Research on Predicting Momentum Changes in Tennis Matches Based on Markov and Random Forest Algorithms

Shuqian Han[1,*], Alia Shayahaxi[2] and Chengzhi Yu[2]

[1]School of Mathematics and Statistics, Jishou University, Hunan, China

[2]School of Communication and Electronic Engineering, Jishou University, Hunan, China

[2]School of Computer Science and Engineering, Jishou University, Hunan, China

Abstract. Nowadays, momentum is widely mentioned in sports competitions and is regarded as an important psychological and emotional state, but its essence and impact on the results of the game are difficult to accurately quantify and analyze. Therefore, studying the momentum effect in sports competitions and predicting momentum changes is particularly important. We first evaluate the momentum effect in the tennis match by calculating the state transfer matrix of the Markov chain about the tennis score. Using the data of 31 matches about 2023 Wimbledon men's singles tennis match, we calculated that the server indeed has a higher probability of winning the serve game. In order to simplify the model, assuming that the transfer probability of the momentum state is equal to the transfer probability of the score state, we have established a momentum model that can capture the progress of tennis matches.Then, we have carried out a series of preprocessing of data, such as One hot encoding. Use a series of indicators of two players as independent variables to build a random forest model. The model has RMSE of 0.15, MAPE is 0.18, and R^2 is 0.2. Then use the SHAP algorithm to find out the factors most relevant to the momentum of the game.

Keywords: Momentum; Markov chain; Random Forest; Machine learning.

1. Introduction

Momentum typically refers to the force in physics that maintains the motion of an object, used to describe the state of motion and inertia of an object. With the development of psychology, especially in the study of cognition and emotions, the idea of momentum has been introduced to describe changes in psychological states, known as psychological momentum[1,2].

As psychologists began to study psychological factors such as confidence, anxiety, and concentration among athletes during competitions[3], the concept of momentum gradually entered the realm of sports competition, used to describe the state and emotional changes of athletes or teams in a game. In different types of sports, the measurement of momentum varies[4]. Figure 1 shows several key factors of psychological momentum in sports competition, which together shape the psychological state and momentum of athletes in games, impacting their competitive performance[5,6].

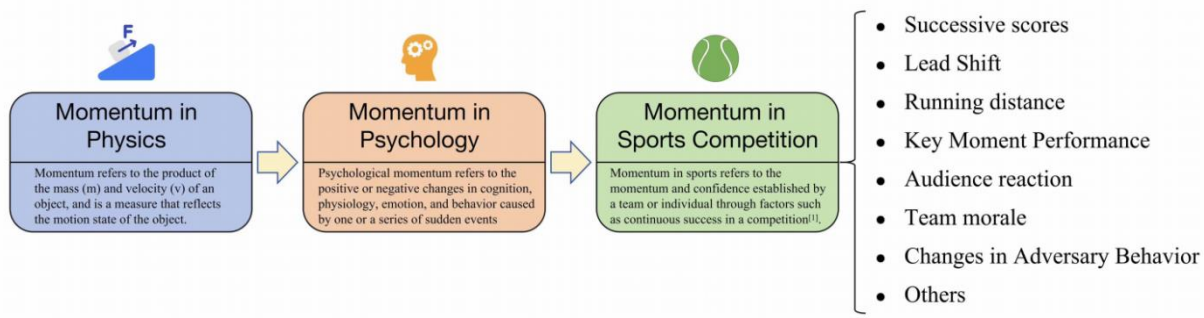

Fig. 1 Definition of momentum in different areas

In this paper, we focuse on the study of momentum changes in sports competitions, utilizing Markov chains to assess and analyze the effect of momentum, and conducting randomness checks. Based on machine learning models, it further predicts changes in game momentum and identifies the

main factors affecting game momentum, to assist coaches and athletes in better understanding and utilizing momentum during competitions.

2. Analysis of Momentum Effects

2.1. Markov model

Markov model is a stochastic model that undergoes transitions from one state to another in a probabilistic manner. The key idea behind a Markov model is the Markov property, which states that the future state of the system depends only on its current state and is independent of how it arrived at that state. A Markov chain consists of the following elements:

The mathematical formula for a Markov chain model is as follows:

Let $\{X_n, n \geq 0\}$ be a stochastic process with a state space S. If for any $n \in N$ and $i_0, i_1, \cdots i_{n+1} \in S$, the following condition holds:

$$P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n) \quad (1)$$

2.2. Analysis of Momentum Effects Based on Markov Model

In this paper, we use a Markov chain to assess whether there is a momentum effect in tennis matches. We define the state space S = $\{S_1, S_2\}$, where S_1 represents Player 1 winning the point and S_2 represents Player 2 winning the point.

Assuming that momentum state is equivalent to score state, we establish the momentum model as follows.

3. State Representation

Let S_t represent the state at time t, where $S_t = 0$ indicates a tie or neutral momentum, $S_t > 0$indicates the home team leading in score and positive momentum, and $S_t < 0$ indicates the away team leading in score and negative momentum.

4. Transition Probabilities

Let $p_{i,j}$ denote the probability of transitioning from state i to state j, where $i, j \in \{-n, -n+1, \ldots, n-1, n\}$and n represents the maximum point differential. Assuming the transition probabilities are related to the score differentials and momentum states as follows:

$$p_{i,j} = \begin{cases} \frac{1}{2(n+1)}, |i-j| = 1 \\ 0, \quad otherwise \end{cases} \quad (2)$$

This assumption implies that in each match, there are only two possible outcomes: either the home team scores or the away team scores. Therefore, score states and momentum states can only transition between adjacent states with equal probabilities.

5. Initial State

Assuming the initial state is $S_0 = 0$, representing a tie or neutral momentum.

6. Momentum Evaluation

Based on the transition probabilities and initial state, we can compute the state S_t at each time point. Let B_t denote the score at time t.

$$B_t = \sum_{i=1}^{t} W_i \cdot X_i, \quad (3)$$

Here, W_i represents the weight of the i-th score, and X_i represents the change in score for the i-th score. For a home team score $X_i = 1$, we can set the weight as $W_i = P_{ij}^i$; for an away team score $X_i = -1$, we can set the weight as $W_i = \left(1 - P_{ij}\right)^i$.

By using B_t and the current state S_t, we can calculate the momentum state M_t as:

$$M_t = \frac{2B_t}{t(t+1)} + S_t \tag{4}$$

This formula indicates that the momentum state is a weighted average of score changes over time, where the weight decreases as time progresses, and it includes the influence of the current state S_t.

Table 1. The transition probability matrix for player A' s consecutive wins

	Win	Scheme 2
Win	0.5426	0.4574
Lose	0.477	0.523

According to the output transition probability data, we can observe the following:

The transition probability from a winning streak state to another winning streak state is 0.5426, while the probability of transitioning from a winning streak state to a non-winning streak state is 0.4574. This implies that in a winning streak state, Player A has a slightly higher chance of continuing the winning streak than transitioning to a non-winning streak state.

The transition probability from a non-winning streak state to a winning streak state is 0.4770, while the probability of transitioning from a non-winning streak state to another nonwinning streak state is 0.5230. This indicates that in a non-winning streak state, Player A has a slightly lower probability of winning than continuing to lose points.

6.1. Evaluation of Randomness

The run test is a statistical testing method based on the number of runs formed by the arrangement of sample signs. It is mainly used to test whether the probability of an event occurring is random. The theoretical principle of the run test model is as follows:

Assumed r be the total number of runs, n1 and n2 represent the numbers of price increases and decreases, respectively. n is the sample size, i.e., n1+n2 = n. In the case of price changes showing random fluctuations, when n is large, r approximately follows a normal distribution. Let $u_r$be the expected value of the run count in a random sequence under the normal distribution, and σ_r^2 be the variance, calculated as follows:

$$u_r = E(r) = (2n_1n_2 + n)/n \tag{5}$$

$$\sigma_r^2 = 2n_1n_2(2n_1n_2 - n)/n^2(n-1) \tag{6}$$

Then, we can obtain the test statistic:

$$\mathrm{z} = \frac{\mathrm{r}-\mathrm{u_r}}{\sigma} \sim \mathrm{N}(0,1) \tag{7}$$

Given a significance level a and knowing the sample size n, we can determine the critical value $Z_{a/2}$ for the test statistic Z. If the obtained value of z from the sample is greater than $Z_{a/2}$, we reject the null hypothesis, indicating that the sequence is not random. Otherwise, we accept the null hypothesis, implying that the sequence is random.

Table 2. The run test for "swings in play" and "runs of success by one player"

match_id	p1_momentum_Rand	p2_momentum_Rand	p1_turning_points_Rand	p2_turning_points_Rand
2023-wimbledon-1301	1	1	0	0
2023-wimbledon-1302	1	1	0	0
2023-wimbledon-1303	1	1	0	0
2023-wimbledon-1304	1	1	0	1
2023-wimbledon-1305	1	1	0	1
2023-wimbledon-1306	1	1	0	0
2023-wimbledon-1307	1	1	0	0
2023-wimbledon-1308	1	1	1	0
2023-wimbledon-1309	1	1	1	1
2023-wimbledon-1310	1	1	0	1
2023-wimbledon-1311	1	1	0	0
2023-wimbledon-1312	1	1	0	0
2023-wimbledon-1313	1	1	0	1
2023-wimbledon-1314	1	1	0	0
2023-wimbledon-1315	1	1	1	0
2023-wimbledon-1316	1	1	0	0

2023-wimbledon-1401	1	1	0	0
2023-wimbledon-1402	1	1	1	1
2023-wimbledon-1403	1	1	0	0
2023-wimbledon-1404	1	1	1	1
2023-wimbledon-1405	1	1	1	0
2023-wimbledon-1406	1	1	1	0
2023-wimbledon-1407	1	1	0	0
2023-wimbledon-1408	1	1	0	0
2023-wimbledon-1501	1	1	0	0
2023-wimbledon-1502	1	1	1	0
2023-wimbledon-1503	1	1	1	0
2023-wimbledon-1504	1	1	1	0
2023-wimbledon-1601	1	1	0	0
2023-wimbledon-1602	1	1	1	0
2023-wimbledon-1701	1	1	1	1

Table 3. Run test statistic

statistic	p1_momentum_Rand	p2_momentum_Rand	p1_turning_points_Rand	p2_turning_points_Rand
count	31	31	31	31
mean	1	1	0.387097	0.258065
std	0	0	0.495138	0.444803
min	1	1	0	0
25%	1	1	0	0
50%	1	1	0	0
75%	1	1	1	0.5
max	1	1	1	1

Based on the Table3 and 4, it can be concluded that:

Analysis of momentum randomness: The momentum of both players in all matches shows non-randomness (value of 1 with no standard deviation). This implies that the changes in momentum for player 1 and player 2 in all matches are not random but influenced by certain systematic factors. These factors may include the players' skills, strategies, psychological state, and the opponent's strength.

Analysis of turning points randomness: The results indicate that for player 1, approximately 38.71% of the matches exhibit randomness in turning points, while for player 2, it is approximately 25.81%. This suggests that in a certain proportion of matches, the occurrence of turning points can be considered random and not influenced by specific systematic factors.

In conclusion, the non-randomness of momentum indicates the presence of identifiable patterns and trends in the matches, providing a basis for game preparation and strategy adjustments. The randomness of turning points signifies the uncertainty and complexity in the matches, highlighting the need for coaches and players to be prepared for unforeseen situations and changes.

7. Prediction model of game momentum change based

7.1. Data Preprocessing

One-hot encoding is a method for processing categorical data. It transforms categorical variables into a series of binary columns, each category represented by an independent column. If a category appears in an observation, the corresponding column is marked as 1; otherwise, it is 0. Using one-hot encoding, the values of discrete features are expanded into Euclidean space, where a particular value of a discrete feature corresponds to a point in Euclidean space.

For the categorical variables in the dataset, "p1_score", "p2_score", "winner_shot_type", "serve_width", "serve_depth", one-hot encoding needs to be applied.

Step 1: List Unique Categories. For each categorical variable, list all unique categories. This step is to determine how many new binary feature columns need to be created. For example, if a variable has three categories (A, B, C), then three new binary feature columns will be created for this variable, representing 'A', 'B', and 'C'.

Step 2: Create Binary Feature Columns. For each unique category, create a new binary feature column. If the record in the original data belongs to that category, mark it as 1 in the corresponding column; if it does not belong to that category, mark it as 0.

Step 3: Repeat the Above Steps. If there are multiple categorical variables in the dataset that need to be one-hot encoded, repeat the above steps until all categorical variables have been processed.

7.2. Evaluation Metrics

In order to test the prediction results of the predictive model, this paper adopts the root mean square error (RMSE), the average absolute percentage error (MAPE) and the coefficient of determination (R^2) as evaluation indicators for establishing a time-based economic forecasting model.

$$\mathrm{RMSE} = \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\right)^{1/2} \tag{8}$$

The above formula is the root mean square error (RMSE), which is expressed as the arithmetic square root of the mean sum of the squares of the predicted value and the true value error. When the difference between the predicted value and the actual value is larger, the RMSE greater it is. The higher the prediction accuracy of the model, the RMSE smaller it is.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \cdot 100\% \tag{9}$$

The above formula is the mean absolute percentage error (MAPE), which represents the mean of the ratio of the error of the predicted value to the true value. When the difference between the predicted value and the actual value is larger, the MAPE larger, the MAPE larger, the lower the accuracy, and vice versa. MAPE can reflect the relative error size of the prediction results, and is usually used to compare the prediction accuracy of the same prediction model at different time points.

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\bar{y}_i - y_i)^2} \tag{10}$$

The above formula is the determination coefficient R^2, which means that the sum of squares of the error is the ratio of the sum of squares of the regression to the sum of the total squares, and the range of values is [0,1], the closer to 1, the closer the sum of the squares of the error is to the sum of the total squares, that is, the better the prediction effect, and the first two evaluation indicators, the larger the value of the index, the higher the prediction accuracy. The coefficient of determination reflects the degree to which the predictive model fits the observed data, and is often used to compare the fitting effect of the predictive model on the same data set.

7.3. Prediction Model Based on Random Forest

3.3.1 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees for classification or regression. Each decision tree is built as a base learner and then combined through ensemble methods. Random Forest introduces randomness in the training process of decision trees, making it more resistant to overfitting and noise.

The randomness in Random Forest is reflected in two aspects:

1.Randomly selecting samples: The sample set of each decision tree in Random Forest is formed by randomly selecting and recombining n training samples from the original dataset using Bootstrap strategy with replacement. This means that the same sample can appear multiple times in the same subset or different subsets.

2.Randomly selecting features: Unlike single decision trees that consider all features to select the optimal feature for splitting nodes, Random Forest randomly examines a certain number of features in the base learners and selects the best feature from them. The randomness in feature selection improves the generalization and learning ability of the model.

The algorithm steps for Random Forest are as follows:

Extract training sets from the original sample set. Each round randomly selects n training samples with replacement using Bootstraping from the original sample set. Conduct k rounds of extraction to obtain k training sets. (The k training sets are independent of each other.)

Use each training set to obtain a model. k training sets yield k models.

For classification problems: use a voting method to combine the k models obtained above to obtain the classification result. For regression problems, calculate the mean of the models as the final result.

3.3.2 SHAP

SHAP (Shapley Additive explanations) is a method for interpreting machine learning model predictions. It is based on the concept of Shapley values from cooperative game theory, used to measure the contribution of each feature to the model's prediction. For each predicted sample, the model generates a predicted value, and SHAP value represents the assigned value for each feature in that sample.

We assume the i-th sample is X_i, the j-th feature of the i-th sample is X_{ij}, the predicted value of the model for that sample is y_i, and the baseline value of the entire model (usually the mean of the target variable for all samples) is y_{base}. Then, the SHAP value follows the following equation:

$$y_i = y_{base} + f(X_{i1}) + f(X_{i2}) + \ldots + f(X_{ik}) \quad (11)$$

Where $f(X_{ij})$ is the SHAP value for X_{ij}. Intuitively, $f(X_{i1})$ represents the contribution of the first feature in the i-th sample to the final predicted value y_i. When $f(X_{y1})>0$, it indicates that the feature enhances the predicted value and has a positive effect. On the other hand, if $f(X_{j1})<0$, it means that the feature decreases the predicted value and has a negative effect.

Traditional feature importance only tells us which features are important, but does not provide insight into how these features affect the prediction. The major advantage of SHAP values is that they reflect the influence of each feature on individual samples, and also indicate the positive or negative impact.

7.4. Model Predictions

In the experiment, we utilize the Random Forest algorithm to predict the changes in game momentum, and compared it with the Neural Network and Gradient Boosting Decision Tree (GBDT) models. The results of the Random Forest model evaluation are shown on figure 2. It shows that the Random Forest model outperforms in all three metrics. It has an RMSE value of 0.15, MAPE of 0.18, and an R^2 of 0.2, all of which are the lowest when compared with the Neural Network and GBDT models. The low RMSE and MAPE values indicate that the Random Forest model has smaller errors in prediction, and the higher R^2 value suggests that the model has a stronger ability to explain data variability.

In contrast, the Neural Network has an RMSE of 6.8 and a MAPE of 7.5, while the GBDT has an RMSE of 0.85 and a MAPE of 0.72, with their R^2 values being 0.65 and 0.2, respectively. Although GBDT's R^2 value is comparable to that of the Random Forest, its performance in error metrics is not as good. In summary, the Random Forest model we proposed shows clear advantages and effectiveness in terms of prediction accuracy and explanatory power. This means that in our experiment, the Random Forest model is more effective in accurately predicting and understanding the data compared to the other two models.

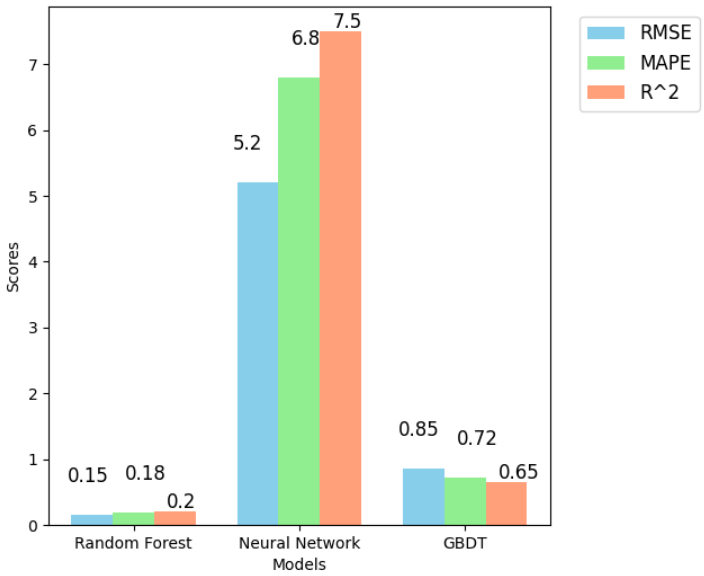

Fig. 2 model evaluation

The distribution plot of important features summarized by SHAP values is showed in figure 3. The two most important features are "p1_distance_run" and "p2_distance_run". This suggests that the distance covered by players on the court has a significant impact on the prediction. It may reflect the importance of endurance and mobility of players in the game.

Other relatively important features indicate that the serve speed, the score of both players, and the stage of the game have a significant influence on the model's prediction. Features such as "p1_break_pt", "p2_break_pt", and "set_victor" have SHAP values close to zero, indicating that they have almost no impact on the model's prediction.

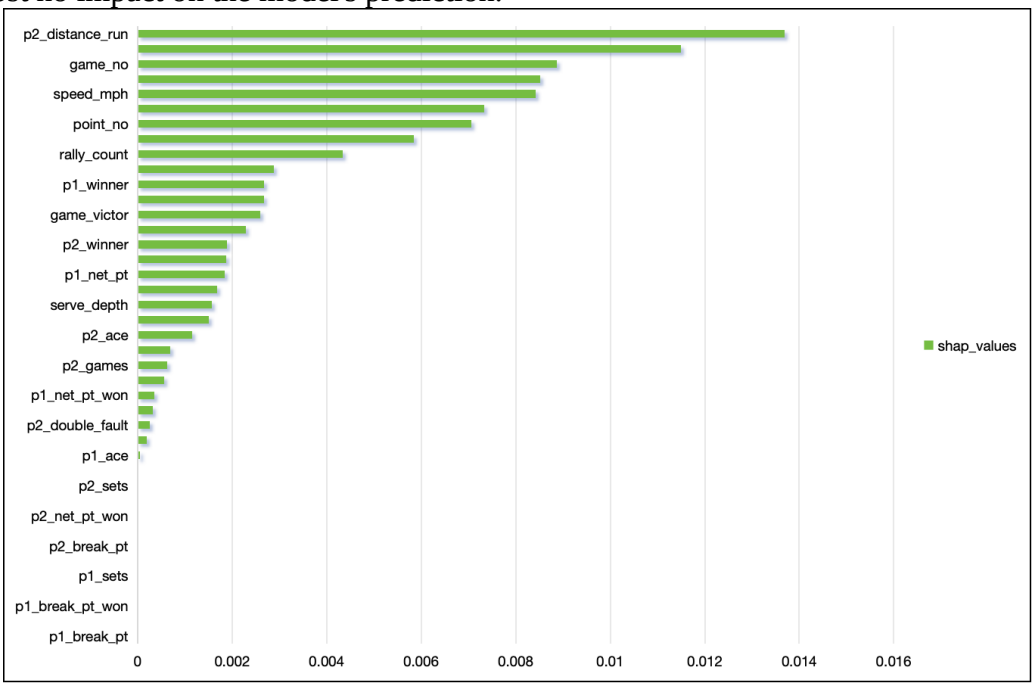

Fig. 3 The distribution plot of important features summarized by SHAP values

8. Summary

In this paper, the parameter estimation and inference for the Markov model we construct can be computed using efficient algorithms, which provide computational efficiency when dealing with large datasets. Additionally, we propose a machine learning-based model for predicting changes in game momentum. It combines accurate predictions and training with historical match data for making

reasonable forecasts. Compared to other models, our prediction model achieves higher accuracy. The momentum change prediction model can be applied to other tournaments, including table tennis and badminton. These models can integrate real-time information and adjust strategies during the game to ensure stability under various conditions.

However, we do not take into account the type of court and certain physiological factors of the players that affect tennis matches. In future research, a more comprehensive set of influencing factors needs to be considered.

References

[1] NOEL, Jordan Truman Paul; FONSECA, Vinicius Prado da; SOARES, Amilcar. A Comprehensive Data Pipeline for Comparing the Effects of Momentum on Sports Leagues. Data, 2024, 9(2): 29.

[2] Taylor J., Demick A. A multidimensional model of momentum in sports. Journal of Applied Sport Psychology, 1994, 6(1): 51-70.

[3] Iso-Ahola S. E., Mobily K. Psychological momentum: A phenomenon and empirical (unobtrusive) validation of its influence in a competitive sport tournament. Psychological Reports, 1980, 46(2): 391-401.

[4] Silva J. M., Hardy C. J., Crace R. K. Analysis of psychological momentum in intercollegiate tennis. Journal of Sport and Exercise Psychology, 1988, 10(3): 346-354.

[5] Richardson P. A., Adler W., Hankes D. Game, set, match: Psychological momentum in tennis. The Sport Psychologist, 1988, 2(1): 69-76.

[6] Dietl H., Nesseler C. Momentum in tennis: Controlling the match. International Journal of Sport Psychology, 2017, 48(365): 459–471.