Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.27812

242 (2021) 226–234
December

A novel method of the water quality system assessment: genetic algorithm optimized back-propagation neural networks evaluation model

Yueqing Ding[a,†], Wei Hong[b,†], Jianhua Yang[a,*]

[a]*School of Automation, Northwestern Polytechnical University, Xi'an 710072, PR China, email: yangjianhua@nwpu.edu.cn (J. Yang)*
[b]*College of Urban and Environmental Sciences, Northwestern University, Xi'an 710069, PR China*

Received 26 August 2021; Accepted 23 September 2021

ABSTRACT

Water quality pollution has become a primary environmental global concern as society is changing. The pollution of environmental resources is an issue of social concern worldwide, particularly the deterioration of water quality. Groundwater quality deterioration has been the focus in the water quality field. A precise predictive model is needed to obtain a clear understanding of the factors controlling groundwater quality conditions, to assess how to reduce risk and to optimize urban water quality management for the purpose of improving water quality in urban groundwater systems. Therefore, how to improve the accuracy of rating prediction calculated by the water quality systematic assessment model becomes the focus of this study. Combining back-propagation (BP) neural network with a genetic algorithm, this paper proposes the water quality, systematic assessment model, based on genetic algorithm optimized BP neural networks and its potential application to a typical urban groundwater quality system assessment in northwest China. In this study, nine main factors were used as indicators of qualifying water quality that encompasses pH, ammonia, chloride, nitrite, dissolved solids, chemical oxygen demand and fluoride ions. The method is illustrated with water quality of 223 variables data from a surveillance system of fifteen groundwater water monitoring sites in Xi'an, Shaanxi Province, China during the 1996–2015 period. Firstly, using Shannon entropy (information gain) to classify the recorded data into five categories. Referring to (in terms of) the groundwater quality criterion in 2017 (GB/T14848-2017), following data pre-processing to provide more abundant information for recognition. Second, correlations were analyzed using the Spearman correlation. Water quality ratings were significantly highly correlated with statistical measures of the chloride, nitrite, dissolved solid and fluoride ions. Then principal component analysis methods are deployed for dimension reduction. On this basis, the evaluation index is refined; the basic structure of the back-propagation neural network (BPNN) is introduced, and a genetic algorithm is used to improve BP neural network. The results indicated that a genetic algorithm-back propagation neural network (GA-BPNN) model accounted for an accuracy of 90.91%, which was much higher than 63.64% on accuracy by BPNN. And thus, it is important to develop GA-BPNN modeling methods to enhance the evaluation accuracy of water quality parameters. This method can be used as a decision support tool to evaluate the impact of urban groundwater utilization on water quality.

Keywords: Groundwater quality; Shannon entropy; Spearman correlation; Back-propagation neural network; Genetic algorithm; Water quality systematic assessment

* Corresponding author.

† These authors contributed equally to this work.

1. Introduction

Environmental pollution, especially water contaminants, has become a serious problem to be urgently overcome. Although about 71% of the earth's surface is covered by water, only 0.03% of freshwater resources can be directly used by human beings, such as freshwater lakes, rivers and shallow groundwater [1]. Groundwater is a prime resource of freshwater globally accounting for roughly 96% of the freshwater reserves of the planet [2]. It is a valuable natural resource that is used for public water systems in some districts of many countries all over the world, such as drinking, irrigation and industrial purposes. Over 50% of the world's largest aquifers show declining trends in groundwater storage 6 and 1.7 billion people inhabit areas where groundwater resources are scarce, and groundwater pollution can constitute a serious risk to human health [3,4]. With the increasing population and development of various industries worldwide, organic matter such as drugs, pesticides, surfactants, and raw chemical materials cause an increasing amount of pollutants in groundwater, there is a growing concern about the deterioration of groundwater quality due to geogenic and anthropogenic activities. Groundwater contamination, in particular, is a critical public concern because it has concealment, irreversibility, diversity of constituent factors and complexity of the system, and its pollution problem is far less intuitive than surface water [5]. It is polluted by harmful elements to varying degrees, and it threatens the ecological environment, human health and drinking water resources. Quality groundwater resources are essential for the socioeconomic development and physical health of the local populations that depend on groundwater [6]. Accurately predicting fluctuations in groundwater quality is the focus of effective management of groundwater [7].

2. Methodology

2.1. Data pre-processing and analysis

2.1.1. Data sources and descriptions

This study is based on the representativeness and availability of data. This paper takes Xi'an, a representative city in Northwest China, as the evaluation object, the groundwater environmental quality data of Xi'an city from 1996 to 2015 were used as input variables to establish the genetic algorithm-back propagation neural network (GA-BPNN) model to make it have better water quality prediction accuracy. The index data are statistical data based on the authoritative data of groundwater monitoring reports of Shaanxi Province, China [8]. Based on the above data, for the missing data, the method of cubic spline interpolation was adopted according to the data before and after the year, part of the missing data was made up, and the groundwater environmental quality of Xi'an was evaluated and analyzed.

2.1.2. Interpolation method

Due to the lack of data in groundwater monitoring, the data were interpolated after processing those for adjacent years. The missing values were interpolated using cubic spline interpolation [5]. Cubic spline interpolation is a smooth curve through a series of shape points. The algorithm is summarized as follows.

Suppose there are $n+1$ data nodes (x_0, y_0) (x_0, y_0)... (x_n, y_0) (x_n, y_0).

(A) Calculate the step size $h_i = x_i + 1 - x_i$ ($i = 0, 1, \ldots, n-1$);
(B) Plug data nodes and specified first endpoint conditions into the matrix equation:

$$A_i = \begin{bmatrix} -h_i & h_0+h_1 & -h_0 & \cdots & \cdots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & & \vdots \\ 0 & h_1 & 2(h_1+h_2) & h_2 & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & -h_{n-1} & h_{n-2}+h_{n-1} & -h_{n-2} \end{bmatrix} \tag{1}$$

(C) Solve the matrix equation and obtain the quadratic differential value. The matrix is a tridiagonal matrix, specifically, where A_i is a $m_i \times m_i$ diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element.
(D) Calculate the coefficient of the spline curve:

$$\begin{aligned} a_i &= y_i \\ b_i &= \frac{y_{i+1}-y_i}{h_i} - \frac{h_i}{2}m_i - \frac{h_i}{6}(m_{i+1}-m_i) \\ c_i &= \frac{m_i}{2} \\ d_i &= \frac{m_{i+1}-m_i}{6h_i} \end{aligned} \tag{2}$$

where $i = 0, 1, \ldots, n-1$; $i = 0, 1, \ldots, n-1$.

For each subinterval $x_i \le x \le x_i$, create an equation:

$$g_i(x) = a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3 \tag{3}$$

To address missing data in Laowatan Village, the northern suburb of Xi'an City, 2013. The data were interpolated by using cubic spline lines to connect and smooth the curves. The 2013 pH of the interpolation is 8.30922195954923.

2.1.3. Standardized treatment

It is different from the scope and degree of groundwater's ecological status, to be fair to analyze data, before on groundwater environment quality ratings, all indexes radial dimension, first of all, need to use chemical indicators of reference state (without human interference in the best state) and ecological threshold level of the index (worst status value) for all kinds of indexes for standardization [10]. The Z-scores method is used to standardize the nine indexes of groundwater environmental quality, namely, pH value, ammonia nitrogen, chloride ion sulfate, nitrate,

nitrous acid, dissolved solids, chemical oxygen demand (COD) and fluoride ion, and the standardized values of the evaluation indexes are assigned to 0 or 1 respectively (the evaluation index value is greater than the maximum standard or less than the minimum standard). Due to the different attributes of the indicators, the standardization methods are different. In this study, pH value is a bidirectional index, and according to the groundwater environmental quality standard, the best range is between 6.5 and 8.5, which is standardized by Eqs. (1) and (2). Other indicators such as ammonia nitrogen are all negative indicators, that is, the smaller the value is, the better the quality is. Eq. (3) is adopted standardization.

$$S_{i,\mathrm{PH}} = \frac{7.5 - D_{i,\mathrm{PH}}}{7.5 - D_{\min}} \left(D_{i,\mathrm{PH}} < 7.5\right)$$
$$S_{i,\mathrm{PH}} = \frac{D_{i,\mathrm{PH}} - 7.5}{D_{\min} - 7.5} \left(D_{i,\mathrm{PH}} > 7.5\right)$$
$$S_{i,j} = \frac{D_{\max} - D_{i,j}}{D_{\max} - D_{\min}} \tag{4}$$

In the formula, $D_{i,\mathrm{pH}}$ and D_{ij} are the original data of the pH index of the ith evaluation object and the jth index respectively, $D_{\min} = \min\{D_{1j}, D_{2j}, \ldots, D_{mj}\}$, $D_{\max} = \max\{D_{1j}, D_{2j}, \ldots, D_{mj}\}$. $S_{i,\mathrm{pH}}$, $S_{i,j}$, $D_{\max}$, $D_{\min}$ are the concentration standardization values, ecological threshold values and monitoring point concentrations of pH, ammonia nitrogen and other factors in groundwater environmental quality indexes.

2.1.4. Water quality classification based on entropy weight method

In order to avoid the interference of subjective factors, information entropy classifies the groundwater environmental quality index based on order degree and utility value. The entropy value method is a kind of objective weighting method, and the determination of entropy weight depends on the contribution degree of index importance [10]. Considering the impact of the fluctuation information of the evaluation value of the evaluation object on the decision-making process of the decision-maker, according to the nature of entropy weight, the indexes with obvious difference in information and large entropy weight and small entropy value are retained, and the evaluation values with relatively stable and very limited information are eliminated [11]. For example, in some extreme cases, the index with the same evaluation value of the evaluated object will not convey valuable information to the decision-maker. In this case, the entropy value and entropy weight value are 1 and 0 respectively, and the index can be eliminated. The entropy weight method quantifies the information of each index of groundwater environmental quality in the study area and gives weight to each factor. The weight of the index is determined objectively by the amount of information provided by the monitoring value of the index. The specific calculation steps are as follows:

(1) Assuming that there are m evaluation objects or samples, and each evaluation object or sample has evaluation indicators (topological parameters), the original data is formed into a judgment matrix R:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & r_{ij} & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}^{c} \tag{5}$$

R_{ij} is the measured value of the jth index of the sample i to be judged (i = 1, 2, …, m; j = 1, 2, …, n).

(2) Using Eqs. (1)–(3) in Z-scores method, B_{ij} is the measured standardized value of the jth index of the sample i to be judged (i = 1, 2, …, m; j = 1, 2, …, n), the judgment matrix R is normalized to form a new matrix $B = (B_{ij})_{m\times n}$, and the value range is (0,1).

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & b_{ij} & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \tag{6}$$

(3) Calculate the information entropy of index e_j:

$$e_j = -K\sum_{i=1}^{m} v_{ij} \cdot \ln v_{ij}$$
$$v_{ij} = \frac{b_{ij}}{\sum_{i=1}^{m} b_{ij}} \tag{7}$$

Among them, i = 1, 2, …, m; j = 1, 2, …, n, in the formula, $K > 0$, in is the natural logarithm, and generally $0 \le e \le 1$. Assuming that when $v_{ij} = 0$, $v_{ij}\ln(v_{ij}) = 0$, the correction v_{ij} is calculated:

$$v_{ij} = \frac{1 + b_{ij}}{\sum^{m}\left(1 + b_{ij}\right)} \tag{8}$$

(4) Calculate the entropy weight of the jth index:

$$W_j = -\frac{d_j}{\sum_{j=1}^{n} d_{ij}} \tag{9}$$

where the information utility value of the jth index $D_j = 1 - e_j$, W_j is the entropy weight of the ith index, which satisfies.

2.2. Data analysis

2.2.1. Principal component analysis

The key to influencing the prediction effect and determining the reliability of the results is to screen the input

variables of the model and simplify its input structure. Principal component analysis (PCA), is a kind of multiple attribute dimension reduction method, the fundamental principle is to extract the original variables, the eigenvalues of the covariance matrix and its related load by the maximum variance method to rotate to generate a new orthogonal linear combination of the variables and the original, with the minimum of the original variable data reduction, provide describe the whole data set the parameters of the most valuable information, improve the accuracy of predicted results [12]. In this paper, the process of implementing the algorithm to reduce the dimension of the original data is as follows: Determine the original indicators and conduct standardization. The original data matrix X was constructed from the original water quality data of the study area.

$$X = \begin{Bmatrix} x_{11} & x_{12} & x_{1j} \\ x_{21} & x_{22} & x_{2j} \\ x_{i1} & x_{i2} & x_{ij} \end{Bmatrix},\ i = 1,2,\ldots,1500;\ j = 1,2,\ldots,20 \tag{10}$$

where x_{ij} represents the jth index of the ith sample unit, the data is standardized by the following Eq. (10):

$$x'_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j} \tag{11}$$

where $\overline{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij},\ S_j = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ij} - \overline{x}_j\right)^2$.

2.2.2. Water quality rating model based on GA-back propagation neural network

Back-propagation (BP) neural network is a multilayer feedforward neural network based on error Back Propagation Algorithm training and composed of highly nonlinear mapping from input to output. As a very classic machine learning algorithm, BP is suitable for pattern recognition, classification, function prediction and other problems [13]. It consists of an output layer, an input layer, and a number of hidden layers. Each layer contains a number of basic units of neurons, and each data layer includes a number of data nodes. The levels of neurons are interconnected through thresholds or weights, and there is no correlation between the states of neurons at the same level. It is a topological network that simulates the process of neural conflict and is established on the basis of biological research. After receiving signals from the outside world at the end of the dendrite of the neural network, they are transmitted to neurons, then processed and fused, and finally transmitted to other neurons.

Its topology is shown in Fig. 1.

For the ith neuron, X_1, X_2, …, X_j is the input of the neuron, and the input is usually the independent variable of the key influence on the system model, W_1, W_2, …, W_j adjusts the weight to weight ratio of each input for the connection weight.

There are many ways to combine signals and input to neurons. The net input of neurons can be obtained by selecting the most convenient linear weighted sum:

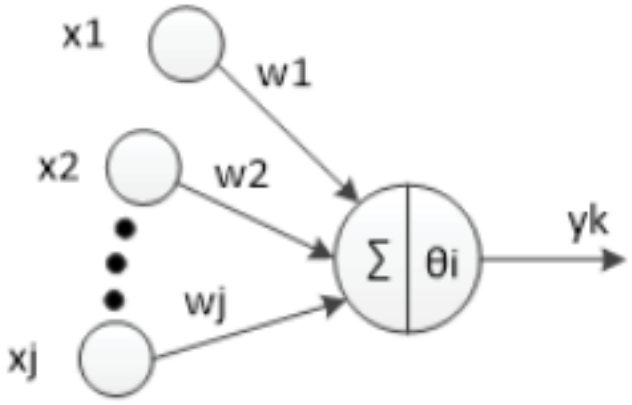

Fig. 1. Neuronal structure.

$$\text{Net}_{\text{in}} = \sum_{i=1}^{n} w_i \times x_i \tag{12}$$

Represents the threshold value of the neuron. According to the knowledge of biology, only when the information received by the neuron reaches the threshold will it be activated. Therefore, compare and, and then process through an activation function to produce the output of the neuron as:

$$y_j = f\left(\sum_{i=1}^{n} w_i \times x_i\right) \tag{13}$$

where $w_0 = -1$, $x_0 = \theta_j$, f represents activation function. Common types include threshold function, logarithmic Sigmoid function, tangent Sigmoid function, linear function, etc. The structure of the single-layer neural network can be expressed as Fig. 2.

BP neural network will reverse transfer the weight of each layer according to the error of each training, to continuously improve the model accuracy until it reaches the invitation. The error function can be expressed by the least square method as follows:

$$e = \frac{1}{2}\sum_{o=1}^{q}\left(d_o(k) - yo_o(k)\right)^2 \tag{14}$$

where d_0 represents the predicted value and yo_o represents the actual value.

Although widely used in various fields, the BP neural network in dealing with complex nonlinear problems have prominent advantages, because of the feedforward learning with the method of partial layers of error correction, which leads to problems such as slow learning speed, sensitive initial weight, sample dependence, easy to trap local optimum and interference of training parameter characteristics on training results [14]. Based on the limitations of BP neural network and genetic algorithm (GA) (based algorithm, genetic algorithm) has outstanding advantages, compared with the general Algorithm has global search, search the advantages of high efficiency, extensibility, through the selection, crossover and mutation operating mechanism,

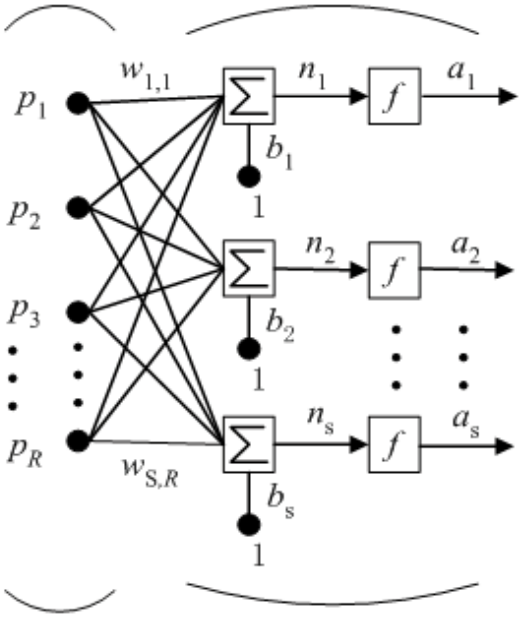

Fig. 2. Single-layer neural network structure.

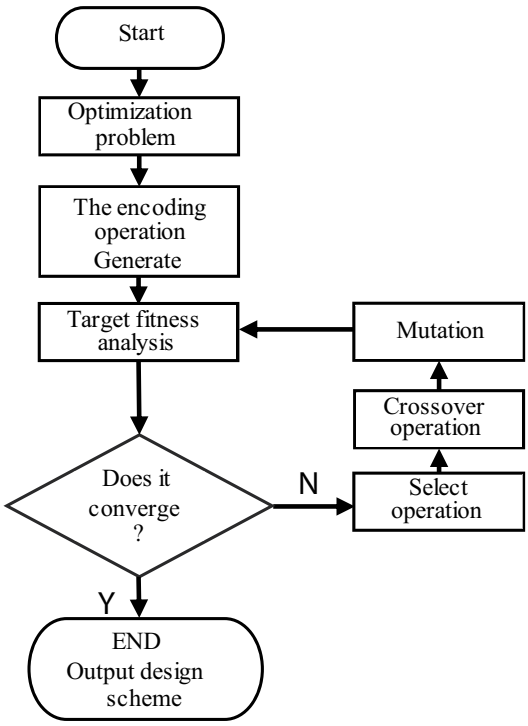

Fig. 3. Genetic algorithm.

make the adaptability of the individuals in the population to improve until convergence, globally search for the solution of the problem area of the optimal solution, the core idea is that natural selection, survival of the fittest [15]. GA is selected to optimize BP neural network. Combining the advantages of the two algorithms, GA is used to find the optimal weight and assign value to the neural network, so as to achieve higher accuracy, and play a fast and accurate effect in solving complex nonlinear problems such as groundwater environmental quality [16].

A genetic algorithm is used to solve the selection problem of machine learning features [17]. The idea of the algorithm is to search for the optimal solution by simulating the natural evolution process. It has good global optimization ability and inherent implicit parallelism can automatically search to the optimal search space and has the characteristics of adaptive adjustment of search direction. The algorithm has the characteristics of unrestricted function continuity and direct manipulation of an object structure. The flow of the genetic algorithm is shown in Fig. 3.

GA is used to optimize the weight of neural network to achieve the minimum error, and its mathematical model can be expressed as:

$$\mathrm{Min}E\left(w_i\right) \tag{15}$$

s.t. $0 \le w_i \le 1$

where E represents the error function of the neural network and represents the weight of each layer.

3. Results and discussion

3.1. Characteristics of groundwater quality in urban areas of Xi'an

Table 1 presents the descriptive summary statistics of groundwater quality data of 15 monitoring sites in Xi'an from 1996 to 2015 in the Shaanxi Groundwater Monitoring Yearbook. The results show that the pH values of the 15 sites are all greater than 7, showing weak alkalinity. Among the other 8 indicators, there are significant differences among different sites, and the standard deviation of most of the indicators are more significant than their mean value, which indicates that there is a significant difference between these indicators among different sites.

Among them, 100% of the sites with Cl^- and NO^{2-} mean values lower than the standard values, but only 46.67% of the sites with NO^{3-} met the standard values.

The results show that the average NO^{3-} concentration of 15 monitoring sites in Xi'an city is high. In order to further compare the differences between different sites, a logarithmic operation was performed on the data of a larger magnitude, and the average value of each index was measured. The average values of pH, SO_4^{2-}, total dissolved solids (TDS) and COD vary little between different sites, but they are basically the same. At the fourth and fifteenth measurement sites, the content of a number of indicators was higher than that of other sites. The contents of SO_4^{2-}, NO^{3-}, NO^{2-} and TDS at the 4th station were significantly higher than those at other stations, and the contents of NH_3–N and COD at the 15th station were both higher than those at other stations. The highest pH, Cl^- and F^- sites were at the 8, 7 and 2 sites, respectively. The results show that the groundwater environmental quality of the main stations in the study area is generally affected by human activities, and the groundwater environmental quality of each station may be affected by the same pollution source such as food factory, sewage treatment plant or paper mill. Among various contaminants, it needs to strictly control COD and ammonia–nitrogen (NH_3–N) emissions to protect water resources [18].

Table 1
Groundwater quality parameters and summary basic statistics of the Xi'an

Parameters	Mean	S.D.	Min.	Max.	Standard	Below standards for all sites (%)	Units
pH	7.81	0.40	7.09	9.08	6.5–8.5	0.00	
NH_3–N	0.49	4.04	0.01	58.50	0.5	86.67	mg/L
Cl^-	85.66	62.01	4.09	435.00	250	100.00	mg/L
SO_4^{2-}	142.81	98.91	1.65	655.00	250	93.33	mg/L
NO^{3-}	71.13	146.05	0.00	880.00	20	46.67	mg/L
NO^{2-}	0.17	0.72	0.001	8.07	1	100.00	mg/L
TDS	825.03	435.35	144	2,724.00	1,000	86.67	mg/L
COD	1.30	1.12	0.32	9.26	3	93.33	mg/L
F^-	0.79	2.74	0.05	41.00	1	93.33	mg/L

3.2. Water quality classification

According to relevant theories in system theory, the groundwater environmental quality assessment system contains information order degree and information utility through information entropy. Based on the entropy weight method, the weight of 9 indicators of groundwater environmental quality in Xi'an is calculated. The results are shown in Table 2. The groundwater quality was further evaluated according to China's current water quality standards, and the groundwater water quality data were divided [19]. After sorting, the top 6 factors were selected as pH, Cl^-, SO_4^{2-}, NO^{3-}, TDS and F^-. Next, each element is graded according to the grade interval set by the standard for underground water quality, and the results are collated.

Finally, the grade divided by each element in each group of data is counted, and the one with the most number of grades in this group of data is identified as the water quality grade evaluation result of this group (if there are two grades in the same group with the same number of times, the classification shall follow the principle of "superior" rather than "inferior". For example, if level 2 occurs twice and level 3 occurs twice, the final result is 2).

15 typical groundwater reservoirs in the study area 223 samples category, class II and class III water quality, four and five respectively 47, 120, 47, 6 copies and 3, respectively accounted for 21.07% of the total groundwater samples, 54.26%, 21.07%, 2.64% and 1.32%, 96% than class III water quality above, suitable for centralized drinking water and industrial and agricultural water; In the groundwater quality classification, the chemical components of category IV, which accounted for 2.64%, were relatively high. Based on the water quality requirements for agriculture and industry as well as a certain level of human health risks, it was suitable for agricultural and part of industrial water and could be used as drinking water after proper treatment. Category V groundwater, which accounts for 1.32% of the groundwater quality classification, has a high chemical component content and is not suitable for domestic reference water sources. It can be selected according to the purpose of use.

3.3. Correlation between water quality variables and grades

Spearman correlation coefficient, also known as Spearman rank correlation coefficient. "Rank" can be understood as an order or a sort of achievement, so it is solved according to the ranking position of the original data, and this representation form does not have the restrictions of finding Pearson's correlation coefficient [20].

Table 2
Weight calculation results of entropy weight method

Parameter	Weight
pH	0.164761
NH_3–N	0.028141
Cl^-	0.130441
SO_4^{2-}	0.113381
NO^{3-}	0.145741
NO^{2-}	0.048108
TDS	0.128198
COD	0.088546
F^-	0.152683

Now X and Y are defined as two groups of data, and the Spearman correlation coefficient calculation formula is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \tag{16}$$

where d_i is the rank difference between X_i and Y_i (the rank of a number is the number of bits in the column sorted from smallest to largest) and r_s is between –1 and 1 $r_s = 1$, perfect positive correlation; $r_s = -1$, completely negative correlation; $r_s > 0$, positive correlation; $r_s < 0$, Negative correlation. The closer r is to 1, the higher the correlation between samples. The closer r is to 0, the lower the correlation between the samples.

Table 3 presents the correlation analysis results of these 9 variables and the grade of groundwater environmental quality. The correlation between pH and the grade of groundwater water quality is low ($r = 0.111$). The negative correlation between pH and Cl^-, SO_4^{2-}, TDS, NO^{3-}, NO^{2-}, and COD is mainly due to the fact that a large amount of rainwater penetrates into the aquifer in the rainy season.

Table 3
Spearman correlation analysis result

	pH (mg/L)	NH_3–N (mg/L)	Cl^- (mg/L)	SO_4^{2-} (mg/L)	NO^{3-} (mg/L)	NO^{2-} (mg/L)	TDS (mg/L)	COD (mg/L)	F^- (mg/L)	Grade
pH	1	0.201**	–0.215**	–0.295**	–0.091	–0.076	–0.271**	–0.146*	0.139*	0.111*
NH_3–N	0.201**	1	0.068	0.053	–0.222**	0.189**	–0.013	0.263**	0.006	0.085
Cl^-	–0.215**	0.068	1	0.700**	0.460**	0.326**	0.843**	0.192**	0.196**	0.588**
SO_4^{2-}	–0.295**	0.053	0.700**	1	0.421**	0.296**	0.855**	0.169**	0.179**	0.558**
NO^{3-}	–0.091	–0.222**	0.460**	0.421**	1	0.349**	0.577**	–0.273**	0.117*	0.528**
NO^{2-}	–0.076	0.189**	0.326**	0.296**	0.349**	1	0.303**	0.143*	0.238**	0.348**
TDS	–0.271**	–0.013	0.843**	0.855**	0.577**	0.303**	1	0.129*	0.173**	0.559**
COD	–0.146*	0.263**	0.192**	0.169**	–0.273**	0.143*	0.129*	1	0.076	0.013
F^-	0.139*	0.006	0.196**	0.179**	0.117*	0.238**	0.173**	0.076	1	0.454**
Grade	0.111*	0.085	0.588**	0.558**	0.528**	0.348**	0.559**	0.013	0.454**	1

Under the interaction of water-rock, the ion concentration increases and the pH value of groundwater is reduced. At the same time, under the influence of negative correlation between NO^{3-} and NH_3–N and COD, NH_3–N and COD are weakly correlated with the grade of underground water quality (r = 0.085).

It can be further obtained from Cl^-, SO_4^{2-}, TDS, NO^{3-}, NO^{2-}, F^- has a strong positive correlation with groundwater quality grade (R-value is 0.348–0.588), indicating that ions from common sources have a significant correlation with groundwater environmental quality grade. Therefore, the groundwater quality in the study area is obviously affected by Cl^-, TDS, F^-, NO^{2-}, SO_4^{2-} and NO^{3-}, as well as potentially affected by NH_3–N and COD, indicating that human activities have an important impact on groundwater quality grade. Among them, Cl^-, SO_4^{2-} and TDS showed a strong positive correlation (0.70–0.855), the results show that the groundwater quality may be affected by the pollution sources such as food factories, sewage treatment plants or paper mills [21].

3.4. Identify the main impact factors via PCA

Use SPSS software to solve and analyze: PCA principal component dimension reduction toolbox in statistical software SPSS 12.0 was used to input the standardized data matrix with covariance matrix and maximum variance method for dimension reduction. The output results of SPSS are summarized in Tables 4 and 5. As shown in Table 1, the cumulative variance contribution rate of the first three components reached 99.71%. Meanwhile, as shown in Fig. 4, the distribution curve of characteristic roots gradually flattens out after the third characteristic root, so it is reasonable to extract the first three principal components.

PCA can effectively eliminate the overlap sampling data, non-homogeneity and periodic trends of correlation information, compression is applied to the matrix, in at the same time retaining as much as possible to reduce the number of matrix d matrix in the main features of the representative of multivariable comprehensive factor to reflect the original variable information as much as possible, will use the original water quality influence factors into explains mainly the main component of information, dimension reduction

Table 4
PCA principal component results

Principal components	A total of	Percentage of variance	Cumulative %
1	13,0516.97	59.40	59.40
2	66,894.27	30.45	89.85
3	21,661.70	9.86	99.71

Table 5
PCA principal component coefficients

Parameters	1	2	3
pH	–0.071	–0.248	–0.240
NH_3–N	–0.053	–0.008	–0.021
Cl^-	0.760	0.370	0.326
SO_4^{2-}	0.348	0.917	0.188
NO^{3-}	0.945	0.244	–0.219
NO^{2-}	0.095	0.074	0.067
TDS	0.767	0.552	0.327
COD (mg/L)	0.004	0.049	0.200
F^- (mg/L)	0.094	0.027	0.124

to improve the modeling speed, so as to achieve the aim of simplifying space and the amount of data, to identify the influence of groundwater environment quality potential pollution sources [22]. According to the above analysis, the spatial distribution of factors after rotation of PCA principal component method will use the three principal components after dimensionality reduction of PCA as input variables to establish a neural network for water quality rating.

3.5. Model solution and comparison

According to the above theory, the establishment of GA-BP neural network water quality rating model, the results after PCA dimension reduction above as the eigenvalues of the neural network training, the input data

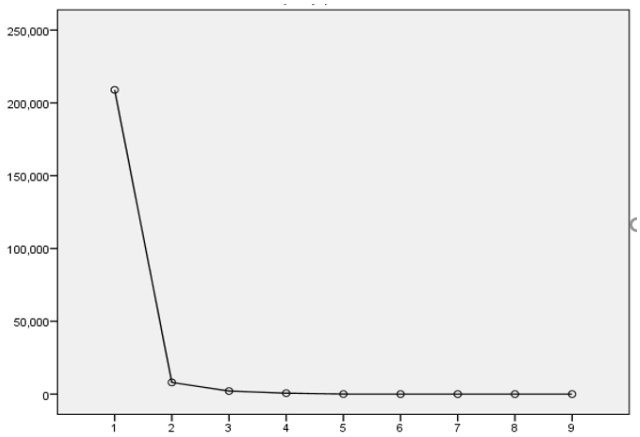

Fig. 4. Gravel figure.

layer water quality indicators for PH, ammonia (NH_3–N) and chloride ions, each group of data of the groundwater environment quality level as the output value of the neural network training, with the GA to optimize the BP neural network until reaching the optimal precision. The parameters of the genetic algorithm and BP neural network are shown in Tables 6 and 7.

In order to make the model more reliable, this paper ran 30 times in MATLAB and took the median value as the resulting model. After the optimal weight is assigned to the neural network, the neural network converges to the 378 generations. The GUI interface of the neural network is shown in Fig. 5.

The rating accuracy of GA-BPNN was 90.91%, and that of the back-propagation neural network (BPNN) was 63.64%. Fig. 5 shows the comparison between the prediction results of the two methods and the actual rating. It can be seen that the accuracy of GA-BPNN is significantly better than that of BPNN.

The results prove that the improved BP neural network established by GA in this paper has higher accuracy, reaching 90.91%, which is obviously better than the traditional BPNN. The genetic algorithm as a kind of from the biological natural selection and genetic mechanism of the random search algorithm, BP neural network optimized for better performance, for water quality evaluation has brought great convenience so that applied to the water quality monitoring of groundwater and surface water, especially in the management of water resources has great reference value in [23,24].

4. Conclusions

In this study, the cubic interpolation method was used to supplement the missing data, and negative indicators such as pH bidirectional index and ammonia nitrogen were standardized. The entropy weight method was used to classify the groundwater quality grade of Xi'an City. The results showed that 4% of the 15 groundwater monitoring points in Xi'an City did not meet the standard of domestic drinking water. In order to analyze the correlation of groundwater environmental quality indexes in the study area, Spearman correlation analysis shows that Cl^-, TDS, F^-, NO^{2-}, SO_4^{2-} and NO^{3-} have a large correlation, while pH, NH_3–N and COD have a slight correlation, among

Table 6
Parameters selected by the genetic algorithm

Parameter	Value
Maximum number of evolutions	500
Population size	30
Crossover probability	0.75
Mutation probability	0.1

Table 7
Selected parameters of GA-BPNN

Parameter	Value
Learning algorithm	Traingdx
Activation function	Input layer tansig, Output layer purelin
Loss function	MSE
Maximum learning frequency	500
Number of hidden layers	15
Vector	0.1
Ratio of training set	80%
Test set ratio	20%

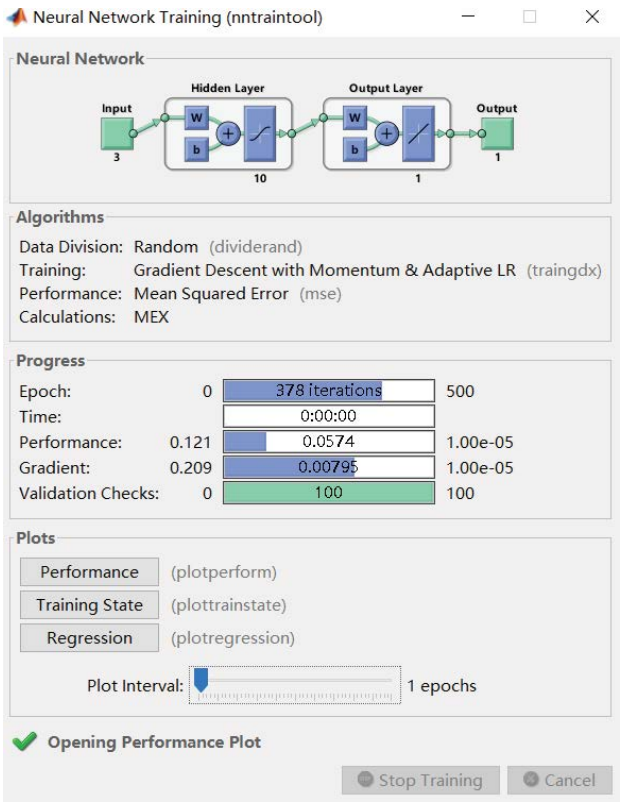

Fig. 5. Neural network training GUI interface.

which Cl^- has a strong positive correlation with SO_4^{2-} and TDS. PCA results show that a city life factory drainage, drainage and other reasons, level of human activities on the groundwater quality in some important influence, the main composition of influence the groundwater environment quality grade is pH, NH_3–N and F^-, the PCA dimensionality reduction after three principal components as input variables, the establishment of GA optimized BP neural network water quality rating model, after two kinds of methods of prediction results and the actual rating contrast, GA-BPNN grading accuracy rate of 90.91%, higher than BPNN 27.27%, therefore, accurate assessment, modeling and prediction of groundwater environmental quality and optimization control of important parameters have brought great convenience for water quality assessment and have a good reference value.

Funding

This study was supported by the Key R&D Program of Shaanxi Province, No. D5140200023 and Xi'an Geological Survey, China Geological Survey, No.211527180141.

References

[1] Z.Q. Wang, A.G. Wu, L. Colombi Ciacchi, G. Wei, Recent advances in nanoporous membranes for water purification, Nanomaterials (Basel), 8 (2018) 65, doi: 10.3390/nano8020065.
[2] A. Molinari, C.M. Mayacela Rojas, A. Beneduci, A. Tavolaro, M.F. Rivera Velasquez, C. Fallico, Adsorption performance analysis of alternative reactive media for remediation of aquifers affected by heavy metal contamination, Int. J. Environ. Res. Public Health, 15 (2018) 980, doi: 10.3390/ijerph15050980.
[3] K.C. Solander, J.T. Reager, Y. Wada, J.S. Famiglietti, R.S. Middleton, GRACE satellite observations reveal the severity of recent water over-consumption in the United States, Sci. Rep., 7 (2017) 8723, doi: 10.1038/s41598-017-07450-y.
[4] S. Das, R. Mandal, V.N. Rabidas, N. Verma, K. Pandey, A.K. Ghosh, S. Kesari, A. Kumar, B. Purkait, C.S. Lal, P. Das, Chronic arsenic exposure and risk of post kala-azar dermal leishmaniasis development in India: a retrospective cohort study, PLoS Negl. Trop. Dis., 10 (2016) e0005060, doi: 10.1371/journal.pntd.0005060.
[5] P. Zhang, W.-M. Wu, J.D. Van Nostrand, Y. Deng, Z. He, T. Gihring, G. Zhang, C.W. Schadt, D. Watson, P. Jardine, C.S. Criddle, S. Brooks, T.L. Marsh, J.M. Tiedje, A.P. Arkin, J. Zhou, Dynamic succession of groundwater functional microbial communities in response to emulsified vegetable oil amendment during sustained in situ U(VI) reduction, Appl. Environ. Microbiol., 81 (2015) 4164–4172.
[6] Q. Zhang, L. Wang, H. Wang, X. Zhu, L. Wang, Spatio-temporal variation of groundwater quality and source apportionment using multivariate statistical techniques for the Hutuo River Alluvial-Pluvial Fan, China, Int. J. Environ. Res. Public Health, 17 (2020) 1055, doi: 10.3390/ijerph17031055.
[7] K. Song, X. Ren, A.K. Mohamed, J. Liu, F. Wang, Research on drinking-groundwater source safety management based on numerical simulation, Sci. Rep., 10 (2020) 15481, doi: 10.1038/s41598-020-72520-7.
[8] J. Chi, Y. Zhang, X. Yu, Y. Wang, C. Wu, Computed tomography (CT) image quality enhancement via a uniform framework integrating noise estimation and super-resolution networks, Sensors (Basel), 19 (2019) 3348, doi: 10.3390/s19153348.
[9] U. Stańczyk, B. Zielosko, G. Baron, Discretisation of conditions in decision rules induced for continuous data, PLoS One, 15 (2020) e0231788, doi: 10.1371/journal.pone.0231788.
[10] Q. Xu, K. Xu, L. Li, X. Yao, Optimization of sand casting performance parameters and missing data prediction, R. Soc. Open Sci., 6 (2019) 181860, doi: 10.1098/rsos.181860.
[11] Y. Suh, Y. Park, D. Kang, Evaluating mobile services using integrated weighting approach and fuzzy VIKOR, PLoS One, 14 (2019) e0222312, doi: 10.1371/journal.pone.0222312.
[12] W. Majeed, M.J. Avison, Robust data driven model order estimation for independent component analysis of fMRI data with low contrast to noise, PLoS One, 9 (2014) e94943, doi: 10.1371/journal.pone.0094943.
[13] M. Wang, H. Wang, J. Wang, H. Liu, R. Lu, T. Duan, X. Gong, S. Feng, Y. Liu, Z. Cui, C. Li, J. Ma, A novel model for malaria prediction based on ensemble algorithms, PLoS One, 14 (2019) e0226910, doi: 10.1371/journal.pone.0226910.
[14] J. Zhang, X. Tan, P. Zheng, Non-destructive detection of wire rope discontinuities from residual magnetic field images using the Hilbert-Huang transform and compressed sensing, Sensors (Basel), 17 (2017) 608, doi: 10.3390/s17030608.
[15] Q. Li, H. Tao, J. Wang, Q. Zhou, J. Chen, W.Z. Qin, L. Dong, B. Fu, J.L. Hou, J. Chen, Z. Wei-Hong, Warfarin maintenance dose prediction for patients undergoing heart valve replacement — a hybrid model with genetic algorithm and back-propagation neural network, Sci. Rep., 8 (2018) 9712, doi: 10.1038/s41598-018-27772-9.
[16] H. Ghazvinian, S.-F. Mousavi, H. Karami, S. Farzin, M. Ehteram, M.S. Hossain, C.M. Fai, H.B. Hashim, V.P. Singh, F.C. Ros, A.N. Ahmed, H. Abdulmohsin Afan, S.H. Lai, A. El-Shafie, Integrated support vector regression and an improved particle swarm optimization-based model for solar radiation prediction, PLoS One, 14 (2019) e0217634, doi: 10.1371/journal.pone.0217634.
[17] H.W. Darwish, M.I. Attia, A.S. Abdelhameed, A.M. Alanazi, A.H. Bakheit, Comparative ANNs with different input layers and GA-PLS study for simultaneous spectrofluorimetric determination of melatonin and pyridoxine HCl in the presence of melatonin's main impurity, Molecules, 18 (2013) 974–996.
[18] K. Song, X. Ren, A.K. Mohamed, J. Liu, F. Wang, Research on drinking-groundwater source safety management based on numerical simulation, Sci. Rep., 10 (2020) 15481, doi: 10.1038/s41598-020-72520-7.
[19] T. Hong, Geological Environment Monitoring Station of Shaanxi Province, Groundwater Monitoring Yearbook of Shaanxi Province, China University of Geosciences Press, Wuhan, 2016, pp. 1–45.
[20] B.S. Raccor, A.J. Claessens, J.C. Dinh, J.R. Park, D.S. Hawkins, S.S. Thomas, K.W. Makar, J.S. McCune, R.A. Totah, Potential contribution of cytochrome P450 2B6 to hepatic 4-hydroxycyclophosphamide formation in vitro and in vivo, Drug Metab. Dispos., 40 (2012) 54–63.
[21] M.A. Martín Del Campo, M.V. Esteller, J.L. Expósito, R. Hirata, Impacts of urbanization on groundwater hydrodynamics and hydrochemistry of the Toluca Valley aquifer (Mexico), Environ. Monit. Assess., 186 (2014) 2979–2999.
[22] H. Akramifard, M.A. Balafar, S.N. Razavi, A.R. Ramli, Emphasis learning, features repetition in width instead of length to improve classification performance: case study—Alzheimer's disease diagnosis, Sensors (Basel), 20 (2020) 941, doi: 10.3390/s20030941.
[23] X. Liu, X. Pei, N. Li, Y. Zhang, X. Zhang, J. Chen, L. Lv, H. Ma, X. Wu, W. Zhao, T. Lou, Improved glomerular filtration rate estimation by an artificial neural network, PLoS One, 8 (2013) e58242, doi: 10.1371/journal.pone.0058242.
[24] W. Yang, Y. Zhao, D. Wang, H. Wu, A. Lin, L. He, Using Principal Components Analysis and IDW Interpolation to determine spatial and temporal changes of surface water quality of Xin'anjiang River in Huangshan, China, Int. J. Environ. Res. Public Health, 17 (2020) 2942, doi: 10.3390/ijerph17082942.