Management forecast based on big data fusion DEA and RBF algorithm

Yan Zhu *
Xi'an jiaotong university city college, 710018, Xi'an, China
*Corresponding author e-mail: jdesxy@xjtu.edu.cn

Abstract. With the rapid development of information technology, how to use big data for effective analysis has become the focus of various industries. Big data will inevitably bring data noise and data redundancy. The traditional big data preprocessing methods do not take into account the functional relationship between variables, and the commonly used big data modeling tools cannot effectively model big data with complex nonlinear relationships. This paper puts forward a prediction method combining DEA and RBF, uses DEA to preprocess big data to select the effective data, then uses RBF to model the data, and compares the prediction accuracy with the original environmental management fund data without pre-processing. The results show that the DEA-RBF method has higher prediction efficiency and better prediction accuracy.

Keywords: big data; Radial basis function (RBF); Data Envelopment Analysis (DEA); Environmental Protection Management Fund.

1. Introduction
With the development of information technology, data is filled with every corner. Big data has become a research hotspot in various industries. As big data has four characteristics: massive data scale, rapid data flow, various data types, and low-value density, how to deal with big data effectively poses new challenges to various industries. Big data's source, quantity, different structural forms, real-time, and other diversified characteristics make it cover a high value, but its value density is very low, often only hundreds or even dozens of pieces of data are valuable in thousands of data. Thus it can be seen that there may be more redundant data and noise data in big data, which will have an inestimable negative impact on data analysis. At present, effective methods to overcome these negative effects include data cleaning algorithms (such as trajectory cleaning algorithm) and data reduction algorithms (such as Pearson correlation coefficient, PPMCC). These algorithms only unilaterally eliminate outliers or measure the correlation between two variables X and Y, but fail to consider the functional relationship between multiple variables while eliminating outliers and redundant values. Big data set contains not only the data value itself but also its function relationship. If the function relationship is not considered, the analysis result may have deviated. Therefore, it is very necessary to consider the function relationship in the data set. Data Envelopment Analysis ((DEA),) can effectively deal with the deviation caused by the functional relationship between variables. When DEA evaluates the efficiency of the decision-making unit (DMU), it is not necessary to predict the functional relationship between input and output variables and set weights in advance. Through the obtained weights, we can understand the
functional relationship between variables, eliminate outliers and redundant values, and reduce the number of data without changing the quality of data. It is an effective way of data preprocessing that can be applied to machine learning. At present, some scholars have initially used DEA to deal with and analyze big data. [1]

The radial basis function (RBF) has simple learning rules and a powerful memory function. It is a kind of function model with good nonlinear approximation. It can deal with big data with the complex variable relationship. Its output is independent of initial weight and can overcome the problems existing in traditional prediction models. [2] Some studies aimed at the actual sample big data, using RBF method in project evaluation, geology, power industry and other fields with the development of information technology, data-filled with every corner, big data in various industries has become a research hotspot. As big data has four characteristics: massive data scale, rapid data flow, various data types, and low-value density, how to deal with big data effectively poses new challenges to various industries. Big data's source, quantity, different structural forms, real-time, and other diversified characteristics make it cover a high value, but its value density is very low, often only hundreds or even dozens of pieces of data are valuable in thousands of data. Thus it can be seen that there may be more redundant data and noise data in big data, which will have an inestimable negative impact on data analysis. At present, effective methods to overcome these negative effects include data cleaning algorithms (such as trajectory cleaning algorithm) and data reduction algorithms (such as Pearson correlation coefficient, PPMCC). These algorithms only unilaterally eliminate outliers or measure the correlation between two variables X and Y, but fail to consider the functional relationship between multiple variables while eliminating outliers and redundant values. Big data set contains not only the data value itself but also its function relationship. If the function relationship is not considered, the analysis result may have deviated. Therefore, it is very necessary to consider the function relationship in the data set. Data Envelopment Analysis ((DEA),) can effectively deal with the deviation caused by the functional relationship between variables. When DEA evaluates the efficiency of decision-making units (DMU), it is not necessary to predict the functional relationship between input and output variables and set weights in advance. [3] Through the obtained weights, we can understand the functional relationship between variables, eliminate outliers and redundant values, and reduce the number of data without changing the quality of data. It is an effective way of data preprocessing that can be applied to machine learning.

2. DEA-RBF method

The datasets used in big data's background are usually large and complex, so this work involves many steps. First of all, we use DEA to preprocess the data, reduce the number of data on the premise of ensuring the quality of the data, and maintain the monotonicity and universality of the data set. Secondly, RBF is used to train and test the processed data to judge the effectiveness of the RBF prediction model. Finally, the effectiveness of the hybrid method (DEA-RBF) is analyzed. The DEA-RBF method proposed in this paper can effectively deal with complex data sets and improve the prediction accuracy of the model. [4,5]

2.1. Data preprocessing based on DEA

In the process of data preprocessing, the variable index should be selected first. In this paper, we use the method of literature for reference to select the relevant variables and determine the input and output variables. According to the input and output data provided, the relative efficiency of DMU is calculated and ranked by the DEA model, and the incomplete or redundant data are deleted. This paper selects the classical CCR model to obtain the efficiency value of DMU.

Suppose there are Do the same type of DMU, and each DMU contains P input indicators and S output indicators, and the traditional CCR model is used to calculate their DEA efficiency. Then the input index of the unit can be expressed by matrix \( Q = q_{wt} \) \((t = 1,2,...; P; d = 1,2,...,D)\), \( q_{wt}\) Then the input index of the unit can be expressed by matrix \( d \) th DMU. The same matrix \( Z = z_{rd} \) \((r = 1,2,...; S; d = 1,2,...,D)\) represents the output index data. Per the regulations. The
efficiency $DMU_n$ is expressed by the linear weighted combination of output indicators and input indicators, and the formula is as follows:

$$\theta_n = \frac{\sum_{i=1}^{s} \mu_i z_{r_{io}}}{\sum_{i=1}^{p} \beta_i q_{t_{io}}}$$  \hspace{1cm} (1)$$

Among them, $\theta_n$ indexes are called the efficiency evaluation index of $DMU_n$ while $\beta$ indexes represent the input index weight vector and $\mu_i$ indicator represents the output index weight vector.

The traditional $CCR$ model calculates the following questions for all $DMU_n$:

$$\text{Max} \theta_n = \frac{\sum_{i=1}^{s} \mu_i z_{r_{io}}}{\sum_{i=1}^{p} \beta_i x_{t_{io}}}$$  \hspace{1cm} (2)$$

DEA identifies the frontier of the best DMU and measures the relative efficiency of other DMUs according to the distance to the frontier. All the efficiency values obtained by DMU are less than or equal to 1, and the efficiency value of DMU on the best front is 1. The efficiency value calculated by the traditional CCR model is to compare the existing DMU with the effective DMU on the frontier, so it can effectively distinguish the invalid DMU. In this paper, all DMU, with efficiency values of 1 is selected as the most effective layer for further training and testing.

2.2. Construction of RBF model

The dataset preprocessed by DEA is used to build the RBF model. RBF randomly selects an appropriate amount of data to learn, and the rest of the data is used to test the accuracy of the RBF model. If the accuracy does not reach the target, it can be optimized by adjusting the model parameters. RBF is an effective method for nonlinear modeling with high learning efficiency and can deal with data sets with unknown relationships between input and output variables. The algorithm is essentially an interpolation process, which can deal with a large number of discrete data and construct functions through discrete sampling points to predict unknown points. [6]

Suppose there are $k$ input variables and $n$ output variables. Then the sample data set can be represented as the collection $\{X_{km}, Y_{kn}\}$, The expression of RBF radial basis function model is:

$$f(x) = \sum_{i=1}^{k} \omega_i \beta_i (x)$$  \hspace{1cm} (3)$$

And satisfy the interpolation condition:

$$f(x) = F(x)$$  \hspace{1cm} (4)$$

Among them, $h$ represents the number of interpolation nodes, $\omega_i$ nodes, the interpolation condition is solved by the matrix to get; $\beta_i (x)$ nodes as the basis function, the Gaussian function is often selected, and its expression is as follows:

$$\beta_i (x) = \exp \left( -\frac{\|x - c_i\|^2}{2\sigma_i^2} \right)$$  \hspace{1cm} (5)$$

In formula (4), $c_i$ represents the center vector of the $i$th hid node; $\sigma_i^2$ represents the variance or width of the radial basis function and is used to adjust the sensitivity of the model.
\[ G \times \omega_i = F(x_i) \]  

As a result, the only solution can be obtained:

\[ \hat{\omega}_i = G^{-1} \times F(x_i) \]  

RBF determines the relevant weights through the training data, and optimizes the regression prediction of the data by adjusting the weights; applies a group of test data to the adjusted model to get its predicted value; uses the error between the real value and the predicted value to judge whether the accuracy of the model reaches the standard or not. If the accuracy is not up to standard, the model can be modified by constantly adjusting the parameters.

2.3. DEA-RBF method

DEA-RBF method combines the advantages of both DEA and RBF and can deal with complex data sets of unknown variable attributes. In general, the size of the amount of data will have an impact on the training time of RBF. With the increase of the amount of data, the storage cost and operation cost of data increase geometrically, and too large an amount of data will even cause the system to crash. Therefore, in the face of big data, to reduce the training time and improve the running efficiency of RBF, the data set should be reduced to a certain extent. However, the reduction of data sets may lead to changes in the accuracy of RBF training. The traditional data sampling method is only simple sampling, does not take into account the relationship between variables in the data set, and can not guarantee that the filtered data set can retain the characteristics of the original data. This may harm the experiments using this data set for prediction. DEA can eliminate some useless data and interference data without changing the universality of the data, to filter out the most effective layer of data in the data set. As there are certain differences in the size of the data represented by different variables, to prevent data distortion and small data from being swallowed by big data, this paper normalizes the selected effective data, which can achieve a better prediction effect. Normalization formula (8):

\[ G = \frac{\alpha - \alpha_{\text{min}}}{\alpha_{\text{max}} - \alpha_{\text{min}}} \]  

Where \( \alpha \) represents the original data, \( \alpha_{\text{min}} \) and \( \alpha_{\text{max}} \) are the maximum and minimum values in the original data, respectively. Then the filtered data is divided into training data and test data, and the effectiveness of the hybrid method is verified by learning and testing the multi-output RBF model. The DEA-RBF model proposed in this paper can not only reduce the amount of data, but also improve the accuracy of model prediction, greatly shorten the training time, and improve the effectiveness of data processing.

3. Experimental test and analysis

3.1. Experimental design

The data set selected in this paper comes from the National Bureau of Statistics of China (http://www.stats.gov.cn/), which is used to invest in environmental protection funds from 2000 to 2020. The revised data include four input variables and two output variables. As shown in Table 1. Before using the DEA method, we first delete the incomplete data in the sample data set and then proceed to the next step.

| Table 1. Input variables and output variables |
|---------------------------------------------|
| **Input variable**                         | **Variable name** | **Numbering** |
| GDP                                         | X1               |
| Number of environmental protection companies | X2               |
| Total rural population                      | X3               |
| Total urban population                      | X4               |
| Output variable                             |                   |
| Town funds for environmental protection     | Y1               |
| Funds for rural environmental protection    | Y2               |
The implementation environment of this lab is MATLAB2017. First of all, 200 original records are preprocessed by the DEA model, and finally, 66 DMU with efficiency values of 1 are obtained. Then these screened data with the highest efficiency are normalized to eliminate the influence of small data on big data, which is used in RBF model modeling. Then 10 pieces of data are randomly selected as training data, and the other 10 data are used as test data. The prediction accuracy of the RBF model is obtained by comparing the predicted value with the real value.

3.2. Experiment analysis

According to the above experimental process, the predicted values of the output variables $Y_1$ and $Y_2$ in the test data set are obtained and compared with the real values of the variables.

According to the prediction result diagram of figure 1 and figure 2, it can be seen that except for individual points, there is little difference between the predicted value and the real value obtained by the DEA-RBF method, and the prediction result is better. To compare the results more accurately, Table 2 shows the error comparison between the predicted and real values of the output of 26 test units. The formula for calculating the error rate is:
\[
\text{Error} = \left| \frac{\text{Actual} - \text{Predictive}}{\text{Actual}} \right| \quad (9)
\]

Table 2. Error rate between Y1 and Y2 predicted value and true value (unit: %)

| Year | Errorrate | Averageerror |
|------|-----------|--------------|
|      | \( Y_1 \) | \( Y_2 \)    |                |
| 2000 | 5.51      | 3.13         | 4.32           |
| 2001 | 33.04     | 30.13        | 31.585         |
| 2002 | 9.28      | 7.88         | 8.58           |
| 2003 | 6.72      | 0.48         | 3.6            |
| 2004 | 2.52      | 1.12         | 1.82           |
| 2005 | 4.1       | 0.82         | 2.46           |
| 2006 | 5.87      | 2.83         | 4.35           |
| 2007 | 0.65      | 6.45         | 3.55           |
| 2008 | 11.6      | 3.25         | 7.425          |
| 2009 | 12.6      | 1.75         | 7.175          |
| 2010 | 2.31      | 14.78        | 8.545          |
| 2011 | 3.35      | 18.81        | 11.08          |
| 2012 | 3.95      | 0.28         | 2.115          |
| 2013 | 1.62      | 2.26         | 1.94           |
| 2014 | 0.68      | 0.25         | 0.465          |
| 2015 | 8.18      | 2.54         | 5.36           |
| 2016 | 6.63      | 4.29         | 5.46           |
| 2017 | 0.86      | 2.94         | 1.9            |
| 2018 | 1.02      | 0.25         | 0.635          |
| 2019 | 5.45      | 3.46         | 4.455          |
| 2020 | 6.92      | 7.82         | 7.37           |

The second column and the third column in Table 2 are the error rates of the output variables Y1 and Y2, respectively, and the fourth column is the average error rate of each sample. In 26 test units, except for a few points, the prediction accuracy of each unit is basically within the acceptable range, the total error is less than 7%, and the fitting effect is good. Among them, the two points numbered 2 and 18 have a large error because they are on the edge. Generally speaking, RBF does not guarantee that the prediction accuracy of all points is 100%. There are usually some errors in the edge areas, and the points with large errors can be improved, but it may take dozens of times to reach the acceptable range at the expense of RBF learning time.

To facilitate the comparison, this paper also uses RBF for regression prediction of the original data without DEA preprocessing, and the prediction error results are shown in Table 3.

Table 3. The error between the predicted value and the actual value of the original data (unit: %)

| Output | Maximum error | Minimum error | Average error | Total error |
|--------|---------------|---------------|---------------|-------------|
| \( Y_1 \) | 212.59        | 0.02          | 11.6          | 12.32       |
| \( Y_2 \) | 230.77        | 0.27          | 13.03         |             |

As can be seen from Table 3, the prediction effect of RBF on the data not preprocessed by DEA is not ideal. Although the average error is close to 10%, the gap between the maximum error and the minimum error is too large, and the number of samples with an overall prediction error rate of more than 30% is relatively large, and the predicted results are unstable and unreliable. Generally speaking, with
the reduction of the amount of data, the prediction accuracy of RBF will decrease to a certain extent, this is because the learning of RBF needs enough data to ensure high prediction accuracy. However, in this paper, DEA removes the outliers that will harm the prediction accuracy of RBF, thus improving the prediction accuracy when the amount of data is reduced.

4. Conclusion
The DEA-RBF prediction model proposed in this paper combines the advantages of DEA preprocessing and RBF regression prediction. In the process of processing big data, the learning and training time is reduced, the prediction accuracy of the model is improved, and the prediction effect is better. The results of theoretical analysis and empirical analysis show that the prediction model based on DEA and RBF method can greatly shorten the processing time of big data while ensuring good prediction accuracy, which has important theoretical and practical significance to improve the processing efficiency of big data. To provide a relevant basis for predicting the data of environmental protection funds.

References
[1] Schuessler N, Axhausen K W. Processing Raw Data From Global Positioning Systems Without Additional Information [J]. Transportation Research Record, 2009, 2105(1).
[2] Pendharkar P C. A Data Envelopment Analysis-based Approach for Data Preprocessing [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(10).
[3] Kotsiantis S B, Kanellopoulos D, Pintelas P E. Data Preprocessing for Supervised Learning [J]. International Journal of Computer Science, 2006, 1(2).
[4] Chen I F. Evaluate the Performance of Cardholders' Repayment Behaviors Using Artificial Neural Networks and Data Envelopment Analysis [C]. The 6th International Conference on Networked Computing and Advanced Information Management, 2010.
[5] Misiunas N, Oztekin A, Chen Y, et al. DEANN: A Healthcare Analytic Methodology of Data Envelopment Analysis and Artificial Neural Networks for the Prediction of Organrecipient Functional Status [J]. Omega, 2016, (58).
[5] Zhou Xiaoqian, Chen Enene. Forecasting method based on big data fusion DEA and RBF[J]. Statistics and Decision, 2020, v.36; No.562(22):38-41.