CSE514A Datamining – Fall 2018 Course Project

Grouping:

- You form a group of 3 students for this project – if you cannot find your teammates, post an open request on piazza. In case you need to have a group of more or less students, talk to the instructor.

Objectives:

- Learn how to formulate a real-world data analytic problem as a datamining task or a series of tasks and solve it.
- Discover interesting models/patterns/relationships (e.g., parts or community structures) in a large dataset.
    - Reveal and understand that real systems have structures, patterns and relationships among entities within.
    - Gain new insights/knowledge from the problem domain.
    - Visualization if possible
- Improve the accuracy and/or running time of some existing data mining algorithms.
- Develop new datamining methods

Time schedule:

- 10/2 - Form your team and decide your project (project 1 or project 2 below); submit a one-page report
- 10/9 – Submit your letter-of-intent if you propose your own project. (skip this if you choose project 1)
- 10/23 – Submit a short review-type essay (see below)
- 11/15 – Submit a midterm progress report
- 12/4 and 12/6 – In-class presentation; you need to submit your presentation slides before your presentation.
- 12/6 (midnight) – Submit your final project report
- Note – the instructor/TAs may request to have your implementation and/or ask you to demonstrate your software for the project

Possible project 1 – Identification and analysis of functional modules in large networks

- Datasets - Stanford Large Network Dataset Collection (snap.stanford.edu/data/)
    - Study the datasets in other categories closely. Choose at least two categories for study in your project.
    - Tip: for preparation and test purpose, use the datasets in "Network with ground-truth communities" category. Compare the results from whatever methods you choose with the ground truth network modules to evaluate their performance. So this category cannot be used in your discovery
- The objectives here are to 1) identify good module or community structures embedded in the data, 2) determine (as much as you can) the possible functions of the modules.
- What to do:

- Read the papers in the network_community.zip – read one review paper and the abstracts of all of the papers in the package, and read closely 3-5 papers whose methods you choose to use.
- Note: you are highly encouraged to find more recent papers on community detection and choose methods in some of these recent publications. You will get a few bonus points if you find/choose some good methods not in the package above
- Based on your reading, propose what you plan to do:
    - Select at least 3 methods to be used. Note: if you propose to develop a new method and it works reasonably well, you get an A+.
    - Define your objectives based on the data categories you choose
    - Discuss the problem(s) and methods among yourselves
    - Write and submit a short review-type essay (~5 pages without references) to demonstrate your understanding of the problem(s) and methods.
    - Form your plan– decide how to achieve the objectives, e.g., determine the steps of your (comparative) analysis.
    - Write and submit a proposal – the proposal must include all of the following items:
        - the problem(s)
        - your objectives
        - the data and methods to be used, as well as a briefly explanation why you choose these methods over other existing ones
        - a detailed description of your work plan (including work distribution within the group, work integration and collaboration)
        - your evaluation criteria (e.g., solution quality, time and space complexity and other performance metrics that may be relevant to your work) and an overall evaluation plan (e.g., how 4 methods should be compared)
        - and a brief description of what you expect to discover
        - time schedule and milestones for all individual tasks
    - Perform the actual work
    - Submit a progress report in the middle
    - prepare a presentation and a project report. Present your work to the class

Possible Project 2 – Your own project.

Read description of project 1 closely. Follow the same or a similar principles in Project 1 above and consider all the items discussed there. Furthermore, you need to consider the following additional issues:

- What is the problem or what are the problems to be address?
- What datasets will you use? Are the data available to you if you work on your own project?
- Submit a letter-of-intent that include above two items plus your rough idea what methods to use.
- Find 3 or more existing methods for this problem – you need to understand them and write a review-type essay on them. Note: if you propose to develop a new method and it works reasonably well, you get an A+.
- How will you do it?
    - Are any pre-processing steps needed?
    - What algorithm(s) do you plan to use?
    - Why do you choose the specific algorithm(s) instead of the others?

- What kind of difficulties do you expect?