CI6227 Data Mining Project Assignment

Due Date: 31 Oct, 2015 (Saturday)

Reminders

- You are NOT allowed to COPY code/report directly from others / Internet (unless specified for special cases). Any plagiarism case will be seriously punished!
- For late submissions, a penalty of **1 mark** per day will be applied after the deadline. The assignment will not be accepted if more than **7-day delay**. Please remember to submit your assignment before the deadline.
- Operating System Platform: Windows / Linux

Marking Scheme (Total: 20 marks)

Project Objective

This project aims to provide an opportunity for students to apply data mining techniques and tools to solve real-world problems. The students are expected to practice hand-on skills for how to perform a real-world KDD (knowledge discovery from data) task from the beginning (data pre-processing - data collection, cleaning, etc), to model building and the final stage (data post-processing - evaluation and presentation, etc).

Assignment Grading Criteria

- **Problem Innovation**
  How innovative is the problem you are going to address? Any existing work in data mining community has addressed the similar problems? What is the difference between your problem and the problems studied by existing work?
- **Novelty of Methodology**
  Have you proposed a new method to solve the problem? What is the novelty of the proposed method in comparison to other existing work in literature?
- **Technical Depth**
  How challenging is your selected problem? How difficult is your proposed methodology/solution? Is it trivial to implement the proposed idea? What kinds of tools/knowledge/code required in order to implement such a solution?
- **Significance of Experimental Results**
  Are your experimental results significant? Can your results answer your research question or achieve the objectives of your application?
- **Project Report/Draft Paper**
  This is to evaluate the quality of your project report, including the organization, presentation, and comprehensiveness of the write-up.

What should be included in your project report?

Cover page: your group ID, your team members and their student ID

- **Abstract**
  (use no more 300 words to summarize your whole project)
- **Problem Description**
  - Motivation
  - Problem Definition
  - Related Work
- **Approach**
  - Methodology
  - Algorithms
- **Implementations**
- **Experimental Results and Analysis**
  - Experimental Setup
  - Comparison Schemes
  - Results and Analysis
- **Discussions of Props and Cons**
- **Conclusions**
  - Summary of project achievements
  - Future Directions for improvements
- **Appendix: (optional)**
  - **Data sets** *(if you collect your own data sets)*
  - **Source Codes** *(if you implement your own codes)*
  - **Implementation Guidelines** (instructions on using any tools)

Project Topics for Data Analytics and Mining

The following are some example topic for your consideration. However, you are strongly encouraged to propose your own new data analytics and mining problem.

1. Market Basket Analysis via Associate Rule Mining

Project P1: PEP marketing analysis (Dataset: D1 PEP market.csv)

Description: The marketing department of a financial firm keeps records on customers, including demographic information and number of type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. Based on this store of prior experience, the managers decide to use data mining techniques to build customer profile models. In this particular problem we are interested only in deriving association rules from the data.

Task: perform association rule discovery on the data set.

Project P2: Retail Market Basket Analysis (Dataset: D2 retail MBA.data)

Description: The data set contains the retail market basket data from an anonymous Belgian retail supermarket store. For the data description, please refer to http://fimi.ua.ac.be/data/retail.pdf. As a data scientist, you are going to do a market basket analysis.

Task: Conduct a market basket analysis for the supermarket.

2. Summarization via Cluster Analysis

Project P3: USDA plants' summarization (Dataset: D3 USDA plants.data)

Description: The United States Department of Agriculture (USDA) collected a dataset containing all plants (species and genera) and the states of USA and Canada where they occur. They want a summarization of the plant information. The data is in the transactional form. It contains the Latin names (species or genus) and state abbreviations. Each row contains a Latin name (species or genus) and a list of state abbreviations.

Task: Give USDA a summarization of their plant database.

Hint: Hämäläinen, W. and Nykänen, M.: Efficient discovery of statistically significant association rules. Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), pp. 203-212.

Project P4: Synthetic Control Chart Time Series (Dataset: D4 synthetic control.data)

Description: This dataset contains 600 examples of control charts synthetically generated by the process in Alcock and Manolopoulos (1999). Each row denotes a control chart. There are six different classes of control charts:

Rows	Classes
1-100	Normal
101-200	Cyclic
201-300	Increasing trend
301-400	Decreasing trend
401-500	Upward shift
501-600	Downward shift

Task: Summarize the data.

Hint: Alcock R.J. and Manolopoulos Y. Time-Series Similarity Queries Employing a Feature-Based Approach. The 7th Hellenic Conference on Informatics. August 27-29. Ioannina,Greece 1999.

3. Challenge the top data mining experts

The followings are the KDD 2014, KDD 2015 and ICDM 2014 paper lists. These papers have good quality.

http://www.kdd.org/kdd2014/program.html

http://www.kdd.org/kdd2015/program.html

http://icdm2014.sfu.ca/program_accepted_papers.html

You can do the followings to check if you can do better than the top data mining experts.

1) Find a paper that your group members are particularly interested in
2) Try to see if you can get their data sets (download or check with authors). If not, you may find other papers
3) Explore your research to get better results.

4. Any New real-world data mining tasks defined by yourselves

You are strongly encouraged to propose new real-world data mining tasks, which could be different or twisted from the existing problems.

You can either use existing data sets or your own data sets.

Submission Guidelines

What To Submit:

1. Project Report: a file called **ProjectReportGroupXX.pdf.** Please show your group members' names and IDs in the cover page of your project report. The report should elaborate clearly each members' work distribution in the first page – there could be small differences on the scores for students within the same group.
2. Source code (if any)
    - Well-commented code
    - Include Makefiles if necessary
    - Remove the binary executable program if any
3. A README file. Please name it **README.txt** This file should include three sections:
    - Your group ID and group member names
    - Detailed instructions on how to re-produce your results using any toolbox.
    - If you write your own code, please state clearly the program language used and instructions on how to run your programs.
4. Presentation Slides, with audio/video explanation

The page limit of the report is 15 pages (written in Times New Roman, font size 12). The over-length case may be penalized. Please do not simply attach your source code in the report. However, if necessary, you can show some code segment or pseudo code to describe your key algorithm.

Submission Instructions

1. Please package all of your files (including report "ProjectReportGroupXX.pdf", the README.txt file, presentation slides with audio explanation, and source code if any) into a ZIP file, named as "ProjectGroupXX.zip", where XX is your group ID.

2. Submit the package file with the Subject "**CI6227 PROJECT SUBMISSION GROUP XX**" to the following course E-mail: xlli@ntu.edu.sg, where XX in the email subject is your group ID. If Subject title is not correct, then it may not be received

3. If your package file is too big (>10M), then you should upload them into a website so that I can download

Appendix A: Data Analytics and Mining Tools

- **WEKA Toolbox**
- **R**

Appendix B: On-line Data Set Repository for Data Mining

- **UCI machine learning database**
  http://archive.ics.uci.edu/ml/