CSE 8331 Spring 2008 FINAL PROJECT
DUE DATES:
Project Choice: 3/31/08
Presentation Slides: 4/21/08 8:00 am
Inclass Presentations: 4/21-28/08
Writeup: 5/10/08 8:00 am

General Description:

The objective of this project is to give each student the chance to perform some original applied research. The project involves surveying previous work, performing experimental analysis comparing at least two different algorithms, writing up results, and presenting results in a classroom presentation.

It is hoped that at least a few of these projects will be used as the basis for publications.

Requirements:

You are to compare the performance of **two different algorithms** using **a dataset of your choice**. Each of the algorithms is to be trained and validated using the chosen datasets You decide how to take the existing dataset and divide into training and validation sets. You then need to compare the performance of the algorithms using **two different metrics** you have chosen.

Grading:

Grading for the projects is based on a total of 100 points and make up 40% of each student's grade in the course. Grading will be performed on the following basis:

Project Identification (10 pts) Due 3/31/08:

- Each student is to submit a brief (1 paragraph) writeup of the project topic and how the experiments will be performed.

Presentation Slides (10 pts) Due 4/21/08 8:00am:

- Each student is to submit the powerpoint (or other electronic) version of presentation.

Presentation (30 pts) 4/21-28/08:

- Give a 20 minute presentation in class between 4/21 and 4/28. Your presentation should provide a summary of your implementation and comparison of your algorithm with the existing one. Grading is as follows:
  - Project overview (5pts)
  - Overall implementation approach & objective (5pts)
  - Overview of second algorithm (5pts)
  - Identification of dataset and metrics (5pts)
  - Implementation overview (tool or language, platform) (5pts)
  - Preliminary results obtained (5pts)

  Note that the results at this point in your study may be preliminary. But you must have implemented enough to present some comparison of the two algorithms. Presentation dates will be assigned based on students' requests in a FCFS manner.

Submission (50 pts) 5/10/08 8:00 am:

You are to submit a paper in IEEE format much as you would a submission to a conference. (http://www.engr.smu.edu/~mhd/8331sp08/ieee.doc). The paper should be about 10-15 pages in length. Although the actual structure of the paper is up to you, it must contain the following parts:

- Problem statement and objectives of paper (5pts)
- Related work (5pts)
- Detailed description of two algorithms being compared (10pts)
- Experimental setup including data, metrics, platform, and experiments performed (10pts)
- Discussion of results of experiments including figures/graphs (10pts)
- Conclusions (5pts)
- Bibliography (5pts) (Note that the bibliography should contain some related work needed to do a brief survey for the related works section. The bibliography should include a minimum of 10 references.)

Anyone found plagiarizing at any step in the process will receive a grade of 0 on the project.

Students are given the choice of two different projects, and alternative projects will be allowed if approved by Professor Dunham.

Alternative Project A – Anomaly Detection in Streaming Data:

This project requires that you compare an anomaly detection (rare event detection) algorithm of your choice against an anomaly detection algorithm implemented using EMM. Recall that EMM (Extensible Markov Model) is a stream modeling technique developed at SMU. Previous published studies have shown the effectiveness of EMM in anomaly detection. You are to compare EMM to an existing anomaly detection algorithm of your choice using data of your choice.

The EMM algorithm and an anomaly detection algorithm have been developed, and a Web site will be made available to you so that you do not have to implement EMM directly.

Related Sites:

- EMM (http://engr.smu.edu/cse/dbgroup/emm.html)
- Rare Event Detection using EMM (http://engr.smu.edu/cse/dbgroup/rare.html)
- "Extensible Markov Model" by Margaret H. Dunham, Yu Meng, and Jie Huang, ICDM'04, (http://engr.smu.edu/~mhd/8331sp08/emm.pdf)
- "Rare Event Detection in a Spatiotemporal Environment," Yu Meng, Margaret Dunham, Marco Marchetti, and Jie Huang, *Proceedings of the IEEE Conference on Granular Computing*, May 2006, pp 629-634, , (http://engr.smu.edu/~mhd/8331sp08/rare.pdf)
- A new Web site is being created which will provide access to EMM code. You are to use this site to execute the EMM algorithm

Alternative Project B – Biodegradation Prediction:

This is an extension of the project performed in CSE 7331 Fall 2007 (http://engr.smu.edu/~mhd/7331f07final.html):

The study of biodegradation of compounds in nature is an important research area in Environmental Engineering. However, the accurate prediction of which compounds actually biodegrade and the speed with which they degrade is a difficult problem yet to be solved. Previous prediction algorithms tend to rely on structural properties of compounds to do the prediction and create somewhat simplistic linear regression models. (Part of the problem with

previous prediction algorithms is the lack of large amounts of reliable data. Unfortunately this will also be a problem with this project.)

Your project requires the development and comparison of data mining classification algorithms to predict biodegradability of compounds. A new Web site is being created which will provide access to biodegradation data as well as results of previous prediction algorithms. You are to use the data on this site to implement your classification algorithm and use the included prediction results for comparison. You may implement any classification/prediction algorithm you choose.