PAPER

Using random forest for the risk assessment of coal-floor water inrush in Panjiayao Coal Mine, northern China

Dekang Zhao[1,2] · Qiang Wu[1,2] · Fangpeng Cui[1,2] · Hua Xu[3] · Yifan Zeng[1,2] · Yufei Cao[4] · Yuanze Du[1,2]

Abstract

Coal-floor water-inrush incidents account for a large proportion of coal mine disasters in northern China, and accurate risk assessment is crucial for safe coal production. A novel and promising assessment model for water inrush is proposed based on random forest (RF), which is a powerful intelligent machine-learning algorithm. RF has considerable advantages, including high classification accuracy and the capability to evaluate the importance of variables; in particularly, it is robust in dealing with the complicated and non-linear problems inherent in risk assessment. In this study, the proposed model is applied to Panjiayao Coal Mine, northern China. Eight factors were selected as evaluation indices according to systematic analysis of the geological conditions and a field survey of the study area. Risk assessment maps were generated based on RF, and the probabilistic neural network (PNN) model was also used for risk assessment as a comparison. The results demonstrate that the two methods are consistent in the risk assessment of water inrush at the mine, and RF shows a better performance compared to PNN with an overall accuracy higher by 6.67%. It is concluded that RF is more practicable to assess the water-inrush risk than PNN. The presented method will be helpful in avoiding water inrush and also can be extended to various engineering applications.

Keywords Water inrush · Risk assessment · Mining · Random forest · China

Introduction

Mining-water inrush constitutes one of the major disasters in coal mine production. Compared with gas outbursts, coal dust explosions and other coal mine accidents, water-inrush incidents result in greater economic loss and the recovery to production is more challenging (Wu et al. 2016; Li et al. 2013; Sun et al. 2012). In North China, the near-surface coal has almost been fully exploited, so with increased mining depths, the water pressure in the underlying limestone aquifer keeps

increasing, and in some coal mines, the pressure is up to as much as 10 MPa or more (Wu et al. 2011; Sun et al. 2012; Li et al. 2015). According to statistics, more than 16 billion tons of coal resources are at risk of coal-floor water inrush and up to 60% of coal mines are threatened by the confined karst water to varying degrees (Meng et al. 2012; Wu et al. 2009). Furthermore, the hydrogeological conditions are becoming increasingly complicated; therefore, it is essential to accurately predict the risk of coal-floor water inrush (Shi et al. 2014; Sun et al. 2012; Goldscheider 2005).

In the past several decades, numerous methods and theories for assessing the risk of coal-floor water inrush have been developed. The most widely used method in China is the water–inrush coefficient method, which was proposed in 1964. The water-inrush coefficient is defined as the ratio between the thickness of the aquitard and the water pressure of the underlying aquifer (Wu and Zhou 2008; Shi et al. 2014). The scholars put forward an empirical formula related to the water-inrush coefficient by summarizing lots of measured water-inrush data from several large-scale mines (Meng et al. 2012; Wu et al. 2017; Shi et al. 2014); however, the method does not reflect the complicated water-inrush mechanism, as it only considers two controlling factors. In order to

✉ Qiang Wu
  wuq@cumtb.edu.cn

[1]   China University of Mining & Technology (Beijing),
     Beijing 100083, China

[2]   National Engineering Research Center of Coal Mine Water Hazard
     Controlling, Beijing 100083, China

[3]   Information Engineering College, Beijing Institute of Petrochemical
     Technology, Beijing 102617, China

[4]   Beijing Urban Construction Exploration and Surveying Design
     Research Institute Co. Ltd., Beijing 100101, China

describe the water-inrush process accurately, Wu and Zhou (2008) proposed a vulnerability index method, which incorporates a geographic information system (GIS) and other techniques such as the artificial neural network (ANN), analytic hierarchy process (AHP), and logistic regression analysis. Also, based on the water-inrush coefficient, Meng et al. (2012) discussed the relationship between water inrush and geological conditions including lithology and structure features. Meanwhile, several mathematical theories have gradually been applied to the risk assessment of water inrush, including (1) fuzzy mathematics theory (Wang et al. 2012), (2) attribute measurement theory (Li et al. 2015), (3) gray relational analysis (Li et al. 2015; Qiu et al. 2016; Shi et al. 2014), and (4) unascertained measure theory (Wu et al. 2017).

In recent years, with the rapid development of artificial intelligence technologies, the application of machine-learning algorithms, such as the decision tree (DT), support vector machine (SVM) and artificial neural network (ANN), to risk assessment has gradually become a trend (Harris 2013; Samuel et al. 2017; Pradhan 2013; Naghibi and Dashtpagerdi 2017). However, these methods have certain limitations in practical application—for example, numerous data pretreatment procedures are required for the DT model (Kubal et al. 2009), and it tends to fall into the local optimum; the SVM model has limitations because of its complex mathematical functions; and as for the ANN model, it has the shortcomings of over learning, slow convergence speed and local minimum value.

Considering all the aforementioned problems, this paper proposes a new risk assessment model of coal-floor water-inrush based on random forest (RF), which is a statistical-learning theory introduced by Breiman (2001). Random forest is a nonlinear modeling tool, coupling the main advantages of two major learning techniques: bagging and random feature selection. It has a self-adaptive property and just needs continuous training of sample information, and it is suitable for solving problems with unclear a priori knowledge and incomplete data. Compared with the traditional intelligence algorithms such as ANN and SVM, RF has high prediction accuracy and good tolerance to outliers and noise (Breiman 2001; Bonissone et al. 2010). Moreover, it has less computational cost and can avoid the problem of over-learning (Rodriguez-Galiano et al. 2012). Because of its superior performance, RF has been widely applied in biology, medicine, economics, management, remote sensing and other fields in recent years (Adam et al. 2014; Boulesteix et al. 2012; Eisavi et al. 2015; Hong et al. 2016; Pal 2005; Polishchuk et al. 2009).

In the current study, RF is used to construct a risk assessment model of coal-floor water inrush. The method consists of three main parts: construction of the comprehensive evaluation index system, selection of training and validation samples, and establishment of the RF model. For comparison, a probabilistic neural network (PNN) model was also constructed for risk assessment. PNN is a feed-forward neural network for classification and parallel estimation of probability density proposed by Specht (1990). It is widely used, as its structure is simple. Lastly, the models were applied to Panjiayao Coal Mine in northern China to verify the effectiveness of the proposed RF model.

Study area and data used

Study area

The Panjiayao Coal Mine is situated in the west of Datong Coalfield, approximately 50 km southwest of Datong City, Shanxi Province, in North China (Fig. 1). The mine elevation in the initial planning stage ranged from 1,399.9 to 1,606.0 m. The region has a temperate continental climate with a mean annual precipitation of 433.6 mm.

Topographically, according to drilling data, the main strata include: Ordovician (O), Carboniferous (C), Permian (P), Jurassic (J), Cretaceous (K), Neogene (E), and Quaternary (Q). The coal measure strata in the mining area include the Benxi group of the Middle Carboniferous system, the Taiyuan group of the Upper Carboniferous system, and the Shanxi group of the Lower Permian system. The total thickness of the Taiyuan group strata varies from 91.94 to 137.97 m. It contains 12 coal seams and the average total thickness is about 27.24 m. The main minable coal seams are Nos. 5 and 8, whereby the No. 5 coal seam, studied in this paper, is the thickest and the most stable coal seam in the mine area, with a mean thickness of 11.77 m and is found throughout the entire area except some sporadic areas.

Structurally, the coal seams strike NW–SE to the north of fault F73, and NE–SW to the south of it. The dip angle of the coal seams varies from 3 to 9°. Sixty faults and several collapse columns have been exposed by drilling in the study area; meanwhile, small folds are widely distributed in the study area. The overall structural complexity is moderate. Figure 2 shows a typical geological cross-section of the study area.

Hydrogeologically, the Panjiayao Coal Mine is situated in the runoff area of the Shentou karst spring region. The water level of the Ordovician limestone aquifer is about 1,164–1,180 m above sea level, whereas the elevation of the No. 5 coal floor is 580–810 m. Therefore, the No. 5 coal floor is under pressure in the whole area and the limestone aquifer of the Ordovician system is the main aquifer that threatens the safe mining of No. 5 coal seam (Fig. 2).

Data used

By detailed analysis of the geological conditions and the field-measured data of the Panjiayao Coal Mine, the main controlling factors for coal-floor water inrush were determined. The

Fig. 1 Location of the study area in Shanxi Province, China, and the geological structure of the Panjiayao Coal Mine
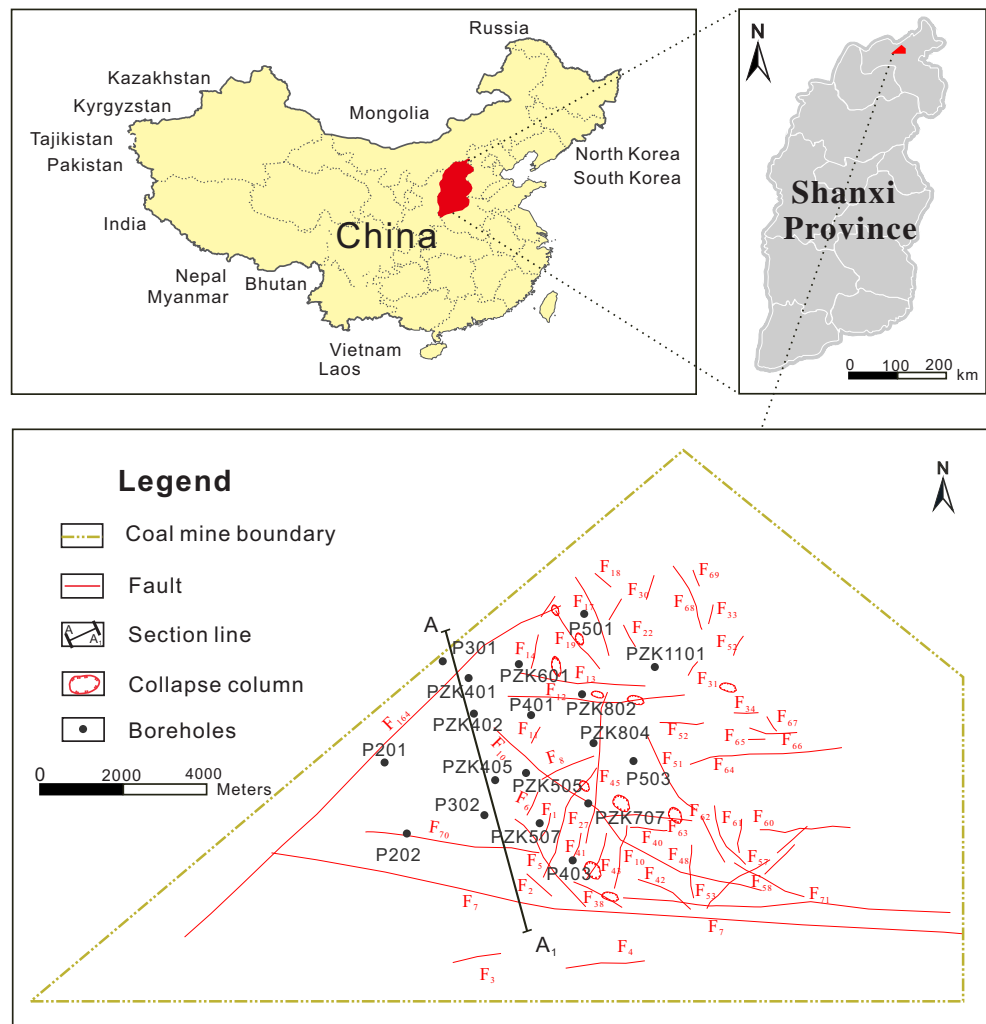

process of constructing the index system controlling water inrush is described in detail in previous works (Wu et al. 2008, 2017). A short overview of the eight main factors is given as follows, and eight thematic maps were obtained by interpolation of the measurement values of the eight factors (Fig. 3):

- *Water pressure (WP).* Water pressure is one of the most important factors resulting in water inrush from the coal floor. The greater the water pressure, the greater the possibility of water inrush. The water pressure is related to the water level and the elevation of the top interface of the Ordovician limestone. The water pressure acting on the No. 5 coal floor is between 2.4 and 5.7 MPa, and it gradually increases from the southeast to the northwest of the mining area (Fig. 3a).
- *Water-yield property (WYP).* The Ordovician karst aquifer underlying the coal seams is the main water source for water inrush. It is directly reflected by the unit inflow, which was calculated from pumping tests. The thematic

map of the water-yield property was generated by interpolation of the unit inflow value of several hydrogeological boreholes (Fig. 3b).
- *Aquitard effective thickness (AET).* The aquitard effective thickness plays an inhibitory role in coal-floor water inrush. The water resistance capacity of the aquitard is related to the thickness, strength and lithologic characteristics of the aquitard. As the aquitard is composed of several different lithological strata, the effective thickness of the aquitard was obtained by converting them into an equivalent effective thickness of sandstone (Coalfield Geological Central Bureau of China 2000; Wu et al. 2016). The aquitard effective thickness ranges from 27 to 67.5 m, with an average value of 46 m (Fig. 3c).
- *Brittle-rock thickness (BRT).* The brittle rock (such as sandstone and limestone) in the aquitard underlying the coal seams plays a key role in the resistance of the water inrush. According to the exploration data, the brittle-rock thickness in the pressure-damaged zone of the No. 5 coal floor varies from 6.6 to 38.2 m (Fig. 3d).
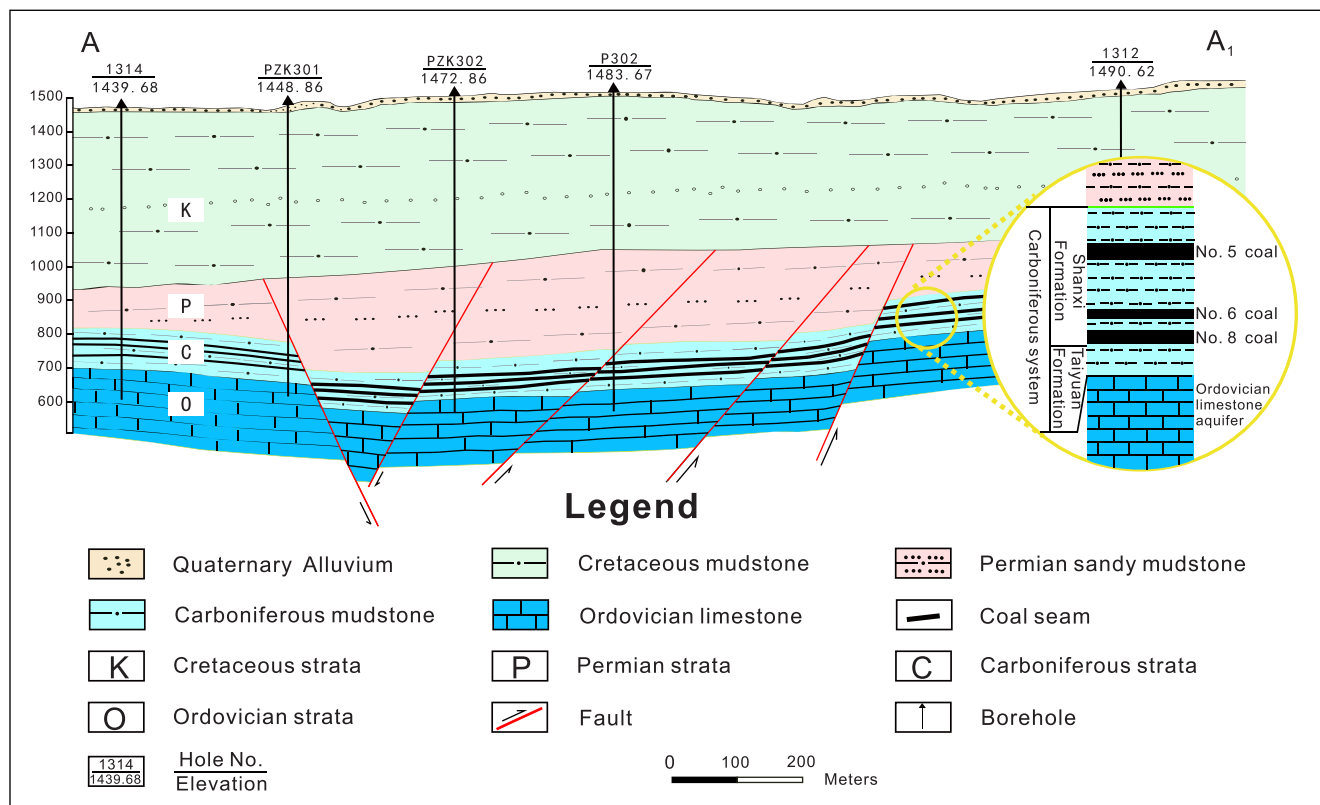
**Fig. 2** Typical geological cross-section (A–A₁, see Fig. 1) of the study area

- *Distribution of faults and folds (DFF).* According to the distance from the fault-lines and fold-axes, the zones around the structures are divided into two parts: affected zones and fractured zones. The fractured zones that are affected by faults and folds are the weak zones which may lead to direct hydraulic connections between the coal seams and the underlying aquifers. The thematic map of the distribution of faults and folds (Fig. 3e) was generated by quantification of the fault-lines and fold-axes distribution as shown in the geological structure map. The fractured zones and affected zones around a structure are quantified with the values of 1.0 and 0.7, respectively, according to the degree of fragmentation.

- *Distribution of collapse columns (DCC).* Collapse columns are the main water-inrush passages in karst areas. As the geological structure map (Fig. 1) shows, several collapse columns were explored in the study area, and they were quantified according to the degree of fragmentation in the impact zones and their hydraulic conductivity (Fig. 3f). The buffer zone, which is an irregular concentric fracture zone around the column, is quantified as 0.8, and the collapse columns as 1.0.

- *Distribution of structural intersections and endpoints (DSIE).* As the intersections and endpoints within structures tend to cause stress concentration and damage the rock mass more easily, hydraulic conductivity in these

areas is apparently greater than the surrounding areas. Additionally, these intersections and endpoints are most likely to lead to water inrush under the influence of mining. They were quantified according to the quantified values of DFF and the fragmentation degree of the rock mass. The thematic map was generated as Fig. 3g.

- *Fault-scale index (FSI).* The fault-scale index is related to the length and throw of the faults, and it refers to the sum of all faults' throw and length per unit area calculated based on the geological structure map (Fig. 1; Wu et al. 2017). It represents the scale and development degree of the faults. Based on the exploration data, the fault-scale index was calculated and interpolated for the whole study area, as in Fig. 3h. Its calculation formula is as follows:

$$F = \frac{\sum\limits_{i=1}^{n} L_i H_i}{S} \tag{1}$$

where F is the fault-scale index, S is the area of the unit, H_i is the fault throw of the i-th fault, L_i is the length of the i-th fault per unit area.
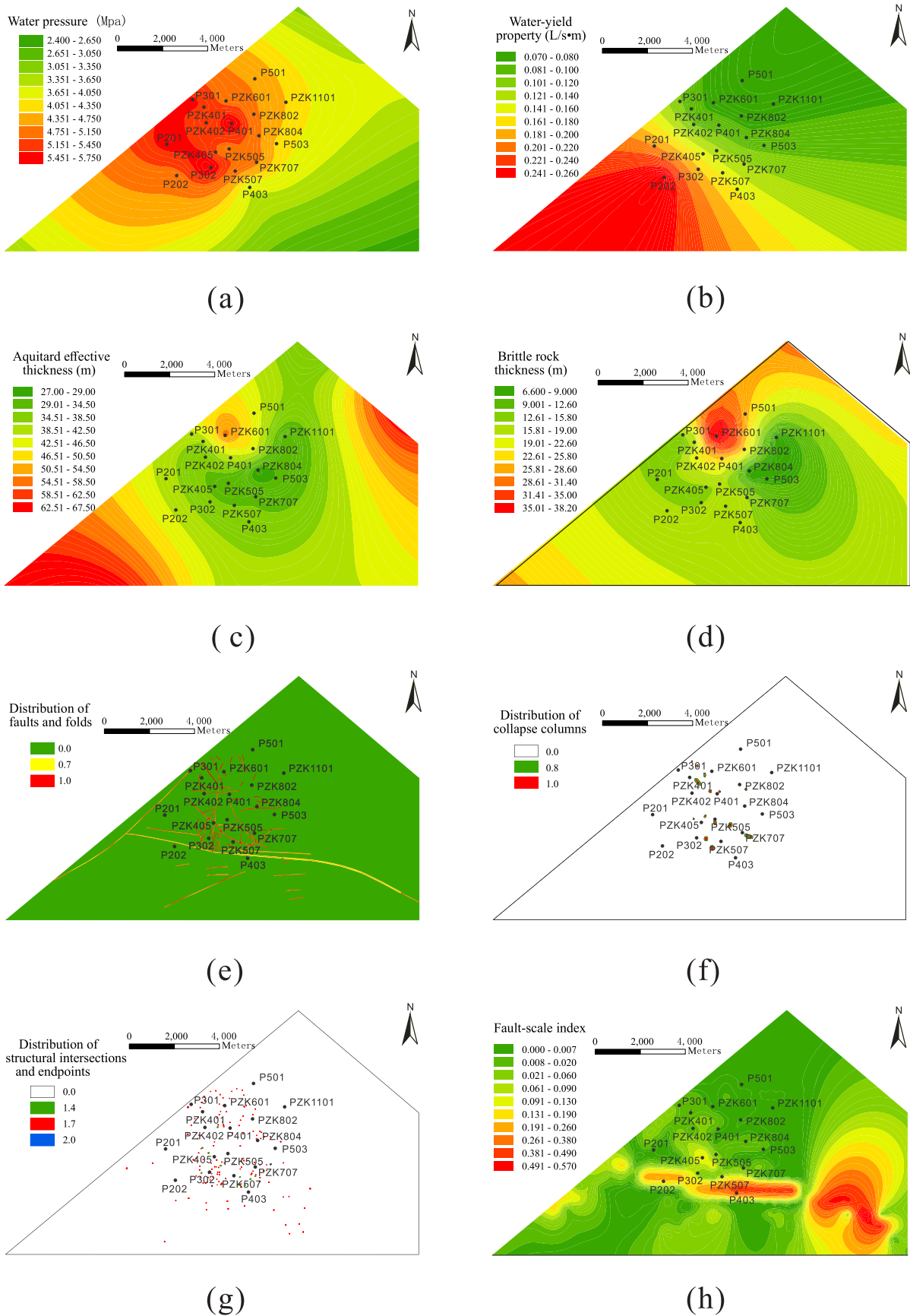
Fig. 3 Thematic map of the eight controlling factors: **a** WP, **b** WYP, **c** AET, **d** BRT, **e** DFF, **f** DCC, **g** DSIE, and **h** FSI

Materials and methodology

Random forest

Random forest is a statistical-learning theory introduced by Breiman ([2001](#)), and it is an ensemble-learning algorithm which is more accurate and robust to noise compared to the single classifiers (Breiman 1996; Dietterich 2000; Tatsumi et al. [2015](#); Krkač et al. [2016](#)). The RF presents some strengths for its application in water-inrush risk assessment, compared with the traditional intelligence algorithms:

- The learning process is very fast.
- For many kinds of data, it produces classifiers with high accuracy.
- It can evaluate the importance of variables when classifying.
- When building a forest, it can internally estimate the generalization error.
- It can estimate the missing data. Even if a large portion of the data is lost, it can still maintain high accuracy.
- For imbalanced classification data sets, it can balance the error.

Random forest is an ensemble of numerous classification and regression trees (CART) to classify or predict the value of a variable (Breiman [2001](#); Rodriguez-Galiano et al. [2012](#); Wang et al. [2012](#)). The CART algorithm was proposed by Breiman et al. ([1984](#)) and has the advantage of handling both numerical and categorical variables. The tree building and pruning procedures of CART are as follows (Death and Fabricius [2000](#); Loh [2011](#)). Firstly, the attributes of the root node are determined by the Gini index method. The Gini index I_G is used to measure the impurity of a given element, and it is the measure for the best split selection. For the dataset, D, I_G can be expressed as:

$$I_G(D) = 1 \sum_{i=1}^{N} p_i^2 \qquad (2)$$

where p_i is the probability of the samples belonging to the i-th leaf and N is the number of the leaves; the lower the Gini index I_G, the purer the samples. Next, in order to split the node into two leaf nodes, one attribute or the combination of several attributes is chosen from the multiple predictive attributes as the split variable. This process is repeated for each new leaf node until a sufficiently large classification tree is built. The complexity of the tree is reflected by the number of terminal nodes in the tree. Lastly, the tree is pruned by deleting the redundant variables to generate a simpler tree, as this is easier to understand and calculate.

The RF model contains two important parameters: the number of classification trees (k), and the number of random variables at each split node (m). The parameters must be optimized to guarantee the least error in data processing. The RF model can increase the diversity of the classification trees in two ways: sampling with replacement, and changing the combination of predictive variables randomly. Each classification tree grows on the basis of a sampling subset (X_i) of the initial data set (X), and the optimal attributes in m attributes are randomly chosen to segment the nodes.

The specific implementation procedures of the RF algorithm are as follows:

1. Draw k samples randomly from the original training set X using the bootstrap resampling method.
2. For k bootstrap samples, k unpruned classification trees are created respectively. In the tree growing process, for each node, m attributes are randomly selected from the total M attributes as internal nodes ($m < M$). Then, according to the minimum Gini index principle, an optimal attribute is selected from m attributes as a split variable to make the branches grow. Thus, each tree will grow without pruning so that the purity of each node is minimal, and k classification results are obtained.
3. According to the k classification results, each decision tree is voted by a simple majority voting method to determine its final classification results.

A schematic diagram of the RF structure is shown in Fig. [4](#).

Out-of-bag error estimate

In an RF model, it is not necessary to use cross validation to establish the unbiased estimate of the errors because the RF estimates the errors during modeling using the out-of-bag error estimate. There are n samples in the original training set X. In the process of sampling from X using the bootstrap resampling method, the probability that each sample would not be drawn is $p = (1 - 1/n)^n$. If n tends to infinity, $p \approx 0.37$, which indicates that about 37% of the samples in the original training set X are not drawn, and these data are called out-of-bag (OOB) data. In the process of building a classification tree, these data are not used. After the classification tree is built, the OOB data are classified. During the operation of the model, each variable of the original data set becomes the OOB data in the creation of $k/3$ classification trees. In the RF model, each decision tree can obtain an OOB error estimate value, and the average value is used as a generalization error estimation to estimate the classification performance of the model (referred to as OOB error estimate). The OOB error is calculated according to the following procedures. Assume that the total number of the OOB data is M, and these data are then used to test the performance of the generated RF. The M OOB data are input to the RF classifier, and then the classification results of the M data can be obtained. As the classification results of the M data have been determined already, the number of the
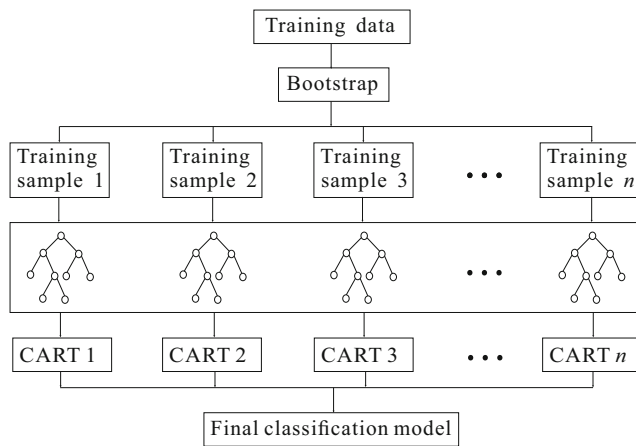
Fig. 4 Principle and flowchart of RF

data with wrong classification results are counted as m. Then the OOB error $e = m/M$. The e has been proved to be unbiased, so it is not necessary to use the cross validation in the RF.

Importance degree measurement

The RF model provides two ways to calculate the degree of importance of each variable index: a mean decrease in the Gini index, and an OOB mean decrease in accuracy. A mean decrease in Gini index means a total impurity decrease of each variable at each tree node and is a measure of the contribution of each variable to the homogeneity of the nodes and leaves (Hong et al. 2016; Menze et al. 2009). The method evaluates the importance of the variables by calculating the Gini index based on Eq. (2), and then accumulates the total impurity decrease of all the trees.

The basic principle of the OOB error estimate method is: when noise is added to a related feature which plays an important role in the accuracy, the prediction accuracy of the RF will decrease significantly. The main procedures are as follows (Ulrike 2009): firstly, for the generated RF, the OOB error e_t of each decision tree is calculated according to the OOB data; secondly, the j-th eigenvalue X^j of the OOB data is changed randomly (namely, the noise interference is added artificially); thirdly, the OOB data with noise is used to test the performance of the RF, and a new OOB error e_t^j is obtained; finally, the importance degree of the variable X^j can be calculated according to Eq. (3):

$$I\left(X^j\right) = \frac{1}{n} \sum_{t=1}^{n} \left(e_t^j - e_t\right) \tag{3}$$

where X^j is the j-th eigenvalue of the OOB data; e_t is the initial OOB error; e_t^j is the OOB error with noise; n is the number of the decision trees; $I(X^j)$ is the importance of the variable X^j. The greater the OOB error caused by the change of the variable X^j, the more the decrease in accuracy, indicating the more important the variable is.

Results and discussion

### Training and validation samples

Figure 5 shows the complete process of RF and the detailed flowchart of the proposed methodology in this study. Before the quantitative model is applied to predict the coal-floor water inrush, it is necessary to select the training sample data for the establishment of the model.

In the past research on machine-learning models, the training sample data only considered the positive samples (water-inrush points data), but ignored the negative samples (non-water-inrush points data) and transitional samples in the prediction of the coal-floor water inrush. However, because of the limited number of measured water-inrush points, the machine-learning models based on these training data are usually less accurate. If only water-inrush samples and non-water-inrush samples are being applied for the model training, the results from the generated machine-learning models cannot accurately reflect the different degrees of risk in the study area. Therefore, in order to reflect the mechanism of water inrush realistically and to subdivide the water-inrush risk accurately, samples in different water-inrush risk degrees are necessary for the training of the machine-learning models.

In this study, firstly, 41 field-measured water-inrush samples were collected from several large-scale coal mines in North China, and the risk grades were quantified according to the amount of water inflow. Secondly, the aforementioned eight evaluation factors were graded based on the risk degree of coal-floor water inrush. The risk-grade classification system is shown as Table 1, which was constructed combining the study area's practical situation and the international research achievements. Simultaneously, the K-means clustering analysis method was used to determine the threshold values of the controlling factors (Wu et al. 2009, 2013, 2017; Wang et al. 2012, 2015). The water-inrush risk of the study area was classified into five grades (safe, relatively safe, transitional, relatively dangerous and dangerous). Finally, in order to increase the number of samples and improve the model accuracy, based on the classification standard of the evaluation factors (Table 1), 20 groups of data were generated in each grade threshold interval of each factor by using the method of uniform-distribution random generation, and a group of 100 ideal samples is obtained for the five grades. Furthermore, a synthetic minority oversampling technique (SMOTE) is introduced to synthesize the data.

SMOTE is a widely used oversampling method proposed by Chawla et al. (2002). In the SMOTE, the class is oversampled according to the similarities in the feature space between the existing sample data using the K-nearest neighbor algorithm, which will avoid the overfitting problem. The specific operation procedure is as follows:

Fig. 5 Detailed flowchart of the proposed methodology for water-inrush risk assessment
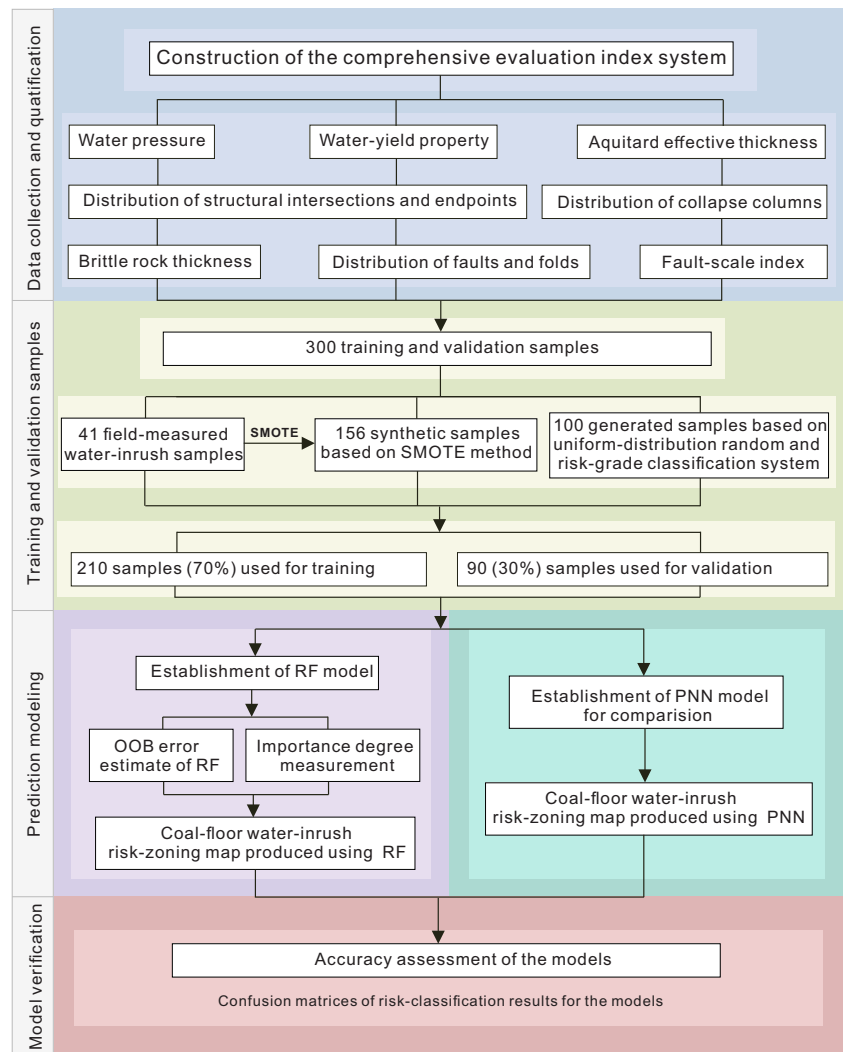
Step 1. Randomly choose two samples in the same class, x_1 and x_2.

Step 2. Calculate the difference of the i-th attribute between x_1 and x_2, namely, $\delta = x_{2i} - x_{1i}$.

Step 3. Calculate the i-th attribute value of the new target sample based on Eq. (4):

$$x_{12i} = x_{1i} + \text{rand}[0, 1] \cdot \delta_i \tag{4}$$

Table 1 Grading of the main controlling factors related to the risk of coal-floor water inrush

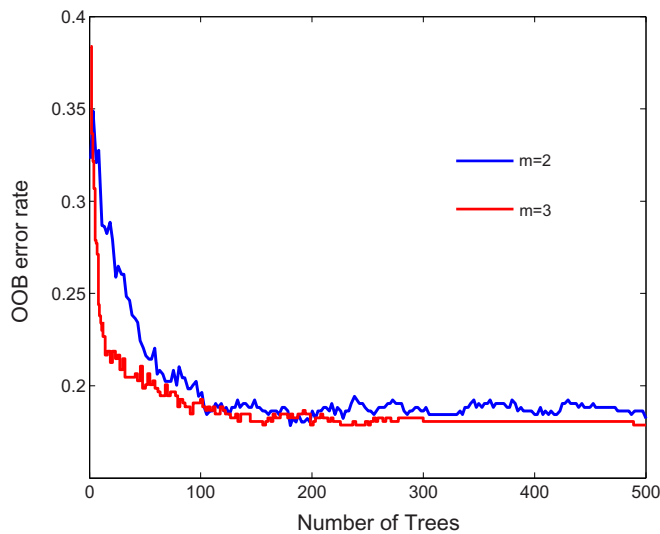| Main controlling factor	Risk grade				
	Safe	Relatively safe	Transitional	Relatively dangerous	Dangerous
WP (MPa)	<1	1–2	2–3	3–4	>4
WYP (L s^{-1} m^{-1})	<0.1	0.1–0.5	0.5–1	1–5	>5
AET (m)	>80	80–60	60–40	40–20	<20
BRT (m)	>15	15–10	10–5	5–0.5	<0.5
DFF	<0.15	0.15–0.4	0.4–0.5	0.5–0.6	>0.6
DCC	<0.2	0.2–0.5	0.5–0.7	0.7–0.9	>0.9
DSIE	<0.5	0.5–1	1–1.2	1.2–1.4	>1.4
FSI	<0.02	0.02–0.05	0.05–0.1	0.1–0.2	>0.2

Fig. 6 The OOB error rate of the RF model

Step 4.   The final synthetic sample generated using x_1 and x_2 can be calculated based on

$$x_{12} = x_1 + \text{rand}[0, 1] \cdot \delta \tag{5}$$

(where) $\delta = (\delta_1, \delta_2, \cdots, \delta_n)$

Step 5.   Repeat the above process for N times, then N synthetic samples are obtained.

According to the SMOTE method, 159 synthetic samples were obtained based on the 41 field-measured water-inrush samples; thus, together with the group of 100 ideal samples generated (mentioned previously), 300 samples were obtained in total, with 210 samples (70%) used for training and 90 (30%) used for validation.
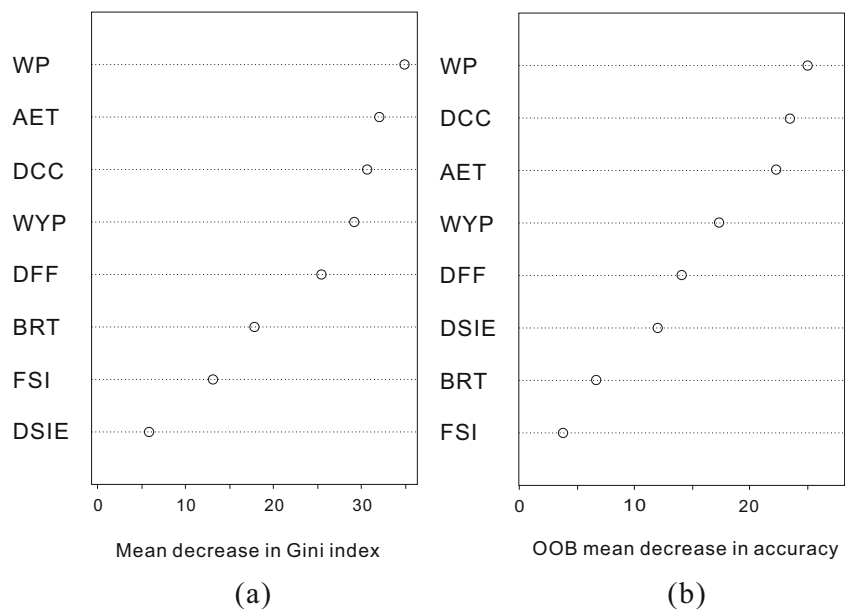
Establishment of the RF model

In the RF, two parameters are required to be defined: the number of trees in the forest (k), and the number of random variables of the split nodes (m). To maximize the model accuracy, it is necessary to optimize the combination of the parameters m and k. When k is defined with a small value, the RF classification error is uncontrollable and the model performance cannot achieve the optimal identification. Conversely, if the parameter k is too large, the computation time and required memory will increase accordingly. By repeated operation, it is found that when $k = 100$, the OOB error of each classification tends to be stable and the model does not tend to over fitting. According to Breiman (2001), $m = \text{sqrt}(M)$. In this case study, there are eight variables, namely $M = 8$. To assess the optimal value of m, two RF models were created for $m = 2$ and $m = 3$ (Fig. 6). Figure 6 shows the change of the OOB error rate depending on the number of trees k. The results show that when $k = 100$ the error rate of the model is stable, and when $m = 3$, the error rate is lower with the value of approximately 0.14. This suggests that when the RF model is applied to the risk prediction of the No. 5 coal-floor water inrush, it will have a result with 86% accuracy, indicating that the RF is a reasonably good model. Therefore, m with value of 3, and k with a value of 100, were selected as input parameters for the RF model.

The contribution of each variable to the generated RF model is shown in Fig. 7. As shown, the importance degree of each variable is measured two ways: mean decrease in Gini index and OOB mean decrease in accuracy. According to the Gini index, water pressure (WP) has the highest importance, followed by AET, DCC, WYP and DFF, while the BRT, FSI and DSIE have the lowest importance. Regarding the OOB mean decrease in accuracy, the most important variables are

Fig. 7 a Mean decrease in Gini index and **b** OOB mean decrease in accuracy (sorted decreasingly from top to bottom) of the RF model

Fig. 8 Risk-zoning maps produced using **a** RF and **b** PNN
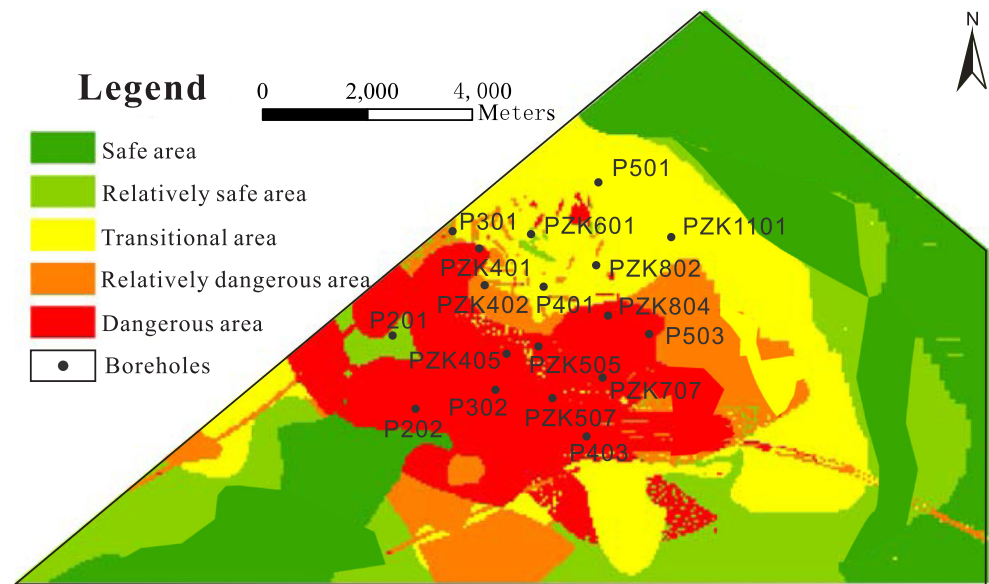
(a)
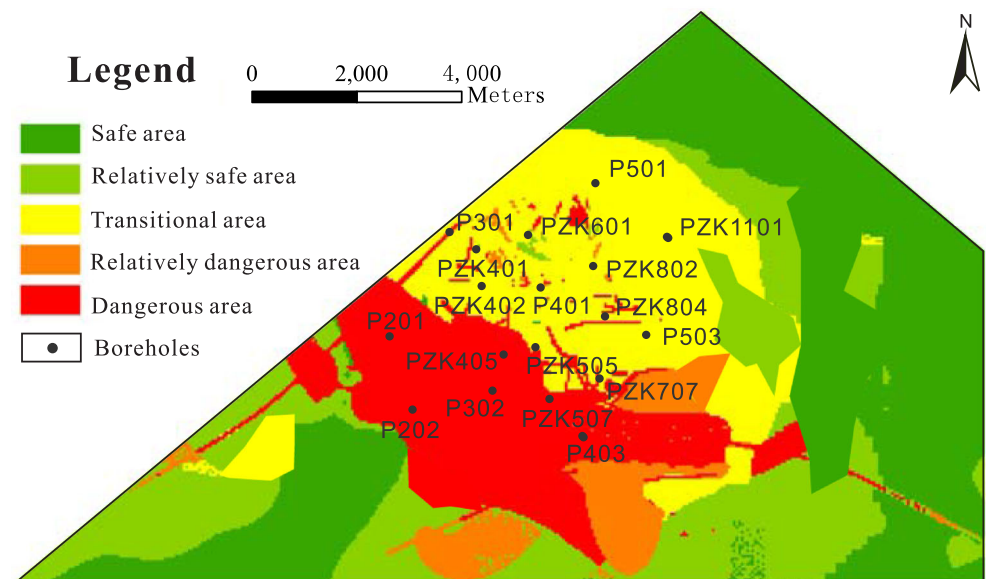
(b)

WP and DCC, and the BRT and FSI are the least important variables in water inrush. Based on both of the features of importance, in the water-inrush risk-index system, WP, DCC and AET are the three most important variables, indicating that they contribute overwhelmingly to the water-inrush risk, and this is in accordance with the past research (Wu et al. 2008, 2011, 2014; Zeng et al. 2016). Even so, all of the eight factors were used to generate the RF model.

Risk classification results

In order to evaluate the risk within the study area and to reflect the spatial difference more exactly, the study area was divided into square grids with a fixed size set to 30 m × 30 m., whereby the study area (117.1 km^2) was divided into 130,119 grids. The RF model, which was constructed based on all the sample data, was applied to the whole study area, making then

Table 2 Risk grading areas for the RF and PNN models (%)

Model	Risk grades				
	Safe	Relatively safe	Transitional	Relatively dangerous	Dangerous
RF	23.1	32.3	12.8	19.5	12.3
PNN	20.9	31.8	14.3	21.8	11.2

possible to calculate the risk grade of each grid (see Fig. 8a for the GIS enabled water-inrush-risk-zoning map of the No. 5 coal floor).

As seen in Fig. 8a, in the study area, the risk of No. 5 coal-floor water-inrush shows a transitional trend from east to west. The entire risk trend gradually increases from the periphery to the center of the study area. The relatively dangerous and dangerous risk zones are mostly located in the areas where the faults, folds and collapse columns are developed. Furthermore, these two risk zones occupy approximately 28.1% of the total area.

A total of 59,358 grids (in green), occupying approximately 53.38 km^2, are classified as safe and relatively safe areas. These areas are mainly distributed in the east and partly west of the mine, where the water abundance of the aquifer is low and the geological conditions are relatively simple. For these areas, researchers need to strengthen the continuous observation of water level, water quantity and water pressure in the hydrogeological boreholes. The relatively dangerous zones (in orange) are mostly in the center and southwest of the mine surrounded by the transitional zones (in yellow). In these areas, the water pressure acting on the aquitard is between 4.5 and 5 MPa; the aquitard effective thickness is of an average low value (30 m), as is the brittle-rock thickness in the mining pressure fractionated zone (15–20 m); several small-scale structures exist in these areas. For the transitional and relatively dangerous zones, focusing the investigation on the geological structures is necessary, especially the zones of the structural intersections and endpoints, guaranteeing that there are no water-inrush precursors in these zones. The dangerous

zones are mainly located in the center of the mine around the boreholes P302, PZK507 and PZK505. According to the exploration data and the water pressure thematic map, the water pressures of these three boreholes are the largest in the whole study area. In these dangerous zones: the water pressures are all above 5 MPa; the geological structures are more developed; the water yield of the Ordovician limestone is high; some large-scale faults (such as faults F73 and F164) have large displacements; and most of the collapse columns are exposed. For the dangerous zones, it will be necessary to depressurize the aquifer by placing large-diameter dewatering wells in the subsurface and enhance the water-resistance capacity of the aquitard.

Comparison with the PNN model

For comparison, the PNN model was also used for the risk assessment. PNN is a feed-forward neural network proposed by Specht in 1990. It is a neural network for classification and parallel estimation of probability density (Specht 1990, 2002). It adopts the estimation method of union probability-density distribution and a Bayesian optimization rule based on the Gauss function (Iounousse et al. 2015; Singh et al. 2013). The main idea of PNN is to separate the decision space in the multidimensional input space by using the Bayes decision rule (Iounousse et al. 2015; Tabatabaei 2016). It is a widely used artificial neural network, and it is simple in structure and easy in training; therefore, PNN not only has the superior characteristics of general neural networks, but also learns quickly.

In this study, the PNN model was constructed using the same training data aforementioned. The radial basis function was adopted to generate a cycle-training algorithm program in the model because of its simple structure, fast convergence, and the ability of approximating any nonlinear function. After repeated operation, the PNN model achieved optimal performance when the distribution density of the radial basis function and the spread value were determined as 0.1 and 0.5. The water-inrush risk assessment map based on PNN is shown in Fig. 8b.

Table 3 Confusion matrix of the RF model (model overall accuracy 88.89%)

Risk grades	Safe	Relatively safe	Transitional	Relatively dangerous	Dangerous	Row total	User's accuracy
Safe	12	1	0	0	0	13	92.31%
Relatively safe	1	13	1	0	0	15	86.67%
Transitional	0	1	20	2	0	23	86.96%
Relatively dangerous	0	0	1	19	1	21	90.48%
Dangerous	0	0	0	2	16	18	88.89%
Column total	13	15	22	23	17	90	–
Producer's accuracy	92.31%	86.67%	90.91%	82.61%	94.12%	–	–

Table 4 Confusion matrix of the PNN model (model overall accuracy 82.22%)

Risk grades	Safe	Relatively safe	Transitional	Relatively dangerous	Dangerous	Row total	User's accuracy
Safe	10	1	0	0	0	11	90.91%
Relatively safe	2	12	1	1	0	16	75.00%
Transitional	1	2	19	1	1	24	79.17%
Relatively dangerous	0	0	2	19	2	23	82.61%
Dangerous	0	0	0	2	14	16	87.50%
Column total	13	15	22	23	17	90	–
Producer's accuracy	76.92%	80.00%	86.36%	82.61%	82.35%	–	–

By comparison of the two assessment results in Fig. 8, it could easily be found that the two model results are practically consistent, with the exception of some small differences in certain local regions. The correlation coefficient of the two zoning maps is 0.84, which indicates that the results from the two models are similar in the majority of the areas, and both of the models are reasonable for the water inrush assessment.

The percentages of grading area for the RF and PNN models are shown in Table 2. The main differences between the two results are mainly concentrated in the fourth and fifth grade (dangerous and relatively dangerous areas). For the RF model, the relatively dangerous and dangerous areas are 12.3 and 23.1%, respectively, while for the PNN model, they are 11.2 and 20.9%, respectively. This indicates that more areas were classified as relatively dangerous and dangerous in the RF model, especially in the areas where the geological structures are relatively developed.

In fact, according to the detailed analysis of the geological structure map and the risk zoning map by the RF, it is found that the areas in which there are fault lines, structural endpoints and endpoints located, are most likely to be classified as dangerous, and that the RF model is effective in describing the effect of these structures on the coal-floor water inrush. For example, the areas around boreholes P503 and PZK804 are classified as dangerous in the RF model, while they are classified as transitional grade in the PNN model. Just as shown in the geological structures map (Fig. 1) and structural thematic maps (Figs. 3a–h), the structures in these areas are highly

developed; however, what is worse, is that the water pressure is up to 5.2 MPa, while in these areas the aquitard effective thickness (30 m) and the brittle-rock thickness (20 m) are the lowest of the whole study area. Because there tends to be water inrush in these areas, the risk grade should be classified as dangerous, just as the RF model has shown. This demonstrates that besides the water pressure and aquitard thickness, the geological structures also have a great influence on the occurrence of coal-floor water inrush, as is consistent with previous conclusions (Wu et al. 2013, 2017). Compared with the PNN model, the RF model is more applicable for areas with developed structures, especially when considering point and linear elements, such as faults lines and the structural intersections and endpoints.

Accuracy assessment

For machine-learning models, model validation is necessary to evaluate the accuracy of the generated model. The RF can internally estimate the generalization error using OOB error rate, as shown in Fig. 6. The OOB error estimate is as accurate as the estimation based on the test subset, while it should be pointed out that compared with the test estimation used in traditional machine-learning models, the OOB error estimate is unbiased (Breiman 2001; Rodriguez-Galiano et al. 2012). So for the RF model, it is unnecessary to use an independent test data set to do a test estimation; nevertheless, in this study, in order to compare the accuracy of the RF and PNN models, the confusion matrices of risk-classification results for the two

Table 5 Confusion matrix of the actual field-measured samples (model overall accuracy 81.25%)

Risk grades	Safe	Relatively safe	Transitional	Relatively dangerous	Dangerous	Row total	User's accuracy
Safe	3	0	0	0	0	3	100%
Relatively safe	0	1	0	0	0	1	100%
Transitional	0	1	2	0	1	4	50%
Relatively dangerous	0	0	1	1	1	3	33.33%
Dangerous	0	0	0	1	4	5	80%
Column total	3	2	3	2	6	16	–
Producer's accuracy	100%	50%	66.67%	50%	66.67%	–	–

models are presented using the independent test data set as Tables 3 and 4, respectively. The confusion matrix contains the information about actual and predicted classification done by a classification system. It has two dimensions: one dimension (in rows) is indexed by the actual class of the risk grade, and the other dimension (in columns) is indexed by the risk grade that the generated model predicts.

Table 3 shows that the overall accuracy of the RF model is up to 88.89%, and this is consistent with the OOB error rate of 86% as shown in Fig. 6. On the other hand, the PNN model performed with a relatively lower overall accuracy of 82.22% (Table 4). Both RF and PNN classification effects are satisfactory, and the prediction accuracy of the RF is 6.67% higher than the PNN model. Generally, both of them are applicable to evaluate the risk of coal-floor water inrush. There are 16 field-measured samples out of the 90 validation samples. To verify the effectiveness of the RF model more intuitively, the confusion matrix of the actual 16 field-measured samples is shown as Table 5. The results show that the overall accuracy of the actual section of the test set is up to 81.25%. It is also demonstrated that the proposed RF model has a satisfactory accuracy, and the synthetic samples generated by both the SMOTE method and uniform-distribution random method are relatively reasonable.

Conclusion

Coal-mine water inrush is a prominent problem threatening coal mine safety. This study proposes a risk assessment model of coal-floor water inrush based on RF. For the water-inrush evaluation, there are many objective problems such as the lack of data or samples and the unbalanced classification of data. Random forest is a robust machine-learning method with numerous advantages, especially its high classification accuracy, and it is well suited to problems with unclear prior knowledge and incomplete data. With regard to the water-inrush problem, the RF model has finally been demonstrated to be able to maintain a high degree of accuracy.

The RF model was applied to Panjiayao Coal Mine, North China. The eight main water-inrush controlling factors were selected and 300 samples were created for the RF model training and validation. The water pressure and aquitard effective thickness are the top two most important out of the eight variables, according to the importance degree measurement by the mean decrease in the Gini index and the OOB mean decrease in accuracy. The results show that the overall risk in the study area is relatively high, and the entire risk trend gradually increases from the periphery to the center of the study area. For comparison, the PNN model was used for the risk assessment based on the same training samples and validation samples. Confusion matrices were used to compare the performance of the RF and PNN models, and the overall accuracy of the RF and PNN models were 88.89 and 82.22%, respectively. Subsequently, this study analyzed the local differences of the two results in detail. The results demonstrate that the RF has a better prediction ability than the PNN. This study shows the potential of a novel approach to water-inrush risk assessment. Evaluation results provide a reference for coal-floor water-inrush risk management, prevention, and reduction in the study area.

References

Adam E, Mutanga O, Odindi J, Abdelrahman EM (2014) Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers. Int J Remote Sens 35(10):3440–3458

Bonissone P, Cadenas JM, Garrido MC, Díaz-Valladares RA (2010) A fuzzy random forest. Int J Approx Reason 51(7):729–747

Boulesteix A, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mining Knowl Discov 2(6):493–507

Breiman L (2001) Random forest. Mach Learn 45:5–32

Breiman L, Friedman JH, Olsen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont, CA

Chawla NV, Bowyer KW, Hall HO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16(1):321–357

Coalfield Geological Central Bureau of China (2000) Coalfield hydrogeology of China. Coal Industry Publishing House of China, Beijing, pp 15–59

Death G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for the analysis of complex ecological data. Ecology 81(11):3178–3192

Eisavi V, Homayouni S, Yazdi AM, Alimohammadi A (2015) Land cover mapping based on random forest classification of multitemporal spectral and thermal images. Environ Monit Assess 187(5):291

Goldscheider N (2005) Karst groundwater vulnerability mapping: application of a new method in the Swabian Alb, Germany. Hydrogeol J 13(4):555–564

Harris T (2013) Quantitative credit risk assessment using support vector machines: broad versus narrow default definitions. Expert Syst Appl 40(11):4404–4413

Hong HY, Pourghasemi HR, Pourtaghi ZS (2016) Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology 259:105–118

Iounousse J, Er-Raki S, El Motassadeq A, Chehouani H (2015) Using an unsupervised approach of Probabilistic Neural Network (PNN) for land use classification from multitemporal satellite images. Appl Soft Comput 30(C):1–13

Krkač M, Špoljarić D, Bernat S, Arbanas SM (2016) Method for prediction of landslide movements based on random forests. Landslides 14(3):947–960

Kubal C, Haase D, Meyer V, Scheuer S (2009) Integrated urban flood risk assessment: adapting a multicriteria approach to a city. Nat Hazards Earth Syst Sci 9(6):1881–1895

Li LP, Zhou ZQ, Li SC, Xue YG, Xu ZH (2015) An attribute synthetic evaluation system for risk assessment of floor water inrush in coal mines. Mine Water Environ 34(3):288–294

Li T, Mei TT, Sun XH, Lv YG, Sheng JQ, Cai M (2013) A study on a water-inrush incident at Laohutai coalmine. Int J Rock Mech Mining Sci 59(5):151–115

Loh WY (2011) Classification and regression trees. WIREs Data Mining Knowl Discov 1(1):14–23

Meng ZP, Li G, Xie X (2012) A geological assessment method of floor water inrush risk and its application. Eng Geol 143–144:51–60

Menze BH, Kelm MB, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprech FA (2009) A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinform 10(1):1–16

Naghibi SA, Dashtpagerdi MM (2017) Evaluation of four supervised learning methods for groundwater spring potential mapping in Khalkhal region (Iran) using GIS-based features. Hydrogeol J 25(1):169–189

Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222

Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'Min VE (2009) Application of random forest approach to QSAR prediction of aquatic toxicity. J Chem Inform Model 49(11):2481–2488

Pradhan B (2013) A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput Geosci 51(2):350–365

Qiu M, Shi LQ, Teng C, Zhou Y (2016) Assessment of water inrush risk using the fuzzy delphi analytic hierarchy process and grey relational analysis in the Liangzhuang Coal Mine, China. Mine Water Environ. https://doi.org/10.1007/s10230-016-0391-7

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. Isprs J Photogram Remote Sens 67(1):93–104

Samuel OW, Asogbon GM, Sangaiah AK, Peng F, Li G (2017) An integrated decision support system based on ANN and fuzzy-AHP for heart failure risk prediction. Expert Syst Appl 68:163–172

Shi LQ, Qiu M, Wei WX, Xu DJ, Han J (2014) Water inrush evaluation of coal seam floor by integrating the water inrush coefficient and the information of water abundance. Int J Min Sci Technol 24(5):677–681

Singh KP, Gupta S, Rai P (2013) Predicting carcinogenicity of diverse chemicals using probabilistic neural network modeling approaches. Toxicol Appl Pharmacol 272(2):465–475

Specht DF (1990) Probabilistic neural networks. Neural Netw 3(1):109–118

Specht DF (2002) Probabilistic neural networks and the polynomial Adaline as complementary techniques for classification. IEEE Trans Neural Netw 1(1):111–121

Sun WJ, Wu Q, Dong DL, Jiao J (2012) Avoiding coal–water conflicts during the development of China's large coal-producing regions. Mine Water Environ 31(1):74–78

Tatsumi K, Yamashiki Y, Torres MAC, Taipe CLR (2015) Crop classification of upland fields using random forest of time-series Landsat 7 ETM+ data. Comput Electron Agric 115:171–179

Tabatabaei S (2016) A probabilistic neural network based approach for predicting the output power of wind turbines. JETAI 29(2). https://doi.org/10.1080/0952813X.2015.1132272

Ulrike G (2009) Variable importance assessment in regression: linear regression versus random forest. Am Stat 63(4):308–319

Wang Y, Yang WF, Li M, Liu X (2012) Risk assessment of floor water inrush in coal mines based on secondary fuzzy comprehensive evaluation. Int J Rock Mech Mining Sci 52(6):50–55

Wang ZL, Lai CG, Chen XH, Yang B, Zhao SW, Bai XY (2015) Flood hazard risk assessment model based on random forest. J Hydrol 527:1130–1141

Wu Q, Zhou WF (2008) Prediction of groundwater inrush into coal mines from aquifers underlying the coal seams in China: vulnerability index method and its construction. Environ Geol 56(2):245–254

Wu Q, Zhou WF, Wang JH, Xie SH (2009) Prediction of groundwater inrush into coal mines from aquifers underlying the coal seams in China: application of vulnerability index method to Zhangcun Coal Mine, China. Environ Geol 57(5):1187–1195

Wu Q, Liu YZ, Liu DH, Zhou WF (2011) Prediction of floor water inrush: the application of GIS-based AHP vulnerable index method to Donghuantuo Coal Mine, China. Rock Mech Rock Eng 44(5):591–600

Wu Q, Fan SK, Zhou WF, Liu SQ (2013) Application of the analytic hierarchy process to assessment of water inrush: a case study for the no. 17 coal seam in the Sanhejian Coal Mine, China. Mine Water Environ 32(3):229–238

Wu Q, Fan ZL, Zhang ZW, Zhou WF (2014) Evaluation and zoning of groundwater hazards in Pingshuo no. 1 underground coal mine, Shanxi Province. Chin Hydrogeol J 22(7):1693–1705

Wu Q, Liu YZ, Wu XL, Liu SQ, Sun WJ, Zeng YF (2016) Assessment of groundwater inrush from underlying aquifers in Tunbai Coal Mine, Shanxi Province, China. Environ Earth Sci 75(9):1–13

Wu Q, Zhao DK, Wang Y, Shen JJ, Mu WP, Liu HL (2017) Method for assessing coal-floor water-inrush risk based on the variable-weight model and unascertained measure theory. Hydrogeol J. https://doi.org/10.1007/s10040-017-1614-0

Zeng YF, Wu Q, Liu SQ, Zhai YL, Zhang W, Liu YZ (2016) Vulnerability assessment of water bursting from Ordovician limestone into coal mines of China. Environ Earth Sci 75(22):1431