## 

**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ**

**ΕΡΓΑΣΤΗΡΙΟ 2**

## WEKA: ASSOCIATION RULES (<HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA>)

1. Αποθηκεύστε σε έναν φάκελο τo αρχείo *adult.arff*. Ανοίξτε το με το Νotepad, για να δείτε τι περιέχει. Πρόκειται για δημογραφικά στοιχεία 48842 ενηλίκων Αμερικανών. Στόχος είναι η πρόβλεψη αν το ετήσιο εισόδημά τους ξεπερνάει τα $50,000.
2. Ο αλγόριθμος APRIORI δουλεύει μόνο με ονοματικές μεταβλητές. Οπότε χρειάζεται να ακολουθήσετε την εξής προεπεξεργασία στα δεδομένα σας:

Βήμα 1: Μείωση του μεγέθους του dataset με τυχαία δειγματοληψία. Στο Preprocess, στα φίλτρα, επιλέξτε το φίλτρο 'ResampleFilter', βάλτε το ποσοστό 'sampleSizePercentage' στο 15%, και επιλέξτε Apply. Τα δεδομένα σας τώρα είναι ένα υποσύνολο (το 15%) του αρχικού dataset.

Βήμα 2: Μετασχηματισμός αριθμητικών χαρακτηριστικών. Διακριτοποιείστε το χαρακτηριστικό 'age' διαλέγοντας το φίλτρο 'DiscretizeFilter'. Επιλέξτε το 'first' σαν τιμή για το 'attributeIndices' (μια και το 'age' είναι το πρώτο χαρακτηριστικό), επιλέξτε τα bins να είναι 4 ή 5 (ο αριθμός των διακριτών τιμών), και όλες τις άλλες τιμές 'False'. Κάντε Apply και σε αυτό το φίλτρο.

Βήμα 3: Διαγραφή των υπόλοιπων αριθμητικών χαρακτηριστικών. Επιλέξτε το φίλτρο 'RemoveRange', βάλτε στο 'invertSelection' την τιμή 'False', και εισάγετε τα indices όλων των υπόλοιπων αριθμητικών χαρακτηριστικών (3,5,11-13). Κάντε Apply και σε αυτό το φίλτρο.

1. Στην καρτέλα *Associate* πατήστε στο *Associator*. Ορίστε το 'numRules' σε 25, και αφήστε τις υπόλοιπες παραμέτρους στις προεπιλεγμένες τους τιμές. Πατήστε 'Start'.
2. Παρατηρήστε τα αποτελέσματα. Πώς σας φαίνονται; Είναι χρήσιμα;
3. Κάποια χαρακτηριστικά ίσως να μην πρέπει να εξετάζονται ταυτόχρονα γιατί οδηγούν σε προφανή αποτελέσματα (αυτονόητους κανόνες). Από την καρτέλα *Preprocess* επιλέξτε τα χαρακτηριστικά που θέλετε να διαγράψετε και πατήστε 'Remove'.
4. Ξανατρέξτε τον αλγόριθμο APRIORI. Είναι πιο χρήσιμοι οι καινούριοι κανόνες;
5. Επιλέξτε άλλους συνδυασμούς χαρακτηριστικών για διαγραφή, και ξανατρέξτε τον αλγόριθμο APRIORI, ψάχνοντας για ενδιαφέροντες κανόνες.
6. Πειραματιστείτε με διαφορετικές τιμές του κάτω ορίου του support. Το Weka ξεκινά από το άνω όριο του support (προεπιλεγμένη τιμή 1) και με σταδιακά μειώνει το support κατά σταθερά διαστήματα (δ, προεπιλεγμένη τιμή 0.05) μέχρι το κάτω όριο. Σταματάει όταν φτάσει τον καθορισμένο αριθμό κανόνων ή όταν φτάσει το κάτω όριο του support.
7. Πειραματιστείτε με διαφορετικές μετρικές κατάταξης των κανόνων (lift, leverage, conviction).