**Assignment - Data Mining Practice and Analysis**

**Due date:** 5pm Wednesday 12 September 2018

**This assignment can be worked either as a group (two students at maximum) or as an individual. If you work as a group, then group members must equally contribute to the group work. Also, all group members must participate in the presentation.**

**Aims**

* Familiarise with some well-known data mining techniques in order to understand their working principles;
* Apply data mining techniques to domain-specific datasets;
* Review cutting-edge data mining techniques to gain good overview on current data mining technology;

**Requirements (Tasks)**

The whole task of this assignment consists of the following procedural steps.

**Step 1**

Set up (by your imagination of a real-like business situation or by applying an actual analysis problem case) a scenario in which you are given a set of domain-specific dataset and asked to analyze the given dataset. The purpose of the analysis might be to understand (overview or learn about) the given data or to solve a specific analytical problem – depending on the scenario you made up.

**Step 2**

Find and get your own domain-specific dataset to fit for the scenario you made up. The dataset could be unique or publicly available. Public data sets are available from:

* + <http://archive.ics.uci.edu/ml/>
  + <http://service.re3data.org/search/>
  + <https://dataverse.harvard.edu/>
  + <http://catalog.data.gov/dataset>
  + <http://dataportals.org/>
  + <http://mldata.org/>
  + <http://oad.simmons.edu/oadwiki/Data_repositories>
  + <https://www.quandl.com/>
  + <http://www.google.com/publicdata>
  + <http://www.kdnuggets.com/datasets/index.html>
  + <http://lib.stat.cmu.edu/datasets/>
  + <http://webscope.sandbox.yahoo.com/catalog.php>
  + <http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html>
  + <https://github.com/caesar0301/awesome-public-datasets>
  + [https://www.reddit.com/r/datasets/](https://www.reddit.com/r/datasets/%20)

**Step 3**

Choose appropriate data mining techniques (algorithms) – see more details for each option in Step 4 below.

\*\* Note: The procedural order of the above three steps can be alternated. For example, you may find an interesting dataset first and then set up a specific data-mining scenario which fits for the analysis on the dataset chosen. \*\*

**Step 4**

You can select either of two options for this assignment.

* **Option (1) – Programming-intensive Assignment**
* Once you have your own domain-specific dataset and chosen data mining algorithm, then you need to design and implement the chosen algorithm in your preferred programming language.
* A series of preprocessing will be required at this step. The preprocessing procedure should be designed carefully (considering what kind of processing will be required? How? Why?) to make your data ready to be fed to your program. Some parts of this preprocessing procedure can be included in your program as a part of “pre-data-mining module”.
* Your final program must become a stand-alone data-mining tool designed for your own purpose of data analysis. It is expected that your program should include the following modules (and may include more sub-modules if needed);
  + - * 1. pre-data-mining module – designed for necessary preprocessing and for getting the data ready to be fed to the next module (data-mining module). You don’t need to include all required pre-processing in this module. It is assumed that some initial preprocessing (e.g. cleaning noise data) can be done externally using other software tools (e.g. Excel or Weka).
        2. data-mining module – the chosen data mining algorithm is implemented. You can directly borrow the algorithm from one popular existing data mining method, or you can design your own algorithm (by amending the existing one)
        3. post-mining module – this module is for presenting/reporting the output result produced through previous modules. The result can be made in a simple text report or additionally in a non-text visualization way (e.g. graph, chart or diagram).
* This programming-intensive assignment still requires an analysis. Try to find all the patterns you can detect with your implemented algorithm. Try to compare and contrast the result using your chosen preprocessing scheme and algorithm with using other existing algorithm or with using other preprocessing methods.

Note: in particular for the comparison the result using your program with using other existing algorithm, you can use other existing data mining tools (e.g. Weka) to get the result using other algorithm.

* **Option (2) – Analysis-intensive Assignment**
* Once you have your own domain-specific dataset chosen, you need to design your own data-mining analysis scheme. This analysis scheme can consist of multiple steps of procedures:
  + - * 1. Set up a strategy for preprocessing on your data.

A series of preprocessing will be required and need to be designed carefully (considering what kind of processing will be required? How? Why?). You may include multiple different preprocessing schemes for the comparison analysis.

* + - * 1. Set up a strategy for data-mining.

you need to select one data mining areas (clustering, classification, association rules mining) of your choice and select AT LEAST TWO existing data mining algorithms in your chosen data mining area. For example, if you chose Clustering as your data mining area, you can apply two algorithms; DBScan and K-mean and compare the two results. Alternatively you can design a combined algorithm which applies multiple algorithms from same/different data mining areas in a series. Your strategy also can be designed to apply different parameters for one algorithm. Another strategy you can set up is to apply multiple preprocessing (attribute selection) schemes for one algorithm.

* You can choose one data mining tool (e.g. Weka) to analyze your chosen dataset. Apply the data-mining strategy (you had set up) on your chosen data (preprocessed) using the data mining tool and try to find all the patterns you can detect.
* Do various comparison experiments either by applying different data mining algorithms (or strategy) to the same chosen dataset or by applying a same algorithm to the differently pre-processed datasets.
* Critically analyze experimental results and discuss/demonstrate why a chosen algorithm (strategy) is superior/inferior to other algorithm (strategy).

**Step 5**

* You need to present an in-class presentation (15 minutes presentation + 5 minutes question) based on your chosen algorithm (strategy) and experimental test, and also you need to write a scientific paper as an experimental report.
  + The presentation must generally include a good overview on your project, aims/objectives, reasons of your choice, brief overview of strategy/algorithm you chosen, findings, comparison including experimental results and conclusion.
  + You need to write a research report paper of minimum 10~15 pages (for CP3300 students) or 15~20 pages (for CP5605/CP5634 students) in length on your project to summarise your algorithm and experimental results. The report should contain all topics listed above for presentation but with more details. For CP5605/CP5634 students, you need to add in your report one additional section for a brief (mini) literature review about the data mining methods (strategy, algorithm and/or preprocessing methods) you chose for your project. Please refer to the following link if you need to get further idea of “literature review”:

<http://www-public.jcu.edu.au/libcomp/assist/training/JCUPRD_026326>

* The research paper must follow the generally accepted format of research article consisting of introduction, related work (brief review of methodologies (algorithm/strategy used), a summarized description of your experimental settings and procedures (description of data, justification of chosen data mining area, justification of chosen algorithm, preprocessing details, etc.), comparison, discussion, issues, conclusion, possible future work and a list of references. (you may add more sections if needed)
* In addition to the general components listed above, the report from “**Programming-intensive option**” should include a summary of your program (including the program structure, implementation details, a summarized algorithm for the main modules etc. including code if necessary).
* For “**Analysis-intensive option**”, it is required to include a more in-depth analysis on the investigation and experimental comparison made through the project.

**Submission**

* + Due for the report submission: Tuesday 16 May 2017 (in week 9)
  + Presentation: Before the report is due, usually during week 8.
  + You need to submit your final report as a single document file (MS Word or PDF format) to LearnJCU.
  + For the “Programming intensive option”, you need to submit the source code and executable file of your program accompanied to your report. Please make a zip file including all necessary files (report document and program files).

**Useful Links**

• http://www.kdnuggets.com/

• http://www.cs.waikato.ac.nz/ml/weka/

• http://mlearn.ics.uci.edu/MLRepository.html

• http://kdd.ics.uci.edu/

• http://www.sigkdd.org/

Writing Skills: <http://www-public.jcu.edu.au/learningskills/resources/wsonline/>

Scientific Report Writing:

<http://unilearning.uow.edu.au/report/2b.html>

<http://writing.wisc.edu/Handbook/ScienceReport.html>

and more on the Web.

**Marking Criteria**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **REQUIREMENT (TASK)** | **CRITERIA** | | **MARKS** | **COMMENTS** |
| Data Mining Practice  (20%)   * + Common to both options | Dataset choice – appropriateness and complexity | | \_\_\_/2 |  |
| Relevance of chosen data mining area and algorithm | | \_\_\_/2 |  |
| Preprocessing – appropriateness and complexity | | \_\_\_/3 |  |
| Investigation/justification/practices for appropriate application of the chosen algorithm. | | \_\_\_/3 |  |
|  |
| **[Programming-intensive option only]** | Structurised experimental setting for comparison analysis | | \_\_\_/2 |  |
| Program design/implementation/completeness | | \_\_\_/6 |  |
| Result Analysis | | \_\_\_/2 |  |
| **[Analysis-intensive option only]** | Structurised experimental setting for comparison analysis | | \_\_\_/4 |  |
| Result Analysis | | \_\_\_/6 |  |
| Written report  (20%) |  | **CP3300** | **CP5605**  **CP5634** |  |
| Readability & presentation | \_\_\_/5 | \_\_\_/4 |  |
| Structure & organisation | \_\_\_/5 | \_\_\_/4 |  |
| Scientific/Technical quality (including novelty&innovation) | \_\_\_/4 | \_\_\_/3 |  |
| Literature Review | N/A | \_\_\_/4 |  |
| **[Programming-intensive** option only]  Summary of the program implemented (description about design, structure, components, algorithms etc.) | \_\_\_/6 | \_\_\_/5 |  |
| **[Analysis-intensive option only]**  Comparison analysis | \_\_\_/6 | \_\_\_/5 |  |
| Class presentation  (10%) | Speech | | \_\_\_/3 |  |
| Visual aids | | \_\_\_/3 |  |
| Structure | | \_\_\_/3 |  |
|  | Question handling | | \_\_\_/1 |  |
| Total | | | \_\_\_/50 |  |