The Evaluation on Balance of Economic Benefit and Water Pollution of Enterprise Based on Kmeans-RBF Neural Network

Minna Chen[1,a], Tiehan Zhu[2,b*], Jinlan Guan[3,c]

{719778750@qq.com[a], zhuth@126.com[b], 1036905838@qq.com[c]}

Guangdong Polytechnic of Environmental Protection Engineering, Foshan, Guangdong, China[1]
Guangdong Industry Polytechnic, Guangzhou, Guangdong, China[2]
Guangdong Agricultural Industry Business Polytechnic, Guangzhou, Guangdong, China[3]

Abstract. According to the requirements of the Ministry of Ecology and Environment of the People's Republic of China and the evaluation of excellent enterprises, a new evaluation index system for the balance between economic benefits and water pollution protection of enterprises is established. The new indicator system established based on the principles of purposiveness, scientificity, systematicity, and operability has been demonstrated by experts in environmental protection. The index system includes 14 indicators, including Total industrial output value, Annual normal production time, Number of wastewater treatment facilities, Treatment capacity of wastewater treatment facilities, Investment cost of enterprise sewage treatment facilities, Discharge quantity of industrial wastewater, Chemical oxygen demand, Ammonia nitrogen from wastewater, Arsenic effluent from wastewater, Plum-bum from wastewater, Cadmium discharge from wastewater, Mercury emissions from wastewater, Total chromium discharge from wastewater, Discharge of hexavalent chromium in wastewater. At the same time, 241 enterprises were selected as the research objects through a focused investigation, and the Kmeans-RBF neural network was constructed using R software. The result matrix of the neural network model was constructed and the results were visualized. The results showed that the classifier had strong stability and effectively classified enterprises into three categories: well balanced enterprises, excellent balanced enterprises, and average balanced enterprises.

Keywords: water pollution index system, balance degree, Kmeans-RBF neural network, classifier, R software

1. Introduction

Any substance in nature must be governed by a certain equilibrium relationship during its movement and change, and its equilibrium principle can be applied to various fields and scenarios. In environmental and ecological protection research, the principle of balance can be used to measure different economic entities, investment in environmental protection equipment, environmental pollution situation, resource allocation, and other aspects, in order to achieve the best results. At present, there are two kinds of theories on environmental pollution assessment: First, using traditional assessment models such as data envelopment, such as "Evaluating the efficiency of green economic production and Environmental pollution

IEDM 2024, February 23-25, Kuala Lumpur, Malaysia

DOI 10.4108/eai.23-2-2024.2345905

control in China"[1] by Cao Yuequn et al; "Environmental performance evaluation of heavy polluting enterprises"[2] by Yongxin Gao; "Appraisal of Environmental, ecological and carcinogenic risk due to heavy metals in a sewage and solid waste contaminated area"[3]by Das Shreya. Second, environmental pollution assessment, such as "A review on radionuclide pollution in global soils with environmental and health hazards evaluation"[4]by Chandra Krishno, "Groundwater contamination assessment in Tarantula City, Mongolia with combined use of hydro chemical, environmental isotopic, and statistical approaches"[5]by Bayartungalag Batsaikhan. Starting from the balance between the economic benefits of enterprises and the environmental protection of pollution sources, the paper introduces the evaluation method of artificial intelligence, and establishes a grade evaluation model of economic benefits and environmental protection balance, which takes the economic development of enterprises as the goal and takes into account the responsibility of environmental protection. At present, there is no similar research, which is more conducive to the sustainable development of society.

2. Research Foundation

At present, there are various methods for evaluating water pollution, including Nemerow Index method, Grey clustering method, Fuzzy comprehensive evaluation method, etc. [6], and neural networks are gradually being applied in the field of environmental science. Back propagation neural network (BP) and radial basis function neural network (RBF) are two important branches of neural networks. Both are nonlinear multilayer feedforward networks, which have the ability to approach any nonlinear continuous mapping and have wide applications in modeling and control of nonlinear systems[7]. However, BP networks have some drawbacks, mainly slow convergence speed and often converge to local minima.The number of hidden layer nodes in the network depends on experience and experimental data, and its numerical stability is poor, making it difficult to obtain the optimal network[8].RBF networks have the best approximation performance and global optimal characteristics, with only one hidden layer, simple structure, and fast training speed, which can effectively overcome the shortcomings of BP networks.

RBF network is a single hidden layer feedforward neural network that uses radial basis function (usually Gaussian kernel function) as the activation function of hidden layer neurons, mapping each sample point to an infinite dimensional feature space, so that the originally linear and indivisible data is linearly separable. The determination of the center point of the radial basis hidden layer neuron can most affect the effectiveness of the neural network [9][10]. At present, the common methods for determining the center are direct calculation (random selection of RBF centers), self-organizing learning (such as nearest neighbor clustering, K-MEANS clustering, LMS), supervised selection of RBF centers (gradient descent method), orthogonal least squares method for selecting RBF centers[11], etc. RBF networks have multiple learning methods based on different methods for selecting radial basis function centers. The traditional selection of the center point often uses the input sample as the data center, which can easily lead to a sharp increase in the number of data centers and the training time with the increase of the sample size. However, using kmeans algorithm to select the RBF neural network data center and using Euclidean distance to measure the center point has higher efficiency and accuracy. And due to the highly complex nonlinear mapping relationship between various environmental pollution indicators, as well as between enterprise economic

benefits and different pollution indicators, the sample data of each enterprise is distributed without obvious representativeness. The proposed Kmeans based RBF algorithm for evaluating the balance between enterprise economic benefits and environmental protection is a fast and effective method.

3.Principle and steps of Kmeans-RBF neural network

3.1 Principle of Kmeans-RBF neural network

RBF neural network is a forward neural network using radial basis function as activation function first proposed by D.S.Broomhead and Lowe. Through proper training, it can realize data classification and function approximation [12]. The network includes an input layer, a hidden layer and an output layer, which simulates the neural network structure of local adjustment, mutual coverage and acceptance in human brain. The radial basis function constitutes the nonlinear activation function $\varphi_j(x)$ of hidden layer neurons. Each hidden layer node contains a central vector c , c and x with the same dimension as the input vector, w_j is the connection weight between the hidden layer and the output layer. The output of the network is realized by the following weighting function:

$$y(x) = \sum_{j=1}^{m} w_j \varphi_j(x) \tag{1}$$

The selection of the center vector c can be selected randomly, or by self-organizing learning methods such as nearest neighbor clustering and K-means clustering, or by supervised learning.

3.1.1 Radial basis function

Radial basis function is a common function form, which is widely used in machine learning, data mining, signal processing and other fields. The essence of radial basis function is a distance based function. Its value only depends on the real value function of the distance from the origin, that is $\varphi(x) = \varphi(\|x\|)$, any function that meets this characteristic $\varphi(x)$ is called radial basis function, which is usually defined as the Euclidean distance between any point x in space and a center c, that is $r = \|x - c_i\|$ (c_i the center point). Common radial basis functions include Gauss function: $\varphi(r) = \exp(-\frac{r^2}{2\sigma^2})$, σ is called the expansion function (variance) of radial basis function. It reflects the width of the function image, As σ becomes smaller and narrower in width, the function becomes more selective.

3.1.2 Radial basis function neural network structure

The RBF network is a three-layer forward network[13].The first layer is the input layer composed of signal source nodes. The second layer is the hidden layer, Its transformation function is a radial basis function of non negative nonlinear function(RBF). The number of hidden units is selected according to the needs of specific problems. The third layer is the

output layer, which is the response to the input mode and the linear combination of the second layer neuron output. The network structure is shown in Fig.1.

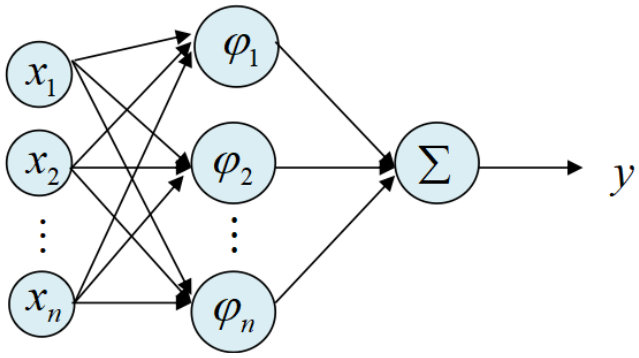

Fig.1 The RBF network structure diagram

where X is the input vector, x_i is the *i-th* factor (feature), and the neuron node where X is located is the input layer; φ is the Gaussian radial basis function, the neuron node where it is located is the hidden layer, and φ_j is the Gaussian radial basis function of the *j-th* neuron;Σ is the sum function, The neuron node of the Σ is the output layer; y is the output of the network.

3.1.3 K-means clustering algorithm

K-means clustering was first proposed by James Macqueen in 1967 as an iterative unsupervised learning clustering algorithm. Compared with hierarchical clustering, density clustering and other clustering methods, it has strong advantages in selecting the center point. The specific process is as follows [14]:

① Randomly select k data points from the data set as centroid (k can be recommended by using 26 judgment criteria provided by Nbclust package of R software);

② For each data point in the dataset, calculate the distance from the k centroids (generally select the Euclidean distance);

③ In the divided k categories, the center point of the data point belonging to each category is calculated as the new k centroids;

④ Calculate the distance between the new centroid point and the old centroid point, denoted as R. If R is less than a set threshold, it means that the recalculated centroid position does not change much, and the model tends to stabilize. The results show that our clustering has achieved the expected results, and the algorithm terminates;

⑤ If R changes significantly, steps②-④ need to be iterated.

3.2 Step of Kmeans-RBF neural network

The Kmeans-RBF neural network used in this paper is composed of three layers of neurons: input layer, hidden layer and output layer. The adaptive moment is used to estimate the Adam

optimization function, the mean square error (MSE) loss function, and L2 loss is added to prevent over fitting. At the same time, Kmeans-RBF neural network has good ability of learning nonlinear mapping, so when using this method to analyze the enterprise water pollution data, the traditional fitting modeling can be combined with the enterprise economic benefit data to transform into the fitting modeling considering the enterprise scale, so as to avoid the disadvantage of ignoring the characteristics of the data itself in pure mathematical fitting.

Specifically, when using the method proposed in this paper for analysis, firstly, the training samples composed of enterprise economic and water pollution detection data are input into the Kmeans-RBF neural network, and the radial basis function of the neural network is initialized by using the Kmeans clustering algorithm. After the data is propagated forward through the hidden layer and the output layer, the Adam optimization function [15][16] is used to reverse optimize the model according to the loss function, and after several rounds of supervised learning and training, Get the trained neural network model of enterprise water pollution balance evaluation, then use the test sample data composed of enterprise economic data and water pollution data of the test point to verify the accuracy of the neural network. Finally, the neural network model is used to build the regional enterprise water pollution balance evaluation data, and the accuracy information of the model is evaluated according to the test data.the specific process is shown in Fig. 2.

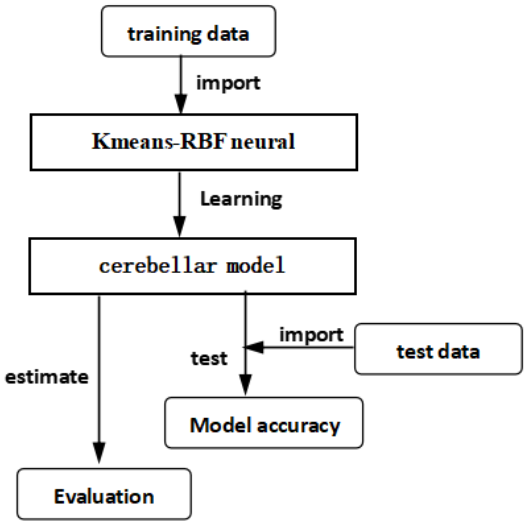

Fig.2 Schematic representation of the Kmeans-RBF neural network flow

4. Model simulation process

4.1 Modeling index setting and data preprocessing

To determine the balance index between the economic benefits of enterprises and water pollution protection, this paper selects the total industrial output value, annual normal production time, the number of wastewater treatment facilities, the treatment capacity of

wastewater treatment facilities, the operation cost of wastewater treatment facilities, the amount of industrial wastewater treatment, the discharge of chemical oxygen demand The evaluation index system is composed of 14 indicators such as ammonia nitrogen emissions and arsenic emissions from wastewater [17][18][19][20][21]. The balance evaluation index is shown in Table 1.

Table 1 Evaluation index of enterprise economic benefit and water pollution protection balance

Index and unit	Abbreviation	Index and unit	Abbreviation
Gross industrial output value (ten thousand yuan)	GDP	Ammonia nitrogen from wastewater (ton)	NH3
Annual production period (hours)	Time	Arsenic effluent from wastewater (kg)	As
Number of wastewater treatment facilities (sets)	Facility	Plum-bum from wastewater (kg)	Pb
Treatment capacity of wastewater treatment facilities (ton / day)	Capacity	Cadmium discharge from wastewater (kg)	Cd
Investment cost of enterprise sewage treatment facilities (ten thousand yuan)	Cost	Mercury emissions from wastewater (kg)	Hg
Discharge quantity of industrial wastewater (ton)	Quantity	Total chromium discharge from wastewater (kg)	Cr
Chemical oxygen demand (ton)	COD	Discharge of hexavalent chromium in wastewater (kg)	Cr6+

The experimental area is located in Nanhai District, Foshan City, Guangdong Province. According to the preliminary investigation, 241 enterprises whose industrial wastewater is directly discharged into the environment are selected as the experimental objects. Through the statistical questionnaire, on-site sampling and the results of our school's sewage testing laboratory, the data of each enterprise's total industrial output value, annual production period, the number of wastewater treatment facilities, the treatment capacity of wastewater treatment facilities, the investment cost of enterprise's sewage treatment facilities, the discharge quantity of industrial wastewater, chemical oxygen demand, etc. are obtained.(Table 2)

4.2 Center selection of K-means clustering algorithm

In this paper, R language is selected as the experimental environment, and the number of clustering clusters is determined by using wssplot( ) and NbClust( ) in R[12] (Fig. 3), and then the final clustering scheme is obtained by kmeans( ) and the clustering centroid is output. Because the output cluster centroid is based on the standardized data, the aggregate function and the members of the cluster are used to obtain the variable mean of each cluster in the original data (Table 3).

Table 2 List of enterprises

	GDP	Time	Facility	Capacity	Cost	Quantity	COD
Max	2628728.0	8664.0	4.0	11000.0	360.0	1808072.0	55.3
Min	20.2	300.0	0.0	0.0	0.0	0.0	0.0
Averag	41773.0	3436.8	1.1	515.6	27.6	85836.4	5.8

e							
SD	230541.3	2013.0	0.5	1148.9	53.4	181775.8	10.1
	NH3	As	Pb	Cd	Hg	Cr	Cr6+
Max	18.7	5.0	9.4	1.9	0.0	62.1	58.3
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Averag e	0.7	0.1	0.1	0.0	0.0	1.4	0.5
SD	2.0	0.5	0.8	0.2	0.0	5.9	4.0

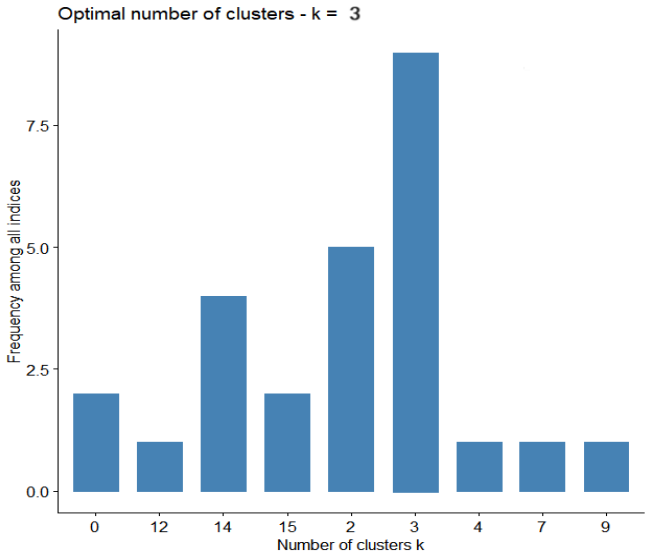

Fig. 3 Yields the recommended number of clustered clusters obtained using the 26 judgment criteria provided by the NbClust package

Table 3 Cluster centroids of the three types of clusters

	GDP	Time	Facility	Capacity	Cost	Quantity	COD
1	174286.33	5402.32	1.68	2573.82	134.00	367252.84	19.029
2	27266.95	3218.97	1.06	284.98	15.57	49536.46	4.206
3	288788.60	7680.00	1.00	5500.00	301.70	1808072.00	54.636
	NH3	As	Pb	Cd	Hg	Cr	Cr6+
1	3.378	0.1393	1.0466	0.2120	2.2727e-04	4.2294	1.2302
2	0.426	0.0716	0.0370	0.0080	1.8348e-05	0.8546	0.1570
3	5.696	0.8025	0	0	0	62.1331	58.2898

4.3 Kmeans-RBF model establishment and performance evaluation

4.3.1 Kmeans-RBF model establishment

The Kmeans-RBF neural network generates three RFB curves according to the three centroids obtained by Kmeans clustering, and the superposition of the three RBF curves is a curve that can smoothly fit the original data points. The radial basis function neural network mainly uses the characteristic (locality) that the RFB function only affects the adjacent area of the center, so as to use a radial basis function to make up for each local area, and finally achieve global fitting.

The data set of this model contains the statistical values of 241 enterprises, including 60 enterprises with good balance between economic benefits and water pollution protection, 13 enterprises with excellent balance and 168 enterprises with average balance. There are 16 variables in the data set, including the name of the enterprise, the total industrial output value, the annual normal production time, the number of wastewater treatment facilities, the treatment capacity of wastewater treatment facilities, the operation cost of wastewater treatment facilities, the treatment capacity of industrial wastewater, the discharge of chemical oxygen demand, the discharge of ammonia nitrogen, the discharge of arsenic in wastewater, the discharge of lead in wastewater, the discharge of cadmium in wastewater, the discharge of mercury in wastewater, the discharge of total chromium in wastewater, the discharge of hexavalent chromium in wastewater, and the category. The first variable enterprise name is not included in the data analysis, and the last variable (category) is the output variable (coded as good=versioncolor, excellent=virginica, general=setosa).

First, the data set is randomly divided into training set and test set, and the training set contains 169 observations (70%); The test set contains 72 observations (30%); Second, the neuralnet function of the neural network is used to construct the result matrix of the neural network model and visualize the results (Fig.4).The model expression from the output is:

$$y(x) = -2.05 * \exp(-[38.52 * dist(x, 0.44)]^2) + 0.46 * \exp(-[38.52 * dist(x, -0.35)]^2) + 2.31 * \exp(-[38.52 * dist(x, -0.85)]^2)$$

Then the 16 weights involved in the model were generalized, and the results showed that all the generalized values were close to 0, indicating that the covariates had little effect on the classification results; Finally, Generate the relevant prediction probability matrix and call the fusion matrix function to predict performance, and the results show that the stability of the model is 98%.

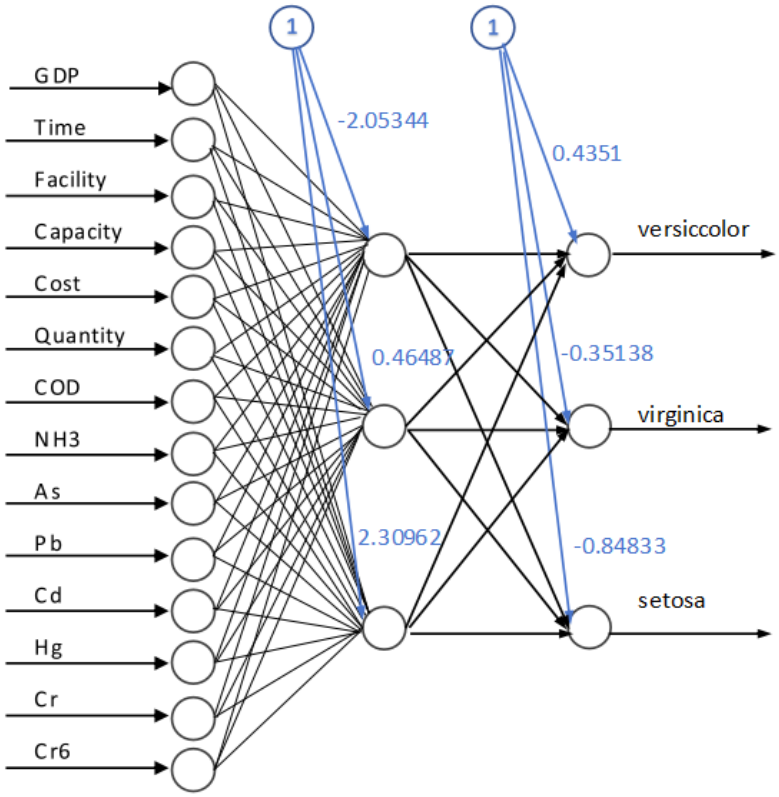

Fig.4 The neural network model

4.3.2 Performance evaluation

The training methods of RBF neural network are also different according to the selection method of RBF central point. The common methods to determine the central point are random selection of RBF center, self-organizing learning selection of RBF center, OLS method selection of RBF center, ESA algorithm selection of RBF center, K-means algorithm selection of center, etc. The following experiment compares three different RBF neural networks: Gradient RBF neural network, OLS-RBF neural network and Kmeans-RBF neural network. The results obtained from the approximation of sine function by these three RBF neural networks in the same input sample data are shown in Fig. 5-fig. 10. From Fig. 5 to Fig. 7(Input x:Sampling points;Input y:Sin x),it can be seen that the three RBF neural networks can better approximate sine functions. Figure 8-10 shows the convergence of three kinds of network training.

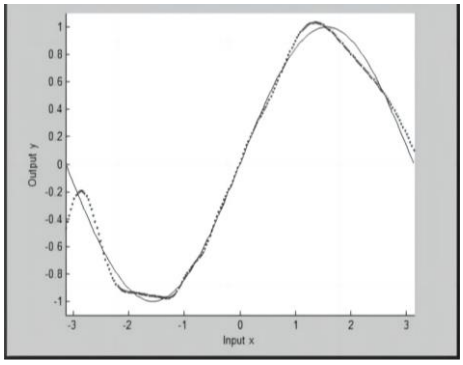

Fig.5 Kmeans-RBF neural network

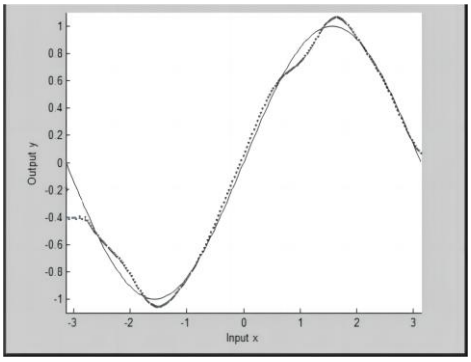

Fig.6 Gradient RBF neural network

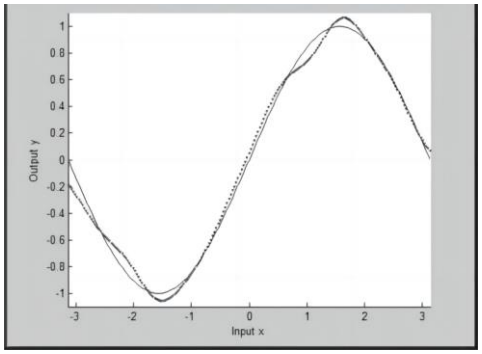

Fig.7 OLS-RBF neural network

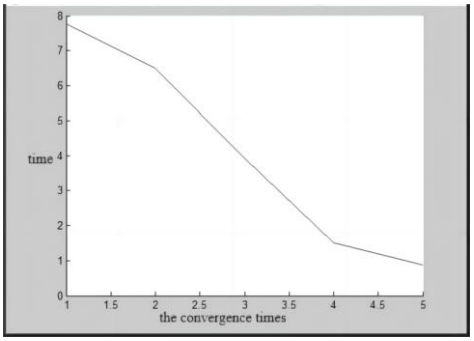

Fig.8 Training Error of Kmeans-RBF

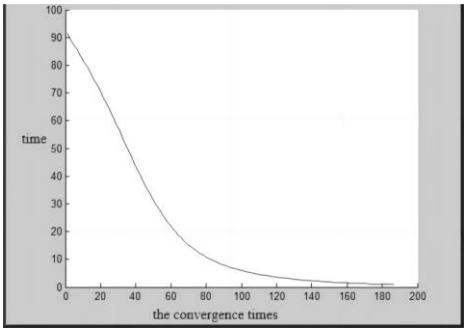

Fig.9 Training Error of Gradient RBF

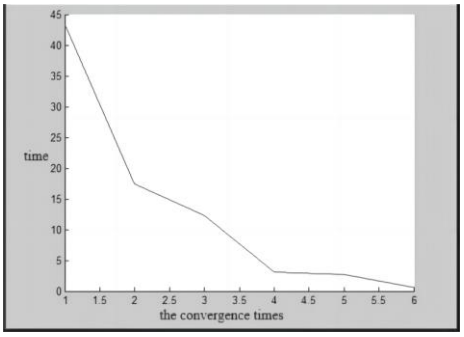

Fig.10 Training Error of OLS-RBF

From the X-axis, it is known that the convergence times are 5,183and 6 respectively, that is, Gradient RBF neural network is the slowest convergence speed, while the training time of Kmeans-RBF neural network and OLS-RBF neural network is the shortest, and the convergence speed of the network is the fastest, and Kmeans-RBF does not need to define the number of hidden layer nodes in advance.

5 Conclusion

The water pollution discharge of enterprises is affected by many factors, such as regional economy, facility investment, surrounding environment and so on. It is difficult to express it with a general mathematical model, and the machine learning model can better make up for these shortcomings. Kmeans-RBF artificial neural network has many advantages, such as simple training and fast learning convergence, which is very suitable for establishing evaluation model. In this study, by formulating the evaluation system of economic benefit and water pollution protection balance degree, the Kmeans-RBF neural network algorithm is used to establish a classifier of enterprise environment and economic balance degree, and the effectiveness of the algorithm is evaluated. Finally, the following conclusions are obtained:

(1) Starting from the policies and regulations related to environment and economy, this paper attempts to establish a set of index system suitable for evaluating the degree of environmental protection and economic balance of enterprises, and selects 241 sample enterprises by means of key investigation. The "index system" is quantitatively analyzed by kmeans RBF artificial neural network algorithm to form a classifier.

(2) The classifier is effective. The sample enterprises can be divided into three categories: enterprises with good balance, enterprises with excellent balance and enterprises with general balance.

(3) Kmeans-RBF artificial neural network can accurately classify other enterprises, saving time and effort.

Topic Acknowledgments: 2022 Education and Teaching Reform Project of Guangdong Internship Guidance Committee of Vocational Colleges (Project No.: A2022044)

References

[1] Guo Ke;CaoYuequn..Evaluating the efficiency of green economic production and environmental pollution control in China[J]. Environmental Impact Assessment Review..2023,10(104)：255-268;

[2] Yongxin Gao.Environmental performance evaluation of heavy polluting enterprises[J].Academic Journal of Business & Management. 2022,4(18):950-965;

[3] Das Shreya;Sengupta Sudip;Patra Prasanta Kumar.Appraisal of environmental, ecological and carcinogenic risk due to heavy metals in a sewage and solid waste contaminated area [J].Soil and Sediment Contamination: An International Journal. 2023,5(32): 591-614；

[4] Chandra Krishno;Proshad Ram;Dey Hridoy Chandra;Idris Abubakr M，A review on radionuclide pollution in global soils with environmental and health hazards evaluation.Environmental geochemistry and health. 2023(8):201-221;

[5]Bayartungalag Batsaikhan;Seong-TaekYun;Kyoung-HoKim.Groundwater contamination assessment in Tarantula City, Mongolia with combined use of hydro chemical, environmental isotopic, and statistical approaches [J] .Science of the Total Environment.. 2020(765).:14-27;

[6] Piroozfar Parisa ;Alipour Samad ;Modabberi Soroush .Using multivariate statistical analysis in assessment of surface water quality and identification of heavy metal pollution sources in Sarough watershed, NW of Iran[J].Ernvironmental monitoing and assessment .2021,9(193): 564-564;

[7] Zhao Qinglin; Guo Yanbing; Mei Qiang. Review of the methods to determine the center point of the RBF neural network [J]. Guangdong Automation and information Engineering.2002(02):13-15;
[8] Broomhead;D.S.; Lowe.D.Multivariable functional interpolation and adaptive networks[J].Complex
System.1988,2(3): 321-355;
[9] Chen S.;Grant P.M.;Cowan C.F.N.Orthogonal least-squares algorithm for training multioutput radial basis function networks[J].IEE proceedings.Radar, sonar and navigation.1992,6(139): 378-378;
[10], He Yingsheng, Duan Mingxiu.A new RBF neural network design based on an improved kmeans clustering method [J]. Journal of Shaoyang University (Natural Science Edition).2008(02):48-50;
[11] Chung Sang Ryu;Chae Bong Son;Eun Soo Kim .Multi-target Data Association System using RBF and Hopfield Networks[J].ICONIP : International Conference On Neural Information Processing.
1994,1(3):110-120;
[12] Hiroyuki Shibayama ;Toshimichi Saito .A Realistic RBF Approximation of Chaotic Dynamics[J].ITC-CSCC :International Technical Conference on Circuits Systems, Computers and Communications.1996,1(1):145-159;
[13] Thomas A.;Cheong C.S.;Tellam J.H.Development of a GIS based model for assessment of groundwater contamination through sewage networks in urban environments[J].Pollution Research.2006,1(25): I-VIII;
[14] Kabacoff ， Robert I.R in Action, Third Edition R[M].Manning Publications .2021(8):350-369;
[15] Mubera Sosthene ;Ogada PhilipOchieng ;Cigoja.Dragan .Assessment of Industrlal Wsatewaiter Pollution in Developing Countries-Current Polluttion Level in Rwanda [J].International Journal of Advanced Research.2016,10(4): 716-723;
[16] Ratika Agarwal ;Preeti Chhabra ;Aparna Prashant Goyal ;Sanjay Srivastava .Predictive Modelling to Assess Groundwater Pollution and Integration with Water Quality Index[J].International Journal of Engineering and Advanced Technology (IJEAT).2019,5(8):1076-1084;
[17] Environmental Engineering Assessment Center, Ministry of Environmental Protection. Technical methods for environmental impact assessment [M]. China Environment Press.2013(3)：208-227；
[18] Yue Wang;Song Xue ;Junming Ding .Research on water pollution prediction of township enterprises based on support vector regression machine[J].E3S Web of Conferences.2021,1(228): 02-14;
[19] Huang Taozhen ;Zheng Wei .Water Pollution Prevention and Control of Chemical Enterprises Based on Cooperative Game[J].Chemical Engineering Transactions (CET Journal).
2018,9(67):289-301;
[20] Hassan Yousefi ;Alireza Taghavi Kani ;Iradj Mahmoudzadeh Kani .Multiscale RBF-based central high resolution schemes for simulation of generalized thermoelasticity problems[J].Frontiers of Structural and Civil Engineering.2019,2(13): 429-455;
[21] Giuseppe Ciaburro;Balaji Venkateswaran.Neural Networks with R[M].Chia Machine Press.
2018(7):119-126.