

ΘΕΜΑ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ: Διαχείριση Λογαριασμών Πελατών

Σύνολο δεδομένων

- Δεδομένα εκπαίδευσης(training set)

2528 υποδείγματα

39 χαρακτηριστικά(δυναμικά, ακέραια, πραγματικά)& ζητούμενο(record label, good / bad)

- Δεδομένα επαλήθευσης(quiz set)

1265 υποδείγματα

39 χαρακτηριστικά

- Δεδομένα εξέτασης(test set)

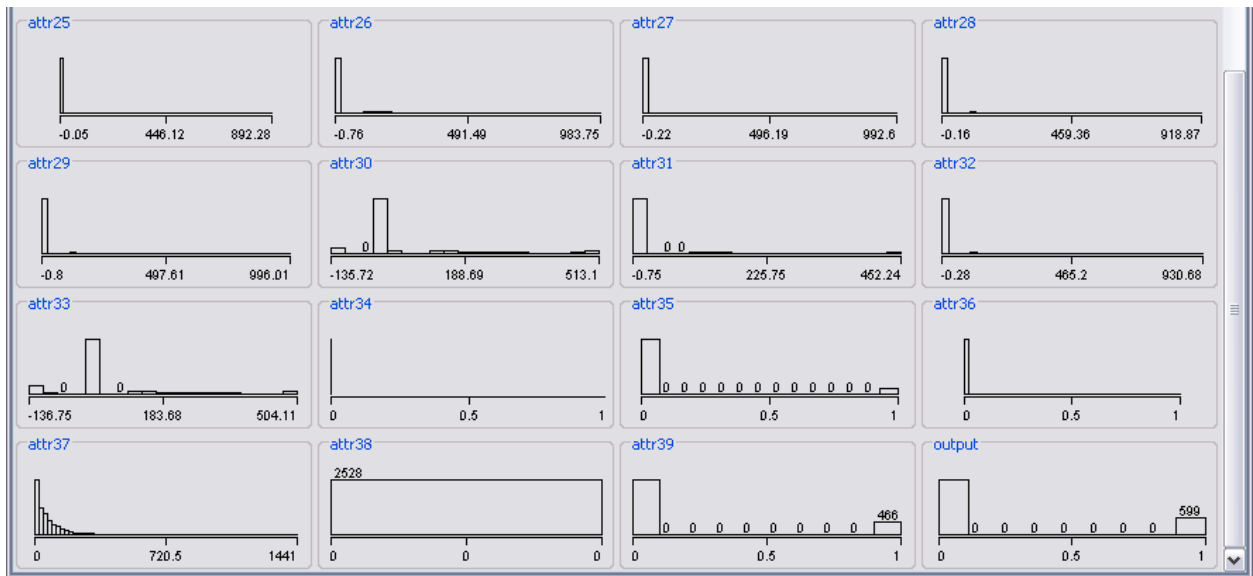
1265 υποδείγματα

39 χαρακτηριστικά

Οπτική διερεύνηση των δεδομένων

Σε πρώτο στάδιο του θέματος ερευνάται οπτικά το σύνολο δεδομένων





Από την διερεύνηση των δεδομένων φαίνεται αμέσως ότι το χαρακτηριστικό attr38 έχει μια μοναδική τιμή για κάθε υπόδειγμα, οπότε δεν πρόκειται να προσφέρει κάτι στην ανάλυσή μας και για αυτό τον λόγο εξαιρείται όπως και το χαρακτηριστικό attr34 που εμφανίζει μόνο μια μη μηδενική τιμή σε ένα υπόδειγμα ενώ σε όλα τα υπόλοιπα έχει μηδενική. Επίσης, χρήσιμη ήταν και η μετατροπή της κλάσης σε nominal (μέσω του φίλτρου Discretize στα φίλτρα χωρίς επίβλεψη) ώστε να γίνει πιο ευκρινής η κατανομή της πρακτικά δυαδικής τάξης output σε 0 και 1 στην κατανομή των χαρακτηριστικών στα instances.

Επιλογή Χαρακτηριστικών

Από το tab Select Attributes επιλέγονται σε πρώτη φάση για την αξιολόγηση των χαρακτηριστικών οι αλγόριθμοι CfsSubsetEval (μέθοδος διήθησης), ReliefAttributeEval (αποτίμηση μοναδιαίου χαρακτηριστικού), ClassifierSubsetEval (μέθοδος ενσωμάτωσης) σε συνδυασμό με τις εξής μεθόδους αναζήτησης BestFirst, GeneticSearch, GreedyStepwise. Στη συνέχεια για την εφαρμογή των ακόλουθων μεθόδων ταξινόμησης με τη βοήθεια του φίλτρου Discretize γίνεται μετατροπή της κλάσης output από numeric σε nominal:

ChiSquaredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, SVMAttributeEval (αποτίμηση μοναδιαίου χαρακτηριστικού) σε συνδυασμό με τη μέθοδο αναζήτησης Ranker, καθώς και η ConsistencySubsetEval (μέθοδος διήθησης) σε συνδυασμό με τις μεθόδους αναζήτησης BestFirst, GeneticSearch, GreedyStepwise. Μετά από εφαρμογή των παραπάνω αλγορίθμων έγινε επιλογή των εξής χαρακτηριστικών: 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 32, 33, 35, 39

Εξαγωγή Συμπερασμάτων- Προβλέψεων

A. Προκειμένου να αποφευχθεί κατά το δυνατό περισσότερο η υπερπροσορμογή του μοντέλου και των προβλέψεων στα δεδομένα εκπαίδευσης χρησιμοποιούνται δεδομένα επαλήθευσης (quiz set).

B. Εφαρμογή απλών αλγορίθμων

Αλγόριθμοι βασισμένοι σε δένδρα απόφασης:

J48, Decision Stump, REPTree

Αλγόριθμοι βασισμένοι σε κανόνες ταξινόμησης που έχουν προκύψει με κατάλληλη προσαρμογή από δένδρα απόφασης:

M5Rules, 1R

Για τους J48 και 1R η κλάση διακριτοποιήθηκε σε nominal

Από τους παραπάνω αλγορίθμους επιλέγονται οι M5Rules, J48 που έχει με το μικρότερο σφάλμα και τη καλύτερη επίδοση για το επόμενο βήμα, οι οποίοι στηρίζονται και οι δύο στη μέθοδο C4.5 με μετά-κλάδεμα, παρόλο που ο M5Rules υπάγεται στους κανόνες ταξινόμησης που έχουν προκύψει από δένδρα απόφασης. Τη καλύτερη επίδοση είχε βέβαια ο M5Rules 0,938 και ακόλουθα σφάλματα

Correlation coefficient	0
Mean absolute error	0.2719
Root mean squared error	0.4618
Relative absolute error	114.7704 %
Root relative squared error	194.9128 %
Total Number of Instances	1265

Γ. Συνδυασμός απλών αλγορίθμων με μετά-αλγόριθμους και Αποτελέσματα

Υιοθέτηση του Μετά-αλγορίθμου RegressionByDiscretization: Ένα σχήμα παλινδρόμησης που μπορεί να συνεργαστεί με οποιοδήποτε αλγόριθμο ταξινόμησης το οποίο εφαρμόζει διακριτοποίηση στη τάξη. Στη συνέχεια αυτός ο αλγόριθμος συνεργάζεται με τον αλγόριθμο Ενδυνάμωσης AdaBoostM1 που βελτιώνει αισθητά την απόδοση ενός οποιουδήποτε απλού αλγορίθμου βάσης και στη συγκεκριμένη περίπτωση του J48. Η απόδοση του συγκεκριμένου αλγορίθμου είναι: 0,956 και τα σχετικά σφάλματα:

Correlation coefficient	0
Mean absolute error	0.2489
Root mean squared error	0.4955
Relative absolute error	105.0617 %
Root relative squared error	209.1196 %
Total Number of Instances	1265

Για την απομάκρυνση περισσότερου θορύβου υιοθετείται και ένα σχήμα εμφωλίαςσης Bagging, το οποίο αποδίδει ικανοποιητικά όταν ο αλγόριθμος βάσης στηρίζεται σε δένδρα απόφασης.

Τα αποτελέσματα είναι ένα ικανοποιητικό σκορ:0,957 και τα παρακάτω σφάλματα

Correlation coefficient	0
Mean absolute error	0.2491
Root mean squared error	0.4826
Relative Absolute error	105.1472 %
Root relative squared error	203.6782 %
Total Number of Instances	1265

Τέλος γίνεται υποβολή αποτελεσμάτων που έχουν προκύψει με supplied test to test set του θέματος με τα εξής σφάλματα

Mean absolute error	0.2278
Root mean squared error	0.4584
Relative absolute error	96.1335 %
Root relative squared error	193.4659%
Total Number of Instances	1265

Γίνεται αντιληπτό, ότι οι απλοί αλγόριθμοι μπορούν να επιφέρουν μεγάλη απόδοση, η οποία μπορεί να βελτιωθεί περαιτέρω με σχήματα εμφωλίας, παλινδρόμησης και ενδυνάμωσης.