Explore Machine Learning's Prediction of Football Games

Bomao Pan

Chongqing Bashu Ivy School, Chongqing, 400000, China

Abstract. The aim of this study is to predict the outcome and score of football matches. To achieve this goal, this paper employs a variety of machine learning models, including Random Forest, support vector classifiers (SVC), and Logistic Regression, and conducts in-depth analysis of the data. The results show that home teams have a significantly higher win rate than away teams. In addition, the score changes show a high degree of randomness, reflecting that the game is affected by a variety of factors. The prediction performance of these models is different, and the prediction accuracy of the random forest model is better than the other two models. Through the prediction of the winning rate, this paper aims to provide more scientific reference for the majority of fans and deepen the understanding of the strength of each team and the influence of external factors on the result of the game. This study not only helps to improve the analysis ability of football matches, but also provides a theoretical basis for the optimization of game strategies.

1 Introduction

E-sports uses electronic equipment to complete mechanical movements and human brain competition. E-sports has risen rapidly in recent years and has become an important part of the global entertainment industry. According to the latest statistics, the market size of e-sports has exceeded billions of dollars, the number of viewers is growing at a rate of tens of millions every year, and the scale and influence of professional events are becoming increasingly significant.

E-sports can train some abilities in people's lives, such as communication skills, calculation skills, and reaction skills. Finally, it can help people improve their ability to cooperate with the team. Some e-sports are still helpful for players' physical strength. To give an example, in the first half of 2024, the number of e-sports users will be approximately 490 million, and the operating income will also be very considerable. E-sports products are steadily increasing and the types of competitions are becoming more abundant. For my own favorite e-sports competition, the number of online viewers reached 1.11 million, and the total number of viewers reached 35.93 million. For example, some competitions require players to compete for a long time and play some games. Like Honor of Kings, this game requires 5 players to play together. It is a great test of cooperation and tacit understanding

Corresponding author:gary.panbomao@biacademy.cn

between players. Sometimes the other four people may not be able to play normally due to one person's mistake. The complexity of this area cannot be underestimated, with match outcomes often influenced by a variety of factors, including individual player skills, team tactics and real-time decision-making during the match, which together shape the dynamics of the match. and unpredictability. For the development of e-sports, people mainly need to pay attention to winning and losing, so the importance of predicting winning and losing at this time is very important.

However, there are still many challenges in predicting the outcome of e-sports. Traditional methods often rely on subjective judgment and lack scientific data support, resulting in uncertainty and insufficient accuracy of prediction results. In addition, existing analysis tools often have difficulty processing complex game data and cannot fully reflect various influencing factors. Therefore, a new methodology is urgently needed to improve the accuracy and reliability of winning and losing predictions and to help players, coaches, and analysts better understand and grasp the dynamic changes in the game.

Against this background, the rise of machine learning technology provides new possibilities for e-sports victory and defeat prediction. As a powerful data analysis and pattern recognition tool, machine learning has made remarkable achievements in many fields in recent years, including finance, medical care, and traditional sports. In these fields, machine learning not only improves the efficiency of data analysis, but also greatly improves the accuracy of predictions. This shows that machine learning has broad prospects for application in e-sports. Through in-depth analysis of massive game data, machine learning can help us identify key factors and reveal potential patterns, thereby providing more scientific and reliable support for e-sports victory and defeat predictions.

The theme of this article is to explore machine learning technology to predict the outcome of e-sports competitions to improve the accuracy and reliability of predictions. Data collection: First, this article will identify and discuss some inevitable factors, such as players' mental state, the environment and technical issues on the day of the game, etc., which may have a significant impact on the game. Secondly, machine learning models suitable for prediction of e-sports wins and losses will be found and comparisons between different models will be made to determine which model can more effectively capture the patterns in the game data. Finally, this article provides effective strategies and suggestions for e-sports victory and defeat prediction, providing a reference for teams, coaches and analysts, and promoting the further development of the e-sports field.

2 Data and Methods

2.1 Method

The data presented in this paper are obtained from the kaggle. The data included the following variables: home goal away goal home corner away corner home attack home shots, ht-diff at-diff ht-result at-result total -corner.

There are some common classified models: logistic regression, Random Forest, decision tree, SVM, XGBoost**,** LightGBM, and neural network. These methods all can be used in predicting the consequence of competition, and each of these has different advantages and disadvantages.Common models of machine learning algorithms include XGBoost, decision trees, random forests, neural networks, logistic regression, and LightGBM.

XGBoost is a highly accurate algorithm that can effectively handle missing values (such as 0 values and null values in the database), is suitable for large-scale data processing, and can accelerate the training process. It is very flexible, supports multiple loss functions, and can handle regression problems. However, the disadvantage of XGBoost is that the training

process is restrictive and cannot be trained multiple times at the same time. You must wait until the results of one training are out before starting the next training.

The performance of neural networks is highly dependent on the amount of data. As the data increases, the performance of neural networks will gradually improve. It can process large amounts of data and perform parallel calculations to achieve synchronous training. However, the disadvantage of neural networks is that they require a large amount of data to improve the calculation accuracy, and their computational complexity is high, and the training time is long.

Logistic regression is a simple and efficient algorithm with low computational complexity, good interpretability, and a wide range of applications (it can also be used for multi-classification problems). However, logistic regression is sensitive to outliers, requires careful inspection before training data, and performs poorly when dealing with complex relationships. Although it can handle binary and multi-classification problems, logistic regression often has difficulty in effectively classifying when faced with nonlinear relationships.

Random forest has high accuracy and fast processing speed, which can improve training efficiency. As a random computing model, it is very suitable for our needs. However, the disadvantage of random forest is that multiple similar decision trees may obscure the true situation.

Nowadays, the most of predictions for football competitions still come from the predictor's analysis of their experience. Predictors usually focus on the performance and ability of each player and consider the strategy between each team. These data are too large for both collector and computer, even some performance cannot be presented by simple data, as a result, this paper just focused on the total data of each competition. Some predictions from computers use neural networks, random forests, and SVM. Because of the lack of computing power, this paper didn't consider neural networks.

2.2 Data processing process

Because there are only two competitions between two specific teams, the number of directive data is small. Therefore, any noise will affect the consequence notably. Instead of filling out the blank with the mean value, the better choice is deleting the range which includes blank data.

What's more the dataset includes every level tournament of England football. However, most people just focus on the Premier League, so this paper just keeps the data in the Premier League and delete the useless data from another tournament.

Another decision tree also has a problem. The problem is that it is possible to overfit, especially in the condition, which does not provide a large amount of data.

The lack of ability to solve muti-classify problems lets this paper does not use the SVM method.

Because of the low requestions for computing power and the fast speed, this paper adds logistic regression to our model. However, its performance in the nonlinear prediction is not very well.

To avoid the problem, this paper add a random forest into our model. These two models analyze data together.

The dataset, people can use NumPy to draw some pictures of every kind of data, such as the ratio of the score of both the home team and away team [1-3]; show spare of the number of shooting and attack, which can let us get a generous comprehension[4,5].

These data visualizations not only let us know the general range of the data but also directly show the advantage of the home team. This means this paper should not just consider the two teams but also consider the identity of the team. Then, the amount of perfect data

between each team was reduced to 10, because each team only meets two times each year, and the identity of the team would change [6,7].

3 Result analysis

As the paper mentions in the background, the analysis for the football game cannot avoid the analysis from humans, the analysis was separated into two steps, the first one is by computer-using training data and testing data. The second step is an analysis by the previous experience and analysis method- to compare the difference between football lovers and computers.

This paper decided that 20 presents of data to be the testing data, and the rest of these be used to train the model. The accuracy of each model is nearly 50%. Although it seems low, this paper believes it is a good program. The reason is there are only three consequences, and none of the probability is low, which means that the accuracy immutably be high.

What's more, if the program predicts the probability of win is 70%, and the consequence is win too. The accuracy of this prediction will be 70% too, although it presents an excellent prediction.

The second part is human evaluation. To ensure that the prediction is useful, a simple way is to consider the model can get the right prediction between the team which exits a gap in strength.

This paper chosed the "big six" in the premier league, and they are the top-six classically powerful teams (Man City, Man United, Chelsea, Arsenal, Liverpool, Tottenham), and analyzed the prediction between these and some weaker teams [8-10]. The consequence shows that all of these comply with human experience.

However, because of the lack of timeliness of the model, it cannot match the present change of power of each team. For example, a team named Brighton has become powerful in recent years, and Chelsea always lost in this year since the change of coaches. This is a fact that every football lover knows, but the program cannot match the change and makes some invalid predictions. The specific results are shown in the Fig.below。

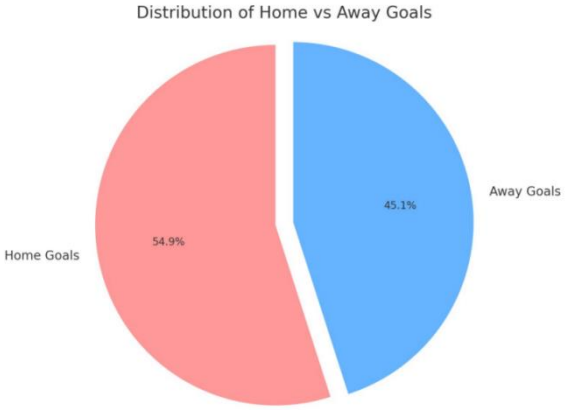

Fig.1 Distribution of Home vs Away Goals (Photo/Picture credit : Original)

From Fig.1, the difference between home goal and away goal, and the winning rate of home goals is higher than that of away goals.

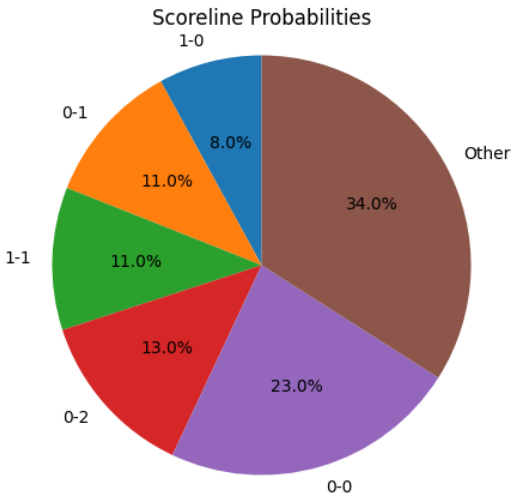

Fig. 2 Scoreline Probabilities (Photo/Picture credit : Original)

Fig.2 shows the odds of the score, and you can see more other and 0-0.

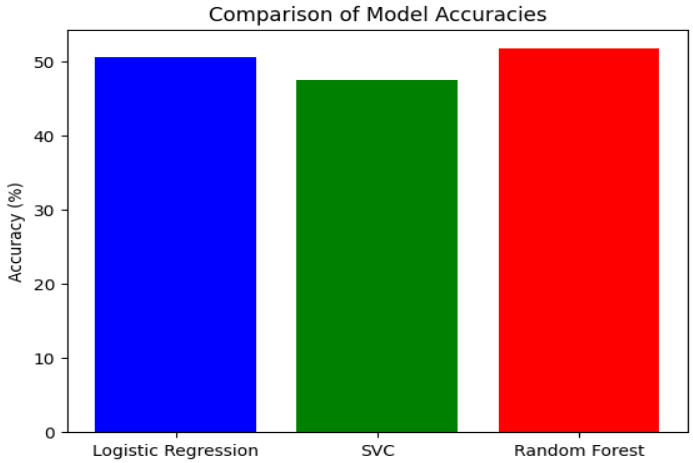

Fig. 3 Comparison of Model Accuracies (Photo/Picture credit : Original)

This paper uses three models to evaluate the win rate. As can be seen from Fig.3, the accuracy of logistic regression, SVC, random forest3 models, random forest greater than logistic regression greater than SVC. So it can be seen that the prediction result of random forest is better than the other two.

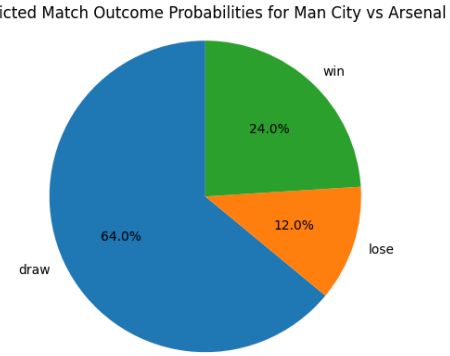

Fig.4 Predicted Match Outcome Probabilities for Man City vs Arsenal (Photo/Picture credit : Original)

Fig.4 shows the predictions from the man city and arsenal game.man city The odds of winning are 24 percent, 12 percent losing and 64 percent drawing.

By training models using logistic regression and random forests, this paper can effectively predict the outcome of each match between two participating teams. The results are presented in three graphs: one graph showing the accuracy of each model [5]; one graph showing the probability of winning, drawing, and losing [6]; and the last graph showing the probability of each score. Since the number of score combinations is theoretically infinite, this paper grouped scores with a probability less than or equal to 3% [7]. Other analyses, such as the number of shots and corners, are not only less meaningful but also difficult to predict, so other types of outputs were not considered.

4 Conclusion

The aim of this study is to predict the outcome and score of football matches. To achieve this goal, this paper employs a variety of machine learning models, including Random Forest, support vector classifiers (SVC), and Logistic Regression, and conducts in-depth analysis of the data. The results show that home teams have a significantly higher win rate than away teams. In addition, the score changes show a high degree of randomness, reflecting that the game is affected by a variety of factors. The prediction performance of these models is different, and the prediction accuracy of the random forest model is better than the other two models. Through the prediction of the winning rate, this paper aims to provide more scientific reference for the majority of fans, and deepen the understanding of the strength of each team and the influence of external factors on the result of the game. This study not only helps to improve the analysis ability of football matches, but also provides a theoretical basis for the optimization of game strategies. In this study, the field of football match prediction has been deeply explored, and after rigorous thinking and analysis, a suitable model has been selected for application. In this paper, the data are cleaned effectively, and the features of the original data are displayed by visual means. In addition, the corresponding program is written and a variety of evaluation methods are applied. The results show that the model can make effective prediction in most cases.

However, the limitation of the model is that it cannot reflect the dynamic changes of each team's strength in a timely manner. Establishing a relationship between match results and recent performance may help mitigate this problem. Although this direction has not been explored in depth in this paper due to feasibility considerations and training time constraints, this paper believe that its potential improvements will help improve the performance of the model.

References

1. W. Zhang and G. Huang, Classification and prediction of 2006 World Cup football matches using graph recognition and neural networks, Ph.D. thesis, (2006).
2. Kang et al., MOBA game situation trend prediction model based on sequence-to-sequence structure, (2023).
3. Z. Jiang, Z. Huang, and F. Wu, The influence of technical and tactical performance of teams in the Super League on the outcome of the game under multiple game scenarios, J. Phys. Educ. / Tiyu Xuekan 25, 2 (2018).
4. L. Jin and X. Pan, Data-driven prediction of e-sports game results and comparative study of methods, Coll. Econ. 22, 154-155 (2020).
5. H. Jiang, Y. Yu, and W. Long, Study on the analysis of influencing factors of football games based on fuzzy grey correlation analysis, Comput. Digit. Eng. 51, 3 : 555-560 (2023).
6. Z. Guo, Research on football attack and defense data and game wins and losses - Taking the 2017-2019 European Champions League as an example, M.S. thesis, National Taiwan Normal University, Taiwan (2021).
7. Y. Shen et al., AI in game intelligence - from multi-role game to game, J. Intell. Sci. Technol. 2, 3: 205-213 (2020).
8. Y. Sun et al., A review of intelligent games: the inspiration of game AI for combat simulation, J. Intell. Sci. Technol. 4, 2: 157-173 (2022).
9. Z. Chen et al., Performance analysis of the 2019 Asian Cup football qualifying tournament - taking the Chinese men's team as an example, Natl. Taiwan Normal Univ. Sports Res. 27: 47-49 (2020).
10. H. Wang and C. Sun, Exploring and predicting players' strategy styles in real-time strategy games, Ph.D. thesis, (2006).