IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Hybrid Data Mining Method for Tunnel Engineering Based on Real-Time Monitoring Data from Tunnel Boring Machines

Shuo Leng[1], Jia-Rui Lin[1], Zhen-Zhong Hu[1, 2, *], and Xuesong Shen[3]

[1] Department of civil engineering, Tsinghua University, Beijing, China, 100084
[2] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, 518055
[3] School of Civil and Environmental Engineering, the University of New South Wales, Sydney, Australia, NSW 2052

Corresponding author: Zhen-Zhong Hu (e-mail: huzhenzhong@tsinghua.edu.cn).

This research is supported by the National Natural Science Foundation of China (No. 51778336, No. 51908323) and the Tsinghua University-Glodon Joint Research Centre for Building Information Model (RCBIM). The authors also acknowledge the Metro Protection Department of the Guangzhou Metro for providing the data support and thank Prof. Lucio Soibelman (University of Southern California) for his valuable comments.

ABSTRACT Tunnel engineering is one of the typical megaprojects given its long construction period, high construction costs and potential risks. Tunnel boring machines (TBMs) are widely used in tunnel engineering to improve work efficiency and safety. During the tunneling process, large amount of monitoring data has been recorded by TBMs to ensure construction safety. Analysis of the massive real-time monitoring data still lacks sufficiently effective methods and needs to be done manually in many cases, which brings potential dangers to construction safety. This paper proposes a hybrid data mining (DM) approach to process the real-time monitoring data from TBM automatically. Three different DM techniques are combined to improve mining process and support safety management process. In order to provide people with the experience required for on-site abnormal judgement, association rule algorithm is carried out to extract relationships among TBM parameters. To supplement the formation information required for construction decision-making process, a decision tree model is developed to classify formation data. Finally, the rate of penetration (ROP) is evaluated by neural network models to find abnormal data and give early warning. The proposed method was applied to a tunnel project in China and the application results verified that the method provided an accurate and efficient way to analyze real-time TBM monitoring data for safety management during TBM construction.

INDEX TERMS Data mining, monitoring data, tunnel boring machine (TBM), tunnel construction, underground structure.

I. INTRODUCTION

Tunnel boring machines (TBMs) have been widely used as an effective tool for tunnel construction which is characterized by its large scale, high cost and long project lifecycle, and the construction process is always associated with complex technical problems and potential risks [1]. Modern TBMs usually integrate a large number of sensing and monitoring methods and record a series of operation parameters [2] such as stress, current, flow and gas pressure, etc. These monitoring data indicate the working state of TBM, and can be used to reduce construction risks, optimize construction operations and ensure construction safety. With the development of sensing technology, some cyber-physical platforms have been built to achieve effective data management [3], [4], including automatic data collection, storage and visualization. However, manual analysis falls short of consistent standards and poses risks for tunnel construction. In fact, some tragic construction accidents have occurred due to the lack of timely and accurate data analysis. A serious water inrush accident occurred in a tunnel project in Guangdong, China on February 2018, resulting in more than 10 fatalities and major economic losses. The cause of the accident was that the TBM entered into a permeable layer that had not been identified by geological investigation. This could be avoided if the formation anomalies were recognized in advance, for example, by analyzing the real-time monitoring data. However, the real-time automatic analysis

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

of data still faces many difficulties due to the following characteristics of TBM monitoring data.

(1) Variables related to critical parameters are difficult to determine. As megaprojects with enormous uncertainties, tunnel construction needs to record data from every aspect and subsystem to reduce risk. The number of monitoring parameters can be more than one hundred, and the relationships among them are difficult to understand and model manually [5]. When critical parameters need to be predicted and analyzed, it is difficult to determine all related variables from the set of massive parameters. As a result, the accuracy of data analysis is affected.

(2) The heterogeneity of soil makes the data not accurate. Compared to projects above the ground, the geological information plays an extremely important role in underground tunnel construction. However, the geological data are obtained through the borehole sampling and the formations between the sampling points are usually calculated by linear fitting. The fitting results often differ from the actual situation and affects the accuracy of data analysis. Some weak formations may also not be sampled and discovered, bringing hidden dangers to the tunnel construction process.

(3) The complex nonlinear relationship between monitoring variables is challenging for data analysis and modeling. When a TBM is working, soil, shield and internal machinery interact with each other, making the correlation between parameters complicated and difficult to be described in a simple mathematical formula [6]. As a result, those commonly used statistical analysis methods do not perform well.

The aim of this study is to conquer the above problems and achieve automatic and efficient analysis of TBM monitoring data to ensure the construction safety. To solve this problem, DM techniques are introduced in this paper. DM is a data-driven analysis method which is especially suitable for extracting the required patterns from massive, fluctuating and complex data [7]. It has been validated that DM methods can promote the value of monitoring data in engineering projects [8]. However, the application of DM in the field of TBMs has encountered challenges due to strict efficiency and accuracy requirements. A novel hybrid DM method is proposed in this paper to overcome these challenges. The originality of this paper contains the following 5 aspects. (1) Association rules are extracted to supplement people's experience and assist in judgement of anomalies. (2) Formation data is refined by classification analysis to support the decision-making process of TBM operators. (3) The performance of TBM is evaluated by ANN models and abnormal conditions can be discovered by comparing the evaluation value with actual value. (4) Improvements have been made to existed DM algorithms based on TBM characteristics to improve performance. (5) The three DM algorithms are not only designed to solve single problems, they are also combined to form a hybrid DM framework to improve data mining process. The extracted association rules are used for parameter selection in subsequent algorithms, and the classified formation is applied to improve data quality in rate of penetration (ROP) prediction. Verified by a case study, the efficiency and accuracy of algorithms is improved by the proposed hybrid DM framework. The mining results, including construction laws, refined formation and reviewed ROP, are combined to provide support for safety management in tunnel engineering.

The remainder of this paper is organized as follows. Related studies of TBM data analysis are first reviewed. The framework of the proposed hybrid DM method is then introduced. The following three sections investigate three different data mining methods. The association rule algorithm is used to model the relationship between monitoring parameters. The classification algorithm is then introduced to estimate the formation where TBM is located. Finally, the artificial neural network (ANN) is applied to predict the ROP. The last two sections discuss and conclude main findings of the study.

II. LITERATURE REVIEW

A. DATA ANALYSIS IN TBM AND TUNNEL ENGINEERING

When it comes to the field of shield tunneling, how to assess the performance of TBMs accurately remains one of the most challenging issues for both practitioners and researchers. Accurate judgement of the TBM performance can assist selecting machine and forecasting project duration, leading to a reduction in project costs [9]. In general, to predict the performance of TBMs is to estimate certain parameters which include ROP and advance rate (AR), while input variables include rock properties and machine parameters [10]. Typically, the prediction methods can be divided into three categories: theoretical methods, empirical methods, and numerical methods. Rostami [10] elaborated theoretical and empirical methods in a recent review.

The theoretical methods build models based on force balance among rock, cutters and internal machinery. One of the most frequently used theoretical models is the Colorado School of Mines (CSM) model developed by Rostami and Ozdemir [11]. The model is based on basic principles of rock cutting with disc cutters, and considers the influence of rock mechanical properties, disc cutter geometrical parameters and cutting parameters. Cheema [12] and Ramezanzadeh et al. [13] modified the CSM model by introducing other rock mass parameters.

The empirical models are based on engineering experience involving a large number of laboratory tests, field measurements and construction records. A commonly used empirical model is the Norwegian University of Science and Technology (NTNU) model [14]. Several revisions and improvements have been made to the NTNU model, and the latest work was done by Macias [15]. Field Penetration Index (FPI) model is another popular empirical model to predict

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

TBM's performance. Recent modifications to the model were done by Hassanpour et al. [16] and Delisio and Zhao [17].

With rapid advancement of computation theory and devices in recent years, utilization of the numerical simulation method has drawn increasing attentions in the literature. Nonlinear regression analysis is a basic but effective simulation method, by which a fitting formula between TBM performance and input parameters can be obtained [18], [19]. In the field of artificial intelligence (AI) methods, artificial neural network (ANN) is a popular way to establish models for its strong learning and nonlinear fitting ability [20], [21]. Meanwhile, fuzzy logic is also used in modeling [22] and has been further combined with ANN to become a neuro-fuzzy method [23]. The use of other algorithms, such as support vector machines (SVM) [24], gene expression programming (GEP) [25] and particle swarm optimization (PSO) [26], has also been tried to predict the performance of TBM. . Isam and Zhang [5] investigated the applications of soft computing techniques in TBM performance prediction, and concluded that soft computing methods shows good performance in dealing with complex relationships among TBM parameters.

In addition to the evaluation of TBM performance, the monitoring data also reflect some other working conditions of TBMs which are worthy of attention. The calculation of the interaction between soil and TBMs can predict the risk of the machine and surrounding environment, and help to optimize the design and selection of TBMs. Acaroglu [27] developed a fuzzy logic model to determine the cutting forces and energy cost for disc cutters. The model was based on experience and linear cutting tests, and its input parameters included rock properties and mechanical dimensions. Festa et al. [28] established a prediction model to quantify the magnitude of the driving force and its temporal and spatial distribution by mining the TBM logged data. Moreover, a dynamic load prediction model was proposed based on the random forest algorithm [29]. Excavation in soft soil can lead to ground settlements, affecting the surrounding environment and buildings. Broere and Festa [30] proposed a theoretical model which linked soil displacements to the dynamic and geometrical characteristics of TBMs. And the amplitude and spatial distribution of the soil displacement can be then obtained with data records. Geological information is a key factor in safety management of tunnel construction, and different methods including ANN [31] and support vector classifier (SVC) [32] has been proposed to predict geological formation based on TBM operating data. Shi et al. [33], [34] proposed a fuzzy c-means algorithm to cluster TBM monitoring data. Formation information and operating behavior can then be derived from data clusters. Besides, Salimi et al. [35] constructed a decision tree model to develop a rock mass classification system. The system can provide a classification criterion for geological data in TBM performance evaluation. These studies were used in the process of automation in shield tunneling, demonstrating the feasibility and practicality of TBM monitoring data analysis.

B. APPLICATION OF DM IN AEC INDUSTRY

DM is a collection of data analysis technologies that are designed to extract unknown knowledge from large data sets. According to mining tasks, DM algorithms can be further divided into several categories, such as classification, regression, clustering, and discovery of associations [36]. The general process of DM includes data selection, data preprocessing, application of DM algorithms, and interpretation of mining results.

The Architecture, Engineering and Construction (AEC) industry has accumulated a large amount of data for a long time. The introduction of sensors and Internet of Things (IoT) has further increased the amount of data [37]. Some important reasons for practitioners to adopt DM include sustainability, improving process, acquiring intelligence, identifying costs and reducing costs [38]. Currently, DM has been applied at all stages of the building's lifecycle.

In the design phase, Kim et al. [39] used DM techniques to evaluate building design options. Energy-related impacts of multiple building components, including roofs, walls, heating, ventilation and air conditioning (HVAC) system and building orientation were analyzed. The key patterns were extracted to help the project team to improve building design. Petrova et al. [40] also used DM methods to predict the energy saving outcomes of the building. A semantic graph was constructed to extract information from textual design data. The association rule algorithm was then applied to investigate the connection between design options and building performance after construction.

In the construction phase, DM was used to evaluate construction risks, find construction defects and predict project costs. Cheng et al. [41] used the decision tree model to predict the probability of casualties. The involved parameters of the model include accident type, project type, the age and gender of the workers, etc. Ayhan and Tokdemir [42] used ANN and Case-Based Reasoning (CBR) methods to evaluate the risk of construction in megaprojects. Heterogeneous data and multiple parameters were considered in model construction, and a Latent Class Clustering Analysis (LCCA) process was applied to reduce the data size. The model finally outputs the risk degree of certain construction behavior. Lin and Fan [43] used the association rule algorithm to analyze the relationship between construction defects and inspection indicators. The rules could help inspectors pay more attention to key indicators. In the operation and maintenance (O&M) phase, DM has shown promising results in the field of energy efficiency analysis. Geronazzo et al. [44] applied the decision tree model to explore what climate response strategies will create a comfortable indoor environment. The structure of the generated decision tree could clearly explain the impact of each factor. Ashouri et al. [45] used a hybrid data mining

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

method to give advisory on energy saving behaviors. Association rules and cluster analysis were used to analyze occupants' energy consuming records and give recommendations. ANN was then involved to predict the amount of saved energy. In addition to energy analysis, DM methods could be also applied to maintenance records. Peng et al. [46] and Wen et al. [47] conducted an innovative research on extracting hidden patterns in maintenance records of large public buildings using hybrid DM methods. The extracted rules could be provided to building managers as recommendations.

C. DISCUSSION

A variety of models and algorithms has been applied to tunnel construction data to model the complex nonlinear relationship of TBM parameters. These data-driven methods have successfully improved tunneling efficiency and reduced labor hours in many cases. However, there are still many deficiencies in the current literature. Most research on TBM performance prediction are not designed for real-time analysis. The data used in these studies are mainly limited to the rock property and TBM geometry with few real-time monitoring data involved. The analysis process also involves a lot of manual participation and cannot meet the efficiency requirements of real-time analysis. As a result, these studies can only be used for evaluations before or after a project, instead of real-time analysis of TBM status. Besides, tunnel construction will record numerous parameters in real time. But the selection of parameters is mostly based on experiences and insufficient understanding of the meanings of those parameters. It also happens that related parameters are not considered, and thus the model accuracy is reduced. In addition, the formation parameters of TBM have an important influence on the tunneling performance. However, at present, formation data can only be obtained by borehole sampling, which makes formation information between sampling points inaccurate. Since the performance of DM depends on the quality of data, this inaccurate data will reduce the accuracy of DM. Furthermore, these studies are mostly designed to solve a specific predictable problem with a single method. A systematical study on TBM parameters and their relationships can seldom be found in the literature. The combination of multiple analysis methods to solve complex problems is also not common. Therefore, new methods need to be designed to consider the influence of various parameters based on real-time TBM monitoring data.

Many research literatures have demonstrated the effectiveness of DM in AEC projects. However, DM is still considered a semi-automated process because of the manual participation in the preprocess and postprocess steps [48]. Excessive manual participation will reduce the efficiency of DM and bring uncertainty to mining results. Some researchers have proposed methods of using other DM techniques to replace manual work in the current DM process, and this is the core idea of hybrid DM approaches [49]. A typical hybrid DM method uses one or more DM algorithms as preprocessors to find useful sub datasets for the main DM algorithm [46]. These sub datasets have better data quality and more obvious features, and are expected to provide better results [50]. In fact, some studies have proved benefits from a hybrid DM approach [42], [43]. Unsupervised algorithms, such as association rules and cluster analysis, can be performed first, followed by supervised algorithms including classification and prediction. It is believed that early exploratory analysis will help make sense of data patterns, identify data characteristics, and point directions for in-depth analysis [44].

Therefore, a hybrid DM method is proposed in this paper to solve the above problems and ensure the safety of TBM construction. By introducing the hybrid DM approach, multiple deficiencies can be addressed: (1) Analysis methods with heavy labor participation cannot meet the efficiency requirement of real-time data analysis. The hybrid DM method replaces manual work in data preprocessing and postprocessing with automated DM algorithms, which improves working efficiency and enables real-time data analysis. (2) The parameter selection based on experience will reduce the accuracy of analysis results. The proposed hybrid DM method replace this process with an association rule discovery algorithm in the preprocess step. Parameters can be then selected based on association rules instead of experiences to get a better result. (3) Inaccurate formation data will affect the effect of DM. To solve this problem, a classification algorithm is designed in the hybrid DM framework to calibrate formation information. (4) Most studies are designed to solve single specific problems. By applying the hybrid DM algorithm, multiple safety management objectives in tunnel construction is achieved. It should be noted that these objectives are not independent. The results of the previous analysis tasks can provide support for the subsequent DM processes. The proposed approach benefits from the combination of multiple DM methods and obtains better accuracy and efficiency of data analysis.

The characteristics of TBMs are also considered in the construction of the hybrid DM method. These characteristics will be used as priori knowledge to improve the general DM process. The framework of the hybrid DM method will be detailed in the following section.

III. FRAMEWORK OF THE HYBRID DM APPROACH

The framework of the proposed approach is presented in Fig. 1. Three objectives in different aspects of TBM safety management were achieved by the proposed method. A unified database that integrated monitoring data and formation information was first established. Three DM methods towards different safety management goals were then performed to carry out the analyses of real-time TBM monitoring data. Each method followed the general process of DM including data collection, data preprocessing, core algorithm application and result interpretation [51].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

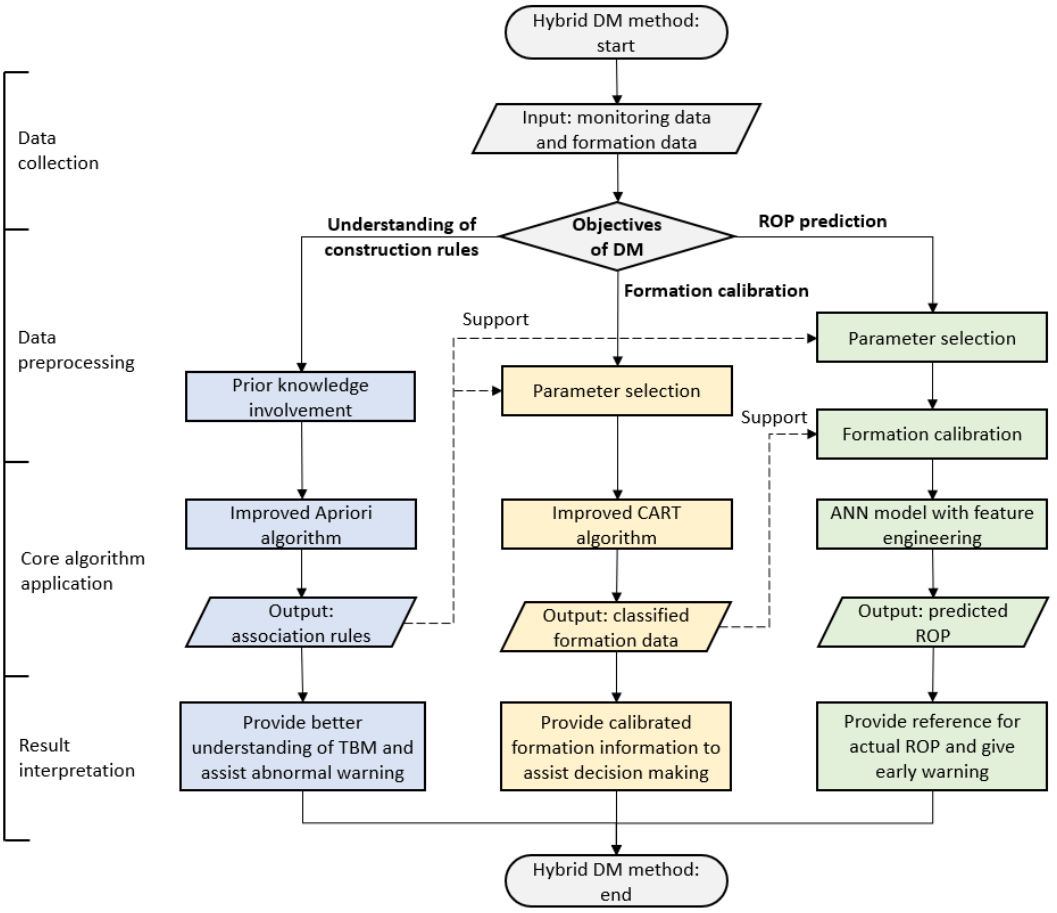


FIGURE 1. **Framework of the hybrid DM method**

In order to provide people with a better understanding of TBM operating laws, association rule discovery was first performed. Monitoring data with multiple parameters was input into the model, and the association rules among parameters were extracted as output. Meanwhile, the prior knowledge of TBM parameters was considered to improve the calculation efficiency. The mining results reveal the relationship among TBM parameters, and could be used to give early warning about abnormal changes as acquired experiences. When the change of TBM parameters is inconsistent with association rules, anomalies may exist and early warnings should be given. The extracted rules can also be used for parameter selection for the following procedures. Due to the complex interaction between soil and TBM, factors affecting dependent variables are hard to be fully considered. Association rules reveal potential relationships among TBM parameters, which can be used to select input parameters for DM models. For example, when determining the input parameters of the formation classification model, the involved parameters only need to be looked up in the association rules that contain the formation.

Classification analysis of formation was subsequently performed to address the problem of inaccurate formation information. Input parameters of the classification model can be determined by association rules to get a better classification result. The trained classification model could then be used to provide people with current formation information which is essential in the decision-making process for TBM safety management. When the calibrated formation differs from borehole sampling results, operators should conduct careful inspections to ensure safety and take corresponding construction strategies. The calibrated formation can also provide more accurate data for ROP prediction, and is expected to improve the performance of ROP prediction model.

Finally, an ANN was selected to carry out real-time ROP prediction. The parameter selection of the model was also based on association rules, and the calibrated formation data was applied to improve the accuracy. During the modeling process, the characteristics of monitoring data were considered, and a special network structure that considered two adjacent monitoring records was designed. The predicted ROP could be used as a reference for the actual monitoring data and help with construction decisions.

A comparison between the proposed method and some related data-driven researches in the field of TBMs is shown

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

TABLE 1
COMPARISON BETWEEN THE PROPOSED METHOD AND RELATED RESEARCHES

	Objectives	Data Types	Parameter Selection	Models
Yagiz et al. [19]	ROP prediction	Formation data	By simple regression	Use nonlinear regression and ANN independently
Acaroglu [27]	Load prediction	Formation data and machine properties	Selected manually	Fuzzy logic regression
Mahdevari et al. [24]	ROP prediction	Formation data and machine properties	Not mentioned	Support vector regression
Zhang et al. [52]	Fault diagnosis	Failure and maintenance records	Selected manually	Integrate DFT and Bayesian network as a hybrid model
Broere and Festa [30]	Settlement prediction	Real-time monitoring data	Selected manually	Force Analysis Model
Sun et al. [29]	Load prediction	Formation data and operation data	Not mentioned	Random forest
Proposed method	Association extraction, formation classification and ROP prediction	Formation data and real-time monitoring data	By association rules	Integrate association rule, decision tree and ANN as a hybrid model

in Table 1. The contributions of the proposed approach include: (1) Aiming at three objectives of TBM safety management, a hybrid DM method is proposed to improve mining process and achieve multi-objective analysis (2) Both formation data and real-time monitoring data were involved, and the formation data were calibrated to improve accuracy; (3) Association rules are adopted in the parameter selection process to improve the accuracy of DM.

IV. CONSTRUCTION LAW EXTRACTION WITH ASSOCIATION RULES

A. TBM PARAMETERS AND THEIR RELATIONSHIPS

During the excavation process, the soil, shield and machinery interact with each other, resulting in a complex relationship between TBM parameters. Some qualitative descriptions can be given by experience although it is difficult to describe these relationships in mathematical formulas. For example, the thrust and the cutter torque will affect the ROP and the cutter head speed as well. The strength of the soil formation determines the required thrust and torque to maintain a constant tunneling speed. And the mud circulation will change the nature of the formation, making the soil softer and easier to cut through. These experiences have been validated by a large number of engineering practices, and are correct in most cases. TBM operators can use this knowledge to judge the safety status of the machine and warn about abnormal situations that are not compliant with experiences. In fact, there are many hidden unknown construction laws that have not been discovered and summarized. Similar to people's experience, these laws can be applied to judge anomalies and assist in the safety management of tunnel construction. Therefore, an improved association rule discovery algorithm is developed in this section to discover hidden relationships among parameters and extract construction laws. The extracted rules can assist operators with risk identification and ensure construction safety.

The extracted association rules can also help with the parameter selection process of the other two DM algorithms. The complex interaction between soil, shield and internal machinery makes it difficult to model the relationships among TBM parameters. As a result, the factors affecting a parameter are hard to be fully considered. In previous studies, input variables for TBM data analysis are mainly based on people's experience, such as inherited from previous models or determined by domain experts. This method is quite useful with a small number of parameters. However, when the number of parameters is large, determining relationships between parameters manually will become inefficient and error-prone. A novel method for parameter selection based on association rules is proposed in this paper. Parameters that exist in the same association rule are considered relevant and selected as candidates for parameter selection. The process of parameter selection is detailed in the following two sections.

B IMPROVED ASSOCIATION RULES FOR CONTINUOUS DATA

The association rule algorithm is developed to find all the strong association rules among the attribute combination, where both support and confidence requirements needs to be met. Originally, the support requirement can be indicated as (1):

$$Support\left(X \Rightarrow Y\right) = \frac{\left|X \cup Y\right|}{\left|D\right|} \times 100\% \geq S_{min} \tag{1}$$

where X and Y are item sets with certain parameters taking a specific value; $X \Rightarrow Y$ is the association rule where X is the result of Y; $|X \cup Y|$ is the number of events where X and Y occur simultaneously; $|D|$ is the total number of events; and S_{min} is the minimum threshold for support set by user. Similarly, the confidence requirement can be represented as (2):

$$Confidence\left(X \Rightarrow Y\right) = \frac{Support\left(X \cup Y\right)}{Support\left(X\right)} \times 100\% \geq C_{min} \tag{2}$$

where X and Y have the same meaning as in (1), and C_{min} is the minimum threshold for confidence determined by user.

Association rules are originally designed to describe relationships among parameters with discrete values. However, most of the TBM monitoring data is time series

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

data with continuous values. To apply the association rule algorithm to these data, an improved association rule is proposed in this paper. Suppose A and B are two time series parameters with n records:

$$A=\{a_1,a_2,\cdots a_{t-1},a_t,\cdots a_n\} \tag{3}$$

$$B=\{b_1,b_2,\cdots b_{t-1},b_t,\cdots b_n\} \tag{4}$$

where a_t and b_t is the data recorded at a certain moment. At the moment t, parameter A or B taking a specific value i is defined as

$$f(x_t,i,m)=\begin{cases} true, & \frac{i}{m}\le\frac{x_t-x_{min}}{x_{max}-x_{min}}<\frac{i+1}{m} \\ false, & other\ conditions \end{cases} \tag{5}$$

where x_t stands for a_t or b_t, and m is a manually specified value which is set to 3 in the proposed method. X and Y are events that A and B takes a specific value i and j. The support value of parameter X and Y is calculated as:

$$Support(X\Rightarrow Y)=\frac{|f(a_t,i,3)\,f(b_t,j,3)|}{|D|}\times 100\%\ge S_{min} \tag{6}$$

Concisely, the support value of $X\Rightarrow Y$ is the proportion of records where A and B are in the specified value range. The calculation of confidence is still based on (2). Similarly, the improved association rules can be further extended to relationships between continuous values and discrete values. Assuming X is the event that a continuous variable A takes value i, and Y represents a discrete parameter taking a specific value, the support value can be calculated as:

$$Support(X\Rightarrow Y)=\frac{|f(a_t,i,3)\cup Y|}{|D|}\times 100\%\ge S_{min} \tag{7}$$

In this case, support is the proportion of records where A is in a certain value range and Y takes a certain value. In brief, the support requirement guarantees the frequency of association rules, and the confidence requirement ensures the reliability of association rules.

C IMPROVED APRIORI ALGORITHM

According to the definition, the association rule algorithm can be divided into the following two steps:

(1) Mining process 1 is to find all item sets which meet support requirement, and those item sets are called the frequent item sets.

(2) Mining process 2 is to generate association rules in all frequent item sets with confidence requirement as the criterion.

Specifically, the Apriori algorithm [53] is a classical algorithm in association rule extraction. The flow of the algorithm is shown in Fig. 2. The Apriori algorithm uses an iterative method to generate candidate item sets with k + 1 parameters from frequent item sets with k parameters. The frequent item sets with k + 1 parameters will be output after support test. And association rules with k + 1 parameters can then be obtained by applying confidence test.

The efficiency of the original Apriori algorithm is relatively low. During the generation of frequent item sets with k + 1 parameters, the data set needs to be scanned for every combination of parameter item, which leads to a large amount of time consumption. Therefore, an improved Apriori algorithm is proposed to improve computing efficiency. As shown in Fig. 2, the improved algorithm uses prior knowledge to filter candidate item sets and reduce the combination of parameter items. Specifically, TBM parameters are divided into groups based on experiences. There is no relationship between the parameters of different groups, and the association rules where parameter come from different groups cannot meet the confidence requirements. Candidate item sets with parameters from different groups can then be filtered. For example, there may be a potential relationship between the jack thrust in different directions and the orientation parameters. But the total thrust of the machine, which is the sum of all jack thrusts, must not be related to its orientation due to the spatial symmetry. Therefore, the total thrust and the orientation parameters can be divided into two different groups, and frequent item sets with total thrust and orientation can be filtered.

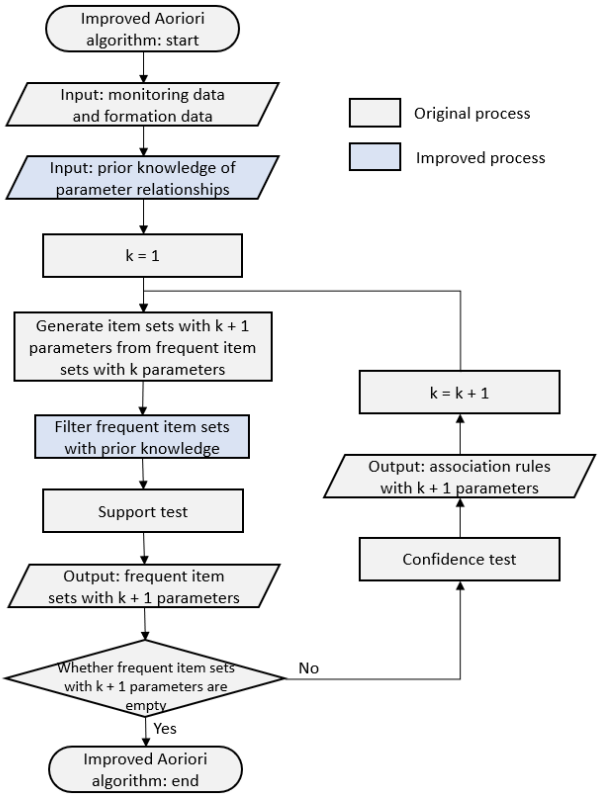


FIGURE 2. Mining processes of the improved Apriori algorithm

The two key parameters of the algorithm are the support and confidence threshold. Generally, the higher the support, the higher the frequency of association rules, indicating that the rules are more universal. The higher the confidence, the higher the accuracy of the association rules, indicating that the rules are more accurate. Increasing the support and confidence thresholds will result in higher quality of association rules, but the number of rules that meet the

requirements will be reduced. A trial and error process were carried out in this paper to ensure that sufficient quantity and quality of association rules are generated.

V. REAL-TIME REFINEMENT OF FORMATION BY CLASSIFICATION ANALYSIS

The formation property is one of the most important construction parameters during the shield tunneling process, affecting a series of construction processes such as cutter head cutting, mud circulation and simultaneous grouting. As a result, the formation properties are associated with many parameters in TBM monitoring data, and become a key factor in the safety management of tunnel construction. Almost all studies in the literature have considered geological parameter as input variables to the analysis model. However, the accuracy of the formation is still low because at present, the common approach to obtaining geological information is drilling exploration before construction. The method samples boreholes at a certain distance to obtain the geological information of a point, and linearly interpolates between sampling points to estimate the geological conditions of all locations [29]. Due to the lack of direct survey data, the exploration results between sampling points are relatively rough, and there are often unexplored formations encountered during construction. The accuracy requirements of the construction and monitoring data analysis are often difficult to meet in complicated formation. Therefore, a classification method was proposed in this section to estimate the current formation through the real-time monitoring data of TBM. The refined formation can improve the accuracy of geological data and help constructors adopt the correct safety management strategy. The calibrated formation data can also be integrated into the hybrid DM framework to improve the performance of subsequent mining algorithms.

A. PARAMETER SELECTION

In the literature, formation parameters have only been used for independent variables rather than dependent variables in the analysis model. One of the main reasons is that there are so many parameters related to the formation properties. Many parameters can be affected by the properties of the formation, and the formation parameters can be thus used as influencing factors to participate in the analysis. However, the estimation of the formation needs to determine all the important parameters related to the formation properties, which is rather difficult considering the large number of TBM parameters. The application of association rules could provide a solution to the problem. The relationship between variables can be explained through the association rules, and the parameters related to the stratum can be then determined.

In this study, the association rules containing the formation data is selected. The other attributes in the association rules were used as candidates. Further sensitivity analysis was conducted to improve the efficiency of model training, while the remaining parameters were used for formation evaluation. In addition, since the TBM monitoring data and the geological data are heterogeneous, a data integration process needs to be carried out before the algorithm starts, which will be detailed in the case study.

B. CLASSIFICATION AND REGRESSION TREE

A decision tree is selected as the formation classification model in this approach. The model is a tree-based classification structure, where the internal nodes are classification criteria based on the monitoring data, and the leaves are the predicted formation results. The classification model is constructed based on the classification and regression tree (CART) algorithm [54]. The CART model could map input data to multiple values and is suitable for multi-result formation classification. In general, three following steps are taken to construct a CART model:

(1) Decision Tree Generation: A decision tree is built with the training set, and each node in the decision tree is expected to be as pure as possible. The purity here means that all samples belong to the same class and can be judged by the Gini index as defined in (3):

$$i(t)=\sum_{i\neq j} p(i\,|\,t)\,p(j\,|\,t) \tag{8}$$

Where t is a given node, and $p(i \mid t)$ and $p(j \mid t)$ are proportions of records belonging to class i and j in node t respectively.

(2) Decision Tree pruning: The nodes of the generated tree are sequentially pruned to avoid overfitting. The pruning order is calculated based on the complexity cost:

$$\alpha=\frac{C(t)-C(T_t)}{|T_t|-1} \tag{9}$$

where t is a given node, T_t is the sub-tree rooted at t, $C(T)$ is the sum of Gini index of all nodes in tree T, and $|T_t|$ is the number of nodes in T_t. Nodes with smaller complexity cost will be preferentially pruned. The final pruning result is an ordered sub-tree sequence with a decreasing number of nodes.

(3) Decision Tree evaluation: The validation set is used to find the tree with the best classification result.

The key parameters of the CART model include maximum tree depth, minimum number of samples required for node partition, maximum number of leaf nodes, and minimum number of leaf nodes samples. These parameters are used to control the tree generation process. If the tree is generated without restriction, the model will become very complicated and take a large amount of training time. At the same time, there is a high probability that the performance of the model will deteriorate due to overfitting. In this paper, a grid search method was applied to exhaustively search each combination of the preset parameter values. For each parameter value combination, a ten-fold cross validation process was performed. The parameter combination with the best performance will be used to generate the final decision tree.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

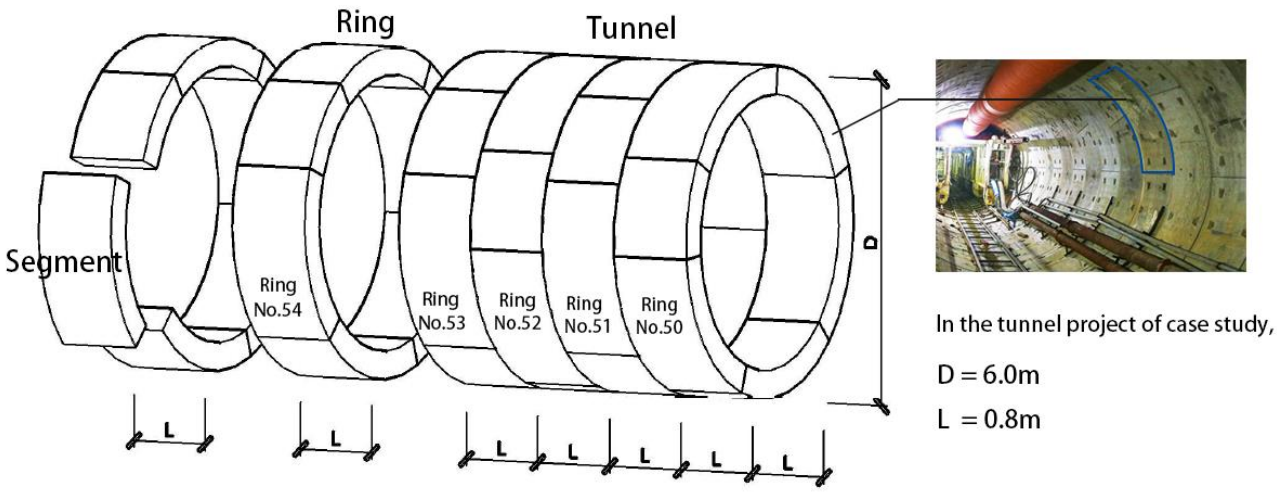


FIGURE 3. **Ring in tunnel construction**

C. IMPROVED CART MODEL WITH ENSEMBLE RESULTS

The proposed classification model made improvements to the CART model based on the characteristics of TBMs and the idea of ensemble learning. An ensemble process is designed to get more accurate results. Specifically, data from the same "ring" are grouped together and vote for the final result. As shown in Fig. 3, the ring is the basic unit for construction progress in tunneling, and the position of the TBM can be determined by the ring number. It takes several hours for a TBM to install a ring piece, and hundreds to thousands of data can be recorded during the period. Each piece of ring is less than one meter in length, and within such a short distance, the formation change can be ignored in most cases. The ideal situation is that all data points of the same ring are classified into the same formation for classification results, but there will always be a misclassification. Therefore, the dominant formation results in the data set are taken as the ensemble results of the ring to avoid the influence of some erroneous data points, and the final estimation results are presented in ring segments.

A quantitative indicator to measure the quality of the ensemble classification results is proposed. The proportion of the dominant formation result, which indicates the number of misclassified data points, can be used to judge the quality of classification results. The proposed indicator applies the Gini index shown in (3), where t is the set of all the data in a ring piece, and $p(i \mid t)$ and $p(j \mid t)$ are proportions of classification results in formation i and j respectively. Ideally, a perfect model has all data in the same result, where the Gini index is 0. A large Gini index value means the dominant classification result is weak and there are many misclassified data, which indicates that the classification model is not fully trained. In this case, the data from this ring can be added to the training set and the model needs to be retrained.

VI. ROP EVALUATION BY ARTIFICIAL NEURAL NETWORK

As illustrated in the section of literature review, the evaluation of ROP is one of the most important topics for a tunneling project. Various methods have been proposed to estimate the ROP. However, most of these studies made their predictions before construction begins for machine selection and construction scheme optimization, and the data involved were mainly incomplete or inaccurate rock properties explored by borehole sampling. In fact, evaluation of the real-time tunneling speed is also important. Although the TBM records its actual ROP during tunneling, the judgement of whether this value is normal or not still needs manual participation. Usually this process is carried out by the TBM operator, and there are no quantified indicators for reference, bringing risk and uncertainty to the tunnel construction. A threshold of ROP can be preset for consulting, but it is too complicated to determine and consult the value for every working condition. Therefore, an ANN method was proposed in this section to review the real-time ROP. Historical monitoring data under normal construction conditions are applied to develop the ANN model. And real-time ROP can be then used to evaluated real-time ROP through real-time monitoring data. The evaluated ROP can be applied as a reference for the actual monitoring parameter, providing assistance for construction personnel to make decisions.

A. ANN IN TUNNEL CONSTRUCTION

ANN is a kind of mathematical model that imitates the structure of human brain and performs distributed parallel calculations. With strong learning and nonlinear fitting ability, ANN has been widely used in various disciplines as one of the most effective tools for solving numerical simulation problems [55]. There are also many applications of ANN in the field of TBM data analysis. Several kinds of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

neural networks have been developed in the literature such as backpropagation (BP) neural networks, Hopfield networks, Kohonen networks, Elman networks, etc. Among them, BP neural network is most widely used for its simple structure and strong adaptability.

Recently, the concept of deep learning has emerged and greatly improved the learning ability of ANN [56]. Deep neural networks usually have multiple hidden layers to improve prediction ability, and has been successfully applied in multiple disciplines. Some typical deep neural networks include the Convolutional Neural Network (CNN) [57], Generative Adversarial Network (GAN) [58] and Recurrent Neural Networks (RNN) [59]. In tunnel engineering, RNN has been applied to evaluate TBM performance [60] and predict TBM operating parameters during next period [59]. Erharter et al. [61] compares the performance of two ANNs including multilayer perception (MLP) and RNN in TBM data classification, and concludes that both models are capable of classifying TBM data. Besides RNN, CNN also has strong fitting ability and realizes automatic feature extraction through hidden convolution layers, which is expected to achieve good performance in ROP evaluation.

In this section, two network models for ROP evaluation including BP neural network and CNN are constructed separately. In the case study, modeling results of the two models are compared, and the usage of predicted ROP is discussed.

B. FEATURE ENGINEERING BASED ON TBM CHARACTERISTICS

The network model was integrated into the hybrid DM framework to improve its performance. Mining results of Apriori and CART were involved in the data preprocessing process of ROP evaluation. To determine the input variables from numerous monitoring parameters, association rules were used for parameter selection. The extracted rules reveal hidden relationships among TBM parameters, and those parameters related to ROP were selected as input variables. And the geological parameters were modified by the classification results and represented by three mechanical indicators. The classification results refined inaccurate formation data between boreholes, and was expected to improve the accuracy of the ROP evaluation model. In this study, six parameters were determined as the input parameters for the model, as shown in Fig. 4.

To improve the training effect of ANN, feature engineering is carried out based on the characteristics of TBMs. Feature engineering is the process of extracting features from raw data to improve algorithm performance. Specifically, the actual ROP and its influencing factors of the last record are selected as features and input to the model to calculate the current ROP, as shown in Fig. 4. The feature is designed to consider the influence of unquantified factors such as the operation of the TBM operator. ROP cannot be perfectly evaluated due to these factors. But the prediction errors of adjacent records can be considered roughly unchanged because some of these factors, such as operating patterns of the same TBM operator, rarely change. Therefore, features from the previous record can be designed to estimate the errors caused by unquantified factors. And the performance of model can be improved by these features.

C. STRUCTURE OF BP NEURAL NETWORK

There are 10 input parameters including parameters from raw data and designed features. The number of neurons in the input and output layers was equal to the number of variables, which were 10 and 1 respectively. And the structure of the hidden layer was carefully designed for its crucial impact on the prediction results., A simple hidden layer will lead to reduction in fitting ability, and a complex layer will result in overfitting and long training time. In the case study, the neural network structure $(10 \times 12 \times 1)$ was selected after parameter adjustment, as shown in Fig. 4. Other factors determine learning outcomes include learning rate and activation functions. Low learning rates can reduce neural network training efficiency, while excessive rates may lead to fluctuation. Thus, the adaptive moment estimation (Adam) [62] optimizer with adaptive learning rate is applied in this study. The hyper parameter β_1 and β_2 in Adam follow the default values 0.9 and 0.999 which are suggested in the original paper. The original learning rate is determined by a parameter adjustment process. Some alternative activation functions include Sigmoid, Tanh and Relu. These functions are all tested in a trial and error to determine the best parameter.

D. STRUCTURE OF CNN

CNN is a typical deep neural network with multiple convolutional layers to achieve automatic feature extraction. The most common and successful application of CNN is image recognition. However, ROP evaluation has different characteristics with image recognition. Image recognition is a classification task with two-dimensional input data, while ROP evaluation is a regression task with one-dimensional data. Therefore, the structure of CNN needs to be carefully designed.

As illustrated in Fig. 5, the proposed CNN model has 6 weight layers with 39,425 weight parameters. The input parameters are also determined by association rules and are the same as the BP neural network. Four convolution (conv.) layers are designed to extract high-dimensional features from the input data. Different from image recognition, the convolution kernel here is one-dimensional to adapt to the format of monitoring data. As the dimensions of monitoring data are not high, small convolution kernels with size 2×1 are adopted. A max pooling layer with 2×1 filters is added after the convolution layers to down-sample the features. The dropout layer is subsequently added to solve overfitting problems caused by deep layers. Then, two fully connected (FC) layers with 128 and 1 channels are applied to the model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

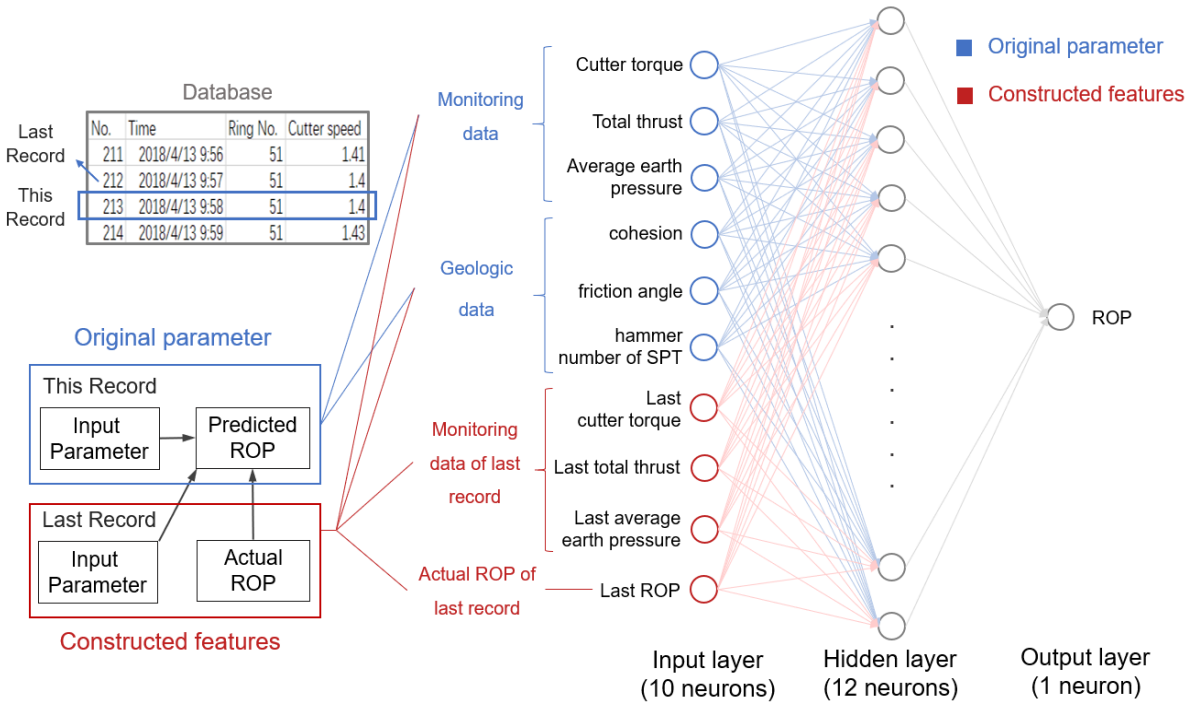


FIGURE 4. **Mining processes of the association rule algorithm**

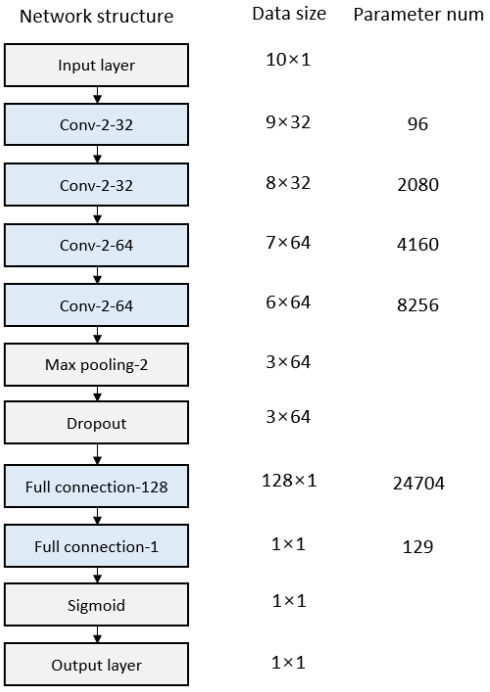


FIGURE 5. **Structure of the CNN model**

The size of the second FC layer is set to 1 to match the format of output. The last layer replaces the commonly used SoftMax function with the Sigmoid function to perform regression tasks.

The Adam optimizer is also applied in the CNN model to adjust learning rate with training process. The original learning rate of the optimizer is determined by a trial and error process. The activation function Relu is selected for convolution layers and FC layers to solve the vanishing gradient problem which is common in deep networks. The dropout rate in the dropout layer are determined by the degree of overfitting. The best value is derived from the parameter adjustment process.

VII. A CASE STUDY

A. PROJECT OVERVIEW AND DATA PREPROCESSING

The tunnel project locates in Guangzhou, the capital city of Guangdong Province in southern China. The tunnel is a twin-tube tunnel constructed for public transportation with a length of 16.1 km and a duration of 4 years. After the construction, the quantity of daily passenger-flow is expected to reach 1 million. Over ten TBMs work simultaneously to ensure the progress of the project. Monitoring data from a slurry TBM of the project was selected for data analysis. A schematic formation profile of its construction section is provided in Fig. 6. The diameter of the tunnel is 6m, and the depth of the tunnel varies from 10 to 13m. The formation in which the tunnel passed through consists of weathered limestone, silty clay, and sand of various particle sizes and densities (Fig. 6). The TBM is in a diameter of 6.26m and can provide a maximum thrust of 46000 kN and a maximum cutter torque of 5442 kN·m. Equipped with a sensing system, the TBM can automatically record real-time monitoring data with over 100 parameters at regular intervals and integrate them into an online database, from which the monitoring data used in this approach are derived in a spreadsheet form.

The database was established with the monitoring data of the tunnel for over three months. The number of raw data exceeds 40,000, and more than 10,000 valid data have been

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

retained after pre-processing. The criteria to select valid data here is that the data should be recorded during the tunneling process rather than stopping or segment assembling process and without a sensor failure. A total of 26 TBM parameters are selected, which include speed, mechanical properties, attitude and current. The parameter selection process is carried out with the help of TBM operators and managers. Only parameters potentially involved in the safety management process of tunnel construction are selected, because too many parameters will cause mining results of interesting parameters hidden in a large number of irrelevant parameters, and the efficiency of DM will also be reduced. The selected parameters are considered as the major indicators to judge the working status of TBM, and will be most concerned by TBM operators. DM on these parameters is expected to provide valuable information for safety management. Basic descriptive statistics of some major and frequently used parameters are given in Table 2. Besides real-time monitoring data, formation data are also involved in this approach to take the effects of formation into account.

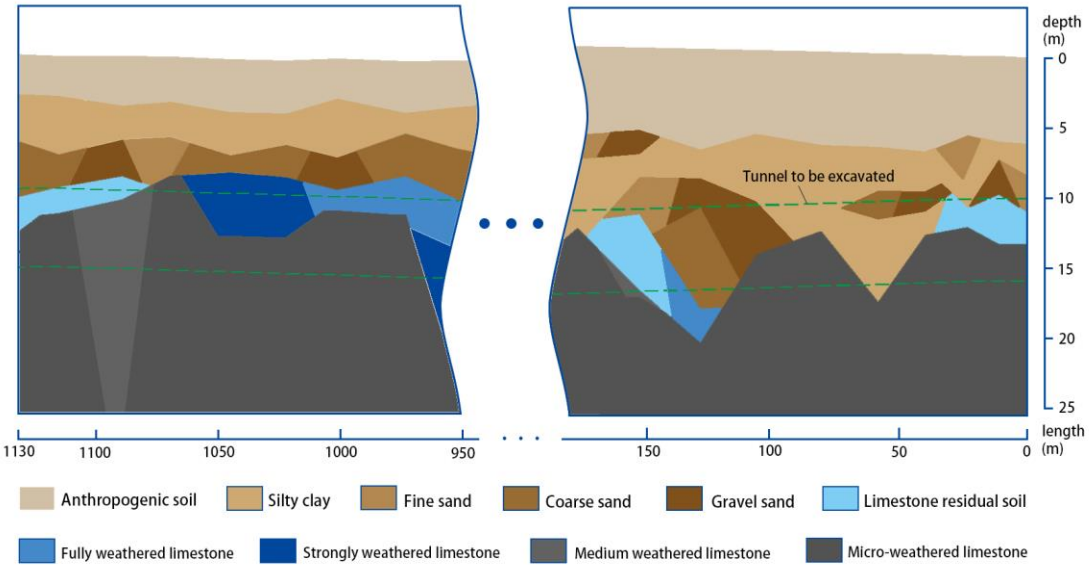


FIGURE 6. **Geological profile of the south tube**

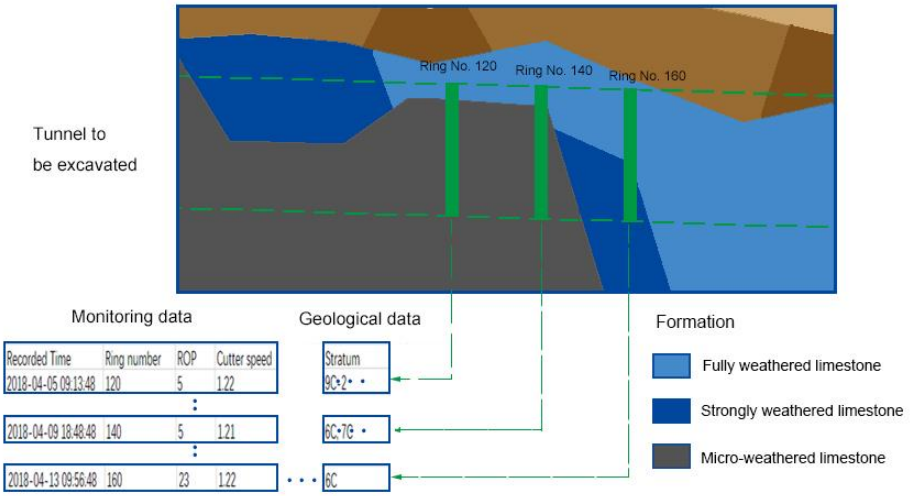


FIGURE 7. **The process of data integration**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

TABLE 2
BASIC STATISTICS OF KEY TBM MONITORING PARAMETERS

	ROP (mm/min)	Cutterhead speed (rpm)	Cutterhead torque (kN·m)	Total thrust (kN)	Cutter motor current (A)	Average earth pressure (bar)
Maximum	55.67	1.84	3704.0	21270.0	1076.0	2.56
Minimum	1.00	0.81	92.0	320.0	329.0	1.40
Average	16.19	1.37	727.1	10861.2	411.3	1.85
Median	15.33	1.36	595.0	10360.0	385.0	1.84
Standard deviation	8.43	0.12	426.0	2282.0	70.1	0.12

TABLE 3
PARAMETERS TO BE MINED IN ASSOCIATION RULES

No.	Parameters			Total
	Geological	Mechanical	Operational	
1	Cohesion, internal friction angle, hammer number of SPT	Total thrust, Cutterhead torque, Cutterhead speed	ROP, Cutter total current, average earth pressure, grouting amount, slurry pump current	11
2	-	Jack thrust in 4 directions, Jack speed in 4 directions	Front, middle and back horizontal deviation, front, middle and back vertical deviation Earth pressure in 4 directions	18

TABLE 4
DETAILS OF SOME TYPICAL ASSOCIATION RULES

No.	Condition	⇒	Result	Support	Confidence
1	Total thrust	⇒	ROP	0.631	0.882
2	ROP	⇒	Total thrust	0.631	<0.8
3	Cutter torque	⇒	Total thrust	0.346	0.964
4	Cutter speed, Cutter torque, Cutter total current, Total thrust, Average earth pressure, Slurry pump current, Cohesion, Internal friction angle, Hammer number of SPT		ROP	0.123	0.844

Three numerical indices include cohesion, internal friction angle, and hammer number of standard penetration test (SPT) were selected to measure the mechanical properties of formation. The distribution of rocks is obtained from the drilling experiment on site. And the mechanical dataset of the rock is established by both in-situ and laboratory tests. With a data integration method, the monitoring data and formation data are combined together to constitute the database for DM.

Data preprocessing was performed before DM. Specifically, two types of preprocessing technologies, namely data cleaning and data integration, were adopted in the approach. The task of data cleaning was to clean up invalid or erroneous data in the database. These data accounted for a large part of the database since the TBM did not always operate in the state of excavation. Data integration was required to integrate heterogeneous data into the same database because the monitoring data was stored in spreadsheets while the formation information was imported from geological profiles. The process of data integration is illustrated in Fig. 7. The integration is based on the correspondence in spatial position between the monitoring data and the formation. The geological information is first obtained from the profile and then recorded in the monitoring database. If a formation is dominant on the excavation section, it will be recorded separately, otherwise all formations on the section will be recorded together. The mechanical properties of the formation can then be derived from the formation code and marked in the database.

B. MODELING AND RESULTS

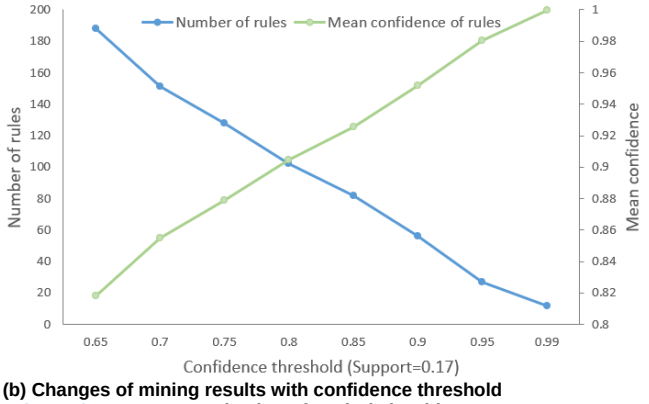
The proposed DM methods were implemented in Python 3.6. The toolbox scikit-learn [63] was applied to construct the decision tree and BP neural network. Another toolbox TensorFlow [64] was used to train the CNN model. The python programs are executed in a computer with the Windows system. The computer has a medium hardware configuration with a 2.8 GHz processor and 8 GB of RAM.

Association rule algorithm was first executed. A total of 310 valid data points from mining database were taken as input, with each dataset containing 29 attribute items to be mined. Based on the characteristics of TBM data, those parameters to be mined were divided into two groups as prior knowledge to improve mining process. As illustrated in Table 3, the first group of data is the holistic data without directionality, while the second group of data are the data related to direction and spatial distribution. Due to spatial symmetry, there is no correlation between the two sets of parameters. The support and confidence threshold are determined by a trial and error process. As shown in Fig. 8, as the support threshold increases, the mean support of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

association rules increases, which indicates that extracted rules are more universal, but the number of rules decreases. Similarly, increasing the confidence threshold will increase the mean confidence and improve the accuracy of association rules, but the number of association rules will be reduced. In the proposed approach, the support and confidence threshold were set to 0.17 and 0.8 to balance the number of rules, universality and accuracy. After calculating, a total of 102 strong association rules were found by the program, indicating the complex association between TBM parameters. Some typical mining results are detailed in Table 4.

(a) Changes of mining results with support threshold



(b) Changes of mining results with confidence threshold

FIGURE 8. Parameter selection of Apriori algorithm

The extracted association rules can provide people with a better understanding of construction laws and assist in safety management. The first association rule "total thrust ⇒ ROP" showed that ROP tended to be large when total thrust was large, which is a verification of people's experience. However, the reverse rule "ROP ⇒ total thrust" is not a strong association rule as the confidence is less than the threshold. The reason is that other parameters, such as formation property and grouting amount also have an effect on ROP. The conclusion is contrary to people's intuition, and is a correction of misunderstandings. The third rule "Cutter torque ⇒ total thrust" shows relationships of two parameters that are not directly related. The rule is reasonable because TBM is often in hard formation when cutter torque is large. And in order to keep penetration in hard formation, operators need to increase the thrust. The rule reveals laws that has not been grasped by people. These association rules verify people's experience, correct misunderstandings and reveal new laws, which can be used for quick judgement in on-site anomalies to ensure safety. Some complex association rules may contain more potential information and could be used for further numerical analysis. As shown in the third rule, the ROP was associated with nine parameters simultaneously. In fact, these parameters were the candidate parameters to be input to the neural network model for the prediction of ROP.

Classification analysis was then carried out after the determination of the relationship between TBM parameters. Seven parameters included ROP, cutter speed, cutter torque, total thrust, total cutter current, average earth pressure, and slurry pump current were selected as input variables according to the association rules. The training dataset consisted of the data in the location where strata appeared for the first time to simulate the on-site construction conditions, and had a total number of 1122. As shown in Table 5, a grid search process is carried out to determine the best parameters for CART model. The final tree is generated based on the best values. The depth of the generated tree is 12, with a total of 153 nodes. And the overall classification accuracy of the model reached 0.86.

The rest of the monitoring data with a number of 9025 data points was used to review the formation. The estimated results of these data were integrated into 126 ring slices. Table 6 shows the classification results for several rings. Taking the first record as an example, it means that the formation is a mixed formation of 4N-2 and 9C-2 when the TBM is in the construction of the 49th ring, where 4N-2 and 9C-2 here are the formation codes of silty clay and slight weathered limestone respectively. In the 53rd ring, the reviewed formation is different from the explored formation, and the formation data will be calibrated based on classification results. The Gini indicator is calculated to measure the quality of the ensemble results. A smaller Gini index indicates that there are more data voting to support the dominant result in the ring, and the result is more credible. In the worst case, the Gini index will reach a maximum value of 0.5. The gini indicator in the table is relatively small, indicating a high-quality classification result.

Further explanation of the mining results was performed by means of data visualization. In this paper, building information modeling (BIM) technology is adopted as a visualization tool for its convenience of data integration and good interactivity [65]. A plug-in was developed in a BIM software Autodesk Revit to enable automatic modeling of the estimated formation based on the geometric mapping rules. Fig. 9a shows the geological model drawn by the Revit plug-in. The BIM geological model uses a visual approach to transform the analysis results into 3D entities distinguished by color, which gives a more intuitive description on formation compared with the DM results in Table 6. The formation model based on the geological exploration was synchronously drawn to verify the prediction effect of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

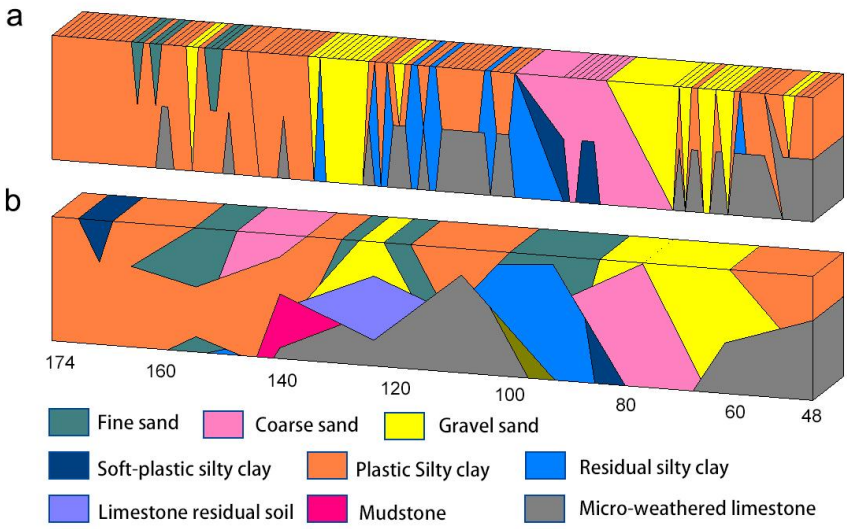


FIGURE 9. **Data visualization of case project: (a) Geological model based on classification results; (b) Geological model based on drilling exploration**

TABLE 5
PARAMETER ADJUSTMENT OF CART MODEL

Parameter	Adjustment Range	Best value
Maximum tree depth	{3, 6, 9, 12, 15, not limited}	12
Minimum number of samples required for node partition	{5, 10, 15, ..., 50, not limited}	40
Maximum number of leaf nodes	{50, 100, 150, ..., 500, not limited}	Not limited
Minimum number of leaf node samples	{5, 10, 15, ..., 50, not limited}	10

TABLE 6
EXAMPLES OF FORMATION CLASSIFICATION RESULTS

Ring No.	Formation obtained by classification	Formation obtained by exploration	Gini indicator of the ring
50	4N-2 9C-2	4N-2 9C-2	0.034
51	4N-2 9C-2	4N-2 9C-2	0.064
52	4N-2 9C-2	4N-2 9C-2	0.137
53	3-3 9C-2	4N-2 9C-2	0.149

The formation in the table is in the form of formation codes. The correspondence between codes and formation is: 3-3: Gravel sand; 4N-2: Silty clay; 9C-2: slight weathered limestone.

classification algorithm, as shown in Fig. 9b. The numbers below represent the rings of the corresponding position of the model. It could be concluded that the two models are aligned with each other generally, but different in some local parts. In the part where the model in Fig. 9b is processed by linear interpolation, the formation of the classification model is more detailed and varied, indicating that the classification model can find formation that haven't been discovered by exploration due to the sampling interval of the borehole.

ROP evaluation was then carried out based on ANN. The classification results of CART were applied to calibrate formation data, and the calibrated data were expected to improve performance of neural networks. The input variables of the two model were determined by association rules, and have been illustrated in Fig. 3. 80% of the data were selected randomly to train the models and the rest of the data were used as test set. The process of determining the model parameters is shown in Table 7. Note that the training epochs of CNN is set to 50 and the batch size is 128. The structure of BP neural network is a key factor determining learning outcomes, and the test results of some candidate structures are illustrated in Fig. 10. The performance of the model increases with the number of hidden neurons when the structure is simple. But when the scale of the hidden layer reaches a certain size, the performance will not be further improved. And increasing the number of hidden layers does not significantly improve the outcome of the model. Therefore, the structure with single hidden layer (10×12×1) is selected to generate the final model.

The prediction results of the two models are summarized in Table 8. Three indicators, including R-squared value (R^2), mean squared error (MSE), and mean absolute error (MAE) are selected to measure the performance of the model. These indicators can be calculated according to the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{10}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{11}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994115, IEEE Access

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (12)$$

where x_i and y_i are the i^{th} actual and evaluated value in the data group respectively, and $\bar{x}$ is the average of all actual values. It can be concluded that the prediction effect of BP neural network and CNN is approximately equivalent. The reason why CNN does not work better than BP neural network is that the dimensions of data are not high enough. The convolution operations are designed to extract hidden features from high-dimensional data. In the field of image recognition, the input data is two-dimensional data with hundreds of pixels in each dimension. However, TBM monitoring data is one-dimensional data, and the number of involved parameters is also small.

To verify the effectiveness of the proposed hybrid DM method, comparative experiments are conducted simultaneously. ANN models without formation calibration and feature construction are trained, and performances of these model are compared in Table 8. Without formation calibration, the performances of the two models are both reduced, indicating that the formation classification process in the hybrid DM method improves data quality and model accuracy. The fitting effect of models without feature construction is also reduced. Revealing the necessity of feature engineering. The results indicate that the accuracy of the ANN model can be improved by combining hybrid DM method and considering the feature of TBM dataset.

TABLE 7
PARAMETER ADJUSTMENT OF ANN MODEL

Models	Parameter	Adjustment Range	Best value
BP neural network	Model structure	{(10×i×1), (10×i×j×1)}, i, j = 2, 4, 6, ..., 20	(10×12×1)
	Learning rate	{0.1, 0.01, 0.001, 1×10^{-4}, 1×10^{-5}}	0.001
	Activation function	{Sigmoid, Tanh, Relu}	Sigmoid
CNN	Learning rate	{0.1, 0.01, 0.001, 1×10^{-4}, 1×10^{-5}}	0.001
	Dropout rate	{0.1, 0.2, 0.3, 0.4, 0.5}	0.2

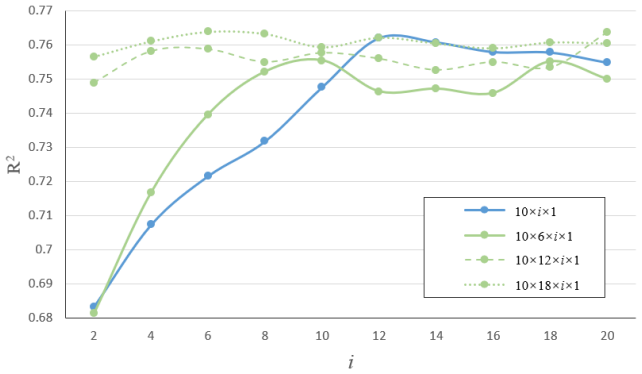


FIGURE 10. Model structure selection of BP neural network

TABLE 8
PERFORMANCE OF THE ANN MODELS

DM Models	R^2	MSE	MAE
BP neural network	0.762	3.109	16.792
CNN	0.745	3.295	18.687
BP neural network without formation calibration	0.713	3.464	20.50
BP neural network without feature engineering	0.618	4.019	26.724
CNN without formation calibration	0.736	3.302	18.587
CNN without feature engineering	0.678	3.585	21.914

The ROP output by neural network can be further compared with the actual ROP to evaluate the working state of the TBMs. A big difference between the two values indicates that the actual ROP exceeds the expected ROP threshold, and an early warning could be send to the TBM operator. The prediction results of the BP neural network are selected to illustrate this process for its best performance. The data is grouped according to the ring which it belongs to, and the difference in ROP is measured by MAE. In general, the difference between the predicted and actual rate is relatively stable, which fluctuates around 4 mm/min. However, there is a significant peak at the 130th ring, where the expected error reaches 18 mm/min. The location of the 130th ring is between two boreholes, and formation information is obtained by linear fitting. In fact, a solutional cave was explored at the nearby borehole, which needs to pay special attention to in tunnel construction. The large evaluation error indicated that the TBM was in an abnormal working state which was later proved to be caused by the scale of the cave exceeding expectations. Early warnings were sent out based on unexpected ROP, and operators suspended work and took inspection measures to ensure construction safety.

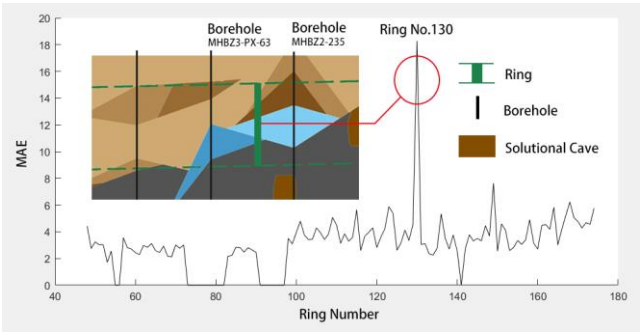


FIGURE 11. Prediction error of ROP for each tunnel ring

C. DISCUSSION

In this study, three data mining methods were investigated to analyze the real-time monitoring data of the TBM. Association rule mining was used to discover the relationships among parameters. Classification analysis was

applied to determine the formation where the TBM is located. And the ANN method helped evaluate the real-time ROP of the TBM. The hybrid DM method is expected to contribute to tunnel engineering, improving construction efficiency and ensuring safety. Association rule discovery can help construction managers better understand the relationship between tunneling parameters. The extracted association rules can verify people's experience and correct people's intuitive misunderstandings. Moreover, some hidden unknown relationships can also be discovered. Formation classification analysis can be used to supplement and refine formation information among sampling boreholes, providing a more accurate decision basis for tunnel construction. The ANN model can help operators control real-time ROP and ensure construction safety. The output ROP can be used as a reference for the actual monitoring value, and abnormal conditions can be found in advance to avoid danger. These improvements are made based on data-driven approaches, which is expected to have important implications for current management methods.

In the case study, more than 40,000 pieces of data recorded in a tunnel project were involved. Actually, this amount of data is still far from the true amount of big data. Only data recorded from one TBM was analyzed, while the megaproject has over ten TBMs working together. This indicates that the ability of the proposed approach to process massive data may not be fully tested. However, the mining algorithms used in this paper were designed for dealing with a large amount of data, and they took only a few seconds to process such amount of data. Furthermore, the purpose of the proposed method is to get the working status for a specific machine. Even if data from all TBMs are integrated together, algorithms for eliminating heterogeneity need to be executed. Finally, data from different TBMs need to be analyzed separately.

As mentioned in the literature review, the challenges of applying DM to TBM data mainly comes from its strict efficiency and accuracy requirements. The proposed approach provides a solution to the problem by combining hybrid DM methods and features of TBM monitoring data. During tunnel boring process, TBM records large number of parameters, resulting in numerous analytical works or domain experience in the parameter selection process. A novel parameter selection method based on association rules was proposed in this paper. An improved association rules discovery algorithm was applied as a preprocessing process in the hybrid DM system, and the range of related parameters to be determined will be greatly reduced. The discovered rules can also can be used in diagnosis of on-site anomalies to ensure safety. The inaccuracy of formation information is an important factor affecting the performance of TBM data analysis and safety of tunnel construction. A formation classification model was constructed in the hybrid DM method to solve this problem. In addition, based on the characteristics of TBM data, some modifications were applied to the general DM models for improvement. An extended association rule is proposed to apply association rule analysis to continuous values. And prior knowledge based on TBM characteristics is involved in the algorithm to reduce the amount of calculation. In formation classification, ensemble results in rings is proposed to improve the performance of the model. In the process of ANN construction, a feature involving adjacent data records was designed to improve the training effect. The proposed hybrid DM model are expected to meet the actual construction requirements of the DM efficiency and accuracy.

The monitoring data involved in this study were collected from a tunnel project, but in fact, massive data has been recorded in different projects of building industry. These data have different sources, formats, and practical meanings. The specific means of processing and analyzing them are also different. However, the characteristics of these data are similar. They are large, heterogeneous and noisy, but contain potentially valuable information for the project. Under these characteristics, traditional analytical methods, such as manual processing or statistical analysis, will no longer be accurate and efficient. DM is a kind of analysis methods for processing large amounts of data, and is suitable for processing data in megaprojects. Regardless of the types of projects, the basic process of DM is following the same workflow. The proposed methods in this paper will provide reference for DM applications in other projects.

After careful design and adequate training, the DM system is expected to achieve better accuracy than manual judgement and meet the standards of practical applications. However, there are still some limitations with the current methods. On the one hand, the amount of data available for modeling is insufficient. The systematic approach of recording construction data in building industry has only begun in recent years, and manual recording is still one of the main means of data acquisition, thus DM in building industry lacks sufficient high-quality data. This has been observed from error results in DM such as obviously unreasonable association rules and formation. On the other hand, manual work involved in the current DM process is still too much. The data interfaces of different TBMs and monitoring platforms are different, which makes it difficult for automatic data collection. The integration of heterogeneous data is also done manually. These labor-intensive tasks have increased the cost of data mining. In addition, further explanation is needed for DM results. The tabular results can only be understood for specialists involved in the data mining process, and are obscure for TBM operators. In the near future, the proposed method will be improved in the following aspects:

(1) The time complexity of the algorithm should be optimized and the model should be trained using more data. A program for automatically collecting monitoring data can be executed to obtain a steady flow of data.
(2) The TBM monitoring platform should be integrated. The trained model can be applied to the platform where the

modeling data comes from. The integration is expected eliminate the data transfer process and realize real-time data analysis. In addition, the platform can visualize the mining results and help explain the mining results.

(3) Other techniques can be combined to improve data mining effects. Some data-driven platforms, such as BIM and building automation systems (BAS), can provide massive mining data and visualization paths. The Internet of Things (IoT) technology is expected to provide solutions for data collection and remote control. The cloud platform can be also utilized to increase the computing ability of DM.

VIII. CONCLUSION

In this paper, a hybrid DM approach is proposed to achieve accurate and efficient real-time TBM monitoring data mining for safety analysis during tunneling construction. Three DM methods are combined together to improve mining process and extract useful patterns to support safety management: (1) Association rule discovery are applied to extract relationships among parameters, and the extracted rules are used for parameter selection for subsequent algorithms. (2) Classification analysis can be used to estimate the current formation, and the refined formation is used to improve data quality for ROP prediction. (3) ANN models are constructed to validate the current ROP based on the classified formation and real-time monitoring data. The proposed approach is validated by real-time monitoring data from a tunnel project in China, and the result shows its effectiveness in safety management. To provide people with a better understanding of TBM, 102 association rules are extracted by an improved Apriori algorithm. These rules reveal construction laws and can be used for early warning of anomalies as acquired experiences. To refine the formation between exploration boreholes, formation is classified by an improved CART model. The refined formation improves the accuracy of geological data and can help operators adopt corresponding strategies in different formation to ensure safety. To review real-time ROP, two ANN models are developed to evaluate ROP. The evaluated ROP can be used as a reference of the actual value and give early warning in abnormal situations. The proposed method provides a feasible way for analyzing real-time TBM monitoring data. The analysis results can provide reliable assistance for project decision, and lead to the improvements of construction methods. The proposed hybrid DM method shows good accuracy and efficiency and meets the requirements of actual tunnel projects. The analysis results can provide reliable assistance for project decision and safety management in tunnel construction.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (No. 51778336, No. 51908323) and the Tsinghua University-Glodon Joint Research Centre for Building Information Model (RCBIM). The authors also acknowledge the Metro Protection Department of the Guangzhou Metro for providing the data support and thank Prof. Lucio Soibelman (University of Southern California) for his valuable comments.

REFERENCES

[1] L. Ding, L. Zhang, X. Wu, M. J. Skibniewski, and Y. Qunzhou, "Safety management in tunnel construction: Case study of Wuhan metro construction in China," *Saf. Sci.*, vol. 62, pp. 8–15, 2014.

[2] Y. L. Zheng, Q. B. Zhang, and J. Zhao, "Challenges and opportunities of using tunnel boring machines in mining," *Tunn. Undergr. Sp. Technol.*, vol. 57, pp. 287–299, 2016.

[3] C. Koch, A. Vonthron, and M. König, "A tunnel information modelling framework to support management, simulations and visualisations in mechanised tunnelling projects," *Autom. Constr.*, vol. 83, pp. 78–90, 2017.

[4] J. Ninić, C. Koch, and J. Stascheit, "An integrated platform for design and numerical analysis of shield tunnelling processes on different levels of detail," *Adv. Eng. Softw.*, vol. 112, pp. 165–179, 2017.

[5] S. Isam and Z. Wengang, "Use of soft computing techniques for tunneling optimization of tunnel boring machines," *Undergr. Sp.*, 2020.

[6] D. Festa, W. Broere, and J. W. Bosch, "Kinematic behaviour of a Tunnel Boring Machine in soft soil: Theory and observations," *Tunn. Undergr. Sp. Technol.*, vol. 49, pp. 208–217, 2015.

[7] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 6, pp. 866–883, 1996.

[8] J. Chen, J. E. Taylor, and H.-H. Wei, "Modeling building occupant network energy consumption decision-making: The interplay between network structure and conservation," *Energy Build.*, vol. 47, pp. 515–524, 2012.

[9] M. Alber, "Advance rates of hard rock TBMs and their effects on project economics," *Tunn. Undergr. Sp. Technol.*, vol. 15, no. 1, pp. 55–64, 2000.

[10] J. Rostami, "Performance prediction of hard rock Tunnel Boring Machines (TBMs) in difficult ground," *Tunn. Undergr. Sp. Technol.*, vol. 57, pp. 173–182, 2016.

[11] J. Rostami and L. Ozdemir, "New model for performance production of hard rock TBMs," in *Proceedings of the 1993 Rapid Excavation and Tunneling Conference*, 1993, pp. 793–809.

[12] S. N. Cheema, "Development of a rock mass boreability index for the performance of tunnel boring machines.," Colorado School of Mines, Golden, CO, USA, 2001.

[13] A. Ramezanzadeh, J. Rostami, and D. Tadic, "Impact of rock mass characteristics on hard rock tunnel boring machine performance," in *13th Australian Tunnelling Conference 2008*, 2008, p. 213.

[14] O. T. Blindheim, "Boreability Predictions for Tunneling," The Norwegian Institute of Technology, 1979.

[15] F. J. Macias, "Hard rock tunnel boring: Performance predictions and cutter life assessments," 2016.

[16] J. Hassanpour, J. Rostami, M. Khamehchiyan, and A. Bruland, "Developing new equations for TBM performance prediction in carbonate-argillaceous rocks: a case history of Nowsood water conveyance tunnel," *Geomech. Geoengin. An Int. J.*, vol. 4, no. 4, pp. 287–297, 2009.

[17] A. Delisio and J. Zhao, "A new model for TBM performance prediction in blocky rock conditions," *Tunn. Undergr. Sp. Technol.*, vol. 43, pp. 440–452, 2014.

[18] Q. Gong and J. Zhao, "Development of a rock mass characteristics model for TBM penetration rate prediction," *Int. J. Rock Mech. Min. Sci.*, vol. 46, no. 1, pp. 8–18, 2009.

[19] S. Yagiz, C. Gokceoglu, E. Sezer, and S. Iplikci, "Application of two non-linear prediction tools to the estimation of tunnel boring machine performance," *Eng. Appl. Artif. Intell.*, vol. 22, no. 4–5, pp. 818–824, 2009.

[20] A. G. Benardos and D. C. Kaliampakos, "Modelling TBM performance with artificial neural networks," *Tunn. Undergr. Sp. Technol.*, vol. 19, no. 6, pp. 597–605, 2004.

[21] G. Javad and T. Narges, "Application of artificial neural networks to the prediction of tunnel boring machine penetration rate," *Min. Sci. Technol.*, vol. 20, no. 5, pp. 727–733, 2010.

[22] R. Mikaeil, M. Z. Naghadehi, and F. Sereshki, "Multifactorial fuzzy approach to the penetrability classification of TBM in hard rock conditions," *Tunn. Undergr. Sp. Technol.*, vol. 24, no. 5, pp. 500–505, 2009.

[23] M. A. Grima, P. A. Bruines, and P. N. W. Verhoef, "Modeling Tunnel Boring Machine Performance by Neuro-Fuzzy Methods," *Tunn. Undergr. Sp. Technol. Technol.*, vol. 15, no. 3, pp. 259–269, 2000.

[24] S. Mahdevari, K. Shahriar, S. Yagiz, and M. Akbarpour Shirazi, "A support vector regression model for predicting tunnel boring machine penetration rates," *Int. J. Rock Mech. Min. Sci.*, vol. 72, pp. 214–229, 2014.

[25] M. Z. Naghadehi, M. Samaei, M. Ranjbarnia, and V. Nourani, "State-of-the-art predictive modeling of TBM performance in changing geological conditions through gene expression programming," *Measurement*, vol. 126, pp. 46–57, 2018.

[26] K. Elbaz, S.-L. Shen, W.-J. Sun, Z.-Y. Yin, and A. Zhou, "Prediction model of shield performance during tunneling via incorporating improved Particle Swarm Optimization into ANFIS," *IEEE Access*, vol. 8, pp. 39659–39671, 2020.

[27] O. Acaroglu, "Prediction of thrust and torque requirements of TBMs with fuzzy logic models," *Tunn. Undergr. Sp. Technol.*, vol. 26, no. 2, pp. 267–275, 2011.

[28] D. Festa, W. Broere, and J. W. Bosch, "An investigation into the forces acting on a TBM during driving - Mining the TBM logged data," *Tunn. Undergr. Sp. Technol.*, vol. 32, pp. 143–157, 2012.

[29] W. Sun, M. Shi, C. Zhang, J. Zhao, and X. Song, "Dynamic load prediction of tunnel boring machine (TBM) based on heterogeneous in-situ data," *Autom. Constr.*, vol. 92, no. March, pp. 23–34, 2018.

[30] W. Broere and D. Festa, "Correlation between the kinematics of a Tunnel Boring Machine and the observed soil displacements," *Tunn. Undergr. Sp. Technol.*, vol. 70, no. July, pp. 125–147, 2017.

[31] J. Zhao *et al.*, "A Data-Driven Framework for Tunnel Geological-Type Prediction Based on TBM Operating Data," *IEEE Access*, vol. 7, pp. 66703–66713, 2019.

[32] Q. Zhang, Z. Liu, and J. Tan, "Prediction of geological conditions for a tunnel boring machine using big operational data," *Autom. Constr.*, vol. 100, pp. 73–83, 2019.

[33] M. Shi, L. Zhang, W. Sun, and X. Song, "A fuzzy c-means algorithm guided by attribute correlations and its application in the big data analysis of tunnel boring machine," *Knowledge-Based Syst.*, vol. 182, p. 104859, 2019.

[34] M. Shi, T. Zhang, L. Zhang, W. Sun, and X. Song, "A fuzzy c-means algorithm based on the relationship among attributes of data and its application in tunnel boring machine," *Knowledge-Based Syst.*, vol. 191, p. 105229, 2020.

[35] A. Salimi, J. Rostami, C. Moormann, and J. Hassanpour, "Examining feasibility of developing a rock mass classification for hard rock TBM application using non-linear regression, regression tree and generic programming," *Geotech. Geol. Eng.*, vol. 36, no. 2, pp. 1145–1159, 2018.

[36] S. S. Anand and J. G. Hughes, "Hybrid data mining systems: the next generation," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1998, pp. 13–24.

[37] M. Bilal *et al.*, "Big Data in the construction industry: A review of present status, opportunities, and future trends," *Adv. Eng. informatics*, vol. 30, no. 3, pp. 500–521, 2016.

[38] V. Ahmed, Z. Aziz, A. Tezel, and Z. Riaz, "Challenges and drivers for data mining in the AEC sector," *Eng. Constr. Archit. Manag.*, vol. 25, no. 11, pp. 1436–1453, 2018.

[39] H. Kim, A. Stumpf, and W. Kim, "Analysis of an energy efficient building design through data mining approach," *Autom. Constr.*, vol. 20, no. 1, pp. 37–43, 2011.

[40] E. Petrova, P. Pauwels, K. Svidt, and R. L. Jensen, "In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data," in *Advances in Informatics and Computing in Civil and Construction Engineering*, Springer, 2019, pp. 19–26.

[41] C.-W. Cheng, S.-S. Leu, Y.-M. Cheng, T.-C. Wu, and C.-C. Lin, "Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry," *Accid. Anal. Prev.*, vol. 48, pp. 214–222, 2012.

[42] B. U. Ayhan and O. B. Tokdemir, "Safety assessment in megaprojects using artificial intelligence," *Saf. Sci.*, vol. 118, pp. 273–287, 2019.

[43] C.-L. Lin and C.-L. Fan, "Examining association between construction inspection grades and critical defects using data mining and fuzzy logic," *J. Civ. Eng. Manag.*, vol. 24, no. 4, pp. 301–317, 2018.

[44] A. Geronazzo, G. Brager, and S. Manu, "Making sense of building data: New analysis methods for understanding indoor climate," *Build. Environ.*, vol. 128, pp. 260–271, 2018.

[45] M. Ashouri, F. Haghighat, B. C. M. Fung, A. Lazrak, and H. Yoshino, "Development of building energy saving advisory: A data mining approach," *Energy Build.*, vol. 172, pp. 139–151, 2018.

[46] Y. Peng, J.-R. Lin, J.-P. Zhang, and Z.-Z. Hu, "A hybrid data mining approach on BIM-based building operation and maintenance," *Build. Environ.*, vol. 126, pp. 483–495, 2017.

[47] Q. Wen, J.-P. Zhang, Z.-Z. Hu, X.-S. Xiang, and T. Shi, "A Data-Driven Approach to Improve the Operation and Maintenance Management of Large Public Buildings," *IEEE Access*, vol. 7, pp. 176127–176140, 2019.

[48] S. S. Anand and A. G. Büchner, *Decision support using data mining*. Financial Times Management, 1998.

[49] O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook," 2005.

[50] X. Song, M. Shi, J. Wu, and W. Sun, "A new fuzzy c-means clustering-based time series segmentation approach and its application on tunnel boring machine analysis," *Mech. Syst. Signal Process.*, vol. 133, p. 106279, 2019.

[51] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework.," in *KDD*, 1996, vol. 96, pp. 82–88.

[52] L. Zhang, X. Wu, and M. J. Skibniewski, "Simulation-based analysis of tunnel boring machine performance in tunneling excavation," *J. Comput. Civ. Eng.*, vol. 30, no. 4, p. 4015073, 2015.

[53] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, 1994, vol. 1215, pp. 487–499.

[54] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York: Routledge, 1984.

[55] M. Lu, S. M. AbouRizk, and U. H. Hermann, "Sensitivity analysis of neural networks in spool fabrication productivity studies," *J. Comput. Civ. Eng.*, vol. 15, no. 4, pp. 299–308, 2001.

[56] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[58] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[59] X. Gao, M. Shi, X. Song, C. Zhang, and H. Zhang, "Recurrent neural networks for real-time prediction of TBM operating parameters," *Autom. Constr.*, vol. 98, pp. 225–235, 2019.

[60] B. Gao, R. Wang, C. Lin, X. Guo, B. Liu, and W. Zhang, "TBM penetration rate prediction based on the long short-term memory neural network," *Undergr. Sp.*, 2020.

[61] G. H. Erharter, T. Marcher, and C. Reinhold, "Comparison of artificial neural networks for TBM data classification," in *Rock*

Mechanics for Natural Resources and Infrastructure Development: Full Papers: Proceedings of the 14th International Congress on Rock Mechanics and Rock Engineering (ISRM 2019), 2019.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[63] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[64] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[65] M. Kassem, G. Kelly, N. Dawood, M. Serginson, and S. Lockley, "BIM in facilities management applications: a case study of a large university complex," *Built Environ. Proj. Asset Manag.*, vol. 5, no. 3, pp. 261–277, 2015.

SHUO LENG was born in Henan, China in 1996. He received the B.S. degree in civil engineering from the Department of Civil Engineering, Tsinghua University, China in 2018.

He is currently a Ph.D. Student at the Department of Civil Engineering, Tsinghua University. His research interests are building information model (BIM) and information technologies in civil engineering.

JIA-RUI LIN received the B.S. and Ph.D. degrees from the Department of Civil Engineering, Tsinghua University, China, in 2011 and 2016.

He is currently a Research Assistant Professor with the Department of Civil Engineering, Tsinghua University. His research interests are information technology for building and civil engineering, including BIM, augmented reality (AR), cloud computing and internet of things (IoT).

ZHEN-ZHONG HU was born in Guangdong, China. He received the B.S. and Ph.D. degrees from the Department of Civil Engineering, Tsinghua University, China, in 2005 and 2009, respectively.

He is currently an Associate Professor with the Tsinghua Shenzhen International Graduate School and the Department of Civil Engineering, Tsinghua University. His research interests focus on information technology in civil engineering, BIM, and digital disaster prevention and mitigation.

XUESONG SHEN received the B.S. degree in automation engineering and the M.S. degree in precision instruments and mechanology from Nanjing University of Aeronautics and Astronautics, China, in 2002 and 2005, respectively, and the Ph.D. degree in Construction Engineering and Management from The Hong Kong Polytechnic University, Hong Kong, China in 2010.

He is currently a Senior Lecturer in the School of Civil and Environmental Engineering, the University of New South Wales. His research interests focus on construction automation and robotics, artificial intelligence in rngineering applications, and digital twins.