Data Mining- Assignment 3	
Due	1400/02/14
Grading	This assignment will be graded from 0 to 10.
Grading Measures	• The quality of your data mining strategy and results • The argumentation, validity, and clarity of your report
How to deliver:	• Groups of 1 students are allowed. • Write down your report to this assignment in a .pdf file with the following name "<your student number><your name>.pdf", e.g., "012345JanJansen.pdf". • Upload your result to LMS system.

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 application areas including Web usage mining, bioinformatics, etc. In this assignment you have to use the association rule mining module of the Weka system to mine association rules from the "churn" and "marketBasket" datasets.

WEKA

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) you need about the weka could be find on following links:

http://www.cs.waikato.ac.nz/~ml/weka/index.html
Data Mining: Practical Machine Learning Tools and Techniques (Chapter_9)

Datasets

1. Churn Dataset

The churn dataset deals with telecommunications customers and the data pertinent to the telephone calls they make. Churn is a term used to indicate a customer leaving the service of one company in favor of another company. The data set contains 20 variables worth of information, along with an indication of whether or not that customer churned (left the company). You can download the churn dataset and also find more detail information about the dataset in the following links:

Discovering knowledge in data: an introduction to data mining(Chapter 3)
http://dataminingconsultant.com/DKD.htm
http://www.sgi.com/tech/mlc/db/

2. Market Basket Dataset

Given:

- a set I of 11 items: {*fruitveg, freshmeat, dairy, cannedveg, cannedmeat, frozenmeal, beer, wine, softdrink, fish, confectionery*}.
- a database of 1000 transactions T s.t. $T \subseteq I$.

Find:

interesting association rules that explain customer behavior.

Guidelines

· Due to the representation of frequent itemsets in Weka, this system may run out of memory when mining datasets with as few as a dozen attributes. Run several experiments with your data and the system varying the parameters until you obtain a collection of association rules that represent your data well.

· Use the Weka system to mine the association rules as well as for preparing the data and presenting the results. Code by yourself any functionality that you need for manipulating the data and that is not offered in the Weka system.

· You can restrict your experiments to a subset of the dataset if Weka cannot handle the whole dataset. But remember that the more representative the association rules you mine from the data, the better.

· 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 other ”view” of the data given by generalizing/specializing your data according to the concept hierarchies and/or by selecting different portions of the data.

· Assume that you as the user/miner you 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 satisfies this objective.

Reports

Your report should contain the following sections with the corresponding discussions:

- 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 acknowledge any sources of information/code you used.
- Experiments:
  - Describe what the objective of your analysis is. Is it to understand the data better? If so, what 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 ran 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 project.