Checking the Pulse and Temperature of Higher Education Based on K-means

Menghui Chu, Zekai Wang, Zirui Zhao

Electrical Information Department, Shandong University of Science and Technology, Jinan, Shandong, 250031, China

Abstract: *This paper is here to develop a model that can assess the health status of the higher education system in any country. Then select a country and propose a set of policies that will move that country from its current state to its target state. Five indicators for the evaluation of the education system are formulated. Then, Python crawler is used to capture the data of different countries under these five indicators from the Internet. Then, K-means++ clustering method is used to classify countries into three classes. The significance and specific results of classification are shown in 4.2.1 and Table3. Then, according to different types of countries, TOPSIS based on entropy weight method is used to establish a system that can evaluate the health status of higher education in any country.*

Keywords: *Python crawler, K-means++ clustering, TOPSIS, PCA, BP neural network, Fuzzy comprehensive evaluation*

1. Background

Higher education is of great importance to every country. From Germany to the United States, from Japan to Australia, we see different ways of higher education in different countries. Each country not only trains its own students but also attracts a large number of international students every year. The diversity of students and higher education means the exchange and integration of different cultures around the world. In such an ecological environment, it is particularly important to establish a health assessment standard for higher education. Suggestions for countries with low health level of higher education and implementation are of great help to the improvement of a country's comprehensive strength.

There are two main tasks that are required of us in this question. Task 1: Develop a model that can assess the health of the higher education system in any country. Task 2: Select a country based on a series of analyses, determine a healthy and sustainable state (target state), and come up with a set of policies that can move the country from its current state to its ideal state. it will need to collect the data of several indicators of the evaluation system in different countries by using crawler and other methods, and then pre-process the data. Then it can classify countries by K-means clustering according to the collected data, and develop a system that can evaluate the health status of higher education in any country by using TOPSIS evaluation model based on entropy weight method according to different types of countries. In this way, the country types will then be judged and can then be evaluated under the respective categories.

2. National higher education health assessment model

2.1 Data analysis

According to differences in the levels of each country's higher education as well as references [1], and the model of higher education level evaluation index can be divided into five indicators, including every country of the higher education popularity, overseas students proportion, the proportion of expenditure on education in each country, the enrollment of higher education, the proportion of people receiving higher education. For the required data, an automatic program (namely Python crawler) that captures data information from the Internet is used here to collect the required data in the reference [2], and the corresponding countries of the missing data in different years are excluded one by one to obtain the original data. Combined with the horizontal analysis method mentioned in the time series index analysis in reference [3], the data of different years were processed and the key data supporting the establishment of the model were obtained

$$\overline{X} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{\sum x}{n}$$

2.2 Establish the health status evaluation model of higher education system

Based on the principle of clustering algorithm, the collected data are first used to divide countries into three categories according to the level of education investment, and each category has its own evaluation system. After the weight of each index of each country is calculated, any country can be classified according to the principle of clustering algorithm. (Use entropy weights) After determining which country is in the corresponding category, the country can be put into the data set of the corresponding category and scored by TOPSIS, so as to complete the evaluation.

2.2.1 K-means clustering was used to classify countries

In order to make the model universal, the health status of the higher education system of any country can be assessed, and the 96 countries collected can be classified by Kmeans++clustering algorithm in combination with references. The classification results are the proportion of higher education investment in GDP (category A), the proportion of secondary education investment in GDP (category B), and the proportion of low education investment in GDP (category C). The idea of clustering is as follows:

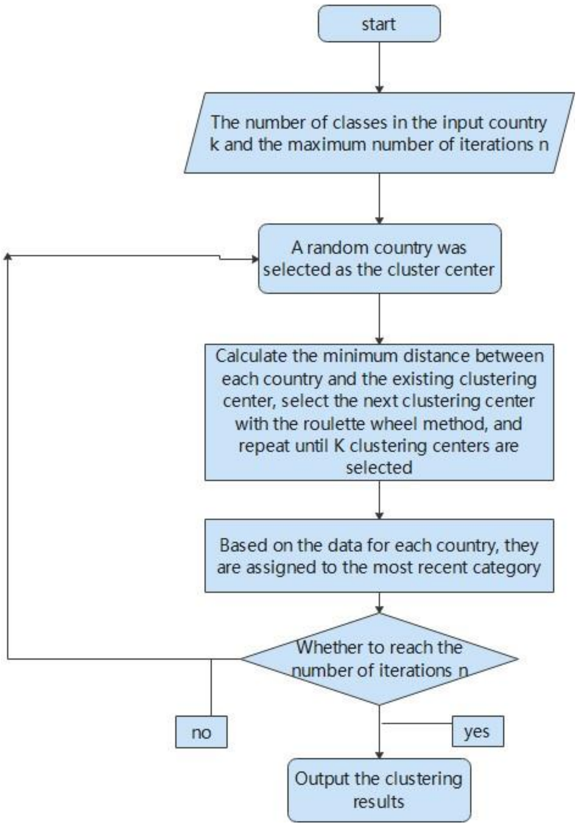
start
The number of classes in the input country k and the maximum number of iterations n
A random country was selected as the cluster center
Calculate the minimum distance between each country and the existing clustering center, select the next clustering center with the roulette wheel method, and repeat until K clustering centers are selected
Based on the data for each country, they are assigned to the most recent category
Whether to reach the number of iterations n
no
yes
Output the clustering results

Figure 1: Cluster analysis ideas

In order to ensure the reliability of the results, all indicators were standardized before clustering, and then the data standardized results were obtained, and then the standardized results were clustered.

Step 1: Build the make blobs module in the sklearn. datasets and the K-Means module in the sklearn. cluster into the Python environment

Step 2: According to the above reference [4], set k=3 and n=20, and then find the initial clustering center according to the model established in the previous step.

Step 3: Calculate the minimum distance between each country and the clustering center selected in the previous step. Select the next clustering center according to the roulette wheel method, and continue to cycle until 3 clustering centers are selected. The results are as follows:

Table 1: the final cluster centers

	clustering		
	C	B	A
Zscore (Higher Education Penetration by Country)	-1.17024	-0.49511	1.03137
Zscore (Ratio of International Students by Country)	-2.61296	0.14436	0.37452
Zscore (The proportion of expenditure on education in each country)	0.88518	0.5581	-0.28844
Zscore (Enrolment rate of higher education)	-1.16981	-0.46086	0.97843
Zscore (The proportion of the population receiving higher education)	-1.16981	-0.46086	1.08669

Step 4: Assign all countries to the nearest class according to their data under different indicators until the set number of iterations n=20 is reached, and then output the clustering results. The results are as follows:

Table 2: Clustering results

Type	Country	Distance from the sample to the center of the class
A	Australia	1.20549
A	Bulgaria	1.01575
A	Canada	2.63432
A	Japan	1.57486
A	Russia	1.28785
…	…	…
A	United States	2.81053

The distances between the categories and the number of countries in each category were counted

Table 3: Clustering results statistical

clustering	C	8.000
	B	54.000
	A	35.000
effective		97.000
missing		0

2.2.2 Based on entropy weight method, five indexes of three kinds of countries are given weight respectively

Step 1: Determine whether there is a negative number in the input matrix, and if there is, re-normalize to a non-negative interval

Assume that there are n objects to be evaluated (in this case, the country to be evaluated), and the forward matrix composed of m evaluation indicators (which have been positive change) is as follows

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

$xn1$ … xnm

So, the normalized matrix is Z, each of the entries in Z:

$$Z_{ij} = \frac{xij}{\sqrt{\sum_{i=1}^{n} x_{ij}^{2}}}$$

Determine if there are any negative numbers in the Z matrix, and if so, use another normalized method for X

The matrix X is normalized once to obtain the $\tilde{Z}$matrix, and the normalization formula is

$$\tilde{Z}_{ij} = \frac{x_{ij} - \min\{x_{1j}, x_{2j}, \ldots x_{nj}\}}{\max\{x_{1j}, x_{2j} \ldots x_{nj}\} - \min\{x_{1j}, x_{2j}, \ldots x_{nj}\}}$$

Step 2: Calculate the proportion of the I country in the j index, and regard it as the probability used in the relative entropy calculation

Suppose there are n countries to be evaluated and m evaluation indicators (i.e. the indicators that determine the health of the education system), and the non-negative matrix obtained

Calculate the probability matrix P, where the calculation formula of each element p_{ij} in P is as follows:

$$p_{ij}=\frac{\tilde{z}_{ij}}{\sum_{i=1}^{n}\tilde{z}_{ij}}$$

Easy to verify: $\sum_{i=1}^{n}p_{ij}=1$ In other words, the probability sum corresponding to each index is guaranteed to be 1.

Step 3: For the jTH index, its information entropy can be calculated as follows:

$$e_j=\frac{-1}{\ln n}\sum_{i=1}^{n}\ln(p_{ij})\ \ (j=1,2\ldots m)$$

Make the information utility value

$$d_j=1-e_j,$$

d_j normalized so that the entropy weight of each index can be obtained:

$$w_j=\frac{d_j}{\sum_{j=1}^{m}d_j}$$

Finally, we use MATLAB to get the index weights of the three types of countries.

2.2.3 The establishment of TOPSIS evaluation model based on entropy weight method

Step 1: Unify metric types ------ The original matrix is going forward. The forward expression of the original matrix is to convert all the indicators into extremely large ones. Here, the share of overseas students is a very small indicator which should be transformed into a very large indicator. The algorithm is as follows:

tran=max-xi

Max represents the maximum value of the share of overseas students, xi represents the share of overseas students of each country, and Tran represents the result after the index is positive.

Step 2: Normalize the forward matrix--- The purpose of standardization is to eliminate the influence of different dimensions of indicators. In order to ensure a higher accuracy of the evaluation system, the forward matrix is standardized here, with 96 objects to be evaluated and 5 indicators to be evaluated. The forward matrix composed is as follows:

$$X=\begin{bmatrix}x_{11} & \cdots & x_{1m}\\ \vdots & \ddots & \vdots\\ x_{n1} & \cdots & x_{nm}\end{bmatrix}$$

So, for the normalized matrix, it is going to call it Z, so for each of the entries in Z

$$z_{ij}=x_{ij}/\sum_{i=1}^{n}x_{ij}{}^{2}$$

This makes the evaluation result more accurate.

Step 3: Calculate and normalize the score-- A standardized matrix of 96 objects to be evaluated and 5 evaluation indexes is provided:

$$Z=\begin{bmatrix}z_{11} & \cdots & z_{1m}\\ \vdots & \ddots & \vdots\\ z_{n1} & \cdots & z_{nm}\end{bmatrix}$$

Defining a maximum z+ =(z1+, z2+, …zm+)=(max{z11, z21,…, zn1}, max{z12, z22,…, zn2},…, max{z1m, z2m,…, znm}).

Define minimum z-=(z1-, z2-, …zm-)=(max{z11, z21,…, zn1}, max{z12, z22,…, zn2},…, max{z1m, z2m,…, znm}).

And this vector means to minimize each of these indices, w_j is the entropy weight in 4.3.

Define the i (i= 1,2..., n) Distance between the evaluation object and the maximum value

$$\mathrm{D_i^+}=\sqrt{\sum_{j=1}^{m} w_j\left(\mathrm{z_j^+}-\mathrm{z_{ij}}\right)^2}$$

Here, the sum of Euclidean distances between each index in each country and the maximum value of that index

Define the i (i = 1,2..., n) Distance between the evaluation objects and the minimum

Value

$$Di^- = \sqrt{\sum_{j=1}^{m} w_j\left(\mathrm{z_j^-}-\mathrm{z_{ij}}\right)^2}$$

Here, the sum of Euclidean distances between each index in each country and the minimum value of that index can calculate the i (i = 1,2..., n) countries with unnormalized scores:

$$\mathrm{S}=\frac{D_i^-}{D_i^+ + D_i^-}$$

Score normalization processing:

$$s_i^{\sim} = s_i / \sum_{i=1}^{n} s_{i=1}^{\sim}$$

2.3 Solving the evaluation model

Step1: is to use the algorithm principle of 4.1 to classify the 96 countries according to 2.2.1 Divided into three categories A, B and C, the clustering results are shown in Table 1 in 2.2.1.

Step 2: After classifying the categories, we will use the model in 4.2 to comprehensively score and rank the countries in each category according to the weight of each index. The specific results are as follows (due to the large amount of data, some of them are selected for analysis)

Table 4: The ranking results

A			B		C
United States	0.060755312	China	0.03567704	Gabon	0.261917915
Canada	0.055722679	Croatia	0.035150264	Zimbabwe	0.144299546
New Zealand	0.046122808	Austria	0.034869719	Namibia	0.133035854
Australia	0.04309378	Malaysia	0.034103021	Mauritania	0.114272892
Russia	0.04108928	Argentina	0.034056466	Mauritius	0.113641291
Ukraine	0.040419345	Slovakia	0.033920427	Rwanda	0.090591301
Sweden	0.038104114	Italy	0.032218395	Niger	0.08772546
…			…		…
Japan	0.037862651	Colombia	0.02982548	Malawi	0.054515741

The results showed,

A Top of the list is the United States score: 0.0607

B Top of the list is the China score: 0.035677

C Top of the list is the Gabon score: 0.263

(Scores: Because the weights of A, B and C are different, the evaluation system is different, and the scores are normalized. The scores of countries in each category can only be compared under their respective categories, but cannot be compared across categories. And according to the clustering principle of 4.2.1, in principle, the health degree of the national education system is A>B> C.) Through the analysis of the results, the ranking results are consistent with the reality that the model in 4.2 can be used to evaluate the health of the higher education system of any country.

2.4 Solving question two

According to the model and results of the first question, we have applied the model to many countries, and the results are referred to 4.3. According to the results of the above model, it is found that Vietnam belongs to B, that is, the proportion of education level investment to GDP in a medium type of country, and Vietnam scores the lowest in B (to sort out the ranking and write the ranking).

Through searching relevant literature, it is found that the Vietnamese government and society attach great importance to the development and improvement of education, and the scale and investment of education increase rapidly, which greatly promotes the expansion of education in Vietnam. However, Vietnam's overall level of economic and social development is limited, and its educational investment is relatively small, while the scale of education continues to expand, which puts forward higher requirements for the financial management of school education and the improvement of the utilization efficiency of financial investment. With reference to the overview in Reference [5], we chose Vietnam as a country with a higher education system that still has room for improvement

3. Conclusion

The comprehensive evaluation of each country according to the health evaluation sys-tem of higher education found that the comprehensive strength of a country can not completely replace its health level of higher education. After the dimensionality reduction of the influence index, it can be seen that the penetration rate of higher education has a great impact on the health status of a country & apos;s higher education. The formulation of policies requires the prediction of the results to ensure the high efficiency of the implementation of policies. Policy implementation also needs a long time, adhere to the implementation of the corresponding policies can be relatively accelerated to reach the tar-get state. Higher education is closely related to each of us, and we should actively respond to the policy of improving the health level of higher education formulated by the country.

References

[1] Wang Zhengqing, Wang Yin, Sun Xinyan. Research on the Measurement and Related Factors of Higher Education Competitiveness Level of Countries along the Belt and Road [J]. Journal of southwest university (social science edition), 2021, 47(01): 112-123+227. DOI:10.13718/j.cnki.xdsk.2021.01.011
[2] Zhang Tianshu. Excel based time series index analysis method in statistics [J]. Office automation, 2020, 25(22): 45-46+37.
[3] https://ourworldindata.org/global-education
[4] Liang Qian. Three challenges of quality assurance in higher education and their solutions [J]. Educational Research, Tsinghua University, 2020, 41(01): 142-148.