Football Match Result Prediction Using the Random Forest Classifier

Pakawan Pugsee, Pattarachai Pattawong

Innovative Network and Software Engineering Technology Laboratory

Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, THAILAND

pakawan.p@chula.ac.th; splitario_por@hotmail.com

ABSTRACT

Nowadays, people are turning their attention to football as the business investment more and more, especially invest in joining football club. In addition, If the results of some matches do not meet the goal of the clubs, the investors will not invest in the club and the club may be loss a lot of money that they should be. So, we have developed the web system for a football match result prediction method in order to help making investment decisions for investors and generating the guidance for developing their football teams. The objective of this project is to predict the football match results for the English Premier League, and to analyze factors affecting the outcome of the match for guiding team improvement. This project collected previous three-season match information from www.premierleague.com to predict the current league season match results. All collected data were analyzed by the machine learning technique for building a football match result prediction model, and for finding factors affecting on football match results to give advice for improving their football teams. The testing results of the prediction are shown that the accuracy and the precision are more than 70%. Therefore, this system can help the user get the guidance for improving the football team and the precise prediction of football match results.

CCS Concepts

• **Information systems~Data analytics** • Computing methodologies~Machine learning approaches**.**

Keywords

Football match result prediction; Random forests; The English Premier League.

1. INTRODUCTION

The development of the football match prediction system started by researchers having a passion for football games and have followed the sporting events for a long time, especially the famous football league, called the English Premier League. The English Premier League is the highest-level football league in England with famous world-class football teams. In addition, this league has the competition for succession among football clubs higher than other leagues. The football competition system is that all teams will meet all the other teams for home and away matches (when a team is serving as host of a contest, called as the home team; the opposing team is the visiting team, named as the away team). There are composed of 20 football teams in the league, so each team must compete in 38 matches for the whole season. The scoring system is that the winning team will get 3 points, while each team get 1 point, when two teams draw, and there is no point for the losing team. At the end of the season, the team with the highest score will be the champion. In addition, the first four teams with the highest score will compete in the UEFA (Union of European Football Associations) Champions League, which is the biggest football league among football clubs in Europe.

ICBDT2019, August 28–30, 2019, Jinan, China

ACM ISBN 978-1-4503-7192-6/19/08…$15.00

https://doi.org/10.1145/3358528.3358593

There are some articles about the prediction of football match results. For examples, the paper [1] applied statistical learning techniques to predict the outcome of a football match. The frequency counts of in-game events from the Manchester City Analytics program were used to generate predictive models with multinomial regression for the 2011-2012 Premier League season. The articles [2] and [3] used data gathered from video game FIFA (Federation International Football Association) to improve prediction quality. The prediction model in [3] was generated by logistic regression using the training data from the 2010-2011 season until the 2015-2016 season. The variables of match results that were "Home Offense", "Home Defense", "Away Offense", and "Away Defense" were features to learn and predict sport match results. The highest accuracy of experiments by altering seasons of training data was 69.5%.

Moreover, the research [4] predicted the results of football matches using several machine learning techniques, such as Naive Bayes, Bayesian networks, the k-nearest neighbors, random forest, artificial neural networks. Many experiments were run to optimize the combination of features ("Home Win", "Draw", "Away Win") and previous classifiers were tested for developing the prediction system. The satisfactory capability of classification was shown in the results with the accuracy about 50%-69%. The next research [5] built a goal score prediction model and this model was tested with the results of the FIFA World Cup 2014. Latent features obtained from matrix factorization process were used for the model generation using Naive Bayes Classifier based on the betting quotas. The researchers claimed that their algorithm can be used to estimate accuracy of an expert knowledge-based system for match result prediction. Another predictive system for football match results [6] was improved the performance of prediction accuracy by using only related features to the football match results. The system was implemented by the artificial neural network and logistic regression techniques with the prediction accuracy of 85% and 93%, respectively. The feature sets of the system were a lot of data concerning the match history record, the performance index record and the football spreadsheet,

such as Home and Away goals, Home and Away shorts, Home and Away corner, Home and Away Odds, Home and Away attack strength, Home and Away Players' performance index, Home and Away Managers' performance index, Home and Away managers' win, Home and Away streaks.

Furthermore, due to the FIFA approval of the use of Electronic Performance and Tracking Systems (EPTS) during competition, there is the availability of novel data regarding physical player performance. The data analysis [7] provided competitive advantages of football teams by predicting match performance from training and physical information. Some machine learning with feature selection techniques and Principal Component Analysis (PCA) are applied to generate regression models. The analysis result revealed that the amount number of variables of physical information can be reduced for the period analysis because some specific variables can represent the set of highly correlated data.

Although all previous research found that there have been various techniques for implementing the prediction models of football match results with different the number of considered factors, this problem has still been an interesting challenge because of widespread and popular sports. In addition, if the results of football matches do not meet the expectations of football clubs, some bad effects on the football team can be occurred, for an example; the team is dropped out of the league. Therefore, our research tried on generating football match prediction model with the simpler machine learning technique and minimal information without the biased judgment on the person considered primarily. To design the football match result prediction, this proposed method used data from www.premierleague.com [8], which is the main official website of the English Premier League. In our research, two selected machine learning techniques are random forests based on decision trees and the deep learning with multilayer perceptron. Moreover, to create a guidance for increasing competitiveness of their football teams, the correlation analysis is applied to analyze related factors that affect competition results.

2. BACKGROUND KNOWLEDGE

2.1 Standardization for a Machine Learning

Standardization for machine learning is making all the information on the same scale level. The average value of the attribute is 0 and the standard deviation is 1. In each attribute, the average value and the standard deviation of the data are determined. Then, the original data are replaced with the mean and divided by the standard deviation following (1). Each attribute data will be in the same scale after process standardization, which help to improve the quality of training data affecting to better machine learning efficiency.

$$x' = \frac{x - \bar{x}}{\sigma} \qquad (1)$$

2.2 Random Forest Classifier (RFC)

The random forest classifier is one machine learning technique to create classification models starting from randomize training data into many different data sets. Then, each training dataset is executed to create an individual classification model by the decision tree, which will display the data classification as a tree and identify data in nodes with impurity of data. The variation of classification depends on each data set and designed attributes for avoiding overfitting problems of only one decision tree. The impurity measurement of entropy data will be used for data classification with information gain to calculate the knowledge of each attribute. The set of attributes with the most knowledge value will be used to divide the data and a random forest is built by combining multiple decision trees. SourceTree [9] and Scikit-learn [10] are tools implementing this random forest classifier for predicting football match results.

2.3 Multilayer Perceptron (MLP)

Multilayer perceptron is one of the supervised learning, which is a type of artificial neural network in the field of deep learning technique. The neural network structure consists of one input layer getting input data, one or more hidden layers with different calculation functions, and one output layer displaying the answer. The sending back values are used to adjust the weight values in the neural network, called backpropagation method. The signal data will be forwarded from one layer to another layer, and a backward pass will adjust connected weight values between two layers according to the error correction criteria. The criteria are defined by the difference between the actual response value and the target response value. Deep learning for classifying football match results in this research using MLP is deployed by the library of MLP in Scikit-learn [10].

2.4 Correlation Coefficient

The correlation coefficient is an indicator of the relationship of data with values in the range +1.0 to - 1.0. If the value is close to +1.0 or -1.0, it means one variable are highly correlated to another variable. On the other hand, if the value near 0. It means that both variables are not correlated. To find the relationship between competition results and related information, the correlation analysis by RapidMiner [11] is applied to identify factors for guiding football team improvement.

3. THE PROPOSED METHOD

Our proposed method uses the historical football match results for previous seasons, which are useful information, to predict the current competition results. Two selected machine learning techniques (the random forest classifier vs. the multilayer perceptron) have also been implemented with two different sets of data features (the information of three previous league seasons vs. the information of five previous league seasons). The results found that the three previous league season is more suitable than the five previous seasons, and the performance of the random forest classifier is better than the multilayer perceptron.

3.1 Data Collection

Information about previous matches of the football teams from www.premierleague.com [8] were collected to predict football competition results. This information consists of the following manner.

- The football team names of all 20 teams were collected as home team and away team.
- The names of the football referees were kept as referee.
- All previous encounter competition data on both teams were collected as the number of times that each team loses, wins or draws by dividing into five data features, including the number of home team wins, visiting team wins, and draws from all matches, the number of home team wins as the home team, the number of the away team wins as the away team.
- The competition results of the last five matches of the home team and away team with other teams were saved as the result of each team's competition.

Referee	HTW	ATW	D	HW	AW	PH1	PH2	PH3	PH4	PH5	PA1	PA2	PA3	PA4	PA5	HomeTean	AwayTeam
J Moss	13	2	3	7	1	W	W	W	W	W	D	D	L	W	W	Arsenal	Crystal Palace
M Jones	4	7	13	3	3	W	W	W	W	D	W	L	L	W	L	Leicester	Everton
M Dean	8	3	2	4	1	D	W	L	W	L	W	L	W	W	L	Man Unite	Swansea
C Pawson	0	2	0	0	1	W	D	W	L	W	L	L	L	D	L	QPR	Hull
A Taylor	7	3	6	4	2	W	W	L	D	W	L	L	W	L	D	Stoke	Aston Villa
N Swarbric	10	3	7	7	1	L	L	L	W	L	L	W	W	W	W	West Brom	Sunderland
C Foy	14	21	9	10	9	L	W	L	L	L	W	L	W	W	D	West Ham	Tottenham
M Clattenb	18	10	10	10	3	W	D	L	W	W	D	W	W	D	L	Liverpool	Southampton
M Atkinsor	7	22	7	6	9	L	W	L	L	L	W	W	W	W	W	Newcastle	Man City
M Oliver	1	4	2	0	2	D	W	W	W	L	W	D	W	L	W	Burnley	Chelsea
M Dean	8	22	14	6	8	W	L	L	W	L	L	L	W	L	L	Aston Villa	Newcastle

Figure 1. Examples of data for predicting the competition results

Hometeam	Awayteam	Hbp	Abp	Htotp	Atotp	Hpsucc	Apsucc	Hf	Af	Hc	Ac	Ho	Ao
Arsenal	Leicester	70	30	632	263	85	63	9	12	9	4	5	3
Watford	Liverpool	45.6	54.4	395	477	70	73	14	8	3	3	3	1
West Brom	Bournemo	28.7	71.3	242	612	64	86	15	3	8	2	2	0
Southampt	Swansea	59.6	40.4	518	365	83	77	10	13	13	0	0	1
Everton	Stoke	61.6	38.4	497	292	78	72	13	10	6	7	2	6
Crystal Pal	Huddersfie	56.7	43.3	391	306	77	65	7	19	12	9	0	2
Chelsea	Burnley	61.9	38.1	521	320	85	75	16	11	8	5	2	1
Brighton	Man City	21.8	78.2	213	768	61	89	6	9	3	10	6	1
Newcastle	Tottenham	26.9	73.1	245	702	64	89	6	10	5	7	2	1
Man utd	West Ham	55.4	44.6	493	397	84	79	19	7	11	1	1	4
Swansea	Man utd	41	59	406	601	77	87	12	11	3	5	1	0

Figure 2. Examples of data for guiding the football team improvement

Therefore, there will be a total of 18 attributes to identify the football match results, including the home team name, the away team name, the referee's name, the number of home team wins, the number of away team wins, the number of draws, the number of home team wins as the home team, the number of away team wins as the away team, and the last 5 match results of the home team and away team with others (examples of attributes shown in Figure 1). In addition, there is a collection of related information used to analyze the factors affecting competition results (examples of these information revealed in Figure 2). To define the recommendation list for improving the performance of football teams, examples of interesting data from www.premierleague.com [8] were the percentage of ball possession and passing the ball successfully, the number of ball passes, the number of fouls, the number of corner kicks, the number of offsides, the number of total shots, shots on target, and woodwork hits, the number of blocked shots, the number of shooting in/outside the penalty area, the number of hits, the number of yellow cards and red cards.

3.2 The Experiment of Selecting the Machine Learning Techniques

To select the machine learning techniques, the information of three league seasons, i.e. the 2014-2015, 2015-2016, and 2016-2017 seasons are learned and evaluated the prediction results. The total number of matches was 1,140 matches, which consisted of 516 home team wins, 340 away team wins, and 284 draws. The 10-fold cross-validation method with the confusion matrix were executed to measure the efficiency of each classification technique.

Table 1. A confusion matrix

Actual class	Predicted class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

A confusion matrix is shown in Table 1 and all evaluated values for the classification performance of football match results are expressed in Table 2. The experimental results by two machine learning techniques are demonstrated in Table 3 (results of the random forest classifier) and Table 4 (results of the multilayer perceptron). The accuracy value will display the performance of classification, while the precision will present the efficiency of each category classification model. In addition, the recall value will show the efficiency of the classification model that can be fully classified (without losing data).

Table 2. The percentage of performance value

Class	Accuracy	Precision	Recall
Positive	(TP+TN) / (TP+TN+FP+FN) *100	TP / (TP+FP) *100	TP / (TP+FN) *100
Negative		TN / (TN+FN) *100	TN / (TN+FP) *100

Table 3. The performance of the random forest classifier

Class	Accuracy	Precision	Recall
Home Win	64.47%	58.47%	74.22%
Draw	70.96%	34.64%	18.66%
Away Win	70.70%	50.90%	49.70%
Average	68.71%	48%	47.53%

Table 4.The performance of the multi-layer perceptron

Class	Accuracy	Precision	Recall
Home Win	61.22%	54.92%	80.03%
Draw	73.24%	37.03%	10.56%
Away Win	69.21%	48.20%	43.52%
Average	67.89%	46.72%	44.70%

According to Table 3 and Table4, all average performance values of the random forest classifier are higher than those of the multilayer perceptron a little bit. Therefore, the proposed method has applied the random forest technique to identify football competition results.

3.3 Overview of Prediction

The overview of prediction system is separated into 3 parts as displayed in Figure 3: the preprocessing of collected information, the prediction of football competition results, the guidance of football team improvement.

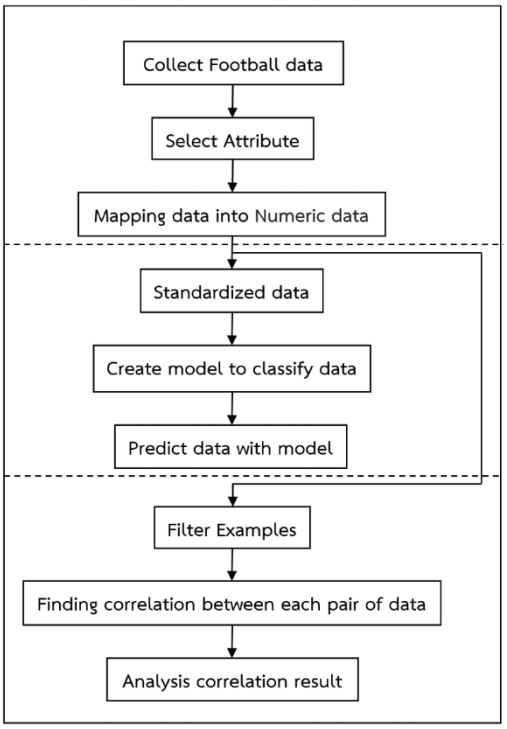

Figure 3. Overview of prediction system

3.3.1 The preprocessing of collected information

In the preprocessing, many experiments on different sets of data features were tried to classify the football match results. The best data set consisted of 18 features described in section 3.1 data collection. Then, all collected data for predicting competition results were converted into numbers, i.e. the names of the home team and the away team to number 1-20, the referee names to different individual numbers, and the football match results (win to 1, draw to 2, lose to 3). Examples of converted data are presented in Figure 4.

HomeTeam	AwayTeam	Referee	HTW	ATW	D	HW	AW	PH1	PH2	PH3	PH4	PH5	PA1	PA2	PA3	PA4	PA5
1	6	1	13	2	3	7	1	1	1	1	1	1	2	2	3	1	1
9	7	14	4	7	13	3	3	1	1	1	1	2	1	3	3	1	3
12	16	4	8	3	2	4	1	2	1	3	1	3	1	3	1	1	3
22	23	6	0	2	0	0	1	1	2	1	3	1	3	3	3	2	3
15	21	5	7	3	6	4	2	1	1	3	2	1	3	3	1	3	2
19	24	7	10	3	7	7	1	3	3	3	1	3	3	1	1	1	1
20	17	22	14	21	9	10	9	3	1	3	3	3	1	3	1	1	2
10	14	19	18	10	10	10	3	1	2	3	1	1	2	1	1	2	3
13	11	3	7	22	7	6	9	3	1	3	3	3	1	1	1	1	1
4	5	0	1	4	2	0	2	2	1	1	1	3	1	2	1	3	1
21	13	4	8	22	14	6	8	1	3	3	1	3	3	3	1	3	3

Figure 4. Examples of converted data into numbers

3.3.2 The prediction of football competition results

Before inputting data sets to train, create the classification model, and identify the competition results, all data numbers were transformed by standardization, which believed that this can improve the machine learning efficiency. Examples of transformed data are expressed in Figure 5.

```
array([[-1.77252329e+00, -6.11798627e-01, -6.23997297e-01, ...,
         1.15302005e+00, -1.16049472e+00, -1.11508523e+00],
       [-1.48234213e+00, -3.21617460e-01, -1.29858897e+00, ...,
        -1.15099898e+00, -1.16049472e+00,  3.01374387e-02],
       [-1.19216096e+00, -1.33725154e+00, -2.86701461e-01, ...,
         1.01053466e-03, -1.16049472e+00,  1.17536011e+00],
       ...,
       [ 9.84197791e-01,  8.39107208e-01,  1.06248188e+00, ...,
         1.15302005e+00,  1.16253246e+00,  3.01374387e-02],
       [-1.77252329e+00, -1.19216096e+00, -4.55349379e-01, ...,
         1.01053466e-03, -1.16049472e+00, -1.11508523e+00],
       [ 5.48926041e-01,  9.84197791e-01, -6.23997297e-01, ...,
         1.15302005e+00, -1.16049472e+00,  3.01374387e-02]])
```

Figure 5. Examples of transformed data

All input features were trained to create the classification model for the football competition results using the random forest method. The output classification model will be used to predict the competition results in advance. The experimental results will be described in section 4. The experimental results.

3.3.3 The guidance of football team improvement

All collected data for analyzing the factors affecting football competition results were calculated by the correlation analysis. The relevant features, which have a high positive correlation with the full-time result feature (FTR), are focused as the valuable information to generate recommendation list for football team improvement. However, only the home team features will be analyzed to create some advice for the home team, and only the away team information for the recommendation of the away team.

4. THE EXPERIMENTAL RESULTS

4.1 Experimental Data

To evaluate the performance of the proposed method, the football match results of the three league seasons: the 2014-2015, 2015-2016, and 2016-2017 seasons (1,140 matches) have been the training data, and the 2017-2018 season (only 220 matches) has been the unseen test data. The total of 220 matches has been divided into the home teams win 95 matches, both teams draw 59 matches, and the away teams win 66 matches.

4.2 Experimental Results of Prediction

The classification models of competition results have been generated by the random forests technique to identify football match results into home win, draw, or away win. The results of classifying the match results by the proposed method are represented by three confusion matrixes in Table 5, Table 6, and Table 7, respectively. The overall performance of match result prediction is also shown in Table 8.

Table 5. A confusion matrix of the home team win prediction

Actual class	Predicted class	
	Home Win	Others
Home Win	84	11
Others	35	90

Table 6. A confusion matrix of the draw prediction

Actual class	Predicted class	
	Draw	Others
Draw	24	35
Others	5	156

Table 7. A confusion matrix of the away team win prediction

Actual class	Predicted class	
	Away Win	Others
Away Win	46	20
Others	26	128

Table 8. The performance of match result prediction

Class	Accuracy	Precision	Recall
Home Win	79.09%	70.58%	88.42%
Draw	81.81%	82.75%	40.67%
Away Win	79.09%	63.88%	69.69%
Average	80.00%	72.40%	66.26%

According to Table 5, Table 6 and Table 7, there are 84 of 95 matches correctly predicted as the home team wins, and 24 of 59 matches accurately identified as the draws, including 46 of 66 matches properly defined as the away team wins.

Referring to Table 8, the accuracy rates of all competition results (home win, draw, away win) are higher than 79%, and the precision rates of them are about 60-80%. However, the recall rate of the draws is about 40%, while those of the home team wins and the away team wins are about 88% and 70%, respectively. One reason of that may be, there are the draws less than the home team wins and away team wins in the training data. Although the average recall value is about 66%, the average accuracy and precision are higher than 72%, especially 80% average accuracy rate. The overall performance of match result prediction is considered satisfactory. Therefore, the proposed classification model of the football match results is effective and useful for predicting the competition results.

4.3 Recommendation List for Improving the Team Performance

The guidance of football team improvement has been generated from the variables highly related to the football match results. The correlation coefficients are as indicators of data relationship between the match results and other features. The different football teams have been affected by the different variables depending on the correlation between the full-time result and each other feature. The winning team and the losing team have also got the alteration of the recommendation list. The examples of recommendation list are “Increase ball possession / Increase ball possession training”, “Increase pass attempted / Increase pass training”, “Decrease fouls”, “Increase corner attempted / Increase corner training”, “Avoid offside”, “Increase shot attempted Increase shoot training”, “Avoid shoot off target / Increase shoot training”, “Avoid shot woodwork / Increase shoot training”, “Avoid block / Increase shoot training”, “Increase shot inside box attempted / Increase shoot inside box training”, “Increase shot outside box attempted / Increase shoot outside box training”, “Increase tackle attempted / Increase tackle training”, “Decrease yellow card”, and “Decrease red card”.

5. CONCLUSION

The research proposed the prediction of football match results using the random forests. There are 3 main parts of processing method that are preprocessing collected data, predicting the match results by the machine learning technique, and guiding the recommendation list using the correlation analysis. The experimental results found that the efficiency of football competition result prediction is good in the accuracy and the precision rates, while the recall rates are acceptable. The classification model for the football match prediction is expected to perform better when there are three types of match results (the home team win, the away team win, and the draw) in the same ratio for running the training data. Another research result is shown that some suggestions about improving the football team performance, which made from calculating the correlation coefficients, are beneficial to increase the football team efficiency.

6. REFERENCES

[1] Snyder, J. A.L. 2013. *What actually wins soccer matches: Prediction of the 2011-2012 Premier League for fun and profit*. Thesis, University of Washington, WA: Department of Computer Science.

[2] Shin, J. and Gasparyan, R. 2014. *A novel way to soccer match prediction*. Stanford University, Department of Computer Science.

[3] Prasetio , D. and Harlili D. 2016. Predicting football match results with logistic regression. In *Proceedings of the International Conference on Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 1-5.

[4] Hucaljuk, J., Rakipovi, A. 2011. Predicting football scores using machine learning techniques, In *Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1623 – 1627.

[5] Dobravec, S. 2015. Predicting sports results using latent features: A case study. In *Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1267 – 1272.

[6] Igiri, Peace, C., Nwachukwu, Okechukwu, E. 2014. An Improved Prediction System for Football a Match Result. *IOSR Journal of Engineering (IOSRJEN)*. 4, 12 (Dec. 2014), 12-20.

[7] Fernández, J., Medina, D., Gómez, A.; Arias, M., Gavaldá, R. 2016. In *Proceedings of16th International Conference on Data Mining Workshops (ICDMW)*, 136-143.

[8] Premierleague.com. (2019). Premier League Football News, Fixtures, Scores & Results. https://www.premierleague.com.

[9] SourceTree. 2019. SourceTree | Free Git GUI for Mac and Windows. https://www.sourcetreeapp. com/.

[10] Scikit-learn. 2019. Scikit-learn: machine learning in Python. https://github.com/scikit-learn/scikit-learn.

[11] RapidMiner. 2019. RapidMiner Studio. https://rapidminer.com/.