Pol. J. Environ. Stud. Vol. 31, No. 2 (2022), 1769-1781
DOI: *10.15244/pjoes/140170* ONLINE PUBLICATION DATE: 2022-01-25

Original Research

Groundwater Quality Evaluation Based on PCA-PSO-SVM Machine Learning in Xinzhou City, China

Qihang Ni, Leihua Yao*, Chenguang Song, Chengya Hua

School of Engineering and Technology, China University of Geosciences (Beijing), Beijing 100083, China

Received: 17 December 2020
Accepted: 12 July 2021

Abstract

The scientific evaluation of water quality change trends and pollution characteristics is of great significance to improving the current situation of water resources. The particle swarm optimization support vector machine based on principal component analysis (PCA-PSO-SVM) was used to conduct a comprehensive evaluation of groundwater quality in Xinzhou city, and the results were compared with those of a variety of traditional water quality evaluation methods. The evaluation results show that the water quality evaluation model based on PCA-PSO-SVM is more comprehensive and objective. The overall groundwater quality situation in Xinzhou city is generally good. The upper pore water quality is mainly class II, and the lower karst water is mainly class III. The water quality shows significant spatial differences. Compared with the traditional water quality evaluation method, the improved SVM algorithm compensates for the defects of the traditional method. The model structure is stable, and the accuracy and calculation efficiency are high, which makes the method worthy of promotion and application. The research results can provide a scientific basis and reference value for the water quality evaluation of groundwater-related projects.

Keywords: machine learning, particle swarm optimization (PSO), support vector machine (SVM), water quality evaluation, neural network

Introduction

Water is an indispensable natural resource for human development and the material basis for human beings and all living things to survive. China is a country with severe droughts and water shortages. The per capita water resources of the country are only 2,200 m^3, which is only 1/4 of the world average level. China is one of the 13 countries with the poorest water resources per capita in the world [1, 2]. As an important part of water resources, groundwater only accounts for one-third of the total domestic water resources [3] and shows a pattern of "more in the south and less in the north". Located in Shanxi Province in North China, Xinzhou city is a mining city with relatively poor groundwater resources compared to the national average. In recent years, with the acceleration of urbanization, the rapid growth of the urban population, and the rapid development of industry and agriculture, Xinzhou

*e-mail: leihuayao@163.com

city's groundwater pollution has increased daily, and the mining intensity in some areas has been excessive. Xinzhou city is facing serious water pollution and water shortage problems [4, 5]. In addition, Xinzhou is located near the basin of the Yellow River, China's "mother river". Therefore, the comprehensive evaluation of the groundwater quality in Xinzhou city, the identification of prominent pollutants and areas with prominent pollution, and then the proposal of reasonable pollution control measures have become important parts of Xinzhou's current water environment research. In addition,, the research results can also provide a more accurate grasp of the overall water environment of Xinzhou city and provide a scientific basis for preventing the deterioration of the water quality in the assessment area, formulating water resources management plans, and conducting the next hydrogeological survey.

Common methods for groundwater quality evaluation include the single factor index method, the F value scoring method [6], the fuzzy comprehensive evaluation method (FCE) [7], and the grey system evaluation method [8]. These traditional methods mostly refer to the water quality category standard, calculate the weight of conventional water resource pollution factors based on multivariate statistics, and establish comprehensive evaluation indexes [9-10]. These methods are simple and have been widely used, but they fail to solve the uncertainty and nonlinearity of the pollutants in water quality evaluation [11], and there are certain deficiencies in each method. For example, the FCE needs to design the membership function of each evaluation index to all levels of standards and the weight of each index, and the evaluation results are easily affected by subjective factors [12]. The selection of the whitening function and the determination of the clustering weight in the grey system evaluation method are often different from each other, and the evaluation model is difficult to universally use [13]. With the development of artificial intelligence technology, artificial neural networks [14], Dempster-Shafer theory of evidence (D-S evidence theory) [15], genetic algorithms, extreme learning machines [16], SVMs and other methods have been applied due to their strong learning ability and performance. However, these methods still have some limitations in their application. For example, a neural network needs a large number of samples, the network structure is subjective and easy to learn, and the generalization ability is poor [17]; D-S evidence theory calculations are large, and the mass function is difficult to determine [18]; and the genetic algorithm has a certain dependence on the selection of the initial population, and it easily produces the premature convergence problem [19]. SVM is a kind of learning machine developed from statistical learning theory. It is based on the principle of structural risk minimization and has the ability to approximate complex nonlinear systems, a strong learning generalization ability and good classification performance [20]. It requires fewer samples and possesses convenient modeling, simple calculations, short learning and training times, and strong versatility; therefore, it can be used to solve the groundwater quality evaluation problem belonging to pattern recognition [21]. In SVM applications, its performance is directly affected by the selection of the model parameters [22]; therefore, the optimal search of the best parameters is particularly important.

Based on this, on the basis of PCA, this paper uses a particle swarm optimization-based support vector machine (PSO-SVM) to conduct a comprehensive evaluation of the groundwater quality in Xinzhou city and compares the results with those a variety of traditional water quality evaluation methods. The research results can provide certain referential value for the rational development and utilization of urban groundwater resources.

Materials and Methods

Sampling and Analysis

Based on the hydrogeological structure of Xinzhou city, this paper selects local loose rock pore water and carbonate fissure karst water as the research objects. There are 14 groundwater sampling points in the study area, including 12 diving samples and 9 confined water samples. The sampling points are all water wells in use, the depth of the diving wells is 5-40 m, and the depth of the pressurized water wells is 110-210 m. The sampling wells are basically evenly distributed throughout the study area. GPS is used to locate each sampling point. The distribution of the sampling points is shown in Fig. 1. Among them, 1#~12# correspond to the upper pore water, and 13#~21# correspond to the lower karst water.

The water samples were collected in accordance with the requirements of the "Technical Specifications for Groundwater Environmental Monitoring" (HJ/T 164-2004). The collected water samples were filtered through a 0.45 μm filter membrane and collected in precleaned and sterilized 5 L polyethylene bottles. The water samples were divided into two parts: one part was used for the anion test without any reagents, and the other part was used for cation analysis by adding premium grade HNO_3 to a pH value of less than 2. After the water samples were sealed, they were transported back to the laboratory within 24 h and stored at 0-4ºC, and the determination was completed within 48 h. Fig. 2 shows photos of the field test.

The test items of the water samples taken are mainly based on indicators that have potential risks to the health of local residents, including the following: nitrite (calculated by N), nitrate (calculated by N), sulfate, chloride, fluoride, Fe, total dissolved solids (TDS) and total hardness (TH). The soluble metal Fe in the water sample after acidification was analyzed by inductively coupled plasma mass spectrometry

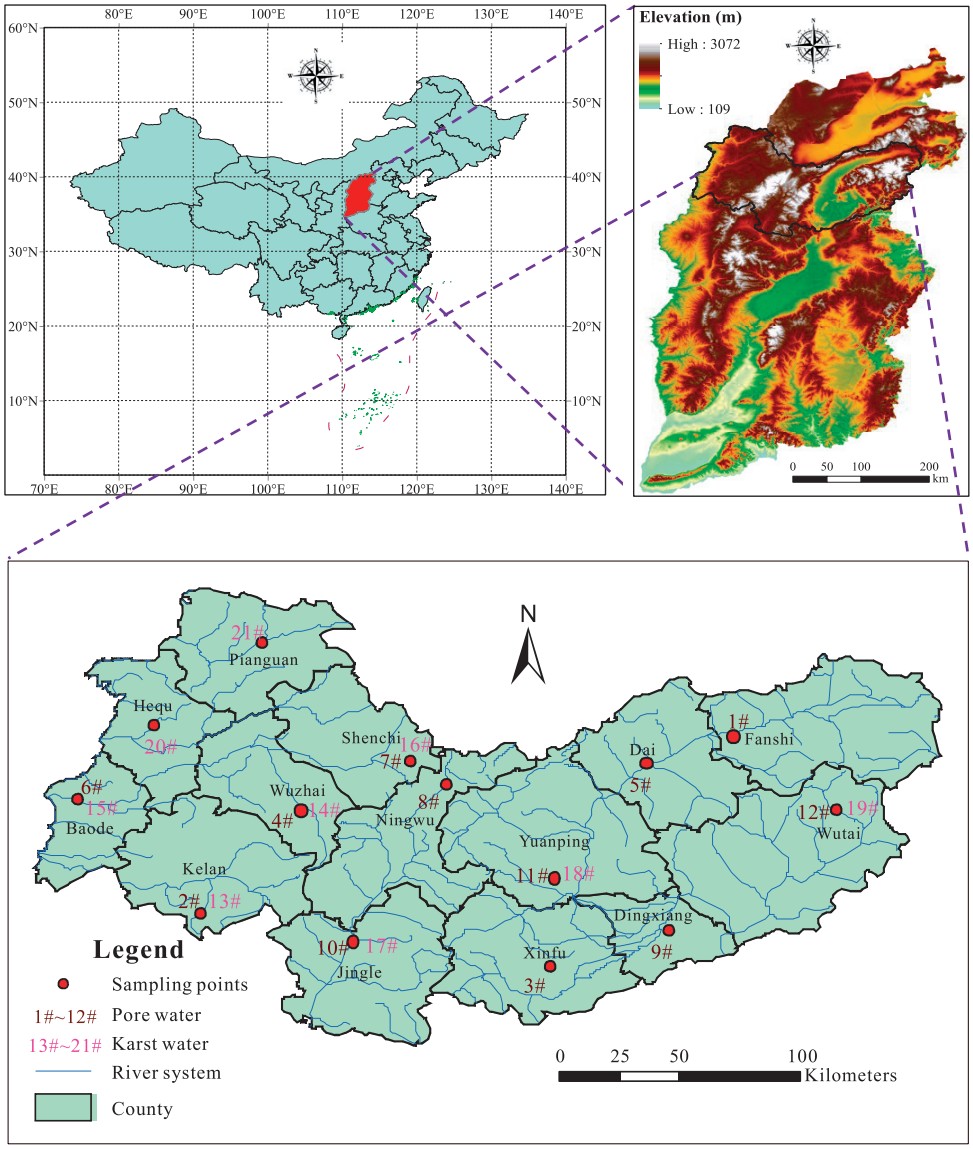

Fig. 1. Geographical location of Xinzhou city.

(ICP-MS) (Themofisher XII, USA). Cl^-, SO_4^{2-} and F^- in unacidified water samples were determined by ion chromatography (Metrohm 761/813). The concentrations of NO_2^- and NO_3^- in the water samples were determined by a DR2800 portable spectrophotometer. TDS was measured by the gravimetric method, and the total hardness (TH) was determined by EDTA titration. All instruments were corrected before the experiment, and the average analysis error of ICP-MS was less than ±10%. The indexes of the water samples were determined by the laboratory of the Institute of Geology and Geophysics, Chinese Academy Sciences.

Data Analysis Techniques

The charge balance error of anions and cations in all test water samples should be guaranteed to be within 5% [23]. Through calculation, the charge balance error of anions and cations in water samples is less than 5%, indicating that the detection accuracy of water chemistry is high.

The water quality grade evaluation adopted the "Groundwater Quality Standard" (GB/T14848-2017). The data not detected in the test were replaced by 1/2 of the detection limit of the instrument. The analysis of the anion and cation charge balance error of each water sample was completed using SPSS 26.0; the water quality evaluation method was completed using MATLAB R2018a and Excel 2016; and illustrations were made by software such as Origin 2019b, ArcMap 10.7 and Surfer 17.0.

Water Quality Evaluation Method

This paper uses the PCA-PSO-SVM algorithm to evaluate groundwater quality data. The basic idea of the algorithm is as follows: First, considering that water sample test indicators may cause collinearity due

Fig. 2. Field test photos a) Drilling, b) Installing PVC pipe, c) Filling with filter material, d) Washing well before sampling, e) Sample collection, and f) Parameter determination.

to their diversity, this paper uses principal component analysis (PCA) to reduce the dimensionality of the multiple water sample test indicators. Second, PSO is used to optimize the selection of the key parameters of SVM (kernel function parameter g and penalty factor C) to construct a PSO-SVM model. Finally, the water sample test data after the dimensionality reduction process are inserted into the PSO-SVM model for training and testing, and the results are output and analyzed.

PCA

PCA obtains new variables with fewer dimensions that are irrelevant by constructing the linear combination of the original variables. In other words, it uses the dimension reduction idea to reduce multiple indicators to a few independent comprehensive indicators. In this process, the information overlap between different indexes is fully considered. On the basis of retaining the original information as much as possible, the dimension reduction of multidimensional data is conducted, and the independent comprehensive factors are selected more objectively. Subjective arbitrariness is avoided, and the data structure is simplified. The intuitiveness of the analysis is greatly improved, and the running time of the program is greatly reduced.

SVM

The basic principle of SVM is to minimize structural risks. The idea is to find an optimal superflat in the sample space and divide the sample space into two categories so that the distance between the two-class sample set and the optimal superflat is the largest [24]. Before using SVM, the classification problem is divided into two types: one is linearly separable, and the other is linearly inseparable. Most of the actual problems are the latter.

We introduce slack variables and increase the kernel function K(x, x_j) and penalty factor C for analysis. The input data in low-dimensional space are transformed into high-dimensional space through nonlinear transformation, making it a linear sample, so as to find the optimal classification hyperplane.

The solution of the hyperplane is transformed into an optimization problem:

$$\begin{cases} \min Q(a) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K\left(x_i, x_j\right) - \sum_{i=1}^{n} a_i \\ s.t.\ \sum_{i=1}^{n} y_i a_i = 0 \quad 0,\ a_i,\ c \end{cases} \tag{1}$$

This paper selects the RBF radial basis kernel function $K\left(x, x_j\right) = \exp\left(-\gamma\left\|x - x_j\right\|^2\right)$, γ>0 and solves $\alpha^*, \mathrm{b}^* = y_i - \sum_{i=1}^{l} y_i \alpha_i^* K\left(x_i, x_j\right)$.

The discriminant function is:

$$f(x) = sng\{(\omega \cdot x) + b\} = \operatorname{sgn}\left\{\sum_{i=1}^{n} \alpha_i^* y_i k\left(x_i \cdot x\right) + b^*\right\}.$$

PSO

PSO was originally derived from the ideas of artificial life and evolutionary computing theory. It was first proposed by two American doctors, James Kennedy and Russell Eberhart, and imagined such a scenario. A group of birds randomly search for food in a fixed area. There is only one piece of food in this area, and all the birds do not know where the food is; however, they know how far the piece of food is from their current location. Therefore, in order to search for food better and faster, the birds farther from the food need to keep getting closer to the birds closer to the food. Each bird is constantly searching for the closest bird to itself, and the final result is that the entire flock of birds flies to unknown food under one control. The particle swarm algorithm is inspired by this kind of bird predation behaviour and used in optimization problems.

(1) Mathematical description of the particle swarm algorithm

The particle swarm algorithm is a kind of bionic evolutionary computing technology. The mathematical description of the PSO process is as follows [25]:

Assuming there is an N-dimensional space, the position vector of the *i*th particle in the N-dimensional space is:

$$x_i = \left(x_{i1}, x_{i2}, \ldots, x_{iN}\right) \tag{2}$$

The flight speed vector is:

$$v_i = \left(v_{i1}, v_{i2}, \ldots, v_{iN}\right) \tag{3}$$

Each particle can calculate the corresponding fitness value according to the fitness function, update their position and speed in N-dimensional space according to their own and group experience, and constantly adjust their moving speed and position by comparing the fitness value of the new position, individual extremum and group extremum. The updated formula is:

$$v_i^{k+1} = \omega v_i^k + c_1 r_1\left(p_i - x_i^k\right) + c_2 r_2\left(g_i - x_i^k\right) \tag{4}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{5}$$

where ω is the inertia weight; i = 1, 2..., N is the size of the particle population; k is the number of iterations; p_i is the best local position of particles, also known as the individual extrema; g_i is the best position of the whole population at present, also known as the global extremum; r_1 and r_2 are random numbers that are uniformly distributed from [0,1]; and c_1 and c_2 are learning factors, usually between 0 and 2. In the velocity update formula, v_i^k represents the size and direction of the velocity of the particle in the previous iteration; and $c_1 r_1(p_i - x_i^k)$ belongs to the "cognitive" part of the particle, namely, the distance vector between the current position and its optimal position. $c_2 r_2(g_i - x_i^k)$ is the distance vector between the current position of the particle and the optimal neighbour, which can reflect the cooperation, information and sharing ability between particles.

The particle swarm algorithm is an optimization algorithm, and the fitness function refers to the optimization objective function in the primary algorithm. The fitness function directly reflects the individual's survivability and adaptability to the environment. In general, the larger the fitness value is, the stronger the individual's adaptability, and the smaller the fitness value is, the weaker the individual's adaptability.

(2) Comparison of PSO and other optimization algorithms

Traditional optimization algorithms can solve simple problems, but for nonlinear and complex problems, the optimization time is often too long. However, modern optimization algorithms are an important branch of artificial intelligence and can solve optimization problems well. Among the modern optimization algorithms, the most important intelligent optimization algorithms include the particle swarm algorithm, genetic algorithm, ant colony algorithm, evolutionary programming algorithm, etc. These optimization algorithms have many similarities but also have obvious differences.

Compared with the genetic algorithm (GA), the optimization of the particle swarm algorithm does not crossover and mutate particles. The algorithm is simple, and the next position is obtained through a global and local search mode; therefore, the search speed is relatively fast and the efficiency is high. The ant colony algorithm (ACO) is the same as the particle swarm algorithm. It only needs to calculate the size of the fitness value for selection. The main problem is that the implementation of the ant colony algorithm is complex, which will cause stagnation in the search process; however, the particle swarm algorithm will not cause such a problem. The iterative process of the evolutionary programming algorithm (EP) is affected by the random function and may be solved in any direction; however, the particle swarm algorithm will move based on the local optimum and the group optimum in the iterative process, making it smarter, rather than unconsciously bumping.

Various algorithms have their own strengths in solving optimization problems, but the application of PSO in practical applications is more flexible. Its advantages lie in its simple implementation and powerful functions, and it is feasible in terms of the

parameter optimization of mathematical methods for groundwater quality evaluation [26, 27].

PCA-PSO-SVM

The specific operating steps of the PCA-PSO-SVM algorithm are as follows:

(1) Dimension reduction of influence factors

Since each class of water quality assessment level can only correspond to a set of water pollution index limits (i.e., critical values), only five pairs of training samples can be provided according to the standard requirements. This is far from meeting the sample requirements of SVM. Due to the lack of training samples, the data cannot reasonably reflect the internal rules of water quality, which will cause the shortcomings of the poor generalization ability, low recognition accuracy and weak robustness of the evaluation model and decrease the model's practical value. Therefore, the number of training samples must be increased. Combined with the test results of water samples, the standard threshold of class V was set to be twice that of Class IV. Each water quality level randomly generates 100 8-dimensional water quality data points, for a total of 500 data points. We normalize the initial data and then set the training set (400) and test set (100) according to the ratio of 4:1.

Based on the aforementioned PCA principle, we use the auxiliary function (pcaForSVM.m) in the libsvm-3.1 [Faruto Ultimate3.1 Mcode] toolbox [28] of MATLAB to conduct dimensionality reduction preprocessing, and the function interface program is [train_pca, test_pca]=pcaForSVM(train, test, threshold). By entering the training data (train), test data (test), and threshold parameters (i.e., cumulative variance contribution rate, which is 85% in this example), the reduced training data (train_pca) and test data (test_pca) can be obtained. The principal components obtained after dimensioning are used for subsequent analysis.

(2) Parameter optimization based on PSO

The selection of kernel function parameter g and penalty factor C is very important to the classification performance of SVM, and there are many ways to optimize the key parameters (kernel function parameter g and penalty factor C) of SVM. The most common algorithm is the grid optimization algorithm. Other methods include genetic algorithms and particle swarm algorithms. In this paper, in order to improve the speed of model parameter optimization and to obtain more reasonable optimal parameters, PSO is selected for parameter optimization. The main process can be summarized into the following two steps:

1) Make any particle close to the individual and obtain the global optimal solution (q_{best} and p_{best}), and

2) Iteratively update q_{best} and p_{best} to obtain the global optimal solution.

The SVM model is mainly established using the LIBSVM toolbox, and the interface program model=svmtrain (training_label_vector, training_instance_matrix, ['libsvm_options']) is called and mainly used for training the model. Then, [predicted_label, accuracy, r]=svmpredict (test_label_vector, test_instance_matrix, model), which is mainly used to predict data, is called. The input training data are the extracted impact factor matrix and label (grade 5 water quality level) with 400 training samples.

PSO-SVM mainly uses the auxiliary function (psoSVMcgForClass.m) in the libsvm-3.1 [Faruto

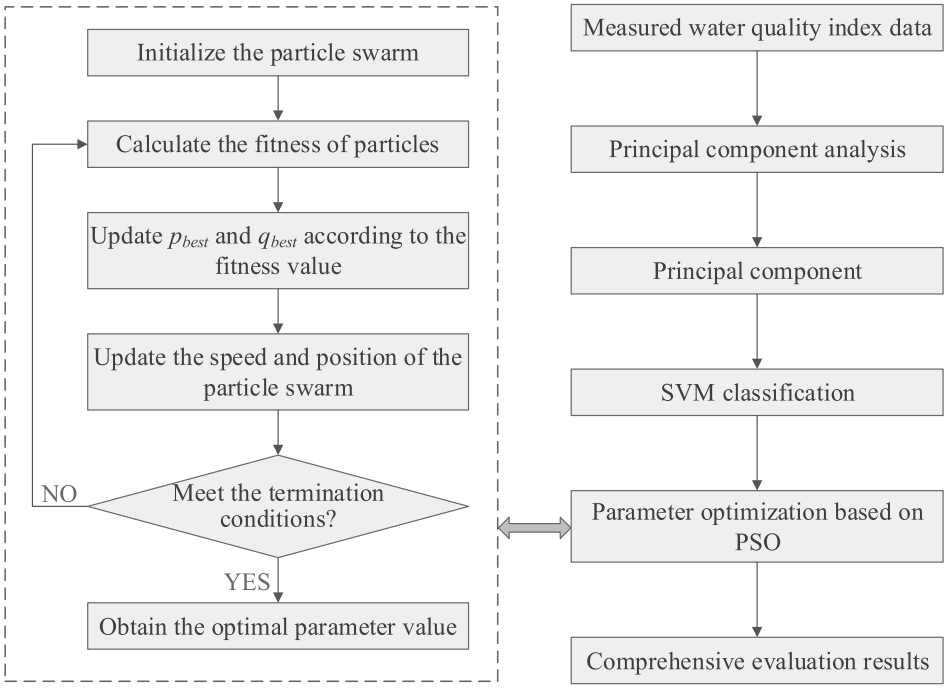

Fig. 3. Flowchart of the research methodology.

Table 1. Test results of pore water and karst water samples.

NO.	NO_2^- (mg/L)	NO_3^- (mg/L)	SO_4^{2-} (mg/L)	Cl^- (mg/L)	F^- (mg/L)	Fe^{3+} (mg/L)	TDS (mg/L)	TH (mg/L)
1#	0.011	13.300	48.03	23.16	0.2	0.80	346	200
2#	0.003	11.800	40.83	21.38	0.3	1.00	310	233
3#	0.003	19.700	24.01	12.47	0.5	0.10	267	200
4#	0.003	17.300	16.81	14.25	0.2	1.10	267	193
5#	0.002	18.800	69.64	17.73	0.5	0.20	341	273
6#	0.066	25.100	288.18	146.07	0.6	0.05	942	388
7#	0.016	35.800	4.80	14.18	0.5	0.05	319	255
8#	0.001	27.300	158.50	23.16	0.2	0.05	463	365
9#	0.005	4.000	21.61	17.81	0.4	0.05	323	203
10#	0.004	1.300	45.63	12.47	0.2	0.05	267	223
11#	0.019	24.000	86.45	19.59	0.3	0.05	366	253
12#	0.005	13.600	28.82	10.69	0.3	0.05	247	200
13#	0.013	17.100	12.01	12.47	0.2	0.05	237	193
14#	0.002	19.600	16.81	14.18	0.2	0.50	255	198
15#	0.013	5.900	96.06	33.85	0.6	0.05	404	300
16#	0.002	35.600	24.02	19.59	0.4	0.10	358	288
17#	0.034	25.100	45.63	17.73	0.2	0.05	306	248
18#	0.024	6.400	175.31	8.91	0.1	0.05	410	320
19#	0.005	3.300	379.44	8.91	0.5	0.05	713	570
20#	0.014	12.300	86.45	53.44	0.3	0.05	420	290
21#	0.005	12.400	19.21	10.69	0.5	0.05	244	195

Ultimate 3.1 Mcode] toolbox in MATLAB to optimize the kernel function parameter g and penalty factor C. The function interface program is [bestCVaccuracy, bestc, bestg, pso_option]=psoSVMcgForClass (train_label, train, pso_option). The optimal C and g can be obtained by inputting training data (train), training data labels (train_label) and PSO parameters.

The flowchart of the PCA-PSO-SVM algorithm shown in Fig. 3.

Results and Discussion

General Results

The specific monitoring values of the pore water and karst water samples are shown in Table 1. Based on these data, SVM and PCA-PSO-SVM are used to evaluate water quality.

SVM

There is no universally acknowledged best method to optimize the SVM parameters. At present, the commonly used method is to let the values of C and g be in a certain range. For the selected C and g, the training set is used as the original data set, and the accuracy of training set verification classification is obtained by using the K-fold Cross Validation (K-CV) method [29]. Finally, C and g that resulted in the highest accuracy of training set verification classification were selected as the best parameters. This section uses the 3-fold cross validation method.

The cross-validation method is adopted. First, the ranges of C and g are set as 2^{-12}, 2^{-9}, ... , 2^{10}; and the optimal values of parameters C and g are preliminarily roughly obtained as C = 0.000097656 and g = 0.75786. Then, the ranges of C and g are narrowed and refined. The ranges of C and g are set as 2^{-12}, $2^{-11.5}$, ... , 2^{4}. Finally, the optimal parameters are C = 0.00024414 and g = 0.0625, and the classification accuracy of the test set is 90%. In this case, the parameters used are the best in a certain sense.

PCA-PSO-SVM

The PCA-PSO-SVM is used to conduct water quality evaluation. First, the relevant data are

standardized to eliminate the influence of dimensions and orders of magnitude between different indicators. Kaiser-Meyer-Olkin (KMO) test statistics and Bartlett sphericity tests were used to determine the correlation between indicators [30] to determine whether the original variables are suitable for factor analysis. The calculated KMO value is 0.959, which means that the data be used for factor analysis. When the significance of the Bartlett sphericity test is 0<0.05, there is a correlation between the original variables, which means that the variables can be analyzed by principal component analysis.

By calculating and selecting the principal components whose eigenvalues are greater than or equal to 1, it is found that the cumulative variance contribution rate of the first four principal components reaches 88.76%, which can basically reflect the attributes of the source data. The distribution of the gravel map and principal component analysis double-plot is shown in Fig. 4.

By dividing the data in the loading matrix of the initial factor of the principal component by the square root of the corresponding eigenvalue of the principal component, the eigenvector corresponding to the principal component is obtained. That is, the corresponding coefficient of each index, multiplied by the standardized data, and the corresponding expression of the principal component F can be obtained (Eq. (6), Eq. (7), Eq. (8), and Eq. (9)).

$$F_1 = 0.36162X_1 + 0.03015X_2 + 0.44315X_3 + 0.39077X_4 + 0.25658X_5 - 0.19784X_6 + 0.49581X_7 + 0.41041X_8 \quad (6)$$

$$F_2 = 0.40867X_1 + 0.62685X_2 - 0.3603X_3 + 0.38797X_4 + 0.0961X_5 + 0.02704X_6 - 0.04031X_7 - 0.38485X_8 \quad (7)$$

$$F_3 = 0.22325X_1 - 0.20278X_2 + 0.12675X_3 + 0.24797X_4 - 0.52942X_5 + 0.72414X_6 + 0.15273X_7 - 0.05922X_8 \quad (8)$$

$$F_4 = -0.17943X_1 - 0.35094X_2 - 0.13672X_3 + 0.23091X_4 + 0.74542X_5 + 0.42462X_6 + 0.03234X_7 - 0.18876X_8 \quad (9)$$

where F_i is the ith principal component; and X_i (i = 1~8) are the concentrations of NO_2^-, NO_3^-, SO_4^{2-}, Cl^-, F^-, Fe^{3+}, TDS and TH, respectively.

The principal component formula shows that the indexes closely related to the first principal component are TDS and TH, which mainly reflect the total amount of dissolved solids and total hardness, respectively, and can characterize the overall degree of pollution of the water body. Moreover, since the variance contribution rate of the first principal component reaches 48.42%, which is much larger than that of other principal components, the first principal component plays a decisive role in the evaluation of the water quality. Among the second principal components, the loadings of NO_2^- and NO_3^- are relatively large, which can indicate the degree of pollution of the water body by nitrogen salts. The index closely related to the third principal component is Fe^{3+}, which mainly reflects the influence of Fe^{3+} on water bodies. The index closely related to the fourth principal component is F^-, which mainly reflects the degree of dissolution of fluorite in the groundwater aquifer.

Second, PSO is used to optimize the parameters of SVM. The parameters that need to be optimized are the following: the kernel function g parameter and penalty factor C. The kernel function parameter g and penalty factor C of SVM are used as the particles of PSO, and the classification accuracy of SVM is used as the objective function of PSO. The parameters of PSO are set as follows: the maximum number of evolutionary generations T is 200, the population popsize is 20, and the learning factors are c1 = 1.5 and c2 = 1.7. The support vector classification machine is selected, and the radial basis kernel function is used as the kernel function of the support vector classification machine. The value interval of the penalty parameter c is set to [0.1, 100], and the value interval of the radial basis kernel function g is set to [0.01, 100]. The three-fold cross validation method is used. Through PSO-

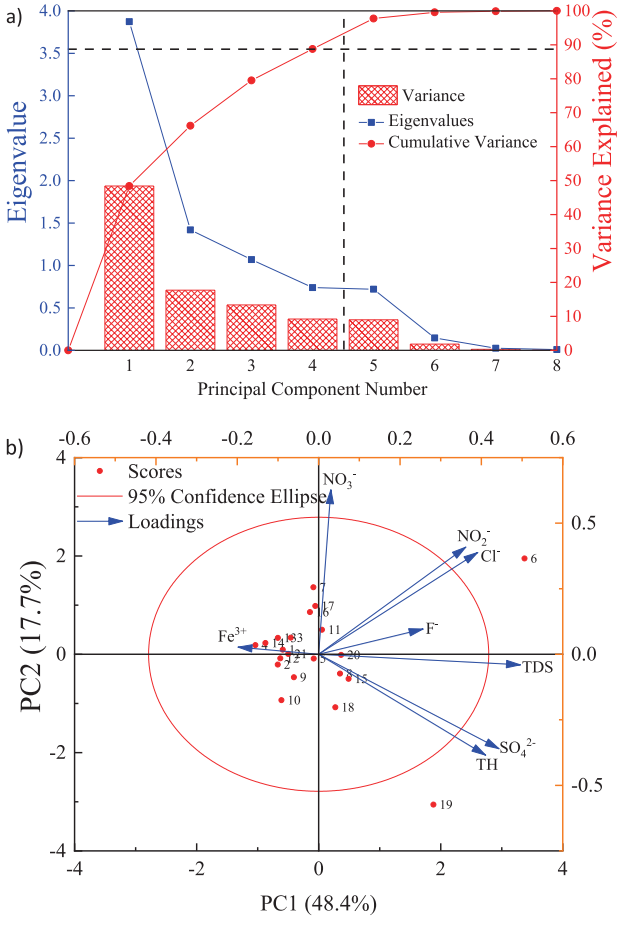

Fig. 4. a) Scree plot and b) Biplot.

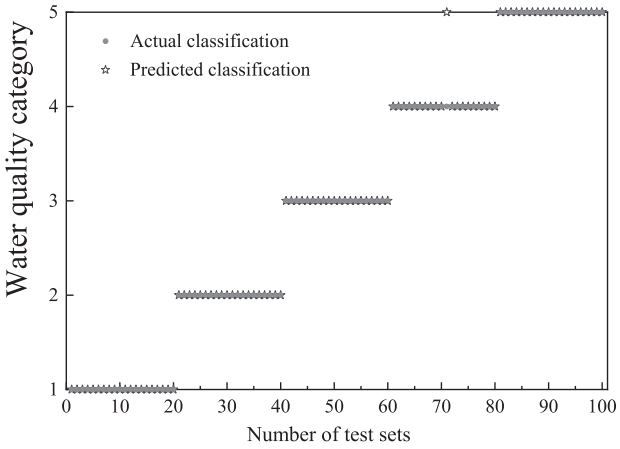

Fig. 5. Classification results of the actual test set and predicted test set.

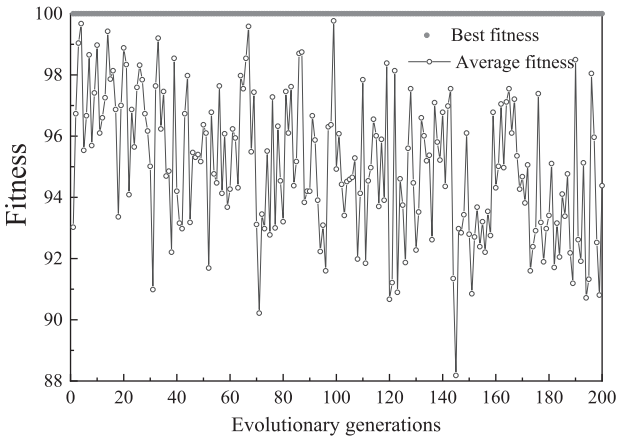

Fig. 6. Fitness curve.

SVM, the optimal parameter combination is obtained as follows: C = 0.1 and g = 0.01. The optimal cross-validation accuracy is 99%. The classification results are shown in Fig. 5, and the obtained fitness curve is shown in Fig. 6.

The final classification results are shown in Table 3.

Comparison of the Results

First, part of the data obtained by interpolation between standards at all levels was selected as the verification set data, and the accuracy of each method was tested. The results are shown in Table 2. Since FCE, the F value scoring method and the single factor index method have no training process, of course, there is no corresponding verification process; therefore, these methods are not included in the following table.

The table shows that the accuracy of water quality prediction based on SVM is significantly higher than that of the BP neural network, and the SVM method based on PSO can meet the demand of water quality prediction, regardless of whether the principal component extraction is used or not. The advantages of

Table 2. Comparison of the performance of each method

Methods	Total number	Number of correct evaluations	Accuracy
PCA-PSO-SVM	100	99	99.00%
PSO-SVM	100	96	96.00%
SVM	100	90	90.00%
BP neural network	100	52	52.00%

PCA-PSO-SVM over PSO-SVM are mainly reflected in the running time.

Liao, Xu and Wang proposed a water quality evaluation method based on the combination of SVM and a genetic algorithm with an average prediction accuracy of more than 80% [31]. Modaresi and Arghinejad studied and compared the water quality classification performance of three supervised classification methods: SVM, the probabilistic neural network (PNN), and k-nearest neighbours (KNN). Among these three methods, SVM has the best performance with an average accuracy rate of 90.6% in the test of 5 data sets [32]. In comparison, the water quality evaluation method adopted in this paper has certain advantages in evaluation accuracy and has promotional value.

Table 3. Summary of the results of each methoIn addition, according to the calculated water quality evaluation grade (Table 3), the water quality distribution map of each layer of water is drawn, as shown in Fig. 7.

The figures show that there is a certain correlation between the upper pore water and the lower karst water in the central and western water quality levels, and the eastern region has a large contrast, which is inconsistent with the conventional understanding that the lower water is usually less disturbed and the water quality of lower water is better than that of the upper water. Through observation, this may be because in the karst water aquifer, the area near Wutai is large, and there are few sampling points. When the kriging interpolation method is used, the water quality around Wutai is extrapolated and interpolated based on Wutai. Therefore, the water quality around Wutai has a strong correlation with the results, which also causes poor water quality in the eastern karst water aquifer. Therefore, in future research, the number of monitoring points and monitoring frequency should be increased.

Discussion

Based on the data of 21 groundwater monitoring water samples in Xinzhou city, the improved SVM method was used to comprehensively evaluate the groundwater quality, and the results were compared with the results a variety of water quality evaluation models. The evaluation results of the single factor index method and the F value score method are relatively close, the results are better than those of the other evaluation

Table 3. Summary of the results of each method

NO.	PCA-PSO-SVM	PSO-SVM	SVM	BP neural network	FCE	F value scoring method	Single factor index method
1#	II	II	II	II	III	IV	IV
2#	II	II	II	II	I	IV	IV
3#	II	II	II	II	I	II	III
4#	II	II	II	II	III	IV	IV
5#	III	III	III	II	I	II	III
6#	V	V	V	III	III	IV	IV
7#	III	III	III	III	V	IV	V
8#	III	III	III	III	II	IV	IV
9#	II	II	II	I	I	II	III
10#	I	I	I	I	I	II	III
11#	III	III	III	III	I	IV	IV
12#	II	II	II	I	I	II	III
13#	I	I	I	I	I	II	III
14#	II	II	II	II	III	IV	IV
15#	III	III	III	II	II	II	III
16#	III	III	III	III	V	IV	V
17#	III	III	III	III	I	IV	IV
18#	III	III	III	III	II	II	III
19#	IV	IV	IV	III	V	V	V
20#	III	III	III	III	II	II	III
21#	II	II	II	I	I	II	III

models, and the evaluations are relatively conservative. This is because the single factor evaluation method determines the water quality category using the single index with the worst water quality, so the water quality category evaluated is inferior. The F value scoring method can reflect the overall situation of water quality, but the evaluation results highlight the level of excessive pollution indicators, and the evaluation results are not continuous. Compared with the results of the single factor index method and F value score method, the results of FCE are more volatile, which may be caused by the influence of FCE in which the pollution factors in a few monitoring indexes exceed the standards. In addition, this algorithm fully considers the fuzziness of the boundaries of different water quality categories and the influence of evaluation factors on the water quality weights in the calculation process, and the evaluation results are the results of the interaction between the two.

SVM and improved SVMs (PSO-SVM and PCA-PSO-SVM) in the evaluation results are consistent. The results are better than the results of the traditional evaluation methods (single factor index method and F value score method), more consistent with the actual research, and more objective. The evaluation results of the BP neural network are slightly different from those of SVM and the improved SVM models (PSO-SVM and PCA-PSO-SVM), which are basically consistent. However, the artificial intervention of the BP neural network structure is large, and the verification accuracy of the results is low. When only a single method is used for evaluation, the credibility of the results is general. As a comparison, SVM and the improved SVM models (PSO-SVM and PCA-PSO-SVM) have high accuracy, but the SVM has certain blindness in searching the best parameters based on cross-validation, and it takes a certain amount of time. PSO-SVM compensates for this defect, and the performance of the model is greatly improved. However, in addition, there are shortcomings such as low computational efficiency and large memory occupation. In contrast, the PCA-PSO-SVM algorithm established in this paper improves the operating efficiency of the program and reduces the amount of data occupying memory under the premise of ensuring the accuracy by reducing the dimensionality. The application effect is good.

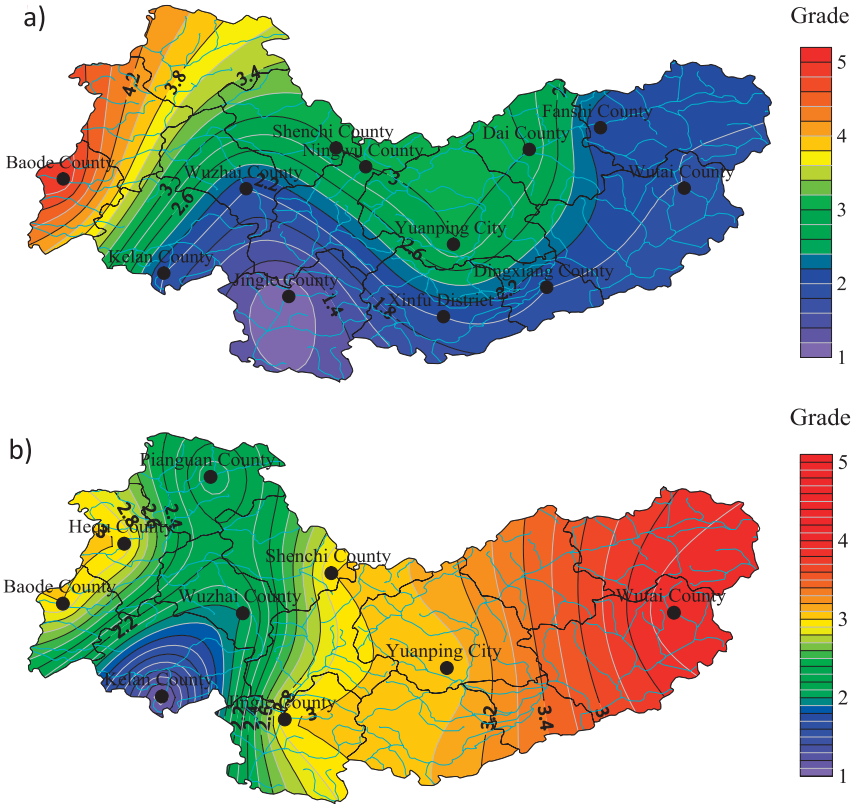

Fig. 7. Water quality distribution map: a) Pore water, b) Karst water.

Since the training set of the various new artificial intelligence algorithms selected in this paper is composed of random data, the corresponding training models are different, which causes each test result to not be static, although this change is small. In addition, in this paper, the threshold range point of the class V water quality is set to twice that of class IV water quality, which has certain applicability and representativeness. However, for poor water quality, it is far beyond the threshold range point of class V water quality; and although it exceeds the standard of class V water quality, the two situations of small exceeding range are generally applicable and the training effect is general. Therefore, more appropriate threshold ranges should be set for different research water bodies.

Conclusions

(1) The overall condition of the groundwater quality in Xinzhou city is generally good. The upper pore water quality is mainly class II, the lower karst water is mainly of class III, and the water quality shows significant spatial differences.

(2) Overall, the water quality in Kelan, Wuzhai, and Shenchi is relatively good while the water quality in Baode and Wutai is poor; and water quality management is urgently needed. However, the water quality of Yuanping, Xinfu, and Dingxiang fluctuates greatly, and the monitoring frequency should be increased to treat pollution sources in time.

(3) PCA-PSO-SVM is more comprehensive and objective in the evaluation results and has high accuracy and calculation efficiency. It is worthy of promotion and use.

Acknowledgments

This research was supported by the project "Compilation of Comprehensive Hydrogeological Map and Spatial Hydrogeological Information System in Xinzhou City, Shanxi Province". We sincerely thank the anonymous reviewers for their time and effort devoted to improving the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

1. XIA J., ZUO Q.T. Utilization and Protection of Water Resources in China: 1978=2018. Chinese Journal of Urban and Environmental Studies, **6** (04), 1850022, **2018**.
2. WANG Y.X., WANG Y., SU X.L., QI L., LIU M. Evaluation of the comprehensive carrying capacity of

interprovincial water resources in China and the spatial effect. Journal of Hydrology, **575**, 794, **2019**.
3. JIA Z.M., CAI Y.P., CHEN Y., ZENG W.H. Regionalization of water environmental carrying capacity for supporting the sustainable water resources management and development in China. Resources, Conservation and Recycling, **134**, 282, **2018**.
4. ZHAO L., HOU H., SHANGGUAN Y.X., CHENG B., XU Y.F., ZHAO R.F., ZHANG Y.G., HUA X.Z., HUO X.L., ZHAO X.F. Occurrence, sources, and potential human health risks of polycyclic aromatic hydrocarbons in agricultural soils of the coal production area surrounding Xinzhou, China. Ecotoxicology and Environmental Safety, **108**, 120, **2014**.
5. DAI D., XU X.Q., SUN M.D., HAO C.L., LV X.B., LEI K. Decrease of both river flow and quality aggravates water crisis in North China: A typical example of the upper Yongding River watershed. Environmental Monitoring and Assessment, **192**, 1, **2020**.
6. MUHAREMI F., LOGOFĂTU D., LEON F. Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication, **3** (3), 294, **2019**.
7. LIU L., ZHOU J.Z., AN X.L., ZHANG Y.C., YANG L. Using fuzzy theory and information entropy for water quality assessment in Three Gorges region, China. Expert Systems with Applications, **37** (3), 2517, **2010**.
8. YAN F., QIAO D.Y., QIAN B., MA L., XING X.G., ZHANG Y., WANG X.G. Improvement of CCME WQI using grey relational method. Journal of Hydrology, **543**, 316, **2016**.
9. YUAN Y., WANG J. Advanced Grey Relational Analysis Method and Its Application in Water Quality Evaluation of the Lake-type Wetland. Journal of Landscape Research, **9** (4), **2017**.
10. WU Z.S., ZHANG D.W., CAI Y.J., WANG X.L., ZHANG L., CHEN Y.W. Water quality assessment based on the water quality index method in Lake Poyang: The largest freshwater lake in China. Scientific reports, **7** (1), 1, **2017**.
11. SUN Y., LIANG X.J., XIAO C.L. Assessing the influence of land use on groundwater pollution based on coefficient of variation weight method: a case study of Shuangliao City. Environmental Science and Pollution Research, **26** (34), 34964, **2019**.
12. ASTUTI A.D., ARIS A., SALIM M.R., AZMAN S., SALMIATI., SAID M.I.M. Artificial Intelligence Approach to Predicting River Water Quality: A Review. Journal of Environmental Treatment Techniques, **8** (3), 1093, **2020**.
13. LI R.R., ZOU Z.H., AN Y. Water quality assessment in Qu River based on fuzzy water pollution index method. Journal of environmental sciences, **50**, 87, **2016**.
14. AL-OBAIDI B.H., MAHMOOD R.S., KADHIM R.A. Water quality assessment and sodium adsorption ratio prediction of tigris river using artificial neural network. Journal of Engineering Science and Technology, **15** (5), 3055, **2020**.
15. YANG L.K., ZHAO X.H., PENG S., LI X. Water quality assessment analysis by using combination of Bayesian and genetic algorithm approach in an urban lake, China. Ecological Modelling, **339**, 77, **2016**.
16. YU T.T., YANG S., BAI Y., GAO X., LI C. Inlet water quality forecasting of wastewater treatment based on kernel principal component analysis and an extreme learning machine. Water, **10** (7), 873, **2018**.
17. ISIYAKA H.A., MUSTAPHA A., JUAHIR H., PHIL-EZE P. Water quality modelling using artificial neural network and multivariate statistical techniques. Modeling Earth Systems and Environment, **5** (2), 583, **2019**.
18. LI L., JIANG P., XU H., LIN G., GUO D., WU H. Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China. Environmental Science and Pollution Research, **26** (19), 19879, **2019**.
19. SOTOMAYOR G., HAMPEL H., VÁZQUEZ R.F. Water quality assessment with emphasis in parameter optimisation using pattern recognition methods and genetic algorithm. Water research, **130**, 353, **2018**.
20. SHAN W., CAI S.S., LIU C. A new comprehensive evaluation method for water quality: improved fuzzy support vector machine. Water, **10** (10), 1303, **2018**.
21. ABOBAKR YAHYA A.S., AHMED A.N., BINTI OTHMAN F., IBRAHIM R.K., AFAN H.A., EL-SHAFIE A., FAI C.M., HOSSAIN M.S., EHTERAM M., ELSHAFIE A. Water quality prediction model based support vector machine model for Ungauged River catchment under dual scenarios. Water, **11** (6), 1231, **2019**.
22. KAMYAB-TALESH F., MOUSAVI S.F., KHALEDIAN M., YOUSEFI-FALAKDEHI O., NOROUZI-MASIR M. Prediction of Water Quality Index by Support Vector Machine: a Case Study in the Sefidrud Basin, Northern Iran. Water Resources, **46** (1), 112, **2019**.
23. EL BABA M., KAYASTHA P., HUYSMANS M., DE SMEDT F. Evaluation of the Groundwater Quality Using the Water Quality Index and Geostatistical Analysis in the Dier al-Balah Governorate, Gaza Strip, Palestine. Water, **12** (1), 262, **2020**.
24. LEONG W.C., BAHADORI A., ZHANG J., AHMAD Z. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). International Journal of River Basin Management, 1-8, **2019**.
25. AGHEL B., REZAEI A., MOHADESI M. Modeling and prediction of water quality parameters using a hybrid particle swarm optimization-neural fuzzy approach. International Journal of Environmental Science and Technology, **16** (8), 4823, **2019**.
26. LI M.S., WU W., CHEN B.S., GUAN L.X., WU Y. Water quality evaluation using back propagation artificial neural network based on self-adaptive particle swarm optimization algorithm and chaos theory. Computational Water, Energy, and Environmental Engineering, **6** (03), 229, **2017**.
27. JAHANDIDEH-TEHRANI M., BOZORG-HADDAD O., LOÁICIGA H.A. Application of particle swarm optimization to water management: an introduction and overview. Environmental Monitoring and Assessment, **192**, 1, **2020**.
28. LIBSVM: a library for support vector machines, 2001. Available online: http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed on 31/12/2020).
29. LI W., YANG M.Y., LIANG Z.W., ZHU Y., MAO W., SHI J.Y., CHEN Y.X. Assessment for surface water quality in Lake Taihu Tiaoxi River Basin China based on support vector machine. Stochastic environmental research and risk assessment, **27** (8), 1861, **2013**.
30. POPUGAEVA D., KREYMAN K., RAY A.K. Assessment of Khibiny Alkaline Massif groundwater quality using statistical methods and water quality index. The Canadian journal of chemical engineering, **98** (1), 205, **2020**.

31. LIAO Y., XU J.Y., WANG Z.W. Application of biomonitoring and support vector machine in water quality assessment. Journal of Zhejiang University Science B, **13** (4), 327, **2012**.

32. MODARESI F., ARAGHINEJAD S. A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water resources management, **28** (12), 409, **2014**.