Advancements of Football Data Analysis Based on Machine Learning Algorithms

Qishu Wen[a]
Data Science and Big Data Technology, Guilin Tourism University, Guilin, China

Keywords: Football, Machine Learning, Random Forest.

Abstract: Football is the most influential sport in the world and has a significant impact on the global economy. With the development of big data technology, the application of artificial intelligence in the statistical analysis of football data has become increasingly widespread. This article summarizes relevant research on machine learning in football, with a special focus on prediction methods based on football match data. These methods include hybrid learning models, binary classification and regression, TOPSIS methods, as well as expert systems and ensemble learning. By iterative training on historical match data, an estimate of the ability of each team in the training data can be obtained. A regression-based model is used to calculate the number of goals scored by each team. Sensitivity analysis can be used to assess the impact of different weighting schemes or criteria choices on player rankings. Principal Component Analysis (PCA) can help identify underlying patterns and relationships in player and team performance metrics. Support Vector Machine (SVM) classifiers can effectively learn decision boundaries based on relevant features. These models predict various outcomes by analyzing historical match data and player performance with some degree of success. However, the subjective nature of football matches and refereeing decisions can lead to inaccurate model predictions, emphasizing the importance of explaining model decisions. To improve the applicability of the model, it is suggested that the AI model be applied to different levels of football leagues. Real-time data analytics tools such as Apache Spark can be used to address the challenges of real-time data and processing diversity.

1 INTRODUCTION

There is no doubt that football is the most influential sport in the world, and no matter which country it is played in, there are countless avid fans who follow everything that happens on the field. Football holds immense commercial value in the sports industry, whether it is through sponsoring advertisements on jerseys, broadcasting rights on online platforms, or sports tourism projects. All of these factors have a significant impact on related industries and become a vital driving force for global economic development. In modern football, statistical analysis has become a crucial aspect of the sport. The 22 players on the field generate a large amount of data in 90 minutes. To react promptly to the different situations that arise in the field, it is crucial to analyze the data swiftly and precisely to obtain projected outcomes. This analysis is then shared with the target audience, including coaches, commentators, and spectators. In the age of big data, the use of various football prediction tools has become necessary as data collection and analysis techniques continue to advance.

Artificial Intelligence (AI) has received much attention in recent years, and great progress has been made previously. Currently, AI algorithms are used in various fields, including finance, healthcare, education, transport, etc., and are being used more and more widely in sports. In particular, in football, AI is also involved in many predictions based on match data (Bunker, 2019). The following is a progression of football result predictions based on match data. The authors apply the proposed hybrid machine learning models to the full data of Euro 2004-2016. By iterative training on historical match data before each UEFA EURO, it is possible to obtain an estimate of the ability of each team in each UEFA EURO in the training data. (Groll, 2021). The authors transformed the ternary classification problem into a binary classification problem in order to predict

[a] https://orcid.org/0009-0002-4551-0824

Wen, Q.
Advancements of Football Data Analysis Based on Machine Learning Algorithms.
DOI: 10.5220/0012902400004508
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 67-71
ISBN: 978-989-758-713-9
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

which team will win a match. The number of goals scored by each team would be calculated by the authors using a regression-based model. Random Forest and Gradient Boosting models outperformed the bookies' forecasts by a large margin among all the machine learning models that were tested (Baboota, 2019). The authors utilized the Technique of Ordering Performance with Similarity to Ideal Solutions (TOPSIS) (hwang, 1981) approach to rank players based on a specific set of measures, with the aim of selecting the most suitable players through a single experiment. To ensure systematic ranking of each of the four groups of players, mean and standard deviation were employed. As a result of this process, the ultimate objective of selecting the players was achieved (Qader, 2017). The authors proposed an expert system. The system integrates a number of techniques, such as SVM classifiers, PCA, and nonparametric statistical analysis, to create an ensemble of machine learning techniques. The aim is to predict whether or not the hockey team will win the match (Gu, 2019).

The remainder of this article is as follows. Firstly, providing an overview of relevant research on machine learning in football in Section 2. Then, analyzing the application of artificial intelligence algorithms in processing match data and discussing possible new approaches for the future in Section 3. Lastly, conclusions are presented in Section 4.

2 METHODS

2.1 Hybrid Machine Learning Models for Tournament Prediction

The authors used hybrid machine learning models such as random forests (Groll, 2019), extreme gradient boosting, current ability ranking based on historic matches, Bookmaker consensus model, and Plus-minus player ratings (Hvattum, 2019) to analyze the entire UEFA EURO 2004-2016 data. In addition to standard covariate data, extra ability and rating factors were used to train the cforest model. Subsequently, UEFA EURO 2020 was simulated 100,000 times to ascertain the winning odds for each of the 24 teams who took part. The machine learning model primarily relies on the covariate information of the participating teams, with the ability parameters providing an adequate estimate of the current team strength. An estimate of each participating team's ability for each UEFA EURO can be obtained by training the historical match data iteratively before each tournament. The authors validated the performance of different models in predicting the results. The best model was selected in terms of polynomial likelihood, classification rate, and RPS (Groll, 2021).

2.2 Binary Classification and Regression

To forecast if a football club will win or lose, the authors reduced the complexity of the ternary classification problem to a binary classification problem. The authors tested various machine learning models including Gaussian Naive Bayes (Sudhaman, 2022), SVM, Random Forests and Gradient Boosting (Bent é jac, 2021). The results showed that the Random Forest and Gradient Boosting models performed better than the bookies' predictions. In order to determine the goals scored by each team, the authors intend to employ regression-based models (Liu, 2020). This will enable a more thorough examination and analysis of football games (Baboota, 2019).

In order to determine the best feature and classifier combination for precisely predicting Premier League match results, the authors performed research. The authors experimented using various feature and classifier combinations, including match records from the previous two seasons of the Premier League. Through several experiments, the authors found K-NN algorithm to be the most accurate in predicting matches (Haruna, 2021).

2.3 TOPSIS Approach for Player Ranking

A technique called TOPSIS (Çelikbilek, 2020) is used to rank options based on their proximity to the ideal solution, using a multi-criteria decision-making method. In the context of football player ranking, the authors employ TOPSIS to systematically evaluate and rank players based on specific performance measures. In TOPSIS, the ideal solution represents the best possible performance across all criteria, while the anti-ideal solution represents the worst performance (Qader, 2017). In the context of football player ranking, the ideal solution may correspond to the highest values for goals scored, assists, successful passes, etc., while the anti-ideal solution represents the lowest values for these metrics. In order to make sure that the ranking outcomes are strong and reliable, sensitivity analysis (Li, 2021) was carried out to evaluate the influence of different weighting schemes or criteria choices on the player rankings. Additionally, the accuracy and consistency of the ranking outcomes can be confirmed by using historical match data or expert opinion.

2.4 Expert Systems and Ensemble Learning

The authors advocate for the utilization of expert systems and ensemble learning techniques to enhance the accuracy and reliability of football match outcome predictions. The expert system proposed integrates various machine learning methodologies to form an ensemble, thereby leveraging the strengths of each individual model while mitigating their weaknesses. In the context of football prediction, PCA can help identify the underlying patterns and relationships within player and team performance metrics, enabling more efficient modeling and prediction. Nonparametric statistical analysis techniques, such as kernel density estimation and rank-based tests, are employed to capture complex and nonlinear relationships in football data. SVM classifiers can effectively learn decision boundaries between different classes of match outcomes (e.g., win, lose, draw) based on historical match data and relevant features. The authors employ ensemble learning methods, such as bagging, boosting, and stacking, to aggregate predictions from diverse machine learning models and improve overall performance. By leveraging the collective wisdom of multiple models, ensemble learning (Mienye, 2022) enhances prediction reliability and generalization to unseen data, thereby increasing the effectiveness of football match outcome forecasts. Finally, the expert system proposed by the authors incorporates mechanisms for real-time adaptation and learning, allowing the prediction model to continuously evolve and improve based on incoming match data and performance feedback. By dynamically updating model parameters and adjusting prediction strategies, the system can adapt to changing match conditions, team dynamics, and other relevant factors, thereby maintaining high prediction accuracy over time (Gu, 2019).

3 RESULTS AND DISCUSSIONS

Groll et al. presented the average outcomes of four prediction techniques, namely ranger, cforest, xgboost, and lasso, along with bookmaker performance metrics for 144 matches played within 90 minutes during four editions of the UEFA EURO 2004 and 2016 (Table 1). The cforest technique was found to be slightly superior to other methods in terms of polynomial likelihood, with a score of 0.382, followed by the xgboost method with a score of 0.380. In terms of classification rate, both the cforest and xgboost methods exhibited the best performance, with a score of 0.486, which is quite close to the bookmaker's performance as a natural benchmark. All four methods performed equally well in terms of rank probability score, with the lasso technique having a slightly higher score of 0.210. The authors also evaluated the methods' performance in predicting the mean absolute error between the actual and predicted number of goals scored per game and per team, as well as the mean absolute error between the actual and predicted number of goals scored and goal difference (Table 2). The results for all four methods were generally similar, with the lasso method performing slightly better, with scores of 0.846 for goals scored and 1.148 for goals against (Groll, 2021).

Table 1: The results of four prediction methods - ranger, cforest, xgboost, and lasso - are averaged and presented.

	Likelihood	Class.Rate	RPS
ranger	0.372	0.458	0.216
cforest	**0.382**	**0.486**	0.213
xgboost	0.380	**0.486**	0.217
lasso	0.379	0.458	**0.210**
bookmakers	0.400	0.493	0.203

Table 2: Comparison of accurate goal scoring and goal gap predictin methods based on mean absolute error.

	Goals	Goal Difference
ranger	0.862	1.176
cforest	0.862	1.166
xgboost	0.883	1.162
lasso	**0.846**	**1.148**

Haruna et al. provided five prediction models, namely Naïve Bayes, K-NN, SVM, Random Forest, and Logistic Regression. Based on the results of all the experiments conducted so far, it was found that the combination of features used in Experiment 2 (which includes home team club, away team club, away team's average goals per game, home team's average goals per game, home team's ranking, away team's ranking, home team's attack, away team's attack, home team's defence, away team's defence, and number of goals scored), and the logistic regression classifiers used in the experiments, had the lowest prediction accuracy of 5.00% (Figure 1). On the other hand, the combination of features used in Experiment 5 (which includes home club, away club,

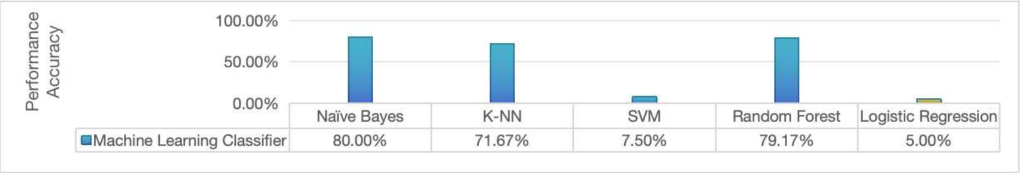

Figure 1: Results achieved in Experiment-II (Haruna, 2021).

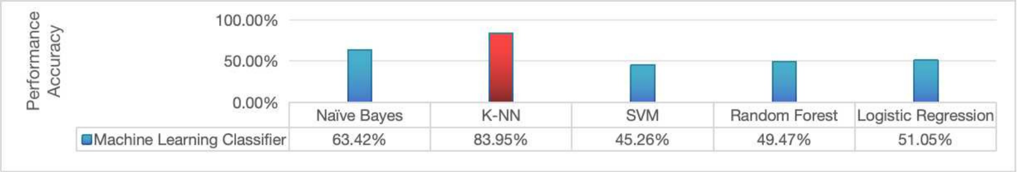

Figure 2: Results achieved in Experiment-V (Haruna, 2021).

22 players, and goals against) and the K-NN classifier produced the highest prediction accuracy of 83.95% (Figure 2) (Haruna, 2021).

Experiments 2 and 5, presented in the article, are considered the most convincing out of the five experiments due to their varying scales of dataset. Experiment 2 utilized data from 38 Manchester United matches in a season, while Experiment 5 used data from 380 matches of 20 teams in the Premier League's season. These two experiments can be compared to demonstrate how the dataset size impacts the accuracy of experimental results. Experiment 2 showcased a significant difference in classifier accuracy, whereas Experiment 5, when compared to Experiment 2, displayed a smaller difference in accuracy between each method, making it easier to determine the best performing machine learning classifier.

The game of football is subjective in nature, and the referee's decisions during the game may not always match the expectations of the fans, leading to chaos. This subjectivity makes it difficult to predict the outcome of football matches accurately. To address this, AI models are being introduced. However, if the model's predictions do not match the actual results, interpretability becomes crucial. Explaining the model's decision-making process is essential, as it helps establish trust and transparency.

To maximize the usefulness of AI models in football, they need to be applied to different levels of football leagues. While the English Premier League is one of the most competitive football leagues globally, extending the model to other leagues, such as the Chinese Super League or the Spanish LaLiga2, is necessary. This experiment will validate the model's performance in different environments and reveal its adaptability and robustness, ensuring it produces similar results in diverse match scenarios. However, predicting football matches accurately is not foolproof. With the explosion of data from football matches, models need to be trained faster and more efficiently to keep up with the pace and changes in the game. There is a new challenge of real-time analysis and prediction of large-scale data.

To solve this problem, Apache Spark emerges as a potential solution. Apache Spark is an open-source big data processing framework that is highly parallel and scalable (Salloum, 2016). Its built-in Spark Streaming module makes real-time data stream processing possible (Mehmood, 2020). By importing football match data into the Spark cluster, real-time analysis of large-scale data can be achieved to update and optimize models more quickly. In practical applications, real-time data from football matches can be transferred to the Spark cluster for feature extraction and model prediction in real-time. In this way, the model can reflect the changes in the match promptly and improve the accuracy of the prediction. Spark's distributed computing capability also allows models to be trained in parallel on large-scale datasets, speeding up the training process and improving model efficiency. It is possible to interpret the model and help users understand the decision basis of the model with the MLlib library in Spark (Özgüven, 2021).

The challenge is not only the real-time nature of the data but also how to handle the diversity and quality of the data. Football matches involve data from multiple sources, including team statistics, player performance, weather, and field conditions. To better predict the outcome of a match, these multiple sources of data need to be considered together, and complex models need to be constructed using some advanced solutions (Lu, 2023, Qiu, 2019). It is also crucial to ensure the quality of the data so that it does not negatively affect the model's predictions.

In the future, football match prediction models will face increasingly complex and diverse data and changes in the match environment. Therefore, it is an ongoing challenge to continuously optimize and update the models to improve their robustness and adaptability. The continuous development of technology will provide new data sources and analysis methods, offering more possibilities for further improvement of football match prediction models.

4 CONCLUSIONS

In this article, a review of the field of Machine Learning to analyse football data and predict outcomes is provided. The review covers a range of methods including Hybrid Machine Learning Models, Binary Classification, Regression, TOPSIS, Expert Systems, Ensemble Learning, and others. It was found that while these models can be useful, there may be some issues with their interpretability and usefulness, as well as limitations in terms of fast feedback. In order to establish trust with users, the model's interpretability needs to be further enhanced to clearly communicate the rationale and reasoning behind each prediction. Additionally, the quality and diversity of data can impact the models' effectiveness when dealing with complex football match situations. To improve the models, future researches are recommended to focus on expanding the dataset to include more information on leagues, teams, and players. Researchers shall also explore emerging technologies and methods to cope with the ever-changing field of football match data analysis.

REFERENCES

Baboota, R. & Kaur, H. 2019. Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), 741-755.

Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

Bunker, R. P. & Thabtah, F. 2019. A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), 27-33.

Çelikbilek, Y. & Tüysüz, F. 2020. An in-depth review of theory of the TOPSIS method: An experimental analysis. Journal of Management Analytics, 7(2), 281-300.

Gu, W., Foster, K., Shang, J. & Wei, L. 2019. A game-predicting expert system using big data and machine learning. Expert Systems with Applications, 130, 293-305.

Groll, A., Hvattum, L. M., Ley, C., Popp, F., Schauberger, G., Van Eetvelde, H. & Zeileis, A. 2021. Hybrid machine learning forecasts for the UEFA EURO 2020. arXiv preprint arXiv:2106.05799.

Groll, A., Ley, C., Schauberger, G. & Van Eetvelde, H. 2019. A hybrid random forest to predict soccer matches in international tournaments. Journal of Quantitative Analysis in Sports, 15(4), 271-287.

Haruna, U., Maitama, J. Z., Mohammed, M. & Raj, R. G. 2021. Predicting the outcomes of football matches using machine learning approach. In International Conference on Informatics and Intelligent Applications. Cham: Springer International Publishing, 92-104.

Hvattum, L. M. 2019. A comprehensive review of plus-minus ratings for evaluating individual players in team sports. International Journal of Computer Science in Sport, 18(1), 1-23.

Hwang, C. L., Yoon, K., Hwang, C. L. & Yoon, K. 1981. Methods for multiple attribute decision making. Multiple attribute decision making: methods and applications a state-of-the-art survey, 58-191.

Li, Y., Wang, L. & Li, F. 2021. A data-driven prediction approach for sports team performance and its application to National Basketball Association. Omega, 98, 102123.

Liu, Y. & Wang, J. 2020. Research on post-match score mechanism of players based on artificial intelligence and clustering regression model. Journal of Intelligent & Fuzzy Systems, 39(4), 4869-4879.

Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X., & Zheng, W. 2023. The multi-modal fusion in visual question answering: a review of attention mechanisms. PeerJ Computer Science, 9, e1400.

Mehmood, E. & Anees, T. 2020. Challenges and solutions for processing real-time big data stream: a systematic literature review. IEEE Access, 8, 119123-119143.

Mienye, I. D. & Sun, Y. 2022. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129-99149.

Özgüven, Y. M., Gönener, U. & Eken, S. 2021. A Dockerized big data architecture for sports analytics.

Qader, M. A., Zaidan, B. B., Zaidan, A. A., Ali, S. K., Kamaluddin, M. A. & Radzi, W. B. 2017. A methodology for football players selection problem based on multi-measurements criteria analysis. Measurement, 111, 38-50.

Qiu, Y., Chang, C. S., Yan, J. L., Ko, L., & Chang, T. S. 2019. Semantic segmentation of intracranial hemorrhages in head CT scans. In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) (pp. 112-115). IEEE.

Salloum, S., Dautov, R., Chen, X., Peng, P. X. & Huang, J. Z. 2016. Big data analytics on Apache Spark. International Journal of Data Science and Analytics, 1, 145-164.

Sudhaman, K., Akuthota, M. & Chaurasiya, S. K. 2022. A Review on the Different Regression Analysis in Supervised Learning. Bayesian Reasoning and Gaussian Processes for Machine Learning Applications, 15-32