A Study of Tennis Momentum Based on Markov Chains and Lgbms

Yilan Xie *

College of Energy, Xiamen University, Xiamen, China

* Corresponding Author Email: 32420222202045@stu.emu.edu.cn

Abstract. The heightened focus on competition has captured people's attention, elevating the significance of athletic endeavors. This study aimed to explore the effects of changes in sports competition. Through data preprocessing and principal component analysis, 13 factors influencing posture were confirmed and dynamic expressions were derived to measure athlete performance using entropy weighting. We further investigated the effect of competition fluctuations on the performance of future athletes, revealing the close relationship between situation size and subsequent scoring. Specifically, we used two methods, Markov Chain and Light Gradient Boosting Machine (LGBM), to predict match swings. Overall, this study deeply dissected the dynamic changes in athlete performance, reveals the interaction between posture and athlete outcomes, and provides practical suggestions for optimizing athlete status and competition strategies.

Keywords: Principal component analysis; Entropy weighting method; Markov chain; LGBM.

1. Introduction

Interest in professional sports has risen due to improved quality of life and higher fitness standards. Therefore, it is crucial to improve the predictive ability of coaches to effectively guide athletes and optimize players' performance in professional sports [1]. Based on the real game data, firstly, this paper divides the game data into 4 clusters by principal component analysis and retains their key information respectively. Then the entropy weighting method is applied to derive the weights of each principal component index, which in turn leads to the expression of momentum. Secondly, in order to investigate whether the fluctuations in the game would affect the future performance of the athletes, we first hypothesized that "the magnitude of momentum is related to whether the next goal is scored" and used the chi-square test to verify the hypothesis. Finally, we used both Markov chain and Light Gradient Boosting Machine (LGBM) to predict the fluctuation of the game. The results show that "whether the last goal is scored or not" has a significant impact on player performance, and recommendations are made for coaches accordingly.

2. Quantitative Momentum

The model is based on the assumption that all tennis players win with no accidents; all matches follow uniform rules ("three wins of five sets"); the spectators, referees and coaches do not affect the players. Based on the above assumptions, we will reduce the 13 factors to four key variables using principal component analysis (PCA) to achieve dimensional reduction and modeling of player performance in tennis [2]. The clustering results are shown in Table 1.

Table 1. Clusters of the factors

Cluster	Factor	Definition
Match Flow	X_1	Number of set-wins
	X_2	Number of opponent's set-wins
	X_3	Difference of game-wins
	X_4	Scoring rate
Mentality	X_5	Whether scored in the previous round
	X_6	Whether serving and failing in the first serve
	X_7	Whether made an error in the previous round
	X_8	Whether ace in the previous serve
	X_9	Winner-shot-type
Ability	X_{10}	Running distance in the previous round
	X_{11}	Speed of the previous serve
	X_{12}	Serve-depth or return-depth in the previous round
Objective Advantage	X_{13}	Whether serving

Here, we use PCA to construct indexes of match flow, mentality, ability, and objective advantage. Where Y_4 consists of only one factor, so $Y_4 = X_{13}$.

First, we normalized the data.

$$X_{ip} = \frac{x_{ip} - \mu_p}{\sigma_p} \tag{1}$$

Among them, μ_p is the mean of the $p-th$ factor, σ_p is the standard deviation of the $p-th$ factor, X_{ip} is the normalized data for the $p-th$ factor.

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_P \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_P \\ \ldots \\ Z_P = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_P \end{cases} \tag{2}$$

In Eq. (2), for each i, $c_{i1}^2 + c_{i2}^2 + \cdots + c_{ip}^2 = 1$ and $[c_{11}, c_{12}, \ldots, c_{1p}]$ maximizes the variance of Z_1. $[c_{21}, c_{22}, \ldots, c_{2p}]$ is orthogonal to $[c_{11}, c_{12}, \ldots, c_{1p}]$, and it maximizes the variance of Z_2; and (3)$[c_{31}, c_{32}, \ldots, c_{3p}]$ is orthogonal to $[c_{11}, c_{12}, \ldots, c_{1p}]$ and $[c_{21}, c_{22}, \ldots, c_{2p}]$, and it maximizes the variance of Z_3. Similarly, we can get all the p and principal components. The results of the principal component analysis are shown in Table 2.

Table 2. Results of principal component analysis of Y_1

No	Characteristic root	Variance contribution rate	Cumulative variance Contribution rate
X_1	1.46	36.38%	36.38%
X_2	1.19	29.73%	66.12%
X_3	0.85	21.35%	87.47%
X_4	0.50	12.53%	100.00%

$$\begin{cases} Z_1 = 0.28512\mathrm{X}_1 - 0.28515\mathrm{X}_2 + 0.574929\mathrm{X}_3 + 0.711921\mathrm{X}_4 \\ Z_2 = -0.707107\mathrm{X}_1 - 0.7071061\mathrm{X}_2 \\ \mathrm{Z}_3 - 0.46764\mathrm{X}_1 + 0.4676405\mathrm{X}_2 + 0.7208128\mathrm{X}_3 - 0.2074922\mathrm{X}_4 \end{cases} \tag{3}$$

Mentality as Z_4, Z_5, Z_6, Z_7, and Ability as Z_8, Z_9, Z_{10}. Combining the principal components, we derive the composite evaluation indexes Y_1, Y_2, Y_3.Combining the principal components we arrive at a composite evaluation index for Match Flow, Mentality, Ability.

$$Y = \sum_{i=1}^{m} a_i \times Z_i \tag{4}$$

Among them, m is the number of principal components for each index, a_i is the contribution of the principal component to the index.

Finally, we arrive at the expression for Y:

$$\begin{cases} Y_1 = 0.36Z_1 + 0.30Z_2 + 0.21Z_3 \\ Y_2 = 0.38Z_4 + 0.22Z_5 + 0.20Z_6 + 0.16Z_7 \\ Y_3 = 0.38Z_8 + 0.31Z_9 + 0.30Z_{10} \end{cases} \tag{5}$$

Next, we use the entropy weighting method to define the weights of the indexes Y_1, Y_2, Y_3, Y_4 for constructing the momentum index M [3].

Standardize data: in order to eliminate differences in data dimensions between indicators, standardized data processing is first carried out prior to data analysis:

$$x'_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \tag{6}$$

Determine indicator weights based on EWM.

Calculate the weight of point i on indicator j :

$$p_{ij} = \frac{1 + x_{ij}}{\sum_{i=1}^{1+x_{ij}} (1 + x_{ij})} \tag{7}$$

Calculate the coefficient of variation of entropy and index j:

$$e_j = -(\ln\ m)^{-1} \sum_{i=1}^{m}\ p_{ij} \tag{8}$$

$$g_j = 1 - e_j \tag{9}$$

The final weight of indicator j can be calculated:

$$w_j = \frac{g_j}{\sum_{i=1}^{n} g_j} \tag{10}$$

The weight set is as follows:

$$W = [w_1, w_2, w_3, w_4] \tag{11}$$

Using the entropy weighting method described above, we can get the weights of indicators Y_1, Y_2, Y_3, and Y_4, which are 0.039, 0.24, 0.035, and 0.6836, respectively.

Bringing in the values give:

$$M = 0.039Y_1 + 0.24Y_2 + 0.035Y_3 + 0.6836Y_4 \tag{12}$$

Substituting the values of the four key variables at each moment into the equation, the momentum of the player at the corresponding moment can be solved. The higher the momentum, the better the player will perform.

We visualized the data for one of the matches, visualizing the M of the two players and their trends at various time points of the match.



Fig. 1 Average moment of each game in set 1 and set 2

From Fig. 1 we can see that in four out of seven games in the first set the athlete with high momentum won, and in 11 out of 13 games in the second set the athlete with high momentum won. Overall, the trend matches the trend of the match score.

3. Correlation Analysis of Momentum and Success

It is clear that M is closely related to swings in play. Therefore, we can prove or disprove the coach's view by exploring the correlation between M and the player's future performance.

In statistics, the Pearson correlation coefficient r is used to measure the linear correlation between two variables and is calculated as [4].

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{13}$$

The closer the correlation coefficient is to -1 or 1.

In order to investigate the correlation between "whether M is higher than 0.5 and whether the next ball is scored or not" in the case of the distribution of unknown variables, we divided M into two groups of more than 0.5 and not more than 0.5 and adopted the Chi-square test method. The formula of Chi-square test is:

$$\chi^2 = \sum_{i=1}^{r}\frac{(n_i-np_i)^2}{np_i} \tag{14}$$

Among them, n_i is the actual number of balls scored on the next ball when M > 0.5, np_i is the theoretical number of balls scored on the next goal.

The solution of χ^2 is 731.812, the significance level is less than 0.001.

We calculated the average momentum per set for each of the two players over a total of 1109 sets in 29 matches. To better measure the performance of the players in each game, we divided the specific performance. As shown in Table 3. Where a higher number means that the player performed better in the game.

Table 3. Players' wins and losses in different situations

No	Performance of game	Definition
1	thorough defeat	lose in dominant game
2	narrow defeat	lose in locked game
3	narrow victory	win in locked game
4	thorough victory	win in dominant game

Dominant game: the number of rounds is less than or equal to 6, meaning that they end without experiencing a 40-40 scoreline.

Holding game: with a round count of no less than 6, meaning that a 40-40 score was experienced.

The already quantified per-game performances were combined with the corresponding M to obtain a total of 2,218 data sets. Similarly, we used the chi-square test again to investigate the correlation between "the size of the average momentum of the innings and the performance of the innings" by dividing the M into four groups, namely, less than 0.25, between 0.25 and 0.5, between 0.5 and 0.75, and greater than 0.75, and solved for the correlation coefficient of the two groups, which is 66.1%. The Pearson correlation coefficient is 0.68.

In Fig. 2, we can see that the performance of through defeat is much better when $0 \le M < 0.25$, and the performance of through victory is much better when $0.75 \le M < 1$. We can conclude that "M is significantly associated with better performance".

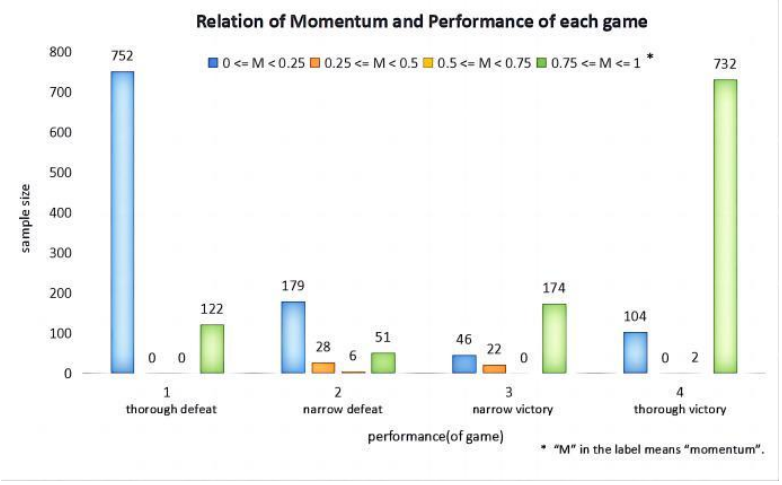


Fig. 2 Relation of momentum and performance of each game

Taken together, we reach the final conclusion that a player's momentum is closely related to his future performance, and that more momentum on the part of a player in a game lead to a higher likelihood of winning the game. Therefore, the coach's argument that fluctuations during a game can affect a player's future performance is inappropriate.

4. Swings Prediction

4.1. Markov Chain

To predict momentum in a match, we define momentum through swings linked to a player's past performance changes. Performance is categorized into defeat, narrow defeat, narrow victory, and victory states with values 1, 2, 3, and 4. The average of P over two consecutive games (AP) measures player performance, ensuring fairness. Using the last 20 games' data (70%), a model was trained, solving for P and AP in each game for each player. Next it is necessary to define how large a change in AP is a swing. We define: AP_{t-1} at the moment of $t-1$, if:

$$|AP_{t-1} - AP_t| \geq 1.5 \tag{15}$$

Then a swing occurs (symbolized by S). i.e.

$$S_t = \begin{cases} 1, |AP_{t-1} - AP_t| \geq 1.5 \\ 0, |AP_{t-1} - AP_t| < 1.5 \end{cases}, t \in N^+ \tag{16}$$

Accordingly, we wrote a program to solve for S for every two games.

Next, we need to make predictions using our model. We solved for the Markov chain state transfer matrix for the P of each game using the Markov Chain approach [5].

$$P_1 = \begin{pmatrix} 0.1831 & 0.0662 & 0.1606 & 0.5901 \\ 0.2613 & 0.0905 & 0.1558 & 0.4925 \\ 0.5317 & 0.1268 & 0.0829 & 0.2585 \\ 0.5986 & 0.1571 & 0.0643 & 0.1800 \end{pmatrix} \tag{17}$$

On this basis, based on the definition of swing, the probability of a swing occurring in the state of the performance of each game at each moment in time can be found, as shown in Fig. 3.

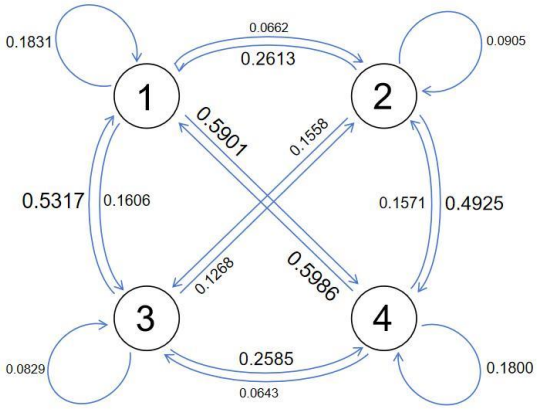


Fig. 3 State transfer probability of performance

Based on the state transfer matrix of performance, the state transfer matrix of average performance is further solved for both games. Two games are a unit of time. Where, if the performance of the first game is a and the second game is b, the state of the two games is noted as: $\overline{ab}$. Based on this, the probability of a shift (P_s) occurring in the state at each moment can be found based on the definition of shift shown in Table 4.

Table 4. The probability that a swing occurs.

State	$\overline{11}$	$\overline{12}$	$\overline{13}$	$\overline{14}$	$\overline{21}$	$\overline{22}$	$\overline{23}$	$\overline{24}$	$\overline{31}$	$\overline{32}$	$\overline{33}$	$\overline{34}$	$\overline{41}$	$\overline{42}$	$\overline{43}$	$\overline{44}$
P_s	0.75	0.21	0.28	0.46	0.21	0.04	0.03	0.11	0.11	0.03	0.04	0.23	0.46	0.26	0.19	0.82

Predictions are made based on the above probabilities. When P_s > 0.5, it is predicted that a swing will occur, otherwise it is predicted as not occurring. i.e.

$$S_{predict} = \begin{cases} 1, p_s \geq 0.5 \\ 0, p_s < 0.5 \end{cases} \tag{18}$$

4.2. The Model of LGBM and Feature Importance Assessment

Analyzing 13 factors, including "serve or not," on P to predict fluctuations. Using four models (LGBM, SCV, Logistic, MLO), we enhance accuracy and assess their impact [6]. We selected one part of the data as a training sample and the other part as a test sample, and calculated the values of accuracy A and the reconciled average F1 Score of Precision and Recall for each of the four models, where Accuracy A is used to denote the accuracy of a model's prediction, and F1 Score is used to consider the model's precision and robustness in a single metric, and their calculations are based on the F1 Score. and robustness, and their formulas are given below:

$$A = \frac{Number\ of\ correct\ predictions}{Total\ sample\ size} = \frac{n}{N} \tag{19}$$

$$P = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \tag{20}$$

$$R = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FN)} \tag{21}$$

$$F = 2 \times \frac{P \times R}{P + R} \tag{22}$$

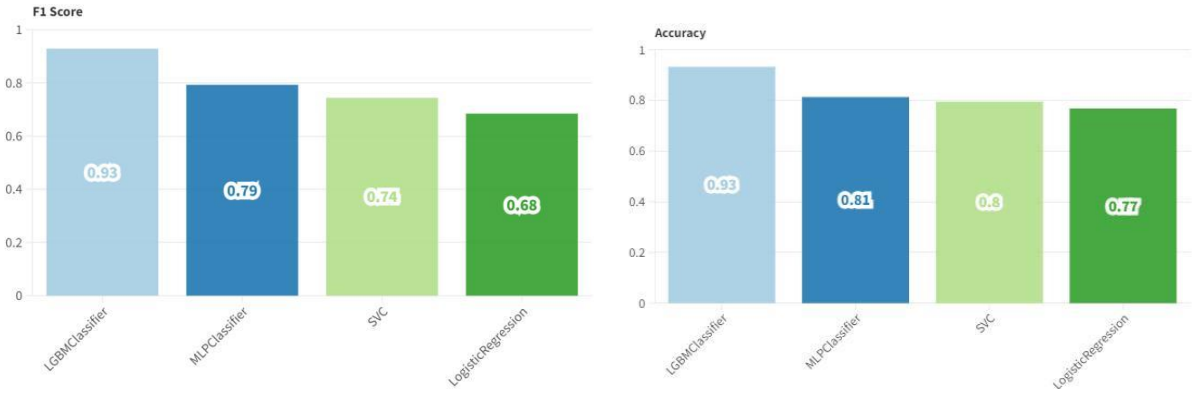


Fig. 4 Accuracy and F1 score of four models.

Through Fig. 4 we can clearly find that the Accuracy and F1 Score of the LGBM model are greater than the other models, i.e., the model has the optimal accuracy and, precision and robustness, so we choose the LGBM multiple classification model to explore the relationship between performance and $X_1 \sim X_{13}$.

In multiple classification problems, LGBM uses a SoftMax function to compute the probability of each category, and in this problem, the model will select the P with the highest probability as the final output. Its goal is to minimize the logarithmic loss of multiple categories.

For a classification problem containing K categories given a sample, the probability of predicting it into category k is:

$$P(Y = k \mid x) = \frac{e^{s_k(x)}}{\sum_{j=1}^{K} e^{s_j(x)}} \tag{23}$$

Where $s_k(x)$ is the score of the model for sample x belonging to the category.

The logarithmic loss function for multiple classification is:

$$L = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log P(Y = k \mid x_i) \tag{24}$$

Where N is the number of samples and y_{ik} is an indicator function that is 1 if sample i belongs to category k and 0 otherwise.

In order to assess the importance of the 13 factors affecting P, we used the method of fea-ture ranking importance. Feature arrangement importance is a method used to assess the im-portance of features in a machine learning model. The basic idea of this method is to measure the extent to which each feature contributes to the model performance by changing the values of the features (i.e., arranging the features) and observing the changes in the model performance. The results are shown in Fig. 5.



Fig. 5 The importance of X

From Fig. 5 we can learn:

Firstly, the factor X_5 affecting mentality is most likely to cause a swing in the game situation, and the importance of whether or not the last ball was scored is as high as 0.5, which is much higher than the rest of the influencing factors. From this we can learn that whether or not the last ball was scored greatly affects the psychology of the players, which in turn affects the performance of a game, causing the dynamics of the game to shift from one player to another.

Secondly, whether X_{13} serves or not is an objective advantage that exists in the game, and it will also have a greater impact on the shift of the game state.

Thirdly, X_{10} and X_{11} show the strength of the athlete's physical fitness, and X_{12} shows the strength of the athlete's own game. Both of these factors are not likely to change significantly in a short period of time, and therefore have little or no effect on the change of game dynamics.

Finally, $X_1 \sim X_4$ are summaries of the overall situation at the end of the match and have no effect on the shift of the player's posture in the current match.

4.3. Recommendations

Combining Markov Chain's average performance of the first two games to predict whether or not a swing occurs, and LGBM's analysis of the importance of X, we make the following recommendations:

Advice1: Based on predicting performance at the point of momentum shift in a match, coaches should focus on specific match-turning point indicators, such as whether or not the previous ball was scored, whether or not they have the serve, etc., that may signal an impending change in the flow of the match.

Advice2: Train players to recognize and capitalize on turning points in the game, such as mental preparation and tactical adjustments when scoring consecutive points or facing important points.

Advice3: analyzes game data from different opponents to respond to different game situations with customized strategies.

5. Conclusion

In order to investigate how various events in a match create or change momentum and how momentum affects the flow of the match, a model was developed to quantify momentum, analyze the causes and effects of momentum, and predict changes in momentum. First, the weights of each principal component metric were derived through principal component analysis and entropy weighting, and then momentum expressions were derived to quantify an athlete's performance at specific moments of a match. Then, we investigated whether fluctuations in the game affect players' future performance by chi-square test and found that players' motivation is closely related to their future performance and affects the direction of the game and players' performance. Finally, we use Markov chain and Light Gradient Boosting Machine (LGBM) to predict game swings. In the Markov chain approach, we divide the game data into training and validation sets, solve the performance state transfer matrix using the former data, and calculate the probability of game swings for different performance states of the players through Markov chain. We found the relationship between 13 factors and performance through LGBM and solved the expression of performance. Then, we use the importance feature ranking method to understand that "whether the last goal is scored or not" has a great influence on players' performance and make suggestions to coaches accordingly.

References

[1] Den Hartigh R J R, Niessen A S M, Frencken W G P, et al. Selection procedures in sports: Improving predictions of athletes' future performance [J]. European journal of sport science, 2018, 18(9): 1191-1198.

[2] Giles B, Peeling P, Kovalchik S, et al. Differentiating movement styles in professional tennis: A machine learning and hierarchical clustering approach [J]. European Journal of Sport Science, 2023, 23(1): 44-53.

[3] Zhu Y, Tian D, Yan F. Effectiveness of entropy weight method in decision-making [J]. Mathematical Problems in Engineering, 2020, 2020: 1-5.

[4] Armstrong R A. Should Pearson's correlation coefficient be avoided? [J]. Ophthalmic and Physiological Optics, 2019, 39(5): 316-327.

[5] Gensbittel F, Renault J. The value of Markov chain games with incomplete information on both sides [J]. Mathematics of Operations Research, 2015, 40(4): 820-841.

[6] Ge C, Deng F, Chen W, et al. Machine learning for early prediction of sepsis-associated acute brain injury [J]. Frontiers in Medicine, 2022, 9: 962027.