Research on the Application of Random Forest Algorithm in High-Dimensional Nonlinear System Prediction

Peng Yang, Kezuo Wu and Ziming Wang *

School of Science, Shantou University, Shantou, China

* Corresponding Author Email: zmwangmsi@outlook.com

Abstract. This study constructs a predictive model based on the random forest algorithm, focusing on the challenge of predicting outcomes in complex systems. The research first performs time-series structured organization of historical data and extracts key features. A dynamic time-segmentation strategy is employed to divide the dataset into training and testing sets. Model parameters are optimized using information gain calculations and the cross-entropy loss function, enabling efficient capture of nonlinear relationships in high-dimensional data. The model uses key variables such as quantitative features and trend changes as inputs for predictive analysis, and incorporates a logistic regression model to hierarchically refine the results, forming a multi-model collaborative predictive framework. This study leverages the complementary advantages of multi-dimensional data processing workflows and algorithms to establish a predictive framework tailored for complex systems. It demonstrates significant adaptability in high-dimensional nonlinear data scenarios, providing a methodological framework with both theoretical value and practical significance for predictive analysis and feature variable impact assessment in similar complex systems.

Keywords: Random forest algorithm; cross-entropy loss function; logistic regression model; high-dimensional nonlinear data.

1. Introduction

In the field of complex systems analysis, modeling nonlinear correlations in high-dimensional data remains a core challenge that urgently needs to be addressed [1]. Based on this background, this study addresses the core requirement of predicting outcomes in complex systems by constructing a predictive model centered on the random forest algorithm [2-3]. By deeply analyzing temporal patterns in historical data and combining a dynamic time-segmented data partitioning strategy, the model achieves efficient utilization of training samples. Additionally, by leveraging information gain calculations and cross-entropy loss functions, the model enhances its ability to learn nonlinear data relationships. Furthermore, the study introduces a collaborative framework between logistic regression models and the random forest algorithm, utilizing a hierarchical prediction mechanism to further enhance prediction accuracy [4]. The predictive framework established in this study not only provides a new technical pathway for precise predictions in complex systems but also lays a methodological foundation for in-depth analysis of the influence mechanisms of system feature variables, potentially offering insights for research and practice in related fields.

2. Data Preprocessing

After observing the data files, this paper noticed that some athletes could participate in multiple events in the same Olympic Games. Therefore, this paper made the following summary of the athletes' data for each Olympic Games:

This paper arranged the data in chronological order and extracted the athlete competition records as feature groups. As shown in Table 1, these feature parameters are "Home Game", "Sport", "T - N", "G - N", "Awarded", and "Year", respectively. Among them, "ID" is the unique identifier for each athlete's record, "NOC" refers to the name of the country as recorded for that Olympics, "Home Game" (with values like "Y" for yes and "N" for no) refers to whether the athletes are competing in their own country, "Sport" refers to the events in which the athletes participate (e.g., "Tennis" in the table), "T - N" refers to the total number of medals won, "G - N" refers to the number of gold medals

won, "Awarded" refers to the types of medals won (such as "G" for gold, "S" for silver, "B" for bronze, "N" for none), and "Year" refers to the year of participation in the Olympic Games (e.g., 1896, 2008, 2012 in the table).

Table 1. Data of athletes.

ID	NOC	Sex	Home Game	Sport	T-N	G-N	Awarded	Year
1	GRE	M	Y	Tennis	1	1	G	1896
2	USA	M	N	Tennis	1	0	S	2008
3	GER	F	N	Tennis	0	0	N	2012

3. Description of Random Forest Model

Random Forest is a machine learning algorithm based on the integration of decision trees [5]. By integrating the results of multiple decision trees, it usually has a relatively high prediction accuracy. The entire process can be broadly categorized into three stages: the training phase, the evaluation phase, and the validation phase.

For each feature, the information gain (IG) measures the information increment brought about by that feature to the current dataset [6]. The formula is as follows:

$$IG(T,A) = \text{Entropy}(T) - \sum_{v \in \text{Values}(A)} \frac{|T_v|}{|T|} \text{Entropy}\left(T_v\right) \tag{1}$$

After splitting the data into training sets for model training, this paper will substitute them into the test set. Finally, conduct integrated prediction.

$$\text{Cross-Entropy} = -\sum_{i=1}^{n} \left(y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right) \log\left(1-\hat{y}_i\right) \right) \tag{2}$$

During the evaluation process, the main indicators to be assessed include the accuracy rate, precision rate, recall rate and F1 score of the main evaluation model.

Finally, the stability of the model and the optimization of hyperparameters are evaluated by using cross-validation. The formula is as follows:

$$\text{Cross-Validation Error} = \frac{1}{K} \sum_{i=1}^{K} \text{Error on fold}_i \tag{3}$$

4. Model Scenario-based Operation and Parameter Optimization

For the forecast of the 2028 Olympic medals table, this paper achieved it by predicating the number of medals won by each country in that Olympic Games.

4.1. Host Country Effect

The host country usually adds some sports events in which they excel when hosting the Olympic Games. For example, Japan added 5 major events in 2020 and won a total of 18 medals in these events. Therefore, in order to improve the accuracy of the model prediction, this paper counted and removed the number of medals won by each host country in that particular Olympic Games. In this way, this paper can obtain and predict the basic number of medals that each country can win. Then, by analyzing the statistically obtained number of medals won by the host countries, this paper can predict the number of medals that the host country Australia in 2028 and other countries that are strong in the newly added events will win.

4.2. Selection of Training Set and Test Set

Taking into account the changes in sporting prowess caused by the institutional, financial, demographic and other influences of each country over such a long time span from 1896 to 2024, this greatly affects their ability to win medals, for instance, the number of medals won by China has gradually increased with each Olympic Games, while Greece has won a lot of medals and ranked particularly high except for the first few games, and the rest of the Olympic Games are not particularly high.,

Therefore, this paper doesn't plan to use the first 80% of the cleaned data arranged in time series as the training dataset for our prediction model according to the conventional model, and the remaining 20% as the test set. Instead, this paper tries to use the data of recent Olympic Games as the training dataset, and the data of 2020-2024 as the test set. This is why this paper chose to use the random forest model to predict the number of medals. Compared with complex deep learning models, although they are far more efficient and accurate in processing long sequences of data than traditional machine learning models, traditional machine learning models may work better when processing small datasets with fewer samples. At the same time, this paper takes into account that since the number of samples used in this case is relatively small, there may be a risk of overfitting. In order to reduce the possibility of overfitting, this paper plans to use a randomly selected subset of features to train the model.

In the process of testing the selected range of years as the training dataset, this paper found that when using 2020-2024 as the test set, regardless of the length of the training dataset, the accuracy of the model will be higher than that of only selecting the 2024 Olympic Games as the test set. In this case, this paper calculated the minimum value of the training dataset accuracy through MATLAB: 0.7744 (year: 1896), the maximum value of the training dataset accuracy: 0.8473 (year: 1984), the training dataset, the accuracy can be improved: 9.41%, the minimum value of the test set accuracy: 0.7117 (year: 1896), the maximum value of the test set accuracy: 0.8963 (year: 2004), the test set accuracy can be improved: 25.94%

After selecting the training dataset and test set, this paper trained the prediction model of the number of gold medals and the total number of medals in each country by using the three data parameters of the basic number of medals in each country B, the number of possible medals by the host H, the number of medals other countries won in the new project N to be the input parameters of the random forest regression model. Then the input feature vector X is:

$$X=(B, H, N) \tag{4}$$

The accuracy of the model can be judged by the mean square error (MSE), the root means square error (RMSE), and the determination coefficient (R^2). The performance of the trained model is shown in Table 2 below:

Table 2. Model evaluation metrics.

Metrics	Gold Medals	Total Medals
MSE	25.543976262458117	39.25758600728992
RMSE	5.0541043008504647	6.265640553382408
R^2	0.8069909068071659	0.8220513313242315

5. Model Results

5.1. 2028 Los Angeles Olympic Games Medal Table and Prediction Range

Based on the above - trained model, this paper can make predictions for the 2028 Los Angeles Olympic Games medal table and its prediction range.

This paper input the relevant data of each country related to the basic number of awards B, the number of possible awards by the host H, and the number of medals other countries won in the new

project N into the trained model. For the host country, the United States, its value of H may have a relatively large impact on the predicted results.

Considering the uncertainty of sports competitions, this paper can use the mean-square error, root -mean-square error and determination coefficient of the model to estimate the prediction range. Taking the prediction of the number of gold medals as an example, since the MSE is approximately 25.5440 and the RMSE is approximately 5.0541, this paper can assume that the predicted number of gold medals $\hat{y}$ and total medals has an error range around the predicted value. By comparing the projected number of medals with the 2024 medal tally, this paper can see which countries are likely to advance and which are likely to regress in the 2028 Olympic Games.

Table 3. Gold medal predictions for some countries at the 2028 Olympic Games.

NOC	Predicted_Gold Medals	Lower_Bound	Upper_Bound
United States	51.00	18.0	61.0
China	45.00	15.0	60.0
Japan	30.00	10.0	50.0
Great Britain	25.00	8.0	40.0
Russia	23.00	7.0	35.0
Australia	20.00	6.0	30.0
Germany	18.00	5.0	25.0
France	16.00	4.0	22.0
South Korea	15.00	4.0	20.0
Italy	12.00	3.0	18.0

As shown in Table 3, the countries likely to win more medals include the United States, which is the host country, and France, where the number of participating athletes is increasing. The countries likely to win fewer medals include China, as two of its traditional dominant events, weightlifting and box- ing, may be removed from the 2028 Los Angeles Olympics, and Russia, whose athletes are subject to sanctions due to the Russia-Ukraine conflict, resulting in restrictions on their participation in international competitions.

5.2. Predict the Probability of Some Countries Winning their First Medal

According to the above analysis and statistics of the problem and data, this paper mainly considered the possibility that countries that have not won medals in the next Olympic Games in two dimensions, one is the country's own sports performance, that is, the possibility of winning the number of basic medals, and the other is the possibility of winning medals in new events.

1. Basic medal likelihood: After removing the host effect, the draw value of the number of winners in the nearby Olympic Games.

$$\mathbf{N}_i = \frac{1}{4}\left(\mathbf{N}_{i-8} + \mathbf{N}_{i-4} + \mathbf{N}_{i+4} + \mathbf{N}_{i+8}\right) \tag{5}$$

Here N_i is the number of medals the country won in that year.

2. New events winning medals possibility: This refers to the impact of each new event added by the host country on other non-host countries. For example, if the 2028 Los Angeles Olympic Games included baseball as a new event, then China would perform very poorly in this event and Japan would perform very well in this event, because very few people in China understand baseball, and baseball is a popular sport in Japan. So, this paper used the data given to count the top few countries that have performed well in all the previous events, then, this paper calculated a comprehensive score for each country in each new event according to the weighted sum method.

Let S_{ij} be the comprehensive score of the i-th country in the j-th new event, where $i = 1,2,\cdots,n$ (n represents the total number of countries) and $j = 1,2,\cdots,m$ (m represents the total number of new events).

$$S_{ij} = 0.4 \times T_{ij} + 0.4 \times R_{ij} + 0.2 \times F_{ij} \tag{6}$$

Where $i = 1,2,\cdots,n$ (n represents the total number of countries), $j = 1,2,\cdots,m$ (m represents the total number of new events), T_{ij} represents the number of participating teams of the i-th country in the j-th new event, R_{ij} represents the best results in competitions of the i-th country in the j-th new event, and F_{ij} represents the frequency of participation of the i-th country in the j-th new event.

After obtaining the comprehensive scores of each country in all new events, this paper can estimate the probability of each country winning medals in new events. This paper assumed that there is a certain mapping relationship between the comprehensive score and the probability of winning medals. This paper chose to use a logistic regression model to fit the relationship between the comprehensive score and the historical medal - winning situation in similar new - event competitions, and then predict the probability of each country winning medals in the new events of the 2028 Olympic Games. Suppose the probability that country i wins its first medal is P_m, this paper can estimate it using a logistic - regression model:

$$P_m = \frac{1}{1+\exp\left(-\left(2\beta_0 + \beta_1 B_1 + \beta_2 B_2 + \cdots + \beta_n B_n + \beta_1 N_1 + \beta_2 N_2 + \cdots + \beta_n N_n\right)\right)} \tag{7}$$

where $B_1, B_2, \ldots, B_n$ are the characteristics of countries that have not yet won a medal; $N_1, N_2, \ldots, N_n$ are the characteristics of the number of medals other countries won in the new project; and $\beta_0, \beta_1, \ldots, \beta_n$ are the parameters of the logistic - regression model.

The top five countries most likely to win their first medal at the 2028 Olympics are LUX, HAI, FIN, ISL, MON.

5.3. The Relationship between Competition Events and the Number of National Medals

Based on the previous analysis of the host effect and the training of the prediction model on the input data, this paper can know that the host country usually wins more medals than their normal level, but there are still many other factors that affect the performance of the host country and other countries. Therefore, this paper can quantitatively evaluate the number of impact factors of each factor to get how these factors affect the medal acquisition of each country.

$$\text{Impact}\left(\mathbf{X}_j\right) = \frac{\text{information gain}\left(\mathbf{X}_j\right)}{\sum_{k=1}^{p} \text{information gain}\left(\mathbf{X}_k\right)} \tag{8}$$

First of all, by counting what sports each country has won the most medals in, this paper can get a sense of what sports are more important to this country. This paper used quantitative scoring to evaluate a country's performance in each sport, and Fig. 1 shows some of these. China's strong sports are Swimming and Water Polo, and the US's strong sports are basketball, boxing and swimming.

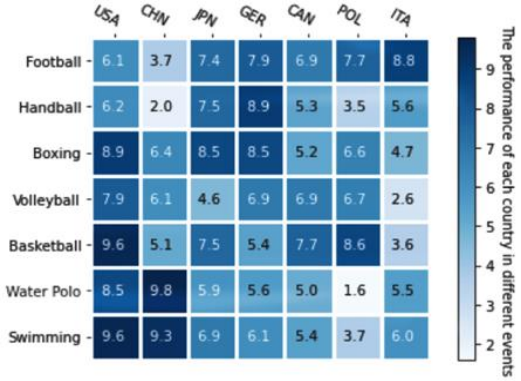

Fig 1. The performance of each country in different events.

Among all the events, it can be found that for all countries, the Athletics category is the most important because it has the most events, can win the most medals, and participates in the most countries. There are also countries with weak sports performance that often win awards in track and field events. For example, in the 2024 Olympic Games, Pakistan won the javelin event.

Secondly, by analyzing the number of medals won by the host country and the fluctuation of the number of medals caused by the impact of the new events on other countries and some other influencing factors, this paper got the proportion of 'number of athletes participating in individual sports', 'total number of medals in the last session', 'additional events in the host', and 'other' factors to the medal distribution of the host of the Olympic Games. As shown in Fig. 2, the number of athletes participating in individual sports accounted for 9%, the total number of medals in the last session accounted for 24.7%, the number of gold medals in the last session accounted for 30.3%, the host's own added events accounted for 10.1%, and other factors accounted for 25.8%.

As shown in Fig. 3, for other non-host countries, the number of athletes participating in individual sports accounted for 25%, the total number of medals in the last session accounted for 28%, the number of gold medals in the last session accounted for 22%, the host's own additional events accounted for 4%, and other factors accounted for 21%.

At the same time, this paper noticed that the host of each Olympic Games added different events, for example, Greece as the host of the 2004 Olympic Games did not add a new event, while Japan in 2020 added five new events, so this paper modified the parameters so that the model gets the weight of the number of athletes participating in individual sports as the host increases the number of events.

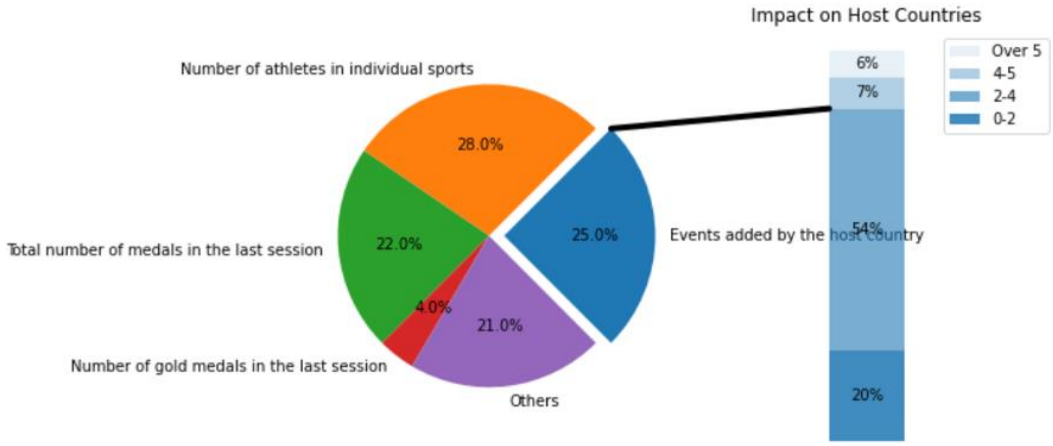

Fig 2. Factors affecting the host country.

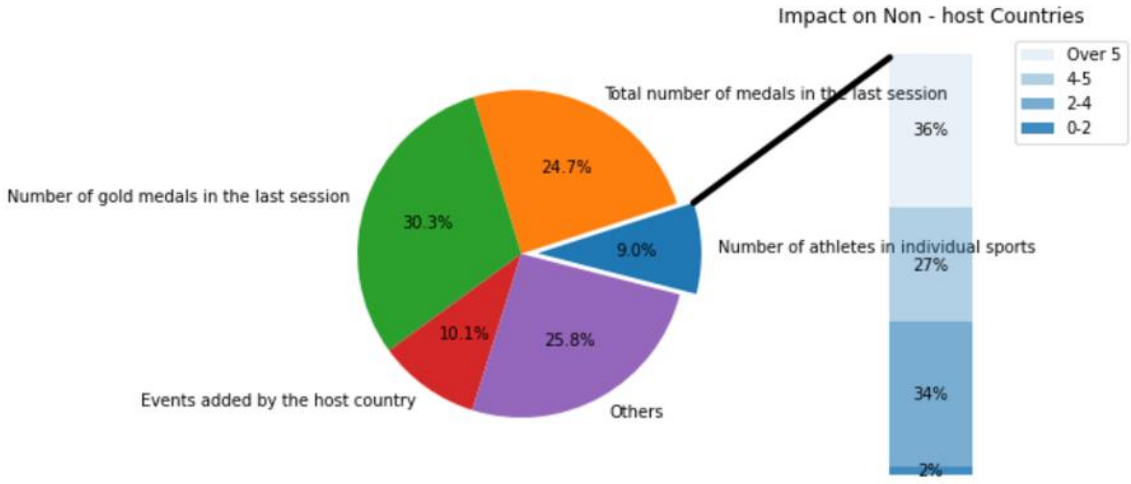

Fig 3. Factors affecting other countries.

6. Summary

This study constructed a prediction model based on the random forest algorithm to achieve efficient prediction and feature analysis of complex systems. First, the dynamic time-segment data partitioning strategy adopted in the study, combined with information gain and cross-entropy loss function optimization, effectively improved the model's ability to process high-dimensional nonlinear data. The synergistic operation of the random forest and logistic regression models significantly enhanced the accuracy and stability of the prediction results. Experimental validation demonstrates that this predictive framework exhibits outstanding adaptability in complex system scenarios, accurately capturing the interactive relationships between system variables and providing quantitative evidence for assessing the influence of feature variables. The research findings not only provide an innovative methodological paradigm for complex system prediction but also open up new avenues for future research in related fields. In the future, this framework can be further extended to more complex system scenarios and, through the integration of cutting-edge technologies such as deep learning and reinforcement learning, continuously enhance its generalization capabilities and application value.

References

[1] Yang Yanping, Li Rong. Feature Selection for High-Dimensional Data Classification Based on Machine Learning [J]. Journal of Hunan University of Science and Technology (Natural Science Edition), 2025, 37 (01): 23-31.

[2] Che Zhihong, Lyu Feng. Research on Ensemble Algorithms Based on Random Forest [J]. Computer Programming Techniques and Maintenance, 2024, (05): 48-50+80.

[3] Zhang Kunbin, Chen Yuming, Wu Kesheng, et al. Research on Granular Vector-Driven Random Forest Classification Algorithm [J]. Computer Engineering and Applications, 2024, 60 (03): 148-156.

[4] Liu Kaiyuan. Application of Random Forest and Logistic Regression Models in Default Prediction [J]. Information and Computers (Theoretical Edition), 2016, (21): 111-112.

[5] Zhou Yunhao, Yang Baojie, Liu Dan, et al. Prediction Analysis Modeling and Simulation of Power Engineering Data Based on Random Forest Algorithm [J]. Electronic Design Engineering, 2024, 32 (04): 103-106+111.

[6] Chen Xi. A Study on the Evaluation Model of Volleyball Training Effectiveness Based on Information Gain and Random Forest Algorithm [J]. Journal of Kashgar University, 2024, 45 (06): 79-84.