ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ & ΔΙΟΙΚΗΣΗΣ

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

"Επιλογή χαρακτηριστικών σε προβλήματα ταξινόμησης μέσω της θεωρίας των προσεγγιστικών συνόλων και τεχνικών επαναληπτικής δειγματοληψίας"

ΧΡΗΣΤΑΚΗΣ ΓΙΩΡΓΟΣ

Επιβλέπων Καθηγητής
ΔΟΥΜΠΟΣ ΜΙΧΑΛΗΣ

ΧΑΝΙΑ 2003

Στην οικογένεια μου

“ - Παππού αγαπημένε, είπα, δώσ’ μου μια προσταγή.
- Φτάσε όπου μπορείς, παιδί μου…
- Παππού, φώναξα τώρα πιο δυνατά, δώσ’ μου μια πιο δύσκολη, πιο κρητικιά προσταγή.
- Φτάσε όπου δεν μπορείς! ”

ΝΙΚΟΣ ΚΑΖΑΝΤΖΑΚΗΣ

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω τον κ. Μιχάλη Δούμπο για την αμέριστη συμβολή του στην παρούσα εργασία. Οι γνώσεις του, η μεταδοτικότητα, η υπευθυνότητα και η καθοδήγησή του ήταν πολύτιμη βοήθεια για μένα.

Επίσης, χρωστάω ένα μεγάλο ευχαριστώ στους ανθρώπους που συνάντησα μέσα και έξω από τις αίθουσες του Πολυτεχνείου, και οι οποίοι με βοήθησαν να μάθω πολλά, αλλά και να ζήσω πολλές ευχάριστες στιγμές κατά τη διάρκεια της φοιτητικής μου ζωής στα Χανιά.

Περιεχόμενα

1. Εισαγωγή

Ένα πρόβλημα της επιχειρησιακής έρευνας είναι αυτό της ταξινόμησης, δηλαδή το πρόβλημα της ένταξης ορισμένων εναλλακτικών δραστηριοτήτων ή αντικείμενων σε προκαθορισμένες ομοιογενείς κατηγορίες. Προβλήματα που συχνά απαιτούν την υιοθέτηση της προβληματική της ταξινόμησης βρίσκονται σε διάφορα πεδία. Στην *ιατρική,* για παράδειγμα, ο ιατρός καλείται να διαχωρίσει τους ασθενείς σε κατηγορίες (παθήσεις) ανάλογα με τα συμπτώματα που παρουσιάζουν [Tsumoto (1998), Belacel (2000)]. Στην *χρηματοοικονομική διοίκηση και οικονομική πολιτική*, μια τράπεζα επεξεργάζεται χρηματοοικονομικούς, και μη δείκτες, κατηγοριοποιώντας τις επιχειρήσεις σε πτωχευμένες και μη [Zopounidis (1998), Zopounidis και Doumpos (1998)]. Άλλοι τομείς που απαιτούν την κατηγοριοποίηση κάποιων αντικειμένων σε προκαθορισμένες ομοιογενείς κατηγορίες είναι οι: *αναγνώριση προτύπων, διαχείριση ανθρώπινου δυναμικού, διαχείριση παραγωγικών συστημάτων, μάρκετινγκ, περιβαλλοντική και ενεργειακή διαχείριση* [Δούμπος και Ζοπουνίδης (2001)].

Κατά τη διάρκεια πολλών χρόνων μελέτης του αντικειμένου της ταξινόμησης οι μεθοδολογίες που έχουν εφαρμοστεί στην συντριπτική τους πλειοψηφία ακολουθούν τη γενική φιλοσοφία της παλινδρόμησης. Προσπαθούν δηλαδή να αξιοποιήσουν τη

διαθέσιμη γνώση και πληροφορία από ταξινομήσεις που έχουν γίνει στο παρελθόν και να τις εφαρμόσουν σε νέα αντικείμενα. Για παράδειγμα στη χρηματοοικονομική διοίκηση η εξέταση των χαρακτηριστικών των επιχειρήσεων που πτώχευσαν στο παρελθόν σε αντιπαράθεση με τα χαρακτηριστικά των υγιών επιχειρήσεων μπορεί να οδηγήσει σε χρήσιμα συμπεράσματα σχετικά με τον κίνδυνο πτώχευσης των επιχειρήσεων σήμερα.

Για την κατανόηση των μεθοδολογιών ταξινόμησης κρίνεται απαραίτητο να ορισθούν κάποια χαρακτηριστικά μεγέθη της ταξινόμησης. Ως U, ορίζεται ένα πεπερασμένο σύνολο m εναλλακτικών δραστηριοτήτων ή αντικειμένων (objects). Κάθε αντικείμενο περιγράφεται από ένα σύνολο χαρακτηριστικών $Q=\{g_1, g_2, ..., g_n\}$. Η περιγραφή του αντικειμένου x_i στο χαρακτηριστικό g_j, στο εξής θα συμβολίζεται ως a_{ij}. Ορίζεται επίσης η εξαρτημένη μεταβλητή C, η οποία αφορά ένα σύνολο διακριτών επιπέδων καθένα από τα οποία αντιστοιχεί σε μία κατηγορία $c_1,c_{2,...,}c_q$, με q το πλήθος των κατηγοριών. Ακόμα έχουμε το δείγμα των παρατηρήσεων (ή δείγμα εκμάθησης, ή δείγμα αναφοράς), όπου αποτελείται από m ζεύγη της μορφής (x_i, c_i) καθένα από τα οποία αντιστοιχεί σε ένα αντικείμενο x_i (ως $c_i \in C$ συμβολίζεται η ταξινόμηση του αντικειμένου x_i) [Δούμπος και Ζοπουνίδης (2001)].

Σε ένα πρόβλημα πρόβλεψης της πτώχευσης των επιχειρήσεων όπως στο παράδειγμα του Πίνακα 1.1, το δείγμα εκμάθησης περιλαμβάνει ένα σύνολο επιχειρήσεων U (επιχειρήσεις $x_1, x_2, ..., x_{10}$). Κάθε επιχείρηση περιγράφεται από ένα σύνολο χαρακτηριστικών $Q=\{g_1, g_2, ..., g_6\}$ και ταξινομείται σε δύο κατηγορίες c_1="πτωχευμένη" και c_2="μη πτωχευμένη".

	Χαρακτηριστικά						**Κατάσταση Επιχείρησης**
Επιχειρήσεις	g_1	g_2	g_3	g_4	g_5	$g_6 = g_n$	
x_1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	Πτωχευμένη
x_2	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	Μη Πτωχευμένη
x_3	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	Μη Πτωχευμένη
x_4	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	Πτωχευμένη
x_5	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	Πτωχευμένη
x_6	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	a_{66}	Μη Πτωχευμένη
x_7	a_{71}	a_{72}	a_{73}	a_{74}	a_{75}	a_{76}	Μη Πτωχευμένη
x_8	a_{81}	a_{82}	a_{83}	a_{84}	a_{85}	a_{86}	Πτωχευμένη
x_9	a_{91}	a_{92}	a_{93}	a_{94}	a_{95}	a_{96}	Μη Πτωχευμένη
$x_{10} = x_m$	a_{101}	a_{102}	a_{103}	a_{104}	a_{105}	a_{106}	Μη Πτωχευμένη

Πίνακας 1.1. Το δείγμα παρατηρήσεων - δείγμα εκμάθησης

Το πρόβλημα της ταξινόμησης επιλύεται όταν βρεθεί η συνάρτηση f η οποία θα έχει ως είσοδο το σύνολο $U \times Q$ και θα δίνει σαν έξοδο την εξαρτημένη μεταβλητή $\hat{C}$, δηλαδή $f(U \times Q) \rightarrow \hat{C}$. Το σύμβολο $\hat{C}$ είναι η εκτιμώμενη ταξινόμηση, ενώ το C η δεδομένη ταξινόμηση. Η διαφορά τους έγκειται στο ότι η συνάρτηση f δεν επιτυγχάνει με απόλυτη επιτυχία να αντιστοιχήσει το σύνολο $U \times Q$ με το C. Στόχος φυσικά είναι να ελαχιστοποιηθεί ένα μέτρο των διαφορών που εντοπίζονται μεταξύ της εκτιμώμενης ταξινόμησης $\hat{C}$, και της δεδομένης ταξινόμησης C. Ανάλογα την μεθοδολογία ελαχιστοποίησης των διαφορών μεταξύ C και $\hat{C}$ έχουν αναπτυχθεί διάφορα μοντέλα ταξινόμησης.

Μια διαδεδομένη μέθοδος για την αντιμετώπιση των προβλημάτων ταξινόμησης είναι η θεωρία των προσεγγιστικών συνόλων (rough set theory). Η ιδέα στη οποία στηρίζεται η θεωρία αυτή είναι ότι: "για κάθε αντικείμενο υπάρχει κάποια διαθέσιμη πληροφορία, δεδομένο, ή γνώση" [Krawiec et al., 1998]. Βάσει της διαθέσιμης πληροφορίας η θεωρία των προσεγγιστικών συνόλων προσπαθεί να εντοπίσει τις αλληλοεξαρτήσεις μεταξύ διαφόρων αντικειμένων, αλλά και μεταξύ των χαρακτηριστικών. Στόχος του μοντέλου αυτού είναι ο εντοπισμός των σημαντικών χαρακτηριστικών, με απώτερο σκοπό τη μείωση της απαραίτητης πληροφορίας· παράλληλα η μεθοδολογία αυτή αντιμετωπίζει προβλήματα με ασαφή και μη συνεπή δεδομένα.

Στην εργασία αυτή γίνεται εφαρμογή της θεωρία των προσεγγιστικών συνόλων. Με την χρήση της σε ένα δείγμα παρατηρήσεων με n χαρακτηριστικά επιτυγχάνεται η δημιουργία των «ελαχίστων συνόλων». Τα ελάχιστα σύνολα (reducts) είναι υποσύνολα των Q χαρακτηριστικών. Η δημιουργία των ελαχίστων συνόλων γίνεται με σκοπό να μειωθεί η πληροφορία που είναι απαραίτητη για την ταξινόμηση των αντικειμένων. Επομένως, για το παράδειγμα που απεικονίζεται στον Πίνακα 1.1, τα ελάχιστα σύνολα (εάν υπάρχουν) είναι υποσύνολα του $Q=\{g_1, g_2, \ldots, g_6\}$, μπορεί να είναι για παράδειγμα τα υποσύνολα (g_1, g_2, g_6), (g_1, g_3, g_5, g_6). Εν τέλει, η δημιουργία των ελαχίστων συνόλων έχει τη λογική ότι χρησιμοποιώντας μόνο τα μειωμένα χαρακτηριστικά ενός ελαχίστου συνόλου μπορούμε να κατηγοριοποιήσουμε τις επιχειρήσεις επιτυγχάνοντας εξίσου καλό ή καλύτερο $\hat{C}$ από ότι με το σύνολο των χαρακτηριστικών.

Φυσικά, οι δυνατοί συνδυασμοί που μπορούν να προκύψουν εάν από το σύνολο των χαρακτηριστικών χρησιμοποιηθεί ένα υποσύνολο τους είναι πολλοί. Ακόμα και στο

παράδειγμα της πτώχευσης των επιχειρήσεων, με n=6, είναι 63! Είναι δηλαδή, το άθροισμα όλων των συνδυασμών, χωρίς διάταξη, που προκύπτουν εάν επιλέξουμε από τα 6 χαρακτηριστικά 1,ή 2,…,ή 6 από αυτά, $\binom{6}{1}+\binom{6}{2}+\binom{6}{3}+\binom{6}{4}+\binom{6}{5}+\binom{6}{6}=63$.

Κάποια (ίσως και κανένα) από αυτά τα υποσύνολα χαρακτηριστικών θα επιλεγούν από τη θεωρία των προσεγγιστικών συνόλων με συγκεκριμένη μεθοδολογία με στόχο με τη χρήση των μειωμένων χαρακτηριστικών να επιτυγχάνεται εξίσου καλό ή καλύτερο $\hat{C}$ από ότι με το σύνολο των χαρακτηριστικών. Τα υποσύνολα των χαρακτηριστικών που θα επιλεχθούν σύμφωνα με τη θεωρία των προσεγγιστικών συνόλων είναι τα «ελάχιστα σύνολα». Το ερώτημα είναι κατά πόσο τα ελάχιστα σύνολα περιέχουν επαρκή χαρακτηριστικά.

Στόχος αυτής της εργασία είναι να εξεταστεί εάν ο περιορισμός της ανάλυσης της πλέον σημαντικής πληροφορίας, όπως αυτή αποτυπώνεται στα ελάχιστα σύνολα, παρέχει αποτελέσματα εξίσου καλά ή και καλύτερα σε σχέση με τα αποτελέσματα που επιτυγχάνονται με βάση το σύνολο της διαθέσιμης πληροφορίας (σύνολο των χαρακτηριστικών). Η χρησιμότητα των ελαχίστων συνόλων ως μέσο μείωσης της πληροφορίας κατά την ανάπτυξη μοντέλων ταξινόμησης εξετάζεται όχι μόνο με την εφαρμογή της θεωρίας των προσεγγιστικών συνόλων (rough set theory) αλλά και σε συνδυασμό με άλλες μεθόδους, όπως: τα νευρωνικά δίκτυα (neural networks), ο αλγόριθμος του πλησιέστερου γείτονα (nearest neighbor), η γραμμική διακριτική ανάλυση (linear discriminant analysis), η τετραγωνική διακριτική ανάλυση (quadratic discriminant analysis) και τα δέντρα ταξινόμησης και παλινδρόμησης (classification and regression trees).

Στα κεφάλαια που ακολουθούν γίνεται αναφορά στην ταξινόμηση στη γενική της ιδέα αλλά και σε συγκεκριμένα μοντέλα ταξινόμησης (Κεφάλαιο 2). Επίσης στο Κεφάλαιο 3 γίνεται μια εκτενή αναφορά στη θεωρία των προσεγγιστικών συνόλων και σε εφαρμογές τους που χρησιμοποιήθηκαν. Ακολουθούν η πειραματική ανάλυση (Κεφάλαιο 4) και τα Συμπεράσματα (Κεφάλαιο 5).

2. Ταξινόμηση

2.1. Εισαγωγικά

Η ταξινόμηση αποτελεί κομμάτι της επιστήμης των αποφάσεων, της επιστήμης που έχει στόχο την αντιμετώπιση πρακτικών προβλημάτων λήψης αποφάσεων. Καθώς η επιστήμη των αποφάσεων αποτελεί ένα ευρύ πεδίο έρευνας σε θεωρητικό αλλά και σε πρακτικό επίπεδο η προσπάθεια κατηγοριοποίησης των προβλημάτων της μπορεί να οδηγήσει στη δημιουργία πολλών κατηγοριών. Μια όμως γενική κατηγοριοποίηση συνίσταται στη διάκριση μεταξύ των διακριτών προβλημάτων και των συνεχών προβλημάτων. Στα συνεχή προβλήματα δεν είναι δυνατό να καθοριστεί ένα πεπερασμένο σύνολο εναλλακτικών δραστηριοτήτων αλλά η οριοθέτηση του χώρου στον οποίον βρίσκονται. Η επιλογή της κατάλληλης λύσης γίνεται μόνο μέσω της αναλυτικής διερεύνησης του χώρου των εφικτών λύσεων. Σε αντιπαράθεση με τα συνεχή, τα διακριτά προβλήματα αφορούν σύνολα $\boldsymbol{U}$, πεπερασμένων, συγκεκριμένων εναλλακτικών δραστηριοτήτων. Τα διακριτά προβλήματα έχουν ως εξής:

- Επιλογή της καλύτερης εναλλακτικής δραστηριότητας.

- Κατάταξη των εξεταζόμενων εναλλακτικών δραστηριοτήτων από τις καλύτερες προς της χειρότερες.
- Περιγραφή των εναλλακτικών με στόχο τον εντοπισμό των βασικών τους χαρακτηριστικών και ιδιοτήτων.
- Ταξινόμηση των εναλλακτικών δραστηριοτήτων σε προκαθορισμένες κατηγορίες.

Τα προβλήματα της ταξινόμησης διαχωρίζονται επιμέρους στα προβλήματα της: διάκρισης (discrimination) - ταξινόμησης (classification), και της διατεταγμένης ταξινόμησης (sorting). Οι δύο πρώτοι όροι αναφέρονται στα προβλήματα όπου οι κατηγορίες στις οποίες εντάσσονται οι εναλλακτικές δραστηριότητες ορίζονται ονομαστικά (nominal). Οι όροι «διάκριση» και «ταξινόμηση» χρησιμοποιούνται από στατιστικολόγους και ερευνητές του πεδίου της τεχνητής νοημοσύνης (νευρωνικά δίκτυα, μηχανική μάθηση, κλπ.). Αντίθετα, ο όρος «διατεταγμένη ταξινόμηση» αναφέρεται σε προβλήματα όπου οι κατηγορίες είναι διατεταγμένες από την καλύτερη στη χειρότερη και χρησιμοποιείται από ερευνητές της πολυκριτήριας ανάλυσης αποφάσεων.

Ένας άλλος διαχωρισμός που πρέπει να αποσαφηνισθεί είναι αυτός μεταξύ της ταξινόμησης και της ομαδοποίησης (clustering). Η διαφορά τους έγκειται στο ότι, στα προβλήματα της ταξινόμησης οι κατηγορίες στις οποίες εντάσσονται οι εναλλακτικές δραστηριότητες είναι αυστηρά προκαθορισμένες. Σε αντίθεση με τα προβλήματα της ταξινόμησης, στην περίπτωση της ομαδοποίησης σκοπός της ανάλυσης είναι ο σχηματισμός των ομάδων, έτσι ώστε οι δραστηριότητες σε κάθε ομάδα να παρουσιάζουν παρόμοια χαρακτηριστικά [Δούμπος και Ζοπουνίδης (2001)].

2.2. Ανασκόπηση των τεχνικών ταξινόμησης

Για την επίλυση των προβλημάτων ταξινόμησης υπήρξε σημαντικό ενδιαφέρον από τις αρχές του 20^{ου} αιώνα μέχρι σήμερα και έχει δημιουργηθεί ένα ευρύ φάσμα μεθόδων ταξινόμησης. Η συντριπτική τους πλειοψηφία εντάσσεται σε δύο βασικές κατηγορίες: α) στις στατιστικές και οικονομετρικές προσεγγίσεις, β) στις μη παραμετρικές προσεγγίσεις.

Οι **στατιστικές και οικονομετρικές προσεγγίσεις** είναι οι παλαιότερες τεχνικές ταξινόμησης που αναπτύχθηκαν. Το 1936 ο Fisher, στην ερευνητική του εργασία, ανέπτυξε την πρώτη πολυδιάστατη στατιστική μέθοδο, τη **γραμμική διακριτική ανάλυση** (linear discriminant analysis). Αργότερα, το 1947, ο Smith επέκτεινε την εργασία του Fisher, αναπτύσσοντας την **τετραγωνική διακριτική ανάλυση** (quadratic discriminant analysis).

Οι δύο αυτές τεχνικές είναι ιδιαίτερα διαδεδομένες καθώς: εφαρμόζονται εύκολα και αποτελούν το αποδεδειγμένα βέλτιστο αποτέλεσμα ταξινόμησης όταν οι μεταβλητές ακολουθούν την πολυμεταβλητή κανονική κατανομή και οι πίνακες διακύμανσης συνδιακύμανσης των κατηγοριών είναι γνωστοί.

Από την άλλη, οι μέθοδοι της διακριτικής ανάλυσης, δεν βοηθούν στον προσδιορισμό της συνεισφοράς του κάθε χαρακτηριστικού στην ταξινόμηση των εναλλακτικών δραστηριοτήτων. Παράλληλα, στην περίπτωση όπου οι μεταβλητές δεν ακολουθούν την πολυμεταβλητή κανονική κατανομή και οι πίνακες διακύμανσης-συνδιακύμανσης δεν είναι εύκολο να προσδιοριστούν, φαινόμενα που συναντώνται στην πλειοψηφία των πρακτικών περιπτώσεων, δημιουργούνται θεωρητικά κενά στην εφαρμογή τους.

Τα παραπάνω προβλήματα των μεθόδων της διακριτικής ανάλυσης αποτέλεσαν το βασικό κίνητρο για την ανάπτυξη των οικονομετρικών προσεγγίσεων: του **γραμμικού υποδείγματος πιθανότητας** (linear probability model), του **λογιστικού υποδείγματος πιθανότητας** (logit analysis) και του **κανονικού υποδείγματος πιθανότητας** (probit analysis). Τα τελευταία δύο παρουσιάζουν σημαντικές ομοιότητες ενώ έχουν και ένα σημαντικό πλεονέκτημα σε σχέση με το γραμμικό υπόδειγμα, αντιμετωπίζουν προβλήματα ταξινόμησης με παραπάνω από δύο κατηγορίες.

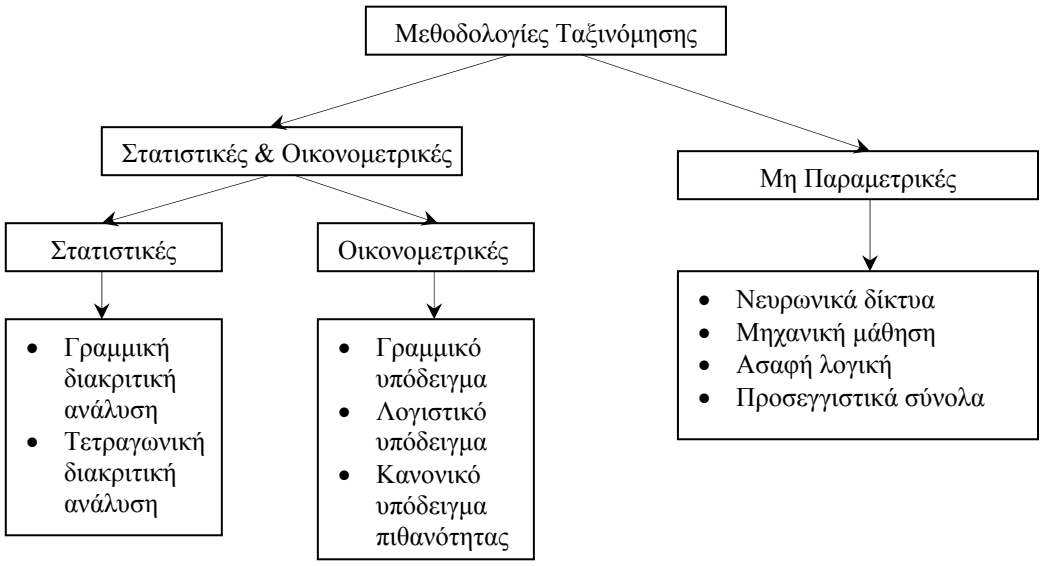
Το λογιστικό και το κανονικό υπόδειγμα πιθανότητας γνώρισαν ιδιαίτερη διάδοση μετά τις εργασίες του πρόσφατα βραβευμένου με Νόμπελ Οικονομίας Daniel McFadden (1974, 1980). Τα υποδείγματα αυτά με τη βοήθεια μιας μη γραμμικής συνάρτησης υπολογίζουν την πιθανότητα των εναλλακτικών δραστηριοτήτων να ανήκουν σε κάθε μια από τις υπό εξέταση κατηγορίες. Σημαντικά πλεονεκτήματα των μεθοδολογιών αυτών είναι ότι: α) μπορεί να προσδιορισθεί η σημαντικότητα των χαρακτηριστικών μέσω στατιστικών ελέγχων όπως το t-τεστ, β) είναι δυνατό για ταξινομήσεις που αφορούν περισσότερο από δύο κατηγορίες να εφαρμόσουν την

πολλαπλή ονομαστική (multinomial) και τη διατεταγμένη μορφή (ordered), και γ) δεν υπόκεινται σε στατιστικούς περιορισμούς.

Όμως, παρά το ότι δεν πραγματοποιούνται υποθέσεις για τις στατιστικές ιδιότητες των δεδομένων, όπως στην διακριτική ανάλυση, άλλες μορφές στατιστικών υποθέσεων υφίστανται στις οικονομετρικές προσεγγίσεις. Αξιοσημείωτο δε είναι ότι, παρά της φιλοδοξίες των ερευνητών, τα αποτελέσματα των υποδειγμάτων ταξινόμησης δεν δείχνουν σημαντική βελτίωση σε σχέση με αυτά της γραμμικής και τετραγωνικής διακριτικής ανάλυσης.

Γενικά οι στατιστικές και οικονομετρικές προσεγγίσεις, πετυχαίνοντας πολύ καλά αποτελέσματα στην ταξινόμηση, αποτελούν ένα σημείο αναφοράς και σύγκρισης με τις νέες τεχνικές ταξινόμησης.

Αναζητώντας, οι επιστήμονες, μεθόδους ανεξάρτητες των στατιστικών ιδιοτήτων των εξεταζόμενων δεδομένων οδηγήθηκαν στην ανάπτυξη των **μη παραμετρικών προσεγγίσεων**. Οι προσεγγίσεις αυτές δεν βασίζονται σε στατιστικές υποθέσεις και δεν απαιτούν την ανάλυση των στατιστικών ιδιοτήτων των δεδομένων. Σημαντικές μη παραμετρικές μέθοδοι, οι περισσότερες εκ των οποίων θα αναλυθούν στην Ενότητα 2.3, είναι: τα νευρωνικά δίκτυα (neural networks), η μηχανική μάθηση (machine learning), η ασαφής λογική (fuzzy logic), και (Κεφάλαιο 3) τα προσεγγιστικά σύνολα (rough set theory). Στο Σχήμα 2.1 παρουσιάζονται γραφικά οι κύριοι διαχωρισμοί των μεθόδων ταξινόμησης.



Σχήμα 2.1. Παρουσίαση των κύριων μεθόδων ταξινόμησης

2.3. Διάφορες μέθοδοι ταξινόμησης

Για την ανάλυση των μεθόδων της ταξινόμησης θα χρησιμοποιηθούν οι ακόλουθοι ορισμοί.

- $U=\{x_1, x_2, ..., x_m\}$, ορίζεται ένα πεπερασμένο σύνολο m εναλλακτικών δραστηριοτήτων ή αντικειμένων (objects).
- $Q=\{g_1, g_2, ..., g_n\}$, ορίζεται ένα σύνολο n χαρακτηριστικών (attributes). Στο εξής η περιγραφή του αντικειμένου x_i στο χαρακτηριστικό g_j θα συμβολίζεται ως a_{ij}.
- C, ορίζεται η εξαρτημένη μεταβλητή η οποία αφορά ένα σύνολο διακριτών επιπέδων καθένα από τα οποία αντιστοιχεί σε μία κατηγορία $c_1,c_2,...,c_q$, με q το πλήθος των κατηγοριών.
- Ακόμα έχουμε $(C,U\times Q)$, το δείγμα των παρατηρήσεων (ή δείγμα εκμάθησης, ή δείγμα αναφοράς), όπου αποτελείται από m ζεύγη της μορφής (x_i, c_i) καθένα από τα οποία αντιστοιχεί σε ένα αντικείμενο x_i (ως $c_i \in C$ συμβολίζεται η ταξινόμηση του αντικειμένου x_i)

	Χαρακτηριστικά						**Ταξινόμηση**
Εναλλακτικές	g_1	g_2	…	…	…	g_n	
x_1							c_1
x_2							c_2
…			$U\times Q$				…
…							…
x_m							c_m

Πίνακας 2.1. Το δείγμα παρατηρήσεων - δείγμα εκμάθησης

2.3.1. Η γραμμική και τετραγωνική διακριτική ανάλυση

Η γραμμική διακριτική ανάλυση [Fisher 1936] καθώς και η τετραγωνική διακριτική ανάλυση [Smith 1947] αποτελούν παρόμοιές μεθοδολογικές προσεγγίσεις, για αυτό εξετάζονται παράλληλα. Οι μέθοδοι αυτοί χρησιμοποιούν ως δείγμα εκμάθησης ένα σύνολο εναλλακτικών δραστηριοτήτων που έχει γνωστή ταξινόμηση.

Η **γραμμική διακριτική ανάλυση** πραγματοποιείται σε q κατηγορίες αναπτύσσοντας q-1 γραμμικές συναρτήσεις της μορφής:

$$Z_{kl} = a_{kl} + \sum_{j=1}^{n} b_{klj} g_j$$

Έχουμε, $g_1, g_2, \ldots, g_n$ τα χαρακτηριστικά του δείγματός παρατηρήσεων, a_{kl} μια σταθερά, και $b_{kl1}, b_{kl2}, \ldots, b_{kln}$ τους συντελεστές των συναρτήσεων της διακριτικής ανάλυσης. Επίσης, οι δείκτες k και l αναφέρονται στις κατηγορίες c_k, c_l, αντίστοιχα.

Ο υπολογισμός των συντελεστών b_{kl} και του σταθερού όρου a_{kl} βασίζεται στο ότι οι εναλλακτικές δραστηριότητες στα εξεταζόμενα χαρακτηριστικά ακολουθούν την πολυμεταβλητή κανονική κατανομή. Παράλληλα, στηρίζεται στην υπόθεση ότι οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών είναι ίσοι. Βάσει αυτών των υποθέσεων οι υπολογισμοί γίνονται ως εξής:

$$b_{kl} = S^{-1} \cdot [\mu_k - \mu_l]$$

$$a_{kl} = -[\mu_\kappa + \mu_l]' \cdot b_{kl}/2$$

Ως μ_k συμβολίζεται το διάνυσμα των μέσων τιμών των χαρακτηριστικών για τις εναλλακτικές δραστηριότητες της κατηγορίας c_k. Ενώ ως S συμβολίζεται ο κοινός πίνακας διακύμανσης-συνδιακύμανσης των κατηγοριών ο οποίος υπολογίζεται από την παρακάτω σχέση.

$$S = \frac{\sum_{k=1}^{q} \sum_{\forall x_i \in c_k} [g_i - \mu_k] \cdot [g_i - \mu_k]'}{m - q}$$

Η ταξινόμηση της δραστηριότητας x_i στην κατηγορία c_k θα γίνει εάν για όλες τις άλλες κατηγορίες c_l ισχύει:

$$Z_{kl}(g_i) \geq \ln \frac{K(k|l)\pi_l}{K(l|k)\pi_k}$$

Ως π_k και π_l συμβολίζονται οι εκ των πρότερων πιθανότητες και ως $K(k \mid l)$ το κόστος εσφαλμένων ταξινομήσεων.

Αντίστοιχα για την **τετραγωνική διακριτική** ανάλυση έχουμε την παρακάτω τετραγωνική συνάρτηση.

$$Z_{kl} = a_{kl} + \sum_{j=1}^{n} b_{klj} g_j + \sum_{j=1}^{n} \sum_{h=1}^{n} c_{kljh} g_j g_h$$

Σε αυτή την περίπτωση οι πίνακες διακύμανσης-συνδιακύμανσης μεταξύ των κατηγοριών είναι άνισοι και υπολογίζονται από μια νέα σχέση, όμοια και ο σταθερός όρος a_{kl}, και οι συντελεστές b_{klj}, c_{kljh}. Ο υπολογισμός των συντελεστών, όπως και στην γραμμική διακριτική ανάλυση, βασίζεται στο ότι οι εναλλακτικές δραστηριότητες στα

εξεταζόμενα χαρακτηριστικά ακολουθούν την πολυμεταβλητή κανονική κατανομή. Υπενθυμίζεται ότι οι δείκτες *k* και *l* αναφέρονται στις κατηγορίες c_k, c_l, αντίστοιχα.

$$b_{kl} = -2[\mu_k' S_k^{-1} - \mu_l' S_l^{-1}$$

$$a_{kl} = \mu_k^{'} S_k^{-1} \mu_k - \mu_l^{'} S_l^{-1} \mu_l - \ln\left|S_l S_k^{-1}\right|$$

$$c_{kl} = S_k^{-1} - S_l^{-1}$$

και,

$$S_k = \frac{\sum_{\forall x_i \in c_k} [g_i - \mu_k][g_i - \mu_k]'}{m-1}$$

Η εναλλακτική x_i ταξινομείται στην κατηγορία c_k εάν για όλες τις άλλες c_l ισχύει:

$$Z_{kl}(g_i) \geq -2 \ln \frac{K(k|l)\pi_l}{K(l|k)\pi_k}$$

Αξίζει να αναφερθεί πως ο καθορισμός των πρότερων πιθανοτήτων π_k και του κόστους εσφαλμένων ταξινομήσεων $K(k|l)$ είναι μια δύσκολη διαδικασία. Ο τρόπος αντιμετώπισης της δυσκολίας αυτής είναι να προσδιορίζονται τα όρια που διαχωρίζουν τις κατηγορίες μέσω της διαδικασίας δοκιμής και σφάλματος, με στόχο να ελαχιστοποιηθεί ο συνολικός αριθμός των εσφαλμένων ταξινομήσεων και να υπάρχει μια ισορροπία στον αριθμό των εσφαλμένων ανα κατηγορία.

Αναφερόμενοι στην γραμμική και διακριτική ανάλυση υπενθυμίζεται ότι είναι ιδιαίτερα διαδεδομένες μέθοδοι και αποτελούν σημείο αναφοράς και σύγκρισης με τις νέες τεχνικές ταξινόμησης καθώς: εφαρμόζονται εύκολα, και αποτελούν το αποδεδειγμένα βέλτιστο αποτέλεσμα ταξινόμησης όταν οι μεταβλητές ακολουθούν την πολυμεταβλητή κανονική κατανομή και οι πίνακες διακύμανσης συνδιακύμανσης των κατηγοριών έχουν τη μορφή που ταιριάζει στην αντίστοιχη μέθοδο.

2.3.2. Νευρωνικά δίκτυα

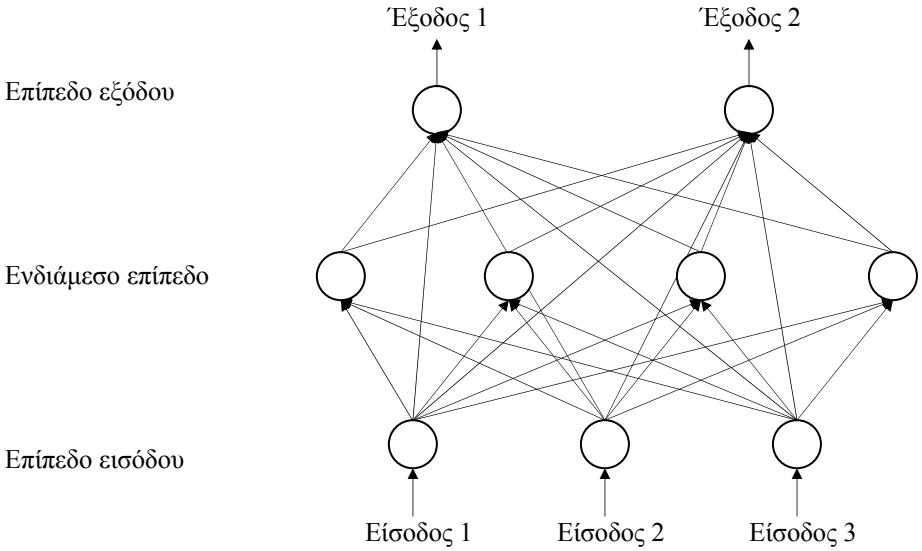
Μια από τις πιο δημοφιλείς μεθόδους εύρεσης παραμέτρων σε μη γραμμικά προβλήματα είναι τα νευρωνικά δίκτυα (neural networks) ή τεχνητά νευρωνικά δίκτυα (artificial neural networks). Το όνομα τους φανερώνει ότι προσπαθούν να εξομοιώσουν τη λειτουργία του ανθρώπινου εγκεφάλου. Ο τρόπος λειτουργίας του ανθρώπινου

εγκεφάλου στηρίζεται σε ένα ιδιαίτερα πολύπλοκο δίκτυο πολυπληθών οργάνων, των νευρώνων. Κάθε νευρώνας επεξεργάζεται σήματα που λαμβάνει είτε από τους αισθητήρες του εξωτερικού περιβάλλοντος του ανθρώπου, είτε από άλλους νευρώνες. Ύστερα από την επεξεργασία αυτών των σημάτων εισόδου, σήματα εξόδου στέλνονται προς άλλους νευρώνες για περαιτέρω επεξεργασία ή παράγονται σήματα εξόδου από το νευρωνικό δίκτυο. Ακολουθεί μια απεικόνιση της παραπάνω διαδικασίας (Σχήμα 2.2).

Όπως φανερώνει το σχήμα 2.2 η επεξεργασία των σημάτων είσοδο του δικτύου γίνεται σε διάφορα επίπεδα:

- Το επίπεδο εισόδου: αποτελείται από μια σειρά κόμβων, ένα για κάθε είσοδο του νευρωνικού δικτύου. Ενώ, κάθε είσοδος του δικτύου αντιπροσωπεύει ένα χαρακτηριστικό $g_1, g_2, \ldots, g_n$ του προβλήματος της ταξινόμησης.
- Το ενδιάμεσο επίπεδο: αποτελείται από μια ή περισσότερες σειρές κόμβων. Ο αριθμός των κόμβων προσδιορίζεται μέσα από την διαδικασία δοκιμής και σφάλματος. Για τα προβλήματα της ταξινόμησης, έρευνες των Patuwo et al. (1993) και Subramanian et al. (1993) έδειξαν πως ικανοποιητικά αποτελέσματα επιτυγχάνονται μεταξύ των αριθμών q και $2n+1$.
- Το επίπεδο εξόδου: αποτελείται από έναν ή περισσότερους κόμβους, ανάλογα με τη φύση του προβλήματος. Για την επίλυση μεθοδολογιών ταξινόμησης για q κατηγορίες ταξινόμησης ο αριθμός των κόμβων του επιπέδου εξόδου είναι ο αμέσως μεγαλύτερος ακέραιος του αριθμού $\log_2 q$ [Subramanian et al. (1993)]. Εναλλακτικά όμως μπορεί κάθε κόμβος στο επίπεδο εξόδου να αντιπροσωπεύει μια κατηγορία ταξινόμησης. Για παράδειγμα, για δυο κατηγορίες ταξινόμησης, μπορεί να υπάρχει ένας κόμβος εξόδου, που να λαμβάνει την τιμή 1 για την c_1 και την τιμή 2 για την c_2. Από την άλλη μπορεί να υπάρχουν δύο κόμβοι, ένας για κάθε κατηγορία.

Γενικότερα, σε ένα νευρωνικό δίκτυο, όταν δύο κόμβοι ανήκουν σε διαφορετικό επίπεδο συνδέονται μεταξύ τους. Μεταξύ δύο κόμβων κάθε σύνδεση έχει ένα βάρος. Τα βάρη αυτά υπολογίζονται μέσω διαδικασιών ελαχιστοποίησης των αποκλίσεων μεταξύ των αποτελεσμάτων του δικτύου και των πραγματικών. Κατά αναλογία με την στατιστική παλινδρόμηση, ως μέτρο των αποκλίσεων υπολογίζεται το άθροισμα των τετραγώνων των σφαλμάτων.



Σχήμα 2.2. Απεικόνιση ενός νευρωνικού δικτύου

Η εξήγηση της λειτουργίας των νευρωνικών δικτύων είναι συνήθως πολύ δύσκολη. Τα νευρωνικά δίκτυα έχουν χαρακτηριστεί ως «μαύρο κουτί», καθώς δεν μπορούν να φανερώσουν γιατί μια εναλλακτική με συγκεκριμένα χαρακτηριστικά ταξινομείται σε συγκεκριμένη κατηγορία. Αυτό αποτελεί και ένα βασικό μειονέκτημα της μεθόδου ιδιαίτερα στην υποστήριξη μιας απόφασης. Ένα άλλο αρνητικό των νευρωνικών δικτύων αποτελεί ο αυξημένος υπολογιστικός φόρτος κατά την εκμάθηση του δικτύου.

Αντιθέτως, τα νευρωνικά δίκτυα έχουν τη δυνατότητα να αναπαριστούν έντονα μη γραμμικά προβλήματα με θεωρητικά άπειρη ακρίβεια. Επίσης πλεονέκτημα τους είναι ότι παρέχουν τη δυνατότητα παράλληλης επεξεργασίας.

Κατά την εφαρμογή των νευρωνικών δικτύων στα προβλήματα της ταξινόμησης οι επιστήμονες δεν έχουν συναντήσει τα αναμενόμενα αποτελέσματα. Σύμφωνα με τις έρευνες του Subramanian et al. (1993) τα αποτελέσματα ήταν ικανοποιητικά σε σχέση με τις μεθόδους της διακριτικής ανάλυσης, ιδιαίτερα σε πολύπλοκα προβλήματα με μεγάλο αριθμό κριτηρίων και κατηγοριών. Όμως, οι έρευνες του Patuwo et al (1993) έδειξαν σημαντικά χειρότερα αποτελέσματα σε σχέση με τις ίδιες μεθόδους. Μόνο οι Archer και Wang (1993), εφαρμόζοντας κατάλληλους περιορισμούς μονοτονίας κατά

την εκμάθηση του δικτύου, πέτυχαν υψηλότερη ακρίβεια σε σχέση με την γραμμική διακριτική ανάλυση.

2.3.3. Μηχανική Μάθηση

Η μηχανική μάθηση αποτελεί μια σημαντική θεωρία του χώρου της τεχνικής νοημοσύνης. Μια χαρακτηριστική και διαδεδομένη μεθοδολογία μηχανικής μάθησης είναι η επαγωγική μάθηση. Κατά την εφαρμογή της επαγωγικής μάθησης αναλύονται οι ταξινομήσεις από σύνολα δεδομένων. Η εκμετάλλευση αυτής της γνώσης δεν οδηγεί σε οποιαδήποτε μορφής στατιστική ανάλυση ή σε δημιουργία συνάρτησης ταξινόμησης όπως στις πιο πολλές μεθοδολογίες. Η επαγωγική μάθηση λειτουργεί μέσω κανόνων της μορφής:

ΕΑΝ *στοιχειώδες συνθήκες* ΤΟΤΕ *συμπεράσματα*

Οι κάνονες βασίζονται στη λογική: εάν επαληθεύονται όλες οι συνθήκες του κανόνα τότε η εναλλακτική δραστηριότητα ταξινομείται στην κατηγορία που υποδεικνύεται στο συμπέρασμα.

Ενδιαφέρον είναι ο τρόπος με τον οποίο δημιουργούνται οι κανόνες αυτοί. Η ανάλυση που ακολουθεί είναι βασισμένη στον αλγόριθμο C4.5 (Quinlan 1993), μια ιδιαίτερα γνωστή τεχνική στο χώρο της μηχανικής μάθησης. Για την δημιουργία των κανόνων η μεθοδολογία επαναλαμβάνει τρία βήματα. Σε κάθε επανάληψη υπολογίζεται η εντροπία της ταξινόμησης του g_i χαρακτηριστικού. Η σχέση που υπολογίζει την εντροπία είναι:

$$I(D) = -\sum_{h=1}^{t} \frac{v_h}{m} \sum_{k=1}^{q} p(D_h / C_k) \log_2 [p(D_h / C_k)]$$

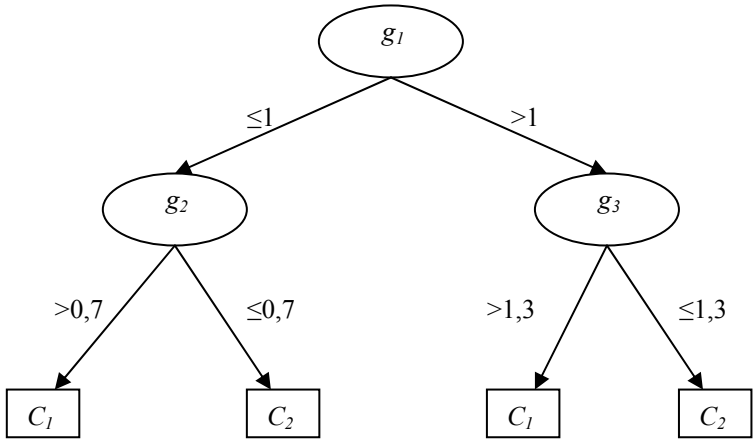
Όπου: m οι εναλλακτικές δραστηριότητες του συνόλου δεδομένων, t τα υποσύνολα $D_1, D_2, \ldots, D_t$ στα οποία διαχωρίζονται οι εναλλακτικές ανάλογα με την τιμή τους στο χαρακτηριστικό που εξετάζεται, και v_h ο αριθμός των εναλλακτικών στο υποσύνολο D_h (h=1,2, …, t).

Τα τρία βήματα που ακολουθούνται είναι:

1) Υπολογισμός της εντροπίας της ταξινόμησης των χαρακτηριστικών.
2) Επιλογή του χαρακτηριστικού με την μικρότερη εντροπία.

3) Διαχωρισμός των εναλλακτικών σε υποσύνολα $D_1,D_2,...,D_t$, όσες και οι διαφορετικές τιμές που λαμβάνει το χαρακτηριστικό που επιλέχθηκε (για ποιοτικά κριτήρια), ή όσα τα σημεία διαχωρισμού (για ποσοτικά κριτήρια).

Για κάθε νέο υποσύνολο επαναλαμβάνονται τα τρία παραπάνω βήματα μέχρι τελικά να επιτευχθεί η σωστή ταξινόμηση όλων των εναλλακτικών δραστηριοτήτων του συνόλου δεδομένων. Το αποτέλεσμα είναι η δημιουργία ενός δέντρου. Ένα απλό παράδειγμα τέτοιου δέντρου παρουσιάζεται στο σχήμα 2.3.



Σχήμα 2.3. Αναπαράσταση ενός δέντρου αποφάσεων

Εύκολα γίνεται κατανοητό ότι η δημιουργία ενός δέντρου αποφάσεων όπως το παραπάνω είναι δυνατό να αναλυθεί σε κανόνες ταξινόμησης. Στο παράδειγμα του Σχήματος 2.3 ένας τέτοιος κανόνας είναι: «ΕΑΝ μια εναλλακτική x_i λαμβάνει τιμή μεγαλύτερη του 1 για το g_1 και μεγαλύτερη του 1,3 για το g_3 ΤΟΤΕ ταξινομείται στην κατηγορία C_1».

Η μεθοδολογία που περιγράφηκε είναι δυνατόν να οδηγήσει σε ένα ιδιαίτερα πολύπλοκο δένδρο με μεγάλο αριθμό φύλλων. Υπάρχει περίπτωση ένα φύλλο να αντιπροσωπεύει μια μονό εναλλακτική δραστηριότητα. Αυτό έχει το θετικό ότι η επαγωγική μάθηση μπορεί να προσαρμοστεί πλήρως στην ταξινόμηση που ακολουθεί το σύνολο δεδομένων. Όμως παράλληλα με την ανάπτυξη του δέντρου εφαρμόζονται διάφορες μέθοδοι περικοπής των φύλλων του με σκοπό την γενίκευση των ιδιοτήτων του (Quinlan 1993, Breiman et al. 1984, Gelfand et al. 1991).

Κλείνοντας την αναφορά στη μηχανική μάθηση, τα σημαντικά πλεονεκτήματα που χαρακτηρίζουν τις μεθόδους του χώρου αυτού είναι: α) η δυνατότητα διαχείρισης ποιοτικών δεδομένων, β) η δυνατότητα διαχείρισης δεδομένων με ελλιπή στοιχεία, γ) η δυνατότητα διαχείρισης ιδιαίτερα μεγάλων συνόλων δεδομένων και δ) η εύκολη κατανόηση του αναπτυσσόμενου μοντέλου ταξινόμησης.

3. Θεωρία προσεγγιστικών συνόλων

3.1. Εισαγωγικά

Η θεωρία των προσεγγιστικών συνόλων (rough set theory) αναπτύχθηκε από τον Pawlak το 1982. Σύμφωνα με τον ίδιο η ιδέα στη οποία στηρίζεται η θεωρία αυτή είναι ότι: «για κάθε αντικείμενο υπάρχει κάποια διαθέσιμη πληροφορία, δεδομένο, ή γνώση». Ένα παράδειγμα ταξινόμησης το οποίο παραθέτει για την κατανόηση της παραπάνω ιδέας είναι: «Έστω ότι τα αντικείμενα είναι ασθενείς που νοσούν από μια συγκεκριμένη αρρώστια, τα συμπτώματα της αρρώστιας διαμορφώνουν την πληροφορία που γνωρίζουμε για τους ασθενείς. Αντικείμενα τα οποία χαρακτηρίζονται από την ίδια πληροφορία είναι δυσδιάκριτα βάσει της διαθέσιμης πληροφορίας. Η σχέση της δυσδιακριτότητας που προσδιορίζεται με αυτόν τον τρόπο είναι η μαθηματική βάση της θεωρία των προσεγγιστικών συνόλων» [Pawlak 1997]. Εκ πρώτης όψεως, για τους μη γνώστες της συγκεκριμένης μεθοδολογίας, το παραπάνω παράδειγμα δεν διευκολύνει άμεσα στην κατανόηση της μεθοδολογίας. Με την ανάλυση της Ενότητας 3.2 θα φανεί ότι το παράδειγμα του Pawlak αποτελεί μια αρκετά συμπυκνωμένη άποψη για τη θεωρία των προσεγγιστικών συνόλων.

Η θεωρία των προσεγγιστικών συνόλων ξεκίνησε ως μια νέα μεθοδολογία στον χώρο της μηχανικής μάθησης (Ενότητα 2.3.3) καθώς εμφανίζει αρκετές ομοιότητες ως προς τη θεωρία αυτή. Ένα από τα κοινά χαρακτηριστικά τους είναι και η δημιουργία κανόνων ταξινόμησης: «ΕΑΝ συνθήκες ΤΟΤΕ ταξινόμηση». Όμως, σταδιακά αποτέλεσε έναν αυτοτελή χώρο έρευνας με σημαντικές διαφορές από τη μηχανική μάθηση. Μάλιστα, θεωρείται ότι η θεωρία των προσεγγιστικών συνόλων βρίσκεται στη διασταύρωση της ασαφής λογικής, της θεωρίας των σημείων, των νευρωνικών δικτύων, των δικτύων Petri, και πολλών άλλων θεωριών της τεχνικής νοημοσύνης, της λογικής και των μαθηματικών [Beynon et al. 2000].

Το ιδιαίτερο χαρακτηριστικό της θεωρίας των προσεγγιστικών συνόλων είναι ότι περιγράφει τις αλληλεξαρτήσεις μεταξύ των χαρακτηριστικών των εναλλακτικών δραστηριοτήτων. Με την διαδικασία αυτή βοηθάει στον εντοπισμό των σημαντικών χαρακτηριστικών και στη μείωση της απαραίτητης πληροφορίας για την ταξινόμηση. Σημαντική είναι και η συνεισφορά της θεωρίας των προσεγγιστικών συνόλων για την αντιμετώπιση προβλημάτων με ασαφή και μη συνεπή δεδομένα.

3.2. Ανάλυση της θεωρίας των προσεγγιστικών συνόλων

Όπως έχει διαφανεί από το παράδειγμα που αναφέρθηκε προηγούμενα, η θεωρία των προσεγγιστικών συνόλων αναλύει περιπτώσεις αντικειμένων (εναλλακτικών δραστηριοτήτων) για τα οποία τίθεται κάποια πληροφορία, υπάρχει γνώση για τα χαρακτηριστικά τους. Για την ανάλυση της κρίνεται αναγκαίο αρχικά να ορίστουν κάποια χαρακτηριστικά μεγέθη.

- $U=\{x_1, x_2, ..., x_m\}$, ένα πεπερασμένο σύνολο m εναλλακτικών δραστηριοτήτων ή αντικειμένων (objects).
- $Q=\{g_1, g_2, ..., g_n\}$, ένα σύνολο n χαρακτηριστικών (attributes) ή υπό συνθήκη χαρακτηριστικών (condition attribute).
- C, η εξαρτημένη μεταβλητή η οποία αφορά ένα σύνολο διακριτών επιπέδων καθένα από τα οποία αντιστοιχεί σε μία κατηγορία $c_1,c_{2,...,}c_q$, με q το πλήθος των κατηγοριών. Οι κατηγορίες $c_1,c_{2,...,}c_q$ ονομάζονται και χαρακτηριστικά απόφασης (decision attributes).

- $(C,U\times Q)$, το δείγμα των παρατηρήσεων (ή δείγμα εκμάθησης, ή δείγμα αναφοράς), όπου αποτελείται από m ζεύγη της μορφής (x_i, c_i) καθένα από τα οποία αντιστοιχεί σε ένα αντικείμενο x_i (ως $c_i \in C$ συμβολίζεται η ταξινόμηση του αντικειμένου x_i).
- $V=\bigcup_{i\in Q} V_i$, με V_i το πεδίο τιμών του κάθε χαρακτηριστικού g_i[1].
- $f(U\times Q)\rightarrow V$, η συνάρτηση πληροφορίας, τέτοια ώστε $f(x_j,g_i)\in V_i$ για κάθε $g_i \in Q$, $x_j \in U$. Η συνάρτηση αυτή δίνει την τιμή του χαρακτηριστικού g_i που λαμβάνει η εναλλακτική δραστηριότητα x_i.

3.2.1. Ανάλυση μέσω παραδείγματος

Για την κατανόηση της θεωρίας των προσεγγιστικών συνόλων η ανάλυση της θα γίνει μέσω ενός παραδείγματος (Πίνακας 3.1). Στο παράδειγμα είναι επτά αντικείμενα με τα χαρακτηριστικά τους να αποτελούνται από 0 και 1 (π.χ. απαντήσεις Ναι/Όχι) και να εντάσσονται στις κατηγορίες «Α» ή «Θ» (π.χ. Αρσενικό/Θηλυκό) [Beynon et al. 2000].

		Χαρακτηριστικά						**Ταξινόμηση Αντικειμένων**
		g_1	g_2	g_3	g_4	g_5	g_6	
Αντικείμενα	x_1	1	1	1	1	1	1	Α
	x_2	1	0	1	0	1	1	Α
	x_3	0	0	1	1	0	0	Α
	x_4	1	1	1	0	0	1	Θ
	x_5	1	0	1	0	1	1	Θ
	x_6	0	0	0	1	1	0	Θ
	x_7	1	0	1	0	1	1	Θ

Πίνακας 3.1. Παράδειγμα ταξινομημένων αντικειμένων – Πίνακας δεδομένων

Η αρχή στην οποία στηρίζεται η θεωρία των προσεγγιστικών συνόλων είναι ότι: τα αντικείμενα με τις ίδιες τιμές χαρακτηριστικών ταξινομούνται στην ίδια κατηγορία. Όμως, στον παραπάνω πίνακα υπάρχουν αντικείμενα των οποίων τα χαρακτηριστικά

[1] Στην κλασική θεωρία των προσεγγιστικών συνόλων οι τιμές που παίρνει κάθε χαρακτηριστικό g_i είναι διακριτές. Σε περίπτωση που οι τιμές είναι συνεχείς πραγματοποιείται μια διακριτοποίηση. Ως διακριτοποίηση του πεδίου τιμών [a, b] σε h υποδιαστήματα εννοείται η διάσπαση του στα $[a_1,a_2)$, $[a_2,a_3)$, …, $[a_{h-1},a_h]$, με $a_1=a$ και $a_h=b$. Όμως, νεότερες διαδικασίες εφαρμογής της θεωρίας των προσεγγιστικών συνόλων δεν απαιτούν την διακριτοποίηση των συνεχών τιμών των χαρακτηριστικών.

έχουν τις ίδιες τιμές αλλά ταξινομούνται σε διαφορετική κατηγορία, π.χ. τα αντικείμενα x_2, x_5, x_7. Για την επίλυση αυτού του προβλήματος ορίσθηκε ένα πολύ σημαντικό μέγεθος στο οποίο βασίζεται η θεωρία των προσεγγιστικών συνόλων, η **δυσδιακριτότητα**. Τα αντικείμενα x_2, x_5, x_7 ονομάζονται δυσδιάκριτα επειδή λαμβάνουν τις ίδιες τιμές στα χαρακτηριστικά τους, ενώ τα αντικείμενα x_1, x_3, x_4, x_6 ονομάζονται διακριτά επειδή λαμβάνουν διαφορετικές τιμές στα χαρακτηριστικά τους.

Τα παραπάνω μπορούν να παρουσιαστούν και με μαθηματική μορφή. Έτσι, δυο εναλλακτικές δραστηριότητες x_j και x_l λέγεται ότι είναι δυσδιάκριτες εάν και μόνο εάν χαρακτηρίζονται από ακριβώς την ίδια πληροφορία, δηλαδή $f(x_j,g_i)=f(x_l,g_i)$ για κάθε $g_i \in P \subseteq Q$. Ως P ορίζεται ένα υποσύνολο του Q, καθώς η δυσδιακριτότητα έχει νόημα να υπολογίζεται και για ένα υποσύνολο των συνολικών χαρακτηριστικών. Για κάθε P λοιπόν, μεταξύ δύο εναλλακτικών δραστηριοτήτων x_j και x_l, αντιστοιχεί μια σχέση δυσδιακριτότητας που συμβολίζεται I_P.

$$I_p = \{(x_j, x_l) \in U \times U : f(x_j, g_i) = f(x_l, g_i), \forall g_i \in P)$$

Επιπλέον, κάθε σύνολο δυσδιάκριτων αντικειμένων ονομάζεται στοιχειώδες σύνολο (elementary set). Το σύνολο των δυσδιάκριτων αντικειμένων που περιλαμβάνει το αντικείμενο $x_j \in U$ συμβολίζεται ως $I_P(x_j)$.

Παράλληλα, παρατηρείται στο παράδειγμα του Πίνακα 3.1. ότι τα αντικείμενα x_1, x_3, x_4, x_6 μπορούν να ταξινομηθούν με βεβαιότητα, ενώ τα αντικείμενα x_2, x_5, x_7 δεν μπορούμε να τα ταξινομήσουμε με βεβαιότητα. Το σύνολο των αντικειμένων το οποίο περιλαμβάνει όλα τα αντικείμενα τα οποία ταξινομούνται με βεβαιότητα σε κάποια κατηγορία ονομάζεται **κάτω προσέγγιση** (lower approximation) της κατηγορίας. Από την άλλη, **άνω προσέγγιση** (upper approximation) ονομάζεται το σύνολο των αντικειμένων το οποίο περιλαμβάνει όλα τα αντικείμενα τα οποία πιθανόν να ανήκουν στην εξεταζόμενη κατηγορία.

Η μαθηματική διατύπωση των παραπάνω, έχει ως εξής:

$$\underline{P}Y = \{x_j \in Y : I_p(x_j) \subseteq Y)$$

και

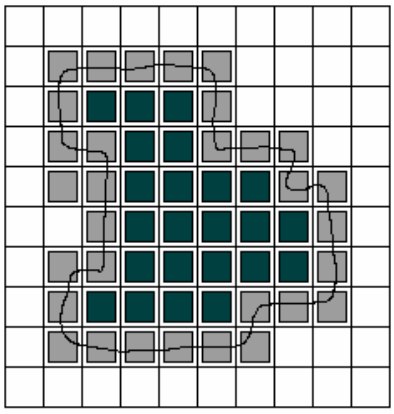
$$\overline{P}Y = \bigcup_{x_j \in Y} I_p(x_j)$$

Τα $\underline{PY}$ και $\overline{PY}$ αντιπροσωπεύουν την κάτω και άνω προσέγγιση αντίστοιχα. Το P είναι ένα υποσύνολο του Q και το Y είναι ένα υποσύνολο του U.

Η διαφορά μεταξύ της κάτω και άνω προσέγγισης ονομάζεται περιοχή αμφιβολίας (boundary region) του προσεγγιστικού συνόλου και συμβολίζεται ως:

$$BN_p(Y) = \overline{PY} - \underline{PY}$$

Σχηματικά τα παραπάνω μπορούν να αναπαρασταθούν ως έχει στο Σχήμα 3.1.



- Τα τετράγωνα αντιπροσωπεύουν τα δυσδιάκριτα αντικείμενα
- Το $\underline{PY}$ είναι τα μαύρα τετράγωνα
- Το $\overline{PY}$ είναι τα γκρι και μαύρα τετράγωνα
- Το $BN_P(Y)$ είναι τα γκρι τετράγωνα
- Το Y είναι η περιοχή που εσωκλείεται από την κλειστή γραμμή

Σχήμα 3.1. Αναπαράσταση των προσεγγιστικών συνόλων

Επίσης: α) ένα σύνολο με περιοχή αμφιβολίας $BN_P(Y)=Ø$ ονομάζεται ακριβές σύνολο (crisp set), ενώ β) ένα σύνολο με $BN_P(Y)\neq Ø$ ονομάζεται προσεγγιστικό σύνολο (rough set).

Η παραπάνω θεωρία μπορεί να αναλυθεί στο παράδειγμα του Πίνακα 3.1. Έστω ότι $P=\{g_1, g_2, g_3\}$ και $Y=\{x_5, x_6\}$. Τα σύνολα των δυσδιάκριτων αντικειμένων είναι τα εξής:

$$I_P(x_1)=I_P(x_4)=\{x_1, x_4\},\ I_P(x_2)=I_P(x_5)=I_P(x_7)=\{x_2, x_5, x_7\},\ I_P(x_3)=\{x_3\},\ I_P(x_6)=\{x_6\}$$

Από τα παραπάνω σύνολα μόνο το $I_P(x_6)$ ταξινομείται με βεβαιότητα στο Y. Σε αυτή την περίπτωση $\underline{PY}=\{x_6\}$ (πρέπει $I_p(x_j)\subseteq Y$). Για την άνω προσέγγιση χρησιμοποιούμε τα $I_P(x_j)$ που έχουν κοινά αντικείμενα με το Y, έτσι $\overline{PY}=$ $=I_P(x_5)\cup I_P(x_6)=\{x_2, x_5, x_6, x_7\}$. Επίσης, γνωρίζοντας την άνω και κάτω προσέγγιση, $BN_P(Y)=\{x_2, x_5, x_7\}$.

Μια ακόμα βασική αρχή στην οποία στηρίζεται η θεωρία των προσεγγιστικών συνόλων είναι η **ποιότητα της ταξινόμησης**. Για κάθε κατηγορία $c_1, c_2, \ldots, c_q$

προσδιορίζεται η επιτυχία της ταξινόμησης ως ο αριθμός των αντικειμένων που κατηγοριοποιούνται με βεβαιότητα σε αυτή την κατηγορία προς τον αριθμό αυτών που πιθανόν ταξινομούνται στην ίδια κατηγορία. Ο δείκτης που διαμορφώνεται κατά αυτό τον τρόπο ονομάζεται **ακρίβεια του προσεγγιστικού συνόλου**. Όσο αφορά το σύνολο των κατηγοριών, ο δείκτης αυτός ονομάζεται ποιότητα της ταξινόμησης. Η σχέση που δίνει την ακρίβεια του προσεγγιστικού συνόλου είναι:

$$a_P(Y) = \frac{\left|\underline{PY}\right|}{\left|\overline{PY}\right|}$$

Αντίστοιχα, η σχέση της ποιότητας της ταξινόμησης δεν είναι τίποτα άλλο από το άθροισμα των q κάτω προσεγγίσεων προς το σύνολο των αντικειμένων (m), δηλαδή είναι:

$$\gamma_P(Y) = \frac{\sum_{k=1}^{q}\left|\underline{PY}_k\right|}{m}$$

Η ακρίβεια του προσεγγιστικού συνόλου καθώς και η ποιότητα της ταξινόμησης μπορούν να αναλυθούν στο παράδειγμα του Πίνακα 3.1. Έστω ότι $P=\{g_1, g_2, g_3\}$ και $Y=\{x_1, x_2, x_3\}$. Τα σύνολα των δυσδιάκριτων αντικειμένων παραμένουν:

$$I_P(x_1)=I_P(x_4)=\{x_1, x_4\},\ I_P(x_2)=I_P(x_5)=I_P(x_7)=\{x_2, x_5, x_7\},\ I_P(x_3)=\{x_3\},\ I_P(x_6)=\{x_6\}$$

Το σύνολο $I_P(x_3)$ ταξινομείται με βεβαιότητα στο Y σύνολο. Σε αυτή την περίπτωση $\underline{PY}=\{x_3\}$. Για την άνω προσέγγιση χρησιμοποιούνται τα $I_P(x_j)$ που έχουν κοινά αντικείμενα με το Y, έτσι $\overline{PY} = I_P(x_1)\cup I_P(x_2)\cup I_P(x_3)=\{x_1, x_2, x_3, x_4, x_5, x_7\}$. Η ακρίβεια του προσεγγιστικού συνόλου σε αυτήν την περίπτωση είναι 1/6. Επιπλέον με $P=\{g_1, g_2, g_3\}$ και $Y=\{x_4, x_5, x_6, x_7\}$, τα σύνολα των δυσδιάκριτων αντικειμένων είναι:

$$I_P(x_1)=I_P(x_4)=\{x_1, x_4\},\ I_P(x_2)=I_P(x_5)=I_P(x_7)=\{x_2, x_5, x_7\},\ I_P(x_3)=\{x_3\},\ I_P(x_6)=\{x_6\}$$

Αυτή τη φορά το σύνολο $I_P(x_6)$ ταξινομείται με βεβαιότητα στο Y σύνολο, $\underline{PY}=\{x_6\}$. Για την άνω προσέγγιση $\overline{PY} = I_P(x_4)\cup I_P(x_5)\cup I_P(x_6)\cup I_P(x_7)=\{x_1, x_2, x_4, x_5, x_6, x_7\}$. Έτσι, η ακρίβεια του προσεγγιστικού συνόλου είναι 1/6. Όσο αφορά την ποιότητα της ταξινόμησης για το U, έχουμε $\gamma_P(Y) = \frac{\sum_{k=1}^{q}\left|\underline{PY}_k\right|}{m} = \frac{2}{7}$.

Όμοια εργαζόμενοι για το σύνολο των χαρακτηριστικών (P=Q) βρίσκεται ότι η ποιότητα της ταξινόμησης είναι 4/7.

3.2.2. Η δημιουργία των ελαχίστων συνόλων

Έχει ήδη επισημανθεί, στην Ενότητα 3.1 αλλά και στην Εισαγωγή της εργασίας, πως μια ιδιαίτερα σημαντική ιδιότητα της θεωρίας των προσεγγιστικών συνόλων είναι η δημιουργία των «ελαχίστων συνόλων». Η θεωρία των προσεγγιστικών συνόλων αποτελεί τη μόνη μέθοδο ταξινόμησης όπου αντιμετωπίζει το πρόβλημα της συσχέτισης της πληροφορίας. Προσπαθεί δηλαδή να βρει τα σημαντικά χαρακτηριστικά, τα οποία θα αποτελέσουν τα χαρακτηριστικά των ελαχίστων συνόλων και τα οποία μπορούν να αποδώσουν το ίδιο αποτέλεσμα ταξινόμησης με το σύνολο της πληροφορίας που είναι διαθέσιμη.

Τα ελάχιστα σύνολα είναι υποσύνολα του Q. Στο παράδειγμα του Πίνακα 3.1 είναι δυνατόν να δημιουργηθούν πολλά υποσύνολα του Q, τόσα δηλαδή όσα το άθροισμα όλων των συνδυασμών, χωρίς διάταξη, που μπορούν να προκύψουν εάν επιλεγεί από το σύνολο των 6 χαρακτηριστικών ένα υποσύνολο των 1,ή 2,…,ή 6 από αυτών, ίσο με: $\binom{6}{1}+\binom{6}{2}+\binom{6}{3}+\binom{6}{4}+\binom{6}{5}+\binom{6}{6}=63!$ Αυτός είναι και ο μέγιστος αριθμός των ελαχίστων συνόλων που μπορούν να υπάρξουν σε σύνολο δεδομένων με έξι χαρακτηριστικά.

Πιο συγκεκριμένα, για να ονομαστεί ένα υποσύνολο του Q ως ελάχιστο σύνολο πρέπει η ποιότητα της ταξινόμησης για το υποσύνολο αυτό των χαρακτηριστικών να είναι ίση με την ποιότητα της ταξινόμησης του συνόλου των χαρακτηριστικών. Η μαθηματική έκφραση των παραπάνω έχει: έστω P ένα σύνολο χαρακτηριστικών και R ένα υποσύνολο του P, εάν $\gamma_P(Y)=\gamma_R(Y)$, τότε το R ονομάζεται Y-ελάχιστο σύνολο ή απλά **ελάχιστο σύνολο** (reduct) και συμβολίζεται $\mathrm{RED}_Y(P)$.

Στην περίπτωση όπου υπάρχουν περισσότερα του ενός ελάχιστα σύνολα, τότε η τομή τους ονομάζεται **πυρήνας** των ελαχίστων συνόλων και συμβολίζεται $\mathrm{CORE}_Y(P)$. Δηλαδή: $\mathrm{CORE}_Y(P) = \cap\, \mathrm{RED}_Y(P)$. Ο πυρήνας έχει την ιδιότητα να αποτελείται από τα πλέον σημαντικά χαρακτηριστικά τα οποία δεν μπορούν να αγνοηθούν χωρίς να υπάρξει μείωση της ποιότητας της ταξινόμησης.

Εφαρμόζοντας την παραπάνω θεωρία στο παράδειγμα του Πίνακα 3.1, για $P=\{g_1, g_2, g_3\}$ και $U=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ πιθανά ελάχιστα υποσύνολα είναι τα: $R_1=\{g_1\}$, $R_2=\{g_2\}$, $R_3=\{g_3\}$, $R_4=\{g_1, g_2\}$, $R_5=\{g_1, g_3\}$, $R_6=\{g_2, g_3\}$, $R_7=\{g_1, g_2, g_3\}$. Αρκεί, η ποιότητα της ταξινόμησης κάποιων από τα R_1, R_2, R_3, R_4, R_5, R_6, R_7 να είναι ίση με την ποιότητα της ταξινόμησης του P, η οποία έχει βρεθεί στην Ενότητα 3.2.1 ίση με 2/7, έτσι ώστε τα σύνολα αυτά να ονομαστούν ελάχιστα σύνολα.

- Για το $R_1=\{g_1\}$ τα σύνολα των δυσδιάκριτων αντικειμένων έχουν:

$$I_P(x_1)=\{x_1, x_2, x_4, x_5, x_7\},\ I_P(x_3)=\{x_3, x_6\}$$

Για $Y=\{x_1, x_2, x_3\}$ κανένα από αυτά δεν ταξινομείται με βεβαιότητα. Σε αυτή την περίπτωση το $\underline{PY}$ είναι το κενό σύνολο. Ομοίως, για $Y=\{x_4, x_5, x_6, x_7\}$ το $\underline{PY}$ είναι το κενό σύνολο. Έτσι, η ποιότητα της ταξινόμησης $\gamma_P(Y)=\frac{\sum_{k=1}^{q}\left|\underline{PY}_k\right|}{m}$ είναι 0/7 (m=7).

- Για το $R_2=\{g_2\}$ τα σύνολα των δυσδιάκριτων αντικειμένων έχουν:

$$I_P(x_1)=\{x_1, x_4\},\ I_P(x_2)=\{x_2, x_3, x_5, x_6, x_7\}$$

Καθώς για $Y=\{x_1, x_2, x_3\}$, αλλά και για $Y=\{x_4, x_5, x_6, x_7\}$ το $\underline{PY}$ είναι το κενό σύνολο, η ποιότητα της ταξινόμησης είναι 0/7.

- Για το $R_3=\{g_3\}$ τα σύνολα των δυσδιάκριτων αντικειμένων έχουν:

$$I_P(x_1)=\{x_1, x_2, x_3, x_4, x_5, x_7\},\ I_P(x_6)=\{x_6\}$$

Για $Y=\{x_4, x_5, x_6, x_7\}$ το $I_P(x_6)=\{x_6\}$ ταξινομείται με βεβαιότητα ($\underline{PY}=\{x_6\}$), ενώ με $Y=\{x_1, x_2, x_3\}$ είναι $\underline{PY}=\emptyset$. Η ποιότητα της ταξινόμησης έτσι, είναι 1/7.

- Για το $R_4=\{g_1, g_2\}$ τα σύνολα των δυσδιάκριτων αντικειμένων έχουν:

$$I_P(x_1)=\{x_1, x_4\},\ I_P(x_2)=\{x_2, x_5, x_7\},\ I_P(x_3)=\{x_3, x_6\}$$

Υπολογίζεται ότι η ποιότητα της ταξινόμησης είναι 0/7.

- Για το $R_5=\{g_1, g_3\}$ τα σύνολα των δυσδιάκριτων αντικειμένων έχουν:

$$I_P(x_1)=\{x_1, x_2, x_4, x_5, x_7\},\ I_P(x_3)=\{x_3\},\ I_P(x_6)=\{x_6\}$$

Για $Y=\{x_1, x_2, x_3\}$ είναι $\underline{P}Y=\{x_3\}$, για $Y=\{x_4, x_5, x_6, x_7\}$ βρίσκεται $\underline{P}Y=\{x_6\}$. Συνεπώς, η ποιότητα της ταξινόμησης υπολογίζεται 2/7.

- Για το $R_6=\{g_2, g_3\}$ τα σύνολα των δυσδιάκριτων αντικειμένων έχουν:

$$I_P(x_1)=\{x_1, x_4\},\ I_P(x_2)=\{x_2, x_3, x_5, x_7\},\ I_P(x_6)=\{x_6\}$$

Σε αυτή την περίπτωση η ποιότητα της ταξινόμηση υπολογίζεται ότι είναι 1/7.

- Για το $R_7=P$ η ποιότητα της ταξινόμησης είναι 2/7.

Ύστερα από αυτή την ανάλυση συμπεραίνεται ότι τα ελάχιστα σύνολα του P είναι τα $R_5=\{g_1, g_3\}$, $R_7=\{g_1, g_2, g_3\}$ και ο πυρήνας τους είναι $R_5 \cap R_7=\{g_1, g_3\}$.

3.2.3. Η δημιουργία των κανόνων ταξινόμησης

Ένα σημαντικό χαρακτηριστικό της θεωρίας των προσεγγιστικών συνόλων είναι η δημιουργία των κανόνων ταξινόμησης. Μάλιστα μπορεί να θεωρηθεί ότι η μέχρι τώρα ανάλυση αποτελεί το κομμάτι που αφορά αποκλειστικά τη θεωρία των προσεγγιστικών συνόλων, το κομμάτι που ακολουθεί αποτελεί μια συχνή εφαρμογή των μηχανών μάθησης. Οι κανόνες ταξινόμησης έχουν τη μορφή:

«ΕΑΝ *συγκεκριμένες συνθήκες* ΤΟΤΕ *ταξινόμηση σε συγκεκριμένη/ες κατηγορία/ες*»

Πρέπει να διευκρινιστεί ότι οι κανόνες ταξινόμησης προέρχονται αποκλειστικά από ένα ελάχιστο σύνολο, το οποίο προσδιορίστηκε από τη διαδικασία που αναφέρθηκε στην προηγούμενη ενότητα. Το ελάχιστο σύνολο μπορεί να επιλέχθηκε από τα υπόλοιπα σύμφωνα με την κρίση του αποφασίζοντα ή μέσω απλών ευρετικών διαδικασιών [Slowinski και Zopounidis, 1995].

Έστω, ότι για το παράδειγμα του Πίνακα 3.1 επιλέγεται το ελάχιστο σύνολο $R_5=\{g_1, g_3\}$, για να δημιουργηθούν οι αντίστοιχοι κανόνες ταξινόμησης. Σε αυτή την περίπτωση έχουμε την ανάλυση μέσω του αντίστοιχου Πίνακα 3.2.

	g_1	g_3	Ταξινόμηση
x_1	1	1	Α
x_2	1	1	Α
x_3	0	1	Α
x_4	1	1	Θ
x_5	1	1	Θ
x_6	0	0	Θ
x_7	1	1	Θ

Πίνακας 3.2. Πίνακας δεδομένων του ελαχίστου συνόλου $R_5=\{g_1, g_3\}$.

Μια πλήρη αντιστοίχηση του παραπάνω πίνακα σε κανόνες ταξινόμησης θα αποφέρει τους εξής κανόνες:

1) ΕΑΝ $g_1 = 1$ και $g_3 = 1$ ΤΟΤΕ ταξινόμηση = Α ή Θ
2) ΕΑΝ $g_1 = 0$ και $g_3 = 1$ ΤΟΤΕ ταξινόμηση = Α
3) ΕΑΝ $g_1 = 0$ και $g_3 = 0$ ΤΟΤΕ ταξινόμηση = Θ

Οι παραπάνω κανόνες μπορούν να απλοποιηθούν στους εξής κανόνες:

1) ΕΑΝ $g_1 = 0$ και $g_3 = 1$ ΤΟΤΕ ταξινόμηση = Α
2) ΕΑΝ $g_3 = 0$ ΤΟΤΕ ταξινόμηση = Θ
3) ΕΑΝ $g_1 = 1$ ΤΟΤΕ ταξινόμηση = Α ή Θ

Ενώ, μπορούν με τα ίδια δεδομένα να ελαχιστοποιηθούν οι κανόνες που δημιουργούνται, χωρίς να είναι ισοδύναμοι σε πληροφορία με τους παραπάνω, στους εξής δύο:

1) ΕΑΝ $g_3 = 0$ ΤΟΤΕ ταξινόμηση = Θ
2) ΕΑΝ $g_1 = 0$ και $g_3 = 1$ ΤΟΤΕ ταξινόμηση = Α

Από το παραπάνω παράδειγμα διαφαίνεται ότι για την ανάπτυξη των κανόνων ταξινόμησης μπορούν να εφαρμοστούν οι ακόλουθες στρατηγικές:

- Ανάπτυξη ενός εξαντλητικού συνόλου αποτελούμενο από όλους τους δυνατούς κανόνες.
- Ανάπτυξη του ελαχίστου συνόλου που καλύπτει όλες τις εναλλακτικές δραστηριότητες του δείγματος εκμάθησης.
- Ανάπτυξη ενός συνόλου ισχυρών κανόνων, ακόμη και μερικώς διακριτικών κανόνων, οι οποίοι δεν καλύπτουν απαραίτητα όλες τις εναλλακτικές δραστηριότητες του δείγματος εκμάθησης.

Οι κανόνες είναι δυνατόν να συνοδεύονται από διάφορους χαρακτηρισμούς. Αυτοί, οι οποίοι καλύπτουν μόνο εναλλακτικές δραστηριότητες που ανήκουν στην κατηγορία στην οποία υποδεικνύει ο κάθε κανόνας (θετικά παραδείγματα) ονομάζονται διακριτικοί (discriminant). Ενώ, κανόνες οι οποίοι καλύπτουν ακόμα και εναλλακτικές δραστηριότητες που δεν ανήκουν στην κατηγορία που υποδεικνύει ο κανόνας (αρνητικά παραδείγματα), ονομάζονται μερικώς διακριτικοί. Οι μερικώς διακριτικοί κανόνες συνοδεύονται από ένα συντελεστή συνέπειας, ο οποίος ονομάζεται επίπεδο

διάκρισης και ισούται με το λόγο των θετικών προς των αρνητικών παραδειγμάτων. Επιπλέον, συνεπής (consistent) ονομάζεται ο κανόνας όπου όλες οι εναλλακτικές δραστηριότητες που επαληθεύουν το μέρος των συνθηκών ανήκουν στην κατηγορία που υποδεικνύεται από το αποτέλεσμα του κανόνα. Εάν ένας συνεπής κανόνας υποδεικνύει μόνο μια κατηγορία ονομάζεται **ακριβής** (exact), διαφορετικά ονομάζεται **προσεγγιστικός**.

Όταν ένα αντικείμενο ανήκει σε έναν προσεγγιστικό κανόνα τότε ανήκει σε κάποια από τις κατηγορίες όπου υποδεικνύει ο κανόνας. Για να προσδιοριστεί όμως σε ποια από αυτές ακολουθείται μια διαδικασία. Κάθε κανόνας συνοδεύεται από ένα μέτρο της ισχύος του. Μάλιστα, για τους προσεγγιστικούς κανόνες προσδιορίζεται η ισχύς της κάθε κατηγορίας που υποδεικνύει ο κανόνας. Ανάλογη με την ισχύ του κανόνα είναι και η επιτυχία της ταξινόμησης που επιφέρει. Επίσης, κανόνες με μεγάλη ισχύ είναι συνήθως λιγότερο εξειδικευμένοι και το τμήμα των συνθηκών τους περιλαμβάνει ένα μικρό αριθμό στοιχειωδών συνθηκών.

Μετά τη δημιουργία των κανόνων ολοκληρώνεται η μεθοδολογία της θεωρίας των προσεγγιστικών συνόλων με την εφαρμογή τους και την ταξινόμηση νέων αντικειμένων στις κατηγορίες του συνόλου *C*. Σε αυτή την περίπτωση εμφανίζονται οι ακόλουθες περιπτώσεις [Slowinski και Stefanowski, 1994]:

I. Το νέο αντικείμενο τηρεί τις προϋποθέσεις ενός ακριβή κανόνα.
II. Το νέο αντικείμενο τηρεί τις συνθήκες πολλών ακριβών κανόνων που υποδεικνύουν την ίδια κατηγορία.
III. Το νέο αντικείμενο τηρεί τις συνθήκες ενός προσεγγιστικού κανόνα ή πολλών ακριβών κανόνων που υποδεικνύουν διαφορετικές κατηγορίες.
IV. Το νέο αντικείμενο δεν τηρεί τις συνθήκες κανενός κανόνα.

Για τις περιπτώσεις I και II είναι προφανές ότι δεν υπάρχουν προβλήματα ταξινόμησης, η κατηγοριοποίηση γίνεται στην υποδεικνυόμενη κατηγορία. Για την περίπτωση III το πρόβλημα της κατηγοριοποίησης αντιμετωπίζεται λαμβάνοντας υπόψη την ισχύ των κανόνων ταξινόμησης. Η ταξινόμηση μπορεί να γίνει όπως υποδεικνύει ο ισχυρότερος κανόνας ή λαμβάνοντας υπόψη την κατηγορία με τη μεγαλύτερη ισχύ για την περίπτωση προσεγγιστικών κανόνων. Αντίθετα με τις παραπάνω η περίπτωση IV παρουσιάζει σημαντικές δυσκολίες στην αντιμετώπιση της.

Προτεινόμενη λύση είναι η διερεύνηση των κανόνων οι οποίοι τηρούνται μερικώς από το νέο αντικείμενο σε συνδυασμό με την ισχύ τους, σύμφωνα με το σύστημα LERS [Grzymala-Busse, 1992].

Γενικότερα διαδικασίες ανάπτυξης και εφαρμογής των κανόνων ταξινόμησης υπάρχουν πολλές. Μια από τις πλέον διαδεδομένες διαδικασίες ανάπτυξης κανόνων είναι ο αλγόριθμος LEM2 [Grzymala-Busse, 1992]. Ο αλγόριθμος αυτός προϋποθέτει ότι έχει εφαρμοστεί μια διαδικασία διακριτοποίησης των τιμών. Οι συνθήκες των κανόνων που αναπτύσσονται μέσω του αλγορίθμου LEM2 έχουν την μορφή $g_i=v_i$, με $v_i \in V_i$. Γενικώς, η εφαρμογή του LEM2 οδηγεί στην ανάπτυξη του ελαχίστου συνόλου κανόνων, που είναι πλήρες και δεν πλεονάζει. Το σύνολο των κανόνων είναι πλήρες ή μη πλεονασματικό όταν με την αφαίρεση ενός οποιουδήποτε κανόνα από το σύνολο, πλέον δεν καλύπτονται όλες οι εναλλακτικές δραστηριότητες του δείγματος εκμάθησης.

4. Πειραματική ανάλυση

4.1. Εισαγωγή

Ένα ιδιαίτερο χαρακτηριστικό της θεωρίας των προσεγγιστικών συνόλων, όπως έχει ήδη επισημανθεί, είναι ο εντοπισμός των «ελαχίστων συνόλων»· η ικανότητα τους αυτή συμβάλει στη μείωση της απαραίτητης πληροφορίας για την ταξινόμηση χωρίς να απαιτεί τη μείωση του ποσοστού επιτυχίας στην κατηγοριοποίηση. Σκοπός της εργασίας αυτής είναι να εξεταστεί το παραπάνω, δηλαδή εάν ο περιορισμός της πληροφορίας, όπως αυτή αποτυπώνεται στα ελάχιστα σύνολα, παρέχει αποτελέσματα εξίσου καλά ή και καλύτερα σε σχέση με τα αποτελέσματα που επιτυγχάνονται με βάση το σύνολο της διαθέσιμης πληροφορίας (σύνολο των χαρακτηριστικών).

Πιο συγκεκριμένα, στην εργασία αυτή γίνεται εφαρμογή της θεωρίας των προσεγγιστικών συνόλων για την ανεύρεση των ελαχίστων συνόλων. Με βάση τα χαρακτηριστικά τα οποία αποτελούν τα ελάχιστα σύνολα δημιουργούνται τα δείγματα εκμάθησης, και από αυτά οι κανόνες ταξινόμησης. Σύμφωνα με την επιτυχία αυτών των κανόνων στα δείγματα ελέγχου κρίνεται και η επιτυχία των αντίστοιχων ελαχίστων συνόλων στην ταξινόμηση. Παράλληλα με την παραπάνω μεθοδολογία χρησιμοποιείται και η μέθοδος του cross-validation [Stone, (1974)]. Με τη μέθοδο του cross-validation

πραγματοποιείται η πλησιέστερη αξιολόγηση της ικανότητας γενίκευσης των κανόνων που αναπτύσσονται με την παραπάνω διαδικασία. Έτσι, ελέγχεται εάν το ελάχιστο σύνολο με την μεγαλύτερη δυνατότητα γενίκευσης επιτυγχάνει αποτελέσματα ταξινόμησης εξίσου καλά ή και καλύτερα σε σχέση με τη χρήση του συνόλου των χαρακτηριστικών.

Επίσης, στην εργασία εξετάζεται η χρησιμότητα των ελαχίστων συνόλων ως μέσο μείωσης της πληροφορίας σε συνδυασμό και με άλλες μεθόδους, όπως: τα νευρωνικά δίκτυα (neural networks), ο αλγόριθμος του πλησιέστερου γείτονα (nearest neighbor), η γραμμική διακριτική ανάλυση (linear discriminant analysis), η τετραγωνική διακριτική ανάλυση (quadratic discriminant analysis) και τα δέντρα ταξινόμησης και παλινδρόμησης (classification and regression trees). Και σε αυτή την περίπτωση χρησιμοποιήθηκε η μέθοδος του cross-validation για να ελεγχθεί η δυνατότητα γενίκευση των αποτελεσμάτων της ταξινόμησης που επιτυγχάνονται.

4.2. Δεδομένα και μεθοδολογία

4.2.1. Η μεθοδολογία του cross-validation

Με τη χρήση της μεθοδολογίας του cross-validation [Stone (1974)] αξιολογείται ο βαθμός γενίκευσης των αποτελεσμάτων των μεθοδολογιών ταξινόμησης. Υπάρχουν δύο σημαντικοί λόγοι για να μετρηθεί ο βαθμός γενίκευσης των αποτελεσμάτων: α) για να εκτιμηθεί η αποτελεσματικότητα μιας μεθόδου ταξινόμησης, β) για να συγκριθούν διάφορες μεθοδολογίες ταξινόμησης μεταξύ τους και να επιλεχθεί η πλέον κατάλληλη.

Στην εργασία αυτή, έγινε εφαρμογή της μεθοδολογίας του cross-validation με σκοπό να αξιολογηθεί η αποτελεσματικότητα της χρήσης των ελαχίστων συνόλων. Κατά την εφαρμογή της θεωρίας των προσεγγιστικών συνόλων, αλλά και σε συνδυασμό με άλλες μεθοδολογίες που εφαρμόσθηκαν στην εργασία αυτή, δημιουργήθηκαν τα δείγματα εκμάθησης και τα δείγματα ελέγχου. Στην περίπτωση ενός δεδομένου δείγματος εκμάθησης και ενός δείγματος ελέγχου, το δείγμα εκμάθησης χρησιμοποιείται για την ανάπτυξη ενός κατάλληλου μοντέλου ταξινόμησης σύμφωνα με τις αρχές της εκάστοτε μεθόδου, ενώ το δείγμα ελέγχου χρησιμοποιείται για τον έλεγχο της αποτελεσματικότητας του μοντέλου. Όμως, το γεγονός ότι με αυτό τον τρόπο πραγματοποιείται μόνο ένας έλεγχος της αποτελεσματικότητας, αφήνει το

ενδεχόμενο τα αποτελέσματα που λαμβάνονται να ήταν τυχαία (μη αντιπροσωπευτικά για τη μεθοδολογία). Για την αντιμετώπιση του προβλήματος αυτού, χρησιμοποιήθηκε η επαναληπτική διαδικασία ελέγχου «cross-validation».

Κατά την εφαρμογή της διαδικασίας cross-validation, το σύνολο των αντικειμένων U, που αποτελείται από m αντικείμενα, χωρίζεται σε K αμοιβαίως αποκλειόμενα μικρότερου μεγέθους δείγματα U_1, U_2, …, U_K περίπου ίδιου μεγέθους d. Σε κάθε επανάληψη t αναπτύσσεται ένα μοντέλο, έχοντας ως δείγμα εκμάθησης το δείγμα U εκτός του U_t, και ως δείγμα ελέγχου το αποκλειόμενο δείγμα U_t. Συνήθως ο αριθμός των επαναλήψεων K κυμαίνεται μεταξύ του 1 και του 20. Όμως μπορεί να τεθεί ακόμα και ίσος με m (leave-one-out cross-validation). Μελέτες έδειξαν ότι μια τέτοια επιλογή μπορεί να οδηγήσει σε αποτυχία υπολογισμού της πραγματικής αποτελεσματικότητας του μοντέλου ταξινόμησης, ενώ αυξάνει και η διακύμανση των υπολογισμών [Kohavi (1995), Shao (1997), Brieman (1996)]. Αντίθετα, αν το K είναι μικρό, είναι πιθανό ο υπολογισμός του σφάλματος να είναι υπερβολικά απαισιόδοξος, λόγω της διαφοράς στο μέγεθος των δειγμάτων εκμάθησης και ελέγχου που διαμορφώνονται σε κάθε επανάληψη της διαδικασίας cross-validation. Το πρόβλημα γίνεται ακόμα πιο σημαντικό όταν το σύνολο του δείγματος είναι μικρό. Στην περίπτωση αυτή, η επιλογή ενός μικρού αριθμού επαναλήψεων, οδηγεί στη χρήση ανεπαρκών δειγμάτων για την ανάπτυξη του μοντέλου, αφού ο αριθμός των παρατηρήσεων στο σύνολο αναφοράς είναι αρκετά περιορισμένος. Βάσει των παραπάνω παρατηρήσεων, η πιο ευρέως χρησιμοποιούμενη τιμή για τον αριθμό των επαναλήψεων είναι 10, 10-fold cross-validation (Σχήμα 4.1) [Καπλάνης (2003)].

10-fold cross-validation

1. Διαχωρισμός του U σε 10 περίπου ίδιου μεγέθους δείγματα.
2. Για το δείγμα U_t, εκμάθηση της μεθόδου ταξινόμησης από τα υπόλοιπα 9 δείγματα και υπολογισμός της αποτελεσματικότητας της μεθόδου από το δείγμα ελέγχου U_t.
3. Εφαρμογή του βήματος 2 για t=1,2, ..., 10

Σχήμα 4.1. Εφαρμογή του 10-fold cross-validation στο σύνολο των αντικειμένων *U*

Από εδώ και στο εξής, όταν γίνεται αναφορά στη μεθοδολογία του cross-validation πρόκειται για την εφαρμογή του 10-fold cross-validation.

4.2.2. Τα δεδομένα της ανάλυσης

Η εφαρμογή της θεωρίας των προσεγγιστικών συνόλων (ΘΠΣ) σε συνδυασμό με την μεθοδολογία του cross-validation (CV) έγινε σε μια ομάδα διαφορετικών δεδομένων. Αυτά τα δεδομένα προέρχονται από μια πολύ γνωστή βάση δεδομένων για μηχανική μάθηση, την *UCI repository of machine learning databases* του πανεπιστημίου της Καλιφόρνια [Blake και Merz (1998)]. Από τη βάση αυτή επιλέχθηκαν δέκα σύνολα δεδομένων, τα οποία και παρουσιάζονται στον παρακάτω πίνακα μαζί με τα ιδιαίτερα χαρακτηριστικά τους.

Σύνολα δεδομένων	**Αντικείμενα**	**Υπό συνθήκη χαρακτηριστικά**	**Πλήθος κατηγοριών**	**Ποσοτικά – Ποιοτικά Χαρακτηριστικά**
Echocardiogram	131	8	2	Ποσοτικά & Ποιοτικά
Glass	214	9	6	Ποσοτικά
Heart-Disease (Pro Hungarian)	294	13	2	Ποσοτικά & Ποιοτικά
Iris	150	4	3	Ποσοτικά
Liver-Disorders	345	6	2	Ποσοτικά
Prima-Indians-Diabetes	768	8	2	Ποσοτικά
Thyroid Disease (Ann Train)	3772	21	3	Ποσοτικά & Ποιοτικά
Thyroid Disease (New-Thyroid)	215	5	3	Ποσοτικά
Tic-Tac-Toe	958	9	2	Ποιοτικά
Zoo	101	17	7	Ποσοτικά & Ποιοτικά

Πίνακας 4.1. Τα εξεταζόμενα δεδομένα

Τα χαρακτηριστικά τα οποία αναζητήθηκαν από τα παραπάνω σύνολα δεδομένων είναι: α) να περιέχουν τουλάχιστον 100 αντικείμενα, β) ο αριθμός των ελαχίστων συνόλων τους να είναι πάνω από ένα. Το μέγεθος των δεδομένων αναζητήθηκε να είναι πάνω από 100 αντικείμενα για να υπάρχει ένα ικανό πλήθος αντικειμένων κατά την εφαρμογή των ΘΠΣ και του CV, ενώ ο αριθμός των ελαχίστων συνόλων αναζητήθηκε να είναι μεγαλύτερος ή ίσος με δύο ώστε να έχει νόημα η σύγκριση διαφορετικών ελαχίστων συνόλων. Επίσης, καθώς κάποιες τιμές των συνόλων δεδομένων απουσίαζαν, χρειάσθηκε να αφαιρεθούν στήλες των υπό συνθήκη χαρακτηριστικών ή να προστεθούν κάποιες τιμές (τέτοιες ώστε να απέχουν σημαντικά από το πεδίο τιμών του χαρακτηριστικού και να αγνοούνται από το σύστημα κατά την

ανάλυση). Ακόμα, οι τιμές των ποιοτικών χαρακτηριστικών που δεν ήταν αριθμοί αντικαταστάθηκαν με αριθμούς.

Στα σύνολα δεδομένων του Πίνακα 4.1 εφαρμόσθηκε η μεθοδολογία της ΘΠΣ έτσι ώστε να βρεθούν τα ελάχιστα σύνολα. Ο εντοπισμός των ελαχίστων συνόλων έγινε με τη χρήση του λογιστικού προγράμματος «RSES» [Bazan και Szczuka (2001)]. Με την χρήση του εξαντλητικού (και άλλοτε του γενετικού) αλγορίθμου που προσφέρει το πρόγραμμα RSES, χωρίς να γίνει διακριτοποίηση των τιμών, βρέθηκαν τα ελάχιστα σύνολα του κάθε συνόλου δεδομένων (Πίνακας 4.2). Μάλιστα, σύμφωνα με τον αλγόριθμο που χρησιμοποιεί το πρόγραμμα RSES τα ελάχιστα σύνολα έχουν πάντα λιγότερα χαρακτηριστικά (έστω r) από το σύνολο των χαρακτηριστικών (έστω n).

Σύνολα δεδομένων	**Αριθμός ελαχίστων συνόλων**
Echocardiogram	22
Glass	17
Heart-Disease (Pro Hungarian)	16
Iris	4
Liver-Disorders	9
Prima-Indians-Diabetes	26
Thyroid Disease (Ann Train)	14
Thyroid Disease (New-Thyroid)	6
Tic-Tac-Toe	9
Zoo	7

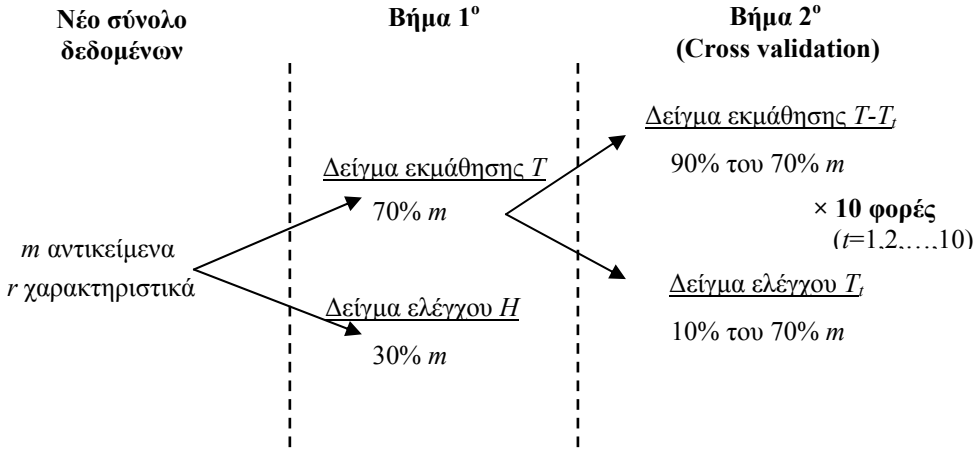
Πίνακας 4.2. Ο αριθμός των ελαχίστων συνόλων για τα εξεταζόμενα σύνολα δεδομένων

Για κάθε ελάχιστο σύνολο δημιουργήθηκε ένα νέο σύνολο δεδομένων με τα r χαρακτηριστικά του κάθε ελαχίστου συνόλου. Έτσι, π.χ. για το σύνολο δεδομένων Iris δημιουργήθηκαν 4 νέα σύνολα δεδομένων για τα 4 ελάχιστα σύνολα του. Έπειτα, οι τιμές των νέων συνόλων διακριτοποιήθηκαν με τη χρήση του προγράμματος RSES. Η περαιτέρω διαδικασία που εφαρμόστηκε αναλύεται στα παρακάτω βήματα.

Βήμα 1. Το κάθε νέο σύνολο δεδομένων χωρίζεται σε δύο υποσύνολα: α) το δείγμα εκμάθηση T (training set), που αποτελείται από το 70% των αντικειμένων, και β) το δείγμα ελέγχου H, που αποτελείται από το 30% των αντικειμένων. Ο διαχωρισμός σε αυτά τα δύο υποσύνολα γίνεται με τυχαίο τρόπο (για το σκοπό αυτό χρησιμοποιήθηκε μια γεννήτρια τυχαίων αριθμών).

Βήμα 2. Σε κάθε δείγμα εκμάθησης *T* εφαρμόζεται η μεθοδολογία του cross-validation (10-fold cross validation). Η μέθοδος αυτή: α) διασπά το κάθε δείγμα εκμάθησης *T* σε δέκα περίπου ίδιου μεγέθους δείγματα T_t, με t=1, 2, …, 10, β) χρησιμοποιείται το σύνολο T-T_t ως δείγμα εκμάθησης και το T_t ως δείγμα ελέγχου, διαδοχικά για t=1, 2, ..., 10.

Για την καλύτερη κατανόηση των δυο πρώτων βημάτων, ας θεωρηθεί ότι υπάρχει ένα νέο σύνολο δεδομένων με *m* αντικείμενα και *r* χαρακτηριστικά. Σχηματικά η διάσπαση του σύμφωνα με τα βήματα 1 και 2, απεικονίζεται στο Σχήμα 4.2.



Σχήμα 4.2. Διάσπασης του νέου συνόλου δεδομένων σε δείγματα εκμάθησης και ελέγχου

Βήμα 3. Εφαρμόζονται οι αλγόριθμοι LEM2 και LERS [Grzymala - Busse, 1992]. Για την ανάπτυξη των κανόνων της ΘΠΣ χρησιμοποιείται ο αλγόριθμος LEM2 στα δείγματα εκμάθησης T-T_t, ενώ ο LERS χρησιμοποιείται για την εφαρμογή των κανόνων της ΘΠΣ στα δείγματα ελέγχου T_t, για t=1, 2, …, 10.

Βήμα 4. Υπολογίζεται η μέση επιτυχία της ταξινόμησης στα δείγματα ελέγχου T_t.

Βήμα 5. Εφαρμόζονται, οι αλγόριθμοι LEM2 και LERS στο δείγμα εκμάθησης *T* και στο δείγμα ελέγχου *H*. Ο αλγόριθμος LEM2 χρησιμοποιείται για την ανάπτυξη των κανόνων της ΘΠΣ στο δείγμα εκμάθησης *T*, ενώ ο LERS για την εφαρμογή των κανόνων της ΘΠΣ στο δείγμα ελέγχου *H*.

Βήμα 6. Εφαρμόζονται οι αλγόριθμοι LEM2 και LERS στο σύνολο των δεδομένων με το σύνολο των χαρακτηριστικών (*n* χαρακτηριστικά). Ο αλγόριθμος LEM2 χρησιμοποιείται για την ανάπτυξη των κανόνων της ΘΠΣ στο δείγμα εκμάθησης *T*, ενώ ο LERS για την εφαρμογή των κανόνων της ΘΠΣ στο δείγμα ελέγχου *H*.

4.2.3. Παράδειγμα της διαδικασίας

Το σύνολο δεδομένων Iris αποτελείται από ένα σχετικά μικρό αριθμό αντικειμένων (150) και υπό συνθήκη χαρακτηριστικών (4)· είναι λοιπόν εύκολο να παρατεθεί στις επόμενες σελίδες και στο Παράρτημα Α η πλήρη εφαρμογή της θεωρίας των προσεγγιστικών συνόλων σε συνδυασμό με τη μεθοδολογία του cross-validation σε αυτό.

Το σύνολο δεδομένων Iris είναι ένας πίνακας με 5 στήλες και 150 γραμμές (Παράρτημα Α, Πίνακας Α.1). Αφού αντικαταστάθηκαν οι κατηγορίες ταξινόμησης με αριθμούς (Iris-setosa=1, Iris-versicolor=2 και Iris-virginica=3), ο νέος πίνακας 150×5 καταχωρήθηκε στο πρόγραμμα RSES. Το πρόγραμμα RSES υπολόγισε τα ελάχιστα σύνολα του Iris, τα οποία και είναι $RED_1=\{g_1, g_2, g_3\}$, $RED_2=\{g_1, g_3, g_4\}$, $RED_3=\{g_1, g_2, g_4\}$ και $RED_4=\{g_2, g_3, g_4\}$. Σύμφωνα με τα χαρακτηριστικά του κάθε ελάχιστου συνόλου δημιουργήθηκαν 4 νέα σύνολα δεδομένων. Το κάθε νέο σύνολο δεδομένων είναι πίνακας 150×4, με τις πρώτες 3 στήλες του να αντιστοιχούν στις τιμές που λαμβάνουν τα 150 αντικείμενα στα 3 χαρακτηριστικά του κάθε ελαχίστου συνόλου, και με την $4^{η}$ στήλη να αντιστοιχεί στην ταξινόμηση των 150 αντικειμένων. Το νέο σύνολο δεδομένων που δημιουργήθηκε από το ελάχιστο σύνολο $RED_2=\{g_1, g_3, g_4\}$ παρατίθεται στο Παράρτημα Α (Πίνακας Α.2).

Οι τιμές του Πίνακας Α.2 μέσω του προγράμματος RSES διακριτοποιήθηκαν. Ύστερα ακολουθήθηκαν τα βήματα που περιγράφηκαν στην Ενότητα 4.2.2 για κάθε νέο σύνολο δεδομένων. Ενδεικτικά παρουσιάζονται τα βήματα που εφαρμόστηκαν για το νέο σύνολο δεδομένων που προήλθε από το ελάχιστο σύνολο $RED_2=\{g_1, g_3, g_4\}$.

Βήμα 1. Το νέο σύνολο δεδομένων, που παρουσιάζεται στον Πίνακα Α.2, διασπάστηκε σε δύο υποσύνολα: α) το δείγμα εκμάθησης *T*, που αποτελείται από 90 αντικείμενα, και β) το δείγμα ελέγχου *H*, που αποτελείται από 60 αντικείμενα. Η διάσπαση αυτή παρουσιάζεται στον Πίνακα Α.3. Η $5^{η}$ στήλη του Πίνακα Α.3 αποτελεί το διαχωρισμό

των αντικειμένων στα δύο υποσύνολα. Έτσι, τα αντικείμενα που βαθμολογούνται στην 5η στήλη με 1 αποτελούν το δείγμα ελέγχου *H*, ενώ τα αντικείμενα που βαθμολογούνται με 0 αποτελούν το δείγμα εκμάθησης *T*. Η δημιουργία της 5ης στήλης έγινε με τη βοήθεια της γεννήτριας τυχαίων αριθμών του Excel.

Βήμα 2. Στα 90 αντικείμενα του δείγματος εκμάθησης *T* εφαρμόστηκε η μεθοδολογία του cross-validation. Έτσι: α) τα 90 αντικείμενα διαχωρίσθηκαν τυχαία σε 10 δείγματα T_t των 9 αντικειμένων, β) τα 9 δείγματα χρησιμοποιήθηκαν ως δείγμα εκμάθησης T-T_t και το T_t ως δείγμα ελέγχου, για t=1, 2, …, 10.

Βήμα 3. Εφαρμόστηκαν οι αλγόριθμοι LEM2 και LERS [Grzymala - Busse, 1992]. Για την ανάπτυξη των κανόνων της ΘΠΣ χρησιμοποιήθηκε ο αλγόριθμος LEM2 στο δείγμα εκμάθησης T-T_t, 81 αντικειμένων, ενώ ο LERS χρησιμοποιήθηκε για την εφαρμογή των κανόνων της ΘΠΣ στο δείγμα ελέγχου T_t, 9 αντικειμένων. Τα αποτελέσματα που λήφθηκαν μετά το βήμα 3 παρουσιάζονται στον Πίνακα 4.3. Η πρώτη στήλη φανερώνει την 100% επιτυχία της ταξινόμησης στα 10 δείγματα εκμάθησης T-T_t (με t=1, 2, …, 10). Η δεύτερη στήλη παρουσιάζει την επιτυχία της ταξινόμησης στα 10 δείγματα ελέγχου T_t. Τα ποσοστά αυτά έχουν σημαντική διαβάθμιση από 77,78% έως 100%.

Βήμα 4. Υπολογίσθηκε η μέση επιτυχία της ταξινόμησης των 10 δειγμάτων ελέγχου T_t, ως ο μέσος όρος της δεύτερης στήλης του Πίνακα 4.3 (93,3%).

RED_2	
Επιτυχία της ταξινόμησης στα δείγματα εκμάθησης T-T_t	**Επιτυχία της ταξινόμησης στα δείγματα ελέγχου T_t**
100.0%	88.9%
100.0%	100.0%
100.0%	100.0%
100.0%	100.0%
100.0%	100.0%
100.0%	100.0%
100.0%	77.8%
100.0%	77.8%
100.0%	100.0%
100.0%	88.9%

Πίνακας 4.3. Τα αποτελέσματα της εφαρμογής του βήματος 3

Βήμα 5. Εφαρμογή του αλγορίθμου LEM2 για την ανάπτυξη των κανόνων της ΘΠΣ στο δείγμα εκμάθησης *T* (90 αντικειμένων και 3 υπό συνθήκη χαρακτηριστικών), και εφαρμογή του αλγορίθμου LERS για την ταξινόμηση των αντικειμένων, σύμφωνα με τους κανόνες που αναπτύχθηκαν, στο δείγμα ελέγχου *H* (60 αντικειμένων και 3 υπό συνθήκη χαρακτηριστικών). Η επιτυχία της ταξινόμησης στο δείγμα ελέγχου *H* είναι 96,67%.

Βήμα 6. Εφαρμογή των αλγορίθμων LEM2 και LERS στο σύνολο των δεδομένων (150 αντικειμένων και 4 χαρακτηριστικών). Χρησιμοποιήθηκε ο αλγόριθμος LEM2 για την ανάπτυξη των κανόνων της ΘΠΣ στο δείγμα εκμάθησης *T* και ο LERS για την εφαρμογή των κανόνων της ΘΠΣ στο δείγμα ελέγχου *H*. Η επιτυχία της ταξινόμησης στο δείγμα ελέγχου *H* για το σύνολο των χαρακτηριστικών είναι 98,33%.

Iris		RED_1	RED_2	RED_3	RED_4
Cross - validation	1	100.0%	88.9%	88.9%	100.0%
	2	100.0%	100.0%	100.0%	100.0%
	3	88.9%	100.0%	88.9%	100.0%
	4	100.0%	100.0%	88.9%	100.0%
	5	88.9%	100.0%	100.0%	100.0%
	6	100.0%	100.0%	100.0%	100.0%
	7	66.7%	77.8%	88.9%	77.8%
	8	77.8%	77.8%	66.7%	88.9%
	9	100.0%	100.0%	77.8%	88.9%
	10	88.9%	88.9%	88.9%	88.9%
Μέση επιτυχία CV		91.1%	93.3%	88.9%	94.4%
Επιτυχία ελαχίστων συνόλων		93.3%	96.7%	95.0%	98.3%
Επιτυχία συνόλου χαρακτηριστικών		98.3%	98.3%	98.3%	98.3%

Πίνακας 4.4. Συνολικά αποτελέσματα για τα δεδομένα Iris

Στον Πίνακα 4.4 παρουσιάζονται τα αποτελέσματα που λήφθηκαν και για τα 4 ελάχιστα σύνολα του συνόλου δεδομένων Iris. Στις πρώτες δέκα γραμμές παρουσιάζεται η επιτυχία της ταξινόμησης στα δείγματα ελέγχου T_t, όπου και εφαρμόσθηκε η μεθοδολογία cross-validation. Η $11^{η}$ γραμμή παρουσιάζει την μέση επιτυχία του cross-validation. Ενώ, η $12^{η}$ γραμμή παρουσιάζει την επιτυχία της ΘΠΣ να ταξινομήσει σωστά τα αντικείμενα του δείγματος ελέγχου *H* με τη χρήση των ελαχίστων συνόλων. Η τελευταία γραμμή παρουσιάζει την επιτυχία της ταξινόμησης

μέσω της ΘΠΣ χρησιμοποιώντας την πληροφορία που δίνει το σύνολο των χαρακτηριστικών.

4.2.4. Ανάλυση αποτελεσμάτων

Με όμοιο τρόπο με αυτόν που αναλύθηκε στην προηγούμενη ενότητα για το Iris εφαρμόσθηκε η θεωρία των προσεγγιστικών συνόλων σε συνδυασμό με τη μέθοδο του cross-validation. Τα αποτελέσματα που προέκυψαν για τα σύνολα δεδομένων Echocardiogram, Glass, Heart-Disease (Pro Hungarian), Liver-Disorders, Prima-Indians-Diabetes, Thyroid Disease (Ann Train), Thyroid Disease (New-Thyroid), Tic-Tac-Toe και Zoo παρουσιάζονται στους Πίνακες 4.5-4.13. Στους πίνακες αυτούς είναι διαχωρισμένες με χρώματα κάποιες τιμές. Με το πιο σκούρο γκρι και τα λευκά γράμματα εμφανίζεται η επιτυχία της ταξινόμησης στο δείγμα ελέγχου *H* για το σύνολο των χαρακτηριστικών. Το λιγότερα έντονο γκρι χρώμα με τα μαύρα γράμματα στην $11^{η}$ γραμμή παρουσιάζει το ελάχιστο σύνολο με το υψηλότερο ποσοστό ακρίβειας στη διαδικασία CV. Τέλος, το έντονο γκρι χρώμα με τα μαύρα γράμματα παρουσιάζει τη μέγιστη επιτυχία της ταξινόμησης (ακρίβεια) που λήφθηκε με τη χρήση των ελαχίστων συνόλων.

Echocardiogram		**RED_1**	**RED_2**	**RED_3**	**RED_4**	**RED_5**	**RED_6**	**RED_7**	**RED_8**	**RED_9**	**RED_{10}**
Cross - validation	1	67%	44%	44%	56%	44%	44%	44%	78%	78%	44%
	2	100%	78%	78%	89%	67%	78%	100%	89%	100%	100%
	3	89%	89%	89%	89%	89%	78%	78%	89%	100%	78%
	4	44%	67%	67%	67%	44%	67%	56%	78%	44%	78%
	5	56%	67%	56%	67%	78%	89%	78%	89%	56%	78%
	6	78%	78%	78%	78%	89%	89%	89%	89%	78%	78%
	7	44%	44%	44%	67%	44%	44%	44%	44%	67%	44%
	8	67%	56%	78%	89%	67%	56%	100%	67%	89%	78%
	9	67%	89%	78%	67%	56%	78%	67%	67%	67%	89%
	10	78%	67%	100%	67%	67%	78%	67%	89%	100%	78%
Μέση επιτυχία CV		69%	68%	71%	73%	64%	70%	72%	78%	78%	74%
Επιτυχία ελαχίστων συνόλων		56%	61%	66%	46%	66%	59%	63%	68%	73%	54%
Επιτυχία συνόλου χαρακτηριστικών		68%	68%	68%	68%	68%	68%	68%	68%	68%	68%

Echocardiogram		RED_{11}	RED_{12}	RED_{13}	RED_{14}	RED_{15}	RED_{16}	RED_{17}	RED_{18}	RED_{19}	RED_{20}
Cross - validation	1	67%	56%	44%	44%	56%	44%	67%	44%	44%	44%
	2	78%	89%	89%	100%	78%	100%	89%	78%	89%	89%
	3	100%	78%	100%	56%	78%	89%	67%	100%	78%	67%
	4	67%	44%	78%	44%	78%	44%	67%	67%	67%	56%
	5	78%	78%	67%	67%	56%	78%	56%	78%	67%	78%
	6	78%	78%	89%	89%	78%	67%	78%	67%	78%	89%
	7	56%	22%	33%	56%	78%	44%	56%	67%	44%	44%
	8	78%	67%	67%	67%	89%	89%	67%	89%	100%	100%
	9	89%	67%	67%	67%	67%	44%	67%	56%	56%	67%
	10	89%	100%	89%	100%	89%	89%	100%	89%	100%	100%
Μέση επιτυχία CV		78%	68%	72%	69%	74%	69%	71%	73%	72%	73%
Επιτυχία ελαχίστων συνόλων		66%	61%	61%	59%	61%	66%	61%	59%	56%	71%
Επιτυχία συνόλου χαρακτηριστικών		68%	68%	68%	68%	68%	68%	68%	68%	68%	68%

Echocardiogram		RED_{21}	RED_{22}
Cross - validation	1	56%	56%
	2	89%	100%
	3	78%	78%
	4	67%	56%
	5	33%	67%
	6	89%	78%
	7	67%	33%
	8	100%	89%
	9	67%	78%
	10	89%	100%
Μέση επιτυχία CV		73%	73%
Επιτυχία ελαχίστων συνόλων		56%	66%
Επιτυχία συνόλου χαρακτηριστικών		68%	68%

Πίνακας 4.5. Συνολικά αποτελέσματα για τα δεδομένα Echocardiogram

Glass		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6	RED_7	RED_8	RED_9	RED_{10}
Cross - validation	1	71%	57%	50%	57%	57%	71%	79%	43%	36%	57%
	2	57%	43%	29%	50%	43%	57%	43%	57%	36%	57%
	3	36%	50%	50%	50%	36%	36%	29%	43%	50%	64%
	4	71%	50%	50%	50%	29%	50%	64%	57%	57%	57%
	5	79%	86%	79%	50%	57%	50%	50%	57%	79%	71%
	6	57%	86%	71%	50%	64%	50%	50%	71%	79%	57%
	7	50%	57%	57%	43%	50%	36%	57%	43%	50%	79%
	8	71%	93%	79%	93%	29%	71%	93%	71%	64%	64%
	9	71%	64%	50%	71%	43%	29%	71%	50%	64%	71%
	10	64%	64%	57%	57%	50%	21%	57%	64%	64%	71%
Μέση επιτυχία CV		63%	65%	57%	57%	46%	47%	59%	56%	58%	65%
Επιτυχία ελαχίστων συνόλων		64%	67%	70%	55%	59%	52%	69%	55%	70%	61%
Επιτυχία συνόλου χαρακτηριστικών		67%	67%	67%	67%	67%	67%	67%	67%	67%	67%

Glass		RED_{11}	RED_{12}	RED_{13}	RED_{14}	RED_{15}	RED_{16}	RED_{17}
Cross - validation	1	43%	71%	64%	86%	64%	64%	71%
	2	57%	64%	57%	71%	57%	71%	50%
	3	50%	43%	29%	43%	50%	36%	36%
	4	50%	50%	43%	36%	36%	43%	50%
	5	86%	86%	79%	86%	79%	71%	79%
	6	57%	64%	71%	64%	64%	86%	50%
	7	50%	57%	64%	64%	64%	50%	57%
	8	64%	71%	50%	64%	64%	50%	43%
	9	50%	79%	71%	64%	79%	57%	36%
	10	64%	57%	57%	71%	57%	57%	43%
Μέση επιτυχία CV		57%	64%	59%	65%	61%	59%	51%
Επιτυχία ελαχίστων συνόλων		56%	52%	56%	63%	59%	52%	50%
Επιτυχία συνόλου χαρακτηριστικών		67%	67%	67%	67%	67%	67%	67%

Πίνακας 4.6. Συνολικά αποτελέσματα για τα δεδομένα Glass

Heart-Disease (Pro Hungarian)		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6	RED_7	RED_8	RED_9	RED_{10}
Cross - validation	1	74%	79%	63%	84%	95%	74%	89%	47%	89%	89%
	2	47%	42%	37%	68%	63%	68%	74%	42%	53%	32%
	3	68%	63%	68%	89%	89%	84%	74%	74%	79%	79%
	4	58%	68%	68%	89%	79%	74%	63%	63%	84%	74%
	5	58%	32%	58%	74%	63%	84%	79%	58%	89%	63%
	6	53%	68%	42%	79%	74%	74%	74%	63%	68%	58%
	7	68%	68%	63%	74%	68%	79%	84%	42%	53%	47%
	8	68%	84%	53%	84%	74%	68%	74%	47%	84%	74%
	9	53%	42%	53%	53%	53%	74%	74%	32%	68%	58%
	10	74%	68%	58%	89%	89%	84%	68%	58%	63%	47%
Μέση επιτυχία CV		62%	62%	56%	78%	75%	76%	75%	53%	73%	62%
Επιτυχία ελαχίστων συνόλων		70%	58%	74%	85%	82%	77%	80%	65%	78%	64%
Επιτυχία συνόλου χαρακτηριστικών		80%	80%	80%	80%	80%	80%	80%	80%	80%	80%

Heart-Disease (Pro Hungarian)		RED_{11}	RED_{12}	RED_{13}	RED_{14}	RED_{15}	RED_{16}
Cross - validation	1	53%	84%	58%	84%	84%	68%
	2	53%	63%	58%	63%	58%	68%
	3	63%	63%	74%	84%	79%	74%
	4	47%	74%	63%	79%	79%	74%
	5	63%	74%	84%	84%	74%	79%
	6	68%	79%	63%	89%	84%	68%
	7	53%	74%	42%	63%	79%	58%
	8	74%	74%	79%	74%	79%	63%
	9	47%	63%	68%	79%	63%	63%
	10	68%	79%	53%	68%	74%	63%
Μέση επιτυχία CV		59%	73%	64%	77%	75%	68%
Επιτυχία ελαχίστων συνόλων		66%	74%	68%	76%	77%	73%
Επιτυχία συνόλου χαρακτηριστικών		80%	80%	80%	80%	80%	80%

Πίνακας 4.7. Συνολικά αποτελέσματα για τα δεδομένα Heart-Disease (Pro Hungarian)

Liver-Disorders		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6	RED_7	RED_8	RED_9
Cross - validation	1	46%	63%	54%	54%	71%	67%	75%	75%	58%
	2	42%	67%	58%	46%	58%	50%	83%	67%	58%
	3	54%	67%	58%	54%	50%	54%	58%	58%	46%
	4	63%	67%	38%	67%	67%	71%	50%	58%	67%
	5	63%	50%	54%	46%	50%	63%	75%	75%	54%
	6	54%	46%	75%	63%	50%	67%	50%	58%	58%
	7	67%	67%	42%	67%	58%	58%	71%	46%	67%
	8	46%	71%	63%	54%	54%	58%	75%	42%	54%
	9	42%	75%	63%	50%	46%	58%	67%	54%	67%
	10	58%	63%	67%	79%	75%	46%	71%	63%	83%
Μέση επιτυχία CV		53%	63%	57%	58%	58%	59%	67%	60%	61%
Επιτυχία ελαχίστων συνόλων		58%	65%	59%	58%	63%	59%	66%	57%	55%
Επιτυχία συνόλου χαρακτηριστικών		65%	65%	65%	65%	65%	65%	65%	65%	65%

Πίνακας 4.8. Συνολικά αποτελέσματα για τα δεδομένα Liver-Disorders

Prima-Indians-Diabetes		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6	RED_7	RED_8	RED_9	RED_{10}
Cross - validation	1	75%	69%	67%	71%	57%	67%	65%	47%	59%	69%
	2	78%	73%	69%	80%	63%	76%	76%	69%	69%	75%
	3	75%	76%	65%	69%	73%	65%	69%	59%	55%	55%
	4	55%	61%	53%	63%	45%	57%	63%	49%	51%	51%
	5	73%	71%	73%	65%	59%	75%	76%	61%	59%	69%
	6	76%	78%	69%	80%	63%	75%	67%	53%	63%	49%
	7	65%	61%	57%	67%	57%	63%	63%	65%	67%	59%
	8	71%	61%	69%	76%	65%	76%	71%	63%	67%	63%
	9	71%	69%	75%	71%	67%	69%	65%	53%	67%	61%
	10	80%	67%	71%	71%	55%	63%	69%	49%	45%	51%
Μέση επιτυχία CV		72%	68%	66%	71%	60%	68%	68%	57%	60%	60%
Επιτυχία ελαχίστων συνόλων		74%	76%	74%	74%	68%	77%	67%	66%	64%	66%
Επιτυχία συνόλου χαρακτηριστικών		73%	73%	73%	73%	73%	73%	73%	73%	73%	73%

Prima-Indians-Diabetes		RED_{11}	RED_{12}	RED_{13}	RED_{14}	RED_{15}	RED_{16}	RED_{17}	RED_{18}	RED_{19}	RED_{20}
Cross - validation	1	73%	73%	80%	63%	78%	67%	80%	67%	73%	65%
	2	75%	67%	65%	80%	69%	73%	82%	71%	76%	65%
	3	59%	61%	67%	63%	73%	78%	71%	59%	69%	65%
	4	55%	55%	55%	55%	69%	67%	57%	61%	65%	55%
	5	65%	61%	63%	65%	78%	73%	75%	75%	69%	71%
	6	71%	61%	71%	73%	75%	69%	71%	69%	69%	69%
	7	65%	63%	69%	61%	61%	63%	73%	71%	69%	67%
	8	65%	61%	65%	67%	69%	78%	73%	69%	67%	69%
	9	71%	88%	67%	67%	73%	76%	73%	69%	75%	67%
	10	57%	59%	61%	55%	78%	80%	76%	76%	63%	57%
Μέση επιτυχία CV		65%	65%	66%	65%	72%	72%	73%	68%	69%	65%
Επιτυχία ελαχίστων συνόλων		65%	69%	69%	67%	77%	78%	74%	77%	76%	69%
Επιτυχία συνόλου χαρακτηριστικών		73%	73%	73%	73%	73%	73%	73%	73%	73%	73%

Prima-Indians-Diabetes		RED_{21}	RED_{22}	RED_{23}	RED_{24}	RED_{25}	RED_{26}
Cross - validation	1	71%	61%	69%	71%	67%	67%
	2	84%	63%	67%	76%	73%	73%
	3	76%	65%	61%	63%	55%	55%
	4	71%	55%	69%	59%	63%	63%
	5	73%	61%	63%	69%	65%	65%
	6	73%	61%	75%	71%	71%	71%
	7	71%	65%	65%	69%	69%	69%
	8	69%	63%	57%	69%	61%	61%
	9	75%	59%	67%	75%	71%	71%
	10	69%	59%	63%	63%	51%	51%
Μέση επιτυχία CV		73%	61%	65%	68%	64%	64%
Επιτυχία ελαχίστων συνόλων		76%	69%	69%	69%	64%	64%
Επιτυχία συνόλου χαρακτηριστικών		73%	73%	73%	73%	73%	73%

Πίνακας 4.9. Συνολικά αποτελέσματα για τα δεδομένα Prima-Indians-Diabetes

Thyroid Disease (Ann Train)		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6	RED_7	RED_8	RED_9	RED_{10}
Cross - validation	1	96%	92%	97%	97%	98%	95%	98%	97%	95%	95%
	2	95%	89%	96%	95%	96%	91%	96%	94%	91%	91%
	3	97%	89%	96%	97%	97%	93%	97%	97%	90%	93%
	4	98%	94%	97%	96%	97%	95%	97%	97%	94%	95%
	5	96%	91%	97%	94%	95%	91%	95%	95%	89%	91%
	6	98%	92%	95%	95%	98%	95%	98%	97%	94%	95%
	7	98%	93%	98%	97%	96%	93%	97%	97%	92%	94%
	8	95%	93%	96%	97%	97%	93%	97%	96%	93%	93%
	9	97%	95%	98%	95%	96%	93%	97%	96%	92%	93%
	10	98%	93%	95%	95%	97%	93%	99%	96%	93%	94%
Μέση επιτυχία CV		97%	92%	96%	96%	97%	93%	97%	96%	92%	93%
Επιτυχία ελαχίστων συνόλων		96%	92%	96%	96%	97%	92%	96%	96%	92%	93%
Επιτυχία συνόλου χαρακτηριστικών		99%	99%	99%	99%	99%	99%	99%	99%	99%	99%

Thyroid Disease (Ann Train)		RED_{11}	RED_{12}	RED_{13}	RED_{14}
Cross - validation	1	97%	97%	97%	98%
	2	96%	95%	97%	97%
	3	97%	97%	98%	97%
	4	97%	98%	98%	98%
	5	95%	97%	97%	95%
	6	97%	97%	97%	97%
	7	98%	98%	96%	98%
	8	96%	96%	96%	97%
	9	97%	97%	98%	97%
	10	97%	97%	97%	97%
Μέση επιτυχία CV		97%	97%	97%	97%
Επιτυχία ελαχίστων συνόλων		97%	96%	97%	97%
Επιτυχία συνόλου χαρακτηριστικών		99%	99%	99%	99%

Πίνακας 4.10. Συνολικά αποτελέσματα για τα δεδομένα Thyroid Disease (Ann Train)

Thyroid Disease (New-Thyroid)		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6
Cross - validation	1	100%	93%	93%	86%	86%	79%
	2	93%	100%	93%	93%	79%	86%
	3	86%	93%	100%	79%	86%	93%
	4	100%	93%	100%	93%	86%	79%
	5	86%	100%	100%	93%	100%	86%
	6	93%	100%	100%	93%	93%	86%
	7	93%	100%	100%	71%	79%	86%
	8	93%	86%	100%	100%	86%	79%
	9	79%	100%	100%	93%	100%	100%
	10	86%	100%	100%	86%	100%	86%
Μέση επιτυχία CV		91%	96%	99%	89%	89%	86%
Επιτυχία ελαχίστων συνόλων		96%	97%	95%	92%	88%	91%
Επιτυχία συνόλου χαρακτηριστικών		100%	100%	100%	100%	100%	100%

Πίνακας 4.11. Συνολικά αποτελέσματα για τα δεδομένα Thyroid Disease (New-Thyroid)

Tic-Tac-Toe		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6	RED_7	RED_8	RED_9
Cross - validation	1	83%	88%	86%	89%	81%	94%	83%	84%	88%
	2	89%	88%	92%	91%	88%	92%	73%	89%	89%
	3	89%	84%	97%	88%	86%	88%	78%	86%	92%
	4	86%	78%	88%	89%	78%	91%	70%	80%	81%
	5	84%	91%	94%	94%	81%	94%	72%	88%	92%
	6	77%	83%	89%	89%	94%	89%	80%	88%	91%
	7	81%	86%	92%	92%	89%	91%	78%	80%	95%
	8	73%	88%	94%	84%	88%	88%	73%	78%	92%
	9	84%	89%	95%	88%	83%	95%	84%	81%	84%
	10	86%	88%	92%	89%	88%	92%	69%	89%	89%
Μέση επιτυχία CV		83%	86%	92%	89%	85%	91%	76%	84%	89%
Επιτυχία ελαχίστων συνόλων		87%	82%	92%	91%	86%	90%	75%	88%	90%
Επιτυχία συνόλου χαρακτηριστικών		99%	99%	99%	99%	99%	99%	99%	99%	99%

Πίνακας 4.12. Συνολικά αποτελέσματα για τα δεδομένα Tic-Tac-Toe

Zoo		RED_1	RED_2	RED_3	RED_4	RED_5	RED_6	RED_7
Cross - validation	1	100%	100%	86%	100%	100%	86%	86%
	2	100%	100%	100%	100%	100%	86%	100%
	3	71%	71%	57%	57%	71%	71%	71%
	4	100%	86%	86%	86%	100%	71%	86%
	5	100%	100%	100%	100%	100%	100%	100%
	6	100%	100%	71%	100%	100%	100%	100%
	7	100%	86%	86%	86%	86%	100%	86%
	8	100%	100%	100%	100%	100%	100%	100%
	9	100%	86%	100%	100%	100%	86%	100%
	10	100%	100%	100%	100%	100%	86%	100%
Μέση επιτυχία CV		97%	93%	89%	93%	96%	89%	93%
Επιτυχία ελαχίστων συνόλων		100%	97%	100%	100%	97%	94%	97%
Επιτυχία συνόλου χαρακτηριστικών		97%	97%	97%	97%	97%	97%	97%

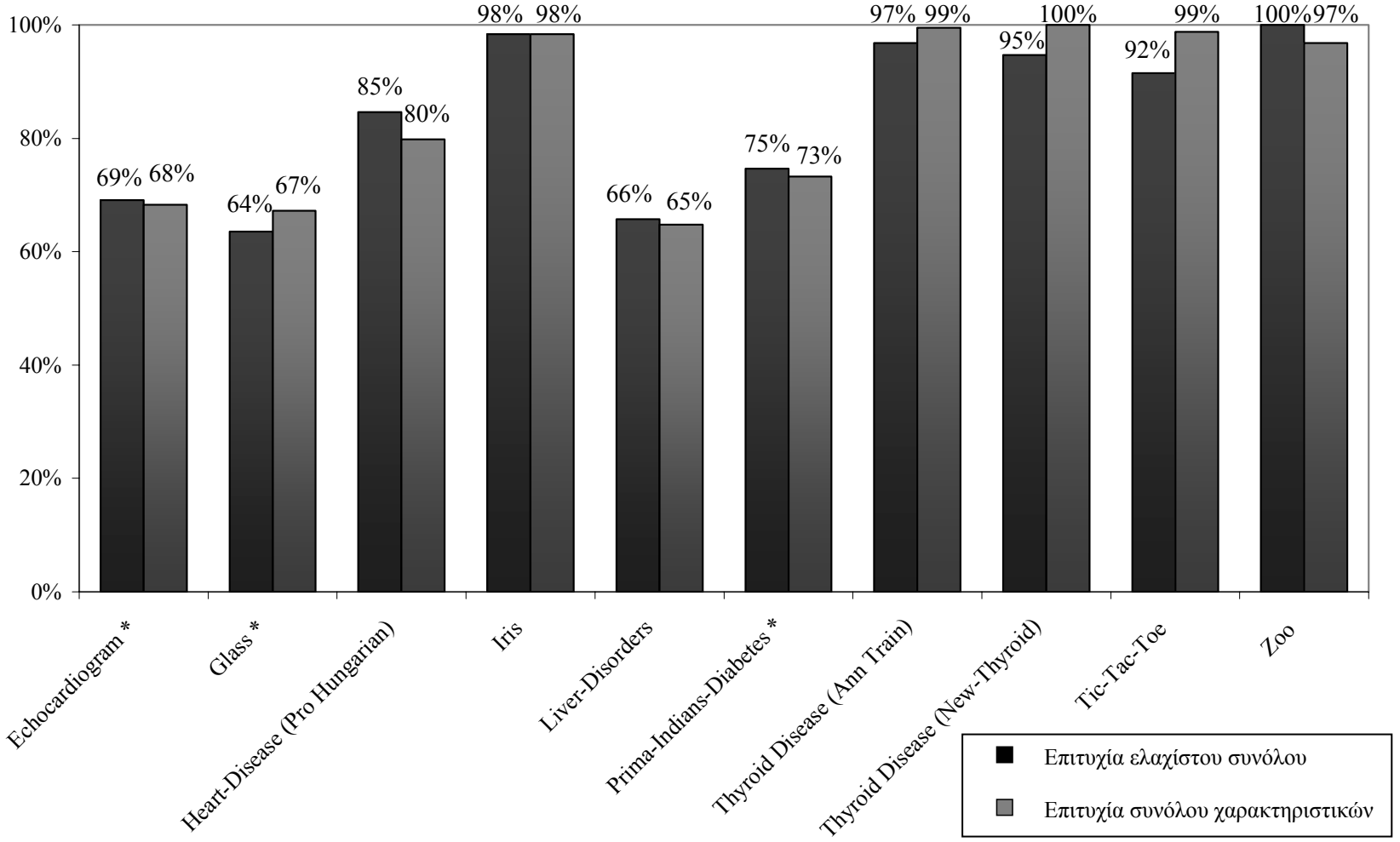
Πίνακας 4.13. Συνολικά αποτελέσματα για τα δεδομένα Zoo

Στόχος της εργασίας αυτής είναι η καλύτερη δυνατή επιτυχία της ταξινόμησης με τη χρήση ελαχίστων συνόλων (γραμμή «επιτυχία ελαχίστων συνόλων»). Ο δείκτης που χρησιμοποιήθηκε για να προσδιοριστεί το ελάχιστο σύνολο που έχει τη μεγαλύτερη δυνατή επιτυχία της ταξινόμησης είναι ο μεγαλύτερος μέσος όρος τιμών του cross-validation («μέση επιτυχία CV»). Ο δείκτης αυτός χρησιμοποιήθηκε επειδή μέσω της μεθοδολογίας του CV είναι δυνατόν να ελεγχθεί η ακρίβεια που πετυχαίνει η θεωρία των προσεγγιστικών συνόλων δέκα φορές για κάθε ελάχιστο σύνολο. Με αυτόν τον τρόπο ελέγχεται η γενικότητα των αποτελεσμάτων που λαμβάνονται με τη χρήση των ελαχίστων συνόλων.

Κύριος στόχος της εργασίας αυτής είναι να συγκρίνει δύο τιμές. Η πρώτη τιμή είναι η επιτυχία της ταξινόμησης του ελαχίστου συνόλου με τη μεγαλύτερη μέση ακρίβεια στη διαδικασία του cross-validation. Η δεύτερη τιμή, αναφέρεται στην επιτυχία της ταξινόμησης, που επιτυγχάνεται με τη χρήση του συνόλου των χαρακτηριστικών. Καθώς οι τιμές που παρουσιάζονται στους παραπάνω πίνακες είναι πολλές, στο Σχήμα 4.3 πραγματοποιείται μια απλούστερη σύνοψη των αποτελεσμάτων. Για τις περιπτώσεις όπου το ελάχιστο σύνολο με τη μεγαλύτερη μέση επιτυχία CV δεν είναι ένα, λαμβάνεται ο μέσος όρος των τιμών και τα σύνολα δεδομένων σημειώνονται με αστεράκι «*» στο Σχήμα 4.3.

Από το Σχήμα 4.3 φαίνεται ότι τα ποσοστά επιτυχίας της ταξινόμησης με τη χρήση των ελαχίστων συνόλων είναι συγκρίσιμα με αυτά του συνόλου των χαρακτηριστικών. Σε έξι περιπτώσεις, Echocardiogram, Heart-Disease (Pro Hungarian), Iris, Liver-Disorders, Prima-Indians-Diabetes, Zoo, τα αποτελέσματα που προκύπτουν από την χρήση των ελαχίστων συνόλων είναι καλύτερα ή εξίσου καλά από αυτά με τη χρήση του συνόλου των χαρακτηριστικών. Μάλιστα, για την περίπτωση του Heart-Disease (Pro Hungarian) η επιτυχία της ταξινόμησης βελτιώνεται κατά 4,81%. Ενώ, ο μέσος όρος βελτίωσης της ταξινόμησης για τις έξι περιπτώσεις είναι 1,86%. Για τα δεδομένα Glass, Thyroid Disease (Ann Train), Thyroid Disease (New-Thyroid) και Tic-Tac-Toe η επιτυχία της ταξινόμησης μειώνεται με τη χρήση των ελαχίστων συνόλων έως και 7,23%, με μέσο όρο 4,71%. Λαμβάνοντας υπόψη και τα δέκα σύνολα δεδομένων η επιτυχία της ταξινόμησης μειώνεται 7,7% κατά μέσο όρο.

Για να αναδειχθεί η αξία της χρήσης των ελαχίστων συνόλων ως μέσο μείωσης της απαραίτητης πληροφορίας αρκεί να γίνει μια σύγκριση των χαρακτηριστικών που



Σχήμα 4.3. Συγκεντρωτικά αποτελέσματα της μεθοδολογίας ΘΠΣ (RST) σε συνδυασμό με cross-validation

χρησιμοποιήθηκαν από τα ελάχιστα σύνολα που επιλέγονται με βάση την παραπάνω διαδικασία, με το σύνολο των υπό συνθήκη χαρακτηριστικών για τα σύνολα των δεδομένων. Ο Πίνακας 4.14 παρουσιάζει την σημαντική μείωση των χαρακτηριστικών. Για τις περιπτώσεις των Echocardiogram, Glass, Heart-Disease (Pro Hungarian), που παρουσιάζουν δύο ή τρία ελάχιστα σύνολα με το μέγιστο μέσο όρο τιμών του CV, λήφθηκε υπόψη το μεγαλύτερο σε πλήθος χαρακτηριστικών ελάχιστο σύνολο.

Επίσης, σημαντική παρατήρηση της επιτυχίας της μεθοδολογίας που ακολουθήθηκε είναι ότι με τη χρήση του cross-validation αποφεύχθηκαν ελάχιστα σύνολα με σημαντικά μικρότερα ποσοστά επιτυχίας στην ταξινόμηση του δείγματος ελέγχου· στην περίπτωση του Heart-Disease (Pro Hungarian) για το ελάχιστο σύνολο RED_4 είναι 85%, όμως για το ελάχιστο σύνολο RED_2 η αντίστοιχη επιτυχία φτάνει το 58%! Φαίνεται λοιπόν ότι η μεθοδολογία του cross-validation αποτελεί ένα πολύ χρήσιμο εργαλείο στα χέρια των επιχειρησιακών ερευνητών για την επιλογή του καλύτερου ελαχίστου συνόλου.

	Σύνολο υπό συνθήκη χαρακτηριστικών	**Χαρακτηριστικά ελαχίστων συνόλων**	**Ποσοστό χρήσης της πληροφορίας**
Echocardiogram	8	3	37,50%
Glass	9	3	33,33%
Heart-Disease (Pro Hungarian)	13	5	38,46%
Iris	4	3	75,00%
Liver-Disorders	6	3	50,00%
Prima-Indians-Diabetes	8	4	50,00%
Thyroid Disease (Ann Train)	21	3	14,29%
Thyroid Disease (New-Thyroid)	5	3	60,00%
Tic-Tac-Toe	9	8	88,88%
Zoo	17	1	5,89%

Πίνακας 4.14. Ο αριθμός των συνολικών χαρακτηριστικών και ο αριθμός των χαρακτηριστικών των ελαχίστων συνόλων που παρουσίαζαν τη μεγαλύτερη μέση τιμή CV

4.3. Εφαρμογή σε άλλες μεθοδολογίες ταξινόμησης

4.3.1. Γενικά

Η εφαρμογή των ελαχίστων συνόλων, παράλληλα με τη χρήση της μεθόδου του cross-validation, στις μεθοδολογίες νευρωνικά δίκτυα (ANN), αλγόριθμος του πλησιέστερου γείτονα (1NN), δέντρα ταξινόμησης και παλινδρόμησης (CART), γραμμική διακριτική ανάλυση (LDA), τετραγωνική διακριτική ανάλυση (QDA) έγινε

πάνω σε δεδομένα της βάσης δεδομένων για μηχανική μάθηση, *UCI repository*. Μάλιστα, τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι τα Echocardiogram, Glass, Heart-Disease (Pro Hungarian), Iris, Liver-Disorders, Prima-Indians-Diabetes, Thyroid Disease (Ann Train), Thyroid Disease (New-Thyroid), Tic-Tac-Toe, Zoo, για να υπάρχει άμεση σύγκριση με τα αποτελέσματα της Ενότητας 4.2.4.

Η διαδικασία που ακολουθήθηκε παρουσιάζει σημαντικές ομοιότητες με αυτή που περιγράφτηκε στην Ενότητα 4.2.2. Μελετήθηκαν τα ίδια ελάχιστα σύνολα, από τα οποία και προέκυψαν τα νέα σύνολα δεδομένων με τα r χαρακτηριστικά του κάθε ελαχίστου συνόλου. Ύστερα, ακολουθήθηκε με όμοιο τρόπο και η διαδικασία της διάσπασης των νέων συνόλων δεδομένων σε δείγματα ελέγχου T και δείγματα εκμάθησης H, όπως και η διαδικασία του cross–validation που εφαρμόσθηκε για τα δείγματα εκμάθησης T-T_t και τα δείγματα ελέγχου T_t (βήματα 1 και 2 Ενότητας 4.2.2).

Οι διαφοροποιήσεις της μεθοδολογίας που εφαρμόσθηκε σε αυτήν την ενότητα παρουσιάζονται στα αντίστοιχα βήματα 3 έως 6 της Ενότητας 4.2.2. Κατά την εφαρμογή της μεθοδολογίας αυτής της ενότητας δεν εφαρμόστηκαν οι αλγόριθμοι της θεωρίας των προσεγγιστικών συνόλων LEM2 και LERS, αλλά οι αντίστοιχοι αλγόριθμοι των μεθοδολογιών: νευρωνικά δίκτυα, αλγόριθμος του πλησιέστερου γείτονα, δέντρα ταξινόμησης και παλινδρόμησης, γραμμική διακριτική ανάλυση, τετραγωνική διακριτική ανάλυση. Οι αλγόριθμοι αυτοί χρησιμοποιήθηκαν στα δείγματα εκμάθησης T – ελέγχου H και στα δείγματα εκμάθησης T-T_t – ελέγχου T_t.

4.3.2. Ανάλυση αποτελεσμάτων

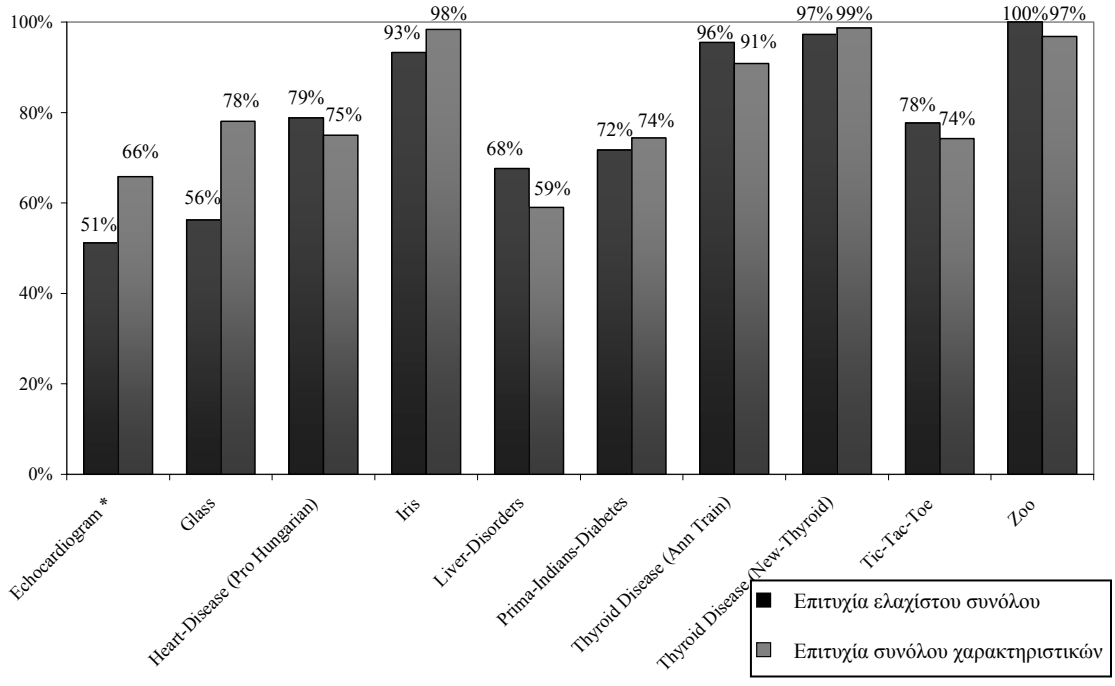
Στόχος της εφαρμογής αυτής είναι να διερευνηθεί κατά πόσο τα ελάχιστα σύνολα, που επιλέγονται με τη χρήση της μεθοδολογίας του cross-validation, είναι χρήσιμα όχι μόνο στη μεθοδολογία της θεωρίας των προσεγγιστικών συνόλων, αλλά και για άλλες μεθόδους.

Στα γραφήματα των επόμενων σελίδων παρουσιάζονται για κάθε μέθοδο που εφαρμόσθηκε τα ποσοστά επιτυχία της με τη χρήση των ελαχίστων συνόλων και με τη χρήση του συνόλου των χαρακτηριστικών.

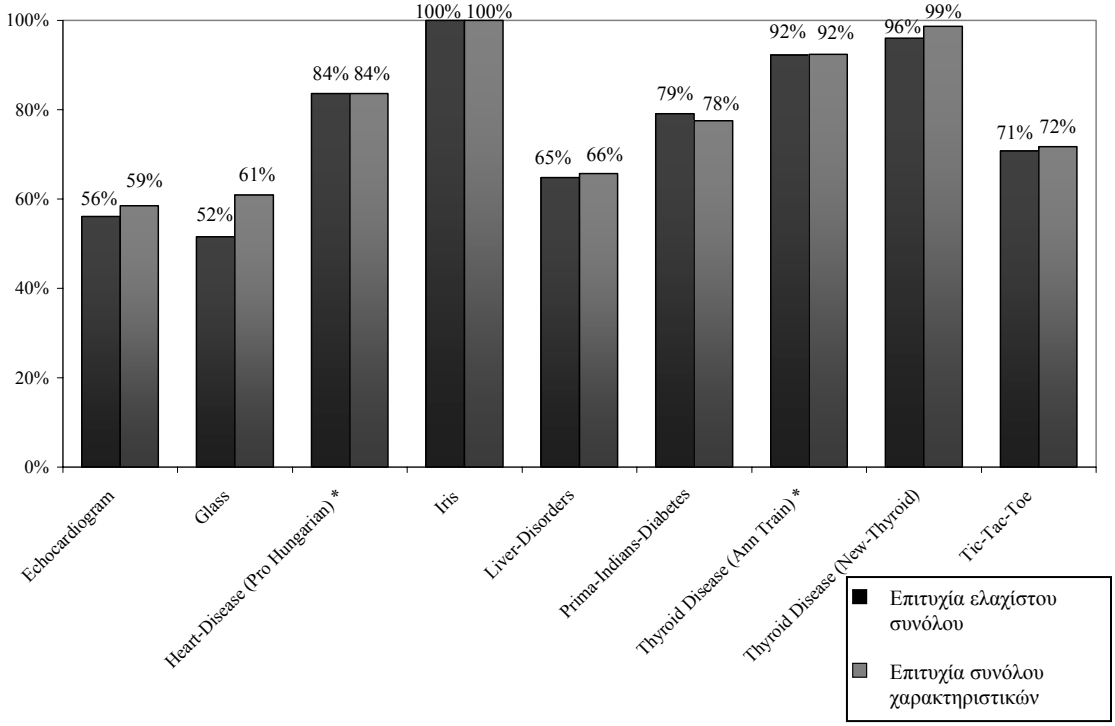
Όπως διαφαίνεται και μέσα από τα Σχήματα 4.4 - 4.8 η χρήση των ελαχίστων συνόλων επιτυγχάνει αξιόλογα αποτελέσματα σε σχέση με την χρήση του συνόλου των χαρακτηριστικών. Χαρακτηριστικά, στο Σχήμα 4.4 για το σύνολο δεδομένων Liver-

Disorders παρουσιάζεται η κατά 9% βελτίωση του ποσοστού επιτυχίας της ταξινόμησης με τη χρήση της μεθοδολογίας του πλησιέστερου γείτονα. Στον αντίποδα, στο ίδιο σχήμα, η επιτυχία της ταξινόμησης στο σύνολο δεδομένων Glass μειώνεται σημαντικά, κατά 22%, με τη χρήση των μειωμένων χαρακτηριστικών.

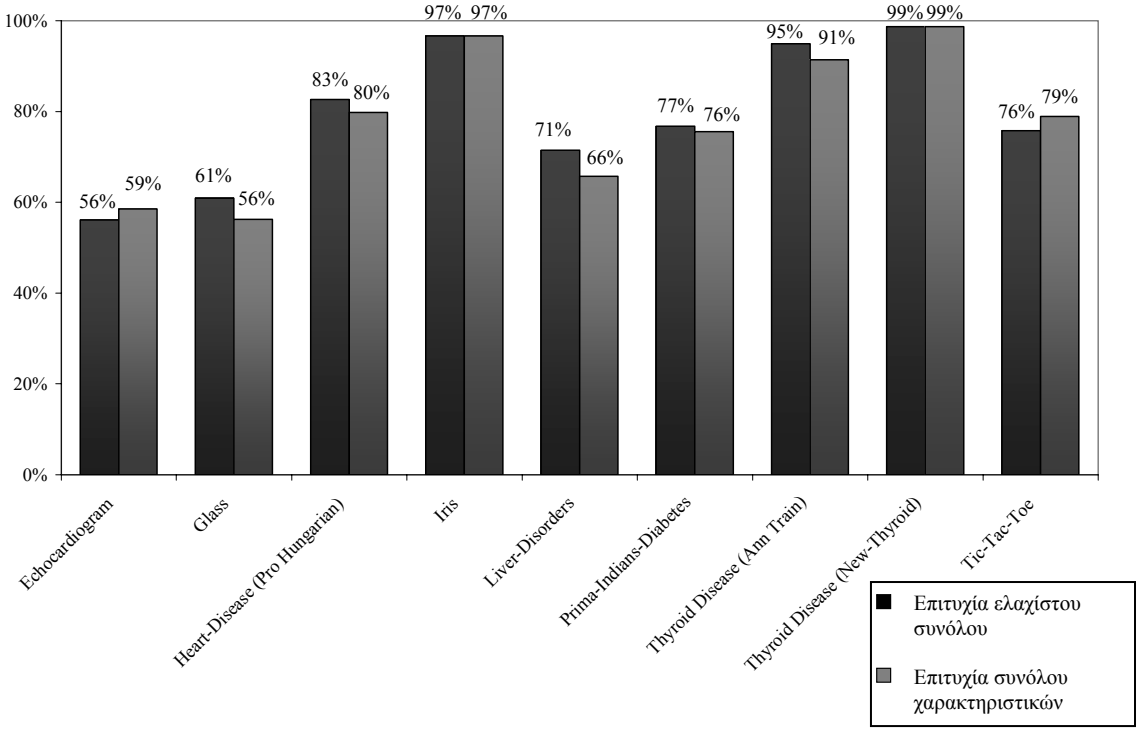
Για μια πιο εμπεριστατωμένη άποψη, σχετικά με την επιτυχία της χρήσης των ελαχίστων συνόλων, δημιουργήθηκαν οι Πίνακες 4.15 και 4.16. Στον Πίνακα 4.15 εμφανίζεται η ποσοστιαία μεταβολή της επιτυχίας της ταξινόμησης με τη χρήση των ελαχίστων συνόλων, σε συνδυασμό με τη μέθοδο του cross-validation, ως προς τη χρήση του συνόλου των χαρακτηριστικών. Σημειώνεται, πως για τις μεθόδους της διακριτικής ανάλυσης (LDA, QDA) για το σύνολο δεδομένων Zoo, δεν υπάρχουν αποτελέσματα καθώς λόγω της μορφής των δεδομένων παρουσιάστηκαν δυσκολίες στον υπολογισμό των πινάκων διακύμανσης/συνδιακύμανσης των κατηγοριών.



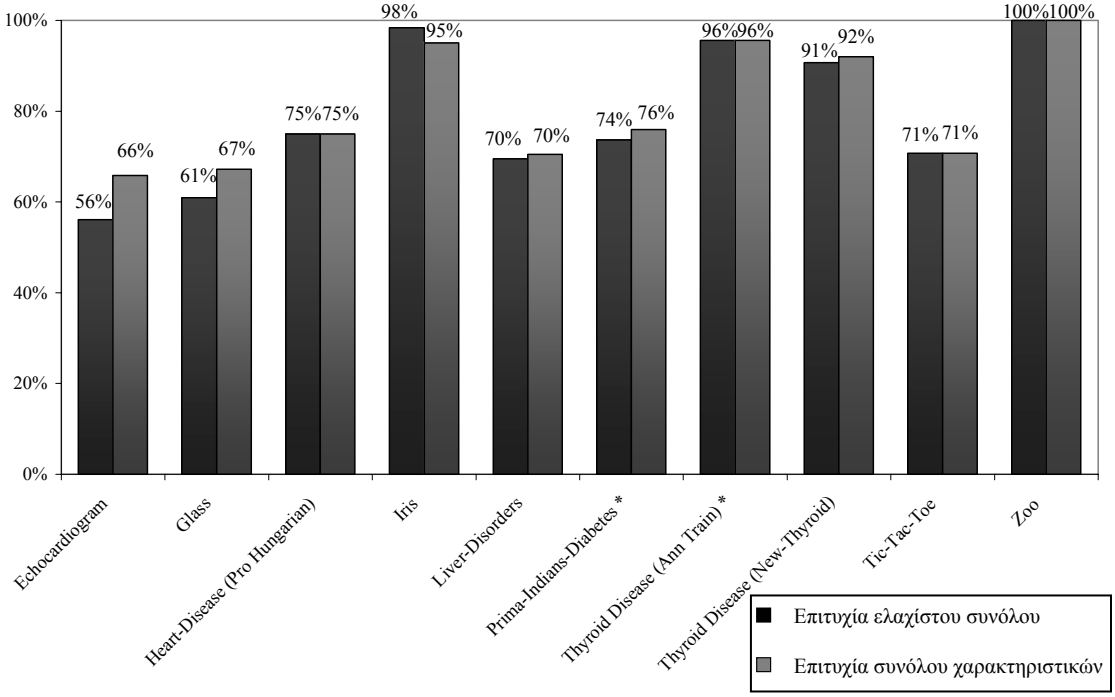
Σχήμα 4.4. Αποτελέσματα της ανάλυσης για τον αλγόριθμο του πλησιέστερου γείτονα (1NN).



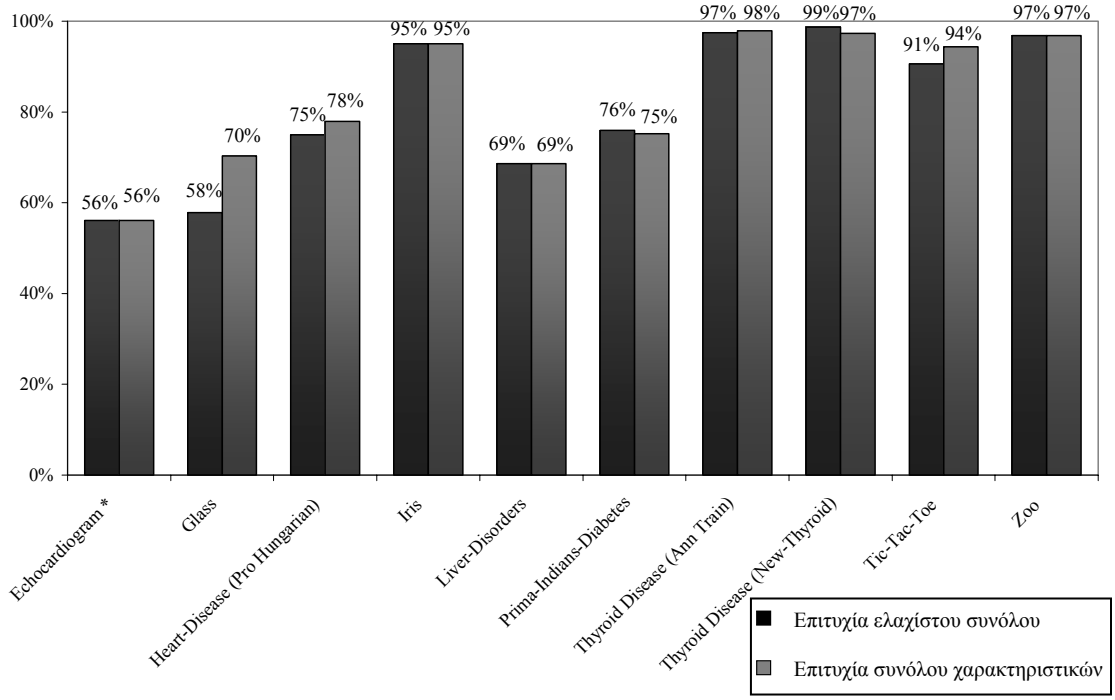
Σχήμα 4.5. Αποτελέσματα της ανάλυσης για τη γραμμική διακριτική ανάλυση (LDA).



Σχήμα 4.6. Αποτελέσματα της ανάλυσης για την τετραγωνική διακριτική ανάλυση (QDA).



Σχήμα 4.7. Αποτελέσματα της ανάλυσης για τα δέντρα ταξινόμησης και παλινδρόμησης (CART).



Σχήμα 4.8. Αποτελέσματα της ανάλυσης για τα νευρωνικά δίκτυα (ANN).

	RST	1NN	LDA	QDA	CART	ANN
Echocardiogram *	0,81%	-14,63%	-2,44%	-2,44%	-9,76%	0,00%
Glass *	-3,65%	-21,88%	-9,38%	4,69%	-6,25%	-12,50%
Heart-Disease (Pro Hungarian)	4,81%	3,85%	0,00%	2,88%	0,00%	-2,89%
Iris	0,00%	-5,00%	0,00%	0,00%	3,33%	0,00%
Liver-Disorders	0,95%	8,57%	-0,95%	5,72%	-0,95%	0,00%
Prima-Indians-Diabetes *	1,36%	-2,71%	1,55%	1,16%	-2,33%	0,77%
Thyroid Disease (Ann Train)	-2,63%	4,73%	-0,18%	3,50%	0,00%	-0,44%
Thyroid Disease (New-Thyroid)	-5,33%	-1,33%	-2,67%	0,00%	-1,33%	1,33%
Tic-Tac-Toe	-7,23%	3,46%	-0,94%	-3,15%	0,00%	-3,77%
Zoo	3,23%	3,23%	-	-	0,00%	0,00%

Πίνακας 4.15. Ποσοστιαίες μεταβολές της επιτυχίας της ταξινόμησης με τη χρήση των ελαχίστων συνόλων σε σχέση με τη χρήση του συνόλου των χαρακτηριστικών, για το σύνολο των μεθόδων

Για την καλύτερη κατανόηση των αποτελεσμάτων που παραθέτει ο Πίνακας 4.15 σχηματίστηκε ο Πίνακας 4.16. Στον τελευταίο παρουσιάζεται η μέγιστη, η ελάχιστη και η μέση τιμή της ποσοστιαίας μεταβολής της επιτυχίας της ταξινόμησης. Επίσης, στις πρώτες δέκα γραμμές, η μονάδα αντιπροσωπεύει την θετική ή μηδενική μεταβολή της επιτυχίας της ταξινόμησης, ενώ το μηδέν την αρνητική μεταβολή. Η αντιστοιχία αυτή έχει γίνει για να φανεί ξεκάθαρα σε πόσες περιπτώσεις η χρήση των ελαχίστων συνόλων επιτυγχάνει καλύτερα ή ισοδύναμα αποτελέσματα από τη χρήση του συνόλου των χαρακτηριστικών.

	RST	1NN	LDA	QDA	CART	ANN
Echocardiogram *	1	0	0	0	0	1
Glass *	0	0	0	1	0	0
Heart-Disease (Pro Hungarian)	1	1	1	1	1	0
Iris	1	0	1	1	1	1
Liver-Disorders	1	1	0	1	0	1
Prima-Indians-Diabetes *	1	0	1	1	0	1
Thyroid Disease (Ann Train)	0	1	0	1	1	0
Thyroid Disease (New-Thyroid)	0	0	0	1	0	1
Tic-Tac-Toe	0	1	0	0	1	0
Zoo	1	1	-	-	1	1
Άθροισμα	6	5	3	7	5	6
Μέγιστη αρνητική μεταβολή	-7,23%	-21,88%	-9,38%	-3,15%	-9,76%	-12,50%
Μέγιστη θετική μεταβολή	4,81%	8,57%	1,55%	5,72%	3,33%	1,33%
Μέση μεταβολή	-0,77%	-2,17%	-1,67%	1,37%	-1,73%	-1,75%

Πίνακας 4.16. Συγκεντρωτικός πίνακας επιτυχίας της χρήσης των ελαχίστων συνόλων.

Ενδιαφέρον παρουσιάζουν τα αποτελέσματα της άθροισης των μονάδων αυτών. Με αυτόν τον τρόπο διαφαίνεται πώς η χρήση των ελαχίστων συνόλων επιφέρει

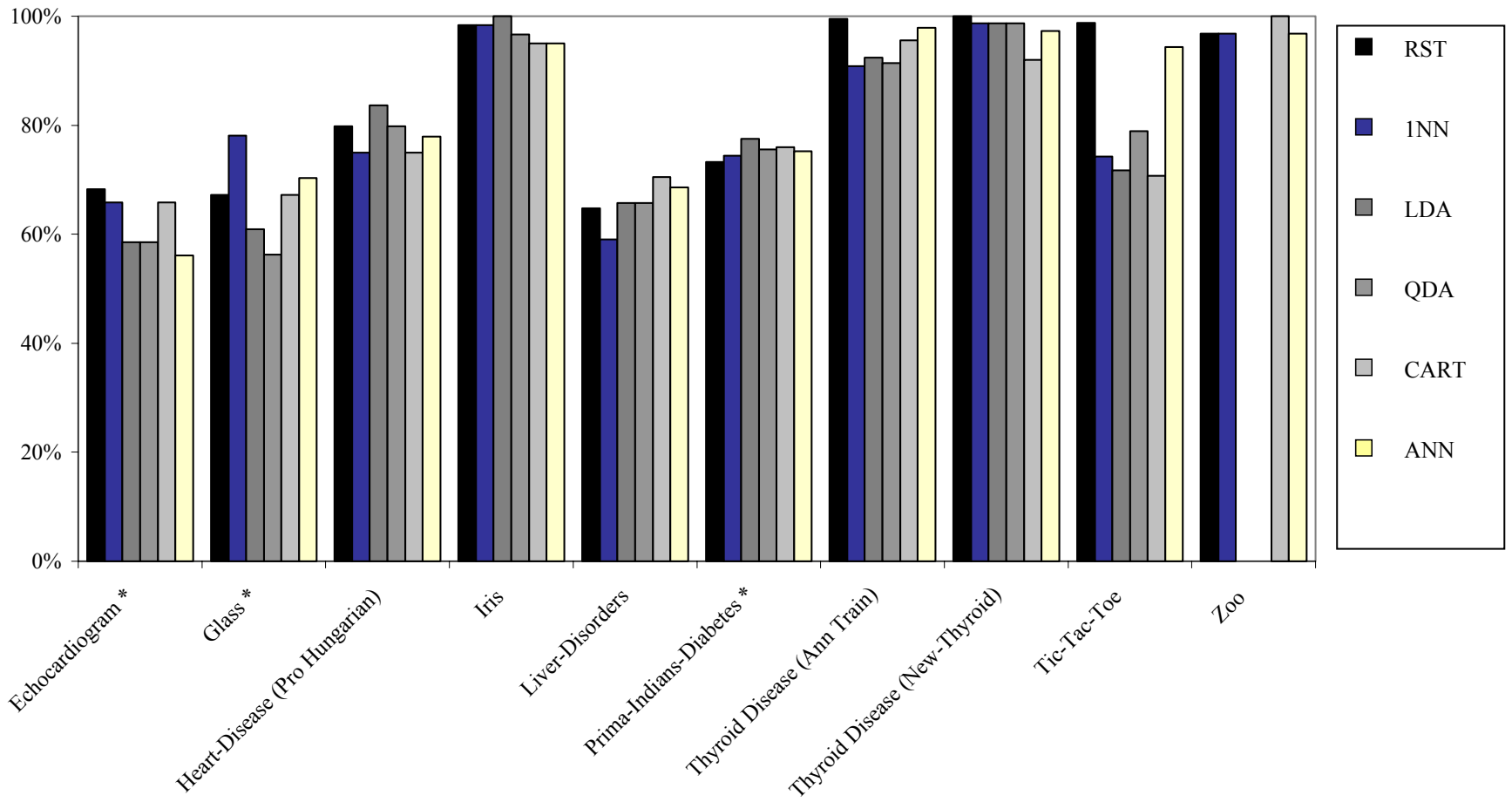
τουλάχιστον εξίσου καλά αποτελέσματα με το σύνολο των χαρακτηριστικών στην πλειοψηφία των συνόλων των δεδομένων για τις 5 από τις 6 μεθόδους (εξαιρείται η γραμμική διακριτική ανάλυση). Περισσότερο θετικά είναι τα αποτελέσματα στην περίπτωση της τετραγωνικής διακριτικής ανάλυσης, όπου στα 7 από τα 9 σύνολα δεδομένων που εφαρμόσθηκε η μέθοδος επιτεύχθηκαν καλύτερα ή εξίσου καλά αποτελέσματα.

Δύο σημαντικοί δείκτες που εμφανίζονται στον Πίνακα 4.16, είναι η μέγιστη θετική μεταβολή και η μέγιστη αρνητική μεταβολή. Είναι ευνόητο, ότι οι θετικές μεταβολές που μπορούν να εμφανιστούν με την χρήση των ελαχίστων συνόλων είναι ευπρόσδεκτες, όμως σχετικά απαγορευτικές είναι οι μεγάλες αρνητικές μεταβολές. Είναι ένα σημαντικό ζήτημα για τον αποφασίζοντα η χρήση μιας μεθοδολογίας η οποία μπορεί να του επιφέρει μέχρι και 21,88% χειρότερα αποτελέσματα (π.χ. στο Glass για τον αλγόριθμο του πλησιέστερου γείτονα). Σε αυτό το ενδεχόμενο τα καλύτερα αποτελέσματα είχαν οι μέθοδοι της τετραγωνική διακρατική ανάλυση (-3,15%) και της θεωρία των προσεγγιστικών συνόλων (-7,23%).

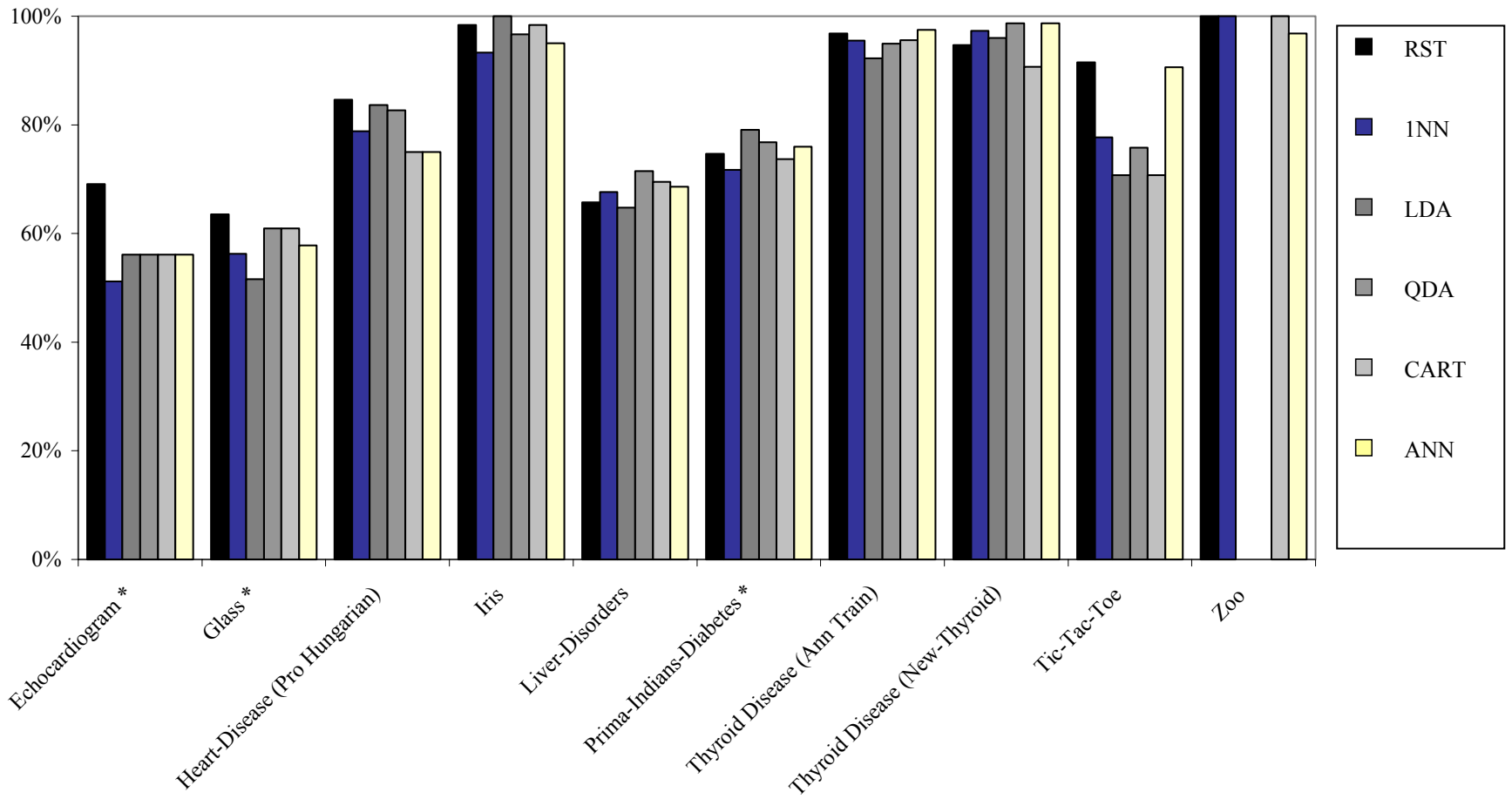
Ιδιαίτερα ενθαρρυντικά είναι τα αποτελέσματα που παρουσιάζονται στον Πίνακα 4.16 και αφορούν την μέση ποσοστιαία μεταβολή. Σύμφωνα με αυτά τα αποτελέσματα, χρησιμοποιώντας τα ελάχιστα σύνολα αντί του συνόλου των χαρακτηριστικών η επιτυχία της ταξινόμησης μειώνεται κατά μέσο όρο στο -2,17% για τον αλγόριθμο του πλησιέστερου γείτονα· ενώ για τη θεωρία των προσεγγιστικών συνόλων μόλις στο -0,77%. Χαρακτηριστική είναι η θετική κατά μέσο όρο μεταβολή στο 1,37% για την τετραγωνική διακριτική ανάλυση.

Σύμφωνα με τα στοιχεία του Πίνακα 4.16 λοιπόν, η τετραγωνική διακριτική ανάλυση παρουσιάζει σημαντικά πλεονεκτήματα κατά την χρήση των ελαχίστων συνόλων. Παράλληλα, και η μεθοδολογία της θεωρίας των προσεγγιστικών συνόλων επιτυγχάνει καλά αποτελέσματα κατά την χρήση των ελαχίστων συνόλων. Όμως, είναι σημαντικό να μελετηθεί ποιο είναι το ποσοστό επιτυχίας της ταξινόμησης που επιτυγχάνουν οι δυο αυτές μέθοδοι. Τα συγκεντρωτικά ποσοστά επιτυχίας της ταξινόμησης, για όλες τις μεθοδολογίες, παρουσιάζονται στα Σχήματα 4.8 και 4.9.

Στο Σχήμα 4.9 φαίνεται ότι η θεωρία των προσεγγιστικών συνόλων παρουσιάζει τα μεγαλύτερα ποσοστά επιτυχίας σε σχέση με τις άλλες μεθόδους, σε τέσσερα σύνολα δεδομένων Echocardiogram, Thyroid Disease (Ann Train), Thyroid Disease (New-



Σχήμα 4.9. Συγκεντρωτικά ποσοστά επιτυχίας της ταξινόμησης για τις μεθόδους που χρησιμοποιήθηκαν όσο αφορά το σύνολο των χαρακτηριστικών



Σχήμα 4.10. Συγκεντρωτικά ποσοστά επιτυχίας της ταξινόμησης για τις μεθόδους που χρησιμοποιήθηκαν όσο αφορά τα ελάχιστα σύνολα

Thyroid) και Tic-Tac-Toe. Στην περίπτωση μάλιστα που χρησιμοποιούνται τα ελάχιστα σύνολα (Σχήμα 4.10) τα σύνολα δεδομένων στα οποία η ίδια μέθοδος παρουσιάζει τα μεγαλύτερα ποσοστά επιτυχίας είναι πέντε (Echocardiogram, Glass, Heart-Disease (Pro Hungarian), Tic-Tac-Toe και Zoo). Επειδή όμως τα καλύτερα ποσοστά επιτυχίας δεν τα πετυχαίνει μια μεθοδολογία, σχηματίσθηκε ο Πίνακας 4.17 που παρουσιάζει το μέσο όρο της επιτυχία της ταξινόμησης για τις μεθόδους.

	RST	**1NN**	**LDA**	**QDA**	**CART**	**ANN**
Μ. Ο. επιτυχίας με χρήση των ελαχίστων συνόλων	84%	79%	77%	79%	79%	81%
Μ. Ο. επιτυχίας για το σύνολο των χαρακτηριστικών	85%	81%	79%	78%	81%	83%

Πίνακας 4.17 Οι μέσοι όροι επιτυχίας της ταξινόμησης για τις μεθοδολογίες που εφαρμόσθηκαν

Από τον παραπάνω πίνακα διαφαίνεται η επιτυχία της θεωρίας των προσεγγιστικών συνόλων ακόμα και χωρίς την χρήση των ελαχίστων χαρακτηριστικών. Όμως, με την χρήση των ελαχίστων συνόλων η επιτυχία της ΘΠΣ είναι σημαντικά καλύτερη από τις υπόλοιπες μεθόδους. Έτσι, εάν συγκριθεί η ΘΠΣ με την τετραγωνική διακριτική ανάλυση η διαφορά των 7 και 5 ποσοστιαίων μονάδων είναι σημαντική υπέρ της ΘΠΣ.

5. Συμπεράσματα

Πληθώρα μελετών έχει γίνει πάνω στη θεωρία των προσεγγιστικών συνόλων. Ένα μεγάλο μέρος αυτών των μελετών έχουν προσπαθήσει και έχουν πετύχει να μειώσουν τον αριθμό των κανόνων που απαιτούνται για μια επιτυχημένη ταξινόμηση. Όμως, οι προσπάθειες αυτές έχουν συναντήσει προβλήματα στην αντιμετώπιση ενός αυξημένου αριθμού ασυσχέτιστων χαρακτηριστικών. Ακόμα και ο αλγόριθμος C4.5, ένας ιδιαίτερα επιτυχημένος αλγόριθμος περικοπής των φύλλων ενός δέντρου αποφάσεων, εκφυλίζεται σημαντικά κατά την εφαρμογή του όταν εισέρχεται στο δείγμα εκμάθησης ένα ασυσχέτιστο χαρακτηριστικό και ένα θορυβώδες χαρακτηριστικό [Kohavi και Fresca (1995)].

Η θεωρία του Pawlak έχει να αντιπροτείνει τη χρήση της δυσδιακριτότητας μεταξύ των αντικειμένων με σκοπό να αναδείξει τα πλεονάζοντα και εξαρτώμενα χαρακτηριστικά. Με τη χρήση της θεωρίας των προσεγγιστικών συνόλων είναι δυνατό να δημιουργηθούν υποσύνολα χαρακτηριστικών (τα ελάχιστα σύνολα) τα οποία, προβαλλόμενα σε ένα σύνολο δεδομένων, πετυχαίνουν με υψηλά ποσοστά ακρίβειας την πρόβλεψη της ταξινόμησης. Μειώνοντας τα χαρακτηριστικά ενός συνόλου

δεδομένων είναι δυνατόν να μειωθεί αντίστοιχα ο θόρυβος και η πολυπλοκότητα της ταξινόμησης.

Στην εργασία αυτή αναλύθηκε η επιτυχία της ταξινόμησης με τη χρήση των ελαχίστων συνόλων. Σημείο αναφοράς ήταν η επιτυχία της ταξινόμησης με τη χρήση του συνόλου των χαρακτηριστικών. Ο δείκτης που χρησιμοποιήθηκε για να προβλεφθεί ποιο από τα ελάχιστα σύνολα έχει τη μεγαλύτερη δυνατότητα γενίκευσης είναι ο μέσος όρος αποτελεσμάτων μιας μεθοδολογίας που εφαρμόσθηκε στα πλαίσια του cross-validation.

Τα αποτελέσματα της εργασίας έδειξαν ότι τα ελάχιστα σύνολα μπορούν να πετύχουν εξίσου, αλλά και καλύτερα αποτελέσματα σε σχέση με την χρήση του συνόλου των χαρακτηριστικών. Είναι εφικτό λοιπόν σε ένα σύνολο δεδομένων να χρησιμοποιήσουμε ένα μέρος της διαθέσιμης πληροφορίας, μειώνοντας δραστικά το χρόνο επεξεργασίας των δεδομένων και επιτυγχάνοντας ταυτόχρονα εξίσου ικανοποιητικά αποτελέσματα.

Σημαντική βοήθεια στην επιτυχία της μεθοδολογίας που εφαρμόσθηκε προσέφερε η μεθοδολογία του cross-validation. Είναι φανερό από τα αποτελέσματα που συλλέχτηκαν ότι η θεωρία των προσεγγιστικών συνόλων αποκτά μια «φωτογραφική μνήμη» κατά την εκπαίδευση της. Όμως, στα δείγματα ελέγχου στα οποία εφαρμόσθηκε η μεθοδολογία του cross-validation διαφάνηκε η δυνατότητα γενίκευσης των κανόνων για τα διάφορα ελάχιστα σύνολα. Με αυτόν τον τρόπο αποφεύχθηκε η χρήση ελαχίστων συνόλων με ποσοστά επιτυχίας της τάξης του 58% και αντί αυτών χρησιμοποιήθηκαν ελάχιστα σύνολα με ποσοστά επιτυχίας που έφτασαν το 85%.

Παράλληλα, σε αυτή την εργασία δοκιμάστηκε η χρησιμότητα και επιτυχία της εφαρμογής των ελαχίστων συνόλων, σε συνδυασμό με τη μέθοδο του cross-validation, και σε άλλες μεθόδους: νευρωνικά δίκτυα, αλγόριθμος του πλησιέστερου γείτονα, δέντρα ταξινόμησης και παλινδρόμησης, γραμμική διακριτική ανάλυση και τετραγωνική διακριτική ανάλυση. Σύμφωνα με αυτή την εφαρμογή η χρήση των ελαχίστων συνόλων, σε μεθόδους διαφορετικές της θεωρίας των προσεγγιστικών συνόλων, επιφέρει αξιόλογα αποτελέσματα. Για τη μεθοδολογία της τετραγωνικής διακριτικής ανάλυσης μάλιστα, ο μέσος όρος βελτίωσης των ποσοστών επιτυχίας της ταξινόμησης, έναντι της χρήσης του συνόλου των χαρακτηριστικών, αγγίζει το 1,37%.

Συμπερασματικά, η εργασία αυτή αποδεικνύει ότι η θεωρία των προσεγγιστικών συνόλων αποτελεί μια επιτυχημένη μεθοδολογία μείωσης της απαραίτητης πληροφορίας στα προβλήματα της επιστήμης των αποφάσεων που αφορούν την ταξινόμηση. Καθώς η θεωρία των προσεγγιστικών συνόλων είναι από τις λίγες μεθοδολογίες που πετυχαίνουν τη μείωση των απαραίτητων υπό συνθήκη χαρακτηριστικών, ενδιαφέρον θα είχε η σύγκριση της άλλες μεθοδολογίες επιλογής χαρακτηριστικών. Επίσης, χρήσιμη μελέτη θα ήταν και αυτή που συνδυάζει τις τεχνικές μείωσης χαρακτηριστικών με τη μέθοδο bootstrap, μια τεχνική δειγματοληψίας διαφορετική της cross-validation. Παράλληλα, σημαντικές θα ήταν οι έρευνες που εφαρμόζουν τις μεθοδολογίες ταξινόμησης σε πραγματικά δεδομένα. Σε περίπτωση μάλιστα που τέτοιες εφαρμογές παρουσιάσουν επιτυχή αποτελέσματα, σημαντική προσφορά στην επιστήμη της λήψης των αποφάσεων θα είχε η ενσωμάτωση των επιτυχημένων μεθοδολογιών ταξινόμησης σε Συστήματα Υποστήριξης Αποφάσεων.

Παράρτημα Α

Παρουσίαση του συνόλου δεδομένων Iris κατά την εφαρμογή της θεωρίας των προσεγγιστικών συνόλων σε συνδυασμό με τη μέθοδο του cross-validation.

g_1	g_2	g_3	g_4	Ταξινόμηση
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa
4.3	3.0	1.1	0.1	Iris-setosa
5.8	4.0	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa
5.1	3.5	1.4	0.3	Iris-setosa
5.7	3.8	1.7	0.3	Iris-setosa
5.1	3.8	1.5	0.3	Iris-setosa
5.4	3.4	1.7	0.2	Iris-setosa
5.1	3.7	1.5	0.4	Iris-setosa
4.6	3.6	1.0	0.2	Iris-setosa
5.1	3.3	1.7	0.5	Iris-setosa
4.8	3.4	1.9	0.2	Iris-setosa
5.0	3.0	1.6	0.2	Iris-setosa
5.0	3.4	1.6	0.4	Iris-setosa

g_1	g_2	g_3	g_4	Ταξινόμηση
5.2	3.5	1.5	0.2	Iris-setosa
5.2	3.4	1.4	0.2	Iris-setosa
4.7	3.2	1.6	0.2	Iris-setosa
4.8	3.1	1.6	0.2	Iris-setosa
5.4	3.4	1.5	0.4	Iris-setosa
5.2	4.1	1.5	0.1	Iris-setosa
5.5	4.2	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.0	3.2	1.2	0.2	Iris-setosa
5.5	3.5	1.3	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
4.4	3.0	1.3	0.2	Iris-setosa
5.1	3.4	1.5	0.2	Iris-setosa
5.0	3.5	1.3	0.3	Iris-setosa
4.5	2.3	1.3	0.3	Iris-setosa
4.4	3.2	1.3	0.2	Iris-setosa
5.0	3.5	1.6	0.6	Iris-setosa
5.1	3.8	1.9	0.4	Iris-setosa
4.8	3.0	1.4	0.3	Iris-setosa
5.1	3.8	1.6	0.2	Iris-setosa
4.6	3.2	1.4	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.5	2.3	4.0	1.3	Iris-versicolor
6.5	2.8	4.6	1.5	Iris-versicolor
5.7	2.8	4.5	1.3	Iris-versicolor
6.3	3.3	4.7	1.6	Iris-versicolor
4.9	2.4	3.3	1.0	Iris-versicolor
6.6	2.9	4.6	1.3	Iris-versicolor
5.2	2.7	3.9	1.4	Iris-versicolor
5.0	2.0	3.5	1.0	Iris-versicolor
5.9	3.0	4.2	1.5	Iris-versicolor
6.0	2.2	4.0	1.0	Iris-versicolor
6.1	2.9	4.7	1.4	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3.0	4.5	1.5	Iris-versicolor
5.8	2.7	4.1	1.0	Iris-versicolor
6.2	2.2	4.5	1.5	Iris-versicolor
5.6	2.5	3.9	1.1	Iris-versicolor
5.9	3.2	4.8	1.8	Iris-versicolor
6.1	2.8	4.0	1.3	Iris-versicolor
6.3	2.5	4.9	1.5	Iris-versicolor
6.1	2.8	4.7	1.2	Iris-versicolor
6.4	2.9	4.3	1.3	Iris-versicolor
6.6	3.0	4.4	1.4	Iris-versicolor
6.8	2.8	4.8	1.4	Iris-versicolor
6.7	3.0	5.0	1.7	Iris-versicolor
6.0	2.9	4.5	1.5	Iris-versicolor
5.7	2.6	3.5	1.0	Iris-versicolor
5.5	2.4	3.8	1.1	Iris-versicolor
5.5	2.4	3.7	1.0	Iris-versicolor
5.8	2.7	3.9	1.2	Iris-versicolor
6.0	2.7	5.1	1.6	Iris-versicolor
5.4	3.0	4.5	1.5	Iris-versicolor
6.0	3.4	4.5	1.6	Iris-versicolor
6.7	3.1	4.7	1.5	Iris-versicolor
6.3	2.3	4.4	1.3	Iris-versicolor
5.6	3.0	4.1	1.3	Iris-versicolor
5.5	2.5	4.0	1.3	Iris-versicolor
5.5	2.6	4.4	1.2	Iris-versicolor

g_1	g_2	g_3	g_4	Ταξινόμηση
6.1	3.0	4.6	1.4	Iris-versicolor
5.8	2.6	4.0	1.2	Iris-versicolor
5.0	2.3	3.3	1.0	Iris-versicolor
5.6	2.7	4.2	1.3	Iris-versicolor
5.7	3.0	4.2	1.2	Iris-versicolor
5.7	2.9	4.2	1.3	Iris-versicolor
6.2	2.9	4.3	1.3	Iris-versicolor
5.1	2.5	3.0	1.1	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
7.1	3.0	5.9	2.1	Iris-virginica
6.3	2.9	5.6	1.8	Iris-virginica
6.5	3.0	5.8	2.2	Iris-virginica
7.6	3.0	6.6	2.1	Iris-virginica
4.9	2.5	4.5	1.7	Iris-virginica
7.3	2.9	6.3	1.8	Iris-virginica
6.7	2.5	5.8	1.8	Iris-virginica
7.2	3.6	6.1	2.5	Iris-virginica
6.5	3.2	5.1	2.0	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3.0	5.5	2.1	Iris-virginica
5.7	2.5	5.0	2.0	Iris-virginica
5.8	2.8	5.1	2.4	Iris-virginica
6.4	3.2	5.3	2.3	Iris-virginica
6.5	3.0	5.5	1.8	Iris-virginica
7.7	3.8	6.7	2.2	Iris-virginica
7.7	2.6	6.9	2.3	Iris-virginica
6.0	2.2	5.0	1.5	Iris-virginica
6.9	3.2	5.7	2.3	Iris-virginica
5.6	2.8	4.9	2.0	Iris-virginica
7.7	2.8	6.7	2.0	Iris-virginica
6.3	2.7	4.9	1.8	Iris-virginica
6.7	3.3	5.7	2.1	Iris-virginica
7.2	3.2	6.0	1.8	Iris-virginica
6.2	2.8	4.8	1.8	Iris-virginica
6.1	3.0	4.9	1.8	Iris-virginica
6.4	2.8	5.6	2.1	Iris-virginica
7.2	3.0	5.8	1.6	Iris-virginica
7.4	2.8	6.1	1.9	Iris-virginica
7.9	3.8	6.4	2.0	Iris-virginica
6.4	2.8	5.6	2.2	Iris-virginica
6.3	2.8	5.1	1.5	Iris-virginica
6.1	2.6	5.6	1.4	Iris-virginica
7.7	3.0	6.1	2.3	Iris-virginica
6.3	3.4	5.6	2.4	Iris-virginica
6.4	3.1	5.5	1.8	Iris-virginica
6.0	3.0	4.8	1.8	Iris-virginica
6.9	3.1	5.4	2.1	Iris-virginica
6.7	3.1	5.6	2.4	Iris-virginica
6.9	3.1	5.1	2.3	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
6.8	3.2	5.9	2.3	Iris-virginica
6.7	3.3	5.7	2.5	Iris-virginica
6.7	3.0	5.2	2.3	Iris-virginica
6.3	2.5	5.0	1.9	Iris-virginica
6.5	3.0	5.2	2.0	Iris-virginica
6.2	3.4	5.4	2.3	Iris-virginica
5.9	3.0	5.1	1.8	Iris-virginica

Πίνακας Α.1. Αρχικό σύνολο δεδομένων Iris όπως λαμβάνεται από τη βάση δεδομένων *UCI repository*

g_1	g_3	g_4	Ταξινόμηση
51	35	14	1
49	30	14	1
47	32	13	1
46	31	15	1
50	36	14	1
54	39	17	1
46	34	14	1
50	34	15	1
44	29	14	1
49	31	15	1
54	37	15	1
48	34	16	1
48	30	14	1
43	30	11	1
58	40	12	1
57	44	15	1
54	39	13	1
51	35	14	1
57	38	17	1
51	38	15	1
54	34	17	1
51	37	15	1
46	36	10	1
51	33	17	1
48	34	19	1
50	30	16	1
50	34	16	1
52	35	15	1
52	34	14	1
47	32	16	1
48	31	16	1
54	34	15	1
52	41	15	1
55	42	14	1
49	31	15	1
50	32	12	1
55	35	13	1
49	31	15	1
44	30	13	1
51	34	15	1
50	35	13	1
45	23	13	1
44	32	13	1
50	35	16	1
51	38	19	1
48	30	14	1
51	38	16	1
46	32	14	1
53	37	15	1
50	33	14	1
70	32	47	2
64	32	45	2
69	31	49	2
55	23	40	2
65	28	46	2
57	28	45	2
63	33	47	2
49	24	33	2
66	29	46	2
52	27	39	2
50	20	35	2
59	30	42	2
60	22	40	2
61	29	47	2

g_1	g_3	g_4	Ταξινόμηση
56	29	36	2
67	31	44	2
56	30	45	2
58	27	41	2
62	22	45	2
56	25	39	2
59	32	48	2
61	28	40	2
63	25	49	2
61	28	47	2
64	29	43	2
66	30	44	2
68	28	48	2
67	30	50	2
60	29	45	2
57	26	35	2
55	24	38	2
55	24	37	2
58	27	39	2
60	27	51	2
54	30	45	2
60	34	45	2
67	31	47	2
63	23	44	2
56	30	41	2
55	25	40	2
55	26	44	2
61	30	46	2
58	26	40	2
50	23	33	2
56	27	42	2
57	30	42	2
57	29	42	2
62	29	43	2
51	25	30	2
57	28	41	2
63	33	60	3
58	27	51	3
71	30	59	3
63	29	56	3
65	30	58	3
76	30	66	3
49	25	45	3
73	29	63	3
67	25	58	3
72	36	61	3
65	32	51	3
64	27	53	3
68	30	55	3
57	25	50	3
58	28	51	3
64	32	53	3
65	30	55	3
77	38	67	3
77	26	69	3
60	22	50	3
69	32	57	3
56	28	49	3
77	28	67	3
63	27	49	3
67	33	57	3
72	32	60	3
62	28	48	3
61	30	49	3

g_1	g_3	g_4	Ταξινόμηση
64	28	56	3
72	30	58	3
74	28	61	3
79	38	64	3
64	28	56	3
63	28	51	3
61	26	56	3
77	30	61	3
63	34	56	3
64	31	55	3
60	30	48	3
69	31	54	3
67	31	56	3
69	31	51	3
58	27	51	3
68	32	59	3
67	33	57	3
67	30	52	3
63	25	50	3
65	30	52	3
62	34	54	3
59	30	51	3

Πίνακας Α.2. Το νέο σύνολο δεδομένων του Iris που δημιουργήθηκε από το ελάχιστο σύνολο $RED_2=\{g_1,g_3,g_4\}$ (πριν τη διακριτοποίηση)

g_1	g_3	g_4	Ταξινόμηση	Στήλη διαχωρισμού
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	1
0	0	0	1	1
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	0
1	0	0	1	0
1	0	0	1	0
0	0	0	1	0
0	0	0	1	0
1	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	1
0	0	0	1	1
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	1
0	0	0	1	1
0	0	0	1	1

g_1	g_3	g_4	Ταξινόμηση	Στήλη διαχωρισμού
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	1	0
0	0	0	1	0
3	1	0	2	0
3	1	0	2	1
3	1	0	2	0
0	1	0	2	0
3	1	0	2	1
1	1	0	2	0
3	1	0	2	1
0	1	0	2	1
3	1	0	2	0
0	1	0	2	0
0	1	0	2	1
1	1	0	2	0
2	1	0	2	0
3	1	0	2	0
0	1	0	2	0
3	1	0	2	1
0	1	0	2	0
1	1	0	2	0
3	1	0	2	1
0	1	0	2	0
1	1	2	2	0
3	1	0	2	1
3	1	0	2	1
3	1	0	2	0
3	1	0	2	0
3	1	0	2	1
3	1	0	2	0
3	2	1	2	0
2	1	0	2	1
1	1	0	2	1
0	1	0	2	1
0	1	0	2	1
1	1	0	2	1
2	3	0	2	0
0	1	0	2	0
2	1	0	2	0
3	1	0	2	1
3	1	0	2	0
0	1	0	2	0
0	1	0	2	1
0	1	0	2	0
3	1	0	2	1
1	1	0	2	0
0	1	0	2	0
0	1	0	2	1
1	1	0	2	0
1	1	0	2	0
3	1	0	2	1
0	1	0	2	0

g_1	g_3	g_4	Ταξινόμηση	Στήλη διαχωρισμού
1	1	0	2	0
3	3	2	3	0
1	3	2	3	1
3	3	2	3	0
3	3	2	3	0
3	3	2	3	1
3	3	2	3	0
0	1	1	3	1
3	3	2	3	0
3	3	2	3	0
3	3	2	3	0
3	3	2	3	0
3	3	2	3	1
3	3	2	3	1
1	2	2	3	0
1	3	2	3	1
3	3	2	3	1
3	3	2	3	0
3	3	2	3	1
3	3	2	3	1
2	2	0	3	0
3	3	2	3	1
0	1	2	3	0
3	3	2	3	1
3	1	2	3	0
3	3	2	3	0
3	3	2	3	1
3	1	2	3	0
3	1	2	3	1
3	3	2	3	1
3	3	0	3	0
3	3	2	3	0
3	3	2	3	0
3	3	2	3	1
3	3	0	3	0
3	3	0	3	0
3	3	2	3	0
3	3	2	3	1
3	3	2	3	1
2	1	2	3	1
3	3	2	3	1
3	3	2	3	0
3	3	2	3	1
1	3	2	3	0
3	3	2	3	0
3	3	2	3	1
3	3	2	3	0
3	2	2	3	0
3	3	2	3	1
3	3	2	3	1
1	3	2	3	0

Πίνακας Α.3. Το νέο σύνολο δεδομένων του Iris που δημιουργήθηκε από το ελάχιστο σύνολο $RED_2=\{g_1,g_3,g_4\}$ (μετά τη διακριτοποίηση) μαζί με την στήλη του διαχωρισμού του σε δείγμα εκμάθηση και σε δείγμα ελέγχου

Βιβλιογραφία

Ελληνική:

Δούμπος, Μ και Ζοπουνίδης, Κ. (2001), *Πολυκριτήριες Τεχνικές Ταξινόμησης: Θεωρία και Εφαρμογές*, Κλειδάριθμος, Αθήνα.

Καπλάνης, Λ και Δούμπος, Μ. (2003), *Επίδραση των χρηματιστηριακών μεγεθών στη χρηματιστηριακή συμπεριφορά των επιχειρήσεων: Η περίπτωση του Χρηματιστηρίου Αξιών Αθηνών*, διπλωματική εργασία, Πολυτεχνείο Κρήτης, Χανιά.

Ξένη:

Archer, N.P. and Wang, S. (1993), "Application of the back propagation neural networks algorithm with monotonicity constraints for two-group classification problems", *Decision Sciences*, 24, 60-75.

Balacel, N. (2000), "Multicriteria assignment method PROAFTN: Methodology and medical application", *European Journal of Operational Research*, 125, 175-183.

Bazan, J.G. and Szczuka, M. (2001), "RSES and RSESlib – A Collection of Tools for Rough Set Computations", in: W. Ziarko and Y.Y. Yao (eds.), *Rough Sets and Current Trends in Computing*, Springer, Berlin.

Beynon, M., Curry, B. and Morgan, P. (2000), "Classification and rule induction using rough set theory", *Expert Systems*, 17, 3, 136-148.

Blake, C.L. and Merz, C.J. (1998), UCI Repository of machine learning, [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science.

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

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

Gelfand, S., Ravishankar, C. and Delp, E. (1991), "An iterative growing and pruning algorithm for classification tree design", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13/2, 163-174.

Grzymala-Busse, J.W. (1992), "LERS: A system for learning from examples based on rough sets", in: R. Slowinski (ed.), *Intelligent Decision Support. Handbook of Application and Advances of the Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht, 3-18.

Kohavi, R. and Frasca, B. (1995), "Useful Feature Subsets and Rough Set Reducts", in: T.Y. Lin and A. Wildberger (eds.), *Soft Computing: Rough Sets, Fuzzy Logic, Neural Networks, Uncertainty Management, Knowledge Discovery*, Simulation Councils, San Diego.

Krawiec, K., Slowinski, R. and Vanderpooten, D. (1998), "Learning decision rules from similarity based rough approximation", in: L. Polkowski and A. Skowron (eds.), *Rough Sets in Knowledge Discovery 2: Application, Case Studies and Software Systems*, Heidelberg: Physica, 37-54.

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

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

Patuwo, P.M., Siskos, Y. and Zopounidis, C. (1993), *Advances in Multicriteria Analysis*, Kluwer Academic Publishers, Dordrecht.

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

Pawlak, Z. (1997), "Rough sets approach to knowledge-based decision support", *European Journal of Operational Research*, 99, 48-57.

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

Slowinski, R. and Stefanowski, J. (1994), "Rough classification with valued closeness relation", in E. Diday et al (eds.), *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin, 482-488.

Slowinski, R. and Zopounidis, C. (1995), "Application of the rough set approach to evaluation of bankruptcy risk", *International Journal of Intelligent Systems in Accounting*, Finance and Management, 4, 27-41.

Smith, C. (1947), "Some examples of discrimination", *Annals of Eugenics*, 13, 27-41.

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

Subramanian, V., Hung, M.S. and Hu, M.Y. (1993), "An experimental evaluation of neural networks for classification", *Computers and Operation Research*, 20/7, 769-782.

Tsumoto, S. (1998), "Automated extraction of medical expert system rules from clinical databases on rough set theory", *Information Sciences, 112, 67-84.*

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

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