Εργασία 2: Κανόνες Συσχέτισης

1. Apriori με το χέρι

Εφαρμόστε τον αλγόριθμο Apriori στις συναλλαγές που εικονίζονται στον παρακάτω πίνακα, με ελάχιστη υποστήριξη (support), s = 40% και ελάχιστη εμπιστοσύνη (confidence), c = 70%. Διευκρινίστε την τεχνική που χρησιμοποιείτε για την παραγωγή των υποψήφιων στοιχειοσυνόλων (itemsets) και παρουσιάστε τα υποψήφια και τα συχνά στοιχειοσύνολα σε κάθε βήμα του αλγόριθμου. Δώστε όλα τα συχνά στοιχειοσύνολα και τους αντίστοιχους κανόνες συσχέτισης που προκύπτουν ταξινομημένους ως προς την εμπιστοσύνη. Για τους κανόνες που ικανοποιούν τους περιορισμούς (δηλαδή τους κανόνες με εμπιστοσύνη μεγαλύτερη από την ελάχιστη εμπιστοσύνη), υπολογίστε επίσης και το lift. Σχολιάστε ποιοι θεωρείτε ότι είναι οι συνολικά πιο ισχυροί κανόνες και γιατί.

Transaction ID	Items
T1	beer, eggs, diapers, bread, cheese
T2	eggs, diapers
T3	beer, eggs, milk
T4	eggs, beer, cheese
T5	beer, milk, diapers

2. Apriori με το Weka

Θεωρείστε το αρχείο δεδομένων house-votes-84.csv το οποίο βασίζεται στις καταχωρήσεις των ψηφοφοριών του Αμερικανικού Κογκρέσου το 1984 (1984 United Stated Congressional Voting Records) οι οποίες είναι διαθέσιμες από το UCI Machine Learning Repository (https://doi.org/10.24432/C5C01P).

Τα δεδομένα περιλαμβάνουν 16 κατηγορικά γνωρίσματα που περιγράφουν τις ψήφους των μελών του Κογκρέσου σε 16 καίρια θέματα, και ένα γνώρισμα κλάσης που αναπαριστά αν το μέλος του Κογκρέσου είναι του Ρεπουμπλικανικού ή του Δημοκρατικού κόμματος. Για τα 16 γνωρίσματα, θεωρείστε ότι είναι δυαδικά γνωρίσματα με τιμές 1 (ψήφισε ναι), και 0 (ψήφισε όχι), ενώ υπάρχουν και κάποιες άγνωστες τιμές (?). Χρησιμοποιείστε το WEKA για να βρείτε κανόνες συσχέτισης για αυτό το σύνολο δεδομένων.

Συγκεκριμένα:

1) Δημιουργείστε ένα κατάλληλο .arff αρχείο.

2) Αντικαταστήστε τις τιμές που λείπουν (τις άγνωστες τιμές) με την τιμή που εμφανίζεται πιο συχνά σε κάθε κατηγορία (δηλαδή για κάθε γνώρισμα).

3) Χρησιμοποιείστε τον Apriori για την εξόρυξη κανόνων συσχέτισης με ελάχιστη υποστήριξη *s=0.5* και ελάχιστη εμπιστοσύνη *c=0.9.*

4) Χρησιμοποιείστε τον Apriori για την εξόρυξη κανόνων συσχέτισης με ελάχιστη υποστήριξη *s=0.5* και ελάχιστο lift, *l=1.5.*

5) Συγκρίνετε τα αποτελέσματα των ερωτημάτων 3 και 4 για τους κορυφαίους-20 κανόνες (δηλαδή, πόσο όμοιοι είναι, ποιοι είναι οι συνολικά πιο ισχυροί κανόνες, κτλ)

6) Χρησιμοποιείστε τον Apriori με κατάλληλη υποστήριξη και εμπιστοσύνη Use Apriori to mine association rules για την εξόρυξη κανόνων συσχέτισης που θα μπορούσαν να χρησιμοποιηθούν για να κατηγοριοποιήσουν έναν ψηφοφόρο ως Ρεπουμπλικάνο ή ως Δημοκρατικό. Σχολιάστε τα αποτελέσματα σας.

Θα παραδώσετε μια αναφορά με τη λύση για το πρόβλημα 1, και όλα σας τα σχόλια για το πρόβλημα 2.

Για το πρόβλημα 2 θα παραδώσετε επίσης το .arff file που χρησιμοποιήσατε, καθώς και τα αρχεία των αποτελεσμάτων για τα ερωτήματα 3, 4 και 6 (αποθηκεύστε την έξοδο του weka ως αρχεία κειμένου .txt).