A Study on Momentum Prediction for Tennis Matches Based on Random Forest and Support Vector Machine Algorithms

Yizhi Wang *, Jiaming Cao and RuiHu

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

* Corresponding Author Email: wangyizhi2003@nuaa.edu.cn

Abstract. In this study, based on the 2023 Wimbledon Championship finals data, we use the random forest model and support vector machine to quantify the match momentum. Firstly, we built a random forest regression prediction model and used data preprocessing and integrated learning methods to improve the accuracy of the prediction. By using the pandas and Scikit-learn libraries in Python for model training and prediction, we successfully quantify the effect of momentum on race outcomes. Secondly, we demonstrate that the volatility of match results is not random and illustrate the importance of momentum in competitive matches, showing the correlation between momentum and match wins and losses through correlation tests and support vector machine algorithms. Finally, we verified the relationship between momentum calculation and match results by the Kolmogorov-Smirnov test, summarised the strengths and weaknesses of the model, and presented an outlook for future improvement. In summary, this study provides a new approach to understand the influencing factors of match momentum and provides a useful reference for future related studies.

Keywords: Random forest; support vector machine; Kolmogorov-Smirnov.

1. Introduction

Momentum is widely recognised as a key concept in sports competitions and is widely believed to have a significant impact on the outcome of a match. The heated duel in the final of the 2023 Wimbledon Championships between Carlos Alcaraz and Novak Djokovic triggered an in-depth discussion on the role of momentum in a match. Using this final as a case study, this study aims to quantify and analyse the performance of momentum in matches by applying machine learning techniques such as Random Forest Models and Support Vector Machines. Momentum is not only a reflection of the state of the game, but also a comprehensive reflection of the players' psychological state and game trends, which is important for understanding the key moments during the game and the outcome of the winners and losers.

In this paper, we aim to reveal the importance of momentum in a match and its correlation with the match results through a data-driven approach. First, we develop a random forest regression prediction model to quantify momentum and optimise the model to improve prediction accuracy. Second, we use correlation tests and support vector machine algorithms to validate the relationship between momentum and match trends and to explore the mechanism of momentum's role in matches. Finally, we further validate the validity of the model and the association between momentum and match winners and losers through the Kolmogorov-Smirnov test. This research aims to provide new perspectives and methods for analysing momentum in sports matches, to provide more in-depth tactical guidance for athletes and coaches, and to contribute to the development of the field of sports data science.

2. Winner Prediction and Momentum Quantification in Random Forests

This paper collects data from the 2023 Wimbledon Championship finals and processes that data. This paper defines several new metrics, including "oppo_set", "opp_game", "opp_point", "next_winner", "ace", "winner", "double_fault ", "unf_err" and "win_streak".

2.1. Random Forest Algorithms for Predicting Winners and Quantifying Momentum

According to several debugging of the existing feature metrics, for tennis player 1, the feature metrics such as server, p1_ ace, p2_ ace, points_winner, p1_winner, p2_winner, and type of winner's stroke have a significant effect on the results of each match. In order to improve the accuracy of Random Forest prediction, we pre-processed the data reasonably. In this paper, we use random forest regression prediction model. Random forest is an algorithm that integrates multiple trees through ensemble learning ideas [1]. Its basic unit is a decision tree, and its essence belongs to an important branch of machine learning - ensemble learning method [2].

Suppose that there are N tennis players, and there are M features. Each sample has a corresponding marker y_i ,then T decision trees are built, where each tree is randomly selected on a sample subset and only a part of features is considered in the construction of each tree Each feature is called an independent variable [3]. All independent variables can be expressed as $\{X_1, X_2 \ldots X_{m-1}, X_m\}$. The training set includes feature data of all athletes in the 2023 Wimbledon competition, and the test set is 2023-wimbledon-1701. The specific steps of random forest are shown in Fig. 1 below:

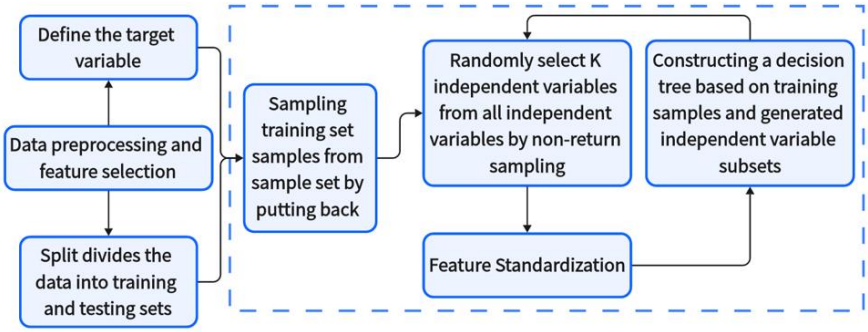


Fig. 1 Random Forest Algorithm

Feature importance and quantified momentum is visualized to dig into the inherent rules, which is help for subsequent modeling. Fig. 2 shows the proportion of different features in quantifying momentum. It can be seen from Fig. 2 that the server S_e has the greatest influence on the momentum and the outcome of the game. At the same time, the opportunity in point (game or set) also has a great influence on the momentum and the outcome of the game.

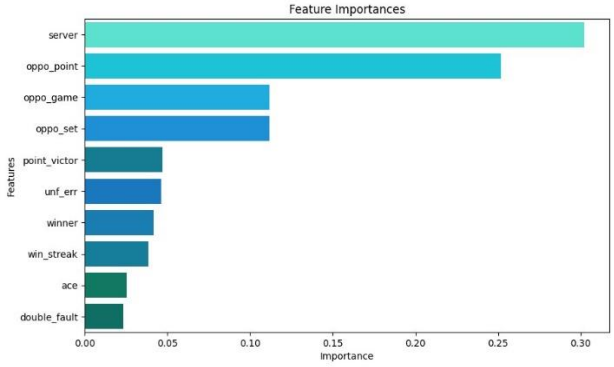


Fig. 2 Feature Importance

2.2. The Model Solution

Through the regression model established through random forest, we can obtain the instantaneous value of momentum at a certain moment after quantification and the total momentum obtained through accumulation. Suppose that p is some characteristic indicators of a tennis player at time t,

and $p = [p_1, p_2, \ldots, p_m]^T$,where x_i is the value of the $i\text{-}th$ feature obtained from the dataset. The formula for calculating momentum is as follows:

$$m_t = w_T p \tag{1}$$

In the formula, $w_T = [w_{T1}, w_{T2}, \ldots, w_{Tm}]$, where w_{Ti} is the coefficient corresponding to the $i\text{-}th$ feature index obtained by the random forest algorithm.

And $M_{(t_1,t_2)}$ represents the total momentum is the sum of all momentum in a period of time, which can reflect the comprehensive performance of tennis players. Its calculation formula is as follows:

$$M_{(t_1,t_2)} = \begin{cases} \sum_{i=0}^{n} m_{ti}, Time\ variables\ are\ discrete\ variables \\ \int_{t_1}^{t_2} m_t, Time\ variables\ are\ continuous\ variables \end{cases} \tag{2}$$

Fig. 3 shows the change of momentum with time after quantification. It can be seen that the player's momentum fluctuates greatly with time in this game. Fig. 4 shows the change of total momentum with time, which can reflect the change trend of tennis players' performance in one game.

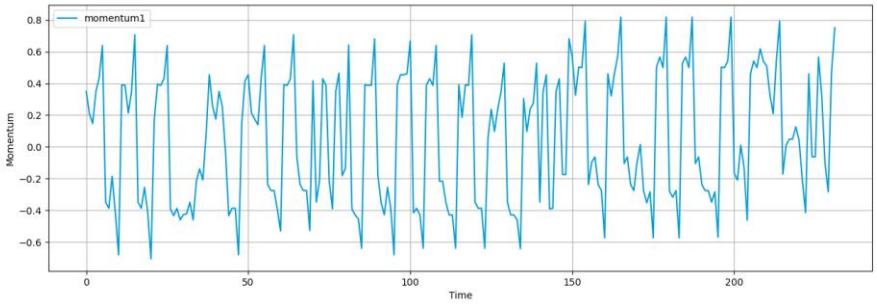


Fig. 3 Momentum-Time Diagram

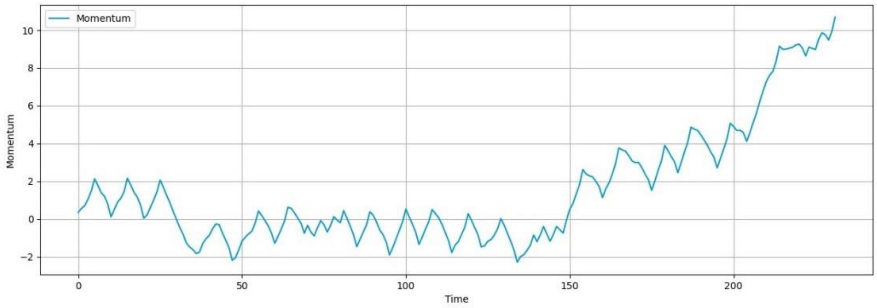


Fig. 4 Momentum-Time Diagram

After evaluating the model, we can obtain an accuracy of 0.656142759, which can effectively predict the trend of the competition. We can use the predicted results as an indicator to quantify the momentum. Fig. 5 shows the confusion matrix, and Fig. 6 shows the ROC, both of which indicate that the momentum quantified by the random forest model can reflect the flow of the play.

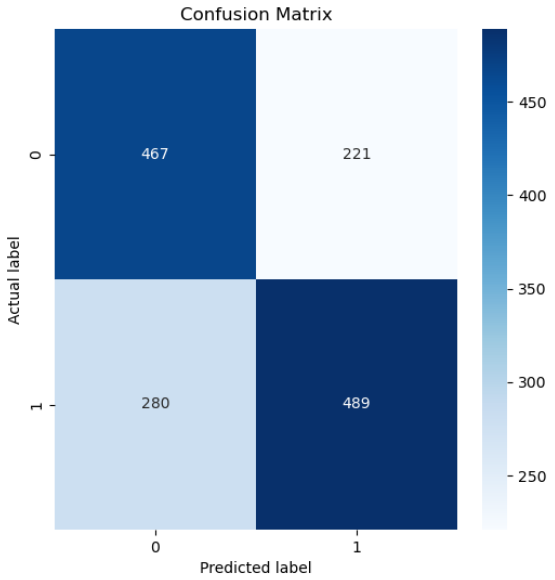


Fig. 5 Confusion Matrix

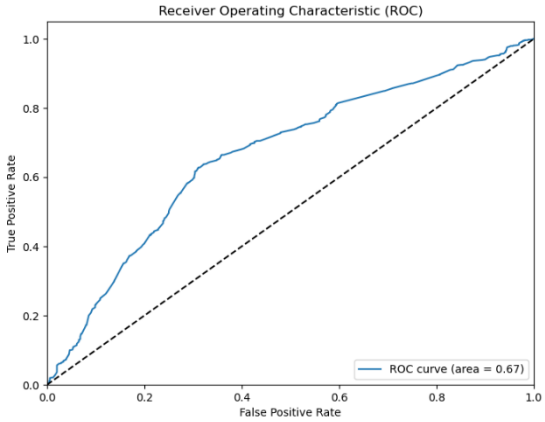


Fig. 6 Receiver Operating Characteristic

3. Proof of Momentum Existence and Regularity Based on SVM and Spearman Test

3.1. Proof of Momentum Existence and Regularity Based on Spearman Test

We need to prove that swings in play and runs of success are not random before proving that momentum plays a role in competition. We use correlation coefficient test to prove that the randomly generated data has no correlation with momentum. Because Spearman correlation coefficient does not require all the distribution of variables, this article will use Spearman correlation coefficient Test [4].

Data can be visualized before correlation test. Fig. 7 shows random data and momentum. It can be seen from the figure that there is no strong correlation between random data and momentum.

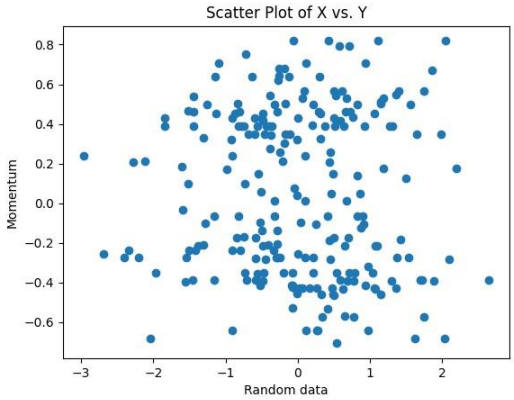


Fig. 7 Random Data-Momentum Figure

Spearman rank correlation coefficient is used to explain the close degree and correlation direction of linear correlation between variables by rank correlation coefficient r_s. Sort the original data x_i and y_i in ascending order, denote x'_i. Record x'_i, y'_i as the original x_i, y_i's position in the sorted list, x'_i, y'_i as the rank of x_i, y_i, and the difference in rank $d_i = x'_i - y'_i$. d_i reflects the consistency of the rank arrangement of X and Y variables. The formula of Spearman rank correlation coefficient is:

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2-1)} \tag{3}$$

In the formula, n is the number of samples.

Finally, the Spearman correlation coefficient is 0.03125976. It can be seen that the correlation coefficient is small, and it can be concluded that the two variables are not directly related, so the change of available momentum is not random.

3.2. Proving the Regularity of Runs of Success by Support Vector Machine

To demonstrate that momentum plays any role in the match, we use the Support Vector Machine algorithm. Support Vector Machine (SVM) is a binary classification model. Its basic model is a linear classifier defined in the feature space with the largest interval: support vector machines also include kernel techniques, making it essentially a nonlinear classifier. The learning strategy of support vector machines is to maximize the interval, which can be formalized as a problem for solving convex quadratic programming. The learning algorithm of support vector machine is an optimization algorithm for solving convex quadratic programming [5].

The basic formula of SVM mainly involves the distance from sample point to hyperplane and the decision boundary of hyperplane. Specifically, given sample point $\boldsymbol{x}$ and normal vector $\boldsymbol{w}$ and intercept b of hyperplane, the distance from sample point to hyperplane can be expressed as:

$$r = \frac{|w^T x + b|}{\|w\|} \tag{4}$$

In the above formula, $\|w\| = \sqrt{w_1^2 + \cdots w_n^2}$

For Support Vector Machines, the goal is to find a hyperplane so that the hyperplane can correctly classify sample points and maximize the distance from the nearest sample point (support vector) to the hyperplane. This can be described by the following formula:

For positive class samples ($y_i = 1$), there are: $w^T x_i + b \geq +1$.

For the negative class sample ($y_i = -1$), there are: $w^T x_i + b \geq -1$.

Here y_i represents the label of the sample, and point x_i represents the eigenvector of the sample point. These formulas describe the relationship between the hyperplane and the sample points and ensure that the distance from the support vector to the hyperplane is at least $\frac{1}{\|w\|}$.

Fig. 8 shows the use of support vector machines to prove that tennis players' runs of success has regularity.

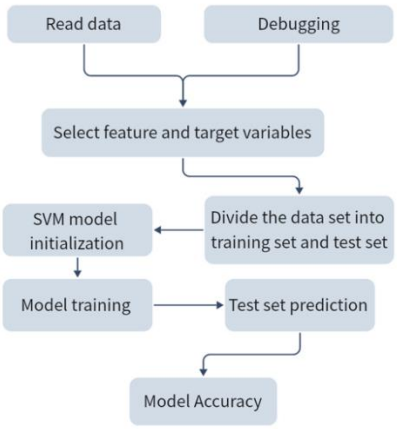


Fig. 8 Support Vector Machine algorithm

In the debugging stage, we choose some feature indexes to make the model accurate through continuous testing. These feature indexes are "server", "winner", "double_fault", "unf_err","ace","point_victor","point_no","p1_distance_run","oppo_set","oppo_game" and "oppo_point".

The accuracy of SVM model is 0.785175, which shows that the runs of success have a great relationship with the above feature indexes. At the same time, the above indexes can express the momentum through Model I, so swings in play and runs of success are regular.

4. Determination of Indicators Reflecting the Flow of Play Through Model Visualization and KS Test

Model visualization is a powerful tool that can intuitively reflect data relationships. Visualizing the next_winner and momentum (m_t of player1 in the 1307th Wimbledon game of 2023 yields Fig. 9.

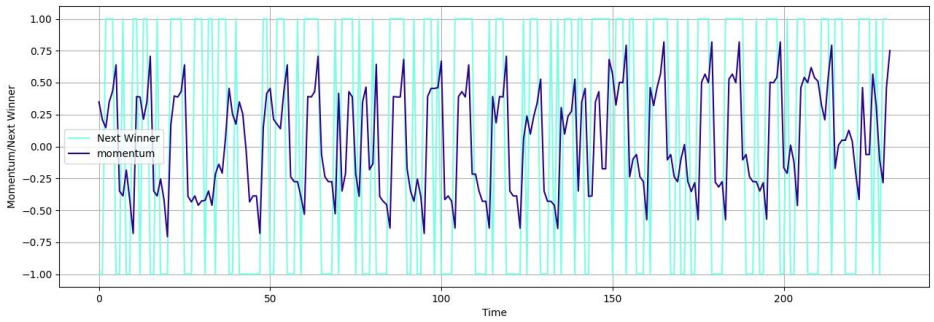


Fig. 9 Next Winner and Momentum-Time Figure

From the Fig. 9, we can intuitively see that when the instantaneous momentum is high, the player is more likely to win in the next game.

4.1. Kolmogorov-Smirnov Test Determining Indicators of the Flow of Play

Since the distribution of data on the winner of the next game is unknown, we use the KS test method to verify whether the momentum calculated using multiple indicators in Model 1 and Model 2 is related to the game victory. Kolmogorov Smirnov is a test method for comparing a frequency distribution $f\ (x)$ with a theoretical distribution $g\ (x)$ or two observed value distributions. Its original assumption H0: Two data distributions are consistent, or the data conforms to the theoretical distribution. $D = max\ |\ f\ (x)\ -\ g\ (x)\ |$, when the actual observed value $D > D\ (n, \alpha)$ Reject $H0$, otherwise accept $H0$ assumption [6].

The calculation of the KS Statistic can be done using the library Pandas, Scipy, Statsmodels. Having call the above library, you can directly use instruction "ks_stat = ks_2samp (,)" to calculate. After KS test, we obtained the KS Statistic of both, with a value of 0.5281385281. The Kolmogorov Smirnov test measures the maximum difference between the cumulative distribution function of the sample data and the cumulative distribution function of another sample. The value of the statistic is between 0 and 1, where 0 indicates that the two distributions are completely consistent and 1 indicates that the two distributions are completely different. 0.5281385281 indicates a certain degree of inconsistency between the cumulative distribution function of the sample data and the theoretical distribution (or the cumulative distribution function of another sample), but this value is not very large.

As a result, to some extent, momentum can reflect when the flow of play is about to change from favoring one player to the other. In Model 1 and Model 2, the server S_e accounts for the largest proportion in momentum, which, together with many other factors, determines the magnitude of the momentum and thus affects the flow of the play.

5. Model Application and Analysis

The above models are all based on the 2023 Wimbledon data as the training set and the final data as the test set. To verify the accuracy of the model, we will apply the resulting model to other matches.

Here are the test results of our model in another game. Fig. 10 shows the momentum change chart of player 1 in the selected match, and his performance fluctuates greatly over time. According to the total momentum reflected in Fig. 11, this player's performance has fallen into a low valley for a period of time. Combined with specific data, he did indeed lose points seriously at that stage.

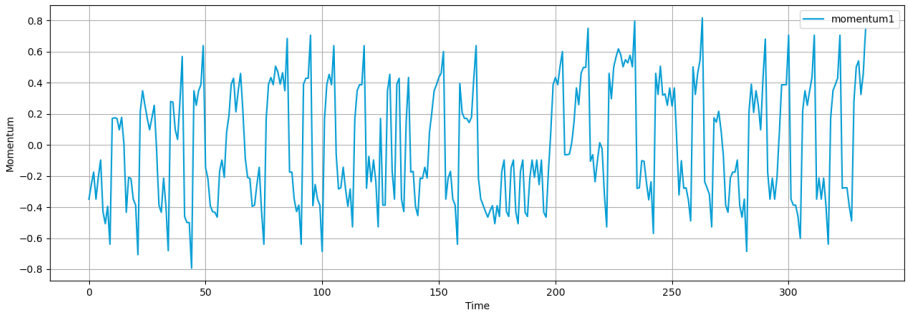


Fig. 10 Momentum-Time Figure

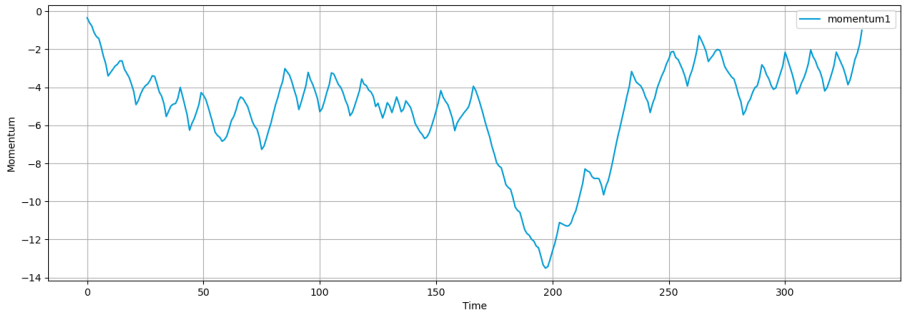


Fig. 11 Total Momentum-Time Figure

The model we have developed is based on machine learning, so quantifying the momentum of a tennis ball to predict trends in matches will be more accurate and relatively common in tennis, both for men and women, provided there is a sufficient training set and an effective method for identifying the important influencing metrics. However, the effect of time on scoring efficiency is an important factor in basketball and football games, with only a 0.5 to 0.6 probability in previous studies. Therefore, in order to be generalisable to other types of games, more factors need to be considered, such as time, scoring efficiency, and timeouts.

6. Summary

In this study, we provide an in-depth analysis and quantification of match momentum in the final of the 2023 Wimbledon Championships through machine learning techniques such as Random Forest Models and Support Vector Machines. By modelling and validating the role of momentum in the match, we draw the following conclusions:

Firstly, we successfully built a random forest regression model for predicting momentum and improved the accuracy of the model through data preprocessing and integrated learning methods. The model has high accuracy in predicting momentum in matches and provides an effective tool for understanding momentum changes in matches.

Second, we demonstrated that momentum fluctuations in matches are not random, and that momentum plays an important role in competitive matches. Through correlation tests and support vector machine algorithms, we verified the association between momentum and the outcome of the match, providing empirical support for the importance of momentum in the match.

Finally, we verified the validity of the model using the Kolmogorov-Smirnov test, which further demonstrated the relationship between momentum computation and match wins and losses. This study provides new methods and perspectives for momentum analysis in sports competition and

provides more in-depth tactical guidance and decision support for athletes, coaches, and policy makers.

In summary, this study not only expands the understanding of game momentum, but also provides a useful example of using machine learning techniques to analyse sports competition. Future research can further optimise the models and algorithms to explore the effects of more match data and metrics on momentum to improve the understanding and prediction of momentum changes in sports competition. This will help improve the performance of athletes and teams in competitions and advance the development and application of the field of sports data science.

References

[1] GAN Ruiping, REN Xinmin, JIANG Jun, et al. Daily energy consumption prediction of ship special coating repair based on random forest regression[J]. Big Data,2024,10(01):170-184.

[2] JIANG Gen,XIAO Shirong,YANG Xuanche,et al. Evaluation of landslide susceptibility in Badong County, Hubei Province based on machine learning[J/OL]. Water Resources and Hydropower Express,1-13[2024-03-30]. http://kns.cnki.net/kcms/detail/42.1142.TV.20240326.1153.004.html.

[3] XU Fei, ZHU Guangxi,WANG Kexin. Construction of a prediction model based on decision tree algorithm for the risk of developing radioactive oral mucositis after radiotherapy in nasopharyngeal carcinoma patients[J/OL]. Journal of Otorhinolaryngology and Ophthalmology of Shandong University,1-11[2024-03-30]. http://kns.cnki.net/kcms/detail/37.1437.R.20240327.1459.002.html.

[4] ZHAO Shengzhi,SHEN Yinhong,ZHOU Yinghuai. A study on the credibility of post-level qualification of power grid enterprises between zones based on Spearman's correlation coefficient[C]//China's power enterprise management innovation practice (2022). Yunnan Power Grid Limited Liability Company; 2024:4. DOI: 10.26914/c.cnkihy.2024.001241.

[5] F.Y. Zhang, L.L. Lv, L.C. Han, et al. Structured least squares twin support vector machine for intuitionistic fuzzy[J]. Journal of Applied Science,2024,42(02):350-363.

[6] Fang Xiangyu,Lei Yiju,Ou Zujun. Isomorphism discrimination of U-design combinations based on KS test[J]. Journal of Lanzhou College of Arts and Sciences (Natural Science Edition),2022,36(01):6-9+69.DOI: 10.13804/j.cnki.2095-6991.2022.01.005.