ΤΕΧΝΟΛΟΓΙΕΣ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ ΚΑΙ ΡΟΜΠΟΤΙΚΗ
Project B 2022-2023

ΓΕΝΙΚΑ

Σκοπός της εργασίας είναι να εξοικειωθείτε με διάφορες ευφυείς μεθόδους που μπορούν να χρησιμοποιηθούν για επίλυση προβλημάτων ταξινόμησης (classification) και να αποκτήσετε και εμπειρίες από τη σύγκρισή τους. Επίσης, να εξοικειωθείτε με τον τρόπο δημιουργίας ευφυών συστημάτων ταξινόμησης από εμπειρικά δεδομένα.

Η εργασία είναι ατομική ή ομαδική (μέχρι 2 άτομα) και αφορά τα εξής:

1. Περιγραφή αλγορίθμων εξαγωγής κανόνων
2. Εξαγωγή μοντέλου κανόνων μέσω του εργαλείου WEKA.
3. Δημιουργία έμπειρου συστήματος (ΕΣ) σε CLIPS με βάση τους κανόνες από το WEKA.
4. Δημιουργία νευρωνικού δικτύου ταξινόμησης.
5. Σύγκριση των παραπάνω ταξινομητών.

Το Σύνολο Δεδομένων (ΣΔ)

Δίνεται ένα σύνολο δεδομένων (dataset) σε κάθε ομάδα, όπως φαίνεται στον πίνακα αναθέσεων, και το οποίο 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. Οι περιγραφές των συνόλων δεδομένων δίνονται είτε στο eclass είτε στο UCI machine learning repository. Στην παραπάνω βάση 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.

Το σύνολο αυτό δεδομένων μπορεί να χρειαστεί να το προ-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. Μπορείτε να χρησιμοποιήσετε μεθόδους αξιολόγησης των παραμέτρων εισόδου για μείωση της διαστατικότητας του ΣΔ από το WEKA. 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 μεθόδους (φίλτρα) προεπεξεργασίας δεδομένων, τις οποίες μπορείτε να χρησιμοποιήσετε, αφού τις περιγράψετε συνοπτικά (πράγμα που θα μετρήσει θετικά).

Το σύνολο που θα προκύψει, θα πρέπει να χωριστεί σε δύο υποσύνολα, το σύνολο εκπαίδευσης (ΣΕΚ) (συνήθως 2/3 του ΣΔ) και το σύνολο ελέγχου (ΣΕΛ) (συνήθως το 1/3 του ΣΔ). Με το ΣΕΚ γίνεται εκπαίδευση του μοντέλου ταξινόμησης και με το ΣΕΛ γίνεται αξιολόγηση του μοντέλου. Επειδή ένας και μοναδικός τέτοιος διαχωρισμός μπορεί να μην είναι υπολογιστικά «ουδέτερος», για να έχουμε καλύτερα αποτελέσματα αξιολόγησης μιας μεθόδου, χρησιμοποιούμε τη μέθοδο k-fold cross validation. Το WEKA διαθέτει έτοιμη επιλογή για τέτοια περίπτωση, στην οποία μοναδική παράμετρος είναι το k (default τιμή 10, αλλά πρέπει να προσαρμόζεται στο ΣΔ).

1. Περιγραφή Αλγορίθμων Εξαγωγής Κανόνων

Υπάρχουν δύο κατηγορίες αλγορίθμων εξαγωγής κανόνων από δεδομένα: αλγόριθμοι εξαγωγής δέντρων απόφασης και αλγόριθμοι απ' ευθείας εξαγωγής κανόνων. Στην πρώτη κατηγορία ανήκουν αλγόριθμοι όπως οι C4.5 (J4.8), RandomTree, LMT, REPTree, ενώ στη δεύτερη αλγόριθμοι όπως οι JRIP, PART, OneR. Στο τμήμα αυτό της εργασίας θα περιγράψετε ΑΝΑΛΥΤΙΚΑ

a) Τους αλγορίθμους που σας έχουν ανατεθεί στον πίνακα αναθέσεων (πλην του J48), παραθέτοντας σχετική βιβλιογραφία, και ένα αναλυτικό παράδειγμα εφαρμογής για τον καθένα.
b) Τις παραμέτρους ρύθμισης κάθε αλγορίθμου που είναι διαθέσιμοι γι' αυτούς στο εργαλείο WEKA.

2. Εξαγωγή Κανόνων (WEKA)

Στο τμήμα αυτό θα δημιουργήσετε ένα σύνολο κανόνων που θα μοντελοποιούν το ΣΔ και θα προκύψουν από την εφαρμογή των αλγορίθμων που σας έχουν ανατεθεί μέσω του WEKA. (Οδηγίες για τη χρήση του WEKA στις διαφάνειες του μαθήματος).

Για να το κάνετε αυτό θα χρειαστεί να διασπάσετε το ΣΔ σε ΣΕΚ και ΣΕΛ με τον τρόπο που αναφέρθηκε στην εισαγωγή, δηλ. τη μέθοδο k-fold cross validation. Θα προσπαθήσετε ρυθμίζοντας διάφορες παραμέτρους των αλγορίθμων να επιτύχετε όσο το δυνατόν καλύτερα αποτελέσματα, δηλ. καλύτερη ταξινόμηση των παραδειγμάτων του ΣΔ, με βάση τις μετρικές

που αναφέρονται στο αρχείο 'ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ.pdf', που μπορείτε να τις αντιγράψετε από τον πίνακα αποτελεσμάτων του WEKA.

Θα συγκρίνετε τους αλγορίθμους, με βάση τις μετρικές, και θα προσδιορίσετε το καλύτερο σύνολο κανόνων. Καλό θα είναι να μελετήσετε σε ποια παραδείγματα του ΣΔ είναι καλύτερος ο κάθε αλγόριθμος.

3. Δημιουργία έμπειρου συστήματος σε CLIPS (CLIPS)

Εδώ θα δημιουργήσετε ένα απλό έμπειρο σύστημα (ΕΣ), αποτελούμενο από ένα σύνολο κανόνων (όχι ασαφών) για διάγνωση/ταξινόμηση, όχι μέσω συνεντεύξεων με εμπειρογνώμονα ή μέσω βιβλιογραφίας, αλλά ημι-αυτόματα μέσω του συνόλου κανόνων που προσδιορίσατε ως καλύτερο στο 2.

Θα καταγράψετε το σύνολο κανόνων που προσδιορίσατε ως καλύτερο στο 2. Στη συνέχεια θα υλοποιήσετε τους κανόνες στο CLIPS ώστε να δημιουργήσετε τη βάση γνώσης του ΕΣ. Θα τρέξετε το ΕΣ στο ΣΔ και θα υπολογίσετε τις τιμές των παραπάνω μετρικών αξιολόγησης. Αν κάνετε σωστά την υλοποίηση θα πρέπει να βρίσκετε τα ίδια αποτελέσματα με αυτά του WEKA όσον αφορά τις μετρικές. 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 ή/και τη λεπτομερή μελέτη των αποτελεσμάτων στα παραδείγματα του ΣΔ.

Αφού καταλήξετε στο σύστημα που σας δίνει τα καλύτερα αποτελέσματα, θα κρατήσετε τις αντίστοιχες μετρικές για σύγκριση.

4. Δημιουργία Νευρωνικού Δικτύου (ΝΔ)

Δημιουργήστε ένα Νευρωνικό Δίκτυο (ΝΔ) τύπου Multi-layer Perceptron που να ταξινομεί όσο το δυνατόν καλύτερα τα δεδομένα στο ΣΔ. Χρησιμοποιείστε το εργαλείο WEKA. (Οδηγίες για τη χρήση του WEKA στις διαφάνειες του μαθήματος), με μέθοδο εκπαίδευσης-αξιολόγησης την k-fold cross validation. Θα πρέπει να κάνετε διάφορες δοκιμές μεταβάλλοντας τον αριθμό των κρυφών επιπέδων και νευρώνων (hidden layers), τον αριθμό των επαναλήψεων εκπαίδευσης (training time), τον ρυθμό μάθησης (learning rate) κλπ ώστε να πετύχετε ένα ΝΔ που να έχει το δυνατόν καλύτερα αποτελέσματα, βλέποντας τις μετρικές από τον πίνακα του WEKA. Παρουσιάστε τα αποτελέσματα των δοκιμών σ' ένα πίνακα ή/και σε γραφήματα. Αφού καταλήξετε στο καλύτερο δυνατό ΝΔ θα καταγράψετε τις τιμές των μετρικών, τις οποίες θα κρατήσετε για σύγκριση.

5. Σύγκριση Ταξινομητών

Εδώ θα κάνετε σύγκριση των αποτελεσμάτων των παραπάνω συστημάτων/ταξινομητών με βάση τις μετρικές, προσπαθώντας να εξηγήσετε τις διαφορές τους.

ΠΑΡΑΔΟΤΕΑ

1. ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ (που θα περιέχει αναλυτικά τις απαντήσεις στα παραπάνω ερωτήματα)
2. Κώδικας CLIPS
3. Παρουσίαση (διαφάνειες σε pptx).

Τα ΠΑΡΑΔΟΤΕΑ θα συμπιεστούν σε ένα αρχείο rar ή zip, το οποίο θα ανεβάσετε στο eclass/Εργασίες/ΕΡΓΑΣΙΑ ΑΝΑΠΤΥΞΗΣ ΕΣ ΚΑΙ ΤΑΞΙΝΟΜΗΤΩΝ. Οι ομάδες δύο ατόμων θα ανεβάσουν ένα κοινό παραδοτέο για τα δύο μέλη.