

A Study on Visualization of Tennis Players' Game Performance Based on Wavelet Analysis and ARIMA-LightGBM Model

Teng Wang*

Business school, Beijing Technology and Business University, Beijing, 102401

*Corresponding author: 13086856901@163.com

Abstract. With the continuous progress of science and technology, scientific prediction methods gradually penetrate into the field of competitive sports, and the performance of athletes on the field can be visualized by a series of indicators to further predict the outcome of the match. In order to visualize player performance, this paper first establishes a performance scoring model and quantifies the fluctuation of the matching situation and the player's performance, and the quantitative results are consistent with the trend of the actual matching. Considering the data characteristics of its time series, we carried out multi-scale decomposition on the basis of discrete wavelet transform, and finally reconstructed and synthesized the original sequence using multi-scale wavelet to obtain the corresponding feature matching trend graph. In order to verify whether there is a relationship between consecutive scores and players winning matches, this paper uses Spearman correlation analysis and Kruskal-Wallis test to verify that there is a significant correlation between the two. At the same time, by synthesizing the relevant data indicators of each tennis match, this paper combines the advantages of ARIMA and LightGBM to construct a more accurate prediction model.

Keywords: Wavelet Analysis, Performance Visualization, Correlation Analysis.

1. Introduction

Adjusting the performance of tennis players on the court to increase the likelihood of scoring more points in a match is one of the most talked about topics in the sports world. It is crucial to build a model to accurately predict the match situation and to visualize the players' performance on the court.

Many researchers and scholars at home^[1] and abroad have done a lot of research in this area. João Gustavo Claudino et al.^[2] at the University of São Paulo used artificial intelligence and artificial neural networks to analyze individual athletes' abilities and applied artificial neural networks, decision tree classifiers and support vector machines to predict the risk of sports injuries. These techniques were successfully applied to predict game performance in soccer and basketball, providing an intelligent upgrade to team sports. Adam Maszczyk and Artur Golas, on the other hand, used nonlinear regression models and neural network models in javelin throw, and the results showed that the neural network models had a higher prediction accuracy. The team led by Rocio de la Iglesia^[3] biologically perspective to analyze the genetics of athletes and to identify genetically gifted athletes through the Genetic Hardening Predictive Score (GES) to improve shortcomings and improve athlete performance.

Although data mining techniques and related algorithms are widely used in the analysis and prediction of sports event data, we found during our review that these literatures seldom consider the psychological factors of athletes when making predictions. However, we found through our research that the psychological factors of athletes are crucial to the outcome of the game. In this paper, we synthesize the previous research results and add a more comprehensive analysis for the indicator aspect of the prediction model. The psychological factors of the players on the court are quantitatively taken into account^[4], so that the performance of the players on the court is accurately visualized in order for the players to be able to adjust their state in time in order to win more scores. Therefore, we analyze the data based on the real data of many tennis matches, and at the same time, we use these real match data to repeatedly test and experiment the model, and we find that the images obtained from these experiments match the real situation of the matches to a high degree.

2. Performance Score Model Based on Wavelet Analysis

In this study, a score model is first constructed to present the matching streams, and then time series feature mining based on wavelet analysis is performed, and feature development based on time series analysis is performed.

2.1. Player Performance Scoring System

By observing the data of many tennis matches in the past to construct a model, the model mainly refers to three scoring indicators, including the number of sets, the number of games and the number of points. It also incorporates factors with high probability of scoring on the serve side factors with high probability of scoring on the serve side^{[5][6]}. This created a player performance scoring system. The scoring performance *PerS* can be expressed as

$$PerS = \theta_1(S_1 - S_2) + \theta_2(G_1\mu_1^{F_1} - G_2\mu_2^{F_2}) + \theta_3(P_1\mu_2^{F_1} - P_2\mu_2^{F_2}), \quad (1)$$

$$\sum_{i=1}^3 \theta_i = 1, \quad \sum_{i=1}^2 F_i = 1, \quad \prod_{i=1}^2 F_i = 0, \quad \mu_1 < \mu_2, \quad F_i = \begin{cases} 1, & \text{-serving} \\ 0, & \text{no serving} \end{cases}, \quad (i = 1, 2).$$

In the expression of *PerS*, the objective fact that exists is that the disparity in S_i , G_i , and P_i between players, affects the performance score to a lesser extent in that order. Therefore, we take θ_i into account as an impact multiplier in our modeling. At the same time, we believe that the serve has an effect on the match score. In order to build the model more accurately, we quantify it into the index μ . By observing the historical match data and finding related literature, we synthesize that the results of the model fit better with the actual match score when we take the value of μ_1 to 1.1, μ_2 to 1.3. In addition, when $F_i = 1$ it means that the i th player holds the serve at this moment, we establish a mathematical relationship between μ and the serving F_i , in time in order to better improve the model.

The data from all Wimbledon 2023 mens matches was substituted into the above model, which in turn yielded performance scores for the entire match. We drew a time series figure of the players' match scores with time as the independent variable in Figure1. The Figure1 shows how the situation changed during the match and the specific performance of the players. This figure corresponds to player performance in each set of the match, with positive and negative values each corresponding to the dominance held by two players. This approach quantifies the value of the advantage specifically as the distance of the curve from the x-axis at each moment. Then, we found the scoring swings calculated by the above model perfectly match the match flow described in the actual situation.

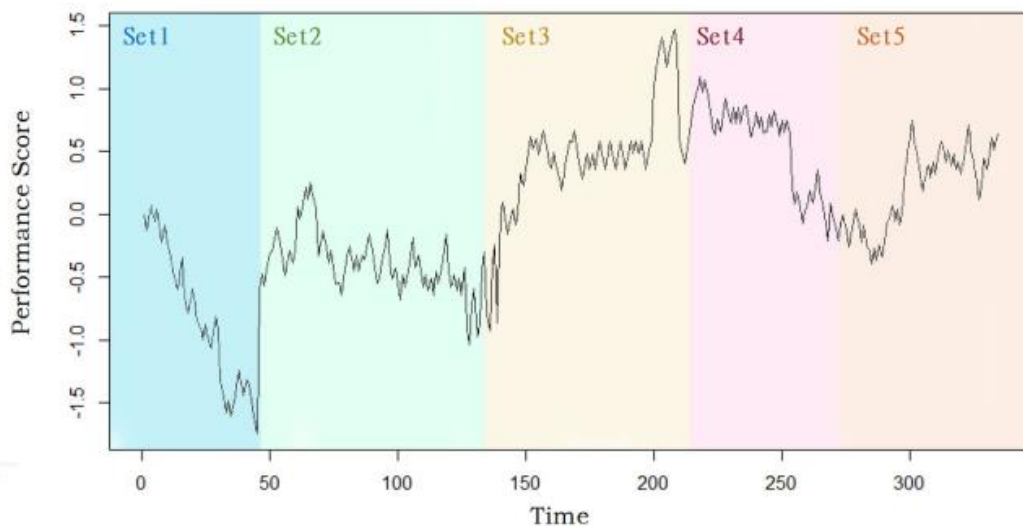


Figure 1. Performance score model based match flow

2.2. Wavelet Transform Based Timing Feature Mining

By looking at the autocorrelation and partial autocorrelation plots of the score time series, the autocorrelation plot shows a significant trailing autocorrelation coefficient and the partial autocorrelation coefficients in the partial autocorrelation plots, except for the delayed first, third and seventh order, fall within two times the standard deviation. We then performed a unit root test on the model and the p-value of each type of statistic was greater than the significance level of 0.05, so we concluded that the series was non-stationary. For the non-stationarity of this sequence, this paper is based on wavelet analysis for feature mining of time series. We use wavelet analysis to filter the original time series. The wavelet transform is utilized to fully extract and separate various hidden cycles and nonlinear relationships in the scoring time series. This approach enables the monitoring of situation transitions, the identification of periodic components, and the analysis of multiple time scales in the scored time series.

Then performing wavelet reconstruction, we just get the coefficient maps at multiple scales are synthesized into the original sequence, and perform the Inverse Discrete Wavelet Transform (IDWT), which aims at rounding off some detailed factors that have little influence. So we calculate the discrete wavelet variation coefficients, plot the discrete wavelet coefficients at multiple scales, and finally get the discrete wavelet transform coefficient map in Figure2. These coefficients are used to analyze the time-frequency variation characteristics of the time series.

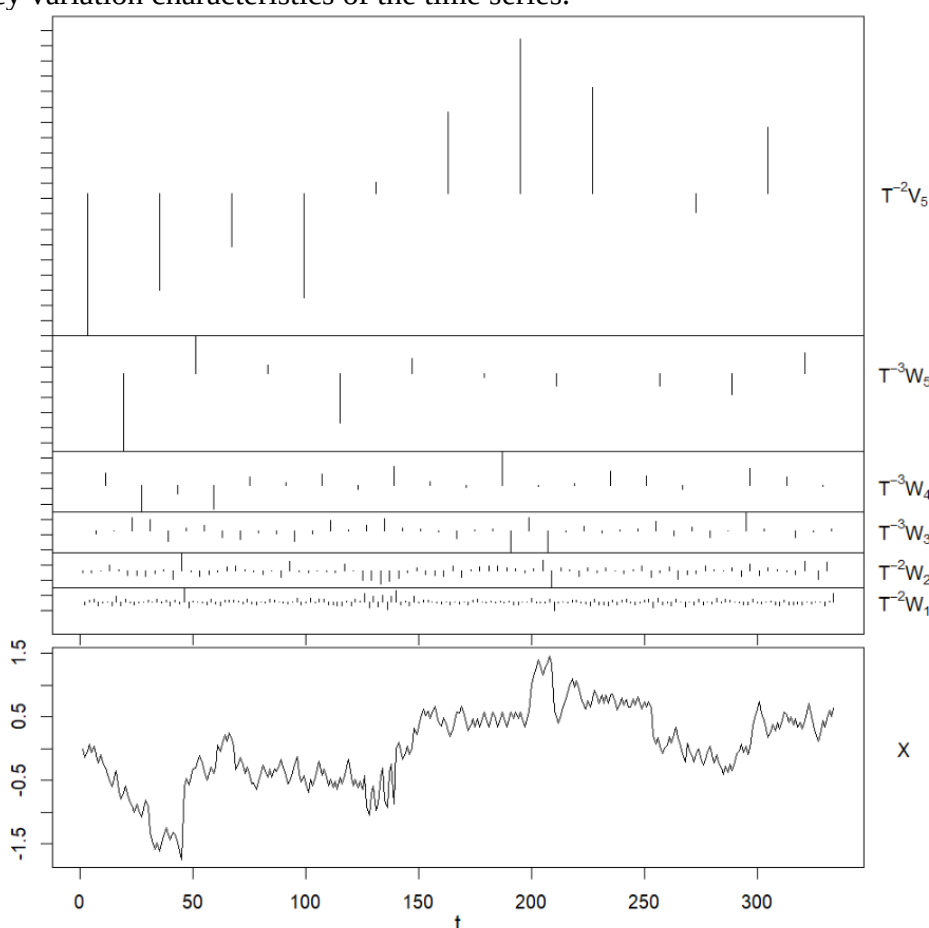


Figure 2: Discrete wavelet transform coefficient map

According to the trend of wavelet coefficients over time, we can then analyze the characteristics of the fluctuating cycle of player performance changes and mutation points under different time scales. In order to study the match flow in more depth, we analyzed the wavelet power and periodic component phases and plotted the wavelet power mean time distribution and wavelet power spectrum image in Figure 3, periodic component phase map and periodic component phase mirror image in Figure 4.

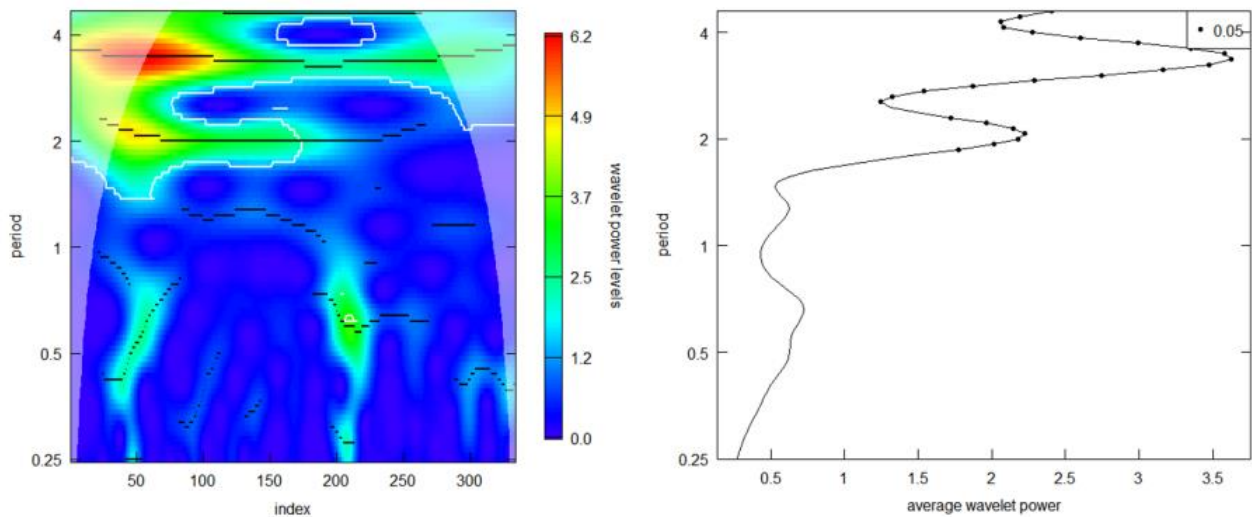


Figure 3: The wavelet power mean time distribution and wavelet power spectrum image

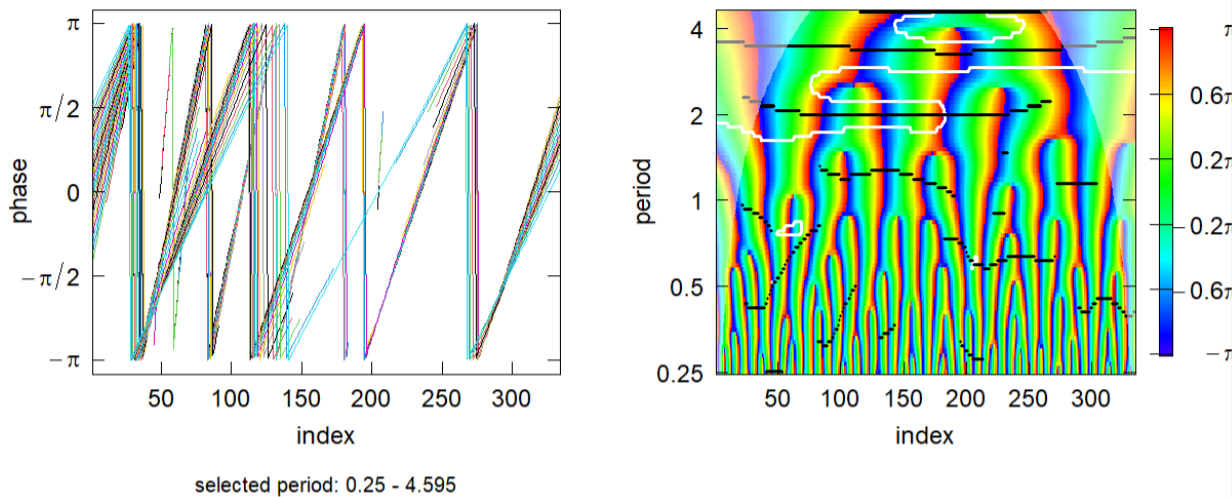


Figure 4: Periodic component phase map and periodic component phase mirror image

3. Correlating player swings in play with runs of success

3.1. Spearman correlation test

Because these data are not normally distributed, we use the non-parametric correlation spearman correlation test. We performed the spearman correlation test by taking the *PerS* with *Rp*, *Rg* and *Rs*, in groups of two. Finally, we reject H_0 and accept H_1 , and *PerS* are correlated with each of the indicators of runs of success, i.e., *Rp*, *Rg* and *Rs*, respectively. Their correlation coefficients are $r_s=0.23$, $r_g=0.41$ and $r_p=0.69$. Therefore, we conclude through spearman correlation test, that there is a correlation between a player's performance scores in the match and the runs of success.

3.2. Kruskal-Wallis testing

In a similar way as above, perform a Kruskal-Wallis test. In order to better compare the degree of influence of the three metrics^[6], we put together box plots of all three in Figure 5. Through these box plots we find that the runs of success are significantly correlated to a player's score, and that as the number of consecutive successes increases, the degree to which the players excel is higher. The runs of

Points *Rs* is the favorable factor that has the greatest impact on a player's performance level in these three metrics, followed by *Rg*, *Rp*, respectively.

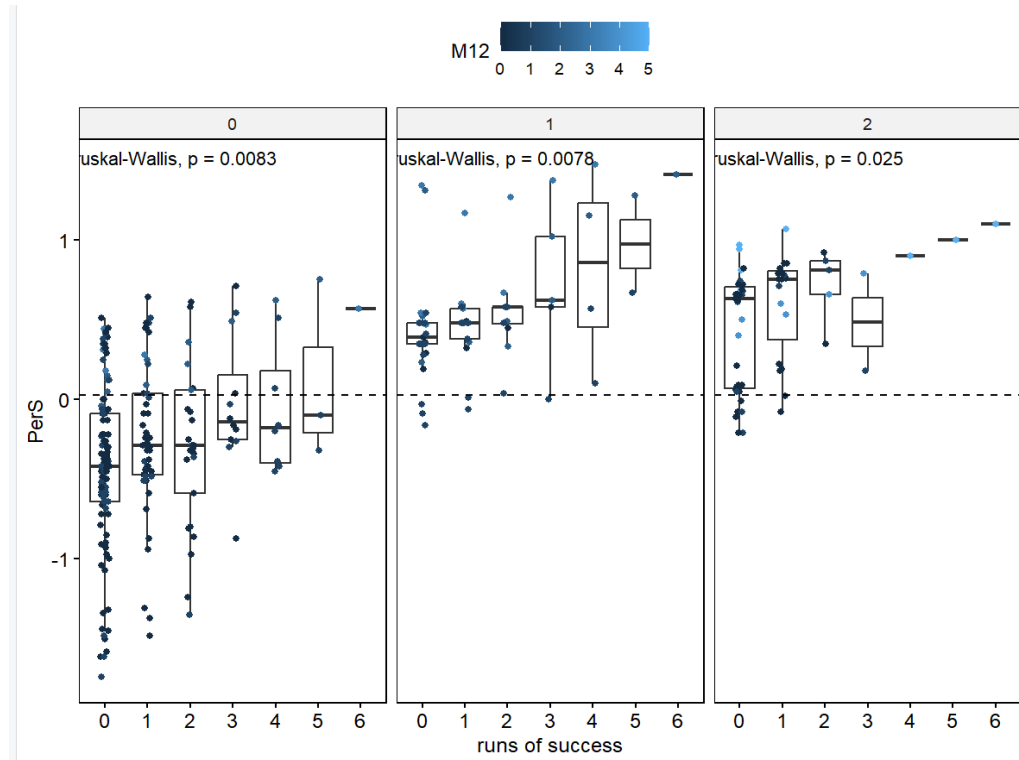


Figure 5: Comprehensive Box Line Diagram

4. Match Swing Recognition Prediction Based on ARIMA-LightGBM Model

Choose this model because it has advantages in the extraction and prediction of online nonlinear relationships, and it can recognize the mutation situation. Compared with the deep learning model, the model is shorter and more interpretable, which is more relevant to the actual situation of this problem.

4.1. Prediction of PerS with ARIMA modeling

The ARIMA model is able to capture the relevant trend factors in the time series^[8]. By inputting relevant training data and continuously tuning in the iterations, it is able to capture the nonlinear relationship between the match swing and the corresponding influencing factors, which can then be used for match swing prediction. The ARIMA model (p,d,q) ^{[8][9]} time series formula can be expressed as

$$I_t: y_t = \mu + \sum_{i=1}^p \gamma_i y_{t-i} + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (2)$$

In the expression, y_t is the current predicted match swing, μ is a constant term, ε_t is the white noise sequence, p is the autoregressive order, d is the number of differences, q is the moving average order, γ_i is the autoregressive coefficient, and θ_i is the weighting factor of the white noise sequence. We consider that $PerS$ is non-smooth, so we use Time series forecasting methods in R. The model is identified and fixed order. The final result is obtained. $p=2$, $d=1$, $q=2$. These data passed the significance test.

4.2. Prediction of $\Delta PerS$ with LightGBM Model

Considering that there are multiple factors affecting match swing, we refer to the related literature and utilize historical race data for machine learning. $\Delta PerS$ and multiple influencing factors are identified and extracted.

4.2.1 Momentum Scoring System

Quantify the momentum of a player on the field of play and construct a momentum scoring system. For a given player, the momentum MoS_t at a given moment can be expressed as

$$MoS_t = \sum_{i=1}^4 \omega_i \cdot R_{it} + P_t \cdot \Phi^{\sum_{i=1}^4 E_{it}} \cdot \Psi^{K_t} \quad (i = 1,2,3,4) \quad (3)$$

In this expression, we can see that momentum consists of the addition and combination of two parts, the continuity factor term and the criticality factor term. For the continuity factor term, it is again composed of two factors, the resultant and the state nature of continuous events. And we think that the occurrence of continuous events of statehood has a greater impact than the occurrence of continuous events of outcome. That is $\omega_{3,4} > \omega_{1,2}$.

The resultant nature of a continuous event consists of consecutive points scored R_1 and consecutive points lost R_2 , and the status of a continuous event consists of consecutive successful serves R_3 and consecutive service errors R_4 . By analyzing the data, we get $\omega_1=0.4, \omega_2=-0.4, \omega_3=0.6$ and $\omega_4=-0.6$. For the criticality factor, it is further composed of two factors: key skills behavior E_i and key ball round K . E_i is affected by four factors, namely, ACE ball launch, winning points, double faults and unforced errors, and has three values: 0, 1 and 2. The value of E_i is 2 when double faults and aces occur, 1 when winning points and unforced errors occur, and 0 when none of the above four factors occur.

For the key ball round K , when $K=0$, the moment is not a key ball. When $K=1$, the score situation is a 40:40 draw. When K takes the value 2, the player is on the receiving side and is behind the opponent, the score situation is 40:AD, 15:40 or 30:40, or the player is on the serving side and is ahead of the opponent, the score situation is AD:40, 40:15 or 40:30. When K takes the value 3, the player is on the receiving side and is ahead of the opponent, which is the break point, or the player is on the serving side and is behind the opponent, which is the opponent's break point. ω is the impact factor of the continuity

factor term. Φ and Ψ are the impact factors of key skills behavior and key ball turn, respectively. By analyzing the data, we make both Φ and Ψ equal to 1.2. The latter two influence factors are multiplied together to influence the critical factor. In other words, if the values of both factors are not zero, the multiplication will have a superimposed effect on the key factor term. Interpretation of Momentum Scores in Figure 6, and Account for Mos System in Figure 7.

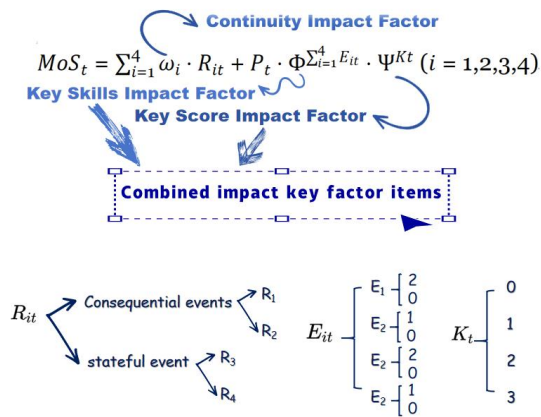


Figure 6: Interpretation of Momentum Scores

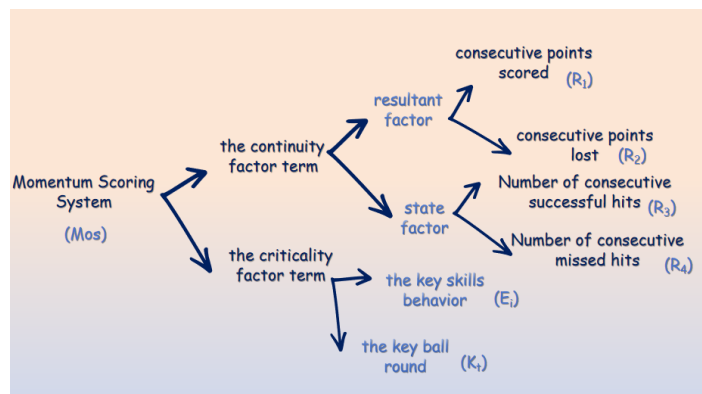


Figure 7: Account for Mos System

4.2.2 Feature Engineering for Match Swing with Light GBM model

Having an in-depth analysis of the chronology chart of the performance of the players^[10], we want to see the trend of the game through the image. Also prior to this, a preliminary correlation coefficient matrix analysis was performed on the data in Figure 8.

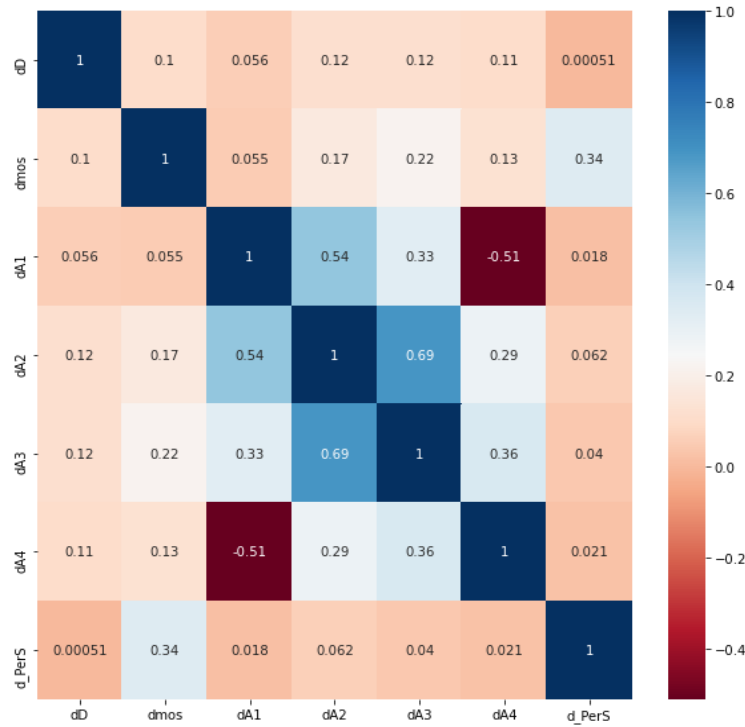


Figure 8: correlation-coefficient-matrix-analysis-before-feature-engineering

In terms of the definition of the derivative, that is, the derivative of $PerS$ with respect to the time t . However, we take into account that computationally, if we select this indicator for the next step of the analysis there will be a point of no significance, so we will turn it into another indicator: when the situation at the moment t to turn to the $\Delta PerSt$, which we defined as follows

$$\Delta PerSt = PerSt - PerSt_{-1}. \quad (4)$$

When $\Delta PerSt > 0$, the match advantage will be shifted from the opponent to that player, and when $\Delta PerSt < 0$, the match advantage will be shifted from that player to the opponent. For its influencing factors, we have reviewed the relevant literature and summarized them as the momentum gap between two players ΔMoS , the physical exertion gap ΔD and the skill-tactics gap ΔA . Their calculation formulas are as follows

$$\begin{cases} \Delta MoS = MoS_1 - MoS_2 \\ \Delta D = D_1 - D_2 \\ \Delta A_1 = A_{11} - A_{12} \\ \Delta A_2 = A_{21} - A_{22} \\ \Delta A_3 = A_{31} - A_{32} \\ \Delta A_4 = A_{41} - A_{42} \end{cases} \quad (5)$$

In the above equation, we quantify the physical exertion gap as the difference in distance run by two players. When $\Delta D > 0$, it proves that the player has a large consumption and a small physical advantage, and when $\Delta D < 0$, it is the opposite. For the skill and tactical indicators, A_1 stands for the success rate of the first serve, A_2 stands for the scoring rate of the first serve, A_3 stands for the scoring rate of the second serve, and A_4 stands for the probability of successfully receiving the ball from the server and scoring a point.

4.3. Combined match swing identification forecast

We trained the two models separately and constructed the combined prediction model as follows
First, define the prediction error of the time series model as

$$\text{err}^{(1)} = \text{RMSE}^{(1)} = \left(\sqrt{\frac{1}{n} \cdot \sum_{i=1}^n \left(y_i^{(1)} - y_i'^{(1)} \right)^2} \right) \quad (6)$$

In this formula, y_i is the real match swing, $y_i(t)$ is the predicted match swing. Similarly, the prediction error of l_t in this model is defined as

$$\text{err}^{(2)} = \text{RMSE}^{(2)} = \left(\sqrt{\frac{1}{n} \cdot \sum_{i=1}^n \left(y_i^{(2)} - y_i'^{(2)} \right)^2} \right) \quad (7)$$

The weights assigned to the predicted values of the time series model are defined as

$$\rho_{(1)} = 1 - \frac{\text{err}^{(1)}}{\text{err}^{(1)} + \text{err}^{(2)}} \quad (8)$$

Then, since the sum of the weights of the two models is 1, the weights assigned to the predicted match flows are as follows

$$\rho_{(2)} = 1 - \rho_{(1)} = \frac{\text{err}^{(1)}}{\text{err}^{(1)} + \text{err}^{(2)}} \quad (9)$$

Finally, we get the combined prediction model as follows

$$F(x) = \rho_{(1)} I_t + \rho_{(2)} L_t \quad (10)$$

4.4. Model feature importance output

Through the data processing above, the model feature importance is outputted in Figure 9. Through this figure, we come to the following conclusion: compared to other factors, *Mos* has the greatest impact on match volatility with an importance level of more than 0.35, followed by the *D* factor with an importance level of more than 0.3. The importance of *A1*, *A2*, *A4* and *A3* decreases in order. *A3* has the least influence on match fluctuation.

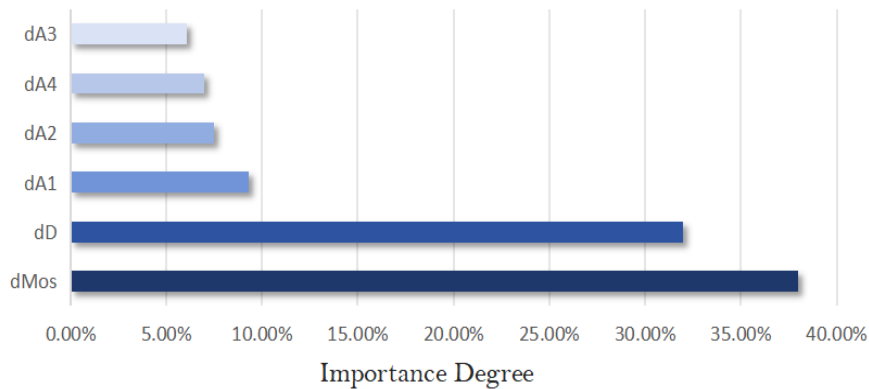


Figure 9: Feature Importance of LightGBM

4.5. Model Evaluation and Testing

In this paper, select the data of the men's singles semifinal match between Novak Djokovic and Sinner, match ID 1602, for testing the model, and we processed the data according to the process described in the model above, and then obtained the prediction results for this match. The curve identified as *F* in the image is the final result of the combined model. At the same time, we carried out a comparison with the original data of the game based on this curve and found that our predicted results were extremely similar to the facts of what actually happened in Figure 10. Therefore the model we built passed this test.

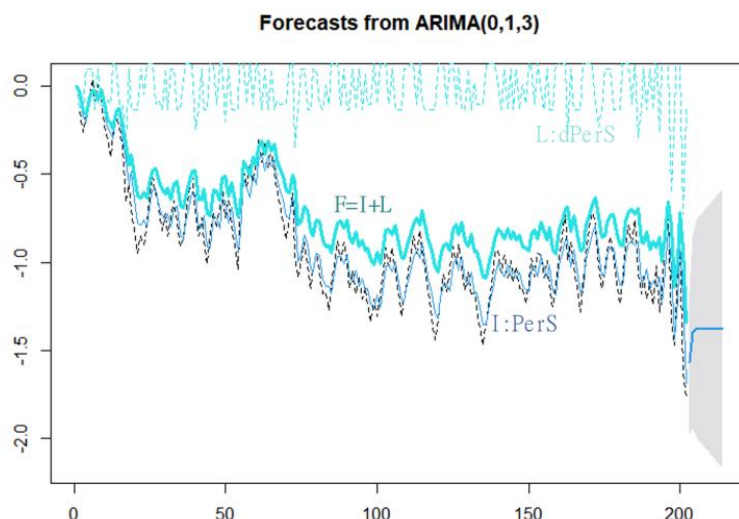


Figure 10: Chart of predicted results for contest ID 1602

5. Conclusions

For the scored time series we used wavelet analysis to filter the original time series, enabling the monitoring of situational transitions, the identification of periodic components and the analysis of multiple time scales. In addition, the use of wavelet analysis is highly adaptive to the data.

For the Kruskal-Wallis test we use the main advantage is that it does not require the sample to satisfy the normality assumption and is more widely applicable. It is also more generalizable since it only relies on the rank sum of each set of data and is not affected by extreme values.

The benefits of using the ARIMA-LightGBM model^[8] for match swing prediction are that it saves a lot of time, is more interpretable, and has advantages in the extraction and prediction of nonlinear relationships. The model combines the advantages of Random Forest Model, Deep Learning Model and avoids their disadvantages. The most important thing is that the model has practical significance for the accurate prediction of this question. At the same time, we verified the prediction accuracy of the model using other groups of historical game data.

Through the deep mining of big data, we quantify the player's momentum into several core indicators, and we also quantify the player's psychological factors into parameters as much as possible, which are added to the model for evaluation and prediction. For the prediction model, we made a greater innovation, synthesized the advantages of most previous models, and verified the accuracy and universality of its model with actual data. Closer to the truth of the data in terms of predictive effectiveness, with relatively small deviations.

References

- [1] WANG Qiusi, GALIN GAHENNUR, DING Lan et al. Research progress based on big data analysis of open source information of professional tennis[C]//Chinese Society of Sports Science. Abstracts Collection of the 13th National Sports Science Conference - Wall Communication (Athletic Training Section) (II). Beijing Sport University; 2023:3. DOI:10.26914/c.cnkihy.2023.074111.
- [2] Claudino, J.G., Capanema, D.d., de Souza, T.V. et al. Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: A Systematic Review. *Sports Med - Open* 5, 28 (2019).
- [3] de la Iglesia R, Espinosa-Salinas I, Lopez-Silvarrey FJ, Ramos-Alvarez JJ, Segovia JC, Colmenarejo G, Borregon-Rivilla E, Marcos-Pasero H, Aguilar-Aguilar E, LoriaKohen V, Reglero G and Ramirez-de Molina A (2020) A Potential Endurance Algorithm Prediction in the Field of Sports Performance. *Front. Genet.* 11:711.

- [4] ZHAO Mingnan, JIANG Haoqi, QI Bing et al. Theoretical modeling and scale development of emotional intelligence in Chinese tennis players [J]. Psychological Research, 2024, 17(01): 42-53. DOI:10.19988/j.cnki.issn.2095-1159.2024.01.005.
- [5] Huang ZY. Research on tennis match performance prediction model based on BP neural network[D]. Fujian Normal University,2022.DOI:10.27019/d.cnki.gfjsu.2022.000427.
- [6] Mccullagh, John. (2010). Data Mining in Sport: A Neural Network Approach. Int. J. of Sports Science and Engineering. 4.
- [7] Xu Q. Zhang Yankui. (2022). Exploration of dynamic sales forecasting model for retail apparel based on ARIMA model and random forest model. China Management Informatization(06),100-104.
- [8] Ye Li-Ming, Shi Shi-Liang, Lu Yi, Li He Zeng Ming-Sheng. (2022). Coal mine safety posture prediction based on ARIMA and XGBoost algorithm. Safety (02), 53-59. doi:10.19737/j.cnki.issn1002-3631.2022.02.008.
- [9] Wang, Jiancheng Cai, Yanguang. (2019).Comparison of the performance of LightGBM algorithm and ARIMA algorithm in population flow prediction applications. Journal of Dongguan Institute of Technology (05), 27-32+44. doi: 10.16002/j.cnki.10090312.2019.05.005.
- [10] Guo Wenxia. Construction and application of tactical diagnostic index system for hard court tennis men's singles matches [D]. Beijing sports university, 2018.