SEIS 734 – Data Mining

Project Descriptions and Requirements

Project Plan (Check Class Schedule for Due Date)

The report must include the following information:

- *Name* for ***each team member***
- *Project Title*
- *Problem Statement*.
  - A detailed description of a data mining problem is required (250 – 350 words, excluding surveys).
  - Please indicate the type of problem your group intends to solve: association, classification, clustering, temporal or spatial analysis.
  - Specify the goals of your mining project and its potential benefits for the users.
  - Explain the ***significance*** and ***innovation*** of the problem your team intends to solve.
  - Survey existing methods with ***annotated references***.
- *Data Selection*.
  - Please indicate where does your team ***find*** or ***generate*** the target dataset.
  - Describe your dataset ***by using one or few E-R diagram***.
  - Provide a short description of each field (attribute), ***including*** its use in the data mining process (input / target / none).
- *Data Pre-Processing*. Describe the operations your team has performed on the dataset to re-organize it, e.g. database queries and joins, transformations, encoding nominal data, deriving new attributes from the stored ones, treating missing values, etc.
- *Datasets*.
  - If your dataset is too big, please do **NOT** send it via email.
  - ***Zip*** original and processed datasets with pre-processing programs.
  - If your datasets are not stored in an Excel file, ***please convert your datasets into this format (Excel) and submit them with other datasets***.
  - If your team creates programs to generate datasets, please also zip those ***data-generation programs with extensive comments and documents***.

Final project report (Check Class Schedule for Due Date)

The report will include the following sections:

- *Algorithm.*
  - Explain your choice of specific algorithms.
  - List the main ***assumptions*** of each algorithm and discuss their applicability and ***weakness*** to the problem.
  - Zip ***all data-mining programs with extensive comments and documents***.
  - Explain what aspect of the project is ***innovative, interesting or difficult***. ***Please compare every point your team tries to make with your annotative references***.
- *Tools Selection.*
  - Describe the software tools your team has selected.
  - Describe the reasons of your selection.
  - The ***source*** of each tool should be clearly indicated so the instructor can verify your results.
  - Provide / include all the programs you developed using those tools.
- *Data Mining Results.*
  - Represent the complete results of each algorithm as rule lists, tables, graphs, trees, or in any other appropriate (and easily understandable) form.
  - Provide all the necessary explanation.
  - Explain the meaning of results from user viewpoint.
  - Explain why the discovered knowledge is non-trivial, interesting and potentially useful.
  - Explain pitfalls you experienced during the mining process.
  - You may refer to the references listed in the syllabus.
- *Comparison of Algorithms.*
  - Compare (qualitatively or statistically) between different results (e.g. ***accuracy, performance***)
  - ***Explain the differences*** in the results your team observed.
  - If you use datasets from KDD CUP, explain how does your method(s) compare with other methods, such as ones used by the KDD CUP winner.

- Suggest possible ways of improving your results.

Submission Guidelines

- Please copy all the zipped reports onto a digital media (i.e. CD / DVD / USB/Cloud Drive) and submit to the instructor in class. Please do **NOT** submit your report via email as it will likely jam the instructor's e-mail box.
- **Project Plan**.
  - Any ***confidential information*** (names, SSNs, etc.) may be omitted or replaced with codes.
  - **Zipping all files together is highly recommended!**
- **Final Report.**
  - The source and the executable created by your team should be attached in the report ***with a simple user manual***.
  - The data files used for the analysis should be sent again (in the Excel format).
  - ***Slides for the final presentation***.
  - Please remember to use Zip (like in your project plan)!
- ***File Names***. Use your project title + 1 or 2 as the name of your Zip file. ***1 is used for the project plan, and 2 is used for the final report.*** **For example**: *AirplanCollision1.zip* for the project plan and *AirplanCollision2.zip* for the final report.

Following information is for reference only. Many links may not exist anymore. Please google to find dataset for your project.

Links to KDD CUP datasets

(Thanks to *Nga Nguyen* for preparing this table)

Type	Name	URL-link	Note
Classification	Japanese Vowels	http://kdd.ics.uci.edu/databases/JapaneseVowels/JapaneseVowels.html	
	Microsoft Anonymous Web Data	http://kdd.ics.uci.edu/databases/msweb/msweb.html	
	The Insurance Company Benchmark	http://kdd.ics.uci.edu/databases/tic/tic.html	Regression and Description
	KDD Cup 2002	http://www.biostat.wisc.edu/~craven/kddcup/	
	KDD Cup 2001	http://www.kdnuggets.com/datasets/kdd-cup-2001k.html	
	KDD Cup 1999	http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html	
	KDD Cup 1998	http://www.kdnuggets.com/meetings/kdd98/kdd-cup-98.html	
	KDD Cup 1997	http://www.kdnuggets.com/datasets/kddcup.html#1997	Data set not found
Association Rules	KDD Cup 2000	http://www.kdnuggets.com/datasets/kdd-cup-2000k.html	
Clustering	Synthetic Control	http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.html	

Other datasets

- UCI Machine Learning Repository
- Data Mining Gateway
- Web Traces
- UCI KDD (Data Mining) Data Set
- **EachMovie** Movie Voting Data from Digital/Compaq Research Lab
    1. EachMovie Project Web Site
    2. A subset of the data for class use, extracted by Henry
        1. Henry Zhang's Description of the subset, and
        2. The subset itself, where the data have been neatly converted to MS Excel formats
    3. A research paper from Microsoft Research on using the data for evaluation

Some Data Mining Tools

- Cubist (Linux, instance-based learning, 1-2)
- DBMiner (relational databases, Dr. Cook will present) Han et al, KDD, 250-255, 1996 IBM Intelligent Miner / Text Miner
- IBM Intelligent Miner / Text Miner (Windows NT or Solaris, a suite of data mining algorithms, 2-3)
- JAM (Java, meta learning over distributed databases, 2-3)
- PolyAnalyst Lite (Windows, statistical analysis, regression, prediction, 1-2)
- TextAnalyst (Windows, text analysis, 1-2)
- Timbl (Linux, memory-based learning, 1-2)