An Exploration of Momentum Changes in Tennis Players Based on CART Decision Regression Trees

Junjie Huang [1, *, #], Qiwei Li [2, #], Lisha Luo [1, #]

[1] School of Mechanical and Energy Engineering, Guangdong Ocean University, Yangjiang, China, 529500

[2] School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou, China, 510006

* Corresponding Author Email: 13427763688@163.com

[#]These authors contributed equally.

Abstract. As the sport of tennis is booming, the study of players' technical and tactical level and game status is becoming more and more important. For this reason, for the sake of the research of imposing changes. First of all, from the six tennis technical and tactical evaluation indexes of scoring change, error change, serve-receive scoring change, technical change, momentum change, and streak change, this paper derives the four characteristic indexes that are most relevant to the target variables, i.e., serve-receive scoring change, technical change, scoring change and streak change, through the CART algorithm. Among them, the correlation between the change in serve-receive score and the change in momentum was the greatest. Then, a decision regression tree model based on the CART algorithm was established. Then the appropriate decision tree model parameters were obtained after 10 rounds of data validation by the 10-fold cross-validation method. Among them, the depth is 5 and the minimum number of leaf nodes is 20. Meanwhile, to prevent the model from overfitting, this paper carries out a pre-pruning operation on the decision tree model. That is, estimation is performed before each node division, and if the division of the current node cannot bring the decision tree generalization performance improvement, the division is stopped and the current node is labeled as a leaf node to improve the applicability of the model. Finally, through the mean square error (MSE) and coefficient of determination (R-squared) test, it is found that the mean square error (MSE) is very close to 0, and the coefficient of determination R-squared is very close to 1, which indicates that the model has good prediction effect.

Keywords: Tennis Player, Momentum Change, Decision Regression Tree, Cross-Validation, CART Algorithm.

1. Introduction

With the rapid development of today's sports, in which tennis is popular, tennis players have higher and higher requirements for their performance. If you want to realize a breakthrough in the current performance, you can not leave the great role brought by information technology. Only through the description of information technology can provide the direction for the next decision, to make scientific decisions, need to use a variety of statistical means to explore the hidden laws and characteristics of the huge data[1].

Andrew Evans[2] and others studied the relationship between athletes' psychological momentum and performance, but the influence of performance is not limited to athletes' psychological factors. Wang Dongxue[3] and others analyzed the causative factors of Thiem's victory in the 2020 U.S. Open men's singles tournament. To make the study more generalizable, this paper should also extend the study to different matches and players in other leagues. Meng Fanming[4] and others used the decision tree method to study the winning factors of female tennis players, but the ordinary decision tree algorithm is less flexible and less efficient. For this reason, this paper finds that Zhang Kun[5] and others analyzed the college students' physical fitness test data by using the CART decision regression tree model. The decision tree model based on the CART algorithm is more flexible and efficient.

In summary, this paper will analyze the data related to the 2023 Wimbledon Men's Open by using the CART decision regression tree model, to explore the relationship between the changes in the data of tennis technical and tactical evaluation indexes and the changes in athletes' momentum.

2. CART Decision Regression Tree Modeling

2.1. Data source

The data were obtained from the following website, as shown in Table 1.

Table 1: Data source table

Data description	Data sources
Competition data	https://www.sofascore.com/zh/
Competition data	https://www.contest.comap.com/undergraduate/contests/

2.2. Fundamentals of CART Decision Regression Tree Modeling

The decision tree is an important classification model, that uses the principle of information theory to analyze and summarize the attributes of a large number of samples to produce a tree structure of classification rules, which uses the attributes of the samples as nodes and the attributes as branches CART (Categorical Regression Tree) Algorithm Tree, also known as Categorical Regression Tree, can be used for classification as well as regression[6]. The CART Decision Tree constructs the dataset by recursively dividing the dataset into two parts tree structure, with each leaf node representing either a category (classification task) or a value (regression task). Since the data metrics are continuous data, probabilistic assessments such as the Gini coefficient and entropy cannot be used as assessment metrics at this time[7]. Because continuous data can not use the classical generalization to find out the probability, this paper adopts the mean square error as the good or bad feature division, the division of each node after the sum of all the samples of the mean square error and the node not divided before the difference between the mean square error instead of the Gini coefficient.

(1)Select the optimal cut feature "j" and cut point "s"

$$\min_{j,s}[\min_{c1}\sum_{xi\in R_1(j,s)}(y_i-c_1)^2+\min_{c2}\sum_{xi\in R_2(j,s)}(y_i-c_2)^2] \tag{1}$$

"j" denotes the different features and "s" is the division point that corresponds to the available choices for each feature.

(2) Divide the region for (j, s) and decide the corresponding output values

$$R_1(j,s)=\{x\mid x^{(j)}\le s\} \tag{2}$$

$$R_2(j,s)=\{x\mid x^{(j)}> s\} \tag{3}$$

$$c_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i \tag{4}$$

$R_1(j,s)$ and $R_2(j,s)$ the data is divided into two nodes according to the cutoff point. c_m is the sum of the mean square error for each node of the solution and $x\in R_{m,m=1,2}$.

(3) Continue with step 1,2 for both sub-areas until the stopping conditions are met.

(4) The input space is divided into M regions, $R_1,R_2,\cdots\cdots R_M$ and finally, a decision tree is generated.

$$f(x)=\sum_{i=1}^{M}c_m I(x\in R_m) \tag{5}$$

$f(x)$ is the predicted value, which is derived using the mean value of the same node. $I(x \in R_m)$ is the region into which it is divided.

2.3. Calculate the correlation between each feature and the target variable

The dataset of "momentum" and other related feature indicators is divided and organized by Matlab, and the data are processed with missing values and outliers. Meanwhile, considering that the data of the decision tree model is not affected by feature scaling and is limited by linear relationships, the data of this model does not need to be normalized and standardized.

The correlation between each feature and the target variable is calculated by the CART algorithm, and based on the scores of these correlations are ranked, and finally, a few feature metrics with the highest correlation are selected[8]. In this paper, by importing the SelectKBest class from the Python library sci-kit-learn feature library, which is used for feature selection, it helps to select the K most relevant features to the target variable from the set of features to evaluate the contribution of each feature metrics to the mean squared error (MSE) and the ranking operation[9]. As can be seen in Figure 1, the top four feature metrics associated with momentum change are ranked in order of importance and relevance as follows: change in serve/receive scores, change in technique, change in scoring, and change in number of consecutive sets.

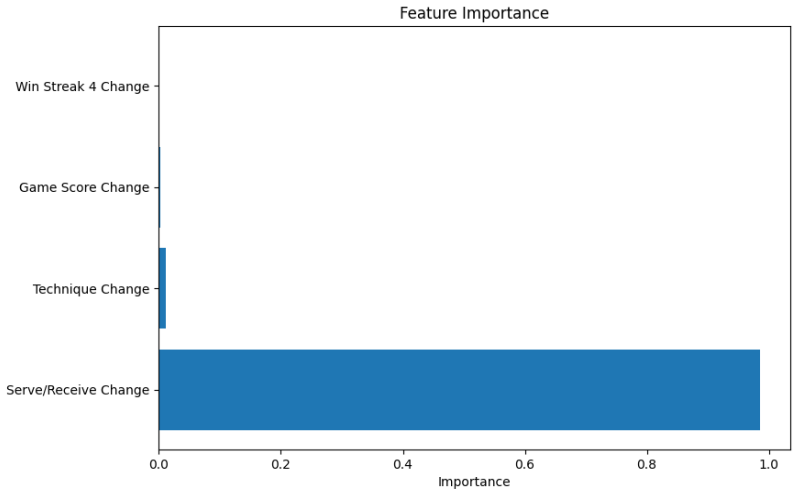

Figure 1: Importance of the indicator

To control the complexity of the decision tree, prevent overfitting of the training data, and improve the model's ability to generalize to new data, the model performs a pre-pruning operation on the decision tree, which reduces the generation of branch nodes by limiting the depth of the model tree and the smallest leaf node.

2.4. Pre-pruning and parameter tuning

To control the complexity of the decision tree, prevent overfitting of the training data, and improve the model's ability to generalize to new data, the model performs a pre-pruning operation on the decision tree[10]. Which reduces the generation of branching nodes by restricting the depth of the model tree and the smallest leaf node. This paper uses a 10-fold cross-validation method to derive the most appropriate training data with a depth of 5 and minimum leaf nodes of 20.

The mean square error (MSE) indicates the degree of difference between the estimated and the estimated quantity. Let t be the overall parameter determined from the subsample θ. The mathematical expectation of the estimate of t is called the mean square error of the estimate t. Usually, the smaller the value, the better the fit of the model to the data. It is equal to, where and are the variance and bias of t, respectively.

$$MSE(\theta) = E(\theta - t)^2 \tag{6}$$

Coefficient of Determination (R-Squared): A numerical characterization of the relationship between a random variable and multiple random variables and a statistical indicator of the reliability of a regression model in explaining changes in the dependent variable. The R-squared, which ranges from 0 to 1, indicates the proportion of the variance of the dependent variable that can be explained by the independent variables. A coefficient of determination of 1 indicates that the model fits the data perfectly, while a coefficient of determination of 0 indicates that the model does not explain any variation in the dependent variable.

$$r^2 = \frac{\sum_{i=1}^{n}(\mathrm{Y}_i - Y)^2}{\sum_{i=1}^{n}(\mathrm{Y}_i - \overline{Y})^2} \tag{7}$$

$\sum_{i=1}^{n}(\mathrm{Y}_i - Y)^2$ is the residual squared and $\sum_{i=1}^{n}(\mathrm{Y}_i - \overline{Y})^2$ is the residual squared sum.

As shown in Figure 2 and Figure 3, when the depth of the model tree is 5 and the smallest leaf node is 20, the mean square error (MSE) is very close to 0, and the coefficient of determination R-square is very close to 1. It can be verified that the parameters of the decision tree are adjusted very appropriately, and the model works very well.

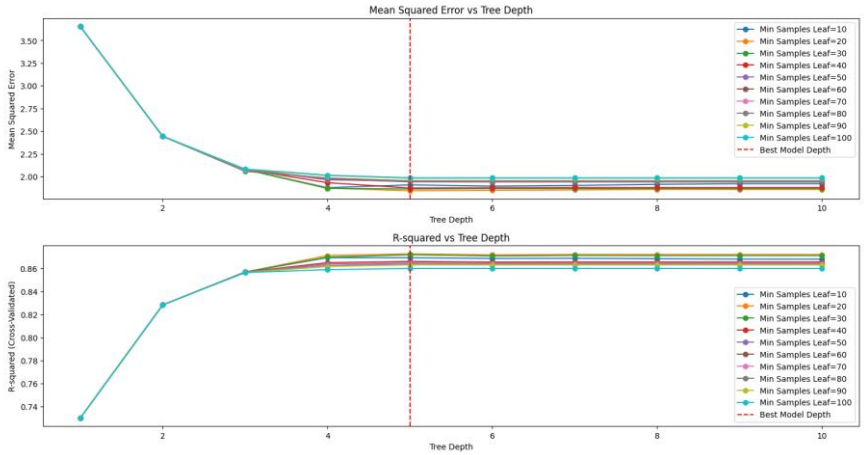

Figure 2: Mean square error and coefficient of determination for different parameters

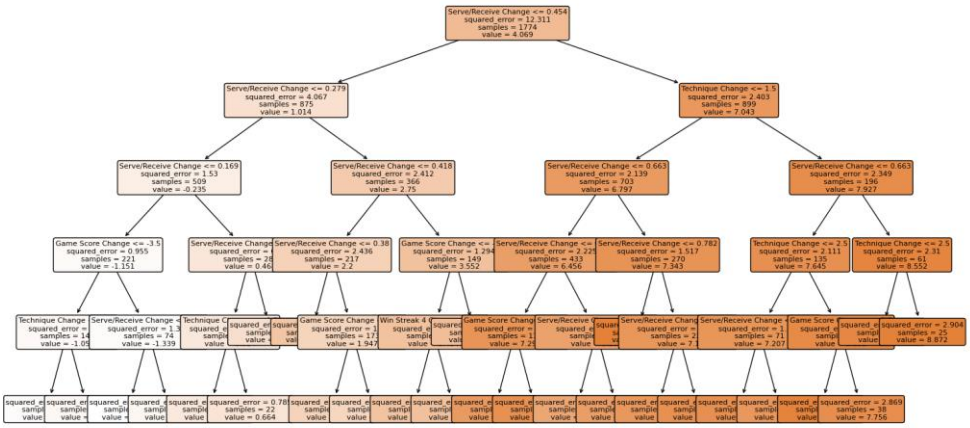

Figure 3: Decision tree

3. Results

3.1. Correlations between features and target variables

Of all the characteristics, change in serve/receive score is the one that correlates to the greatest extent with change in player momentum. This was followed by changes in technique, then changes in scoring, and finally changes in the number of consecutive sets.

3.2. Decision tree related parameters

Decision tree model parameters, as shown in Table 2.

Table 2: Model parameter

parameters	Data
depths	5
leaf minima node	20

3.3. Advice for the players

(1) Pay attention to the timing of momentum changes of both sides in a match

Pay attention to the timing of changes in momentum of both sides in previous matches, especially at key moments or key innings. If there are obvious changes in momentum and key factors, tactics can be adjusted according to these timing and factors to maintain or change the trend of the match.

(2) Analyzing the characteristics of momentum change nodes in opponent's matches

Understand the performance of the opponent in different situations by studying the opponent's historical match data and momentum change characteristics. Identify the opponent's weaknesses to exploit them in the match. At the same time, analyze the opponent's strong points to take corresponding countermeasures.

(3) Increase the pair serve training

For the best men's professional tennis players in the world today, all of whom possess excellent serving techniques and tactics, holding serve has become easier and easier. The focus of the match is how to break the opponent's serve. If a player is dominant on serve, then they can be more aggressive and aggressive in their attack during the match. Utilizing powerful serves and clever serving positions can increase the chances of winning. If a side has an advantage in serve reception, they can focus more energy and win more points in a serve-receive environment, making up for their deficiencies in serve in a sensible way, and building on their strengths and weaknesses.

4. Conclusions

This paper analyzes the data of the 2023 Wimbledon Men's Open and concludes that the change of serve/receive score is the most relevant technical and tactical evaluation index with the change of player's momentum among all the technical and tactical evaluation indexes of tennis. A decision tree model based on the CART algorithm with a depth of 5 and a minimum leaf node of 20 is successfully established, and the model is verified by the correlation test, which shows that the model is very effective. This paper expands the scope of analysis, not only analyzing the data of a certain game but also analyzing the data of the whole league. The CART algorithm is flexibly adopted to optimize the decision tree so that the model is more flexible and applicable. At the same time, to prevent the model from overfitting, this paper also carries out a pre-cutting operation on the model, which improves the generality of the model.

Using the model established in this paper can well help tennis players predict the momentum change of themselves or their opponents in advance during the match and make timely adjustments to improve the winning rate of the match. Of course, this paper only studied the relevant data of men's singles, and the follow-up should be followed by studying the data of other forms of matches such as

women's singles. To improve the shortcomings of the model, and improve the generality of the model, to make more scientific predictions for tennis players.

References

[1] Zhang Bing. Analysis of Tennis Technique and Tactics under Statistical Perspective[D]. Shandong University of Finance and Economics,2022.

[2] E. A E, Paul C, Y. S S. Psychological momentum among non-experts: Evidence from club golfers[J]. Journal of Behavioral and Experimental Economics,2023,104.

[3] Wang, D. X. Analysis of Thiem's winning factors in the 2020 U.S. Open men's singles tournament[D]. Harbin Institute of Physical Education and Sports, 2022.

[4] MENG Fanming, HUANG Wenmin. Quantitative analysis of winning factors of female tennis players based on decision tree method[J]. Journal of Jilin Institute of Physical Education,2019,35(05):43-47.

[5] Kun Zhang. Research on the application of decision tree CART in the analysis of college students' physical fitness test data[J]. Sichuan Sports Science,2022,41(01):41-45.

[6] H Wang, Q Zhang, Y Wang, et al. An improved CART decision tree regression algorithm based on ELM[J]. Computer System Applications,2021,30(02):201-206.

[7] Yafen Liu. GA-based CART decision tree improvement algorithm and application[D]. Guangzhou University,2021.

[8] JU Meixin, ZHOU Yuxin, TANG Weining, et al. A CART decision tree power usage anomaly detection method based on data imbalance[J]. Jilin Electric Power,2024,52(01):53-56.

[9] Jian Song, Kaiqiang Meng, Shuwen Shen, et al. Constructing machine learning based on Python to explore the regularity of the decision tree of the pathway of traditional Chinese medicine for gastric cancer[J]. World Journal of Integrative Medicine,2024,19(02):225-232.

[10] Zheng Lijia,Song Bing. Pre-pruning and optimization of decision tree classification algorithms[J]. Automation Instrumentation,2023,44(05):56-62.