Data Mining Project Guidelines

This document provides some guidelines for writing your project proposal (project preparation) and then your final project report.

It should be examining real-world data sets and an associated problem. This is an application-based project. Ideally you should try to do something a bit interesting, so that it is novel or your approach is novel. Perhaps it is something of great interest to you. You should make sure that your analysis is not trivial. For example, running a data set through WEKA and spending an hour on the analysis and then doing a quick write-up would be considered trivial.

There are some steps listed under the full proses that will give you ideas how to expand the project.

Project Prep. Presentation & Project Prep. Reports Your project preparation report must be 1-2 pages long, single spaced. The purpose of the prep. presentation and report is to make sure that you are on the right track and to give me enough information so that I can give you useful feedback.

In your preps you should cover the following items:

• Preliminary title and list of students working on the project (1 or 2 students per project)

• **Abstract:** Similar to the abstract that will ultimately appear in your paper. It should be one paragraph long, for now perhaps only 5-15 lines. It should provide a high-level summary of your project and outline your main goals.

• Brief description of what you plan to do.

  o What problem are you trying to solve?

  o How do you formulate the problem as a data mining problem (e.g., is it classification, association rule mining, etc.)? What exactly are you trying to predict (for prediction tasks) and how will you evaluate your results. How will you know if your results are good? What can you compare them to? It is critical that your problem is well-defined.

  o What data sets do you plan to use? If you must do significant work to get the data or convert it into the proper format, then describe the process and approximate effort required. How many examples are in the data set? How many features?

  o What learning tools do you plan to use (e.g., WEKA. Python Scikit) and what algorithms do you plan to use (e.g., decision trees, neural networks, etc.)?

  o If sufficiently valuable, list a few related research papers.

Project Presentations & Project Reports Students should do an oral presentation of their project and submit project report.

Below are some things that you should consider including, that are often included in the best papers. But it depends on how complex your project is. For more straightforward projects, you will need to go out of your way to do more, so the project is not trivial.

• **Parameter tuning:** for prediction algorithms, try different setting and perhaps even show how the results vary. For example, the most obvious setting for kNN is the value of k, so consider varying it.

• **Visualize the distribution of key features** and how they relate to the class variable. Minimally this can be done with a set of scatter plots that include the feature under consideration and the class variable.

• Consider **feature selection.**

• If your project involves building a predictive model, definitely try to explain/describe the **model**, not just present the results. This is easier for some models (e.g., decision trees) than others (e.g., neural networks). Some methods also will generate a measure of feature importance, so you can at least rank order the importance of the features.

The report need not be organized exactly as described below, but it should be quite similar, since the outline below is generally what is used for most conference papers in data mining related conferences. There is no precise page or word length requirements. The key is that the report clearly and concisely describes the project, and the project should be substantive enough so that there is a fair amount to discuss. The paper should be written to an audience who has basic knowledge of data mining, such as those completing one course on the topic.

Here is the suggested outline:

• **Abstract:** summarizes the report and the goals of the work. It should be limited to a single paragraph and should be a maximum of 500 words. It should not provide a comprehensive summary of the report. Rather it should motivate the problem, define it, and briefly discuss the general approach. It may or may not include some basic results, but any discussion of results should be limited to 1-2 sentences.

• **Introduction:** Introduces the project and what you are trying to do. Should motivate the problem, quickly define it and the approach taken, and may discuss some highly related work. Probably should mention and contributions of the work. Usually about ½ page to one page.

• **Background:** Depending on the project, you may want a separate background section, depending on how much background you want to include. For example, it may provide domain information for the domain that you are studying. If the domain is not complex, then this section may not be needed. This is generally not about related work.

• **Experiment Methodology:** Describes the experiments and the experiment methodology. Will describe the data sets, evaluation metrics, data mining algorithms used, the precise methodology related to the setup of experiments, and any other details related to the experiments. There will usually be a subsection for each of the sub-topics just mentioned. If you use only well-known algorithms, you can usually just cover them all in a single paragraph. But if some are less well known, maybe include a paragraph on each. Should identify key parameters. If you try various parameters, may want to discuss that here. Results do not go into this section.

• **Results:** Presents the experiment results and a discussion and analysis of the results. Normally a separate discussion section is not necessary. If there are a lot of results, try to break it down into two or three subsections if there are different types of experiments. Make good use of Tables and Figures. Figures are better than tables when you want to show something like a trend. It may make sense to have some results that are relatively low level, and then separate tables/figures to show higher level results for different

experimental setups that can easily be compared. For example, you may want to summarize the results for the best set of parameters in a separate table. You should make effective use of Tables and Figures. Please label all Figure axes precisely and provide appropriate caption names.

• **Related Work:** A description of related work, with citations to relevant papers. You are doing an application report, where you analyze some data, you are less likely to rely on a lot of related work. Nonetheless, there almost always should be some related work discussed. If there is not that much related work to discuss, it may be possible to include the related work in the introduction (since it may provide a motivation and context for the project). If there is a lot of related work, you may want to provide the bulk of it in this section, but include some in the introduction too. In general, every paper should mention a minimum of 3-5 related work papers. There almost always will be related work, even if it is not a perfect match. As an example, lets say you are doing a paper on using data mining to classify Pokemon characters. You could look at other data mining papers on Pokemon, then other data mining papers on similar video games, then possible related work on sports. All related work papers should be cited in the paper and then should appear in the reference section. All items in the reference section should be explicitly cited in the paper.

• **Conclusion:** Provide your conclusion (perhaps summarize your main results). Normally will also discuss limitations and avenues for future work.

• **References**: Each paper should have a references section. This should include references to related work, but also references unrelated to related work. For example, if you are using Weka there should be a Weka reference (same for Python Scikit), possibly references to specific algorithms and metrics. Generally speaking, you should have at least 3 related work references and at least 3 non-related work references.

Sample Data Mining Projects

• Air Pressure System Failure Prediction in Scania Trucks (2019)

• Predicting Kickstarter Campaign Success (2019)

• Bike Sharing Rental Prediction (2019)

• Landscape Image Classification Using Unordered Color Data (2019)

• Spam Email Trigger Words (2019)

• Gender Prediction from OkCupid Profiles using Text Mining & Ensemble Methods (2019)

• Pokemon Type Classification (2019)

• Using Statistical Averages to Predict Preferred Roles for Overwatch League Players (2019)

• NCAA March Madness Result Prediction Model (2015)

• Authorship of Federalist Papers using SAS text mining (2009)