Momentum Analysis of Tennis Matches Based on Logistic Regression

Huanyu Liu*

College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 211106

*Corresponding author: liuhuanyu2002@163.com

Abstract. The paper first selected ten evaluation indicators such as first-serve scoring rate, double fault rate, and ace rate through a literature review. Then, using the entropy weight method, weights were assigned to each evaluation indicator. The top seven indicators with larger weights were selected as independent variables, with the pointing situation as the dependent variable, to establish a binary logistic regression model. The model achieved an accuracy rate of 94.3%. The paper multiplied the regression-derived weight matrix with the seven indicators, normalizing the resulting values to represent the probability of a player pointing probability. The slope of the pointing probability obtained in this way was defined as momentum. Based on the real-time changes in pointing probability during matches, the article calculated the average momentum for each match. By conducting a chi-square test between the direction (positive or negative) of the average momentum and the corresponding match results, with a significance level less than 0.001, the paper demonstrated a strong correlation between the two, suggesting that match outcomes could be predicted by calculating the average momentum.

Keywords: Tennis Match, Momentum, Entropy Weight Method, Binary Logistic Regression, Chi-square Test.

1. Introduction

With the rapid progress of science and technology, as well as the rapid development of the sports industry, data and information are growing exponentially. Scientific forecasting methods have been widely used in various fields. In recent years, sports events have employed various forecasting methods, such as the gray forecasting method and statistical regression analysis. The research of sports competitions through scientific prediction methods can provide valuable assistance to athletes in winning competitions. Therefore, the prediction of important sports event results has garnered increasing attention from sports researchers across the globe. As the second largest ball sport in the world, tennis enjoys immense popularity worldwide due to the exciting ups and downs in the competition process and the strong appreciation it receives. In large-scale events, the index data is highly detailed, allowing for the assessment of the technical level of the athletes on both sides in each match. By combining several mathematical models for match prediction, objective and accurate prediction results can be obtained, making it easier for people to appreciate and understand the matches. Currently, the main tennis prediction model consists of four main components: data collection, data processing, model training, and model evaluation.

Zhang [1] proposed a prediction method based on the Markov prediction model, which first utilizes factor analysis to extract important factors, then evaluates the strength of competitors, and conducts Markov prediction. Huang [2] established a tennis performance prediction based on the BP neural network algorithm. Zhao [3] proposed a prediction model based on the hitting rate per round and provided a formula for the player's winning rate.

Ma et al. [4] conducted a study on the 127 men's singles matches in the 2023 Australian Open Tennis Tournament. They utilized logistic regression analysis to establish a predictive model, achieving an accuracy rate of 89.4%. Therefore, this paper also employs logistic regression to construct a predictive model and explores the correlation between momentum and match outcomes based on this model.

2. Win Prediction Model Based on Entropy Weight Method and Logistic Regression

2.1. Data Source

The data was obtained from Problem C of the 2024 Mathematical Contest in Modeling (MCM), accessible at https://www.comap.com. It comprises data from the 2023 Wimbledon men's tournament, documenting a wide array of information including serving, scoring, aces served, and match outcomes.

2.2. The Process of Developing the Model

This paper proposes a prediction model of the pointing probability based on the entropy weight method [5] and logistic regression. First, the entropy weight method is used to screen the prediction indicators to determine the key indicators that have a greater impact on the win rate. Then, the data set was transformed into the training set and the test set, and the logistic regression model was used for predictive analysis of the screened indicators, and the winner was divided into one. After that, the value of $h(x)$ was normalized to the range of $(0,1)$, and its value was defined as the win rate of p2, and the predictive indicator at each point was calculated, the real-time win probability was calculated and the results were visualized. The construction process of the model is shown in the following Figure 1.

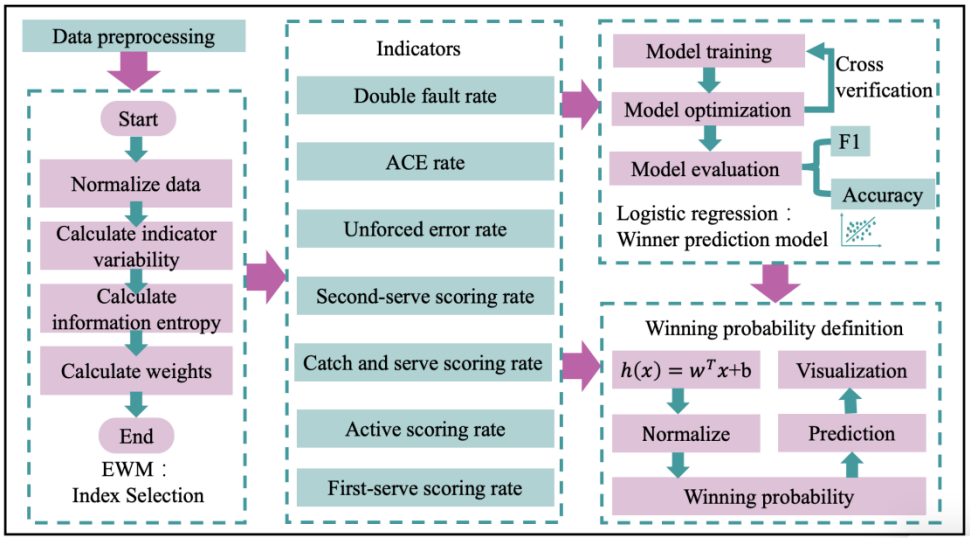


Figure 1. Win prediction model flow chart

2.3. Index Selection Based on Entropy Weight Method

In tennis, many technical indicators can have an important impact on the result of the match in different aspects. On the official website of the Professional Tennis Federation, there is match data for each player, which lists a total of 18 statistical indicators for singles matches. These indicators are divided into the following three categories according to different technical links.

Serving statistics include the number of aces, double faults, first-serve success rate, second-serve success rate, break points faced, breakpoint saving success rate, service game winning percentage, and number of service games played. Return statistics consist of scoring rates for one return and two returns, break chances, break success rate, service receiving game winning percentage, and number of service-receiving games played. Scoring statistics encompass serve scoring rate, serve receiving scoring rate, and total scoring rate.

According to the obtained data, the paper takes each set as a group and calculates the following ten indicators: serve percentage, serve success rate, second serve percentage, return percentage, total

percentage, ace service percentage, double fault rate, unforced error rate, active score percentage and break rate, totally 113 data set.

For these indicators, the entropy weight method of objective evaluation is used to assign weights to the values, and the results are as Table 1.

Table 1. Result of entropy weight method

	$P1$	$P2$
Double fault rate	0.38149	0.36212
ACE rate	0.20753	0.22108
Unforced error rate	0.09187	0.08603
Second-serve scoring rate	0.08547	0.08750
Catch-and-serve scoring rate	0.07011	0.08209
Active scoring rate	0.06552	0.70935
First-serve scoring rate	0.04181	0.03900
First-delivery success rate	0.03389	0.02854
Total scoring rate	0.01678	0.01735
Service break rate	0.00554	0.00535

Therefore, the following indexes are chosen: double fault rate, ACE rate, unforced error rate, second-serve scoring rate, catch-and-serve scoring rate, active scoring rate, and first-serve scoring rate. The following Figure 2 shows the correlation coefficient [6] of the indicator.

	First-serve scoring rate	Second-serve scoring rate	Catch-serve scoring rate	ACE rate	Double fault rate	Unf_err	Active scoring rate
First-serve scoring rate	1	0.2441	0.6057	-0.4533	-0.9144	-0.6747	0.5662
Second-serve scoring rate	0.2441	1	-0.2427	-0.5199	-0.4367	-0.2891	-0.2288
Catch-serve scoring rate	0.6057	-0.2427	1	-0.3412	-0.4605	-0.3195	0.4871
ACE rate	-0.4533	-0.5199	-0.3412	1	0.6459	0.1208	0.2654
Double fault rate	-0.9144	-0.4367	-0.4605	0.6459	1	0.5561	-0.3761
Unf_err	-0.6747	-0.2891	-0.3195	0.1208	0.5561	1	-0.2518
Active scoring rate	0.5662	-0.2288	0.4871	0.2654	-0.3761	-0.2518	1

Figure 2. The coefficient between the ten indicators

2.4. Logistic Regression Model Construction

Logistic regression [7] is a statistical method mainly used in binary classification problems, so it is very reasonable to predict the outcome of tennis matches. In general, the logistic regression model [8] looks like this:

$$h(x) = w^T x + b \tag{1}$$

$$\sigma(x) = \frac{1}{1 + e^{-h(x)}} \tag{2}$$

Here x is the indicator selected in 3.1, w is the indicator weight, and b is the constant.

The paper selected the first 70 sets of data as the training set and calculated the w of indicators in Table 2.

Table 2. The weights of the indicators

	$P1$	$P2$
Double fault rate	0.38149	0.36212
ACE rate	0.20753	0.22108
Unforced error rate	0.09187	0.08603
Second-serve scoring rate	0.08547	0.08750
Catch and serve scoring rate	0.07011	0.08209
Active scoring rate	0.06552	0.70935
First-serve scoring rate	0.04181	0.03900
First-delivery success rate	0.03389	0.02854
Total scoring rate	0.01678	0.01735
Service break rate	0.00554	0.00535

Using the remaining 33 data for prediction, the result is shown in Table 3.

Table 3. Result of Regression

Player	Real	Predicted	Accuracy
$P1$	19	18	94.7%
$P2$	16	15	93.8%

As can be seen from the figure, the prediction accuracy is 94.3%, and the F1 score obtained by calculation is 0.9375. Draw the AUC-ROC as following Figure 3.

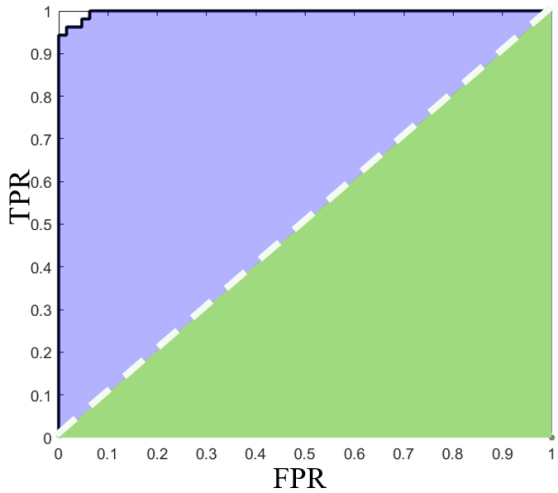


Figure 3. AUC-ROC of logistic regression model

2.5. Winning probability definition

Considering that the basic idea of logistic regression is to solve parameter values through maximum likelihood estimation and bring them into logical functions, it is reasonable to believe that

there is a strong correlation between the parameter values solved by maximum likelihood estimation and the outcome of the game, so p2 winning probability is defined as the normalized result of $h(x)$.

$$R_{P2} = \frac{h(x) - \min[h(x)]}{\max[h(x)] - \min[h(x)]} \tag{3}$$

And p1 winning percentage is

$$R_{P1} = 1 - \frac{h(x) - \min[h(x)]}{\max[h(x)] - \min[h(x)]} \tag{4}$$

At the same time, the paper calculates the indicators of each point and brings them into the formula to calculate the real-time victory ratio. The results are as follows. Figure 4 can be used to see the changes in the value intuitively.

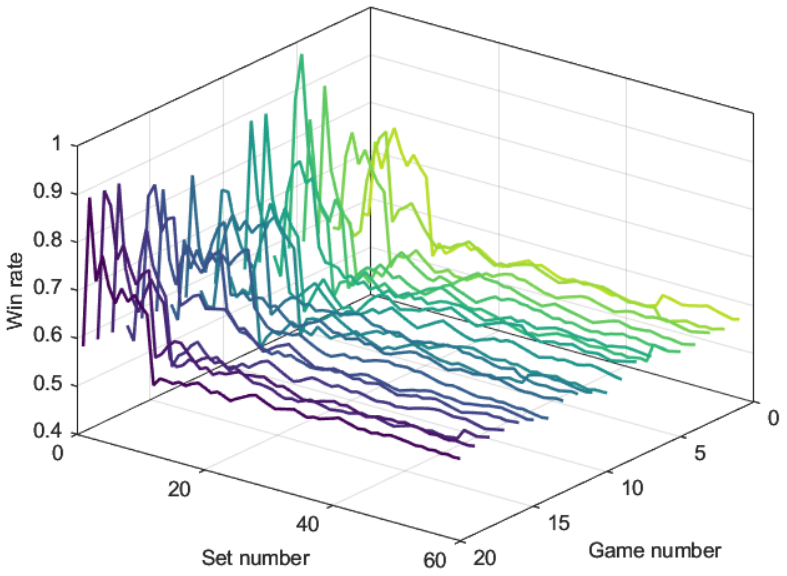


Figure 4. The result of the win prediction

3. Correlation Prediction based on the Chi-square Test

3.1. Analysis Approach

Another objective is to demonstrate whether momentum is related to winning the game. Therefore, the paper first gives the definition of momentum, calculates the average of momentum for each game, and uses the method of Chi-square test to explore whether momentum ownership is related to winning the game.

3.2. Momentum Definition

In physics, momentum refers to the tendency of an object to keep moving in the direction of its movement. Therefore, momentum is defined in tennis here as the tendency of a player to keep a winning state, determined by the ratio of the change in the winning rate of two adjacent points to the change in time. The formula is as follows:

$$I_{p1} = \frac{\Delta R_{p1}}{\Delta t} \tag{5}$$

$$I_{p2} = \frac{\Delta R_{p2}}{\Delta t} \tag{6}$$

The paper calculates the momentum evaluation factor of each point and calculates the average value of momentum in the same game. If the momentum evaluation factor p1 is greater than p2, the momentum in the game is considered to be p1, and vice versa. The results of the calculation are shown in Figure 5.

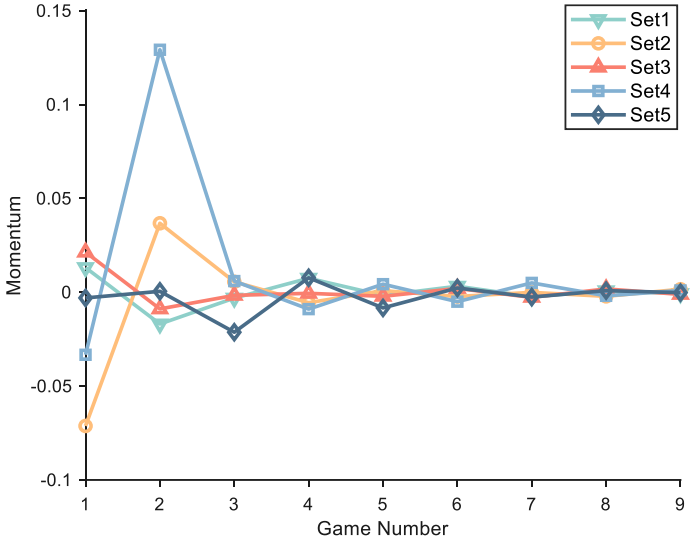


Figure 5. Momentum change

3.3. Correlation Prediction

The Chi-square test is a data correlation analysis method [9] based on the Chi-square distribution. Its basic idea is to compare the degree of coincidence or goodness of fit between theoretical frequency and actual frequency, which is mainly used to compare the difference between actual observed frequency and expected frequency. Therefore, the Chi-square test is chosen as the method to explore the correlation between momentum and winner.

Chi-square test steps include the following steps: calculate the expected frequency of each category, and calculate the actual frequency and expected frequency ratio, finally according to the ratio of chi-square value calculation.

The paper applied 1188 games to the chi-square test found in Table 4.

Table 4. Result of the Chi-square test

Player	Correct	Wrong	Pearson's chi-squared	Significance
$P1$	464	135	315.35	$P < 0.001$
$P2$	437	152		

Asymptotic significance is less than 0.001, so a strong correlation between winning and momentum can be demonstrated [10].

4. Conclusions

In addressing the predictive issue of tennis match outcomes, the paper established a tennis score prediction model based on the entropy weighting method and logistic regression, with an accuracy exceeding 90%. Subsequently, the paper defined momentum and conducted chi-square tests on the

direction of momentum and match outcomes, yielding a significance level below 0.001, thus verifying the correlation between momentum and tennis match results.

In practical applications, it is also necessary to consider the influence of other factors such as the weather on the day of the match, the physical condition of the players, and the surface of the court on the match. Therefore, it is advisable to introduce additional indicators for further investigation.

References

[1] Zhang Rong. Tennis match result forecast and player analysis [D]. Yunnan University, 2022.

[2] Huang Zhiying. Research on the Prediction Model of Tennis Match Results Based on BP Neural Network [D]. Fujian Normal University, 2022.

[3] Zhao Bin. The Bout System Method Application Forecast Model Research Tennis Men's Singles Competition [D]. Guangzhou Sport University, 2014.

[4] Ma Jiacai, Chen Guoqing, Liang Cairong, Zhou Zichen. Model Construction and Research Analysis of Winning Factors of Professional Tennis Players in Competitions [J]. Bulletin of Sport Science and Technology, 2023, 31(4): 66-68,118.

[5] Si Shoukui, Sun Xijing, Sun Zhaoliang, Zhou Gang, Si Wanling, Kang Shugui. Mathematical Modeling Algorithms and Applications 3rd Edition [M]. Beijing: National Defense Industry Press, 2021.

[6] Dufera, Abdisa G., Tiantian Liu, and Jin Xu. "Regression Models of Pearson Correlation Coefficient." Statistical Theory and Related Fields, vol. 7, no. 2, 2023, pp. 97-106.

[7] Pal, Ankita. "Logistic Regression: A Simple Primer." Cancer Research, Statistics, and Treatment (Online), vol. 4, no. 3, 2021, pp. 551-554.

[8] Harris, Jenine K. "Primer on Binary Logistic Regression." Family Medicine and Community Health, vol. 9, no. Suppl 1, 2021, pp. e001290.

[9] McHugh, Mary L. "The Chi-Square Test of Independence." Biochemia Medica, vol. 23, no. 2, 2013, pp. 143-149.

[10] Whatley, Melissa. "One-Way ANOVA and the Chi-Square Test of Independence." Introduction to Quantitative Analysis for International Educators. Springer International Publishing AG, Switzerland, 2022.