CISC520 Data Engineering and Mining

Final Project Instruction

1. Choose data mining problem and data set

For this project, you must choose your own dataset. It can be one found from an on-line source, one of your own, or one of the ones from the UCI repository (http://archive.ics.uci.edu/ml/). A list of additional dataset sources is provided at the end of this document. If you would like to use data collection API to curate data, the attached document provides an example of using R to collect Twitter data.

Some rules/tips about choosing data sets:

a. Do not choose the datasets that we have already analyzed in class.
b. It should not be a small or made-up dataset. For this semester, "small" is defined as fewer than 1000 examples in the dataset.
c. Choose a data set that does not require excessive data preprocessing.

2. Experiment design

Define a problem on the dataset and describe it in terms of its real-world organizational or business application. The complexity level of the problem should be at least comparable to one homework assignment.

The problem may use at least TWO different types of data mining algorithms that we have studied this semester such as Classification, Clustering and Association Rules, in an investigation of the analytics solution to the problem.

This investigation must include some aspects of experimental comparison: depending on the problem, you may choose to experiment with different types of algorithms, e.g. different types of classifiers, and some experiments with tuning parameters of the algorithms. Alternatively, if your problem is suitable, you may use multiple algorithms (Clustering + Classification, etc.). If there are a larger number of attributes, you can try some type of feature selection to reduce the number of attributes. You may use summary statistics and visualization techniques to help you explain your findings.

3. Final project paper

To complete this project, write a final report that conforms to general research paper format. See (Pang, Lee, and Vaithyanathan, 2002) as an example. Your report should be within 6 pages, 1 inch margin on all sides, and at least 12 point Arial or Times New

Roman. Remember that your project paper serves as the tour guide for your readers to be able to repeat your data mining process and discover the same patterns as you did. It is very important to cite and paraphrase relevant work appropriately.

Submit your final project report to the Moodle. This is the final report that will be graded.

References

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP 2002*, 79-86.

Additional Sources of Data Sets for Final Project

1. http://socialcomputing.asu.edu/pages/datasets (Social Computing Data Repository ASU)

2. http://snap.stanford.edu/data/index.html (Stanford Repository)

3. https://www.kaggle.com/datasets
   Amazon Fine Food Reviews
   World Food Facts
   Reddit Comments
   US baby names

4. https://www.yelp.com/academic_dataset
   Dataset containing the reviews of business

5. http://openflights.org/data.html
   Airline, airport data

6. http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

   - http://archive.ics.uci.edu/ml/datasets/Internet+Advertisement (Internet Advertisement Dataset)
   - http://osmot.cs.cornell.edu/kddcup/datasets.html (Particle Physics Dataset)
   - http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/ (4 university dataset)

7. http://www.kdnuggets.com/datasets/kddcup.html
   Contains KDD cup datasets

8. http://www.kdnuggets.com/datasets/index.html
   Contains links to multiple datasets.