CP3300: Data Mining and Knowledge Discovery

Assignment, Semester 51, 2013, School of Business (IT), JCU

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

- Design and implement some well-known data mining techniques in order to understand their working principles;
- Apply data mining techniques to publicly available datasets or your own datasets;
- Choose the best data mining algorithm for a given dataset;
- Review cutting-edge data mining techniques to gain good overview on current data mining technology;

Tasks

- Assume that you are given a set of domain-specific dataset (might be geospatial data, might be web data, might be bioinformatics data etc), and asked to develop a data mining algorithm to analyse the given dataset.
- You can choose your favourite domain-specific dataset and also you can choose your favourite data mining techniques.
- You can choose datasets from the UCI machine learning repository (http://archive.ics.uci.edu/ml/), Web, or you can make-up your own dataset. If you choose a dataset publically available (the UCI repository or Web), then you have to compare and contrast your mining results with those who have done the experiment with that dataset. For instance, if you choose the Wine dataset (http://archive.ics.uci.edu/ml/datasets/Wine+Quality) from the UCI repository, then you need to compare and contrast your mining results with, for instance, the paper written by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis titled "Modelling wine preferences by data mining from physicochemical properties", Decision Support Systems, Elsevier, 47(4): 547-553.
- Once you have your domain-specific datasets at hand and your data mining algorithms, then you need to design and implement the chosen algorithms in your preferred programming language. Those who do not have strong background in programming, then you might form a group with somebody else who has programming background, or you can use Visual Basic Application within Excel spreadsheet to program and visualise your experiment. (Do NOT simply reuse some existing code in the Web. I can easily spot on this code reuse. You need to implement in your own.) If a group decides not to implement its own algorithm, but to use existing algorithms, then the group needs to do more extra work. Details will be found later.
- One you have your chosen algorithm implemented, then analyse your dataset with your algorithm. Try to mine all the patterns you can find.
- Also, you may need to compare and contrast your chosen algorithm with other existing algorithms. At this point, you do NOT really have to implement all other methods to test, but you can use some existing data mining tools to test

(experimentally compare and contrast) your algorithm. There are some freely available data mining tools and you might use one of the following tools.

  - WEKA
  - RapidMiner
  - You can find more at http://www.kdnuggets.com/software/suites.html

- Critically analyse experimental results and demonstrate why your chosen algorithm is superior/inferior to other existing algorithms with your datasets.
- You need to present an in-class presentation based on your chosen algorithm and experimental test, and also you need to write a scientific paper on that.

Weighting & Composition

This assignment is worth 50% of your overall mark.

- Component 1: Programming assignment (**Group assignment**) (20%)
  - Select one domain-specific dataset to analyse;
  - Select one data mining area (clustering, classification, association rules mining) of your choice for the domain-specific dataset;
  - Select AT LEAST ONE existing data mining algorithm that you believe the best for the dataset of your choice in your chosen data mining area;
  - Design and implement the chosen algorithm in your preferred programming language;
  - Investigate if you can improve the chosen algorithm in terms of efficiency and effectiveness;
  - If your group decides not to implement your own code, but to use one of existing codes, then your implementation mark will be assigned to the investigation into an improvement of your chosen algorithm for the chosen dataset.
- Component 2: Class presentation (**Group** assignment) (10%)
  - 20 minutes (15 minutes talk + 5 minutes question time) in-class presentation on your project.
  - The talk must generally include a good overview on your project, aims and objectives, reasons of your choice, literature review, findings, comparison including experimental results and conclusion.
  - In this presentation, assume that you are one of authors of the algorithm of your choice and you are presenting it in front of audience.
- Component 3: Written assignment (**Group assignment**) (20%)
  - A research paper of 10-15 pages in length on your project that summaries your algorithm and experimental results including references. You need to use a proper referencing technique such as Harvard referencing.
  - The research paper must follow the generally accepted format of research article consisting of introduction, related work, comparison, discussion, issues, conclusion, future work and a list of references.

Due

- TBA

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/

Assessment criteria:

1. Programming-intensive project (implementing his/her own code)
    - Programming assignment (**Group** assignment) [20 marks]:
        - Difficulty of the algorithm: 5
        - Improvement of the algorithm in terms of efficiency or effectiveness: 10
        - Execution (including correctness): 3
        - Uniqueness of dataset: 2
    - Class presentation (**Group**) [10 marks]:
        - Speech: 2
        - Visual aids: 2
        - Structure: 2
        - Group participation: 2 (equal participation: 2, unequal participation: 0)
        - Questions & answers: 2
    - Written assignment (**Group** assignment) [20 marks]:
        - Readability & presentation: 5
        - Novelty & innovation: 5
        - Scientific & technical quality: 5
        - Structure & organisation: 5
2. Investigation-intensive project (NOT implementing his/her own code)
    - Programming assignment (**Group** assignment) [20 marks]:
        - Justification of the choice of the algorithm: 5
        - Improvement of the algorithm in terms of efficiency or effectiveness: 10
        - Uniqueness of dataset: 5
    - Class presentation (**Group**) [10 marks]:
        - Speech: 2
        - Visual aids: 2

    - Structure: 2
    - Group participation: 2 (equal participation: 2, unequal participation: 0)
    - Questions & answers: 2
- Written assignment (**Group** assignment) [20 marks]:
    - Readability & presentation: 5
    - Novelty & innovation: 5
    - Scientific & technical quality: 5
    - Structure & organisation: 5