

A study on the influence factors of athletes' momentum based on RF model and BP neural network

Rongqi Lan ^{1, *, #}, Youzhen Chen ^{1, #}, Chengbin Lin ^{2, #}

¹ College of Science, Minzu University of China, Beijing, China, 100081

² School of Information Engineering, Minzu University of China, Beijing, China, 100081

* Corresponding author: 18846483638@163.com

#These authors contributed equally.

Abstract. Momentum is used in the dictionary to mean "the acquisition of power through motion or a series of events". In sports, a team or a player may feel that they have momentum in the sport they are competing in. Athletes with "momentum" are often considered to have an advantage. To study the influencing factors of athletes' momentum, this paper uses the data of the men's match of Wimbledon Open 2023, adopts the Random Forest Model to filter the features affecting momentum, and obtains the five most important factors affecting the momentum of tennis players; and adopts the BP neural network to fit the athletes' movement state, and predicts the fluctuation of the situation. Studying the characteristics of momentum influence, understanding the development of momentum and the timing of its formation can provide a clearer understanding of the reality of the athlete's situation, which can allow the athlete and the coach to analyze and train in a more targeted way to achieve the effect of improving sports performance.

Keywords: Momentum, Athlete Performance, Random Forest, BP Neural Network.

1. Introduction

Momentum, also known as the athletic state, is affected by many factors, including the athlete's mentality, personal technique, physical fitness, and so on. Since the theory of athletes' competitive state was proposed by the Soviet scholar Matveyev in 1964, the competitive state has been used as an important basis for athletes' preparation for training and competition and for judging sports performance [1]. The study of athletes' competitive state has always been a hotspot of attention in the field of sports research, and exploring the intrinsic connection between momentum and athletic performance can help us understand the reality of athletes in a deeper way, to promote the development of sports training practice [2].

There have been several literature studies on the relationship between momentum state and athletes' performance as well as the change cycle of athletes' momentum [3]. According to Xiong Yan, momentum is the state of preparation and reality for athletes to participate in training and competitions, and the most appropriate state of preparation for athletes to create excellent performance is the optimal competitive state, which is the highest stage of the competitive state [4]. According to Prof Tian Maijiu, momentum is the appropriate state of readiness for athletes to create superior athletic performance [5]. The athletic state determines the system of athletic ability and is a reflection of the structure of athletic ability [6]. The athletic state is a cyclical change, week after week spiral enhancement, any training is to make the athlete in the competition period achieve the best athletic state, to strive for excellent performance, the enhancement of the athletic state is closely related to the monitoring of the body's biochemical indicators, the scientific design of the training program.

However, there is a lack of research in the direction of the factors affecting momentum. In this regard, this paper adopts the random forest model and BP neural network to mainly study the influencing factors affecting the momentum of tennis players. The random forest model can handle a large number of data features and effectively obtain the importance of the influencing factors [8], and the BP neural network can continuously optimize the model parameters through the back-propagation algorithm to improve the accuracy of the model [7]. Combining the results of the two models further

improves the reliability and stability of the research results [9]. Therefore, the use of the BP neural network and random forest model for the exploration of influencing factors has high analytical efficiency and accuracy, which helps to deeply understand the law and causal relationship behind the problem [10].

2. The basic fundamental of RF Model BP neural network

2.1. Random Forest Model

The random forest model can filter a data set containing many features and filter out the features that have a greater impact on the labels. In this paper, we use the random forest model to filter the most important key features that affect the momentum change during the player's game, get the ranking of the importance of the features, and get the factors that affect the player's performance during the game based on the results of the model.

In this model, we selected the features that affect the player's momentum change during the race, p1_distance, p1_games, etc. The features and their related meanings are shown in Table 1:

Table 1. Characteristics and what it means

notation	hidden meaning
p1_distance	Total distance run in the game
p1_games	Number of discs won by the player
p1_missed	Number of missed opportunities to win this game
p1_unf_err	Number of unforced errors by players
p1_sets	Number of games won by the player in the set
p1_scores	Runs scored by the player in the current innings
p1_ser	Whether or not the player serves in this round
p1_ser_ace	percentage of first-pass points
p1_ser_err	Number of service errors in the game
p1_net_won	percentage of points at the net (volleyball, tennis, etc)

The formula for the Gini index is:

$$G_m = \sum_{k=1}^K p_{mk}(1 - p_{mk}) \quad (1)$$

Where G_m is the m th feature of the Gini index, where K is the number of categories in the sample set, which is 2 in our tennis match model, and p_{mk} is the node m belongs to the k th class of the probability estimate.

The amount of change in the Gini index before and after node branching is: the

$$VIM_{jm} = GI_m - GI_l - GI_r \quad (2)$$

Where VIM is Variable Importance Measure for each feature. When we have m features, where VIM_j is the importance score of the j th feature, which represents the average change in the impurity of node splitting of the j th feature in all decision trees of the random forest. Where GI_l and GI_r are the average change in the impurity of node splits in all decision trees in the random forest for the j th feature m of the two new nodes split by the node Gini indices.

Suppose the variable appears in the i tree M times, then the importance of the variable in the first tree is of importance:

$$VIM_{ij} = \sum_{m=1}^M VIM_{jm} \quad (3)$$

The variable in the random forest Gini Importance is defined as:

$$VIM_j = \frac{1}{n} \sum_{i=1}^n VIM_{ij} \quad (4)$$

Where n is the number of classification trees in the random forest.

During the feature importance screening process of the random forest model, the Gini A higher importance indicates that the variable contains a richer amount of information, is better able to differentiate between different categories of samples, and has an extremely important contribution to the prediction of the target variable.

2.2. BP Neural and Network Model

BP neural network is a common artificial neural network model used for tasks such as pattern recognition, classification, and regression analysis. The principle is as follows: first, there is forward propagation, which passes the input data through each layer of the network, from the input layer to the output layer; then there is backpropagation of the error, which propagates the error forward from the output layer to the hidden and input layers by calculating the error in the output layer. The backpropagation uses a gradient descent algorithm to adjust the weights and biases in the network to minimize the error; followed by weight updating, which uses a gradient descent algorithm to update the weights and biases in the network based on the error calculated from the backpropagation. The training process of the BP neural network is an iterative process, whereby the weights and biases are continually adjusted to approximate the objective function to classify or regress input data for Prediction.

The mathematical principle is as follows; a three-layer BP neural network is shown in Figure 1:

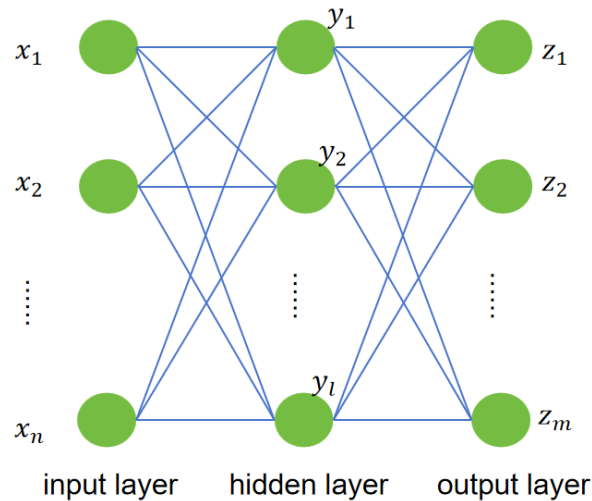


Figure 1. Neural network structure

In the BP neural network structure in the above figure, the input layer has a total of n nodes, the hidden layer consists of a total of l nodes, and the output layer uses m nodes. Assume that the weights from the input layer to the hidden layer are w_{ij} and the bias value is a_j and the weight of the hidden layer to the output layer is w_{jk} and the bias value is b_j . The learning rate of this neural network is η and the excitation function is $g(x)$ is the Sigmoid function, then

$$g(x) = \frac{1}{1+e^{-x}} \quad (5)$$

In the neural network shown above, after the feature values of the samples have entered the hidden layer from the input layer, the hidden layer outputs are

$$H_j = g(\sum_{i=1}^n w_{ij}x_i + a_i) \quad (6)$$

After passing the output of the hidden layer through the output layer the output is obtained as

$$O_k = \sum_{j=1}^l H_j w_{jk} + b_k \quad (7)$$

In general, the error formula for a neural network is

$$E = \frac{1}{2} \sum_{k=1}^m (Y_k - O_k)^2 \quad (8)$$

In each iteration of training, the parameters of each neural network layer are updated according to the following formula:

$$w_{ij}(t) = w_{ij}(t-1) + \eta H_j(1-H_j)x_i \sum_{k=1}^m w_{jk}(t-1)e_k \quad (9)$$

$$W_{jk}(t) = e_{jk}(t-1) + \eta H_j e_k \quad (10)$$

After many training iterations, the parameters of the neural network will iteratively converge to the optimum, after which the neural network will be able to make predictions on the test set.

In our work, we take each round of the opponent's data as a sample and use data from multiple matches for training and prediction to obtain a BP neural network model that predicts a player's match momentum and predicts the overall win/loss of a match-by-match characteristics.

3. Results

3.1. Data pre-processing

Outlier handling: duplicate data and missing values are removed to ensure data integrity and accuracy. Secondly, we process individual abnormal data to improve the robustness and stability of the model.

Feature Generation: The raw data is accumulated and calculated to generate the required features, such as the current number of missed balls, the rate of scoring at the net, the rate of scoring on one shot, etc.

Data normalization: scaling the data to a uniform range reduces the bias between features, improves the convergence speed of the algorithm, and simplifies the training process.

3.2. Momentum equation results

The dictionary defines momentum as "the force acquired through motion or a series of events". In sports, momentum affects the performance of athletes, and in tennis, the player who serves has a higher likelihood of winning a point. Therefore, let the influence factor of the serve on the score be k ($0 < k < 1$), the momentum be p , and the scores of the two players in the course of the match be α_i and β_i respectively. Let

$$\vartheta = \begin{cases} 0 & \text{not the server} \\ 1 & \text{be the server} \end{cases} \quad (11)$$

Define momentum to be determined by the combination of the number of consecutive points scored and whether or not the ball is on the serving side. Then the momentum equation can be expressed as:

$$\begin{cases} p_1(t) = \sum \alpha_i(t) + \vartheta k \\ p_2(t) = \sum \beta_i(t) + \vartheta k \end{cases} \quad (12)$$

3.3. Random Forest modelling results

To study the importance of various factors of "momentum", this paper establishes a random forest model. By analyzing the game data, we have filtered the features that affect the "momentum" of the players. In Random Forest, feature filtering is achieved by comparing the importance of features, so the model can be used to evaluate the importance of features, and thus determine the degree of influence of each feature on "momentum".

In this way, we filter the characteristics that influence the player's "momentum" and obtain the importance of each influence factor as shown in Figure 2:

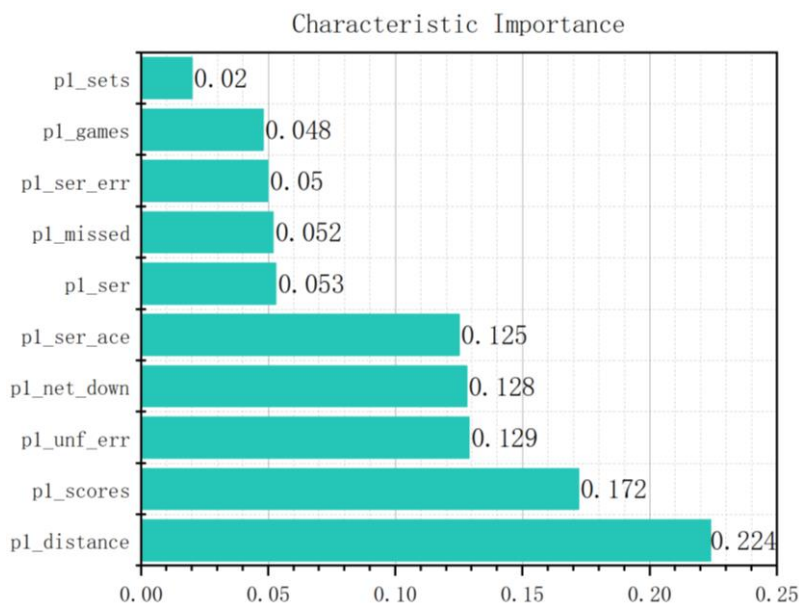


Figure 2. Random forest model results

From the figure, it can be seen that the factors affecting the fluctuation of the game in tennis are in the following order of importance: p1_distance>p1_scores>p1_unf_err>p1_net_pt_won>p1_ser_ace>p1_ser>p1_break_pt_ missed>p1_ser_err>p1_games>p1_sets. Therefore, physical fitness is the factor that has the greatest impact on the performance of the players, and factors related to mentality such as points scored, the total number of balls missed, and the percentage of points scored at the net also have a significant impact on the situation of the game.

3.4. BP Neural Network Results

In this paper, to investigate the influence of "momentum" in the game, principal component analysis was conducted and new features were generated, and the first five principal components were retained based on the principle that the cumulative contribution was greater than 80%.

With these five principal components, we constructed a neural network model with five nodes corresponding to the five principal components in the input layer; in the hidden layer 1, we set 16 nodes and extracted the features from the five principal components in the input layer; in the hidden layer 2, we set 16 nodes and nonlinearly mapped the features extracted in the first hidden layer; in the output layer, since the our task is a binary classification task, we set the number of nodes in the output layer to 1 and use the Sigmoid function to map the output values to the range [0,1] to classify the problem.

During the process of performing network training, we record the change in the accuracy and loss values of the training set as the number of training sessions increases, as shown in Figure 3 below:

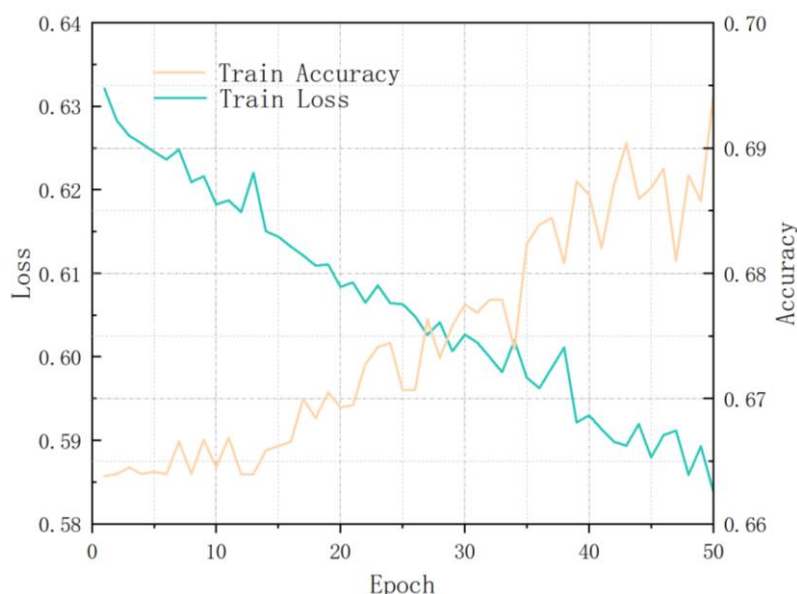


Figure 3. Change in loss value, accuracy

As can be seen from the above figure, with the increase in the number of training times, the loss value of the BP network model on the training set gradually decreases from 0.63 to near 0.58, and the accuracy of the model gradually increases from 0.66 to 0.7, which shows that the model's degree of fit to the results of the game is constantly improving.

The test set is then substituted and the model fit is examined for comparison with the support vector machine model, and the results of the model evaluation on the test set are obtained as shown in Table 2.

Table 2. Results of the model assessment

	BP	SVM	KNN
accuracy	0.655	0.650	0.638
precision	0.655	0.650	0.695
recall	0.974	1.000	0.789
f1	0.844	0.788	0.739

As can be seen from Table 2, the accuracy and precision of the three models are all greater than 0.6, and the recall is greater than 0.9, indicating that the feature selection is in line with reality, and the selected indicators can help determine the changes in the match situation. The f1 value of the BP neural network model is higher than the other models, and the other indicators generally perform better than SVM, KNN, and other models, indicating that the BP neural network model is suitable for tennis match situation prediction with a high degree of fit.

4. Conclusion

This paper establishes the momentum equation and constructs a random forest model to analyze the importance of the factors affecting the momentum, and concludes that the most important factors affecting the momentum of a tennis player are the total running distance, the current set score, the number of errors, the rate of the net score, and the rate of the first serve equal points. In addition, this paper establishes the BP neural network to predict the player's performance, and compares the results with SVM, KNN, and other models, and concludes that the BP neural network model has a better prediction effect. It seems that the athlete's physical ability and whether to seize the first opportunity, priority advantage and other mental factors have a great influence on the momentum of the athlete. In the usual training time, we can pay more attention to training athletes' physical fitness and mental quality, so that they can calmly deal with the scores whether they are at a disadvantage or an advantage in the process of the competition.

There are some shortcomings in this study. Firstly, comparisons between samples from different sites were missing as the study was conducted on a single-site sample only. Future studies may consider increasing the sample size and introducing more influencing factors to improve the reliability and generalizability of the findings.

References

- [1] Yang Z, Ke P, Zhang Y, et al. Quantitative analysis of the dominant external factors influencing elite speed Skaters' performance using BP neural network [J]. *Frontiers in Sports and Active Living*, 2024, 61227785 - 1227785.
- [2] Oliveira P D T, Newell J. A hierarchical approach for evaluating athlete performance with an application in elite basketball. [J]. *Scientific reports*, 2024, 14 (1): 1717 - 1717.
- [3] Vogel K, Larsen B, McLellan C, et al. Female Athletes and the Menstrual Cycle in Team Sports: Current State of Play and Considerations for Future Research [J]. *Sports*, 2023, 12 (1).
- [4] Oliveira P D T, Newell J. A hierarchical approach for evaluating athlete performance with an application in elite basketball. [J]. *Scientific reports*, 2024, 14 (1): 1717 - 1717.
- [5] Yuta K, Toru I, Keita K. Balancing academics and athletics: school-level athletes' results are positively associated with their academic performance. [J]. *Trends in neuroscience and education*, 2023, 33100210 - 100210.
- [6] Cintineo P H, Bello L M, Walker J A, et al. Monitoring training, performance, biomarkers, and psychological state throughout a competitive season: a case study of a triathlete. [J]. *European journal of applied physiology*, 2024,
- [7] Muñoz M S A, Alvis G I Á, Martínez B F I. A random forest model to predict soil organic carbon storage in mangroves from Southern Colombian Pacific coast [J]. *Estuarine, Coastal and Shelf Science*, 2024, 299108674-.
- [8] Kiehas F, Reiter M, Torres P J, et al. Predicting Ductile-Brittle transition temperatures for polyolefins using convolutional neural networks and instrumented notched Charpy experiments [J]. *Polymer*, 2024, 296126797-.
- [9] Muñoz M S A, Alvis G I Á, Martínez B F I. A random forest model to predict soil organic carbon storage in mangroves from Southern Colombian Pacific coast [J]. *Estuarine, Coastal and Shelf Science*, 2024, 299108674-.
- [10] David W, Balaji R, Edith Z. Long-Lead Forecasting of Runoff Season Flows in the Colorado River Basin Using a Random Forest Approach [J]. *Journal of Water Resources Planning and Management*, 2024, 150 (4).