Geological Type Recognition by Machine Learning on In-Situ Data of EPB Tunnel Boring Machines

Qian Zhang,1 Kaihong Yang,1 Lihui Wang,2 and Siyang Zhou1

1Key Laboratory of Modern Engineering Mechanics, School of Mechanical Engineering, Tianjin University, Tianjin 300072, China
2Department of Military Vehicle, Academy of Military Transportation, Tianjin 300161, China

Correspondence should be addressed to Siyang Zhou; xiaodaidaizyy@163.com

Received 6 November 2019; Revised 25 March 2020; Accepted 30 March 2020; Published 27 April 2020

Academic Editor: Akhil Garg

Copyright © 2020 Qian Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, many large-scale engineering equipment can obtain massive in-situ data at runtime. In-depth data mining is conducive to the real-time understanding of equipment operation status or recognition of service environment. This paper proposes a geological type recognition system by the analysis of in-situ data recorded during TBM tunneling to address geological information acquisition during TBM construction. Owing to high dimensionality and nonlinear coupling between parameters of TBM in-situ data, the dimensionality reduction feature engineering and machine learning methods are introduced into TBM in-situ data analysis. The chi-square test is used to screen for sensitive features due to the disobedience to common distributions of TBM parameters. Considering complex relationships, ANN, SVM, KNN, and CART algorithms are used to construct a geology recognition classifier. A case study of a subway tunnel project constructed using an earth pressure balance tunnel boring machine (EPB-TBM) in China is used to verify the effectiveness of the proposed geological recognition method. The result shows that the recognition accuracy gradually increases to a stable level with the increase of input features, and the accuracy of all algorithms is higher than 97%. Seven features are considered as the best selection strategy among SVM, KNN, and ANN, while feature selection is an inherent part of the CART method which shows a good recognition performance. This work provides an intelligent path for obtaining geological information for underground excavation TBM projects and a possibility for solving the problem of engineering recognition of more complex geological conditions.

1. Introduction

With the rapid development of sensor detection technology, an increasing number of large-scale engineering equipment are made available to capably provide rich monitoring data in real time during construction. These data contain a large number of control rules related to equipment operation. An intelligent analysis of engineering monitoring data can provide a new path for the research on complex engineering problems and offer a decision-making basis for intelligent control of the engineering equipment.

A tunnel boring machine (TBM) is a type of large-scale engineering equipment that is widely used in tunneling construction. This equipment combines the functions of soil cutting, soil debris conveying, and tunnel supporting to achieve full mechanized construction of tunnel engineering [1] with a high degree of safety and construction efficiency [2]. A schematic diagram of the construction of a TBM is shown in Figure 1. A TBM consists of many parts, including a cutter head, TBM head, gripper, guniting, jumbolter, and belt conveyor [3]. Functions such as excavation, support, and guidance need to be carried out by different mechanical components and the synergy of the overall mechanical system. During the process of TBM excavation, it is necessary to continuously adjust the construction strategy based on the operational state and the information of the surrounding geological environment, which is an important basis for the safe and efficient operation of the machine [4]. However, due to the special characteristic of the underground excavation by a TBM, the
poor service conditions and complex muddy environment make it very inconvenient to observe the construction state of a TBM. Therefore, the monitoring data acquired and recorded by various sensors loaded on the main parts of the machine form an important information basis for understanding the working state of the equipment [5–8]. Current TBMs can simultaneously monitor hundreds of operation parameters, such as the tunneling rate, cutter head rotational speed, cylinder thrust, cutter head torque, and sealed chamber pressure. These in-situ data contain rich information on interactions between the machine and the surrounding environment. The rapid development of technologies such as big data and artificial intelligence in recent years provides a more effective method and path for in-depth and sufficient exploration of information provided by in-situ data to realize the informatization and intelligence of TBM construction [9].

In the early stage, the engineering data of TBM was used to establish some empirical models used in solving practical problems conveniently. For example, Krause [10] used the data of hundreds of TBM construction from Germany and Japan to analyze and give the empirical prediction range of tunneling load. In addition, classic empirical models include the NTNU model [11] developed by the Norwegian University of Science and Technology and the improved model of Bruland [12], which are often used in engineering for the prediction of the rate of penetration. On the basis of empirical models, many scholars have further given parameter prediction models based on a statistical analysis of engineering data. For example, Zhang et al. [13] established tunneling load prediction model for the earth pressure balance tunnel boring machine (EPB-TBM) by combining regression analysis with dimensional analysis from engineering data. Avunduk and Copur [14] established a nonlinear regression model of rate of penetration by several soil property parameters such as particle size distribution and natural water content. Macias et al. [15] analyzed the change rule of prediction curve of rate of penetration of a hard rock TBM under different fracturing conditions, and the fracturing coefficient was determined as an effective index of the influence of rock fracturing on tunneling performance. The single index affecting the rate of penetration was regressed by Armetti et al. [16] to analyze the influence degree of different parameters in the empirical model on the tunneling performance. Vergara and Saroglou [17] established the regression relationship between the weighted rock mass rating and mixed-face penetration index under the condition of mixed geology, considering the proportion of rock and soil in the tunneling face. Yagiz et al. [18] set up regression equation to predict the tunneling performance under the condition of joint fault rock based on the rock properties such as distance between planes of weakness and orientation of discontinuities in rock mass. The works based on TBM in-situ data mainly focus on the basic statistical regression of key tunneling performance parameters, with some limitations on the applicable problems and the number of features that can be considered. Statistical regression can extract and describe the rules in the data. But TBM is a complex engineering system with hundreds of parameters collected in the process. Moreover, TBM in-situ data are often characterized by many influencing factors and nonlinear coupling between parameters, which makes it difficult for in-depth data mining [19]. The valuable information hidden within the massive monitoring data remains to be explored.

In recent years, machine learning algorithms have been developed rapidly. Because of their excellent nonlinear expression ability and adaptability to massive data, they provide powerful tools for TBM in-situ data analysis. Some typical works are as follows: Bouayad and Emeriault [20] established a prediction model of ground settlement caused by shield machine based on earth pressure balance through the principal component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS). Mahdevri et al. [21] used a support vector machine and artificial neural network to predict the tunnel convergence caused by ground compression and verify the output results and measured data of the model through engineering examples. Hyun et al. [22] combined fault tree analysis (FTA) and analytic hierarchy process (AHP) to analyze the risk and probability of shield construction and constructed a risk management system according to good consistency. Salimi et al. [23] used nonlinear regression and artificial intelligence algorithms to predict the performance of the hard rock TBM. Sun et al. [24] established a model by random forest to predict the dynamic load of shield tunneling. Gholamnejad and Tayarani [25] used an artificial neural network to predict the rate of penetration with three rock mass parameters of uniaxial tensile strength, rock quality index, and weak face spacing and tried to evaluate the results with different hidden-layer settings. Adoko et al. [26] proposed a Bayesian method to select the performance of different tunneling machines. Seker and Oçak [27] compared the application effect of random forest and other ensemble learning algorithms in the prediction of the rate of penetration. Gao et al. [28] used several kinds of recurrent neural networks to analyze the sequence rule of TBM performance parameters, so as to predict the important performance parameters in advance.
Previous studies have shown that machine learning methods can be used in multiparameter analysis of TBM data. In addition, these results indicate that changes in the geological types during TBM driving will be reflected in the in-situ data through tunneling between the machine and the geology. Due to the characteristics of underground excavation in TBM, various geological conditions may be faced in TBM tunneling. The geology varies greatly between different projects, such as soft soil, hard rock, and composite ground. Therefore, geological conditions are the important factors affecting the project and geological type recognition is one of the major tasks in TBM engineering. Therefore, it is a feasible way to identify geology category by digging into the relationship between TBM in-situ data and geological conditions. Furthermore, it may be a feasible way to analyze the relationship between TBM in-situ data and geological conditions, so as to identify different construction geological types.

In this paper, feature selection and machine learning methods are introduced into the engineering data analysis to propose a geological recognition system based on in-situ data analysis during tunneling. The proposed method provides an effective way to acquire geological information for construction decision-making. The influence parameters sensitive to the change of geological type are selected as input features by the feature engineering algorithm for dimension reduction. While four machine learning classification algorithms KNN (k-nearest neighbor), SVM (support vector machine), ANN (artificial neural network), and CART (classification and regression tree) are selected to train different geological type labels. And the recognition performance is evaluated in an independent test set. Through the above steps, the sensitive features are extracted from in-situ data, and the geological recognition system is established. In this paper, a subway tunnel project constructed by the tunnel boring machine (EPB-TBM) is taken as a case to discuss the effectiveness of the above methods. The procedure of the proposed TBM geological recognition system is shown in Figure 2.

2. Methods

The geological recognition system proposed in this paper mainly includes the following three steps. First, normalization preprocessing is performed to reduce the dominant effects generated by the difference in dimensions and order of magnitude between different parameters in the TBM in-situ data. Second, the chi-square test, which is the non-parametric test method in feature selection, is used to select the key parameters that are highly sensitive to geological variation as input features. Third, several typical machine learning classification algorithms are used to train the data sets with geological labels to obtain the geological recognition classifier, which is used to perform the geological type recognition. The test set data are used to validate the accuracy of the geological recognition system and evaluate the effectiveness of the method.

2.1. Data Preprocessing. During the TBM excavation process, numerous types of information related to machine operation, such as hundreds of different types of engineering parameters, including the cylinder thrust, motor torque, cutter head rotational speed, advance rate, guiding attitude, and sealed chamber pressure, can be recorded in real time in the data acquisition system. These engineering parameters have various dimensions, and the corresponding numerical magnitudes are very different. For example, the cylinder thrust can reach tens of thousands of kN, while the advance rate is usually only tens of millimeters per minute, both of which are important factors reflecting the features of the operating states of the machine in different geological conditions.

Considering that most feature selection and machine learning algorithms are not invariant to scale, to prevent certain parameters from playing a dominant role in data mining due to differences in the order of magnitude, all the
parameters in this work are min-max normalized before machine learning classification. The calculation method is

\[ x_{\text{pre}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]

where \( x_{\text{pre}} \) is the dimensionless form after normalization pretreatment, \( x_{\text{min}} \) is the minimum value in the recorded data of this parameter, and \( x_{\text{max}} \) is the maximum value in the recorded data of this parameter.

Min-max normalization can convert parameters from dimensional to dimensionless and map the parameters to the interval of 0 to 1, so that parameters with different dimensions and orders of magnitude can be treated as equally possible in the subsequent analysis. In addition, the use of normalization in the actual solution is beneficial for improving the convergence speed and results.

2.2. Feature Engineering. As mentioned above, the in-situ data include the records of hundreds of engineering parameters, whereas in the existing engineering experience, only a few parameters such as cylinder thrust and motor torque are used to analyze the geological conditions [29, 30]. However, it is of great concern to fully investigate the variation of the parameters in the data with the geological conditions, thus achieving effective geological recognition. To this end, it is necessary to more comprehensively consider and select the parameters that are highly sensitive to geological changes as the input features for the subsequent machine learning, namely, to conduct feature engineering. Through this step, redundant parameters with low correlation with geological changes can be removed, while the informative parameters are retained, which is conducive to improving the recognition accuracy, reducing the empirical risk and avoiding the overfitting problems caused by incorrect generalization due to the accidental nature of certain parameters in engineering.

The engineering data often do not follow the common data distribution forms, and the relationships among many parameters cannot be explained by independent statistical analysis. Instead, the target variable is influenced by a combination of parameters [31]. Therefore, this work uses the chi-square test algorithm for feature engineering. The chi-square test is a nonparametric test method that represents the degree of the deviation between the observed value and the theoretical value based on the independence assumption, and it does not make assumptions on the data distribution. Hence, this method is suitable for the analysis of the engineering data in this research. Its basic principle is to evaluate the parameter independence by calculating the deviation between the theoretical value and the expected one. The specific calculation formula is

\[ \chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i} \]

where \( \chi^2 \) is the chi-square value of the parameter, \( k \) is the number of recorded values, \( N_i \) is the actual value, and \( E_i \) is the expected value. \( \chi^2 \) is a measure of the degree to which the expected value and the actual value deviate from each other.

The high value of \( \chi^2 \) indicates that the independent hypothesis is incorrect, that is, the parameter as an input feature is helpful to judge whether a certain kind of event occurs or not.

The geological type recognition problem to be solved in this work is essentially a type of supervised classification problem. For the training set data, different geological types are marked with the supervised learning label in the construction area of known geological information. Using the geological label as the target, the chi-square test is performed on the training set data to yield the chi-square value of each parameter under the given geological label. The values are sorted from the largest to the smallest, and the first few parameters, that is, those with the highest sensitivity, are selected as the input features of the subsequent recognition algorithm. After testing and validation, the input features with the best recognition performance are selected as the optimal input features.

Due to the long distance of TBM construction and irregular geological changes, the in-situ data of TBM are massive and with the nonuniform distribution of information. To more effectively evaluate the impact of different feature selection strategies on the performance of the geological recognition system, this paper uses the 10-fold cross-validation method [32], with its basic idea given in Figure 3. The dataset is divided into ten subsets with similar amounts of data. A subset is selected as the test set successively without repetition, and the remaining nine subsets are used as the training sets until all the subsets have been validated as test sets once. Finally, the evaluation values using the ten test sets are averaged and taken as the final evaluation value of the 10-fold cross-validation method. Thus, the contingency and randomness problems caused by the use of a single test set are avoided as much as possible, giving an insight on how the model will generalize to an independent dataset.

2.3. Applied Algorithms and Classification Metrics. Considering the characteristics of TBM in-situ data, including high dimensionality, nonlinear coupling of parameters, and high noise, four commonly used supervised classification algorithms, namely, KNN (k-nearest neighbor), SVM (support vector machine), ANN (artificial neural network), and CART (classification and regression tree), are selected in this study to express the relationships between the input features and geological labels and to establish several corresponding geological type recognition systems.

As shown in Figure 4(a), the k-nearest-neighbor (KNN) [33] algorithm is an example-based method, which makes decisions on prediction by the properties of \( k \) sample points closest to the prediction points in the feature space. The principle is simple, and it can adapt to multiclassification tasks.

Support vector machine (SVM) [34] illustrated in Figure 4(b) is a geometric method to find the optimal separating hyperplane through support vector. In the nonlinear case, SVM maps the nonlinear problems in the original space to the high-dimensional space through the
kernel function, which only needs fewer support vectors to make decisions and adaptability to the high-dimensional problems, making it one of the most widely used machine learning methods.

The basic principle of the artificial neural network (ANN) is shown in Figure 4(c), which is a nonlinear fitting model inspired by the biological neural system [35]. It is mainly composed of input layer, hidden layer, and output layer. In the hidden layer, it is endowed with nonlinear properties by complex network structure and activation function. Because of its strong nonlinear expression ability, it has become one of the most popular fields in machine learning methods in recent years.

In Figure 4(d), the classification and regression tree (CART) [36] method is one of the decision tree methods. Through the Gini index, it constantly searches for the best feature and the best segmentation point and divides the binary tree, so as to complete the classification of the whole data set. The biggest characteristic of the cart algorithm is that it can provide a clear and even visual decision-making process, thus providing useful guidance in practical engineering.
To quantify the quality of predictions, there are several metrics that are adopted to assess the prediction accuracy. Among the supervised classification problems in machine learning, the accuracy (AR), precision (PR), recall (RE), and $F_1$-score ($F_1$) are the most commonly used indices to evaluate the performance of classifiers. Besides, the confusion matrix is a format used to show classification results. For example, the confusion matrix for the binary classification problem is shown in Table 1, where true positive (TP) is a prediction of a positive class as a positive class, true negative (TN) is a prediction of a negative class as a negative class, false positive (FP) is a prediction of a negative class as a positive class, which is a type I error, and false negative (FN) is a prediction of a positive class as a negative class, which is a type II error.

Based on a given confusion matrix, the accuracy, precision, and recall can be calculated. The accuracy is the most common classification evaluation, which represents the number of correctly classified samples divided by the total number of samples. The precision represents the percentage of samples that are correctly classified in the samples that are determined to be of a certain class. The recall is a measure of the covering surface and represents the proportion of correctly classified samples in the samples that should be classified as a certain class. Since precision and recall sometimes conflict with each other, high precision is usually accompanied by a low recall, and vice versa, while the $F_1$-score is a comprehensive evaluation of these two parameters. In this paper, these four indices are used to evaluate geological recognition results. The calculation method of each evaluation index is as follows:

\[
\begin{align*}
AR &= \frac{TP + TN}{TP + TN + FP + FN} \\
PR &= \frac{TP}{TP + FP} \\
RE &= \frac{TP}{TP + FN} \\
F_1 &= \frac{2 \times PR \times RE}{PR + RE}
\end{align*}
\]

### 3. Results and Discussion

The method proposed in this paper is applied to the geological recognition of the actual tunnel engineering, and the applicability and effectiveness of the method and the recognition performance of different classification algorithms are discussed in this section. As a preliminary study to use the machine learning method to recognize geological types, in order to test the feasibility of this method, Tianjin Metro Line 9 and Tianjin Metro Line 3 are discussed in this paper, which are mainly composed of soft soil. A few types are involved in the section of the data, such as muddy clay and silt and silty clay. The section of Tianjin Metro Line 9 is approximately 1104 m long, constructed using an EPB-TBM. The construction area of this project mainly passed through soft soil such as silty clay, muddy clay, and silty soil.

| Table 1: Confusion matrix for a binary classification problem. |
|-----------------|-----------------|-----------------|
|                 | Positive        | Negative        |
| True            | True positive (TP) | True negative (TN) |
| False           | False positive (FP) | False negative (FN) |

The engineering data used in this paper have 357 parameters recorded by the data acquisition system during construction. The sampling frequency was approximately set to every 30 s (approximately advanced by 17 mm). In the application, the dataset is divided into training sets and test sets according to certain proportions. The training set data are used to establish the geological recognition system, while the test set data are not involved in the training process but used for independent testing of the recognition results. To obtain the geological recognition labels of the supervised classification algorithm, the geological survey report obtained from the geological exploration is used as the prior information. Table 2 lists the basic statistical characteristics of some representative TBM tunneling parameters in Tianjin Metro Line 9.

3.1. Implementation of Feature Selection and Performance of the Geological Recognition System. In this section, the recognition accuracy and computational time of the four geological classifiers are discussed, with different numbers of features selected using the aforementioned engineering example.

For feature engineering dealing with high-dimensional problems, it is necessary to comprehensively consider the issues of training precision, computational cost, and possible overfitting in the selection of the appropriate number of features as the effective input for classifier training. Therefore, the effect of the number of different features on the recognition accuracy of the four types of geological classifiers is discussed first.

The hyperparameters of the algorithms used are set as follows, the number of hidden layers is 4 and the number of nodes in each layer is 10 in the ANN. The distance metric is Euclidean distance for the KNN. The kernel function of SVM is radial basis function and the criterion in CART is Gini coefficient.

In the Tianjin Metro Line 9, the variation results of the geological recognition by KNN, SVM, ANN, and CART are shown in Figure 5. All the points are given by the accuracy from 10-fold cross validation. In this figure, it is shown that as the features are added to the feature input according to the chi-square values, the accuracy of the geological recognition models gradually increases. Eventually, the algorithms have good recognition performance, with the accuracy exceeding 96% after $K > 4$. Performance of algorithms is discussed based on the results in Figure 5. Among the four algorithms, the recognition performance of KNN is significantly good and the classification result reaches an accuracy of 99.9% when $K = 3$ in KNN, probably because the TBM construction is a continuous process so that KNN can find similar samples for decision in the high-density TBM data collection more effectively. While the performance of SVM is inferior to
other algorithms, which may be caused by the difficulty in the selection of the hyperparameters resulting from the complex distribution characteristics and noise phenomenon of TBM in-situ data.

Figure 5 can also provide some references for the number of inputs for this multiple input problem. For most algorithms, the recognition results are generally good when $K = 7$. Subsequently, with the increasing number of features, the accuracy only slightly improves. Combined with the consideration of the calculation and feature acquisition costs and the complexity of the recognition system, $K = 7$ is used for feature combination as the optimal feature selection strategy for the geological recognition operation in this work. In addition, since feature selection is an inherent part of the CART algorithm, it does not participate in the discussion of the chi-square test and the number of input features can be controlled by adjusting the depth of the tree. The top 7 features selected by this method and their chi-square values and $P$-values are shown in Table 3.

To discuss the dependence of the proposed geological recognition system on the amount of data in the training set, 10% of the samples are randomly selected from the engineering datasets as the training set, and the remaining 90% samples are used as an independent test set. The above feature selection results are used as input for the training classifiers to validate the recognition accuracy again using the independent test set. The computational costs of the training and predicting for these four types of the classifier are compared. The results are shown in Table 4. For the SVM, ANN, and KNN classifiers, the computational time of the chi-square test is excluded, and the duration of each algorithm from the training set fitting to the test set prediction is measured. It should be noted that the feature selection of CART is included in its training process. Table 4 demonstrates that even only 10% of the samples are used for training, the optimal feature combination selected by the feature selection algorithm still retains excellent recognition performance when 90% of the samples are used for prediction and validation.

In the Tianjin Metro Line 9, the computational cost of the CART-based geological classifier is significantly smaller than those of the other three. The prediction time of the KNN classifier is the longest and is significantly longer than those of the other three classifiers, since KNN is an instance-based algorithm and the training process of KNN is only a storing process. Moreover, each prediction requires the calculation of the distances between the point to be predicted and all the sample points in the training set, resulting in a longer prediction time with regard to a large amount of data. In addition, for other classifiers, the prediction of the test set is relatively fast after the training is completed.

3.2. Generalization Ability of Geological Recognition Systems. The generalization ability is an important indicator to evaluate whether a learner has the overfitting phenomena, which is a prerequisite for the practical application of the proposed method in engineering problems. The generalization ability generally refers to the adaptability of the machine learning method for predicting the new data, that is, whether a reasonable output can still be achieved when a dataset outside the training set is given. In this work, instead of using the data from Tianjin Metro Line 9, Tianjin Metro Line 3 is used as the engineering example for generalization validation. The statistical characteristics of its dataset involved in the calculation are shown in Table 5.

This project and Tianjin Metro Line 9 were both constructed using the same EPB shield, and they are located in the same city with similar geological conditions. In this section, the generalization of the geological recognition system proposed in this paper is investigated by using the geological recognition system established on the basis of the in-situ data of the Tianjin Metro Line 9 project to the geological recognition of Tianjin Metro Line 3, which is another project independent of Tianjin Metro Line 9. Considering the difference of the parameters in
different engineering datasets of Line 9 and Line 3, the feature selection method introduced in Section 2.2 is used to select the important features in both engineering datasets, as the input for the training of the geological recognition system. Based on the training set data from Tianjin Metro Line 9, three classification algorithms (ANN, SVM, and KNN) are used to establish the geological recognition system. The recognition performance is verified using the test set data from Line 3. The evaluation indicators for recognition performance are shown in Figure 6. The geological recognition system trained using the engineering data of Line 9 can effectively recognize the similar geology in the Line 3 project, such as muddy clay and silty clay. The recognition accuracies of the three types of classifiers are all above 90%. Among the classifiers, the ANN outperforms the other two algorithms, and algorithms based on KNN and SVM have similar prediction results. The results show that the geological recognition system based on the existing training engineering data can give a reasonable output when applied to new datasets from different projects with similar geological conditions, demonstrating a good performance.

### Table 3: Chi-square value and P-value of top 7 features in Tianjin Metro Line 9.

<table>
<thead>
<tr>
<th>Name</th>
<th>Chi-square value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accumulated rotation number of screw machine</td>
<td>4913.95</td>
<td>&lt;1e−275</td>
</tr>
<tr>
<td>Primary voltage of high-voltage transformer</td>
<td>4360.52</td>
<td>&lt;1e−275</td>
</tr>
<tr>
<td>Filling mud pressure no. 2</td>
<td>3692.76</td>
<td>&lt;1e−275</td>
</tr>
<tr>
<td>Initial pitch angle y</td>
<td>3626.04</td>
<td>&lt;1e−275</td>
</tr>
<tr>
<td>Upper control limit of soil discharge quantity</td>
<td>3339.61</td>
<td>&lt;1e−275</td>
</tr>
<tr>
<td>Standard deviation of soil discharge quantity</td>
<td>3311.44</td>
<td>&lt;1e−275</td>
</tr>
<tr>
<td>Target value of soil pressure</td>
<td>3162.43</td>
<td>4.23e−275</td>
</tr>
</tbody>
</table>

### Table 4: The recognition performance of classifiers in Tianjin Metro Line 9 when the ratio of training and test set is 1:9.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Accuracy (%)</th>
<th>Training time (s)</th>
<th>Test time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>99.8</td>
<td>0.29</td>
<td>81.45</td>
</tr>
<tr>
<td>SVM</td>
<td>95.9</td>
<td>3.21</td>
<td>5.60</td>
</tr>
<tr>
<td>ANN</td>
<td>97.3</td>
<td>5.76</td>
<td>0.03</td>
</tr>
<tr>
<td>CART</td>
<td>99.9</td>
<td>0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>

### Table 5: Statistical properties of main parameters from the selected section of Tianjin Metro Line 3.

<table>
<thead>
<tr>
<th>parameter</th>
<th>Max</th>
<th>Min</th>
<th>Average</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advance rate (mm/min)</td>
<td>39.80</td>
<td>18</td>
<td>31.53</td>
<td>0.53</td>
</tr>
<tr>
<td>Cylinder thrust (kN)</td>
<td>27447.14</td>
<td>15511.25</td>
<td>21198.53</td>
<td>3814.17</td>
</tr>
<tr>
<td>Cutter head torque (kNm)</td>
<td>1539.93</td>
<td>1030.53</td>
<td>1279.04</td>
<td>86.89</td>
</tr>
<tr>
<td>Cutter head rotational speed (r/min)</td>
<td>0.8</td>
<td>0.5</td>
<td>0.65</td>
<td>0.15</td>
</tr>
</tbody>
</table>

![Figure 6: Recognition performance in Tianjin Metro Line 3 using geological classifiers trained by data from Tianjin Metro Line 9.](image)
generalization ability of the proposed recognition systems.

4. Conclusions

This paper proposes a method based on the in-situ data recorded during TBM construction to conduct geological type recognition. The main conclusions of this research can be summarized as follows:

1. The proposed method consists of feature engineering and machine learning classification methods. The recognition method based on the analysis of TBM in-situ data can effectively mine the internal influence law between variables and provide an effective way to obtain geological information for construction decision-making.

2. In feature engineering, considering the disobedience of TBM in-situ data to common distributions, the chi-square test method is chosen for feature selection. Four machine learning classification algorithms, ANN, SVM, KNN, and CART, are used for the nonlinear coupling between features.

3. The proposed method is applied to the geological recognition of urban metro projects constructed with EPB-TBM in China. The comparison between the recognition results and the measured geology types shows that proposed method is effective. The recognition accuracy gradually increases with the increase of input and eventually reaches a flat level when the accuracy of all algorithms is higher than 97%. Based on this trend, a selection strategy for optimal input features is also given that the optimal number of input variables for this validation case is seven.

4. Studies regarding more advanced applications would be worthwhile: a database with more comprehensive geological types (such as hard rock, composite ground) is recommended to be established and analyzed though the presented learning procedure. Moreover, intelligent visual interface could be conducted based on the proposed system for more convenient applications.

Data Availability

The data used in this paper are available from the relevant engineering enterprises, which have not been released for commercial reasons.

Conflicts of Interest

All authors declare that there are no conflicts of interest.

Acknowledgments

This research was supported by the National Key R&D Program of China (grant no. 2018YFB1702505), National Natural Science Foundation of China (grant no. 11872269), and Natural Science Foundation of Tianjin (grant no. 18JCYBJC19600). Support from Prof. Wei Guo in the School of Computer Science and Technology of Tianjin University for the research in this paper is greatly appreciated.

References


