CSE5002 Intelligent Data Analysis

Mini Project 2022

Background

The Cora dataset consists of 2708 scientific publications classified into **one of seven classes**, including “Case Based”, “Genetic Algorithms”, “Neural Networks”, “Probabilistic Method”, “Reinforcement Learning”, “Rule Learning” and “Theory”. The citation network consists of 5429 links. Each publication in the dataset is described by a **0/1-valued word vector** indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 1433 unique words. The README file in the cora dataset provides more details.
Dataset URLlink: https://linqs-data.soe.ucsc.edu/public/lbc/cora.tgz

Related papers:

Qing Lu, and Lise Getoor. "Link-based classification." ICML, 2003.

Prithviraj Sen, et al. "Collective classification in network data." AI Magazine, 2008.

Problem Specification

You can download the dataset or utilize the preprocess dataset we provide, which consists of

- attr.csv: id of paper, id of presence of the corresponding word 1, … (2708 rows)
- adj_list.csv: id of paper, id of citing paper 1, id of citing paper 2, … (2708 rows)
- label.csv: id of paper, class (2708 rows)

Assume some of there papers miss their labels, i.e., one of seven classes. Your objective is to preprocess the data, **utilize node attributes, or network topology, or both**, for training a classifier, then divide the training set and testing set, so as to finally make the good prediction for the missing labels in testing set.

Requirement

Report

You need to submit a formal report. As tips, some key points are as follows.

- What is the problem to solve?
- How do you preprocess the data before feeding to a classifier?
- What are the models of classifiers you would like to have a try? Please discuss each model w.r.t. the problem.
- How do you evaluate the models?
- How do you conduct experiments?
- Please compare and discuss your results.
- What are the limitations and how would you address them in the future.

- Your report is NOT to simply answer these questions. You should properly organize them in a formal report.
- Language: English or 中文

Source Code

- Your code.
- A README file about how to set up the environment, and how to run your code.

Attention

1. How to submit: You should compress your report (in a pdf) and source code (in a folder) into one zip file. The zip file with "ID_name", e.g., "10101010_张三", should be submitted to Blackboard.
2. **Plagiarism, 0%**. You could discuss with your classmate about the mini project, but please remember no plagiarism. We will check your report and source code.
3. Score: 70 pts (report) + 30 pts (source code)
4. Deadline: **May 30, 23:59.** No late submission