AUT UNIVERSITY COMPUTING + MATHEMATICAL SCIENCES

# MASTER OFCOMPUTER AND INFORMATION SCIENCES

COMP809

**Data Mining & Machine Learning**

## ASSIGNMENT TWO

## DATA MINING PROJECT

**Semester 1, 2019**

Due**:** Wednesday June 12 at midnight

Weighting**: 70%**

This assignment represents the major piece of work on this course, and accordingly carries 70% of your course mark. The major requirements involved are: *experimentation* with one or more Mining packages (not necessarily restricted to the ones used in this course) and *post experiment analysis*.

The results and analysis of your investigation will be presented in written form. The written report will need to include the following:

- Full description of the application domain, including a clear statement of the overall purpose of the Mining exercise that you have performed.

- A description of the data set that you targeted, together with any transformations that were required to pre-process the data for the Mining exercise. Ensure that you present samples of both raw data and pre-processed data.

- An explanation of how your selected Mining algorithms work (no more than a page for each algorithm). Your description of algorithms should not just be a passive description of how the algorithms work but should also **include an examination of its strengths and limitations.**

- A rationale (justification) for selecting the algorithms (you must select at least **three**, **at least one of which must be an algorithm not covered in the lecture course**), together with your reasons for rejecting the other algorithms that you initially considered. Note that your justifications can cite literature as supporting evidence but these **must be backed up by your own reasoning**.

- Detailed experimental study of the *performance* of the selected Mining algorithms on your selected data set. Include actual outputs from your Mining software as supporting evidence. **You will need to give your own definition of "performance" which may include a variety of factors. Your performance measures must take into account aspects of accuracy and time.**

  In your experimental study you should run several different experiments that will consist of various combinations of algorithms, and pre-processing methods (e.g. different methods of feature selection, different combinations of parameter values). For example, if you experiment with *a* different algorithms, *f* different feature selection methods and *p* different sets of parameter values and *b* different boosting methods then you will run a total of t=a*f*p*b different experiments. The number t can be quite large depending on your choice of a, f, p and b. Thus, as part of your report you will need to **produce an experimental plan describing what strategy you used to keep the total number of experiments down to a manageable number whilst not sacrificing performance.**

- An analytical (this can include statistical methods) comparison of the performance of the algorithms, together with an explanation of the superior performance of the winner. You may use the Experimenter module in Weka for this purpose. Your analysis should also include suitable visualizations (model diagrams, PRC curves, whatever is appropriate) that compare the performances of your winner and runner-up algorithms. **Your winner should then be compared to any significant (data mining) work previously undertaken on the data set you selected (if any).** In your experimental study you will have defined a number of different performance measures and these measures should (a) be used on their own and (b) combined into a single measure. To combine several measures into one use a **linear weighted model, with weights to be supplied by yourself, backed up by suitable justification.**

Your **report size should not exceed 15 pages** (with standard single spacing and a font size of 12). Summarize results in the form of tables, graphs and other methods of visualization. There is no need to include detailed Weka, R or Python screen screenshots.

## Choice of Datasets

A choice of 4 datasets is available, from which you must pick ***one*****. No other dataset should be used.**

*Each of these datasets have their own challenges. I will point out the challenges involved with each dataset but it is up to you to study these datasets carefully and make the right choice depending on your interests and pilot tests done on them.*

The datasets that are available for use are:

1. **Secom** (1567 samples, 591 features)

   http://archive.ics.uci.edu/ml/datasets/SECOM

   This dataset is obtained from sensor readings operating taken from a large semi-conductor manufacturing facility. The measured signals contain a combination of useful information, irrelevant information as well as noise.

   Each signal is to be considered as a feature and thus one of the primary challenges in this dataset is to perform feature selection. Apart from feature there are two other challenges to be overcome in this dataset: (1) dealing with a large percentage of missing values (2) class imbalance.

   Each data instance represents a single production entity with associated measured features and a class label that represents a simple pass/fail test, where –1 corresponds to a pass and +1 corresponds to a fail and the data time stamp is for that specific test point.

   The dataset will need to be segmented into train and test splits (do NOT use 10 fold cross validation as it is an imbalanced dataset) in order to lift the rate of detection of the minority (+1) class. However, also keep in mind other factors in your performance metric. You may use either Weka or Python platforms for mining this dataset.

2. **Daily and Sports Activities Classification** (9120 samples, 5625 features)

   http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities

   This dataset represents a collection of data related to sports and other activities. The major challenge in this dataset is feature selection as there are 5625 features and so the feature selection challenge here is a bigger one than with the “Secom” dataset above. However, it does not have the imbalanced data problem that exists with the “Secom” dataset.

   To solve the feature selection issue you can use feature selections methods (from Weka, R, or Python) but you may also need to look at creative ways of reducing the number of features by using methods such as aggregating feature categories into a single feature by using their arithmetic average (or some other statistical summary measure) taken over the category. *Feature selection is critical in this dataset as your computer could run out of memory and/or your execution time could be very long unless the number of features is reduced to a manageable number.*

   This is a classification problem with 19 classes, denoting the number of activities. The data is organized into a hierarchy and some scripts in R or Python or some other language will be required to flatten the data so that it can be represented as a single file.

   The hierarchy is as follows. Each of the 19 activities is performed by 8 subjects (4 males and 4 females) for 5 minutes. Readings are taken from sensors attached to the body of the subject at a sampling rate of 25 Hz, which means that 125 readings are taken from each subject in the trial. Each subject trains in 60 different segments (at different times of the day). Thus the total number of samples is 19*8*60=9120. The total number of features is 5*25*45=5625 as there are 9 sensors worn by each subject and each sensor produces 5 different measurements.

   More details of the dataset and the data is available at the url given above. For this dataset you do NOT need to separate it into train and test datasets. Use a single dataset for training using 10-fold CV or use a holdout method of using 66% of the training and 34% for testing. You may use either Weka or Python platforms for mining this dataset. If using Weka for mining you will need to either write a script in Python (or R) to flatten the hierarchy and do some averaging as pre-processing to reduce the number of features prior to mining in Weka.

3. **Bitcoin price movement dataset**

   This dataset tracks the movement of bitcoin over a period of time spanning over a period form 01/01/2017 to 25/03/2018 (10776 data samples). Data is recorded in hourly intervals and the captured features include timestamp, opening price, maximum trading price during the 1 hour period, minimum trading price, trade volume and weighed price. The objective in mining this dataset would be to predict the weighted trading price in advance. Since this is a time series problem a recurrent neural network such as the LSTM should be experimented with. Experiment with the LSTM with different time windows of data and compare your results with generic regression methods such as Support Vector Regression (SVR) and Linear Regression (LR). One of your performance metrics could be the Root Mean Square Error. You need to use two other metrics for comparison in addition to RMSE. Use the first 66% of the data for training and use the rest for testing. Note that you should not randomize the data as it would destroy the time series nature of the patterns.

This project will suit someone who is comfortable with Python programming and is willing to learn TensorFlow. Starter code for configuring the LSTM will be provided and placed in the Assessment 2 folder.

### 4. Earthquake aftershock analysis

This is another interesting dataset that contains data on aftershock activity that followed after the large Darfield earthquake that occurred on 4 September 2011 in Canterbury, New Zealand. The objective in aftershock activity in general is to obtain a spatiotemporal (i.e. in both space and time simultaneously) understanding of the key parameters of aftershock activity which are magnitude (intensity) and depth (i.e. at which depth measured from the earth's surface) at which the shock occurred. For the particular dataset that you will be mining you will only conduct a spatial analysis and remove the time component. This means that you will track shock magnitude and shock depth by latitude and longitude.

The data is obtained from earthquake catalog data from GNS Science New Zealand https://www.gns.cri.nz. Some preprocessing of the data has been done by myself and this version can be found in Blackboard in the Assessment2 folder. You will need to do further preprocessing as per the paragraph above.

The data will be mined through a clustering approach. Experiment with 3 different clustering algorithms, one of which must be the Self Organizing Map (SOM). If you are using Weka, please note that this algorithm does not come built into Weka and so you will need to download it using the Package Manager.

One you have preprocessed the data apply each of the clustering algorithms and visualize (using Weka's clustering visualizer or any other visualization tool) in order to gain a visual understanding of the quality of the clusters produced – bear in mind that we seek a spatial understanding of aftershock activity.

For a fair comparison we will keep the number of clusters the same across all clustering algorithms. In order to set the number of clusters, first apply SOM. SOM will automatically determine the number of clusters from the grid that you specify and you can then use this number with your other clustering methods (for e.g. k means).

Once you have compared the 3 algorithms on the basis of visualization you need to compare cluster quality on the basis of two quantitative measures:

1. Average sum of squares taken over all clusters.

2. Cluster Silhouette measure, (see https://en.wikipedia.org/wiki/Silhouette_(clustering) for an explanation on how this is computed).

Now compare the 3 algorithms on these two measures. In your analysis section of your report you can comment on consistency between the visual comparison and the quantitative comparison.

In the last part of your experimentation take the winner algorithm (i.e. the best clusterer) and perform a classification analysis in order to find out the spatial location of the largest aftershocks. In order to classify aftershocks we can convert clusters to classes by applying the AddCluster filter in Weka. Once the conversion is done apply a classifier such J48 or Naïve Bayes to generate a model and answer the question: Where in space (i.e. latitude and/or longitude) does the highest magnitude earthquakes lie?

This particular project will not make use of either Bagging or Boosting methods. *It will require programming skills in Python or R in order to do the quantitative cluster comparison mentioned above.*

## Some Notes on Experimentation: - only applies for Mining Secom or Daily activities datasets only

1. When choosing algorithms you may wish to do some initial exploration and then draw up a short list of algorithms for further experimentation. The size a, of this shortlist must be at least 3. Parameters *f* (number of feature selection methods) and *p* (number of parameters tuned for each algorithm) must be at least 2, while *b* (number of different boosting methods used) must be at least 2.

2. Your boosting methods must include Bagging, Boosting and another method that uses two or more different algorithms to build a model. Note that pure Bagging and Boosting only use one algorithm as the base method, so your third method needs to go beyond using pure Bagging or Boosting methods. Explore the different options available in Classify/meta tab.

3. If using Weka, not that although it is capable of using both .arff and .csv files, I recommend that you use .arff files only. To convert a .txt file into .arff, two steps are involved. First read the .txt file into MS Excel, format it into columns, and then save the formatted version using the .csv option. Read the .csv file

into Weka and then save it as an .arff file. Now open the .arff file and start working.

4. Because of the number of experiments and the written report you must not leave it till too late. Start work in the week that the assessment is handed out and do some initial experimentation as soon as possible to get a feel for the datasets and/or algorithms that you would like to work with.

5. Read through the accompanying document (Experimental Framework) for suggestions on a generic experimental plan (for the Daily Activities dataset, ignore the first 2 steps in the plan). The experimental plan is the key to successfully completing this project on time. The plan suggested needs to be adapted to the dataset you have selected as not all activities in the plan may be applicable for the dataset that you have selected to mine.

## Marking Scheme

The marking for this assessment will be done against four major headings, namely *Overall Document Quality*, *Description of the Mining schemes used*, *Experimental Study*, and *Post-Processing Analysis*. The detailed breakdown of the mark is as follows:

| |
|---|
| 1. Overall Document Quality (10%) |
| • Structure<br>• Clarity of presentation<br>• Referencing |
| 2. Mining Schemes used (20%)<br>• Justification for using schemes<br>• Description of Mining algorithms used - include necessary theory, algorithm strengths, limitations |
| 3. Experimental Study (50%)<br>• Overall Experimental Plan<br>• Pre-Processing (Feature Selection, Missing Value Estimation, Normalization, etc, as appropriate)<br>• Parameter Tuning<br>• Definition of Performance metric |

| |
|---|
| • Use of performance boosting techniques (if applicable to your project) |
| 4. Post Processing Analysis (20%)<br>• Analytical techniques used for performance comparison (between the algorithms you used)<br>• Your interpretation of why your winner algorithm was better than your runner up. You must include your justification based on appropriate performance metrics<br>• The success (or otherwise) of the application of performance boosting methods<br>• Comparison of your results with previous research on the same dataset (if appropriate). |