης, με χειρότερη στα συγκεκριμένα δεδομένα τη μέθοδο NaiveBayes με ποσοστό 26.4%. Οι τεχνικές ταξινόμησης RBFNetwork και IBk, σημείωσαν ποσοστά της τάξεως 41.5% και 61%, αντίστοιχα. Η τεχνική ταξινόμησης IBk μπορεί να παρουσιάζει πάντοτε τις μικρότερες επιδόσεις στην αντιμετώπιση προβλημάτων ταξινόμησης, αλλά κυμαίνεται πάντα στα ίδια επίπεδα, ακόμα και για διαφορετικού μεγέθους ή χαρακτηριστικών ή ιδιοτήτων δεδομένων ταξινόμησης.

Τέλος, σε χαμηλά επίπεδα κυμαίνονται και οι συνδυασμοί ταξινομητών NaiveBayes – RBFNetwork – SVM και NaiveBayes – LMT – RBFNetwork, οι οποίοι σημείωσαν για το κριτήριο F ποσοστά της τάξεως 61% και 63.6%, αντίστοιχα.



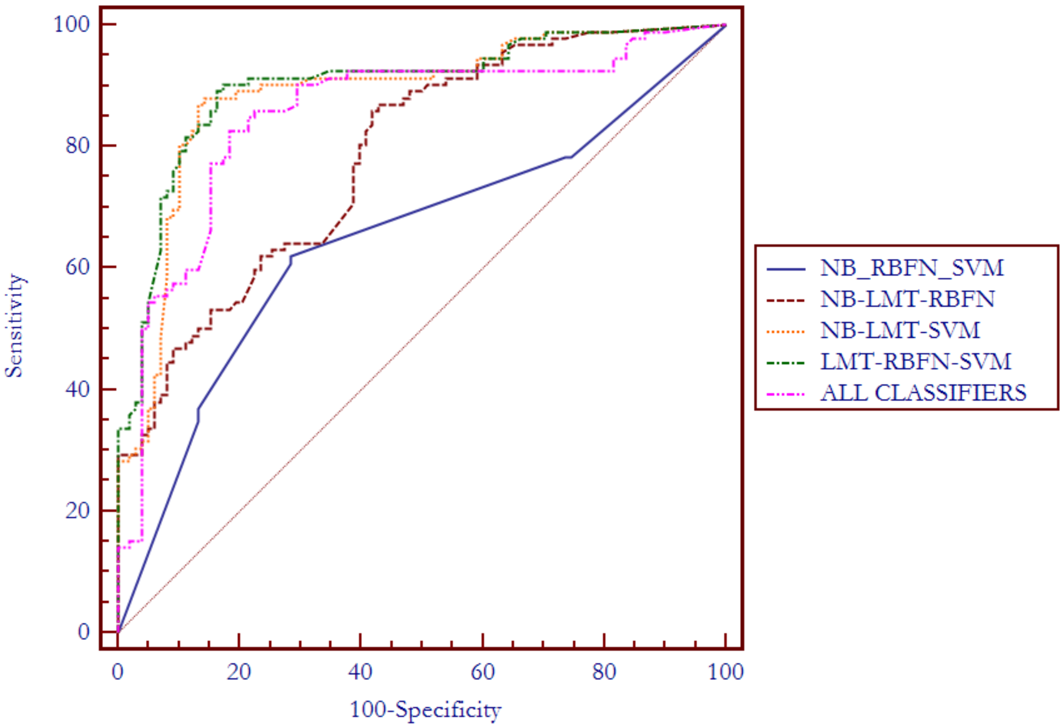
Σχήμα 6. 15: Εκτίμηση F-Score για τα δεδομένα Breast Cancer (Fusion)

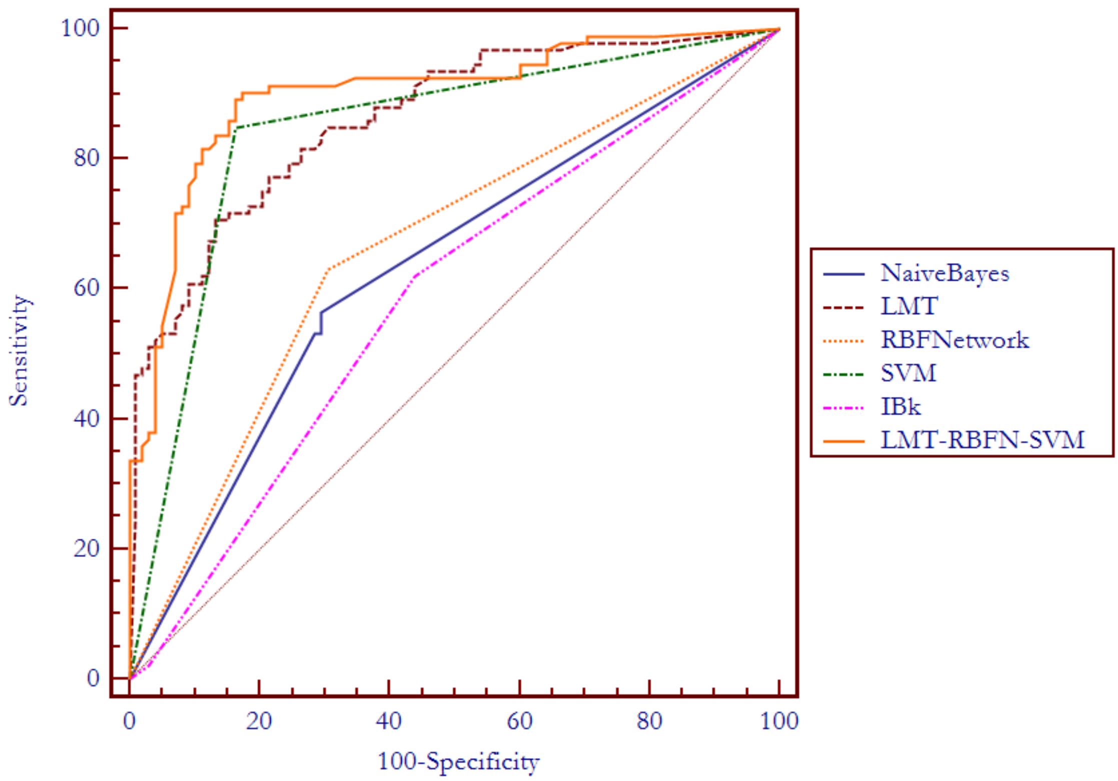
Στα σχήματα 6.16 εξετάζεται η αποτελεσματικότητα των υποδειγμάτων ταξινόμησης με τη βοήθεια του εμβαδού της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC). Προφανώς, το εμβαδόν αυτό για ένα άριστο υπόδειγμα είναι ίσο με τη μονάδα, ενώ για ένα ανεπαρκές υπόδειγμα είναι μικρότερο από 0.5. Στο πρώτο παρουσιάζονται σχηματικά οι καμπύλες ROC για διάφορους συνδυασμούς των τεχνικών ταξινόμησης, όπως αυτοί επιλέχθηκαν με την χρήση του κριτηρίου Q Statistic και συνδυάστηκαν με τη βοήθεια της τεχνικής συνδυασμού ταξινομητών Average Rule (Κεφάλαιο 5). Στο δεύτερο σχήμα παρουσιάζονται σχηματικά οι καμπύλες ROC των διαφόρων τεχνικών ταξινόμησης μεμονωμένα εν συγκρίσει με τον συνδυασμό ταξινομητών, με βάσει το πρώτο σχήμα, που σημείωσε την υψηλότερη τιμή για το εμβαδόν κάτω από την καμπύλη ROC.

Από τα σχήματα 6.16 και τον πίνακα 6.6 παρατηρείται ευκρινώς ότι ο συνδυασμός NaïveBayes – RBFNetwork – SVM και NaïveBayes – LMT – RBFNetwork δεν αποδίδουν τόσα καλά αποτελέσματα όσο οι υπόλοιποι συνδυασμοί. Την καλύτερη απόδοση με υψηλότερη τιμή εμβαδού σημειώνει ο συνδυασμός των ταξινομητών LMT – RBFNetwork – SVM (AUC=0.9), όπως ακριβώς και στα δεδομένα Pima Indians Diabetes. Ακολουθούν οι συνδυασμοί NaiveBayes – LMT – SVM (AUC=0.886) και όλοι οι ταξινομητές μαζί με AUC=0.85.

Την απόδοση ταξινόμησης του συνδυασμού όλων των ταξινομητών έρχεται να «κοντράρει», η τεχνική LMT η οποία σημειώνει ελάχιστα υψηλότερη τιμή εμβαδού (AUC=0.862) και η τεχνική SVM με ελάχιστα χαμηλότερη τιμή εμβαδού (AUC=0.842). Οι δύο αυτές τεχνικές παρουσίασαν παρόμοιες τιμές για το εμβαδόν κάτω από την καμπύλη ROC, είτε εφαρμόστηκαν στα δεδομένα Pima Indians Diabetes είτε εφαρμόστηκαν στα δεδομένα Breast Cancer.

Όσον αφορά τις υπόλοιπες εξεταζόμενες τεχνικές ταξινόμησης η μέθοδος IBk δίνει την χαμηλότερη τιμή εμβαδού (AUC=0.585), γεγονός που αποδεικνύει την αδυναμία της μεθόδου να αποδώσει υψηλά ποσοστά ακρίβειας. Ακολουθούν οι τεχνικές NaiveBayes (AUC=0.633) και RBFNetwork (AUC=0.662) με μέτριες αποδόσεις ακρίβειας.





Σχήμα 6. 16: Σχηματική απεικόνιση καμπύλης ROC για τα δεδομένα Breast Cancer (Fusion)

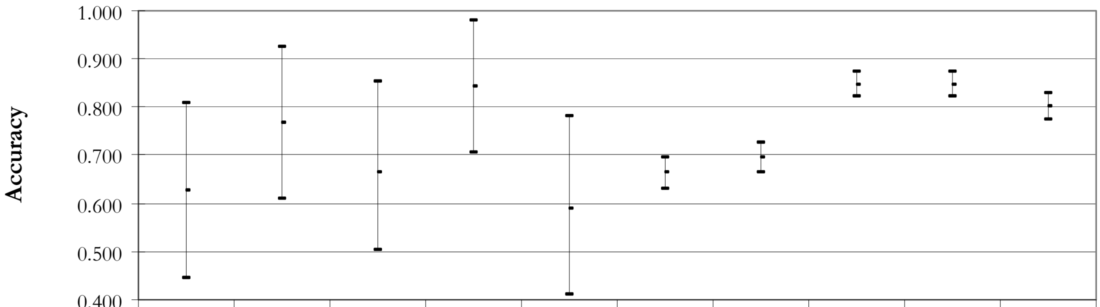
Πίνακας 6. 6: Αποτελέσματα καμπύλης ROC για τα δεδομένα Breast Cancer (Fusion)

CLASSIFIERS ARE COMPARED WITH THE BEST CLASSIFIER FUSION	AUC	95% CI
NaiveBayes	0.633	0.560 to 0.701
LMT	0.862	0.804 to 0.907
RBFNetwork	0.662	0.590 to 0.729
SVM	0.842	0.783 to 0.891
IBk	0.585	0.512 to 0.656
NB_RBFN_SVM	0.648	0.576 to 0.716
NB_LMT_RBFN	0.784	0.719 to 0.841
NB_LMT_SVM	0.886	0.832 to 0.928
LMT_RBFN_SVM	0.900	0.848 to 0.939
ALL_CLASSIFIERS	0.850	0.791 to 0.898

Στο σύνολο των σχημάτων 6.17 εξετάζονται τα διαστήματα εμπιστοσύνης της ακρίβειας ταξινόμησης των εξεταζόμενων υποδειγμάτων ταξινόμησης. Για τον υπολογισμό των διαστημάτων εμπιστοσύνης, όπως και προαναφέρθηκε, χρησιμοποιήθηκαν δύο μέθοδοι εκτίμησης αυτών (Wald Method και Wilson Score Method) και υπολογίστηκαν τα διαστήματα εμπιστοσύνης για συντελεστή εμπιστοσύνης 95%. Από τα παρακάτω σχήματα προκύπτουν ότι οι δύο εξεταζόμενες μέθοδοι εκτίμησης των διαστημάτων εμπιστοσύνης, Wald Method και Wilson Score Method, παρουσιάζουν παρόμοιες εκτιμήσεις.

Όλες οι τεχνικές ταξινόμησης σημειώνουν μεγάλο εύρος τιμών των διαστημάτων εμπιστοσύνης της ακρίβειας ταξινόμησης, εν συγκρίσει με τους συνδυασμούς αυτών όπου το εύρος των διαστημάτων μειώνεται αισθητά. Αυτό σημαίνει ότι με πιθανότητα 95% η ακρίβεια ταξινόμησης μπορεί να κυμαίνεται στο διάστημα εμπιστοσύνης, για παράδειγμα για την τεχνική SVM, μεταξύ 70.4% με 98%. Αυτό βέβαια δεν μπορεί να εγγυηθεί την αξιοπιστία της εν λόγω τεχνικής ταξινόμησης στην εκτίμηση της ακρίβειας ταξινόμησης. Αντιθέτως, οι συνδυασμοί ταξινομητών παρουσιάζουν μικρού εύρους διαστήματα εμπιστοσύνης, γεγονός που βεβαιώνει τον υπολογισμό πιο αξιόπιστων και ακριβών εκτιμήσεων ακρίβειας ταξινόμησης. Και σε αυτή την περίπτωση είναι εμφανές ότι οι συνδυασμοί ταξινομητών NaiveBayes – LMT – SVM ή LMT – RBFNetwork – SVM παρουσιάζουν υψηλές αποδόσεις ταξινόμησης καθώς σημειώνουν την υψηλότερη ακρίβεια ταξινόμησης (84.7%) και μικρότερο εύρος διαστημάτων εμπιστοσύνης [0.823, 0.872]. Συνεπώς, όπως και παρατηρήθηκε στα δεδομένα Pima Indians Diabetes, οι τεχνικές ταξινόμησης LMT και SVM αποδίδουν υψηλές ακρίβειες ταξινόμησης όταν εφαρμοστούν μεμονωμένα σε ένα δείγμα ελέγχου, αλλά επίσης ο συνδυασμός αυτών αποφέρει εξίσου πολύ ικανοποιητικά αποτελέσματα ταξινόμησης και πιο αξιόπιστα.

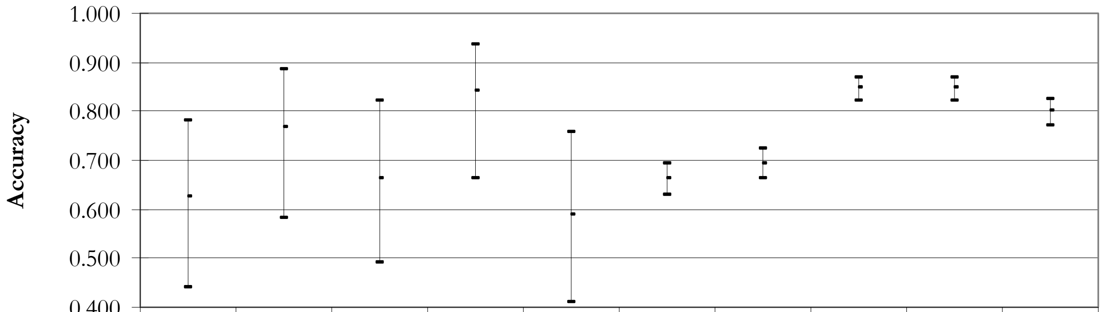
Confidence Intervals with Wald Method



	Naivebayes	LMT	RBFNetwork	SVM	IBk	NB-RBFN-SVM	NB-LMT-RBFN	NB-LMT-SVM	LMT-RBFN-SVM	ALL CLASSIFIERS
Low CI	0.444	0.609	0.502	0.704	0.412	0.631	0.664	0.823	0.823	0.773
High CI	0.809	0.923	0.854	0.980	0.781	0.695	0.726	0.872	0.872	0.827
Accuracy	0.626	0.768	0.663	0.842	0.589	0.663	0.695	0.847	0.847	0.800

Classifiers

Confidence Intervals with Wilson Score Method



	Naivebayes	LMT	RBFNetwork	SVM	IBk	NB-RBFN-SVM	NB-LMT-RBFN	NB-LMT-SVM	LMT-RBFN-SVM	ALL CLASSIFIERS
Low CI	0.439	0.582	0.489	0.664	0.411	0.631	0.663	0.822	0.822	0.772
High CI	0.782	0.884	0.822	0.935	0.758	0.694	0.725	0.870	0.870	0.825
Accuracy	0.626	0.768	0.663	0.842	0.589	0.663	0.695	0.847	0.847	0.800

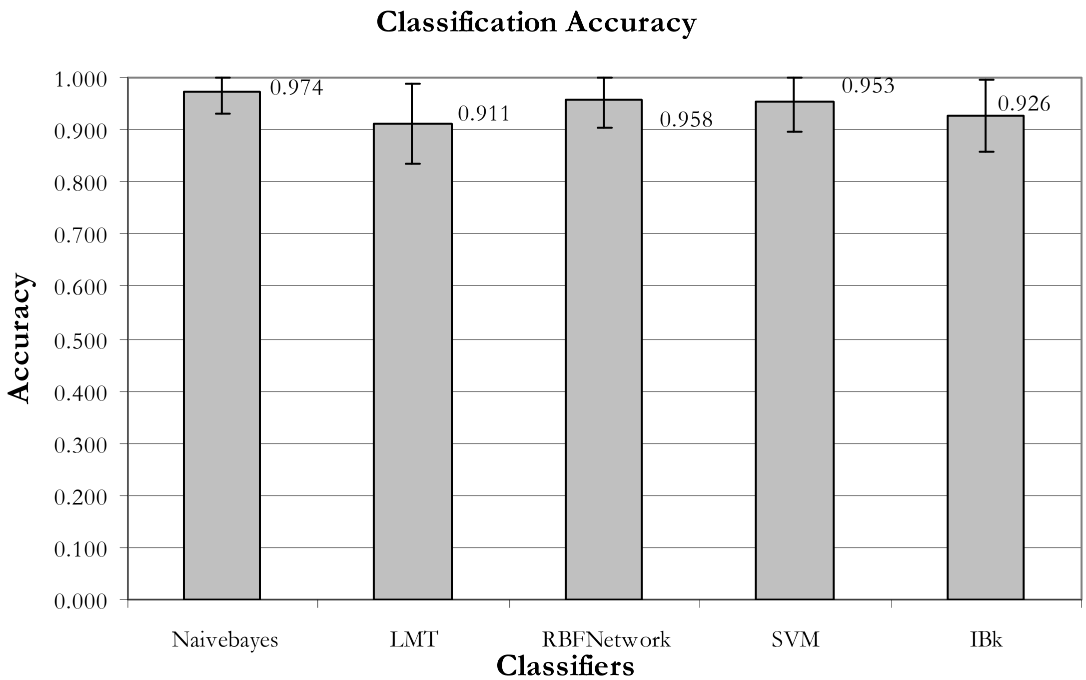
Classifiers

Σχήμα 6. 17: Εκτίμηση των διαστημάτων εμπιστοσύνης σύμφωνα με τις μεθόδους Wald και Wilson Score για επίπεδο σημαντικότητας 5% για τα δεδομένα Breast Cancer (Fusion)

6.3.3 Δεδομένα Leukemia

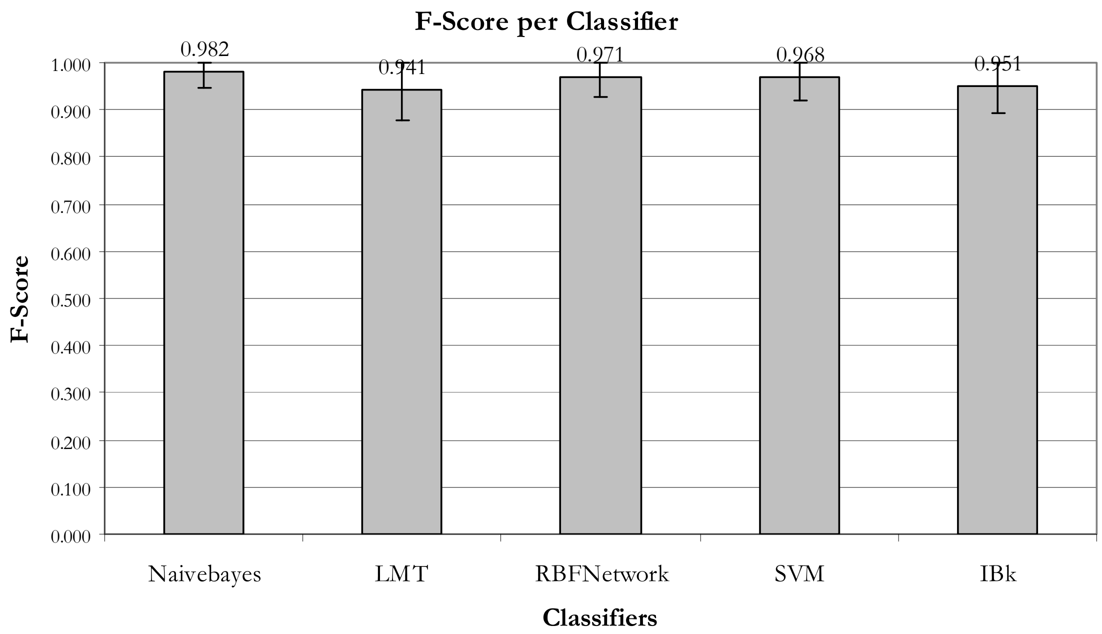
6.3.3.1 Ανάλυση αποτελεσμάτων μεμονωμένα για κάθε τεχνική ταξινόμησης (Δεδομένα: 38x715)

Στο σχήμα 6.18 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων όλων των εξεταζόμενων τεχνικών στο δείγμα εκπαίδευσης. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζει η προσέγγιση Naïve Bayes (97.4%) έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Σημαντικά μεγαλύτερο σφάλμα ταξινόμησης από την Naïve Bayes (97.4%) παρουσιάζει η προσέγγιση LMT (91.1%). Αυτό όμως δεν σημαίνει ότι η μέθοδος LMT δεν αποδίδει ικανοποιητικά αποτελέσματα ταξινόμησης, καθώς σημειώνει πολύ υψηλά ποσοστά ακρίβειας. Ακολουθούν οι τεχνικές RBFNetwork (95.8%) και SVM (95.3%) με μικρές διαφοροποιήσεις όσον αφορά την ακρίβεια ταξινόμησης και IBk (92.6%).



Σχήμα 6. 18: Ποσοστό ακρίβειας ταξινόμησης για τα δεδομένα Leukemia

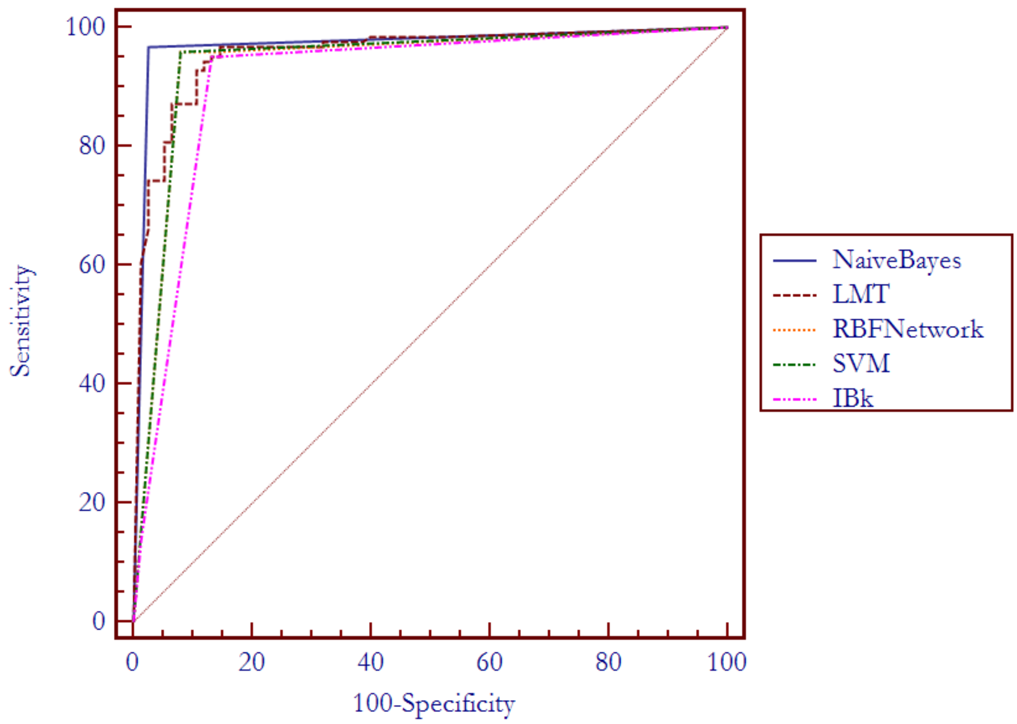
Στο σχήμα 6.19 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με την ικανότητα ταξινόμησης όλων των εξεταζόμενων τεχνικών στο δείγμα εκπαίδευσης, όπως εκτιμάται με τη βοήθεια του μέτρου F. Τα αποτελέσματα δείχνουν την υψηλή ικανότητα ταξινόμησης που παρουσιάζει η προσέγγιση Naïve Bayes (98.2%) έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Τη μικρότερη τιμή του μέτρου παρουσιάζει η προσέγγιση LMT (94.1%). Οι υπόλοιπες τρεις προσεγγίσεις με αυστηρή σειρά, RBFNetwork (97.1%), SVM (96.8%) και IBk (95.1%), παρουσιάζουν μία μέση απόδοση στην ικανότητα ταξινόμησης.



Σχήμα 6. 19: Εκτίμηση F-Score για τα δεδομένα Leukemia

Στο σχήμα 6.20 εξετάζεται η αποτελεσματικότητα των υποδειγμάτων ταξινόμησης με τη βοήθεια του εμβαδού της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC). Προφανώς, το εμβαδόν αυτό για ένα άριστο υπόδειγμα είναι ίσο με τη μονάδα, ενώ για ένα ανεπαρκές υπόδειγμα είναι μικρότερο από 0.5. Όσον αφορά τις εξεταζόμενες τεχνικές ταξινόμησης, οι εκτιμήσεις του εμβαδού της καμπύλης ROC έδωσαν ικανοποιητικά αποτελέσματα (πίνακας 6.7). Γενικότερα, με αυστηρή σειρά, οι τιμές που σημειώθηκαν για κάθε ταξινομητή είναι Naïve Bayes

(AUC=0.971), LMT (AUC=0.956), SVM (AUC=0.940) και RBFNetwork (AUC=0.939) με μικρές διαφοροποιήσεις και τέλος, ακολουθεί με εξίσου καλή απόδοση η IBk (AUC=0.913).



Σχήμα 6. 20: Εμβαδόν κάτω από την καμπύλη ROC για τα δεδομένα Leukemia

Πίνακας 6. 7: Αποτελέσματα καμπύλης ROC για τα δεδομένα Leukemia

Classifiers	AUC	95% CI
NaiveBayes	0.971	0.937 to 0.989
LMT	0.956	0.917 to 0.980
RBFNetwork	0.939	0.897 to 0.968
SVM	0.940	0.898 to 0.969
IBk	0.913	0.865 to 0.948

Στο σύνολο των σχημάτων 6.21 εξετάζονται τα διαστήματα εμπιστοσύνης της ακρίβειας ταξινόμησης των εξεταζόμενων υποδειγμάτων ταξινόμησης. Για τον υπολογισμό των διαστημάτων εμπιστοσύνης, όπως και προαναφέρθηκε, χρησιμοποιήθηκαν δύο μέθοδοι εκτίμησης αυτών (Wald Method και Wilson Score Method) και υπολογίστηκαν τα διαστήματα εμπιστοσύνης για διάφορους συντελεστές εμπιστοσύνης. Συγκεκριμένα, στα σχήματα απεικονίζονται τα διαστήματα εμπιστοσύνης της ακρίβειας των ταξινομητών για συντελεστή εμπιστοσύνης 99.5% (α=0.005, 1η δεκάδα κάθε διαγράμματος), 97.5% (α=0.025, 2η δεκάδα κάθε διαγράμματος), 95% (α=0.05, 3η δεκάδα κάθε διαγράμματος) και 90% (α=0.1, 4η και τελευταία δεκάδα κάθε διαγράμματος). Από τα παρακάτω σχήματα προκύπτουν οι εξής διαπιστώσεις:

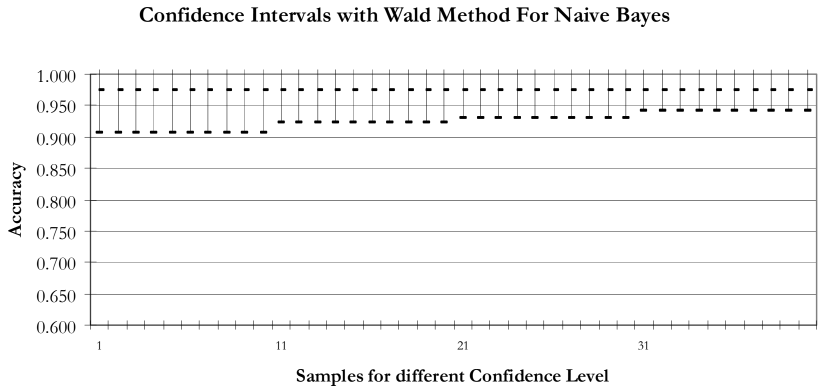
1. Οι δύο εξεταζόμενες μέθοδοι εκτίμησης των διαστημάτων εμπιστοσύνης, Wald Method και Wilson Score Method, παρουσιάζουν μικρές διαφοροποιήσεις στην εκτίμηση των διαστημάτων εμπιστοσύνης, που αποδεικνύουν όμως την προτίμηση της μεθόδου Wald. Η μέθοδος Wald δίνει πιο υψηλό και μικρότερο εύρος τιμών ακρίβειας από ότι η μέθοδο Wilson Score, όπως και παρατηρείται από τα παρακάτω διαγράμματα.

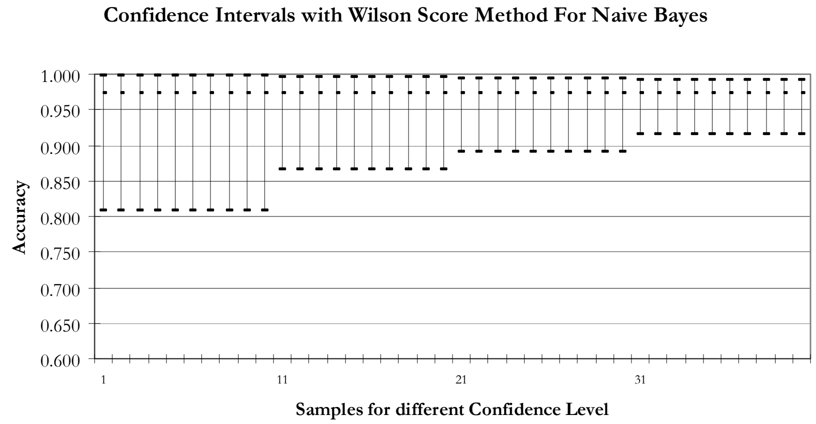
2. Όσο μικραίνει ο συντελεστής εμπιστοσύνης τόσο μικραίνει και το εύρος των διαστημάτων εμπιστοσύνης. Συνεπώς, ο συντελεστής εμπιστοσύνης επηρεάζει τα άκρα του διαστήματος εμπιστοσύνης και όσο αυξάνεται ο συντελεστής, αυξάνονται και οι πιθανότητες να βρίσκεται μέσα στο διάστημα εμπιστοσύνης η πραγματικής τιμή της εκτιμώμενης παραμέτρου.

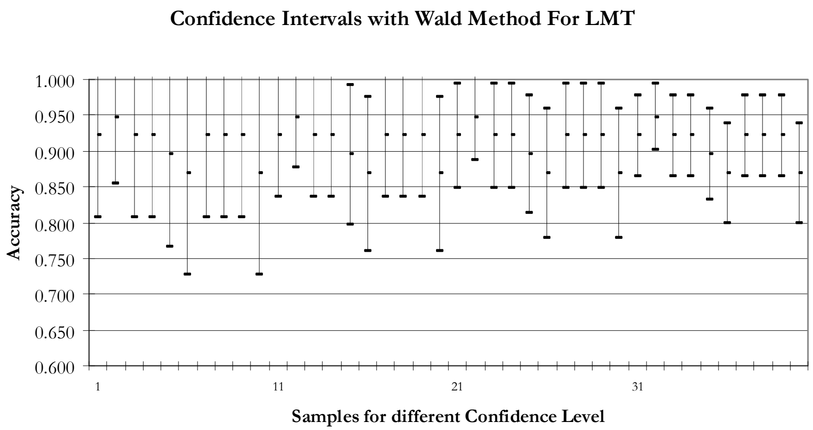
3. Όσον αφορά τις χρησιμοποιούμενες τεχνικές ταξινόμησης και για συντελεστή εμπιστοσύνης 95% με τη χρήση της μεθόδου Wald, η πραγματική διαφορά στην εκτίμηση της μέγιστης από την ελάχιστη τιμή ακρίβειας, που μπορεί να αποδώσει το κάθε υπόδειγμα ταξινόμησης για τα πραγματικά δεδομένα Leukemia, βρίσκεται στο διάστημα, κατά προσέγγιση μέση τιμής, [0.931, 1] για την τεχνική Naïve Bayes, [0.835, 0.986] για την LMT, [0.906, 1] για την τεχνική RBFNetwork, [0.896, 1] για την SVM και [0.857, 0.996] για την τεχνική IBk.

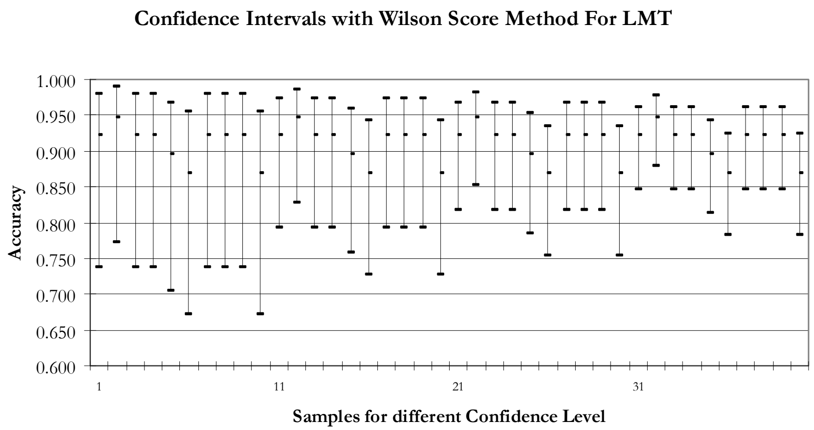
4. Η τεχνική ταξινόμησης LMT παρουσιάζει αισθητά διαφορετικές εκτιμήσεις διαστημάτων εμπιστοσύνης ως προς την ακρίβεια ταξινόμησης για διαφορετικά δείγματα, ίδιου μεγέθους και από το ίδιο σύνολο δεδομένων.

5. Η τεχνική Naïve Bayes παρουσιάζει σταθερή μεταβολή των άκρων των διαστημάτων εμπιστοσύνης και σημειώνει το μικρότερο εύρος διαστήματος εμπιστοσύνης της ακρίβειας από όλες τις τεχνικές ταξινόμησης.

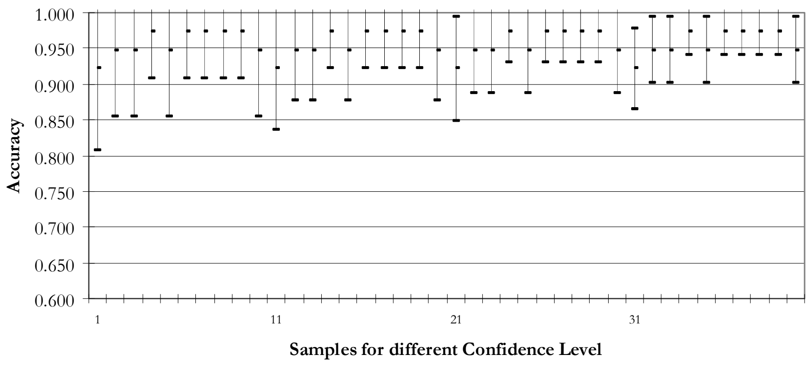

Confidence Intervals with Wald Method For Naive Bayes
1.000
0.950
0.900
0.850
0.800
0.750
0.700
0.650
0.600
Accuracy
1
11
21
31
Samples for different Confidence Level


Confidence Intervals with Wilson Score Method For Naive Bayes
1.000
0.950
0.900
0.850
0.800
0.750
0.700
0.650
0.600
Accuracy
1
11
21
31
Samples for different Confidence Level

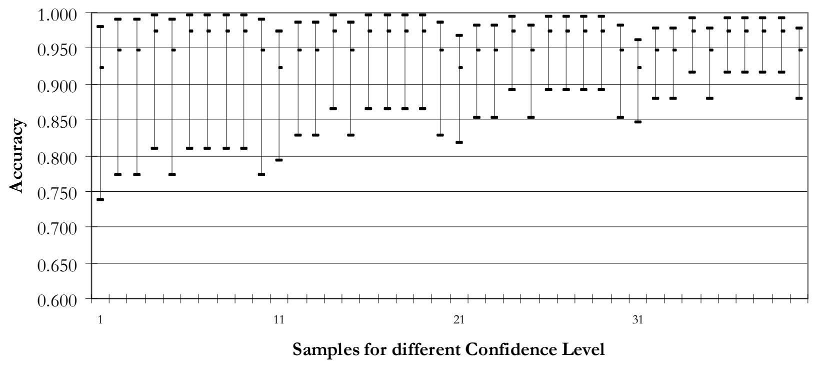

Confidence Intervals with Wald Method For LMT
1.000
0.950
0.900
0.850
0.800
0.750
0.700
0.650
0.600
Accuracy
1
11
21
31
Samples for different Confidence Level


Confidence Intervals with Wilson Score Method For LMT
1.000
0.950
0.900
0.850
0.800
0.750
0.700
0.650
0.600
Accuracy
1
11
21
31
Samples for different Confidence Level

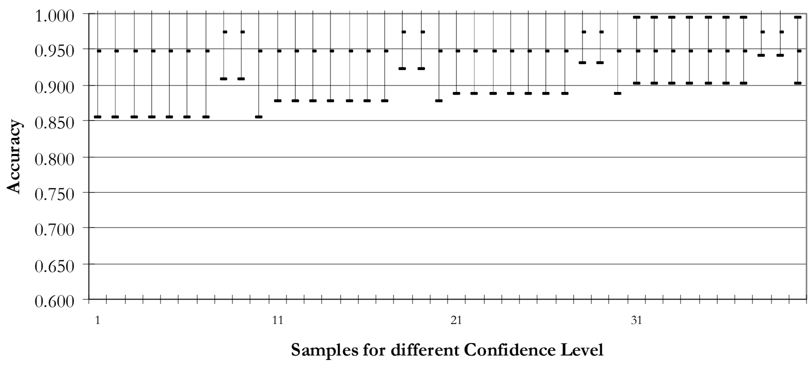
Confidence Intervals with Wald Method For RBFNetwork



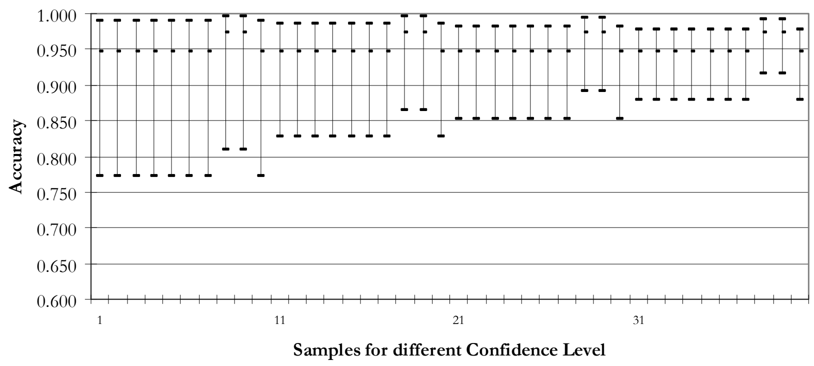
Confidence Intervals with Wilson Score Method For RBFNetwork

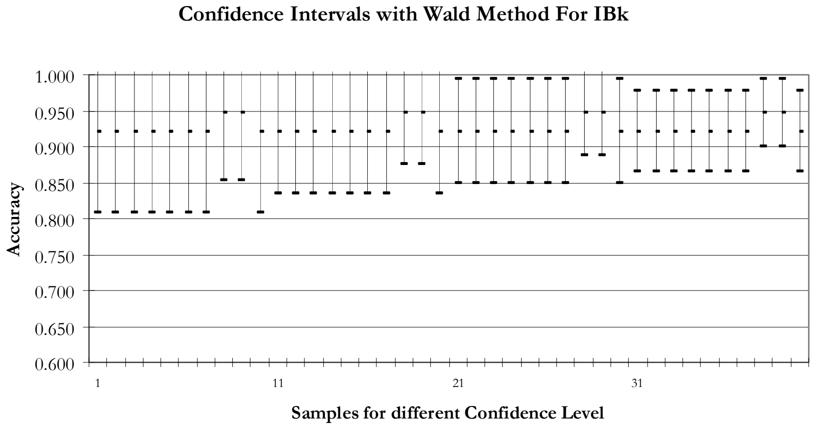


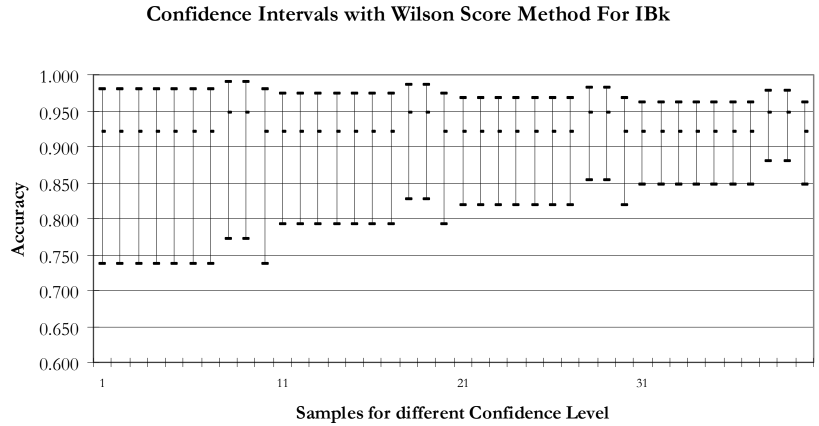
Confidence Intervals with Wald Method For SVM

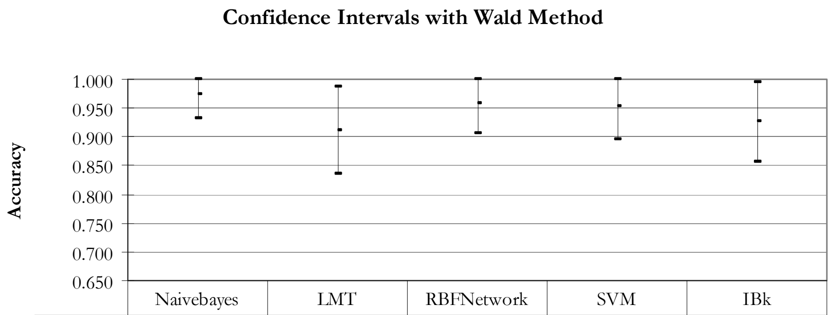


Confidence Intervals with Wilson Score Method For SVM



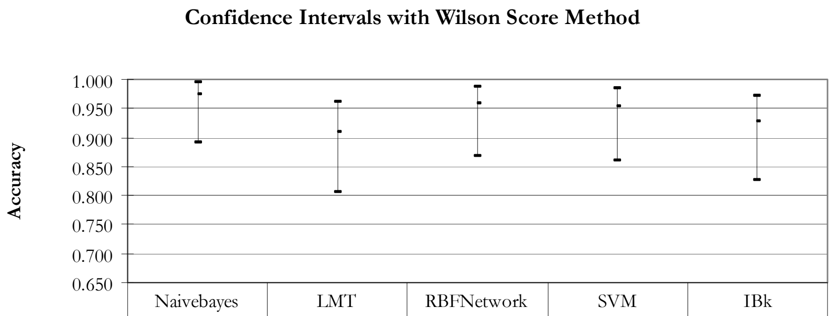






	Naivebayes	LMT	RBFNetwork	SVM	IBk
Low CI	0.931	0.835	0.906	0.896	0.857
High CI	1.000	0.986	1.000	1.000	0.996
Accuracy	0.974	0.911	0.958	0.953	0.926

Classifiers



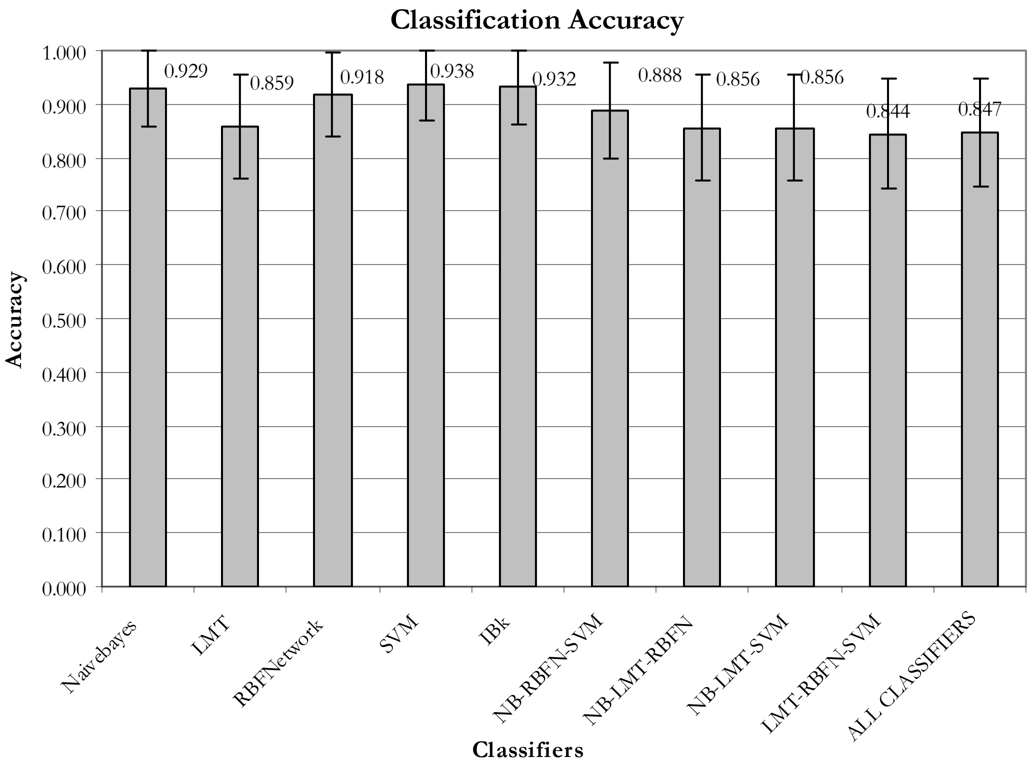
	Naivebayes	LMT	RBFNetwork	SVM	IBk
Low CI	0.890	0.805	0.868	0.860	0.825
High CI	0.994	0.961	0.987	0.985	0.971
Accuracy	0.974	0.911	0.958	0.953	0.926

Classifiers

Σχήμα 6. 21: Εκτίμηση των διαστημάτων εμπιστοσύνης σύμφωνα με τις μεθόδους Wald και Wilson Score για τα δεδομένα Leukemia

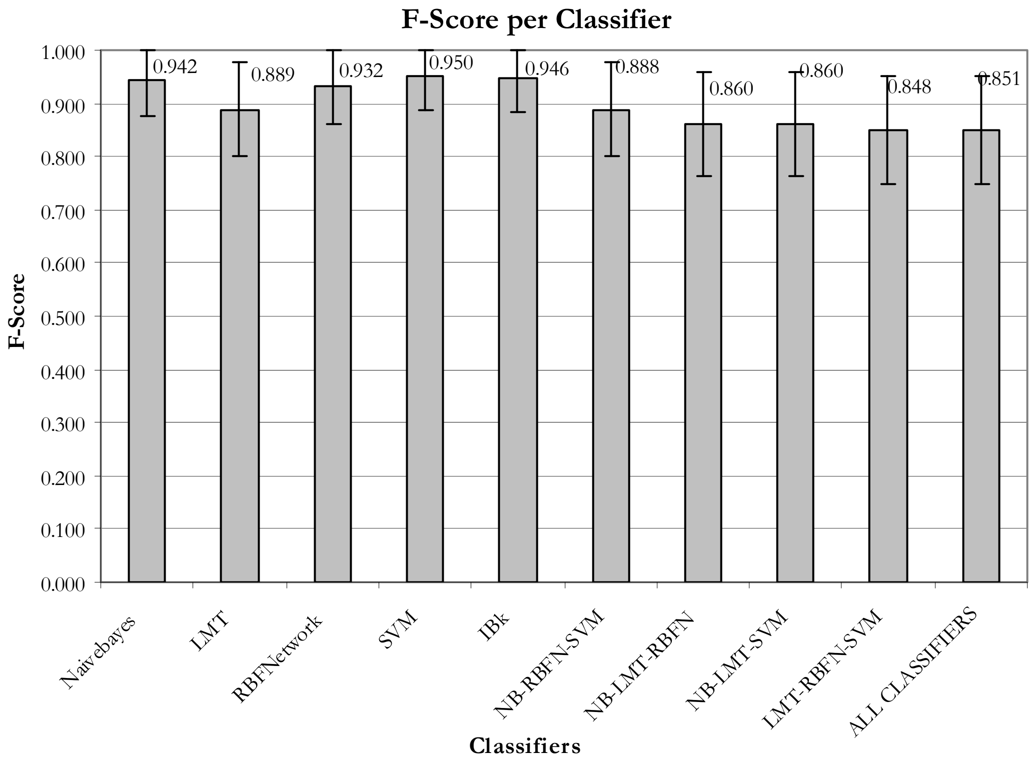
6.3.3.2 Ανάλυση αποτελεσμάτων της διαδικασία Fusion για κάθε τεχνική ταξινόμησης (Δεδομένα: 34x715)

Στο σχήμα 6.22 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό των σωστών ταξινομήσεων των μεμονωμένα εξεταζόμενων τεχνικών στο δείγμα ελέγχου (independent test set) εν συγκρίσει με τους πέντε (5) επιλεγμένους συνδυασμούς ταξινομητών. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζουν οι πέντε (5) συνδυασμούς ταξινομητών, αλλά την εξίσου υψηλή ακρίβεια ταξινόμησης των υποδειγμάτων ταξινόμησης, στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Συγκεκριμένα, η υψηλότερη ακρίβεια ταξινόμησης σημειώθηκε από τη τεχνική SVM (93.8%) και ακολουθούν IBk (93.2%), NaiveBayes (92.9%), RBFNetwork (91.8%), οι συνδυασμοί NaiveBayes – LMT – RBFNetwork (77.2%), NaiveBayes – RBFNetwork – SVM (88.8%), η τεχνική LMT (85.9%), οι συνδυασμοί NaiveBayes – LMT – SVM ή NaiveBayes – LMT – RBFNetwork (85.6%), όλοι οι ταξινομητές (84.7%) και τέλος, ο συνδυασμός LMT – RBFNetwork – SVM (84.4%). Γενικότερα, παρατηρήθηκαν υψηλές αποδόσεις με μικρές διαφοροποιήσεις.



Σχήμα 6. 22: Ποσοστό ακρίβειας ταξινόμησης για τα δεδομένα Leukemia (Fusion)

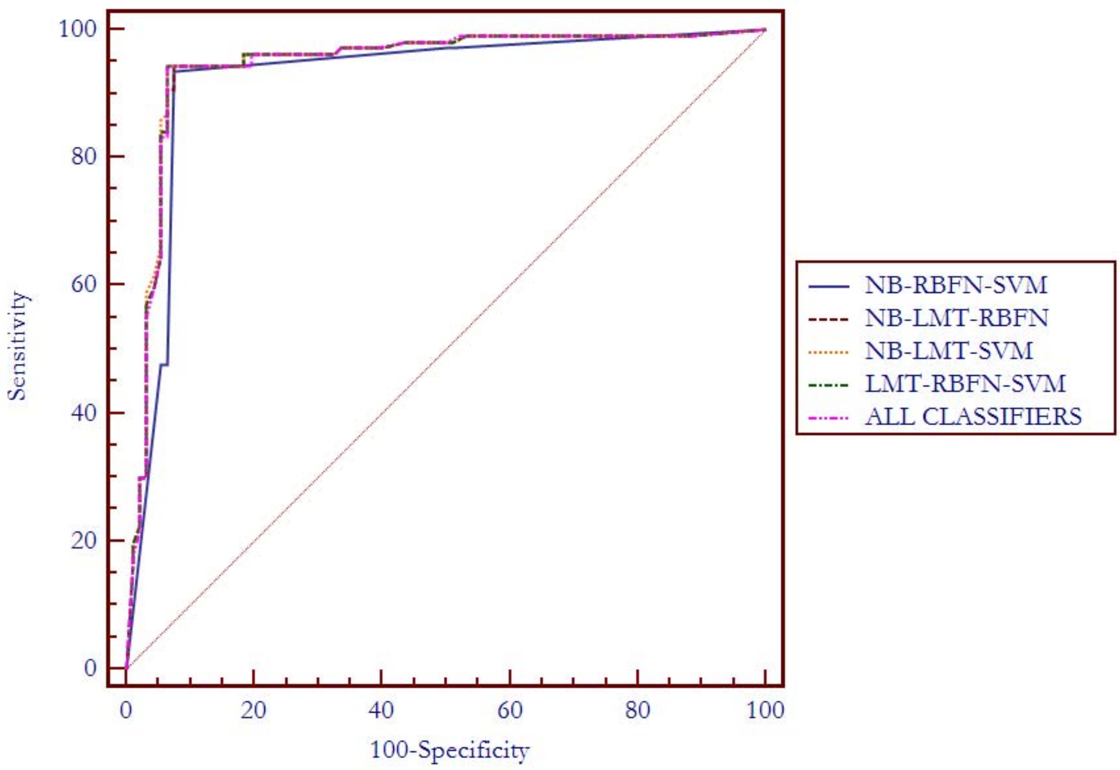
Στο σχήμα 6.23 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με την ικανότητα ταξινόμησης όλων των εξεταζόμενων τεχνικών στο δείγμα ελέγχου (independent test set) εν συγκρίσει με την επίδραση του συνόλου των ταξινομητών, ύστερα από εφαρμογή της διαδικασίας Fusion, σε κάθε υπόδειγμα ταξινόμησης, αντίστοιχα. Η παραπάνω εκτίμηση πραγματοποιείται με τη βοήθεια του μέτρου F. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζουν όλες οι προσεγγίσεις, με την τεχνική SVM (95%) να σημειώσει την υψηλότερο τιμή για το μέτρο F, έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων. Οι υπόλοιπες προσεγγίσεις, με αυστηρή σειρά σημειώνουν τα εξής ποσοστά: IBk (94.6%), Naïve Bayes (94.2%), RBFNetwork (93.2%) και LMT (88.9%). Και ακολουθούν οι συνδυασμοί των ταξινομητών με την ακόλουθη σειρά, NaiveBayes – RBFNetwork – SVM (88.8%), NaiveBayes – LMT – SVM ή NaiveBayes – LMT – RBFNetwork (86%), όλοι οι ταξινομητές (84.8%) και τέλος, ο συνδυασμός LMT – RBFNetwork – SVM (85.1%). Τα αποτελέσματα των συνδυασμών των ταξινομητών παραμένουν σταθερά ανεξαρτήτου συνόλου δεδομένων. Αντιθέτως οι τεχνικές ταξινόμησης εκτιμούν διαφορετικά το κάθε σύνολο δεδομένων.

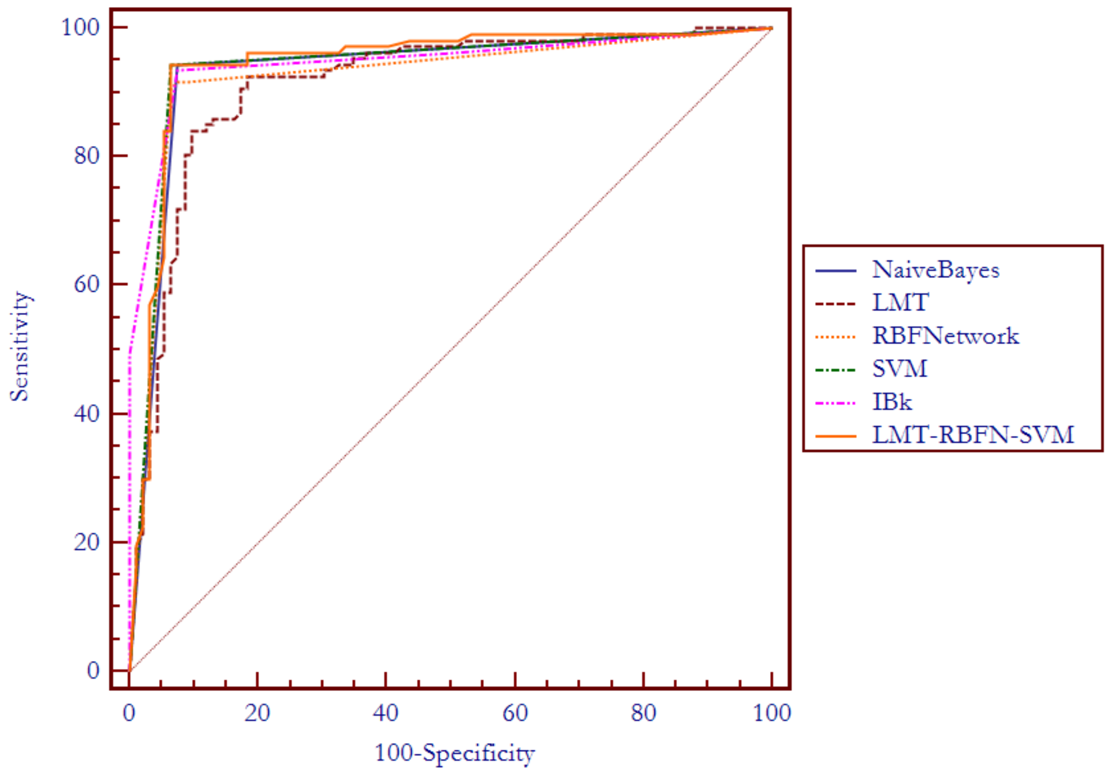


Σχήμα 6. 23: Εκτίμηση F-Score για τα δεδομένα Leukemia (Fusion)

Στα σχήματα 6.24 εξετάζεται η αποτελεσματικότητα των υποδειγμάτων ταξινόμησης με τη βοήθεια του εμβαδού της περιοχής κάτω από την καμπύλη ROC (Area Under Curve, AUC). Προφανώς, το εμβαδόν αυτό για ένα άριστο υπόδειγμα είναι ίσο με τη μονάδα, ενώ για ένα ανεπαρκές υπόδειγμα είναι μικρότερο από 0.5. Στο πρώτο παρουσιάζονται σχηματικά οι καμπύλες ROC για διάφορους συνδυασμούς των τεχνικών ταξινόμησης, όπως αυτοί επιλέχθηκαν με την χρήση του κριτηρίου Q Statistic και συνδυάστηκαν με τη βοήθεια της τεχνική συνδυασμού ταξινομητών Average Rule (Κεφάλαιο 5). Στο δεύτερο σχήμα παρουσιάζονται σχηματικά οι καμπύλες ROC των διαφόρων τεχνικών ταξινόμησης μεμονωμένα εν συγκρίσει με τον συνδυασμό ταξινομητών, με βάσει το πρώτο σχήμα, που σημείωσε την υψηλότερη τιμή για το εμβαδόν κάτω από την καμπύλη ROC.

Από τα σχήματα 6.24 και τον πίνακα 6.8 παρατηρείται ευκρινώς ότι όλοι οι συνδυασμοί ταξινομητών, καθώς και οι τεχνικές ταξινόμησης μεμονωμένα σημειώνουν υψηλές αποδόσεις αποτελεσματικότητας ταξινόμησης με μικρές διαφοροποιήσεις μεταξύ τους.





Σχήμα 6. 24: Σχηματική απεικόνιση καμπύλης ROC για τα δεδομένα Leukemia (Fusion)

Πίνακας 6. 8: Αποτελέσματα καμπύλης ROC για τα δεδομένα Leukemia (Fusion)

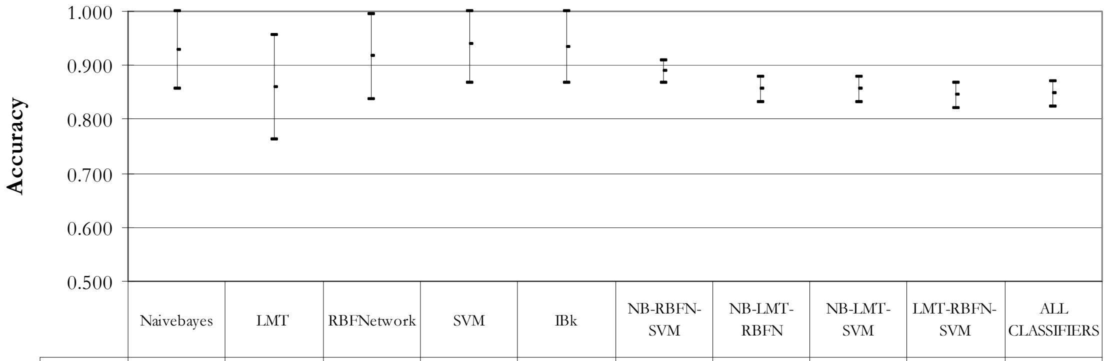
CLASSIFIERS ARE COMPARED WITH THE BEST CLASSIFIER FUSION	AUC	95% CI
NaiveBayes	0.934	0.890 to 0.964
LMT	0.913	0.865 to 0.948
RBFNetwork	0.924	0.878 to 0.957
SVM	0.939	0.897 to 0.968
IBk	0.947	0.906 to 0.974
NB_RBFN_SVM	0.923	0.876 to 0.956
NB_LMT_RBFN	0.943	0.901 to 0.971
NB_LMT_SVM	0.944	0.902 to 0.971
LMT_RBFN_SVM	0.943	0.901 to 0.971
ALL_CLASSIFIERS	0.943	0.901 to 0.971

Στο σύνολο των σχημάτων 6.25 εξετάζονται τα διαστήματα εμπιστοσύνης της ακρίβειας ταξινόμησης των εξεταζόμενων υποδειγμάτων ταξινόμησης. Για τον υπολογισμό των διαστημάτων εμπιστοσύνης, όπως και προαναφέρθηκε, χρησιμοποιήθηκαν δύο μέθοδοι εκτίμησης αυτών (Wald Method και Wilson Score Method) και υπολογίστηκαν τα διαστήματα εμπιστοσύνης για συντελεστή εμπιστοσύνης 95%. Από τα παρακάτω σχήματα προκύπτουν ότι οι δύο εξεταζόμενες μέθοδοι εκτίμησης των διαστημάτων εμπιστοσύνης, Wald Method και Wilson Score Method, παρουσιάζουν παρόμοιες εκτιμήσεις, με μικρές διαφοροποιήσεις μεταξύ τους.

Όλες οι τεχνικές ταξινόμησης σημειώνουν μεγάλο εύρος τιμών των διαστημάτων εμπιστοσύνης της ακρίβειας ταξινόμησης, εν συγκρίσει με τους συνδυασμούς αυτών όπου το εύρος των διαστημάτων μειώνεται αισθητά. Αυτό σημαίνει ότι με πιθανότητα 95% η ακρίβεια ταξινόμησης για τη Wald Method μπορεί να κυμαίνεται στο διάστημα εμπιστοσύνης, για παράδειγμα για την τεχνική SVM, μεταξύ 86.9% με 100%. Αυτό βέβαια δεν μπορεί να εγγυηθεί την αξιοπιστία της εν λόγω τεχνικής ταξινόμησης στην εκτίμηση της ακρίβειας ταξινόμησης. Αντιθέτως, οι συνδυασμοί ταξινομητών παρουσιάζουν μικρού εύρους διαστήματα εμπιστοσύνης, γεγονός που βεβαιώνει τον υπολογισμό πιο αξιόπιστων και ακριβών εκτιμήσεων ακρίβειας ταξινόμησης.

Γενικότερα, στα συγκεκριμένα δεδομένα ελέγχου δεν παρουσιάζονται σημαντικές διαφοροποιήσεις μεταξύ της εφαρμογής των τεχνικών ταξινόμησης μεμονωμένα ή του συνδυασμού αυτών. Για όλα τα υποδείγματα ταξινόμησης και τους συνδυασμούς αυτών, κατά τη χρήση των δεδομένων ελέγχου Leukemia, σημειώθηκαν υψηλές ακρίβειες ταξινόμησης, υψηλές τιμές του κριτηρίου F Score και υψηλές τιμές του εμβαδού κάτω από την καμπύλη ROC με μικρές διαφοροποιήσεις μεταξύ τους. Η διαφορά μεταξύ της εφαρμογής των ταξινομητών μεμονωμένα στην αντιμετώπιση του προβλήματος ταξινόμησης και του συνδυασμού αυτών, η οποία θεωρείται και σημαντική, είναι η εγκυρότητα των αποτελεσμάτων στις τιμές της ακρίβειας ταξινόμησης που σημειώνεται με βάσει το συνδυασμό των ταξινομητών, γιατί όσο το εύρος του διαστήματος μικραίνει (στενεύει), γίνεται καλύτερη εκτίμηση της ακρίβειας ταξινόμησης. Γενικότερα, όπως παρατηρήθηκε και στα δεδομένα εκπαίδευσης, εάν αυξηθεί ο συντελεστής εμπιστοσύνης, που σημαίνει ότι απαιτείται μια ασφαλέστερη πρόβλεψη, τότε το εύρος του διαστήματος εμπιστοσύνης αυξάνεται. Αυτό συμβαίνει γιατί με το ίδιο δείγμα τιμών απαιτείται ο υπολογισμός ενός ασφαλέστερου άνω και κάτω ορίου για την ακρίβεια ταξινόμησης. Για να ελαττωθεί λοιπόν, η πιθανότητα η ακρίβεια ταξινόμησης να μην βρίσκεται εντός των ορίων του διαστήματος εμπιστοσύνης, αυτό που γίνεται είναι ότι αυξάνεται του εύρος του διαστήματος εμπιστοσύνης.

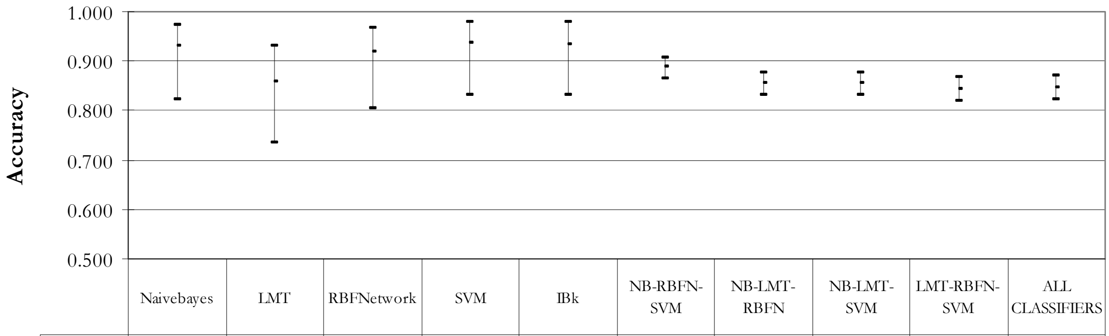
Confidence Intervals with Wald Method



	Naivebayes	LMT	RBFNetwork	SVM	IBk	NB-RBFN-SVM	NB-LMT-RBFN	NB-LMT-SVM	LMT-RBFN-SVM	ALL CLASSIFIERS
Low CI	0.858	0.762	0.837	0.869	0.868	0.867	0.832	0.832	0.820	0.823
High CI	1.000	0.957	0.994	1.000	1.000	0.909	0.879	0.879	0.868	0.871
Accuracy	0.929	0.859	0.918	0.938	0.932	0.888	0.856	0.856	0.844	0.847

Classifiers

Confidence Intervals with Wilson Score Method



	Naivebayes	LMT	RBFNetwork	SVM	IBk	NB-RBFN-SVM	NB-LMT-RBFN	NB-LMT-SVM	LMT-RBFN-SVM	ALL CLASSIFIERS
Low CI	0.822	0.736	0.803	0.831	0.831	0.865	0.831	0.831	0.818	0.821
High CI	0.974	0.930	0.966	0.978	0.978	0.908	0.878	0.878	0.867	0.870
Accuracy	0.929	0.859	0.918	0.938	0.932	0.888	0.856	0.856	0.844	0.847

Classifiers

Σχήμα 6. 25: Εκτίμηση των διαστημάτων εμπιστοσύνης σύμφωνα με τις μεθόδους Wald και Wilson Score για επίπεδο σημαντικότητας 5% για τα δεδομένα Leukemia (Fusion)

6.4 Συγκεντρωτικά αποτελέσματα και βασικές επισημάνσεις

Ο πειραματικός σχεδιασμός που παρουσιάστηκε παραπάνω συνέβαλε στην αναλυτική διερεύνηση της αποτελεσματικότητας του συνδυασμού διαφόρων υποδειγμάτων ταξινόμησης σε σύγκριση με την εφαρμογή αυτών μεμονωμένα στη αντιμετώπιση προβλημάτων ταξινόμησης. Τα αποτελέσματα του πειραματικού σχεδιασμού βοήθησαν στην εξαγωγή χρήσιμων πληροφοριών και συνολικά τα βασικά συμπεράσματα και οι κύριες επισημάνσεις του πειραματικού αυτού σχεδιασμού παρουσιάζονται παρακάτω.

Στον πίνακα 6.9 παρουσιάζονται τα συγκεντρωτικά αποτελέσματα ταξινόμησης του κάθε ταξινομητή μεμονωμένα για το δείγμα εκπαίδευσης. Από τα αποτελέσματα συμπεραίνεται ότι όσο μειώνεται ο αριθμός των ασθενών (παρατηρήσεις) και αυξάνεται ο αριθμός των γονιδίων (χαρακτηριστικά), τόσο αυξάνεται η απόδοση των υποδειγμάτων ταξινόμησης και συνεπώς, αυξάνεται και η ακρίβεια ταξινόμησης. Εξετάζοντας βέβαια, τα αποτελέσματα του δείγματος δεδομένων Pima Indians Diabetes, τα οποία αποτελούνται από πολύ μεγαλύτερο αριθμό, προς εκτίμηση ταξινόμησης παρατηρήσεων, από ότι τα δύο άλλα σύνολα δεδομένων, διαπιστώνεται η υψηλή αποτελεσματικότητα που παρουσιάζει η προσέγγιση RBFNetwork στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησής τους. Σημαντικά μεγαλύτερο σφάλμα ταξινόμησης από την RBFNetwork παρουσιάζει η προσέγγιση IBk. Οι υπόλοιπες τρεις προσεγγίσεις Naïve Bayes, LMT και SVM παρουσιάζουν ικανοποιητικά αποτελέσματα με μικρές διαφοροποιήσεις όσον αφορά την ακρίβεια ταξινόμησης. Επομένως, μπορεί να θεωρηθεί ότι οι εν λόγω τεχνικές ταξινόμησης αποδίδουν παρόμοια αποτελέσματα. Αξίζει να σημειωθεί ότι η μέθοδος που παρουσιάζει σταθερά, σε όλα τα δεδομένα που εξετάστηκαν, καλύτερα αποτελέσματα εν συγκρίσει με τις άλλες τεχνικές ταξινόμησης και καλύτερα διαστήματα εμπιστοσύνης είναι η Naïve Bayes. Ελάχιστες είναι οι φορές που η εν λόγω μέθοδος κατατάχθηκε στη δεύτερη θέση με μικρή έως μηδαμινή διαφορά από την πρώτη, σε κατάταξη τεχνική, σύμφωνα πάντοτε με την εκτίμηση της ακρίβειας ταξινόμησης.

Όσον αφορά τα διαστήματα εμπιστοσύνης συμπεραίνεται ότι η τεχνική LMT δεν είναι ιδιαίτερα αξιόπιστη ως προς την αντιμετώπιση προβλημάτων ταξινόμησης, καθώς παρουσιάζει σημαντικές διακυμάνσεις από δείγμα σε δείγμα. Σαφώς καλύτερα αποτελέσματα αποδίδουν οι υπόλοιπες τεχνικές, με την SVM να υπερτερεί έναντι των τεχνικών Naïve Bayes και RBFNetwork με μικρή διαφορά. Γενικότερα, οι τεχνικές SVM, Naïve Bayes και RBFNetwork λειτουργούν αρκετά καλά στην αντιμετώπιση προβλημάτων ταξινόμησης. Η LMT δεν έχει σταθερή απόδοση ταξινόμησης, καθώς επηρεάζεται από τα διάφορα δείγματα.

Η μέθοδος που παρουσιάζει τις μικρότερες ακρίβειες στην εκτίμηση της αποτελεσματικότητας στην προβληματική της ταξινόμησης είναι ο αλγόριθμος IBk. Γενικότερα, ο αλγόριθμος IBk, μπορεί να παρουσιάζει το λιγότερο υπολογιστικό φόρτο, αλλά δεν θεωρείται και ούτε έχει αποδείξει σε παλαιότερες έρευνες, την ικανότητά του στην αντιμετώπιση προβλημάτων ταξινόμησης. Γενικότερα, παρουσιάζει πρόβλημα όταν χρησιμοποιείται μεμονωμένα σε προβλήματα ταξινόμησης.

Επίσης, παρατηρείται ότι σε όλα τα σύνολα των δεδομένων εκπαίδευσης οι δύο εξεταζόμενες μέθοδοι εκτίμησης των διαστημάτων εμπιστοσύνης, Wald Method και Wilson Score Method, παρουσίασαν παρόμοιες εκτιμήσεις με μηδαμινές διαφοροποιήσεις μεταξύ τους. Με μικρή διαφορά υπερτερεί η Wald Method, η οποία υπολογίζει ελάχιστα μικρότερα διαστήματα εμπιστοσύνης από ότι η Wilson Score Method. Επιπλέον, παρατηρήθηκε ότι όσο μικραίνει ο συντελεστής εμπιστοσύνης τόσο μικραίνει και το εύρος των διαστημάτων εμπιστοσύνης. Στην παρούσα διατριβή ελέγχθηκαν τέσσερις (4) διαφορετικές τιμές του συντελεστή εμπιστοσύνης 99.5%, 97.5%, 95% και 90%. Από την ανάλυση διαπιστώθηκε ότι υπήρχαν μεγάλες διαφοροποιήσεις ως προς το εύρος των διαστημάτων για τους διάφορους συντελεστές. Μικρές διαφοροποιήσεις σημειώθηκαν για συντελεστή εμπιστοσύνης μεταξύ 95% και 90% και σαφώς καλύτερο εύρος διαστήματος από τους υπόλοιπους δύο (2) συντελεστές. Λόγω των μικρών διαφοροποιήσεων και βάσει της διεθνής βιβλιογραφίας επιλέχθηκε η υπόλοιπη ανάλυση του δείγματος ελέγχου να πραγματοποιηθεί για συντελεστή εμπιστοσύνης 95% και με τη βοήθεια της μεθόδου Wald, ως λιγότερο πολύπλοκη και χρονοβόρα στον υπολογισμό της.

Πίνακας 6. 9: Συγκεντρωτικός πίνακας αποτελεσμάτων για το δείγμα εκπαίδευσης

TRAIN SET	**Pima Indian Diabetes (770x3)**		**Breast Cancer (78x605)**		**Leukemia (38x715)**	
Classification Accuracy	Naïve Bayes	76.5%	Naïve Bayes	86.2%	Naïve Bayes	97.4%
	LMT	76.4%	LMT	80%	LMT	91.1%
	RBFNetwork	76.9%	RBFNetwork	85.3%	RBFNetwork	95.8%
	SVM	76.4%	SVM	87.6%	SVM	95.3%

	IBk	70.5%	IBk	81.3%	IBk	92.6%
F Score	Naïve Bayes	62.6%	Naïve Bayes	88.2%	Naïve Bayes	98.2%
	LMT	62%	LMT	82.7%	LMT	94.1%
	RBFNetwork	63.4%	RBFNetwork	87.6%	RBFNetwork	97.1%
	SVM	61%	SVM	89.3%	SVM	96.8%
	IBk	57.3%	IBk	84.1%	IBk	95.1%
AUC ROC Curve	Naïve Bayes	0.843	Naïve Bayes	0.874	Naïve Bayes	0.971
	LMT	0.842	LMT	0.869	LMT	0.956
	RBFNetwork	0.828	RBFNetwork	0.899	RBFNetwork	0.939
	SVM	0.713	SVM	0.890	SVM	0.940
	IBk	0.701	IBk	0.814	IBk	0.913
Confidence Interval Average (συντελεστής 95% - Wald Method)	Naïve Bayes	[0.740, 0.790]	Naïve Bayes	[0.797, 0.926]	Naïve Bayes	[0.931, 1]
	LMT	[0.739, 0.789]	LMT	[0.726, 0.874]	LMT	[0.835, 0.986]
	RBFNetwork	[0.744, 0.794]	RBFNetwork	[0.787, 0.918]	RBFNetwork	[0.906, 1]
	SVM	[0.739, 0.789]	SVM	[0.814, 0.937]	SVM	[0.896, 1]
	IBk	[0.678, 0.732]	IBk	[0.740, 0.885]	IBk	[0.857, 0.996]

Στον πίνακα 6.10 παρουσιάζονται τα συγκεντρωτικά αποτελέσματα ταξινόμησης του κάθε ταξινομητή μεμονωμένα για το δείγμα ελέγχου και του συνδυασμού των ταξινομητών με τη μεγαλύτερη διαφορετικότητα, όπως αυτοί επιλέχθηκαν με τη βοήθεια του pairwise κριτηρίου διαφορετικότητας Q Statistic. Υπενθυμίζεται ότι ο συνδυασμός των επιλεγμένων ταξινομητών πραγματοποιήθηκε με τη χρήση της Soft Labeling μεθόδου Average Rule (Κεφάλαιο 5).

Από τα αποτελέσματα συμπεραίνεται ότι όσο μειώνεται ο αριθμός των ασθενών (παρατηρήσεις) και αυξάνεται ο αριθμός των γονιδίων (χαρακτηριστικά), τόσο αυξάνεται η απόδοση των υποδειγμάτων ταξινόμησης και συνεπώς, αυξάνεται και η ακρίβεια ταξινόμησης. Στο μεγαλύτερο ποσοστό των δειγμάτων ελέγχου, υψηλή ακρίβεια ταξινόμησης και μικρότερου εύρος διαστήματα εμπιστοσύνης εν συγκρίσει με την εφαρμογή μεμονωμένων ταξινομητών σημειώνεται όταν γίνουν οι συνδυασμοί NaiveBayes – LMT – SVM ή LMT – RBFNetwork – SVM, ανεξαρτήτου αριθμού παρατηρήσεων Ο συνδυασμός όλων των ταξινομητών παρουσιάζει μηδαμινή διαφοροποίηση εν συγκρίσει με την εφαρμογή μεμονωμένων τεχνικών ταξινόμησης. Το γεγονός ότι συνδυάστηκαν όλοι οι ταξινομητές, ενώ η αποτελεσματικότητα ταξινόμησης δεν βελτιώθηκε αισθητά, οφείλεται στη συμμετοχή της μεθόδου IBk στη διαδικασία Fusion, η οποία εμποδίζει τη βελτίωση και αντιμετώπιση του προβλήματος ταξινόμησης. Αυτό άλλωστε φαίνεται και από τα ποσοστά ακρίβειας ταξινόμησης που σημειώνει η τεχνική IBk στην εφαρμογή των δεδομένων ελέγχου και εκπαίδευσης. Παρατηρείται επίσης, ότι οι συνδυασμοί NaïveBayes – RBFNetwork – SVM και NaïveBayes – LMT – RBFNetwork δεν αποδίδουν τόσα καλά αποτελέσματα εν συγκρίσει με τους υπόλοιπους συνδυασμούς ή μεμονωμένα με κάθε ταξινομητή.

Από τα δείγματα ελέγχου και το συνδυασμό των ταξινομητών προκύπτει ότι ενώ η τεχνική LMT παρουσίαζε σημαντικές διακυμάνσεις στο εύρος των διαστημάτων εμπιστοσύνης και όχι πάντοτε υψηλές αποδόσεις από δείγμα σε δείγμα, συνδυάζεται επιτυχώς με άλλες τεχνικές ταξινόμησης και βελτιώνει κατά πολύ την αποτελεσματικότητα της ακρίβειας ταξινόμησης. Συγκεκριμένα, παρατηρήθηκε ότι η μέθοδος LMT αποδίδει υψηλότερα αποτελέσματα εάν συνδυαστεί με το υπόδειγμα ταξινόμησης SVM. Οι δύο αυτές τεχνικές σε συνδυασμό αποδίδουν υψηλές ακρίβειες, ενώ προέρχονται από διαφορετικά πεδία εφαρμογής και με υψηλή διαφορετικότητα. Συγκεκριμένα, ο συνδυασμός ενός προβλήματος βελτιστοποίησης τετραγωνικού προγραμματισμού (SVM) και ενός δέντρου λογιστικής παλινδρόμησης μπορεί να αποδώσει, όπως διαπιστώθηκε υψηλή αποτελεσματικότητα.

Επίσης, παρατηρήθηκε ότι όταν οι δύο αυτές τεχνικές SVM και LMT συνδυαστούν είτε με το υπόδειγμα ταξινόμησης NaïveBayes είτε με το RBFNework αποδίδουν ακριβώς τα ίδια αποτελέσματα. Συνεπώς, οι δύο αυτές τεχνικές, NaïveBayes και RBFNework, λειτουργούν ακριβώς κατά το ίδιο τρόπο, κάτι το οποίο ήταν αναμενόμενο καθώς ακολουθούν ίδιες υποθέσεις κατά την αντιμετώπιση της προβληματικής της ταξινόμησης. Συγκεκριμένα, και για τους δύο ταξινομητές θεωρείται πρώτον ότι, οι τιμές των χαρακτηριστικών είναι κανονικά κατανεμημένες και δεύτερον ότι, η συνάρτηση μάθησης είναι Gaussian.

Όσον αφορά την τεχνική ταξινόμησης IBk μπορεί να παρουσιάζει πάντοτε τις μικρότερες επιδόσεις στην αντιμετώπιση προβλημάτων ταξινόμησης, αλλά κυμαίνεται πάντα στα ίδια επίπεδα, ακόμα και για διαφορετικού μεγέθους παρατηρήσεων ή χαρακτηριστικών ή ιδιοτήτων δεδομένων ταξινόμησης.

Η διαφορά μεταξύ της εφαρμογής των ταξινομητών μεμονωμένα στην αντιμετώπιση του προβλήματος ταξινόμησης και του συνδυασμού αυτών, η οποία θεωρείται και σημαντική, είναι η εγκυρότητα των αποτελεσμάτων στις τιμές της ακρίβειας ταξινόμησης που σημειώνεται με βάσει το συνδυασμό των ταξινομητών, γιατί όσο το εύρος του διαστήματος μικραίνει (στενεύει), γίνεται καλύτερη εκτίμηση της ακρίβειας ταξινόμησης. Γενικότερα, όπως παρατηρήθηκε και στα δεδομένα εκπαίδευσης, εάν αυξηθεί ο συντελεστής εμπιστοσύνης, που σημαίνει ότι απαιτείται μια ασφαλέστερη πρόβλεψη, τότε το εύρος του διαστήματος εμπιστοσύνης αυξάνεται. Αυτό συμβαίνει γιατί με το ίδιο δείγμα τιμών απαιτείται ο υπολογισμός ενός ασφαλέστερου άνω και κάτω ορίου για την ακρίβεια ταξινόμησης. Για να ελαττωθεί λοιπόν, η πιθανότητα η ακρίβεια ταξινόμησης να μην βρίσκεται εντός των ορίων του διαστήματος εμπιστοσύνης, αυτό που γίνεται είναι ότι αυξάνεται του εύρος του διαστήματος εμπιστοσύνης. Όλες οι τεχνικές ταξινόμησης σημειώνουν μεγάλο εύρος τιμών των διαστημάτων εμπιστοσύνης της ακρίβειας ταξινόμησης, εν συγκρίσει με τους συνδυασμούς αυτών όπου το εύρος των διαστημάτων μειώνεται αισθητά, γεγονός που βεβαιώνει τον υπολογισμό πιο αξιόπιστων και ακριβών εκτιμήσεων ακρίβειας ταξινόμησης.

Πίνακας 6. 10: Συγκεντρωτικός πίνακας για το ανεξάρτητο δείγμα ελέγχου

TEST SET	**Pima Indian Diabetes (600x3)**		**Breast Cancer (19x605)**		**Leukemia (34x715)**	
%Classification Accuracy	Naïve Bayes	76.6%	Naïve Bayes	62.6%	Naïve Bayes	92.9%
	LMT	76.3%	LMT	76.8%	LMT	85.9%
	RBFNetwork	76.8%	RBFNetwork	66.3%	RBFNetwork	91.8%
	SVM	76.4%	SVM	84.2%	SVM	93.8%
	IBk	70.2%	IBk	58.9%	IBk	93.2%
	NB_RBFN_SVM	70.8%	NB_RBFN_SVM	66.3%	NB_RBFN_SVM	88.8%
	NB_LMT_RBFN	77.2%	NB_LMT_RBFN	69.5%	NB_LMT_RBFN	85.6%
	NB_LMT_SVM	77.1%	NB_LMT_SVM	84.7%	NB_LMT_SVM	85.6%
	LMT_RBFN_SVM	77.1%	LMT_RBFN_SVM	84.7%	LMT_RBFN_SVM	84.4%
	ALL_CLASSIFIERS	76.8%	ALL_CLASSIFIERS	80%	ALL_CLASSIFIERS	84.7%

F Score	Naïve Bayes	62.4%	Naïve Bayes	26.4%	Naïve Bayes	94.2%
	LMT	61.3%	LMT	73.7%	LMT	88.9%
	RBFNetwork	63.2%	RBFNetwork	41.5%	RBFNetwork	93.2%
	SVM	60.2%	SVM	82.4%	SVM	95%
	IBk	56.2%	IBk	61%	IBk	94.6%
	NB_RBFN_SVM	70.7%	NB_RBFN_SVM	63.6%	NB_RBFN_SVM	88.8%
	NB_LMT_RBFN	77.1%	NB_LMT_RBFN	66.3%	NB_LMT_RBFN	86%
	NB_LMT_SVM	77%	NB_LMT_SVM	83.8%	NB_LMT_SVM	86%
	LMT_RBFN_SVM	77%	LMT_RBFN_SVM	84%	LMT_RBFN_SVM	84.8%
	ALL_CLASSIFIERS	76.7%	ALL_CLASSIFIERS	79.1%	ALL_CLASSIFIERS	85.1%
AUC ROC Curve	Naïve Bayes	0.823	Naïve Bayes	0.633	Naïve Bayes	0.934
	LMT	0.851	LMT	0.862	LMT	0.913
	RBFNetwork	0.841	RBFNetwork	0.662	RBFNetwork	0.924
	SVM	0.773	SVM	0.842	SVM	0.939
	IBk	0.684	IBk	0.585	IBk	0.947
	NB_RBFN_SVM	0.713	NB_RBFN_SVM	0.648	NB_RBFN_SVM	0.923
	NB_LMT_RBFN	0.842	NB_LMT_RBFN	0.784	NB_LMT_RBFN	0.843
	NB_LMT_SVM	0.842	NB_LMT_SVM	0.886	NB_LMT_SVM	0.944
	LMT_RBFN_SVM	0.848	LMT_RBFN_SVM	0.900	LMT_RBFN_SVM	0.943
	ALL_CLASSIFIERS	0.837	ALL_CLASSIFIERS	0.850	ALL_CLASSIFIERS	0.943
Confidence Interval Average (συντελεστής 95% - Wald Method)	Naïve Bayes	[0.738 0.795]	Naïve Bayes	[0.444 0.809]	Naïve Bayes	[0.858 1]
	LMT	[0.734 0.791]	LMT	[0.609 0.923]	LMT	[0.762 0.957]
	RBFNetwork	[0.737 0.794]	RBFNetwork	[0.502 0.854]	RBFNetwork	[0.837 0.994]
	SVM	[0.735 0.792]	SVM	[0.704 0.980]	SVM	[0.869 1]
	IBk	[0.670 0.731]	IBk	[0.412 0.781]	IBk	[0.868 1]
	NB_RBFN_SVM	[0.677 0.738]	NB_RBFN_SVM	[0.631 0.695]	NB_RBFN_SVM	[0.867 0.909]
	NB_LMT_RBFN	[0.744 0.8]	NB_LMT_RBFN	[0.664 0.726]	NB_LMT_RBFN	[0.832 0.879]
	NB_LMT_SVM	[0.742 0.799]	NB_LMT_SVM	[0.823 0.872]	NB_LMT_SVM	[0.832 0.879]
	LMT_RBFN_SVM	[0.743 0.799]	LMT_RBFN_SVM	[0.823 0.872]	LMT_RBFN_SVM	[0.820 0.868]
	ALL_CLASSIFIERS	[0.739 0.76]	ALL_CLASSIFIERS	[0.773 0.827]	ALL_CLASSIFIERS	[0.823 0.871]

Κ ε φ ά λ α ι ο 7^{ο}

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Το πρόβλημα της ταξινόμησης παρουσίαζε ανέκαθεν αυξημένο ερευνητικό και πρακτικό ενδιαφέρον. Η διαπίστωση αυτή επιβεβαιώνεται από την πληθώρα των πρακτικών εφαρμογών που η προβληματική της ταξινόμησης παρουσιάζει και οι οποίες μπορούν να αντιμετωπιστούν πραγματοποιώντας απόλυτες συγκρίσεις με προκαθορισμένα δείγματα εκμάθησης και δείγματα ελέγχου. Οι έρευνες για την αντιμετώπιση των προβλημάτων ταξινόμησης επικεντρώνονται πρώτον, στην εφαρμογή κατάλληλων τεχνικών για την ανάπτυξη υποδειγμάτων ταξινόμησης (Κεφάλαιο 3), δεύτερον, στην επιλογή κατάλληλων αλγορίθμων επιλογής χαρακτηριστικών (Κεφάλαιο 3), οι οποίοι σε συνδυασμό με διαδεδομένες τεχνικές ταξινόμησης μπορούν να αποδώσουν μεγάλες ακρίβειες ταξινόμησης και τρίτον, στο συνδυασμό των διαφόρων υποδειγμάτων ταξινόμησης (Κεφάλαιο 5).

Επί δεκαετίες η ανάπτυξη των υποδειγμάτων ταξινόμησης βασίζονταν σε στατιστικές προσεγγίσεις. Οι προσεγγίσεις αυτές συνέβαλαν στην κατανόηση της προβληματικής της ταξινόμησης, των χαρακτηριστικών και των ιδιαιτεροτήτων που παρουσιάζει, καθώς επίσης και των κανόνων που ακολουθούν τα αναπτυσσόμενα υποδείγματα ταξινόμησης. Το πρόβλημα, κατά την εφαρμογή των στατιστικών τεχνικών ταξινόμησης, έγκειται στις περιοριστικές υποθέσεις που διέπουν την κάθε μια από αυτές τις τεχνικές, με αποτέλεσμα τη δημιουργία κινήτρου για τη ανάπτυξη εναλλακτικών προσεγγίσεων.

Ένα από τα σημαντικότερα θέματα που ανακύπτουν κατά τη διαδικασία ανάπτυξης υποδειγμάτων ταξινόμησης αφορά την επιλογή του κατάλληλου συνόλου χαρακτηριστικών, δηλαδή των παραγόντων που περιγράφουν το εξεταζόμενο φαινόμενο και καθορίζουν την ταξινόμηση των αντικειμένων. Στόχος είναι ο εντοπισμός ενός περιορισμένου συνόλου χαρακτηριστικών τα οποία περιγράφουν επαρκώς το φαινόμενο και μπορούν να οδηγήσουν στην ανάπτυξη αξιόπιστων υποδειγμάτων ταξινόμησης. Η χρησιμότητα τέτοιων αναλύσεων γίνεται άμεσα εμφανής σε διάφορα επίπεδα, καθώς περιορίζεται ο όγκος της πληροφορίας που πρέπει να εξεταστεί τόσο κατά την ανάπτυξη όσο και κατά τη χρήση του υποδείγματος, μειώνεται ο υπολογιστικός φόρτος που απαιτείται για την ανάπτυξη του υποδείγματος και περιορίζεται ο θόρυβος που πιθανόν υπάρχει στα δεδομένα λόγω της ύπαρξης χαρακτηριστικών που δεν αποδίδουν αξιόπιστη πληροφορία. Στη συγκεκριμένη πειραματική ανάλυση χρησιμοποιήθηκαν αλγόριθμοι επιλογής χαρακτηριστικών που ακολουθούν διαδικασίες filter (Κεφάλαιο 3).

Η παρούσα, βέβαια, εργασία δεν επικεντρώθηκε μόνο στη μαθηματική διατύπωση των χρησιμοποιούμενων μεθόδων και στη μείωση της εξεταζόμενης πληροφορίας, αλλά πραγματοποιήθηκε και μία διεξοδική πειραματική ανάλυση για τη διερεύνηση της δυνατότητας του συνδυασμού με άλλες εναλλακτικές μεθοδολογίες. Ο συνδυασμός των τεχνικών ταξινόμησης (Κεφάλαιο 5) αποτελεί χρήσιμο εργαλείο για την αντιμετώπιση προβλημάτων ταξινόμησης κατά τα οποία εντοπίζονται προβλήματα τόσο ως προς τον ορισμό των κατηγοριών ταξινόμησης όσο και ως προς τις επιδόσεις των εναλλακτικών δραστηριοτήτων στα κριτήρια αξιολόγησης. Για να εντοπιστούν όμως, εκείνοι οι ταξινομητές που όταν συνδυαστούν αποδίδουν υψηλή απόδοση ταξινόμησης απαιτείται η χρήση διαφόρων κριτηρίων διαφορετικότητας (diversity measures). Στην παρούσα εργασία χρησιμοποιήθηκε μόνο το κριτήριο pairwise Q Statistic. Και αφού εντοπίστηκαν οι πλέον κατάλληλοι ταξινομητές εφαρμόστηκε η μέθοδος Soft labeling Average Rule, η οποία και επιλέχθηκε από μια πληθώρα μεθόδων συνδυασμού ταξινομητών (Κεφάλαιο 5).

Ο εκτενής αυτός πειραματικός σχεδιασμός (Κεφάλαιο 6) επέτρεψε τη μέτρηση της ακρίβειας της εξεταζόμενης προσέγγισης ως προς την ταξινόμηση των εναλλακτικών δραστηριοτήτων, οι επιδόσεις των οποίων έχουν προκαθορισμένη και ελεγχόμενη μορφή. Η εκτίμηση της αποτελεσματικότητας των μεμονωμένων ταξινομητών αλλά και του συνδυασμού αυτών δεν περιορίστηκε μόνο στον υπολογισμό της ακρίβειας ταξινόμησης (Κεφάλαιο 2), αλλά εκτιμήθηκαν η τιμή του κριτηρίου F Score (Κεφάλαιο 2) ως αρμονικός μέσος για την εκτίμηση της ικανότητας ταξινόμησης των υποδειγμάτων ή του συνδυασμού αυτών, το εμβαδόν κάτω από την καμπύλη ROC ως διαδικασία ελέγχου της αποτελεσματικότητας ταξινόμησης, αλλά και τα διαστήματα εμπιστοσύνης με τη Wald και Wilson Score μέθοδο (Κεφάλαιο 4) ως εκτιμητές της ακρίβειας ταξινόμησης.

Η διερεύνηση των παραπάνω ερευνητικών κατευθύνσεων συμβάλλει ουσιαστικά στην πληρέστερη αντιμετώπιση των προβλημάτων λήψης αποφάσεων που απαιτούν τη χρησιμοποίηση της προβληματικής της ταξινόμησης, μέσω ενός ολοκληρωμένου μεθοδολογικού πλαισίου το οποίο θα είναι σε θέση να οδηγήσει στην ανάπτυξη υποδειγμάτων υψηλής αποτελεσματικότητας/ ακρίβειας, παρέχοντας παράλληλα ένα διευρυμένο σύνολο πληροφοριών για την υποστήριξη των αποφάσεων υπό καθεστώς βεβαιότητας και αβεβαιότητας. Κύριος στόχος είναι η διερεύνηση εκείνων των συνθηκών υπό τις οποίες οι διάφοροι αλγόριθμοι σε σχέση με τεχνικές ταξινόμησης και την επιλογή κατάλληλων μεθόδων συνδυασμού των εξεταζόμενων τεχνικών ταξινόμησης θα παρουσιάζουν βελτιωμένα αποτελέσματα. Απαραίτητος είναι και ο προσδιορισμός εκείνων των συνδυασμών που θα συμβάλλουν στην υψηλότερη αποτελεσματικότητα ταξινόμησης. Η πειραματική ανάλυση της επίδρασης του συνδυασμού των αλγορίθμων στο αποτέλεσμα της ταξινόμησης, συμβάλλει στην καλύτερη κατανόηση της λειτουργίας των εξεταζόμενων αλγορίθμων ταξινόμησης και στην εξαγωγή χρήσιμων συμπερασμάτων ως προς τις πρακτικές τους εφαρμογές.

Συνοψίζοντας, θα μπορούσαν να χρησιμοποιηθούν επιπρόσθετοι αλγόριθμοι ταξινόμησης, όπως τα νευρωνικά δίκτυα, η γραμμική και τετραγωνική ανάλυση, τα δέντρα ταξινόμησης και παλινδρόμησης, κ.ά., όπως επίσης και επιπρόσθετοι αλγόριθμοι επιλογής χαρακτηριστικών, όπως γενετικοί αλγόριθμοι, ο αλγόριθμος του κλάδου και φράγματος, καθώς και τεχνικές γραμμικού προγραμματισμού και οι οποίοι να ακολουθούν, εκτός από διαδικασίες filter και διαδικασίες wrapper (Κεφάλαιο 3) για την επιλογή του πλέον βέλτιστου συνόλου χαρακτηριστικών. Ενδιαφέρον επίσης, θα ήταν να συνδυαστούν, εκτός από τεχνικές ταξινόμησης, και διάφοροι αλγόριθμοι επιλογής χαρακτηριστικών για τη βελτίωση των αποτελεσμάτων όσον αφορά το λόγο ακρίβειας, όπως επίσης και την επιλογή των καλύτερων χαρακτηριστικών κατά τη διαδικασία της ταξινόμησης και την επίτευξη απόλυτης συμφωνίας μεταξύ των αλγορίθμων επιλογής χαρακτηριστικών. Επιπλέον, ιδιαίτερα ενδιαφέρον θα ήταν να συγκριθούν τα διάφορα κριτήρια μέτρησης της διαφορετικότητας των ταξινομητών και οι διάφοροι μέθοδοι συνδυασμού των υποδειγμάτων ταξινόμησης, έτσι ώστε να διαπιστωθεί η χρησιμότητα του κάθε κριτηρίου και η αποτελεσματικότητα της κάθε μεθόδου, αντίστοιχα στη βελτίωση της αποτελεσματικότητας της προβληματικής της ταξινόμησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Almuallim, H. and Dietterich, T.G. (1991), "Learning with many irrelevant features", in: Proceedings of the 9th National Conference on Artificial Intelligence, volume 2, Anaheim, CA, AAAI Press, pp. 547-552.

[2] Almuallim, H. and Dietterich, T.G. (1994), "Learning boolean concepts in the presence of many irrelevant features", Artificial Intelligence, 69(1-2), pp. 279-305.

[3] Alpaydin Ethem (2004), Intorduction to Machine Learning, MIT Press.

[4] Altman, E. I., Marco, G. and Varetto, F. (1994), "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (The Italian experience)", Journal of Banking and Finance 18, pp.505-529.

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

[6] Ben-Bassat, M. (1982), "Use of distance measures, information measures and error bounds in feature evaluation", in: Krishnaiah, P.R. and Kanal, L.N. (eds.), Handbook of Statistics, North Holland, pp. 773-791.

[7] Bishop M. Christopher (2006), Pattern Recognition and Machine Learning, Springer.

[8] Blum, A.L. and Langley, P. (1997), "Selection of relevant features and examples in machine learning", Artificial Intelligence, 97(1-2), pp. 245-271.

[9] Blyth, C.R. and Still, H.A. (1983), "Binomial confidence intervals", Journal of the American Statistical Association, 78, pp. 108-116.

[10] Boomsma, A. (2005), "Confidence intervals for a binomial proportion", Department of Statistics & Measurement Theory, University of Groningen. Available at http://www.ppsw.rug.nl/~boomsma/confbin.pdf

[11] Boutsikas M.V. (2005), «Διαστήματα Εμπιστοσύνης», Σημειώσεις Στατιστικής ΙΙΙ, Τμήμα Οικονομικής Επιστήμης, πανεπιστήμιο Πειραιώς. Available at http://www.unipi.gr/faculty/mbouts/statIII/Ch5_Confidence_Intervals_v2.pdf

[12] Breiman, L., Friedman, J.H., Olsen, R.A. and Stone, C.J. (1984), Classification and Regression Trees, Pacific Grove, California. Available at http://www.ifpri.org/pubs/microcom/micro3.pdf.

[13] Brown, L.D., Cai, T.T. and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion", Statistical Science 16(2), pp. 101-117.

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

[15] Cardie, C. (1993), "Using decision trees to improve case-based learning", in: Utgoff (ed.), Proceedings of the 10th International Conference on Machine Learning, Morgan Kaufmann, pp. 25-32.

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

[17] Chang, C. Y. (1973), "Dynamic programming as applied to feature selection in a pattern recognition system", IEEE Transactions on Systems, Man and Cybernetics, 3, pp. 166-171.

[18] Choubey, S. K., Deogun, J. S., Raghavan, V. V. and Sever, H. (1996), "A comparison of feature selection algorithms in the context of rough classifiers", in: Proceedings of the 5th IEEE International Conference on Fuzzy Systems (vol. 2), New Orleans, LA, pp. 1122-1128.

[19] Clopper, C. and Pearson, S. (1934), "The use of confidence or fiducial limits illustrated in the case of the binomial", Biometrika 26, pp. 404-413.

[20] Dash, M. and Liu, H. (1998), "Hybrid search of feature subsets", in: Lee, H.Y. and Motoda, H. (eds.), Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence, Springer Verlag, Singapore, pp. 22-27.

[21] Devijver, P.A. and Kittler, J. (1982), Pattern Recognition: A Statistical Approach, Prentice Hall, London.

[22] Diakoulaki, D.; Zopounidis, C.; Mavrotas, G.; and Doumpos, M. (1999) "The Use of a Preference Disaggregation Method in Energy Analysis and Policy Making," Energy 24(2), pp. 157-166.

[23] Doak, J. (1992), "An evaluation of feature selection methods and their application to computer security", Technical report CSE-92-18, University of California, Department of Computer Science, Davis, CA.

[24] Duda, R.O., Hart, P.E. and Stork, D.G. (2001), Pattern Classification (2nd Edition), John Wiley, New York.

[25] Duin, R. (2002), "The Combining Classifier: To Train or Not to Train?", ICR16, Proceedings 16th International Conference on Pattern Recognition, IEEE Computer Society Press, Los Alamitos, Volume II, pp. 765-770.

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

[27] Efron, B. (1983), "Estimating the error rate of a prediction rule: Improvement of cross-validation", Journal of the American Statistical Association, 78, pp. 316-330.

[28] Eisenbeis, R. (1977), "The pitfalls in the application of discriminant analysis in business, finance and economics", The Journal of Finance 32, pp. 723-739.

[29] Elazmeh W, Japkowicz N. and Matwin S. (2006), "Confidence Interval for the difference in classification error", American Association for Artificial Intelligence. Available at https://www.aaai.org/Papers/Workshops/2006/WS-06-06/WS06-06-008.pdf

[30] Fawcett, T. (2003), "ROC Graphs: Notes and Practical Considerations for Researchers", Tech Report HPL-2003-4, HP Laboratories. Available at http://www.hpl.hp.cpm/personal/Tom_Fawcett/papers/ROC101.pdf.

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

[32] Ghosh, B.K. (1979), "A comparison of some approximate confidence intervals for the binomial parameter", Journal of the American Statistical Association, 74, pp. 894-900.

[33] Giacinto G. and Roli F. (2001), "Design of effective neural network ensembles for image classification processes", Image Vision and Computing Journal, 19(9/10), pp. 699-707.

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

[35] Goldberg E. David (1989). Genetic algorithms in search, optimization and machine learning. Addison-Wesleym Professional.

[36] Hall, M.A. (1999), Correlation-Based Feature Selection for Machine Learning, PhD thesis, University of Waikato. Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.4643

[37] Hand, D.J. (1997) "Construction and Assessment of Classification Rules," John Wiley & Sons Ltd., Baffins Lane, Chichester.

[38] Ho T.K. (1998), "The random space method for constructing decision forests", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), pp. 832-844.

[39] John, G.H., Kohavi, R. and Pfleger, K. (1994), "Irrelevant features and the subset selection problem", in: Cohen, W. and Hirsh, H. (eds.), Machine Learning: Proceedings of the 11th International Conference, Morgan Kaufmann, San Francisco, pp. 121-129.

[40] Kira, K. and Rendell, L. (1992), "The feature selection problem: Traditional methods and a new algorithm", in: Proceedings of AAAI-92, AAAI Press, pp. 129-134.

[41] Kirkby, R., Frank, E. and Reutemann, P. (2008) "WEKA Explorer User Guide for Version 3-5-1", 2002-2008 University of Waikato. Available at http://www.docstoc.com/docs/2138851/WEKA-Explorer-User-Guide-for-Version-3-5-1

[42] Klein, M. and Methlie, L. B. (1995), "Expert systems: A decision support approach with applications in management and finance", Addison-Wesley, Wokingham.

[43] Kononenko, I. (1994), "Estimating attributes: Analysis and extensions of Relief", in: De Raedt, L. and Bergadano, F. (eds.), Proceedings of the European Conference on Machine Learning, Springer-Verlag, New York, pp. 171-182.

[44] Kordatoff, Y. and Michalski, R.S. (1990), "Machine Learning: An Artificial Intelligence Approach", Volume III, Morgan Kaufmann Publishers, Los Altos, California.

[45] Kuncheva L.I, Whitaker C.J., Shipp C.A. and Duin R.P.W (2003), "Limits on the majority vote accuracy in classifier Fusion", pp. 22-31.

[46] Kuncheva L.I, Whitaker C.J. and Shipp C.A. (2000), "Is independence good for combining classifiers?", Pattern Recognition. Available at http://ict.ewi.tudelft.nl/~duin/papers/icpr_00_cclass.pdf

[47] Kuncheva L.I and Whitaker C.J. (2003), "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy", Machine Learning, volume 51, pp. 181-207.

[48] Landwehr Niels, Hall Mark and Frank Eibe (2005), "Logistic Model Trees", Volume 59, Issue 1-2, pp. 161-205.

[49] Liu, H. and Motoda, H. (1998), "Feature Selection for Knowledge Discovery and Data Mining", Kluwer Academic Publishers, London, GB.

[50] Liu, H. and Setiono, R. (1996a), "A probabilistic approach to feature selection: A filter solution", in: Saitta, L. (ed.), Proceedings of the 13th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, pp. 319-327.

[51] Liu, H. and Setiono, R. (1996b), "Feature selection and classification: A probabilistic wrapper approach", in: Proceedings of the 9th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Morgan Kaufmann, pp. 129-135.

[52] Liu, H. and Setiono, R. (1998a), "Incremental feature selection", Applied Intelligence, 9(3), pp. 217-230.

[53] Liu, H. and Setiono, R. (1998b), "Scalable feature selection for large sized databases", in: Proceedings of the 4th World Congress on Expert Systems, Morgan Kaufmann, pp. 68-75.

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

[55] Mitchell M. Tom (1997), "Machine Learning", McGraw Hill.

[56] Molina, L.C., Belanche, L. and Nebot, A. (2002), "Feature selection algorithms: A survey and experimental evaluation", in: Proceedings of the 2002 IEEE International Conference on Data Proceedings of the 2002 IEEE International Conference on Data Mining, IEEE Computer Society, pp. 306-313.

[57] Narendra, P. and Fukunaga, K. (1977), "A branch and bound algorithm for feature subset selection", IEEE Transactions on Computers, 26(9), pp. 917-922.

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

[59] Pawlak, Z. (1982), "Rough sets", International Journal of Information and Computer Sciences, 11, pp. 341–356.

[60] Pudil, P., Novovicova, J. and Kittler, J. (1994), "Floating search methods in feature selection", Pattern Recognition Letters, 15(11), pp. 1119-1125.

[61] Quinlan, J.R. (1983), "Learning efficient classification procedures and their application to chess and games", in: R.S. Michalski, J.G. Carbonell and T.M. Mitchell (eds.), Machine Learning: "An Artificial Intelligence Approach", Tioga Publishing Company, Palo Alto, CA.

[62] Quinlan, J.R. (1986), "Induction of decision trees", Machine Learning 1, p.p. 81-106.

[63] Quinlan, J.R. (1993), "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Los Altos, California.

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

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

[66] Ruta, D. and Gabryst, B. (2005), "Classifier selection for Majority Voting", Information Fusion, Elsevier. Available at http://cis.paisley.ac.uk/ruta-ci0/downloads/paper11.pdf.

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

[68] Shipp, C.A. and Kuncheva L.I (2002), "Relationships between combination methods and measures of diversity in combining classifiers", Information Fusion, Elsevier, pp. 135-148.

[69] Siskos, Y.; Grigoroudis, E.; Zopounidis, C.; and Saurais, O. (1998) "Measuring Customer Satisfaction using a Survey Based Preference Disaggregation Model," Journal of Global Optimization 12(2), pp. 175-195.

[70] Skalak D.B. (1996), "The sources of increased accuracy for two proposed boosting algorithms", In Proceedings of American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop, pp. 120-125.

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

[72] Stone, M. (1974), "Cross-validation choice and assessment of statistical predictions", Journal of the Royal Statistical Society B, 36, pp. 111-147.

[73] Tafti, M. H. A. and Nikbakht, E. (1993), "Neural networks and experts systems: New horizons in business finance applications", Information Management & Computer Security 1, pp. 22-28.

[74] Theodoridis S. and Koutroumbas K. (1999), Pattern Recognition, Academic Press.

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

[76] Vapnik, V.N. (1998), Statistical Learning Theory, John Wiley, New York.

[77] Vapnik, V.N. (2000) "The Nature of Statistical Learning Theory," Springer, New York, Second Edition.

[78] Wald, A. and Wolfowitz, J. (1939), "Confidence limits for continuous distribution functions", The Annals of Mathematical Statistics, 10, pp. 105-118.

[79] Wilson, E.B. (1927), "Probable inference, the law of succession and statistical inference", Journal of the American Statistical Association 22, pp. 209-212.

[80] Yohannes, Y. and Webb, P. (1999), "Classification and Regression Trees, CART", A user manual for identifying indicators of vulnerability to famine and chronic food insecurity, Microcomputers in Policy Research, International Food Policy Research Institute. Available at http://www.ifpri.org/pubs/microcom/micro3.pdf

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

[82] Zopounidis, C. and Doumpos, M. (1998), "Developing a multicriteria decision support system for financial support classification problems: The FINCLAS system", Optimization Methods and Software 8, pp. 277-304.

[83] Αποστολάκης Ι., Καστανιά Α. και Πιερράκου Χρ. (2003), «Στατιστική Επεξεργασία Δεδομένων στην Υγεία», Εκδόσεις Παπαζήση.

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

[85] Καλαπανίδας, Η. (2003), «Περιβαλλοντική πρόβλεψη με μεθόδους Μηχανικής Μάθησης», Διδακτορική Διατριβή, Τμήμα Ηλεκτρολόγων Μηχανικών & Τεχνολογίας Υπολογιστών, Πανεπιστήμιο Πατρών.

[86] Κολυβά-Μαχαιρά Φ. και Μπόρα-Σέντα Ε. (1999), «Στατιστική: Θεωρία-Εφαρμογές», Εκδόσεις Ζήση.

[87] Παπούλιας Αθανάσιος (1994), «Πιθανότητες, Τυχαίες Μεταβλητές και Στοχαστικές Διαδικασίες», Εκδόσεις Τζιόλα.