



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ



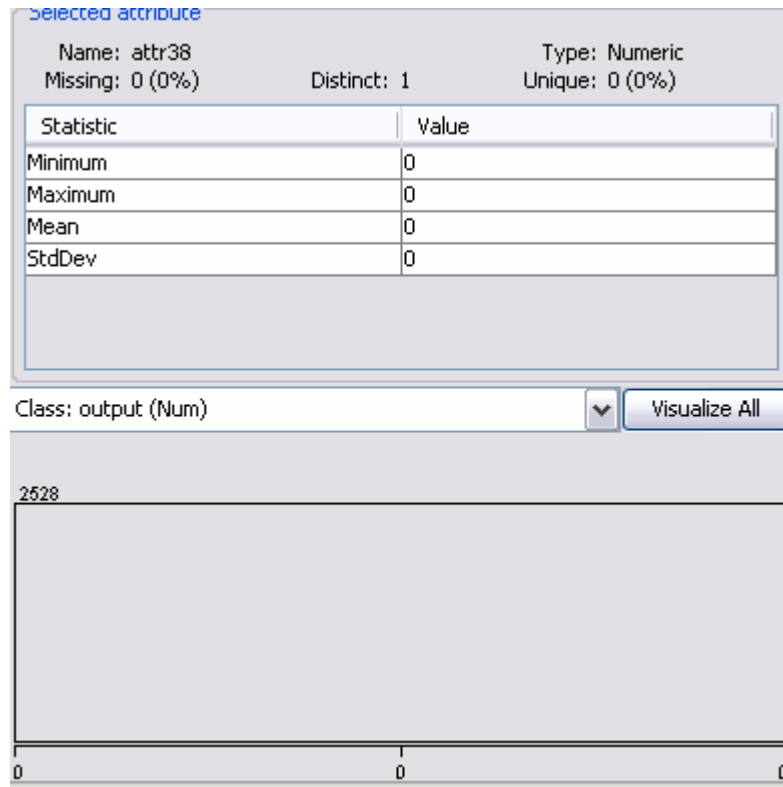
**ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ
ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

**ΔΠΜΣ : ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ
ΡΟΗ ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ**

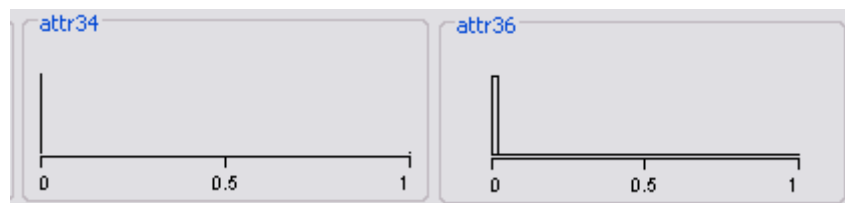
ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΥΥΕΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΕΡΓΑΣΙΑ 08: ΕΙΡΗΝΗ ΛΥΓΚΩΝΗ

1^ο ΣΤΑΔΙΟ: Πριν εφαρμόσουμε οποιοδήποτε αλγόριθμο στο training set , προεπεξεργαζόμαστε πρώτα τα δεδομένα. Παρατηρούμε ότι το training set αποτελείται από 39 numeric attributes των 2528 instances. Επειδή στην attribute38 όλα τα instances είναι 0, την αφαιρούμε από το training set με τη εντολή **Remove**.



Όμως παρόμοια διαπιστώνουμε ότι οι attribute36 και η attribute34 έχουν 15 και 1 αντίστοιχα διαφορετικά του μηδενός instances. Άρα θεωρούμε σκόπιμο να τις αφαιρέσουμε.



Για να επαληθεύσουμε αυτή την κίνηση που κάναμε, τρέξαμε κάποιους αλγόριθμους και παρατηρούμε ότι τα σφάλματα (Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error) είναι μικρότερα αν έχουμε αφαιρέσει τις attributes34,36,38 απ'ότι μόνο την attribute38. Έτσι αποθηκεύουμε το

καινούριο training set με το όνομα train343638.arff. Επίσης υλοποιήσαμε και τον αλγόριθμο attribute selection και παρατηρούμε ότι τα attributes μειώνονται από 39 σε 11 και συγκεκριμένα είναι τα εξής:

Attributes: 11
attr06
attr07
attr12
attr13
attr23
attr24
attr27
attr28
attr32
attr35
output

Όμως, υλοποιώντας οποιοδήποτε αλγόριθμο έχοντας σαν training set αυτό με τα 10 attributes παρατηρούμε ότι τα σφάλματα είναι μεγαλύτερα σε σχέση με το train343638.arff οπότε κρατάμε το δεύτερο.

Επιπλέον μπορούμε σε ένα training set γενικά μπορούμε να χρησιμοποιήσουμε αρκετούς κανόνες ταξινόμησης (classification rules), κανόνες συσχέτισης (association rules) και κανόνες ομαδοποίησης. Επειδή όμως το δικό μας training set όλες οι attributes είναι αριθμητικές (numeric) δεν μπορούμε να εφαρμόσουμε όλους αυτούς τους κανόνες. Υπάρχει βέβαια η δυνατότητα αλλαγής της κλάσης των χαρακτηριστικών από numeric -> nominal πράγμα που δεν το εφαρμόζουμε γιατί μπορεί να μεγαλώσει το σφάλμα.

2° ΣΤΑΔΙΟ: είναι η επανάληψη όλων των μεθόδων που ακολουθήσαμε στο quiz set και συγκεκριμένα η αφαίρεση των attributes 34,36,38 και η αποθήκευση του καινούριου quiz set με το όνομα quiz343638.arff .

3^ο ΣΤΑΔΙΟ: είναι η επιλογή του κατάλληλου αλγορίθμου με το μικρότερο σφάλμα για training set και η αντίστοιχη εφαρμογή σαν supply test στο quiz set. Δοκιμάζοντας αρκετούς αλγόριθμους επιλέξαμε εκείνον που έδινε το μεγαλύτερο ποσοστό επιτυχίας και τον τρέξαμε για το test set έτσι ώστε να πετύχουμε το βέλτιστο ποσοστό επιτυχίας.

Παρακάτω ακολουθούν κάποιοι από τους πολλούς αλγόριθμους που εφαρμόστηκαν στο training set με χαμηλό σφάλμα και μεγάλο ποσοστό επιτυχίας στο quiz set.

1. Bagging->RegressionByDiscretization->Reptree

==== 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

2. RandomSubspace->Bagging-> RegressionByDiscretization-> Reptree

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9219
Mean absolute error	0.0758
Root mean squared error	0.1672
Relative absolute error	20.9631 %
Root relative squared error	39.2965 %
Total Number of Instances	2528

3. Bagging ->RandomSubspace->Reptree

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9159
Mean absolute error	0.0914
Root mean squared error	0.1759
Relative absolute error	25.2535 %
Root relative squared error	41.3445 %
Total Number of Instances	2528

4. RegressionByDiscretization->J48

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.8852
Mean absolute error	0.0494
Root mean squared error	0.2012
Relative absolute error	13.6587 %
Root relative squared error	47.3025 %
Total Number of Instances	2528

5. Bagging->RepTree

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9018
Mean absolute error	0.0778
Root mean squared error	0.1843
Relative absolute error	21.5171 %
Root relative squared error	43.3226 %
Total Number of Instances	2528

6. Random subspace-> RegressionByDiscretization->J48

==== Cross-validation ====

==== Summary ====

Correlation coefficient	0.9154
Mean absolute error	0.0721
Root mean squared error	0.1722
Relative absolute error	19.9263 %
Root relative squared error	40.4707 %
Total Number of Instances	2528

7. **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

4^ο ΣΤΑΔΙΟ: Επιλέγουμε τον αλγόριθμο με Correlation coefficient κοντά στο 1 αρκετά και ποσοστό επιτυχίας για το quiz set 0.955 που είναι ο **RandomSubspace->RegressionByDiscretization->J48graft** και τον εφαρμόζουμε στο test set. Αν δεν υπάρξει overtraining στο training set εξαιτίας του ότι χρησιμοποιήσαμε 2 meta αλγόριθμους το success rate θα αναμέντε υψηλό.