Advanced Machine Learning NFL

Adithya Seshadri, Alexander Sobran, and Anand Vijayaraghavan

October 1, 2015

Purpose

The NFL is highly unpredictable, being highly notorious for unexpected results. As a consequence, NFL prediction is a fertile ground for assessing the performance of various machine learning tools.

The purpose of this project is to explore the effectiveness of advanced machine learning techniques using multiple tiers of NFL data. This project proposes that machine learning techniques can be used to predict the next play given a series of previous plays, the winner of a game, and the performance of a player in a single game. These predictions could be useful when applied to live play calling strategy, personnel analysis, sports betting, and fantasy football.

This project will also compare various machine learning algorithms and assess their strengths and weaknesses. We will compare optimization techniques for parameter and structure optimization. We will also examine the effect of data aggregation as data can be trained on a play level, drive level, quarter level, half level, game level, or season level.

Previous Work

Las Vegas correctly selects the correct favorite at a rate of 61.2-74.7% from 1989 - 2012 [3]. Previous work [1] has achieved a success rate of 64%. The NFL also has one of the lowest rates of upsets among all sports at 36.4% [4].

There is no history of research for play prediction based a play by play data. There is also no research predicting the performance of players all though many fantasy football pundits may offer projections. This has been omitted as aggregating a useful dataset for comparison is not feasible.

Data Sets

Play by Play Data Consists of a summary of each play in each game including, time, down, yards to go, pass/run, distance, current score, as well as other metrics. From 2009-present.

NFL Combine Data A set of metrics used to measure the athleticism of players as well as one test to measure mental aptitude.

Coach History Head coach, offensive coordinator, and defensive coordinator for each team.

Rosters The players that make up a team.

Weather Weather of games.

Injury History Football is a contact sport with high a frequency of injuries.

Team Offense/Defensive Average aggregated data viewed as team statistics.

Temperature Differences Temperature differences of team's cities. Used in previous work [1]. Gathered from weatherunderground.com

Algorithms

Deep Belief Nets Structure of deep belief nets will be varied and compared to determine the most useful structure. Use in play by play data can be represented as a time-series. Application of deep belief nets has been successfully applied to time-series data in previous work using time-series data [8] by treating the previous historical data as the feature set.

SVMs with Kernel Transformations - The NFL game prediction has been previously done using SVM and kernel transformations in [5] and it it uses linear, polynomial and tangent kernels. This can be applied to perform play prediction.

Hidden Markov Model Markov Models can be used to predict plays by assessing the current state if the game. This can be used as another strategy to predict plays without considering the sequence of previous plays. A Markov model of football has be described in [9].

Logistic Regression This will be used as a benchmark as previous work has mainly used logistic regression for prediction.

Ensemble Ensemble of the above. After optimization of the the above models an ensemble will be explored using weights.

Optimizations

Particle Swarm Optimization This optimization was used to optimize deep belief net structure and parameters in previous work [8]. We will apply it to deep belief nets as well as the other techniques listed above.

Differential Evolution This optimization technique will be used on the above algorithms. Comparisons will be made between this and particle swarm optimization.

Study Plan

1. Data aggregation
2. Data pre-processing: Parsing of play by play data. Mapping coaches.
3. ML Implementation
4. Optimization Implementation
5. Ensemble Implementation
6. Evaluation of models
7. Comparison of performance of different tiers of data

Testing Validity

1. Accuracy of predicting type of next play(run, pass, punt, field goal), strategic location of next play(left, right, inside, deep, short, middle)
2. Accuracy of predicting player performance in game
3. Accuracy of predicting winners to be measured against the previous research of 64% [1], the historic rate for bookies, as well as the upset rate for teams.
4. Regression accuracy of predicting scores.
5. Analysis of bias and variance of various models.

Potential Challenges

The size of the dataset is relatively small as each teams plays 16 regular season games and run approximately 50-60 players per game. We hope to mitagate this issue by training on the historical data going back 6 years. It could be assumed coaches talented enough to earn multi million dollar contracts do not have predictable tendencies as they would be exploited by opposing coaches.

Deliverables

A report, source code, and data will be delivered. This project will be tracked and managed at a publicly available Github repository: `https://github.com/aisobran/Adv-ML-NFL`

References

1. http://www.cs.cornell.edu/courses/cs6780/2010fa/projects/warner_cs6780.pdf
2. http://ttic.uchicago.edu/~kgimpel/papers/machine-learning-project-2006.pdf
3. http://www.inpredictable.com/2013/08/is-nfl-betting-market-getting-more.html
4. http://www.mpipks-dresden.mpg.de/~federico/myarticles/sports-jqas.pdf
5. http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf
6. http://www.incrediblemolk.com/aws-machine-learning-example-nfl-player-positions/
7. http://www.engineering.leeds.ac.uk/e-engineering/documents/JackBlundell.pdf
8. http://www.sciencedirect.com/science/article/pii/S0925231213007388
9. http://archive.advancedfootballanalytics.com/2011/09/markov-model-of-football.html