Computer Science- CpSc 405/605

DATA MINING AND DATA ANALYSIS

Slippery Rock University

*Fall 2016*

Term project: (35% of the grade)

Proposal due date: Monday, Oct. 10th. 8:00 AM

Final submission due date: Monday, Dec. 12, 5:00 PM

The project can be done alone or in groups of 2 students.

**Project description:**

The project is the core component of the course, accounting for 35% of the grade. Students are required to work on a research project that involves writing code, running experiments, and submitting a project report.

**Students have to pick one of the following two options:**

1. Survey of Data Mining algorithms for solving a specific problem with experimental evaluation on several datasets.
2. Develop your own algorithm for solving a given problem.

**Option 1:**

1. **Survey of Data Mining algorithms for solving a specific problem with experimental evaluation on several datasets.**

The goal of this project is to analyze the behavior of Data Mining algorithms for solving a specific problem with experimental evaluation on several datasets, you should use the Weka tool to run the algorithms and demonstrate their effectiveness on real data. You can download the Weka tool, and find the documentation in the following link: <http://www.cs.waikato.ac.nz/ml/weka/>.

For the datasets, you need to find at least two real datasets, make sure that the datasets suitable for the problem you are trying to solve, many real datasets are available in <http://archive.ics.uci.edu/ml/> (avoid datasets that do not have any description). To use the datasets in Weka, the format of your dataset should be ARRF. If your dataset does not have the ARRF format, write a program that generates the ARRF format. You can find information about this format in <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>

For the algorithms, you should compare and analyze the performance of at least two algorithms (could be any classification or clustering algorithms), you need to decide which algorithms suitable for the problem you are trying to solve.

After applying the picked algorithms on the datasets, you need to analyze the results and generalize the findings. For example discuss questions like: which choices lead to best results? Why Algorithm A outperforms Algorithm B? How reliable are the results for each algorithm? What evaluation measure did you use to compare the results and why? Justify any decision you make at the end of the study… etc

Here are some comparative study research papers that you can look into to give you a brief idea about how to conduct a comparative study research:

* [**http://research.ijcaonline.org/volume74/number5/pxc3889673.pdf**](http://research.ijcaonline.org/volume74/number5/pxc3889673.pdf)
* [**http://research.ijcaonline.org/volume75/number16/pxc3890862.pdf**](http://research.ijcaonline.org/volume75/number16/pxc3890862.pdf)
* [**http://www.computerscijournal.org/vol8no1/a-comparative-study-of-classification-techniques-in-data-mining-algorithms/**](http://www.computerscijournal.org/vol8no1/a-comparative-study-of-classification-techniques-in-data-mining-algorithms/)
* [**http://www.ijcst.com/vol22/2/milan.pdf**](http://www.ijcst.com/vol22/2/milan.pdf)

**Option 2:**

1. **Develop your own algorithm for solving a given problem.**

The goal of this project is to propose a new Data Mining algorithm for solving a specific problem, and demonstrate the effectiveness of the proposed algorithm on a dataset. You should implement your algorithm using R, Python or any programming language. For the dataset, you can use a dataset from the UCI repository <http://archive.ics.uci.edu/ml/> (avoid datasets that do not have any description), make sure that the datasets suitable for the problem you are trying to solve.

After you implement your proposed algorithm and run it on a dataset, you need to pick an existing algorithm, you can use Weka, and run the algorithm on the same dataset and then compare the results to your proposed algorithm’s results to measure the effectiveness/accuracy of the proposed algorithm.

You can download the Weka tool, and find the documentation in <http://www.cs.waikato.ac.nz/ml/weka/>.

To use the datasets in Weka, the format of the dataset should be ARRF. If your dataset does not have the ARRF format, write a program that generates the ARRF format. You can find information about this format in <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>

**Research topics that can be used for your project:**

* Using Clustering as a pre-process step for improving an algorithm performance.
* Outliers Detection.
* Multi-objective clustering.
* Market-basket analysis on a real data
* A Music Recommendation System
* A Data Mining Based Approach to Determining Causal Associations between Drugs and Conditions
* Book Recommendation System
* Using Twitter tweets' sentiments to predict stock price change
* Text mining (Finding interesting patterns from unstructured textual corpus)
* Cancer prediction from biological data (classification task)
* Clustering gene expressions

You can choose your project outside of this list as long as the main theme of the project fits the course.

**Deliverable for both options 1 and 2:**

1- Project Proposal

2- A comparative study report explaining the findings

3- The source code of any algorithm you wrote, if you did.

1. **Project Proposal**

For the project proposal, you are required to submit at least a 1-page write up explaining the motivations for the problem you are addressing and the different approaches proposed in the literature. While writing your proposal, you should cite at least 3 relevant references and include the followings:

1. Cover page (Project name, and Group members)
2. What are you trying to do
3. What is the problem and why it is important
4. What are the existing techniques
5. What datasets will you be using
6. What measures of assessment
7. What are you going to submit in your report
8. **A research paper explaining the findings:**

The paper must be double-column in IEEE format (see IEEE template) and limited to 8 pages

Typical sections include:

Abstract

Introduction

Related work

Methodology

Conclusion

Guidelines on writing a research paper.

[**http://writing.wisc.edu/Handbook/PlanResearchPaper.html#intro**](http://writing.wisc.edu/Handbook/PlanResearchPaper.html#intro)

1. **The source code:**

Option 1: submit the source code of any algorithm you wrote, if you did.

Option 2: submit the source code for your proposed algorithm.

**This project counts 35% of your grade. The breakdown is as follows.**

  Project proposal - 15%

Final Paper - 20%