Research on Prediction Methods Based on the KAN-LSTM Hybrid Model

Kaitong Sun *, Haoyuan Liu and Xiaoyu Pi

School of science, North China University of Science and Technology, Tangshan, China

* Corresponding Author Email: 3515475846@qq.com

Abstract. This study investigates a prediction method based on the KAN-LSTM hybrid model, aiming to achieve classification, prediction, and impact assessment of complex systems. Firstly, this study uses the K-means clustering algorithm to classify multi-feature datasets, constructs a feature matrix by selecting numerical indicators, and determines the optimal number of clusters using the Calinski-Harabasz index to systematically reveal the differentiated performance characteristics of different sample groups in time series. Secondly, to address the need for non-linear trend prediction, the KAN-LSTM hybrid model is constructed, integrating binary features, numerical features, and time-series data. The LSTM network captures the dynamic trends of the sequence, while the KAN algorithm calculates sample similarity and performs weighted averaging. Additionally, the Ada-Boost ensemble strategy is employed to optimise the prediction output. Finally, the difference-in-differences (DID) model and regression discontinuity design (RDD) model are employed to conduct causal inference analysis on the impact effects of intervention events. Through the integration of multiple methods, the study forms a complete analytical chain of 'data classification - trend prediction - impact assessment,' providing methodological references for modelling complex systems.

Keywords: K-means clustering algorithm; Calinski-Harabasz index; KAN-LSTM hybrid model; difference-in-differences (DID) model; regression discontinuity design (RDD) model.

1. Introduction

In the field of data analysis and modelling for complex systems, effectively integrating multi-dimensional features, capturing dynamic patterns, and revealing causal relationships are core challenges for both academia and engineering. To address this issue, this study investigates a prediction method based on a KAN-LSTM hybrid model. First, the study introduces the K-means clustering algorithm to perform unsupervised classification of multi-feature data, optimising clustering performance using the Calinski-Harabasz index to address the issue of analysing differentiated features across sample groups. Secondly, to address the need for non-linear trend prediction, the KAN-LSTM hybrid model is constructed, combining the temporal modelling capabilities of Long Short-Term Memory (LSTM) networks with the similarity calculation advantages of K-Nearest Neighbor (KAN) neural networks. Enhancing prediction robustness through the AdaBoost ensemble strategy. Finally, the study employs difference-in-differences (DID) and regression discontinuity design (RDD) models to quantify the causal effects of intervention events on target variables, addressing the limitations of traditional correlation analysis in causal inference.

The study follows a technical approach of 'feature clustering - hybrid prediction - causal inference' to construct a multi-method collaborative analysis framework, aiming to provide multi-method collaborative solutions for complex system analysis, enhance the interpretability and predictive accuracy of data-driven modelling, and provide methodological references for research in similar data scenarios.

2. National Classification: K-means Clustering and Evaluation

The K-means clustering algorithm is a common unsupervised learning method used to di-vide a dataset into different clusters. This paper uses the K-means algorithm to classify the countries participating in the Olympics.

2.1. Feature Selection and Data Preprocessing

Extracted from the data related to the medal and the number of entries in characteristics, such as various countries won gold medals, silver, bronze medals number, total number of medals, the number and performance stability and volatility.

Performance stability: Performance stability can be measured by the standard deviation of the number of medals won.

$$\text{Stability} = \sigma_{\text{Gold}} \tag{1}$$

Ranking volatility: this paper calculated each year (or) in each Olympic Games medal changes in amplitude, and through the standard deviation to measure volatility. Assume that each country in each Olympic Games medal is the ranking volatility can be calculated as follows:

$$\text{Ranking Fluctuation} = \sigma\left(\Delta M_t\right) \tag{2}$$

Where $\Delta M_t = \left|M_{t2} - M_{t1}\right|, \left|M_{t3} - M_{t2}\right|$ represents the change in the number of medals won.

2.2. K-means Clustering

For samples and cluster centers, the Euclidean distance is calculated as follows:

$$d\left(x_i, \mu_k\right) = \sqrt{\sum_{j=1}^{n}\left(x_{ij} - \mu_{kj}\right)^2} \tag{3}$$

Where $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{in}\right)$ $\mu_k = \left(\mu_{k1}, \mu_{k2}, \ldots, \mu_{kn}\right)$.

Feature matrix:

$$X = [\ Gold, Silver, Bronze, Total\ Medals, Participation\ Count, Stability, Ranking\ Fluctuation\ ] \tag{4}$$

Construct a good input into the feature matrix of above formula to obtain the classification results.

2.3. Analysis of Clustering Results and Evaluation

After using K-means clustering, this paper analyzed the characteristics of each cluster and give their interpretation: Category A: these countries have many competitions and stable medal results, with low fluctuations. Category B: these countries many times, but the medals volatile, unstable performance. Category C: These countries have only started competing in recent years, and their medal count is small but gradually improving. Category D: These are countries that have competed fewer times and have won zero or very few medals.

This article uses Calinski-Harabasz Index to evaluate K-means clustering effect. The algorithm calculates based on the variance between clusters and clusters to effectively measure the compactness and separation degree of clusters. When CH index is larger, the greater the degree of separation between clusters, cluster sample within the closer, the better clustering effect. CH index calculation formula is as follows:

$$CH = \frac{\text{tr}\left(B_k\right)}{\text{tr}\left(W_k\right)} \times \frac{N - K}{K - 1} \tag{5}$$

$tr(B_k)$ Is the trace of the cluster dispersion matrix (total deviation) and indicates the degree of separation between clusters. It calculates the deviation between the cluster center and the mean of the entire sample. $tr(W_k)$ Is within the cluster dispersion matrix trace, said samples of tightness in the cluster, calculates the deviation between the sample and the cluster center.

By calculating this value, this paper judged whether the clustering effect is good or bad. The larger the CH index, the better the clustering effect.

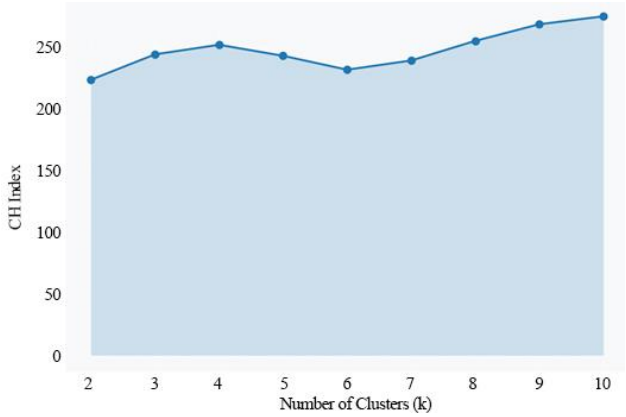

Fig 1. CH index analysis for determining the optimal number of clusters.

When too much is not convenient to data classification statistics and forecast, so this article wants to classified as little as possible. As shown in Fig. 1, when $k = 4$, CH value of max-imum points, this paper constructed model is more reasonable.

Based on the K-means clustering algorithm and using CH index evaluation to build a perfect and steady point of classification model, this model is based on classified in the Olympic Games show differences in different countries. This analysis is not only helpful to understand different types of national teams, but also can provide data support for subsequent sports management and policy making.

3. Medal Prediction: KAN-LSTM Hybrid Model

This paper predicts the number of Olympic medals based on several constraints, including the host country, the number of events in each Olympic Games, and the division of sports.

3.1. Features

This paper constructs input data based on the following features: whether for the host country, number of athletes, and number of new events added, historical Medals (Gold, Silver, Bronze, and Total) and entry number. $Host$ Indicates whether it is the host country or not (the results are denoted by 0 or 1). A Denotes the number of athletes. P Is the number of sports added? G, S, B Represents the number of gold, silver and bronze medals, respectively. T Denotes the total number of medals. N Denotes the number of entries. And combine these features into the input data:

$$X = [\ Host, A, P, G, S, B, T, N] \tag{6}$$

These features will be input as the input to the LSTM and KAN neural network to predict.

3.2. LSTM Model

LSTMS process time series data and are used to capture trends in past medal data and make future predictions. In this paper, the history of gold medals and total medals as a time series data input into the LSTM, to predict the future of gold medals and total medals.

Forget gate:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{7}$$

Input gate:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{8}$$

Candidate cell states:

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{9}$$

Cell status update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{10}$$

Output gate:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{11}$$

Hidden state:

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{12}$$

The LSTM goes through these gates to capture the time-series patterns of Olympic gold medals and medal totals and make predictions

3.3. KAN (K-Nearest Neighbor Neural Network) Model

KAN predicts the target (e.g., the number of gold medals and total medals) by calculating the similarity between the input and other samples in the historical data. In this task, we calculate the similarity of each country's historical performance (e.g., number of gold, silver and bronze medals, number of games played) to other countries and make the prediction by weighted average. KAN's core formula is as follows.

Calculate the Euclidean distance for the input and training samples:

$$d\left(x, x_i\right) = \sqrt{\sum_{j=1}^{m}\left(x_j - x_i\right)^2} \tag{13}$$

Select the most similar samples:

$$\left\{x_{i_1}, x_{i_2}, \ldots, x_{i_K}\right\} = \textit{ Top K nearest neighbors of } x \tag{14}$$

Weighted average (regression task):

$$\hat{y} = \frac{\sum_{k=1}^{K} w_k \cdot y_{i_k}}{\sum_{k=1}^{K} w_k} \tag{15}$$

3.4. AdaBoost Ensemble Algorithm

In this paper, the prediction results of LSTM and KAN are integrated by AdaBoost. AdaBoost improves the overall performance by weighting the results of multiple weak models. Specifically, AdaBoost aggregates the prediction results of LSTM and KAN to obtain the final prediction.

AdaBoost formula:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\grave{o}_t}{\grave{o}_t}\right) \tag{16}$$

$$H(x) = \sum_{t=1}^{T} \alpha_t \cdot h_t(x) \tag{17}$$

3.5. Uncertainty Estimation

In this paper, Monte Carlo simulation is used to estimate the uncertainty of model prediction results. Monte Carlo simulation trains the model multiple times to obtain different predictions and computes the standard deviation of these predictions.

Train the model multiple times and get the predictions:

$$\hat{y}_i^{\text{mean}} = \frac{1}{N}\sum_{n=1}^{N}\hat{y}_i^{(n)} \tag{18}$$

Calculate the standard deviation:

$$\hat{\sigma}_i = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{y}_i^{(n)} - \hat{y}_i^{\text{mean}}\right)^2} \tag{19}$$

3.6. Medal Prediction Results and Analysis

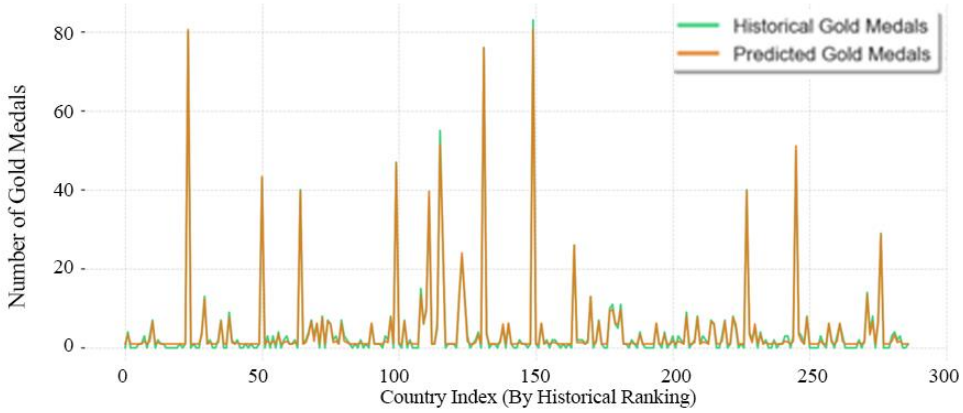

Fig 2. Comparative analysis of the prediction of Olympic gold medals.

Fig. 2 shows the comparative analysis chart of the Olympic gold medal prediction, showing the trend comparison between the actual value and the predicted value. Through the above, can be directly compared gold medals and national history model predicts gold medals of differences and trends. After observation found that the real value and predicted value towards the basic consistent, that model has good prediction ability. The fusion of LSTM and KAN neural network and the AdaBoost ensemble algorithm can effectively predict the gold medal ranking and the total medal ranking of 2028 Olympic Games. Combined with the Monte Carlo simulation to estimate the uncertainty of prediction results, further help us to assess the degree of reliability of prediction. In the end, using MSE and R^2 score evaluation index to measure the prediction ability of model.

After evaluation, the MSE of the model is 0.8001, which indicates that the MSE between the predicted value and the real value is at a low level. The R^2 score was 0.9886, which was close to 1, indicating that the model had an excellent fitting effect on the data, could well explain the changes of the dependent variable, and had high prediction accuracy and reliability.

Finally, the medal list interval was obtained by prediction, and the average of the interval was taken as the number of medals. The specific results are shown in Table 1.

Table 1. Predicted gold medal list and total medal list

Rank	Gold Medal	Total
America	52	155
China	33	105
England	16	75
France	14	65
Japan	11	55
Australia	9	50
German	8	45
Italy	7	40
Netherlands	6	35
Canada	5	30

The table shows that the UK, Germany, South Korea and Sweden are likely to decline. The UK predicts 16 gold medals, Germany eight, South Korea six and Sweden five. The loss of athletes to other countries or regions, changes in their training systems and increased international competition may make it difficult for these countries to maintain their previous results and reduce the number of medals at the 2028 Games.

4. Coach Impact Analysis: DID and RDD Models

In the Olympic Games, coaches play a crucial role, and the influence of "famous and handsome effect" on the number of medals won by various countries has attracted much attention. This study aims to double difference (DID) model and time series breakpoint regression (RDD) two models quantify the influence of the coach, by comparing the two kinds of model, further to explore the internal mechanism of "famous handsome effect", the influence degree, and provide targeted decision-making reference for national Olympic committee.

4.1. Double Difference (DID) Model Construction

1) Model setting

$$T_{ijt} = \beta_1 D_{ijt} + \beta_2 Q_{ijt} + \beta_3 \left(D_{ijt} \times Q_{ijt} \right) + \gamma \text{ Host }_{it} + \delta E_{jt} + \theta A_{ijt} \tag{20}$$

2) Solution steps

Firstly, data grouping: divide the data into experimental group and control group. The experimental group selected the country-project combination with the changed coach, while the control group selected the same type combination without the changed coach, and accurately matched according to the historical medal count and project type to ensure that the two groups of data were comparable in other conditions

Then, the parameters of the model were estimated by ordinary least squares (OLS) method $\beta_1, \beta_2, \beta_3, \gamma, \delta, \theta$

Derivation of the formula:

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y \tag{21}$$

Significance test: The test is performed $\beta_3 \neq 0$ by calculating the t statistics and p values. If $p < 0.05$, declined to the original assumption, that coach to replace medal significant impact.

Effect of quantitative: coach effect through formula: $coach\ effect\ = \widehat{\beta_3} \times Q$

4.2. Regression Discontinuity Design (RDD) Model Construction

1) Model specification

$$T_{ijt} = \alpha + \beta_1 \left(t - t_c \right) + \beta_2 D_{ijt} + \beta_3 \left(t - t_c \right) \times D_{ijt} + \gamma \text{ Host }_{it} + \delta E_{jt} \tag{22}$$

2) Solution steps

Data selection near the breakpoint: The data of each Olympic Games before and after the change of coach are selected k to ensure that the data has local continuity near the breakpoint, to accurately capture the impact brought by the change of coach.

Parameter estimation: Piecewise linear regression was used to fit the model before ($t < t_c$) and after ($t \geq t_c$) the breakpoint.

Before the breakpoint: $T = \alpha + \beta_1 (t - t_c)$

After the breakpoint: $T = (\alpha + \beta_2) + (\beta_1 + \beta_3)(t - t_c)$

Effect test: Compare the difference in intercept and slope before and after the breakpoint. If $\beta_2 > 0$ or $\beta_3 > 0$ indicates that the number of medals increased after the change of coach.

4.3. Model Solution and Result Analysis

Extract coach-effect estimates from individual models, such as the $\widehat{\beta_3}$ in DID model. At the same time, the prediction interval was calculated, and the confidence interval of DID model was calculated by Bootstrap method to evaluate the uncertainty of model prediction. The results of different models were compared, and the influence degree and significance of “famous effect" were comprehensively analyzed. According to the results, the "famous handsome effect" indeed helps potential countries to win more gold medals.

Fig. 3 shows the visualisation results of model analysis, including model assessment index comparison, comparison of coach effect, DID model residual figure and RDD model residual figure four parts.

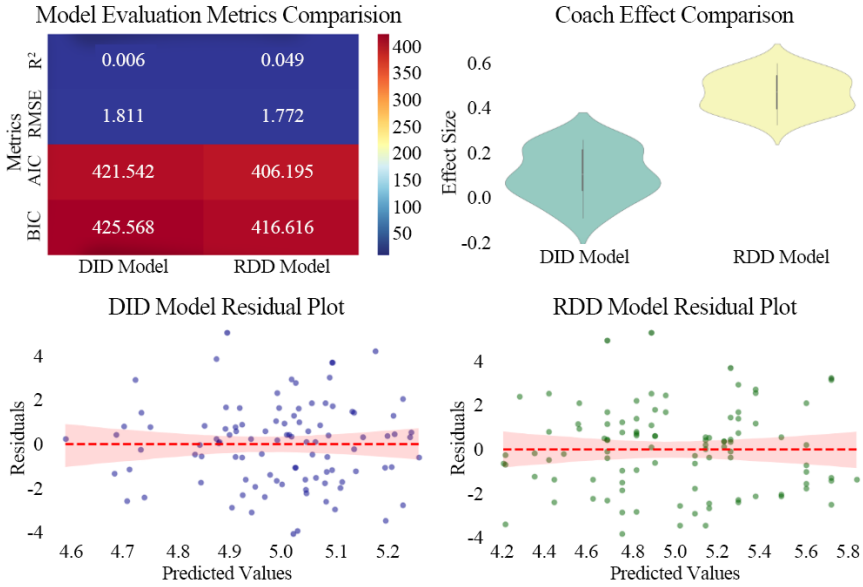

Fig 3. Model analysis results visualization.

Through the analysis of these charts, RDD model performed better than DID model on multiple evaluation index, the estimate of the effect of coach, and on the residual error performance have different characteristics, these results to evaluate the effectiveness and applicability of model and understand the coach effect provides an important reference.

5. Summary

This study establishes a multi-dimensional analysis framework for complex systems based on a KAN-LSTM hybrid model, achieving end-to-end modelling from data classification, trend prediction to causal effect assessment. Firstly, the study uses the K-means clustering algorithm combined with the Calinski-Harabasz index to divide multi-feature datasets into sample groups with significant temporal feature differences, providing a structured data foundation for subsequent modelling. Second, the KAN-LSTM hybrid model is constructed by integrating the temporal dynamic capture capabilities of LSTM with the sample similarity calculation advantages of KAN, combined with the AdaBoost ensemble strategy, significantly improving the prediction accuracy of non-linear trends and validating the effectiveness of multi-model collaboration in complex system prediction. Finally, DID model and RDD model were employed to quantify the causal effects of intervention events on target variables through comparative analysis between experimental and control groups and identification of discontinuity effects at intervention timepoints, overcoming the limitations of traditional correlation analysis in causal logic inference. Future research could further incorporate attention mechanisms or graph neural networks to enhance the processing capabilities for high-dimensional heterogeneous data, thereby expanding the application scope of this framework across broader domains.

References

[1] Ren Liqiang, Jia Shuyi, Wang Haipeng, et al. A Review of Time Series Classification Based on Deep Learning [J]. Journal of Electronics and Information Technology, 2024, 46 (08): 3094-3116.

[2] Guan Zheng, Yin Yongqiang, Zhang Xiaoxiang, et al. Assessment of flash flood hazard suscep-tibility in small watersheds based on K-means clustering and ensemble learning algorithms [J]. Journal of Applied Sciences, 2024, 42 (03): 388-404.

[3] Tang Qingwei, Xiang Yue, Dai Jiakun, et al. A Power Transfer Prediction Method for Wind Farms Based on CNN–LSTM [J]. Engineering Science and Technology, 2024, 56 (02): 91-99.

[4] Cheng Runkun, Wang Hui, Liu Da, et al. Integrated Forecasting of Spot Electricity Price Het-erogeneous Models Using the RSDE Framework and KAN Algorithm [J]. Journal of Electrical Engineering of China, 2024, 44 (24): 9645-9658.

[5] Shi, Pengzhen, Wei, Xia, Zhang, Chunmei. et al. Short-term wind power forecasting based on VMD-BOA-LSSVM-AdaBoost [J]. Journal of Solar Energy, 2024, 45 (01): 226-233.

[6] Qu Shuiling, Zhang Yue, Wang Qiqi, et al. Breakpoint Regression Method and Its Application Implementation [J]. Journal of Environmental Health, 2024, 14 (01): 1-7.

[7]