



ΑΝΔΡΟΥΛΑΚΗΣ ΜΑΝΟΣ

A.M. 09470015

ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΥΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Διδάσκων: Γιώργος Τζιραλής

ΔΠΜΣ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

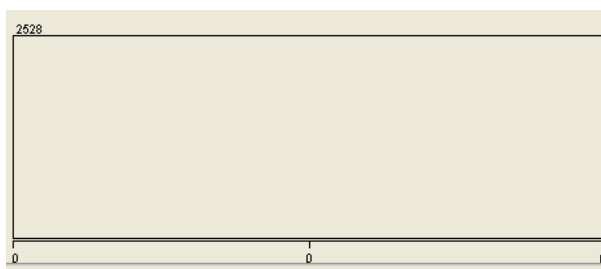
Στάδιο 1^ο. Προεπισκόπηση-προεπεξεργασία δεδομένων:

Δίδονται τα παρακάτω δεδομένα που έχουμε προς επεξεργασία και ανάλυση και τα οποία είναι διαχωρισμένα σε τρία μέρη:

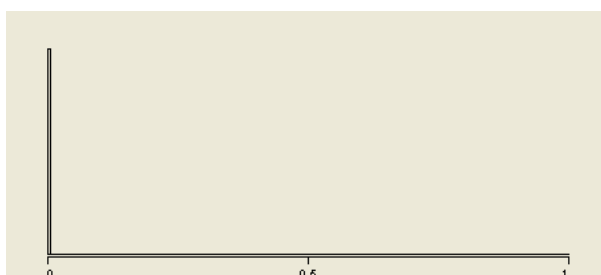
1. Σε δεδομένα εκπαίδευσης (training set) που αποτελούνται από 2528 υποδείγματα και 39 χαρακτηριστικά.
2. Σε δεδομένα επαλήθευσης (quiz set) που αποτελούνται από 1265 υποδείγματα και 39 χαρακτηριστικά.
3. Σε δεδομένα εξέτασης (test set) που αποτελούνται από 1265 υποδείγματα και 39 χαρακτηριστικά.

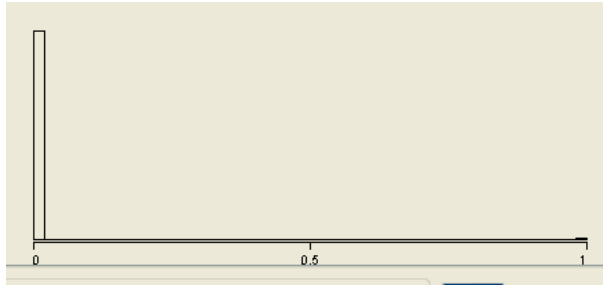
Γνωρίζουμε ότι κατά την ανάλυση ενός μεγάλου όγκου δεδομένων, ώστε να βρεθούν σχέσεις, αναμενόμενες και μη, ανάμεσα στα δεδομένα, είναι αρκετά χρήσιμο να γίνει μια πρώτη προεπεξεργασία, πριν την οποιαδήποτε εφαρμογή αλγορίθμων. Όσον αφορά το dataset 'training set', με το οποίο και θα δουλέψουμε αρχικά, παρατηρούμε τα εξής:

- ✓ Στο attribute38, κάθε instance έχει τιμή 0. Αυτό σημαίνει ότι είναι καλό να αφαιρεθεί. Παρατηρώντας το παρακάτω ιστόγραμμα για το attribute 38, ενισχύεται η άποψή μας:



- ✓ Τα attributes 36 και 34, έχουν από 15 και 1 αντίστοιχα καταχωρήσεις με 1, ενώ όλες οι υπόλοιπες είναι 0. Παραθέτουμε και τα δύο ιστογράμματα, για τα attributes 34 και 36 αντίστοιχα:





- ✓ Αυτό σημαίνει ότι ίσως θα πρέπει να εκτελεστούν κάποιοι μέθοδοι παλινδρόμησης ώστε έπειτα να αποφασιστεί αν θα παραμείνουν τα συγκεκριμένα attributes, ή θα πρέπει να τα αφαιρέσουμε.
- ✓ Παρατηρούμε επίσης ότι missing values δεν υπάρχουν, αν και αυτό δε θα αποτελούσε πρόβλημα, καθότι στο Weka υπάρχει η δυνατότητα αντικατάστασης τους.

Οπότε οδηγούμαστε στη δημιουργία ενός καινούργιου αρχείου, του train1, στο οποίο έχουμε αφαιρέσει τα attributes 34, 36, και 38. Εφαρμόζουμε στη συνέχεια linear regression, και στο train και στο train1, και συγκρίνουμε τα αποτελέσματα, για να ελέγξουμε αν καλώς πράξαμε:

Για το train προκύπτει το ακόλουθο log-file:

Correlation coefficient	0.743
Mean absolute error	0.2056
Root mean squared error	0.2847
Relative absolute error	56.8375 %
Root relative squared error	66.9218 %
Total Number of Instances	2528

Συνεχίζουμε με το train1 και παίρνουμε τα παρακάτω αποτελέσματα:

Correlation coefficient	0.7423
Mean absolute error	0.2059
Root mean squared error	0.285

Relative absolute error	56.918 %
Root relative squared error	66.9906 %
Total Number of Instances	2528

Άρα είναι προφανές από τα σφάλματα παλινδρόμησης και τις τιμές των correlation coefficients ότι δεν επηρεάζει το μοντέλο η αφαίρεση των προαναφερθέντων attributes, άρα συνεχίζουμε έχοντας ως βάση το train1.

Στάδιο 2^ο . Εφαρμογή αλγορίθμων στα training και quiz sets.

Σε αυτό το σημείο, θα αναφέρουμε τους καλύτερους αλγόριθμους που χρησιμοποιήσαμε, βάσει των τιμών του correlation coefficient, των σφαλμάτων που καθένας δίνει και του ποσοστού πρόβλεψης. Η διαδικασία που ακολουθήσαμε είναι η παρακάτω:

Χρησιμοποιήσαμε διάφορους αλγορίθμους στο Weka, πρώτα στο training set και έπειτα στο quiz. Να σημειώσουμε ότι κατάρχην, το training set χρησιμοποιείται για τη δημιουργία μοντέλου, ενώ με τη χρήση cross-validation μπορεί να προκύψει και αξιόπιστη εκτίμηση του σφάλματος. Ωστόσο, το μοντέλο που προκύπτει δεν μπορεί να χρησιμοποιηθεί στο Weka για την εκπόνηση πρόβλεψης σε dataset διαφορετικού αριθμού χαρακτηριστικών, όπως το quiz set. Για αυτό το λόγο είναι αναγκαία η προσθήκη ενός πλαστού χαρακτηριστικού 'prediction', το οποίο και περιέχει το πρόσθετο attribute 'prediction' με τιμή ίση με 0 σε όλα τα instances του quiz. Αυτό το αρχείο το αποθηκεύσαμε ως quizextended. Έπειτα, εφαρμόσαμε σε αυτό όλα τα βήματα προεπεξεργασίας που κάναμε στο training set, και συγκεκριμένα αφαιρέσαμε τις μεταβλητές 34, 36, και 38. Το καινούργιο αρχείο το αποθηκεύσαμε ως quizextended1 και με βάση αυτό θα δουλέψουμε. Εφαρμόσαμε τους αλγορίθμους πρώτα στο training set και έπειτα στο quizextended1 set οπότε και αποθηκεύσαμε τα predictions.

Παρατήρηση: Δώσαμε μεγάλη σημασία στη συμβατότητα των επιλεγμένων αλγορίθμων με το class attribute, το οποίο και είναι numeric. Αυτό για το λόγο ότι

αρκετοί έχουν σαν προϋπόθεση για να μπορούν να χρησιμοποιηθούν, την ύπαρξη nominal μεταβλητής στόχου. Υπάρχουν τρόποι για να αντιμετωπιστεί αυτό, όπως με χρήση κάποιου φίλτρου, αλλά δε συνίσταται η αλλαγή του class attribute, οπότε και δεν έγινε κάτι τέτοιο.

Έπειτα λοιπόν από αρκετούς πειραματισμούς και συνδυασμούς αλγορίθμων, καταλήξαμε στην παρακάτω καλύτερη τετράδα. Καταγράφουμε επίσης σε κάθε αλγόριθμο τα αποτελέσματα που πήραμε στο training set.

Αλγόριθμος 1

EnsembleSelection→Backward Elimination

Correlation coefficient	0.8951
Mean absolute error	0.082
Root mean squared error	0.1901
Relative absolute error	22.658 %
Root relative squared error	44.6943 %
Total Number of Instances	2528

Αλγόριθμος 2

Bagging→RegressionByDescretization→J48

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
---------------------------	------

Αλγόριθμος 3

RandomSubSpace→RegressionByDescretization→J48

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

Αλγόριθμος 4

RandomSubSpace→RegressionByDescretization→J48graft

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

Οι παραπάνω αλγόριθμοι, έπειτα από την εφαρμογή τους στο quizextended1 set, θέτοντάς το ως supplied test set, έδωσαν τα εξής success rates:

Αλγόριθμος 1: 0,948

Αλγόριθμος 2: 0,953

Αλγόριθμος 3: 0,953

Αλγόριθμος 4: 0,955

Οπότε οδηγηθήκαμε στην επιλογή του αλγορίθμου

RandomSubSpace→RegressionByDiscretization→J48graft,

καθότι μας έδωσε το υψηλότερο correlation coefficient, τα μικρότερα σφάλματα και το καλύτερο success rate.

Στάδιο 3°. Εφαρμογή του επιλεγμένου αλγορίθμου στο test set.

Πριν την εφαρμογή του παραπάνω αλγορίθμου στο test set, πρέπει και σε αυτό να αφαιρεθούν οι μεταβλητές 34, 36 και 38, καθώς και να προσθέσουμε ένα πρόσθετο attribute 'prediction' με τιμή ίση με 0 σε όλα τα instances του test set. Το τελικό αρχείο το αποθηκεύσαμε ως testextendedteliko. Οπότε χρησιμοποιήσαμε τον αλγόριθμο στο training set με cross-validation, και έπειτα τον εφαρμόσαμε πάλι, θέτοντας ως supplied test set το test και αναμένουμε τα αποτελέσματα. Το μόνο μας πρόβλημα, είναι ότι ο συνδυασμός των δύο meta αλγορίθμων, ίσως οδηγήσει σε overtraining, αλλά εμένουμε στην επιλογή μας, λόγω των πολύ καλών αποτελεσμάτων που πήραμε.