



**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΤΕΧΝΟΛΟΓΙΕΣ ΕΥΦΥΟΥΣ ΔΙΑΧΕΙΡΙΣΗΣ
ΑΝΘΡΩΠΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ**

WEKA: ASSOCIATION RULES

([HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA](http://www.cs.waikato.ac.nz/ml/weka))

Στόχος αυτού του εργαστηρίου είναι

- η εξοικείωση με τον αλγόριθμο APRIORI για την εξαγωγή κανόνων συσχέτισης
- η εξοικείωση με τεχνικές προεπεξεργασίας των δεδομένων

1. Αποθηκεύστε σε έναν φάκελο το αρχείο *adult.arff*. Ανοίξτε το με το Wordpad, 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. Περιγραφή των χαρακτηριστικών:

- age – The age of the individual
- type_employer – The type of employer the individual has. Whether they are government, military, private, and so on.
- fnlwgt – The \# of people the census takers believe that observation represents*
- education – The highest level of education achieved for that individual
- education_num – Highest level of education in numerical form
- marital – Marital status of the individual
- occupation – The occupation of the individual
- relationship – A bit more difficult to explain. Contains family relationship values like husband, father, and so on, but only contains one per observation.
- race – descriptions of the individuals race. Black, White, Eskimo, and so on
- sex – Biological Sex
- capital_gain – Capital gains recorded
- capital_loss – Capital Losses recorded
- hr_per_week – Hours worked per week
- country – Country of origin for person
- income – Boolean Variable. Whether or not the person makes more than \">\$50,000 per annum income.

3. 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. Διακριτοποιείτε 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.

4. Στην καρτέλα *Associate* πατήστε στο *Associator*. Ορίστε το 'numRules' σε 25, και αφήστε τις υπόλοιπες παραμέτρους στις προεπιλεγμένες τους τιμές. Πατήστε 'Start'.
5. Παρατηρήστε τα αποτελέσματα. Πώς σας φαίνονται; Είναι χρήσιμα;
6. 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'.
7. Ξανατρέξτε τον αλγόριθμο APRIORI. Είναι πιο χρήσιμοι οι καινούριοι κανόνες;
8. Επιλέξτε άλλους συνδυασμούς χαρακτηριστικών για διαγραφή, και ξανατρέξτε τον αλγόριθμο APRIORI, ψάχνοντας για ενδιαφέροντες κανόνες.
9. 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.
10. Πειραματιστείτε με διαφορετικές μετρικές κατάταξης των κανόνων (lift, leverage, conviction).

*** Description of fnlwtg (final weight)**

The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:

- A single cell estimate of the population 16+ for each state.
- Controls for Hispanic Origin by age and sex.
- Controls by Race, age and sex.

We use all three sets of controls in our weighting program and "rake" through them 6 times so that by the end we come back to all the controls we used. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.