ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
ΓΡΑΠΤΗ ΕΡΓΑΣΙΑ 2
Rule-Based Learning

Σκοπός της συγκεκριμένης εργασίας είναι η εξοικείωση με την υλοποίηση των αλγορίθμων παραγωγής κανόνων, μέσω της βιβλιοθήκης Orange της Python. Η εφαρμογή τους έγινε με χρήση του DataSet “wine” που προσφέρεται από την ίδια τη βιβλιοθήκη. Χρησιμοποιήθηκαν 3 παραλλαγές του αλγορίθμου, οι λεπτομέρειες των οποίων θα αναλυθούν παρακάτω. Η αξιολόγηση και η σύγκριση της απόδοσης των αλγορίθμων έγινε, σύμφωνα με τις οδηγίες της εκφώνησης, με χρήση των ακόλουθων μετρικών (στον υπολογισμό των μετρικών επιλέχθηκε τιμή παραμέτρου average=’macro’):

- **precision**: Ορίζεται ως TP / (TP+FP)
- **recall**: Ορίζεται ως TP / (TP+FN)
- **f1**: Ορίζεται ως 2*precision*recall / (precision+recall)

Α. Περιγραφή και μελέτη των αλγορίθμων

Σύμφωνα με τις απαιτήσεις της εκφώνησης χρησιμοποιήθηκαν 3 διαφορετικές παραλλαγές του αλγορίθμου:

- **Rule ordering: Ordered και Evaluator: Entropy**: Επιλέχθηκε ο προσφερόμενος από τη βιβλιοθήκη CN2Learner(), που είναι Ordered ως προς το rule ordering και χρησιμοποιεί εξ' ορισμού Entropy evaluator.
- **Rule ordering: Unordered και Evaluator: Entropy**: Επιλέχθηκε ο προσφερόμενος από τη βιβλιοθήκη CN2UnorderedLearner(), που είναι Unordered ως προς το rule ordering, αλλά χρησιμοποιεί εξ' ορισμού Laplace evaluator, οπότε αυτός ορίστηκε με την κατάλληλη εντολή σε Entropy.
- **Rule ordering: Ordered και Evaluator: Laplace**: Επιλέχθηκε ο προσφερόμενος από τη βιβλιοθήκη CN2Learner(), που είναι Ordered ως προς το rule ordering αλλά χρησιμοποιεί εξ' ορισμού Entropy evaluator, οπότε αυτός ορίστηκε με την κατάλληλη εντολή σε Laplace.

Η συγκριτική μελέτη των αλγορίθμων βασίστηκε στην μελέτη της επίδρασης στην απόδοσή τους 3 βασικών παραμέτρων: beam_width, min_rule_coverage και max_rule_length. Για κάθε παράμετρο μελετήθηκε ένα ευρύ φάσμα τιμών, μέχρι να καταλήξω στην βέλτιστη τριάδα τιμών, που μεγιστοποιεί την απόδοση του κάθε αλγορίθμου.

Αυτό έγινε με επαναληπτική εκτέλεση του κάθε αλγορίθμου με όλους τους δυνατούς συνδυασμούς των παραπάνω παραμέτρων. Ο κώδικας υλοποίησης της επαναληπτικής εκτέλεσης υπάρχει στο συνοδευτικό αρχείο *3_Algorithms_Evaluation.py*. Τα συνολικά αποτελέσματα, που οδήγησαν και στον καθορισμό των βέλτιστων τιμών των παραμέτρων, φαίνονται στο συνοδευτικό αρχείο *Parameter examination data.xls*.

Β. Εξαγωγή των κανόνων που παράγει κάθε αλγόριθμος στις βέλτιστες συνθήκες

Αφού καθορίστηκαν οι βέλτιστες τιμές των 3 παραμέτρων, αυτές χρησιμοποιήθηκαν για να μελετηθούν οι κανόνες που παράγει ο κάθε αλγόριθμος, όταν εφαρμοστεί πάνω στο μελετούμενο DataSet. Οι κανόνες αυτοί εξάγονται και εκτυπώνονται στην κονσόλα. Για το σκοπό αυτό χρησιμοποιήθηκε ο κώδικας που υπάρχει στο συνοδευτικό αρχείο *liapikos_ge2.py*.

Οι βέλτιστες τιμές παραμέτρων, οι βέλτιστες τιμές των μετρικών απόδοσης καθώς και οι κανόνες που παράγει ο κάθε αλγόριθμος στις βέλτιστες συνθήκες φαίνονται συγκεντρωτικά στο συνοδευτικό αρχείο *RuleBased_Results.xlsx*.

Γ. Σύγκριση των αλγορίθμων

Αναφορικά με τις τιμές των παραμέτρων γίνεται αμέσως φανερό ότι ο Unordered βελτιστοποιείται σε πολύ μεγαλύτερες τιμές beam_width σε σχέση με τους Ordered αλγορίθμους. Στις υπόλοιπες παραμέτρους οι 3 αλγόριθμοι παρουσιάζουν παρόμοια μεταξύ τους συμπεριφορά, απαιτώντας μεγάλες τιμές min_rule_coverage και χαμηλές τιμές max_rule_length.

Αναφορικά με τη βέλτιστη απόδοση, αυτή ήταν παρόμοια και στους 3 αλγορίθμους, ανεξαρτήτως της μετρικής που χρησιμοποιείται.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Υλικό μαθήματος.

2. Müller, A.C & Guido, S. - Introduction to Machine Learning with Python. A Guide for Data Scientists, O' Reilly, 2017.

3. Τεκμηρίωση από τον ιστότοπο της βιβλιοθήκης Orange.