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

Αλγόριθμοι Επιλογής Χαρακτηριστικών σε Προβλήματα Ταξινόμησης: Μία Πειραματική Ανάλυση

Διατριβή που υπεβλήθη για τη μερική ικανοποίηση
των απαιτήσεων για την απόκτηση μεταπτυχιακού
διπλώματος από τη

Σαλάππα Αθηνά

Χανιά 2005

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

Ζοπουνίδης Κωνσταντίνος	Καθηγητής
Δούμπος Μιχάλης	Λέκτορας
Γρηγορούδης Ευάγγελος	Λέκτορας

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 επιβέποντα Καθηγητή κ. Κωνσταντίνο Ζοπουνίδη και τον Λέκτορα κ. Μιχάλη Δούμπο τόσο για την σημαντική βοήθεια που μου παρείχε για την διεκπεραίωση της εργασίας, όσο και για τις γνώσεις που μου προσέφερε κατά τη διάρκεια του μεταπτυχιακού μου.

Επίσης, ευχαριστώ την οικογένειά μου για την υλική και κυρίως ηθική υποστήριξή της όλα αυτά τα χρόνια που λείπω από κοντά της και στην οποία αφιερώνεται η παρούσα ερευνητική εργασία.

Τέλος, θα ήθελα να ευχαριστήσω τους κοντινούς μου φίλους, οι οποίοι μου συμπαραστάθηκαν κατά τη διάρκεια της διατριβής μου.

Σαλάππα Αθηνά
Χανιά
Μάρτιος 2005

Κεφάλαιο 1ο

ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της έρευνας

Η διαδικασία επιλογής χαρακτηριστικών (Feature Selection – FS) παίζει σημαντικό ρόλο στην ανάπτυξη αξιόπιστων υποδειγμάτων ταξινόμησης. Οι διαδικασίες FS αποσκοπούν στην επιλογή του πλέον κατάλληλου συνόλου χαρακτηριστικών, τα οποία θα περιγράφουν με ικανοποιητικό βαθμό την ταξινόμηση ενός αντικειμένου. Βασικός σκοπός της συγκεκριμένης ερευνητικής διατριβής είναι η διεξαγωγή μιας πλήρης ανάλυσης της επίδοσης και αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών (Feature Selection Algorithms – FSAs) στην αντιμετώπιση προβλημάτων ταξινόμησης. Η αξιολόγηση της αποτελεσματικότητας των FSAs καλύπτει ένα ευρύ πεδίο θεμάτων, όπως α) την ικανότητα των FSAs να εντοπίζουν τα σχετικά χαρακτηριστικά, β) την αποτελεσματικότητα των υποδειγμάτων ταξινόμησης που αναπτύσσονται μετά την εφαρμογή των FSAs, γ) το ποσοστό μείωσης του συνόλου των χαρακτηριστικών, δ) τις αλληλεπιδράσεις μεταξύ των FSAs και των διαφόρων τεχνικών ταξινόμησης. Η όλη διαδικασία ξεκινάει με ένα αρχικό δείγμα πληροφοριών, το οποίο βαθμιαία μειώνεται, καθώς απαλείφονται όλες εκείνες οι πληροφορίες (χαρακτηριστικά) που εμπεριέχουν θόρυβο ή είναι ακατάλληλες, ελέγχοντας ταυτόχρονα το ποσοστό μείωσης της ποιότητας της ταξινόμησης. σε συνδυασμό με διαδεδομένες τεχνικές ταξινόμησης, όπως είναι η γραμμική διακριτική ανάλυση, η λογιστική παλινδρόμηση, τα πιθανοτικά νευρωνικά δίκτυα, ο αλγόριθμος του πλησιέστερου γείτονα, τα δέντρα ταξινόμησης και οι μηχανές διανύσματος υποστήριξης. 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.

1.1.1 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 $X = \{x_1, ..., x_m\}$ 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.

1.1.2 Προτεινόμενοι αλγόριθμοι επιλογής χαρακτηριστικών

Στα πλαίσια της συγκεκριμένης ερευνητικής διατριβής εξετάζονται αλγόριθμοι επιλογής χαρακτηριστικών που ανήκουν σε διαδικασίες filter και wrapper, σύμφωνα με τον τρόπο λειτουργίας τους.

Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στην πρώτη κατηγορία 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 σύνολο χαρακτηριστικών που επιλέγεται και της τεχνικής ταξινόμησης που χρησιμοποιείται για την ανάπτυξη του υποδείγματος ταξινόμησης. Οι εξεταζόμενοι αλγόριθμοι αυτής της κατηγορίας είναι ο FOCUS (Almuallim and Dietterich, 1991, 1994), ο RELIEF (Kira and Rendell, 1992) και οι αλγόριθμοι Las Vegas (Liu and Setiono, 1996a, 1998b), sequential forward generation (Pudil et al., 1994) και sequential backward generation (Choubey et al., 1996).

Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στη δεύτερη κατηγορία 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). Οι εξεταζόμενοι αλγόριθμοι αυτής της κατηγορίας είναι οι αλγόριθμοι Las Vegas (Liu and Setiono, 1996a, 1998b), sequential forward generation (Pudil et al., 1994) και sequential backward generation (Choubey et al., 1996).

1.2 Προτεινόμενη μεθοδολογική προσέγγιση και στόχοι της έρευνας

Για την αντιμετώπιση του προβλήματος επιλογής χαρακτηριστικών, στη διεθνή βιβλιογραφία έχουν προταθεί διάφοροι αλγόριθμοι και μεθοδολογίες. Παρά όμως τη σημαντική έρευνα που έχει πραγματοποιηθεί στο χώρο αυτό δεν υπάρχει μια ολοκληρωμένη έρευνα σχετικά με την αποτελεσματικότητα των προτεινόμενων μεθοδολογιών και αλγορίθμων. Μια τέτοια έρευνα πρέπει να λάβει υπόψη τις «αλληλεπιδράσεις» μεταξύ των διαδικασιών επιλογής χαρακτηριστικών και των τεχνικών που χρησιμοποιούνται για την ανάπτυξη υποδειγμάτων ταξινόμησης. Οι μέχρι σήμερα έρευνες περιορίζονται σε συγκεκριμένες τεχνικές ταξινόμησης (συνήθως δέντρα ταξινόμησης και τεχνικές εξαγωγής κανόνων απόφασης). Δεδομένου όμως του πλήθους των διαφορετικών τεχνικών ταξινόμησης

που είναι σήμερα διαθέσιμες (νευρωνικά δίκτυα και συναφείς τεχνικές, μηχανική μάθηση, μαθηματικός προγραμματισμός, κ.ά.), είναι εμφανές ότι απαιτείται μια πιο ολοκληρωμένη ανάλυση.

Βάσει της διαπίστωσης αυτής, στην παρούσα εργασία πραγματοποιείται μια ολοκληρωμένη έρευνα της αποτελεσματικότητας και των δυνατοτήτων που παρέχουν οι αλγόριθμοι επιλογής χαρακτηριστικών, καλύπτοντας σημαντικά θέματα όπως: (α) τη δυνατότητα των αλγορίθμων να εντοπίσουν την πραγματικά χρήσιμη πληροφορία (χαρακτηριστικά), (β) την αποτελεσματικότητα των υποδειγμάτων ταξινόμησης τα οποία αναπτύσσονται με τη χρήση αλγορίθμων επιλογής χαρακτηριστικών, (γ) το βαθμό μείωσης της εξεταζόμενης πληροφορίας και τη σχέση του με την αποτελεσματικότητα των υποδειγμάτων ταξινόμησης, και (δ) τις αλληλεπιδράσεις στα παραπάνω θέματα μεταξύ των χρησιμοποιούμενων τεχνικών για την ανάπτυξη υποδειγμάτων ταξινόμησης και των αλγορίθμων επιλογής χαρακτηριστικών.

Για τη διερεύνηση των παραπάνω θεμάτων θα ληφθούν υπόψη διάφοροι αλγόριθμοι επιλογής χαρακτηριστικών αντιπροσωπευτικοί των υπαρχόντων προσεγγίσεων του προβλήματος. Ταυτόχρονα θα εξεταστούν διάφορες τεχνικές ανάπτυξης υποδειγμάτων ταξινόμησης από το χώρο της μηχανικής μάθησης, της επιχειρησιακής έρευνας και της στατιστικής. Για τον σκοπό αυτό απαιτείται η χρήση πειραματικών δεδομένων.

1.3 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 του προβλήματος της ταξινόμησης και των βασικών σημείων έρευνας στο χώρο αυτό. Επίσης, πραγματοποιείται μια γενική περιγραφή των εξεταζόμενων τεχνικών ταξινόμησης, της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης και της σπουδαιότητας του προβλήματος τόσο σε θεωρητικό, όσο και πρακτικό επίπεδο. Τέλος, πραγματοποιείται μια συνοπτική αναφορά στη διαδικασία ελέγχου cross-validation, ως μιας τεχνικής επαναληπτικής δειγματοληψίας για τον έλεγχο της αποτελεσματικότητας των αναπτυσσόμενων υποδειγμάτων ταξινόμησης.

Στο τρίτο κεφάλαιο παρουσιάζεται μια συνοπτική παρουσίαση των αλγορίθμων επιλογής χαρακτηριστικών. Οι εξεταζόμενοι αλγόριθμοι παρουσιάζονται σύμφωνα με τη στρατηγική διερεύνησης των χαρακτηριστικών που ακολουθούν, τη διαδικασία επιλογής των χαρακτηριστικών και τα κριτήρια αξιολόγησης της ποιότητας των επιλεγμένων χαρακτηριστικών. Τέλος, παρουσιάζονται, από τη διεθνή βιβλιογραφία, κάποιες συγκριτικές έρευνες για την αξιολόγηση των υπαρχόντων προσεγγίσεων και δίνονται συνοπτικά κάποια αποτελέσματα και συμπεράσματα που προέκυψαν από τη διεξαγωγή των σχετικών ερευνών πάνω στους αλγόριθμους επιλογής χαρακτηριστικών.

Στο τέταρτο κεφάλαιο πραγματοποιείται μια γενική περιγραφή των εξεταζόμενων αλγορίθμων επιλογής χαρακτηριστικών και εξετάζεται η αποτελεσματικότητά τους σε συνδυασμό με ευρέως διαδεδομένες τεχνικές ταξινόμησης. Η πειραματική ανάλυση εστιάζεται, όχι μόνο σε τεχνητά, αλλά και σε πραγματικά δεδομένα. Σκοπός της πειραματικής ανάλυσης είναι η αξιολόγηση της αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών στην ανάπτυξη κατάλληλων υποδειγμάτων ταξινόμησης.

Τέλος, στο πέμπτο 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 τους αλγορίθμους επιλογής χαρακτηριστικών.

Κεφάλαιο 2ο

ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΤΑΞΙΝΟΜΗΣΗΣ

2.1 Εισαγωγή στο πρόβλημα της ταξινόμησης

Πολλά προβλήματα λήψης αποφάσεων σε διάφορα επιστημονικά και πρακτικά πεδία μοντελοποιούνται ως προβλήματα ταξινόμησης. Γενικά, το πρόβλημα της ταξινόμησης αναφέρεται στην ταξινόμηση ενός συνόλου αντικειμένων σε προκαθορισμένες ομοιογενείς κατηγορίες. Ουσιαστικά, τα προβλήματα αυτής της μορφής αφορούν τον εντοπισμό ενός υποδείγματος ταξινόμησης $f(\mathbf{x}) \rightarrow C$ το οποίο συνδυάζει τα χαρακτηριστικά (διάνυσμα $\mathbf{x}$) που περιγράφουν τα αντικείμενα και αποδίδει την ταξινόμηση των αντικειμένων σε κάποια από τις κατηγορίες του συνόλου C (Zopounidis and Doumpos, 2002).

Ο προσδιορισμός του υποδείγματος ταξινόμησης f βασίζεται στην ανάλυση ενός δείγματος εκμάθησης το οποίο αποτελείται από n ζεύγη της μορφής $(\mathbf{x}_i, c_i)$, $i = 1, ..., n$, όπου $\mathbf{x}_i \in \mathbb{R}^m$ είναι η περιγραφή του αντικειμένου i σε ένα σύνολο m χαρακτηριστικών $x_1, x_2, ..., x_m$ και $c_i \in C$ είναι η ταξινόμηση του αντικειμένου i σε κάποια από τις προκαθορισμένες κατηγορίες. Δεδομένου ενός τέτοιου δείγματος εκμάθησης ο προσδιορισμός του βέλτιστου υποδείγματος f γίνεται με στόχο τη μεγιστοποίηση της αναμενόμενης ακρίβειας ταξινόμησης. Η αναμενόμενη ακρίβεια ταξινόμησης αναφέρεται στις διαφοροποιήσεις που παρατηρούνται μεταξύ της εκτιμώμενης ταξινόμησης που προσδιορίζεται από το υπόδειγμα f και της πραγματικής ταξινόμησης ενός οποιουδήποτε συνόλου αντικειμένων.

Μέσα σε αυτό το μεθοδολογικό πλαίσιο η ανάπτυξη ενός κατάλληλου υποδείγματος ταξινόμησης μπορεί να επιτευχθεί μέσα από διάφορες γνωστές τεχνικές ταξινόμησης, οι οποίες μεταξύ των άλλων περιλαμβάνουν στατιστικές και οικονομετρικές τεχνικές (διακριτική ανάλυση, λογιστική παλινδρόμηση) και μη παραμετρικές τεχνικές (νευρωνικά δίκτυα, δέντρα ταξινόμησης, μηχανές διανύσματος υποστήριξης, προσεγγιστικά σύνολα, πολυκριτήριες μέθοδοι, κ.ά.).

Τα βασικά σημεία έρευνας στο χώρο των προβλημάτων ταξινόμησης αφορούν θέματα όπως: (α) τη μορφή του αναπτυσσόμενου υποδείγματος ταξινόμησης, (β) τα κριτήρια και τις διαδικασίες βελτιστοποίησης του υποδείγματος, (γ) τις διαδικασίες ελέγχου της αναμενόμενης αποτελεσματικότητας των υποδειγμάτων που αναπτύσσονται, και (δ) το συνδυασμό διαφορετικών υποδειγμάτων.

2.1.1 Η διαδικασία ανάπτυξης υποδειγμάτων ταξινόμησης

Γενικότερα, 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.

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 Y με ένα διάνυσμα μεταβλητών X βάσει της ανάλυσης ενός συνόλου δεδομένων παρατηρήσεων $(X,\ Y)$. 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 (X, 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 $f(X) \to C$ 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 2.1: Γενικό περίγραμμα της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης

Σύμφωνα με τους Fayyad (1996) και Simoudis (1996), η διαδικασία ανάπτυξης υποδειγμάτων ταξινόμησης περιλαμβάνει πέντε κύρια στάδια: τη διαμόρφωση ενός δείγματος εκμάθησης, την προεπεξεργασία του δείγματος, το μετασχηματισμό των δεδομένων και τη μείωσή τους, τη βελτιστοποίηση ενός υποδείγματος ταξινόμησης, και τέλος, την αξιολόγηση του υποδείγματος. Το πρώτο στάδιο αφορά την συλλογή των απαραίτητων δεδομένων από τις διαθέσιμες πηγές πληροφοριών. Στη φάση της

προεπεξεργασίας αντιμετωπίζονται θέματα όπως η ύπαρξη ακραίων περιπτώσεων οι οποίες πιθανόν εισάγουν θόρυβο, η ύπαρξη ελλιπών δεδομένων, κλπ. Στην τρίτη φάση (μετασχηματισμός και μείωση των δεδομένων) ελέγχεται η αναγκαιότητα πραγματοποίησης μετασχηματισμών στα δεδομένα (μετατροπή ποιοτικών δεδομένων σε ποσοτικά, διακριτοποίηση ποσοτικών δεδομένων, ορισμός νέων χαρακτηριστικών, κ.ά.) και διερευνάται η μείωση των διαστάσεων των δεδομένων μέσω της επιλογής των κατάλληλων υποσυνόλων χαρακτηριστικών. Το στάδιο της βελτιστοποίησης ενός υποδείγματος αναφέρεται στην εφαρμογή κάποιας τεχνικής ταξινόμησης με στόχο την ανάπτυξη του κατάλληλου υποδείγματος ταξινόμησης βάσει των δεδομένων που διαμορφώθηκαν στο προηγούμενο στάδιο. Το τελευταίο στάδιο της διαδικασίας αναφέρεται στην εκτίμηση της αναμενόμενης αποτελεσματικότητας του υποδείγματος που αναπτύχθηκε στο προηγούμενο στάδιο.

Η επιλογή κατάλληλων χαρακτηριστικών, η οποία ενσωματώνεται στην τρίτη φάση της παραπάνω διαδικασίας, είναι ένα από τα πλέον κρίσιμα στάδια με ιδιαίτερη σημασία όσον αφορά τον υπολογιστικό φόρτο της βελτιστοποίησης του υποδείγματος, την αναμενόμενη αποτελεσματικότητά του, καθώς και το χρόνο/ κόστος χρήσης του υποδείγματος. Βέβαια, η επιλογή κατάλληλων χαρακτηριστικών δεν είναι ένα απλό πρόβλημα, καθώς η εξαντλητική διερεύνηση όλων των πιθανών υποσυνόλων χαρακτηριστικών απαιτεί αυξημένο υπολογιστικό χρόνο της τάξης $O(2^m T)$, όπου m είναι το πλήθος των χαρακτηριστικών και T είναι ο χρόνος που απαιτείται για τη βελτιστοποίηση του υποδείγματος ανάλογα με την τεχνική ταξινόμησης που επιλέγεται. Προφανώς, μια εξαντλητική διαδικασία επιλογής χαρακτηριστικών είναι πρακτικά εφικτή μόνο εάν το πλήθος των εξεταζόμενων χαρακτηριστικών είναι μικρό και ο υπολογιστικός φόρτος της χρησιμοποιούμενης τεχνικής ταξινόμησης είναι περιορισμένος. Για την αντιμετώπιση του θέματος αυτού στη διεθνή βιβλιογραφία έχουν προταθεί διάφοροι αλγόριθμοι και τεχνικές επιλογής χαρακτηριστικών (feature selection algorithms, FSAs).

2.1.2 Σπουδαιότητα του προβλήματος της 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 και 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 και 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 και 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 και Doumpos και Zopounidis, 1998).

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

2.2 Τεχνικές ταξινόμησης

Οι τεχνικές που χρησιμοποιήθηκαν στην παρούσα έρευνα είναι αντιπροσωπευτικές όλων 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.

2.2.1 Γραμμική διακριτική ανάλυση

Η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis – LDA) αποτέλεσε την πρώτη πολυδιάστατη μέθοδο ταξινόμησης και αναπτύχθηκε αρχικά από τον Fisher (1936). Σκοπός της μεθόδου είναι η ανάπτυξη μιας σειράς διακριτικών συναρτήσεων οι οποίες μεγιστοποιούν τη διακύμανση μεταξύ των κατηγοριών σε σχέση με την διακύμανση εντός των κατηγοριών, χρησιμοποιώντας ως δείγμα εκμάθησης ένα σύνολο εναλλακτικών δραστηριοτήτων η ταξινόμηση των οποίων είναι γνωστή. Στην περίπτωση των δύο κατηγοριών (C_1 και C_2), η οποία εξετάζεται στην παρούσα ανάλυση, η LDA οδηγεί στην ανάπτυξη μιας διακριτικής συνάρτησης της μορφής::

$$F(\mathbf{x}) = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m$$

όπου $\mathbf{x} = (x_1, x_2, ..., x_m)$ είναι το διάνυσμα των χαρακτηριστικών που περιγράφουν τις εναλλακτικές δραστηριότητες, a είναι μια σταθερά και $b_1, b_2, ..., b_m$ είναι οι συντελεστές των χαρακτηριστικών στη συνάρτηση.

Ο υπολογισμός του σταθερού όρου a και του διανύσματος $\mathbf{b} = (b_1, b_2, ..., b_m)^T$ βασίζεται στην υπόθεση ότι οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών είναι ίσοι και ότι οι επιδόσεις των εναλλακτικών δραστηριοτήτων στα εξεταζόμενα χαρακτηριστικά ακολουθούν την πολυμεταβλητή κανονική κατανομή. Βάσει των υποθέσεων αυτών οι υπολογισμοί των a και $\mathbf{b}$ πραγματοποιούνται ως εξής:

$$\mathbf{b} = \mathbf{\Sigma}^{-1} \cdot \left[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right] \quad \text{και} \quad a = -\left[\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right]' \cdot \mathbf{b} / 2$$

όπου $\boldsymbol{\mu}_1$ και $\boldsymbol{\mu}_2$ είναι τα διανύσματα των μέσων τιμών των χαρακτηριστικών για τις εναλλακτικές δραστηριότητες των κατηγοριών C_1 και C_2, αντίστοιχα, και Σ είναι ο πίνακας διακύμανσης-συνδιακύμανσης μεταξύ των κατηγοριών.

Η ταξινόμηση κάθε αντικειμένου i σε μια εκ των προκαθορισμένων κατηγοριών πραγματοποιείται βάσει του σκορ διάκρισης $F(\mathbf{x}_i)$ του αντικειμένου όπως αυτό υπολογίζεται από τη διακριτική συνάρτηση. Συγκεκριμένα, ο κανόνας ταξινόμησης έχει την ακόλουθη μορφή:

$$F(\mathbf{x}_i) \geq \ln \frac{K(1/2)\pi_1}{K(2/1)\pi_2} \Rightarrow i \in C_1$$

$$F(\mathbf{x}_i) < \ln \frac{K(1/2)\pi_1}{K(2/1)\pi_2} \Rightarrow i \in C_2$$

Ως $K(1/2)$ συμβολίζεται το κόστος της εσφαλμένης ταξινόμησης ενός αντικειμένου, της κατηγορίας C_1 στην κατηγορία C_2, ενώ ως π_1 συμβολίζεται η εκ των προτέρων πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα στην κατηγορία C_1. Θεωρώντας ίσα τα κόστη εσφαλμένων ταξινομήσεων και τις εκ των προτέρων πιθανότητες, ο γραμμικός κανόνας ταξινόμησης για τη διάκριση μεταξύ δύο κατηγοριών μπορεί να αποδοθεί γραφικά μέσω του παρακάτω σχήματος.

Σχήμα 2.2: Σχηματική απεικόνιση του κανόνα ταξινόμησης της γραμμικής διακριτικής ανάλυσης

Δεδομένου ότι ο καθορισμός των εκ των προτέρων πιθανοτήτων και του κόστους των εσφαλμένων ταξινομήσεων είναι συχνά δύσκολος, το όριο που διαχωρίζει τις κατηγορίες

καθορίζεται συνήθως μέσω διαδικασιών δοκιμής και λάθους, ώστε να ελαχιστοποιηθεί ο συνολικός αριθμός των εσφαλμένων ταξινομήσεων και παράλληλα να υπάρχει μια ισορροπία στον αριθμό των εσφαλμένων ταξινομήσεων ανά κατηγορία.

2.2.2 Λογιστική παλινδρόμηση

Η λογιστική ανάλυση (Logit Regression – LR) αποτελεί μια εναλλακτική μέθοδο ταξινόμησης, η οποία πλεονεκτεί της διακριτικής ανάλυσης, τόσο σε θεωρητικό επίπεδο, όσο και στην αποτελεσματικότητα των αναπτυσσόμενων υποδειγμάτων. Προέρχεται από τον χώρο της οικονομετρίας και αν και αποτελεί μια από τις πιο παλιές μεθοδολογίες γνώρισε ιδιαίτερη διάδοση (μαζί με το γραμμικό υπόδειγμα πιθανότητας – linear probability model και το κανονικό υπόδειγμα πιθανότητας – probit probability model) κατά τη δεκαετία του 1970 στην ανάπτυξη της θεωρίας της διακριτής επιλογής (discrete choice).

Η λογιστική παλινδρόμηση είναι ένα πολυμεταβλητό υπό συνθήκη πιθανοτικό μοντέλο και βασίζεται σε μία αθροιστική συνάρτηση πιθανότητας, η τιμή της οποίας δίνει την πιθανότητα ένα αντικείμενο να ανήκει σε μια εκ των προκαθορισμένων κατηγοριών.

Στο λογιστικό υπόδειγμα πιθανότητας, η πιθανότητα ένα αντικείμενο $\mathbf{x}_i$ να ανήκει στην κατηγορία C_1 είναι: $P_i = F\left(a + b\mathbf{x}_i\right)$, όπου $F\left(a + b\mathbf{x}_i\right)$ είναι η αθροιστική λογιστική συνάρτηση: $F\left(a + b\mathbf{x}_i\right) = \dfrac{1}{1 + e^{-(a+b\mathbf{x}_i)}}$.

Ο υπολογισμός του σταθερού όρου a και του διανύσματος b, το οποίο περιέχει τους συντελεστές των χαρακτηριστικών, πραγματοποιείται χρησιμοποιώντας τεχνικές μέγιστης πιθανοφάνειας, και πιο συγκεκριμένα μεγιστοποιώντας την ακόλουθη συνάρτηση:

$$\ln L = \sum_{\forall \mathbf{x}_i \in C_2} \ln\left(P_i\right) + \sum_{\forall \mathbf{x}_i \in C_1} \ln\left(1 - P_i\right)$$

Από τη μορφή της συνάρτησης αυτής εύλογα εξάγεται το συμπέρασμα ότι η εκτίμηση των παραμέτρων του λογιστικού υποδείγματος ανάγεται σε ένα πρόβλημα μη γραμμικής παλινδρόμησης, η επίλυση του οποίου πολλές φορές καθίσταται ιδιαίτερα δύσκολη, ιδίως στην περίπτωση του κανονικού υποδείγματος. Μάλιστα σε περιπτώσεις όπου είναι δυνατή η ανάπτυξη ενός γραμμικού συνδυασμού των χαρακτηριστικών, που να διαχωρίζει απόλυτα τις δύο κατηγορίες, τότε η διαδικασία βελτιστοποίησης δεν θα συγκλίνει με αποτέλεσμα να μην είναι δυνατός ο υπολογισμός των παραμέτρων του λογιστικού υποδείγματος (Altman et al., 1981).

Το λογιστικό υπόδειγμα πιθανότητας παρέχει την πιθανότητα ένα αντικείμενο να ανήκει σε μια εκ των δύο κατηγοριών. Με βάση τη σύγκριση αυτής της πιθανότητας με μια πιθανότητα-όριο, κατατάσσεται ένα αντικείμενο, στοχεύοντας ταυτόχρονα στην ελαχιστοποίηση των σφαλμάτων τύπου I (ένα αντικείμενο της κατηγορίας C_1 να ταξινομηθεί στην κατηγορία C_2) και τύπου II (ένα αντικείμενο της κατηγορίας C_2 να ταξινομηθεί στην κατηγορία C_1). Οι συντελεστές του μοντέλου υπολογίζονται μεγιστοποιώντας τη λογαριθμική συνάρτηση πιθανότητας.

Η λογιστική παλινδρόμηση έχει γνωρίσει ιδιαίτερη διάδοση για την αντιμετώπιση προβλημάτων ταξινόμησης σε διάφορα επιστημονικά πεδία, αντικαθιστώντας σταδιακά τη διακριτική ανάλυση. Κάποιες σχετικές μελέτες που έγιναν δεν απέδειξαν κάποια μεγαλύτερη ακρίβεια στα αποτελέσματα ταξινόμησης, σε σχέση με τα αποτελέσματα τα οποία επιτυγχάνονται μέσω της διακριτικής ανάλυσης.

2.2.3 Ο αλγόριθμος του πλησιέστερου γείτονα

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 K 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).

2.2.4 Πιθανοτικά νευρωνικά δίκτυα

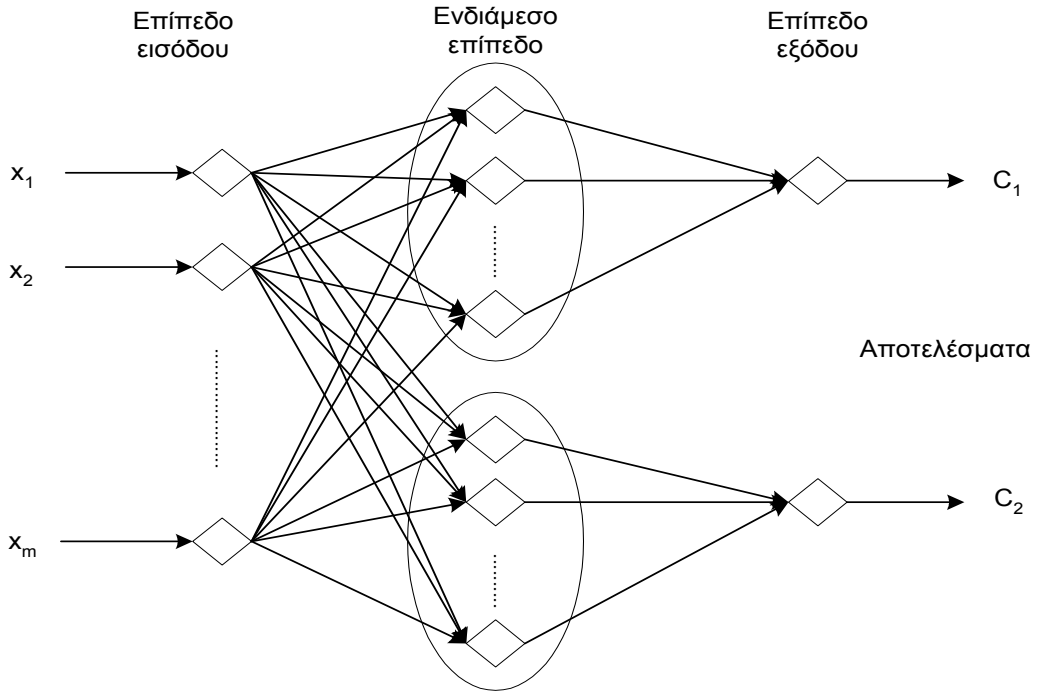
Τα πιθανοτικά νευρωνικά δίκτυα (Probabilistic Neural Networks – PNN) είναι μία κατηγορία νευρωνικών δικτύων, τα οποία συνδυάζουν τις ιδιότητες της στατιστικής αναγνώρισης προτύπων (Pattern Recognition), και των συμβατικών νευρωνικών δικτύων. Τα πιθανοτικά νευρωνικά δίκτυα εισήχθησαν από τον Donald Specht στο τέλος της δεκαετίας του 1980 και αναπτύχθηκαν ως τεχνικές αξιολόγησης για προβλήματα ταξινόμησης (Parzen window method, Duda et al., 2001). Έχουν παρόμοια οργανωτική δομή με τα νευρωνικά δίκτυα και η μεθοδολογία ταξινόμησης συνδυάζει την υπολογιστική δομή και την ευελιξία των τεχνητών νευρωνικών δικτύων.

Κάθε PNN είναι ένα δίκτυο παράλληλων μονάδων επεξεργασίας οι οποίες είναι οργανωμένες σε μια σειρά επιπέδων (layers). Το σχήμα 2.3 δείχνει την τυπική αρχιτεκτονική ενός PNN για προβλήματα ταξινόμησης σε δύο κατηγορίες, αλλά μπορεί να αναχθεί και σε πολλαπλές κατηγορίες ταξινόμησης ανάλογα με τις απαιτήσεις του εκάστοτε προβλήματος. Η αρχιτεκτονική δομή ενός PNN αποτελείται από τα εξής επίπεδα όπως φαίνεται και στο σχήμα (4.1).

1. Ένα επίπεδο εισόδου (input layer) αποτελούμενο από μια σειρά κόμβων, έναν για κάθε είσοδο.

2. Ένα επίπεδο εξόδου (output layer) το οποίο αποτελείται από τόσους κόμβους, όσο και το πλήθος των κατηγοριών.

3. Μια σειρά ενδιάμεσων επιπέδων (hidden layers) χωρισμένα σε ομάδες. Κάθε ομάδα αντιστοιχεί και σε μια κατηγορία.



Σχήμα 2.3: Σχηματική απεικόνιση της δομής ενός πιθανοτικού νευρωνικού δικτύου

Όπως φαίνεται στο σχήμα 2.3 το επίπεδο εισόδου αποτελείται από m κόμβους, έναν για κάθε μία είσοδο (χαρακτηριστικό). Οι ενδιάμεσοι κόμβοι (pattern nodes) αντιστοιχούν στα αντικείμενα του δείγματος εκπαίδευσης (υπάρχουν n ενδιάμεσοι κόμβοι) και είναι χωρισμένοι σε ομάδες, μία ομάδα για κάθε κατηγορία. Στην περίπτωση των δύο κατηγοριών, οι κόμβοι του ενδιάμεσου επιπέδου χωρίζονται σε δύο ομάδες: η πρώτη αφορά τα αντικείμενα του δείγματος εκπαίδευσης που ανήκουν στην κατηγορία C_1, ενώ η δεύτερη αφορά τα αντικείμενα του δείγματος εκπαίδευσης που ανήκουν στην κατηγορία C_2. Θεωρώντας ότι το δίκτυο χρησιμοποιείται για την ταξινόμηση ενός άγνωστου αντικειμένου $\mathbf{x}$, τα χαρακτηριστικά του αντικειμένου αυτού χρησιμοποιούνται ως είσοδοι στο δίκτυο. Σε κάθε κόμβο i του ενδιάμεσου επιπέδου προσδιορίζεται ο βαθμός ομοιότητας του αντικειμένου $\mathbf{x}$ προς το αντικείμενο $\mathbf{x}_i$ του δείγματος εκπαίδευσης. Ο βαθμός ομοιότητας προσδιορίζεται από τη συνάρτηση ενεργοποίησης (activation function) του κόμβου i, η οποία καθορίζει την έξοδο o_i από τον κόμβο αυτό. Η πλέον διαδεδομένη μορφή της συνάρτησης ενεργοποίησης είναι η εκθετική συνάρτηση:

$$o_i = \exp\left(-\frac{\left\| \mathbf{x}_i - \mathbf{x} \right\|^2}{2\sigma^2} \right)$$

Η τιμή του σ μπορεί να είναι η μέση τιμή της απόστασης των διανυσμάτων που ανήκουν στην ίδια ομάδα ή η μέση τιμή της απόστασης των διανυσμάτων της μιας ομάδας από τα κοντινότερα διανύσματα της άλλης ομάδας.

Οι έξοδοι από των ενδιάμεσων κόμβων που προσδιορίζονται κατά τον τρόπο αυτό συνδυάζονται στους κόμβους εξόδων. Σε κάθε κόμβο εξόδου k αθροίζονται οι έξοδοι των ενδιάμεσων κόμβων που αντιστοιχούν στην κατηγορία C_k, ώστε να προσδιορίστεί μια βαθμολογία $f_k(\mathbf{x})$ η οποία αναπαριστά το συνολικό βαθμό ομοιότητας του αντικειμένου $\mathbf{x}$ ως προς τα αντικείμενα του δείγματος εκπαίδευσης που ανήκουν στην κατηγορία C_k. Το αντικείμενο τελικά ταξινομείται στην κατηγορία σε σχέση με την οποία παρουσιάζει τον υψηλότερο βαθμό ομοιότητας.

2.2.5 Δέντρα ταξινόμησης και παλινδρόμησης

Η μέθοδος CART (Classification and Regression Trees – CART, Breiman et al., 1984) είναι μια μη παραμετρική προσέγγιση που αναπτύχθηκε για την ανάλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Σε κάθε περίπτωση το μοντέλο ταξινόμησης ή παλινδρόμησης που αναπτύσσεται μέσω της μεθόδου CART αναπαρίσταται με τη μορφή ενός δέντρου αποφάσεων. Στην περίπτωση της ταξινόμησης κύριος σκοπός της μεθόδου CART 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 CART 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, η μέθοδος CART χρησιμοποιεί μια πιθανοθεωρητική προσέγγιση η οποία μπορεί να υλοποιηθεί με τρεις τρόπους: (α) προσδιορισμός των *a priori* πιθανοτήτων των κατηγοριών από τα δεδομένα: $\pi_i = n_i / n$, όπου π_i η *a priori* της κατηγορίας C_i, n ο αριθμός των αντικειμένων στο δείγμα, και n_i ο αριθμός των αντικειμένων της κατηγορίας C_i , (β) θεώρηση των *a priori* πιθανοτήτων των κατηγοριών ως ίσων , και (γ) προσδιορισμός των *a priori* πιθανοτήτων των κατηγοριών μέσω μιας υβριδικής προσέγγισης θεωρώντας τον μέσο όρο των δύο εκτιμήσεων που υπολογίζονται από τις προηγούμενες δύο προσεγγίσεις.

Η ανάπτυξη ενός δέντρου ταξινόμησης 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.

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.

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 $x_{ij} \le d_j$, όπου x_j είναι ένα χαρακτηριστικό και d_j είναι ένα όριο διαχωρισμού. Ειδικότερα, μια παρατήρηση i τοποθετείται στον κόμβο t_L εάν $x_{ij} \le d_j$, 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:

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

όπου: s ο διαχωρισμός των αντικειμένων από τον κανόνα που αναπτύσσεται, p_L η αναλογία των περιπτώσεων του κόμβου t που καταλήγουν στον αριστερό κόμβο t_L, p_R η αναλογία των περιπτώσεων του κόμβου t που καταλήγουν στο δεξί κόμβο t_R, $i(t_L)$ η ομοιογένεια του αριστερού κόμβου και $i(t_R)$ η ομοιογένεια του δεξιού κόμβου.

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.

Οι Breiman et al. (1984) και Steinberg και Colla (1995) τονίζουν ως βασικά πλεονεκτήματα της μεθόδου CART τα ακόλουθα σημεία:

1. Δεν πραγματοποιείται καμία στατιστική υπόθεση όσον αφορά τα χαρακτηριστικά.
2. Είναι δυνατή η χρησιμοποίηση τόσο ποιοτικών όσο και ποσοτικών χαρακτηριστικών.
3. Είναι δυνατή η ανάπτυξη δέντρων ακόμα και από δεδομένα που δεν είναι πλήρη.
4. Τα αποτελέσματα της μεθόδου CART δεν επηρεάζονται από την ύπαρξη ακραίων δεδομένων (outliers), από φαινόμενα πολυσυγγραμμικότητας (multicollinearity) ή άλλα στατιστικά προβλήματα.
5. Η CART έχει τη δυνατότητα να αναζητά και να αποκαλύπτει τις αλληλεπιδράσεις των μεταβλητών μέσα στο σύνολο των δεδομένων.

6. Τα αποτελέσματα της μεθόδου παραμένουν αμετάβλητα ανεξάρτητα από πιθανούς μονότονους μετασχηματισμούς των δεδομένων.

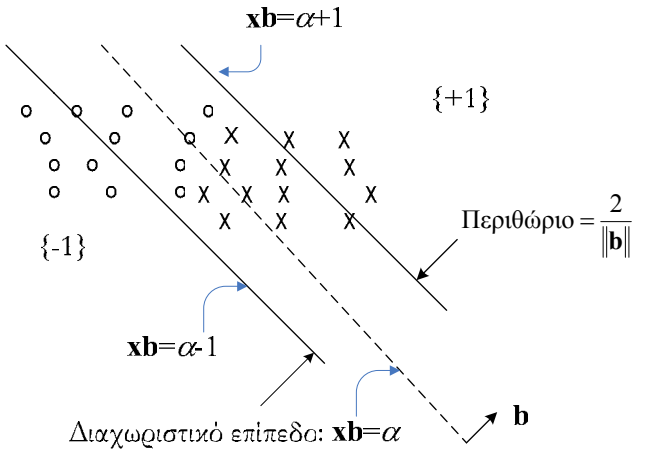
7. Μπορεί να παράγει χρήσιμα αποτελέσματα από ένα μεγάλο αριθμό μεταβλητών που παρέχονται προς ανάλυση, χρησιμοποιώντας μόνο ελάχιστες σημαντικές μεταβλητές.

8. Η κατανόηση των δέντρων ταξινόμησης της μεθόδου CART είναι ιδιαίτερα εύκολη.

2.2.6 Μηχανές διανύσματος υποστήριξης

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 2.4 όπου απεικονίζεται ένα πρόβλημα ταξινόμησης n αντικειμένων οι οποίες περιγράφονται βάσει m χαρακτηριστικών, σε δύο κατηγορίες οι οποίες συμβολίζονται ως +1 και -1.



Σχήμα 2.4: 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 **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):

$$\left.\begin{aligned} &\min_{\mathbf{b},\alpha,\mathbf{d}} \quad \nu\mathbf{e}^T\mathbf{d}+\frac{1}{2}\mathbf{b}^T\mathbf{b} \\ &\text{υπό}: \\ &\qquad \mathbf{D}\left(\mathbf{x}\mathbf{b}-\mathbf{e}\alpha\right)+\mathbf{d}\geq\mathbf{e} \\ &\qquad \mathbf{d}\geq 0 \end{aligned}\right\} \quad (4.6)$$

Ο τετραγωνικός όρος $\mathbf{b}^T\mathbf{b}$ στην αντικειμενική συνάρτηση του προβλήματος (4.6) μεγιστοποιεί το περιθώριο μεταξύ των δυο υπερεπιπέδων $\mathbf{x}\mathbf{b}-\alpha=+1$ και $\mathbf{x}\mathbf{b}-\alpha=-1$, 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 $\mathbf{x}\mathbf{b}=\alpha+1$ περικλείει όλα τα αντικείμενα της κατηγορίας +1, ενώ το επίπεδο $\mathbf{x}\mathbf{b}=\alpha-1$ περικλείει όλα τα αντικείμενα της κατηγορίας -1.

Με την επίλυση του προβλήματος (4.6) και τον προσδιορισμό των $\mathbf{b}$ και α που καθορίζουν το βέλτιστο υπερεπίπεδο, η ταξινόμηση κάθε αντικειμένου μπορεί εύκολα να πραγματοποιηθεί ως εξής:

$$\text{Εάν} \quad \mathbf{x}_i\mathbf{b}-\alpha\begin{cases} >0, & \text{τότε} \quad \mathbf{x}_i\in\{+1\}, \\ <0, & \text{τότε} \quad \mathbf{x}_i\in\{-1\}, \\ =0, & \text{τότε} \quad \mathbf{x}_i\in\{+1\} \quad \text{ή} \quad \mathbf{x}_i\in\{-1\} \end{cases} \quad (4.7)$$

Το κύριο μειονέκτημα του προβλήματος βελτιστοποίησης (4.6) 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):

$$\left.\begin{aligned} &\min_{\mathbf{b},\alpha,\mathbf{d}} \ \nu\frac{1}{2}\|\mathbf{d}\|^2+\frac{1}{2}\left(\mathbf{b}^T\mathbf{b}+\alpha^2\right) \\ &\text{υπό :} \\ &\qquad \mathbf{D}\left(\mathbf{A}\mathbf{b}-\mathbf{e}\alpha\right)+\mathbf{d}\geq\mathbf{e} \end{aligned}\right\} \quad (4.8)$$

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 (4.6)). 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 πρόβλημα (4.8) απλοποιείται ακόμα περισσότερο εάν ο περιορισμός ανισότητας μετατραπεί σε περιορισμό ισότητας ως εξής::

$$\left.\begin{aligned} &\min_{\mathbf{b},\alpha,\mathbf{d}} \ \nu\frac{1}{2}\|\mathbf{d}\|^2+\frac{1}{2}\left(\mathbf{b}^T\mathbf{b}+\alpha^2\right) \\ &\text{υπό :} \\ &\qquad \mathbf{D}\left(\mathbf{A}\mathbf{b}-\mathbf{e}\alpha\right)+\mathbf{d}=\mathbf{e} \end{aligned}\right\} \quad (4.9)$$

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 $\Phi: R^n \to H$. 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 εντιμετώπιση 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 πρόβλημα βελτιστοποίησης (4.9) διατυπώνεται ως εξής:

$$\left.\begin{aligned}&\min_{\mathbf{u},\alpha,\mathbf{d}}\ \nu\frac{1}{2}\|\mathbf{d}\|^2+\frac{1}{2}\left(\mathbf{u}^T\mathbf{u}+\alpha^2\right)\\&\text{υπό:}\\&\qquad\mathbf{D}\left(K(\mathbf{A},\mathbf{A})\mathbf{D}\mathbf{u}-\mathbf{e}\alpha\right)+\mathbf{d}=\mathbf{e}\end{aligned}\right\}\quad(4.10)$$

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

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

2.3 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) για τον έλεγχο της αποτελεσματικότητας υποδειγμάτων ταξινόμησης και παλινδρόμησης βάσει ενός συνόλου δεδομένων (σύνολο αναφοράς). 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 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 $A_1, A_2, ..., A_T$ μεγέθους n/T.

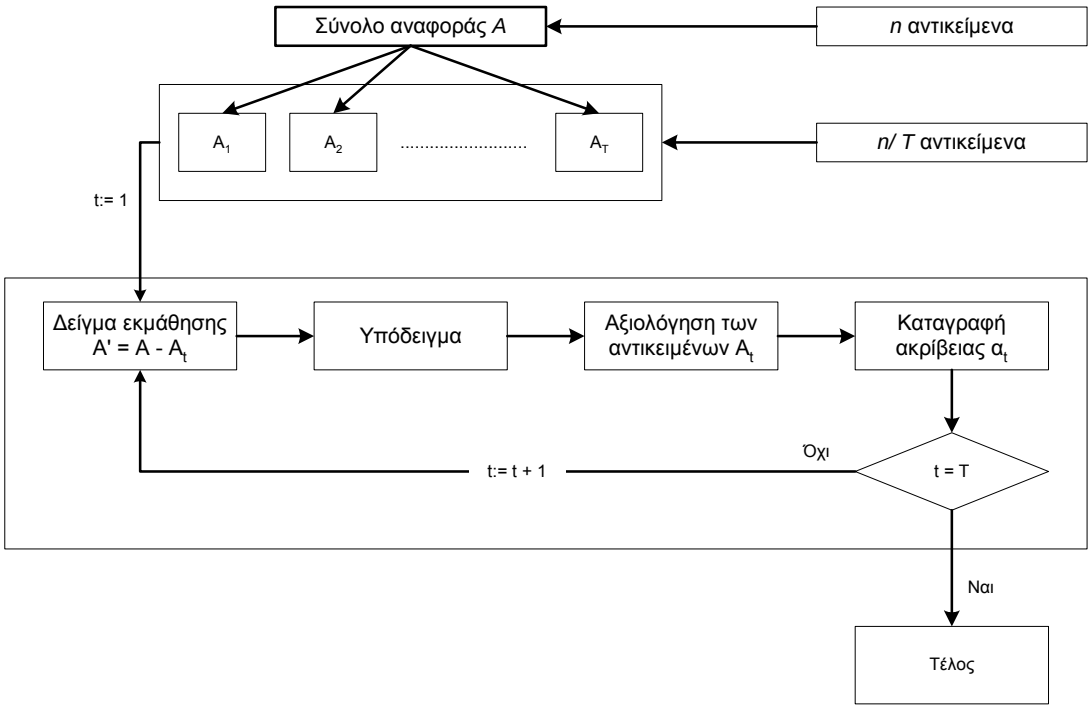
2. Για την πρώτη επανάληψη τίθεται $t = 1$.

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 2.5: Σχηματική απεικόνιση της διαδικασίας ελέγχου k-fold cross validation

2.4 Περίληψη

Σε αυτό το κεφάλαιο, έγινε, μια συνοπτική εισαγωγή του προβλήματος της ταξινόμησης και των βασικών σημείων έρευνας στο χώρο του προβλήματος της ταξινόμησης. Επίσης, πραγματοποιήθηκε μια συνοπτική περιγραφή της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης και της σπουδαιότητας του προβλήματος της ταξινόμησης τόσο σε θεωρητικό, όσο και πρακτικό επίπεδο. Επιπλέον, έγινε μια γενική περιγραφή των εξεταζόμενων τεχνικών ταξινόμησης. Τέλος, έγινε μια σύντομη αναφορά στη διαδικασία cross-validation, η οποία είναι μια τεχνική επαναληπτικής δειγματοληψίας για τον έλεγχο της αποτελεσματικότητας των αναπτυσσόμενων υποδειγμάτων ταξινόμησης. Εν συνεχεία, το επόμενο κεφαλαίο αναφέρεται στη λειτουργία των αλγορίθμων επιλογής χαρακτηριστικών και των ιδιοτήτων τους.

Κ ε φ ά λ α ι ο 3 ο

ΑΛΓΟΡΙΘΜΟΙ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

3.1 Εισαγωγή

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.

3.2 Στρατηγικές διερεύνησης

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 $O(2^m)$[1]. 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.

3.3 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 Q της ποιότητας κάθε υποσυνόλου χαρακτηριστικών. Έτσι με δεδομένο ένα υποσύνολο χαρακτηριστικών

[1] Χωρίς να λαμβάνεται υπόψη ο υπολογιστικός φόρτος της τεχνικής ταξινόμησης που χρησιμοποιείται.

$X_t \subset X$ το οποίο έχει επιλεχθεί στο στάδιο t της εμπρόσθιας διαδικασίας, το νέο υποσύνολο X_{t+1} στο επόμενο στάδιο διαμορφώνεται έτσι ώστε:

$$X_{t+1} = X_t \cup \{x_i \in X \setminus X_t \mid x_i = \arg\max Q(X_t \cup x_i)\}$$

Η επαναληπτική διαδικασία τερματίζεται όταν όλα τα χαρακτηριστικά επιλέγονται ή εναλλακτικά, εάν το κριτήριο αξιολόγησης Q δεν παρουσιάζει αισθητή βελτίωση σε μια ακολουθία συνεχόμενων επαναλήψεων ή πάρει μικρότερη τιμή από ένα προκαθορισμένο όριο Q_0. Ο υπολογιστικός φόρτος τέτοιων διαδικασιών είναι $O(m)$, αλλά βασικό τους μειονέκτημα είναι ότι δεν λαμβάνουν υπόψη πιθανές αλληλεπιδράσεις μεταξύ των 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 Q. Έτσι με δεδομένο ένα υποσύνολο χαρακτηριστικών $X_t \subset X$ το οποίο έχει επιλεχθεί στο στάδιο t της διαδικασίας, το νέο υποσύνολο X_{t+1} στο επόμενο στάδιο διαμορφώνεται έτσι ώστε:

$$X_{t+1} = X_t \setminus \{x_i \in X_t \mid x_i = \arg\max Q(X_t \setminus x_i)\}$$

Η επαναληπτική διαδικασία τερματίζεται όταν $|X_t| = 1$ ή εναλλακτικά, όταν το κριτήριο αξιολόγησης Q πάρει μικρότερη τιμή από ένα προκαθορισμένο όριο Q_0. Όπως και στην περίπτωση εμπρόσθιων διαδικασιών, ο υπολογιστικός φόρτος μιας ανάστροφης προσέγγισης είναι $O(m)$, στην πράξη όμως η υλοποίησή τους απαιτεί συνήθως μεγαλύτερο χρόνο σε σύγκριση με τις εμπρόσθιες διαδικασίες.

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. Ο παράγοντας της τυχαιότητας μπορεί να ενσωματωθεί και στις προηγούμενες προσεγγίσεις, αλλά αυτές σε κάθε περίπτωση βασίζονται σε κάποια προκαθορισμένα κριτήρια που αφορούν, είτε τον τρόπο διαμόρφωσης των νέων λύσεων, είτε την αξιολόγηση της ποιότητας των λύσεων.

3.4 Κριτήρια αξιολόγησης

Κάθε στρατηγική διερεύνησης και διαδικασία επιλογής χαρακτηριστικών βασίζεται σε κάποιο κριτήριο αξιολόγησης $Q(X') \to \mathbb{R}$ για τη μέτρηση της ποιότητας του κάθε υποσυνόλου χαρακτηριστικών $X' \subset X$ που επιλέγεται.

Το πλέον δημοφιλές κριτήριο αφορά τη μεγιστοποίηση της αναμενόμενης ακρίβειας ταξινόμησης η οποίο ορίζεται ως εξής (Devijver and Kittler, 1982):

$$Q = \sum_{i=1}^{q} p(C_i) \int_{f(\mathbf{x}) \in C_i} p\left(\mathbf{x} \mid C_i\right) d\mathbf{x}$$

όπου $p(C_i)$ είναι η *a priori* πιθανότητα της κατηγορίας C_i $(i = 1, ..., q)$ και $p\left(\mathbf{x} \mid C_i\right)$ είναι η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας της κατηγορία C_i.

Ένα εναλλακτικό κριτήριο αφορά την επιλογή χαρακτηριστικών, ώστε να μεγιστοποιούνται οι διαφορές στις κατανομές πυκνότητας πιθανότητας των κατηγοριών. Στην περίπτωση δύο κατηγοριών, το κριτήριο αυτό έχει την ακόλουθη γενική μορφή:

$$Q = \int d\left[p\left(\mathbf{x} \mid C_1\right), p\left(\mathbf{x} \mid C_2\right)\right] d\mathbf{x}$$

όπου d είναι μια πραγματική συνάρτηση οριζόμενη έτσι ώστε: (α) η Q να παίρνει πάντα θετικές τιμές, (β) $Q = 0$, όταν οι κατανομές πυκνότητας πιθανότητας των κατηγοριών ταυτίζονται, και (γ) η μέγιστη τιμή της συνάρτησης Q να λαμβάνεται όταν δεν υπάρχει επικάλυψη 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 $X' \subset X$ θεωρείται συνεπές, όταν δεν υπάρχουν αντικείμενα διαφορετικών κατηγοριών με την ίδια περιγραφή στα χαρακτηριστικά του συνόλου X'.

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 Αλγόριθμοι επιλογής χαρακτηριστικών

Οι αλγόριθμοι επιλογής χαρακτηριστικών 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.

3.5.1 Ενσωματωμένες διαδικασίες

Στην κατηγορία αυτή 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).

3.5.2 Διαδικασίες filter

Οι αλγόριθμοι επιλογής χαρακτηριστικών που εντάσσονται στην κατηγορία 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 αυτής της κατηγορία 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 αλγόριθμος αυτής της κατηγορίας 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).

3.5.3 Διαδικασίες wrapper

Αλγόριθμοι επιλογής χαρακτηριστικών της κατηγορίας αυτής 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).

Διάφοροι αλγόριθμοι της κατηγορίας αυτής παρουσιάζονται και αναλύονται στην εργασία των Kohavi and John (1997). Στην ίδια κατηγορία μπορούν να ενταχθούν και τεχνικές που βασίζονται στη χρήση γενετικών αλγορίθμων, οι οποίες χρησιμοποιούνται σε συνδυασμό με τεχνικές ταξινόμησης, όπως τα νευρωνικά δίκτυα και οι μηχανές διανύσματος υποστήριξης. Κύριο πλεονέκτημα των γενετικών αλγορίθμων αποτελεί η ικανότητά τους να διερευνούν μεγάλα σύνολα χαρακτηριστικών, ενώ ταυτόχρονα έχει παρατηρηθεί ότι παρουσιάζουν περιορισμένη ευαισθησία στην ύπαρξη θορύβου (Vafai and De Jong, 1992). Χαρακτηριστικό παράδειγμα των δυνατοτήτων που προσφέρουν τέτοιες τεχνικές είναι ο πολυκριτήριος γενετικός αλγόριθμος που χρησιμοποιήθηκε από τους Oliveira et al. (2002) σε συνδυασμό με ένα νευρωνικό δίκτυο για την αναγνώριση γραπτών κειμένων.

3.6 Συγκριτικές έρευνες

Στη διεθνή βιβλιογραφία έχουν παρουσιαστεί διάφορες συγκριτικές έρευνες για την αξιολόγηση των υπαρχόντων προσεγγίσεων. Στον πίνακα που ακολουθεί συνοψίζονται κάποιες χαρακτηριστικές έρευνες στο θέμα αυτό. Για κάθε έρευνα καταγράφονται οι

αλγόριθμοι επιλογής χαρακτηριστικών που εξετάστηκαν, το πλήθος των δεδομένων που χρησιμοποιήθηκαν (πραγματικά και τεχνητά) και οι τεχνικές ταξινόμησης που εξετάστηκαν.

Πίνακας 3.1: Συγκριτικές έρευνες για την αξιολόγηση των υπαρχόντων προσεγγίσεων

Έρευνες	FSAs	Δεδομένα (πραγματικά/ τεχνητά)	Τεχνικές ταξινόμησης
Salzberg (1992)	Combined Stepwise Selection	4/-	Πλησιέστερος γείτονας, νευρωνικά δίκτυα, CART και ID3
John et al. (1994)	Διαδικασίες wrapper: Backward stepwise selection και Forward stepwise selection Διαδικασίες filter: Relief	4/4	ID3, C4.5
Choubey, et al. (1996)	Sequential Backward Selection algorithms: Best fit SBS, Hybrid heuristic SBS, Alternating heuristic SBS και K-level best SBS	8/5	Προσεγγιστικά σύνολα
Liu and Setiono (1996a)	Las Vegas Filter	4/5	ID3, C4.5
Liu and Setiono (1998a)	Las Vegas Filter	4/6	ID3, Naïve Bayes
Hall and Holmes (2000)	Information gain attribute ranking (Dumais et al., 1998), Relief, Principal components, Correlation-based feature selection (Hall, 1999), Consistency-based subset evaluation (Almuallim and Dietterich, 1991, Liu and Setiono, 1996b) και Wrapper subset evaluation	15/-	C4.5, Naïve Bayes
Yu and Liu (2003)	Fast correlation-based filter, ReliefF, CorrSF και ConsSF	10/-	C4.5 και NBC (Witten and Frank, 2000)

Σε μια πρώτη έρευνα που πραγματοποιήθηκε από τον Salzberg (1992), διαπιστώθηκε ότι η αποτελεσματικότητα του αλγορίθμου του πλησιέστερου γείτονα βελτιώνεται χρησιμοποιώντας ένα περιορισμένο σύνολο χαρακτηριστικών, ενώ αντίθετα για άλλες τεχνικές ταξινόμησης δεν παρατηρήθηκε ουσιαστική βελτίωση. Οι John et al. (1994) διαπίστωσαν ότι στην περίπτωση πραγματικών δεδομένων η επιλογή των χαρακτηριστικών με διαδικασίες wrapper δεν απέφερε ικανοποιητική βελτίωση στην ακρίβεια υποδειγμάτων ταξινόμησης που αναπτύχθηκαν με τους αλγορίθμους ID3 και C4.5. Αντίθετα, σε κάποια από τα τεχνητά δεδομένα που εξέτασαν την εφαρμογή του αλγορίθμου επιλογής

χαρακτηριστικών backward stepwise selection, οδήγησε σε σημαντική μείωση του σφάλματος ταξινόμησης. Σε κάθε περίπτωση πάντως, παρατηρήθηκε σημαντική μείωση της πολυπλοκότητας των αναπτυσσόμενων υποδειγμάτων ταξινόμησης.

Οι Choubey et al. (1996) εξέτασαν διάφορους αλγορίθμους επιλογής χαρακτηριστικών χρησιμοποιώντας τα προσεγγιστικά σύνολα ως τεχνική ανάπτυξης των υποδειγμάτων ταξινόμησης. Τα αποτελέσματά τους έδειξαν ότι στην πλειοψηφία των περιπτώσεων παρατηρείται βελτίωση της ακρίβειας των υποδειγμάτων ταξινόμησης, όταν χρησιμοποιούνται αλγόριθμοι επιλογής χαρακτηριστικών. Οι Liu and Setiono (1996a) παρατήρησαν ότι η επιλογή ενός μικρού αριθμού χαρακτηριστικών από το αρχικό δείγμα οδήγησε στη μείωση του σφάλματος ταξινόμησης τόσο στα τεχνικά, όσο και στα πραγματικά δεδομένα για υποδείγματα ταξινόμησης που αναπτύσσονται με τους αλγορίθμους ID3 και Naïve Bayes. Οι ίδιοι ερευνητές το 1998 (Liu and Setiono, 1998a), παρατήρησαν ότι ο αλγόριθμος επιλογής χαρακτηριστικών Las Vegas Filter σε αρκετές περιπτώσεις οδηγεί σε αύξηση του σφάλματος ταξινόμησης, φαινόμενο το οποίο αποδόθηκε από τους συγγραφείς στην ανεπάρκεια των εξεταζόμενων δεδομένων. Επίσης, παρατηρήθηκε ότι ο αλγόριθμος C4.5 δίνει συγκριτικά καλύτερα αποτελέσματα για ένα μικρό αριθμό χαρακτηριστικών, σε σχέση με τον Naïve Bayes.

Οι Hall and Holmes (2000) διαπίστωσαν ότι η χρήση διαδικασιών wrapper συμβάλλει στην βελτίωση της ακρίβειας των υποδειγμάτων που αναπτύσσονται με τον αλγόριθμο Naïve Bayes, ενώ αντίθετα τα υποδείγματα ταξινόμησης που αναπτύσσονται με τον C4.5 παρουσιάζουν υψηλότερη ακρίβεια, όταν χρησιμοποιηθεί ο αλγόριθμος επιλογής χαρακτηριστικών Relief. Γενικά, παρατηρήθηκε ότι ο αλγόριθμος Naïve Bayes, σε αντίθεση με τον C4.5, δεν αποδίδει ικανοποιητικά αποτελέσματα χωρίς την εφαρμογή κάποιου αλγορίθμου επιλογής χαρακτηριστικών. Σε μια άλλη έρευνα, οι Yu and Liu (2003) πρότειναν έναν νέο αλγόριθμο επιλογής χαρακτηριστικών, τον fast correlation-based filter, ο οποίος βελτίωσε την ακρίβεια δύο τεχνικών ταξινόμησης, παρουσιάζοντας ταυτόχρονα μικρότερο υπολογιστικό φόρτο σε σχέση με άλλους αλγόριθμους επιλογής χαρακτηριστικών, όπως είναι ο ReliefF (Kononenko, 1994), ο CorrSF (Hall, 1999) και ο ConsSF. Από τους αλγόριθμους αυτούς διαπιστώθηκε ότι μόνο ο CorrSF οδηγεί σε βελτίωση της αποτελεσματικότητας των αναπτυσσόμενων υποδειγμάτων ταξινόμησης.

3.7 Περίληψη

Σε αυτό το κεφάλαιο έγινε μια συνοπτική αναφορά των αλγορίθμων επιλογής χαρακτηριστικών βάσει των ιδιοτήτων που τους χαρακτηρίζει. Η ανασκόπηση των αλγορίθμων παρουσιάστηκε σύμφωνα με τη στρατηγική διερεύνησης των χαρακτηριστικών που ακολουθούν, τη διαδικασία επιλογής των χαρακτηριστικών και τα κριτήρια αποτίμησης της ποιότητας των επιλεγμένων χαρακτηριστικών. Η ομαδοποίηση των αλγορίθμων έγινε σε τρεις κατηγορίες ανάλογα με τον τρόπο λειτουργίας τους, σε ενσωματωμένες, filter ή wrapper διαδικασίες. Τέλος, παρουσιάστηκαν, από τη διεθνή βιβλιογραφία, κάποιες συγκριτικές έρευνες για την αξιολόγηση των υπαρχόντων προσεγγίσεων και δίνονται συνοπτικά κάποια αποτελέσματα και συμπεράσματα που προέκυψαν από τη διεξαγωγή των σχετικών ερευνών πάνω στους αλγόριθμους επιλογής χαρακτηριστικών. Παρακάτω παρουσιάζονται αναλυτικά τα αποτελέσματα της αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών (κεφάλαιο 3ο) στην ανάπτυξη κατάλληλων υποδειγμάτων ταξινόμησης (κεφάλαιο 2ο), αφού γίνεται αρχικά κάποια πιο αναλυτική αναφορά στους επιλεγμένους αλγορίθμους επιλογής χαρακτηριστικών υπό τη μορφή ψευδοκώδικα.

Κεφάλαιο 4ο

ΠΕΙΡΑΜΑΤΙΚΗ ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ

4.1 Σκοπός της έρευνας

Στόχος της προτεινόμενης έρευνας είναι να πραγματοποιήσει μια συνολική και διεξοδική ανάλυση της αποτελεσματικότητας των μεθοδολογιών επιλογής χαρακτηριστικών καλύπτοντας ένα ευρύ φάσμα μεθόδων ταξινόμησης και δεδομένων. Η ανάλυση αυτή θα συμβάλλει ουσιαστικά, στον εντοπισμό των αλληλεπιδράσεων που πιθανόν να υπάρχουν μεταξύ:

1. Της αποτελεσματικότητας των μεθόδων επιλογής χαρακτηριστικών και των τεχνικών ανάπτυξης υποδειγμάτων ταξινόμησης.

2. Της αποτελεσματικότητας των μεθόδων επιλογής χαρακτηριστικών και των ιδιοτήτων των εξεταζόμενων δεδομένων.

Η έννοια της αποτελεσματικότητας αναφέρεται στην αναμενόμενη ακρίβεια των αναπτυσσόμενων υποδειγμάτων ταξινόμησης, στη δυνατότητα εντοπισμού των πραγματικά χρήσιμων χαρακτηριστικών και στο ποσοστό μείωσης της εξεταζόμενης πληροφορίας. Με την υλοποίηση της έρευνας αυτής θα εξαχθούν χρήσιμα συμπεράσματα τα οποία αφορούν όχι μόνο τις μεθοδολογίες επιλογής χαρακτηριστικών, αλλά και τις συνθήκες υπό τις οποίες η χρήση διαφόρων τεχνικών ταξινόμησης μπορεί να ωφεληθεί από την επιλογή κατάλληλων χαρακτηριστικών, στοιχείο το οποίο δεν έχει καλυφθεί σε υπάρχουσες έρευνες (Blum and Langley, 1997).

4.2 Αλγόριθμοι επιλογής χαρακτηριστικών (FSAs)

Βασικό χαρακτηριστικό της συγκεκριμένης έρευνας είναι η διερεύνηση της αποτελεσματικότητας ενός ικανοποιητικού αριθμού αλγορίθμων επιλογής χαρακτηριστικών. Συγκεκριμένα, στην παρακάτω πειραματική ανάλυση χρησιμοποιούνται εννιά αλγόριθμοι, οι οποίοι εφαρμόζουν τόσο διαδικασίες filter, όσο και διαδικασίες wrapper για την επιλογή του πλέον αξιόπιστου και ικανοποιητικού υποσυνόλου χαρακτηριστικών. Για όλους τους αλγόριθμους filter, η εκτίμηση της ποιότητας των χαρακτηριστικών βασίζεται στον υπολογισμό ενός κριτηρίου απόκλισης $J = tr\left(S_w^{-1}S_b\right)$, όπου S_w είναι ο πίνακας διασποράς εντός των κατηγοριών και S_b ο πίνακας διασποράς μεταξύ των κατηγοριών (Fukunaga,1990). Το κριτήριο αυτό είναι μονοτονικό και παρέχει δυνατότητα χρήσης τόσο ποσοτικών, όσο και ποιοτικών χαρακτηριστικών. Όσο υψηλότερη είναι η τιμή αυτού του κριτηρίου, τόσο καλύτερη και αξιόπιστη είναι η ταξινόμηση των αντικειμένων του δείγματος βάσει των επιλεγμένων κριτηρίων. Συνεπώς, τα χαρακτηριστικά που σημειώνουν υψηλή τιμή για το κριτήριο απόκλισης, θεωρούνται και πιο κατάλληλα.

Η πρώτη κατηγορία των FSAs που χρησιμοποιούνται κατά την πειραματική ανάλυση περιλαμβάνει τους αλγορίθμους Las Vegas, οι οποίοι είναι, κατά κύριο λόγο, ο αλγόριθμος Las Vegas Filter (LVF), ο αλγόριθμος Las Vegas Incremental (LVI) και ο αλγόριθμος Las Vegas Wrapper (LVW).

Ο αλγόριθμος LVF ακολουθεί μια στρατηγική τυχαίας διερεύνησης για την επιλογή χαρακτηριστικών, βελτιστοποιώντας το κριτήριο ποιότητας των επιλεγμένων χαρακτηριστικών. Ο τρόπος λειτουργίας του αλγορίθμου LVF κατά τη φάση της ανάλυσης φαίνεται στο σχήμα 4.1 υπό τη μορφή ψευδοκώδικα. Η χρήση του αλγορίθμου αυτού βασίζεται στις παραμέτρους a και β. Η παράμετρος a τίθεται ίση με 0.9 και χρησιμοποιείται για να διασφαλίσει ότι κάθε επιλεγμένο υποσύνολο χαρακτηριστικών θα έχει απόκλιση τουλάχιστον ίση με 90% της απόκλισης που υπολογίζεται με βάση όλα τα χαρακτηριστικά του δείγματος. Η δεύτερη παράμετρος β τίθεται επίσης, ίση με 0.9 και χρησιμοποιείται για να διασφαλίσει ότι ένα υποψήφιο υποσύνολο χαρακτηριστικών λιγότερο αξιόπιστο, συγκρινόμενο με το τρέχων υποσύνολο επιλεγμένων χαρακτηριστικών, θα επιλεχθεί έναν και μόνο εάν η τιμή απόκλισης που θα προκύψει, είναι τουλάχιστον ίση με το 90% της εκτιμώμενης απόκλισης από το τρέχων υποσύνολο των επιλεγμένων χαρακτηριστικών. Ο μέγιστος αριθμός των επαναλήψεων, κατά τη διαδικασία εφαρμογής του αλγορίθμου LVF

στην επιλογή του πλέον κατάλληλου και αξιόπιστου συνόλου χαρακτηριστικών T τίθεται ίσος με $50n$, όπου n είναι το πλήθος των χαρακτηριστικών. Ο αλγόριθμος LVF χρησιμοποιείται επίσης, σε διαδικασίες wrapper, όπου σε αυτή την περίπτωση ως κριτήριο εκτίμησης της ποιότητας των επιλεγμένων χαρακτηριστικών λαμβάνεται η αναμενόμενη ακρίβεια ταξινόμησης που υπολογίζεται χρησιμοποιώντας τη διαδικασία ελέγχου 5-fold cross validation. Στην προκείμενη περίπτωση, ο μέγιστος αριθμός επαναλήψεων T τίθεται ίσος με $5n$, ενώ ο παράμετροι a και β είναι ίσοι με τη μονάδα.

Input:

T – Μέγιστος αριθμός επαναλήψεων

A – Δείγμα εκμάθησης αποτελούμενο από ένα σύνολο m αντικειμένων και n χαρακτηριστικών

$J(A, X)$ – Κριτήριο αξιολόγησης της ποιότητας του συνόλου των χαρακτηριστικών X στο δείγμα εκμάθησης A

Output:

X_s – Σύνολο των χαρακτηριστικών που επιλέγονται

Initialization: $X_S := X$, $J_0 := J(A, X)$, $t := 1$

While $(t < T)$ and $(|X_S| > 1)$ **do**

  $X' :=$ τυχαίο υποσύνολο X

  If $(|X'| < |X_S|)$ and $(J(A, X') \geq \alpha J_0)$ and $(J(A, X') \geq \beta J(A, X_S))$ **then**

    $X_S := X'$

  end

  $t := t + 1$

end

Σχήμα 4.1: Ο αλγόριθμος Las Vegas Filter

Ο αλγόριθμος LVI είναι προέκταση του αλγορίθμου LVF και χρησιμοποιεί ένας μέρος του δείγματος εκμάθησης, το μέγεθος του οποίου αυξάνεται σταδιακά εξαρτώμενο από την ποιότητα των επιλεγμένων χαρακτηριστικών. Ο τρόπος λειτουργίας του αλγορίθμου LVI κατά τη φάση της ανάλυσης φαίνεται στο σχήμα 4.2 υπό τη μορφή ψευδοκώδικα. Το μέρος του δείγματος εκμάθησης που χρησιμοποιείται από τον αλγόριθμο LVI καθορίζεται από μια παράμετρο p, η τιμή της οποίας είναι 20%. Είναι σημαντικό να σημειωθεί ότι ο αλγόριθμος LVI χρησιμοποιεί τον αλγόριθμο LVF ως μέρος της διαδικασίας επιλογής χαρακτηριστικών.

Input:

T – Μέγιστος αριθμός επαναλήψεων
A – Δείγμα εκμάθησης αποτελούμενο από ένα σύνολο m αντικειμένων και n χαρακτηριστικών
p – ποσοστό επί του συνόλου των m αντικείμενων (σε %)
$J(A, X)$ – Κριτήριο αξιολόγησης της ποιότητας του συνόλου των χαρακτηριστικών X στο δείγμα εκμάθησης A

Output:

X_s – Σύνολο των χαρακτηριστικών που επιλέγονται

Initialization: $X_S := X$, $A_0 := p\%$ of A τυχαία επιλεγμένο, $A_1 := A \setminus A_0$, $J_0 := J(A_0, X)$
Loop
$X_S := LVF(A_0, J, T)$
If $\left(J(A, X_S) \geq \alpha J_0\right)$ **then** Stop
else
$C :=$παρατηρήσεις του συνόλου A_1 με αρνητική επίδραση στο κριτηρίο αξιολόγησης J
$A_0 := A_1 \cup C$
$A_1 := A_1 \setminus C$
end
end

Σχήμα 4.2: Ο αλγόριθμος Las Vegas Incremental

Ο αλγόριθμος FOCUS ανήκει στην κατηγορία των διαδικασιών filter και ακολουθεί εκθετική διερεύνηση για την επιλογή ενός κατάλληλου υποσυνόλου χαρακτηριστικών, βελτιστοποιώντας το κριτήριο ποιότητας των επιλεγμένων χαρακτηριστικών. Ο τρόπος λειτουργίας του αλγορίθμου FOCUS περιγράφεται με τη βοήθεια του ψευδοκώδικα που παρουσιάζεται στο σχήμα 4.3. Για τη μείωση του υπολογιστικού φόρτου και την εφαρμογή του αλγορίθμου σε διαδοχικά χαρακτηριστικά έγιναν κάποιες τροποποιήσεις αυτού σε σχέση με την αρχική προτεινόμενη μορφή του από τους Almuallim και Dietterich (1991). Συγκεκριμένα, αντί ο αλγόριθμος να εξετάζει την ποιότητα όλων των δυνατών υποσυνόλων των χαρακτηριστικών, ξεκινώντας αρχικά από την επιλογή ενός χαρακτηριστικού και αυξάνοντας σταδιακά το πλήθος αυτών που επιλέγονται, η υλοποίηση του αλγορίθμου που χρησιμοποιείται στην παρούσα εργασία ξεκινά με το σύνολο των χαρακτηριστικών και επαναληπτικά αφαιρούνται εκείνα τα χαρακτηριστικά που θεωρούνται μη ουσιαστικά και τα οποία δεν επιφέρουν σημαντική μείωση στο κριτήριο ποιότητας J σε σχέση με το σύνολο των χαρακτηριστικών. Ένα υποσύνολο χαρακτηριστικών θεωρείται αποδεκτό και αξιόπιστο, εάν το κριτήριο ποιότητας των επιλεγμένων χαρακτηριστικών δεν είναι μικρότερο αυτού που προκύπτει βάσει του συνόλου των χαρακτηριστικών και μεγαλύτερο του προεπιλεγμένου ως καλύτερου

υποσυνόλου. Η υπόθεση αυτή ελέγχεται βάσει μιας παραμέτρου a, η οποία παίρνει την τιμή 0.9, όπως ακριβώς και στους αλγόριθμους Las Vegas που περιγράφθηκαν παραπάνω. Η διερεύνηση ενός αποδεκτού υποσυνόλου χαρακτηριστικών επαναλαμβάνεται T φορές (στην παρακάτω πειραματική ανάλυση το T παίρνει την τιμή 100).

Input:
T – Μέγιστος αριθμός επαναλήψεων
A – Δείγμα εκμάθησης αποτελούμενο από ένα σύνολο m αντικειμένων και n χαρακτηριστικών
$J(A, X)$ – Κριτήριο αξιολόγησης της ποιότητας του συνόλου των χαρακτηριστικών X στο δείγμα εκμάθησης A

Output:
X_s – Σύνολο των χαρακτηριστικών που επιλέγονται

Initialization: $X_S := X$, $J_0 := J(A, X)$
While $(t < T)$ **do**
  $X' :=$ Τυχαίο υποσύνολο των X_s μεγέθους το πολύ $|X_s| - 1$
  If $(J(A, X') \geq \alpha J_0)$ and $(J(A, X') > J(A, X_S))$ **then**
    $X_S := X'$
    $t := 1$
  else
    $t := t + 1$
  end
end

Σχήμα 4.3: Ο αλγόριθμος FOCUS

Εκτός από τους προηγούμενους αλγόριθμους, οι οποίοι ακολουθούν στρατηγικές τυχαίας ή εκθετικής διερεύνησης για την επιλογή αξιόπιστων χαρακτηριστικών, εξετάστηκαν και αλγόριθμοι που ακολουθούν στρατηγικές σειριακής διερεύνησης. Σε αυτή την κατηγορία ανήκουν οι αλγόριθμοι sequential forward και backward generation filter, καθώς και οι αντίστοιχοι αλγόριθμοι wrapper (SFGF, SBGF, SFGW, SBGW). Οι αλγόριθμοι SFGF και SBGF ως filter διαδικασίες εφαρμόζουν εμπρόσθια και ανάστροφη, αντίστοιχα, διαδικασία διερεύνησης χαρακτηριστικών. Οι αλγόριθμοι SFGW και SBGW είναι οι αντίστοιχες wrapper παραλλαγές των δύο τελευταίων αλγορίθμων. Ο αλγόριθμος SFGF προσθέτει βαθμιαία χαρακτηριστικά στο αρχικό υποσύνολο με σκοπό τη βελτίωση του κριτηρίου J. Η ίδια διαδικασία εφαρμόζεται και με τον αλγόριθμο SBGF με τη μόνη διαφορά ότι τώρα αφαιρούνται βαθμιαία εκείνα τα χαρακτηριστικά που δεν έχουν σημαντική επίδραση στο κριτήριο J. Στα σχήματα 4.4 και 4.5 φαίνονται υπό μορφή ψευδοκώδικα οι διαδικασίες επιλογής χαρακτηριστικών των αλγορίθμων SFGF και SBGF.

Σύμφωνα με τον αλγόριθμο SFGF, ένα χαρακτηριστικό προστίθεται στο σύνολο των προεπιλεγμένων χαρακτηριστικών, εάν αυτό βελτιώνει αισθητά την ήδη υπάρχουσα λύση. Η βελτίωση αυτή ελέγχεται μέσω μιας παραμέτρου a, η οποία παίρνει την τιμή 10%. Η ίδια παράμετρος λαμβάνεται υπόψη και στον αλγόριθμο SBGF. Σε αυτή την περίπτωση, η παράμετρος a προσδιορίζει τη μέγιστη αποδεκτή μείωση στη ποιότητα των επιλεγμένων χαρακτηριστικών, εν συγκρίσει με την ποιότητα J_0 του αρχικού πλήθους των χαρακτηριστικών. Στον αλγόριθμο SFGF χρησιμοποιείται και μια επιπλέον παράμετρος β για τον τερματισμό της εμπρόσθιας διαδικασίας, όταν η ήδη υπάρχουσα λύση δεν βελτιώνεται μετά από την πραγματοποίηση $\beta - 1$ διαδοχικών επαναλήψεων του αλγορίθμου. Στην συγκεκριμένη πειραματική ανάλυση θεωρείται ότι το β είναι ίσο με 4. Στις αντίστοιχες διαδικασίες wrapper των παραπάνω αλγορίθμων εφαρμόζονται διαδικασίες ελέγχου 5-fold cross validation, η ακρίβεια των οποίων χρησιμοποιείται ως κριτήριο αξιολόγησης J με παράμετρο a ίση με το μηδέν.

Input:

A – Δείγμα εκμάθησης αποτελούμενο από ένα σύνολο m αντικειμένων και n χαρακτηριστικών

$J(A, X)$ – Κριτήριο αξιολόγησης της ποιότητας του συνόλου των χαρακτηριστικών X στο δείγμα εκμάθησης A

Output:

X_s – Σύνολο των χαρακτηριστικών που επιλέγονται

Initialization: $X_S := \varnothing$, $J_0 := 0$, $t := 0$

While $\left(\left|X_S\right| < \left|X\right|\right)$ and $\left(t < \beta\right)$

$X' = X_S \cup \left\{x_j \in X \setminus X_S \mid x_j = \arg\max J\left(A, X_S \cup x_j\right)\right\}$

If $J\left(A, X'\right) > (1+\alpha) J_0$ **then**

$X_S^{best} := X'$

$X_S := X'$

$J_0 := J(A, X')$

$t := 0$

else

$X_S := X'$

$t := t + 1$

end

end

$X_S := X_S^{best}$

Σχήμα 4.4: Ο αλγόριθμος sequential forward generation

Input:

A – Δείγμα εκμάθησης αποτελούμενο από ένα σύνολο m αντικειμένων και n χαρακτηριστικών

$J(A, X)$ – Κριτήριο αξιολόγησης της ποιότητας του συνόλου των χαρακτηριστικών X στο δείγμα εκμάθησης A

Output:

X_s – Σύνολο των χαρακτηριστικών που επιλέγονται

Initialization: $X_S := X$, $J_0 := J(A, X)$

While $\left(\left|X_S\right| > 1\right)$

$X' = X_S \setminus \left\{x_j \in X_S \mid x_j = \arg\max J\left(A, X_S \setminus x_j\right)\right\}$

If $J\left(A, X'\right) \geq (1-\alpha) J_0$ **then**

$X_S := X'$

end

end

Σχήμα 4.5: Ο αλγόριθμος sequential backward generation

Τέλος, στην πειραματική ανάλυση χρησιμοποιείται και ο αλγόριθμος RELIEF, ο οποίος ακολουθεί διαδικασία filter βασισμένη σε μια στρατηγική τυχαίας διερεύνησης χαρακτηριστικών, μεγιστοποιώντας την ευκλείδεια απόσταση d που διαχωρίζει τις δύο κατηγορίες. Ο αλγόριθμος ξεκινάει με μια τυχαία επιλογή ενός αντικειμένου από το σύνολο αναφοράς A και προσδιορίζει το πλησιέστερο αντικείμενο αυτού από την ίδια κατηγορία (near hit) και το πλησιέστερο αντικείμενο αυτού από την άλλη κατηγορία (near miss). Βασικός σκοπός του αλγορίθμου RELIEF είναι να εκτιμήσει την ποιότητα των χαρακτηριστικών σύμφωνα με το πόσο καλά αυτά διαχωρίζουν τα αντικείμενα τα οποία βρίσκονται σε διαφορετικές κατηγορίες. Το αποτέλεσμα που προκύπτει από τον αλγόριθμο είναι ένα διάνυσμα βαρών, το οποίο προσδιορίζει τη σημαντικότητα του κάθε χαρακτηριστικού. Στο σχήμα 4.6 φαίνεται, υπό μορφή ψευδοκώδικα, η διαδικασία επιλογής χαρακτηριστικών του αλγορίθμου RELIEF. Βάσει του διανύσματος βαρών που προκύπτει, προσδιορίζεται το σύνολο των χαρακτηριστικών χρησιμοποιώντας τη διαδικασία που προτάθηκε από τους Molina et al. (2002). Πρέπει επίσης να τονιστεί ότι, για υπολογιστικούς λόγους, σε μεγάλα σύνολα δεδομένων ο αλγόριθμος RELIEF εφαρμόζεται μόνο σε ένα τυχαίο κομμάτι του συνόλου αναφοράς, το οποίο δεν ξεπερνάει τα 1000 αντικείμενα.

Input:

A – Δείγμα εκμάθησης αποτελούμενο από ένα σύνολο m αντικειμένων και n χαρακτηριστικών

Output:

X_s – Σύνολο των χαρακτηριστικών που επιλέγονται

Initialization: $t := 1$, $\mathbf{w} := 0$

While $(t < m)$

Επιλέγεται τυχαία ένα αντικείμενο $\mathbf{x}$ από το σύνολο αναφοράς A

$\mathbf{x}_{nh} = \text{Near hit}(A, \mathbf{x})$

$\mathbf{x}_{mis} = \text{Near miss}(A, \mathbf{x})$

For $j = 1$ **to** n

$$w(j) := w(j) + \frac{1}{m} d_j(\mathbf{x}, \mathbf{x}_m) - \frac{1}{m} d_j(\mathbf{x}, \mathbf{x}_h)$$

end

end

Σχήμα 4.6: Ο αλγόριθμος RELIEF

4.3 Εξεταζόμενες τεχνικές ταξινόμησης

Όπως προαναφέρθηκε, ένα ακόμα καινοτόμο στοιχείο της προτεινόμενη 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 τη γραμμική διακριτική ανάλυση και τη λογιστική παλινδρόμηση (McLachlan, 1992). 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 παλινδρόμησης και ταξινόμησης (classification and regression trees, Breiman et al., 1984), οι μηχανές διανύσματος υποστήριξης (support vector machines, Vapnik, 1998) και τα πιθανοτικά νευρωνικά δίκτυα (probabilistic neural networks, Specht, 1990).

4.4 Σχεδιασμός του πειράματος

Η υλοποίηση της ανάλυσης πραγματοποιείται σε δύο φάσεις. Αρχικά, γίνεται ένας εκτεταμένος πειραματικός σχεδιασμός ώστε να διερευνηθεί η αποτελεσματικότητα των αλγορίθμων επιλογής χαρακτηριστικών σε δεδομένα με προκαθορισμένες ιδιότητες (τεχνητά δεδομένα). Η ανάλυση αυτή βοηθάει στην εξαγωγή χρήσιμων συμπερασμάτων όσον αφορά τη σχέση που πιθανόν να υπάρχει μεταξύ της αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών και των ιδιοτήτων των δεδομένων. Σε δεύτερη φάση, εξετάζονται πραγματικά δεδομένα από τη βάση δεδομένων UCI Machine Learning Repository (Blake et al., 1998) τα οποία χρησιμοποιούνται ευρέως σε έρευνες σχετικές με θέματα ταξινόμησης.

Η πραγματοποίηση της ανάλυσης τόσο σε πειραματικά, όσο και σε πραγματικά δεδομένα συμβάλλει στη διαμόρφωση μιας ολοκληρωμένης εικόνας για την αποτελεσματικότητα των διαδικασιών επιλογής χαρακτηριστικών, των δυνατοτήτων που παρέχουν και τις υπάρχουσες αλληλεπιδράσεις με διάφορες τεχνικές ανάπτυξης υποδειγμάτων ταξινόμησης.

4.4.1 Εξεταζόμενοι παράγοντες

Τα δεδομένα αυτά αναπτύσσονται με τέτοιο τρόπο έτσι ώστε να διαθέτουν συγκεκριμένες ιδιότητες, βάσει των οποίων κάθε εναλλακτική (αντικείμενο) ταξινομείται σε δύο προκαθορισμένες κατηγορίες. Οι παράγοντες που εξετάζονται κατά το σχεδιασμό των τεχνητών δεδομένων περιλαμβάνουν:

1. Τη μορφή διάκρισης των κατηγοριών.
2. Το πλήθος των χαρακτηριστικών.
3. Τη μορφή των ακατάλληλων χαρακτηριστικών.
4. Το ποσοστό των κατάλληλων χαρακτηριστικών σε σχέση με το σύνολο των χαρακτηριστικών.

Για κάθε παράγοντα που ενσωματώνεται στην ανάλυση καθορίζονται κατάλληλα σενάρια βάσει των οποίων κατασκευάστηκαν τα τεχνητά δεδομένα με τις προκαθορισμένες ιδιότητες.

Ο πρώτος παράγοντας καθορίζει τη μορφή της διάκρισης των κατηγοριών. Οι περιπτώσεις που εξετάζονται αφορούν τη γραμμική και τη μη γραμμική (τετραγωνική) διάκριση των κατηγοριών. Οι συντελεστές των χαρακτηριστικών στη συνάρτηση διάκρισης ορίζονται ως τυχαία ομοιόμορφα κατανεμημένοι στο διάστημα [1, 2] για τους γραμμικούς όρους και στο διάστημα [0, 0.25] για τους τετραγωνικούς όρους. Στη συνάρτηση διάκρισης λαμβάνεται ένα τυχαίο κανονικά κατανεμημένο σφάλμα ταξινόμησης με μηδενική μέση τιμή και διακύμανση ίση με τη μονάδα.

Ο δεύτερος παράγοντας αναφέρεται στο πλήθος των χαρακτηριστικών που υπάρχουν σε κάθε ένα από τα σύνολα δεδομένων, τα οποία κυμαίνονται μεταξύ 10, 15 και 20 χαρακτηριστικών.

Ο τρίτος παράγοντας προσδιορίζει την ύπαρξη ακατάλληλων ή περιττών χαρακτηριστικών στο σύνολο των δεδομένων. Και στις δυο περιπτώσεις, κατάλληλα χαρακτηριστικά θεωρούνται εκείνα που έχουν μη μηδενικούς συντελεστές στη συνάρτηση ταξινόμησης. Αναξιόπιστα (μη κατάλληλα) χαρακτηριστικά θεωρούνται εκείνα που έχουν την ίδια απόκλιση και για τις δυο κατηγορίες και γι' αυτό, δεν χρησιμοποιούνται στην ανάπτυξη της συνάρτησης. Από την άλλη, ως περιττά χαρακτηριστικά θεωρούνται εκείνα που συσχετίζονται με τα κατάλληλα (χρήσιμα) χαρακτηριστικά, αλλά ούτε αυτά λαμβάνονται υπόψη στην ανάπτυξη της συνάρτησης ταξινόμησης.

Τέλος, ο τέταρτος παράγοντας προσδιορίζει την αναλογία των κατάλληλων χαρακτηριστικών στο σύνολο των χαρακτηριστικών. Η αναλογία αυτή κυμαίνεται μεταξύ 40% και 80%, με βήμα 20%.

Ο παρακάτω πίνακας παρουσιάζει τον τρόπο με τον οποίο χρησιμοποιούνται οι παραπάνω παράγοντες κατά την παραγωγή των τεχνητών δεδομένων.

Πίνακας 4.1: Εξεταζόμενοι παράγοντες για την παραγωγή των τεχνητών δεδομένων

Παράγοντες		Επίπεδα
F_1	Μορφή διάκρισης των κατηγοριών.	1. Γραμμική. 2. Μη γραμμική (τετραγωνική).
F_2	Πλήθος χαρακτηριστικών	1. 10 χαρακτηριστικά. 2. 15 χαρακτηριστικά. 3. 20 χαρακτηριστικά.
F_3	Μορφή ακατάλληλων χαρακτηριστικών	1.Ύπαρξη μη κατάλληλων χαρακτηριστικών. 2.Ύπαρξη περιττών χαρακτηριστικών.
F_4	Ποσοστό κατάλληλων χαρακτηριστικών σε σχέση με το σύνολο των χαρακτηριστικών	40% ή 60% ή 80%

Τέτοιες πειραματικές αναλύσεις με τεχνητά δεδομένα έχουν χρησιμοποιηθεί και σε προηγούμενες έρευνες για την αξιολόγηση της αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών. Όμως το πλήθος των εξεταζόμενων δεδομένων ήταν ιδιαίτερα περιορισμένο με αποτέλεσμα να είναι δύσκολη η εξαγωγή ασφαλών και ολοκληρωμένων συμπερασμάτων. Αντίθετα, στην παρούσα έρευνα εξετάζεται ένας μεγάλος όγκος τεχνητών δεδομένων, ώστε να είναι δυνατή η εξαγωγή πληρέστερων αποτελεσμάτων. Τα αποτελέσματα αυτά αφορούν την επίδραση που έχουν οι διάφοροι παράγοντες και οι αλληλεπιδράσεις τους στην αποτελεσματικότητα των αλγορίθμων επιλογής χαρακτηριστικών.

4.4.2 Διαδικασία παραγωγής τεχνητών δεδομένων

Βασικό σημείο αυτού του πειραματικού σχεδιασμού είναι η παραγωγή των τεχνητών δεδομένων των εναλλακτικών δραστηριοτήτων, έτσι ώστε αυτές να ακολουθούν όλες τις παραπάνω στατιστικές ιδιότητες. Τα δεδομένα σχεδιάστηκαν με τέτοιο τρόπο έτσι ώστε να ακολουθούν κανονική κατανομή με μέση τιμή μηδέν και διακύμανση μονάδα. Στην περίπτωση αυτή δεν υφίσταται καμία δυσκολία, καθώς υπάρχουν κατάλληλες γνωστές διαδικασίες.

Απώτερος στόχος της υλοποίησης αυτής της διαδικασίας είναι η παραγωγή ενός διανύσματος αποτελούμενου από n τυχαίες μεταβλητές, οι οποίες διαθέτουν τις παραπάνω στατιστικές

ιδιότητες. Στο συγκεκριμένο πειραματικό σχεδιασμό, στο διάνυσμα αυτό αντιστοιχεί το σύνολο των χαρακτηριστικών, καθένα από τα οποία θεωρείται ως μια τυχαία μεταβλητή.

Για κάθε συνδυασμό των παραγόντων F_1 έως και F_4, η παραπάνω διαδικασία χρησιμοποιείται για την παραγωγή δύο συνόλων δεδομένων. Το πρώτο χρησιμοποιείται ως δείγμα εκμάθησης και το δεύτερο ως δείγμα ελέγχου. Το μέγεθος των δειγμάτων, εκμάθησης και ελέγχου, περιλαμβάνει 500 αντικείμενα (εναλλακτικές), ενώ το πλήθος των χαρακτηριστικών καθορίζεται από τον παράγοντα F_2. Σε κάθε ένα από τα παραπάνω δείγματα εισάγεται θόρυβος στο 10% των δεδομένων. Επίσης, τόσο στο δείγμα ελέγχου, όσο και στο δείγμα εκμάθησης, ο αριθμός των αντικειμένων των δύο κατηγοριών είναι ίδιος.

Ο παραπάνω πειραματικός σχεδιασμός επαναλαμβάνεται 10 φορές για κάθε συνδυασμό των παραγόντων F_1 έως και F_4 (36 δυνατοί συνδυασμοί). Συνολικά ελέγχονται 360 διαφορετικά δείγματα εκμάθησης, τα οποία αντιστοιχούν σε ισάριθμα δείγματα ελέγχου. Για κάθε δείγμα εκμάθησης παράγεται αντίστοιχα δεδομένα (συνολικά 360 δεδομένα) από τυχαίες μεταβλητές 0 ή 1, που καθορίζει το πλήθος των χαρακτηριστικών που είναι κατάλληλα κατά τη διαδικασία της ταξινόμησης. Εκείνα τα χαρακτηριστικά που θεωρούνται ως κατάλληλα παίρνουν την τιμή 1, ενώ αυτά που θεωρούνται μη χρήσιμα ή περιττά παίρνουν την τιμή 0. Τα δεδομένα αυτά συγκρίνονται κατά τη διάρκεια της πειραματικής ανάλυσης με το σύνολο των επιλεγμένων χαρακτηριστικών που προέκυψαν από τη χρήση των αλγορίθμων επιλογής χαρακτηριστικών, έτσι ώστε να προσδιοριστεί η αναλογία εκείνων των χαρακτηριστικών που θεωρούνται κατάλληλα και επιλέχθηκαν από τους αλγόριθμους, όπως επίσης και η αναλογία εκείνων των χαρακτηριστικών που θεωρούνται ακατάλληλα ή περιττά και όμως, επιλέχθηκαν από τους αλγόριθμους.

Η υλοποίηση του πειραματικού σχεδιασμού, τόσο για τη παραγωγή των τεχνητών δεδομένων, όσο και για την μοντελοποίηση των τεχνικών ταξινόμησης και την ανάπτυξη των αλγορίθμων επιλογής χαρακτηριστικών, πραγματοποιήθηκε στο περιβάλλον του Matlab 6.0. Η ανάλυση και επεξεργασία των αποτελεσμάτων πραγματοποιήθηκε σε Excel.

4.4.3 Πραγματικά δεδομένα από το UCI ML Repository

Η συλλογή των πραγματικών δεδομένων πραγματοποιείται κατά τη δεύτερη φάση του πειράματος από τη βάση δεδομένων UCI machine learning repository. Το πλήθος των δεδομένων που συλλέχθηκαν είναι 15 στον αριθμό. Τα χαρακτηριστικά των πραγματικών

δεδομένων συνοψίζονται στον πίνακα 4.2. Όλα τα παρακάτω δεδομένα αναλύονται μέσω της διαδικασίας ελέγχου 10-fold cross validation.

Πίνακας 4.2: Πραγματικά δεδομένα από τη βάση δεδομένων UCI machine learning repository

Σύνολο δεδομένων	Αντικείμενα	Χαρακτηριστικά
Bupa liver disorders	350	6
Hepatitis	160	19
Credit screening	690	14
Ionosphere	350	33
Mushroom	8120	6
Pima Indians Diabetes	770	8
Tic tac toe	960	9
Thyroid	2800	26
Breast cancer Wisconsin	570	30
Voting	440	16
German credit	1000	20
Heart disease	270	13
Monks-1	550	6
Monks-2	600	6
Monks-3	550	6

Η ανάλυση αυτού του δεύτερου σταδίου συμπληρώνει την πειραματική ανάλυση της πρώτης φάσης με στόχο τη διερεύνηση της πρακτικής ισχύος των συμπερασμάτων που προκύπτουν από την ανάλυση των τεχνητών δεδομένων. Ταυτόχρονα, η ανάλυση αυτή επιτρέπει τη σύγκριση των αποτελεσμάτων με αυτά προηγούμενων ερευνών.

4.5 Ανάλυση αποτελεσμάτων

4.5.1 Τεχνητά δεδομένα

Η πειραματική ανάλυση των τεχνητών δεδομένων επιτρέπει την εκτίμηση της αποτελεσματικότητας των υποδειγμάτων ταξινόμησης, από την χρήση των αλγορίθμων επιλογής χαρακτηριστικών, συναρτήσει των παραγόντων που λήφθηκαν υπόψη κατά την κατασκευή των δεδομένων αυτών.

Οι παράγοντες της ύπαρξης ακατάλληλων ή περιττών χαρακτηριστικών (F_3) και το ποσοστό των πλέον αξιόπιστων χαρακτηριστικών για την ανάπτυξη ενός ικανοποιητικού υποδείγματος ταξινόμησης (F_4) οδήγησαν σε υψηλό λόγο ακρίβειας ταξινόμησης. Στον πίνακα 4.3 συνοψίζονται τα αντίστοιχα αποτελέσματα για κάθε FSA. Τα αποτελέσματα που παρουσιάζονται αφορούν τους λόγους ακρίβειας των υποδειγμάτων που αναπτύσσονται με την εφαρμογή των FSAs σε σχέση με τους λόγους ακρίβειας των υποδειγμάτων που αναπτύσσονται με το πλήρες σύνολο χαρακτηριστικών. Ο μέσος όρος των πειραμάτων που απέδωσαν καλύτερα αποτελέσματα με την χρήση των FSAs δίνονται για κάθε FSA εντός των παρενθέσεων.

Βάσει των αποτελεσμάτων φαίνεται ότι η αποτελεσματικότητα των FSAs μειώνεται όταν στο δείγμα δεδομένων υπάρχουν περιττά χαρακτηριστικά, σε αντίθεση βέβαια, με την ύπαρξη μη κατάλληλων χαρακτηριστικών. Κατά μέσο όρο, περίπου στο 54% των περιπτώσεων που περιλαμβάνουν μη αξιόπιστα χαρακτηριστικά τα υποδείγματα που αναπτύσσονται με την χρήση των FSA υπερτερούν έναντι αυτών που δεν εφαρμόζουν διαδικασίες επιλογής χαρακτηριστικών. Αντιθέτως, περίπου στο 31% των περιπτώσεων που περιλαμβάνουν περιττά χαρακτηριστικά τα υποδείγματα που αναπτύσσονται με την χρήση των FSA υπερτερούν έναντι αυτών που δεν εφαρμόζουν διαδικασίες επιλογής χαρακτηριστικών. Από τα παραπάνω, διαπιστώνεται ότι τα FSAs είναι πιο αποτελεσματικά στην αναγνώριση ακατάλληλων χαρακτηριστικών από ότι περιττών. Επιπλέον, οι FSAs είναι πιο αποτελεσματικοί στις περιπτώσεις όπου τα κατάλληλα χαρακτηριστικά αποτελούν ένα μικρό μέρος του συνόλου των χαρακτηριστικών. Όσο αυξάνεται το ποσοστό του συνόλου των κατάλληλων χαρακτηριστικών στο σύνολο του δείγματος τόσο μειώνεται ο λόγος ακρίβειας των αναπτυσσόμενων υποδειγμάτων με τη χρήση των FSAs. Ειδικότερα, οι FSAs είναι αποτελεσματικοί περίπου στο 65% των περιπτώσεων όπου μόνο το 40% των χαρακτηριστικών είναι κατάλληλα, ενώ μόνο στο 26% των περιπτώσεων με ποσοστό 80% των χρήσιμων χαρακτηριστικών, οι FSAs δίνουν ικανοποιητικά αποτελέσματα.

Γενικότερα, τα FSAs wrapper φαίνεται να δίνουν καλύτερα αποτελέσματα από ότι τα FSAs filter, αλλά δεν είναι πάντα αποτελεσματικά. Όπως φαίνεται στον πίνακα 4.3 ο αλγόριθμος LVW είναι πιο αποτελεσματικός στον εντοπισμό των περιττών χαρακτηριστικών από ότι των μη κατάλληλων, καθώς στο 50% των περιπτώσεων σημειώνεται υψηλός λόγος ακρίβειας ταξινόμησης που κυμαίνεται περίπου στο 99%. Αξιοσημείωτο, επίσης, είναι ότι όταν το ποσοστό των κατάλληλων χαρακτηριστικών είναι σχετικά μικρό, ο αλγόριθμος LVW είναι αρκετά αποτελεσματικός με μέσο λόγο ακρίβειας περίπου στο 100%, ενώ μικρές διαφοροποιήσεις παρουσιάζονται καθώς ο λόγος αυτός αυξάνεται. Παρατηρείται δηλαδή ότι όταν το ποσοστό κατάλληλων χαρακτηριστικών κυμαίνεται από 60% και πάνω, ο λόγος της ακρίβειας ταξινόμησης διατηρείται σε σταθερό επίπεδο περίπου στο 98% για το 33% του συνόλου των περιπτώσεων. Ανάλογες διακυμάνσεις παρουσιάζονται και για τους δύο άλλους αλγόριθμους wrapper, SFGW και SBGW. Αυτό που θα πρέπει να τονιστεί είναι ότι ο αλγόριθμος SBGW στο 100% των περιπτώσεων που περιλαμβάνουν ποσοστό 60% κατάλληλων χαρακτηριστικών παρουσιάζουν υψηλό λόγο ακρίβειας ταξινόμησης που έγκειται στο 100%.

Σε αντίθεση με τους αλγορίθμους wrapper, οι αλγόριθμοι filter δεν παρουσιάζουν κάτι το σημαντικό. Όλοι οι αλγόριθμοι αυτής της κατηγορίας παρουσιάζουν παρόμοια αποτελέσματα με μικρές διαφοροποιήσεις. Συνεπώς, οι αλγόριθμοι filter είναι αποτελεσματικοί όταν τα δεδομένα περιλαμβάνουν ποσοστό 40% των χρήσιμων χαρακτηριστικών και λειτουργούν καλύτερα στον εντοπισμό κατάλληλων χαρακτηριστικών από ότι περιττών.

Πίνακας 4.3: Λόγοι ακρίβειας ταξινόμησης που προκύπτουν από την χρήση των αλγορίθμων επιλογής χαρακτηριστικών έναντι της μη εφαρμογής αυτών συναρτήσει των παραγόντων F_3 και F_4 (τεχνητά δεδομένα).

	F_3		F_4		
	Ακατάλληλα χαρακτηριστικά	Περιττά χαρακτηριστικά	40%	60%	80%
LVF	101.96 (66.7)	99.07 (16.7)	101.74 (50.0)	100.16 (33.3)	99.31 (16.7)
LVI	101.75 (66.7)	99.30 (16.7)	101.73 (66.7)	100.19 (33.3)	99.37 (16.7)
FOCUS	101.05 (66.7)	100.03 (33.3)	100.74 (50.0)	100.52 (50.0)	100.23 (33.3)
SBGF	103.46 (66.7)	98.73 (16.7)	102.92 (66.7)	100.68 (33.3)	99.15 (16.7)
SFGF	103.69 (66.7)	99.74 (16.7)	103.25 (83.3)	101.33 (66.7)	100.10 (33.3)
RELIEF	94.48 (16.7)	98.12 (0.0)	97.55 (33.3)	95.61 (0.0)	95.45 (0.0)
LVW	98.83 (0.0)	99.54 (50.0)	100.04 (66.7)	99.10 (33.3)	98.42 (33.3)
SBGW	101.22 (83.3)	100.66 (66.7)	101.60 (83.3)	100.85 (100.0)	100.30 (50.0)
SFGW	100.89 (50.0)	100.70 (66.7)	102.22 (83.3)	100.54 (66.7)	99.51 (33.3)
Μέσος όρος	100.81 (53.7)	99.54 (31.5)	101.31 (64.8)	99.89 (46.3)	99.09 (25.9)

Οι παράγοντες της μορφής διάκρισης των κατηγοριών (F_1) και το πλήθος των χαρακτηριστικών (F_2) οδήγησαν σε αρκετά υψηλούς λόγους ακρίβειας ταξινόμησης με μικρές διαφοροποιήσεις. Στον πίνακα 4.4 συνοψίζονται τα αντίστοιχα αποτελέσματα για κάθε FSA. Και σε αυτή την περίπτωση, τα αποτελέσματα που παρουσιάζονται αφορούν τους λόγους ακρίβειας των υποδειγμάτων που αναπτύσσονται με την εφαρμογή των FSAs σε σχέση με τους λόγους ακρίβειας που αναπτύσσονται με το πλήρες σύνολο χαρακτηριστικών. Ο μέσος όρος των πειραμάτων που απέδωσαν καλύτερα αποτελέσματα με την χρήση των FSAs δίνονται για κάθε FSA εντός των παρενθέσεων. Από τους μέσους όρους των αποτελεσμάτων διαπιστώνεται ότι όλοι οι αλγόριθμοι σημειώνουν ικανοποιητικά αποτελέσματα στις περισσότερες των περιπτώσεων όσον αφορά το λόγο ακρίβειας ταξινόμησης και ιδιαίτερα, σε περιπτώσεις που οι κατηγορίες διακρίνονται γραμμικά, ενώ το πλήθος των χαρακτηριστικών είναι μικρό. Αντιθέτως, όταν αυξάνεται το πλήθος των χαρακτηριστικών ή οι κατηγορίες διακρίνονται μη γραμμικά, ο λόγος ακρίβειας ταξινόμησης των αναπτυσσόμενων υποδειγμάτων με τη χρήση των FSAs παρουσιάζει μικρή μείωση, αλλά όχι ιδιαίτερα αισθητή. Και σε αυτή την περίπτωση θα πρέπει να τονιστεί ότι ο αλγόριθμος SBGW παρουσιάζει σταθερά αποτελέσματα, ανεξαρτήτως του πλήθους των χαρακτηριστικών που υπάρχουν, στο 83% των περιπτώσεων με μέσο λόγο ακρίβειας 101%.

Πίνακας 4.4: Λόγοι ακρίβειας ταξινόμησης που προκύπτουν από την χρήση των αλγορίθμων επιλογής χαρακτηριστικών έναντι της μη εφαρμογής αυτών συναρτήσει των παραγόντων F_1 και F_2 (τεχνητά δεδομένα).

	F_1		F_2		
	Μορφή διάκρισης των κατηγοριών		Πλήθος χαρακτηριστικών		
	Γραμμική	Μη γραμμική (τετραγωνική)	10	15	20
LVF	100.54% (33.33%)	100.29% (33.33%)	100.55% (33.33%)	100.44% (33.33%)	100.25% (16.67%)
LVI	100.53% (33.33%)	100.35% (33.33%)	100.16% (33.33%)	100.98% (33.33%)	100.19% (16.67%)
FOCUS	100.61% (50.0%)	100.38% (33.33%)	100.45% (66.67%)	100.64% (50.0%)	100.41% (33.33%)
SBGF	101.10% (50.0%)	100.76% (33.33%)	100.43% (33.33%)	100.78% (50.0%)	101.65% (50.0%)
SFGF	101.94% (66.67%)	101.76% (33.33%)	101.11% (66.67%)	101.59% (50.0%)	102.06% (66.67%)
RELIEF	96.27% (33.33%)	96.18% (0.0%)	95.49% (0.0%)	96.12% (16.67%)	97.12% (33.33%)
LVW	99.64% (16.67%)	98.74% (33.33%)	99.22% (33.33%)	99.28% (33.33%)	99.08% (33.33%)
SBGW	100.96% (83.33%)	100.89% (83.33%)	100.49% (83.33%)	101.11% (83.33%)	101.20% (83.33%)
SFGW	100.90% (33.33%)	100.64% (66.67%)	100.42% (83.33%)	101.14% (66.67%)	100.78% (33.33%)
Μέσος όρος	100.28% (44.44%)	100.0% (38.89%)	98.81% (48.15%)	100.23% (46.30%)	100.30% (40.74%)

Εκτός από την εκτίμηση της αποτελεσματικότητας των FSAs στο λόγο ακρίβειας ταξινόμησης, εξετάζεται η ικανότητά τους στον εντοπισμό των πλέων κατάλληλων χαρακτηριστικών. Αυτό πραγματοποιείται βάσει δύο κριτηρίων. Το πρώτο κριτήριο αφορά το πλήθος των κατάλληλων χαρακτηριστικών που επιλέγεται, εκφρασμένο σε ποσοστό επί του συνόλου των κατάλληλων χαρακτηριστικών. Αντίστοιχα, το δεύτερο κριτήριο αφορά το πλήθος των ακατάλληλων/ περιττών χαρακτηριστικών που επιλέγονται, εκφρασμένο σε ποσοστό επί του συνόλου των ακατάλληλων/ περιττών χαρακτηριστικών στο δείγμα. Η διαφορά μεταξύ των δύο κριτηρίων παρουσιάζεται με τον υπολογισμό ενός βαθμού επιτυχίας (hit rate) για κάθε FSA, ο οποίος παίρνει τιμές στο διάστημα [-100%, 100%]. Στην περίπτωση που το hit rate πάρει υψηλή θετική τιμή, αυτό σημαίνει ότι ένας FSA επιλέγει μόνο κατάλληλα χαρακτηριστικά και όχι περιττά ή μη κατάλληλα. Στην αντίθετη περίπτωση, όταν το hit rate πάρει τιμή κοντά στο -100%, αυτό σημαίνει ότι ένας FSA επιλέγει μόνο περιττά ή μη κατάλληλα χαρακτηριστικά και όχι κατάλληλα.

Στον πίνακα 4.5 δίνονται τα ποσοστά επιτυχίας (hit rates) των FSAs για τους τέσσερις παράγοντες που χρησιμοποιούνται στην πειραματική ανάλυση. Από τα αποτελέσματα φαίνεται ότι όλοι οι παράγοντες παρουσιάζουν σημαντικές επιπτώσεις στα τελικά αποτελέσματα. Μεγαλύτερες διαφορές παρατηρούνται για τους παράγοντες F_3 και F_4. Συγκεκριμένα, τα FSAs είναι περισσότερο αποτελεσματικά στο να διακρίνουν τα κατάλληλα από τα μη κατάλληλα χαρακτηριστικά και λιγότερο τα κατάλληλα από τα περιττά χαρακτηριστικά. Αυτό είναι σύμφωνο με τα αποτελέσματα που περιγράφονται προηγούμενα σχετικά με το λόγο ακρίβειας ταξινόμησης από την εφαρμογή των FSAs. Όταν, όμως, η διαδικασία ταξινόμησης γίνεται πιο περίπλοκη (μη γραμμική μορφή διάκρισης κατηγοριών, μεγάλο πλήθος χαρακτηριστικών, μικρό ποσοστό κατάλληλων χαρακτηριστικών), τότε οι τιμές των hit rates μειώνονται.

Αξιοσημείωτο είναι το γεγονός ότι, τα hit rates κατά την εφαρμογή των FSAs filter σημειώνουν υψηλότερους λόγους από ότι τα FSAs wrapper. Από τα αποτελέσματα φαίνεται ότι οι αλγόριθμοι FOCUS και RELIEF είναι αυτοί που επιλέγουν περισσότερα ακατάλληλα ή περιττά χαρακτηριστικά, όταν κατά τη διαδικασία επιλογής αυτών υφίσταται μικρό ποσοστό κατάλληλων χαρακτηριστικών. Συγκεκριμένα, σημειώνουν αρνητικές τιμές, -5.9% και -0.5% για τα hit rates, αντίστοιχα. Ο FOCUS, βέβαια, δεν παρουσιάζει ιδιαίτερα προβλήματα στην επιλογή των πλέον κατάλληλων χαρακτηριστικών συναρτήσει των άλλων παραγόντων, σε αντίθεση με τον RELIEF, του οποίου η παραπάνω παρατήρηση δεν είναι το μόνο μειονέκτημα, καθώς οι τιμές των hit rates δεν είναι ιδιαίτερα υψηλές και για τους άλλους

παράγοντες. Γενικά, ο αλγόριθμος αυτός παρουσιάζει προβλήματα όσον αφορά τον εντοπισμό των χρήσιμων χαρακτηριστικών. Οι υπόλοιποι αλγόριθμοι παρουσιάζουν μικρές διαφοροποιήσεις στους διάφορους παράγοντες. Τέλος, οι αλγόριθμοι SBGF και SFGF είναι σαφώς πιο αποτελεσματικοί από τους υπόλοιπους, καθώς σημειώνουν υψηλά ποσοστά επιλογής κατάλληλων χαρακτηριστικών για όλους τους παράγοντες που λαμβάνονται υπόψη κατά το σχεδιασμό των τεχνητών δεδομένων.

Πίνακας 4.5: Βαθμός επιτυχίας (hit rate) κατά τη διαδικασία επιλογής κατάλληλων χαρακτηριστικών με τη χρήση των FSAs συναρτήσει των προκαθορισμένων παραγόντων (τεχνητά δεδομένα).

	F_1		F_2			F_3		F_4		
	1	2	1	2	3	1	2	1	2	3
LVF	31.7%	29.3%	38.9%	28.6%	24.0%	40.9%	20.1%	15.2%	29.8%	46.5%
LVI	31.8%	26.7%	35.8%	28.6%	23.5%	38.5%	20.1%	8.8%	29.5%	49.6%
FOCUS	30.0%	25.2%	30.9%	27.0%	25.0%	31.1%	24.1%	-5.9%	28.5%	60.3%
SBGF	40.9%	33.1%	42.4%	35.2%	33.5%	51.0%	23.0%	26.8%	36.8%	47.5%
SFGF	40.2%	33.5%	42.7%	34.7%	33.2%	51.0%	22.7%	24.3%	37.3%	49.0%
RELIEF	18.2%	16.4%	18.9%	17.4%	15.5%	18.6%	16.0%	-0.5%	17.5%	35.0%
LVW	22.7%	17.8%	26.6%	18.8%	15.4%	25.5%	14.9%	4.8%	19.9%	36.0%
SBGW	29.7%	27.9%	33.4%	28.4%	24.7%	33.3%	24.3%	4.1%	28.7%	53.7%
SFGW	30.6%	24.5%	35.4%	26.5%	20.7%	35.8%	19.2%	15.0%	27.3%	40.2%
Μέσος όρος	30.6%	26.1%	33.9%	27.2%	23.9%	36.2%	20.5%	10.3%	28.4%	46.4%

Στον πίνακα 4.6 παρουσιάζονται κάποια επιπρόσθετα αποτελέσματα των hit rates για τα FSAs wrapper σε συνδυασμό με τις επιλεγμένες τεχνικές ταξινόμησης. Βάσει των αποτελεσμάτων φαίνεται ότι ο αλγόριθμος LVW παρουσιάζει χειρότερα αποτελέσματα σε σχέση με αυτά που προκύπτουν από τους αλγόριθμους SBGW και SFGW. Στις περισσότερες των περιπτώσεων αποδεικνύεται ότι η χρήση του αλγορίθμου SBGW, ακολουθούμενος από τον αλγόριθμο SFGW, οδηγεί σε αισθητά καλύτερα αποτελέσματα. Αυτό που θα πρέπει να τονιστεί είναι ότι

οι παρακάνω αλγόριθμοι συνεργάζονται πολύ καλά με την τεχνική ταξινόμησης SVM και λιγότερο καλά με την CART.

Πίνακας 4.6: Βαθμός επιτυχίας (hit rate) κατά τη διαδικασία επιλογής κατάλληλων χαρακτηριστικών με τη χρήση των FSAs wrapper (τεχνητά δεδομένα).

	CART	LDA	PNN	NN	SVM	LR
LVW	13.30%	22.33%	11.90%	23.55%	27.43%	22.95%
SBGW	23.19%	30.59%	29.61%	28.56%	29.78%	31.19%
SFGW	15.88%	29.27%	24.22%	27.36%	38.38%	30.00%

Στον πίνακα 4.7 παρουσιάζονται τα αποτελέσματα των hit rates για τα FSAs filter. Εδώ θα πρέπει να σημειωθεί ότι ένα επιλεγμένο σύνολο χαρακτηριστικών εφαρμόζεται αυτομάτως σε όλες τις τεχνικές ταξινόμησης, εφόσον όπως και προαναφέρθηκε, κατά τις διαδικασίες filter οι τεχνικές ταξινόμησης δεν χρησιμοποιούνται ως κριτήρια για την επιλογή αυτών. Βάσει των αποτελεσμάτων φαίνεται ότι ο αλγόριθμος SBGF, ακολουθούμενος από τον αλγόριθμο SFGF, είναι πιο αποτελεσματικός στον εντοπισμό των χρήσιμων χαρακτηριστικών. Ο λιγότερο αποτελεσματικός με μέσο ποσοστό 17.30% θεωρείται ο RELIEF. Οι υπόλοιποι αλγόριθμοι σημειώνουν ικανοποιητικά αποτελέσματα με μικρές διαφοροποιήσεις μεταξύ τους.

Πίνακας 4.7: Βαθμός επιτυχίας (hit rate) κατά τη διαδικασία επιλογής κατάλληλων χαρακτηριστικών με τη χρήση των FSAs filter (τεχνητά δεδομένα).

	CART – LDA – PNN – NN – SVM – LR
LVF	30.50%
LVI	29.27%
FOCUS	27.63%
SBGF	37.01%
SFGF	36.87%
RELIEF	17.30%

4.5.2 Δεδομένα του UCI ML Repository

Αρχικά, η πειραματική ανάλυση επικεντρώνεται στην εκτίμηση του λόγου ακρίβειας των υποδειγμάτων ταξινόμησης που αναπτύσσονται με βάσει τα επιλεγμένα χαρακτηριστικά που προκύπτουν από τους αλγόριθμους επιλογής χαρακτηριστικών σε σύγκριση με εκείνα τα υποδείγματα που αναπτύσσονται με βάσει το σύνολο των χαρακτηριστικών. Στον πίνακα 4.8 δίνονται οι μέσες τιμές, σε ποσοστά επί τοις εκατό, του λόγου ακρίβειας των αναπτυσσόμενων μοντέλων από τα FSAs έναντι της ακρίβειας των τεχνικών ταξινόμησης χωρίς την χρήση των FSAs. Οι τιμές που παρουσιάζονται εντός των παρενθέσεων δείχνουν τον αριθμό των δεδομένων των οποίων ο μέσος όρος του λόγου ακρίβειας ταξινόμησης των 10-fold cross validation ξεπερνάει το αντίστοιχο λόγο ακρίβειας των υποδειγμάτων που χρησιμοποιούν όλη τη διαθέσιμη πληροφορία (αρχικό σύνολο χαρακτηριστικών) κατά τη διαδικασία της ταξινόμησης των δεδομένων.

Από τις τιμές των αποτελεσμάτων φαίνεται ότι οι τεχνικές CART και SVM πλεονεκτούν έναντι των υπολοίπων τεχνικών, σημειώνοντας το μεγαλύτερο λόγο ακρίβειας ταξινόμησης. Γενικότερα, παρατηρείται ότι η SVM συνδυασμένη με οποιοδήποτε FSA, αποδίδει υψηλό λόγο ακρίβειας στην πλειοψηφία του συνόλου των δεδομένων της ανάλυσης. Χαρακτηριστικό παράδειγμα αποτελεί ο αλγόριθμος επιλογής χαρακτηριστικών SFGW, ο οποίος βελτιώνει το λόγο ακρίβειας του μοντέλου SVM σε 14 από τα 15 δεδομένα. Παρόμοια αποτελέσματα προκύπτουν και κατά την εφαρμογή της τεχνικής ταξινόμησης CART. Σε αυτή την περίπτωση βέβαια, δεν προκύπτουν ικανοποιητικά αποτελέσματα από όλους τους FSAs. Χαρακτηριστικό παράδειγμα αποτελεί η χρήση του αλγορίθμου LVI, ο οποίος αν εφαρμοστεί στη τεχνική CART οδηγεί σε χαμηλούς λόγους ακρίβειας σε 12 από τα 15 δεδομένα.

Όσον αφορά τα υπόλοιπα υποδείγματα ταξινόμησης, οι λόγοι ακρίβειας που προκύπτουν από τα FSAs είναι ελαφρώς υποδεέστεροι από αυτούς που προκύπτουν χωρίς την χρήση των FSAs, αλλά οι διαφορές, σε πολλές περιπτώσεις, είναι ελάχιστες. Συγκεκριμένα, οι μέσοι λόγοι ακρίβειας από την χρήση των FSAs έναντι της μη χρήσης αυτών είναι 99.1% για την LDA, 98.6% για την LR, 95.4% για τον NN και 95.1% για τα PNN. Αξιοσημείωτο είναι το γεγονός ότι, κατά μέσο όρο, οι αλγόριθμοι επιλογής χαρακτηριστικών που ακολουθούν διαδικασίες wrapper υπερτερούν έναντι των αλγορίθμων που ακολουθούν διαδικασίες filter. Συγκεκριμένα, ο αλγόριθμος SFGW βελτιώνει το λόγο ακρίβειας των αναπτυσσόμενων υποδειγμάτων ταξινόμησης, κατά μέσο όρο σε 10 από τα 15 δεδομένα.

Όπως και παρατηρήθηκε, από όλες τις εξεταζόμενες τεχνικές εκείνη που μειονεκτεί περισσότερο έναντι των υπολοίπων είναι η τεχνική ταξινόμησης PNN, η οποία παρουσιάζει περιθώρια βελτίωσης μόνο εάν συνδυαστεί με τον αλγόριθμο SFGW αποδίδοντας υψηλούς λόγους ακρίβειας σε 10 από τα δεδομένα που εξετάστηκαν. Ανάλογοι συνδυασμοί μπορούν να πραγματοποιηθούν και για τα υπόλοιπα υποδείγματα ταξινόμησης προκειμένου να βελτιωθεί η ακρίβεια των αντίστοιχων υποδειγμάτων. Για παράδειγμα, οι συνδυασμοί που θα ήταν δυνατόν να πραγματοποιηθούν, έτσι ώστε να υπάρξει βελτίωση στο λόγο ακρίβειας των υποδειγμάτων ταξινόμησης είναι: CART με SBGW, LDA και PNN με SFGW, NN με SBGW ή SFGW, SVM με SFGW και LR με LVI ή SFGW.

Πίνακας 4.8: Λόγοι ακρίβειας ταξινόμησης που προκύπτουν από την χρήση των αλγορίθμων επιλογής χαρακτηριστικών έναντι της μη εφαρμογής αυτών (πραγματικά δεδομένα).

	CART	LDA	PNN	NN	SVM	LR
LVF	98.7 (7)	99.6 (5)	91.5 (5)	92.8 (6)	105.6 (9)	98.8 (6)
LVI	98.8 (3)	100.2 (6)	94.4 (6)	94.3 (7)	104.9 (8)	99.7 (8)
FOCUS	100.1 (7)	99.6 (7)	96.9 (8)	97.1 (7)	101.5 (8)	99.3 (6)
SBGF	98.9 (5)	98.9 (6)	91.4 (7)	91.7 (6)	106.8 (7)	98.7 (5)
SFGF	98.7 (5)	99.5 (6)	92.5 (6)	92.5 (6)	107.1 (10)	99.1 (5)
RELIEF	93.8 (6)	95.0 (4)	90.6 (5)	86.5 (3)	101.8 (8)	93.2 (4)
LVW	102.0 (9)	99.6 (5)	95.8 (5)	97.0 (5)	107.4 (12)	99.7 (6)
SBGW	101.9 (10)	100.0 (5)	100.4 (5)	103.8 (9)	106.2 (10)	99.2 (5)
SFGW	101.5 (7)	99.7 (9)	102.1 (10)	102.9 (9)	111.6 (14)	99.4 (8)
Μέσος όρος	99.4	99.1	95.1	95.4	105.9	98.6

Σε δεύτερη φάση, η πειραματική ανάλυση των πραγματικών δεδομένων εστιάζεται στην εκτίμηση του πλήθους των χαρακτηριστικών που επιλέγονται από κάθε FSA και τον υπολογιστικό φόρτο που απαιτείται για την εφαρμογή των αλγορίθμων αυτών. Ο πίνακας 4.9 δείχνει κατά μέσο όρο το πλήθος των επιλεγμένων χαρακτηριστικών ως ποσοστό του συνόλου των χαρακτηριστικών, καθώς επίσης και τις μέσες τιμές των CPU χρόνων σε δευτερόλεπτα για τους έξι αλγόριθμους επιλογής χαρακτηριστικών που ακολουθούν διαδικασίες filter.

Παρόμοια αποτελέσματα παρουσιάζονται και στον πίνακα 4.10 για τα FSAs που ανήκουν στις διαδικασίες wrapper, όπου και παρατηρείται ότι οι τιμές των αποτελεσμάτων διαφέρουν ανάλογα με την τεχνική ταξινόμησης στην οποία το κάθε FSA εφαρμόζεται.

Πίνακας 4.9: Ποσοστό του πλήθους των επιλεγμένων χαρακτηριστικών και των χρόνων CPU των αλγορίθμων επιλογής χαρακτηριστικών που ακολουθούν διαδικασίες filter (πραγματικά δεδομένα).

	Ποσοστό του πλήθους των επιλεγμένων χαρακτηριστικών	CPU time
LVF	49.16	1.89
LVI	61.10	7.79
FOCUS	78.48	0.36
SBGF	43.79	0.48
SFGF	44.46	0.17
RELIEF	41.76	2.34

Πίνακας 4.10: Ποσοστό του πλήθους των επιλεγμένων χαρακτηριστικών και των χρόνων CPU των αλγορίθμων επιλογής χαρακτηριστικών που ακολουθούν διαδικασίες wrapper (πραγματικά δεδομένα)

	Ποσοστό του πλήθους των επιλεγμένων χαρακτηριστικών			CPU time		
	LVW	SBGW	SFGW	LVW	SBGW	SFGW
CART	36.09	55.17	39.22	15.9	15.7	30.8
LDA	50.87	62.08	42.23	1.8	3.7	2.8
PNN	44.39	65.57	53.99	19.5	69.7	61.2
NN	47.42	69.91	52.20	11.2	26.6	22.3
SVM	22.29	73.53	34.07	5.4	15.7	9.1
LR	44.53	69.22	43.47	19.2	109.1	46.4
Μέσος όρος	40.93	65.91	44.20	12.2	40.1	28.8

Βάσει των τιμών που αναγράφονται στους δύο παραπάνω πίνακες, φαίνεται ότι οι αλγόριθμοι LVI, FOCUS και SBGW είναι λιγότερο αποτελεσματικοί όσον αφορά το πλήθος των επιλεγμένων χαρακτηριστικών. Αντιθέτως, οι υπόλοιποι αλγόριθμοι επιλογής χαρακτηριστικών παρουσιάζουν σαφώς καλύτερα αποτελέσματα, καθώς ο μέσος αριθμός των επιλεγμένων χαρακτηριστικών είναι μικρότερος από το 50% του συνόλου των χαρακτηριστικών. Όσον αφορά, τον υπολογιστικό φόρτο που απαιτείται για την εφαρμογή των FSAs, οι αλγόριθμοι SBGF και SFGF, όπως επίσης και ο αλγόριθμος FOCUS υπερτερούν έναντι των υπολοίπων αλγορίθμων.

Στους FSAs wrapper, όπως αναμένονταν, ο υπολογιστικός φόρτος στην επιλογή των χαρακτηριστικών ήταν αισθητά πιο μεγάλος και ειδικότερα, σε εκείνες τις περιπτώσεις όπου τα σύνολα των δεδομένων περιείχαν πολλά χαρακτηριστικά. Παρόλα αυτά, όπως παρατηρήθηκε προηγούμενα, ο αυξημένος υπολογιστικός φόρτος που παρατηρήθηκε κατά την εφαρμογή των FSAs wrapper αντισταθμίζεται από τις μεγάλες τιμές των λόγων ακρίβειας ταξινόμησης που σημειώθηκαν. Τέλος, θα πρέπει να τονιστεί ότι οι μεγαλύτερες και περισσότερο σημαντικές μειώσεις στο σύνολο των επιλεγμένων χαρακτηριστικών παρατηρήθηκαν για τις τεχνικές ταξινόμησης CART και SVM στις οποίες και εφαρμόστηκαν οι αλγόριθμοι επιλογής χαρακτηριστικών LVW και SFGW. Για τον ίδιο συνδυασμό τεχνικών ταξινόμησης και FSAs wrapper σημειώθηκαν, επίσης, υψηλοί λόγοι ακρίβειας, όπως και φαίνεται στον πίνακα 4.8, καθώς επίσης, απαιτήθηκε μικρός υπολογιστικός φόρτος για την επιλογή των χαρακτηριστικών.

Είναι επίσης, σημαντικό να ελεγχθούν οι ομοιότητες των επιλεγμένων συνόλων χαρακτηριστικών από κάθε FSAs. Στον πίνακα 4.11 δίνονται κάποια ενδεικτικά αποτελέσματα της παραπάνω επισήμανσης. Για κάθε ζεύγος FSAs, a και b, ο πίνακα 4.11 παρουσιάζει τον δείκτη συμφωνίας $R(a,\ b)$, ο οποίος υπολογίζεται ως ακολούθως:

$$R(a,\ b)=\frac{1}{n}\sum_{j=1}^{n} I\left(x_j^a,\ x_j^b\right)$$

Όπου $I\left(x_j^a,\ x_j^b\right)$ είναι ένας δείκτης, ο οποίος παίρνει την τιμή 1, εάν και οι δύο αλγόριθμοι a και b συμφωνούν στην επιλογή ή όχι ενός χαρακτηριστικού x_j και την τιμή -1, εάν οι δύο αλγόριθμοι διαφωνούν όσον αφορά την καταλληλότητα του χαρακτηριστικού x_j. Εάν το σύνολο των επιλεγμένων χαρακτηριστικών και για τους δύο αλγόριθμους συμπίπτει τότε ο

δείκτης συμφωνίας $R(a, b)$ παίρνει την τιμή 1. Στην αντίθετη περίπτωση εάν υπάρχει απόλυτη διαφωνία μεταξύ των δύο αλγορίθμων όσον αφορά την επιλογή των χαρακτηριστικών, τότε ο δείκτης συμφωνίας $R(a, b)$ παίρνει την τιμή -1.

Τα αποτελέσματα που παρουσιάζονται στον πίνακα 4.11 είναι ο μέσος όρος των δεικτών συμφωνίας που προέκυψαν από τα 10-fold cross validation για κάθε σύνολο δεδομένων. Βάσει των αποτελεσμάτων του πίνακα 4.11 είναι εμφανές ότι υπάρχουν σημαντικές διαφορές στα σύνολα των χαρακτηριστικών που επιλέγονται από κάθε FSA. Μεγαλύτερη συμφωνία στην επιλογή χαρακτηριστικών σημειώνεται μεταξύ των αλγορίθμων SBGF και SFGF με δείκτη συμφωνίας 0.83, όπως επίσης και του LVF και των δύο προηγούμενων αλγορίθμων. Επιπλέον, συμφωνία σε πάνω από το 50% των επιλεγμένων χαρακτηριστικών παρατηρείται μεταξύ των αλγορίθμων LVF – LVI, LVF – SBGF και LVF – SFGF, με τιμές του δείκτη συμφωνίας 0.51, 0.68 και 0.63, αντίστοιχα. Στις περισσότερες των περιπτώσεων, παρατηρούνται σημαντικές διαφορές στα σύνολα των επιλεγμένων χαρακτηριστικών από την χρήση των εξεταζόμενων FSAs. Επιπλέον, υπάρχουν περιπτώσεις με διαφωνία στην επιλογή χαρακτηριστικών. Αυτό αποδεικνύεται από την ύπαρξη αρνητικών τιμών του δείκτη συμφωνίας. Η διαφωνία αυτή έγκειται στις διαφορετικές στρατηγικές διερεύνησης των FSAs. Χαρακτηριστικά παραδείγματα διαφωνίας σημειώνουν οι αλγόριθμοι FOCUS με τους RELIEF ή SFGW ή LVW, ο LVI με τον LVW όταν χρησιμοποιείται στην τεχνική ταξινόμησης SVM και ο LVW με τον SBGW, ιδιαίτερα όταν ο δεύτερος χρησιμοποιείται για την ανάπτυξη υποδειγμάτων στις τεχνικές SVM και LR.

Πίνακας 4.11: Συμφωνία των FSAs στην επιλογή χαρακτηριστικών (πραγματικά δεδομένα)

		Filter FSAs					LVW						SBGW						SFGW					
		LVI	FOCUS	SBGF	SFGF	RELIEF	CART	LDA	PNN	NN	SVM	LR	CART	LDA	PNN	NN	SVM	LR	CART	LDA	PNN	NN	SVM	LR
Filter FSAs	LVF	0.51	0.31	0.68	0.63	0.21	0.16	0.23	0.14	0.09	0.21	0.27	0.12	0.30	0.18	0.14	0.22	0.27	0.27	0.30	0.14	0.15	0.24	0.29
	LVI		0.42	0.47	0.45	0.14	0.01	0.20	0.04	0.01	-0.02	0.15	0.20	0.31	0.19	0.19	0.33	0.34	0.11	0.20	0.10	0.10	0.08	0.19
	FOCUS			0.21	0.23	-0.02	-0.17	0.13	-0.07	-0.04	-0.30	0.01	0.14	0.30	0.27	0.30	0.42	0.39	-0.10	0.01	0.07	0.05	-0.15	0.03
	SBGF				0.83	0.23	0.28	0.23	0.20	0.16	0.32	0.36	0.17	0.31	0.16	0.07	0.11	0.26	0.37	0.38	0.17	0.22	0.36	0.37
	SFGF					0.24	0.31	0.25	0.22	0.18	0.30	0.35	0.19	0.29	0.18	0.10	0.12	0.25	0.40	0.38	0.20	0.25	0.37	0.38
	RELIEF						0.10	0.04	0.15	0.07	0.20	0.03	0.16	0.09	0.16	0.06	-0.01	-0.01	0.14	0.20	0.13	0.09	0.14	0.06
LVW	CART							0.14	0.36	0.32	0.38	0.28	0.24	0.08	0.12	0.07	-0.14	-0.03	0.52	0.31	0.26	0.23	0.41	0.29
	LDA								0.05	0.14	0.17	0.35	0.20	0.29	0.04	0.09	0.14	0.23	0.17	0.33	0.02	0.08	0.22	0.32
	PNN									0.33	0.22	0.10	0.13	0.04	0.32	0.15	-0.05	-0.03	0.37	0.17	0.38	0.29	0.25	0.15
	NN										0.15	0.16	0.20	0.06	0.24	0.23	-0.03	0.00	0.32	0.13	0.27	0.32	0.21	0.20
	SVM											0.30	0.06	0.03	-0.06	-0.12	-0.15	-0.11	0.35	0.37	0.05	0.07	0.52	0.30
	LR												0.14	0.22	0.02	0.07	0.03	0.19	0.32	0.43	0.08	0.14	0.32	0.42
SBGW	CART													0.14	0.14	0.18	0.03	0.12	0.31	0.20	0.22	0.21	0.12	0.16
	LDA														0.16	0.19	0.23	0.35	0.13	0.29	0.12	0.15	0.10	0.26
	PNN															0.45	0.22	0.17	0.17	0.05	0.37	0.31	0.08	0.05
	NN																0.27	0.25	0.13	0.00	0.23	0.19	0.02	0.05
	SVM																	0.36	-0.09	0.02	0.01	0.01	0.00	0.06
	LR																		0.01	0.16	0.07	0.09	0.00	0.21
SFGW	CART																			0.35	0.33	0.28	0.38	0.35
	LDA																				0.17	0.13	0.42	0.41
	PNN																					0.34	0.17	0.14
	NN																						0.14	0.15
	SVM																							0.36

4.6 Βασικές επισημάνσεις

Η αξιολόγηση της αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών για την ανάπτυξη αξιόπιστων υποδειγμάτων ταξινόμησης αποτελεί σημαντικό αντικείμενο έρευνας. Αυτή η μελέτη παρουσίασε τα πειραματικά αποτελέσματα της εφαρμογής επιλεγμένων FSAs σε διαδεδομένες τεχνικές ταξινόμησης, χρησιμοποιώντας τεχνητά και πραγματικά δεδομένα.

Βάσει της παραπάνω ανάλυσης φαίνεται ότι οι περισσότεροι αλγόριθμοι επιλογής χαρακτηριστικών οδηγούν σε σημαντικές μειώσεις του πλήθους των χαρακτηριστικών ενός δείγματος δεδομένων, χωρίς να μειώνεται η αποτελεσματικότητα των αναπτυσσόμενων υποδειγμάτων ταξινόμησης. Γενικότερα, οι διαδικασίες wrapper ήταν πιο αποτελεσματικές όσον αφορά το λόγο ακρίβειας ταξινόμησης, έχοντας βέβαια, όπως ήταν και αναμενόμενο, υψηλό υπολογιστικό φόρτο. Παρατηρήθηκε, επίσης, ότι η τεχνική ταξινόμησης SVM συνεργάζεται καλύτερα με τους FSAs, σε σχέση με τις υπόλοιπες τεχνικές. Όσον αφορά, βέβαια, την αποτελεσματικότητα των αλγορίθμων να διακρίνουν τα πλέον κατάλληλα χαρακτηριστικά, η πειραματική ανάλυση έδειξε ότι οι διαδικασίες filter δίνουν αισθητά καλύτερα αποτελέσματα από τις διαδικασίες wrapper. Εντούτοις, θα πρέπει να δοθεί έμφαση στις διαφορές που υπάρχουν στα σύνολα των χαρακτηριστικών που επιλέγονται από τα FSAs για ένα κοινό δείγμα δεδομένων, αποδεικνύοντας ότι η εφαρμογή ενός μόνο FSA μπορεί να οδηγήσει στην επιλογή διαφορετικών χαρακτηριστικών από αυτά που θα επιλεχθούν από τα υπόλοιπα FSAs. Θα ήταν, λοιπόν πιο αποτελεσματικός ο συνδυασμός δύο ή περισσότερων FSAs, με σκοπό την καλύτερη επιλογή χαρακτηριστικών και την επίτευξη απόλυτης συμφωνίας μεταξύ των διαφόρων FSAs.

Συνοψίζοντας, θα μπορούσαν να χρησιμοποιηθούν επιπρόσθετοι αλγόριθμοι επιλογής χαρακτηριστικών, όπως είναι οι γενετικοί αλγόριθμοι, ο αλγόριθμος κλάδου και φράγματος, καθώς και τεχνικές γραμμικού προγραμματισμού. Ενδιαφέρον επίσης, θα ήταν να εξεταστεί ο συνδυασμός διαφόρων FSAs για τη βελτίωση των αποτελεσμάτων όσον αφορά το λόγο ακρίβειας ταξινόμησης, όπως επίσης την επιλογή των καλύτερων χαρακτηριστικών κατά τη διαδικασία ταξινόμησης.

4.7 Περίληψη

Σε αυτό το κεφάλαιο, έγινε μια γενική περιγραφή των εξεταζόμενων αλγορίθμων επιλογής χαρακτηριστικών και εξετάστηκε η αποτελεσματικότητά τους σε συνδυασμό με ευρέως διαδεδομένες τεχνικές ταξινόμησης. Η πειραματική ανάλυση εστιάστηκε, όχι μόνο σε τεχνητά, αλλά και σε πραγματικά δεδομένα. Στην ανάλυση των αποτελεσμάτων σκοπός ήταν η αξιολόγηση της αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών στην ανάπτυξη κατάλληλων υποδειγμάτων ταξινόμησης. Στο επόμενο κεφάλαιο, θα γίνει μια συνολική αναφορά στα βασικά αποτελέσματα που επιτεύχθηκαν από την έρευνα που πραγματοποιήθηκε και θα προταθούν μελλοντικές ερευνητικές κατευθύνσεις, οι οποίες θα συμβάλλουν στην καλύτερη αντιμετώπιση του προβλήματος της ταξινόμησης.

Κεφάλαιο 5ο

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 (κεφάλαιο 2ο) και οι οποίες μπορούν να αντιμετωπιστούν πραγματοποιώντας απόλυτες συγκρίσεις με προκαθορισμένα δείγματα εκμάθησης και δείγματα ελέγχου. Οι έρευνες για την αντιμετώπιση των προβλημάτων ταξινόμησης επικεντρώνεται πρώτον, στην εφαρμογή κατάλληλων τεχνικών για την ανάπτυξη υποδειγμάτων ταξινόμησης (κεφάλαιο 2ο) και δεύτερον, στην επιλογή κατάλληλων αλγορίθμων επιλογής χαρακτηριστικών (κεφάλαιο 3ο), οι οποίοι σε συνδυασμό με διαδεδομένες τεχνικές ταξινόμησης μπορούν να αποδώσουν μεγάλες ακρίβειες ταξινόμησης.

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.

Σε πρόσφατες έρευνες μόνο ένα μικρό μέρος των ερευνών που πραγματοποιούνται για προβλήματα ταξινόμησης επικεντρώθηκε σε αλγόριθμους επιλογής χαρακτηριστικών. Σκοπός των ερευνών αυτών ήταν η μείωση των χαρακτηριστικών που απαιτούταν για την επίτευξη μιας επιτυχημένης ταξινόμησης. Όμως, οι προσπάθειες αυτές συνάντησαν προβλήματα στην αντιμετώπιση ενός αυξημένου αριθμού ακατάλληλων ή περιττών χαρακτηριστικών. Ακόμα και ο αλγόριθμος C4.5, ένας ιδιαίτερα σημαντικός αλγόριθμος στο χώρο της ταξινόμησης, εκφυλίστηκε σημαντικά όταν στο δείγμα εκμάθησης υπήρχαν μη σχετικά ή περιττά χαρακτηριστικά (Kohavi και Fresca, 1995).

Στη διεθνή βιβλιογραφία έχουν προταθεί διάφοροι αλγόριθμοι και μεθοδολογίες. Παρά όμως τις σχετικές έρευνες που έχουν διεξαχθεί στο χώρο αυτό, δεν υπάρχει μια ολοκληρωμένη έρευνα σχετικά με την αποτελεσματικότητα των προτεινόμενων μεθοδολογιών και αλγορίθμων. Μια τέτοια έρευνα πρέπει να λάβει υπόψη τις αλληλεπιδράσεις μεταξύ των διαδικασιών επιλογής χαρακτηριστικών και των τεχνικών που χρησιμοποιούνται για την ανάπτυξη υποδειγμάτων ταξινόμησης. Οι μέχρι σήμερα έρευνες περιορίζονται σε συγκεκριμένες τεχνικές ταξινόμησης (συνήθως δέντρα ταξινόμησης και τεχνικές εξαγωγής κανόνων απόφασης). Δεδομένου όμως του πλήθους των διαφορετικών τεχνικών ταξινόμησης που είναι σήμερα διαθέσιμες (νευρωνικά δίκτυα και συναφείς τεχνικές, μηχανική μάθηση, μαθηματικός προγραμματισμός, κ.ά.), είναι εμφανές ότι απαιτείται μια πιο ολοκληρωμένη ανάλυση.

Βάσει της διαπίστωσης αυτής, η παρούσα εργασία επικεντρώθηκε στην ανάπτυξη κατάλληλων αλγορίθμων επιλογής χαρακτηριστικών για την επιλογή των πλέον κατάλληλων χαρακτηριστικών για την ανάπτυξη υποδειγμάτων ταξινόμησης, αλλά πραγματοποιήθηκε και μια διεξοδική πειραματική ανάλυση (κεφάλαιο 4ο) για την αξιολόγηση της

αποτελεσματικότητας των αναπτυσσόμενων αλγορίθμων σε διαδεδομένες τεχνικές ταξινόμησης. Σκοπός της παρούσας μελέτης ήταν η πραγματοποίηση μιας ολοκληρωμένης έρευνας της αποτελεσματικότητας και των δυνατοτήτων που παρέχουν οι αλγόριθμοι επιλογής χαρακτηριστικών, καλύπτοντας σημαντικά θέματα όπως: (α) τη δυνατότητα των αλγορίθμων να εντοπίσουν την πραγματικά χρήσιμη πληροφορία (υποσύνολο χαρακτηριστικών), (β) την αποτελεσματικότητα των υποδειγμάτων ταξινόμησης τα οποία αναπτύσσονται με τη χρήση αλγορίθμων επιλογής χαρακτηριστικών, (γ) το βαθμό μείωσης της εξεταζόμενης πληροφορίας και τη σχέση του με την αποτελεσματικότητα των υποδειγμάτων ταξινόμησης, και (δ) τις αλληλεπιδράσεις στα παραπάνω θέματα μεταξύ των χρησιμοποιούμενων τεχνικών για την ανάπτυξη υποδειγμάτων ταξινόμησης και των αλγορίθμων επιλογής χαρακτηριστικών.

Η έρευνα που πραγματοποιήθηκε παρέχει τις απαραίτητες βάσεις για την πλήρη κατανόηση των αλγορίθμων επιλογής χαρακτηριστικών και την ικανότητα αυτών να διαχωρίζουν τα κατάλληλα από τα μη σχετικά ή περιττά χαρακτηριστικά και επιτρέπει τη αποτίμηση της αποτελεσματικότητας των αλγορίθμων αυτών σε προβλήματα ταξινόμησης. Για τη διερεύνηση των παραπάνω θεμάτων λήφθηκαν υπόψη διάφοροι αλγόριθμοι επιλογής χαρακτηριστικών αντιπροσωπευτικοί των υπαρχόντων προσεγγίσεων του προβλήματος. Ταυτόχρονα, εξετάστηκαν διάφορες τεχνικές ανάπτυξης υποδειγμάτων ταξινόμησης από το χώρο της μηχανικής μάθησης, της επιχειρησιακής έρευνας και της στατιστικής. Η πραγματοποίηση της συγκεκριμένης πειραματικής ανάλυσης, τόσο σε πειραματικά όσο και σε πραγματικά δεδομένα, συνέβαλλε στη διαμόρφωση μιας ολοκληρωμένης εικόνας για την αποτελεσματικότητα των διαδικασιών επιλογής χαρακτηριστικών, των δυνατοτήτων που παρέχουν και τις υπάρχουσες αλληλεπιδράσεις με διάφορες τεχνικές ανάπτυξης υποδειγμάτων ταξινόμησης.

Γενικότερα, η έρευνα της αποτελεσματικότητας των αλγορίθμων επιλογής χαρακτηριστικών παρουσιάζει σημαντικές προοπτικές περαιτέρω έρευνας. Οι κύριες από τις προοπτικές αυτές και οι αντίστοιχες ερευνητικές κατευθύνσεις εντοπίζονται στη διερεύνηση της δυνατότητας του συνδυασμού των προτεινόμενων αλγορίθμων επιλογής χαρακτηριστικών για την επιλογή ενός πιο αποτελεσματικού συνόλου χαρακτηριστικών. Κύριος στόχος είναι η διερεύνηση εκείνων των συνθηκών υπό τις οποίες οι διάφοροι αλγόριθμοι σε σχέση με τεχνικές ταξινόμησης θα παρουσιάζουν παρόμοια αποτελέσματα. Απαραίτητος είναι και ο

προσδιορισμός εκείνων των αλγορίθμων που θα συμβάλλουν στην υψηλότερη αποτελεσματικότητα των υποδειγμάτων ταξινόμησης. Η πειραματική ανάλυση της επίδρασης των αλγορίθμων στο αποτέλεσμα της ταξινόμησης, συμβάλλει στην καλύτερη κατανόηση της λειτουργίας των εξεταζόμενων αλγορίθμων επιλογής χαρακτηριστικών και στην εξαγωγή χρήσιμων συμπερασμάτων ως προς τις πρακτικές τους εφαρμογές. Επίσης, ο συνδυασμός των διαφόρων αλγορίθμων θα αποτελέσει χρήσιμο εργαλείο για την αντιμετώπιση προβλημάτων ταξινόμησης στα οποία εντοπίζονται προβλήματα ως προς τις επιδόσεις των αντικειμένων στα κριτήρια αξιολόγησης.

Β Ι Β Λ Ι Ο Γ Ρ Α Φ Ι Α

[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, 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), 279-305.

[3] Altman, E.I., Avery, R., Eisenbeis, R. and Stinkey, J. (1981), Application of Classification Techniques in Business, Banking and Finance, Contemporary Studies in Economic and Financial Analysis, Vol. 3, JAI Press, Greenwich.

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

[5] 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, 773-791.

[6] Blake, C., Keogh, E. and Merz, C. J. (1998), "UCI repository of machine learning data bases", University of California, Department of Information and Computer Science, Irvine, CA [http://www.ics.uci.edu/~mlearn/MLRepository.html].

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

[8] Breiman, L., Friedman, J.H., Olsen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*, Pacific Grove, California.

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

[10] 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, 25-32.

[11] 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).

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

[13] 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, 1122-1128.

[14] 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, 22-27.

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

[16] 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).

[17] 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.

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

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

[20] Dumais, S., Platt, J., Heckerman, D. and Sahami, M. (1998), "Inductive learning algorithms and representations for text categorization", in: *Proceedings of the International Conference on Information and Knowledge Management*, ACM Press, 148-155.

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

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

[23] 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.

[24] Fayyad, U.M. (1996), "Data mining and knowledge discovery: Making sense out of data", *IEEE Expert*, 10*,* 20-25.

[25] Fisher, R.A. (1936) "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics 7,* (pp. 179-188).

[26] Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition (2nd edition)*, Academic Press, San Diego.

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

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

[29] Hall, M. A. and Holmes, G. (2000), "Benchmarking attribute selection techniques for data mining", Department of Computer Science, University of Waikato Hamilton, New Zealand.

[30] Hall, M.A. (1999), *Correlation-Based Feature Selection for Machine Learning*, PhD thesis, University of Waikato.

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

[32] 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, 121-129.

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

[34] Kohavi, R. and John, G.H. (1997), "Wrappers for feature subset selection", *Artificial Intelligence*, 97(1-2), 273–324.

[35] Koller, D. and Sahami, M. (1996), "Toward optimal feature selection", in: Saitta, L. (ed.), *Proceedings of the 13th International Conference on Machine Learning,* Morgan Kaufmann, San Francisco, 284-292.

[36] 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, 171-182.

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

[38] 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, 319-327.

[39] 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, 129-135.

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

[41] 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, 68-75.

[42] Martin, D. (1977), *Early warning of bank failure: A logit regression approach,* Journal of Banking and Finance 1, 249-276.

[43] McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

[44] 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, 306-313.

[45] Moody, J. (1992), "The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems", in: Moody, J., Hanson, S. and Lippmann, R. (eds), *Advances in Neural Information Processing Systems*, Morgan Kaufmann, 847-854.

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

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

[48] Ohlson, J. A. (1980), *Financial ratios and the probabilistic predictions of bankruptcy,* Journal of Accounting Research, 109-131.

[49] Oliveira, L.S., Sabourin, R., Bortolozzi, F. and Suen, C.Y. (2002), "Feature selection using multi-objective genetic algorithms for handwritten digit recognition", in: *Proceedings of the 16th International Conference on Pattern Recognition,* 568-571.

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

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

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

[53] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning,* Morgan Kaufmann.

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

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

[56] Salzberg, S. (1992), "Improving classification methods via feature selection", Tech. Rep. JHU-92/12, Johns Hopkins University, Department of Computer Science.

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

[58] Simoudis, E. (1996), "Reality check for data mining", *IEEE Expert*, 10, 26-33.

[59] 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).

[60] Specht, D.F. (1980), "Probabilistic neural networks", *Neural Networks*, 3, 109-118.

[61] Steinberg, D. and Colla, P. L., (1995), CART: Tree-Structured Nonparametric Data Analysis, San Diego, CA: Salford Systems.

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

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

[64] Vafai, H. and De Jong, K. (1992), "Genetic algorithms as a tool for feature selection in machine learning", in: *4th International Conference on Tools with Artificial Intelligence*, IEEE Computer Society Press, 200-203.

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

[66] Vapnik, V.N. (2000) "The Nature of Statistical Learning Theory," Springer, New York, second edition.

[67] Witten, I. and Frank, E. (2000), "Data mining – practical machine learning tools and techniques with JAVA implementations", Morgan Kaufmann Publishers.

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

[69] Yu, L. and Liu, H. (2003), "Feature selection for high-dimensional data: A fast correlation-based filter solution", in: *Proceedings of the 20th International Conference on Machine Learning*, Washington DC.

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

[71] Zopounidis, C. and Doumpos, M. (2002), "Multicriteria classification and sorting Methods: A literature review", *European Journal of Operational Research*, 138(2), 229-246.