ΟΜΑΔΕΣ ΤΑΞΙΝΟΜΗΤΩΝ ΓΙΑ ΤΗΝ ΑΥΞΗΣΗ ΤΗΣ ΑΚΡΙΒΕΙΑΣ ΤΩΝ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Σωτήρης Β. Κωτσιαντής

Διδακτορική διατριβή

Πανεπιστήμιο Πατρών
Σχολή Θετικών Επιστημών
Τμήμα Μαθηματικών
Πάτρα

Επιβλέπων Καθηγητής: Παναγιώτης Ε. Πιντέλας

Σεπτέμβριος 2005

Στους γονείς μου

Πρόλογος

Η μηχανική μάθηση έχει ως σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν, δηλαδή ικανών να βελτιώνουν την απόδοση τους αξιοποιώντας προγενέστερη γνώση και εμπειρία. Αν και απέχουμε πολύ από τη δυνατότητα δημιουργίας μηχανών ικανών να μαθαίνουν τόσο αποτελεσματικά όπως οι άνθρωποι, εντούτοις έχουν αναπτυχθεί αλγόριθμοι που αποδίδουν πολύ ικανοποιητικά σε συγκεκριμένες περιοχές/εφαρμογές. Η εξόρυξη γνώσης είναι ένα άλλο ερευνητικό αντικείμενο που έχει συνδεθεί με την μηχανική μάθηση. Χωρίς την ύπαρξη και χρήση κατάλληλων μεθόδων μηχανικής μάθησης, ο αναλυτής αδυνατεί να εξάγει χρήσιμα συμπεράσματα από μεγάλα μεγέθη πληροφορίας (βάσεις δεδομένων).

Ένας αρκετά γενικός ορισμός που θα μπορούσε να δοθεί για τη μηχανική μάθηση είναι: “Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P , αν η απόδοση του σε εργασίες από το T , όπως μετριέται από το P , βελτιώνεται μέσω της εμπειρίας E .”

Πως ακριβώς μπορεί η μηχανική μάθηση να προβλέψει πράγματα που δεν ξέρουμε ή που θα συμβούν στο μέλλον; Η τεχνική που χρησιμοποιείται για να επιτευχθεί αυτό λέγεται μοντελοποίηση. Την σκέψη της δημιουργίας μοντέλων την είχαν οι άνθρωποι εδώ και πολύ καιρό και σίγουρα πριν την έλευση των υπολογιστών και της τεχνολογίας της εξόρυξης γνώσης. Για παράδειγμα, αν αναζητούσαμε ένα βυθισμένο αρχαίο πλοίο που μετέφερε θησαυρό το πρώτο πράγμα που ίσως σκεφτόμασταν θα ήταν να ερευνήσουμε όλες τις περασμένες περιπτώσεις εύρεσης θησαυρών από άλλους. Ίσως λοιπόν να παρατηρούσαμε ότι υπήρχαν κάποιες συγκεκριμένες πορείες που ακολουθούσαν οι καπετάνιοι των πλοίων αυτών εκείνη την εποχή. Με αυτό το μοντέλο (υπόθεση) θα αρχίζαμε το ψάξιμο σε περιοχές που είναι πιθανό να υπήρξε μια παρόμοια περίπτωση στο παρελθόν. Αν το μοντέλο μας είναι καλό ο θησαυρός θα βρεθεί.

Ο τρόπος της δημιουργίας των μοντέλων από τους υπολογιστές δεν διαφέρει πολύ από τον τρόπο με τον οποίο οι άνθρωποι δημιουργούν μοντέλα. Το ζητούμενο στην περίπτωση της μάθησης υπό επίβλεψη είναι η κατασκευή ενός μοντέλου (ή αλλιώς μιας υπόθεσης) που αναπαριστά τη γνώση που παρέχεται μέσω της εμπειρίας E και το οποίο στη συνέχεια χρησιμοποιείται για την αξιολόγηση νέων στιγμιοτύπων. Κατά κανόνα, οι προβλέψεις του προκύπτοντος μοντέλου (οι τιμές της συνάρτησης που προσεγγίζει τη συνάρτηση-στόχο) θα επαληθεύονται (θα ισούνται με την τιμή της συνάρτησης-στόχου) για την πλειοψηφία από τα στοιχεία που περιλαμβάνονται στην E , τα οποία λέγονται και στιγμιότυπα εκπαίδευσης.

Μία θεμελιώδης υπόθεση στην οποία στηρίζονται οι περισσότεροι αλγόριθμοι και η θεωρία μηχανικής μάθησης είναι πως η κατανομή των στιγμιοτύπων εκπαίδευσης είναι αντιπροσωπευτική της γενικής κατανομής των στιγμιοτύπων στον υπό μοντελοποίηση χώρο. Όπως είναι φυσικό, οι προβλέψεις ενός μοντέλου (ταξινομητή) για μελλοντικά (άγνωστα) στιγμιότυπα είναι περισσότερο αξιόπιστες αν τα στιγμιότυπα εκπαίδευσης ακολουθούν παρόμοια κατανομή με αυτή των μελλοντικών.

Η διεργασία της ανακάλυψης γνώσης και της μηχανικής μάθησης είναι μία διαδικασία πολλαπλών σταδίων, η οποία περιλαμβάνει την προετοιμασία των δεδομένων, την αναζήτηση προτύπων και την αξιολόγηση της γνώσης που ανακτάται από τα δεδομένα. Πρέπει σε αυτό το σημείο να αναφερθεί ότι στην περίπτωση της μηχανικής μάθησης τα δεδομένα είναι δομημένα και περιγράφονται από μία συγκεκριμένη ενιαία δομή όπου κάθε στιγμιότυπο ενός προβλήματος ορίζεται μέσω ενός συγκεκριμένου και σταθερού συνόλου χαρακτηριστικών (μεταβλητών). Ο βαθμός δυσκολίας ενός προβλήματος εξόρυξης γνώσης εξαρτάται στην μεταβλητότητα των τιμών των χαρακτηριστικών για στιγμιότυπα (αντικείμενα) της ίδιας κατηγορίας σε σχέση με τη διαφορά μεταξύ τιμών για στιγμιότυπα άλλων κατηγοριών. Γενικά, αν η απόδοση που λαμβάνεται από ένα δοθέν σύνολο χαρακτηριστικών είναι ανεπαρκής, τότε είναι λογικό να θεωρήσουμε ότι πρέπει να

προσθέσουμε νέα χαρακτηριστικά. Ειδικά τέτοια χαρακτηριστικά που θα βοηθήσουν στο διαχωρισμό των στιγμιοτύπων που ήταν δύσκολο να ταξινομηθούν σωστά.

Οι διάφοροι επιτηρούμενοι αλγόριθμοι μάθησης που έχουν αναπτυχθεί έως σήμερα διαφέρουν ως προς την υποκείμενη αναπαράσταση του χώρου των δυνατών υποθέσεων, και κατά συνέπεια και του τρόπου που οργανώνουν την αναζήτηση σε αυτό το χώρο. Μερικά παραδείγματα αναπαραστάσεων είναι τα δέντρα αποφάσεων (decision trees), τα τεχνητά νευρωνικά δίκτυα (artificial neural networks), τα δίκτυα Bayes και άλλα. Είναι πλέον αποδεκτό από όλη την επιστημονική κοινότητα ότι διαφορετικές τεχνικές επιτηρούμενης μηχανικής μάθησης είναι κατάλληλες για τη μάθηση διαφορετικών ειδών συναρτήσεων-στόχων και συνεπώς δεν είναι δυνατόν να κατασκευαστεί ένας αλγόριθμος που να υπερτερεί όλων των άλλων σε όλα τα πεδία εφαρμογών (Brazdil και άλλοι, 2003).

Η **επιστημονική συνεισφορά της διδακτορικής διατριβής** μπορεί να συνοψιστεί στα ακόλουθα:

- Πραγματοποιείται μια κριτική μελέτη των πλεονεκτημάτων και των μειονεκτημάτων κάθε κατηγορίας μεθόδων μηχανικής μάθησης ώστε ο χρήστης δεδομένου ενός προβλήματος εφαρμογής, να είναι σε θέση να επιλέξει την βέλτιστη τεχνική βάσει των προτεραιοτήτων του.
- Στην περίπτωση της αναζήτησης της βέλτιστης ακρίβειας από ένα σύστημα εξόρυξης γνώσης είναι δύσκολο αν όχι αδύνατο ένας αλγόριθμος που βασίζεται σε μια και μόνο μέθοδο μηχανικής μάθησης να υπερτερεί σε ακρίβεια μιας ομάδας ταξινομητών. Είναι εύκολα κατανοητό ότι αν όλοι οι ταξινομητές συμφωνούν στην ταξινόμηση ενός στιγμιότυπου τότε δεν υπάρχει κανένα πρόβλημα. Αν υποθέσουμε όμως το αντίθετο, πώς θα μπορούσε ένας υπερ-ταξινομητής να εξάγει τις απαραίτητες ενδείξεις από τις διάφορες συνιστώσες ώστε να επιτύχει βέλτιστη απόδοση; Η περισσότερη ερευνητική δουλειά της παρούσας διδακτορικής διατριβής κινήθηκε προς αυτή την κατεύθυνση. Προτείναμε ένα πλήθος νέων μεθόδων παραγωγής ομάδων ταξινομητών.
- Είναι γεγονός ότι αύξηση της ακρίβειας ταξινόμησης μπορούμε να πετύχουμε εκτός μέσω της δημιουργίας ομάδας ταξινομητών και από την προσεκτική επιλογή των μεταβλητών του προβλήματος. Προτείναμε μια υβριδική τεχνική επιλογής μεταβλητών που συνδυάζει τα πλεονεκτήματα των δύο γνωστών μεθόδων επιλογής μεταβλητών: των φίλτρων και των αλγορίθμων περιτυλίγματος.
- Επίσης εξετάσαμε την χρήση των μεθόδων μηχανικής μάθησης σε προβλήματα ειδικής δυσκολίας, όπως σε ανομοιογενή δεδομένα (imbalanced datasets) σε προβλήματα πραγματικού χρόνου (online learning) και σε προβλήματα βαθμωτής συνάρτησης στόχου (ordinal classification problems). Προτείναμε μερικούς νέους αλγορίθμους αντιμετώπισης αυτών των προβλημάτων.
- Τέλος, εξετάσαμε την δυνατότητα χρησιμοποίησης μεθόδων μηχανικής μάθησης για εκπαιδευτικούς σκοπούς, όπως την πρόβλεψη της επίδοσης των φοιτητών στο Ανοικτό Πανεπιστήμιο. Για το σκοπό αυτό κατασκευάσαμε και μερικά εργαλεία υποστήριξης των αποφάσεων τα οποία βρίσκουν εφαρμογή και σε πλήθος άλλων προβλημάτων.

Δομή της διδακτορικής διατριβής

Στο πρώτο κεφάλαιο γίνεται μια θεωρητική εισαγωγή στη μηχανική μάθηση δίνοντας περισσότερη έμφαση στην επιτηρούμενη μάθηση που απασχολεί περισσότερο την παρούσα διδακτορική διατριβή. Επίσης, γίνεται αναφορά πως συνδέεται η εξόρυξη γνώσης με την μηχανική μάθηση.

Όπως προείπαμε, η διαδικασία λύσεως ενός προβλήματος μηχανικής μάθησης είναι μία αλληλεπιδραστική επαναληπτική διαδικασία, η οποία περιλαμβάνει πλήθος βημάτων στα

οποία θα πρέπει να ληφθούν αποφάσεις από τον χρήστη. Στο δεύτερο κεφάλαιο αναλύονται τα βασικότερα βήματα της διαδικασίας αυτής όπως η επιλογή των δεδομένων, ο καθαρισμός τους κτλ.

Στο τρίτο κεφάλαιο αναφέρονται συνοπτικά οι διαφορετικές μέθοδοι επιτηρούμενης μηχανικής μάθησης, όπως τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα, τα δίκτυα Bayes και οι μηχανές υποστήριξης διανυσμάτων. Φυσικά, ένα κεφάλαιο όσο μεγάλο και αν είναι δεν μπορεί να αποτελεί μια πλήρης ανασκόπηση όλων των εποπτευόμενων αλγορίθμων μηχανικής μάθησης. Παρόλα αυτά, ελπίζουμε ότι οι αναφορές που αναφέρονται καλύπτουν τα σημαντικότερα θεωρητικά ζητήματα και καθοδηγούν τον ερευνητή σε ενδιαφέρουσες ερευνητικές κατευθύνσεις.

Στο τέταρτο κεφάλαιο γίνεται μια ανασκόπηση των σημαντικότερων μεθόδων συνδυασμού αλγορίθμων μάθησης και αναλύονται διάφοροι νέοι τρόποι συνδυασμού των αποφάσεων των αλγορίθμων μάθησης που αναπτύχθηκαν στα πλαίσια αυτής της διδακτορικής διατριβής.

Παρά το προφανές πλεονέκτημα της αύξησης της ακρίβειας από το συνδυασμό των ταξινομητών, υπάρχουν τουλάχιστον δύο αδυναμίες: (1) αύξηση του χρόνου εκπαίδευσης και (2) μειωμένη δυνατότητα κατανόησης των αποτελεσμάτων. Αύξηση της ακρίβειας μπορούμε να πετύχουμε εκτός από την δημιουργία ομάδας ταξινομητών και από την προσεκτική επιλογή των μεταβλητών του προβλήματος. Στο πέμπτο κεφάλαιο γίνεται μια ανασκόπηση των σημαντικότερων μεθόδων επιλογής των μεταβλητών και αναλύονται κάποιοι υβριδικοί τρόποι επιλογής μεταβλητών που αναπτύχθηκαν στα πλαίσια της διδακτορικής διατριβής.

Σε πολλές εφαρμογές οι αλγόριθμοι καλούνται να μάθουν από ένα σύνολο δεδομένων, που περιέχει πολύ περισσότερα παραδείγματα (στιγμιότυπα) από τη μια κλάση σε σχέση με την άλλη. Το έκτο κεφάλαιο είναι αφιερωμένο στην χρήση μεθόδων μηχανικής μάθησης σε προβλήματα ειδικής δυσκολίας όπως τα ανομοιογενή δεδομένα (imbalanced datasets). Επίσης αναλύεται η χρήση μεθόδων μάθησης σε προβλήματα πραγματικού χρόνου (online learning) και σε προβλήματα βαθμωτής συνάρτησης στόχου (ordinal classification problems). Τέλος, αναλύονται κάποιοι νέοι αλγόριθμοι που αναπτύχθηκαν στα πλαίσια της διδακτορικής διατριβής για την αντιμετώπιση αυτών των προβλημάτων.

Το έβδομο κεφάλαιο είναι αφιερωμένο στην δυνατότητα χρησιμοποίησης μεθόδων μηχανικής μάθησης για εκπαιδευτικούς σκοπούς, όπως την πρόβλεψη της επίδοσης των φοιτητών στο ανοιχτό πανεπιστήμιο. Στο ίδιο κεφάλαιο ακόμα αναφέρουμε πως οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν σε άλλα προβλήματα όπως αυτά της χημικής μηχανικής: έλεγχος της ποιότητας του κρασιού και της ποιότητας των χαρτομάντιλων.

Παρότι η παρούσα διατριβή ασχολείται κυρίως με την επιτηρούμενη μηχανική μάθηση θα ήταν παράληψη να μην αναφερθεί τίποτα για την μη επιτηρούμενη μηχανική μάθηση. Στη μη επιτηρούμενη μάθηση, δεν παρέχεται κάποια εμπειρία E για να καθοδηγήσει τη μάθηση, αλλά ο στόχος είναι να αναδειχθεί η δομή οργάνωσης των δεδομένων μέσω κάποιου ή κάποιων κατάλληλα επιλεγμένων κριτηρίων “ομοιότητας”. Στο όγδοο κεφάλαιο αναφέρονται συνοπτικά οι διάφορες κατηγορίες μεθόδων μη επιτηρούμενης μηχανικής μάθησης.

Στο τελευταίο κεφάλαιο συνοψίζεται η διδακτορική διατριβή και παρουσιάζονται κάποια ανοιχτά επιστημονικά ζητήματα του χώρου. Γίνεται επίσης μια εισαγωγή στην πρόσφατα εμφανιζόμενη εξόρυξη γνώσης από πολυμεσικό υλικό που αποτελεί πιθανή προέκταση των μεθόδων που αναπτύχθηκαν στην διατριβή.

Τέλος την διδακτορική διατριβή ολοκληρώνουν 4 παραρτήματα. Στο πρώτο περιγράφονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για τις αναγκαίες συγκρίσεις των προτεινόμενων μεθόδων με τις υπάρχουσες. Στο δεύτερο αναλύεται η χρήση του καλύτερου, κατά την άποψη του συγγραφέα, ελεύθερα διαθέσιμου εργαλείου εξόρυξης γνώσης. Στο τρίτο παρουσιάζονται οι ιστοσελίδες διάθεσης όλων των εργαλείων εξόρυξης τα οποία διατίθενται προς πώληση. Τέλος, το τελευταίο παράρτημα είναι το γλωσσάρι των ελληνικών όρων που υποδεικνύει την αντιστοίχιση των ελληνικών με τους αγγλικούς όρους.

Για την ολοκλήρωση του προλόγου παραθέτουμε τις εργασίες που πραγματοποιήθηκαν στα πλαίσια της διδακτορικής διατριβής.

Εργασίες στα πλαίσια της διδακτορικής διατριβής

1. S. Kotsiantis, C. Pierrakeas, P. Pintelas, Predicting Students' Performance in Distance Learning Using Machine Learning Techniques, Applied Artificial Intelligence (AAI), Volume 18, Number 5 / May-June 2004, pp. 411 - 426.
2. S. Kotsiantis, P. Pintelas, Increasing the Classification Accuracy of Simple Bayesian Classifier, Lecture Notes in Artificial Intelligence, Springer-Verlag Vol 3192, 2004, pp. 198-207¹.
3. S. Kotsiantis, P. Pintelas, Logitboost of Simple Bayesian Classifier, Computational Intelligence in Data mining Special Issue of the Informatica Journal, Vol 29 (1), pp. 53-59, 2005.
4. S. Kotsiantis, P. Pintelas, Local Voting of Weak Classifiers, Local Voting of Weak Classifiers, International Journal of Knowledge-Based & Intelligent Engineering Systems (KES), 2005, pp. 239 - 248.
5. S. Kotsiantis, P. Pintelas (2004), An Online Ensemble Of Classifiers, The Fourth International Workshop on Pattern Recognition in Information Systems (PRIS), In conjunction with 6th International Conference on Enterprise Information Systems, pp 59-68, Porto - Portugal 14-17, April 2004².
6. S. Kotsiantis, P. Pintelas, Selective Voting, IEEE 4th International Conference on Intelligent Systems Design and Applications (ISDA 2004), August 26-28, 2004, Budapest, Hungary, pp. 397-402.
7. S. Kotsiantis, P. Pintelas (2004), Hybrid Feature Selection instead of Ensembles of Classifiers in Medical Decision Support, Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), July 4-9, Perugia - Italy, pp. 269-276³.
8. S. Kotsiantis, I. Zaharakis, V. Tampakas, P. Pintelas, On Constructing a Financial Decision Support System, Proceedings of International Conference on Enterprise Systems and Accounting 2004, September 3-4, 2004, Thessaloniki, Greece, pp 319-331.
9. S. Kotsiantis, P. Pintelas, Recent Advances in Clustering: A Brief Survey, WSEAS Transactions on Information Science and Applications, Vol 1, No 1 (73-81), 2004⁴.
10. S. Kotsiantis, P. Pintelas, Combining Bagging and Boosting, International Journal of Computational Intelligence, Vol. 1, No. 4 (324-333), 2004.

Οι εργασίες είναι αριθμημένες και με την αρίθμηση τους αναφέρονται στο κείμενο σε αντίθεση με τις υπόλοιπες αναφορές που επισημαίνονται με το όνομα του πρώτου συγγραφέα.

¹ Η συγκεκριμένη εργασία ήταν υποψήφια για το βραβείο «Best paper award» στο συνέδριο: 11th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, 2-4 September, Varna (Bulgaria).

² Η συγκεκριμένη εργασία έχει και αναφορά από: Grinaldo Oliveira, Manoel Mendonça, ExperText: Uma Ferramenta de Combinação de Múltiplos Classificadores Naive Bayes, 4th Ibero-American Symposium on Software Engineering and Knowledge Engineering (JIISIC'04), pp. 317-332.

³ Η συγκεκριμένη εργασία έχει και αναφορά από: Edwige Fangseu Badjio and Francois Poulet, Dimension Reduction for Visual Data Mining, Applied Stochastic Models and Data Analysis (ASMDA 2005), pp. 266-275.

⁴ Η συγκεκριμένη εργασία προτείνεται από τον καθηγητή Gedas Adomavicius για επιπλέον μελέτη του μαθήματος «Data Mining for Business Research» στο Πανεπιστήμιο της Μινεσότα και από τον καθηγητή Tom Bylander για επιπλέον μελέτη του μαθήματος «Machine Learning» στο Πανεπιστήμιο του Τέξας.

Ευχαριστίες

Αρκετοί ήταν αυτοί που συνέβαλαν με τον ένα ή τον άλλο τρόπο για την πραγματοποίηση της παρούσας διδακτορικής διατριβής και κατά συνέπεια νιώθω την ανάγκη να τους ευχαριστήσω.

Καταρχήν, ευχαριστώ τον Καθηγητή κ. Π. Πιντέλα για την υποστήριξή του σε όλη την προσπάθεια καθώς τόσο σε επιστημονικό όσο και σε προσωπικό επίπεδο έδινε ακούραστα τις συμβουλές του. Επίσης, ευχαριστώ τα άλλα δύο μέλη της συμβουλευτικής επιτροπής τους καθηγητές κ. Μ. Βραχάτη και κ. Σ. Λυκοθανάση για τις εύστοχες παρατηρήσεις τους.

Κατά την διάρκεια της εκπόνησης της διατριβής μου είχα την χαρά να συνεργαστώ με τον Δρ. Ιωάννη Ζαχαράκη, τον Δρ. Χ. Πιερρακέα και τον καθηγητή Β. Ταμπακά και γι' αυτό τους ευχαριστώ θερμά.

Επίσης αισθάνομαι την ανάγκη να ευχαριστήσω τα άλλα τέσσερα μέλη της επταμελούς εξεταστικής επιτροπής της παρούσας διατριβής, τους καθηγητές: Δ. Χριστοδουλάκη, Α. Τσακαλίδη, Ι. Μανωλόπουλο και τον επίκουρο καθηγητή Β. Βουτσινά.

Το μεγαλύτερο «ευχαριστώ» όμως ανήκει στους γονείς μου που ανιδιοτελώς μου προσφέρουν και θα μου προσφέρουν την αγάπη τους. Σε αυτούς αφιερώνω αυτή την εργασία.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ	19
1.1 Λόγοι που κάνουν την εξόρυξη γνώσης με μηχανική μάθηση όλο και πιο δημοφιλή ..	22
1.2 Εφαρμογές της μηχανικής μάθησης και της εξόρυξης γνώσης	23
ΚΕΦΑΛΑΙΟ 2. Η διαδικασία επίλυσης ενός προβλήματος Μηχανικής Μάθησης	25
2.1 Επιλογή δεδομένων/Αξιολόγηση των δεδομένων.....	26
2.1.1 Καθαρισμός και επεξεργασία των δεδομένων	26
2.2 Εξαγωγή χρήσιμων χαρακτηριστικών/μεταβλητών.....	33
2.3 Επιλογή του κατάλληλου αλγορίθμου μάθησης	35
2.4 Αξιολόγηση της παραγόμενης γνώσης	39
ΚΕΦΑΛΑΙΟ 3. Επιτηρούμενη μηχανική μάθηση	41
3.1 Δέντρα Αποφάσεων.....	42
3.1.1 Επιλογή μεταβλητών για την δημιουργία του δέντρου απόφασης	44
3.1.2 Αντιμετώπιση του υπερτρεψιμότητας στα δέντρα αποφάσεων.....	48
3.1.3 Σύγκριση αλγορίθμων κατασκευής δέντρων απόφασης.....	51
3.2 Κανόνες ταξινόμησης.....	53
3.2.1 Σύγκριση αλγορίθμων παραγωγής κανόνων ταξινόμησης	56
3.2.2 Χρήση γενετικών αλγορίθμων για την παραγωγή κανόνων ταξινόμησης.....	57
3.2.3 Κανόνες πρώτης τάξης	59
3.3 Μάθηση βασισμένη στα στιγμιότυπα	62
3.3.1 Σύγκριση οκνηρών αλγορίθμων μάθησης.....	69
3.4 Τεχνητά Νευρωνικά Δίκτυα.....	70
3.4.1 Το μοντέλο του τεχνητού νευρώνα	70
3.4.2 Νευρωνικά δίκτυα	73
3.4.3 Σύγκριση αλγορίθμων	81
3.5 Μπεϋζιανοί ταξινομητές	83
3.5.1 Βέλτιστος ταξινομητής Bayes.....	84
3.5.2 Αφελής ταξινομητής	84
3.5.3 Σύγκριση αλγορίθμων	88
3.5.4 Μπεϋζιανά Δίκτυα.....	89
3.5.5 Σύγκριση αλγορίθμων κατασκευής Μπεϋζιανών δικτύων	93
3.6 Μηχανές διανυσμάτων υποστήριξης.....	94
3.6.1 Διαχωρίσιμη περίπτωση.....	94
3.6.2 Μη γραμμικές μηχανές υποστήριξης διανυσμάτων	98
3.6.3 Σύγκριση πολυωνυμικών πυρήνων	102
3.7 Θεωρία υπολογιστικής μάθησης	104
3.7.1 Διάσταση VC	106
ΚΕΦΑΛΑΙΟ 4. Ομάδες ταξινομητών και συνεισφορά της διατριβής στη δημιουργία νέων μεθόδων.....	108
4.1 Τεχνικές που βασίζονται σε έναν αλγόριθμο μάθησης και την δειγματοληψία των δεδομένων εκπαίδευσης	108
4.1.1 Προτεινόμενη μέθοδος συνδυασμού σακουλιάσματος και προώθησης.....	112
4.1.2 Προτεινόμενη μέθοδος τοπικής προώθησης.....	119
4.1.3 Προτεινόμενη μέθοδος τοπικού σακουλιάσματος	123
4.1.4 Προτεινόμενη μέθοδος προώθησης επιλεκτικών απλών Μπεϋζιανών ταξινομητών	125
4.1.5 Προτεινόμενη μέθοδος συνδυασμού της λογιστικής προώθησης με τον απλό μπεϋζιανό αλγόριθμο	130
4.2 Συνδυασμός αποφάσεων διαφορετικών ταξινομητών	135

4.2.1	Συνδυάζοντας τους αλγόριθμους NB και C4.5	139
4.2.2	Προτεινόμενη μέθοδος ψηφοφορίας μέσα σε σάκο	141
4.2.3	Προτεινόμενη μέθοδος επιλεκτικής ψηφοφορίας	145
4.2.4	Προτεινόμενη μέθοδος βελτιωμένης συσσωρευμένης γενίκευσης.....	150
4.2.5	Προτεινόμενη μέθοδος τοπικής ψηφοφορίας απλών ταξινομητών	154
4.2.6	Προτεινόμενη μέθοδος γρήγορης ομάδας ταξινομητών	159
4.3	Αλλαγή της συνάρτησης-στόχου	163
4.4	Εισαγωγή τυχαιότητας στον αλγόριθμο μάθησης.....	165
ΚΕΦΑΛΑΙΟ 5. Επιλογή μεταβλητών και συνεισφορά της διατριβής στη δημιουργία υβριδικής μεθόδου		166
5.1	Σύγκριση των πιο γνωστών μεθόδων επιλογής μεταβλητών περιτυλίγματος	166
5.2	Προτεινόμενη υβριδική τεχνική επιλογής μεταβλητών	174
ΚΕΦΑΛΑΙΟ 6. Μάθηση σε ειδικά προβλήματα και συνεισφορά της διατριβής σε αυτά		177
6.1	Προτεινόμενη μέθοδος μάθησης σε ανομοιογενή (imbalanced) δεδομένα	178
6.2	Προτεινόμενη μέθοδος μάθησης σε πραγματικό χρόνο.....	183
6.3	Προτεινόμενη μέθοδος μάθησης βαθμωτής συνάρτησης στόχου.....	187
ΚΕΦΑΛΑΙΟ 7. Εφαρμογές των προτεινόμενων μεθόδων αλλά και των μεθόδων μηχανικής μάθησης γενικότερα σε πραγματικά προβλήματα		194
7.1	Εφαρμογές μεθόδων μηχανικής μάθησης για εκπαιδευτικούς σκοπούς.....	194
7.1.1	Πρόβλεψη της εγκατάλειψης των σπουδών από το ελληνικό ανοιχτό πανεπιστήμιο	194
7.1.2	Πρόβλεψη απόδοσης φοιτητών.....	197
7.1.3	Εύρεση των μεταβλητών που επηρεάζουν την πρόβλεψη επίδοσης.....	199
7.1.4	Εργαλείο υποστήριξης αποφάσεων για τους καθηγητές του ανοιχτού πανεπιστήμιου	199
7.1.5	Πρόβλεψη βαθμού σπουδαστή.....	204
7.2	Άλλες εφαρμογές χρήσης των μεθόδων μάθησης.....	206
7.2.1	Χρήση μεθόδων μάθησης για την εκτίμηση της ποιότητας του κρασιού.....	207
7.2.2	Χρήση μεθόδων μάθησης για εκτίμηση ποιότητας χαρτομάντιλων	208
ΚΕΦΑΛΑΙΟ 8. Μη επιτηρούμενη μηχανική μάθηση		211
8.1	Αναπαράσταση των δεδομένων	212
8.2	Μέτρα Ομοιότητας.....	213
8.3	Τεχνικές συγκρότησης	214
8.3.1	Ιεραρχικοί αλγόριθμοι.....	214
8.3.2	Διαμεριστικοί αλγόριθμοι	216
8.3.3	Αυτό-Οργανούμενοι Χάρτες (SOM).....	218
8.3.4	Αλγόριθμοι βασισμένοι στη πυκνότητα (Density based)	220
8.3.5	Αλγόριθμοι βασισμένοι στον Κάνναβο (Grid-based Algorithms).....	221
8.4	Αναπαράσταση των συστάδων	221
8.5	Εκτίμηση του αποτελέσματος.....	223
ΚΕΦΑΛΑΙΟ 9. Επίλογος.....		224
9.1	Επεκτάσεις της παρούσας εργασίας.....	225
Αναφορές		228
Παράρτημα Α: Δεδομένα.....		238
Παράρτημα Β: Το εργαλείο εξόρυξης γνώσης Weka (ένας μικρός οδηγός χρήσης)		240
Χρήση του Weka για αυτόματη ταξινόμηση		241
Χρήση του Weka για αυτόματη ομαδοποίηση.....		243
Αυτόματη αναζήτηση των πιο σημαντικών χαρακτηριστικών		245
Παράρτημα Γ: Άλλα εργαλεία εξόρυξης γνώσης		246
Παράρτημα Δ: Αντιστοίχιση ελληνικών και αγγλικών όρων		248

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1. Δεδομένα εκπαίδευσης για ένα τραπεζικό σύστημα αξιολόγησης πελατών για δανειοδότηση	20
Πίνακας 2. Ταξινόμηση Μεταβλητών	26
Πίνακας 3. Έλεγχος για σφάλματα στα δεδομένα	27
Πίνακας 4. Ο αρχικός πίνακας δεδομένων με ελλιπή δεδομένα (απουσία τιμών).....	28
Πίνακας 5. Ο πίνακας δεδομένων μετά την διαγραφή των στιγμιοτύπων που περιέχουν ελλιπή δεδομένα.....	29
Πίνακας 6. Ο πίνακας δεδομένων μετά την ακτικατάσταση των ελλιπών δεδομένων με την διάμεσο ή την μέση τιμή της αντίστοιχης μεταβλητής	29
Πίνακας 7. Ο πίνακας δεδομένων μετά την ακτικατάσταση των ελλιπών δεδομένων με την διάμεσο ή την μέση τιμή της αντίστοιχης μεταβλητής λαμβάνοντας υπόψη τα στιγμιότυπα μόνο της ίδιας κλάσης	29
Πίνακας 8. Ο πίνακας δεδομένων μετά την ακτικατάσταση των ελλιπών δεδομένων με όλους τους πιθανούς συνδυασμούς τιμών	30
Πίνακας 9. Ο πίνακας δεδομένων μετά την ακτικατάσταση των ελλιπών δεδομένων με όλους τους πιθανούς συνδυασμούς τιμών λαμβάνοντας υπόψη μόνο τα στιγμιότυπα της ίδιας κλάσης.....	31
Πίνακας 10. Ο αρχικός πίνακας δεδομένων που περιέχει μια συνεχή μεταβλητή.....	31
Πίνακας 11. Ο πίνακας δεδομένων μετά την κατάτμηση του διαστήματος τιμών της συνεχούς μεταβλητής σε διαστήματα που περιέχουν ίδιο αριθμό περιπτώσεων.....	32
Πίνακας 12. Ο πίνακας δεδομένων μετά την κατάτμηση του διαστήματος τιμών της συνεχούς μεταβλητής με την εφαρμογή της μεθόδου της εντροπίας.....	33
Πίνακας 13. Δεδομένα εκπαίδευσης	43
Πίνακας 14. Σύγκριση αλγορίθμων κατασκευής δέντρων απόφασης	52
Πίνακας 15. Σύγκριση αλγορίθμων δημιουργίας κανόνων.....	57
Πίνακας 16. Δεδομένα εκπαίδευσης.....	64
Πίνακας 17. Πελάτης προς ταξινόμηση	64
Πίνακας 18. Κανονικοποίηση των δεδομένων εκπαίδευσης	64
Πίνακας 19. Μετρικές απόστασης στιγμιοτύπων	65
Πίνακας 20. Σύγκριση οκνηρών αλγορίθμων μάθησης	70
Πίνακας 21. Λογικό XOR	74
Πίνακας 22. Σύγκριση αλγορίθμων μάθησης νευρωνικών δικτύων	82
Πίνακας 23. Δεδομένα εκπαίδευσης	85
Πίνακας 24. Σύγκριση απλοϊκών ταξινομητών Μπέντζ.....	89
Πίνακας 25. Τα δεδομένα εκπαίδευσης του παραδείγματος μας.....	92
Πίνακας 26. Σύγκριση αλγορίθμων κατασκευής Μπεϋζιανών δικτύων	93
Πίνακας 27. Πρόβλημα OR.....	99
Πίνακας 28. Πρόβλημα XOR.....	100
Πίνακας 29. Σύγκριση πολυωνυμικών πυρήνων.....	103
Πίνακας 30. Σχετική βελτίωση της ακρίβειας των ταξινομητών μετά την χρήση σακουλιάσματος και προώθησης	111
Πίνακας 31. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον C4.5.....	114
Πίνακας 32. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον DS	115
Πίνακας 33. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον NB.....	117
Πίνακας 34. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον OneR	118

Πίνακας 35. Σύγκριση της προτεινόμενης μεθόδου τοπικής προώθησης χρησιμοποιώντας ως βασικό ταξινομητή τον DS με τους υπόλοιπους αλγορίθμους.....	121
Πίνακας 36. Σύγκριση της προτεινόμενης μεθόδου τοπικής προώθησης χρησιμοποιώντας ως βασικό ταξινομητή τον OneR με τις υπόλοιπες μεθόδους.....	122
Πίνακας 37. Σύγκριση της προτεινόμενης μεθόδου τοπικού σακουλιάσματος με άλλες μεθόδους.....	124
Πίνακας 38. Σύγκριση της προτεινόμενης μεθόδου (BoostFSNB) με άλλες μεθόδους	128
Πίνακας 39. Σύγκριση της προτεινόμενης μεθόδου (BoostFSNB) με άλλες ομάδες ταξινομητών	129
Πίνακας 40. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλες μπεϋζιανές τεχνικές.....	132
Πίνακας 41. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλες κοινές ομάδες ταξινομητών που χρησιμοποιούν τον NB ως βασικό αλγόριθμο.....	133
Πίνακας 42. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλους μετα-ταξινομητές	134
Πίνακας 43. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλες ομάδες ταξινομητών	135
Πίνακας 44. Σύγκριση της εξεταζόμενης μεθόδου με άλλες ομάδες ταξινομητών που βασίζονται είτε στον C4.5 είτε στον NB.....	140
Πίνακας 45. Σύγκριση της προτεινόμενης μεθόδου ψηφοφορίας μέσα σε σάκο με άλλες ομάδες ταξινομητών που χρησιμοποιούν σαν βασικούς ταξινομητές είτε τον NB είτε τον C4.5	143
Πίνακας 46. Σύγκριση της προτεινόμενης μεθόδου ψηφοφορίας μέσα σε σάκο με άλλους αλγορίθμους που συνδυάζουν τις αποφάσεις των NB και C4.5	144
Πίνακας 47. Σύγκριση της επιλεκτικής ψηφοφορίας με όλους τους αλγόριθμους βάσης.....	146
Πίνακας 48. Σύγκριση της προτεινόμενης μεθόδου με άλλες ομάδες ταξινομητών	148
Πίνακας 49. Εκτέλεση της μεθόδου επιλεκτικής ψηφοφορίας σε υποσύνολα των δεδομένων	150
Πίνακας 50. Σύγκριση της προτεινόμενης μεθόδου Stacking' με τους βασικούς αλγόριθμους και ομάδες ταξινομητών των αλγορίθμων βάσης	152
Πίνακας 51. Σύγκριση της προτεινόμενης μεθόδου (Stacking') με άλλες ομάδες ταξινομητών	153
Πίνακας 52. Η σύγκριση του αλγορίθμου τοπικής ψηφοφορίας με τους αλγόριθμους βάσης χρησιμοποιώντας καθολική ή τοπική μάθηση	156
Πίνακας 53. Η σύγκριση του αλγορίθμου τοπικής μάθησης με άλλες ομάδες ταξινομητών	157
Πίνακας 54. Η σύγκριση του αλγορίθμου τοπικής ψηφοφορίας με άλλες περίπλοκες ομάδες ταξινομητών	158
Πίνακας 55. Η σύγκριση του προτεινόμενου αλγορίθμου VoteDVN με άλλους ισχυρούς αλγόριθμους μάθησης	162
Πίνακας 56. Η σύγκριση της τεχνικής VoteDVN με άλλες γρήγορες ομάδες ταξινομητών	163
Πίνακας 57. Διάσπαση ενός προβλήματος με 4 κλάσεις χρησιμοποιώντας την τεχνική ενός ταξινομητή ανά κατηγορία.....	164
Πίνακας 58. Διάσπαση ενός προβλήματος με 4 κλάσεις χρησιμοποιώντας την τεχνική ECOC	164
Πίνακας 59. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον NB.....	169
Πίνακας 60. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον C4.5.....	170
Πίνακας 61. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον RIPPER.....	171
Πίνακας 62. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον 3NN.....	173
Πίνακας 63. Σύγκριση της ακρίβειας των μεθόδων μετά την εφαρμογή της προτεινόμενης υβριδικής επιλογής των μεταβλητών.	176

Πίνακας 64. Σύγκριση της προτεινόμενης μεθόδου (ABKD) με άλλες που χρησιμοποιούν ως βασικό ταξινομητή τον NB	181
Πίνακας 65. Σύγκριση της προτεινόμενης μεθόδου (ABKD) με άλλες που χρησιμοποιούν ως βασικό ταξινομητή τον C4.5	182
Πίνακας 66. Σύγκριση της προτεινόμενης μεθόδου (ABKD) με άλλες που χρησιμοποιούν ως βασικό ταξινομητή τον 5NN	183
Πίνακας 67. Συγκρίνοντας την προτεινόμενη μέθοδο (Voting online ensemble) με τους αλγορίθμους βάσης	186
Πίνακας 68. Σύγκριση της προτεινόμενης μεθόδου (Voting online ensemble) με άλλους γνωστούς ταξινομητές.....	187
Πίνακας 69. Σύγκριση της προτεινόμενης μεθόδου με άλλες που χρησιμοποιούν ως βασικό ταξινομητή δέντρα αποφάσεων.....	190
Πίνακας 70. Σύγκριση της προτεινόμενης μεθόδου με άλλες που χρησιμοποιούν ως βασικό ταξινομητή αλγόριθμο παραγωγής κανόνων	191
Πίνακας 71. Σύγκριση της προτεινόμενης μεθόδου με άλλες που χρησιμοποιούν ως βασικό ταξινομητή σκληρό αλγόριθμο.....	193
Πίνακας 72. Περιγραφή των δεδομένων.....	195
Πίνακας 73. Σύγκριση των αλγορίθμων μάθησης για την πρόβλεψη της εγκατάλειψης των σπουδών	196
Πίνακας 74. Ακρίβεια των αλγορίθμων κατά το πρώτο πείραμα	198
Πίνακας 75. Ακρίβεια των αλγορίθμων κατά το δεύτερο πείραμα.....	198
Πίνακας 76. Ταξινόμηση των μεταβλητών βάσει του μέσου κέρδους πληροφορίας για την πρόβλεψη της επίδοσης των σπουδαστών	199
Πίνακας 77. Τα αποτελέσματα των συγκρίσεων	206
Πίνακας 78. Μετρικές αξιολόγησης μεθόδων παλινδρόμησης.....	208
Πίνακας 79. Σύγκριση των αλγορίθμων για την βαθμολόγηση του αρώματος του κρασιού	208
Πίνακας 80. Σύγκριση των αλγορίθμων για την βαθμολόγηση της γεύσης του κρασιού	208
Πίνακας 81. Ακρίβεια των μεθόδων στα δεδομένα μας	209
Πίνακας 82. Σύγκριση της μεθόδου συνδυασμού τυχαίων δέντρων απόφασης με άλλες μεθόδους δημιουργίας ομάδας ταξινομητών.....	210
Πίνακας 83. Σύγκριση της μεθόδου συνδυασμού τυχαίων δέντρων απόφασης με άλλες μεθόδους.....	210
Πίνακας 84. Τα κατηγορικά δεδομένα που χρησιμοποιήσαμε στις συγκρίσεις	239
Πίνακας 85. Τα δεδομένα παλινδρόμησης που χρησιμοποιήσαμε στις συγκρίσεις	239
Πίνακας 86. Άλλα εργαλεία μάθησης.....	246
Πίνακας 87. Διευθύνσεις στο διαδίκτυο των διαθέσιμων εργαλείων εξόρυξης γνώσης	247

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1. Ο κύκλος εφαρμογής της μηχανικής μάθησης	25
Εικόνα 2. Διαγνωστικό διάγραμμα	27
Εικόνα 3. Αλγόριθμος διακριτοποίησης	32
Εικόνα 4. Επιλογή των πιο σημαντικών χαρακτηριστικών	34
Εικόνα 5. Διασταυρωμένη επικύρωση	37
Εικόνα 6. Δειγματοληπτικός έλεγχος	37
Εικόνα 7. Εκτίμηση της ακρίβειας	38
Εικόνα 8. Γραμμικά διαχωρίσιμο πρόβλημα (Linear decision boundary).....	41
Εικόνα 9. Μη γραμμικά διαχωρίσιμο πρόβλημα Non-linear decision boundary	42
Εικόνα 10. Δέντρο Απόφασης.....	43
Εικόνα 11. Ο αλγόριθμος C4.5	47
Εικόνα 12. Το φαινόμενο του υπερταϊρίσματος	48
Εικόνα 13. Κλάδεμα δέντρου απόφασης	48
Εικόνα 14. Ορθογώνια κατηγοριοποίηση του χώρου	50
Εικόνα 15. Παραγωγή κανόνων από ένα δέντρο απόφασης.....	52
Εικόνα 16. Σχηματική περιγραφή των αλγορίθμων ακολουθιακής κάλυψης.....	54
Εικόνα 17. Αλγόριθμος παραγωγής κανόνων ταξινόμησης	55
Εικόνα 18. Η γραφική περιγραφή των γενετικών τελεστών	59
Εικόνα 19. Κανόνας πρώτης τάξης (κατηγορηματικής λογικής).....	59
Εικόνα 20. Ο αλγόριθμος FOIL	61
Εικόνα 21. Η εφαρμογή του k-NN αλγόριθμου	63
Εικόνα 22. Ο διαχωρισμός που προκύπτει από τον αλγόριθμο του πλησιέστερου γείτονα (1NN) στο επίπεδο και στο χώρο	67
Εικόνα 23. Περικοπή Πλησιέστερου Γείτονα.....	68
Εικόνα 24. Σχηματική αναπαράσταση ενός νευρώνα.....	71
Εικόνα 25. Σύγκλιση του αλγορίθμου.....	71
Εικόνα 26. Η γραφική περιγραφή των βαρών στις συνάψεις	72
Εικόνα 27. Παράδειγμα Νευρωνικού Δικτύου	73
Εικόνα 28. Συναρτήσεις μεταφοράς.....	74
Εικόνα 29. Γραφική παράσταση του XOR	75
Εικόνα 30. Το εκπαιδευμένο νευρωνικό δίκτυο για το πρόβλημα XOR	75
Εικόνα 31. Γενική περιγραφή εκπαίδευσης νευρωνικού δικτύου.....	75
Εικόνα 32. Ο αλγόριθμος BP	77
Εικόνα 33. Τοπικό και ολικό ελάχιστο σφάλματος	78
Εικόνα 34. Ανταγωνιστικό δίκτυο	80
Εικόνα 35. Το δίκτυο Bayes του παραδείγματος	90
Εικόνα 36. Παράδειγμα άπληστης αναζήτησης βέλτιστου Μπεϋζιανού δικτύου	91
Εικόνα 37. Το δίκτυο του παραδείγματος.....	92
Εικόνα 38. Γραμμικά υπερεπίπεδα για τη διαχωρίσιμη περίπτωση. Τα διανύσματα υποστήριξης είναι μέσα σε κύκλο	95
Εικόνα 39. Υπερταϊρίσματος	96
Εικόνα 40. (α) Στην περίπτωση αυτή, ένα σημείο πέφτει στην περιοχή του διαχωρισμού, αλλά στην σωστή πλευρά του επιπέδου απόφασης (β) Στην περίπτωση αυτή, το σημείο πέφτει στην λάθος πλευρά του επιπέδου απόφασης	97
Εικόνα 41. Πρόβλημα OR.....	100
Εικόνα 42. Πρόβλημα XOR.....	102
Εικόνα 43. Τρία σημεία στον R^2 που διαχωρίζονται από προσανατολισμένες γραμμές.....	107
Εικόνα 44. Η γραφική αναπαράσταση του σακουλιάσματος (bagging).....	109
Εικόνα 45. Ψευδοκώδικας του σακουλιάσματος (bagging).....	109

Εικόνα 46. Ψευδοκώδικας της προώθησης (boosting)	110
Εικόνα 47. Η γραφική περιγραφή της προτεινόμενης μεθόδου Vote B&B.....	112
Εικόνα 48. Ο προτεινόμενος αλγόριθμος τοπικής προώθησης.....	119
Εικόνα 49. Ο προτεινόμενος αλγόριθμος τοπικού σακουλιάσματος.....	123
Εικόνα 50. Ο προτεινόμενος αλγόριθμος BoostFSNB	127
Εικόνα 51. Ο προτεινόμενος αλγόριθμος LogitboostNB.....	131
Εικόνα 52. Η γραφική αναπαράσταση της ψηφοφορίας των ταξινομητών.....	136
Εικόνα 53. Η κατασκευή των μετα-δεδομένων	137
Εικόνα 54. Η εκπαίδευση του μετα-ταξινομητή	138
Εικόνα 55. Σχηματική αναπαράσταση της μεθόδου	139
Εικόνα 56. Ψευδογλώσσα της προτεινόμενης μεθόδου ψηφοφορίας μέσα σε σάκο (Bagged Voting)	142
Εικόνα 57. Η γραφική αναπαράσταση της επιλεκτικής ψηφοφορίας.....	145
Εικόνα 58. Καμπύλη μάθησης	149
Εικόνα 59. Σχηματική αναπαράσταση της βελτιωμένης συσσωρευμένης γενίκευσης (Stacking').....	151
Εικόνα 60. Ο ψευδοκώδικας της βελτιωμένης συσσωρευμένης γενίκευσης.....	151
Εικόνα 61. Ο προτεινόμενος αλγόριθμος τοπικής ψηφοφορίας	154
Εικόνα 62. Η γραφική αναπαράσταση της μεθόδου τοπικής ψηφοφορίας.....	154
Εικόνα 63. Συγκρίνοντας την τοπική ψηφοφορία με την απλή ψηφοφορία.....	159
Εικόνα 64. Συγκρίνοντας την τοπική ψηφοφορία με την συσσωρευμένη γενίκευση (stacking)	159
Εικόνα 65. Ο συνδυασμός γραμμικών διαχωρίσεων (linear decision boundaries) δίνει μην γραμμική διαχώριση (non-linear boundary).....	160
Εικόνα 66. Η γραφική αναπαράσταση της μεθόδου VoteDVN.....	161
Εικόνα 67. Ο αλγόριθμος ECOC σε ψευδογλώσσα, όπου μ είναι ο αριθμός των κατηγοριών	164
Εικόνα 68. Σχηματική αναπαράσταση της υβριδικής τεχνικής επιλογής μεταβλητών	174
Εικόνα 69. Καμπύλη Ροκ (ROC curve)	177
Εικόνα 70. Η γραφική αναπαράσταση της προτεινόμενης μεθόδου συνδυασμού ευφύων πρακτόρων.....	180
Εικόνα 71. Η γραφική αναπαράσταση της προτεινόμενης μεθόδου επαυξητικής μάθησης (Voting Online Ensemble)	185
Εικόνα 72. Το πρωτότυπο διαδικτυακό εργαλείο πρόβλεψης της εγκατάλειψης των σπουδών στο ΕΑΠ.....	197
Εικόνα 73. Αρχιτεκτονική του εργαλείου	200
Εικόνα 74. Επιλέγοντας τον αλγόριθμο μάθησης.....	201
Εικόνα 75. Το παραγόμενο μοντέλο	202
Εικόνα 76. Η παρουσίαση των μεταβλητών ως προς την επιρροή τους στην μεταβλητή στόχο	202
Εικόνα 77. Ο αλγόριθμος A-priori σε ψευδογλώσσα	204
Εικόνα 78. Παραγόμενοι κανόνες συσχέτισης.....	204
Εικόνα 79. Το εργαλείο πρόβλεψης βαθμών σπουδαστή	206
Εικόνα 80. Αλγόριθμος συνδυασμού τυχαίων δέντρων απόφασης (Bagging random trees)	209
Εικόνα 81. Ομαδοποίηση δεδομένων (data clustering)	211
Εικόνα 82. Αλγοριθμικό σχήμα της ιεραρχικής ομαδοποίησης.....	214
Εικόνα 83. Λειτουργία των ιεραρχικών αλγορίθμων.....	215
Εικόνα 84. α) Ομαδοποίηση απλού-συνδέσμου. β) Ομαδοποίηση πλήρους-συνδέσμου	216
Εικόνα 85. Δύο ομοκεντρικές συστάδες	216

Εικόνα 86. Ο αλγόριθμος K-μέσων (K-means).....	217
Εικόνα 87. Ανακατανομή στοιχείων σε συστάδες και επαναληπτικός υπολογισμός κεντροειδών με τον αλγόριθμο k-means.....	217
Εικόνα 88. Ευαισθησία του K-means αλγορίθμου στην αρχική επιλογή διαμέρισης	218
Εικόνα 89. Νευρωνικό δίκτυο Kohonen	219
Εικόνα 90. Εκπαίδευση δικτύου Kohonen.....	220
Εικόνα 91. Κβαντικοποίηση του χώρου σε ένα πεπερασμένο αριθμό κελιών	221
Εικόνα 92. Αναπαράσταση των συστάδων από σημεία.....	222
Εικόνα 93. (a) συστάδες σε χώρο 2 διαστάσεων, (b) ομαδοποίηση με δένδρο, (c) ομαδοποίηση με λογικούς συζευκτικούς τελεστές.	222
Εικόνα 94. Εικόνα του Weka Explorer.	241
Εικόνα 95. Εικόνα του Weka Explorer μετά την επιλογή των δεδομένων.	242
Εικόνα 96. Εικόνα του Weka Explorer μετά την μετάβαση στην στήλη "Classify".	242
Εικόνα 97. Εικόνα του Weka μετά την εφαρμογή του αλγορίθμου J48 στο σύνολο δεδομένων	243
Εικόνα 98. Άποψη της στήλης "Cluster" του Weka Explorer.	244
Εικόνα 99. Μη επιτηρούμενη εκπαίδευση	245

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

Το κεφάλαιο αυτό παρέχει μια εισαγωγή στις έννοιες της μηχανικής μάθησης που σχετίζονται με το αντικείμενο της διατριβής. Η Μηχανική Μάθηση (M.M.) αποτελεί έναν από τους παλαιότερους τομείς έρευνας της Τεχνητής Νοημοσύνης (Dutton & Conroy, 1996). Στόχος της είναι η δημιουργία συστημάτων τα οποία θα μπορούν να βελτιώνουν της απόδοσή τους στην εργασία που επιτελούν εκμεταλλευόμενα προηγούμενη εμπειρία από την εκτέλεση της εργασίας.

Δεν θα μπορούσε βέβαια να διατυπωθεί ο ισχυρισμός ότι η επιστήμη έχει φθάσει σε σημείο να ερμηνεύσει τη διαδικασία μάθησης του ανθρώπου, πόσο μάλλον να την αποτυπώσει σε πρόγραμμα εκτελέσιμο από υπολογιστή. Αυτό που έχει επιτευχθεί όμως είναι η δημιουργία αλγορίθμων οι οποίοι μπορούν να αυτοματοποιήσουν την κατασκευή ευφών συστημάτων χρησιμοποιώντας δεδομένα εκπαίδευσης. Το γεγονός αυτό μαρτυρεί μια πληθώρα επιστημονικών αλλά και εμπορικών εφαρμογών που έχουν αναπτυχθεί τα τελευταία χρόνια, επωφελούμενα από την χρήση των παραπάνω αλγορίθμων (Witter & Frank, 2000). Σε αυτό το σημείο πρέπει να διευκρινιστεί ότι η εξόρυξη γνώσης είναι κατά κάποιο τρόπο η χρήση των μεθόδων μηχανικής μάθησης σε μεγάλες βάσεις δεδομένων. Όταν πρωτοεμφανίστηκε η μηχανική μάθηση το πρόβλημα που αντιμετώπιζαν οι ερευνητές ήταν η έλλειψη πλήθους δεδομένων εκπαίδευσης. Σήμερα με τις τεράστιες βάσεις δεδομένων στις περισσότερες περιπτώσεις το πρόβλημα έχει μετατοπιστεί στο χειρισμό αυτού του πλήθους των δεδομένων από τους αλγόριθμους μηχανικής μάθησης.

Η προβλεπτική εξόρυξη γνώσης περιλαμβάνει την ταξινόμηση (classification) και την παλινδρόμηση (regression). Και στις δύο τεχνικές, στόχος είναι η δημιουργία ενός μοντέλου που θα επιτρέπει την πρόβλεψη της τιμής μιας μεταβλητής (εξαρτημένη μεταβλητή- response –dependent-target variable) από τις γνωστές τιμές άλλων μεταβλητών (ανεξάρτητες μεταβλητές – independent- predictor variables). Αξίζει να σημειωθεί ότι ο όρος πρόβλεψη, στην εξόρυξη γνώσης, δεν έχει χρονική διάσταση όπως όταν αναφερόμαστε σε βραχυπρόθεσμες, μεσοπρόθεσμες ή μακροπρόθεσμες προβλέψεις. Η ειδοποιός διαφορά μεταξύ ταξινόμησης και παλινδρόμησης είναι ότι στην μεν ταξινόμηση η εξαρτημένη μεταβλητή είναι κατηγορική στην δε παλινδρόμηση είναι ποσοτική.

Πολύ συνοπτικά η μηχανική μάθηση χρησιμοποιείται κυρίως για τρεις σκοπούς:

- Πρόβλεψη τάσεων και συμπεριφορών. Για παράδειγμα, η μελέτη παλαιότερων σεισμικών ακολουθιών μπορεί ίσως να οδηγήσει στην πρόβλεψη, με υψηλά ποσοστά επιτυχίας, σεισμικής δραστηριότητας.
- Αναγνώριση. Οι τυποποιημένες μορφές ανάμεσα στα δεδομένα μπορούν να χρησιμοποιηθούν για να αποκαλύψουν την ύπαρξη ενός γεγονότος, μια δραστηριότητας. Για παράδειγμα οι εισβολείς στη προσπάθεια να σπάσουν ένα σύστημα ασφαλείας μπορούν να αναγνωριστούν από τα προγράμματα που εκτέλεσαν, τα αρχεία που προσπέλασαν και τον χρόνο που απασχόλησαν την CPU.
- Ταξινόμηση. Η εξόρυξη γνώσης μπορεί να διαχωρίσει έτσι τα δεδομένα ώστε να προκύψουν διαφορετικές κλάσεις ή κατηγορίες βάσει κάποιων παραμέτρων. Για παράδειγμα οι πελάτες ενός καταστήματος μπορούν να χωριστούν σε κατηγορίες, όπως φίλοι-των-εκπτώσεων, παρορμητικοί, πιστοί-κανονικοί, και σπάνιοι πελάτες. Αυτή η κατηγοριοποίηση μπορεί να χρησιμοποιηθεί στην ανάλυση των πωλήσεων ώστε να μπορεί για παράδειγμα ο μάνατζερ να λάβει αποφάσεις για να προσελκύσει σε μεγαλύτερο βαθμό κάποια από τις παραπάνω κατηγορίες.

Το πρώτο στάδιο της σχεδίασης ενός συστήματος εξόρυξης γνώσης (ή μηχανικής μάθησης) συνίσταται στον προσδιορισμό της γνώσης που θα χρησιμοποιηθεί κατά την εκπαίδευσή του.

Ιδιαίτερη σημασία στην ποιότητα του σώματος εκπαίδευσης αποδίδεται στην κατανομή των παραδειγμάτων από τα οποία θα αντληθεί η γνώση από το σύστημα Μ.Μ., καθώς θα πρέπει να είναι όμοια με την αντίστοιχη κατανομή των περιπτώσεων που θα χρησιμοποιηθούν κατά τον έλεγχο της απόδοσης, ακόμα δε περισσότερο κατά τη λειτουργία του.

Η πλέον συνηθισμένη αναπαράσταση του σώματος εκπαίδευσης είναι αυτή που παρέχει το μοντέλο του διανυσματικού χώρου (Kleinberg, 2000). Σύμφωνα με αυτό το μοντέλο, οι οντότητες αναπαρίστανται ως διανύσματα, τα στοιχεία των οποίων αναπαριστούν τα χαρακτηριστικά (features ή attributes) της οντότητας που έχουν επιλεγεί ως σχετικά για το συγκεκριμένο πρόβλημα. Τα χαρακτηριστικά μπορούν να παίρνουν συμβολικές ή αριθμητικές τιμές. Για παράδειγμα, αν οι οντότητες αντιπροσωπεύουν μανιτάρια και το ζητούμενο είναι το αν αυτά είναι δηλητηριώδη, το διάνυσμα που αντιστοιχεί σε κάθε μανιτάρι είναι δυνατόν να περιλαμβάνει χαρακτηριστικά όπως την οσμή του, την προέλευσή του, το βάρος του κ.α. Οι τιμές της συνάρτησης-στόχου μπορεί να είναι πρακτικά οτιδήποτε: αριθμητικές ή συμβολικές, διακριτές ή συνεχείς, βαθμωτές ή διανυσματικές, κ.ο.κ.

Επόμενο στάδιο στη σχεδίαση του συστήματος αποτελεί ο τρόπος με τον οποίο θα διαχειριστεί το τελευταίο τη γνώση που αποκομίζει κατά την εκπαίδευση, με σκοπό την αποδοτικότερη λειτουργία του. Υιοθετώντας το μοντέλο της διανυσματικής αναπαράστασης, αναζητούμε ουσιαστικά μια συνάρτηση f' η οποία προσεγγίζει όσο το δυνατόν περισσότερο μια ιδανική συνάρτηση με την οποία δύναται να μοντελοποιηθεί το πρόβλημα, τη συνάρτηση στόχο (target function) f . Οι δύο συναρτήσεις έχουν ελεύθερη μεταβλητή ένα τυχαίο διάνυσμα x , πεδίο ορισμού το χώρο των στιγμιότυπων, το δε σύνολο τιμών καθορίζεται από την εκάστοτε εφαρμογή. Έτσι, η επίλυση ενός προβλήματος Μ.Μ. ανάγεται στην επίλυση ενός προβλήματος προσέγγισης των τιμών μιας συνάρτησης (function approximation). Ο παραπάνω ορισμός που δόθηκε για τη μηχανική μάθηση αναφέρεται στην πραγματικότητα στην περίπτωση της μάθησης υπό επίβλεψη (supervised learning), όπως λέγεται, υπό την έννοια πως η διαδικασία της μάθησης μπορεί να θεωρηθεί ότι επιβλέπεται από ειδικούς που γνωρίζουν την τιμή της συνάρτησης-στόχου.

Με άλλα λόγια, η διαδικασία της επιτηρούμενης μάθησης «καθοδηγείται» από τα στιγμιότυπα εκπαίδευσης υπό την έννοια ότι υποδεικνύουν στον αλγόριθμο μάθησης την τιμή της συνάρτησης στόχου που επιδιώκει να προσεγγίσει. Για παράδειγμα ο Πίνακας 1 παρουσιάζει πιθανά δεδομένα εκπαίδευσης για ένα τραπεζικό σύστημα αξιολόγησης πελατών για δανειοδότηση.

Πελάτης	Οφειλές	Εισόδημα	Παντρεμένος/η	Πελάτης;
1	Υψηλές	Υψηλό	Ναι	Καλός
2	Χαμηλές	Υψηλό	Όχι	Κακός
3	Χαμηλές	Υψηλό	Ναι	Καλός
4	Υψηλές	Χαμηλό	Ναι	Κακός
5	Χαμηλές	Χαμηλό	Ναι	Κακός

Πίνακας 1. Δεδομένα εκπαίδευσης για ένα τραπεζικό σύστημα αξιολόγησης πελατών για δανειοδότηση

Πιο φορμαλιστικά, δοθέντων:

- ενός προβλήματος με N κλάσεις: $C_1, C_2, \dots, C_N$ όπου κάθε στιγμιότυπο του προβλήματος έχει m χαρακτηριστικά (μεταβλητές): $A_1, A_2, \dots, A_m$

- και ενός συνόλου στιγμιότυπων του προβλήματος για τα οποία γνωρίζουμε εκ των προτέρων σε ποια κλάση ανήκουν - το σύνολο αυτό είναι γνωστό ως σύνολο εκπαιδευτικών στιγμιότυπων (training set)

το ζητούμενο είναι

- η δημιουργία ενός μοντέλου για την ταξινόμηση νέων άγνωστων στιγμιότυπων του προβλήματος. Με τον όρο ταξινόμηση εννοούμε την τοποθέτηση ενός στιγμιότυπου σε μία από τις προκαθορισμένες κλάσεις του προβλήματος.

Η επιτυχής έκβαση της ταξινόμησης εξαρτάται από δύο βασικούς παράγοντες:

- το σαφή καθορισμό των κλάσεων του προβλήματος - οι κλάσεις είναι προκαθορισμένες και δεν μεταβάλλονται κατά τη διάρκεια της ταξινόμησης.
- την «ποιότητα» του συνόλου των στιγμιότυπων εκπαίδευσης – τα στιγμιότυπα αυτά θα πρέπει να είναι αντιπροσωπευτικά του προβλήματος.

Αν υποθέσουμε την απλοϊκή περίπτωση ότι έχουμε στη διάθεσή μας μία συλλογή δεδομένων εκπαίδευσης η οποία περιέχει n παραδείγματα, που το κάθε ένα είναι ένα διάνυσμα m αριθμών τότε αυτά μπορούν να θεωρηθούν ως σημεία σε ένα m -διάστατο χώρο. Ένας απλός τρόπος να κατασκευαστεί ένας δυαδικός ταξινομητής (binary classifier) είναι η δημιουργία ενός υπερεπιπέδου, το οποίο θα διαχωρίζει τα μέλη της μίας κατηγορίας από εκείνα της άλλης.

Ολοκληρώνοντας τη σύντομη αυτή αναφορά στις θεμελιώδεις έννοιες της Μ.Μ., κρίνεται σκόπιμη η παρουσίαση ενός συνόλου παραγόντων καθοριστικής σημασίας για τη σχεδίαση ενός αποδοτικού συστήματος επιτηρούμενης Μ.Μ. Από τους πλέον σημαντικούς θεωρείται η ικανότητα γενίκευσης (generalization ability) που χαρακτηρίζει ένα τέτοιο σύστημα, η ικανότητά του δηλαδή να χρησιμοποιεί την εμπειρία που απέκτησε από τα στιγμιότυπα εκπαίδευσης όσο το δυνατόν αποτελεσματικότερα επί του συνόλου των άγνωστων στιγμιότυπων ελέγχου. Σε μια πρώτη προσέγγιση ακούγεται αρκετά λογικό πως κάθε υποψήφιο προς επιλογή μοντέλο θα πρέπει να επαληθεύεται από όλα τα στιγμιότυπα εκπαίδευσης, ή όπως λέγεται, το μοντέλο θα πρέπει να είναι συνεπές (consistent). Στην πράξη, πέρα από το γεγονός πως δεν είναι βέβαιο ότι υπάρχει ακριβώς ένα τέτοιο μοντέλο, ακόμα κι αν υπάρχει και βρεθεί, δεν είναι σίγουρα και η καλύτερη λύση. Η αιτία είναι το φαινόμενο του υπερταυρίσματος (overfitting), το οποίο θα μπορούσε να αποδοθεί ως το υπερβολικό ταίριασμα με τα δεδομένα εκπαίδευσης.

Μία υπόθεση η λέγεται πως υπερταυρίζει (overfits) με τα δεδομένα εκπαίδευσης αν υπάρχει μια άλλη υπόθεση h' τέτοια ώστε η h να έχει μικρότερο σφάλμα από την h' για τα δεδομένα εκπαίδευσης, αλλά η h' να έχει μικρότερο σφάλμα από την h για τη συνολική κατανομή των στιγμιότυπων. Η h' δηλαδή είναι καλύτερη προσέγγιση του πραγματικού μοντέλου από την h . Οι κύριοι λόγοι εμφάνισης του υπερταυρίσματος είναι οι εξής:

- Ο μεγάλος αριθμός παραμέτρων του μοντέλου, ή πιο γενικά η ικανότητα του αλγορίθμου μάθησης να κατασκευάζει ιδιαίτερα πολύπλοκα μοντέλα.
- Η μη κατάλληλη επιλογή των χαρακτηριστικών αναπαράστασης, καθώς και ο θόρυβος στα δεδομένα εκπαίδευσης, δηλαδή τα τυχαία λάθη που είναι δυνατόν να περιέχονται στα δεδομένα. Αν και θα θέλαμε να είχαμε απολύτως αξιόπιστα δεδομένα τα οποία να χρησιμοποιούσαμε για την κατασκευή του ταξινομητή, στην πράξη αυτό δεν είναι πάντα εφικτό. Για παράδειγμα, μπορεί τα δεδομένα να είναι σήματα από βιντεοκάμερες ή μικρόφωνα αλλοιωμένα από τυχαίο ηλεκτρομαγνητικό θόρυβο, ή να προέρχονται από ανακριβείς πειραματικές μετρήσεις σε μη ελεγχόμενο περιβάλλον, όπως αυτές που γίνονται στο διάστημα. Αξίζει να σημειωθεί πως η πιο κοινή πηγή θορύβου είναι ο “ανθρώπινος παράγοντας”, π.χ. στην εισαγωγή των δεδομένων. Είναι επομένως λογικό πως ένας ταξινομητής προσαρμοσμένος απόλυτα ή πολύ κοντά στα (θορυβώδη) δεδομένα

εκπαίδευσης, δεν αναμένεται να διατηρήσει την υψηλή του απόδοση σε νέα μη παρατηρηθέντα δεδομένα, ή όπως λέγεται δεν θα έχει μεγάλη ακρίβεια γενίκευσης (generalization accuracy).

Το υπερταΐριασμα είναι μια σημαντική πρακτική δυσκολία για πολλούς αλγορίθμους μάθησης. Για τη μετρίασή του έχουν επινοηθεί μέθοδοι, τόσο προσαρμοσμένες σε καθέναν από αυτούς, όσο και ανεξάρτητες αλγορίθμου. Βασικός οδηγός στην αποφυγή του είναι η “αρχή” του ξυραφιού του Occam (Occam’s Razor): “μεταξύ όλων των ικανοποιητικών λύσεων, προτιμήστε την απλούστερη”.

Μία οπτική γωνία απ’ την οποία μπορεί κανείς να δει τη μηχανική μάθηση είναι αυτή της αναζήτησης, σε ένα πολύ μεγάλο χώρο δυνατών υποθέσεων, μιας υπόθεσης που ταιριάζει “αρκετά καλά” με τα δεδομένα εκπαίδευσης και την τυχόν εκ των προτέρων (a priori) γνώση. Οι διάφοροι επαγωγικοί αλγόριθμοι μάθησης (inducers, για συντομία EAM) που έχουν αναπτυχθεί έως σήμερα διαφέρουν ως προς την υποκείμενη αναπαράσταση του χώρου των δυνατών υποθέσεων, και κατά συνέπεια και του τρόπου που οργανώνουν την αναζήτηση σε αυτό το χώρο. Μερικά παραδείγματα αναπαραστάσεων είναι οι κανόνες ταξινόμησης, τα δέντρα απόφασης (decision trees), τα τεχνητά νευρωνικά δίκτυα (artificial neural networks), κ.α. Διαφορετικές αναπαραστάσεις είναι κατάλληλες για τη μάθηση διαφορετικών ειδών συναρτήσεων-στόχων. Για κάθε μια από αυτές τις αναπαραστάσεις, ο αντίστοιχος EAM εκμεταλλεύεται τη διαφορετική υποκείμενη δομή για να οργανώσει την αναζήτηση στο χώρο των υποθέσεων. Μία θεμελιώδης ιδιότητα που χαρακτηρίζει κάθε EAM είναι η επαγωγική προδιάθεση ή κλίση του (inductive bias). Κάθε EAM απαιτεί κάποιου είδους προδιάθεση για να μπορέσει να γενικεύσει πέρα από τα παρατηρηθέντα δεδομένα. Διαφορετικά, ένας πλήρως αμερόληπτος αλγόριθμος (bias-free learner) ο οποίος δεν κάνει καμιά υπόθεση σχετικά με την ταυτότητα της συνάρτησης-στόχου, δεν έχει κανένα λογικό έρεισμα για να αποφασίσει την τιμή κάποιου άγνωστου στιγμιότυπου.

Με διαφορετική διατύπωση, η επαγωγική κλίση ενός EAM L είναι ένα ελάχιστο σύνολο από υποθέσεις, οι οποίες σε συνδυασμό με ένα δοθέν σύνολο εκπαίδευσης και ένα άγνωστο στιγμιότυπο προς κατάταξη, μπορούν να οδηγήσουν παραγωγικά (deductively) στην πρόβλεψη που δίνει ο L για αυτό το στιγμιότυπο. Οι υποθέσεις αυτές άλλοτε περιορίζουν το χώρο των υπό θεώρηση μοντέλων (restriction bias ή language bias), άλλοτε επιβάλλουν μια συγκεκριμένη στρατηγική αναζήτησης στο χώρο αυτό, επιβάλλοντας έτσι την προτίμηση κάποιων υποθέσεων ως προς άλλες (search ή preference bias) και άλλοτε συνδυάζουν και τα δύο.

Αντίθετα με την επιτηρούμενη μηχανική μάθηση, στην περίπτωση της μάθησης χωρίς επίβλεψη, ο αλγόριθμος καλείται να ανακαλύψει δομικούς σχηματισμούς στα στιγμιότυπα εκπαίδευσης, βάσει κριτηρίων τα οποία δεν του παρέχονται άμεσα. Παραδείγματα προβλημάτων που εντοπίζονται σε αυτήν την περιοχή είναι: η ομαδοποίηση οντοτήτων (clustering), όπου επιχειρείται ομαδοποίηση των στιγμιότυπων σε κλάσεις τις οποίες θα συμπεράνει ο αλγόριθμος, στηριζόμενος στην ομοιότητα των μελών κάθε ομάδας.

1.1 Λόγοι που κάνουν την εξόρυξη γνώσης με μηχανική μάθηση όλο και πιο δημοφιλή

Αρκετοί είναι οι λόγοι που κάνουν την εξόρυξη γνώσης με μηχανική μάθηση όλο και πιο δημοφιλή. Ανάμεσα σε αυτούς τους λόγους είναι ο αυξανόμενος όγκος δεδομένων, οι έμφυτοι περιορισμοί της δυνατότητας ανθρώπινης ανάλυσης, το χαμηλότερο κόστος για την εκμάθηση μιας μηχανής από την εκπαίδευση ενός συνόλου ειδικών και τέλος η ανάγκη των επιχειρήσεων για επιχειρηματική νοημοσύνη.

Τα τελευταία χρόνια, οι δυνατότητες μας να παράγουμε και να συλλέγουμε δεδομένα έχουν αυξηθεί σημαντικά. Η ευρεία χρήση των υπολογιστών στις συναλλαγές σε όλους τους τομείς της σύγχρονης κοινωνίας (στο χώρο των επιχειρήσεων, της βιομηχανίας, των επιστημών) καθώς και τα πολλαπλά πλεονεκτήματα που παρέχουν τα διάφορα εργαλεία συλλογής και επεξεργασίας δεδομένων έχουν οδηγήσει στη συγκέντρωση μεγάλου όγκου πληροφορίας.

Ο βασικός λόγος για την ανάγκη των αυτοματοποιημένων συγκροτημάτων ηλεκτρονικών υπολογιστών για την ευφυή ανάλυση δεδομένων είναι ο τεράστιος όγκος των υπαρκτών και μελλοντικών δεδομένων που απαιτούν περαιτέρω επεξεργασία. Ο όγκος των δεδομένων που συσσωρεύονται κάθε ημέρα από διάφορες επιχειρήσεις, επιστημονικές και κυβερνητικές οργανώσεις σε όλο τον κόσμο αυξάνεται με εντυπωσιακό ρυθμό. Σύμφωνα με έρευνες, μόνο οι επιστημονικές οργανώσεις επεξεργάζονται και αποθηκεύουν κάθε ημέρα για 1 TB (terabyte) νέων δεδομένων. Γίνεται αδύνατο για τους αναλυτές να εξάγουν χρήσιμη πληροφορία από τόσο μεγάλες βάσεις δεδομένων.

Δύο άλλα προβλήματα που εμφανίζονται όταν αναλυτές επεξεργάζονται δεδομένα είναι η ανεπάρκεια του ανθρώπινου εγκεφάλου να επεξεργάζεται σύνθετες εξαρτήσεις στα δεδομένα, και η έλλειψη αντικειμενικότητας σε μια τέτοια ανάλυση. Ένας εμπειρογνώμονας επηρεάζεται σχεδόν πάντοτε από την εμπειρία του στην ανάλυση άλλων συστημάτων. Μερικές φορές αυτό βοηθά, μερικές φορές βλάπτει, αλλά είναι σχεδόν αδύνατο να ξεφύγει από αυτό το γεγονός.

Ένα πρόσθετο όφελος από τη χρήση αυτοματοποιημένων συστημάτων εξόρυξης δεδομένων είναι ότι αυτή η διαδικασία έχει πολύ χαμηλότερο κόστος από το να προσλάβει μια εταιρία έναν ιδιαίτερα εκπαιδευμένο (και πολύ καλά αμειβόμενο) αριθμό από εμπειρογνώμονες. Ενώ η εξόρυξη γνώσης δεν εξαλείφει εντελώς την ανθρώπινη συμμετοχή στη λύση ενός προβλήματος, απλοποιεί σημαντικά την εργασία αυτή.

Στον επιχειρησιακό κόσμο, η εξόρυξη γνώσης παρέχει μια πλήρως καινούρια ικανότητα, τη δυνατότητα να βελτιστοποιηθεί η λήψη αποφάσεων χρησιμοποιώντας τις αυτοματοποιημένες μεθόδους που μαθαίνουν από παρελθοντικές ενέργειες. Εάν μια επιχείρηση ήξερε τι προϊόντα ενδιαφέρουν ένα πελάτη τότε θα μπορούσε να λάβει μέτρα για να εξασφαλίσει ότι ο πελάτης θα πραγματοποιήσει την αγορά από αυτή την επιχείρηση. Για παράδειγμα, η επιχείρηση θα μπορούσε να στείλει με το ταχυδρομείο ειδικές προσφορές για το συγκεκριμένο είδος προϊόντος στο συγκεκριμένο πελάτη (Ling & Li, 1998). Χωρίς τη προσέγγιση αυτή, η επιχείρηση θα σπαταλούσε χρήματα στέλνοντας προσφορές σε πελάτες που δεν θα ήθελαν το συγκεκριμένο προϊόν.

1.2 Εφαρμογές της μηχανικής μάθησης και της εξόρυξης γνώσης

Η μηχανική μάθηση χρησιμοποιείται σήμερα σε πολλά και διαφορετικά πεδία εφαρμογών όπως η αναγνώριση προτύπου (pattern recognition), ταυτοποίηση (identification), ταξινόμηση (classification) κ.λπ. Εν συνεχεία, αναφέρονται ενδεικτικά μερικές περιοχές εφαρμογών (Michalski και άλλοι, 1998):

- Αεροπορία: Υψηλής απόδοσης αυτόματοι πιλότοι αεροπλάνων, προσομοιωτές πτήσης, συστήματα αυτομάτου ελέγχου αεροπλάνων, συστήματα ανίχνευσης βλαβών.
- Αυτοκίνηση: Αυτοκινούμενα συστήματα αυτόματης πλοήγησης.
- Τραπεζικές εφαρμογές: Συστήματα αξιολόγησης αιτήσεων δανειοδότησης.
- Άμυνα: Ανίχνευση στόχων, σόναρ, ραντάρ, αναγνώριση σήματος / εικόνας.
- Ηλεκτρονική: Πρόβλεψη ακολουθίας κωδίκων, διάγνωση βλαβών ολοκληρωμένων κυκλωμάτων, μηχανική όραση.
- Οικονομία: Οικονομική ανάλυση, πρόβλεψη τιμών συναλλάγματος.
- Κοινωνική ασφάλιση: Αξιολόγηση εφαρμοζόμενης πολιτικής, βελτιστοποίηση παραγωγής.

- Βιομηχανία: Βιομηχανικός έλεγχος διεργασιών, συστήματα ποιοτικού ελέγχου, διάγνωση βλαβών διεργασιών και μηχανών.
- Ιατρική: Ανάλυση καρκινικών κυττάρων, ανάλυση Ηλεκτροεγκεφαλογράφηματος και Ηλεκτροκαρδιογραφήματος.
- Γεωλογικές έρευνες: Εντοπισμός πετρελαίου και φυσικού αερίου.
- Ρομποτική: Έλεγχος τροχιάς και σύστημα όρασης ρομπότ.
- Επεξεργασία φωνής: Αναγνώριση φωνής, σύνθεση φωνής από κείμενο.
- Χρηματιστηριακές εφαρμογές: Ανάλυση αγοράς, πρόβλεψη τιμών μετοχών.
- Τηλεπικοινωνίες: Αυτοματοποιημένες υπηρεσίες πληροφοριών, μετάφραση πραγματικού χρόνου.
- Μεταφορές: Συστήματα διάγνωσης βλαβών φρένων, συστήματα δρομολόγησης.

ΚΕΦΑΛΑΙΟ 2. Η διαδικασία επίλυσης ενός προβλήματος Μηχανικής Μάθησης

Στην ενότητα αυτή περιγράφεται η πορεία που ακολουθείται κατά τη σχεδίαση ενός συστήματος που χρησιμοποιεί τεχνικές μηχανικής μάθησης. Ο σκοπός αυτής της ενότητας είναι αφ' ενός μεν να καταδειχθεί η θέση της μηχανικής μάθησης στο γενικότερο πεδίο της ταξινόμησης/πρόβλεψης και αφ' ετέρου να παρουσιαστούν οι σημαντικότερες επιλογές που υπάρχουν σε κάθε βήμα της διαδικασίας σχεδίασης.

Η διαδικασία λύσεως ενός προβλήματος Μηχανικής Μάθησης είναι μία αλληλεπιδραστική και επαναληπτική διαδικασία, η οποία περιλαμβάνει πλήθος βημάτων στα οποία θα πρέπει να ληφθούν αποφάσεις από τον χρήστη. Στην εργασία μας [8] αναλύονται τα βασικότερα βήματα της διαδικασίας αυτής. Συνοπτικά, τα βασικά στάδια της σχεδίασης ενός συστήματος Μ.Μ. φαίνονται στην Εικόνα 1.

Αρχικά, κατανοούμε όλες τις πλευρές του συγκεκριμένου προβλήματος συμπεριλαμβανομένης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα και τους στόχους των τελικών χρηστών για την αρχική επιλογή των δεδομένων εκπαίδευσης. Η κατανόηση όλων των πτυχών του προβλήματος είναι σημαντικό βήμα για τη περαιτέρω συνέχιση της όλης διαδικασίας.

Εν συνεχεία, σημαντικός είναι ο ρόλος της προετοιμασίας των δεδομένων που αφορά την εξάλειψη των απλών σφαλμάτων και την ορθή αποκατάσταση των εγγραφών (στιγμιότυπων). Η διαδικασία αυτή είναι πολύ σημαντική διότι ασυνεπή και λανθασμένα δεδομένα οδηγούν σε ανακριβή αποτελέσματα. Μια περίπτωση που εμφανίζεται συχνά η ανάγκη διόρθωσης των δεδομένων είναι τα πεδία που περιέχουν κείμενο. Ένα όνομα πόλης, για παράδειγμα, μπορεί να είναι γραμμένο με ποικίλους τρόπους (π.χ. διάφορες συντομεύσεις). Ένα άλλο παράδειγμα στο οποίο εμφανίζεται η ανάγκη μετατροπής των δεδομένων, είναι τα δεδομένα πωλήσεων που πραγματοποιούνται από πολυεθνική εταιρία σε διηπειρωτικό επίπεδο. Τα Ευρωπαϊκά δεδομένα θα είναι εκφρασμένα σε μονάδες όπως μέτρα, κιλά, λίτρα, ευρώ, ενώ τα Αμερικανικά σε ίντσες, λίβρες και δολάρια.

Εικόνα 1. Ο κύκλος εφαρμογής της μηχανικής μάθησης

Στις επόμενες υποενότητες αναλύουμε τα βήματα που παρουσιάστηκαν στην Εικόνα 1.

2.1 Επιλογή δεδομένων/Αξιολόγηση των δεδομένων

Επιλέγουμε το σύνολο των δεδομένων ή επικεντρώνουμε το ενδιαφέρον μας στις μεταβλητές ή τα δεδομένα στα οποία πρόκειται να γίνει η όλη επεξεργασία και να εξαχθεί η γνώση και τα τελικά συμπεράσματα. Όπως είπαμε κάθε στοιχείο αποτελεί ένα πολυδιάστατο διάνυσμα, όπου κάθε διάσταση εκφράζει και μια μεταβλητή. Ο Πίνακας 2 παρουσιάζει τις κατηγορίες των μεταβλητών.

Ποσοτικές μεταβλητές

- i. Συνεχείς (π.χ.: το βάρος ενός ανθρώπου)
- ii. Διακριτές (π.χ.: το πλήθος των μαθητών)

Ποιοτικές μεταβλητές

- i. Κατηγορικές (π.χ.: η οικογενειακή εξέταση)
- ii. Διάταξης (π.χ.: η επίδοση ενός μαθητή: Καλώς/Λίαν καλώς/ Άριστα)

Πίνακας 2. Ταξινόμηση Μεταβλητών

Ο καθορισμός και η επιλογή του συνόλου των μεταβλητών που περιγράφουν τα δεδομένα αποτελεί κρίσιμο βήμα στη διαδικασία μάθησης αφού σε μεγάλο βαθμό επηρεάζει την αναπαράσταση των δεδομένων και τα παραγόμενα αποτελέσματα,

2.1.1 Καθαρισμός και επεξεργασία των δεδομένων

Ιδιαίτερη προσοχή θα πρέπει επίσης να δοθεί στην ορθότητα των δεδομένων που χρησιμοποιούνται για την εκπαίδευση, καθώς δεν είναι σπάνια η περίπτωση εισχώρησης σφαλμάτων στις τιμές των χαρακτηριστικών κάποιων στιγμιотύπων (Hernandez & Stolfo, 1998). Το φαινόμενο αυτό ονομάζεται θόρυβος (noise) και συναντάται κατά τη συλλογή δεδομένων από πειραματικές μετρήσεις, αλλά και γενικότερα, όπου επεμβαίνει ο ανθρώπινος παράγοντας στη δημιουργία δεδομένων εκπαίδευσης. Όπως είναι αναμενόμενο, η εκτεταμένη παρουσία θορύβου οδηγεί στον αποπροσανατολισμό του αλγορίθμου μάθησης και στην επιλογή μιας υποβέλτιστης υπόθεσης, ελάχιστα συνεπούς με τα δεδομένα ελέγχου (Bruha & Franek, 1996). Αυτό ωστόσο δε συμβαίνει και στην περίπτωση που παρουσιάζεται θόρυβος της ίδιας μορφής και στα δεδομένα αξιολόγησης, οπότε χρειάζεται να μοντελοποιηθεί και αυτός, καθώς ανάγεται πλέον σε «χαρακτηριστικό» του συγκεκριμένου προβλήματος μάθησης (Jerzy και άλλοι, 2000).

Για την αντιμετώπιση του προβλήματος ανίχνευσης σφαλμάτων υπάρχουν δύο θεμελιώδεις προσεγγίσεις (Zhu & Wu, 2004):

Τύπος 1 – Καθορισμός σφαλμάτων χωρίς την προγενέστερη γνώση των στοιχείων.

Η προσέγγιση αυτή επεξεργάζεται τα στοιχεία υποθέτοντας ότι ανήκουν σε μια στατική κατανομή. Είναι η απλή προσέγγιση φιλτραρίσματος κατά την οποία εκείνες οι τιμές που φαίνονται ύποπτες σε σχέση με μια συγκεκριμένη κατανομή πιθανότητας αφαιρούνται. Για παράδειγμα για μια κανονική κατανομή με μέσο όρο 5 και τυπική απόκλιση 3, η τιμή 15 σε μία εγγραφή θεωρείται ύποπτη και αφαιρείται. Ο Πίνακας 3 παρουσιάζει παραδείγματα πως κάποια μεταδεδομένα μπορούν να βοηθήσουν στην ανίχνευση διάφορων πιθανών προβλημάτων στα δεδομένα μας.

Προβλήματα	Μεταδεδομένα	Παραδείγματα/ευρετικά
Απίθανες τιμές	πληθάριθμος	Για παράδειγμα, αν το φύλο εμφανίζει > 2 πιθανές τιμές τότε υπάρχει πρόβλημα
	μέγιστο, ελάχιστο	Δεν πρέπει οι τιμές μιας μεταβλητής να υπερβαίνουν ένα άνω ή κάτω όριο (π.χ στον αριθμό των έκνων δεν μπορώ να έχω αρνητικό αριθμό)
	Τυπική απόκλιση	Η μεγάλη απόκλιση μεταξύ της μέσης τιμής και της εξεταζόμενης τιμής μπορεί να είναι ένδειξη θορύβου
Ορθογραφικά λάθη	Τιμές των μεταβλητών	Ταξινομούμε τις μεταβλητές ως προς τις τιμές τους και ανιχνεύουμε λάθη κατά την εισαγωγή των δεδομένων

Πίνακας 3. Έλεγχος για σφάλματα στα δεδομένα

Οι ακραίες τιμές/θόρυβος (outliers/noise) μπορούν επίσης να αναγνωρισθούν με τη χρησιμοποίηση διαγνωστικών διαγραμμάτων (boxplots) (Batista & Monard, 2003). Το παρακάτω σχήμα παρουσιάζει τα διαγράμματα για την πρώτη μεταβλητή του συνόλου δεδομένων iris⁵.

Εικόνα 2. Διαγνωστικό διάγραμμα

Κάθε κιβώτιο (boxplot) αντιπροσωπεύει τη σειρά των τιμών των χαρακτηριστικών για κάθε μια από τις τρεις κατηγορίες (κλάσεις). Η διάμεσος για τα δεδομένα υποδεικνύεται από τη μαύρη κεντρική γραμμή (μέσα στα παραλληλόγραμμα), και τα πρώτα και τρίτα τεταρτημόρια είναι οι άκρες των παραλληλόγραμμων. Οι μαύρες γραμμές επάνω από και κάτω από τα παραλληλόγραμμα αντιπροσωπεύουν τόση απόσταση όση είναι η διαφορά του πρώτου από

⁵ Πολύ γνωστό σύνολο δεδομένων. Μια αναλυτικότερη περιγραφή του παρουσιάζεται στο Παράρτημα Α

το τρίτο τεταρτημόριο επί 1.5. Τα σημεία έξω από τις μαύρες γραμμές αντιπροσωπεύουν πιθανό θόρυβο.

Παρ' όλο που τα διαγνωστικά διαγράμματα δεν αποκαλύπτουν πάντοτε τα αίτια εμφάνισης ανεπιθύμητων τιμών, απλοποιούν σε ένα βαθμό το πρόβλημα, εστιάζοντας σε μία ορισμένη ομάδα μεταβλητών. Σε αυτό το σημείο πρέπει επίσης να τονιστεί ότι πριν από την εκπαίδευση του μοντέλου, τα δεδομένα θα πρέπει, επίσης, να ελεγχθούν για να διαπιστωθεί αν υπάρχουν μεταβλητές με σταθερές τιμές. Όπως είναι φυσικό είναι απαραίτητο η μεταβλητότητα των μεταβλητών, που λαμβάνονται υπόψη στην ανάλυση, να είναι διαφορετική του μηδενός. Αυτό γίνεται υπολογίζοντας την τυπική απόκλιση ή τους συντελεστές συσχέτισης των μεταβλητών. Οι μεταβλητές με μηδενική απόκλιση πρέπει να παραλειφθούν από τα δεδομένα εκπαίδευσης.

Τύπος 2 - Αυτή η προσέγγιση είναι ανάλογη με την εποπτευμένη μάθηση. Τα παραδείγματα που ταξινομούνται λανθασμένα από ένα πλήθος ταξινομητών εκλαμβάνονται ως σφάλματα ανά κατηγορία.

Τα συμπεράσματα από τα πειράματα των (Zhu & Wu, 2004), όσον αφορά την επίδραση του θορύβου, μπορούν να συνοψιστούν ως εξής:

1. Η εξάλειψη των περιπτώσεων που περιέχουν θόρυβο κατηγορίας (θόρυβο στην εξαρτημένη μεταβλητή) ενισχύει την ακρίβεια ταξινόμησης.
2. Σε σύγκριση με το θόρυβο κατηγορίας, ο θόρυβος στις ανεξάρτητες μεταβλητές είναι συνήθως λιγότερο επιβλαβής, αλλά και αυτός μπορεί να προκαλέσει σοβαρά προβλήματα στην εκπαίδευση των αλγορίθμων. Η εξάλειψη του θορύβου θα ενισχύσει στις περισσότερες περιπτώσεις την ακρίβεια των ταξινομητών.
3. Σε σύγκριση με το χειρισμό θορύβου στο σύνολο εκπαίδευσης, ο καθαρισμός του θορύβου από τα δεδομένα ελέγχου/ταξινόμησης συνήθως δίνει περισσότερα οφέλη (από άποψη ακρίβειας ταξινόμησης), ακόμα κι αν ο ταξινομητής έχει μάθει από ένα αλλοιωμένο από θόρυβο σύνολο εκπαίδευσης.
4. Στις περισσότερες περιπτώσεις, ο θόρυβος ανά μεταβλητή επηρεάζει διαφορετικά την απόδοση των συστημάτων. Όσο υψηλότερος είναι ο συσχετισμός μεταξύ της μεταβλητής και της κατηγορίας, τόσο πιο αρνητικός είναι ο αντίκτυπος στην ακρίβεια ταξινόμησης.
5. Για να προσδιορίσουμε και να διορθώσουμε το θόρυβο, μπορούμε να εκπαιδεύσουμε μερικούς αλγορίθμους και να κατασκευάσουμε ένα φίλτρο θορύβου (Brodley & Friedl, 1999).

Παρόμοιο φαινόμενο είναι και αυτό της απουσίας τιμών (missing values), περιπτώσεων δηλαδή όπου οι τιμές ορισμένων χαρακτηριστικών δεν αναφέρονται για κάποια από τα στιγμιότυπα.

Έστω ότι από τον παρακάτω πίνακα δεδομένων λείπουν αρχικά κάποιες τιμές (αναφέρονται ως undefined):

	Diploma	Experience	French	Reference	Decision
1	MBA	Undefined	Yes	Undefined	Accept
2	MBA	Low	Yes	Neutral	Reject
3	Undefined	Low	Yes	Good	Reject
4	MSc	High	Undefined	Neutral	Accept
5	MSc	Medium	Yes	Neutral	Reject
6	MSc	High	Yes	Undefined	Accept
7	Undefined	High	No	Good	Accept
8	MCE	Low	No	Excellent	Reject

Πίνακας 4. Ο αρχικός πίνακας δεδομένων με ελλιπή δεδομένα (απουσία τιμών)

Η πιο απλή λύση στο πρόβλημα της απουσίας τιμών είναι η διαγραφή των στιγμιοτύπων που δεν περιέχουν όλες τις τιμές. Για παράδειγμα, μετά την εφαρμογή αυτής της τεχνικής ο πίνακας δεδομένων θα γίνει ως εξής:

	Diploma	Experience	French	Reference	Decision
1	MBA	Low	Yes	Neutral	Reject
2	MSc	Medium	Yes	Neutral	Reject
3	MCE	Low	No	Excellent	Reject

Πίνακας 5. Ο πίνακας δεδομένων μετά την διαγραφή των στιγμιοτύπων που περιέχουν ελλιπή δεδομένα

Ο Πίνακας 5 παρουσιάζει τα δεδομένα μετά την αφαίρεση των εγγραφών με ελλιπή δεδομένα. Ο αριθμός των εγγραφών του πίνακα έχει μειωθεί σημαντικά γιατί υπήρχαν πολλές ελλιπείς εγγραφές.

Μια άλλη λύση στο πρόβλημα της απουσίας τιμών είναι η αντικατάσταση των ελλিপών δεδομένων με την διάμεσο ή τη μέση τιμή του αντίστοιχου χαρακτηριστικού. Για παράδειγμα, μετά την εφαρμογή αυτής της τεχνικής ο πίνακας δεδομένων θα γίνει ως εξής:

	Diploma	Experience	French	Reference	Decision
1	MBA	Low	Yes	Neutral	Accept
2	MBA	Low	Yes	Neutral	Reject
3	MSc	Low	Yes	Good	Reject
4	MSc	High	Yes	Neutral	Accept
5	MSc	Medium	Yes	Neutral	Reject
6	MSc	High	Yes	Neutral	Accept
7	MSc	High	No	Good	Accept
8	MCE	Low	No	Excellent	Reject

Πίνακας 6. Ο πίνακας δεδομένων μετά την αντικατάσταση των ελλিপών δεδομένων με την διάμεσο ή την μέση τιμή της αντίστοιχης μεταβλητής

Αυτός ο αλγόριθμος αντικαθιστά τα κενά πεδία του πίνακα με την τιμή εκείνη που εμφανίζεται με μεγαλύτερη συχνότητα στο αντίστοιχο πεδίο αν πρόκειται για κατηγορική μεταβλητή. Όταν τα δεδομένα είναι αριθμητικά η αντικατάσταση γίνεται από τη μέση τιμή του αντίστοιχου πεδίου.

Μια άλλη λύση στο πρόβλημα της απουσίας τιμών είναι η αντικατάσταση των ελλিপών δεδομένων με την διάμεσο ή τη μέση τιμή του αντίστοιχου χαρακτηριστικού αν λάβουμε υπόψη τα στιγμιότυπα της ίδιας κλάσης. Για παράδειγμα, μετά την εφαρμογή αυτής της τεχνικής ο πίνακας δεδομένων θα γίνει ως εξής:

	Diploma	Experience	French	Reference	Decision
1	MBA	High	Yes	Neutral	Accept
2	MSc	High	Yes	Neutral	Accept
3	MSc	High	Yes	Neutral	Accept
4	MSc	High	No	Good	Accept
5	MBA	Low	Yes	Neutral	Reject
6	MBA	Low	Yes	Good	Reject
7	MSc	Medium	Yes	Neutral	Reject
8	MCE	Low	No	Excellent	Reject

Πίνακας 7. Ο πίνακας δεδομένων μετά την αντικατάσταση των ελλিপών δεδομένων με την διάμεσο ή την μέση τιμή της αντίστοιχης μεταβλητής λαμβάνοντας υπόψη τα στιγμιότυπα μόνο της ίδιας κλάσης

Ο αλγόριθμος αυτός είναι τροποποίηση του προηγούμενου. Ανάλογα με την τιμή του πεδίου απόφασης της συγκεκριμένης εγγραφής υπολογίζεται η μέση τιμή (ή η δημοφιλέστερη τιμή) των εγγραφών. Αυτή η τιμή αντικαθιστά το κενό πεδίο της συγκεκριμένης εγγραφής. Πρέπει να τονιστεί ότι και στους δύο προηγούμενους αλγόριθμους ο αριθμός των εγγραφών δεν μεταβάλλεται.

Μια άλλη λύση στο πρόβλημα της απουσίας τιμών είναι η αντικατάσταση των ελλιπών δεδομένων με όλους τους πιθανούς συνδυασμούς τιμών. Για παράδειγμα, μετά την εφαρμογή αυτής της τεχνικής ο πίνακας δεδομένων παρουσιάζεται στον Πίνακα 8. Μία ελλιπής εγγραφή επεκτείνεται σε πολλές άλλες εγγραφές καλύπτοντας όλους τους πιθανούς συνδυασμούς των κενών πεδίων της. Όπως παρατηρούμε ο αριθμός των εγγραφών του πίνακα μπορεί να αυξηθεί πολύ. Στο παράδειγμά μας ο αριθμός των εγγραφών έχει σχεδόν τριπλασιαστεί.

	Diploma	Experience	French	Reference	Decision
1	MBA	Medium	Yes	Excellent	Accept
2	MBA	Medium	Yes	Neutral	Accept
3	MBA	Medium	Yes	Good	Accept
4	MBA	Low	Yes	Excellent	Accept
5	MBA	Low	Yes	Neutral	Accept
6	MBA	Low	Yes	Good	Accept
7	MBA	High	Yes	Excellent	Accept
8	MBA	High	Yes	Neutral	Accept
9	MBA	High	Yes	Good	Accept
10	MBA	Low	Yes	Neutral	Reject
11	MBA	Low	Yes	Good	Reject
12	MCE	Low	Yes	Good	Reject
13	MSc	Low	Yes	Good	Reject
14	MSc	High	Yes	Neutral	Accept
15	MSc	High	No	Neutral	Accept
16	MSc	Medium	Yes	Neutral	Reject
17	MSc	High	Yes	Excellent	Accept
18	MSc	High	Yes	Neutral	Accept
19	MSc	High	Yes	Good	Accept
20	MBA	High	No	Good	Accept
21	MCE	High	No	Good	Accept
22	MSc	High	No	Good	Accept
23	MCE	Low	No	Excellent	Reject

Πίνακας 8. Ο πίνακας δεδομένων μετά την αντικατάσταση των ελλιπών δεδομένων με όλους τους πιθανούς συνδυασμούς τιμών

Μια άλλη λύση στο πρόβλημα της απουσίας τιμών είναι η αντικατάσταση των ελλιπών δεδομένων με όλους τους πιθανούς συνδυασμούς τιμών λαμβάνοντας υπόψη τα στιγμιότυπα μόνο της ίδιας κλάσης. Για παράδειγμα, μετά την εφαρμογή αυτής της τεχνικής ο πίνακας δεδομένων θα γίνει ως εξής:

	Diploma	Experience	French	Reference	Decision
1	MBA	High	Yes	Neutral	Accept
2	MBA	High	Yes	Good	Accept
3	MSc	High	Yes	Neutral	Accept
4	MSc	High	No	Neutral	Accept
5	MSc	High	Yes	Neutral	Accept
6	MSc	High	Yes	Good	Accept
7	MBA	High	No	Good	Accept
8	MSc	High	No	Good	Accept
9	MBA	Low	Yes	Neutral	Reject
10	MBA	Low	Yes	Good	Reject
11	MCE	Low	Yes	Good	Reject
12	MSc	Low	Yes	Good	Reject
13	MSc	Medium	Yes	Neutral	Reject
14	MCE	Low	No	Excellent	Reject

Πίνακας 9. Ο πίνακας δεδομένων μετά την ακτικατάσταση των ελλিপών δεδομένων με όλους τους πιθανούς συνδυασμούς τιμών λαμβάνοντας υπόψη μόνο τα στιγμιότυπα της ίδιας κλάσης

Ο αλγόριθμος αυτός είναι όμοιος με τον προηγούμενο με τη διαφορά ότι δεν δημιουργούνται τόσο πολλές καινούριες εγγραφές. Αυτό γιατί για κάθε κενό πεδίο μιας εγγραφής δημιουργούνται τόσες εγγραφές όσες και οι διαφορετικές τιμές του συγκεκριμένου πεδίου στις εγγραφές που έχουν την ίδια τιμή στο πεδίο απόφασης με τη συγκεκριμένη εγγραφή. Όπως είναι αναμενόμενο, με την χρήση των δύο τελευταίων αλγορίθμων ο αριθμός των εγγραφών μπορεί να αυξηθεί κατά πολύ, ανάλογα με τον αριθμό των κενών πεδίων.

Σε αυτό το σημείο πρέπει να σημειωθεί ότι ο βαθμός δυσκολίας του προβλήματος ταξινόμησης εξαρτάται από την μεταβλητότητα των τιμών των χαρακτηριστικών για αντικείμενα της ίδιας κατηγορίας σε σχέση με τη διαφορά μεταξύ τιμών για αντικείμενα άλλων κατηγοριών (Mitchell, 1997). Για αυτό το λόγο πολλοί αλγόριθμοι μάθησης ευνοούνται από την διακριτοποίηση (discretization) μιας συνεχούς μεταβλητής σε κατηγορική (Liu και άλλοι, 2002).

Πολυάριθμες μέθοδοι διακριτοποίησης (κατηγοριοποίησης) έχουν εξεταστεί όπως η κατάτμηση του διαστήματος τιμών μιας συνεχούς μεταβλητής σε ίσου μεγέθους K διαστήματα καθώς και η κατάτμηση του διαστήματος τιμών μιας συνεχούς μεταβλητής σε διαστήματα που να περιέχουν ίδιο αριθμό περιπτώσεων (Berka & Bruha, 1998). Για παράδειγμα παρουσιάζουμε στον Πίνακα 10 ένα σύνολο δεδομένων που περιέχει μια συνεχή μεταβλητή την ηλικία (age).

	Diploma	Experience	French	Reference	Age	Decision
1	MBA	Medium	Yes	Excellent	25	Accept
2	MBA	Low	Yes	Neutral	32	Reject
3	MCE	Low	Yes	Good	23	Reject
4	MSc	High	Yes	Neutral	40	Accept
5	MSc	Medium	Yes	Neutral	41	Reject
6	MSc	High	Yes	Excellent	27	Accept
7	MBA	High	No	Good	29	Accept
8	MCE	Low	No	Excellent	30	Reject

Πίνακας 10. Ο αρχικός πίνακας δεδομένων που περιέχει μια συνεχή μεταβλητή

Με την πιο απλή τεχνική διακριτοποίησης τα διαστήματα που δημιουργούνται είναι τέτοια ώστε σε κάθε διάστημα να περιλαμβάνεται ο ίδιος αριθμός εγγραφών. Για παράδειγμά, αν δημιουργήσουμε 3 τέτοια διαστήματα, το αποτέλεσμα παρουσιάζεται στον Πίνακα 11.

	Diploma	Experience	French	Reference	Age	Decision
1	MBA	Medium	Yes	Excellent	[*, 28)	Accept
2	MBA	Low	Yes	Neutral	[28, 36)	Reject
3	MCE	Low	Yes	Good	[*, 28)	Reject
4	MSc	High	Yes	Neutral	[36, *)	Accept
5	MSc	Medium	Yes	Neutral	[36, *)	Reject
6	MSc	High	Yes	Excellent	[*, 28)	Accept
7	MBA	High	No	Good	[28, 36)	Accept
8	MCE	Low	No	Excellent	[28, 36)	Reject

Πίνακας 11. Ο πίνακας δεδομένων μετά την κατάτμηση του διαστήματος τιμών της συνεχούς μεταβλητής σε διαστήματα που περιέχουν ίδιο αριθμό περιπτώσεων

Μια συνοπτική αναφορά σε όλες τις μεθόδους κατηγοριοποίησης και μια εμπειρική μελέτη που συγκρίνει την απόδοσή τους δίνεται στην εργασία (Elomaa & Rousu, 1999). Οι ίδιοι ερευνητές επίσης κατέληξαν στο συμπέρασμα ότι η μέθοδος που βασίζεται στην εντροπία έδωσε λίγο καλύτερα αποτελέσματα από τις υπόλοιπες. Η μέθοδος που βασίζεται στην εντροπία είναι μια εποπτευμένη επαυξητική μέθοδος που επιλέγει κατ' επανάληψη τα σημεία περικοπής του διαστήματος τιμών μιας συνεχούς μεταβλητής που ελαχιστοποιούν την εντροπία, έως ότου ένα κριτήριο τερματισμού βασισμένο στο κριτήριο ελαχίστου μήκους περιγραφής ικανοποιηθεί (Fayyad & Irani, 1993) ⁶. Ο αλγόριθμος σχηματικά παρουσιάζεται στην Εικόνα 3.

Εικόνα 3. Αλγόριθμος διακριτοποίησης

Ο Πίνακας 12 παρουσιάζει τα αποτελέσματα της εφαρμογής της μεθόδου της εντροπίας στο παράδειγμα μας.

⁶ Περισσότερα στοιχεία για την εντροπία δίνονται στο κεφάλαιο 3.1

	Diploma	Experience	French	Reference	Age	Decision
1	MBA	Medium	Yes	Excellent	[*, 41)	Accept
2	MBA	Low	Yes	Neutral	[*, 41)	Reject
3	MCE	Low	Yes	Good	[*, 41)	Reject
4	MSc	High	Yes	Neutral	[*, 41)	Accept
5	MSc	Medium	Yes	Neutral	[41, *)	Reject
6	MSc	High	Yes	Excellent	[*, 41)	Accept
7	MBA	High	No	Good	[*, 41)	Accept
8	MCE	Low	No	Excellent	[*, 41)	Reject

Πίνακας 12. Ο πίνακας δεδομένων μετά την κατάτμηση του διαστήματος τιμών της συνεχούς μεταβλητής με την εφαρμογή της μεθόδου της εντροπίας

Επίσης, η μετατροπή των δεδομένων στην ίδια κλίμακα μεγέθους ή η κανονικοποίησή τους ως προς μια μονάδα μέτρησης είναι συχνά στάδια προετοιμασίας των δεδομένων για την παραπέρα ανάλυση. Περισσότερα στοιχεία για την κανονικοποίηση δίνονται στην ενότητα 3.3.

Τέλος, πρέπει να σημειωθεί ότι η επιλογή περιπτώσεων δεν χρησιμοποιείται μόνο για να αντιμετωπιστεί ο θόρυβος αλλά και για την δυνατότητα σχετικά ταχείας μάθησης από πολύ μεγάλα σύνολα δεδομένων (Blum & Langley, 1997). Η επιλογή περιπτώσεων είναι σε αυτήν την περίπτωση ένα πρόβλημα βελτιστοποίησης που προσπαθεί να διατηρήσει την ποιότητα της εξαγόμενης γνώσης ελαχιστοποιώντας το μέγεθος των δεδομένων (Liu και Metoda, 2001). Μειώνοντας τα δεδομένα επιτρέπουμε σε έναν αλγόριθμο μάθησης να λειτουργήσει και να εργαστεί πιο αποτελεσματικά σε τεράστιες βάσεις δεδομένων (Provost & Kolluri, 1999). Υπάρχουν πολλές γνωστές διαδικασίες δειγματοληψίας από μεγάλα σύνολα δεδομένων. Οι πιο γνωστές είναι (Reinartz, 2002):

- η τυχαία δειγματοληψία που επιλέγει ένα υποσύνολο των περιπτώσεων τυχαία
- η στρωματοποιημένη δειγματοληψία που κυρίως χρησιμοποιείται όταν δεν διανέμονται ομοιόμορφα οι τιμές ανά κατηγορία στο σύνολο κατάρτισης. Σε αυτή την δειγματοληψία οι περιπτώσεις για τις κατηγορίες μειοψηφίας επιλέγονται με μεγαλύτερη συχνότητα για την κατασκευή του συνόλου δεδομένων.

2.2 Εξαγωγή χρήσιμων χαρακτηριστικών/μεταβλητών

Η ποιότητα ενός συνόλου δεδομένων μπορεί συνήθως να χαρακτηριστεί από δύο πηγές πληροφοριών: μεταβλητές (ιδιότητες) και ετικέτες κατηγορίας. Η ποιότητα των ιδιοτήτων προσδιορίζει πόσο καλά οι ιδιότητες χαρακτηρίζουν τις περιπτώσεις για το σκοπό της ταξινόμησης και η ποιότητα των ετικετών κατηγορίας αντιπροσωπεύει εάν η κατηγορία σε κάθε μια από τις περιπτώσεις ορίζεται σωστά.

Ένα πρόβλημα που εμφανίζεται σε αρκετές εφαρμογές μηχανικής μάθησης είναι ο μεγάλος αριθμός χαρακτηριστικών που θεωρούνται, κατ'αρχήν τουλάχιστον, σχετικά με τη συνάρτηση-στόχο που επιδιώκεται να προσεγγιστεί. Όπως προείπαμε, καθοριστική σημασία στην ποιότητα του σώματος εκπαίδευσης φέρει η ορθή επιλογή των χαρακτηριστικών τα οποία θα συστήσουν το χώρο του προβλήματος, καθώς μέσω αυτών καθίσταται δυνατή η αναπαράσταση της γνώσης που απαιτείται για την αποδοτική λειτουργία του συστήματος. Δύο είναι οι βασικοί λόγοι για τους οποίους η μεγάλη διαστασιμότητα (dimensionality) αποτελεί πρόβλημα, και μάλιστα τόσο σημαντικό που πολλές φορές αναφέρεται ως κατάρα της διαστασιμότητας (curse of dimensionality). Ένας λόγος είναι η αναπόφευκτη αύξηση της υπολογιστικής πολυπλοκότητας για την εκπαίδευση και τη λειτουργία του ταξινομητή. Αν και το επιπλέον υπολογιστικό κόστος ενδεχομένως να γινόταν αποδεκτό σε διάφορες εφαρμογές, αν μπορούσε να εγγραφεί την παράλληλη αύξηση της ακρίβειας του προκύπτοντος ταξινομητή, στην πράξη όμως, η μεγάλη διαστασιμότητα συχνά όχι μόνο δεν

αυξάνει την ακρίβεια, αλλά την μειώνει (Dash & Liu, 1997). Η αιτία είναι η εμφάνιση του φαινομένου του υπερταϊριάσματος. Περισσότερα διαθέσιμα χαρακτηριστικά σημαίνουν μεγαλύτερη πιθανότητα να αναγνωριστούν λανθασμένα πολλά από αυτά ως ουσιώδη για τον προσδιορισμό της τιμής της συνάρτησης-στόχου, ενώ στην πραγματικότητα αποτελούν συμπτωματικά καλούς προσδιοριστές, για το συγκεκριμένο διαθέσιμο σύνολο στιγμιότυπων εκπαίδευσης. Έτσι, ο ταξινομητής που προκύπτει έχει υψηλή ακρίβεια για το σύνολο εκπαίδευσης, αλλά σημαντικά χαμηλότερη για άγνωστα στιγμιότυπα.

Οι δύο μέθοδοι που εφαρμόζονται για την επιλογή των πιο σημαντικών χαρακτηριστικών είναι του περιτυλίγματος (wrapper) και του φίλτρου (filter) (Dash & Liu, 1997). Η σχηματική αναπαράσταση της διαδικασίας επιλογής των σημαντικότερων χαρακτηριστικών παρουσιάζεται στην Εικόνα 4.

Εικόνα 4. Επιλογή των πιο σημαντικών χαρακτηριστικών

Η μέθοδος του περιτυλίγματος χρησιμοποιεί τον ίδιο τον αλγόριθμο μάθησης ως μέσο για την καθοδήγηση της αναζήτησης στο χώρο των μεταβλητών/χαρακτηριστικών (Kohavi & John, 1997). Ξεκινώντας από ένα αρχικό σύνολο μεταβλητών, δημιουργούνται διαδοχικά νέα σύνολα μέσω κατάλληλα ορισμένων τελεστών που προσθέτουν ή διαγράφουν μεταβλητές. Καθένα απ' αυτά χρησιμοποιείται για την εκπαίδευση ενός ταξινομητή, ο οποίος στη συνέχεια ελέγχεται πάνω σε ένα σύνολο δεδομένων επικύρωσης (validation set). Όταν εκπληρωθεί η συνθήκη τερματισμού της αναζήτησης, το καλύτερο σύνολο ως προς την απόδοση στο σύνολο δεδομένων επικύρωσης επιστρέφεται. Το πλεονέκτημα της μεθόδου είναι πως συντονίζει το σύνολο των μεταβλητών της αναπαράστασης στο συγκεκριμένο επαγωγικό αλγόριθμο μάθησης. Το μειονέκτημα της είναι η απαγορευτική πολυπλοκότητά της αν επιχειρηθεί εξαντλητική αναζήτηση του χώρου αναζήτησης, γεγονός που αντιμετωπίζεται μερικώς μέσω τεχνικών ευριστικής αναζήτησης. Περισσότερα για τις μεθόδους περιτυλίγματος στο κεφάλαιο 5.

Η μέθοδος του φίλτρου είναι περισσότερο εφικτή υπολογιστικά και πραγματοποιείται με την επιλογή των μεταβλητών που έχουν υψηλότερη τιμή σε μια προκαθορισμένη αριθμητική συνάρτηση που εκτιμά τη βαρύτητα της μεταβλητής για το έργο της κατηγοριοποίησης (Kudo & Sklansky, 2000). Αρκετές πολύπλοκες συναρτήσεις έχουν χρησιμοποιηθεί, προερχόμενες από τη θεωρία πληροφορίας, και οι οποίες γενικά προσπαθούν να εκφράσουν την πεποίθηση πως οι καλύτερες μεταβλητές για μια κατηγορία είναι αυτές που έχουν τη μεγαλύτερη διαχωριστική ικανότητα ως προς την κατηγορία, δηλαδή οι τιμές τους κατανέμονται όσο γίνεται πιο διαφορετικά για τα θετικά και τα αρνητικά στιγμιότυπα της κατηγορίας. Στις

συναρτήσεις αυτές περιλαμβάνονται το πληροφοριακό κέρδος (information gain), η αμοιβαία πληροφορία (mutual information), η χ^2 (chi-square), ο λόγος πιθανοτήτων (odds ratio), ο βαθμός σχετικότητας (relevancy score) και ο συντελεστής συσχέτισης (Liu & Jain, 1998). Για παράδειγμα ο συντελεστής συσχέτισης της μεταβλητής X και της κλάσης Y υπολογίζεται ως:

$$\rho(X, Y) \equiv \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

$$\text{όπου } E(X) \approx \bar{X} = \frac{1}{M} \sum_{m=1}^M x_m, \quad E(Y) \approx \bar{Y} = \frac{1}{M} \sum_{m=1}^M c_m, \quad E(XY) \approx \frac{1}{M} \sum_{m=1}^M x_m c_m$$

$$\text{και } Var(X) \approx \frac{1}{M} \sum_{m=1}^M (x_m - \bar{X})^2, \quad Var(Y) \approx \frac{1}{M} \sum_{m=1}^M (c_m - \bar{Y})^2$$

Μια άλλη προσέγγιση φίλτρου εξετάζει τη διαχωριστική ικανότητα υποσυνόλων χαρακτηριστικών, η οποία υλοποιείται με τους αλγορίθμους: επιλογής των χαρακτηριστικών βασισμένης στην αλληλοσυσχέτιση (Correlation-based Feature Selection ή CFS) και της αξιολόγησης της συνέπειας των υποσυνόλων των χαρακτηριστικών (Consistency Subset Evaluation ή CSE). Βάσει αυτού του αλγορίθμου (Hall, 2000), αναζητείται όχι μόνο το καλύτερο υποσύνολο χαρακτηριστικών του οποίου τα μέλη παρουσιάζουν μεγάλο βαθμό συσχέτισης με την τιμή της συνάρτησης στόχου, αλλά και αυτό με το μικρότερο βαθμό συσχέτισης μεταξύ τους.

2.3 Επιλογή του κατάλληλου αλγορίθμου μάθησης

Η επιλογή του καταλληλότερου αλγορίθμου μάθησης για το δοθέν πρόβλημα περιλαμβάνει αποφάσεις για το ποιο μοντέλο και ποιες παράμετροι είναι οι πλέον κατάλληλες να χρησιμοποιηθούν. Είναι γνωστό ότι ένα πρόβλημα μπορεί να λυθεί συγκριτικά καλά χρησιμοποιώντας διάφορες μεθόδους, π.χ. διαφορετικούς τύπους τεχνητών νευρωνικών δικτύων, δέντρων απόφασης, Μπεϋζιανών ταξινομητών, κ.λπ. Επιπλέον, μπορεί να συμβεί μια μέθοδος να προβλέπει καλύτερα ορισμένα μέρη του διαστήματος των περιπτώσεων από τις άλλες. Κατά συνέπεια, η επιλογή της καταλληλότερης μεθόδου για την τελική λύση είναι ένα περίπλοκο πρόβλημα.

Ακολουθώς περιγράφονται οι πιο συνήθεις μέθοδοι εκτίμησης ακρίβειας για την επιλογή του βέλτιστου μοντέλου. Η ακρίβεια (accuracy) του μοντέλου προκύπτει από τη διαίρεση του αθροίσματος των στιγμιοτύπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία και των στιγμιοτύπων που σωστά απορρίφθηκαν από την συγκεκριμένη κατηγορία, με το άθροισμα των στιγμιοτύπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία, των στιγμιοτύπων που λανθασμένα ανατέθηκαν στην συγκεκριμένη κατηγορία, των στιγμιοτύπων που λανθασμένα απορρίφθηκαν από την συγκεκριμένη κατηγορία και των στιγμιοτύπων που σωστά απορρίφθηκαν από την συγκεκριμένη κατηγορία.

Η πιο απλή μέθοδος αξιολόγησης της ακρίβειας είναι η εκτίμηση μέσω δείγματος ελέγχου. Σε αυτήν, η αρχική συλλογή των δεδομένων χωρίζεται σε δύο ξένα υποσύνολα, ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο ελέγχου (test set). Ο αλγόριθμος μάθησης δέχεται το πρώτο για την εκπαίδευσή του και “εξετάζεται” στο δεύτερο, χωρίς να του αποκαλύπτονται φυσικά οι γνωστές (και θεωρούμενες σωστές) απαντήσεις. Στην περίπτωση που υπάρχουν παράμετροι του αλγορίθμου που πρέπει να συντονισθούν, έτσι ώστε να επιλεγθούν αυτές που δίνουν την υψηλότερη αποτελεσματικότητα, το σύνολο εκπαίδευσης πρέπει να χωριστεί επιπλέον σε ένα “πραγματικό” σύνολο εκπαίδευσης και ένα σύνολο επικύρωσης (validation set), το οποίο χρησιμοποιείται “εσωτερικά” από τον αλγόριθμο για

την εκτίμηση των βέλτιστων παραμέτρων. Μόνο όταν αυτές σταθεροποιηθούν, εξετάζεται ο αλγόριθμος πάνω στο σύνολο ελέγχου.

Σε αυτήν την περίπτωση η στατιστική μπορεί να δώσει κάποια διαστήματα μέσα στα οποία ανήκει η ακρίβεια με πιθανότητα $(1-a) \cdot 100\%$, τα οποία καλούνται διαστήματα εμπιστοσύνης. Με εφαρμογή της διαδικασίας που δίνει τα διαστήματα εμπιστοσύνης μπορεί να δοθεί ένα διάστημα, όπου με πιθανότητα 95% η τιμή πραγματικού σφάλματος ανήκει σ' αυτό. Το διάστημα εμπιστοσύνης λοιπόν του σφάλματος (Mitchell, 1997) είναι

$$\left[error_s(h) - 1.96 \sqrt{\frac{error_s(h)(1 - error_s(h))}{n}}, error_s(h) + 1.96 \sqrt{\frac{error_s(h)(1 - error_s(h))}{n}} \right]$$

Άρα η έκφραση που πρέπει να εκτιμηθεί για την σύγκριση αλγορίθμων δίνεται από την έκφραση $error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$ όπου η έκφραση $L(S)$ συμβολίζει τον ταξινομητή που εξάγεται από έναν αλγόριθμο L που εκπαιδεύτηκε πάνω σε ένα σύνολο εκπαίδευσης S και ο συμβολισμός $S \subset D$ σημαίνει ότι η επιλογή του S γίνεται σύμφωνα με την κατανομή D . Για να εκτιμηθεί η διαφορά του σφάλματος υπολογίζεται λοιπόν η μέση τιμή $\bar{\delta} \equiv E_{S \subset D_0} [error_D(L_A(S)) - error_D(L_B(S))]$ δίνοντας ένα διάστημα εμπιστοσύνης για τη

διαφορά του σφάλματος $\bar{\delta} \pm t_{N,k-1} s_{\bar{\delta}}$ όπου $t_{N,k-1}$ είναι μια σταθερά της στατιστικής και

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2} \text{ (Mitchell, 1997). Αφού οι υποθέσεις προέρχονται από εκπαίδευση}$$

πάνω στα ίδια σύνολα, οι μόνες διαφορές που παρατηρούνται στις υποθέσεις οφείλονται μόνο στους αλγορίθμους μάθησης.

Το πρόβλημα με τη προηγούμενη μέθοδο είναι πως εξαρτάται έντονα από τον τυχαίο διαχωρισμό σε δύο σύνολα. Αν τύχει το δείγμα που θα επιλεγεί για σύνολο εκπαίδευσης να είναι αρκετά αντιπροσωπευτικό του συνόλου ελέγχου, τα αποτελέσματα θα είναι υπερεκτιμημένα, ενώ το αντίθετο θα συμβεί αν δεν είναι αντιπροσωπευτικό. Μια μέθοδος που μειώνει το πρόβλημα είναι η διασταυρωμένη επικύρωση (cross-validation – CV). Σε αυτήν, η συλλογή χωρίζεται τυχαία σε k ξένα μεταξύ τους τμήματα ίσου (περίπου) μεγέθους (folds). Στην συνέχεια γίνονται k επαναλήψεις. Σε κάθε επανάληψη, ένα από τα τμήματα είναι το σύνολο ελέγχου και τα υπόλοιπα $k-1$ αποτελούν το σύνολο εκπαίδευσης. Αφού κληθεί ο αλγόριθμος μάθησης για κάθε τμήμα, η εκτιμώμενη τιμή του μέτρου είναι ο μέσος όρος του για όλα τα τμήματα (βλέπε Εικόνα 5). Με τον τρόπο αυτό, ακόμα κι αν σε κάποια τμήματα γίνεται υπερεκτίμηση, για μεγάλο k αναμένεται ισόποση σχεδόν υποεκτίμηση, με αποτέλεσμα ο μέσος όρος να είναι πιο κοντά στην πραγματική τιμή.

Ειδική περίπτωση της διασταυρωμένης επικύρωσης είναι η εξαντλητική μέθοδος (leave-one-out), για την οποία τα σύνολα ελέγχου αποτελούνται από ένα μόνο στιγμιότυπο. Μια άλλη παραλλαγή είναι η στρωματοποιημένη διασταυρωμένη επικύρωση (stratified CV), στην οποία τα τμήματα δε δημιουργούνται τελείως τυχαία, αλλά διατηρούν την ίδια σχεδόν κατανομή στιγμιότυπων για κάθε κλάση με αυτή της αρχικής συλλογής.

Εικόνα 5. Διασταυρωμένη επικύρωση

Μια άλλη μέθοδος είναι η δειγματοληπτική. Σε αυτήν, αν η συλλογή αποτελείται από m στιγμιότυπα, κατασκευάζονται b δειγματοληπτικά (bootstrap) δείγματα (bootstrap samples), επιλέγοντας τυχαία για το καθένα n στιγμιότυπα με επανάθεση. Εφόσον τα δεδομένα δειγματολειτουργούν με επανάθεση, η πιθανότητα να μην περιέχεται ένα στιγμιότυπο στο δείγμα είναι 36.8% για μεγάλο n . Επομένως, κάθε δείγμα περιέχει περίπου 63.2% διαφορετικά στιγμιότυπα (Breiman, 1996). Κάθε τέτοιο δείγμα αποτελεί το σύνολο εκπαίδευσης και όσα στιγμιότυπα δεν περιέχονται σε αυτό αποτελούν το σύνολο ελέγχου (βλέπε Εικόνα 6).

Εικόνα 6. Δειγματοληπτικός έλεγχος

Το εκτιμώμενο λάθος είναι ο μέσος όρος του λάθους για τις b επαναλήψεις (βλέπε Εικόνα 7). Ένα πλεονέκτημα που έχει αυτή η μέθοδος έναντι της διασταυρωμένης επικύρωσης είναι πως το b μπορεί να είναι οσοδήποτε μεγάλο, παρέχοντας καλύτερη εκτίμηση, ενώ στη διασταυρωμένη επικύρωση ο μέγιστος αριθμός τμημάτων είναι το πλήθος των παραδειγμάτων. Από την άλλη, η διασταυρωμένη επικύρωση έχει το πλεονέκτημα πως τα σύνολα ελέγχου είναι ανεξάρτητα μεταξύ τους (αφού δεν περιέχουν κοινά στιγμιότυπα), με αποτέλεσμα οι k μετρήσεις να είναι πιο ανεξάρτητες σε σχέση με τη δειγματοληπτική (bootstrap), όπου τα σύνολα ελέγχου επικαλύπτονται μερικώς.

Εικόνα 7. Εκτίμηση της ακρίβειας

Μια άλλη σημαντική πτυχή της αξιολόγησης είναι η σύγκριση διαφορετικών μεθόδων με όρους στατιστικής (Salzberg, 1997). Το πρόβλημα αποτελεί μία περίπτωση ελέγχου υποθέσεων και διατυπώνεται ως το κατά πόσο μια παρατηρούμενη διαφορά στην αποτελεσματικότητα μεταξύ δύο μεθόδων αντικατοπτρίζει την πραγματική υπεροχή της μιας από την άλλη ή είναι τυχαία, ή όπως συνήθως λέγεται, το κατά πόσο η διαφορά είναι στατιστικά σημαντική (statistically significant). Συνήθως αυτό αποφασίζεται από την πιθανότητα να παρατηρηθεί διαφορά υποθέτοντας πως οι μέθοδοι δε διαφέρουν πραγματικά (μηδενική υπόθεση –null hypothesis). Η διαφορά θεωρείται σημαντική (οπότε απορρίπτουμε τη μηδενική υπόθεση) αν και μόνο αν αυτή η πιθανότητα είναι κάτω από ένα όριο p , η επιλογή του οποίου γίνεται εμπειρικά, με συνήθη τιμή το 0.05. Η εκτίμηση της πιθανότητας γίνεται μέσω κάποιου στατιστικού ελέγχου, όπως η βελτιωμένη δοκιμή t (Nadeau & Bengio, 2003).

Τέλος, αν η απόδοση που λαμβάνεται από ένα δοθέν σύνολο χαρακτηριστικών είναι ανεπαρκής για όλα τα μοντέλα, τότε είναι λογικό να θεωρήσουμε ότι πρέπει να προσθέσουμε νέα χαρακτηριστικά. Ειδικά αυτά που θα βοηθήσουν στο να ξεχωρίσουμε τα ζευγάρια των κατηγοριών που είναι πιο συχνά δύσκολο να ταξινομηθούν χωρίς μπέρδεμα. Αν και η αύξηση του αριθμού των χαρακτηριστικών αυξάνει το κόστος και την πολυπλοκότητα του ταξινομητή, είναι συχνά λογικό να υποθέσουμε ότι η απόδοση θα βελτιωθεί.

2.4 Αξιολόγηση της παραγόμενης γνώσης

Τυπικά, μετά και από την επαγωγική κατασκευή του ταξινομητή, ένα σύστημα σε λειτουργικό στάδιο έχει ολοκληρώσει το έργο της εκπαίδευσής του και είναι έτοιμο να χρησιμοποιηθεί στην πράξη. Κατά την πειραματική εξέλιξη του συστήματος, ωστόσο, μένει το πολύ σημαντικό βήμα της αξιολόγησης. Αυτό έχει τρεις κυρίως σκοπούς:

1. Την εκτίμηση της χρησιμότητας του συστήματος για τους υποψήφιους χρήστες του, με τη βοήθεια κατάλληλα ορισμένων μέτρων αποτελεσματικότητας και τεχνικών εκτίμησης των μέτρων αυτών.
2. Τη ρύθμιση διαφόρων παραμέτρων του συστήματος έτσι ώστε να βελτιστοποιηθεί η απόδοσή του ως προς τα επιλεχθέντα μέτρα αποτελεσματικότητας.

Ακολούθως περιγράφονται τα σημαντικότερα ζητήματα σχετικά με την αξιολόγηση. Το πιο σημαντικό κριτήριο για την αξία ενός συστήματος κατηγοριοποίησης είναι φυσικά η αποτελεσματικότητά του (effectiveness) στο έργο της κατηγοριοποίησης, η οποία αντικατοπτρίζει την ακρίβεια των προβλέψεων. Άλλα κριτήρια μέτρησης της απόδοσης είναι τα:

- Recall: προκύπτει από τη διαίρεση του πλήθους των στιγμιοτύπων (instances) που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία, με το άθροισμα των στιγμιοτύπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία και των στιγμιοτύπων που λανθασμένα απορρίφθηκαν από την συγκεκριμένη κατηγορία.
- Precision: προκύπτει από τη διαίρεση του πλήθους των στιγμιοτύπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία, με το άθροισμα των στιγμιοτύπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία και των στιγμιοτύπων που λανθασμένα ανατέθηκαν στην συγκεκριμένη κατηγορία.
- Fallout: προκύπτει από τη διαίρεση του πλήθους των στιγμιοτύπων που λανθασμένα ανατέθηκαν στην συγκεκριμένη κατηγορία, με το άθροισμα των στιγμιοτύπων που λανθασμένα ανατέθηκαν στην συγκεκριμένη κατηγορία και των στιγμιοτύπων που σωστά απορρίφθηκαν από την συγκεκριμένη κατηγορία.
- Error: Η τιμή του μέτρου αυτού, προκύπτει από τη διαίρεση του αθροίσματος των στιγμιοτύπων που λανθασμένα ανατέθηκαν στην συγκεκριμένη κατηγορία και των στιγμιοτύπων που λανθασμένα απορρίφθηκαν από την συγκεκριμένη κατηγορία, με το άθροισμα των στιγμιοτύπων που σωστά ανατέθηκαν στην συγκεκριμένη κατηγορία, των στιγμιοτύπων που λανθασμένα ανατέθηκαν στην συγκεκριμένη κατηγορία, των στιγμιοτύπων που λανθασμένα απορρίφθηκαν από την συγκεκριμένη κατηγορία και των στιγμιοτύπων που σωστά απορρίφθηκαν από την συγκεκριμένη κατηγορία.

Άλλα κριτήρια, όπως π.χ. η αποδοτικότητα (efficiency) η οποία αναφέρεται στην χρονική και χωρική πολυπλοκότητα των αλγορίθμων, θεωρούνται δευτερεύουσας σημασίας. Ωστόσο είναι χρήσιμα για συγκρίσεις μεταξύ συστημάτων με παρόμοια αποτελεσματικότητα, καθώς και για εφαρμογές με μεγάλους όγκους δεδομένων ή/και πραγματικού χρόνου που απαιτούν γρήγορη απόκριση.

Πρέπει τέλος να σημειωθεί ότι η τελική απόδοση ενός συστήματος είναι συνάρτηση πολλών παραμέτρων. Στην ιδανική περίπτωση αυτές οι παράμετροι είναι ανεξάρτητες μεταξύ τους, γεγονός που απλοποιεί ουσιαστικά την πειραματική βελτιστοποίηση του τελικού συστήματος. Ο λόγος είναι πως μπορεί να μελετηθεί κάθε παράμετρος χωριστά από τις υπόλοιπες, να εκτιμηθεί η βέλτιστη τιμή της και στη συνέχεια να μελετηθούν οι υπόλοιπες θεωρώντας αυτήν ως σταθερά με τιμή τη βέλτιστη. Έτσι, το πλήθος των βημάτων που απαιτούνται για την πειραματική βελτιστοποίηση του συστήματος είναι όσο και το άθροισμα του πλήθους των διαφορετικών τιμών που παίρνουν οι παράμετροί του (αν κάποιες είναι συνεχείς, είναι αναγκαία η διακριτοποίηση τους).

Στην πράξη όμως, οι παράμετροι είναι συχνά αλληλοεξαρτώμενες. Στη χειρότερη περίπτωση που όλες οι παράμετροι αλληλεξαρτώνται, οι δυνατοί συνδυασμοί που πρέπει να μελετηθούν είναι το καρτεσιανό γινόμενο των συνόλων τιμών των παραμέτρων, γεγονός που ανεβάζει το πλήθος των βημάτων για τη βελτιστοποίηση στο γινόμενο του πλήθους των τιμών των παραμέτρων. Κάτι τέτοιο είναι συνήθως απαγορευτικό από άποψη χρόνου. Μία πρακτική αντιμετώπιση του προβλήματος είναι να γίνουν κάποιες υποθέσεις ανεξαρτησίας μεταξύ συγκεκριμένων παραμέτρων, οι οποίες στην καλύτερη περίπτωση επιβεβαιώνονται μέσω δειγματοληπτικής επιλογής κάποιων τιμών των θεωρούμενων παραμέτρων και διενέργειας πειραμάτων με τις τιμές αυτές. Αν οι υποθέσεις επιβεβαιωθούν, προσεγγιστικά έστω, υιοθετούνται για τη συνέχεια της βελτιστοποίησης. Αφού σταθεροποιηθεί μία (υποτιθέμενη) ανεξάρτητη παράμετρος στη βέλτιστη, κατά τα πειράματα, τιμή της, η ίδια διαδικασία πρότασης και υιοθέτησης υποθέσεων μπορεί να συνεχιστεί για τις υπόλοιπες παραμέτρους, στο βαθμό που επιβεβαιώνεται από τα πειραματικά δεδομένα. Επιχειρείται έτσι μια ευρετική διερεύνηση του χώρου των παραμέτρων αντί μιας εξαντλητικής αναζήτησης του βέλτιστου συνδυασμού.

Μια διάκριση που μπορεί να γίνει ως προς το ρόλο των παραμέτρων είναι αν αυτές προέρχονται από τις ίδιες τις απαιτήσεις του προβλήματος (“εξωτερικές”) ή είναι “εσωτερικές” της μεθόδου επίλυσής του. Οι πρώτες μπορούν να θεωρηθούν ανεξάρτητες, επειδή καθορίζονται από το χρήστη κατά βούληση, ενώ οι δεύτερες εξαρτώνται από την επιλογή των πρώτων και αποφασίζονται από το σχεδιαστή του συστήματος.

Είναι γεγονός ότι κάποια προβλήματα μάθησης μπορούν να επιλυθούν με τη χρήση αλγορίθμων που είναι όμως μη πρακτικοί. Για παράδειγμα, μπορεί να θέλουμε να θέσουμε ετικέτες σε εικόνες με 20x20 εικονοστοιχεία για γραφική αναγνώριση χαρακτήρων (OCR) και να χρησιμοποιήσουμε ένα τεράστιο πίνακα για κάθε πιθανή περίπτωση και στη συνέχεια να ταξινομούμε τα εισερχόμενα πρότυπα. Ενώ θεωρητικά μπορεί να έχουμε μηδενικό σφάλμα ταξινόμησης, ο χρόνος για την «ετικετοποίηση» και οι απαιτήσεις σε αποθηκευτικό χώρο είναι απαγορευτικός αφού θα απαιτούνταν «ετικετοποίηση» $2^{20 \times 20}$ προτύπων. Έτσι γίνεται σαφές ότι πρέπει τέτοια θέματα να παίρνονται σοβαρά υπόψη σε πρακτικά προβλήματα.

Στη βιβλιογραφία υπάρχει μια μεγάλη ποικιλία μοντέλων (αλγορίθμων Μ.Μ.), οι οποίοι χωρίζονται σε κατηγορίες. Στο επόμενο κεφάλαιο αναφέρονται συνοπτικά οι διάφορες κατηγορίες. Φυσικά, ένα κεφάλαιο όσο μεγάλο και αν είναι δεν μπορεί να αποτελεί μια πλήρη ανασκόπηση όλων των εποπτευμένων αλγορίθμων μηχανικής μάθησης. Παρόλα αυτά, ελπίζουμε ότι οι αναφορές που αναφέρονται καλύπτουν τα σημαντικότερα θεωρητικά ζητήματα και καθοδηγούν τον ερευνητή σε ενδιαφέρουσες ερευνητικές κατευθύνσεις.

ΚΕΦΑΛΑΙΟ 3. Επιτηρούμενη μηχανική μάθηση

Έστω ότι έχουμε τέσσερα παραδείγματα, για όλους τους συνδυασμούς δυαδικών τιμών για τις μεταβλητές εισόδου A και B:

#	A	B	Έξοδος ($A \wedge \neg B$)
1	0	0	0
2	0	1	0
3	1	0	1
4	1	1	0

Σε διδιάστατο διάγραμμα μπορούμε να σχεδιάσουμε μια ευθεία η οποία να διαχωρίζει τα τέσσερα παραδείγματα (βλέπε Εικόνα 8):

Εικόνα 8. Γραμμικά διαχωρίσιμο πρόβλημα (Linear decision boundary)

Στην παραπάνω εικόνα με λευκούς κύκλους φαίνονται τα παραδείγματα που αντιστοιχούν στην έξοδο 0 και με μαύρο κύκλο το παράδειγμα που αντιστοιχεί στην έξοδο 1. Τα παραδείγματα είναι φανερά γραμμικώς διαχωρίσιμα. Η διακεκομμένη ευθεία κλίσης 45° διαχωρίζει σωστά τα παραδείγματα (και μάλιστα το κάνει με τον καλύτερο δυνατό τρόπο).

Όμως εν γένει, τα προβλήματα ταξινόμησης στην μηχανική είναι μη γραμμικώς διαχωρίσιμα (βλέπε Εικόνα 9). Όπως είναι λογικό τα μη γραμμικώς διαχωρίσιμα προβλήματα είναι και πιο δύσκολα.

Ένα πλήθος μεθόδων όπως τα τεχνητά νευρωνικά δίκτυα, τα δέντρα αποφάσεων, οι Μπεϋζιανοί ταξινομητές, κ.λπ. μπορούν να επιλύσουν μη γραμμικώς διαχωρίσιμα προβλήματα. Η παρουσίαση των αλγορίθμων επιτηρούμενης μηχανικής μάθησης σε αυτή την ενότητα αρχίζει με τα Δένδρα Αποφάσεων (Decision Trees). Τα Δένδρα Αποφάσεων αντιπροσωπεύουν κανόνες, οι οποίοι μπορούν εύκολα να διατυπωθούν σε φυσική γλώσσα ώστε να είναι εύκολα κατανοητοί από τους ανθρώπους ή να διατυπωθούν σε μία γλώσσα προσπέλασης βάσεων δεδομένων π.χ. σε SQL.

Εικόνα 9. Μη γραμμικά διαχωρίσιμο πρόβλημα Non-linear decision boundary

3.1 Δέντρα Αποφάσεων

Μια ευρέως χρησιμοποιούμενη μέθοδος Μ.Μ. είναι και εκείνη που βασίζεται σε δένδρα απόφασης, κατά την οποία επιχειρείται η προσέγγιση μιας κατηγορικής συνάρτησης στόχου, ακολουθώντας την τεχνική του «διαίρει και βασίλευε» (Divide and Conquer). Ο χώρος του προβλήματος χωρίζεται σε περιοχές από στιγμιότυπα που φέρουν την ίδια τιμή ως προς κάποιο χαρακτηριστικό, και η διαδικασία επαναλαμβάνεται αναδρομικά, αναπαριστώντας με τον τρόπο αυτό το παραγόμενο μοντέλο ως δένδρο απόφασης.

Γενικά, ένα δένδρο απόφασης αντιπροσωπεύει μια σειρά από Αν-Τότε κανόνες (IF-THEN) που συνδυάζονται μεταξύ τους από τη ρίζα του δένδρου προς τα φύλλα. Οι κόμβοι του δένδρου χαρακτηρίζονται με τα ονόματα των χαρακτηριστικών, οι ακμές ονομάζονται με τις δυνατές τιμές που μπορεί να πάρει ένα χαρακτηριστικό και τα φύλλα με τις διάφορες κλάσεις. Τα αντικείμενα ταξινομούνται ακολουθώντας ένα μονοπάτι που οδηγεί προς τα κάτω στο δένδρο, λαμβάνοντας τις ακμές που αντιστοιχούν στις τιμές των χαρακτηριστικών ενός αντικειμένου.

Αναλυτικότερα για την ταξινόμηση, μία εγγραφή εισέρχεται στο δένδρο από τον κόμβο της κορυφής. Στην ρίζα, εφαρμόζεται έλεγχος για να καθορισθεί ποιο κόμβο παιδί θα ακολουθήσει στην συνέχεια η εγγραφή. Η επεξεργασία αυτή επαναλαμβάνεται μέχρι η εγγραφή να φτάσει στο κόμβο φύλλο. Όλες οι εγγραφές οι οποίες καταλήγουν σε ένα συγκεκριμένο φύλλο ταξινομούνται με τον ίδιο τρόπο. Υπάρχει ένα μοναδικό μονοπάτι που οδηγεί από την ρίζα σε κάθε φύλλο. Το μονοπάτι αυτό είναι μία έκφραση του κανόνα που χρησιμοποιείται για να ταξινομήσουμε τις εγγραφές.

Κατά την εκπαίδευση, το δένδρο χτίζεται με την επαναλαμβανόμενη διάσπαση του δοσμένου συνόλου δεδομένων σύμφωνα με τις διάφορες ανεξάρτητες μεταβλητές. Η «σειρά» με την οποία χρησιμοποιούνται οι ανεξάρτητες μεταβλητές στη δόμηση του δένδρου εξαρτάται από τη δυνατότητα ταξινόμησης της κάθε ανεξάρτητης μεταβλητής. Υπάρχουν διάφοροι αλγόριθμοι για την επιλογή της σειράς (Murthy, 1998), αλλά ο στόχος είναι πάντα ο ίδιος, δηλαδή, να επιλέξουμε την μεταβλητή εκείνη που διαχωρίζει καλύτερα τις τελικές κλάσεις. Ο αλγόριθμος σταματά όταν φτάσει σε κόμβο από τον οποίο δεν είναι δυνατό να ξεκινήσει μία νέα διάσπαση. Τότε ο κόμβος αυτός δεν έχει παιδιά και αποτελεί φύλλο του δένδρου.

Στην Εικόνα 10 παρουσιάζεται ένα δένδρο απόφασης το οποίο χτίστηκε χρησιμοποιώντας σαν δεδομένα εκπαίδευσης τα δεδομένα του παρακάτω πίνακα με εξαρτημένη μεταβλητή το «play».

outlook	humidity	play
sunny	high	no
sunny	high	no
overcast	high	yes
rainy	high	yes
rainy	normal	yes
rainy	normal	no
overcast	normal	yes
sunny	high	no
sunny	normal	yes
rainy	normal	yes
sunny	normal	yes
overcast	high	yes
overcast	normal	yes
rainy	high	no

Πίνακας 13. Δεδομένα εκπαίδευσης

Εικόνα 10. Δέντρο Απόφασης

Ως προς τον τρόπο ανάπτυξής τους, τα δένδρα αποφάσεων διακρίνονται σε

- Δυαδικά: από κάθε κόμβο διακλαδίζονται δύο νέοι κόμβοι
- Σύνθετα: από κάθε κόμβο διακλαδίζονται δύο ή περισσότεροι νέοι κόμβοι

Αναλυτικότερα η επιλογή της «σειράς» των μεταβλητών για την δημιουργία του δέντρου απόφασης παρουσιάζεται στην επόμενη υποενότητα.

3.1.1 Επιλογή μεταβλητών για την δημιουργία του δέντρου απόφασης

Κάθε κόμβος και προφανώς και η ρίζα του δέντρου θέτουν ένα ερώτημα: «Το χαρακτηριστικό A τι τιμή παίρνει;» Μια εύλογη απορία είναι λοιπόν πιο χαρακτηριστικό θα χρησιμοποιηθεί σε κάθε κόμβο. Ο τρόπος επιλογής του χαρακτηριστικού συχνά είναι η μόνη διαφορά μεταξύ των αλγορίθμων. Το κυριότερο μέτρο της απόδοσης ενός χαρακτηριστικού είναι η εντροπία του. Η τιμή της εντροπίας εκφράζει το κατά πόσο είναι διαταραγμένο το σύνολο εκπαίδευσης. Η μείωση της εντροπίας εκφράζει την μείωση της διατάραξης, δηλαδή κάποια μορφή τάξης που αρχίζει να επικρατεί στο σύνολο. Έτσι σκοπός των μεθόδων των δέντρων αποφάσεων είναι η μείωση της εντροπίας καθώς ακολουθούμε ένα μονοπάτι από τη ρίζα ως τα φύλλα ενός δέντρου.

Ορισμός εντροπίας: Έστω το σύνολο S των στιγμιοτύπων πάνω στα οποία θα γίνει η εκπαίδευση. Η εντροπία σε δυαδικά προβλήματα ορίζεται σαν το αντίθετο της πιθανότητας εμφάνισης ενός θετικού στιγμιότυπου (p_1) επί τον δυαδικό λογάριθμο της πιθανότητας αυτής μείον την πιθανότητα να εμφανιστεί ένα αρνητικό στιγμιότυπο (p_2) επί τον δυαδικό λογάριθμο αυτής της πιθανότητας:

$$Entropy(S) \equiv -p_1 \log_2 p_1 - p_2 \log_2 p_2.$$

Εύκολα επεκτείνεται ο ορισμός και για την περίπτωση ενός συνόλου όπου οι κλάσεις δεν είναι μόνο δύο. Έτσι ο ορισμός γίνεται: $Entropy(S) \equiv \sum_{i=1}^c -p_i \log_2 p_i$ όπου το c είναι το πλήθος του συνόλου των κλάσεων κατάταξης των στιγμιοτύπων.

Τα περισσότερα μέτρα επιλογής του κατάλληλου χαρακτηριστικού λαμβάνουν υπ' όψη τους την εντροπία του συνόλου εκπαίδευσης, αλλά αναλόγως και την εφαρμογή για την οποία κατασκευάστηκαν, επηρεάζονται και από άλλα χαρακτηριστικά του προβλήματος. Ένα πολύ διαδεδομένο είναι το μέτρο επιλογής - κέρδος (Gain) - από τη μείωση της εντροπίας που αποκτάμε εκχωρώντας το συγκεκριμένο χαρακτηριστικό στον κόμβο του δέντρου που αναπτύσσουμε. Το κέρδος για ένα χαρακτηριστικό A ενός συνόλου S δίνεται από τον τύπο

$$Gain(S, A) \equiv Entropy(S) - \sum_{u \in Values(A)} \frac{|S_u|}{|S|} Entropy(S_u) \text{ όπου } S_u = \{s \in S \mid A(s) = u\}.$$

Βάσει των δεδομένων εκπαίδευσης του Πίνακα 13, η μεταβλητή στόχος (play) έχει δύο τιμές, οι οποίες δεν είναι ισοπίθανες. Έχουμε 14 παραδείγματα εκπαίδευσης, από τα οποία τα 9 αντιστοιχούν στην τιμή play=yes και τα 5 στην τιμή play=no. Άρα η εκ των προτέρων πιθανότητα για την τιμή play=yes είναι $P(\text{yes})=0.64$, ενώ η πιθανότητα play=no είναι $P(\text{no})=0.36$.

Η εντροπία είναι:

$$\begin{aligned} Entropy &= -P(\text{no}) \cdot \log_2 P(\text{no}) - P(\text{yes}) \cdot \log_2 P(\text{yes}) = \\ &= -0.64 \cdot \log_2 0.64 - 0.36 \cdot \log_2 0.36 = 0.64 \cdot 0.64 + 0.36 \cdot 1.47 = 0.94 \end{aligned}$$

Για να αποφασίσουμε ποια θα είναι η μεταβλητή στη ρίζα του δέντρου, πρέπει να δούμε πόσο είναι το αναμενόμενο κέρδος ύστερα από κάθε πιθανή πρώτη ερώτηση.

Εάν η πρώτη ερώτησή μας αφορά το πεδίο «outlook», τότε υπάρχουν τρεις πιθανές απαντήσεις, οι sunny με πιθανότητα $P(\text{sunny})=5/14=0.36$, «overcast» με πιθανότητα $P(\text{overcast})=4/14=0.28$ και «rainy» με πιθανότητα $P(\text{rainy})=5/14=0.36$.

Στην περίπτωση που η απάντηση στην ερώτηση «outlook» είναι «sunny» έχουμε $3/5=0.6$ πιθανότητα η τελική απάντηση να είναι «no» και $2/5=0.4$ πιθανότητα η τελική απάντηση να είναι «yes». Η εντροπία σε αυτή την περίπτωση είναι:

$$\text{Entropy}_{\text{outlook=sunny}} = -0.6 \cdot \log_2 0.6 - 0.4 \cdot \log_2 0.4 = 0.6 \cdot 0.74 + 0.4 \cdot 1.32 = 0.97$$

Στην περίπτωση που η απάντηση στην ερώτηση «outlook» είναι «overcast» έχουμε $4/4=1$ πιθανότητα η τελική απάντηση να είναι «yes» και $0/4=0$ πιθανότητα η τελική απάντηση να είναι «no». Η εντροπία σε αυτή την περίπτωση είναι προφανές ότι είναι:

$$\text{Entropy}_{\text{outlook=overcast}} = 0$$

Στην περίπτωση που η απάντηση στην ερώτηση «outlook» είναι «rainy» έχουμε $3/5=0.6$ πιθανότητα η τελική απάντηση να είναι «yes» και $2/5=0.4$ πιθανότητα η τελική απάντηση να είναι «no». Η εντροπία σε αυτή την περίπτωση είναι (όπως ακριβώς και στην περίπτωση της sunny) είναι:

$$\text{Entropy}_{\text{outlook=rainy}} = 0.97 \text{ bits}$$

Άρα η εντροπία μετά την πρώτη ερώτηση, εφόσον αυτή αφορά το πεδίο «outlook», είναι:

$$\begin{aligned} \text{Entropy}_{\text{outlook}} &= P(\text{sunny}) \cdot \text{Entropy}_{\text{outlook=sunny}} + P(\text{overcast}) \cdot \text{Entropy}_{\text{outlook=overcast}} + P(\text{rainy}) \cdot \\ &\cdot \text{Entropy}_{\text{outlook=rainy}} = 0.36 \cdot 0.97 + 0.28 \cdot 0 + 0.36 \cdot 0.97 = 0.70 \end{aligned}$$

Το κέρδος από την ερώτηση «outlook» είναι λοιπόν:

$$\text{Gain}_{\text{outlook}} = \text{Entropy} - \text{Entropy}_{\text{outlook}} = 0.94 - 0.70 = 0.24$$

Εάν η πρώτη ερώτηση αφορά το πεδίο «humidity», υπάρχουν δύο δυνατές απαντήσεις, οι «high» με πιθανότητα $P(\text{high}) = 7/14 = 0.5$ και η «normal» με πιθανότητα $P(\text{normal}) = 7/14 = 0.5$. Στην περίπτωση που η απάντηση στην ερώτηση «humidity» είναι «high» έχουμε $4/7=0.57$ πιθανότητα η τελική απάντηση να είναι «no» και $3/7=0.43$ πιθανότητα η τελική απάντηση να είναι «yes». Η εντροπία σε αυτή την περίπτωση είναι:

$$\text{Entropy}_{\text{humidity=high}} = -0.57 \cdot \log_2 0.57 - 0.43 \cdot \log_2 0.43 = 0.57 \cdot 0.81 + 0.43 \cdot 1.22 = 0.99$$

Στην περίπτωση που η απάντηση στην ερώτηση «humidity» είναι «normal» έχουμε $6/7=0.86$ πιθανότητα η τελική απάντηση να είναι «yes» και $1/7=0.14$ πιθανότητα η τελική απάντηση να είναι «no». Η εντροπία σε αυτή την περίπτωση είναι:

$$\text{Entropy}_{\text{humidity=normal}} = -0.86 \cdot \log_2 0.86 - 0.14 \cdot \log_2 0.14 = 0.86 \cdot 0.22 + 0.14 \cdot 2.84 = 0.59$$

Άρα η εντροπία μετά την πρώτη ερώτηση, εφόσον αυτή αφορά το πεδίο «humidity», είναι:

$$\begin{aligned} \text{Entropy}_{\text{humidity}} &= P(\text{high}) \cdot I_{\text{humidity=high}} + P(\text{normal}) \cdot I_{\text{humidity=normal}} \\ &= 0.5 \cdot 0.99 + 0.5 \cdot 0.59 = 0.79 \end{aligned}$$

Το κέρδος από την ερώτηση «humidity» είναι:

$$\text{Gain}_{\text{humidity}} = \text{Entropy} - \text{Entropy}_{\text{humidity}} = 0.94 - 0.79 = 0.15$$

Από τα παραπάνω φαίνεται ότι η ερώτηση που θα πραγματοποιηθεί στη ρίζα του δένδρου αφορά τη μεταβλητή «outlook».

Οι ιδιότητες του κέρδους το καθιστούν εκ πρώτης όψεως ιδανικό ως μέτρο για την επιλογή του χαρακτηριστικού εκείνου που διαχωρίζει το χώρο των στιγμιότυπων κατά βέλτιστο τρόπο, σε κάθε εσωτερικό κόμβο του δένδρου. Στην πράξη όμως το κέρδος έχει προδιάθεση να επιλέγει χαρακτηριστικά με μεγάλο σύνολο τιμών, σε σχέση με άλλα που λαμβάνουν λιγότερες τιμές.

Ας υποθέσουμε ότι ένα χαρακτηριστικό A χωρίζει το σύνολο εκπαίδευσης σε πάρα πολλά υποσύνολα με πολύ μικρή εντροπία το καθένα. Αν χρησιμοποιούσαμε το κέρδος σαν μέτρο, το χαρακτηριστικό αυτό θα εκχωρούταν στην ρίζα και θα είχαμε ένα πάρα πολύ πλατύ δέντρο. Ένα τέτοιο δέντρο θα μπορούσε εύκολα να καταχωρήσει και με μεγάλη ακρίβεια οποιοδήποτε στιγμιότυπο του συνόλου εκπαίδευσης. Θα αποτύγχανε όμως σε ένα άλλο σύνολο πέραν του συνόλου εκπαίδευσης λόγω του υπερταϊριάσματος, καθώς το χαρακτηριστικό αυτό μπορεί να διαδραματίζει σημαντικό ρόλο για την κατάταξη του συνόλου εκπαίδευσης, όμως δεν μπορεί απαραίτητα να γενικευθεί και να καταχωρήσει ένα οποιοδήποτε στιγμιότυπο.

Προκειμένου να καταδείξουμε τον κίνδυνο που ενέχει η επιλογή ενός τέτοιου χαρακτηριστικού, ας θεωρήσουμε ένα πρόβλημα ταξινόμησης, τα στιγμιότυπα του οποίου προσδιορίζονται μοναδικά από την τιμή ενός χαρακτηριστικού, έστω του αύξοντα αριθμού (ID). Κατά το σχηματισμό του δένδρου απόφασης, το χαρακτηριστικό αυτό θα έχει τη μεγαλύτερη τιμή του κέρδους, καθώς θα διαχωρίσει το σώμα εκπαίδευσης σε ισάριθμα των στιγμιότυπων του μονοσύνολα, με την εντροπία καθενός από αυτά να ισούται με μηδέν. Γίνεται επομένως αντιληπτό ότι η παραγωγή ενός τέτοιου δένδρου (βάθους 1 και παράγοντα διακλάδωσης ισάριθμο των στιγμιότυπων εκπαίδευσης), δεν εξυπηρετεί κανένα σκοπό, καθώς το τελευταίο στερείται της απαιτούμενης ικανότητας γενίκευσης που θα του επέτρεπε την ορθή ταξινόμηση ενός άγνωστου στιγμιότυπου.

Την αδυναμία αυτή του κέρδους έρχεται να καλύψει ο Λόγος του Κέρδους Πληροφορίας (Information Gain Ratio – GR) (Quinlan, 1993), ο οποίος υιοθετεί το μέτρο της Πληροφορίας Διαχωρισμού (Split Information - SI), με σκοπό να αποθαρρύνει την επιλογή χαρακτηριστικών που παραπλανούν λόγω του μεγάλου κέρδους πληροφορίας τους, τείνοντας να δημιουργήσουν δένδρα με μεγάλο παράγοντα διακλάδωσης.

Έτσι ορίστηκε ένα καινούργιο μέγεθος, ο χωρισμός πληροφορίας που ορίζεται από τη σχέση

$$SplitInformation(S, A) \equiv - \sum_{i=1}^c \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Πρόκειται για ένα μέγεθος που εκφράζει το πόσο πολύ χωρίζεται ένα σύνολο από ένα χαρακτηριστικό. Προφανώς δεν μπορεί από μόνο του να εκφράσει το συνολικό κέρδος που μας δίνει ένα χαρακτηριστικό. Έτσι έχουμε το λόγο πληροφορίας που δίνεται από τον τύπο

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)} \text{ που συνδυάζει το κέρδος που προαναφέραμε με}$$

το χωρισμό της πληροφορίας.

Το μέτρο του Λόγου του Κέρδους Πληροφορίας (Gain Ratio) εγγυάται αποτελεσματική διαμέριση του χώρου στην περίπτωση που η πληθικότητα των υποσυνόλων διαφέρει μεταξύ τους. Σε αντίθετη περίπτωση, η τιμή του μέτρου Πληροφορίας Διαχωρισμού γίνεται πολύ μικρή με αποτέλεσμα ο Λόγος του Κέρδους Πληροφορίας να καθίσταται απροσδιόριστος ή πολύ μεγάλος. Μια ευριστική μέθοδος που μετριάζει την παρενέργεια αυτή, υπολογίζει το μέτρο του Λόγου του Κέρδους Πληροφορίας των χαρακτηριστικών εκείνων που η τιμή του κέρδους είναι μεγαλύτερη της μέσης τιμής του μέτρου αυτού όλων των χαρακτηριστικών.

Ο πιο γνωστός αλγόριθμος για την κατασκευή δέντρων απόφασης που χρησιμοποιεί το Λόγο του Κέρδους Πληροφορίας είναι ο C4.5 (Quinlan, 1993). Μια από τις πιο πρόσφατες έρευνες που συγκρίνουν τα δέντρα απόφασης με άλλους αλγορίθμους εκμάθησης (Sien Lim και

άλλοι, 2000) έδειξε ότι ο C4.5 έχει έναν πολύ καλό συνδυασμό ακρίβειας και ταχύτητας εκπαίδευσης. Στη συνέχεια, παρουσιάζουμε σε ψευδοκώδικα τον αλγόριθμο C4.5 (Εικόνα 11), ο οποίος χρησιμοποιήθηκε ως εκπρόσωπος των δέντρων απόφασης στο πειραματικό μέρος της διατριβής.

1. Επέλεξε το χαρακτηριστικό που πετυχαίνει τον καλύτερο διαχωρισμό μεταξύ των κατηγοριών χρησιμοποιώντας το μέτρο του Λόγου του Κέρδους Πληροφορίας.
2. Χώρισε τα δεδομένα σε υποσύνολα με βάση της τιμές του χαρακτηριστικού αυτού.
3. Για κάθε υποσύνολο που περιέχει περισσότερες από μία κατηγορίες, επανέλαβε τη διαδικασία.
4. Σταμάτησε εφόσον δεν υπάρχουν υποσύνολα που περιέχουν περισσότερες από μία κατηγορίες ή έχουν χρησιμοποιηθεί όλα τα χαρακτηριστικά.

Εικόνα 11. Ο αλγόριθμος C4.5

Η παρουσίαση του αλγορίθμου που ως τώρα αναπτύξαμε, προϋποθέτει τη χρήση κατηγορικών χαρακτηριστικών. Ο αλγόριθμος C4.5 ωστόσο και οι διάφορες επεκτάσεις του, έχουν τη δυνατότητα να διαχειριστούν και συνεχή χαρακτηριστικά, εφαρμόζοντας στην αρχή κάθε αναδρομικού βήματος μια διαδικασία μετατροπής τους σε ένα σύνολο διακριτών λογικών (Boolean) χαρακτηριστικών, γνωστή ως Διακριτοποίηση (Discretization). Η διαδικασία αυτή ξεκινά με την διάταξη των στιγμιότυπων του υποσυνόλου του σώματος εκπαίδευσης που πρόκειται να χρησιμοποιηθούν στο τρέχον βήμα κατά αύξουσα σειρά, βάσει της τιμής που φέρουν για ένα συνεχές χαρακτηριστικό a_i . Με αυτό τον τρόπο εντοπίζονται όλα τα διαδοχικά στιγμιότυπα x_k , x_l τα οποία ανήκουν σε διαφορετική μεταξύ τους κλάση. Η τιμή του a_i στο δεύτερο κατά σειρά στιγμιότυπο κάθε τέτοιου ζεύγους ανατίθεται σε ένα κατώφλι (threshold), t_{ij} , με βάση το οποίο ορίζεται ένα λογικό χαρακτηριστικό a_{ij} . Το νέο χαρακτηριστικό παίρνει την τιμή 1 αν η τιμή του αρχικού a_i υπερβαίνει την τιμή του κατωφλίου t_{ij} , διαφορετικά η τιμή του είναι 0. Έτσι το χαρακτηριστικό a_i αντικαθίσταται από ένα σύνολο λογικών χαρακτηριστικών $a_{i1}, a_{i2}, \dots, a_{im}$ με αντίστοιχα κατώφλια.

Ο αλγόριθμος δημιουργίας του δέντρου σταματά όταν δεν είναι δυνατή η παραγωγή δύο νέων κλαδιών που το καθένα να περιέχει αριθμό εγγραφών ίσο ή μεγαλύτερο από έναν ελάχιστο αριθμό παρατηρήσεων σε κάθε κλαδί, τον οποίο έχει καθορίσει ο χρήστης στην αρχή της διαδικασίας.

Ένα πρόβλημα που καλούνται να λύσουν οι αλγόριθμοι των δέντρων απόφασης είναι το υπερταίριασμα (Murthy, 1998) (βλέπε Εικόνα 12).

Εικόνα 12. Το φαινόμενο του υπερταίριασματος

3.1.2 Αντιμετώπιση του υπερταίριασματος στα δέντρα αποφάσεων

Υπάρχουν δύο κοινές προσεγγίσεις που οι αλγόριθμοι επαγωγής δέντρων απόφασης χρησιμοποιούν για να αποφύγουν το υπερταίριασμα: i) Τερματίζοντας τον αλγόριθμο κατάρτισης προτού φθάσει σε ένα δέντρο το οποίο ταιριάζει τέλεια στα στοιχεία κατάρτισης, ii) περικόπτοντας (κλαδεύοντας) το δέντρο απόφασης εκ των υστέρων.

Ο απλούστερος τρόπος είναι να μην αφεθεί το δέντρο να αυξηθεί στο πλήρες μέγεθός του (Cherkauer & Shavlik, 1996). Η επιλογή ενός κριτηρίου λήξης όπως ο ορισμός κατώτατου ορίου για την ποιότητα των χαρακτηριστικών μπορεί να χρησιμοποιηθεί σε αυτή την περίπτωση. Όμως, οι ταξινομητές δέντρων απόφασης, υιοθετούν συνήθως τεχνικές περικοπής που αξιολογούν την απόδοση των δέντρων απόφασης κλαδεύοντας το δέντρο χρησιμοποιώντας ένα σύνολο επικύρωσης (βλέπε Εικόνα 13). Ο Elomaa (1999) έπειτα από ένα πλήθος συγκρίσεων κατέληξε στο συμπέρασμα ότι δεν υπάρχει μέθοδος περικοπής που να λειτουργεί αποτελεσματικότερα σε όλα τα προβλήματα.

Εικόνα 13. Κλάδεμα δέντρου απόφασης

Ολοκληρώνοντας την παρουσίαση του αλγορίθμου C4.5, κρίνεται σκόπιμη η σύντομη αναφορά μας στη μέθοδο που χρησιμοποιεί για την αποφυγή του φαινομένου του υπερταίριασματος στα δεδομένα εκπαίδευσης. Η μέθοδος που χρησιμοποιεί ο C4.5 βασίζεται στο «κλάδεμα» του δέντρου, αφού αυτό έχει φθάσει στην πλήρη ανάπτυξή του. Δύο από τις χρησιμοποιούμενες προσεγγίσεις για κλάδεμα είναι αυτές της Αντικατάστασης Υποδέντρου (Subtree Replacement) και της Ανύψωσης Υποδέντρου (Subtree Raising). Σε κάθε κόμβο του δέντρου, ο αλγόριθμος καλείται να αποφασίσει αν θα αφήσει το υποδέντρο που ξεκινά απ' αυτόν ως έχει, αν θα το αντικαταστήσει με έναν κόμβο φύλλο (Αντικατάσταση Υποδέντρου)

ή αν θα τον αντικαταστήσει με κάποιο από τους άμεσους απογόνους του, αναδιατάσσοντας τα στιγμιότυπα του αρχικού υποδένδρου στο νέο που μόλις δημιουργήθηκε (Ανύψωση Υποδένδρου).

Σε κάθε περίπτωση, απαιτείται ο προσδιορισμός του αναμενόμενου σφάλματος ταξινόμησης που ενδέχεται να οδηγήσει κάθε μια εκ των αποφάσεων. Μια τεχνική που εφαρμόζεται για το σκοπό αυτό, το Κλάδεμα Μειωμένου Σφάλματος (Reduced Error Pruning):

$$\frac{err'(t)}{data(t)} - \frac{err'(T_t)}{data(T_t)} \leq \sqrt{\frac{err'(T_t) \cdot (data(t) - err'(T_t))}{data(t)}}$$

Όπου:

το υπο-δέντρο T_t αντικαθιστάται από τη ρίζα του t , που γίνεται φύλλο,

$$err'(t) = err(t) + 0.5$$

$$err'(T_t) = \sum_{s \in leaves(T_t)} err(s) + 0.5$$

και $err(t)$: αριθμός λαθών στο t , $data(t)$: αριθμός παραδειγμάτων στο t .

Αυτή η τεχνική προϋποθέτει τη χρήση ενός μέρους του σώματος εκπαίδευσης αποκλειστικά για τη διεξαγωγή του παραπάνω ελέγχου, με το τίμημα βέβαια το δένδρο που θα αναπτυχθεί να βασίζεται σε λιγότερα δεδομένα εκπαίδευσης. Υπάρχουν βέβαια και άλλες εναλλακτικές προσεγγίσεις που κάνουν χρήση στατιστικών μεθόδων (Breslow & Aha, 1997).

Αν και οι αλγόριθμοι κατασκευής δέντρων απόφασης μπορούν να διαχειριστούν συνεχείς μεταβλητές, οι Elomaa & Rousu (1999) απέδειξαν ότι χρησιμοποιώντας δυαδική διακριτοποίηση (discretization) με τον C4.5 αλγόριθμο επιτυγχάνεται κατά μέσον όρο διπλάσια ταχύτητα από την χρησιμοποιούμενη πολυ-διακριτοποίησης. Επιπλέον, σύμφωνα με τα πειράματά τους η πολυ-διακριτοποίηση των αριθμητικών χαρακτηριστικών δεν επιφέρει οποιοδήποτε αύξηση στην ακρίβεια πρόβλεψης σε σχέση με τον δυαδική διακριτοποίηση.

Σε αυτό το σημείο πρέπει να τονιστεί ότι η διαμέριση του πληθυσμού σε ένα κόμβο του δένδρου απόφασης γίνεται χωρίς να λαμβάνεται υπόψη η επίδραση που αυτή μπορεί να έχει σε επόμενες διαμερίσεις. Μετά τη διαμέριση του πληθυσμού σε ένα κόμβο, το δένδρο αναπτύσσεται χωρίς να επανεξετασθεί ο κόμβος. Επιπλέον, οι διαμερίσεις γίνονται σειριακά (κάθε διαμέριση εξαρτάται από την προηγούμενή της και όλες βεβαίως από την πρώτη διαμέριση στην ρίζα του δένδρου).

Πρόσφατα, ο Ruggieri (2001) παρουσίασε μια αναλυτική αξιολόγηση του χρόνου εκτέλεσης του C4.5 αλγορίθμου. Με βάση αυτή την αναλυτική αξιολόγηση, υλοποίησε μια αποδοτικότερη έκδοση του αλγορίθμου, αποκαλούμενη EC4.5. Υποστηρίζει ότι η εφαρμογή του συγκεκριμένου αλγορίθμου υπολογίζει τα ίδια δέντρα απόφασης με τον C4.5 χρησιμοποιώντας έως και πέντε φορές λιγότερο χρόνο για εκπαίδευση.

Οι περισσότεροι αλγόριθμοι δέντρων απόφασης χρησιμοποιούν μια μόνο ανεξάρτητη μεταβλητή κάθε φορά για τη διαμόρφωση του κανόνα διαμέρισης (univariate split). Ένα πρόβλημα που προκύπτει σε αυτή την περίπτωση είναι η ορθογώνια κατηγοριοποίηση του χώρου (βλέπε Εικόνα 14).

Εικόνα 14. Ορθογώνια κατηγοριοποίηση του χώρου

Υπάρχουν βέβαια ειδικοί αλγόριθμοι οι οποίοι επιτρέπουν την διαμόρφωση κανόνων διαμέρισης με περισσότερες από μια ανεξάρτητες μεταβλητές (multivariate split). Παραδείγματος χάριν, ο Zheng (1998) βελτιώνει την ακρίβεια ταξινόμησης των δέντρων απόφασης με την κατασκευή νέων δυαδικών χαρακτηριστικών γνωρισμάτων συνδυάζοντας τα υπάρχοντα χρησιμοποιώντας λογικούς τελεστές όπως η ένωση, η άρνηση, και η τομή. Επιπλέον, ο Zheng (2000) πρότεινε την δημιουργία νέων χαρακτηριστικών, με την υπόθεση ότι η τιμή της νέας μεταβλητής είναι θετική εάν τουλάχιστον το μ από τα n χαρακτηριστικά είναι θετικά. Επίσης, οι Gama και Brazdil (1999) κατασκεύασαν ένα δέντρο απόφασης χρησιμοποιώντας νέα χαρακτηριστικά που υπολογίζονται ως γραμμικοί συνδυασμοί των προηγούμενων.

Οι Markovitch και Rosenstein (2002) παρουσίασαν τον αλγόριθμο FICUS, ο οποίος λαμβάνει τα αρχικά δεδομένα εκπαίδευσης και τα χρησιμοποιεί για να παράγει ένα σύνολο από παραγόμενα χαρακτηριστικά. Ενώ ο αλγόριθμος FICUS είναι παρόμοιος σε μερικά σημεία με άλλους αλγορίθμους κατασκευής χαρακτηριστικών, η κύρια δύναμή του είναι η γενικότητά του και η ευελιξία. Ο FICUS είναι σχεδιασμένος να εκτελεί την παραγωγή χαρακτηριστικών σύμφωνα με μια γραμματική παραγωγής γενικού σκοπού.

Πρέπει επίσης να σημειωθεί ότι ο C4.5 υποθέτει ότι τα δεδομένα κατάρτισης χωρούν στη μνήμη, γι' αυτό το λόγο ο Gehrke και άλλοι (1998) πρότειναν το Rainforest, ένα πλαίσιο για την ανάπτυξη γρήγορων αλγορίθμων για την κατασκευή δέντρων απόφασης που προσαρμόζονται στο ποσό της διαθέσιμης κύριας μνήμης.

Επίσης, είναι γεγονός ότι στους περισσότερους αλγορίθμους των δέντρων απόφασης ουσιαστικός χρόνος σπαταλιέται στη διάρκεια της κατασκευής των κλαδιών του δέντρου που στη συνέχεια θα κοπούν κατά τη φάση της περικοπής. Γι' αυτό οι Rastogi & Shim (2000) πρότειναν τον PUBLIC, ένα βελτιωμένο αλγόριθμο δέντρων απόφασης που ολοκληρώνει τη δεύτερη φάση περικοπής κατά την αρχική κατασκευαστική φάση. Στον αλγόριθμο PUBLIC, ένας κόμβος δεν επεκτείνεται κατά την κατασκευαστική φάση εάν είναι αναμενόμενο ότι θα απομακρυνθεί στην φάση της περικοπής.

Παράλληλα, ο Zaki και άλλοι (1999) πέτυχαν τον παραλληλισμό της αναζήτησης χώρου ενός αλγορίθμου δέντρου απόφασης. Εκμεταλλευόμενοι ένα κοινό πολυεπεξεργαστή είναι σε θέση να αποφύγουν την αντιγραφή ή την επικοινωνία όλου του συνόλου δεδομένων μεταξύ των επεξεργαστών. Άλλη μια παράλληλη υλοποίηση έχει προταθεί από τους Srivastava και άλλους (1999). Τέλος, οι Baik & Bala (2004) πρότειναν έναν αλγόριθμο κατασκευής δέντρων απόφασης με δυνατότητα εκπαίδευσης από κατανεμημένα δεδομένα.

Στο επόμενο υποκεφάλαιο παραθέτουμε την σύγκριση γνωστών αλγορίθμων κατασκευής δέντρων απόφασης σε ένα πλήθος από δεδομένα εκπαίδευσης⁷.

3.1.3 Σύγκριση αλγορίθμων κατασκευής δέντρων απόφασης

Υπάρχει μια πληθώρα αλγορίθμων που αναλαμβάνουν να φτιάξουν δένδρα αποφάσεων. Ο αλγόριθμος Decision stump (Iba & Langley, 1992) είναι στην ουσία ένα εκφυλισμένο δέντρο απόφασης. Ένα μόνο χαρακτηριστικό χρησιμοποιείται για την ταξινόμηση. Ο αλγόριθμος NBTree (Kohavi, 1996) χτίζει ένα δέντρο απόφασης με την διαφορά ότι στα φύλλα του δέντρου δεν αποδίδεται η τιμή ταξινόμησης αλλά εκπαιδεύεται ένας απλός Μπεϋζιανός ταξινομητής⁸. Ο αλγόριθμος RandomTree (Breiman, 2001) χτίζει ένα δέντρο απόφασης χωρίς κλάδεμα διαλέγοντας τυχαία σε κάθε βήμα την μεταβλητή κόμβο. Ο αλγόριθμος LMT (Landwehr και άλλοι, 2003) χτίζει ένα δέντρο απόφασης με την διαφορά ότι στα φύλλα του δέντρου δεν αποδίδεται η τιμή ταξινόμησης αλλά χρησιμοποιείται λογιστική παλινδρόμηση. Βάσει των πειραμάτων μας (Πίνακας 14) οι αλγόριθμοι NBTree και LMT σε ορισμένες περιπτώσεις δίνουν καλύτερα αποτελέσματα από τον C4.5 αλλά χρειάζονται περισσότερο χρόνο για εκπαίδευση αφού είναι υβριδικοί. Θέματα πολυπλοκότητας των αλγορίθμων δέντρων απόφασης αναλύονται στην εργασία των (Buhman & Wolf, 2002).

Συνοπτικά, στα θετικά σημεία των δέντρων απόφασης συγκαταλέγονται:

- Η ευρωστία που επιδεικνύει αναφορικά με το θόρυβο που ενδέχεται να υπάρχει στα δεδομένα εκπαίδευσης.
- Η ανοχή στην απουσία τιμών (missing values), σε κάποια χαρακτηριστικά του σώματος εκπαίδευσης.
- Η χρήση ακόμα και συνεχών (μη διακριτών) χαρακτηριστικών και η προσέγγιση μη διακριτών συναρτήσεων στόχου, μέσω εξειδικευμένων τεχνικών που αναλαμβάνουν τη διακριτοποίησή τους, τη διαδικασία δηλαδή της μετατροπής συνεχών αριθμητικών χαρακτηριστικών σε κατηγορικά. Αυτά τα δένδρα ονομάζονται και δέντρα παλινδρόμησης (regression trees) και θα αναλυθούν αργότερα.
- Η δυνατότητα μεταφοράς του παραγόμενου μοντέλου από δένδρο απόφασης σε ένα σύνολο κανόνων συμπερασμού (if – then rules), προς διευκόλυνση της κατανόησής του (βλέπε Εικόνα 15).

⁷ Περισσότερα λεπτομέρια για τα σύνολα δεδομένων εκπαίδευσης στο Παράρτημα Α.

⁸ Περισσότερα στοιχεία για τον απλό Μπεϋζιανό ταξινομητή στην ενότητα 3.5.2

Δεδομένα	C4.5	RandomTree	NBTree	LMT	Decision stump
autos	72.94	64.42	74.40	72.23	44.09
badges	100.00	86.79	100.00	100.00	100.00
balance-scale	78.18	77.10	75.96	89.87	58.48
breast-cancer	72.03	68.71	70.89	71.63	68.74
breast-w	94.37	93.66	95.97	95.88	91.30
colic	84.04	71.85	81.24	82.12	80.03
Credit_g	72.59	65.71	73.35	75.06	70.00
credit-a	85.49	74.45	85.49	85.33	85.79
diabetes	72.86	67.73	74.77	77.76	71.51
haberman	71.85	64.63	71.38	73.58	72.62
heart-c	77.66	71.66	81.07	83.78	74.36
heart-h	77.72	75.22	81.41	82.91	79.52
heart-statlog	77.10	74.03	79.39	83.20	71.54
hepatitis	78.67	77.89	80.93	80.19	79.04
ionosphere	87.83	86.41	90.01	90.44	82.63
iris	94.90	93.33	93.73	95.10	66.67
labor	79.98	80.37	85.51	92.20	82.51
lymphography	75.15	69.02	78.99	81.52	73.00
monk1	72.68	67.90	84.80	70.07	72.93
monk2	56.16	53.71	62.76	61.53	60.68
monk3	89.90	67.09	93.74	94.47	73.81
relation	78.51	78.74	78.05	78.67	77.81
sonar	72.46	66.92	73.30	75.42	70.60
students	87.20	75.34	85.39	88.14	88.32
vehicle	69.45	62.74	70.35	82.50	40.78
vote	95.13	89.79	93.37	95.60	95.47
wine	90.76	85.50	95.72	96.37	59.64
zoo	92.33	83.04	94.73	93.24	60.22

Πίνακας 14. Σύγκριση αλγορίθμων κατασκευής δέντρων απόφασης

Rule Set

- r1: (P=No,Q=No) ==> -
- r2: (P=No,Q=Yes) ==> +
- r3: (P=Yes,R=No) ==> +
- r4: (P=Yes,R=Yes,Q=No) ==> -
- r5: (P=Yes,R=Yes,Q=Yes) ==> +

Εικόνα 15. Παραγωγή κανόνων από ένα δέντρο απόφασης

3.2 Κανόνες ταξινόμησης

Μια συγγενική μεθοδολογία επαγωγικής κατασκευής ταξινομητών με αυτή των δένδρων απόφασης αποτελεί η αυτόματη εκμάθηση κανόνων, χαρακτηριστική για την ικανότητά της να παράγει ιδιαίτερα εύληπτα μοντέλα με τη μορφή κανόνων συμπερασμού (if-then rules).

Υπάρχουν πολλές περιπτώσεις που η παραγωγή κανόνων είναι περισσότερο επιθυμητή από την παραγωγή οποιοδήποτε άλλου μοντέλου. Για παράδειγμα στην περίπτωση της πολιτικής πίστωσης, η τράπεζα θα ήθελε να γνωρίζει πότε μπορεί να δίνει δάνειο σε κάποιο πελάτη χωρίς να λαμβάνει μεγάλο ρίσκο. Εφαρμόζοντας τη διαδικασία της ταξινόμησης το πρόβλημα θα μπορούσε να οριστεί ως εξής:

- Οι πελάτες χωρίζονται στις κλάσεις: “άριστος”, “καλός”, “μέτριος” και “κακός” ανάλογα με την πιθανότητα κάθε πελάτη να εξοφλήσει το δάνειο.
- Κάθε πελάτης χαρακτηρίζεται από την ηλικία του, την εκπαίδευσή του, το ετήσιο εισόδημά του, κ.α.
- Υπάρχουν δεδομένα πελατών που έχουν δανειστεί από την τράπεζα στο παρελθόν.

Ένα ενδεχόμενο αποτέλεσμα της ταξινόμησης για τις κλάσεις “άριστος” και “καλός” θα μπορούσε να είναι το ακόλουθο:

- Για κάθε πελάτη P , με $P.πτυχίο = \text{μεταπτυχιακό}$ και $P.εισόδημα > 30.000 \text{ €}$ $\rightarrow P.κλάση = \text{άριστος}$
- Για κάθε πελάτη P , με $P.πτυχίο = \text{πτυχίο πανεπιστημίου}$ και $P.εισόδημα \geq 25.000 \text{ €}$ $\rightarrow P.κλάση = \text{καλός}$

Αξιοποιώντας τέτοιους κανόνες ταξινόμησης η τράπεζα αναμένεται να μειώσει το ρίσκο ο πελάτης στον οποίο χορήγησε κάποιο δάνειο να είναι ασυνεπής ως προς την εξόφληση του δανείου.

Ένα άλλο παράδειγμα είναι η περίπτωση οργάνωσης μιας διαφημιστικής καμπάνιας. Η κάθε εταιρία θα ήθελε να γνωρίζει ποιοι πελάτες είναι πιο πιθανό να απαντήσουν θετικά στην καμπάνια. Στόχος της κάθε εταιρίας είναι να προωθήσει την καμπάνια μόνο σε (πιθανά) ενδιαφερόμενα άτομα μειώνοντας έτσι το συνολικό κόστος. Εφαρμόζοντας τη διαδικασία της ταξινόμησης το πρόβλημα ορίζεται ως εξής:

- Οι πελάτες χωρίζονται στις κλάσεις: θετικοί και αρνητικοί αποδέκτες διαφημιστικών φυλλαδίων.
- Κάθε πελάτης χαρακτηρίζεται από το όνομά του, την ηλικία του, το επάγγελμά του κ.α.
- Υπάρχουν δεδομένα πελατών που είχαν απαντήσει σε παλαιότερες διαφημιστικές καμπάνιες της εταιρίας.

Ένα ενδεχόμενο αποτέλεσμα ταξινόμησης για την κλάση των θετικών αποδεκτών διαφημιστικών φυλλαδίων θα μπορούσε να είναι το ακόλουθο:

Για κάθε πελάτη P , με $(P.ηλικία > 25 \text{ και } P.ηλικία < 55)$ και $Περιοχή = N. Προάστια \rightarrow P.κλάση = \text{θετικός αποδέκτης}$

Η εταιρία συνεπώς θα μπορούσε να αξιοποιήσει τον κανόνα αυτό αποστέλλοντας το νέο διαφημιστικό υλικό μόνο στους θετικούς αποδέκτες μειώνοντας έτσι το συνολικό κόστος της διαφημιστικής καμπάνιας.

Όπως ήδη αναφέραμε, κανόνες συμπερασμού μπορούν εύκολα να προκύψουν από την εκμάθηση ενός δένδρου απόφασης και τη μετατροπή του σε κανόνες προτασιακής λογικής. Μια εναλλακτική προσέγγιση, ικανή να δώσει ένα κατά πολύ απλούστερο σύνολο κανόνων από την προηγούμενη, βασίζεται στη στρατηγική της εκμάθησης ενός κανόνα και της

απομάκρυνσης από το σώμα εκπαίδευσης όλων των στιγμιοτύπων τα οποία καλύπτει (βλέπε την ακολουθία βημάτων στην Εικόνα 16). Η ίδια διαδικασία επαναλαμβάνεται μέχρις ότου η ακρίβεια του συστήματος φθάσει μια επιθυμητή τιμή. Για την αποφυγή του φαινομένου του υπερταϊσμού, η τιμή αυτή για την ακρίβεια ταξινόμησης των στιγμιοτύπων εκπαίδευσης είναι πρέπει να είναι συνήθως χαμηλότερη του 100%. Οι αλγόριθμοι που ακολουθούν την παραπάνω προσέγγιση ονομάζονται Ακολουθιακοί Αλγόριθμοι Κάλυψης (Sequential Covering Algorithms). Ο Furnkranz (1999) παρέχει μια άριστη ανασκόπηση της υπάρχουσας βιβλιογραφίας για τις μεθόδους παραγωγής κανόνων ταξινόμησης.

Εικόνα 16. Σχηματική περιγραφή των αλγορίθμων ακολουθιακής κάλυψης

Βασική διαφορά αυτών των αλγορίθμων ακολουθιακής κάλυψης από τους υπόλοιπους αλγορίθμους μηχανικής μάθησης είναι πως δεν προσπαθούν να κατασκευάσουν έναν ταξινομητή που να μπορεί να επαληθεύσει ολόκληρο το σύνολο εκπαίδευσης, αλλά ένα όσο πιο μεγάλο μέρος αυτού. Η λογική τους είναι να κατασκευάσουν έναν κανόνα που να μπορεί να επαληθεύει ένα μεγάλο σύνολο στιγμιοτύπων του συνόλου εκπαίδευσης. Αναζητείται

λοιπόν, ένας κανόνας που θα έχει μεγάλη ακρίβεια αλλά όχι απαραίτητα μεγάλη κάλυψη. Δηλαδή θα πρέπει να μπορεί να έχει πολύ μεγάλη ακρίβεια πάνω στα στιγμιότυπα που περιγράφονται από χαρακτηριστικά που ο κανόνας αυτός προβλέπει, αλλά δεν είναι απαραίτητο να μπορεί να έχει εφαρμογή σε άλλα στιγμιότυπα. Στη συνέχεια αφαιρεί από το σύνολο εκπαίδευσης τα στιγμιότυπα τα οποία καλύπτει ο κανόνας που μόλις κατασκευάστηκε και επαναλαμβάνει τη διαδικασία ωσότου κατασκευάσει κανόνες που να καλύπτουν και το υπόλοιπο σύνολο εκπαίδευσης.

Οι κανόνες ταξινόμησης αντιπροσωπεύουν κάθε κατηγορία χρησιμοποιώντας την διαζευκτική κανονική μορφή (DNF). Ο στόχος είναι να κατασκευαστεί το μικρότερο σύνολο κανόνων που είναι σύμφωνο με τα στοιχεία κατάρτισης (δεδομένα εκπαίδευσης). Ο μεγάλος αριθμός κανόνων ταξινόμησης είναι συνήθως ένα σημάδι ότι ο αλγόριθμος μάθησης προσπαθεί να θυμηθεί το σύνολο κατάρτισης, αντί να ανακαλύπτει τις υποθέσεις που το κυβερνούν.

Ένας γενικός αλγόριθμος για δημιουργία κανόνων ταξινόμησης δίνεται παρακάτω σε ψευδογλώσσα:

function ΜάθησηΚανόνων(Στόχος, Μεταβλητές, Παραδείγματα, Κατώφλι):

Κανόνες := $\emptyset$

Κανόνας := ΜάθησηΚανόνα(Στόχος, Μεταβλητές, Παραδείγματα)

while Απόδοση(Κανόνας, Παραδείγματα) > Κατώφλι, **do**

Κανόνες := Κανόνες $\cup$ {Κανόνας}

Παραδείγματα := Παραδείγματα - {Παραδείγματα που ταξινομήθηκαν σωστά από τον κανόνα}

Κανόνας := ΜάθησηΚανόνα (Στόχος, Μεταβλητές, Παραδείγματα)

ταξινόμησε τους Κανόνες σύμφωνα με την Απόδοση τους

return Κανόνες

Εικόνα 17. Αλγόριθμος παραγωγής κανόνων ταξινόμησης

Ο τρόπος επιλογής των χαρακτηριστικών συχνά είναι μια διαφορά μεταξύ των αλγορίθμων παραγωγής κανόνων ταξινόμησης. Και για αυτούς τους αλγόριθμους το κυριότερο μέτρο της απόδοσης ενός χαρακτηριστικού είναι η εντροπία του. Πληθώρα άλλων μέτρων έχουν χρησιμοποιηθεί (Lavrac, 1999). Οι An & Cercone (2001) μετά από ένα σύνολο πειραμάτων απέδειξαν ότι δεν υπάρχει ένα μέτρο που να λειτουργεί καλύτερα σε όλες τις περιπτώσεις.

Ο παραπάνω αλγόριθμος (Εικόνα 17), όπως και οι υπόλοιποι που εξάγουν άμεσα κανόνες για την ταξινόμηση στιγμιότυπων, μαθαίνουν έναν – έναν τους κανόνες σε αντίθεση με τους αλγόριθμους δέντρων αποφάσεων, οι οποίοι μαθαίνουν πολλούς κανόνες ταυτόχρονα. Γι' αυτό το λόγο οι τελευταίοι καλούνται και αλγόριθμοι ταυτόχρονης κάλυψης. Το ερώτημα που εγείρεται είναι πιο είδος προτιμάται σε κάθε περίπτωση. Για την επιλογή του καταλληλότερου αλγορίθμου πρέπει να ληφθεί υπ' όψη το πλήθος των δεδομένων του συνόλου εκπαίδευσης. Αν αυτά είναι αρκετά τότε μπορεί κανείς να χρησιμοποιήσει αλγορίθμους ακολουθιακής κάλυψης. Αντιθέτως, αν υπάρχουν λίγα δεδομένα ο «διαμοιρασμός» των αποφάσεων μπορεί να αποδειχθεί πιο αποτελεσματικός. Οι αλγόριθμοι ακολουθιακής κάλυψης κάνουν αρκετά περισσότερους υπολογισμούς από αυτούς της ταυτόχρονης κάλυψης. Επιπλέον, θα πρέπει να ληφθεί υπ' όψη αν έχει νόημα η επανεξέταση του ίδιου χαρακτηριστικού για διαφορετικούς κανόνες όπως γίνεται στους αλγόριθμους ταυτόχρονης κάλυψης σε αντίθεση με αυτούς της ακολουθιακής.

Όπως και στα δέντρα αποφάσεων έτσι και στους κανόνες ταξινόμησης υπάρχουν δύο κοινές προσεγγίσεις που οι αλγόριθμοι χρησιμοποιούν για να αποφύγουν το υπερταίριασμα (Furnkranz, 1997): i) Τερματίζοντας τον αλγόριθμο κατάρτισης προτού φθάσει σε ένα σύνολο κανόνων το οποίο ταιριάζει τέλεια στα στοιχεία κατάρτισης, ii) περικόπτοντας περιττούς κανόνες εκ των υστέρων.

Ο Bramer (2002) παρατήρησε ότι η παρουσία θορύβου στο σύνολο εκπαίδευσης κάνει τους αλγόριθμους να παράγουν μεγαλύτερο πλήθος κανόνων ταξινόμησης από ότι χρειάζεται. Πρότεινε μια μέθοδο περικοπής κανόνων (J-pruning) που βελτιώνει τα αποτελέσματα.

Ο RIPPER είναι ένας πολύ γνωστός αλγόριθμος παραγωγής κανόνων ταξινόμησης (Cohen, 1995). Διαμορφώνει τους κανόνες μέσω μιας διαδικασίας επαναλαμβανόμενης ανάπτυξης και περικοπής. Κατά τη διάρκεια της αυξανόμενης φάσης οι κανόνες γίνονται πιο περιοριστικοί προκειμένου να ταιριάζουν με τα στοιχεία κατάρτισης όσο το δυνατόν περισσότερο. Κατά τη διάρκεια της φάσης περικοπής, οι κανόνες γίνονται λιγότερο περιοριστικοί για να αποφύγουν το υπερταίριασμα, το οποίο μπορεί να προκαλέσει την κακή απόδοση σε νέες περιπτώσεις (καινούρια στιγμιότυπα).

Υπάρχουν πολυάριθμοι άλλοι αλγόριθμοι εκμάθησης βασισμένοι στους κανόνες. Μια πιο εξελιγμένη μορφή των αλγορίθμων ακολουθιακής κάλυψης είναι να οργανωθεί η αναζήτηση στο χώρο υποθέσεων με μια δομή όπως αυτή της κατασκευής δέντρων αποφάσεων. Ο αλγόριθμος κατασκευάζει ένα δέντρο με τη διαφορά πως δεν αναπτύσσει όλους τους κόμβους του κάθε επιπέδου αλλά μόνο αυτόν που κρίνει ότι θα δώσει τα καλύτερα αποτελέσματα. Ή ακόμα καλύτερα αντί να αναπτύσσει μόνο ένα κόμβο αναπτύσσει τους k πιο «υποσχόμενους» κόμβους του κάθε επιπέδου. Ο αλγόριθμος PART είναι για παράδειγμα ένας αλγόριθμος που διαμορφώνει κανόνες από τμήματα των δέντρων απόφασης. Μόλις χτιστεί ένα επί μέρους δέντρο, ένας ενιαίος κανόνας εξάγεται από αυτό (Frank και Witten, 1998).

Ο αλγόριθμος OneR είναι ο πιο απλός αλγόριθμος παραγωγής κανόνων ταξινόμησης μιας και δημιουργεί μόνο έναν κανόνα και σε αυτόν βασίζεται όλες τις αποφάσεις του (Holte, 1993). Ο Mansuri και άλλοι (1991) πρότειναν εκτός από την παραγωγή κανόνων και την παραγωγή ενός συνόλου εξαιρέσεων. Αν το στιγμιότυπο προς ταξινόμηση δεν ανήκει στο σύνολο των εξαιρέσεων τότε ταξινομείται βάσει του συνόλου των παραγόμενων κανόνων.

Αν και οι περισσότεροι αλγόριθμοι παραγωγής κανόνων ταξινόμησης μπορούν να χειριστούν συνεχείς μεταβλητές οι An & Cercone (1999) ισχυρίστηκαν ότι η διακριτοποίηση των συνεχών τιμών των μεταβλητών μειώνει τους κανόνες χωρίς να βλάπτει την ακρίβεια.

3.2.1 Σύγκριση αλγορίθμων παραγωγής κανόνων ταξινόμησης

Σε αυτό το υποκεφάλαιο παραθέτουμε την σύγκριση των πιο γνωστών αλγορίθμων παραγωγής κανόνων ταξινόμησης σε ένα πλήθος από σύνολα δεδομένων εκπαίδευσης.

Δεδομένα	OneR	PART	RIPPER	Ridor
autos	53.46	70.92	71.50	64.89
badges	28.64	100.00	100.00	100.00
balance-scale	60.60	80.96	80.54	80.78
breast-cancer	68.22	69.96	68.95	70.78
breast-w	91.34	95.34	95.67	94.96
colic	80.03	83.80	84.28	81.23
credit-a	85.79	83.02	86.13	83.70
credit-g	66.88	71.41	70.56	71.06
diabetes	73.78	72.63	74.58	73.62
haberman	73.01	71.95	72.24	72.42
heart-c	73.78	77.38	77.87	78.15
heart-h	79.52	78.72	77.62	77.02
heart-statlog	71.54	78.41	77.65	77.54
hepatitis	79.81	81.54	78.85	77.32
ionosphere	83.72	88.17	87.74	86.33
iris	94.51	94.90	95.29	94.51
labor	77.40	82.06	88.73	78.70
lymphography	72.40	77.55	75.57	75.75
monk1	70.07	89.52	84.76	69.13
monk2	60.51	63.09	58.38	60.31
monk3	76.90	89.19	80.25	82.21
relation	77.81	79.04	78.08	77.22
sonar	62.01	75.13	72.30	70.34
students	88.32	83.85	86.26	84.63
vehicle	52.09	70.43	68.03	70.46
vote	95.47	95.54	95.47	93.30
wine	76.63	91.09	89.78	92.43
zoo	41.93	92.33	85.85	91.25

Πίνακας 15. Σύγκριση αλγορίθμων δημιουργίας κανόνων

Βάσει των πειραμάτων μας ο αλγόριθμος RIPPER είχε την καλύτερη ακρίβεια και γι' αυτό χρησιμοποιήθηκε ως επί το πλείστον στο πειραματικό μέρος της διατριβής.

Σε αυτό το σημείο πρέπει να τονιστεί ότι η ακρίβεια ταξινόμησης των αλγορίθμων εκμάθησης κανόνα μπορεί να βελτιωθεί συνδυάζοντας χαρακτηριστικά (όπως στα δέντρα απόφασης) χρησιμοποιώντας ως υπόβαθρο τη γνώση του χρήστη (Flach και Lavrac, 2000) ή αλγόριθμους για αυτόματη κατασκευή χαρακτηριστικών (Markovitch και Rosenstein, 2002).

3.2.2 Χρήση γενετικών αλγορίθμων για την παραγωγή κανόνων ταξινόμησης

Για τη παραγωγή κανόνων ταξινόμησης έχουν επίσης χρησιμοποιηθεί γενετικοί αλγόριθμοι (GAs). Ο Dejong και άλλοι (1993) χρησιμοποίησαν ένα γενετικό αλγόριθμο για να την παραγωγή δυαδικών κανόνων. Υποθέτοντας δύο δύτιμα χαρακτηριστικά X_1 , X_2 και μια δυαδική κλάση C , η δυαδική αναπαράσταση κανόνων σε χρωμόσωμα είναι:

Εάν X_1 =σωστό X_2 =λάθος TOTE C =σωστό AN X_1 =λάθος X_2 =σωστό TOTE C =λάθος
10 01 1 01 10 0

Πρέπει σε αυτό το σημείο να σημειωθεί ότι το μήκος της δυαδικής αναπαράστασης κάθε κανόνα είναι σταθερό. Η δυαδική αναπαράσταση έχει το πλεονέκτημα ότι χρησιμοποιείται ανεξάρτητα από το πρόβλημα. Μπορούν όμως να χρησιμοποιηθούν και άλλες μορφές, ανάλογα με τη φύση του προβλήματος, όπως πραγματικοί αριθμοί, ακέραιοι, αλφαριθμητικά κ.λ.π (Mitchell, 1997). Δυστυχώς δεν υπάρχει ένας τρόπος αναπαράστασης που να είναι ο καλύτερος για όλα τα προβλήματα.

Ο στόχος του γενετικού αλγορίθμου είναι να βρεθούν τα καλύτερα χρωμοσώματα/κανόνες. Το πρώτο βήμα στην υλοποίηση ενός Γενετικού Αλγόριθμου είναι η δημιουργία ενός αρχικού πληθυσμού από πιθανές (και σίγουρα όχι τις καλύτερες) λύσεις του προβλήματος (χρωμοσώματα). Η δημιουργία τους γίνεται συνήθως με τυχαίο τρόπο. Δεν υπάρχουν αυστηροί κανόνες για το μέγεθος του πληθυσμού. Γενικά επιλέγονται τιμές από 100 έως 200. Μεγάλοι πληθυσμοί πιθανόν να οδηγήσουν σε καλύτερες λύσεις, όμως χρησιμοποιούν περισσότερους πόρους του συστήματος.

Η καλή ποιότητα ενός χρωμοσώματος αντιπροσωπεύεται στους γενετικούς αλγόριθμους από μια συνάρτηση, η οποία ονομάζεται συνάρτηση ικανότητας (fitness function) (Banzhaf και άλλοι, 1998). Για την ταξινόμηση, η συνάρτηση ικανότητας είναι η ακρίβεια ταξινόμησης πάνω σ'ένα σύνολο παρεχόμενων περιπτώσεων κατάρτισης.

Στην καρδιά του γενετικού αλγόριθμου υπάρχουν διαδικασίες, που παίρνουν τον πληθυσμό από την αρχική γενιά και παράγουν τον πληθυσμό της επόμενης γενιάς κατά τέτοιο τρόπο ώστε η γενική ικανότητα του πληθυσμού να αυξάνεται. Αυτές οι λειτουργίες επαναλαμβάνονται έως ότου ικανοποιείται κάποιο κριτήριο διακοπής, όπως ένας ορισμένος αριθμός χρωμοσωμάτων επεξεργάστηκε ή ένα χρωμόσωμα ορισμένης ποιότητας έχει παραχθεί. Τρεις διαδικασίες (γενετικοί τελεστές) παίρνουν τον πληθυσμό στην γενιά t και παράγουν το νέο πληθυσμό στην γενιά $t + 1$: επιλογή, διασταύρωση και μετάλλαξη (Banzhaf και άλλοι 1998).

3.2.2.1 Επιλογή

Η επιλογή καθορίζει ποια μέλη του πληθυσμού θα μεταφέρουν τις πληροφορίες που περικλείονται στη δομή τους, στα μέλη της επόμενης γενιάς. Σκοπός της μεθόδου επιλογής είναι να δώσει στα μέλη με καλύτερο δείκτη ποιότητας περισσότερες πιθανότητες επιλογής για αναπαραγωγή.

Υπάρχουν πολλές εναλλακτικές διαδικασίες επιλογής. Η πιο συνηθισμένη είναι με χρήση τροχού ρουλέτας (roulette wheel). Σύμφωνα με αυτήν, τμήματα του τροχού της ρουλέτας αντιστοιχούν σε μέλη του πληθυσμού αναλογικά με την τιμή της ποιότητάς τους. Η επιλογή γίνεται με το «γύρισμα» του τροχού. Η διαδικασία είναι τυχαία αλλά τα άτομα με υψηλό βαθμό ποιότητας έχουν μεγαλύτερες πιθανότητες να επιλεγούν. Σε σχετικά μικρό πληθυσμό, που συνήθως χρησιμοποιείται σε τυπικές εφαρμογές, μια σειρά από μη επιθυμητές επιλογές μπορεί να δημιουργήσει ένα δυσανάλογα μεγάλο αριθμό από απογόνους με χαμηλή συγκριτικά ποιότητα.

3.2.2.2 Διασταύρωση

Είναι ένας τρόπος αναζήτησης λύσεων, βασισμένος σε πληροφορίες που πάρθηκαν από υπάρχουσες λύσεις. Αν δύο άτομα με υψηλό βαθμό ποιότητας ανταλλάξουν μεταξύ τους μερικά γονίδια, μπορεί να δημιουργηθούν απόγονοι με ακόμη καλύτερη ποιότητα. Μπορεί να υλοποιηθεί με πολλούς τρόπους, όπως ανασυνδυασμός ενός σημείου, δύο σημείων, n -σημείων, ενιαίος (uniform) κ.ά (βλέπε Εικόνα 18).

Εικόνα 18. Η γραφική περιγραφή των γενετικών τελεστών

Στην πιο απλή μορφή, του ενός σημείου, επιλέγεται τυχαία ένα σημείο κοπής και ο απόγονος που δημιουργείται παίρνει τα γονίδια δεξιά από το σημείο κοπής από τον ένα γονέα και τα αριστερά από τον άλλο. Έτσι δημιουργείται ένα νέο άτομο, διαφορετικό από τους γονείς του έχοντας όμως χαρακτηριστικά και των δύο γονέων του. Για να αποφύγουμε το ενδεχόμενο οι ανασυνδυασμοί να δημιουργούν απογόνους με χειρότερη ποιότητα, δίνουμε τη δυνατότητα να μην υλοποιηθεί, με κάποιο μικρή συνήθως πιθανότητα και έτσι να περάσουν οι γονείς τις γενετικές πληροφορίες τους αναλλοίωτες στην επόμενη γενιά.

3.2.2.3 Μετάλλαξη

Επιλέγονται κάποια γονίδια και μεταβάλλεται τυχαία (ή μή) η τιμή τους. Αυτή η λειτουργία αποτρέπει την σύγκλιση ενός πληθυσμού και οδηγεί σε νέες περιοχές με την προσδοκία καλύτερων λύσεων. Εφαρμόζεται συνήθως με πολύ μικρή πιθανότητα.

3.2.3 Κανόνες πρώτης τάξης

Οι αλγόριθμοι παραγωγής κανόνων ταξινόμησης έχουν ένα βασικό μειονέκτημα. Οι κανόνες που εξάγουν δεν μπορούν να έχουν μεταβλητές σαν όρους αλλά μόνο τα χαρακτηριστικά των στιγμιотύπων. Κάτι τέτοιο όμως δεν επιτρέπει τη γενίκευση των κανόνων με αποτέλεσμα αυτοί να μπορούν να προβλέπουν μόνο καταστάσεις τις οποίες έχουν εντοπίσει στο σύνολο εκπαίδευσης. Αντιθέτως οι κανόνες πρώτης τάξης μπορούν να δεχθούν και μεταβλητές με αποτέλεσμα να είναι πιο εκφραστικοί (βλέπε Εικόνα 19). Επιπλέον μπορούν εύκολα να μεταφερθούν στην Prolog, υλοποιώντας την ιδέα του επαγωγικού λογικού προγραμματισμού.

Κανόνας πρώτης τάξης (κατηγορηματικής λογικής):

- If Father (x,y) and Female(y) then Daughter(x,y)

Αντίστοιχος κανόνας πρωτασιακής λογικής

- If (Father1=Bob) and (Name2 = Bob) and (Femal1 = True) then (Daughter1,2) = True

Εικόνα 19. Κανόνας πρώτης τάξης (κατηγορηματικής λογικής)

Με άλλα λόγια οι προηγούμενες μέθοδοι που βασίζονταν στη λογική μπορούν να μάθουν υποθέσεις που μπορούν να εκφραστούν στην προτασιακή λογική. Οι ταξινομητές του επαγωγικού λογικού προγραμματισμού χρησιμοποιούν το πλαίσιο κατηγορηματικής λογικής πρώτης τάξης (De Raedt, 1996). Οι Mantaras & Armengol (1998) παρουσίασαν μια ανασκόπηση στον επαγωγικό λογικό προγραμματισμό.

Πρέπει να σημειωθεί ότι αρχικά οι περισσότεροι από τους αλγορίθμους παραγωγής κανόνων πρώτης τάξης προτιμούν εκείνες τις προτάσεις που καλύπτουν όσο το δυνατόν περισσότερες περιπτώσεις, κατόπιν προτιμούν εκείνες τις προτάσεις που έχουν έναν μικρότερο αριθμό

χαρακτηριστικών γνωρισμάτων και τελικά προτιμούν εκείνες τις προτάσεις που έχουν το μικρότερο συνολικό αριθμό τιμών στις εσωτερικές συνδέσεις ($\vee$, $\wedge$).

Υπάρχουν τρεις τρόποι με τους οποίους ένα πρόγραμμα λογικής να γενικεύει την υπόθεση του: 1) αντικαταθιστώντας μερικούς όρους σε μια πρόταση του προγράμματος με μεταβλητές, 2) αφαιρώντας κατηγορήματα μιας πρότασης, 3) προσθέτοντας μια πρόταση στο πρόγραμμα. Ανάλογα, υπάρχουν τρεις τρόποι με τους οποίους ένα πρόγραμμα λογικής μπορεί να ειδικεύει την υπόθεση του: 1) αντικαταθιστώντας μερικές μεταβλητές σε μια πρόταση προγράμματος με όρους, 2) προσθέτοντας κατηγορήματα στην πρόταση, 3) αφαιρώντας μια πρόταση από το πρόγραμμα.

Ένας από τους πιο χαρακτηριστικούς αλγορίθμους για την εκμάθηση κανόνων πρώτης τάξης είναι ο FOIL (Quinlan, 1995) ο οποίος παρουσιάζεται παρακάτω σε ψευδογλώσσα (Εικόνα 20). Αν και ο εσωτερικός βρόγχος του αλγορίθμου λειτουργεί εκ πρώτης όψεως όπως και το βασικό κομμάτι των αλγορίθμων ακολουθιακής κάλυψης υπάρχει μια μικρή διαφορά. Ο FOIL αντί να εντοπίζει και να προσπαθεί να καλύψει και θετικά και αρνητικά στιγμιότυπα ενδιαφέρεται να εντοπίσει μόνο τα θετικά αδιαφορώντας για τα αρνητικά στιγμιότυπα. Η διαφορά στην αναζήτηση του FOIL οφείλεται στην προσπάθειά του να ενσωματώνει μεταβλητές στους κανόνες που παράγει, καθώς επίσης και στην προσπάθειά του να κατασκευάζει αναδρομικούς κανόνες.

Ο αλγόριθμος PROGOL (Muggleton, 1995) είναι ένας άλλος γνωστός αλγόριθμος μάθησης των κανόνων πρώτου βαθμού

Ένα σημαντικό μέρος της έρευνας του επαγωγικού λογικού προγραμματισμού τώρα υπάγεται στον τίτλο Σχεσιακή Ανάσυρση Δεδομένων (Dzeroski και Lavrac, 2001). Αυτοί οι αλγόριθμοι μπορούν να γίνουν κατανοητοί ως επαγωγή δέντρων απόφασης, όπου κάθε κόμβος του δέντρου απόφασης είναι ο ίδιος μια υποκατηγορία δέντρου απόφασης και κάθε υποκατηγορία δέντρου απόφασης αποτελείται από τους κόμβους που κάνουν τις δυαδικές διασπάσεις σε διάφορα χαρακτηριστικά χρησιμοποιώντας τις υπάρχουσες σχέσεις (Blockeel & Raedt, 1998).

Η παραγωγή κανόνων πρώτης τάξης διαφοροποιείται από τις περισσότερες άλλες μορφές μηχανικής μάθησης και από τη χρήση μιας γλώσσας με μεγαλύτερη εκφραστικότητα και από τη δυνατότητά να χρησιμοποιεί λογικά κωδικοποιημένη προϋπάρχουσα γνώση. Αυτό έχει επιτρέψει τις επιτυχείς εφαρμογές του επαγωγικού λογικού προγραμματισμού σε περιοχές όπως της μοριακής βιολογίας και της φυσικής γλώσσας που και οι δύο έχουν πλούσιες πηγές προϋπάρχουσας γνώσης (Muggleton, 1999).

Εικόνα 20. Ο αλγόριθμος FOIL

Κατά τη σύγκριση των τεχνικών του επαγωγικού λογικού προγραμματισμού και της παραγωγής προτασιακών κανόνων ταξινόμησης, υπάρχει μια δυναμική ισορροπία μεταξύ εκφραστικής δύναμης και αποδοτικότητας. Οι επαγωγικές τεχνικές λογικού προγραμματισμού είναι χαρακτηριστικά πιο εκφραστικές αλλά και λιγότερο αποδοτικές. Εάν τα διαθέσιμα στοιχεία έχουν μια τυποποιημένη συνοπτική μορφή, με τις σειρές να είναι τα μεμονωμένα στιγμιότυπα (περιπτώσεις κατάρτισης) και τις στήλες τα χαρακτηριστικά χρησιμοποιήθηκαν για να περιγράψουν τα στοιχεία, ένας αναλυτής δεν έχει λόγο να χρησιμοποιήσει τον επαγωγικό λογικό προγραμματισμό δεδομένου ότι η τυποποιημένη μάθηση μηχανών είναι πιο αποτελεσματική σε αυτή την περίπτωση. Ο επαγωγικός λογικός προγραμματισμός συμφέρει να χρησιμοποιηθεί για την εξόρυξη γνώσης από δεδομένα που αποθηκεύονται σε σχεσιακές βάσεις δεδομένων και υπάρχει άφθονη ειδική γνώση (προϋπάρχουσα γνώση). Χαρακτηριστικά, κάθε κατηγορία αντιστοιχεί σε μια σχέση στη σχεσιακή βάση δεδομένων. Ανάλογα, κάθε γεγονός σε μια ερμηνεία είναι μια περίπτωση στη βάση δεδομένων, και μια ερμηνεία αντιστοιχεί σε ένα μέρος της βάσης δεδομένων (ένα σύνολο περιπτώσεων).

3.3 Μάθηση βασισμένη στα στιγμιότυπα

Οι βασισμένες στα στιγμιότυπα (instance-based, για συντομία IB) μέθοδοι μάθησης έχουν μια θεμελιώδη διαφορά από τις άλλες μεθόδους μάθησης που έχουν αναπτυχθεί: δεν κατασκευάζουν ένα μοντέλο που προσεγγίζει τη συνάρτηση-στόχο καθολικά. Το μόνο που κάνουν στη φάση της εκπαίδευσης είναι να αποθηκεύουν τα δεδομένα, γι' αυτό είναι γνωστές και ως μέθοδοι βασισμένες στη μνήμη (memory-based). Η γενίκευση για νέα δεδομένα γίνεται κάθε φορά που εμφανίζεται ένα νέο στιγμιότυπο προς κατάταξη. Τότε, ένα σύνολο από σχετιζόμενα με αυτό γνωστά στιγμιότυπα ανακαλείται από τη μνήμη και χρησιμοποιείται για την κατάταξη του νέου στιγμιότυπου. Έτσι, αυτό που συμβαίνει ουσιαστικά είναι να παρέχεται μια τοπική προσέγγιση της συνάρτησης-στόχου αντί μία καθολική. Το κύριο πλεονέκτημα των IB μεθόδων είναι πως μπορούν να προσεγγίσουν πολύ καλύτερα από άλλες μεθόδους τη συνάρτηση-στόχο αν αυτή είναι πολύπλοκη καθολικά, αλλά μπορεί να περιγραφεί από τοπικές προσεγγίσεις. Το κύριο μειονέκτημα τους είναι πως το υπολογιστικό κόστος κατά την ταξινόμηση νέων στιγμιότυπων είναι πολύ υψηλό. Ο λόγος είναι πως σχεδόν όλοι οι υπολογισμοί λαμβάνουν χώρα τότε και όχι κατά τη φάση εκπαίδευσης.

Το βιβλίο του Aha (1997) και το άρθρο των De Mantaras & Armengol (1998) παρουσιάζουν μια ανασκόπηση για τους IB ταξινομητές. Παρακάτω περιγράφεται ο πιο διαδεδομένος αλγόριθμος της κατηγορίας, αυτός των k κοντινότερων γειτόνων και οι διάφορες παραλλαγές του. Ο αλγόριθμος ταξινόμησης με βάση τους k κοντινότερους γείτονες (k -Nearest Neighbor Algorithm – k -NN) είναι η πιο βασική IB 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 “γείτονές” του.

Με άλλα λόγια, κάθε απόφαση για ταξινόμηση ενός δείγματος x παίρνεται αποκλειστικά με βάση τις ετικέτες των k πλησιέστερων γειτόνων του (Εικόνα 21).

Εικόνα 21. Η εφαρμογή του k-NN αλγόριθμου

Όπως προείπαμε, η αναζήτηση σύμφωνα με τον κανόνα του k πλησιέστερου γείτονα ξεκινάει από το σημείο ελέγχου x και μεγαλώνει μία σφαιρική περιοχή ώσπου να περιλαμβάνει k δείγματα εκπαίδευσης. Το σημείο ελέγχου παίρνει την ετικέτα που έχει η πλειοψηφία αυτών των δειγμάτων. Στην παραπάνω περίπτωση όπου $k = 5$, το σημείο ελέγχου x παίρνει την ετικέτα της κατηγορίας των μαύρων σημείων.

Τρία ζητήματα πρέπει να αποφασιστούν προκειμένου να καθοριστεί πλήρως ο αλγόριθμος:

- Ο ορισμός της απόστασης μεταξύ δύο στιγμιοτύπων, δηλαδή μιας μετρικής πάνω στο χώρο των στιγμιοτύπων (instance space), που θα εκφράζει την εγγύτητα, ή αλλιώς την “ομοιότητα” μεταξύ των στιγμιοτύπων.
- Ο τρόπος συνδυασμού των τιμών των k κοντινότερων γειτόνων.
- Η τιμή του k .

Για το πρώτο ζήτημα, υπάρχουν πολλές εναλλακτικές επιλογές. Η απόφαση εξαρτάται από τα ειδικά χαρακτηριστικά του χώρου στιγμιοτύπων του προβλήματος. Ιδιαίτερη σημασία έχει το αν στην αναπαράσταση των στιγμιοτύπων περιλαμβάνονται αριθμητικά ή συμβολικά χαρακτηριστικά. Στον “παραδοσιακό” k-NN αλγόριθμο, στον οποίο τα στιγμιότυπα θεωρούνται πως ανήκουν στον n -διάστατο χώρο R_n , μια μετρική που υιοθετείται συχνά είναι η γνωστή Ευκλείδεια απόσταση.

Εν συνεχεία, αφού προσδιορισθούν μέσω κάποιας μετρικής οι k κοντινότεροι γείτονες ενός νέου στιγμιότυπου, οι τιμές της συνάρτησης-στόχου που έχει ο καθένας από αυτούς πρέπει να συνδυαστούν για να δώσουν την εκτιμώμενη τιμή για το νέο. Και εδώ είναι δυνατές διάφορες επιλογές. Στην περίπτωση που η συνάρτηση-στόχος παίρνει διακριτές τιμές, η πιο συνηθισμένη τακτική είναι να επιλέγεται η πιο συχνή από τις τιμές των γειτόνων. Σε περίπτωση ισοβαθμιών επιλέγεται εκ των ισοβαθμούντων μια τιμή, είτε τυχαία ή η καθολικά πιο συχνή τιμή (για το σύνολο των στιγμιοτύπων εκπαίδευσης).

Αν για παράδειγμα έχουμε τα παρακάτω δεδομένα εκπαίδευσης για την απόφαση χορήγησης δανείου (Πίνακας 16)

Πελάτης	Ηλικία	Εισόδημα χιλιάδες	σε	Αριθμός Καρτών	Πιστωτικών	Απάντηση
Γιάννης	35	35		3		Όχι
Ρένα	22	50		2		Ναι
Ηλίας	63	200		1		Όχι
Θωμάς	59	170		1		Όχι
Νίκη	25	400		4		Ναι

Πίνακας 16. Δεδομένα εκπαίδευσης

Η απόφαση χορήγησης δανείου στον παρακάτω πελάτη (Πίνακας 17), θα βασιζόταν στους υπολογισμούς της απόστασης των χαρακτηριστικών του νέου πελάτη από τα χαρακτηριστικά των παλιών πελατών.

Πελάτης	Ηλικία	Εισόδημα χιλιάδες	σε	Αριθμός Καρτών	Πιστωτικών	Απάντηση
Τάσος	37	50		2		?

Πίνακας 17. Πελάτης προς ταξινόμηση

Ακολουθούν οι υπολογισμοί της Ευκλείδειας απόστασης των χαρακτηριστικών του νέου πελάτη από τα χαρακτηριστικά των υπόλοιπων πελατών.

Απόσταση από Γιάννη: $[(35-37)^2 + (35-50)^2 + (3-2)^2]^{1/2} = 15.16$

Απόσταση από Ρένα: $[(22-37)^2 + (50-50)^2 + (2-2)^2]^{1/2} = 15$

Απόσταση από Ηλία: $[(63-37)^2 + (200-50)^2 + (1-2)^2]^{1/2} = 152.23$

Απόσταση από Θωμά: $[(59-37)^2 + (170-50)^2 + (1-2)^2]^{1/2} = 122$

Απόσταση από Νίκη: $[(25-37)^2 + (40-50)^2 + (4-2)^2]^{1/2} = 15.74$

Συνεπώς από τον αλγόριθμο 3-NN η απάντηση θα ήταν «Ναι» (αφού Γιάννης: Όχι, Ρένα: Ναι, Νίκη: Ναι).

Πολλές φορές, λόγω του γεγονότος ότι τα δεδομένα στις περισσότερες περιπτώσεις αναφέρονται σε διαφορετικές μονάδες πρέπει να κανονικοποιηθούν στην κλίμακα 0-1 ώστε να αντιμετωπίζονται ισόβαρα από τις μετρικές (ο Πίνακας 18 παρουσιάζει πως κανονικοποιούνται οι τιμές των μεταβλητών του προηγούμενου προβλήματος).

Πελάτης	Ηλικία	Εισόδημα χιλιάδες	σε	Αριθμός Καρτών	Πιστωτικών	Απάντηση
Γιάννης	55/63= 0.55	35/200= 0.175		3/4= 0.75		Όχι
Ρένα	22/63=0.34	50/200= 0.25		2/4= 0.5		Ναι
Ηλίας	63/63= 1	200/200=1		1/4= 0.25		Όχι
Θωμάς	59/63= 0.93	170/200=0.85		1/4= 0.25		Όχι
Νίκη	25/63= 0.39	40/200= 0.2		4/4= 1		Ναι

Πίνακας 18. Κανονικοποίηση των δεδομένων εκπαίδευσης

Η απλότητα αλλά και η γενικότητα του παραπάνω αλγορίθμου παρέχει στον χρήστη την ευκολία να τον τροποποιήσει, προσθέτοντας δικά του μέτρα απόστασης, που θα επηρεάσουν την ευαισθησία του αλγορίθμου ως προς τους k γείτονες και θα φέρουν τα αποτελέσματα πιο κοντά στις επιθυμητές τιμές. Έτσι, κάθε συνάρτηση που πληροί τα κριτήρια μετρικής είναι δυνατόν να επιλεγεί αντί της Ευκλείδειας.

Μια μετρική πρέπει να έχει τέσσερις ιδιότητες: Για οποιαδήποτε διανύσματα (στιγμιότυπα) a , b και c αυτές οι ιδιότητες είναι οι εξής:

- Θετικότητα: $D(a, b) \geq 0$.
- Αντανakλαστικότητα: $D(a, b) = 0$ εάν και μόνο εάν $a = b$.
- Συμμετρικότητα: $D(a, b) = D(b, a)$.
- Τριγωνική Ανισότητα: $D(a, b) + D(b, c) \geq D(a, c)$.

Σε αυτό το σημείο πρέπει να τονιστεί ότι στην περίπτωση που τα χαρακτηριστικά είναι συμβολικά, 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 (βλέπε Πίνακα 19). Πολυάριθμες άλλες μετρικές που έχουν προταθεί παρουσιάζονται στον Πίνακα 19.

Minkowsky: $D(x, y) = \left(\sum_{i=1}^m x_i - y_i ^r \right)^{1/r}$	Euclidean: $D(x, y) = \left(\sum_{i=1}^m x_i - y_i ^2 \right)^{1/2}$
Manhattan: $D(x, y) = \sum_{i=1}^m x_i - y_i $	Camberra: $D(x, y) = \sum_{i=1}^m \frac{ x_i - y_i }{ x_i + y_i }$
Chebychev: $D(x, y) = \max_{i=1}^m x_i - y_i $	Kendall's Rank Correlation: $D(x, y) = 1 - \frac{2}{m(m-1)} \sum_{i=j}^m \sum_{j=1}^{i-1} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j)$

Πίνακας 19. Μετρικές απόστασης στιγμιότυπων

Ένα μειονέκτημα που παρουσιάζουν τα προηγούμενα παραδείγματα μετρικών είναι πως όλα τα χαρακτηριστικά 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 των χαρακτηριστικών 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 επιμηνυνθούν 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 αρκετά συχνά χρησιμοποιούμενη (Wettschereck και άλλοι, 1997).

Σχετικά με το πώς υπολογίζονται τα βάρη, υπάρχουν δύο κύριες προσεγγίσεις. Η πιο απλή και άμεση προσέγγιση είναι να προσδιοριστούν τα βάρη μέσω της βελτιστοποίησης της αποτελεσματικότητας του ταξινομητή. Επιλέγονται τυχαία κάποια από τα δεδομένα εκπαίδευσης για να εκπαιδεύσουν τον αλγόριθμο μάθησης (training set) και τα υπόλοιπα χρησιμοποιούνται για έλεγχο (test set). Πάνω σε αυτά τα δεδομένα, τα βάρη των χαρακτηριστικών επιλέγονται έτσι ώστε να ελαχιστοποιήσουν το ποσοστό λαθών (ή κάποιο άλλο μέτρο αξιολόγησης). Στη συνέχεια, η όλη διαδικασία μπορεί να επαναληφθεί αρκετές φορές για διαφορετικούς διαμερισμούς των δεδομένων σε σύνολα εκπαίδευσης και ελέγχου, με σκοπό τη μεγαλύτερη προσέγγιση της πραγματικής κατανομής των στιγμιότυπων. Τέλος, υπολογίζονται οι μέσοι όροι των βαρών για το σύνολο των επαναλήψεων, οι οποίοι αποτελούν και τα τελικά βάρη.

Αυτή η προσέγγιση περιτυλίγματος, αν και άμεση, δεν είναι αποδοτική, για το λόγο ότι απαιτεί κατά την αναζήτηση σε ένα μεγάλο χώρο παραμέτρων την κλήση του αλγορίθμου μάθησης τόσες φορές, όσα είναι τα βήματα της αναζήτησης. Μια άλλη διαδεδομένη προσέγγιση στην εκτίμηση παραμέτρων είναι αυτή του φίλτρου (filter approach). Σύμφωνα με αυτήν, η εκτίμηση γίνεται χωρίς τη χρήση του αλγορίθμου μάθησης, με τη βοήθεια των δεδομένων μόνο και μιας αριθμητικής συνάρτησης που εκτιμά τη “σημαντικότητα” της παραμέτρου όπως η μετρική του κέρδους που περιγράφηκε στα δέντρα αποφάσεων. Εκτός από την μετρική του κέρδους όπως και στα δέντρα αποφάσεων μπορεί να χρησιμοποιηθούν και άλλες μετρικές (Aha, 1997).

Μία άλλη παραλλαγή του k-NN 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. Η βαρύτητα του j γείονα μπορεί να υπολογιστεί για παράδειγμα:

$$w_j = \begin{cases} \frac{d_k - d_j}{d_k - d_1}, & \text{αν } d_k \neq d_j \text{ αλλιώς } 1 \end{cases}$$

Υπάρχουν δυο βασικοί λόγοι για τους οποίους ο μπορεί ο k – κοντινότερος ταξινομητής να ταξινομήσει λανθασμένα ένα υπο-εξέταση στιγμιότυπο:

- Όταν υπάρχει θόρυβος στη τοποθεσία γύρω υπό την εξεταζόμενη περίπτωση. Οι λανθασμένες κατατάξεις λόγω θορύβου μπορεί να νικούν κατά την ψηφοφορία ταξινόμησης, δίνοντας ως αποτέλεσμα λανθασμένη πρόβλεψη. Η επιλογή σε αυτή την περίπτωση μιας μεγαλύτερης τιμής για το k θα μπορούσε να λύσει το πρόβλημα
- Όταν η περιοχή καθορισμού της κατηγορίας ή περιορισμού της κατηγορίας είναι τόσο μικρή ώστε οι περιπτώσεις που ανήκουν στην κατηγορία που περιβάλλει τον περιορισμό, νικούν κατά την ψηφοφορία ταξινόμησης. Η επιλογή μιας μικρότερης τιμής για το k θα μπορούσε να λύσει το πρόβλημα σε αυτή την περίπτωση.

Συνεπώς, η επιλογή του k είναι ένα σύνθετο πρόβλημα. Ο Wettschereck και άλλοι (1997) ερευνήσαν καταρχήν την συμπεριφορά του k-NN στην παρουσία θορύβου. Τα πειράματα έδειξαν ότι η απόδοση του k-NN δεν είναι ευαίσθητη στην ακριβή επιλογή του k όταν το k είναι μεγάλο. Βρήκαν επίσης ότι για μικρές τιμές του k ο k-NN αλγόριθμος ήταν περισσότερο εύρωστος απ’ ότι ο μοναδικός κοντινότερος γείτονας (1NN) για την πλειοψηφία των μεγάλων σε πλήθος δεδομένων εκπαίδευσης που εξετάστηκαν. Όμως, βρέθηκε ότι η απόδοση του k-NN ήταν κατώτερη αυτής που επιτεύχθηκε από τον 1NN σε δεδομένα με λίγα παραδείγματα (< 100).

Σε αυτό το σημείο πρέπει να σημειώσουμε ότι στις δύο διαστάσεις, ο αλγόριθμος του πλησιέστερου γείτονα (1NN) οδηγεί σε ένα διαχωρισμό του χώρου εισόδου σε κελιά Voronoï, καθένα από τα οποία έχει ως ετικέτα την κατηγορία του σημείου εκπαίδευσης που

περιέχει (βλέπε την Εικόνα 22- αριστερή). Στις τρεις διαστάσεις, τα κελιά είναι τρισδιάστατα και το όριο απόφασης μοιάζει με την επιφάνεια ενός κρυστάλλου (βλέπε την Εικόνα 22- δεξιά).

Εικόνα 22. Ο διαχωρισμός που προκύπτει από τον αλγόριθμο του πλησιέστερου γείτονα (1NN) στο επίπεδο και στο χώρο

Μια παραλλαγή του k-NN αλγορίθμου, με την ονομασία IB1 (Aha, 1997), εμφανίζεται στο λογισμικό MBT/TiMBL. Η παραλλαγή αυτή παρουσιάζει δυο βασικές διαφορές:

1. Η τιμή του k αφορά τις k μικρότερες αποστάσεις και όχι τα k κοντινότερα στιγμιότυπα. Μπορεί, επομένως, να χρησιμοποιηθούν κατά την κατάταξη περισσότεροι από k γείτονες, αν κάποιοι από αυτούς απέχουν το ίδιο από το αντικείμενο προς κατάταξη.
2. Σε περίπτωση ισοψηφίας μεταξύ των γειτόνων, αν δηλαδή οι γείτονες ισοκατανέμονται στις δύο κατηγορίες, το k αυξάνεται κατά 1 και η απόφαση λαμβάνεται βάσει του νέου k.

Η IB μάθηση αναφέρεται και ως οκνηρή μάθηση (lazy learning), ακριβώς για το λόγο ότι αναβάλλει τους υπολογισμούς μέχρι την αίτηση για κατάταξη ενός νέου στιγμιότυπου. Έτσι, ένα σημαντικό πρακτικό ζήτημα είναι η ανάπτυξη τεχνικών αποδοτικής ευρετηριοποίησης των στιγμιότυπων εκπαίδευσης, για να μειωθεί ο χρόνος ανάκτησης τους κατά τη φάση κατάταξης. 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 επιλεγθούν στη συνέχεια τα k κοντινότερα. Κάτι τέτοιο έχει υψηλότατο κόστος για μεγάλα σύνολα δεδομένων.

Έστω ότι υπάρχουν n δείγματα εκπαίδευσης με ετικέτες σε d διαστάσεις και γίνεται αναζήτηση για να βρεθεί το πλησιέστερο ως προς ένα σημείο ελέγχου x ($k = 1$). Σε μια πιο απλοϊκή προσέγγιση, ανιχνεύεται κάθε σημείο με τη σειρά του, υπολογίζεται η Ευκλείδεια απόστασή του από το x και διατηρείται η ταυτότητα μόνο του τρέχοντος πιο κοντινού σημείου. Ο υπολογισμός κάθε απόστασης απαιτεί $O(d)$ υπολογισμούς και επομένως συνολικά η αναζήτηση απαιτεί $O(dn)$. Γι' αυτό το λόγο υπάρχει μια γενική αλγοριθμική τεχνική για τη μείωση του υπολογιστικού φόρτου στις αναζητήσεις των πλησιέστερων γειτόνων: η προκατασκευή.

Στις μεθόδους προκατασκευής, κατασκευάζεται κάποιας μορφής δέντρο αναζήτησης στο οποίο τα πρότυπα είναι επιλεκτικά συνδεδεμένα. Κατά τη διάρκεια της ταξινόμησης, υπολογίζεται η απόσταση του σημείου ελέγχου από ένα ή περισσότερα αποθηκευμένα πρότυπα (πρότυπα «εισόδου» ή πρότυπα «ρίζας») και στη συνέχεια μελετώνται μόνο τα πρότυπα που είναι συνδεδεμένα με αυτό. Εάν το δέντρο είναι σωστά δομημένο, ο συνολικός αριθμός των προτύπων που χρειάζεται να ανιχνευτούν μειώνεται.

Ένα συνδυασμό των δέντρων αποφάσεων και των αλγόριθμων του κοντινότερου γείτονα αποτελεί ο αλγόριθμος IGTREE (Daelemans και άλλοι, 1997). Η διαφορά είναι πως η αξιολόγηση των ιδιοτήτων γίνεται μόνο μία φορά στην αρχή και όλοι οι κόμβοι του δένδρου που βρίσκονται στο ίδιο βάθος ελέγχουν την ίδια ιδιότητα. Το δένδρο απόφασης που προκύπτει διαχωρίζει το αντικείμενα εκπαίδευσης σε ομάδες. Κατά το στάδιο της κατάταξης αντικειμένων των οποίων η κατηγορία είναι άγνωστη, εντοπίζεται πρώτα η «κοντινότερη» ομάδα χρησιμοποιώντας το δέντρο απόφασης και το αντικείμενο κατατάσσεται στην κατηγορία που πλειοψηφεί σε εκείνη την ομάδα αντί (όπως στην περίπτωση του k-NN) για την κατηγορία που πλειοψηφεί στο σύνολο των αντικειμένων εκπαίδευσης. Με αυτήν την έννοια ο IGTREE αποτελεί μια προσέγγιση του k-NN. Το κυριότερο πλεονέκτημα του IGTREE είναι ότι η κατάταξη γίνεται πολύ γρηγορότερα, αφού δεν χρειάζεται να εξεταστούν όλα τα παραδείγματα εκπαίδευσης.

Επίσης, οι αλγόριθμοι TRIBL και TRIBL2 (Daelemans και άλλοι, 2003) αποτελούν συνδυασμούς του IGTREE και του IB1. Σε γενικές γραμμές, και οι δύο χρησιμοποιούν τον IGTREE για να κατασκευάσουν ένα δένδρο απόφασης, που διαχωρίζει τα αντικείμενα εκπαίδευσης σε ομάδες. Κατά την κατάταξη νέων αντικειμένων, επιλέγεται πρώτα η «κοντινότερη» ομάδα χρησιμοποιώντας το δένδρο απόφασης ως κάποιο βάθος. Στη συνέχεια, βρίσκονται οι k κοντινότεροι γείτονες του νέου αντικειμένου μέσα στην ομάδα, και το νέο αντικείμενο κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των k κοντινότερων γειτόνων, όπως στην περίπτωση του IB1. Η διαφορά μεταξύ των TRIBL και TRIBL2 έγκειται στο πότε ακριβώς γίνεται κατά την κατάταξη η μετάβαση από τη χρήση του δένδρου απόφασης στη χρήση του IB1. Ο TRIBL χρησιμοποιεί το δένδρο απόφασης ως κάποιο συγκεκριμένο (το ίδιο πάντα) βάθος, ενώ ο TRIBL2 χρησιμοποιεί το δένδρο απόφασης ως το βάθος όπου δεν βρίσκει κλαδί που να ταιριάζει με τις τιμές των ιδιοτήτων του αντικειμένου προς κατάταξη.

Μια άλλη μέθοδος για τη μείωση της πολυπλοκότητας της αναζήτησης του πλησιέστερου γείτονα είναι η απαλοιφή των «άχρηστων» προτύπων κατά τη διάρκεια της εκπαίδευσης, μια τεχνική που είναι ευρύτερα γνωστή ως περικοπή ή συμπύκνωση. Μία απλή μέθοδος για να ελαττωθεί η $O(n)$ πολυπλοκότητα του χώρου είναι να απαλειφθούν πρότυπα τα οποία είναι περιτριγυρισμένα από σημεία εκπαίδευσης που ανήκουν στην ίδια κατηγορία (βλέπε Εικόνα 23). Αυτό αφήνει τα όρια απόφασης, και επομένως και το λάθος, ανεπηρέαστα, ενώ ταυτόχρονα μειώνει το χρόνο απόκρισης.

```

αρχικοποίησε  $j \leftarrow 0$ ,  $D \leftarrow$  σύνολο δεδομένων,  $n \leftarrow$  αριθμός προτύπων
κάνε  $j \leftarrow j + 1$ ; για κάθε πρότυπο  $x'_j$ 
  βρες τους γείτονες του  $x'_j$ 
  Εάν κάποιος γείτονας δεν ανήκει στην ίδια κλάση με το  $x'_j$  τότε σημείωσε το  $x'_j$ 
μέχρι  $j = n$ 
απέρριψε όλα τα σημεία που δεν είναι σημειωμένα
  
```

Εικόνα 23. Περικοπή Πλησιέστερου Γείτονα

Οι Brighton & Mellish (2002) βρήκαν ότι ο αλγόριθμος τους (ICF) και ο αλγόριθμος RT3 (Wilson & Martinez 2000) επιτυγχάνουν τον υψηλότερο βαθμό μείωσης περιπτώσεων χωρίς

απώλεια στην ακρίβεια ταξινόμησης. Ο βαθμός μείωσης περιπτώσεων που οι αλγόριθμοι επιτυγχάνουν είναι αρκετά εντυπωσιακός: με μια μέση μείωση της τάξης του 80% των παραδειγμάτων. Μια άλλη μέθοδος μείωσης των περιπτώσεων, είναι να τροποποιήσουμε τις περιπτώσεις χρησιμοποιώντας λιγότερα νέα παραδείγματα (κατασκευασμένα από τα παλιά)-πρωτότυπα (Sebban και άλλοι, 1999). Ο Lam και άλλοι (2002) συνδύασαν τις τεχνικές μείωσης των στιγμιότυπων με την παραγωγή πρωτοτύπων.

Όμως, ο χρόνος ταξινόμησης της υπό έρευνα περίπτωσης είναι στενά συνδεδεμένος εκτός από τον αριθμό των αποθηκευμένων περιπτώσεων και από τον αριθμό των μεταβλητών που χρησιμοποιούνται για να περιγράψουν την κάθε περίπτωση (στιγμιότυπο). Η μέθοδος του φίλτρου επιλογής μεταβλητών πραγματοποιείται συνήθως σε αυτή την κατηγορία αλγορίθμων με την επιλογή των μεταβλητών που έχουν υψηλή τιμή σε μια προκαθορισμένη αριθμητική συνάρτηση που εκτιμά τη βαρύτητα της μεταβλητής για το έργο της ταξινόμησης. Υπάρχει μια πληθώρα σκληρών αλγορίθμων μάθησης. Ο αλγόριθμος Kstar (John και άλλοι, 1995) διαφοροποιείται από τους υπόλοιπους γιατί χρησιμοποιεί μετρική απόστασης βασισμένη στην εντροπία. Ο αλγόριθμος NNge (Martin, 1995) χρησιμοποιεί πρωτότυπα τα οποία περιγράφονται από κανόνες. Ο αλγόριθμος LBR (Zhenh & Webb, 2000) χρησιμοποιεί σκληρή μάθηση με την διαφορά ότι δεν αποδίδεται η τιμή ταξινόμησης βάση της πλειοψηφίας της κλάσης των γειτονικών παραδειγμάτων αλλά εκπαιδεύεται ένας απλός Μπεϋζιανός ταξινομητής με τα κοντινά παράδειγμα ως σύνολο εκπαίδευσης⁹. Στο επόμενο υποκεφάλαιο παραθέτουμε την σύγκριση των πιο γνωστών αλγορίθμων σε ένα πλήθος από σύνολα δεδομένων εκπαίδευσης.

3.3.1 Σύγκριση σκληρών αλγορίθμων μάθησης

Η παράμετρος που πρέπει να διερευνηθεί ως προς την επίδρασή της στην απόδοση του ταξινομητή είναι το k . Δυστυχώς τα αποτελέσματα δεν ήταν τόσο ξεκάθαρα (βλέπε Πίνακα 20 για $k=1$, $k=3$, $k=5$ με την ευκλείδεια απόσταση ως μετρική απόστασης). Βάσει των πειραμάτων μας ο αλγόριθμος 3NN είχε την καλύτερη ακρίβεια (όχι όμως στατιστικά σημαντική) και γι' αυτό χρησιμοποιήθηκε ως επί το πλείστον στο πειραματικό μέρος της διατριβής. Γενικά, η τιμή του k , πέραν από τη μεγάλη σημασία που έχει η επιλογή του για διαφορετικά σενάρια και η οποία εξηγείται επαρκώς διαισθητικά, εξαρτάται και από τη διαστασιμότητα κατά τρόπο μη ερμηνεύσιμο.

Βάσει των πειραμάτων μας, ο αλγόριθμος LBR σε ορισμένες περιπτώσεις δίνει καλύτερα αποτελέσματα από τους υπολοίπους αλλά χρειάζεται περισσότερο χρόνο κατά την ταξινόμηση αφού είναι υβριδικός.

Συνοπτικά, ο k -NN είναι ένας πολύ αποτελεσματικός αλγόριθμος μάθησης, τόσο για αριθμητικά όσο και για συμβολικά 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.

Τέλος, ο Breiman (1996) αναφέρει ότι η σταθερότητα των k NN ταξινομητών τους ξεχωρίζει από τα δέντρα αποφάσεων και από τα νευρωνικά δίκτυα. Μια μέθοδος μάθησης χαρακτηρίζεται «ασταθής» εάν μικρές αλλαγές στα δεδομένα εκπαίδευσης έχουν σαν αποτέλεσμα μεγάλες αλλαγές στα αποτελέσματα του ταξινομητή.

⁹ Περισσότερα στοιχεία για τον απλό Μπεϋζιανό ταξινομητή στην ενότητα 3.5.2

Δεδομένα	1NN	LBR	NNge	3NN	5NN	Kstar
autos	71,92	69,2	66,58	63,14	58,81	71,31
badges	100	100	100	100	100	90,32
balance-scale	82,56	74,27	79,79	83,69	87,65	87,84
breast-c	72,86	72,43	67,59	72,95	74,08	73,99
breast-w	95,59	97,23	95,92	96,26	96,3	94,96
colic	78,84	81,64	77,95	80,92	81,24	73,54
credit-a	80,46	86,13	81,27	84,51	85,62	78,8
credit-g	70,74	73,56	70,56	72,82	72,65	69,47
diabetes	70,02	74,85	71,67	74,08	74,2	70,14
haberman	64,93	72,62	65,71	68,69	69,74	69,54
heart-c	76,31	84,27	78,93	81,55	83,78	73,59
heart-h	78,82	83,01	79,83	81,92	80,42	76,82
heart-statlog	75,69	82,87	74,8	79,94	82,01	74,37
hepatitis	80,94	83,02	80,17	82,1	84,37	78,85
ionosphere	86,91	89,42	90,1	86,32	84,73	82,71
iris	94,71	94,12	96,27	95,29	95,69	94,12
labor	86,12	88,14	87,62	89,67	88,2	92,23
lymphography	83,76	82,16	78,93	84,16	81,58	85,34
monk1	71,74	81	80,28	75,53	75,76	73,87
monk2	57,38	62,41	53,23	57,03	60,15	57,72
monk3	80,06	93,74	92,52	88,69	87,74	84,4
relation	78,74	78,28	66,37	78,7	78,51	77,71
sonar	87,28	75,15	68,93	82,91	80,37	85,3
students	79,46	87,54	83,93	83,5	84,19	82,38
vehicle	69,17	67,58	62,97	70,01	68,65	69,8
vote	92,16	93,85	94,93	92,83	92,29	92,29
wine	94,07	97,7	94,9	95,05	95,06	97,37
zoo	95,07	94,44	94,15	93,24	90,6	97,34

Πίνακας 20. Σύγκριση σκληρών αλγορίθμων μάθησης

3.4 Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks) αποτελούν μια άλλη μέθοδο μοντελοποίησης σύνθετων προβλημάτων πρόβλεψης με μεγάλο αριθμό εξάρτητων μεταβλητών. Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι.

3.4.1 Το μοντέλο του τεχνητού νευρώνα

Μια σημαντική μονάδα από την οποία αποτελούνται τα τεχνητά νευρωνικά δίκτυα καλείται νευρώνας (perceptron) (Εικόνα 24). Έχει σαν είσοδο ένα διάνυσμα πραγματικών τιμών και σχηματίζει ένα γραμμικό συνδυασμό αυτών των τιμών. Έπειτα παίρνει τιμή 1 αν ο συνδυασμός αυτός είναι μεγαλύτερος από κάποιο όριο και -1 σε κάθε άλλη περίπτωση. Συνεπώς, αν $o(x_1, \dots, x_n)$ είναι ο νευρώνας που έχει σαν είσοδο το διάνυσμα $(x_1, \dots, x_n)$ τότε μπορούμε να γράψουμε :

$$o(x_1, \dots, x_n) = \begin{cases} 1 & \text{αν } w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n > 0 \\ -1 & \text{σε κάθε άλλη περίπτωση} \end{cases}$$

όπου οι τιμές w_i καλούνται βάρη και ορίζουν την επιρροή που έχει κάθε είσοδος x_i στη διαμόρφωση του γραμμικού συνδυασμού. Για συντομία γράφουμε τον νευρώνα σαν τον γραμμικό συνδυασμό της παραπάνω κλαδικής συνάρτησης ή ακόμα πιο απλά $o(\vec{x}) = \text{sgn}(\vec{w} \cdot \vec{x})$ όπου sgn η συνάρτηση πρόσημου. Ένα βασικό χαρακτηριστικό των perceptron είναι ότι χρησιμοποιούνται κυρίως και για την εκμάθηση λογικών συναρτήσεων.

Εικόνα 24. Σχηματική αναπαράσταση ενός νευρώνα

Ένας απλός τρόπος για την εκπαίδευση του νευρώνα είναι η εκχώρηση στα βάρη του τυχαίων τιμών και εν συνέχεια με διαρκή εφαρμογή των στιγμιοτύπων να μεταβάλλονται οι εκτιμήσεις για τις τιμές των βαρών ώστε αυτά να δίνουν την καλύτερη δυνατή προσέγγιση της πραγματικής τιμής της συνάρτησης στόχου.

Ο αλγόριθμος που αποτελεί βάση για την ενημέρωση των βαρών δίνεται από τον τύπο: $w_i \leftarrow w_i + \Delta w_i$ όπου $\Delta w_i = \eta(t - o)x_i$ και t είναι το πραγματικό αποτέλεσμα (αφού για τα στοιχεία στο σύνολο εκπαίδευσης, οι τιμές αυτές είναι γνωστές) και o είναι η προσεγγιστική τιμή που δίνει στο συγκεκριμένο βήμα ο νευρώνας. Το η είναι μια θετική σταθερά που καλείται βαθμός μάθησης. Ο ρόλος αυτής της σταθεράς είναι να ελέγχει το ρυθμό που μεταβάλλονται τα βάρη. Η σύγκλιση του αλγορίθμου έρχεται μετά από ένα πλήθος επαναλήψεων (βλέπε Εικόνα 25).

Εικόνα 25. Σύγκλιση του αλγορίθμου

Η ταχύτητα σύγκλισης του αλγορίθμου εκπαίδευσης επηρεάζεται από την ύπαρξη ακραίων διανυσμάτων εισόδου (outliers) στο σύνολο εκπαίδευσης. Ως ακραία εννοούμε εδώ διανύσματα των οποίων τα επιμέρους στοιχεία (συνιστώσες) διαφέρουν πολύ από τις τιμές των υπολοίπων διανυσμάτων του συνόλου εκπαίδευσης. Τούτο συμβαίνει διότι, όταν ένα

ακραίο διάνυσμα εμφανίζεται στην είσοδο του δικτύου επιφέρει σημαντικές αλλαγές στα συνοπτικά βάρη, σε τέτοιο μάλιστα βαθμό που τα άλλα διανύσματα χρειάζεται να εμφανισθούν αρκετές φορές στην είσοδο του δικτύου (σε πολλούς κύκλους εκπαίδευσης) προκειμένου να επαναπροσαρμόσουν τα συνοπτικά βάρη στα μέτρα τους.

Οι νευρώνες συνήθως εκπαιδεύονται σταδιακά. Ένας άλλος τρόπος είναι να εκπαιδευτούν κατά δέσμες. Σε αυτή την περίπτωση οι προσαρμογές των συνοπτικών βαρών γίνονται αφού περάσουν στο δίκτυο όλα τα παραδείγματα του συνόλου εκπαίδευσης και πραγματοποιούνται με συνυπολογισμό (άθροισμα) των επιμέρους αλλαγών που επιφέρουν τα παραδείγματα όταν εμφανίζονται στην είσοδο του δικτύου. Δεν είναι όμως σίγουρο ότι ένα τέτοιος αλγόριθμος εκπαίδευσης θα συγκλίνει.

Στην Εικόνα 26 παρουσιάζεται ένας εκπαιδευμένος νευρώνας βάση των δεδομένων που εμφανίζονται στον πίνακα της ίδιας εικόνα.

Εικόνα 26. Η γραφική περιγραφή των βαρών στις συνάψεις

Εύκολα παρατηρούμε ότι το πρόβλημα είναι γραμμικώς διαχωρίσιμο:

$$Y = I(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4 > 0)$$

$$\text{όπου } I(z) = \begin{cases} 1 & \text{αν } z \text{ είναι αλήθεια} \\ 0 & \text{αλλιώς} \end{cases}$$

Οι Littlestone & Warmuth (1994) τροποποίησαν τον αλγόριθμο εκπαίδευσης του νευρώνα μεταβάλλοντας τα βάρη χρησιμοποιώντας πολλαπλασιασμό αντί για πρόσθεση ανά βήμα ($w_i = w_i \cdot \alpha$). Απέδειξαν ότι αυτή η τροποποίηση βελτιώνει την ταχύτητα σύγκλισης και ονόμασαν τον νέο αλγόριθμο Winnow.

Οι νευρώνες εφαρμόζονται με απόλυτη επιτυχία σε προβλήματα ταξινόμησης, στα οποία ο χώρος των διανυσμάτων εισόδου είναι γραμμικά διαχωρίσιμος. Αν δεν συμβαίνει αυτό, τότε ο αλγόριθμος εκπαίδευσης του νευρώνα δεν θα οδηγήσει ποτέ το δίκτυο σε σημείο να διαχωρίζει πλήρως τα δεδομένα. Η λύση σε αυτό το πρόβλημα είναι ο «συνδυασμός» πολλών νευρώνων για την δημιουργία τεχνητών νευρωνικών δικτύων.

Μια άλλη λύση προτάθηκε από τους (Freund & Schapire, 1999) με τον αλγόριθμο Voted Perceptron που αποθηκεύει περισσότερες πληροφορίες κατά τη διάρκεια της εκπαίδευσης για να παράγει καλύτερες προβλέψεις. Οι πληροφορίες που διατηρούνται κατά τη διάρκεια της εκπαίδευσης είναι ο κατάλογος όλων των διανυσμάτων πρόβλεψης (το διάνυσμα βαρών του νευρώνα) που παρήχθησαν μετά από κάθε λάθος. Για κάθε τέτοιο διάνυσμα, ο αλγόριθμος μετρά τον αριθμό επαναλήψεων που το διάνυσμα "επέζησε" έως ότου έγινε το επόμενο λάθος και αυτό είναι η "επιρροή" του διανύσματος πρόβλεψης. Για να υπολογίσει ο αλγόριθμος μια πρόβλεψη συνδυάζει όλες τις προβλέψεις αυτών των διανυσμάτων των βαρών με

σταθμισμένη ψηφοφορία ανάλογη με την επιρροή του κάθε διανύσματος. Τα "καλά" διανύσματα πρόβλεψης τείνουν να επιζούν για πολύ και έχουν έτσι μεγαλύτερη επιρροή στην ψηφοφορία.

3.4.2 Νευρωνικά δίκτυα

Η δομή των νευρωνικών δικτύων είναι ανάλογη με αυτή της Εικόνα 27. Η αρχιτεκτονική (τοπολογία) των νευρωνικών δικτύων χαρακτηρίζεται από ένα γράφημα (δίκτυο), του οποίου οι κόμβοι κατανομούνται σε: α) ένα επίπεδο εισόδου (input layer), β) ένα επίπεδο εξόδου (output layer) και γ) ένα ή περισσότερα ενδιάμεσα κρυμμένα επίπεδα (hidden layers). Οι Neocleous & Schizas (2002) προσφέρουν μια πρόσφατη σύντομη ανασκόπηση τη υπάρχουσας βιβλιογραφίας στα Τεχνητά Νευρωνικά Δίκτυα.

Εικόνα 27. Παράδειγμα Νευρωνικού Δικτύου

Στο νευρωνικό δίκτυο, κάθε ένας από τους κόμβους εισόδου αναπαριστά και μια ανεξάρτητη μεταβλητή (μεταβλητή εισόδου). Εν συνεχεία, κάθε κόμβος εισόδου συνδέεται με όλους τους κόμβους στο πρώτο κρυμμένο επίπεδο. Οι κόμβοι του κρυμμένου επιπέδου συνδέονται με κόμβους ενός άλλου κρυμμένου επιπέδου ή με κόμβους στο επίπεδο εξόδου. Οι κόμβοι στο επίπεδο εξόδου αναπαριστούν μια ή περισσότερες εξαρτημένες μεταβλητές (μεταβλητές εξόδου) βάσει του προβλήματος.

Οι κόμβοι του νευρωνικού δικτύου ονομάζονται επίσης νευρώνες, ενώ οι δεσμοί ονομάζονται «συνάψεις». Σε κάθε σύναψη αντιστοιχεί ένα βάρος που ονομάζεται «συνοπτικό βάρος». Συνεπώς, η αρχιτεκτονική ή τοπολογία ενός νευρωνικού δικτύου προσδιορίζεται από τον αριθμό των κόμβων, τον αριθμό των κρυμμένων επιπέδων και τον τρόπο που οι κόμβοι συνδέονται μεταξύ τους. Ο αριθμός των κρυμμένων επιπέδων και κόμβων και ενδεχομένως τα όρια μέσα στα οποία θα κυμαίνονται τα συνοπτικά βάρη είναι στοιχεία που προσδιορίζονται στη φάση σχεδιασμού του νευρωνικού δικτύου.

Συνήθως δεν χρησιμοποιούνται σε ένα νευρωνικό δίκτυο παραπάνω από τρία επίπεδα (ένα κρυφό επίπεδο) καθώς έχει παρατηρηθεί ότι η χρήση παραπάνω επιπέδων αυξάνει κατά πολύ το χρόνο εκπαίδευσης χωρίς να συνοδεύεται από μια αντίστοιχη αύξηση της απόδοσης του δικτύου (Mitcell, 1997).

Όπως προείπαμε, κάθε κόμβος (νευρώνας) στο μεσαίο επίπεδο είναι πλήρως συνδεδεμένος με τις εισόδους, γεγονός που σημαίνει ότι το κρυφό πεδίο βασίζεται σε όλες τις εισόδους τις οποίες και συνδυάζει στις τιμές εξόδου. Ο συνδυασμός αυτός καλείται συνάρτηση ενεργοποίησης του κόμβου. Η συνάρτηση ενεργοποίησης έχει δύο μέρη. Το πρώτο μέρος είναι η *συνάρτηση σύνδεσης (combination function)* η οποία συνδυάζει όλες τις εισόδους σε μία απλή τιμή. Πρέπει να σημειωθεί ότι κάθε είσοδος έχει το δικό της βάρος. Η πιο κοινή συνάρτηση σύνδεσης είναι το άθροισμα όλων των εισόδων πολλαπλασιασμένων με το αντίστοιχο βάρος τους ($X_1 \cdot W_1 + X_2 \cdot W_2 + \dots + X_N \cdot W_N$). Σε ορισμένες περιπτώσεις είναι χρήσιμες άλλες συναρτήσεις και περιλαμβάνουν το μέγιστο των εισόδων

πολλαπλασιασμένων με το βάρος τους, το ελάχιστο, ή το λογικό ΚΑΙ (AND) ή Η (OR) των τιμών. Ωστόσο, η συνάρτηση που βασίζεται στο άθροισμα των εισόδων πολλαπλασιασμένων με τα βάρη τους δουλεύει συνήθως καλύτερα στην πράξη.

Το δεύτερο μέρος της συνάρτησης ενεργοποίησης είναι η *συνάρτηση μεταφοράς (transfer function)*, η οποία μεταφέρει την τιμή της συνάρτησης σύνδεσης στην έξοδο. Υπάρχουν αρκετά είδη συναρτήσεων μεταφοράς: η *σιγμοειδής*, *γραμμική* και η συνάρτηση *υπερβολικής εφαπτομένης (hyperbolic tangent)* και άλλες (βλέπε Εικόνα 28). Η γραμμική συνάρτηση έχει περιορισμένη πρακτική σημασία αντίθετα με τις άλλες δύο (μη γραμμικές συναρτήσεις) οι οποίες παρουσιάζουν μη γραμμική συμπεριφορά.

Εικόνα 28. Συναρτήσεις μεταφοράς

Στην ειδική περίπτωση που εφαρμόζεται γραμμική συνάρτηση ενεργοποίησης και δεν υπάρχουν ενδιάμεσα κρυμμένα επίπεδα, το νευρωνικό δίκτυο είναι ισοδύναμο με το μοντέλο γραμμικής παλινδρόμησης. Παρόμοια, για συγκεκριμένους μη γραμμικούς τύπους συναρτήσεων ενεργοποίησης, τα νευρωνικά δίκτυα είναι ισοδύναμα με το μοντέλο λογιστικής παλινδρόμησης

Σε αυτό το σημείο πρέπει να τονιστεί ότι η μονόδρομη αρχιτεκτονική δικτύων επιτρέπει στα σήματα να κινούνται προς την μια κατεύθυνση, από τη είσοδο στην έξοδο. Δεν υπάρχουν αμφίδρομες καταστάσεις (loops), έτσι το αποτέλεσμα ενός επιπέδου δεν επηρεάζει το ίδιο το επίπεδο. Η μορφή αυτή των νευρωνικών δικτύων χρησιμοποιείται συχνότερα.

3.4.2.1 Εκπαίδευση νευρωνικού δικτύου

Όπως προείπαμε τα τεχνητά νευρωνικά δίκτυα μπορούν να λύσουν μη γραμμικά προβλήματα. Το πιο απλό μη γραμμικό πρόβλημα είναι το λογικό XOR (βλέπε Πίνακα 21).

x_1	x_2	$x_1 \text{ XOR } x_2$
-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

Πίνακας 21. Λογικό XOR

Στην γραφική παράσταση του XOR (βλέπε Εικόνα 29), τα ζευγάρια 1 και -1 απεικονίζονται με τους πράσινους και κόκκινους κύκλους, αντίστοιχα. Αυτές οι δύο κατηγορίες (πράσινες και κόκκινες) δεν μπορούν να χωριστούν χρησιμοποιώντας μια γραμμή. Πρέπει να χρησιμοποιήσουμε δύο γραμμές, όπως εκείνες που απεικονίζονται με μπλε (βλέπε Εικόνα 29).

Εικόνα 29. Γραφική παράσταση του XOR

Το νευρωνικό δίκτυο με δύο κρυμμένους κόμβους πραγματοποιεί αυτόν τον μη γραμμικό χωρισμό, όπου κάθε κρυμμένος κόμβος περιγράφει μια από τις δύο μπλε γραμμές (βλέπε Εικόνα 30).

Εικόνα 30. Το εκπαιδευμένο νευρωνικό δίκτυο για το πρόβλημα XOR

Η εκπαίδευση του νευρωνικού δικτύου πραγματοποιείται με συστηματική τροποποίηση των συνοπτικών βαρών μέχρι να ληφθεί η επιθυμητή έξοδος (βλέπε Εικόνα 31). Δηλαδή τα συνοπτικά βάρη είναι οι άγνωστοι παράμετροι που εκτιμώνται μέσω μιας διαδικασίας εκπαίδευσης.

Εικόνα 31. Γενική περιγραφή εκπαίδευσης νευρωνικού δικτύου

Το νευρωνικό δίκτυο θεωρείται ότι έχει εκπαιδευτεί, όταν τα συνοπτικά βάρη, με τις τιμές που έχουν λάβει, δίδουν την επιθυμητή έξοδο.

Ένας βασικός τύπος εκπαίδευσης είναι ο FFB (Feed-forward backpropagation) που υλοποιείται γενικά σε δύο φάσεις (εδώ θεωρούμε ένα μόνο κρυμμένο επίπεδο):

1. Φάση προώθησης (Feed-forward): Λαμβάνεται ένα σύνολο αρχικών συνοπτικών βαρών και με αυτά προσδιορίζεται η τιμή στον κόμβο εξόδου
2. Φάση οπισθοδρόμησης (back-propagation): υπολογίζεται το σφάλμα στην έξοδο, δηλαδή η διαφορά μεταξύ της τιμής που υπολογίστηκε και της επιθυμητής τιμής (πραγματικής τιμής που βρίσκεται στο σύνολο δοκιμής). Στη συνέχεια το σφάλμα αυτό κατανέμεται στους κόμβους του κρυμμένου επιπέδου αναλογικά ως προς τα συνοπτικά βάρη. Με τη διαδικασία αυτή σε κάθε κόμβο εξόδου και κάθε κρυμμένο κόμβο αντιστοιχίζεται ένα σφάλμα. Στη συνέχεια το σφάλμα του κάθε κόμβου χρησιμοποιείται για να προσαρμοστεί το εισερχόμενο στον κόμβο αυτό συνοπτικό βάρος.

Η παραπάνω διαδικασία εκπαίδευσης επαναλαμβάνεται για κάθε παράδειγμα του συνόλου εκπαίδευσης. Όταν εξαντληθεί το σύνολο εκπαίδευσης λέμε ότι συμπληρώνεται ένα κύκλος (μια εποχή) εκπαίδευσης (epoch). Το σύνολο εκπαίδευσης χρησιμοποιείται επαναληπτικά (πραγματοποιούνται διαδοχικοί κύκλοι εκπαίδευσης) έως ότου το σφάλμα στην έξοδο σταματήσει να μειώνεται. Στο σημείο αυτό το νευρωνικό δίκτυο θεωρείται ότι έχει εκπαιδευτεί και ότι μπορεί να αναπαράγει τις τιμές εξόδου στο σύνολο δοκιμής.

Πριν αναλυθεί περισσότερο ο αλγόριθμος Backpropagation κρίνεται απαραίτητο να εισαχθούν κάποιοι συμβολισμοί:

x_{ji} η i -οστή μονάδα εισόδου του j νευρώνα

w_{ji} το βάρος που σχετίζεται με την προαναφερθείσα μονάδα εισόδου

o_j η προσεγγιστική τιμή της μονάδας j

t_j η πραγματική τιμή της μονάδας j

Παρακάτω δίνεται ο αλγόριθμος Back propagation σε ψευδογλώσσα:

Εικόνα 32. Ο αλγόριθμος BP

Κρίσιμο σημείο στη διαδικασία εκπαίδευσης είναι το πότε πρέπει αυτή να σταματήσει. Ένας μεγάλος αριθμός κύκλων εκπαίδευσης είναι πολύ πιθανό να οδηγήσει σε ένα υπερ-προσαρμοσμένο (υπερ-ταιριασμένο) νευρωνικό δίκτυο, προσαρμοσμένο δηλαδή σε μεγάλο βαθμό στα δεδομένα του συνόλου εκπαίδευσης, που δεν μπορεί να δουλέψει καλά με άλλα δεδομένα (τα δεδομένα του συνόλου δοκιμής για παράδειγμα).

Το πρόβλημα εμφανίζεται σε νευρωνικά δίκτυα με πολύ μεγάλους αριθμούς νευρώνων στο κρυφό επίπεδο, τα οποία εκπαιδεύονται μέχρι να ελαχιστοποιηθεί το συνολικό σφάλμα σε πολύ μεγάλο βαθμό. Μερικές μέθοδοι αντιμετώπισης του προβλήματος είναι οι εξής (Saad, 1998):

- Τροποποίηση της συνάρτησης σφάλματος, έτσι ώστε να προτιμούνται οι μικροί συντελεστές
- Χρήση στατιστικών μεθόδων για εκτίμηση του πλήθους των εσωτερικών συνδέσεων που χρειάζονται για την εκπαίδευση του δικτύου.
- Μέθοδος του πρόωρου τερματισμού, με καταμερισμό των διαθέσιμων δεδομένων σε δεδομένα εκπαίδευσης, δεδομένα επαλήθευσης και δεδομένα ελέγχου. Το ποσοστό των παραδειγμάτων που θα χρησιμοποιηθούν στο σύνολο εκπαίδευσης ορίζεται παραμετρικά.

Επίσης ορίζεται παραμετρικά και ο αριθμός των κύκλων εκπαίδευσης που αλγόριθμος θα συνεχίσει να εκτελεί από τη στιγμή που το σφάλμα σταματήσει να μειώνεται. Εάν η τιμή της παραμέτρου αυτής είναι αρκετά μεγάλη μπορεί να βοηθήσει τον αλγόριθμο να απεγκλωβιστεί από ένα τοπικό ελάχιστο και να συνεχίσει με περαιτέρω μείωση του σφάλματος (βλέπε Εικόνα 33).

Εικόνα 33. Τοπικό και ολικό ελάχιστο σφάλματος

Σύμφωνα με τη μέθοδο πρόωρου τερματισμού που χρησιμοποιείται συχνότερα, τα διαθέσιμα δεδομένα διαιρούνται σε τρεις ομάδες:

- Τα δεδομένα εκπαίδευσης (training data)
- Τα δεδομένα επαλήθευσης (validation data)
- Τα δεδομένα ελέγχου (test data)

Ένας συνήθης ποσοστιαίος διαχωρισμός είναι 50-25-25. Η εκπαίδευση γίνεται μόνο με τα δεδομένα εκπαίδευσης. Κατά την εκπαίδευση παρακολουθείται και το σφάλμα στα δεδομένα επαλήθευσης. Εάν κατά την εκπαίδευση συμβεί το σφάλμα στα δεδομένα επαλήθευσης να αυξηθεί για συγκεκριμένο αριθμό συνεχόμενων εποχών, η εκπαίδευση διακόπτεται και επιστρέφονται οι παράμετροι που αντιστοιχούσαν στο μικρότερο σφάλμα των δεδομένων επαλήθευσης. Τα δεδομένα ελέγχου χρησιμοποιούνται για να ελέγχουν την ισοκατανομή των δεδομένων επαλήθευσης. Εάν τα δεδομένα ελέγχου παρουσιάσουν ελάχιστο σφάλμα σε αρκετά διαφορετικό σημείο της εκπαίδευσης σε σχέση με τα δεδομένα επαλήθευσης, πραγματοποιείται νέος χωρισμός των δεδομένων σε κατηγορίες.

Η τεχνική πρόωρου τερματισμού δεν έχει ως αποτέλεσμα την κατασκευή καλύτερου νευρωνικού δικτύου, αλλά την ανίχνευση προβλημάτων γενίκευσης κατά την εκπαίδευση με οποιαδήποτε μέθοδο εκπαίδευσης. Γενικά, η επιφάνεια που ορίζεται από τη συνάρτηση σφάλματος στα μη γραμμικά νευρωνικά δίκτυα είναι περισσότερο πολύπλοκη από την αντίστοιχη επιφάνεια στα γραμμικά δίκτυα. Αυτό έχει ως συνέπεια την ύπαρξη πολλών τοπικών βέλτιστων, γεγονός που αυξάνει την πιθανότητα εγκλωβισμού του αλγόριθμου σε ένα από αυτά. Αυτό μπορεί να εξαρτάται ακόμη από τις αρχικές τιμές που δίδονται στα συνοπτικά βάρη και τις τάσεις. Προκειμένου να διασφαλισθεί ότι ο αλγόριθμος εκπαίδευσης θα δώσει τα βέλτιστα συνοπτικά βάρη, είναι αναγκαίο να επιχειρηθούν διάφοροι κύκλοι επανεκπαίδευσης με διαφορετικά αρχικά συνοπτικά βάρη ή ακόμη να επιστρατευθούν άλλες τεχνικές διατάραξης της λύσης προκειμένου ο αλγόριθμος εκπαίδευσης να απεγκλωβιστεί από ενδεχόμενο τοπικό ελάχιστο. Επίσης, η ποιότητα των μη γραμμικών δικτύων εξαρτάται από την τοπολογία τους, δηλαδή τον αριθμό των κρυμμένων επιπέδων και τον αριθμό των νευρώνων σε κάθε επίπεδο (Yam & Chow, 2001).

Μία άλλη προσέγγιση στο πρόβλημα της βελτιστοποίησης του μεγέθους ενός δικτύου είναι να ορίσουμε ένα δίκτυο που είναι μεγάλο ή μεγαλύτερο απ' ότι είναι αναγκαίο για να κάνει

τη δουλειά, και ύστερα να χρησιμοποιήσουμε ένα γενετικό αλγόριθμο για να ορίσουμε ποιοι συνδυασμοί συνδέσεων είναι ικανοί να μάθουν γρήγορα και με ακρίβεια να εκτελούν ένα έργο χρησιμοποιώντας προς τα πίσω διάδοση (Banzhaf και άλλοι, 1998).

Εν συνέχεια αναφέρονται ενδεικτικά μερικές πιο εξελιγμένες μέθοδοι εκπαίδευσης:

- Δυναμική μέθοδος: Ορίζεται μια αρχική τοπολογία του δικτύου και στη συνέχεια όσο η εκπαίδευση εξελίσσεται με την προσαρμογή των συνοπτικών βαρών, η αρχική τοπολογία τροποποιείται με προσθήκη ή/ και διαγραφή νευρώνων από τα κρυμμένα επίπεδα (Young και Downs, 1998).
- Μέθοδος πολλαπλών δικτύων: Η εκπαίδευση γίνεται όπως και με την δυναμική μέθοδο, με τη διαφορά ότι ορίζονται αρχικά περισσότερα του ενός δίκτυα με διαφορετικές τοπολογίες, τα οποία στη συνέχεια εκπαιδεύονται παράλληλα (ψευδο-παράλληλα για την ακρίβεια). Ως τελικό δίκτυο λαμβάνεται εκείνο με το μικρότερο σφάλμα.
- Μέθοδος αποκοπής (Castellano και άλλοι, 1997): Ορίζεται αρχικά ένα μεγάλο δίκτυο (τοπολογία με πολλά επίπεδα και πολλούς νευρώνες) και στη συνέχεια, όσο η εκπαίδευση εξελίσσεται, αποκόπτονται οι ασθενέστεροι νευρώνες των κρυμμένων επιπέδων και στοιχεία από το επίπεδο εισόδου. Η όλη διαδικασία προσδιορίζεται μέσω των τιμών διαφόρων παραμέτρων, όπως: α) Αριθμός κρυμμένων νευρώνων που θα αποκόπτονται ταυτόχρονα σε κάθε φάση αποκοπής, β) Αριθμός φάσεων αποκοπής που θα συνεχίσουν να γίνονται από τη στιγμή που σταματήσει να βελτιώνεται το δίκτυο (απεγκλωβισμός από τοπικό ελάχιστο), γ). Αριθμός στοιχείων εισόδου που θα αποκόπτονται ταυτόχρονα σε κάθε φάση αποκοπής στην είσοδο και δ) Αριθμός κύκλων εκπαίδευσης που ο αλγόριθμος θα συνεχίσει να εκτελείται από τη στιγμή που το σφάλμα σταματήσει να μειώνεται και πριν περάσει σε φάση αποκοπής.

Όπως και στους νευρώνες (perceptron), υπάρχουν γενικά δύο τύποι επιβλεπόμενης μάθησης-εκπαίδευσης: η εκπαίδευση κατά δέσμες (batch training) και η σταδιακή εκπαίδευση (incremental training). Στην πρώτη περίπτωση οι αλλαγές στα συνοπτικά βάρη γίνονται στη βάση μια δέσμης διανυσμάτων εισόδου. Σύμφωνα με τη σταδιακή εκπαίδευση, τα συνοπτικά βάρη τροποποιούνται κάθε φορά που ένα νέο διάνυσμα τιμών εισόδου δίδεται στο δίκτυο. Η σταδιακή εκπαίδευση είναι γνωστή και ως εκπαίδευση ανοικτής γραμμής (on-line training) ή προσαρμοστική (adaptive) εκπαίδευση.

Όπως προείπαμε, ένα πρόβλημα είναι η επιλογή του μεγέθους του κρυμμένου επιπέδου, επειδή ο μικρός αριθμός νευρώνων μπορεί να οδηγήσει σε μικρή ικανότητα γενίκευσης, ενώ οι υπερβολικοί κόμβοι μπορεί να οδηγήσουν σε υπερβολική προσαρμογή και υπερταΐριασμα. Ο αριθμός των απαιτούμενων κρυφών νευρώνων αναλύεται από τους (Camargo & Yoneyama, 2001). Ενώ οι Kon & Plaskota (2000) μελέτησαν το θέμα των χαμηλότερων ορίων στον αριθμό των νευρώνων και των απαραίτητων περιπτώσεων εκπαίδευσης.

Τα νευρωνικά δίκτυα ακτινικής βάσης (RBF, Radial Basis Function), αποτελούν έναν άλλο τύπο νευρωνικών δικτύων (Acosta, 1995). Ένας τυπικός αλγόριθμος εκπαίδευσης νευρωνικών δικτύων RBF αποτελείται από δύο στάδια: Στο πρώτο στάδιο χρησιμοποιείται ο αλγόριθμος k-means¹⁰ για τον καθορισμό της τοποθεσίας των κέντρων των νευρώνων της κρυφής στοιβάδας. Στο δεύτερο, καθορίζονται τα συνοπτικά βάρη του δικτύου, χρησιμοποιώντας γραμμική παλινδρόμηση ανάμεσα στις εξόδους των νευρώνων της κρυφής στοιβάδας και στις πραγματικές εξόδους του συστήματος που μοντελοποιείται.

Αναλυτικότερα, οι νευρώνες του ακτινικού επιπέδου είναι τόσοι όσα και τα παραδείγματα εκπαίδευσης. Κάθε ακτινικός νευρώνας λοιπόν αντιστοιχεί σε ένα παράδειγμα εκπαίδευσης. Τα βάρη στις εισόδους κάθε νευρώνα του ακτινικού επιπέδου είναι ίδια με τις τιμές του αντίστοιχου παραδείγματος εκπαίδευσης. Οι εξοδοί των νευρώνων του ακτινικού επιπέδου

¹⁰ Αναλυτική περιγραφή για τον αλγόριθμο k-means βρίσκεται στην ενότητα 8.3.2

πολλαπλασιάζονται με τα βάρη στις εισόδους των νευρώνων του γραμμικού επιπέδου και τελικά δίνουν τις εξόδους του δικτύου.

Το ακτινικό δίκτυο δεν χρειάζεται εκπαίδευση. Τα βάρη μπορούν να υπολογιστούν αυτόματα από τα παραδείγματα εκπαίδευσης. Ειδικότερα, τα βάρη στις εισόδους των ακτινικών νευρώνων αντιστοιχούν στις εισόδους των παραδειγμάτων εκπαίδευσης (ένα παράδειγμα για κάθε νευρώνα). Στη συνέχεια τα βάρη στις εξόδους των ακτινικών νευρώνων υπολογίζονται έτσι ώστε το ακτινικό δίκτυο να απαντά απολύτως σωστά σε όλα τα παραδείγματα εκπαίδευσης. Ο υπολογισμός των βαρών αυτών βασίζεται στην επίλυση συστημάτων N γραμμικών εξισώσεων με N αγνώστους, όπου N το πλήθος των παραδειγμάτων (και άρα και των ακτινικών νευρώνων) αλλά και το πλήθος των βαρών στις εισόδους κάθε γραμμικού νευρώνα.

Σε σχέση με ένα απλό νευρωνικό δίκτυο, ένα ακτινικό δίκτυο πλεονεκτεί γιατί:

- Δεν χρειάζεται εκπαίδευση
- Παρουσιάζει καλύτερη συμπεριφορά σε προβλήματα με πάρα πολλά παραδείγματα

Και μειονεκτεί γιατί:

- Έχει περισσότερους νευρώνες
- Έχει μεγαλύτερο χρόνο υπολογισμού κατά τη λειτουργία του (λόγω ακριβώς των περισσότερων νευρώνων).

Επίσης, τα ανταγωνιστικά δίκτυα μπορούν να χρησιμοποιηθούν και για επιβλεπόμενη κατηγοριοποίηση. Αυτά τα δίκτυα ονομάζονται Learning Vector Quantization (LVQ) ή Counter-propagation networks (CPN) (Kohonen, 1997). Αποτελούνται από ένα (κρυφό) επίπεδο ανταγωνιστικών νευρώνων και ένα δεύτερο επίπεδο με γραμμικούς νευρώνες (επίπεδο εξόδου). Οι νευρώνες του γραμμικού επιπέδου είναι σημαντικά λιγότεροι αυτών του ανταγωνιστικού επιπέδου και ίσοι με τον αριθμό των προκαθορισμένων κατηγοριών.

Τα διανύσματα εισόδου κατηγοριοποιούνται δύο φορές. Το ανταγωνιστικό επίπεδο τα κατανέμει σε υπο-κατηγορίες. Οι υπο-κατηγορίες του ανταγωνιστικού επιπέδου αντιστοιχίζονται στις κατηγορίες του γραμμικού επιπέδου. Κάθε νευρώνας του κρυφού επιπέδου ενεργοποιεί έναν και μόνο έναν νευρώνα του επιπέδου εξόδου (βλέπε Εικόνα 34, βάρος σύνδεσης προς αυτόν τον νευρώνα=1, προς τους υπόλοιπους νευρώνες=0).

Εικόνα 34. Ανταγωνιστικό δίκτυο

Αρχικά καθορίζεται μια αντιστοίχιση των νευρώνων του κρυφού επιπέδου στους νευρώνες εξόδου, βάσει ποσοστών που δίνονται από τον χρήστη (π.χ., 50% των νευρώνων του κρυφού επιπέδου θα ενεργοποιούν τον πρώτο νευρώνα εξόδου και 50% θα ενεργοποιούν τον δεύτερο νευρώνα εξόδου). Τα ποσοστά πρέπει να ανταποκρίνονται στην κατανομή των παραδειγμάτων στις προκαθορισμένες κατηγορίες. Η αντιστοίχιση των νευρώνων του κρυφού επιπέδου στους νευρώνες του επιπέδου εξόδου είναι σταθερή, καθορίζεται από την κατασκευή του δικτύου LVQ και δεν αλλάζει κατά τη διάρκεια της εκπαίδευσης. Στη συνέχεια ξεκινά η εκπαίδευση, η οποία σκοπό έχει να αντιστοιχίσει τα παραδείγματα εκπαίδευσης στους "σωστούς" νευρώνες του κρυφού επιπέδου.

Για κάθε παράδειγμα εκπαίδευσης υπάρχουν πολλοί νευρώνες του κρυφού επιπέδου στους οποίους μπορεί να αντιστοιχηθεί. Οι νευρώνες αυτοί ανταγωνίζονται μεταξύ τους. Για κάθε παράδειγμα εκπαίδευσης βρίσκεται ο νευρώνας-νικητής μεταξύ όλων των νευρώνων.

- Εάν ο νευρώνας-νικητής αντιστοιχίζει το παράδειγμα στη σωστή κατηγορία εξόδου, τα βάρη εισόδου του νευρώνα αλλάζουν ώστε να πλησιάσουν το παράδειγμα.
- Εάν ο νευρώνας-νικητής αντιστοιχίζει το παράδειγμα σε λάθος κατηγορία εξόδου, τα βάρη εισόδου του νευρώνα αλλάζουν ώστε να απομακρυνθούν από το παράδειγμα.

Συνοψίζοντας, οι ιδιότητες των νευρωνικών δικτύων μας επιτρέπουν να τα χρησιμοποιούμε στην περίπτωση που υπάρχει αρκετός θόρυβος στα δεδομένα. Τέτοιες περιπτώσεις είναι τα δεδομένα αισθητήρων του περιβάλλοντος. Επιπλέον τα νευρωνικά δίκτυα χρησιμοποιούνται σε περιπτώσεις που έχουμε σύνθετα ή και πολλά χαρακτηριστικά, όπως διάνυσματα διακριτών και πραγματικών χαρακτηριστικών ή ακόμα και διάνυσμα σαν χαρακτηριστικό. Δυστυχώς όμως υπάρχουν και κάποια μειονεκτήματα αν αποφασίσουμε να εκπαιδεύσουμε νευρωνικά δίκτυα. Το κυριότερο είναι πως εν γένει τα τεχνητά νευρωνικά δίκτυα χρειάζονται πολύ περισσότερο χρόνο από άλλες μεθόδους για την εκπαίδευσή τους. Για να βρει κανείς τις βέλτιστες παραμέτρους του δικτύου σε ένα συγκεκριμένο πρόβλημα δεν είναι εύκολο μίας και η ακρίβεια μπορεί να απαιτήσει μεγάλο υπολογιστικό κόστος. Εάν κάποιος είναι διατεθειμένος να χρησιμοποιήσει έναν υπέρ-υπολογιστή ο οποίος θα εργαζόταν στο πρόβλημα για άπειρο καιρό η βέλτιστη τοπολογία δικτύου θα ήταν αυτή με παραμέτρους το ελάχιστο στόχο σφάλματος, θα χρησιμοποιούσε τον μέγιστο αριθμό δειγμάτων προς εκπαίδευση κτλ. Ωστόσο σε πραγματικά προβλήματα αυτό είναι κάτι αδύνατο και σκοπός είναι πώς μπορεί κάποιος να θυσιάσει την ακρίβεια για χάρη του υπολογιστικού κόστους. Τέλος ο ταξινομητής που δίνουν συχνά είναι από δυσνόητος έως ακατανόητος για τον άνθρωπο. Θα πρέπει λοιπόν να μας ενδιαφέρει κυρίως η σωστή ταξινόμηση και όχι η αναγνωσιμότητα και κατανόηση αυτής.

3.4.3 Σύγκριση αλγορίθμων

Σε αυτό το υποκεφάλαιο παραθέτουμε την σύγκριση των πιο συχνά χρησιμοποιούμενων αλγορίθμων νευρωνικών δικτύων σε ένα πλήθος από σύνολα δεδομένων εκπαίδευσης. Οι αλγόριθμοι Voted Perceptron και Winnow εφαρμόζονται απευθείας μόνο σε προβλήματα δυο κλάσεων. Για αυτόν τον λόγο, μετατρέψαμε τα πολυ-κατηγορικά (multi-class) προβλήματα σε ένα σύνολο δυαδικών προβλημάτων. Έχουμε χρησιμοποιήσει την απλή μέθοδο μια κλάση εναντίον όλων (One Per Class). Ο ταξινομητής i εκπαιδεύεται με όλα τα παραδείγματα της κατηγορίας i ως θετικές ετικέτες, και όλα τα άλλα παραδείγματα ως αρνητικές ετικέτες. Η τελική απόφαση είναι η κατηγορία με τις περισσότερες ψήφους. Για την εκπαίδευση των νευρωνικών δικτύων με τον αλγόριθμο BP χρησιμοποιήσαμε ένα κρυφό επίπεδο.

Δεδομένα	Winnow	Voted Perceptron	BP
Autos	44.28	13.06	54.45
Badges	100.00	80.02	100.00
balance-scale	61.01	87.46	87.11
breast-c	65.00	71.53	71.25
breast-w	95.44	90.47	96.11
colic	74.84	63.56	82.25
credit-a	78.41	63.42	85.81
credit-g	67.48	70.00	72.90
diabetes	68.97	66.03	75.85
haberman	71.44	73.33	74.20
heart-c	76.22	64.02	81.18
heart-h	72.93	68.54	81.74
heart-statlog	79.30	63.93	82.07
hepatitis	71.95	78.85	80.86
ionosphere	85.21	85.50	85.10
iris	82.00	74.27	96.67
labor	76.07	72.63	89.33
lymphography	66.23	79.74	81.20
monk1	67.83	70.02	86.85
monk2	55.89	55.10	75.70
monk3	77.79	88.58	88.32
relation	61.07	77.85	78.44
sonar	71.43	70.33	74.22
vehicle	47.20	31.75	51.58
vote	92.21	95.17	94.37
wine	90.11	38.37	84.97
zoo	77.72	81.09	61.34

Πίνακας 22. Σύγκριση αλγορίθμων μάθησης νευρωνικών δικτύων

Πρέπει να σημειωθεί ότι ο αλγόριθμος BP χρησιμοποιήθηκε ως επί το πλείστον στο πειραματικό μέρος της διατριβής ως εκπρόσωπος των νευρωνικών δικτύων.

Συμπεράσματα για τον αλγόριθμο BP

- Όσο αυξάνεται ο αριθμός των δειγμάτων τόσο βελτιώνεται η εκπαίδευση, η ακρίβεια που επιτυγχάνεται ανεξάρτητα από την τοπολογία του δικτύου.
- Σε πολυεπίπεδα δίκτυα ο αριθμός των νευρώνων που απαιτείται σε κάθε επίπεδο, προκειμένου να επιτευχθεί συγκεκριμένη ακρίβεια δεν μπορεί να προβλεφθεί. Περισσότερες από μία τοπολογίες θα μπορούσαν να επιτύχουν το ίδιο αποτέλεσμα.
- Πέρα από ένα κατώφλι η αύξηση του αριθμού των νευρώνων δεν θα οδηγήσει απαραίτητα σε βελτίωση της επιθυμητής ακρίβειας.
- Όταν τίθεται ακριβέστερος στόχος σφάλματος στην εκπαίδευση τότε επιτυγχανόταν και καλύτερο μέσο σφάλμα στα αποτελέσματα.
- Δεν είναι απαραίτητη η επιδίωξη καλύτερου στόχου σφάλματος από ένα σημείο και ύστερα διότι δεν άξιζε ο υπολογιστικός χρόνος που ήταν αναγκαίος για την διεκπεραίωση της εκπαίδευσης.
- Η βέλτιστη τοπολογία εξαρτάται από την διαθέσιμη υπολογιστική ισχύ.

- Δίκτυα με μικρό αριθμό νευρώνων είναι πολύ ασταθή και η πιθανότητα για απόκλιση είναι πολύ μεγάλη.

Αν και τα νευρωνικά δίκτυα και τα δέντρα απόφασης είναι δύο πολύ διαφορετικές τεχνικές για την ταξινόμηση, μερικοί ερευνητές όπως ο Michie και άλλοι (1994) και ο Tjen - Sien Lim και άλλοι (2000) διεξήγαγαν μερικές εμπειρικές συγκριτικές μελέτες. Μερικά από τα γενικά συμπεράσματα που συνάγονται είναι ότι: 1) τα νευρωνικά δίκτυα είναι συνήθως ευκολότερο να παρέχουν επαυξητική μάθηση απ' ό,τι τα δέντρα απόφασης (Saad, 1998), 2) ο χρόνος εκπαίδευσης για ένα νευρωνικό δίκτυο είναι περισσότερος από το χρόνο εκπαίδευσης ενός δέντρου απόφασης, και 3) τα νευρικά δίκτυα συνήθως αποδίδουν όπως και τα δέντρα απόφασης και σπάνια καλύτερα.

Πρέπει επίσης να σημειωθεί ότι υπάρχουν Μπεϋζιανοί μέθοδοι για εκπαίδευση νευρωνικών δικτύων. Οι Vivarelli & Williams (2001) συγκρίνουν δύο Μπεϋζιανές μεθόδους για νευρωνικά δίκτυα. Περισσότερα για την Μπεϋζιανή προσέγγιση στην επόμενη ενότητα.

Τέλος υπάρχουν και αμφίδρομα δίκτυα (σε αντίθεση με τα μονόδρομα) όπου τα σήματα ταξιδεύουν και προς τις δυο κατευθύνσεις. Τα αμφίδρομα δίκτυα είναι αρκετά ισχυρά αλλά και πολύ πιο περίπλοκα και χρονοβόρα. Χαρακτηρίζονται από τη διαρκή τους εξέλιξη μέχρι να φτάσουν σε κάποια κατάσταση ισορροπίας. Αυτή η κατάσταση διατηρείται ως ότου νέα δεδομένα προστεθούν οπότε το δίκτυο αναζητεί μια νέα κατάσταση ισορροπίας.

3.5 Μπεϋζιανοί ταξινομητές

Η Μάθηση κατά Μπέυζ (Bayes) αποτελεί μια άλλη ιδιαίτερα δημοφιλή προσέγγιση για την επαγωγική κατασκευή ταξινομητών, αφενός διότι εκπορεύεται από τον χώρο των Πιθανοτήτων και αφετέρου διότι έχει επιδείξει σημαντικά αποτελέσματα σε ένα ευρύτατο φάσμα εφαρμογών. Η λειτουργία αυτής της κατηγορίας αλγορίθμων στηρίζεται στην υπόθεση ότι η υπό εκμάθηση έννοια σχετίζεται άμεσα με την κατανομή των πιθανοτήτων που παρουσιάζουν τα στιγμιότυπα του προβλήματος αναφορικά με την κλάση στην οποία ανήκουν. Συνεπώς, υπάρχει μια τελείως διαφορετική αντιμετώπιση του χώρου υποθέσεων. Δεν κατασκευάζεται ένας ταξινομητής ο οποίος βελτιώνεται σύμφωνα με τις επιδόσεις του στο σύνολο υποθέσεων, αλλά αναζητείται η υπόθεση με την μεγαλύτερη πιθανότητα να ταξινομεί σωστά τα στιγμιότυπα του συνόλου εκπαίδευσης.

Εν συνεχεία δίνονται κάποιες από τις βασικότερες έννοιες στο χώρο της μηχανικής μάθησης που βασίζονται στην στατιστική. Το σημαντικότερο από αυτά είναι το θεώρημα του Μπέυζ. Έστω μια υπόθεση h ενός χώρου υποθέσεων H και D το σύνολο δεδομένων που χρησιμοποιείται στην εκπαίδευση. Η πιθανότητα η υπόθεση h να ταξινομεί σωστά (ή τουλάχιστον με την επιζητούμενη ακρίβεια) τα στιγμιότυπα συμβολίζεται με $P(h)$ ενώ η πιθανότητα η υπόθεση να ταξινομεί σωστά τα στιγμιότυπα του D συμβολίζεται με $P(h|D)$ και καλείται δεσμευμένη πιθανότητα. Ο Bayes διατύπωσε το θεμελιώδες για την στατιστική θεώρημα που είναι γνωστό με το όνομά του και συνοψίζεται στον τύπο:

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

Παρατηρείται ότι η πιθανότητα της υπόθεσης που μελετάται να ταξινομεί σωστά το σύνολο εκπαίδευσης αυξάνει όταν αυξάνεται η πιθανότητα να ταξινομεί σωστά όλα πιθανά στιγμιότυπα. Στη συνέχεια θα δοθούν κάποιοι ορισμοί εννοιών που συναντώνται συχνά στις μεθόδους μηχανικής μάθησης που έχουν ως βάση το θεώρημα του Μπέυζ.

Μια υπόθεση λέγεται MAP (μέγιστη a posteriori) και συμβολίζεται h_{MAP} αν και μόνο αν

$$h_{MAP} \equiv \arg \max_{h \in H} P(h|D), \text{ ή αν εφαρμοστεί το θεώρημα Bayes } h_{MAP} = \arg \max_{h \in H} P(D|h)P(h).$$

Αν όλες οι υποθέσεις θεωρούνται ισοπίθανες, είτε λόγω της φύσης του προβλήματος, είτε για λόγους απλοποίησης, τότε το δεύτερο μέλος του προηγούμενου τύπου γίνεται $\arg \max_{h \in H} P(D|h)$ και η υπόθεση καλείται υπόθεση μέγιστης πιθανότητας h_{ML} .

Έστω λοιπόν μια συνάρτηση στόχος $c: X \rightarrow \{0,1\}$ η οποία προσεγγίζεται με τη βοήθεια ενός συνόλου εκπαιδεύσεως $D = \{\langle x_1, d_1 \rangle, \dots, \langle x_m, d_m \rangle\}$ ή απλούστερα $D = \{x_1, \dots, x_m\}$ όπου d_i η προσεγγιστική τιμή στις συνάρτησης στόχου που αντιστοιχεί στο στιγμιότυπο x_i . Η διατήρηση σταθερής σειράς στα στιγμιότυπα δεν επηρεάζει τους ισχυρισμούς. Άρα το σύνολο εκπαίδευσης μπορεί γραφεί ακόμα πιο απλά σαν $D = \langle d_1 \dots d_m \rangle$. Σύμφωνα με τον αλγόριθμο MAP η εύρεση της μέγιστης δεσμευμένης πιθανότητας h_{MAP} είναι μια απλή διαδικασία αλλά έχει τεράστιο υπολογιστικό κόστος:

- Για κάθε υπόθεση $h \in H$ πρέπει να υπολογιστεί η δεσμευμένη πιθανότητα

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

- Και εν συνεχεία να θέσουμε σαν h_{MAP} την υπόθεση με τη μεγαλύτερη υπόθεση:

$$h_{MAP} \equiv \arg \max_{h \in H} P(h|D)$$

3.5.1 Βέλτιστος ταξινομητής Bayes

Στην προηγούμενη παράγραφο δόθηκε απάντηση στο ερώτημα της πιο πιθανής υπόθεσης. Τώρα θα δοθεί απάντηση και στο ερώτημα της καταλληλότερης ταξινόμησης. Μια υπόθεση πρέπει να καλύπτει το σύνολο των στιγμιότυπων όμως κατά την αναζήτηση της κατάλληλης υπόθεσης είναι πολύ πιο απλό να κατασκευαστεί μια υπόθεση που να μπορεί να ταξινομή τα περισσότερα (και όχι όλα) στιγμιότυπα σωστά. Κατά την ταξινόμηση ενός στιγμιότυπου, προτεραιότητα έχει η ορθή ταξινόμηση του συγκεκριμένου στιγμιότυπου, χωρίς να έχει κάποια σημασία η επίδοση του ταξινομητή στο σύνολο των στιγμιότυπων.

Έστω ένα τυχαίο παράδειγμα όπου h_i είναι οι συνεπής υποθέσεις με το σύνολο υποθέσεων D .

Τότε οι ταξινόμηση ενός τυχαίου στιγμιότυπου που ανήκει στη κλάση v_i θα είναι

$$P(v_i|D) = \sum_{h_i \in H} P(v_i|h_i)P(h_i|D).$$

Η μέγιστη από αυτές τις ταξινομήσεις καλείται βέλτιστη ταξινόμηση Bayes και είναι $\arg \max_{v_i \in V} \sum_{h_i \in H} P(v_i|h_i)P(h_i|D)$.

Δεν υπάρχει μέθοδος που να μπορεί να ταξινομήσει στιγμιότυπα με την καλύτερη ακρίβεια (κατά μέσο όρο). Κατά συνέπεια ο ταξινομητής αυτός μπορεί να μας δώσει ένα άνω φράγμα ακρίβειας για όλες τις μεθόδους μηχανικής μάθησης. Επιπλέον, παρά τον συμβολισμό, μπορεί να εφαρμοστεί ακόμα και αν οι συνεπής υποθέσεις δεν περιέχονται στο χώρο υποθέσεων. Δυστυχώς όμως απαιτεί μεγάλο αριθμό πράξεων.

3.5.2 Αφελής ταξινομητής

Το μοντέλο του βέλτιστου κατά Μπέυζ ταξινομητή που παρουσιάστηκε στην προηγούμενη παράγραφο, παρ' ότι είναι σε θέση να μας υποδείξει ένα άνω φράγμα των επιδόσεων ενός συστήματος ταξινόμησης, καθίσταται ιδιαίτερα δύσκολο να χρησιμοποιηθεί στην πράξη, καθώς προϋποθέτει τον υπολογισμό ενός μεγάλου αριθμού πιθανοτήτων (Friedman, 1997). Αυτός ο υπολογισμός τις περισσότερες φορές είναι στην πράξη μη εφικτός, όχι μόνο λόγω του πλήθους των συναρτήσεων, αλλά και εξ αιτίας του μικρού μεγέθους του σώματος εκπαίδευσης που συχνά διαθέτουμε. Μια πιο πρακτική προσέγγιση της συλλογιστικής που

αναπτύχθηκε είναι δυνατή μέσω του Αφελούς Μπεϋζιανού Ταξινομητή (Naive Bayes ή NB Classifier).

Αν $\langle a_1, a_2, \dots, a_n \rangle$ είναι το σύνολο των χαρακτηριστικών των στιγμιότυπων, σύμφωνα με την προσέγγιση του Μπέυζ η κλάση ενός τυχαίου στιγμιότυπου είναι $v_{MAP} = \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n)$ ή αλλιώς $v_{MAP} = \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j) P(v_j)$. Αν

τώρα υποθεθεί ότι οι τιμές των χαρακτηριστικών είναι ανεξάρτητες μεταξύ τους, τότε η $\arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$ είναι μια απλοποιημένη μορφή με λιγότερο κόστος για τον

υπολογισμό της πιο πιθανής ταξινόμησης και καλείται αφελής ταξινομητής Μπέυζ. Το όνομα του ο τύπος το οφείλει στο γεγονός ότι η τιμή του ενός χαρακτηριστικού δεν επηρεάζει την τιμή που θα πάρει το άλλο κάτι που συχνά δεν ισχύει στην πραγματικότητα. Παρόλα αυτά οι μέθοδοι που χρησιμοποιούν τον αφελή ταξινομητή Μπέυζ έχουν απόδοση συγκρίσιμη με αυτή άλλων αλγορίθμων. (Domingos & Pazzani, 1997).

Για να γίνει περισσότερο κατανοητή η χρήση του απλοϊκού (αφελή) ταξινομητή Μπέυζ παραθέτουμε ένα παράδειγμα χρήσης του. Έστω ότι έχουμε στη διάθεσή μας στιγμιότυπα από τη βάση δεδομένων ενός καταστήματος με ηλεκτρονικά είδη (Πίνακας 23). Κάθε στιγμιότυπο αποτελείται από τα γνωρίσματα «age», «income», «student», «credit rating» και ανήκει σε κάποια από τις δύο κλάσεις του προβλήματος («Yes», «No»).

age	income	student	Credit_rating	Class: buys_computer
<=30	high	No	fair	No
<=30	high	No	excellent	No
31...40	high	No	fair	Yes
>40	medium	no	Fair	Yes
>40	low	yes	Fair	Yes
>40	low	yes	Excellent	No
31...40	low	yes	Excellent	Yes
<=30	medium	no	fair	No
<=30	low	yes	Fair	Yes
>40	medium	yes	fair	Yes
<=30	medium	yes	Excellent	Yes
31...40	medium	no	Excellent	Yes
31...40	high	yes	fair	Yes
>40	medium	no	Excellent	No

Πίνακας 23. Δεδομένα εκπαίδευσης

Έστω λοιπόν ότι θέλουμε να ταξινομήσουμε το ακόλουθο άγνωστο στιγμιότυπο του προβλήματος:

$X = (\text{age} = "<=30", \text{income} = "\text{medium}", \text{student} = "\text{yes}", \text{credit_rating} = "\text{fair}"))$

Αρχικά υπολογίζουμε τις εκ των προτέρων (a priori) πιθανότητες των δύο κλάσεων του προβλήματος:

$P(\text{buys_computer} = "\text{yes}") = 9/14 = 0.643$

$P(\text{buys_computer} = "\text{no}") = 5/14 = 0.357$

Στη συνέχεια, υπολογίζουμε τις υπό συνθήκη πιθανότητες για κάθε μεταβλητή και όλες τις κλάσεις του προβλήματος:

$$P(\text{age} \leq 30 | \text{buys_computer} = \text{"yes"}) = 2/9 = 0.222$$

$$P(\text{age} \leq 30 | \text{buys_computer} = \text{"no"}) = 3/5 = 0.600$$

$$P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"yes"}) = 4/9 = 0.444$$

$$P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.400$$

$$P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$$

$$P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"no"}) = 1/5 = 0.200$$

$$P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$$

$$P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.400$$

Χρησιμοποιώντας τις παραπάνω πιθανότητες υπολογίζουμε την πιθανότητα το άγνωστο στιγμιοτύπο X να ανήκει σε κάποια από τις δύο κλάσεις του προβλήματος:

$$P(X | \text{buys_computer} = \text{"yes"}) = 0.222 * 0.444 * 0.667 * 0.667 = 0.044$$

$$P(X | \text{buys_computer} = \text{"no"}) = 0.600 * 0.400 * 0.200 * 0.400 = 0.019$$

$$P(X | \text{buys_computer} = \text{"yes"}) * P(\text{buys_computer} = \text{"yes"}) = 0.044 * 0.643 = 0.028$$

$$P(X | \text{buys_computer} = \text{"no"}) * P(\text{buys_computer} = \text{"no"}) = 0.019 * 0.357 = 0.007$$

Αφού $P(X | \text{buys_computer} = \text{"yes"}) * P(\text{buys_computer} = \text{"yes"}) > P(X | \text{buys_computer} = \text{"no"}) * P(\text{buys_computer} = \text{"no"})$ η κλάση "Yes", είναι η απάντηση που ψάχνουμε.

Παρ' όλη την απλοποίηση των υπολογισμών που μας εξασφαλίζει η υπόθεση της στοχαστικής ανεξαρτησίας, δεν αντιμετωπίζει το πρόβλημα της απουσίας συγκεκριμένων τιμών των χαρακτηριστικών από τα στιγμιοτύπα εκπαίδευσης κάποιας κλάσης. Αυτό το ενδεχόμενο οδηγεί στο μηδενισμό της αντίστοιχης δεσμευμένης πιθανότητας και κατ' επέκταση ολόκληρου του γινομένου, καθιστώντας έτσι τις προβλέψεις του ταξινομητή μη αξιόπιστες. Η δυσκολία αυτή μπορεί να ξεπεραστεί ωστόσο, με τη χρήση του εκτιμητή Laplace, ο οποίος ορίζεται ως ακολούθως:

➤ Αν $n_c(a_i = a)$ το πλήθος των στιγμιοτύπων που έχουν σαν τιμή του a_i την τιμή a και ανήκουν στην κλάση c και n_c το πλήθος των στιγμιοτύπων που ανήκουν στην κλάση

αυτή, τότε ο εκτιμητής Laplace είναι $\frac{n_c(a_i = a) + \mu p}{n_c + \mu}$ όπου p είναι η εκ των προτέρων

εκτίμηση της προς εκτίμησης πιθανότητας και μ καλείται ισοδύναμο μέγεθος δείγματος και συμβολίζει τη βαρύτητα της εκ των προτέρων πιθανότητας σε σχέση με την πιθανότητα που δίνεται στην τιμή του χαρακτηριστικού κατά την εκπαίδευση.

Γενικά, ο εκτιμητής Laplace μπορεί να βοηθήσει στην περίπτωση που οι τιμές ενός χαρακτηριστικού είναι λίγες στο πλήθος. Όμως, ο προσδιορισμός των εκ των υστέρων πιθανοτήτων $Pr(a_i | c_j)$ μέσω του εκτιμητή Laplace είναι εφικτός μόνο για χαρακτηριστικά που παίρνουν ένα μικρό σύνολο από διακριτές τιμές. Στην περίπτωση που τα χαρακτηριστικά είναι συνεχή, θεωρούμε πως οι τιμές τους ακολουθούν σε κάθε κλάση μια πιθανοτική κατανομή, με αποτέλεσμα οι πιθανότητες $Pr(a_i | c_j)$ να αποτελούν στην πραγματικότητα συναρτήσεις πυκνότητας πιθανότητας (*probability density functions* ή *pdfs*). Όταν δεν γνωρίζουμε την κατανομή που ακολουθούν οι τιμές ενός χαρακτηριστικού, θεωρούμε, χωρίς

βλάβη της γενικότητας, εξ ορισμού την κανονική κατανομή. Ως μέση τιμή $\mu_{i,cj}$ και διασπορά $\sigma_{i,cj}$ της λαμβάνουμε τη μέση τιμή και τη διασπορά αντίστοιχα των τιμών του δεδομένου χαρακτηριστικού a_i στα παρατηρηθέντα στιγμιότυπα του σώματος εκπαίδευσης που ανήκουν στην κλάση c . Οι δύο αυτές παράμετροι μπορούν να προσεγγιστούν με τη βοήθεια του *Εκτιμητή Μέγιστης Πιθανοφάνειας* (*Maximum Likelihood Estimator*).

Αν $X = \{x_1, x_2, \dots, x_n\}$ είναι το σύνολο των τιμών κάποιου χαρακτηριστικού και αυτά ακολουθούν μια κατανομή με συνάρτηση πυκνότητας πιθανότητας $f(\vec{x}; \vec{\theta})$ όπου $\vec{\theta}$ το άγνωστο διάνυσμα των παραμέτρων που χαρακτηρίζουν την κατανομή, τότε σύμφωνα με τη μέθοδο μέγιστης πιθανοφάνειας $\hat{\theta}_{ML} = \arg \max_{\theta} \prod_{i=1}^n f(x_i; \vec{\theta})$ και κατά συνέπεια η πιθανότητα που αποδίδει ο αφελής εκτιμητής Bayes σε ένα χαρακτηριστικό c είναι

$$c_{NB} = \arg \max_{c_j \in C} \Pr(c_j) \prod_{i=0}^{n-1} f(a_i; \mu_{i,c_j}, \sigma_{i,c_j}) = \arg \max_{c_j \in C} \Pr(c_j) \prod_{i=0}^{n-1} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2}$$

Μια άλλη μέθοδος αντιμετώπισης των συνεχών τιμών στα χαρακτηριστικά είναι η διακριτοποίηση τους πριν την εκπαίδευση (Yang & Webb, 2003).

Μια επέκταση του Αφελή Ταξινομητή Bayes είναι ο Ευέλικτος Ταξινομητής Bayes (Flexible Bayes ή FB Classifier). Αυτός ο ταξινομητής είναι ικανός να προσδιορίσει με μεγαλύτερη ακρίβεια την άγνωστη κατανομή τιμών των συνεχών χαρακτηριστικών ενός προβλήματος, επιχειρώντας να προσεγγίσει την πυκνότητα πιθανότητάς τους μέσω κανονικών πυρήνων (Gaussian Kernels). Με αυτό τον τρόπο, ουσιαστικά επιτυγχάνεται να υπολογιστούν με μεγαλύτερη ακρίβεια η μέση τιμή και η διασπορά της κατανομής των τιμών κάθε χαρακτηριστικού για την εκάστοτε κλάση.

Όπως γίνεται αντιληπτό, η χρήση του FB πλεονεκτεί κατά πολύ εκείνης του NB, στην περίπτωση που η κατανομή των συνεχών τιμών των χαρακτηριστικών του χώρου ενός προβλήματος δεν προσεγγίζεται από την κανονική κατανομή. Ως αντιστάθμισμα της βελτίωσης στην απόδοση, θα πρέπει να ληφθεί υπ' όψη η αυξημένη πολυπλοκότητα της μεθόδου κατά το στάδιο της ταξινόμησης ενός άγνωστου στιγμιότυπου σε:

- χώρο (όσο χρειάζεται για να αποθηκεύσει όλες τις διαφορετικές τιμές που συναντά στα στιγμιότυπα εκπαίδευσης μιας συγκεκριμένης κλάσης που αντιστοιχούν σε κάθε χαρακτηριστικό, καθώς αποτελούν τη μέση τιμή της εκάστοτε επιμέρους κατανομής), και
- σε χρόνο (όσο χρειάζεται για τον υπολογισμό των τιμών των επιμέρους συναρτήσεων πυκνότητας πιθανότητας του κάθε χαρακτηριστικού, για κάθε στιγμιότυπο εκπαίδευσης της δεδομένης κλάσης).

Το βασικό αφελές μοντέλο Bayes έχει επίσης τροποποιηθεί με ποικίλους τρόπους στην προσπάθεια να βελτιωθεί η απόδοση του. Απόπειρες να ξεπεραστεί η υπόθεση της ανεξαρτησίας των μεταβλητών είναι κυρίως βασισμένες στο να συμπεριληφθούν μερικές εξαρτήσεις μεταξύ των μεταβλητών, π.χ (Friedman και άλλοι, 1997). Σ' αυτή την περίπτωση, κάθε χαρακτηριστικό σχετίζεται το πολύ με ένα ακόμα χαρακτηριστικό.

Μια εναλλακτική γενίκευση που λέγεται «επιλεκτικός Μπεϋζιανός ταξινομητής» έχει αναπτυχθεί από τον (Kontkanen και άλλοι, 1999). Αυτός περιέλαβε ένα στάδιο επιλογής χαρακτηριστικών, έτσι ώστε το τελικό τους μοντέλο να μην χρειάζεται να περιλαμβάνει όλα τα χαρακτηριστικά. Η ιδέα αυτού του αλγόριθμου είναι να βελτιώσει την απόδοση της ταξινόμησης με την απομάκρυνση των περιττών χαρακτηριστικών. Άλλοι ερευνητές έχουν διερευνήσει τη δυνατότητα να χρησιμοποιηθεί ένας αλγόριθμος δέντρων απόφασης ως προ-επεξεργασία των δεδομένων για να ανακαλύψει τα χρήσιμα υποσύνολα χαρακτηριστικών για τον NB. Μια απλή μέθοδος που χρησιμοποιεί τον C4.5 ως δέντρο απόφασης για να επιλέξει τα χαρακτηριστικά έχει περιγραφεί από τους (Ratanamahatana & Gunopulos, 2003).

Μια επαναληπτική προσέγγιση του απλού ταξινομητή Bayes παρουσιάζεται στην εργασία (Gama, 2000). Η επαναληπτική προσέγγιση χτίζει τους πίνακες πιθανοτήτων με τον απλό Μπεϋζιανό ταξινομητή και έπειτα τους ανανεώνει πραγματοποιώντας κύκλους του αλγορίθμου σε όλα τα παραδείγματα εκπαίδευσης χρησιμοποιώντας μια τεχνική αναρρίχησης λόφου. Η πειραματική αξιολόγηση της επαναληπτικής προσέγγισης παρουσίασε μικρή αλλά συνεπής αύξηση στην ακρίβεια ταξινόμησης σε σχέση με τον απλό Μπεϋζιανό ταξινομητή (Gama, 2000). Εντούτοις, οι πίνακες πιθανότητας ενημερώνονται επαυξητικά κάθε φορά που φαίνεται ένα παράδειγμα εκπαίδευσης, το οποίο υπονοεί ότι η διάταξη των παραδειγμάτων θα μπορούσε να επηρεάσει την τελική απόφαση.

3.5.3 Σύγκριση αλγορίθμων

Σε αυτό το υποκεφάλαιο παραθέτουμε την σύγκριση των γνωστών αφελών Μπεϋζιανών αλγορίθμων σε ένα πλήθος από σύνολα δεδομένων εκπαίδευσης. Έχουμε χρησιμοποιήσει για το αφελή ταξινομητή με την αντιμετώπιση των συνεχών χαρακτηριστικών με: α) την κανονική κατανομή, β) την διακριτοποίηση, γ) των κανονικών πυρήνων.

Οι κανονικοί πυρήνες φαίνεται να αποδίδουν ελαφρά καλύτερα βάση των πειραμάτων μας (βλέπε Πίνακας 24). Επίσης, εξετάστηκε η ακρίβεια του αλγόριθμου AODE (Webb και άλλοι, 2002) που εκπαιδεύει μια ομάδα μπεϋζιανών ταξινομητών χρησιμοποιώντας ένα υποσύνολο των δεδομένων εκπαίδευσης για τον καθένα. Ο συνδυασμός των μοντέλων μέσω της ψηφοφορίας αποδίδει καλύτερα από τους υπόλοιπους εξεταζόμενους αλγόριθμους αλλά ανεβάζει παράλληλα το υπολογιστικό κόστος.

Το μεγάλο πλεονέκτημα του αφελή ταξινομητή Μπεϋζ είναι ο μικρός υπολογιστικός χρόνος που απαιτείται για την εκπαίδευσή του. Παράλληλα όμως ο απλοϊκός ταξινομητής Μπεϋζ, 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. Στην εργασία των (Domingos & Pazzani, 1997) παρέχεται μια ενδιαφέρουσα ανάλυση για το ευτυχές αυτό φαινόμενο. Ένα ακόμα πλεονέκτημα του NB είναι η σχετική απλότητα των μοντέλων που κατασκευάζει, τα οποία μπορούν να γίνουν εύκολα κατανοητά από τον άνθρωπο, ιδιαίτερα μέσω της οπτικοποίησης (Becker και άλλοι, 1997).

Η κατανόηση του μοντέλου αυξάνει γενικά την εμπιστοσύνη των χρηστών σε ένα σύστημα σε σχέση με τη θεώρηση του τελευταίου ως “μαύρου κουτιού” που δέχεται στην είσοδο στιγμιότυπα και επιστρέφει στην έξοδο προβλέψεις για αυτά. Παράλληλα, οι χρήστες μπορούν να απορρίψουν προβλέψεις του μοντέλου ή και ολόκληρο το μοντέλο αν κρίνουν πως αυτό βασίζεται σε ασήμαντους ή άσχετους παράγοντες ή αγνοεί άλλους περισσότερο κρίσιμους.

Δεδομένα	NB (κανονική κατανομή)	NB (με διακριτοποίηση)	AODE	NaiveBayes (κανονικοί πυρήνες)
autos	55.92	60.11	70.91	61.55
badges	99.80	100.00	100.00	99.90
balance-scale	89.82	74.22	74.08	89.82
breast-c	72.73	72.73	71.60	72.73
breast-w	96.09	97.23	96.97	97.35
colic	78.83	79.71	82.04	79.31
credit-a	77.26	86.26	86.73	80.80
credit-g	75.74	73.65	74.94	74.56
diabetes	74.96	74.85	75.38	74.92
haberman	73.77	72.62	72.62	73.29
heart-c	84.37	84.27	83.69	84.66
heart-h	83.71	83.41	83.21	83.71
heart-statlog	84.40	82.98	83.20	83.53
hepatitis	81.53	83.22	84.36	83.24
ionosphere	83.05	89.17	91.27	92.03
iris	96.08	94.12	94.12	96.86
labor	93.28	89.17	91.67	91.23
lymphography	82.14	81.76	84.74	81.54
monk1	68.65	68.65	78.39	68.65
monk2	56.86	56.86	60.14	56.86
monk3	93.74	93.74	93.75	93.74
relation	78.20	78.20	78.40	78.20
sonar	67.92	75.85	75.99	72.30
students	87.54	87.54	87.89	87.54
vehicle	44.95	59.74	68.76	59.32
vote	89.04	89.04	94.12	89.04
wine	96.71	97.70	97.37	96.54
zoo	96.14	94.44	95.90	96.44

Πίνακας 24. Σύγκριση απλοϊκών ταξινομητών Μπέυζ

3.5.4 Μπεϋζιανά Δίκτυα

Ο αφελής Μπεϋζιανός ταξινομητής κάνει την περιοριστική υπόθεση ότι τα χαρακτηριστικά των στιγμιοτύπων δεν αλληλεξαρτώνται. Κάτι τέτοιο όμως στην πράξη συμβαίνει σπάνια. Ένα Μπεϋζιανό δίκτυο (Bayesian Network) περιγράφει τη δομή αλληλεξάρτησης των χαρακτηριστικών, ενώ μπορεί και να περιγράψει χαρακτηριστικά που είναι ανεξάρτητα με μια πιο κατανοητή και βολική δομή απ' ότι ο αφελής Μπεϋζιανός ταξινομητής (Jordan, 1998).

Το Μπεϋζιανό δίκτυο μπορεί να οριστεί ως ένα ζεύγος (G, p) όπου $G=(V, E)$ είναι ένα κατευθυνόμενο άκυκλο γράφημα (directed acyclic graph-DAG) του οποίου οι κόμβοι αναπαριστούν τυχαίες μεταβλητές και οι σύνδεσμοι αντιπροσωπεύουν τις αιτιολογικές επιδράσεις μεταξύ των μεταβλητών. Επίσης, p είναι μια διακριτή συνάρτηση πιθανότητας. Πρέπει να σημειωθεί λοιπόν ότι η ισχύς της επίδρασης αναπαρίσταται από δεσμευμένες πιθανότητες. Στο γράφημα αυτό οι μεταβλητές είναι υπό συνθήκη ανεξάρτητες από αυτές με τις οποίες δεν ενώνονται, δεδομένων των γονιών τους.

Αν X, Y, Z είναι τρεις τυχαίες διακριτές μεταβλητές, τότε λέμε ότι η X είναι υπό συνθήκη ανεξάρτητη της Y δεδομένης της Z αν $(\forall x_i, y_j, z_k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$ ή για συντομία $P(X | Y, Z) = P(X | Z)$.

Γενικότερα, δεδομένου ενός κατευθυνόμενου άκυκλου γραφήματος G και μιας από κοινού κατανομής P των διακριτών μεταβλητών $X_1, X_2, X_3, X_4, \dots, X_n$ λέμε ότι το G αντιπροσωπεύει την P μόνο εάν υπάρχει μια '1-1' σχέση μεταξύ των μεταβλητών X και των κόμβων του G , έτσι ώστε το P να παίρνει την εξής μορφή: $P(X_1, X_2, X_3, X_4, \dots, X_n) = \prod_{i=1}^n (P(X_i | pa_i))$

Όπου pa_i (γονείς) είναι οι κομβοί που βρίσκονται στο αμέσως προηγούμενο επίπεδο από την μεταβλητή X και συνδέονται με αυτήν, στο δίκτυο G .

Στην Εικόνα 35 παρουσιάζουμε μια δομή δικτύου για 5 μεταβλητές και παρακάτω δείχνουμε πως υπολογίζονται οι δεσμευμένες πιθανότητες επιρροής μιας μεταβλητής σε μια άλλη.

Εικόνα 35. Το δίκτυο Bayes του παραδείγματος

$$\begin{aligned}
 p(a | c) &= p(a) \\
 p(s | c, a) &= p(s) \\
 p(b | c, a, s) &= p(b | c, a) \\
 p(d | c, a, s, b) &= p(d | c, s, b)
 \end{aligned}$$

$$\begin{aligned}
 p(c | a, s, b, d) &= \frac{p(c, a, s, b, d)}{p(a, s, b, d)} = \frac{p(c, a, s, b, d)}{\sum_{c'} p(c', a, s, b, d)} \\
 &= \frac{p(c)p(a)p(s)p(b | c, a)p(d | c, s, b)}{\sum_{c'} p(c')p(a)p(s)p(b | c', a)p(d | c', s, b)} \\
 &= \frac{p(c)p(b | c, a)p(d | c, s, b)}{\sum_{c'} p(c')p(b | c', a)p(d | c', s, b)}
 \end{aligned}$$

Μέσα στο γενικό πλαίσιο της εκμάθησης δικτύων Bayes υπάρχουν δύο σενάρια:

- Να είναι γνωστή η δομή του δικτύου (π.χ. ένας ειδικός ξέρει τις εξαρτήσεις των μεταβλητών)
- Να είναι άγνωστη η δομή

Στο πρώτο σενάριο, η δομή του δικτύου είναι δεδομένη (π.χ. από έναν ειδικό) και θεωρείται ότι είναι σωστή και τα δεδομένα περιέχουν παρατηρήσεις για όλα τα χαρακτηριστικά. Στην

περίπτωση αυτή απλώς υπολογίζονται οι δεσμευμένες πιθανότητες από τις τιμές των μεταβλητών όπως στο προηγούμενο παράδειγμα (Jensen 1996).

Σε ένα υπο-σενάριο της πρώτης περίπτωσης η δομή του δικτύου δίνεται και θεωρείται ότι είναι σωστή αλλά τα δεδομένα εκπαίδευσης δεν είναι πλήρη. Οι Ramoni & Sebastiani (2001) πρότειναν μια μέθοδο που την αποκαλούν εύρωστο εκτιμητή Bayesian (RBE), ο οποίος μαθαίνει από ατελή δεδομένα.

Παρόλη την αξιοσημείωτη δύναμη των δικτύων Bayes, υπάρχει ένας έμφυτος περιορισμός. Αυτή είναι η υπολογιστική δυσκολία να βρεθεί το βέλτιστο δίκτυο στην περίπτωση που αυτό δεν είναι γνωστό. Δοθέντος ενός προβλήματος που περιγράφεται από n μεταβλητές, ο αριθμός των πιθανών δικτύων είναι εκθετικός του n . Στην περίπτωση που η δομή είναι άγνωστη, μια προσέγγιση είναι να προτείνουμε ένα παράγοντα βαθμολόγησης (ή ένα σκορ) που εκτιμά την καταλληλότητα των δικτύων σε σχέση με τα δεδομένα εκπαίδευσης, και έπειτα να ψάξουμε για το καλύτερο δίκτυο (σύμφωνα μ' αυτό το σκορ). Πολλοί ερευνητές έχουν αποδείξει πειραματικά ότι η επιλογή μιας μόνης καλής υπόθεσης χρησιμοποιώντας άπληστη αναζήτηση συχνά αποφέρει ακριβές προβλέψεις (Heckerman και άλλοι, 1995), (Chickering, 1996). Σε κάθε βήμα της άπληστης αναζήτησης είτε προστίθεται μια ακμή για την σύνδεση δύο μεταβλητών είτε σβήνεται μια ήδη υπάρχουσα είτε αντιστρέφεται η φορά της εξάρτησης (βλέπε Εικόνα 36). Εκτός της άπληστης αναζήτησης πλήθος άλλων αλγορίθμων έχουν προταθεί στην βιβλιογραφία (Blanco και άλλοι, 2003).

Εικόνα 36. Παράδειγμα άπληστης αναζήτησης βέλτιστου Μπεϋζιανού δικτύου

Όπως προείπαμε, ορίζεται λοιπόν μια βαθμολογία σφάλματος και θέλουμε να βρούμε τη δομή που ελαχιστοποιεί τη βαθμολογία αυτή. Μια τέτοια βαθμολογία δίνεται από τον τύπο:

$$L'(\Xi, B) = - \sum_{i=1..m} \log p(v_i) + \frac{|B| \log m}{2}$$

με όπου m = Αριθμός εμφανίσεων

$\log$ = ο λογάριθμος με βάση το δύο

$|B|$ = Αριθμός των διαφορετικών συνδυασμών τιμών αλήθειας

$p(v_i)$ = Η από κοινού πιθανότητα για κάθε διαφορετικό συνδυασμό τιμών αλήθειας

Για να γίνει περισσότερο πως υπολογίζεται η βαθμολογία της δομής ενός δικτύου παραθέτουμε ένα παράδειγμα. Έστω τα δεδομένα εκπαίδευσης του Πίνακας 25.

G	M	B	L	Αριθμός εμφανίσεων
True	True	True	True	54
True	True	True	False	1
True	False	True	True	7
True	False	True	False	27
False	True	True	True	3
False	False	True	False	2
False	False	False	True	4
False	False	False	False	2

Πίνακας 25. Τα δεδομένα εκπαίδευσης του παραδείγματος μας

Αν λοιπόν βαθμολογήσουμε το δίκτυο της Εικόνα 37 θα πάρουμε:

Εικόνα 37. Το δίκτυο του παραδείγματος

$$\begin{aligned}
L'(\Xi, B) = & - \sum_{i=1..54} \log(p(G|B)p(M|B,L)p(B)p(L)) - \sum_{i=55} \log(p(G|B)p(M|B,\sim L)p(B)p(\sim L)) - \\
& - \sum_{i=56..62} \log(p(G|B)p(\sim M|B,L)p(B)p(L)) - \sum_{i=63..89} \log(p(G|B)p(\sim M|B,\sim L)p(B)p(\sim L)) - \\
& - \sum_{i=90..92} \log(p(\sim G|B)p(M|B,L)p(B)p(L)) - \sum_{i=93..94} \log(p(\sim G|B)p(\sim M|B,\sim L)p(B)p(\sim L)) - \\
& - \sum_{i=95..98} \log(p(\sim G|\sim B)p(\sim M|\sim B,L)p(\sim B)p(L)) - \sum_{i=99..100} \log(p(\sim G|\sim B)p(\sim M|\sim B,\sim L)p(\sim B)p(\sim L)) + \\
& + \frac{8 \log 100}{2} = 223.26
\end{aligned}$$

Πρέπει να σημειωθεί ότι πληθώρα άλλων μαθηματικών τύπων βαθμολογίας έχουν προταθεί στην βιβλιογραφία (Cheng. & Greiner, 2001). Κανένας όμως δεν λειτουργεί καλύτερα από τους άλλους για όλα τα προβλήματα.

Πρόσφατα η προσέγγιση που κυριαρχεί χρησιμοποιεί παράλληλα με το σκορ τοπικές μεθόδους έρευνας στο διάστημα των κατευθυνόμενων άκυκλων γραφήματων (Provost & Fawcett, 2001). Οι συνηθισμένες επιλογές καθορισμού στοιχειωδών τροποποιήσεων (τοπικές αλλαγές) οι οποίες μπορούν να εφαρμοστούν είναι η προσθήκη ακμής, η αφαίρεση ακμής και η αναστροφή της φοράς της ακμής. Οι Acid & de Campos (2003) πρότειναν μια μέθοδο τοπικής αλλαγής.

Ένας άλλος τρόπος αναζήτησης της δομής του δικτύου είναι μαθαίνοντας τις υπό προϋπόθεσης ανεξάρτητες σχέσεις ανάμεσα στις μεταβλητές (Cheng και άλλοι, 2002). Χρησιμοποιώντας μερικά στατιστικά τεστ (όπως X^2 και κέρδους εντροπίας) κάποιος μπορεί να βρει τις υπό προϋποθέσεις ανεξάρτητες σχέσεις ανάμεσα στις μεταβλητές και να χρησιμοποιήσει αυτές τις σχέσεις σαν πληροφορίες για να δομήσει ένα Μπεϋζιανό δίκτυο.

Αυτοί οι αλγόριθμοι αναφέρονται και ως περιοριστικοί αλγόριθμοι (CI-based ή constraint-based).

Μια σύγκριση των μεθόδων που βασίζονται στη βαθμολογία και των περιοριστικών, παρουσιάζεται στην εργασία (Heckerman και άλλοι, 1999), η οποία δείχνει ότι οι μέθοδοι που βασίζονται στη βαθμολογία έχουν πλεονεκτήματα σε μικρά σύνολα δεδομένων εκπαίδευσης. Ωστόσο, ο Friedman και άλλοι (1997) έχει δείξει χρησιμοποιώντας θεωρητική ανάλυση ότι οι γενικές μέθοδοι που βασίζονται στη βαθμολογία μπορεί να έχουν ως αποτέλεσμα φτωχούς σε ακρίβεια ταξινομητές.

3.5.5 Σύγκριση αλγορίθμων κατασκευής Μπεϋζιανών δικτύων

Σε αυτό το υποκεφάλαιο παραθέτουμε την σύγκριση των γνωστών τεχνικών δημιουργίας Μπεϋζιανών δικτύων σε ένα πλήθος από σύνολα δεδομένων εκπαίδευσης. Έχουμε χρησιμοποιήσει: α) την προσέγγιση του ολικού σκορ (βαθμολογία), β) την προσέγγιση του ολικού σκορ με τοπικές αλλαγές, γ) την περιοριστική προσέγγιση.

Δεδομένα	Ολικό σκορ	Τοπικό σκορ	Περιοριστική προσέγγιση
badges	100	100	100
balance-scale	71.56	71.56	69.59
breast-c	71.96	74.73	74.75
breast-w	96.90	97.20	87.29
colic	82.39	81.55	76.66
credit-a	85.99	86.20	74.33
credit-g	74.96	73.39	69.56
diabetes	75.27	75.25	73.53
grub-damage	49.59	49.66	38.76
haberman	71.57	71.57	71.57
heart-c	82.65	83.24	79.13
heart-h	84.67	84.09	76.29
heart-statlog	83.26	82.56	79.33
hepatitis	82.47	84.19	75.85
iris	93.80	93.20	93.13
labor	88.50	89.70	76.57
lymphography	85.98	84.23	72.57
monk1	72.67	72.04	71.07
monk2	56.19	62.13	62.13
monk3	93.45	93.45	93.45
relation	77.86	77.86	78.95
students	86.67	87.25	84.99
vote	94.21	90.14	90.39
wine	98.43	98.65	82.09
zoo	94.65	94.56	87.99

Πίνακας 26. Σύγκριση αλγορίθμων κατασκευής Μπεϋζιανών δικτύων

Βάσει των πειραμάτων μας η τοπική βαθμολογία αποδίδει καλύτερα. Πάντως, όταν ο αριθμός μεταβλητών είναι μεγάλος, οι περιοριστικοί αλγόριθμοι είναι συνήθως αποδοτικότεροι (Friedman & Koller, 2003). Όταν όμως το μέγεθος των δεδομένων εκπαίδευσης είναι μικρό

και υπάρχει θόρυβος στα δεδομένα, οι αλγόριθμοι που βασίζονται στη βαθμολογία μπορούν συχνά να δώσουν τα ακριβέστερα αποτελέσματα (Zhang, 1998).

Ο χρόνος εκπαίδευσης του απλού Μπεϋζιανού ταξινομητή είναι σημαντικά λιγότερος σε σχέση αυτόν ενός Μπεϋζιανού δικτύου. Όμως, αν η εκπαίδευση έχει λάβει χώρα, ο χρόνος απόκρισης για τη διαδικασία εξαγωγής συμπεράσματος για την κλάση ενός νέου παραδείγματος είναι ισότιμος. Μάλιστα, όταν έχουμε πολλές μεταβλητές τα Μπεϋζιανά δίκτυα υπερτερούν διότι εξετάζουν μόνο εκείνες τις μεταβλητές που επηρεάζουν την κλάση, οπότε στη χειρότερη περίπτωση που την επηρεάζουν όλες, τα Μπεϋζιανά δίκτυα αποκρίνονται το πολύ στον ίδιο χρόνο με τον απλό Μπεϋζιανό ταξινομητή.

3.6 Μηχανές διανυσμάτων υποστήριξης

Θα ολοκληρώσουμε την παρουσίαση των αλγορίθμων μάθησης με τις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines ή SVMs), ένα είδος συγκεκριμένου γραμμικών μοντέλων και μάθησης βασισμένης σε στιγμιότυπα (Scholkopf, 1999). Στόχος της μεθόδου αυτής είναι η επιλογή ενός μικρού αριθμού στιγμιοτύπων εκπαίδευσης από κάθε κλάση, των διανυσμάτων υποστήριξης (support vectors), που συνορεύουν στο χώρο του προβλήματος με στιγμιότυπα άλλων κλάσεων. Τα επιλεγμένα στιγμιότυπα χρησιμοποιούνται για την κατασκευή μιας γραμμικής συνάρτησης διάκρισης (discriminant function), ικανής να τα διαχωρίσει όσο το δυνατόν καλύτερα. Μια εξαιρετική ανασκόπηση για τα SVMs μπορεί να βρεθεί στην εργασία του (Burges, 1998) και σε ένα πιο πρόσφατο βιβλίο από τους (Cristianini και Shawe–Taylor, 2000).

Ξεκινάμε την παρουσίαση των Μηχανών Διανυσμάτων Υποστήριξης με την υπόθεση γραμμικώς διαχωρίσιμης λύσης.

3.6.1 Διαχωρίσιμη περίπτωση

Υιοθετώντας για μια ακόμη φορά το μοντέλο της διανυσματικής αναπαράστασης του χώρου ενός προβλήματος, με $\{x_i, y_i\}$, $i=1, \dots, k$, $y_i \in \{-1, 1\}$ και λόγω της γραμμικής διαχωρισιμότητας των κλάσεων που υποθέσαμε, μπορούμε να βρούμε ένα υπερεπίπεδο Π το οποίο να διαχωρίζει κατά βέλτιστο τρόπο τα διανύσματα εκπαίδευσης, με εξίσωση $w \cdot x + b = 0$.

Έστω d_+ και d_- η ελάχιστη απόσταση του υπερεπιπέδου διαχώρισης από το πλησιέστερο θετικό ή αρνητικό παράδειγμα αντίστοιχα. Για τη γραμμική διαχωρίσιμη περίπτωση ο αλγόριθμος εκπαίδευσης διανυσμάτων υποστήριξης αναζητά τα υπερεπίπεδα με τη μεγαλύτερη απόσταση (margin) μεταξύ των κλάσεων. Το παραπάνω παίρνει την εξής μορφή προβλήματος:

$$x_i \cdot w + b \geq +1 \quad \text{για } y_i = +1 \quad (1)$$

$$x_i \cdot w + b \leq -1 \quad \text{για } y_i = -1 \quad (2)$$

τα οποία μπορούν να συνδυαστούν σε μια πιο συμπαγή μορφή:

$$y_i (x_i \cdot w + b) - 1 \geq 0 \quad \forall i \quad (3)$$

Τα «οριακά» σημεία που ορίζονται από την (1) βρίσκονται στο υπερεπίπεδο H_1 : $x_i \cdot w + b = +1$. Ανάλογα τα «οριακά» σημεία που ικανοποιούν την (2) βρίσκονται στο υπερεπίπεδο H_2 : $x_i \cdot w + b = -1$. Έτσι d_+ , $d_- = 1/\|w\|$ και το περιθώριο είναι $2/\|w\|$. Σημειώνουμε δε ότι τα H_1 , H_2 είναι προφανώς παράλληλα και ότι πλέον είναι πολύ εύκολο να βρούμε ένα ζεύγος από υπερεπίπεδα που δίνει τη μεγαλύτερη απόσταση μεταξύ των κλάσεων ελαχιστοποιώντας το $\|w\|^2$.

Όπως περιμένει λοιπόν ο καθένας η λύση μιας τυπικής περίπτωσης σε δύο διαστάσεις έχει τη μορφή που παρουσιάζεται στο (Εικόνα 38). Τα σημεία μάθησης για τα οποία η ανισότητα (3) ισχύει και η αφαίρεση των οποίων θα αλλάξει τη λύση που βρίσκουμε καλούνται διανύσματα

υποστήριξης (Support Vectors) και παρουσιάζονται στην Εικόνα 38 μέσα σε κύκλους. Θα πρέπει τέλος να σημειώσουμε ότι για την περίπτωση που εξετάζουμε, στην περιοχή που ορίζεται από τα H_1 και H_2 , και που ονομάζεται περιθώριο (margin), δεν αντιστοιχίζεται κανένα διάνυσμα εκπαίδευσης.

Εικόνα 38. Γραμμικά υπερεπίπεδα για τη διαχωρίσιμη περίπτωση. Τα διανύσματα υποστήριξης είναι μέσα σε κύκλο

Τώρα θα αντιμετωπίσουμε το πρόβλημα της ελαχιστοποίησης του $\|w\|^2$ με τη βοήθεια των πολλαπλασιαστών Lagrange. Θα αναρωτηθεί κανείς γιατί είναι απαραίτητη μια τέτοια θεώρηση. Υπάρχουν δύο βασικοί λόγοι για τους οποίους κάνουμε αυτήν την θεώρηση .

1. Οι περιορισμοί (3) θα αντικατασταθούν από πολλαπλασιαστές Lagrange που είναι πολύ πιο εύκολο να αντιμετωπιστούν κατά τη διάρκεια λύσης του προβλήματος.
2. Με αυτή την αναδιατύπωση του προβλήματος τα δεδομένα εκπαίδευσης θα παρουσιαστούν υπό τη μορφή κουκκίδων μεταξύ διανυσμάτων. Αυτή είναι μια κρίσιμη ιδιότητα η οποία και θα μας επιτρέψει να γενικεύσουμε τη διαδικασία και στη μη γραμμική περίπτωση.

Εισάγουμε λοιπόν τους θετικούς πολλαπλασιαστές Lagrange α_i με $i=1,\dots,k$ δηλαδή έναν για κάθε ένα από τους περιορισμούς του (3). Έτσι παίρνουμε τη Lagrangian

$$L_p \equiv 1/2 \|w\|^2 - \sum_{i=1}^k \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^k \alpha_i \quad (4).$$

Για την ελαχιστοποίηση του L_p ως προς w και b πρέπει να ισχύει $\frac{\partial L_p(w,b)}{\partial w} = 0$ και

$$\frac{\partial L_p(w,b)}{\partial b} = 0. \text{ Κατά συνέπεια } w = \sum_{i=1}^k \alpha_i y_i x_i \text{ και } \sum_{i=1}^k \alpha_i y_i = 0.$$

Με αντικατάσταση των παραπάνω εκφράσεων στην Lagrangian προκύπτει ένα δυϊκό πρόβλημα μεγιστοποίησης της έκφρασης $L_d(a) = \sum_{i=1}^k \alpha_i - \sum_{i=1}^k \sum_{j=1}^k \alpha_i \alpha_j y_i y_j x_i x_j$ δεδομένου ότι

$\alpha_i \geq 0$ και $\sum_{i=1}^k \alpha_i y_i = 0$. Εν συνεχεία, με την παραδοχή ότι $b=0$, το υπερεπίπεδο διαχωρισμού

περνάει από την αρχή των αξόνων, ο περιορισμός $\sum_{i=1}^k \alpha_i y_i = 0$ δεν εμφανίζεται.

Υπάρχει λοιπόν ένας πολλαπλασιαστής Lagrange για κάθε στιγμιότυπο για τους οποίους ισχύει $a_i > 0$ για κάθε στιγμιότυπο που αντιστοιχεί σε ένα διάνυσμα υποστήριξης και $a_i = 0$ για όλα τα άλλα. Άρα αν sv το σύνολο όλων διανυσμάτων υποστήριξης (δηλαδή τα σημεία για τα οποία ισχύει $a_i > 0$), τότε $y_i(x_i w + b) = 1 \forall i \in sv$. Άρα $w = \sum_{i=1}^k a_i y_i x_i = \sum_{i \in sv} a_i y_i x_i$ και κατά συνέπεια $y_i \left(\sum_{j \in sv} a_j y_j x_j + b \right) = 1 \Rightarrow y_i \sum_{j \in sv} a_j y_j x_j = 1 - y_i b$. Έτσι το $\|w\|^2$, που ζητείται να ελαχιστοποιηθεί γίνεται $\|w\|^2 = w^T w = \sum_{i \in sv} a_i y_i x_i \sum_{j \in sv} a_j y_j x_j = \sum_{i \in sv} a_i y_i \sum_{j \in sv} a_j y_j x_i^T x_j$ και τέλος, αφού $\|w\|^2 = \sum_{i \in sv} a_i$, η απόσταση των υπερεπιπέδων από το κανονικό είναι $\|w\| = \left(\sum_{i \in sv} a_i \right)^{\frac{1}{2}}$.

Σε αυτό το σημείο πρέπει να τονιστεί ότι οι συνθήκες Karach Kuhn Tucker (ή για συντομία KKT) παίζουν κυρίαρχο ρόλο στη βελτιστοποίηση με περιορισμούς. Έτσι, η επίλυση του προβλήματος των μηχανών διανυσμάτων υποστήριξης είναι ισοδύναμη με την εύρεση λύση στο πρόβλημα KKT. Ουσιαστικά καταλήξαμε σε ένα πρόβλημα κυρτού τετραγωνικού προγραμματισμού.

Στην περίπτωση λοιπόν των γραμμικά διαχωρισμένων δεδομένων, μόλις το κατάλληλο υπέρ-πλάνο διαχωρισμού βρεθεί, τα σημεία δεδομένων που βρίσκονται στο υπέρ-πλάνο και είναι γνωστά ως διανύσματα υποστήριξης τελικά χρησιμοποιούνται. Η λύση παρουσιάζεται ως γραμμικός συνδυασμός μόνο αυτών των σημείων. Τα υπόλοιπα σημεία μπορούν να αγνοηθούν. Έτσι η πολυπλοκότητα του σχεδιασμού των SVMs δεν επηρεάζεται από τον αριθμό των δεδομένων εισόδου (ο αριθμός των διανυσμάτων που επιλέγεται από τον SVM αλγόριθμο είναι συνήθως μικρός). Γι' αυτό το λόγο, οι SVMs ενδείκνυνται σε ζητήματα μάθησης όπου ο αριθμός των μεταβλητών είναι μεγάλος σε σχέση με τον αριθμό των δεδομένων εισόδου.

Όμως, παρόλο που το μέγιστο περιθώριο επιτρέπει στον αλγόριθμο SVM να επιλέγει ανάμεσα σε πολλαπλά υποψήφια υπέρ-πλάνα, σε πολλές περιπτώσεις δεδομένων ο SVM μπορεί να μην είναι ικανός να βρει το κατάλληλο υπέρ-πλάνο διαχωρισμού γιατί τα δεδομένα περιέχουν θόρυβο. Στην κάπως εξεζητημένη περίπτωση της Εικόνας 39 (αριστερά), παρατηρούμε ότι όλα τα διανύσματα εκπαίδευσης αποτελούν διανύσματα υποστήριξης, με αποτέλεσμα το υπερεπίπεδο μέγιστου περιθωρίου να είναι μια οριζόντια γραμμή στο μέσο του χώρου που ορίζουν οι δύο κλάσεις. Αν ωστόσο υπεισέλθει θόρυβος στα δεδομένα εκπαίδευσης (Εικόνα 39, δεξιά), το επαγόμενο μοντέλο ενδέχεται να αλλάξει δραματικά.

Εικόνα 39. Υπερταίριασμα

Το πρόβλημα που παρουσιάστηκε λόγω της παρουσίας θορύβου μπορεί να διευθετηθεί χρησιμοποιώντας ένα περιθώριο που αποδέχεται μερικές μη ορθές ταξινομήσεις των περιπτώσεων εκπαίδευσης (Veropoulos και άλλοι, 1999). Βλέπε για παράδειγμα Εικόνα 40.

Εικόνα 40. (α) Στην περίπτωση αυτή, ένα σημείο πέφτει στην περιοχή του διαχωρισμού, αλλά στην σωστή πλευρά του επιπέδου απόφασης (β) Στην περίπτωση αυτή, το σημείο πέφτει στην λάθος πλευρά του επιπέδου απόφασης

Οι ανισότητες που αποτελούν τους αρχικούς περιορισμούς για την διαχωρίσιμη περίπτωση θα εφαρμοστούν και στην συγκεκριμένη περίπτωση με μια μικρή τροποποίηση. Οι περιορισμοί γίνονται

$$\begin{aligned} \mathbf{x}_i \mathbf{w} + b &\geq 1 - \xi_i \text{ για } y_i = 1 \\ \mathbf{x}_i \mathbf{w} + b &\geq -1 + \xi_i \text{ για } y_i = -1 \\ \xi_i &\geq 0 \end{aligned}$$

Για να καταχωρηθεί ένα στιγμιότυπο λάθος θα πρέπει τα ξ_i να ξεπερνούν τη μονάδα και κατά συνέπεια η έκφραση $\sum_{i=1}^k \xi_i$ αποτελεί άνω φράγμα για το σφάλμα της μεθόδου. Για να αποδοθεί μια τιμή για το κόστος του σφάλματος ζητείται να ελαχιστοποιηθεί η τιμή $\frac{\|\mathbf{w}\|^2}{2} + C \left(\sum_{i=1}^k \xi_i \right)^k$ αντί της $\frac{\|\mathbf{w}\|^2}{2}$ που ελαχιστοποιήθηκε προηγουμένως. Η C είναι μια παράμετρος που επιλέγεται ξεχωριστά για κάθε περίπτωση. Όσο μεγαλύτερη είναι η τιμή της, τόσο μεγαλύτερο είναι το σφάλμα. Με αυτήν την μορφή το πρόβλημα είναι ένα πρόβλημα κυρτού προγραμματισμού.

Οι συνθήκες KKT γίνονται $\frac{\partial L_p}{\partial \mathbf{w}_v} = \mathbf{w}_i - \sum_{j=1}^k a_j y_j \mathbf{x}_{ji} = 0$ με $i \in \{1, 2, \dots, n\}$, $\frac{\partial L_p}{\partial b} = -\sum_{i=1}^k a_i y_i = 0$,

$$\frac{\partial L_p}{\partial \xi_i} = C - a_i - \mu_i = 0, \quad y_i (\mathbf{x}_i \mathbf{w} + b) - 1 + \xi_i \geq 0 \quad \text{με } i \in \{1, 2, \dots, n\}, \quad \xi_i \geq 0, \quad a_i \geq 0, \quad \mu_i \geq 0,$$

$a_i (y_i (\mathbf{x}_i \mathbf{w} + b) - 1 + \xi_i) = 0$ με $\forall i \in \{1, 2, \dots, k\}$ και $\mu_i \xi_i = 0$. Από τις δύο τελευταίες σχέσεις

μπορούν να προσδιοριστούν το b . Επίσης, η εξίσωση $\frac{\partial L_p}{\partial \xi_i} = C - a_i - \mu_i = 0$ σε συνδυασμό με

την $\mu_i \xi_i = 0$ δίνουν ότι $\xi_i = 0$ αν $0 \leq a_i \leq C$.

Γενικά, η παράμετρος ξ_i επιτρέπει στο αντίστοιχο διάνυσμα εκπαίδευσης να βρεθεί στην περιοχή του περιθωρίου εφόσον είναι μεγαλύτερη του μηδενός, ενώ η παράμετρος C , η οποία

πρέπει να προσδιοριστεί από το χρήστη, εκφράζει την αυστηρότητα που αναμένεται να επιδείξει ο αλγόριθμος στην ανοχή στιγμιότυπων στο περιθώριο, κατά την εύρεση του βέλτιστου υπερεπιπέδου.

Επανερχόμενοι στην Εικόνα 39, η προσθήκη τεσσάρων ακόμα διανυσμάτων εκπαίδευσης στο μέσο του χώρου δεν επιφέρει καμία ουσιαστική αλλαγή στη συμπεριφορά του αλγορίθμου, αν επιλεγεί μικρή τιμή για το C , καθώς τον κάνει ανεκτικό σε μικρό αριθμό σφαλμάτων εκπαίδευσης. Μεγάλη τιμή του C ωστόσο θα ωθήσει τον αλγόριθμο να ταξινομήσει όσο το δυνατόν περισσότερα διανύσματα εκπαίδευσης σωστά, οδηγώντας στο διαγώνιο διαχωρισμό του επιπέδου, όπως φαίνεται στο σχήμα δεξιά, που απέχει πολύ από τον επιθυμητό (υπερταίριασμα).

Στο σημείο αυτό θα πρέπει να τονίσουμε ότι το προαναφερθέν ενδεχόμενο θεωρείται σχετικά σπάνιο να παρατηρηθεί σε έναν ταξινομητή SVM, αφού τα διανύσματα υποστήριξης αποτελούν ένα πολύ μικρό ποσοστό των διανυσμάτων εκπαίδευσης, όχι όμως και εντελώς απίθανο.

Είναι όμως γεγονός ότι τα περισσότερα πραγματικά προβλήματα περιλαμβάνουν μη-διαχωρίσιμα δεδομένα για τα οποία δεν υπάρχουν καθόλου υπέρ-πλάνα τα οποία να διαχωρίζουν επιτυχημένα τις θετικές από τις αρνητικές περιπτώσεις. Μια λύση στο πρόβλημα αυτό είναι να καθορίσουμε τα δεδομένα σ' ένα διάστημα υψηλότερων διαστάσεων και να ορίσουμε ένα υπέρ-πλάνο διαχωρισμού εκεί.

3.6.2 Μη γραμμικές μηχανές υποστήριξης διανυσμάτων

Μέχρι τώρα μελετήθηκαν περιπτώσεις όπου η συνάρτηση απόφασης είναι γραμμική. Παρατηρείται ότι το γινόμενο είναι η μοναδική έκφραση στην οποία εμφανίζονται τα στιγμιότυπα του εκάστοτε προβλήματος μηχανικής μάθησης στις εξισώσεις

$$L_d(a) = \sum_{i=1}^k a_i - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k a_i a_j y_i y_j x_i x_j, \quad 0 \leq a_i \leq C \quad \text{και} \quad \sum_{i=1}^k a_i y_i = 0.$$

Ας υποθεθεί ότι τα δεδομένα απεικονίζονται σε ένα Ευκλείδειο χώρο H με τη χρήση μιας απεικόνισης $\Phi: \mathcal{R}^n \rightarrow H$. Άρα το γινόμενο $\mathbf{x}_i \cdot \mathbf{x}_j$ απεικονίζεται στο γινόμενο $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$.

Μάλιστα αν υπάρχει ένας πυρήνας K τέτοιος ώστε $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, τότε θα χρησιμοποιούνταν μόνο αυτός ο πυρήνας στον αλγόριθμο και δεν θα υπήρχε καν η ανάγκη να είναι γνωστή η απεικόνιση Φ . Αν ο χώρος H είναι απειροδιάστατος, τότε είναι δύσκολο να γίνουν οι πράξεις με την απεικόνιση Φ , ενώ είναι πολύ πιο εύκολο αυτές να γίνουν μόνο με την χρήση αυτού του πυρήνα K στον αλγόριθμο. Έτσι κατασκευάζεται μια μηχανή διανυσμάτων υποστήριξης που ανήκει σε ένα απειροδιάστατο χώρο. Έτσι επίσης ισχύουν όλες οι παραδοχές που έχουν γίνει στις προηγούμενες παραγράφους εφόσον και πάλι γίνεται γραμμικός διαχωρισμός των στιγμιότυπων, μόνο που τώρα η διεργασία αυτή γίνεται σε ένα διαφορετικό χώρο.

Έστω λοιπόν ότι ο χώρος στιγμιότυπων είναι ο $\mathcal{R}$ και ένας πυρήνας K τέτοιος ώστε $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^2$. Στην συγκεκριμένη περίπτωση είναι εύκολο να βρεθεί ένας χώρος H και μια απεικόνιση Φ από τον $\mathcal{R}^2$ στον H , τέτοια ώστε $(\mathbf{x}_i \cdot \mathbf{x}_j)^2 = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Για παράδειγμα μια απεικόνιση είναι η $\Phi(\mathbf{x}) = [x_1^2, \sqrt{2}x_1x_2, x_2^2]^T$. Δεν είναι εύκολη όμως πάντα η εύρεση ενός πυρήνα και της αντίστοιχης απεικόνισης Φ . Απάντηση σε αυτό δίνει η συνθήκη του Mercer.

Συνθήκη του Mercer για την επιλογή πυρήνα: Βασική προϋπόθεση για την επιλογή του πυρήνα είναι να υπάρχει ζεύγος $\{H, \Phi\}$. Σύμφωνα με την συνθήκη του Mercer,

υπάρχει μια απεικόνιση Φ και επέκταση $K(x, y) = \sum_{i=1}^k \Phi(x)_i \cdot \Phi(y)_i$ αν και μόνον αν για ένα $g(x)$, τέτοιο ώστε $\int (g(x))^2 dx$ πεπερασμένο, ισχύει $\int K(x, y) g(x) g(y) dx dy \geq 0$. Η επαλήθευση της συνθήκης Mercer δεν είναι πάντα απλή διαδικασία.

Ωστόσο αποδεικνύεται ότι κάθε $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^p$ με p θετικό ικανοποιεί την συνθήκη Mercer. Για να ισχύει η συνθήκη πρέπει $\int \sum_{i=1}^n (x, y)^p g(x) g(y) dx dy \geq 0$. Ο όρος $\left(\sum_{i=1}^n (x, y) \right)^p$ συνεισφέρει έναν όρο της μορφής $\frac{p!}{r_1! r_2! \dots r_n! (p - r_1 - r_2 - \dots - r_n)} \int x_1^{r_1} x_2^{r_2} \dots x_n^{r_n} y_1^{r_1} y_2^{r_2} \dots y_n^{r_n} g(x) g(y) dx dy$ στην αριστερή πλευρά της ανισότητας $\int K(x, y) g(x) g(y) dx dy \geq 0$. Μια απλή συνέπεια των παραπάνω είναι ότι οποιοσδήποτε πυρήνας μπορεί να εκφραστεί ως $K(x, y) = \sum_{p=0}^{+\infty} C_p (\mathbf{x} \cdot \mathbf{y})^p$, όπου C_p είναι θετικοί πραγματικοί συντελεστές η σειρά συγκλίνει ομοιόμορφα και ο K ικανοποιεί τη συνθήκη του Mercer.

Παρακάτω παραθέτουμε ένα γραμμικό και ένα μη γραμμικό πρόβλημα ώστε να παρουσιαστεί ο τρόπος χρήσης των μηχανών διανυσμάτων υποστήριξης. Έστω λοιπόν το γραμμικό πρόβλημα διαχωρισμού του λογικού H' (OR).

Είσοδος, x	Επιθυμητή έξοδος, y
(-1,-1)	-1
(-1,+1)	-1
(+1,-1)	-1
(+1,+1)	1

Πίνακας 27. Πρόβλημα OR

Θεωρούμε: $K(x, x_i) = (1 + x^T x_i)$ όπου $x = [x_1, x_2]^T$, $x_i = [x_{i1}, x_{i2}]^T$.

$$\text{Έτσι } K(x, x_i) = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

$$L(a) = \sum_{i=2}^N a_i - \frac{1}{2} \sum_{i=2}^N \sum_{j=2}^N a_i a_j y_i y_j K(x_i, x_j) = a_2 + a_3 + a_4 - \frac{1}{2} (3a_2^2 - 2a_2 a_3 - 2a_2 a_4 + 3a_3^2 - 2a_3 a_4 + 3a_4^2)$$

Lagrange multipliers a_i :

$$\begin{cases} 3a_2 - a_3 - a_4 = 1 \\ -a_2 + 3a_3 - a_4 = 1 \\ -a_2 - a_3 + 3a_4 = 1 \end{cases}$$

Η λύση μας δίνει: $a_2 = a_3 = a_4 = 1$.

$$K(x, x_i) = \varphi(x) \varphi(x_i)$$

$$\varphi(x_i) = [1, x_{i1}, x_{i2}]^T$$

$$w_0 = \sum_{i=1}^N a_i y_i \varphi(x_i)$$

$$= -\varphi(x_2) - \varphi(x_3) + \varphi(x_4)$$

$$= \begin{bmatrix} -\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

$w_0^T \varphi(x) = 0$, που δίνει,

$$\begin{bmatrix} -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = 0$$

το οποίο μας δίνει: $x_1 + x_2 - 1 = 0$ (βλέπε Εικόνα 41).

Εικόνα 41. Πρόβλημα OR

Παρακάτω παραθέτουμε το μη γραμμικό πρόβλημα διαχωρισμού του λογικού XOR.

Είσοδος, x	Επιθυμητή έξοδος, y
$(-1, -1)$	-1
$(-1, +1)$	+1
$(+1, -1)$	+1
$(+1, +1)$	-1

Πίνακας 28. Πρόβλημα XOR

Έστω $K(x, x_i) = (1 + x^T x_i)^2$

όπου $x = [x_1, x_2]^T$, $x_i = [x_{i1}, x_{i2}]^T$.

Τότε παίρνουμε

$$K(x, x_i) = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

$$\begin{aligned} L(a) &= \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j K(x_i, x_j) \\ &= a_1 + a_2 + a_3 + a_4 - \frac{1}{2} (9a_1^2 - 2a_1a_2 - 2a_1a_3 + 2a_1a_4 + 9a_2^2 + 2a_2a_3 - 2a_2a_4 + 9a_3^2 - 2a_3a_4 + 9a_4^2) \end{aligned}$$

Lagrange multipliers a_i :

$$\begin{cases} 9a_1 - a_2 - a_3 + a_4 = 1 \\ -a_1 + 9a_2 + a_3 - a_4 = 1 \\ -a_1 + a_2 + 9a_3 - a_4 = 1 \\ a_1 - a_2 - a_3 + 9a_4 = 1 \end{cases}$$

Δίνει την λύση: $a_1=a_2=a_3=a_4=1/8$, και $Q(a) = 1/4$.

Αφού $K(x, x_i) = \varphi(x)\varphi(x_i)$ παίρνουμε

$$\varphi(x_i) = [1, x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}]^T$$

$$\begin{aligned} w_0 &= \sum_{i=1}^N a_i y_i \varphi(x_i) = \frac{1}{8} [-\varphi(x_1) + \varphi(x_2) + \varphi(x_3) - \varphi(x_4)] \\ &= \frac{1}{8} \left[- \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ -\sqrt{2} \\ -\sqrt{2} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -\sqrt{2} \\ 1 \\ -\sqrt{2} \\ \sqrt{2} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -\sqrt{2} \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ \sqrt{2} \end{bmatrix} \right] = \begin{bmatrix} 0 \\ 0 \\ -1/\sqrt{2} \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

$$w_0^T \varphi(x) = 0, \text{ που είναι,}$$

$$\begin{bmatrix} 0, 0, \frac{-1}{\sqrt{2}}, 0, 0, 0 \end{bmatrix} \begin{bmatrix} 1 \\ x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \end{bmatrix} = 0$$

το οποίο μας δίνει: $-x_1x_2=0$ (το βέλτιστο υπερ-πλάνο σύμφωνα με το SVM – βλέπε Εικόνα 42).

Εικόνα 42. Πρόβλημα XOR

Ο Genton (2001) περιέγραψε αρκετές κατηγορίες πυρήνων, ωστόσο δεν απάντησε στην ερώτηση για το ποια κατηγορία είναι καταλληλότερη για κάθε πρόβλημα μάθησης. Είναι κοινή πρακτική να υπολογίζουμε ένα εύρος πιθανών πυρήνων και να χρησιμοποιούμε την ακρίβεια ταξινόμησης στα δοκιμαστικά δεδομένα ώστε να βρούμε τον καλύτερο. Γι' αυτό το λόγο ένας περιορισμός των SVM είναι η αργή εκπαίδευση. Η επιλογή του κατάλληλου πυρήνα μπορεί να θεωρηθεί ως ένα παρόμοιο πρόβλημα όπως η επιλογή του αριθμού των κρυμμένων νευρώνων σ' ένα νευρωνικό δίκτυο.

Οι πιο δημοφιλείς πυρήνες είναι οι πολυωνυμικοί: $K(x, y) = (x \cdot y + 1)^p$ Στο επόμενο κεφάλαιο παραθέτουμε μια σύγκριση για διάφορες τιμές του p .

3.6.3 Σύγκριση πολυωνυμικών πυρήνων

Ο SMO είναι ένας πρόσφατος αλγόριθμος για τα SVMs, που προτάθηκε από τον Platt (1999). Ο Platt χωρίζει το μεγάλο πρόβλημα KKT σε μικρά υπό-προβλήματα κάτω από ορισμένες συνθήκες και λύνει κάθε υπό-πρόβλημα χωριστά.

Χρησιμοποιήσαμε τον αλγόριθμο SMO για τις συγκρίσεις μας με $p=1$, $p=2$, $p=3$ στον πολυωνυμικό πυρήνα που περιγράψαμε στην προηγούμενη ενότητα. Τα πειράματα (βλέπε Πίνακα 29) έδειξαν ότι οι πολυωνυμικοί πυρήνες δεύτερης τάξης αποδίδουν συνήθως καλύτερα.

Πρέπει να σημειωθεί ότι οι Keerthi και Gilbert (2002) πρότειναν δύο τροποποιημένες εκδόσεις του SMO που είναι αξιοσημείωτα γρηγορότερες από το αυθεντικό SMO στις περισσότερες περιπτώσεις. Επίσης, ο Duan και άλλοι (2003) μελέτησαν πως μπορούν να βελτιστοποιηθούν και οι άλλοι παράμετροι πλην των πυρήνων του αλγόριθμου SMO για την βέλτιστη απόδοση σε ένα σύνολο δεδομένων.

Δεδομένα	Πολυωνυμικός πυρήνας 1 ^{ης} τάξης	Πολυωνυμικός πυρήνας 2 ^{ης} τάξης	Πολυωνυμικός πυρήνας 3 ^{ης} τάξης
autos	71.34	76.84	76.64
badges	100.00	100.00	100.00
balance-scale	87.57	91.38	90.64
breast-c	69.52	67.33	66.94
breast-w	96.75	96.37	95.71
colic	82.66	76.74	76.87
credit-a	84.88	84.88	81.71
credit-g	75.09	69.28	70.90
diabetes	76.80	77.32	76.74
haberman	73.40	73.27	73.23
heart-c	83.86	81.27	77.62
heart-h	82.74	82.52	78.41
heart-statlog	83.89	81.22	79.70
hepatitis	85.77	81.63	82.46
ionosphere	88.07	90.69	88.09
iris	96.27	95.53	92.87
labor	92.97	92.50	92.10
lymphography	86.48	83.45	82.79
monk1	79.58	100.00	99.85
monk2	58.70	70.33	66.30
monk3	93.45	86.29	87.61
relation	77.60	79.05	79.05
sonar	76.60	83.80	84.82
students	86.72	78.97	77.85
vehicle	74.08	80.26	83.25
vote	95.77	94.50	95.01
wine	98.76	97.97	96.96
zoo	96.05	97.05	94.60

Πίνακας 29. Σύγκριση πολυωνυμικών πυρήνων

Εκτός από τους πολυωνυμικούς μερικοί άλλοι δημοφιλείς πυρήνες είναι οι ακόλουθοι :

- $K(x, y) = e^{-\|x-y\|^2 / 2\sigma^2}$
- $K(x, y) = \tanh(\kappa x \cdot y - \delta)^p$

Γενικά, τα SVM φτάνουν απαραίτητα στο γενικό ελάχιστο αποφεύγοντας τα τοπικά ελάχιστα, σε αντίθεση με τα νευρωνικά δίκτυα. Ωστόσο, οι SVM μέθοδοι είναι δυαδικές και στην περίπτωση των προβλημάτων με περισσότερες από δύο κλάσεις, πρέπει να μειωθεί το πρόβλημα σε μια ομάδα πολλαπλών δυαδικών προβλημάτων ταξινόμησης¹¹.

Ένα άλλο πλεονέκτημα των SVMs είναι η ανεκτικότητα που παρουσιάζουν όσον αφορά στο πλήθος των στιγμιοτύπων εκπαίδευσης, ιδιαίτερα όταν υπάρχει διαφορά στο πλήθος

¹¹ Περισσότερα για το πως ένας δυαδικός ταξινομητής μπορεί να εφαρμοστεί σε πολυ-κατηγορικά δεδομένα στην ενότητα 4.3

παραδειγμάτων ανά κλάση, καθώς τα SVMs δεν επιδιώκουν να ελαχιστοποιήσουν το σφάλμα των δεδομένων εκπαίδευσης, αλλά να τα διαχωρίσουν αποτελεσματικά σε ένα χώρο μεγάλης διάστασης. Όσον αφορά τέλος στους χρόνους εκπαίδευσης των αλγορίθμων αυτής της κατηγορίας, αυτοί αποδεικνύονται κάπως αυξημένοι, ιδιαίτερα όταν η διάσταση του χώρου είναι μεγάλη, ή όταν η συνάρτηση διάκρισης δεν είναι γραμμική.

3.7 Θεωρία υπολογιστικής μάθησης

Κάτω από ποιες συνθήκες μπορεί ένας αλγόριθμος να δώσει ικανοποιητικά αποτελέσματα; Ποια είναι τα χαρακτηριστικά που περιγράφουν ένα πρόβλημα μηχανικής μάθησης και από ποιες παραμέτρους εξαρτάται ένα πρόβλημα;

Όσον αφορά το πλήθος των στιγμιότυπων που είναι απαραίτητα αυτό εξαρτάται από δυο κυρίως παράγοντες. Ένας απ' τους βασικότερους παράγοντες είναι το μέγεθος και η πολυπλοκότητα του χώρου υποθέσεων. Όσο περισσότερο πολύπλοκος και μεγάλος είναι ο χώρος των υποθέσεων, τόσο περιπλοκότερες υποθέσεις μπορεί να εξάγει ο αλγόριθμος. Έτσι τα στιγμιότυπα που χρησιμοποιούνται πρέπει να είναι μεγάλου πλήθους ώστε να μπορέσει ο αλγόριθμος να μάθει τις τιμές των παραμέτρων της συνάρτησης που καλείται να βρει.

Ένας δεύτερος εξίσου σημαντικός παράγοντας είναι η επιθυμητή ακρίβεια του ταξινομητή. Είναι προφανές ότι ο αλγόριθμος εκπαίδευσης δεν αρκείται σε ένα πλήθος στιγμιότυπων για να δώσει μια εκτίμηση για τις τιμές των παραμέτρων, αλλά προκειμένου να έχει και μια ικανοποιητική ακρίβεια θα πρέπει το πλήθος των στιγμιότυπων του να είναι τέτοιο που θα του επιτρέψει να δώσει μια εκτίμηση με την καλύτερη δυνατή ακρίβεια. Έτσι, όσο μεγαλύτερη ακρίβεια ζητείται, τόσο αυξάνουν οι ανάγκες του αλγορίθμου για στιγμιότυπα.

Έτσι λοιπόν κατά τη θεωρητική μελέτη της μηχανικής μάθησης γίνεται προσπάθεια να καλυφθούν ερωτήματα που αφορούν το πλήθος των απαραίτητων στιγμιότυπων, το υπολογιστικό κόστος που θα έχει η εκπαίδευση και φυσικά το φράγμα του σφάλματος. Ένα μοντέλο για τη μελέτη των αλγορίθμων είναι το «Πιθανόν Περίπου Σωστό» (Probably Approximately Correct ή για συντομία PAC) (Vapnik, 1998). Σκοπός της εκπαίδευσης δεν είναι η κατασκευή ενός ταξινομητή που θα ταξινομεί σωστά κάθε στιγμιότυπο, καθώς κάτι τέτοιο θα απαιτούσε το σύνολο εκπαίδευσης να περιέχει όλα τα δυνατά στιγμιότυπα. Όποτε ο αλγόριθμος αρκείται να κατασκευάζει έναν ταξινομητή που με μια πιθανότητα δ θα δίνει σωστή τιμή σε ένα στιγμιότυπο.

Έστω μια κλάση υποθέσεων C που ορίζεται σε ένα σύνολο στιγμιότυπων X μεγέθους n και ένα αλγόριθμο εκπαίδευσης L που χρησιμοποιεί το χώρο υποθέσεων H . Η κλάση C λέγεται εκπαιδύσιμη κατά PAC από τον L χρησιμοποιώντας το H αν για κάθε $c \in C$, υπάρχει ε

τέτοιο ώστε $0 < \varepsilon < \frac{1}{2}$ και δ τέτοιο ώστε $0 < \delta < \frac{1}{2}$ ο αλγόριθμος L εξάγει με πιθανότητα

$(1 - \delta)$ μια υπόθεση $h \in H$ τέτοια ώστε $error_D(h) \leq \varepsilon$ σε χρόνο που είναι πολυωνυμική συνάρτηση των $\frac{1}{\varepsilon}, \frac{1}{\delta}, n$ και $size(c)$.

Είναι εμφανές ότι η εκπαιδευσιμότητα κατά PAC εξαρτάται από το πλήθος των στιγμιότυπων που χρειάζεται ένας αλγόριθμος για να εκπαιδευτεί. Το πλήθος των απαραίτητων στιγμιότυπων για την εκπαίδευση αποτελεί το βασικότερο παράγοντα για την ικανότητα και την ακρίβεια του αλγορίθμου. Στην πράξη αυτό μεταφράζεται με περιορισμό της ακρίβειας του ταξινομητή όταν δεν υπάρχουν αρκετά στιγμιότυπα διαθέσιμα.

Σε ένα χώρο υποθέσεων ποιες υποθέσεις είναι συνεπής; Οι υποθέσεις αυτές έχουν ιδιαίτερη σημασία για τη μηχανική μάθηση και ομαδοποιούνται σαν ένα σύνολο που καλείται Χώρος Συνέπειας (Version Space). Τυπικά Χώρος Συνέπειας του χώρου υποθέσεων H και των στιγμιότυπων εκπαίδευσης D ονομάζεται ένα υποσύνολο του H που είναι συνεπές με τα

στιγμιότυπα του D: $VS_{H,D} \equiv \{h \in H \mid \text{Consistent}(h, D)\}$ ή
 $VS_{H,D} = \{h \in H \mid (\forall \langle x, c(x) \rangle \in D)(h(x) = c(x))\}.$

Όπως προαναφέρθηκε, είναι αδύνατον στην πράξη να κατασκευαστεί ένας ταξινομητής που θα έχει μηδενικό σφάλμα. Γι' αυτό και ορίζεται ένα φράγμα ε που αποτελεί ένα ανεκτό σφάλμα για τις απαιτήσεις του προβλήματος. Αυτό το φράγμα ορίζει μια καινούρια έννοια. Ο Χώρος Συνέπειας $VS_{H,D}$ καλείται ε-εξαντλημένος δεδομένων των c και D , αν κάθε υπόθεση h στο $VS_{H,D}$ έχει σφάλμα ε . Ο ορισμός αυτός περιγράφεται από τη σχέση $(\forall h \in VS_{H,D}) \text{error}_D(h) < \varepsilon$.

ΘΕΩΡΗΜΑ

Αν ένας χώρος υποθέσεων H είναι πεπερασμένος και D είναι μια ακολουθία από $m \geq 1$ ανεξάρτητα τυχαία στιγμιότυπα κάποιας συνάρτησης-στόχου c , τότε για κάθε $0 \leq \varepsilon \leq 1$, η πιθανότητα ο Version Space να μην είναι ε-εξαντλημένος (δεδομένου του c) είναι μικρότερη ή ίση από την τιμή $|H|e^{-\varepsilon m}$.

Απόδειξη: Έστω $h_1, h_2, \dots, h_k$ όλες οι υποθέσεις του H που έχουν σφάλμα μεγαλύτερο από ε δεδομένου c . Ο Version Space δεν είναι ε-εξαντλημένος αν και μόνο αν τουλάχιστον μια από τις υποθέσεις είναι συνεπής με όλα τα m ανεξάρτητα τυχαία στιγμιότυπα του συνόλου εκπαίδευσης. Η πιθανότητα μια υπόθεση, που έχει πραγματικό σφάλμα μεγαλύτερο από ε , να είναι συνεπής με ένα τυχαίο στιγμιότυπο είναι το πολύ $(1-\varepsilon)$. Γι' αυτό το λόγο η πιθανότητα ότι αυτή η υπόθεση θα είναι συνεπής με m ανεξάρτητα στιγμιότυπα είναι το πολύ $(1-\varepsilon)^m$. Δεδομένου ότι υπάρχουν k υποθέσεις με σφάλμα μεγαλύτερο από ε , η πιθανότητα τουλάχιστον μία από αυτές να είναι συνεπής με όλα τα στιγμιότυπα είναι το πολύ $k(1-\varepsilon)^m$. Προφανώς $k \leq |H|$ και κατά συνέπεια $|H|(1-\varepsilon)^m$. Τέλος είναι γνωστή η σχέση $(1-\varepsilon) \leq e^{-\varepsilon}$ απ' την οποία εξάγεται το τελικό συμπέρασμα $k(1-\varepsilon)^m \leq |H|(1-\varepsilon)^m \leq |H|e^{-\varepsilon m}$.

Το φράγμα, της πιθανότητας ο Χώρος Συνέπειας να μην είναι ε-εξαντλημένος, που μόλις βρέθηκε αποτελεί βάση για να βρεθεί το πλήθος των στιγμιότυπων που είναι απαραίτητα για να εκπαιδευτεί με ικανοποιητική ακρίβεια ένας ταξινομητής. Στην ουσία ζητείται η πιθανότητα μιας αποτυχημένης ταξινόμησης να είναι μικρότερη από δ . Δηλαδή $|H|e^{-\varepsilon m} \leq \delta$,

η οποία αν λυθεί ως προς m δίνει τη σχέση $m \geq \frac{1}{\varepsilon} \left(\ln |H| + \ln \left(\frac{1}{\delta} \right) \right)$. Η σχέση αυτή αποτελεί

ένα ικανοποιητικό φράγμα αν και σε πολλές περιπτώσεις μπορούν να δοθούν πολύ πιο στενά φράγματα.

Έστω ένας συνεπής αλγόριθμος L που χρησιμοποιεί το χώρο υποθέσεων H . Ζητείται το πλήθος των στιγμιότυπων που απαιτούνται για να κατασκευαστεί ένας ταξινομητής ο οποίος θα έχει μέγιστο σφάλμα ε με πιθανότητα $(1-\delta)$. Από τον τύπο $m \geq \frac{1}{\varepsilon} \left(\ln |H| + \ln \left(\frac{1}{\delta} \right) \right)$ μένει

να υπολογιστεί ο πληθάριθμος του χώρου υποθέσεων. Έστω ότι ο χώρος υποθέσεων αποτελείται από συζεύξεις n λογικών μεταβλητών. Ο πληθάριθμος του χώρου είναι λοιπόν 3^n , καθώς μια μεταβλητή είτε περιέχεται αυτούσια, είτε η άρνησή της είτε δεν υπάρχει μέσα στην υπόθεση που έχει επιλεγεί σαν ταξινομητής. Άρα ένα φράγμα για το πλήθος των

στιγμιότυπων που απαιτούνται σαν σύνολο εκπαίδευσης με την ακρίβεια είναι $m \geq \frac{1}{\varepsilon} \left(n \ln 3 + \ln \left(\frac{1}{\delta} \right) \right)$.

Υπάρχουν όμως κλάσεις με λογικές συναρτήσεις οι οποίες δεν μπορούν να βρεθούν από κάποιο αλγόριθμο σε πολυωνυμικό χρόνο. Έστω ότι υπάρχει ένας χώρος υποθέσεων που αποτελείται από υποθέσεις που περιγράφουν το σύνολο στιγμιότυπων X . Το σύνολο του χώρου υποθέσεων ταυτίζεται με το δυναμοσύνολο του X . Άρα ο πληθάριθμός του είναι $2^{|X|}$ και αν τα στιγμιότυπα του X αποτελούνται από n λογικούς όρους. Ο πληθάριθμος του X είναι 2^n και κατά συνέπεια ο χώρος υποθέσεων έχει πληθάριθμο 2^{2^n} . Άρα φράγμα όπως υπολογίστηκε και στις προηγούμενες περιπτώσεις υπολογίζεται και σε αυτήν την περίπτωση

από τον τύπο $m \geq \frac{1}{\varepsilon} \left(\ln |H| + \ln \left(\frac{1}{\delta} \right) \right)$ που για το συγκεκριμένο σύνολο εκπαίδευσης δίνει την

έκφραση $m \geq \frac{1}{\varepsilon} \left(2^n \ln 2 + \ln \left(\frac{1}{\delta} \right) \right)$. Όμως η έκφραση του φράγματος έχει εκθετική σχέση με

το n και κατά συνέπεια δεν υπάρχει γραμμική σχέση με το μέγεθος του συνόλου στιγμιότυπων, όπως απαιτεί ο ορισμός του PAC-εκπαιδευσιμου προβλήματος.

3.7.1 Διάσταση VC

Η εφαρμογή του PAC έχει δύο μειονεκτήματα. Κατά πρώτον αν ο χώρος υποθέσεων είναι πολύ μεγάλος, τα φράγματα που δίνονται αδυνατούν να περιορίσουν το πλήθος των στιγμιότυπων και συνεπώς υπερεκτιμούν το απαραίτητο μέγεθος του συνόλου εκπαίδευσης. Επιπλέον ο τύπος αυτός δεν μπορεί να χρησιμοποιηθεί όταν ο χώρος υποθέσεων δεν είναι πεπερασμένος. Στην περίπτωση αυτή γίνεται χρήση ενός διαφορετικού μέτρου της πολυπλοκότητας του χώρου υποθέσεων που καλείται διάσταση Vapnik-Chervonenkis ή για συντομία διάσταση VC (Vapnik, 1998). Η διάσταση VC δεν δίνει μόνο καλύτερα φράγματα για το πλήθος των απαραίτητων στιγμιότυπων, αλλά μπορεί να δώσει φράγματα και για προβλήματα με χώρους υποθέσεων άπειρου πλήθους (Mossel & Umans, 2001).

Η διάσταση VC μετράει την πολυπλοκότητα του χώρου υποθέσεων όχι σύμφωνα με τον πληθάριθμο, αλλά με το πλήθος των διαφορετικών στιγμιότυπων που μπορούν να διαχωριστούν πλήρως με τη χρήση του χώρου υποθέσεων.

Ένα σύνολο στιγμιότυπων λέγεται ότι διαμερίζεται από ένα χώρο υποθέσεων αν και μόνο αν για κάθε διχοτόμηση του συνόλου στιγμιότυπων υπάρχει μια υπόθεση από το χώρο υποθέσεων που είναι συνεπής με αυτή τη διχοτόμηση, δηλαδή να ταξινομεί τα στιγμιότυπα και να τα χωρίζει σε δύο υποσύνολα που ταυτίζονται με τα σύνολα της διαμέρισης. Η διάσταση VC, λοιπόν, μετράει την ικανότητα του χώρου υποθέσεων να αναπαριστά όσων το δυνατόν περισσότερες διχοτομήσεις του συνόλου εκπαίδευσης.

Έτσι, η διάσταση Vapnik-Chervonenkis, $VC(H)$, του χώρου υποθέσεων H που ορίζεται σε ένα χώρο στιγμιότυπων X είναι το μέγεθος του μεγαλύτερου υποσύνολου του X που διαμερίζεται από το H . Έστω ότι ο χώρος των δεδομένων μας είναι ο $\mathbf{R}^2$ και το σύνολο $\{f(a)\}$ αποτελείται από προσανατολισμένες ευθείες γραμμές, έτσι ώστε για μια δεδομένη γραμμή, όλα τα σημεία από τη μία πλευρά είναι κλάσης 1 και όλα τα σημεία στην άλλη πλευρά είναι: -1. Ο προσανατολισμός παρουσιάζεται στην Εικόνα 43 με ένα τοξάκι υποδεικνύοντας την πλευρά που είναι τα σημεία κλάσης 1. Ενώ είναι πάντα δυνατό να βρει κανείς τρία σημεία τα οποία μπορούν να διαχωριστούν από αυτό το σύνολο συναρτήσεων, δεν είναι δυνατό να βρει κανείς τέσσερα με την ίδια ιδιότητα. Έτσι η VC-διάσταση των προσανατολισμένων υπερειπέδων στον $\mathbf{R}^2$ είναι 3.

Εικόνα 43. Τρία σημεία στον $\mathbb{R}^2$ που διαχωρίζονται από προσανατολισμένες γραμμές

Πρέπει να σημειωθεί ότι αν τυχαία μεγάλα πεπερασμένα υποσύνολα του X μπορούν να διχοτομηθούν από τον H , τότε $VC(H) \equiv +\infty$.

Το μέτρο της διάστασης Vapnik-Chervonenkis μπορεί να χρησιμοποιηθεί για να δώσει ένα νέο φράγμα για το πλήθος των ελάχιστων στιγμιοτύπων που απαιτούνται για την κατασκευή ενός αλγορίθμου που πληροί την ακρίβεια που ο εκπαιδευτής ζητάει. Το νέο αυτό φράγμα δίνεται από την έκφραση $m \geq \frac{1}{\varepsilon} \left(4 \log_2 \left(\frac{2}{\delta} \right) + 8VC(H) \log_2 \left(\frac{13}{\varepsilon} \right) \right)$. Αν ο χώρος H απαιτεί 2^d

διακριτές υποθέσεις για το διαμερισμό d στιγμιοτύπων, τότε $VC(H) = d$ αλλά και $2^d \leq |H|$.

Άρα $d = VC(H) \leq \log_2 |H|$. Κατά συνέπεια το φράγμα είναι αυστηρότερο, όμως ταυτόχρονα το σφάλμα δεν αποτελεί γραμμική συνάρτηση του σφάλματος ε , αλλά λογαριθμική.

Υπάρχει και ένας τύπος που δίνει ένα φράγμα κάτω από το οποίο δεν μπορεί να επιτευχθεί η επιθυμητή ακρίβεια. Έστω μια οποιαδήποτε κλάση C τέτοια ώστε $VC(C) \geq 2$, ένας

αλγόριθμος L , ένα $0 < \varepsilon < \frac{1}{8}$ και ένα $0 < \delta < \frac{1}{100}$. Τότε υπάρχει μια κατανομή D και μια

συνάρτηση-στόχος c , όπου αν ο αλγόριθμος L εκπαιδευτεί σε λιγότερα από $\max \left[\frac{1}{\varepsilon} \log \left(\frac{1}{\delta} \right), \frac{VC(C)-1}{32\varepsilon} \right]$ στιγμιότυπα, τότε με πιθανότητα τουλάχιστον δ , ο αλγόριθμος L

εξάγει ένα ταξινομητή h με σφάλμα $error_D(h) > \varepsilon$.

ΚΕΦΑΛΑΙΟ 4. Ομάδες ταξινομητών και συνεισφορά της διατριβής στη δημιουργία νέων μεθόδων

Στην περίπτωση που ενδιαφερόμαστε για την καλύτερη δυνατή ακρίβεια ταξινόμησης, είναι δύσκολο (αν όχι αδύνατο) να βρεθεί ένας ταξινομητής που να αποδίδει όπως ένας καλός συνδυασμός ταξινομητών (Dietterich, 2001). Η βασική ιδέα πίσω από τις ομάδες ταξινομητών είναι πως, δοθείσας μιας εργασίας που απαιτεί γνώσεις ειδικού (expert) για να εκτελεστεί με επιτυχία, πολλοί ειδικοί είναι αποτελεσματικότεροι από έναν, αν οι ατομικές τους αποφάσεις συνδυαστούν κατάλληλα.

Πολλές μέθοδοι κατασκευής ομάδων ταξινομητών έχουν προταθεί, άλλες που εφαρμόζονται σε συγκεκριμένους αλγόριθμους μάθησης και άλλες πιο γενικές. Οι τελευταίες χωρίζονται σε τέσσερις κατηγορίες (Roli και άλλοι, 2001), (Villada & Drissi, 2002):

- Στις τεχνικές που βασίζονται σε έναν αλγόριθμο μάθησης και για να παράγουν ομάδα ταξινομητών χρησιμοποιούν επαναληπτική δειγματοληψία στα δεδομένα εκπαίδευσης.
- Στις τεχνικές που χρησιμοποιούν μόνο το αρχικό σύνολο δεδομένων εκπαίδευσης αλλά συνδυάζουν αποφάσεις ταξινομητών που προέκυψαν από την εκπαίδευση διαφορετικών αλγορίθμων μάθησης.
- Στις τεχνικές που βασίζονται σε έναν αλγόριθμο μάθησης και για να παράγουν ομάδα ταξινομητών αλλάζουν την συνάρτηση στόχο (αυτή η τεχνική εφαρμόζεται μόνο στα πολυκατηγορικά προβλήματα).
- Στις τεχνικές που βασίζονται σε έναν αλγόριθμο μάθησης και για να παράγουν ομάδα ταξινομητών χρησιμοποιούν διαφορετική παραμετροποίηση ανά εκτέλεση του αλγορίθμου.

4.1 Τεχνικές που βασίζονται σε έναν αλγόριθμο μάθησης και την δειγματοληψία των δεδομένων εκπαίδευσης

Σε αυτή την κατηγορία οι διαφορετικοί ταξινομητές προκύπτουν από εκπαίδευση ενός αλγορίθμου σε διαφορετικά υποσύνολα των διαθέσιμων δεδομένων. Αυτή η κατηγορία μεθόδων εκμεταλλεύεται την αστάθεια που παρουσιάζουν διάφοροι αλγόριθμοι μάθησης, την υπερευαισθησία δηλαδή που επιδεικνύουν στις μικρές μεταβολές των δεδομένων της εισόδου. Σκοπός είναι η διαδοχική δημιουργία μοντέλων ικανών να αλληλοσυμπληρώνονται, υπό την έννοια ότι το κάθε ένα αποδίδει τα μέγιστα σε ένα υποσύνολο του σώματος εκπαίδευσης το οποίο τα υπόλοιπα δεν μπορούν να αξιοποιήσουν αποτελεσματικά. Όπως είναι αναμενόμενο, η μέθοδος δουλεύει ιδιαίτερα καλά για ασταθείς αλγορίθμους μάθησης – αλγορίθμους οι οποίοι παράγουν ταξινομητές αρκετά διαφορετικούς με μικρές αλλαγές του συνόλου εκπαίδευσης. Τα δέντρα απόφασης και τα νευρωνικά δίκτυα είναι παραδείγματα ασταθών αλγορίθμων, ενώ αντίθετα ο k-NN και ο απλός Μπεϋζιανός ταξινομητής (NB) είναι γενικά πολύ σταθεροί.

Η πιο απλή μέθοδος αυτής της κατηγορίας λέγεται σακούλιασμα (bagging ή Bootstrap AGGregation). Η μέθοδος αυτή συνίσταται στην παραγωγή ενός αριθμού μοντέλων (υπο-ταξινομητών), προερχόμενων από έναν κοινό αλγόριθμο μάθησης, χρησιμοποιώντας όμως διαφορετική διαμέριση του σώματος εκπαίδευσης (επιλογή με επανατοποθέτηση) για κάθε ένα εξ αυτών (Breiman, 1996). Η σχηματική αναπαράσταση της μεθόδου παρατίθεται στην Εικόνα 44, ενώ η αναλυτική περιγραφή σε ψευδογλώσσα παρουσιάζεται στην Εικόνα 45. Για τη λήψη της τελικής απόφασης ακολουθείται συνήθως η πλειοψηφική λογική. Η τελική απόφαση του συστήματος δηλαδή συμπίπτει με την απόφαση της πλειοψηφίας.

Εικόνα 44. Η γραφική αναπαράσταση του σακουλιάσματος (bagging)

Δημιουργία του μοντέλου

Έστω n είναι ο αριθμός των παραδειγμάτων των δεδομένων εκπαίδευσης

Κάνε για t επαναλήψεις:

Διάλεξε n παραδείγματα με επανατοποθέτηση από τα δεδομένα εκπαίδευσης.

Εφάρμοσε το αλγόριθμο μάθησης στο δείγμα δεδομένων που προέκυψε.

Αποθήκευσε το μοντέλο/ταξινομητή προέκυψε από τον αλγόριθμο.

Ταξινόμηση

Για καθένα από τα t μοντέλα:

Πρόβλεψε την κλάση του παραδείγματος.

Επέστρεψε την κλάση που τα περισσότερα μοντέλα επέλεξαν.

Εικόνα 45. Ψευδοκώδικας του σακουλιάσματος (bagging)

Παρόμοια τεχνική δανεισμένη από την τεχνική στατιστικής επικύρωσης είναι η επιτροπή διασταυρωμένης επικύρωσης (cross-validated committees), όπου τα υποσύνολα των δεδομένων στα οποία εκπαιδεύονται οι ταξινομητές καθορίζονται από τη γνωστή μέθοδο διασταυρωμένης επικύρωσης (Parmanto και άλλοι, 1996).

Μία πιο πολύπλοκη μέθοδος είναι η μέθοδος της λεγόμενης προώθησης (boosting) (Freund και Schapire, 1996). Εν συνέχεια, θα αναπτύξουμε εν συντομία τον αλγόριθμο AdaBoost.M1 (Freund και Schapire, 1996), ο οποίος αποτελεί τη βάση μιας ολόκληρης οικογένειας αλγορίθμων προώθησης. Ο αλγόριθμος αναθέτει στα στιγμιότυπα εκπαίδευσης ένα βάρος, ενδεικτικό της δυσκολίας που παρουσιάζει το υπό εκμάθηση τρέχον μοντέλο στην

ταξινόμησή του και κατ' επέκταση της βαρύτητας που θα πρέπει να δοθεί σε αυτό κατά την παραγωγή του επόμενου μοντέλου, προκειμένου να αναγνωρισθεί σωστά.

Αρχικά λοιπόν, όλα τα στιγμιότυπα εκπαίδευσης φέρουν το ίδιο βάρος $w = 1/N$, όπου N η πληθικότητα του σώματος εκπαίδευσης. Εν συνεχεία, το πρώτο μοντέλο εκπαιδεύεται και η απόδοσή του αξιολογείται, υπολογίζοντας το ζυγισμένο σφάλμα του e_1 επί του σώματος εκπαίδευσης ως το άθροισμα των βαρών των εσφαλμένως ταξινομηθέντων στιγμιοτύπων προς το συνολικό άθροισμα των βαρών. Παράλληλα, τα βάρη των στιγμιοτύπων που ταξινομήθηκαν λανθασμένα πολλαπλασιάζονται με τον παράγοντα $e/(1-e)$ και ακολουθεί κανονικοποίηση όλων των βαρών, ώστε να αθροίζονται στη μονάδα.

Αποτέλεσμα της παραπάνω διαδικασίας αποτελεί η αύξηση του βάρους των λανθασμένα ταξινομηθέντων στιγμιοτύπων και η αντίστοιχη μείωση του βάρους εκείνων που ταξινομήθηκαν σωστά. Αυτή η ακολουθία βημάτων επαναλαμβάνεται έως της δημιουργίας ενός καθορισμένου αριθμού μοντέλων, έστω b . Στην περίπτωση που το σφάλμα του τρέχοντος μοντέλου ξεπεράσει την τιμή 0.5 ή εξισωθεί με το μηδέν, τότε ο αλγόριθμος τερματίζει και το μοντέλο αυτό δε λαμβάνεται υπ' όψη.

Κατά την ταξινόμηση ενός άγνωστου στιγμιοτύπου, κάθε ένα από τα παραγόμενα μοντέλα προβαίνει στην εκτίμησή του, η οποία ωστόσο συμμετέχει με διαφορετική βαρύτητα στην τελική απόφαση. Πιο συγκεκριμένα, σε κάθε μοντέλο ανατίθεται ένας συντελεστής βαρύτητας ανάλογα με το σφάλμα του μοντέλου. Αυτός ο συντελεστής πολλαπλασιάζεται με την εκτίμηση του μοντέλου ώστε να αναδειχθεί η τελική απόφαση της επιτροπής.

Αναλυτικότερα, ο αλγόριθμος της προώθησης παρατίθεται σε ψευδογλώσσα στην Εικόνα 46.

Δημιουργία του μοντέλου

Θέσε όμοια βάρη σε όλα τα δεδομένα εισόδου.

Για t επαναλήψεις:

Εφάρμοσε τον αλγόριθμο μάθησης στα δεδομένα εκπαίδευσης και αποθήκευσε το μοντέλο.

Υπολόγισε το σφάλμα e του μοντέλου και αποθήκευσε το.

Αν το e ισούται με 0 ή είναι μεγαλύτερο ή ίσο του 0.5:

Τερμάτισε τις επαναλήψεις.

Για κάθε παράδειγμα των δεδομένων:

Αν το παράδειγμα ταξινομήθηκε σωστά από το μοντέλο:

Πολλαπλασίασε το βάρος του παραδείγματος με $e / (1 - e)$.

Ομαλοποίησε (normalize) τα βάρη όλων των παραδειγμάτων.

Ταξινόμηση

Θέσε το βάρος της κάθε κλάσης 0

Για καθένα από τα t (ή και λιγότερα) μοντέλα:

Πρόσθεσε $-\log(e / (1 - e))$ στο βάρος της κάθε κλάσης που προβλέφθηκε από το μοντέλο.

Επέστρεψε την κλάση με το μεγαλύτερο βάρος.

Εικόνα 46. Ψευδοκώδικας της προώθησης (boosting)

Ένα αξιοσημείωτο χαρακτηριστικό του αλγορίθμου της προώθησης αποτελεί το γεγονός ότι η εκτέλεση ενός αρκετά μεγάλου αριθμού επαναλήψεων του δείχνει να αποφέρει ευεργετικά αποτελέσματα, ακόμα και όταν το συνολικό σφάλμα του τελικού μοντέλου επί των δεδομένων εκπαίδευσης έχει προ πολλού ελαχιστοποιηθεί (Bauer & Kohavi, 1999).

Μια προσπάθεια θεωρητικής ερμηνείας της ιδιότητας αυτής (Schapire και άλλοι, 1998), κάνει χρήση της έννοιας των περιθωρίων ταξινόμησης (classification margins), σύμφωνα με την

οποία: όσο μεγαλύτερη είναι η απόσταση μεταξύ της πιθανότητας επιλογής της πραγματικής κλάσης ενός στιγμιοτύπου και της πιθανότητας εσφαλμένης επιλογής της πλέον επικρατούσας κλάσης, τόσο μεγαλύτερη εμπιστοσύνη δίδεται στις προβλέψεις ενός μοντέλου.

Το σακούλιασμα (bagging), η προώθηση (boosting) και διάφορες παραλλαγές τους συγκρίνονται πειραματικά από τους (Kohavi & Bauer, 1999). Το σακούλιασμα φαίνεται ότι δίνει πάντα καλύτερα αποτελέσματα από την χρήση μόνο του βασικού αλγορίθμου, αλλά η προώθηση κατά μέσο όρο αυξάνει περισσότερο την ακρίβεια του βασικού μοντέλου. Πρέπει όμως να σημειωθεί ότι η προώθηση για ορισμένα δεδομένα εκπαίδευσης δίνει χειρότερα αποτελέσματα από τον βασικό αλγόριθμο μάθησης λόγω του υπερ-ταιριάσματος.

Τα αποτελέσματα άλλων ερευνητών που συγκρίναν εμπειρικά το σακούλιασμα με την προώθηση παρατίθενται στον Πίνακα 30.

Αναφορά	Ταξινομητής	Σχετική βελτίωση (%)	
		Σακούλιασμα	Προώθηση
Freund & Schapire (1996)	C4.5	20.0	24.8
Quinlan (1996)	C4.5	10.0	15.01
Opitz & Maclin (1997)	C4.5	18.5	22.0
Opitz & Maclin (1997)	BP	13.3	17.1
Bauer & Kohavi (1999)	C4.5	14.5	27.0
Bauer & Kohavi (1999)	BP	3.0	24.0

Πίνακας 30. Σχετική βελτίωση της ακρίβειας των ταξινομητών μετά την χρήση σακουλιάσματος και προώθησης

Η απόφαση σχετικά με τον περιορισμό του αριθμού των υπο-ταξινομητών είναι σημαντική για πρακτικές εφαρμογές. Για να είναι ανταγωνιστικές οι μέθοδοι είναι σημαντικό οι αλγόριθμοι να τρέχουν σε λογικό χρόνο. Ο Quinlan (1996) χρησιμοποίησε μόνο 10 υπο-ταξινομητές, ενώ οι Bauer & Kohavi (1999) χρησιμοποίησαν 25 υπο-ταξινομητές, και οι Freund & Schapire (1996) χρησιμοποίησαν 100. Φαίνεται ότι οι μέθοδοι συνεχίζουν να βελτιώνουν το σφάλμα τους μέχρι περίπου τους 25 υπο-ταξινομητές (Opitz & Maclin, 1999).

Ο αλγόριθμος Arcing (Breiman, 1998) είναι ένας άλλος αλγόριθμος που μοιάζει στον AdaBoost. Όπως AdaBoost, ο αλγόριθμος δημιουργεί διαδοχικά ταξινομητές αλλά αντίθετα με τον AdaBoost, οι παραγόμενοι ταξινομητές έχουν ίδιο βάρος για την τελική απόφαση.

Ο αλγόριθμος LogitBoost ανάγει τη μέθοδο της προώθησης στην παραγωγή ενός μοντέλου Προσθετικής Λογιστικής Παλινδρόμησης (Additive Logistic Regression) μέσω της κατά στάδια βελτιστοποίησης της λογαριθμικής πιθανοφάνειας του Bernoulli (Friedman και άλλοι, 2000)¹².

Ο αλγόριθμος MultiBoosting (Webb, 2000) είναι μια άλλη μετα-μέθοδος της ίδιας κατηγορίας που μπορεί να θεωρηθεί ως ανατροφοδοτούμενες ομάδες (wagging ensembles) που δημιουργούνται από τον αλγόριθμο AdaBoost. Οι ανατροφοδοτούμενες ομάδες είναι μια παραλλαγή του σακουλιάσματος (bagging). Αντί να επιλέγονται τα παραδείγματα κάθε επανάληψης με επανάθεση, μεταβάλλονται τυχαία τα βάρη του κάθε παραδείγματος.

Ένας άλλος αλγόριθμος που χρησιμοποιεί έναν αλγόριθμο μάθησης για να χτίσει μια ομάδα είναι ο DECORATE (Melville & Mooney, 2003). Η ομάδα σχηματίζεται με την προσθήκη

¹² Περισσότερα για αυτόν τον αλγόριθμο στην ενότητα 4.1.5

διαφορετικών τυχαία κατασκευασμένων παραδειγμάτων στα δεδομένα, κατά την οικοδόμηση των νέων μελών της ομάδας.

Η αποτελεσματικότητα του συνδυασμού των ταξινομητών βασίζεται στο παρακάτω θεώρημα (Lam, 2000):

Υποθέτοντας $n \geq 3$ το πλήθος των ταξινομητών. Τότε:

Αν κάθε ταξινομητής βάσης έχει πιθανότητα ακριβής πρόγωσης $p > 0.5$, τότε η πιθανότητα ορθής ταξινόμησης από την ομάδα των ταξινομητών χρησιμοποιώντας την απλή ψηφοφορία είναι: $PC(n) \rightarrow 1$ όταν $n \rightarrow \infty$.

4.1.1 Προτεινόμενη μέθοδος συνδυασμού σακουλιάσματος και προώθησης

Αρχικά συνδυάσαμε το σακούλιασμα (bagging) με την προώθηση (boosting) με ένα τρόπο διαφορετικό από αυτόν του αλγόριθμου MultiBoosting. Το προτεινόμενο μοντέλο (Vote B&B) χρησιμοποιώντας το δεδομένα εκπαίδευσης χτίζει 10 υπο-ταξινομητές (sub-classifiers) με τη χρήση του σακουλιάσματος και άλλους 10 με τη χρήση του αλγόριθμου Adaboost και κατόπιν συνδυάζει τις αποφάσεις τους με απλή ψηφοφορία. Στην Εικόνα 47 δίνεται η γραφική περιγραφή της προτεινόμενης μεθόδου.

Εικόνα 47. Η γραφική περιγραφή της προτεινόμενης μεθόδου Vote B&B

Τα πειράματα απέδειξαν ότι η προτεινόμενη μέθοδος δίνει καλύτερα αποτελέσματα από το σακούλιασμα, την προώθηση, τον αλγόριθμο Multiboosting και τον αλγόριθμο DECORATE με 25 υπο-ταξινομητές (δηλαδή χρησιμοποιώντας 5 υπο-ταξινομητές περισσότερους από την προτεινόμενη μέθοδο). Για τα πειράματα χρησιμοποιήσαμε σαν βασικούς αλγορίθμους τέσσερις γνωστούς αλγορίθμους που έχουμε περιγράψει ήδη στο τέταρτο κεφάλαιο, τους: C4.5, Decision Stump (DS), αφελή Μπεϋζιανό αλγόριθμο (NB) και OneR.

Για τη διεξαγωγή των πειραμάτων χρησιμοποιήθηκε η πλατφόρμα Μηχανικής Μάθησης WEKA¹³, η οποία αναπτύχθηκε στο τμήμα της Επιστήμης Υπολογιστών του Πανεπιστημίου του Waikato της Νέας Ζηλανδίας. Το συγκεκριμένο πρόγραμμα έχει υλοποιηθεί σε Java και διατίθεται υπό μορφή βιβλιοθήκης, μαζί με τον πηγαίο κώδικά του, καθιστώντας δυνατή τη χρησιμοποίηση των επιμέρους τμημάτων του από άλλα προγράμματα υψηλότερου επιπέδου σαν αυτά που αναπτύξαμε και εμείς.

¹³ Περισσότερα στοιχεία για το πρόγραμμα WEKA στο Παράρτημα

Σε όλα τα πειράματα μας χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Πιο συγκεκριμένα, κάθε σύνολο δεδομένων διαιρέθηκε τυχαία σε δέκα ίσα μέρη. Κάθε πείραμα επαναλήφθηκε δέκα φορές, με διαφορετικό ανά φορά τμήμα για τον έλεγχο και με τα υπόλοιπα εννέα να συστήνουν το σύνολο εκπαίδευσης. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων.

Η μέθοδος στατιστικού ελέγχου που χρησιμοποιήθηκε ήταν η βελτιωμένη δοκιμή- t κατά ζεύγη. Η δοκιμή- t αποδίδει την πιθανότητα δυο δείγματα να προέρχονται από δύο ίδιους πληθυσμούς με την ίδια μέση τιμή, υποθέτοντας πως οι πληθυσμοί ακολουθούν την κατανομή t . Η κατά ζεύγη εκδοχή της δοκιμής- t είναι κατάλληλη όταν υπάρχει ένα-προς-ένα αντιστοιχία μεταξύ των τιμών των δειγμάτων. Στη συγκεκριμένη περίπτωση, η αντιστοιχία αφορά την επίδοση δύο υπό σύγκριση αλγορίθμων στο ίδιο τμήμα (fold) της διασταυρωμένης επικύρωσης, δηλαδή όταν έχουν εκπαιδευθεί πάνω στο ίδιο σύνολο εκπαίδευσης και εξεταστεί πάνω στο ίδιο σύνολο ελέγχου. Η τιμή p είναι μια πιθανότητα, με αξία που κυμαίνεται από μηδέν ως ένα. Εάν η τιμή p είναι μικρή, συμπεραίνουμε ότι η διαφορά μεταξύ των μέσων όρων είναι απίθανο να αποτελεί σύμπτωση. Αντ' αυτού, θα συμπεράνουμε ότι οι πληθυσμοί έχουν διαφορετικούς μέσους όρους (Nadeau & Bengio, 2003). Εάν για παράδειγμα η τιμή p είναι 0,01, αυτό σημαίνει ότι υπάρχει μια πιθανότητα 1% να αποφανθούμε ότι υπάρχει διαφορά όταν οι μέσοι όροι των δύο πληθυσμών είναι ίδιοι (η μηδενική υπόθεση ισχύει).

Βάσει των πειραμάτων μας (βλέπε Πίνακα 31) ο προτεινόμενος αλγόριθμος χρησιμοποιώντας σαν βασικό ταξινομητή τον C4.5 είναι στατιστικά ακριβέστερος από τον ταξινομητή C4.5 στα 7 από τα εξεταζόμενα σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Πρέπει να τονίσουμε ότι στον Πίνακα 31, συμβολίζεται με "*" ότι η συγκεκριμένη ομάδα ταξινομητών υστερεί σε επίδοση από τον ταξινομητή βάσης. Δηλαδή ο ταξινομητή βάσης έδωσε στατιστικά καλύτερα αποτελέσματα από συγκεκριμένη ομάδα ταξινομητών σύμφωνα με τη δοκιμή- t κατά ζεύγη με $p < 0.01$. Επιπλέον, στον πίνακα, συμβολίζεται με "v" ότι ο ταξινομητή βάσης έδωσε στατιστικά χειρότερα αποτελέσματα από συγκεκριμένη ομάδα ταξινομητών σύμφωνα με τη δοκιμή- t κατά ζεύγη με $p < 0.01$. Σε όλες τις άλλες περιπτώσεις, δεν υπάρχει καμία σημαντική στατιστική διαφορά μεταξύ των αποτελεσμάτων (ισοπαλία). Συγκεντρωτικά, στην τελευταία γραμμή του πίνακα παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c). Το «a» σημαίνει ότι η συγκεκριμένη ομάδα ταξινομητών είναι στατιστικά πιο ακριβείς από τον αλγόριθμο βάσης σε «a» σύνολα δεδομένων, το «c» σημαίνει ότι ο αλγόριθμος βάσης είναι στατιστικά ακριβέστερος από τη συγκεκριμένη ομάδα ταξινομητών σε «c» σύνολα δεδομένων, ενώ το «b» ότι δεν υπάρχει στατιστική διαφορά μεταξύ των αποτελεσμάτων.

Αντίστοιχα, η τεχνική σακουλιάσματος (bagging) χρησιμοποιώντας σαν βασικό ταξινομητή τον C4.5 είναι στατιστικά ακριβέστερη από τον C4.5 σε μόλις 3 από τα εξεταζόμενα σύνολα δεδομένων ενώ και αυτή η τεχνική δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Η τεχνική προώθησης έχει στατιστικά χαμηλότερα ποσοστά λάθους στα 8 από τα εξεταζόμενα σύνολα δεδομένων από τον C4.5, ενώ είναι στατιστικά λιγότερο ακριβής σε 3 σύνολα δεδομένων. Επίσης, ο αλγόριθμος Multiboost είναι στατιστικά ακριβέστερος από τον C4.5 στα 7 από τα εξεταζόμενα σύνολα δεδομένων και έχει στατιστικά υψηλότερα ποσοστά λάθους σε ένα σύνολο δεδομένων. Τέλος, ο αλγόριθμος DECORATE έχει στατιστικά χαμηλότερα ποσοστά λάθους στα 6 από τα εξεταζόμενα σύνολα δεδομένων από τον C4.5, ενώ είναι στατιστικά λιγότερο ακριβής σε 1 σύνολο δεδομένων.

Για να συνοψίσουμε, η απόδοση της προτεινόμενης μεθόδου είναι καλύτερη από όλες τις άλλες γνωστές μεθόδους που χρησιμοποιούν τον C4.5 ως βασικό ταξινομητή (καλύτερη

διαφορά «a»-«c»). Η προτεινόμενη μέθοδος μπορεί να επιτύχει μια μείωση του ποσοστού λάθους περίπου της τάξης του 15% έναντι του απλού C4.5.

Δεδομένα	C4.5	Bagging C4.5	Adaboost C4.5	Multiboost C4.5	Decorate C4.5	Vote B&B C4.5
anneal	98.57	98.83	99.61 v	99.59 v	98.75	99.62 v
audiology	77.26	81.29	84.62 v	85.31 v	82.06	84.92 v
autos	81.77	83.89	86.05	85.75	83.86	85.7
badges	100	100	100	100	100	100
balance-scale	77.82	82.33 v	76.91	79.26	81.31 v	79.16
breast-cancer	74.28	73.37	66.5 *	68.27	70.97	69.79
breast-w	95.01	96.31	96.51	96.51	96.5	96.24
colic	85.16	85.23	82.01	83.13	84.82	83.53
credit-g	71.25	74.17	72.79	74.35	72.99	72.63
diabetes	74.49	75.67	72.81	74.21	75.28	73.1
dimin	97.09	97.07	96.03 *	96.17 *	97.2	96.6
glass	67.63	74.86	77.3 v	77.1 v	73	75.85 v
haberman	71.05	72.06	71.12	71.12	73.42	72.16
heart-c	76.94	79.54	79.6	79.87	79.12	79.22
heart-h	80.22	79.91	78.28	80.11	79.3	79.4
heart-statlog	78.15	81.11	80.15	81.37	80.41	80.33
hepatitis	79.22	81.63	82.74	83.59	82.45	82.78
hypothyroid	99.54	99.58	99.67	99.68	98.58 v	99.65
ionosphere	89.74	92.23	93.62	93.54	92.68	93.02
iris	94.73	94.8	94.47	94.47	94.93	94.33
kr-vs-kp	99.44	99.45	99.62	99.62	99.25	99.61
lymphography	75.84	79.14	83.09	83.24	79.09	80.67
monk1	80.61	82.32	96.54 v	94.36 v	89.94	93.87 v
monk2	57.75	61.15	61.86	61.28	58.33	60.97
primary-tumor	41.39	44.4	41.65	41.65	44.57	43.28
segment	96.79	97.54	98.42 v	98.36 v	98.09 v	98.15 v
sick	98.72	98.86	99.06	99.08	98.49 v	99
sonar	73.61	79.03	83.03 v	82.54	82.31	80.08
soybean	91.78	92.78	93.19	93.21	93.86	93.12
students	86.75	86.49	81.44 *	81.68	82.14 *	82.63
titanic	78.55	77.93	78.89	78.71	78.85	78.24
vote	96.57	96.53	95.24	95.51	95.4	96.2
vowel	80.2	91.75 v	95.42 v	94.94 v	96.57 v	93.34 v
waveform	75.25	82.81 v	83.32 v	83.73 v	81.12 v	82.12 v
wine	93.2	95.5	96.62	96.84	97.28	96.56
zoo	92.61	93.29	95.38	95.77	92.96	95.08
		3/33/0	8/25/3	7/28/1	6/29/1	7/29/0

Πίνακας 31. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον C4.5

Όπως προείπαμε στο δεύτερο πείραμα χρησιμοποιήσαμε σαν βασικό αλγόριθμο τον DS (βλέπε Πίνακα 32). Ο προτεινόμενος αλγόριθμος χρησιμοποιώντας σαν βασικό ταξινομητή τον DS είναι στατιστικά ακριβέστερος από τον ταξινομητή DS στα 14 από τα εξεταζόμενα

σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Δεδομένα	DS	Bagging DS	Adaboost DS	Multiboost DS	Decorate DS	VOTE B&B DS
anneal	77.17	82.62 v	83.63 v	83.63 v	76.89	82.81 v
audiology	46.46	46.46	46.46	46.46	46.46	46.46
autos	44.9	44.9	44.9	44.9	51.81	44.9
badges	100	100	100	100	100	100
balance-scale	56.72	68.88 v	71.77 v	71.77 v	81.25 v	72.6 v
breast-cancer	69.27	73.44	71.55	71.76	75.18 v	73.03
breast-w	92.33	92.63	95.28 v	95.05 v	95.78 v	95.12 v
colic	81.52	81.52	82.72	82.9	82.03	82.42
credit-g	70	70	72.6	71.8	69.71	70.65
diabetes	71.8	72.55	75.37	75.19	76.09 v	74.7 v
dimin	59.31	59.31	59.31	59.31	64.75 v	59.31
glass	44.89	44.99	44.89	44.89	53.12 v	44.81
haberman	71.57	72.74	74.06	73.8	71.61	72.95
heart-c	72.93	75.64	83.11 v	83.54 v	72.43	82.45 v
heart-h	81.78	81.37	82.42	81.91	81.78	81.81
heart-statlog	72.3	75.04	81.81 v	82.89 v	81.48 v	81.52 v
hepatitis	77.62	80.72	81.5	82.21	80.02	81.51
hypothyroid	95.39	95.39	92.97 *	92.97 *	95.39	95.38
ionosphere	82.57	82.54	92.34 v	90 v	90.4 v	87.41 v
iris	66.67	70.33	95.07 v	94.73 v	93.93 v	95.4 v
kr-vs-kp	66.05	66.05	95.08 v	93.9 v	90.43 v	94.09 v
lymphography	75.31	74.63	75.44	74.96	72.25 *	75.44
monk1	73.41	73.41	69.79 *	70.37	70.94 *	73.41
monk2	59.58	61.31	53.99	54.19	61.95	56.67
primary-tumor	28.91	28.91	28.91	28.91	29.09	28.91
segment	28.52	56.54 v	28.52	28.52	53.91 v	54.53 v
sick	96.55	96.55	97.07	97.14	96.57	97.11 v
sonar	72.25	73.16	81.06 v	77.58	72.91	75.56
soybean	27.96	27.88	27.96	27.96	41.42 v	27.74
students	87.22	87.22	87.16	86.95	87.1	87.22
titanic	77.6	77.6	77.83	77.62	77.6	77.6
vote	95.63	95.63	96.41	95.63	95.59	95.22
vowel	17.47	23.52 v	17.47	17.47	32.08 v	21.72 v
waveform	56.82	57.41	67.68 v	66.44 v	68.7 v	59.47 v
wine	57.91	85.16 v	91.57 v	91.17 v	96.45 v	88.7 v
zoo	60.43	60.63	60.43	60.43	61.96	60.43
		5/31/0	11/23/2	10/25/1	15/19/2	14/22/0

Πίνακας 32. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον DS

Αντίστοιχα, η τεχνική σακουλιάσματος χρησιμοποιώντας σαν βασικό ταξινομητή τον DS είναι στατιστικά ακριβέστερη από τον DS σε μόλις 5 από τα εξεταζόμενα σύνολα δεδομένων αλλά και αυτή η τεχνική δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο

δεδομένων. Η τεχνική προώθησης έχει στατιστικά χαμηλότερα ποσοστά λάθους σε 11 από τα εξεταζόμενα σύνολα δεδομένων από τον DS, ενώ είναι στατιστικά λιγότερο ακριβής σε 2 σύνολα δεδομένων. Επίσης, ο αλγόριθμος Multiboost είναι στατιστικά ακριβέστερος από τον DS στα 10 από τα εξεταζόμενα σύνολα δεδομένων και έχει στατιστικά υψηλότερα ποσοστά λάθους σε ένα σύνολο δεδομένων. Τέλος, ο αλγόριθμος DECORATE έχει στατιστικά χαμηλότερα ποσοστά λάθους σε 15 από τα εξεταζόμενα σύνολα δεδομένων από τον DS, ενώ είναι στατιστικά λιγότερο ακριβής σε 2 σύνολα δεδομένων.

Για να συνοψίσουμε, η απόδοση της προτεινόμενης μεθόδου είναι καλύτερη από όλες τις άλλες γνωστές μεθόδους που χρησιμοποιούν τον DS ως βασικό ταξινομητή (base-learner). Η προτεινόμενη μέθοδος μπορεί να επιτύχει μια μείωση του ποσοστού λάθους περίπου της τάξης του 16% έναντι του απλού DS.

Όπως προείπαμε στο τρίτο πείραμα χρησιμοποιήσαμε σαν βασικό αλγόριθμο τον NB (βλέπε Πίνακα 33). Ο προτεινόμενος αλγόριθμος χρησιμοποιώντας σαν βασικό ταξινομητή τον NB είναι στατιστικά ακριβέστερος από τον ταξινομητή NB σε 8 από τα εξεταζόμενα σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Η τεχνική σακουλιάσματος χρησιμοποιώντας σαν βασικό ταξινομητή τον NB δεν είναι αποτελεσματική αφού δίνει περίπου τα ίδια αποτελέσματα (αυτό ήταν αναμενόμενο αφού ο αλγόριθμος Bagging δεν δίνει καλύτερα αποτελέσματα αν συνδυαστεί με σταθερό βασικό μοντέλο). Η τεχνική προώθησης χρησιμοποιώντας σαν βασικό ταξινομητή τον NB έχει στατιστικά χαμηλότερα ποσοστά λάθους στα 6 από τα εξεταζόμενα σύνολα δεδομένων από τον NB, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Επίσης, ο αλγόριθμος Multiboost είναι στατιστικά ακριβέστερος από τον NB στα 7 από τα εξεταζόμενα σύνολα δεδομένων ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Τέλος, ο αλγόριθμος DECORATE έχει στατιστικά χαμηλότερα ποσοστά λάθους μόλις σε 1 από τα εξεταζόμενα σύνολα δεδομένων από τον NB, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Για να συνοψίσουμε, η απόδοση της προτεινόμενης μεθόδου είναι καλύτερη από όλες τις άλλες γνωστές μεθόδους που χρησιμοποιούν τον NB ως βασικό ταξινομητή (base-learner). Η προτεινόμενη μέθοδος μπορεί να επιτύχει μια μείωση του ποσοστού λάθους περίπου της τάξης του 9% έναντι του απλού NB.

Δεδομένα	NB	Bagging NB	Adaboost NB	Multiboost NB	Decorate NB	VOTE B&B NB
anneal	86.59	86.94	95.2 v	94.13 v	86.59	92.64 v
audiology	72.64	72.1	78.2	80.1 v	72.16	79.22 v
autos	57.41	57.15	57.12	57.12	57.82	57.04
badges	99.66	99.69	99.66	99.66	96.73	99.66
balance-scale	90.53	90.32	92.11	92.29	90.63	91.28
breast-cancer	72.7	72.7	68.57	69.01	72.94	72.45
breast-w	96.07	96.11	95.55	95.58	95.94	96.07
colic	78.7	78.59	77.46	79.28	78.05	80.44
credit-g	75.16	75.08	75.09	74.71	74.7	75.54
diabetes	75.75	75.7	75.88	76.2	75.3	76.03
dimin	92.86	92.77	92.93	93.95	92.86	94.52 v
glass	49.45	50.05	49.63	49.63	49.5	50.14
haberman	75.06	74.86	73.94	73.94	74.83	75.25
heart-c	83.34	83.44	83.14	83.56	83.41	83.54
heart-h	83.95	83.99	84.67	84.8	84.05	84.36
heart-statlog	83.59	83.78	82.3	82.7	83.67	83.74
hepatitis	83.81	83.93	84.23	84.67	82.92	85.89
hypothyroid	95.3	95.47	95.27	95.33	95.3	95.31
ionosphere	82.17	82.36	91.12 v	91.66 v	83.08	88.41 v
iris	95.53	95.73	95.07	95.07	94.87	95.73
kr-vs-kp	87.79	87.77	95.1 v	95.22 v	87.66	93.04 v
lymphography	83.13	83.14	80.67	82.64	82.98	82.98
monk1	73.38	73.35	72.37	72.42	75.65	73.21
monk2	56.83	56.49	56.83	56.83	57.02	56.78
primary-tumor	49.71	49.62	49.71	49.71	49.18	49.32
segment	80.17	80.26	80.17	80.17	80.1	80.31
sick	92.75	92.77	93.7	93.65	92.75	93.17
sonar	67.71	68.05	81.21 v	81.5 v	67.42	74.58 v
soybean	92.94	92.83	92.02	93.15	92.62	93.73
students	85.7	85.59	85.12	85.47	85.09	85.99
titanic	77.85	77.86	77.86	77.88	78.31 v	77.86
vote	90.02	90.05	95.19 v	95.36 v	89.93	92.88 v
vowel	62.9	63.36	81.32 v	79.42 v	62.08	76.54 v
waveform	80.01	80	80.01	80.29	80.4	80.01
wine	97.46	97.52	96.57	96.57	96.51	96.62
zoo	94.97	95.07	97.23	97.23	94.68	97.43
		0/36/0	6/30/0	7/29/0	1/35/0	8/28/0

Πίνακας 33. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον NB

Όπως προείπαμε στο τέταρτο πείραμα χρησιμοποιήσαμε σαν βασικό αλγόριθμο τον OneR (βλέπε Πίνακα 34). Ο προτεινόμενος αλγόριθμος χρησιμοποιώντας σαν βασικό ταξινομητή τον OneR είναι στατιστικά ακριβέστερος από τον ταξινομητή OneR σε 9 από τα εξεταζόμενα

σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Δεδομένα	OneR	Bagging OneR	Adaboost OneR	Multiboost OneR	Decorate OneR	VOTE B&B OneR
anneal	83.63	83.63	85.68	85.06	83.63	83.63
audiology	46.46	46.46	46.46	46.46	46.46	46.46
autos	61.57	64.72	65.97	66.06	70.51 v	65.6
Badges	28.55	28.55	28.55	28.55	43.1 v	28.55
balance-scale	57.09	68.28 v	72.84 v	72.84 v	60.94	71.78 v
breast-cancer	66.91	68.81	70.07	70	66.87	69.89
breast-w	92.01	92.95	95.52 v	95.45 v	92.38	94.77 v
Colic	81.52	81.52	81.01	80.32	66.13	81.49
credit-g	66.23	68.02	64.31	64.98	71.98	67.53
diabetes	71.71	72.07	69.65	69.95	89.53	73.1
dimin	89.53	89.53	96.39 v	94.67 v	58.76	89.51
glass	57.09	59.69	56.43	56.81	71.97	59.52
haberman	72.79	72	71.3	71.74	72.83	72.4
heart-c	72.53	76.13	73.31	76.24	80.36	78.06 v
heart-h	80.69	80.87	76.75	77.67	74.93	81.31
heart-statlog	71.26	75.85	71.56	75.26	80.94	77.11 v
hepatitis	82.49	82.41	77.56	78.73	78.15	80.92
hypothyroid	96.43	96.54	96.65	96.72	96.38	96.78
ionosphere	82.48	84.98	89.17 v	88.91 v	85.59	87.77 v
iris	93.4	93.73	94	94.53	93.47	94.27
kr-vs-kp	66.91	67.06	94.43 v	93.55 v	66.91	85.04 v
lymphography	74.77	74.7	80.89	79.47	70.66	75.97
monk1	73.41	73.41	71.1	70.44	74.35	73.16
monk2	58.35	58.4	54.71	54.73	57.22	58.46
primary-tumor	27.74	26.76	27.25	27.25	27.88	27.25
segment	64	65.28	84.41 v	80.83 v	68.38 v	71.55 v
sick	96.29	96.34	96.61	96.08	96.24	96.61
sonar	62.36	70.59 v	64.38	65.69	62.44	65.38
soybean	39.75	40.13	40.63	40.63	40.28	40.63
students	87.22	87.22	86.37	87.04	87.04	87.22
titanic	77.6	77.6	77.74	77.64	77.6	77.6
vote	95.63	95.63	96.48	95.91	95.63	95.54
vowel	32.95	35.85	31.33	31.33	40.6 v	31.33
waveform	53.74	55.01	73.76 v	71.28 v	55.68	60.58 v
wine	78.03	79.05	92.08 v	89.7 v	79.15	85.56 v
zoo	42.59	42.59	42.59	42.59	42.59	42.59
		2/34/0	8/28/0	8/28/0	4/32/0	9/27/0

Πίνακας 34. Σύγκριση της προτεινόμενης μεθόδου Vote B&B με τις υπόλοιπες μεθόδους χρησιμοποιώντας ως βασικό ταξινομητή τον OneR

Αντίστοιχα, η τεχνική σακουλιάσματος χρησιμοποιώντας σαν βασικό ταξινομητή τον OneR είναι στατιστικά ακριβέστερη από τον OneR σε μόλις 2 από τα εξεταζόμενα σύνολα δεδομένων ενώ και αυτή η τεχνική δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε

κανένα σύνολο δεδομένων. Η αλγόριθμοι Adaboost και Multiboost χρησιμοποιώντας σαν βασικό ταξινομητή τον OneR έχουν στατιστικά χαμηλότερα ποσοστά λάθους στα 8 από τα εξεταζόμενα σύνολα δεδομένων από τον OneR, ενώ δεν έχουν στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Τέλος, ο αλγόριθμος DECORATE έχει στατιστικά χαμηλότερα ποσοστά λάθους μόλις σε 4 από τα εξεταζόμενα σύνολα δεδομένων από τον OneR, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Για να συνοψίσουμε, η απόδοση της προτεινόμενης μεθόδου είναι καλύτερη από όλες τις άλλες γνωστές μεθόδους που χρησιμοποιούν τον OneR ως βασικό ταξινομητή (base-learner). Η προτεινόμενη μέθοδος μπορεί να επιτύχει μια μείωση του ποσοστού λάθους περίπου της τάξης του 9% έναντι του απλού OneR.

Εκτός από καλύτερη απόδοση σε όλα τα πειράματα η προτεινόμενη τεχνική χρησιμοποιεί και λιγότερο χρόνο για εκπαίδευση. Αυτό συμβαίνει γιατί χρησιμοποιεί 20 υπο-ταξινομητές και όχι 25 που χρησιμοποιούν οι άλλες μέθοδοι.

Τέλος, πρέπει να αναφερθεί ότι η προτεινόμενη μεθοδολογία μπορεί εύκολα να εκτελεστεί παράλληλα εκτός από σειριακά χρησιμοποιώντας μια ομάδα ταξινομητών ανά μηχανή. Γενικά ένας αλγόριθμος μηχανικής μάθησης εκμεταλλευόμενος την παράλληλη ή/και καταναμημένη εκτέλεση μπορεί να αυξήσει την ταχύτητά του και το εύρος των εφαρμογών όπου μπορεί να χρησιμοποιηθεί (επειδή μπορεί να επεξεργαστεί περισσότερα δεδομένα εκπαίδευσης).

4.1.2 Προτεινόμενη μέθοδος τοπικής προώθησης

Στη συνέχεια της έρευνας μας, μελετήσαμε το συνδυασμό της τοπικής μάθησης με την προώθηση. Τα πειράματα απέδειξαν ότι αυτός ο συνδυασμός δίνει καλύτερα αποτελέσματα από την απλή προώθηση. Ο κύριος λόγος είναι ότι οι τοπικές μέθοδοι μπορούν να προσεγγίσουν πολύ καλύτερα από άλλες μεθόδους τη συνάρτηση-στόχο αν αυτή είναι πολύπλοκη καθολικά, αλλά μπορεί να περιγραφεί ως μια συλλογή λιγότερο σύνθετων τοπικών προσεγγίσεων (Atkeson και άλλο, 1997a).

Είναι γνωστό ότι όταν το μέγεθος του συνόλου κατάρτισης είναι μικρό έναντι στην πολυπλοκότητα του ταξινομητή, η μηχανή μάθησης υπερταυριάζει τον θόρυβο που βρίσκεται στα δεδομένα κατάρτισης (εκπαίδευσης). Συνεπώς, ο αποτελεσματικός έλεγχος της πολυπλοκότητας του ταξινομητή διαδραματίζει βασικό ρόλο στην επίτευξη καλής γενίκευσης. Μερικά θεωρητικά και πειραματικά αποτελέσματα (Vapnik, 1998) δείχνουν ότι ένας τοπικός αλγόριθμος μάθησης (που εκπαιδεύεται σε ένα υποσύνολο των δεδομένων κατάρτισης) παρέχει μια αποτελεσματική λύση σε αυτό πρόβλημα.

Γι' αυτό το λόγο συνδυάσαμε την τοπική μάθηση με την προώθηση σε μια προσπάθεια περαιτέρω βελτίωσης των τοπικών μεθόδων στις περιπτώσεις που τα δεδομένα εκπαίδευσης είναι περιορισμένα. Παρακάτω ακολουθεί ο προτεινόμενος αλγόριθμος.

1. Καθόρισε την κατάλληλη μετρική απόστασης (distance metric).
2. Βρες τους k κοντινότερους γείτονες χρησιμοποιώντας αυτή την μετρική.
3. Εφάρμοσε την τεχνική της προώθησης (boosting) σε ένα απλό ταξινομητή χρησιμοποιώντας τα k παραδείγματα ως δεδομένα εκπαίδευσης.
4. Η απάντηση της ομάδας των ταξινομητών που δημιουργήθηκαν από την προώθηση είναι η κλάση του παραδείγματος που έχουμε προς ταξινόμηση.

Εικόνα 48. Ο προτεινόμενος αλγόριθμος τοπικής προώθησης

Ο προτεινόμενος αλγόριθμος έχει μερικές ελεύθερες παραμέτρους όπως η μετρική απόστασης και το πλήθος των γειτόνων. Στα πειράματά μας, χρησιμοποιήσαμε την πιο συχνά

χρησιμοποιούμενη μετρική - την Ευκλείδεια απόσταση. Χρησιμοποιήσαμε επίσης το 50 ως την τιμή των k πλησιέστερων γειτόνων δεδομένου ότι αυτό το μέγεθος περιπτώσεων αποδείχθηκε κατάλληλο βάσει της εργασίας (Frank και άλλοι, 2003), ώστε ένας απλός αλγόριθμος να χτίζει ένα ακριβές μοντέλο. Ο συχνά χρησιμοποιούμενος αλγόριθμος AdaBoost (Freund και Schapire, 1996) ήταν η επιλογή μας για μέθοδο προώθησης. Τέλος, χρησιμοποιήσαμε 10 επαναλήψεις για την προώθηση προκειμένου να μην επιβαρυνθεί πολύ ο χρόνος για την ταξινόμηση μιας νέας περίπτωσης. Για τα πειράματα χρησιμοποιήσαμε 2 ταξινομητές ως βασικούς, τους: Decision Stump (DS) και OneR.

Αρχικά, συγκρίναμε την προτεινόμενη μέθοδο χρησιμοποιώντας σαν βασικό ταξινομητή τον DS με τους ακόλουθους αλγόριθμους:

- 50-κοντινότερους γείτονες (50-nearest neighbors) αφού η προτεινόμενη μέθοδος χρησιμοποιεί τα κοντινότερα 50 παραδείγματα. Επίσης, χρησιμοποιήσαμε έναν ακόμα τοπικό αλγόριθμο μάθησης τον Kstar.
- Τοπική μάθηση χρησιμοποιώντας τον DS με τα 50 κοντινότερα παραδείγματα (Local weighted DS)
- Bagging DS, Boosting DS, MultiBoost DS και DECORATE DS (με 25 υπο-ταξινομητές)

Σε όλα τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή- t ($p < 0.05$). Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 35 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επιπλέον στην τελευταία γραμμή του Πίνακα 35 παρουσιάζεται η μέση ακρίβεια κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Ο προτεινόμενος αλγόριθμος χρησιμοποιώντας σαν βασικό ταξινομητή τον DS είναι στατιστικά ακριβέστερος από τον ταξινομητή DS σε 18 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 35). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging DS σε 14 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Ομοίως, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον ταξινομητή 50NN σε 11 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 5 σύνολα δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Adaboost DS σε 9 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Multiboost DS σε 11 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων. Ακόμα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον DECORATE DS σε 11 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Τελειώνοντας με τις συγκρίσεις, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Local DS σε 3 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Τέλος, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον ταξινομητή Kstar σε 6 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων.

Συνοψίζοντας η απόδοση της παρουσιαζόμενης μεθόδου είναι ακριβέστερη από τις άλλες μεθόδους. Η προτεινόμενη μέθοδος επιτυγχάνει μια αύξηση στην ακρίβεια ταξινόμησης της τάξης του 17% σε σχέση με τον απλό αλγόριθμο DS. Παράλληλα, η μέση σχετική βελτίωση ακρίβειας της προτεινόμενης μεθοδολογίας είναι από 1.5% έως 12.5% σε σχέση με τις υπόλοιπες μεθόδους. Αυτό δείχνει ότι είναι δυνατό να πάρουμε μια εφικτή λύση στα

προβλήματα της ταξινόμησης με την τοπική προώθηση όταν η καθολική προσέγγιση της συνάρτησης στόχου είναι δύσκολη δεδομένου ότι δεν υπάρχουν αρκετά δεδομένα κατάρτισης διαθέσιμα.

Δεδομένα	Local DS	Boost DS	Kstar	Local DS	DS	50NN	Boost DS	Bagging DS	Multiboost DS	Decorate DS
audiology	73.12		80.32 v	72.68	46.46 *	35.95 *	46.46 *	46.46 *	46.46 *	46.46 *
autos	75.49		72.01	74.82	44.9 *	48.18 *	44.9 *	44.95 *	44.9 *	51.81 *
balance-scale	84.82		88.72 v	84.16	56.72 *	89.01 v	71.77 *	68.21 *	71.77 *	81.25 *
breast-cancer	72.75		73.73	72.68	69.27	70.75	71.55	73.38	71.76	75.18
breast-w	95.9		95.35	96.4	92.33 *	95.9	95.28	92.56 *	95.05	95.78
colic	78.91		75.71	80.87	81.52	84.04 v	82.72	81.52	82.9	82.03
credit-rating	84.94		79.1 *	83.61	85.51	86.16	85.57	85.51	85.39	85.28
diabetes	73.77		70.19 *	73.2	71.8	74.68	75.37	72.45	75.19	76.09
Glass	70.39		75.31	70.58	44.89 *	56.16 *	44.89 *	45.08 *	44.89 *	53.12 *
haberman	68.9		70.27	69.81	71.57	72.91	74.06 v	73.07	73.8 v	71.61
heart-c	80.4		75.18 *	78.29	72.93 *	81.58	83.11	75.26	83.54	72.43 *
heart-h	79.06		77.83	79.17	81.78	83.98 v	82.42	81.41	81.91	81.78
heart-statlog	78.15		76.44	76.33	72.3	83.74 v	81.81	75.33	82.89 v	81.48
hepatitis	84.45		80.17	83.04	77.62 *	79.38	81.5	80.61	82.21	80.02
ionosphere	90.01		84.64 *	88.24	82.57 *	71.65 *	92.34	82.66 *	90	90.4
Iris	94.47		94.67	94	66.67 *	90.53	95.07	68.87 *	94.73	93.93
Labor	90.47		92.03	85.3	78.77 *	64.67 *	90.57	81.97	89.97	91.07
lymphography	84.6		85.08	76.67 *	75.31 *	80.59	75.44 *	74.5 *	74.96 *	72.25 *
monk1	86.38		80.27	77.22 *	73.41 *	59.8 *	69.79 *	73.41 *	70.37 *	70.94 *
monk2	54.08		58.35	48.75	59.58	62.13 v	53.99	61.13	54.19	61.95
monk3	90.68		86.22	93.44	76.01 *	82.46 *	90.92	82.41	92.3	93.45
Primary-tumor	43.22		38.02	43.22	28.91 *	39.26	28.91 *	28.91 *	28.91 *	29.09 *
sonar	83.89		85.11	76.62 *	72.25 *	68.25 *	81.06	73.21 *	77.54 *	72.91 *
Titanic	79.05		77.56 *	79.05	77.6 *	77.56 *	77.83	77.6 *	77.62 *	77.6 *
Vehicle	70.98		70.22	69.58	39.81 *	63.47 *	39.81 *	40.14 *	39.81 *	47.35 *
Vote	96.02		93.22 *	95.4	95.63	90.41 *	96.41	95.63	95.63	95.59
Wine	97.47		98.72	96.79	57.91 *	96.46	91.57 *	86.27 *	91.22 *	96.45
			2/19/6	0/24/3	0/9/18	5/11/11	1/17/9	0/13/14	2/14/11	0/16/11
Μέση ακρίβεια	80,09		79,05	78,52	68,67	73,69	74,26	71,20	74,07	75,09

Πίνακας 35. Σύγκριση της προτεινόμενης μεθόδου τοπικής προώθησης χρησιμοποιώντας ως βασικό ταξινομητή τον DS με τους υπόλοιπους αλγορίθμους

Όπως προείπαμε, συγκρίναμε την προτεινόμενη μέθοδο με τους ίδους αλγορίθμους χρησιμοποιώντας αυτή τη φορά σαν βασικό ταξινομητή τον OneR. Ο προτεινόμενος αλγόριθμος χρησιμοποιώντας σαν βασικό ταξινομητή τον OneR είναι στατιστικά ακριβέστερος από τον ταξινομητή OneR σε 16 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων (βλέπε Πίνακα 36). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging OneR σε 14 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Ομοίως, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον ταξινομητή 50NN σε 11 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 5 σύνολα δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Adaboost OneR σε 8 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Multiboost OneR σε 10 σύνολα δεδομένων, ενώ έχει

στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Ακόμα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον DECORATE OneR σε 14 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Τελειώνοντας με τις συγκρίσεις, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Local OneR σε 3 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Τέλος, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον ταξινομητή Kstar σε 5 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων.

Δεδομένα	Local Boost OneR	Kstar	Local OneR	OneR	50NN	Boost OneR	Bagging OneR	Multiboost OneR	Decorate OneR
audiology	74.53	80.32 v	72.28	46.46 *	35.95 *	46.46 *	46.46 *	46.46 *	46.46 *
autos	80.87	72.01 *	76.49	61.77 *	48.18 *	65.47 *	65.26 *	66.06 *	70.51 *
balance-scale	88.3	88.72	87.36	57.09 *	89.01	72.81 *	67.95 *	72.84 *	60.94 *
breast-cancer	72.58	73.73	72.16	66.91 *	70.75	69.49	68.45	70	66.87 *
breast-w	96	95.35	96.3	92.01 *	95.9	95.55	92.81 *	95.45	92.38 *
colic	80.05	75.71	81.17	81.52	84.04	81.17	81.52	80.32	78.15
credit-rating	82.45	79.1	83.59	85.51	86.16 v	81.86	85.51	81.87	85.51
diabetes	70.2	70.19	69.76	71.98	74.68 v	69.91	72.08	69.95	71.98
Glass	71.7	75.31	68.52	56.84 *	56.16 *	56.56 *	60.85 *	56.81 *	58.76 *
haberman	65.73	70.27	68.39	72.53 v	72.91 v	71.65 v	72.19 v	71.74 v	71.97 v
heart-c	77.33	75.18	79.23	72.53	81.58	72.61	76.08	76.24	72.83
heart-h	78.04	77.83	79.14	80.69	83.98 v	76.59	80.49	77.67	80.36
heart-statlog	77.78	76.44	76.3	71.26	83.74 v	73.59	75.59	75.26	74.93
hepatitis	82.62	80.17	82.02	82.05	79.38	77.41	82.33	78.73	80.94
ionosphere	90.09	84.64 *	88.21	82.59 *	71.65 *	88.43	85.04 *	88.91	85.59
Iris	93.33	94.67	94	93.53	90.53	93.67	93.6	94.53	93.47
Labor	90.2	92.03	82.47 *	72.97 *	64.67 *	89.37	77.93 *	87.53	79.57 *
lymphography	84.63	85.08	78.31 *	74.77 *	80.59	79.83	74.57 *	79.47	70.66 *
monk1	85.35	80.27	80.78	73.41 *	59.8 *	70.35 *	73.41 *	70.44 *	74.35 *
monk2	62.36	58.35	56.74	58.35	62.13	55.36	58.76	54.73	57.22
monk3	90.07	86.22	92.15	77.88 *	82.46 *	90.67	77.57 *	91.82	93.45
Primary-tumor	44.04	38.02 *	44.57	27.74 *	39.26	27.38 *	27.05 *	27.25 *	27.88 *
sonar	77.85	85.11	73.57	62.12 *	68.25 *	65.43 *	70.59	65.69 *	62.44 *
Titanic	78.97	77.56 *	79.03	77.6 *	77.56 *	77.74	77.6 *	77.64 *	77.6 *
Vehicle	70.89	70.22	67.37 *	52.36 *	63.47 *	51.63 *	52.85 *	51.45 *	52.71 *
Vote	95.79	93.22 *	95.52	95.63	90.41 *	96.62	95.63	95.91	95.63
Wine	96.46	98.72	94.83	77.93 *	96.46	92.07	79.38 *	89.7 *	79.15 *
Μέση ακρίβεια	79,93	79,05	78,53	71,33	73,69	73,69	73,02	73,87	72,68

Πίνακας 36. Σύγκριση της προτεινόμενης μεθόδου τοπικής προώθησης χρησιμοποιώντας ως βασικό ταξινομητή τον OneR με τις υπόλοιπες μεθόδους

Συνοψίζοντας η απόδοση της παρουσιαζόμενης μεθόδου είναι ακριβέστερη από τις άλλες μεθόδους. Η προτεινόμενη μέθοδος επιτυγχάνει μια αύξηση στην ακρίβεια ταξινόμησης της τάξης του 12% σε σχέση με τον απλό αλγόριθμο OneR. Παράλληλα, η μέση σχετική βελτίωση ακρίβειας της προτεινόμενης μεθοδολογίας είναι από 1% έως 10% σε σχέση με τις υπόλοιπες μεθόδους.

Το μόνο αρνητικό της τοπικής προώθησης είναι ότι δύσκολα μπορεί να εκτελεστεί παράλληλα ώστε να μην απαιτεί πολύ χρόνο σε μεγάλα δεδομένα. Γι'αυτό το λόγο εξετάζουμε στην επόμενη υπο-ενότητα το τοπικό σακούλιασμα, που μπορεί ευκολότερα να

εκτελεστεί παράλληλα και μπορεί συνεπώς να είναι αποτελεσματικότερο σε μεγάλα σύνολα δεδομένων (datasets).

4.1.3 Προτεινόμενη μέθοδος τοπικού σακούλιασματος

Τα πρώτα πειράματα της έρευνας μας απέδειξαν ότι το τοπικό σακούλιασμα χρησιμοποιώντας σαν βασικό ταξινομητή τον DS δίνει καλύτερα αποτελέσματα από όλες τις μέχρι τώρα χρησιμοποιούμενες μεθόδους αύξησης της ακρίβειας του DS ταξινομητή. Παρακάτω ακολουθεί ο προτεινόμενος αλγόριθμος σε ψευδογλώσσα.

1. Καθόρισε την κατάλληλη μετρική απόστασης (distance metric).
2. Βρες τους k κοντινότερους γείτονες χρησιμοποιώντας αυτή την μετρική.
3. Εφάρμοσε την τεχνική σακούλιασματος στον ταξινομητή Decision Stump χρησιμοποιώντας τα k παραδείγματα ως δεδομένα εκπαίδευσης.
4. Η απάντηση της ομάδας των ταξινομητών που δημιουργήθηκαν από το σακούλιασμα είναι η κλάση του παραδείγματος που θέλουμε να ταξινομήσουμε.

Εικόνα 49. Ο προτεινόμενος αλγόριθμος τοπικού σακούλιασματος

Και αυτός ο προτεινόμενος αλγόριθμος έχει μερικές ελεύθερες παραμέτρους όπως η μετρική απόστασης. Στα πειράματά μας, χρησιμοποιήσαμε την πιο συχνά χρησιμοποιούμενη μετρική - την Ευκλείδεια απόσταση. Χρησιμοποιήσαμε επίσης το 50 ως την τιμή του k δεδομένου ότι αυτό το μέγεθος περιπτώσεων είναι κατάλληλο ώστε ένας απλός αλγόριθμος να χτίζει ένα ακριβές μοντέλο βάσει των πειραμάτων στην εργασία (Frank και άλλοι, 2003)

Συγκρίναμε την προτεινόμενη μέθοδο χρησιμοποιώντας σαν βασικό ταξινομητή τον DS με τις ακόλουθες τεχνικές:

- 50-κοντινότερους γείτονες (50-nearest neighbors) αφού η προτεινόμενη μέθοδος χρησιμοποιεί τα κοντινότερα 50 παραδείγματα. Επίσης, χρησιμοποιήσαμε έναν ακόμα τοπικό αλγόριθμο μάθησης τον Kstar.
- Τοπική μάθηση με βασικό ταξινομητή τον DS (Local weighted) DS χρησιμοποιώντας τα 50 κοντινότερα παραδείγματα
- Bagging DS, Boosting DS and LogitBoost DS (με 25 υπο-ταξινομητές)

Σε όλα τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή- t ($p < 0.05$). Συγκεντρωτικά, στην τελευταία γραμμή του Πίνακα 37 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Local DS σε 4 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 37). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον 50NN σε 9 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Kstar σε 17 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging DS σε 21 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Ο προτεινόμενος αλγόριθμος είναι ακόμα στατιστικά ακριβέστερος από τον Boosting DS σε 13

σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Τέλος, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Logitboost DS σε 4 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων.

Δεδομένα	Local Bagging DS	Local DS	50NN	Kstar	Bagging DS	Boosting DS	LogitBoost DS
anneal	99,48	99,34	91,18 *	95,69 *	82,96 *	83,63 *	99,29
autos	76,37	74,82	48,18 *	72,01	44,95 *	44,9 *	81,47
badges	99,93	100	99,69	90,27 *	100	100	100
breast-c	73,66	72,68	70,75	73,73	73,38	71,55	68,62 *
breast-w	96,45	96,4	95,9	95,35	92,56 *	95,28	95,94
colic	81,77	80,87	84,04	75,71 *	81,52	82,72	82,15
credit-a	85,01	83,61	86,16	79,1 *	85,51	85,57	86,32
Credit-g	73,68	71,02 *	71,96	70,17 *	70	72,6 *	74,84
diabetes	74,54	73,2	74,68	70,19 *	72,45	75,37	75,09
Glass	70,96	70,58	56,16 *	75,31	45,08 *	44,89 *	72,52
haberman	71,18	69,81	72,91	70,27	73,07	74,06	72,73
heart-c	80,97	78,29	81,58	75,18 *	75,26	83,11	79,98
heart-h	79,69	79,17	83,98 v	77,83	81,41	82,42	80,72
heart-statlog	78,07	76,33	83,74 v	76,44	75,33	81,81	81,63
hepatitis	84,59	83,04	79,38 *	80,17	80,61	81,5	82,49
ionosphere	88,89	88,24	71,65 *	84,64 *	82,66 *	92,34 v	92,19
iris	93,87	94	90,53	94,67	68,87 *	95,07	93,53
kr-vs-kp	98,51	98,45	91,07 *	96,91 *	66,05 *	95,08 *	94,69 *
labor	87,57	85,3	64,67 *	92,03	81,97	90,57	91,37
lymphography	80,78	76,67	80,59	85,08	74,5	75,44	83,33
monk1	79,8	77,22	59,8 *	80,27	73,41 *	69,79 *	71,85 *
monk3	93,45	93,44	82,46 *	86,22 *	82,41 *	90,92	92,47
Nursery	97,75	97,52 *	96,05 *	96,88 *	66,25 *	66,25 *	91,6 *
primary-tumor	43,98	43,22	39,26	38,02 *	28,91 *	28,91 *	45,67
segment	96,97	96,68	90,43 *	97,09	56,54 *	28,52 *	97,16
sick	97,46	97,64	94,84 *	95,72 *	96,55 *	97,07	97,85
sonar	82,21	76,62 *	68,25 *	85,11	73,21 *	81,06	81,45
soybean	93,05	92,56	62,34 *	87,97 *	27,83 *	27,96 *	93,5
splice	91,65	89,6 *	88,89 *	78,84 *	62,38 *	86,24 *	95,81 v
titanci	78,99	79,05	77,56 *	77,56 *	77,6 *	77,83	77,83
vehicle	71,47	69,58	63,47 *	70,22	40,14 *	39,81 *	74,36
vote	95,72	95,4	90,41 *	93,22 *	95,63	96,41	96,39
wine	97,8	96,79	96,46	98,72	86,27 *	91,57 *	97,4
zoo	90,11	88,84	55,11 *	96,03 v	60,53 *	60,43 *	94,09
		0/30/4	2/13/19	1/16/17	0/13/21	1/20/13	1/29/4

Πίνακας 37. Σύγκριση της προτεινόμενης μεθόδου τοπικού σακουλίσματος με άλλες μεθόδους

Όπως φαίνεται από τα πειράματα μας το τοπικό σακούλισμα αποδίδει καλύτερα από τις άλλες μεθόδους. Δεν κατορθώνει βέβαια να ξεπεράσει την τοπική προώθηση (που προτάθηκε στο προηγούμενο υποκεφάλαιο) αλλά μπορεί να εκτελεστεί παράλληλα (ένας υπο-ταξινομητής ανά μηχανήμα) και κατ' αυτόν τον τρόπο μπορεί να είναι μια αποτελεσματικότερη τεχνική από άποψη χρόνου ταξινόμησης σε μεγάλα σύνολα δεδομένων.

4.1.4 Προτεινόμενη μέθοδος προώθησης επιλεκτικών απλών Μπεϋζιανών ταξινομητών

Εν συνεχεία συνδυάσαμε την μέθοδο της προώθησης (boosting) με τον απλό Μπεϋζιανό αλγόριθμο (NB). Για να γίνει πιο ασταθής αλλά να μην χάσει ακρίβεια ο NB ώστε να έχει αποτέλεσμα η προώθηση (boosting), σε κάθε βήμα της προώθησης εφαρμόσαμε ένα φίλτρο για την επιλογή των πιο χρήσιμων μεταβλητών. Τα πειράματα απέδειξαν ότι η προτεινόμενη μέθοδος αυξάνει την ακρίβεια πρόγνωσης του NB περισσότερο από κάθε άλλη προσπάθεια βελτίωσής του.

Ξεκινώντας την ανασκόπηση των προσπαθειών για βελτίωση της ακρίβειας των απλών Μπεϋζιανών ταξινομητών πρέπει να σταθούμε στην διακριτοποίηση (discretization), την πιο γνωστή και απλή μέθοδο για τη βελτίωση της απόδοσης. Η διακριτοποίηση των συνεχών μεταβλητών, αντί της χρησιμοποίησης της κανονικής κατανομής από τον τύπο του Bayes για να υπολογισθούν οι πιθανότητες για τις συνεχείς μεταβλητές, δίνει καλύτερα αποτελέσματα βάσει των πειραμάτων των (Dougherty και άλλοι, 1995).

Πολυάριθμες μέθοδοι διακριτοποίησης (discretization) έχουν εξεταστεί όπως η κατάτμηση του διαστήματος τιμών μιας συνεχούς μεταβλητής σε ίσου μεγέθους K διαστήματα και η κατάτμηση του διαστήματος τιμών μιας συνεχούς μεταβλητής σε διαστήματα που περιέχουν ίδιο αριθμό περιπτώσεων. Μια συνοπτική αναφορά για τις μεθόδους διακριτοποίησης και μια εμπειρική μελέτη που συγκρίνει την απόδοση τους δίνεται στην εργασία (Dougherty και άλλοι, 1995). Η ίδια εργασία κατέληξε στο συμπέρασμα ότι όλες οι δοκιμασμένες μέθοδοι διακριτοποίησης μπορούν να ξεπεράσουν την απόδοση του απλού Μπεϋζιανού ταξινομητή που υποθέτει την κανονική κατανομή για τις συνεχείς μεταβλητές. Οι ίδιοι ερευνητές επίσης κατέληξαν στο συμπέρασμα ότι η μέθοδος που βασίζεται στην εντροπία (Fayyad & Irani, 1993) έδωσε λίγο καλύτερα αποτελέσματα από τις υπόλοιπες. Η μέθοδος που βασίζεται στην εντροπία είναι μια εποπτευμένη επαυξητική μέθοδος που επιλέγει κατ' επανάληψη τα σημεία περικοπής του διαστήματος τιμών μιας συνεχούς μεταβλητής ώστε να ελαχιστοποιείται η εντροπία έως ότου ένα κριτήριο τερματισμού ικανοποιηθεί.

Η απόδοση του απλού Μπεϋζιανού ταξινομητή σε δεδομένα με περιττές μεταβλητές εισόδου μπορεί να βελτιωθεί επίσης με την αφαίρεση αυτών των περιττών χαρακτηριστικών. Ο επονομαζόμενος "επιλεκτικός Μπεϋζιανός ταξινομητής", έχει αποτελέσει αντικείμενο έρευνας πολλών ερευνητών όπως (Langley & Sage, 1994) και (Pazzani, 1995). Ένα γνωστό φίλτρο που υπολογίζει το πληροφοριακό κέρδος (information gain) μεταξύ των μεταβλητών εισόδου και της μεταβλητής εξόδου, και απορρίπτει τις μικρής αξίας (πληροφορίας) μεταβλητές χρησιμοποιήθηκε από τους (Singh & Provan, 1996).

Σε αντίθεση με τη χρήση φίλτρων για την επιλογή των καταλληλότερων μεταβλητών, μια στρατηγική περιτυλίγματος προς τα εμπρός για να επιλεγθούν τα χαρακτηριστικά που θα χρησιμοποιηθούν από τον NB περιγράφηκε στην εργασία (Langley & Sage, 1994). Επίσης, ο Pazzani (1997) χρησιμοποίησε την εποικοδομητική δημιουργία και διαγραφή χαρακτηριστικών (αλγόριθμος BSEJ) για να μειώσει το πρόβλημα της αλληλεξάρτησης των χαρακτηριστικών στον απλό Μπεϋζιανό ταξινομητή. Ο αλγόριθμος BSEJ χρησιμοποιεί την μέθοδο περιτυλίγματος για να ενώσει και να διαγράψει μεταβλητές. Με άπληστη αναζήτηση, σε κάθε βήμα, είτε διαγράφει ένα χαρακτηριστικό είτε δημιουργεί ένα νέο μέσω της ένωσης (που παράγει το καρτεσιανό γινόμενο) δύο χαρακτηριστικών. Άλλοι ερευνητές (Kohavi & John, 1997) χρησιμοποίησαν την βέλτιστη προς τα πίσω μέθοδο επιλογής (best first backward selection) ως μέθοδο αναζήτησης για την προσέγγιση περιτυλίγματος. Έδειξαν ότι αυτός ο συνδυασμός δίνει πολύ καλά αποτελέσματα για τον NB.

Πιο πρόσφατες απόπειρες να υπερνικηθεί η υπόθεση της ανεξαρτησίας μεταξύ των μεταβλητών είναι βασισμένες στη προσθήκη νέων όρων που θα περιγράφουν τη σχέση μεταξύ των χαρακτηριστικών (Friedman και άλλοι, 1997). Οι απόπειρες αυτές είναι κυρίως βασισμένες στην προσθήκη επιπλέον ακμών στο δίκτυο (ερχόμενες έτσι πιο κοντά στα

Μπεϋζιανά δίκτυα) για να συμπεριληφθούν μερικές εξαρτήσεις μεταξύ των μεταβλητών. Στην εργασία για τον αλγόριθμο TAN (Friedman και άλλοι, 1997) κάθε χαρακτηριστικό σχετίζεται το πολύ με ένα ακόμα χαρακτηριστικό πλην της κλάσης.

Μια άλλη προσπάθεια για τη βελτίωση του NB ήταν τα Μπεϋζιανά δέντρα (Zheng & Webb, 2000). Ένας Μπεϋζιανός αλγόριθμος μάθησης δέντρο απόφασης χτίζει ένα δέντρο απόφασης, και παράγει έναν τοπικό απλό Μπεϋζιανό ταξινομητή σε κάθε φύλλο. Οι έλεγχοι, που οδηγούν σε ένα φύλλο, μπορούν να περιορίσουν τις αλληλεξαρτήσεις των χαρακτηριστικών για τον τοπικό απλό Μπεϋζιανό ταξινομητή. Οι ίδιοι ερευνητές πρότειναν επίσης την εφαρμογή των οκνηρών τεχνικών μάθησης στην Μπεϋζιανή επαγωγή δέντρων απόφασης και παρουσίασαν τον οκνηρό Μπεϋζιανό αλγόριθμο μάθησης δέντρων, αποκαλούμενο LBR. Για κάθε παράδειγμα δοκιμής, χτίζεται ένας κατάλληλος κανόνας δέντρου απόφασης.

Όπως προείπαμε στον προτεινόμενο αλγόριθμό έχουμε επιλέξει να ενσωματώσουμε ένα φίλτρο επιλογής μεταβλητών για να γίνει πιο ασταθής ώστε να έχει καλύτερο αποτέλεσμα η προώθηση (boosting). Το γνωστό φίλτρο που υπολογίζει το πληροφοριακό κέρδος (information gain) μεταξύ των χαρακτηριστικών και της κατηγορίας (Singh & Provan, 1996) και απορρίπτει τα μικρής αξίας χαρακτηριστικά τροποποιήθηκε για τον αλγόριθμό μας. Το κριτήριο του κέρδους πληροφορίας (gain) έχει μια ισχυρή προκατάληψη υπέρ των χαρακτηριστικών με πολλές διαφορετικές τιμές (Zaffalon & Hutter, 2002). Αποκαθιστάμε αυτήν την προκατάληψη με ένα είδος κανονικοποίησης (gain ratio), που μερικές φορές χρησιμοποιείται και από τους αλγορίθμους δέντρων απόφασης (Quinlan, 1993). Τα χαρακτηριστικά επιλέγονται με την κράτηση εκείνων για τα οποία η αναλογία κέρδους (gain ratio) υπερβαίνει ένα σταθερό κατώτατο όριο ϵ . Θέσαμε το ϵ ίσον με 0.02, σε μία προσπάθεια να απορριφθούν μόνο τα χαρακτηριστικά με αμελητέο αντίκτυπο στις προβλέψεις. Εντούτοις, ακόμα και ένα τόσο χαμηλό κατώτατο όριο μπορεί να απορρίψει πολλά χαρακτηριστικά.

Οι ομάδες απλών Μπεϋζιανών ταξινομητών παραδοσιακά δεν έχουν αποτελέσει αντικείμενο εκτενής έρευνας. Ο λόγος είναι ότι ο απλός Μπεϋζιανός ταξινομητής είναι ένας εξαιρετικά σταθερός αλγόριθμος μάθησης, και η τοποθέτηση σε σάκο (bagging) είναι κυρίως τεχνική μείωσης διασποράς, και κατά συνέπεια δεν μπορεί να ωφελήσει στην περίπτωση του απλού Μπεϋζιανού ταξινομητή. Εντούτοις, ο απλός Μπεϋζιανός ταξινομητής μπορεί να χρησιμοποιηθεί αποτελεσματικότερα με τεχνικές, που εκτελούν επίσης μείωση της προκατάληψης (bias), όπως η προώθηση (boosting). Οι ερευνητές (Bauer & Kohavi, 1999) αναφέρουν ότι υπάρχει ένα πρόβλημα με την προώθηση που είναι η μη ανεκτικότητα στο θόρυβο. Αυτό είναι αναμενόμενο επειδή τα θορυβώδη παραδείγματα τείνουν να κατηγοριοποιηθούν λάθος, και κατά συνέπεια το βάρος τους θα αυξηθεί. Για αυτόν τον λόγο, επιλέξαμε να χρησιμοποιήσουμε τον αλγόριθμο Adaboost με μόνο 10 επαναλήψεις.

Συνολικά, ο προτεινόμενος αλγόριθμος (BoostFSNB) χρησιμοποιεί τον αλγόριθμο Adaboost, με τη διαφορά ότι σε κάθε επανάληψη του αλγόριθμου Adaboost, χρησιμοποιείται η μέθοδος της εντροπίας για διακριτοποίηση (discretization) των συνεχών μεταβλητών και αφαιρούνται τα περιττά χαρακτηριστικά χρησιμοποιώντας την αναλογία κέρδους ως φίλτρο επιλογής μεταβλητών.

Στην Εικόνα 50 παρουσιάζεται ο προτεινόμενος αλγόριθμος σε ψευδογλώσσα [2]:

```

Θέσε βάρος=1 σε κάθε δεδομένο εκπαίδευσης
Για i=1 έως 10
{ Για i=1 έως τον αριθμό του πλήθους των μεταβλητών
  {
    S= Ταξινομημένες τιμές της i μεταβλητής;
    Χώρισε(S)
    Αν το Entropy-MDL κριτήριο ικανοποιείται τότε σταμάτα
    αλλιώς {
      T = ΠάρεΤοκαλύτεροΣημείοΚοπής(S); S1 = ΠάρεΤοΑριστερόΜέρος(S,T);
      S2 = ΠάρεΤοΔεξίΜέρος(S,T);
      Χώρισε(S1)
      Χώρισε(S2) }
    }
  Για i=1 έως τον αριθμό του πλήθους των μεταβλητών
  { G= Gain Ratio της i μεταβλητής
    Αν (G<0.02) τότε διώξε την i μεταβλητή από τα δεδομένα
  }
Εφάρμοσε τον αλγόριθμο μάθησης στα δεδομένα εκπαίδευσης και αποθήκευσε το μοντέλο.
Υπολόγισε το σφάλμα του μοντέλου e και αποθήκευσε το.
Αν το e ισούται με 0 ή είναι μεγαλύτερο ή ίσο του 0.5:
  Τερμάτισε τις επαναλήψεις.
Για κάθε παράδειγμα των δεδομένων:
  Αν το παράδειγμα ταξινομήθηκε σωστά από το μοντέλο:
    Πολλαπλασίασε το βάρος του παραδείγματος με e / (1 - e).
  Ομαλοποίησε (normalize) τα βάρη όλων των παραδειγμάτων.}
}

Ταξινόμηση:
Θέσε το βάρος της κάθε κλάσης 0
Για καθένα από τα t (ή και λιγότερα) μοντέλα:
  Πρόσθεσε -log(e / (1 - e)) στο βάρος της κάθε κλάσης που προβλέφθηκε από το μοντέλο.
Επέστρεψε την κλάση με το μεγαλύτερο βάρος.

```

Εικόνα 50. Ο προτεινόμενος αλγόριθμος BoostFSNB

Εν συνεχεία, συγκρίναμε την προτεινόμενη μέθοδο με τους ακόλουθους αλγόριθμους:

- Αφελή Μπεϋζιανό ταξινομητή (NB).
- Τον αφελή ταξινομητή Μπέιζ με χρήση διακριτοποίησης (DiscNB)
- Τον επιλεκτικό ταξινομητή Μπέιζ (WrapperNB) (Langley & Sage, 1994)
- Τον αλγόριθμο LBR
- Και τους γνωστούς αλγόριθμους άλλων μεθόδων μάθησης: BP, SMO και 3NN

Σε όλα τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 38 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του

Πίνακα 38 παρουσιάζεται το μέσο σφάλμα κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον NB σε 12 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 5 σύνολα δεδομένων (βλέπε Πίνακα 38). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους DiscNB και WrapperNB σε 13 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 4 σύνολα δεδομένων. Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους TAN και BP σε 9 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 7 σύνολα δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους SMO και 3NN σε 8 και 15 σύνολα δεδομένων, αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 7 και 6 σύνολα δεδομένων, αντίστοιχα. Τέλος, ο προτεινόμενος αλγόριθμος έχει παραπλήσια απόδοση με τον LBR αλλά χρησιμοποιεί πολύ λιγότερο χρόνο για εκπαίδευση και ταξινόμηση.

Δεδομένα	Boost FSNB	NB	DiscNB	Wrapper-NB	TAN	LBR	BP	SMO	3NN
anneal	99.31	86.59*	95.9 *	89.57 *	98.11*	98.01 *	94.97 *	96.89 *	97.29 *
autos	75.6	57.41 *	65.17 *	63.61 *	74.78	73.8	48.84 *	56.55 *	67.23 *
balance	81.9	90.53 v	71.56 *	90.53 v	70.8*	72.17 *	85.67 v	87.62 v	86.74 v
breast-cancer	72.17	72.7	72.7	71.64	69.28*	72.35	72.95	69.92	73.13
breast-w	95.98	96.07	97.2 v	95.5 *	97 v	97.21 v	96.35	96.81 v	96.61 v
colic	82.63	78.7 *	79.54*	83.91 v	82.8	82.33	83.07	82.69	80.95 *
credit-a	85.9	77.86 *	86.22	85.3	86.07	86.1	85.94	84.91	84.96 *
credit-g	72.83	75.16 v	75.04 v	73.79 v	74.3 v	74.9 v	74.86 v	75.11 v	72.21
diabetes	75.38	75.75	75.26	76.06	75.36	75.38	77.04 v	77.07 v	73.86 *
heart-c	82.42	83.34	83.47 v	78.75 *	82.97	83.54 v	82.98	84.03	81.82
heart-h	83.51	83.95	84.2	82.63	84.84 v	84.54 v	84.16	83.26	82.33 *
heart-statlog	82.19	83.59 v	82.56	82.78	82.63	82.59	83.3 v	83.81	79.11 *
hepatitis	84.14	83.81	84.28	82.55	84.83	84.91	84.29	85.03	80.85 *
hypothyroid	98.72	95.3 *	98.19 *	95.19 *	99.16 v	99.12 v	93.44 *	93.49 *	93.21 *
ionosphere	92.14	82.17 *	89.29 *	90.14 *	91.03 *	90 *	87.07 *	87.93 *	86.02 *
iris	94.2	95.53 v	93.33	94.27	92.93 *	93.2 *	84.8 *	84.87 *	95.2 v
kr-vs-kp	94.58	87.79 *	87.79 *	94.34	92.28 *	96.79 v	98.92 v	95.78 v	96.56 v
labor	91.33	93.57	88.57 *	85 *	92.1	87.5 *	88.93	94	87.83 *
Lymph/rapy	83.69	83.13	85.1	81.22 *	86.3 v	85.45	82.26	85.94	81.74
mushroom	100	95.76 *	95.76 *	99.62 *	99.96 *	99.96 *	99.97 *	100	100
prim.-tumor	46.49	49.71 v	49.71 v	43.45 *	47.76 v	48.85 v	24.9 *	29.2 *	44.98 *
segment	93.08	80.16 *	91.16 *	89.17 *	95.16 v	93.94 v	91.35 *	89.72 *	96.12 v
sonar	78.19	67.71 *	76.23	71.25 *	76.41 *	76.04*	78.67	77.88	83.76 v
vote	95.08	90.02 *	90.02 *	95.63	94.62	94.11*	96.32 v	96.23 v	93.08 *
waveform	81.61	80.01 *	79.97 *	81.94 v	80.1 *	83.42 v	86.56 v	86.94 v	77.67 *
zoo	95.64	94.97	93.21 *	92.67 *	95.45	93.21 *	60.43 *	93.16 *	92.61 *
		5/9/12	4/9/13	4/9/13	7/10/9	9/8/9	7/10/9	7/11/8	6/5/15
Μέσο σφάλμα	14.66	17.62	16.32	16.52	16.11	15.02	18.15	16.2	15.73

Πίνακας 38. Σύγκριση της προτεινόμενης μεθόδου (BoostFSNB) με άλλες μεθόδους

Συνοπτικά, κατορθώσαμε να βελτιώσουμε την απόδοση του απλού Μπεϋζιανού ταξινομητή λαμβάνοντας καλύτερη ακρίβεια από άλλες γνωστές μεθόδους που έχουν προσπαθήσει να βελτιώσουν την απόδοση του απλού Μπεϋζιανού αλγορίθμου. Ο προτεινόμενος αλγόριθμος έδωσε επίσης καλύτερη ακρίβεια από άλλους πιο περίπλοκους αλγορίθμους μηχανικής μάθησης στα περισσότερα από τα σύνολα δεδομένων που χρησιμοποιήσαμε για την συγκριτική μελέτη. Επιπλέον, ο BoostFSNB χρειάστηκε 31 φορές λιγότερο χρόνο για να

εκπαιδεύσει (κατά μέσο όρο) στα δοκιμασμένα σύνολα δεδομένων από τον BP και 2.5 φορές λιγότερο χρόνο από τον SMO.

Ακολουθώς επειδή η προτεινόμενη τεχνική είναι συνδυασμός ταξινομητών, την συγκρίναμε με τους ακόλουθους αλγόριθμους μάθησης χρησιμοποιώντας και με αυτούς 10 υπο-ταξινομητές:

- Bagging NB, Boosting NB
- Bagging C4.5, Boosting C4.5

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging NB σε 12 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 5 σύνολα δεδομένων (βλέπε Πίνακα 39). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Boosting NB σε 10 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 8 σύνολα δεδομένων.

Δεδομένα	BoostFSNB	AdaboostC4.5	BaggingC4.5	BaggingNB	AdaboostNB
anneal	99.31	99.59 v	98.82 *	87.27 *	95.2 *
autos	75.6	85.46 v	81.48 v	57.71 *	57.12 *
balance	81.9	78.35 *	82.68	90.29 v	92.11 v
breast-cancer	72.17	66.89 *	73.1	72.73	68.57 *
breast-w	95.98	96.08	96.18	96.07	95.55 *
colic	82.63	81.63	84.99 v	78.94 *	77.46 *
credit-a	85.9	84.01 *	86.38	77.81 *	81.16 *
credit-g	72.83	70.75 *	73.38	75.13 v	75.09 v
diabetes	75.38	71.69 *	75.23	75.57	75.88
heart-c	82.42	78.79 *	79.86 *	83.24	83.14
heart-h	83.51	78.68 *	80.24 *	84.16	84.67 v
heart-statlog	82.19	78.59 *	80.93	83.41 v	82.3
hepatitis	84.14	82.38 *	82.05 *	84.39	84.23
hypothyroid	98.72	99.65 v	99.57 v	95.53 *	95.27 *
ionosphere	92.14	93.05 v	91.63	81.94 *	91.12
iris	94.2	94.33	94.73	95.53 v	95.07 v
kr-vs-kp	94.58	99.59 v	99.43 v	87.8 *	95.1 v
labor	91.33	87.17 *	83.67 *	93.4	88.93
lymphotherapy	83.69	80.87 *	78.44 *	83.5	80.67 *
mushroom	100	100	100	95.73 *	100
primary-tumor	46.49	41.65 *	42.81 *	49.47 v	49.71 v
segment	93.08	98.12 v	97.25 v	80.41 *	80.16 *
sonar	78.19	79.13	77.25	67.96 *	81.21 v
vote	95.08	95.51	96.25 v	90.09 *	95.19
waveform	81.61	81.4	81.63	80 *	80.01 *
zoo	95.64	95.18	93.7 *	95.07	97.23 v
<i>Μέσο σφάλμα</i>	14.66	6/8/12 15.44	6/14/8 14.94	5/9/12 17.57	8/8/10 16.07

Πίνακας 39. Σύγκριση της προτεινόμενης μεθόδου (BoostFSNB) με άλλες ομάδες ταξινομητών

Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging C4.5 σε 8 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 6 σύνολα δεδομένων. Παράλληλα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Boosting C4.5 σε 12 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε

6 σύνολα δεδομένων. Συνοπτικά, η προτεινόμενη μέθοδος αποδίδει καλύτερα από τις ομάδες ταξινομητών που συγκρίναμε.

Επειδή ο προτεινόμενος αλγόριθμος είναι ένας συνδυασμός ταξινομητών, ο παραχθείς ταξινομητής είναι λιγότερο ερμηνεύσιμος από τον απλό ταξινομητή Μπέιζ. Εντούτοις, ο αλγόριθμος BoostFSNB ως μια ομάδα απλών ταξινομητών Μπέιζ, παραμένει πολύ περισσότερο ερμηνεύσιμος από τα νευρωνικά δίκτυα και τους ταξινομητές διανυσμάτων υποστήριξης (Ridgeway και άλλοι, 1998). Και άλλοι ερευνητές (Ting & Zheng, 1999) έδειξαν ότι η προώθηση βελτιώνει την ακρίβεια του απλού Μπεϋζιανού ταξινομητή σε 9 από τα 14 σύνολα δεδομένων που χρησιμοποίησαν. Όμως, κατέληξαν στο συμπέρασμα ότι η μέση σχετική μείωση του λάθους λόγω της ώθησης σε σχέση με τον απλό Μπεϋζιανό ταξινομητή στα 14 σύνολα δεδομένων ήταν μόνο 1%, δείχνοντας έτσι μια πολύ οριακή βελτίωση. Αντίθετα, ο προτεινόμενος αλγόριθμος έχει μέση σχετική μείωση λάθους 20% σε σχέση με τον απλό Μπεϋζιανό ταξινομητή στα δοκιμασμένα σύνολα δεδομένων. Ο κύριος λόγος είναι ότι η ενσωματωμένη τεχνική επιλογής χαρακτηριστικών καθιστά τον απλό ταξινομητή Μπέιζ ελαφρώς πιο ασταθή και κατά συνέπεια καταλληλότερο για την τεχνική προώθησης.

4.1.5 Προτεινόμενη μέθοδος συνδυασμού της λογιστικής προώθησης με τον απλό μπεϋζιανό αλγόριθμο

Αποδείξαμε στην προηγούμενη ενότητα με τα πειράματά μας ότι ο απλός Μπεϋζιανός ταξινομητής μπορεί να χρησιμοποιηθεί αποτελεσματικά με τεχνικές, που εκτελούν μείωση της προκατάληψης (bias), όπως ο Adaboost. Ένας αλγόριθμος ακόμα πιο δραστηκός στην μείωση της προκατάληψης είναι αυτός της λογιστικής προώθησης (Logitboost). Όμως, όπως είναι ευρέως γνωστό, ο Logitboost απαιτεί έναν αλγόριθμο παλινδρόμησης ως βασικό ταξινομητή. Για αυτόν τον λόγο, τροποποιήσαμε ελαφρώς τον απλό Μπεϋζιανό ταξινομητή προκειμένου να είναι σε θέση να λειτουργήσει ως μέθοδος παλινδρόμησης.

Ο αφελής ταξινομητής Μπέιζ πρωτοεφαρμόστηκε σε προβλήματα παλινδρόμησης στην εργασία (Frank και άλλοι, 2000) χωρίς καλά αποτελέσματα. Στην εργασία μας [3] διακριτοποιήσαμε την μεταβλητή στόχο χρησιμοποιώντας την τεχνική ίσου πλάτους. Συγκεκριμένα, χωρίσαμε την μεταβλητή στόχο σε δέκα διαστήματα ίσου-πλάτους, και εφαρμόζουμε τον αφελή ταξινομητή Bayes για την ταξινόμηση της δεκάτιμης πλέον μεταβλητής στόχου. Για την τελική πρόβλεψη, η προβλεφθείσα τιμή είναι η αναμενόμενη τιμή της κατηγορίας για κάθε διάστημα πολλαπλασιασμένη επί τις προβλεφθείσες πιθανότητες από τον Naïve Bayes για κάθε διάστημα.

Συνοπτικά, στην Εικόνα 51 παρουσιάζεται ο προτεινόμενος αλγόριθμος σε ψευδογλώσσα.

Βήμα 1: Αρχικοποίηση

συνάρτηση επιτροπής: $F^{(0)}(x) \equiv 0$
αρχική πιθανότητα: $p^{(0)}(x) \equiv 1/2$

Βήμα 2: LogitBoost επαναλήψεις

Για $m=1,2,\dots,10$ επανέλαβε:

A. Ταίριαζε τον ταξινομητή

Υπολόγισε τις απαντήσεις και τα βάρη για $i=1,\dots,m$

$$w_i^{(m)} = p^{(m-1)} \cdot (1 - p^{(m-1)})$$
$$z_i^{(m)} = \frac{y_i - p^{(m-1)}(x_i)}{w_i^{(m)}}$$

1. Διακριτοποίησε την τιμή της μεταβλητής στόχου με την τεχνική 10 διαστημάτων ίσου πλάτους (10 equal-width intervals)
2. Ταίριαζε την τροποποιημένη έκδοση του NB με τη μέθοδο ελαχίστων τετραγώνων (weighted least squares)

$$f^{(m)} = \arg \min_f \sum_{i=1}^n w_i^{(m)} (z_i^{(m)} - f(x_i))^2$$

B. Ανανέωσε την απάντηση του ταξινομητή

$$F^{(m)}(x_i) = F^{(m-1)}(x_i) + \frac{1}{2} f^{(m)}(x_i)$$

$$C^{(m)}(x_i) = \text{Sign}(F^{(m)}(x_i))$$

$$p^{(m)}(x_i) = \frac{1}{1 + e^{-2F^{(m)}(x_i)}}$$

Εικόνα 51. Ο προτεινόμενος αλγόριθμος LogitboostNB

Στα πειράματα, συγκρίναμε την αποτελεσματικότητα της προτεινόμενης μεθόδου με τις ακόλουθες μπεϋζιανές τεχνικές:

- Αφελή Μπεϋζιανό ταξινομητή (NB).
- AODE (Webb και άλλοι, 2004) που υπολογίζει τον μέσο όρο όλων των μοντέλων από μια περιορισμένη κατηγορία ταξινομητών που επιτρέπουν την εξάρτηση από μόνο μια μεταβλητή της κλάσης.

Επίσης συγκρίναμε τον προτεινόμενο αλγόριθμο με τον αλγόριθμο κατασκευής δικτύου Μπέϋζ K2 (Cooper & Herskovits, 1992) που είναι μια επέκταση του απλού Μπεϋζιανού ταξινομητή. Ο K2 αρχικά υποθέτει ότι κάθε κόμβος δεν έχει κανέναν γονέα, και έπειτα προσθέτει επαυξητικά εκείνο τον γονέα του οποίου η προσθήκη αυξάνει περισσότερο την ακρίβεια του προκύπτοντος δικτύου. Οι γονείς προστίθενται ένας-ένας σε κάθε κόμβο έως ότου η προσθήκη κάποιου γονέα δεν αυξάνει πλέον την ακρίβεια του δικτύου.

Σε όλα τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 40 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του

Πίνακα 40 παρουσιάζεται η μέση ακρίβεια κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Βάσει των πειραμάτων μας (βλέπε Πίνακα 40) ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον NB σε 6 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον K2 σε 6 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον AODE σε 2 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων.

Δεδομένα	LogitboostNB	NB	Bayesian Network (K2)		AODE
Autos	75.40	57.41	*	67.26	74.76
Badges	99.80	99.66		100.00	100.00
balance-scale	91.19	90.53		71.56	69.96
breast-cancer	66.67	72.7		72.59	73.05
breast-w	96.18	96.07		97.20	97.05
Colic	80.74	78.7		80.98	82.45
credit-g	72.73	75.16		74.97	75.83
diabetes	74.11	75.75		75.25	75.70
haberman	71.48	75.06		71.57	71.57
heart-c	78.15	83.34		83.34	82.87
Heart-h	79.55	83.95		84.57	84.33
heart-statlog	79.93	83.59		82.56	82.70
hepatitis	84.67	83.81		84.18	85.36
ionosphere	92.71	82.17	*	89.54	91.09
iris	94.87	95.53		93.20	93.07
labor	93.23	93.57		90.60	88.43
lymphography	84.65	83.13		85.64	86.86
monk1	85.33	73.38	*	73.46	82.32
monk2	59.92	56.83		56.78	59.62
monk3	91.87	93.45		93.45	93.21
relation	77.99	77.85		77.86	78.21
sonar	84.41	67.71	*	76.71	77.05
students	83.04	85.70		85.76	86.08
vehicle	70.91	44.68	*	61.05	70.32
vote	95.20	90.02	*	90.23	94.28
wine	98.14	97.46		98.65	98.21
zoo	96.91	94.97		94.37	94.66
		0/21/6		0/21/6	1/23/2
Μέση ακρίβεια	83.70	81.19		81.98	83.30

Πίνακας 40. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλες μπεϋζιανές τεχνικές

Επίσης, συγκρίναμε την προτεινόμενη μέθοδο με τις ακόλουθες γνωστές ομάδες ταξινομητών που βασίζονται στον απλό μπεϋζιανό ταξινομητή (χρησιμοποιώντας 25-υποταξινομητές):

- Adaboost NB
- Bagging NB
- DECORATE NB

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Adaboost NB σε 3 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 41). Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging NB σε 6 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον DECORATE NB σε 7 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων.

Δεδομένα	LogitboostNB	Adaboost NB	Bagging NB	DECORATE NB
autos	75.40	57.12 *	57.12 *	57.82 *
badges	99.80	99.66	99.69	96.73 *
balance-scale	91.19	91.68	90.29	90.55
breast-cancer	66.67	68.68	73.12 v	72.97 v
breast-w	96.18	95.55	96.04	95.97
colic	80.74	77.62	78.73	78.05
credit-g	72.73	75.14	75.20	74.72
diabetes	74.11	75.86	75.64	75.33
haberman	71.48	73.91	74.90	74.83
heart-c	78.15	82.97	83.37	83.51 v
Heart-h	79.55	84.81	84.13	83.98
heart-statlog	79.93	82.59	83.59	83.74
hepatitis	84.67	84.62	84.13	82.99
ionosphere	92.71	91.06	82.00 *	83.08 *
iris	94.87	94.80	95.53	94.87
Labor	93.23	89.60	93.73	92.87
lymphography	84.65	83.76	81.27	82.98
monk1	85.33	72.68 *	73.22 *	75.90 *
monk2	59.92	56.83	56.56	57.08
monk3	91.87	90.90	93.37	93.29
relation	77.99	77.86	77.88	78.31
sonar	84.41	80.77	68.21 *	67.65 *
students	83.04	85.18	85.73	85.09
vehicle	70.91	44.68 *	45.58 *	46.77 *
vote	95.20	95.01	90.02 *	89.93 *
wine	98.14	96.18	97.36	96.51
zoo	96.91	97.23	95.07	94.68
		0/24/3	1/20/6	2/18/7
Μέση ακρίβεια	83.70	81.73	81.17	81.12

Πίνακας 41. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλες κοινές ομάδες ταξινομητών που χρησιμοποιούν τον NB ως βασικό αλγόριθμο

Επιπλέον, συγκρίναμε την προτεινόμενη μέθοδο με άλλες ομάδες ταξινομητών που βασίζονται στον απλό μπεϋζιανό ταξινομητή και έχουν αναλυτικά περιγραφεί σε προηγούμενες ενότητες:

- BoostFSNB
- NBTree
- LBR

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον BoostFSNB σε 4 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 42). Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον NBTree σε 1 σύνολο δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον LBR σε 2 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων.

Δεδομένα	LogitboostNB	BoostFSNB	NBTree	LBR
autos	75.40	76.33	77.18	74.04
badges	99.80	100.00	100.00	100.00
balance-scale	91.19	81.98 *	75.83 *	72.17 *
breast-cancer	66.67	72.07	70.99	72.35
Breast-w	96.18	96.05	95.97	97.23
colic	80.74	82.35	81.88	82.33
credit-g	72.73	71.36	74.07	74.90
Diabetes	74.11	75.10	75.18	75.38
haberman	71.48	73.08	71.97	71.57
heart-c	78.15	82.15	80.60	83.54
heart-h	79.55	83.61	81.33	84.54
heart-statlog	79.93	82.26	80.59	82.59
Hepatitis	84.67	84.81	81.36	84.97
ionosphere	92.71	91.86	89.49	89.92
iris	94.87	93.47	93.53	93.20
labor	93.23	88.03	91.70	87.50
lymphography	84.65	82.99	80.89	85.45
monk1	85.33	69.67 *	91.78	94.91 v
monk2	59.92	61.43	63.72	60.40
monk3	91.87	92.88	92.94	93.45
relation	77.99	77.23	78.00	78.31
sonar	84.41	77.95 *	77.16	76.47 *
students	83.04	86.20	84.62	85.38
vehicle	70.91	62.14 *	71.03	69.43
vote	95.20	95.26	95.03	94.11
wine	98.14	96.84	96.57	98.71
zoo	96.91	95.75	94.55	93.21
		0/23/4	0/26/1	1/26/2
Μέση ακρίβεια	83.70	82.70	83.26	83.56

Πίνακας 42. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλους μετα-ταξινομητές

Τέλος, συγκρίναμε την απόδοση του προτεινόμενου αλγορίθμου με τους AdaBoost C4.5 και Bagging C4.5 που έχουν δώσει πολύ καλά αποτελέσματα σε πολλά προβλήματα μηχανής μάθησης (Quinlan, 1996). Παρόμοια, με τον προτεινόμενο αλγόριθμο, ο Quinlan (1996) χρησιμοποίησε 10 υπο-ταξινομητές για την κατασκευή κάθε ομάδας ταξινομητών. Επίσης συγκρίναμε τον προτεινόμενο αλγόριθμο με τον Logitboost DS, αφού τον DS χρησιμοποίησαν σαν βασικό ταξινομητή οι ερευνητές που πρότειναν τον Logitboost (Friedman και άλλοι, 2000).

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον AdaBoost C4.5 σε 3 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων (βλέπε Πίνακα 43). Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging C4.5 σε 3 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων. Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον LogitboostDS σε 3 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων.

Δεδομένα	LogitboostNB	AdaBoostC4.5	Bagging C4.5	LogitboostDS
Autos	75.40	85.46	82.24 v	79.22
Badges	99.80	100.00	100.00	100.00
balance-scale	91.19	78.35 *	82.04 *	87.34 *
breast-cancer	66.67	66.89	72.71 v	71.42
Breast-w	96.18	95.55	95.17	95.63
Colic	80.74	81.63	85.34 v	82.75
credit-g	72.73	70.75	73.89	71.68
diabetes	74.11	71.69	75.65	74.54
haberman	71.48	71.12	72.78	73.61
heart-c	78.15	78.79	78.88	81.59
heart-h	79.55	78.68	79.93	81.44
heart-statlog	79.93	78.59	80.59	82.22
hepatitis	84.67	82.38	80.73 *	81.58
ionosphere	92.71	93.05	92.17	90.83
iris	94.87	94.33	94.67	94.93
labor	93.23	87.17	82.60	92.33
lymphography	84.65	80.87 *	77.25 *	82.36
monk1	85.33	94.10 v	82.10	71.63 *
monk2	59.92	60.82	59.80	55.60
monk3	91.87	90.01	92.38	93.37
relation	77.99	78.86	78.10	77.83
sonar	84.41	79.13 *	78.51 *	77.17 *
students	83.04	81.70	86.20	86.73 v
vehicle	70.91	75.59 v	74.48	70.73
vote	95.20	95.51	96.27	95.49
wine	98.14	96.45	95.16	97.86
zoo	96.91	95.18	93.21	95.06
		2/22/3	2/22/3	1/23/3
Μέση ακρίβεια	83.70	83.06	83.07	83.15

Πίνακας 43. Σύγκριση της προτεινόμενης μεθόδου LogitboostNB με άλλες ομάδες ταξινομητών

Συγκεντρωτικά, ο προτεινόμενος αλγόριθμος είναι πιο ακριβής από τους όλους τους συγκρινόμενους αλγορίθμους.

4.2 Συνδυασμός αποφάσεων διαφορετικών ταξινομητών

Μέχρι τώρα αναφερόμασταν στο συνδυασμό ταξινομητών που προκύπταν από την εκπαίδευση του ίδιου αλγορίθμου μάθησης σε τροποποιημένα δεδομένα εισόδου. Μια άλλη

κατηγορία μεθόδων συνδυάζει διαφορετικούς αλγορίθμους μάθησης που έχουν εκπαιδευτεί στα ίδια δεδομένα εισόδου.

Στο συνδυασμό διαφορετικών αλγορίθμων, διάφοροι όροι έχουν αναφερθεί ως επιθυμητές ιδιότητες για τους ταξινομητές που συνδυάζονται, μεταξύ αυτών είναι η ορθογωνιότητα (orthogonality), η συμπληρωματικότητα και η ανεξαρτησία. Η ορθογωνιότητα χρησιμοποιείται για να δείξει την τάση των ταξινομητών να λαμβάνουν διαφορετικές αποφάσεις. Είναι γεγονός ότι οι διαφορετικοί ταξινομητές έχουν διαφορετικά μεταξύ τους δυνατά και αδύνατα σημεία, και κατά συνέπεια ο συνδυασμός τους μπορεί να έχει συμπληρωματική επίδραση (συμπληρωματικότητα). Εντούτοις, πρέπει να αναγνωριστεί ότι δυο αυτοί όροι στερούνται ακριβή τρόπο μέτρησης. Η ανεξαρτησία γίνεται καλύτερα κατανοητή, λόγω της συχνής της χρήσης στη θεωρία πιθανοτήτων.

Όσον αφορά τον τρόπο συνδυασμού των αποφάσεων των ταξινομητών, ο πιο απλός κανόνας είναι η πλειοψηφική ψηφοφορία (majority voting), κατά την οποία επιλέγεται η κλάση που προβλέπει η πλειοψηφία των ταξινομητών (Chuanyi & Sheng, 1997) (βλέπε Εικόνα 52). Μια βελτίωση αυτού του σχήματος είναι δυνατή αν κάθε ταξινομητής δίνει στην έξοδό του την εμπιστοσύνη (confidence) του ως προς τις αποφάσεις του, οπότε εκλέγεται η κλάση που κέρδισε τη μέγιστη μέση εμπιστοσύνη (Chuanyi & Sheng, 1997).

Εικόνα 52. Η γραφική αναπαράσταση της ψηφοφορίας των ταξινομητών

Άλλες τεχνικές συνδυασμού αποφάσεων ταξινομητών εκτιμούν το βάρος κάθε ταξινομητή στην επιτροπή, ανάλογα με την αποτελεσματικότητά του στο σύνολο δεδομένων επικύρωσης (Lam & Suen, 1995). Αυτά τα βάρη χρησιμοποιούνται στη συνέχεια, π.χ. μέσω γραμμικού συνδυασμού, για την πρόβλεψη της κλάσης από την επιτροπή. Μία ακόμα πολιτική είναι να επιλέγεται δυναμικά, ανά στιγμιότυπο προς κατάταξη, ο ταξινομητής εκείνος ο οποίος είχε την καλύτερη επίδοση στα N στιγμιότυπα επικύρωσης που συγγενεύουν περισσότερο με το άγνωστο και αυτός να αποφασίζει (dynamic classifier selection) (Giacinto & Roli, 2000). Οι δύο προηγούμενες ιδέες μπορούν να συνδυαστούν, έτσι ώστε να λαμβάνονται υπόψη όλοι οι ταξινομητές, αλλά η απόφασή τους να μετράει με βάση την αποτελεσματικότητά τους στα N στιγμιότυπα επικύρωσης που μοιάζουν περισσότερο με το άγνωστο (adaptive classifier combination) (Giacinto & Roli, 2000).

Μια ακόμα μέθοδος συνδυασμού αποφάσεων διαφορετικών ταξινομητών είναι η λεγόμενη συσσωρευμένη γενίκευση (stacked generalization), που βασίζεται σε μια διεπίπεδη ιεραρχία ταξινομητών χρησιμοποιώντας ένα υψηλού επιπέδου μοντέλο (ταξινομητή) για να συνδυάσει χαμηλότερου επιπέδου μοντέλα με σκοπό την επίτευξη μεγαλύτερης ακρίβειας πρόβλεψης

(Wolpert, 1992). Αναλυτικότερα, η γενική ιδέα της μεθόδου έχει ως εξής: δύο ή περισσότεροι αλγόριθμοι μάθησης εκπαιδεύονται αρχικά σε ένα πλήθος από υποσύνολα των αρχικών δεδομένων, παράγοντας αντίστοιχο αριθμό μοντέλων. Στη συνέχεια, κάθε στιγμιότυπο από τα αρχικά δεδομένα αντιστοιχίζεται σε ένα νέο, το οποίο αναπαριστά την πρόβλεψη κάθε μοντέλου για το αρχικό στιγμιότυπο, καθώς και την πραγματική τιμή της συνάρτησης-στόχου (βλέπε Εικόνα 53). Στο βήμα αυτό πρέπει να εξασφαλιστεί πως τα μοντέλα δημιουργούνται από σύνολα εκπαίδευσης που δεν περιλαμβάνουν το στιγμιότυπο που αναπαρίσταται (για το οποίο δηλαδή κάνουν πρόβλεψη), ακριβώς όπως ισχύει στη διασταυρωμένη επικύρωση.

Εικόνα 53. Η κατασκευή των μετα-δεδομένων

Εν συνεχεία, τα νέα δεδομένα σχηματίζουν ένα νέο πρόβλημα μάθησης, και στο δεύτερο βήμα της μεθόδου ένας αλγόριθμος μάθησης καλείται να το λύσει, παράγοντας το μοντέλο του δεύτερου επιπέδου (βλέπε Εικόνα 54). Σύμφωνα με την ορολογία του Wolpert, τα αρχικά δεδομένα και τα μοντέλα που κατασκευάζονται για αυτά στο πρώτο βήμα αναφέρονται ως δεδομένα επιπέδου-0 και μοντέλα επιπέδου-0, αντίστοιχα, ενώ τα νέα δεδομένα που προέρχονται από τις προβλέψεις των μοντέλων επιπέδου-0 και ο αλγόριθμος μάθησης που καλείται στο δεύτερο βήμα αναφέρονται αντίστοιχα ως δεδομένα επιπέδου-1 και ταξινομητής επιπέδου-1. Σχηματικά, τα μοντέλα επιπέδου-0 μπορούν να θεωρηθούν ως τα “απλά μέλη” της επιτροπής ταξινομητών, ενώ το μοντέλο που κατασκευάζεται από τα δεδομένα επιπέδου-1 αντιστοιχεί στον “πρόεδρο” της επιτροπής, ο οποίος εκπαιδεύεται πάνω στις προβλέψεις των μελών για κάθε στιγμιότυπο. Στη φάση της κατάταξης ενός νέου στιγμιότυπου, κάθε μέλος κάνει ανεξάρτητα την πρόβλεψή του και στη συνέχεια ο πρόεδρος παίρνει την τελική απόφαση, λαμβάνοντας υπόψη τις γνώμες των μελών σύμφωνα με το μοντέλο που έχει κατασκευαστεί.

Εφόσον ο πρόεδρος της επιτροπής χρειάζεται να εκπαιδευθεί πάνω στις κρίσεις των μελών, το σύνολο εκπαίδευσης πρέπει να χωριστεί σε ένα τμήμα εκπαίδευσης των μελών και ένα τμήμα εξέτασης των μελών (επικύρωσης). Αν αυτός ο διαμερισμός γίνει μία μόνο φορά, ο πρόεδρος θα έχει στη διάθεση του μόνο τις προβλέψεις των μελών στο σύνολο επικύρωσης, δηλαδή ένα υποσύνολο των διαθέσιμων δεδομένων. Με τη διασταυρωμένη επικύρωση όμως, διαφορετικοί ταξινομητές επιπέδου-0 (μέλη) εκπαιδεύονται πάνω σε διαφορετικά υποσύνολα των αρχικών δεδομένων και εξετάζονται στα υπόλοιπα, έτσι ώστε τελικά ο πρόεδρος να αποκτήσει το ίδιο πλήθος δεδομένων με τα αρχικά.

Εικόνα 54. Η εκπαίδευση του μετα-ταξινομητή

Μερικές από τις επιλογές που είναι κρίσιμες για την αποτελεσματικότητα της συσσωρευμένης γενίκευσης είναι:

- i) Οι αλγόριθμοι μάθησης για την κατασκευή των μελών.
- ii) Ο αλγόριθμος μάθησης για την κατασκευή του προέδρου.
- iii) Το πλήθος των μελών.
- iv) Η αναπαράσταση των δεδομένων επιπέδου-1.

Για τα ζητήματα αυτά, τα οποία κατά την πρώτη δημοσίευση της μεθόδου ο Wolpert (1992) χαρακτήρισε ως “μαύρη τέχνη”, δεν έχουν δοθεί ακόμα πλήρως αποδεκτές απαντήσεις. Θεωρητικά, οποιοσδήποτε αλγόριθμος μάθησης μπορεί να χρησιμοποιηθεί στα (i) και (ii), και πράγματι έχουν δοκιμαστεί διάφοροι γνωστοί αλγόριθμοι χωρίς να προκύψουν ξεκάθαρα συμπεράσματα. Στο άρθρο των (Ting & Witten, 1999) υποστηρίζεται πειραματικά πως μεταξύ του C4.5, του NB, του IB1 και του MLR (προσαρμογή ενός αλγορίθμου γραμμικής παλινδρόμησης ελαχίστων τετραγώνων), μόνο ο τελευταίος είναι κατάλληλος για πρόεδρος της επιτροπής στα περισσότερα προβλήματα μάθησης που δοκιμάστηκαν.

Ως προς το (iii), δεν έχει βρεθεί ο βέλτιστος αριθμός αν και πολλοί ταξινομητές επιβαρύνουν τον χρόνο εκπαίδευσης. Ως προς το (iv), πέραν από την περίπτωση που περιγράφηκε παραπάνω, κατά την οποία τα δεδομένα επιπέδου-1 είναι απλά οι προβλέψεις των μελών, μία άλλη επιλογή είναι δυνατή αν τα μέλη βγάζουν ως έξοδο το βαθμό εμπιστοσύνης τους για την κατάταξη του στιγμιότυπου σε κάθε μία από τις κλάσεις. Σε αυτή την περίπτωση, τα δεδομένα του επιπέδου-1 μπορούν να αποτελούνται από το βαθμό εμπιστοσύνης που δίνει το κάθε μέλος σε κάθε μία κλάση (Ting & Witten, 1999).

Μια άλλη μέθοδος που μοιάζει με την συσσωρευμένη γενίκευση είναι αυτή της βαθμολόγησης (grading) (Seewald & Furnkranz, 2001). Ένας ταξινομητής επιπέδου-1 εκπαιδεύεται για προβλέπει ποιους ταξινομητές επιπέδου-0 να εμπιστεύεται κατά τον συνδυασμό των αποφάσεων. Ο ταξινομητής επιπέδου-1 χρησιμοποιεί εκτός από τα δεδομένα εισόδου για είσοδο και ένα μετα-χαρακτηριστικό που του υποδεικνύει αν ο ταξινομητής επιπέδου-0 ήταν σωστός ή όχι στην ταξινόμηση του συγκεκριμένου παραδείγματος. Μόνο οι ταξινομητές επιπέδου-0 που προβλέπει ο ταξινομητής επιπέδου-1 ότι κάνουν σωστή πρόβλεψη συμμετέχουν στην τελική ψηφοφορία.

Πληθώρα άλλων αξιόλογων μεθόδων συνδιασμού ομάδων ταξινομητών έχουν παρουσιαστεί (Gama & Brazdil, 2000), (Merz, 1999), (Todorovski & Dzeroski, 2003). Αν και, ή ίσως επειδή, πολλές μέθοδοι δημιουργίας ομάδων ταξινομητών έχουν προταθεί, δεν υπάρχει μέχρι σήμερα καμία σαφής εικόνα ποια μέθοδος είναι καλύτερη. Αυτό μας ώθησε στην δημιουργία δικών μας νέων μεθόδων συνδιασμού των αποφάσεων διαφορετικών ταξινομητών.

4.2.1 Συνδυάζοντας τους αλγόριθμους NB και C4.5

Σε αυτή την ενότητα πειραματικά αποδεικνύουμε ότι ένας απλός συνδυασμός όπως η μέθοδος τριπλής διασταυρωμένης επιλογής (3-cross validation) του καλύτερου ταξινομητή χρησιμοποιώντας τους αλγόριθμους NB και C4.5 δίνει στις περισσότερες περιπτώσεις καλύτερη ακρίβεια χρησιμοποιώντας λιγότερο χρόνο για εκπαίδευση από άλλες ομάδες ταξινομητών που χρησιμοποιούν είτε τον NB είτε τον C4.5 ως βασικό ταξινομητή.

Η βασική ιδέα των πειραμάτων ήταν να συνδυαστούν δύο αλγόριθμοι μάθησης που είχαν ήδη χρησιμοποιηθεί σε πολλές εφαρμογές. Αυτό έγινε με σκοπό να εκτιμηθεί το κατά πόσο οι ταξινομητές που κατασκευάζονται με βάση τους NB και C4.5 είναι ασυσχέτιστοι μεταξύ τους. Μόνο στην περίπτωση που η συσχέτισή τους κρινόταν μικρή, όπως και έγινε, θα χρησιμοποιούντουσαν για περαιτέρω έρευνα.

Ένας απλός, αλλά ενδεικτικός, τρόπος να εκτιμηθεί στην πράξη η εν λόγω συσχέτιση ήταν να μετρηθούν τα λάθη ταξινόμησης που κάνουν από κοινού οι ταξινομητές και τα λάθη που κάνει μόνο ο ένας εκ των δύο. Αν τα τελευταία ήταν αρκετά περισσότερα από τα πρώτα, ένας αλγόριθμος θα είχε αρκετά περιθώρια να “μάθει” να επιλέγει το σωστό μέλος κάθε φορά. Αντίθετα, αν οι ταξινομητές έκαναν συνήθως τα ίδια λάθη, με δεδομένη τη γενικά υψηλή τους ακρίβεια στο σύνολο επικύρωσης, οποιοσδήποτε αλγόριθμος θα μάθαινε να εμπιστεύεται σχεδόν πάντα τη γνώμη τους όταν αυτοί συμφωνούν και κατά συνέπεια θα αναπαρήγαγε τα λάθη τους.

Η μεθοδολογία της μεθόδου διασταυρωμένης επικύρωσης που ακολουθήσαμε (BestCV) αποτελείται από τέσσερα βήματα (παρόμοια μέθοδο χρησιμοποίησε ο Schaffer (1993) για το συνδυασμό ενός νευρωνικού δικτύου με ένα δέντρο απόφασης):

- Το νέο σύνολο δεδομένων διαιρείται τυχαία σε τρία ίσα μέρη
- Τα δύο τρίτα των δεδομένων χρησιμοποιούνται για την εκπαίδευση των αλγορίθμων και το υπόλοιπο ένα τρίτο είναι το σύνολο ελέγχου. Η διαδικασία επαναλαμβάνεται 3 φορές.
- Από τα αποτελέσματα των τριών πειραμάτων υπολογίζουμε το μέσο όρο της ακρίβειας κάθε αλγορίθμου
- Ο αλγόριθμος με την μεγαλύτερη κατά μέσο όρο ακρίβεια επιλέγεται για να χτίσει το τελικό μοντέλο χρησιμοποιώντας το σύνολο των δεδομένων εκπαίδευσης

Η σχηματική αναπαράσταση της μεθόδου ακολουθεί:

$$\sigma\phi\alpha\lambda\mu\alpha_s(h) = \sum_{i=1}^f \frac{|S_i|}{|S|} \sigma\phi\alpha\lambda\mu\alpha_{s_i}(h_i)$$

Εικόνα 55. Σχηματική αναπαράσταση της μεθόδου

Συγκρίναμε αυτή την μέθοδο με τις ακόλουθες τεχνικές χρησιμοποιώντας 25 υπο-ταξινομητές:

- Bagging NB, Boosting NB και Multiboost NB
- Bagging C4.5, Boosting C4.5 και Multiboost C4.5

Σε όλα τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Συγκεντρωτικά, στην τελευταία γραμμή του Πίνακα 44 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα προηγούμενα πειράματα.

Δεδομένα	<i>BestCV</i>	<i>Bagging</i> <i>C4.5</i>	<i>Boosting</i> <i>C4.5</i>	<i>Multiboost</i> <i>C4.5</i>	<i>Bagging</i> <i>NB</i>	<i>Boosting</i> <i>NB</i>	<i>Multiboost</i> <i>NB</i>
anneal	98.57	98.83 v	99.61 v	99.59 v	87.27 *	95.2 *	94.13 *
autos	81.77	83.89	86.05 v	85.75 v	57.71 *	57.12 *	57.12 *
balance	90.53	82.33 *	76.91 *	79.26 *	90.29	92.11 v	92.29 v
breast-cancer	72.35	73.37	66.5 *	68.27 *	72.73	68.57 *	69.01 *
breast-w	96.02	96.31	96.51	96.51	96.07	95.55	95.58
colic	85.16	85.23	82.01 *	83.13 *	78.94 *	77.46 *	79.28 *
credit-a	85.57	86.23	85.75	86.13	77.81 *	81.16 *	81.04 *
credit-g	75	74.17	72.79 *	74.35	75.13	75.09	74.71
diabetes	75.48	75.67	72.81 *	74.21	75.57	75.88	76.2
heart-c	83.34	79.54 *	79.6 *	79.87 *	83.24	83.14	83.56
heart-h	83.79	79.91 *	78.28 *	80.11 *	84.16	84.67	84.8
heart-statlog	83.59	81.11 *	80.15 *	81.37 *	83.41	82.3 *	82.7
hepatitis	82.89	81.63	82.74	83.59	84.39	84.23	84.67
hypothyroid	99.54	99.58	99.67 v	99.68 v	95.53 *	95.27 *	95.33 *
ionosphere	89.74	92.23 v	93.62 v	93.54 v	81.94 *	91.12	91.66 v
iris	95.27	94.8	94.47	94.47	95.53	95.07	95.07
kr-vs-kp	99.44	99.45	99.62 v	99.62 v	87.8 *	95.1 *	95.22 *
labor	93.37	83.83 *	89.1	88.1 *	93.4	88.93 *	90.07 *
lymphotherapy	83.13	79.14 *	83.09	83.24	83.5	80.67	82.64
mushroom	100	100	100	100	95.73 *	100	100
primary-tumor	49.71	44.4 *	41.65 *	41.65 *	49.47	49.71	49.71
segment	96.79	97.54 v	98.42 v	98.36 v	80.41 *	80.16 *	80.16 *
sonar	71.51	79.03 v	83.03 v	82.54 v	67.96	81.21 v	81.5 v
vote	96.57	96.53	95.24 *	95.51 *	90.09 *	95.19 *	95.36 *
waveform	80.01	82.81 v	83.32 v	83.73 v	80	80.01	80.29 v
zoo	93.3	93.29	95.38	95.77 v	95.07 v	97.23 v	97.23 v
		5/14/7	8/8/10	9/8/9	1/15/10	3/12/11	5/11/10

Πίνακας 44. Σύγκριση της εξεταζόμενης μεθόδου με άλλες ομάδες ταξινομητών που βασίζονται είτε στον C4.5 είτε στον NB

Ο αλγόριθμος BestCV είναι στατιστικά ακριβέστερος από τον Bagging NB σε 10 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων (βλέπε Πίνακα 44). Αντίστοιχα, ο αλγόριθμος BestCV είναι στατιστικά ακριβέστερος από τον AdaBoost NB σε 11 σύνολα δεδομένων, ενώ έχει στατιστικά

υψηλότερα ποσοστά λάθους σε 3 σύνολα δεδομένων. Επιπλέον, ο αλγόριθμος BestCV είναι στατιστικά ακριβέστερος από τον Multiboost NB σε 10 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 5 σύνολα δεδομένων.

Αντίστοιχα, ο αλγόριθμος BestCV είναι στατιστικά ακριβέστερος από τον Bagging C4.5 σε 7 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 5 σύνολα δεδομένων. Επίσης, ο αλγόριθμος BestCV είναι στατιστικά ακριβέστερος από τον Boosting C4.5 σε 10 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 8 σύνολα δεδομένων. Επιπλέον, ο αλγόριθμος BestCV είναι στατιστικά ακριβέστερος από τον multiboost C4.5 σε 9 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε άλλα 9 σύνολα δεδομένων.

Από τα πειράματα φάνηκε ότι η μέθοδος τριπλής διασταυρωμένης επιλογής (3-cross validation) του καλύτερου ταξινομητή χρησιμοποιώντας τους αλγορίθμους NB και C4.5 δίνει στις περισσότερες περιπτώσεις καλύτερη ακρίβεια από άλλες ομάδες ταξινομητών που χρησιμοποιούν είτε τον NB είτε τον C4.5 ως βασικό ταξινομητή. Τα περιθώρια υψηλής απόδοσης, λοιπόν, είναι μεγάλα για μια επιτροπή (βασιζόμενη στους NB και C4.5 ταξινομητές) που θα μάθει να επιλέγει το σωστό μέλος ανά περίπτωση όταν αυτά διαφωνούν, έστω κι αν αποτυγχάνει σε όλα τα στιγμιότυπα που αποτυγχάνουν και τα δύο μέλη μαζί.

4.2.2 Προτεινόμενη μέθοδος ψηφοφορίας μέσα σε σάκο

Στη συνέχεια λοιπόν, αντί για την μέθοδο διασταυρωμένης (3-cross validation) επιλογής του καλύτερου ταξινομητή [3] χρησιμοποιήσαμε διαφορετικά υποσύνολα του ίδιου συνόλου δεδομένων εκπαίδευσης με την ταυτόχρονη χρήση μιας μεθοδολογίας ψηφοφορίας που συνδυάζει τους NB και C4.5 ταξινομητές. Τα πειράματα απέδειξαν την περαιτέρω βελτίωση στην ακρίβεια πρόγνωσης στα περισσότερα δεδομένα εκπαίδευσης - επικύρωσης.

Το σακούλιασμα (τοποθέτηση σε σάκο ή bagging) χρησιμοποιεί μια τεχνική ψηφοφορίας που δεν λαμβάνει υπόψη την ετερογένεια του διαστήματος κάθε περίπτωσης. Όταν η πλειοψηφία των ταξινομητών βάσης δίνει λανθασμένη πρόβλεψη για την περίπτωση ελέγχου είναι αυτονόητο ότι η ψηφοφορία πλειοψηφίας θα οδηγήσει σε λανθασμένη πρόβλεψη. Το πρόβλημα μπορεί είτε να συνίσταται στην απόρριψη των ταξινομητών βάσης που είναι ιδιαίτερα ακριβείς σε μια περιορισμένη περιοχή του διαστήματος περιπτώσεων επειδή αυτή η ακρίβεια υπερκαλύπτεται από την ανακρίβειά τους έξω από την περιορισμένη αυτή περιοχή. Είτε μπορεί να συνίσταται στη χρήση των ταξινομητών που είναι ακριβείς στο μεγαλύτερο μέρος του διαστήματος αλλά συγχέουν ολόκληρη την επιτροπή ταξινόμησης σε μερικούς περιορισμένους τομείς του διαστήματος.

Για να υπερνικήσουμε αυτό το πρόβλημα προτείνουμε να χρησιμοποιούνται διαφορετικά υποσύνολα του ίδιου συνόλου δεδομένων εκπαίδευσης με την ταυτόχρονη χρήση όμως μιας μεθοδολογίας ψηφοφορίας που συνδυάζει διαφορετικούς αλγορίθμους και όχι όμοιους αλγορίθμους. Στον αλγόριθμό μας, οι ψηφοφόροι (NB και C4.5) εκφράζουν το βαθμό προτίμησής τους χρησιμοποιώντας ως αποτέλεσμα την εμπιστοσύνη στην πρόβλεψή τους. Έπειτα οι τιμές εμπιστοσύνης προστίθενται για κάθε κλάση και η κλάση με το υψηλότερο ποσό κερδίζει την εκλογή.

Αναλυτικότερα, η προτεινόμενη μέθοδος περιγράφεται με ψευδογλώσσα στην Εικόνα 56. Πρέπει να σημειωθεί ότι η πολυπλοκότητα χρόνου της προτεινόμενης μεθόδου αυξάνεται δύο φορές σε σχέση με αυτή του σακουλιάσματος (bagging). Αυτή η αύξηση, εντούτοις, αντισταθμίζεται από το γεγονός ότι η προτεινόμενη μέθοδος (τοποθετημένη σε σάκο ψηφοφορία) παρέχει καλύτερη απόδοση γενίκευσης με πολύ λιγότερες επαναλήψεις. Γι' αυτόν τον λόγο, χρησιμοποιήσαμε μόνο 10 επαναλήψεις για τον προτεινόμενο αλγόριθμο.

ΔΗΜΙΟΥΡΓΙΑ ΜΟΝΤΕΛΟΥ

Αν n είναι ο αριθμός των δεδομένων εκπαίδευσης.

Κάνε για 10 επαναλήψεις:

- Πάρε n παραδείγματα από τα δεδομένα εκπαίδευσης χρησιμοποιώντας τυχαία δειγματοληψία με επανατοποθέτηση.
- Δημιούργησε δυο ταξινομητές (NB, DT) από αυτό το δείγμα
- Εφάρμοσε την αθροιστική ψηφοφορία στους 2 ταξινομητές
- Αποθήκευσε τα μοντέλα

ΤΑΞΙΝΟΜΗΣΗ

Για καθένα από τα μοντέλα που παράχθηκαν:

Πρόβλεψε την κλάση.

Η τελική απάντηση προκύπτει από την αθροιστική ψηφοφορία

Εικόνα 56. Ψευδογλώσσα της προτεινόμενης μεθόδου ψηφοφορίας μέσα σε σάκο (Bagged Voting)

Αρχικά, συγκρίναμε την προτεινόμενη μέθοδο με τις ακόλουθες τεχνικές που χρησιμοποιούν σαν βασικούς ταξινομητές είτε τον NB είτε τον C4.5:

- Bagging NB, Boosting NB και Multiboost NB (χρησιμοποιώντας 25 επαναλήψεις – υποταξινομητές)
- Bagging C4.5, Boosting C4.5 και Multiboost C4.5 (χρησιμοποιώντας 25 επαναλήψεις – υποταξινομητές)

Σε όλα τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή- t ($p < 0.05$). Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 45 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του Πίνακα 45 παρουσιάζεται το μέσο σφάλμα κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging NB σε 10 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων (βλέπε Πίνακας 45). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Boosting NB σε 8 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Multiboost NB σε 7 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων.

Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Bagging C4.5 σε 4 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Boosting C4.5 σε 8 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 4 σύνολα δεδομένων. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Multiboost C4.5 σε 6 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε άλλα 4 σύνολα δεδομένων.

Δεδομένα	Bagged Voting	Bagging NB	Boosting NB	Multi boost NB	Bagging C4.5	Boosting C4.5	Multiboost C4.5
anneal	97.93	87.27*	95.2 *	94.13 *	98.83	99.61 v	99.59 v
badge	100	99.69	99.66	99.66	100	100	100
balance	86.68	90.29 v	92.11 v	92.29 v	82.33*	76.91*	79.26 *
breast-cancer	72.08	72.73	68.57	69.01	73.37	66.50*	68.27
breast-w	96.15	96.07	95.55	95.58	96.31	96.51	96.51
colic	83.24	78.94 *	77.46 *	79.28	85.23	82.01	83.13
credit-g	75.63	75.13	75.09	74.71	74.17	72.79	74.35
diabetes	76.17	75.57	75.88	76.2	75.67	72.81 *	74.21
dimin	97.56	92.77*	92.93 *	93.95 *	97.07*	96.03 *	96.17 *
eucalyptus	63.51	54.87*	59.06 *	59.24 *	65.02	64.89	64.73
haberman	75.2	74.86	73.94	73.94	72.06	71.12 *	71.12 *
heart-c	83.04	83.24	83.14	83.56	79.54	79.6	79.87
heart-h	84.03	84.16	84.67	84.8	79.91 *	78.28 *	80.11 *
heart-statlog	84.37	83.41	82.3	82.7	81.11	80.15	81.37
hepatitis	83.61	84.39	84.23	84.67	81.63	82.74	83.59
ionosphere	90.66	81.94*	91.12	91.66	92.23	93.62 v	93.54
iris	95.4	95.53	95.07	95.07	94.8	94.47	94.47
kr-vs-kp	99.31	87.80*	95.10*	95.22 *	99.45	99.62 v	99.62 v
labor	90.87	93.4	88.93	90.07	83.83*	89.1	88.1
lymphotherapy	82.06	83.5	80.67	82.64	79.14	83.09	83.24
monk1	81.03	73.35 *	72.37 *	72.42 *	82.32	96.54 v	94.36 v
mushroom	100	95.73*	100	100	100	100	100
primary-tumor	47.34	49.47	49.71	49.71	44.4	41.65 *	41.65 *
soybean	93.57	92.83	92.02	93.15	92.78	93.19	93.21
student	86.81	85.59	85.12	85.47	86.49	81.44 *	81.68 *
titanic	78.22	77.86 *	77.86 *	77.88 *	77.93	78.89	78.71
waveform	82.1	80.00*	80.01 *	80.29 *	82.81	83.32	83.73 v
white-clover	66.81	62.79	70.74	71.14	68.69	72.71	71.55
wine	97.18	97.52	96.57	96.57	95.5	96.62	96.84
zoo	96.23	95.07	97.23	97.23	93.29	95.38	95.77
		1/19/10	1/21/8	1/22/7	0/26/4	4/18/8	4/20/6
Μέσο σφάλμα	15.11	17.14	16.26	15.93	16.14	16.01	15.71

Πίνακας 45. Σύγκριση της προτεινόμενης μεθόδου ψηφοφορίας μέσα σε σάκο με άλλες ομάδες ταξινομητών που χρησιμοποιούν σαν βασικούς ταξινομητές είτε τον NB είτε τον C4.5

Επίσης, συγκρίναμε την προτεινόμενη μέθοδο με τις ακόλουθες ομάδες που συνδυάζουν τους βασικούς ταξινομητές NB και C4.5:

- Συσσωρευμένη γενίκευση (stacking) με ταξινομητή επιπέδου-1 τον MLR όπως οι Ting & Witten (1999)
- Απλή ψηφοφορία (voting)
- Βαθμολόγηση (grading) με ταξινομητή επιπέδου-1 τον 10-NN όπως έκαναν οι Seewald & Furnkranz (2001)

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από την συσσωρευμένη γενίκευση χρησιμοποιώντας τους ίδιους υπο-ταξινομητές σε 2 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους 1 κανένα σύνολο δεδομένων (βλέπε Πίνακα

46). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από την απλή ψηφοφορία (voting) σε 4 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από την βαθμολόγηση (grading) χρησιμοποιώντας τους ίδιους υποταξινομητές σε 2 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους 1 σύνολο δεδομένων.

	Bagged Voting	Stacking	Voting	Grading
Anneal	97.93	98.51	98.7	98.86
Badge	100	100	100	100
Balance	86.68	90.23 v	78.41*	90.3 v
breast-cancer	72.08	71.35	71.66	72.04
breast-w	96.15	96.01	95.74	96.08
Colic	83.24	85.02	83.07	84.07
credit-g	75.63	74.81	72.58*	74.06
Diabetes	76.17	76.39	75.9	75.63
Dimin	97.56	97.2*	97.26 *	97.25 *
Eucalyptus	63.51	63.6	61.63	61.29
Haberman	75.2	73.88	74.34	72.78
heart-c	83.04	83.11	80.1	81.46
heart-h	84.03	83.58	83.95	82.76
heart-statlog	84.37	84.04	80.89	81.63
Hepatitis	83.61	84.00	81.98	82.51
ionosphere	90.66	89.97	88.04	92.51
Iris	95.4	95.33	95.53	94.87
kr-vs-kp	99.31	99.44	99.38	99.3
Labor	90.87	93.57	86.87	90.33
Lymphotherapy	82.06	81.58	78.66	82.39
monk1	81.03	81.02	81.28	79.22
Mushroom	100	100	100	100
primary-tumor	47.34	45.93	45.64	47.5
Soybean	93.57	93.06	93.47	93.31
Student	86.81	86.9	86.73	85.99
Titanic	78.22	77.76	77.94	77.9
Waveform	82.1	81.84	77.32*	79.55 *
white-clover	66.81	56.43*	65.31	65.93
Wine	97.18	97.01	95.84	96.51
Zoo	96.23	95.95	95.18	96.15
		1/27/2	0/26/4	1/27/2
<i>Μέσο σφάλμα</i>	<i>15.11</i>	<i>15.42</i>	<i>16.55</i>	<i>15.59</i>

Πίνακας 46. Σύγκριση της προτεινόμενης μεθόδου ψηφοφορίας μέσα σε σάκο με άλλους αλγορίθμους που συνδυάζουν τις αποφάσεις των NB και C4.5

Συνοψίζοντας, η παρουσιαζόμενη μεθοδολογία του συνδυασμού των NB και C4.5 αλγορίθμων είναι μια αποτελεσματική μέθοδος με χαμηλό λάθος ταξινόμησης για έναν πρόβλημα όπου δεν υπάρχει πρότερη γνώση διαθέσιμη.

Πρέπει επίσης να αναφερθεί ότι η προτεινόμενη μεθοδολογία μπορεί εύκολα να εκτελεστεί παράλληλα, εκτός από σειριακά. χρησιμοποιώντας έναν αλγόριθμο μάθησης ανά μηχανή.

Εκμεταλλευόμενος ο αλγόριθμος την παράλληλη ή και κατανεμημένη εκτέλεση μπορεί να αυξήσει την ταχύτητά του και το εύρος των εφαρμογών όπου μπορεί να χρησιμοποιηθεί επειδή για παράδειγμα μπορεί να χειριστεί πολύ μεγαλύτερες βάσεις δεδομένων (Hall και άλλοι, 2000).

4.2.3 Προτεινόμενη μέθοδος επιλεκτικής ψηφοφορίας

Είναι αναμενόμενο ότι όταν ένας ταξινομητής δεν αποδίδει καθόλου καλά σε ένα συγκεκριμένο πρόβλημα τότε επηρεάζει αρνητικά την συνολική απόφαση της ομάδας των ταξινομητών (Erp και άλλοι, 2002). Γι' αυτό το λόγο σκεφτήκαμε να χρησιμοποιήσουμε επιλεκτική ψηφοφορία κατά την οποία οι ταξινομητές που έχουν στατιστικά χειρότερη ακρίβεια πρόγνωσης δεν θα παίρνουν μέρος στην τελική ψηφοφορία. Τα πειράματα της μελέτης μας απέδειξαν ότι η επιλεκτική ψηφοφορία δίνει καλύτερα αποτελέσματα από την απλή πλειοψηφική ψηφοφορία αλλά και άλλες πιο περίπλοκες μεθόδους όπως η συσσωρευμένη γενίκευση.

Η προτεινόμενη μέθοδος επιλεκτικής ψηφοφορίας (selective voting) αποτελείται από έξι βήματα [6]:

- Από το αρχικό σύνολο δεδομένων επιλέγεται τυχαία περίπου το 20% των δεδομένων εκπαίδευσης
- Το νέο σύνολο δεδομένων διαιρείται τυχαία σε τρία ίσα μέρη
- Τα δύο τρίτα των δεδομένων χρησιμοποιούνται για την εκπαίδευση των αλγορίθμων και το υπόλοιπο ένα τρίτο είναι το σύνολο ελέγχου. Η διαδικασία επαναλαμβάνεται 3 φορές
- Από τα αποτελέσματα των τριών πειραμάτων υπολογίζουμε το μέσο όρο της ακρίβειας
- Οι αλγόριθμοι που έχουν στατιστικά χειρότερη ακρίβεια (σύμφωνα με την δοκιμή t με $p < 0.05$) από τον ακριβέστερο αλγόριθμο δεν χρησιμοποιούνται στην ομάδα
- Οι υπόλοιποι αλγόριθμοι αφού εκπαιδευτούν στο σύνολο των δεδομένων εκπαίδευσης εν συνεχεία αποτελούν την ομάδα των ταξινομητών που αποφασίζει χρησιμοποιώντας την απλή ψηφοφορία.

Σχηματικά, η διαδικασία ταξινόμησης περιλαμβάνει δύο φάσεις: (1) φάση εκπαίδευσης και (2) φάση εφαρμογής (βλέπε Εικόνα 57). Κατά τη διάρκεια της φάσης εκπαίδευσης, κάθε ταξινομητής βάσης (ταξινομητές $h_1 \dots h_n$) εκπαιδεύεται. Για την ταξινόμηση νέων παραδειγμάτων οι αντίστοιχες ταξινομήσεις των ταξινομητών βάσης συνδυάζονται με την εκλεκτική ψηφοφορία $h^* = F(h_1, h_2, \dots, h_n)$ για να παραγάγουν την τελική ταξινόμηση του παραδείγματος (y^*).

Εικόνα 57. Η γραφική αναπαράσταση της επιλεκτικής ψηφοφορίας

Σαν αλγόριθμους βάσης χρησιμοποιήσαμε με την προτεινόμενη μέθοδο έναν αλγόριθμο ανά κατηγορία μεθόδων μηχανικής μάθησης:

- Τον NB ως εκπρόσωπο των Μπεϋζιανών αλγορίθμων
- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των οκνηρών αλγορίθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMO ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης
- Τον RIPPER ως εκπρόσωπο των κανόνων ταξινόμησης

Στο πρώτο μας πείραμα αξιολογήσαμε την απόδοση της προτεινόμενης τεχνικής σε σχέση με κάθε αλγόριθμο βάσης. Για τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$).

Δεδομένα	Selective Voting	C4.5	3NN	NB	BP	SMO	RIPPER
autos	81.77	81.77	67.23 *	57.41 *	48.84 *	56.55 *	74.05 *
badges	100	100	100	99.66 *	100	100	100
breast-cancer	72.67	74.28	73.13	72.7	71.18	69.52	71.65
breast-w	96.71	95.01 *	96.61	96.07	95.97	96.75	95.72
credit-a	86.2	85.57	84.94	77.86 *	86.07	84.88	85.33
credit-g	75.3	71.25 *	72.21 *	75.16	72.75	75.15	71.86 *
diabetes	76.59	74.49	73.86	75.75	76.56	76.8	75.22
Glass	72.52	67.63	70.02	49.45 *	51.99 *	57.37 *	66.94 *
haberman	73.13	71.05	69.77	75.06	74.64	73.37	72.43
heart-c	84.45	76.94 *	81.82	83.34	81.39 *	83.89	79.05 *
heart-h	83.79	80.22	82.33	83.95	81.37	82.81	79.26 *
heart-statlog	84.19	78.15 *	79.11 *	83.59	82.11	83.89	78.7 *
hepatitis	84.65	79.22	80.9	83.81	81.3	85.7	77.43 *
ionosphere	90.69	89.74	86.02 *	82.17 *	85.84 *	88.07 *	89.3
Iris	95.6	94.73	95.2	95.53	96.27	96.2	93.93
Labor	93.4	78.6 *	87.83	93.57	88.57	92.97	83.37
lymphography	83.72	75.84	81.74	83.13	80.24 *	86.41	77.36
monk1	92.48	80.61 *	78.97 *	73.38 *	86.69 *	79.58 *	86.74 *
monk2	75.7	57.75 *	54.74 *	56.83 *	75.7	58.52 *	55.91 *
monk3	93.45	92.95	86.72 *	93.45	88.69	93.45	83.99 *
Mushroom	100	100	100	95.76 *	99.97	100	100
sonar	83.76	73.61 *	83.76	67.71 *	78.67 *	77.88 *	75.45 *
soybean	93.71	91.78 *	91.2 *	92.94	33.43 *	93.1	91.93 *
Vehicle	74.18	72.28 *	70.21 *	44.68 *	50.07 *	74.08	68.29 *
Vote	96.16	96.57	93.08 *	90.02 *	94.6	95.77	95.7
vowel	96.99	80.2 *	96.99 *	62.9 *	30.09 *	70.61 *	71.17 *
wine	98.47	93.2 *	95.85	97.46	83.08 *	98.76	92.48
		0/15/12	0/16/11	0/15/12	0/16/11	0/20/7	0/13/14
<i>Μέση ακρίβεια</i>	86.68	81.98	82.75	79.38	76.89	82.67	81.23

Πίνακας 47. Σύγκριση της επιλεκτικής ψηφοφορίας με όλους τους αλγόριθμους βάσης

Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 47 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του Πίνακα 47 παρουσιάζεται η μέση ακρίβεια κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους NB και C4.5 σε 12 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους BP και 3NN σε 11 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον RIPPER και τον SMO σε 14 και 7 σύνολα δεδομένων αντίστοιχα, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Συνοψίζοντας, όπως ήταν αναμενόμενο η απόδοση της παρουσιαζόμενης τεχνικής είναι πιο ακριβής από όλους τους αλγόριθμους βάσης. Η μέση σχετική βελτίωση της ακρίβειας ταξινόμησης της προτεινόμενης μεθοδολογίας είναι από 5% έως 13% σε σχέση με τους αλγορίθμους βάσης.

Επίσης, συγκρίναμε την προτεινόμενη μέθοδο χρησιμοποιώντας τους ίδιους αλγόριθμους βάσης με άλλες τεχνικές συνδυασμού ταξινομητών όπως:

- Ψηφοφορία (voting)
- Βαθμολόγησης (grading)
- Συσσωρευμένης γενίκευσης (stacking) χρησιμοποιώντας την μία φορά τον αλγόριθμο MLR ως αλγόριθμο μετα-μάθησης όπως πρότειναν στην εργασία τους οι (Ting & Witten, 1999) και την άλλη τον αλγόριθμο M5' (έναν αλγόριθμο παλινδρόμησης που βασίζεται στα δέντρα αποφάσεων) όπως πρότειναν στην εργασία τους οι (Dzeroski & Zenko, 2002).
- Την απλή μέθοδο της τριπλής διασταυρωμένης επιλογής (BestCV) που περιγράφηκε στην ενότητα 4.2.1.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά πιο ακριβής από την απλή ψηφοφορία (voting), την βαθμολόγηση (grading) και την συσσωρευμένη γενίκευση (stacking) χρησιμοποιώντας σαν αλγόριθμο επιπέδου-1 τον MLR αλγόριθμο, σε 3 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 48). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά πιο ακριβής από την απλή μέθοδο της τριπλής διασταυρωμένης επιλογής (BestCV), και την συσσωρευμένη γενίκευση (stacking) χρησιμοποιώντας σαν αλγόριθμο επιπέδου-1 τον M5' αλγόριθμο, σε 1 σύνολο δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Συνοψίζοντας, η απόδοση της προτεινόμενης τεχνικής είναι πιο ακριβής από τις άλλες γνωστές τεχνικές δημιουργίας ομάδας ταξινομητών, αν και χρησιμοποιεί λιγότερο χρόνο για εκπαίδευση. Η εκλεκτική ψηφοφορία μπορεί να επιτύχει μια αύξηση στην ακρίβεια ταξινόμησης 1.5% σε σχέση με την απλή ψηφοφορία. Επίσης, η μέση σχετική βελτίωση της ακρίβειας της προτεινόμενης μεθοδολογίας είναι από 1% έως 4% σε σχέση με τις υπόλοιπες μεθόδους.

Δεδομένα	Selective voting	voting	Grading	Stacking (MLR)	Stacking (M5')	BestCV
autos	81.77	81.62	79.08	50.33 *	77.2	79
badges	100	100	100	100	100	100
breast-cancer	72.67	71.13	72.73	72.17	72.56	71.65
breast-w	96.71	96.71	96.84	96.72	96.67	96.45
credit-a	86.2	86.14	85.75	85.8	85.41	84.77
credit-g	75.3	75.63	75.55	75.48	75.18	74.61
diabetes	76.59	77.17	76.78	76.95	75.98	76.8
Glass	72.52	71.98	71.56	50.79 *	69.25	68.21
haberman	73.13	72.58	73.57	72.78	72.09	73.43
heart-c	84.45	82.7	83.2	76.4	81.86	82.51
heart-h	83.79	83.48	83.48	66.03 *	81.81	83.04
heart-statlog	84.19	84.04	83.41	84.33	84.3	83.59
hepatitis	84.65	84.07	83.24	82.97	82.71	82.85
ionosphere	90.69	91.94	91.46	91.57	91.19	89.06
Iris	95.6	95.8	95.33	94.93	95.53	94.93
Labor	93.4	94.7	94.13	91.6	92.07	91.2
lymphography	83.72	84.56	85.15	82.03	82.03	84.24
monk1	92.48	86.33*	85.51*	86.31*	86.77*	83.88*
monk2	75.7	60.7 *	66.1 *	79.67	74.29	73.41
monk3	93.45	93.28	93.45	93.45	92.87	93.45
Mushroom	100	100	100	100	100	100
sonar	83.76	81.66	82.9	85.48	84.1	82.23
soybean	93.71	93.65	94.17	91.5	93.34	92.84
Vehicle	74.18	74.43	75.17	75.84	74.97	73.85
Vote	96.16	95.77	95.86	96.06	95.31	95.84
vowel	96.99	92.92*	93.66 *	93.12	96.33	96.99
wine	98.47	97.97	98.42	97.41	97.3	97.69
		0/24/3	0/24/3	0/24/3	0/26/1	0/26/1
<i>Μέση ακρίβεια</i>	86.68	85.59	85.80	83.32	85.60	85.43

Πίνακας 48. Σύγκριση της προτεινόμενης μεθόδου με άλλες ομάδες ταξινομητών

Σε αυτό το σημείο πρέπει να τονιστεί ότι η μείωση των μεγάλων συνόλων δεδομένων σε συμπαγέστερα αντιπροσωπευτικά υποσύνολα, διατηρώντας την ίδια εκμεταλλεύσιμη γνώση, θα μπορούσε να επιταχύνει την διαδικασία μάθησης και να μειώσει τις απαιτήσεις για αποθήκευση. Η δειγματοληψία είναι γενικά αποδεκτή από την στατιστική κοινότητα, η οποία παρατηρεί ότι "μια προσεγμένη μέθοδος που λειτουργεί σε ένα υποσύνολο των στοιχείων μπορεί στην πραγματικότητα να παρέχει ανώτερη ακρίβεια από μια λιγότερο προσεγμένη μέθοδο χρησιμοποιώντας ολόκληρη τη βάση δεδομένων" (Friedman, 1997).

Στην πράξη, καθώς το πλήθος των δεδομένων αυξάνεται, το ποσοστό αύξησης στην ακρίβεια επιβραδύνεται, διαμορφώνοντας την καμπύλη μάθησης (learning curve) (βλέπε Εικόνα 58). Αν η δειγματοληψία είναι αποτελεσματική εξαρτάται από το πόσο εντυπωσιακά το ποσοστό αύξησης επιβραδύνεται.

Εικόνα 58. Καμπύλη μάθησης

Είναι ενδιαφέρον ότι τα αποτελέσματά μας δείχνουν ότι η προτεινόμενη μέθοδος παραγάγει καλά αποτελέσματα όσον αφορά την καμπύλη μάθησης. Από ένα σύνολο δεδομένων μεγέθους n , απλά επιλέγουμε t τυχαίες περιπτώσεις (ένα ποσοστό των περιπτώσεων n) από το σύνολο δεδομένων χωρίς επανατοποθέτηση. Έπειτα αυτές οι t περιπτώσεις χρησιμοποιούνται για την εκπαίδευση της ομάδας των ταξινομητών μας. Στις τελευταίες σειρές του Πίνακα 49 κάποιος μπορεί να δει τα συγκεντρωτικά αποτελέσματα χρησιμοποιώντας διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης για τον προτεινόμενο αλγόριθμο (χρησιμοποιώντας το 33% και το 50% των περιπτώσεων του κάθε συνόλου δεδομένων).

Η προτεινόμενη μέθοδος χρησιμοποιώντας το πλήρες σύνολο κατάρτισης είναι στατιστικά ακριβέστερη από την ίδια μέθοδο χρησιμοποιώντας το 33% του συνόλου κατάρτισης σε 8 από τα 27 σύνολα δεδομένων. Πρέπει όμως να σημειωθεί ότι η απόδοση της προτεινόμενης μεθόδου χρησιμοποιώντας το 33% του συνόλου δεδομένων είναι συγκρίσιμη με την απόδοση των ταξινομητών βάσης χρησιμοποιώντας το σύνολο των δεδομένων. Κατά συνέπεια, εκμεταλλευόμενη την δυνατότητα λειτουργίας με λιγότερα δεδομένα και την παράλληλη ή καταναμημένη εκτέλεση (ένας αλγόριθμος βάσης ανά μηχανή) η επιλεκτική ψηφοφορία μπορεί να αυξήσει την ταχύτητα ενός συστήματος μάθησης χωρίς απώλειες στην ακρίβεια ταξινόμησης..

Η προτεινόμενη μέθοδος χρησιμοποιώντας το πλήρες σύνολο κατάρτισης είναι στατιστικά ακριβέστερη από την ίδια μέθοδο χρησιμοποιώντας το 50% των περιπτώσεων κατάρτισης σε μόνο 3 από τα 27 σύνολα δεδομένων. Η απόδοση της προτεινόμενης μεθόδου χρησιμοποιώντας το 50% του συνόλου κατάρτισης είναι αρκετά καλύτερη από την απόδοση των ταξινομητών βάσης χρησιμοποιώντας το σύνολο των δεδομένων. Συνεπώς, ακόμη και στην περίπτωση της τυχαίας επιλογής υποσυνόλων παραδειγμάτων εκπαίδευσης, όπου οποιοσδήποτε μεμονωμένος ταξινομητής που δημιουργείται σε ένα υποσύνολο αποδίδει συχνά στατιστικά χειρότερα από τον ίδιο ταξινομητή που εκπαιδεύεται σε ολόκληρο το σύνολο δεδομένων, η εκλεκτική ψηφοφορία μπορεί να βελτιώσει την απόδοση.

Είναι σαφές ότι η αναζήτηση του βέλτιστου ποσοστού δειγματοληψίας του συνόλου δεδομένων πρέπει να λάβει υπόψη της όχι μόνο τις ιδιότητες (meta-characterizations) του συνόλου των δεδομένων αλλά και τις προτιμήσεις του χρήστη όσον αφορά τα κριτήρια απόδοσης όπως το χρόνο εκπαίδευσης, την ακρίβεια και το μέγεθος του παραγόμενου μοντέλου. Περισσότερη έρευνα χρειάζεται προς αυτήν την κατεύθυνση.

Δεδομένα	Selective Voting (100%)	Selective Voting (33%)	Selective Voting (50%)
autos	81.77	54.19 *	61.92 *
badges	100	100	100
breast-cancer	72.67	71.17	71.95
breast-w	96.71	96.35	96.82
credit-a	86.2	85.16	85.77
credit-g	75.3	73.24	74.47
diabetes	76.59	75.93	76.37
Glass	72.52	64.38 *	68.66
haberman	73.13	73.88	73.59
heart-c	84.45	83.15	83.45
heart-h	83.79	82.42	82.15
heart-statlog	84.19	81.7	83
hepatitis	84.65	81.8	82.4
ionosphere	90.69	88.69	89.15
Iris	95.6	94.6	94.67
Labor	93.4	85.23	87.97
lymphography	83.72	79.91	82.34
monk1	92.48	78.85 *	84.17
monk2	75.7	53.26 *	57.06 *
monk3	93.45	89.96	92.21
Mushroom	100	99.98	100
sonar	83.76	71.3 *	77.54
soybean	93.71	88.37 *	91.79
Vehicle	74.18	69.86 *	71.84
Vote	96.16	95.45	95.51
vowel	96.99	46.15 *	64.69 *
wine	98.47	96.13	97.46
		0/19/8	0/24/3
<i>Μέση ακρίβεια</i>	86.68	80.04	82.48

Πίνακας 49. Εκτέλεση της μεθόδου επιλεκτικής ψηφοφορίας σε υποσύνολα των δεδομένων

4.2.4 Προτεινόμενη μέθοδος βελτιωμένης συσσωρευμένης γενίκευσης

Στην πορεία της διατριβής δημιουργήσαμε ένα υβριδικό σύστημα υποστήριξης απόφασης που συνδυάζει ένα νευρωνικό δίκτυο, ένα δέντρο απόφασης και έναν Μπεϋζιανό αλγόριθμο χρησιμοποιώντας μια βελτιωμένη μεθοδολογία συσσωρευμένης γενίκευσης. Συγκρίναμε την προτεινόμενη μέθοδο με άλλες τεχνικές χρησιμοποιώντας τους ίδιους ταξινομητές σαν βασικούς ταξινομητές σε διάφορα σύνολα δεδομένων και πήραμε καλύτερη ακρίβεια στις περισσότερες περιπτώσεις.

Στην προτεινόμενη τεχνική τα δεδομένα επιπέδου-1 δεν είναι απλά οι προβλέψεις των μελών, αλλά τα μέλη δίνουν ως έξοδο το βαθμό εμπιστοσύνης τους για την κατάταξη του στιγμιότυπου σε κάθε μία από τις κλάσεις. Ακολουθήσαμε δηλαδή την πρόταση των (Ting & Witten, 1999). Στον αλγόριθμο μας όμως, τα δεδομένα του επιπέδου-1 αποτελούνται μόνο από το βαθμό εμπιστοσύνης που δίνει το κάθε μέλος στην πραγματική κλάση και όχι σε όλες τις κλάσεις σε αντίθεση με τους (Ting & Witten, 1999). Έτσι μειώνουμε το χρόνο εκπαίδευσης του μετα-ταξινομητή και επιπλέον μειώνοντας τα χαρακτηριστικά που λαμβάνει

υπ' όψιν του είναι αναγκαία λιγότερα παραδείγματα για εκπαίδευση. Τέλος χρησιμοποιήσαμε ως πρόεδρο τον αλγόριθμο M5' και όχι τον MLR (Seewald, 2002). Η σχηματική αναπαράσταση της προτεινόμενης τεχνικής παραθέτεται στην Εικόνα 59.

Εικόνα 59. Σχηματική αναπαράσταση της βελτιωμένης συσσωρευμένης γενίκευσης (Stacking')

Η προτεινόμενη μέθοδος είναι γενική και μπορεί εφαρμοστεί με οποιουσδήποτε ταξινομητές επιπέδου-0. Για τα πειράματά μας εμείς χρησιμοποιήσαμε τον BP ως εκπρόσωπο των νευρωνικών δικτύων, τον C4.5 ως εκπρόσωπο των δέντρων απόφασης και τον NB ως εκπρόσωπο των Μπεϋζιανών μεθόδων. Ο κύριος στόχος ήταν να καταδειχθούν οι δυνατότητες της συγκεκριμένης συνδυαστικής χρήσης ταξινομητών και το συγκριτικό πλεονέκτημα της τελευταίας ως προς κάθε μέλος χωριστά. Διαφορετικές επιλογές ταξινομητών βάσης, πάντως, είναι πιθανό να οδηγούν σε ακόμα υψηλότερες επιδόσεις. Ακολουθεί ο αλγόριθμος της βελτιωμένης συσσωρευμένης γενίκευσης (Stacking') σε ψευδογλώσσα.

- Ένα τμήμα είναι το σύνολο ελέγχου TSi, στο οποίο θα εξεταστεί ο πρόεδρος και τα υπόλοιπα 9 αποτελούν το σύνολο εκπαίδευσης TRi. Στη συνέχεια γίνεται εσωτερικά στο TRi διασταυρωμένη επικύρωση 3 (γενικά n) σημείων.
- Στο βήμα j της εσωτερικής διασταυρωμένης επικύρωσης:
 - Το 1/3 του TRi αποτελεί το σύνολο επικύρωσης V_{ij} και τα υπόλοιπα 2/3 το σύνολο εκπαίδευσης T_{ij} των μελών της επιτροπής.
 - Τα μέλη της επιτροπής (NB, C4.5 και BP) εκπαιδεύονται ξεχωριστά πάνω στο T_{ij} και οι ταξινομητές C_{ij} που κατασκευάζονται εξετάζονται στο V_{ij}.
 - Για κάθε στιγμιότυπο του V_{ij} κατασκευάζεται ένα νέο, το οποίο περιλαμβάνει τους βαθμούς εμπιστοσύνης των C_{ij} ως προς την σωστή κλάση
- Έτσι σχηματίζεται το σύνολο V'_{ij} (δεδομένα επιπέδου-1).
- Ο πρόεδρος εκπαιδεύεται στο TR'_i

Εικόνα 60. Ο ψευδοκώδικας της βελτιωμένης συσσωρευμένης γενίκευσης

Κατά το πρώτο πείραμα ελέγχου της απόδοσης της προτεινόμενης μεθόδου, συγκρίναμε την προτεινόμενη μέθοδο με κάθε αλγόριθμο βάσης αλλά και τις γνωστές τεχνικές σακουλιάσματος και προώθησης σε συνεργασία με κάθε αλγόριθμο βάσης. Αναλυτικότερα, συγκρίναμε τον προτεινόμενο αλγόριθμο με τους ακόλουθους:

- C4.5, NB, BP
- Bagging C4.5, Bagging NB, Bagging BP (με 25 υπο-ταξινομητές)
- AdaBoost C4.5, AdaBoost NB, AdaBoost BP (με 25 υπο-ταξινομητές)

Για τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 50 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του Πίνακα 50 παρουσιάζεται η μέση ακρίβεια κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Ο προτεινόμενος αλγόριθμος (Stacking') είναι στατιστικά ακριβέστερος από τους NB, Bagging NB, BP και AdaBoost BP σε 7 σύνολα δεδομένων (datasets), ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 50).

Δεδομένα	Stacking'	C4.5	NB	BP	Bagging C4.5	Bagging NB	Bagging BP	Boost C4.5	Boost NB	Boost BP
Anneal	98.79	98.57	86.59*	92.78*	98.83	86.94*	93.52*	99.60	95.20*	92.81*
audiology	74.92	77.26	72.64	43.82*	81.29	72.10	46.50*	84.62v	78.20	43.74*
autos	80.95	81.77	57.41*	54.38*	83.85	57.15*	56.39*	86.05	57.12*	56.36*
balance	90.37	77.82*	90.53	87.09*	82.33*	90.29	86.18*	76.91*	92.11	87.19*
breast-w	96.12	95.01	96.07	95.97	96.31	96.07	96.44	96.51	95.55	95.90
colic	84.72	85.16	78.70*	82.58	85.23	78.94*	83.75	82.01	77.46*	79.44*
credit-g	74.77	71.25*	75.16	72.75	74.17	75.13	74.97	72.79	75.09	73.96
diabetes	76.41	74.49	75.75	76.56	75.67	75.57	76.84	72.81*	75.88	76.5
haberman	73.20	71.05	75.06	74.64	72.06	74.86	75.69	71.12	73.94	74.51
heart-c	83.08	76.94*	83.34	81.39	79.54	83.24	82.81	79.60	83.14	80.99
heart-h	83.75	80.22	83.95	81.37	79.91	84.16	83.27	78.28	84.67	81.30
heart-stat	83.48	78.15	83.59	82.11	81.11	83.41	83.22	80.15	82.30	81.07
hepatitis	83.09	79.22	83.81	81.30	81.63	84.39	84.25	82.74	84.23	82.64
ionosphere	91.31	89.74	82.17*	85.84*	92.23	81.94*	87.09	93.62	91.12	89.09
iris	95.67	94.73	95.53	96.27	94.80	95.53	96.67	94.47	95.07	95.47
lymp/rapy	81.15	75.84	83.13	80.24	79.14	83.50	82.75	83.09	80.67	80.70
mushroom	100.0	100.0	95.76*	100.0	100.0	95.56*	100.0	100.0	100.0	100.0
vehicle	74.09	72.28	44.68*	50.07*	74.91	45.73*	55.03*	77.16	44.68*	50.32*
sonar	76.90	73.61	67.71	73.30	79.03	67.96	76.98	83.03	81.21	79.77
vote	96.64	96.57	90.02*	94.60	96.53	90.09*	95.17	95.24	95.19	95.45
zoo	97.13	92.61	94.97	61.21*	93.29	95.07	72.97*	95.38	97.23	61.21*
		0/18/3	0/14/7	0/14/7	0/20/1	0/14/7	0/15/6	1/18/2	0/17/4	0/14/7
Μέση ακρίβεια	85.55	82.97	80.79	78.49	84.85	80.84	80.50	84.99	82.86	78.97

Πίνακας 50. Σύγκριση της προτεινόμενης μεθόδου Stacking' με τους βασικούς αλγόριθμους και ομάδες ταξινομητών των αλγορίθμων βάσης

Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους C4.5 και Bagging C4.5 σε 3 και 1 σύνολα δεδομένων, αντίστοιχα, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους Bagging BP και AdaBoost NB σε 6 και 4 σύνολα

δεδομένων, αντίστοιχα, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Τέλος, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον AdaBoost C4.5 σε 2 σύνολα δεδομένων, αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 1 σύνολο δεδομένων.

Επίσης, συγκρίναμε την προτεινόμενη μέθοδο χρησιμοποιώντας τους ίδιους αλγόριθμους βάσης με άλλες τεχνικές συνδυασμού ταξινομητών όπως:

- Ψηφοφορία (voting)
- Βαθμολόγησης (grading)
- Συσσωρευμένης γενίκευσης (stacking) χρησιμοποιώντας την μία φορά τον αλγόριθμο MLR ως αλγόριθμο μετα-μάθησης όπως πρότειναν στην εργασία τους οι (Ting & Witten, 1999) και την άλλη τον αλγόριθμο M5' (έναν αλγόριθμο παλινδρόμησης που βασίζεται στα δέντρα αποφάσεων) όπως πρότειναν στην εργασία τους οι (Dzeroski & Zenko, 2002).
- Την απλή μέθοδο της τριπλής διασταυρωμένης επιλογής (BestCV) που περιγράφηκε στην ενότητα 4.2.1.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους υπόλοιπους σε 1 σύνολο δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 51).

Δεδομένα	Stacking'	BestCV	Voting	Grading	Stacking (MLR)	Stacking (M5')
anneal	98.79	98.57	98.92	98.90	98.53	98.85
audiology	74.92	77.26	74.93	74.31	72.83	72.41
autos	80.95	81.77	78.11	78.04	82.40	80.37
balance	90.37	90.53	89.87	90.08	87.11*	93.29
breast-w	96.12	95.84	96.45	96.55	96.20	96.10
colic	84.72	84.99	84.18	84.54	84.56	84.69
credit-g	74.77	75.11	74.75	74.52	74.76	74.74
diabetes	76.41	75.74	76.77	75.94	76.63	76.41
haberman	73.20	73.99	74.70	73.70	72.55	73.20
heart-c	83.08	82.85	82.57	82.75	83.21	83.08
heart-h	83.75	83.95	83.24	83.31	83.95	83.82
heart-statlog	83.48	83.44	83.07	82.93	84.22	83.52
hepatitis	83.09	82.63	82.97	82.64	83.16	83.09
ionosphere	91.31	89.77	92.25	92.05	91.31	90.92
iris	95.67	94.93	96.00	95.93	95.13	95.73
lymphotherapy	81.15	81.56	82.32	82.40	80.63	80.15
mushroom	100.0	100.0	100.0	100.0	100.0	100.0
vehicle	74.09	72.28	72.72	72.25*	74.23	71.05*
sonar	76.90	73.38	78.71	78.71	76.71	76.71
vote	96.64	96.57	96.18	95.91	96.57	96.52
zoo	97.13	93.59*	94.78*	96.15	96.05	96.83
W/D/L		0/20/1	0/20/1	0/20/1	0/20/1	0/20/1
Μέση ακρίβεια	85.55	85.18	85.40	85.31	85.27	85.31

Πίνακας 51. Σύγκριση της προτεινόμενης μεθόδου (Stacking') με άλλες ομάδες ταξινομητών

4.2.5 Προτεινόμενη μέθοδος τοπικής ψηφοφορίας απλών ταξινομητών

Στη συνέχεια της διατριβής χρησιμοποιήσαμε τοπική ψηφοφορία μιας ομάδας ταξινομητών. Ο κύριος λόγος για την σκέψη μας αυτή ήταν ότι η τοπική ψηφοφορία μπορεί να προσεγγίσει πολύ καλύτερα από άλλες μεθόδους τη συνάρτηση-στόχο αν αυτή είναι πολύπλοκη καθολικά, αλλά μπορεί να περιγραφεί ως μια συλλογή λιγότερο σύνθετων τοπικών προσεγγίσεων. Τα πειράματα της μελέτης μας απέδειξαν ότι η προτεινόμενη μέθοδος δίνει καλύτερα αποτελέσματα από την απλή ψηφοφορία και άλλες πιο περίπλοκες μεθόδους [4]. Στην Εικόνα 61 παρουσιάζεται ο προτεινόμενος αλγόριθμος.

1. Καθόρισε την κατάλληλη μετρική απόστασης (distance metric).
2. Βρες τους k κοντινότερους γείτονες χρησιμοποιώντας αυτή την μετρική.
3. Εφάρμοσε την τεχνική της ψηφοφορίας σε απλούς ταξινομητές χρησιμοποιώντας τα k παραδείγματα ως δεδομένα εκπαίδευσης.
4. Η απάντηση της ομάδας των ταξινομητών που δημιουργήθηκαν είναι η κλάση του παραδείγματος που θέλουμε να ταξινομήσουμε.

Εικόνα 61. Ο προτεινόμενος αλγόριθμος τοπικής ψηφοφορίας

Ο προτεινόμενος αλγόριθμος έχει μερικές ελεύθερες παραμέτρους όπως η μετρική απόστασης. Στα πειράματά μας, χρησιμοποιήσαμε την πιο συχνά χρησιμοποιούμενη μετρική - την Ευκλείδεια απόσταση. Χρησιμοποιήσαμε το 50 ως την τιμή του k, δεδομένου ότι αυτό το μέγεθος των περιπτώσεων είναι κατάλληλο ώστε ένας απλός αλγόριθμος να χτίζει ένα ακριβές μοντέλο (Frank και άλλοι, 2003).

Η προτεινόμενη μέθοδος είναι γενική και μπορεί εφαρμοστεί με οποιουσδήποτε ταξινομητές βάσης. Για τα πειράματα χρησιμοποιήσαμε τρεις πολύ γνωστούς απλούς αλγορίθμους μάθησης: OneR, Decision Stump και NB. Κάθε ταξινομητής (NB, OneR, DS) παράγει μια υπόθεση h_1 , h_2 , h_3 αντίστοιχα, χρησιμοποιώντας τα 50 κοντινότερα παραδείγματα στο υπο-ταξινόμηση στιγμιότυπο (βλέπε Εικόνα 62). Αν οι πιθανότητες πρόβλεψης που παράγονται από τους μεμονωμένους ταξινομητές είναι αντίστοιχα $p_1(i)$, $p_2(i)$, $p_3(i)$ για κάθε κλάση i , τότε η κατηγορία που συγκεντρώνει την μεγαλύτερη τιμή πιθανότητας λαμβάνονται ως απάντηση (predictive class) της ψηφοφορίας (h^*). Αναλυτικότερα, η κατηγορία υπολογίζεται από τον

κανόνα:
$$\text{Predictive Class} = \arg \max_i \sum_{j=1, j=1}^{\text{number of classes, } j=3} p_j(i)$$

Εικόνα 62. Η γραφική αναπαράσταση της μεθόδου τοπικής ψηφοφορίας

Όπως είναι γνωστό, οι μέθοδοι ψηφοφορίας αποτυγχάνουν εάν οι απλοί ταξινομητές δεν μπορούν να επιτύχουν ακρίβεια τουλάχιστον 50% για το συγκεκριμένο σύνολο δεδομένων. Για αυτόν τον λόγο, η προτεινόμενη μέθοδος μετατρέπει τα πολυ-κατηγορικά (multi-class) προβλήματα σε ένα σύνολο δυαδικών προβλημάτων. Έχουμε χρησιμοποιήσει την μέθοδο μια κλάση εναντίων όλων (One Per Class). Ο ταξινομητής i εκπαιδεύεται με όλα τα παραδείγματα της κατηγορίας i ως θετική κλάση, και όλα τα άλλα παραδείγματα ως αρνητική κλάση¹⁴. Η τελική απόφαση είναι η κλάση με την μεγαλύτερη εμπιστοσύνη πρόγνωσης (confidence value).

Κατά το πρώτο πείραμα, συγκρίναμε την προτεινόμενη μέθοδο με τους αλγόριθμους βάσης (χρησιμοποιώντας καθολική ή τοπική μάθηση). Επίσης, χρησιμοποιήσαμε και την τοπική μέθοδο 50NN. Αναλυτικότερα, συγκρίναμε την προτεινόμενη μέθοδο με τους αλγόριθμους:

- NB, DS, OneR
- Local NB, Local DS, Local OneR
- 50NN

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους Local DS και Local OneR σε 11 και 10 σύνολα δεδομένων (datasets), αντίστοιχα, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 52). Ομοίως, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους DS και OneR σε 21 και 20 σύνολα δεδομένων, αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους και από τους δυο αλγόριθμους σε μόλις 1 σύνολο δεδομένων. Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους NB και Local NB σε 10 και 6 σύνολα δεδομένων (datasets), αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 5 και 4 σύνολα δεδομένων, αντίστοιχα. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον 50NN σε 14 σύνολα δεδομένων αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων.

Για τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 52 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του Πίνακα 52 παρουσιάζεται η μέση ακρίβεια κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

¹⁴ Περισσότερα στοιχεία για την μέθοδο μετατροπής των πολυ-κατηγορικών προβλημάτων σε ένα σύνολο δυαδικών προβλημάτων στην ενότητα 4.3

Δεδομένα	Local voting	Local NB	Local DS	Local OneR	NB	DS	OneR	50NN
audiology	82,27	77,25 *	70,59 *	72,31 *	72,64 *	46,46 *	46,46 *	35,95 *
autos	81,07	75,93 *	69,86 *	76,49	57,41 *	44,9 *	61,77 *	48,18 *
Badges	99,86	96,54 *	99,97	28,55 *	99,66	100	28,55 *	99,69
balance-scale	88,7	90,18 v	82,9 *	87,73	90,53 v	56,72 *	57,09 *	89,01
breast-cancer	73,59	72,82	73,38	72,69	72,7	69,27	66,91 *	70,75
breast-w	96,45	96,32	96,22	96,27	96,07	92,33 *	92,01 *	95,9
colic	82,39	81,77	81,3	81,17	78,7	81,52	81,52	84,04
Credit-g	72,76	75,23 v	71,54	70,93 *	75,16	70 *	66,13 *	71,96
credit-rating	85,39	82,87 *	82,35 *	83,59	77,86 *	85,51	85,51	86,16
diabetes	71,9	71,84	72,36	69,76 *	75,75 v	71,8	71,98	74,68
Glass	75,48	72,66	69,44 *	68,52 *	49,45 *	44,89 *	56,84 *	56,16 *
grub-damage	43,86	44,7	44,28	39,01	49,76	33,2 *	42,15	44,21
haberman	69,89	71,3	70,65	68,88	75,06 v	71,57	72,53	72,91
heart-c	80,53	81,36	78,68	79,23	83,34	72,93 *	72,53 *	81,58
heart-h	79,68	80,66	78,31	79,14	83,95	81,78	80,69	83,98 v
heart-statlog	78,85	80,41	74,63	76,3	83,59 v	72,3 *	71,26 *	83,74 v
hepatitis	84,71	86,18	83,76	82,02	83,81	77,62 *	82,05	79,38
ionosphere	88,24	81,91 *	87,56	88,24	82,17 *	82,57 *	82,59 *	71,65 *
Iris	94,53	95,67	93,8	94	95,53	66,67 *	93,53	90,53
lymphotherapy	82	82,95	75,61 *	79,6	83,13	75,31	74,77	80,59
monk3	92,4	91,66	92,64	92,4	93,45	76,01 *	77,88 *	82,46 *
Primary-tumor	44,9	44,46	43,28	43,8	49,71 v	28,91 *	27,74 *	39,26 *
Sonar	82,45	86,6	77,81 *	73,57 *	67,71 *	72,25 *	62,12 *	68,25 *
soybean	94,86	93,29 *	91,98 *	92,25 *	92,94 *	27,96 *	39,75 *	62,34 *
Students	84,6	83,38	84,02	84,36	85,7	87,22 V	87,22 v	78,89 *
Titanic	78,95	78,95	78,94	78,95	77,85	77,6 *	77,6 *	77,56 *
Vehicle	72,01	74,94 v	69,12 *	67,37 *	44,68 *	39,81 *	52,36 *	63,47 *
Vote	96	95,49	95,88	95,47	90,02 *	95,63	95,63	90,41 *
Vowel	91,23	94,56 v	36,35 *	74,81 *	62,9 *	17,47 *	33,05 *	8,88 *
Wine	97,07	98,92	95,67	94,83	97,46	57,91 *	77,93 *	96,46
Zoo	97,11	97,53	88,05 *	42,59 *	94,97	60,43 *	42,59 *	55,11 *
Μέση ακρίβεια	82,06	4/21/6	0/20/11	0/21/10	5/16/10	1/9/21	1/10/20	2/15/14
		81,88	77,77	75,32	78,18	65,76	66,48	71,75

Πίνακας 52. Η σύγκριση του αλγορίθμου τοπικής ψηφοφορίας με τους αλγόριθμους βάσης χρησιμοποιώντας καθολική ή τοπική μάθηση

Εν συνεχεία, συγκρίναμε την προτεινόμενη μέθοδο με την μέθοδο της απλής προώθησης και την μέθοδο της τοπικής προώθησης (Local Boost) κάθε αλγορίθμου βάσης που προτείναμε στην ενότητα 4.1.2. Αναλυτικότερα συγκρίναμε την προτεινόμενη μέθοδο τις ακόλουθες τεχνικές:

- Adaboost NB, Adaboost DS, Adaboost OneR
- Local boost NB, Local boost DS, Local boost OneR

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους Adaboost DS και Adaboost NB σε 11 και 6 σύνολα δεδομένων (datasets), αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 και 4 σύνολα δεδομένων, αντίστοιχα (βλέπε Πίνακα 53). Ομοίως, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους Adaboost OneR και Local Boost DS σε 15 και 5 σύνολα δεδομένων, αντίστοιχα, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Παράλληλα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους Local Boost NB και Local Boost OneR σε 8 και 5 σύνολα δεδομένων (datasets), αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 και κανένα σύνολα δεδομένων, αντίστοιχα.

Δεδομένα	Local Voting	Adaboost DS	Adaboost NB	Adaboost OneR	Local boost DS	Local boost NB	Local boost OneR
Audiology	82,27	46,46 *	78,2	46,46 *	73,12 *	77,22 *	74,53 *
Autos	81,07	44,9 *	57,12 *	65,47 *	75,49	77,48	80,87
Badges	99,86	100	99,66	28,55 *	100	93,68 *	28,55 *
balance-scale	88,7	71,77 *	92,11 V	72,81 *	84,82 *	84,33 *	88,3
breast-cancer	73,59	71,55	68,57	69,49	72,75	68,36 *	72,58
breast-w	96,45	95,28	95,55	95,55	95,9	95,91	96
Colic	82,39	82,72	77,46	81,17	78,91	77,47 *	80,05
Credit-g	72,76	72,6	75,09	64,35 *	72,92	71,09	71,46
credit-rating	85,39	85,57	81,16 *	81,86 *	84,94	80,17 *	82,45
Diabetes	71,9	75,37 v	75,88 v	69,91	73,77	71,22	70,2
Glass	75,48	44,89 *	49,63 *	56,56 *	70,39	71,88	71,7
grub-damage	43,86	33,2 *	49,9	41,22	42,43	40,48	39,65
Haberman	69,89	74,06	73,94	71,65	68,9	68,47	65,73
heart-c	80,53	83,11	83,14	72,61 *	80,4	78,62	77,33
heart-h	79,68	82,42	84,67 v	76,59	79,06	78,49	78,04
heart-statlog	78,85	81,81	82,3	73,59	78,15	78,07	77,78
Hepatitis	84,71	81,5	84,23	77,41	84,45	83,98	82,62
ionosphere	88,24	92,34 v	91,12	88,43	90,01	86,81	90,09
Iris	94,53	95,07	95,07	93,67	94,47	95,87	93,33
lymphography	82	75,44	80,67	79,83	84,6	85,58	84,63
monk3	92,4	90,92	91,07	90,67	90,68	87,21	90,07
primary-tumor	44,9	28,91 *	49,71 v	27,38 *	43,22	44,45	44,04
Sonar	82,45	81,06	81,21	65,43 *	83,89	86,47	77,85
Soybean	94,86	27,96 *	92,02 *	40,48 *	92,99 *	91,95 *	93,05 *
Students	84,6	87,16	85,12	86,52	82,86	80,4 *	82,83
Titanic	78,95	77,83	77,86	77,74	79,05	78,95	78,97
Vehicle	72,01	39,81 *	44,68 *	51,63 *	70,98	77,77 v	70,89
Vote	96	96,41	95,19	96,62	96,02	94,62	95,79
Vowel	91,23	17,47 *	81,32 *	31,13 *	76,46 *	95,66 v	81,03 *
Wine	97,07	91,57 *	96,57	92,07 *	97,47	98,65	96,46
Zoo	97,11	60,43 *	97,23	42,59 *	88,84 *	97,31	42,59 *
		2/18/11	4/21/6	0/16/15	0/26/5	2/21/8	0/26/5
Μέση ακρίβεια	82,06	70,63	79,60	68,05	80,26	80,60	76,11

Πίνακας 53. Η σύγκριση του αλγορίθμου τοπικής μάθησης με άλλες ομάδες ταξινομητών

Τέλος, συγκρίναμε την προτεινόμενη μέθοδο χρησιμοποιώντας τους ίδιους αλγόριθμους βάσης με άλλες τεχνικές συνδυασμού ταξινομητών όπως:

- Ψηφοφορία (voting)
- Συσσωρευμένης γενίκευσης (stacking) χρησιμοποιώντας τον αλγόριθμο MLR ως αλγόριθμο μετα-μάθησης όπως πρότειναν στην εργασία τους οι (Ting & Witten, 1999).
- Την απλή μέθοδο της τριπλής διασταυρωμένης επιλογής (BestCV) που περιγράφηκε στην ενότητα 4.2.1.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Best-CV σε 8 σύνολα δεδομένων (datasets), ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 4 σύνολα δεδομένων (βλέπε Πίνακα 54). Ομοίως, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από την συσσωρευμένη γενίκευση και την ψηφοφορία σε 15 και 5 σύνολα δεδομένων, αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 και 1 σύνολα δεδομένων, αντίστοιχα.

Δεδομένα	Local Voting	Best-3CV	Stacking (MLR)	Voting
audiology	82,27	72,64 *	66,11 *	46,46 *
Autos	81,07	57,43 *	73,78 *	66,45 *
Badges	99,86	99,83	100	99,66
balance-scale	88,7	90,53 v	90,03	57,28 *
breast-cancer	73,59	71,19	72,24	67,26 *
breast-w	96,45	96,07	95,99	94,48
Colic	82,39	81,3	83,88	81,52
Credit-g	72,76	75,16	75,13	69,82
credit-rating	85,39	85,51	85,51	85,51
Diabetes	71,9	75,7 v	75,62 v	72,72
Glass	75,48	55,22 *	63,48 *	60,21 *
grub-damage	43,86	49,76	48,48	41,89
haberman	69,89	74,4	73,07	72,56
heart-c	80,53	83,34	83,51	73,19 *
heart-h	79,68	83,95	83,95	81,58
heart-statlog	78,85	83,59 v	83,63 v	72,67
Hepatitis	84,71	83,28	84,32	82,18
ionosphere	88,24	82,02 *	89,89	90,03
Iris	94,53	95,13	95,13	93,53
lymphography	82	83,13	82,25	75,91
monk3	92,4	93,45	93,45	78,45 *
primary-tumor	44,9	49,71 v	46,64	28,21 *
Sonar	82,45	68,74 *	69,75 *	68,98 *
Soybean	94,86	92,94 *	93,06	64,2 *
Students	84,6	86,98	86,99	87,22 V
Titanic	78,95	77,76	77,83	77,6 *
Vehicle	72,01	52,4 *	58,06 *	46,37 *
Vote	96	95,63	95,56	95,63
Vowel	91,23	62,9 *	64,14 *	35,74 *
Wine	97,07	97,46	97,63	86,02 *
Zoo	97,11	94,97	94,87	95,76
		4/19/8	2/23/6	1/16/14
Μέση ακρίβεια	82,06	79,10	80,13	72,55

Πίνακας 54. Η σύγκριση του αλγορίθμου τοπικής ψηφοφορίας με άλλες περίπλοκες ομάδες ταξινομητών

Η παρακάτω εικόνα παρουσιάζει τη βελτίωση της τοπικής ψηφοφορίας σε σχέση με την απλή ψηφοφορία υπολογίζοντας: (ακρίβεια της τοπικής ψηφοφορίας)/(ακρίβεια της απλής ψηφοφορίας) ως συνάρτηση του κάθε συνόλου δεδομένων. Στα δεδομένα όπου η κουκίδα εμφανίζεται πάνω από τη γραμμή, η τοπική ψηφοφορία είναι καλύτερη από την απλή ψηφοφορία και το αντίστροφο.

Εικόνα 63. Συγκρίνοντας την τοπική ψηφοφορία με την απλή ψηφοφορία

Όπως φαίνεται στην Εικόνα 63, η ακρίβεια της τοπικής ψηφοφορίας είναι σχεδόν ομοιόμορφα καλύτερη από αυτή της απλής ψηφοφορίας.

Ομοίως, η παρακάτω εικόνα παρουσιάζει τη βελτίωση της τοπικής ψηφοφορίας σε σχέση με την συσσωρευμένη γενίκευση υπολογίζοντας: $(\text{ακρίβεια της τοπικής ψηφοφορίας}) / (\text{ακρίβεια της συσσωρευμένης γενίκευσης})$ ως συνάρτηση κάθε συνόλου δεδομένων. Στα δεδομένα όπου η κουκίδα εμφανίζεται πάνω από τη γραμμή, η τοπική ψηφοφορία είναι καλύτερη από τη συσσωρευμένη γενίκευση και αντίστροφα.

Εικόνα 64. Συγκρίνοντας την τοπική ψηφοφορία με την συσσωρευμένη γενίκευση (stacking)

Σαν γενικό συμπέρασμα, η προσέγγισή μας αποδίδει καλύτερα από τις υπάρχοντες μεθόδους κατασκευής ομάδων ταξινομητών. Η μέση σχετική βελτίωση ακρίβειας της προτεινόμενης μεθοδολογίας είναι τουλάχιστον 2% σε σχέση με τις υπόλοιπες μεθόδους.

4.2.6 Προτεινόμενη μέθοδος γρήγορης ομάδας ταξινομητών

Όπως έχουμε ήδη αναφέρει, μια σημαντική ερευνητική περιοχή είναι η εύρεση τεχνικών μάθησης ικανών να λειτουργούν στις όλο και μεγαλύτερες βάσεις δεδομένων, για τις οποίες οι υπάρχοντες συνδυασμοί ταξινομητών έχουν μεγάλο υπολογιστικό κόστος. Στην πορεία της

ερευνάς μας, προσπαθήσαμε να γεφυρώσουμε το χάσμα χρησιμοποιώντας πολύ γρήγορους και απλούς αλγόριθμους για την παραγωγή μιας γρήγορης ομάδας ταξινομητών.

Το κίνητρο για να χρησιμοποιήσουμε απλούς ταξινομητές αντί ισχυρούς είναι ότι είναι υπολογιστικά ταχύτεροι. Ο χρόνος εκπαίδευσης για την παραγωγή πολλαπλάσιων απλών ταξινομητών είναι συχνά λιγότερος σε σχέση με την εκπαίδευση ενός ισχυρού ταξινομητή. Αυτό είναι επειδή, οι ισχυροί ταξινομητές ξοδεύουν την πλειοψηφία του χρόνου εκπαίδευσης στην εκλέπτυνση του επιθυμητού ορίου απόφασης (decision boundary), ενώ οι απλοί ταξινομητές παράγουν μόνο μια χοντρική προσέγγιση (rough approximation) του ορίου απόφασης. Έτσι, αν ομάδες απλών ταξινομητών μπορούσαν να επιτύχουν ικανοποιητική γενίκευση, θα παρείχαν μια καλή λύση στο πρόβλημα της ταυτόχρονης επίτευξης καλής επίδοσης και αποδοτικότητας όσον αφορά στον χώρο αλλά και στον χρόνο. Ένα άλλο κίνητρο για να χρησιμοποιήσουμε απλούς ταξινομητές αντί ισχυρούς είναι ότι δεν χρειάζονται παραμετροποίηση.

Μια ομάδα απλών ταξινομητών πρέπει να ικανοποιεί τους ακόλουθους δύο όρους: α) κάθε απλός ταξινομητής πρέπει να κάνει καλύτερη πρόγνωση από την τυχαία εικασία και β) το σύνολο ταξινομητών πρέπει να έχει αρκετή υπολογιστική δυνατότητα να μάθει το πρόβλημα. Ο πρώτος όρος εξασφαλίζει ότι κάθε απλός ταξινομητής κατέχει την ελάχιστη υπολογιστική δυνατότητα. Ο δεύτερος όρος εξασφαλίζει ότι οι μεμονωμένοι απλοί ταξινομητές μαθαίνουν διαφορετικά μέρη του προβλήματος έτσι ώστε η ομάδα των απλών ταξινομητών να μπορεί να μάθει το συνολικό πρόβλημα.

Στην Εικόνα 65 δείχνουμε πως ο συνδυασμός δυο απλών γραμμικών ταξινομητών μπορεί να διαχωρίσει μη γραμμικώς διαχωρίσιμα προβλήματα.

Εικόνα 65. Ο συνδυασμός γραμμικών διαχωρίσεων (linear decision boundaries) δίνει μη γραμμική διαχώριση (non-linear boundary)

Για τον αλγόριθμο μας (VoteDVN) χρησιμοποιήσαμε την αθροιστική ψηφοφορία για τον συνδυασμό των τριών πιο γνωστών απλών αλγόριθμων μάθησης: Decision Stump, VFI¹⁵ (Demiroz & Guvenir, 1997) και NB (βλέπε Εικόνα 66). Κάθε ταξινομητής (DS, VFI, NB) παράγει μια υπόθεση h_1 , h_2 , h_3 αντίστοιχα. Αν οι πιθανότητες πρόβλεψης που παράγονται από

¹⁵ Από τις περιπτώσεις κατάρτισης, ο αλγόριθμος VFI κατασκευάζει διαστήματα για κάθε χαρακτηριστικό (διακριτοποίηση). Για την ταξινόμηση μιας νέας περίπτωσης χρησιμοποιείται η ψηφοφορία κάθε χαρακτηριστικού χωριστά για τις τιμές των μεταβλητών της νέας περίπτωσης.

τους μεμονωμένους ταξινομητές είναι αντίστοιχα $p_1(i)$, $p_2(i)$, $p_3(i)$ για κάθε κλάση i , τότε η κατηγορία που συγκεντρώνει την μεγαλύτερη τιμή πιθανότητας λαμβάνονται ως απάντηση της ψηφοφορίας (h^*). Αναλυτικότερα, η κατηγορία (predictive class) υπολογίζεται από τον

$$\text{Predictive Class} = \arg \max_i \sum_{j=1, j=1}^{\text{number of classes, } j=3} p_j(i)$$

Εικόνα 66. Η γραφική αναπαράσταση της μεθόδου VoteDVN

Με άλλα λόγια, η διαδικασία ταξινόμησης περιλαμβάνει δύο φάσεις: (1) φάση εκπαίδευσης και (2) φάση ταξινόμησης. Κατά τη διάρκεια της φάσης εκπαίδευσης, τρεις ταξινομητές βάσης εκπαιδεύονται. Για την ταξινόμηση νέων παραδειγμάτων οι αντίστοιχες ταξινομήσεις των ταξινομητών βάσης συνδυάζονται με απλή αθροιστική ψηφοφορία για να παράγουν την τελική ταξινόμηση του παραδείγματος.

Για τα πειράματά μας εκτελέσαμε μια μεγάλης κλίμακας σύγκριση με άλλους γνωστούς αλγορίθμους μάθησης και η προτεινόμενη μέθοδος έδωσε καλύτερη ακρίβεια χρησιμοποιώντας πολύ λιγότερο χρόνο για εκπαίδευση. Αναλυτικότερα, συγκρίναμε την προτεινόμενη μέθοδο με τους ακόλουθους ισχυρούς αλγόριθμους μάθησης:

- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των οκνηρών αλγορίθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMO ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης
- Τον RIPPER ως εκπρόσωπο των κανόνων ταξινόμησης

Για τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Συγκενρωτικά, στην προτελευταία γραμμή του Πίνακα 55 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του Πίνακα 55 παρουσιάζεται η μέση ακρίβεια κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους RIPPER και C4.5 σε 8 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 3 σύνολα δεδομένων (βλέπε Πίνακα 55). Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους BP και 3NN σε 6 και 7 σύνολα δεδομένων, αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους και από τους δυο αλγόριθμους σε μόλις 1 σύνολο δεδομένων. Επιπλέον, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον SMO σε 2 σύνολα δεδομένων αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε ένα σύνολο δεδομένων.

Δεδομένα	VoteDVN	C4.5	3NN	RIPPER	BP	SMO
breast-cancer	71.77	74.28	73.13	71.65	71.18	69.52
breast-w	96.45	95.01	96.61	95.72	95.97	96.75
colic	81.42	85.16	80.95	84.97	82.58	82.66
credit-g	75.28	71.25*	72.21*	71.86*	72.75	75.15
diabetes	76.04	74.49	73.86	75.22	74.64	76.80
grub-damage	49.62	35.96*	44.05	34.73*	42.00	47.52
haberman	74.57	71.05	69.77*	72.43	76.56	73.37
heart-c	83.04	76.94*	81.82	79.05	81.39	83.89
heart-h	85.01	80.22*	82.33	79.26*	81.37	82.81
heart-statlog	83.81	78.15*	79.11*	78.70*	82.11	83.89
hepatitis	85.37	79.22	80.85	77.43*	81.30	85.70
hypothyroid	95.77	99.54v	93.21*	99.39v	94.33*	93.58*
ionosphere	85.04	89.74	86.02	89.30	85.84	88.07
iris	95.13	94.73	95.20	93.93	96.27	96.20
labor	95.53	78.60*	92.83	83.37*	88.57	92.97
lymphotherapy	81.52	75.84	81.74	77.36	80.24	86.41
mushroom	99.00	100.0v	100.0v	100.0v	99.45v	100.0v
primary-tumor	50.56	41.39*	44.98*	38.59*	28.11*	46.97
sick	97.75	98.72v	96.21*	98.30v	96.78*	93.88*
soybean	92.84	91.78	91.20	91.93	33.43*	93.10
waveform	79.28	75.25*	77.67 *	79.14	59.62 *	86.48v
zoo	97.03	92.61	92.61	87.03*	61.21*	96.05
		3/11/8	1/14/7	3/11/8	1/15/6	1/19/2
Μέση ακρίβεια	83.27	80.00	81.20	79.97	75.71	83.26

Πίνακας 55. Η σύγκριση του προτεινόμενου αλγορίθμου VoteDVN με άλλους ισχυρούς αλγόριθμους μάθησης

Επίσης, συγκρίναμε την προτεινόμενη μέθοδο με άλλες γρήγορες ομάδες ταξινομητών (με 25 υπο-ταξινομητές), όπως:

- Adaboost DS
- MultiBoost Decision Stump
- Bagging NB
- Adaboost NB

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τους Adaboost DS και Bagging NB σε 7 και 4 σύνολα δεδομένων, αντίστοιχα, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε σχέση με τις δυο άλλες μεθόδους σε ένα σύνολο δεδομένων (βλέπε Πίνακα 56).

Δεδομένα	VoteDVN	Adaboost DS	MultiBoost DS	Bagging NB	Adaboost NB
breast-cancer	71.77	71.55	71.03	72.70	68.57
breast-w	96.45	95.28	95.25	96.11	95.55
colic	81.42	82.72	80.96	78.59	77.46
credit-g	75.28	72.60	72.24	75.08	75.09
diabetes	76.04	75.37	76.70	75.70	75.88
grub-damage	49.62	33.20*	36.23	49.71	49.90
haberman	74.57	74.06	72.02	74.86	73.94
heart-c	83.04	83.11	83.98	83.44	83.14
heart-h	85.01	82.42	82.20	83.99	84.67
heart-statlog	83.81	81.81	80.54	83.78	82.30
hepatitis	85.37	81.50	81.51	83.93	84.23
hypothyroid	95.77	92.97*	93.64*	95.47*	95.27*
ionosphere	85.04	92.34 _v	89.83	82.36*	91.12 _v
iris	95.13	95.07	95.69	95.73	95.07
labor	95.53	90.57	93.00	94.27	88.93*
lymphotherapy	81.52	75.44	73.14*	83.14	80.67
mushroom	99.00	98.73	97.77*	95.76*	100.0 _v
primary-tumor	50.56	28.91*	29.39*	49.62	49.71
sick	97.75	97.07*	97.19	92.77*	93.70*
soybean	92.84	27.96*	28.21*	92.83	92.02
waveform	79.28	67.68*	66.44*	80.00 _v	80.01 _v
zoo	97.03	60.43*	61.47*	95.07	97.23
		1/14/7	0/15/7	1/17/4	3/16/3
Μέση ακρίβεια	83.27	75.49	75.38	82.50	82.48

Πίνακας 56. Η σύγκριση της τεχνικής VoteDVN με άλλες γρήγορες ομάδες ταξινομητών

Επίσης, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον MultiBoost DS σε 7 σύνολα δεδομένων, αντίστοιχα, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Τέλος, αν και ο αλγόριθμος έχει παρόμοια επίδοση με τον Adaboost NB χρησιμοποιεί πολύ λιγότερο χρόνο για εκπαίδευση.

Συνεπώς, όλα τα πειράματα έδειξαν ότι η προτεινόμενη ομάδα ταξινομητών απέδωσε, κατά μέσον όρο, καλύτερα από όλους τους εξεταζόμενους αλγορίθμους μάθησης χρησιμοποιώντας επίσης και πολύ λιγότερο χρόνο για εκπαίδευση.

Τέλος, πρέπει να αναφερθεί ότι η προτεινόμενη μέθοδος μπορεί εύκολα να εκτελεστεί παράλληλα εκτός από σειριακά χρησιμοποιώντας έναν αλγόριθμο μάθησης ανά μηχανή. Εκμεταλλευόμενη την παράλληλη ή την κατανεμημένη εκτέλεση μπορεί να αυξήσει την ταχύτητά της ακόμα περισσότερο.

4.3 Αλλαγή της συνάρτησης-στόχου

Μια άλλη κατηγορία μεθόδων κατασκευής ομάδας ταξινομητών πραγματοποιεί αλλαγή στην συνάρτηση στόχο και εκπαιδεύει τους ταξινομητές σε αυτά τα νέα προβλήματα. Όπως είναι αναμενόμενο αυτή η κατηγορία μεθόδων εφαρμόζεται μόνο σε πολυ-κατηγορικά προβλήματα (multi-class).

Η πιο απλή μέθοδος αυτής της κατηγορίας είναι η «ένας ταξινομητής ανά κατηγορία (one per class)». Ο ταξινομητής εκπαιδεύεται (τόσες φορές όσες και οι κλάσεις) με όλα τα παραδείγματα της κατηγορίας i ως θετικά, και όλα τα άλλα παραδείγματα ως αρνητικά (βλέπε

Πίνακα 57). Η τελική απόφαση είναι η κατηγορία που έχει τις περισσότερες επιλογές ή την μεγαλύτερη βεβαιότητα (Allwein και άλλοι, 2000).

Κλάση	Διάνυσμα κλάσης
A	1000
B	0100
C	0010
D	0001

Πίνακας 57. Διάσπαση ενός προβλήματος με 4 κλάσεις χρησιμοποιώντας την τεχνική ενός ταξινομητή ανά κατηγορία

Μια άλλη παρόμοια μέθοδος με την προηγούμενη είναι αυτή της δημιουργίας ενός δυαδικού προβλήματος ανά δυο κλάσεις (pairwise classification) (Furnkranz, 2001). Τα παραδείγματα που δεν ανήκουν στις δυο εξεταζόμενες κλάσεις ανά περίπτωση απλά δεν λαμβάνονται υπόψιν.

Σε αυτή την κατηγορία μεθόδων ανήκει επίσης και μια πρωτότυπη τεχνική που ονομάζεται κωδικοποίηση διόρθωσης λαθών εξόδου (Error-Correcting Output Coding ή ECOC) (Dietterich & Bakiri, 1995). Και αυτή μετασχηματίζει το αρχικό πρόβλημα πολλών κλάσεων σε πολλά προβλήματα δύο κλάσεων, για καθένα από τα οποία εκπαιδεύεται ένας ταξινομητής. Καθεμιά από τις αρχικές κλάσεις αντιστοιχίζεται σε μία συμβολοσειρά δυαδικών ψηφίων (bit-string) και το ίδιο συμβαίνει για κάθε άγνωστο στιγμιότυπο που κατατάσσεται από τους εκπαιδευθέντες ταξινομητές. Ο αλγόριθμος ECOC παρουσιάζεται σε ψευδογλώσσα στην Εικόνα 67.

Φάση εκπαίδευσης

1. δημιουργείται δυαδικός πίνακας M (όπως στον Πίνακα 58 για 4 κλάσεις)
2. κάθε κατηγορία αντιστοιχίζεται σε μια σειρά του πίνακα M
3. εκπαίδευση τόσων ταξινομητών βάσης όσα τα δυαδικά προβλήματα (ένας ταξινομητής ανά στήλη του διανύσματος κλάσης του πίνακα M)

Φάση ταξινόμησης

1. εφαρμογή όλων των δυαδικών ταξινομητών στο παράδειγμα ταξινόμησης
2. συνδυασμός των προβλέψεων ώστε να δημιουργηθεί μια δυαδική σειρά μήκους αντίστοιχου των δυαδικών προβλημάτων
3. ταξινόμηση στην κατηγορία εκείνη της οποίας η κωδική λέξη είναι πιο κοντά (κατά απόσταση Hamming) στη δική του

Εικόνα 67. Ο αλγόριθμος ECOC σε ψευδογλώσσα, όπου μ είναι ο αριθμός των κατηγοριών

Κλάση	Διάνυσμα κλάσης
a	1111111
b	0000111
c	0011001
d	0101010

Πίνακας 58. Διάσπαση ενός προβλήματος με 4 κλάσεις χρησιμοποιώντας την τεχνική ECOC

Πρέπει να τονιστεί ότι οι δυαδικοί αλγόριθμοι όπως οι μηχανές διανυσμάτων υποστήριξης (SVM) για να εφαρμοστούν σε πολυ-κατηγορικά προβλήματα κάνουν χρήση των μεθόδων αυτής της κατηγορίας. Γι' αυτό το λόγο πλήθος πειραμάτων σύγκρισης των μεθόδων αυτών έχουν πραγματοποιηθεί στην βιβλιογραφία χωρίς ξεκάθαρα αποτελέσματα (Hsu & Lin, 2002), (Crammer & Singer, 2002). Διαφαίνεται ότι η τεχνική ένας ταξινομητής ανά κλάση έχει την καλύτερη απόδοση ανά χρόνο εκπαίδευσης (Rifkin & Klautau, 2004).

4.4 Εισαγωγή τυχαιότητας στον αλγόριθμο μάθησης

Τέλος, η εισαγωγή τυχαιότητας μπορεί να χρησιμοποιηθεί σε αλγορίθμους μάθησης που είτε χρησιμοποιούν κάποιες αρχικές παραμέτρους (π.χ. στα νευρωνικά δίκτυα, τα αρχικά βάρη των συνάψεων μπορούν να επιλεγούν τυχαία, δίνοντας αρκετά διαφορετικούς ταξινομητές) ή αποτελούν επεκτάσεις ντετερμινιστικών αλγορίθμων (π.χ. (Dietterich, 2000) για το γνωστό αλγόριθμο δέντρων απόφασης C4.5).

Έτσι για παράδειγμα μπορούμε να εκπαιδεύσουμε μια ομάδα νευρωνικών δικτύων χρησιμοποιώντας διαφορετικές αρχικές παραμέτρους ανά ταξινομητή και εν συνεχεία να συνδυάσουμε τις αποφάσεις τους (Maclin & Shavlik, 1995).

Μια άλλη προσέγγιση είναι η επιλογή διαφορετικών μεταβλητών για την εκπαίδευση πολλών ταξινομητών χρησιμοποιώντας τον ίδιο αλγόριθμο μάθησης (Kuncheva & Whitaker, 2001). Η τεχνική αυτή μοιάζει πολύ με την τοποθέτηση μέσα σε σάκο (bagging). Αντί των περιπτώσεων, χρησιμοποιεί τυχαία δειγματοληψία των μεταβλητών. Ο Ho (1998) έδειξε ότι η απλή τυχαία επιλογή υποσυνόλων των μεταβλητών (subspace) μπορεί να είναι μια αποτελεσματική τεχνική σε πολλές περιπτώσεις.

Επίσης, ο Ho (1998) και ο Tsymbal και άλλοι (2002) έδειξαν ότι η επιλογή διαφορετικών μεταβλητών για την εκπαίδευση πολλών ταξινομητών χρησιμοποιώντας τον ίδιο αλγόριθμο μάθησης ενδείκνυται όταν τα δεδομένα εκπαίδευσης έχουν πολλές μεταβλητές.

ΚΕΦΑΛΑΙΟ 5. Επιλογή μεταβλητών και συνεισφορά της διατριβής στη δημιουργία υβριδικής μεθόδου

Σε πολλές εφαρμογές δεν είναι διόλου απίθανο να πρέπει να αντιμετωπίσουμε προβλήματα με πενήντα ή και παραπάνω χαρακτηριστικά. Μπορεί βέβαια να πιστεύουμε ότι κάθε χαρακτηριστικό είναι χρήσιμο για τουλάχιστον κάποιες από τις διαφοροποιήσεις, αλλά μπορεί να αμφιβάλλουμε για το αν κάθε χαρακτηριστικό παρέχει ανεξάρτητη πληροφορία (είναι αυτονόητο ότι με δική μας πρόθεση τα περισσότερα χαρακτηριστικά δεν θα έχουν συμπεριληφθεί). Υπάρχουν δύο θέματα λοιπόν που πρέπει να αντιμετωπιστούν. Το πιο σημαντικό είναι το πόσο η απόδοση της ταξινόμησης εξαρτάται από τις πολλές μεταβλητές, και το δεύτερο είναι η υπολογιστική πολυπλοκότητα του σχεδιασμού του ταξινομητή (Liu & Motoda, 1998).

Η μεγάλη διαστασιμότητα (dimensionality) αποτελεί πρόβλημα, και μάλιστα τόσο σημαντικό που πολλές φορές αναφέρεται ως κατάρα της διαστασιμότητας (curse of dimensionality). Ένας λόγος είναι η αναπόφευκτη αύξηση της υπολογιστικής πολυπλοκότητας για την εκπαίδευση και λειτουργία του ταξινομητή. Ωστόσο, το επιπλέον υπολογιστικό κόστος ενδεχομένως να γινόταν αποδεκτό σε διάφορες εφαρμογές, αν μπορούσε να εγγυηθεί την παράλληλη αύξηση της ακρίβειας του προκύπτοντος ταξινομητή. Στην πράξη, η μεγάλη διαστασιμότητα συχνά όχι μόνο δεν αυξάνει την ακρίβεια, αλλά τη μειώνει (Pyle, 1999).

Γενικά, οι μεταβλητές χωρίζονται (Bell & Wang, 2000) σε:

- σχετικές: μεταβλητές που έχουν μεγάλη επιρροή στην ακρίβεια του ταξινομητή και ο ρόλος τους δεν μπορεί να υποκατασταθεί από καμία άλλη μεταβλητή του συνόλου εκπαίδευσης
- άσχετες: μεταβλητές που δεν έχουν οποιαδήποτε επιρροή στην ακρίβεια του ταξινομητή.
- περιττές: μεταβλητές που έχουν μεγάλη επιρροή στην ακρίβεια του ταξινομητή αλλά ο ρόλος τους μπορεί να υποκατασταθεί από άλλες μεταβλητές του συνόλου εκπαίδευσης.

Όπως έχει ξανααναφερθεί μια από τις μεθόδους που εφαρμόζονται για την επιλογή μεταβλητών είναι αυτή του περιτυλίγματος (wrapper). Η μέθοδος του περιτυλίγματος χρησιμοποιεί τον ίδιο τον αλγόριθμο μάθησης ως μέσο για την καθοδήγηση της αναζήτησης στο χώρο των μεταβλητών. Ξεκινώντας από ένα αρχικό σύνολο μεταβλητών, δημιουργούνται διαδοχικά νέα σύνολα μέσω κατάλληλα ορισμένων τελεστών που προσθέτουν ή διαγράφουν μεταβλητές. Καθένα απ' αυτά χρησιμοποιείται για την εκπαίδευση ενός ταξινομητή, ο οποίος στη συνέχεια ελέγχεται πάνω σε ένα σύνολο δεδομένων επικύρωσης (validation set). Όταν εκπληρωθεί η συνθήκη τερματισμού της αναζήτησης, το καλύτερο σύνολο ως προς την απόδοση στο σύνολο δεδομένων επικύρωσης επιστρέφεται. Το πλεονέκτημα της μεθόδου είναι πως επιλέγει το σύνολο των μεταβλητών με την καλύτερη απόδοση για τον συγκεκριμένο επαγωγικό αλγόριθμο μάθησης.

Στην επόμενη ενότητα συγκρίνουμε τις πιο γνωστές μεθόδους επιλογής μεταβλητών περιτυλίγματος.

5.1 Σύγκριση των πιο γνωστών μεθόδων επιλογής μεταβλητών περιτυλίγματος

Η διαδικασία της επιλογής χαρακτηριστικών αποτελεί κατ' ουσία ένα πρόβλημα αναζήτησης του καταλληλότερου υποσυνόλου στο χώρο όλων των δυνατών υποσυνόλων που ορίζονται από τα χαρακτηριστικά. Στην πράξη, η εξαντλητική αναζήτηση του χώρου αυτού κρίνεται από άποψη πολυπλοκότητας απαγορευτική, γεγονός που μας ωθεί στην υιοθέτηση ευριστικών μεθόδων αναζήτησης, προκειμένου να μειωθεί ο χρόνος εκτέλεσης.

Οι αλγόριθμοι επιλογής χαρακτηριστικών περιτυλίγματος γενικά αποτελούνται από δύο συστατικά: έναν αλγόριθμο επιλογής που παράγει τα προτεινόμενα υποσύνολα των χαρακτηριστικών προσπαθώντας να βρει το βέλτιστο υποσύνολο και έναν αλγόριθμο αξιολόγησης που αξιολογεί τα παραγόμενα υποσύνολα βάσει ενός επιλεγμένου κριτηρίου. Εντούτοις, χωρίς μια κατάλληλη συνθήκη τερματισμού η διαδικασία επιλογής χαρακτηριστικών μπορεί να τρέξει εξαντλητικά χωρίς εμφανές κέρδος στην απόδοση του συστήματος. Η αναζήτηση μπορεί να τερματίσει: (i) είτε όταν η προσθήκη (ή διαγραφή) οποιουδήποτε χαρακτηριστικού δεν παράγει ένα καλύτερο υποσύνολο είτε (ii) ένα βέλτιστο υποσύνολο σύμφωνα με κάποια συγκεκριμένη συνθήκη βρεθεί.

Υπάρχουν διάφοροι αλγόριθμοι επιλογής μεταβλητών περιτυλίγματος. Η απλούστερη μέθοδος είναι η προς τα εμπρός επιλογή (Forward Selection-FS) (Caruana & Freitag, 1994) που αρχίζει με το κενό σύνολο και προσθέτει μια μεταβλητή τη φορά. Σε κάθε βήμα επιλέγει την μεταβλητή που, όταν προστεθεί στο τρέχον σύνολο, παράγει την υπόθεση που γενικεύει καλύτερα. Σε αυτό το σημείο πρέπει να τονιστεί ότι αν προστεθεί μια μεταβλητή στο σύνολο σε ένα βήμα δεν μπορεί σε επόμενο βήμα να αφαιρεθεί. Η διαδικασία τερματίζει όταν η προσθήκη καμιάς μεταβλητής δεν αυξάνει επιπλέον την ακρίβεια.

Αντίθετα, η προς τα πίσω επιλογή (Backward Selection-BS) αρχίζει με όλες τις μεταβλητές και αφαιρεί μια μεταβλητή τη φορά (Dash & Liu, 1997). Αντίστοιχα με την προς τα εμπρός επιλογή, η προς τα πίσω επιλογή αφαιρεί σε κάθε βήμα τη μεταβλητή της οποίας η αφαίρεση παράγει το σύνολο των μεταβλητών που επιτυγχάνει την καλύτερη γενίκευση. Επίσης όπως στην προς τα μπρος επιλογή, μόλις αφαιρεθεί μια μεταβλητή, δεν μπορεί σε επόμενο βήμα να προστεθεί πίσω στο σύνολο.

Ένα πρόβλημα με την προς τα εμπρός επιλογή είναι ότι μπορεί να αποτύχει να περιλάβει τις μεταβλητές που είναι αλληλοεξαρτώμενες, δεδομένου ότι προσθέτει μια μεταβλητή τη φορά. Εντούτοις, μπορεί να εντοπίσει μικρά αποτελεσματικά υποσύνολα αρκετά γρήγορα, δεδομένου ότι οι πρόωρες αξιολογήσεις, που περιλαμβάνουν σχετικά λίγες μεταβλητές, είναι γρήγορες. Αντίθετα, στην προς τα πίσω επιλογή οι αλληλεξαρτήσεις αντιμετωπίζονται καλά, αλλά οι πρόωρες αξιολογήσεις είναι σχετικά ακριβές (Liu & Motoda, 1998).

Η διαδοχική προς τα εμπρός επιλογή μεταβλητών (SFFS) και η διαδοχική προς τα πίσω επιλογή μεταβλητών (SBFS) είναι μια παραλλαγή των προαναφερθέντων μεθόδων αφού χαρακτηρίζονται από μεταβαλλόμενο αριθμό χαρακτηριστικών που περιλαμβάνονται ή εξαλείφονται σε κάθε στάδιο της διαδικασίας (Pudil και άλλοι, 1994). Ένας παρόμοιος τρόπος ακολουθείται και από την βέλτιστη αναζήτηση (best first), σύμφωνα με την οποία, σε κάθε βήμα της αναζήτησης παράγονται διαδοχικά όλα τα υποσύνολα του τρέχοντος βέλτιστου υποσυνόλου που προκύπτουν με την προσθήκη ενός ακόμα χαρακτηριστικού σε αυτά, και στη συνέχεια επιλέγεται εκείνο που επιτυγχάνει την καλύτερη απόδοση. Αν σε κάποιο βήμα ο αλγόριθμος οδηγηθεί σε αδιέξοδο, οπισθοδρομεί και επιλέγει το αμέσως επόμενο κατά σειρά καταλληλότητας υποσύνολο. Ένα κριτήριο τερματισμού της αναζήτησης που συναντάται στις διάφορες υλοποιήσεις του αλγορίθμου και που έχει ως σκοπό να αποτρέψει την εξαντλητική αναζήτηση του χώρου, αποτελεί ο περιορισμός των διαδοχικών οπισθοδρομήσεων σε ένα αριθμό που καθορίζεται από το χρήστη. Ο αναγνώστης μπορεί για επιπλέον πληροφορίες να ανατρέξει στο άρθρο (Kohavi & John, 1997). Πρέπει επίσης να τονιστεί ότι η βέλτιστη αναζήτηση (best first) μπορεί να συνδυαστεί και με την προς τα εμπρός και με την προς τα πίσω επιλογή μεταβλητών.

Ένας άλλος τρόπος επιλογής του καλύτερου υποσυνόλου μεταβλητών είναι να αρχίσει η αναζήτηση από ένα τυχαία επιλεγμένο υποσύνολο και να προστίθεται ή να διαγράφεται ένα χαρακτηριστικό τυχαία ανά επανάληψη (Liu & Setiono, 1996). Αναλυτικότερα, σε κάθε ένα από τα βήματα του, ο συγκεκριμένος αλγόριθμος συστήνει τυχαία ένα υποσύνολο, έστω S , και εφόσον τα στοιχεία του είναι λιγότερα από εκείνα του τρέχοντος καλύτερου S' , τότε

ελέγχεται αν ο ρυθμός ασυνέπειας¹⁶ που εμφανίζει το σώμα εκπαίδευσης (inconsistency rate). Αν ο ρυθμός ασυνέπειας είναι μικρότερος ενός προκαθορισμένου, το τρέχον υποσύνολο αντικαθίσταται από το S. Αν όμως η πληθικότητα των S και S' ταυτίζεται και το κριτήριο ασυνέπειας ικανοποιείται για το S, τότε και αυτό γίνεται ένα από τα βέλτιστα υποψήφια σύνολα χαρακτηριστικών που θα προτείνει ο αλγόριθμος για χρήση.

Μια πιο ενημερωμένη τυχαία επιλογή χαρακτηριστικών γνωρισμάτων πραγματοποιείται με τη χρήση των γενετικών αλγορίθμων (Yang & Honavar, 1998). Το πλήθος των μεταβλητών περιγράφεται από μια σταθερή δυαδική σειρά τόσο μήκους όσο το πλήθος των μεταβλητών - η τιμή κάθε θέσης στη σειρά αντιπροσωπεύει την παρουσία (1) ή την απουσία (0) ενός χαρακτηριστικού. Ο αλγόριθμος στη συνέχεια είναι μια επαναληπτική διαδικασία όπου κάθε διαδοχική γενιά παράγεται με την εφαρμογή των γενετικών τελεστών της διασταύρωσης και της μετάλλαξης στα μέλη της τρέχουσας γενεάς (στην ενότητα 3.2.2 περιγράφηκε αναλυτικότερα ο τρόπος λειτουργίας των γενετικών αλγορίθμων).

Όπως προείπαμε, οι μέθοδοι περιτυλίγματος επιλέγουν το σύνολο των μεταβλητών με την καλύτερη απόδοση για τον συγκεκριμένο αλγόριθμο μάθησης που χρησιμοποιούμε. Εν συνεχεία, παρουσιάζεται η σύγκριση των πέντε πιο γνωστών μεθόδων επιλογής μεταβλητών περιτυλίγματος: η προς τα εμπρός επιλογή, η προς τα πίσω επιλογή, η βέλτιστη προς τα εμπρός, η βέλτιστη προς τα πίσω και η επιλογή μεταβλητών με χρήση γενετικών αλγορίθμων. Για τα πειράματά μας χρησιμοποιήσαμε τέσσερις γνωστούς αντιπροσωπευτικούς αλγορίθμους μάθησης:

- Τον NB ως εκπρόσωπο των Μπεϋζιανών αλγορίθμων
- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των οκνηρών αλγορίθμων
- Τον RIPPER ως εκπρόσωπο των κανόνων ταξινόμησης

Ο Πίνακας 59 δείχνει την ακρίβεια του αλγορίθμου NB καθώς επίσης και τον αριθμό των χρησιμοποιούμενων μεταβλητών σε κάθε σύνολο δεδομένων πριν (Without Selection-WS) και μετά από την εφαρμογή των αλγορίθμων επιλογής μεταβλητών. Για τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Επιπλέον, στον πίνακα, συμβολίζεται με "v" ότι ο ταξινομητής με την συγκεκριμένη μέθοδο περιτυλίγματος έδωσε στατιστικά καλύτερα αποτελέσματα από τον ίδιο τον ταξινομητή χωρίς την εφαρμογή αλγορίθμου επιλογής μεταβλητών. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Σύμφωνα με τα πειράματά μας, η βέλτιστη προς τα πίσω επιλογή (BFBS) είναι ελαφρώς η καλύτερη μέθοδος επιλογής χαρακτηριστικών για τον ταξινομητή NB αν και η προς τα εμπρός αναζήτηση (FS) χρησιμοποιεί πολύ λιγότερα χαρακτηριστικά για την επαγωγή και είναι λιγότερο δαπανηρή μέθοδος όσον αφορά τον χρόνο εκπαίδευσης. Πρέπει επίσης να αναφερθεί ότι όλοι οι αλγόριθμοι επιλογής βελτιώνουν την ακρίβεια του ταξινομητή NB κατά τουλάχιστον 3%.

¹⁶ Δύο στιγμιότυπα θεωρούνται ασυνεπή εφόσον ταυτίζονται σε όλα τα χαρακτηριστικά εκτός από την κλάση τους. Ως βαθμός ασυνέπειας (inconsistency count) ενός συνόλου n ασυνεπών στιγμιοτύπων ορίζεται η διαφορά $n - l$, όπου l ο αριθμός εκείνων που ανήκουν στην πλέον πολυπληθή κλάση. Ο ρυθμός ασυνέπειας ορίζεται ως το πηλίκο όλων των βαθμών ασυνέπειας προς τον ολικό αριθμό των στιγμιοτύπων.

Δεδομένα	WS	FS	BS	BFFS	BFBS	GS
anneal	86.59	93.21 v (7/38) ¹⁷	86.59 (38/38)	93.66 v (8/38)	93.65 v (28/38)	92.5 v (15/38)
autos	57.41	64.82 v (5/25)	65.62 v (17/25)	67.9 v (9/25)	68.13 v (10/25)	66.62 v (9/25)
balance	90.53	90.53 (4/4)	90.53 (4/4)	90.53 (4/4)	90.53 (4/4)	90.53 (4/4)
breast-cancer	72.7	74.58 (2/9)	72.99 (8/9)	74.58 (2/9)	75.41 v (3/9)	75.41 v (3/9)
breast-w	96.07	96.37 (3/9)	96.64 v (6/9)	96.58 v (7/9)	96.64 v (6/9)	96.64 v (6/9)
colic	78.7	85.35 v (4/22)	80.95 v (16/22)	85.11 v (6/22)	86.29 v (5/22)	85.77 v (8/22)
credit-a	77.86	87.38 v (8/15)	87.38 v (8/15)	87.38 v (8/15)	87.38 v (8/15)	87.35 v (9/15)
credit-g	75.16	75.67 (5/20)	75.14 (17/20)	76.18 v (10/20)	76 v (11/20)	76.08 v (12/20)
diabetes	75.75	76.4 (2/8)	77.58 v (5/8)	76.89 v (5/8)	77.58 v (5/8)	77.58 v (5/8)
heart-c	83.34	85.05 (6/13)	84.66 (10/13)	85.05 (6/13)	84.72 (10/13)	85.19 (7/13)
heart-h	83.95	85.14 (6/13)	84.77 (8/13)	85.14 (6/13)	84.77 (8/13)	85.14 (6/13)
heart-statlog	83.59	86.07 v (11/13)	86.07 v (11/13)	86.07 v (11/13)	86.07 v (11/13)	86.07 v (11/13)
hepatitis	83.81	85.77 (3/19)	87.75 v (12/19)	84.63 (4/19)	87.75 v (12/19)	89.08 v (8/19)
ionosphere	82.17	90.92 v (5/34)	93.2 v (12/34)	93 v (10/34)	93.2 v (12/34)	91.94 v (13/34)
iris	95.53	95.87 (3/4)	95.87 (3/4)	95.87 (3/4)	95.87 (3/4)	95.67 (3/4)
labor	93.57	91.23 (2/16)	94.77 (9/16)	92.8 (5/16)	93.3 (9/16)	91.27 (7/16)
Lymphotherapy	83.13	82.98 (3/18)	85.66 v (16/18)	84.07 (8/18)	87.51 v (11/18)	86.77 v (12/18)
sonar	67.71	77.16 v (5/60)	75.65 v (36/60)	76.86 v (10/60)	75.35 v (40/60)	74.25 v (34/60)
vote	90.02	94.37 v (5/16)	95.86 v (2/16)	96.32 v (3/16)	95.86 v (2/16)	96.09 v (3/16)
zoo	94.97	95.15 (8/17)	96.83 v (15/17)	95.15 (8/17)	97.12 v (8/17)	96.04 (7/17)
Μέση ακρίβεια	82.63	85.7	85.73	86.19	86.66	86.3

Πίνακας 59. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον NB

Ο Πίνακας 60 δείχνει την ακρίβεια του αλγορίθμου C4.5 καθώς επίσης και τον αριθμό των χρησιμοποιούμενων μεταβλητών σε κάθε σύνολο δεδομένων πριν και μετά από την εφαρμογή των αλγορίθμων επιλογής μεταβλητών.

¹⁷ (7/38) σημαίνει ότι 7 από τις 38 μεταβλητές εν τέλει χρησιμοποιήθηκαν

Δεδομένα	WS	FS	BS	BFFS	BFBS	GS
anneal	98.51	98.75 (9/38)	98.75 (9/38)	98.75 (9/38)	98.93 v (19/38)	98.82 v (20/38)
autos	81.82	79.21 (9/25)	81.77 (18/25)	81.55 (13/25)	81.77 (18/25)	83.32 (14/25)
balance	77.82	77.82 (4/4)	77.82 (4/4)	77.82 (4/4)	77.82 (4/4)	77.82 (4/4)
breast-cancer	74.28	70.3 * (0/9)	75.72 (3/9)	70.3 * (0/9)	75.72 (3/9)	75.72 (3/9)
breast-w	95.01	96.04 v (3/9)	94.82 (6/9)	96.04 v (3/9)	95.92 v (5/9)	95.44 (4/9)
colic	85.16	85.86 v (2/22)	85.86 v (2/22)	85.86 v (2/22)	86.02 v (5/22)	85.78 (7/22)
credit-a	85.57	85.51 (1/15)	87 v (9/15)	85.51 (1/15)	87 v (9/15)	86.57 (8/15)
credit-g	71.25	73.88 v (5/20)	73.27 v (12/20)	73.88 v (5/20)	73.27 v (12/20)	73.2 v (10/20)
diabetes	74.49	74.66 (3/8)	74.66 (3/8)	74.66 (3/8)	74.66 (3/8)	74.66 (3/8)
heart-c	76.87	77.4 (3/13)	82.1 v (6/13)	82.74 v (6/13)	82.1 v (6/13)	82.89 v (5/13)
heart-h	80.26	81.32 (1/13)	83 v (2/13)	81.32 (1/13)	83 v (2/13)	83 v (2/13)
heart-statlog	77.96	85.7 v (3/13)	85.7 v (3/13)	85.7 v (3/13)	85.7 v (3/13)	85.56 v (4/13)
hepatitis	79.22	84.47 v (2/19)	85.71 v (7/19)	84.47 v (2/19)	86.15 v (7/19)	88.3 v (6/19)
ionosphere	89.46	94.27 v (4/34)	90.6 (25/34)	94.27 v (4/34)	90.63 (25/34)	91.88 v (18/34)
iris	94.73	94.8 (2/4)	94.8 (2/4)	94.8 (2/4)	94.8 (2/4)	94.8 (2/4)
labor	78.27	83.1 v (3/16)	81.5 (6/16)	83.1 v (3/16)	82.03 v (6/16)	83.1 v (4/16)
Lymphotherapy	76.04	82.66 v (4/18)	78.46 (15/18)	82.66 v (4/18)	79.6 v (12/18)	83.54 v (8/18)
sonar	73.61	75.93 (5/60)	76.89 (29/60)	81.01 v (10/60)	77.69 (28/60)	78.32 v (22/60)
vote	96.57	95.63 * (1/16)	96.43 (5/16)	95.63 * (1/16)	96.43 (5/16)	96.73 (5/16)
zoo	91.31	94.45 v (6/17)	97.05 v (7/17)	94.45 v (6/17)	97.05 v (7/17)	96.04 v (7/17)
Μέση ακρίβεια	82,91	84,59	85,1	85,23	85,31	85,77

Πίνακας 60. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον C4.5

Η χρήση γενετικών αλγορίθμων (GS) είναι η καλύτερη μέθοδος επιλογής χαρακτηριστικών σύμφωνα με τα πειράματά μας για το C4.5 ταξινομητή εντούτοις, η προς τα εμπρός αναζήτηση επιλογής (FS) χρησιμοποιεί πολύ λιγότερα χαρακτηριστικά για την επαγωγή. Πρέπει επίσης να αναφερθεί ότι ο C4.5 βελτιώνεται λιγότερο από την επιλογή

χαρακτηριστικών σε σχέση με τον NB, επειδή ο C4.5 ενσωματώνει ένα είδος επιλογής χαρακτηριστικών στον κώδικα του.

Ο Πίνακας 61 δείχνει την ακρίβεια του αλγορίθμου RIPPER καθώς επίσης και τον αριθμό των χρησιμοποιούμενων μεταβλητών σε κάθε σύνολο δεδομένων πριν και μετά από την εφαρμογή των αλγορίθμων επιλογής μεταβλητών.

Δεδομένα	WS	FS	BS	BFFS	BFBS	GS
anneal	98.09	98.33 (7/38)	98.09 (38/38)	98.33 (7/38)	98.81 v (7/38)	98.61 v (22/38)
autos	73.37	77.36 v (6/25)	73.57 (20/25)	77.65 v (6/25)	74.67 (20/25)	73.93 (15/25)
balance	80.96	80.96 (4/4)	80.96 (4/4)	80.96 (4/4)	80.96 (4/4)	80.96 (4/4)
breast-cancer	72.53	75.12 v (3/9)	72.63 (6/9)	74.97 v (3/9)	75.12 v (3/9)	75.12 v (3/9)
breast-w	95.65	95.87 (7/9)	95.58 (7/9)	95.81 (7/9)	95.78 (5/9)	95.74 (4/9)
colic	84.91	86.27 v (4/22)	84.85 (12/22)	85.91 v (6/22)	84.85 (12/22)	85.78 (8/22)
credit-a	85.38	85.51 (1/15)	85.77 (12/15)	85.51 (1/15)	85.7 (12/15)	85.86 (9/15)
credit-g	72.25	71.01 (2/20)	72.02 (18/20)	71.43 (5/20)	72.29 (11/20)	72.15 (16/20)
diabetes	74.99	75.66 (3/8)	74.99 (8/8)	75.79 (3/8)	75.66 (3/8)	74.99 (8/8)
heart-c	79.51	76.57 * (1/13)	82.45 v (7/13)	83.96 v (5/13)	84.26 v (5/13)	83.93 v (5/13)
heart-h	78.92	81.98 v (1/13)	81.98 v (1/13)	81.98 v (1/13)	81.98 v (1/13)	81.04 v (6/13)
heart-statlog	78.7	84.7 v (4/13)	81.41 v (7/13)	84.7 v (4/13)	84.85 v (3/13)	84.7 v (4/13)
hepatitis	77.1	84.22 v (2/19)	82.2 v (15/19)	84.22 v (2/19)	84.4 v (5/19)	84.93 v (4/19)
ionosphere	89.89	91.77 v (3/34)	89.84 (26/34)	91.82 v (7/34)	88.95 (23/34)	91.34 (13/34)
iris	94.47	94.4 (3/4)	94.53 (2/4)	94.53 (2/4)	94.4 (3/4)	94.4 (3/4)
labor	81.53	90.07 v (3/16)	87.87 v (5/16)	90.07 v (3/16)	87.7 v (5/16)	90.5 v (2/16)
Lymphotherapy	76.9	80.17 (2/18)	77.33 (15/18)	79.12 (6/18)	79.89 (8/18)	78.69 (10/18)
sonar	75.1	77.68 (5/60)	75.43 (41/60)	78.21 (9/60)	75.15 (39/60)	77.06 (27/60)
vote	95.63	95.63 (1/16)	95.95 (8/16)	95.63 (1/16)	96.2 v (6/16)	95.84 (10/16)
zoo	86.85	93.26 v (5/17)	94.16 v (10/17)	93.26 v (5/17)	93.16 (7/17) v	92.78 v (5/17)
Μέση ακρίβεια	82,63	84,83	84,08	85,19	84,74	84,92

Πίνακας 61. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον RIPPER

Η βέλτιστη προς τα εμπρός επιλογή (BFFS) είναι η καλύτερη μέθοδος επιλογής χαρακτηριστικών σύμφωνα με τα πειράματά μας για τον ταξινομητή RIPPER χρησιμοποιώντας μόνο λίγα περισσότερα χαρακτηριστικά από την προς τα εμπρός αναζήτηση (FS). Μια άλλη χρήσιμη επίδραση όλων των τεχνικών επιλογής χαρακτηριστικών για τους αλγόριθμους παραγωγής κανόνων ταξινόμησης είναι η μείωση των παραχθέντων κανόνων.

Ο Πίνακας 62 δείχνει την ακρίβεια του αλγορίθμου 3NN καθώς επίσης και τον αριθμός των χρησιμοποιούμενων μεταβλητών σε κάθε σύνολο δεδομένων πριν και μετά από την εφαρμογή των αλγορίθμων επιλογής μεταβλητών. Σύμφωνα με τα πειράματά μας, η βέλτιστη προς τα εμπρός επιλογή (BFFS) είναι ελαφρώς η καλύτερη μέθοδος επιλογής χαρακτηριστικών για τον ταξινομητή 3NN αν και η προς τα εμπρός αναζήτηση χρησιμοποιεί πολύ λιγότερα χαρακτηριστικά για την επαγωγή.

Όπως ήταν αναμενόμενο, κανένας από τους αλγορίθμους επιλογής χαρακτηριστικών δεν είναι ανώτερος από όλους τους άλλους σε όλα τα σύνολα δεδομένων και για όλους τους αλγόριθμους μηχανικής μάθησης. Κάθε αλγόριθμος έχει μερικά πλεονεκτήματα και μειονεκτήματα (Yu & Liu, 2004).

Γενικά, οι προς τα πίσω στρατηγικές επιλογής μεταβλητών είναι πολύ χρονοβόρες για τα μεγάλα σύνολα δεδομένων, τα οποία έχουν πολλές μεταβλητές αρχικά. Οι προς τα εμπρός μέθοδοι επιλογής είναι λιγότερο ικανές να βελτιώσουν την απόδοση ενός δεδομένου ταξινομητή, αλλά είναι λιγότερο χρονοβόρες από την άποψη της υπολογιστικής πολυπλοκότητας και επίσης χρησιμοποιούν λιγότερα χαρακτηριστικά για την επαγωγή, κάνοντας τον ταξινομητή πιο κατανοητό. Τέλος, η επιλογή μεταβλητών με τη χρήση γενετικών αλγορίθμων απαιτεί έναν μεγάλο αριθμό αξιολογήσεων για να φθάσει σε τοπικό ελάχιστο.

Σαν γενικό συμπέρασμα, αν και οι μέθοδοι περιτυλίγματος επιλέγουν το σύνολο των μεταβλητών της αναπαράστασης με την καλύτερη απόδοση για τον συγκεκριμένο επαγωγικό αλγόριθμο μάθησης έχουν το βασικό μειονέκτημα της απαγορευτικής πολυπλοκότητας σε προβλήματα με εκατοντάδες μεταβλητές. Για τον λόγο αυτό στη συνέχεια, εξετάζουμε τη χρήση μιας υβριδικής τεχνικής επιλογής μεταβλητών (συνδυασμός μεθόδων περιτυλίγματος και φίλτρου) στην προσπάθεια να βελτιωθεί η ακρίβεια ταξινόμησης χωρίς να αυξηθεί δραματικά το υπολογιστικό κόστος.

Δεδομένα	WS	FS	BS	BFFS	BFBS	GS
anneal	97.29	98.76 v (11/38)	97.29 (38/38)	98.76 v (11/38)	98.62 v (14/38)	98.86 v (18/38)
autos	67.23	71.84 v (2/25)	67.23 (25/25)	71.84 v (2/25)	74.35 v (17/25)	73.06 v (12/25)
balance	86.74	86.74 (4/4)	86.74 (4/4)	86.74 (4/4)	86.74 (4/4)	86.74 (4/4)
breast-cancer	73.13	73.78 (2/9)	76.77 v (7/9)	73.78 (2/9)	76.77 v (7/9)	75.51 v (5/9)
breast-w	96.61	96.97 (6/9)	97.08 v (8/9)	97.08 v (8/9)	97.08 v (8/9)	97.08 v (8/9)
colic	80.95	86.57 v (4/22)	84.7 v (17/22)	86.73 v (4/22)	84.7 v (17/22)	87.25 v (8/22)
credit-a	84.96	85.51 (1/15)	87.33 v (11/15)	87.09 v (4/15)	86.55 v (9/15)	87.41 v (6/15)
credit-g	72.21	74.36 v (3/20)	72.3 (16/20)	74.36 v (3/20)	74.32 v (16/20)	73.07 (11/20)
diabetes	73.86	72.41 (2/8)	73.55 (6/8)	73.71 (5/8)	73.55 (6/8)	74.24 (6/8)
heart-c	81.82	82.31 (3/13)	83.57 (9/13)	84.17 v (7/13)	83.57 (9/13)	83.87 v (9/13)
heart-h	82.33	83.59 (4/13)	83.86 v (9/13)	83.59 (4/13)	83.86 v (9/13)	85.21 v (7/13)
heart-statlog	79.11	83.96 v (3/13)	83.22 v (8/13)	83.96 v (3/13)	83.22 v (8/13)	83.33 v (6/13)
hepatitis	80.85	88.29 v (6/19)	86.79 v (14/19)	88.29 v (6/19)	86.79 v (14/19)	87.56 v (11/19)
ionosphere	86.02	93.45 v (3/34)	88.66 v (26/34)	93.45 v (3/34)	93.16 v (10/34)	92.76 v (12/34)
iris	95.2	96 (1/4)	96.27 v (3/4)	96.27 v (3/4)	96.27 v (3/4)	96 (1/4)
labor	92.83	96.07 (4/16)	94.37 (11/16)	96.07 (4/16)	94.37 (11/16)	96.77 v (9/16)
Lymphotherapy	81.74	82.24 (3/18)	85.49 v (12/18)	85.93 v (11/18)	85.55 v (9/18)	86.12 v (13/18)
sonar	83.76	86.01 (6/60)	87.52 v (46/60)	90.7 v (15/60)	88.13 v (41/60)	88.33 v (25/60)
Vote	93.08	96.48 v (4/16)	96.66 v (10/16)	96.48 v (4/16)	96.66 v (10/16)	96.09 v (3/16)
Zoo	92.61	96.04 v (6/17)	94.27 v (14/17)	96.04 v (6/17)	96.04 v (6/17)	96.04 v (8/17)
Μέση ακρίβεια	84.12	86,57	86,18	87,25	87,01	86,75

Πίνακας 62. Σύγκριση των αλγορίθμων επιλογής περιτυλίγματος για τον 3NN

5.2 Προτεινόμενη υβριδική τεχνική επιλογής μεταβλητών

Έχουμε ενσωματώσει ένα φίλτρο επιλογής μεταβλητών στον αλγόριθμο που παρουσιάζεται σε αυτή την ενότητα. Το γνωστό φίλτρο που υπολογίζει τις εμπειρικές αμοιβαίες πληροφορίες μεταξύ των χαρακτηριστικών και της κλάσης (information gain), και απορρίπτει τα μικρής αξίας χαρακτηριστικά τροποποιήθηκε για τον αλγόριθμό μας. Αυτό συνέβηκε γιατί το αμοιβαίο κριτήριο κέρδους πληροφοριών (gain) έχει μια ισχυρή προκατάληψη υπέρ των χαρακτηριστικών με πολλές διαφορετικές τιμές. Αποκαταστάθηκε λοιπόν αυτή η προκατάληψη με ένα είδος κανονικοποίησης (gain ratio). Τα χαρακτηριστικά επιλέγονται με την κράτηση εκείνων για τα οποία η αναλογία κέρδους (gain ratio) υπερβαίνει ένα σταθερό κατώτατο όριο ϵ . Θέσαμε το ϵ ίσον με 0.02, σε μία προσπάθεια να απορριφθούν μόνο τα χαρακτηριστικά με αμελητέο αντίκτυπο στις προβλέψεις. Εντούτοις, ακόμα και ένα τόσο χαμηλό κατώτατο όριο μπορεί να απορρίψει πολλά χαρακτηριστικά σε μεγάλα προβλήματα. Εν συνεχεία, ο αλγόριθμος μας χρησιμοποιεί την βέλτιστη προς τα πίσω επιλογή μεταβλητών. Η βέλτιστη αναζήτηση αρχίζει με το κενό σύνολο χαρακτηριστικών και εν συνεχεία παράγει όλα τις πιθανά υποσύνολα του συνόλου των χαρακτηριστικών. Το υποσύνολο με την υψηλότερη αξιολόγηση επιλέγεται και η διαδικασία συνεχίζεται ακολουθώντας την ίδια λογική. Εάν η επιλογή ενός υποσυνόλου δεν οδηγήσει σε καμία βελτίωση, η αναζήτηση γυρνάει πίσω μέχρι τρία βήματα (όσα τις έχουμε ορίσει εμείς) και συνεχίζει από εκεί. Το καλύτερο υποσύνολο επιστρέφεται όταν ολοκληρωθεί η αναζήτηση. Ακολουθεί η σχηματική αναπαράσταση της μεθόδου.

Εικόνα 68. Σχηματική αναπαράσταση της υβριδικής τεχνικής επιλογής μεταβλητών

Τα πειράματα της εργασίας [7] με ιατρικά δεδομένα έδειξαν ότι τα αποτελέσματα της υβριδικής διαδικασίας επιλογής μεταβλητών είναι ελπιδοφόρα δεδομένου ότι αυτή η μεθοδολογία δίνει τόσο καλή ακρίβεια όσο μια από τις καλύτερες μεθοδολογίες συνδυασμού ταξινομητών (stacking). Ένα ισχυρό κίνητρο επιλογής μεταβλητών στα ιατρικά προβλήματα είναι ότι (Kononenko, 2001):

1. μερικές μεταβλητές έχουν συχνά υψηλό κόστος
2. ο χρόνος που καταναλώνεται για την εύρεση των τιμών τους είναι συνήθως μεγάλος
3. η διαδικασία εύρεση των τιμών τους μπορεί να είναι επίπονος (ακόμα και επιβλαβής) για τους ασθενείς.

Τα πειραματικά αποτελέσματα παράχθηκαν χρησιμοποιώντας 6 γνωστούς αντιπροσωπευτικούς εποπτευμένους αλγορίθμους μάθησης:

- Τον NB ως εκπρόσωπο των Μπεϋζιανών αλγορίθμων
- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των σκνηρών αλγορίθμων
- Τον RIPPER ως εκπρόσωπο των κανόνων ταξινόμησης
- Τον SMO ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης
- Τον BP ως εκπρόσωπο των νευρωνικών δικτύων

Για την συσσωρευμένη γενίκευση χρησιμοποιήσαμε όλους τους αλγόριθμους ως ταξινομητές βάσης και τον αλγόριθμο MLR ως πρόεδρο.

Ο Πίνακας 63 δείχνει την ακρίβειας των μεθόδων πριν και μετά την εφαρμογή της προτεινόμενης υβριδικής επιλογής των μεταβλητών.

<i>Δεδομένα</i>	<i>Stacking</i>	<i>NB</i>	<i>C4.5</i>	<i>3NN</i>	<i>RIPPER</i>	<i>BP</i>	<i>SMO</i>
breast-cancer		74.58 ¹⁸ (2/9) ¹⁹	75.19 (3/9)	73.78 (2/9)	74.91 (4/9)	76.13 (2/9)	72.35 (2/9)
	72.38	72.70 ²⁰	74.28	73.13	71.65	72.95	69.92
breast-w		96.64 (6/9)	95.92 (5/9)	97.08 (8/9)	95.78 (5/9)	96.85 (5/9)	97.05 (5/9)
	96.70	96.07	95.01	96.61	95.72	96.35	96.81
colic		86.10 (6/22)	85.86 (5/22)	81.11 (18/22)	85.18 (13/22)	86.87 (5/22)	83.53 (14/22)
	84.59	78.70	85.16	80.95	84.97	83.07	82.69
diabetes		76.52 (5/8)	74.66 (3/8)	72.37 (6/8)	75.66 (3/8)	76.77 (5/8)	77.03 (6/8)
	76.12	75.75	74.49	73.86	75.22	77.04	77.07
heart-h		83.79 (8/13)	83.00 (2/13)	81.12 (8/13)	81.98 (1/13)	83.91 (7/13)	83.00 (7/13)
	84.75	83.95	80.22	82.33	79.26	84.16	83.26
heart-statlog		86.07 (11/13)	85.70 (3/13)	81.78 (8/13)	84.85 (4/13)	84.52 (7/13)	84.85 (9/13)
	84.22	83.59	78.15	79.11	78.70	83.3	83.81
hepatitis		87.68 (7/19)	85.68 (5/19)	85.12 (14/19)	82.14 (6/19)	87.44 (14/19)	86.76 (14/19)
	84.62	83.81	79.22	80.85	77.43	84.29	85.03
hypothyroid		95.29 (11/29)	99.42 (7/29)	97.94 (4/29)	99.12 (5/29)	93.68 (7/29)	93.37 (12/29)
	94.48	95.30	99.54	93.21	99.39	93.44	93.49
lymphothe- rapy		87.51 (11/18)	79.53 (12/18)	85.55 (9/18)	78.02 (9/18)	85.25 (10/18)	87.15 (13/18)
	82.59	83.13	75.84	81.74	77.36	82.26	85.94
primary- tumor		49.71 (14/17)	45.96 (10/17)	46.53 (14/17)	44.88 (15/17)	25.19 (14/17)	44.19 (15/17)
	45.01	49.71	41.39	44.98	38.59	24.90	29.20
Μέση ακρίβεια μετά την επιλογή μεταβλητών		82.39	81.09	80.24	80.25	79.66	80.93
Μέση ακρίβεια πριν την επιλογή μεταβλητών	80.55	80.27	78.33	78.68	77.83	78.18	78.72

Πίνακας 63. Σύγκριση της ακρίβειας των μεθόδων μετά την εφαρμογή της προτεινόμενης υβριδικής επιλογής των μεταβλητών.

¹⁸ Ακρίβεια πρόγνωσης μετά την εφαρμογή της προτεινόμενης μεθόδου επιλογής μεταβλητών

¹⁹ (2/9) σημαίνει ότι 2 από τις 9 μεταβλητές εν τέλει χρησιμοποιήθηκαν

²⁰ Ακρίβεια πρόγνωσης πριν την εφαρμογή της προτεινόμενης μεθόδου επιλογής μεταβλητών

ΚΕΦΑΛΑΙΟ 6. Μάθηση σε ειδικά προβλήματα και συνεισφορά της διατριβής σε αυτά

Οι αλγόριθμοι μηχανικής μάθησης είναι επιθυμητό να είναι σε θέση να γενικεύσουν για οποιασδήποτε κλάση με ίδια ακρίβεια. Δηλαδή σε ένα πρόβλημα δύο κλάσεων - θετικών και αρνητικών περιπτώσεων - ο αλγόριθμος να προβλέπει με την ίδια ακρίβεια και τα θετικά και τα αρνητικά παραδείγματα. Αυτό είναι φυσικά η ιδανική κατάσταση. Σε πολλές εφαρμογές οι αλγόριθμοι καλούνται να μάθουν από ένα σύνολο στοιχείων, το οποίο περιέχει πολύ περισσότερα παραδείγματα από τη μια κλάση σε σχέση με την άλλη. Εν γένει, οι επαγωγικοί αλγόριθμοι είναι σχεδιασμένοι να ελαχιστοποιούν τα σφάλματα. Ως συνέπεια οι κλάσεις που περιέχουν λίγες περιπτώσεις μπορούν να αγνοηθούν κατά ένα μεγάλο μέρος επειδή το κόστος λανθασμένης ταξινόμησης της υπερ-αντιπροσωπευόμενης κλάσης ξεπερνά το κόστος λανθασμένης ταξινόμησης της μικρότερης κλάσης.

Έως τώρα, αντιμετωπίζαμε ένα πρόβλημα μάθησης – και συγκεκριμένα κατηγοριοποίησης – ως ένα πρόβλημα ελαχιστοποίησης της πιθανότητας εσφαλμένης ταξινόμησης. Η προσέγγιση αυτή ωστόσο κρίνεται ανεπαρκής για ένα πλήθος προβλημάτων κατηγοριοποίησης, καθώς αποδίδει την ίδια σημασία σε όλα τα είδη σφαλμάτων που δύνανται να εμφανιστούν. Η μοντελοποίηση της έννοιας του κόστους λανθασμένης ταξινόμησης αναφορικά με τη βαρύτητα ενός σφάλματος, αποτελεί αντικείμενο της Ευαίσθητης σε Κόστη Μάθησης (Cost-sensitive Learning).

Δύο από τα πλέον χρησιμοποιούμενα μέτρα αξιολόγησης συστημάτων ταξινόμησης είναι η ακρίβεια (Acc) και το σφάλμα (Err). Γίνεται εύκολα αντιληπτό ότι η χρήση των παραπάνω μέτρων για την εκτίμηση της αποτελεσματικότητάς του ταξινομητή σε ανομοιογενή προβλήματα κρίνεται μάλλον ανεπαρκής. Βάσει των παραπάνω, τα μέτρα της ακρίβειας και του σφάλματος δίνουν τη θέση τους σε αυτό του γεωμετρικού μέσου – το γινόμενο του ποσοστού των ορθά ταξινομημένων θετικών παραδειγμάτων επί το ποσοστού των ορθά ταξινομημένων αρνητικών παραδειγμάτων ($\text{TruePositiveRate} \times \text{TrueNegativeRate}$) για τα δυαδικά προβλήματα (εύκολα γενικεύεται στα πολυ-κατηγορικά προβλήματα) (Kubat & Holte, 1998). Επίσης κατασκευάζουμε το διάγραμμα συσχέτισης Ροκ (Roc curve) που περιγράφει την σχέση του ποσοστού των ορθά ταξινομημένων θετικών παραδειγμάτων με το ποσοστού των ορθά ταξινομημένων αρνητικών παραδειγμάτων (βλέπε Εικόνα 69). Όσο πιο κοντά είναι η γραφική παράσταση στον άξονα τον y τόσο καλύτερη η απόδοση του ταξινομητή.

Εικόνα 69. Καμπύλη Ροκ (ROC curve)

Το πρόβλημα των ανομοιογενή συνόλων δεδομένων εμφανίζεται και σε πολλές πραγματικές εφαρμογές όπως στην ιατρική διάγνωση, στη ρομποτική, στις διαδικασίες βιομηχανικής

παραγωγής, στην ανίχνευση λαθών δικτύων επικοινωνίας, στην αυτοματοποιημένη δοκιμή του ηλεκτρονικού εξοπλισμού, και σε πολλές άλλες περιοχές (Chawla και άλλοι, 2002).

6.1 Προτεινόμενη μέθοδος μάθησης σε ανομοιογενή (*imbalanced*) δεδομένα

Η πιο απλή μέθοδος που μπορεί να χρησιμοποιηθεί όταν τα σύνολα δεδομένων είναι ανομοιογενή είναι να δώσουμε βαρύτητα στα παραδείγματα εκπαίδευσης σύμφωνα με το συνολικό κόστος που θέλουμε να προσδώσουμε σε κάθε κατηγορία (Domingos, 1998). Για παράδειγμα, αν υποθεθεί ότι για ένα δυαδικό πρόβλημα οι περιπτώσεις της θετικής κατηγορίας είναι τέσσερις φορές περισσότερες από τις περιπτώσεις της αρνητικής κατηγορίας. Η απλή αυτή μέθοδος (ευαίσθητη στο κόστος) αυξάνει το κόστος ταξινόμησης των αρνητικών περιπτώσεων 4 φορές. Κατόπιν ο αλγόριθμος μάθησης, που στοχεύει να μειώσει τον αριθμό λαθών ταξινόμησης, θα βρει έναν ταξινομητή που είναι πιο προσεκτικός στην ταξινόμηση κάθε αρνητικού παραδείγματος αφού κάθε λάθος έχει 4 φορές μεγαλύτερη επίπτωση στην ακρίβεια ταξινόμησης.

Ένα ιδιαίτερα σημαντικό, πλην όμως υποκειμενικό ζήτημα για την απόδοση οποιουδήποτε μοντέλου μάθησης, αποτελεί ο προσδιορισμός της τιμής της παραμέτρου λ (που είχε την τιμή 4 στο προηγούμενο παράδειγμα), καθώς εκφράζει τη σημασία που αποδίδει ο χρήστης στην λανθασμένη ταξινόμηση ενός αντικειμένου της μικρότερης κλάσης (Japkowicz, 2000).

Σε πολυ-κατηγορικά προβλήματα το κόστος λαμβάνει τη μορφή ενός πίνακα, όπου το κόστος ταξινόμησης από την πραγματική κατηγορία j στην κατηγορία i αντιστοιχεί στο στοιχείο λ_{ij} του πίνακα. Αυτός ο πίνακας εκφράζει συνήθως το κόστος ανά λανθασμένη ταξινόμηση στο συγκεκριμένο πρόβλημα. Τα διαγώνια στοιχεία είναι μηδέν, αφού η σωστή ταξινόμηση ενός στοιχείου δεν έχει κανένα κόστος.

Μια άλλη απλή μέθοδος είναι αυτή του κατωφλιού που αλλάζει το όριο της απόφασης υπέρ της κλάσης μειοψηφίας (Witten & Frank, 2000).

Η Japkowicz (2002) μελέτησε την επίδραση της δυσαναλογίας σε ένα σύνολο δεδομένων. Δύο μέθοδοι επανακατασκευής (αναδιάταξης) των δεδομένων εκπαίδευσης εξετάστηκαν. Η τυχαία αύξηση των παραδειγμάτων της μικρότερης κατηγορίας έως ότου αυτή αποτελείται από τόσα δείγματα όσα η κατηγορία της πλειοψηφίας και η στοχευόμενη αύξηση των παραδειγμάτων της μικρότερης κατηγορίας επαναλαμβάνοντας μόνο τα παραδείγματα εκείνα που είναι στο όριο (boundary) ταξινόμησης (μεταξύ των κατηγοριών μειοψηφίας και πλειοψηφίας). Επίσης εξετάστηκε η τυχαία προς τα κάτω δειγματοληψία (under-sampling), η οποία υπο-επιλέγει τα παραδείγματα της πλειοψηφίας τυχαία έως ότου ο αριθμός τους ισούται με τον αριθμό των παραδειγμάτων της μειοψηφίας καθώς και η στοχευόμενη μείωση των παραδειγμάτων της μεγαλύτερης κατηγορίας απομακρύνοντας τα παραδείγματα που είναι εύκολα στην ταξινόμηση τους. Η ερευνητρια σημείωσε ότι και οι δύο προσεγγίσεις δειγματοληψίας ήταν αποτελεσματικές, αλλά επίσης παρατήρησε ότι η χρησιμοποίηση περίπλοκων τεχνικών δειγματοληψίας (στοχευόμενες δειγματοληψίες) δεν έδωσαν οποιοδήποτε επιπλέον κέρδος.

Μια άλλη γνωστή τεχνική είναι η MetaCost (Domingos, 1999). Η διαδικασία αρχίζει να εκπαιδεύει ένα αλγόριθμο βάσης με την εφαρμογή της ευαίσθητης στο κόστος μεθόδου. Κατόπιν, υπολογίζει τις πιθανότητες ανά κλάση χρησιμοποιώντας το σακούλιασμα (bagging) και αλλάζει κλάση στις περιπτώσεις κατάρτισης με αυτή που ελαχιστοποιεί το αναμενόμενο κόστος ταξινόμησης. Εν τέλει, εκπαιδεύει τον αλγόριθμο βάσης χρησιμοποιώντας το τροποποιημένο σύνολο κατάρτισης. Άλλοι ερευνητές (Fan και άλλοι, 1999) χρησιμοποίησαν αλγόριθμο προώθησης (boosting) στη θέση του σακούλιασματος (bagging) με παρόμοια αποτελέσματα.

Σε πολλά πραγματικά σύνολα δεδομένων που χρησιμοποιούμε σε αυτή την ενότητα οι περισσότερες περιπτώσεις κατανέμονται σε μια κλάση και πολύ λιγότερες σε μια

μικρότερη, συνήθως πιο ενδιαφέρουσα κλάση. Η προτεινόμενη μέθοδος αντιμετώπισης του προβλήματος των ανομοιογενών δεδομένων είναι βασισμένη στην τεχνολογία των ευφύων πρακτόρων για την ανακάλυψη της γνώσης (ABKD). Καταρχήν, η υιοθέτηση των ευφύων πρακτόρων διευκολύνει τον παραλληλισμό, βελτιώνοντας την ταχύτητα και την αποδοτικότητα των μηχανών μάθησης. Κατά δεύτερον, η τεχνολογία των ευφύων πρακτόρων βοηθάει στο σχεδιασμό καταναεμημένων συστημάτων μάθησης.

Συνοπτικά, η προσέγγισή μας είναι να χρησιμοποιήσουμε τρεις πράκτορες (ο πρώτος εκπαιδεύεται με τον NB αλγόριθμο, ο δεύτερος με τον C4.5 και ο τρίτος με τον 5NN) σε μια φιλτραρισμένη έκδοση των δεδομένων εκπαίδευσης και να συνδυάσουμε τις προβλέψεις τους. Η τεχνική μας προσπαθεί να επιτύχει την ποικιλομορφία στα λάθη ταξινόμησης με τη χρησιμοποίηση των διαφορετικών αλγορίθμων μάθησης.

Χρησιμοποιήσαμε επίσης αλγόριθμο επιλογής χαρακτηριστικών επειδή στα μικρά σύνολα δεδομένων (με πολλές μεταβλητές) το ποσό δυσαναλογίας έχει μεγαλύτερη επίπτωση στην επαγωγή και έτσι η επιλογή χαρακτηριστικών καθιστά το πρόβλημα λιγότερο δύσκολο. Τα χαρακτηριστικά επιλέγονται με την κράτηση εκείνων για τα οποία η αναλογία κέρδους (gain ratio) υπερβαίνει ένα σταθερό κατώτατο όριο ϵ . Θέσαμε το ϵ ίσον με 0.02, σε μία προσπάθεια να απορριφθούν μόνο τα χαρακτηριστικά με αμελητέο αντίκτυπο στις προβλέψεις. Εντούτοις, ακόμα και ένα τόσο χαμηλό κατώτατο όριο μπορεί να απορρίψει πολλά χαρακτηριστικά σε μεγάλα προβλήματα.

Αναλυτικότερα, το προτεινόμενο σύνολο ειδικών πρακτόρων αποτελείται από έναν πράκτορα «πρόεδρο» που φιλτράρει τα χαρακτηριστικά του συνόλου εκπαίδευσης (με την χρήση του αλγόριθμου επιλογής μεταβλητών που ήδη περιγράφηκε) και περνά ένα αντίγραφο των περιπτώσεων στους μαθητευόμενους πράκτορες (learning agents). Κατόπιν, κάθε μαθητευόμενος πράκτορας αναδιατάσσει τα σύνολα δεδομένων που έλαβε έτσι ώστε η σχέση κόστους ταξινόμησης μεταξύ των αρνητικών και των θετικών παραδειγμάτων να είναι παρόμοια (ευαίσθηση στο κόστος μέθοδος). Μετά την εκπαίδευση των μαθητευόμενων πρακτόρων και κατά την διάρκεια της ταξινόμησης κάθε μαθητευόμενος πράκτορας δίνει την πρόβλεψη του για την νέα περίπτωση στον πράκτορα «πρόεδρο». Τέλος, ο πράκτορας «πρόεδρος» παίρνει την τελική πρόβλεψη σύμφωνα με την πλειοψηφική ψηφοφορία.

Ακολουθεί η σχηματική αναπαράσταση της μεθόδου μας.

Εικόνα 70. Η γραφική αναπαράσταση της προτεινόμενης μεθόδου συνδυασμού ευφυών πρακτόρων

Στον Πίνακα 64, κάποιος μπορεί να δει τις συγκρίσεις του προτεινόμενου συνόλου ειδικών πρακτόρων με τις άλλες μεθόδους που έχουν προσπαθήσει να λάβουν την καλύτερη απόδοση σε ανομοιογενή σύνολο δεδομένων χρησιμοποιώντας τον αφελή Μπεϋζιανό ταξινομητή (NB) ως βασικό ταξινομητή. Πέντε γνωστοί αλγόριθμοι χρησιμοποιήθηκαν για τη σύγκριση: η μέθοδος του κατωφλιού (ThresNB) που αλλάζει το όριο της απόφασης υπέρ της κλάσης μειοψηφίας (Witten & Frank, 2000), η ευαίσθητη σε κόστος μέθοδος (CostNB) (Domingos, 1998), η εξισορρόπηση (αναδιάταξη) του δείγματος (ReWNB) (Japkowicz, 2002), ο Adaboost σε συνεργασία με την ευαίσθητη σε κόστος μέθοδο (AdabcostNB) (Schapire και άλλοι, 1998) και ο αλγόριθμος Metacost (Domingos, 1999). Παρουσιάζουμε επίσης την ακρίβεια του απλού αλγορίθμου NB ως μέτρο σύγκρισης.

Πρέπει να αναφερθεί, ότι στον Πίνακα 64 εκτός από το γεωμετρικό μέσο όρο (g) παρουσιάζουμε επίσης το ποσοστό των ορθά ταξινομημένων θετικών παραδειγμάτων (α^+ : true positive rate), και το ποσοστό των ορθά ταξινομημένων αρνητικών παραδειγμάτων (α^- : true negative rate). Πρέπει σε αυτό το σημείο να σημειωθεί ότι η θετική κατηγορία για τα πειράματά μας θεωρείται ότι είναι η κατηγορία της πλειοψηφίας. Στην τελευταία σειρά του Πίνακα 64, υπολογίζεται η μέση τιμή των γεωμετρικών μέσων για όλα τα σύνολα δεδομένων. Για τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Επιπλέον, στον πίνακα, συμβολίζεται με "v" ότι η προτεινόμενη μέθοδος έδωσε στατιστικά καλύτερα αποτελέσματα από τον συγκρινόμενο αλγόριθμο, ενώ με "*" συμβολίζεται το αντίθετο. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$).

Δεδομένα		ABKD	ThresNB	ReWNB	CostNB	AdabcostNB	MetacostNB	NB
breast-cancer	g	0.62	0.63v	0.66v	0.66v	0.63v	0.65v	0.6
	α^+	0.72	0.62	0.74	0.74	0.72	0.79	0.85
	α^-	0.53	0.65	0.58	0.58	0.56	0.54	0.43
credit-g	g	0.67	0.71v	0.72v	0.72v	0.71v	0.66	0.65
	α^+	0.7	0.69	0.75	0.75	0.75	0.77	0.86
	α^-	0.65	0.74	0.69	0.69	0.67	0.57	0.49
diabetes	g	0.7	0.72	0.73	0.73	0.73	0.70	0.71
	α^+	0.72	0.65	0.78	0.78	0.77	0.75	0.84
	α^-	0.69	0.8	0.68	0.68	0.69	0.66	0.6
haberman	g	0.61v	0.59v	0.56v	0.56v	0.56v	0.57v	0.44
	α^+	0.77	0.64	0.89	0.89	0.88	0.87	0.94
	α^-	0.49	0.55	0.35	0.35	0.36	0.38	0.21
hepatitis	g	0.79	0.76	0.8	0.8	0.78	0.81v	0.78
	α^+	0.84	0.87	0.83	0.83	0.86	0.79	0.87
	α^-	0.75	0.67	0.78	0.78	0.71	0.84	0.7
ionosphere	g	0.9v	0.88v	0.82	0.82	0.91v	0.77*	0.83
	α^+	0.96	0.93	0.78	0.78	0.93	0.68	0.8
	α^-	0.84	0.81	0.87	0.87	0.9	0.88	0.86
eap	g	0.83	0.83	0.85	0.85	0.83	0.85	0.84
	α^+	0.87	0.86	0.87	0.87	0.85	0.88	0.9
	α^-	0.8	0.81	0.83	0.83	0.82	0.83	0.78
sick	g	0.93v	0.76*	0.86	0.86	0.87	0.8*	0.86
	α^+	0.95	0.98	0.82	0.82	0.88	0.73	0.94
	α^-	0.91	0.59	0.9	0.9	0.86	0.87	0.78
Μέση τιμή	g	0.76	0.74	0.75	0.75	0.75	0.73	0.71

Πίνακας 64. Σύγκριση της προτεινόμενης μεθόδου (ABKD) με άλλες που χρησιμοποιούν ως βασικό ταξινομητή τον NB

Γενικά, όλες οι τεχνικές δίνουν καλύτερα αποτελέσματα από την απλή εφαρμογή του αφελή ταξινομητή Μπέιζ. Η πιο αξιοπρόσεκτη βελτίωση προέρχεται από την τεχνική μας (ABKD), αν και η μέθοδος του κατωφλιού (Threshold method) δίνει, κατά μέσον όρο, την καλύτερη ακρίβεια για την κατηγορία (κλάση) μειοψηφίας. Ο αλγόριθμος Metacost δεν μπορεί να βελτιώσει τα αποτελέσματα του NB όπως υποπτευόταν ο κατασκευαστής του. Πρέπει να αναφερθεί ότι για τον ταξινομητή NB, η ευαίσθητη στο κόστος μέθοδος (Cost Sensitive method) είναι ισοδύναμη με τη μέθοδο που αναδιατάσσει τις περιπτώσεις κατάρτισης έτσι ώστε η σχέση μεταξύ των ψεύτικων αρνητικών και των ψεύτικων θετικών δαπανών να είναι αντεστραμμένη από την αρχική. Επιπλέον, ο Adaboost με την προσθήκη ευαισθησίας κόστους δεν μπορεί να δώσει τα καλύτερα αποτελέσματα από την ευαίσθητη στο κόστος μέθοδο και τη μέθοδο αναδιάταξης, αν και είναι μια περισσότερο χρονοβόρα μέθοδος.

Στον Πίνακα 65, κάποιος μπορεί να δει τις συγκρίσεις του προτεινόμενου συνόλου ειδικών πρακτόρων με τις άλλες μεθόδους που έχουν προσπαθήσει να λάβουν την καλύτερη απόδοση σε ανομοιογενή σύνολο δεδομένων χρησιμοποιώντας τον αλγόριθμο C4.5 ως βασικό ταξινομητή.

Δεδομένα		ABKD	ThresC4.5	ReWC4.5	CostC4.5	Adabcost C4.5	Metacost C4.5	C4.5
breast-cancer	g	0.62v	0.45*	0.57v	0.5	0.56v	0.55v	0.5
	α^+	0.72	0.8	0.72	0.85	0.77	0.84	0.95
	α^-	0.53	0.25	0.45	0.3	0.41	0.36	0.26
credit-g	g	0.67v	0.64v	0.66v	0.61v	0.62v	0.64v	0.58
	α^+	0.7	0.7	0.67	0.82	0.81	0.76	0.85
	α^-	0.65	0.58	0.65	0.46	0.47	0.54	0.4
diabetes	g	0.7	0.7	0.72	0.72	0.67*	0.73v	0.7
	α^+	0.72	0.69	0.72	0.78	0.79	0.78	0.82
	α^-	0.69	0.71	0.73	0.67	0.57	0.67	0.6
haberman	g	0.61v	0.56v	0.63v	0.58v	0.57v	0.62v	0.52
	α^+	0.77	0.61	0.68	0.66	0.76	0.76	0.85
	α^-	0.49	0.51	0.58	0.51	0.43	0.52	0.32
hepatitis	g	0.79v	0.62v	0.73v	0.64v	0.7v	0.68v	0.58
	α^+	0.84	0.78	0.62	0.86	0.9	0.83	0.9
	α^-	0.75	0.49	0.85	0.48	0.55	0.56	0.37
ionosphere	g	0.9v	0.88	0.89	0.88	0.90	0.9	0.88
	α^+	0.96	0.95	0.94	0.94	0.94	0.98	0.94
	α^-	0.84	0.81	0.85	0.82	0.86	0.82	0.82
eap	g	0.83	0.69*	0.81	0.83	0.79*	0.82	0.83
	α^+	0.87	0.91	0.86	0.94	0.85	0.89	0.94
	α^-	0.8	0.53	0.77	0.74	0.74	0.76	0.74
sick	g	0.93	0.92	0.97v	0.96v	0.95	0.96v	0.93
	α^+	0.95	0.99	0.99	0.99	1	0.98	0.99
	α^-	0.91	0.85	0.95	0.92	0.9	0.95	0.87
Μέση τιμή	g	0.76	0.68	0.75	0.72	0.72	0.74	0.69

Πίνακας 65. Σύγκριση της προτεινόμενης μεθόδου (ABKD) με άλλες που χρησιμοποιούν ως βασικό ταξινομητή τον C4.5

Παρόμοια με το προηγούμενο πείραμα, η μεθοδός μας έχει την καλύτερη απόδοση από τις άλλες τεχνικές. Εντούτοις, ο αλγόριθμος Metacost έχει πραγματικά καλύτερη απόδοση συνδυαζόμενος με τον C4.5 ταξινομητή σε σχέση με τον NB. Πρέπει όμως να αναφερθεί ότι η μέθοδος του κατωφλιού δίνει χειρότερη απόδοση από τον απλό C4.5. Για τον C4.5 η ευαίσθητη στο κόστος μέθοδος είναι λιγότερο αποδοτική από την αναδιάταξη των περιπτώσεων εκπαίδευσης έτσι ώστε η σχέση μεταξύ των ψεύτικων αρνητικών και ψεύτικων θετικών να είναι ανεστραμμένη. Ο λόγος μπορεί να είναι ότι τα κλαδεμένα δέντρα απόφασης δεν μπορούν να υπολογίσουν πολύ καλά την πιθανότητα πρόβλεψης ανά κατηγορία. Ο Adaboost συνδυασμένος με την ευαίσθητη στο κόστος μέθοδο, βάσει του προηγούμενου πειράματος, δεν μπορεί να δώσει καλύτερα αποτελέσματα από την απλή ευαίσθητη στο κόστος μέθοδο αν και χρησιμοποιεί περισσότερο χρόνο για την εκπαίδευσή του.

Στον Πίνακα 66, κάποιος μπορεί να δει τις συγκρίσεις του προτεινόμενου συνόλου ειδικών πρακτόρων με τις άλλες μεθόδους που έχουν προσπαθήσει να λάβουν την καλύτερη απόδοση σε ανομοιογενή σύνολο δεδομένων χρησιμοποιώντας τον αλγόριθμο 5NN ως βασικό ταξινομητή.

Δεδομένα		ABKD	Thres 5NN	ReW 5NN	Cost 5NN	Adabcos 5NN	Metacost 5NN	5NN
breast-cancer	g	0.62v	0.6v	0.62v	0.61v	0.61v	0.51v	0.45
	α^+	0.72	0.57	0.73	0.72	0.7	0.86	0.96
	α^-	0.53	0.63	0.52	0.52	0.53	0.3	0.21
credit-g	g	0.67v	0.59	0.66v	0.66v	0.63v	0.63v	0.57
	α^+	0.7	0.84	0.69	0.69	0.7	0.73	0.89
	α^-	0.65	0.42	0.63	0.63	0.56	0.55	0.37
diabetes	g	0.7	0.69	0.71v	0.71v	0.66	0.71v	0.68
	α^+	0.72	0.79	0.69	0.69	0.71	0.75	0.83
	α^-	0.69	0.61	0.74	0.74	0.62	0.68	0.56
haberman	g	0.61v	0.58v	0.57v	0.57v	0.53v	0.59v	0.39
	α^+	0.77	0.52	0.68	0.68	0.68	0.66	0.9
	α^-	0.49	0.65	0.47	0.47	0.41	0.52	0.17
hepatitis	g	0.79v	0.68	0.69v	0.73v	0.58*	0.8v	0.66
	α^+	0.84	0.91	0.79	0.85	0.8	0.84	0.94
	α^-	0.75	0.51	0.6	0.62	0.42	0.76	0.46
ionosphere	g	0.9v	0.82v	0.83v	0.83v	0.83v	0.79	0.78
	α^+	0.96	0.97	0.97	0.97	0.95	0.98	0.98
	α^-	0.84	0.7	0.71	0.71	0.72	0.63	0.62
eap	g	0.83v	0.79	0.8	0.8	0.78	0.77	0.78
	α^+	0.87	0.83	0.84	0.84	0.79	0.87	0.9
	α^-	0.8	0.75	0.76	0.76	0.77	0.69	0.68
sick	g	0.93v	0.62	0.84v	0.84v	0.87v	0.79v	0.61
	α^+	0.95	0.99	0.89	0.89	0.98	0.9	0.99
	α^-	0.91	0.39	0.79	0.79	0.77	0.7	0.37
Μέση τιμή	g	0.76	0.67	0.72	0.72	0.69	0.7	0.62

Πίνακας 66. Σύγκριση της προτεινόμενης μεθόδου (ABKD) με άλλες που χρησιμοποιούν ως βασικό ταξινομητή τον 5NN

Όλες οι τεχνικές έδωσαν καλύτερα αποτελέσματα από τον απλό 5NN αλγόριθμο. Εντούτοις, η προτεινόμενη μέθοδος έχει τον καλύτερο γεωμετρικό μέσο όρο ακριβειών και σε αυτό το πείραμα. Πρέπει να αναφερθεί ότι ο αλγόριθμος Adaboost σε συνδυασμό με την ευαίσθητη στο κόστος μέθοδο και ο αλγόριθμος Metacost είναι εξαιρετικά αργοί στην εκπαίδευση εάν συνδυαστούν με τον σκληρό αλγόριθμο 5NN, χωρίς προσφορά θεαματικής βελτίωσης στην απόδοση. Για τον ταξινομητή 5NN, η μέθοδος τροποποίησης του ορίου απόφασης είναι το ίδιο αποδοτική με την μέθοδο της αναδιάταξης των δεδομένων εκπαίδευσης. Τέλος, η μέθοδος του κατωφλιού δίνει, κατά μέσον όρο, τη λιγότερη βελτίωση.

6.2 Προτεινόμενη μέθοδος μάθησης σε πραγματικό χρόνο

Σε όλες τις προηγούμενες περιπτώσεις, ο εκπαιδευόμενος ταξινομητής έχει εκ των προτέρων όλα τα διαθέσιμα δεδομένα εκπαίδευσης, και τα χρησιμοποιεί για να κατασκευάσει μια υπόθεση, η οποία χρησιμοποιείται έκτοτε για την ταξινόμηση χωρίς περαιτέρω τροποποίηση. Στην επανζητική μάθηση, ο ταξινομητής τροποποιεί συνεχώς την υπόθεσή του δεδομένου ότι δεν έχει όλα τα διαθέσιμα δεδομένα από την αρχή της εκπαίδευσής του αλλά τα λαμβάνει κατά την διάρκεια της λειτουργίας του. Οι αλγόριθμοι μάθησης σε πραγματικό χρόνο

προσπαθούν να ξεχάσουν τις άσχετες πληροφορίες αντί να συνθέσουν όλες τις διαθέσιμες πληροφορίες (όπως πραγματοποιούν οι κλασικοί αλγόριθμοι μάθησης).

Η επαυξητική δυνατότητα μάθησης είναι πολύ σημαντική σε ορισμένα προβλήματα μάθησης για δύο λόγους. Πρώτον, σε ορισμένα προβλήματα είναι ασύμφορο να συλλεχθούν όλα τα χρήσιμα παραδείγματα μάθησης προτού τεθεί το σύστημα σε χρήση. Επομένως όταν νέα παραδείγματα είναι διαθέσιμα, η μηχανή μάθησης πρέπει να έχει τη δυνατότητα να κάνει μερικές μικρές αλλαγές στην υπόθεση του συστήματος έτσι ώστε η νέα γνώση που υπάρχει στα νέα παραδείγματα να ενσωματωθεί. Δεύτερον, η τροποποίηση ενός εκπαιδευμένου συστήματος είναι πολύ λιγότερο χρονοβόρα διαδικασία από το χτίσιμο ενός νέου συστήματος από την αρχή, ειδικά στις εφαρμογές πραγματικού χρόνου.

Παράδειγμα όπου είναι αναγκαία η χρήση επαυξητικής μάθησης, είναι η ανίχνευση των προτιμήσεων των χρηστών στα ηλεκτρονικά καταστήματα. Όπως είναι εύκολα αντιληπτό είναι αναμενόμενη η αλλαγή των προτιμήσεων των πελατών σε νέα προϊόντα και υπηρεσίες που διατίθενται. Οι αλγόριθμοι μάθησης πρέπει λοιπόν να συγκλίνουν γρήγορα και με ικανοποιητική ακρίβεια στις αλλαγές της συνάρτησης στόχου, ενώ να είναι αποδοτικοί όσον αφορά το χρόνο και το χώρο εκπαίδευσης.

Τα επιθυμητά χαρακτηριστικά ενός επαυξητικού αλγόριθμου μάθησης είναι λοιπόν:

1. η δυνατότητα να ανιχνεύει γρήγορα μια αλλαγή στην υπόθεση του συστήματος.
2. η δυνατότητα να προσαρμόζει την υπόθεση του συστήματος άμεσα σε οποιαδήποτε αλλαγή (concept drift).

Σήμερα, αν και ο συνδυασμός των ταξινομητών προτείνεται ως νέα κατεύθυνση για τη βελτίωση της ακρίβειας ταξινόμησης, εντούτοις, οι περισσότεροι αλγόριθμοι συνδυασμού ταξινομητών δεν λειτουργούν σε πραγματικό χρόνο (Fern & Givan, 2000). Για αυτόν τον λόγο, προτείνουμε μια ομάδα ταξινομητών που λειτουργεί σε πραγματικό χρόνο και συνδυάζει μια επαυξητική έκδοση του NB, τον Voted Perceptron και τον Winnow αλγόριθμο χρησιμοποιώντας τη μεθοδολογία ψηφοφορίας (βλέπε Εικόνα 71) [5].

Δεν υπάρχει κάποια ιδιαίτερη δυσκολία που να αποτρέπει τη χρησιμοποίηση του αλγόριθμου NB για να μάθει επαυξητικά. Στο βήμα # 1, αρχικοποιούμε όλες τις τιμές των αθροιστών (counters) εμφάνισης κάθε τιμής μιας μεταβλητής σε 0 και έπειτα ο αλγόριθμος χρησιμοποιεί ένα παράδειγμα κατάρτισης τη φορά. Κάθε παράδειγμα κατάρτισης, όπως είναι γνωστό, δίνεται ως διάνυσμα των χαρακτηριστικών και της τιμής της κλάσης. Ο αλγόριθμος περνώντας από το διάνυσμα αυξάνει τους κατάλληλους αθροιστές (counters) για κάθε τιμή της εκάστοτε μεταβλητής. Στο βήμα # 2, αυτοί οι αθροιστές μετατρέπονται σε πιθανότητες με τη διαίρεση κάθε αθροιστή με τον αριθμό παραδειγμάτων κατάρτισης της ίδιας κατηγορίας. Το τελικό βήμα (#3) υπολογίζει τις προγενέστερες (prior) πιθανότητες $p(k)$ ως κλάσμα όλων παραδειγμάτων κατάρτισης που ανήκουν στην ίδια κατηγορία (k).

Οι αλγόριθμοι Voted Perceptron και Winnow είναι επαυξητικοί αλγόριθμοι αλλά εφαρμόζονται απευθείας μόνο σε προβλήματα δυο κλάσεων. Για αυτόν τον λόγο, η προτεινόμενη μέθοδος μετατρέπει τα πολυ-κατηγορικά (multi-class) προβλήματα σε ένα σύνολο δυαδικών προβλημάτων. Έχουμε χρησιμοποιήσει την απλή μέθοδο μια κλάση εναντίον όλων (One Per Class). Όπως έχουμε ξανααναφέρει ο ταξινομητής i εκπαιδεύεται με όλα τα παραδείγματα της κατηγορίας i ως θετικά παραδείγματα, και όλα τα άλλα παραδείγματα ως αρνητικά παραδείγματα. Η τελική απόφαση είναι η κατηγορία με τις περισσότερες ψήφους.

Εικόνα 71. Η γραφική αναπαράσταση της προτεινόμενης μεθόδου επανζητικής μάθησης (Voting Online Ensemble)

Στο πρώτο πείραμα συγκρίναμε την προτεινόμενη μέθοδο με καθένα από τους αλγόριθμους βάσης. Για τα πειράματα χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten-fold stratified cross validation), προκειμένου να προσεγγιστεί κατά το δυνατόν η πραγματική απόδοση των αλγορίθμων. Τα τελικά αποτελέσματα προέκυψαν από το μέσο όρο των δέκα επιμέρους αποτελεσμάτων. Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η κατά ζεύγη δοκιμή-t ($p < 0.05$). Συγκεντρωτικά, στην προτελευταία γραμμή του Πίνακα 67 παρουσιάζονται συνοπτικά τα αποτελέσματα με την μορφή (a/b/c) όπως και στα πειράματα της προηγούμενης ενότητας. Επίσης, στην τελευταία γραμμή του Πίνακα 67 παρουσιάζεται η μέση ακρίβεια κάθε αλγορίθμου σε όλα τα εξεταζόμενα σύνολα δεδομένων.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον WINNOW σε 23 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων (βλέπε Πίνακα 67). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον Voted Perceptron σε 5 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων. Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον επανζητικό NB σε 6 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε μόλις ένα σύνολο δεδομένων.

Συγκεντρωτικά, τα αποτελέσματα έδειξαν λοιπόν την ανωτερότητα της προτεινόμενης μεθόδου.

Δεδομένα	Voting online ensemble	WINNOW	VOTED PERCEPTRON	NB
audiology	72.64	40.24*	48.71*	72.64
badges	99.93	100	99.83	99.66
balance	85.34	61.65 *	75.12 *	90.53 v
breast-cancer	72.48	61.96 *	71.67	72.7
breast-w	96.61	94.68	96.3	96.07
Colic	82.23	76.61 *	82.2	78.7
Credit-a	84.06	76.78 *	85.52	77.86 *
credit-g	75.07	64.55 *	74.41	75.16
Diabetes	75.56	68.62 *	75.5	75.75
Haberman	73.1	71.77	72.41	75.06
heart-c	83.73	69.38 *	80.79	83.34
heart-h	84.5	65.38 *	81.98	83.95
heart-statlog	83.41	77.22 *	81.85	83.59
Hepatitis	84.01	74.91 *	82.93	83.81
Ionosphere	91.08	87.82 *	89.8	82.17 *
Iris	95.33	80.6 *	94.27	95.53
Labor	90.07	79.03 *	86.13	93.57
Lymphotherapy	83.01	67.55 *	78.74	83.13
Monk3	92.63	79.57 *	88.19	93.45
primary-tumor	49.77	20.73 *	13.52 *	49.71
Sick	96.95	92.16 *	97.56	92.75 *
sonar	73.93	67.78	74.22	67.71 *
Soybean	92.93	68.91 *	78.76 *	92.94
titanic	77.85	63.09 *	77.93	77.85
vote	94.04	91.26	94.82	90.02 *
Waveform	82.36	65.81 *	81.8	80.01 *
wine	97.63	89.63 *	95.61	97.46
Zoo	95.07	79.91 *	87.02 *	94.97
		0/5/23	0/23/5	1/21/6
Μέση ακρίβεια	84.48	72.77	80.27	83.57

Πίνακας 67. Συγκρίνοντας την προτεινόμενο μέθοδο (Voting online ensemble) με τους αλγορίθμους βάσης

Κατά τη διάρκεια του δεύτερου πειράματος, ένας αντιπροσωπευτικός αλγόριθμος για κάθε μια τεχνική μάθησης συγκρίθηκε με τον προτεινόμενο αλγόριθμο. Χρησιμοποιήσαμε τους γνωστούς αλγορίθμους ως ανώτερο μέτρο της ακρίβειας της μάθησης των επαυξητικών αλγορίθμων. Οι περισσότερες από τις επαυξητικές εκδόσεις των ίδιων αλγορίθμων δεν είναι χωρίς απώλειες (Utgoff και άλλοι, 1997), (Saad, 1998) και (Widmer & Kubat, 1996). Ένας χωρίς απώλειες επαυξητικός αλγόριθμος μάθησης θα ήταν εκείνος ο αλγόριθμος που θα επέστρεφε την ίδια υπόθεση με αυτή που ο αντίστοιχος αλγόριθμος (batch) θα επέστρεφε χρησιμοποιώντας το σύνολο των δεδομένων κατάρτισης.

Ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον C4.5 σε 9 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε 2 σύνολα δεδομένων (βλέπε Πίνακα 68). Αντίστοιχα, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον 3NN σε 3 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε ένα σύνολο δεδομένων. Παρόμοια, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος

από τον RIPPER σε 8 σύνολα δεδομένων, ενώ έχει στατιστικά υψηλότερα ποσοστά λάθους σε μόλις ένα σύνολο δεδομένων. Τέλος, ο προτεινόμενος αλγόριθμος είναι στατιστικά ακριβέστερος από τον BP σε 7 σύνολα δεδομένων, ενώ δεν έχει στατιστικά υψηλότερα ποσοστά λάθους σε κανένα σύνολο δεδομένων.

Δεδομένα	Voting online ensemble	C4.5	3NN	RIPPER	BP
audiology	72.64	77.26	67.97	72.57	43.82 *
badges	99.93	100	100	100	100
balance	85.34	77.82*	86.74	80.91*	85.67
breast-cancer	72.48	74.28	73.13	71.65	71,18
breast-w	96.61	95.01*	96.61	95.72	95,97
Colic	82.23	85.16	80.95	84.97	82,58
credit-a	84.06	85.57	84.96	85.33	85.94
credit-g	75.07	71.25*	72.21	71.86	72,75
Diabetes	75.56	74.49	73.86	75.22	74.64
Haberman	73.1	71.05	69.77	72.43	76,56
heart-c	83.73	76.94*	81.82	79.05	81,39
heart-h	84.5	80.22*	82.33	79.26 *	81,37
heart-statlog	83.41	78.15*	79.11	78.70	82,11
Hepatitis	84.01	79.22	80.85	77.43 *	81,3
Ionosphere	91.08	89.74	86.02*	89.30	85,84*
Iris	95.33	94.73	95.20	93.93	96,27
Labor	90.07	78.60	92.83	83.37	88,57
Lymphotherapy	83.01	75.84	81.74	77.36	80,24
Monk3	92.63	92.95	86.72	83.99*	88.69
primary-tumor	49.77	41.39*	44.98*	38.59 *	28,11*
Sick	96.95	98.72 v	96.21	98.30 v	96,78
sonar	73.93	73.61	83.76 v	75.45	73.3
Soybean	92.93	91.78	91.2	91.93	33,43*
titanic	77.85	78.55	78.9	77.97	78.25
vote	94.04	96.57v	93.08	95.7	96.32
Waveform	82.36	75.25 *	77.67*	79.14*	59.62*
wine	97.63	93.2 *	95.85	92.48*	83.08 *
Zoo	95.07	92.61	92.61	87.03 *	61,21*
		2/17/9	1/24/3	1/19/8	0/21/7
Μέση ακρίβεια	84.48	82.14	83.11	81.77	77.32

Πίνακας 68. Σύγκριση της προτεινόμενης μεθόδου (Voting online ensemble) με άλλους γνωστούς ταξινομητές

Συνολικά, τα αποτελέσματα έδειξαν και σε αυτή την περίπτωση την ανωτερότητα της προτεινόμενης μεθόδου.

6.3 Προτεινόμενη μέθοδος μάθησης βαθμωτής συνάρτησης στόχου

Μια κλάση προβλημάτων μεταξύ της ταξινόμησης και της παλινδρόμησης, όπου ο αλγόριθμος μάθησης καλείται να προβλέπει διατεταγμένες κλάσεις, δηλαδή κλάσεις με γραμμική διάταξη, δεν έχει λάβει μέχρι σήμερα πολλή προσοχή, ακόμα κι αν υπάρχουν πολλά προβλήματα στον πραγματικό κόσμο που ανήκουν σε αυτή την κατηγορία. Στην

περίπτωση των διαταγμένων κλάσεων, μας ενδιαφέρει όχι μόνο η μεγιστοποίηση της ακρίβειας ταξινόμησης, αλλά και η ελαχιστοποίηση των αποστάσεων μεταξύ των πραγματικών και των κλάσεων πρόβλεψης.

Οι αλγόριθμοι ταξινόμησης μπορούν να εφαρμοστούν στα προβλήματα βαθμωτής συνάρτησης στόχου με την απόρριψη της διάταξης στην μεταβλητή στόχο. Εντούτοις, όμως όπως είναι εύκολα αντιληπτό κάποιες πληροφορίες που θα μπορούσαν να βελτιώσουν την απόδοση του συστήματος χάνονται όταν γίνεται αυτό.

Η χρήση των αλγορίθμων παλινδρόμησης έχει επίσης χρησιμοποιηθεί για την επίλυση προβλημάτων βαθμωτής συνάρτησης στόχου (Kramer και άλλοι, 2001). Για αυτήν την προσέγγιση κάθε τιμή της μεταβλητής στόχου πρέπει να αναπαρασταθεί με μια αριθμητική τιμή. Σε αυτή την περίπτωση όμως δεν είναι εύκολο να αναπαρασταθεί η απόσταση μεταξύ των κλάσεων με την απόσταση μεταξύ δύο αριθμών.

Μια άλλη προσέγγιση είναι να περιοριστεί το πρόβλημα ταξινόμησης πολλαπλών κλάσεων (multi-class) σε ένα σύνολο δυαδικών προβλημάτων ταξινόμησης χρησιμοποιώντας την ένα-εναντίων-όλων (one-per-class) προσέγγιση. Όπως, έχει αναφερθεί σε προηγούμενη ενότητα, στην ένα-εναντίων-όλων προσέγγιση, εκπαιδεύουμε έναν ταξινομητή για κάθε μια από τις κατηγορίες χρησιμοποιώντας για θετικά παραδείγματα τα παραδείγματα κατάρτισης που ανήκουν σε εκείνη την κατηγορία, και ως αρνητικά όλα τα άλλα παραδείγματα κατάρτισης. Συνδυάζονται έπειτα οι εκτιμήσεις που δίνονται από κάθε δυαδικό ταξινομητή προκειμένου να ληφθούν οι εκτιμήσεις πιθανότητας για κάθε τιμή της μεταβλητής στόχου για το πρόβλημα ταξινόμησης πολλαπλών κλάσεων.

Παρόμοια με την προηγούμενη μέθοδο, μια πιο εξελιγμένη μέθοδος μετατρέπει το αρχικό πρόβλημα σε μια σειρά δυαδικών προβλημάτων, όμως, για να προβλέψει την τιμή της κλάσης μιας νέας περίπτωσης αυτός ο αλγόριθμος υπολογίζει τις πιθανότητες για κάθε κλάση από τις K χρησιμοποιώντας $K-1$ μοντέλα (Frank & Hall, 2001). Παραδείγματος χάριν, για ένα πρόβλημα τριών κλάσεων, η εκτίμηση της πιθανότητας για την πρώτη στη σειρά κλάση υπολογίζεται από τον ταξινομητή βρίσκοντας την πιθανότητα: $P(\text{Target} < \text{πρώτα τιμή})$. Επίσης, η εκτίμηση της πιθανότητας για την τελευταία στη σειρά κλάση υπολογίζεται από το βασικό ταξινομητή βρίσκοντας την πιθανότητα: $P(\text{Target} > \text{δεύτερη τιμή})$. Τέλος, για την μεσαία κλάση, η πιθανότητα εξαρτάται από το ζευγάρι των προηγούμενων ταξινομητών και δίνεται από την πιθανότητα: $P(\text{Target} > \text{πρώτα τιμή}) * (1 - P(\text{Target} > \text{δεύτερη τιμή}))$.

Σε αυτή την ενότητα προτείνουμε μια τεχνική που χρησιμοποιεί σταθερό αλλά άνισο κόστος για κάθε λάθος ταξινόμησης μεταξύ των κλάσεων. Η τεχνική αυτή μπορεί να αποτελέσει μια πιο εύρωστη λύση στο πρόβλημα βαθμωτής συνάρτησης στόχου επειδή ελαχιστοποιεί τις αποστάσεις μεταξύ των πραγματικών και προβλεφθεισών κλάσεων, χωρίς να βλάπτει αλλά στην πραγματικότητα ελαφρώς να βελτιώνει την ακρίβεια ταξινόμησης.

Αναλυτικότερα, έχει ενσωματωθεί στη λήψη αποφάσεων ένα σταθερό και άνισο κόστος λανθασμένης ταξινόμησης μεταξύ των κατηγοριών. Το κόστος αυτό λαμβάνει τη μορφή ενός πίνακα, όπου το κόστος ταξινόμησης από την πραγματική κατηγορία j στην κατηγορία i αντιστοιχεί στο στοιχείο λ_{ij} του πίνακα. Τα διαγώνια στοιχεία είναι μηδέν, αφού η σωστή ταξινόμηση ενός στοιχείου δεν έχει κανένα κόστος. Μπορούμε να καθορίσουμε τον υπό όρους κίνδυνο για μια απόφαση a_i όπως:

$$R(a_i | x) = \sum_j \lambda_{ij} P(v_j | x)$$

Η εξίσωση δηλώνει ότι ο κίνδυνος επιλογής της κατηγορίας i καθορίζεται από το σταθερό κόστος ταξινόμησης ανά κατηγορία και την αβεβαιότητα της γνώσης μας για την αληθινή κατηγορία του παραδείγματος x . Ο στόχος στην ταξινόμηση είναι να ελαχιστοποιηθεί το κόστος της λάθος ταξινόμησης, το οποίο μπορεί να πραγματοποιηθεί με την επιλογή της κατηγορίας (v_j) με τον ελάχιστο υπό όρους κίνδυνο. Κατά συνέπεια, ένας πίνακας κόστους

για ένα πρόβλημα ταξινόμησης βαθμωτής συνάρτησης στόχου τριών κατηγοριών θα μπορούσε να είναι:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

Σε ένα πρόβλημα N κλάσεων ο πίνακας κόστους θα μπορούσε να είναι:

$$\begin{bmatrix} 0 & 1 & 2 & \dots & n-1 \\ 1 & 0 & 1 & \dots & n-2 \\ \dots & \dots & \dots & \dots & \dots \\ n-2 & \dots & 1 & 0 & 1 \\ n-1 & \dots & 2 & 1 & 0 \end{bmatrix}$$

Ένα κύριο χαρακτηριστικό της μεθόδου μας είναι ότι δεν απαιτεί οποιαδήποτε τροποποίηση του χρησιμοποιούμενου αλγορίθμου μάθησης εφ' όσον αυτός παράγει εκτιμήσεις πιθανότητας ανά κατηγορία. Παρακάτω, αξιολογούμε εμπειρικά την απόδοση της προσέγγισής μας σε σχέση με τις άλλες γνωστές τεχνικές, χρησιμοποιώντας και την ακρίβεια ταξινόμησης αλλά και το μέσο απόλυτο σφάλμα ως κριτήρια. Το Μέσο Απόλυτο Σφάλμα (ΜΑΣ) ορίζεται ως: $\frac{|p_1 - a_1| + \dots + |p_n - a_n|}{n}$, όπου p: οι προβλεθίσεις τιμές και a: οι πραγματικές

τιμές.

Πρέπει να σημειωθεί ότι τα χρησιμοποιούμενα σύνολα δεδομένων αντιπροσώπευαν προβλήματα παλινδρόμησης που μετατρέψαμε τις αριθμητικές τιμές στόχου σε διατεταγμένες κατηγορικές χωρίζοντας τις τιμές σε τρία ίσα διαστήματα (βλέπε Παράρτημα Α). Αυτή η μέθοδος επιλέχτηκε λόγω της έλλειψης συνόλων δεδομένων που περιλαμβάνουν πραγματικά βαθμωτά προβλήματα.

Από τα αποτελέσματα στον Πίνακα 69, φαίνεται ότι η προτεινόμενη τεχνική χρησιμοποιώντας σαν βασικό ταξινομητή ένα δέντρο αποφάσεων είναι ελαφρώς καλύτερη στην ακρίβεια ταξινόμησης από τις υπόλοιπες προσεγγίσεις που χρησιμοποιούν επίσης δέντρο αποφάσεων ως ταξινομητή. Χρησιμοποιήσαμε τον αλγόριθμο M5' (Frank και άλλοι, 1998) ως αλγόριθμο παλινδρόμησης που βασίζεται στα δέντρα αποφάσεων, τον αλγόριθμο C4.5 μόνο του, καθώς και τον αλγόριθμο C4.5 σε συνδυασμό με την δική μας μέθοδο (C4.5-COST) και σε συνδυασμό με τη μέθοδο των (Frank & Hall, 2001) (C4.5-ORD).

Η προτεινόμενη τεχνική κατορθώνει να ελαχιστοποιήσει τις αποστάσεις μεταξύ των πραγματικών και προβλεφθεισών κατηγοριών. Η μείωση του μέσου απόλυτου σφάλματος είναι περίπου 19% έναντι του απλού C4.5 και του C4.5-ORD, ενώ υπερβαίνει το 27% έναντι του M5'. Πρέπει να σημειωθεί ότι η τεχνική C4.5-ORD ξεπερνά τον απλό C4.5 μόνο στην ακρίβεια ταξινόμησης. Δεν κατορθώνει να ελαχιστοποιήσει την απόσταση μεταξύ της πραγματικής και της προβλεφθείσας κατηγορίας. Επιπλέον, ο αλγόριθμος M5' φαίνεται να δίνει τα χειρότερα μέσα αποτελέσματα σύμφωνα με τα πειράματά μας αν και σε μερικά σύνολα δεδομένων η απόδοσή του είναι πολύ καλύτερη από την απόδοση των υπόλοιπων αλγορίθμων.

Δεδομένα		C4.5-COST	C4.5	M5'	C4.5-ORD
auto93	ακρίβεια	79.71	80.83	81.24	78.31
	ΜΑΣ	0.13	0.16	0.15	0.17
autoHorse	ακρίβεια	96.78	96.78	94.74	96.93
	ΜΑΣ	0.02	0.03	0.06	0.03
autoMpg	ακρίβεια	82.24	82.09	82.36	82.79
	ΜΑΣ	0.12	0.17	0.16	0.17
autoPrice	ακρίβεια	89.25	89.25	89.5	89.68
	ΜΑΣ	0.07	0.08	0.11	0.08
basketball	ακρίβεια	71.29	70.64	71.46	67.9
	ΜΑΣ	0.19	0.26	0.26	0.27
bodyfat	ακρίβεια	98.37	98.37	99.12	98.37
	ΜΑΣ	0.01	0.02	0.06	0.02
cleveland	ακρίβεια	69.5	69.57	75.09	68.9
	ΜΑΣ	0.2	0.23	0.22	0.23
cloud	ακρίβεια	90.53	90.54	88.76	91.12
	ΜΑΣ	0.06	0.08	0.11	0.07
cpu	ακρίβεια	97.19	96.95	97.14	97.42
	ΜΑΣ	0.02	0.02	0.03	0.02
echoMonth	ακρίβεια	63.46	62.46	66.08	65.54
	ΜΑΣ	0.24	0.29	0.31	0.31
fishcatch	ακρίβεια	97.72	97.65	96.84	97.91
	ΜΑΣ	0.02	0.02	0.06	0.01
housing	ακρίβεια	80.34	78.22	82.33	79.07
	ΜΑΣ	0.15	0.17	0.17	0.16
hungarian	ακρίβεια	80.22	80.22	82.44	80.22
	ΜΑΣ	0.13	0.2	0.17	0.2
lowbwt	ακρίβεια	60.09	60.09	61.3	60.04
	ΜΑΣ	0.27	0.3	0.3	0.3
pbc	ακρίβεια	55.25	55.53	57.4	54.27
	ΜΑΣ	0.3	0.34	0.34	0.35
pharynx	ακρίβεια	70.52	69.4	43.97	73.59
	ΜΑΣ	0.2	0.26	0.39	0.25
pwLinear	ακρίβεια	78.05	76.55	78.3	78.5
	ΜΑΣ	0.16	0.18	0.23	0.17
sensory	ακρίβεια	64.24	64.24	64.24	64.05
	ΜΑΣ	0.24	0.35	0.33	0.35
strike	ακρίβεια	99.21	99.21	99.21	99.21
	ΜΑΣ	0.01	0.01	0.01	0.01
veteran	ακρίβεια	91.26	91.26	91.26	91.26
	ΜΑΣ	0.06	0.11	0.11	0.11
Μέση τιμή	ακρίβεια	80.76	80.49	80.14	80.75
	ΜΑΣ	0.13	0.16	0.18	0.16

Πίνακας 69. Σύγκριση της προτεινόμενης μεθόδου με άλλες που χρησιμοποιούν ως βασικό ταξινομητή δέντρα αποφάσεων

Από τα αποτελέσματα στον Πίνακα 70, φαίνεται ότι η προτεινόμενη τεχνική χρησιμοποιώντας σαν βασικό ταξινομητή κανόνες ταξινόμησης είναι ελαφρώς καλύτερη

στην ακρίβεια ταξινόμησης από τις υπόλοιπες προσεγγίσεις που χρησιμοποιούν επίσης κανόνες ταξινόμησης.

Δεδομένα		PART-COST	PART	M5rules	PART-ORD
auto93	ακρίβεια	81.31	80.23	79.9	78.2
	ΜΑΣ	0.13	0.14	0.15	0.15
autoHorse	ακρίβεια	96.78	96.78	93.9	96.93
	ΜΑΣ	0.02	0.03	0.06	0.03
autoMpg	ακρίβεια	81.99	80.03	82.14	80.76
	ΜΑΣ	0.13	0.15	0.16	0.15
autoPrice	ακρίβεια	89.38	89.38	90.5	90.44
	ΜΑΣ	0.07	0.08	0.09	0.07
basketball	ακρίβεια	67.96	66.6	70.72	60.39
	ΜΑΣ	0.22	0.26	0.26	0.28
bodyfat	ακρίβεια	98.37	98.13	98.41	98.37
	ΜΑΣ	0.01	0.02	0.04	0.02
cleveland	ακρίβεια	69.69	69.73	75.02	68.91
	ΜΑΣ	0.2	0.21	0.22	0.22
cloud	ακρίβεια	90.86	89.75	88.97	90.84
	ΜΑΣ	0.07	0.08	0.1	0.08
cpu	ακρίβεια	97.57	97.05	97.42	97.42
	ΜΑΣ	0.02	0.02	0.02	0.02
EchoMonth	ακρίβεια	63.69	62.77	65.69	63.38
	ΜΑΣ	0.24	0.28	0.31	0.29
fishcatch	ακρίβεια	97.79	97.65	97.08	97.91
	ΜΑΣ	0.02	0.02	0.04	0.01
Housing	ακρίβεια	81.27	78.74	82.07	79.68
	ΜΑΣ	0.14	0.15	0.16	0.15
hungarian	ακρίβεια	81.07	81.11	82.4	80.9
	ΜΑΣ	0.13	0.17	0.17	0.17
lowbwt	ακρίβεια	56.66	56.66	61.51	56.34
	ΜΑΣ	0.29	0.31	0.3	0.31
pbc	ακρίβεια	52.74	52.86	57.62	52.45
	ΜΑΣ	0.32	0.34	0.34	0.34
pharynx	ακρίβεια	70.56	70.61	43.38	71.45
	ΜΑΣ	0.2	0.25	0.38	0.25
pwLinear	ακρίβεια	79.55	76.55	76.9	77.8
	ΜΑΣ	0.16	0.16	0.23	0.16
sensory	ακρίβεια	60.99	57.92	64.38	60.41
	ΜΑΣ	0.28	0.33	0.33	0.32
strike	ακρίβεια	99.21	99.21	99.11	99.21
	ΜΑΣ	0.01	0.01	0.01	0.01
veteran	ακρίβεια	88.78	87.18	90.82	87.34
	ΜΑΣ	0.09	0.11	0.11	0.11
Μέση τιμή	ακρίβεια	80.31	79.45	79.90	79.46
	ΜΑΣ	0.14	0.16	0.17	0.16

Πίνακας 70. Σύγκριση της προτεινόμενης μεθόδου με άλλες που χρησιμοποιούν ως βασικό ταξινομητή αλγόριθμο παραγωγής κανόνων

Χρησιμοποιήσαμε τον αλγόριθμο M5rules ως αλγόριθμο παλινδρόμησης (Frank και άλλοι, 1998) που βασίζεται σε κανόνες, τον αλγόριθμο PART μόνο του, καθώς και τον αλγόριθμο PART σε συνδυασμό με την δική μας μέθοδο (PART-COST) και σε συνδυασμό με τη μέθοδο των (Frank & Hall, 2001) (PART-ORD).

Η μείωση του μέσου απόλυτου σφάλματος είναι περίπου 12% έναντι του απλού PART και του PART-ORD, ενώ υπερβαίνει το 17% έναντι του M5rules. Πρέπει να σημειωθεί ότι η τεχνική PART-ORD δεν κατορθώνει να ξεπεράσει τον απλό PART. Αντίθετα, ο αλγόριθμος M5rules φαίνεται να δίνει την καλύτερη ακρίβεια ταξινόμησης αλλά και το χειρότερο μέσο απόλυτο σφάλμα σύμφωνα με τα πειράματά μας.

Από τα αποτελέσματα στον Πίνακα 71, φαίνεται ότι η προτεινόμενη τεχνική χρησιμοποιώντας σαν βασικό ταξινομητή έναν σκληρό αλγόριθμο είναι ελαφρώς καλύτερη στην ακρίβεια ταξινόμησης από τις υπόλοιπες προσεγγίσεις που χρησιμοποιούν επίσης σκληρό αλγόριθμο. Χρησιμοποιήσαμε τον αλγόριθμο LWR (Atkeson και άλλοι, 1997b) ως σκληρό αλγόριθμο παλινδρόμησης, τον αλγόριθμο 3-NN μόνο του, καθώς και τον αλγόριθμο 3-NN σε συνδυασμό με την δική μας μέθοδο (3NN-COST) και σε συνδυασμό με τη μέθοδο των (Frank & Hall, 2001) (3NN-ORD).

Στον Πίνακα 71 φαίνεται ότι η προτεινόμενη δίνει παρόμοια αποτελέσματα με τον 3NN-ORD, αλλά λειτουργεί ταχύτερα. Επίσης, η μείωση του μέσου απόλυτου σφάλματος είναι περίπου 12% έναντι του απλού 3NN και του LWR. Πρέπει να αναφερθεί ότι το LWR φαίνεται να δίνει τα χειρότερα μέσα αποτελέσματα στα πειράματά μας αν και σε μερικά σύνολα δεδομένων η απόδοσή του είναι πολύ καλύτερη από την απόδοση των υπόλοιπων αλγορίθμων.

Δεδομένα		3NN-COST	3NN	LWR	3NN-ORD
auto93	ακρίβεια	78.46	78.36	84.01	78.46
	ΜΑΣ	0.14	0.15	0.12	0.14
autoHorse	ακρίβεια	88.33	88.33	93.56	88.33
	ΜΑΣ	0.08	0.09	0.05	0.08
autoMpg	ακρίβεια	82.41	82.41	81.44	82.41
	ΜΑΣ	0.12	0.14	0.14	0.12
autoPrice	ακρίβεια	86.93	86.23	88.94	86.23
	ΜΑΣ	0.09	0.1	0.09	0.09
basketball	ακρίβεια	60.36	59.52	60.66	60.36
	ΜΑΣ	0.26	0.29	0.29	0.26
bodyfat	ακρίβεια	86	86	79.72	86
	ΜΑΣ	0.09	0.12	0.15	0.09
cleveland	ακρίβεια	71.12	72.32	69.2	71.12
	ΜΑΣ	0.19	0.2	0.21	0.19
cloud	ακρίβεια	81.93	81.93	85.69	81.93
	ΜΑΣ	0.12	0.12	0.12	0.12
cpu	ακρίβεια	96.42	97.05	98.48	97.05
	ΜΑΣ	0.02	0.02	0.01	0.02
echoMonths	ακρίβεια	64.15	56.31	55	64.15
	ΜΑΣ	0.24	0.31	0.32	0.24
fishcatch	ακρίβεια	95.84	94.83	94.44	95.84
	ΜΑΣ	0.03	0.05	0.06	0.03
Housing	ακρίβεια	78.34	78.04	78.2	78.34
	ΜΑΣ	0.14	0.17	0.17	0.14
hungarian	ακρίβεια	82.33	82.33	78.84	82.33
	ΜΑΣ	0.12	0.14	0.15	0.12
lowbwt	ακρίβεια	56.13	55.81	58.67	56.13
	ΜΑΣ	0.29	0.31	0.3	0.29
pbc	ακρίβεια	51.36	50.78	49.53	51.36
	ΜΑΣ	0.32	0.35	0.35	0.32
pharynx	ακρίβεια	62.88	63.43	58.53	64.26
	ΜΑΣ	0.25	0.26	0.33	0.24
pwLinear	ακρίβεια	72.15	70.45	62.65	71.1
	ΜΑΣ	0.19	0.24	0.25	0.19
sensory	ακρίβεια	62.54	60.63	52.57	62.21
	ΜΑΣ	0.25	0.33	0.32	0.25
strike	ακρίβεια	99.21	99.21	98.77	99.21
	ΜΑΣ	0.01	0.01	0.01	0.01
veteran	ακρίβεια	88.08	88.08	87.05	88.08
	ΜΑΣ	0.08	0.12	0.11	0.08
Μέση τιμή	ακρίβεια	77.25	76.60	75.80	77.24
	ΜΑΣ	0.15	0.17	0.17	0.15

Πίνακας 71. Σύγκριση της προτεινόμενης μεθόδου με άλλες που χρησιμοποιούν ως βασικό ταξινομητή σκληρό αλγόριθμο

ΚΕΦΑΛΑΙΟ 7. Εφαρμογές των προτεινόμενων μεθόδων αλλά και των μεθόδων μηχανικής μάθησης γενικότερα σε πραγματικά προβλήματα

Σε ένα μέρος της έρευνας μας ασχοληθήκαμε και με την δυνατότητα χρησιμοποίησης μεθόδων μηχανικής μάθησης σε πραγματικά δεδομένα που συλλέξαμε εμείς οι ίδιοι. Επίσης, κατασκευάστηκε ένα ολοκληρωμένο πρωτότυπο εργαλείο το οποίο επιτρέπει στους χρήστες να εισάγουν δεδομένα για οποιοδήποτε πρόβλημα μάθησης. Στη συνέχεια, το εργαλείο μπορεί να κατασκευάσει τον ταξινομητή για το οποιοδήποτε πρόβλημα χρησιμοποιώντας ένα πλήθος αλγορίθμων μάθησης.

7.1 Εφαρμογές μεθόδων μηχανικής μάθησης για εκπαιδευτικούς σκοπούς

Καταρχήν, ασχοληθήκαμε και με την δυνατότητα χρησιμοποίησης των μεθόδων μηχανικής μάθησης για εκπαιδευτικούς σκοπούς.

7.1.1 Πρόβλεψη της εγκατάλειψης των σπουδών από το ελληνικό ανοιχτό πανεπιστήμιο

Τα Ανοικτά Πανεπιστήμια μπορούν να προσφέρουν εκπαίδευση πανεπιστημιακού επιπέδου με τη μέθοδο της εξ αποστάσεως εκπαίδευσης και σε πολίτες οι οποίοι για διάφορους λόγους, όπως είναι η αδυναμία της φυσικής τους παρουσίας σε συγκεκριμένο τόπο, δεν είχαν τη δυνατότητα να σπουδάσουν. Δίνουν, επίσης, την ευκαιρία για σπουδές και σε ευαίσθητες ομάδες του πληθυσμού, όπως ανέργους, άτομα με ειδικές ανάγκες, μητέρες, απόδημους, μετανάστες κλπ.

Από το Ελληνικό Ανοικτό Πανεπιστήμιο (ΕΑΠ) υιοθετήθηκε το αρθρωτό (modular) σύστημα, με βασική λειτουργική μονάδα τη Θεματική Ενότητα (ΘΕ), διάρκειας ενός ακαδημαϊκού έτους. Για την ολοκλήρωση των σπουδών σε μια ΘΕ, ο φοιτητής πρέπει να εκπονήσει 4 γραπτές εργασίες, η βαθμολογία των οποίων καθορίζει το 30% του τελικού βαθμού, και να προσέλθει σε τελική εξέταση, η βαθμολογία της οποίας καθορίζει το 70% του τελικού βαθμού της ΘΕ. Προαιρετικά, τέλος, μπορεί να συμμετέχει σε 4 Ομαδικές Συμβουλευτικές Συναντήσεις (ΟΣΣ).

Η εγκατάλειψη των σπουδών από τους σπουδαστές πριν την ολοκλήρωση τους εμφανίζεται αρκετά συχνά στα πανεπιστήμια που παρέχουν εκπαίδευση από απόσταση και το ποσοστό εγκατάλειψης είναι σίγουρα υψηλότερο από εκείνο των συμβατικών πανεπιστημίων (Xenos και άλλοι, 2002). Η κυριότερη αιτία διακοπής της φοίτησης στο ΕΑΠ είναι οι υποχρεώσεις των ενηλίκων και ειδικότερα (Xenos και άλλοι, 2002):

- τόσο για τους άντρες όσο και για τις γυναίκες ο φόρτος εργασίας και οι επαγγελματικές υποχρεώσεις,
- κυρίως για τις γυναίκες οι οικογενειακές υποχρεώσεις.

Μια άλλη αιτία διακοπής της φοίτησης φαίνεται ότι είναι η λανθασμένη εκτίμηση του διαθέσιμου χρόνου για μελέτη, σε συνάρτηση με αλλαγές στην εργασιακή και καθημερινή ζωή, που προκύπτουν κατά τη μετάβαση από τις σπουδές στην επαγγελματική και οικογενειακή ζωή, καθώς και η υποτίμηση του βαθμού δυσκολίας των σπουδών. Οι φοιτητές στην περίπτωση αυτή, δεν είχαν συνειδητοποιήσει στον απαιτούμενο βαθμό ότι οι σπουδές στο ΕΑΠ προϋποθέτουν αρκετή μελέτη και επομένως διαθέσιμο χρόνο.

Η δυνατότητα να προβλεφθεί η εγκατάλειψη ενός σπουδαστή θα μπορούσε συνεπώς να είναι χρήσιμη ώστε να δοθεί επιπλέον βοήθεια στον συγκεκριμένο σπουδαστή για να μην

εγκαταλείπει με την πρώτη δυσκολία. Αλλά επιπλέον θα είναι και στο ΕΑΠ για διάφορα πρακτικά προβλήματα όπως ο προγραμματισμός των αιθουσών για τις εξετάσεις κτλ.

Ο πρώτος στόχος της μελέτης μας ήταν να μελετήσει εάν η χρήση των τεχνικών μηχανικής μάθησης θα μπορούσε να είναι χρήσιμη στον εντοπισμό των φοιτητών με μεγάλη πιθανότητα εγκατάλειψης των σπουδών (dropout prone students). Πρέπει σε αυτό το σημείο να τονιστεί ότι η πρόβλεψη είναι ίδια με την ταξινόμηση εκτός από το ότι τα αρχεία είναι ταξινομημένα σύμφωνα με κάποια μελλοντική συμπεριφορά. Στην εργασία της πρόβλεψης, ο μόνος τρόπος να ελεγχθεί η ακρίβεια της ταξινόμησης είναι η αναμονή των τελικών αποτελεσμάτων. Οποιοσδήποτε από τις τεχνικές που χρησιμοποιούνται για την κατηγοριοποίηση μπορούν να προσαρμοστούν για να χρησιμοποιηθούν και στην πρόβλεψη.

Στον Πίνακα 72, παρουσιάζονται τα δεδομένα που χρησιμοποιήθηκαν για την πρόβλεψη στο συγκεκριμένο πρόβλημα.

Δημογραφικά στοιχεία	
Φύλο	άνδρας, γυναίκα
ηλικία	ηλικία<32, ηλικία≥32
Οικογενειακή κατάσταση	άγαμος, έγγαμος
Αριθμός Τέκνων	κανένα, ένα, δύο ή περισσότερα
Εργασία	όχι, ημιαπασχόληση, πλήρης απασχόληση
Γνώση υπολογιστών	όχι, ναι
Εργασία συσχετιζόμενη με υπολογιστές	όχι, απλός χρήστης, προχωρημένος χρήστης
Στοιχεία επίδοσης φοιτητών	
ΟΣΣ1	απών, παρών
Πρώτη γραπτή εργασία	βαθμός<3, 3≤ βαθμός ≤6, βαθμός >6
ΟΣΣ2	απών, παρών
Δεύτερη γραπτή εργασία	βαθμός<3, 3≤ βαθμός ≤6, βαθμός >6
ΟΣΣ3	απών, παρών
Τρίτη γραπτή εργασία	βαθμός<3, 3≤ βαθμός ≤6, βαθμός >6
ΟΣΣ4	απών, παρών
Τέταρτη γραπτή εργασία	βαθμός<3, 3≤ βαθμός ≤6, βαθμός >6

Πίνακας 72. Περιγραφή των δεδομένων

Τα πειράματα πραγματοποιήθηκαν σε δύο διακριτές φάσεις. Κατά τη διάρκεια της πρώτης φάσης (φάση εκπαίδευσης) οι αλγόριθμοι εκπαιδεύθηκαν χρησιμοποιώντας τα στοιχεία που συλλέχθηκαν από το ακαδημαϊκό έτος 2000-1. Επειδή είναι σημαντικό για τους εκπαιδευτές να αναγνωρίσουν τους επιρρεπείς για εγκατάλειψη φοιτητές πριν από τα μέσα της περιόδου, οι μεταβλητές που μπορούν να συλλεχθούν πριν από τα μέσα της περιόδου χρησιμοποιήθηκαν.

Έτσι, η φάση εκπαίδευσης διαιρέθηκε σε 5 διαδοχικά βήματα. Το 1ο βήμα (Δημογραφικά) περιέλαβε τα δημογραφικά στοιχεία και την μεταβλητή στόχου (εγκατάλειψη ή όχι). Το 2ο βήμα (ΟΣΣ1) περιέλαβε τα δημογραφικά στοιχεία μαζί με τα στοιχεία από την πρώτη συνάντηση (ΟΣΣ) και την μεταβλητή στόχου. Το 3ο βήμα (Πρώτη γραπτή εργασία) περιέλαβε τα στοιχεία που χρησιμοποιήθηκαν για το 2ο βήμα καθώς και τα στοιχεία από την πρώτη γραπτή εργασία. Το 4ο βήμα (ΟΣΣ 2) περιέλαβε τα στοιχεία που χρησιμοποιήθηκαν για το 3ο βήμα καθώς και τα στοιχεία από την δεύτερη συνάντηση (ΟΣΣ) και τέλος το 5ο βήμα (Δεύτερη γραπτή εργασία) περιέλαβε όλες τις διαθέσιμες μεταβλητές μέχρι και την δεύτερη γραπτή εργασία.

Στη συνέχεια, οι εκπαιδευτές συνέλεξαν δέκα ομάδες στοιχείων από το νέο ακαδημαϊκό έτος (2001-2). Καθεμία αυτών των 10 ομάδων χρησιμοποιήθηκε για να μετρήσει την ακρίβεια πρόβλεψης μέσα σε αυτές τις ομάδες (φάση ελέγχου). Η φάση ελέγχου πραγματοποιήθηκε επίσης σε 5 βήματα σε αντιστοιχία με τα 5 βήματα της φάσης εκπαίδευσης. Κατά τη διάρκεια της φάσης ελέγχου κάθε βήμα επαναλήφθηκε 10 φορές (μία φορά για το σύνολο δεδομένων κάθε εκπαιδευτή).

Στη συνέχεια της μελέτης μας, προσπαθήσαμε να προσδιορίσουμε τον πιο κατάλληλο αλγόριθμο μάθησης για την πρόβλεψη της εγκατάλειψης των σπουδών. Για τα πειράματα χρησιμοποιήσαμε έναν αλγόριθμο ανά κατηγορία μεθόδων μηχανικής μάθησης:

- Τον NB ως εκπρόσωπο των Μπεϋζιανών αλγορίθμων
- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των σκνηρών αλγορίθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMO ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης

Στον Πίνακα 73, παρουσιάζεται η μέση ακρίβεια των αλγορίθμων σε κάθε εξεταστικό βήμα.

	<i>NB</i>	<i>C4.5</i>	<i>BP</i>	<i>SMO</i>	<i>3NN</i>
Δημογραφικά	61.49	63.02	63.06	58.09	59.44
ΟΣΣ1	61.66	62.80	62.65	58.40	59.47
Πρώτη γραπτή εργασία	71.55	70.19	71.04	72.33	62.87
ΟΣΣ2	72.93	70.39	72.94	72.33	65.82
Δεύτερη γραπτή εργασία	82.89	83.75	83.89	82.96	78.17

Πίνακας 73. Σύγκριση των αλγορίθμων μάθησης για την πρόβλεψη της εγκατάλειψης των σπουδών

Βάσει των πειραμάτων μας, αποδείχθηκε ότι οι αλγόριθμοι μάθησης θα μπορούσαν να επιτρέψουν στους καθηγητές, ακόμη και από την αρχή ενός ακαδημαϊκού έτους χρησιμοποιώντας μόνο τα δημογραφικά στοιχεία των σπουδαστών (όπως η ηλικία, το φύλο κ.λπ.), να εντοπίσουν με ικανοποιητική ακρίβεια τους επιρρεπείς για εγκατάλειψη φοιτητές. Η ακρίβεια φθάνει το 63% στις αρχικές προβλέψεις βασισμένες μόνο στα δημογραφικά στοιχεία και υπερβαίνει 83% πριν τα μέσα της ακαδημαϊκής περιόδου. Μεταξύ άλλων σημαντικών συμπερασμάτων βρέθηκε ότι ο αφελής αλγόριθμος Μπέιζ είναι ο πιο κατάλληλος αφού έχει περισσότερο από ικανοποιητική ακρίβεια και είναι ένας από τους απλούστερους αλγόριθμους στην υλοποίηση.

Ένα πρωτότυπο διαδικτυακό εργαλείο υποστήριξης που μπορεί αυτόματα να αναγνωρίσει τους σπουδαστές με υψηλή πιθανότητα εγκατάλειψης, κατασκευάστηκε εφαρμόζοντας αυτόν τον αλγόριθμο (βλέπε Εικόνα 72).

PREDICTION DROP-OUT TOOL

Sex: female	Job associated with computer: junior-user
Age: <32	FTOF-1: no
Marital status: married	WRI-1: 5=<mark<8
Children: >1	FTOF-2: yes
Occupation type: part-time	WRI-2: 5=<mark<8
Computer Literacy: yes	

Predict Clear

Probability of drop-out = 16.25 %

Εικόνα 72. Το πρωτότυπο διαδικτυακό εργαλείο πρόβλεψης της εγκατάλειψης των σπουδών στο ΕΑΠ

7.1.2 Πρόβλεψη απόδοσης φοιτητών

Αλλά και μετά την αναγνώριση των σπουδαστών με υψηλή πιθανότητα εγκατάλειψης, η δυνατότητα της πρόβλεψης της απόδοσης ενός σπουδαστή θα μπορούσε να είναι χρήσιμη για επιπλέον προσωποποίηση (personalization) της εκπαιδευτικής διαδικασίας. Τα βασικά δημογραφικά χαρακτηριστικά των σπουδαστών και οι βαθμοί τους στις γραπτές εργασίες μπορούν να αποτελέσουν το σύνολο δεδομένων για έναν εποπτευμένο αλγόριθμο μάθησης μηχανών και σε αυτή την περίπτωση.

Ο αλγόριθμος μάθησης θα μπορούσε έπειτα να προβλέπει την απόδοση των νέων σπουδαστών δίνοντας έτσι ένα χρήσιμο εργαλείο για την αναγνώριση των αδύναμων φοιτητών (που δεν εγκαταλείπουν αλλά είναι «κάτω από το 5»). Σκοπός της εργασίας [1] ήταν να συγκριθούν μερικοί από τους αλγορίθμους μάθησης και σε αυτό το πρόβλημα.

Πραγματοποιήθηκαν δύο πειράματα. Κατά τη διάρκεια του πρώτου πειράματος οι αλγόριθμοι εκπαιδεύθηκαν χρησιμοποιώντας τα στοιχεία που συλλέχθηκαν από ολόκληρο το ακαδημαϊκό έτος 2000-1. Στο δεύτερο πείραμα αντί να χρησιμοποιήσουμε το σύνολο των δεδομένων από το ακαδημαϊκό έτος, χρησιμοποιήσαμε τα σύνολα των δεδομένων που μπορούσε να μαζέψει ένας καθηγητής από κάθε Θ.Ε. (περίπου 30 παραδείγματα). Το δεύτερο πείραμα έγινε για να δούμε αν και κατά πόσο ένας καθηγητής χωρίς συνεργασία με άλλους (για συλλογή επιπλέον δεδομένων) θα μπορούσε να χρησιμοποιήσει ένα αλγόριθμο μάθησης επιτυχώς.

Πρέπει να σημειωθεί ότι η φάση της εκπαίδευσης διαιρέθηκε σε 9 διαδοχικά βήματα (και για τα δύο πειράματα). Το 1ο βήμα (Δημογραφικά) περιελάμβανε τα δημογραφικά στοιχεία και την μεταβλητή στόχο (προβηβάσιμο βαθμό ή όχι). Το 2ο βήμα (ΟΣΣ1) περιελάμβανε τα δημογραφικά στοιχεία μαζί με τα στοιχεία από την πρώτη συνάντηση (ΟΣΣ) και την μεταβλητή στόχου. Το 3ο βήμα (Πρώτη γραπτή εργασία) περιελάμβανε τα στοιχεία που χρησιμοποιήθηκαν για το 2ο βήμα καθώς και τα στοιχεία από την πρώτη γραπτή εργασία. Το 4ο βήμα (ΟΣΣ 2) περιελάμβανε τα στοιχεία που χρησιμοποιήθηκαν για το 3ο βήμα καθώς και τα στοιχεία από την δεύτερη συνάντηση (ΟΣΣ) και ούτω καθεξής μέχρι το 9ο βήμα το οποίο περιελάμβανε όλες τις διαθέσιμες μεταβλητές.

Στη συνέχεια, οι εκπαιδευτές συνέλεξαν δέκα ομάδες στοιχείων από το ακαδημαϊκό έτος (2001-2). Καθεμία αυτών των 10 ομάδων χρησιμοποιήθηκε για να μετρήσει την ακρίβεια πρόβλεψης μέσα σε αυτές τις ομάδες. Η φάση ελέγχου πραγματοποιήθηκε επίσης σε 9

βήματα σε αντιστοιχία με τα 9 βήματα της φάσης εκπαίδευσης (και για τα δύο πειράματα). Κατά τη διάρκεια της φάσης ελέγχου κάθε βήμα επαναλήφθηκε 10 φορές (μία φορά για το σύνολο δεδομένων κάθε εκπαιδευτή).

Για τα πειράματα χρησιμοποιήσαμε έναν αλγόριθμο ανά κατηγορία μεθόδων μηχανικής μάθησης:

- Τον NB ως εκπρόσωπο των Μπεϋζιανών αλγορίθμων
- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των οκνηρών αλγορίθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMO ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης

Ο Πίνακας 74 παρουσιάζει την μέση ακρίβεια των αλγορίθμων σε κάθε βήμα για το πρώτο πείραμα.

	NB	C4.5	BP	SMO	3-NN
Δημογραφικά	62.95%	61.65%	61.85%	64.47%	58.84%
ΟΣΣ1	62.72%	61.56%	61.14%	64.47%	59.12%
1 ^η γραπτή εργασία	66.23%	65.35%	63.62%	63.11%	60.21%
ΟΣΣ2	69.78%	62.93%	68.18%	68.58%	62.41%
2 ^η γραπτή εργασία	75.36%	74.16%	75.55%	75.99%	68.45%
ΟΣΣ3	75.38%	72.44%	76.38%	76.22%	68.78%
3 ^η γραπτή εργασία	78.58%	79.22%	80.11%	77.71%	72.62%
ΟΣΣ4	79.20%	74.84%	78.02%	78.37%	75.14%
4 ^η γραπτή εργασία	82.14%	77.80%	82.14%	80.68%	76.77%

Πίνακας 74. Ακρίβεια των αλγορίθμων κατά το πρώτο πείραμα

Μεταξύ άλλων σημαντικών συμπερασμάτων βρέθηκε ότι και σε αυτή την περίπτωση ο αφελής αλγόριθμος Μπέιζ είναι ο πιο κατάλληλος. Έχει περισσότερο από ικανοποιητική ακρίβεια και είναι ένας από τους απλούστερους αλγόριθμους.

Στο δεύτερο πείραμα, όπως προείπαμε, χρησιμοποιήσαμε τα σύνολα των δεδομένων που μπορούσε να μαζέψει ένας καθηγητής από κάθε Θ.Ε. (περίπου 30 παραδείγματα) ως δεδομένα εκπαίδευσης. Ο Πίνακας 75 παρουσιάζει την μέση ακρίβεια των αλγορίθμων σε κάθε βήμα για το δεύτερο πείραμα. Και σε αυτό το πείραμα ο αφελής αλγόριθμος Μπέιζ παρήγαγε τα καλύτερα αποτελέσματα.

	NB	C4.5	BP	SMO	3-NN
Δημογραφικά	56.29%	57.73%	53.37%	54.97%	55.67%
ΟΣΣ1	56.47%	57.42%	50.71%	55.28%	55.95%
1 ^η γραπτή εργασία	57.58%	57.58%	51.56%	58.87%	55.77%
ΟΣΣ2	61.20%	58.99%	50.46%	60.25%	59.33%
2 ^η γραπτή εργασία	66.75%	62.21%	59.69%	65.08%	64.08%
ΟΣΣ3	69.60%	62.76%	62.94%	67.24%	67.39%
3 ^η γραπτή εργασία	74.69%	68.99%	71.50%	73.53%	71.75%
ΟΣΣ4	76.29%	71.20%	73.87%	73.65%	73.31%
4 ^η γραπτή εργασία	79.51%	73.99%	74.88%	76.96%	76.60%

Πίνακας 75. Ακρίβεια των αλγορίθμων κατά το δεύτερο πείραμα

7.1.3 Εύρεση των μεταβλητών που επηρεάζουν την πρόβλεψη επίδοσης

Ένα άλλο πεδίο μελέτης της μελέτης μας ήταν να προσδιοριστούν τα χαρακτηριστικά των σπουδαστών που, ενδεχομένως, επηρεάζουν περισσότερο την επαγωγή των αλγορίθμων μάθησης στην πρόβλεψη της επίδοσης των σπουδαστών.

Στον Πίνακα 76 ταξινομούνται οι μεταβλητές ως προς την επιρροή τους στην τελική πρόβλεψη. Σαν μέτρο χρησιμοποιήθηκε το μέσο κέρδος πληροφορίας (Average Information Gain), που έχει αναλυτικά περιγραφεί σε προηγούμενη ενότητα. Βρέθηκαν μερικά προφανή και μερικά λιγότερο προφανή χαρακτηριστικά των σπουδαστών να έχουν ισχυρή συσχέτιση με την απόδοσή τους. Όπως ήταν αναμενόμενο πιο χρήσιμες μεταβλητές είναι αυτές που αναφέρονται στις γραπτές εργασίες. Λιγότερο ρόλο διαδραματίζει η ηλικία, το φύλο και τα παιδιά.

Μεταβλητές	Μέσο κέρδος πληροφορίας (Average Information Gain)
3 ^η γραπτή εργασία	0.385 ± 0.018
4 ^η γραπτή εργασία	0.382 ± 0.014
2 ^η γραπτή εργασία	0.261 ± 0.011
ΟΣΣ4	0.217 ± 0.016
ΟΣΣ3	0.158 ± 0.010
1 ^η γραπτή εργασία	0.138 ± 0.010
ΟΣΣ2	0.082 ± 0.009
Εργασία συσχετιζόμενη με υπολογιστές	0.061 ± 0.008
Γνώση υπολογιστών	0.030 ± 0.004
Οικογενειακή κατάσταση	0.008 ± 0.002
Αριθμός Τέκνων	0.007 ± 0.003
Φύλο	0.005 ± 0.003
Εργασία	0.002 ± 0.002
ΟΣΣ1	0.001 ± 0.001
Ηλικία	0.000 ± 0.001

Πίνακας 76. Ταξινόμηση των μεταβλητών βάσει του μέσου κέρδους πληροφορίας για την πρόβλεψη της επίδοσης των σπουδαστών

7.1.4 Εργαλείο υποστήριξης αποφάσεων για τους καθηγητές του ανοιχτού πανεπιστημίου

Στα πλαίσια της έρευνας μας αναπτύχθηκε λοιπόν ένα εργαλείο το οποίο είναι σε θέση να αυτοματοποιεί τις προβλέψεις που περιγράφηκαν στις προηγούμενες ενότητες και σχετίζονται με τους σπουδαστές του ΕΑΠ. Η γενική αρχιτεκτονική του εργαλείου παρουσιάζεται στην Εικόνα 73 και αποτελείται από έξι συστατικά (components):

- Το συστατικό μηχανικής μάθησης, το οποίο περιλαμβάνει τους πιο γνωστούς αλγορίθμους μάθησης. Είναι επίσης εξοπλισμένο με ένα είδος έξυπνου συμβούλου, που καθοδηγεί τον εκπαιδευτικό στην ανεύρεση του καταλληλότερου αλγορίθμου μάθησης για το πρόβλημά του
- Το συστατικό στατιστικών στοιχείων, το οποίο είναι σε θέση να παρουσιάσει χρήσιμες πληροφορίες για το σύνολο δεδομένων, όπως η απουσία τιμών σε κάποια χαρακτηριστικά, η συχνότητα κάθε τιμής κλπ. Μπορεί επίσης να παρουσιάσει τις μεταβλητές όσον αφορά προς την επιρροή τους στην τελική πρόβλεψη σε αύξουσα σειρά
- Το συστατικό επιλογής χαρακτηριστικών, το οποίο περιλαμβάνει τους πιο γνωστούς αλγορίθμους περιτυλίγματος για την επιλογή των χαρακτηριστικών. Είναι επίσης

εξοπλισμένο με ένα είδος συμβούλου, ο οποίος καθοδηγεί τον εκπαιδευτή στην επιλογή του καταλληλότερου αλγόριθμου επιλογής χαρακτηριστικών για τον επιλεγμένο αλγόριθμο μάθησης

- Το συστατικό εξισορρόπησης των δεδομένων. Ένα σύνολο δεδομένων μπορεί να είναι ανομοιογενές, π.χ. οι περισσότεροι σπουδαστές συνεχίζουν τις σπουδές τους και αρκετά λιγότεροι εγκαταλείπουν αλλά είναι σημαντικότερο να προβλέψουμε σωστά ότι ένας εκπαιδευόμενος είναι κοντά στην εγκατάλειψη. Το εργαλείο λοιπόν μπορεί να χειριστεί το πρόβλημα δυσαναλογίας με την κράτηση της σχέσης μεταξύ των αρνητικών και θετικών παραδειγμάτων σε βέλτιστη αναλογία.
- Το συστατικό κανόνων συσχέτισης, το οποίο αναλαμβάνει να ενημερώσει τους εκπαιδευτικούς για τους χρήσιμους συσχετισμούς μεταξύ των μεταβλητών που χαρακτηρίζουν τους εκπαιδευόμενους (αναλυτικότερη περιγραφή τι είναι η κανόνες συσχέτισης δίνεται στην ενότητα 7.1.4.3).
- Τον εξαγωγέα αλληλεπιδραστικών ιστοσελίδων (HTML), ένα συστατικό που επιτρέπει στον εκπαιδευόμενο να εξαγάγει μια δυναμική ιστοσελίδα (παρόμοια με αυτή της ενότητας 7.1.1) η οποία μπορεί να χρησιμοποιηθεί είτε από τους καθηγητές είτε από τους ίδιους τους σπουδαστές για να παρακολουθήσουν την πρόοδό τους. Η εξαγόμενη ιστοσελίδα χρησιμοποιεί Javascript προκειμένου να υποστηρίξει την αλληλεπίδραση.

Εικόνα 73. Αρχιτεκτονική του εργαλείου

Στις επόμενες ενότητες περιγράφουμε αναλυτικότερα το κάθε συστατικό.

7.1.4.1 Το συστατικό μηχανικής μάθησης

Το συστατικό μηχανικής μάθησης περιλαμβάνει ένα αντιπροσωπευτικό αλγόριθμο ανά κατηγορία μεθόδων μάθησης:

- Τον NB ως εκπρόσωπο των Μπεϋζιανών αλγορίθμων
- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των σκληρών αλγορίθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMO ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης
- Τον RIPPER ως εκπρόσωπο των κανόνων ταξινόμησης

Βάσει των προηγούμενων πειραμάτων ο αλγόριθμος NB είχε την καλύτερη ακρίβεια. Αλλά επειδή η αξιολόγηση των αλγορίθμων στηρίζονταν μόνο σε ένα ακαδημαϊκό έτος, κρίθηκε σκόπιμο το εργαλείο είναι σε θέση να βρίσκει τον πιο αποτελεσματικό αλγόριθμο μάθησης (Auto Model Selection) για τυχαία δεδομένα εισόδου (βλέπε Εικόνα 74) .

Η μεθοδολογία που χρησιμοποιήθηκε είναι η ίδια με αυτή της ενότητας 4.2.1 που αποτελείται από τέσσερα βήματα:

- Το νέο σύνολο δεδομένων διαιρείται τυχαία σε τρία ίσα μέρη
- Τα δύο τρίτα των δεδομένων ν χρησιμοποιούνται για την εκπαίδευση των αλγορίθμων και το υπόλοιπο ένα τρίτο είναι το σύνολο ελέγχου. Η ίδια διαδικασία επαναλαμβάνεται 3 φορές.
- Από τα αποτελέσματα των τριών πειραμάτων υπολογίζουμε το μέσο όρο της ακρίβειας κάθε αλγορίθμου
- Ο αλγόριθμος με την μεγαλύτερη κατά μέσο όρο ακρίβεια επιλέγεται για να χτίσει το τελικό μοντέλο χρησιμοποιώντας το σύνολο των δεδομένων εκπαίδευσης

Εικόνα 74. Επιλέγοντας τον αλγόριθμο μάθησης

Το εργαλείο μας δίνει επίσης την δυνατότητα να δούμε το παραγόμενο μοντέλο. Στην Εικόνα 75 βλέπουμε ένα παραχθέν δέντρο απόφασης.

Εικόνα 75. Το παραγόμενο μοντέλο

7.1.4.2 Συστατικό εντοπισμού των πιο χρήσιμων μεταβλητών

Σε μερικές περιπτώσεις ο διδακτικός στόχος μπορεί να περιορίζεται στον εντοπισμό των παραγόντων που επιδρούν περισσότερο στην εξέλιξη του φαινομένου. Στην Εικόνα 76 βλέπουμε πως το εργαλείο παρουσιάζει την ταξινόμηση των μεταβλητών βάσει της επιρροής τους στην μεταβλητή στόχο (χρησιμοποιώντας το φίλτρο ποσοστού κέρδους πληροφορίας (Gain Ratio), που έχει αναλυτικά περιγραφεί σε προηγούμενες ενότητες..

Εικόνα 76. Η παρουσίαση των μεταβλητών ως προς την επιρροή τους στην μεταβλητή στόχο

Επίσης το συστατικό επιλογής χαρακτηριστικών περιλαμβάνει τους πιο γνωστούς αλγορίθμους περιτυλίγματος (που περιγράφηκαν στην ενότητα 5.1) για την επιλογή των χαρακτηριστικών. Ο χρήστης μπορεί να επιλέξει μόνος του ποιον αλγόριθμο θέλει να χρησιμοποιήσει αλλά το εργαλείο είναι επίσης εξοπλισμένο με ένα είδος συμβούλου, ο οποίος καθοδηγεί τον χρήστη/εκπαιδευτή στην επιλογή του καταλληλότερου αλγόριθμου επιλογής χαρακτηριστικών για τον επιλεγμένο αλγόριθμο μάθησης και τα συγκεκριμένα δεδομένα εισόδου. Η μεθοδολογία εντοπισμού της βέλτιστης τεχνικής επιλογής μεταβλητών αποτελείται από τα ίδια τέσσερα βήματα που περιγράφηκαν στην ενότητα 7.1.4.1.

7.1.4.3 Το συστατικό εξαγωγής κανόνων συσχέτισης

Η ανακάλυψη κανόνων συσχέτισης (association rules discovery) είναι άλλη μια δημοφιλής τεχνική εξόρυξης γνώσης (Agrawal & Srikant, 1994). Αν, για λόγους ευκολίας υποθέσουμε ότι όλες οι μεταβλητές (στήλες ενός πίνακα δεδομένων) είναι δυαδικές, ένας κανόνας συσχέτισης έχει τη μορφή:

AN A =1 KAI B =1 TOTE C =1

ή συμβολικά

$(A = 1, B = 1) \rightarrow C = 1$

όπου A, B και C είναι δυαδικές μεταβλητές.

Η πιθανότητα $p_s(A=1, B=1, C=1)$ ονομάζεται στήριξη (support) ή διάδοση (prevalence) του συνδέσμου και εκφράζει την συχνότητα με την οποία ένας συγκεκριμένος σύνδεσμος εμφανίζεται στη βάση δεδομένων.

Η δεσμευμένη πιθανότητα p_a ονομάζεται, στην ορολογία της εξόρυξης γνώσης, ακρίβεια (accuracy) ή εμπιστοσύνη (confidence) του συνδέσμου και εκφράζει τη σχετική συχνότητα εμφάνισης των στοιχείων και των συνδυασμών τους και ορίζεται ως εξής:

ακρίβεια $[(A = 1, B = 1) \rightarrow C = 1] = p(A = 1, B = 1, C = 1) / p(A = 1, B = 1)$

Τι συμβολίζει όμως η εμπιστοσύνη πραγματικά; Λέγοντας ότι ο κανόνας “αν B και C συνεπάγεται A” έχει εμπιστοσύνη 0,33 εννοούμε το ότι όταν εμφανίζονται το B και το C, υπάρχει πιθανότητα 33% το A να εμφανίζεται επίσης. Δηλαδή μια φορά στις 3 το A εμφανίζεται με το B και το C, και τις άλλες 2 όχι.

Τυπικά, ο στόχος της ανακάλυψης κανόνων συσχέτισης είναι να εντοπισθούν όλοι οι σύνδεσμοι που έχουν ακρίβεια μεγαλύτερη από ένα κατώφλι p_a και στήριξη μεγαλύτερη από p_s . Σε ότι αφορά την αποδοτικότητα της διαδικασίας αναζήτησης των κανόνων συσχέτισης, διαπιστώνεται εύκολα ότι είναι κρίσιμο ζήτημα αφού ο αριθμός των δυνατών κανόνων, αν περιοριστούμε σε δυαδικές μεταβλητές, είναι της τάξης $O(m^{2^m-1})$.

Ας σημειωθεί ότι αν $p(A=1) \leq p_s$ ή $p(B=1) \leq p_s$ τότε είναι φανερό ότι θα είναι και $p(A=1, B=1) \leq p_s$. Αξιοποιώντας αυτή την παρατήρηση, μπορεί αρχικά να γίνει αναζήτηση των μεμονωμένων στοιχείων (π.χ. $A=1$) κατά στήλη που εμφανίζουν συχνότητα (στήριξη) μεγαλύτερη του κατωφλίου p_s . Αυτό επιτυγχάνεται με μια γραμμική σάρωση όλου του πίνακα. Ένα στοιχείο (ή ένα σύνολο στοιχείων) που η συχνότητα εμφάνισής του στο πίνακα δεδομένων ξεπερνά το κατώφλι p_s ονομάζεται «συχνό». Μπορούμε λοιπόν να θεωρήσουμε όλα τα δυνατά ζευγάρια των συχνών στοιχείων τάξεως 1 ως υποψήφια συχνά στοιχεία τάξεως 2. Η διαπίστωση του ποια από αυτά τα σύνολα στοιχείων τάξεως 2 είναι πράγματι συχνά γίνεται με νέα σάρωση του πίνακα δεδομένων.

Στο εργαλείο που κατασκευάσαμε χρησιμοποιήθηκε ο πιο γνωστός αλγόριθμος παραγωγής κανόνων συσχέτισης: A-priori. Ο ψευδοκώδικας του αλγορίθμου δίνεται στην Εικόνα 77.

1. Όλα στοιχεία (σύνολα τάξεως 1, δηλαδή μονοσύνολα) είναι αρχικά υποψήφια για χαρακτηρισμό ως συχνά.
2. Γίνεται μια σάρωση του πίνακα δεδομένων για να βρεθούν τα στοιχεία που είναι πράγματι συχνά, δηλαδή εκείνα των οποίων η συχνότητα εμφάνισης (στήριξη) ξεπερνά το κατώφλι p_s .
3. Όλα τα δυνατά ζευγάρια των συχνών στοιχείων τάξης 1 είναι υποψήφια για χαρακτηρισμό ως συχνά στοιχεία τάξης 2. Γίνεται σάρωση (2η) του πίνακα δεδομένων για να βρεθούν τα στοιχεία τάξης 2 που είναι πράγματι συχνά.
4. Υποψήφιο συχνό στοιχείο τάξης 3 είναι το σύνολο που προέρχεται από συνδυασμό των συχνών στοιχείων τάξης 2 κτλ
5. Στο τελευταίο στάδιο, γίνεται έλεγχος της ακρίβειας των συνδέσμων στις λίστες των συχνών στοιχείων με βάση το κατώφλι ακρίβειας p_a για να εξαχθούν οι σύνδεσμοι υπό μορφή κανόνων.

Εικόνα 77. Ο αλγόριθμος A-priori σε ψευδογλώσσα

Στην Εικόνα 78 παρουσιάζεται ένα σύνολο κανόνων συσχέτισης που προκύπτουν από τα δεδομένα του ΕΑΠ.

Εικόνα 78. Παραγόμενοι κανόνες συσχέτισης

Σχετικά με την πολυπλοκότητα του αλγόριθμου A-priori, είναι φανερό ότι στην χειρότερη περίπτωση που όλα τα δυνατά σύνολα αποδεικνύονται συχνά, ο αλγόριθμος είναι εκθετικού χρόνου. Όμως στην πράξη, επειδή οι πίνακες δεδομένων στους οποίους κυρίως εφαρμόζεται ο αλγόριθμος A-priori χαρακτηρίζονται από σποραδικότητα, η μεγαλύτερη τάξη συχνών στοιχείων (πληθάριθμος των συνόλων στη λίστα τελευταίας τάξης) είναι πολύ μικρή σε σχέση με το m (πλήθος μεταβλητών). Αυτό ενισχύεται επιπλέον όταν το κατώφλι στήριξης είναι αρκετά μεγάλο.

7.1.5 Πρόβλεψη βαθμού σπουδαστή

Όπως έχει πάλι αναφερθεί, η κατηγοριοποίηση αναφέρεται στην εκμάθηση μιας συνάρτησης από το σύστημα που απεικονίζει ένα στιγμιότυπο σε μία από κάποιες κλάσεις που έχουν ορισθεί εκ των προτέρων. Από την άλλη μεριά η παλινδρόμηση (regression) αναφέρεται στην εκπαίδευση μίας συνάρτησης η οποία αντιστοιχεί τα δεδομένα σε μία μεταβλητή πρόβλεψης (prediction variable) πραγματικής τιμής. Εκτός από τις κλασσικούς τύπους της γραμμικής παλινδρόμησης, τα δένδρα αποφάσεων και τα νευρωνικά δίκτυα αποτελούν διαδεδομένες τεχνικές προβλεπτικής όρυξης δεδομένων (Witten & Frank, 2000).

Ένας άλλος στόχος της μελέτης μας ήταν να βρεθεί ποιος από τους αλγόριθμους παλινδρόμησης είναι ο βέλτιστος για να χρησιμοποιηθεί για την πρόβλεψη των βαθμών των φοιτητών στο Ελληνικό Ανοικτό Πανεπιστήμιο.

Όπως προείπαμε, οι αλγόριθμοι μάθησης θα μπορούσαν να επιτρέψουν στους εκπαιδευτές, ακόμη και από την αρχή ενός ακαδημαϊκού έτους χρησιμοποιώντας μόνο τα δημογραφικά στοιχεία των σπουδαστών (όπως η ηλικία, το φύλο κ.λ.π.), να εντοπίσουν με ικανοποιητική ακρίβεια τους επιρρεπείς για εγκατάλειψη φοιτητές. Η ακρίβεια φθάνει το 63% στις αρχικές προβλέψεις βασισμένες μόνο στα δημογραφικά στοιχεία και υπερβαίνει 83% πριν τα μέσα της ακαδημαϊκής περιόδου. Μετά από τα μέσα της περιόδου, μπορούμε λοιπόν να χρησιμοποιήσουμε τις υπάρχουσες τεχνικές παλινδρόμησης προκειμένου να προβλέψουμε τους βαθμούς των σπουδαστών για να προσωποποιηθεί ακόμα περισσότερο η εκπαιδευτική διαδικασία.

Και σε αυτή την περίπτωση, τα πειράματα πραγματοποιήθηκαν σε δύο διακριτές φάσεις. Κατά τη διάρκεια της πρώτης φάσης (φάση εκπαίδευσης) οι αλγόριθμοι εκπαιδεύθηκαν χρησιμοποιώντας τα στοιχεία από το ακαδημαϊκό έτος 2000-1. Η φάση εκπαίδευσης διαιρέθηκε με τη σειρά της σε 5 διαδοχικά βήματα. Το 1ο βήμα περιελάμβανε τα δημογραφικά στοιχεία, τα στοιχεία από τις δύο πρώτες συμβουλευτικές συναντήσεις, τις δυο πρώτες γραπτές εργασίες και την μεταβλητή στόχο (τελικό βαθμό). Το 2ο βήμα περιελάμβανε πρόσθετα τα στοιχεία για την τρίτη συνάντηση. Το 3ο βήμα περιελάμβανε επιπρόσθετα τα στοιχεία από την τρίτη γραπτή εργασία. Το 4ο βήμα περιελάμβανε επιπρόσθετα τα στοιχεία από την τέταρτη συνάντηση και τέλος το 5ο βήμα που περιελάμβανε όλες τις διαθέσιμες μεταβλητές.

Στη συνέχεια, δέκα ομάδες στοιχείων από το επόμενο ακαδημαϊκό έτος (2001-2) συλλέχθηκαν από 10 εκπαιδευτές. Καθεμία από αυτές τις 10 ομάδες χρησιμοποιήθηκε για να μετρηθεί η ακρίβεια (φάση ελέγχου). Η φάση ελέγχου πραγματοποιήθηκε επίσης σε 5 βήματα. Κατά τη διάρκεια του 1ου βήματος, τα δημογραφικά στοιχεία καθώς επίσης και οι δύο πρώτες συμβουλευτικές συναντήσεις και οι δύο πρώτες γραπτές εργασίες του επόμενου ακαδημαϊκού έτους χρησιμοποιήθηκαν για να προβλέψουν τον τελικό βαθμό κάθε σπουδαστή. Αυτό το βήμα επαναλήφθηκε 10 φορές (για τα στοιχεία κάθε εκπαιδευτή). Κατά τη διάρκεια του 2ου βήματος τα στοιχεία από το πρώτο βήμα μαζί με τα στοιχεία από την τρίτη συνάντηση χρησιμοποιήθηκαν προκειμένου να προβλεφθεί η βαθμολογία κάθε σπουδαστή. Κατά τη διάρκεια του 3ου βήματος τα στοιχεία του 2ου βήματος μαζί με τα στοιχεία από την τρίτη γραπτή εργασία χρησιμοποιήθηκαν προκειμένου να προβλεφθεί η βαθμολογία των σπουδαστών, κτλ.

Χρησιμοποιήσαμε ένα αντιπροσωπευτικό αλγόριθμων ανά κατηγορία μεθόδων παλινδρόμησης για τις συγκρίσεις μας:

- Τον M5' (Wang & Witten, 1997) ως εκπρόσωπο των δέντρων αποφάσεων
- Τον LWR (Atkeson και άλλοι, 1997b) ως εκπρόσωπο των οκνηρών αλγόριθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMOreg (Shevade και άλλοι, 2000) ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης
- Τον M5rules ως εκπρόσωπο των κανόνων ταξινόμησης
- Τον LR (Fox, 1997) ως εκπρόσωπο της γραμμικής παλινδρόμησης

Στον παρακάτω πίνακα παρουσιάζεται το ευκολότερα κατανοητό μέτρο - μέσο απόλυτο σφάλμα (mean absolute error)- κάθε αλγόριθμου για όλα τα βήματα ελέγχου.

	<i>M5'</i>	<i>BP</i>	<i>LR</i>	<i>LWR</i>	<i>SMOreg</i>	<i>M5rules</i>
2 ^η γραπτή εργασία	1.83	2.15	1.89	1.84	1.84	1.83
ΟΣΣ3	1.74	2.08	1.83	1.79	1.78	1.74
3 ^η γραπτή εργασία	1.55	1.79	1.6	1.53	1.56	1.55
ΟΣΣ4	1.54	1.8	1.56	1.5	1.55	1.54
4 ^η γραπτή εργασία	1.23	1.65	1.5	1.4	1.44	1.21

Πίνακας 77. Τα αποτελέσματα των συγκρίσεων

Βάσει των πειραμάτων μας βρέθηκε ότι ο αλγόριθμος M5rules είναι ο πιο κατάλληλος αφού έχει περισσότερο από ικανοποιητική ακρίβεια. Για αυτό το λόγο ενσωματώσαμε αυτόν τον αλγόριθμο στο εργαλείο υποστήριξης που περιγράφηκε στην προηγούμενη ενότητα (βλέπε Εικόνα 79).

The screenshot shows a software window titled "Prediction Tool" with a menu bar (File, Options, Data Set, Model, Help) and a toolbar. Below the menu is a "Classification" button. The main area contains a list of input fields with dropdown menus:

- Sex: male
- Age: 33
- Domestic: single
- Children: 2
- Occupation: full-time
- ComputerKnowledge: yes
- ComputerJob: SeniorUser
- FTOF-1: present
- WRI-1: 5
- FTOF-2: present
- WRI-2: 7
- FTOF-3: present
- WRI-3: 8
- FTOF-4: ?
- WRI-4: ?
- FINALMARK: (empty)

At the bottom right, it displays "Prediction: 6.2". At the bottom left, it says "Working Data set: E:\regression tool\regression test4.csv".

Εικόνα 79. Το εργαλείο πρόβλεψης βαθμών σπουδαστή

Τέλος, πρέπει να τονιστεί ότι η εφαρμογή του εργαλείου δεν περιορίζεται στην πρόβλεψη της απόδοσης των φοιτητών, επιτρέπει στους χρήστες να εισάγουν δεδομένα για οποιοδήποτε πρόβλημα μάθησης και χτίζει τον αντίστοιχο ταξινομητή για αυτό το πρόβλημα.

Το εργαλείο δέχεται το σύνολο το δεδομένων εκπαίδευσης ως λογιστικό φύλλο (spreadsheet) της μορφής αρχείων CSV. Το εργαλείο υποθέτει ότι η πρώτη σειρά του αρχείου CSV χρησιμοποιείται για τα ονόματα των μεταβλητών. Δεν υπάρχει οποιοσδήποτε περιορισμός στη διαταξή των μεταβλητών. Όμως, η μεταβλητή στόχος πρέπει να είναι στην τελευταία στήλη. Μόλις εισαχθούν τα δεδομένα στο εργαλείο, κάθε μεταβλητή εξετάζεται αυτόματα ως προς τον τύπο των στοιχείων της (παραδείγματος χάριν, εάν περιέχει αριθμητικές ή συμβολικές τιμές). Οι άγνωστες τιμές δίνονται σαν είσοδο στο εργαλείο με το σύμβολο "?".

7.2 Άλλες εφαρμογές χρήσης των μεθόδων μάθησης

Υπάρχουν πολλές διαδικασίες στην χημική μηχανική, όπου τα ποιοτικά χαρακτηριστικά του προϊόντος δεν μπορούν να μετρηθούν αντικειμενικά είτε λόγω της έλλειψης κατάλληλων αισθητήρων είτε λόγω της απουσίας οποιασδήποτε συσκευής μέτρησης. Σε αυτές τις περιπτώσεις, ένας εμπειρογνώμονας απασχολείται για το χαρακτηρισμό της ποιότητας των προϊόντων, βασισμένος στην εμπειρία και τις αντιλήψεις του. Η διαδικασία της χρήσης του εμπειρογνώμονα για την εκτέλεση της ταξινόμησης της ποιότητας απαιτεί συνήθως τη διακοπή της διαδικασίας προκειμένου να συλλεχθεί ένα δείγμα. Επιπλέον, αυτός ο τρόπος

ελέγχου της ποιότητας των προϊόντων είναι πολύ υποκειμενικός και κατά συνέπεια μπορεί να οδηγήσει σε σημαντικά λάθη, ειδικά όταν δεν απασχολείται πάντα ο ίδιος εμπειρογνώμονας για να εκτελέσει την ταξινόμηση.

Μια διαφορετική προσέγγιση για την αξιολόγηση ποιοτικών παραμέτρων είναι να χρησιμοποιηθούν οι εποπτευμένοι αλγόριθμοι μάθησης προκειμένου να αυτοματοποιηθεί η διαδικασία. Στις επόμενες ενότητες αναλύονται δυο εφαρμογές.

7.2.1 Χρήση μεθόδων μάθησης για την εκτίμηση της ποιότητας του κρασιού

Τα πρόσφατα αποσταγμένα ποτά όπως το ουίσκι και το ρούμι έχουν πικάντικη, δυσάρεστη γεύση. Οι ποιότητα τους βελτιώνεται με την αποθήκευση των αποσταγμάτων σε βαρέλια για αρκετά έτη. Κατά τη διάρκεια αυτής της διαδικασίας, η οποία είναι γνωστή ως διαδικασία ωρίμανσης, πολλές χημικές αντιδράσεις πραγματοποιούνται. Εντούτοις, λόγω του μεγάλου αριθμού παραγόντων που έχουν επίπτωση στη διαδικασία ωρίμανσης όπως το μέγεθος, η φύση και η χρήση του βαρελιού, περιβαλλοντικών συνθηκών κ.λπ., οι μηχανισμοί ωρίμανσης δεν γίνονται κατανοητοί εντελώς. Επιπλέον, δεν υπάρχει κανένας αξιόπιστος χημικός ή φυσικός δείκτης που μπορεί να δείχνει την πρόοδο της διαδικασίας ωρίμανσης.

Η περίπτωση που μελετήσαμε είναι ένα μέρος της διαδικασίας ωρίμανσης της εταιρίας ΜΕΤΑΞΑ (Tsekouras και άλλοι, 2002), μιας διάσημης επιχείρησης παραγωγής ελληνικού κρασιού. Μέχρι σήμερα, ένας εμπειρογνώμονας δοκιμάζει μερικά δείγματα από τα βαρέλια και πραγματοποιεί την ταξινόμηση ποιότητας των προϊόντων, βασισμένη στις αντιλήψεις του. Είναι σαφές ότι ένα σύστημα, που θα μπορεί αυτόματα να προβλέπει τις ιδιότητες των αποσταγμάτων, βασισμένο σε μερικά ακριβώς μετρήσιμα χαρακτηριστικά του αποστάγματος, είναι μεγάλης σημασίας. Εφαρμόσαμε τους πιο γνωστούς αλγόριθμους παλινδρόμησης για να αναπτύξουμε τους ταξινομητές που θα βαθμολογούν το άρωμα και την γεύση του κρασιού, βασιζόμενοι στις ακόλουθες παραμέτρους:

1. x_1 = ο αριθμός ξαναγεμισμάτων κάθε βαρελιού σε έτη (barrelUsage)
2. x_2 = η παλαιότητα του βαρελιού σε έτη (barrelAge)
3. x_3 = η ηλικία του αποστάγματος σε έτη (distillate age)

Τα διαθέσιμα στοιχεία αποτελούνταν από 170 ταξινομημένα δεδομένα εισόδου. Οι τιμές εξόδου στο σύνολο των στοιχείων είναι οι ταξινομήσεις για το άρωμα (y_1) και την γεύση (y_2), που δίνονται από τον εμπειρογνώμονα χρησιμοποιώντας τιμές που κυμαίνονται από 0 έως 10, με ένα βήμα 1, όπου 0 και 10 αντιστοιχούν στη χειρότερη και καλύτερη ποιότητα, αντίστοιχα.

Χρησιμοποιήσαμε ένα αντιπροσωπευτικό αλγορίθμων ανά κατηγορία μεθόδων παλινδρόμησης για τις συγκρίσεις μας:

- Τον M5' (Wang & Witten, 1997) ως εκπρόσωπο των δέντρων αποφάσεων
- Τον LWR (Atkeson και άλλοι, 1997b) ως εκπρόσωπο των οκνηρών αλγορίθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMOreg (Shevade και άλλοι, 2000) ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης
- Τον M5rules ως εκπρόσωπο των κανόνων ταξινόμησης
- Τον LR (Fox, 1997) ως εκπρόσωπο της γραμμικής παλινδρόμησης

Στον παρακάτω πίνακα παρουσιάζονται όλες οι μετρικές αξιολόγησης μεθόδων παλινδρόμησης (όπου p : προβλεθίσεις τιμές, a : πραγματικές τιμές και $\bar{a} = \frac{1}{n} \sum_i a_i$).

Μέσο απόλυτο σφάλμα (Mean absolute error)	$\frac{ p_1 - a_1 + \dots + p_n - a_n }{n}$
Μέσο τετραγωνικό σφάλμα (Root mean squared error)	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$
Μέσο σχετικό σφάλμα (Relative absolute error)	$\frac{ p_1 - a_1 + \dots + p_n - a_n }{ a_1 - \bar{a} + \dots + a_n - \bar{a} }$
Μέσο σχετικό τετραγωνικό σφάλμα (Root relative squared error)	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}}$

Πίνακας 78. Μετρικές αξιολόγησης μεθόδων παλινδρόμησης

Όλες οι εκτιμήσεις ακρίβειας λήφθηκαν με τον υπολογισμό μέσου όρου των αποτελεσμάτων τρέχοντας 10 φορές την στρωματοποιημένη διαγώνια-επικύρωση (10 times stratified 10-fold cross-validation) και για την εκτίμηση του αρώματος του κρασιού και για την εκτίμηση της γεύσης του.

	<i>M5'</i>	<i>BP</i>	<i>LR</i>	<i>LWR</i>	<i>SMOreg</i>	<i>M5rules</i>
Μέσο απόλυτο σφάλμα	0.45	0.83	0.78	0.72	0.76	0.46
Μέσο τετραγωνικό σφάλμα	0.56	1.01	0.93	0.86	0.97	0.56
Μέσο σχετικό σφάλμα	22.48	40.93	38.52	35.71	37.21	22.69
Μέσο σχετικό τετραγωνικό σφάλμα	22.16	40.27	36.86	34.11	37.95	22.34

Πίνακας 79. Σύγκριση των αλγορίθμων για την βαθμολόγηση του αρώματος του κρασιού

	<i>M5'</i>	<i>BP</i>	<i>LR</i>	<i>LWR</i>	<i>SMOreg</i>	<i>M5rules</i>
Μέσο απόλυτο σφάλμα	0.62	0.88	0.76	0.75	0.75	0.64
Μέσο τετραγωνικό σφάλμα	0.79	1.10	0.96	0.95	0.98	0.81
Μέσο σχετικό σφάλμα	25.31	36.09	31.23	30.85	31.02	26.29
Μέσο σχετικό τετραγωνικό σφάλμα	28.06	39.37	34.42	33.75	35.02	29.08

Πίνακας 80. Σύγκριση των αλγορίθμων για την βαθμολόγηση της γεύσης του κρασιού

Ως τελικό συμπέρασμα μπορούμε να πούμε ότι οι κανόνες που εξάγονται από τον *M5'* αλγόριθμο είναι τουλάχιστον τόσο ακριβείς όσο οι αποφάσεις των εμπειρογνομόνων.

7.2.2 Χρήση μεθόδων μάθησης για εκτίμηση ποιότητας χαρτομάντιλων

Άλλη μια εφαρμογή των μεθόδων μάθησης είναι η εκτίμηση της ποιότητας των χαρτομάντιλων. Τα χαρτομάντιλα είναι ένας ειδικός τύπος χαρτιού, που χρησιμοποιούνται κυρίως για την φροντίδα του προσώπου, έτσι τα ποιοτικά ζητήματα στην παραγωγή τους είναι σημαντικότερα απ'ό,τι σε οποιοδήποτε άλλο τύπο χαρτιού. Υπάρχουν διάφορες παράμετροι, οι οποίες έχουν επιπτώσεις στη ποιότητα των χαρτομάντιλων, συμπεριλαμβανομένου της κατασκευής του πολτού, του καθαρισμού του πολτού και του πάχους του αποθέματος στη μηχανή χαρτιού.

Προσπαθήσαμε να αυτοματοποιήσουμε την διαδικασία της αξιολόγησης της ποιότητας χρησιμοποιώντας τις τεχνικές μηχανικής μάθησης. Τα διαθέσιμα δεδομένα αποτελούνταν από 375 ταξινομημένα δεδομένα εισόδου. Χρησιμοποιήσαμε την μαλακότητα του χαρτιού ως μεταβλητή εξόδου και τις ακόλουθες παραμέτρους ως μεταβλητές εισόδου (Corboy, 1991), (Tsekouras και άλλοι, 2002):

- x1: διαγώνιας κατεύθυνσης αντοχή του χαρτιού (cross directional tensile of tissue)
- x2: μηχανική αντοχή του χαρτιού (machine directional tensile of tissue)
- x3: κατευθυντικό τέντωμα μηχανών του ιστού (machine directional stretch of tissue)
- x4: Ποσοστό επιστρώματος (Yankee coating rate - ml/min)
- x5: Ποσοστό απελευθέρωσης (Yankee release rate - ml/min)

Χρησιμοποιήσαμε έναν αντιπροσωπευτικό αλγόριθμο ανά κατηγορία μεθόδων μάθησης:

- Τον NB ως εκπρόσωπο των Μπεϋζιανών αλγορίθμων
- Τον C4.5 ως εκπρόσωπο των δέντρων αποφάσεων
- Τον 3NN ως εκπρόσωπο των οκνηρών αλγορίθμων
- Τον BP ως εκπρόσωπο των Νευρωνικών Δικτύων
- Τον SMO ως εκπρόσωπο των μηχανών διανυσμάτων υποστήριξης
- Τον RIPPER ως εκπρόσωπο των κανόνων ταξινόμησης

Όλες οι εκτιμήσεις ακρίβειας λήφθηκαν με τον υπολογισμό μέσου όρου των αποτελεσμάτων από τη στρωματοποιημένη 10-φορών διαγώνια-επικύρωση στο σύνολο δεδομένων μας. Κάποιος μπορεί να δει από τον παρακάτω πίνακα ότι αυτό το πρόβλημα ταξινόμησης είναι πολύ δύσκολο, δεδομένου ότι η ακρίβεια όλων των γνωστών ταξινομητών είναι κάτω από το 68% σύμφωνα με το σύνολο δεδομένων μας.

	NB	C4.5	3NN	BP	RIPPER	SMO
Ακρίβεια	54.93%	67.20%	60.53%	58.93%	64.27%	60%

Πίνακας 81. Ακρίβεια των μεθόδων στα δεδομένα μας

Για αυτόν τον λόγο, υπήρχε η ανάγκη για μια τεχνική που θα έδινε μεγαλύτερη ακρίβεια. Στην παρακάτω εικόνα παρουσιάζεται ένας αλγόριθμος συνδυασμού τυχαίων δέντρων απόφασης.

ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ ΜΟΝΤΕΛΟΥ

Για $i = 1$ έως n :

- Πάρε ένα τυχαίο δείγμα (bootstrap sample) από το σύνολο των δεδομένων
- Δημιούργησε ένα "τυχαίο" δέντρο, όπου κάθε κόμβος, είναι μια τυχαία επιλεγμένη μεταβλητή. Αύξησε το δέντρο στο μέγιστο μέγεθος (δεν χρησιμοποιείται κλάδεμα).
- Αποθήκευσε τα παραγόμενα δέντρα

ΤΑΞΙΝΟΜΗΣΗ

Για κάθε ένα από τα n δέντρα:

- Πάρε την πρόγνωση
- Επέστρεψε την κατηγορία που έχει προβλεφθεί από τα περισσότερα δέντρα

Εικόνα 80. Αλγόριθμος συνδυασμού τυχαίων δέντρων απόφασης (Bagging random trees)

Το προτεινόμενο μοντέλο έχει μια ελεύθερη παράμετρο: τον αριθμό των δέντρων απόφασης (n) που θα χτιστούν. Στα πειράματά μας, χρησιμοποιήσαμε 90 δέντρα δεδομένου ότι οι

πολλοί ερευνητές χρησιμοποιούν για επίτευξη βέλτιστης απόδοσης 100 ταξινομητές (Freund & Schapire, 1996) και θέλαμε η τεχνική μας να χρησιμοποιεί λιγότερο χρόνο για την εκπαίδευσή της.

Συγκρίναμε την παρουσιαζόμενη τεχνική με τις άλλες γνωστές τεχνικές που χτίζουν ομάδες ταξινομητών. Όλες οι εκτιμήσεις ακρίβειας λήφθηκαν με τον υπολογισμό μέσου όρου των αποτελεσμάτων από τη στρωματοποιημένη 10 φορές διαγώνια-επικύρωση στο σύνολο δεδομένων μας. Η παρουσιαζόμενη τεχνική (βλέπε Πίνακα 82) είναι στατιστικά ακριβέστερη (δοκιμή-τ με $p < 0.05$) από τον αλγόριθμο Bagging C4.5 με τη χρήση 100 μοντέλων, τον αλγόριθμο Boosting C4.5 με τη χρήση 100 μοντέλων, τον αλγόριθμο Boosting NB με τη χρήση 100 μοντέλων, τον αλγόριθμο Random Rorest (Breiman, 2001) με 100 δέντρα και τον αλγόριθμο Decorate C4.5 με 100 επαναλήψεις.

	Bagging random trees	Bagging C4.5 (100 models)	Boosting C4.5 (100 models)	Boosting NB (100 models)	Random Rorest (100 trees)	Decorate C4.5
Ακρίβεια	74.40%	69.60%	69.60%	62.67%	69.07%	69.07%

Πίνακας 82. Σύγκριση της μεθόδου συνδυασμού τυχαίων δέντρων απόφασης με άλλες μεθόδους δημιουργίας ομάδας ταξινομητών

Επιπλέον, η προτεινόμενη τεχνική (βλέπε Πίνακα 83) είναι στατιστικά ακριβέστερη από τον αλγόριθμο Multiboost C4.5 με τη χρήση 100 μοντέλων, τον αλγόριθμο Logitboost DS με τη χρήση 100 μοντέλων, τον αλγόριθμο Bagging BP με τη χρήση 100 μοντέλων, τον αλγόριθμο Boosting BP με τη χρήση 100 μοντέλων και τον αλγόριθμο LMT (Landwehr & Hall, 2003).

	Bagging random trees	Multiboost C4.5 (100 models)	Logitboost DS (100 models)	Bagging BP (100 models)	Boosting BP (100 models)	LMT
Ακρίβεια	74.40%	69.60%	60.27%	61.07%	58.67%	68.27%

Πίνακας 83. Σύγκριση της μεθόδου συνδυασμού τυχαίων δέντρων απόφασης με άλλες μεθόδους

Για να συνοψίσουμε, η απόδοση της προτεινόμενης τεχνικής είναι ακριβέστερη από τους πιο γνωστούς ταξινομητές και τις πιο γνωστές ομάδες ταξινομητών. Η προτεινόμενη τεχνική μπορεί να επιτύχει μια αύξηση στην ακρίβεια ταξινόμησης από 11% έως 35% έναντι των γνωστών ταξινομητών. Επιπλέον, η μέση σχετική βελτίωση ακρίβειας της προτεινόμενης μεθοδολογίας είναι από 7% έως 26% σε σχέση με τις γνωστές ομάδες ταξινομητών. Συνεπώς, είναι δυνατόν να δοθεί μια εφικτή λύση στο πρόβλημα εκτίμησης της ποιότητας των χαρτομάντιλων με την προτεινόμενη τεχνική.

Πρέπει επίσης να αναφερθεί ότι η προτεινόμενη μεθοδολογία μπορεί εύκολα να εκτελεστεί παράλληλα εκτός από σειριακά χτίζοντας ένα δέντρο απόφασης ανά μηχανή, μειώνοντας το χρόνο εκπαίδευσής της.

ΚΕΦΑΛΑΙΟ 8. Μη επιτηρούμενη μηχανική μάθηση

Παρότι η παρούσα διατριβή ασχολείται κυρίως με την επιτηρούμενη μηχανική μάθηση θα ήταν παράληψη να μην αναφερθεί τίποτα για την μη επιτηρούμενη μηχανική μάθηση. Αυτό το κεφάλαιο αφιερώνεται λοιπόν στην μη επιτηρούμενη μηχανική μάθηση (Jain και άλλοι, 1999).

Δεν είναι όλα τα προβλήματα μάθησης επιβλεπόμενα – ένα παράδειγμα μη επιβλεπόμενης (unsupervised) μάθησης είναι αυτό της ομαδοποίησης εγγράφων (document clustering), κατά το οποίο το ζητούμενο είναι να ομαδοποιηθούν τα έγγραφα σε κατηγορίες, άγνωστες εκ των προτέρων. Στη μη επιτηρούμενη μάθηση, δεν παρέχεται κάποια εμπειρία E για να καθοδηγήσει τη μάθηση, αλλά ο στόχος είναι να αναδειχθεί η δομή οργάνωσης των δεδομένων μέσω κάποιου ή κάποιων κατάλληλα επιλεγμένων κριτηρίων “ομοιότητας”.

Με άλλα λόγια, η ανάλυση συστάδων (cluster analysis) ή πιο απλά ομαδοποίηση (clustering) είναι η οργάνωση μιας συλλογής από δεδομένα σε συστάδες (clusters) με βάση κάποιο μέτρο ομοιότητας. Στοιχεία που ανήκουν στην ίδια ομάδα παρουσιάζουν μεγαλύτερη ομοιότητα από ότι στοιχεία που ανήκουν σε διαφορετικές ομάδες. Ένα παράδειγμα ομαδοποίησης φαίνεται στην Εικόνα 81. Τα δεδομένα είναι αυτά που φαίνονται στο (α) και η ομαδοποίηση τους φαίνεται στο (β). Στοιχεία που ανήκουν στην ίδια ομάδα φέρουν κοινό αριθμό. Στην μη επιτηρούμενη μηχανική μάθηση, επαφίεται σε εμάς να καθορίσουμε την σημασία που θα έχει κάθε μια από τις συγκροτήσεις που προκύπτουν.

Εικόνα 81. Ομαδοποίηση δεδομένων (data clustering)

Υπάρχει ένας τουλάχιστον σημαντικός λόγος για τον οποίο είναι σημαντική η ομαδοποίηση. Υπάρχουν συχνά περιπτώσεις όπου η συλλογή δεδομένων είναι σχετικά “φθηνή”, όμως ο προσδιορισμός της αντίστοιχης κατηγορίας είναι “ακριβός”, π.χ. σε εφαρμογές αναγνώρισης φωνής.

Στη συνέχεια παραθέτουμε κάποια ενδεικτικά παραδείγματα εφαρμογής της ομαδοποίησης:

- Τμηματοποίηση του πληθυσμού μιας χώρας σε διάφορα μορφωτικά επίπεδα. Στην περίπτωση αυτή κάθε στιγμιότυπο του προβλήματος έχει ιδιότητες όπως ηλικία, τόπος κατοικίας, οικονομική κατάσταση, μόρφωση, κ.α., οπότε εφαρμόζοντας τμηματοποίηση μπορούμε να διασπάσουμε τον πληθυσμό με βάση τα γνωρίσματα αυτά σε μια προσπάθεια να βρούμε το προφίλ του κάθε τμήματος.

- Διαχωρισμός των χρηστών ενός δικτυακού τόπου με βάση τα κινηματογραφικά τους ενδιαφέροντα. Στην περίπτωση αυτή κάθε στιγμιότυπο του προβλήματος έχει ιδιότητες όπως ηλικία, προηγούμενες προτιμήσεις σε ταινίες, επάγγελμα, μόρφωση, κ.α., οπότε μέσω της ομαδοποίησης μπορούμε να διασπάσουμε τους χρήστες σε τμήματα σε μια προσπάθεια να βρούμε το προφίλ του κάθε τμήματος.

Συνοπτικά τα βήματα που ακολουθούνται κατά την διαδικασία της συγκρότησης (ομαδοποίησης) είναι:

1. Αναπαράσταση των δεδομένων: αφορά στον αριθμό των διαθέσιμων δεδομένων και στον αριθμό και τύπο των χαρακτηριστικών τα οποία περιγράφουν το πρόβλημα.
2. Ορισμός του μέτρου ομοιότητας (Similarity measure): Σε κάθε συστάδα (cluster) τα στοιχεία που περιέχονται σε αυτή παρουσιάζουν ομοιότητα μεταξύ τους. Η διαφορετικότητα είναι βασική για να ορισθεί μια ξεχωριστή συστάδα. Το μέτρο ομοιότητας μεταξύ των δεδομένων καθορίζεται από μια συνάρτηση απόστασης. Δεδομένης της μεγάλης ποικιλίας στα χαρακτηριστικά των στοιχείων, η επιλογή του μέτρου ομοιότητας ανά περίπτωση είναι πολύ σημαντική.
3. Η καθεαυτή διαδικασία της συγκρότησης το αποτέλεσμα της οποίας μπορεί να είναι απόλυτα καθορισμένο (ομαδοποίηση των δεδομένων σε συστάδες), ή ασαφές (fuzzy) (όπου κάθε στοιχείο δεν ανήκει μόνο σε μία συστάδα αλλά είναι μέλος σε κάποιο βαθμό όλων των συστάδων).
4. Αφαιρετικότητα (Data abstraction): η διαδικασία κατά την οποία το σύνολο των δεδομένων αποκτά μια απλή και συμπαγή αναπαράσταση. Ο όρος απλή αναπαράσταση μπορεί να εξηγηθεί είτε από την μεριά της αυτοποιημένης ανάλυσης είτε από την μεριά του ανθρώπου. Στην πρώτη περίπτωση θα θέλαμε τα δεδομένα μας να αναπαριστώνται με τέτοιο σαφή και απλό τρόπο ώστε μια περαιτέρω υπολογιστική επεξεργασία να είναι εφικτή. Στην δεύτερη περίπτωση η απλή αναπαράσταση των δεδομένων τα κάνει πιο κατανοητά στους ειδικούς που πρόκειται να τα επεξεργαστούν και να βγάλουν συμπεράσματα. Συνήθως η αφαιρετικότητα στη συγκρότηση είναι μια συνοπτική αναπαράσταση κάθε συστάδας με την βοήθεια κάποιου αντιπροσωπευτικού στοιχείου το οποίο καλείται κεντροειδές (centroid).
5. Προσδιορισμός και εκτίμηση του αποτελέσματος (Assessment of output): η εκτίμηση των αποτελεσμάτων της διαδικασίας που χαρακτηρίζει μια καλή διαδικασία από μια όχι και τόσο επιτυχή. Κυρίως αυτό που εκτιμάται είναι το τελικό αποτέλεσμα, δηλαδή κατά πόσο οι συστάδες που δημιουργήθηκαν έχουν νόημα και κατά πόσο αυτές δεν δημιουργήθηκαν κατά τυχαίο τρόπο εξαιτίας του συγκεκριμένου αλγορίθμου που χρησιμοποιήθηκε.

Στις επόμενες ενότητες τα 5 προαναφερθέντα βήματα αναλύονται σε μεγαλύτερο βάθος.

8.1 Αναπαράσταση των δεδομένων

Όπως σε πολλά ζητήματα που αφορούν την ομαδοποίηση, έτσι και εδώ δεν υπάρχουν σαφείς οδηγίες, αρχές που να προτείνουν την βέλτιστη αναπαράσταση των στοιχείων και την καλύτερη επιλογή χαρακτηριστικών. Ο ρόλος του αναλυτή στην αναπαράσταση των δεδομένων είναι να συγκεντρώσει όσο το δυνατόν περισσότερα στοιχεία (γνώσεις, συμπεράσματα, υποθέσεις) σχετικά με τα δεδομένα. Με βάση την γνώση αυτή για τα δεδομένα είναι περισσότερο εφικτή η κατάλληλη επιλογή και παραγωγή μεταβλητών (χαρακτηριστικών). Μια πολύ προσεκτική ματιά στα σημαντικότερα χαρακτηριστικά και πιθανών κάποιοι μετασχηματισμοί πάνω σε αυτά θα μπορούσαν να οδηγήσουν σε καλύτερα αποτελέσματα.

Είναι γεγονός ότι η καλή αναπαράσταση των στοιχείων συνήθως οδηγεί σε μια απλή και κατανοητή ομαδοποίηση. Αντίθετα μια όχι και τόσο προσεκτική αναπαράσταση μπορεί να οδηγήσει σε μια πολύπλοκη ομαδοποίηση με δομές δύσκολα αντιληπτές και ερμηνεύσιμες.

8.2 Μέτρα Ομοιότητας

Το πιο προφανές μέτρο ομοιότητας δύο στιγμιοτύπων είναι η μεταξύ τους απόσταση. Ένας καλός τρόπος για να ξεκινήσουμε την ομαδοποίηση είναι να ορίσουμε ένα κατάλληλο απόσταση και κατόπιν, να υπολογίσουμε έναν τετραγωνικό πίνακα αποστάσεων μεταξύ όλων των στιγμιοτύπων. Είναι φυσικό να περιμένουμε ότι η απόσταση μεταξύ στιγμιοτύπων της ίδιας συστάδα (σμήνος) να είναι σημαντικά μικρότερη από την απόσταση μεταξύ στιγμιοτύπων σε διαφορετικά σμήνη.

Ας υποθέσουμε ότι δύο στιγμιότυπα ανήκουν στο ίδιο σμήνος εάν η Ευκλείδεια απόσταση ανάμεσά τους είναι μικρότερη από μια κρίσιμη απόσταση d_0 . Είναι προφανές ότι η τιμή του d_0 είναι αποφασιστική. Εάν η κρίσιμη απόσταση d_0 είναι πολύ μεγάλη, τότε όλα τα δείγματα θα τοποθετηθούν σε ένα μόνον σμήνος. Εάν η κρίσιμη απόσταση d_0 είναι πολύ μικρή, τότε το κάθε δείγμα θα σχηματίσει ένα μεμονωμένο σμήνος. Προκειμένου να προκύψουν “φυσικά σμήνη”, η κρίσιμη απόσταση d_0 θα πρέπει να έχει κάποια ενδιάμεση τιμή.

Γενικώς η χρήση της Ευκλείδειας απόστασης δικαιολογείται όταν ο χώρος είναι (περίπου) “ισοτροπικός” και τα δείγματα είναι ομοιόμορφα καταναμημένα στο χώρο. Είναι γεγονός ότι τα σμήνη παραμένουν αμετάβλητα σε μετατοπίσεις ή/και περιστροφές, ωστόσο τα σμήνη μεταβάλλονται από γραμμικούς ή άλλους μετασχηματισμούς που αλλοιώνουν την Ευκλείδεια απόσταση.

Ένας τρόπος άρσης τυχόν μεταβολών είναι να κανονικοποιηθούν τα δεδομένα πριν την ομαδοποίηση. Θεωρούμε, για παράδειγμα, την μετατόπιση/αλλαγή κλίμακας των αξόνων ώστε τα δεδομένα σε κάθε διάσταση να έχουν μηδενική μέση τιμή και μοναδιαία διασπορά. Το επιχείρημα για την προαναφερθείσα κανονικοποίηση είναι ότι αποτρέπει κάποιες συνιστώσες από το να κυριαρχήσουν στους υπολογισμούς επειδή χαρακτηρίζονται από μεγάλες αριθμητικές τιμές. Αφαιρώντας τη μέση τιμή και διαιρώντας δια της τυπικής απόκλισης αναμένεται να είναι αποτελεσματική τεχνική κανονικοποίησης εάν η διακύμανση των τιμών είναι τυχαία. Ωστόσο, η προαναφερθείσα τεχνική κανονικοποίησης αναμένεται να είναι λιγότερο αποτελεσματική εάν η διακύμανση των τιμών οφείλεται στην ύπαρξη σμηνών. Αντί να αλλάζουμε την κλίμακα των αξόνων μπορούμε να αλλάζουμε την απόσταση με διάφορους τρόπους. Για παράδειγμα, μια μεγάλη οικογένεια αποστάσεων είναι οι μετρικές αποστάσεις Minkowski μεταξύ δύο διανυσμάτων $\mathbf{x}$ και $\mathbf{x}'$ όπως φαίνεται στη συνέχεια.

$$d(\mathbf{x}, \mathbf{x}') = \left(\sum_{k=1}^d |\mathbf{x}_k - \mathbf{x}'_k|^q \right)^{1/q},$$

όπου το $q \geq 1$ είναι παράμετρος. Για $q = 2$ προκύπτει η γνωστή Ευκλείδεια απόσταση, ενώ για $q = 1$ προκύπτει η απόσταση Manhattan. Άλλες μετρικές απόστασης που μπορούν να χρησιμοποιηθούν έχουν παρουσιαστεί στην ενότητα 3.3 όταν αναφερόμασταν για την μάθηση βασισμένη στα στιγμιότυπα.

Γενικά, η μετρική της απόστασης πρέπει να ικανοποιεί τα παρακάτω αξιώματα:

1. $D(\mathbf{x}, \mathbf{y}) = 0$
2. $D(\mathbf{x}, \mathbf{y}) = D(\mathbf{y}, \mathbf{x})$
3. $D(\mathbf{x}, \mathbf{y}) \leq D(\mathbf{x}, \mathbf{z}) + D(\mathbf{z}, \mathbf{y})$

8.3 Τεχνικές συγκρότησης

Η ποικιλία τεχνικών έκφρασης της ομοιότητας μεταξύ στοιχείων και ομαδοποίησης των δεδομενων έχει ως αποτέλεσμα την ύπαρξη πολλών μεθόδων συγκρότησης. Στην εργασία μας [9] περιγράφονται οι περισσότερες τεχνικές συγκρότησης.

Οι μέθοδοι συσταδοποίησης χωρίζονται στις ακόλουθες κατηγορίες:

- Ιεραρχικοί αλγόριθμοι
- Διαμεριστικοί αλγόριθμοι
- Αυτο-οργανούμενοι χάρτες
- Αλγόριθμοι βασισμένοι στην πυκνότητα (Density based)
- Αλγόριθμοι βασισμένοι στον κάνναβο (Grid based)

8.3.1 Ιεραρχικοί αλγόριθμοι

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης φθάνουν στον τελικό αριθμό ομάδων συνενώνοντας σταδιακά τα διάφορα σημεία, με κριτήριο προτεραιότητας τις μεταξύ τους αποστάσεις. Οι αλγόριθμοι της κατηγορίας αυτής ξεκινούν από ένα μεγάλο αριθμό ομάδων (μερικοί αλγόριθμοι μάλιστα θεωρούν αρχικά ότι κάθε σημείο αποτελεί και ξεχωριστή ομάδα). Στη συνέχεια υπολογίζονται οι αποστάσεις μεταξύ των ομάδων. Προτεραιότητα για συνένωση αποκτά το ζεύγος των ομάδων που απέχουν ελάχιστα. Ακολουθεί το αλγοριθμικό σχήμα της ιεραρχικής ομαδοποίησης.

- Θέσε $C_j = \{P_j\}, j = 1, \dots, n$
- Επανάλαβε έως ότου μείνει μόνο μια ομάδα
 - ο Προσδιόρισε τις ομάδες C_k και C_h που ελαχιστοποιούν την απόσταση $d(C_i, C_j), i, j = 1, \dots, n$:
$$d(C_k, C_h) = \min \{d(C_i, C_j), i, j = 1, \dots, n\}$$
 - ο Θέσε $C_k = C_k \cup C_h$
 - ο Διέγραψε την ομάδα C_h

Εικόνα 82. Αλγοριθμικό σχήμα της ιεραρχικής ομαδοποίησης

Για παράδειγμα της λειτουργίας των ιεραρχικών αλγορίθμων, παρουσιάζεται το δέντροδιάγραμμα που αντιστοιχεί στα δεδομένα της Εικόνας 83α στην Εικόνα 83β. Χαρακτηριστικό του διαγράμματος είναι ότι αποδίδονται διαφορετικές ομαδοποιήσεις σε διαφορετικά επίπεδα του διαγράμματος .

Εικόνα 83. Λειτουργία των ιεραρχικών αλγορίθμων

Πρέπει να σημειωθεί ότι υπάρχει και μια άλλη προσέγγιση στην ιεραρχική ομαδοποίηση (divisive methods) σύμφωνα με την οποία ο αλγόριθμος ομαδοποίησης ξεκινά από μια ενιαία ομάδα, περιλαμβάνουσα όλα τα σημεία, και στη συνέχεια εξελίσσεται διαιρώντας σταδιακά υπερκλάσεις. Επειδή η μέθοδος αυτή σπάνια χρησιμοποιείται, δεν θα αναπτυχθεί περισσότερο.

Για την εκτίμηση της πολυπλοκότητας του αλγόριθμου θα πρέπει να υπολογίσει κανείς ότι στην αρχή υπάρχουν (n) ομάδες και στο τέλος μένει μόνο μια. Έτσι ο βασικός βρόχος του παραπάνω αλγόριθμου εκτελείται (n) φορές. Σε κάθε επανάληψη k προσδιορίζεται το κοντινότερο ζεύγος ομάδων ανάμεσα σε $(n-k+1)$ ομάδες. Ο προσδιορισμός του ζεύγους αυτού στην πρώτη επανάληψη χρειάζεται χρόνο $O(n^2)$. Έτσι είναι φανερό ότι ο αλγόριθμος δεν είναι εφαρμόσιμος για μεγάλες τιμές του n .

Οι περισσότεροι ιεραρχικοί αλγόριθμοι είναι ακολουθούν είτε την προσέγγιση του *απλού-συνδέσμου* (single-link), είτε την προσέγγιση του *πλήρους-συνδέσμου* (complete-link). Οι διαφορές μεταξύ των προσεγγίσεων αυτών έχει να κάνει με τον τρόπο με τον οποίο ορίζουν την ομοιότητα μεταξύ στοιχείων (στιγμιότυπων) και κατά συνέπεια των συστάδων πριν την συγχώνευσή τους. Στην περίπτωση του απλού-συνδέσμου η απόσταση μεταξύ των δύο συστάδων είναι η ελάχιστη από τις αποστάσεις μεταξύ όλων των ζευγών στοιχείων από τις δύο συστάδες (κάθε ζεύγος περιέχει ένα στοιχείο από τη μία συστάδα και ένα από την άλλη). Στην προσέγγιση πλήρους-συνδέσμου η απόσταση μεταξύ δύο συστάδων είναι η μέγιστη από τις αποστάσεις μεταξύ όλων των ζευγών στοιχείων από τις δύο συστάδες. Και στις δύο περιπτώσεις οι δύο συστάδες συγχωνεύονται για να δημιουργήσουν μια όταν η απόσταση αυτή, όπως και αν ορίζεται είναι ελάχιστη.

Στην Εικόνα 84α φαίνονται μια σειρά από στοιχεία τα οποία ορίζουν δύο συστάδες αλλά χωρίζονται από σημεία τα αποτελούν θόρυβο (x) και δεν μας ενδιαφέρουν. Η προσέγγιση απλού-συνδέσμου δημιουργεί τις συστάδες που φαίνονται στην Εικόνα 84α και η προσέγγιση πλήρους συνδέσμου δημιουργεί τις συστάδες της Εικόνα 84β. Είναι προφανές ότι στην δεύτερη περίπτωση οι συστάδες είναι πιο συμπαγές από την πρώτη στην οποία τα στοιχεία θορύβου έχουν δημιουργήσει ανεπιθύμητα φαινόμενα.

Εικόνα 84. α) Ομαδοποίηση απλού-συνδέσμου.

β) Ομαδοποίηση πλήρους-συνδέσμου

Παρόλα αυτά η προσέγγιση απλού συνδέσμου είναι αρκετά ευέλικτη σε δύσκολες περιπτώσεις. Για παράδειγμα η προσέγγιση απλού-συνδέσμου μπορεί να εξάγει τις ομόκεντρες συστάδες που φαίνονται στην Εικόνα 85 ενώ η προσέγγιση πλήρους-συνδέσμου δεν μπορεί.

Εικόνα 85. Δύο ομόκεντρικές συστάδες

Κατά συνέπεια, καμία από τις δύο προσεγγίσεις δεν είναι πανάκεια. Τελειώνοντας την σύντομη παρουσίαση των ιεραρχικών αλγορίθμων πρέπει να αναφέρουμε ότι οι πιο γνωστοί ιεραρχικοί αλγόριθμοι είναι οι BIRCH (Zhang και άλλοι, 1997), CURE (Guha και άλλοι, 1998), CHAMELEON (Karypis και άλλοι, 1999), ROCK (Guha και άλλοι, 1999) και GRIN (Chien-Yu και άλλοι, 2002).

8.3.2 Διαμεριστικοί αλγόριθμοι

Ένας διαμεριστικός αλγόριθμος έχει ως αποτέλεσμα μια διαμέριση του χώρου των δεδομένων σε αντίθεση με τους ιεραρχικούς αλγορίθμους που δημιουργούν πιο πολύπλοκες δομές που περιγράφονται από δένδρογράμματα. Σε αντίθεση με τις ιεραρχικές μεθόδους κατηγοριοποίησης που βασίζονται στην πληροφορία που αποθηκεύεται στον πίνακα ομοιότητας, οι διαμεριστικοί μέθοδοι εφαρμόζονται απευθείας στα δεδομένα με αποτέλεσμα να έχουν μικρότερες απαιτήσεις σε χώρο και να διαχειρίζονται ευκολότερα πολυπληθή σύνολα δεδομένων. Το κυρίως πρόβλημα των αλγορίθμων αυτών είναι η απόφαση για τον αριθμό των τελικών συστάδων.

Ο αριθμός αυτός καθορίζεται κυρίως από την προσπάθεια βελτιστοποίησης. Στην πραγματικότητα αυτό που γίνεται είναι να τρέχει ο αλγόριθμος για διάφορους αριθμούς από συστάδες και να επιλέγεται εκείνη η τελική κατάσταση η οποία βελτιστοποιεί μια συνάρτηση. Το κριτήριο που κυρίως χρησιμοποιείτε σε διαμεριστικούς αλγορίθμους για την τελική απόφαση του αριθμού των συστάδων είναι το κριτήριο του τετραγωνικού λάθους ή η συνάρτηση τετραγωνικού λάθους (squared error function). Αυτή η συνάρτηση ορίζεται για ένα αποτέλεσμα ομαδοποίησης L με σύνολο στοιχείων S και K συστάδες ως εξής:

$$e^2(K, L) = \sum_{j=1}^K \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$$

όπου $x_i^{(j)}$ είναι το i στοιχείο της j συστάδας, και το c_j είναι το κεντροειδές της j συστάδας.

Ο αλγόριθμος K-μέσων (K-means) είναι ένας πολύ απλός και πολύ διαδεδομένος διαμεριστικός αλγόριθμος που χρησιμοποιεί το κριτήριο του τετραγωνικού λάθους. Ο αλγόριθμος ξεκινά με μια τυχαία διαμέριση και συνεχώς τοποθετεί στοιχεία στις συστάδες με βάση την απόσταση των στοιχείων από το κεντροειδές (γεωμετρικός μέσος) της συστάδας. Αυτό σταματάει όταν ικανοποιηθεί κάποιο κριτήριο όπως η ελαχιστοποίηση της συνάρτησης τετραγωνικού λάθους ή η μη διαφοροποίηση των συστάδων από κάποια επανάληψη και μετά. Αναλυτικότερα ο αλγόριθμος K-μέσων (K-means) περιγράφεται σε ψευδογλώσσα στην παρακάτω.

1. Επιλογή k συστάδων. Τα αντικείμενα της βάσης εντάσσονται αρχικά σε κάποια συστάδα είτε με βάση κάποια συνάρτηση είτε με τυχαίο τρόπο.
2. Τοποθέτηση κάθε αντικειμένου στη πιο κοντινή συστάδα μετά από υπολογισμό της απόστασης του σημείου του αντικειμένου από το κεντροειδές της συστάδας.
3. Υπολογισμός του νέου κεντροειδούς μετά την αλλαγή των αντικειμένων της κάθε συστάδας που έγινε στο βήμα 2.
4. Αν το κριτήριο τερματισμού (αν δηλαδή υπάρχουν αλλαγές στη δομή όλων των συστάδων) δεν ικανοποιείται επανάληψη των βημάτων 2 και 3.

Εικόνα 86. Ο αλγόριθμος K-μέσων (K-means)

Στην Εικόνα 87 παρατίθεται η γραφική αναπαράσταση της ανακατανομής των στοιχείων σε συστάδες και ο επαναληπτικός υπολογισμός των κεντροειδών με τον αλγόριθμο k-means.

Εικόνα 87. Ανακατανομή στοιχείων σε συστάδες και επαναληπτικός υπολογισμός κεντροειδών με τον αλγόριθμο k-means

Αν η επιλογή της τυχαίας διαμέρισης δεν είναι αρκετά προσεκτική τότε το κριτήριο τετραγωνικού λάθους συγκλίνει σε τοπικά ελάχιστο κάνοντας την τελική επιλογή συστάδων ανεπιτυχή. Ας θεωρήσουμε το χώρο της Εικόνας 88 με επτά στοιχεία. Αν η αρχική μας επιλογή είναι τρεις διαμερίσεις με αρχικά στοιχεία το A,B,C στην κάθε μια, το αποτέλεσμα της ομαδοποίησης θα είναι αυτό που φαίνεται στην Εικόνα 88 με τις ελλείψεις. Αντίθετα αν η αρχική επιλογή είναι οι διαμερίσεις με σημεία το A,D,F οι τελικές συστάδες φαίνονται με τα παραλληλόγραμμα. Όπως είναι εύκολα αντιληπτό στην πρώτη περίπτωση το κριτήριο τετραγωνικού λάθους είναι αρκετά μεγαλύτερο από την δεύτερη περίπτωση στην οποία επιλέξαμε διαφορετική αρχική διαμέριση.

Ο αλγόριθμος K-Means αποτελεί ευρεστική (heuristic) προσέγγιση στο πρόβλημα της ομαδοποίησης και η διάδοσή του οφείλεται στην ταχύτητά του και την ευκολία υλοποίησής του. Πρέπει όμως να σημειωθεί ότι, όπως όλοι οι ευρεστικοί αλγόριθμοι, δεν εγγυάται την βέλτιστη λύση (ομαδοποίηση). Ο αλγόριθμος K-Means λοιπόν συχνά καταλήγει σε κάποιο τοπικό ελάχιστο. Το ελάχιστο δε στο οποίο καταλήγει εξαρτάται άμεσα από τις αρχικές συνθήκες, δηλαδή την επιλογή των αρχικών κέντρων. Έτσι στις εφαρμογές συνιστώνται πολλαπλές εφαρμογές του αλγορίθμου, με εκκίνηση κάθε φορά από διαφορετικά αρχικά κέντρα και σύγκριση των αποτελεσμάτων.

Πρέπει επίσης να αναφερθεί ότι οι διαμεριστικές τεχνικές δουλεύουν καλύτερα με συνεχείς και όχι κατηγορικές μεταβλητές.

Εικόνα 88. Ευαισθησία του K-means αλγορίθμου στην αρχική επιλογή διαμέρισης

Τελειώνοντας την σύντομη παρουσίαση των διαμεριστικών αλγορίθμων πρέπει να αναφέρουμε ότι οι δύο πιο σύγχρονοι διαμεριστικοί αλγόριθμοι είναι οι: k-modes (Huang, 1998) και k*-Means (Cheung, 2003)

8.3.3 Αυτό-Οργανούμενοι Χάρτες (SOM)

Το δίκτυο Kohonen είναι ένα ιδιότυπο νευρωνικό δίκτυο που αποτελεί τυπικό παράδειγμα μη επιβλεπόμενης μάθησης (unsupervised learning). Ανήκει σε μια γενικότερη κατηγορία μεθόδων μη επιβλεπόμενης μάθησης, γνωστή ως αυτοοργανωνόμενοι χάρτες χαρακτηριστικών (self-organizing feature maps) και για το λόγο αυτό το δίκτυο Kohonen είναι γνωστό και ως αυτοοργανωνόμενος χάρτης (Kohonen, 1997).

Εικόνα 89. Νευρωνικό δίκτυο Kohonen

Το νευρωνικό δίκτυο Kohonen είναι αρχιτεκτονικής δύο μόνο επιπέδων: ενός επιπέδου εισόδου (input layer) και ενός επιπέδου εξόδου (output layer). Σκοπός του δικτύου Kohonen είναι να απεικονίζει τα N -διανύσματα που εμφανίζονται στην είσοδό του, από τον χώρο των N διαστάσεων σε ένα πλέγμα (grid) μιας ή δύο το πολύ διαστάσεων. Έτσι λοιπόν χρησιμοποιούνται για ομαδοποίηση των διανυσμάτων εισόδου με βάση τα χαρακτηριστικά τους και με την προϋπόθεση ότι δεν είναι εκ των προτέρων γνωστός ο αριθμός των ομάδων ή κάποια άλλη πληροφορία για αυτές. Σύμφωνα με την αρχή λειτουργίας των δικτύων Kohonen, αντικείμενα που ευρίσκονται το ένα κοντά στο άλλο στον N -διάστατο χώρο των διανυσμάτων εισόδου θα πρέπει να ευρίσκονται επίσης κοντά το ένα στο άλλο και στον «χάρτη» του επιπέδου εξόδου.

Η εκπαίδευση του δικτύου Kohonen, όπως συμβαίνει και στα άλλα νευρωνικά δίκτυα, συνίσταται στην προσαρμογή των συνοπτικών βαρών. Δύο βασικά στοιχεία που χαρακτηρίζουν τον αλγόριθμο εκπαίδευσης του δικτύου Kohonen είναι ο κανόνας προσαρμογής των συνοπτικών βαρών και η έννοια της «γειτονιάς» των νευρώνων στο επίπεδο εξόδου.

Οι νευρώνες στο επίπεδο εξόδου λειτουργούν ανταγωνιστικά ο ένας με τον άλλο επιδιώκοντας ο καθένας να κερδίσει το αντικείμενο που εμφανίζεται κάθε φορά στην είσοδο του δικτύου. Ο τρόπος με τον οποίο αναδεικνύεται ο νευρώνας-«νικητής» στη διεκδίκηση ενός διανύσματος $P=(p_1, p_2, \dots, p_N)$ που εμφανίζεται στην είσοδο του δικτύου και για δεδομένες τιμές των συνοπτικών βαρών, είναι ο εξής:

1. Για κάθε νευρώνα j του επιπέδου εξόδου, υπολογίζεται ένας δείκτης ομοιότητας d_j του διανύσματος εισόδου $P=(p_1, p_2, \dots, p_N)$ με το διάνυσμα συνοπτικών βαρών του νευρώνα j , $(w_{1j}, w_{2j}, \dots, w_{Nj})$. Ως δείκτης ομοιότητας λαμβάνεται συνήθως η Ευκλείδεια απόσταση μεταξύ των δύο διανυσμάτων.
2. «Νικητής» της διεκδίκησης αναδεικνύεται ο νευρώνας με την μικρότερη τιμή του δείκτη ομοιότητας.

Αρχικά, ο αλγόριθμος ξεκινά με την αρχικοποίηση των διανυσμάτων αναφοράς (συνοπτικά βάρη). Στη συνέχεια επιλέγεται τυχαία ένα δείγμα εκμάθησης από τα δεδομένα εισόδου, και καθορίζεται ο νευρώνας με το κοντινότερο διάνυσμα αναφοράς ως προς το επιλεγμένο σημείο. Τέλος ενημερώνεται το διάνυσμα αναφοράς καθώς και τα γειτονικά του. Αρχικά η γειτονιά περιλαμβάνει όλους του νευρώνες του επιπέδου εξόδου, όσο όμως η εκπαίδευση του

δικτύου εξελίσσεται, η γειτονιά περιορίζεται σταδιακά μέχρι του σημείου να ορίζεται μόνο από ένα νευρώνα. Η διαδικασία επαναλαμβάνεται μέχρι να μην παρατηρούνται πλέον αλλαγές στα διανύσματα αναφοράς. Όταν ολοκληρωθεί η εκπαίδευση, ο αλγόριθμος αντιστοιχίζει τα δεδομένα εισόδου στους κοντινότερους νευρώνες και επιστρέφει τις συστάδες που έχουν διαμορφωθεί.

Αναλυτικά, το αλγοριθμικό σχήμα της εκπαίδευσης του δικτύου Kohonen περιγράφεται παρακάτω:

- Δώσε αρχικές τιμές στα συνοπτικά βάρη, τον ρυθμό εκπαίδευσης και το εύρος της γειτονιάς - Επανάλαβε έως ότου οι αλλαγές στα συνοπτικά βάρη γίνουν αμελητέες   ο Για κάθε διάνυσμα εισόδου   ▪ Εμφάνισε το διάνυσμα στην είσοδο του δικτύου ▪ Προσδιόρισε τον νευρώνα-«νικητή» ▪ Προσδιόρισε τους νευρώνες στη γειτονιά του «νικητή» ▪ Τροποποίησε τα συνοπτικά βάρη στη γειτονιά του «νικητή» ο Περίορισε τη γειτονιά του νικητή - Μείωσε τον ρυθμό εκπαίδευσης

Εικόνα 90. Εκπαίδευση δικτύου Kohonen

Μετά την εκπαίδευσή του ένα δίκτυο Kohonen (SOM) καταλήγει συνήθως σε ένα μικρό αριθμό νευρώνων που αντιπροσωπεύουν ένα μεγάλο αριθμό εγγραφών (ισχυροί νευρώνες) και αρκετούς νευρώνες που δεν αντιστοιχούν σε καμιά εγγραφή (ασθενείς νευρώνες). Οι ισχυροί νευρώνες και ενδεχομένως κάποιοι άλλοι γειτονικοί νευρώνες αντιπροσωπεύουν τα κέντρα των ομάδων.

Η διαδικασία κατηγοριοποίησης δεδομένων με χρήση SOM μοιάζει αρκετά με τις κλασικές προσεγγίσεις ανάλυσης συστάδων, όπως ο K-means αλγόριθμος. Η βασική διαφορά είναι ότι κατά τη διάρκεια της διαδικασίας εκπαίδευσης ο αλγόριθμος SOM ενημερώνει το σύνολο των διανυσμάτων αναφοράς των γειτονικών νευρώνων. Παρόλα αυτά ο αλγόριθμος SOM δεν μπορεί να ξεπεράσει τα προβλήματα που εμφανίζει και ο K-means, όπως είναι ο καθορισμός του βέλτιστου αριθμού συστάδων

8.3.4 Αλγόριθμοι βασισμένοι στη πυκνότητα (Density based)

Οι μέθοδοι ομαδοποίησης που βασίζονται στην πυκνότητα διερευνούν την τοποθέτηση των δεδομένων στο χώρο προσπαθώντας να ανιχνεύσουν περιοχές υψηλής πυκνότητας που μπορούν να θεωρηθούν φυσικές κατηγορίες- συστάδες ομοειδών δεδομένων. Το βασικό μειονέκτημα αυτών των μεθόδων εντοπίζεται στο γεγονός ότι μπορούν να χρησιμοποιηθούν μόνο για συστάδες σφαιρικού σχήματος στο χώρο.

Σε αυτές τις μεθόδους η έννοια της πυκνότητας και των ορίων διαχωρισμού των συστάδων παίζει κεντρικό ρόλο. Για ευκολότερους υπολογισμούς η χρήση αποδοτικών δομών δεικτοδότησης (R*-trees), κρίνεται απαραίτητη. Οι βασικές παράμετροι για τις μεθόδους κατηγοριοποίησης που βασίζονται στην πυκνότητα περιλαμβάνουν:

- τη γειτνίαση ενός στοιχείου, δηλαδή το σύνολο των στοιχείων με απόσταση μικρότερη ενός δοσμένου αριθμού,
- την ύπαρξη ενός στοιχείου πυρήνα (core object), δηλαδή ένα στοιχείο με βαθμό γειτονίασης- πλήθος γειτονικών στοιχείων τουλάχιστον MinPts σημείων,
- την απόσταση ενός στοιχείου y από ένα στοιχείο πυρήνα και
- τη συνεκτικότητα δυο στοιχείων y, z , τα οποία μπορούν να προσεγγιστούν από ένα κοινό στοιχείο πυρήνα.

Καθορίζοντας τη συνεκτικότητα όλων των στοιχείων (δεδομένων εισόδου), κατηγοριοποιούμε τα δεδομένα σε συστάδες που χαρακτηρίζονται από το αντίστοιχο στοιχείο πυρήνα. Οι πιο γνωστοί αλγόριθμοι αυτής της κατηγορίας είναι οι: DBSCAN (Ester και άλλοι, 1998), GDBSCAN (Sander και άλλοι, 1998) και FDC (Zhou και άλλοι, 1999)

8.3.5 Αλγόριθμοι βασισμένοι στον Κάνναβο (Grid-based Algorithms)

Πρόκειται για μια σειρά αλγορίθμων οι οποίοι έχουν πρόσφατα παρουσιαστεί και χρησιμοποιούν την κβαντικοποίηση του χώρου σε ένα πεπερασμένο αριθμό κελιών (βλέπε Εικόνα 91).

Εικόνα 91. Κβαντικοποίηση του χώρου σε ένα πεπερασμένο αριθμό κελιών

Κύριο χαρακτηριστικό των αλγορίθμων αυτής της κατηγορίας αποτελεί η εξαιρετικά γρήγορη επεξεργασία των δεδομένων, ενώ ως βασικό μειονέκτημά τους εμφανίζεται η περιορισμένη ακρίβεια και ποιότητα των παραγόμενων συστοιχιών, καθώς το σχήμα τους χαρακτηρίζεται από οριζόντια και κάθετα όρια, σύμφωνα με τον κάνναβο που έχει δημιουργηθεί.

Οι πιο γνωστοί αλγόριθμοι αυτής της κατηγορίας είναι οι: STING (Wang και άλλοι, 1997), WaveCluster (Sheikholeslami και άλλοι, 1998) και CLIQUE (Agrawal και άλλοι, 1998).

8.4 Αναπαράσταση των συστάδων

Το αποτέλεσμα της ομαδοποίησης είναι μια διαμέριση των δεδομένων σε συστάδες. Σε πολλές εφαρμογές και κυρίως σε εκείνες λήψεως αποφάσεων είναι ανάγκαια η ύπαρξη μιας συμπαγής και κατανοητής αναπαράστασης των συστάδων. Η αναπαράσταση μιας συστάδας μπορεί να γίνει με διάφορους τρόπους. Μερικοί από αυτούς είναι και οι παρακάτω:

1. Μια συστάδα μπορεί να αναπαρασταθεί από το κεντροειδές σημείου της ή από έναν αριθμό σημείων που είναι τα πιο απομακρυσμένα της συστάδας. (βλέπε πρώτο σχήμα της Εικόνας 92 και δεύτερο σχήμα της ίδιας Εικόνας αντίστοιχα)
2. Οι κόμβοι ενός δένδρου κατηγοριοποίησης μπορούν να αναπαραστήσουν μια συστάδα. (βλέπε Εικόνα 93b)
3. Επίσης μια συστάδα μπορεί να αναπαρασταθεί χρησιμοποιώντας συζευκτικούς λογικούς τελεστές. (βλέπε Εικόνα 93c)

Εικόνα 92. Αναπαράσταση των συστάδων από σημεία

Ο πρώτος τρόπος αναπαράστασης με την χρήση ενός κεντροειδούς είναι ο πιο συνηθισμένος και έχει καλά αποτελέσματα όταν οι συστάδες είναι συμπαγείς και τα στοιχεία κατανέμονται ομοιόμορφα γύρω από το κεντροειδές. Σε αντίθετη περίπτωση ο τρόπος αυτός αναπαράστασης δεν είναι ο πιο κατάλληλος. Σε αυτή την περίπτωση η αναπαράσταση μιας συστάδας από τα συνοριακά σημεία είναι μια πολύ καλή λύση.

Όσο το σχήμα της συστάδας αλλάζει και παίρνει διάφορα σχήματα στο χώρο, η επιλογή των σημείων πρέπει να είναι τέτοια ώστε να περιγράφεται η συστάδα όσο το δυνατόν καλύτερα. Η αναπαράσταση με ένα δένδρο κατηγοριοποίησης είναι ισοδύναμη με την αναπαράσταση μιας συστάδας με λογικούς συζευκτικούς τελεστές.

Εικόνα 93. (a) συστάδες σε χώρο 2 διαστάσεων, (b) ομαδοποίηση με δένδρο, (c) ομαδοποίηση με λογικούς συζευκτικούς τελεστές.

Μέχρι τώρα έχουμε δει ότι όλες οι τεχνικές και οι αλγόριθμοι ομαδοποίησης τοποθετούν ένα στοιχείο σε μια και μόνο συστάδα. Πρόκειται λοιπόν για σκληρούς (crisp) αλγόριθμους και αυτό συνεπάγεται ότι οι συστάδες σε αυτές τις περιπτώσεις είναι ξένες μεταξύ τους σύνολα. Η ασαφής ομαδοποίηση (fuzzy clustering) επεκτείνει την έννοια του «ένα στοιχείο ανήκει σε μια συστάδα» και συνδέει κάθε στοιχείο με όλες τις συστάδες χρησιμοποιώντας ένα βαθμό

συσχέτισης. Ένας αλγόριθμος ασαφής ομαδοποίησης κάνει τα εξής σε γενικές γραμμές (Sato και άλλοι, 1997):

1. Επιλογή μιας ασαφής διαμέρισης των N στοιχείων σε K clusters. Καθορισμός του πίνακα $U=N \times K$ του οποίου κάθε στοιχείο u_{ij} δηλώνει τον βαθμό συμμετοχής του στοιχείου i στο cluster j . Η τιμές των u είναι μεταξύ 0 και 1.
2. Χρησιμοποιώντας τον πίνακα U βρίσκεται η τιμή κάποιας συνάρτησης που αποτελεί και το κριτήριο τερματισμού, και η οποία πρέπει να βελτιστοποιηθεί. Συνεχώς επανατοποθετεί τα στοιχεία στις συστάδες με νέες τιμές συμμετοχής και επαναπροσδιορίζει τον πίνακα U και την τιμή της συνάρτησης.
3. Επαναλαμβάνει το βήμα 2 μέχρι να μην επέρχονται σημαντικές αλλαγές στον πίνακα U και την τιμή της συνάρτησης.

8.5 Εκτίμηση του αποτελέσματος

Η επιλογή του κατάλληλου αλγορίθμου για ομαδοποίηση δεν είναι απλή υπόθεση. Η πληθώρα αλγορίθμων ομαδοποίησης οι οποίοι υπάρχουν στην βιβλιογραφία συνιστούν δυσκολία στην επιλογή του καλύτερου αλγορίθμου για το εκάστοτε πρόβλημα που αντιμετωπίζεται (Ertoz και άλλοι, 2003). Σε κάθε προηγούμενη ενότητα όταν περιγράψαμε τις κατηγορίες μεθόδων ομαδοποίησης παρουσιάσαμε και τα πλεονεκτήματα και τα μειονεκτήματα της κάθε κατηγορίας μεθόδων.

Τα διαφορετικά αποτελέσματα που προκύπτουν ακολουθώντας διαφορετικές μεθόδους ανάλυσης συστάδων, οφείλονται σε μεγάλο ποσοστό στις διαφοροποιήσεις κάθε μεθόδου (όπως περιγράφηκαν σε προηγούμενες ενότητες) και σε μικρότερο στο είδος των δεδομένων. Οι μεγαλύτερες διαφορές των μεθόδων κατηγοριοποίησης εντοπίζονται: α) στη δομή των συστάδων, β) στην ύπαρξη επικαλύψεων και το αντίστοιχο ποσοστό και γ) στη μετρική ομοιότητας που χρησιμοποιείται.

Η δομή των συστάδων αναφέρεται τόσο στο πλήθος των συστάδων, και στο σχήμα τους, όσο και στη δυναμικότητά τους, δηλαδή το πλήθος των στοιχείων που περιλαμβάνουν σε απόλυτο ή σχετικό μέγεθος (Gordon, 1998). Επίσης η ύπαρξη επικάλυψης ανάμεσα σε συστάδες και το ποσοστό επικάλυψης που επιτρέπουμε επηρεάζουν τις παραμέτρους των μεθόδων ομαδοποίησης. Ανάλογα με το είδος των δεδομένων που επεξεργαζόμαστε επιλέγουμε τόσο τη μέθοδο ομαδοποίησης, όσο και την μετρική ομοιότητας.

Το κυρίως πρόβλημα ενός αλγορίθμου ομαδοποίησης είναι ότι δεν μπορεί για όλες τις περιπτώσεις δεδομένων να εφαρμοστεί και να αναδείξει επιτυχώς την ποικιλία δομών που εμφανίζονται ειδικά σε πολυδιάστατα σύνολα δεδομένων. Για τον λόγο αυτό είναι απαραίτητο για κάθε χρήστη ενός αλγορίθμου ομαδοποίησης να γνωρίζει πολύ καλά την τεχνική που ακολουθεί ο αλγόριθμος της επιλογής του, να έχει γνώση των λεπτομερειών στο πως ομαδοποιούνται τα δεδομένα σε συστάδες και να είναι καλός γνώστης της πληροφορίας που πρόκειται να επεξεργαστεί. Όσο περισσότερη πληροφορία για τα δεδομένα έχει στα χέρια του ο χρήστης τόσο καλύτερα θα εκτιμηθεί η διαδικασία της ομαδοποίησης και σωστά συμπεράσματα θα προκύψουν.

Τώρα τελευταία γίνονται οι πρώτες προσπάθειες συνδυασμού αλγορίθμων ομαδοποίησης σε μια προσπάθεια παραγωγής καλύτερων αποτελεσμάτων (Qian & Suen, 2000), (Strehl & Ghosh, 2002) και (Ayad & Kamel, 2003).

ΚΕΦΑΛΑΙΟ 9. Επίλογος

Τα τελευταία χρόνια με την εξάπλωση της χρήσης των υπολογιστών σε όλους τους τομείς της ζωής μας έχουν αυξηθεί σημαντικά οι δυνατότητές μας να παράγουμε και να συλλέγουμε πληροφορίες, γεγονός που οδήγησε στην συγκέντρωση μεγάλου όγκου πληροφορίας. Η αύξηση αυτή κάνει επιτακτική την ανάγκη εύρεσης νέων τεχνικών και εργαλείων που θα υποστηρίζουν την αυτόματη μετατροπή των υπό επεξεργασία πληροφοριών σε χρήσιμη γνώση. Η Μηχανική Μάθηση και η Εξόρυξη Γνώσης αποτελούν δύο από τους κυριότερους τομείς έρευνας προς αυτή την κατεύθυνση.

Σε αυτή την διδακτορική διατριβή αναλύθηκαν κυρίως μέθοδοι επιβλεπόμενης μηχανικής μάθησης. Στόχος του συγκεκριμένου είδους μάθησης είναι η δημιουργία συστημάτων τα οποία θα μπορούν να βελτιώνουν της απόδοσή τους στην εργασία που επιτελούν εκμεταλλευόμενα αυτόματα προηγούμενη εμπειρία από την εκτέλεση της εργασίας.

Σε γενικές γραμμές, η διαδικασία της επιβλεπόμενης μάθησης αποτελείται από τα εξής στάδια:

- Συλλογή των δεδομένων εκπαίδευσης.
- Καθαρισμός και επεξεργασία των δεδομένων.
- Εξαγωγή των χρήσιμων μεταβλητών που χαρακτηρίζουν το πρόβλημα.
- Επιλογή του κατάλληλου αλγορίθμου μάθησης.
- Παραγωγή του ταξινομητή.
- Χρησιμοποίηση των αποτελεσμάτων για την λήψη αποφάσεων.

Γενικά, ο βαθμός δυσκολίας του προβλήματος εξαρτάται από την μεταβλητότητα των τιμών των χαρακτηριστικών για αντικείμενα της ίδιας κατηγορίας σε σχέση με τη διαφορά μεταξύ των τιμών για αντικείμενα άλλων κατηγοριών. Βέβαια, πρέπει να τονιστεί ότι η μεταβλητότητα των τιμών των χαρακτηριστικών των αντικείμενων της ίδιας κατηγορίας μπορεί να οφείλεται στην πολυπλοκότητα αλλά και στο θόρυβο.

Είναι γνωστό ότι ένα πρόβλημα επιβλεπόμενης μηχανικής μπορεί να λυθεί συγκριτικά καλά χρησιμοποιώντας διάφορες μεθόδους μάθησης, π.χ. διαφορετικούς τύπους τεχνητών νευρωνικών δικτύων, δέντρων απόφασης, Μπεϋζιανών ταξινομητών, κ.λπ. Όμως, μπορεί να συμβεί μια μέθοδος να προβλέπει καλύτερα ορισμένα μέρη του διαστήματος των περιπτώσεων από τις άλλες (Kalousis & Gama, 2004). Κατά συνέπεια, η επιλογή της καταλληλότερης μεθόδου για την τελική λύση είναι ένα περίπλοκο πρόβλημα. Διαφορετικοί αλγόριθμοι είναι κατάλληλοι για τη μάθηση διαφορετικών ειδών συναρτήσεων-στόχων.

Η *επιστημονική συνεισφορά της διδακτορικής διατριβής* μπορεί να συνοψιστεί στα ακόλουθα:

- Πραγματοποιήθηκε μια κριτική μελέτη των πλεονεκτημάτων και των μειονεκτημάτων κάθε κατηγορίας μεθόδων μηχανικής μάθησης ώστε ο χρήστης δεδομένου ενός προβλήματος εφαρμογής να είναι σε θέση να επιλέξει την βέλτιστη τεχνική βάση των αναγκών και των προτεραιοτήτων του. Για κάθε κατηγορία μεθόδων επιτηρούμενης μηχανικής μάθησης, εξετάσαμε τους σημαντικότερους βάσει της βιβλιογραφίας.
- Στην περίπτωση που ενδιαφερόμαστε για την καλύτερη δυνατή ακρίβεια ταξινόμησης, είναι δύσκολο αν όχι αδύνατο να βρεθεί ένας ταξινομητής που να αποδίδει όπως ένας καλός συνδυασμός ταξινομητών. Η βασική ιδέα πίσω από τις ομάδες ταξινομητών είναι πως, δοθείσας μιας εργασίας που απαιτεί γνώσεις ειδικού (expert) για να εκτελεστεί με επιτυχία, πολλοί ειδικοί ίσως είναι αποτελεσματικότεροι από έναν, αν οι ατομικές

αποφάσεις τους συνδυαστούν κατάλληλα. Προς αυτή την κατεύθυνση, προτείναμε ένα πλήθος νέων μεθόδων παραγωγής ομάδων ταξινομητών, όπως:

- a. Συνδυάσαμε την μέθοδο της προώθησης (boosting) με τον απλό Μπεϋζιανό αλγόριθμο (NB). Για να γίνει πιο ασταθής αλλά να μην χάσει ακρίβεια ο NB ώστε να έχει αποτέλεσμα η προώθηση (boosting), σε κάθε βήμα της προώθησης εφαρμόσαμε ένα φίλτρο για την επιλογή των πιο χρήσιμων μεταβλητών. Τα πειράματα απέδειξαν ότι η προτεινόμενη μέθοδος αυξάνει την ακρίβεια πρόγνωσης του NB περισσότερο από κάθε άλλη προσπάθεια βελτίωσής του [2].
 - b. Προτείναμε επιλεκτική ψηφοφορία κατά την οποία οι ταξινομητές που έχουν στατιστικά χειρότερη ακρίβεια πρόγνωσης δεν παίρνουν μέρος στην τελική ψηφοφορία. Τα πειράματα απέδειξαν ότι η επιλεκτική ψηφοφορία δίνει καλύτερα αποτελέσματα από την απλή πλειοψηφική ψηφοφορία αλλά και άλλες πιο περίπλοκες μεθόδους όπως η συσσωρευμένη γενίκευση [6].
 - c. Προτείναμε τοπική ψηφοφορία μιας ομάδας ταξινομητών. Τα πειράματα απέδειξαν ότι η προτεινόμενη μέθοδος δίνει καλύτερα αποτελέσματα επειδή μπορεί να προσεγγίσει πολύ καλύτερα από άλλες μεθόδους τη συνάρτηση-στόχο αν αυτή είναι πολύπλοκη καθολικά, αλλά μπορεί να περιγραφεί ως μια συλλογή λιγότερο σύνθετων τοπικών προσεγγίσεων [4].
- Αύξηση της ακρίβειας ταξινόμησης μπορούμε να πετύχουμε εκτός μέσω της δημιουργίας ομάδας ταξινομητών και από την προσεκτική επιλογή των μεταβλητών του προβλήματος. Οι δύο μέθοδοι που εφαρμόζονται για την επιλογή μεταβλητών είναι αυτή του περιτυλίγματος (wrapper) και αυτή του φίλτρου (filter). Το μειονέκτημα των μεθόδων περιτυλίγματος είναι η μεγάλη υπολογιστική πολυπλοκότητά τους που τις καθιστά απαγορευτικές σε μεγάλα προβλήματα. Η μέθοδος του φίλτρου είναι περισσότερο εφικτή υπολογιστικά αλλά δεν δίνει πάντα καλύτερα αποτελέσματα. Για αυτό το λόγο, προτείναμε υβριδική μεθοδολογία επιλογής μεταβλητών που συνδυάζει τα πλεονεκτήματα και των δύο κατηγοριών [7].
 - Επίσης εξετάσαμε την χρήση των μεθόδων μηχανικής μάθησης σε προβλήματα ειδικής δυσκολίας όπως σε ανομοιογενή δεδομένα (imbalanced datasets), σε προβλήματα πραγματικού χρόνου (online learning) και σε προβλήματα βαθμωτής συνάρτησης στόχου. Προτείναμε μερικούς νέους αλγορίθμους με καλύτερη απόδοση στην αντιμετώπιση αυτών των προβλημάτων [5]. Για παράδειγμα για τα προβλήματα βαθμωτής συνάρτησης στόχου προτείναμε μια τεχνική που χρησιμοποιεί σταθερό αλλά άνισο κόστος για κάθε λάθος ταξινόμησης μεταξύ των κλάσεων. Η τεχνική αυτή μπορεί να αποτελέσει μια εύρωστη λύση στο πρόβλημα επειδή ελαχιστοποιεί τις αποστάσεις μεταξύ των πραγματικών και προβλεφθεισών κλάσεων, χωρίς να βλάπτει την ακρίβεια ταξινόμησης.
 - Τέλος, εξετάσαμε την δυνατότητα χρησιμοποίησης μεθόδων μηχανικής μάθησης για εκπαιδευτικούς σκοπούς, όπως την πρόβλεψη της επίδοσης των φοιτητών στο ανοιχτό πανεπιστήμιο αλλά και άλλων εφαρμογών [1]. Για το σκοπό αυτό κατασκευάσαμε και μερικά εργαλεία υποστήριξης των αποφάσεων τα οποία μπορούν να χρησιμοποιηθούν και σε πλήθος άλλων προβλημάτων [8].

9.1 Επεκτάσεις της παρούσας εργασίας

Οι ομάδες ταξινομητών έχουν δώσει μέχρι στιγμής πολύ ενθαρρυντικά αποτελέσματα και αποτελούν μία από τις πλέον ελπιδοφόρες κατευθύνσεις στη μηχανική μάθηση. Το κύριο μειονέκτημα, ωστόσο, αυτών των αλγορίθμων είναι το υψηλό κόστος τους σε υπολογιστικό χρόνο και μνήμη, καθώς περιλαμβάνουν την εκπαίδευση δεκάδων ταξινομητών. Πιθανές επεκτάσεις της παρούσας διατριβής μπορούν να κινηθούν προς τις ακόλουθες κατευθύνσεις:

- Μία λύση σε αυτό το πρόβλημα είναι η χρήση παραλληλίας, στην περίπτωση που κάθε ταξινομητής μπορεί να εκπαιδευθεί ανεξάρτητα από τους άλλους. Περισσότερη έρευνα χρειάζεται προς αυτήν την κατεύθυνση.
- Παράλληλα, καθώς το πλήθος των δεδομένων αυξάνεται, το ποσοστό αύξησης στην ακρίβεια επιβραδύνεται, διαμορφώνοντας την γνωστή καμπύλη μάθησης (βλέπε ενότητα 4.2.3). Συνεπώς γίνεται σαφές ότι η αναζήτηση του βέλτιστου ποσοστού δειγματοληψίας του συνόλου δεδομένων μπορεί να είναι μια άλλη λύση στο πρόβλημα εξόρυξη γνώσης από τεράστιες βάσεις δεδομένων. Η δειγματοληψία πρέπει να λάβει υπόψη της όχι μόνο τις ιδιότητες (meta-characterizations) του συνόλου των δεδομένων αλλά και τις προτιμήσεις του χρήστη όσον αφορά τα κριτήρια απόδοσης όπως το χρόνο εκπαίδευσης, την ακρίβεια και το μέγεθος του παραγόμενου μοντέλου. Περισσότερη έρευνα χρειάζεται και προς αυτήν την κατεύθυνση.
- Μια ακόμα λύση για την εξόρυξη γνώσης από τεράστιες βάσεις δεδομένων είναι η καταναεμημένη εκτέλεση (grid computing). Εκεί λοιπόν που ένας υπολογιστής θα έπρεπε να πραγματοποιήσει 200 εκατομμύρια μαθηματικούς υπολογισμούς, είναι πολύ λιγότερο χρονοβόρο για 200 υπολογιστές να πραγματοποιήσουν ένα εκατομμύριο μαθηματικούς υπολογισμούς.

Επίσης αντικείμενο έρευνας είναι το κατά πόσο μπορεί να μετατραπεί μια επιτροπή ταξινομητών σε μια πιο συμπαγή και κατανοητή μορφή, αφαιρώντας ίσως τα μέλη με υψηλό βαθμό συσχέτισης ή με μετασχηματισμούς της αναπαράστασής της. Μια πρώτη προσέγγιση προς αυτή την κατεύθυνση δίνεται από τους (Wall και άλλοι, 2003).

Μια άμεση επέκταση της παρούσας διατριβής είναι η χρήση των προτεινόμενων τεχνικών συνδυασμού αλγορίθμων μάθησης σε προβλήματα παλλινδρόμησης. Μια απλουστεύμενη αντιμετώπιση είναι η χρήση μοντέλων παλλινδρόμησης αντί για κατηγοριοποίησης ως αλγορίθμων βάσης. Φυσικά και ο αντίστοιχος συνδυασμός αποφάσεων στην απλουστεύμενη περίπτωση θα πρέπει να γίνει με τη συνάρτηση της μέσης τιμής στη θέση της απλής ψηφοφορίας.

Βέβαια στην περίπτωση της εξόρυξης γνώσης στις βάσεις δεδομένων τα δεδομένα είναι δομημένα και περιγράφονται από μία συγκεκριμένη ενιαία δομή όπου κάθε στιγμιότυπο ενός προβλήματος ορίζεται μέσω ενός συγκεκριμένου και σταθερού συνόλου γνωρισμάτων. Αντιθέτως, στην περίπτωση της εξόρυξης γνώσης από πολυμεσικό υλικό τα δεδομένα είναι ημι-δομημένα ή αδόμητα και δεν μπορούν να περιγραφούν μέσω κάποιου σταθερού συνόλου γνωρισμάτων. Μια πιθανή επέκταση των μεθόδων που προτάθηκαν μπορεί να κινηθεί προς αυτή την κατεύθυνση:

- Για παράδειγμα, ένα κείμενο δεν μπορεί να εισαχθεί χωρίς κάποια προεπεξεργασία σε έναν αλγόριθμο κατασκευής ταξινομητή. Από την στιγμή που περιγράφει όμως από μία συγκεκριμένη ενιαία δομή όπου κάθε στιγμιότυπο ορίζεται μέσω ενός συγκεκριμένου συνόλου χαρακτηριστικών, η διαδικασία εξόρυξη γνώσης από κείμενο εκφυλίζεται σε απλό πρόβλημα μηχανικής μάθησης.
- Παρόμοια μια εικόνα δεν μπορεί να εισαχθεί χωρίς κάποια προεπεξεργασία σαν ένα στιγμιότυπο σε έναν αλγόριθμο κατασκευής ταξινομητή. Το στάδιο της εξαγωγής χαρακτηριστικών από τις εικόνες είναι αποφασιστικής σημασίας, αν όχι το σημαντικότερο. Από την στιγμή που οι εικόνες περιγράφουν από μία συγκεκριμένη ενιαία δομή όπου κάθε στιγμιότυπο ενός προβλήματος ορίζεται μέσω ενός συγκεκριμένου και σταθερού συνόλου γνωρισμάτων, η διαδικασία εξόρυξη γνώσης από εικόνες εκφυλίζεται σε απλό πρόβλημα μηχανικής μάθησης. Περισσότερα στοιχεία και μια εφαρμογή ταξινόμησης εικόνων μπορεί να αναζητήσει ο αναγνώστης στην διπλωματική εργασία του Ανυφαντή (2005) που και αυτή πραγματοποιήθηκε υπό την επίβλεψη του συγγραφέα και του επιβλέποντα καθηγητή του.

- Επίσης, η μουσική διαδραματίζει σημαντικό ρόλο στις εφαρμογές πολυμέσων. Η επίθεση (attack), η διάρκεια, ο όγκος (volume), η ταχύτητα (velocity) είναι πληροφορίες που μπορούν να εξαχθούν από ένα μουσικό κομμάτι. Επιπλέον, είναι δυνατό να συνδεθούν με κάθε κομμάτι υψηλού επιπέδου περιγραφές, όπως το είδος οργάνων και ο αριθμός τους. Από την στιγμή που οι ήχοι περιγράφουν από μία συγκεκριμένη ενιαία δομή όπου κάθε στιγμιότυπο ενός προβλήματος ορίζεται μέσω ενός σταθερού συνόλου γνωρισμάτων, η διαδικασία εξόρυξη γνώσης από ήχο εκφυλίζεται σε απλό πρόβλημα μηχανικής μάθησης.

Ολοκληρώνοντας, η εξόρυξη γνώσης από τον Παγκόσμιο Ιστό αποτελεί μια πιθανή επέκταση των μεθόδων που προτάθηκαν σε αυτή τη διατριβή. Η εξατομίκευση στον Ιστό ορίζεται σαν τη διεργασία κατασκευής πληροφοριακών συστημάτων του Ιστού, τα οποία προσαρμόζονται στις εξειδικευμένες ανάγκες του κάθε χρήστη ξεχωριστά. Ο βασικότερος λόγος της έρευνας και μελέτης πάνω στην εξαγωγή γνώσης από τον Παγκόσμιο Ιστό είναι ο μεγάλος όγκος χρησιμών δεδομένων και πληροφοριών που υπάρχουν στον Ιστό.

Παράλληλα, η εξατομίκευση στον Ιστό μπορεί να οδηγήσει σε προσαρμοστικά Διαδικτυακά Εκπαιδευτικά Συστήματα που θα υποστηρίζουν τον εκπαιδευόμενο στη διάρκεια της μελέτης του, προσδίδοντας στο σύστημα τη δυνατότητα να προσαρμόζεται δυναμικά ανάλογα με τον εκπαιδευόμενο και την εξέλιξή του. Οι προτιμήσεις και ο τρόπος μάθησης των σπουδαστών μπορούν να ποικίλουν πολύ. Το ίδιο υλικό εκμάθησης μπορεί να είναι βαρετό για έναν σπουδαστή και συγχρόνως αρκετά ενδιαφέρον για έναν άλλο. Η χρήση αλγορίθμων μάθησης στα προσαρμοστικά εκπαιδευτικά συστήματα μπορεί να κάνει τα συστήματα αυτά να βελτιώνονται αυτόματα με την εμπειρία. Τα βασικά χαρακτηριστικά του υλικού εκμάθησης και οι συνήθειες μάθησης ενός σπουδαστή μπορούν να αποτελέσουν το σύνολο δεδομένων για έναν εποπτευμένο αλγόριθμο εκμάθησης. Ο αλγόριθμος μάθησης θα δημιουργήσει έπειτα ένα μοντέλο απόφασης για κάθε σπουδαστή που θα εξομοιώνει τις αποφάσεις του σχετικά με τις μελλοντικές μονάδες μάθησης (Webb και άλλοι, 2001). Τελικός στόχος, θα μπορούσε να είναι η ανάπτυξη ενός νοήμονος μηχανισμού αξιολόγησης του στυλ μάθησης του εκπαιδευόμενου, ο οποίος θα αξιοποιεί τα δεδομένα που προκύπτουν από τη χρήση του συστήματος από το χρήστη και θα αναγνωρίζει το στυλ μάθησης του εκπαιδευόμενου με βάση την παρατήρηση και την καταγραφή των επιλογών του σε όλη τη διάρκεια αλληλεπίδρασής του με το σύστημα. Οι αλγόριθμοι μάθησης θα μπορούσαν να διευκολύνουν την εφαρμογή της Προσαρμοστικής Πλοήγησης (Adaptive Navigation or Link-level Adaptation), όπου στόχος είναι η υποστήριξη των εκπαιδευομένων στην πλοήγησή τους μέσα στο πεδίο γνώσης του συστήματος, προσαρμόζοντας κατάλληλα τους εμφανείς συνδέσμους (visible links).

Αναφορές

- Acid, S. and de Campos. L.M. (2003). Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs. *Journal of Artificial Intelligence Research* 18: 445-490.
- Acosta F.M.A. (1995), *Radial Basis Function and Related Models: An Overview*. Signal Processing, Elsevier Science Publishers, 45:37-58.
- Aha, D. (1997). *Lazy Learning*. Dordrecht: Kluwer Academic Publishers.
- Agrawal, R., Srikant, R. (1994), Fast Algorithms for Mining Association Rules, *Proc. of the 20th Int. Conference on Very Large Databases*, Santiago, Chile, Sept, pp. 487-499.
- Agrawal R., Gehrke J., Gunopulos D. and Raghavan P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In *Proc. of the 1998 ACM-SIGMOD Conf. On the Management of Data*, 94-105.
- Allwein, E. L., Schapire, R. E., and Singer, Y. (2000), Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research* 1, 113–141.
- An, A., Cercone N. (2001), Rule Quality Measures for Rule Induction Systems: Description and Evaluation, *Computational Intelligence*, Vol. 17, pp. 409-424.
- Atkeson C. G., Moore A. W., and Schaal S. (1997a), Locally weighted learning for control. *Artificial Intelligence Review*, 11:75–113.
- Atkeson C.G., Moore A.W. & Schaal S. (1997b), Locally weighted learning. *Artificial Intelligence Review*, 11(1-5), 11-73.
- Ayad H. and Kamel M. (2003), Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors. In *Multiple Classifier Systems: Fourth International Workshop, MCS 2003*, Guildford, Surrey, United Kingdom, June 11–13.
- Baik S., Bala J. (2004), A Decision Tree Algorithm for Distributed Data Mining: Towards Network Intrusion Detection, *Lecture Notes in Computer Science*, Volume 3046, Pages 206 - 212
- Banzhaf, W., Nordin P., Keller R. E., & Francone, F. D. (1998). *Genetic programming: An introduction*, San Francisco. CA: Morgan Kaufman Publishers, Inc.
- Batista, G., & Monard, M.C., (2003), An Analysis of Four Missing Data Treatment Methods for Supervised Learning, *Applied Artificial Intelligence*, pp.519-533, vol. 17.
- Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36, 105–139.
- Becker B, Kohavi R., Sommerfield D. (1997), Visualizing the Simple Bayesian Classifier. *The KDD Workshop on Issues in the Integration of Data Mining and Data Visualization*.
- Bell D. and Wang H. (2000), A Formalism for Relevance and its Application in Feature Subset Selection. *Machine Learning* Vol. 41(2) 175–195.
- Berka, P. & Bruha, I. (1998). Discretization and grouping: Preprocessing steps for data mining. In: J.M. Zytkow, M. Quafafow (ed.), *Principles of Data Mining and Knowledge Discovery*, 239-245, *Lecture Notes in AI*, Springer.
- Blake, C.L. & Merz, C.J. (1998), *UCI Repository of machine learning databases*. Irvine, CA: University of California, Department of Information and Computer Science. [<http://www.ics.uci.edu/~mlearn/MLRepository.html>]
- Blanco R., Inza I., Larrañaga P. (2003), Learning Bayesian networks in the space of structures by estimation of distribution algorithms, *International Journal of Intelligent Systems*, Volume 18, Issue 2, Pages 205 – 220.
- Blockeel, H. & De Raedt, L. (1998). Top-down induction of first order logical decision trees. *Artificial Intelligence* 101(1-2): 285-297.
- Blum A. and Langley P. (1997), Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271.

- Bramer M. (2002), Using J-Pruning to Reduce Overfitting of Classification Rules in Noisy Domains, Lecture Notes in Computer Science, Volume 2453, Page 433-442.
- Brazdil P., Soares C. and Da Costa J. (2003), Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results, *Machine Learning*, 50: 251-277.
- Breiman L. (1996): Bagging Predictors. *Machine Learning*, 24(3), 123--140. Kluwer Academic Publishers.
- Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, 26(3), 801–849.
- Breiman L. (2001), Random Forests. *Machine Learning* 45 (1):5-32.
- Breslow, L. A. & Aha, D. W. (1997). Simplifying decision trees: A survey. *Knowledge Engineering Review* 12: 1–40.
- Brighton, H. & Mellish, C. (2002), Advances in Instance Selection for Instance-Based Learning Algorithms. *Data Mining and Knowledge Discovery* 6: 153–172.
- Brodley, C.E. & Friedl, M.A. (1999), Identifying mislabeled training data, *Journal of Artificial Intelligence Research*, vol. 11: 131-167.
- Bruha and F. Franek (1996), Comparison of various routines for unknown attribute value processing: covering paradigm. *International Journal of Pattern Recognition and Artificial Intelligence*, 10, 8: 939-955
- Buhrman, H., Wolf, R. (2002), Complexity measures and decision tree complexity: a survey, *Complexity and logic*, Volume 288 , Issue 1, Pages: 21 – 43.
- Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. 2(2):1-47.
- Camargo, L. S. & Yoneyama, T. (2001). Specification of Training Sets and the Number of Hidden Neurons for Multilayer Perceptrons. *Neural Computation* 13: 2673–2680.
- Caruana, R., and Freitag, D. (1994), Greedy Attribute Selection, *Machine Learning: Proceedings of the Eleventh International Conference*, Morgan Kaufman, San Francisco, CA.
- Castellano, G., Fanelli, A., & Pelillo, M. (1997). An iterative pruning algorithm for feedforward neural networks. *IEEE Transactions on Neural Networks* 8: 519–531.
- Chawla N., Bowyer K., Hall L., Kegelmeyer W. (2002), SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research* 16, 321 - 357.
- Cheng, J. & Greiner, R. (2001). Learning Bayesian Belief Network Classifiers: Algorithms and System, In Stroulia, E. & Matwin, S. (ed.), *AI 2001*, 141-151, LNAI 2056,
- Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* 137: 43–90.
- Cherkauer K. and Shavlik J. (1996), Growing simpler decision trees to facilitate knowledge discovery. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press.
- Cheung Yiu-Ming, k*-Means (2003), A new generalized k-means clustering algorithm, *Pattern Recognition Letters* 24: 2883–2893.
- Chickering, D. (1996). Learning equivalence classes of Bayesian-network structures. In *Proceedings of Twelfth Conference on Uncertainty in Artificial Intelligence*. Portland, OR: Morgan Kaufmann.
- Chien-Yu Chen, Shien-Ching Hwang, and Yen-Jen Oyang (2002) An Incremental Hierarchical Data Clustering Algorithm Based on Gravity Theory gravity theory, *Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002*, Taipei, Taiwan, May 6-8, 2002. Springer-Verlag LNCS 2336.
- Chuanyi J. & Sheng M. (1997), “Combinations of weak classifiers”. *IEEE Trans. Neural Net-works*, 8(1):32–42.
- Cohen W. (1995), Fast Effective Rule Induction. In *Proc. ICML-95*. 115-123.
- Cooper, G.F. and Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, Vol. 9, pp 309-347.
- Corboy, W. G. (1991). Yankee dryers. In B. A. Thorp, Paper machine operations. In: *Pulp and paper manufacture*, vol. 7. GA:Joint Textbook Committee of the Paper Industry.

- Crammer, K. & Singer, Y. (2002). On the Learnability and Design of Output Codes for Multiclass Problems. *Machine Learning* 47: 201–233.
- Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge.
- Daelemans, W., Bosch, A., Weijters T. (1997), IGTTree: Using Trees for Compression and Classification in Lazy Learning Algorithms, *Artificial Intelligence Review*, Volume 11 , Issue 1-5 , Pages: 407 – 423.
- Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch (2003). TiMBL: Tilburg memory based learner, version 5.0, reference guide. ILK Technical Report 03-08, Tilburg University. available from <http://ilk.uvt.nl/downloads/pub/papers/ilk.0308.ps>.
- Dash M. and Liu H. (1997). Feature selection methods for classifications. *Intelligent Data Analysis: An International Journal*, 1(3), pp. 131-156.
- De Mantaras & Armengol E. (1998). Machine learning from examples: Inductive and Lazy methods. *Data & Knowledge Engineering* 25: 99-123.
- De Raedt, L. (1996). *Advances in Inductive Logic Programming*. IOS Press.
- Dejong, K. A., Spears, W. M., & Gordon, D. F. (1993). Using genetic algorithms for concept learning. *Machine Learning* 13: 161-188.
- Demiroz, G. and Guvenir, A. (1997), Classification by voting feature intervals, *ECML-97*, pp.85-92..
- Dietterich T.G. and Bakiri G. (1995), Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of AI Research*, 2, pp. 263-286
- Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, *Machine Learning* 40: 139–157.
- Dietterich, T.G. (2001): Ensemble methods in machine learning. In Kittler, J., Roli, F., eds.: *Multiple Classifier Systems*. LNCS Vol. 1857, Springer (2001) 1–15
- Domingos P. & Pazzani M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, 103-130.
- Domingos P. (1998), How to get a free lunch: A simple cost model for machine learning applications. *Proc. AAAI-98/ICML98, Workshop on the Methodology of Applying Machine Learning*, pp1-7.
- Domingos, P. (1999). MetaCost: A General Method for Making Classifiers Cost-Sensitive. *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, 155-164. ACM Press.
- Dougherty J., Kohavi R. and Shami M. (1995), Supervised and unsupervised discretization of continuous features. In *Proceedings of the twelfth International Conference of Machine Learning*. Morgan kaufmann.
- Duan K., Keerthi S., Poo A. (2003), Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing* 51: 41-59.
- Dutton, D. & Conroy, G. (1996), A review of machine learning, *Knowledge Engineering Review* 12: 341-367.
- Dzeroski S., Zenko B. (2002), Is Combining Classifiers Better than Selecting the Best One. *ICML 2002*: 123-130.
- Dzeroski, S., & Lavrac, N. (2001). *Relational Data Mining*. Springer, Berlin.
- Elomaa T. (1999). The biases of decision tree pruning strategies. In D. Hand, J. Kok & M. Berthold (eds.), *Advances in Intelligent Data Analysis, Proc. 3rd IDA* (pp. 63-74). *Lecture Notes in Computer Science* 1642. Springer, 1999.
- Elomaa, T. & Rousu, J. (1999). General and Efficient Multisplitting of Numerical Attributes. *Machine Learning* 36, 201–244.
- Erp Van, Vuurpijl, L.G., & Schomaker, L.R.B. (2002), An overview and comparison of voting methods for pattern recognition. In *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, pp. 195-200,. Niagara-on-the-Lake, Canada.
- Ertöz L., Steinbach M., and Kumar V. (2003), Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of Second SIAM International Conference on Data Mining*, San Francisco, CA, USA.

- Ester, M., Kriegel, H.-P., Sander, J., and Xu X. (1996), A density-based algorithm for discovering clusters in large spatial data sets with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, pp. 226–231.
- Ester, M., Kriegel, H.-P., Sander, J., Wimmer M. and Xu X. (1998), Incremental Clustering for Mining in a Data Warehousing Environment, Proceedings of the 24th VLDB Conference New York, USA, 1998.
- Fan, W., Stolfo, S.J., Zhang, J. & Chan, P.K. (1999). AdaCost: Misclassification costsensitive boosting. Proceedings of the Sixteenth International Conference on Machine Learning, 97-105. San Francisco: Morgan Kaufmann.
- Fayyad U. and Irani K. (1993), Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, pp. 1022-1027.
- Fern, A., & Givan, R. (2000), Online ensemble learning: An empirical study. In Proceedings of the Seventeenth International Conference on ML, pp. 279–286. Morgan Kaufmann.
- Flach, P.A. & Lavrac, N. (2000). The role of feature construction in inductive rule learning. De Raedt, L. & Kramer, S., (ed.), In Proceedings of the ICML2000 workshop on Attribute-Value Learning and Relational Learning: Bridging the Gap, Stanford University.
- Fox, J. (1997), Applied Regression Analysis, Linear Models, and Related Methods, ISBN: 080394540X, Sage Pubns.
- Frank E., Hall M. and Pfahringer B. (2003), Locally weighted naive Bayes. Proc. of the 19th Conference on Uncertainty in Artificial Intelligence. Acapulco, Mexico. Morgan Kaufmann.
- Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. Machine Learning 32, 63-76.
- Frank, E. and Hall M. (2001), A simple approach to ordinal prediction, L. De Raedt and P. Flach (Eds.): ECML 2001, LNAI 2167, pp. 145-156, 2001, Springer-Verlag Berlin.
- Frank, E. & Witten, I. (1998). Generating Accurate Rule Sets Without Global Optimization. In Shavlik, J., (eds), Machine Learning: Proceedings of the Fifteenth International Conference, Morgan Kaufmann Publishers, San Francisco, CA.
- Freund Y. and Schapire R. E. (1996), Experiments with a New Boosting Algorithm, Proceedings: ICML'96, p. 148-156.
- Freund, Y. & Schapire, R. (1999), Large Margin Classification Using the Perceptron Algorithm, Machine Learning 37: 277–296.
- Friedman J. H. (1997), On bias, variance, 0/1-loss and curse-of-dimensionality. Data Mining and Knowledge Discovery, 1: 55-77.
- Friedman J., Hastie T. and Tibshirani R. (2000) Additive Logistic Regression: a Statistical View of Boosting, The Annals of Statistics, 28: 337-374.
- Friedman, N. & Koller, D. (2003). Being Bayesian About Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks. Machine Learning 50(1): 95-125.
- Friedman, N., Geiger, D. & Goldszmidt M. (1997). Bayesian network classifiers. Machine Learning 29: 131-163.
- Furnkranz, J. (1997). Pruning algorithms for rule learning. Machine Learning 27: 139-171.
- Furnkranz, J. (1999), “Separate-and-Conquer Rule Learning”, Artificial Intelligence Review, Vol. 13, pp. 3-54.
- Furnkranz, J. (2001). Round Robin Rule Learning. In Proceedings of the 18th International Conference on Machine Learning (ICML-01), 146-153.
- Gama J., Iterative Bayes (2000). Intelligent Data Analysis, 6: 463 – 473.
- Gama, J. & Brazdil, P. (1999). Linear Tree. Intelligent Data Analysis 3: 1-22
- Gama, J. & Brazdil, P. (2000). Cascade Generalization. Machine Learning 41: 315–343.
- Gehrke, J., Ramakrishnan, R. & Ganti, V. (1998). Rainforest—A framework for fast decision tree classification of large datasets. In Proceedings of the VLDB Conference, 416-427, New York.

- Genton, M. (2001). Classes of Kernels for Machine Learning: A Statistics Perspective. *Journal of Machine Learning Research* 2: 299-312.
- Giacinto G. and Roli F. (2000), Dynamic Classifier Selection, *Proc. of the First Int. Workshop on Multiple Classifier Systems*, June 21-23, Cagliari, Italy, *Lecture Notes In Computer Science* 1857, Springer, Berlin, pp. 177-189.
- Goebel M., Gruenwald L. (1999), "A Survey Of Data Mining And Knowledge Discovery Software Tools", *SIGKDD Explorations*, Vol. 1, No. 1, P. 20-33, June 1999.
- Gordon, A.D. (1998). How many clusters? An investigation of five procedures for detecting nested cluster structure. In Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H., & Baba, Y. (ed.). *Data Science, Classification, and Related Methods*, Tokyo: Springer-Verlag.
- Guha, S, Rastogi, R., Shim K. (1999), "ROCK: A Robust Clustering Algorithm for Categorical Attributes", In the *Proceedings of the IEEE Conference on Data Engineering*.
- Guha, S., Rastogi, R., Shim K. (1998), "CURE: An Efficient Clustering Algorithm for Large Data sets", Published in the *Proceedings of the ACM SIGMOD Conference*.
- Hall L.O., Bowyer K.W., Kegelmeyer W.P., Moore T.E. and Chao C. (2000), Distributed learning on very large datasets. In *ACM SIGKDD Workshop on Distributed and Parallel Knowledge Discovery*.
- Hall, M. (2000), Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, *Proceedings of the Seventeenth International Conference on Machine Learning*, Pages: 359 – 366.
- Heckerman, D., Mamdani, A., & Wellman, M. (1995). Real-world applications of Bayesian networks. *Communications of the ACM* 38: 24-68.
- Heckerman, D., Meek, C. & Cooper, G. (1999). A Bayesian Approach to Causal Discovery. In Glymour, C. and G. Cooper, (ed.), *Computation, Causation, and Discovery*, 141-165. MIT Press.
- Hernandez, M.A.; Stolfo, S.J. (1998), Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. *Data Mining and Knowledge Discovery* 2(1):9-37.
- Ho, T. K. (1998): The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20: 832-844.
- Holte, R. C. (1993), Very simple classification rules perform well on most commonly used datasets', *Machine Learning* 11, 63-90.
- Hsu C. and Lin C. (2002), A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415-425.
- Huang Z. (1998), Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery* 2, 283–304 (1998).
- Iba W., & Langley P. (1992), Induction of one-level decision trees. *Proc. of the Ninth International Machine Learning Conference*, pp. 233-240, Aberdeen, Scotland: Morgan Kaufmann.
- Jain A.K., Murty M.N., Flynn, P.J. (1999), "Data Clustering: A Review", *ACM Computing Surveys*, Vol.31, No3, 264-323.
- Japkowicz N. (2000), The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence, Las Vegas*.
- Japkowicz N. and Stephen, S. (2002), The Class Imbalance Problem: A Systematic Study *Intelligent Data Analysis*, Volume 6, Number 5.
- Jensen F. (1996). *An Introduction to Bayesian Networks*. Springer.
- Jerzy W. Grzymala-Busse, Ming Hu (2000), A Comparison of Several Approaches to Missing Attribute Values in Data Mining, *Rough Sets and Current Trends in Computing : Second International Conference, RSCTC 2000 Banff, Canada*.
- John C. and Trigg L. (1995), K*: An Instance- based Learner Using an Entropic Distance Measure", *Proc. of the 12th International Conference on ML*, pp. 108-114.
- Jordan, M.I. (1998), *Learning in Graphical Models*. MIT Press, Cambridge, MA.

- Kalousis A., Gama, G. (2004), On Data and Algorithms: Understanding Inductive Performance, *Machine Learning* 54: 275–312.
- Karypis G., Han E. H. and Kumar V. (1999), CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, *Computer* 32(8): 68-75, 1999.
- Keerthi, S. & Gilbert, E. (2002). Convergence of a Generalized SMO Algorithm for SVM Classifier Design. *Machine Learning* 46: 351–360.
- Kleinberg E.M. (2000), A Mathematically Rigorous Foundation for Supervised Learning. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems. First International Workshop, MCS 2000, Cagliari, Italy*, volume 1857 of *Lecture Notes in Computer Science*, pages 67–76. Springer-Verlag, 2000.
- Kohavi R., (1996). Scaling up the accuracy of naive-Bayes classifiers: a decision tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97:273-324.
- Kohonen T. (1997), *Self-Organizing Maps*, Second Extended Edition, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997.
- Kon, M. & Plaskota, L. (2000), Information complexity of neural networks, *Neural Networks* 13: 365–375.
- Kononenko I. (2001), Machine learning for medical diagnosis: history, state of the art and perspective, *Artificial Intelligence in Medicine* 23: 89-109.
- Kontkanen, P., Myllymäki, P., Silander, T. & Tirri, H. (1999). On supervised selection of Bayesian networks. In *Proc. of the 15th International Conference on Uncertainty in Artificial Intelligence*.
- Kramer S., Widmer G., Pfahringer B., and DeGroeve M. (2001), Prediction of ordinal classes using regression trees. *Fundamenta Informaticae*.
- Kubat, M., Holte, R. and Matwin, S. (1998), 'Machine Learning for the Detection of Oil Spills in Radar Images', *Machine Learning*, 30:195-215.
- Kudo M., Sklansky J. (2000), Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition* 33: 25-41.
- Kuncheva, L., Whitaker, C. (2001). Feature Subsets for Classifier Combination: An Enumerative Experiment. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems. Second International Workshop, MCS 2001, Cambridge, UK*, volume 2096 of *Lecture Notes in Computer Science*, pages 228–237. Springer-Verlag.
- Lam L. and Suen C. Y. (1997), Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):553–568.
- Lam, L., *Classifier Combinations: Implementations and Theoretical Issues*, *Lecture Notes in Computer Science*, Volume 1857, Jan 2000, Page 77-86.
- Lam W., Chi-Kin Keung, Danyu Liu (2002), Discovering Useful Concept Prototypes for Classification Based on Filtering and Abstraction. *IEEE Trans. Pattern Anal. Mach. Intell.* 24(8): 1075-1090.
- Landwehr N., Hall M., Frank E. (2003), Logistic Model Trees. *ECML 2003*: 241-252
- Langley P. and Sage S. (1994), Induction of selective Bayesian classifiers. In *Proc. of the 10th Conference on Uncertainty in Artificial Intelligence*, Seattle, 399-406.
- Lavrac N. (1999), Rule Evaluation Measures: A Unifying View, *Lecture Notes in Computer Science*, Volume 1634, Page 174-185.
- Ling, C., & Li, C. (1998). Data Mining for Direct Marketing Problems and Solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)* New York, NY. AAAI Press.
- Littlestone, N. & Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation* 108(2): 212–261.
- Liu, H. & Motoda H. (1998), *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer.
- Liu Z., Huang J., Wang Y., and Chen T. (1998), “Audio Feature Extraction and Analysis for Scene Segmentation and Classification,” *Journal of VLSI Signal Processing* 20, pp.61-79.

- Liu, H. and H. Metoda (2001), *Instance Selection and Constructive Data Mining*, Kluwer, Boston, MA.
- Liu, H., Hussain, F. Tan, C. Dash M. (2002), Discretization: An Enabling Technique, *Data Mining and Knowledge Discovery*, Volume 6 , Issue 4, Pages: 393 – 423.
- Liu, H., and Setiono, R., (1996). Feature selection and classification - a Probabilistic Wrapper Approach. In *Proceedings of the 9th International Conference on Industrial and Engineering Applications of AI and ES*.
- Maclin & Shavlik (1995), Combining the Prediction of multiple classifiers: Using competitive learning to initialize ANNs. In the *Proc of the 14th International Joint Conference on AI*, 524-530.
- Mansuri Y., Kim J. G., Compton P., Sammut C. (1991), An evaluation of Ripple-Down Rules. *Proceedings of the IJCAI'91 Knowledge Acquisition Workshop* Pokolbin.
- Markovitch S. & Rosenstein D. (2002), Feature Generation Using General Construction Functions, *Machine Learning* 49: 59-98.
- Martin B., (1995), "Instance-Based learning : Nearest Neighbor With Generalization", Master Thesis, University of Waikato, Hamilton, New Zealand
- Melville P., and Mooney R. (2003), Constructing Diverse Classifier Ensembles using Artificial Training Examples, *Proceedings of the IJCAI-2003*, pp.505-510, Acapulco, Mexico, August 2003.
- Merz, C. J. (1999), Using Correspondence Analysis to Combine Classifiers, *Machine Learning* 36: 33–58.
- Michalski, R. S., I. Bratko, and M. Kubat. (1998). *Machine Learning and Data Mining: Methods and Applications*. New York:
- Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood: London.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Mossel, E., Umans, C. (2001), On the Complexity of Approximating the VC Dimension, *16th Annual Conference on Computational Complexity (CCC'01)*, June 18 - 21.
- Muggleton, S. (1995). Inverse entailment and Progol. *New Generation Computing, Special issue on Inductive Logic Programming*, 13(3-4):245-286.
- Muggleton, S. (1999). Inductive Logic Programming: Issues, results and the challenge of Learning Language in Logic. *Artificial Intelligence* 114: 283–296.
- Murthy, (1998), Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining and Knowledge Discovery* 2: 345–389.
- Neocleous, C. & Schizas, C., (2002), Artificial Neural Network Learning: A Comparative Review, In Vlahavas, I. & Spyropoulos C. (ed.), *SETN 2002*, 300–313, LNAI 2308, Springer-Verlag Berlin Heidelberg.
- Nadeau, C. and Bengio, Y. (2003), Inference for the generalization error. In *Machine Learning* 52:239–281.
- Opitz D. & Maclin R. (1999), Popular Ensemble Methods: An Empirical Study, *Artificial Intelligence Research*, 11: 169-198, Morgan Kaufmann.
- Parmanto B., Munro P., Doyle H. (1996), Reducing Variance of Committee Prediction with Resampling Techniques. *Connect. Sci.* 8(3): 405-426.
- Pazzani M., Searching for dependencies in Bayesian classifiers (1995). In *Proceedings of the Fifth International Workshop on AI and Statistics*.
- Platt, J. (1999). Using sparseness and analytic QP to speed training of support vector machines. In Kearns, M., Solla, S. & Cohn, D. (ed.), *Advances in neural information processing systems*. MIT Press.
- Potharst R. and Bioch J.C. (2000): Decision trees for ordinal classification. *Intelligent Data Analysis* 4: 97-112.
- Provost, F. & Kolluri, V. (1999), A Survey of Methods for Scaling Up Inductive Algorithms. *Data Mining and Knowledge Discovery* 2: 131–169.
- Provost, F. and Fawcett, T. (2001). Robust Classification for Imprecise Environments, *Machine Learning*, 42, 203–231.
- Pudil, P., Novovicova, J., Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Lett.* 15, 1119–1125.

- Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, Los Altos, CA.
- Qian Y. and Suen C. (2000), Clustering combination method. In *International Conference on Pattern Recognition. ICPR 2000*, volume 2, pages 732–735, Barcelona, Spain, September 2000.
- Quinlan J.R. (1993), *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco
- Quinlan, J. R. (1996). Bagging, boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (pp. 725–730). AAAI/MIT Press.
- Quinlan, J.R. (1995). Induction of Logic Programs: FOIL and Related Systems. *New Generation Computing* 13:287-312.
- Ramoni, M. & Sebastiani, P. (2001). Robust Learning with Missing Data. *Machine Learning* 45: 147–170.
- Rastogi, R. & Shim, K. (2000). PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning. *Data Mining and Knowledge Discovery* 4: 315–344.
- Ratanamahatana C. and Gunopulos D. (2003), Feature Selection for the Naive Bayesian Classifier using Decision Trees, *Applied Artificial Intelligence*, 17: 475–487.
- Reinartz T. (2002), A Unifying View on Instance Selection, *Data Mining and Knowledge Discovery*, 6, 191–210, Kluwer Academic Publishers.
- Ridgeway G., Madigan D. and Richardson T. (1998), Interpretable boosted Naive Bayes classification. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, Menlo Park: 101-104.
- Rifkin P., Klautau, A. (2004), In Defense of One-Vs-All Classification, *Journal of Machine Learning Research* 5: 101-141.
- Roli, F., Giacinto, G., Vernazza, G. (2001). Methods for Designing Multiple Classifier Systems. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems. Second International Workshop, MCS 2001*, Cambridge, UK, volume 2096 of *Lecture Notes in Computer Science*, pages 78–87. Springer-Verlag.
- Ruggieri, S. (2001). Efficient C4.5. *IEEE Transactions on Knowledge and Data Engineering* 14 (2): 438-444.
- Saad, D. (1998). *Online learning in neural networks*. London: Cambridge University Press.
- Salzberg, S. (1997), “On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach”, *Data Mining and Knowledge Discovery*, Vol. 1, pp. 317–328.
- Sander J. O., Martin Ester, Hans-Peter Kriegel, Xiaowei Xu (1998), Density-Based Clustering in Spatial Data sets: The Algorithm GDBSCAN and Its Applications, *Data Mining and Knowledge Discovery* 2, 169–194 (1998), Kluwer Academic Publishers.
- Sato M., Sato Y., Jain L. (1997), *Fuzzy Clustering Models and Applications* (Studies in Fuzziness and Soft Computing Vol. 9), ISBN: 3790810266.
- Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning*, 10, 153-178.
- Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26, 1651–1686.
- Scholkopf, C., Burges, J. C. & Smola, A. J. (1999). *Advances in Kernel Methods*. MIT Press.
- Sebban, M., Zighed, D.A. & Di Palma, S. (1999). Selection and statistical validation of features and prototypes. In *Principles of Data Mining and Knowledge Discovery*, 3rd European Conference. Prague, 184–192.
- Seewald A.K. (2002), How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness, in Sammut C., Hoffmann A. (eds.), *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, Morgan Kaufmann Publishers, pp.554-561.
- Seewald, A. K., Furnkranz, J. (2001), An evaluation of grading classifiers. In *Advances in Intelligent Data Analysis: Proceedings of the Fourth International Symposium IDA (2001)*, pages 221–232, Berlin, Springer.
- Sheikholeslami, C., Chatterjee, S., Zhang, A. (1998), "WaveCluster: A-MultiResolution Clustering Approach for Very Large Spatial Data set". *Proc. of 24th VLDB Conference*.
- Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). Improvements to the SMO algorithms for SVM regression. *IEEE Transactions on Neural Networks* 11: 1188–1193.

- Singh M. and Provan G. (1996), Efficient learning of selective Bayesian network classifiers. In Proceedings of the 13th International Conference on Machine Learning (1996): 453-461, Bari, Italy.
- Srivastava, A., Han, E., Kumar, V., Singh, V. (1999). Parallel Formulations of Decision Tree Classification Algorithms. *Data Mining and Knowledge Discovery* 3: 237-261.
- Strehl and J. Ghosh (2002). Cluster ensembles – a knowledge reuse framework for combining partitionings. In Conference on Artificial Intelligence (AAAI 2002), pages 93–98, Edmonton, July 2002. AAAI/MIT Press.
- Ting K. and Zheng Z. (1999), Improving the Performance of Boosting for Naive Bayesian Classification, N. Zhong and L. Zhou (Eds.): PAKDD'99, LNAI 1574, pp. 296-305.
- Ting, K. & Witten, I. (1999). Issues in Stacked Generalization. *Artificial Intelligence Research* 10: 271-289.
- Tjen-Sien Lim, Wei-Yin Loh, Yu-Shan Shih (2000), A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning*, 40, 203–228, 2000, Kluwer Academic Publishers.
- Todorovski, L., Dzeroski, S. (2003). Combining Classifiers with Meta Decision Trees. *Machine Learning* 50: 223–249.
- Tsekouras, G., Sarimveis, H., Raptis, C., Bafas, G. (2002), A fuzzy logic approach for the classification of product qualitative characteristics, *Computers and Chemical Engineering* 26 (2002) 429–438.
- Tsymbol, S. Puuronen and D. Patterson (2002), Feature Selection for Ensembles of Simple Bayesian Classifiers, In Proceedings of ISMIS: 592-600, Lyon, June 27-29.
- Utgoff, P., Berkman, N., & Clouse, J. (1997), Decision tree induction based on efficient tree restructuring. *Machine Learning*, 29: 5–44.
- Vapnik V.N. (1998), *Statistical Learning Theory*, Wiley, New York.
- Veropoulos, K., Campbell, C. & Cristianini, N. (1999). Controlling the Sensitivity of Support Vector Machines. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI99).
- Villada, R. & Drissi, Y. (2002). A Perspective View and Survey of Meta-Learning. *Artificial Intelligence Review* 18: 77–95.
- Vivarelli, F. & Williams, C. (2001). Comparing Bayesian neural network algorithms for classifying segmented outdoor images. *Neural Networks* 14: 427-437.
- Wall, R., Cunningham, P., Walsh, P., Byrne, S. (2003), Explaining the output of ensembles in medical decision support on a case by case basis, *Artificial Intelligence in Medicine*, Volume 28, Issue 2, Pages 191-206.
- Wang, Y. & Witten, I. H. (1997), Induction of model trees for predicting continuous classes, In Proc. of the Poster Papers of the European Conference on ML, Prague (1997) 128–137. Prague: University of Economics, Faculty of Informatics and Statistics.
- Wang W., Yang J. and Muntz R. (1997), STING: A Statistical Information Grid Approach to Spatial Data Mining, Proceedings of the 23rd VLDB Conference Athens, Greece, 1997.
- Webb G. I. (2000), MultiBoosting: A Technique for Combining Boosting and Wagging, *Machine Learning*, 40, pp. 159–196, Kluwer Academic Publishers.
- Webb G. I., Pazzani M. J. and Billsus D. (2001), *Machine Learning for User Modeling, User Modeling and User-Adapted Interaction*, 11: 19-29, Kluwer Academic Publishers.
- Webb, G. I., J. Boughton, and Z. Wang (2002). Averaged One-Dependence Estimators: Preliminary Results. In Proceedings of the Australasian Data Mining Workshop (ADM 02), pages 65-73. University of Technology, Sydney. Canberra
- Wettschereck, D., Aha, D. and Mohri T. (1997). A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. *Artificial Intelligence Review*, 10:1–37.
- Widmer G. and Kubat M. (1996), Learning in the presence of concept drift and hidden contexts. *Machine Learning* 23, 69–101.
- Wilson, D. R. & Martinez, T. (2000). Reduction Techniques for Instance-Based Learning Algorithms. *Machine Learning* 38: 257–286.

- Witten and E. Frank (2000), *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Mateo.
- Wolpert D. (1992), Stacked Generalization. *Neural Networks*, 5(2):241–260.
- Xenos, M., Pierrakeas C. & Pintelas P. (2002). A survey on student dropout rates and dropout causes concerning the students in the course of informatics of the Hellenic Open University, *Computers & Education* (39): 361–377.
- Yam, J. & Chow, W. (2001). Feedforward Networks Training Speed Enhancement by Optimal Initialization of the Synaptic Coefficients. *IEEE Transactions on Neural Networks* 12: 430-434.
- Yang J, Honavar V. (1998), Feature subset selection using a genetic algorithm. *IEEE Int Systems and their Applications*, 13(2): 44–49.
- Yang, Y. and G. I. Webb (2003). On Why Discretization Works for Naive-Bayes Classifiers. In *Proceedings of the 16th Australian Conference on AI (AI 03)*Lecture Notes AI 2903, pages 440-452. Berlin: Springer. Perth, Australia
- Young, S. & Downs, T. (1998). CARVE—A constructive algorithm for real valued examples. *IEEE Transactions on Neural Networks*. 9(6): 1180–1190.
- Yu, L, Liu, H. (2004), Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research* 5: 1205–1224.
- Zaffalon M. and Hutter M. (2002), Robust Feature Selection using Distributions of Mutual Information, *Proc of the 14th International Conference on Uncertainty in Artificial Intelligence*, pp. 577-584.
- Zaki, M.J., Ho, C., & Agrawal, R. (1999). Scalable parallel classification for data mining on shared memory multiprocessors. In *Proc. of IEEE International Conference on Data Engineering*, 198-205.
- Zhang, T., Ramakrishnan, R., and Linvy, M. (1997), BIRCH: An efficient data clustering method for very large data sets. *Data Mining and Knowledge Discovery*, 1(2): 141–182.
- Zhang, N. (1998). Computational Properties of Two Exact Algorithms for Bayesian Networks. *Applied Intelligence* 9:173–183.
- Zheng Z., and Webb G.I. (2000), Lazy learning of Bayesian rules. *Machine Learning*, 41: 53-84
- Zheng, Z. (1998). Constructing conjunctions using systematic search on decision trees. *Knowledge Based Systems Journal* 10: 421–430.
- Zheng, Z. (2000). Constructing X-of-N Attributes for Decision Tree Learning. *Machine Learning* 40: 35–75.
- Zhou Bo, David W.Cheung, and Ben Kao (1999), A Fast Algorithm for Density-Based Clustering in Large Database, *PAKDD'99, LNAI 1574*, pp. 338-349.
- Zhu, X., Wu, X., Class Noise vs. Attribute Noise (2004), A Quantitative Study of Their Impacts, *Artificial Intelligence Review* 22: 177–210.
- Ανυφαντής, Δ. (2005), Εξόρυξη γνώσης από εικόνες, Μεταπτυχιακή Διπλωματική εργασία, Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

Παράρτημα Α: Δεδομένα

Ο Πίνακας 84, δίνει μια περιγραφή των κατηγορικών δεδομένων που χρησιμοποιήθηκαν στις συγκρίσεις μας. Τα περισσότερα προέρχονται από την UCI αποθήκη (Blake & Merz, 1998).

Δεδομένα	Στιγμιότυπα	Κατηγορικές μεταβλητές	Συνεχείς μεταβλητές	Κλάσεις
Anneal	898	32	6	6
audiology	226	69	0	24
autos	205	10	15	6
badge	294	4	7	2
Balance	625	0	4	3
breast-cancer	286	9	0	2
breast-w	699	0	9	2
Colic	368	15	7	2
Credit-a	690	9	6	2
credit-g	1000	13	7	2
Diabetes	768	0	8	2
dimin	3949	12	0	5
glass	214	0	9	6
haberman	306	0	3	2
heart-c	303	7	6	5
heart-h	294	7	6	5
heart-statlog	270	0	13	2
hepatitis	155	13	6	2
hypothyroid	3772	22	7	4
ionosphere	351	34	0	2
iris	150	0	4	3
kr-vs-kp	3196	35	0	2
labor	57	8	8	2
lymphotherapy	148	15	3	4
monk1	124	6	0	2
monk2	169	6	0	2
monk3	122	6	0	2
primary-tumor	339	17	0	21
segment	2310	0	19	7
sick	3772	22	7	2
sonar	208	0	60	2
soybean	683	35	0	19
student	344	11	0	2
titanic	2201	3	0	2
vechicle	846	0	18	4
vote	435	16	0	2
vowel	990	3	10	11
waveform	5000	0	40	3
wine	178	0	13	3
zoo	101	16	1	7

Πίνακας 84. Τα κατηγορικά δεδομένα που χρησιμοποιήσαμε στις συγκρίσεις

Ο Πίνακας 85, δίνει μια περιγραφή των δεδομένων παλινδρόμησης που χρησιμοποιήθηκαν στις συγκρίσεις μας. Τα περισσότερα προέρχονται από την UCI αποθήκη (Blake & Merz, 1998).

Dataset	Αριθμός μεταβλητών	Στιγμιότυπα
auto93	22	93
autoHorse	25	205
autoMpg	7	398
autoPrice	15	159
basketball	4	96
bodyfat	14	252
breastTumor	9	286
cholesterol	13	303
cleveland	13	303
cloud	6	108
cpu	7	209
echoMonths	9	130
fishcatch	7	158
fruitfly	4	125
Housing	13	506
hungarian	13	294
lowbwt	9	189
pbc	18	418
pharynx	11	195
pwLinear	10	200
sensory	11	576
servo	4	167
strike	6	625
veteran	7	137

Πίνακας 85. Τα δεδομένα παλινδρόμησης που χρησιμοποιήσαμε στις συγκρίσεις

Παράρτημα Β: Το εργαλείο εξόρυξης γνώσης Weka (ένας μικρός οδηγός χρήσης)

Σκοπός αυτού του παραρτήματος είναι να παρουσιαστεί το WEKA, ένα ανοιχτού κώδικα πακέτο λογισμικού για εξόρυξη γνώσης, γραμμένο εξ'ολοκλήρου στη γλώσσα προγραμματισμού Java, και υλοποιημένο στους κόλπους του πανεπιστημίου του Waikato στη Νέα Ζηλανδία (Witten & Frank, 2000). Το WEKA αποτελέσε το βασικό εργαλείο τόσο για πειράματα, όσο και για ενσωμάτωση μηχανισμών μηχανικής μάθησης στις εφαρμογές που κατασκευάσαμε.

Πρόκειται για μια ολοκληρωμένη συλλογή αλγορίθμων μηχανικής μάθησης. Εκτός από τους δημοφιλέστερους αλγορίθμους ταξινόμησης, περιλαμβάνει διάφορους αλγορίθμους παλινδρόμησης, αλγόριθμους μετα-μάθησης (π.χ. σακούλιασμα, συσσωρευμένη γενίκευση κ.α.), μεθόδους ομαδοποίησης (clustering), καθώς και μια πληθώρα εργαλείων για την προεπεξεργασία των δεδομένων, όπως μεθόδους διακριτοποίησης και φιλτραρίσματος μεταβλητών.

Το κάθε σύνολο δεδομένων (data set) πρέπει να είναι της μορφής αρχείου ARFF για να υποστηρίζεται από Weka. Η μορφή του αρχείου ARFF είναι η εξής:

```
@RELATION relation_name
@ATTRIBUTE attr1_name attribute_type
@ATTRIBUTE attr2_name attribute_type
@ATTRIBUTE attr3_name attribute_type
.....
@ATTRIBUTE attrn_name attribute_type

@DATA
attr1_value,attr2_value,attr3_value,...,attrn_value
attr1_value,attr2_value,attr3_value,...,attrn_value
attr1_value,attr2_value,attr3_value,...,attrn_value
.....
attr1_value,attr2_value,attr3_value,...,attrn_value
```

Τρέχοντας το Weka, θα έχει ανοίξει ένα παράθυρο με τέσσερα κουμπιά με ονόματα :

- Simple CLI
- Explorer
- Experimenter
- KnowledgeFlow

Κάθε ένα από αυτά τα κουμπιά αποτελεί συντόμευση για τα αντίστοιχα περιβάλλοντα χρήσης του Weka. Συνήθως χρησιμοποιούμε το περιβάλλον Explorer. Αν όλα πάνε σωστά, Θα εμφανιστεί μπροστά σας το γραφικό περιβάλλον Explorer του weka, το οποίο φαίνεται στην παρακάτω εικόνα.

Εικόνα 94. Εικόνα του Weka Explorer.

Το πρώτο πράγμα που πρέπει να γίνει προκειμένου να χρησιμοποιήσουμε τους αλγορίθμους μηχανικής μάθησης, είναι να τροφοδοτήσουμε το πρόγραμμα με ένα σετ δεδομένων. Αυτό γίνεται από την επιλογή "Open file..." στη στήλη "Preprocess" που φαίνεται στην προηγούμενη εικόνα. Αφού φορτώσουμε ένα σετ δεδομένων τότε ενεργοποιούνται και οι υπόλοιπες λειτουργίες του περιβάλλοντος (Classify, Cluster, Associate, Select Attributes και Visualize).

Χρήση του Weka για αυτόματη ταξινόμηση

Φορτώνουμε το σετ δεδομένων πηγαίνουμε στην επόμενη στήλη του προγράμματος που λέγεται "Classify", όπως φαίνεται στις παρακάτω εικόνες.

Εικόνα 95. Εικόνα του Weka Explorer μετά την επιλογή των δεδομένων.

Εικόνα 96. Εικόνα του Weka Explorer μετά την μετάβαση στην στήλη "Classify".

Τώρα μπορούμε να δούμε πώς λειτουργούν διάφοροι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για το πρόβλημα της ταξινόμησης (classification). Επιλέγουμε τον αλγόριθμο `weka.classifiers.trees.J48` (κουμπί "Choose" πάνω αριστερά) και ορίζουμε για παράδειγμα στις επιλογές του πειράματος (test options) το πεδίο percentage split 66%. Όταν ξεκινήσει το πείραμα (κουμπί start), θα χρησιμοποιηθεί ο αλγόριθμος J48 (υλοποίηση του C4.5 από το Weka) στο 66% του σετ δεδομένων, προκειμένου να κατασκευαστεί το δένδρο απόφασης, το οποίο θα αξιολογηθεί ως προς την ορθότητα των αποφάσεων του στο υπόλοιπο 34% του σετ δεδομένων.

Εικόνα 97. Εικόνα του Weka μετά την εφαρμογή του αλγορίθμου J48 στο σύνολο δεδομένων

Αφού τελειώσει η κατασκευή του δένδρου και η αξιολόγηση του στο δεξί μέρος του περιβάλλοντος weka, εμφανίζονται αρκετές πληροφορίες σχετικά με το πείραμα που μόλις έγινε, καθώς και το ίδιο το δένδρο απόφασης που προέκυψε. Η σημαντικότερη πληροφορία είναι η απόδοση του αλγορίθμου η οποία αναφέρεται ως "Correctly Classified Instances" που εμφανίζεται στην έξοδο αλγορίθμου (Classifier output).

Χρήση του Weka για αυτόματη ομαδοποίηση

Πηγαίνουμε στην στήλη του προγράμματος που λέγεται "Cluster", όπως φαίνεται στην παρακάτω εικόνα.

Εικόνα 98. Άποψη της στήλης "Cluster" του Weka Explorer.

Με την επιλογή "Choose" επιλέγουμε τον αλγόριθμο ομαδοποίησης (clustering) και με αριστερό κλικ ακριβώς δίπλα επιλέγουμε τις παραμέτρους που επηρεάζουν την λειτουργία του κάθε αλγορίθμου. Ορίζουμε το "Cluster mode" σε "use training set", και επιλέξτε τον αλγόριθμο `weka.clusterers.SimpleKMeans` (πλήκτρο "Choose"), όπως φαίνεται στην εικόνα:

Εικόνα 99. Μη επιτηρούμενη εκπαίδευση

Πατώντας το πλήκτρο "Start" και αφού τελειώσει ο αλγόριθμος παρακαλούμε την έξοδο του αλγορίθμου (δεξιά στο παράθυρο), για να δούμε ποιες ομάδες ανίχνευσε ο συγκεκριμένος αλγόριθμος. Μπορεί να χρειαστεί να περιμένουμε μερικά λεπτά μέχρι να τελειώσει κάθε αλγόριθμος.

Αυτόματη αναζήτηση των πιο σημαντικών χαρακτηριστικών

Επίσης ενδιαφέρον είναι να ανιχνεύσουμε ποιες είναι οι πιο σημαντικές μεταβλητές οι οποίες οδηγούν σε σωστή απόφαση. Η ανίχνευση τους μπορεί να οδηγήσει αυτόματα σε μείωση των διαστάσεων των δεδομένων το οποίο με τη σειρά του οδηγεί σε αύξηση της ταχύτητας των αλγορίθμων μάθησης. Πηγαίνουμε στην στήλη "Select attributes" και επιλέξτε έναν "attribute evaluator" και ένα "search method". Μπορούμε επίσης να αφήσουμε τις προεπιλεγμένες τιμές. Αφού βεβαιωθούμε ότι το "attribute selection mode" είναι στην επιλογή "use full training set" πατάμε το κουμπί "start" και παρακαλούμε την έξοδο στα δεξιά. Παρατηρούμε ποιες μεταβλητές ανιχνεύονται ως οι πιο σημαντικές, και εξετάζουμε το σύνολο δεδομένων, προκειμένου να αποφανθούμε αν όντως το αποτέλεσμα είναι (φαινομενικά) σωστό.

Παράρτημα Γ: Άλλα εργαλεία εξόρυξης γνώσης

Στον Πίνακα 86 συγκρίνονται άλλα γνωστά εργαλεία εξόρυξης γνώσεων ως προς το πλήθος των αλγορίθμων που προσφέρουν. Επίσης, στον Πίνακα 87 παρατίθενται οι ηλεκτρονικές διευθύνσεις αυτών των εργαλείων.

Εργαλείο	Δέντρα αποφάσεων	Δίκτυα Μπένζ	Οκνηροί αλγόριθμοι	Μηχανές διανυσμάτων	Νευρωνικά Δίκτυα	Κανόνες ταξινόμησης
AC2	X					
Aira Data Mining Tool						X
ALICE d'ISoft	X					
BusinessMiner	X					
CART	X					
Clementine	X				X	X
Cubist						X
Darwin	X		X		X	X
Data Detective			X			
Data Mining Suite						X
Data SURVEYOR	X					
Data Logic/RDS						X
DataMiner 3D			X			
DataMite			X			X
DecisionWORKS					X	
EWA Systems	X	X	X	X	X	X
IBM Intelligent Miner for Data	X				X	X
IDL Interactive Data Language)					X	
KATE-DataMining	X					
KnowledgeSEEKER	X					
KnowledgeSTUDIO	X				X	
Nuggets						X
Oracle Data Mining (ODM)		X		X		
PolyAnalyst			X		X	X
PrudSys Discoverer	X					
S-PLUS	X					
SAS Enterprise Miner	X				X	
Scenario	X					
See 5/C5.0	X					X
Syllogic Data Mining	X					X
The Easy Reasoner (TER)			X			
VisualMine						X
XpertRule Miner	X					X

Πίνακας 86. Άλλα εργαλεία μάθησης

Εργαλείο	Διεύθυνση στο διαδίκτυο
AC2	http://www.alice-soft.com/products/ac2.html
Aira Data Mining Tool	http://www.hycones.com.br/
ALICE d'ISoft	http://www.alice-soft.com/products/alicev50.html
BusinessMiner	http://www.businessobjects.com/products/advanced_analysis_bminer.htm
CART	http://www.salford-systems.com/products.html
Clementine	http://www.spss.com/software/clementine
Cubist	http://www.rulequest.com/cubist-info.html
Darwin	http://www.think.com/html/products/dartechdatasht.htm#one
Data Detective	http://www.smr.nl/
Data Mining Suite	http://www.datamining.com/dmsuite.htm
Data SURVEYOR	http://www.ddi.nl/products/main.html
Data Logic/RDS	http://www.reduct.com/about_the_company/products.htm
DataMiner 3D	http://www.dimension5.sk/products/products.htm
DataMite	http://www.lpa.co.uk/dtm.htm
DecisionWORKS	http://www.asacorp.com/product/index.html
IBM Intelligent Miner for Data	http://www.software.ibm.com/data/iminer/fordata/index.html
IDL Interactive Data Language)	http://www.rsinc.com/idl/idl_dataminer.cfm
KATE-DataMining	http://www.acknosoft.com/fTools.html
KnowledgeSEEKER	http://www.angoss.com/ksprod/kspage.htm
KnowledgeSTUDIO	http://www.angoss.com/kstudio/studio.htm
Nuggets	http://WWW.DATA-MINE.COM/Products.htm
Oracle Data Mining (ODM)	http://www.oracle.com/technology/products/bi/odm/index.html
PolyAnalyst	http://www.megaputer.com/pasystem.html
PrudSys Discoverer	http://www.prudsys.com/discoverer
S-PLUS	http://www.mathsoft.com/splus/
SAS Enterprise Miner	http://www.sas.com/products/miner/index.html
Scenario	http://www.cognos.com/scenario/
See 5/C5.0	http://www.rulequest.com/see5-info.html
Syllogic Data Mining	http://www.perotsystems.nl/
The Easy Reasoner (TER)	http://www.haley.com/0072567834543122/TheEasyReasoner.html
VisualMine	http://www.visualmine.com/Datasheet/datasheet.htm
XpertRule Miner	http://www.attar.com/

Πίνακας 87. Διευθύνσεις στο διαδίκτυο των διαθέσιμων εργαλείων εξόρυξης γνώσης

Πρέπει να αναφερθεί ότι πολλά στοιχεία για τα εργαλεία αυτά πήραμε και από την εργασία (Goebel & Gruenwald, 1999).

Παράρτημα Δ: Αντιστοίχιση ελληνικών και αγγλικών όρων

Α

Αμοιβαία Εξάρτηση	Correlation
Ανακάλυψη Γνώσης	Knowledge Discovery
Αποθήκες δεδομένων	Data warehouses
Ασαφής κανόνες	Fuzzy rules

Γ

Γενετικοί αλγόριθμοι	Genetic algorithms
Γενίκευση	Generalization

Δ

Δειγματοληψία	Sampling
Δέντρα αποφάσεων	Decision trees
Δεξαμενές πληροφορίας	Information repositories
Διαμέρισμα	Partition
Διάνυσμα Χαρακτηριστικών	Feature Vector
Διαχωριστικοί αλγόριθμοι	Divide algorithms

Ε

Εκ των προτέρων πιθανότητα	A priori probability
Εμπιστοσύνη	Confidence
Εξαγωγή κανόνων	Rule induction
Εξόρυξη γνώσης	Data mining
Εξόρυξη γνώσης από κείμενα	Text mining
Εξόρυξη γνώσης από εικόνες	Image mining
Εξόρυξη γνώσης από βίντεο	Video mining
Επιλογή Δεδομένων	Data Selection

Ι

Ιεραρχικό Μοντέλο Δεδομένων	Hierarchical Data Model
-----------------------------	-------------------------

Κ

Καθαρισμός Δεδομένων	Data Cleaning
----------------------	---------------

Κανόνες συσχέτισης	Association rules
Κανονικοποίηση	Normalization
Κατηγοριοποίηση	classification
Κλάση	Class
Κλιμάκωση	Scalability
Κόμβος Γονέας	Parent Node
M	
Μετρική	Metric
Μέτρο ομοιότητας	Similarity measure
Μοντελοποίηση εξαρτήσεων	Dependency modeling
Μεταβλητή	Variable, attribute
Π	
Παλινδρόμηση	Regression
Πολυωνυμικός Χρόνος	Polynomial Time
Πρόβλεψη	Prediction
Πρότυπο σχέδιο	pattern
Σ	
Συγκροτήσεις	clustering
Συνέπεια	Consistency
Σύστημα Υποστήριξης Αποφάσεων	Decision Support System
T	
Ταξινόμηση	Classification
Τερματικός Κόμβος (φύλλο)	Leaf Node
Τεχνικά νευρωνικά δίκτυα	Artificial neural networks
Τμηματοποίηση	partitioning
Υ	
Υποσύνολο	Subset
X	
Χωρικές Βάσεις Δεδομένων	Spatial Databases