Αναγνώριση Προτύπων

1η Εργασία Μαθήματος – Ταξινόμηση

Σκοπός της Εργασίας:
Στο πλαίσιο της παρούσας εργασίας, θα κληθείτε να εφαρμόσετε μια σειρά από τεχνικές ταξινόμησης, να συγκρίνετε τα μοντέλα μεταξύ τους και να αποφασίσετε το βέλτιστο.

Πρόβλημα:
Η πρόβλεψη σφαλμάτων σε έργα λογισμικού αποτελεί ένα πολύ ενδιαφέρον και δύσκολο πρόβλημα. Ένας τρόπος προσέγγισης του προβλήματος είναι η ταξινόμηση των διαφόρων τμημάτων (components) του συστήματος με χρήση μετρικών σε αυτά που έχουν και αυτά που δεν έχουν σφάλμα. Για το πρόβλημα που θα αντιμετωπίσετε δίνεται ένα σύνολο από υπολογισμένες μετρικές σε τμήματα (κλάσεις) του λογισμικού Eclipse. Ζητείται από εσάς να ταξινομήσετε επιτυχώς τις κλάσεις αυτές σε εσφαλμένες και «καθαρές» (μη εσφαλμένες). Σαφείς οδηγίες για τα δεδομένα δίνονται στο αρχείο Dataset.pdf.

Διαδικασία:
Ακολουθήστε την παρακάτω διαδικασία:

Βήμα 1. Εξερευνήστε τα δεδομένα και κάντε οποιοσδήποτε μορφής καθαρισμό/κανονικοποίηση θεωρείτε απαραίτητο. Δημιουργήστε νέα μεγέθη, αν θεωρείτε ότι μπορούν να σας βοηθήσουν.

Βήμα 2. Επιλέξτε τουλάχιστον τρεις τεχνικές ταξινόμησης (Δένδρα, Πιθανοτικούς, Νευρωνικά, SVMs, NNs), και δυο αλγορίθμους από κάθε τεχνική.

Βήμα 3. Αξιολογήστε τα μοντέλα με βάση τα Accuracy, Precision, Recall και F-measure. Κατασκευάστε τον πίνακα σύγχυσης για κάθε μοντέλο.

Βήμα 5. Επιλέξτε το καλύτερο μοντέλο. Δικαιολογήστε την επιλογή σας.

Παρατήρηση: Αν κρίνετε ότι υπάρχει θέμα class imbalance, μπορείτε στο βήμα 3 να ορίσετε και πίνακες κόστους, ενώ στα βήματα 4-6 και meta-classifiers που λαμβάνουν υπόψη τους το cost matrix κατά την κατασκευή του δένδρου. Δείτε το παρακάτω σχόλιο του WEKA (σε περίπτωση που χρησιμοποιήσετε WEKA):

> *"NOTE: The behaviour of the -m option has changed between WEKA 3.0 and WEKA 3.1. -m now carries cost-sensitive evaluation only. For cost-sensitive prediction, use one of the cost-sensitive schemes such as weka.classifiers.meta.CostSensitiveClassifier or weka.classifiers.meta.MetaCost".*

Ομάδες Εργασίας:
Οι ομάδες είναι 3 ατόμων. Θα τις δημιουργήσετε στον δικτυακό τόπο του μαθήματος (και οι συνάδελφοί σας θα κάνουν join), στην Ενότητα "Υποβολή Εργασιών -> Ταξινόμηση".
Προθεσμία δήλωσης ομάδας: Παρασκευή 06/11/2015, 23:59 (**Αυστηρή προθεσμία!**)

Παραδοτέα:

1. Ο πηγαίος κώδικας που θα υλοποιήσετε. Η υλοποίηση μπορεί να γίνει σε Matlab, WEKA, R, Python ή συνδυασμούς αυτών.
2. Όλα τα πειράματα που θα κάνετε.
3. Έγγραφη αναφορά η οποία θα περιέχει: α) σύντομη περιγραφή του προβλήματος που καλείστε να επιλύσετε, β) τις λύσεις που δώσατε (παραμέτρους και πειράματα) και γ) αξιολόγηση των αποτελεσμάτων και συμπεράσματα. Στο εξώφυλλο μη ξεχάσετε να περιλάβετε τον τίτλο της εργασίας, ημερομηνία, καθώς και τα ονόματα και τις ηλεκτρονικές διευθύνσεις των μελών της ομάδας σας.
4. Το τελικό μοντέλο.
5. Το αρχείο bugs_test.csv που θα περιέχει τα αποτελέσματα της εκτέλεσης του τελικού σας μοντέλου στο αρχείο source-code-metrics_test.csv. Το αρχείο που θα παραδώσετε θα πρέπει να έχει την ίδια μορφή με το bugs_train.csv, δηλαδή να περιέχει classid και bugs χωρισμένα με ερωτηματικό (;).

Βαθμολογία:

Όπως έχει ήδη αναφερθεί, η εργασία αυτή αποτελεί το 25% του τελικού βαθμού.

Εξέταση εργασίας:

Η εργασία θα πρέπει να ολοκληρωθεί μόνο κατά την περίοδο Φεβρουαρίου.

Υποβολή:

Τα παραδοτέα πρέπει να γίνουν ένα zip αρχείο, με όνομα Επίθ1_Επίθ2_Επίθ3.zip. Η υποβολή του αρχείου αυτού θα γίνει μέσω του δικτυακού τόπου του μαθήματος.

Προθεσμία υποβολής: Δευτέρα, 23/11/2015, 23:59.

Για περισσότερες πληροφορίες:

Ανδρέας Λ. Συμεωνίδης	**Βάγια Καλτσά**	**Θεμιστοκλής Διαμαντόπουλος**
Επ. Καθηγητής	Μεταπτυχιακή Φοιτήτρια	Μεταπτυχιακός Φοιτητής
E-mail. asymeon@eng.auth.gr	E-mail: vkaltsa@auth.gr	E-mail: thdiaman@auth.gr