An assessment measurement study based on the momentum of the tennis ball in question

Xiangzhi Wang[1, *], Ruize Xu[2], Jinyao Luo[3]

[1]Department of Business, Guilin University of Electronic Technology University, Guilin, China, 541004

[2]Department of Information and Communication, Guilin University of Electronic Technology University, Guilin, China, 541004

[3]Department of Electronic Engineering and Automation, Guilin University of Electronic Technology University, Guilin, China, 541004

* Corresponding Author Email: 17750258448@163.com

Abstract. In tennis, it is very important to always be aware of the movement and trend of the match. Accurately grasping the change of momentum is of great practical significance in predicting the future fluctuation of the match. Based on the relevant match data provided by Wimbledon, three models are constructed in this paper. Model one is a win-loss probability model based on binomial distribution, which deflates the impact of each small game win on the whole game, and accurately calculates the overall win rate of current players by judging the difference in players' level as well as the number of games won and lost. Model 2 is a momentum calculation model, which accurately describes the player's game status by analyzing the number of consecutive scores, score gap, number of key balls, and number of errors in a multi-dimensional analysis, and then accumulating and calculating the momentum data. Model 3 is a match fluctuation prediction model constructed based on Gradient Boosting Tree, which is utilized to predict match fluctuations by comparing predicted fluctuations with actual fluctuations on the test set in the tournament. And it evaluates which factor is the most important in predicting the future trend.

Keywords: Win-loss Probability Model, Momentum Calculation, Gradient Boosting Tree, Kendall Rank Correlation Codification.

1. Introduction

In tennis, it is very necessary to always know the trend and trend of the match, and the results of the match prediction are of great value and significance to the broadcasting of the tournament, sports betting, news reporting [1]. Most of the traditional methods are based on game data to construct a theoretical model of athletes' technical and tactical performance and game-winning [1], which has the defect of low accuracy. In the Grand Slam finals, there have been more than a few occasions of "let's go two, let's go three" reversal drama, when the players are in the development of the game to the critical moment, or the surrounding environment is gradually important, the sense of pressure will be accompanied by the emergence of [2]. This also shows that the game is not only influenced by the score but also by the mentality of the players. Based on this this text propose whether it is possible to analyze the players' momentum, to consider the influence of players' subjective factors to make more accurate predictions of the game.

2. Probabilistic Model of Victory or Defeat

2.1. Establishment of the model

Using the current score of the game, calculate the probability that the player will win in a specific time. In tennis, the general rules are as follows: In tennis, a best-of-five-set format is adopted. To accurately and intuitively describe the quality of a player’s performance in a game, this text only use the quantities most relevant to the player’s performance for interpretation. When a player has a higher

scoring rate in an inning than his opponent, his performance is often better than that of his opponent, so the scoring rate is the most important factor this text need to consider. By analyzing the data of 31 games in the series, this text can find that the probability of a player winning the point when serving is about twice as high as when receiving the serve. One of the most important actions in modern tennis is the serve [3].

When a player is the server, this text multiply his scoring rate by a this weight related to whether he is the server or not to obtain an accurate and true representation of his performance. The this weight generated by whether the server is the server σ can be statistically calculated through the data of the 31-game series. When serving, the value σ is approximately 0.654, and when receiving the serve, it is approximately σ = 1.346. So, by simply dividing the player's score D by the length of the game L is multiplied by the this weight to get the player's true scoring rate S

$$S = \frac{D}{L}\sigma \tag{1}$$

When evaluating player performance, this text need a visual model to help the audience understand the specific process of the game. The audience needs a more accurate and intuitive way to understand the direction of the game in addition to scoring, and the winning rate P can perfectly fulfill this task. Since different players have different skill levels, the winning rate of each game cannot be simply replaced by 50%. this text need X_1 to describe the player's expected winning rate. It can be obtained by the quotient of the player's total score in the game and the sum of the total game scores of both players. After obtaining these data, this text calculated the winning rate from the scale of the game. The standard binomial distribution is based on a sequence of Bernoulli trials. The probability of success is the same for each trial [5].

Through the Bernoulli trial binomial distribution formula, this text can derive the following winning rate formula for in the game.

$$P_r = \sum_{i=6-m}^{12-m-n} \frac{(12-m-n)!\,(X_1)^i(1-X_1)^{12-m-n-i}}{i!\,(12-m-n-i)!} - \frac{(6-m)!\,(6-n)!\,(1-X_1)}{(12-m-n)!} \tag{2}$$

P_r as the winning rate of the current set, it represents the impact of a game on the winning rate of the set. This quantity will be used in subsequent formulas as the intermediate value. After analyzing the winning rate of this set, this text need to analyze the initial winning rate at the beginning of each set. It is affected by the winning or losing of the previous set and can be obtained using the following formula.

$$P_n = \sum_{i=3-a}^{5-a-b} \frac{(5-a-b)!\,(0.5)^{(5-a-b)}}{i!\,(5-a-b-i)!} \tag{3}$$

The winning rate at the beginning of the set. P_n can be regarded as the initial value of the winning rate at the beginning of each set. As the game progresses, this text need to add the impact of each round on the winning rate. This winning rate can be obtained by scaling the winning rate of the current set. The shrinkage coefficient can be obtained as follows.

$$P_1 = U(P_r - 0.5)U(P_n - 0.5)(1 - P_n) \tag{4}$$

$$P_2 = U(P_r - 0.5)U(0.5 - P_n)(0.5 - P_n) \tag{5}$$

$$P_3 = U(0.5 - P_r)U(P_n - 0.5)(P_n - 0.5) \tag{6}$$

$$P_4 = U(0.5 - P_r)U(0.5 - P_n)P_n \tag{7}$$

Due to the existence of the step function, only one of the above four deflation coefficients has a value not equal to 0. This text can compute all the above intermediate quantities, which in turn leads to a real-time win rate per game P

$$P = P_n + 2(P_r - 0.5)(P_1 + P_2 + P_3 + P_4) \quad (8)$$

From this, the audience can obtain more clear and accurate information about the direction of the game through the real-time winning rate of each game.

2.2. Analysis of Results

The range of data this text have used is the Wimbledon tennis tournament numbered 1301 in the competition. The official data set has a total of 31 matches, because the selection of different venues does not have a significant difference in the final results, and the same can be done to achieve data accuracy and reliability. Therefore, this text randomly selected a game as data for analysis to form a visual graph. This can be used to judge the performance of the players at a certain moment and also to determine who has the greater advantage.

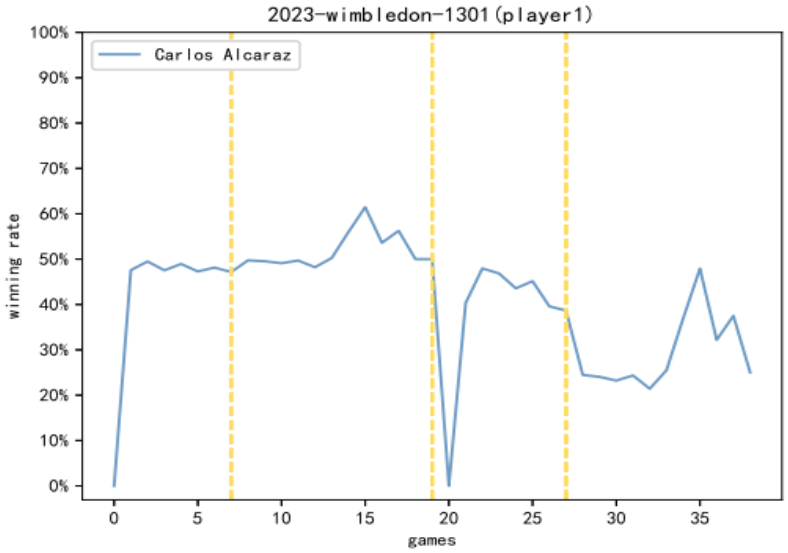


Fig. 1 Winning Rate

The Fig 1 shows the real-time winning rate P, which means that the probability of an athlete winning at a certain moment can be judged. The diagram is divided into four sections with dotted lines, that is, four sets of games. The trend of the above line graph clearly shows the trend of probability of the player winning a certain game.

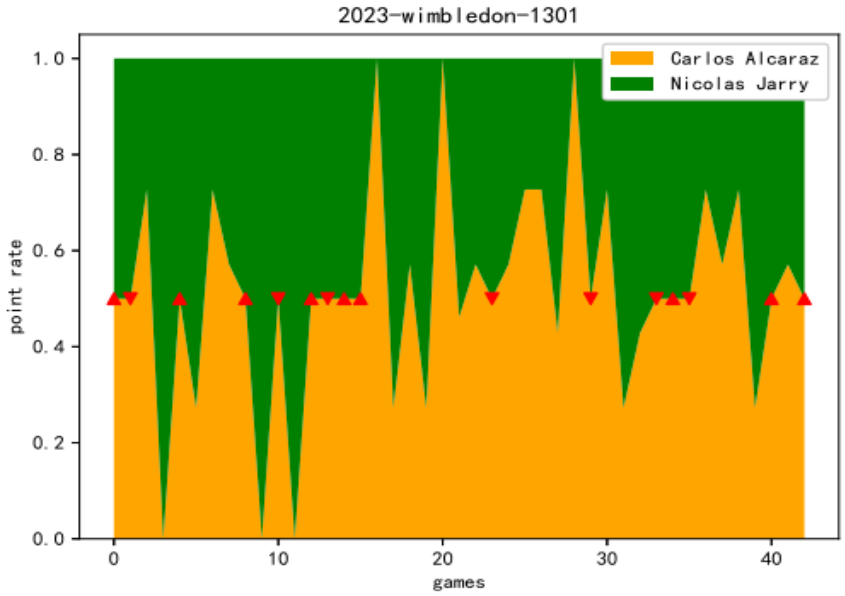


Fig. 2 Score PK

In Fig 2 , the opposite direction of the red triangle indicates which player is winning at a certain moment, and the green and orange parts represent the two players respectively. The figure shows that the factors that serve the server's advantage have been taken into account in the model.

3. Momentum Calculator Model

3.1. Model Structure

A player's momentum M_t is a very difficult quantity to calculate, and involves the complex psychological activities of the player. This text selected a few external conditions that are more likely to have an impact on player psychology for modeling. The momentum of the players in each round changes in real-time, and the momentum of the players in the previous round M_{t-1}will affect the mentality of the players in this round. When a player loses points, his momentum decreases, and when a player scores or scores consecutively, his momentum increases by leaps and bounds. Therefore, this text must first determine whether the player scores α. The player's continuous scoring behavior $C_t$should also have a momentum bonus. Considering that a player's winning streak does not increase linearly after a certain number of games, this text approximately consider the winning streak bonus to be in logarithmic form. Aces and breaks B_r in tennis A_c are exciting, but unforced errors can be frustrating, so by considering these dimensions together this text can come up with the following formula.

$$M_t = M_{t-1} + \alpha[1 + \ln(C_t + 1) + A_c + B_r] + (\alpha - 1)(1 + F) \quad (9)$$

Momentum is M_t defined as the cumulative value over time, α is whether a score is scored in this round. The score value is 1 and cannot be 0. It is a continuous score value, +0.3 for each consecutive score. Is it an ACE ball, if it exists, + 0.05. Is whether the serve is broken, if it exists, + 0.1. F is an active error, +0.1 if present.

3.2. Analysis of Results

Based on the momentum prediction model, the correlation coefficient is plotted in Fig 3. To verify whether there is momentum, this text need to perform a correlation test. Define D-mom and D-score. The chart below shows the correlation between the momentum spread and the point spread. As shown in the table above, the P value is significant ($P < 0.05$). The p value reflects the degree of data compatibility with the null hypothesis [6]. Therefore, it can be determined that the momentum difference and the score difference have a clear correlation. Therefore, the existence of momentum and scores is related, and it can be further deduced that the fluctuations in the game and the success of the players are not random phenomena.

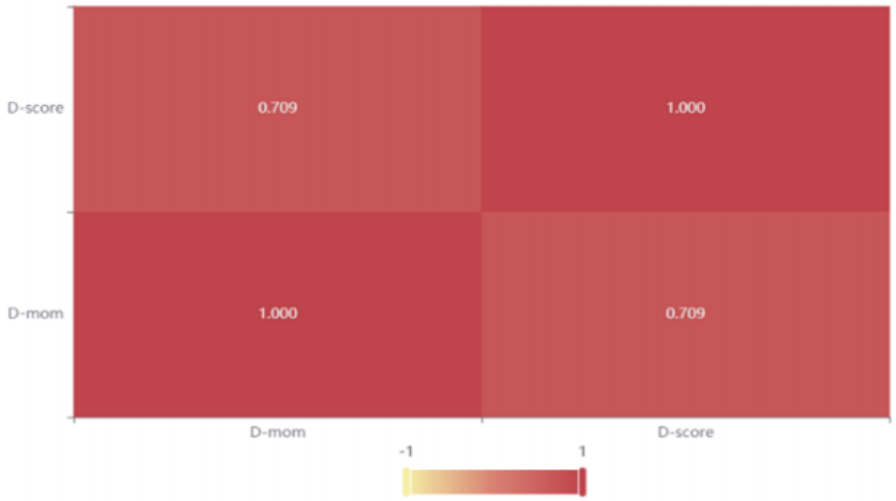


Fig. 3 Correlation coefficient heat map

4. Fluctuation Forecasting Model

4.1. Establishment of the model

The increase in the volume of structured and unstructured data related to more than just sport events leads to the development and increased use of techniques that extract information and employ machine-learning algorithms in predicting process outcomes based on input but not necessarily output data [6]. Different ML models will be assessed, along with different model designs and assumptions, to improve the predictive performance of the model [7].

The fluctuation prediction model mainly uses the following newly defined parameters. For all the following differences, if the value is positive, it means that player P_1 is in the lead; if the value is negative, it means that player P_2 is in the lead. P_1 player lever, use formula

$$M_t = M_{t-1} + \alpha[1 + \ln(C_t + 1) + A_c + B_r] + (\alpha - 1)(1 + F) \tag{10}$$

It is calculated that the Mt value is lever_p1. Player leverage difference is obtained by the difference between lever_p1 and lever_p2. Player momentum changes in the next five games . This data is extremely important. It is the only parameter that describes the fluctuation of the game. It can directly reflect whether a game is about to be reversed. It is also the Y-axis dependent variable in the image. This amount is processed by Excel data, and the player leverage difference in the future fifth round is compared with the player leverage difference in this round. For the last five rounds of the entire game, logical judgment is used to calculate the difference between lever_p1p2 of the last round and this round. P1 player's number of winning streaks Ct1, through Excel data processing, the winning data is judged and accumulated. Player winning streak difference delta Ct between Ct 1 and Ct 2. It can reflect whether a player is on a winning streak or whether the opponent is on a winning streak. It will also affect the player's value. Real-time difference by the difference between the two players' respective scores . which can reflect the score difference in real-time. Key ball difference, the key ball consists of a breaking ball, ace ball, and active fault. The specific algorithm is: (One's own break serve + ace ball + number of opponent's errors) - (one's opponent's break serve + ace ball + one's own number of errors) .This parameter is mainly used to introduce the impact of key balls on players' mentality. Player serving rate in the next five rounds : according to the model, this is the biggest factor affecting the direction of the next few games. It is derived from Excel data processing, using the server parameter and the five intermediate parameters in the table to perform logical operations and accumulation. It describes the p1 player's serving account in the next five games. Through overall data calculation, it can be found that the probability of the server winning the round is as high as 0.673, which is about twice that of the receiving player, so it is not unexpected that it becomes the factor with the greatest impact. CatBoost is a GBDT framework based on a symmetric decision tree algorithm. Gradient Boosting Decision Trees (GBDTs) have become very successful in recent years, with many awards in machine learning and data mining competitions [9]. Compared with GBDT, if the average value of the labels is used to represent the features, there will be a conditional offset problem when the da-ta structure and distribution of the training data set and the test data set are different. CatBoost is a member of the family of GBDT machine learning ensemble techniques [10]. CatBoost proposes that a standard way to improve Greedy TS is to add a prior distribution term, which reduces the effect of noise and low frequency category type data on the data distribution:

$$\hat{x}_k^i = \frac{\sum_{j=1}^{p-1} \left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] Y_{\sigma_j} + a \cdot p}{\sum_{j=1}^{p-1} \left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] + a} \tag{11}$$

Where p is the added prior, CatBoost utilizes a relatively novel method of calculating leaf node values in a way that avoids the problem of overfitting that can occur with direct calculations in multiple dataset alignments.

4.2. Analysis of Results

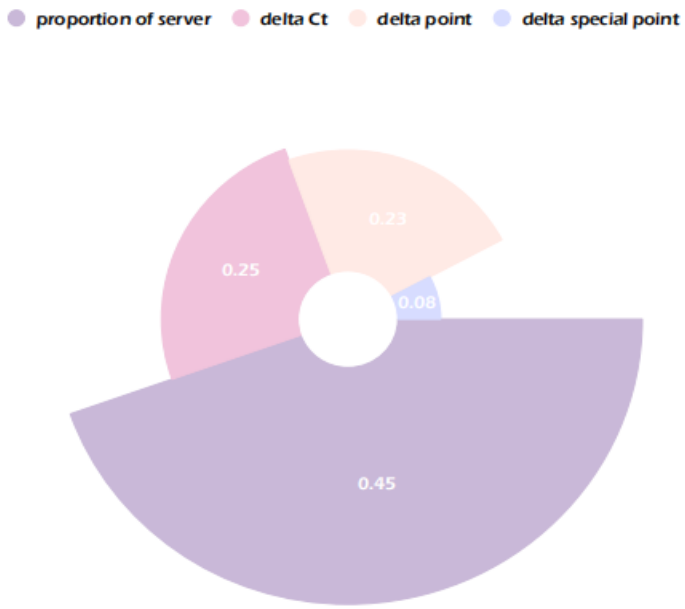


Fig. 4 Feature importance graph

From the results, it can be seen that the proportion of servers has the largest impact on momentum trend up to 44.6%, followed by delta Ct and delta Point. The delta special points are 24.9%, 22.7%, and 7.8% respectively. Therefore, becoming a server will profoundly improve the player's momentum.

5. Conclusions

The three models proposed in this paper to solve the main problem of the topic, some summarized points are as follows.

Model 1 addresses the problem of capturing the flow of momentum in a match, where the two calculations of scoring rate and serve this weights give the player's performance, which is finally represented using a visualization graph. Model 2 addresses the problem of assessing whether momentum is random. Model 3 solves the problem of predicting shifts in momentum during a match. The model used in this paper is generally simple in design, efficient, widely applicable, and innovative. Secondly, the calculation results of the model are highly accurate and stable. Its momentum calculation model fully takes into account the various factors affecting the player's momentum, non-linear processing of the parameters, and stronger practicality. The fluctuation prediction model simulates the fluctuation of the game in real-time for each round through the momentum, which is closer to the real environment of the game.

References

[1] C.Y. Huang. Research on tennis match performance prediction model based on BP neural network [D]. Fujian Normal University, 2022.

[2] Xiu Tingche, Yu Hongyan, Huang Wenyan. Construction and simulation application of table tennis excellent women's singles match result prediction model based on BP neural network[C]//Abstracts Collection (II) of National Sports Training Symposium 2022, Sports [3] Training Branch of the Chinese Society of Sports Science. Department of Physical Education, Shanghai Jiao Tong University;,2022:1.

[3] JIANG Haoqi. Construction of tennis simulation scenarios based on virtual reality technology induced "Choking" phenomenon[C]//Abstracts of the 2022 National Symposium on Athletic Training of the Chinese Society of Sports Science, Athletic Training Branch, China [Publisher's name], 2022:1. [Publisher unknown], 2022:1.

[4] Colomar, Joshua, et al. "Determinant physical factors of tennis serve velocity: a brief review." International Journal of Sports Physiology and Performance 17.8 (2022): 1159-1169.

[5] Malgorzata Miklas-Kalczynska, Taly Dawn Drezner & Zvi Drezner (2024) Experimenting with the generalized binomial distribution, Communications in Statistics: Case Studies, Data Analysis and Applications.

[6] Di Leo, G., Sardanelli, F. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. Eur Radiol Exp 4, 18 (2020).

[7] Horvat T, Job J. The use of machine learning in sport outcome prediction: A review. WIREs Data Mining Knowl Discov. 2020; 10: e1380.

[8] Chakraborty, Sanjay, et al. "Prediction of Winning Team in Soccer Game: A Supervised Machine Learning-Based Approach." Advances on Mathematical Modeling and Optimization with Its Applications. CRC Press 170-186.

[9] Gradient Boosting Decision Trees (GBDTs) have become very successful in recent years, with many awards in machine learning and data mining competitions.

[10] Hancock, J.T., Khoshgoftaar, T.M. CatBoost for big data: an interdisciplinary review. J Big Data 7, 94 (2020).