Mine Water and the Environment (2019) 38:855–862
https://doi.org/10.1007/s10230-019-00637-x

TECHNICAL COMMUNICATION

# Combining the Fisher Feature Extraction and Support Vector Machine Methods to Identify the Water Inrush Source: A Case Study of the Wuhai Mining Area

Donglin Dong[1] · Zhiyuan Chen[1] · Gang Lin[1] · Xiang Li[1] · Ruomeng Zhang[1] · Yuan Ji[1]

Received: 7 May 2019 / Accepted: 24 September 2019 / Published online: 4 October 2019

## Abstract

Discriminating the source of water inrush accurately and efficiently is necessary for water control in the coal mining industry. We combined the Fisher feature extraction and support vector machine (SVM) methods and applied this new model to the Wuhai mining area. The method extracts features from the raw data and integrated SVM, and synthetically considers the influence of geographical factors. Cross-analysis was tested 100 times, which arbitrarily selected 12 samples for the prediction and discrimination process. The results indicate that this new combined model of linear dimension reduction and non-linear dimension elevation was more accurate and efficient in discriminating water inrush sources than the traditional SVM model. Moreover, by reducing the penalty term of SVM model, we analyzed the correlation among the aquifers. We concluded that aquifers II and IV correlated strongly with each other, and that aquifer III was poorly connected with the other aquifers.

**Keywords** Mine water-inrush source · Intelligent recognition · Coal mine

## Introduction

Various machine learning methods have been widely used to identify the source of water of inrush events (Yin et al. 2017). The Fisher linear discriminant (FLD) is an algorithm that projects high-dimensional samples into optimal discriminant vector space, extracts classification information from the data, and reduces the dimension of feature space. In contrast, support vector machine (SVM) models, can solve practical problems such as small sample, nonlinear, and high-dimensional pattern recognition, by projecting low-dimensional data to high-dimensional data. Huang and Wang (2018) modified the FLD method to identify a groundwater source. (Chen et al. 2009) used SVM to perform practical work on water inrush sources. However, abnormal values are likely to be incorrectly identified by FLD, while SVM is sensitive to missing values, as parameters are hard to adjust. Combining the advantages of the FLD and SVM methods provides a new discriminant method that extracts the features as a vector and constructs the SVM model with the combined Fisher features and original data to investigate the water inrush source.

The Wuhai coal mine is located in the arid to semi-arid area of Inner Mongolia (Fig. 1), which has a mean annual precipitation of only 155 mm/year. The geological conditions are complicated, and the coal seams are generally shallowly buried. Many coal seams are repeatedly mined, and water, fire, and gas disasters happen frequently. Many water inrush accidents have happened in the Wuhai Mine, such as the "3.1 (the data of the accident)" water inrush accident that occurred in the Camel Mountain mine. 67,000 $m^3$ of roadway was inundated in 70 min and the peak water inrush volume reached 60,036 $m^3$/h (Wei 2013).

Thus, it is particularly important to effectively and accurately discriminate water sources. The objective of this research was to identify the water inrush source by combining the FFE and SVM method. The method extracted the features from the original data and integrated SVM by Fisher

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10230-019-00637-x) contains supplementary material, which is available to authorized users.

✉ Gang Lin
ling@lreis.ac.cn

✉ Xiang Li
lx_cumtb@163.com

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

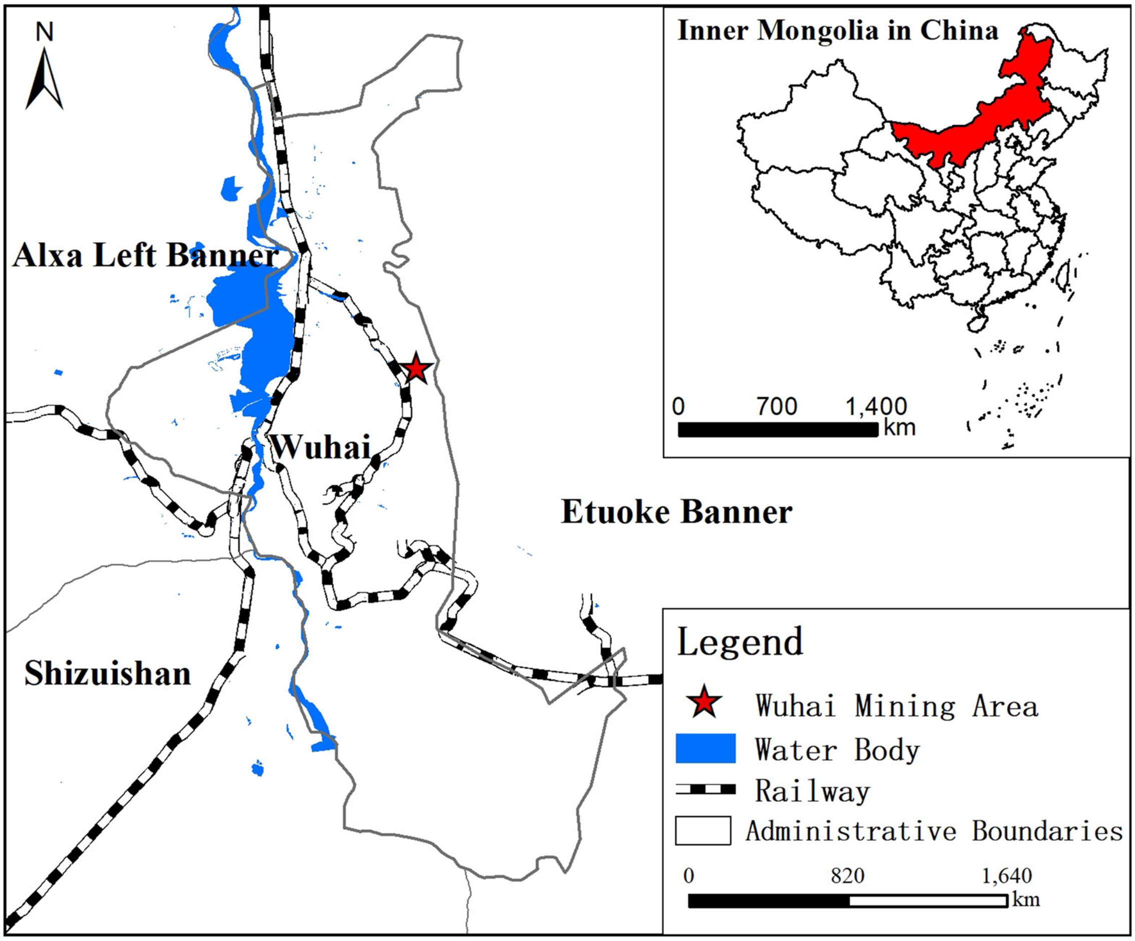

**Fig. 1** Location of the study area

features and original data, and also synthetically considered the influence of geographical factors.

## Methods and Materials

### Data Acquisition

According to the hydrogeological condition, the main aquifers of the Wuhai mine can be divided into three categories based on their lithology, thickness, water features, and burial conditions. The uppermost is the roof sandstone aquifer that consists of variegated medium-grained sandstone, sandy mudstone, and mudstone. The second is the floor sandstone aquifer, composed of gray-white sandstone and dark gray sandy mudstone. The lowest is the Ordovician limestone aquifer, composed of dark cyan limestone, which is the main aquifer in the segment. In addition, goaf water is a very important factor due to the accumulation of water in the old working faces. Thus, the Wuhai mining area contains four main water-filling sources: the floor sandstone aquifer (I), the roof sandstone aquifer (II), the Ordovician limestone aquifer (III), and the goaf water (IV).

Water samples were extracted from the mining district, including 8 groups from type I, 20 groups from type II, 21 groups from type III, and 7 groups from type IV. Supplemental Table 1 shows the results of the water sample analysis. These samples were extracted from the original files.

### Methods

#### Fisher Linear Discriminant (FLD)

FLD is an algorithm for projecting high-dimensional samples into the optimal discriminant vector space, extracting the classification information and compressing the dimension of a feature space. After projection, the sample has the largest inter-class distance and the smallest intra-class distance in the new space (Huang and Wang 2018; Ren and Chang 2005; Shan et al. 2002).

For the data set, X, where $x_j$ is the sample in X, and $X_i$ is the class in X, the class mean vector:

$\mu_i = \frac{1}{N}\sum x_j \epsilon X_i j = 1, 2 \ldots n$. In a discrete degree matrix: $S_w = \sum \left(x_j - \mu_i\right)\left(x_j - \mu_i\right)^T \epsilon X_i j = 1, 2 \ldots n$ between the discrete degree matrix for $S_b = \left(\mu_i - \mu\right)\left(\mu_i - \mu\right)^T$; so: $J = \frac{\omega^T S_b \omega}{\omega^T S_w \omega}$. To find the maximum variance direction, we constructed a Lagrange multiplier cost function and calculated $\omega : \ \omega = S_w^{-1}\left(\mu_i - \mu\right)$; where μ stands for mean vector, $S_w$ stands for intra-class divergence matrix, $S_b$ stands for inter-class divergence matrix and ω is the optimal projection direction vector.

### Support (SVM)

SVM is a machine-learning algorithm with supervised learning that can be used to maximize the learning strategy interval in a feature space. The model is a two-class classification model that can be transformed into a convex quadratic programming problem. SVM handles linear non-separable problems by converting them into high-dimensional ones based on linear mapping, which makes it linearly separable (Zhang et al. 2006). Suppose the data set $T = \left\{\left(x_n, y_n\right)\right\}_{n=1}^{m}$, labeled by the vector $y_n \in \{+1, -1\}$. The SVM separates the training vectors in a $\varphi$ mapped space, with an error cost, $C > 0$, such that: $\min_{\omega} \frac{1}{2}\omega^2 + C\xi_i$; $\text{s.t.} y_i\left(\omega^T \phi\left(x_i\right) + b\right) \le 1 - \xi_i$; and $\xi_i \ge 0, i = 1 \ldots m$.

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

$$\max_{\alpha} \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \kappa\left(x_i, x_j\right) - \sum_{i=1}^{n} \alpha_i \tag{1}$$

$$\text{s.t.} 0 \le \alpha_i \le C, i = 1 \ldots m \tag{2}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3}$$

where $\kappa\left(x_i, x_j\right) = \varphi\left(x_i\right)^T \varphi\left(x_j\right)$ is a linear kernel function. Common examples are the linear kernel function $\kappa\left(x_i, x_j\right) = x_i^T x_j$, the polynomial kernel $\kappa\left(x_i, x_j\right) = \left(x_i^T x_j\right)^n$, and the RBF (Gaussian) kernel $\kappa\left(x_i, x_j\right) = \exp\left(-\frac{x_i - x_j^2}{2\sigma^2}\right)$. By defining the kernel function, the matrix is symmetric and positively semi-definite. After the kernel function is solved, $\omega = \sum_{i=1}^{n} \alpha_i y_i \varphi\left(x_i\right)$, so the decision function for any test vector x is:

$$\text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \kappa\left(x_i, x_j\right) + b\right) \tag{4}$$

where $\alpha_i$ is calculated using the search media optimization (SMO) method and b is calculated through the primal–dual relationship.

### Fisher Reconstruction Feature Matrix and SVM

A large amount of interference information is difficult to exclude by observation because the aquifers are linked to each other. Thus, it is difficult to divide the data using a single Fisher linear discriminant and the SVM machine learning algorithm; the data needs to be characterized to define the integral feature(s) (Kwak 2008; Huang et al. 2011).

First, the diagrams are processed by the Fisher principle to extract the characteristics of the diagrams and the overall data features of each sample are compressed into a value, $F_i = \omega_i x_i + b_i$. Then the Fisher linear discriminant values ($F_i$) are used to reconstruct the feature matrix as $X_i = \left\{x_i, F_i\right\}_{i=1}^{m}$ (Lee et al. 2008).

In addition, according to the Tobler's first law of geography, all things are related in space and nearby things are more related than distant ones (Tobler 1970). That means that the hydrochemical characteristics are affected by their geographical location (Dong et al. 2000; Hu et al. 2010). Thus, in this paper, the geographical location was considered to be a key factor in establishing the feature matrix. Finally, we built four classifiers with SVM, and used the reconstructed feature matrix to train the classifiers.

### Evaluation Model

The performance of the classifier was evaluated by the metrics of the F1-score. The F1-score is the harmonic mean of the precision (P) and recall (R), which are defined by the following two equations (Santhi and Bhaskaran 2014):

$$R = \frac{tp}{tp + fn} \tag{5}$$

$$P = \frac{tp}{tp + fp} \tag{6}$$

where: P is the percentage of positive predictions that are correct; R is the percentage of positive labeled instances that were predicted as positive; tp stands for the true positive of the particular instance; fp stands for the false positive of the particular instance; and fn is the false negative of the particular instance. We can express the value of the F1-score as the following formula:

$$F1 = \frac{2RP}{R + P} \tag{7}$$

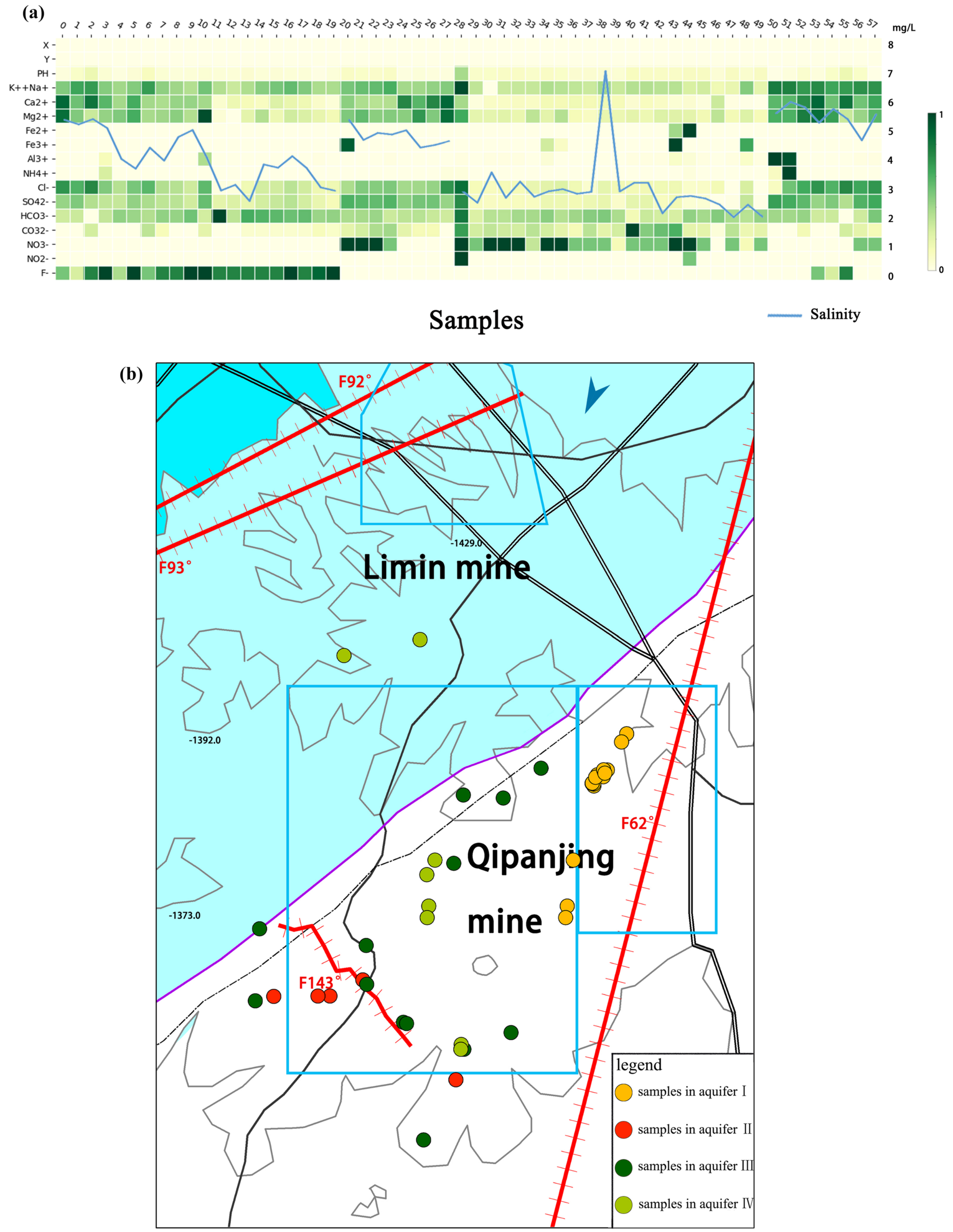

**Fig. 2** The thermodynamic chart analysis. **a** hydrogeological information around the Qipanjing mine and **b** water sampling points

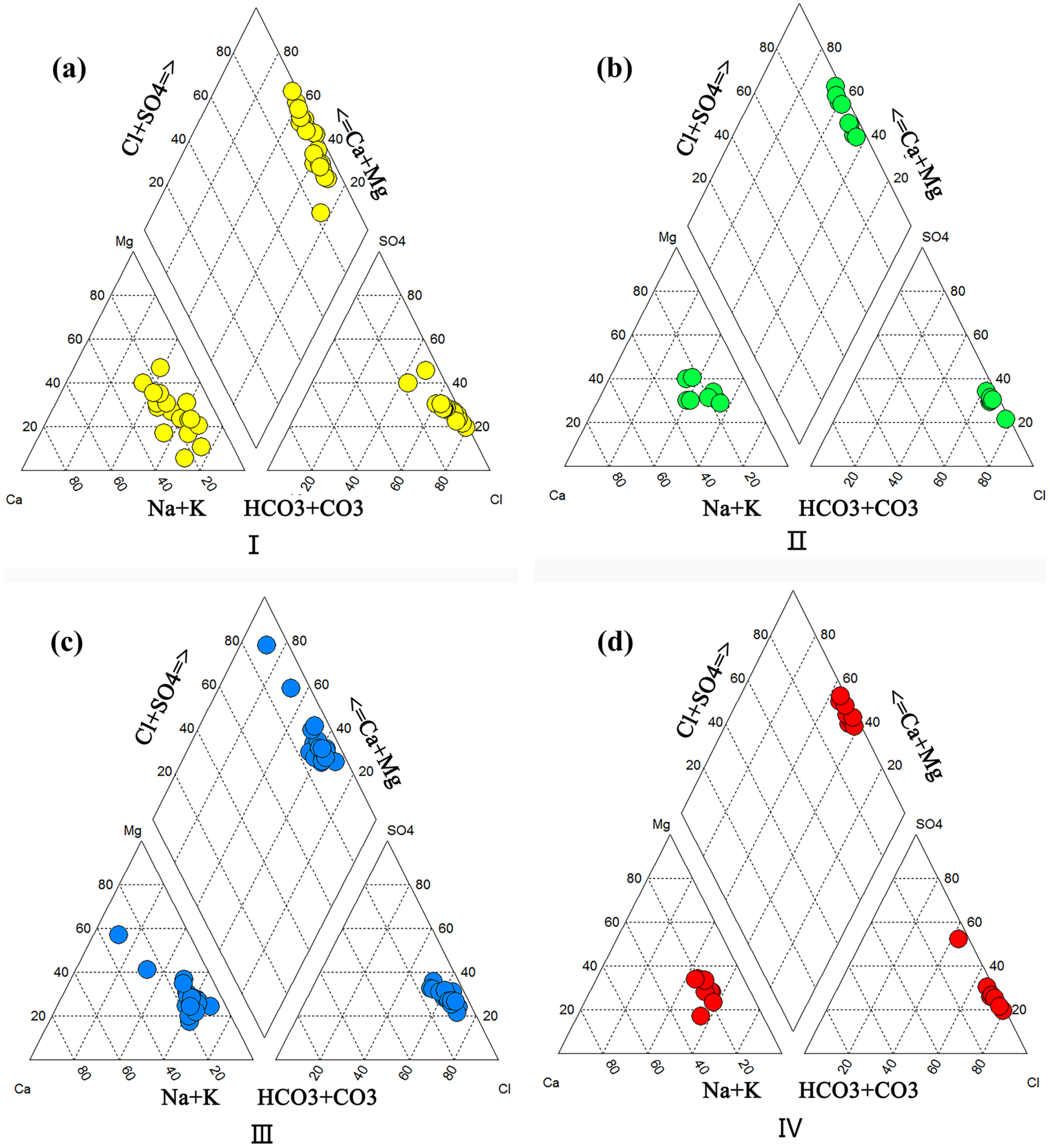

Fig. 3 The Piper trilinear diagram analysis

## Results and Discussions

### Screening of Typical Water Samples

According to the survey report, we ranked the data in coordinates from the northeast to the southwest, and marked the point information on the map (Fig. 2). This showed that the salinity tends to increase in the direction of water flow. We observed that seven main factors covered about 80% of the raw data information and could effectively represent the data information of the original samples.

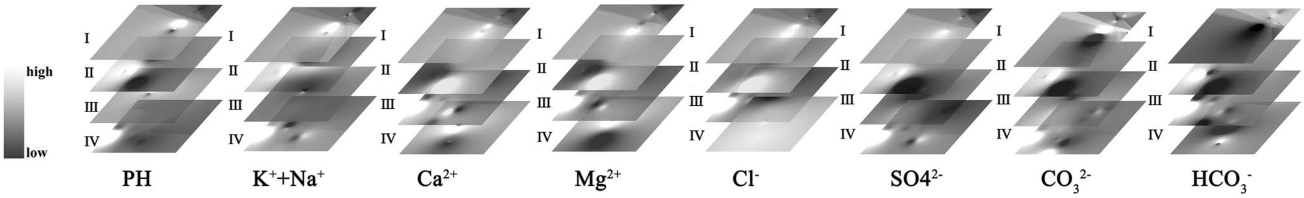

**Fig. 4** The IDW diagram of the icon in each aquifer

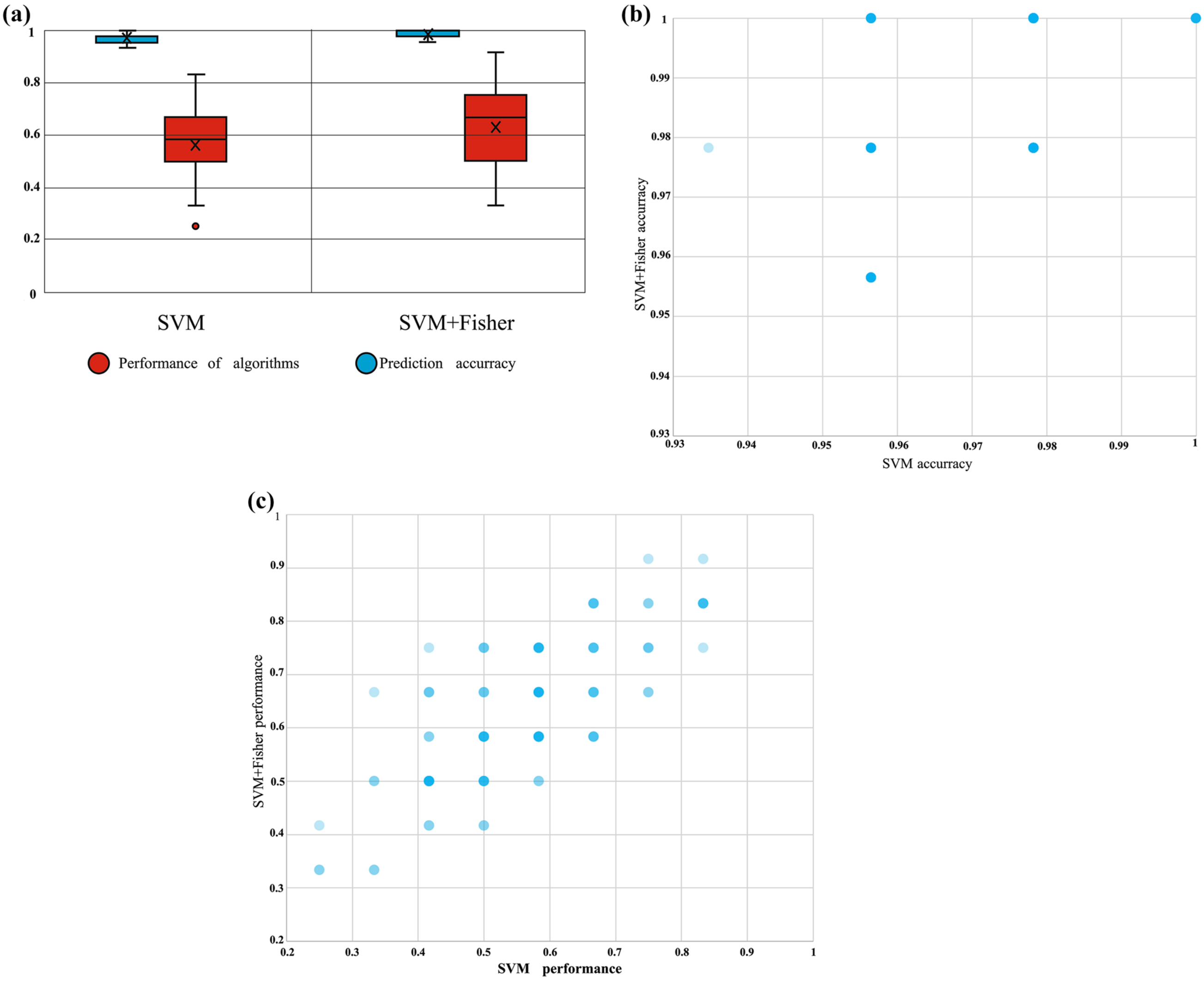

**Fig. 5** **a** A box diagram of performance and accuracy, **b** performance of algorithms, and **c** prediction accuracy

Figure 3 shows that the ion content was relatively stable, except for the floor sandstone aquifer in which the variation ranges of both $Ca^{2+}$ and $HCO_3^- + CO_3^{2-}$ were small. The $Na^+$+$K^+$ and $Cl^-$ levels were high in all aquifers. The water types of the aquifers were Na–Cl–$SO_4$ and Na–Ca–Cl–$SO_4$. Sample numbers 13, 14, 41, 47, and 51 were far from other samples and were thus excluded. Back to the original data, the pH of water samples 2 and 34 were much greater than the others, and the concentrations of other diagnostic ions varied within the aquifers.

In addition, the same icons in each aquifer were compared by inverse distance weighting (IDW) diagrams (Fig. 4). The groundwater not only flowed from the northeast to the southwest, but also from the north to the south due to the complex terrain; this agrees with a former study (Jiang 2015). Based on the hydrochemical characteristics (Wang et al. 2016), the FFE and SVM method were used to rapidly identify the water inrush sources in this study.

## Performance Analysis

To verify the accuracy of the established FFE and SVM water source recognition model, 80% groups of training samples were randomly selected for cross-analysis 100 times. Figure 5 shows the performance and accuracy of the SVM and the combined FFE and SVM classification techniques. The results show that the combined FFE and SVM algorithm performed better and was more balanced.

After calibration of the model, the result showed a 100% fit. Reducing the penalty term to 0.7, we found the model misidentified sample 54 (goaf water) as being from the floor sandstone aquifer. When the punishment term was set at 0.5, all samples in Aquifer IV were misidentified as Aquifer II, and when the punishment term was reduced to 0.45, samples 26 and 24 (Aquifer II) were regarded as Aquifer II. This suggests that goaf water and Aquifer II are well connected and that small joint surfaces or cracks near where samples 26 and 24 were collected connected Aquifers I and II. Aquifer III, which had little surrounding fracture development, was not misidentified.

The above analysis indicated that the established model was successfully applied in this study. The water inrush source was identified based on both parameter optimization and finite data. It was affected by data randomness and accuracy. Thus, the hydrogeological conditions and underground hydrochemical characteristics of the aquifers must be sufficiently understood to apply the model to a single mine.

## Conclusions

In this paper, a combined FFE and SVM method was proposed to identify a water inrush source, after giving due consideration to the influence of geographical factors. The approach outperformed the original and SVM model. Compared to the original SVM, it adds an axis direction to the total sample matrix, which can increase the accuracy and improve the model's generalization. Performing cross-analysis 100 times with 12 random samples as test data showed that the combined method was 12.1% more accurate than SVM alone.

By combining ion concentrations and hydrodynamics with geographic coordinates, we were able to analyze the direction of groundwater runoff and get the general location of groundwater recharge and discharge areas. This work also demonstrated that geographic location plays an important role in water inrush sources.

At this site, the sandstone aquifer and goaf water may be connected by some fractures. Also, there are many cracks between the roof sandstone and floor sandstone aquifers near samples 26 and 24, and some fractures exist in the rocks surrounding aquifer III. A better underground hydrochemical database is needed for future research.

**Acknowledgements** This study was financially supported by the Ministry of Science and Technology of China (2017YFC0804104), the Consultation Project of Chinese Academy of Engineering (2017-ZD-03-05-01), the National Natural Science Foundation (U1710258), and the China Postdoctoral Science Foundation (2018T110134).

## References

Chen ZY, Zhang GZ, Wu CF, Yang SQ (2009) Application of water source identification in mine inflow based on support vector machines. Jiangxi U Sci Tech J 30:10–13 **(In Chinese)**

Dong DL, Wu Q, Sun GM, Qian ZJ (2000) Research on water-filling patterns and countermeasures–a case study of Xiaotun coal mine. Int J Min Sci Tech 1:45–48

Hu WW, Ma ZY, Cao HD, Liu F, Li T, Dou HP (2010) Application of isotope and hydrogeochemical methodsin distinguishing mine bursting water source. J Earth Sci Environ 32:268–271. https://doi.org/10.3969/j.issn.1672-6561.2010.03.009

Huang PH, Wang XY (2018) Piper-PCA-Fisher recognition model of water inrush source: a case study of the Jiaozuo mining area. Geofluids 5:1–10. https://doi.org/10.1155/2018/9205025

Huang YM, Zhang GB, Dong F, Da FP (2011) Multiple features extraction using Gabor wavelet transformation. Fisher faces and integrated SVM with application to facial expression recognition. Appl Res Comp 28:1536–1539 **(In Chinese)**

Jiang SF (2015) Study on the prediction of the water in rush from coal seam floor during open coal mining-take a certain coal mine in Erdos as example. PhD diss, Inner Mongolia University of Science and Technology (**In Chinese**)

Kwak N (2008) Feature extraction for classification problems and its application to face recognition. Pattern Recogn 41:1701–1717

Lee JJ, Uddin MZ, Kim TS (2008) Spatiotemporal human facial expression recognition using Fisher independent component analysis and hidden Markov model. Proc Int Conf IEEE Eng Med Biol Soc. https://doi.org/10.1109/iembs.2008.4649719

Ren H, Chang YL (2005) Feature extraction with modified Fisher's linear discriminant analysis. In: Proceedings SPIE, vol 5995. Chemical and Biological Standoff Detection III, 599506 (4 November 2005), Boston, MA, United States. https://doi.org/10.1117/12.631885

Santhi P, Bhaskaran VM (2014) Detection of objects using fisher SVM with modified adaboost classification technique. J Theor Appl Info Tech 67:18–26

Shan SG, Cao B, Gao W, Zhao DB (2002) Extended Fisherface for face recognition from a single example image per person. IEEE. https://doi.org/10.1109/ISCAS.2002.1010929

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(sup1):234–240. https://doi.org/10.2307/143141

Wang X, 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. https://doi.org/10.1007/s12665-016-5616-3

Wei LY (2013) Study on main problems and control countermeasures of safety production in Shenhua Wuhai coal mine. Shen Hua Sci Technol 11:3–5 (**In Chinese**)

Yin NB, Hao J, Yu SC (2017) Research on establishment method of mine water source identificationmodel base based on artificial intelligence. J North Chin I Sci T 14:24–28 **(in Chinese)**

Zhang H, Berg AC, Maire M, Malk J (2006) SVM-KNN: discriminative nearest neighbor classification for visual category recognition Proc. IEEE Conf on Comput Vis Pattern Recognit 2:2126–2136. https://doi.org/10.1109/CVPR.2006.301