Risk Assessment of Gas Explosion Disaster Based on Random Forest Model

Qing Zhang$^{1,2,*}$

$^1$School of Safety Science and Emergency Management, Wuhan University of Technology, Wuhan, Hubei, 430070, China

$^2$China Research Center for Emergency Management, Wuhan University of Technology, Wuhan, Hubei, 430070, China

$^*$Corresponding author’s e-mail: zq15071244507@163.com

Abstract. The risk assessment of gas explosion is difficult to determine due to multiple factors. To solve this problem, a random forest model is applied to the risk assessment of gas explosion disasters in coal mines. Based on the disaster system theory, the pregnant environment, disaster factors and disaster-receiving bodies were used as secondary indicators, and the corresponding 24 tertiary indicators were selected as the evaluation indicator system. The example analysis shows that the accuracy rate of gas explosion risk category prediction of the random forest model is very high (up to 100%), and it is highly practical for small sample problems with high data dimensions; Coal seam spontaneous ignition period led to an average reduction of 0.2 and caused the Gini index to drop by an average of 1.5. The average accuracy decrease caused by the qualified rate of air volume at the wind-use site and the coal seam gas content is almost 0, and the decrease of the average Gini index is less than 0.15.

1. Introduction

Coal gas explosion, as a sudden and extremely destructive mine accident, will not only cause a large amount of economic loss, but also pose a huge threat to the life and safety of mine workers. It is also extremely unfavorable, so the risk assessment of gas explosion disasters is particularly important, which has practical significance for the prevention of gas explosions. Wang Wenchao$^{(1)}$ calculated the risk of coal mine gas explosion accidents based on a semi-quantitative evaluation model; Sun Bin$^{(2)}$ used a gray correlation method to determine the risk level of a gas explosion hazard source; Wang Jingchun and others$^{(3)}$ reduced the randomness and ambiguity in risk assessment through a two-dimensional cloud model of entropy weight; Wang Wanqing and others$^{(4)}$ established a gas explosion hierarchy analysis model. Zhao Yajun$^{(5)}$ used the Failure Tree Analysis (FTA) to display the relationship of 24 causes, and calculated the minimum cut set, minimum path set, and structural importance. The analytic hierarchy method is simple and requires a small number of samples, but the accuracy is low and the randomness is large; the training accuracy of neural networks and support vector machines is high, but the required samples are large. Based on the disaster system theory, a random forest model is established to achieve a gas explosion risk assessment with high training accuracy on a small sample.
2. Research method

2.1 Decision Tree Overview
Decision tree is a single classifier that is widely used in the field of prediction and classification. It is a very classic prediction model. It reflects the mapping relationship between target feature indicators and target values through the branching of the tree. The decision tree algorithm is CART algorithm. The decision tree as a base classifier exists in a random forest model. Its structure is simple and easy to understand, and it can process high-dimensional data with high accuracy and classification efficiency. However, it also has overfitting and convergence to the local optimal solution.

2.2 Random forest overview
The classification model composed of multiple decision trees \( \{t(X, \emptyset k), k = 1, 2, \ldots \} \) is a random forest model, where \( \{\emptyset k\} \) is an independent and identically distributed parameter set, and X is selected. The independent variable of the classification result of any decision tree is the optimal solution selected by its voting.

In the first step of bootstrap sampling, k samples are put back in the initial training set D, and at the same time, the number of features (m) in the initial training set is the same as that of each sample. In this way, the returned samples reflect the randomness of the random forest on the training set.

The second step is to establish a decision tree model. The k samples extracted in the previous step are used to construct the corresponding k decision tree models, and the corresponding k classification results are obtained \( \{h_1(X), h_2(X), \ldots, h_k(X)\} \). The feature selection of the random forest model is not to select the best feature among all the features, but each sample randomly selects n (n≤m) features to form the feature set used for splitting, and then selects the best feature from the feature set and then It performs node growth. If n < m, it shows the randomness of the random forest on each decision tree fork.

The third step is the final vote. The random forest obtained in the first two steps will be voted according to the k classification results. The final voting decision is the final classification of the random forest model.

\[
f(x_1) = \text{m\_vote}\{h_i(x)\} \quad (i = 1, 2, \ldots, k)
\]

3. Risk assessment model of gas explosion disaster

3.1 Determination of risk index system for gas explosion disaster
A perfect indicator system is essential for the establishment of the model, especially for the gas explosion disaster system. In this paper, the disaster system theory is used as the basis for the construction of the index system. The inducing factors of gas explosions are considered as disaster-causing factors, and the environmental factors that play a role in incubating gas explosions are considered as disaster-preventing environments. As the affected area, the affected area constitutes the three major subsystems of the hazard factor, the environment for the disaster, and the affected body. This article refers to national, provincial, municipal and other departments' coal mine safety evaluation standards, consults the literature, fully considers the characteristics of coal mine gas explosion accidents, selects indicators with high frequency as indicators, and constructs the risk of coal mine gas explosion according to the cause analysis of the gas explosion disaster system. Three-level evaluation index system, the first level is the gas explosion disaster risk target level; the second level index includes the evaluation of the three major subsystems of the gas disaster risk system, the danger of disaster factors, the stability of the pregnant environment, and the Vulnerability; the three-level indicators select the most representative and specific indicator characteristics. The index system constructed is shown in the table.

In the process of coal mine gas explosion disaster risk assessment, the evaluation system does not blindly pursue the number of indicators, and the applicability and appropriateness of the selected evaluation indicators often affect the accuracy of the prediction results. Many model methods use
subjective analytic hierarchy process in combination with the screening process of indicators. Human factors are inevitably added because the subjective judgment of people causes the loss of indicator information. However, in the data preprocessing stage, random forests do not require de-dimensionalization and normalization, and can be directly operated. In addition, the evaluation index selection stage has a strong data set ability to process high-dimensional and large-volume data, does not require feature selection, and can give scores on the importance of each index. In the risk evaluation stage, the influence of human subjective assignment is avoided, and the accuracy is high.

3.2 Risk Assessment Level of Gas Explosion Disaster Determined

Due to the special operating conditions of coal mine production, many attributes of the evaluation indicators are not single items. These indicators can only be divided on a percentage scale, such as the complexity of geological structures, the reasonableness of ventilation methods, and the timeliness of mine rescue. Gas content in coal seams Evaluation indexes such as absolute gas emission need to be measured based on the specific dimensions of their physical and chemical properties. Based on the qualitative descriptions of casualties and economic property losses in coal mine gas explosion accidents based on relevant national regulations, the assignment of three-level indicators for determining the vulnerability of disaster-receiving bodies such as the installation rate of explosion-proof and explosion-proof facilities is determined.

After the evaluation of the indicators of gas explosion accidents in coal mine production is completed, the data needs to be standardized to resolve the differences in the content, dimensions and value ranges of the indicators. At the same time that the data is standardized, it will inevitably cause the loss of data information of each indicator. Therefore, this paper uses relative numbers to describe the meaning and range of each indicator, and establishes the demarcation point of each indicator's risk level. Suppose the number of gas explosion samples is m, which constitutes the sample set W = [W₁, W₂, ..., Wᵢ, ..., Wₘ]ᵀ, and any of the samples Wᵢ = [Wᵢ₁, Wᵢ₂, ..., Wᵢⱼ, ..., Wᵢₘ, uᵢ]ᵀ are all determined by n index attributes and a risk level uᵢ, yᵢ ∈ Y, risk level vector U = [u₁, u₂, ..., uᵢ, ..., uₘ]ᵀ. Let vᵢ(k) (k = 1, 2, ...) be the vector (xᵢ(k) < xᵢ(k+1)) corresponding to the risk level cut-off point xᵢ(k) of the kth indicator. The normal form of the data xᵢ is as follows,

\[ wᵢⱼ = \begin{cases} vᵢ(k+1) + (vᵢ(k) - vᵢ(k+1)) \frac{wᵢ(k+1) - wᵢ(k)}{wᵢ(k+1) + wᵢ(k)}, & wᵢ(k) < wᵢ(k+1) \\ vᵢ(k+1), & wᵢ(k+1) < wᵢ \end{cases} \]

After completing the sample is the data normalization processing, it is necessary to classify the risk of coal mine gas explosion disasters. This process should not only be based on the significance of each index, the scope of the assignment, and the principles of assignment, but also must fully consider the characteristics of the mine operation and the factors affecting its safety in production, and then consider the general classification standards and expert opinions. Developed a coal mine gas disaster risk classification comment set R = {unsafe; less secure; general safety; more secure; safe}. The five items in the comment set R correspond to five wind direction level vectors (1, 2, 3, 4, 5) The classification of the risk model categories for each indicator is shown in the table below.

| Serial number | Feature | Risk level demarcation point |
|---------------|---------|-----------------------------|
|               |         | 1   | 2   | 3   | 4   | 5   |
| 1             | W₁: Geological structure complexity (points) | 20  | 40  | 60  | 80  | 90  |
| 2             | W₂: Danger of coal dust explosion (points)  | 20  | 18  | 15  | 14  | 10  |
| 3             | W₃: Fan running stability (points)          | 60  | 70  | 80  | 90  | 100 |
| 4             | W₄: Fan, blower intact rate (%)             | 70  | 75  | 80  | 85  | 95  |
| 5             | W₅: Reasonable ventilation method (points)  | 60  | 75  | 85  | 90  | 95  |
| 6             | W₆: Qualified rate of wind volume (%)       | 80  | 85  | 90  | 95  | 100 |
4. Model application and result analysis

4.1 Data import
The evaluation index system constructed in accordance with 2.1 collects and organizes the data of various three-level indicators. References to some data in the research done by Li Runqiu, etc., and query the remaining indicator data through the State Administration of Work Safety Administration and the National Coal Mine Safety Production Network, etc., and finally integrate all indicator data to obtain all 20 Sample data.

4.2 Analysis of prediction results
Import 20 samples into the constructed random forest model and SVM model, and then randomly extract 80% of these samples as training samples and the remaining 20% as prediction samples. Use the constructed random forest classification. The model makes classification predictions on the test set, and performs gas explosion disaster risk assessment on the samples in the test set. The accuracy of the classification results is Accuracy = 100%. The classification results are shown in the following table. It can be seen that the risk level of the test set is [4 4 3 4], the classification accuracy of the random forest model is 100%, and the number of false positives is zero. The risk level of the test set obtained by the SVM model is [4 4 3 3]. The accuracy of the classification results of the SVM model is 80%, and the number of false positives is 1. The OOB misclassification rate of out-of-bag data stabilizes at 0.2 after the number of decision trees exceeds 100.
Table 2. Random forest and SVM accuracy evaluation.

| Model     | Test Set Risk Level | Number of false positives | Accuracy |
|-----------|---------------------|---------------------------|----------|
| Random forest | 【4 4 3 4】          | 0                         | 100%     |
| SVM       | 【4 4 3 3】          | 1                         | 80%      |

This article uses a small sample, and in order to establish a perfect coal mine gas explosion disaster system in line with the disaster system theory, many risk wind evaluation indicators are selected in this paper, but the random forest model constructed can still maintain a very low error rate. Random forest has strong generalization ability, strong adaptability to fresh samples, and can process high-dimensional data.

4.3 Index importance analysis

Coal mine gas explosion as a complex coupled disaster system, combined with the disaster system theory, its disaster risk assessment needs to consider a number of indicators in the hazard factor, the pregnancy environment, and the affected body, and the impact of different indicators on the coal mine gas explosion They are not the same. The importance of each index to the coal mine gas explosion grade is of practical significance to the prevention and control of coal mine gas explosion. The SVM model cannot explore the influence of various indicators in the sample on the prediction results, and cannot compare the importance of each indicator. The random forest model constructed in this paper can output the average decline value of the accuracy of the prediction results and the average decline value of the Gini index after the input samples are input, as shown in the following figure, and use this as a basis to evaluate the importance of each indicator. The larger the two values are, the greater the importance of the indicator is, and vice versa.

As can be seen from the figure above; one of the indicators that has the most influence on the accuracy of the test set is the spontaneous combustion period of the coal seam, which caused an average drop of accuracy of 0.1995; secondly, the danger of coal dust explosion, the intact rate of fans and air ducts, the availability of technical personnel and safety The change of the duty execution rate also has a greater impact on the accuracy rate; changing the wind volume qualification rate and safety education and training intensity of the wind use site has a small impact on the prediction results, and the accuracy of the wind volume qualification rate caused by the wind use site decreases on average even At 0, the Gini index dropped by 0.168 on average. From the Gini coefficient, it can also be seen that the coal seam spontaneous ignition period is an index that has the greatest influence on the prediction result, causing the Gini index to decrease by 1.514 on average. The least affected indicators are the qualified rate of air volume at the wind-use site and the coal seam gas content, respectively. The average drop was 0.168 and 0.1423. It can be seen that the coal seam spontaneous ignition period is the most important indicator, the technical personnel operation qualification rate and supervision intensity are second, and the wind site qualification rate and coal seam gas content are the most important.
5. Conclusion
1) Based on random forest theory, construct a random forest model for coal mine gas explosion disaster risk assessment. Through the example verification, it can be seen that the random forest model has a high prediction of the risk category of coal mine gas explosion (the verification of the example in this article reaches 100%). In the analysis of a small sample of large-scale coal mine gas explosion risk evaluation indicators and incomplete data information analysis The adaptability is better than the SVM model.

2) Regularity analysis of the importance index of coal mine gas explosion characteristics, the coal seam spontaneous ignition period is the most important, and it has the greatest influence on the prediction classification results. The wind volume qualification rate and the coal seam gas content are the least important. It is difficult to analyze the possible reasons. It is caused by the gas content of the coal seam in the mine reaching below the gas explosion safety line and the inaccurate scoring of the qualified rate of air volume at the wind-use site.

References
[1] Wang, W. (2005) Mine gas characteristics and risk assessment. J. Coal Mine Blasting, 3: 5-8.
[2] Sun, B. (2005) Application of Grey Relational Analysis to Evaluate Dangerous Sources of Gas Explosion Accidents. J. Coal Science and Technology, 12: 70-73.
[3] Wang, J.C., Zhang, F. (2017) Evaluation of Coal Mine Gas Explosion Based on Entropy Two-dimensional Cloud Model. J. Coal Mining Machinery, 9:166-168.
[4] Wang, W.Q., Lu, S.R., Zhang, Y.D. (2019) Analysis of coal mine gas explosion accident based on IAHP method. J. Coal Technology, 5:124-126.
[5] Zhao, Y.J., Zhang, E.L., Li, C. (2013) Analysis and prevention of coal mine gas explosion accident based on FTA. J. Western Exploration Engineering, 5: 187-190.