WEP Warwick Evans Publishing

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

Tennis Game Swing Prediction Model Based on ID3 Algorithm and Logistic Regression Model

Yufeng Liu[#], Yiming Ma[#], Shichao Liu[*, #]

Arizona College Of Technology At Hebei University Of Technology, Hebei University Of Technology, Tianjin, China, 300401

*Corresponding author: shichaoliu_1@163.com

#These authors contributed equally.

Abstract. Tennis, as one of the most popular sports worldwide, includes exciting matches between players. In some matches, score advantage can swing constantly from one side to another. These score swings may impact players' performance and the result. To accurately predict the swings during a match between players, the article aims at building a score swing prediction model based on ID3 algorithm and logistic regression algorithm. The article will focus on analyzing tennis players' behaviors and performance during a match. The preparation is to form a decision tree to evaluate diverse events' impact on players' mental strength, using ID3 algorithm that can analyze desired indicators with relevant data. By using the results from the decision tree evaluation, indicators with higher priority will be analyzed by the logistic regression model. The following Omnibus test and Hosmer-Lemeshow tests will estimate significance value and prediction accuracy. The results can show the prediction of scores swings in a match and indicate the factors that impact the match more. The study can provide tennis players with theoretical advice on their training activities, which may improve the ability of reversing the match and being consistent under pressure effectively.

Keywords: Decision Tree, ID 3, Logistic Regression Model.

1. Introduction

In the 2023 Wimbledon Gentlemen's single, 20-year-old Carlos Alcaraz from Spain, caused an upset by winning legendary player Novak Djokovic. Spanish lost the first set completely but somehow controlled the next two sets gradually. With Djokovic winning 6 of 9 games in the 4th set, Alcaraz miraculously won the final set at 6-4.

The score swing describes a situation where the score advantage shifts back and forth between players. The score swings during a match can significantly impact the performance of players both psychologically and physically, which further can potentially alternate the match result. Hence, score swing prediction is crucial for players to prepare themselves to maintain performance without being inflicted by score swings. The factors that lead to swings will be analyzed and specifically. Mental strength, which indicates the performance boost by implementing events or motions by players during the match, will be considered in prediction.

The prediction model is established by following two parts, which are ID3 algorithm training for quantifying mental strength levels and logistic regression model for predicting potential game swings. The preparation is to establish related performance index system as the composition of the decision tree. After applying the data corresponding to the index system, ID3 algorithm can build a metric to evaluate various conditions for different outcomes. Based on the outputs from ID3 algorithm, we select indicators that are highly related to game swings and apply logistic regression model to predict swing occurrences.

In the discipline of tennis match prediction, previous studies mainly focused on predicting the future career results of players through their match results and characteristics, and did not predict the final trend of the match and the transformation of the advantages and disadvantages of the match through the characteristics and performance of players in a single match. This article deploys ID3 algorithm

evaluate the advantage level based on their performance. The logistic regression model in this article predicts score swings between advantage and disadvantage players. Logistic regression model integrates and modifies the results of ID3 model to ensure the accuracy of the model's prediction of competition results.

2. Method

2.1. The Build of Decision Tree

The performance decision tree model allows us to evaluate the mental strength and detect the swings of match. Decision tree also can filter the performance index conditions from massive data and quantify the state. Decision tree model is a supervised learning algorithm, including root, leaf, and branch like a physical tree, which build decision-making courses and different results. From the data, we obtain the match information including scores, servers, moving distance, hitting speed, serve width and depth, return depth, etc. The performance of tennis players not only is reflected by scores, but also can be evaluated by multiple factors that constantly vary during a play. To build a relatively reliable model, all factors that impact players' performance will be considered as evaluating indices. To systematically organize and implement the data to build the analyzing model to identify which player performs better and how much better, we deployed the combination of decision tree and ID3 neural network model. By using the model, we implemented the index system and calculate the performance levels for each time interval and the scale of the performance difference.

Before building the decision tree model, index system needs to be established first. During a tennis match, diverse factors can cause the swings of play and even reverse the results. Influencing factors are usually considered as physical strength, psychological strength, strategic skills, and environmental variables, which provide the theoretical standard for the index selection [1]. In order to enhance the generality of the model, environmental variables, including the surface materials of the court, wind direction and speed, etc. are not considered in the model. Also, by referring to the Wimbledon 2023 Gentlemen's singles matches data, indices are selected and further explanation for index understanding is shown below

- Server: Players who make the initial shots in a match are called server. They are given two chances to put balls in play for each point and if they miss them all, the point will be automatically given to the returners. Generally, the servers are thought to have better chances to win the points.

- Percentage of scores won by one player of total scores: Scores can directly reflect the performance of both players in a match. We will use the score percentage of each player in a certain time interval to evaluate the performance of players at a specific time.

- Double fault: A server fails to serve for two consecutive balls and the point will be given to the returner. Double fault is rare in professional matches. However, if this happens, it will put significant stress on the server and decrease the mental strength consequently, increasing the likelihood of swings of the play.

- Moving distance: The physical strength of players is constantly consumed during a match while moving a further distance than the opponent can consume more energy. When reaching physical limits, the player starts to feel exhausted and thus, performance drops.

Some of the indices above are directly recorded during the match, which could be used as index for the evaluating system by implementing a series of calculations, which is to mathematically normalize the data to eliminate different dimensions and units. While some indices are not directly recorded, they can be obtained by comparing scores of two players. For example, to calculate the percentage of scores won by one player of total scores, we will sum up the winning scores in a set and divide it by total scores. Detailed process will be shown in sec.3.1.

After the normalization and standardization, data sets will be imported to ID3 neural network algorithm to organize the order and conditions on the branches of the decision tree. It should be noted

that double fault is discarded by ID 3 training model due to its rarity in a high-level match and thus, negligible impact on analyzing mental strength trend. Ultimately, ID3 helps build the decision tree that quantify the mental strength based on players' real-time performance in the match [2].

The outputs from decision tree will be used as the major reference to evaluate the mental strength during the match. Normalized values of each index are the input for the decision tree. After the judgement for each branch, the output values of mental strength can be obtained as -1, 0, or 1, meaning the relative strength of psychologic condition at each time interval. Mental strength values can be accumulated chronologically to build an overview of mental strength trend. To evaluate the consistency between the mental strength trend and match results, we take the trend curve as the indication of predicted match results and compare them with actual match results. The accuracy will indicate the impact of mental strength on match results. To some extent, it also reflects the game swing as the mental strength indicates players' performance and scores.

2.2. The Build of Logistic Regression Model

To evaluate the game swing occurrences in matches, logistic regression model was chosen as the analytical tool. The logistic regression model, as the variant of generalized linear regression, can deeply study independent variables, whether continuous or categorical, in relation to a dichotomous dependent variable [3]. To enhance the precision of predictions regarding advantage shifts in game, we dived deeper into players' on-court behaviors to select five indicators, which have profound impact on players' mental strength and performance. The indicators include

- Holding serve: Holding serve means that the server, who hits the first shot, wins the game with serving advantage. The reason why it is chosen is that when a server makes full use of serving hit and win the game, serve grows more confidence and mental strength also increases. If at the moment the sever is at a disadvantage on scores, the server can possibly perform better and reverse the game.

- Breaking serve: This is the opposite of holding serve, meaning that the returner wins the game. The achievement greatly encourages the returners as they are at a disadvantageous position in face of servers. On the other hand, servers start to feel upset for losing the advantage, negatively impacting later performance. In this case, the game swing might happen.

- Hitting net: This is the case when players make mistakes, failing to pass the ball and losing point. It rarely happens but has generally large impact on mental strength and decrease the performance potentially.

- Mental strength: Players can gain or loss mental strength during a match due to motions or events. For example, if a server serves successfully and maintains the advantage, he or she might gain mental strength gradually.

- The proportion of when player is one point to win the game with the opponent being the server: When a player is one point to win the game, it is natural to grow more desire and confidence to win the game. Especially the server is the opponent, the winning can bring even more benefit to mental strength growth. This impact has the potential to swing the match and reverse the result [4].

In terms of the indicator selection, the characteristic that logistic regression allows different types of independent variables has been considered, as some indicators have effect when in bigger values and some have effects when in smaller values [5]. It is noteworthy that the logistic regression model is tightly corresponded to the insights gained from the decision tree model that provides essential indicator selection advice. Similarly, we conducted analysis for Wimbledon matches data sets to acquired indicator values, which would be used to train and adjust the regression model. Afterward, we deployed the Omnibus tests and Hosmer-Lemeshow tests, which are basic tools for evaluating the performance and validity of logistic regression models in statistical analysis.

Omnibus test based on Chi-square distribution is a comprehensive assessment of the overall significance of independent variables in predicting dependent variables. Specifically, the significance level is 0.03 which is below the conventional threshold of 0.05, indicating that there is strong evidence

against the null hypothesis. This statistical criterion is indispensable to measure the reliability and robustness of logistic regression models.

Meanwhile, the Hosmer-Lemeshow test plays a key role in evaluating the fitness of logistic regression models. By comparing actual results and expected results at different prediction probability levels, the test assesses the model's ability to accurately predict results. Significance values is 0.82 which is above 0.05, indicating that the actual and expected results are very close, thus affirming the adequacy of the model in capturing the underlying relationships within the data. Thus, the significance value indicates that the prediction accuracy of the model is ideal.

In order to find the quantities that are most correlated with match fluctuations, a significance analysis was done for the model. Significance analysis was used to determine whether each variable was significantly different in both the input and no input scenarios. In the significance analysis, by comparing the significance values more significant indicators would be focused to analyze the impact on the matches for players. These indicators could be evident advice for players to better handle the match swings, which also could be the reference for arranging the concentration of players' training plans. Detailed results for significance analysis will be included in sec. 3.2. To test our model's ability to predict other tournaments, we found data from 3 matches of the French Open, divided the time of each match into 100 equal parts and brought them into our model for prediction. Based on analysis to objective factors that may lead to the differences in prediction accuracy, it is found that environmental factors, including weather and court surface materials, and physical ability differences can contribute to prediction accuracy discrepancy [6].

With the methodology of logistic regression model, our study is dedicated to discover and clarify the complex dynamics in Wimbledon single matches by analyzing the relationship between in-game behaviors of two players and game swings. By incorporating the outcomes from decision tree model and logistic regression algorithm, factors that are most related to the game swings can be found and corresponding game swings based on player moves can be predicted, which can improve public and athletes' understanding to the game process.

3. Results

3.1. Data Processing in Decision Tree Molde and ID3 Algorithm

The server typically has an advantage over the returner. To calculate the benefit from the serving hit, we arrange every 10 time intervals as a set and analyze the serving times of one player respectively. The statistics of events in respect of time were obtained from 2023 Wimbledon Gentlemen's singles final [7]. By adding up the number of serving times of player 1 (t_{p1}), which shows how many times player 1 plays as a server during a set, with total number of serves (t_{total}), the serving percentage (t_i) can be calculated as follows.

$$t_i = \frac{\sum_{i=1}^{n} t_{p1}}{t_{all}} \tag{1}$$

The serving percentage calculated above represents the proportion of the number of player 1 servings to the total number, illustrating the amount of advantage or disadvantage that a player can gain from serving options. The swing of a match can be more apparent if we look into the score comparison of two opponent players. The real-time score comparison can significantly impact the psychological status and further alter the physical performance. To quantify the impact of scores on players, we

calculate to compare one player's score with the sum of player 1 scores (s_{p1}) and player 2 scores (s_{p2}). The equation for calculating the proportion is shown below.

$$s_i = \frac{\sum_{i=1}^{n} s_{p1}}{\sum_{i=1}^{n} s_{p1} + \sum_{i=1}^{n} s_{p2}} \tag{2}$$

This equation directly reflects the score situation with player's performance and also provides theoretical conditions to the detection for flows of a match. Although double fault rarely happens in professional game, it is included because of the noteworthy impact on mental status fluctuation. Just imagining you as a pro tennis player on the court, failing to serve twice, thinking of the free point awarded to your opponent, your focus can be distracted more or less and performance is very likely to drop during next game. This refers to mental recovery capability for players to handle everything might happen during the game with the least loss of performance. Based on the given data, the number of double faults can be acquired as Df_i. We take the reciprocal because of the model feature that the higher the value is, the better. For double fault, the less, the better. In case of Df_i being zero, we choose to add 1 to the denominator.

$$Df_i = \frac{1}{\sum_{i=1}^{n} Df_i + 1} \tag{3}$$

Moving distance also plays a role in performance evaluation because it consumes more energy as the player moves further. To obtain the impact on the performance, we accumulated the moving distance from the beginning of a game to each time interval. Then we gained the total moving distance (L_{p1i}) at any given time. Again, we took the reciprocal of the moving distance to evaluate the performance impact. It should be noted that once the game starts, the moving distance increases from zero. Hence, the denominator does not need the mathematical process that we did to double fault evaluation. Moving distance impact can be expressed as

$$L_i = \frac{1}{L_{p1i}} \tag{4}$$

The performance indicators for different aspects are quantified and calculated in above methods, and by extracting data from match data sets, a standardized evaluation for desired aspects of a match can be completed. All evaluation values will be analyzed by ID3 algorithm to establish the decision tree for mental strength estimation ultimately. ID3 algorithm can analyze the relationship between the amount of occurrence of indicators and evaluation values, accurately determining the impact of each on the mental strength [8]. ID3 algorithm can mathematically determine the order of indicators and the classification standard values in the decision tree [9]. After all these classifications, the final value of mental strength can be reached ranging from -1 to 1. The result of the decision tree is in the fig.1 below.

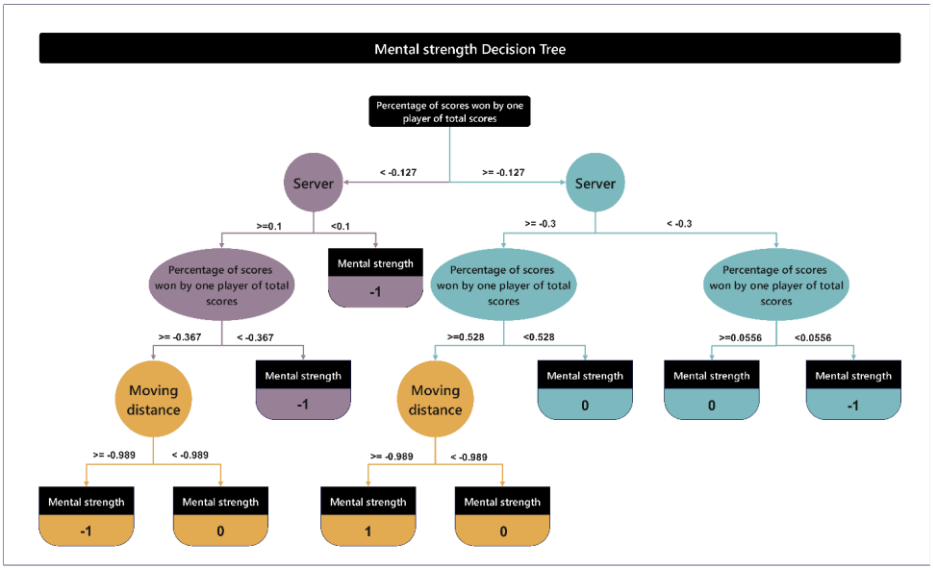

Figure 1: The significance values and parameter values for logistic regression model

By accumulating mental strength values for previous time intervals and repeating for the whole match time, a mental strength-time curve can be obtained, which describes the players' mental strength conditions at different time intervals. On the other hand, the curves can be also perceived as the game swings trend through the match. When the mental strength of a player is lower than the competitor, he/she is probably in a disadvantageous position with less scores. When the mental strength outraces the competitor, he/she could make a score reversal. Compared to the actual score swings, the mental strength evaluation curves can reach over 97% accuracy.

3.2. Data Processing in Logistic Regression Model

In the Omnibus test, the main outcomes are the Chi-square test and significance values, which indicate independent variables can efficiently predict the results of dependent variables if the significance values are less than 0.05. In the Hosmer and Lemeshow test, the fitness of the model is measured, which is basically the consistency between expected values and actual values [10]. If the significance is more than 0.05, the predicting model has a relative low accuracy. The regression model is established based on the following equation (5).

$$ln\left(\frac{p}{1-p}\right) = f(x) = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + \cdots + B_n X_n \tag{5}$$

The parameters (B) in the model equation (1) can be obtained for different indicators by logistic regression. Significance values for each indicator are shown in the fig.2. Significance values can build the relationship between indicators and game swing occurrences by comparing the scale of significance values. Significance analysis was used to determine whether each variable was significantly different in both the input and no input scenarios, and thus the quantity that was most relevant to match fluctuations.

	B	Significance
Holding serve	-2.636	0.043
Hitting net	1.035	0.077
Time when a player is one point to win the game with the opponent being the server	-1.198	0.002
Breaking serve	3.825	0.013
Mental strength	/	0.242

Figure 2: The significance values and parameter values for logistic regression model

Based on the results of our analysis, we find that the indicator of the player being one point to win is the least significant at 0.02. This is followed by the percentage of breaking serve to total pairs and the percentage of holding serves to total serves, which are significant at 0.013 and 0.043. Based on our analysis, we believe that players who are good at breaking serve and holding serves have a stronger ability to reverse the match advantage, while we note that the mental strength indicator has the greatest significance, which means that there is no strong correlation between whether the match is reversed or not and who is currently dominant. These findings based on data analysis may provide training guidance for players to enhance mental strength and ability to reverse the match [11].

In order to assess the reliability and predictive capacity of our model in terms of tennis match outcomes, we conducted an empirical investigation utilizing data from Wimbledon men's and women's matches. Our approach applied in the evaluation, which is similar to model training methodology, includes the data input process reflecting match dynamics, data pre-processing procedures, and the integration with the model structure. Results revealed the notable difference of prediction accuracy between men's and women's matches, with 94% prediction accuracy for Wimbledon men's matches and 82% for women's matches.

The observed discrepancy in prediction performance suggests distinctions in the factors influencing match outcomes across gender divisions within the Wimbledon tournament context. We have assumed that the higher prediction accuracy for men's matches might be attributed to the distinctive nature of the best-of-five games, in contrast to the best-of-three games of women's matches. The difference in rules implies divergent physiological demands and strategic considerations between male and female players, which also suggests that specific modelling strategies are required to precisely predict matches. Compared to women's match in the best-of-three games, higher physical strength requirements for men's match to play best-of-five games significantly contribute to the difference in prediction accuracy. Because the increased game duration and physical demands for male players can generate greater dependency on performance index, including the indicators that have been collected to build the model. Therefore, the greater dependency on such aspects can be more easily captured and predicted by the model. Conversely, women's matches, characterized by a shorter duration and greater emphasis on explosive power, may introduce additional complexities into the indicator selection and prediction model, thereby potentially decreasing prediction accuracy.

Hence, the distinct variability in match dynamics and player strategies across gender divisions further emphasizes the necessity for specifying modeling approaches. For example, factors such as playing style preferences, court surface quality, and in-game strategies may have different influence on match outcomes across men's and women's tournaments. Further research is required to completely locate variabilities to build accurate prediction models for two genders.

On the other hand, to test our model's ability to predict other tournaments, we implemented data from 3 matches of the French Open and divided the time of each match into 100 time intervals. After deploying new data sets in our model, it is obtained that the prediction of French Open reaches 87% in accuracy overall [12]. Compared to the prediction of Wimbledon Championships, the accuracy of the model for the French Open decreased, which is due to the differences between the French Open and Wimbledon Championships. From the perspective of environment factor, French Open is played on red clay courts, while Wimbledon Championships is played on grass courts. Red clay courts provide more friction and balls bounce lower on the courts, meaning that more physical strength and endurance are required for players. Nonetheless, the grass courts provide smoother surface and balls can bounce higher, which means that Wimbledon's courts support faster serves, demanding quicker reaction ability from players.

4. Conclusions

On the tennis courts, game swings between two players occur commonly. We implemented ID3 algorithm to train the player performance metric system in the form of decision tree, by which the score swings through a match can be obtained. The performance metric system especially focused on the impact of diverse players' motions on their mental strength, which can effectively reflect the game swings after verification. Based on the results from decision tree model, we built the logistic regression model to predict potential game swings. By conducting the Omnibus test and significance analysis, the importance of each indicator from the logistic regression model can be quantified. This means that training breaking serves and holding serves might enhance tennis players' abilities to handle game swings.

The evaluation also has been made for two models. The decision tree model based on ID3 algorithm has clear mechanism and can process raw data without complex calculation. However, it is sensitive to data modification, due to overfitting characteristic, leading to the instability of the model. On the other hand, the logistic regression model computes large-scale data with higher stability. It can interpret the parameters as the degree of impact of indicators on classification, illustrating the result with simplicity. Nonetheless, the logistic regression model does not consider the individual physical differences for players, decreasing the prediction accuracy. Therefore, further researches are required to explore methodologies of predicting matches and provide insights into tennis gameplays.

References

[1] Anna Fitzpatrick, Joseph A. Stone, Simon Choppin, John Kelley. How are elite tennis matches won at Wimbledon? [J]. A comparison of close and one-sided contests, 2024, 24(2): 190-199.

[2] Ząbkowski, T. Interactive Decision Tree Learning and Decision Rule Extraction Based on the ImbTreeEntropy and ImbTreeAUC Packages [J]. Processes, 2021, 9(7):1107.

[3] King, E. N., Ryan, T. P. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression [J]. The American Statistician, 2022, 56(3): 163-170.

[4] Reid, Machar, Rob Duffield. The development of fatigue during match-play tennis [J]. British journal of sports medicine, 2014, 48(1): 7-11.

[5] Corey, Cooper P ; Senkbeil, Jason C. Regional to Mesoscale Influences of Climate Indices on Tornado Variability[J]. MDPI, 2023(11):11

[6] Levy, J.J., O'Malley, A.J. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning[J]. BMC Medical Research Methodology, 2020, 20: 171.

[7] ALC 3-2 DJO | Carlos Alcaraz – Novak Djokovic | Summary[EB/OL].(2023-07-16)[2024-04-19]. https://www.flashscore.com/match/dYpF0bum/#/match-summary/match-statistics/0

[8] Pathan, Shabana S. An Approach to Decision Tree Induction for Classification[J]. Karadeniz Technical University Distance Education Research and Application Center, 2021(12):12

[9] Li Y, Jiang Z. L, Wang X, Fang J, Zhang E, Wang X. Securely outsourcing ID3 decision tree in cloud computing[J/OL]. Wireless Communications and Mobile Computing, 2018, 2018(10):1-10.

[10] Lin J, Ruan S, Sun W. A novel score to predict progression in anterior circulation single subcortical infarction patients[J]. Annals of Clinical and Translational Neurology, 2024, 11(3): 791-799.

[11] Selmi W, Hammami A, Hammami R, Ceylan Hİ, Morgans R, Simenko J. Effects of a 6-Week Agility Training Program on Emotional Intelligence and Attention Levels in Adolescent Tennis Players [J]. Applied Sciences, 2024, 14(3): 1070.

[12] ZVE 3=1 ETC | Alexander Zverev – Tomas Martin Etcheverry | Stats[EB/OL].(2023-06-07)[2024-04-20].https://www.flashscore.com/match/MmYG8t6c/#/match-summary/match-statistics/0