Databases and Data Mining

Data Mining Assignment 1

Due: Friday 7-12-2018, 23:59 CET.
Grading: This assignment will be graded from 0 to 10.
Notes:

- Groups of 1-6 students are allowed.
- You are allowed to use any software system (for example Weka, R, etc.) or even your own code (Python, C, etc.) for preparing the data, mine the association rules and presenting the results.
- Write down your **technical** report for this assignment in a *.pdf* file with the following name **"<your student number><your name>_dm1.pdf"**, e.g., **"012345janjansen_dm1.pdf",** or **"012345janjansen_678910_ansjansen_dm1.pdf"** if you are working in a team of 2, etc.
- Send this .pdf file as an attachment of an e-mail with subject **"DBDM_DM1"** to *erwin@liacs.nl* .
- Do not use more than 8 pages (A4-sized and font size ≥ 10 pt) for your report.
- Grading will be based on
    - the quality of your data mining strategy and results
    - the argumentation, validity, and clarity of your report.

Introduction

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. For example, the rule *{onions, potatoes}* → *{beef}* found in the sales data of a supermarket would indicate that if a customer buys *onions* and *potatoes* together, he/she is likely to also buy *beef*. In addition to the above example from market basket analysis, association rules are employed today in many other application areas including Web usage mining, intrusion detection and bioinformatics.
In this assignment you are asked to mine interesting association rules from one of the following datasets:

1. *The Census-Income (KDD) Dataset.*
2. *Global Terrorism Dataset*
3. *UK Traffic Accidents Dataset.*

Datasets

Please select one of the following datasets for your research:

1. The *Census-Income (KDD) Dataset* [1] was originally extracted from the census bureau database found at http://www.census.gov/ . The data set contains 40 attributes such as age, education, work class, race, sex, etc. along with an indication of whether or not that person makes over 50K a year. The main prediction task was to determine whether a person makes over 50K a year. In this assignment you are ***also*** asked to find other interesting rules. You can download the *Census-Income (KDD) Data Set* and find more detailed information about the dataset using one of the following links:
    a. https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD) (obtain the data by following the *Data Folder* link on this page) Download the ***census.tar.gz*** file, unpack it and use the ***.data*** and ***.names*** files for mining. You can ignore the ***.test*** file.
    b. A more pre-processed version of the dataset is available on https://www.kaggle.com/uciml/adult-census-income
    Clearly state in your report which dataset was used for your research.
2. The *Global Terrorism Dataset*, description and terms of use are available via the link https://www.kaggle.com/START-UMD/gtd
3. The *UK Traffic Accidents Dataset* is available via https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales

Data Mining Tools: WEKA, other …

For mining the data set you can use any mining tool(s) of your choice. For example, *Weka* would be a suitable mining tool. *Weka* is a collection of machine learning algorithms for data mining tasks that contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. All the information (i.e., software, documents and book) about *Weka* can be found using the following link and/or reference [2]: http://www.cs.waikato.ac.nz/~ml/weka/index.html

Remarks

- Note that the data has to be pre-processed. The data has to be cleaned, missing values, outliers, etc. have to be handled, and data has to be normalized, etc. For this you may have to develop your own code and/or use a suitable tool.
- After you have cleaned and selected a subset of your data (if necessary), mine association rules using different parameter (confidence, support, etc.) settings. Analyze the resulting rules and repeat the experiment with another "view" of the data given by generalizing/specializing your data according to concept hierarchies and/or by selecting different portions of the data.
- Use explicit pattern evaluation measures to evaluate the quality of your discovered patterns. Use for example the Kulzcynski measure in combination with the imbalance ratio, etc. Explain your choices!
- Assume that you as the user/miner want to obtain association rules for decision support, for understanding the data better, and/or for increasing your company's profit. Mine rules until you obtain a collection of rules that satisfy your objectives.

Report

Your technical report should start with a *title*, the *names and student numbers of the team*, an *abstract and introduction*, and at least contain the following sections with the corresponding discussions using a formal writing style:

Statistical report

- Report the mean, median, minimum, maximum and standard deviation for each of the numerical variables.

Code Description

- *Describe* the code that you used/wrote. Remember to *always* acknowledge any sources of information/code you used. Any actual code/scripts you wrote should be sent in a separate zip-file.

Experiments:

- Describe in each case the objectives of your analysis: Is it to understand the data better? If so, what is it about the data you want to understand? Or is it for decision support? If so, what decisions you need to make based on the data?
- For each experiment you should describe:
    - Instances: What data did you use for the experiments?
    - Any pre-processing done to improve the quality of your results.
    - Your system parameters.
    - Any post-processing done to improve the quality of your results.
    - Analysis of results of the experiment and their significance.

Summary of Results

- What was the best collection of association rules that you obtained? Describe. Discuss the strengths and the weaknesses of your mining and evaluation methods. What are your conclusions?

References

[1] Ron Kohavi, *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

[2] Daniel T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Chapter 3, John Wiley, Chichester, 2004.

[3] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufman, 2005. (3rd Edition 2011, 4th Edition 2016)