Kevin Dombrosky, Jacob Siegel           3/16/15

Project Proposal       CSCI-420

Kevin Dombrosky and Jacob Siegel plan on working as a group for the CSCI-420 Principles of Data Mining project. Because of the small group, both members are expected to complete written requirements, development of the project, especially regarding software design and programming.

The topic our group is exploring is attempting to predict which team will win based on pre-game statistics in the competitive video game, League of Legends. League of Legends is developed by Riot Games, Inc. In the game, two teams of five players choose a unique champion to play as and fight each other in an incredibly strategic game where the ultimate goal is to destroy the opposing team’s base. Often times, these unique champions act in a rock, paper, scissors manner, where one champion may be easier to achieve victory against another, and that champion is better against another, and so on. The community behind League of Legends has already begun collecting the data of which champion is individually better against another, but League of Legends is a team game. While this microcosm one versus one occurs in the match, there are eight other players in the game that could ultimately determine the outcome of the game. Additionally, League of Legends was released in 2007 and with millions of matches having already been played since then, players have had the opportunity to look to how practicing a champion affects their ability to win a game, as well as how their recent performance can affect their ability to play. For example, a player on a losing streak in a session of playing is more likely to lose than one that is not. In summary, the League of Legends community has many statistics related to it, some maintained by Riot Games, and others by the community, that can be used in an attempt to accurately predict the outcome of a game before it is played based on the champions chosen and the players behind them.

Riot Games, Inc has experienced a great amount of success with League of Legends, and as a result has been able to allocate resources to produce and maintain an API to access game and player data. This data ranges anywhere from static data about aspects of the game, to historical data of players past games, to even data for matches in progress. The API has seen steady maintenance to keep it up to date and to provide the most in depth data possible to its users. It is extremely well documented and returns data exactly as described within the documentation, guaranteeing clean data to anyone who uses it. The API allows registered users to issue HTTP GET requests and to receive JSON data back. It is through this API that we intend to get the majority, if not all, of our data.

To solve the problem we face of predicting if a team will win or lose, we will use several different methods to try to classify the game as a win or loss as well as cluster the data received. To cluster the data, we will use DBSCAN as it clusters data based on each records distance to another and the resulting point densities. This seems to suit the data set well, as well as allowing us to identify which games would be more difficult to predict. Additionally, to classify the data, we intend on using the Naive Bayes, linear regression, J48 decision tree, and k-nearest neighbor algorithms. The Naive Bayes and linear regression algorithms will perform well based on the statistical independence of the attributes that we are analyzing. K-nearest neighbors employs use of the distance differences of the many numeric attributes that are being analyzed. Finally, the J48 algorithm has the ability to deal with both continuous and discrete attributes, both of which we will find useful when performing our analysis. We will need to generate test and validation data by parsing many games’ data retrieved from the Riot Games API. Further, the test data games should exclude games with players below a certain ranking threshold, as the game becomes less about team strategy and more about individual mechanical ability and decision making the lower the ranking of players in the game. Finally, data shall be analyzed strictly from the five versus five standard format of the game. To evaluate the successfulness of the classification algorithms, we will evaluate the misclassification rate of the results of those algorithms against a subset of the data.

If we see great success with the algorithm that we implement, it could become a valuable tool for the League of Legends community and could also be a commercial opportunity. May other tools used for League of Legends have entered the mainstream and receive millions of uses a day when they were published as a web application and made open so that anyone could use them. Based on the data that we are analyzing we believe and have high hopes that a useful tool could be created from this project.