ECE657A: Data and Knowledge Modeling and Analysis
Project Description

There are two broad types of projects: (a) Application-oriented projects: You have a problem, perhaps in your field of research, that you like to analyze using the concepts and algorithms of this course, and (b) Algorithm-oriented projects: You select an interesting data mining technique that you want to learn more about it, illustrate and compare its performance against others.

You may choose to develop or extend your own data mining techniques, or simply apply existing techniques (they do not necessarily need to be covered in the class) to data. Similarly, the dataset may be a pre-existing one, or part of your work could be to collect and pre-process the raw data needed for a new dataset.

Note that pure literature surveys are *not* acceptable. There must be a hands-on, experimental and comparative element in your work.

Projects can be performed in groups of at most three. The project consists of proposal, presentation, and a written report. This document explains what is expected in each of these milestones. Marking schemes for the proposal, presentations and final report will be available on the course page in the LEARN system. Combined, these elements constitute 35% of the total course mark.

1 - Recommended Topics

- Application-oriented projects

Sentiment analysis from social media	Topic discovery and analysis from text
Mining of scientific publications	Political event data analysis
Financial data mining	Purchasing behaviour analysis and recommendation systems

- Algorithm-oriented projects

Cost-sensitive classification	Kernel-based clustering
Semi-supervised classification	Clustering ensembles
Frequent pattern mining	Supervised clustering

2 - Proposal

Your proposal should contain:

- Description of the project. You should clearly mention your main goal: is it classification, clustering, or mining association rules?
- A comprehensive review of 3 to 4 well-recognized research papers.

- A paragraph to discuss the expected challenges / difficulties.
- A sketch of your planned approach, algorithms, preprocessing methods, evaluation metrics, etc.
- Description of datasets that you plan to use. It should include a link and a brief description about the properties of the data, such as its features, instances, preprocessing techniques, etc. If you are going to use your own dataset, then a description of its source and preprocessing steps is needed.
- List of key references.

Your proposal should be no longer than 2 pages, and submitted as a PDF file via the LEARN dropbox. It will be graded. If not being approved, you will need to revise it based on our feedback

3 – Presentation

Your twenty minute group presentation would be via projected slides. Please allocate 3-5 minutes for questions. Each presentation should include the following:

- Introduction: Basic definitions, background and terminology used.
- Literature review: Based on papers from your literature search, summarizing common variants of the method and data mining applications being used and the achieved results as claimed in the literature.
- Description of the goal and the use of method in your project, such as types of data mining, representation of the input, training requirements, output representation.
- Report and analyze your comparative experimental results.
- A summary of your work: new findings and potential future directions.
- List of key references.

A copy of presentation slides is to be deposited to the LEARN dropbox before your presentation.

4 – Report

Your report should be in Springer LNCS format, as if you were planning to submit it as a conference paper. The report should be a maximum of ten (10) pages. See Springer's `Information for LNCS Authors' page with LATEX templates and a sample PDF (there is also a Word template but Latex is suggested).

The report should include the following:

- Introduction to methods selected and task applied to.
- Brief review of literature on the selected methods and their application to similar problems.
- Description of the method selected with details on the options and parameters.
- Implementation: Software used, data structures, program structures, data representation and any special set up needed. Please don't put code, only abstract descriptions and diagrams.

- Testing: Test cases on the selected datasets and evaluation of the performance in comparison with base line methods (for example, K-means for clustering, KNN for classification, PCA for data reduction).
- Discussion of results and conclusions: Provide a discussion on the use of the method and its suitability and/or limitations, and discuss the effect of each parameter on the trade-off in performance.
- References to relevant literature you consulted about the data domain or the methods used.

The PDF format of report should be submitted via the LEARN dropbox, attached with related code, datasets, and other supporting materials.

5 - Timelines

Deliverable	Issue	Due	Grade
Proposal	Feb 1	Feb 22	5%
Proposal feedback & revision	Feb 29		
Presentations		March 21,28	10%
Peer Review Participation			5%
Report		April 4	15%

Late submissions up to 3 days are accepted with the penalty of 10% per day. You should not submit a work that you have performed for other class, or have already been developed for your thesis.