



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΙΣ
ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ (Data Mining)

Πανδής Αθηνά

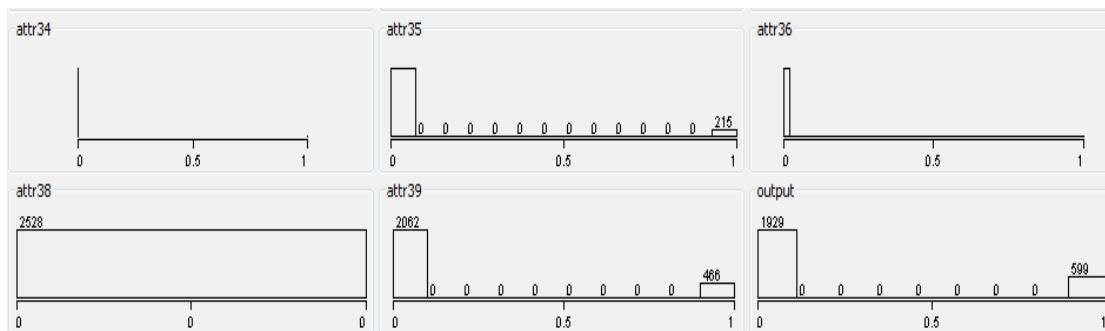
Μάιος 2008

Τα δεδομένα που έχουμε προς επεξεργασία χωρίζονται σε τρία μέρη:

1. Τα δεδομένα εκπαίδευσης (training set) που αποτελούνται από 2528 υποδείγματα και 39 χαρακτηριστικά
2. Τα δεδομένα επαλήθευσης (quiz set) που αποτελούνται από 1265 υποδείγματα και 39 χαρακτηριστικά
3. Τα δεδομένα εξέτασης (test set) που αποτελούνται από 1265 υποδείγματα και 39 χαρακτηριστικά

Πριν όμως από οποιαδήποτε εφαρμογή αλγορίθμων στο train set, κρίνεται απαραίτητη μια προεπεξεργασία και προεπισκόπηση των δεδομένων.

Ανοίγοντας το train set με το λογισμικό Weka επιλέγουμε την εντολή visualize all και παίρνουμε μια πρώτη εικόνα για όλα τα attributes. Όπως βλέπουμε και στο παρακάτω σχήμα η attribute 38 έχει όλα τα instances 0, η 36 έχει 2513 μηδενικά instances ενώ η attribute 34 έχει μόνο ένα μη μηδενικό instance. Επομένως μπορούμε να αφαιρέσουμε αυτές τις τρεις μεταβλητές από το train set χρησιμοποιώντας την εντολή **Remove**.



Στη συνέχεια προχωράμε στην δημιουργία νέου train set χωρίς τις μεταβλητές 34, 36 και 38 το οποίο ονομάζουμε train 11.

Στο train 11 θα εφαρμόσουμε διάφορους αλγόριθμους και θα επιλέξουμε εκείνους με το μικρότερο σφάλμα και υψηλότερο συντελεστή συσχέτισης. Στην περίπτωση μας αρκετοί αλγόριθμοι δεν τρέχουν διότι το class attribute είναι numeric και όχι nominal. Μπορούμε βέβαια να μετατρέψουμε τον τύπο της μεταβλητής στόχου με την εντολή weka filters -> unsupervised-> attribute -> NumericToBinary. Όμως δεν ενδείκνυται να αλλάζουμε την class attribute διότι χάνεται πληροφορία. Συνεπώς δοκιμάζουμε

τους αλγόριθμους που είναι συμβατοί με numeric class. Να σημειώσουμε επίσης ότι στα δεδομένα μας δεν μπορούμε να κάνουμε ομαδοποίηση (cluster) διότι η μεταβλητή στόχος είναι numeric.

Παρακάτω παρουσιάζονται οι αλγόριθμοι που εφαρμόστηκαν στο train 11 και έδωσαν χαμηλό σφάλμα και υψηλό συντελεστή συσχέτισης.

1. Weka.classifiers->RegressionByDiscretization->Bagging->Bagging->PART

Relation: train-weka.filters.unsupervised.attribute.Remove-R34,36,38

Instances: 2528

Time taken to build model: 76.67 seconds

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9251
Mean absolute error	0.0572
Root mean squared error	0.1616
Relative absolute error	15.8132 %
Root relative squared error	37.9856 %

Total Number of Instances 2528

2. Weka.classifiers->RegressionByDiscretization->AdaBoostM1->PART

Relation: train-weka.filters.unsupervised.attribute.Remove-R34,36,38

Instances: 2528

Attributes: 37

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9148
-------------------------	--------

Mean absolute error	0.0326
Root mean squared error	0.1746
Relative absolute error	9.0248 %
Root relative squared error	41.049 %
Total Number of Instances	2528

3. Weka.classifiers->RegressionByDiscretization->AdaBoostM1->Bagging->PART

Correlation coefficient	0.912
Mean absolute error	0.032
Root mean squared error	0.1777
Relative absolute error	8.8583 %
Root relative squared error	41.7661 %
Total Number of Instances	2528

4. Weka.classifiers->RegressionByDiscretization->Bagging->J48

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9123
Mean absolute error	0.0615
Root mean squared error	0.1742
Relative absolute error	17.0029 %
Root relative squared error	40.9434 %
Total Number of Instances	2528

5. Weka.classifiers->RandomSubSpace->RegressionByDiscretization->J48Graft

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9198
-------------------------	--------

Mean absolute error	0.0685
Root mean squared error	0.1678
Relative absolute error	18.9252 %
Root relative squared error	39.4458 %
Total Number of Instances	2528

Στη συνέχεια δημιουργούμε ένα νέο quiz set αφαιρώντας τις μεταβλητές 34, 36 και 38 όπως ακριβώς είχαμε κάνει στο train 11. Ονομάζουμε το νέο αρχείο quiz11.

Τρέχουμε τον αλγόριθμο που έχουμε επιλέξει με cross validation (αφήνουμε την default επιλογή των 10 folds). Όταν ολοκληρωθεί η διαδικασία επιλέγουμε supplied test set και ορίζουμε σαν test set να είναι το quiz11 που έχουμε δημιουργήσει. Τρέχουμε ξανά τον αλγόριθμο και ζητάμε από το weka να μας εμφανίσει τις προβλέψεις (predictions). Επαναλαμβάνουμε αυτήν την διαδικασία και για τους πέντε αλγορίθμους.

Τα success rates που λάβαμε για κάθε αλγόριθμο είναι τα ακόλουθα:

- **Αλγόριθμος 1: 0,955**
- **Αλγόριθμος 2: 0,941**
- **Αλγόριθμος 3: 0,955**
- **Αλγόριθμος 4: 0,953**
- **Αλγόριθμος 5: 0,955**

Για την τελική υποβολή στο test set επιλέγουμε τον Αλγόριθμο 5 με correlation coefficient 0,9198 και success rate 0,955. Να σημειώσουμε σε αυτό το σημείο ότι δεν επιλέξαμε τον Αλγόριθμο 1 παρόλο που μας έδωσε στο train set μεγαλύτερο συντελεστή συσχέτισης (0,9251), διότι η χρησιμοποίηση τριών meta αλγορίθμων (RegressionByDiscretization, Bagging (2)) ενδέχεται να προκαλέσει overtraining στα δεδομένα.

Τέλος πρέπει να αναφερθεί ότι δεν εφαρμόστηκε attribute selection (preprocess->choose filter) στο train set διότι είχαμε αρκετά χαμηλότερο success rate στο quiz set. Για παράδειγμα έχοντας κρατήσει τις μεταβλητές που μας υπόδειξε το φίλτρο

AttributeSelection (6, 7, 12, 13, 23, 24, 27, 28, 32, 35) ο αλγόριθμος 5 έδωσε success rate 0,902 έναντι 0,955 που είχαμε χωρίς τα attribute 34, 36, 38.