Mine Water and the Environment (2022) 41:1106–1117
https://doi.org/10.1007/s10230-022-00884-5

TECHNICAL COMMUNICATION

# Source Discrimination of Mine Water Inrush Using Multiple Combinations of an Improved Support Vector Machine Model

Zhonglin Wei[1] · Donglin Dong[1] · Yuan Ji[1] · Jia Ding[1] · Lujia Yu[1]

Received: 1 January 2022 / Accepted: 10 June 2022 / Published online: 17 September 2022

## Abstract

Accurate and timely identification of water inrush sources is important in controlling mine water hazards. We combined various algorithms to offer an improved inrush source identification method. The algorithms include the Fisher identification method, self-organizing correlation method (SOM), improved principal component analysis method (PCSOM), and grey wolf algorithm (GWO) optimized support vector machine (GWOSVM). The model was used to identify the water inrush source of 47 groups of training samples and 20 groups of analysed samples from different water sources in the Zhaogezhuang coal mine. The results show that PCSOM can reduce the information overlap between discriminant indexes, simplify the model structure, and improve the running speed of the algorithm. The penalty factor c and kernel parameter g of the SVM optimized by the grey wolf algorithm are faster and more stable in parameter optimization, and the discrimination result is more accurate. The proposed model can accurately and quickly identify water inrush sources. Thus, it is helpful for rapidly predicting inrush disasters and can serve as a reference for inrush source identification technology.

**Keywords** Water hazards · Water inrush source identification · Zhaogezhuang coal mine · Grey wolf algorithm

## Introduction

With increased mining depth and more complex hydrogeological conditions, mine water inrush accidents have become more common (Bukowski 2011; Polak et al. 2016). The Zhaogezhuang Mine of the Kailuan Group is one of the deepest mines in mainland China. In 1972, an Ordovician limestone water inrush occurred along the floor of the mine (Wu et al. 2002), with a maximum inflow of 52.7 $m^3$/min. The inrush event caused the mine to stop production for half a year under working face 8 (elevation – 636.10 m). In 1995, a water inrush occurred in the goaf during mining. Stones within the coal seam were washed out for 183 m. The occurrence of inrush accidents severely affect the safety of workers in coal mines. Therefore, it is important to quickly and accurately identify water inrush sources (Dong et al. 2019).

At present, there are three common types of water source discrimination methods: the water temperature and water level method, hydrochemical analysis, and mathematical theory analysis. Wu et al. (2019) established a linear equation of formation buried depth and ground temperature in the Beiyangzhuang Mine to calculate water temperature; they determined the source of water inflow by comparing the calculated water temperature of a water-filled aquifer with the measured water temperature of a water inflow point.

Many scholars have used hydrochemical characteristics to identify water inrush sources at an early stage. Qian et al. (2016) assessed groundwater geochemical evolution and the connectivity of aquifers near a large coal mine in Anhui Province, China, using multivariate statistical analysis of the local hydrochemistry. Wang et al. (2016) analyzed groundwater chemical characteristics of four aquifers in the Pingdingshan coalfield. Based on coalfield geological characteristics, four aquifer-influencing coal-mining operations were categorized according to chemical distributions. Then, a water inrush source discriminant model was developed for each category; the model had strong pertinence and high discriminant accuracy (up to 93.75%). Li et al. (2016) obtained conventional ion concentrations through a field sampling test and combined the test results with a multivariate statistical analysis method to determine seawater infiltration channels and water inrush sources. With the development of

✉ Donglin Dong
ddl@cumtb.edu.cn

[1] Department of Geological Engineering and Environment, China University of Mining and Technology, Beijing (CUMTB), Beijing 100083, China

basic theories and computer technology, a variety of mathematical methods (multivariate statistics, grey system, and fuzzy mathematics) and discrimination models, as well as other computer-based discrimination methods have been developed, i.e. principal component analysis, BP neural network method, Bayes discrimination method, analytic hierarchy process, multiple logistic regression method and Fisher discrimination, support vector machine, and various combinations of such methods (He et al. 2016; Sun et al. 2016; Wang et al. 2016). Ma et al. (2014) established a water inrush source identification method in the Panxie mining area of Huainan based on fuzzy evaluation. This method is more suitable for inrush source identification when the water quality identification ability is poor and the ground temperature difference (buried depth difference) between aquifers is large. Bi et al. (2021) used multivariate statistics and discriminant analysis methods to distinguish 37 water samples from three types of water sources in the Xutuan coal mine and established a Fisher discriminant model of inrush water sources based on fuzzy cluster analysis and factor analysis. The accuracy rate was 91.9%. Zhang et al. (2017) analyzed 118 water samples of four water types in the Qinan coal mine using a multiple logistic regression analysis (MLRA) and Bayes recognition model (BRM); their MLRA-BRM combination model had a recognition accuracy of 95.28%. Dong et al. (2019) combined Fisher feature extraction and support vector machine (SVM) methods to distinguish 56 samples from four main water-filling sources in the Wuhai mining area of China. Their combined model identified water inrush sources more accurately and effectively than a traditional support vector machine model. However, this method requires a large amount of water sample data and cannot identify multiple inrush sources at the same time.

Various methods can be used to model water inrush accidents in mines. The main downside of mathematical function analysis methods, such as multivariate statistics, grey system, and fuzzy mathematics, is that the maximum and minimum operations may lose a large amount of hydrogeochemical information (Qiu et al. 2016). Also, there is a problem of unclear classification. The Bayes discriminant method is suitable if the principal components of samples are evident. Artificial neural networks require too many training samples, and the selection of training samples strongly affect the evaluation results. SVM methods are sensitive to missing values, and it is difficult to modify parameters (Zhang et al. 2017). Also, if aquifers are hydraulically connected, the collected mixed water samples limit the value of this approach. Likewise, the discriminant function of mixed water samples is rarely understood. The Fisher discriminant method can determine when there is too little sample data or if it is too irregularly distributed, so that there is no need to select the initial parameters of the model. Moreover, the discriminant results are good and easy to compute. The PCA and SOM algorithms have the advantage of preserving the original information of sample data by reducing the strong collinearity between sample data dimensions and variables; this prevents information overlap affecting the prediction results. The SVM model can predict and classify non-linear and high-dimensional water samples based on existing water sample information. The algorithm can further improve predictive accuracy by optimizing the parameters of the SVM model.

Although there are many methods to improve the parameter tuning of SVM, such as grid search, genetic algorithm, and particle swarm optimization (Sukawattanavijit et al. 2017; Zhang et al. 2017), the algorithms takes a long time to run. The efficiency and applicability of SVM and other methods are greatly affected by the selection of relevant parameters. Some methods ignore information redundancy between indicators in the selection of discrimination indicators; thus, there is room for further improvement and optimization. Considering the advantages and disadvantages of each method and the limitations of objective conditions, a new and improved combination discriminant method is proposed in order to quickly and accurately judge the water inrush source.

## Hydrogeological Conditions of the Study Area

The Zhaogezhuang Mine is located in the northeast Guye District, Tangshan City, Hebei Province. The topography is hilly in the north and gradually flattens toward the south. The average annual rainfall in the study area is 582.9 mm, mostly concentrated from June to September. The maximum rainfall was 1180.0 mm in 1949. The annual average temperature is 10.5 °C, with the lowest temperature recorded in January (−23.5 °C) and the highest in August (43.0 °C). The maximum evaporation occurs from June to August, with a multiyear average evaporation of 1104.0 mm.

The altitude of the mountainous terrain varies from 52 to 200 m. There is no large water system within the mine field. According to the hydrogeological conditions, the mining area is divided into a piedmont hydrogeological zone (I) and a plain hydrogeological zone (II). The distribution of aquifers and mine-based aquifers is shown in Fig. 1. $I_1$ is a thrust nappe structure area, with a coal-bearing rock series overlying Paleozoic limestone. $I_2$ is the steep slope of the west wing. The development of steep strata in this area makes it easy for water to leak from the coal-bearing rock series. Therefore, it is necessary to prevent potential hazards such as coal-water gushing. $II_1$ is the fault development area of the east wing, with normal faults and reverse faults; however, no fault water diversion accident has occurred. There are normal faults and reverse faults in the $II_2$ area, and water diversion from faults have caused water inrush accidents. The stratum dip in the $II_3$ area is gentle, and the buried depth

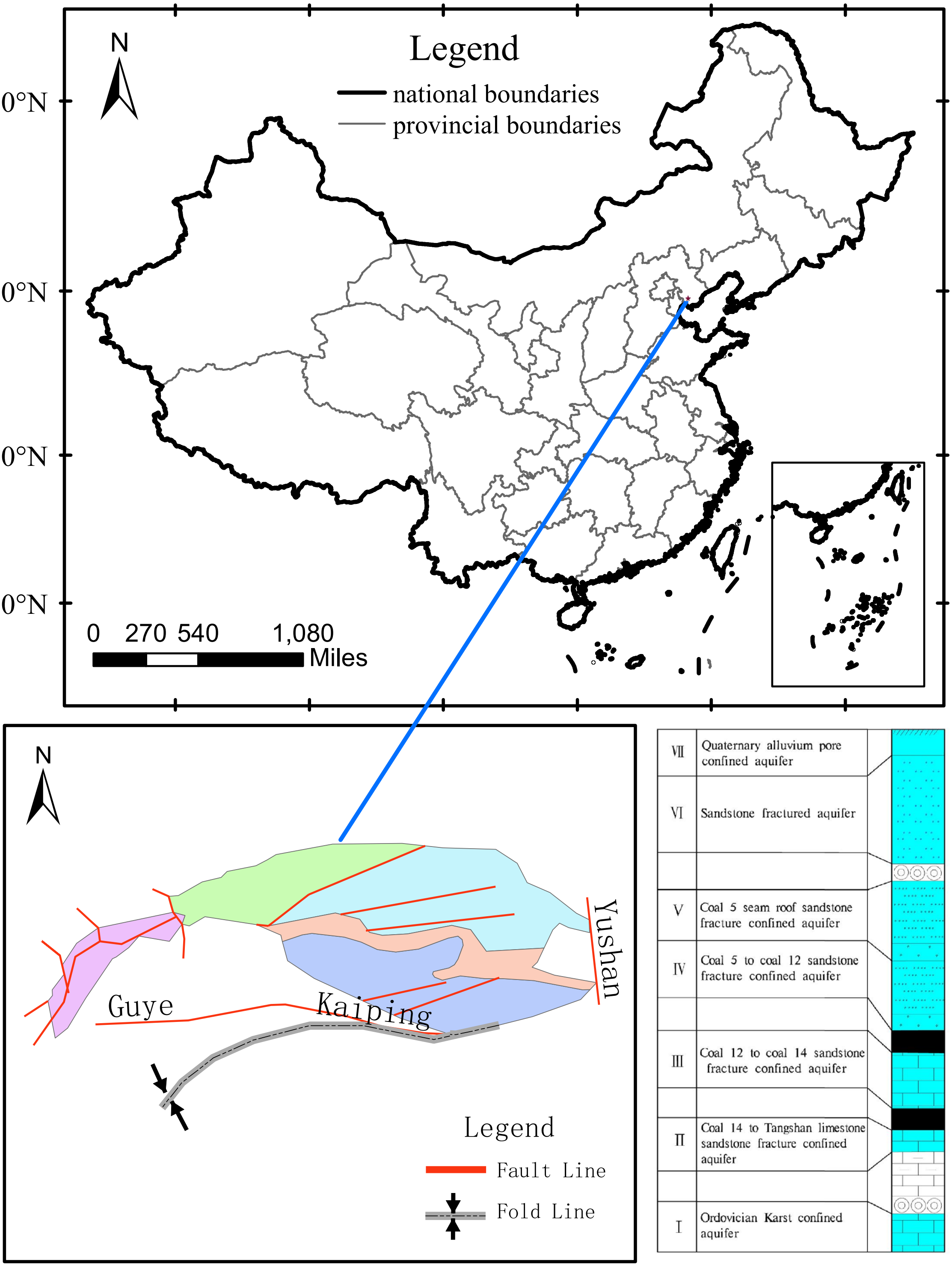
N
Legend
national boundaries
provincial boundaries
50°N
40°N
30°N
20°N
0 270 540 1,080
Miles
N
Guye
Kaiping
Yushan
Legend
Fault Line
Fold Line
Ⅶ Quaternary alluvium pore confined aquifer
Ⅵ Sandstone fractured aquifer
Ⅴ Coal 5 seam roof sandstone fracture confined aquifer
Ⅳ Coal 5 to coal 12 sandstone fracture confined aquifer
Ⅲ Coal 12 to coal 14 sandstone fracture confined aquifer
Ⅱ Coal 14 to Tangshan limestone sandstone fracture confined aquifer
Ⅰ Ordovician Karst confined aquifer

**Fig. 1** Location and structure outline of the study area

is large. The mining depth can reach 1200 m. The top of the Ordovician limestone is close, and the number of faults is small (Lin et al. 2021). Among the various water inrush accidents in the Zhaogezhuang coal mine, most are due to the Ordovician aquifer.

Several groundwater aquifers exist in the Zhaogezhuang mine field: the confined karstic Ordovician limestone (I), the confined fractured Tangshan limestone and sandstone strata just below the14 coal seam (II), the confined fractured sandstone between the 12 and 14 coal seams (III), the confined fractured sandstone between the 12 coal seam and the 5 coal seam (IV), the confined fractured sandstone roof of the 5 coal seam (V), and above this, various confined fractured sandstone strata (VI), and confined quaternary alluvium porous flow aquifers (VII). The I, VI, and VII aquifers are indirect mine-filling aquifers. The II, III, IV, and V aquifers are direct mine-filling aquifers. According to the Brotsky hydrochemical classification, the dominant cations in the goaf samples are $Ca^{2+}$ and $Mg^{2+}$ and the dominant anion is $SO_4^{2-}$. Thus, the water quality type of the two samples is $SO_4$-Ca. In the Ordovician limestone water samples, the dominant cation is $Ca^{2+}$ and the dominant anions are $SO_4^{2-}$ and $HCO_3^-$. The water quality type of the two samples is $SO_4$-$HCO_3$-Ca. In the fractured coal seam roof sandstones water samples, the water quality type of the two samples is $HCO_3$-Ca.

## Methods

### Sampling and Testing

The water sample data were mainly from hydrogeological reports and field investigations of the Zhaogezhuang mining area. Samples were collected through underground drainage holes or surface hydrological observation holes. The underground drainage holes were directly collected from the mine, and the surface hydrological observation holes were collected with self-made deep-water samplers.

Before water sample collection, clean 550 mL plastic bottles and their caps were rinsed three to five times with the sample water. The water samples were stored and processed at low temperatures to inhibit redox and biochemical reactions. Acid base titration was used to determine the concentration of $HCO_3^-$. $Na^+$, $Ca^{2+}$, $Mg^{2+}$, $Cl^-$ and $SO_4^{2-}$ was determined by ion chromatography (ICS-600). Of the 67 water samples, four types were selected for training and testing: goaf water, Ordovician limestone water (from aquifer I), sandstone fissure water from coal seams 12 to 14 (from aquifer III), and sandstone fissure water of coal seam 12 (from a section of aquifer IV). Specific water sample characteristics are shown in supplemental Tables S-1 and S-2; number 1 is goaf water, number 2 is Ordovician limestone water, number 3 is fractured sandstone water (coal seam 13), and number 4 is fractured sandstone water (coal seam 12).

### Fisher Mathematical Principle

The basic principle of the Fisher discriminant analysis method is to project multiple sample data with known group numbers and dimension numbers in a certain direction to clearly distinguish projection data groups. The distinction between groups is realized by constructing several discriminant functions based on analysis of variance (Huang et al 2011; 2018). When calculating the discriminant function, the weight of each index variable is obtained by making the intraclass deviation as small as possible and the interclass deviation as large as possible. In this way, the determined index values of sample *x* can be substituted into the established discriminant to calculate the number of the function. Finally, the result is compared with the dividing point to further complete the identification and classification of *x*.

To achieve this classification, an appropriate projection line must be selected to distinguish the projection direction of various data points to the greatest extent. When high-dimensional data is projected in this direction, the ratio of interclass dispersion $S_b$ to intraclass dispersion $S_w$ is largest. Suppose there are k categories, and $x_j^{(i)}$ represents the $j$-th sample in category $i$. The intraclass dispersion matrix $S_w$ of the sample is expressed as:

$$S_w = \frac{1}{N}\sum\nolimits_{i=1}^{k}\sum\nolimits_{i=1}^{Ni}\left(x_j^{(i)} - x^{(i)}\right)\left(x_j^{(i)} - x^{(i)}\right)^T, \tag{1}$$

where $x^{(i)}$ represents the mean value of class $i$ samples; $N_i$ represents the number of samples of class $i$; and N represents the total number of samples, i.e. $= \sum_{i=1}^{k} N_i$. For this sample, the interclass dispersion matrix $S_b$ can be expressed as:

$$S_b = \frac{1}{N}\sum\nolimits_{i=0}^{k} N_i\left(x^{(i)} - \overline{x}\right)\left(x^{(i)} - \overline{x}\right)^T. \tag{2}$$

The number of categories is k; $x^{(i)}$ represents the mean value of class $i$ samples; $\overline{x}$ is the average of all samples. The criterion function of Fisher linear discriminant is defined as $J(W) = \frac{|W^T S_b W|}{|W^T S_w W|}$. As the discreteness between classes increases and the discreteness within the class decreases, the class severity improves and becomes easier to classify. The vector w* that maximizes the function $J(W)$ is the required optimal vector. W* can be solved by establishing the Lagrange function. After calculating w* and then calculating the threshold Y0, the linear discriminant function *fx* can be determined. Then, any high-dimensional sample X can be entered and classified.

### Self-Organization Map

A self-organizing neural network (a self-organizing map or SOM) is a neural network based on unsupervised learning methods first proposed by Kohen (1982). In hydrology, ecology, and other related fields, scholars usually apply SOM to hydrogeochemical cluster analysis (Nguyen et al. 2015). This study attempts to cluster the hydrochemical parameters of each aquifer through an SOM neural network and to discuss the correlation of each parameter and the necessity of dimension reduction. The process of clustering groundwater samples by an SOM neural network can be subdivided into neuron selection, category selection, and category division. Choosing the right neurons is the key to good clustering. The advantages and disadvantages of network size selection can usually be evaluated using quantitative error (QE) and topological error (TE) indicators to determine the best mapping neurons.

## Principal Component Analysis

PCA is an effective method for analysing statistical data. The objective is to find a set of vectors in data space that best explains the variance of the data. This method projects the original high-dimensional data into lower-dimensional data space through a special matrix, and it retains the main characteristics of the data for convenient processing (Shao and Xu 2015). During the primary component analysis process, feature selection or feature extraction usually occurs. Feature selection refers to the process of transforming data space into feature space, which theoretically has the same dimensions as the original data space. However, when a few effective transformed features can contain the main information of the original variables, it is possible to consider reducing the number of features and extracting the main features by reducing the dimensions of the feature space (called reduction of dimensions). The mathematical model of PCA follows:

Suppose that the line pattern combination of the p vectors of the original data matrix X is Y = AX:

$$\begin{cases} Y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ Y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \quad\quad\quad\quad . \\ \quad\quad\quad\quad . \\ \quad\quad\quad\quad . \\ Y_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{cases}, \tag{3}$$

which simplifies to $Y_i = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{pi}x_p$.

$Y_i$, $Y_j$ are unrelated ($i \neq j$;$i,\ j = 1, 2, \ldots, \mathrm{p}$), and the differences between the groups satisfy the following relationships: Y1 > Y2, Y2 > Y3, Y3 > Y4, and so on.

PCA includes five steps: (1) data standardization; (2) calculating the correlation coefficient matrix; (3) calculating eigenvalues; (4) selecting principal components; and (5) calculating the principal component score.

## Support Vector Machine

SVM is a neural network model that can effectively solve high-dimensional, small sample, and nonlinear problems (Zhang et al. 2006). The basic idea is to map the input vector in low-dimensional space to high-dimensional feature vector space by nonlinear transformation using a nuclear function. Then, one can create an optimal hyperplane in the high-dimensional space, combine the principle of structural risk minimization and VC dimensional theory, classify the sample linearly, and finally realize the nonlinear classification in the low-dimensional input space. Suppose a set $S$ of N training samples is given as $S = \{(x_i, y_i), i = 1, 2, \ldots, \mathrm{N}\}$. The expression and objective function of the classification hyperplane are

$$f(x) = \mathrm{w}x + b, \tag{4}$$

$$\min\varnothing(w) = \frac{1}{2}\|(w)\|^2 + c\sum_{i=1}^{N}\varepsilon_i, \tag{5}$$

$$s.t.y_i(\mathrm{w}.x + b) \geq 1 - \varepsilon_i, \tag{6}$$

$$\varepsilon_i \geq 0, i - 1, 2, \ldots, N, \tag{7}$$

where w is the normal vector of the hyperplane, b is the translation distance of the hyperplane, $\varepsilon_i$ is a non-negative relaxation variable that improves the generalization ability of the model, and c is the penalty factor, which is used to weigh the relationship between classification loss and maximum interval.

Using a Lagrange multiplier, this can be viewed as a Lagrangian dual problem:

$$\max\frac{1}{2}\sum_{\mathrm{j=1}}^{\mathrm{n}}\sum\nolimits_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{a_i}\alpha_{\mathrm{j}}\mathrm{y_i y_j k}\left(\mathrm{x_i}, \mathrm{x_j}\right) - \sum\nolimits_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{a_i}, \tag{8}$$

$$\mathrm{s.t.0} \leq \upalpha_{\mathrm{i}} \leq \mathrm{C}, \mathrm{i} = 1 \ldots \mathrm{m}, \tag{9}$$

$$\sum\nolimits_{\mathrm{i=1}}^{\mathrm{n}}\upalpha_{\mathrm{i}}\mathrm{y_j} = 0, \tag{10}$$

where $k(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ is a linear kernel function. Common examples are the linear kernel function $k(x_i, x_j) = x_i^T x_j$, the polynomial kernel $k(x_i, x_j) = (x_i^T x_j)^n$, and in this paper, the Gaussian kernel is:

$$K(x_i, x_j) = \exp\left(-\frac{\left|x_i - x_j\right|^2}{2g^2}\right), g > 0, \tag{11}$$

where g is the parameter of the kernel function.

Algorithmic performance is closely related to the value of the penalty factor c and nuclear parameter g. If c and g are not selected properly, then the algorithm performance may be poor. However, there is no internationally recognized best way to select SVM parameters. At present, the most used method is to let c and g take values within a certain range.

For the determined c and g, when the training set is from the original dataset, the k-cv method is used to obtain the classification accuracy of the training set under groups c and g. The best parameters are those with the highest classification accuracy in the training set of groups c and g. However, there may be multiple groups of c and g corresponding to the highest verification classification accuracy. In this study, heuristic algorithm grey wolf optimization (GWO), particle swarm optimization (PSO), and grid search (GS) were used to find the best parameters c and g.

The GWO algorithm is a population intelligent optimization algorithm proposed by Mirjalili in 2014.and was inspired by the predation behaviour of grey wolves. The GWO algorithm is roughly divided into five steps: social hierarchy, searching for prey, surrounding prey, hunting, and attacking prey. The grey wolf strictly abides by the hierarchical relationship of social dominance. When the grey wolf searches for desired prey, it will slowly approach the prey and then surround the prey. In each iteration, the three best grey wolves are retained, and the locations of other search agents are updated according to their location information. When the prey is no longer moving, the grey wolf will attack to catch the prey.

PSO, also known as the bird swarm foraging algorithm, is a typical global optimization algorithm. It was first proposed by Kennedy and Eberhart in 1995 and then improved to form the commonly used weighted particle swarm optimization algorithm. The basic idea is to achieve the optimal purpose of the group through cooperation and information sharing among individuals in the group. In searching for optimal parameters, the parameters c and g are regarded as particles in the particle group algorithm. From random solutions, the algorithm continuously iterates to find the optimal solution. This method is simple, easy to implement, and is widely used in neural network optimization, parameter optimization, control system, and other fields.

The GS method is the parameter search method (Liu et al. 2010). In the two-dimensional parameter matrix composed of c and g, all points in the grid are traversed to take on values. For the determined values of c and g, the k-cv method is used to obtain the training set verification classification accuracy under this group of matrices. The c and g groups with the highest verification classification accuracy of the training set are considered the best parameters. Global optimization can be obtained by using the grid-search algorithm, assuming c and g are independent of each other for convenient parallelization.

The Fisher-PCSOM-GWOSVM model.

- Step 1: *Collect and filter water sample data*. When the Fisher and SVM methods are used to distinguish between fast-moving water sources, the collected water samples do not fully represent the actual aquifer. Therefore, we attempt to combine the two methods for identification and eliminate errors caused by the hydraulic connection between aquifers. Water samples that do not belong to one of the four aquifers are excluded. The identification model is established according to the remaining water samples. The Fisher method is first used to screen the original water samples according to the hydrochemical data of the four water sources.
- Step 2: *Dimension-lowering processing, rebuilding the dataset*. The traditional discrimination method does not consider the correlation and information redundancy between the indicators (the six hydrochemical parameters). Directly using the indicator data may impair the discrimination results. The number of indicators also affects the speed of algorithm execution. In this paper, the PCA method combined with the SOM algorithm is used to reduce the dimensions of the discriminant index data.
- Step 3: *Build an optimal combination algorithm model*. GWO, PSO, and GS algorithms are introduced to optimize SVM parameters. Then, the optimal combination algorithm is selected by comparing the discrimination results.

Using the above steps, this paper establishes a forecast model based on Fisher-PCSOM-GWOSVM. The specific process is shown in Fig. 2.

## Results and Discussion

### Fisher Discriminant Analysis

Fisher discrimination was performed on 67 original water samples using SPSS software. The Fisher function was established by considering six main parameters ($Na^+$, $Ca^{2+}$, $Mg^{2+}$, $Cl^-$, $SO_4^{2-}$, $HCO_3^-$). $Na^+$, $Ca^{2+}$, $Mg^{2+}$, $Cl^-$, $SO_4^{2-}$, and $HCO_3^-$ are independent variables, and the actual water source category is the dependent variable. The Fisher functions of 67 original water samples can be seen from the model parameter values listed in Table 1. The mathematical model expressions for the three types of burst water sources are as follows:

$$Y1 = 0.402x_1 + 0.421x_2 + 0.418x_3 - 0.044x_4 + 0.015x_5 + 0.251x_6 - 55.776$$

$$Y2 = 0.4659x_1 + 0.738x_2 + 0.652x3 - 0.631x_4 - 0.477x_5 - 0.481x_6 - 21.974$$

$$Y3 = 0.725x_1 + 0.499x_2 + 0.601x_3 - 0.152x_4 - 0.011x_5 - 0.022x_6 - 54.933$$

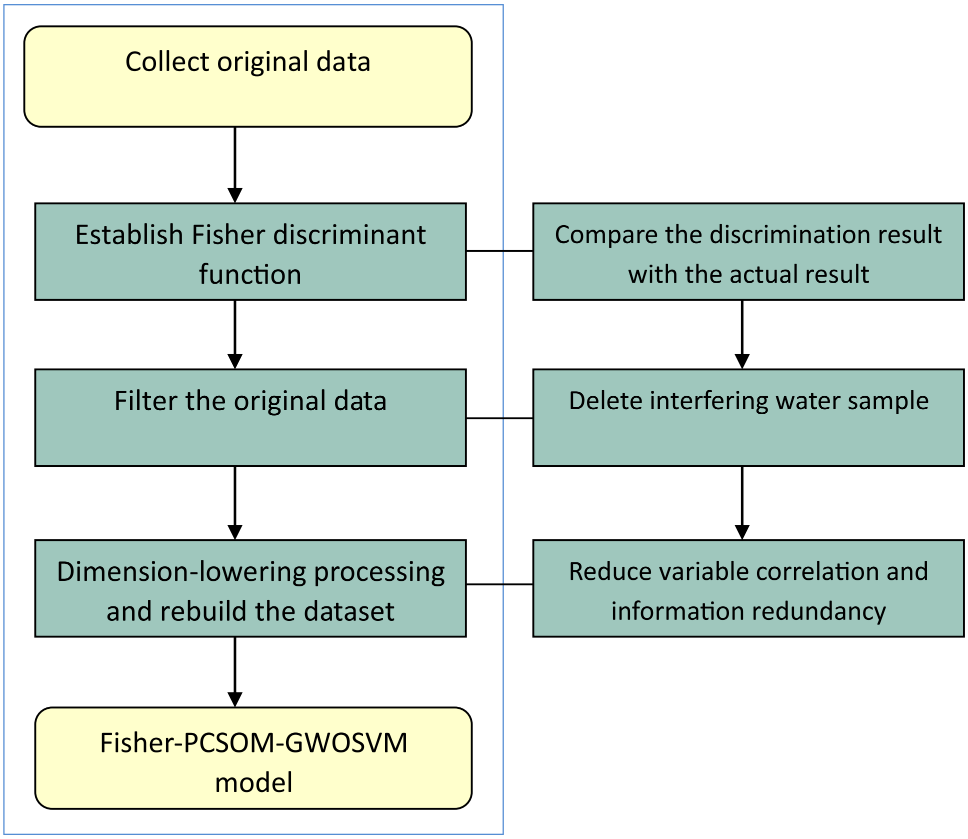
Collect original data
Establish Fisher discriminant function
Compare the discrimination result with the actual result
Filter the original data
Delete interfering water sample
Dimension-lowering processing and rebuild the dataset
Reduce variable correlation and information redundancy
Fisher-PCSOM-GWOSVM model

**Fig. 2** Flowchart of the fisher-PCSOM-GWOSVM

**Table 1** Coefficients of fisher discriminant function

| | Y1 | Y2 | Y3 |
|---|---|---|---|
| $Na^+$ | 0.402 | 0.659 | 0.725 |
| $Ca^{2+}$ | 0.421 | 0.738 | 0.499 |
| $Mg^{2+}$ | 0.418 | 0.652 | 0.601 |
| $Cl^-$ | −0.044 | −0.631 | 0.152 |
| $SO_4^{2-}$ | 0.015 | −0.477 | −.011 |
| $HCO_3^-$ | 0.251 | −0.481 | −.022 |
| (constant) | −55.776 | −21.974 | −54.933 |

Although the mine water inrush sources were hydrochemically different, hydraulic connections between the aquifers meant that the water samples could represent a combination of more than one source. Therefore, it was necessary to identify and eliminate such nonconforming water samples. Of the 67 original water samples tested by the Fisher discriminant, four water samples (21, 26, 37, and 56) did not conform to the actual water sample type. To eliminate the error caused by the hydraulic connection between aquifers, these samples were eliminated (Supplemental Tables S-1 and S-2).

## PCA and SOM Analysis (PCSOM)

SOM neural network training was used for standardization and the SOM neuron diagrams of six hydrochemical parameters were obtained (Fig. 3). The colour depth of each neuron represents the component value of the chemical parameters of the water sample point. The graph intuitively presents the distribution of neuron distance and corresponding colour depth to explain the information and qualitative relationship between various hydrochemical parameters. As shown in Fig. 3, $Mg^{2+}$, $Cl^-$, and $HCO_3^-$ have similar colour gradients, indicating a strong correlation among the three ions. Considering the ionic relationship, $Mg^{2+}$ and $HCO_3^-$ in groundwater are correlated and may originate from the dissolution of calcite, dolomite, and gypsum. In contrast, the colour gradient relationship in the neuron diagram of $SO_4^{2-}$ is opposite that of the other four ions except for $Mg^{2+}$, indicating that there is a strong negative correlation between $SO_4^{2-}$ and the four ions.

The Pearson's correlation analysis of groundwater chemical parameters during the wet and dry periods of the basin was analysed with SPSS software. As can be seen from

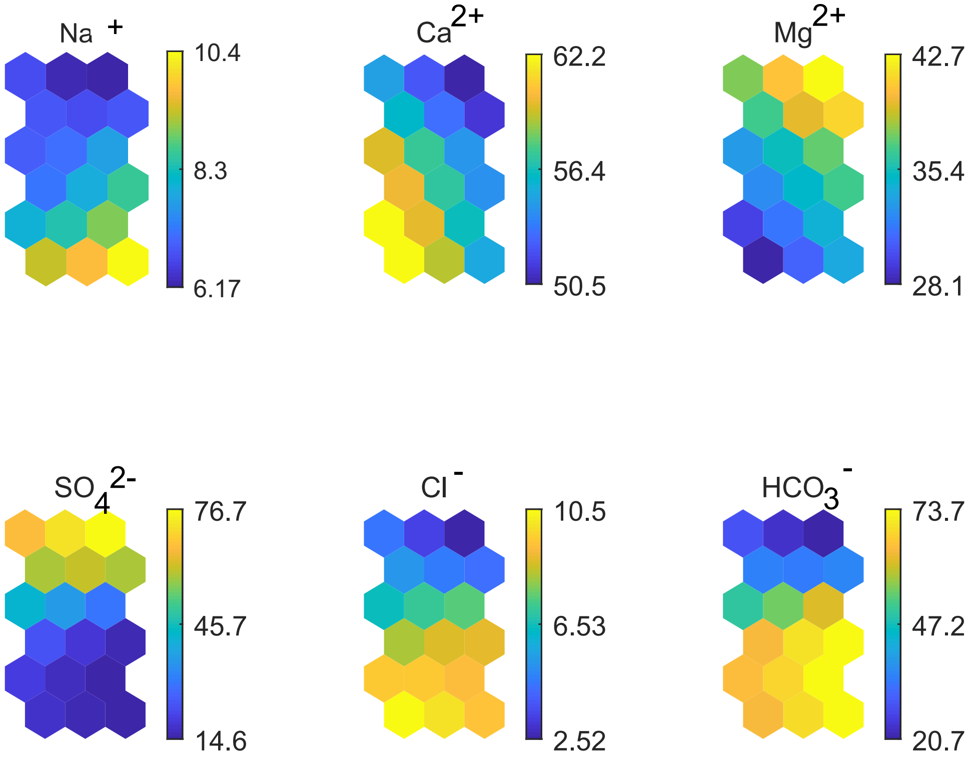

**Fig. 3** SOM composition analysis diagram

**Table 2** Pearson's correlation analysis results

| | $Na^+$ | $Ca^{2+}$ | $Mg^{2+}$ | $Cl^-$ | $SO_4^{2-}$ | – |
|---|---|---|---|---|---|---|
| $Na^+$ | 1 | | | | | |
| $Ca^{2+}$ | −0.348** | 1 | | | | |
| $Mg^{2+}$ | −0.231 | −0.801** | 1 | | | |
| $Cl^-$ | 0.243 | 0.421** | −0.601** | 1 | | |
| $SO_4^{2-}$ | −0.125 | −0.371** | 0.481** | −0.864** | 1 | |
| $HCO_3^-$ | 0.099 | 0.287* | −0.375** | 0.786** | −0.987** | 1 |

**At the 0.01 level (double tail), the correlation is significant; *At the 0.05 level (double tail), the correlation is significant

Table 2, there was a significant negative correlation between $Ca^{2+}$, $Cl^-$, $HCO_3^-$, and $SO_4^{2-}$. $Mg^{2+}$ and $SO_4^{2-}$ have a significant positive correlation, with $Na^+$ showing a significant negative correlation with $Ca^{2+}$. The results of Pearson's correlation analysis of hydrochemical parameters are consistent with those of SOM neuron analysis, indicating the effectiveness of the SOM neuron diagram in characterising the correlation between parameters.

According to the results of SOM and Pearson's correlation analysis, there is overlapping information between various indicators. The six ions cannot be directly used for water source discrimination; otherwise redundant or repeated information will be gathered. This process increases the number of calculations and may also reduce the accuracy of water source discrimination (i.e. false discrimination). Therefore, it is necessary to analyse the main component of the sample. After processing the main component analysis, creating the scree plot of the principal components (Fig. 4), and interpreting the variance ratio table of the main components (Table 3), the first four main components are selected for subsequent water source identification.

## Establishment and Training of the Mine Water Inrush Source Discrimination Model

Using the four principal components extracted by PCSOM as the input vector of the prediction model and the type of water inrush source as the output, the discrimination model of mine water inrush source was established. Using the established predictive model based on Fisher-PCSOM-GWOSVM, Fisher-PCSOM-PSOSVM, and Fisher-PCSOM-GSSVM, the sample data were trained from 1 to

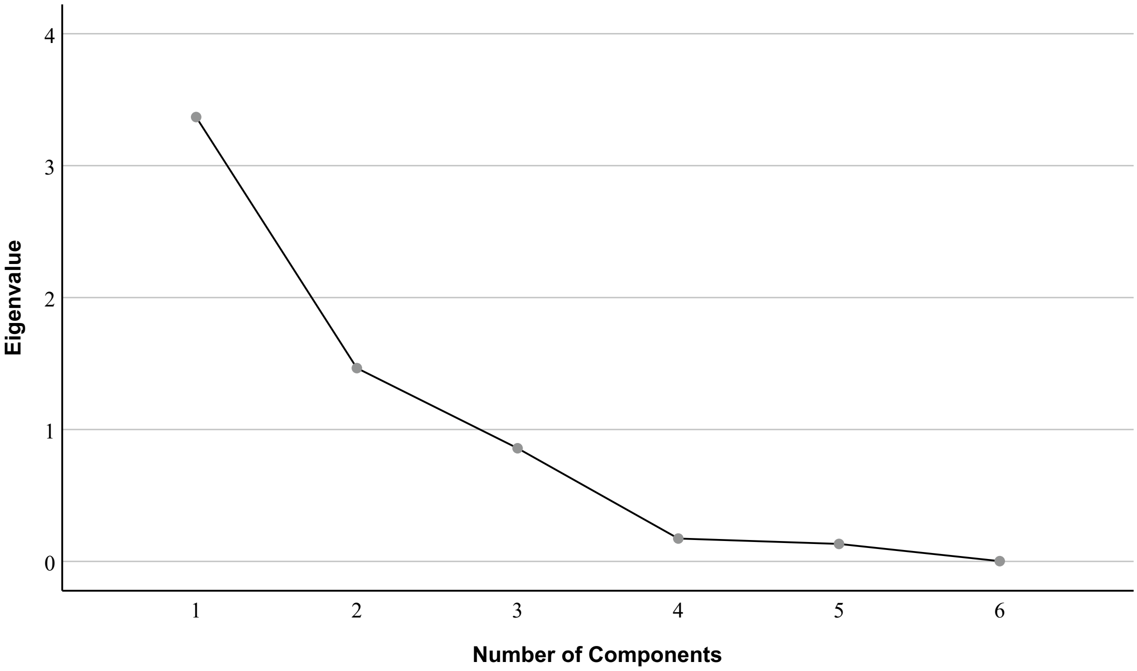

**Fig. 4** Gravel diagram of principal component analysis

**Table 3** Total variance explained

| Component | Initial eigenvalues | Contribution rate | Cumulative contribution rate/ % |
|---|---|---|---|
| 1 | 3.420 | 56.996 | 56.996 |
| 2 | 1.450 | 24.167 | 81.163 |
| 3 | 0.930 | 15.500 | 96.663 |
| 4 | 0.174 | 2.900 | 99.563 |
| 5 | 0.024 | 0.400 | 99.963 |
| 6 | 0.002 | 0.037 | 100.000 |

45. Nineteen additional sets of sample data were identified as predictors.

Another 19 sets of samples are identified by using the trained mine-burst water source identification model (Figs. 5, 6, and 7). To verify the superiority of the prediction model, the optimization quantity of the three models is consistent with the termination algebra. Table 4 shows the parameter optimization results.

The Fisher-PCSOM-GWOSVM model test was 100% accurate, while the Fisher-PCSOM-PSOSVM model test was 94.7% accurate and the Fisher-PCSOM-GSSVM test was 84.6% accurate. The incorrect discrimination results determined the actual water sample of type 2 (Ordovician limestone water) as type 1 (Goaf water). Thus, the hydrochemical ion contents of the two water samples are very similar, and the water in the goaf is well connected with the Ordovician limestone water. The goaf may communicate with the Ordovician limestone aquifer through fractures or faults.

The discrimination results show that the Fisher-PCSOM-GWOSVM model has the fastest convergence speed, and the discrimination accuracy is significantly better than the other two models, with strong generalization and learning ability. The Fisher-PCSOM-GWOSVM model was used to cross-verify the 44 groups of training samples, and the overall recognition rate was 100% (Fig. 8). In addition, to verify the stability of the model, four groups of new water samples from different aquifers were input into the model, and the recognition accuracy of the combined model was 100%. As mentioned above, because the identification model was established based on a limited number of water samples, there were obvious differences in the identification accuracy of water samples from different water inrush sources. However, the traditional model produces multiple errors during the re-discrimination step, and the discrimination rate is less than 90%. Therefore, the Fisher-PCSOM-GWOSVM method was more accurate and more stable, and met the requirements for water source identification.

## Conclusions

This paper combines the advantages of various algorithms to establish a water inrush source identification model. The multiple improved combination methods of Fisher, PCA, SOM, and SVM produce the Fisher-PCSOM-GWOSVM

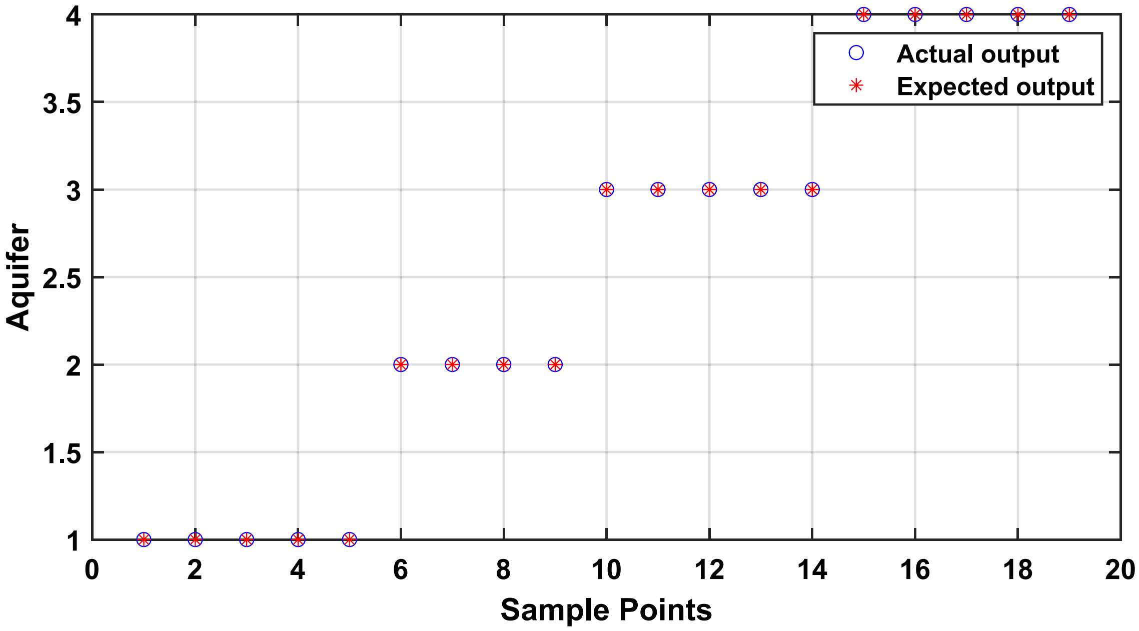

**Fig. 5** Fisher-PCSOM-GWOSVM test results

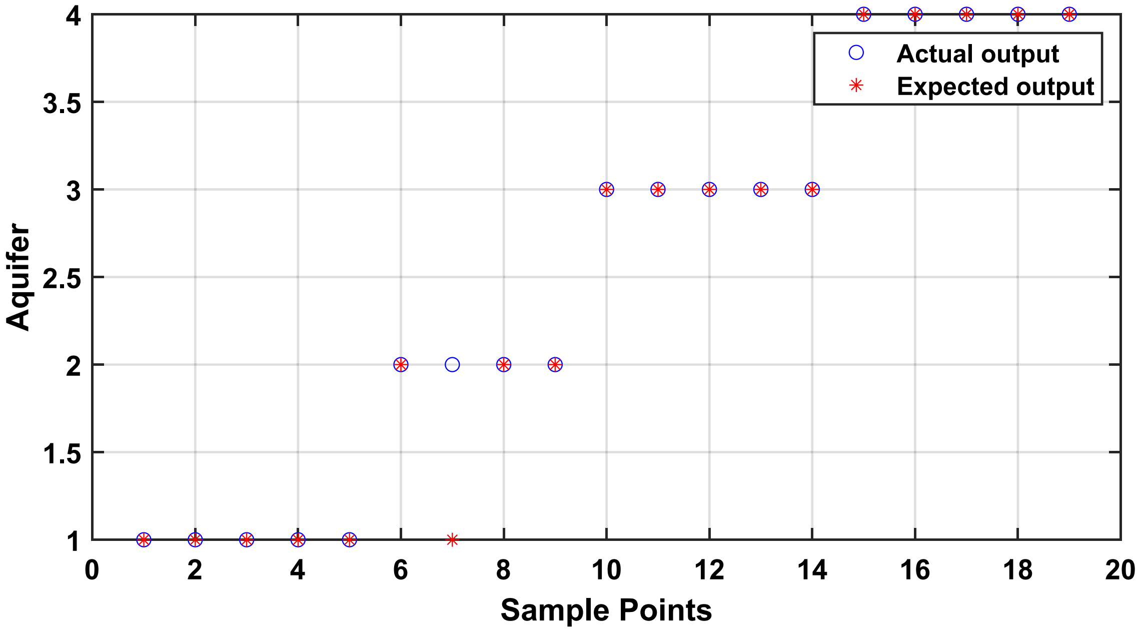

**Fig. 6** Fisher-PCSOM-PSOSVM test results

model for identifying the water inrush source in the Zhaogezhuang mining area. The results indicate the following: the Fisher-PCSOM-GWOSVM model has faster discrimination accuracy, faster running speed, stronger stability, better generalization ability, and superior learning ability. In the process of water source identification, PCA combined with SOM was used to extract the principal components from the original sample data, reduce the information overlap between the discrimination indexes, simplify the model structure, and improve the running speed of the algorithm. The SVM model optimized via the grey wolf algorithm was faster in parameter optimization

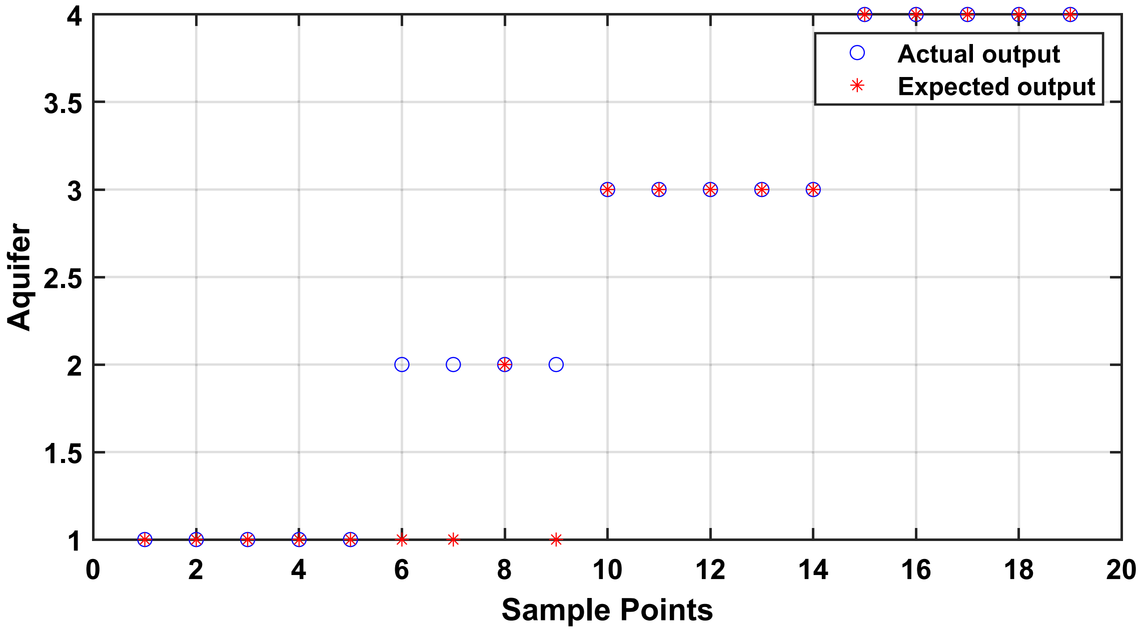

**Fig. 7** Fisher-PCSOM-GSSVM test results

**Table 4** The parameter optimization results

| Method model | c, g optimization results | Accuracy (%) | Discrimination time (s) |
|---|---|---|---|
| Fisher-PCSOM-GWOSVM | bestc:75.8345,bestg:0.3587 | 100 | 1.1255 |
| Fisher-PCSOM-PSOSVM | bestc:11.2135,bestg:0.2902 | 94.7 | 1.5145 |
| Fisher-PCSOM-GSSVM | bestc:13.4949,bestg:0.5219 | 84.6 | 2.8712 |

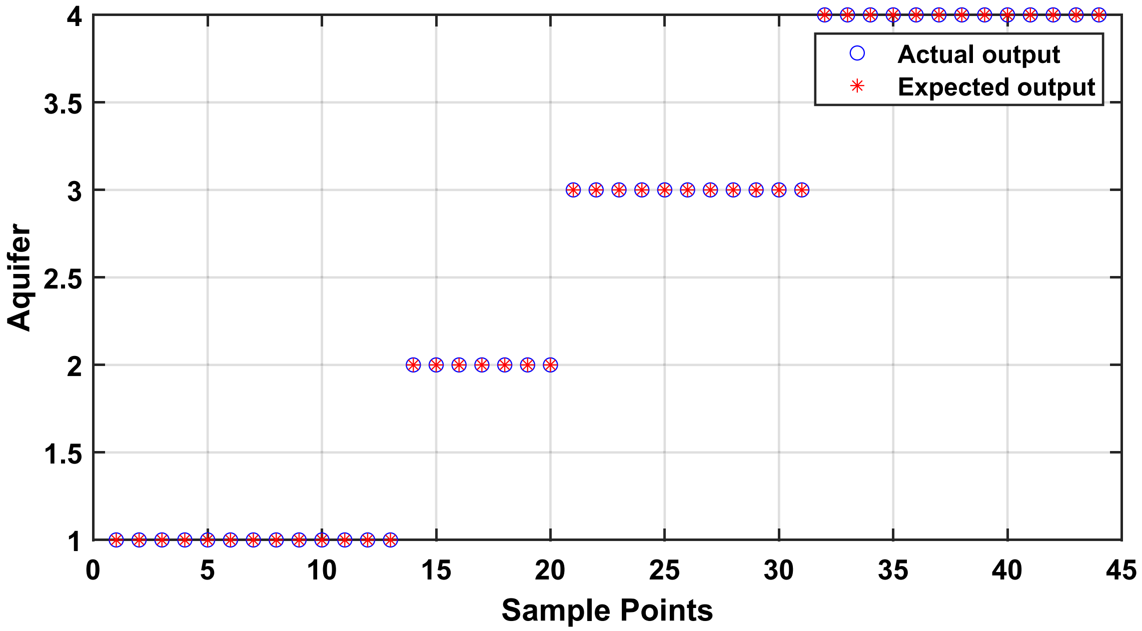

**Fig. 8** Fisher-PCSOM-GWOSVM training results

and more accurate in discrimination results. The established model was successfully applied in this study.

Due to the complexity of hydrogeological conditions in mines, the differences of labelled ions vary. A better underground hydrochemical database is needed. Therefore, future research should test more ion components and further improve the discrimination model to better identify the sources of mine water inrush.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10230-022-00884-5.

**Acknowledgements** This study was financially supported by the National Natural Science Foundation of China (4197020829), the Ministry of Science and Technology of China (2017YFC0804104), and the National Natural Science Foundation (U1710258).

## References

Bi YS, Wu JW, Zhai XR, Wang GT, Shen SH, Qing XB (2021) Discriminant analysis of mine water inrush sources with multi-aquifer based on multivariate statistical analysis. Environ Earth Sci 80:144

Bukowski P (2011) Water hazard assessment in active shafts in upper Silesian Coal Basin mines. Mine Water Environ 30:302–311

Dong DL, Chen ZY, Lin G, Li X, Zhang RM, Ji Y (2019) Combining the Fisher feature extraction and support vector machine methods to identify the water inrush source: a case study of the Wuhai mining area mine. Water Environ 38:855–862

He CY, Zhou MR, Yan PC (2016) Application of the identification of mine water inrush with LIF spectrometry and KNN algorithm combined with PCA. Spectrosc Spectr Anal 36:2234–2237 (**in Chinese**)

Huang PH, Chen JS (2011) Fisher identify and mixing model based on multivariate statistical analysis of mine water inrush sources. J China Coal Soc 36:131–136 (**in Chinese**)

Huang PH, Wang XY (2018) Piper-PCA-Fisher recognition model of water inrush source: a case study of the Jiaozuo mining area. Geofluids 2018:1–10

Jt LU (2012) Recognizing of mine water inrush sources based on principal components analysis and Fisher discrimination analysis method. China Safety Sci J 22:109–115 (**in Chinese**)

Keskin TE, Dugenci M, Kacaroglu F (2015) Prediction of water pollution sources using artificial neural networks in the study areas of Sivas, Karabük and Bartn (Turkey). Environ Earth Sci 73:5333–5347

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69

Li GQ, Meng ZP, Wang XQ, Yang J (2016) Hydrochemical Prediction of Mine Water Inrush at the Xinli Mine, China. Mine Water Environ 36:1–9

Liu XL, Jia DX, Li H, Jiang JY (2010) Research on kernel parameter optimization of support vector machine in speaker recognition. Sci Technol Eng 10:1669–1673

Nguyen TT, Kawamura A, Tong TN, Nakagawa N, Amaguchi H, Gilbuena R (2015) Clustering spatio–seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the Red River Delta, Vietnam. J Hydrol 522:661–673

Polak K, Różkowski K, Czaja P (2016) Causes and effects of uncontrolled water inrush into a decommissoned mine shaft. Mine Water Environ 35:128–135

Qian JZ, Wang L, Ma L, Lu YH, Zhao WD, Zhang Y (2016) Multivariate statistical analysis of water chemistry in evaluating groundwater geochemical evolution and aquifer connectivity near a large coal mine, Anhui, China. Environ Earth Sci 75:747

Qiu M, Shi LQ, Teng C, Zhou Y (2016) Assessment of water inrush risk using the fuzzy Delphi Analytic Hierarchy Process and Grey Relational Analysis in the Liangzhuang Coal Mine, China. Mine Water Environ 36:1–12

Shao LS, Xu B (2015) KPCA-SVM model for predicting Karst collapse tendency level China. Safety Sci J 25:60–65 (**in Chinese**)

Sukawattanavijit C, Chen J, Zhang HS (2017) GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data. IEEE Geosci Remote Sens Lett 14:284–288

Sun WJ, Zhou WF, Jiao J (2016) Hydrogeological classification and water inrush accidents in China's coal mines. Mine Water Environ 35:214–220

Wang XY, Ji HY, Wang Q, Liu XM, Huang D, Yao XP, Chen GS (2016) Divisions based on groundwater chemical characteristics and discrimination of water inrush sources in the Pingdingshan coalfield. Environ Earth Sci 75:872

Wu Q, Liu JT, Zhong YP, Yin ZR, Li JM, Hong YQ, Ye GJ, Tong YD, Dong DL (2002) The numeric simulations of water-bursting time-effect for faults in Zhaogezhuang coal mine, Kailuan, China. J China Coal Soc 27:511–516 (**in Chinese**)

Wu Q, Mu WP, Xing Y, Qian C, Shen JJ, Wang Y, Zhao DK (2019) Source discrimination of mine water inrush using multiple methods: a case study from the Beiyangzhuang Mine, northern China. Bull Eng Geol Environ 78:469–482

Zhang WY, Zhang HG, Liu JH, Li K, Yang DS, Tian H (2017) Weather prediction with multiclass support vector machines in the fault detection of photovoltaic system. IEEE J Automatica Sinica 4:520–525

Lin G, Jiang D, Dong D, Fu J, Li X (2021) A Multilevel Recognition Model of Water Inrush Sources: A Case Study of the Zhaogezhuang Mining Area. Mine Water Environ 40(3):773-782. https://doi.org/10.1007/s10230-021-00793-z

Ma L, Qian JZ, Zhao WD (2014) An approach for quickly identifying water-inrush source of mine based on GIS and groundwater chemistry and temperature. Coal Geo Exploration 42(2):49–53

Zhang H, Berg AC, Maire M, Malik J (2006) SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. IEEE Computer Vision and Pattern Recognition 2:2126–2136