ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ
ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Διπλωματική εργασία με θέμα:

ΣΥΓΚΡΙΤΙΚΗ ΠΕΙΡΑΜΑΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕΘΟΔΩΝ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΚΑΙ ΑΛΓΟΡΙΘΜΩΝ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΑΛΕΞΑΝΔΡΑΤΟΣ ΤΗΛΕΜΑΧΟΣ

Επιβλέπων Καθηγητής:
ΜΑΤΣΑΤΣΙΝΗΣ ΝΙΚΟΛΑΟΣ

Χανιά
Μάρτιος 2008

1 ΕΙΣΑΓΩΓΗ

1.1 Εξόρυξη Δεδομένων και Μηχανική Μάθηση

Με την ψηφιακή επανάσταση, πλήθος πληροφοριών διατίθενται εύκολα και γρήγορα για επεξεργασία, αποθήκευση, διανομή και μετάδοση. Η ταυτόχρονη αλματώδης πρόοδος στην επιστήμη των υπολογιστών και σε άλλες παρεμφερείς τεχνολογίες σε συνδυασμό με την όλο και πιο διαδεδομένη χρήση τους σε όλες τις πτυχές της ζωής, για παράδειγμα το World Wide Web, οδήγησε στη συλλογή και αποθήκευση τεράστιου όγκου δεδομένων σε πολλές και μεγάλες βάσεις δεδομένων. Το ποσό της συγκεντρωμένης πληροφορίας εκτιμάται ότι διπλασιάζεται κάθε είκοσι μήνες [Piatetsky *et al*.1991]

Η ταχύτατη αυτή αύξηση του όγκου των δεδομένων έχει υπερβεί την ανθρώπινη δυνατότητα για κατανόηση χωρίς τη χρήση ισχυρών υπολογιστικών εργαλείων. Έτσι, οι βάσεις στις οποίες αποθηκεύονται γίνονται "τάφοι" δεδομένων που σπάνια δέχονται επισκέψεις. Βασική αιτία για την οποία η εκμετάλλευση αυτών των δεδομένων έχει ελκύσει το ενδιαφέρον της βιομηχανίας της πληροφορίας τα τελευταία χρόνια είναι ακριβώς αυτή η ευρεία διάθεσή τους και η επακόλουθη ανάγκη μετατροπής τους σε χρήσιμες πληροφορίες. Η εξόρυξη γνώσης μέσα από αυτά είναι μία ακόμα πρόκληση. Έχουμε λοιπόν την ανάγκη για ανάπτυξη εργαλείων που αυτόματα θα επεξεργάζονται και θα αναλύουν μεγάλες ποσότητες δεδομένων παράγοντας χρήσιμη γνώση. Τέτοια εργαλεία προσφέρει η μηχανική μάθηση (machine learning), ένας κλάδος του γενικότερου επιστημονικού χώρου της τεχνητής νοημοσύνης.

Ο άνθρωπος προσπαθεί να κατανοήσει το περιβάλλον του παρατηρώντας το και δημιουργώντας μια απλοποιημένη (αφαιρετική) εκδοχή του που ονομάζεται μοντέλο (model). Η δημιουργία ενός τέτοιου μοντέλου, ονομάζεται επαγωγική μάθηση (inductive learning) ενώ η διαδικασία γενικότερα ονομάζεται επαγωγή (induction). Επιπλέον ο άνθρωπος έχει τη δυνατότητα να οργανώνει και να συσχετίζει τις εμπειρίες και τις παραστάσεις του δημιουργώντας νέες δομές που ονομάζονται πρότυπα (patterns). Η δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων, από ένα υπολογιστικό σύστημα, ονομάζεται μηχανική μάθηση (machine learning).

Διάφοροι ορισμοί:

- Carbonell (1987), "... η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης".

- Mitchell (1997), "Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E".

- Witten & Frank (2000), "Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον".

Ως *εξόρυξη γνώσης* (data mining) ορίζεται [*Tan, Steinbach and Kumar, 2005*] η διαδικασία εξαγωγής υποκρύπτουσας, προηγουμένως άγνωστης και πιθανώς χρήσιμης πληροφορίας από δεδομένα ή αλλιώς η διερεύνηση και ανάλυση, με τη βοήθεια αυτόματων ή ημι-αυτόματων μέσων, μεγάλου όγκου δεδομένων με στόχο την ανακάλυψη χρήσιμων προτύπων. Η μηχανική μάθηση αποτέλεσε κινητήριο μοχλό για την ανάπτυξη της παραπάνω περιοχής και αναπόσπαστο τμήμα της. Το ερευνητικό πεδίο αποτελεί τομή μεθόδων και εργαλείων που πηγάζουν από:

- Στατιστική

- Μηχανική Μάθηση

- Βάσεις & αποθήκες δεδομένων

Σύμφωνα με τους Weiss και Indurkhya [Weiss *et al.* 1998] δύο είναι οι μεγάλοι στόχοι των προβλημάτων με τα οποία ασχολείται η μηχανική μάθηση και ως προέκτασή της η Εξόρυξη Δεδομένων:

1. Η πρόβλεψη

2. Η ανακάλυψη γνώσης

Συνήθως η ανακάλυψη γνώσης αποτελεί ένα πρότερο στάδιο της πρόβλεψης, στις περιπτώσεις όπου τα δεδομένα είναι ακατάλληλα για πρόβλεψη. Επίσης δρα συμπληρωματικά στην πρόβλεψη, όντας εγγύτερα στην υποστήριξη αποφάσεων παρά στη λήψη αποφάσεων. Από την άλλη πλευρά η πρόβλεψη είναι απαραίτητη σε πολλές εφαρμογές (π.χ. για τη μετεωρολογία, τη σεισμολογία, την ιατρική, τις οικονομικές επιστήμες) ακόμη και όταν λειτουργεί ως μαύρο κουτί, όταν δηλαδή δεν παρέχει πληροφορίες που να την υποστηρίζουν και να την δικαιολογούν.

1.2 Τεχνικές Μηχανικής Μάθησης - Ταξινόμηση

Είναι πολλά τα εργαλεία που χρησιμοποιούνται στο χώρο της μηχανικής μάθησης και η επιλογή ενός -ή περισσότερων- εξ' αυτών για την εξόρυξη γνώσης από μια εξεταζόμενη βάση δεδομένων αποτελεί από μόνο του ένα πρόβλημα. Οι βασικοί παράγοντες στους οποίους βασίζεται η επιλογή αυτή είναι:

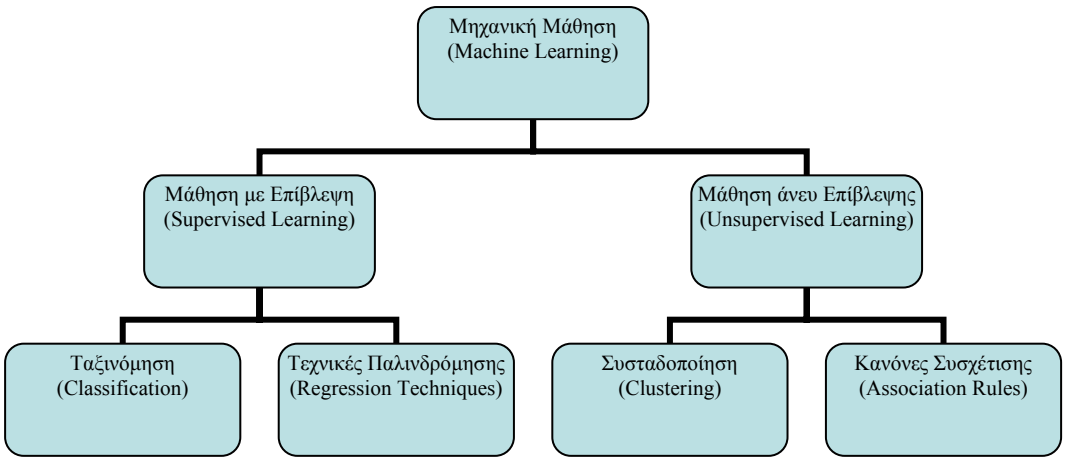
1. Η φύση των παρεχόμενων δεδομένων.

2. Το είδος της πληροφορίας που αναζητά ο χρήστης.

Ένα σετ δεδομένων, αποτελείται από ένα πλήθος στιγμιότυπων, το καθένα από τα οποία αξιολογεί με έναν συγκεκριμένο τρόπο μια σειρά από χαρακτηριστικά. Η επιλογή του κατάλληλου εργαλείου εξαρτάται άμεσα από την ποσότητα των διατιθέμενων στιγμιότυπων, από τη μορφή που έχουν τα χαρακτηριστικά που το αποτελούν, αν δηλαδή λαμβάνουν ονομαστικές ή αριθμητικές τιμές. Σημαντικό ρόλο παίζει επίσης το αν περιλαμβάνει κάποιο χαρακτηριστικό-ταξινόμησης, δηλαδή ένα χαρακτηριστικό που

κατανέμει τα στιγμιότυπα σε κάποιες προκαθορισμένες από το πρόβλημα κατηγορίες και αν αυτό εκφράζεται με ονομαστικές ή συνεχείς αριθμητικές τιμές.

Τα δεδομένα ενός προβλήματος εξόρυξης δεδομένων καθορίζουν σε μεγάλο βαθμό και το είδος της πληροφορίας που μπορεί να εξαχθεί από αυτό. Στην περίπτωση που δεν υπάρχει χαρακτηριστικό ταξινόμησης οι πιο διαδεδομένες πρακτικές είναι η συσταδοποίηση (clustering), δηλαδή η δημιουργία ομάδων από στιγμιότυπα που οι τιμές που λαμβάνουν παρεμφερείς τιμές σε κάποια από τα χαρακτηριστικά, ή η δημιουργία κανόνων συσχέτισης (association rules), δηλαδή κανόνων της μορφής «αιτίου-αποτελέσματος» (ή αλλιώς της μορφής IF-THEN). Οι τεχνικές αυτές που δεν περιλαμβάνουν ταξινόμηση στιγμιότυπων ονομάζονται και «*Μάθηση άνευ Επίβλεψης*» (unsupervised learning). Στην περίπτωση που υπάρχει χαρακτηριστικό ταξινόμησης (ή αλλιώς κλάση) στόχος της μάθησης είναι η εύρεση ενός προτύπου, βασισμένου στα υπάρχοντα δεδομένα, για την ταξινόμηση, βάσει των υπολοίπων χαρακτηριστικών, ενός αταξινόμητου στιγμιότυπου. Στην περίπτωση συνεχούς αριθμητικής κλάσης, το αντικείμενο είναι η εύρεση ενός προτύπου πρόβλεψης της κλάσης ενός αταξινόμητου στιγμιότυπου και για το σκοπό αυτό χρησιμοποιούνται τεχνικές παλινδρόμησης γραμμικής ή μη. Οι παραπάνω τεχνικές που περιγράφεται παραπάνω ορίζονται στην Μηχανική Μάθηση, ως *Μάθηση υπό Επίβλεψη* (Supervised Learning).



Σχήμα 1.1 Σχηματική αναπαράσταση των τεχνικών Μηχανικής Μάθησης

Η ταξινόμηση σε διακριτές, ονομαστικές κλάσεις αποτελεί την πιο διαδεδομένη τεχνική της μηχανικής μάθησης και για τη διενέργεια της έχουν αναπτυχθεί μια σειρά από τεχνικές. Μερικές από τις πιο διαδεδομένες, για την περίπτωση που η κλάση λαμβάνει διακεκριμένες τιμές, είναι τα νευρωνικά δίκτυα, τα δέντρα απόφασης, οι κανόνες απόφασης, οι τεχνικές που χρησιμοποιούν τις πιθανότητες (και πιο συγκεκριμένα το νόμο του Bayes), οι τεχνικές μάθησης βασισμένης στα στιγμιότυπα (instance-based learning) και πολλοί ακόμα υβριδικοί συνδυασμοί αυτών.

Στην πορεία της εργασίας αυτής θα παρουσιαστούν μια σειρά από τεχνικές ταξινόμησης οι οποίες θα χρησιμοποιηθούν στο πρακτικό κομμάτι της εργασίας για την ταξινόμηση ενός σετ δεδομένων αποτελούμενου από 14 χαρακτηριστικά, ονομαστικά και αριθμητικά, σε δύο διακριτές κλάσεις. Πιο συγκεκριμένα θα εξεταστούν:

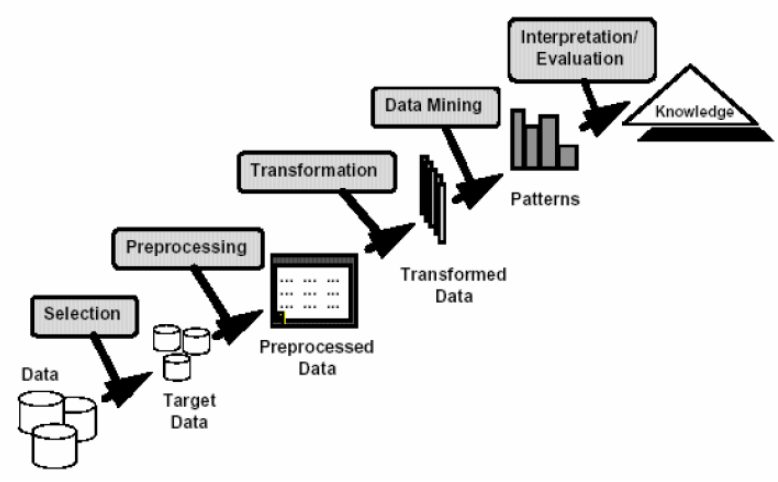
- Δέντρα απόφασης, και πιο συγκεκριμένα οι αλγόριθμοι J.48 (ή C4.5) και NBTree
- Κανόνες απόφασης, με αλγόριθμους όπως ο PART και ο RIDDOR
- Ο αλγόριθμος Naïve Bayes ο οποίος επιτυγχάνει την ταξινόμηση μέσω της χρήσης του νόμου του Bayes
- Ο αλγόριθμος μάθησης «κ-Κοντινότεροι Γείτονες» (k-Nearest Neighbor) ή IBk που βασίζεται στη μάθηση από στιγμιότυπα.

Όλοι οι παραπάνω αλγόριθμοι μπορούν να χειριστούν σετ δεδομένων μεγάλου μεγέθους όπως αυτό που θα εξεταστεί, μπορούν να χειριστούν περιπτώσεις στιγμιότυπων με ασυμπλήρωτες τιμές (ή απολεσθέντες τιμές όπως συνήθως αναφέρονται) και έχουν τη δυνατότητα να επεξεργαστούν τόσο συνεχείς όσο και ονομαστικές τιμές χαρακτηριστικών. Ωστόσο αυτό που αναζητείται είναι ποια από τις προαναφερθείσες μεθόδους έχει τις μεγαλύτερες πιθανότητες να επιτύχει την σωστή ταξινόμηση ενός αταξινόμητου στιγμιότυπου. Η αξιολόγηση της λειτουργίας ενός ταξινομητή και η σύγκριση αυτών μεταξύ τους γίνεται εφικτή με το χωρισμό του σετ δεδομένων σε σετ εκπαίδευσης, το οποίο χρησιμοποιείται για την δημιουργία του σχήματος ταξινόμησης και σε σετ ελέγχου, το οποίο χρησιμοποιείται για τον έλεγχο του σχήματος αυτού. Έτσι

γίνεται εφικτή η σύγκριση της απόδοσης των προαναφερθέντων αλγορίθμων ώστε να διαπιστωθεί ποια από τις παραπάνω μεθόδους είναι η πιο κατάλληλη για το ελεγχόμενο σετ δεδομένων.

1.3 Η Επιλογή Χαρακτηριστικών και η Σημασία της για την Εξόρυξη Δεδομένων

Ωστόσο, η επίτευξη πετυχημένου data mining προϋποθέτει περισσότερα από την απλή επιλογή ενός αλγόριθμου και την εφαρμογή του στα δεδομένα που διαθέτουμε. Η τεχνητή νοημοσύνη των αλγορίθμων πολλές φορές δοκιμάζεται από την ποιότητα των δεδομένων που εισάγονται προς επεξεργασία και είναι δεδομένος ο κίνδυνος κάποια ελαττωματικά δεδομένα να τον οδηγήσουν σε λάθη. Όπως γίνεται κατανοητό λοιπόν, πέραν της επιλογής του αλγορίθμου και των παραμέτρων που θα οριστούν σ' αυτόν, είναι και άλλες ενέργειες που μπορούν να βελτιώσουν ουσιαστικά την εφαρμογή τεχνικών μηχανικής εκμάθησης σε ένα πρακτικό πρόβλημα εξόρυξης γνώσης. Τεχνικές που αποτελούν ένα είδος μηχανικής δεδομένων (data engineering), δηλαδή διαμόρφωση των υπό-εισαγωγή δεδομένων σε μια μορφή πιο κατάλληλη για το σχήμα μάθησης που έχει επιλεγεί ώστε το μοντέλο που θα προκύψει να είναι πιο αποτελεσματικό και πιο έγκυρο.



Σχήμα 1.2 Σχηματική αναπαράσταση της διαδικασίας της μάθησης (47)

Οι τεχνικές αυτές προ-επεξεργασίας κάποιες φορές λειτουργούν ενώ κάποιες άλλες όχι, επιβεβαιώνοντας την άποψη ότι το data mining είναι ένας χώρος στον οποίο οι δοκιμές και τα λάθη αποτελούν τον πιο αξιόπιστο οδηγό.

Η Επιλογή Χαρακτηριστικών (feature selection ή attribute selection) είναι μια διαδικασία προ-επεξεργασίας, που χρησιμοποιείται ευρέως στο χώρο της μηχανικής μάθησης, κατά την οποία μια υποομάδα από τα αρχικά διαθέσιμα χαρακτηριστικά επιλέγεται, με κάποια κριτήρια, για την απώτερη επεξεργασία της από κάποιον αλγόριθμο μηχανικής μάθησης. Η επιλογή χαρακτηριστικών παρέχει πολλά πλεονεκτήματα στην διαδικασία ταξινόμησης. Μειώνει τον αριθμό χαρακτηριστικών, αποκρίνει τα μη σχετικά, τα πλεονάζοντα ή τα θορυβώδη δεδομένα και έχει άμεσα αποτελέσματα τόσο στην επιτάχυνση των αλγορίθμων μάθησης όσο και στην βελτίωση των αποτελεσμάτων τους. Βελτιώνει την ευστοχία ταξινόμησης σε νέα δεδομένα και παρέχει πιο συμπαγή αποτελέσματα καθιστώντας ευκολότερη την κατανόηση και την ερμηνεία του αντικειμένου της μάθησης. Μια τυπική διαδικασία Επιλογής Χαρακτηριστικών αποτελείται από τέσσερα βασικά βήματα:

1. δημιουργία υποομάδας,
2. αξιολόγηση υποομάδας,
3. κριτήριο τερματισμού και
4. επιβεβαίωση αποτελέσματος.

Ανάλογα με τα κριτήρια δημιουργίας και αξιολόγησης υποομάδας, οι αλγόριθμοι επιλογής χαρακτηριστικών χωρίζονται στις παρακάτω δύο κατηγορίες:

- Στους αλγόριθμους ενσωμάτωσης (wrapper algorithms)
- Στους αλγόριθμους διήθησης ή φίλτρου (filter algorithms)

Στην πορεία της εργασίας θα παρουσιαστούν μια σειρά τέτοιων αλγορίθμων οι οποίοι στο πρακτικό μέρος της εργασίας θα βοηθήσουν στην επιλογή μιας καλύτερης –εφ’ όσον αυτή υφίσταται- υποομάδας χαρακτηριστικών η οποία να βελτιώνει το ποσοστό εύστοχων ταξινομήσεων των αλγόριθμων ταξινόμησης στο εξεταζόμενο σετ δεδομένων.

1.4 Σκοπός και Δομή της Εργασίας

Βασικά συστατικά για την πετυχημένη ταξινόμηση και τη μέγιστη εξαγωγή γνώσης από ένα σετ δεδομένων είναι α) η ποιότητα των δεδομένων αυτών, και πιο συγκεκριμένα των χαρακτηριστικών που θα χρησιμοποιηθούν ως δεδομένα εισόδου στους αλγόριθμους ταξινόμησης β) η επιλογή του κατάλληλου αλγόριθμου ταξινόμησης ο οποίος ταξινομεί καλύτερα το εξεταζόμενο σετ δεδομένων. Στόχος της διπλωματικής αυτής είναι η διερεύνηση μιας σειράς διαφορετικών τεχνικών από τον χώρο της μηχανικής μάθησης οι οποίες καθιστούν δυνατή την αξιολόγηση και την κατάταξη των χαρακτηριστικών ενός, μεγάλου σε μέγεθος, σετ δεδομένων. Η βελτίωση της ποιότητας των εισαγόμενων προς ταξινόμηση δεδομένων επιτυγχάνεται μέσα από την εξάλειψη ή την τροποποίηση χαρακτηριστικών του και την δημιουργία μιας υποομάδας από τα αρχικά χαρακτηριστικά. Ωστόσο ο μόνος τρόπος για να ποσοτικοποιηθεί και να ελεγχθεί η βελτίωση της ποιότητας μιας υποομάδας δεδομένων είναι η παρατήρηση της αυξομείωσης του ποσοστού εσφαλμένων ταξινομήσεων σε κάποιον αλγόριθμο ταξινόμησης. Για τον σκοπό αυτό χρησιμοποιούνται μια σειρά διαφορετικών μεθόδων ταξινόμησης και παράλληλα με την αναζήτηση του βέλτιστου σετ δεδομένων αναζητείται και η μέθοδος που ταξινομεί πιο αποτελεσματικά την εξεταζόμενη συλλογή δεδομένων (dataset). Το σετ δεδομένων το οποίο χρησιμοποιείται για την εργασία αυτή είναι το ADULT dataset το οποίο προσφέρεται από την Αμερικανική Υπηρεσία Απογραφής (U.S. Census Bureau).

Η δομή της εργασίας είναι η εξής:

Στο κεφάλαιο 1 γίνεται μια εισαγωγή στο χώρο της μηχανικής μάθησης και της επιλογής χαρακτηριστικών καθώς και παρουσίαση του στόχου και της δομής της εργασίας

Στο κεφάλαιο 2 παρουσιάζεται η διαδικασία της ταξινόμησης μέσα από το χώρο της μηχανικής μάθησης, παρουσιάζονται τα δομικά κομμάτια ενός σχήματος ταξινόμησης και τέλος παρουσιάζονται οι 6 αλγόριθμοι που θα χρησιμοποιηθούν στην πορεία της εργασίας καθώς και το θεωρητικό υπόβαθρο στο οποίο καθένας απ' αυτούς στηρίζεται.

Στο κεφάλαιο 3 παρουσιάζεται το πρόβλημα της επιλογής χαρακτηριστικών, ο τρόπος με τον οποίο λειτουργεί ένας αλγόριθμος επιλογής χαρακτηριστικών και η παρουσίαση των αλγορίθμων που θα εξεταστούν σε αυτή την εργασία.

Στο κεφάλαιο 4 παρουσιάζεται αναλυτικά το σετ δεδομένων το οποίο θα εξεταστεί και τα χαρακτηριστικά από τα οποία αποτελείται, γίνεται σε αυτό εφαρμογή των αλγορίθμων επιλογής χαρακτηριστικών, βρίσκονται νέες υποομάδες χαρακτηριστικών μέσα από αυτό και οι υποομάδες αυτές δοκιμάζονται και ελέγχονται από μια σειρά αλγορίθμων ταξινόμησης. Η βελτίωση της ευστοχίας της ταξινόμησης των αλγορίθμων αυτών είναι και το κριτήριο με το οποίο κρίνεται η ποιότητα της κάθε υποομάδας.

Στο κεφάλαιο 5 παρουσιάζονται τα συμπεράσματα και οι μελλοντικές κατευθύνσεις αυτής της εργασίας.

2 ΤΑΞΙΝΟΜΗΣΗ

2.1 Εισαγωγη

Στην *Εξόρυξη Δεδομένων*, η προβληματική της *Ταξινόμησης* (Classification) έχει οριστεί από τους Witten & Frank (2000) ως εξής: "Η προσπάθεια πρόβλεψης της κατηγορίας σε ήδη κατηγοριοποιημένα δεδομένα μέσω της δόμησης μοντέλου βασισμένου σε κάποιες μεταβλητές πρόβλεψης".

Συνεπώς, η ταξινόμηση είναι η διαδικασία μάθησης του τρόπου ταξινόμησης ενός νέου αντικειμένου σε μια εκ' των προκαθορισμένων κατηγοριών ή κλάσεων. Η ικανότητα της διεξαγωγής της ταξινόμησης και η γνώση που απορρέει από αυτή δίνει την ικανότητα της λήψης αποφάσεων και η δύναμη αυτών των αποφάσεων επηρεάζεται από την επίδοση της ταξινόμησης.

Όπως γίνεται κατανοητό, η επιτυχία της ταξινόμησης εξαρτάται σε πολύ μεγάλο βαθμό από την ποιότητα των δεδομένων που του παρέχονται για εκπαίδευση. Αν τα δεδομένα εισόδου είναι ανεπαρκή ή μη-συναφή αυτό θα αντικατοπτριστεί στην περιγραφή του θέματος από τον αλγόριθμο και θα έχει αποτέλεσμα την εσφαλμένη ταξινόμηση όταν δοκιμαστεί σε νέα δεδομένα.

2.2 Αναπαράσταση Δεδομένων

Στην Μάθηση υπό Επίβλεψη (Supervised Learning), τα δεδομένα αναπαρίστανται σαν πίνακας ταξινομημένων περιπτώσεων που είναι διαθέσιμα για ταξινόμηση. Κάθε περίπτωση περιγράφεται από ένα δεδομένο αριθμό χαρακτηριστικών, τα οποία μπορεί να είναι είτε σημαντικά είτε ασήμαντα, καθώς και ένα χαρακτηριστικό-ταμπέλα το οποίο καταχωρεί την κάθε περίπτωση σε μια απ' τις προκαθορισμένες κλάσεις.

Τα *χαρακτηριστικά* (attributes) μπορούν κατ' αρχή να χωριστούν σε δύο είδη, στα ονομαστικά και στα αριθμητικά. Στα πλαίσια ωστόσο μιας πιο αναλυτικής κατηγοριοποίησης μπορούν να χωριστούν στις εξής κατηγορίες:

- Οι ονομαστικές (nominal) ποσότητες λαμβάνουν τιμές που αποτελούν διακριτά σύμβολα. Οι τιμές από μόνες τους αποτελούν μόνο ταμπέλες-ονόματα. Ανάμεσα στις πιθανές τιμές δεν υπάρχει καμία σχέση κατάταξης ή απόστασης μεταξύ των τιμών. Ένας κανόνας που χρησιμοποιεί ένα τέτοιο χαρακτηριστικό μπορεί να ελέγξει την ισότητα ή μη μεταξύ δύο διαφορετικών απαντήσεων.

- Οι σε σειρά (ordinal) ποσότητες είναι εκείνες που καθιστούν εφικτή την κατάταξη μεταξύ των πιθανών απαντήσεων ενός ονομαστικού χαρακτηριστικού. Ωστόσο αν και υπάρχει η έννοια της κατάταξης δεν υφίσταται η έννοια της απόστασης μεταξύ των τιμών πράγμα που καθιστά αδύνατη την εφαρμογή μαθηματικών πράξεων μεταξύ των τιμών. π.χ. Για τις τιμές: cool, mild και hot ισχύει: cool<mild<hot

- Οι με απόσταση (interval) ποσότητες λαμβάνουν τιμές όχι μόνο διατεταγμένες αλλά και λαμβάνουν μετρήσιμες τιμές σε κοινό σύστημα μονάδων (π.χ. βαθμοί Κελσίου ή Φαρενάιτ, χρονολογίες). Αυτό καθιστά εφικτές κάποιες πράξεις μεταξύ τιμών, τον υπολογισμό μέσου όρου, καθώς και τη συγκέντρωση των απαντήσεων. Ωστόσο αυτό που δεν λαμβάνουν υπόψη τα συγκεκριμένα χαρακτηριστικά είναι το μηδενικό σημείο αναφοράς.

- Οι με αναλογία (ratio) ποσότητες απ' την άλλη, πέρα από όλα τα χαρακτηριστικά των «με απόσταση», ορίζουν έμφυτα το μηδενικό σημείο (π.χ. κατά τη μέτρηση απόστασης ενός αντικειμένου από ένα άλλο, η απόσταση του αντικειμένου από τον εαυτό του αποτελεί ένα φυσικό μηδενικό σημείο).

2.3 Σετ Εκπαίδευσης και Σετ Ελέγχου

Στα προβλήματα ταξινόμησης της Μηχανικής Μάθησης τόσο στις "με επίβλεψη" όσο και στις "άνευ επίβλεψης" εφαρμογές απαιτούνται δύο σετ παραδειγμάτων: το *σετ Εκπαίδευσης* (training set) και *σετ Ελέγχου* (test set).

Το *σετ εκπαίδευσης*, το οποίο είναι η υποομάδα των δεδομένων από τα οποία το μοντέλο εκπαιδεύεται, χρησιμοποιείται για να παραχθεί το αντικείμενο της μάθησης του εξεταζόμενου θέματος ενώ το *σετ ελέγχου* είναι ένα ανεξάρτητο σετ που χρησιμοποιείται τόσο για την καλύτερη ρύθμιση των παραμέτρων του μοντέλου όσο και για την επικύρωση της ευστοχία της ταξινόμησης.

Η ύπαρξη των δύο αυτών υπο-ομάδων είναι επιβεβλημένη σε προβλήματα ταξινόμησης μια και η επίδοση του κάθε ταξινομητή αξιολογείται με βάση τη ποσοστιαία αναλογία λάθος ταξινομήσεων (error rate). Όπως έχει προαναφερθεί ένας βασικός στόχος του μοντέλου που κατασκευάζεται από τα δεδομένα εκπαίδευσης είναι η μελλοντική πρόβλεψη σε νέα δεδομένα. Για την αξιολόγηση της ικανότητας του αυτής δεν μπορούν να χρησιμοποιηθούν τα δεδομένα εκπαίδευσης από τα οποία έχει προκύψει το ίδιο το σχήμα. Στην περίπτωση αυτή τα αποτελέσματα θα είναι προσαρμοσμένα στα δεδομένα δοκιμής με αποτέλεσμα το φαινόμενο της *υπέρ-προσαρμογής* (overfitting). Για τον λόγο αυτό τα δεδομένα δοκιμής δεν πρέπει να έχουν παίξει ρόλο κατά την εκπαίδευση και από κει πηγάζει η ανάγκη ύπαρξης του *σετ ελέγχου*.

Το δίλημμα που τίθεται είναι το εξής: Για τη δημιουργία ενός καλού ταξινομητή χρειάζονται όσο περισσότερα δεδομένα γίνεται, για τον αποτελεσματικό έλεγχο απ' την άλλη πάλι χρειάζονται όσο το δυνατόν περισσότερα δεδομένα Και εκεί τίθεται το δίλημμα του διαχωρισμού των αρχικών δεδομένων. Πόσα από αυτά πρέπει να χρησιμοποιηθούν για εκπαίδευση και πόσα για έλεγχο.

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.

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 αλγόριθμος των επαναλήψεων Κ 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. Αντίθετα αν το Κ είναι μικρό είναι πιθανό ο υπολογισμός του σφάλματος να είναι υπερβολικά αισιόδοξος, λόγω της διαφοράς στο μέγεθος των δειγμάτων εκμάθησης και ελέγχου που διαμορφώνονται σε κάθε επανάληψη της διαδικασίας του 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 και η μέθοδος είναι ευρύτερα γνωστή ως 10-fold-cross-validation.

2.4 Αναπαράσταση Γνώσης και Αλγόριθμοι Μάθησης

Υπάρχουν πολλοί διαφορετικοί τρόποι αναπαράστασης των προτύπων που μπορούν να ανακαλυφθούν από τη μηχανική μάθηση, κάθε ένας από αυτούς υπαγορεύει την τεχνική που θα χρησιμοποιηθεί ώστε να προκύψει αυτή η δομή αποτελέσματος.

Τα βασικά είδη αναπαράστασης γνώσης που οι περισσότερες μέθοδοι μηχανικής μάθησης χρησιμοποιούν είναι τα *δέντρα απόφασης* (Decision Trees) και οι *κανόνες ταξινόμησης* (Classification Rules). Άλλες μέθοδοι όπως οι *ταξινομητές Μπαϊες* (Bayes

classifiers) ερευνούν τις εκ των προτέρων πιθανότητες ταξινόμησης ενός στιγμιότυπου, ενώ τέλος, οι *βασισμένοι στα στιγμιότυπα ταξινομητές* (Instance-Based Classifiers) επικεντρώνονται όχι στη δημιουργία σχημάτων που θα κατατάσσουν τα στιγμιότυπα με βάση τις τιμές των χαρακτηριστικών τους, αλλά στα ίδια τα στιγμιότυπα. Υπάρχουν αρκετοί ακόμα ταξινομητές οι οποίοι είναι είτε παραλλαγές είτε υβριδικοί συνδυασμοί των παραπάνω.

2.4.1 Δέντρα Απόφασης

Τα δέντρα απόφασης αποτελούν μια από τις βασικές μορφές ταξινόμησης και πρόβλεψης και αυτό διότι τα δέντρα απόφασης αναπαριστούν κανόνες. Συνήθως, χρησιμοποιούν αριθμητικά και ονομαστικά δεδομένα. Ο λόγος που αυτός ο τρόπος αναπαράστασης είναι τόσο διαδεδομένος, είναι ότι ο μηχανισμός της διαδικασίας απόφασης είναι διαφανής και επιτρέπει στον εμπειρογνώμονα την πιο εμπεριστατωμένη αποκομιδή γνώσης.

Τα δομικά στοιχεία ενός δέντρου απόφασης είναι οι *κόμβοι* (nodes), τα *κλαδιά* (branches) και τα *φύλλα* (leaves). Οι κόμβοι ενός δέντρου απόφασης αφορούν τον έλεγχο ενός συγκεκριμένου χαρακτηριστικού. Αυτό μπορεί να σημαίνει σύγκριση της τιμής που λαμβάνει το συγκεκριμένο χαρακτηριστικό με κάποιο σταθερό αριθμό ή να σημαίνει κατάταξη σε περίπτωση ονομαστικών χαρακτηριστικών. Σε κάποιες περιπτώσεις δέντρων ένας κόμβος μπορεί ακόμα να σημαίνει και σύγκριση των τιμών δύο διαφορετικών χαρακτηριστικών μεταξύ τους. Οι κόμβοι του δέντρου εκτελούν μια διαδικασία κατάταξης σε όλες τις περιπτώσεις που φτάνουν σε καθέναν από αυτούς με απώτερο σκοπό την τελική ταξινόμηση της καθεμίας από αυτές σε ένα από τα φύλλα, το οποίο αναπαριστά μια από τις προκαθορισμένες «κλάσεις».

Για να ταξινομηθεί μια άγνωστη (αταξινόμητη) περίπτωση ακολουθεί μια πορεία από την κορυφή του δέντρου προς τις ρίζες του μέσα από διαδοχικούς κόμβους που ελέγχουν ο καθένας τις τιμές που λαμβάνει κάποιο από τα χαρακτηριστικά της και ανάλογα την κατατάσσουν σε ενδιάμεσα κλαδιά τα οποία οδηγούν σε νέους κόμβους-παιδιά, ως που να φτάσει σε ένα φύλλο του δέντρου το οποίο θα σημάνει και την ταξινόμησή της.

Αν το χαρακτηριστικό που ελέγχεται σε ένα κόμβο είναι ονομαστικό, ο αριθμός των κλαδιών που θα προκύψουν από αυτόν είναι συνήθως όσες και οι πιθανές απαντήσεις του ελεγχόμενου χαρακτηριστικού. Αν πάλι το χαρακτηριστικό είναι σε αριθμητική μορφή, αυτό που συνήθως ελέγχεται στον κόμβο είναι αν η τιμή που λαμβάνει είναι μεγαλύτερη ή μικρότερη από ένα σταθερό αριθμό –τον οποίο ο ίδιος ο αλγόριθμος προσδιορίζει- οπότε συνήθως προκύπτει διαχωρισμός σε δύο κλαδιά. Στην ίδια περίπτωση, διαχωρισμός σε τρία κλαδιά μπορεί να προκύψει αν εκτός από τα κλαδιά μεγαλύτερο-μικρότερο προστεθεί και το «είναι ίσο», ή αν καταταχθούν από τον αλγόριθμο σε ένα τρίτο κλαδί περιπτώσεις που έχουν άγνωστη τιμή στο ελεγχόμενο χαρακτηριστικό.

Αλγόριθμος C4.5

Μια από τις πλέον γνωστές τεχνικές στο χώρο της μηχανικής μάθησης για την ανάπτυξη ενός συνόλου τέτοιων κανόνων απόφασης είναι ο αλγόριθμος C4.5, ο οποίος αναπτύχθηκε από τον Quinlan (1993), ως εξέλιξη του γνωστού αλγορίθμου ID3 (Quinlan, 1983, 1986). Ο αλγόριθμος J.48 είναι η έκδοση του C4.5 για την πλατφόρμα αλγορίθμων μηχανικής μάθησης WEKA. Βασικά πλεονεκτήματα του αλγορίθμου C4.5 έναντι του προκατόχου του ID3 αποτελούν:

- Η δυνατότητα διαχείρισης ποσοτικών κριτηρίων.
- Η δυνατότητα διαχείρισης δεδομένων με ελλιπή στοιχεία.
- Η αποφυγή της μεγάλης προσαρμογής στα δεδομένα του δείγματος εκμάθησης (overfitting).

Ο ID3 ήταν ο πρώτος αλγόριθμος που χρησιμοποίησε για το κριτήριο καταλληλότητας τεμαχισμού το κέρδος *Gain* από τη θεωρία πληροφορίας.

Αν $Y=\{y_1,\ldots,y_n\}$ το σύνολο των κλάσεων της ποιοτικής εξαρτημένης μεταβλητής Y, $p(yi)$ η πιθανότητα εμφάνισης της y_i κλάσης, τότε η εντροπία του συνόλου υπολογίζεται από τον τύπο:

$$E(Y) = -\sum_{i=1}^{n} p(y_i) \log_2 p(y_i)$$

Η ελάχιστη τιμή εντροπίας $E_{min}(Y)=0$ φανερώνει τη μέγιστη βεβαιότητα (σιγουριά) σχετικά με την πιθανότητα εμφάνισης μίας συγκεκριμένης τιμής y_i από το σύνολο Y. Η μέγιστη τιμή εντροπίας από την άλλη πλευρά επιτυγχάνεται όταν όλες οι πιθανότητες $p(y_i)$ είναι ίσες με $1/n$, οπότε και η εντροπία διαμορφώνεται σε $Emax(Y) = logn$, γεγονός που αυξάνει στο μέγιστο την αβεβαιότητα σχετικά με ποιο μέλος του συνόλου Y θα προκύψει. Πρακτικά για την εφαρμογή ενός τέτοιου κριτηρίου, μικρή τιμή εντροπίας αυξάνουν τις πιθανότητες η τρέχουσα θέση του δέντρου απόφασης δηλαδή ο υπό εξέταση κόμβος t να είναι φύλλο του δέντρου, ενώ αντίθετα μεγάλες τιμές δείχνουν ότι απαιτείται κατασκευή υπό-δέντρου κάτω από τον t κόμβο. Στην περίπτωση που $n=2$, δηλαδή στην περίπτωση των δύο μόνο κλάσεων, έστω οι κλάσεις P και N, και p ο αριθμός των παραδειγμάτων από το σύνολο εκπαίδευσης που ανήκει στην P κλάση, n ο αντίστοιχος αριθμός παραδειγμάτων που ανήκει στην N κλάση, $p/(p+n)$ η πιθανότητα ένα παράδειγμα να ανήκει στην P κλάση, $n/(p+n)$ η πιθανότητα να ανήκει στην N κλάση. Η αναμενόμενη πληροφορία για τον καθορισμό της κλάσης είναι:

$$I(p,n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

Η τιμή του I για τον κόμβο t είναι μία σταθερή τιμή που υπολογίζεται από τον αριθμό θετικών και αρνητικών παραδειγμάτων σύμφωνα με τον προηγούμενο τύπο. Αν ο αλγόριθμος τεμάχιζε το δέντρο κάτω από τον κόμβο t σε m τεμάχια-κλαδιά με βάση την ιδιότητα A, τότε η αναμενόμενη πληροφορία για τον καθορισμό της κλάσης από τον κόμβο t και κάτω με δεδομένο τον τεμαχισμό που αναφέρθηκε θα ήταν:

$$E(Y \mid A) = \sum_{i=1}^{m} \frac{p_i + n_i}{p+n} I(p_i + n_i)$$

Το κέρδος πληροφορίας πριν και μετά τον τεμαχισμό υπολογίζεται σαν η διαφορά μεταξύ της αναμενόμενης πληροφορίας στον κόμβο t πριν τον τεμαχισμό $I(p,n)$, και μετά τον τεμαχισμό $E(Y|A)$:

$$Gain(A) = I(p,n) - E(Y \mid A)$$

Από όλους τους δυνατούς τεμαχισμούς με όλες τις δυνατές ιδιότητες, θα επιλεχθεί αυτός που δίνει το μεγαλύτερο κέρδος, δηλαδή τη μικρότερη εντροπία λόγω τεμαχισμού *E(Y|A)*.

Όταν κάποια χαρακτηριστικά έχουν μεγάλο εύρος πιθανών τιμών οδηγούν σε έναν πιθανό καταμερισμό με πολλά παιδιά-κόμβους. Σε μια ακραία περίπτωση ενός χαρακτηριστικού που λαμβάνει διαφορετική τιμή για την καθεμία εκ' των περιπτώσεων του εξεταζόμενου σετ δεδομένων το μέτρο του κέρδους πληροφορίας για το χαρακτηριστικό αυτό θα μεγιστοποιούταν χωρίς ωστόσο αυτό να αντικατοπτρίζει την πραγματική διαχωριστική του αξία. Για τον λόγο αυτό κρίνεται αναγκαία η δημιουργία ενός νέου μέτρου το οποίο θα λαμβάνει υπόψη τον αριθμό και το μέγεθος των παιδιών-κόμβων που προκύπτουν από το εκάστοτε χαρακτηριστικό. Το μέτρο αυτό ονομάζεται "*αναλογία κέρδους*" (Gain Ratio) και υπολογίζεται ως από το παρακάτω πηλίκο:

H(C) / H(C|A$_i$).

Το κριτήριο αυτό επιλέγει, ανάμεσα στα χαρακτηριστικά με την μικρότερη εντροπία, εκείνο που μεγιστοποιεί την αναλογία κέρδους.

Αλγόριθμος NBTree

Ο NBTree (Kohavi R.,) είναι ένα υβριδικός αλγόριθμος ο οποίος αποτελεί συνδυασμό κλασσικού δέντρου απόφασης με Μπαεϊζιανούς ταξινομητές επιχειρώντας να συνδυάσει προτερήματα των δύο αυτών δημοφιλών μεθόδων ταξινόμησης και πρόβλεψης.

Ο Naïve-Bayes (Langley, Iba, & Thompson 1992) ταξινομητής υπολογίζει μέσω του νόμου του Bayes την πιθανότητα κατάταξης στην κάθε κλάση για την κάθε περίπτωση που εξετάζει. Είναι γενικά ένας γρήγορος στην εφαρμογή του και εύκολος στην κατανόηση και την ερμηνεία των αποτελεσμάτων του αλγόριθμος. Είναι εξαιρετικά ικανός στον χειρισμό "μη-συναφών χαρακτηριστικών" (irrelevant attributes) μια και κατά τη διαδικασία ταξινόμησης λαμβάνει υπόψη τα δεδομένα του συνόλου των ιδιοτήτων. Απ' την άλλη, ο αλγόριθμος αυτός απαιτεί και λαμβάνει ως δεδομένη την

ανεξαρτησία των ιδιοτήτων και όταν αυτό δεν ισχύει η απόδοση του μοντέλου πρόβλεψης μειώνεται.

Απ' την άλλη οι ταξινομητές δέντρων απόφασης είναι επίσης γρήγοροι και ευκολονόητοι αλλά η επαγωγική μέθοδος, βασισμένη στο περιοδικό διαχωρισμό των δεδομένων που χρησιμοποιεί, δημιουργεί πρόβλημα. Μετά από πολλές διαδοχικές διασπάσεις των δεδομένων σε κάθε κόμβο του δέντρου τα δεδομένα που απομένουν στο τέλος είναι συνήθως πολύ λίγα για να βασιστούν σε αυτά αποφάσεις.

Ο NBTree είναι ένας αλγόριθμος παρόμοιος με τα κλασσικά σχήματα περιοδικού διαχωρισμού με τη διαφορά ότι τα φύλλα που δημιουργούνται από τους κόμβους, αντί να προβλέπουν απλά την κλάση, καταλήγουν σε κατηγοριοποιητές Naïve Bayes. Για τις ποσοτικές ιδιότητες υπολογίζεται ένα κατώφλι διαχωρισμού με κριτήριο την ελαχιστοποίηση της εντροπίας όπως γίνεται στα δέντρα απόφασης. Η χρησιμότητα του εκάστοτε κόμβου υπολογίζεται με μετατροπή των δεδομένων σε διακριτά και η εκτίμηση γίνεται μέσω μιας διαδικασίας εκτίμησης ευστοχίας 5-fold cross-validation των χρησιμοποιούμενων εκτιμητών Naïve Bayes που βρίσκονται στα φύλα. Η χρησιμότητα του κάθε μοιράσματος καθορίζεται από το ζυγισμένο άθροισμα της χρησιμότητας των κόμβων του, όπου το βάρος δίνεται στον κάθε κόμβο αναλογικά από τον αριθμό περιπτώσεων που αυτός περιλαμβάνει.

Στη συνέχεια επιχειρείται μια διαδικασία ανάλογη του κλαδέματος (pruning) στα απλά δέντρα απόφασης. Ο αλγόριθμος καλείται λοιπόν να αποφασίσει στον εκάστοτε κόμβο αν μια επιπλέον διάσπαση θα αύξανε την ευστοχία του σχήματος ή είναι προτιμότερη, χάριν της γενίκευσης, η δημιουργία ενός μπαϊεζιανού κατανεμητή στο συγκεκριμένο σημείο. Για να αποφευχθούν διασπάσεις με μικρή αξία, ως σημαντική ορίζεται μια διάσπαση αν η σχετική μείωση στις λάθος ταξινομήσεις είναι μεγαλύτερη του 5% και υπάρχουν τουλάχιστον 30 στιγμιότυπα στον κόμβο.

Σχηματικά ο αλγόριθμος NBTree λειτουργεί ως εξής: Ορίζουμε για είσοδο ένα σετ T από i ιδιότητες

1. Για κάθε ιδιότητα X_i , υπολόγισε την χρησιμότητα $u(X_i)$, ενός διαχωρισμού σε αυτή την ιδιότητα. Για συνεχείς ιδιότητες υπολόγισε το κατώφλι.

2. Όρισε $j = \arg\max_i (u_i)$, *i.e.,* την ιδιότητα με την μεγαλύτερη χρησιμότητα.

3. Αν η u_j είναι σημαντικά καλύτερη από την χρησιμότητα του συγκεκριμένου κόμβου (τα κριτήρια για να θεωρηθεί σημαντική μια διάσπαση περιγράφονται παραπάνω), συνέχισε στο βήμα 4. Αλλιώς, δημιούργησε έναν ταξινομητή Naive Bayes στον συγκεκριμένο κόμβο και ανέφερε.

4. Διαχώρισε το T σύμφωνα με τον έλεγχο στο βήμα 1. Αν η X_i είναι ποιοτική διαχώρισε και κατάταξε τα στιγμιότυπα σε όσες είναι οι πιθανές τιμές της X_i , αν είναι ποσοτική χώρισε στα δύο τα στιγμιότυπα στο κατώφλι που έχει υπολογιστεί στο βήμα 1.

5. Για κάθε παιδί-κόμβο που προκύπτει εφάρμοσε επαναληπτικά τον αλγόριθμο για τις ιδιότητες και τα στιγμιότυπα που αναλογούν σε αυτό.

2.4.2 Κανόνες Απόφασης

Οι κανόνες απόφασης είναι μια μέθοδος αναπαράστασης γνώσης εναλλακτική του δέντρου απόφασης. Η εύρεση των κανόνων γίνεται μέσα από τα δείγματα του σετ εκπαίδευσης και αξιολόγηση αυτών γίνεται στη συνέχεια από το σετ ελέγχου.

Οι κανόνες ταξινόμησης παρουσιάζουν μια συγκεκριμένη δομή η οποία αποτελείται από συνθήκες και αποτέλεσμα. Η συνθήκες του κανόνα είναι μια σειρά από ελέγχους σε κάποια από τα χαρακτηριστικά του σετ ανάλογη με τους ελέγχους που πραγματοποιούνται στους κόμβους ενός δέντρου. Το αποτέλεσμα δίνει την κλάση ή τις κλάσεις στις οποίες κατατάσσεται το δείγμα το οποίο πληροί τις προαναφερθείσες συνθήκες.

Κατά καιρούς έχουν ερευνηθεί διάφορες μέθοδοι κατασκευής κανόνων. Μια μέθοδος είναι η κατασκευή ενός δέντρου απόφασης, η μετατροπή του σε σύνολο κανόνων και η τελική απλοποίηση και διαλογή των καλύτερων απ' αυτούς (Quinlan, 1987a). Μια άλλη

φιλοσοφία δημιουργίας κανόνων είναι η "*separate-and-conquer*" (χώρισε και κατέκτησε) (Paggalo & Haussler, 1990) στρατηγική η οποία εφαρμόστηκε αρχικά στους αλγόριθμους της οικογένειας AQ (Michalski, 1969) και αποτέλεσε στη συνέχεια βάση για πολλά άλλα συστήματα.

Ένας λόγος που οι κανόνες είναι δημοφιλείς είναι το γεγονός ότι καθένας από αυτούς αποτελεί ένα ξεχωριστό κομμάτι γνώσης. Μπορεί κανείς να προσθέσει ή να αφαιρέσει κάποιον απ' αυτούς χωρίς να διαταράξει το συνολικό οικοδόμημα, πράγμα που δεν μπορεί να γίνει με τα δέντρα απόφασης στα οποία κάτι τέτοιο θα αναδιαμόρφωνε την δομή όλου του δέντρου. Ωστόσο από το πλεονέκτημα της ανεξαρτησίας των κανόνων πηγάζει και το μεγαλύτερο τους μειονέκτημα. Είναι πολύ συχνό το φαινόμενο αντικρουόμενων κανόνων ή δειγμάτων τα οποία πληρούν τις συνθήκες σε παραπάνω από έναν κανόνες με διαφορετικό αποτέλεσμα. Για τον λόγο αυτό στην τελική επιλογή των κανόνων τίθεται συνήθως ένα πολύ υψηλό κατώφλι στο ποσοστό ευστοχίας τους.

Αλγόριθμος PART

Η "*separate-and-conquer*" στρατηγική, στην ουσία, λειτουργεί βρίσκοντας τον πιο ισχυρό κανόνα που προκύπτει από τα δεδομένα ξεχωρίζοντας παράλληλα τις περιπτώσεις εκείνες που εμπίπτουν σ' αυτόν. Η διαδικασία αυτή επαναλαμβάνεται για τα εναπομείναντα δεδομένα και δημιουργείται μια λίστα από κανόνες αποκαλούμενη και "decision list" (Rivest, 1987). Οι δύο σημαντικότερες πρακτικές εφαρμογές των παραπάνω μεθόδων είναι οι αλγόριθμοι C4.5 (Quinlan, 1993) και RIPPER (Cohen, 1995). Και οι δύο αυτοί αλγόριθμοι στοχεύουν σε μια διαδικασία *συνολικής βελτιστοποίησης* (global optimization) μέσω της μετέπειτα επεξεργασίας των κανόνων που αρχικά παράγουν.

Στην περίπτωση του C4.5 ο λόγος είναι ο υπερβολικά μεγάλος αριθμός κανόνων που προκύπτει από την μετατροπή του αρχικού δέντρου. Ο αλγόριθμος λοιπόν καλείται αφ' ενός να επιλέξει τους κανόνες με τη μεγαλύτερη ισχύ και παράλληλα να τους βελτιστοποιήσει μέσα από μια διαδικασία περικοπής τυχών πλεονασματικών συνθηκών με στόχο την ελαχιστοποίηση του εκτιμώμενου ποσοστού λάθους του κάθε κανόνα.

Στην περίπτωση πάλι του RIPPER, το κίνητρο είναι η αύξηση της ευστοχίας των αρχικών κανόνων μέσα από την αντικατάσταση ή την ανακατασκευή τους μέσα από μια διαδικασία που ονομάζεται "κλάδεμα" (pruning). Όπως και ο Cohen (1995) το έθεσε: "…οι κανόνες που προκύπτουν τόσο από το C4.5 όσο και από το RIPPER ξεκινούν από ένα αρχικό μοντέλο και το βελτιώνουν συνεχώς χρησιμοποιώντας ευρετική τεχνική (ευρετική: μέθοδος ενεργειών με βάση τις κτηθείσες εμπειρίες)".

Ωστόσο η καθεμία από αυτές τις μεθόδους έχει τα δικά της προβλήματα. Για την C4.5 μεγάλο πρόβλημα αποτελεί η πολυπλοκότητα που παρουσιάζει η οποία την καθιστά εξαιρετικά χρονοβόρα ειδικά σε περιπτώσεις δεδομένων με "θόρυβο". Επιπλέον, παρά την εκτενή βελτιστοποίηση οι κανόνες ακόμα υπόκεινται στο δεσμό τους με τα λάθη που προέρχονται από το αρχικό δέντρο απόφασης. Όσον αφορά τον RIPPER, το πιο μεγάλο πρόβλημα δημιουργείται από τη υπερβολική χρήση του κλαδέματος (over-pruning) η οποία έχει ως αποτέλεσμα την λανθασμένη περικοπή κάποιων κανόνων και είναι η λεγόμενη "*βιαστική γενίκευση*" (hasty generalization).

Ο αλγόριθμος PART αποτελεί ένα συνδυασμό των δύο παραπάνω προσεγγίσεων σε μια προσπάθεια περιορισμού των επιμέρους προβλημάτων τους. Υιοθετεί την στρατηγική "*χώρισε και κατέκτησε*" στο ότι δημιουργεί ένα κανόνα, ξεχωρίζει τις περιπτώσεις που καλύπτονται απ' αυτόν και συνεχίζει να δημιουργεί κανόνες με όσες περιπτώσεις απομένουν.

Ωστόσο, ο PART διαφέρει από τον RIPPER στον τρόπο που δημιουργείται ο κάθε κανόνας. Στην ουσία, για να δημιουργήσει τον καθένα από τους κανόνες, ο αλγόριθμος, κατασκευάζει από τις περιπτώσεις που διαθέτει ένα δέντρο απόφασης, επιλέγει το φύλλο που καλύπτει τις περισσότερες εξ' αυτών και το μετατρέπει σε κανόνα. Στη συνέχεια το δέντρο αυτό διαγράφεται. Με τον τρόπο αυτό αποφεύγεται το βασικό ελάττωμα των στρατηγικών "χώρισε και κατέκτησε": η "βιαστική γενίκευση", μια και με τη δημιουργία του δέντρου εξερευνάται όλο το φάσμα των πιθανών κανόνων και είναι μικρότερος ο κίνδυνος απόκρυψης γνώσης.

Ωστόσο η διαδικασία αυτή θα ήταν εξαιρετικά χρονοβόρα αν για καθέναν από τους κανόνες έπρεπε να κατασκευάζεται ένα ολόκληρο δέντρο απόφασης. Το πρόβλημα αυτό

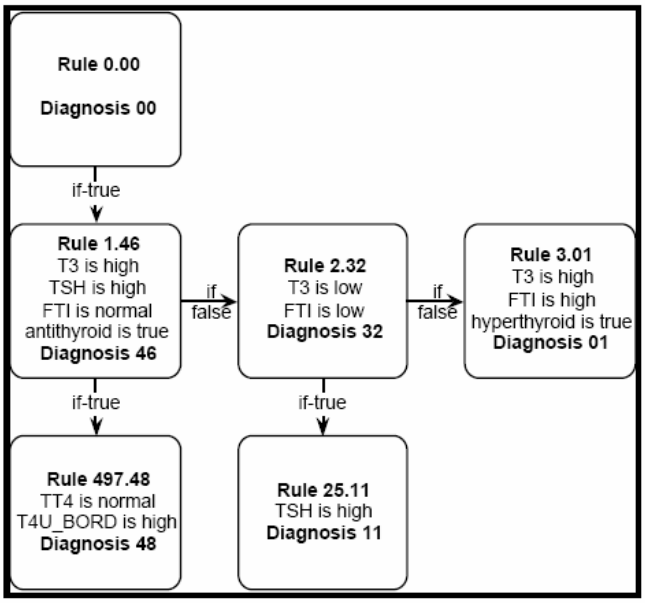
λύνεται με τη χρήση των μερικών δέντρων απόφασης "partial decision trees". Αυτό δεν είναι τίποτε άλλο από ένα δέντρο απόφασης που τα κλαδιά του οδηγούν σε μη-ορισμένα υπό-δέντρα. Για να δημιουργηθεί αυτό το δέντρο πρέπει κατά τη διάρκεια της κατασκευής του δέντρου μέσα από μια παράλληλη διαδικασία κλαδέματος να επιλέγεται η ανάπτυξη μόνο του εκάστοτε καλύτερου υπό-δέντρου εωσότου φτάσει σε κάποιο που να μην αναπτύσσεται παραπέρα. Στο σημείο αυτό η κατασκευή το δέντρου διακόπτεται και γίνεται η εξαγωγή του κανόνα.

Αλγόριθμος Ridor

Οι Compton & Jansen εισήγαγαν την τεχνική εξαγωγής κανόνων "ripple down" ως μεθοδολογία με στόχο την απόκτηση και την υποστήριξη μεγάλων, βασισμένων σε κανόνες, συστημάτων (Compton and Jansen, 1990a, b). Η θεμελιώδης σκέψη στηρίζεται στο ότι οι άνθρωποι αντιμετωπίζουν την απόκτηση και την υποστήριξη πολύπλοκων δομών γνώσης κάνοντας σταδιακές αλλαγές σ' αυτές στα πλαίσια ενός καλά ορισμένου περιβάλλοντος τέτοιου ώστε το αποτέλεσμα των αλλαγών να περιορίζεται τοπικά με έναν συγκεκριμένο τρόπο. Τα καθιερωμένα σχήματα παραγωγής κανόνων δεν διαθέτουν αυτή την ιδιότητα. Η συναρμολογισημότητα των ίδιων των κανόνων δεν αντανακλάται στη συναρμολογισημότητα των συνεπειών των αλλαγών σε αυτούς τους κανόνες. Μικρές αλλαγές μπορούν να οδηγήσουν, μέσω πολύπλοκων αλληλεπιδράσεων, σε μέγιστα αποτελέσματα καθιστώντας την ανάπτυξη και συντήρηση των βασισμένων σε κανόνες συστημάτων πολύ πιο πολύπλοκη απ' ότι αρχικά φαίνεται.

Η τεχνική κανόνων "*ripple down*" δημιουργεί μια σχέση αμφίδρομης εξάρτησης μεταξύ των κανόνων τέτοιας ώστε η ενεργοποίηση ενός κανόνα να εξετάζεται μόνο υπό την προϋπόθεση ότι ένας άλλος κανόνας είναι ενεργός. Αν η προϋπόθεση ενός γονέα-κανόνα πληρείται για ένα συγκεκριμένο χαρακτηριστικό τότε το συμπέρασμα του, εφ' όσον δεν υπάρχουν πιο κάτω άλλοι εξαρτώμενοι απ' αυτό κανόνες, θα βεβαιώνεται γι' αυτό το χαρακτηριστικό. Αν, ωστόσο, ακολουθεί εξαρτώμενος κανόνας της μορφής "*if true*" τότε θα εξεταστεί ο κανόνας αυτός και στη συνέχεια οι υπόλοιποι εξαρτώμενοι από αυτόν κανόνες, εφ' όσων υπάρχουν. Το αρχικό συμπέρασμα θα ισχύει μόνο εφ' όσων οι προϋποθέσεις όλων των κανόνων που ακολουθούν δεν πληρούνται. Αντιθέτως, αν οι

προϋποθέσεις ενός γονέα-κανόνα δεν πληρούνται για ένα υποκείμενο και ακολουθεί εξαρτώμενος κανόνας της μορφής "*if false*", τότε ο κανόνας-γονέας δεν θα γίνει αποδεκτός και θα εξεταστεί στη συνέχεια η ισχύ του παιδιού-κανόνα.



Σχήμα 1. 1 Κάποιοι κανόνες "ripple down" μορφής (4)

Με τον τρόπο αυτό οι κανόνες της μορφής "ripple down" σχηματίζουν ένα δυαδικό δέντρο απόφασης που διαφέρει από τα καθιερωμένα δέντρα απόφασης (Breiman, Friedman, Olshen and Stone, 1984) στο ότι σύνθετες προτάσεις χρησιμοποιούνται για τον καθορισμό του "μοιράσματος σε κλαδιά" (branching) και οι προτάσεις αυτές δεν χρειάζεται να καλύπτουν εξαντλητικά όλες τις περιπτώσεις. Με τον τρόπο αυτό γίνεται εφικτή η λήψη μιας απόφασης σε έναν εσωτερικό κόμβο, πράγμα το οποίο έρχεται σε αντίθεση με τα καθιερωμένα δέντρα απόφασης στα οποία όλες οι αποφάσεις λαμβάνονται στους κόμβους ρίζες. Ωστόσο, το στοιχείο των δέντρων απόφασης που εξακολουθεί να ισχύει είναι ότι μόνο ένας κόμβος απόφασης ενεργοποιείται για την κάθε μια εξεταζόμενη περίπτωση.

Η αυτόματη επαγωγή των κανόνων "ripple down" είναι απλή με τη χρήση στατιστικής μεθοδολογίας για την εμπειρική επαγωγή κανόνων από δεδομένα όπως αυτό περιγράφεται στον αλγόριθμο Induct (Gaines, 1989). Το "Induct" είναι ένα σύστημα με επιδόσεις παρεμφερείς μ' αυτές του C4.5 (Quinlan, 1993) με τη διαφορά ότι εξάγει τους κανόνες απευθείας χρησιμοποιώντας μια προέκταση του αλγόριθμου Prism

(Cendrowska, 1987) για τη χρήση με θορυβώδη δεδομένα. Το σύστημα αυτό εξάγει από μόνο του κανόνες με εξαιρέσεις της μορφής "if-true", "if-false" και η επέκταση του στην εξαγωγή "ripple down" κανόνων θα μπορούσε να θεωρηθεί φυσική εξέλιξη του αλγόριθμου.

Υπάρχουν τρία βήματα στη δημιουργία της πρότασης συλλογισμού της μορφής "ripple down" κανόνων. Κατ' αρχή η πιο συχνά απαντώμενη κλάση, στο κομμάτι της βάσης δεδομένων που εξετάζεται, επιλέγεται ως συμπέρασμα στόχος (target conclusion) και στοιχειοθετείται μια αρχική προϋπόθεση χωρίς επιπλέον όρους. Δεύτερον, ελέγχεται επαναληπτικά κάθε δυνατός συνδυασμός ιδιοτήτων επιλέγεται ο καλύτερος μέσα από στατιστικό έλεγχο. Τρίτον, ο όρος που έχει επιλεγεί προστίθεται στον ήδη υπάρχοντα κανόνα και ελέγχεται, επίσης στατιστικά, αν η προσθήκη αυτή βελτιώνει ή όχι την ισχύ του κανόνα. Στην περίπτωση που κρίνεται σκόπιμη η προσθήκη η διαδικασία επιστρέφει στο δεύτερο στάδιο αναζητώντας την πιθανόν περεταίρω βελτίωση με την προσθήκη κάποιου επιπλέον όρου, διαφορετικά τερματίζει την διαδικασία και δημιουργεί τον κανόνα.

Στην περίπτωση που υπάρχουν χαμένες τιμές (missing values), αυτές λαμβάνονται υπόψη υποθέτοντας ότι αυτές μπορεί να λαμβάνουν κάθε τιμή. Όταν ελέγχεται η επιλογή μιας ιδιότητας η χαμένη τιμή, που τυχόν υπάρχει, λαμβάνεται υπόψη ως τιμή η οποία πληρεί τα κριτήρια επιλογής. Εξετάζοντας στατιστικά το γεγονός αυτό, μια επιλογή βασισμένη σε χαμένες τιμές συμβάλει στην αύξηση των λάθος θετικών (false-positives) γεγονός το οποίο έχει συνέπειες στην απόκτηση γνώσης.

Για τους "ripple down" κανόνες το δίλημμα είναι αν θα επιλεγεί ο πιο γενικός κανόνας που ανταποκρίνεται σε μεγαλύτερο εύρος περιπτώσεων ή αν θα εισαχθεί ένας επιπλέον όρος ο οποίος θα τον εξειδικεύσει. Αυτό που καθιστά ενδιαφέρουσα την απόφαση είναι το γεγονός ότι, με δεδομένη την ικανότητα να διαχειρίζεται εξαιρέσεις, δεν επηρεάζει απαραίτητα την ευστοχία της τελικής βάσης γνώσης. Περισσότερο επηρεάζει τη δομή της θέτοντας την επιλογή κανόνων σε πιο γενικά πλαίσια με πιο πολλές εξαιρέσεις ή πιο συγκεκριμένων κανόνων με λιγότερες. Αυτό θα επηρεάσει ποσοτικά κριτήρια της βάσης γνώσης όπως τους αριθμούς των κανόνων που θα εξαχθούν και των εξαιρέσεων που

αυτοί θα περιλαμβάνουν όπως και θα επηρεάσει επίσης τη δομή της με ένα τρόπο που να ανταποκρίνεται στο γνωστικό τομέα με τον τρόπο που και ένας άνθρωπος ειδικός θα παρουσίαζε τη γνώση.

2.4.3 Μάθηση Βασισμένη σε Στιγμιότυπα

Αλγόριθμος των k Κοντινότερων Γειτόνων (k-Nearest Neighbor)

Ο αλγόριθμος ταξινόμησης με βάση τους *k κοντινότερους γείτονες (k-Nearest Neighbor Algorithm* - *k*-NN) (Belur V. Dasarathy, 1991) είναι η πιο βασική "βασισμένη σε στιγμιότυπα" μέθοδος μάθησης. Η κεντρική ιδέα είναι πως η τιμή της συνάρτησης-στόχου για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των *k* πιο "κοντινών" του στιγμιότυπων εκπαίδευσης, τα οποία αποτελούν τους "γείτονες" του. Τρία ζητήματα πρέπει να αποφασιστούν προκειμένου να καθοριστεί πλήρως ο αλγόριθμος:

1. Ο ορισμός της απόστασης μεταξύ δύο στιγμιότυπων, δηλαδή μιας μετρικής πάνω στο χώρο των στιγμιότυπων (instance space), που θα εκφράζει την εγγύτητα, ή αλλιώς την "ομοιότητα" μεταξύ των στιγμιότυπων.

2. Ο τρόπος συνδυασμού των τιμών των *k* κοντινότερων γειτόνων.

3. Η τιμή του *k*.

Για το πρώτο ζήτημα, υπάρχουν πολλές εναλλακτικές επιλογές. Η απόφαση εξαρτάται από τα ειδικά χαρακτηριστικά του χώρου στιγμιότυπων του προβλήματος. Ιδιαίτερη σημασία έχει το αν στην αναπαράσταση των στιγμιότυπων περιλαμβάνονται αριθμητικά ή συμβολικά χαρακτηριστικά. Στον "παραδοσιακό" *k-Nearest Neighbor* αλγόριθμο, στον οποίο τα στιγμιότυπα θεωρούνται πως ανήκουν στον n-διάστατο χώρο K^n, μια μετρική που υιοθετείται συχνά είναι η γνωστή Ευκλείδεια απόσταση. Πιο συγκεκριμένα, αν τα στιγμιότυπα αναπαρίστανται ως διανύσματα από χαρακτηριστικά που παίρνουν τιμές πραγματικούς αριθμούς, δηλαδή το στιγμιότυπο *x* αναπαρίσταται από το διάνυσμα:

$$\langle a_1(x), a_2(x), ..., a_n(x) \rangle$$

όπου $a_r(x)$ δηλώνει την τιμή του r-οστού feature του x, τότε η απόσταση $d(xi,xj)$ μεταξύ δύο στιγμιότυπων x_i. και x_j ορίζεται ως:

$$d\left(x_i, x_j\right) \equiv \sqrt{\sum_{r=1}^{n}\left(a_r\left(x_i\right) - a_r\left(x_j\right)\right)^2}$$

Φυσικά, κάθε συνάρτηση που πληροί τα κριτήρια μετρικής είναι δυνατόν να επιλεχθεί αντί της Ευκλείδειας.

Στην περίπτωση που τα χαρακτηριστικά είναι ονομαστικά, η Ευκλείδεια απόσταση δεν μπορεί να χρησιμοποιηθεί, αφού δεν έχει νόημα η αφαίρεση συμβολικών ποσοτήτων. Το πιο βασικό μέτρο για αυτήν την περίπτωση είναι το *μέτρο επικάλυψης (overlap metric),* το οποίο αναφέρεται και ως *απόσταση Hamming* ή *απόσταση Manhattan,* και ορίζεται ως εξής:

$$d\left(x_i, x_j\right) \equiv \sum_{r=1}^{n}\delta\left(a_r\left(x_i\right), a_r\left(x_j\right)\right) \qquad \text{Όπου: } \delta\left(x, y\right) \equiv \begin{cases} 0, \ if \ x = y \\ 1, \ if \ x \neq y \end{cases}$$

Το μέτρο αυτό απλά ισούται με τον αριθμό των χαρακτηριστικών στα οποία διαφέρουν τα στιγμιότυπα. Πάνω σε αυτό, μπορούν να οριστούν και άλλα πιο εξελιγμένα μέτρα.

Ένα μειονέκτημα που παρουσιάζουν τα δύο προηγούμενα παραδείγματα μετρικών είναι πως όλα τα χαρακτηριστικά (features) θεωρούνται ισοδύναμα κατά τον υπολογισμό της απόστασης. Αυτό είναι ιδιαίτερα προβληματικό αν στην πραγματικότητα δεν είναι όλα τα χαρακτηριστικά σχετικά με τη συγκεκριμένη συνάρτηση-στόχο που επιδιώκεται να προσεγγιστεί, αλλά και γενικότερα, οποτεδήποτε υπάρχουν σημαντικές διαφορές μεταξύ τους ως προς την αξία τους στον προσδιορισμό της συνάρτησης. Σε μια τέτοια περίπτωση, οι παραπάνω μετρικές είναι παραπλανητικές, από την άποψη πως στιγμιότυπα που πραγματικά σχετίζονται μεταξύ τους, είναι δυνατόν να θεωρούνται απομακρυσμένα λόγω των διαφορών τους σε άσχετα ή ασήμαντα χαρακτηριστικά. Μια λύση σε αυτό το πρόβλημα είναι κάθε χαρακτηριστικό να αποτιμάται διαφορετικά στον υπολογισμό της απόστασης, ανάλογα με την αξία του. Αυτό αντιστοιχεί στο να επιμηκυνθούν οι άξονες στον Ευκλείδειο χώρο για τα σχετικά χαρακτηριστικά και να

συρρικνωθούν για τα λιγότερο σχετικά. Η μέθοδος αυτή λέγεται *αποτίμηση των χαρακτηριστικών (feature -weighting)* και είναι χρήσιμη και σε άλλες περιπτώσεις, πέραν της χρήσης της στη διαμόρφωση της μετρικής για τον *k-Nearest Neighbor*.

Μία βελτιωμένη παραλλαγή του *k-Nearest Neighbor*, όσον αφορά το συνδυασμό των τιμών των γειτόνων, είναι η αποτίμηση της συνεισφοράς καθενός από τους *k* γείτονες με βάση την απόσταση από το προς κατάταξη στιγμιότυπο, δίνοντας μεγαλύτερο βάρος στους κοντινότερους γείτονες. Αυτή αποτελεί τη με βάση την απόσταση *(distance-weighted)* εκδοχή του αλγορίθμου.

Ο *k-Nearest Neighbor* είναι ένας πολύ αποτελεσματικός αλγόριθμος μάθησης, τόσο για αριθμητικά όσο και για ονομαστικά δεδομένα, ιδιαίτερα όταν γίνεται με αποτίμηση χαρακτηριστικών και γειτόνων. Είναι ανθεκτικός σε θορυβώδη στιγμιότυπα εκπαίδευσης, ειδικά για μεγαλύτερες τιμές του *k,* καθώς τα απομονωμένα λανθασμένα δεδομένα "απορροφώνται" κατά τον υπολογισμό του μέσου όρου. Η επαγωγική κλίση του *k-Nearest Neighbor* είναι η υπόθεση πως η τιμή της συνάρτησης-στόχου ενός στιγμιότυπου είναι παρόμοια με αυτή των γειτονικών του.

Ένα πρακτικό θέμα κατά την εφαρμογή του *k-Nearest Neighbor* είναι η αποδοτική ευρετηριοποίηση των στιγμιότυπων στη μνήμη. Σε μια απλή υλοποίηση, η υπολογιστική πολυπλοκότητα για την κατάταξη ενός νέου στιγμιότυπου είναι ανάλογη του αριθμού των στιγμιότυπων εκπαίδευσης, αφού χρειάζεται να υπολογιστεί η απόσταση του νέου με κάθε στιγμιότυπο εκπαίδευσης, για να επιλεχθούν στη συνέχεια τα *k* κοντινότερα. Κάτι τέτοιο έχει υψηλότατο κόστος για μεγάλα σύνολα δεδομένων.

2.4.4 Μπαϊεζιανή Μάθηση

Η *Μπαϊεζιανή συλλογιστική (Bayesian reasoning)* παρέχει μια πιθανοτική προσέγγιση στο πρόβλημα του επαγωγικής συμπερασματικής λογικής. Στηρίζεται στην υπόθεση πως οι υπό μελέτη ποσότητες ακολουθούν πιθανοτικές κατανομές και πως οι βέλτιστες αποφάσεις μπορούν να παρθούν βάσει αυτών των κατανομών και των παρατηρούμενων δεδομένων. Στα πλεονεκτήματα της συγκαταλέγεται η δυνατότητα συνδυασμού της προϋπάρχουσας γνώσης με τα παρατηρούμενα δεδομένα, η θεώρηση πιθανοτικών (μη

ντετερμινιστικών) μοντέλων και η εκτίμηση της καταλληλότητας για κάθε μοντέλο, επιτρέποντας έτσι την εξέταση και εναλλακτικών μοντέλων πέραν του εκτιμώμενου βέλτιστου.

Στη μηχανική μάθηση, συχνά μας ενδιαφέρει να βρούμε την καλύτερη υπόθεση σε ένα χώρο *H* με βάση τα γνωστά δεδομένα *D*. Ένας τρόπος να καθορίσουμε τι εννοούμε λέγοντας καλύτερη είναι να απαιτήσουμε την *πιθανότερη* υπόθεση με βάση τα δεδομένα *D* και την τυχόν προηγούμενη γνώση για τις πιθανότητες των υποθέσεων στο *H*. Το θεώρημα του Μπαϊές (Bayes), το οποίο είναι ο ακρογωνιαίος λίθος της ομώνυμης συλλογιστικής, παρέχει έναν άμεσο τρόπο υπολογισμού της πιθανότητας για μια υπόθεση *h*. Η έκφραση του είναι η εξής:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

όπου: *P(h|D)* είναι η πιθανότητα να ισχύει η υπόθεση h με βάση τα παρατηρηθέντα δεδομένα *D* και καλείται *εκ των υστέρων πιθανότητα (posterior probability)* της *h*. *P(D|h)* είναι η πιθανότητα να παρατηρηθούν τα δεδομένα *D* σε κάποιο κόσμο που η υπόθεση h ισχύει και λέγεται *πιθανοφάνεια (likelihood)* των δεδομένων *D* δοθείσας της *h*. *P(h)* είναι η πιθανότητα να ισχύει η υπόθεση *h* πριν την παρατήρηση των δεδομένων και λέγεται *εκ των προτέρων πιθανότητα (prior probability)* της *h*. *P(D)* είναι η πιθανότητα να παρατηρηθούν τα δεδομένα *D* ανεξαρτήτως της υπόθεσης που ισχύει και λέγεται εκ των προτέρων πιθανότητα των δεδομένων *D*.

Σε πολλές περιπτώσεις, ο αλγόριθμος μάθησης θεωρεί ένα σύνολο υποψήφιων υποθέσεων *H* και αναζητεί την πιο πιθανή από αυτές δοθέντων των δεδομένων εκπαίδευσης. Μια τέτοια υπόθεση *h* λέγεται *μέγιστη εκ των υστέρων (maximum a posteriori* - MAP) υπόθεση. Ένας ευθύς τρόπος εύρεσης των MAP υποθέσεων είναι η εφαρμογή του θεωρήματος του Bayes για κάθε υπόθεση στο *H* και η επιλογή των μέγιστων από αυτές, δηλαδή:

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} = \arg\max_{h \in H} P(D \mid h)P(h)$$

Στο τελευταίο βήμα, το *P(D)* παραλήφθηκε γιατί είναι σταθερά ως προς τις υποθέσεις. Μερικές φορές δεν έχουμε καμιά εκ των προτέρων γνώση για τις υποθέσεις *h* και δεν έχουμε λόγω να πιστεύουμε πως είναι ανισοπίθανες. Τότε μπορούμε να θεωρήσουμε πως και ο όρος *P(h)* είναι σταθερός για όλες τις υποθέσεις και να τον απαλείψουμε και αυτόν από τον τύπο. Έτσι, η MAP υπόθεση θα είναι αυτή που μεγιστοποιεί την πιθανοφάνεια *P(D\h)* και η οποία λέγεται υπόθεση *μέγιστης πιθανοφάνειας (maximum likelihood – ML)*

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

Στην πράξη, περισσότερο από το ποια είναι η πιο πιθανή υπόθεση δοθέντων των δεδομένων μας ενδιαφέρει συνήθως το ποια είναι η πιο πιθανή τιμή της συνάρτησης-στόχου ενός νέου στιγμιότυπου δοθέντων των δεδομένων. Αν και μια απλή προσέγγιση είναι να θεωρήσουμε την τιμή της MAP υπόθεσης ως πιθανότερη τιμή, υπάρχει και καλύτερη λύση. Αυτή προκύπτει αν λάβουμε υπόψη τις προβλέψεις όλων των υποθέσεων, ζυγισμένες κατά την εκ των υστέρων πιθανότητά τους. Έτσι, αν η συνάρτηση-στόχος παίρνει τιμές σε ένα πεπερασμένο σύνολο *V*, τότε η πιθανότητα *P(V$_j$|X,D)* πως η σωστή τιμή για το στιγμιότυπο *x* είναι η *V$_j$*, δίνεται από τη σχέση:

$$P\left(V_j \mid x, D\right) = \sum_{h \in H} P_h\left(V_j \mid x\right) P\left(h \mid D\right)$$

όπου *P$_h$(Vj|*x) είναι η πιθανότητα να έχει το στιγμιότυπο *x* την τιμή *V$_j$* σύμφωνα με την υπόθεση *h.* Η παραπάνω σχέση, όπως φαίνεται, μπορεί να εφαρμοστεί και για μη ντετερμινιστικές υποθέσεις. Η βέλτιστη απόφαση είναι η τιμή *V$_j$* για την οποία το *P(V$_j$|X,D)* μεγιστοποιείται:

$$V_{opt} = \arg\max_{v_j \in V} \sum_{h \in H} P_h\left(V_j \mid x\right) P\left(h \mid D\right)$$

Ένα σύστημα που ταξινομεί τα στιγμιότυπα χρησιμοποιώντας την παραπάνω εξίσωση καλείται *βέλτιστος ταξινομητής Μπαϊεζ (Bayes optimal classifier).*

Απλοϊκός Ταξινομητής Μπαϊεζ (Naive Bayes)

Δύο πρακτικά προβλήματα εμφανίζονται στη χρήση του βέλτιστου ταξινομητή Μπαϊεζ. Το ένα είναι πως έχει γραμμική πολυπλοκότητα ως προς τον πληθυσμό *H* του χώρου υποθέσεων, γεγονός που καθιστά την εφαρμογή του αδύνατη για απειροδιάστατους χώρους και μη αποδοτική για μεγάλους πεπερασμένους χώρους. Το άλλο είναι πως απαιτεί τη γνώση ή την εκτίμηση πάρα πολλών πιθανοτήτων: την πιθανοφάνεια *P(D|h)* των δεδομένων *D* και την εκ των προτέρων πιθανότητα *P(h)* για κάθε υπόθεση *h*. Μία Μπαϊεζιανή μέθοδος που αντιμετωπίζει σε μεγάλο βαθμό αυτές τις δυσκολίες είναι ο απλοϊκός ταξινομητής Μπαϊεζ *(naive Bayes classifier* — NB για συντομία) [Lewis, 1998]).

Ο NB εφαρμόζεται σε προβλήματα μάθησης όπου τα στιγμιότυπα αναπαρίστανται μέσω του μοντέλου του διανυσματικού χώρου, τα χαρακτηριστικά παίρνουν διακριτές τιμές (αν κάποια είναι συνεχή, πρέπει να κβαντιστούν) και η συνάρτηση-στόχος παίρνει τιμές (ετικέτες—labels) σε ένα πεπερασμένο σύνολο *V*. Παρέχεται ένα σύνολο από διανύσματα εκπαίδευσης, βάσει του οποίου ο ταξινομητής πρέπει να προβλέψει την ετικέτα ενός νέου στιγμιότυπου αναπαριστώμενου από το διάνυσμα (α_1 , α_2 ,..., α_n).

Η Μπαϊεζιανή προσέγγιση στην κατάταξη του νέου στιγμιότυπου είναι η ανάθεση σε αυτό της πιο πιθανής τιμής V_{opt}, δεδομένων των τιμών των χαρακτηριστικών του, α_1 , α_2 ,..., α_n:

$$V_{opt} = \arg\max_{v_j \in V} P\left(v_j \mid a_1, a_2, ..., a_n\right)$$

η οποία μέσω του θεωρήματος του Μπαϊεζ εκφράζεται ως:

$$V_{opt} = \arg\max_{v_j \in V} \frac{P\left(a_1, a_2, ..., a_n \mid v_j\right) P\left(v_j\right)}{P\left(a_1, a_2, ..., a_n\right)} = \arg\max_{v_j \in V} P\left(a_1, a_2, ..., a_n \mid v_j\right) P\left(v_j\right)$$

Η εκτίμηση των πιθανοτήτων που εμφανίζονται στην εξίσωση πρέπει να γίνει μέσω των δεδομένων εκπαίδευσης. Οι $P(v_j)$ μπορούν να εκτιμηθούν εύκολα ως η συχνότητα εμφάνισης κάθε ετικέτας v_j στα δεδομένα. Το ίδιο όμως δε μπορεί να γίνει για τις $P(\alpha_1,$

$\alpha_2, ..., \alpha_n | v_j)$, δηλαδή τις πιθανότητες εμφάνισης κάθε δυνατού στιγμιότυπου δεδομένης μιας ετικέτας, αφού για συνηθισμένα μεγέθη συνόλων εκπαίδευσης τα περισσότερα στιγμιότυπα δεν θα έχουν εμφανιστεί, και επομένως η συχνότητα εμφάνισης τους θα είναι μηδέν, που προφανώς δεν είναι αξιόπιστη εκτίμηση της πραγματικής πιθανότητας εμφάνισης τους.

Ο απλοϊκός ταξινομητής Μπαϊεζ βασίζεται στην απλουστευτική υπόθεση πως οι τιμές των χαρακτηριστικών είναι ανεξάρτητες δοθείσας της ετικέτας. Τότε, η πιθανότητα της κοινής εμφάνισης των α_1 , α_2 ,..., α_n, δεδομένης μιας ετικέτας, είναι το γινόμενο των πιθανοτήτων εμφάνισης για καθένα από αυτά:

$$P\left(a_1, a_2, ...a_n \mid v_j\right) = \prod_{i=1}^{n} P\left(a_i \mid v_j\right)$$

Αντικαθιστώντας αυτή την έκφραση στην πιο πάνω εξίσωση έχουμε την έκφραση του NB:

$$V_{NB} = \arg\max_{v_j \in V} P\left(v_j\right) \prod_{i=1}^{n} P\left(a_i \mid v_j\right)$$

Από την εξίσωση αυτή φαίνεται πως το πλήθος των πιθανοτήτων $P(\alpha_i|v_j)$ που πρέπει να εκτιμηθούν επιπλέον των $P(v_j)$ ισούται με το πλήθος των διαφορετικών τιμών των χαρακτηριστικών επί το πλήθος των ετικετών, σημαντικά μικρότερο από αυτό που θα απαιτούνταν για όλες τις $P(\alpha_1 , \alpha_2 ,..., \alpha_n |v_j)$, ακόμα κι αν οι εκτιμήσεις τους ήταν αξιόπιστες. Έτσι, ο NB στη φάση εκπαίδευσής του εκτιμά με βάση τα δεδομένα τις $P(v_j)$ και $P(\alpha_i|v_j)$, το σύνολο των οποίων αποτελούν το μοντέλο ταξινόμησης που μαθαίνει, και στη φάση εξέτασης χρησιμοποιεί την εξίσωση V_{NB} για να κατατάξει κάθε νέο στιγμιότυπο. Ένα ενδιαφέρον χαρακτηριστικό του είναι πως δεν ερευνά το χώρο υποθέσεων για την εντοπισμό της καλύτερης υπόθεσης, όπως κάνουν πολλοί αλγόριθμοι μάθησης, αλλά σχηματίζει άμεσα ένα μοντέλο, απλά μετρώντας τη συχνότητα των συνδυασμών των τιμών των χαρακτηριστικών και των ετικετών μέσα στο σύνολο εκπαίδευσης.

Ο απλοϊκός ταξινομητής Μπαϊέζ, παρά την αρκετά δεσμευτική υπόθεση της υπό συνθήκη ανεξαρτησίας των χαρακτηριστικών, έχει να επιδείξει αναπάντεχα μεγάλη ακρίβεια και σε εφαρμογές που η υπόθεση της ανεξαρτησίας εμφανώς παραβιάζεται. Ένα ακόμα πλεονέκτημα του NB είναι η σχετική απλότητα των μοντέλων που κατασκευάζει, τα οποία μπορούν να γίνουν εύκολα κατανοητά από τον άνθρωπο.

3 ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

3.1 Εισαγωγή

Η Επιλογή Χαρακτηριστικών (feature selection ή attribute selection) είναι μια διαδικασία προ-επεξεργασίας, που χρησιμοποιείται ευρέως στο χώρο της μηχανικής μάθησης, κατά την οποία μια υποομάδα από τα αρχικά διαθέσιμα χαρακτηριστικά επιλέγεται, με κάποια κριτήρια, για την απώτερη επεξεργασία της από κάποιον αλγόριθμο μηχανικής μάθησης.

Η επιλογή χαρακτηριστικών είναι απαραίτητη είτε επειδή είναι μη πρακτική υπολογιστικά η χρήση όλων των δοθέντων χαρακτηριστικών, είτε επειδή τα δείγματα που παρέχει η βάση δεδομένων είναι περιορισμένα σε σχέση με τον αριθμό των χαρακτηριστικών. Το τελευταίο αυτό πρόβλημα είναι γνωστό και σαν "curse of dimensionality" (κατάρα της διαστατικότητας) και αναφέρεται στο γεγονός ότι ο αριθμός των δειγμάτων δεδομένων που απαιτείται για να εκτιμηθεί μια αυθαίρετη κατανομή πιθανοτήτων με πολλές μεταβλητές αυξάνεται εκθετικά όσο οι τιμές των διαστάσεων αυξάνονται γραμμικά. Από θεωρητική άποψης, μπορεί να δειχθεί ότι η βέλτιστη επιλογή χαρακτηριστικών για την διεξαγωγή ταξινόμησης υπό-επίβλεψη, απαιτεί εξαντλητική αναζήτηση όλων των πιθανών υποομάδων χαρακτηριστικών. Ωστόσο αυτό είναι μη πρακτικό στη περίπτωση ύπαρξης μεγάλου αριθμού χαρακτηριστικών. Πρακτικά, στην ταξινόμηση υπό-επίβλεψη η αναζήτηση γίνεται με σκοπό την εύρεση μιας ικανοποιητικής υποομάδας και όχι αναγκαστικά της καλύτερης. Για τον λόγω αυτό αρκετές δημοφιλείς προσεγγίσεις αναφέρονται ως *"greedy hill climbing"* προσεγγίσεις. Μια τέτοια προσέγγιση αποτιμά μια πιθανή υποομάδα χαρακτηριστικών και στη συνέχεια τη τροποποιεί για να δει αν μπορεί να προκύψει μια βελτιωμένη υποομάδα.

Η αποτίμηση των υποομάδων μπορεί να γίνει με πολλούς τρόπους: κάποιες μετρικές χρησιμοποιούνται για την αποτίμηση των χαρακτηριστικών και τους συνδυασμούς αυτών. Δύο δημοφιλείς μετρικές για ταξινόμηση είναι ο *συσχετισμός* (correlation) και *η κοινή πληροφορία* (mutual information). Αυτές οι μετρικές υπολογίζονται μεταξύ ενός υποψηφίου χαρακτηριστικού ή ομάδας χαρακτηριστικών και της επιθυμώμενης κλάσης.

Από τη στιγμή που η εξαντλητική αναζήτηση δεν είναι εφαρμόσιμη, θα πρέπει να οριστεί ένα *σημείο τερματισμού* (stopping point) στο οποίο θα επιλεγεί η υποομάδα χαρακτηριστικών που έχει συγκεντρώσει την υψηλότερη βαθμολογία στην καθορισμένη μετρική. Το θέμα, λοιπόν, της επιλογής του σημείου τερματισμού του αλγορίθμου είναι βασικό και το σημείο τερματισμού διαφέρει ανά αλγόριθμο.

Η επιλογή χαρακτηριστικών παρέχει πολλά πλεονεκτήματα στην διαδικασία ταξινόμησης. Μειώνει τον αριθμό χαρακτηριστικών, αποκρίνει τα μη σχετικά, τα πλεονάζοντα ή τα θορυβώδη δεδομένα και έχει άμεσα αποτελέσματα τόσο στην επιτάχυνση των αλγορίθμων μάθησης όσο και στην βελτίωση των αποτελεσμάτων τους. Βελτιώνει την ευστοχία ταξινόμησης σε νέα δεδομένα και παρέχει πιο συμπαγή αποτελέσματα καθιστώντας ευκολότερη την κατανόηση και την ερμηνεία του αντικειμένου της μάθησης.

Για τους παραπάνω λόγους, η Επιλογή Χαρακτηριστικών αποτελεί έναν γόνιμο χώρο έρευνας και ανάπτυξης από τη δεκαετία του 70 σε χώρους όπως αυτός της Στατιστικής Αναγνώρισης Προτύπων, της Μηχανικής Μάθησης και της Εξόρυξης Δεδομένων και έχει εφαρμογή και σε πολλούς άλλους χώρους.

Ειδικά την τελευταία δεκαετία, έρευνες έχουν δείξει πως ευρέως χρησιμοποιούμενοι αλγόριθμοι επηρεάζονται σε μικρότερο ή μεγαλύτερο βαθμό από μη-σχετικές ή πλεονάζουσες πληροφορίες εκπαίδευσης.

- Ο απλός αλγόριθμος Κοντινότερου Γείτονα (Nearest Neighbor) είναι πολύ ευαίσθητος σε μη-σχετικά χαρακτηριστικά, καθώς η πολυπλοκότητα του δείγματος (ο αριθμός δηλαδή των στιγμιότυπων εκπαίδευσης που απαιτείται για

να προσεγγιστεί ένα δεδομένο επίπεδο ευστοχίας) αυξάνεται εκθετικά σε σχέση με τον αριθμό των μη-σχετικών χαρακτηριστικών.

- Ο Απλοϊκός- Μπαϊεζιανός (Naïve- Bayes) ταξινομητής δεν επηρεάζεται από περιττά χαρακτηριστικά λόγω της υπόθεσης ανεξαρτησίας των χαρακτηριστικών που τον διακρίνει.

- Οι αλγόριθμοι Δέντρου Απόφασης παράγουν σχήματα τα οποία μπορεί κάποιες φορές να προσκολλώνται υπερβολικά στα δεδομένα εκπαίδευσης, σχηματίζοντας από αυτά μεγάλα δέντρα. Σε πολλές λοιπόν περιπτώσεις, η απομάκρυνση περιττών πληροφοριών μπορεί να οδηγήσει στην παραγωγή μικρότερων και πιο ευκολονόητων δέντρων.

Με σκοπό την αποφυγή της προσκόλλησης στα δεδομένα εκπαίδευσης (over fitting) πολλοί αλγόριθμοι χρησιμοποιούν μια μέθοδο αποκαλούμενη και ως στατιστική απόκλιση *Occam's Razor* για την κατασκευή ενός απλού μοντέλου ικανού να πετύχει ένα αποδεκτό επίπεδο απόδοσης. Αυτή η στατιστική απόκλιση οδηγεί έναν αλγόριθμο να επιλέξει ένα μικρό αριθμό χαρακτηριστικών για την πρόβλεψη και να στηριχθεί μόνο στα χαρακτηριστικά με τη μεγαλύτερη συνάφεια. Όπως είναι όμως κατανοητό η μέθοδος αυτή έχει σαν αποτέλεσμα μειωμένα ποσοστά ταξινόμησης σε σχέση με αυτά που θα είχε αν χρησιμοποιούσε όλα τα χαρακτηριστικά.

3.2 Προσεγγίσεις της Επιλογής Χαρακτηριστικών

Μια τυπική διαδικασία Επιλογής Χαρακτηριστικών αποτελείται από τέσσερα βασικά βήματα: δημιουργία υποομάδας, αξιολόγηση υποομάδας, κριτήριο τερματισμού και επιβεβαίωση αποτελέσματος. Παρακάτω παρουσιάζονται το καθένα απ' αυτά αναλυτικά.

3.2.1 Δημιουργία Υποομάδας

Είναι, στην ουσία, μια ευρετική διαδικασία αναζήτησης στην οποία κάθε κατάσταση στο χώρο αναζήτησης προσδιορίζει μια υποψήφια για αξιολόγηση υποομάδα. Η φύση αυτού του βήματος καθορίζεται από δύο βασικά θέματα:

- Το *σημείο εκκίνησης* της αναζήτησης το οποίο διαδοχικά επηρεάζει την *κατεύθυνση αναζήτησης*: η αναζήτηση μπορεί να ξεκινήσει με ένα κενό σύνολο και την σταδιακή προσθήκη χαρακτηριστικών (αυτή η στρατηγική αναζήτησης ονομάζεται *εμπρόσθια επιλογή* (forward selection), ή να ξεκινήσει από ολόκληρο το σετ χαρακτηριστικών και να συνεχίσει με την σταδιακή απομάκρυνση κάποιων εξ' αυτών (backward selection) ή, τέλος, να ξεκινήσει και από τα δύο άκρα και να προσθαφαιρέσει χαρακτηριστικά ταυτόχρονα (bidirectional selection). Η αναζήτηση μπορεί επίσης να ξεκινήσει από μια τυχαία επιλεγμένη υποομάδα με σκοπό το να αποφύγει να παγιδευτεί σε τοπικά μέγιστα.

- Η *στρατηγική αναζήτησης*. Για ένα σετ δεδομένων με *N* χαρακτηριστικά, υπάρχουν 2^N υποψήφια υπό-σετ. Ο χώρος αναζήτησης είναι λοιπόν απαγορευτικός για εξαντλητική αναζήτηση ακόμα και με σχετικά μικρές τιμές του *N*. Γι' αυτό έχουν αναπτυχθεί τρεις διαφορετικές στρατηγικές αναζήτησης: η *πλήρης*, η *ακολουθητική* και η *τυχαία*.

    - Η *Πλήρης Αναζήτηση* εγγυάται την εύρεση του βέλτιστου αποτελέσματος σύμφωνα με το επιλεγμένο κριτήριο αξιολόγησης. Ενώ μια εξαντλητική αναζήτηση είναι πάντα πλήρης, διαφορετικές ευρετικές συναρτήσεις μπορούν να χρησιμοποιηθούν για να μειώσουν τον χώρο αναζήτησης χωρίς ωστόσο να χάσουν την ευκαιρία εύρεσης του βέλτιστου αποτελέσματος. Τέτοιες μέθοδοι είναι οι "branch and bound" (Narendra & Fukunaga, 1977) και "beam search" (Doak, 1992).

    - Η *ακολουθητική αναζήτηση* εγκαταλείπει την πλήρη αναζήτηση ρισκάροντας έτσι να χάσει κάποια βέλτιστη υποομάδα. Υπάρχουν πολλές παραλλαγές στη κλασσική προσέγγιση greedy hill-climbing, όπως η ακολουθητική εμπρόσθια αναζήτηση (sequential forward selection), η ακολουθητική οπισθοδρομική εξάλειψη (sequential backward elimination) και η αναζήτηση δυο κατευθύνσεων (bidirectional selection). Οι αλγόριθμοι με ακολουθητική αναζήτηση είναι απλοί στην εφαρμογή και

γρήγοροι στο να παράγουν αποτελέσματα καθώς ο χώρος αναζήτησης είναι $O(N^2)$ ή λιγότερο.

- Η *τυχαία αναζήτηση* ξεκινάει με ένα τυχαία επιλεγμένο υπό-σετ χαρακτηριστικών και προχωράει με δύο διαφορετικούς τρόπους. Ένας είναι η ακολουθητική αναζήτηση, με εισαγωγή τυχαιότητας στην κλασσική προσέγγιση. Παραδείγματα είναι οι μέθοδοι "random-star hill-climbing" και "simulated annealing". Ο άλλος είναι η δημιουργία του επόμενου υπό-σετ με έναν εντελώς τυχαίο τρόπο, γνωστό και ως "Las Vegas algorithm". Για όλες αυτές τις προσεγγίσεις η χρήση της τυχαιότητας βοηθάει τη διαφυγή από τοπικά μέγιστα στο χώρο αναζήτησης και η καταλληλότητα του επιλεγμένου υπό-σετ εξαρτάται από τους διαθέσιμους πόρους.

3.2.2 Αξιολόγηση Υποομάδας

Το πως αξιολογούνται οι υποομάδες χαρακτηριστικών είναι ο μεγαλύτερος μεμονωμένος διακριτικός παράγοντας μεταξύ όλων των αλγορίθμων επιλογής χαρακτηριστικών. Για παράδειγμα, το *μοντέλο διήθησης* (filter model) λειτουργεί ξεχωριστά από κάθε αλγόριθμο μάθησης και ανεπιθύμητα χαρακτηριστικά απομακρύνονται από τα υπόλοιπα δεδομένα πριν ξεκινήσει η διαδικασία μάθησης.

Αυτοί οι αλγόριθμοι χρησιμοποιούν ευρευτικές μεθόδους βασισμένες στα γενικά χαρακτηριστικά των δεδομένων για να αξιολογήσουν την αξία κάθε υποομάδας χαρακτηριστικών.

Κατά μια άλλη προσέγγιση, τα επιλεγμένα χαρακτηριστικά πρέπει να εξαρτώνται όχι μόνο από τη σχετικότητα των δεδομένων με το θέμα, αλλά και από τον ίδιο τον χρησιμοποιούμενο αλγόριθμο. Αυτή είναι η λεγόμενη προσέγγιση ενσωμάτωσης (wrapper approach).

Και στις δύο, πάντως, περιπτώσεις κάθε εξεταζόμενη υποομάδα αξιολογείται από κάποιο κριτήριο. Τα κριτήρια αξιολόγησης μπορούν να διακριθούν σε δύο ομάδες ανάλογα με το

αν εξαρτώνται ή όχι από τον αλγόριθμο κατάταξης που θα χρησιμοποιηθεί στη συνέχεια στην υποομάδα χαρακτηριστικών που θα επιλεγεί.

- *Ανεξάρτητα κριτήρια* (Independent Criteria). Συνήθως χρησιμοποιούνται σε αλγόριθμους μοντέλου διήθησης καθώς προσπαθούν να αποτιμήσουν την χρησιμότητα ενός χαρακτηριστικού ή μιας υποομάδας χαρακτηριστικών εκμεταλλευόμενοι τα εσωτερικά χαρακτηριστικά των δεδομένων εκπαίδευσης χωρίς να περιλαμβάνουν κάποιον αλγόριθμο εξόρυξης δεδομένων. Κάποια ανεξάρτητα κριτήρια αναφέρονται παρακάτω:

  - *Μέτρα απόστασης (Distance measures)*. Για ένα πρόβλημα δύο κλάσεων ένα χαρακτηριστικό X προτιμάται από ένα άλλο χαρακτηριστικό Y αν το X προκαλεί μεγαλύτερη διαφορά μεταξύ των υπό συνθήκη πιθανοτήτων μεταξύ των δυο κλάσεων από το Y. Ο γενικός στόχος είναι η προσπάθεια να βρεθεί ένα χαρακτηριστικό που να μπορεί να διαχωρίσει τις δύο κλάσεις όσο το δυνατό πιο ευδιάκριτα.

  - *Μέτρα πληροφορίας (Information measures)*. Τυπικά καθορίζουν το κέρδος πληροφορίας για καθένα από τα χαρακτηριστικά. Το κέρδος πληροφορίας για ένα χαρακτηριστικό X ορίζεται σαν η διαφορά ανάμεσα στην a priori και την αναμενόμενη a posteriori αβεβαιότητα όσον αφορά το X στο τελικό υπό-σετ. Το χαρακτηριστικό X λοιπόν, προτιμάται από το χαρακτηριστικό Y εφ' όσον η πληροφορία που προσφέρεται από το X είναι περισσότερη από αυτή που προσφέρεται από το Y.

  - *Μέτρα εξάρτησης (Dependency measures)* ή αλλιώς *μέτρα συσχέτισης (correlation measures)*. Μετρούν την ικανότητα πρόβλεψης της τιμής μιας μεταβλητής από την τιμή μιας άλλης. Στην επιλογή χαρακτηριστικών για ταξινόμηση, ο βασικός στόχος είναι η αναζήτηση του πόσο ισχυρά συνδεδεμένο είναι ένα χαρακτηριστικό με μια από τις κλάσεις. Ένα χαρακτηριστικό X προτιμάται από ένα άλλο χαρακτηριστικό Y αν η σχέση μεταξύ ενός χαρακτηριστικού X και της κλάσης C είναι μεγαλύτερη από αυτή μεταξύ Y και C.

- *Μέτρα συνέπειας (Consistency measures)*. Αυτά τα μέτρα προσπαθούν να εντοπίσουν έναν ελάχιστο αριθμό χαρακτηριστικών που διαχωρίζουν τις κλάσεις με τόση συνέπεια όση και ολόκληρο το σετ χαρακτηριστικών.

- *Εξαρτημένα κριτήρια*. Κάθε εξαρτημένο κριτήριο που χρησιμοποιείται στο μοντέλο ενσωμάτωσης βασίζεται σε ένα γνωστό αλγόριθμο ταξινόμησης για να εκτελέσει την επιλογή χαρακτηριστικών καθώς χρησιμοποιεί την επίδοση αυτού του αλγορίθμου στο υπό-σετ που αξιολογείται για να καθορίσει ποια χαρακτηριστικά θα επιλεγούν. Αυτή η προσέγγιση συνήθως οδηγεί στη βέλτιστη απόδοση καθώς βρίσκει ομάδες χαρακτηριστικών που ταιριάζουν καλύτερα στον συγκεκριμένο αλγόριθμο, ωστόσο έχει μεγάλο υπολογιστικό κόστος και δεν είναι κατάλληλο για όλους τους αλγορίθμους.

3.2.3 Κριτήριο Τερματισμού

Ένας επιλογέας χαρακτηριστικών πρέπει να είναι ικανός να αποφασίσει πότε πρέπει να σταματήσει την αναζήτηση στο χώρο των υποομάδων χαρακτηριστικών. Ανάλογα με την στρατηγική αξιολόγησης, ένας επιλογέας χαρακτηριστικών θα έπρεπε να σταματάει να προσθέτει ή να αφαιρεί χαρακτηριστικά όταν καμία από τις εναλλακτικές δεν βελτιώνει την αξία του δεδομένου υπό-σετ χαρακτηριστικών. Εναλλακτικά, ο αλγόριθμος θα μπορούσε να συνεχίσει να ερευνά όσο η αξία του υπό-σετ δεν μειώνεται. Μια επιπλέον επιλογή θα μπορούσε να είναι η συνέχεια δημιουργίας υποομάδων χαρακτηριστικών ως την προσέγγιση του άλλου άκρου του χώρου αναζήτησης και στη συνέχεια η επιλογή της συνολικά καλύτερης υποομάδας.

Κάποια συχνά χρησιμοποιούμενα κριτήρια τερματισμού είναι τα παρακάτω:

- Ολοκλήρωση αναζήτησης

- Προσέγγιση κάποιου δεδομένου ορίου, όπου το όριο είναι ένα προκαθορισμένο νούμερο, όπως έναν ελάχιστο αριθμό χαρακτηριστικών ή ένα μέγιστο αριθμό επαναλήψεων.

- Η διαδοχική προσθήκη (ή διαγραφή) κάθε χαρακτηριστικού δεν βελτιώνει την υπάρχουσα υποομάδα.

- Μια αρκούντως καλή υποομάδα έχει επιλεγεί (μια υποομάδα μπορεί να χαρακτηριστεί αρκούντως καλή αν το σφάλμα ταξινόμησης της είναι μικρότερο από ένα καθορισμένο επιτρεπτό όριο).

3.2.4 Επικύρωση Αποτελέσματος

Ένας απλός τρόπος για επικύρωση του αποτελέσματος είναι η απευθείας μέτρηση των αποτελεσμάτων χρησιμοποιώντας την προηγούμενη γνώση για τα δεδομένα. Στην περίπτωση που τα σχετικά χαρακτηριστικά είναι εκ' των προτέρων γνωστά θα μπορούσε να συγκριθεί η παραγόμενη υποομάδα με το δεδομένο σετ και να αξιολογηθεί η επιλεκτική ικανότητα του αλγόριθμου. Όμως στα αληθινά προβλήματα τέτοια εκ' των προτέρων γνώση δεν υπάρχει. Γι' αυτό το λόγω, εφικτή είναι μόνο η χρήση έμμεσων μεθόδων μέσω την καταγραφής της απόκλισης της επίδοσης ενός ταξινομητή με την αλλαγή των χαρακτηριστικών.

3.3 Αλγόριθμοι Επιλογής Χαρακτηριστικών

Αναφέρθηκαν πιο πάνω, τα κριτήρια επιλογής και αξιολόγησης υποομάδας σε έναν αλγόριθμο επιλογής χαρακτηριστικών. Με κριτήριο τον διαχωρισμό τους σε εξαρτημένα και ανεξάρτητα κριτήρια οι αλγόριθμοι χωρίζονται στις παρακάτω δύο κατηγορίες:

- Στους αλγόριθμους ενσωμάτωσης (wrapper algorithms)

- Στους αλγόριθμους διήθησης ή φίλτρου (filter algorithms)

3.3.1 Αλγόριθμοι Ενσωμάτωσης

Όπως έχει αναφερθεί σε σχετικές δημοσιεύσεις (Kohavi, John and Pfleger,1994), όταν ο σκοπός είναι η μεγιστοποίηση της ευστοχίας μιας δεδομένης υποομάδας χαρακτηριστικών, τα επιλεγμένα χαρακτηριστικά πρέπει να εξαρτώνται όχι μόνο από σχετικότητα των δεδομένων με το θέμα, αλλά και από τον ίδιο τον χρησιμοποιούμενο

αλγόριθμο. Αυτή είναι η λεγόμενη προσέγγιση ενσωμάτωσης. Συμβαίνει σε αρκετές περιπτώσεις ένα χαρακτηριστικό, αν και φαινομενικά είναι σχετικό με το αντικείμενο της μάθησης, να μην περιλαμβάνεται στην βέλτιστη υποομάδα που μεγιστοποιεί την ευστοχία πρόβλεψης του χρησιμοποιούμενου αλγόριθμου. Πράγμα πού φανερώνει ότι η σχετικότητα ενός χαρακτηριστικού με το θέμα της μάθησης και η συμβολή του στη μεγιστοποίηση της ευστοχίας δεν συμβαδίζουν υποχρεωτικά. Παρακάτω εξετάζονται τόσο ο αλγόριθμος Wrapper όσο και μια παραλλαγή του ο Classifier Subset Evaluation.

Αλγόριθμος Wrapper

Κατά την "προσέγγιση ενσωμάτωσης" ο αλγόριθμος επιλογής αναζητά μια καλή υποομάδα χρησιμοποιώντας τον ίδιο τον επαγωγικό αλγόριθμο που θα χρησιμοποιηθεί για την τελική ταξινόμηση, στη διαδικασία αναζήτησης της. Σε αυτή τη διαδικασία αναζήτησης ο αλγόριθμός Wrapper εκτελεί κάθε φορά τη διαδικασία εκπαίδευσης για κάθε υποομάδα που προτείνεται από τη μηχανή αναζήτησης και την αξιολογεί με τη βοήθεια της μεθοδολογίας 5-fold cross-validation.

Το ελάττωμα της προσέγγισης αυτής είναι το μεγάλο υπολογιστικό κόστος, όμως η συνεχής ανάπτυξη των τεχνικών δυνατοτήτων των υπολογιστών καθιστά δυνατή την εφαρμογή της σε ολοένα μεγαλύτερο πλήθος εφαρμογών. Πολλές προσεγγίσεις έχουν κατά καιρούς προταθεί με σκοπό να ελαφρύνει η υπερφόρτωση που προκαλείται κατά την φάση εκπαίδευσης, κυρίως μέσω της χρήσης πιο απλών Μπαϊεζιανών αλγόριθμων, όμως το πρόβλημα παραμένει και καθιστά τη μέθοδο αυτή δύσχρηστη σε προβλήματα μεγάλης ποσότητας δεδομένων.

Αλγόριθμος Classifier Subset Evaluation

Ένας ακόμα αλγόριθμος της ίδιας φιλοσοφίας με την προσέγγιση ενσωμάτωσης ο οποίος χρησιμοποιεί έναν αλγόριθμο κατάταξης στη διαδικασία επιλογής υποομάδας. Η βασική του διαφορά από τον Wrapper είναι ότι κατά τη διαδικασία αξιολόγησης της εξεταζόμενης υποομάδας δεν χρησιμοποιεί τη μεθοδολογία cross-validation. Διενεργεί την αξιολόγηση ελέγχοντας την εκάστοτε υποομάδα είτε μέσα από τα δεδομένα εκπαίδευσης ή σε κάποια ξεχωριστά δεδομένα ελέγχου που ο χρήστης έχει προβλέψει να

κρατήσει ως δεδομένα ελέγχου. Με τον τρόπο αυτό η συνολική διαδικασία επιταχύνεται αρκετά σε σχέση με τον προαναφερθέντα Wrapper, το αντίτιμο όμως είναι η απώλεια της αξιοπιστίας που προσφέρει η μέθοδος cross-validation.

3.3.2 Αλγόριθμοι Διήθησης

Όταν στη λειτουργία αξιολόγησης δεν χρησιμοποιείται ο αλγόριθμός μάθησης, η κρίση μιας υποομάδας χαρακτηριστικών μπορεί να αξιολογηθεί μόνο βάση των εσωτερικών ιδιοτήτων των δεδομένων και λαμβάνοντας υπόψη μόνο τον στόχο του θέματος προς μάθηση. Αυτή η προσέγγιση επιλογής είναι γνωστή ως προσέγγιση διήθησης και εφαρμόζεται πριν τον αλγόριθμο ταξινόμησης με σκοπό την κατάταξη ή την περικοπή περιττών χαρακτηριστικών. Για το σκοπό αυτών χρησιμοποιούνται τα ανεξάρτητα κριτήρια της καταλληλότητας των χαρακτηριστικών που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, δανεισμένα από τον χώρο της στατιστικής αξιολόγησης προτύπων και της εξόρυξης δεδομένων. Παρακάτω παρουσιάζονται μια σειρά από αλγόριθμους διήθησης που στην πορεία της εργασίας αυτής θα χρησιμοποιηθούν.

Relief Attribute Ranking

Το Relief (Kira and Rendell) είναι ένα σχήμα κατάταξης χαρακτηριστικών που βασίζεται στη στιγμιαία μάθηση. Ο Relief λειτουργεί μέσω της τυχαίας δειγματοληψίας μιας περιπτώσεως από τα δεδομένα και εντοπίζοντας στη συνέχεια τον κοντινότερό του γείτονα από της ίδιας και της αντίθετης κλάσης. Οι τιμές των χαρακτηριστικών των κοντινότερων γειτόνων συγκρίνονται με το δείγμα και χρησιμοποιούνται για να αναβαθμίσουν τα σκορ σχετικότητας για το κάθε χαρακτηριστικό. Η διαδικασία επαναλαμβάνεται για ένα νούμερο περιπτώσεων m το οποίο προσδιορίζεται απ' τον χρήστη. Η λογική στην οποία στηρίζεται αυτό το σχήμα είναι η εξής: ένα χρήσιμο χαρακτηριστικό θα 'πρεπε να διαφοροποιείται μεταξύ περιπτώσεων δειγμάτων που ανήκουν σε διαφορετικές κλάσεις και να λαμβάνει ίδιες τιμές με δείγματα που ανήκουν στην ίδια με αυτό κλάση.

Το Relief ήταν αρχικά ορισμένο για προβλήματα δύο κλάσεων, στη συνέχεια προεκτάθηκε η λειτουργία του, με την ανάπτυξη του ReliefF (Kononenko, 1994), ώστε

να μπορεί να αντιμετωπίσει προβλήματα με περισσότερες κλάσεις και την ύπαρξη θορύβου. Ο ReliefF εξομαλύνει την επιρροή του θορύβου στα δεδομένα λαμβάνοντας υπόψη τον μέσο όρο της συνεισφοράς των κ κοντινότερων γειτόνων από την ίδια και την αντίθετη κλάση αντί για μια περίπτωση απ' την κάθε κλάση. Τα σετ δεδομένων με πολλές κλάσεις αντιμετωπίζονται με την εύρεση κοντινότερων από την καθεμιά από τις κλάσεις και σταθμίζοντας την συνεισφορά τους με την προηγούμενη πιθανότητα κάθε κλάσης. Όταν συγκρίνονται χαρακτηριστικά που λαμβάνουν ονομαστικές τιμές η διαφορά είναι είτε μηδέν (αν έχουν ίδια τιμή) είτε ένα (αν έχουν διαφορετική τιμή), για τα χαρακτηριστικά που λαμβάνουν συνεχείς τιμές η πραγματική απόσταση εξομαλύνεται στο διάστημα [0,1].

Information Gain Attribute Ranking

Αυτή είναι μια από τις πιο απλές και πιο γρήγορες μεθόδους κατάταξης χαρακτηριστικών και χρησιμοποιείται συχνά σε προβλήματα εξόρυξης γνώσης από μεγάλες βάσεις δεδομένων που είναι απαραίτητη η διαλογή των χαρακτηριστικών και το μέγεθος τους αποκλείει τη χρήση πιο πολύπλοκων μεθόδων.

Αν Α είναι ένα εκ των χαρακτηριστικών και C η μια εκ των κλάσεων, η εντροπία της κλάσης δίνεται από την εξίσωση: $H(C) = -\sum_{c \in A} p(c) \log_2 p(c)$. Η εντροπία της κλάσης C με δεδομένη συμπεριφορά του χαρακτηριστικού Α δίνεται αντίστοιχα από την: $H(C \mid A) = -\sum_{a \in A} p(a) \sum_{c \in A} p(c \mid a) \log_2 p(c \mid a)$. Το ποσό της μείωσης της εντροπίας αντικατοπτρίζει την επιπρόσθετη πληροφορία για τον προσδιορισμό της κλάσης C που παρέχεται από το χαρακτηριστικό Α και αποκαλείται "κέρδος πληροφορίας" (information Gain). Για καθένα από τα χαρακτηριστικά A_i υπολογίζεται ένα σκορ βασισμένο στο κέρδος πληροφορίας του A_i και της κλάσης που υπολογίζεται ως εξής:

$$\begin{aligned} IG_i &= H(C) - H(C|A_i) \\ &= H(A_i) - H(A_i|C) \\ &= H(A_i) + H(C) - H(A_i,C) \end{aligned}$$

Τα αριθμητικά χαρακτηριστικά πρώτα κβαντίζονται μέσω της μεθόδου Fayyad and Irani.

Gain Ratio Attribute Ranking

Όταν κάποια χαρακτηριστικά έχουν μεγάλο εύρος πιθανών τιμών οδηγούν σε έναν πιθανό καταμερισμό με πολλά παιδιά-κόμβους. Σε μια ακραία περίπτωση ενός χαρακτηριστικού που λαμβάνει διαφορετική τιμή για την καθεμία εκ' των περιπτώσεων του εξεταζόμενου σετ δεδομένων το μέτρο του κέρδους πληροφορίας για το χαρακτηριστικό αυτό θα μεγιστοποιούταν χωρίς ωστόσο αυτό να αντικατοπτρίζει την πραγματική διαχωριστική του αξία. Για τον λόγω αυτό κρίνεται αναγκαία η δημιουργία ενός νέου μέτρου το οποίο θα λαμβάνει υπόψη τον αριθμό και το μέγεθος των παιδιών-κόμβων που προκύπτουν από το εκάστοτε χαρακτηριστικό. Το μέτρο αυτό ονομάζεται "αναλογία κέρδους" (Gain Ratio) και υπολογίζεται ως από το παρακάτω πηλίκο:

$H(C) / H(C|A_i)$

Δυστυχώς σε κάποιες περιπτώσεις η αναλογία κέρδους μπορεί να οδηγήσει στην επιλογή ενός χαρακτηριστικού απλώς επειδή η εσωτερική του πληροφορία είναι πολύ μικρότερη από των άλλων χαρακτηριστικών. Για τον λόγω αυτό είναι απαραίτητος ο συμψηφισμός του με την τιμή του κέρδους πληροφορίας.

Correlation-based Feature Selection – CFS

Η CFS η πρώτη από μια σειρά μεθόδων που επικυρώνουν ομάδες χαρακτηριστικών αντί για μεμονωμένα. Στην καρδιά του αλγόριθμου βρίσκεται μια ευρετική μέθοδος αξιολόγησης υποομάδων που λαμβάνει υπόψη την χρησιμότητα των ανεξάρτητων χαρακτηριστικών στο να προβλέπουν την κλάση σε συνδυασμό με τον βαθμό συσχέτισης μεταξύ τους. Η παρακάτω ευριστική σχέση δίνει υψηλά σκορ σε υποομάδες οι οποίες περιέχουν χαρακτηριστικά με υψηλό βαθμό συσχέτισης με την κλάση και ταυτόχρονα χαμηλή αλληλο-συσχέτιση μεταξύ τους.

$$Merit_s = \frac{k\overline{r}_{ef}}{\sqrt{k + k(k-1)\overline{r}_{ff}}}$$

Όπου $Merit_s$ είναι η ευριστική αξία μιας υποομάδας αντικειμένων S η οποία περιέχει κ χαρακτηριστικά, $\overline{r}_{ef}$ η μέση τιμή συσχέτισης χαρακτηριστικού – κλάσης και $\overline{r}_{ff}$ η μέση τιμή αλληλο-συσχέτισης χαρακτηριστικού με άλλο χαρακτηριστικό. Ο αριθμητής είναι, κατά κάποιο τρόπο, δείκτης του πόσο καλή ικανότητα πρόβλεψης έχει η ομάδα χαρακτηριστικών που εξετάζεται. Ο παρονομαστής δείχνει το καταπόσο υπάρχει περιττή επανάληψη πληροφοριών άρα πλεονάζοντα χαρακτηριστικά. Η ευρετική μέθοδος που χρησιμοποιείται χειρίζεται επίσης καλά την τυχόν παρουσία άσχετων χαρακτηριστικών οι οποίοι θα εμφανίζονται ως κακοί εκτιμητές της κλάσης. Αν και τα περιττά χαρακτηριστικά διαχωρίζονται αποτελεσματικά μέσω της ισχυρής συσχέτισης που τα χαρακτηρίζει, ωστόσο λόγω του ότι το κάθε χαρακτηριστικό αντιμετωπίζεται ξεχωριστά, ο CFS, δεν μπορεί να αναγνωρίσει αυτά που είναι ισχυρά αλληλεπιδρώντα.

Όπως μπορεί κανείς να καταλάβει ο υπολογισμός της συσχέτισης μεταξύ των χαρακτηριστικών είναι απαραίτητος πριν την εφαρμογή της ευριστικής διαδικασίας. Ο CFS αφού μετατρέπει τα αριθμητικά χαρακτηριστικά σε διακριτά με τη μέθοδο Fayyad and Irani, στη συνέχεια χρησιμοποιεί τη μέθοδο συμμετρικής αβεβαιότητας (symmetrical uncertainty) για να εκτιμήσει το βαθμό συσχέτισης μεταξύ των διακριτών χαρακτηριστικών (X και Y):

$$SU = 2.0 \times \left[\frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)} \right].$$

Μετά τον υπολογισμό του πίνακα συσχετίσεων, ο CFS εφαρμόζει ευριστική στρατηγική αναζήτησης για να βρει μια καλή υποομάδα χαρακτηριστικών στην οποία τα χαρακτηριστικά θα είναι διατεταγμένα σύμφωνα με την συνεισφορά τους στην ικανότητα της υποομάδας.

Consistency-Based Subset Evaluation

Αρκετές προσεγγίσεις στην διαλογή υποομάδας χαρακτηριστικών χρησιμοποιούν την συνάφεια με την κλάση σαν μέτρο αξιολόγησης. Ο βασισμένος στη συνάφεια αξιολογητής υποομάδων που θα εξεταστεί στη συνέχεια προτάθηκε από τους Liu και Setiono μετράει τη συνάφεια με τον παρακάτω τύπο:

$$Consistency_s = 1 - \frac{\sum_{i=0}^{J} |D_i| - |M_i|}{N}$$

Όπου ως *s* ορίζεται μια υποομάδα χαρακτηριστικών, *J* είναι ο αριθμός των ξεχωριστών συνδυασμών τιμών χαρακτηριστικών του *s*, $|D_i|$ είναι ο αριθμός περιστατικών του i-οστού συνδυασμού τιμών χαρακτηριστικών, $|M_i|$ είναι ο αριθμός των στοιχείων του συνόλου της πλειοψηφούσας κλάσης για τον i-οστό συνδυασμό τιμών των χαρακτηριστικών, τέλος, N είναι ο συνολικός αριθμός δειγμάτων του εξεταζόμενου σετ δεδομένων.

Και η μέθοδος αυτή λειτουργεί με ονομαστικά χαρακτηριστικά οπότε τα αριθμητικά χαρακτηριστικά μετατρέπονται σε διακριτά με τη μέθοδο των Fayyad and Irani.

OneR Attribute Ranking

Ο OneR είναι ένας απλός τρόπος για την παραγωγή απλών κανόνων ταξινόμησης από ένα σύνολο στιγμιότυπων. Αποκαλείται OneR από το "1-rule" και παράγει ένα δέντρο απόφασης ενός επιπέδου, το οποίο μπορεί να εκφραστεί στην μορφή ενός συνόλου κανόνων που όλοι ελέγχουν ένα μόνο χαρακτηριστικό. Η μέθοδος 1-rule, είναι μια απλή και γρήγορη υπολογιστικά μέθοδος που συχνά ανακαλύπτει καλούς κανόνες για τον χαρακτηρισμό της δομής των δεδομένων αλλά είναι ταυτόχρονα και ένας καλός τρόπος διαλογής χαρακτηριστικών. Η ιδέα είναι η κατασκευή κανόνων που ελέγχουν ένα απλό χαρακτηριστικό και δημιουργία τόσων κόμβων όσες είναι οι διαφορετικές τιμές που λαμβάνει το κάθε χαρακτηριστικό, η κατάταξη αυτών επιτυγχάνεται με κριτήριο την ικανότητα ταξινόμησης για το κάθε χαρακτηριστικό. Ο ρυθμός λαθών των κανόνων μπορεί να καθοριστεί με την καταμέτρηση των λάθος ταξινομημένων περιπτώσεων στα δεδομένα εκπαίδευσης, δηλαδή, ο αριθμός των περιπτώσεων που δεν έχουν την κλάση της πλειοψηφίας. Κάθε χαρακτηριστικό παράγει ένα διαφορετικό σύνολο κανόνων και ένας κανόνας παράγεται για κάθε τιμή του χαρακτηριστικού. Υπολογίζοντας τον ρυθμό λαθών για το σύνολο κανόνων (rule set) του κάθε χαρακτηριστικού μπορούν αυτά να καταταγούν ιεραρχικά.

Chi-squared Attribute Selection

Μια προσέγγιση βασισμένη στο στατιστικό χ^2 χρησιμοποιείται συχνά για να χαρακτηρίσει τη σημαντικότητα της σχέσης μεταξύ μεταβλητών. Έστω ότι A είναι ένα χαρακτηριστικό και C η ταμπέλα κλάσης. Έστω ότι το dataset περιέχει N_i χαρακτηριστικά με το χαρακτηριστικό A ίσο με A_i, όπου $1 \leq i \leq k$, όπου k ο αριθμός των διακριτών τιμών που παίρνει ο A, τα N_j χαρακτηριστικά ανήκουν στην κλάση C_j, όπου $1 \leq j \leq l$, ;όπου l ο αριθμός των κλάσεων. Έστω ότι ο N_{ij} εκφράζει των αριθμό των χαρακτηριστικών που ανήκουν στην κλάση C_j με το χαρακτηριστικό A να παίρνει την τιμή A_i. Αν η ταμπέλα της κλάσης δεν έχει σχέση με το χαρακτηριστικό A, το αναμενόμενο νούμερο κάθε N_{ij}, που συμβολίζεται με n_{ij} μπορεί να υπολογιστεί από τα N_i και N_j:

$$n_{ij} = \frac{N_i \cdot N_j}{N}$$

Όπου N είναι ο συνολικός αριθμός των χαρακτηριστικών στο dataset. Το χ^2 στατιστικό τώρα δίνεται από τον τύπο

$$x^2 = \sum_i \sum_j \frac{\left(N_{ij} - n_{ij}\right)^2}{n_{ij}} .$$

Όπως είναι προφανές η τιμή της χ^2 εξαρτάται από το μέγεθος του dataset. Ένας τρόπος για να ποσοτικοποιηθεί η δύναμη της σχέσης είναι η χαρτογράφηση της χ^2 σε ένα βολικό διάστημα στο οποίο το αποτέλεσμα δεν είναι εξαρτώμενο της ποσότητας του δείγματος του dataset. Το V του Cramer είναι ένα τέτοιο μέτρο το οποίο ορίζεται ως:

$$V = \sqrt{\frac{x^2}{N \min\left(I-1, J-1\right)}}$$

Όπου I είναι ο αριθμός των χαρακτηριστικών και J ο ρυθμός των κλάσεων. Το V του Cramer έχει την ιδιότητα ότι λαμβάνει τιμές, αποκλειστικά, μεταξύ του μηδέν και του ένα. Το μηδέν αντιστοιχεί στην μη ύπαρξη συσχετισμού ενώ το ένα στην απόλυτη

συνάφεια. Ότι δηλαδή όλα τα χαρακτηριστικά, σε κάθε σειρά, αντιστοιχούν σε μια μοναδική στήλη και το αντίστροφο. Με άλλα λόγια αν το V του Cramer σε έναν πίνακα ενδεχομένων είναι ίσο με ένα, οι τιμές του χαρακτηριστικού που αντιπροσωπεύεται από αυτή τη σειρά καθορίζουν με τρόπο μοναδικό τις κλάσεις των χαρακτηριστικών.

Ανάλογα με τον αλγόριθμο υπάρχουν δύο τρόποι αναπαράστασης των αποτελεσμάτων. Κάποιοι από αυτούς κάνουν κατάταξη των χαρακτηριστικών σύμφωνα με την επίδοση που το καθένα πετυχαίνει σε μια καθορισμένη από τον αλγόριθμο μετρική. Τέτοιοι αλγόριθμοι είναι ο Chi-squared Attribute Evaluation, ο ReliefF, ο OneR, ο Info-Gain, ο Gain-Ratio. Άλλοι πάλι αλγόριθμοι, δεν βαθμολογούν μεμονωμένα χαρακτηριστικά αλλά υποομάδες χαρακτηριστικών και ως αποτέλεσμα δίνουν μια υποομάδα χαρακτηριστικών. Τέτοιοι αλγόριθμοι είναι ο Correlation-Based Feature Selection και ο Consistency-Based Subset Evaluation.

Αν και η προσέγγιση ενσωμάτωσης είναι πιο σύγχρονη και υπερισχύει στο θέμα της αύξησης της ευστοχίας πρόβλεψης, στις μέρες μας, η προσέγγιση φίλτρου χρησιμοποιείται ευρέως από την κοινότητα της εξόρυξης γνώσης κυρίως στην αντιμετώπιση τεράστιων βάσεων δεδομένων που η εφαρμογή της πρώτης είναι αδύνατη.

4 ΠΕΙΡΑΜΑΤΙΚΗ ΑΝΑΛΥΣΗ

4.1 Περιγραφή του Σετ Δεδομένων

4.1.1 Εισαγωγή

Στη παρούσα έρευνα θα χρησιμοποιηθεί το ADULT dataset το οποίο έχει εξαχθεί από τη βάση δεδομένων "Current Population Survey" (CPS) η οποία προσφέρεται από την Αμερικανική Υπηρεσία Απογραφής (U.S. Census Bureau). Η έρευνα CPS διεξάγεται για περισσότερα από 50 χρόνια και συλλέγει πληροφορίες για τα κοινωνικά δημογραφικά και οικονομικά χαρακτηριστικά του εργατικού δυναμικού –από 16 χρονών και πάνω- του πληθυσμού των Η.Π.Α.. Τα δεδομένα που συλλέγονται κάθε μήνα χρησιμοποιούνται για την σύνταξη αναφορών με θέματα όπως η απασχολησημότητα, η ανεργία, το βιοτικό επίπεδο, αλλά και για κοινωνικά δεδομένα όπως το ποσοστό καπνιστών ή το ποσοστό των ενεργών ψηφοφόρων. Τα αποτελέσματα που εξάγονται από τα δεδομένα του CPS χρησιμοποιούνται ως αξιόπιστοι δείκτες της κοινωνικής και οικονομικής κατάστασης του πληθυσμού τόσο από τον ιδιωτικό τομέα και από επιχειρήσεις που βολιδοσκοπούν τις ανάγκες της αγοράς, όσο και από την ίδια την κυβέρνηση και ιδιαίτερα από πολιτικούς αναλυτές ή νομοθέτες για τον σχεδιασμό και την αξιολόγηση κυβερνητικών προγραμμάτων. Τα δεδομένα του CPS είναι διαθέσιμα στο κοινό χωρίς χρέωση γεγονός που ενθάρρυνε την χρήση τους σε σημαντικές κοινωνικές και οικονομικές μελέτες.

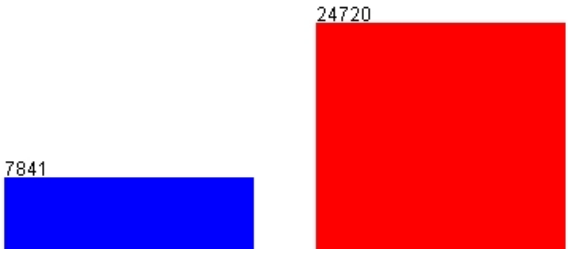
Το adult dataset που θα εξεταστεί στην εργασία αυτή περιλαμβάνει 32.561 στιγμιότυπα και αποτελείται από 14 χαρακτηριστικά, 6 αριθμητικά και 8 ονομαστικά. Το αντικείμενο του σετ είναι η ταξινόμηση των ερωτηθέντων σε εκείνους με ετήσιο εισόδημα μεγαλύτερο ή μικρότερο των 50.000 δολάρια. Στη συνέχεια παρατίθενται ένα προς ένα

τα χαρακτηριστικά του adult dataset συνοδευόμενα από ένα διάγραμμα, προερχόμενο από τη πλατφόρμα αλγορίθμων μηχανικής μάθησης WEKA, το οποίο δείχνει την κατανομή του κάθε χαρακτηριστικού στις πιθανές τιμές του και στο οποίο απεικονίζεται επίσης η κατανομή σε κλάσεις του κάθε χαρακτηριστικού.

4.1.2 Τα χαρακτηριστικά του ADULT dataset

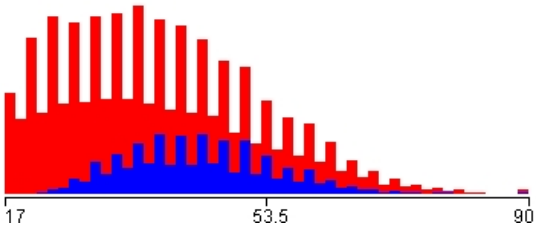
Το adult dataset περιλαμβάνει τα παρακάτω χαρακτηριστικά :

Class: Το χαρακτηριστικό κλάση του adult dataset. Αν δηλαδή ο ερωτούμενος έχει ετήσιο εισόδημα μεγαλύτερο από 50K$ (με μπλε χρώμα) ή μικρότερο (με κόκκινο χρώμα).



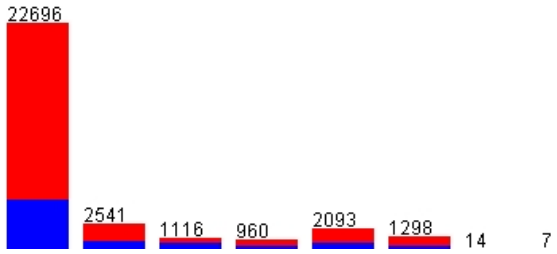
>50 $	7841
<=50 $	24720

Age: Η ηλικία των ερωτηθέντων (από 16 χρονών και πάνω).



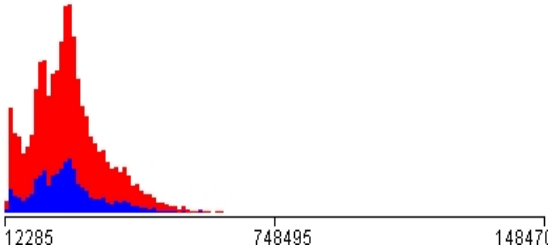
AGE	COUNT
Minimum	17
Maximum	90
Mean	38,58
StdDev	13,64

Workclass: Η κατηγορία επαγγέλματος των ερωτηθέντων (πιθανές απαντήσεις: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked). Παρακάτω δίνεται και σε μορφή πίνακα η κατανομή των αποτελεσμάτων. Στο χαρακτηριστικό αυτό υπάρχουν 1836 στιγμιότυπα με απολεσθέντες τιμές.



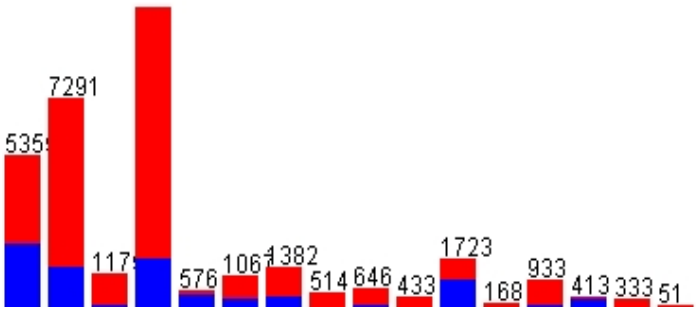
WORKCLASS	COUNT
PRIVATE	22696
SELF EMP NOT INC	2541
SELF EMP INC	1116
FEDERAL GOV	960
LOCAL GOV	2093
STATE GOV	1298
WITHOUT PAY	14
NEVER WORKED	7

Fnlwgt: Το κριτήριο Final weight είναι ένας συντελεστής βαρύτητας που εξαρτάται από τα δημογραφικά χαρακτηριστικά και ειδικότερα τον τόπο διαμονής των ερωτηθέντων που λαμβάνει συνεχείς τιμές από 12285 ως 1484705. Η κατανομή του φαίνεται στο παρακάτω διάγραμμα και τον πίνακα. Να σημειωθεί ότι στο χαρακτηριστικό αυτό 15.330 στιγμιότυπα λαμβάνουν μοναδικές τιμές .



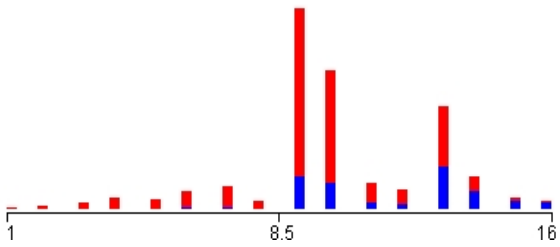
FNLWGT	
Minimum	12285
Maximum	1484705
mean	189778,4
StdDev	105550

Education: Το κριτήριο αυτό αποτυπώνει την μόρφωση των ερωτηθέντων με ονομαστική μορφή. (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool).



EDUCATION	COUNT
Bachelors	5355
Some-college	7291
11th	1175
HS-grad	10501
Prof-school	576
Assoc-acdm	1067
Assoc-voc	1382
9th	514
7th-8th	646
12th	433
Masters	1723
1st-4th	168
10th	933
Doctorate	413
5th-6th	333
Preschool	51

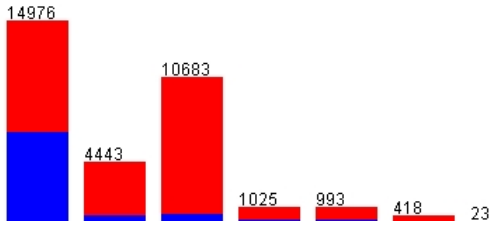
Education num: Μια μεταβλητή που δείχνει, όπως και η προηγούμενη, το μορφωτικό επίπεδο των ερωτηθέντων, αποτυπώνοντας το όμως εδώ με αριθμητική μορφή. Η κατάταξη είναι ως εξής:



EDUCATION	NUMBER	COUNT
Preschool	1	51
1st-4th	2	168
5th-6th	3	333
7th-8th	4	646
9th	5	514
10th	6	933
11th	7	1175
12th	8	433
HS-grad	9	10501
Some-college	10	7291
Assoc-voc	11	1382
Assoc-acdm	12	1067
Bachelors	13	5355
Masters	14	1723
Prof-school	15	576
Doctorate	16	413

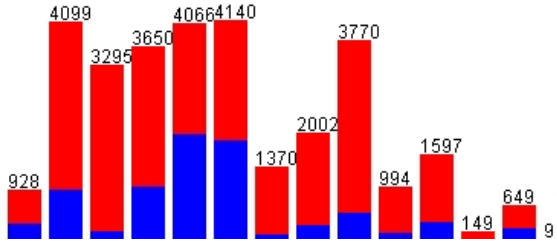
Marital status: Οικογενειακή κατάσταση ερωτηθέντων με πιθανές απαντήσεις:

Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse



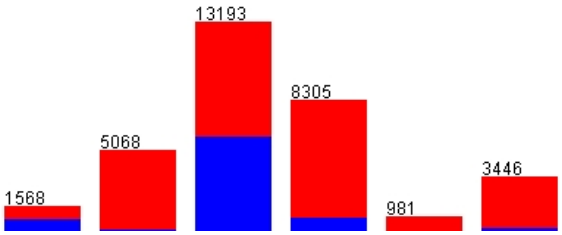
MARITAL STATUS	COUNT
Married-civ-spouse	14976
Divorced	4443
Never-married	10683
Separated	1025
Widowed	993
Married-spouse-absent	418
Married-AF-spouse	23

Occupation: Σε ποιο χώρο δραστηριοποιείται επαγγελματικά ο ερωτούμενος (πιθανές απαντήσεις: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces). Στο χαρακτηριστικό αυτό περιλαμβάνονται 1843 απολεσθέντες τιμές.



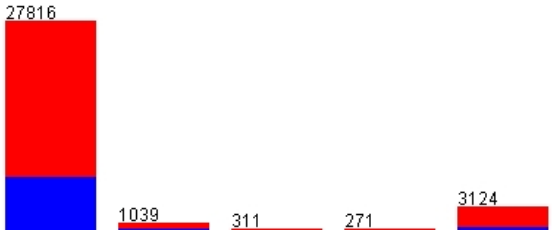
OCCUPATION	COUNT
Tech-support	928
Craft-repair	4099
Other-service	3295
Sales	3650
Exec-managerial	4066
Prof-specialty	4140
Handlers-cleaners	1370
Machine-op-inspect	2002
Adm-clerical	3770
Farming-fishing	994
Transport-moving	1597
Priv-house-serv	149
Protective-serv	649
Armed-forces	9

Relationship: ερώτηση που αφορά την σχέση του ερωτούμενου με τα μέλη της οικογένειας του. Πιθανές απαντήσεις: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.



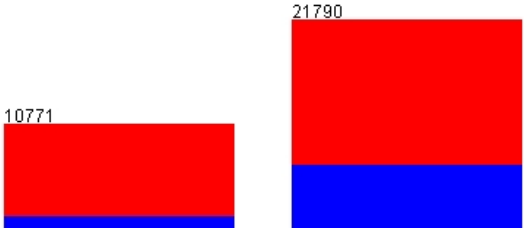
RELATIONSHIP	COUNT
Wife	1568
Own-child	5068
Husband	13193
Not-in-family	8305
Other-relative	981
Unmarried	3446

Race: Το χρώμα δέρματος του ερωτούμενου. Πιθανές απαντήσεις: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black



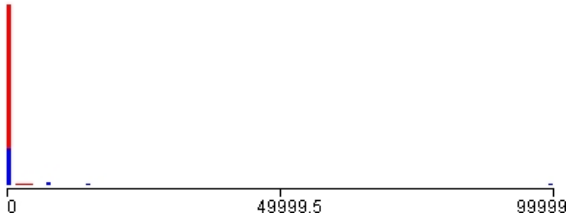
RACE	COUNT
White	27816
Asian-Pac-islander	1039
Amer-Indian-Eskimo	311
Other	271
Black	3124

Sex: Το φύλλο του ερωτούμενου.



SEX	COUNT
Female	10771
Male	21790

Capital-gain: Το χαρακτηριστικό αυτό δίνει τα κέρδη που πιθανόν να έχει ο ερωτούμενος από επενδυμένο κεφάλαιο.

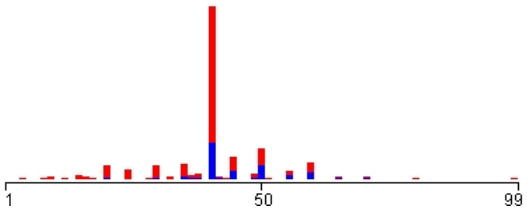


CAPITAL GAIN	COUNT
Minimum	0
Maximum	99999
Mean	1.077,650
StdDev	7385

Capital-loss: Το χαρακτηριστικό αυτό δίνει τη ζημιά που πιθανόν να έχει ο ερωτούμενος από επενδυμένο κεφάλαιο

CAPITAL LOSS	COUNT
Minimum	0
Maximum	4356
Mean	87,304
StdDev	402,96

Hours-per-week: Οι ώρες εργασίας ανά εβδομάδα για τον κάθε ερωτούμενο.



HOURS P/W	COUNT
Minimum	1
Maximum	99
Mean	40,437
StdDev	12,347

Native-country: Η χώρα καταγωγής του ερωτούμενου. Όπως είναι αναμενόμενο οι περισσότεροι από τους ερωτηθέντες είναι γηγενείς Αμερικάνοι και αυτό φαίνεται στο διάγραμμα. Άλλες πιθανές απαντήσεις: {United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand}. Στο χαρακτηριστικό αυτό περιλαμβάνονται 583 απολεσθέντες τιμές.

NATIVE COUNTRY	COUNT	Iran	43
Cambodia	19	Ireland	24
Canada	121	Italy	73
China	75	Jamaika	81
Columbia	59	Japan	62
Cuba	95	Laos	18
Dominican-Republic	70	Mexico	643
Ecuador	28	Nicaragua	34
El-Salvador	106	Outlying US(Guam-USVI-etc)	14
England	90	Peru	31
Fillipines	198	Poland	60
France	29	Portugal	37
Germany	137	Puerto Rico	114
Greece	29	Scotland	12
Guatemala	64	South	80
Haiti	44	Taiwan	51
Holand	1	Thailand	18
Honduras	13	Trinidad&Tobago	19
Hong Kong	20	US	29170
Hungary	13	Vietnam	67
India	100	Yugoslavia	16

4.2 Παρατηρήσεις Πάνω στο ADULT DATASET

Παρατηρώντας ένα προς ένα τα χαρακτηριστικά του εξεταζόμενου dataset, μπορεί κανείς να δει τα εξής:

Τα χαρακτηριστικά EDUCATION NUMBER και EDUCATION περιέχουν την ίδια πληροφορία με δύο διαφορετικές μορφές, ονομαστική και αριθμητική. Είναι δεδομένο ότι ο καθένας από τους αλγόριθμους μηχανικής μάθησης ευνοείται περισσότερο ή λιγότερο από την κάθε μορφή δεδομένων. Όμως, η χρήση και των δύο καθιστά το ένα εκ' των δύο πλεονάζων γεγονός το οποίο δυσχεραίνει, όπως έχει προαναφερθεί, την λειτουργία κάποιων αλγορίθμων. Παρουσιάζει ενδιαφέρον, λοιπόν, η παρατήρηση της λειτουργίας των χρησιμοποιούμενων αλγόριθμων με τη χρήση του καθενός εκ' των δύο χαρακτηριστικών σε σχέση με το αρχικό σετ δεδομένων.

Στα χαρακτηριστικά *Capital-Gain* και *Capital-Loss* παρατηρείται μεγάλη συγκέντρωση στην τιμή μηδέν. Αυτό συμβαίνει επειδή στην τιμή μηδέν προσμετρούνται και οι ερωτηθέντες που δεν έχουν επενδύσει το κεφάλαιο τους. Επίσης σημειώνεται, ότι οι ερωτηθέντες που έχουν μεγαλύτερη του μηδέν τιμή στο ένα χαρακτηριστικό έχουν μηδενική τιμή στο άλλο (πράγμα λογικό, μια και κάποιος ερωτηθέντας που έχει επενδυμένο κεφάλαιο θα έχει είτε κέρδος είτε ζημιά).

Τέλος, στο ονομαστικό χαρακτηριστικό *Native-country* παρατηρείται μεγάλη συγκέντρωση αποτελεσμάτων στην απάντηση United States, γεγονός απόλυτα λογικό με δεδομένο ότι η έρευνα γίνεται εντός των συνόρων των Ηνωμένων Πολιτειών. Επίσης αξιοσημείωτο είναι το μεγάλο εύρος πιθανών απαντήσεων στο συγκεκριμένο χαρακτηριστικό.

Στη διαδικασία Επιλογής Χαρακτηριστικών πέραν των αλγόριθμων η κρίση του ανθρώπινου παράγοντα είναι πολύ σημαντική, αν όχι η σημαντικότερη, για την επεξεργασία και την συμπλήρωση των πληροφοριών που θα προκύψουν.

4.3 Πειραματική Εφαρμογή Αλγορίθμων Μηχανικής Μάθησης

Στη συνέχεια θα εφαρμοστούν μια σειρά από αλγόριθμους Επιλογής Χαρακτηριστικών δανεισμένοι από την πλατφόρμα αλγορίθμων Μηχανικής Μάθησης WEKA η οποία έχει αναπτυχθεί από το Waikato University. Οι αλγόριθμοι αυτοί – των οποίων το θεωρητικό υπόβαθρο έχει αναλυθεί στο προηγούμενο κεφάλαιο- είναι γραμμένοι σε γλώσσα JAVA και προσφέρουν μια πολύ εύχρηστη συλλογή εργαλείων για την αντιμετώπιση απλούστερων ή πιο σύνθετων προβλήματα Data mining. Πρέπει να αναφερθεί ακόμα, ότι λόγω του μεγάλου μεγέθους του εξεταζόμενου dataset και των μεγάλων απαιτήσεων σε επεξεργαστική ισχύ που αυτό συνεπάγεται κάποιες εφαρμογές στάθηκε αδύνατον να εξεταστούν με τα διαθέσιμα τεχνικά μέσα.

Για την εξέταση των δεδομένων θα χρησιμοποιηθούν αρχικά μια σειρά αλγορίθμων διήθησης οι οποίοι ως αποτέλεσμα έχουν την κατάταξη των διαθεσίμων χαρακτηριστικών σύμφωνα με διάφορα κριτήρια. Στη συνέχεια, με βάση την απόδοση των χαρακτηριστικών στους παραπάνω αλγόριθμους ,θα γίνει χειροκίνητα η αφαίρεση και η τροποποίηση κάποιων χαρακτηριστικών, εφ' όσον αυτό κριθεί απαραίτητο, και στη συνέχεια θα γίνει δοκιμή και σύγκριση των υπολοίπων σε έξι διαφορετικούς αλγόριθμους ταξινόμησης. Κριτήριο στην αξιολόγηση της κάθε υποομάδας χαρακτηριστικών θα είναι η μεταβολή της τιμής των λάθος ταξινομημένων περιπτώσεων ενώ θα παρατηρείται παράλληλα και η μεταβολή στο μέγεθος του σχήματος που προκύπτει.

Μετά τους αλγόριθμους αξιολόγησης και κατάταξης μεμονωμένων χαρακτηριστικών θα εφαρμοστούν αλγόριθμοι εύρεσης υποομάδων και θα γίνει αξιολόγηση αυτών μέσα από τη δοκιμή στους ίδιους έξι αλγόριθμους κατάταξης.

Το σετ δεδομένων που χρησιμοποιείται στην εργασία αυτή χαρακτηρίζεται από το μεγάλο του μέγεθος, την ύπαρξη σε αυτό πολλών, διαφορετικών και όχι απαραίτητα χρήσιμων χαρακτηριστικών, την ύπαρξη θορύβου και ελλείπων τιμών. Θα έλεγε κανείς ότι αυτό το σετ δεδομένων είναι πολύ κοντά σε ένα πρόβλημα του πραγματικού κόσμου και για αυτόν τον λόγω παρουσιάζει ενδιαφέρον τόσο η σύγκριση της απόδοσης των σχημάτων μάθησης σε αυτό όσο και η διερεύνηση των περιθωρίων βελτίωσης του.

4.3.1 Εφαρμογή Αλγορίθμων Κατάταξης

Στο αρχικό σετ δεδομένων λοιπόν εφαρμόζονται οι παρακάτω αλγόριθμοι:

- Information Gain Attribute Ranking
- Gain Ratio Attribute Ranking
- OneR Information Attribute Ranking
- Chi-squared attribute selection
- ReliefF Attribute Ranking

Τα αποτελέσματα που προκύπτουν φαίνονται στον πίνακα 4.1. Στον πίνακα αυτό η μετρική του αλγορίθμου OneR είναι ποσοστό επί τοις εκατό σωστών ταξινομήσεων ενώ οι υπόλοιπες τέσσερεις αντιπροσωπεύουν την αριθμητική επίδοση του κάθε χαρακτηριστικού στην χαρακτηριστική εξίσωση του κάθε αλγόριθμου. Και στους τέσσερεις αλγόριθμους ισχύει ότι όσο μεγαλύτερη επίδοση πετυχαίνει ένα χαρακτηριστικό τόσο καλύτερος ταξινομητής είναι. Για να γίνει πιο εύκολη η συνολική παρατήρηση των επιδόσεων των χαρακτηριστικών, στον πίνακα 4.1 έχουν χρωματιστεί με πράσινο χρώμα οι τρεις καλύτερες επιδόσεις στον εκάστοτε αλγόριθμό και με κόκκινο χρώμα οι τρεις χειρότερες.

	ONER	INFO GAIN	GAIN RATIO	CHI-SQUARED	RERIEF-F
01. age:	75,92	0,0975	0,0297	3.437,8	0,0340
02. workclass:	76,31	0,0157	0,0111	798,9	0,0292
03. fnlwgt:	73,86	0,0000	0,0000	0,0	0,0070
04. education:	77,96	0,0936	0,0319	4.429,7	0,1049
05. education-num:	77,96	0,0933	0,0376	4.425,9	0,0260
06. marital-status:	75,92	0,1565	0,0854	6.517,7	0,0931
07. occupation:	75,92	0,0841	0,0248	3.614,1	0,1378
08. relationship:	75,92	0,1654	0,0768	6.699,1	0,1246
09. race:	75,92	0,0084	0,0105	330,9	0,0284
10. sex:	75,92	0,0372	0,0406	1.518,9	0,0199
11. capital-gain:	80,84	0,1145	0,1876	5.559,4	0,0160
12. capital-loss:	78,03	0,0507	0,1165	2.402,0	0,0087
13. hours-per-week:	75,90	0,0581	0,0268	2.541,7	0,0227
14. native-country:	75,92	0,0085	0,0102	311,3	0,0112

Πίνακας 4.1 Αποτελέσματα εφαρμογής αλγορίθμων κατάταξης

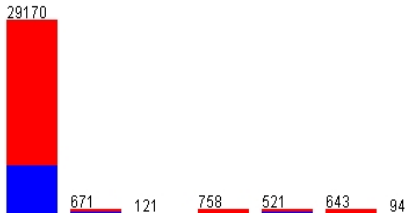
Παρατηρώντας τα πρώτα αυτά δεδομένα ο χρήστης μπορεί να έχει την πρώτη εμπεριστατωμένη εικόνα πάνω στα χαρακτηριστικά του εξεταζόμενου σετ δεδομένων. Στο σετ αυτό είναι φανερό ότι το χαρακτηριστικό-συντελεστής βαρύτητας FNLWGT όπως και το NATIVE COUNTRY έχουν πρόβλημα, συγκεντρώνουν τις χειρότερες αποδόσεις στους πέντε αλγορίθμους και είναι πιθανόν να παρεμποδίσουν στη συνέχεια τη λειτουργία των αλγορίθμων ταξινόμησης. Αντίθετα, χαρακτηριστικά όπως το CAPITAL GAIN και το MARITAL STATUS φαίνονται ότι θα παίξουν σημαντικό ρόλο στην πορεία.

Λαμβάνοντας υπόψη τα αποτελέσματα των αλγορίθμων κατάταξης θα διερευνηθούν οι εξής αλλαγές.

1) Κατάργηση του χαρακτηριστικού-συντελεστής βαρύτητας FNLWGT

2) Εναλλασσόμενη χρησιμοποίηση των χαρακτηριστικών EDUCATION και EDUCATION NUMBER

3) Δημιουργία ενός νέου χαρακτηριστικού με το όνομα GEOGRAPHIC AREA στη θέση του NATIVE COUNTRY. Το νέο αυτό χαρακτηριστικό περιλαμβάνει την ομαδοποίηση και τον περιορισμό των πιθανών απαντήσεων από 41 χώρες σε 7 γεωγραφικά διαμερίσματα. Πιο αναλυτικά, όπως φαινεται και στον πίνακα 4.2, το νέο χαρακτηριστικό περιλαμβάνει τις εξής πιθανές απαντήσεις: {US, ASIA, CANADA, C+S-AMER, EUROPE, MEXICO, OTHER}.



GEOGRAPHIC AREA	COUNT
US	29170
ASIA	671
CANADA	121
C+S AMER	758
EUROPE	521
MEXICO	643
OTHER	94

Πίνακας 4.2 Η κατανομή τιμών του νέου χαρακτηριστικού GEOGRAPHIC AREA

Η τελευταία αυτή αλλαγή στοχεύει αφ' ενός στην απλοποίηση των δεδομένων - πράγμα το οποίο σημαίνει λιγότερη απαιτούμενη υπολογιστική ισχύ, μικρότερους χρόνους εκμάθησης- και αφ' εταίρου στην πιθανή βελτίωση των αποτελεσμάτων

κατά την πρόβλεψη, μιας και ο μεγάλος αριθμός πιθανών τιμών που λαμβάνει αυτό το χαρακτηριστικό μπορεί να μπερδέψει κάποιους αλγόριθμους ή να υποβαθμίσει τη σημασία του σε κάποιους άλλου.

Οι αλγόριθμοι ταξινόμησης στους οποίους οι παραπάνω αλλαγές θα δοκιμαστούν είναι οι:

- J.48 (C4.5)
- NBTree
- PART
- RiDoR
- IBk
- Naïve Bayes

Των οποίων η λειτουργία και το θεωρητικό υπόβαθρο έχει αναλυθεί σε προηγούμενο κεφάλαιο.

Οι υποομάδες χαρακτηριστικών οι οποίες αρχικά θα εξεταστούν είναι οι εξής:

All attributes: Το σύνολο των χαρακτηριστικών του ADULT dataset

No fnlwgt: Εξάλειψη του χαρακτηριστικού FNLWGT

No Educ. Number: Εξάλειψη του χαρακτηριστικού EDUCATION NUMBER

No Educ: Εξάλειψη του χαρακτηριστικού EDUCATION

No Educ. Num, No fnlwgt: Εξάλειψη των χαρακτηριστικών EDUCATION NUMBER και FNLWGT

No Educ., No fnlwgt: Εξάλειψη των χαρακτηριστικών EDUCATION και FNLWGT

Για τον αλγόριθμο J.48 παρουσιάζονται στον πίνακα 4.3: το ποσοστό λάθος ταξινομήσεων για την κάθε ελεγχόμενη υποομάδα, το μέγεθος του δέντρου που προκύπτει και την επί τοις εκατό αύξηση ή μείωση της ευστοχίας ταξινόμησης της εξεταζόμενης υποομάδας σε σχέση με το σύνολο των χαρακτηριστικών:

J.48	incorrectly Classified Inst.	size	leaves	percentage
all atributes	13,7895	710	564	
no fnlwgt	13,7619	607	485	0,200%
no educ. Number	13,9369	480	356	-1,069%
no educ.	13,8816	475	332	-0,668%
no educ. num, no fnlwht	13,9338	387	292	-1,046%
no educ., no fnlwht	13,7465	437	314	0,312%

Πίνακας 4.3 Ποσοστό εσφαλμένων ταξινομήσεων, μέγεθος δέντρου και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον αλγόριθμο J.48

Τα αντίστοιχα δεδομένα για τον αλγόριθμο NBTree παρουσιάζονται στον πίνακα 4.4:

NBTree	incorrectly Classified Inst.	size	leaves	percentage
all atributes	13,9123	223	257	
no fnlwgt	13,9369	273	311	-0,177%
no educ. Number	13,7097	202	228	1,456%
no educ.	13,6697	206	243	1,744%
no educ. num, no fnlwht	13,7465	354	396	1,192%
no educ., no fnlwht	13,6789	120	148	1,678%

Πίνακας 4.4 Ποσοστό εσφαλμένων ταξινομήσεων, μέγεθος δέντρου και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον αλγόριθμο NBTree

Για τους δύο αλγόριθμους δέντρων απόφασης αν και το μέγεθος το οποίο εξετάζεται στα πλαίσια αυτής της εργασίας είναι η ευστοχία ταξινόμησης, παρατίθενται δίπλα από κάθε εξεταζόμενη υποομάδα και στοιχεία για το μέγεθος του δέντρου που προκύπτει. Άλλωστε η βελτίωση της ευκρίνειας των αποτελεσμάτων είναι από τα βασικά ζητούμενα στην εξόρυξη δεδομένων και η μείωση του μεγέθους του δέντρου που προκύπτει από την επιλογή χαρακτηριστικών συμβάλει προς αυτή την κατεύθυνση.

Για τον αλγόριθμο PART φαίνονται στον πίνακα 4.5 το ποσοστό λάθος ταξινομήσεων για την κάθε ελεγχόμενη υποομάδα, η επί τοις εκατό αύξηση ή μείωση της ευστοχίας ταξινόμησης και ο αριθμός των κανόνων που προκύπτουν για την καθεμία από αυτές:

PART	incorrectly Classified Inst.	rules	percentage
all atributes	14,9780	874	
no fnlwgt	14,3669	706	4,080%
no educ. Number	14,7078	907	1,804%
no educ.	14,6832	656	1,968%
no educ. num, no fnlwht	14,4314	773	3,649%
no educ., no fnlwht	14,3055	627	4,490%

Πίνακας 4.5 Ποσοστό εσφαλμένων ταξινομήσεων, αριθμός κανόνων και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον αλγόριθμο PART

Ομοίως για τον αλγόριθμο Ridor στον πίνακα 4.6:

Ridor	incorrectly Classified Inst.	rules	percentage
all atributes	16,8085	23	
no fnlwgt	16,4614	38	2,065%
no educ. Number	17,7359	26	-5,517%
no educ.	16,9897	23	-1,078%
no educ. num, no fnlwht	17,5486	29	-4,403%
no educ., no fnlwht	16,6549	22	0,914%

Πίνακας 4.6 Ποσοστό εσφαλμένων ταξινομήσεων, αριθμός κανόνων και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον αλγόριθμο Riddor

Για την χρήση του αλγόριθμου κοντινότερου γείτονα ή IBκ, σημαντικό ρόλο παίζει ο καθορισμός από τον χρήστη της παραμέτρου *κ*. Ως κ ορίζεται –όπως έχει αναφερθεί στο θεωρητικό μέρος- ο αριθμός των γειτονικών περιπτώσεων οι οποίες λαμβάνονται υπόψη κατά την ταξινόμηση του εκάστοτε δείγματος. Στα πλαίσια της εργασίας αυτής, επιλέχθηκε ο ορισμός *κ*=1. Η ενδεδειγμένη μέθοδος θα ήταν η πειραματική δοκιμή του αλγόριθμου IBκ για διάφορες τιμές του *κ* και η σύγκριση των αποτελεσμάτων με τη χρήση του cross-validation η οποία θα αναδείκνυε την καταλληλότερη τιμή. Ωστόσο η πολυπλοκότητα του αλγόριθμου -και κατά συνέπεια η υπολογιστική ισχύ που απαιτείται- αυξάνεται κατακόρυφα με την αύξηση της τιμής του κ και με δεδομένο το μεγάλο μέγεθος της εξεταζόμενης συλλογής δεδομένων καθίσταται απαγορευτική η χρήση μιας παραμέτρου *κ* μεγαλύτερης της μονάδας για έναν μέσο οικιακό υπολογιστή. Στον πίνακα 4.7 αποτυπώνονται τα αποτελέσματα από την εφαρμογή του αλγόριθμου IBκ στις εξεταζόμενες υποομάδες χαρακτηριστικών καθώς και η επί τοις εκατό βελτίωση/επιδείνωση του ποσοστού εσφαλμένων ταξινομήσεων σε σύγκριση με το αποτέλεσμα -στον ίδιο πάντα αλγόριθμο- της αρχικής ομάδας δεδομένων:

Ibk-1	incorrectly Classified Inst.	percentage
all atributes	20,5829	
no fnlwgt	20,0915	2,387%
no educ. Number	20,5706	0,060%
no educ.	20,3587	1,089%
no educ. num, no fnlwht	20,0792	2,447%
no educ., no fnlwht	19,9441	3,104%

Πίνακας 4.7 Ποσοστό εσφαλμένων ταξινομήσεων και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον αλγόριθμο Ibk-1

Τέλος, στον πίνακα 4.8 καταγράφονται τα αντίστοιχα αποτελέσματα για τον αλγόριθμο Naïve Bayes:

Naïve Bayes	incorrectly Classified Inst.	percentage
all atributes	16,572	
no fnlwgt	16,5996	-0,167%
no educ. Number	17,5793	-6,078%
no educ.	17,653	-6,523%
no educ. num, no fnlwht	17,567	-6,004%
no educ., no fnlwht	17,6407	-6,449%

Πίνακας 4.8 Ποσοστό εσφαλμένων ταξινομήσεων και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον αλγόριθμο PART

Το ποσοστό εσφαλμένων ταξινομήσεων για το αρχικό σετ δεδομένων και για την καθεμία από τις εξεταζόμενες υποομάδες όπως και η επί τοις εκατό μεταβολή του παρουσιάζονται συγκεντρωτικά στον πίνακα 4.9:

	all atributes	no fnlwgt	%	no educ. Number	%	no educ. num, no fnlwht	%	no education	%	no educ., no fnlwht	%
J.48	13,7895	13,7619	0,200%	13,9369	-1,069%	13,9338	-1,046%	13,8816	-0,668%	13,7465	0,312%
NBTree	13,9123	13,9369	-0,177%	13,7097	1,456%	13,7465	1,192%	13,6697	1,744%	13,6789	1,678%
PART	14,9780	14,3669	4,080%	14,7078	1,804%	14,4314	3,649%	14,6832	1,968%	14,3055	4,490%
Ridor	16,8085	16,4614	2,065%	17,7359	-5,517%	17,5486	-4,403%	16,9897	-1,078%	16,6549	0,914%
Naïve Bayes	16,572	16,5996	-0,167%	17,5793	-6,078%	17,567	-6,004%	17,653	-6,523%	17,6407	-6,449%
Ibk	20,5829	20,0915	2,387%	20,5706	0,060%	20,0792	2,447%	20,3587	1,089%	19,9441	3,104%

Πίνακας 4.9 Ποσοστό εσφαλμένων ταξινομήσεων και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον καθένα από τους αλγόριθμους

Παρατηρώντας τη διακύμανση του ποσοστού εσφαλμένων ταξινομήσεων μπορεί κανείς, κατ' αρχήν, να παρατηρήσει τη διαφορετική αντίδραση του καθενός από τους εξεταζόμενους αλγόριθμους στην εκάστοτε μεταβολή του σετ δεδομένων. Ο αλγόριθμος Naïve Bayes για παράδειγμα δουλεύει πιο αποτελεσματικά με το σύνολο των χαρακτηριστικών. Το γεγονός αυτό βέβαια, είναι αναμενόμενο με δεδομένη τη συμπεριφορά που χαρακτηρίζει τον αλγόριθμο αυτό στην αντιμετώπιση μη σχετικών χαρακτηριστικών. Ο J.48 βελτιώνει τα αποτελέσματα του με την εξάλειψη του χαρακτηριστικού FNLWGT και ωφελείται ακόμα περισσότερο με την ταυτόχρονη εξάλειψη του EDUCATION. Ο PART, όπως και οι RIDDOR και IBk, με την εξάλειψη του FNLWGT βελτιώνουν πολύ την απόδοσή τους (+4,08%, 2,06% και 2,39% αντίστοιχα) γεγονός που δείχνει πόσο ευαίσθητοι είναι σε μη σχετικά χαρακτηριστικά. Όσον αφορά τα πλεονάζοντα χαρακτηριστικά EDUCATION και EDUCATION NUMBER, περισσότερο δείχνουν να ωφελούνται οι αλγόριθμοι NBTree, PART και IBk και πιο πολύ ο δεύτερος.

Η επόμενη μετατροπή που θα εξεταστεί είναι η αντικατάσταση του χαρακτηριστικού NATIVE COUNTRY από το νέο χαρακτηριστικό GEOGRAPHIC AREA. Το νέο

χαρακτηριστικό έχει πολύ πιο περιορισμένο αριθμό πιθανών απαντήσεων από το παλιό και είναι ενδιαφέρον να εξεταστεί η αντίδραση του καθενός από τους αλγόριθμους σε αυτή την αλλαγή. Οι υποομάδες που εξετάστηκαν πριν, θα επανεξεταστούν με το νέο χαρακτηριστικό αντί του παλιού για να διαπιστωθεί κατά πόσο αυτό βελτιώνει το σχήμα μάθησης. Να σημειωθεί ότι το ποσοστό αύξησης-μείωσης για την κάθε υποομάδα που εξετάζεται από έναν αλγόριθμο, μετράει τη διαφοροποίηση από την αντίστοιχη υποομάδα με το παλιό χαρακτηριστικό και όχι σε σχέση με το σύνολο των χαρακτηριστικών όπως εξεταζόταν πριν. Παραδείγματος χάρην, η υποομάδα "no educ. geo" παρουσιάζει βελτίωση κατά 0,686% στην ευστοχία ταξινόμησης σε σχέση με την υποομάδα "no educ." κατά την εκτέλεση του αλγόριθμου J.48.

Για τον αλγόριθμο J.48 καταγράφονται στον πίνακα 4.10 το ποσοστό εσφαλμένων ταξινομήσεων που προκύπτουν από τις νέες υποομάδες που προέκυψαν από την μεταβολή του χαρακτηριστικού NATIVE COUNTRY. Καταγράφεται επίσης το μέγεθος του δέντρου που προκύπτει και η ποσοστιαία αύξηση ή ελάττωση του ποσοστού εσφαλμένων ταξινομήσεων:

J.48	incorrectly Classified Inst.	size	leaves	percentage
all atributes, geo	13,8417	691	533	-0,379%
no fnlwgt, geo	13,7312	627	486	0,223%
no educ. Number geo	14,0874	603	449	-1,080%
no educ. geo	13,7864	530	363	0,686%
no educ. num, no fnlwht, geo	14,0014	506	387	-0,485%
no educ., no fnlwht, geo	13,6605	475	332	0,626%

Πίνακας 4.10 Ποσοστό εσφαλμένων ταξινομήσεων, μέγεθος δέντρου και ποσοστό βελτίωσης της ταξινόμησης για τα εξεταζόμενα σετ δεδομένων με την αλλαγή του χαρακτηριστικού NATIVE COUNTRY, για τον αλγόριθμο J.48

Στον πίνακα 4.11 καταγράφονται τα αντίστοιχα δεδομένα που προέκυψαν από την εκτέλεση της ταξινόμησης μέσω του αλγόριθμου NBTree για τις νέες υποομάδες:

NBTree	incorrectly Classified Inst.	size	leaves	percentage
all atributes	13,8939	157	196	0,132%
no fnlwgt	13,8448	168	205	0,661%
no educ. Number	13,639	127	164	0,516%
no educ.	13,5745	140	193	0,696%
no educ. num, no fnlwht	13,7097	154	197	0,268%
no educ., no fnlwht	13,6052	91	124	0,539%

Πίνακας 4.11 Ποσοστό εσφαλμένων ταξινομήσεων, μέγεθος δέντρου και ποσοστό βελτίωσης της ταξινόμησης για τα εξεταζόμενα σετ δεδομένων με την αλλαγή του χαρακτηριστικού NATIVE COUNTRY, για τον αλγόριθμο NBTree

Για τον αλγόριθμο PART αντίστοιχα τα νέα αποτελέσματα παρουσιάζονται στον πίνακα 4.12:

PART	incorrectly Classified Inst.	rules	percentage
all atributes	14,4068	765	3,814%
no fnlwgt	14,2962	664	0,492%
no educ. Number	14,6586	800	0,335%
no educ.	14,5634	633	0,816%
no educ. num, no fnlwht	14,2072	721	1,554%
no educ., no fnlwht	14,3055	574	0,000%

Πίνακας 4.12 Ποσοστό εσφαλμένων ταξινομήσεων, αριθμός κανόνων και ποσοστό βελτίωσης της ταξινόμησης για τα εξεταζόμενα σετ δεδομένων με την αλλαγή του χαρακτηριστικού NATIVE COUNTRY, για τον αλγόριθμο PART

Για τον αλγόριθμο Ridor αντίστοιχα τα νέα αποτελέσματα παρουσιάζονται στον πίνακα 4.13:

Ridor	incorrectly Classified Inst.	rules	percentage
all atributes	17,2292	30	-2,503%
no fnlwgt	17,6991	120	-7,519%
no educ. Number	17,5762	36	0,900%
no educ.	17,3828	25	-2,314%
no educ. num, no fnlwht	17,3060	23	1,382%
no educ., no fnlwht	17,8864	121	-7,394%

Πίνακας 4.13 Ποσοστό εσφαλμένων ταξινομήσεων, αριθμός κανόνων και ποσοστό βελτίωσης της ταξινόμησης για τα εξεταζόμενα σετ δεδομένων με την αλλαγή του χαρακτηριστικού NATIVE COUNTRY, για τον αλγόριθμο Riddor

Για τον αλγόριθμο IBκ -με τιμή για την παράμετρο k ίση με 1- αντίστοιχα τα νέα αποτελέσματα παρουσιάζονται στον πίνακα 4.14:

Ibk-1	incorrectly Classified Inst.	percentage
all atributes	20,715	-0,642%
no fnlwgt	20,2727	-0,902%
no educ. Number	20,7119	-0,687%
no educ.	20,4969	-0,679%
no educ. num, no fnlwht	20,2819	-1,010%
no educ., no fnlwht	20,2174	-1,370%

Πίνακας 4.14 Ποσοστό εσφαλμένων ταξινομήσεων και ποσοστό βελτίωσης της ταξινόμησης για τα εξεταζόμενα σετ δεδομένων με την αλλαγή του χαρακτηριστικού NATIVE COUNTRY, για τον αλγόριθμο IBk-1

Τέλος για τον Αλγόριθμο Naïve Bayes, τα αποτελέσματα από την εφαρμογή ταξινόμησης μέσω αυτού στις νέες υποομάδες φαίνονται στον πίνακα 4.15:

Naïve Bayes	incorrectly Classified Inst.	percentage
all atributes	16,6088	-0,222%
no fnlwgt	16,6303	-0,185%
no educ. Number	17,5762	0,018%
no educ.	17,6039	0,278%
no educ. num, no fnlwht	17,567	0,000%
no educ., no fnlwht	17,6223	0,104%

Πίνακας 4.15 Ποσοστό εσφαλμένων ταξινομήσεων και ποσοστό βελτίωσης της ταξινόμησης για τα εξεταζόμενα σετ δεδομένων με την αλλαγή του χαρακτηριστικού NATIVE COUNTRY, για τον αλγόριθμο Naïve Bayes

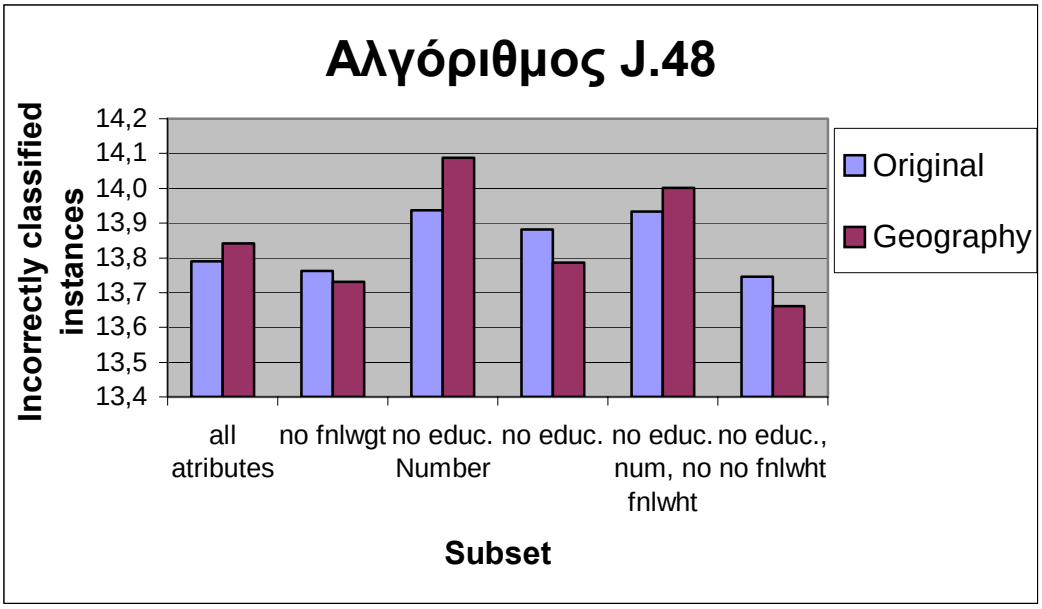
Στον πίνακα 4.16 φαίνεται η μεταβολή που προκαλείται στο ποσοστό εσφαλμένων ταξινομήσεων του κάθε αλγόριθμου από τη μεταβολή του χαρακτηριστικού NATIVE COUNTRY σε GEOGRAPHIC AREA στις εξεταζόμενες υποομάδες χαρακτηριστικών.

geography	all atributes	%	no fnlwgt	%	no educ. Number	%	no education	%	no educ. num, no fnlwht	%	no educ., no fnlwht	%
J.48	13,8417	-0,379%	13,7312	0,223%	14,0874	-1,080%	13,7864	0,686%	14,0014	-0,485%	13,6605	0,626%
NBTree	13,8939	0,132%	13,8448	0,661%	13,639	0,516%	13,5745	0,696%	13,7097	0,268%	13,6052	0,539%
PART	14,4068	3,814%	14,2962	0,492%	14,6586	0,335%	14,5634	0,816%	14,2072	1,554%	14,3055	0,000%
Ridor	17,2292	-2,503%	17,6991	-7,519%	17,5762	0,900%	17,3828	-2,314%	17,3060	1,382%	17,8864	-7,394%
Naïve Bayes	16,6088	-0,222%	16,6303	-0,185%	17,5762	0,018%	17,6039	0,278%	17,567	0,000%	17,6223	0,104%
Ibk	20,715	-0,642%	20,2727	-0,902%	20,7119	-0,687%	20,4969	-0,679%	20,2819	-1,010%	20,2174	-1,370%

Πίνακας 4.16 Ποσοστό εσφαλμένων ταξινομήσεων και ποσοστό βελτίωσης της ταξινόμησης για καθένα από τα εξεταζόμενα σετ δεδομένων με την αλλαγή του χαρακτηριστικού NATIVE COUNTRY, για τον καθένα από τους αλγόριθμους

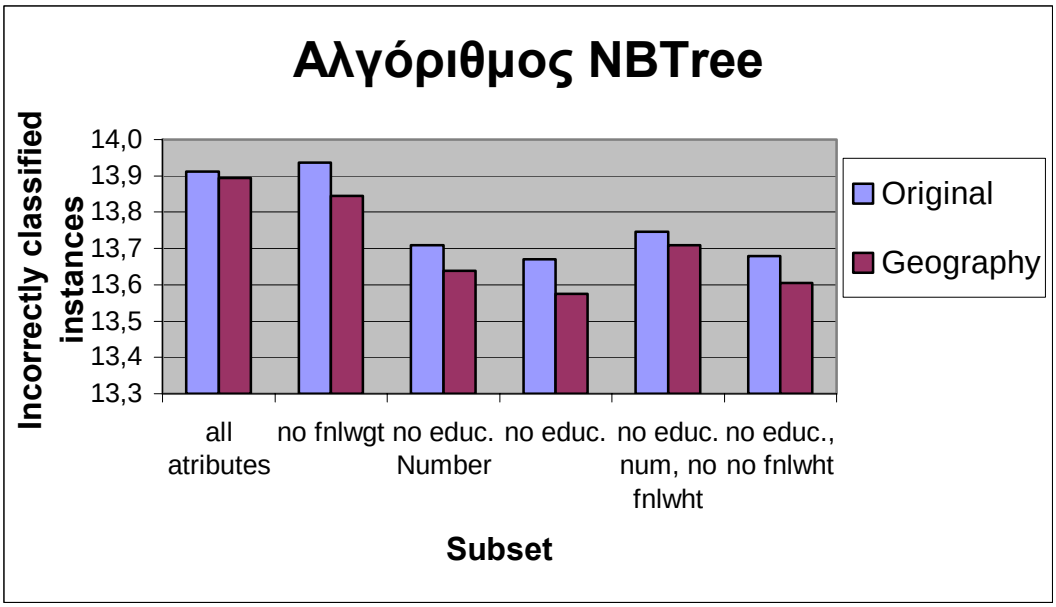
Με βάση τα αποτελέσματα των πινάκων 4.9 και 4.16 κατασκευάζονται μια σειρά από διαγράμματα που αναπαριστούν γραφικά την επίδοση όλων των διερευνηθέντων υποομάδων για καθέναν από τους αλγόριθμους. Με το τρόπο αυτό καθίσταται ευκολότερη η παρατήρηση των αποτελεσμάτων και η εξαγωγή συμπερασμάτων

Στο διάγραμμα 4.1 φαίνεται η διακύμανση της ευστοχίας ταξινόμησης στον αλγόριθμο J.48 για όλες τις εξεταζόμενες υποομάδες. Παρατηρείται λοιπόν, ότι με την προσθήκη του χαρακτηριστικού GEOGRAPHIC AREA αντί του NATIVE COUNTRY η ευστοχία ταξινόμησης αυξάνεται επιπλέον σε συνδυασμό με την παράλειψη των χαρακτηριστικών EDUCATION και FNLWGT.



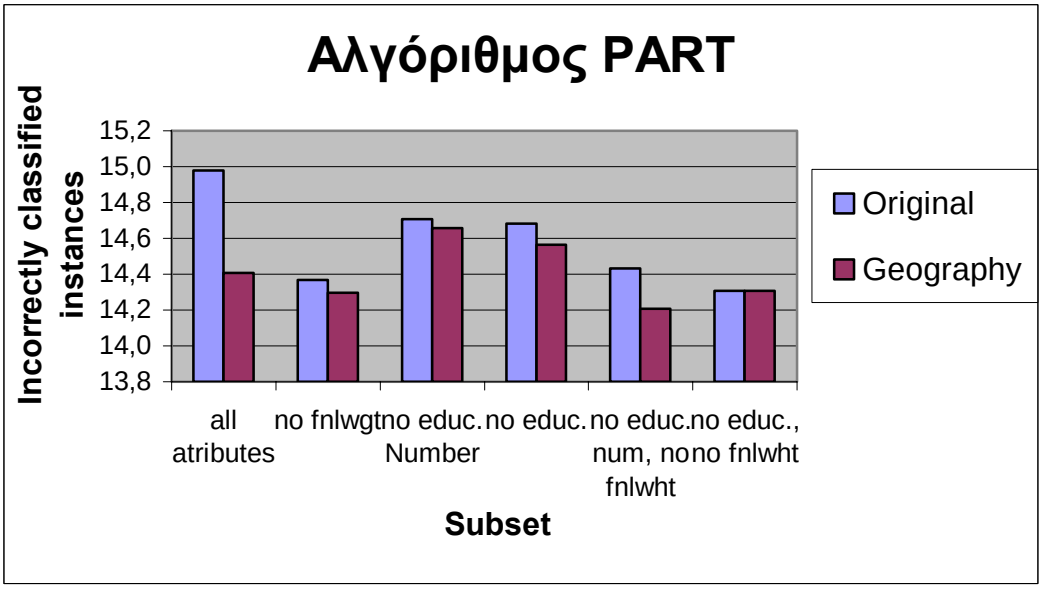
διάγραμμα 4.1 Ποσοστό λάθος ταξινομήσεων για τον αλγόριθμο J.48

Ο υβριδικός αλγόριθμος NBTree απ' την άλλη, λόγω της σχέσης του με τον αλγόριθμο Naïve Bayes, είναι πιο ανθεκτικός στην ύπαρξη μη σχετικών δεδομένων. Φαίνεται από το διάγραμμα 4.2 ότι η εξάλειψη του χαρακτηριστικού FNLWGT αυξάνει το σφάλμα ταξινόμησης. Από την άλλη η απομάκρυνση πλεοναζόντων χαρακτηριστικών (EDUCATION NUMBER ή EDUCATION) φαίνεται ότι αυξάνει κατά πολύ την απόδοση του σχήματος. Τέλος, η αντικατάσταση του NATIVE COUNTRY από το GEOGRAPHIC AREA επιδρά και αυτή θετικά στον αλγόριθμο αυξάνοντας ακόμα περισσότερο το ποσοστό επιτυχημένων ταξινομήσεων.



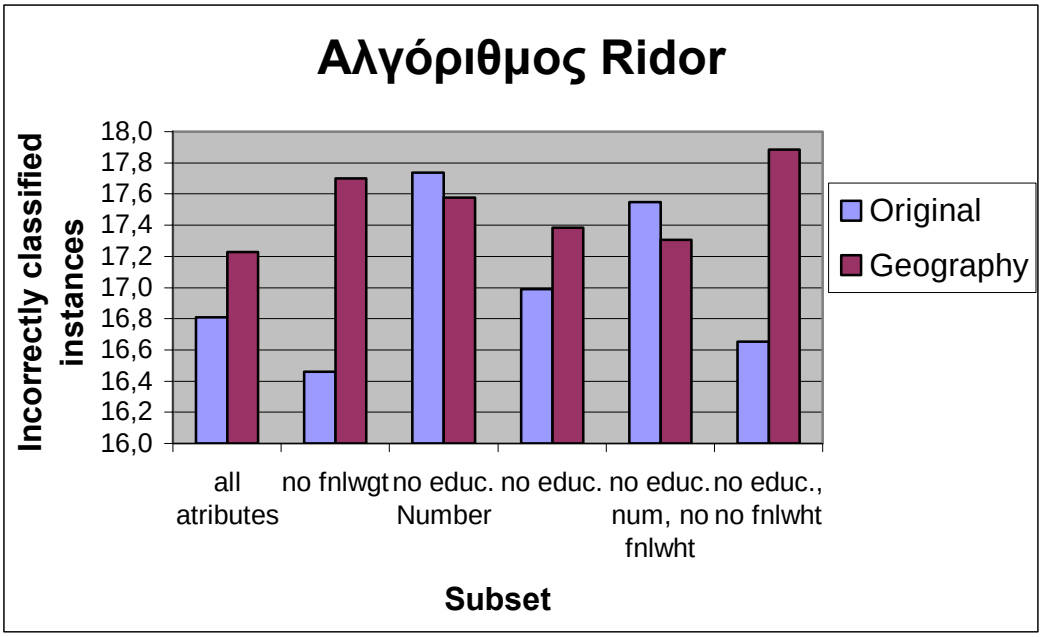
διάγραμμα 4.2 Ποσοστό λάθος ταξινομήσεων για τον αλγόριθμο NBTree

Στον αλγόριθμο PART, το ποσοστό των σωστά ταξινομημένων περιπτώσεων αυξάνεται με τις μετατροπές στα χαρακτηριστικά του dataset. Η μεγαλύτερη αύξηση, σε ποσοστό 4,08%, επιτυγχάνεται με την εξάλειψη του χαρακτηριστικού FNLWGT. Η μετατροπή του GEOGRAPHIC AREA αυξάνει επίσης την ικανότητα κατάταξης του σχήματος, όπως και η εξάλειψη ενός εκ' των EDUCATION NUMBER ή EDUCATION.



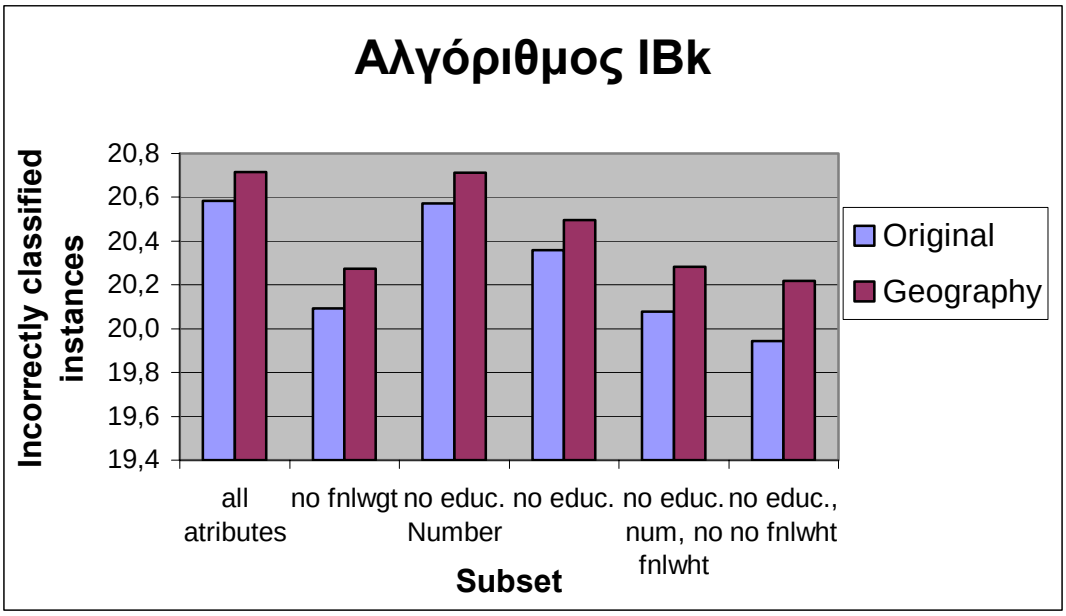
διάγραμμα 4.3 Ποσοστό λάθος ταξινομήσεων για τον αλγόριθμο NBTree

Στον αλγόριθμο Ridor, παρατηρείται ότι η αφαίρεση του FNLWGT επιφέρει αύξηση του ποσοστού εύστοχων ταξινομήσεων. Οι υπόλοιπες αλλαγές λειτουργούν αρνητικά στο συγκεκριμένο αλγόριθμο. Η συμπεριφορά του συγκεκριμένου αλγόριθμου δικαιολογείται λόγω της δομής "if-then" των κανόνων που τον χαρακτηρίζει και η οποία ευνοεί την ύπαρξη πλεοναζόντων χαρακτηριστικών ή χαρακτηριστικών με πολλές πιθανές απαντήσεις όπως το NATIVE COUNTRY.



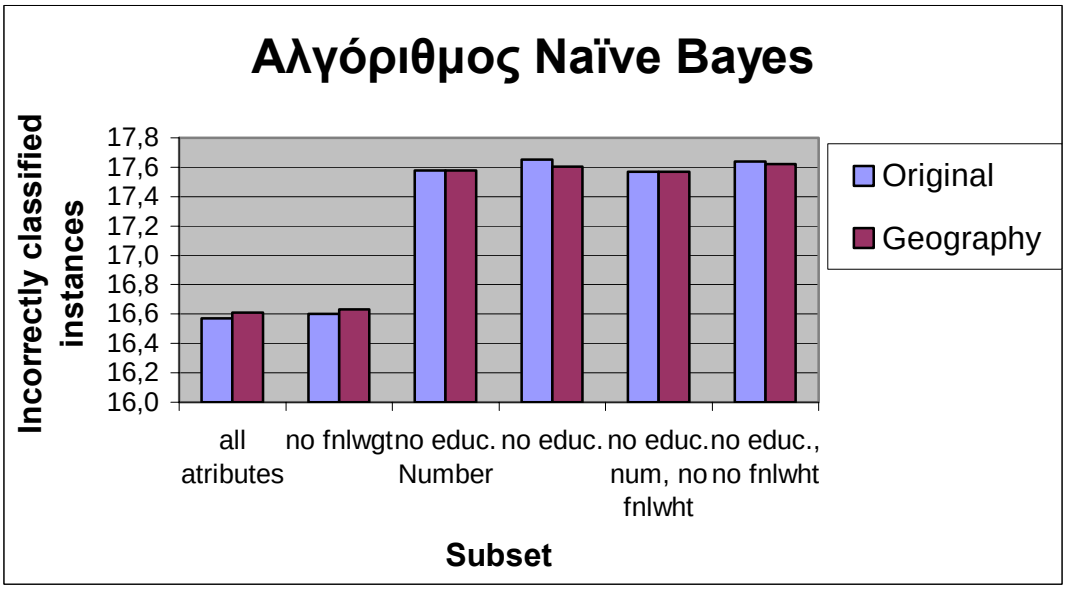
διάγραμμα 4.4 Ποσοστό λάθος ταξινομήσεων για τον αλγόριθμο Ridor

Στον αλγόριθμο "Κοντινότερου Γείτονα" η αφαίρεση ενός μη σχετικού χαρακτηριστικού όπως το FNLWGT αναμενόμενο να οδηγήσει σε αύξηση της διαχωριστικής δύναμης του αλγορίθμου. Επίσης αναμενόμενη ήταν η επιλογή ενός χαρακτηριστικού με αριθμητικές τιμές όπως το EDUCATION NUMBER έναντι ενός ονομαστικού όπως το EDUCATION. Διαπιστώνεται επίσης ότι αντίθετα με άλλους αλγόριθμους ο IBk προτιμά το χαρακτηριστικό NATIVE COUNTRY από το GEOGRAPHIC-AREA.



διάγραμμα 4.5 Ποσοστό λάθος ταξινομήσεων για τον αλγόριθμο IBk

Παρατηρεί κανείς ότι ο αλγόριθμος Naïve Bayes λειτουργεί καλύτερα χρησιμοποιόντας το σύνολο των χαρακτηριστικών. Η εξάλειψη ενός μη σχετικού χαρακτηριστικού, όπως το FNLWGT, αυξάνει την τιμή του ποσοστού λάθους ταξινομήσεων σε μικρό ποσοστό, κοντά 0,1%.



διάγραμμα 4.6 Ποσοστό λάθος ταξινομήσεων για τον αλγόριθμο Naïve Bayes

Η απομάκρυνση ενός χρήσιμου για τη λειτουργία του αλγορίθμου, έστω και πλεονάζοντος, χαρακτηριστικού εκτινάσσει το ποσοστό λάθος ταξινομήσεων σε αύξηση της τάξης του 6%. Η συμπεριφορά αυτή του ταξινομητή μπορεί να εξηγηθεί

από το γεγονός ότι η λειτουργία του στηρίζεται στην ανεξάρτητη αποτίμηση χαρακτηριστικών και ευνοεί την ύπαρξη πολλών, έστω και λιγότερων σχετικών χαρακτηριστικών. Παρατηρείται επίσης ότι η μετατροπή του χαρακτηριστικού NATIVE COUNTRY σε GEOGRAPHIC AREA δεν άλλαξε σχεδόν καθόλου την τα ποσοστά ευστοχίας του σχήματος. Ένα ακόμα γεγονός που δείχνει ότι στον αλγόριθμο Naïve Bayes οι αλλαγές σε χαρακτηριστικά ελάσσονος σημασίας επιφέρουν πολύ μικρές αλλαγές στην ικανότητα ταξινόμηση του σχήματος.

4.3.2 Εφαρμογή Αλγορίθμων Αξιολόγησης Υποομάδων

Οι υπόλοιποι αλγόριθμοι που θα εφαρμοστούν είναι αλγόριθμοι αναζήτησης και αξιολόγησης υποομάδων και όχι χαρακτηριστικών. Οι αλγόριθμοι που θα χρησιμοποιηθούν είναι οι:

- Wrapper Subset Evaluation
- Classifier Subset Evaluation
- Correlation-based Feature Selection
- Consistency-Based Subset Evaluation

Από αυτούς τους αλγόριθμους οι δύο πρώτοι ανήκουν στην κατηγορία ενσωμάτωσης και οι δύο τελευταίοι στην κατηγορία διήθησης.

Ο καθένας από τους αλγόριθμους αυτούς τροφοδοτείται με το σύνολο των χαρακτηριστικών του ADULT dataset και επιλέγει μια υποομάδα χαρακτηριστικών. Αυτή η υποομάδα, στη συνέχεια, ελέγχεται με τη χρήση των ίδιων αλγορίθμων κατάταξης που χρησιμοποιήθηκαν και πριν.

Στην εφαρμογή του αλγόριθμου Wrapper χρησιμοποιείται ο αλγόριθμος J.48 για την αξιολόγηση των ελεγχόμενων υποομάδων.

Στον αλγόριθμο Classifier Subset Evaluation (CSE), οι αλγόριθμοι ταξινόμησης που χρησιμοποιούνται είναι ο J.48, ο PART, ο Naïve Bayes και ο NBTree. Η μέθοδος αναζήτησης υποομάδας που χρησιμοποιήθηκε για την διερεύνηση του "χώρου των υποομάδων χαρακτηριστικών" είναι ο αλγόριθμος Bestfirst με αναζήτηση "προς τα

εμπρός", ξεκινώντας δηλαδή από το κενό υποσύνολο και προσθαφαιρώντας σταδιακά χαρακτηριστικά.

Στον παρακάτω πίνακα φαίνονται οι υποομάδες των χαρακτηριστικών που επελέγησαν από τους τέσσερεις αλγόριθμους.

	Cfs Sub. Eval.	Consistency Sub. Eval.	CSE j.48	CSE PART	CSE NBTree	CSE NB	Wrapper
age:		1	1	1	1	1	
workclass:		1	1	1	1	1	
fnlwgt:			1	1	1		
education:		1		1	1	1	
education-num:	1		1	1		1	1
marital-status:	1	1	1	1		1	1
occupation:		1	1	1	1	1	
relationship:	1	1	1		1	1	
race:		1		1	1	1	1
sex:		1	1	1	1	1	1
capital-gain:	1	1	1	1	1	1	1
capital-loss:	1	1	1	1	1		1
hours-per-week:		1	1	1		1	1
native-country:		1			1	1	

πίνακας 4.17 Πίνακας αλγορίθμων αξιολόγησης υποομάδας

Παρατηρώντας κανείς τις επιλεγμένες υποομάδες των χαρακτηριστικών γίνεται φανερός και στην πράξη ο διαφορετικός τρόπος που λειτουργούν οι αλγόριθμοι αυτοί. Ο "Classifier Subset Evaluation" παράγει μεγάλες σε μέγεθος υποομάδες, έντεκα ή δώδεκα χαρακτηριστικών, αποκλείοντας όμως σε κάθε περίπτωση, ανάλογα με τον χρησιμοποιούμενο αλγόριθμο, διαφορετικά χαρακτηριστικά. Ο αλγόριθμος "Wrapper" απ' την άλλη αν και αναζητά την βέλτιστη υποομάδα μέσω του ίδιου αλγόριθμου με τον "Classifier Subset Evaluation" (μέσω του J.48) ωστόσο παράγει μια εντελώς διαφορετική υποομάδα.

Ένα ακόμα γεγονός που μπορεί κανείς να παρατηρήσει είναι ότι τα χαρακτηριστικά που οι αλγόριθμοι διήθησης ανέδειξαν ως λιγότερο συναφή όπως το FNLWGT, το NATIVE COUNTRY και το RACE περιλαμβάνονται σε πολλά από τα επιλεγμένα subset.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα από την δοκιμή τον επιλεχθέντων υποομάδων και παρατίθενται δίπλα στο ποσοστό λανθασμένων ταξινομήσεων που προκύπτει από την χρήση όλων των χαρακτηριστικών του dataset.

	all atributes	Wrapper j.48	CSE j.48	CSE PART	CSE NBTree	CSE NB	Cfs Sub. Eval.	Consistency Sub. Eval.
J.48	13,7895	14,0874	13,8448	13,7772	14,5419	14,3177	20,0639	13,9338
NBTree	13,9123	14,5911	13,6329	15,3312	13,4118	14,6494	14,2471	13,7465
PART	14,9780	14,3822	14,3331	14,5327	14,4559	15,0947	14,3853	14,4314
Ridor	16,8085	16,8545	17,0173	17,2231	17,5732	17,0480	15,1347	17,5486
Naïve Bayes	16,5720	20,2113	17,7421	17,7267	19,2746	15,9117	14,3300	17,567
Ibk	20,5829	16,1881	20,2819	20,5676	21,0006	20,5676	14,4621	20,0792

Πίνακας 4.18 Ποσοστό εσφαλμένων ταξινομήσεων για καθένα από τα εξεταζόμενα σετ δεδομένων, για τον καθένα από τους αλγόριθμους

Αξιολογώντας τα αποτελέσματα της δοκιμής αυτής και συγκρίνοντας τα με το ποσοστό εσφαλμένων ταξινομήσεων του dataset με όλα τα χαρακτηριστικά, παρατηρεί κανείς τα εξής:

Οι αλγόριθμοι ενσωμάτωσης που χρησιμοποιούν τον J.48 ως κριτήριο επιλογής υποομάδων, θα περίμενε κανείς να έχουν υψηλότερα ποσοστά ευστοχίας σε αυτόν. Κι όμως συμβαίνει το αντίθετο, οι ομάδες που επιλέγηκαν από τον Wrapper και τον CSE έχουν αυξημένο ποσοστό λάθος ταξινομήσεων σε σχέση με το αρχικό σετ. Ωστόσο, ελέγχοντας τα σετ που προκύψαν από αυτούς τους αλγόριθμους ενσωμάτωσης με τους υπόλοιπους αλγόριθμους της δοκιμής παρατηρείται βελτίωση στα αποτελέσματα τόσο του NBTree όσο και του PART. Το γεγονός αυτό δεν αποτελεί έκπληξη με δεδομένη τη σχέση που υπάρχει ανάμεσα στον J.48 και τους δύο αυτούς αλγόριθμους. Ο IBk επίσης βελτιώνει αρκετά το αποτέλεσμα του όταν χρησιμοποιεί την υποομάδα δεδομένων του Wrapper. Αυτό, εξηγείται από το γεγονός ότι ο αλγόριθμος αυτός λειτουργεί πιο αποτελεσματικά όταν εξετάζει λίγα σε αριθμό αλλά με μεγαλύτερο βαθμό σχετικότητας χαρακτηριστικά.

Για την υποομάδα που επιλέχθηκε από τον αλγόριθμο CSE με χρήση του PART ως αλγόριθμου επιλογής, παρατηρείται ότι βελτιώνει την ικανότητα ταξινόμησης του ίδιου του PART καθώς και του J.48. Στους υπόλοιπους αλγόριθμους είτε ελαττώνει, είτε αφήνει στο ίδιο επίπεδο το ποσοστό εύστοχων ταξινομήσεων.

Εφαρμόζοντας, στη συνέχεια, τον CSE με χρήση του αλγόριθμου NBTree και ελέγχοντας την λειτουργία του ίδιου του αλγόριθμου, παρατηρεί κανείς την βελτίωση της λειτουργίας του συγκεκριμένου ταξινομητή. Η ίδια υποομάδα χαρακτηριστικών δοκιμαζόμενη στους υπόλοιπους αλγόριθμους δεν βελτιώνει τη λειτουργία τους, με εξαίρεση τον αλγόριθμό PART.

Ο συνδυασμός του CSE με τον αλγόριθμο Naïve Bayes (NB) είναι αρκετά διαδεδομένος. Οι αλγόριθμοι ενσωμάτωσης που χρησιμοποιούν πιθανοτικούς ταξινομητές απαιτούν μικρότερη υπολογιστική ισχύ, είναι πιο γρήγοροι και συνεπώς μπορούν επεξεργαστούν πολύ μεγαλύτερα και πολύπλοκα σετ δεδομένων απ' ότι άλλοι επαγωγικοί αλγόριθμοι. Στη συγκεκριμένη περίπτωση, το σετ δεδομένων που προκύπτει βελτιώνει αρκετά τη λειτουργία του ίδιου του NB ωστόσο δεν βελτιώνει σημαντικά την λειτουργία κανενός από τους υπόλοιπους αλγόριθμους.

Από τις υποομάδες που επιλέγηκαν από τους αλγόριθμους διήθησης, αυτή του αλγόριθμου Consistency Subset Evaluation ευνοεί τους αλγόριθμους PART και NBTree, όμως εκείνη η οποία προκαλεί εντύπωση είναι αυτή που έχει επιλεγεί από τον αλγόριθμο Correlation-based Feature Selection (Cfs). Είναι προφανές ότι οι αλγόριθμοι Naïve Bayes, IBk και Riddor ευνοούνται από τη χρήση μιας μικρότερης σε μέγεθος ομάδας χαρακτηριστικών και η απόδοση τους με αυτή αυξάνεται κατακόρυφα, του NB κατά 13,5%, του IBκ κατά 30% και του Riddor κατά 10%.

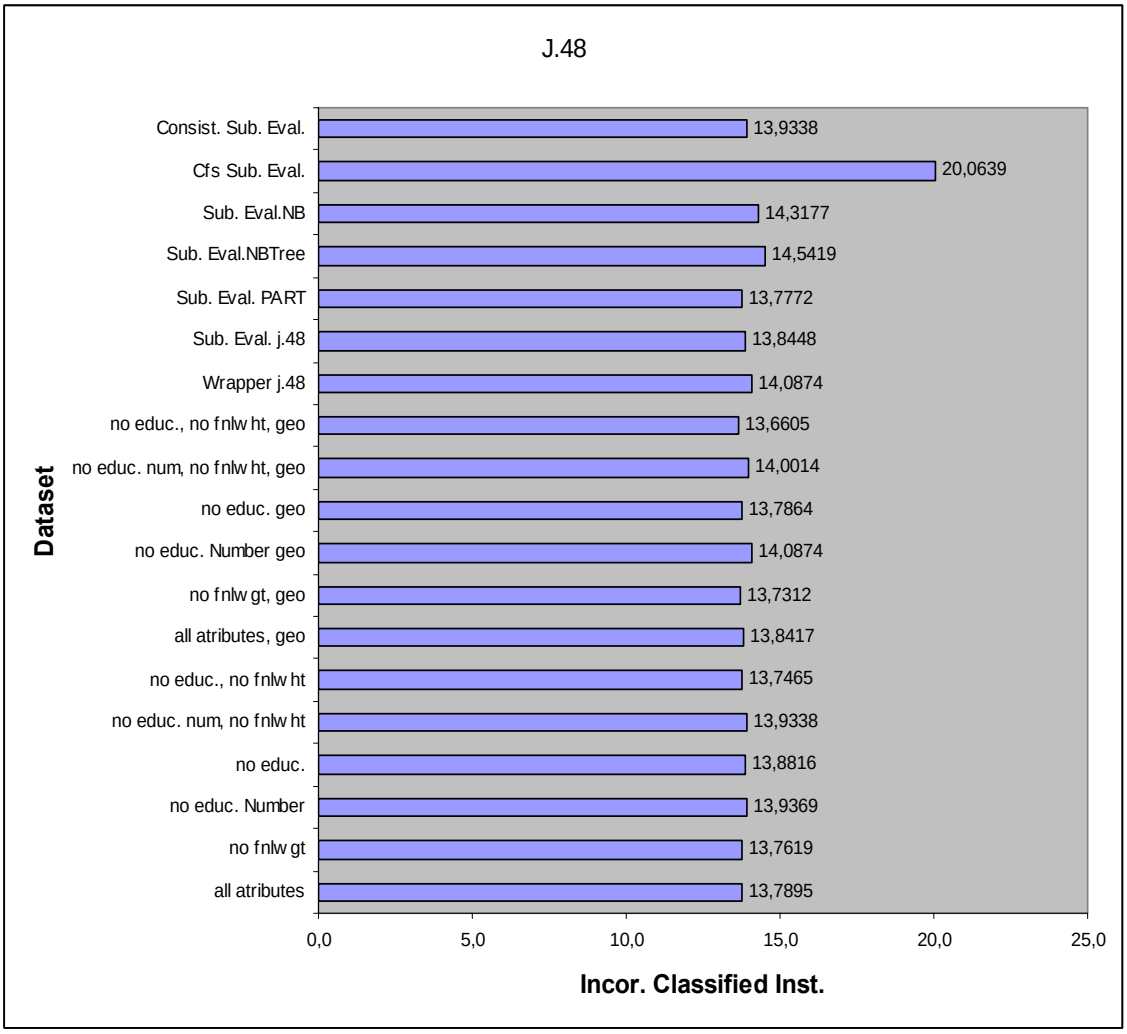
4.4 Συνοπτικά αποτελέσματα

Κατά τη διάρκεια αυτής της εργασίας εξετάστηκαν μια σειρά από μεθόδους επιλογής χαρακτηριστικών, επιλέχτηκαν με τη βοήθεια των μεθόδων αυτών κάποιες υποομάδες, οι οποίες στη συνέχεια τροφοδοτούν τους επιλεγμένους αλγόριθμους ταξινόμησης. Συνολικά δοκιμάστηκαν δεκαοκτώ υποομάδες χαρακτηριστικών από έξι αλγόριθμους ταξινόμησης διαφορετικής φιλοσοφίας. Το συμπέρασμα στο οποίο καταλήγουμε είναι το αναμενόμενο. Δεν υπάρχει υποομάδα χαρακτηριστικών η οποία να μεγιστοποιεί το ποσοστό ευστοχίας ταξινόμησης για το σύνολο των αλγορίθμων ταξινόμησης. Ο κάθε αλγόριθμος έχει συγκεκριμένα πλεονεκτήματα και μειονεκτήματα και το κλασσικό ερώτημα "Ποιος ταξινομητής είναι ο κατάλληλος για το εξεταζόμενο σετ δεδομένων;" ίσως θα έπρεπε να επαναδιατυπωθεί και να

αναρωτιόμαστε: "Ποιες αλλαγές πρέπει να γίνουν στο εξεταζόμενο σετ δεδομένων ώστε να λειτουργήσει καλύτερα ο δεδομένος ταξινομητής;"

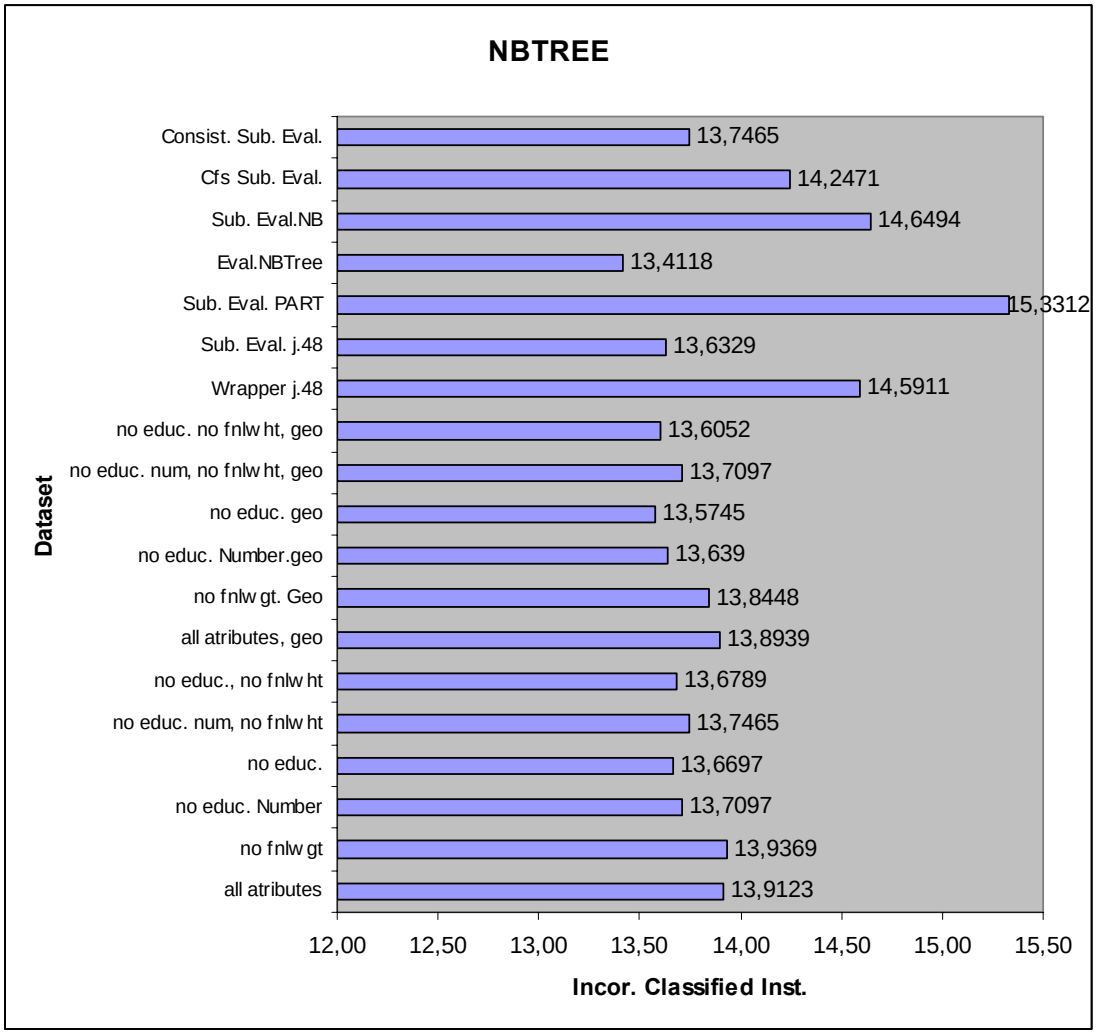
Γίνεται στη συνέχεια με την βοήθεια πινάκων μια σύνοψη στα αποτελέσματα για τον καθένα από τους έξι ταξινομητές.

Ο J.48 δείχνει αρκετά σταθερός και στις περισσότερες από τις εξεταζόμενες υποομάδες παρουσιάζει μικρή διακύμανση στην ευστοχία με το ποσοστό εσφαλμένων ταξινομήσεων να κινείται στις περισσότερες περιπτώσεις μεταξύ 13,7% και 14% τα καλύτερα αποτελέσματα επετεύχθησαν με την υποομάδα χωρίς τα χαρακτηριστικά: EDUCATION, FNLWGT και με το χαρακτηριστικό GEOGRAPHIC AREA αντί του NATIVE COUNTRY (*no educ, no fnlwht, geo*), όπου επιτεύχθηκε ποσοστό εύστοχης ταξινόμησης 86,44% (ή αλλιώς, ποσοστό λάθος ταξινομήσεων 13,66%).



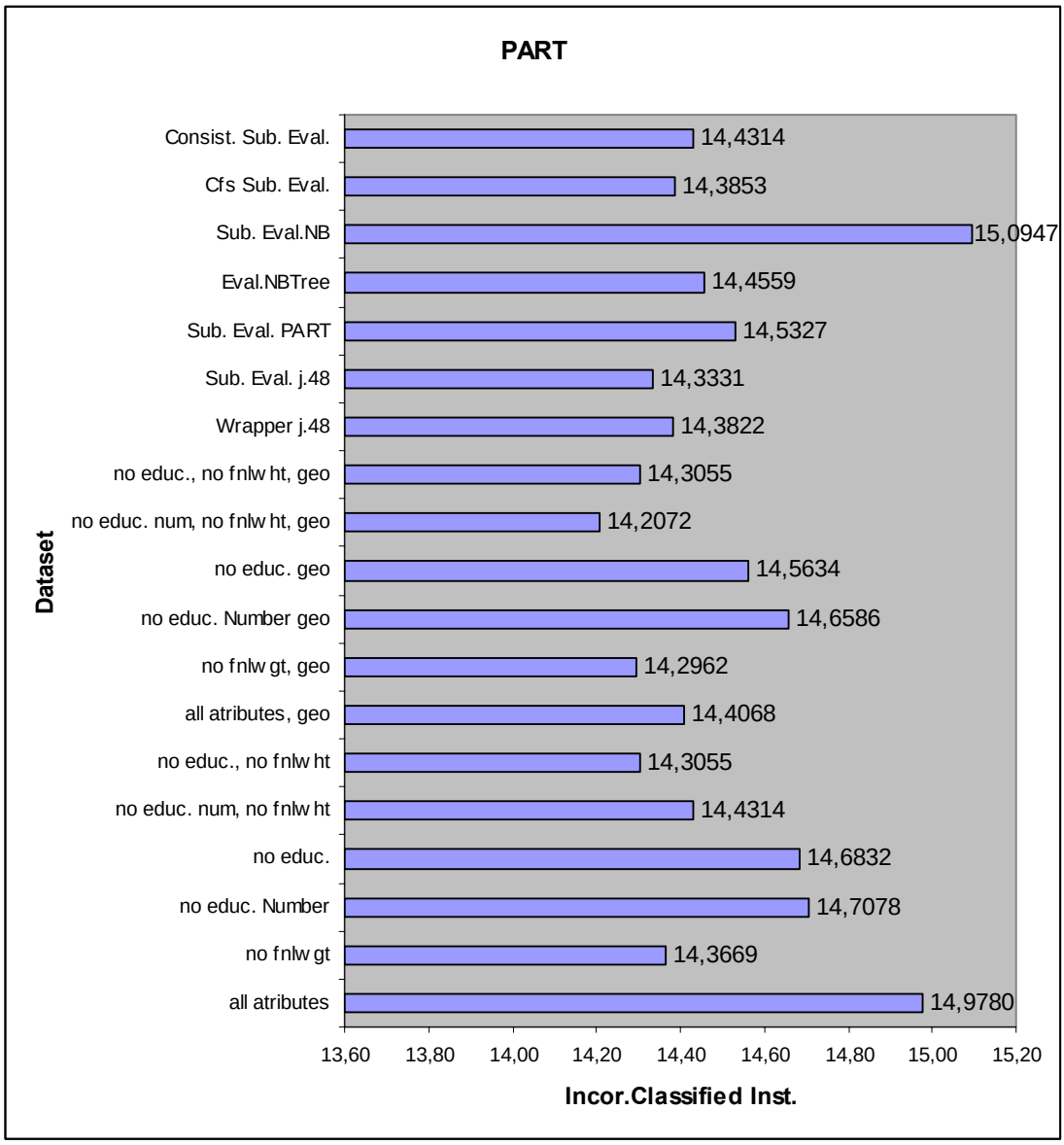
διάγραμμα 4.7 Δοκιμαζόμενες υποομάδες χαρακτηριστικών για τον αλγόριθμο J.48

Ο αλγόριθμος NBTree μάλλον κρίνεται ως ο πιο κατάλληλος για το εξεταζόμενο πρόβλημα. Όχι μόνο πετυχαίνει το υψηλότερο ποσοστό εύστοχων ταξινομήσεων με 86,6% όταν δοκιμάζεται με την υποομάδα που έχει προκύψει από τον αλγόριθμο επιλογής "Classifier Subset Evaluation-NBTree" (*Eval.NBTree*), αλλά έχει και την συνολικά καλύτερη παρουσία στα περισσότερα από τα εξεταζόμενα sub-set με το ποσοστό λάθος ταξινομήσεων να κινείται σε επίπεδα μικρότερα του 14%.



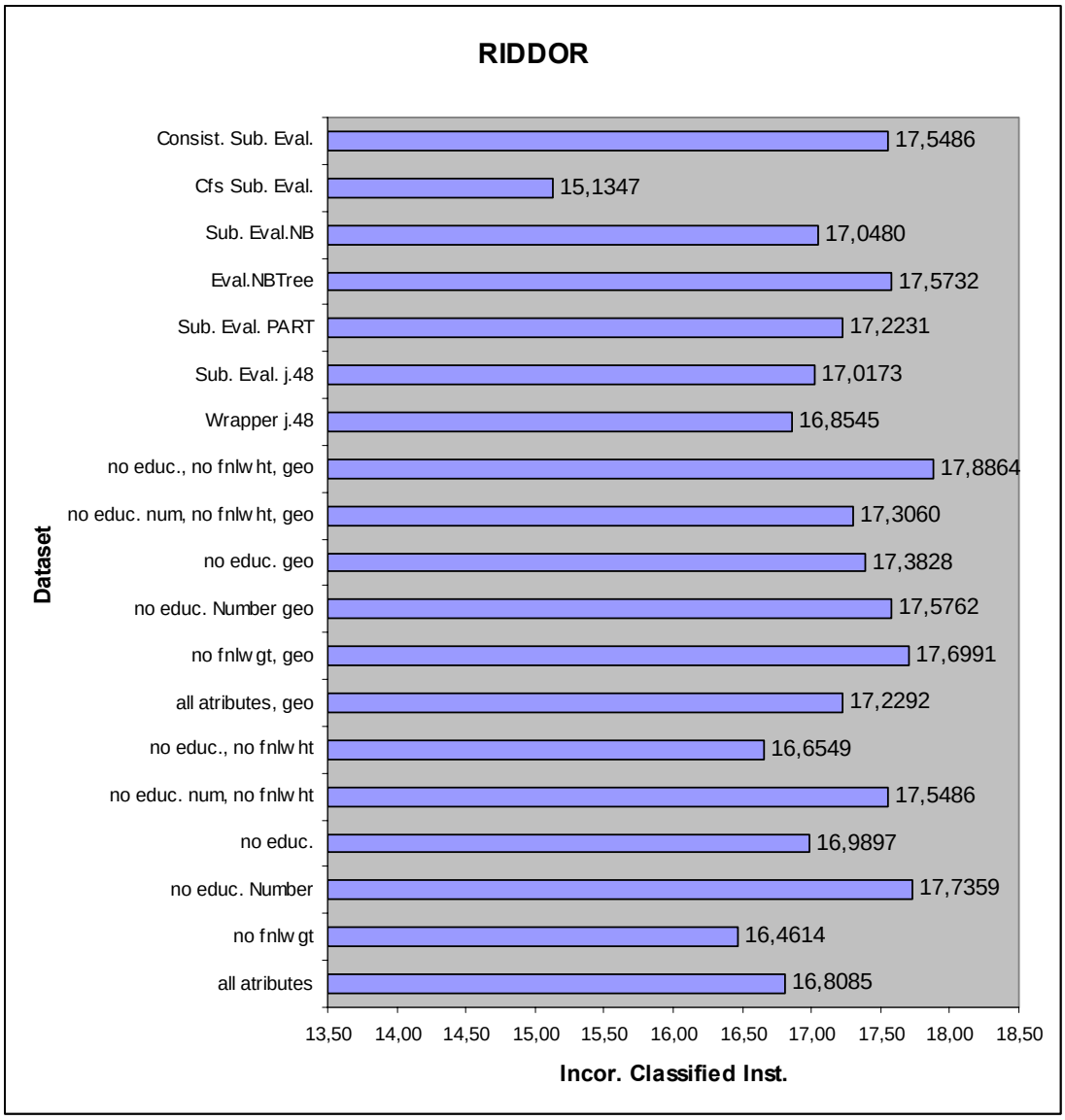
διάγραμμα 4.8 Δοκιμαζόμενες υποομάδες χαρακτηριστικών για τον αλγόριθμο NBTree

Ο PART δεν λειτουργεί το ίδιο καλά με τους δύο προηγούμενους αλγόριθμούς αν και θεωρητικά είναι της ίδιας φιλοσοφίας (δηλαδή στηρίζεται και αυτός εν μέρη στον αλγόριθμο C4.5). Για τον PART τα καλύτερα αποτελέσματα επετεύχθησαν με την υποομάδα χωρίς τα χαρακτηριστικά: EDUCATION NUMBER, FNLWGT και με το χαρακτηριστικό GEOGRAPHIC AREA αντί του NATIVE COUNTRY, όπου επιτεύχθηκε ποσοστό εύστοχης ταξινόμησης 85,8%.



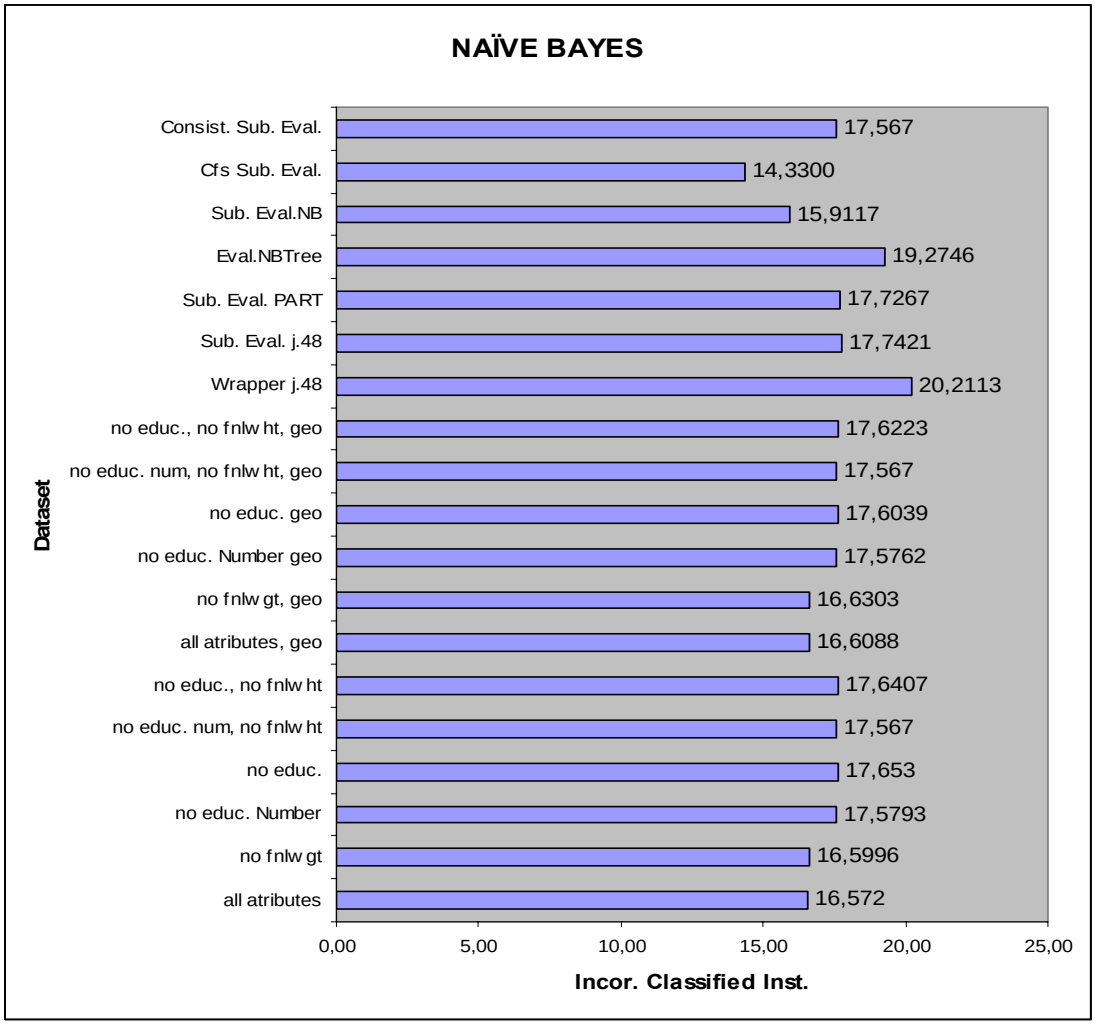
διάγραμμα 4.9 Δοκιμαζόμενες υποομάδες χαρακτηριστικών για τον αλγόριθμο PART

Οι υπόλοιποι τρείς εξεταζόμενοι αλγόριθμοι είναι προφανές ότι δυσκολεύονται να επεξεργαστούν το συγκεκριμένο σετ δεδομένων. Το γεγονός αυτό πιθανόν να είναι αποτέλεσμα της ύπαρξης θορύβου στα δεδομένα, των αρκετών απολεσθέντων τιμών ή κάποιων χαρακτηριστικών που δείχνουν, εκ πρώτης όψεως τουλάχιστον, μη-σχετικά με το χαρακτηριστικό-κλάση.



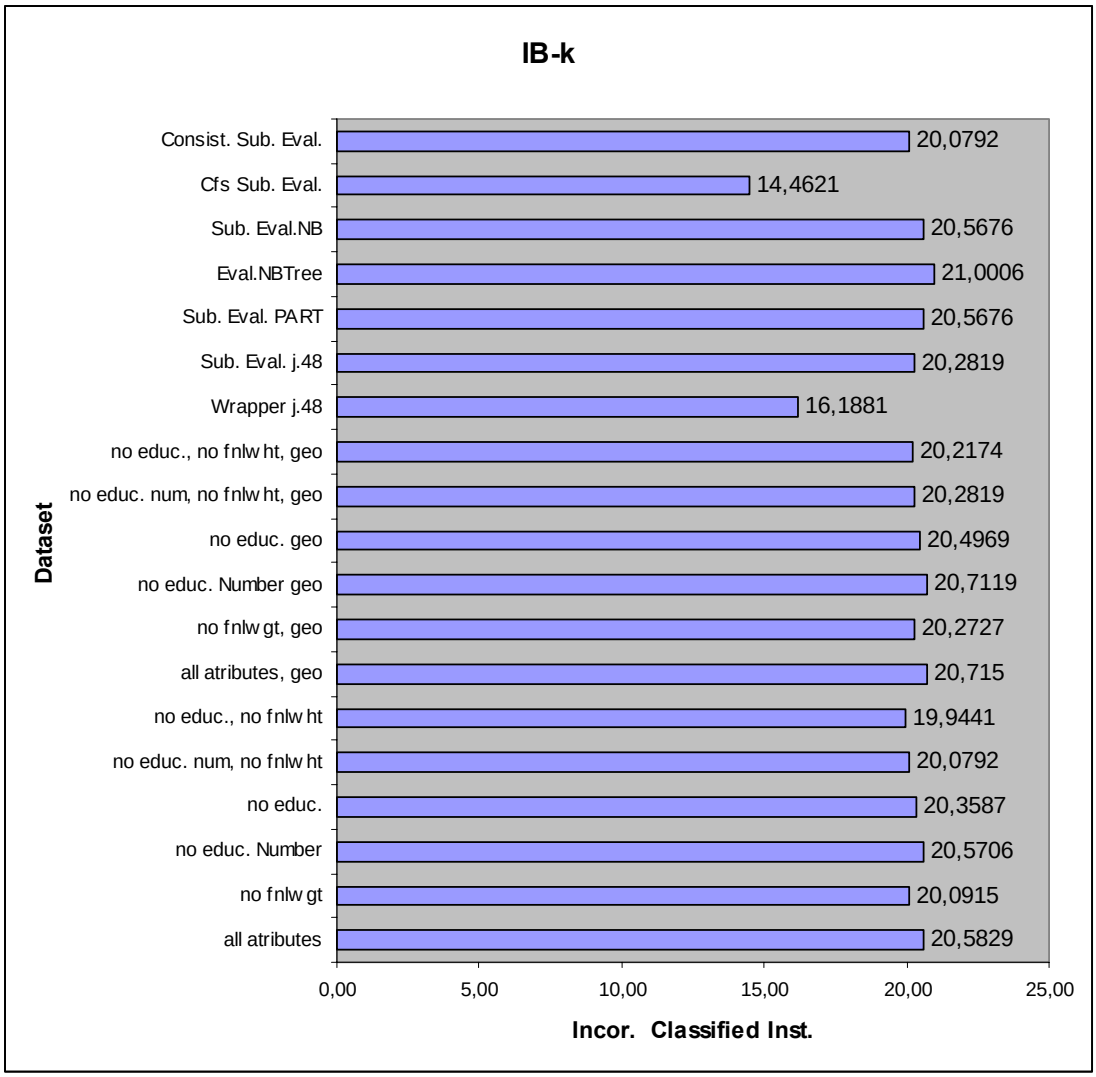
διάγραμμα 4.10 Δοκιμαζόμενες υποομάδες χαρακτηριστικών για τον αλγόριθμο RIDDOR

Ωστόσο και οι τρείς αυτοί αλγόριθμοι φαίνονται να λειτουργούν σαφώς πιο αποτελεσματικά με την υποομάδα δεδομένων που προέρχεται από τον αλγόριθμο “Correlation-based Feature Selection (CFS) Subset Evaluation”. Η ομάδα αυτή χαρακτηριστικών αποτελείται από πέντε μόλις χαρακτηριστικά (EDUCATION NUMBER, MARITAL STATUS RELATIONSHIP, CAPITAL GAIN, CAPITAL LOSS) τα οποία, όπως μπορεί κανείς να διαπιστώσει ανατρέχοντας στα αποτελέσματα που προέκυψαν από τους αλγόριθμους κατάταξης χαρακτηριστικών (πίνακας 4.1), είναι από εκείνα τα οποία συγκεντρώνουν τις υψηλότερες επιδόσεις.



διάγραμμα 4.11 Δοκιμαζόμενες υποομάδες χαρακτηριστικών για τον αλγόριθμο Naïve Bayes

Ο αλγόριθμος RIDDOR (όπως φαίνεται από το διάγραμμα 4.10) με τη χρήση της υποομάδας "CFS Subset Evaluation" φτάνει στο ποσοστό ευστοχίας 84,9%. Ο Naïve Bayes (στο διάγραμμα 4.11) φτάνει στο 85.77%, ενώ ο IBk (από το διάγραμμα 4.12) στο 85,5%.



διάγραμμα 4.12 Δοκιμαζόμενες υποομάδες χαρακτηριστικών για τον αλγόριθμο IBk

Το γεγονός αυτό δείχνει πως οι αλγόριθμοι αυτοί ευνοούνται από τη χρήση λιγότερων, αλλά με μεγαλύτερο βαθμό σχετικότητας, χαρακτηριστικών γεγονός που τους καθιστά μάλλον ακατάλληλους για το δεδομένο πρόβλημα..

5 ΣΥΜΠΕΡΑΣΜΑΤΑ

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

Στην πορεία αυτής της εργασίας παρουσιάστηκαν μια σειρά από αλγόριθμους ταξινόμησης οι οποίοι εντάσσονται στον ευρύτερο χώρο της μηχανικής μάθησης. Καθένας απ' αυτούς προσεγγίζει από διαφορετικές οπτικές γωνίες το πρόβλημα της δημιουργίας ενός σχήματος το οποίο αφ' ενός θα μπορεί να ταξινομήσει σωστά ένα στιγμιότυπο σε προκαθορισμένες κατηγορίες και απ' την άλλη θα μπορεί να προσφέρει πληροφορίες για το ίδιο το αντικείμενο της μάθησης μέσα από τα δεδομένα που χρησιμοποιεί για την εκπαίδευσή του.

Η εξαγωγή γνώσης από μια βάση δεδομένων είναι ένα δύσκολο εγχείρημα, ιδιαίτερα αν αναφερόμαστε σε τεραστίου μεγέθους βάσεις όπως οι περισσότερες που συναντάει κανείς σε προβλήματα του πραγματικού κόσμου. Το εγχείρημα γίνεται ακόμα

δυσκολότερο όταν οι ποιότητα των εξεταζόμενων δεδομένων είναι διαπραγματεύσιμη. Η ύπαρξη πλεοναζόντων και μη-σχετικών χαρακτηριστικών μέσα στα εξεταζόμενα δεδομένα έχουν ως αποτέλεσμα τον εκφυλισμό ακόμα και ανθεκτικών αλγορίθμων, όπως ο C4.5. Οι αλγόριθμοι επιλογής χαρακτηριστικών αποτελούν ένα σημαντικό εργαλείο στην αντιμετώπιση αυτού του προβλήματος. Μέσα από ένα σύνολο διαφορετικών αλγορίθμων είναι εφικτή η επιλογή υποομάδων μέσα από το σύνολο των αρχικών χαρακτηριστικών που συνθέτουν το εξεταζόμενο πρόβλημα τα οποία προβαλλόμενα σε ένα σύνολο δεδομένων μπορούν να πετύχουν 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 εξετάζεται ένα μεγάλο σε μέγεθος σετ δεδομένων αποτελούμενο τόσο από αριθμητικά όσο και από ονομαστικά χαρακτηριστικά το οποίο, βάσει αυτών, ταξινομεί τα δείγματα σε δύο προκαθορισμένες κλάσεις. Η ταξινόμηση του παραπάνω θα γίνει με την χρήση έξι διαφορετικών αλγορίθμων μηχανικής μάθησης των οποίων τα αποτελέσματα θα συγκριθούν -με τη βοήθεια της μεθοδολογίας cross-validation- ως προς το ποσοστό ευστοχίας ταξινόμησης ώστε να διαπιστωθεί ποιος μπορεί να ανταποκριθεί καλύτερα στο συγκεκριμένο πρόβλημα. Το πρόβλημα ωστόσο που μας απασχολεί στην συγκεκριμένη εργασία είναι αν είναι εφικτό μέσα από την απομόνωση κάποιων χαρακτηριστικών ή την μετατροπή κάποιων άλλων να γίνει εφικτή η επιλογή μιας υποομάδας χαρακτηριστικών η οποία θα πετύχει μεγαλύτερη ευστοχία ταξινόμησης στους ίδιους αλγόριθμους από το αρχικό σετ δεδομένων.

Μέσα από μια σειρά αλγορίθμων διήθησης επιχειρήθηκε, κατ' αρχήν, η αξιολόγηση των χαρακτηριστικών ως προς την διαχωριστική τους ικανότητα. Με γνώμονα τα αποτελέσματα της αξιολόγησης αυτής και μέσω της απαλοιφής είτε της τροποποίησης κάποιων χαρακτηριστικών δημιουργήθηκαν υποομάδες από τα αρχικά χαρακτηριστικά. Άλλες υποομάδες προέκυψαν από τη χρήση έτοιμων αλγορίθμων επιλογής υποομάδων. Όλα τα υπό-σετ χαρακτηριστικών που δημιουργήθηκαν ελέγχθηκαν από τους ιδίους έξι ταξινομητές με σκοπό να διαπιστωθεί α) το καταπόσο ωφελεί ή όχι η χρησιμοποίηση κάποιων εξ' αυτών έναντι του συνόλου των

χαρακτηριστικών β) Ποιοι αλγόριθμοι ευνοούνται από αυτές τις μετατροπές, ποιοι μένουν ανεπηρέαστοι και ποιων δυσχεραίνεται η λειτουργία.

Τα αποτελέσματα που προέκυψαν από τις εκτελέσεις των αλγορίθμων ταξινόμησης σε κάποιες περιπτώσεις μπορούν να εξηγηθούν με βάση τις γνώσεις που έχουμε για τη λειτουργία των συγκεκριμένων αλγορίθμων, σε κάποιες άλλες περιπτώσεις τα αποτελέσματα ήταν απρόσμενα. Δεν πρέπει να λησμονεί κανείς ότι οι εσωτερικές διαδικασίες που ακολουθούνται σε έναν αλγόριθμο κατά την εξέταση ενός μεγάλου σε μέγεθος dataset είναι εξαιρετικά πολύπλοκες και όχι πάντα εύκολο να ερμηνευτούν. Όπως πρέπει να συνυπολογίζεται ο παράγοντας τύχη του οποίου η συμβολή στην λήψη του τελικού αποτελέσματος είναι πάντα σημαντική, όσο και αν προσπαθούμε να το περιορίσουμε τη συμμετοχή του.

Συμπερασματικά, η επιλογή χαρακτηριστικών αποτελεί μια επιτυχημένη μεθοδολογία μείωσης της απαραίτητης πληροφορίας στα προβλήματα της επιστήμης που αφορούν την ταξινόμηση. Οι αλγόριθμοι επιλογής χαρακτηριστικών της μηχανικής μάθησης μπορούν να προσφέρουν ένα ακόμα χρήσιμο εργαλείο στο δύσκολο έργο της εξόρυξης δεδομένων.

6 ΒΙΒΛΙΟΓΡΑΦΙΑ

ΞΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Tan, P. N., Steinbach, M. and Kumar, V. *Introduction to Data Mining*. Addison Wesley, 2005.
2. Witten I., Frank E., *Data Mining Practical Machine Learning Tools*, 2000.
3. Kohavi R., *Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision-Tree Hybrid*, 1996.
4. Gaines B. R. and Compton P., *Induction of Ripple-Down Rules Applied to Modeling Large Databases*, 1995.
5. Compton P., Peters L., Edwards G. and Lavers T. G., *Experience with Ripple-Down Rules*. Knowledge-Based System Journal 19(5), 2006.
6. Quinlan J.R.*, Induction of decision trees*. Machine learning, 1986.
7. Quinlan J.R., *C4.5: Programs for Machine Learning*, 1993.
8. Frank E. and Witten I. A., *Generating Accurate Rule Sets Without Global Optimization*. In proc. Of the 15th Int. Conf. on Machine Learning, 1998.
9. Kohavi, R., and John G.H., *Feature Selection for Knowledge Discovery and Data Mining*, 1998
10. Langley P.,Simon H. A., *Applications of machine learning and rule induction*, 1995
11. Kohavi R., *Wrappers for Performance Enhancement and Oblivious Decision Graphs*, Phd thesis, Standford University, 1995
12. Kohavi R. and John G., *Wrappers for Feature Subset Selection*, Artificial Intelligence, Special Issue on relevance, 1996
13. Kohavi R., John G. H. and Pfleger, *Irrelevant Features and the Subset Selection Problem*. Proc. of the 11th International Conference. Morgan Kaufmann, 1994.
14. Kononenko I. and Bratko I., *Information Based Evaluation Criterion for Classifiers Performance*, Machine Learning, 1991
15. Schaffer C., *Selecting a Classification Method by Cross-Validation.* Machine Learning, 1993.
16. Ting K. M., *Common Issues in Instance Based and Naïve Bayesian Classifiers*. University of Sydney, NSW 2006.

17. Dougherty D., Kohavi R. and Sahami M., *Supervised and Unsupervised Discretization of Continuous Features*, Proc. of the 20th International Conference. Morgan Kaufmann, 1995.
18. Setiono R. and Liu H., Chi2: *Feature Selection and Discretization of Numeric Attributes*. In Proc. of the 7th IEEE Int. Conf. on Tools with A.I., 1995
19. Doak J., *An Evaluation of Feature Selection Methods and Their Application to Computer Security Technical Report*. University of California at Davis. Dept. Computer Science, 1992.
20. Liu H. and Motoda h., *Feature Selection for Knowledge Discovery and Data Mining*. Boston, Kluwer Academic, 1998.
21. Hall M.A., *Correlation Based Feature Selection for Descrete and Numeric Class Machine Learning*. Proc. of the 11th International Conference., Machine Learning 1994.
22. Bailey, T., and Jain A.K., *A Note on Distance-Weighted k-Nearest Neighbour Rules*. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-8 (4), 1978.
23. Dudani, S.A., *The Distance-Weighted k-Nearest Neighbour Rule*. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-6, 1976.
24. Narendra P.M. and Fukunaga, *A Brunch and Bound Algorithm for Feature Subset Selection.* IEEE Trans. Computer. Vol. 26, no. 9, Sept 1977.
25. Ben-Bassat M., *Pattern Recognition and Reduction of Dimentionality*. Handbook of Statistics-II. P.R. Krishnaiah and L.N. Kanal. North Holland, 1982.
26. Blum A.L. and Langley P., *Selection of Relevant Features and Examples in Machine Learning*. Artificial Intelligence. vol. 97, 1997.
27. John G.H., Kohavi R. and Pfleger K., *Irrelevant Feature and the Subset Selection Problem*. Proc. 11th International Conf. Machine Learning, 1994.
28. Langley P., *Selection of Relevant Features in Machine Learning*. In proc. Of the AAAI Fall Symposium on Relevance. AAAI Press, 1994.
29. Kononenko I., *Estimating Attributes: Analysis and Extension of RELIEF*. In proc. Of the European Conference on Machine Learning, Catania, Italy, 1994.
30. Sheinvald J., Dom B. and Niblack W., *A Modelling Approach to Feature Selection*. In proc. Of the 10th Int. Conference on Pattern Recognition, 1990.

31. Koller D. and Sahami M., *Toward Optimal Feature Selection*. Machine Learning: proc. 13th International Conference, San Francisco 1996.
32. Mucciardi A.N. and Gose E., *A Comparison of Seven Techiques for Choosing Subsets of Pattern Recognition*. IEEE Transactions on Computers, 1971.
33. Liu H. and Setiono R., *Scalable Feature Selection for Large Sized Databases*. In proc. of the Int. Conference on Machine Learning, 1996.
34. Liu H. and Setiono R., *Scalable Feature Selection for Large Sised Databases*. In proc. of the 4th World Congress on Expert Systems, Mexico City, Mexico, 1998.
35. Liu H. and Setiono R., *Feature Selection and Classification – A Probabilistic wrapper approach*. In proc. of the 9th International Conference of Industrial and Eng. Applications of AI and ES, *1996.*
36. Queiros C. E. and Gelsema E. S., *On Feature Selection*. In proc. Of the 7th Int. conference on Knowledge Discovery and Data Mining, AAAI Press, 1997.
37. Brambilla A. and Masella M., *Feature Subset Selection Using Effective Combine of Filter and Wrapper Approaches*. Dipl. Thesis, Technical Univesrsity of Milan, Italy, 2005.
38. Kotsiantis S. B., Kanellopoulos D. and Pintelas P. E., *Data Preprocessing for Supervised Learning*. Int. Journal of Computer Science, Volume 1 Number 2, ISSN, 2006.
39. Lu H., Sung Y. S. and Lu Y., *On Preprocessing Data for Effective Classification*. In proc. of the ACM SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, 1996.
40. Chen M. S., Han J. and Yu P. S., *Data Mining: An Overview from a Database Perspective*. IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, 1996.
41. Hall M. A. and Holmes G., *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining*. IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 6, 2003.
42. Sondberg-Madsen N., Thomsen C. and Pena J. M., *Unsupervised Subset Selection*. In Proc. of the Workshop on Probabilistic Graphical Models for Classification, 2003.

43. Jain A.K., Duin R.P.W. and Mao J., *Statistical Pattern Recognition: A Review*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, 2000.
44. Portinale L. and Saita L., *Feature Selection*. Dep. Of Information Engineering Univ. of Piemonte Orientale, Italy2002.
45. Almuallin H. and Dietterich T. G., Learning *with many Irrelevant Features*. In Proc. of the AAAI-91, Anhaim, CA, 1991.
46. Bell D.A. and Wang H., *A formalism for relevance and it's application in feature subset selection*. Machine Learning 41, 2000.

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

47. Τζιράλης Γ., *Σημειώσεις για το μάθημα Εξόρυξης Δεδομένων*, Ε.Μ.Π., 2006.
48. Δούμπος Μ., *Πολυκριτήριες Μέθοδοι Ταξινόμησης και Εφαρμογές στη Χρηματοοικονομική Διοίκηση*, 2000.
49. Σάκης Γ., *Αυτόματη Κατάταξη Μηνυμάτων Ηλεκτρονικού Ταχυδρομείου σε Κατηγορίες*. Πτυχιακή Εργασία, Πανεπιστήμιο Αθηνών Τμήμα Πληροφορικής, 2001.
50. Καρασιδέρη Κ., *Εφαρμογή Πολυκριτήριων Μεθόδων Τεχνητής Νοημοσύνης για την Ανάλυση της Συμπεριφοράς των καταναλωτών ελαιολάδου*. Πτυχιακή Εργασία, Πολυτεχνείο Κρήτης Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Χανια 2003.
51. Χρηστάκης Γ., *Επιλογή Χαρακτηριστικών σε Πρόβλημα Ταξινόμησης μέσω της Θεωρίας των Προσεγγιστικών Συνόλων και Τεχνικών Επαναληπτικής δειγματολειψίας*. Πτυχιακή Εργασία, Πολυτεχνείο Κρήτης Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Χανιά 2003.
52. Καλαπανίδας Θ. Η., *Περιβαλλοντική Πρόβλεψη με Μεθόδους Μηχανικής Μάθησης*. Διδακτορική Διατριβή, Πανεπιστήμιο Πατρών Τμήμα Ηλεκτρολόγων Μηχανικών κ Τεχνολογίας Υπολογιστών, Πάτρα 2003.

ΙΣΤΟΣΕΛΙΔΕΣ

53. UC Irvine Machine Learning Repository *(http://archive.ics.uci.edu/ml/)*
54. Waikato University, WEKA, (*http://www.cs.waikato.ac.nz/~ml/weka/*)