ISSN 2616-5775 Vol. 7, Issue 6: 58-64, DOI: 10.25236/AJCIS.2024.070609

Research on the Role of Predicting Momentum in Competition Based on Random Forest Algorithm

Xiaoqi Yu[#], Min Yu[#], Xinping Liu, Rong Guo[*]

School of Mechanical Engineering, Taiyuan University of Science and Technology, Taiyuan, 030024, China
[*]Corresponding author: yuxiaoqi0112@icloud.com
[#]These authors contributed equally.

Abstract: *This paper focuses on a model that can capture match points in real time, record the scoring flow during the match, and determine the highlight moment of each player in real time. We apply this model to one or more competitions to better optimize the model by varying the proportion of weight factors and introducing more complex factors to adjust the value of momentum. Statistical distribution fitting as well as the Pearson correlation coefficient were used to assess the correlation between player momentum and match outcome, thus assessing the stochasticity of the transition between momentum and match situations. Using the random forest distribution algorithm to predict the turning point and evaluate the generalizability of the model gives a reasonable scheme. This article constructs two mathematical models for calculating momentum: a basic model and an optimized and improved model, to obtain the relationship between score and momentum in the game. In the basic model, this article sets up a function model to solve the problem under hypothetical conditions. Introduce weight parameters to adjust the degree of impact of winning or losing on momentum. But there are more complex factors, so we changed the function model to a state space model.*

Keywords: *Random Forest, Weight Factor, PCA, Pearson Correlation Coefficient*

1. Introduction

With the continuous development of technology and social economy, big data technology is becoming more and more important, it can not only to the past and present induction and summary, and can also understand the objective law of the development of things, understand human behavior and can help people change the past way of thinking, establish new data thinking model, to predict and deduction for the future[1]. During the 2014 World Cup, Google, Baidu, Microsoft and Goldman Sachs all launched outcome prediction platforms[2]. Baidu's forecast was 67 percent accurate in 64 matches and 94 percent accurate after entering the knockout rounds. It also means that future sporting events will be dominated by big data predictions[3].

At the same time, we cooperated with China Lottery website Lecai Network and SPdex, the European Bifa index data supplier, introduced the forecast data of the gambling market, and established a prediction model including 199,972 players and 112 million pieces of data, and predicted the results on this basis[4].

However, there are still problems such as unable to monitor the fluctuation of the game in real time, reflecting the game situation and accurate analysis of the players' highlight moment. Therefore, this paper tries to establish a universal visual model and can analyze and predict the real-time competition situation.

2. Research on Momentum Quantification in Competitions

To quantify the momentum in the competition, we can construct a model to assess the contribution of each scoring point to the player's momentum. The quantified value of the defined momentum at the time point t is:M_t

$$M_t = \alpha\Sigma_{i=1}^{t} P_i - \beta\Sigma_{i=1}^{t} L_i \quad (1)$$

2.1 The construction of the optimization model formula

Further, it can consider the introduction of more complex factors to adjust the calculation of

ISSN 2616-5775 Vol. 7, Issue 6: 58-64, DOI: 10.25236/AJCIS.2024.070609

momentum, such as the impact of winning points, important points (such as broken service points), as well as the physical condition and psychological conditions of the players, so as to realize the optimization model:

$$M_t = \alpha\Sigma_{i=1}^{t}(P_i \times W_{pi}) - \beta\Sigma_{i=1}^{t}(L_i \times W_{li}) \quad (2)$$

$$W_{pi}S_i\boldsymbol{C_i}\boldsymbol{B_i}=\times\times(1+\times\text{BallBreakPoint\_i})\times(1+\times\text{UnforcedErrors\_i})R_i \quad (3)$$

$$W_{li}S_i\boldsymbol{C_i}\boldsymbol{B_i}=\times\times(1-\times\text{BallBreakPoint\_i})\times(1-\times\text{UnforcedErrors\_i})R_i \quad (4)$$

Among:

BallBreakPoint_i and UnforcedErrors_i are binary variables based on specific event recordings in the actual competition,BallBreakPoint_i 1 means missed break point, 0 means no missed break point,If UnforcedErrors_i is 1, it means an unforced error, otherwise 0[5].

The momentum change analysis can be performed by comparing the changes in the momentum quantification values at consecutive time points, such as:

$$\Delta M_t = M_t - M_{t-1} \quad (5)$$

ΔM_t Representing the amount of change in momentum between successive time points t and t-1.

By analyzing the changes in momentum throughout the game, you can explore the changing trends in the game and the key events that lead to the momentum change[6].

2.2 State-space model

2.2.1 Foundation of the state-space model construction

Basic idea: Use time series analysis and state space model to capture the process of the game, and consider the influence of players' serve, winning streak, break point and physical decline,Y_tX_tFormula: defined as the game state of the time point t, it is the set of factors affecting the game state, including the current game points, service right, etc. The state-space model can be expressed as follows:

$$Y_t = \text{C} * + \text{D} + X_t\varepsilon_t \quad (6)$$

Y_t: Here can refer to the assessment of the state of the whole game,C: Observation matrix,D: free vector (constant term or control input),ε_t: error in observation,X_t: The state variable vector, which can include the match state of the previous point and any other factors that may affect the current state, the serve power factor is added by adding multiple binary variables: serve weight,: winning streak, getting the break point,unforced error $S_i\boldsymbol{C_i}\boldsymbol{B_i}R_iX_t$,Current and previous scores (p1_score, p2_score),Momentum[7], possibly with the previously calculated cumulative value (M_t_p1, M_t_p2),Service right (via df.server Judgment),Break point situation (possibly p1_break_pt, p2_break_pt),Winning streak status (C_i).

2.2.2 Visualize the momentum calculation

First, define the momentum calculation method, Momentum is based on the concept that a player wins consecutive points in a game. We can set up a momentum score for each player that increases or decreases depending on the points they win or lose. For example, for every one win, momentum is reduced or reset.Second, consider the serve advantage: in tennis, the server usually has a higher probability of scoring[8]. Therefore, in calculating the momentum, we can assign less weight to the points won by the server to reflect this natural advantage. Finally, load the data and calculate the momentum. If the player wins the score on their serve, we can reduce the momentum increase slightly to reflect the serve advantage.Use PCA data dimension reduction techniques while retaining as much information as possible. Principal components are linear combinations of the original features, whose direction is the direction with the largest variance in the data. Dimensionality reduction is achieved by projecting the data onto these principal components.

2.2.3 MATLAB Implementation process

Populate the missing values in the selected value column with a median policy. Specific implementation is

$$\begin{cases} X_{i,j} & if\ X_{i,j}\ is\ not\ missing \\ median X_j & if\ X_{i,j}\ is\ missing \end{cases} \quad (7)$$

Where Z is the populated data matrix, X is the raw data matrix, i is the sample index, and j is the

ISSN 2616-5775 Vol. 7, Issue 6: 58-64, DOI: 10.25236/AJCIS.2024.070609

feature index.

2.2.4 Data standardization

StandardScaler Feature scaling was performed to transform each feature into a standard normal distribution form with mean 0 and standard deviation 1. The normalization formula is as follows:

$$Z_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j} \tag{8}$$

$Z_{i,j} x_{i,j} \mu_j$Here is the normalized value, the original value, the mean of the j th feature, but its standard deviation.

2.3 PCA model establishment and application

The PCA model is initialized without limiting the number of principal components, that is, computing all possible principal components[9]. The core mathematical principle of PCA is based on the eigendecomposition of the covariance matrix or the correlation matrix. For a data matrix X, the goal of PCA is to find a set of orthogonal basis vectors (i. e., principal components) that maximize the corresponding eigenvalues of these vectors, i. e., to solve the following optimization problem as follows:

$$max_{w_1, w_2, \ldots, W_m} \sum_{k=1}^{m} \frac{w_k^T S w_k}{w_k^T w_k} \tag{9}$$

Where S is the covariance matrix of the sample data or the dot product matrix of the normalized data matrix, and m is the number of features and is the direction vector of the k th principal component.w_k

The proportion of the total variance explained by each principal component, i. e., the cumulative explained variance ratio, is expressed as:

$$Explained\ Variance\ \text{Ratio}_k = \frac{\lambda_k}{\sum_{j-1}^{m} \lambda_j} \tag{10}$$

λ_kWhere are the k th eigenvalue, reflecting the importance of the corresponding principal components, determine the number of principal components retained: find the minimum number of principal components required to reach this threshold based on the set cumulative contribution value threshold (95%),actual dimensionality reduction: recreate the PCA model using the determined number of principal components and apply it to the standardized data for data dimension reduction:

$$T = W^T X \tag{11}$$

Results visualization: The following shows the brief results of PCA dimension reduction, as shown in Fig 1 and 2.

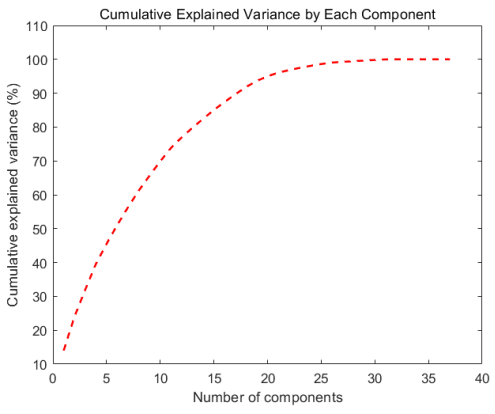

Figure 1: Visualize the variance changes

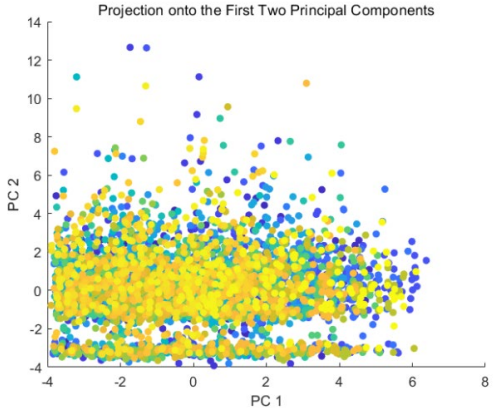

Figure 2: PCA fitted model

2.4 Role analysis of momentum

2.4.1 Momentum changes with the situation

$P_t P_t'$Basic idea: Use Monte Carlo simulations to test the presence and impact of momentum. Using historical data to estimate the odds requires applying a probability model of logistic regression and

ISSN 2616-5775 Vol. 7, Issue 6: 58-64, DOI: 10.25236/AJCIS.2024.070609

calculating the win rate of the player at each time point.P_t, Let represent the probability of a player winning at the time point t, estimated based on historical data[10]. Considering the momentum, the updated win rate is:M_t

$$P_t' = P_t + m(M_{t-1}) \tag{12}$$

M is a function that describes the impact of the previous moment of momentum on the current win rate.M_{t-1}, under different momentum setting, and judge the actual influence of the momentum. A dynamic map of a player's momentum accumulating with the situation is mapped using METLAB as follows Fig 3 and Fig 4:

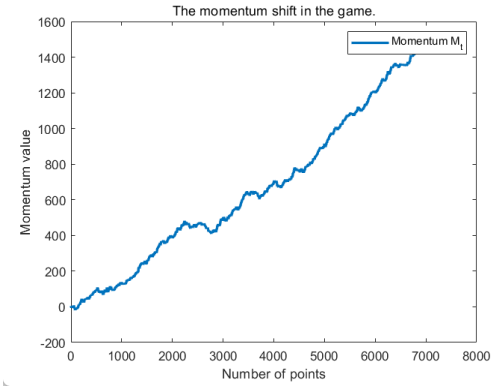

Figure 3: Momentum changes with the situation

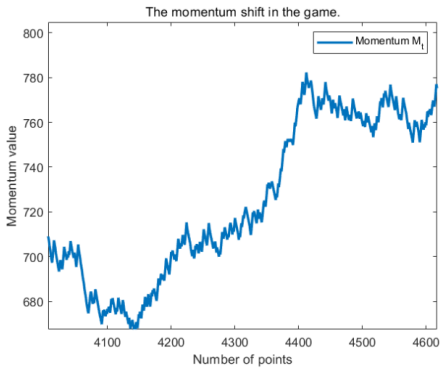

Figure 4: Momentum changes with the situation

The following curve shows the momentum changes of the two players over time, as follows Fig 5:

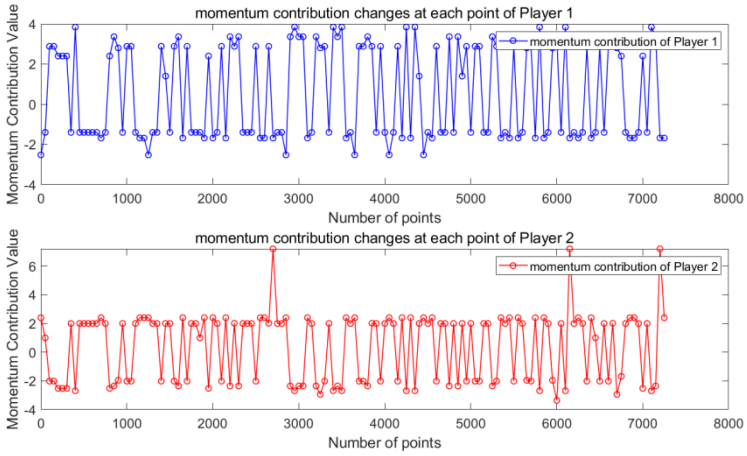

Figure 5: Momentum changes over time

2.4.2 Randomization of momentum conversion

Null hypothesis (H0): Momentum in a game is a random phenomenon, meaning that continuous scoring does not have a substantial impact on the outcome of the game,optional hypothesis (H1): Momentum is not random and has an impact on the outcome.

The Pearson correlation coefficient was used to assess the correlation between player momentum score and game outcome, analyze the running length of a winning streak (or losing streak). If success succession is random, we expect the length of a streak (or losing streak) to follow a specific statistical distribution.

Finally, calculate the running length. For each game calculate the sequence length of the winning streak and losing streak, try to fit the statistical distribution to the run-length data. If the actual data fit well with the theoretical distribution, the coach view is supported. A chi-square test is performed to determine whether the distribution of run lengths is consistent with the random process. The standardization process adopts the StandardScaler method in the sklearn library. The core principle is to convert the original data to a standard normal distribution with mean 0 and standard deviation 1. The specific mathematical expressions are as follows:

For each-column feature X:

ISSN 2616-5775 Vol. 7, Issue 6: 58-64, DOI: 10.25236/AJCIS.2024.070609

$$Z = \frac{x-\mu}{\sigma} \tag{13}$$

Among, μ Represents the mean value of the column features (mean centralization); σ Represents the standard deviation of the column feature (variance scaling);Z is the new value after normalization.

2.4.3 Derived new feature stage-p1_performance index construction

To comprehensively assess the overall on-field performance of player P1, we introduce a synthetic feature, 'p1_performance', which is the sum of all the normalized features described above: p1_performance= ZscoreAD + Zscore +Zserve + Zace + Zdouble_fault + Zdistance_run.

This composite feature forms a comprehensive indicator by calculating the sum of the integration of the player's performance at different technical levels.

2.5 Conclusion on the existence of momentum

Visualization diagram of momentum action, as follows Fig 6:

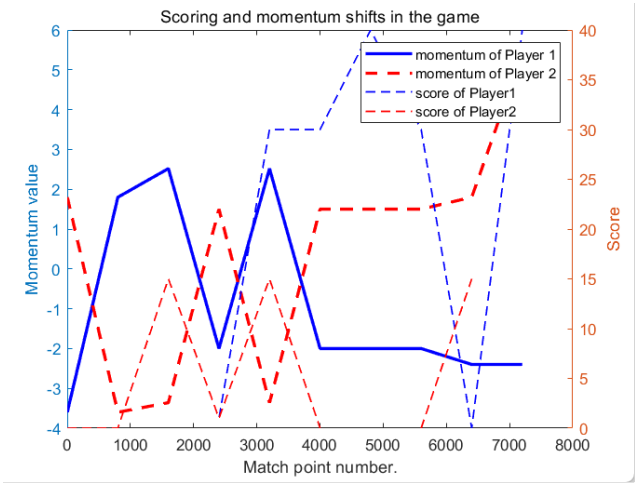

Figure 6: Performance of integration at different technical levels

The relationship reflected by the picture: Correlation coefficient of momentum and score of player 1: -0.198666, p-value: 0.000000. Correlation coefficient of momentum and score of player 2:0.279470, p-value: 0.000000, there was a significant correlation between momentum and score for player 1 (p <0.050000), there was a significant correlation between the momentum and the score of player 2 (p <0.050000). Correlation coefficient of lag momentum and score of player 1:0.461850, p-value: 0.000000. Correlation coefficient of lag momentum and score of player 2:0.290363, p-value: 0.000000. There was a significant correlation between lag momentum and score for player 1 (p <0.050000). There was a significant correlation between lag momentum and score for player 2 (p <0.050000). It can be seen that the relationship between momentum and fraction is not an invariable function, and the changing trend is also different under different interference conditions.

2.5.1 Visual results represent

For player 1, there is a negative correlation between the momentum and the score, that is, the score may be reduced when the momentum value increases. For player 2, there is a positive correlation between the momentum and the score, that is, the score may also improve when the momentum value increases. The results of the correlation test indicated that for our dataset, there was indeed a statistically significant linear correlation between player momentum and their scores. The output showed a significant correlation between the lagged momentum and the score of players 1 and 2, as the p-values for both players were far less than the significance threshold of 0.05.

2.5.2 Analysis conclusion

For both players, the increase in momentum did indeed show an increase in the score later, and this effect may be more pronounced for player 1. That is, this association may be affected by other hidden variables, such as competition stage, competition pressure and other factors, and player 1 takes the momentum longer than player 2.

ISSN 2616-5775 Vol. 7, Issue 6: 58-64, DOI: 10.25236/AJCIS.2024.070609

2.6 Predict game dynamics

2.6.1 Construction of the mathematical model

Use a machine learning model, random forest, to predict turning points in a game.Z_t Formula, set as binary variable, 1 if the momentum is about to change at time t; otherwise 0. The prediction model is:

$$Z_t = h(X_{t-1}, X_{t-2}, \cdots, X_{t-n}; \Phi) \tag{14}$$

And h is the prediction function, and Φ is the model parameter.$X_{t-1}, X_{t-2}, \ldots, X_{t-n}$Is a set of states of the first n time points, used to capture pre-turning patterns.

2.6.2 Mathematical model construction

Characteristic engineering: Extract relevant features from the competition data, which may have an impact on the momentum conversion of the competition.

2.6.3 Model prediction

The predicted target Y is a binary variable, and Y=1 if the game flow turns to another player within the next few points, otherwise Y=0.

2.7 Processing and calculation of data

Selected features are standardized to ensure that all features are on the same scale. The StandardScaler is used, and the formula involved is:

$$Z = \frac{x-\mu}{\sigma} \tag{15}$$

Where x is the original feature value, μ is the mean of that feature, and σ is the standard deviation.

2.7.1 Random forest model establishment

The random forest classifier was initialized and the set parameter n_estimators=100 represents the forest constructed consisting of 100 decision trees. For each decision tree, the optimal features and their thresholds were selected in a recursive fashion to segment the data until the stopping condition was reached,In random forests, the importance of features is often measured based on the ability to 'reduce impurity'. Here the feature_importances_ property is obtained by going through all trees and averaging the importance score for each feature, without an explicit mathematical formulation, but rather evaluated in a statistical sense.

In the prediction phase, the trained random forest model is used to vote or average prediction on the test set (for classification problems, the majority vote is usually used; for regression problems, the average prediction value is taken). For the performance evaluation, the accuracy index is used, which is defined as follows:

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

Among them, TP, TN, FP, and FN represent the number of true, true negative, false positive, and false negative cases, respectively.

This paper uses METLAB for the following mapping as follows Fig 7 and 8:

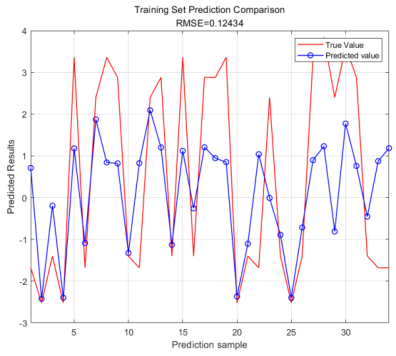

Figure 7: Distribution of predicted and actual results

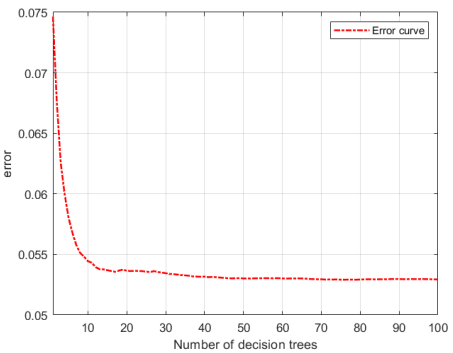

Figure 8: Error curve for the number of decision tree choices

ISSN 2616-5775 Vol. 7, Issue 6: 58-64, DOI: 10.25236/AJCIS.2024.070609

This paper uses bar charts to more intuitively represent the influence of state factors, as follows Fig 9:

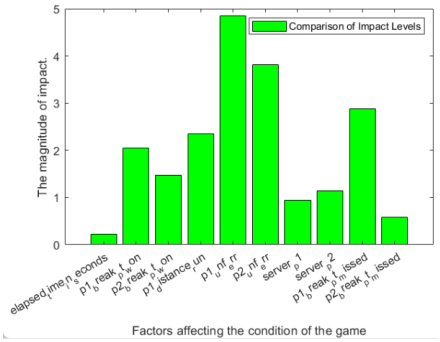

Figure 9: Bar graph of the magnitude of its influence by status factors

3. Conclusions

The model implemented in this paper can capture the game points in real time, record the scoring process during the game, and determine the highlight moment of each player in real time. And can be applied to one or more competitors to better optimize the model by varying the proportion of the weight factors and introducing more complex factors to adjust the momentum values. In general, in our basic model, through the standardization of the data, we can intuitively see the effect of the momentum on the game trends. Massive information can be calculated and analyzed according to big data. The more comprehensive the information, the more information dimensions, the more accurate the analysis results will be. Using the Pearson correlation coefficient to evaluate the correlation between the player's momentum score and the game result, you can avoid some unnecessary data interference. And a random forest algorithm is used to integrate multiple decision trees that recursively select the optimal features and their thresholds to segment the data and train on different randomly selected subsets of features. Model parameters can be adjusted according to different types of data sets, or different machine learning algorithms can be used to adjust the model, so as to better simulate the situation trend and predict the future model. And the performance of the model can be evaluated by cross-validation. Thus, the universality of models under different conditions and different factors is constantly enriched.

References

[1] Shou Kui, Sun Xijing, mathematical modeling algorithm and application [M]. Beijing: National Defense Industry Press, 2021.

[2] Hu Xiaodong, Dong Chenhui, METLAB from entry to mastery [M]. Beijing: People's Posts and Telecommunications Press, 2018.

[3] Wu Xingyong. The Application of Artificial Intelligence in Computer Network Technology in the Era of Big Data [J]. Information and Computer (Theoretical Edition), 2023,35 (22): 167-169

[4] Shuai Anqi. Social Media Data Analysis: Integration of Internet and Big Data [J]. Internet Weekly, 2024 (05): 37-39

[5] Shi Zhilong, Chen Gan, Xie Guoliang. Research on an edge cloud big data analysis algorithm based on Spark [J]. Changjiang Information and Communication, 2024, 37 (02): 183-185.

[6] Du Jiang, Dai Jun, Cao Ruiyuan. Optimization and Design of Practical Teaching System for Statistics in the Context of Artificial Intelligence and Digital Technology [J]. Journal of Higher Education, 2024, 10 (09): 115-118.

[7] Chen Mengna, strengthen the application of artificial intelligence and big data technology in the "champion model" scenario [J]. Shanghai Securities News, 2024,08 (3):78-79

[8] Morgulev E , Azar O H , Galily Y ,et al.The role of initial success in competition: An analysis of early lead effects in NBA overtimes[J].Journal of Behavioral and Experimental Economics (formerly The Journal of Socio-Economics), 2020, 89.DOI:10.1016/j.socec.2020.101547.

[9] Abdoh H , Varela O .Competition and exposure of returns to the C-CAPM[J].Studies in Economics and Finance, 2018, 35(4):525-541.DOI:10.1108/SEF-02-2018-0038.

[10] Yue Feng, the role of psychological quality in tennis competition [J]. Tennis world. 2024(01):58-60