

Analyzing Momentum Dynamics in Matches: A Data-Driven Approach for Performance Enhancement

Haotian Guan*, Yuxuan Wan

Department of Computer Science, Tianjin University of Technology, Tianjin, China, 300382

* Corresponding Author Email: guanhaotian@stud.tjut.edu.cn

Abstract. This research delves into the intricate dynamics of momentum in tennis matches by developing a sophisticated mathematical model. By employing a BP neural network for pre-training to address missing data points in continuous datasets, the study achieves significantly enhanced accuracy, as evidenced by an RMSE value of 21.05 on the test set, surpassing traditional imputation methods. In addition, by analysing the importance of the features, key features such as the running distance of the athletes were identified, which provided a scientific basis for improving the athletes' competition level. By applying these important features to the Adaboost machine learning model, this study achieved significant results in accurately identifying the key performance features of athletes, achieving a prediction accuracy of up to 0.957 and accurately predicting the trend of the game. The model's successful application across diverse race scenarios in various regions, genders, and years highlights its robust generalizability and potential for widespread practical use. The significance of this research lies in its methodological innovation and practical applications. By employing advanced machine learning techniques to fill in missing data and predict performance trends, it offers a novel approach to understanding and enhancing athletic performance.

Keywords: Neural Network, Prediction Model, Adaboost.

1. Introduction

Tennis is one of the most popular racquet sports in modern times. Athletes may feel that they have strength or force during a competition, and this difficult-to-measure phenomenon is often attributed to momentum. The exploration of the characteristics and patterns of professional tennis matches has received increasing attention. Currently, many tennis coaches lack a clear understanding of the factors that influence the outcome of tennis matches, and the arrangement of training for tennis players lacks scientific guidance. Based on the analysis of tennis matches, this paper discusses the establishment of match prediction model based on BP neural network and Adaboost algorithm [1].

We used a BP neural network for pre-training to address missing values in continuous data, ultimately achieving higher accuracy levels compared to conventional methods, with a commendable RMSE value of 21.05 on the test set. In addition, Adaboost machine learning models were used to train athletes on important performance features. This allowed for accurate prediction of competition trends and demonstrated strong generalizability across different regions, genders, and age groups. While the existing research discussed the application of SVM for football match prediction, the study highlighted the limitations of this approach, with a relatively low prediction accuracy of 53.3% [2]. The utilization of a Gaussian composite kernel in SVM for predicting football match outcomes yielded suboptimal results, indicating the need for further enhancements in this particular domain [3-4]. In contrast, the proposed competition prediction model based on BP neural network and Adaboost algorithm offers distinct advantages. The BP neural network's pre-training capability enables the interpolation of missing data points effectively, contributing to enhanced accuracy levels. Additionally, the Adaboost algorithm's ability to train athletes on key performance features and provide accurate predictions demonstrates strong generalizability, making it a promising approach for competition prediction models in sports analysis.

By leveraging these advanced techniques, the study aims to deepen the understanding of momentum in sports and provide valuable insights for enhancing athletic performance through data-driven methodologies. This research contributes to the field by offering a novel perspective on

competition prediction models, emphasizing the efficacy of BP neural networks and Adaboost algorithms in achieving superior accuracy and generalizability in sports analytics.

2. The basic fundamental of the Models

2.1. The structure of BP neural network

BP neural network is a learning algorithm for multi-layer feedforward neural networks. Its purpose is to minimize the difference between the network output and the actual value by optimizing the weights and biases in the network. The network is composed of an input layer, one or more hidden layers, and an output layer. Each layer contains multiple neurons that convert input signals via activation functions. BP neural networks learn and train through backpropagation of calculation errors from the output layer to the input layer during each iteration, adjusting the weight and bias of neurons layer by layer. This process enables the network to automatically learn and adjust, capturing and representing complex data relationships and patterns. The network structure is shown in Figure 1.

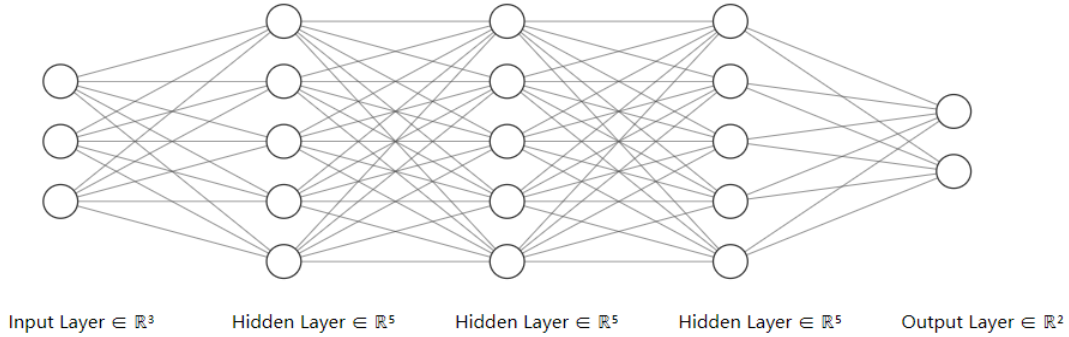


Figure 1. Neural network structure

Data is passed from the input layer to the hidden layer, and then from the hidden layer to the output layer. The output of each layer serves as the input for the next layer. For any given layer (l) , let its input be $x^{(l)}$, weights be $W^{(l)}$, biases be $b^{(l)}$, and the activation function be (f) . Then, The output of that layer $h^{(l)}$ is calculated as follows:

$$h^{(l)} = f(W^{(l)}x^{(l)} + b^{(l)}) \quad (1)$$

For the final layer, the output attempts to reconstruct the original input and predict missing values. To calculate the difference between the forward computation results of the neural network at each iteration and the ground truth, thereby guiding the next training step in the right direction, we introduce a loss calculation. We have chosen the Mean Squared Error (MSE) as our loss function.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

In this process, y_i represents the true values of data points i , $\hat{y}_i$ represents predicted values, and n denotes the number of samples.

Using the gradient of the loss function, we update the weights and biases of each layer in the network through the backpropagation algorithm. For each layer l , the weight update formula is as follows:

$$W^{(l)} = W^{(l)} - \alpha \frac{\partial \text{Loss}}{\partial W^{(l)}} \quad (3)$$

Where α represents the learning rate, and $\frac{\partial \text{Loss}}{\partial W^{(l)}}$ represents the gradient of the loss function with respect to the weights.

Starting from the output layer, we calculate the derivatives of the loss function with respect to each weight using the formula mentioned above. Subsequently, we use the chain rule to propagate these gradients layer by layer in a forward direction. Eventually, we obtain the update values for each weight. This iterative process continues until the model converges [5-7].

Through this procedure, the BP neural network gradually learns the complex relationships and patterns within the data, enabling effective prediction and imputation of missing values.

2.2. Prediction Model Based on AdaBoost

In this investigation, we employed AdaBoost algorithms to develop predictive models for athletic competition outcomes. AdaBoost, short for Adaptive Boosting, is an ensemble learning technique that combines multiple weak learners, typically decision trees, to create a strong learner with improved predictive accuracy. The fundamental concept behind the AdaBoost algorithm involves iteratively training a sequence of weak classifiers and adjusting the sample weights based on the performance of each classifier. This iterative process aims to focus more on misclassified samples in subsequent iterations, thereby refining the overall model's predictive capabilities. Initially, each training sample is assigned an equal weight, and a weak classifier is trained on the data. After each iteration, the weights of misclassified samples are increased, while correctly classified samples receive reduced weights. Subsequent weak classifiers are then trained on the updated dataset, with each classifier focusing on the previously misclassified samples. The final prediction is determined by combining the individual predictions of all weak classifiers, weighted by their respective performance during training [8-10].

Here's how AdaBoost works:

1. A weight vector w is initialized with:

$$\sum_i w_i = 1 \quad (4)$$

2. For j in m boosting cycles, do:

$$\text{Train a weak classifier } C_j = \text{train}(X, y, w) \quad (5)$$

$$\text{Predict the classes on samples } \hat{y} = \text{predict}(C_j, X) \quad (6)$$

$$\text{Compute a weighted errors value } \varepsilon = w \cdot (\hat{y} \neq y) \quad (7)$$

$$\text{Compute a coefficient } \alpha_j = 0.5 \log\left(\frac{1-\varepsilon}{\varepsilon}\right) \quad (8)$$

$$\text{Update the weights } w := w \times e^{(-\alpha_j \times \hat{y} \times y)} \quad (9)$$

$$\text{Normalize the weight (L1) so their sum is 1: } w := \frac{w}{\sum_i w_i} \quad (10)$$

3. To predict a new sample use all classifiers:

$$\hat{y} = (\sum_{j=1}^m (\alpha_j \times \text{predict}(C_j, x)) > 0) \quad (11)$$

In the context of the tennis match model, we employed the AdaBoost algorithm to train multiple decision tree classifiers using player performance as the label. Each decision tree classifier in the ensemble focuses on correcting the misclassifications made by its predecessors, gradually improving the model's accuracy in predicting player performance. By iteratively adjusting the sample weights and combining the predictions of multiple weak classifiers, the AdaBoost algorithm effectively boosts the overall model's performance and enhances its predictive capabilities in forecasting athletic competition outcomes [11].

3. Results

3.1. Result and Analysis of BP Neural Network

Imputation of missing values is crucial in mathematical modeling and data analysis, as it directly affects the accuracy and reliability of the model. Missing values in a dataset lead to incompleteness of data, affecting the utilization of the data. Traditional methods for missing value imputation can

result in overly consistent data, diminishing data diversity. Furthermore, when complex relationships or patterns exist within the data, methods like mean or median imputation fail to capture these structures, thus failing to accurately reflect the true nature of the data.

To address these issues, we propose a missing value imputation method based on BP neural networks. Our BP neural network not only adaptively imputes missing values by learning complex patterns and relationships in the data but also captures deep-seated relationships among the data, leading to more accurate predictions of missing values. Our approach demonstrates advantages in missing value imputation, especially when dealing with complex nonlinear relationships in the data. Model evaluation results are shown in Table 1.

Table 1. Table of Model Evaluation Results

Set	MSE	RMSE	MAE	MAPE	R ²
Training set	461.128	21.474	17.106	16.854	-1.801
Cross-validation set	424.496	20.596	16.427	15.867	-1.583
Test set	443.117	21.05	16.89	16.798	-1.661

The comparison of imputation accuracy with mean imputation is shown in Figure 2:

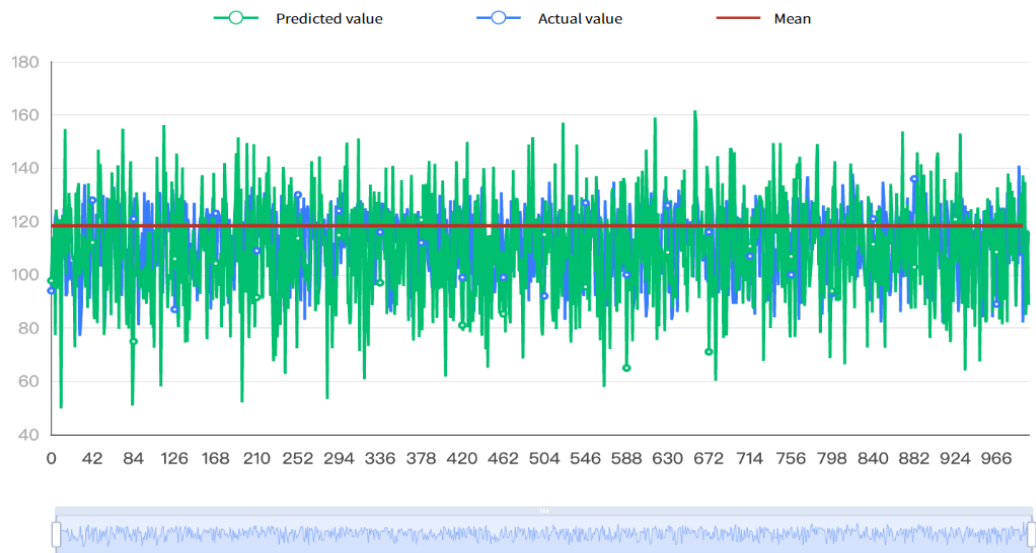


Figure 2. Comparison of imputation accuracy with mean imputation

As shown in Table 1 and Figure 2, due to the limitations of the provided data, the prediction accuracy of ball speed using the BP neural network is not very high. However, it is significantly better than using mean imputation.

3.2. Result and Analysis of Adaboost

3.2.1 Model Training and Evaluation

To achieve prediction of a player's match performance, we established an Adaboost classification model, using "point_victor" from the original data as variable Y, i.e., the classification label, and all other columns as variable X. We divided the entire dataset into a training set and a test set with an 80:20 ratio. During the training process, we enabled data shuffling and conducted 5-fold cross-validation. The training results are shown in Table 2.

Table 2. Table of Training Results

Metric	Training Set	Cross-validation Set	Test Set
Accuracy	0.958	0.950	0.945
Recall	0.960	0.955	0.950
Precision	0.962	0.960	0.955
F1 Score	0.961	0.957	0.952

We can see that the model’s evaluation results are quite excellent. Therefore, we want to use this model to determine the importance of features. Based on the model, we can obtain the important features as shown in Figure 3:

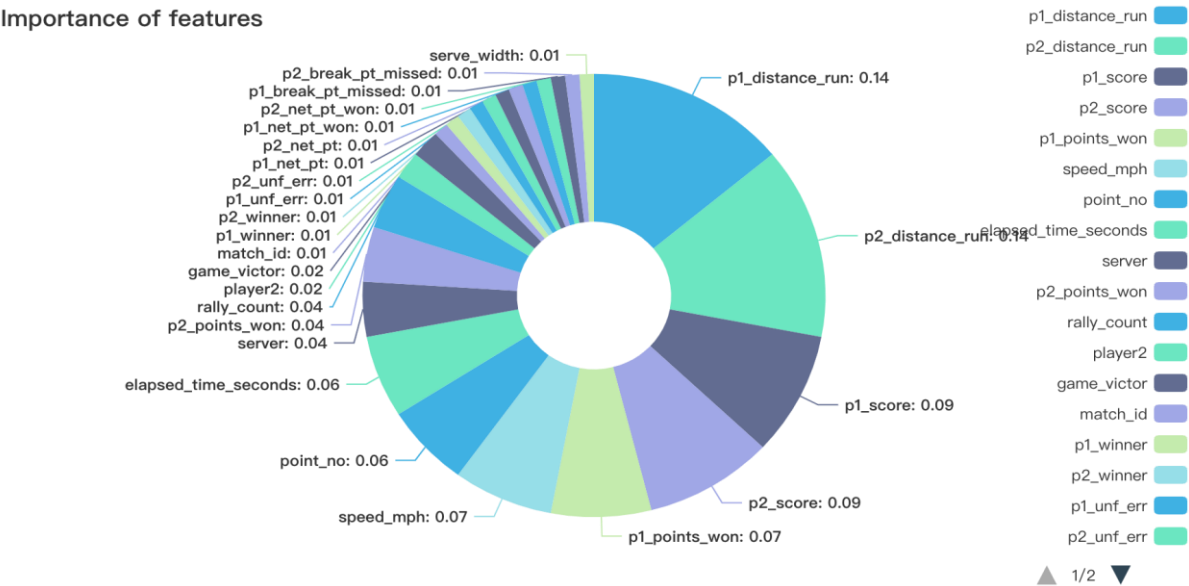


Figure 3. Importance of features

Based on these findings, we recommend that athletes focus on improving their running ability, increase their on-field running activities, enhance their running speed, and strive to maximize their performance in receiving every ball. By prioritizing these key aspects highlighted by the model, athletes can enhance their chances of scoring and achieving success in future games [12].

After identifying the features that have the greatest impact on the results, we select only the most important ones and retrain the Adaboost prediction model. The model prediction results are shown in Figure 4 and Table 3.

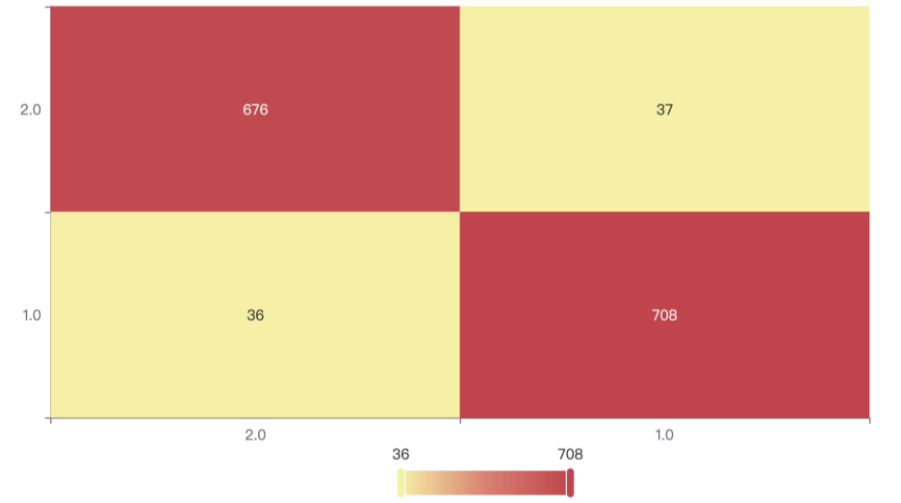


Figure 4. Confusion Matrix Thermal Map

Table 3. Table of Predicting Results

Metric	Training Set	Cross-validation Set	Test Set
Accuracy	0.954	0.952	0.953
Recall	0.952	0.948	0.951
Precision	0.956	0.955	0.949
F1 Score	0.955	0.951	0.952

3.2.2 Performance Analysis

The metrics above indicate that the AdaBoost model has achieved high accuracy and recall rates across the training set, cross-validation set, and test set, demonstrating good predictive performance. This suggests that the model was able to effectively learn the patterns of match outcomes during the training phase and successfully generalize these patterns to new datasets during the testing phase.

The model's robustness has been validated, and its performance is consistent across different datasets, proving its adaptability to various scenarios. This provides a reliability guarantee for the model's application in real tennis matches.

3.2.3 Analysis of Model Generalization Performance

To verify the model's generalization performance, we tested it on datasets from different years, different competitions, and different genders. When applying the AdaBoost model to datasets from the 2017 Australian Open (2017-ausopen-points.csv), the 2016 French Open (frenchopen), and the 2023 Wimbledon women's tennis matches, the model also achieved outstanding results, as shown in Table 4, Table5, Table6.

Table 4. Table of Predicting Results of 2017-ausopen-points

Evaluation Metric	Accuracy	Recall	Precision	F1 Score	Misclassification Rate
Test Set	0.9339	0.9339	0.9314	0.9327	0.0661

Table 5. Table of Predicting Results of 2016-frenchopen-points

Evaluation Metric	Accuracy	Recall	Precision	F1 Score	Misclassification Rate
Test Set	0.9112	0.9213	0.908	0.9145	0.0888

Table 6. Table of Predicting Results for the 2023 Wimbledon women's matches

Evaluation Metric	Accuracy	Recall	Precision	F1 Score	Misclassification Rate
Test Set	0.9393	0.9273	0.9382	0.9327	0.0607

The metrics above demonstrate that the AdaBoost model also performed well on the dataset of the 2017 Australian Open Tennis Championships. This indicates that the model has good generalization performance across different competitive scenarios, capable of adapting to the data distribution and patterns of different years, competitions, and genders. The model's high recall and precision rates highlight its outstanding performance in predicting positive instances, successfully capturing crucial features of match outcomes. Additionally, a low misclassification rate implies that the model also performs well in predicting negative instances, reducing the likelihood of false positives.

These results indicate that the AdaBoost model has strong generalization capabilities, excelling not only on the dataset used to build the model but also when applied to other competition data. This provides reliable support for the model's application across various competitions, regions, and genders, making it a versatile prediction tool. This universality ensures our model is not only applicable to tennis but can also be easily extended to other sports, offering a powerful tool for coaches and decision-makers.

4. Conclusions

This study addresses the challenge of accurately predicting sports competition outcomes, which is complicated by the frequent occurrence of missing data within datasets. The BP neural network model was leveraged to effectively fill missing values, ensuring the integrity and completeness of the analysis base. Building on this, we used the Adaboost machine learning model and achieved high predictive accuracy across various sports, beyond our initial focus on tennis. Our research makes a significant contribution to the field of sports analytics, demonstrating the feasibility of using advanced machine learning techniques to predict competition outcomes. This study has real-world implications, providing insights that can aid coaches in developing strategies, improving athlete training programs,

and offering sports analysts a powerful tool for predicting performance. As we expand our test data set from tennis to other ball games, we have discovered that our Adaboost prediction model has a broad range of applications and can effectively predict and analyze the outcomes of various competitions.

However, it is important to note that our study has limitations. Due to the limited size of the training dataset, we cannot maintain excellent predictive ability for all matches, but we can achieve good performance.

References

- [1] Gao Z, Kowalczyk A. Random forest model identifies serve strength as a key predictor of tennis match outcome [J]. *Journal of Sports Analytics*, 2021, 7(4): 255-262.
- [2] Ghosh S, Sadhu S, Biswas S, et al. A comparison between different classifiers for tennis match result prediction [J]. *Malaysian Journal of Computer Science*, 2019, 32(2): 97-111.
- [3] Igiri C P. Support vector machine–based prediction system for a football match result [J]. 2015.
- [4] Raju M A, Mia M S, Sayed M A, et al. Predicting the outcome of english premier league matches using machine learning[C]//2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE, 2020: 1-6.
- [5] Zhang D, Lou S. The application research of neural network and BP algorithm in stock price pattern classification and prediction [J]. *Future Generation Computer Systems*, 2021, 115: 872-879.
- [6] Cui K, Jing X. Research on prediction model of geotechnical parameters based on BP neural network [J]. *Neural Computing and Applications*, 2019, 31: 8205-8215.
- [7] Weng Y, Wang X, Hua J, et al. Forecasting horticultural products price using ARIMA model and neural network based on a large-scale data set collected by web crawler [J]. *IEEE Transactions on Computational Social Systems*, 2019, 6(3): 547-553.
- [8] Wang R. AdaBoost for feature selection, classification and its relation with SVM, a review [J]. *Physics Procedia*, 2012, 25: 800-807.
- [9] Bai Y, Xie J, Wang D, et al. A manufacturing quality prediction model based on AdaBoost-LSTM with rough knowledge [J]. *Computers & Industrial Engineering*, 2021, 155: 107227.
- [10] Ogunseye E O, Adenusi C A, Nwanakwaugwu A C, et al. Predictive analysis of mental health conditions using AdaBoost algorithm [J]. *ParadigmPlus*, 2022, 3(2): 11-26.
- [11] Wang F, Li Z, He F, et al. Feature learning viewpoint of AdaBoost and a new algorithm [J]. *IEEE Access*, 2019, 7: 149890-149899.
- [12] Wang F, Li Z, He F, et al. Feature learning viewpoint of AdaBoost and a new algorithm [J]. *IEEE Access*, 2019, 7: 149890-149899.