Josh Bhattarai

CSE 482 Big Data Analysis

Dr. Pang-Ning Tan

March 24th 2019

Intermediate Report

Predictive Seeding and Outcome for Ultimate Brackets

Abstract: Michigan State University hosts the largest weekly series of competitive tournaments for the video game Super Smash Brothers out of any other college in the world. Every Friday, an average of 100 entrants need to be seeded into the brackets by the tournament organizers manually. This project's goals would be to collect data from previous MSU tournament results and automate the seeding process for future brackets.

Data Sources: The first source will be smash.gg. The brackets are entered in real time during the tournament and match data such as the players, winner, and set count are recorded. The data will be accessed using the GraphQL API. The second source will be challonge.com. These brackets are also entered in real time, however the website has no API. Instead a JSON file is provided via a download URL and will need to be parsed. There is no constant way of entering player names so they will be messy and require further processing.

Preprocessing Steps: First, the URLs for the different tournaments were entered into a text file manually. Using the smash.gg public REST API, information about the final standings of each tournament were gathered. Smash.gg's GraphQL API is in early alpha and is subject to changes without warning so in the future this project may be updated to make use of the GraphQL API, but for now the REST API is being used, along with a python wrapper. From the standings information, a three-dimensional feature array was created containing the final placement, name and seed disparity from each entrant of the tournament. Second, a dataframe was constructed in order to contain the head-to-head record between each player in the tournament. Using each player as the column and as the row indexes, the value in each intersection is an array of a set object containing each player's score during the match as well as the winner of the match. The same steps were used for the challonge brackets, but the extracted json file had to be parsed through instead of the multiple endpoints. The features extracted are, final placement, seed disparity, head-to-head sets between each player and score for each set.

Analysis Techniques: I'm not quite sure about the exact the machine learning technique that will be best for this project. I've seen many people apply logistic regression to predict the results of different sporting brackets and tournaments. I will most likely pull from many more datasets now that the preprocessing has been automated. I could add University of Michigan and Central Michigan University's brackets into the program to avoid overfitting.

Analysis Concerns: While attempting to automate the data collection and preproccessing portions of this project, loading in 6 brackets took about 30 seconds. I'll either have to optimize my data collection, data preprocessing, find more machines to run my program on, or multithread my program. Another concern is that the only features I have other than the head-to-head data is the match score and the results for past tournaments.

Timeline:

February 11th: Begin work on project. Completed
February 11th – 18th: Obtain API keys for smash.gg and work on challonge parsing. Completed.
February 18th – March 24th: Finish data collection from smash.gg and challonge. Completed.
March 24th – April 21st: Finish data analysis and return a text file of participants seeded in order.
Stretch Goals: Create a preseeded bracket in smash.gg and challonge using the collected data, account for frequently played matches, account for teammates or out of state competitors.