Research On the Role and Influencing Factors of Momentum in Tennis Matches Based on Linear Conditional Probability Model and Support Vector Machine

Jiadong Zhang*, Kaifan Tang and Xuanyi Zhu

School of Artificial Intelligence, Xi 'an Jiaotong University, Xian, China

* Corresponding Author Email: n7242187@gmail.com

Abstract. This research investigates the role and influencing factors of momentum in tennis matches using a Linear Conditional Probability Model (LCPM) and Support Vector Machine (SVM). The study begins with data preprocessing to address outliers and standardize scoring notations, followed by an analysis of the momentum's role in tennis, which is defined in relation to consecutive points, games won, and service breaks. The principal component analysis (PCA) is employed to synthesize key factors reflecting momentum, including consecutive scores, untouchable shots, aces, and net points, with a focus on their Markovian properties. A momentum formula is constructed, accounting for the sliding average of scoring differentials and the mutual cancellation of momentum between players. The study further examines the stochasticity of swings in play and runs of success, challenging the notion that these elements are random. The results of the Wald Wolfowitz Run Test and Granger Causality Test indicate that momentum significantly influences the probability of winning points and is not a random phenomenon. A PCA-SVM-SHAP model is developed to predict momentum fluctuations, achieving an 81.26% accuracy rate in the validation set. The model identifies net skills, serving skills, and untouchable shots as the most influential factors on momentum changes. The research extends the model's application to unknown tennis matches and other sports events, demonstrating its generalization ability with varying degrees of accuracy.

Keywords: Momentum in Tennis; Linear Conditional Probability Model; Support Vector Machine; Principal Component Analysis; Granger Causality Test.

1. Introduction

Tennis, as a highly competitive and dynamic sport, has long fascinated researchers and enthusiasts alike with its strategic depth and the psychological ebbs and flows experienced by players during matches. A particularly intriguing concept within the sport is the notion of momentum, which is often perceived as a sequence of events that can significantly influence the outcome of a game[1, 2, 3, 4]. While anecdotal evidence and intuitive observations have positioned momentum as a pivotal factor in match dynamics, its quantification and understanding remain a subject of considerable debate and investigation.

This study delves into the role of momentum in tennis matches, aiming to provide a comprehensive analysis of its definition, manifestation, and impact on player performance. By leveraging the Linear Conditional Probability Model (LCPM) and Support Vector Machine (SVM)[5], we seek to establish a mathematical framework that not only quantifies the elusive concept of momentum but also predicts its influence on subsequent game points[6].

2. Data Processing

First, we processed the data for outliers. Then, according to the rules of tennis, after getting 40 points first, a player can win the game by winning one more point, so defeating a rival to win a game requires scoring four balls, the winning team’s score at this point is set at "50". AD is the case of winning one or more balls after the fourth ball. Considering the uncertainty of the number of times AD occurs and the overall general probability of AD occurring, we standardize AD as winning five balls. In the case of AD, a player can win the game by winning one more point and winning a game

requires a difference of at least 20 scores between the two players.Therefore, in order to quantify the subsequent indicators, the "AD" in the score is replaced by the numerical value "60".

The textual indicators in the back of the table are converted to numerical format. The status of different shot types is equal, so the discrete unordered variable Forehand(F) is set to 1 and Backhand(B) is set to 2; the unspecified value of speed in speed_mph is set to 0. For serve_width, the unknown value is set to 0, and the other status of different serve widths is equal, so the discrete disorder variables B, BC, BW, C, and W are set to 1 ~5 respectively; similarly, the unknown value of serve_depth is set to 0, while CTL and NCTL are replaced by 1 and 2; the return_depth unknown value set to 0, D, ND replaced by 1, 2.

Two individuals, Jiri Lehecka and Daniil Medvedev, are randomly selected from the entire pool of athletes, and all matches of these two individuals are excluded from all analyses in the first three problems, to be utilized as a test set for the model generalization.

3. Role of Momentum

In some matches consecutive winners are indeed more likely to go on to win the next match. In other words, all that needs to be shown is that the probability of winning the next point under conditions of winning several points in a row is greater than the probability of scoring points affected by the initial serve.

3.1. Quantitative Definition of Momentum

We have selected the situations that may reflect "momentum", including consecutive points scored by two players, hitting untouchable winning serves, double faults, unforced errors, hitting untouchable winning shots, breaks of serve, aces, and net points. The principal component analysis of the two athletes is carried out by setting the number of selected factors to 1. According to the component matrix, it can be found that the momentum is mainly reflected in the four factors of consecutive scores, untouchable shots, the number of aces, and scores made to the net, and their compositions are almost all greater than 0.5.

Considering the momentum embodied in scores has a Markovian property, replacing consecutive scores with a sliding average of every four scoring differentials (since a player can win up to five consecutive rounds), and only considering the current values of the other factors, and also, we consider that the momentum generated by a player and another player in these factors cancel each other out. Consequently, based on the scoring coefficients of the factors, the formula for the momentum is constructed as follows.

$$momentum_1 = 0.230 \cdot p1\_dpoint - 0.229 \cdot p2\_dpoint + 0.128 \cdot p1\_net\_pt\_won - 0.147 \cdot p2\_net\_pt\_won + 0.147 \cdot p1\_ace - 0.152 \cdot p2\_ace + 0.179 \cdot p1\_winner - 0.187 \cdot p2\_winner. \quad (1)$$

$$momentum_2 = -momentum_1. \quad (2)$$

Where $p1\_dpoint$ is a sliding average of the first-order differences in the athletes' winning scores, defined by the following equation.

$$p1\_dpoint = \begin{cases} \frac{1}{4} \cdot p1\_points\_won & i = 0,1,2,3 \\ \frac{k}{4}\sum_{i=4}^{k}(p1\_points\_won_i - p1\_points\_won_{i-4}) & others \end{cases}. \quad (3)$$

Where k is the number of games in a match. Similarly, we get the definition of $p1\_points\_won_i$. With this, we get the value of momentum for all players as it changes with each score.

3.2. Stochasticity Analysis of Swings in play and runs of success

Swings in play and runs of success for any given player are randomized. To evaluate his statement, we interpret runs of success as the probability of winning the next point conditional on winning i points in a row, that is as follows.

$$p(next\ win|win\ ipoints\ in\ a\ row),\quad i = 1,2,3,4,. \tag{4}$$

And swings in play as follows.

$$p(next\ lose|win\ i\ points\ in\ a\ row) = 1 - p(next\ win|win\ i\ points\ in\ a\ row). \tag{5}$$

Since the two athletes win or not is conjunctive, if $p(next\ win|win\ i\ points\ in\ a\ row)$ is not random. Thus, we get that neither swings in play nor runs of success are random, and the converse is also true. We argue that if runs of success are random, then whether or not the next game is won is not affected by whether or not itwas won before, so we get (6).

$$\begin{aligned} p(next\ win|win\ i\ points\ in\ arow) &= p(win\ a\ point) \\ &= p(serve) \times p(win|serve) + p(not\_serve) \times p(win|not\_serve) \\ &= p(serve) \times p(win|serve) + p(rival\_serve) \times p(rival\_loss|rival\_serve) \\ &= p(serve) \times p(win|serve) + 1 - p(rival\_serve) \times p(rival\_win|rival\_serve) \end{aligned} \tag{6}$$

The probability of winning under any circumstance is constant, and thus p is a constant, obeying a uniform distribution. Adding up the number of points won in succession by each person in each match, i can take values from 1 to 11 (consider cross-set wins as also consecutive wins), and using the great likelihood estimation, we get (7).

$$p(nextwin|winipointsinarow) = 1 - \frac{t_i}{t_{i+}}. \tag{7}$$

where t_i is times of consecutive win i points, t_{i+} is times of points won consecutively greater than i.

Wald Wolfowitz Run Test is performed to examine the consistency of the distribution between the true probability distribution and the probability distribution of the randomized scenario. Five individuals were randomly selected to test for consistency of the distribution, and the results are shown in Table 1 below.

Table 1. Distributional Consistency Run Test

Player	z-value	p-value
Carlos Alcaraz	-3.446	0.001
Novak Djokovic	-2.847	0.004
Roman Safiullin	-3.714	0.000
Alexander Zverev	-2.120	0.034
Alexander Bublik	-2.504	0.012

Significance p-values present all less than 0.05, meaning that there is a 95% confidence that swings in play and runs of success are not random.

3.3. Hysteresis Analysis of Effect of Momentum

From the previous section, we have understood that the fluctuations in wins and losses are not randomly distributed, and what follows is an exploration of the relationship between momentum and swings in play.

3.3.1 Visualization of Lagged Effects

As an example, data of the 2023 Wimbledon Gentlemen's final is plotted as a line graph, which depicts the lagged effect between momentum and Alcaraz's score fluctuations in this set, as Fig.1 below.

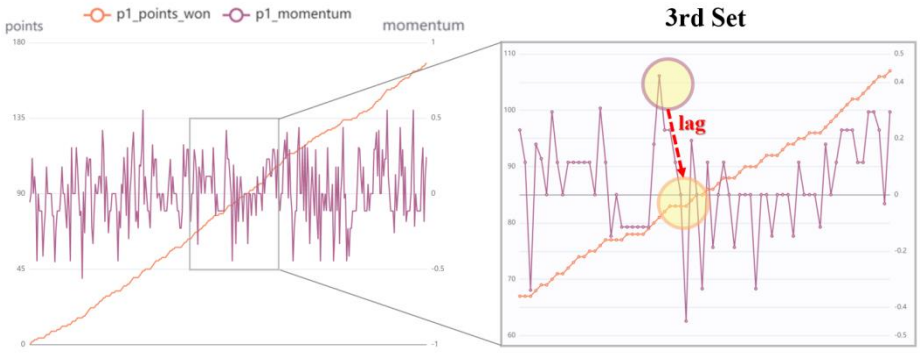


Fig. 1 Trend Plot of Lag Correlation

From the graph, it can be noticed that momentum has a lagged effect on runs of success. Taking the 3rd set as an expressive example, as the circle markers, we intuitively see that momentum first declines causing subsequent losing scores, which is manifested in the unchanged cumulative points, representing opponent has made scores at this stage.

3.3.2 Granger Causality Test

Granger Causality Test is conducted using Carlos Alcaraz as an example. First, the ADF smoothness test results in p1_momentum significance p-value of less than 0.01, which is smooth at 99% confidence, while p1_points_won significance p-value is 0.108, which is not smooth at 90% confidence. After performing first-order differencing on it, the significance p-value of p1_points_won_d is obtained to be less than 0.01, which is smooth at 99% confidence level.

Table 2. Granger Causality Test

Cause-and-effect pairs	F	P
(p1_momentum, p1_points_won_d)	5.992	0.003***
(p1_points_won_d, p1_momentum)	23.334	0.000***

From the above table, when the variable p1_momentum is the cause and p1_points_won_d is the effect, the significance p-value is 0.003, which suggests the original hypothesis can be rejected. The conclusion is obtained that p1_momentum can cause changes in p1_points_won_d. Also, in the case where p1_points_won_d is the cause and p1_momentum is the effect, it appears significant. In conclusion, it shows that p1_points_won_d and p1_momentum interact with each other as cause and effect.

Thus, we can assume with 99% confidence that momentum plays a role in winning and losing, in other words in swings in play and runs of success. Positive momentum is more likely to lead to a win and negative momentum is more likely to lead to a loss.

4. Analysis of Influential Factors and Models for Predicting Scores

Five features are synthesized for predicting the score of a game with Principal Component Analysis, and a visual interpretation of each metric is given based on the weights of the primary features in each of the synthesized metrics. Support Vector Machine is employed to make a prediction of the positive and negative of momentum to get the result of whether or not a point will be scored in a game, and the prediction is visualized.

In the second question, we know that momentum affects swings of play, a positive momentum gives a higher probability of winning the next point, and a negative momentum indicates no score. Positive or negative momentum forms two categories 0,1, using a binary classification model to output the categorical probability, which is the probability of scoring at the next point.

4.1. Feature Extraction

The raw data has many features. As such the following processing is done to get the important features suitable for subsequent analysis.

Step1. Standardization of data: For all the data columns recorded, a total of 37 features, the features that are not related to reflecting momentum and winning or losing are manually filtered out initially, and the features that directly reflect winning or losing and momentum are also screened out to get the remaining 25 features. The z-score is normalized for all these features with the following equation.

$$y_i = \frac{x_i - \bar{x}}{s}. \tag{8}$$

where x_i is the i^{th} original value of the feature, $\bar{x}$ is the mean of the feature in this column, and s is the standard deviation of the feature.

Step2. Spearman's Correlation Analysis: The selected features were subjected to Spearman's Correlation Analysis, and the features whose correlations were all greater than 0.8 or less than 0.3 were excluded to obtain the 20 residual features.

Step3. Principal Component Analysis: Principal component analysis was employed on the selected remaining features with KMO>0.5 and Bartlett's test significance <0.001, yielding eight principal component features.

Table 3. Explanation of the Meaning of Principal Components

No.	Evaluation Indicators	Factors with Higher Loads
1	Physical fitness	p1_distance_run, p2_distance_run, rally_count
2	Opponent's fault	P2_double_fault, p2_unf_err, speed_mph
3	Hitting technique	P1_ace, p1_winner
4	Own fault	p1_double_fault, p1_unf_err, speed_mph
5	Opponent's skill level	P2_ace, p2_winner
6	Serving skill	p1_ace, serve_width, serve_depth
7	Opponent's play at the net	P2_net_pt_won
8	Tecniques at the net	p1_net_pt_won, p2_break_pt_won

Analyzing the rotated matrix component table obtained from PCA, it is found that the first principal component of the player's running distance and the number of rounds factor had large loadings of 0.913, 0.908, and 0.877, which are hypothesized to be associated with physical fitness; the second principal component with the large proportion of player2's double-faults, unforced errors, and the speed of serve, considered to be related to player2's errors. Using the same approach, we infer the implications of the remaining synthetic principal components.

In summary of the results of PCA, we gained five comprehensive evaluation indicators, physical fitness, net skills, errors, hitting skills and serving skills.

4.2. Modeling of Score Wins and Losses

The five extracted features are provided as inputs to the classification model, with two categories consisting of positive and negative current momentum, and then the probability of a score will be predicted. We establish the model as follows.

$$\begin{aligned} Prediction\_score &= \sum_{i=1}^{5} w_i \cdot \phi(feature_i) + b, \\ \min &\ \frac{1}{2} \cdot \| w_i \|_2. \end{aligned} \tag{9}$$

Where $\phi(feature_i)$ is the result of a spatial transformation using a radial basis function, which is an equally spaced independent variable around a fixed-point c with the same function value.

We divide the data into training and testing sets in the ratio of 8:2 and then call the Python scikit-learn package for SVM based on radial basis function to train all the data. The training results of the model are displayed as shown in Figure 8 below and the accuracy on the test set is 81.26%.

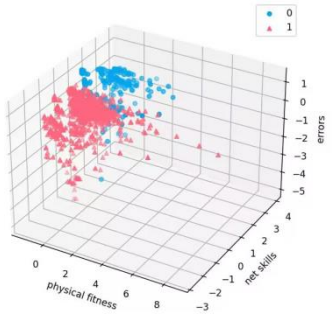


Fig. 2 Visualization of SVM

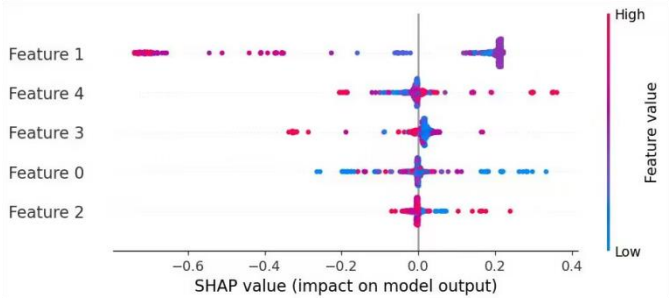
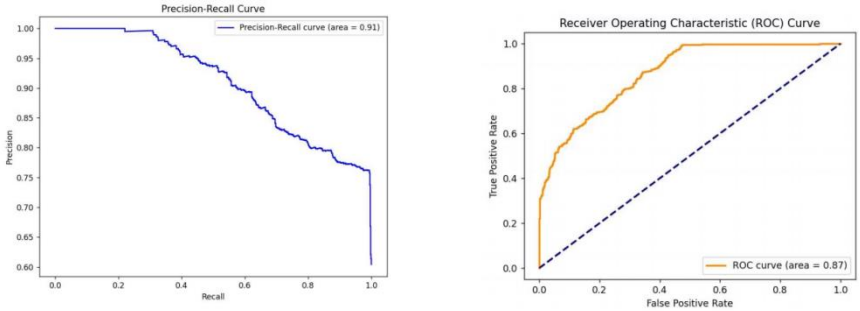


Fig. 3 Characterization Summary Chart

4.3. Evaluation of the Model

According to the classification prediction results of SVM, we plotted the P_R curve and ROC curve as shown in Fig.4. The areas under the curves are both relatively large, reflecting the superior model prediction results.



(a) P_R Curve (b) ROC Curve

Fig. 4 Visualization of Model Accuracy

Since the SVM algorithm is a black-box model and suffers from poor interpretability, we use the SHAP method, hoping to give an order of importance for numerous features. SHAP is an additive interpretive model based on Shapley's value inspiration. When the model is linear, for each predicted sample, the model produces a predicted value, and the SHAP model produces values assigned to each feature of that sample such that the predicted value of the model is the sum of the SHAP values. Since the SVM is essentially a linear model, for the ith sample, the formula for the predicted value and the SHAP value is established as follows.

$$Prediction\_i = Prediction\_base + \sum_{j=1}^{5} f\,(feature\_i\_j). \tag{10}$$

Where $Prediction\_i$ is the prediction result of the i^{th} sample, $Prediction\_base$ is the mean value of the prediction target of all samples, and $f(feature\_i\_j)$ is the SHAP value of the j^{th} feature of the i^{th} sample.

SHAP analysis is performed on the input factors of our model to plot a global SHAP interpretation scatter plot, as shown in Fig.3. Each point in the plot represents a sample, and the vertical axis is the sum of the SHAP values of the corresponding features of all the samples, the higher the vertical axis the greater the impact of the features on the model output. The horizontal axis is the distribution of SHAP values, the further away from 0 the SHAP value is, the greater the impact on the model output. The closer the color of the dot is to a warm tone, the greater the value of that feature for that sample.

Analyzing Fig.3, it can be discovered that the influences on swings of momentum are "Feature1", "Feature4", "Feature3", "Feature0", "Feature2", which means net skills, serving skills, errors, hitting skills and serving skills, and physical fitness, in descending order. The SHAP values of the rest of the features, except for net skills and break of serve, are clustered around 0, indicating that the rest of the features do not have much effect on the swings of momentum.

5. Extension of model

5.1. Modeling Application to Unknown Tennis Match

To further test our model, a duel between two people Jiri Lehecka and Daniil Medvedev, who are not used in the previous training of the model, is chosen to predict the change in the output momentum score. The confusion matrix is obtained as in Table 4.

Table 4. Confusion Matrix for Unknown Tennis Match

Confusion Matrix	True value is Positive	True value is Negative
Predict value is Positive	79 (TP)	0 (FN)
Predict value is Negative	17 (FP)	26 (TN)

Then it is calculated as follows.

$$Precision = \frac{TP}{TP+FP} = 0.82. \tag{11}$$

$$Recall = \frac{TP}{TP+FN} = 1.0. \tag{12}$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} = 3.0. \tag{13}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 0.86. \tag{14}$$

From this, we prove the model's predictions are quite accurate.

Upon further analysis of the confusion matrix, we surprisingly find that the model's prediction of this set of matches has a false-negative number of 0, indicating that the model has not predicted the case of positive momentum as a case of declining momentum, but has predicted the case of declining momentum as a case of positive momentum. Observing the backward and forward factors corresponding to this type of prediction error data, it is found that a portion of the prediction error is not considered in one person's score in AD, suggesting that a person's momentum fluctuates due to psychological reasons at game point. Thus, in the future, psychological factors can also be considered to be added to the model.

5.2. Extension of Model to Other Sports Events

We crawl the partial match data of HARIMOTO TOMOKAZU vs. Puretea in the FIVB Czech Republic 2019 from the FIVB website for a total of 127 matches, including athletes' scores, number of rounds, whether the serve is successful or not, whether the return is faulty or not, whether the serve is scored or not, the score of the long round duel, and the consecutive scoring situation. The data is cleaned by eliminating invalid values and null values, and momentum is calculated according to whether the athlete scored data. Then the PCA-SVM model is employed with the athlete's momentum as the dependent variable and the others as the independent variables, and the prediction accuracy is 0.774. The confusion matrix is shown in Table 5 below.

Table 5. Confusion Matrix for Other Sports Events

Confusion Matrix	True value is Positive	True value is Negative
Predict value is Positive	61 (TP)	40 (FN)
Predict value is Negative	26 (FP)	0 (TN)

5.3. Sensitivity Analysis

For the five extracted principal components, we respectively fluctuate up and down 20% for each component to get the range of the model prediction accuracy. The definition of the range is as follows.

$$R = X_{max} - X_{min}. \tag{15}$$

which reflects the magnitude of the fluctuation range of the model prediction accuracy. The ranges of the five principal components are all less than 3%, as shown in Figure 14, which therefore demonstrates that the model stability is high.

Fig. 5 Sensitivity of Model

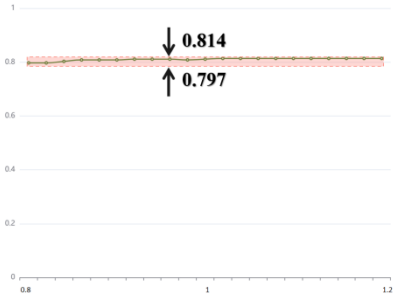


Fig. 6 Stability of Speed

To generalize the model to unknown tennis tournaments, we have to take into account the possible effect of the court on the ball speed [10]. In Wimbledon, the court is grass, and the ball is faster, while the rest of the tournaments, such as the US Open, use hard courts and the French Open uses red clay tennis courts. Therefore, 20% up and down fluctuations were made to the speed of the ball in the given table to analyze the changes in the model performance.

As can be seen from the Figure 6, the resultant float value is 0.026% at maximum. Using the same methodology, models for predicting scores is obtained as 0.029% at maximum, which indicates that the model is very stable and resistant to disturbances.

6. Conclusions

This study conducted an in-depth analysis of momentum and its influencing factors in tennis matches based on the Linear Conditional Probability Model (LCPM) and Support Vector Machine (SVM). By constructing a comprehensive evaluation system, we assessed player performance holistically and quantified the impact of momentum on match outcomes. The research found that momentum plays a significant role in tennis matches, and its effects are not random. The main strengths of the study lie in its comprehensiveness, precision, extensibility, and stability. The high accuracy of the SVM model in predicting match scores further confirms its effectiveness in sports analysis. Moreover, the successful application of the model in different tennis events and other sports demonstrates its strong generalization capability.

However, the study also has some limitations. Although the binary classification model used is effective, to more fully explore the fluctuation trends in score data, future research could consider employing Markov models.

References

[1] Miller,S.and Weinberg,R.(1991), Perceptions of psychological momentum and their relationship to performance.The Sport Psychologist.5.211-222.

[2] Richardson,P A,Adler,W,and Hankes,D,(1988).Game, set, match: psychological momentum in tennis.The Sport Psychologist,2.69-76.

[3] Wardrop,R.(1995). Simpson's paradox and the hot hand in basketball. The American Statistician.49.24-28.

[4] Rees, C. and James, N. (2006). A new approach to evaluating 'streakiness' in golf, a researchpaper presented at the 7h world congress of performance analysis in sport,.Szombathely, Hungary, August. In H. Dancs, M. Hughes and P. O'Donoghue (eds.) Book of Proceedings of the World Congress of Performance Analysis of SportVI, Szombathely: Hungary.pp.329-337

[5] Huang S, Cai N, Pacheco P P, et al. Applications of support vector machine (SVM) learning in cancer genomics[J]. Cancer genomics & proteomics, 2018, 15(1): 41-51.

[6] Meier P, Flepp R, Ruedisser M, et al. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis[J]. Journal of economic psychology, 2020, 78: 102269.