Project Proposal

Chi Jin
Michigan State University
jinchi@msu.edu

Connor Masini
Michigan State University
masinico@msu.edu

October 8th, 2018

1 Introduction

American football, also known as football or gridiron, is the most popular team sport game in the United States[1], attracting hundreds of millions of people all around the world to an ever-growing fan base. Between 2010 and 2013, an estimated average of 3.75 million people played football in some form of organized league. [2]

Like all popular sports, football has a prosperous betting market. Both the bookmakers and the bettors wish to make money by betting on the right team. Before each major football game, a predicted point spread is generated. Bettors can bet whether the total number of points scored between the two teams is higher or lower than a value determined by the bookmakers or on whether the difference in scores between the two teams will be higher or lower than the bookmakers prediction. Therefore, the prediction of the game scores is critical for betting in bookmaking business as both major forms of betting rely on the final scores for the games.

From an academic standpoint, this task can test various mathematical methods of prediction and extrapolation as a benchmark[3]. Due to the high turnover rate in college football (players can only play for a maximum of 4 years), the performance of any given team can vary drastically from year to year. The high turnover rate limiting the usefulness of previous years data in addition to the limited number of games each team plays makes the prediction of college football scores a very interesting and difficult problem.

2 Previous Work

There has been a large amount of work done towards accurately predicting the scores of football games. This can be attributed to the popularity of the sport in America in addition to the potential monetary gain associated with successfully classifying game scores. However, one important distinction about this work needs to be stated: most of the research completed has been done on NFL teams instead of NCAA teams. Predicting NCAA football scores is a much more difficult task than predicting NFL scores, due to the vastly greater number of teams and smaller amount of games in college football [4]. However, the principles used for predicting professional football should also apply to college football. Many previous studies have utilized artificial neural networks to a great degree of success in predicting the outcome of professional football games [4, 5, 6]. In addition, the usage of genetic algorithms to tune neural network parameters and hyper-parameters is a common practice that effectively optimizes parameters and hyper-parameters[7].

3 Method

3.1 Genetic Algorithms

Genetic algorithms (GAs) have long been developed for prediction tasks like predicting upcoming weather patterns or protein structures.[8] Specifically, GAs are used to evolve machine learning system parameters, like weights and hyper-parameters, like learn-

ing rate, number of layers, or number of neurons of a neural network.

3.2 Neural Network Approach

In this project, we aim to design a genetic algorithm to tune a fully connected neural network for predicting the score of a game between two teams. Four hyper parameters are considered for tuning: number of layers, number of neurons per layer, the layer activation function and network optimizer. The initial population is a set of N random networks. We will construct a fitness function using the accuracy of the actual team scores against the predicted scores. For each generation, a certain number of networks will be chosen to have their hyper-parameters mutated. These mutated networks along with the networks that were not selected for mutation will then undergo crossover to generate the next generation. To determine which parents will be selected for reproduction, we will use fitness-proportional selection. Each child will be generated by accepting a random assortment of parameters from its two parents. The entire run will be terminated when the average fitness of a generation reaches a desired value or after a certain number of generations have passed.

4 Data

The data used for this report has already been gathered and is hosted on Kaggle at `https://www.kaggle.com/mhixon/college-football-statistics`. This data set contains detailed game information for every game in the 2005-2013 seasons. This data set contains play-by-play information about each game, in addition to total amounts of plays, yardage gained, points scored, and many other useful pieces of information. Because this data set has play-by-play information, it encapsulates virtually every aspect of a football game. Therefore, we should be able to derive any statistics we want relating to the game that the data set does not already summarize for us. In particular, previous studies have shown that four stats determine the outcome of a football game: yards gained, rushing yards gained, turnover margin, and time of possession.[6].

4.1 Feature Engineering

Because almost every aspect of a football game is recorded in our data set, we have an enormous amount of data to explore and choose from. We plan on using totals of all the major stats throughout the game. As an example, we will look at both the total number of yards rushed in addition to the number of rushing plays a team attempts. This will allow for efficiency patterns to emerge in the data. In addition to the basic statistics present in the data set, we will experiment with using advanced metrics used to analyze teams, including offensive and defensive efficiency, glicko scores, strength of schedule and S&P+ ratings. The combination of these features should provide our neural network with enough information to effectively predict game scores.

References

[1] L. Shannon-Missal. Pro football is still america's favorite sport. 2016.

[2] E. Irick. Ncaa sports sponsorship and participation rates report. 2014.

[3] John wiley. Forecasting methods and applications. 2008.

[4] Andrew D Blaikie, Gabriel J Abud, John A David, and R Drew Pasteur. Nfl & ncaa football prediction using artificial neural networks. In *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics, Denison University, Granville, OH*, 2011.

[5] John A David, R Drew Pasteur, M Saif Ahmad, and Michael C Janning. Nfl prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports*, 7(2), 2011.

[6] Michael C Purucker. Neural network quarterbacking. *IEEE Potentials*, 15(3):9–15, 1996.

[7] M Bashiri and A Farshbaf Geranmayeh. Tuning the parameters of an artificial neural network using central composite design and genetic algorithm. *Scientia Iranica*, 18(6):1600–1608, 2011.

[8] Melanie Mitchell. Genetic algorithms: An overview. *Complexity*, 1(1):31–39, 1995.