CS235 - Data Mining Techniques - Project Description

Instructor: Vagelis Papalexakis, University of California Riverside

- General information
- Project Deliverables:
  - Project Proposal
  - Midterm Progress Report
  - Project Presentation
  - Final Project Deliverable
- Resources
  - Datasets
  - Problem ideas

General information

Projects will be carried out in groups of 3. For MSOL students, the default project group size is 1, but you are highly encouraged to discuss with fellow classmates on iLearn and find projects of mutual interest. Every team will present their work at the end of the quarter, either via a short presentation, or via a poster session, to be determined later in the course.

You can choose from the following project types:

Software: Pick a dataset (see Resources below) and define a problem you want to solve. Select 3-4 data mining techniques that you would like to use in order to solve the problem, *implement them from scratch*, clean and analyze the data, compare the results from the different techniques, and present the findings. Alternatively, you may define a problem associated with data that are on some website or web platform. Write a crawler that scrapes data from that platform, making sure that you respect all the crawling policies that the website has in place (usually looking at the robots.txt file of the website, or looking at their data policy). Choose 3-4 data mining techniques that you want to use; *If you implemented the crawler, it is OK if you implement one less technique*.

Research 1: This is structured exactly like *Software* with the only difference being that in addition to the 3-4 data mining techniques that have been used before (and you will use as baselines for comparison) you have to propose a novel technique (in the sense that it has not been used before in that context) or propose a sufficiently novel modification to an already existing technique for that problem. This project type can earn extra credit.

Research 2: In this project type you can propose your own idea along the lines of your own research (or rather the combination of research interests of all the team members). After you propose the idea, if there are any well known techniques for that problem, as in Research 1, you should use them as baselines, and you should propose a novel solution to that problem. This project type can earn extra credit. **Note**: in the project proposal, you must clearly state what is already a part of your current research project and what will be carried out as the class project (no double dipping! :-)).

Project Deliverables:

Project Proposal

Description:
In the proposal you have to briefly but concisely introduce your project. In particular, you have to clearly define the problem your project proposes to solve. You should be able to distill the essence of your proposal to a statement like:

Given <dataset, website, …> ***Use*** <data mining technique(s)> ***To*** <achieve “KDD outcome”>

For example:
Given Netflix data ***Use*** Collaborative Filtering algorithms ***To*** recommend new movies to users
or
Given Twitter data ***Use*** Matrix Factorization ***To*** detect fake followers

In special cases you may be able to relax the above format for the problem statement, but it is fairly generic and applies to a wide variety of problem statements. In any case, make sure you define what problem you are going to solve, and very importantly, describe how you are planning to *evaluate* your approach.

In addition to the above, make sure you include:

1. The type of the project.
2. The names of all team members.
3. Projected labor division among the members of the teams throughout the quarter.

Depending on the project type you chose, you need to clearly describe your plan on obtaining the data that you will use.

The page limit for the proposal is 2 pages, single column.

If you have chosen type Research 1/2 you must also state why your proposal is novel. Is it because nobody has applied this technique to that particular problem yet? Are you proposing a new algorithm for an already established data mining technique? Are you using a technique from another field to solve the problem? If you chose Research 2, in addition to the novelty justification, you must also state what part of the proposed work relates to your own research / thesis work, and how it will be complementary to what you are already doing for your thesis.

In all cases, remember, your project is subject to approval given the proposal, and the feedback you get from us on the proposal must be taken into account.

Page limit: 2 pages (single column)
How to submit: Submit a .pdf on iLearn.
Grading scheme: 10/100

Midterm Progress Report

Description:
For this milestone all teams have to submit a midterm report that contains the proposal (in the form of a paper introduction), and a part of their related work survey. For all project types, each team member has to read and summarize at least 2 related papers (outlining the pros, the cons, and 1-2 potential extensions).

Page limit: 5 pages (single column)
How to submit: Submit a .pdf on iLearn.
Grading scheme: 15/100

Project Presentation

Description:
The final presentation will be either in the format of a short 5-10 minute talk or a poster presentation. The exact format of the final presentation will be determined later in the quarter.

How to submit: TBD
Grading scheme: 10/100

Final Project Deliverable

Description:
The final project deliverable should include:

1. The project report in .pdf format.
2. The code for your implementation.
3. If you collected any dataset(s) for your project, include it/them in your deliverable. If the dataset is too big, coordinate with the instructor and the TA.

Details for the report:

Your report should resemble a KDD paper (download the ACM "tight" format here http://www.acm.org/publications/proceedings-template) and the page limit is 10 pages in double column format including the references.

For all project types you have to include 1) an **Introduction** where you describe and motivate the problem, give an outline of your contributions and motivate your approach; if you have *Research 1/2* you also have to argue that your proposed approach is sufficiently novel with respect to the state-of-the-art, 2) a **Related Work** section with your literature survey, a 3) **Proposed Method** section where you describe the method(s) you used to solve the problem, 4) an **Experimental Evaluation** section where you compare the methods used; if you have *Research 1/2* you have to further demonstrate that the proposed approach outperforms the baselines (at least in some cases); this can earn extra credit, and 5) a Discussion & Conclusions section where you draw the conclusions of your paper and outline potential future research directions.

For the *code*, make sure you include:

1. All source files you wrote with comments that explain your implementation.
2. A makefile that when invoked it runs your some demo of your code (make sure you include at least a "toy" dataset for this purpose).
3. A README file that describes what each file does.

Page limit: 8 pages (KDD-style double column format, ACM "tight" style)
How to submit: Submit the final report and your code package on iLearn.
Grading scheme: 30/100 = 20/100 (base) + 5/100 (originality of approach as argued in the introduction) + 5/100 (demonstrating that the proposed method outperforms the baselines)

Resources

Problem ideas

You can find ideas for problems in the following links:

1. KDD Cup Archives
2. WSDM 2019 Cup
3. CIKM 2016 Cup
4. Yelp dataset challenge
5. Kaggle

Datasets

In addition to the “Problem idea” resources that contain a fair amount of data, you can find data for your project in the following links:

1. UCR Time Series Archive
2. UCI Machine Learning Repository
3. Stanford SNAP Datasets
4. Aminer Network Data
5. Koblenz Network Data
6. Microsoft Research Asia T-Drive Data