Based on Decision Tree and Improved PCA Algorithm for the Analysis of the Olympic Games in the Future

Yuanhao Zhang(✉) and Jiayi Zhang(✉)

Xi'an Eurasia University, Xi'an, Shaanxi, China
524834748@qq.com, 1332498380@qq.com

Abstract. In recent years, the Olympic Games have been faced with no country or city to host the situation, mainly because of its events, the number of participants and the scale of investment and other aspects of the host country's ability to be very high. In this paper, 68 indexes of 211 countries in the world such as economy, infrastructure, international reputation and national quality are collected. After dimensionality reduction of indexes by improved principal component analysis method, eight effective principal components are obtained. Then, based on the eight principal components, the decision tree model and ID3 algorithm are used to evaluate the ability of 211 countries to host the Olympic Games, so as to provide solutions for the future Olympic Games.

Keywords: Decision tree · Principal component analysis · Olympic Games · ID3 algorithm · Machine learning

1 Introduction

1.1 Background Introduction

Since the 1960s, the number of events, the number of participants and the scale of investment in the Olympic Games have grown rapidly, but the sheer size of the Games has made it unaffordable for host countries and cities, leading to a shift from enthusiasm to apathy. The most urgent task is to come up with a new plan for the Games and raise the profile of the Games. Therefore, this paper proposes a solution based on principal component analysis and decision tree method.

1.2 Index Selection

Firstly, this paper collects various indicators that will affect the holding of the Olympic Games. Based on the analysis of the experience of countries that successfully held the Olympic Games in the past, it considers from the aspects of national economic conditions, traffic conditions, industrial construction capacity, people's education level, national or city prestige, urban modernization degree, etc., to establish a complete and sound, extensive and effective indicator data.

© The Author(s) 2023
D. Kumar et al. (Eds.): IEIT 2023, AHSSEH 10, pp. 1028–1036, 2023.
https://doi.org/10.2991/978-94-6463-230-9_124

On this basis, we collected a total of 68 indicators that may affect the holding of Olympic Games in 211 countries in the world in the past 50 years, and divided them into positive and negative categories to ensure that this index system can basically cover the comprehensive situation of the whole country. They mainly include:

1.3 Latest Progress

There are two possible ways to deal with the current predicament of the Olympic Games. Plan one: Both the Summer and Winter Games should have a fixed location. Such a scheme would put enormous pressure on the host country, so we gradually increased the index requirements on top of the cost of the existing Olympic Games, and found countries that could withstand the pressure in the projection pursuit model. Plan Two: Divide the Olympic Movement into four groups. The second plan should focus on the quality of each country's hosting of the Games, such as urban construction, temperature and other factors. After adjusting the weights of some indicators, it is found that the number of countries able to host the Olympic Games has increased significantly, indicating that the second plan can reduce the cost of hosting the Olympic Games and increase the enthusiasm of all countries to host the Olympic Games. We choose the most appropriate way to host the Olympic Games according to the evaluation results of the projection pursuit method (Table 1 and Fig. 1).

2 Principal Component Analysis Algorithm

Through the above data collection and preliminary analysis, it can be concluded that each index has different impact values on the holding of the Olympic Games. Therefore, it is necessary to screen and analyze the indicators in establishing the index analysis model, and put forward the treatment of abnormal indicators or indicators with low impact [1].

It is necessary to solve the problem of increasing popularity of the Olympic Games, that is, according to the collected data of various countries combined with comprehensive condition analysis of the impact of the Olympic Games, and the use of principal component analysis method to get more important indicators to put forward solutions.

Table 1. Primary index interpretation

Index system	Indicator meaning
National economy	Reflect the economic background of the country hosting the Olympic Games
Urban construction	Reflecting the ability to build Olympic sites and accommodate visitors
Social culture	Reflect the quality of residents and the acceptance of the Olympic Games
Tourism situation	Reflects the ability to accept foreign tourists
International reputation	Reflect the country's ability to promote the Olympic Games
Future development	Reflect the ability to host the future Olympic Games

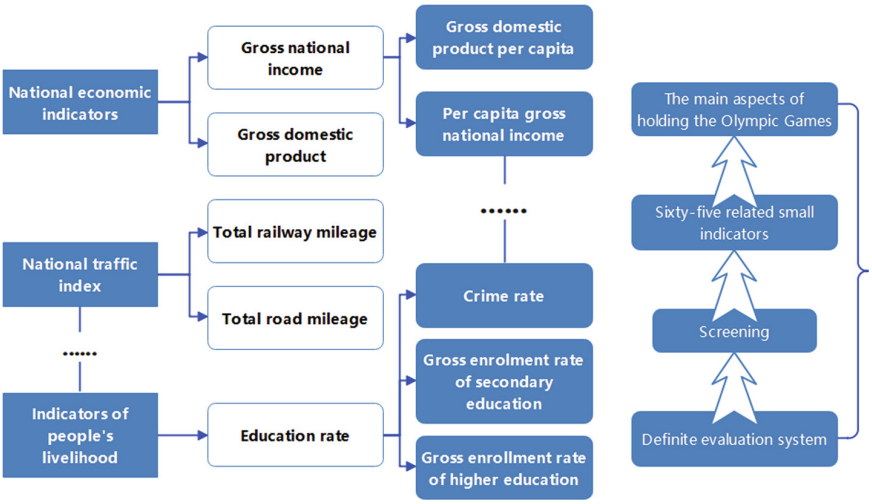

Fig. 1. Evaluation index system

Reasons for choosing principal component analysis: It is necessary to process data with many variables. There are a lot of variables and data, but there may be noise and redundancy. Because some of these variables are related, one of them can be selected from the relevant variables, or several variables can be integrated into one variable as a representative. Use a few variables to represent all variables, to explain the problem to be studied to achieve dimensionality reduction.

PCA can transform a group of possibly correlated variables into a group of linearly unrelated variables through orthogonal transformation, which is called the principal component. In order to analyze a problem comprehensively, many variables (or factors) related to the problem are often presented, because each variable reflects some information to varying degrees. The size of the information is usually measured in terms of the sum of squares of deviation or variance [2].

In the data set, our data set is N-dimensional, with a total of m data $(x^{\{1\}}, x^{\{2\}}, ..., x^{\{m\}})$. The dimensionality of the m data is reduced from n dimension to k dimension, hoping that the m K-dimension data set can represent the original data set as much as possible. In order to avoid the loss of data from n dimension to k dimension, the most suitable solution is obtained through the optimal solution distance in PCA.

Among the selected indicators, some of them belong to high quality indicators, that is, the higher the value, the better the ideal value. In order to deal with the above indicators with different properties and correctly reflect the comprehensive results of different forces, it is necessary to do isochemotaxis treatment.

Assume that each index contains m numerical, x_{new} as a index with the data after the chemotaxis, $x_{\max}$ is the maximum of the index, $x_{\min}$ to the minimum. There are respective formulas for calculating indexes of different properties:

(1) The following formula is adopted for the homochemotaxis of high performance indicators:

$$x_{new} = \frac{x_{\max} - x_i}{x_{\max} - 0}, i = 1, 2, ..., m \tag{1}$$

(2) The following formula is adopted for the homochemotaxis of low optimal indexes:

$$x_{new} = \frac{x_i - x_{\min}}{x_{\max} - 0}, i = 1, 2, ..., m \tag{2}$$

Each index and country in the data set constitute matrix X, and X is standardized to obtain the standardized matrix A:

$$X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nm} \end{pmatrix} \Rightarrow A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{19} \\ a_{21} & a_{22} & ... & a_{29} \\ ... & ... & ... & ... \\ a_{1251} & a_{1252} & ... & a_{1259} \end{pmatrix}$$

Improvement to PAC model: There are two cases. The first is that the distance between sample points and this optimal position is close enough; the second explanation is that the projection of sample points on this line can be separated as far as possible.

If K is extended from one dimension to any dimension, the criterion for dimensionality reduction is that the sample points are close enough to the hyperplane, or that the projections of the sample points on the hyperplane can be separated as far as possible. Based on the above two criteria, two extensions of PCA can be obtained, based on which the optimized PAC model can be obtained:

Where: in the forward processing of data, the distance of the data is defined, and the long distance optimum and short distance optimum are solved separately. The short range solution is based on the minimum projection distance.

Let m n-dimensional data have been centered:

$$\{x_1, x_2, ..., x_m\} \to \sum_{i=1}^{m} x_i = 0 \tag{3}$$

The new coordinate system is obtained by projection: $\{w_1, w_2, ..., w_m\}$.

Where w is the standard orthonormal set: $||w||_2 = 1, w_i^T w_j = 0$. Transforming the data from n-dimensional to W-dimensional requires discarding some of the coordinates in the new coordinates.

The projection of the sample points in the n-dimensional coordinate system is $z_i = \{z_1, z_2, ..., z_m\}^T$.

Where: $z_j = w_j^T x_i$ is the JTH dimensional coordinate of x in the low-dimensional coordinate system.

Using z_i to recover the original data x_i, the recovered data $\overline{x}_i = \sum_{j}^{w} x_j^i w_i = W_z^i$ is obtained, where w is used as the orthogonal basis to form the matrix. Based on this idea, considering the whole sample set, all the samples meeting the requirements are close enough to this hyperplane, and the minimization equation is obtained:

The formula is simplified to obtain:

$$\sum_{i=1}^{m} \|\overline{x_i} - x_i\|_2^2 = \sum_{i=1}^{m} \|Wz_i - x_i\|_2^2 = -tr(W^T XX^T W) + \sum_{i=1}^{m} x_i T x_i \tag{4}$$

$\overline{x}_i = W_z^i$ and sum of squares expansion are used for simplification, and matrix transformation formula is used: $(AB)^T = B^T A^T$, $W^T W = I$ and similar items are merged in the simplification process.

It is worth noting that in the process of summation x_i is the covariance matrix of the data set collected by the Olympic Games, and each vector in W is an orthonormal basis.

$$\underbrace{\arg\min}_{W} -tr(W^T XX^T W), S.T. W^T W = I \tag{5}$$

The derivative of the optimization model with respect to W has $XX^T W + \lambda W = 0$ to obtain: $W^T XX^T W = \lambda W$.

In this way, it can be seen more clearly that W is the matrix composed of n eigenvectors of XX^T, while A is the matrix composed of several eigenvalues of XX^T, the eigenvalues are on the main diagonal, and the rest are O. When we reduce the dataset from n-dimension to W-dimension, we need to find the eigenvector corresponding to the largest w eigenvalues. The matrix w of these W eigenvectors is the matrix we need. For the original data set, the original data set can be reduced to the W-dimensional data set with the minimum projection distance. The Olympic index processing model based on improved PAC is obtained by using the above algorithm optimization as an improvement.

Based on the maximum projection variance (similar to the solution of minimizing distance): Suppose that m n-dimensional data have been centered, a new coordinate system is obtained by projection, and a new coordinate system is obtained by projection, where W is the standard orthogonal intersection: $||w||_2 = 1$, $w_i^T w_j = 0$.

$$\begin{cases} \{x_1, x_2, ..., x_m\} \to \sum_{i=1}^{m} x_i = 0 \\ \{w_1, w_2, ..., w_m\} \end{cases} \tag{6}$$

After the optimization objective based on the minimum projection distance, it can be found that it is exactly the same, as long as the negative number is minimized, otherwise it is maximized, and the Lagrange function is used to obtain:

$$J(W) = tr(W^T XX^T W + \lambda(W^T W - 1)) \tag{7}$$

Taking the derivative of W, we can get: $XX^T W = -\lambda W$.

As above, it can be seen that W is the matrix of n eigenvectors of XX^T, while $(-\lambda)$ is the matrix of a number of eigenvalues of XX^T, with the eigenvalues on the main diagonal and O in the remaining positions [3].

When reducing the dataset from n-dimension to W-dimension, it is necessary to find the eigenvector corresponding to the largest w eigenvalues. The matrix w composed of W eigenvectors is the computational matrix.

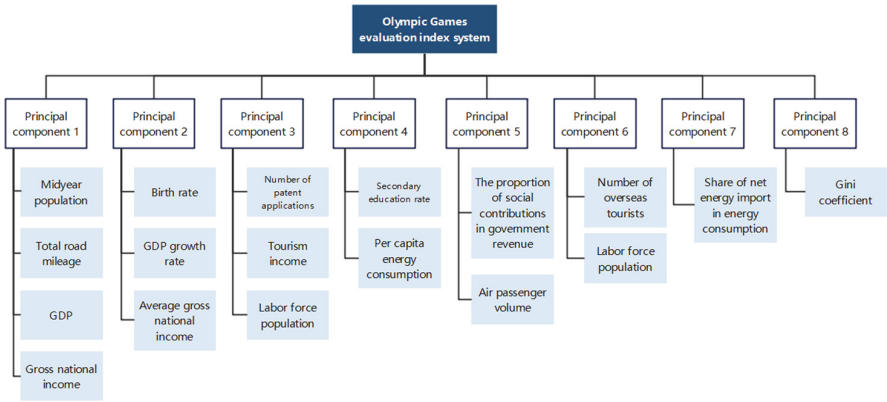

Fig. 2. Principal component analysis results

After optimization, the principal component analysis model is established to obtain the following process:

The principal component of the m-dimensional matrix of sample X is the corresponding matrix W of the first M eigenvalues of the covariance XX^T of the sample set. The PCA dimension reduction can be achieved by doing it for sample X (Fig. 2).

As for the proportion threshold of principal components that cannot be reduced dimension, SPSS statistical software is used to show that some indicators have strong correlation. If these indicators are directly used to grade countries, it will not only cause too much computation, but also cause information overlap and affect the objectivity of classification. Principal component analysis can convert multiple indicators into a few unrelated comprehensive indicators to obtain the results of principal component analysis.

3 Evaluation Model Based on ID3 Decision Tree

Decision Tree is a common algorithm in machine learning. Based on tree structure, decision tree directly simulates the decision-making process of human beings in real life. The final classification results were obtained by ballot. The elements that make up the decision tree are nodes and edges. The node will judge according to the characteristics of the sample. The initial branch points are called root nodes, the rest are called child nodes, and the nodes without branches are called leaf nodes. These nodes represent the classification results of the sample. In general, a decision tree contains a root node, several internal nodes and several leaves [4, 5].

Root node: Contains the full set of samples, and the path from the root node to each leaf node corresponds to a decision test sequence.

Internal node: represents a feature and attribute. Each internal node is a judgment condition and contains the set of data in the data set that satisfies all the conditions from the root node to that node. According to the attribute test results of the internal node, the data set corresponding to the internal node is divided into two or more child nodes.

Leaf node: Represents a class that corresponds to the decision outcome. The leaf node is the final category, and if the data is contained in that leaf node, it belongs to that category.

At present, most decision tree models adopt univariate as the test attribute, which will lead to problems such as large scale of the generated decision tree, difficult to understand classification rules, repeated subtrees of the decision tree, and multiple tests of some conditional attributes.

In this paper, multivariate test in the index system is used to construct an improved decision tree to build an evaluation model [6, 7].

3.1 Decision Tree Model Based on ID3 Algorithm

ID3 algorithm is a descriptive attribute optimization method based on entropy subtraction theory. The attribute to be tested is the one with the highest information value in the current sample set. The sample is divided into as many subsets as possible due to the different values of the attributes to be tested, and new nodes corresponding to the sample are added to the decision tree.

$$I(S_1, S_2, ..., S_m) = -\sum_{i=1}^{m} P_i \log_2(p_i) \tag{8}$$

where, pi is an arbitrary sample belonging to C probability; Use s/s to estimate. The information is coded in binary, so the logarithmic function has a base of 2.

$$E(A) = \sum_{j=1}^{m} \frac{S_{1j}, S_{2j}, ..., S_{mj}}{S} I(S_{1j}, S_{2j}, ..., S_{mj}) \tag{9}$$

where, $I(S_1, S_2, ..., S_m) = -\sum_{i=1}^{m} P_i \log_2(p_i) = S_{ij}/|S_j|$ is taken as the JTH, which is the probability that the sample belongs to Ci in sj. In this way, the information gain obtained by using attribute A to divide the corresponding sample set of the current branch node is:

$$Gain(A) = I(S_{1j} + S_{2j} + ...S_{mj}) - E(A) \tag{10}$$

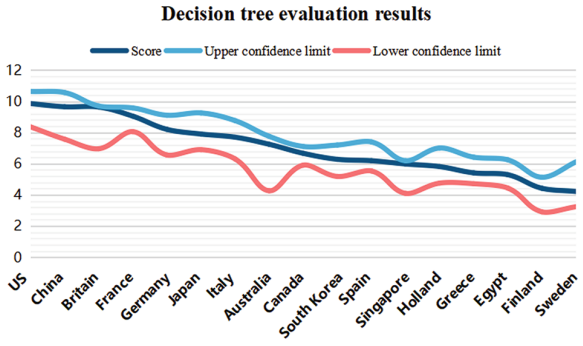

Fig. 3. Decision tree evaluation results

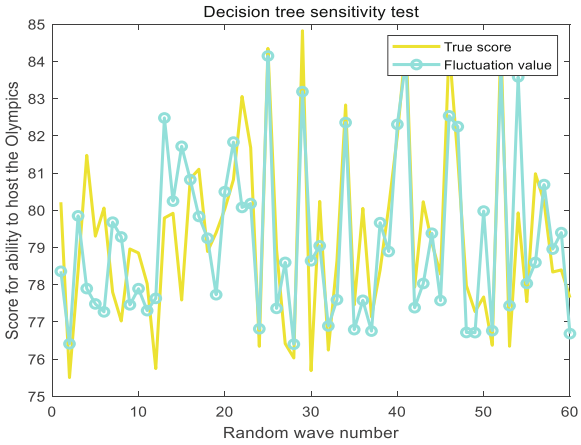

Fig. 4. Sensitivity test

Based on the information theory, ID3 algorithm uses the uncertainty of the sample set after partitioning as a measure of the quality of sample subset partitioning, and uses "information gain value" to measure the uncertainty – the greater the information gain value, the smaller the uncertainty, which prompts us to find a good non-leaf node for partitioning. Through ID3 algorithm, the information gain value of each influence factor is calculated, and a decision tree of each influence factor is gradually established [8, 9].

4 Conclusion

The evaluation scores of 211 countries are solved in MATLAB:

The score value obtained is between 0 and 100. The closer the value is to 100, the higher the degree of the evaluation unit is to the optimal level; otherwise, the worse it is.

At the same time, considering the possible errors, we make the upper and lower limits of the score into consideration, so as to analyze the score results more intuitively (Figs. 3 and 4).

We tested the sensitivity of the decision tree model by adjusting the data for each country in the annex for the parameters of each indicator known in the data set.

It can be seen from the figure above that TOPSIS evaluation model has strong stability [10].

It can be found that only the United States and China, with a decision tree score above 80, are most capable of hosting the Olympics, followed by France, Britain, India, Germany, Russia, Japan, Brazil and Australia.

References

1. Shuangfeng on the improvement of comprehensive evaluation of Principal component Analysis Huangshan 200111
2. Xu Yajing, Improvement of Application Method of Principal Component Analysis, Zhengzhou Institute of Light Industry, 2006

3. Sun Liupin, Improvement of Comprehensive Method Based on Principal PCA, Nanjing University of Aeronautics and Astronautics, 2009
4. Luan Lihua, Decision Tree Classification. Nanjing Normal University. 2004
5. Liu Xiaohu, Optimization Algorithm of Decision Tree, Harbin Institute of Technology, 1998
6. Zou Huanggang ID3 Decision Tree Algorithm Analysis and Application in Automobile Detection University of Shanghai for Science and Technology 2022
7. Xiong Yan Formative Evaluation Decision Tree Model and Countermeasures, University of Science and Technology Liaoning, 2023
8. Jing Meiyuan Research on Decision Tree Algorithm Based on Decision Path Shandong University of Technology 2023
9. Huang Xinxiong, Decision Tree Method based on Principal Component Analysis, China University of Petroleum, 2019
10. Ding Huanfeng, The Asymmetric Impact of Hosting the Olympic Games on the Economic growth of the host country, South China University of Technology, 2022

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.