CSE 445 Machine Learning
2018 Summer
September 11 2018
Phase: Final

Using Machine Learning to Predict the Outcome of Football Matches in the Premier League

Team: El Equipo

Name: T.M Rezoan Tamal
ID: 1411671042
Section: 2

Name: Nawar Chisty
ID: 1411060042
Section: 1

Name: Muhibur Rahim
ID: 1331073040
Section: 1

Abstract—**Football, as we know it, is the most popular game in the world. With an estimated 3.5 billion fans around the world, it is enjoyed by half the population in the world []. The goal of this project is to predict the team-agnostic match outcome. The dataset includes the premier league table from season 2014/15 to 2017/18. It includes points, possession, goal scored, results. The models presented include Neural Network (NN), Support Vector Machine (SVM) and Decision Tree.**

Keywords—Support Vector Machine, Decision Tree, Neural Network, machine learning, football, J48.

I. Introduction

Football is a multi-billion dollar sports industry played by over 250 million football players in over 200 countries. Football is the national sport of many countries in the world and the love of the sport transcends national and international borders[1]. The Premier League, is the top level of the football league system in England. It is played by 20 clubs, it operates on a system of promotion and relegation based on points earned over a season. In this project, we have a multi-class classifier to predict the outcome of a match over a set of input data that is a database containing league points, possession, goal scored and actual outcome.

We have used SVM with the implementation of Sequential Minimal Optimization (SMO) and J. Ross Quinlan's implementation of Decision Tree (J48) and also Neural Network as the learning algorithms.

II. Related Work

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Similar research in the area of soccer match prediction falls generally into two categories: time-series/Markov based, and non-time-series. The time-series models include that by Rue and Salveson[2], which uses a Monte Carlo time-series to simulate match results using only goals scored in prior games.

Koopman and Lit[3] use a bivariate Poisson distribution with coefficients varying over the course of the season according to a custom stochastic process.

The non-time-series research includes that by Goddard and Asimakopoulos[4], who studied the impact of match importance, distance travelled, and recent indicators of form on match outcome. But they focused on testing economic price efficiency of the betting markets rather than match prediction.

Titman et. al.[5] explore the interplay between yellow and red card issuance ("bookings") and goals in a non-time-series manner, but primarily concern themselves with the effect of those bookings on the outcome of the game at the moment of occurrence, rather than predicting prior to match start.

Ulmer and Fernandes[6] sought to do team-specific prediction rather than making team-agnostic predictions. For example, they sought to predict whether Tottenham Hotspur would win a given game they played, whereas we seek to predict any game correctly, regardless of the participating teams.

Kevin Bishop[7] made team-agnostic prediction based on average values of 16 features over the previous season of play for each team and each statistic, respectively.

We achieved a better success rate on the three-class win vs. loss vs. draw problem (58%) compare to Kevin Bishop's three-class classifier (52%). We also had far superior results (83%) in terms of successfully classifying two-class win vs. loss/draw compared to what Ulmer and Fernandes did for the same (48%).

III. Methodology

Although there seems to be an abundance of data on soccer games available online at first look, obtaining a data set amenable to machine learning proved challenging[8]. First we collected a dataset from kaggle[9] that contains Premier League match statistics of the past 4 seasons. From that dataset we collected only a few attributes that we were interested in, including the name of the team, percentage of ball possession, scored goals. From the number of scored goals of both the home and the away team, we obtained the match result. After that we used points from the previous season's league table as a measure of the strength of a particular team. Name of the team was then replaced by their respective points. The strength of the opposing team was later added to highlight the difference between the home and away side. League table data were obtained from ESPN's website[10]. After the data processing was done we were left with five attributes for each data: strength of the team, strength of the opposing team, the team's overall possession of the ball, number of scored goals, and the outcome of the match from the team's perspective.

We used three class classification (win/draw/loss) to begin with. After analyzing the results of it, a two class classification (win/draw or loss) was also implemented. We used SVM, Decision Tree, and Neural Network as our classifiers as they offered better overall accuracy.

A. Decision Tree

Decision trees classifiers are simple and prompt data classifiers as supervised learning means with the potential of generating comprehensible output, usually used in data mining to study the data and generate the tree and its rules that will be used to formulate predictions [11].

This project represents an implementation of a J48 algorithm analysis tool on the data collected. J48 is one of the most used Weka classification algorithms that offers a superior stability between precision, speed and interpretability of results [11].

B. Support Vector Machine

In the last few years, there has been a surge of interest in Support Vector Machines (SVMs)[12][13]. SVMs have empirically been proved to give good generalization performance on a wide variety of problems such as handwritten character recognition[14], face detection[15], pedestrian detection [16], and text categorization[17].

The SVM algorithm that is used in this project is SMO. This algorithm is conceptually simple, easy to implement, generally faster and has better scaling properties for hard SVM problems than the standard SVM training algorithm.

C. Neural Network

NNs have been used to learn patterns in the data. A NN is modelled in the way a human body pass or suppress signals. Weights are associated with each of the input and activation functions are applied to their weighted sum (including a bias term) to get new sets of inputs. There can be multiple such hidden layers before reaching a final layer with the required classification [18].

As we will see in the results section, the confusion matrix shows better percentages of accuracy in the neural network than SVM and Decision Tree.

IV. Results

As expected, it was quite difficult to predict the exact outcome of a football match without a really detailed set statistics of a match. On the other hand that would defeat the purpose of the project, which is to predict the outcome of a match based on publicly available data.

We initially started off with a three class classification problem. The results were promising. From 10 fold cross validation, we obtained test accuracy of 56.45% for decision tree, 59.20% for SVM, and 59.88% for neural network. But we noticed some interesting trend from the obtained confusion matrices.

Actual vs Prediction	Predicted Defeat	Predicted Draw	Predicted Win	TP Rate
Actual Defeat	376	163	99	58.9%
Actual Draw	309	153	154	24.8%
Actual Win	25	74	593	85.5%

Table 1: Confusion Matrix for Decision Tree

Actual vs Prediction	Predicted Defeat	Predicted Draw	Predicted Win	TP Rate
Actual Defeat	340	214	84	53.3%
Actual Draw	238	245	133	39.8%
Actual Win	30	73	535	83.9%

Table 2: Confusion Matrix for SVM

Actual vs Prediction	Predicted Defeat	Predicted Draw	Predicted Win	TP Rate
Actual Defeat	379	144	115	59.4%
Actual Draw	291	150	175	24.4%
Actual Win	22	12	602	94.7%

Table 3: Confusion Matrix for Neural Network

It was evident from the confusion matrices that the classification algorithms were able to predict winning classes more precisely from our dataset in relation to the chosen attributes. But they struggled to predict draws, misclassifying in most cases which resulted in accuracy rates ranging from 24.4% to 39%, almost close to random probability of 33%. Intuitively, this makes sense considering that wins and losses can be quite lopsided, while draws are by nature very tight affairs.

At this point we decided to classify for two classes, win and not win (draw/defeat), to see whether that improves classification accuracy on both sides. As expected, the classification accuracy improved indeed. From 10 fold cross validation, we obtained test accuracy of 83.87% for decision tree, 83.88% for SVM, and 83.98% for neural network. And as usual, the confusion matrices offered some interesting insight.

Actual vs Prediction	Predicted Draw/Defeat	Predicted Win	TP Rate
Actual Draw/Defeat	1126	128	89.8%
Actual Win	177	461	72.3%

Table 4: Confusion Matrix for Decision Tree (Binary)

Actual vs Prediction	Predicted Draw/Defeat	Predicted Win	TP Rate
Actual Draw/Defeat	1153	101	91.9%
Actual Win	204	434	68.0%

Table 5: Confusion Matrix for SVM (Binary)

Actual vs Prediction	Predicted Draw/Defeat	Predicted Win	TP Rate
Actual Draw/Defeat	1045	209	83.3%
Actual Win	94	544	85.3%

Table 6: Confusion Matrix for Decision Tree (Binary)

With binary classification, for decision tree, and SVM, classification accuracy decreased in terms of predicting wins, but improved massively in terms of classifying draw/defeats. This is due to the fact that due to the nature of the dataset, there are more data available with class label draw or defeat that the class label win. More training data resulted in more accurate prediction for such difference based classifiers. On the other hand, neural network learned to pick up little patterns, and after a bit of optimization, produced impressive accuracy for both classes.

Overall, the Neural Network performed better. It had better accuracy for each class of both 3-class classifier and 2-class (binary) classifier. This was in contrast with the evidence from our related literatures, which suggested that the SVM was the most effective method. But for us, a more optimized neural network produced far better results.

V. Conclusion

Football matches are difficult to predict. Because Football, at its core, is a game of incredible moments mostly created by glorious match settling goals. And yet, goals are not always indicative of the overall match. For instance, one effective counterattack over a 90 minute game, combined with some stellar defense and luck, can win the game in the face of massive statistical dominance. In 2011-12, Chelsea won the Champions league with a tactically, and statistically inferior team. In 2016-17 season, Leicester City won the league with statistics much lower the top level teams. Our little endeavor was to get a better understanding of the method to that madness. Our tests have shown that irrespective of possession of the ball, the strength of a team, and the ability to score goals were more accurate indicators of the outcome of the game. Whereas possession, even though perceived as an indicator of dominance, were not indicative of how the results may turn out to be.

Our three class classification showed that draws are the hardest to predict, and massively impact prediction accuracy. Hence our subsequent attempt to classify with two classes significantly increased our accuracy.

In terms of predicting the outcome of a game, Neural Network produced the best results. But the other algorithms performed admirably too as they produced significant prediction accuracy, and perhaps helped us understand football just a little better.

References

[1] Schumaker, R. P., Jarmoszko, A. T., & Labedz, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. Decision Support Systems.

[2] Rue, H. and Salvesen, O. (2000), “Prediction and Retrospective Analysis of Soccer Matches in a League”. Journal of the Royal Statistical Society: Series D (The Statistician), 49: 399–418. doi:10.1111/1467-9884.00243

[3] Koopman, S. J. and Lit, R. (2015), “A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League”. J. R. Stat. Soc. A, 178: 167–186. doi:10.1111/rssa.1204

[4] Goddard, J. and Asimakopoulos, I. (2004), “Forecasting football results and the efficiency of fixed-odds betting”. J. Forecast., 23: 51–66. doi:10.1002/for.877

[5] Titman, A. C., Costain, D. A., Ridall, P. G. and Gregory, K. (2015), “Joint modelling of goals and bookings in association football”. J. R. Stat. Soc. A, 178: 659–683. doi:10.1111/rssa.12075

[6] Ulmer, B., Fernandez, M. (2014), “Predicting Soccer Results in the English Premier League”. Web. 21 November 2016. https://goo.gl/Zdj87n.

[7] Bishop, K., Data-driven insights into football match results, 2016.

[8] Smolka, S, BeatingtheBookies: Predicting the Outcome of Soccer Games, 2016.

[9] https://www.kaggle.com/shubhmamp/english-premier-league-match-data.

[10] http://www.espn.in/football/table

[11] V. P. B. felean, "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment," 2007 29th International Conference on Information Technology Interfaces, Cavtat, 2007, pp. 51-56. doi: 10.1109/ITI.2007.4283743

[12] Vapnik, V., Estimation of Dependences Based on Empirical Data, Springer-Verlag, (1982).

[13] Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, (1995).

[14] LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P. and Vapnik, V., "Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition," Neural Networks: The Statistical Mechanics Perspective, Oh, J. H., Kwon, C. and Cho, S. (Ed.), World Scientific, 261-276, (1995).

[15] Osuna, E., Freund, R., Girosi, F., "Training Support Vector Machines: An Application to Face Detection," Proc. Computer Vision and Pattern Recognition ’97, 130-136, (1997).

[16] Oren, M., Papageorgious, C., Sinha, P., Osuna, E., Poggio, T., "Pedestrian Detection Using Wavelet Templates," Proc. Computer Vision and Pattern Recognition ’97, 193-199, (1997).

[17] Joachims, T., "Text Categorization with Support Vector Machines", LS VIII Technical Report, No. 23, University of Dortmund, ftp://ftp-ai.informatik.unidortmund.de/pub/Reports/report23.ps.Z, (1997).

[18] Arabzad, S. Mohammad & Araghi, Me & Soheil, Sadi-Nezhad & Ghofrani, Nooshin. (2014). Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. International Journal of Applied Research on Industrial Engineering. 1. 159-179.