

Data Mining Techniques – Εργασία 1

Ημερομηνία Παράδοσης: 13/11/2016

Η εργασία είναι ομαδική (ομάδες των 2 ατόμων)

Διδάσκοντας: Ι. Κωτίδης (kotidis@aueb.gr)

Βοηθός: Ι. Φιλιππίδου (filippidou@aueb.gr)

Περιγραφή

Στην παρούσα εργασία καλείστε να επεξεργαστείτε ένα data set, χρησιμοποιώντας μεθόδους και τεχνικές όπως αυτές παρουσιάστηκαν στο μάθημα με σκοπό να εξάγετε χρήσιμες πληροφορίες και ενδιαφέροντα συμπεράσματα που δεν είναι προφανή από τα πρωτογενή δεδομένα. Συγκεκριμένα πρέπει να επιλέξετε και να χρησιμοποιήσετε μεθόδους από τις γενικότερες κατηγορίες association rule mining, dimensionality reduction, clustering και classification. Για την εφαρμογή τους σας προτείνεται να χρησιμοποιήσετε το λογισμικό Weka και τους αντίστοιχους αλγόριθμους που περιλαμβάνονται. Το Weka μπορεί να χρησιμοποιηθεί μέσω του GUI που προσφέρεται, ή του αντίστοιχου API. Εναλλακτικά μπορείτε να χρησιμοποιήσετε οποιοδήποτε άλλο πρόγραμμα ή framework επιλέξετε ή ακόμα και τις δικές σας υλοποιήσεις των αντίστοιχων αλγορίθμων και τεχνικών αφού όμως πρώτα έρθετε σε σχετική επικοινωνία με τον βοηθό του μαθήματος.

Παραδοτέο

Το παραδοτέο σας πρέπει να είναι ένα αρχείο pdf, το οποίο προτείνεται να περιέχει τα ακόλουθα κεφάλαια:

- 1) **Εισαγωγή/Περίληψη:** Θα πρέπει να παρουσιάζονται τα αρχικά στοιχεία του dataset που χρησιμοποιήθηκε (attributes, number of instances, missing values etc.), και οι αλγόριθμοι που εφαρμόστηκαν πάνω σε αυτά.
- 2) **Προεπεξεργασία και Μετατροπή Δεδομένων:** Θα πρέπει να περιγράφετε οποιαδήποτε μετατροπή και επεξεργασία έγινε στην μορφή των αρχικών δεδομένων καθώς και τα attributes, instances που επιλέχθηκαν να χρησιμοποιηθούν ή να απορριφθούν αντίστοιχα.

3) Εφαρμογή Μεθόδων και Τεχνικών Data Mining: Θα αναφέρονται αναλυτικά οι μέθοδοι και οι παράμετροι που χρησιμοποιήθηκαν για κάθε τεχνική. Επίσης θα πρέπει να περιγραφούν οι μέθοδοι και οι παράμετροι που δοκιμάστηκαν αλλά δεν χρησιμοποιήθηκαν τελικά και τα προβλήματα που προέκυψαν κατά την διαδικασία και πως αντιμετωπίστηκαν. Για την εργασία απαιτείται η χρήση και η παρουσίαση τουλάχιστον δύο από τις μεθόδους που παρουσιάστηκαν στο μάθημα καθώς και οποιαδήποτε επέκταση τους κρίνεται απαραίτητη ή παρουσιάζει σημαντικές διαφοροποιήσεις στα αποτελέσματα.

4) Παρουσίαση Συμπερασμάτων: Στο κεφάλαιο αυτό είναι επιθυμητό να παρουσιάσετε τα συμπεράσματά σας χρησιμοποιώντας κατάλληλα διαγράμματα και σχόλια.

Μπορείτε να συνοδέψετε το pdf με ένα συμπιεσμένο αρχείο με scripts, προγράμματα ή ότι άλλο θεωρείτε χρήσιμο ως συνοδευτικό υλικό. Στην πρώτη σελίδα της εργασίας θα πρέπει να αναφέρεται το ονοματεπώνυμο σας και ο ΑΜ σας. Το παραδοτέο σας θα πρέπει να υποβληθεί με email στον βοηθό του μαθήματος, εντός της προθεσμίας που αναφέρεται στην αρχή του εγγράφου.

Τα **dataset** που θα χρησιμοποιήσετε στην εργασία θα σας ανατεθούν από τον βοηθό του μαθήματος αφού του στείλετε email με την σύνθεση της ομάδας σας (ονοματεπώνυμο και Α.Μ.).