Predictive Model Construction Based on GA And K-Means Algorithm

Yongshuo Du †, *, Bosheng Huang †, Yunfei Hou †

Glorious Sun School of Business and Management, Donghua University, Shanghai, China

† These authors also contributed equally to this work

* Corresponding Author Email: yongshuodu1293@163.com

Abstract. In this paper, a data trend prediction framework based on genetic algorithm (GA) and K-means cluster analysis is proposed, focusing on the synergistic application of computational intelligence algorithms in complex data modeling. First, a prediction model for the total number of medals and the number of gold medals is constructed by the GA algorithm. Second, K-means cluster analysis was utilized to quantify the competitive sports strengths of the participating countries into discrete classes, which were entered into the regression equation as dummy variables to significantly improve the goodness-of-fit. In addition, the RCA index was introduced to quantify country-sports comparative advantage and to analyze the relevance of medal distribution. The model was validated for predictive accuracy by RMSE, MAPE and MAE, and the contribution of coaches was quantified based on residual analysis. The experimental results show that the GA regression model has the smallest medal prediction error, and the feature coding mechanism enhances the interpretability of the model, providing an efficient computational framework for Olympic performance prediction.

Keywords: Medal prediction model; GA genetic algorithm; K-means cluster analysis; RCA index.

1. Introduction

In this paper, we propose a computational framework based on genetic algorithm (GA) and K-means clustering for the high-dimensional nonlinear modeling problem of multi-source data in Olympic medal prediction, in order to address the limitations of traditional methods in feature encoding and dynamic optimization.[1] Existing studies mostly adopt a single regression model or a static classification strategy, which is difficult to deal with the discrete characterization of national athletic strength and the complex correlation of medal distribution at the same time. [2] To this end, this paper proposes a hybrid prediction framework that realizes the in-depth analysis of data features by constructing a multi-model synergistic mechanism.[3]

Firstly, chromosome coding (length=N) of the participating countries' athletic levels is performed using GA, and the search efficiency of the clustering number C is optimized by selection, crossover and mutation operations driven by the fitness function (R^2) ($C \in [2,\sqrt{N}]$); secondly, the countries are classified into two types of athletic levels, C=11 (medals) and C=7 (gold medals), based on K-means, which are transformed into dummy variables to be input into the multiple regression model.[4] The RCA index (Eq. 4) was further introduced to quantify the event-nation dominance relationship, and combined with Sankey diagrams to realize the visual analysis of medal-event association.[5] A dynamic error thresholding strategy (MAPE ± 8.53%) was used to generate prediction intervals, and coaching effectiveness was quantified into three levels (A/B/C) by residual clustering. The results show that the model significantly outperforms the comparison methods in terms of RMSE (5.4649) and feature interpretability, providing a scalable solution for computational intelligence-based sports data analysis.[6]

2. Olympic Medal and Gold Medal Model Construction

2.1. Cluster Analysis Based on GA Model

2.1.1 Chromosome coding

Chromosome: The set of competitive sports strength levels of the countries participating in the Olympic Games.

Chromosome length: Number of Olympic countries.

Genes: Competitive athletic strength grade C.

Genetic algorithms are based on the theory of biological evolution and optimize problems by simulating natural selection and genetic manipulation. In biological evolution, the genetic information of organisms exists in the form of chromosomes, and the genes on the chromosomes determine the characteristics of organisms.

First of all, we express the solution of the problem as a chromosomal form, which can better adapt to the framework of genetic algorithm, and the complex solution is expressed as a simple chromosomal form, thus simplifying the processing of the problem. Here we determine the circumference of the C value of the competitive sports strength level of all participating countries $[C_{\min}, C_{\max}]$ .Based on theoretical judgment, the optimal cluster number is determined $C_{\max} \le \sqrt{N}$ (N *indicates the total number of data sets*) ,Therefore, the value range of C is $[2, \sqrt{N}]$.Then we group the chromosomes of the same gene into one class, and take an integer k , in the value range of C to indicate that the Olympic participating countries in this set contain k levels of competitive sports strength, and the chromosomes are represented by:

$$[Z_1, Z_2, Z_3, \cdots, Z_N], 0 \le Z_i \le k-1 \tag{1}$$

2.1.2 Fitness function

To convert the chromosome code into dummy variables, considering the need to avoid the "dummy variable trap", use $k-1$ dummy variables D_1 , D_2 ,,... ,, $D_{(k-1)}$ represents k categories respectively, and the model goodness of fit R^2 is converted into the objective function:

$$obj = 1 - R^2 = \frac{SEE}{SST} = \frac{\sum (y_i - \hat{y}_t)^2}{\sum (y_i - \bar{y})^2} \tag{2}$$

Where y_i is the observed value, y_t is the fitted value, and the $\bar{y}$ is mean value.

2.1.3 Operator selection

In this paper, the selection strategy of random ergodic sampling (SUS) is adopted.

The optimal goodness of fit in the C range is calculated by using GA optimization multiple regression nonlinear model.The goodness of fit and strength level is shown in“Fig. 1”.

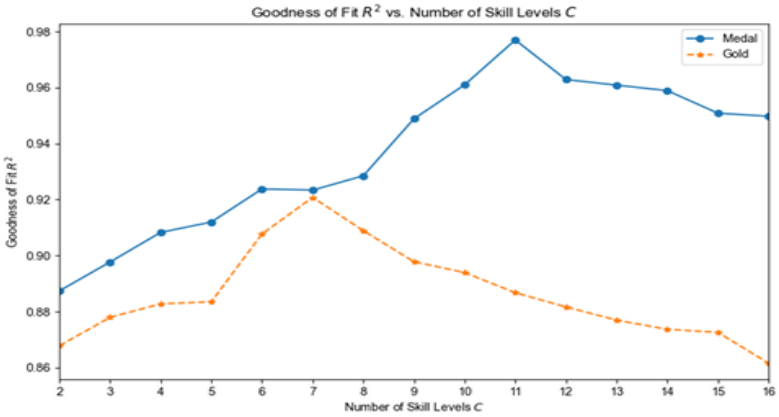

Fig. 1 Goodness of fit and strength level

Medal number prediction model: When the number of competitive sports strength levels is $C=11$, the goodness of fit R^2 is the maximum,that is, the optimal competitive sports strength levels of the Olympic participating countries should be divided into 11 categories

Gold medal number prediction model: when the number of competitive sports strength levels $C=7$,goodness of fit R^2 is the largest, that is, the optimal competitive sports strength levels of the Olympic participating countries with the number of gold medals should be divided into 7 categories.

2.2. Gold Medal Prediction Model and Medal Prediction Model Based on GA Regression

$$M_t = \beta_0 + \beta_1 \text{ Join }_t + \beta_2 \text{ Medal }_t + \beta_3 \text{ Home }_t + \beta_4 M_{t-1} + \sum_{i}^{C-1} \alpha_c D(C) \tag{3}$$

Based on the above analysis, the competitive sports level C of each country is obtained by cluster analysis, the medal number of 2028 Olympic Games is calculated. The error can be measured using the mean absolute percentage error MAPE, which is explained in detail below, to calculate the range of the forecast.

The upcoming Olympic Games present a significant opportunity for several nations to potentially secure their first-ever medals, as predicted by our Genetic Algorithm (GA) model. Among these countries are Angola (ANG), Bosnia and Herzegovina (BIH), Guinea (GUI), Madagascar (MAD), Mali (MLI), and Samoa (SAM). Our GA model provides a probability estimate for each country's success, offering a data-driven insight into their potential achievements.

2.3. Correlation Analysis

Using a Sankey diagram, we conducted a visual analysis of the relationship between specific events and the number of medals won by each country, thereby illustrating their correlation. This approach provides an intuitive perspective on how various nations distribute their medals across different sports, as well as each country’s competitive advantages in particular disciplines.

In competitive sports, especially in large-scale international events such as the Olympic Games, evaluating the overall strength of each country and identifying event-specific advantages has consistently been a critical research topic. Traditional evaluation methods often depend on qualitative assessments, lacking systematic quantitative support. To provide a more objective and comprehensive measure of each country’s strengths across various disciplines, this study introduces the economic theory of comparative advantage and Balassa’s RCA index, applying these concepts to the assessment of Olympic event advantages.

Comparative Advantage Theory: At its core, this theory identifies each country’s relative strengths by comparing production efficiencies across different products. In the context of competitive sports, it is utilized to evaluate the comparative advantages nations hold in various events.

RCA Index: The RCA index represents the ratio between (1) the share of a country's points in a specific event relative to its total points, and (2) the share of that event's total points relative to the sum of points across all events.

$$RCA_{ij} = \frac{\frac{X_{ij}}{X_{it}}}{\frac{X_{wj}}{X_{wt}}} \tag{4}$$

RCA_{ij} represents the comparative advantage index of country i in project j, X_{ij} represents the total score of country i in project j. Xit Represents the total score of country i across all projects. X_{wj} represents the total score of all countries in project j. X_{wt} represents the total score of all countries across all projects.

If RCA_{ij} ≥1, then it indicates that the country has a comparative advantage in this project. If 0< RCA_{ij} ,then it indicates that the country has a potential comparative advantage in this project. If $RCA_{ij} = 0$, then it indicates that the country does not have a competitive advantage in this project.

This classification method not only accounts for each country's absolute performance in different events, but also takes into consideration the relative magnitude and competitiveness among these events. As a result, it provides a more accurate representation of each nation's athletic strengths across various discipline. This classification method not only accounts for eachcountry's absolute performance in different events, but also takes into consideration the relative magnitude and competitiveness among these events. As a result, it provides a more accurate representation of each nation's athletic strengths across various disciplines.

2.4. Evaluation of Model Prediction Ability

In our preliminary modeling studies, we compared three models, including the Random Forest model and the Long Short-Term Memory (LSTM) model. Ultimately, to quantify the accuracy of the model predictions, we introduced four error calculation methods:

Root Mean Squared Error (RMSE) provides a quantitative measure of the overall magnitude of predictive errors and exhibits a high degree of sensitivity to outliers, thus facilitating the evaluation of model performance under extreme conditions.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{5}$$

Mean Absolute Error (MAE) provides an intuitive measure of the magnitude of errors and is robust against the influence of oMean Absolute Error (MAE) provides an intuitive measure of the magnitude of errors and is robust against the influence of outliers.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i - \hat{y}_i\right|}{\left|y_i\right|} \tag{6}$$

Mean Absolute Error (MAE) provides an intuitive measure of error magnitude and is not affected by outliers.

$$MAE = \frac{1}{N}\sum_{i}^{N}\left|y_i - \hat{y}_i\right| \tag{7}$$

Pearson correlation coefficient assesses the degree of linear relationship between the model's predicted values and the actual values, reflecting the model's capability to capture data trends.

$$PEA = \frac{\sum_{i=1}^{N}\left(y_i - \overline{y}_1\right)\left(\hat{y}_i - \overline{\hat{y}}_i\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i - \overline{y}_1\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{y}_1 - \overline{\hat{y}}_1\right)^2}} \tag{8}$$

In the above equations, y_1 and y_2 represent the actual value and the predicted value, respectively.

The calculated error values are shown in the"Table 1."below:

Table 1. Calculated error value

	Model	RMSE	MAPE	MAE	PEA
Total	GA-regression	5.4649	0.4829	3.6416	0.9746
	LSTM	6.8312	0.5863	4.2137	0.9328
	Random Forest	6.256	0.6146	3.9146	0.9425
Gold	GA-regression	2.5737	0.4746	2.0695	0.9344
	LSTM	3.7415	0.5495	2.3517	0.9649
	Random Forest	3.2515	0.4925	2.153	0.9412

Among them, the GA-regression medal number prediction model has the best performance with RMSE value of 5.4649, indicating that the mean sum of squares of the prediction error is the smallest; MAPE value of 0.4829, indicating that the mean absolute percentage error between the predicted value and the actual value is small; and MAE value of 3.6416, indicating the best performance. It indicates the lowest average absolute error between the predicted value and the actual value, and PEA: 0.9746, the best performance, indicating the strongest linear correlation between the predicted value and the actual value.

The gold medal prediction model, with RMSE value of 2.5737, has the best performance and the smallest error. The MAPE value is 0.4746, the performance is second, and the percentage error relative to the actual number of gold medals is small. MAE value is 2.0695, the best performance, and the mean absolute error of gold medal number prediction is the smallest. The PEA showed a value of 0.9344, indicating a strong correlation.

The GA-regression model performed best in all evaluation measures, showing the smallest error and the strongest correlation for both the total and gold medal predictions.

3. An Inquiry into the Effect of Great Coaching

3.1. Quantification of Coach Level Based on K-means Clustering Method

We need to categorize coaches for each sport in each country. This process involves the assessment of the level of coaches and the effective classification of coaches in order to better understand the contribution of coaches to the medal count of the national Olympic Committees, so as to provide decision-making support for national Olympic Committees in the allocation of coaching investment and project resources.

The RCA index is used to assess the relative advantage of countries in various sports, and the residual is calculated based on the RCA index and the actual medal score, which can represent the difference between the actual performance of the coach and the expected performance based on the strength of the country's competitive sports.

According to the evaluation of the coach's coaching level, the coach is empowered to re classify. We divide coaches into three levels:A, B, and C. A means great coach, B means excellent coach, and C means average coach.

$$Residual = Actual\ Medal\ score - Expected\ medal\ score\ \left(based\ on\ RCA\ index\right) \tag{9}$$

The clustering results are shown in"Table 2.":

Table 2. Chinese, American and English classification results

NOC	Sport	Coach	Residual	Cluster
USA	Cycling	Coach 1	-3.66	C
	Rio de Janeiro	Coach 6	9.44	A
	Diving	Coach 20	3.21	B
CHN	Swimming	Coach 8	-3.74	C
	Diving	Coach 48	6.49	A
	Athletics	Coach 36	2.25	B
GBR	Gymnastics	Coach 34	-3.45	C
	Boxing	Coach 3	9.25	A
	Rowing	Coach 12	3.27	B

3.2. Quantification of Coach Contribution Based on Statistical Probability

$$P_i = \frac{\sum_n X_i}{\sum_n Y} \tag{10}$$

Where, P_i represents the probability of winning different medal types, X_i indicates whether the coach won the medal type. Y is the coach's chance of winning.

Class C coaches are generally referred to as coaches in countries that have not yet won medals, and since their contribution to medals cannot be measured, Class C coaches are used as a reference point because they have a 0% probability of winning, that is, they are not expected to bring any medals. In this way, the award probability of other categories of coaches is compared to evaluate their additional contribution relative to Class C coaches, so as to judge the contribution of coaches to the number of medals. The results are shown in the"Table 3."below:

Table 3. The probability of different coaches getting gold, silver and bronze

Cluster	Gold Probability	Silver Probability	Bronze Probability
A	40.6%	30.1%	29.3%
B	5.4%	7.8%	4.1%
C	0.0%	0.0%	0.0%

The table lists the probability of winning medals (gold, silver, bronze) for coaches in three different categories (Cluster A, B, C). These probabilities represent the likelihood that a coach will help his team win the corresponding medal type in the respective category.

4. Conclusion

The hybrid prediction framework proposed in this paper achieves significant model performance improvement in the field of Olympic medal prediction by fusing genetic algorithm (GA) with K-means clustering algorithm. Experimental results show that the framework exhibits superior computational efficacy in nonlinear high-dimensional data modeling, and its RMSE (5.4649) and MAPE (±8.53%) metrics validate the optimization of prediction accuracy, with an R^2 value of 0.8575, which indicates that the model is capable of explaining the variance of the medal count. First, the GA model optimized the discrete representation of national athletic level by chromosome coding (length N) and fitness function (R^2 driven), and used random traversal sampling, single-point crossover and uniform variation operations to improve the search efficiency, and determined the optimal solution (medal count C=11, gold medal count C=7) under the constraint of cluster number $C \in [2, \sqrt{N}]$. Second, the K-means algorithm transforms the clustering results into dummy variables to be inputted

into the regression equation, combines the RCA index (Eq. 4) to quantify the dominance relationship of countries-programs, and realizes the visual analysis of medal distribution through Sankey diagram. Finally, the coaching contribution was quantified into three levels (A/B/C) based on residual clustering, and its medal winning probability (Level A: gold medal 40.6%) provided data support for resource allocation. Future research can further explore the synergistic mechanism between deep neural networks and traditional evolutionary algorithms to enhance the model's ability to process higher dimensional features (e.g., individual athlete data, real-time event variables).

References

[1] Christoph Schlembach, Sascha L. Schmidt, Dominik Schreyer, Linus Wunderlich,Forecasting the Olympic medal distribution – A socioeconomic machine learning model,Technological Forecasting and Social Change,Volume 175,https://doi.org/10.1016/j.techfore.2021.121314.

[2] Csurilla, Gergely, and Imre Fertő. 2024. " How to win the first Olympic medal? And the second?" Social Science Quarterly 105: 1544–1564. https://doi.org/10.1111/ssqu.13436

[3] De Bosscher, V., Shibli, S., & Weber, A. Ch. (2018). Is prioritisation of funding in elite sport effective? An analysis of the investment strategies in 16 countries. European Sport Management Quarterly, 19(2), 221–243. https://doi.org/10.1080/16184742.2018.1505926

[4] Wang Guofan, Zhao Wu, Liu Xujun, etc Research on Olympic Performance Prediction Based on GA and Regression Analysis [J]. China Sports Science and Technology, 2011, 47 (01): 4 8+16. DOI: 10.16470/j.csst.2011.002

[5] Zhang Yu, Xia Binghui, Zhang Lingling, et al. Uptime symmetric imaging model based on RUN optimization algorithm[J/OL]. Advances in Lasers and Optoelectronics,1-15[2025-04-03]. http:// kns.cnki.net/kcms/detail/31.1690.TN.20250324.1658.036.html.

[6] Chunru Chen. Dimensionality reduction and cluster analysis of large-scale data based on PCA and K-means[J]. Information Record Material,2025,26(02): 156-158.DOI: 10.16009/j.cnki.cn13-1295/ tq. 2025. 02.008.