Evaluation of Higher Education System based on PCA-GA-PB Neural Network

Na Li\textsuperscript{1,a}, Haoqi Tian\textsuperscript{1,b} and Kailuan Yang\textsuperscript{1,c}

\textsuperscript{1}North China University of Science and Technology, Tangshan 063200 of China

\textsuperscript{a}2509505148@qq.com, \textsuperscript{b}1176409256@qq.com, \textsuperscript{c}ykh06618@163.com

Abstract

In the past 40 years, the development of higher education has received extensive attention from all over the world. The quality of higher education has had a profound impact on all aspects of society. The higher education system not only has the value of the industry itself, but also has the value of providing society with well-trained citizens. When we look around the world, we will see that different countries have different approaches to higher education. At the same time, the outbreak of COVID-19 has brought unprecedented challenges to higher education systems in various countries. Different countries' higher education systems have their own advantages and disadvantages. Therefore, it is necessary to establish a higher education health evaluation system to evaluate the higher education level of different countries in order to improve the national higher education level and make it healthier. Shock proof. This paper establishes a higher education health assessment model based on principal component analysis and genetic algorithm improved BP neural network. The national higher education rating is divided into 5,4,3,2,1 from high to low. It also assessed the health level of higher education in the United States, Australia, Germany, Japan, India and South Africa. This paper selects 12 indicators such as the number of papers and degrees as secondary indicators. Then, through principal component analysis, these twelve indicators are extracted as three primary indicators of the past, present and future of national higher education. Then, with three indicators as input and higher education health evaluation as output, GA-BP neural network evaluation model is used to construct a higher education health evaluation model. And for the BP neural network has the shortcomings of slow convergence speed and low accuracy, we use genetic algorithm to improve it. Finally, the health level of higher education in six countries was obtained, and the German higher education health system was discussed.

Keywords

National higher education; GA-BP neural network evaluation model; Principal component analysis.

1. Introduction

In the past 40 years, the development of higher education has received widespread attention worldwide. The quality of a country's higher education system has had a profound impact on all aspects of the country's society. Around the world, more and more people are receiving higher education. According to data compiled by UNESCO, the global gross tertiary enrollment rate (GTER) increased from 10% in 1972 to 32% in 2012, and is now growing at a rate of 1% every year. By 2014, 64 countries in the world had a gross enrolment rate of 50% in higher education, compared to only five countries 20 years ago. There are 14 countries in the world with a gross enrollment ratio of more than 80%. Scholars represented by Martin Trow believe
that higher education in the popularization stage will shape the country, the new relationship between education and society.
At the same time, the emergence of COVID-19 has also severely affected the higher education systems of various countries.
So, what will happen to higher education in the process of popularization?
It is important to further discuss and explain these issues by constructing a health evaluation model of the country's higher education system.

2. Our Work

2.1. Noun Explanation
Higher education: Higher education is professional education and vocational education conducted on the basis of completing secondary education, and is the main social activity for cultivating senior specialized talents and professional personnel. Higher education is one of the important components of the education system that are interrelated. It usually includes various educational institutions with high-level learning and training, teaching, research and social services as their main tasks and activities.
System health: measure the stability and sustainability of the system. We usually think that a healthy system has good development prospects and stable development status, and a healthy system is powerful and outstanding to a certain extent.
Higher education system: an organizational structure composed of higher education institutions (colleges, universities, etc.).
It also requires personnel and infrastructure to educate students above the intermediate level.

2.2. Assumptions
1. It is assumed that the education data collected by all countries/regions are true and reliable.
2. Assuming that only the 12 factors mentioned in this article are used to represent all the factors that may affect higher education.
3. Assume that the six countries mentioned in this article only represent the countries to be analyzed.
4. Higher education institutions are complex and diverse. Assuming that the analysis process is not affected by extreme data, because a university lacks an indicator, there will be no abnormal nationwide evaluation results.
5. It is assumed that in the process of research and testing of key principal components, only the influence of changes of key principal components on the evaluation results is considered, and the interaction between affecting principal components is temporarily excluded.

2.3. Principal Component Analysis Model Establishment
2.3.1. Multi-level Index Screening
In the data collection of this article, we divide the indicator set into the past, present and future directions of national higher education statistics. There are 12 indicators in total:
National History of Higher Education (HE): The past performance of the national higher education system is mainly manifested by past achievements. A country with a healthy higher education system must have many academic achievements such as scientific research. A school with a glorious past is likely to perform well now, and vice versa. The university was not built in a day. The school’s past education will be reflected in all aspects of the school’s current performance. Therefore, we will evaluate the health of the country’s current higher education system based on the school’s past achievements.
The main indicators are: the number of papers, the value of the degree, the number of Nobel Prize winners, the number of Fields Medal winners and the education index.

Current State of Higher Education (NE): Education now focuses on the data of students currently attending universities. Education is student-centered, and the most important indicator is the number of students. At the same time, education needs financial support. The proportion of national investment in education, the number of universities and the average tuition of higher education will directly determine whether education can proceed smoothly.

The future of national higher education (FE): The future of education will focus on the graduation prospects of students who are in school after graduation and those who are preparing to enroll.

Student prospects (employment rate), education opportunities (enrollment rate) and equality (gender ratio) will determine the future level of higher education in a country.

2.3.2. Sample Selection

To sum up, we selected 3 indicator categories and 12 indicator sets, and analyzed the data of a single country and a single year as a sample.

2.3.3. Data Preprocessing

In statistical analysis, there are differences in the order of magnitude or measurement units between different dimensions, which leads to incomplete variables and in turn leads to calculation results that have no practical significance.

Therefore, we need to standardize the data first to eliminate the influence of different dimensions to ensure the correct operation of the neural network.

Then standardize the data. The linear function is used to convert the original data linearization method into the range of [0, 1]. The normalization formula is as follows:

\[ X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \]  

(1)

Where, \( X_{\text{norm}} \) is the normalized data, \( X \) is the original data.

\( X_{\text{max}} \) and \( X_{\text{min}} \) are the maximum and minimum values of the original data set respectively.

This method realizes equal scaling of the original data.

2.3.4. Main Index Extraction

For the 12 indicators in the above quantified sample, if they are directly analyzed, they will be affected by noise and extreme data. At the same time, the excessive number of input layers and data will seriously affect the convergence speed. Therefore, it is necessary to adjust and reduce the size of all indicators. Since this article classifies the index set into index categories, it is only necessary to perform principal component extraction on the indexes in each index category.
Because the extraction of each index set has nothing to do with the year. Therefore, 30 samples of 5 years data from 6 countries were directly analyzed through principal component analysis. Applicability test of principal component analysis:

Let $r(X_i, X_j)$ be the correlation coefficient between the indicator $X_i$ and the indicator $X_j$, then

$$ r(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}[X_i] \text{var}[X_j]}} $$

Where $\text{Cov}(X_i, X_j)$ is the covariance of $X_i$ and $X_j$, $\text{Var}[X_i]$ is the variance of $X_i$, $\text{Var}[X_j]$ is the variance of $X_j$. We can judge whether the principal component analysis method is suitable according to the correlation coefficient of all indicator variables. If most of $r(X_i, X_j)$ are greater than or equal to 0.2, it indicates that the principal component analysis method is initially feasible.

Next "KMO" ("Kaiser-Meyer-Olkin") test:
The statistics of this test are used to compare simple correlation and partial correlation coefficients between variables. The KMO value is between 0 and 1, and the closer it is to 1, it indicates that the sum of squares of simple correlation coefficients among all variables is much larger than the sum of squares of partial correlation coefficients, and the more suitable it is for principal component analysis.

Among them, Kaiser gave a KMO test standard: $KMO>0.9$, very suitable; $0.8<KMO<0.9$, suitable; $0.7<KMO<0.8$, general; $0.6<KMO<0.7$, not very suitable; $KMO<0.6$, not suitable.

For this problem environment, the calculated KMO value is 0.561.

In conclusion, almost all the coefficients of correlation coefficient test are greater than 0.2, and the "KMO" test value is 0.561, which is suitable for principal component analysis.

Principal component extraction process of principal component analysis:

There are three methods to extract principal components:

First, according to the size of the eigenvalue, if the characteristic root of the principal component is greater than 1, it can be selected as the princess component.

Secondly, according to the cumulative contribution rate, if the cumulative variance contribution rate of the principal component is greater than 80%, the extraction result of the main component can be selected.

Third, according to the scattered distribution of the gravel graph, the principal component eigenvalues before the turning point in the graph are all large, which can explain the original variables more completely. Therefore, several principal components before the inflection point are selected as the principal component extraction results.

No variables are used to explain the total variance through principal component analysis, which is extracted according to the eigenvalue and the percentage of variance. The first principal component and the principal component of the variance percentage are 87.764%, 72.625%, 75.951%, respectively. Therefore, it can be thought of other than three variables, the influence of variance is very small and can be ignored. Therefore, the national higher education evaluation is divided into three main components, namely the past, present and future national higher education.

Principal component analysis of the calculation of scoring process:

Factor score coefficient and original variable data after realization of standardization are the basis to obtain the score of each component.

The formula can be expressed as:
\[ F_i = \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots + \beta_{in}X_n \] (3)

\( F_i \) (i = 1, 2, ..., m) is the score of factor Fi on variable \( X_p \).

The composite score of the principal component is obtained by multiplying the score of each component with the contribution rate of the principal component after rotation.

The specific expression is as follows:

\[
\begin{align*}
HE &= 0.091 \times F_1 + 0.438 \times F_2 + 0.374 \times F_3 - 0.065 \times F_4 - 0.187 \times F_5 \\
NE &= 0.468 \times F_6 - 0.323 \times F_7 + 0.434 \times F_8 + 0.017 \times F_9 \\
FE &= 0.092 \times F_{10} + 0.438 \times F_{11} + 0.374 \times F_{12}
\end{align*}
\] (4)

\( F_i \) (i = 1, 2, ..., m) is the score of each component, and \( \beta_i \) (i = 1, 2, ..., m) is the contribution rate of each component.

As can be seen from the above, through the principal component analysis, the national higher education indicators are divided into three categories, and the weight of each index in each component is different. In order to replace the secondary index with the first-level index constructed, the expression to calculate the first-level index should be established according to the above relationship.

### 2.4. Evaluation Model of Health System of National Higher Education

#### 2.4.1. Model Building

In this part, the evaluation system of health degree of higher education based on BP neural network is established according to the three first-level indexes extracted in the previous part. In the absence of data and training sets, we attempted to use standard data for training and to evaluate the health status of higher education in five countries.

In the evaluation, we classified the health of higher education in the United States as a 5 level and in South Africa as a 1 level, graded by level.

We analyzed the applicability of BP neural network and obtained the correlation of each index as follows.

<table>
<thead>
<tr>
<th>Numble</th>
<th>HE</th>
<th>NE</th>
<th>FE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>1.000</td>
<td>.213</td>
<td>-719</td>
</tr>
<tr>
<td>NE</td>
<td>.213</td>
<td>1.000</td>
<td>-216</td>
</tr>
<tr>
<td>FE</td>
<td>-719</td>
<td>-216</td>
<td>1.000</td>
</tr>
</tbody>
</table>

#### 2.4.2. Overview of BP Neural Network

BP neural network is a forward propagation signal and backward propagation error learning algorithm.

Generally, a neural network has a multi-layer structure, including an input layer, a hidden layer and an output layer. The more layers, the more accurate the calculation result, but the longer it takes. Layers should be designed according to requirements.

When the neural network is constructed, the transfer function and transformation function of its neurons have been determined. In the learning process, the transformation function cannot be changed. Therefore, if the output size of the network is to be changed, it can only be achieved by changing the input of the weighted sum.

Since the neuron can only process the input signal of the network accordingly, the weight parameters of the neuron can only be modified if the weighted input of the network is to be
changed, so the learning of the neural network is the process of changing the weight matrix. The structure diagram of BP neural network will be given below.

![Figure 2. Schematic diagram of BP neural network](image)

In Figure 2, $X_1, X_2, ..., X_n$ are the input values of the BP neural network, $Y_1, Y_2, ..., Y_m$ are the predicted values of the BP neural network, and $\omega_{ij}$ and $\omega_{jk}$ are the weights of the BP neural network.

As can be seen from Figure 2, BP neural network can be regarded as a nonlinear function, and the network input value and predicted value are independent variables and dependent variables of the function respectively. When the number of input nodes $n$ is and the number of output nodes $m$, the BP neural network expresses the functional mapping relationship from $n$ independent variables to $m$ variables.

The BP neural network must be trained before prediction, through which the network has associative memory and predictive ability. The training process of BP neural network includes the following steps.

Step 1: initialize network. According to the input and output sequence of the system, determine the number of input layer nodes, the number of hidden layer nodes, the number of output layer nodes, etc.

Step 2: Calculate hidden layer
According to the input vector $X$, the connection weight between the input layer and the hidden layer, and the hidden layer threshold $A$, the hidden layer output $H$ is calculated.

$$H_j = f \left( \sum_{i=1}^{n} \omega_{ij} X_i - a_j \right) \quad j = 1, 2, ..., l$$  \hspace{1cm} (5)

Where, $l$ is the number of nodes in the hidden layer; $f$ is the excitation function of hidden layer, which can be expressed in various forms. The function selected in this chapter is:

$$f(x) = \frac{1}{1+e^x}$$  \hspace{1cm} (6)

Step 3: Compute output layer.
According to the hidden layer output $H$, connection weight $\omega_{jk}$ and threshold $b$, BP neural network prediction output $O$ is calculated.
\[ O_k = \sum_{j=1}^{i} H_j \omega_{jk} - b_k \quad k = 1, 2, ..., m \] (7)

Step 4: Error calculation. According to the network forecast output O and the expected output Y, the network forecast error e is calculated.

\[ e_k = Y_k - O_k \quad k = 1, 2, ..., m \] (8)

Step 5: Update weights. Update the network connection weight based on the network prediction error.

\[ \omega_{ij} = \omega_{ij} + \eta H_j (1-H_j) x(i) \sum_{k=1}^{m} \omega_{jk} e_k \quad i = 1, 2, ..., n; j = 1, 2, ..., l \] (9)

\[ \omega_{jk} = \omega_{jk} + \eta H_j e_k \quad j = 1, 2, ..., l; k = 1, 2, ..., m \] (10)

Where, \( \eta \) is the learning rate.

Step 6: Threshold update. Update the network node threshold a, b through the network error e.

\[ a_j = a_j + \eta H_j (1-H_j) \sum_{k=1}^{m} \omega_{jk} e_k \quad j = 1, 2, ..., l \] (11)

\[ bk = bk + ek \quad k = 1, 2, ..., m \] (12)

Step 7: Determine the iteration. If the iteration is not over, return to step 2.

2.4.3. BP Model Establishment

The modeling of speech feature signal classification algorithm based on BP neural network includes three steps: BP neural network construction, BP neural network training and BP neural network evaluation. The algorithm flow is shown in Figure 2.

Figure 3. Algorithm flow chart
BP neural network to build according to system input and output data characteristics determine the structure of the BP neural network, due to the input signal for 3 d level indicators, for evaluation of the national higher education health there are five classes, so the structure of the BP neural network for the 3-5-1, namely the input layer has three nodes, hidden layer with 5 nodes, 1 output layer nodes.

When constructing the evaluation model of BP neural network, it is necessary to have a certain number of learning samples to establish the evaluation system.

In order to minimize the error of evaluation, this paper considers using behavioral anchoring quantization method to define different horizontal scales (as shown in the table below) and construct data sets that meet the conditions as training samples.

<table>
<thead>
<tr>
<th>Numble</th>
<th>HE</th>
<th>NE</th>
<th>FE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>HE ≥ 0.82</td>
<td>NE ≥ 0.72</td>
<td>-1.42 ≥ FE</td>
</tr>
<tr>
<td>4</td>
<td>0.82 ≥ HE ≥ 0.12</td>
<td>0.72 ≥ NE ≥ 0.02</td>
<td>-0.72 ≥ FE ≥ -1.42</td>
</tr>
<tr>
<td>3</td>
<td>0.12 ≥ HE ≥ -0.58</td>
<td>0.02 ≥ NE ≥ -0.68</td>
<td>-0.02 ≥ FE ≥ -0.72</td>
</tr>
<tr>
<td>2</td>
<td>-0.58 ≥ HE ≥ -1.28</td>
<td>-0.68 ≥ NE ≥ -1.38</td>
<td>0.68 ≥ FE ≥ -0.02</td>
</tr>
<tr>
<td>1</td>
<td>-1.28 ≥ HE</td>
<td>-1.38 ≥ NE</td>
<td>FE ≥ 0.68</td>
</tr>
</tbody>
</table>

Then insert data from the United States and South Africa into the neural network to verify whether it is consistent with the preset classification.

2.5. Improvement of GA on Neural Network

In the repeated tests of BP neural network, we find that the established BP neural network is not stable, and the convergence time is often too long, and the weight value and threshold value of each test differ greatly. This indicates that it is difficult for a simple BP neural network to achieve a good evaluation effect under this data. In fact, the determination of network weights and thresholds is always a difficulty in network training.

Randomly selected parameters will seriously affect the accuracy of the network. Therefore, it is considered to add genetic algorithm into the neural network, and use the optimal individuals trained by the genetic algorithm to optimize the weight and threshold of the network.

The specific steps are as follows:

Step 1: Individual coding and population initialization. The individual contains the ownership values and thresholds of the entire neural network. In this paper, individuals are coded by real number coding. The encoding length is:

\[ S = n \times m + m \times l + m + l \]  

Step 2: Determine the fitness function. Set the reciprocal of the BP neural network error sum of squares as the fitness function.

\[ f = \frac{1}{SE} \]  

Among them, is the sum of squared errors between the predicted output of the neural network and the expected output.

Step 3: Individual selection. Individual selection can be made according to the probability value, the formula is as follows:
\[ P_i = \frac{f_i}{\sum_{i=1}^{k} f_i} \]  

Among them: \( f_i \) is the fitness value of the individual \( i \); \( k \) is the number of individuals.

Step 4: Crossover and mutation. The optimal individual is not a crossover operation, but directly copied to the next generation. For other individuals, use the mutation probability \( p_m \) for mutation create another new individual. In this experiment, \( p_c = 0.4 \), \( p_m = 0.07 \), and the evolution algebra is 100.

Step 5: Loop the operation.

Training error of GA-BP neural network is shown in the figure below:

![Training error of GA-BP neural network](image)

Figure 4. GA-BP neural network
3. Results

Based on our model for evaluating the health status of national higher education, we can obtain specific ratings for six countries as follows:

<table>
<thead>
<tr>
<th>countries</th>
<th>US</th>
<th>Germany</th>
<th>Japan</th>
<th>Australia</th>
<th>South Africa</th>
<th>India</th>
</tr>
</thead>
<tbody>
<tr>
<td>level</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

3.1. Analysis of Higher Education System in Germany as An Example

Let’s take Germany as an example to analyze the German higher education system.

**Figure 5. Some indicators of Germany**

**Figure 6. Some indicators of Germany**
A separate analysis of Germany from 2013 to 2018 shows that Germany's education index, the number of Fields Medals, the number of universities, and the ratio of male to female enrollment are all on the rise. Among them, the number of university students in Germany has increased significantly, indicating that Germany attaches more importance to higher education. The ratio of men to women is close to 1, which shows that higher education in Germany is relatively fair. Men and women have the same opportunity to receive higher education, and there is no or less gender discrimination.

Similarly, there is a certain upward trend in the number of German universities, enrollment rates, and Nobel Prize winners. Among them, the Nobel Prize, the number of papers, and the number of universities in Germany are maintained at a high level.

As can be seen from our previous results, Germany’s higher education health rating is Level 4, and it has the potential to move towards Level 5 in the future.

Horizontal comparison, it is better than India, South Africa and other countries in all aspects, but there is still a gap with the United States.

Low enrolment rates and student numbers are Germany’s handicap, though cheap tuition is a big advantage.

Longitudinally, the number of students is growing fast, but the enrollment rate is growing very slowly. Compared with the enrollment rate of the United States and even Australia, which is over 100 percent, the enrollment rate of Germany is always at 70 percent, which is extremely low.

We are trying to further improve the German higher education system. If the German enrollment rate is adjusted, based on the time series analysis, it will be increased at an annual growth rate of 5%. Three years later, the future of German higher education The energy will be increased from 0.67 to 0.73, and then brought into the evaluation model, and the health of German higher education can reach level 5.

The targeted plan we proposed is only to modify and improve the data. In reality, if it is necessary to increase the enrollment rate by an annual increase of 5%, it will inevitably bring more influencing factors:
1. The tuition fees in Germany are relatively cheap in countries at the same level. The increase in enrollment rate means that the country needs to increase investment in education, and the government needs to pay additional education expenditures. A 5% increase in a year may affect Germany The economy has caused quite a shock.

2. The world ranking of German schools is inseparable from the selection of students. A substantial increase in enrollment may result in a decline in the quality of students, which will affect the future scores of the German higher education foundation, and also Make the ranking cause a certain degree of decline.

3. The infrastructure construction of German schools, the maximum number of students that the colleges can accommodate cannot meet the annual enrollment rate of 5%. If you need to greatly increase the enrollment rate, you need government help to open more colleges and universities for enrollment.

4. Under the influence of the epidemic, German schools are facing restrictions on closure and online teaching. Although it has facilitated the increase in the number of enrollment to a certain extent, it has also increased the difficulty of enrollment and the difficulty of opening a school.

4. **Strengths and Weaknesses**

4.1. **Strengths**

1. The models and data in this article look at the global higher education situation from both horizontal and vertical dimensions, selecting countries at different levels for analysis, and the selection of data and models is representative.

2. In terms of empirical models and empirical methods, this article has absorbed advanced models and methods in related research fields around the world, combined with US data, and conducted empirical tests on each part.

3. The combination of genetic algorithm and neural network provides high classification accuracy.

4.2. **Weaknesses**

1. Since the past year, the world has been affected by the new crown epidemic, and the economies of all countries are in a haze. The first promotion of online higher education and immature response methods have caused a lot of noise in last year's data. This noise will not disappear in the past two years. At the same time, the noise data is relatively small, it is difficult to use the noise data to predict future development, and the model has certain limitations.

2. Due to the lack of data on higher education in specific countries and specific majors, analysis and empirical research cannot be conducted for specific majors.

3. The error and convergence speed are very slow during network training. In the basic BP neural network, the learning rate is a fixed value, and in order to ensure the stability of the neural network, this value is generally relatively small, which causes the error and convergence speed of the neural network model to be very slow during the training process.

5. **Modification of the model**

In addition to index reasons, the model accuracy of the PCA-GA-BP higher education evaluation system we constructed is mainly due to the influence of BP neural network. In the construction of BP neural network, the number of hidden layer nodes in the neural network often has a relatively large impact on the prediction accuracy. Too many or too few nodes will lead to inaccurate predictions. So we can try to find the optimal number of hidden layer nodes. In addition, the BP neural network will also be affected by the learning rate. The greater the learning rate, the greater the weight modification, and the faster the network learning speed,
but an excessive learning rate will cause oscillations in the learning process, and the learning rate will be smaller. The slower the network convergence speed, the harder the weights tend to stabilize. Therefore, we try to replace it with a changing learning rate and optimize the learning rate to make the running result better.

Generalization of the model

The higher education health evaluation model established in this paper can be evaluated not only in higher education, but also in all walks of life, such as politics, economy, military, agriculture, etc.

Combined with the two evaluation systems, the evaluation results are more accurate and objective, and have more reference value.

6. Conclusion

To sum up, firstly, we use principal component analysis to determine the indicators, and secondly, we use GA-BP neural network to rate the status of higher education system in six countries, namely the United States, Germany, Japan, Australia, South Africa and India, and the results are 5, 4, 3, 2, 1, 1 respectively.

Then we analyze the higher education system in Germany.

Germany could strengthen its higher education system by increasing enrolment.

Finally, we analyze the advantages and disadvantages of the model.

References


Ma Fashun, Xing Anqi, Zhao Mengxue, Ge Yujie, Wang Mohan, Song Yuwei, Liang Xiuli.

[6] Principal component analysis and cluster analysis were used to determine the evaluation index of AA broiler productivity [J]. Animal Husbandry and Veterinary Medicine, 201, 53(02): 13-17.


