Milestone 1 Progress Evaluation

Student Performance Analysis

Shreyas Ugemuge sugemuge2014@my.fit.edu

Yaqeen AlKathiri yalkathiri2013@my.fit.edu

Mohammed AlHabsi malhabsi2013@my.fit.edu

Shiru Hou shou2015@my.fit.edu

Faculty Sponsor: Dr. Phillip Chan pkc@cs.fit.edu

February 23, 2017

Contents

1 Progress of current Milestone

Task	Completion %	Mohammed	Shreyas	Yaqeen	Shiru	To do
1. Select Collaboration Tools	100	20	40	20	20	n/a
2. Investigate packages, languages and tools	100	25	25	25	25	n/a
3. Familiarize with the required technologies	100	25	25	25	25	n/a
4. Requirement Document	90	-	100	-	-	Document will be revised for later milestones
5. Design Document	100	33	33	-	33	n/a
6. Test Plan	50	-	10	40	-	Test plan will be revised for milestone 2
7. Progress Evaluation	100	10	70	10	10	n/a
8. GIT test push	100	25	25	25	25	n/a
9. Python hello world	100	25	25	25	25	n/a

1.1 Discussion of each task

1.1.1 Select Collaboration Tools

This task included of setting up the following collaboration tools:

1. GIT repository

2. Webpage
3. Slack team communication
4. Slack shared calendar
5. Slack shared and personal to do list
6. Google slides
7. Google docs

This task was finished 100%
links: `github.com/shreyasugemuge`

1.1.2 Investigate packages, languages and tools

The following data mining tools were explored and pros and cons listed:

Technology	**Notes**	**Pros**	**Cons**
Advanced Miner	see 4.1.1	see 4.1.1	see 4.1.1
Ghost Miner	see 4.1.2	see 4.1.2	see 4.1.2
GNOME	see 4.2	see 4.2	see 4.2
SPAD	see 4.2	see 4.2	see 4.2
ESTARD	see 4.3.1	see 4.3.1	see 4.3.1
Statistica	Not freeware, Changed to WEKA (see 4.3)	see 4.3.2	see 4.3.2
Tanagra	see 4.4	see 4.4	see 4.4
Rapid Miner	see 4.4	see 4.4	see 4.4

1.1.3 Familiarize with required technologies

This task included of familiarizing with GIT, google docs, Python and 8 of the assigned tools. Hello world like programs were pushed on the repository for python and git.

1.1.4 Requirements, Progress Evaluation and Design Documents, Test Plan

The documents were created according to the guidelines provided on `http://cs.fit.edu/~pkc/classes/seniorProjects/` and can be found on the website as well as repository

1.1.5 GIT test push and Python Hello World

The corresponding files were pushed on to the repository.

1.2 Discussion of team member contribution

1.2.1 Shreyas

Created GIT repository and project website and set up Slack team communication. Researched and provided brief report on AdvancedMiner and GhostMiner. Authored the progress evaluation and the Requirements document. Contributed to Design and Test Plan document and presentation.

1.2.2 Shiru

Set up calendar and todo list plugin for slack. Researched and provided report on GNOME and SPAD. Contributed to Design document and presentation. Explored new technologies including GIT.

1.2.3 Yaqeen

Provided detailed report on Tanagra, SPAD and Rapid Miner. Set up google slides. Authored Test Plan.

1.2.4 Mohammed

Provided report on ESTARD and statistica. Set up google docs. Contributed to all 4 documents and presentation.

2 Plan for the next milestone

Task	Shreyas	Yaqeen	Shiru	Mohammed
Finalize conceptualizing 8 behaviors	25%	25%	25%	25%
Finalize packages, languages and tools	25%	25%	25%	25%
Implement and test framework to get input from syllabus	25%	25%	25%	25%
Select data mining techniques for each behavior	25%	25%	25%	25%
Provide report explaining behaviours and corresponding data mining methods	25%	25%	25%	25%
Implement and extract data for 2 behaviors	25%	25%	25%	25%
Update Requirements and test document	25%	25%	25%	25%

2.1 Discussion of each task

2.1.1 Finalize conceptualizing 8 behaviors

Conceptualizing a behavior will include examining the data and thinking of common behaviors that can be modelled from the data. There is no programming involved here, more of common sense and process analysis.

2.1.2 Finalize packages, languages and tools

Select between Java, Python or suitable frameworks to integrate both. Research 1 more tool per person and finalize tools for each behavior.

2.1.3 Implement and test framework to get input from syllabus

A program that will allow the user to enter data like due dates, test dates etc.

2.1.4 Select data mining techniques for each behavior

Select a data mining technique for each of the behaviors explored in 2.1.1. Specify language, packages and tools.

2.1.5 Provide report explaining behaviours and corresponding data mining methods

Provide report of work from 2.1.4 and 2.1.1

2.1.6 Implement and extract data for 2 behaviors

Implement atleast two behaviour extractions.

2.1.7 Update Requirement document

Update requirements document with all the specific data mining techniques that have been finalized in 2.1.4

3 Sponsor feedback on each task for current milestone

4 Appendices

4.1 Shreyas' Tool research report

4.1.1 Advanced Miner

Advanced miner seems to be an all in one tool that has features ranging from preparing data to creating specialized reports like risk analysis. This tool is a great resource for statistical analysis, but lacks the more lower level commands.
The language support includes java hadoop and SQL which are result of its big data specialization. The tool is also capable of creating formatted reports.
This may be a good fit for statistical aspects of the product but the community version has limited functionality in terms of memory usage thus restricting the scope which the free version can be used for. This tool may be effective in order to attain abstractions for some math pivoted data mining methods.

4.1.2 Ghost Miner

Developed by fujitsu Ghost Miner is a great tool that supports databases like spreadsheets as well as complex machine lerning algorithms.
The provided packages are highly flexible in terms of different data. Ghost miner also provides visualization support. Ghost miner has an intuitive GUI that facilitates ease of use.

4.2 Shiru's Tool research report

Gnome data-mine tools is growing collection of tools packaged to provide a freely available single collection of data mining tools. Sometimes, gone data-mine includes underlying command-line data mining applications and some other GUIs for these applications. Gnome-data-mine-tools works on the Linux platform, and support C/Python language. Gnome-data-mine-tools is free download online. It is work on Linux platform. After we installed it. There are includes some tools that in the package. For example, Apriori Association Rules, Bayes Classifier, Decision trees. GDM Apriori Association Rules is a Gnome2 GUI for building association rules from transaction data. GDM Bayes is a Gnome2 GUI for building Bayes classifiers from data. GDM Decision Tree is a Gnome2 GUI for building decision trees from data. Actually, I dont think GNOME is a good way data mining for our project. Because there have some problems for this tool. The first one is GDM does not provide API for program to extend function, it means programmers need to read and extract the source code by themselves. It is inconvenient. The second reason is GDM only work on Linux platform, cannot work on other platforms. I think one tools that support multiple platforms is better.

4.3 Mohammed's tool research report

4.3.1 ESTARD

ESTARD Data Miner is an easy to use tool that is designed for both skilled and regular users. One of the main strengths of ESTARD is predictive analysis which would be very

helpful in this project, because predicting the behaviors that lead to a better grade is one of the main tasks in this project. Another strength of ESTARD is that it is able to find hidden relations in both structured and unstructured sets of data. ESTARD has many useful features that provide different types of output, some of these features include What-if statistics, decision rules, and decision trees. There are not any significant disadvantages of using ESTARD for this project. However, more tools are necessary for better and more accurate information extracted in the project.

4.3.2 WEKA

WEKA contains machine learning algorithms written in Java for data mining tasks. This tool is fairly easy to use; it can perform many data mining tasks such as data preprocessing, visualization, clustering, regression and feature selection. Weka is widely used in several different programs such as algorithms for data analysis and predictive modeling. One disadvantage of Weka is that it does not support multi-relations data mining. Overall, this tool would be a good addition to our data mining tools selection as it is rated one of the top tools for finding data mining.

4.4 Yaqeen's tool research report

Finding ways to extract the information that we need from the data is one of the important steps toward completing the goal of this project. The two tools that I have been exploring are Tanagra and SPAD.

SPADis a mature data mining suite that provides powerful exploratory analyses and data mining tools, including PCA, MCA, clustering, interactive decision trees, discriminant analyses, text mining and more, all via user-friendly GUI. SPAD data mining approaches are Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Text Mining, Outlier Discovery, Data Visualization, Discovery Visualization, Social Network Analysis. Although SPAD is a powerful data mining software, it is a commercial licensed software that comes in French. These two points in SPAD will delay the process of extracting information. Also, the only website that we can download SPAD from is www.coheris.com, which is in French, too. In my opinion, SPAD is not the data mining software that we want to be using in this project.

Tanagra, data mining suite, is a free software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. Tanagra is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. Moreover, its an "open source project" as every researcher can access to the source code, and add his/her own algorithms. Also, Tanagras main purpose is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development

in this domain (especially in the design of its GUI and the way to use it), and allowing to analyze either real or synthetic data. In the other hand, TANAGRA does not include what makes all the strength of the commercial software in this domain; a wide set of data sources, direct access to datawarehouses and databases, data cleansing, interactive utilization ..etc. The last thing, Tanagra originally comes in French as well, but we can find it in English, too. From this concept, I think using Tanagra is better than using SPAD, but I wouldnt recommend it.

While exploring SPAD and Tanagra, I researched rapid miner. Weka, a software that I think is helpful in data mining, which is originally in JAVA. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from the researcher own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU (General Public License). Also, one of the helpful things about Weka is that you can apply it to big data.

In general, all the data mining softwares above are useful, but in order to choose a software that is most helpful we have to consider the main goal of the project. In my opining, from all the three data mining softwares, Weka is what I think most helpful. SPAD is powerful commercial tool that is not free nor is in English, which makes it hard to work with. Also, Tanagra doesnt include all what a commercial software would offer which makes it less useful in our case. In conclusion, Weka what I see best from what I gathered from my research.

5 Sponsor Evaluation

Sponsor: Please detach this page and return to Dr. Shoaff

Score (0-10) for each member: circle a score (or circle two adjacent scores for .25 or write down a real number between 0 and 10)

Shreyas Ugemuge	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10
Yaqeen AlKathiri	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10
Mohammed AlHabsi	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10
Shiru Hou	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10

Faculty Sponsor

_____

Signature

_____

Date