Research on an Identification Method for Gas Disaster Risk Based on the Selective Ensemble Classification Model

Rong Liang,* Qiaolin Lv, Pengtao Jia, Zhilong Zhao, and Chengyixiong Xu

ABSTRACT: To improve the accuracy of gas disaster risk identification, a selective ensemble classification model is proposed based on clustering selection and a new degree of combination fitness (CS−NDCF). First, nine base classifiers for gas disasters are constructed on the training data set, including the backpropagation (BP) neural network classifier, naive Bayes (NB) classifier, K-nearest neighbor (KNN) classifier, logistic regression (LR) classifier, decision tree (DT) classifier, support vector machine (SVM) classifier, SVM classifier with cross-validation (SVMCV), random forest (RF) classifier, and gradient boosting DT (GBDT) classifier. Second, the K-means clustering algorithm is used to cluster the base classifiers according to their classification performance. Then, the best performing classifier in each cluster is selected to compose the first selection set. Third, the degree of combination fitness is used to filter the first selection set again to obtain the optimal base classifier result set. Finally, an ensemble classification model is constructed with the optimal base classifier result set. The experimental results on actual mine monitoring data show that compared with the BP, NB, KNN, LR, DT, SVM, SVMCV, RF, and GBDT classifiers, the accuracy of CS−NDCF increases by 7.34, 34.83, 8.28, 12.94, 5.51, 11.72, 6.47, 1.31, and 1.20%, respectively, and CS−NDCF achieves the best forecasting results. Thus, CS−NDCF is an effective method for identifying gas disasters and has a good application value.

1. INTRODUCTION

With the gradual increase in the depth and intensity of coal mining, gas disasters have the highest frequency, the largest number of casualties, and the most serious losses and seriously threaten the production safety in coal mines. To effectively prevent and control gas disasters, gas disaster prediction technology has become a research focus in the field of coal mine safety.1

In recent years, artificial intelligence methods have become an important research topic in the field of gas disaster prediction. Some intelligent algorithms, such as neural networks, grey theory, support vector machine (SVMs), and genetic algorithms, have been applied to gas disaster predictions.2−5 These methods have improved the prediction performance of gas disasters to a certain extent, but the accuracy and generalization ability of a single prediction model for gas disaster risk prediction need to be improved.

As ensemble learning methods have significant advantages in improving the generalization ability of learners, research on ensemble learning methods has become a hot issue in the field of machine learning.6 At present, ensemble learning is widely used in many fields, such as image processing, disease diagnosis, and circuit fault diagnosis,7−13 but its application in the field of coal mining is still relatively rare.

As early as 2002, Zhou et al. first proposed the concept of selective ensemble learning.14 This study pointed out that the classification or prediction effect achieves better results when multiple base learners are used to construct an ensemble model.15

Cite This: ACS Omega 2021, 6, 14059−14067
Read Online

Download Options

ACCESS
Metrics & More
Article Recommendations

Received: January 23, 2021
Accepted: May 14, 2021
Published: May 25, 2021

© 2021 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acsomega.1c00426
ACS Omega 2021, 6, 14059−14067
learning model rather than when all of the base learners are used. The selective ensemble idea has been widely used in the problem of ensemble learning, and how to choose better base learners is still the research focus of selective ensemble learning.15–17

In recent years, many studies have focused on the selection of base classifiers.18–23 The classification accuracy of individual classifiers in the base classifier set and the differences among individual classifiers are very important to improve the classification performance of an ensemble classification model. Therefore, the accuracy of base classifiers and the diversity of the base classifier set are taken as the criteria for selecting base classifiers.24,25 Hao et al. proposed an improved heuristic classifier selection method.26 This method uses an improved difference metric index to guide the base classifier selection algorithm to find the optimal base classifier set. To balance the relationship between the accuracy and diversity of base learners in ensemble learning, Mi et al. used a clustering algorithm based on a confusion matrix to select more suitable classifiers to construct the ensemble classification model.27 Li et al. studied and compared different ensemble learning strategies, such as the ensemble forward sequential selection, ensemble backward sequential selection, clustering selection, and hill-climbing selection strategies.28 These studies lay a good foundation for the selection of base classifiers.

To improve the prediction performance of coal mine gas disaster prediction methods, it is necessary to improve the selection efficiency of base classifiers. Thus, this paper proposes a base classifier selection strategy based on clustering selection and a new degree of combination fitness (CS–NDCF). First, a clustering algorithm is used to perform a preliminary screening of the base classifier set to obtain a new base classifier subset, which reduces the scale of the initial base classifiers. Then, a new degree of combination fitness (NDCF) method is used to search the base classifier subset to realize the secondary selection of base classifiers. The base classifier set filtered by the CS–NDCF method considers both the diversity of the base classifiers and the classification accuracy of the individual classifiers. The experimental results show that CS–NDCF improves the efficiency of not only the base classifiers but also the ensemble classification model to better identify the risk of coal mine gas disasters.

2. SELECTIVE ENSEMBLE CLASSIFICATION MODEL

2.1. Idea of the Model. A selective ensemble classification model consists of base classifiers, a base classifier selection strategy, and a base classifier combination rule, in which the base classifier selection strategy plays a crucial role in the construction of the selective ensemble classification model.

A good selection strategy should be to select the classifiers with large differences as much as possible. Therefore, we use a clustering method to cluster the base classifiers into different sets according to their differences. Then, the base classifier with the best classification performance is selected from each base classifier set. The clustering method realizes the preliminary screening of the base classifiers.

To select the best classifier from the clustering sets, NDCF is used. The classification performance of the selective ensemble classification model is related to the differences among the individual classifiers in the base classifier set and the classification accuracy of the individual classifiers. NDCF is defined by a function with two factors as variables. NDCF realizes the secondary selection of the base classifier and further improves the performance of the selection ensemble classification model.

Thus, we proposed a new selective ensemble classification model based on the CS–NDCF strategy.

2.2. Model Principle. Assume that the initial base classifier set is \( C_n = \{C_1,C_2,...,C_n\} \), where \( n \) represents the number of initial base classifiers. Assume that the clustering method is \( y_{\text{clus}}(C_{clus}, k) \), where \( k \) represents the clustering number. The clustering selection of the base classifiers is shown in eq 1.

\[
C_{\text{clus}} = y_{\text{clus}}(C_{\text{all}}, k)
\]  

where \( C_{\text{clus}} \) is the clustering result and is a set of multiple base classifier sets. \( C_{\text{clus}} = \{C_{clus1},C_{clus2},...,C_{clusk}\} \).

A new base classifier set is constructed by selecting a base classifier with the best classification performance from each base classifier \( C' \) of set \( C_{clus} \). Among them, the classifier performance evaluation method is expressed as \( z_{\text{eval}}(C_{clus}) \). The classification performance evaluation is described in formula 2.

\[
R_{\text{clus}} = z_{\text{eval}}(C_{\text{clus}})
\]  

where \( R_{\text{clus}} \) is a classifier set composed of the optimal base classifiers selected from each cluster and is named the clustering result set.

The NDCF method \( f_{\text{com}}(R_{\text{clus}}) \) is used to filter the clustering result set \( R_{\text{clus}} \) for the second time. The description of the NDCF method is shown in eq 3.

\[
R_{\text{opt}} = f_{\text{com}}(R_{\text{clus}})
\]  

where \( R_{\text{opt}} \) is the optimal base classifier result set.

When selecting base classifiers, the CS–NDCF selection strategy considers the classification performance of base classifiers and considers the differences between classifiers. By selecting the base classifier twice, the size of the base classifier set is reduced, and the classification performance of the selective ensemble classification method is further improved.

2.3. Clustering Selection. The clustering selection method performs the preliminary screening of base classifiers by clustering classifier samples. To assist the base classifier selection and better compare the performance between two base classifiers, the classifier selection criteria are defined.

2.3.1. Classifier Sample Clustering. The concept of selective ensemble learning is that the classification or prediction effect is not the best when all base learners participate in building an ensemble learning machine. If all base classifiers are used to construct an ensemble classification model, not only can the classification accuracy not be guaranteed but also the classification efficiency can be reduced. Therefore, the candidate classifier set must be optimized before building an ensemble learning model. In this paper, the K-means clustering algorithm is used to initially screen the base classifiers.

The clustering selection method uses the distance formula in ref 29 to measure the distance between two base classifiers. Its definition is shown in formula 4.

\[
d(C_i,C_j) = 1 - \frac{D(L_{i,j}(C_i,C_j))}{D(L)}
\]  

where \( L_{i,j} \) represents the number of samples in which classifier \( C_i \) is misclassified, \( L_{i,j} \) represents the number of samples in which classifier \( C_j \) is misclassified, and \( D(L) \) represents the total number of samples.

The base classifier selection strategy based on K-means clustering is designed according to the above distance formula.
First, base classifier models are built on the same data set, and the classification results of each base classifier are collected to form a vector space $O$, which is taken as the sample set of the $K$-means clustering algorithm. Among them, each sample of vector space $O$ represents the classification results of classifiers $C_i (i = 1, 2, \ldots, n)$. Second, the vector space $O$ is clustered, and the clustering result is $k$ group base classifier subsets. Finally, a classifier with the best classification performance is selected from each group classifier set to form the base classifier result set for clustering selection.

2.3.2. Classifier Selection Criteria. After clustering, the initial classifier set is divided into $k$ group classifier subsets. One of the best classifiers is selected from each classifier as the preliminary selection result of the base classifiers. The classifiers in the same cluster are different. The optimal classifier in each cluster is selected according to the classification performance of the classifiers. The definition of the performance evaluation index is shown in eq 5.

$$P = \frac{R}{T}$$  \hspace{1cm} (5)

where $R$ represents the classification accuracy of a single classifier, $T$ represents the run time, and $c$ is a constant with a value of 0 or 1. The larger the value of $P$, the better the classification performance of the classifier. To select the optimal classifier in each cluster, the classifiers in each cluster are compared in pairs. If the classification accuracies $R$ are different, let $c = 0$ and then the base classifiers are sorted according to the classification accuracy $R$. If the classification accuracy $R$ is the same, let $c = 1$ and a classifier that requires less time is regarded as a better classifier. According to the performance evaluation index $P$, base classifiers are sorted to obtain the optimal classifier in each cluster. Finally, the base classifier set of clustering selection can be determined.

2.4. NDCF Method. The base classifier selection strategy based on clustering selection is used to remove the redundant classifiers with high similarity to reduce the scale of the initial classifier set. To ensure that the ensemble classification model has higher classification accuracy, the clustering results are further analyzed by the NDCF method. Therefore, the base classifier sets are optimized again.

The NDCF method comprehensively considers the diversity of the base classifiers and the classification accuracy of the individual classifiers. The definition of the NDCF method is shown in eq 6.

$$D_{Q} = e^{\alpha x k - \beta x \rho}$$  \hspace{1cm} (6)

where $(\alpha + \beta) = 1, 0 \leq \alpha \leq 1, 0 \leq \beta \leq 1$. $R$ is the classification accuracy of a single classifier, and $\rho$ is the average of the correlation coefficients between a single classifier and each classifier of the base classifier set. The greater the difference between a single classifier and a classifier set (the smaller the $\rho$ value is), the higher the classification accuracy $R$ and the greater the NDCF $D_{Q}$ of the classifier and the classifier set.

Generally, there are correlation coefficients and $Q$ statistics methods to measure the difference between classifiers. The proposed method uses correlation coefficients to measure the difference between individual classifiers and the base classifier set. Its definition is shown in formula 7.

$$\rho_{ij} = \frac{N_{11}N_{00} - N_{01}N_{10}}{\sqrt{(N_{11} + N_{10})(N_{01} + N_{00})(N_{11} + N_{01})(N_{10} + N_{00})} \hspace{1cm} -1 \leq \rho_{ij} \leq 1}$$  \hspace{1cm} (7)

where $N_{ij}$ denotes the number of samples correctly classified by $C_i$ and $C_j$, $N_{10}$ denotes the number of samples correctly classified by $C_i$ and incorrectly classified by $C_j$, $N_{01}$ denotes the number of samples incorrectly classified by $C_i$ and correctly classified by $C_j$, and $N_{00}$ denotes the number of samples incorrectly classified by two classifiers at the same time. When the numbers of samples correctly classified or incorrectly classified by two classifiers are the same, the value of $\rho$ is 1. When the classification results of the two classifiers are different, the value is negative. Typically, the average of the correlation coefficients between individual classifiers in the base classifier set represents the output correlation coefficient of the base classifier set, and the smaller the value, the greater the difference in individual classifiers in the classifier set.

2.5. Model Framework. The implementation framework of the CS–NDCF selection strategy can be divided into three layers. First, the base classifiers in the initial base classifier set are modeled on the experimental data set. The $K$-means clustering algorithm is used to cluster the base classifiers, and the base classifiers are clustered into $K$ clusters. Then, the classifiers in each cluster are sorted according to the performance evaluation index. An optimal classifier is selected from each classifier cluster to obtain a set of base classifiers for preliminary screening. Finally, the NDCF method is used to filter the clustering selection base classifier set to obtain the optimal base classifier.
result set. The implementation process of the CS–NDCF method is shown in Figure 1.

2.6. Combination Rules Based on the Voting Method. In the ensemble classification, the voting method adopts the decision rule of majority voting, which is used in solving contradictory problems in human sociology. The decision result formula based on voting rules is as follows

$$E(x) = \begin{cases} \omega_i, & T_k(x \in \omega_i) = \max_{k} T_k(x \in \omega_i) \geq \alpha \times K \\ M + 1, & \text{otherwise} \end{cases}$$

(8)

Among them

$$T_k(x \in \omega_i) = \sum_{i=1}^{K} T_k(x \in \omega_i), \quad i = \ldots, M$$

(9)

$$\omega(0 \leq \alpha \leq 1)$$ is a constant, which means that the category must reach above the ratio value can be regarded as the category to which the sample belongs.

3. ALGORITHM DESCRIPTION

In the implementation of the CS–NDCF method, the base classifiers are first clustered. Second, the best classifier is selected from each clustering result set. Then, NDCF is used as a judgment basis to select the clustering result set. Finally, the optimal base classifier result set is determined. The algorithm screens the base classifiers twice by combining clustering selection and the NDCF method and improves the classification performance of the selective ensemble classification model.

The algorithm implementation of the CS–NDCF selection strategy is shown in Table 1.

4. RESULTS AND DISCUSSION

4.1. Data Description. The experimental data are the monitoring data from sensors in a coal mine in Shanxi Province in July 2020, and the data collection granularity is one minute. A description of the data is shown in Table 2.

As seen in Table 2, the experimental data contain seven monitoring indicators, namely, seven types of sensors. The number of sensor monitoring equipment corresponding to different detection indexes is different. Among them, the same type of sensor monitored the same environmental index at different locations in the underground coal mine, so the data set contains a total of 24 attributes. Since the three indexes of the instantaneous mixing flow are under standard conditions, the mixing accumulation under standard conditions and the instantaneous mixing flow under industrial and mining conditions are not limited here, and the value ranges of these three monitoring indexes are not given in this paper.

The accuracy and time are used to evaluate the performance of the gas disaster risk prediction model. The accuracy definition is shown in formula 11.

$$\text{Acc: } R_{\text{acc}} = \frac{n_i}{n}$$

(11)

where \(n\) represents the total number of samples, and \(n_i\) is the number of samples with the same gas disaster risk level label in the prediction result and the level label value in the real data.

Table 1. Algorithm CS–NDCF Selection Strategy

| Algorithm CS–NDCF Selection Strategy |
|--------------------------------------|
| **Input:** the number of classifiers \(n\), the base classifier \(C_i\), the base classifier \(C_i = 1, 2, \ldots, n\), the clustering number \(k\). |
| **Output:** the base classifier result set with optimal performance \(S_i\), the classification accuracy \(\max R\). |
| **Initialization:** the classifier set \(\mathcal{S} = \{1, 2, \ldots, n\}\), the classifier set \(S_i = \emptyset\). |
| Construct vector space \(\sigma\), each vector \(\sigma = (1, \ldots, K)\) in the vector space is the classification result of \(C_i\). |
| The vector space \(\sigma\) is taken as a sample set, and the K-means clustering algorithm is used for clustering to obtain the \(k\) group classifier set \(S_k = \{1, 2, \ldots, k\}\). |
| for \(j = 1, k\) |
| acc = 0; |
| for each classifier \(C_i\) in cluster \(S_k\) |
| Calculate the classification performance evaluation index \(P\) of each classifier; |
| if \(P > \text{acc}\) |
| acc = \(P\); |
| \(S_i = S_{\text{acc}}\); |
| end if |
| end for |
| \(S_k = S_{\text{acc}}\); |
| end for |
| Find the classifier \(C_{\text{acc}}\) with the highest classification accuracy in \(S_k\); |
| \(S_k = S_{\text{acc}}\); |
| \(S_{\text{acc}} = S_{\text{acc}}\); |
| for \(p = 1: k+1\) |
| max = 0; |
| for \(i = 1: n\) |
| Calculate the degree of combination fitness \(D_{\text{acc}}\) of the classifier \(C_i\) and \(S_{\text{acc}}\); |
| if \(D_{\text{acc}} > \text{max}\) |
| max = \(D_{\text{acc}}\); |
| \(C_{\text{acc}} = C_i\); |
| end if |
| \(S_{\text{acc}} = S_{\text{acc}}\); |
| \(S_{\text{acc}} = S_{\text{acc}}\); |
| end for |
| \(S_{\text{acc}} = S_{\text{acc}}\); |
| end for |
| Calculate the classification accuracy \(R\) of the ensemble classifier constructed by \(S_{\text{acc}}\); |
| if \(R > \text{acc}\) |
| acc = \(R\); |
| \(S_{\text{acc}} = S_{\text{acc}}\); |
| end if |
| end if |
| return \(S_{\text{acc}}\); |

The gas disaster risk level is a label attribute and is obtained through the threshold judgment method. To reasonably evaluate the safety level of coal mines, this thesis divides the
gas disaster risk level into three levels, namely, no gas disaster risk, gas disaster threat, and gas disaster risk. The numbers “0”, “1”, and “2” are used to identify the three danger levels of gas disasters.

There are 4761 records with three categories, of which category label 0 has 1403 records, label 1 has 3299 records, and label 2 has 59 records. The categories are extremely unbalanced. The bootstrap method is used for resampling to balance the types of data. The resampled data set has 9895 records, of which label 0 has 3299 records, label 1 has 3299 records, and label 2 has 3297 records. The data set is divided into a training set and a test set at a ratio of 4:1.

### 4.2. Data Preprocessing

A typical data set for the task addressed in this paper consists of gas concentration data and related data collected by the monitoring equipment over a certain period of time. However, these data will generally contain various missing values and outliers due to factors, such as process alteration, equipment failure, or other factors related to human activity. Therefore, it is necessary to preprocess the monitoring data that are collected in real time to improve the accuracy of mine gas disaster forecasting.

The first step in data preprocessing is to apply the Laida criterion to process the noise in the monitoring data.

The Laida criterion is one of the most common and simplest criteria for discriminating errors. Suppose that the monitoring time series data obtained by a sensor is \( X = \{x_1, x_2, \ldots, x_n\} \), then the process of finding outliers with the Laida criterion is as follows.

If the absolute value of a residual is larger than 3 times the standard deviation, meaning that inequality (13) is true when combined with formula (12), then this error is considered to be excessively large, and the corresponding measured value is considered to be an outlier that should be rejected; this results in a missing value for the corresponding measurement time.

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \quad (i = 1, 2, \ldots, n) 
\]

\[
|x_i - \bar{x}| > 3\sigma \quad (13)
\]

where \( \bar{x} \) is the average value of the data set and \( \sigma \) is the standard deviation of the data set.

The second step is to apply Lagrange interpolation to fill in the resulting missing values of the monitoring data sequence to obtain a complete set of processed monitoring data.

Finally, when predicting the risk of gas disasters, it is considered that the value ranges of different parameters may vary greatly. If the monitoring data are directly involved in the calculation, it will affect the prediction result. To eliminate the influence of the difference in the value range and make the data comparable, minimum–maximum normalization is used to normalize the data, and the value is mapped to \([0,1]\). The normalization formula is shown in eq. 14.

\[
\text{norm}(x_i) = \frac{x_i - \text{min}(X)}{\text{max}(X) - \text{min}(X)} \quad (14)
\]

where \( \text{norm}(x_i) \) is the normalized value of \( x_i \), \( \text{min}(X) \) is the smallest value in data set \( X \), and \( \text{max}(X) \) is the largest value in data set \( X \).

### 4.3. Attribute Reduction

There are 23 inputs and one class attribute in the experimental data set \( D_{\text{set}} = \{T1, T2, C1, C2, S1, S2, S3, S4, S5, S6, P1, P2, P3, P4, P5, P6, SS, W1, W2, W3, W4, W5, gas, level\} \), and the correlation among these attributes has a certain impact on the classification results. Therefore, correlation analysis is used to analyze the correlation among the attributes of the original data set, and the sample attributes in the monitoring data are reduced according to the correlations between each attribute and the gas concentration.

Through the normal test of coal mine monitoring data, it is found that the data do not have a normal distribution. Therefore, the Spearman method is used to calculate the correlation coefficient \( r \). The Spearman correlation coefficient \( r \) is defined in formula 15.

\[
r = 1 - \frac{6 \sum_{i=1}^{n} (R_i - Q_i)^2}{n(n^2 - 1)} \quad (15)
\]

where \( R_i \) and \( Q_i \) are the ranks of \( x_i \) and \( y_i \), respectively. \( -1 \leq r \leq 1 \). \( r > 0 \) indicates a positive correlation, \( r < 0 \) indicates a negative correlation, and \( |r| = 0 \) indicates that there is no linear relationship.

\( 0 < |r| \leq 1 \) represents the different degrees of the linear correlation between two variables: \( |r| \leq 0.3 \) indicates no linear correlation; \( 0.3 < |r| \leq 0.5 \) indicates a low linear correlation; \( 0.5 < |r| \leq 0.8 \) indicates a significant linear correlation; \( |r| \geq 0.8 \) indicates a high linear correlation; and \( |r| = 1 \) indicates a complete linear correlation.

By analyzing the correlation degree of each sample attribute in \( D_{\text{set}} \) with the gas concentration, the correlation vector of the sample attribute is as follows:

\[
p = [-0.16860, -0.09747, -0.70467, -0.60151, 1.00000, -0.54862, -0.32789, -0.48926, -0.36574, -0.64087, -0.05516, -0.02539, 0.45129, 0.06134, -0.02898, -0.39783, 0.06732, 0.44400, -0.48754, 0.27261, -0.48909, -0.48345, 0.33333] \]

The attributes with correlation coefficients greater than 0.3 are selected from the sample attribute set as the experimental sample attributes. The new sample attribute set is as follows:

\[
D_{\text{new}} = \{T1, T2, gas, P1, P2, P3, P4, P5, P6, SS, W1, W2, W3, W4, W5, level\}
\]

After the correlation analysis method is used for attribute reduction and sample attribute sets \( D_{\text{set}} \) and \( D_{\text{new}} \) are compared, it can be concluded that the new sample attribute set \( D_{\text{new}} \) has eight fewer sample attributes than the original sample attribute set \( D_{\text{set}} \).

### 4.4. Experimental Steps

The implementation steps of the CS–NDCF selection strategy are as follows.

1. The base classifier models are constructed on data set \( D_{\text{new}} \).
(2) The candidate base classifiers \( C_i \) \((i = 1,2,...,n)\) are clustered with the K-means algorithm, and the number of clusters is denoted as \( k \).

(3) The classifiers \( C_m \) with the best classification performance in each cluster \( C_j \) \((j = 1,2,...,k)\) are selected to join classifier set \( S_m \).

(4) The base classifier \( C_m \) with the highest classification accuracy in \( S_m \) is determined, \( C_m \) is added to empty set \( S_r \), and \( C_m \) is deleted from \( S_m \).

(5) The degrees of combination fitness \( D_{Qj} \) between each classifier contained in \( T \) and the classifier set \( S(i = 1,2,...,k) \) are calculated. Classifier \( C_m \) with the greatest degree of combination fitness is selected and added to \( S_r \) to obtain classifier set \( S_{r+1} \). Additionally, \( C_m \) is deleted from \( S_m \).

(6) If \( S_m \) is not empty, step (5) is repeated; otherwise, go to step (7).

(7) The optimal base classifier result set is selected from \( k \) classifier sets \( S(i = 1,2,...,k) \) to obtain \( S_o \).

(8) The classifiers in \( S_o \) are combined with a base classifier combination rule based on the voting method to construct a selective ensemble classification model.

4.5. Experimental Results and Analysis. In the selective ensemble classification, the diversity of the base classifier set is taken as one of the criteria for selecting base classifiers so that heterogeneous base classifiers would be chosen, that is, the types of base classifiers are different. Nine common classifiers are selected as base classifiers, including the backpropagation (BP) neural network classifier, naive Bayes (NB) classifier, k-nearest neighbor (KNN) classifier, logistic regression (LR) classifier, decision tree (DT) classifier, SVM classifier, SVCM classifier using cross-validation (SVCMCV), random forest (RF) classifier, and gradient boosting DT (GBDT) classifier. The differences in these nine base classifiers are relatively large, and they are commonly used, which is more suitable for gas disaster risk identification.

When clustering \( k \) parameters \( \alpha \) and \( \beta \) take different values and the CS–NDCF method obtains different base classifier result sets. The ensemble classification model is constructed by selecting the base classifier set with the best classification performance. To verify the effectiveness of the proposed method, the performance of the ensemble classification model is compared with those of five selection methods, including the exhaustive method, ensemble forward sequential selection strategy, heuristic method, clustering selection strategy, and degree of combination fitness method.

4.5.1 Classification Performance of the Base Classifiers. The classification performance of the base classifiers in the test data set before resampling and after resampling is shown in Table 3.

As shown in Table 3, after resampling, the recognition rate of BP increased by 4.68%, that of NB increased by 0.92%, that of KNN decreased by 3.94%, that of LR increased by 2.20%, that of DT decreased by 0.24%, that of SVM increased by 3.59%, that of SVCMCV increased by 1.51%, that of RF increased by 3.89%, and that of GBDT increased by 5.39%. Except for the KNN and DT classifiers, the classification effect of the rest of the classifiers on the data set after resampling is better than the effect before resampling. It is proven that the classifiers can achieve better classification effects on the data set with balanced categories. After resampling, the running times of BP, SVM, and SVCMCV increased and those of NB, KNN, LR, and RF decreased.

### Table 3. Classification Performance of the Base Classifiers

| classifier | accuracy before resampling (%) | running time before resampling (s) | accuracy after resampling (%) | running time after resampling (s) |
|------------|---------------------------------|-----------------------------------|-------------------------------|----------------------------------|
| BP         | 82.56                           | 14.179                            | 86.42                         | 31.111                           |
| NB         | 68.17                           | 0.129                             | 68.80                         | 0.005                            |
| KNN        | 89.18                           | 0.431                             | 85.67                         | 0.133                            |
| LR         | 80.36                           | 1.512                             | 82.13                         | 0.100                            |
| DT         | 88.13                           | 0.038                             | 87.92                         | 0.046                            |
| SVM        | 80.15                           | 1.439                             | 83.03                         | 6.668                            |
| SVCMCV     | 85.82                           | 35.969                            | 87.12                         | 246.79                           |
| RF         | 88.13                           | 0.391                             | 91.56                         | 0.120                            |
| GBDT       | 86.97                           | 2.882                             | 91.66                         | 4.489                            |

Before resampling, the best classifier was KNN, with an accuracy rate of 89.18%. After resampling, GBDT achieved the best classification effect of 91.66%, and RF also achieved a better classification effect of 91.56%. RF and GBDT are ensemble classification methods, and their classification effects are significantly higher than that of a single classifier method, indicating that the ensemble learning effect is better than that of a single learner under the condition of class balance.

4.5.2 Selection Result of the CS–NDCF Method. When the CS–NDCF method is used to select the base classifiers, the clustering algorithm first selects the base classifiers to obtain the clustering result set. Then, the clustering result set is further filtered by the NDCF method to obtain the optimal base classifier result set. The selection results of the CS–NDCF selection strategy for the base classifiers with different values of \( k \) are shown in Figure 2.

Figure 2 shows that the difference in the \( k \) value affects the final choice of classifier.

4.5.3. Analysis of the Ensemble Classification Results Based on the CS–NDCF Selection Strategy. To better understand the impact of the clustering number on the classification result of the selective ensemble classifier, the selective ensemble classification results under different clustering numbers are compared, as shown in Figure 3.

Figure 3 shows that when the clustering number \( k = 5, \alpha = 0.8 \), and \( \beta = 0.2 \), the classification accuracy of the selective ensemble classifier based on CS–NDCF is the highest. Under different values of \( k \), the average time is 4.137 s.

By combining Table 2 and Figure 3, it can be seen that the classification effect of the selected ensemble method proposed in this paper is significantly better than that of a single classifier. After resampling, the best single classifier is DT, and its effect on the test set is 87.92%. The best integrated classifier is GBDT, with an effect of 91.66%, while the effect of the classifier proposed in this paper is 92.76%. Compared with DT and GBDT, our model increased the result by 5.51 and 1.20%, respectively.

5. CONCLUSIONS AND THE FUTURE WORK

To reduce the economic damage caused by coal mine gas disasters and protect the lives of miners, it is necessary to use modern technology to study an effective gas disaster risk identification method based on real-time monitoring data from coal mines. Thus, in this paper, a gas disaster risk ensemble classification model based on CS–NDCF is proposed.

First, nine base classifiers are constructed based on monitor data. Second, the K-means clustering algorithm is used to initially screen the base classifiers. After this, the classifiers with
large differences are retained, and the number of base classifiers is reduced. It effectively improves the base classifier search efficiency. Then, the NDCF selection strategy is used for secondary selection. NDCF not only considers the differences among the base classifiers but also considers the classification accuracy of the base classifiers, which effectively improves the generalization performance of ensemble classifiers. The experimental results show that the effectiveness of the CS−NDCF
The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

This work was supported by the State Key Research Development Program of China (Grant No. 2018YFC0808303), the National Natural Science Foundation of China (Grant No. 51974236), the Natural Science Foundation of Shaanxi Province (Grant No. 2019JLM-10), and the Scientific and Technology Program Funded by Xi’an City (Grant No. 2020KJRC0069)

**REFERENCES**

(1) Li, R. Q.; Shi, S. H. L.; Nian, Q. F.; Zhu, H. P.; Zhu, C. H. Q. Risk pattern recognition of coal mine gas explosion disaster based on PSO-SVM. *China Saf. Sci. J.* 2013, 23, 38–43.

(2) Li, S. P.; Wang, X. H. Prediction of gas outburst danger degree based on IPSO-SVM. *Mod. Electron. Technol.* 2016, 39, 21–25.

(3) Wen, T. X.; Yu, F. G.; Shao, L. S. H. Coal and gas outburst probability neural network prediction model based on grey relational entropy. *Appl. Res. Comput.* 2018, 35, 3326–3329.

(4) Li, F.; Chen, X.; Lin, W.; Pan, H.; Jin, X.; Hua, X. Coal and gas outburst prediction based on improved PSO optimized LS-SVM parameters. *Surf. Coat. Technol.* 2017, 319, 129–135.

(5) Liu, Y. J.; Zhao, Q.; Hao, W. L. Research on gas concentration prediction based on BP neural network optimized by genetic algorithm. *Min. Saf. Environ. Protect.* 2015, 42, 56–60.

(6) Jia, P. T.; Li, Y. Facial expression recognition research based on selective ensemble classifiers. *Appl. Res. Comput.* 2017, 34, 3825–3827, 3833.

(7) Ni, D.; Feng, G.; Shen, L.; Zhang, X. Selective ensemble classification of image steganalysis via deep Q Network. *IEEE Signal Process. Lett.* 2019, 26, 1065–1069.

(8) Li, W.; Liu, H.; Wang, Y.; Li, Z.; Jia, Y.; Gui, G. Deep learning-based classification methods for remote sensing images in Urban built-Up areas. *IEEE Access* 2019, 7, 36274–36284.

(9) Huo, D. X.; Liu, H.; Shang, Z. H. A method of pediatric disease diagnosis based on heterogeneous integration learning. *Comput. Appl. Software* 2018, 35, No. 157.

(10) Wang, S. L.; Yin, Y. L.; Gao, G. B.; Wei, B. Z.; Zheng, Y. J.; Yang, G. P. Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing* 2015, 149, 708–717.

(11) Hosni, M.; Abnane, I.; Idri, A.; Carrillo de Gea, J. M.; Fernández Alemán, J. L. Reviewing ensemble classification methods in breast cancer. *Comput. Methods Programs Biomed.* 2019, 177, 89–112.

(12) ALzubi, J. A.; Bharathikannan, B.; Tanwar, S.; Manikandan, R.; Khanna, A.; Thaventhiran, C. Boosted neural network ensemble classification for lung cancer disease diagnosis. *Appl. Soft Comput.* 2019, 80, 579–591.

(13) Johnson, J.; Giraud-Carrier, C. Diversity accuracy and efficiency in ensemble learning. An unexpected result. *Intell. Data Anal.* 2019, 23, 297–311.

(14) Zhou, Z. H.; Wu, J. X.; Tang, W. Ensembling neural network: Many could be better than all. *Artif. Intell.* 2002, 137, 239–263.

(15) Wang, Z. Y.; Lu, C.; Zhou, B. Fault diagnosis for rotary machinery with selective ensemble neural networks. *Mech. Syst. Signal Process.* 2018, 113, 112–130.

(16) Zhang, Y.; Liu, B.; Yu, J. X. A selective ensemble learning approach based on evolutionary algorithm. *J. Intell. Fuzzy Syst.* 2017, 32, 2365–2373.

(17) Liu, L.; Wang, B. S.; Yu, B.; Zhong, Q. X. In *A Novel Selective Ensemble Learning Based on K-Means and Negative Correlation*, *Proceedings of 2nd International Conference on Cloud Computing and Security*, Springer Verlag, 2016; pp 578–588.

(18) Ruta, D.; Gabrys, B. Classifier selection for majority voting. *Inf. Fusion* 2005, 6, 63–81.

(19) Lin, C.; Chen, W. Q.; Qiu, C.; Wu, Y.; Krishnan, S.; Zou, Q. Ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing* 2014, 123, 424–435.

(20) Tama, B. A.; Im, S.; Lee, S.; Guaricci, A. I. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble[J]. *BioMed Res. Int.* 2020, 2020, 1–10.

(21) Zhang, J. N.; Xia, K. W.; He, Z. P.; Yin, ZhX.; Wang, S. J. Semi-Supervised Ensemble Classifier with Improved Sparrow Search Algorithm and Its Application in Pulmonary Nodule Detection[J]. *Math. Probl. Eng.* 2021, 2021, 1–18.

(22) Chandra, Y. D.; Saurabh, P. An Experimental Study of Diversity of Diabetes Disease Features by Bagging and Boosting Ensemble Method with Rule Based Machine Learning Classifier Algorithms[J]. *SN Comput. Sci.* 2021, 2, No. 50.
(23) Yang, X. Y.; Zhang, Y. F.; Lv, W.; Wang, D. Image recognition of wind turbine blade damage based on a deep learning model with transfer learning and an ensemble learning classifier [J]. Renewable Energy 2021, 163, 386–397.

(24) Niu, P.; Wei, W.; Li, J. J.; Guo, J. G. Classifier selection method. Comput. Eng. 2010, 36, 163–165.

(25) Aizhong, M. I.; Yao, L. U. Classifier selection method with hybrid diversity measure. J. Front. Comput. Sci. Technol. 2018, 12, 1522–1530.

(26) Hao, H. W.; Chen, Z. Q. New heuristic classifier selection method. Comput. Eng. 2008, 34, 206–208.

(27) Mi, A. Z.; Zhang, P. In A Method of Classifier Selection Based on Confusion Matrix. J. Henan Polytechnic University, 2017.

(28) Li, K.; Han, Y. X. Research on ensemble learning methods based on selective strategies. Comput. Eng. Appl. 2011, 47, 26–30.

(29) Giacinto, G.; Roli, F. An approach to the automatic design of multiple classifier system. Pattern Recognit. Lett. 2001, 22, 25–33.