Nirdesh Bhandari
Senior Capstone Research
Spring 2017

Project Proposal

1. Backgrounds and Motivation:

Data is the currency of this century**.** With more than 2.5 quintillion bytes of data being created every day, there exists immense policy and financial implications to managing, analyzing and learning from this data [1]. Data can be analyzed to test and discover trends and patterns. Many of the largest companies these days are data driven in their operations. Machine learning algorithms are commonly used to learn from a given dataset and make predictions. While most machine learning algorithms are based around classification, clustering, categorizing or creating associative rules, they remain subject to errors in the dataset itself. At the end, the predictive accuracy of these machine-learning algorithms is largely susceptible to anomalies in the dataset itself. Furthermore, datasets can often contain variables that are correlated to each other or whose errors are occurring in a certain pattern.

I believe that by employing statistical tests and analysis to 'correct' a dataset by removing these anomalies or outliers and picking the right inputs, it is possible to improve the accuracy prediction when machine-learning algorithms are implemented on them.

2. Project Description

For this project, I plan on developing an interface to improve the predictive accuracy of some of the machine learning algorithms by bettering a given dataset. My proposed solution involves first reading data from a dataset, running statistical tests on it, cleaning up the dataset and then feeding this processed dataset to the correct machine learning algorithms. Tests will be have to be done on both the raw and the cleaned dataset to see whether or not I have made a difference in the predictive accuracy of these algorithms. Finally, I will plan on properly visualizing the results and the dataset.

2.1 Selecting a dataset

My first task will be to find an appropriate and large dataset to work with. This will involve looking for datasets that interest me from various data repositories. I will have to decide between discrete or time series data from a number. Next, I will have to decide whether I want to work with categorical or numeric data. My choices here will decide which machine-learning algorithm I will be using because various machine-learning algorithms are data type dependent. Furthermore, I plan on splitting this dataset such that I work with 75% of it and save the remaining 25% for testing. This will be done to avoid bias in testing when the dataset is fed into algorithms to see check how well the machine predicts data points it has never seen before.

2.2 Pre-Processing

The second part of my project will involve working heavily with RStudio- an open source user interface for the R language. Using RStudio I plan on automating the task of running statistical analysis on the input datasets. I will be using a series of hypothesis testing, autocorrelation correction and Prias-Winston procedure based on tests such as White's General test, Breusch-Pagan test, Glejser Test and Park Test. I will be using the Variance Inflation Factor, hettests, t-stat values and R squared values to decide which method of correct would work best. If working with numeric data, these tests will also tell me whether I should Generalize Least Squared regression (GLS), Ordinary Least Squared Regression (OLS), or any other regression model.

Overall, my goal during this phase will be to find the best way to alter the dataset. Removing certain values, assigning weights or even transforming certain variables, could help me achieve this goal. How well I am able to do on this phase and how much of these tests I automate will decide how much more the algorithms in the next phase can learn from the improved dataset.

2.3 Machine learning

Based on the type of dataset I selected in the first part, I will have to decide the correct algorithm for machine learning. I will be working with K-clustering, support vector machine (SMV) and Logistic Regression algorithms for machine learning from a dataset. All these algorithms are available under the Python scikit-learn module along with various other libraries for data processing.

2.4 Testing

After the improved dataset has been fed to the machine-learning algorithm, I will have to test whether or not the preprocessing made a difference. I believe the best approach to do this will be to have the machine learning algorithm learn first from the raw dataset and then test how well it predicts the values for the test-set (25%) we initially split off. Then, the same process can be repeated, but this time to check predictions after learning from the improved dataset.

The results I obtain here will need to be quantified and properly analyzed. Here, I will have to use various matrices of error classification to properly assess my results.

2.4 Visualization:

Lastly, my software module will then work with the Tableau interface to visualize the results that I obtained in my tests.

2.5 Library Research

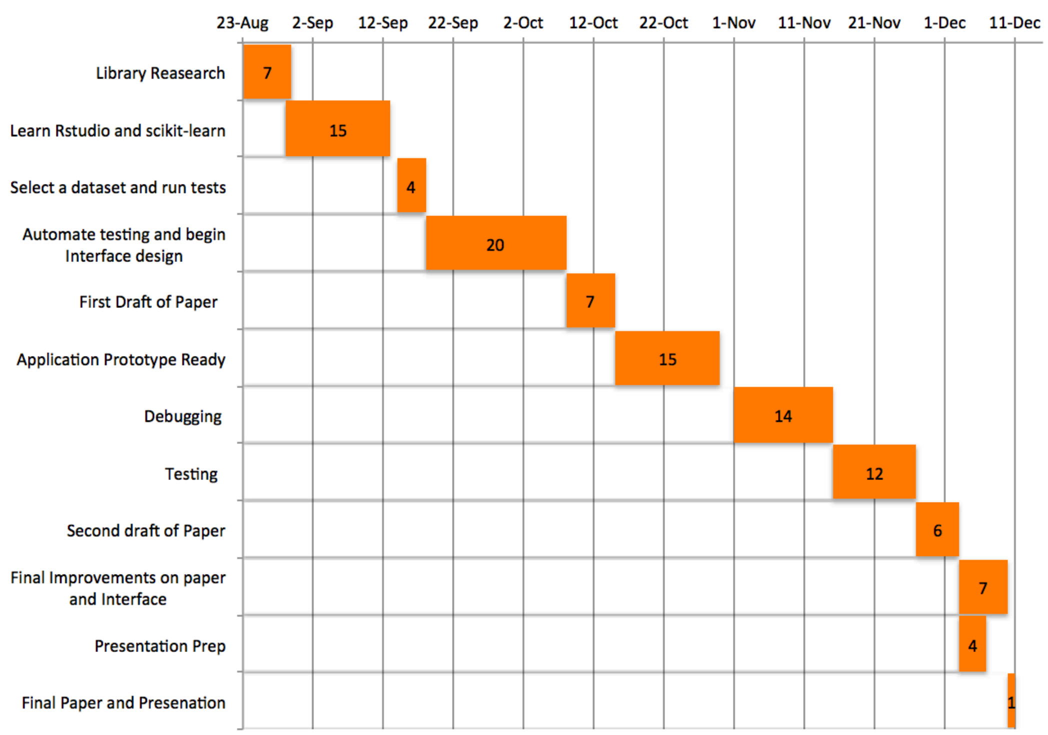
For this project, I will have to learn RStudio, Python scikit-learn module, R programming language and some intermediate statistics. I will have to look into some similar work people might have done in the past and how they have implemented these tools. I have saved up some citations from the papers that I read for my Survey paper, which I will have to dig into more as well. A good design flow would definitely help get started on the right foot.

2.6 The Final Product

The final product I plan on making will be an interface to automate all my sub modules into one. This will possibly work with Python and RStudio at the same time and will feed the results into an excel sheet that Tableau will use to generate visualizations. It will contain buttons to select what tests to run on RStudio, how to improve the dataset, which algorithm to use and how to view the results.

4. Plan

For the successful completion of this project, I have worked on allocating my time based on how much effort I anticipate I will need on each phase. I have also left ample space to catch up if I fall behind on any phase. The fall break is at Nov 18-26 and the early semester break is at Oct 12-15. I have set these times to work on the first draft of my paper as well as the debugging and testing so that I can allocate extra time on these segments when I don't have classes going on. Furthermore, I plan on getting some bits of my Library research done during this summer as well. The following is my planned project timeline:



References:

1. X. Wu, X. Zhu, G.-Q. Wu, W. Ding, "Data mining with big data", *Knowledge and Data Engineering IEEE Transactions on*, vol. 26, no. 1, pp. 97-107, Jan 2014.