ISSN 2616-5775 Vol. 7, Issue 8: 27-32, DOI: 10.25236/AJCIS.2024.070805

Study on Dynamic Prediction Model for Tennis Matches Based on Multi-Dimensional Indicator Evaluation

Dishen Yang[#], Qingyang Zhang[#], Baihe Luan[#]

Faculty of Science and Technology, University of Macau, Macau, 999078, China
[#]These authors contributed equally.

Abstract: *In modern tennis, data analytics and artificial intelligence are key for predicting crucial match turning points and forming strategies. This study enhances the understanding of tennis momentum by focusing on strategic turning points and players' diverse skills. The study uses the TOPSIS method to quantify player attributes and integrate these with key momentum indicators using Principal Component Analysis and Factor Analysis. This results in nine principal components, with three particularly influential ones—emphasizing the importance of service quality, physical fitness, and mental stability in match outcomes. The model is highly effective, demonstrated by a Kendall coordination coefficient of 0.933. The study also suggests potential improvements by including factors like environmental conditions to increase model generalizability. Our findings improve insights into player performance dynamics and enhance the predictive accuracy of match results, supporting the development of sophisticated strategies for players and coaches. The research highlights the significance of advanced analytical methods in leveraging the subtle yet impactful elements of momentum in professional tennis.*

Keywords: *TOPSIS, Principal Component Analysis, Factor Analysis, Prediction Model, Kendall Consistency*

1. Introduction

In modern tennis matches, the outcome often depends on players' performances at critical moments. Effectively predicting potential turning points during the match and developing corresponding strategies is crucial for both players and coaches. The performance is influenced not only by technical skills but also by psychological state, physical condition, and match strategies. However, with the advancement of data analysis and artificial intelligence technologies, using data-driven methods to analyze and predict dynamic changes in tennis matches has become an important research direction.

Technics for predicting the tennis match flow swings are becoming more sophisticated. TOPSIS[1] can objectively screen out important indicators, which have been used to evaluate the contribution quality of table tennis matches[2]. The algorithms of principle component analysis have already been used to evaluate the external factors influencing the athlete[3]. Therefore, these methods are appropriate to be used in predicting the match flow.

The study references data from https://www.comap.com and employs the TOPSIS method to objectively and hierarchically weigh data, which quantitatively evaluate players' characteristics through a weighted linear combination of six evaluation dimensions, and visualize these characteristics with a radar chart. Additionally, PCA and FA are used to reduce the dimensionality of 15 indicators to mitigate the impact of noise and preserve essential data features. PCA is applied first to determine the optimal number of principal components, followed by FA to analyze the relationships and significance of sub-factors within these components, culminating in the creation of a heatmap of the factor loading matrix. The prediction model is then established by defining the “swings”of a match, demonstrating a high Kendall coordination coefficient.

2. Comprehensive Assessment of Tennis Player Performance Factors

In this case, the study uses TOPSIS, PCA, and FA to acquire indicators that can help determine when the flow of play is about to change from favoring one player to the other, which is practical during the competition. The study abstracts it as a combination of evaluation and prediction problem.

ISSN 2616-5775 Vol. 7, Issue 8: 27-32, DOI: 10.25236/AJCIS.2024.070805

2.1 Quantitative Evaluation Model for Indicators

Recognizing the differences in skill levels and playing strategies of the players, here first analyze the players' characteristics on six general dimensions, namely Mental Factor, Defense Factor, Offense Factor, Physical Factor, Technical Factor, and Stability Factor, and then quantify them for evaluation. Here takes all relevant data of the sub-factors in the given data set. The specific extraneous sub-factors used for evaluation are listed in Table 1.

Table 1: Sub-factors of evaluation

Factors	Sub-factors	Explanation
Mental Factor	reverse victory	number of reverse victories/numbers of victory
	break saving rate	number of breaks that were saved/number of breaks
	success rate of tiebreaker	number of successful tie breaks/number of tie breaks
Defense Factor	return depth D	number of D in return depth/number of D and ND in return depth
Offense Factor	rate of winner	number of winner/total points gained
	the success rate of break	number of successful breaks/total number of breaks by the competitor
Physical Factor	distance run	the average running distance at each point in a match
Technical Factor	success rate of net_pt	sum of net_pt/sum of net_pt won
Stability Factor	rate of unf_err	number of unf_err/total points gained

First, all the data are normalized and regularized. Eight sub-factors are then weighted hierarchically after weighting objectively using the TOPSIS method. As for factors affected by one single sub-factor, the obtained sub-factor weights are equal to the dimension weights. For evaluation dimensions that are affected by multiple sub-factors, the weights are re-assigned in proportion to the weights of the sub-factors within the same dimension. The weighting and linear combinations are first performed within that dimension. The weight of such evaluation dimension is the sum of the weights of its sub-factors. Finally, the weighted linear combination of the six dimensions is performed to obtain the quantitative evaluation of the players' characteristics, followed by Figure 1.

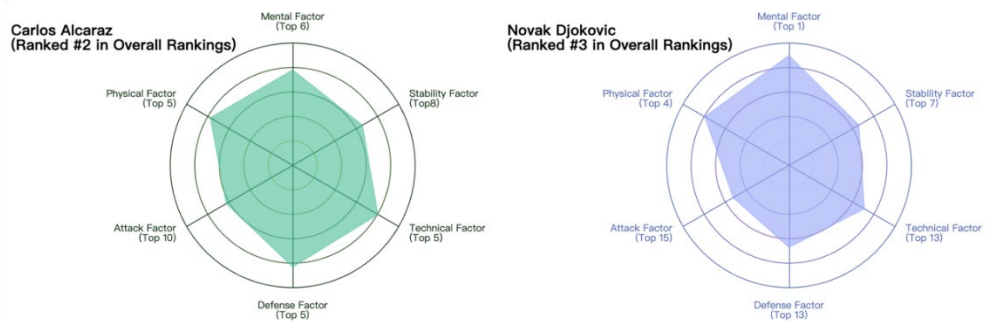

Figure 1: Radar Chart of N. Djokovic and C. Alcaraz

Using the two finalists, Carlos Alcaraz and Novak Djokovic, as an example, their player characteristics radar chart metrics are evenly distributed, and their indicator rankings are on the high side. Moreover, the overall rankings of Carlos Alcaraz and Novak Djokovic are in second and third place among all the 32 players, separately, which is in line with the actual results of the match.

After examining the processed data, here use the 14 indicators mentioned in Problem 1 and add Characteristic as the 15th indicator. Obviously, with too many evaluation criteria, noise may have a large impact on the model. Moreover, the covariance, significance, and other exclusionary relationships of the degree of the dimension itself should be taken into account. As such, here use Principal Component Analysis and Factor Analysis to downscale the normalized data while maximizing the preservation of the original data features.

Calculate the sample correlation coefficient matrix R, the eigenvalues, and the eigenvectors of this matrix. It is found that the cumulative variance explained reaches 82.652% at a component number of 9

ISSN 2616-5775 Vol. 7, Issue 8: 27-32, DOI: 10.25236/AJCIS.2024.070805

and slows down thereafter, so the study ultimately takes 9 as the number of principal components based on Principal Component Analysis (PCA).

Figure 2: Factor loading matrix heat map

By calculating the factor loading coefficients, the correlation between the 15 sub-factors is related to the principal components obtained in the previous step, and the importance of the sub-factors in each principal component is derived. After that, here introduce the common degree to describe the amount of information that each extracted principal component contains about the original data. The factor loading matrix heat map is shown in Figure 2.

It is clear from the calculation results that the common degree of all 15 factors is close to 0.7 or greater than 0.7, indicating that the obtained principal components can better cover the information provided by the original data.

3. Evaluating and Optimizing Swing Prediction Models

3.1 Prediction Model for Swings

According to the rules of tennis, a player wins a game when reaching 4 points but must win by two, so this study considers that a swing occurs when the absolute value of the difference between a player's point difference at point i and the point difference at point (i+6) is greater than or equal to 4, i.e.

$$|D_{i+6} - D_i| \geqslant 4 \tag{1}$$

where D_i denotes the difference of points between two players at point i.

If a swing occurs, then it assigns the value to 1. Otherwise, the value is defined to be 0.

The principal components obtained from the previous analysis were fed into the LM-BP neural network for learning to quantify the likelihood of a swing occurring, and the predicted results are shown in Figure 3:

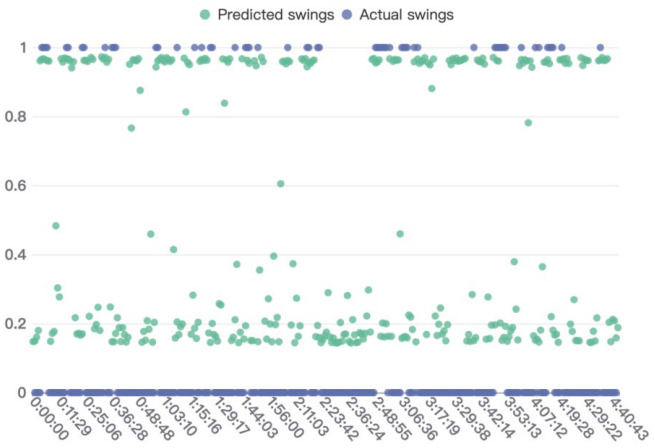

Figure 3: Visualization of predicted swings and actual swings

ISSN 2616-5775 Vol. 7, Issue 8: 27-32, DOI: 10.25236/AJCIS.2024.070805

As can be seen from the comparisons shown in the scatterplot, the high likelihood points for the occurrence of fluctuations are very close to the points with value 1. They also coincide very closely with when the actual wave occurred. This indicates that the neural network accurately captures the patterns associated with swing occurrence.

Similarly, the low likelihood points also exhibit a consistent pattern. They align closely with the points assigned a value of 0, indicating that the neural network correctly identifies instances where swings are less likely to occur. This suggests that the network has learned to differentiate between swing and non-swing instances, thereby providing a reliable measure of likelihood.

The network effectively learns and predicts swing occurrences, as evidenced by the close alignment between high-likelihood points and actual wave events. The consistency between low likelihood points and non-swing instances further supports the network's accuracy.

3.2 Strategy Suggestions

Based on the LM-BP neural network learning results, this study obtained 9 principal component weights. From the calculation results, it can be seen that principal components 1, 4, and 6 have the highest weights, which are 5.9213, 7.1024, and 7.2767, respectively. Depending on the correlation between the sub-factors and the principal components, this study can conclude that service, energy level, and psychological stability are the most relevant indicators.

Consequently, strategic recommendations have been formulated based on the circumstances of the tennis match. When entering a new match against a different player, several suggestions are provided below following the models and evaluation system:

1) Before the match: Use the characteristic evaluation system to make a rough assessment of the competitor, based on their previous performance.

2) During the match: When the critical indicators, namely, serving situation, energy level, and mentality change, adjust the strategies in case any string of matches happens.

3) After the match: Using the momentum evaluation and flow capturing model to have an overview of the match and make reflections on their performance.

3.3 Swings Prediction Model Analysis

This study considers analyzing the prediction model for swings obtained in the last section concerning three dimensions: prediction accuracy, level of generalization, and sensitivity analysis.

Apply the swings prediction model to the two matches in the given data that were not subjected to LM-BP neural network learning. Subsequently, Kendall's consistency test[4] was introduced to analyze the prediction results against the actual data.

As can be seen from the calculated data, the P-value is less than 0.05, which presents significance, and it is considered that there is consistency between the two sets of data, the predicted and the true values. Meanwhile, Kendall's W coefficient value is 0.933, so the degree of correlation is high with almost complete consistency.

Although this study optimized the accuracy of the predictive model by considering as many factors as possible, the model also sacrificed generalizability for this purpose. When attempting to test the model in other competitions, factors had to be added or removed.

On the one hand, data with the same level of detail and comprehensiveness as Wimbledon_featured_matches.csv is not readily available, which means that it is not entirely applicable to other tennis tournaments such as the U.S. Open, Australian Open, or Women's Tennis, as the venue and gender factors should be taken into account as well.

On the other hand, considering that tennis has different rules than other sports, some factors such as "breakpoint" may not always be present. They are not general enough to be applied or tested in other sports.

If the prediction effectiveness is rather poor sometimes, the following factors may be possible indicators:

Firstly, test every player's average oxygen uptake VO_2 (mL * kg^{-1} * min^{-1})[5] for a game. VO_2 can be

ISSN 2616-5775 Vol. 7, Issue 8: 27-32, DOI: 10.25236/AJCIS.2024.070805

a better guide to the intensity of tennis matches, rather than the data of running distance when reflecting the stamina of players.

Secondly, detect the contestant's recent match frequency[6]. The player has participated in matches in the past short term so the fatigue state of the player will be considered.

Thirdly, collect the data on weather factors. The momentum of a match may swing due to wind or rain.

Sensitivity analysis is a method to quantitatively describe the importance of the input variables to the output variables of the model and determine the influence of variables on the output of the model through the sensitivity coefficient, which helps to solve the problem and optimize the model. This study chooses a single sample sensitivity coefficient to reflect the input variable to the output variable of the model and uses the full sample sensitivity coefficient to directly reflect the rank of contribution rate of the input variable to the output variable.

The input of the *k*th neuron of the output layer at the sample point is expressed as:

$$r_k^t = \sum_{l=1}^{L} v_{lk} a_l^t = \sum_{l=1}^{L} v_{lk} f\left(\sum_{n=1}^{N} w_{nl} x_n^t\right) \tag{2}$$

Represent the single sensitivity coefficient of *t*he sample's *i*th input variable x^t_i to *k*th output variable y^t_k:

$$s_{i,k}^t = \frac{\partial y_k}{\partial x_i} = g'(r_k^t) \sum_{l=1}^{L} v_{lk} f'(q_l^t) w_{il} \tag{3}$$

The full sample sensitivity coefficient in total *T* samples is expressed as:

$$S_{i,k} = \sqrt{\sum_{t=1}^{T} (s_{i,k}^t)^2 / T} \tag{4}$$

Here $S_{i,k}$ is the sensitivity coefficient of input factor x_i to output factor y_k, and T is the number of samples.

By calculation, the full sample sensitivity coefficients for the 14 indicators are shown in Figure 4:

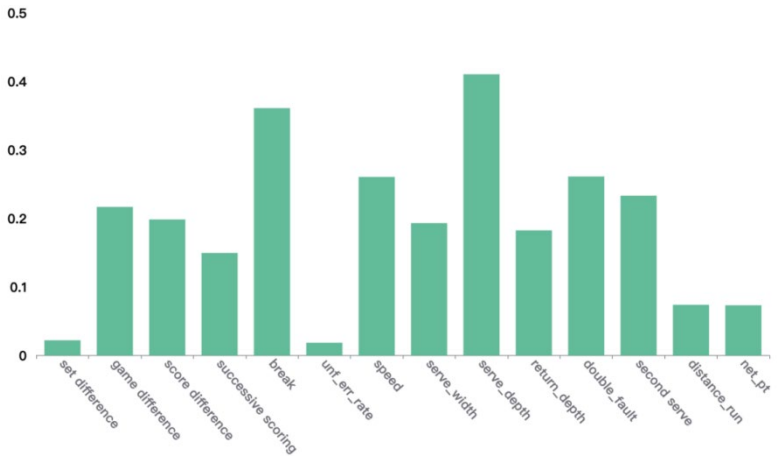

Figure 4: Full Sample Sensitivity Coefficient of the 14 Indicators

As can be seen from the figure, the sensitivity coefficients of all the factors are distributed more evenly, with the difference not exceeding 0.4, with 'serve-depth' being the most sensitive.

Because the sensitivity coefficient explains the relative importance of each feature to the model output, a higher sensitivity coefficient indicates that the feature has a greater impact on the model output, while a lower sensitivity coefficient indicates that the feature has a lesser impact on the model output. Therefore, it can be obtained that the difference in the effect of each factor on the model output is not significant, then the selection of features in this study is reasonable and the model prediction is relatively stable.

ISSN 2616-5775 Vol. 7, Issue 8: 27-32, DOI: 10.25236/AJCIS.2024.070805

4. Conclusion

In summary, this paper establishes a prediction model based on quantitative evaluation and the LM-BP neural network, which can be used to predict the emergence of swings in tennis matches. The modeling is based on player characteristics and actual performance on the court. TOPSIS and hierarchical weighting are applied to evaluate the player characteristics from six dimensions, i.e., Mental, Physical, Stability, Offense, Defense, and Technical, and visualized as radar charts. To eliminate inter-dimensional covariance and noise effects, this paper introduces Principal Component Analysis (PCA) and Factor Analysis (FA) to downscale the influencing factors and quantitatively evaluate the players' game performance. The appearance of swing is determined by the difference between the score difference of two players in six adjacent games, which is combined with the previous factors to establish a prediction model based on the LM-BP neural network. Comparing the prediction results with the actual results according to the available data, the Kendall coefficient W is 0.933, which shows good accuracy. Meanwhile, the calculation results of the single sample sensitivity coefficient show that the sensitivity coefficients of most of the factors are more average, indicating favorable stability of the model.

This paper provides a research framework based on a data-driven approach to analyzing and predicting the dynamics in tennis matches, which contributes to solving the problem in the flow of sports events toward related fields. Given the background of the digital era, the accuracy and availability of match data ensure the feasibility of the methodology proposed in this paper.

References

[1] CHAKRABORTY S. TOPSIS and Modified TOPSIS: A comparative analysis[J/OL]. Decision Analytics Journal, 2022, 2: 100021.
[2] YIN H, CHEN X, ZHOU Y, et al. Contribution quality evaluation of table tennis match by using TOPSIS-RSR method - an empirical study[J/OL]. BMC Sports Science, Medicine and Rehabilitation, 2023, 15(1): 132.
[3] OLIVA-LOZANO J M, ROJAS-VALVERDE D, GÓMEZ-CARMONA C D, et al. Impact of contextual variables on the representative external load profile of Spanish professional soccer match-play: A full season study[J/OL]. European Journal of Sport Science, 2021, 21(4): 497-506.
[4] MOSLEM S, DEVECI M, PILLA F. A novel best-worst method and Kendall model integration for optimal selection of digital voting tools to enhance citizen engagement in public decision making[J/OL]. Decision Analytics Journal, 2024, 10: 100378.
[5] REID M, DUFFIELD R. The development of fatigue during match-play tennis[J/OL]. British Journal of Sports Medicine, 2014, 48(Suppl 1): i7-i11.
[6] SMEKAL G, VON DUVILLARD S P, RIHACEK C, et al. A physiological profile of tennis match play:[J/OL]. Medicine and Science in Sports and Exercise, 2001, 33(6): 999-1005.