Dynamic Prediction of Rock Mass Classification in the Tunnel Construction Process based on Random Forest Algorithm and TBM in situ Operation Parameters

S K Hou¹, Y R Liu¹, C Y Li¹,², P X Qin¹,²
¹State Key Laboratory of Hydroscience and Hydraulic Engineering, Tsinghua University, Beijing 100084, China
²Huaneng Yarlung Zangbo River Hydropower and Investment Co., Ltd., Chengdu 610093, China

Corresponding author’s e-mail: liuyaoru@tsinghua.edu.cn

Abstract. Accurately predicting the rock mass classification is of great significance to ensure the safe and efficient construction of tunnel boring machines (TBMs). On the basis of the TBM in situ operation data recorded during tunnel construction, a prediction model for rock mass classification using the random forest (RF) algorithm is proposed. Through data preprocessing, 7538 TBM excavation cycles were obtained to form a data set. Each data sample included 195 operation parameters and corresponding information on mileage and rock mass classification. Furthermore, 6784 samples were randomly selected as the training set and the remaining 754 samples as the test set. According to the changing characteristics of operation parameters, each TBM excavation cycle was divided into the empty-push phase, the rising phase, and the stable phase. On the basis of the mean decrease Gini index, seven machine parameters highly correlated with rock mass classification were selected. Then, the variation characteristics (i.e., mean value and linear fitting slope) of the seven operation parameters in the first 30 s of the rising phase were used as the input features of the RF model. Additionally, the hyperparameters in the RF model were analyzed. The quantitative results show that the prediction accuracy is up to 87.27%, indicating that the proposed model is effective for the prediction of rock mass classification.

1. Introduction

As tunnel engineering continues to be an interesting and useful engineering discipline, tunnel boring machines (TBMs) have played an increasingly important role in the process of tunnel excavation [1]. TBMs are sensitive to the change of rock mass conditions. For the tunnels constructed in a complex geological environment, the uncertainty of surrounding rock properties will be more prominent [2]. Therefore, it is of great significance to propose an effective and accurate method to predict the rock mass classification ahead of the tunnel face, which can ensure the safety and efficiency of TBM construction.

The first rock mass classification schemes had their origin in 1987 and were used to determine the support requirements of a tunnel [3]. On this basis, many rock mass classification systems have been developed from single-parameter classification schemes to multi-parameter classification schemes [4, 5]. In tunnel engineering, the commonly used classification systems are the rock mass rating, rock mass quality, and geological strength index. To determine the rock mass classification using these methods, we need to conduct field sampling before construction, laboratory experiments, or some
advanced geological prediction means [6], all of which, however, entail additional costs and equipment. Moreover, field sampling is discontinuous and relatively rough, and it cannot predict the rock mass classification in front of the tunnel face in real time. Many studies have shown that the operating parameters of TBMss have a strong correlation with the rock mass conditions [7]. On the basis of the analysis of the change of TBM operation parameters during the excavation process, the rock mass classification can be obtained. In recent years, because machine learning is highly advantageous in solving nonlinear problems, machine leaning-based models have been widely used in rock engineering, such as tunnels [8-10].

In this paper, a prediction model for rock mass classification using the random forest (RF) algorithm is proposed. On the basis of the 802-day data of the Songhua River water conveyance project in China, a data set having 7538 TBM excavation cycles with 195 operation parameters is formed. Each excavation cycle is divided into three working phases: the empty-push phase, the rising phase, and the stable phase. By calculating the mean decrease Gini index, seven operation parameters are selected as the high-correlation features with rock mass classification, namely, advance rate (v), penetration (Pr), cutterhead torque (T), total thrust (F), cutterhead power (P), cutterhead rotational speed (n), and pressure of gripper shoes (Ps). Furthermore, the variation characteristics (i.e., mean value and linear fitting slope) of the seven features in the data of the first 30 s of the rising phase were used as the inputs of the RF model. Then, 6784 and 754 samples were randomly selected as the training set and the test set, respectively, to test the prediction performance of the RF model on rock mass classification. Additionally, the influence of the hyperparameters in the RF model is analyzed.

2. Methodology

2.1. Decision tree

A decision tree is a commonly used machine learning algorithm that uses a simple and comprehensible structure for data analysis to solve the classification and regression problems [11]. A decision tree contains one root node and several branches (also called sub-trees). Before a decision tree model is used, the attribute variables and the target variable should be determined, and all data are input to the root node before splitting. Then, the root node is split into two or more branches on the basis of the specific split rule. Each branch is composed of some decision nodes and terminal nodes. The terminal node corresponds to a decision result, and the decision node corresponds to an attribute test, which will be split into new terminal nodes (or a new terminal node and a new decision node) on the basis of the specific split rule.

According to the different selection criteria of attribute variables in decision nodes, the decision tree algorithm can be divided into the ID3 algorithm, the C4.5 algorithm, and the classification and regression tree (CART) algorithm [12, 13]. In this study, the CART algorithm is used to conduct the attribute selection of each decision node. In the CART algorithm, the Gini impurity is used to select the attribute, to reduce the problem of large information gain caused by too many eigenvalues.

Given the data set \( S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \), suppose that the category quantity of the attribute variable is \( m \). The Gini impurity is calculated as follows:

\[
\text{Gini}(S) = \sum_{k=1}^{m} p_k (1 - p_k) = 1 - \sum_{k=1}^{m} p_k^2
\]  

(1)

where \( p_k \) represents the probability that the sample points belong to the class \( k \). The smaller the value of Gini impurity is, the lower the probability that the selected samples in the set will be misclassified, which means that the classification effect is better.

2.2. Random forest

A decision is a single classifier. For the training set, its classification performance is always good. However, decision trees are prone to overfitting when some classification problems are being dealt with, resulting in poor performance on the classification of a full data set. To solve the above problems, in 2001, Breiman [14] proposed the RF algorithm, which is a supervised ensemble learning algorithm
suitable for classification and regression problems. The idea of RF, which essentially is a multi-decision tree model, is to combine multiple CARTs with certain rules. Figure 1 shows the topological structure of the RF model.

![Figure 1](image_url)

Figure 1. The topological structure of the RF model.

Compared with the decision tree, RF introduces the idea of bagging and random subspace. According to the findings of many theoretical and case studies, the RF models have good prediction accuracy and do not easily overfit [15]. As shown in Figure 1, first, bagging is carried out; that is, bootstrap sampling is applied for the data set to extract $l$ subsets from the original data set as the training data set of each decision tree. The samples of the $l$ subsets are called in-bag data, and the unselected ones are contained in a subset called out of bag (OOB) [16]. When the RF model is training, the OOB data are used to estimate the internal error. The lower the OOB error rate, the better the performance of the RF model. When the RF model is used to deal with a classification problem, each decision tree obtains a classification result. Then, the final classification results are obtained by majority voting from all decision trees.

3. Data preparation and preprocessing

3.1. Project description and data acquisition

The study area of the tunnel in this paper is the No. 4 bid section of the Songhua River water conveyance project in Jilin Province, China. An open TBM constructs the main body of the project with a diameter of 8.03 m. The total length of the tunnel is about 22,955 m, of which about 20,198 m is for the TBM section, and 2757 m is for the drilling and blast section. In this paper, we choose the TBM section as the research object. During the tunnel construction process, the 195 TBM operation parameters are collected once a second. Finally, 802-day data are collected, and about 86,400 pieces of data are collected every day as a storage unit. Figure 2 shows the variation of four TBM operation parameters in a day. It can be seen that there are a large number of no-operation sections in a day. Through geological exploration and tests in the engineering site, the geological conditions of rock mass classification and its corresponding mileage are obtained. The engineering site mainly consists of surrounding rock samples belonging to four grades: grade III make up most of the samples, grade IV samples come second, and grades II and V samples are relatively few. The proportions of grades II–V rock mass are 8.13%, 66.73%, 20.03%, and 5.11%, respectively.
3.2. Data preprocessing

A TBM takes the excavation cycle as a working unit. Figure 3 shows a complete TBM excavation cycle. Each TBM excavation cycle can be divided into three working phases: the empty-push phase, the rising phase, and the stable phase. In the empty-push phase, the TBM is a start-up, and the cutterhead rotation speed is made to reach the set value. Then the advance rate increases gradually. When the TBM comes into contact with the tunnel face, the main control room will generate obvious vibration sense data. Then the driver will instantly reduce the advance rate, resulting in a falling edge in the advance rate curve. In the rising phase, the TBM operation parameters will increase sharply until they reach stable values. In the stable phase, TBM tunneling moves forward with each parameter maintained at their relatively stable states.

As shown in Figure 2, the collected data have many zero values, which are of no use to a machine learning model. Additionally, the data of the rising phase can reflect the interaction between the TBM and surrounding rock. Therefore, we use the data of the first 30 s of the rising phase to predict the rock mass classification, specifically by calculating the mean value and linear fitting slope of the first 30-s data points in the rising phase as the inputs of the RF model. First, we eliminate these useless data by constructing a state discriminant function [17]:

\[ S = f(v) \cdot f(n) \cdot f(F) \cdot f(T) \]  

(2)
\[ f(x) = \begin{cases} 0, & x = 0 \\ 1, & x \neq 0 \end{cases} \]  

(3)

\[ S = \begin{cases} 0, & \text{useless data} \\ 1, & \text{data of TBM excavation cycle} \end{cases} \]  

(4)

where \( S \) is the state discriminant function and the other symbols are as described above.

On the basis of the data preprocessing, 7538 TBM excavation cycles are obtained. Then, the data of the first 30 s of the rising phase are obtained by identifying the falling edge of the advance rate curve. The starting point \( t_s \) of the falling edge is a maximum extreme value so that it can be found by the \textit{findpeaks} function in MATLAB. Furthermore, because the duration of the falling edge is short, the ending point \( t_e \) of the falling edge can be determined by finding the minimum value point of \([t_s, t_s + 29]\). After that, the data of the first 30 s of the rising phase can be collected.

Because the inputs and outputs of the machine learning model are numerical data, it is necessary to conduct the quantitative disposal process for the rock mass classification. In this paper, the one-hot encoding method [18] is used to deal with rock mass classification. The one-hot encoding results are shown in Table 1.

| Rock mass classification | One-hot encoding result |
|--------------------------|-------------------------|
| II                       | (1, 0, 0, 0)            |
| III                      | (0, 1, 0, 0)            |
| IV                       | (0, 0, 1, 0)            |
| V                        | (0, 0, 0, 1)            |

4. Case study and analysis

4.1. Setting of predictors

As mentioned above, a total of 195 TBM operation parameters have been recorded. For machine learning models, too many redundant input features will reduce their calculation accuracy in some cases and will increase the calculation time. Therefore, it is necessary to carry out the feature selection process to determine the useful input parameters of machine learning models. The RF algorithm uses two indices for feature selection, namely, the mean decrease Gini and mean decrease accuracy [19]. In this study, the mean decrease Gini index is used to conduct the feature selection process. During the training process of the RF model, the mean value of the Gini impurity of each feature in all decision trees is calculated as the mean decrease Gini index. The larger the value of the mean decreases the Gini index, the more important the feature is. By calculating the mean decrease Gini, the following seven TBM operation parameters with a mean decrease Gini value of more than 14 are selected as the input features of the RF model: the pressure of gripper shoes \((P_{gs})\), cutterhead power \((p)\), cutterhead rotational speed \((n)\), total thrust \((F)\), penetration \((Pr)\), advance rate \((v)\), and cutterhead torque \((T)\). The mean decrease Gini value of the seven selected features is shown in Figure 4.

The RF model cannot directly deal with the time series data. Therefore, we calculate the mean value and linear fitting slope of the first 30-s data of the seven selected TBM operation parameters as the inputs of the RF model. That is, the input data of the RF model are composed of 14 features. Furthermore, 6784 (90\%) samples are randomly selected as the training set to train the model, and the remaining 754 (10\%) samples are selected as the test set to test the prediction performance of the model.
Figure 4. The mean decrease in the Gini value of seven selected features.

4.2. Hyperparameter analysis of the RF model
In the RF model, ntree and mtry are two important hyperparameters that can influence the prediction performance of the RF model. ntree is the number of decision trees. Generally, the more the number of decision trees are, the better the training effect of the model is. mtry is the number of random sampling variables when constructing the branch of the decision tree. Choosing the appropriate value of mtry can reduce the prediction error rate of the RF model. Generally, the value of mtry is set as 66% of the total number of variables.

First, for the value of ntree, we set the number of decision trees to 1, 10, 50, 100, 200, 300, 400, and 500, and the value of mtry is set as 9. The prediction accuracy under different values of ntree is shown in Figure 5. It can be seen that the prediction accuracy reaches a stable value when the number of decision trees exceeds 100. Second, we set the value of mtry to 5, 6, 7, 8, 9, and 10. The number of decision trees is set as 300. Figure 6 shows the prediction accuracy of the RF model under the different values of mtry, which, as can be seen, has little difference. Moreover, when the value of mtry equals to 8 or 9, the RF model demonstrates superior prediction performance. Using the above analysis on the hyperparameters, we set the values of ntree and mtry as 300 and 9, respectively.

Figure 5. Prediction accuracy of the RF model under different values of ntree.  
Figure 6. Prediction accuracy of the RF model under different values of mtry.

4.3. Prediction results and discussion
As discussed, 7538 TBM excavation cycles were extracted, of which 6784 (90%) were randomly selected as the training set and the remaining 754 (10%) as the test set. The established RF model is used to predict rock mass classification. The setting of the RF model is as mentioned above. All the calculations are performed in MATLAB and CPU environments, and the prediction accuracy is used as the evaluation index of model performance.

Figure 7 shows the prediction results of rock mass classification. Table 2 lists the prediction accuracy of each rock mass classification. It can be seen that the total prediction accuracy for rock mass
classification is up to 87.27% (658/754). For each rock mass classification, the prediction effect on grade III rock mass is highest with a prediction accuracy of 99.12%. The prediction accuracy of grade IV rock mass is 57.36% (74/129). However, the prediction effect on grades II and V rock mass is relatively low, with a prediction accuracy of 38.64% (17/44) and 35.71% (5/14), respectively.

The difference in prediction accuracy of the different grades of the rock mass is mainly due to the large difference in the number of samples of different rock mass classification. As mentioned in Section 3.1, the proportions of grades II–V rock mass are 8.13%, 66.73%, 20.03%, and 5.11%, respectively. Grade III rock mass accounts for the overwhelming majority of the samples, and grades II and V rock mass samples are very few. Therefore, the trained RF model has good prediction performance on grade III rock mass and relatively low prediction performance on grades II and V rock mass.

Overall, the proposed RF model is useful for the prediction of rock mass classification, and the result is also acceptable. Furthermore, it can be expected that the prediction performance of the model will be further improved with the increase of samples with different rock mass classification.

| Rock mass classification | II  | III | IV  | V   | Total       |
|--------------------------|-----|-----|-----|-----|-------------|
| Correct prediction/total samples | 17/44 | 562/567 | 74/129 | 5/14 | 658/754 |
| Prediction accuracy (%)  | 38.64 | 99.12 | 57.36 | 35.71 | 87.27 |

5. Conclusions
In this paper, a rock mass classification prediction model is proposed on the basis of the RF algorithm. With data from the No. 4 bid section of the Songhua River water conveyance project in China, a database was established, including 195 TBM operation parameters and rock mass classification corresponding to mileage. By data analysis and preprocessing, 7538 TBM excavation cycles were extracted to make up the data set. Seven TBM operation parameters were selected as the input features of the RF model on the basis of the mean decrease Gini index. Then, 6784 samples were randomly selected as the training set, and the remaining 754 samples were selected as the test set to test the prediction performance. The conclusion of this study are as follows:

(1) The data of the rising phase in a TBM excavation cycle can effectively reflect the interaction between the machine and its surrounding. This study proposes a method to extract the data of the rising phase by using the variation characteristics (i.e., mean value and linear fitting slope) of the first 30-s data of the rising phase as the inputs of the RF model.

(2) The prediction accuracy of the rock mass classification is 87.27%. However, because of the large difference in the number of samples of different grades of the rock mass, the trained model has
different prediction performance on different rock mass classification. The prediction performance of the RF model on grade III rock mass is good with a prediction accuracy of 99.12%, whereas that on grades II and V rock mass is relatively low with a prediction accuracy of 38.64% and 35.71%, respectively.

The case study shows that the proposed RF model herein is useful for the prediction of rock mass classification. However, the prediction performance is limited because of the small number of samples with grades II and V rock mass. In future work, we will increase the number of new samples from similar tunnel projects in the database to improve the model.

Acknowledgments
This work was supported by the State Key Laboratory of Hydrosience and Engineering (Grant No. 2019-KY-03). Additionally, we sincerely thank the data support received from the National Program on Key Basic Research Project (973 Program, No. 2015CB058100) of China, China Railway Engineering Equipment Group Corporation, and the Survey and Design Institute of Water Conservancy of Jilin Province.

References
[1] Liu Y R, Hou S K, Li C Y, Zhou H W, Jin F, Qin P X and Yang Q 2020 Study on support time in double-shield TBM tunnel based on self-compacting concrete backfilling material Tunn. Undergr. Space Technol. 96 103212
[2] Jung J H, Chung H, Kwon Y S and Lee I M 2019 An ANN to predict ground condition ahead of tunnel face using TBM operational data KSCE J. Civ. Eng. 23 3200–06
[3] Ritter W 1879 Die Statik der Tunnelgewölbe (Berlin: Springer)
[4] Barton N R, Lien R and Lunde J 1974 Engineering classification of rock masses for the design of tunnel support Rock Mech. 6 189–239
[5] Bieniawski Z T 1989 Engineering rock mass classifications: a complete manual for engineers and geologists in mining, civil, and petroleum engineering (New York: John Wiley & Sons)
[6] Rahmannnejad R and Mohammadi H 2007 Comparison of rock mass classification systems J. Min. Sci. 43 404–408
[7] Liu B, Wang R, Zhao G, Guo X, Wang Y, Li J and Wang S 2020 Prediction of rock mass parameters in the TBM tunnel based on BP neural network integrated simulated annealing algorithm Tunn. Undergr. Space Technol. 95 103103
[8] Koopialipoor M, Fahimifar A, Ghaleini E N, Momenzadeh M and Armaghani D J 2020 Development of a new hybrid ANN for solving a geotechnical problem related to tunnel boring machine performance Eng. Comput. 36 345–357
[9] Gao X J, Shi M L, Song X G, Zhang C and Zhang H W 2019 Recurrent neural networks for real-time prediction of TBM operating parameters Autom. Constr. 98 225–235
[10] Hasanpour R, Rostami J, Schmitt J, Ozelik Y and Sohrabian B 2020 Prediction of TBM jamming risk in squeezing grounds using Bayesian and artificial neural networks J. Rock Mech. Geotech. Eng. 12 21–31
[11] Salimi A, Rostami J and Moormann C 2017 Evaluating the suitability of existing rock mass classification systems for TBM performance prediction by using a regression tree Procedia Eng. 191 299–309
[12] Myles A J, Feudale R N, Liu Y, Woody N A and Brown S D 2004 An introduction to decision tree modeling J. Chemometr. 18 275–285
[13] Quinlan J R 1986 Induction of decision trees Mach. Learn. 1 81–106
[14] Breiman L 2001 Random forests Mach. Learn. 45 5-32
[15] Yu B, Wang H, Shan W and Yao B 2018 Prediction of bus travel time using random forests based on near neighbors Comput.-Aided Civil Infrastruct. Eng. 33 333-350
[16] Ramosaj B and Pauly M 2019 Consistent estimation of residual variance with random forest Out-Of-Bag errors Stat. Probab. Lett. 151 49–57
[17] Zhang Q L, Liu Z Y and Tan J 2019 Prediction of geological conditions for a tunnel boring...
machine using big operational data *Autom. Constr.* **100** 73–83

[18] Potdar K, Pardawala T S, Pai C D 2017 A comparative study of categorical variable encoding techniques for neural network classifiers *Int. J. Comput. Appl.* **175** 7–9

[19] Sandri M and Zuccolotto P 2010 Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms *Stat. Comput.* **20** 393–407