ISSN 2616-5775 Vol. 7, Issue 4: 103-109, DOI: 10.25236/AJCIS.2024.070414

Research on Tennis Match Strategies Based on LightGBM and BP Neural Network Models

Jia Yang[1,*]

[1]*College of Science, Wuhan University of Science and Technology, Wuhan, China*
[*]*Corresponding author: 495402166@qq.com*

Abstract: *This paper focuses on the influence of "momentum" in tennis and analyzes the scoring pattern of athletes using LightGBM model and BP neural network model. First, the selected influencing factors were subjected to PCA dimensionality reduction. Then, we built the LightGBM model to evaluate the real-time performance of the players. Secondly, we quantified "momentum" using the entropy weighting method and confirmed the importance of "momentum" in the game by comparing the cumulative "momentum" with the actual win/loss situation and conducting the chi-square test. The importance of "momentum" in the game was confirmed by comparing the cumulative "momentum" with the actual winning and losing situations and conducting chi-square test. Finally, a BP neural network model was developed to predict momentum fluctuations and to identify the factors associated with these fluctuations.*

Keywords: *Momentum, LightGBM, Bayesian Variable Point Detection, BP Neural Networks*

1. Introduction

In sports competitions, it is often seen that a team or a player performs strongly at a certain stage of the game, and this phenomenon is called "momentum". Momentum plays an extremely important role in tennis matches, and sometimes it can influence the development of the whole match. Through systematic training and timely change of match strategy and tactics, the winning rate of the match can be effectively improved [1]. How to build up the momentum and give full play to its positive effect is a problem that needs to be considered in sports training, which can help players better cope with high-intensity and high-pressure matches.

In this paper, based on the men's singles dataset of 2023 Wimbledon Open, we firstly analyzed the "momentum" and selected 12 relevant factors that may affect the "momentum". Due to the excessive number of factors, we used PCA to reduce the dimensionality and constructed a LightGBM model to evaluate the real-time performance of players. Then, we quantified the "momentum" using entropy weighting method to obtain the specific performance of "momentum". By comparing the cumulative "momentum" of players during the game with their actual win/loss situations, and conducting a chi-square test, we determined that "momentum" plays a significant role in the game. Finally, we interpret "the flow of the game from favoring one player to another" as a turn in the player's performance curve. Using Bayesian Variable Point Detection and visualization techniques, we can clearly identify these turns in the game. We then build a BP neural network model to predict the fluctuations in the game and ultimately identify the factors correlated with these fluctuations. This provides valuable information and insights to athletes, coaches, and administrators to improve the competitiveness and management of the game.

2. Data processing and LightGBM modeling

2.1 Data processing

Data missing processing: if a large amount of data is missing in the same contest, it is deleted to ensure the accuracy of subsequent modeling; if some data is missing in the same contest, Lagrange interpolation is used to process the missing values. Lagrange interpolation is a method to estimate the value at a given point by constructing a polynomial from a known data point. The Lagrange interpolation polynomial has the following form:

$$L(x) = \Sigma_{i=0}^{n} y_i \Pi_{j=0, j\neq i}^{n} \frac{x-x_j}{x_i-x_j} \quad (1)$$

ISSN 2616-5775 Vol. 7, Issue 4: 103-109, DOI: 10.25236/AJCIS.2024.070414

Data outlier processing: when dealing with outliers, quartiles and IQR(Interquartile Range) can be used for identification and processing. IQR is used as follows:

First, Q_1(upper quartile) and Q_3(lower quartile) are obtained, then IQR=$Q_3 - Q_1$ is calculated, and then the upper and lower boundary of the threshold is defined. The calculation result is as follows:

$$IQR = Q_3 - Q_1 \tag{2}$$

Then we define the lower and upper bounds of the threshold, which are computed as

$$Lower\ Bound = Q_1 - 1.5 * IQR \tag{3}$$

$$Upper\ Bound = Q_3 + 1.5 * IQR \tag{4}$$

Next, if the data is less than $Lower\ Bound$ or more than $Upper\ Bound$, mark it as an exception. The results before and after processing are shown in Figure 1.

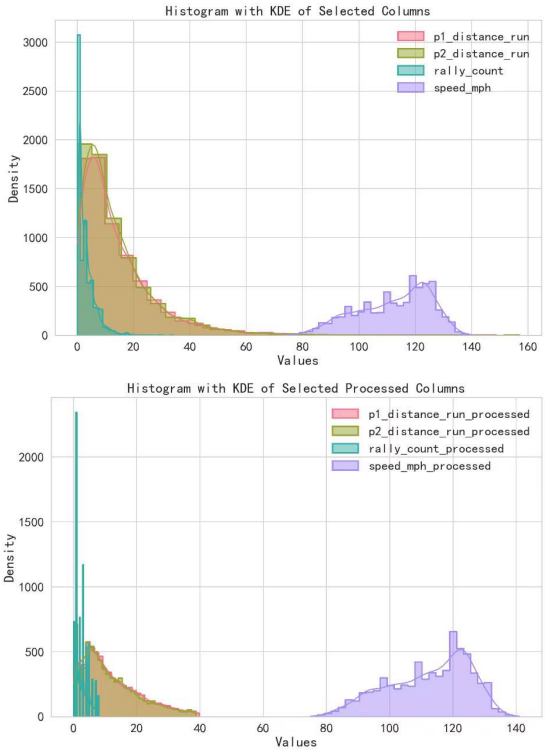

Figure 1: Comparison of outliers before and after processing

Data normalization: in order to reduce the adverse impact of sample data singularity, the data is limited to [0,1]. For the bigger-is-better indicator, it is normalized to:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5}$$

Where x_{min} is the minimum value, x_{max} is the maximum value, and $x_{normalized}$ is the standardized data.

2.2 Feature extraction and dimensionality reduction

In this paper, 12 initial related indexes were selected from three perspectives: individual technical ability, player fatigue and real-time mentality. Finally, we decided to perform PCA (principal component analysis) to reduce the dimensionality of the indicators, which is a multivariate statistical method to replace the original indicators by transforming the original indicators with certain correlation into a set of comprehensive indicators unrelated to each other through dimensionality reduction technique [2]. The obtained PCA downscaling scree plot is shown in the Figure 2.

ISSN 2616-5775 Vol. 7, Issue 4: 103-109, DOI: 10.25236/AJCIS.2024.070414

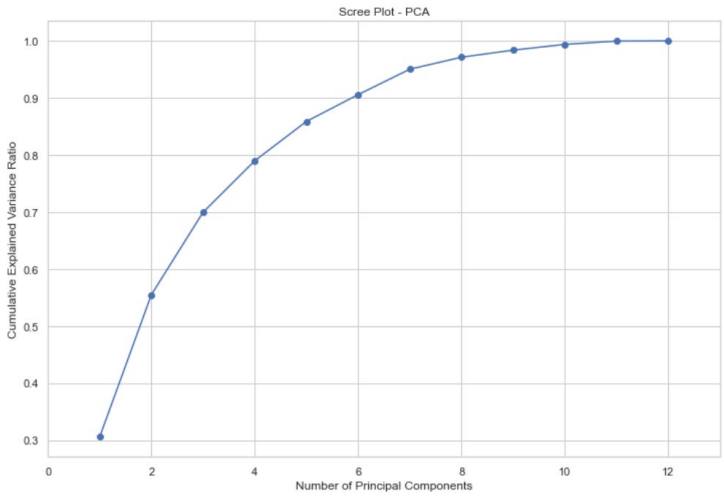

Figure 2: Indicator's scree plot

We can see that 6 principal components can account for more than 90% of the data, so in order to reduce the complexity of the modeling process, we can reduce the data to 6 dimensions.

2.3 LightGBM model building and solving

LightGBM is a new GBDT (Gradient Boosting Decision Tree) algorithm proposed by Ke [3] in 2017. GBDT has the functional characteristics of gradient boosting and the decision tree has the advantages of good training effect and is not easy to overfit.

We use the above data and input feature sequences to train the model with a ratio of the training set to test set of 4:1.To determine the reliability of the LightGBM model, this paper applies the knowledge of statistics to analyze the fit evaluation of the Logistic Regression, LightGBM, Random Forest, and SVM models using the AUC (Area Under the Curve), ACC (accuracy), REC (Recall), PRE (Precision) and F1 (score) are used to measure the area under the ROC curve to evaluate the overall performance of these models. A comparison of the four model fits is shown in Figure 3.

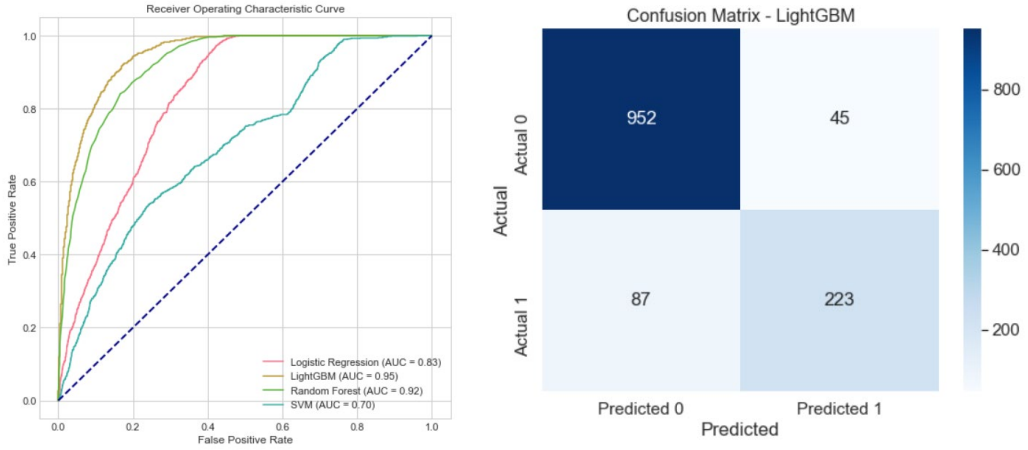

Figure 3: Evaluation of the fit of the LightGBM model

From the above model fitting, it is easy to see that the LightGBM model has the best fitting effect. Therefore, this model can be used to simulate the real-time "momentum" of athlete 1 and athlete 2 in the race. The fitting results are shown in Figure 4.

ISSN 2616-5775 Vol. 7, Issue 4: 103-109, DOI: 10.25236/AJCIS.2024.070414

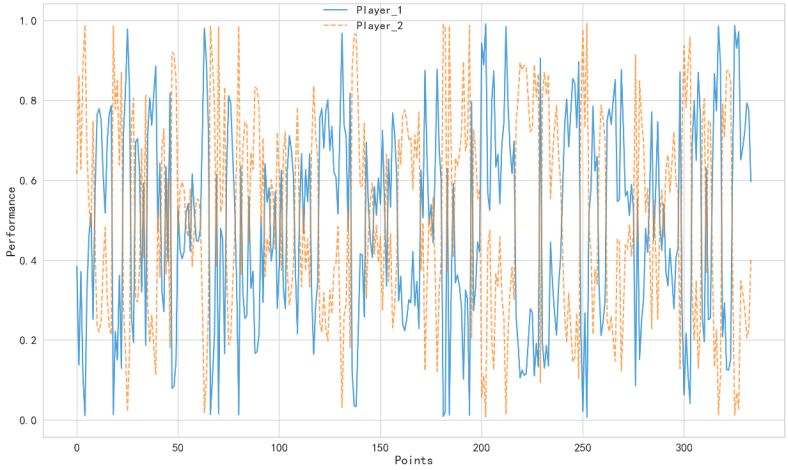

Figure 4: Classic matchups real time chart

Obviously through the charts, we can clearly see that at the beginning of the match, Player_2 had the upper hand, scoring more points than Player_1, but at the end of the match, Player_1 outscored Player_2, and the above charts also depict the trend of the classic match in which the rising star, Carlos Alcaraz, defeated Novak Djokovic.

3. Construction and testing of momentum model

3.1 Momentum modeling

To truly determine whether momentum plays a role in a match, this paper quantifies momentum by visualizing data, comparing it to the win-loss ratio of the match outcome, and calculating the correlation between momentum and the win-loss ratio of each match for analysis.

Firstly, the entropy weight method is used to determine the index weight.

After standardization, we calculated the weight of the $i-th$ sample value under the $j-th$ indicator in relation to the total indicator as:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \tag{6}$$

The entropy value of the $j-th$ indicator is calculated as:

$$e_j = \frac{1}{\ln n} \sum_{i=1}^{n} P_{ij} \ln\left(P_{ij}\right) \tag{7}$$

Calculate information entropy redundancy:

$$d_j = 1 - e_j \tag{8}$$

The weights of the indicators were calculated as:

$$W_i = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{9}$$

According to the results of the weighting of the above indicators, we can quantify the "momentum" by weighting the indicators. Momentum" is large, which means the player's performance is good, and "momentum" is small, which means the player's performance is not so good.

3.2 Randomness test

Next, we calculated the cumulative momentum for a total of 31 matches from 1301-1701, And after calculation, we have predicted 31 matches by using players' "momentum", and got 27 correctly verified matches, while 4 failed to be verified, with a correct rate of 87.1%. Therefore, in order to further verify whether there is any relationship between players' "momentum" and success in a match, we introduced the chi-square test to determine whether the two are related.

The Chi-Square test is a statistical method used to test whether there is a correlation between two

ISSN 2616-5775 Vol. 7, Issue 4: 103-109, DOI: 10.25236/AJCIS.2024.070414

categorical variables [4]. It determines whether two variables are independent by comparing the difference between the observed frequencies and the expected frequencies. The chi-square test is commonly used to analyze the correlation between two categorical variables.

Where the statistic of chi-square test is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{10}$$

In this case, the original and alternative hypotheses of the chi-square test are: original hypothesis: H0: the player's "momentum" has nothing to do with winning the game. Alternative hypothesis: H1: player "momentum" is associated with game winning.

Choose the Significance Level α: this is the probability that we are willing to make a mistake before rejecting the null hypothesis, choose 0.05.

Table 1: Chi-Square test series

Comparison	Winner of the competition	Loser of the competition
Cumulative momentum greater than 0	17	1
Cumulative of momentum less than 0	3	10
Total	20	11

Calculated by the result in Table 1, we get the value of chi-square statistic is 13.821, p-value is 0.0002, obviously the p-value is very small, so we reject the original hypothesis. That is to say, it is sufficiently clear that the player's "momentum" in the game is going to affect the game trend.

4. Momentum prediction model based on BP neural network

4.1 BP Neural Network prediction model based on Bayesian variable point detection

The Bayesian Change Point Detection algorithm can detect change points in the data stream in real-time [5], detecting changes in the data distribution promptly, and is suitable for finding the inflection points of fluctuations during each game.

First, we divide the game performance of the two players in the game into $x_1, x_2 \ldots x_T$ non-overlapping product partitions. The divisions between partitions are called change points. For each partition p, the data in it is from some probability distribution $p(x_t|\eta_p)$.We denote the set of consecutive observations between score point a and score point b as $x_{a:b}$.

From a given variable length γ_t, using edge probability density summation, it follows that:

$$p(x_{t+1}|x_{1:t}) = \sum_{\gamma_t} p(x_{t+1}|r_t, x_t^r)\ p(r_t|x_{1:t}) \tag{11}$$

Therefore, the posterior probability is:

$$p(r_t|x_{1:t}) = \frac{p(r_t, x_{1:t})}{p(x_{1:t})} \tag{12}$$

Thus:

$$\begin{aligned} p(r_t \mid x_{1:t}) &= \sum_{\gamma_{t-1}} p(r_t, r_{t-1}, x_{1:t}) \\ &= \sum_{\gamma_{t-1}} p(r_t \mid x_{t-1}) p\left(x_t \mid r_t, x_{t-1} x_t^{(r)}\right) p(r_{t-1}, x_{1:t-1}) \end{aligned} \tag{13}$$

Where the predictive distribution $p(x_t|r_{t-1}, x_{1:t})$ depends only on the most recent data x.

Assuming that the value of the probability before each sequence is known, the probability of it depends on two possibilities, i.e., the change point occurs or the change point does not occur when the length of the trip is added to 1. To make a judgment about these two possibilities and to calculate the probability in both cases, it is necessary to compute the Hazard Function.

$$p(r_t|r_{t+1}) = \begin{cases} H(r_{t-1}+1) & if r_t = 0 \\ 1 - H(r_{t-1}+1) & if r_t = r_{t-1}+1 \\ 0 & otherwise \end{cases} \tag{14}$$

Bayesian variable point detection was used to find the variance of performance scores between two players over the course of a game score in Figure 5.

ISSN 2616-5775 Vol. 7, Issue 4: 103-109, DOI: 10.25236/AJCIS.2024.070414

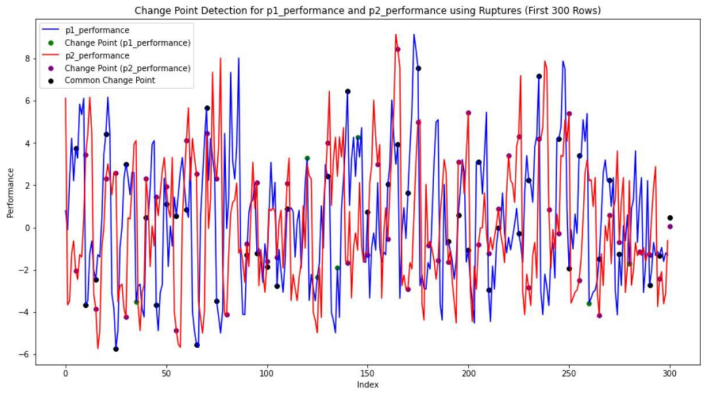

Figure 5: Changing points in the game

For the problem of predicting changes in momentum fluctuations, the data derived from 12 distinct characteristic equations form the input space, while the 12 momentum weight distributions constitute the output space. The essence of the problem lies in establishing a mapping from a 12-dimensional input space to a 12-dimensional output space. A survey of existing literature indicates that, for modeling this finite-dimensional space mapping, typically only one hidden layer is necessary for constructing a neural network [6,7]. Consequently, the BP neural network utilized in this model comprises three layers: input, hidden, and output layers. Given that both the input and output spaces are 12-dimensional, the number of neurons in both the input and output layers is set to 12. Following the empirical formulas provided in the literature, the number of neurons in the hidden layer is determined by equation:

$$N_h = \frac{N_s}{\alpha(N_i+N_o)}, \alpha \in [2,10] \tag{15}$$

So far, we have completed the construction of the BP neural network.

In the training of the BP neural network, we randomly selected the training set according to 85%. After several tests, we found that the neural network obtained from each training always has unacceptable errors at very few samples in the test set. Moreover, the repetition rate of samples with large prediction errors was not high for different neural networks. Therefore, we decided to adopt the Bagging algorithm to integrate multiple BP neural networks. Then, the final results are derived through a hard voting mechanism, which has the effect of reducing the generalization error of the prediction results.

We took 90% of the data as a training set to train the model and carried out BP neural network fitting of 12 indicators on the inflection point data detected by Bayesian variable points in 2023-wimbledon-1301 competition, and the fitting results were shown in Figure 6.

4.2 Model generalization ability test

In order to test the generalization ability of the model, we tested the prediction effect of the developed model in other data sets, and the results and predictions were shown in Figure 7.

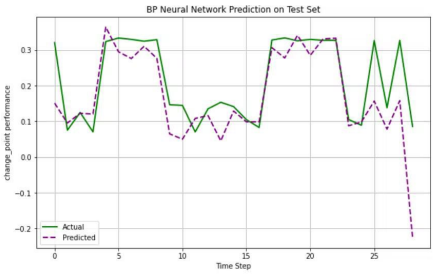

Figure 6: BP neural network prediction on test set

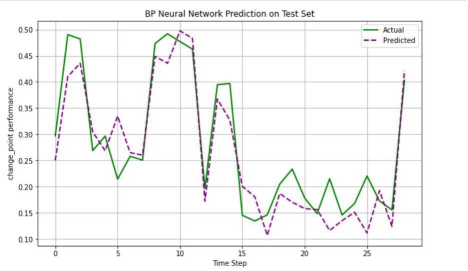

Figure 7: BP neural network prediction

We introduce the criteria for evaluating the model: MSE, RMSE, MAE, MAPE, where MSE (the expected value of the square of the difference between the predicted and the actual value), RMSE (the square root of the MSE), MAE (the mean of the absolute error), and MAPE (the mean absolute percentage error), where the lower the value of the above metrics, the higher the accuracy of the model. R^2 compares the predicted values to the mean only, the closer the result is to 1, the more accurate the model is.

ISSN 2616-5775 Vol. 7, Issue 4: 103-109, DOI: 10.25236/AJCIS.2024.070414

The above metrics are used to measure the prediction effect of the BP neural network. Among them, the evaluation metrics of the cross-validation set can continuously adjust the hyperparameters to obtain a reliable and stable model. The results of the measurement are shown in Table.2.

Table 2: Model evaluation

	MSE	RMSE	MAE	MAPE	R^2
Training set	1.971	1.404	1.124	289.824	0.878
Test set	10.289	2.36	2.162	535.158	0.515

The results in Table.2 shows that the model has a high prediction accuracy in similar scoring rule matches other than Wimbledon matches, which can indicate that the model has some general applicability in other matches such as Women's matches, tournaments, court surfaces, and other sports.

4.3 Sensitivity Test

To test the stability of the BP neural network prediction model based on Bayesian variable point detection, in solving the linear programming model, here we choose to float 6 feature works in the feature importance by 5%, and then use the perturbed data to make predictions, and compare with the original prediction results. By observing the changes in the prediction results, we observe whether the calculation results will appear large changes, using MATLAB to test the results are shown in Table 3.

Table 3: Sensitivity analysis table

Feature Engineering	S1	S7	S10	S4	S3	S5
Average predict impact	$1.32*10^{-3}$	$0.78*10^{-3}$	$0.53*10^{-3}$	$0.42*10^{-3}$	$0.51*10^{-4}$	$0.24*10^{-4}$

As can be seen from the Table.3, the Average predicted impact obtained from these feature engineering is less than 0.1%, which shows that the model is more stable.

5. Conclusions

Based on data from previous matches, this paper constructs mathematical models to assess athletes with advantage or momentum in a match and visualize how various events in a match create or change momentum. First, the PCA dimensionality reduction process was utilized for major factor selection, and a LightGBM model was built to assess player performance in real time and demonstrate the performance excellence time period. Second, the importance of momentum in the game was confirmed by quantifying momentum through entropy weighting method and calculating the weighted average. The key role of momentum in the game was verified by comparing the theoretical calculation results of the momentum model with the actual win/loss situation and performing the chi-square test. Finally, the momentum fluctuation prediction is carried out by using the BP neural network model, and the turn-turns in the game are identified by Bayesian variable point detection and visualization techniques, and the simulation test illustrates that the model has good generalization ability and robustness. Thus, it provides athletes, coaches and managers with game strategies and tactics to help players better cope with high-intensity and high-pressure games.

References

[1] Dietl, Helmut, and Cornel Nesseler. "Momentum in tennis: Controlling the match." UZH Business Working Paper Series, 365 (2017).
[2] Li, Tao, et al. "A PCA-based method for construction of composite sustainability indicators." The International Journal of Life Cycle Assessment, 17 (2012): 593-603.
[3] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems, 30 (2017).
[4] Vierra, Andrew, Abdul Razzaq, and Athena Andreadis. "Categorical Variable Analyses: Chi-square, Fisher Exact, and Mantel–Haenszel." Translational Surgery. Academic Press, (2023). 171-175.
[5] Nitzan, Eyal, Topi Halme, and Visa Koivunen. "Bayesian methods for multiple change-point detection with reduced communication." IEEE Transactions on Signal Processing, 68 (2020): 4871-4886.
[6] Merk, Timon, et al. "Machine learning based brain signal decoding for intelligent adaptive deep brain stimulation." Experimental Neurology, 351 (2022): 113993.
[7] Benth, Fred Espen, Nils Detering, and Luca Galimberti. "Neural networks in Fréchet spaces." Annals of Mathematics and Artificial Intelligence, 91.1 (2023): 75-103.