Improved Ensemble Feature Selection Based on DT for KPI Prediction

FULIN GAO, SHUAI TAN, HONGBO SHI, YANG TAO, AND BING SONG

Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

Corresponding author: Shuai Tan (tanshuai@ecust.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC1522502 and Grant 2020YFC1522505, and in part by the National Natural Science Foundation of China under Grant 62073141.

ABSTRACT In the production process of large-scale machinery and complex industries, the key performance indicator (KPI) prediction is an essential part of project scheduling and cost estimation. The continuous enrichment of sensor types and functions brings us massive soft-sensing parameters for regression, but also brings severe challenges to algorithm learning. In this paper, an improved ensemble feature selection based on decision tree (EFS-DT) strategy for KPI prediction is developed. On the one hand, the ensemble of multi-criteria filtering results broadens the selector’s perspective without the time cost of superposition. On the other hand, credibility and similarity analysis are designed to eliminate the concerns of Dempster’s combination rule about conflict. After re-evaluating the variable scores, more high-quality variables can be selected to build a more accurate and robust KPI prediction model. Finally, a realistic shield tunnel case in China is used to evaluate the feasibility and effectiveness of proposed approach.

INDEX TERMS Key performance indicator, soft-sensing, ensemble feature selection, decision tree, conflict management.

I. INTRODUCTION

From the perspective of safety, schedule and cost, key performance indicators (KPIs), e.g., the core parameters of major equipments and product quality variables, are vital during construction or production [1], [2]. In recent years, the development of sensors and data analysis technology has made it possible to evaluate KPI, which has attracted widespread attention from scholars [3]–[5]. Sun and Ge [6] combined ensemble and semi-supervised learning to effectively use unlabeled data for KPI prediction. Si et al. [7] divided the kernel matrix into KPI-related and KPI-unrelated subspaces. Then, two statistics are designed for process monitoring based on the improved kernel partial least squares (KPLS) method. Shao et al. [8] applied decision tree (DT) to KPI prediction in the medical field and achieved satisfactory results. DT has become an excellent tool for KPI prediction because it is easy to understand and interpret [9], [10]. However, due to complex working environment, indirect control and difficulty in real-time acquisition of key indicators, it is still extremely necessary and challenging to identify positive soft-sensing parameters and realize accurate KPI prediction.

With the continuous enrichment of sensor types and functions, there are numerous variables, also known as parameters or features, available in DT-based modeling for KPI prediction [11], [12]. However, invalid parameters containing irrelevant or redundant information not only increase the overfitting risk and pruning difficulty, but also adversely affect convergence and performance [13]–[15]. Therefore, how to achieve effective and robust dimensionality reduction becomes particularly critical. Filter models are popular in feature selection of high-dimensional data due to their high efficiency and less computation compared with wrapper and embedded models [16]. However, the accuracy of the separate filter model is not guaranteed due to its one-sided perspective and independence from the algorithm, so the hybrid serial feature selection methods emerge [17], [18]. Based on the feature subset selected by gain ratio (GR), Saengsiri et al. [19] further pruned the candidate features through greedy search (GS) modeled on the support vector machine (SVM). Akadi et al. [20] used genetic algorithm (GA), a wrapper model, to obtain more condensed candidate feature set from feature subset selected by minimum redundancy maximum relevance (mRMR).
However, the serial refinement is strict with the initial selection strategy and shelves the time consuming, so the ensemble methods have aroused interest [21]–[23]. Bolón-Canedo and Alonso-Betanzos [24] introduced the basic concepts of ensemble feature selection, and discussed the state-of-the-art advances and future trends. Jin et al. [25] made two improvements to the objective function in the common space model (CSP), and fused the candidate feature subsets obtained under the two objectives through the Dempster-Shafer (DS) evidence theory. Although ensemble learning has become very popular for classification, there are still rare subjects in the case of ensemble feature selection. Moreover, most of the existing methods are not universal and extensible. Therefore, this study develops an improved ensemble feature selection based on DT (EFS-DT) strategy for KPI prediction. The ensemble feature selection fuses multi-criteria filtering results to retain more informative variables at less time cost.

Ensemble feature selection is to fuse the variable scores from different selectors, i.e., feature selection models. It is helpful to enhance the robustness of feature selection and obtain closer unbiased prediction results [21], [26]. Decision fusion is a strong tool to realize ensemble feature selection. DS evidence theory shines in this field due to its compact computation and no prior knowledge [27], [28]. However, it is always subject to counterintuitive results when dealing with inevitable conflicts of non-homologous decisions [29]. Unfortunately, differences in criteria between selectors often cause significant divergences in feature scores, known as evidence conflicts in decision fusion. This makes reckless fusion infeasible. Therefore, conflict management has received widespread attention from scholars [30], [31]. Li et al. [32] took the consistency of evidences as the basis for weighting, and modified the fusion rules. Song and Deng [33] developed a conflict management method based on evidence belief divergence and information volume. However, most of the existing conflict management methods only focus on the evidence itself, but ignore the credibility of the evidence sources. Since the performance of each selector is different, it is not reasonable to treat all evidences equally. Therefore, this study proposes a novel weight distribution strategy that combines the credibility and similarity of evidences for conflict management. Credibility analysis based on the evaluation results of the selectors puts the fusion eye on evidences with high reliability. Similarity analysis based on support puts the fusion eye on evidences with high-reputation. The stacking of the two makes the final decision more accurate and robust.

The main contributions of this paper are threefold: (1) An improved EFS-DT strategy is developed for KPI prediction, which implements preliminary evaluation and revaluation of variable scores. (2) Ensemble feature selection is utilized to fuse multi-criteria filtering results to retain more informative variables with less time cost. (3) A novel weight distribution strategy is proposed, which uses credibility analysis and similarity analysis for conflict management.

The remainder of this paper is structured as follows: In Section II, variables are scored and modeled based on the multi-criteria selectors. Section III implements KPI prediction based on ensemble feature selection. Section IV gives an overview of the overall framework. Section V applies the proposed method to a real shield tunneling case in China and produces the experimental results. Section VI concludes this paper.

II. MULTI-CRITERIA SCORING AND MODELING

In this section, multiple selectors with different criteria evaluated the quality of variables to obtain variable scores. Then, feature selection is performed based on the scoring results of each selector. Finally, the selected variable sets are used for modeling, and the prediction performance of the models is evaluated.

A. MULTI-CRITERION FILTERING

Multi-criteria filtering is the process in which multiple filter models with different criteria obtain their respective variable scores on the training set.

Given the input data normalized by Z-score, it is tabulated to N samples and M variables $X = \{x_{ij}, i = 1, 2, \ldots, M \}$ and output data $y = \{y_i, i = 1, 2, \ldots, N \}$. The data is split into training, validation and test sets. The ratio of each set is defined as 60%, 20%, and 20%, which are commonly used.

In regression task, the quality of variables can be evaluated according to information, distance and dependence [34]. Therefore, three selectors based on the above criteria are used in this study. They are mRMR, RReliefF and Pearson’s correlation coefficient (PCC) method. In essence, they are all striving to preserve relevance ($x_i,y$) and remove redundancy ($x_i,x_j$)[35]. Their respective evaluation criteria are as follows:

$$\max_{x_i \in X-S_m} \left[ I(x_j; y) - \frac{1}{m} \sum_{x_i \in S_m} I(x_j; x_i) \right]$$

where $S_m$ contains the selected $m$ variables, and the remaining variable set is $X-S_m$. $I(\mathbf{a}; \mathbf{b})$ represents the mutual information between $\mathbf{a}$ and $\mathbf{b}$ [36]. This method, information-based mRMR, adopts an incremental search strategy, so its variable ranking, denoted as $\text{Rank}_m$, is the order in which variables are selected.

$$w_{V} = \frac{P_{diffL|diffV}P_{diffV}}{p_{diffL}} - \frac{(1 - P_{diffL|diffV})P_{diffV}}{1 - p_{diffL}}$$

where $w_{V}$ is the quality score of variable $V$. $P_{diffL|diffV}$, $P_{diffV}$ are probabilistic approximation terms based on relative distance [37]. The variable ranking of this method, distance-based RReliefF, is the descending order of $w$, denoted as $\text{Rank}_R$.

$$r_{x_i,y} = \frac{\sum_{i=1}^{N} (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$
where \( r_{x,y} \) represents PCC between \( x \) and \( y \). \( \bar{x} \) and \( \bar{y} \) are the mean of \( x \) and \( y \) respectively. The variable ranking of this method, dependence-based PCC method, is the descending order of \( r \), denoted as \( \text{Rank}_{p} \).

The preliminary scores of the above selectors for variables can be calculated by Eqs. 1-3. However, the scoring rules are inconsistent due to differences in criteria and perspectives of filter models. Specifically, the incremental selection strategy of mRMR cannot guarantee that each newly selected variable will score lower than the previous one. The scores and gaps of both RReliefF and PCC method gradually decrease with the ranking. Therefore, the reciprocal of the ranking is used as the variable score in this study, which possesses the property that the score and the gap gradually decrease with the ranking, calculated as follows:

\[
\text{Score}_{sm} = \frac{1}{\text{Rank}_{sm}} \tag{4}
\]

where \( \text{Rank}_{sm} \) and \( \text{Score}_{sm} \) represent the ranking and reciprocal score of the filter model \( s \) for the \( m \)th variable, respectively, \( s \) belongs to \{mRMR, RReliefF, PCC method\} in this study.

The variable scores obtained from selectors with different criteria will broaden the one-sided perspective of separate filter model through subsequent ensemble.

**B. MODELING AND EVALUATING**

According to the scores evaluated by each selector, different subsets of variables can be determined for modeling. Evaluating a selector is to evaluate the performance of the variable set selected by it in modeling and prediction.

Classification and regression tree (CRAT) makes DT competent for KPI prediction. The growth of regression tree is a recursive process of constructing piecewise constant approximation based on minimizing error [38]. To be self-content, we summarize the main steps of regression DT as follows:

First, traverse the variable \( j \) and the splitting point \( s \) to find the optimal splitting variable and point \( (j, s) \) that meet the following conditions:

\[
\min_{j, s} \left[ \min_{c_1} \sum_{x_j \in R_1(j, s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_j \in R_2(j, s)} (y_i - c_2)^2 \right] \tag{5}
\]

where \( x_j \) is the \( j \)th variable value of sample \( x_i \). \( R_1(j, s) \) and \( R_2(j, s) \) are the left and right regions of the \( j \)th variable divided by \( s \). \( c_1 \) and \( c_2 \) are the predictions on the above two regions.

Then, the partition and the corresponding outputs are determined by \( (j, s) \):

\[
\hat{\gamma}_d = \frac{1}{N_d} \sum_{x_i \in R_d} y_i, \quad x \in R_d, \quad d = 1, 2 \tag{6}
\]

where \( N_d \) is the number of samples in \( R_d \). \( \hat{\gamma}_d \) represents the optimal prediction of the \( d \)th region, i.e., the mean of labels in this region.

Repeat the above steps in the new sub-regions until the termination conditions are met. Finally, the input space is divided into \( D \) partitions, and the predicted output is denoted as follows:

\[
\hat{y}_i = \sum_{d=1}^{D} \hat{\gamma}_d I(x_i \in R_d) \tag{7}
\]

where \( I(x \in R_d) \) is the indicator function. If the \( i \)th sample is in \( R_d \), the value is 1, otherwise it is 0.

In the training set, the top \( h \) \((h \leq M)\) variables in the ranking results of each filter model in Section II-A are selected to build the DT-based KPI prediction model. Their respective prediction error metrics on the validation set are obtained to form the performance metrics matrix (PMM):

\[
PMM = \begin{bmatrix}
p_{11} & p_{12} & \cdots & p_{1J} \\
p_{21} & p_{22} & \cdots & p_{2J} \\
\vdots & \vdots & \ddots & \vdots \\
p_{S1} & p_{S2} & \cdots & p_{SJ}
\end{bmatrix} \tag{8}
\]

where \( S \) and \( J \) are the number of filter models and performance metrics respectively. \( p_{ij} \) represents the \( j \)th performance metric of filter model \( s \). Three minimal metrics are used in this study: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), defined as follows:

\[
\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{9}
\]

\[
\text{MAE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{10}
\]

\[
\text{MAPE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)} \right| \tag{11}
\]

where \( \epsilon \) is an arbitrarily small but strictly positive number to avoid undefined results. \( y \) and \( \hat{y} \) are the ground truth and estimated target values respectively.

PMM is the evaluation of filtering results, which reflects the quality of each selector. However, the separate selector has limited effect on improving the prediction accuracy. During iteration, the wrapper model recalibrates the selected variables based on the prediction errors to ensure accuracy. Inspired by this, we incorporate the resulting PMM into the subsequent ensemble analysis to assist in the revaluation of variable scores. This not only points the way in favor of ensemble, but is less cumbersome than the wrapper models.

**III. KPI PREDICTION BASED ON ENSEMBLE FEATURE SELECTION**

In this section, the filtering results of the selectors with different criteria are integrated. Integration is completed by decision fusion. Conflict management, i.e., revision, is required before fusion. The process of re-evaluating variable scores through fusion and selecting high-quality variables for modeling is ensemble feature selection. Finally, KPI prediction is implemented based on the result of ensemble feature selection.
A. REVISION OF EVIDENCES

In the revision stage, credibility analysis and similarity analysis pioneered by us will be used to manage the inevitable conflicts between variable scores.

1) CREDIBILITY ANALYSIS

Credibility analysis utilizes the PMM in Section II-B to analyze the performance differences of the selectors. Forward the minimal metrics to obtain new PMM:

\[
PMM' = \begin{bmatrix}
p'_{11} & p'_{12} & \cdots & p'_{1J} \\
p'_{21} & p'_{22} & \cdots & p'_{2J} \\
\vdots & \vdots & \ddots & \vdots \\
p'_{S1} & p'_{S2} & \cdots & p'_{SJ}
\end{bmatrix}
\]

(12)

where \(p'_{sj}\) represents the \(j^{th}\) positive performance metric of the filter model \(s\). Specifically, if \(j = \text{RMSE}\), \(p'_{sj} = 1/p_{sj}\) makes the metric forward. Then, each column is divided by its own 2-Norm to obtain the standardized PMM:

\[
PMM = \begin{bmatrix}
\tilde{p}_{11} & \tilde{p}_{12} & \cdots & \tilde{p}_{1J} \\
\tilde{p}_{21} & \tilde{p}_{22} & \cdots & \tilde{p}_{2J} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{p}_{S1} & \tilde{p}_{S2} & \cdots & \tilde{p}_{SJ}
\end{bmatrix}
\]

(13)

where

\[
\tilde{p}_{sj} = \frac{p'_{sj}}{\sqrt{\sum_{s=1}^{S} (p'_{sj})^2}}
\]

(14)

Determine ideal and nadir solutions:

\[
p^+ = [p^+_1, p^+_2, \cdots, p^+_J]
\]

(15)

\[
p^- = [p^-_1, p^-_2, \cdots, p^-_J]
\]

(16)

where the ideal and nadir solutions correspond to the maximum and minimum values of each column respectively. Specifically, \(p^+_j = \max \{p'_{1j}, p'_{2j}, \cdots, p'_{sj}\}\).

The credibility of the selector can be described by the relative distance between the performance metrics and the ideal and nadir solutions, which is calculated as follows:

\[
w_{\text{cre}}^s = \frac{\text{Dis}_{p^+,p^-}}{\text{Dis}_{p^+,p^+} + \text{Dis}_{p^-,p^-}}
\]

(17)

where \(w_{\text{cre}}^s\) represents the credibility of the filter model \(s\), the value range is \([0,1]\), the closer to 1, the more credible it is. \(p_s = [\tilde{p}_{s1}, \tilde{p}_{s2}, \cdots, \tilde{p}_{sj}]\) and the distance is calculated by

\[
\text{Dis}_{a,b} = \sqrt{\sum_{l=1}^{L} (a_l - b_l)^2}
\]

(18)

where \(L\) is the length of \(a\) and \(b\). The larger the distance, the lower the proximity and similarity.

The variable score vector modified by the credibility analysis is as follows:

\[
\text{Score}'_s = w_{\text{cre}}^s \times \text{Score}_s
\]

(19)

Therefore, credibility analysis puts more bets on selectors with high credibility. In other words, the higher the historical accuracy of the selector, the more dominant the results it produces, and the more accurate the KPI prediction model based on the ensemble results will be.

2) SIMILARITY ANALYSIS

Similarity analysis utilizes distance to measure the support between evidences. The distance metrics matrix (DMM) between variable score pairs after credibility analysis can be obtained through Eq. 18:

\[
\text{DMM} = \begin{bmatrix}
0 & d_{12} & \cdots & d_{1S} \\
d_{21} & 0 & \cdots & d_{2S} \\
\vdots & \vdots & \ddots & \vdots \\
d_{S1} & d_{S2} & \cdots & 0
\end{bmatrix}
\]

(20)

where \(d_{sj}\) represents the distance between the \(s^{th}\) variable score and the \(j^{th}\) variable score, i.e., \(d_{sj} = \text{Dis}_{\text{Score}'_s, \text{Score}'_j}\).

Then, the average difference between the \(s^{th}\) variable score and other variable scores is calculated as follows:

\[
\bar{d}_s = \frac{\sum_{j=1,j\neq i}^{S} d_{sj}}{S-1}
\]

(21)

The correction weight of similarity analysis can be calculated by normalizing the support degree as follows:

\[
\text{Sup}_s = \frac{1}{d_s}
\]

(22)

\[
w_{\text{sim}}^s = \frac{\text{Sup}_s}{\sum_{s=1}^{S} \text{Sup}_s}
\]

(23)

The variable score vector modified by the similarity analysis is as follows:

\[
\text{Score}''_s = w_{\text{sim}}^s \times \text{Score}'_s
\]

(24)

Therefore, the similarity analysis puts more bets on the variable scores with high-reputation (supported by more evidences). In other words, the more a piece of evidence is supported by more other evidences, the more dominant it is. The divergence between the revised evidences is weakened to make the ensemble results more robust.

B. FUSION OF EVIDENCES

In the fusion stage, a unified conclusion is obtained by decision fusion. Before this, the revised variable scores need to be summed and normalized, as follows:

\[
\text{Score} = \text{Norm} \left\{ \sum_{s=1}^{S} \text{Score}''_s \right\}
\]

(25)

where \(\text{Score} = [\text{Score}''_1, \text{Score}''_2, \cdots, \text{Score}''_m]\), and \(\sum_{m=1}^{M} \text{Score}_m = 1\). The purpose of normalization is to convert variable score into probability distribution for subsequent fusion.

DS evidence theory is more and more widespread in decision fusion. It combines multiple specific evidences to derive
abstract fusion conclusions within the framework of discernment [27]. The fusion rule of DS evidence theory, also known as Dempster’s combination rule, is as follows:

$$M(A) = \begin{cases} \frac{1}{1-K} \sum_{B \in C = A} M_1(B)M_2(C), & A \neq \Phi \\ 0, & A = \Phi \end{cases}$$

$$K = \sum_{B \cap C = \Phi} M_1(B)M_2(C)$$

where $\Phi$ is the empty set and $K$ is conflict factor. The range of $M_1(B)$ and $M_2(C)$ is [0, 1], which represent the confidence of the $1^{th}$ and $2^{th}$ evidences (probability distributions) on proposition $B$ and $C$. $M(A)$ is the trust of proposition $A$ resulting from the fusion of two evidences. In this study, the score for individual variable is the confidence of this variable proposition. Specifically, $\text{Score}_m$ is the quality score of the $m^{th}$ variable, i.e., $M(m)$.

Without correction, $S - 1$ fusion needs to be performed, so fusion is still performed $S - 1$ times after correction for fairness, as follows:

$$F(\text{Score}) = \text{Score} \oplus \text{Score} \oplus \cdots \oplus \text{Score}$$

where $\oplus$ represents the fusion operation of Eq. 26. $F(\text{Score})$ is the comprehensive quality score of all variables.

So far, the ensemble has been completed by decision fusion. The variables that stand out in the comprehensive variable rankings will be used to build KPI prediction models.

**C. KPI PREDICTION**

The variable subsets with cardinality $h$ can be selected from the training, validation and test sets of $X$ according to $F(\text{Score})$, denoted as $\mathbf{I}_{\text{train}}^h$, $\mathbf{I}_{\text{val}}^h$, and $\mathbf{I}_{\text{test}}^h$. The corresponding real KPIs are $y_{\text{train}}^h$, $y_{\text{val}}^h$, and $y_{\text{test}}^h$, respectively. $h$ is determined at the lowest point of the curve where the performance of the model on the validation set varies with the number of retained variables.

In the prediction phase, $\mathbf{I}_{\text{train}}^h$ and $y_{\text{train}}^h$ are fed into the DT model for training. The model is then verified on $\mathbf{I}_{\text{val}}^h$ and $y_{\text{val}}^h$. After adjusting $h$, retraining and revalidation are performed to obtain the model performance curve with the number of retained variables. The $h$ corresponding to the lowest point of the curve is the number of variables to be retained. When unseen data $\mathbf{I}_{\text{test}}^h$ arrives, the predicted output $y_{\text{test}}^h$ of KPI can be obtained through Eq. 7. Ultimately, the prediction performance can be evaluated and compared in terms of Eqs. 9-11 and time consumption.

**IV. SOLUTION OVERVIEW**

The flowchart of improved EFS-DT for KPI prediction is depicted in Fig. 1, which mainly includes multi-criteria filtering, modeling and evaluating, revision and fusion of evidences, and KPI prediction. In order to avoid the risks of dimensionality disasters, overfitting caused by high model complexity, and high computational cost of serial feature selection, the ensemble idea runs through the whole paper. In the stage of multi-criteria filtering, multiple filter models with different criteria evaluate and score variables according to their own standards. In the stage of modeling and evaluating, the variables selected according to the filtering results are fed into the DT model, and the PMM is obtained. In the stage of evidence revision, conflict management is performed through credibility analysis and similarity analysis. In the stage of evidence fusion, the revised variable scores are fused by Dempster’s combinational rule to determine the variable subset that contributes more to KPI prediction. Finally, the DT-based KPI prediction model is built and the prediction performance is evaluated.

**V. EXPERIMENTAL STUDIES**

**A. CASE BACKGROUND**

Tunnel boring machines (TBMs) are widely used in underground projects due to their efficiency, safety, and economy. Since the introduction of TBMs in the 1950s, engineers and contractors have been committed to study their mechanisms,
operation analysis of core components, and the TBM-soil interactions. The development of sensors and data analysis technology has made it possible to evaluate the KPI of TBM, i.e., advance rate (AR). However, the operator cannot control AR directly, but indirectly by adjusting other parameters, such as thrust, torque, etc. Therefore, identifying positive soft-sensing parameters and realizing accurate KPI prediction are critical to project safety and construction schedule.

The application case in this study is the right line of the double-lane tunnel crossing the Q river in Hangzhou, China, which connects the north and south shores. The design mileage is YK0+000.000∼YK3+586.566, of which the length of the shield section is 1830m (YK1+000.000∼YK2+830.000). Fig. 2 depicts the cross-section and overview of the Q river tunnel project. According to Article 3.1.1∼3.1.4 of “Geotechnical Engineering Survey Code” (2009 edition): This project is a river-crossing tunnel with level I (importance). The site has diverse rock and soil types, poor uniformity, and variable geology, so the foundation complexity is II (medium).

B. EXPERIMENT RESULTS AND DISCUSSION
The performance of TBM is closely related to the complex geological environment. However, precise geological surveys, limited by above-ground construction, water depth and cost, are a luxury. Therefore, this study divides the geology into different soil units based on geological exploration report. It lays the foundation for the implementation of the feature engineering and data analysis in Sections II-III for different soil units. Next, the analysis and comparison of experiments will be conducted from three aspects: stratigraphic division, feature selection, and KPI prediction. The experiments are performed on a Windows Server with a dual-3.60-GHz CPU and a RAM of 16GB.

1) THE STUDY OF STRATIGRAPHIC DIVISION
In this project, the exploration points are 20∼25m apart, and the hole depth is 1.5∼2 times the tunnel diameter below the tunnel buried depth. However, the difference in the sampling frequency of long-interval geological exploration and operating parameters (1 sample every 6 sec) makes data supplement or alignment infeasible. Nevertheless, TBM often exhibit distinct performance on different cross-sections, but analogous behavior on similar cross-sections at different locations. Therefore, this study divides the strata into four major engineering soil units (ESUs) based on the geological exploration report. The sections S1 to S4 of ESUs are presented in Fig. 2. As shown in the figure, there are 915 rings in the project. S1 is composed of descending section from 0∼147 rings and ascending section from 801∼915 rings, S2 is 148∼380 rings, S3 is 381∼554 rings and S4 is 555∼800 rings. The geological composition of each ESU is diverse, but the soil properties within individual units are similar. On this basis, ensemble feature selection can be implemented and KPI prediction models can be built in different soil units. Geotechnical conditions are also implicit in the analysis.

2) THE STUDY OF FEATURE SELECTION
The parameters collected by TBM have direct or indirect influence on the tunneling quality mainly concentrated in three modules. They are the main drive electrical module, the propulsion module and the mud conveying module. Parameters that do not contribute to the TBM KPI prediction are excluded from candidates, such as binary power flags, component positions, and parameters that are valid only in a specific mode. In particular, weekend mode is rarely used on-site. The final candidates for feature selection are the remaining 99 sensor parameters. The sensor types involved are listed in Table 1, one of which may correspond to multiple sensors in different locations.

However, feeding all the candidate parameters into the KPI prediction model is computationally intensive. Therefore, the improved EFS-DT strategy is implemented to find out the raw parameters in each ESU that contribute to their KPI prediction. The number of retained parameters in each ESU is determined by the performance on the validation set. The metrics of predictive performance include RMSE, MAE and MAPE. The results are shown in Fig. 3. In order to facilitate
the discovery of the comprehensive optimum, the three minimal prediction performance metrics are mapped to $[0, 1]$ by Max-Min normalization. The comprehensive optimum, i.e., minimum sum, are obtained at $S_1 = 32$, $S_2 = 24$, $S_3 = 22$, and $S_4 = 30$ respectively. The dimension is reduced to less than $1/3$ of the original number.

The variable scores and selection results under the four ESUs are shown in Fig. 4. The green bars indicate the selected variables, and the # in the horizontal axis ticks separate the sensor type and project number. As shown in Fig. 4, the scores of variables ranked below 40 are generally less than 0.1. Increasing low-quality variables will not only adversely affect the prediction, but also increase the burden of learning algorithm. Moreover, the number of retained variables is too small to provide sufficient information, which will also lead to performance degradation. Therefore, for a clearer view,
Fig. 3 only shows the fusion performance curve with the number of retained variables between 10 and 40, while Fig. 4 only shows the fusion scores of top 40 variables with positive contributions.

In Fig. 4, there are significant diversities in the number and type of retained variables in different units, indicating the geological differences and the necessity of distinctive analysis. Variable scores and gaps gradually decrease with ranking, which is also consistent with intuitive perception. In all units, the score of AR (F11 in Table 1) is the highest, indicating that it contributes the most to KPI prediction. Since the data fed into the model are current samples, it is not surprising that the KPI prediction for the next moment appears autoregressive, i.e., F11 is retained. Overall, the proposed method effectively reduces the data dimension, thus revealing more granularly how the raw parameters affect KPI.

3) THE STUDY OF KPI PREDICTION

After ensemble, the variable subset of each ESU used to build the KPI prediction model is determined. In order to verify the effectiveness of this strategy, the proposed method is compared with other separate filter models, conventional serial feature selection methods. The results on the test set are listed in Table 2. In the time columns, + separate the time loss of two stages. They are the time cost for feature selection and the time required for modeling and evaluation on the validation set using the selected variables. In order to ensure the fairness of the comparative experiment, the proposed method and other methods participating in the comparison all use DT model with default hyperparameters for KPI prediction.

As shown in Table 2, the absence of feature selection saves time in the screening process. However, it shifts the burden to modeling and cannot eliminate the negative impact of irrelevant or redundant information on prediction accuracy. The prediction errors of the three separate filter models in all ESUs are reduced compared with no feature selection. The optimization of time cost in modeling and evaluation stage is also very obvious, about 2~3 times more. Interestingly, the most mundane method of the three filter models,
the PCC method, not only consumes less selection time, but also achieves considerable performance. This indicates that variables with strong linear correlation with AR in this study are not only easier to establish the mapping relationship with the labels, but also contribute greatly to the prediction.

Compared with the filter model alone, the prediction errors of the serial methods are further reduced. This is attributed to the serial feature selection which relieves the one-sidedness of the separate filter model. However, they cannot get rid of the inherent defects, i.e., the time-for-accuracy and the dependence on the reliability of the initial feature selection result. Compared with the most time-consuming filter models, the proposed method only takes about 4s to achieve greater improvement, while the serial methods take 3 times or even 10 times more time. This is because once the revision weights are determined in the validation phase, the computation required for fusion will be minimal. The proposed method is far ahead in prediction accuracy. Moreover, since the proposed method is the fusion of filter models with natural stability, it is not as elusive as the heuristic method. Therefore, it is more conducive to lock the soft-sensing parameters with positive effects. In summary, the effectiveness and feasibility of the proposed method in feature selection are verified.

In order to verify the effectiveness of the proposed revision and fusion scheme, the comparison results with other revision and fusion methods are listed in Table 3. Since the fusion time is minimal, time cost is not involved in this comparison. In the table, soft voting and DS evidence theory directly integrate the variable scores of the three filter models, i.e., evidences, without revision. Compare with the results in Table 2, the reckless fusion is worse than the separate filter models, even the RMSE of DS evidence theory in $S_1$ and $S_2$ is close to no selection. The revision methods have hardly improved or even slightly impairs the prediction accuracy. This is because the revision strategy is adopted without considering the reliability of the evidence sources, and the less credible evidence has brought a devastating blow to the fusion. The proposed method implements evidence revision through credibility analysis and similarity analysis. The results demonstrate that improved EFS-DT strategy retains more informative parameters that the filter models pay attention to from different perspectives, and achieves more accurate predictions. Overall, the effectiveness and feasibility of the proposed method in conflict management and fusion are verified.

VI. CONCLUSION

In this paper, an improved EFS-DT strategy is developed for KPI prediction. On the one hand, it integrates multi-criteria filtering results to retain more informative variables with less time cost. On the other hand, it uses credibility analysis and similarity analysis for conflict management to achieve more accurate and robust prediction. Finally, the feasibility and effectiveness of the proposed method are evaluated with a real shield tunnel case in China. The experimental results indicate that the proposed method can achieve effective dimensionality reduction and accuracy improvement with as little time overhead as possible.

ACKNOWLEDGMENT

The authors would like to thank China Railway 14th Bureau Group Shield Engineering Company Ltd. for providing actual...
data about the operation process of the shield machine, and Shouzhi Guo, Lei Guo, and Yao Ma for their kind help.

REFERENCES

[1] H. Yu, Y. Li, and L. Li, “Evaluating some dynamic aspects of TBMs performance in uncertain complex geological structures,” Tunnelling Underground Space Technol., vol. 96, Feb. 2020, Art. no. 103216.

[2] J. Li, F. Li, D. Guo, X. Li, and Z. Chen, “Advanced prediction of tunnel boring machine performance based on big data,” Geosci. Frontiers, vol. 12, no. 1, pp. 331–338, Jan. 2021.

[3] J. Rostami, “Performance prediction of hard rock tunnel boring machines (TBMs) in difficult ground,” Tunnelling Underground Space Technol., vol. 57, pp. 173–182, Aug. 2016.

[4] M. Z. Naghadehi and A. Ramezanzadeh, “Models for estimation of TBM performance in granitic and mica gneiss hard rocks in a hydropower tunnel,” Bull. Eng. Geol. Environ., vol. 76, no. 4, pp. 1627–1641, Nov. 2017.

[5] M. Koopaliipoor, H. Tootoonchi, D. J. Armaghani, E. T. Mohamad, and A. Hedayat, “Application of deep neural networks in predicting the penetration rate of tunnel boring machines,” Bull. Eng. Geol. Environ., vol. 78, no. 8, pp. 6347–6360, Dec. 2019.

[6] Q. Sun and Z. Ge, “Deep learning for industrial KPI prediction: When ensemble learning meets semi-supervised data,” IEEE Trans. Ind. Informat., vol. 17, no. 1, pp. 260–269, Jan. 2021.

[7] Y. Si, Y. Wang, and D. Zhou, “Key-performance-indicator-related process monitoring based on improved kernel partial least squares,” IEEE Trans. Ind. Electron., vol. 68, no. 3, pp. 2626–2636, Mar. 2021.

[8] C. Shao, K. Liu, C. Li, Z. Cong, L. Hu, J. Luo, Y. Diao, Y. Xu, S. Ji, Y. Qiang, and Y. Shen, “C-Reactive protein to albumin ratio is a key indicator in a predictive model for anastomosis leakage after esophagectomy: Application of classification and regression tree analysis,” Thoracic Cancer, vol. 10, no. 4, pp. 728–737, Apr. 2019.

[9] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcír, E. Herrera-Viedma, H. Fujita, and N. A. M. Ghani, “Multiclass prediction model for student grade prediction using machine learning,” IEEE Access, vol. 9, pp. 95608–95621, 2021.

[10] Y.-Y. Zhang, W. Wu, and H. Liu, “Factors affecting variations of soil pH in different horizons in hilly regions,” PLoS ONE, vol. 14, no. 4, Jun. 2019, Art. no. e0218563.

[11] Y. Feng, Z. Chen, H. Luo, S. Wang, Y. Zhao, L. Liu, D. Ling, and L. Ling, “Tunnel boring machines (TBMs) performance prediction: A case study using big data and deep learning,” Tunnelling Underground Space Technol., vol. 110, Apr. 2021, Art. no. 103636.

[12] S. Mokhtari and M. A. Mooney, “Predicting EPBM advance rate per-operation, and fault monitoring and diagnosis,” IEEE Access, vol. 10, Oct. 2020, Art. no. 103520.

[13] M. Shafiq, S. Tian, A. K. Bashir, D. G. Yoon, and M. Guizani, “CorrAUC: ensemble learning meets semi-supervised data,” PLoS ONE, vol. 16, no. 4, pp. 1–15, Apr. 2021.

[14] A. Hedayat, “Application of deep neural networks in predicting the penetration rate of tunnel boring machines,” Bull. Eng. Geol. Environ., vol. 78, no. 8, pp. 6347–6360, Dec. 2019.

[15] G. Cerruela-García, A. de Haro-García, J. P.-P. Toledano, and N. García-Pedrajas, “Improving the combination of results in the ensembles of prototype selectors,” Neural Netw., vol. 118, pp. 175–191, Oct. 2019.

[16] F. Xiao, “GIQ: A generalized intelligent quality-based approach for fusing multisource information,” IEEE Trans. Fuzzy Syst., vol. 29, no. 7, pp. 2018–2031, Jul. 2021.

[17] Z. Wang and F. Xiao, “An improved multi-source data fusion method based on the belief entropy and divergence measure,” Entropy, vol. 21, no. 6, pp. 611, Jun. 2019.

[18] Z. Su, F. Wang, H. Xiao, H. Yu, and S. Dong, “A fault diagnosis model based on singular value manifold features, optimized SVMs and multi-sensor information fusion,” Meas. Sci. Technol., vol. 31, no. 9, Sep. 2020, Art. no. 095002.

[19] L. Fei, Y. Deng, and Y. Hu, “DS-VIKOR: A new multi-criteria decision-making method for supplier selection,” Int. J. Fuzzy Syst., vol. 21, no. 1, pp. 157–175, 2019.

[20] M. N. Khan and S. Anwar, “Paradox elimination in Dempster–Shafer combination rule with novel entropy function: Application in decision-level multi-sensor fusion,” Sensors, vol. 19, no. 21, pp. 4810, Nov. 2019.

[21] F. Xiao, “Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy,” Entropy, vol. 46, pp. 23–32, Mar. 2019.

[22] Y. Li, A. Wang, and X. Yi, “Fire control system operation status assessment based on information fusion: Case study,” Sensors, vol. 19, no. 10, p. 2222, May 2019.

[23] Y. Song and Y. Deng, “A new method to measure the divergence in evidential sensor data fusion,” Int. J. Distrib. Sensor Netw., vol. 15, no. 4, Apr. 2019, Art. no. 155014771984129.

[24] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, “Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 13, no. 5, pp. 971–989, Sep. 2016.

[25] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” Neurocomputing, vol. 300, pp. 70–79, Jul. 2018.

[26] E. Hancer, B. Xue, and M. Zhang, “Differential evolution for filter feature selection based on information theory and feature ranking,” Knowl.-Based Syst., vol. 140, pp. 103–119, Jan. 2018.

[27] B. Li, X. Xu, L. Zhang, J. Han, C. Bian, G. Li, J. Liu, and L. Jin, “Above-ground biomass estimation and yield prediction in potato by using UAV-based RGB and hyperspectral imaging,” ISPRS J. Photogramm. Remote Sens., vol. 162, pp. 161–172, Apr. 2020.

[28] W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, D. T. Bui, Z. Duan, and J. Ma, “A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility,” Catena, vol. 151, pp. 147–160, Apr. 2017.

[29] FULIN GAO received the B.S. degree in electrical engineering and automation from Ludong University, China, in 2019. He is currently pursuing the master’s degree in control science and engineering with the School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China. His research interests include information fusion, operation status assessment of industrial process, and fault monitoring and diagnosis.
SHUAI TAN received the B.S. degree in automation and the Ph.D. degree in control theory and control engineering from Northeastern University, China, in 2005 and 2012, respectively. She is currently an Associate Professor with the East China University of Science and Technology, Shanghai, China. Her research interests include operation state evaluation for complex industrial process, fault monitoring and diagnosis, and machine learning of image information.

HONGBO SHI received the B.E. degree in chemical automation and the Ph.D. degree in control theory and control engineering from the East China University of Science and Technology, China, in 1986 and 2000, respectively. He was the 2003 Shu Guang Scholar of Shanghai. He is currently a Professor with the East China University of Science and Technology. His research interests include modeling of industrial process and advanced control technology, theory and methods of integrated automation systems, and condition monitoring and fault diagnosis of industrial process.

YANG TAO received the B.E. degree in automation from Zhengzhou University, Zhengzhou, China, in 2015. He is currently pursuing the Ph.D. degree in control theory and control engineering with the East China University of Science and Technology, Shanghai, China. His research interests include feature extraction, process operating performance assessment, fault detection, and fault diagnosis.

BING SONG received the B.E. degree in automation and the Ph.D. degree in control theory and control engineering from the East China University of Science and Technology, Shanghai, China, in 2012 and 2017, respectively. He is currently an Associate Professor with the Department of Automation, East China University of Science and Technology. His research interests include feature extraction, fault detection, fault diagnosis, and multimode process monitoring.