ISSN 2618-1576 Vol. 3, Issue 3: 74-78, DOI: 10.25236/FSR.2021.030314

Predict soccer match outcome based on player performance

Wei Yang[1,a,*]

[1] *School of Business Administration, Xi'an Eurasia University, Xi'an, China*
[a] *yangwei@eurasia.edu*

Abstract: *In the field of sports, due to the unpredictability of soccer match outcome and the existence of sports betting industry, the prediction of soccer match results has been highly concerned by the news media, fans, sports experts and other stakeholders, and it is also a hot topic in academic circles. The strength of the team is made up of players. Player performance is a key factor in winning a soccer match. This article builds a model to predict the outcome of the match based on the performance of the players. In experiments,* the statistics of *the AUC, F1 and prediction accuracy of the model were 0.8597, 0.6973 and 0.7965* respectively *on the verification data. The experimental results also show the feasibility of this method.*

Keywords: *player performance, match outcome prediction, LSVC classifier*

1. Introduction

In the field of sports, due to the unpredictability of soccer match outcome and the existence of sports betting industry, the prediction of soccer match results has been highly concerned by the news media, fans, sports experts and other stakeholders, and it is also a hot topic in academic circles. Because there are certain luck factors in soccer matches, and referees often have an impact on the outcome of the match, so the score may not be a real reflection of the strength of the team. There are some shortcomings in using the score of historical match results to predict the upcoming match. A better way is therefore needed to evaluate the true strength of the team and to predict the outcome of the match based on an analysis of the team's strength.

The strength of the team is made up of players. Players are a key factor in winning a football match. So, it is more accurate to predict the outcome of a match by the performance of a player than by historical results and the relationship between a team's victory and defeat. This article builds a model based on the performance of the players to predict the outcome of the match.

2. Methods

2.1. Model prediction framework

Figure 1 shows the process of using player performance data to predict soccer match outcome. The framework realizes the prediction of soccer match results through data standardization and processing, model training and optimization, prediction and other steps.

ISSN 2618-1576 Vol. 3, Issue 3: 74-78, DOI: 10.25236/FSR.2021.030314

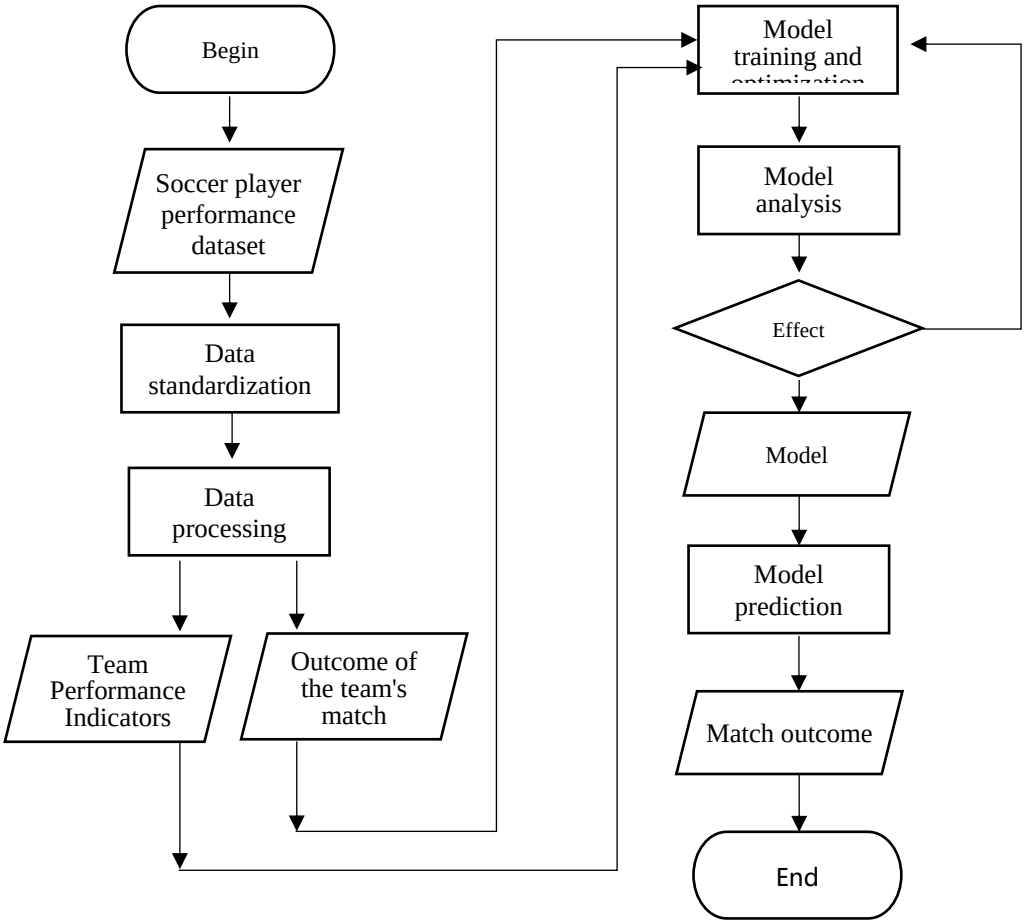

Figure 1: Model prediction framework

2.2. Soccer player performance dataset

Each row in this dataset corresponds to the performance of one player at one match. Each player's performance is described by simple statistics (e.g. goals, assists, passes, etc.) and metrics derived from network theory (e.g. betweenness centrality) that are taken by understanding the match from a network perspective. All of these were computed out of event data (around 2 million events=rows) of the 2017-18 Seasons of German Bundesliga, the World Cup 2018 in Russia and Euro 2016 in France. Table 1 shows the number of players, the number of teams and the number of matches in different competitions.

Table 1: number of players, teams and matches of performance dataset

Competition	Numbers		
	match	team	player
Bundesliga 2017-18	300	18	441
Euro 2016	49	24	390
World Cup 2018	63	32	527

Table 2 classifies all 29 variables and explains the types of variables.

Table 2: Variables of performance dataset

Category	Expression	variables
laction_related	category	pos, is_home_team
attack- related	number	touches, assists, interceptions, tackles, countattack
foul-related	number	fouls, wasfouled,ycards,rcards,offsides
shot-related	number	stop_shots, shots_ontarget , shots_offtarget, shotsblocked
dribble-related	number	dribbled_past, drib_success, drib_unsuccess
pass-related	number	keypasses,passes_acc,passes_inacc,pass_lost,
cross-related	number	crosses_acc,crosses_inacc
network--related	float	degree_centrality,betweenness_centrality,closeness_centrality, flow_centrality

ISSN 2618-1576 Vol. 3, Issue 3: 74-78, DOI: 10.25236/FSR.2021.030314

2.3. Data standardization and processing

2.3.1. Data standardization

Restricting each variable within a certain range through data standardization helps to eliminate the influence of singular samples and at the same time helps to reduce the time for subsequent model training.

First, the two categorical variables, the player's position(pos) in the match and the home and away(is_home_team) variables, are dumb-coded. Then, because the ranges of different variables are different, they need to be standardized to make their change ranges consistent. The formula is below.

$$x_j' = \frac{x_j' - \overline{x_j'}}{\sigma_j},$$

where σ_j is the standard deviation of the variable, $\overline{x_j'}$ is the mean value of this variable.

2.3.2. Data processing

The outcome of the match is processed as 0 or 1, where 1 means the match is won, and 0 means the match is tied or lost. The processed match outcome variable will be used as the output of model training, denoted as vector O.

Because the outcome of the soccer match is the most important feature to measure the performance of all players in the team after a match. Therefore, here we calculate team performance indicators as vector $X^{C,M,T}(\overline{x_1}, \overline{x_2}, \dots)$, which means performance of team T in a certain match M of a certain competition C.

$$X^{C,M,T}(\overline{x_1}, \overline{x_2}, \dots) = \frac{1}{n}\sum_{P=1}^{n} X^{C,M,T,P}(x_1, x_2, \dots)$$

Where n is the number of players playing in the match, the vector $X^{C,M,T,P}(x_1, x_2, \dots)$ represents the performance of player P in the team T after a match M in a competition C.

2.4. Model training and optimization

Various indicators of team performance will affect the outcome of the match. We take the rating vector $O^{C,M,T}$as input, and the match outcome vector $O^{C,M,T}$ as output. The classification model of machine learning is used for training.

Here, a linear support vector classifier (LSVC) is used to do the two-class training of the match outcome. LSVC is used to solve unconstrained optimization problems. The loss function here is represented by $\xi(W, X^{C,M,T}, O^{C,M,T})$, and the objective function is:

$$\min_W = \frac{1}{2}W^T W + C\sum_{T=1}^{l} \xi(W, X^{C,M,T}, O^{C,M,T})$$

Where l is the number of teams.

Through the method of cross-validation, the effect of the model is evaluated. If the model effect is not good, optimize the model parameters to improve its effect, and then get the final model.

Figure 2 shows the ROC of each fold and the mean ROC. The statistics of the AUC, F1 and prediction accuracy were 0.8597, 0.6973 and 0.7965 respectively after training and validating the LSVC model, which were higher than the predictive result of chance (AUC = 0.5).

ISSN 2618-1576 Vol. 3, Issue 3: 74-78, DOI: 10.25236/FSR.2021.030314

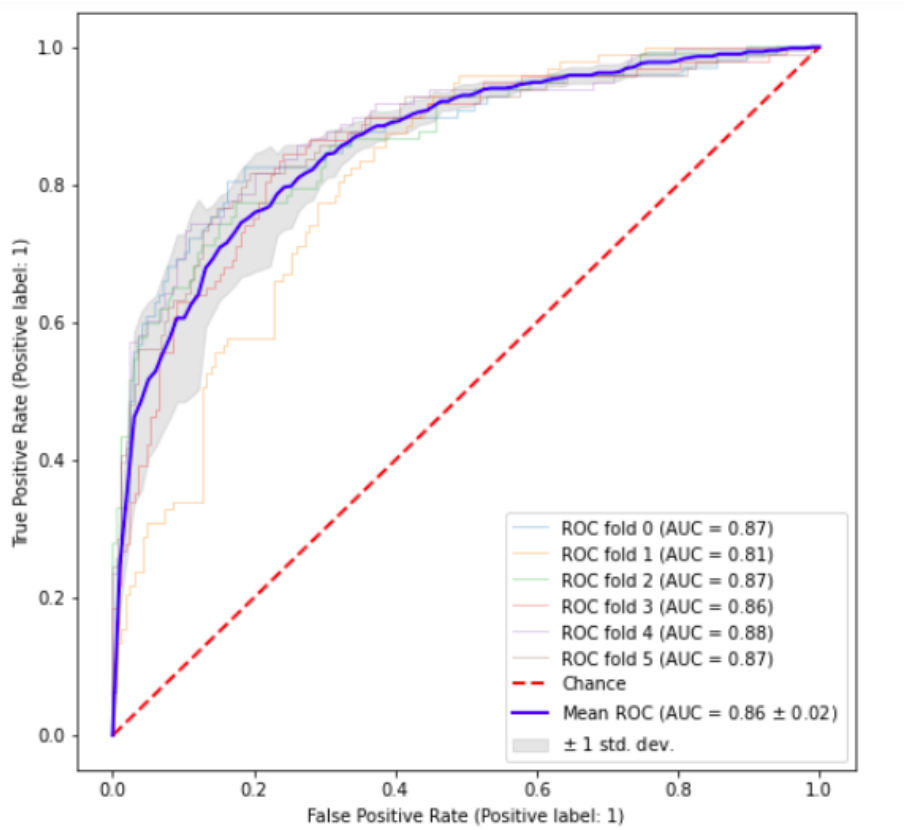

Figure 2: ROC of each fold and the mean ROC

2.5. Match outcome prediction

Use the model to predict the outcome of matches and compare it with the actual outcome to get the confusion matrix in Table 3. It can be seen from Table 2 that the model's TPR is 62.3% and FPR is 8.35%.

Table 3: Confusion matrix

		predicted outcome	
		win	tied or lost
actual outcome	win	364	220
	tied or lost	83	911

The model can predict that the match will not win more accurately, and the accuracy of the prediction of the match win needs to be further improved.

3. Conclusions

The experimental results also show the feasibility of this method. The results of this research can be used not only for soccer matches, but also for basketball, volleyball and other matches with a slight improvement.

Acknowledgements

Routine project of Shaanxi Provincial Sports Bureau: Research on the prediction model of football match outcome based on player rating.

References

[1] Engist, O.; Merkus, E.; Schafmeister, F. The Effect of Seeding on Tournament Outcomes: Evidence from a Regression-Discontinuity Design. J. Sports Econ. 2021, 22, 115–136.

ISSN 2618-1576 Vol. 3, Issue 3: 74-78, DOI: 10.25236/FSR.2021.030314

[2] Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Number 2; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
[3] Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2020, 2, 56–67.
[4] Musa, R.M.; Majeed, A.A.; Taha, Z.; Abdullah, M.; Maliki, A.H.M.; Kosni, N.A. The application of Artificial Neural Network and k-Nearest Neighbour classification models in the scouting of high-performance archers from a selected fitness and motor skill performance parameters. Sci. Sport 2019, 34, e241–e249.
[5] Huifeng, W.; Kadry, S.N.; Raj, E.D. Continuous health monitoring of sportsperson using IoT devices based wearable technology. Comput. Commun. 2020, 160, 588–595.
[6] Zahran, L.; El-Beltagy, M.; Saleh, M. A Conceptual Framework for the Generation of Adaptive Training Plans in Sports Coaching. In International Conference on Advanced Intelligent Systems and Informatics; Springer: Berlin/Heidelberg, Germany, 2019; pp. 673–684.
[7] Choraś, M.; Pawlicki, M. Intrusion Detection Approach based on Optimised Artificial Neural Network. Neurocomputing 2020, in press. Horvat, T.; Job, J. The use of machine learning in sport outcome prediction: A review. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2020, 10, e1380.
[8] Behravan, I.; Razavi, S.M. A novel machine learning method for estimating football players' value in the transfer market. Soft Comput. 2020, 1–13.