CP465, Databases II

Term Project (Option 1)

Data Mining

Project due date: March 31, 2020, 10:00 am
Max. Score: 30 + 5 (optional bonus)

1 Introduction and PHP website component (optional) of the project

OPTIONAL PART: intended for a bonus score of (5 points--- this is apart from the 10% meant for other bonus activities)

- The objective of this project is to create an impressive PHP website that would carry details of this project. It is left to you to be creative with the website (the extent of functionalities you provide to it, level of associated interactivity in terms of user inputs and outputs possible). The website in the most basic sense should carry at least your project code, link to the datasets used and all the outputs.

MANDATORY PART:

- You have to implement three Data Mining algorithms and test your implementations with a reasonably large dataset. The term *reasonably large* is to be interpreted as anywhere from a few hundreds to a few thousands instances (tuples).

2 **Algorithms** (30 marks = 10 for each of the 3 algorithms)

You need to implement any **three** out of the four Data Mining algorithms seen in class:

1. the C4.5 algorithm, based on information gain ratio, to build decision trees;
2. the Apriori algorithm for mining association rules;
3. the k-Means clustering algorithm;
4. the Supervised genetic learning algorithm;

3 Datasets

You need to test your implementations for each Data Mining algorithm with a dataset:

1. one (reasonably large) dataset that you will find on your own.
   A large selection of datasets is available for instance at the Machine Learning Repository http://archive.ics.uci.edu/ml/

4 Data Mining Resources

This is a selection of some useful on-line Data Mining resources:

- A portal for KDD/DM tools, news etc. http://www.kdnuggets.com
- Data Mining tutorials and other goodies: http://www.the-data-mine.com/
- ACM Special Interest Group (SIG) on KDD http://www.sigkdd.org/

5 Submission of your work

By the project due date, you need to submit the following items:

1. The cover sheet, (Names/Student IDs of group members, Course Number, Date)
2. The design document, which should include:
   - (a) description of what features work and what features don't.
   - (b) description and justification of the design choices made.
   - (c) description of data structures used.
   - (d) instructions on how to compile and run the code.
3. The test document which should show the results of running (with your implementation) each algorithm on the dataset, as specified above. Include a copy of the dataset you found on your own.
4. The code in electronic form only. Well-commented code will count for more marks than code without comments.

==

- All project submissions must be typesetted.
- Late project submissions will not be accepted and will be marked with 0.
- Students can work in the groups of 3 to 5.
- All submissions to be made over MyLS.
- Only one group member should submit the file (in PDF) and ensure that names of all the group members with student IDs are mentioned on the front page of the file.
- Each member must clearly document their contribution in the work done on the first page. This will help the grader to determine participation level of the group members.