Research on Predicting the Winning Rate of Momentum Tennis Athletes

Jie He[1, *], Chengjie Yang[1], Mingcheng Cai[2]

[1]School of science, East China University of Technology, Nanchang, China, 330013

[2]Teachers' college, East China University of Technology, Nanchang, China, 330013

* Corresponding Author Email: hejie_ecut1@163.com

Abstract. In tennis, "momentum" usually refers to the momentum or trend formed by a series of events (such as consecutive scoring) in a game, which can have a significant impact on the outcome of the game. After gaining motivation, players or teams often exhibit higher morale and confidence, and are more likely to win consecutive points or games. Specifically, momentum may be reflected in indicators such as continuous scores and break points. In a game, the server is usually more likely to win points, so evaluating momentum requires considering the influence of various factors. The paper believe that athletes are primarily influenced by the psychological impact of momentum, but psychological factors are also an abstract concept. Therefore, the paper have turned our attention to the various behaviors of athletes on the field. Therefore, the paper used five indicators to quantify the momentum score of athletes. To address the coach's concerns, the paper use random testing and correlation analysis to verify the statistical significance of attempting to score and win games, rather than random events. This article uses learning algorithms such as decision tree models to predict the value of target variables at turning points based on feature data. Analyze the contribution of each feature to advantage transformation and ultimately identify the relevant factors that affect advantage transformation.

Keywords: Momentum Quantification, Permutation Test, Correlation Analysis, Decision Tree.

1. Introduction

Research has shown that tennis players can benefit psychologically or physiologically from momentum, including the number of rounds in a timed tour or psychological feedback ratings after various behaviors. However, most studies have focused on the correlation between athletes in one match and another, and the results are also based on collected on-site data rather than evaluating model results[1]. Here, this article uses permutation testing to generate multiple random datasets, and then uses a decision tree model to predict the results. Tennis matches are of great significance in the sports industry, as they not only represent sports competitions, but also represent the value of teamwork and personal growth. Tennis competition is a model of sportsmanship, encouraging people to challenge their limits and pursue excellence through competition. Tennis matches also demonstrate the importance of teamwork, including collaboration between players and coaches, medical teams, and other support personnel. Athlete's momentum usually refers to their performance and state during the competition. A good athlete's momentum usually indicates that they are in good condition, performing well, and may win the game. If the athlete's motivation is not good, it may indicate that their condition is not good and they need to adjust and recover. The issue of athlete motivation requires a comprehensive consideration of many factors, including physical fitness, technical skills, psychological well-being, and past competition results. In actual competitions, the importance of momentum is self-evident. The data in this article comes from: https://www.comap.com/.

2. Research on the Role of Momentum in Competitions

Momentum is very mysterious and it is difficult to quantify this concept. In the analysis of this article, it is believed that athletes are mainly influenced by momentum psychologically, but psychological factors are also an abstract concept. Therefore, this article focuses on the various

behaviors of athletes on the field. Because psychological factors that affect exercise behavior are often reflected in specific contexts. The corresponding positive feedback behavior will give athletes higher motivation.

2.1. Model building and solving

Summarize and calculate the momentum scores of athletes at each scoring point using the following formula:

It should be noted that at each scoring point, the composition of the above scores does not necessarily exist for every item, and calculation needs to be made based on the specific behavior of each game[2].

$$S_{momentum} = S_{foundation} + S_{holding} + S_{breaking} + S_{streak} + S_{important} \quad (1)$$

Where Smk is the momentum score at the kth scoring points and N represents the Total number of scoring points.

$$S_{summary} = \sum_{K=1}^{N} S_{mK} \quad (2)$$

Using the obtained formula, this article calculated the men's singles final of the 2023 Wimbledon Tennis Championships, which is the 1701 match described in this article. After quantifying various data, Fig.1 was drawn.

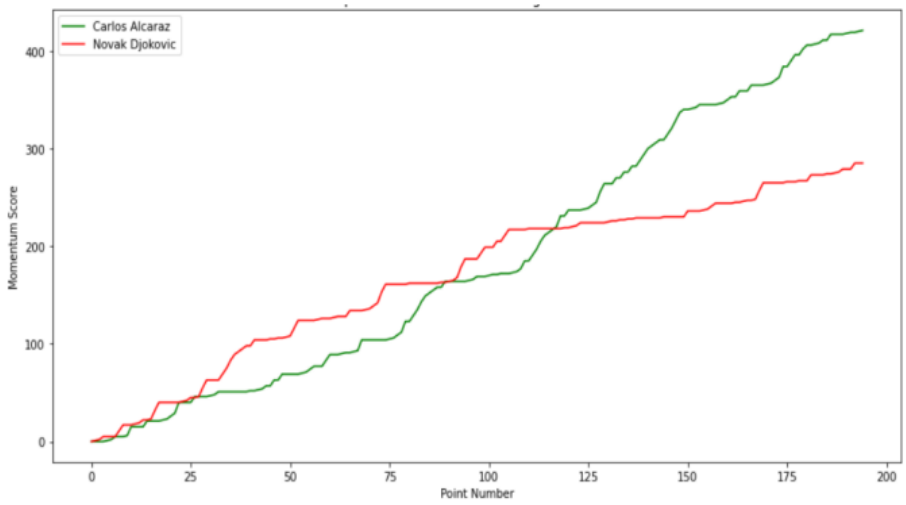


Fig. 1 Exhibition of Competition 1701

From Fig.1, it can be seen that the change in momentum is very consistent with the actual combat situation of the two individuals. From the beginning, Novak Djokovic had an advantage, until halfway through the game, Carlos Alcaras turned the tide and began to counterattack Novak Djokovic, ultimately achieving victory[3].

2.2. The Relationship between Momentum and Competition Victory

To address the coach's concerns, this article needs to explore whether the momentum calculated in the model is consistent with the game victory, which is a random phenomenon. Is there a relationship between momentum and game results? It is necessary to investigate whether there is a statistically significant correlation between momentum and corresponding matching results. This article adopts two methods for this:

This article applies statistical tests to compare the differences between simulated game data and actual game data, in order to evaluate the non randomness of momentum. The statistical testing method uses permutation testing. Check if the change in momentum score will exceed the level of change in simulated competition results under random conditions .

During the competition, use the momentum score formula to calculate the momentum score. And conduct statistical correlation analysis with the game results to indicate whether momentu m has an impact on the game results.

2.3. Model building and solving

Permutation test, also known as randomization test or re randomization test, is a statistical inference method mainly used to compare the differences between two or more samples and evaluate whether these differences are statistically significant. Taking the 1701 competition as an example[4].

The specific steps are as follows:

(1) Definition of null hypothesis: equality between sample groups, alternative hypothesis: differences between sample groups.

(2) Calculate the test statistic T_{origin} for the original observation data. The mean difference between two sets of data is taken as the original statistic.

$$T_{origin} = \bar{M} - \bar{W} \tag{3}$$

Where $\overline{M}$ and $\overline{W}$ are the average of the athlete's momentum score and the average of the number of wins, respectively.

(3) All sample data are merged into one large dataset.

(4) Randomly assign identical samples from each original grouping to each processing.

(5) Calculate and record the test statisticT_{new} for new observed data.

(6) Repeat steps 4 and 5 for each possible random allocation until all possible combinations of allocations are considered.

(7) Arrange the test statistics of each allocation combination in ascending order, which is the test distribution based on sample data.

(8) Calculate the p-value. If the p-value is less than 0.05, then the paper reject the null hypothesis that the differences between sample groups are statistically significant.

$$p = \frac{X}{\mathrm{Y}} \tag{4}$$

Where X represents the dataset $T_{origin} \leq T_{new}$ number of permutations for new, where Y represents the total number of permutations.

The basic idea of permutation testing is to randomly reassign samples, generate a large number of virtual samples, and then make statistical inferences based on these virtual samples to determine the statistical significance of the original hypothesis[5].

After calculation, the permutation test results showed a P-value of 0.1, indicating that in all permutation tests, the proportion of the highest momentum score consistent with the winner in the randomly arranged dataset was very small compared to the same proportion in the original data.

Meanwhile, the raw statistical data equal to 1 indicates that the proportion of the maximum momentum score matched with the winner of the competition in the raw data reaches 100%[6], as shown in Fig.2.

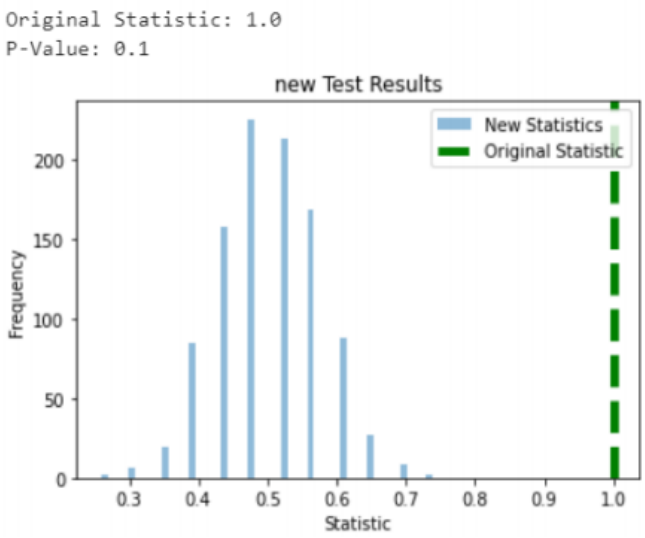


Fig. 2 Permutation test result

A P-value of 0.1 is very small, indicating that in the case of permutation testing, it is difficult to observe a proportion of the highest momentum score consistent with the winner, which is as high as the original data. This indicates that the original observation is unlikely to occur by chance, and the consistency between the momentum score and the winner of the competition is statistically significant rather than random. The results indicate that there is no random relationship between momentum score and game victory, and the role of momentum does indeed have an impact on the outcome of tennis matches[7].

2.4. association analysis

This article has fitted the momentum score and number of wins of Carlos Alcaras.

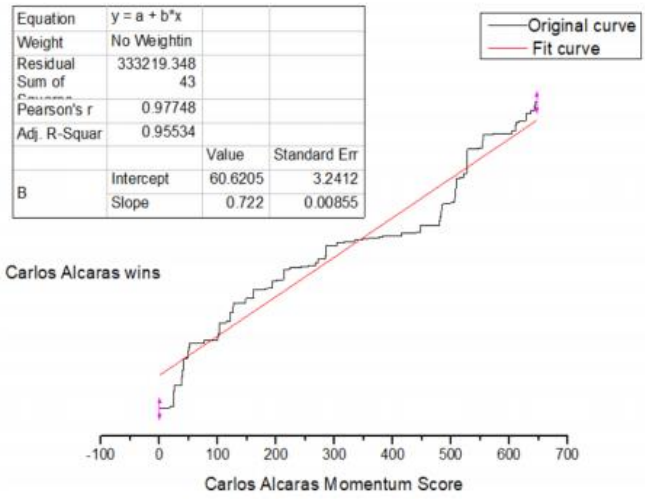


Fig. 3 Fitting curve

As shown in Fig.3, the R^2 statistical value of the curve is 0.95534, indicating a strong correlation between the two. The paper select game data, calculate momentum score data for each scoring point, then determine the winner of each game, calculate the maximum momentum score that two players can achieve in each game, and extract the maximum value for each group. Finally, analyze whether there is a correlation between momentum score and the winner of the competition[8].

$$M_{A1} = max(M_{A1}^K) \tag{5}$$

$$M_{A2} = max(M_{A2}^K) \tag{6}$$

In a game, the momentum scores of Athlete 1 and Athlete 2 at the k-th scoring point are MA1, MA1, so this article takes their maximum momentum scores as max (MAK1), max (MAK2).

Then, this article drew a scatter plot of the maximum number of pitches and the winner of the game for the two players. The red dot indicates that the winner of the game has the highest momentum score, while the green dot indicates the opposite.

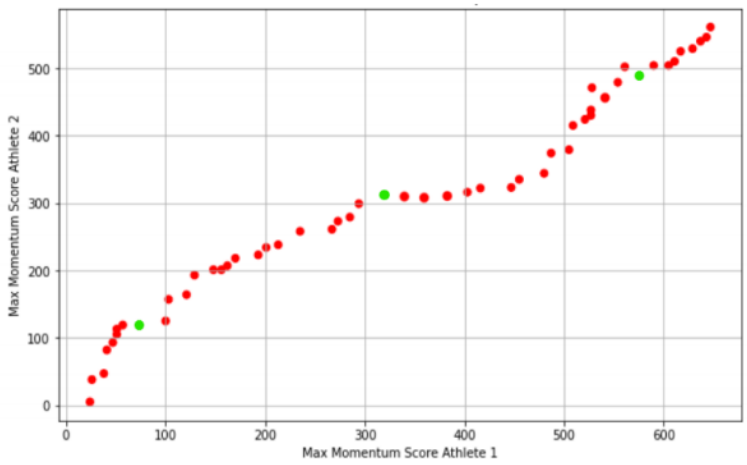


Fig. 4 Winners and Momentum Scores

As shown in Fig.4, most red dots indicate that high momentum scores are consistent with the winner, which also reflects a clear correlation between momentum scores and victory[9].

3. Advantage prediction

On this issue, this article needs to develop a model to predict the points of advantage transition in the competitive process. Here, learning algorithms are used to treat some feature data during the competition process as independent variables, and advantage turning points are treated as dependent variables. Subsequently, a decision tree model is used to predict the value of the target variable based on feature data. After the final analysis is completed, analyze the contribution of each feature to advantage transformation to obtain the relevant factors that affect advantage transformation. Finally, based on this, provide relevant suggestions for athletes.

3.1. Model construction and solution

Taking the 1701 competition as an example, a graph was drawn showing the relationship between t and the turning point.

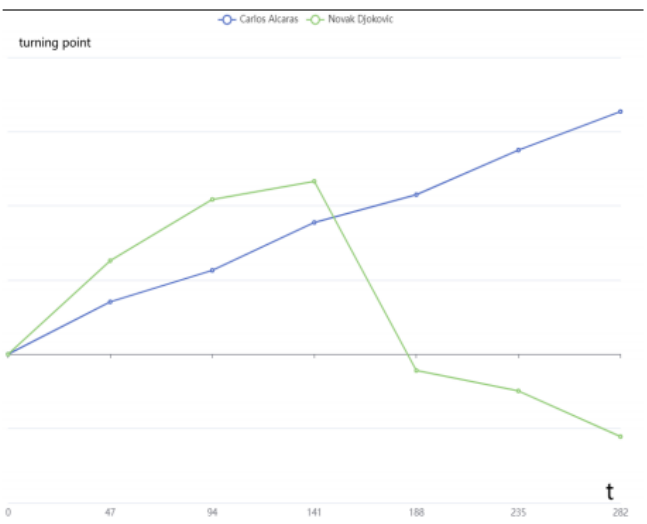


Fig. 5 Turning point

From Fig.5, it can be seen that the turning points of both are similar to the description of the problem, and the model is consistent.

3.2. Decision tree prediction

Decision tree is a non parametric supervised learning method primarily used for classification and regression. The goal is to create a model that predicts the value of the target variable by learning simple decision rules inferred from data features. The specific process is as follows:

Information gain, subtracting the information entropy before and after partitioning. The criteria for each division should choose the attribute that maximizes information gain to ensure that the data can be maximally distinguished[10].

Information entropy before partitioning:

$$H(D) = -\sum_{i=1}^{v} p_i log p_i \tag{7}$$

Information entropy after partitioning:

$$H(D^*) = \sum_{i=1}^{v} \frac{|D^i|}{|D|} H(D^i) \tag{8}$$

Among them, v represents the number of sub cases generated by the partition, while i represents the proportion of samples in class i to the total number of samples. Di represents the number of samples for the i-th value of the attribute.

$$Gain(D, a) = H(D) - H(D^*) \tag{9}$$

Information entropy represents the uncertainty of random variables in a dataset. A represents the properties used in this partition, which are various features mentioned in the previous section.

The gain ratio corrects the information gain preference, making the partition not biased towards attributes with more values.

$$Gain_{ratio(D,a)} = \frac{Gain(D, a)}{IV(a)} \tag{10}$$

where

$$IV(a) = -\sum_{\mathrm{i}=1}^{V} \frac{|D^i|}{|D|} log_2 \frac{|D^i|}{|D|} \tag{11}$$

The Gini index randomly selects two samples from a sample set with different probabilities and divides them using the same Gini index. The optimal partition properties also need to distinguish them and select the partition with the most index reduction.

$$Gini(D) = \sum_{k=1}^{n} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{\mathrm{n}} p_k^2 \tag{12}$$

Where pk represents the probability of selecting random samples as category k

$$Gini(D, a) = Gini(D) - \sum_{i=1}^{v} \frac{|D^i|}{|D|} Gini(D^i) \tag{13}$$

Further partition the generated child nodes and perform the above operations. Pruning, in order to avoid the endless growth of the decision tree, removes attribute forks with smaller contributions, after training, this article first obtained the relevant parameters of the model, as shown in Table.1and Table.2.

Table. 1 Related parameters

Parameter Name	Parameter values
Data segmentation	0.7
Node splitting evaluation criteria	Gini
Selection criteria for feature partitioning points	best
Maximum feature ratio considered during partitioning	None
The minimum number of samples for internal node splitting	2
The minimum number of samples for leaf nodes	1
The minimum weight of samples in leaf nodes	0
Maximum number of leaf nodes	50
The maximum depth of the tree	10
The threshold for impure node partitioning	0

Export feature contribution:

Table. 2 Feature contribution

Feature values	percentage
inning	2.9%
set	3.6%
serve	10.5%
consecutive point	11.5%
breaking point	12.5%
momentum difference	16.7%
breaking win	20.8%
turn point	21.5%

Comparison of Visual Function Contributions:

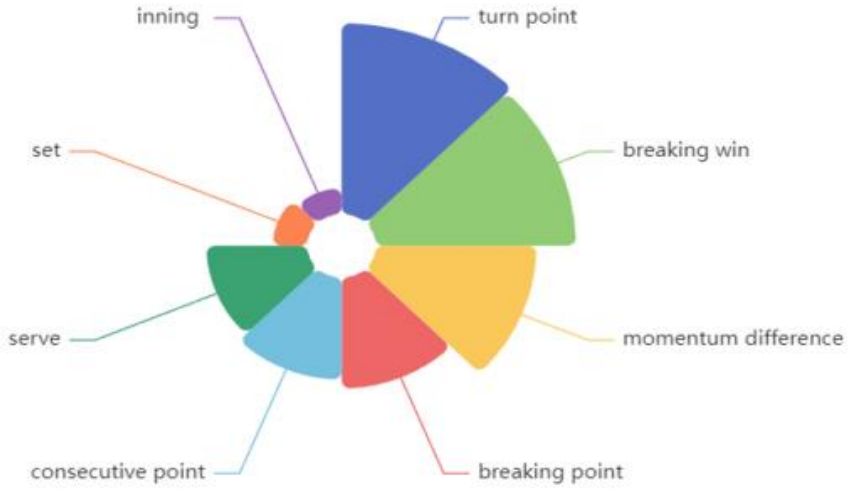


Fig. 6 feature contribution

Therefore, the factors that affect the transformation of competitive advantage and its contribution are shown in Fig.6.

Model evaluation:

Table 3. Evaluation

	accuracy	recall	precision ratio
Training set	1	1	1
Test set	0.99	0.99	0.98

According to Table 3, the classification evaluation indicators for the training and testing sets are quantified to measure the classification performance of the decision tree on the training and testing data.

4. Conclusions

This article uses momentum score and decision tree prediction models to study the results presented by momentum in tennis, which is more intuitive and makes it easier for coaches to explain to players, making players more motivated and informative in the game. The selection of influencing parameters is more scientific and reasonable because there is relatively less data required for training and prediction, making the implementation of the solution more convenient. Therefore, this article adopts a decision tree model to study this problem more reasonably.

In order to improve the accuracy of predicting the momentum of a tennis ball, the following methods can be used in the future to combine other types of algorithms with decision tree models, such as ensemble learning. This is a way to improve prediction performance by combining multiple models. For example, random forests can be used, which is an ensemble learning method consisting of multiple decision trees that improves the stability and accuracy of the overall model by averaging or voting on the predicted results of each decision tree.

References

[1] Dietl H M, Nesseler c. Momentum in tennis: Controlling the match [J]. international journal of sport psychology, 2017, 48(4): 459-471.

[2] Briki w. Rethinking the relationship between momentum and sport performance: Toward an integrative perspective [J]. psychology of sport and exercise, 2017, 30: 38-44.

[3] Kim I, Neykov M, Balakrishnan S,Wasserman L.local permutation tests for conditional independence [J]. annals of statistics, 2023, 50(6): 3388-3414.

[4] Hemerik J, Thoresen M,Finos L.Permutation testing in high-dimensional linear models: an empirical investigation [J]. journal of statistical computation and simulation, 2021, 91(5): 897-914.

[5] Dehghan S. Permutation tests of multivariate location using data depth [J]. journal of statistical computation and simulation, 2024.

[6] Johnson E N, Richter S J .Permutation tests for mixed paired and two-sample designs[J]. computational statistics, 2022,37(2):739-750.

[7] Toth D. A permutation test on complex sample data [J]. journal of survey statistics and methodology, 2020, 8(4): 772-791.

[8] Hu J,Chen Y, Leng C L, Tang C Y. Applied regression analysis of correlations for correlated data [J]. annals of applied statistics, 2024, 18(1): 184-198.

[9] Lipovetsky s. Canonical concordance correlation analysis [J]. mathematics, 2023, 11(1).

[10] Lu H F, Ma X, Hybrid decision tree-based machine learning models for short-term water quality prediction [J]. chemosphere, 2020, (249).