Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Based on the LGBM model, a dynamic analysis was conducted to study the momentum of athletes in tennis matches

Haoming Tian, Ye Zhang, Runduo Ding*

Faculty of Electrical and Control Engineering, L.N TECHNICAL UNIVERCITY Huludao, China, 123032

*Corresponding author: 15704130526@163.com

Abstract. The changes in momentum in a tennis tournament have a significant impact on players' performance and the outcome of the match. The application of big data and machine learning techniques can effectively enhance the accuracy of predicting these dynamic changes. This article aims to delve into the impact of momentum on matches and how it can be leveraged to help players achieve success. Firstly, this paper determines the confidence level of each indicator through multiple logistic regression and uses the LGBM algorithm to train the data to predict the real-time momentum of an athlete, confirming that momentum is real and will have a great impact on the competition result. The paper trains the obtained data using the SVR model. By comprehensively applying the information gain method, this paper derive importance scores for each indicator. This measure was found to have the biggest impact on momentum in the progress in scoring.

Keywords: Light Gradient Boosting Machine, Support Vector Regression, Momentum, Multiple logistic regression, Tennis tournament.

1. Introduction

In tennis, momentum is often considered to be one of the key factors affecting the outcome of a match. The change of momentum can affect the mental state and performance of the players, thus changing the direction of the game at a crucial moment. Therefore, how to accurately analyze and predict the momentum change in the game has become an important topic in sports science research. With the development of big data technology and machine learning algorithms, these advanced technologies are widely used in the analysis of sports competition data to improve the accuracy of the prediction of game results, and thus improve the win rate.

For this study, this paper started to use LGBM and SVR algorithms. To prove that momentum is real and determine the extent of its impact on the game, this paper introduced the Kumar mas LGBM algorithm used in the field of Cluster computing. On this basis, this paper integrate multiple logistic regression algorithms into the LGBM algorithm to judge the significance of each index and improve the accuracy of prediction results. To make better use of momentum to help athletes win games, this paper introduces the PSO-SVR model used by Zhou H in the field of Tunnelling and underground space technology. This paper draws on the advantages of SVR big data model prediction and uses the ID3 algorithm to calculate index weights to make momentum prediction more accurate.

Firstly, in this work, this study predicted the momentum of each player based on the LGBM model. Key indicators describing player momentum were extracted using multivariate logistic regression. Initially selected the LGBM model with a higher AUC value based on the ROC curve and further optimized it using the LGBM algorithm through machine learning training. Then this paper conducted a quantitative analysis of player momentum using the pre-trained model. Secondly, this paper conducted an analysis of momentum shifts in the game based on the SVR model utilized to predict momentum shifts in the game. The paper analyzed momentum shifts in the game using the SVR model and identified several factors with the highest correlation.

2. Data Preprocessing and Preparation

2.1. Data source and preprocessing

Data in this study are derived from www.comap.com. The statistics table Wimbledon featured matchs.csv contains two problems: One is Some samples have outliers, which deviate from the data group, and the other is There were missing values in the data of pace. This paper deals with outliers in the data to prevent deviations from the actual scenarios in subsequent analysis. For the missing pace data, utilized cubic spline interpolation to fill in the missing values, ensuring the reasonableness and accuracy of numerical predictions. To better evaluate the model and explore the relationship between real-time momentum and scores, as well as the relationship between fluctuations and success, This paper treated the data labels in the Wimbledon featured matches.csv table as indicators and then performed dimensionless processing on each indicator column to mitigate the impact of different dimensions on the final results.

2.2. Description of tennis tournament format

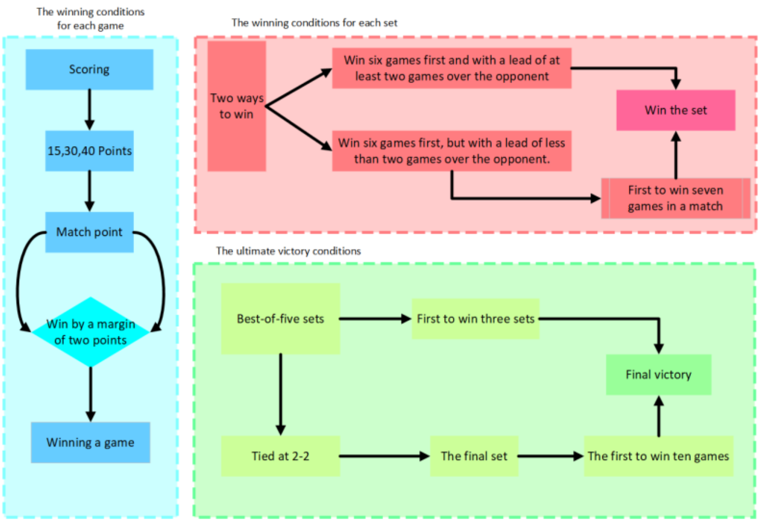

Figure 1: Tennis match rules

•As shown in figure 1 given that the match consists of a best-of-five set format, this paper initially designate two players as Player1 and Player2, recording their respective match scores as Score1 and Score2. At the beginning of the match, the initial score is set as Score1:Score2=0:0.

• During the process of receiving a serve, after each point, the scoring team's score is incremented by 1, and the score difference between the two sides is denoted as P'.

• At the end of each match, it is necessary to judge whether any player has won, and the judging conditions are: If Score1=40 or Score2=40, and the difference between the two is greater than or equal to 2, the office ends; If Condition 1 is not met, the match will continue until the next point appears.

• If the match ends and Player1 wins, then Player1 wins the match, then record the number of games won by Player1, game1, the initial value of match2 is 0; If Player2 wins, the number of wins is recorded in the same way[3].

• "To continue the match, reset Score1 and Score2 to 0 and begin the next set. Typically, when a Player reaches 6 points and $|game1\text{-}game2|\geq 2$, that player wins the set. In certain circumstances, if both sides reach a 6-6 tie, a tiebreaker is initiated. The conditions for the tiebreaker are as follows: when game1≥6 or game2≥6, the first player to reach 7 points wins the set and the match, i.e., game1=7 or game2=7."

• Similarly, record the score of each set for the two players and designate the initial set scores for Player1 and Player2 as set1=0 and set2=0. Ultimately, this paper can determine the conditions for the match's conclusion: when set1≥3 or set2≥3, the match ends, and the player who wins three sets first achieves the first victory in the match. Otherwise, the match continues to the next set, until one side wins three sets.

3. Predicting Tennis Momentum with LGBM

3.1. Determine index weights with Multiple Logistic Regression

To better characterize the momentum of various players given knowledge of match-winning conditions, this paper analyze all their eigenvalues and employ multiple logistic regression to extract the most representative eigenvalues.

Initially, to facilitate subsequent understanding, this paper identified 14 feature values associated with player momentum. The symbols and corresponding meanings are detailed in the following table 1:

Table 1 Four eigenvalues representing player momentum

Three-level index	Symbol
Whether it is the server	X1
The score of the match is ahead of schedule	X2
The number of matches won in this set	X3
Whether there were unforced turnovers in this match	X4
Whether there is a double fault in this match	X5
Whether to serve or not to score (no touch)	X6
Whether or not to score a return shot (no touch)	X7
No touch score forehand and backhand	X8
Total mileage in this match	X9
The ratio of internet usage time to internet score	X10
The ratio of points scored on the opponent's serve to total points	X11
Total mileage in the last three points	X12
The last point ran the graph mileage	X13
Real-time Serving Speed	X14

As different feature values undoubtedly impact a player's momentum to varying degrees, it is imperative to analyze these values and select those with a high correlation to momentum. In this section, this paper employ multivariate logistic regression to investigate the relationship between feature values and momentum. Multivariate logistic regression extends beyond binary classification, encompassing multiple logistic functions, making it more suitable for characterizing various player features. The hypothesis function of the multivariate logistic regression model is as follows:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_8 x_8) \tag{1}$$

where g(z) is the softmax function:

$$g(z) = e^{zk} / (1 + \sum_{i=1}^{8} e^{zi}) \tag{2}$$

In multivariate logistic regression, θ represents the model parameters, which denote the weights of the feature values corresponding to categories. Specially, θ includes an intercept term θ0, analogous to the bias in logistic regression; while θ1, θ2, ..., θn comprises the coefficients of the feature values, indicating each feature's contribution to classification. Upon input of a new feature vector x = [x1, x2, ..., xn], the model calculates the significance levels of different features based on these values and their corresponding θ values[2].

Figure 2: The confidence level of each feature value of the athlete

Based on Figure 2, it is evident that features such as x2 and x9 exhibit high confidence levels. As higher confidence in feature values corresponds to increased confidence in the results, it implies lower risk. To mitigate risk and enhance confidence in the results, thus obtaining a more accurate confidence interval, this paper choose eight features with high confidence, including x2 and x4, as influential factors in describing the momentum of athletes.

3.2. Training of the model based on the LGBM algorithm

To compare the training effects of different models, this paper trained the dataset's eigenvalues and constructed five models: XGB, MLP, LR, LGBM, and SVC. Upon inputting a specific player's eigenvalue into the assumption function of multiple logistic regression, this paper obtained the AUC values of the three models, as depicted in Figure 3. Notably, the LGBM model exhibited the highest AUC value, suggesting superior performance. Consequently, this paper selected the LGBM model for predicting player momentum and trained it using the LGBM algorithm.

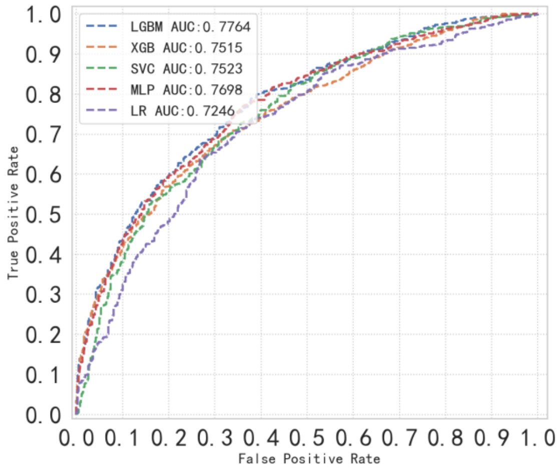

Figure 3: AUC curves for the three models

Light GBM is a gradient-boosting framework that uses tree tree-based learning algorithm. Assuming that for a certain player, their momentum is characterized by the real-time trend of the game, denoted as P. Based on the assessment metrics considered earlier, the trend of the game for different athletes can be represented as:

$$P_i = \sum_{i=1}^{n} *(Y_i, Y_i'(t-1) + f_t(x_i)) + C \tag{3}$$

Building an ensemble learning decision tree based on the LGBM model. Let f(xi)represent a single decision tree function. The feature values for a single decision tree have been extracted in Section 3.1. The paper selected 8 confident feature values to serve as branches integrated multiple decision trees and utilized the LGBM model to obtain the objective function for m decision trees:

$$Y_l' = \sum_{k=1}^{m} f_k(x_i) \tag{4}$$

Updating the real-time trend function for athletes. Next, this paper will use the GBDT algorithm to train the model. As our machine learning model is the LGBM model, the GBDT algorithm can utilize the negative gradient of the loss function to replace residuals and ft the residuals to obtain base learners:

$$W_t = \arg\min w_t \sum_{i=1}^{n} (Yi' - ht(x; wt))^2 \tag{5}$$

Where Yi represents the negative gradient. Through the iterative process outlined above, each iteration produces a weak classifier. Each weak classifier is trained using the negative gradient of the previous classifier until the best model is obtained. This yields the updated optimized model for LGBM, representing the real-time trend function for each player[1]:

$$P = \sum_{i=1}^{n} [2(Y_i'^{t-1} - Y_i) f_t(x_i) + f_t(x_i)^2] \tag{6}$$

3.3. Quantitative Analysis of Athlete Momentum

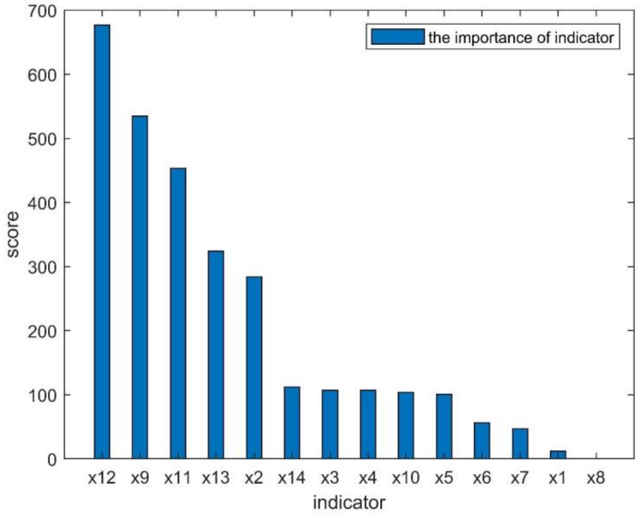

Figure 4: Graph illustrating the significance of momentum indicators in player scoring

Regarding the selection of momentum indicators for players, according to Figure 4, this paper can observe that the importance scores of the feature values such as the total mileage of the latest point (x12) and the total mileage of the match (x9) are relatively high, indicating that the total mileage of the athlete in the whole game is closely related to his momentum, which can indicate that the greater the mileage of the athlete, the greater the mileage of the match. The more tired they are, the more they can affect the momentum of the game. The score-leading progress (x2) of the game is also strongly related to the momentum, indicating that the athlete's score-leading state can affect the trend of the game, and the momentum is more inclined toward the athlete with a higher score. In addition, it is difficult to influence the momentum of the game with low-importance values such as forehand and backhand without touch (x8) and whether there are two turnovers in the game (x5). This paper selected the five high-scoring indicators x12, x9, x11, x13, and x2 to study the correlation between momentum and success.

According to the optimized LGBM model, this paper can analyze the momentum of athletes at specific times, as depicted in the following figure 5:

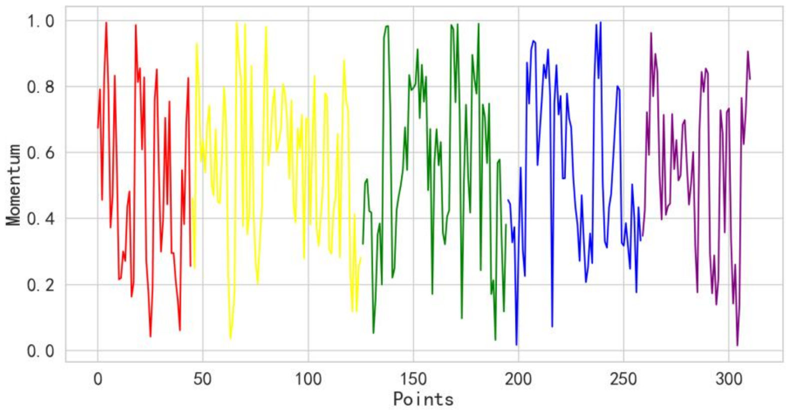

Figure 5: Real-time representation of athletes' momentum changes during a match

4. SVR Application in Tennis Momentum Analysis

4.1. Establishment of the SVR model

Selection of the kernel function. As the relationships between data points are non-

linear, this paper opted for the RBF kernel function, known for effectively handling non-linear correlations:

$$k(x,x') = e^{-gamma*\|x-x'\|^2} \tag{7}$$

In the given formula, x and x ′ represent two sample points. In this section, our goal is to achieve both excellent predictive performance for changes in player momentum and a high level of interpretability for the model. Regression Prediction of Game Momentum Shifts. This paper input the dataset into the SVR model and train it. Following this, our objective is to locate a hyperplane and define a perturbation term ε. Subsequently, this paper set the residual to 0 for data points inside the dashed line. Simultaneously, striving to minimize the total deviation of all sample points from the hyperplane, the SVR model aims to identify the optimal bar range. Regression is then applied to points outside this range to predict the game momentum shift.

4.2. Anticipation of Momentum Shifts in a Player's Game

In this section, this paper have chosen to provide a real-time visualization of a classic match featuring the 20-year-old Spanish star against the 36-year-old Novak Djokovic. The paper find Carlos Alcaraz's real-time momentum chart during the match:

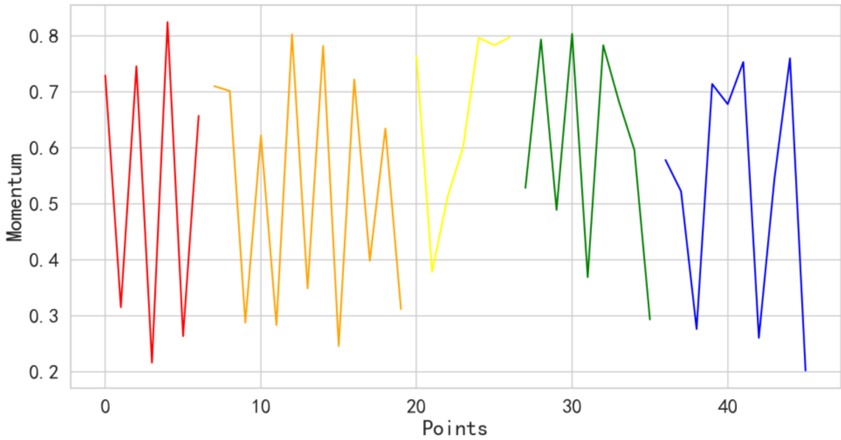

Figure 6: Real-Time Momentum Chart for Carlos Alcaraz in the Match

According to Figure 6, it is evident that Carlos Alcaraz's real-time momentum undergoes significant fluctuations throughout the match, with the highest volatility observed in the first two sets. The stability of the player's momentum initially rises and

Then declines, reaching its peak in the third set. This suggests that the player performed exceptionally well in this match, aligning with the conclusion drawn in the initial analysis. As the SVR model is unable to calculate the importance score of the index under the RBF kernel function, this paper employed the ID3 algorithm to determine the importance score of each index. Before computing the information gain, it is necessary to calculate the entropy and conditional entropy of each indicator influencing the momentum transition. To assess the selectivity of indicators, this paper calculate the information gained for different word features.

$$g(D,A) = H(D) - H(D \mid A) \tag{8}$$

Next, this paper applied the ID3 algorithm to assign importance scores to various factors influencing the player's momentum shift. The result highlighted several factors strongly correlated with this transition:

Table 2: Evaluation indicators of the five models

indicator	name	score
X2	The scoring lead for the game	0.09675
X7	Whether or not the game was scored on a return shot (no touch)	0.02478
X4	The ratio of internet surfing frequency to internet surfing score	0.01783
X14	Is this for real-time interaction with server-related items in this field	0.00526
X6	Whether the game is scored by serving (no touch)	0.00505
X5	Whether there were unforced turnovers in this game	0.00037
X3	Number of games won in the current set	0.00028

Based on Table 2, this paper observed that the factors with significant impact on momentum shift include the progress in scoring, the ability to counter-attack points (without contact), and the absence of unforced errors in the game. Among these, this paper selected the progress in scoring during the game as the most influential factor affecting momentum transition, furthermore, there is a shift in momentum from one player to another at the end of the first set.

Considering the varying dynamics of momentum in tennis matches, this paper recommend Players minimize unforced errors, manage psychological pressure when trailing in points, actively shift their mindset, and seize opportunities in serving, especially in the initial two sets to conserve energy. In the latter three sets, players should gradually unleash their full potential[8]. At the beginning of each game, as the opponent's momentum is likely at its lowest, utilizing this timeframe for optimal state adjustment is advisable. Additionally, players should be particularly attentive as each game approaches its conclusion because this is when the opponent's momentum tends to peak. Avoiding being influenced by the opponent's momentum during this critical phase is crucial in preventing a shift.

5. Conclusions

This paper predict Athlete Momentum Based on the LGBM Model. This paper conduct a multivariate logistic regression to assess the significance of the extracted dataset metric. Upon comparing significance, it is evident that the server is more likely to win. After plotting ROC curves for multiple machine learning algorithms, this paper select the LGBM model with the highest AUC value. Using the Gradient Boosting Decision Tree algorithm as the foundation, this paper train the model and generate a real-time momentum dynamic graph to describe the match flow. This paper use LGBM to score each indicator in the training data, with the scores reflecting the players' performance. Carlos Alcaraz performs better in the second, third, and fifth sets.

This paper analyze the dataset's significance using multivariate logistic regression, achieving a regression accuracy of 71.4%. Notably, previously insignificant metrics in the initial problem become significant. Following this, By inputting the aggregated metric dataset, this paper test the SVR model and identify a shift in player momentum at the end of the first set by predicting real-time swings in the match. Additionally, utilizing the ID3 algorithm to determine the importance scores for each metric, this paper find that the leading score progression in this game has the greatest impact on shifts in match momentum.

This study reveals that Momentum is real and has a big impact on the outcome of tennis matches which can help to advise tennis players during the match to scientifically improve the winning rate of the match.

References

[1] Kumar M A, Shweta P . Mitigating cyber threats through integration of feature selection and stacking ensemble learning: the LGBM and random forest intrusion detection perspective[J].Cluster computing, 2023.

[2] Kim S, Yoon H C, Lim J T . Productivity prediction in the Wolfcamp A and B using weighted voting ensemble machine learning method[J].Gas Science and Engineering, 2023.

[3] Ahamed B S, Arya M S . LGBM Classifier-based Technique for Predicting Type-2 Diabetes[J]. 2021(3).

[4] Zhang Q, Jiang Y, Duan S, et al.LGBM-improved-MLSE equalizer for 50Gb/s PAM4 IM-DD PON with 10G class optical transceivers[C]//Asia Communications and Photonics Conference.2020.DOI:10.1364/ACPC.2020.T4B.7.

[5] Pae H . Research Note: Differential Item Functioning and Unidimensionality in the Pearson Test of English Academic[J]. 2022.

[6] Tran T V H , Duong C D , Nguyen T H ,et al.UPPS impulsivity, entrepreneurial self- efficacy and entrepreneurial intentions among university students: ADHD symp- toms as a moderator[J].Journal of Applied Research in Higher Education, 2023, 15(2):473-495.DOI:10.1108/JARHE-12-2021-0464.

[7] Zhou H, Huang S, Zhang P, et al. Prediction of jacking force using PSO-BPNN and PSO-SVR algorithm in curved pipe roof [J]. Tunneling and underground space tech- nology, 2023.

[8] Munirathnam R , Vidya Y S , Manjunatha H C ,et al.Multifunctional Properties ofOcimumsanctumLinnLeavesMediatedSynthesisofNanoceria[J].MaterialsOpen, 2023, 01.DOI:10.1142/S2811086223500061.