Yang *et al.*
Chinese Journal of Mechanical Engineering (2023) 36:128
https://doi.org/10.1186/s10033-023-00932-3

Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE

Open Access

Vote-Based Feature Selection Method for Stratigraphic Recognition in Tunnelling Process of Shield Machine

Liman Yang[1], Xuze Guo[1], Jianfu Chen[1,2*], Yixuan Wang[1], Huaixiang Ma[3], Yunhua Li[1], Zhiguo Yang[1] and Yan Shi[1*]

Abstract

Shield machines are currently the main tool for underground tunnel construction. Due to the complexity and variability of the underground construction environment, it is necessary to accurately identify the ground in real-time during the tunnel construction process to match and adjust the tunnel parameters according to the geological conditions to ensure construction safety. Compared with the traditional method of stratum identification based on staged drilling sampling, the real-time stratum identification method based on construction data has the advantages of low cost and high precision. Due to the huge amount of sensor data of the ultra-large diameter mud-water balance shield machine, in order to balance the identification time and recognition accuracy of the formation, it is necessary to screen the multivariate data features collected by hundreds of sensors. In response to this problem, this paper proposes a voting-based feature extraction method (VFS), which integrates multiple feature extraction algorithms FSM, and the frequency of each feature in all feature extraction algorithms is the basis for voting. At the same time, in order to verify the wide applicability of the method, several commonly used classification models are used to train and test the obtained effective feature data, and the model accuracy and recognition time are used as evaluation indicators, and the classification with the best combination with VFS is obtained. The experimental results of shield machine data of 6 different geological structures show that the average accuracy of 13 features obtained by VFS combined with different classification algorithms is 91%; among them, the random forest model takes less time and has the highest recognition accuracy, reaching 93%, showing best compatibility with VFS. Therefore, the VFS algorithm proposed in this paper has high reliability and wide applicability for stratum identification in the process of tunnel construction, and can be matched with a variety of classifier algorithms. By combining 13 features selected from shield machine data features with random forest, the identification of the construction stratum environment of shield tunnels can be well realized, and further theoretical guidance for underground engineering construction can be provided.

Keywords Shield machine, Tunneling parameters, Feature selection, Stratigraphic recognition

*Correspondence:
Jianfu Chen
49608764@qq.com
Yan Shi
shiyan@buaa.edu.cn
Full list of author information is available at the end of the article

1 Introduction

With the development of large-scale machinery and construction technology in our country, shield machine plays an increasingly important role in underground tunnel engineering, and the complex and changeable geological environment puts forward higher requirements for the safety of shield machine [1–3]. Since the Shield Machine can only advance and can not retreat during the tunnel excavation process, if it encounters geological disasters during the construction process, it will be accompanied by accidents such as landslides and falling blocks, which will delay the construction period and increase the construction cost. Therefore, it is very important to monitor and identify the geological environment of the tunnel excavation face, and adjust the shield machine parameters in real-time to match the geological environment. The shield machine should maintain good working conditions and ensure the safety and efficiency of underground construction [4–6]. In the traditional excavation method, the geological conditions of the construction route must first be investigated and studied, usually by drilling holes at certain intervals, and inferring the overall geological conditions according to the drilling data. With the development of artificial intelligence and data mining technology, data-driven formation identification methods have been widely used in underground construction due to their superior performance in real-time, accuracy and other aspects. The stratum can be identified through the mapping rule between the geological environment and the data parameters of the equipment's tunneling operation [6, 7], and different strata conditions can be reflected in real-time, so as to make real-time adjustments to the tunneling parameters and ensure the safety of the shield tunneling machine under different geological conditions. Since there are hundreds of shield machine data features, to complete the accurate mapping of data features and formation features, it is necessary to select the most representative data features from many data features as the input of the recognition model, so as to improve the calculation efficiency and shorten the recognition time. The data characteristic parameters commonly used in formation identification are mainly divided into two categories, the working parameters of shield machine equipment, such as excavation speed, pressure and temperature of each working chamber, and the parameters measured by sensors when interacting with the geological environment, such as cutter head Torque, propulsion, penetration, etc. The manual selection of features is a commonly used method at present. Engineers select effective data features through geological theory and engineering experience [8, 9], such as digging speed, total thrust, cutterhead speed, cutterhead torque, penetration force, etc, which are often used as input features for geological identification [10–12]. Although the manual selection of features has certain pertinence and reliability, it is not conducive to further improving the recognition accuracy, and it also needs to rely on rich engineering experience and professional knowledge, which does not have wide applicability. In contrast, the feature extraction method in machine learning (ML) requires less professional experience and knowledge, and it is more objective and efficient to use data analysis techniques to select data features with wider coverage [13–15].

In the actual operation of the shield tunneling machine, we usually maintain the tunneling direction and attitude of the shield tunneling machine by constantly adjusting tunneling parameters to ensure its good rock-breaking efficiency. At present, a more practical method in engineering is to extract effective features from a large number of data sets through machine learning, and establish a mapping relationship between relevant tunneling parameters of the shield machine and the actual working state, so as to realize real-time monitoring and adjustment of the working state of the shield machine. Feature extraction is a widely used research direction in machine learning [16–18]. Current feature extraction methods (FSM) are usually divided into three categories: Filter FSM, Wrapper FSM and Embedded FSM. Filter FSM first performs feature extraction on the data set, and then trains the classification model. The feature extraction process and the subsequent classification model are relatively independent two processes. Compared with Filter FSM, Wrapper FSM directly uses the performance indicators of the final classifier as the evaluation criteria for feature subsets, and optimizes for a given classification model, so the performance of Wrapper's classification model is better than Filter, but Wrapper is in the feature extraction process. The classifier needs to be trained multiple times, and its time complexity is much more complicated than that of the Filter. The embedded FSM integrates the feature extraction process and the classification model training process, both of which are the same optimization process. In recent years, in order to solve the problem of local optima of a single FSM, a new feature extraction method, ensemble learning-based feature extraction (ELFS) [19–22], has been studied. The main principle of ELFS is to evaluate and select a better subset of features in an attribute-weighted or prioritized manner from the results of multiple feature extractors. From the analysis of the difficulty of integration, Wrapper and Embedded are more difficult to integrate into ELFS than Filter, so we usually consider integrating Filter into ELFS. The data features selected by the filter FSM mainly include

feature variance, the correlation between data features and target values, such as Pearson correlation coefficient, Spearman correlation coefficient, Kendall correlation coefficient, information gain, gain ratio, symmetric uncertainty, RF, One- R and Chi-square. Olsson et al. [23–25] used multiple Filter FSMs to process software failure datasets PROMISE database and NASA metrics database. The weights of all features were calculated through each FSM, and finally, the weights of each feature were accumulated. After averaging and sorting, it was found that the subset of features selected by the combination of information gain and PCA has the best recognition results. However, in the field of underground construction, due to the high hybridity and redundancy of shield machine massive data, there is currently no data feature machine for processing shield machine Learning feature extraction method. Therefore, in order to reduce the time complexity of the ELFS algorithm, improve the applicability of the algorithm, and at the same time give full play to the advantages of ELFS to easily extract data features with larger weights and ensure higher recognition accuracy, this paper studies and proposes A VFS (Voting based Feature Selection) algorithm based on feature frequencies in all FSMs is proposed to filter shield machine raw data and reduce its hybrid redundancy.

The accuracy of stratigraphic identification not only requires characteristic data with high correlation with geological conditions, but also depends on the correct establishment of identification models. Classification algorithms are often closely related to the accuracy of the recognition model. Currently, the commonly used classification algorithms mainly include linear discriminant analysis (LDA) [26], quadratic discriminant analysis (QDA), support vector machine (SVM) [27], K-nearest neighbor (KNN) [28], neural network (NN) [29], naive Bayes (NB) [30], AdaBoost (AdaB) [31], gradient boosted decision tree (GBDT) [32] and random forest (RF) [33], etc. In a single geological environment, the formation recognition achieved good recognition results. However, since there are only 60 sets of data available for training and testing, the generalization ability of the model in complex geological environments needs to be further tested. In order to test the effectiveness of the VFS algorithm proposed in this paper, this paper selects 10 commonly used classification algorithms to match with VFS, performs stratigraphic identification in 6 different geological structures, and compares them with the basic feature set. Through the matching and comparison of multi-class classification algorithms, the best recognition model combined with VSF is obtained, which also verifies the wide applicability of the VFS algorithm.

Feature screening of shield machine data is a very important research direction. Due to a large number of data features of shield machine itself, it is of great significance to reduce the dimensionality of the data and select the optimal feature subset for formation identification to reduce the computational cost and improve the identification accuracy.

Therefore, in the actual construction process, it is in urgent need of a feature subset screening method that can be applied together with a variety of classification algorithms at a high speed, so as to quickly extract the parameter feature set that is most relevant to the working state of the shield tunneling machine, establish an accurate mapping relationship, and improve the accuracy of stratum identification. In response to this problem, this paper proposes a voting-based feature extraction (VFS) method, which integrates 4 filter FSMs to obtain the optimal feature subset according to the frequency of features in all FSMs. After matching with 10 commonly used classification algorithms, it is proved that the proposed VFS can effectively shield the highly redundant and highly hybrid data, and has the advantages of short time-consuming, strong adaptability and wide applicability. The organization of this paper is as follows: Section 2 introduces the research object and shield machine construction data characteristics and data sources; Section 3 introduces the principle and design process of the VSF algorithm; Section 4 designs several experiments to verify the superiority of VSF, Data preprocessing and results are discussed; Section 5 presents the conclusions of this paper and future work.

2 Related Work

When the shield machine is excavated in the horizontal direction underground, it comes into contact with different geological environments in turn. Due to the relatively large diameter of the mud-water balance shield machine, it often contacts one or more sedimentary geological layers in the longitudinal direction. In the current construction process, the difference between different geological environments is mainly reflected in the combination of different sedimentary geological layers. Common geological structures include a silty sand structure layer, silty clay sandwiched silty soil layer, silty fine sand layer, and pebbleslayers, gravel-sand layers, and combinations of layers, etc. When the super-diameter shield machine comes into contact with strata with different geological structures, its own excavation parameters will change to a certain extent. Therefore, by studying the correlation between the changes of the excavation parameters and the changes of the geological environment, we can establish the mapping law between the geological environment and the parameters of the equipment excavation operation, so as to construct a data-driven identification model, reasonably match the excavation parameters of the shield, and ensure the construction quality and safe operation.

Super-large mud-water shield machine pumps a certain concentration of mud into the mud-water shield chamber. As the soil residue and groundwater cut by the cutter head flow into the excavation chamber along the cutter groove, the mud concentration and pressure in the mud-water chamber gradually increase, and balance with the earth pressure and water pressure on the excavation face, forming a mud film or a permeable wall formed by mud-water pressure on the excavation face, and forming a stable excavation on the excavation face. Super-large mud-water balance shield machine is generally composed of an air cushion, mud-water tank, cutter head system, main drive, scouring device, crusher, slurry pipe suction port, pipe fitting assembly machine, main engine row, slurry pump, etc. Its three-dimensional model is shown in Figure 1 below. The data used in this paper is the sensor data of a super-large diameter mud shield machine in a tunnel construction in my country. There are a total of 197 data features, including 5 data features commonly used in existing studies, namely, travel speed, total propulsion force, cutter head speed, cutter head torque, and penetration. In addition, the 201 features also include the volume of the mud conveying mud circulation system and the mud discharge volume, the grouting pump capacity of the grouting system, the rated torque and output torque of the milling head of the drive system, as well as the rotational speed, working chamber pressure, and rotational speed range, cylinder working pressure, etc.

This paper obtained 800 data sample points from 6 geological layers during the construction of a certain tunnel, as shown in Table 1. Figure 2 depicts the correlation between each feature and a graph of five commonly used features.

Figure 2(a) shows the correlation thermal map of the commonly used five features, and the depth of color represents the correlation difference among the five features. Figure 2(b) shows the correlation thermal diagram of 197 features; The following Figure is the graph of these five features and the classification limits of 800 sample data. Label 1, label 2, label 3, label 4, label 5 and label 6 represents respectively the six kinds of geological conditions.

After data preprocessing, 197 original shield data features were initially obtained. If all the above features are applied to formation identification, it will increase the complexity and calculation time of the identification model, which is not conducive to the real-time and efficient completion of subsequent data-driven geological identification. Therefore, it is necessary to complete feature extraction through appropriate feature selection methods, realize data dimensionality reduction, and select core data features for subsequent stratigraphic identification.

3 Proposed Method

3.1 Design Principle of VFS Method

The minimum number of features selected by different Filter FSMs is different, so it is necessary to integrate the advantages of each feature selection method. In this way, the selected feature subset is not affected by the characteristics of the feature method, and a more reasonable feature subset is selected. Similar to the principle of ensemble learning, this paper proposes a feature selection (VFS) method through voting, as shown in Figure 3.

Figure 3 illustrates the whole process of the calculation using the VFS method in detail. First, after the original feature set is cleaned, m features can be obtained, and the m features are quantified and numbered from 1 to m. For example, 135 represents the 135th feature. Assuming that there are n types of feature selection methods, each of the first p features with better priority levels can be used to obtain n feature subsets. These subsets form a feature pool (FP). In FP, each feature is selected at most n times, and the features are prioritized according to the number of times the features appear in the FP, and finally n-Rank feature subsets can be obtained. Assuming that a certain

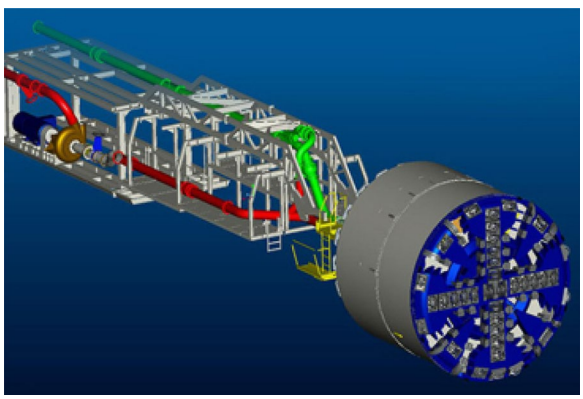

Figure 1 Slurry balance shield machine

Table 1 Classify the tunnelling data into classification labels

Label name	Target value	Sample numbers	Geologic layer type
label 1	1	65	Muddy silty clay
label 2	2	146	Mixed layer of clay and silt
label 3	3	220	Fine sand layer
label 4	4	235	Mixed layer of fine sand and gravel
label 5	5	67	Fine sand, gravel sand and gravel
label 6	6	67	Mixture of fine sand and gravel

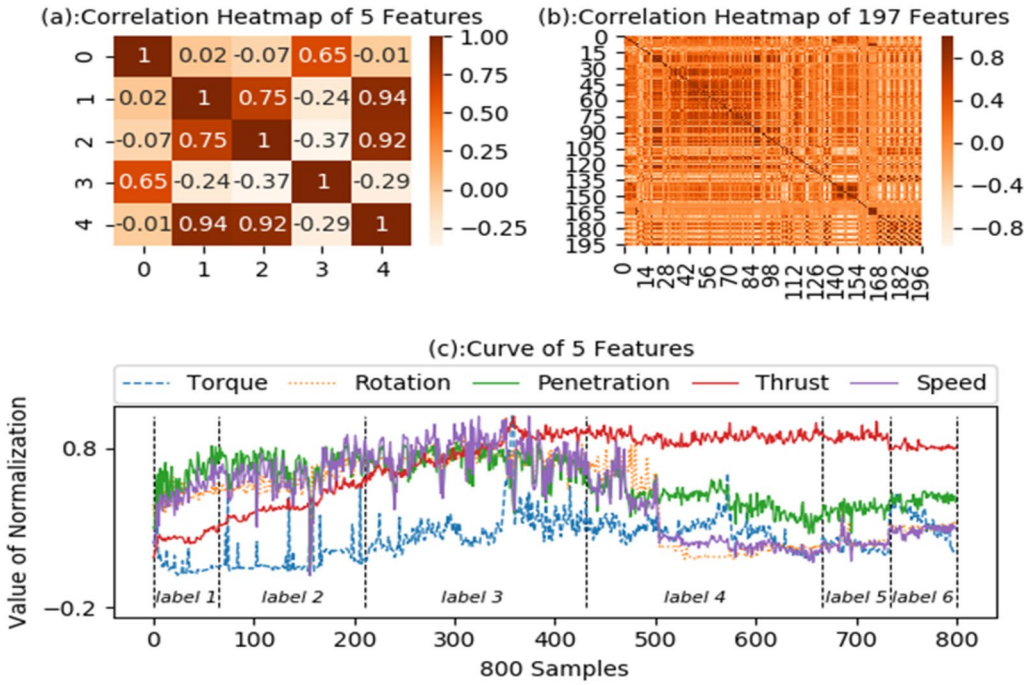

Figure 2 Correlation heatmap of features and graphs of five features

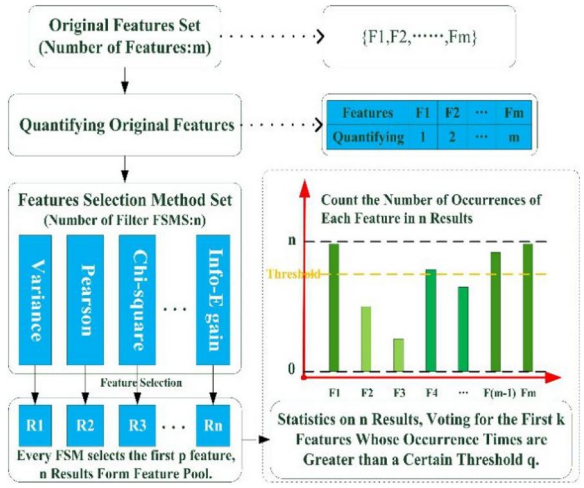

Figure 3 Process of VFS method

feature in FP appears more than q times, it is considered to be an effective feature that can be selected. The number k of the effective feature subsets finally selected by VFS depends on whether the selected feature subsets can reach the accuracy of the classification model. Therefore, the parameters p, q, and k in VFS are adjusted according to the prediction accuracy of the classifier.

The advantages of VFS are as follows: Other ELFS based on feature weights need to iteratively calculate the weights of all features, and filter the effective feature subsets based on the weights, but it often requires hundreds or even thousands of iterations to achieve the accuracy of the classifier, and the time is complicated. The VFS greatly simplifies the weight calculation process, and can directly obtain the level of all features in the feature pool in one step, reducing the computational complexity a lot, at the same time, compared to a single Filter FSM, the evaluation indicators of each Filter are comprehensively considered in the VSF, and the characteristics obtained by the screening are excellent in each FSM, and its practicability and popularization are high, which meets the requirements of shield construction. The goal of stratum recognition can be accomplished well, and the time spent on feature selection is reduced.

The pseudo code of the VFS algorithm is shown below.

Pseudo code of VFS algorithm:
Assume:
- *there are m Feature Selection Methods(FSMs) Set;*
- *initialize set G={},used to store selected feature subsets;*

For $fsm_{(i)}$ *in FSMs(i=1,2,...,m):*
- *select top* $k_{(i)}$ *features subset in* $fsm_{(i)}$*;*
- *add top* $k_{(i)}$ *features subset to G;*

N *=kinds of features in G*
N'*= number of features in G=k*m*

$$k \le N \le N'$$

Count the occurrences of every feature that appears in G:
- *for* j *in range(N):*
  - $f(j)$*= occurrence that feature(j) appears in G*
  - *if* $f(j) = m$*:*

Feature$(j) \in$ *VFS feature subset =Rank1 feature subset*
- *else* $f(j) = 1,2,3,\ldots,m-1$*:*

feature$(j) \in$ *Rank(m-* $f(j)+1$*) feature subset*
End.

In this paper, $m = 197$, $n = 4$ Filter FSMs with smaller computational complexity, and $p = 30$, the four feature subsets form the feature pool of VFS. Each feature in the feature pool appears at most $a = 4$ times, so all the features in the feature pool are divided into four levels of feature subsets. Here, the feature subsets of Rank 1 and Rank 2 are selected to be the input of the classification model algorithm 3.2 Four FSMs of VFS introduces the process of feature selection in detail.

3.2 Four FSMs of VFS

In this paper, the filtering feature selection method is as follows: variance, Pearson correlation coefficient, Chi-square and information entropy gain.

3.2.1 Variance

The variance of features measures the dispersion degree of each feature, and the larger variance of each feature used in the classification algorithm is, the greater fluctuation of the feature is. So the larger variance of feature is, the larger the information contained in feature is. The samples of each class may have great differences in the six kinds of geological layers, which can make the classifier easier to distinguish between various geological layers and to separate the strata of different categories. In Eq. (1), x_i represents one of the 800 features, s^2 is variance of the feature x_i, N shows the number of total samples.

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2, \tag{1}$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, N = 800$.

3.2.2 Pearson Correlation Coefficient

The classifier uses the feature data with higher correlation with the formation to get better accuracy. Here refers to the Pearson Correlation Coefficient between each feature and target value. In Eq. (2), the $r_{x,y}$ represents the correlation coefficient between feature x and target value y (values of y are 1, 2, 3, 4, 5, 6).

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}. \tag{2}$$

Figure 4(a) shows the variance distribution diagram calculated by 197 features according to Eq. (2), and Figure 4(b) shows the correlation coefficient distribution diagram calculated by 197 features according to Eq. (2). Among them, the distribution diagram form of specific parameters can better capture more representative features.

3.2.3 Information Entropy Gain

In this paper, Information Entropy Gain is one of the filtering techniques to select the features. The concept of this technique is used to choose the best feature to construct the tree in Decision Tree algorithm. Each feature gains a specific score and features (maybe more than one) with the highest scores will be selected to be the optimal subset of features that are considered input variables of the classifier. We can find the explanation in more detail in Ref. [34].

3.2.4 Chi-Square

Chi-square (χ^2) test is the deviation degree between the actual observed value and the theoretical inferred value of statistical samples. The deviation degree between the actual observed value and the theoretical inferred value determines the size of chi-square value. On the contrary, the smaller the deviation between the two; if the two values are completely equal, the chi-square value is 0, indicating that the theoretical value is completely consistent. In Eq (3), f_i represents the number of samples on interval, the number of intervals are l. p_i represents the probability of a sample appearing on the interval i [35].

$$\boldsymbol{\chi}^2 = \sum_{i=1}^{l}\frac{(\boldsymbol{f_i} - \boldsymbol{Np_i})^2}{\boldsymbol{Np_i}}. \tag{3}$$

Figure 5(a) shows the Chi-square distribution diagram calculated by 197 features according to Eq. (3), and Figure 5(b) shows the Information entropy gain distribution diagram calculated by 197 features according to Eq. (3). Among them, the distribution diagram

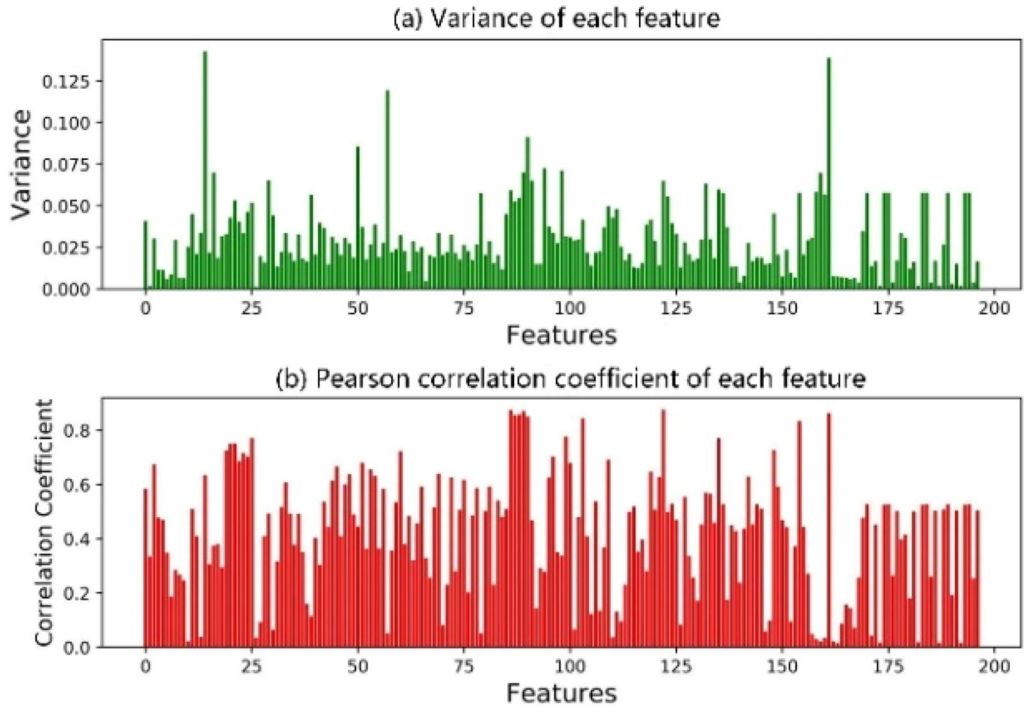

Figure 4 Variance and correlation coefficients of 197 features

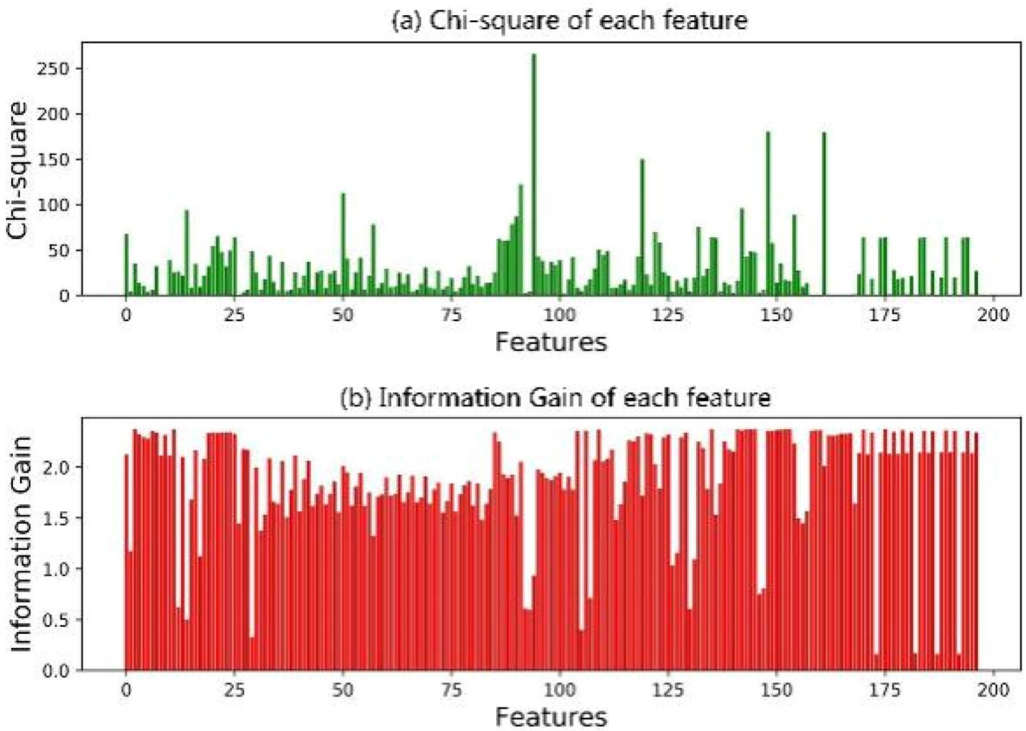

Figure 5 Chi-square and information gain of 197 features

form of specific parameters can better capture more representative features.

This paper calculates the four attributes of the feature, including variance, Pearson correlation coefficient, chi-square and information entropy gain. The first 30 features of the four attributes are obtained respectively, and thus 4 feature subsets are obtained. The feature subset is shown in Table 2. In this table, the number represents the serial number of each feature.

Only one feature (Feature numbered 135, which is bold in Table 3) appears in the four feature subsets, ranking this feature as Rank 1. The features that appear three times, two times and one time in the four feature subsets are listed as Ranks 2-4 successively. After statistical calculation, the feature subset of the four ranks is shown in Table 3.

4 Experiments and Discussion

4.1 Experimental Design and Baseline

In order to verify the superiority of the VFS algorithm, this paper chooses 10 common classification algorithms, random forest (RF), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machine (SVM), K nearest neighbor (KNN), Neural Network (NN), Naive Bayes (NB), Bagging (Bag), Ada-Boost (Ada) and Gradient Boosting Tree (GBDT). Random forest is an algorithm that integrates multiple trees through the idea of ensemble learning. Its basic unit is a decision tree, and its essence belongs to ensemble learning, a big branch of machine learning. Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant approach, which uses statistics, pattern recognition and machine learning methods to try to find a linear combination of the characteristics of two classes of objects or events in order to be able to characterize or distinguish them. The resulting combination can be used as a linear classifier or, more commonly, to reduce dimensions for subsequent classifications. Similar to the linear discriminant analysis, the quadratic discriminant analysis is obtained from the perspective of the probability distribution. The difference is that the linear discriminant analysis assumes the same covariance matrix for each category, while the quadratic discriminant analysis has a different covariance matrix for each category. Support Vector Machine (SVM) is a generalized linear classifier that classifies data according to supervised learning; its decision boundary is the maximum-margin hyperplane for solving the learning sample. SVM uses hinge loss function to calculate empirical risk and adds regularization terms into the solving system to optimize structural risk, which is a classifier with sparse and robust. SVM is one of the common kernel learning methods, which can be used for nonlinear classification by kernel method. *k*-Nearest Neighbor (KNN) classification algorithm is a relatively mature method in theory and one of the simplest machine learning algorithms. The idea of this method is as follows: in the feature space, if most of the k nearest samples near a sample belong to a certain category, the sample also belongs to this category. Neural network (NN) is a mathematical model or computational model that imitates the structure and function of biological neural network. Neural networks are computed by a large number of artificial neuron connections. In most cases, artificial neural network can change the internal

Table 2 Four feature subsets of Top 30 in four FSMs

Attributes of features	Top 30 of the 197 features
Variance	14, 161, 57, 90, 50, 94, 98, 89, 16, 159, 29, 91, 122, 132, **135**, 86, 158, 189, 193, 170, 183, 174, 184, 175, 194, 154, 136, 79, 160, 39
correlation coefficient	122, 86, 89, 161, 88, 87, 90, 103, 154, 99, **135**, 25, 21, 20, 148, 19, 60, 23, 96, 24, 109, 22, 51, 100, 2, 45, 53, 119, 69, 48
Chi-square	94, 148, 161, 119, 91, 50, 142, 14, 154, 90, 89, 57, 132, 122, 0, 21, **135**, 189, 184, 194, 175, 170, 25, 193, 183, 174, 136, 86, 88, 87
information gain	153, 152, 145, 143, 141, **135**, 2, 175, 144, 11, 170, 151, 109, 160, 159, 150, 179, 189, 142, 106, 194, 184, 148, 158, 149, 104, 6, 191, 186, 177

Table 3 Feature subset of Rank1, Rank 2, Rank 3 and Rank 4

Feature rank	Feature subset	Features number
Rank 1	Only 135	1
Rank 2	194,189,184,175,170,161,154,148,122,90,89, 86	12
Rank 3	193, 183, 174, 160, 159, 158, 142, 136, 132, 119, 109, 94, 91, 88, 87, 57, 50, 25, 21, 14, 2	21
Rank 4	191, 186, 179, 153, 152, 151, 150, 149, 145, 144, 143, 141, 106, 104, 103, 100, 99, 98, 96, 79, 69, 60, 53, 51, 48, 45, 39, 29, 24, 23, 22, 20, 19, 16, 11, 6, 0	37

structure on the basis of external information, which is an adaptive system. Naive Bayes (NB) is one of the most widely used classification algorithms, which is a classifier method based on Bayesian definition and feature condition independent assumption. Bagging(Bag) is short for Bootstrap Aggregating. In short, the method of bootstrap sampling was used to construct several different training sets. Then, the corresponding base learners are trained on each training set. Finally, these base learners are aggregated to get the final model. Adaboost is an iterative algorithm whose core idea is to train different classifiers for the same training set, and then assemble these weak classifiers to form a stronger final classifier. Gradient Boosting Decision Tree(GBDT), also known as MART (Multiple Additive Regression Tree), is an iterative decision tree algorithm consisting of multiple decision trees. The conclusions of all the trees are added together to make the final answer. The parameters of the 10 classifiers are adjusted as much as possible to the optimal value. In addition, we select the five basic features of shield machine (mining speed, total thrust, tool speed, tool torque, penetration force) to be combined with the above ten classification algorithms, and the classification results are used as the Baseline of the VFS algorithm. In order to minimize the error in the training and testing process, we adopt the 10-fold cross-validation (Cross-Validation is usually abbreviated as CV) method to compare and analyze the test accuracy and test time of each classifier.

The test results of the five basic features under 10 classifiers are shown in Table 4.

4.2 Experimental Results

VFS has obtained four characteristic grades, as shown in Table 3. In order to compare the pros and cons of different grade features of VFS, each grade feature was used as the input of 10 classification algorithms, and the test accuracy was used as the evaluation index for comparative analysis. Considering that there is only one feature in the Rank1 feature set, the feature subsets of Rank1 and 2 are put together as the new feature subset Rank1+2, and the final four levels of feature subsets are Rank1+2, Rank2, Rank3 and Rank4; At the same time, a comparative analysis of VFS and four Filter FSMs was carried out to detect the superiority of VFS compared to a single FSM.

4.2.1 Comparative Analysis of Four Rank Feature Subsets of VFS

Tables 5 and 6 show the test results of different levels of feature subsets. For a clearer comparison and analysis, Tables 4, 5, 6 are visualized as Figure 6.

From Tables 4, 5, 6, it is found that the accuracy of the random forest classifier is mostly higher than the other 9 classifiers, and the accuracy of the random forest is relatively stable. The accuracy of QDA and NN is very low; Bag, AdaB, and GBDT have good classification results,

Table 4 Accuracy baseline of ten classifiers (10 and 5 Fold CV)

Feature Rank	RF	LDA	QDA	SVM	KNN
5 CV	0.738	0.623	0.724	0.627	0.710
10 CV	0.845	0.715	0.789	0.674	0.830
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
5 CV	0.544	0.742	0.747	0.628	0.734
10 CV	0.637	0.798	0.838	0.733	0.825

Table 5 Testing results of different features rank (5 Fold CV)

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.867	0.878	0.682	0.848	0.844
Rank2	0.864	0.875	0.704	0.846	0.838
Rank3	0.8017	0.767	0.081	0.676	0.705
Rank4	0.7993	0.833	0.746	0.641	0.798
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.826	0.786	0.841	0.702	0.874
Rank2	0.294	0.783	0.849	0.710	0.859
Rank3	0.709	0.769	0.734	0.580	0.835
Rank4	0.625	0.751	0.771	0.523	0.804

Table 6 Testing results of different features rank (10 Fold CV)

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.926	0.894	0.784	0.871	0.902
Rank2	0.918	0.889	0.778	0.871	0.896
Rank3	0.921	0.808	0.120	0.748	0.774
Rank4	0.902	0.893	0.883	0.752	0.909
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.797	0.860	0.911	0.699	0.933
Rank2	0.293	0.854	0.921	0.705	0.914
Rank3	0.669	0.801	0.812	0.601	0.902
Rank4	0.615	0.856	0.894	0.550	0.875

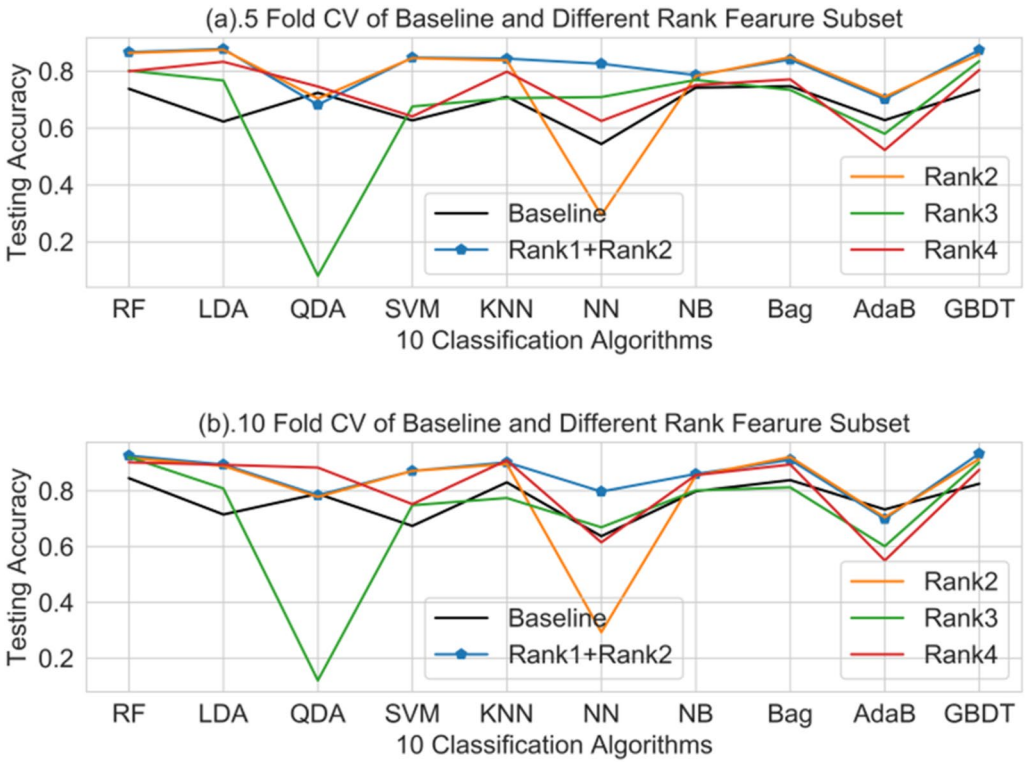

Figure 6 Visualization of Tables 4, 5 and 6

but their accuracy is not as stable as random forest. Furthermore, the test accuracy of most classifiers decreases as the feature subset level degrades.

4.2.2 Comparative Analysis of VFS and Four Filter FSMs

The four feature subsets in Table 3 have 71 features in total. The first 71 features obtained by the four Filter methods are sliced, and Top 13, Top 14–33, and Top 34–71 in a single FSM feature are obtained, which are compared with the Rank1 obtained by VFS. +Rank2, Rank3 and Rank4 for comparison. Tables 7, 8, 9, 10, 11, 12 show the test results of four single FSM and VFS: Tables 7 and 8 show the test results of the top 13 features of each FSM. Tables 9 and 10 show the test results of each FSM, and the results of the 14th to 33rd features are obtained by screening. Tables 11 and 12 show the results of the 34th to 71st features obtained by the screening of each FSM.

It is found in Tables 7, 8, 9, 10, 11, 12 that in most cases, the random forest algorithm has higher accuracy and the best stability than other classifiers. In the case of the same number of features, the test results of the features

Table 7 5 fold CV of feature set Top 13

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.867	0.878	0.682	0.848	0.844
Rank2	0.821	0.631	0.081	0.622	0.718
Rank3	0.865	0.879	0.723	0.739	0.818
Rank4	0.921	0.808	0.278	0.748	0.770
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.826	0.786	0.830	0.702	0.873
Rank2	0.639	0.689	0.764	0.802	0.766
Rank3	0.664	0.761	0.795	0.713	0.852
Rank4	0.767	0.755	0.816	0.678	0.855

Table 8 10 fold CV of feature set Top 13

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.926	0.894	0.784	0.871	0.902
Rank2	0.938	0.733	0.081	0.672	0.806
Rank3	0.913	0.901	0.750	0.799	0.883
Rank4	0.980	0.873	0.302	0.821	0.839
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.797	0.860	0.921	0.705	0.932
Rank2	0.784	0.754	0.882	0.798	0.866
Rank3	0.673	0.841	0.911	0.707	0.916
Rank4	0.776	0.777	0.881	0.869	0.927

Table 9 5 fold CV of feature set Top 14–33

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.801	0.767	0.081	0.676	0.705
Rank2	0.891	0.810	0.470	0.694	0.764
Rank3	0.803	0.778	0.536	0.719	0.812
Rank4	0.870	0.868	0.512	0.674	0.784
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.709	0.769	0.767	0.580	0.835
Rank2	0.715	0.781	0.808	0.705	0.830
Rank3	0.759	0.794	0.819	0.412	0.807
Rank4	0.651	0.783	0.784	0.744	0.796

selected by the VFS are generally better than the results of the features selected by a single Filter FSM. For a clearer comparison and analysis, Tables 9, 10, 11, 12 are visualized in Figure 7.

After looking up Tables 7, 8, 9, 10, 11, 12, 13 features selected from the VFS can be determined, as shown in Table 13 below.

In addition to the accuracy of the algorithm, the consuming time is also one of the evaluation indicators to measure the quality of the algorithm. At the end of this section, we use the cross-validation method to compare the time taken by the 10 classification algorithms, as shown in Figure 8. In Figure 8, the abscissa is a different classification algorithm, and the ordinate is consuming time. The testing device is a ThinkPad notebook computer, the processor is AMD A8-7100 Radeon R5, 1.8 Hz. It can be found that the time taken by the random forest algorithm is slightly longer than that of LDA, QDA,

Table 10 10 fold CV of feature set Top 14–33

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.921	0.808	0.120	0.748	0.774
Rank2	0.931	0.888	0.556	0.800	0.850
Rank3	0.867	0.851	0.561	0.791	0.871
Rank4	0.922	0.890	0.577	0.715	0.866
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.669	0.801	0.815	0.601	0.902
Rank2	0.845	0.828	0.882	0.706	0.935
Rank3	0.794	0.863	0.865	0.437	0.877
Rank4	0.767	0.829	0.871	0.839	0.891

Table 11 5 fold CV of feature set Top 34–71

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.799	0.833	0.746	0.641	0.798
Rank2	0.878	0.890	0.862	0.846	0.892
Rank3	0.854	0.874	0.772	0.699	0.809
Rank4	0.804	0.804	0.759	0.727	0.810
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.625	0.751	0.810	0.529	0.804
Rank2	0.603	0.852	0.900	0.595	0.854
Rank3	0.575	0.681	0.774	0.576	0.841
Rank4	0.723	0.755	0.810	0.515	0.765

Table 12 10 fold CV of feature set Top 34–71

Feature Rank	RF	LDA	QDA	SVM	KNN
Rank1+2	0.902	0.893	0.883	0.752	0.909
Rank2	0.938	0.932	0.946	0.898	0.919
Rank3	0.895	0.900	0.771	0.719	0.852
Rank4	0.900	0.888	0.865	0.851	0.905
Feature Rank	**NN**	**NB**	**Bag**	**Ada**	**GBDT**
Rank1+2	0.615	0.856	0.911	0.550	0.875
Rank2	0.697	0.919	0.925	0.510	0.905
Rank3	0.641	0.731	0.867	0.619	0.832
Rank4	0.766	0.854	0.905	0.556	0.796

SVM, KNN and NB, but the accuracy of the random forest is higher. Although the accuracy of Bagging, Adab and GBDT algorithms is almost the same as that of random forest, these algorithms take much more time than random forest, and the time consumed by these four algorithms is not very stable.

Through the above experiments, we can see that the new method proposed in this paper can effectively shield highly redundant and highly mixed data, and has the advantages of short time consuming, strong adaptability and wide applicability. It can be well adapted to various classification algorithms, proving that it can be used to process complex data sets of large equipment such as shield tunneling machines. In this way, the real-time identification of the stratum and the real-time adjustment of the working state of the shield machine can be realized in practical application.

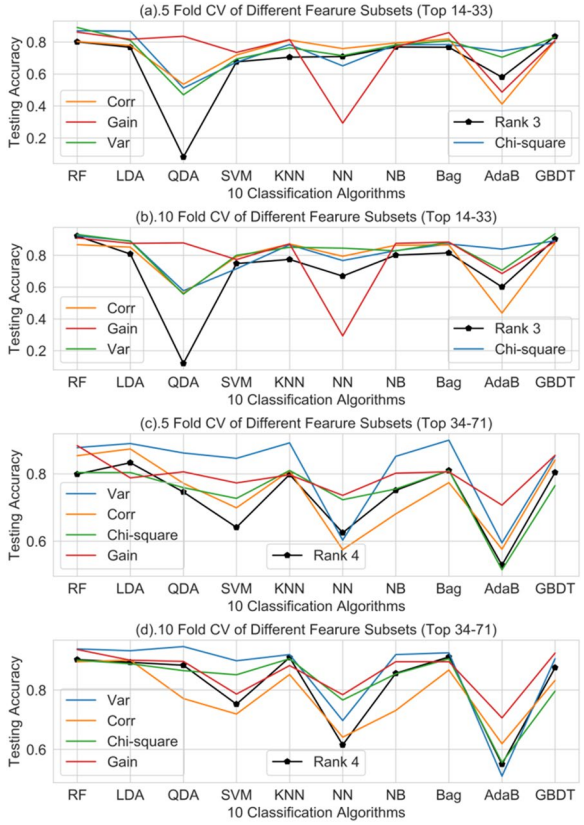

Figure 7 Visualization of Tables 9, 10, 11, 12

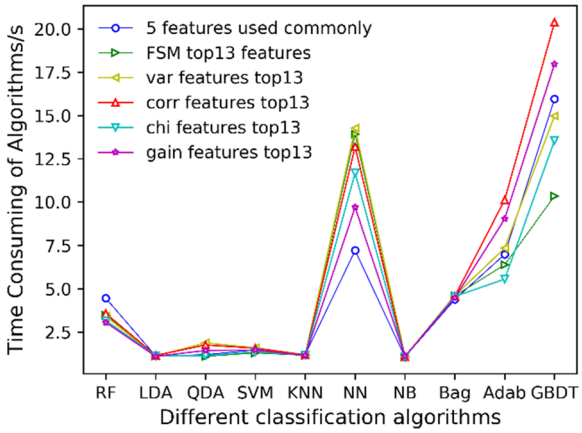

Figure 8 Time spent by different classification algorithms

Table 13 The actual feature name corresponding to the FSM Top 13 features

Feature	Feature name	Feature level
135	Total thrust of hydraulic cylinder	Rank1
194	Rotation Speed of No. 6 main drive motor	Rank2
189	Rotation Speed of No. 5 main drive motor	Rank2
184	Rotation Speed of No. 4 main drive motor	Rank2
175	Rotation Speed of No. 2 main drive motor	Rank2
170	Rotation Speed of No. 1 main drive motor	Rank2
161	Screw machine torque	Rank2
154	Frequency of No. 4 main drive motor	Rank2
148	Frequency of No. 6 main drive motor	Rank2
122	Suction pressure of cylinder	Rank2
90	Pressure of No. 5 sensor in excavation warehouse	Rank2
89	Pressure of No. 3 sensor in excavation warehouse	Rank2
86	Pressure of No. 1 sensor in excavation warehouse	Rank2

5 Conclusions

Aiming at the feature selecting problem of shield machine excavation data with high redundancy and confounding characteristics, this paper proposes a Voting-based Feature Selection method (VFS), which integrates multiple FSMs to screen and fuse the optimal shield data subsets according to the frequentness that the features occur in the feature pool. And it simplifies the process of feature sorting, thereby reducing the time complexity of feature selection. The five basic features and the features obtained by VFS are respectively combined with 10 common classification models to verify the superiority of VFS.

The test results indicate that the highest test accuracy of the feature combination of Rank1 and Rank2 in the classification is 92.6% (under the ten-fold cross-validation, the accuracy of the random forest algorithm), which successfully completes the task of stratum identification. The accuracy rate is higher than the highest accuracy rate of 84.5% based on the five basic features. Furthermore, the combination of VFS and random forest classifier can make the algorithm more effective. Compared to other classification models, random forests have shorter recognition time, stronger stability and easy promotion.

The three parameters in VFS will be further optimized in our future work to adapt to more complex geological environments. The VFS proposed in this article only uses Filter FSM. If combined with Wrapper FSM and Embedded FSM, the classification effect may be better while taking no account of consuming time. The 13 characteristics obtained by VFS have good applicability and can provide the theoretical guidance for future underground shield machine constructions. To sum up, the method proposed in this paper can effectively solve the problems of miscellaneous and high complexity of shield tunneling machine construction data, quickly screen out the most representative features in a large number of tunneling parameter data

sets, and greatly improve the accuracy of stratum identification related to tunneling parameters. Due to its universality with various classification algorithms, it can be used as a key auxiliary decision-making method to help shield tunneling machine maintain a better working state in future construction.

Author contributions

LY, YS, YW and YL were in charge of the whole trial; XG and JC wrote the manuscript; HM and ZY assisted with sampling and laboratory analyses. All authors read and approved the final manuscript.

Authors' information

Liman Yang, born in 1975, is an associate professor in *School of Automation Science and Electrical Engineering, Beihang University, China*.
Xuze Guo, born in 1995, receives his master degree from *School of Automation Science and Electrical Engineering, Beihang University, China*.
Jianfu Chen, born in 1982, is currently studying in *School of Automation Science and Electrical Engineering, Beihang University, China*. As the same time, he is a chief engineer of *China Railway 14th Bureau Group Mega Shield Construction Engineering Co., Ltd, Nanjing of Jiangsu Province, China*.
Yixuan Wang, born in 1989, is a lecturer in *Engineering Training Centre, Beihang University, China*. His research interests include fuel and power systems of UAV, fluid control, measurement and control system.
Huaixiang Ma, born in 1985, is currently studying in *School of Automation Science and Electrical Engineering, Beihang University, China*.
Yunhua Li, born in 1963, is a professor in *School of Automation Science and Electrical Engineering, Beihang University, China*.
Zhiguo Yang, born in 1997, is currently studying in *School of Automation Science and Electrical Engineering, Beihang University, China*.
Yan Shi, born in 1981, is a professor in *School of Automation Science and Electrical Engineering, Beihang University, China*. He received his doctoral degree in mechanical engineering from *Beihang University, China*. His research interests include intelligent medical devices and energy-saving technologies of pneumatic systems.

Funding

Supported by National Natural Science Foundation of China and Shanxi Coal-based Low Carbon Joint Fund (Grant No. U1910211); National Natural Science Foundation of China (Grant Nos. 51975024 and 52105044); National Key Research and Development Project (Grant No. 2019YFC0121700).

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article.

Declarations

Competing interests

The authors declare no competing financial interests.

Author Details

[1]School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. [2]China Railway 14th Bureau Group Mega Shield Construction Engineering Co., Ltd., Nanjing 211800, China. [3]School of Mechanical Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China.

Received: 18 March 2022 Revised: 26 July 2023 Accepted: 27 July 2023
Published online: 24 October 2023

References

[1] S Heydari, J K Hamidi, M Monjezi, et al. An investigation of the relationship between muck geometry, TBM performance, and operational parameters: A case study in Golab II water transfer tunnel. *Tunnelling and Underground Space Technology*, 2019, 88(JUN.): 73–86.
[2] H S Jung, J M Choi, B S Chun, et al. Causes of reduction in shield TBM performance – A case study in Seoul. *Tunnelling and Underground Space Technology*, 2011, 26(3): 453–461.
[3] H Y Yang, X H Zhou, G F Gong. Perspectives in intelligentization of tunnel boring machine (TBM). *Tunnel Construction,* 2018, 38(12): 1919–1926.
[4] Q M Gong, L J Yin, H S Ma, et al. TBM tunnelling under adverse geological conditions: An overview. *Tunnelling and Underground Space Technology*, 2016, 57: 4–17.
[5] L Soldo, M Vendramini, A Eusebio. Tunnel design and geological studies. *Tunnelling and Underground Space Technology*, 2018, 84: 82–98.
[6] B D Zhu, G F Gong, R L Zhou, et al. Identification of strata with BP neural network based on parameters of shield driving. *Journal of Zhejiang University*, 2011, 45(5): 851–857. (in Chinese)
[7] L H Wang. Identification of load dynamics parameters of industrial robots based on artificial neural network. *2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA)*, 2020. https://doi.org/10.1109/ICSGEA51094.2020.00033.
[8] E Farrokh, J Rostami. Correlation of tunnel convergence with TBM operational parameters and chip size in the Ghomroud tunnel, Iran. *Tunnelling and Underground Space Technology*, 2008, 23(6):700–710.
[9] Q S Liu, X Huang, Q M Gong, et al. Application and development of hard rock TBM and its prospect in China. T*unnelling and Underground Space Technology,* 2016, 57(Aug.): 33–46.
[10] D Bouayad, F Emeriault. Modeling the relationship between ground surface settlements induced by shield tunneling and the operational and geological parameters based on the hybrid PCA/ANFIS method. *Tunnelling and Underground Space Technology*, 2017, 68(Sep.): 142–152.
[11] Q Wang, X Y Xie, I Shahrour. Deep learning model for shield tunneling advance rate prediction in mixed ground condition considering past operations. *IEEE Access*, 2020, 8: 215310–215326.
[12] Y Z Hernández, A D Farfán, A P D Assis. Three-dimensional analysis of excavation face stability of shallow tunnels. *Tunnelling and Underground Space Technology*, 2019, 92: 103062.
[13] R Malhotra, A Budhiraja, A K Singh, et al. A novel feature selection approach based on binary particle swarm optimization and ensemble learning for heterogeneous defect prediction. *APIT 2021: 2021 3rd Asia Pacific Information Technology Conference,* 2021. https://doi.org/10.1145/3449365.3449384.
[14] R Panthong, A Srivihok. Wrapper feature subset selection for dimension reductionbased on ensemble learning algorithm. *Procedia Computer Science*, 2015, 72(Complete): 162–169.
[15] B Pes, N Dessì, M Angioni. Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. *Information Fusion*, 2017, 35: 132–147.
[16] C Zhen, Y Zhang. A survey of data preprocessing in data mining. *International Core Journal of Engineering,* 2019, 5(9):133–137.
[17] L Jedlinski, J Gajewski. Optimal selection of signal features in the diagnostics of mining head tools condition. *Tunnelling & Underground Space Technology*, 2019, 84: 451–460.
[18] S Fallahpour, M H Zadeh, E N Lakvan. Comparison of wrapper and filtering approaches for corporate failure prediction. *First International Conference on Networks & Soft Computing*. IEEE, 2014. https://doi.org/10.1109/CNSC.2014.6906698.
[19] D J Dittman, T M Khoshgoftaar, R Wald, et al. Comparing two new gene selection ensemble approaches with the commonly-used approach. *International Conference on Machine Learning & Applications.* IEEE Computer Society, 2012. https://doi.org/10.1109/ICMLA.2012.175.
[20] A Woznica, P Nguyen, A Kalousis. Model mining for robust feature selection. *ACM SIGKDD International Conference on Knowledge Discovery and DataMining*.2012. https://doi.org/10.1145/2339530.2339674.
[21] B Y Zhang, A K Qin, T Sellis. Evolutionary feature subspaces generation for ensemble classification. *The Genetic and Evolutionary Computation Conference.* ACM, 2018. https://doi.org/10.1145/3205455.3205638.
[22] S Liang, A J Ma, S Yang, et al. A review of matched-pairs feature selection methods for gene expression data analysis. *Computational & Structural Biotechnology Journal*, 2018. https://doi.org/10.1016/j.csbj.2018.02.005.
[23] J S Olsson, D W Oard. Combining feature selectors for text classification. *Proceedings of the 2006 ACM CIKM International Conference on Information*

and Knowledge Management, Arlington, Virginia, USA, November 6–11, 2006. ACM, 2006. https://doi.org/10.1145/1183614.1183736.
[24] S S Rathore, A Gupta. A comparative study of feature-ranking and feature-subset selection techniques for improved fault prediction. *ACM International Conference Proceeding Series (2014)*, 2014: 1–10. https://doi.org/10.1145/2590748.2590755.
[25] M H Rahman, S Sharmin, S M Sarwar, et al. Software defect prediction using feature space transformation. *The International Conference*. ACM, 2016. https://doi.org/10.1145/2896387.2900324
[26] J Kiani, C Camp, S Pezeshk. On the application of machine learning techniques to derive seismic fragility curves. *Computers & Structures*, 2019, 218(Jul.): 108–122.
[27] K Mahmodi, M Mostafaei, E Mirzaee-Ghaleh. Detection and classification of diesel-biodiesel blends by LDA, QDA and SVM approaches using an electronic nose. *Fuel*, 2019, 258: 116114.
[28] C Bo, C Zang. Artificial immune pattern recognition for structure damage classification. *Computers & Structures*, 2009, 87(21–22): 1394–1407.
[29] S S Lin, S L Shen, N Zhang, et al. Modelling the performance of EPB shield tunnelling using machine and deep learning algorithms. *Earth Science Frontiers,* 2021(5): 81–92
[30] R M Kynch, P D Ledger. Resolving the sign conflict problem for hphexahedral Ndlec elements with application to eddy current problems.*Pergamon*, 2017. https://doi.org/10.1016/J.COMPSTRUC.2016.05.021.
[31] Y Wu, Y Ke, Z Chen, et al. Application of alternating decision tree with adaboost and bagging ensembles for landslide susceptibility mapping. *CATENA*, 2020, 187: 104396.
[32] R Sun, G Y, Wang, W Y Zhang, et al. A gradient boosting decision tree based GPS signal reception classification algorithm. *Applied Soft Computing*, 2020: 86.
[33] K Proniewska, A Pregowska, K P Malinowski. Identification of human vital functions directly relevant to the respiratory system based on the cardiac and acoustic parameters and random forest. *Elsevier Masson*, 2021, 42(3): 174–179.
[34] S Oreski, D Oreski, G Oreski. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications*, 2012, 39(16):12605–12617. https://doi.org/10.1016/j.eswa.2012.05.023.
[35] H Zhang, D Miao, R Wang. A modified Chi2 Algorithm Based on the Significance of Attribute//*IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology Workshops. ACM, 2007.* https://doi.org/10.1109/WI-IATW.2006.13.