www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

Arch. Min. Sci. **64** (2019), 2, 321-333

Electronic version (in color) of this paper is available: http://mining.archives.pl

DOI 10.24425/ams.2019.128686

QIONG JIANG*, WEIDONG ZHAO**,*#, YONG ZHENG**, JIAJIA WEI**, CHAO WEI**

A SOURCE DISCRIMINATION METHOD OF MINE WATER-INRUSH BASED ON 3D SPATIAL INTERPOLATION OF RARE CLASSES

ANALIZA DYSKRYMINACYJNA ŹRÓDEŁ WYCIEKÓW WODY DO KOPALNI NA PODSTAWIE TRÓJWYMIAROWEJ INTERPOLACJI DANYCH O ZDARZENIACH RZADKICH

When the distribution of water quality samples is roughly balanced, the Bayesian criterion model of water-inrush source generally can obtain relatively accurate results of water-inrush source identification. However, it is often difficult to achieve desired classification results when training samples are imbalanced. Sample imbalance is common in the source identification of mine water-inrush. Therefore, we propose a three-dimensional (3D) spatial resampling method based on rare water quality samples, which achieves the balance of water quality samples. Based on the virtual water sample points distributed by the 3D grid, the method uses the 3D Inverse Distance Weighting (IDW) method to interpolate the groundwater ion concentration of the virtual water samples to achieve oversampling of rare water samples. Case study in Gubei Coal Mine shows that the method improves overall discriminant accuracy of the Bayesian criterion model by 5.26%, from 85.26% to 90.69%. In particular, the discriminative precision of the rare class is improved from 0% to 83.33%, which indicates that the method can improve the discriminant accuracy of the rare class to large extent. In addition, this method increases the Kappa coefficient of the model by 19.92%, from 52.26% to 72.19%, increasing the degree of consistency from “general” to “significant”. Our research is of significance to enriching and improving the theory of prevention and treatment of mine water damage.

Keywords: source discrimination; water-inrush; water quality; Bayesian classifier; rare class

W przypadku zrównoważonych danych o jakościowym rozkładzie próbek, zastosowanie kryterium Bayesowskiego do modelowania źródeł wycieków daje stosunkowo dokładne wyniki w analizie dyskryminacyjnej źródeł wycieków wody kopalnianej. Jednakże w przypadku niezrównoważonych danych, pożądane efekty kategoryzacji są niezmiernie trudne do uzyskania. Dane o składzie próbek są w znacznej mierze niezrównoważone, i jest to powszechny problem napotykany przy identyfikacji źródeł wycieków. W obecnej pracy zaproponowano więc trójwymiarową (3D) metodę powtórnego próbkowania z wykorzystaniem próbek wód z kategorii zdarzeń rzadkich, tak by uzyskać zrównoważony zbiór danych. W oparciu

* SCHOOL OF CIVIL & HYDRAULIC ENGINEERING, HEFEI UNIVERSITY OF TECHNOLOGY
** SCHOOL OF RESOURCE & ENVIRONMENTAL ENGINEERING, HEFEI UNIVERSITY OF TECHNOLOGY
# Corresponding author: zhaowd66@163.com

o wirtualne punkty na trójwymiarowej siatce, wykorzystano trójwymiarową metodę średniej ważonej odległością (Inverse Distance Weighing – IDW) do interpolacji stężenia jonów w wodach gruntowych w wirtualnych próbkach wody, w celu nadpróbkowania dla kategorii zdarzeń rzadkich. Studium przypadku kopalni węgla Gubei pokazuje, że metoda poprawia dokładność dopasowania modelu w oparciu o kryterium Bayesowskie o 5.25% (z 85.26% na 90.96 %). W szczególności, dokładność rozróżniania i dyskryminacji próbek należących do kategorii zdarzeń rzadkich wzrasta od 0% do 83.33%, co oznacza bardzo znaczna poprawę. Ponadto, wartość współczynnika Kappa wzrasta o 19.92%, od 52.26 % do 72.19%, tym samym podnosząc poziom zgodności metody z poziomu ogólnego na „znaczący". Prowadzone przez nas badania mają poważne znaczenie z punktu widzenia udoskonalenia teorii leżących u podstaw metod i technik zapobiegania i kontroli wycieków wód kopalnianych.

Słowa kluczowe: analiza dyskryminacyjna źródeł wycieków, wyciek wód, jakość wód, kryterium Bayesowskie, kategoria zdarzeń rzadkich

1. Introduction

Quick and effective identification of mine water-inrush is one of the important technical means to ensure coal mine safety production. Using hydrogeochemical information to quickly and accurately identify water-inrush sources is a hot topic in water disaster prevention research of coal mine in recent years. Various attempts have been made to quickly and accurately discriminate the source of water-inrush by use of groundwater hydrogeochemical information (Ma, 2010; Rina et al., 2012; Bagyaraj et al., 2013; Pantaleoni et al., 2013). Manzi et al. (2012) predicted possible passages of deep groundwater and methane gas of Witwatersrand Gold Mine in South Africa through 3D edge detection seismic attributes. Vincenzi et al. (2009) evaluated the effectiveness of drainage systems of underground tunnels by tracer tests and hydrological observation methods. However, few overseas studies focus on source discrimination of mine water-inrush by use of groundwater hydrogeochemical information. In China, Zhang applied the Bayesian discrimination method to source discrimination of water-inrush in Guqiao Coal Mine, and Ben et al. (2006) applied comprehensive fuzzy evaluation method to source discrimination of water-inrush of subsided columns in a mine in Shanxi Province. The discrimination methods mentioned above have strong dependence of training samples, and do not consider the imbalance of distribution of number of training samples. There always exists problems of imbalanced distribution of number of training samples when source discrimination of water-inrush is conducted by use of hydrogeochemical information, that is, the number of some training samples is significantly less than that of other classes (Umar et al., 2013). For convenience, this kind of sample that its number of samples is significantly less than that of other classes is called "rare class sample" in this paper.

For the problem of unbalanced classification, the main research focuses on feature selection, data distribution adjustment, and improvement of model training algorithms. The main feature selection methods include SYMON (Moayedikia et al., 2017), FAST (Feature Assessment by Sliding Thresholds) (Chen et al., 2008), etc. Chan et al. (2007) studied a lightweight intrusion detection system by use of feature selection approach. The method of data distribution adjustment mainly includes methods such as resampling and data grouping. The resampling method improves the recognition rate of rare classes by classifiers by increasing the number of training samples of rare classes and reducing the number of training samples of most classes, so that the imbalanced sample distribution becomes roughly balanced. The early method of increasing the number of rare class training samples was to copy rare class samples directly and randomly, but this would lead to over-fitting problems. Two methods proposed by Chawla et al. (2002) to

generate a few samples by synthetic methods, SMOTE and SMOTEBoost, are improvements to the early random oversampling technique, which can avoid the problem of model overfitting to some extent. The under-sampling method of most classes mainly includes CNN (Condensed Nearest Neighbor) (Alegria et al., 2000; Liang et al., 2017) and NCR (Neighborhood Cleaning Rule) (Li et al., 2009; Mishra et al., 2018). In the improvement of model training algorithm, the current research mainly focuses on cost-sensitive learning (Chai et al., 2004) and integrated learning methods (Yuan et al., 2013). Zou et al. (2011) used cost-sensitive learning methods to classify customers, help companies lock in high-end customers and dynamically adjust regional market strategies. The above research improves the classification accuracy of the unbalanced classification problem by different methods. However, in previous studies, no research has been done on the sample imbalance in the source identification of mine water-inrush. Therefore, we propose a rare class oversampling method by 3D spatial interpolation of rare class to try to solve the sample imbalance problem in the source identification of mine water-inrush to some extent.

2. The basic principles and method

2.1. The main principles of rare class resampling

The so-called imbalanced classification problem refers to the pattern classification problems that the distribution of number of training samples among different classes is imbalanced. When the number of samples of a specific class is far less than that of other types of classes (generally 10% less than other samples), the class is called "rare class" in this paper. Imbalanced sample distribution always leads to absolute or relative scarcity of rare class samples. Absolute scarcity refers to the fact that the absolute number of rare class samples is too small to reflect the hydrogeochemical information of groundwater. The study of artificial experimental data showed that the error rate of identifying the rare class with absolute scarcity is much higher than the average error rate (Weiss et al., 1995). Relative scarcity means that the absolute number of rare class samples is quite large, but its proportion is too small compared to other classes (i.e., generally 10% less than other samples). In this case, to identify a rare class is very difficult just like to find a needle in a haystack. The advent of the rare class with relative scarcity always decreases the effect of the greedy heuristic search method (Ye et al., 2009). The rare class always appears when we identify the source of mine water-inrush based on the Bayesian method and hydrogeochemical information of mine groundwater, which often seriously affects the discriminant accuracy of rare classes.

The resampling method mainly achieves the approximate balance of training samples by increasing the number of rare class training samples or reducing the number of training samples of most classes, so as to improve the discriminant accuracy of rare class. The most primitive method of increasing the number of rare class samples is to copy the samples of rare class directly and randomly, but this method always leads to "over learning" and cannot dramatically improve the recognition rate of rare class (Zhu et al., 2004). Because of the spatial continuity of water quality changes in groundwater systems with uniform hydraulic connections, it is reasonable to improve the balance of distribution of training samples of rare class by resampling method. Therefore, the main goal of this paper is to achieve the balance of rare class samples by interpolating water quality of rare class, thus improving the discriminant accuracy of rare classes.

2.2. 3D interpolation method for rare classes

The technical process of our method is show in figure 1. First, we should make sure that there is no less than one rare class in the water samples. Second, virtual water samples of rare classes should be generated by 3D spatial resampling method. Finally, the virtual water samples of rare class and other actual water samples are added to the Bayesian classifier for the source discrimination of mine water-inrush.

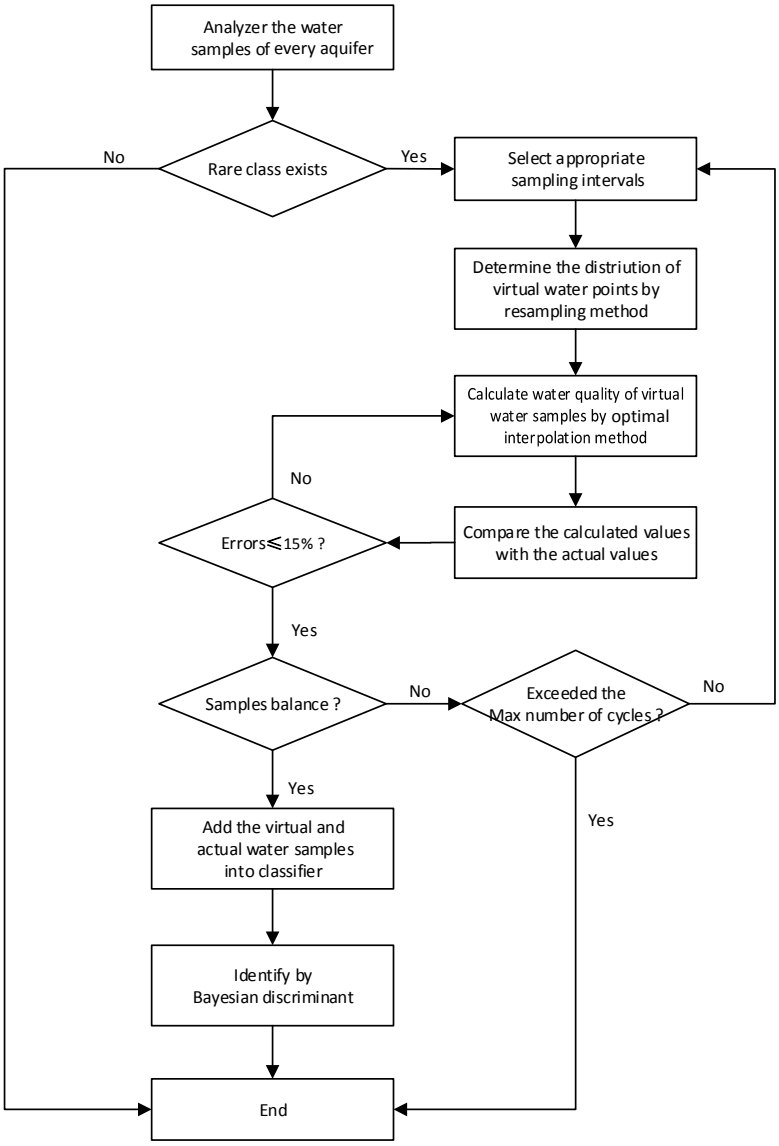

Fig. 1. The technical process of the method

In addition, in order to ensure a better interpolating effect, we should try to increase the number of rare water samples in the interpolated neighborhood. It is generally required that the number of rare samples should be at least 3 to 6 or more, and these samples should be evenly distributed in the interpolated neighborhood to achieve better interpolation. In order to minimize the error caused by interpolation, only those virtual water samples with less than 15% of the interpolation error can be added to the Bayesian classifier. When the number of all water samples (including virtual water samples of rare class and other actual water samples) is balanced, the interpolation of rare classes will end. In order to prevent the occurrence of an infinite loop in the resampling process, generally a maximum number of cycles should be set according to the number of samples. If this condition is reached, the cycle calculation will end.

2.2.1. Generation of virtual samples of rare class

In this study, appropriate amounts of virtual water samples are generated by 3D spatial resampling method and added into the Bayesian classifier as training samples, which makes the number of different kinds of samples balanced. These virtual samples are generated according to the following principles: (1) it is necessary to generate virtual samples from the virtual samples matrix of hydrogeochemical field of rare class in terms of the category of virtual samples; (2) the total number of different types of water samples should be equal in terms of the number of virtual samples. In other words, the total number of virtual and actual water samples of each rare class should be nearly equal to the number of other normal classes, and (3) the spatial distribution of sample points should be relatively uniform in the study area. The optimal planar and vertical sample interval are determined according to the above principles and the matrix of virtual water samples can be established by resampling method.

2.2.2. Interpolation of water quality of virtual water samples

The water quality of virtual water samples can be calculated out using GIS spatial interpolation methods after the matrix of virtual water samples has been established. The IDW method is adopted because of its advantages of simple concept, high speed, etc. (Wang et al., 1995). The most important reason is that it can be easily extended to 3D space. The IDW method is given by:

$$u(x) = \sum_{i=0}^{N} \frac{w_i(x)u_i}{\Sigma_{j=0}^{N} w_j(x)}$$

In which,

$$w_i(x) = \frac{1}{d(x, x_i)^p}$$

Where x represents the virtual point. $u(x)$ is the interpolation results of attribute values of the virtual point x. x_i represents the i-th sampling point. u_i is the actual value of the i-th sampling point. $W_i(x)$ is the weight of the point x_i; $d(x, x_i)$ is the Euclidean distance between x_i and x. P is an exponent of the distance $d(x, x_i)$. Its value varies under different conditions, and its value is two in our study.

In particular, under the condition of two-dimensional spatial interpolation,

$$d(x, x_i) = \sqrt{(X_x - X_{xi})^2 + (Y_x - Y_{xi})^2}$$

Under the condition of three-dimensional spatial interpolation,

$$d(x,x_i)=\sqrt{(X_x-X_{xi})^2+(Y_x-Y_{xi})^2+(Z_x-Z_{xi})^2}$$

Where $X(X_x, Y_x, Z_x)$ and X_i (X_{xi}, Y_{xi}, Z_{xi}) represents the 3D coordinates of the two points respectively, and $X(X_x, Y_x)$ and X_i (X_{xi}, Y_{xi}) represents the two-dimensional coordinates of the two points respectively.

2.2.3. The Bayesian discrimination method based on balanced samples

The virtual water samples and other actual water samples except the water samples need to be identified are all taken as training samples, and they are used to discriminate the source of water-inrush by the Bayesian methods. The Bayesian discrimination is a method based on probability and statistics. It requires a certain understanding of the research object and uses prior probability to describe this understanding. Based on the prior probability, Bayesian model uses the probability density of multivariate normal distribution to calculate the posterior probability of a sample, and calculates the probability that the sample falls into each category, and considers that the class with the maximum posterior probability is the category to which the sample belongs. For a given m-variate aquifer categories $G_1, ..., G_K$ $(K > 2)$ with K aquifers, the vector of their average and their covariance matrices are $\mu^{(i)}$, $\sum_i (i=1,2,\ldots,K)$. For any given water sample $X = (x_1, ..., x_m)'$, the formula for calculating the posterior probability $P(t|X)$ of it belonging to an aquifer G_t is as follow:

$$P(t\mid X)=\frac{\exp\left(-0.5D_t^2(x)\right)}{\sum_{i=1}^{k}\exp\left(-0.5D_i^2(x)\right)} \tag{1}$$

Where: $D_i^2(x)$ is the generalized squared distance from X to the i-th aquifer. The generalized squared distance of X to the t-aquifer $D_i^2(X)$ is calculated as follows:

$$D_t^2(X)=(X-\mu)'\sum{}^{-1}(X-\mu)+g_1(t)+g_2(t) \tag{2}$$

Where: $\mu = (\mu_1, ..., \mu_m)'$ is the mean vector of groundwater ion concentration, $\sum = (\sigma_{ij})_{m\times m}$ is the covariance matrix of groundwater ion concentration.

$$g_1(t)=\begin{cases}\ln|S_t|, & \textit{If the covariance matrices } \sum_i \textit{ of each aquifer are not equal}\\ 0, & \textit{If the covariance matrices } \sum_i \textit{ of the aquifers are exactly equal}\end{cases} \tag{3}$$

$$g_2(t)=\begin{cases}-2\ln|q_t|, & \textit{If the prior probability } q_t \textit{ of each aquifer are not equal}\\ 0, & \textit{If the prior probability } q_t \textit{ of the aquifers are exactly equal}\end{cases} \tag{4}$$

Where: S_t is the covariance matrix of the water sample in the t-th aquifer, and q_t is the prior probability of the water sample of the t-th aquifer. The Bayesian model uses the criterion of posterior

probability to determine $X \in G_t$, when $P(t|X) > P(i|X)$, $i \neq t (i = 1, ..., k)$. In other words, the criterion for the Bayesian model to discriminate the source of a water-inrush is that the water sample should belong to the aquifer with the highest posterior probability. The main purpose of this study is to use the Bayes model and the water quality of the mine water-inrush to determine which aquifer the source of the water-inrush originated from. Therefore, the probability of occurrence of six major ion concentrations ($m = 6$) in different aquifers, such as $K^+ + Na^+$, Ca^{2+}, Mg^{2+}, HCO^{3-}, Cl^-, SO_4^{2-}, in groundwater is used to comprehensively determine the source of water-inrush. The source of the water -inrush contains three groundwater aquifers ($K = 3$).

According to the results of discrimination, the confusion matrix and the Kappa coefficient can be calculated out. The Kappa coefficient is a method of classification accuracy based on the confusion matrix. The Kappa coefficient can reflect the extent that the discriminant accuracy of the Bayesian method is superior to the discriminant accuracy through random assigning a specific category to each point in the statistical sense. In other words, it can be used to evaluate the classification accuracy of the Bayesian method. The classification and evaluation criteria of the Kappa coefficient proposed by Cohen in 1968 (Xu et al., 2011) is shown in table 1 and adopted to evaluate the classification accuracy of the Bayesian method.

TABLE 1

The classification criteria of the Kappa coefficient

Kappa	<0.00	0.00 ~ 0.20	0.21 ~ 0.40	0.41 ~ 0.60	0.61 ~ 0.80	0.81 ~ 1.00
Degree of consistency	Poor	Micro-Weak	Weak	General	Notable	Optimal

3. Case Study

3.1. Study area

The Gubei Mine is located at about 23 km northwest of Fengtai County of Huainan City in Anhui Province, and located in the western Huainan Coalfield. It is 7.5 km long, 4.5 km wide and covers an area of about 34 km^2. There are significant differences of hydrogeological conditions between its shallow aquifers and deep aquifers. The hydrogeological conditions controlled by regional tectonic and new tectonic movement are relatively complex, which make it difficult to effectively prevent and treat the disaster of mine water.

The main aquifers of the Gubei Mine include the Cenozoic loose aquifers, the Permian sandstone aquifers, and the Ordovician karst aquifers. The thickness of the Cenozoic loose aquifers that directly overlay the Permian coal measures ranges from 390.35 m ~ 509.10 m. According to the penetration of saturated rocks, the Cenozoic loose stratums are divided into five aquifers, five water-resisting layers, and a "gravel layer" (also called "Red Layer") from top to bottom. The five aquifers are as follows: (1) the top segment of Upper Aquifer of the Cenozoic loose aquifers (hereinafter referred to as "Upper Aquifer"); (2) the bottom segment of the Upper Aquifer; (3) the top segment and bottom segment of Middle Aquifer of the Cenozoic loose aquifers, and (4) the Bottom Aquifer of the Cenozoic loose aquifers ("Bottom Aquifer"). The water in the Bottom Aquifer is our main study object. The upper boundary of the Bottom Aquifer

is a water-resisting layer containing the light gray-green, gray-green thick layer of consolidated clay and sandy clay. The lower boundary is the water-resisting layer of the Bottom Aquifer composed of gray-green, reddish brown consolidated clay, sandy clay, etc. The Bottom Aquifer consists of light gray fine and silt sand layer sandwiched between the main purple gravel and clay. According to the regional pumping data, its water level ranges from 26.18 m to 26.45 m, and it is a rich aqueous rock stratum.

3.2. Data processing and analysis

We collected 95 water samples from the Gubei Mine, including 6 samples of the Bottom Aquifer, 79 samples of the Coal Measure Aquifer, and 10 samples of the Taiyuan Formation Limestone Aquifer ("Tai Limestone"). From the point of view of absolute quantity, the quantity of water samples of both the Bottom Aquifer and the Tai Limestone are less than 10, which should be considered as the absolute scarcity of sample numbers. From the point of view of relative quantity, the ratio of quantity of water samples of the Bottom Aquifer and the Coal Measure Aquifer is 1:13.2, while the ratio of the Tai Limestone and the Coal Measure Aquifer is 1:7.9. This situation should be regarded as the relative scarcity of sample numbers. Therefore, the classification problem is a typical imbalance classification problem and both the Bottom Aquifer and the Tai Limestone are considered as rare class. There are 77 virtual water samples (see Fig. 2) within the scope of the study area when the sampling interval is 900 m. The sample numbers of the Bottom Aquifer, the Coal Measure Aquifer and the Tai Limestone are 83, 79, and 87 respectively when the virtual samples are added into the classifier, which make the three kinds of samples approximately balanced.

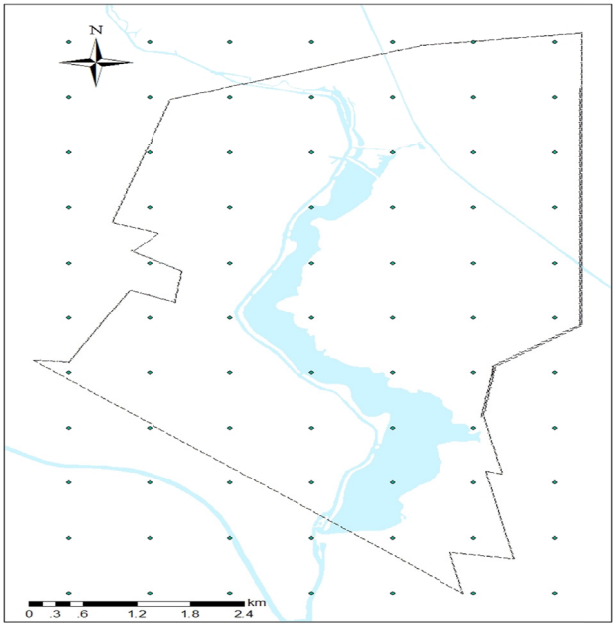

Fig. 2. Distribution of water samples

The spatial distribution of water quality of the Bottom Aquifer changes continuously because that the Bottom Aquifer is a pore aquifer system with unified groundwater level. Therefore, it is reasonable to interpolate the water quality of different water samples in the Bottom Aquifer by use of the IDW method. It should be noted that the distance from a virtual point to the actual point should be three-dimensional distance. In order to improve the interpolation accuracy, we collected water samples of the Gubei Mine and its surrounding mines because that the quantity of water samples of the Gubei Mine is too few to cover its whole territory. All of the water samples of the Bottom Aquifer of the Gubei Mine and its surrounding mines are shown in table 2.

TABLE 2

Water samples of the Bottom Aquifer of the Gubei Mine and its surrounding mines (Unit: mg/L)

Mine Name	id	*X* (m)	*Y* (m)	Ca^{2+}	Mg^{2+}	$K^+ + Na^+$	HCO_3^-	Cl^-	SO_4^{2-}
Gubei Mine	1	39458387	3635848	41.4	23.64	839.06	305.4	1031.91	292.17
…	…	…	…	…	…	…	…	…	…
Gubei Mine	6	39456985	3636106	35.47	12.4	415.89	190.99	529.27	135.21
Xieqiao Mine	7	39442403	3629692	47.33	20.64	785.43	260.86	1027.54	228.38
North Zhangji Mine	8	39450180	3628741	56.61	33.79	691.37	168.87	966.6	264.19
North Zhangji Mine	9	39450086	3628754	54.65	34.38	828.86	268.95	1087.26	307.39
Dingji Mine	10	39467179	3641136	26.51	11.7	741.64	334.81	786.9	274.57
…	…	…	…	…	…	…	…	…	…
Dingji Mine	20	39459388	3637582	22.48	14.88	979.57	253.18	1021.61	567.26
Zhuji Mine	21	39485800	3636953	53.44	30.44	728.48	283.71	954.17	240.37
Zhuji Mine	22	39480394	3639128	67.33	36.94	784.02	300.75	1031.11	296.76

The interpolation results of concentration of main ions of groundwater of the Bottom Aquifer are shown in table 3. As shown in table 3, the average errors of interpolation results are about 15%, which indicates that the interpolation results of the concentration of each ion combination

TABLE 3

The average errors of interpolated water quality of the Bottom Aquifer (Unit: mg/L)

ID	Ca^{2+}		Mg^{2+}		$K^+ + Na^+$		HCO_3^-		Cl^-		SO_4^{2-}	
	Interpolated value	Actual value	Interpolated value	Actual value	Interpolated value	Actual value	Interpolated value	Actual Value	Interpolated value	Actual value	Interpolated Value	Actual value
1	39.80	41.4	19.97	23.64	775.01	839.06	300.79	305.4	930.23	1031.91	287.67	292.17
2	38.80	31.06	19.94	24.83	789.71	964.97	295.63	374.66	930.21	1170.08	318.81	307.81
3	39.85	42.89	20.31	19.7	730.23	830.86	275.90	347.2	893.43	1009.62	263.06	275.15
4	40.89	40.88	18.56	26.02	696.49	803.04	276.45	289.23	857.93	979.84	242.33	290.38
5	36.61	35.47	14.88	12.4	510.22	415.89	218.67	190.99	637.32	529.27	172.68	135.21
6	46.82	49.5	21.93	21.4	837.27	841.2	319.72	322.19	1034.01	1047.55	290.37	287.91
Average Error (%)	7.42		14.91		12.38		10.46		12.67		9.10	

of virtual water samples are reliable. Therefore, it is reasonable to consider the virtual water samples as actual water samples, which can play important role in increasing the quantity of training samples of the rare class.

3.3. Results and discussion

The results of source discrimination of water-inrush by use of the Bayesian discrimination method based on unbalanced samples are shown in table 4. The actual aquifers represent the aquifers to which the water samples belong. For example, the first water sample (i.e., ID = 1) actually belongs to the Bottom Aquifer, but it has been falsely identified as the water samples coming from the Tai Limestone. The results of source discrimination of water-inrush by use of the Bayesian discrimination method based on balanced samples are shown in table 5.

TABLE 4

The results of the Bayesian discrimination method based on unbalanced samples (Unit: mg/L)

ID	Ca^{2+}	Mg^{2+}	$K^+ + Na^+$	HCO^{3-}	Cl^-	SO_4^{2-}	Actual aquifers	Discrimination aquifers
1	41.4	23.64	839.06	305.4	1031.91	292.17	1	3*
2	31.06	24.83	964.97	374.66	1170.08	307.81	1	3*
3	42.89	19.7	830.86	347.2	1009.62	275.15	1	3*
4	40.88	26.02	803.04	289.23	979.84	290.38	1	3*
5	49.5	21.4	841.2	322.19	1047.55	287.91	1	3*
…	…	…	…	…	…	…	…	…
93	29.15	25.65	862.2	439.34	1019.89	244.84	3	3
94	30.36	0	858.68	317.3	1039.88	280.64	3	3
95	40.08	0	602.78	349.33	554.94	477.34	3	2*

Note: * Indicates a wrong identification of water samples, 1 – the Bottom Aquifer, 2 – the Coal Measure Aquifer, 3 – the Tai Limestone.

According to the above discrimination results, the confusion matrix are calculated out and shown in table 6. The results of comparative analysis of the two discrimination methods are shown in figure 3. As can be seen from table 6, though the overall accuracy of the Bayesian method based on unbalanced samples is about 85.26%, its Kappa coefficient is only about 52.26%, indicating that its degree of consistency is general level. Especially, the discrimination effect of the rare class is not ideal. For example, the error rate of discrimination of the Bottom Aquifer is 100% and that of the Tai Limestone is about 60%, which indicates that the discrimination effect of the rare class is not ideal by use of the Bayesian method based on unbalanced samples though its overall discrimination effect is relatively good. In addition, the Bottom Aquifer is the main source of water-inrush of the Gubei Mine. Therefore, the overall discrimination effect of the Bayesian method is not ideal indeed, especially for the discrimination of the two rare classes.

Compared with the Bayesian method based on unbalanced samples, the overall accuracy of the Bayesian method based on balanced samples improves about 5% and achieves to around 90.69%. In addition, the Kappa coefficient increases by about 20% and achieves to about 72.19%, indicating that its degree of consistency is significant level. Especially, the discrimination accuracy

TABLE 5

The results of the Bayesian discrimination method based on balanced samples

ID	Ca^{2+}	Mg^{2+}	$K^+ + Na^+$	HCO^{3-}	Cl^-	SO_4^{2-}	Actual aquifers	Discrimination aquifers
1	41.4	23.64	839.06	305.4	1031.91	292.17	1	1
2	31.06	24.83	964.97	374.66	1170.08	307.81	1	1
3	42.89	19.7	830.86	347.2	1009.62	275.15	1	1
4	40.88	26.02	803.04	289.23	979.84	290.38	1	1
5	49.5	21.4	841.2	322.19	1047.55	287.91	1	1
6	35.47	12.4	415.89	190.99	529.27	135.21	1	3*
…	…	…	…	…	…	…	…	…
93	29.15	25.65	862.2	439.34	1019.89	244.84	3	3
94	30.36	0	858.68	317.3	1039.88	280.64	3	3
95	40.08	0	602.78	349.33	554.94	477.34	3	3

Note: * Indicates a wrong identification of water samples, 1 – the Bottom Aquifer, 2 – the Coal Measure Aquifer, 3 – the Tai Limestone.

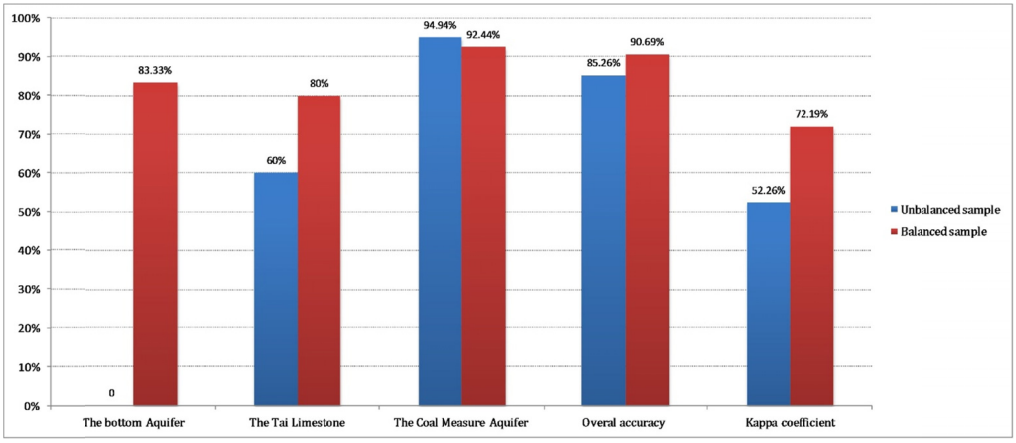

Fig. 3. The results of comparative analysis of unbalanced samples and balanced samples

TABLE 6

The confusion matrix of unbalanced samples and balanced samples

Results / Actual aquifer	Results of unbalanced samples				Results of balanced samples			
	1	2	3	Total	1	2	3	Total
1	0	1	5	6	5	0	1	6
2	3	75	1	79	2	73	4	79
3	3	1	6	10	2	0	8	10
Total	6	77	12	95	9	73	13	95

Note: 1 – the Bottom Aquifer, 2 – the Coal Measure Aquifer, 3 – the Tai Limestone.

of the Bottom Aquifer greatly increases from 0% to about 83.33% and that of the Tai Limestone increases by 20%. Although the discrimination accuracy of the Coal Measure Aquifer slightly decreased from about 94.94% down to around 92.41%, it still has a high discrimination accuracy. In summary, the Bayesian method based on balanced samples improves the overall discrimination accuracy, in particular, greatly improves the discrimination accuracy of the rare class.

4. Conclusion

In order to solve the imbalanced classification problem of source discrimination of water-inrush in the Gubei Mine, we propose a three-dimensional (3D) spatial resampling method based on rare water quality samples, which achieves the balance of water quality samples. Our study shows that the resampling method can effectively increase the virtual water samples of the rare classes and improve the balance of different classes. The study results of the Gubei Mine show that the Bayesian method based on balanced samples not only increases the overall discrimination accuracy but also improves the Kappa coefficient in the absence of any additional training samples. In particular, it greatly improves the discrimination accuracy of the rare classes, such as the Bottom Aquifer and the Tai Limestone, indicating that our method can increase the discrimination accuracy of the rare class to some extent. Our study is a beneficial attempt to improve the discrimination accuracy of mine water-inrush, and whether this method has universal applicability for other kinds of aquifers needs further research.

References

Aleegria F.C., Serra A.C., 2000. *Computer vision applied to the automatic calibration of measuring instruments* [J]. Measurement **28** (3), 185-195.

Bagyaraj M., Ramkumar T., Venkatramanan S., 2013. *Application of Remote Sensing and GIS Analysis for Identifying Groundwater Potential Zone in Parts of Kodaikanal Taluk* [J]. Frontiers of Earth Science **7** (1), 65-75.

Ben Xudong, Guo Haiying, XIE Yiwei, 2006. *The Application of Fuzzy Comprehensive Evaluation to Discrimination of Mine Water-Inrush Source* [J]. Mining Safety & Environmental Protection (03), 57-59 (in Chinese).

Chai X., Deng L., Yang Q., 2004. *Test-cost sensitive naïve bayes classification* [C]. IEEE International Conference on Data Mining, 2004(ICDM'04). IEEE,51-58.

Chawla N. V., Bowyer K.W., Hall L.O., 2002. *SMOTE:synthetic minority over-sampling technique* [J]. Journal of artificial intelligence researth **16** (1), 321-357.

Chen X., Wasikowski M., 2008. *Fast:a roc-based feature selection metric for small samples and imbalanced data classification problems* [C]. 14th ACM SIGKDD International Coference on Knowledge Discovery and Data Mining. ACM, 124-132.

Chen You, Chen Xueqi, Li Yang, 2007. *Lightweight Intrusion Detection System Based on Feature Selection* [J]. Journal of Software (07): 1639-1651 (in Chinese).

Li B., Jia Z., 2009. *Some results on condition numberss of the scaled total least squares problem* [J]. Linear Algebra&Its Applications **435**(3), 674-686.

Liang Tianming, Xu Xinzheng, Xiao Pengcheng, 2017. *A new image classification method based on modified condensed nearest neighbor and convolutional neural networks* [J]. Pattern Recognition Letters **94**, 105-111.

Ma Lei, 2010. *A GIS-Based System for Mine Water-Inrush Source Quick Discrimination with Comprehensive Information* [J]. Hei University of Technology (in Chinese).

Manzi M., Durrheim R.J., Hein K., 2012. *3D Edge Detection Seismic Attributes Used to Map Potential Conduits for Water and Methane in Deep Gold Mines in the Witwatersrand Basin* [J]. Geophysics **77** (5), 133-147 (South Africa).

Mishra B.K., Shkla P., Madhu S.V., 2018. *Prevalence of double diabetes in youth onset diabetes patients fromeast Delhi and neighboring NCR region* [J]. Diabetes & Metabolic Syndrome: Clinical Research & Reviews **12**, 839-842.

Moayedikia A., Ong K.L., Boo Y., 2017. *Feature selection for high dimensional imbalanced class data using harmony search* [J]. Engineering Application of Artificial Intelligence **55**, 38-49.

Pantaleoni E., 2013. *Combining a Road Pollution Dispersion Model with GIS to Determine Carbon Monoxide Concentration in Tennessee* [J]. Environmental Monitoring & Assessment **185** (3), 2705-2722.

Rina K., Datta P., Singh C., 2012. *Characterization and Evaluation of Processes Governing the Groundwater Quality in Partso of the Sabarmati Basin,Gujarat Using Hydrochemistry Integrated with GIS* [J]. Hydrological Processes **26** (10), 1538-1551.

Umar M., Waseem A., Sabir M. A., 2013. *The Impact of Geology of Recharge Areas On Groundwater Quality: A Case Study of Zhob River Basin, Pakistan* [J]. Clean-soil Air Water **41** (2), 119-127.

Vincenzi V., Gargini A., Goldscheider N., 2009. *Using Tracer Tests and Hydrological Observations to Evaluate Effects of Tunnel Drainage On Groundwater and Surface Waters in the Northern Apennines* [J]. Hydrogeology Journal **17** (1), 135-150 (Italy).

Wen Yimin, Li Jian, Du Feiming, 2009. *Using Ensemble Learning Strategy to Handle Class Imbalance Problems* [J]. Computing Technology and Automation **110** (02): 103-106 (in Chinese).

Weiss G.M., 1995. *Learning with rare cases and small disjuncts //Proceedings of the 12th International Conference on Machine Learning* [J]. San Francisco: Morgan Kaufmann, 58-565.

Wang Dachun, Zhang Renquan, Shi Yihong, 1995. *Basis Hydrogeology Beijing: Geological Publishing House*, 113 -115(in Chinese).

Xu Fei, Zheng Changjiang, Yang Cheng, 2012. *Identification Method of Traffic Congestion Based on Resampling* [J]. Journal of Highway and Transportation Research and Development **203** (11), 140-144 (in Chinese).

Xu Wenning, Wang Pengxin, Han Ping, 2011. *Application of Kappa coefficient to accuracy assessments of drought forecasting model:a case study of Guanzhong Plain* [J]. Journal of Natural Disasters (06), 81-86 (in Chinese).

Ye Zhifei, Wen Yimin, Lv Baoliang, 2009. *A survey of imbalanced pattern classification problems* [J]. CAAI Transactions on Intelligent Systems **16** (02), 148-156 (in Chinese).

Yuan Xingmei, Yang Ming, Yang Yang, 2013. *A structured SVM integrated classifier for unbalanced data* [J]. PR&AI **26** (3), 215-320 (in Chinese).

Zhang Chunlei, Qian Jiazhong, Zhao Weidong, 2010. *The Application of Bayesian Approach to Discrimination of Mine Water-Inrush Source* [J]. Coal Geology & Exploration **220** (04), 34-37 (in Chinese).

Zhao Nan, Zhang Xiaofang, Zhang Lijun, 2018. *A Survey of Unbalanced Data Classification Research* [J]. Computer Science **45** (6a), 22-27 (in Chinese).

Zou Peng, Yu Bo, Wang Xianquan, 2011. *Cost-Sensitive Learning Method with Data Drift in Customer Segmentation* [J]. Journal of Harbin Institute of Technology **43** (01), 119-124 (in Chinese).

Zhu Qiuan, Zhang Wangchang, Yu Yunhui, 2004. *The Spatial Interpolations in GIS* [J]. Journal of Jiangxi Normal University (Natural Sciences Edition) (02), 183-188 (in Chinese).