UNIVERSITY OF COLOMBO, SRI LANKA

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

IS4003/SCS4104/CS4104 – Data Analytics

Academic Year 2021 - Semester 2
Assignment 2
(Submit on or before August 20, 2021)

Your task for this assignment is to identify and perform an association rule mining task. This involves

1. Selecting an appropriate data set
2. Preparing and preprocessing the data
3. Finding rules, including appropriate parameter setting
4. Determining which of the resulting rules are interesting
5. Figuring out how the interesting rules could be useful

Select a dataset from https://archive.ics.uci.edu/ml/datasets.html, excluding IRIS and bank dataset used in the practical. This contains many data sets, not all of which are appropriate for association rules, so you'll need to do some thinking. You are also welcome to identify data from other sources, especially those that you find personally of interest.

Assignment Report

The Assignment report should contain the following:

1. Objectives: What is the domain and what are the potential benefits to be derived from association rule mining. This is high level - **not find patterns**, but what would improve because of the use of the patterns.

2. Data set description: What is in the data, and what preprocessing was done to make it amenable for association rule mining. Where choices were made (e.g., parameter settings for discretization, or decisions to ignore an attribute), describe your reasoning behind the choices.

3. Rule mining process: Parameter settings, choice of algorithm and the time required.

4. Resulting rules: Summary (number of rules, general description), and a selection of those you would show to a client.

5. Recommendations: What should the client do because of the rules discovered?

Also turn in (likely as a separate plain-text file) a complete listing of the rules found, and instructions (preferably machine-readable/executable) for recreating your results. WEKA provides several ways to do this, from command-line scripts to Explorer.

If you iterate over different attribute sets / parameter settings / etc., only turn in the rule list and scripts for your final iteration. You should include a description of the iterations, and why you needed to make changes from your initial choices, in the project description.

Marking Scheme

Marking will be based on:

- Your reasoning behind choice of data set
- Preparation and preprocessing
- Rule generation
- Choice of interesting rules
- Evaluation / use of rules
- Overall quality of report, including readability/clarity

Extra points will be given for making the problem more challenging (provided you do so appropriately - no extra credit for doing something the hard way when an easy way is available.) Examples could include implementing an algorithm other than apriori that you think will be faster than apriori on your data, or accessing data directly from a database (JDBC) rather than as comma-separated value or ARFF formats. Extending the problem to utilize other techniques like clustering, for example, would be added-value.