Data Mining Course Project 2 Description

2017 Autumn

General Goal

In this project, you will have an opportunity to apply the data mining algorithms and techniques you learned in the class to some real-world problems. You can choose any problem that you are interested in, and formalize it into a data mining task. Then you get some data related to the task and apply some data mining algorithms to your data. Also you need to evaluate and compare your algorithms. And finally, you should submit a report, together with your data and code.

Typical steps:

- identifying a data set and problem domain
- deciding on what you want to achieve with data mining
- choosing appropriate methods and algorithms
- implementing and testing your methods
- evaluating your techniques on your data sets
- reporting conclusions

Here are the project deliverables and due dates:

Project Deliverable	Due Date
1. Project Proposal	**Written proposal, due in class 2017-10-13. Also submit electronic copy to me.**
2. Progress Report 1	**Written report, due in class 2017-11-3. Also submit electronic copy to me.**
3. Progress Report 2	**Written report, due in class 2017-12-1. Also submit electronic copy to me.**
5. Final Project Report	**Last week.**

Project Proposal

Proposals should be relatively brief but informative. About 3 well-written pages should be sufficient.

- 1) Problem and goal
  What do you want to solve?
  Why do you think it is important?
  What results do you expect?
- 2) Formalization into data mining task
  E.g., Frequent pattern mining, classification, and clustering.
- 3) Data plan
  What kind of data?

Where and how do you get the data?
Make sure get data in time

- 4) Schedule: detailed plan of your project.

Your proposal will be reviewed and you will receive one of 3 responses: (1) approved without any changes, (2) approved with minor suggestions for changes, (3) major revision required. Typically most (if not all) proposals are in categories (1) or (2). You may begin work on your project after submittal, but be aware that you may get feedback from me after your proposal is submitted (within a week or so).

Final Report

A final report, data, and code:
Problem introduction, formalization, algorithms, experiment results, etc.

Collaboration Rules

- Every member in a team gets the same score (encourage teamwork, 1-2 students per team)
  Exception: the team has the right to claim someone as a free rider, and we will lower his/her score
- Form Groups : Group name; Group members; Group leader
- A table describing your division

An example:

Task	People
1. Collecting and preprocessing data	Student A
2. Implementing Algorithm 1	Student A
3. Implementing Algorithm 2	Student B
4. Evaluating and comparing algorithms	Student B
5. Writing report	Student A and B

- Peer evaluation

Total Credit

- 1) Group formation
- 2)Proposal
- 3) Data and code
  - Any programming language
  - Documentation
  - Own implementation
- 4) Final report

At least two algorithms and two evaluation methods

- 5)Additional features(extra points)
  - Novelty of the problem
  - Your own data
  - More than two algorithms/evaluation methods

- Other innovative features (e.g., new algorithm)

Datasets

- UCI Machine Learning Repository
  http://archive.ics.uci.edu/ml/
- Microsoft Learning-to-Rank data sets(search engine result ranking for 30k queries):
  http://research.microsoft.com/en-us/projects/mslr/

A Simple Example: Email Classification

Problem

- Determine whether a given email is spam or not

Data Mining Task

- Binary classification

Data

UCI spam data set(http://archive.ics.uci.edu/ml/datasets/Spambase)

- Number of instances: 4601
- Number of attributes: 57
- The last column denotes whether the e-mail was spam (1) or not (0)
- Partition it into training set and test set

Algorithms

- Naive Bayesian classifier
- Artificial neural network
- AdaBoost

Evaluation and Comparison

- Error rate
- ROC and AUC
- Speed

Suggested Outline and Contents for your Proposal

Project Definition: provide a clear description of the problem you plan to address.
Motivation: briefly explain why this problem is worth addressing (eg. addresses an important open research question, tackles an important practical application, etc).
Background: provide a brief summary of what is known in the literature about this problem. Provide at least 2 or 3 relevant references if you can. At this early stage of your project proposal I don't expect a full literature review (but will expect this by the time of the final report), but you should have some idea of what is known in the literature about your proposed project (for example,

try to make sure that your project has not already been done/published earlier).

Data sets: briefly describe the data set (or sets) that you will use in your project. You will need to describe your proposed data set in more detail and provide some rationale as to why you are choosing this data set.

Software: describe what algorithms and software you plan to use in your project. Algorithms that you will develop yourself? if so, sketch out how you think these algorithms will work, or at least your current ideas. Algorithms that have already been published and implemented? if so, provide brief descriptions and references. If you plan to use someone else's implementation in software, provide references and details to the extent that you can.

Evaluation method: you need to define how you will evaluate/measure/test your results or technique. For prediction problems, such as classification or regression, this is typically straightforward, since techniques such as "out of sample" accuracy are well-defined and a good indicator of your method. For techniques such as clustering or pattern-finding, it is not so clear how to measure success - you might for example try a clustering algorithm on a data set where the class labels are known, but you remove them and see if the algorithm can recover them. For such problems you may want to look in the literature to see what evaluation techniques are typically used. For a visualization project for example you might want to have human users use your system and generate subjective evaluations, e.g., comparing your method with a baseline approach in a "blind" manner where the testers do not know which is baseline and which is the new method.

Milestones and Plan: sketch out what you think will be the major intermediate milestones that you will need to achieve, e.g., for each progress report and for the final report. A bulleted list would be fine. You don't need to include every small detail, but give a general idea of what you plan to be doing for the next 6 or 7 weeks. And of course this plan may change as you learn more about the data and the algorithms. Feel free to identify potential risks in your proposal, i.e., items that might cause delay or problems.

Project Report Format

Below are guidelines on how to write-up your report for the final project. Of course, for a short class project, all of the comments may not be relevant. However, please use it as a general guide in structuring your final report.

A "standard" experimental data mining paper consists of the following sections:

1. Introduction

Motivate and abstractly describe the problem you are addressing and how you are addressing it. What is the problem? Why is it important? What is your basic approach? A short discussion of how it fits into related work in the area is also desirable. Summarize the basic results and conclusions that you will present.

2. Problem Definition and Algorithm

2.1 Task Definition

Precisely define the problem you are addressing (i.e. formally specify the inputs and outputs). Elaborate on why this is an interesting and important problem.

2.2 Algorithm Definition

Describe in reasonable detail the algorithm you are using to address this problem. A psuedocode

description of the algorithm you are using is frequently useful. Trace through a concrete example, showing how your algorithm processes this example. The example should be complex enough to illustrate all of the important aspects of the problem but simple enough to be easily understood. If possible, an intuitively meaningful example is better than one with meaningless symbols.

3. Experimental Evaluation

3.1 Methodology

What are criteria you are using to evaluate your method? What specific hypotheses does your experiment test? Describe the experimental methodology that you used. What are the dependent and independent variables? What is the training/test data that was used, and why is it realistic or interesting? Exactly what performance data did you collect and how are you presenting and analyzing it? Comparisons to competing methods that address the same problem are particularly useful.

3.2 Results

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data. Are they statistically significant?

3.3 Discussion

Is your hypothesis supported? What conclusions do the results support about the strengths and weaknesses of your method compared to other methods? How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

4. Related Work

Answer the following questions for each piece of related work that addresses the same or a similar problem. What is their problem and method? How is your problem and method different? Why is your problem and method better?

5. Future Work

What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.

6. Conclusion

Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? How will your results improve future research and applications in the area?

Bilbiography

Be sure to include a standard, well-formated, comprehensive bibliography with citations from the text referring to previously published papers in the scientific literature that you utilized or are related to your work.

Chen Xiangtao

2017-09-15

Project 汇报 PPT 写作要求

一、研究背景（要求简明扼要，特别地，将问题写明白）

二、研究思路（最好按照自己的理解，说明研究的主要思路）

三、算法过程（该部分为报告的重点，要求思路清晰，算法明了，若原文有示例，写明示例）

四、实验结果与讨论（指明算法的优势）

五、主要结论（要求简明扼要，并指明存在的问题，拟解决的思路等）