




Article

A Machine Learning Method for Engineering Risk Identification of Goaf

Haiping Yuan ¹, Zhanhua Cao ¹, Lijun Xiong ¹, Hengzhe Li ¹ and Yixian Wang ^{1,2,*}¹ School of Civil Engineering, Hefei University of Technology, Hefei 230009, China² State Key Laboratory of Explosion Science and Technology, Beijing Institute of Technology, Beijing 100081, China

* Correspondence: wangyixian2012@hfut.edu.cn

Abstract: The risk evaluation indexes of goaf are multi-source and have complex mutual internal correlations, and there are great differences in the risk identification of goaf from different mines among the various influencing factors. This paper mainly focuses on principal component analysis (PCA) and the differential evolution algorithm (DE), while a multi-classification support vector machine (SVM) is adopted to classify the risks of goaf. Then, the *K*-fold cross-validation method is used to prevent the overfitting of selection in the model. After the analysis, nine factors affecting the risk identification of goaf in a certain area of East China were determined as the primary influencing factors, and 120 measured goafs were taken as examples for classifying the risks. More specifically, the classification results show that: (1) SVM has the useful ability of generalization, especially when solving the problems of overfitting, and it is easy to fall into the local minima under the conditions of small samples; (2) PCA is employed to realize the intelligent dimensionality reduction and denoising of multi-source impact indicators for goaf risk identification, which immensely improves the prediction accuracy and classification efficiency of the model; (3) after using the DE, the optimal solutions of the problems to be optimized are automatically obtained through the global optimization search mechanism, namely, the kernel function parameter, ' γ ', and the penalty factor, '*C*', of the SVM, which further verifies that the characteristics of clear logic, strong convergence, and good robustness can be found in the DE. As demonstrated, this method has the advantages of guiding significance and application value for goaf risk identification.

Keywords: goaf; risk assessment; support vector machine (SVM); principal component analysis (PCA); differential evolution algorithm (DE)



Citation: Yuan, H.; Cao, Z.; Xiong, L.; Li, H.; Wang, Y. A Machine Learning Method for Engineering Risk Identification of Goaf. *Water* **2022**, *14*, 4075. <https://doi.org/10.3390/w14244075>

Academic Editor: Dongmei Han

Received: 16 October 2022

Accepted: 8 December 2022

Published: 13 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent decades, with the continuous exploitation of mines, the number and volume of goaf formations have been constantly rising, which has brought a tremendous increase in the potential risks regarding the safety of mines [1–5]. Therefore, research on the risk identification of goaf is of great significance to ensure the safe development of mines [6–9]. In general, there are many factors affecting the stability of goaf, such as the engineering and hydrogeological conditions, the exploitation depth, the ore block constituent elements, the goaf mining height, the pillar situation, the goaf formation time, the measured volume, the impact of blasting on the ore body, the goaf treatment rate, the distribution and scale of structural plane, the goaf treatment mode, the development of geological structure, the strength of the rock surrounding the goaf, the goaf's shape, the maximum exposure area, the maximum exposure height, the thickness of the roof and bottom protection, the mining method, etc., and each influencing factor will exert huge effects on the stability of mine goaf among the various areas [10–12]. Thus, the evaluation indexes of goaf in different mines should be explicitly analyzed according to the specific circumstances. In this paper, the risk types that are assessed mainly include the roof caving in, rockslides, falls due to collapses

in the goaf, and other risks. Therefore, nine parameters that are closely related to these risk categories, such as the exploitation depth, the mining method, the goaf mining height, the maximum exposure area, the maximum exposure height, the maximum exposure span, the pillar locations, the measured volume, and the treatment rate are selected as the central influencing factors. However, when considering the numerous factors and when there is noise in the data, this will cause enormous inconvenience in terms of the analysis. Hence, the dimensionality reduction method can be used to process the data; as yet, the most widely used method in dimensionality reduction has been principal component analysis (PCA) [13]. Briefly, PCA is adopted to preprocess the input data, and the main information contained in the data is still retained in the principal component. Above all, not only can it reduce the dimension of the data, but also play a significant role in denoising, so as to make the prediction results more accurate.

In general, the analysis and research methods in the field of mine goafs regarding risk assessment have been applied by several scholars, the techniques used mainly include the fuzzy synthetic assessment method [14], the set pair analysis theory [15], the grey relation analysis method [16], the uncertainty measurement theory [17], etc. Nonetheless, all these non-machine learning algorithms have their own respective application scope; the fairly strong abilities of promotion and improvement are yet to be enhanced, and some of them need to further improve their generalization performance. Recently, data-driven machine learning models have been broadly utilized to work out the nonlinear problems, and some scholars have applied the techniques to the study of goaf risk. Hu et al. [18] proposed a Bayesian discrimination method (BDM) for the risk identification of goaf, and established the corresponding BDM for further research; Feng et al. [19] combined PCA and a neural network to evaluate the risk of goaf, so as to reduce the input variables, eliminate the correlation among variables, and improve the prediction accuracy; Wang et al. [20] constructed multi-classification support vector machine (SVM) models for goaf stability classification, according to the SVM theory and the 'one-against-one' method; Wang et al. [21] used the directed acyclic graph method to constitute the multi-classification SVM and acquired SVM models for goaf stability classification.

Currently, SVMs have been extensively utilized under nonlinear conditions and have achieved numerous satisfactory results. They apply the structural risk minimization principle to replace certain empirical risk minimization principles in traditional machine learning methods, which provides the useful ability of generalization. Compared with the neural network, there are evident advantages to the SVM in terms of solving the problems of overfitting and it is easy to fall into the local minima in the case of small samples, which is considered to be a superior theory for predictive learning [22]. Moreover, the SVM is a typical binary classification problem, and the multi-classification cases can be constructed using the 'one-against-all' method [23], the 'one-against-one' method [24], the directed acyclic graph method [25], the binary tree method [26], etc., while the cross-validation method can be used to choose the appropriate model. In this paper, the parameters of SVMs are optimized, based on the DE, which is an efficient global optimization algorithm on the basis of population and can realize a global search through competitive selection and differential mutation. Additionally, the technique of differential mutation can also avoid the problem of falling into local optimization due to a lack of mutation in the genetic algorithm (GA). On the whole, the DE is comprehensively applied in optimization problems owing to its clear structure, strong convergence, and good robustness.

In summary, the PCA adopted in this paper aims to preprocess the data in order to reduce the input variables and eliminate the correlation among them, so that the data-processing speed can be accelerated, and the prediction accuracy can be improved simultaneously. Subsequently, a multi-classification SVM is employed to train and predict the data, and the cross-validation principle is used to select the most preferable model. Furthermore, the DE is employed to optimize the parameters of the SVM. In the meantime, we have decided to take 120 measured goafs from a mine in a certain area of East China as an

example to be classified and compare it with the phenomena after using the PCA of the Naive Bayes classification and the BP neural network.

Consequently, conclusions can be drawn that the classification results in this paper ideally reflect the actual classification of goat stability, which has obvious beneficial guiding significance and application value for engineering.

2. Principal Component Analysis

2.1. Basic Principles

Principal component analysis (PCA) is a dimensionality reduction method that transforms the data represented by multiple related variables into a few unrelated ones through orthogonal mapping. After transformation, the variables are defined as ‘principal components’. Among some multi-variable problems, the variables may be related to each other. Normally, a complex direct analysis may demonstrate inaccurate results. Therefore, the PCA method is adopted to replace the original multiple variables with a few principal components, while retaining most of the information involved in the data, so that it is convenient for further analysis.

2.2. Mathematical Model

Assuming that the m variables, cover x_1, x_2, \dots, x_m , are included in the n samples, the original data matrix can be obtained, as follows:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}. \tag{1}$$

Generally, there are different dimensions in the variables among the research questions, which may create some new problems. Thus, the original data should be standardized with Equation (2) before the PCA takes place:

$$x_{ij}^* = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{2}$$

where x_{ij} represents the data before standardization; x_{ij}^* represents the data after standardization; $\max(x_j)$ and $\min(x_j)$ are the maximum and minimum values in the column j data, respectively.

After data standardization, the original data will become the values between 0 and 1 in subsequent calculations and analysis.

The matrix after data standardization is represented by y ; namely, the m variables of $x = (x_1, x_2, \dots, x_m)^T$ are denoted as the m new ones, and the new variables can be linearly expressed by the original ones as x_1, x_2, \dots, x_m , that is:

$$\begin{cases} y_{1m} = u_{11}x_1 + u_{12}x_2 + \cdots + u_{1m}x_m \\ y_{2m} = u_{21}x_1 + u_{22}x_2 + \cdots + u_{2m}x_m \\ \vdots \\ y_{nm} = u_{n1}x_1 + u_{n2}x_2 + \cdots + u_{nm}x_m \end{cases} \tag{3}$$

where $y_{1m}, y_{2m}, \dots, y_{nm}$ signifies the variables sourced through PCA, and u is the correlation coefficient matrix among the variables, which need to satisfy the following conditions:

- (1) $u_{k1}^2 + u_{k2}^2 + \cdots + u_{km}^2 = 1 (k = 1, 2, \dots, n)$;
- (2) $\text{cov}(y_i, y_j) = 0 (i \neq j; i, j = 1, 2, \dots, m)$, namely, the components of principal analysis are independent and there is no overlapping information;
- (3) $\text{var}(y_1) \geq \text{var}(y_2) \geq \dots \geq \text{var}(y_m)$, namely, the principal components are sorted according to the standard deviation, where: y_1, y_2, \dots, y_m , obtained through the

above process, can be determined as the principal components of $1, 2, \dots, m$ of the original variables.

In general, the cumulative variance contribution rate represents the amount of original data information. However, one of the first of several principal component factors determines the number of principal components. In order to reduce the amount of calculation needed, the cumulative variance contribution rate can be about 80%.

2.3. Geometric Interpretation

In short, PCA is employed to acquire the linearly independent variables that are defined as the principal components from the correlated ones through orthogonal transformation. Specifically, by means of rotating and transforming the original coordinate system, the data is represented in the new system. As depicted in Figure 1, it shows the new coordinate system and the corresponding principal components obtained by rotating the data expressed by the two variables.

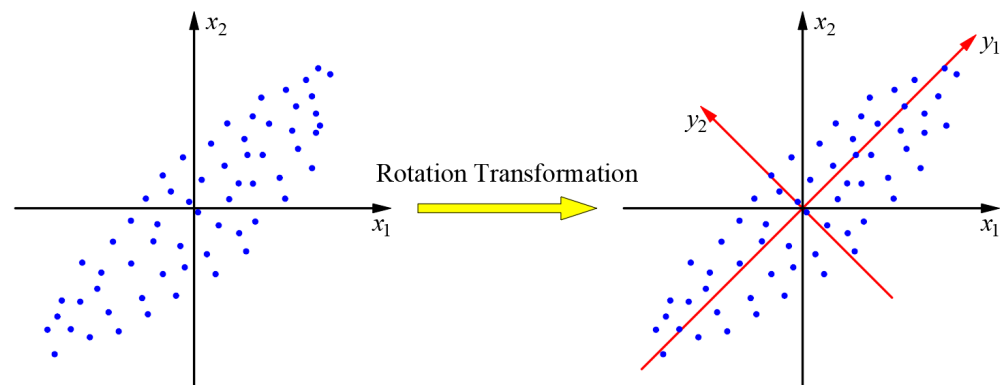


Figure 1. Geometric interpretation of PCA.

3. Multi-Classification Support Vector Machine (SVM)

3.1. Basic Principles of SVM

Notably, the learning strategy of the support vector machine (SVM) is to minimize the structural risk, based on a sound ability in terms of generalization. According to the machine learning theory [27], the basic model of the SVM is a binary classifier for linearly separable data sets, which can be applied to nonlinear problems by introducing the 'kernel' function. When linearly separating the data sets, the SVM is normally adopted to solve the optimal hyperplane that linearly separates the data sets according to interval maximization or the corresponding convex quadratic programming problems that need to be solved promptly.

Hence, the general optimal hyperplane can be determined:

$$wx + b = 0. \quad (4)$$

The corresponding classification decision function is:

$$f(x) = \text{sign}(wx + b). \quad (5)$$

To conclude, the basic model is denoted as the 'linear separable support vector machine', and interval maximization refers to 'hard interval maximization'.

Figure 2 demonstrates the classification problem within the two-dimensional feature space. In particular, the distance between the lines ' H_1 ' and ' H_2 ' is signified as 'interval', the size of which is related to the normal vector ' w ' of line ' H_0 ', and where the value is equal to $2/||w||$.

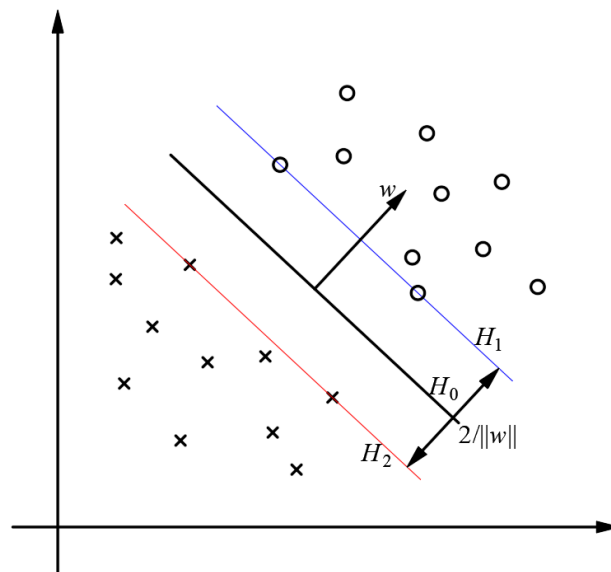


Figure 2. Schematic diagram of the linear separable support vector machine.

Considering the realistic conditions, the data sets cannot be completely linearly separated due to the existence of some special points among the data. As a consequence, the linear support vector machine can be structured by introducing the relaxation variables of ‘ $\xi_i \geq 0$ ’. Thus, the interval maximization at this moment can be signified as ‘soft interval maximization’, the corresponding convex quadratic programming problem of which is:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(wx_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N \\ & \xi_i \geq 0, \quad i = 1, 2, \dots, N \end{aligned} \tag{6}$$

where C refers to the ‘penalty factor’. The larger the value of C , the more prone it is to overfitting, and vice versa.

Regarding Equation (6) as the initial circumstance and using the ‘Lagrange duality’ to obtain the dual problem, the optimal solution of the original problem can be achieved by solving the dual problem. Indeed, the ‘Lagrange’ function of the original problem is determined as follows:

$$\begin{aligned} L(w, b, \xi, \alpha, \mu) \equiv \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i - \\ & \sum_{i=1}^N \alpha_i (y_i(wx_i + b) - 1 + \xi_i) - \sum_{i=1}^N \mu_i \xi_i \end{aligned} \tag{7}$$

Among them, α_i refers to the ‘Lagrange multiplier’; $\alpha_i \geq 0$ ($i = 1, 2, \dots, N$); $\mu_i \geq 0$. The dual problem in the original problem is:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i x_j) - \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N \end{aligned} \tag{8}$$

Assuming $\alpha^* = (\alpha^* 1, \alpha^* 2, \dots, \alpha^* N)^T$ is the solution of the dual problem, if there is a

component $\alpha^* j$ of α^* that satisfies $0 < \alpha^* j < C$, then the optimal solution of the original problem can be determined as:

$$\begin{cases} w^* = \sum_{i=1}^N \alpha_i^* y_i x_i \\ b^* = y_j - \sum_{i=1}^N y_i \alpha_i^* (x_i x_j) \end{cases} \quad (9)$$

After that, the classification decision function can be obtained as:

$$f(x) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i (x_i x) + b^*\right). \quad (10)$$

On account of the nonlinear real data, it is feasible to take advantage of the kernel technique to set up the nonlinear support vector machine, the classification decision function of which is:

$$f(x) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i K(x, x_i) + b^*\right) \quad (11)$$

where $K(x, z)$ represents the function of the positive definite kernel.

According to the SVM theory, there are four types of kernel function that are commonly used, namely, the ‘linear’ kernel function, the ‘polynomial’ kernel function, the ‘Gaussian radial basis’ function (the ‘RBF’ kernel function), and the ‘Sigmoid’ kernel function, respectively [28]. In particular, the kernel function adopted in this paper is the radial basis function:

$$K(x, z) = \exp(-\gamma \|x - z\|^2). \quad (12)$$

In addition, the corresponding classification decision function can be defined as:

$$f(x) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i \exp(-\gamma \|x - z\|^2) + b^*\right). \quad (13)$$

3.2. Constructing a Multi-Classification Support Vector Machine

As a matter of fact, the risk grade evaluation of goaf is a multi-classification problem rather than a simple binary classification problem. Generally, the most common construction methods of a multi-classification support vector machine include the ‘one-against-one’ method, the ‘one-against-all’ method, the directed acyclic graph method, the binary tree method, etc. In this paper, the ‘one-against-one’ method is implemented to construct the multiple classifiers. In this way, the basic principle of the ‘one-against-one’ method is to establish a binary classification support vector machine between any two types when there are n different classifications in the training set. Therefore, $n(n - 1)/2$ binary classifiers will be accurately obtained. In addition, the test set is input into each classifier, and the category with the most votes will be selected as the final output category of the classifier. For instance, when the number of classifiers equals 4, the 6 binary classifiers will be established; the classification structure diagram is depicted in Figure 3.

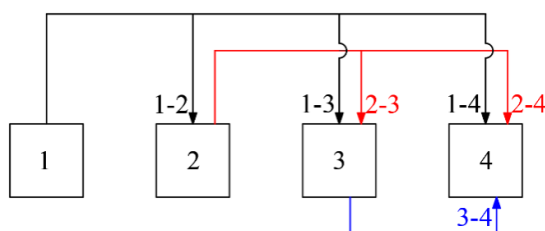


Figure 3. Schematic diagram of the classification structure.

3.3. Cross-Validation

In order to obtain a superior model under the conditions of small samples, a cross-validation technique can be implemented to make full use of the data for training and testing. As a rule, the most broadly used cross-validation method, namely, K -fold cross-validation, is adopted in this paper. To account for this, the specific principles can be summed up according to the following aspects: in the first place, it is viable to divide the data set into K subsets. Subsequently, each subset has the same size and does not intersect with any other subset. Later, one subset is taken as the test set and the remaining $K - 1$ subsets are chosen as the training sets. In the end, after repeating the training for K times, the model with the smallest error among the K tests will be selected as the optimal choice.

3.4. Parameter Optimization of the Differential Evolution Algorithm

In fact, the differential evolution algorithm (DE) is an efficient global optimization algorithm, based on the population in question, which can realize a global search via competitive selection and differential mutations in the population. First and foremost, the DE is expected to be encoded and then randomly initializes the population of $M = (M_1, M_2, \dots, M_N)$, where n denotes the size of the population. Note that the upper and lower bounds of the parameters should be set before initialization, then afterward the intermediate population can be accessibly obtained by mutating and crossing the parameters. Moreover, a greedy strategy is adopted to select the method of one-against-one between the two populations to obtain the new generation.

In this paper, the DE/rand/1/bin [29] is chosen as the form of DE, wherein the coding method adopts the real coding. One typical advantage of the real coding method that should be noted is that it does not need frequent coding and decoding, which can improve the accuracy and convergence speed when solving problems, and can effectively avoid some additional problems, such as 'Hamming cliffs', etc. Thus, the variation, crossover, and selection operations of this form can be concluded as follows.

Variation: An individual from the population will be randomly selected as the basis vector, and then it is necessary to take both other individuals as the difference vectors. After that, the variation operation can be performed:

$$n_i^{G+1} = m_{a_1}^G + F(m_{a_2}^G - m_{a_3}^G) \quad (14)$$

where $a_1, a_2, a_3 \in \{1, 2, \dots, N\}$, and $a_1 \neq a_2 \neq a_3$; N is the size of the population; F refers to the scaling factor, the value of which is a positive real number and also generally a random one between (0, 1), which can control the evolution rate of the population; G represents the current population, and $G + 1$ denotes the next generation.

Crossover: in order to increase the population diversity, the DE employs the binomial distribution crossover method to generate new individuals using predetermined parent individuals and mutated ones:

$$l_{i,j}^{G+1} = \begin{cases} n_{i,j}^{G+1}, & \text{rand} \leq C \text{ or } j = j_{rand} \\ m_{i,j}^G, & \text{otherwise} \end{cases} \quad (15)$$

where, in this paper, $j \in \{1, 2\}$; $rand$ is a random number between [0, 1]; j_{rand} signifies an integer that is randomly generated in {1, 2} to ensure that at least one optimization parameter will mutate; C refers to the crossover factor, which is normally a random number between [0, 1].

Selection: A pair of survivors are competing for selection, and those with high fitness will be chosen for the next generation. In this case, the principle is that each individual in the population can only compare their fitness with those in the same position in another population. Hence, this can be expressed as:

$$m_i^{G+1} = \begin{cases} l_i^{G+1}, & f(l_i^{G+1}) \leq f(m_i^G) \\ m_{i,j}^G, & f(l_i^{G+1}) > f(m_i^G) \end{cases} \quad (16)$$

where $f(x)$ signifies the fitness function.

To sum up, through the operations of mutation, crossover, and selection, the population will evolve to form the next generation. In this way, it can reach the optimal after cycling and the optimal solution of the problem can be tackled. When the kernel function of SVM selects the radial basis function, the kernel parameter, ' γ ', and the penalty factor, ' C ', in the SVM model need to be determined. Therefore, the average score of cross-validation is taken as the goal to be optimized, and the parameters, such as ' C ' and ' γ ', are selected as the decision variables. Furthermore, the DE is adopted to find the optimal solution to the problem, and the optimal parameters can be obtained automatically, according to the optimization algorithm. Ultimately, the comprehensive workflow chart is displayed in Figure 4.

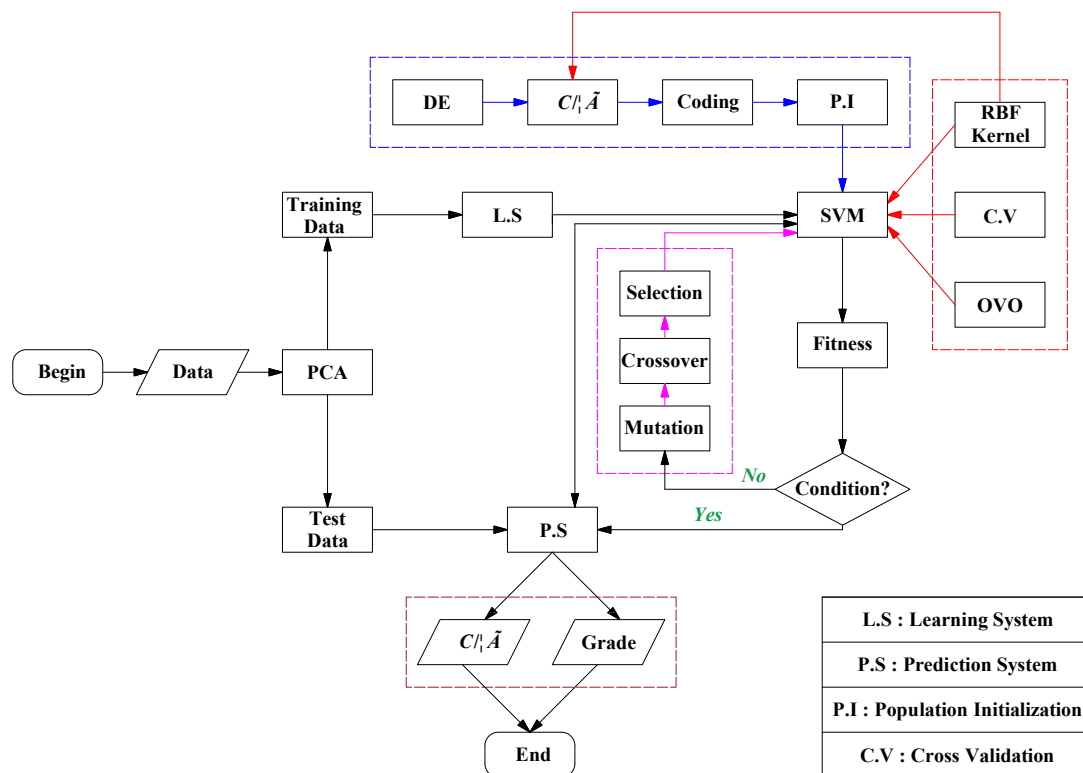


Figure 4. Integrated workflow chart.

4. Engineering Examples

After more than 20 years of mining, a certain amount of goaf from an iron mine in East China has been formed since the site was put into production. In this paper, 120 samples of sub-goaf are chosen as the data set and 9 parameters, such as the mining depth, the mining method, the goaf mining height, the maximum exposure area, the maximum exposure height, the maximum exposure span, the pillar situation, the measured volume, and the treatment rate are selected as the central influencing factors; sample data for goaf are listed in Table 1.

Table 1. Influencing factors of goaf stability and partial samples of the risk rank.

Sample Serial Number	Exploitation Depth X1/m	Mining Methods X2	Goaf Mining Height X3/m	Maximum Exposure Area X4/m ²	Maximum Exposure Height X5/m	Maximum Exposed Span X6/m	Pillar Situation X7	Measured Volume X8/m ³	Governance Rate X9	Risk Rank
1	130	1	35	3589	35	39	0	57,481.1	0.0	2
2	130	1	20	1208	0.99	24	1	12,141.3	94.4	1

Table 1. *Cont.*

Sample Serial Number	Exploitation Depth X1/m	Mining Methods X2	Goaf Mining Height X3/m	Maximum Exposure Area X4/m ²	Maximum Exposure Height X5/m	Maximum Exposed Span X6/m	Pillar Situation X7	Measured Volume X8/m ³	Governance Rate X9	Risk Rank
3	130	1	35	1735	5.97	28	0	31,595.7	96.3	1
4	130	1	35	1644	35	32	2	17,144.4	100.0	1
5	130	1	25	2489.5	25	39	2	19,377.7	100.0	1
119	220	1	15	349	15	17	0	3200	0.0	1
120	220	1	15	259	15	10	0	2867	0.0	1

When using the SVM to analyze data, the non-data factors can be transformed into data factors with the purpose of facilitating learning. Therefore, this paper deals with two factors: the mining method and the pillar situation. Based on this focus, the processing methods can be generalized into two aspects: (1) the mining method—the shallow mining method is recorded as +1, while the medium/deep-hole method is recorded as −1; (2) the pillar situation—a no-pillar situation is recorded as 0, while the boundary, intermediate, and mixed pillars are recorded as 1, 2, and 2, respectively.

According to the analysis and determinations of the field professionals, the risk level can be divided into three ranks, namely, 1, 2, and 3. In other words, the higher the rank, the more serious the risk that will occur.

As an illustration, the 9 factors affecting the stability of goaf are taken as the input factors, and the risk level is chosen to be the output factor. Afterward, the data in Table 1 were analyzed using the SPSS software. Meanwhile, the correlation coefficient adopts the Pearson correlation coefficient, while the two-tailed t-test is applied to the significance test.

The Pearson correlation coefficient matrix of each factor and the heat map of the correlation matrix are shown in Table 2 and Figure 5, respectively. It can be seen from Table 2 that several factors among the input factors have a strong correlation with each other. As a result, it is necessary to conduct the PCA using the input data.

Table 2. Pearson correlation coefficient matrix of each factor.

Index	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1.000								
X2	−0.366	1.000							
X3	−0.389	0.273	1.000						
X4	0.001	−0.432	−0.084	1.000					
X5	−0.325	−0.045	0.512	0.089	1.000				
X6	−0.050	−0.465	0.097	0.695	0.227	1.000			
X7	−0.342	0.163	0.296	0.086	0.236	0.103	1.000		
X8	−0.046	−0.150	0.098	0.594	0.019	0.370	0.110	1.000	
X9	0.104	0.010	0.153	−0.039	−0.309	−0.093	−0.005	0.061	1.000

It is evident that there are different dimensions in the variables of the research issues, which may lead to some new problems. Therefore, the original data should be standardized using Equation (2) before the PCA takes place.

After using the SPSS software to conduct a PCA on the standardized data, the principal component gravel diagram (Figure 6) and the principal component list (Table 3) can be obtained promptly. According to Figure 6 and Table 3, the eigenvalues of the first five factors differ vastly from each other, and the cumulative contribution rate of the total variance is equal to 83.643%, which fulfills the requirement that the variance of the principal components accounts for 75–85% of the overall variance [30]; namely, the most information on the overall variance can be precisely summarized by the first five factors. Hence, it is practicable to select the first five components as the principal components to replace the original variables for analysis.

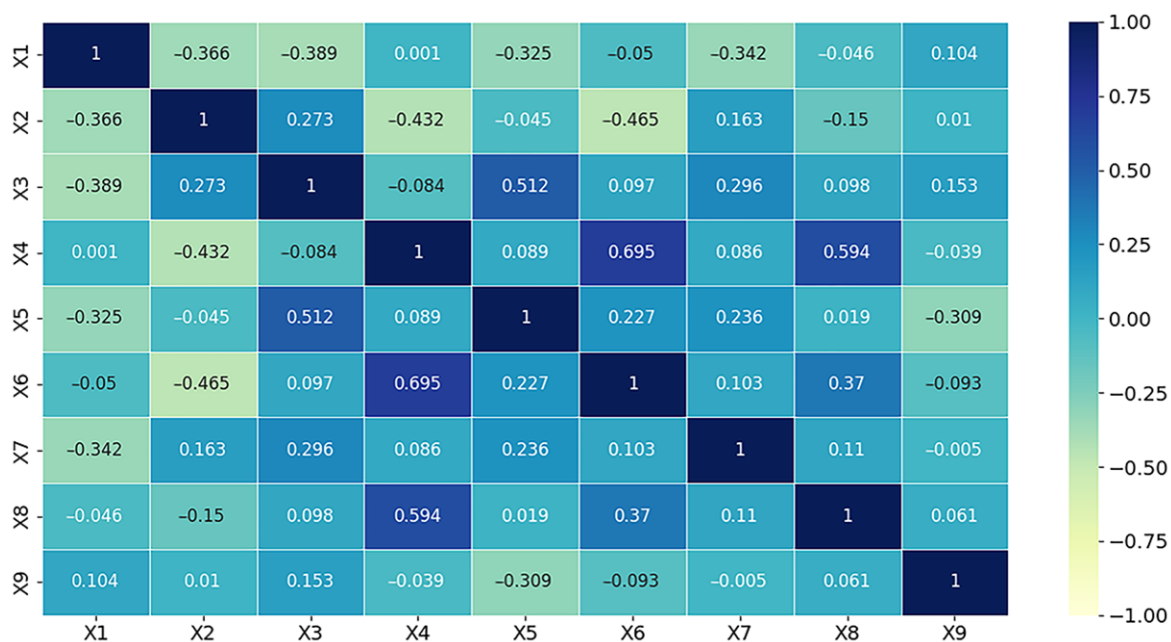


Figure 5. Correlation matrix heat map.

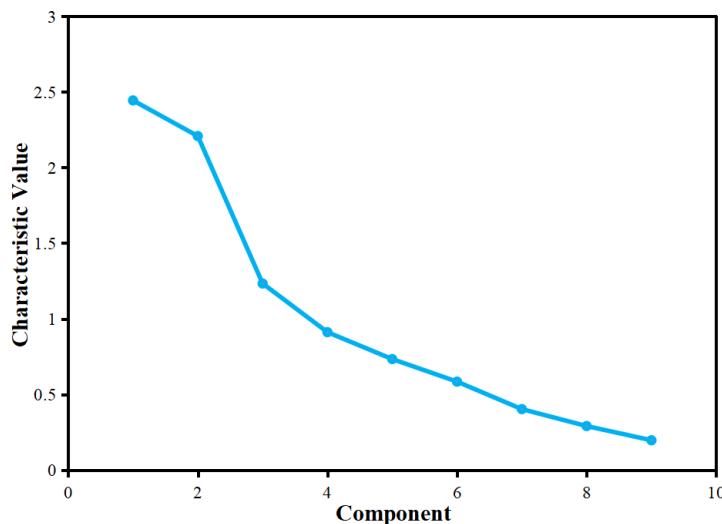


Figure 6. Gravel diagram of PCA.

Table 3. List of principal components.

Component	Initial Characteristic Value			Sum of Squares of Extracted Loads		
	Total	Percentage Variance	Accumulation/%	Total	Percentage Variance	Accumulation/%
1	2.444	27.150	27.150	2.444	27.150	27.150
2	2.208	24.528	51.678	2.208	24.528	51.678
3	1.233	13.698	65.377	1.233	13.698	65.377
4	0.911	10.127	75.504	0.911	10.127	75.504
5	0.733	8.139	83.643	0.733	8.139	83.643
6	0.584	6.493	90.136			
7	0.402	4.470	94.607			
8	0.290	3.217	97.824			
9	0.196	2.176	100.000			

As displayed in Table 4, the load matrix of the main component factors can be adopted to ascertain the relationship between the main component factors, Y1, Y2, Y3, Y4, and Y5,

and the original variables. Hence, the expressions of the corresponding factors can be listed as follows:

$$\begin{aligned}
 Y1 &= -0.059X1 - 0.565X2 + 0.100X3 + 0.884X4 + 0.321X5 + 0.858X6 + 0.188X7 + 0.664X8 - 0.118X9; \\
 Y2 &= -0.751X1 + 0.519X2 + 0.766X3 - 0.158X4 + 0.638X5 - 0.019X6 + 0.583X7 - 0.001X8 - 0.123X9; \\
 Y3 &= -0.031X1 + 0.219X2 + 0.211X3 + 0.110X4 - 0.454X5 - 0.065X6 + 0.191X7 + 0.405X8 + 0.846X9; \\
 Y4 &= 0.296X1 - 0.363X2 + 0.451X3 - 0.173X4 + 0.367X5 + 0.103X6 - 0.246X7 - 0.288X8 + 0.413X9; \\
 Y5 &= 0.117X1 - 0.251X2 - 0.182X3 - 0.043X4 - 0.067X5 + 0.060X6 + 0.697X7 - 0.343X8 + 0.096X9.
 \end{aligned}$$

Table 4. Load matrix of the principal component factors.

Index	Principal Component				
	Y1	Y2	Y3	Y4	Y5
X1	-0.059	-0.751	-0.031	0.296	0.117
X2	-0.565	0.519	0.219	-0.363	-0.251
X3	0.100	0.766	0.211	0.451	-0.182
X4	0.884	-0.158	0.110	-0.173	-0.043
X5	0.321	0.638	-0.454	0.367	-0.067
X6	0.858	-0.019	-0.065	0.103	0.060
X7	0.188	0.583	0.191	-0.246	0.697
X8	0.664	-0.001	0.405	-0.288	-0.343
X9	-0.118	-0.123	0.846	0.413	0.096

According to the calculated factor expressions, it is necessary to conduct the PCA and calculation on the standardized data at a later time. This can be seen most obviously in Table 5 for the partial calculated data.

Table 5. Partial data after principal component calculation.

Sample Serial Number	Y1	Y2	Y3	Y4	Y5	Risk Rank
1	0.9692	1.8113	0.1422	0.3396	-0.6124	2
2	-0.1286	1.0236	1.2293	0.1273	0.0872	1
3	0.0849	1.2046	1.2507	0.5023	-0.4067	1
4	0.5255	2.3128	1.0500	0.6140	0.2771	1
5	0.6189	1.8470	1.1160	0.3447	0.3573	1
			...			
119	-0.2744	0.6973	0.0772	0.0678	-0.2842	1
120	-0.4014	0.7017	0.0838	0.0565	-0.2910	1

In view of the standardized data after PCA, the multi-classification SVM test will be carried out next. To begin with, the ‘one-against-one’ method was selected to construct the multi-classification classifiers. After adopting the cross-validation method of the K-fold, K = 5 is taken. Simultaneously, the principal components ‘Y1’, ‘Y2’, ‘Y3’, ‘Y4’, and ‘Y5’ are used as the input factors and the risk level is chosen as the output factors. Eventually, all 120 samples can be divided into the training sets and the test sets, in which the number of the former and the latter are equal to 80 and 40, respectively.

After optimizing the parameters of SVM by means of the DE, the real-number coding method has been employed. In this case, the number of individuals in the population and the maximum evolution algebra can be taken to be a value of 20 and 30, respectively. In the end, the final calculation result is compared with the classification results of SVM without PCA, and simultaneously, the results of the naive Bayes and the BP neural network after PCA are compared with each other.

As a matter of fact, the data prediction accuracy of BP neural network classification (Method 1) and Naive Bayes classification after PCA can both achieve 87.5% while the accuracy of the classification methods (Method 3) without PCA can reach 90%. Conversely, the data prediction accuracy of the method used in this paper (Method 4) can attain 92.5%.

More specifically, the relevant parameters after optimization by DE are $C = 145.50$ and $\gamma = 0.26$. As exhibited, the iterative effect of DE is depicted in Figure 7, and the confusion matrices for the test sets among each method are shown in Figure 8. In addition, the comparison of prediction accuracy is displayed in Figure 9.

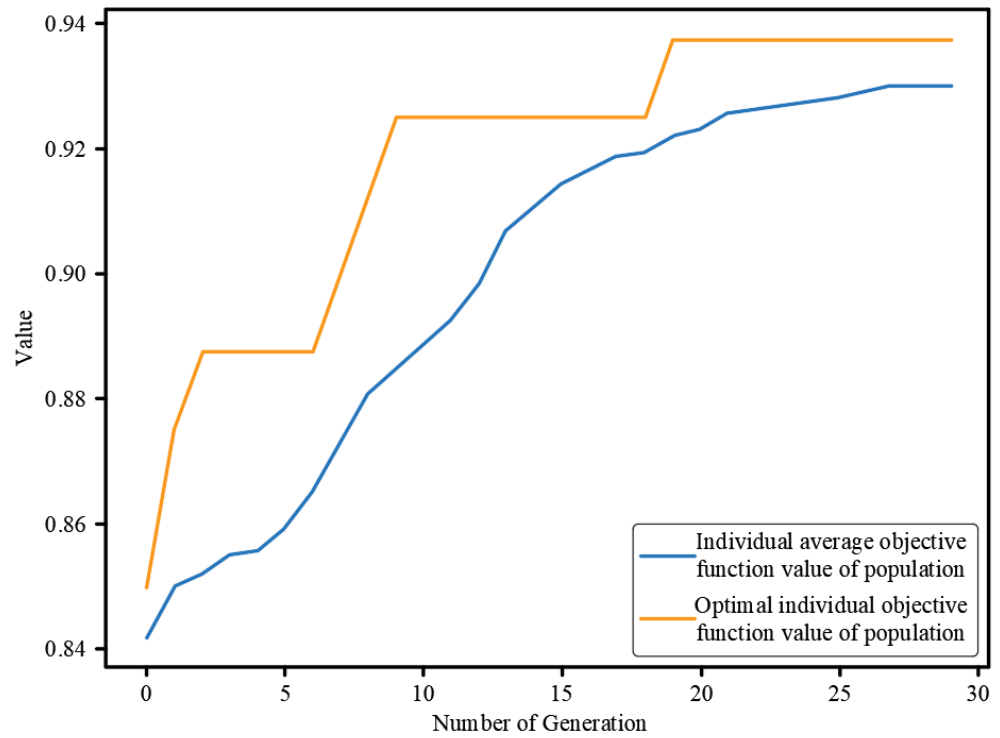


Figure 7. Iterative effect of the DE.

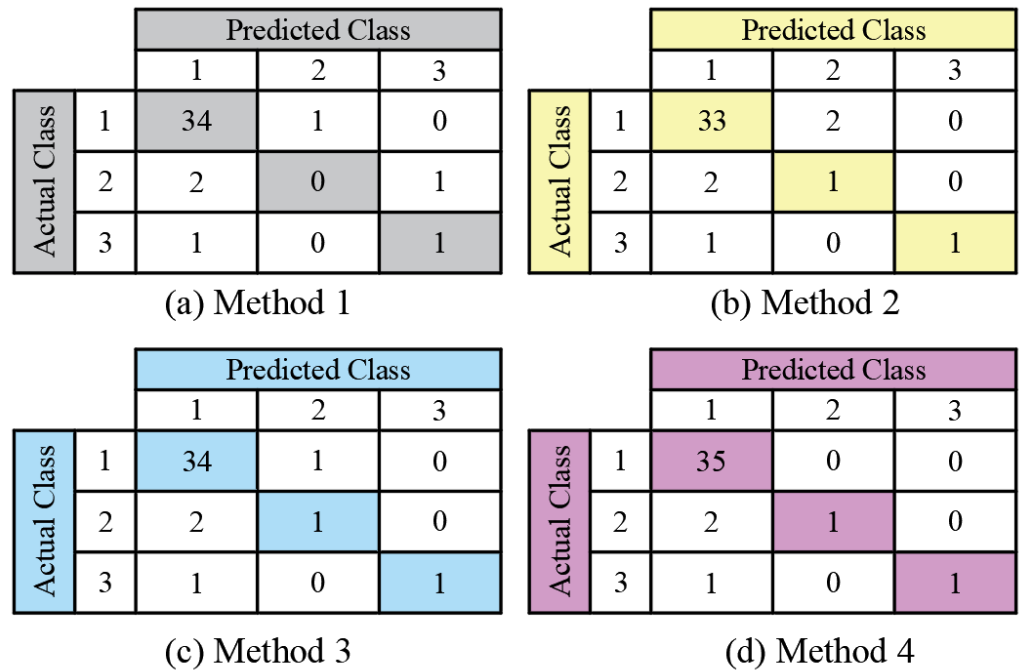


Figure 8. Confusion matrices for the test sets.

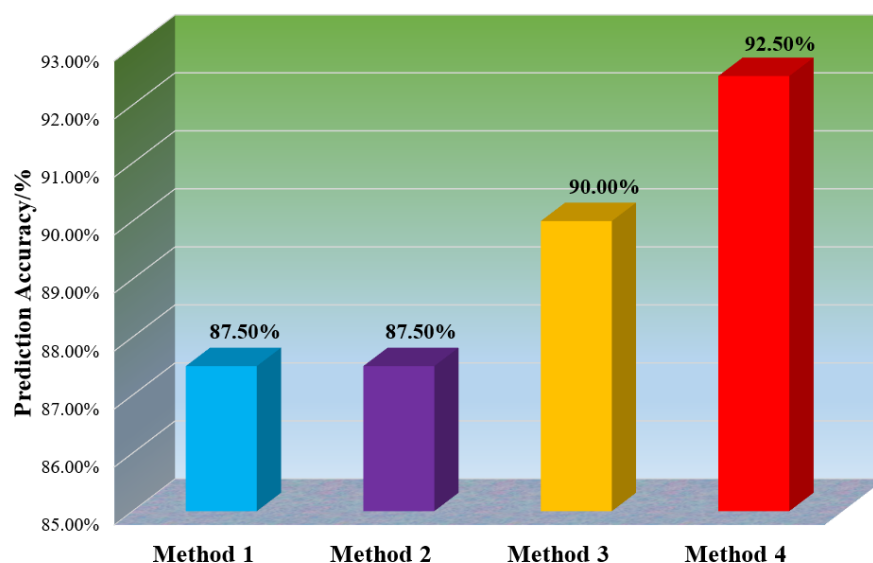


Figure 9. Comparison of the prediction accuracy.

As demonstrated, in the meantime, the PCA method can not only retain the most information involved in the original data but also play a vital role in denoising. Owing to the effects of the dimension reduction, the running time can be shortened, and the training efficiency can be enhanced. When there is noise in the large data sets, the PCA will significantly improve the tests' efficiency and accuracy. Although the neural network can deal with nonlinear issues due to the insufficient samples in practical application, the visible problems that will proceed, such as the local minimum, overfitting, etc., have reflected the limitation of weak robustness and low recognition accuracy.

5. Conclusions

In this paper, the support vector machine (SVM) based on the principal component analysis (PCA) and the differential evolution algorithm (DE) is adopted to identify the risk level of goaf, and the primary findings can be drawn as follows:

- (1) The 'one-against-one' method is used to construct a multi-classification SVM. In order to prevent the overfitting of the model, the K -fold cross-validation method will be employed to select it. Above all, the research results reveal that the SVM has the desirable ability of generalization. Compared with the neural network, the apparent advantages lie in solving the problems of overfitting and it is easy to fall into the local minimum that can be detected in the SVM under the conditions of small samples.
- (2) PCA is used to preprocess the original data of multi-source impact indicators for goaf risk identification, which can realize the dimensionality reduction and data denoising, and can simultaneously improve the prediction accuracy and classification efficiency while retaining the most information.
- (3) Using the strategy of DE and a global optimization search mechanism, the optimal solution of the problems to be optimized will be automatically obtained, namely, the kernel function parameter of SVM, ' γ ', and the penalty factor, ' C '. Moreover, the engineering calculation example further verifies that the DE has the characteristics of clear logic, strong convergence, and good robustness.

In simpler terms, the method embraced in this study is then discussed. Compared with some machine learning algorithms, the support vector machine (SVM) can be converted into high-dimensional feature space through nonlinear transformation, which can subtly avoid the curse of the dimensionality problem. In addition, since the algorithm can be transformed into a convex quadratic programming problem, it has an obvious ascendancy over the neural network in terms of eliminating the local extremum issues and small sample data sets. Therefore, the work efficiency of safety production in mines will be

enormously improved by means of introducing this method into the domain of the risk identification of mine goafs, which has significant engineering instructive significance and widespread application values. Nevertheless, except for the aforementioned advantages, some limitations can be examined in SVM. For example, in contrast to other machine learning algorithms, when a huge level of capacity can be found in the sample data sets, SVM will perform inefficiently, and it is sensitive to missing the data among the data sets. Furthermore, standard algorithms cannot reasonably reflect the probability characteristics. Consequently, for further research routes, a number of rational improvements can be prompted for the SVM, for instance, introducing the probability characteristics can refine it into a probabilistic SVM, while replacing a single kernel function with a family of ones is enhanced as the multiple kernel SVM, etc.

Author Contributions: Conceptualization, H.Y. and Y.W.; methodology and software, Z.C.; validation, L.X.; investigation, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by the National Natural Science Foundation of China (51874112;42077249, 51774107), the University Synergy Innovation Program of Anhui Province (Grant. No.GXXT-2020-055), the Opening Project of the State Key Laboratory of Explosion Science and Technology, and the Beijing Institute of Technology (KFJJ21-03Z).

Data Availability Statement: The data supporting the findings of this study are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yi, H.; Zhang, X.; Yang, H.; Li, M.; Gao, Q.; Jinke. Goaf collapse vibration analysis and disposal based on a experiment of heavy ball touchdown. *Explos. Shock Waves* **2019**, *39*, 91–103.
2. Zhao, Y.; Tang, J.; Chen, Y.; Zhang, L.; Wang, W.; Wan, W.; Liao, J. Hydromechanical coupling tests for mechanical and permeability characteristics of fractured limestone in complete stress-strain process. *Environ. Earth Sci.* **2017**, *76*, 24. [[CrossRef](#)]
3. Zhao, Y.; Luo, S.; Wang, Y.; Wang, W.; Zhang, L.; Wan, W. Numerical Analysis of Karst Water Inrush and a Criterion for Establishing the Width of Water-resistant Rock Pillars. *Mine Water Environ.* **2017**, *36*, 508–519. [[CrossRef](#)]
4. Liao, Y.; Yu, G.; Liao, Y.; Jiang, L.; Liu, X. Environmental Conflict Risk Assessment Based on AHP-FCE: A Case of Jiuhua Waste Incineration Power Plant Project. *Sustainability* **2018**, *10*, 4095. [[CrossRef](#)]
5. Wu, H.; Jia, Q.; Wang, W.; Zhang, N.; Zhao, Y. Experimental Test on Nonuniform Deformation in the Tilted Strata of a Deep Coal Mine. *Sustainability* **2022**, *13*, 13280. [[CrossRef](#)]
6. Du, K.; Li, X.; Liu, K.; Zhao, X.; Zhou, Z.; Dong, L. Comprehensive evaluation of underground goaf risk and engineering application. *J. Cent. South Univ.* **2011**, *42*, 2802–2811.
7. Yuan, Z.; Zhai, J.; Li, S.; Jiang, Z.; Huang, F. A Unified Solution for Surrounding Rock of Roadway Considering Seepage, Dilatancy, Strain-Softening and Intermediate Principal Stress. *Sustainability* **2022**, *14*, 8099. [[CrossRef](#)]
8. Liu, Y.; Hao, Y.; Lu, Y. Improved Design of Risk Assessment Model for PPP Project under the Development of Marine Architecture. *J. Coast. Res.* **2021**, *9*, 74–80. [[CrossRef](#)]
9. Liu, S.; Nie, Y.; Hu, W.; Ashiru, M.; Li, Z.; Zou, J. The Influence of Mixing Degree between Coarse and Fine Particles on the Strength of Offshore and Coast Foundations. *Sustainability* **2022**, *14*, 9177. [[CrossRef](#)]
10. Chen, W.; Wan, W.; Zhao, Y.; Peng, W. Experimental Study of the Crack Predominance of Rock-Like Material Containing Parallel Double Fissures under Uniaxial Compression. *Sustainability* **2020**, *12*, 5188. [[CrossRef](#)]
11. Zhang, Y.; Chang, X.; Liang, J. Comparison of different algorithms for calculating the shading effects of topography on solar irradiance in a mountainous area. *Environ. Earth Sci.* **2017**, *76*, 295. [[CrossRef](#)]
12. Feng, T.; Chen, H.; Wang, K.; Nie, Y.; Zhang, X.; Mo, H. Assessment of underground soil loss via the tapering grikes on limestone hillslopes. *Agric. Ecosyst. Environ.* **2020**, *5*, 297. [[CrossRef](#)]
13. Chen, J.; Liu, L.; Zhou, Z.; Yong, X. Optimization of mining methods based on combination of principal component analysis and neural networks. *J. Cent. South Univ.* **2010**, *41*, 1967–1972.
14. Wang, X.; Duan, Y.; Peng, X. Fuzzy Synthetic Assessment of the Danger Degree of Mined-out Area Disaster. *Min. Res. Dev.* **2005**, *25*, 83–85.
15. Zhou, J.; Shi, X. Evaluation of the alternatives for mined out area disposal based on the identical degree of set pair analysis. *Met. Mine* **2009**, *396*, 10–13.
16. Wang, X.; Ding, D.; Duan, Y. Applications of the grey relation analysis in the evaluation of the risk degree of the underground mined-out stopes. *J. Saf. Sci. Technol.* **2006**, *2*, 35–39.

17. Gong, F.; Li, X.; Dong, L.; Liu, X. Underground goaf risk evaluation based on uncertainty measurement theory. *Chin. J. Rock Mech. Eng.* **2008**, *27*, 323–330.
18. Hu, Y.; Li, X. Bayes discriminant analysis method to identify risky of complicated goaf in mines and its application. *Trans. Nonferrous Met. Soc. China* **2012**, *22*, 425–431. [[CrossRef](#)]
19. Feng, Y.; Wang, X.; Cheng, A.; Zhang, Q.; Zhao, J. Method optimization of underground goaf risk evaluation. *J. Cent. South Univ.* **2013**, *44*, 2881–2888.
20. Wang, Z.; Guo, J.; Wang, L. Recognition of goaf risk based on support vector machines method. *J. Chongqing Univ.* **2015**, *38*, 85–90.
21. Wang, H.; Li, X.; Dong, L.; Liu, K.; Tong, H. Classification of goaf stability based on support vector machine. *J. Saf. Sci. Technol.* **2014**, *10*, 154–159.
22. Fang, X.; Ding, Z.; Shu, X. Hydrogen yield prediction model of hydrogen production from low rank coal based on support vector machine optimized by genetic algorithm. *J. China Coal Soc.* **2010**, *35*, 205–209.
23. Hsu, C.-W.; Lin, C.-J. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425. [[PubMed](#)]
24. Keel, U. Pairwise Classification and Support Vector Machines. In *Advances in Kernel Methods-Support Vector Learning*; MIT Press: Cambridge, MA, USA, 1999; pp. 255–258.
25. Platt, J.C.; Cristianini, N.; Shawe-Taylor, J. Large Margin DAGs for Multi-class Classification. *Adv. Neural Inf. Process. Syst.* **2000**, *12*, 547–553.
26. Bennett, K.P.; Blue, J.A. A support vector machine approach to decision tree. *Rensselaer Polytech. Inst.* **1997**, *3*, 2396–2401.
27. Zhou, Z. *Machine Learning*; Tsinghua University Press: Beijing, China, 2017; pp. 121–139.
28. Liang, N.; Tuo, Y.; Deng, Y.; Jia, Y. Classification model of ice transport and accumulation in front of channel flat sluice based on PCA-SVM. *Chin. J. Theor. Appl. Mech.* **2021**, *53*, 703–713.
29. Chen, X.; Yang, G.; Huang, M. Real-coded Quantum Differential Evolution Algorithm. *J. Chin. Comput. Syst.* **2013**, *34*, 1141–1146.
30. Xu, Z.; Zhou, D.; Luo, Y. Fuzzy Neural Network Based on Principal Component. *Comput. Eng. Appl.* **2006**, *42*, 34–36.