Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/31/20230118

Research on predicting football matches based on handicap data and BPNN

Jiahao Hu

Dalian University of Technology, Dalian, Liaoning 116620, China

hu1124666055@163.com

Abstract. Football is one of the most influential sports in the world, and billions of people around the globe pay much attention to the football matches. With the growing popularity of football and the continuous development of the football betting industry, the prediction of the outcomes of football matches has become a hot topic in the commercial operations of sports especially footballs in recent years. It is also an important subject of academic research. In this paper, we develop a football match result prediction model based on the back propagation neural network. We take the German Bundesliga competitions as the research object in this paper. In addition to utilizing historical statistic data and team attributes from previous matches, we also incorporate a new dataset, known as handicap data, which refers to the odds data of the football matches, as the input layer of the BPNN (back propagation neural networks) for prediction. We also innovatively use varying numbers of hidden nodes, which greatly improves the prediction accuracy and stability of the model. Experimental results indicate that the average prediction accuracy of this football match prediction model is around 57.2%, with the highest prediction accuracy reaching 59.8% and the lowest prediction accuracy at 53.8%. The prediction model demonstrates relative stability, with no significant fluctuations in prediction accuracy.

Keywords: Football Match Win-Draw-Lose Prediction, Back Propagation Neural Networks, Handicap Data, Machine Learning.

1. Introduction

Football has a far-reaching influence worldwide, engaging numerous individuals involved in football-related work. The football industry has experienced rapid development and is currently the outdoor sport with the highest production value, greatest influence, and highest public attention in the world. In the field of sports, football has achieved a leading position far ahead of other sports. According to relevant data statistics, football, hailed as the "17th largest economy in the world," has an annual industry production value accounting for 43% of the total output value of the sports industry, reaching $500 billion, surpassing the Gross Domestic Product (GDP) of many developed countries and regions. Undoubtedly, it is the world's number one sport, far surpassing other sports such as basketball, golf, baseball, and F1 racing. According to Fédération Internationale de Football Association (FIFA) statistics, as of July 2016, there were over 200 million athletes from 1.5 million football teams worldwide participating in various football competitions. In addition, the number of individuals working in football-related jobs has reached 30 million. It is estimated that those involved in football and related occupations account for approximately 3% of the world's total population.[1] Since the 21st century, people's living standards have greatly improved, and sports-related industries have also advanced rapidly and steadily. Sports lottery, as one of the sports-related industries, has also gained wide support and attention from various countries. Over the past few decades, this sport has experienced steady development, and as a result, the business revenue in the related betting industry has significantly grown. However, due to the high unpredictability of football matches, it is a highly complex task to achieve stable and accurate predictions of football match outcomes. Without analyzing match data and statistical information, the success rate of betting on football matches is typically low.

However, with the rapid development of computer technology, the emergence of machine learning has opened up new opportunities for addressing this issue. Machine learning, particularly the widespread adoption of neural networks, has recently reached its pinnacle. The results of football matches can be predicted based on various machine learning algorithms and various characteristic data, such as the strength of both teams, their rankings, and their performance in home and away matches. In this paper, we utilize BP neural networks, multiple match feature data, and handicap data to predict football match results. And we also innovatively use varying numbers of hidden nodes of the BP neural network, which greatly improves the prediction accuracy and stability of the model.

The remainder of this paper is organized as follows. Section 2 provides a brief introduction of football betting rules. And we also provide a brief summary of previous related work in this section. Section 3 mainly elaborates on the work we have conducted, including data acquisition, data processing, and data mining and predicting. It mainly describes the methods used in our research and the process of building the BP neural network. Section 4 discusses and interprets the prediction results of the BP neural network model. We also explore the limitations of the prediction model and the areas where our work can be improved in this section. Finally, in Section 5, we summarize the manuscript and provide prospects for future work.

2. Background and literature review

2.1. Background

Back Propagation (BP) neural network is one of the most popular neural network architectures, not only because of its implementation complexity, but also because of its high productivity and efficiency. It is a multi-layer feedforward neural network trained by the error backpropagation algorithm. The BP neural network generally consists of three parts: input layer, hidden layer, and output layer, each of which is composed of several neurons, that is, several feature data. The hidden layer may also include multiple layers of neurons, and its quantity is one of the decisive factors affecting the efficiency and accuracy of the neural network. The learning of the BP neural network is a supervised learning process. The model continuously adjusts the parameters in the neural network by learning and backpropagating the error to correct the error, ultimately achieving or approaching the expected mapping relationship between input and output.

Sports betting is a practice that involves predicting the outcome of a sporting event, placing a certain amount of money, and potentially earning profits if the prediction is correct. However, if the prediction is incorrect, the individual stands to lose the entire amount of money placed on the bet. In the case of football matches, there are typically three possible outcomes: a win, a draw, or a loss. This makes the probability of randomly guessing the correct result only 1/3, rendering random betting highly likely to result in financial losses. Therefore, it is important to conduct thorough analysis and research to improve the accuracy of predictions and increase the chances of making a profit in sports betting. By analyzing match data and statistical information, individuals can make informed decisions and potentially increase their chances of success in sports betting.

2.2. Literature review

In recent years, the application of machine learning in the sports field has significantly increased to help predict sports match results and make decisions. In 2019, Rahul Baboota, Harleen Kaur, and others used data from the English Premier League to create a feature set for determining the most important factors for predicting football match results. They also used machine learning to create a highly accurate predictive system. Their best model achieved a performance of 0.2156 on the Ranked Probability Score (RPS) metric for game weeks in the English Premier League, but this performance still fell short of bookmaker predictions.[2] In 2022, Fátima Rodrigues, Ângelo Pinto, and others focused on the profitability of football betting and used data from 1900 matches in five seasons (2013/2014 to 2018/2019) of the English Premier League. They conducted correlation analysis on 31 variables and compared the predictive accuracy of several machine learning algorithms, such as K-nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM), using the selected highly correlated variables. The RF model achieved the highest prediction accuracy of 65.26%.[3]

Research on using BP neural networks for predicting football match results is also a hot topic in the academic community. For example, in 2018, Fu Yu utilized neural networks and feature data such as shots on target, free kicks, and corner kicks from 24 matches of FC Barcelona during the 2015-2016 and 2016-2017 seasons to predict match outcomes. The dataset was divided into training and testing sets in a 3:1 ratio. His Multi-Layer Perceptron neural network model consisted of 4 input layers and 3 output layers, with a fixed hidden layer of 6 neurons. He implemented the research using the R programming language and assessed the prediction accuracy through a confusion matrix, which indicated reliable overall prediction accuracy.[4] Another example is the research conducted by Yiqi Wang, Hongrun Zhao, and others in 2021. They employed social network analysis and BP neural networks to predict match outcomes based on player behaviors on the football field. Their study considered passing success rate, participation in passing, player interaction, and core player dependence as passing factors. They also included attacking-defending strategy, ball possession rate, successful interception rate, and defensive interception rate as attacking-defending factors. Furthermore, they incorporated team injury occurrences, coach performance, and home/away advantage as player status factors. By utilizing these three sets of factors, their BPNN achieved a prediction accuracy of 81.8%.[5] In the same year, Yi Huang utilized BP neural networks to identify the most influential factors on match results and used these selected feature factors for predicting football match outcomes. The research achieved a prediction accuracy of 75%.[6]

There is also a case study that utilized handicap data for football match result prediction. In 2016, Xichen Ao and others used Python tools to scrape odds data from some matches in the English League One, and they built a model using principal component analysis and logistic regression. The model's prediction accuracy was verified to be above 70.1% [7].

It is worth mentioning that the researchers mentioned above used different feature data and different datasets, resulting in significant variations in prediction accuracy.

3. Method

3.1. Data Description

The data used in this study consists of 612 records from a single season (2021/2022) of the German Bundesliga, which includes a total of 306 football matches. We extracted eight types of feature data related to football matches from the “All Football” app:

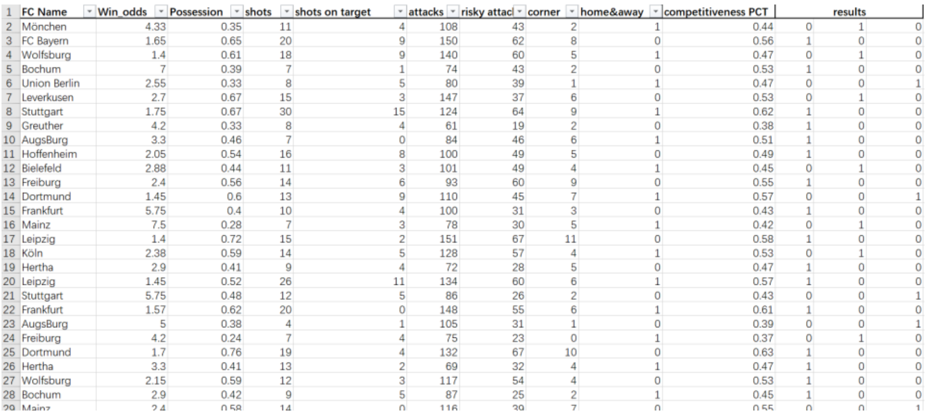
(1) Possession rate of both teams.
(2) Number of shots by both teams.
(3) Number of shots on target by both teams.
(4) Number of attacks by both teams.
(5) Number of dangerous attacks by both teams.
(6) Number of corner kicks by both teams.
(7) Home and away match indicators.
(8) Team strength index of both teams.

The team strength index is a comprehensive indicator calculated by the “All Football” app, taking into account recent team performance, goals scored and conceded, and market value.

For obtaining handicap data, similar to [7], we used a web scraping tool, Python, to extract odds data from the Bet365 company for the 2021/2022 German Bundesliga season, sourced from the website https://odds.500.com/.

(9) Win-draw-loss odds data from Bet365 company.

For missing values, we manually supplemented the data. In the end, we obtained a total of 612 original data records, as shown in Fig. 1.



1	FC Name	Win_odds	Possession	shots	shots on target	attacks	risky attacl	corner	home&away	competitiveness PCT	results		
2	Mönchen	4.33	0.35	11	4	108	43	2	1	0.44	0	1	0
3	FC Bayern	1.65	0.65	20	9	150	62	8	0	0.56	1	0	0
4	Wolfsburg	1.4	0.61	18	9	140	60	5	1	0.47	0	1	0
5	Bochum	7	0.39	7	1	74	43	2	0	0.53	1	0	0
6	Union Berlin	2.55	0.33	8	5	80	39	1	1	0.47	0	0	1
7	Leverkusen	2.7	0.67	15	3	147	37	6	0	0.53	0	1	0
8	Stuttgart	1.75	0.67	30	15	124	64	9	1	0.62	1	0	0
9	Greuther	4.2	0.33	8	4	61	19	2	0	0.38	1	0	0
10	AugsBurg	3.3	0.46	7	0	84	46	6	1	0.51	1	0	0
11	Hoffenheim	2.05	0.54	16	8	100	49	5	0	0.49	1	0	0
12	Bielefeld	2.88	0.44	11	3	101	49	4	1	0.45	0	1	0
13	Freiburg	2.4	0.56	14	6	93	60	9	0	0.55	1	0	0
14	Dortmund	1.45	0.6	13	9	110	45	7	1	0.57	0	0	1
15	Frankfurt	5.75	0.4	10	4	100	31	3	0	0.43	1	0	0
16	Mainz	7.5	0.28	7	3	78	30	5	1	0.42	0	1	0
17	Leipzig	1.4	0.72	15	2	151	67	11	0	0.58	1	0	0
18	Köln	2.38	0.59	14	5	128	57	4	1	0.53	0	1	0
19	Hertha	2.9	0.41	9	4	72	28	5	0	0.47	1	0	0
20	Leipzig	1.45	0.52	26	11	134	60	6	1	0.57	1	0	0
21	Stuttgart	5.75	0.48	12	5	86	26	2	0	0.43	0	0	1
22	Frankfurt	1.57	0.62	20	0	148	55	6	1	0.61	1	0	0
23	AugsBurg	5	0.38	4	1	105	31	1	0	0.39	0	0	1
24	Freiburg	4.2	0.24	7	4	75	23	0	1	0.37	0	1	0
25	Dortmund	1.7	0.76	19	4	132	67	10	0	0.63	1	0	0
26	Hertha	3.3	0.41	13	2	69	32	4	1	0.47	1	0	0
27	Wolfsburg	2.15	0.59	12	3	117	54	4	0	0.53	1	0	0
28	Bochum	2.9	0.42	9	5	87	25	2	1	0.45	1	0	0
29	Mainz	2.4	0.58	14	0	116	39	7	0	0.55	0	0	1

Figure 1. Football match raw data.

3.2. *Data Processing*

In Section 3.1, we normalized the nine types of feature data. Among them, (1), (7), and (8) were already normalized and can be directly used as input to the neural network. As for (2), (3), (4), (5), (6), and (9), we grouped the data pairwise for both teams in a football match and applied the following formula for normalization:

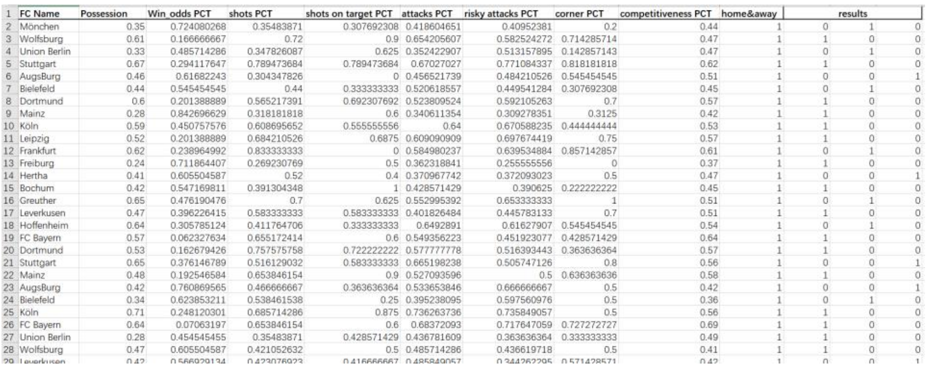
$$p = \frac{data1}{data1 + data2} \qquad (data1 \neq 0 \text{ 或} data2 \neq 0)$$

$$p = 0.5 \quad (data1 = data2 = 0)$$

Among the feature data, data1 represents the features of the home team, and data2 represents the features of the opposing team. Both data1 and data2 are normalized to the range [0, 1], which facilitates their use as input for the neural network.

Regarding the representation of football match results, similar to [4], we adopted a categorical label approach. We represent the outcomes of "win," "draw," and "loss" using category labels as follows: "win" is represented as "1 0 0," "draw" is represented as "0 1 0," and "loss" is represented as "0 0 1." These labels serve as the output of the neural network.

In the end, we obtained 612 records of normalized data, as shown in Fig. 2.



1	FC Name	Possession	Win_odds PCT	shots PCT	shots on target PCT	attacks PCT	risky attacks PCT	corner PCT	competitiveness PCT	home&away	results		
2	Mönchen	0.35	0.724080268	0.35483871	0.307692308	0.418604651	0.40952381	0.2	0.44	1	0	1	0
3	Wolfsburg	0.61	0.166666667	0.72	0.9	0.654205607	0.582524272	0.714285714	0.47	1	1	0	0
4	Union Berlin	0.33	0.485714286	0.347826087	0.625	0.352422907	0.513157895	0.142857143	0.47	1	0	1	0
5	Stuttgart	0.67	0.294117647	0.789473684	0.789473684	0.67027027	0.771084337	0.818181818	0.62	1	1	0	0
6	AugsBurg	0.46	0.61682243	0.304347826	0	0.456521739	0.484210526	0.545454545	0.51	1	0	0	1
7	Bielefeld	0.44	0.545454545	0.44	0.333333333	0.520618557	0.449541284	0.307692308	0.45	1	0	1	0
8	Dortmund	0.6	0.201388889	0.565217391	0.692307692	0.523809524	0.592105263	0.7	0.57	1	1	0	0
9	Mainz	0.28	0.842696629	0.318181818	0.6	0.340611354	0.309278351	0.3125	0.42	1	1	0	0
10	Köln	0.59	0.450757576	0.608695652	0.555555556	0.64	0.670588235	0.444444444	0.53	1	1	0	0
11	Leipzig	0.52	0.201388889	0.684210526	0.6875	0.609090909	0.697674419	0.75	0.57	1	1	0	0
12	Frankfurt	0.62	0.238964992	0.833333333	0	0.584980237	0.639534884	0.857142857	0.61	1	0	1	0
13	Freiburg	0.24	0.711864407	0.269230769	0.5	0.362318841	0.255555556	0	0.37	1	1	0	0
14	Hertha	0.41	0.605504587	0.52	0.4	0.370967742	0.372093023	0.5	0.47	1	0	0	1
15	Bochum	0.42	0.547169811	0.391304348	1	0.428571429	0.390625	0.222222222	0.45	1	1	0	0
16	Greuther	0.65	0.476190476	0.7	0.625	0.552995392	0.653333333	1	0.51	1	0	1	0
17	Leverkusen	0.47	0.396226415	0.583333333	0.583333333	0.401826484	0.445783133	0.7	0.51	1	1	0	0
18	Hoffenheim	0.64	0.305785124	0.411764706	0.333333333	0.6492891	0.61627907	0.545454545	0.54	1	0	1	0
19	FC Bayern	0.57	0.062327634	0.655172414	0.6	0.549356223	0.451923077	0.428571429	0.64	1	1	0	0
20	Dortmund	0.53	0.162679426	0.757575758	0.722222222	0.577777778	0.516393443	0.363636364	0.57	1	1	0	0
21	Stuttgart	0.65	0.376146789	0.516129032	0.583333333	0.665198238	0.505747126	0.8	0.56	1	0	0	1
22	Mainz	0.48	0.192546584	0.653846154	0.9	0.527093596	0.5	0.636363636	0.58	1	1	0	0
23	AugsBurg	0.42	0.760869565	0.466666667	0.363636364	0.533653846	0.666666667	0.5	0.42	1	0	0	1
24	Bielefeld	0.34	0.623853211	0.538461538	0.25	0.395238095	0.597560976	0.5	0.36	1	0	1	0
25	Köln	0.71	0.248120301	0.685714286	0.875	0.736263736	0.735849057	0.5	0.56	1	1	0	0
26	FC Bayern	0.64	0.07063197	0.653846154	0.6	0.68372093	0.717647059	0.727272727	0.69	1	1	0	0
27	Union Berlin	0.28	0.454545455	0.35483871	0.428571429	0.436781609	0.363636364	0.333333333	0.49	1	1	0	0
28	Wolfsburg	0.47	0.605504587	0.421052632	0.5	0.485714286	0.436619718	0.5	0.41	1	1	0	0
29	Leverkusen	0.42	0.566929134	0.423076923	0.416666667	0.485849057	0.344262295	0.571428571	0.42	1	0	0	1

Figure 2. Football match processed data.

3.3. Data mining and predicting

In Section 3.2, we used the normalized feature data obtained in Section 3.1 as the input layer of the neural network. The corresponding category labels for different football match results were used as the output layer of the neural network. The dataset was divided into a training set and a test set in a 7:3 ratio, with 428 records for training and 184 records for testing.

As for the activation function of the BP neural network, we chose the "tansig" function. The "tansig" function allows the output to maintain a nonlinear monotonic relationship with the input, which is suitable for the gradient calculation in the BP network and exhibits good fault tolerance. The function expression is as follows:

$$y = \frac{2}{1+e^{-2x}}\text{-}1$$

In the case of selecting the number of neurons in the hidden layer, there is no standard method so far. However, a rough range can be determined using the following formula:

$$h = \sqrt{m+n} + a \quad (1 \leq a \leq 10, a\epsilon Z)$$

Based on the range of hidden neurons mentioned earlier, we iterate from 4 to 13 hidden nodes and calculate the Mean Squared Error (MSE) for each training set. The hidden layer node count that yields the minimum MSE is selected as the optimal configuration for testing the model on the test set.

We utilize the trained model to make predictions, similar to the approach in [4]. We generate a confusion matrix using the confusionMat() function to visualize the experimental results and calculate the prediction accuracy based on the confusion matrix.

The pseudocode introduction of my MATLAB code is provided below:

```
1. Read data from 'input.xlsx' file and split it into input and output data.
2. Define the number of training and testing samples.
3. Split the data into training and testing sets.
4. Initialize the mean square error (MSE) to a high value.
5. Define the activation and training functions for the neural network.
6. For each hidden layer neuron number between sqrt(inputnum+outputnum)+1 and sqrt(inputnum+outputnum)+10:
    a. Create a neural network with the defined architecture.
    b. Set the training parameters.
    c. Train the neural network.
    d. Compute the mean square error of the training set.
    e. If the MSE is lower than the previous value, keep the current number of hidden neurons as the best.
7. Create the final neural network with the best number of hidden neurons.
8. Set the training parameters.
9. Train the neural network using the training set.
10. Use the trained neural network to predict the output of the testing set.
11. Convert the predicted output to class labels.
12. Convert the true output to class labels.
13. Compute the confusion matrix.
14. Compute the accuracy of the model.
15. Print the confusion matrix and the accuracy.
```

4. Discussion

The proposed BP neural network prediction model in this study achieves an average prediction accuracy of approximately 57.2%. The highest prediction accuracy is around 59.8%, while the lowest prediction accuracy is around 53.8%. The prediction model is relatively stable, with no significant fluctuations in accuracy. Furthermore, the prediction accuracy significantly surpasses the 33.3% probability of randomly predicting football match results without considering match and analysis data. The main contribution of this study lies in incorporating a new dataset - odds data - as a feature for predicting football match results. Compared to [4], this study also innovatively uses varying numbers of hidden nodes, which greatly improves the prediction accuracy and stability of the model.

During practical predictions, since we cannot obtain data such as ball possession, shot attempts, and corner kicks for upcoming matches in advance, we can substitute these missing data with the average values of the team's previous 10 matches. This allows us to make predictions for actual football match results.

However, this study also has several limitations. Firstly, due to limited data collection capabilities, our dataset is still relatively small, which directly affects the final prediction accuracy. Secondly, we have a limited number of feature variables and have not performed correlation testing on the feature data. Future work can involve increasing the variety of feature data and conducting correlation testing to select relevant features and improve prediction accuracy. Lastly, in building the neural network model, we did not conduct in-depth research and exploration of parameters such as learning rate. We only used commonly used values for parameter tuning and did not depict the learning rate's variation curve with respect to loss to select the optimal parameters for tuning.

5. Conclusion

This paper describes the development of a football match result prediction model using BP neural networks. It utilizes a dataset of 612 records from 306 matches of the 2021/2022 German Bundesliga. The dataset consists of two different sources: football match-related data and team strength indices extracted from the "All Football" app, and odds data obtained from Bet365. A total of 9 feature variables are collected as inputs for the neural network, and categorical labels representing "win," "draw," and "loss" are used as the network's outputs. By varying the number of hidden layers, the best configuration is determined through validation on the test set, achieving an average prediction accuracy of approximately 57.2%, with a highest accuracy of around 59.8% and a lowest accuracy of around 53.8%. These results significantly outperform the 33.3% prediction probability achieved without considering match and analysis data.

The model exhibits great potential for further development. As future work, expanding the dataset and increasing the variety of feature variables can be considered. Additional factors such as passing attempts, passing accuracy, tackles, clearances, and the physiological and psychological states of players can be included to enhance the model's prediction accuracy for football match results. Moreover, integrating the prediction model with other machine learning models into a decision support system can help evaluate betting risks and provide gamblers with insights into the risks involved. This integration can create more value for the sports betting industry and have a greater impact on the world of football.

References

[1] Xia Fei. (2017). Application research of BP neural network in predicting football match outcomes. Chongqing Normal University Press, Chongqing.

[2] Rahul Baboota, Harleen Kaur. (2019). Predictive analysis and modelling football results using machine learning approach for English premier league. International Journal of Forecasting, 35:741-755.

[3] Fátima Rodrigues, Ângelo Pinto. (2022). Prediction of football match results with machine learning. Procedia Computer Science, 204:463-470.

[4] Fu Yu. (2018). Neural network for predicting football match outcomes. Science and Technology Review, 23:238.

[5] Wang Yiqi, Zhao Hongrun, Zhao Hongwen, Xu Xichen. (2021). Research on football match prediction based on social network analysis and BP neural network. Journal of Weifang Engineering Vocational College, 34(3):104-108.

[6] Huang Yi. (2021). Prediction of football match results by using neural network. Microcomputer Applications, 37(11):137-140.

[7] Ao Xiqin, Gong Yujie, Li Jian. (2016). Prediction of football match outcomes based on odds data. Journal of Chongqing Technol Business Univ (Nat Sci Ed), 33(6):85-89.