ΔΕΥΤΕΡΗ ΕΡΓΑΣΙΑ
ΣΤΟ ΜΑΘΗΜΑ
«ΤΕΧΝΟΛΟΓΙΕΣ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ ΚΑΙ ΡΟΜΠΟΤΙΚΗ»
2016-17

Γενικά

Στην εργασία αυτή καλείστε να δημιουργήσετε ένα ευφυές/έμπειρο σύστημα με εξαγωγή κανόνων από ένα σύνολο δεδομένων.

Δίνονται εννέα (9) σύνολα δεδομένων (ΣΔ) από την βάση UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/), στο τέλος αυτού του εγγράφου. Κάθε ΣΔ έχει ένα αριθμό (1-9). Η εργασία είναι ατομική. Σε καθένα ανατίθεται το ΣΔ με αριθμό ίδιο με το τελευταίο ψηφίο του Α.Μ. του. Αν το τελευταίο ψηφίο είναι '0' (μηδέν) τότε λαμβάνεται υπ' όψιν το προτελευταίο κ.ο.κ. Προφανώς, κάποιοι θα έχετε κοινό ΣΔ. Θα πρέπει όμως να εργαστείτε ανεξάρτητα, ώστε τα συστήματά σας να μην είναι παρόμοια. Για οποιεσδήποτε ερωτήσεις μπορείτε να έρχεστε στο γραφείο μου τα απογεύματα 5.30-7.30. Η προτεινόμενη προθεσμία παράδοσης είναι η **Τρίτη 4 Ιουλίου 2017**.

Τα παραδοτέα είναι:

1. Αρχεία .arff (WEKA)
2. Πρόγραμμα CLIPS
3. Τεχνική Αναφορά

Η δομή της Τεχνικής Αναφοράς περιγράφεται στο τέλος αυτού του εγγράφου.

Η εργασία αυτή θα συμβάλλει κατά 60% στο βαθμό του μαθήματος.

Θα γίνει προφορική εξέταση της εργασίας την Πέμπτη ή Παρασκευή 6 ή 7 Ιουλίου 2017.

Το Σύνολο Δεδομένων (ΣΔ)

Κάθε ΣΔ αναφέρεται σε κάποιο πρόβλημα ταξινόμησης. Κάθε τέτοιο πρόβλημα σχετίζεται με ένα αριθμό παραμέτρων (παράμετροι εισόδου) που παίζουν ρόλο στην εξαγωγή της απόφασης (παράμετρος εξόδου). Για κάθε συνδυασμό τιμών των παραμέτρων εισόδου εξάγεται και μια τιμή της παραμέτρου εξόδου, που ονομάζεται κλάση εξόδου. Θεωρούμε ότι ένα τέτοιο σύνολο αποτελείται από Ν παραδείγματα. Κάθε παράδειγμα είναι ένα σύνολο τιμών για τις παραμέτρους εισόδου και την παράμετρο εξόδου. Στην παραπάνω βάση (UCI), τα σύνολα δεδομένων δίνονται με δύο αρχεία. Το ένα (κατάληξη .names ή .doc) περιγράφει το πεδίο που αφορά το σύνολο και δίνει πληροφορίες για τις παραμέτρους εισόδου και τις κλάσεις εξόδου. Το άλλο (κατάληξη .data) είναι το πραγματικό σύνολο δεδομένων.

Στο σύνολο αυτό δεδομένων μπορεί να χρειαστεί να το επεξεργαστείτε πριν το χρησιμοποιήσετε. Π.χ. μπορεί να υπάρχουν παραδείγματα με ελλιπείς τιμές. Αυτές ή θα τις συμπληρώσετε (π.χ. με το μέσο όρο από γειτονικά παραδείγματα ή με την πιο κοινή τιμή

από γειτονικά παραδείγματα ή με κάποια συστηματική μέθοδο-πράγμα που θα μετρήσει θετικά) ή θα αφαιρέσετε εντελώς τα παραδείγματα αυτά. Επίσης, μπορεί να χρειαστεί να αφαιρέσετε κάποιες παραμέτρους διότι δεν παίζουν ρόλο (αυτό θα είναι ξεκάθαρο ή θα αναφέρεται στην περιγραφή του συνόλου δεδομένων ή θα το διαπιστώσετε εσείς με κάποιο τρόπο-πράγμα που επίσης θα μετρήσει θετικά). Π.χ. μια τέτοια παράμετρος μπορεί να είναι κάποια που παίζει το ρόλο αριθμού μητρώου ή απλώς δείκτη ή έχει σταθερή τιμή για όλα τα παραδείγματα. Επίσης, μπορείτε να μειώσετε τον αριθμό κλάσεων εξόδου, αν είναι μεγάλος, αφαιρώντας κάποιες τιμές από την παράμετρο εξόδου και μαζί τα αντίστοιχα παραδείγματα από το σύνολο δεδομένων. Π.χ. μια τέτοια περίπτωση είναι θεμιτή όταν κάποιες κλάσεις αντιπροσωπεύουν πολύ λιγότερα παραδείγματα από τις υπόλοιπες. Το εργαλείο WEKA περιλαμβάνει τέτοιες μεθόδους προεπεξεργασίας δεδομένων, τις οποίες μπορείτε να χρησιμοποιήσετε, αφού τις περιγράψετε συνοπτικά (πράγμα που θα μετρήσει θετικά). Θα σας είναι χρήσιμο να αποκτήσετε μια «εικόνα» του ΣΔ που έχετε, χρησιμοποιώντας τα εργαλεία οπτικοποίησης για ΣΔ του WEKA και παρουσιάζοντας τα αποτελέσματά τους.

Το σύνολο που θα προκύψει, θα πρέπει να χωριστεί σε δύο υποσύνολα, το σύνολο εκπαίδευσης και το σύνολο ελέγχου, όπως αναφέρεται και στη συνέχεια. Συνήθως, το *σύνολο εκπαίδευσης* (ΣΕΚ) (training set) είναι τα 2/3 και το *σύνολο ελέγχου* (ΣΕΛ) (test set) το 1/3 του αρχικού συνόλου. Σ' αυτό τον χωρισμό προσπαθούμε να κρατήσουμε την ίδια αναλογία μεταξύ των διαφορετικών εξόδων στα δύο σύνολα. Π.χ. αν στο αρχικό σύνολο η έξοδος (παράμετρος/κλάση εξόδου) παίρνει τιμές yes-no και έχουμε π.χ. από τις 100 περιπτώσεις 60 yes και 40 no, τότε στα σύνολα εκπαίδευσης και ελέγχου προσπαθούμε να κρατήσουμε αυτή την αναλογία (περίπου).Το ίδιο ισχύει και για περισσότερες κλάσεις εξόδου.

1. Δημιουργία δέντρου αποφάσεων (WEKA)

Κατ' αρχήν, θα δημιουργήσετε ένα σύνολο κανόνων που θα μοντελοποιούν το ΣΕΚ και θα προκύψουν από τη δημιουργία ενός δέντρου αποφάσεων (ΔΑ), χρησιμοποιώντας ένα αντίστοιχο αλγόριθμο μέσω του εργαλείου WEKA (http://www.cs.waikato.ac.nz/ml/weka/). Ένας τέτοιος αλγόριθμος, που παράγει δέντρα αποφάσεων από δεδομένα στο WEKA, είναι ο *J48* (μια υλοποίηση του *C4.5*). (Οδηγίες για τη χρήση του WEKA και πιο συγκεκριμένα για την εφαρμογή του *J48* μπορείτε να βρείτε στο επισυναπτόμενο αρχείο 'weka_explorer.pdf').

Για να το κάνετε αυτό θα χρειαστεί να διασπάσετε το ΣΔ από τη UCI Repository σε ΣΕΚ και ΣΕΛ με τον τρόπο που αναφέρθηκε πιο πάνω. Θα προσπαθήσετε, ρυθμίζοντας διάφορες παραμέτρους του αλγορίθμου, να επιτύχετε όσο το δυνατόν καλύτερα αποτελέσματα, δηλ. καλύτερη ταξινόμηση των παραδειγμάτων του ΣΕΚ, με βάση τις μετρικές που αναφέρονται στο αρχείο 'ΜΕΤΡΙΚΕΣ_ΑΞΙΟΛΟΓΗΣΗΣ.pdf', που μπορείτε να υπολογίσετε από το confusion matrix, που επιστρέφει το WEKA. Επίσης, προσπαθήστε τα ΔΑ να περιέχουν ικανό αριθμό από τις παραμέτρους εισόδου. **Θα παρουσιάσετε τα δύο καλύτερα ΔΑ και θα τα σχολιάσετε**. Στη συνέχεια θα εφαρμόσετε τα δύο αυτά μοντέλα (ΔΑ) στο ΣΕΛ και θα υπολογίσετε πάλι τις μετρικές, τις οποίες θα κρατήσετε. **Σχολιάστε τα αποτελέσματα**. Στην πρώτη περίπτωση (σύνολο ΣΕΚ) επιλέγετε "Use training set" στην καρτέλα "Classify", ενώ στη δεύτερη (σύνολο ΣΕΛ) επιλέγετε "Supplied test set", και ανεβάζετε τα αντίστοιχα αρχεία.

2. Δημιουργία έμπειρου συστήματος σε CLIPS (CLIPS)

Εδώ θα δημιουργήσετε ένα απλό έμπειρο σύστημα (ΕΣ), δηλαδή ένα σύνολο κανόνων για διάγνωση/ταξινόμηση, όχι μέσω συνεντεύξεων με εμπειρογνώμονα ή μέσω βιβλιογραφίας, αλλά ημι-αυτόματα μέσω ενός ΔΑ που προσδιορίσατε στο 1.

Θα καταγράψετε τους κανόνες που παράγονται από το καλύτερο ΔΑ στο 1. Στη συνέχεια θα **υλοποιήσετε τους κανόνες σε CLIPS** ώστε να δημιουργήσετε τη βάση γνώσης του ΕΣ. Θα χωρίσετε το σύνολο δεδομένων στα ίδια σύνολα ΣΕΚ και ΣΕΛ, όπως παραπάνω. Κατόπιν θα πρέπει να βρείτε ένα τρόπο να εισάγετε αυτόματα τα δεδομένα στο CLIPS για να τρέξετε το ΕΣ στο ΣΕΚ και να υπολογίσετε τις τιμές των παραπάνω μετρικών αξιολόγησης. Αν κάνετε σωστά την υλοποίηση θα πρέπει να βρίσκετε τα ίδια αποτελέσματα με το WEKA όσον αφορά τις μετρικές. Για τον υπολογισμό των μετρικών μπορείτε να γράψετε κώδικα σε CLIPS (με κανόνες και συναρτήσεις μέσα στο ΕΣ) ή σε άλλη γλώσσα προγραμματισμού. Στη συνέχεια, προσπαθήστε να κάνετε αλλαγές στους κανόνες (π.χ. διαγραφή ή προσθήκη κάποιων συνθηκών ή αλλαγές σε κάποιες συνθήκες, αφαίρεση κανόνων κλπ) ώστε να επιτύχετε τα καλύτερα δυνατά αποτελέσματα. Σ' αυτό θα βοηθηθείτε από τη μελέτη βιβλιογραφίας για το πρόβλημα και τη λεπτομερή μελέτη των αποτελεσμάτων στα παραδείγματα του ΣΕΚ.

Αφού καταλήξετε στο σύστημα που σας δίνει τα καλύτερα αποτελέσματα στις μετρικές, με βάση το ΣΕΚ, θα το εφαρμόσετε στο ΣΕΛ και θα υπολογίσετε πάλι τις τιμές των μετρικών, τις οποίες και θα κρατήσετε. **Συζητήστε τα αποτελέσματα** σε σχέση με αυτά του WEKA.

Δομή Τεχνικής Αναφοράς

1. Περιγραφή του προβλήματος που διαπραγματεύεται το ΕΣ
   1.1. Πρόβλημα, παράμετροι, κλάσεις εξόδου
   1.2. Προεπεξεργασία συνόλου δεδομένων.
2. Δημιουργία Δέντρων Αποφάσεων
   2.1. Περιγραφή εξαγωγής δέντρων αποφάσεων μέσω WEKA
   2.2. Περιγραφή προσπαθειών βελτίωσης δέντρων
   2.3. Αποτελέσματα μετρικών στο σύνολο εκπαίδευσης
   2.4. Αποτελέσματα μετρικών στο σύνολο ελέγχου
3. Υλοποίηση σε CLIPS
   3.1. Εξαγωγή κανόνων από WEKA
   3.2. Περιγραφή υλοποίησης των κανόνων σε CLIPS
   3.3. Αλλαγές και «ρύθμιση» του ΕΣ με βάση το σύνολο εκπαίδευσης
   3.4. Αποτελέσματα μετρικών στο σύνολο εκπαίδευσης
   3.5. Αποτελέσματα μετρικών στο σύνολο ελέγχου
4. Γενικές παρατηρήσεις και συμπεράσματα.

ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ

1. Liver Disorders
2. Blood Transfusion Service Center
3. Mammographic Mass
4. Pima Indians Diabets
5. Echocardiogram
6. Credit Approval
7. Statlog (Australian Credit Approval)
8. Statlog (Heart)
9. **Breast Cancer**