Project: Solve a Real Data Mining Problem

Report Due: May 7 11:59pm

In this project, you will practice what you learn in class to solve a real-world data mining problem. You can choose any problem that you are interested in as long as it can be formulated as a data mining task. Each student needs to submit an individual project report, but you are encouraged to discuss your idea with the other students.

Complete the following tasks:

1. Pick a real-world application that data mining may help.

2. Formulate it as a data mining problem (clustering, classification, association analysis, anomaly detection, recommendation, or a combination of these tasks).

3. Collect relevant datasets. Some possible sources:
   - https://archive.ics.uci.edu/ml/datasets.php
   - https://kdd.ics.uci.edu/
   - https://www.data.gov/
   - http://www.kdnuggets.com/datasets/index.html

4. Preprocess the datasets into the format that can be used by data mining algorithms if necessary.

5. Apply your implemented algorithms or any existing software package to solve the proposed problem.

6. Discuss the data mining results you obtain and evaluate the results.

7. Prepare for a short report (2-4 pages) based on the key points of your project. Name it as project.pdf or project.doc or project.docx, and submit it on Brightspace.

Your report should include the following components.

- Introduction: What data mining problem you are trying to solve? What impact it will bring if the problem is solved?
- Formulation: Which data mining task it can be formulated into? What's the input and the expected output?
- Datasets: Where do you get the datasets? Give some statistics about the data. How do you preprocess the data?
- Algorithm: Which data mining algorithm do you apply?
- Experiments: Evaluate the output using appropriate evaluation metrics and case studies. Show the results you get and discuss whether they are meaningful.
- (Optional) Comparison: How's the performance of different algorithms on the dataset? How's the performance of an algorithm on different datasets? Can you explain why an algorithm performs well (or not) on a dataset?
- (Optional) Challenges: What challenges do you find in the data? How do you tackle these challenges?

Scoring of the project:

- Is the data mining problem you identify is formulated well?
- Have you applied necessary preprocessing steps, and chosen appropriate data mining algorithms? Have you justified why?
- Have you specified clearly the experimental setting (such as parameters and evaluation metrics)?
- Have you demonstrated the experimental results in both evaluation metric and case studies? Have you correctly interpreted the obtained results?
- Have you presented everything clearly in the report?
- (Bonus): Have you compared more than three algorithms or on more than three datasets? Have you explained the results well? Have you explained on what types of the data the algorithm works well or does not work well?
- (Bonus): Have you tried some advanced techniques to improve the algorithm based on the characteristics of the dataset?