ARTICLE IN PRESS

Applied Soft Computing Journal xxx (xxxx) xxx

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model

Pin Zhang

Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
School of Civil Engineering & Built Environment, Science and Engineering Faculty, Queensland University of Technology (QUT), Brisbane, Qld 4001, Australia

ARTICLE INFO

Article history:
Received 31 July 2019
Received in revised form 2 October 2019
Accepted 10 October 2019
Available online xxxx

Keywords:
Shield tunnel
Settlement
Machine learning
Sensitivity analysis
Feature selection
Principal components analysis

ABSTRACT

Feature selection (FS) is vitally important for determining the optimum subsets of features with effective information and maximizing the model accuracy. This study proposes a novel FS method based on global sensitivity analysis (GSA) for effectively determining the most relevant feature subsets and improving prediction performance of machine learning (ML)based models. Feature ranking is determined based on the results obtained from three global sensitivity analysis (GSA) including *Pearson*, *Sobol'* and PAWN. This novel GSA-based FS method is applied to engineering practice with the combination of ML algorithm random forest (RF) to predict tunnelling-induced settlement prediction model. Meanwhile, the feature extraction method principle component analysis (PCA) is also used to develop RF-based model for comparing the performance of proposed GSA-based FS method. The results indicate the novel GSA-based FS method effectively determines the significance of input variables. The prediction performance of RF-based model with the integration of GSA-based FS methods is enhanced dramatically, and obviously outperforms the model with the integration of PCA-based dimensionality reduction method.

1. Introduction

Machine learning (ML) algorithms have been extensively used in engineering practice because of the strong applicability of addressing non-linear and high-dimensional problems [1–3]. However, these data with high dimensions pose a challenge for data analysis and decision-making. Feature selection (FS) method has thus been proposed to process high-dimensional data [4,5]. FS approach refers to determine the best set of features with maximum information to maximize the model accuracy. Cai, et al. [6] gave a list of the FS method classification according to different criteria. In general, the FS method is categorized into three groups: filter, wrapper and embedded models, considering their relationship with learning methods.

Filter methods such as correlation coefficient choose feature subsets based on the relationships among parameters, which means that such methods are independent of any learning algorithm, completing FS as a pre-processing step before start a learning process. Filter methods are thus characterized by low computational cost and instability due to the independency with models. Wrapper methods such as recursive feature elimination require integration with a ML-based model to perform FS. Various combinations of features are used to establish ML-based models [7]. The optimum feature subset is determined as the ML-based model achieves the best performance. Therefore, wrapper methods tend to achieve better performance and have a smaller dimensionality of parameters, but the interaction with ML-based model and the numerous combinations of features inevitably increase computational cost. Embedded models lie between filter and wrapper methods, in which the FS is integrated with the training process of a ML-based model. The most typical embedded methods-based ML algorithm is decision tree [8]. FS and the training of ML-based models are synchronically completed, thereby the computational cost of embedded methods is less than wrappers with the sacrifice of model performance, because the noisy or irrelevant parameters cannot be eliminated in embedded methods. Such factors as mentioned above indicate a novel FS method that can overcome these problems deserves to be developed.

Sensitivity analysis (SA) aims to investigate how model output uncertainty can be apportioned to the uncertainty in each input variable [9], thereby determine the significance of input variable to the output variable. SA methods is predominantly classified into two types: qualitative and quantitative methods [10], as shown in Fig. 1. Scatter plot is the representative of the qualitative method, which presents visual indication of the influence of individual inputs on an output [11]. The failure to objectively ranking the influence of input variables limits the application of

E-mail address: zhangp@hnu.edu.cn.

https://doi.org/10.1016/j.asoc.2019.105859
1568-4946/

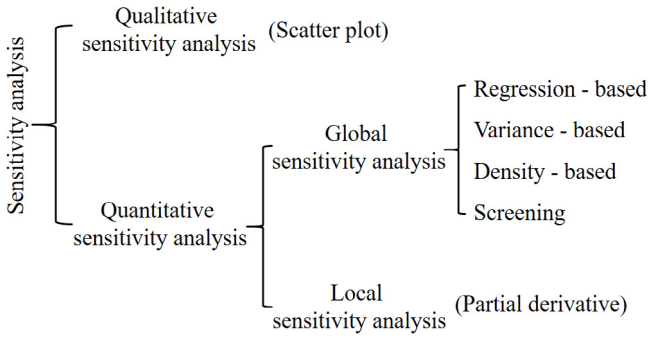

Fig. 1. Category of sensitivity analysis methods.

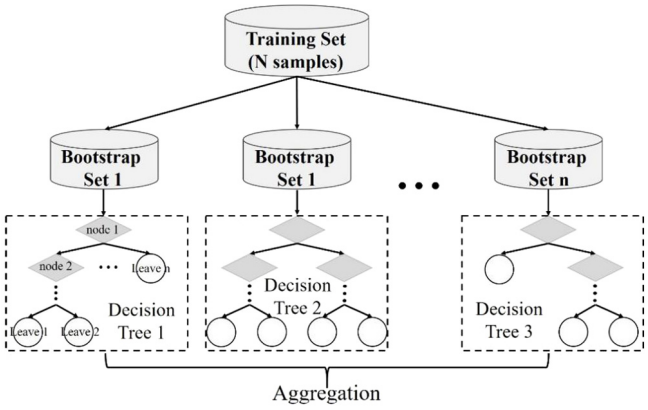

Fig. 2. The schematic view of random forest algorithm.

this method, but it still can be used to complement the results of quantitative methods for better representation [12].

Quantitative SA involves local and global methods. The emphasis of local sensitivity analysis (LSA) is on the model itself, which is implemented by taking partial derivatives of the output function with respect to input variables [9]. LSA is thus not suitable for the problems that input and output variables exist non-linear relationships or the correlations exist among input variables. To eliminate the interaction among input variables on the output variable, global sensitivity analysis (GSA) methods that can account for the whole variable space was proposed to reveal real correlation between input and output variables and have been successfully used in many domains [13–15]. Therefore, GSA is a reliable and method to objectively evaluate the significance of input variables to output variables in a given system.

The model developed upon ML methods is generally termed as "black box" because the physical mechanism of the model is not clear. The integration of GSA and ML methods is helpful to understand the meaning behind the "black box". Zhang, et al. [16] proposed a ML-based model to predict tunnelling-induced building damage, and demonstrated the importance of input variables to the outputs using extended Fourier amplitude sensitivity analysis (EFAST). Liu, et al. [17] considered the same ML and SA methods to evaluate the influence of input variables on the structural safety of tunnel segment linings. Nevertheless, the main characteristic in these published research works is that GSA is merely utilized for evaluating the importance of input variables to the output variables after establish a ML-based model. The attempt of identifying the optimum combination of input features based on SA results for enhancing prediction performance of model is not conducted, although it is more important in engineering practice. Meanwhile, aiming at a specific engineering problem, there is no research to compare the performance of FS and feature extraction approaches such as principle component analysis (PCA) with respect to improving model accuracy, although both of these two approaches are developed for reducing the dimension of variables [18].

To resolve these problems, a new FS method based on GSA results is proposed for improving accuracy of ML-based prediction models and determining the optimum input variables in a faster way. Four GSA approaches including scatter plot, *Pearson* correlation coefficient, *Sobol'* variance-based sensitivity index [19] and PAWN method [20] are utilized to evaluate the significance of input variables to the output variables. To reduce the uncertainty of GSA approaches, a new feature ranking criteria is proposed based on the sum of scores of input variables calculated by these GSA methods. Thereafter, hybrid filter-wrapper modelling process is conducted to determine the optimum feature subsets. This proposed FS method is applied to a practical tunnel engineering with the integration with ML algorithm random forest (RF) for predicting tunnelling-induced settlement. Meanwhile, the integration of RF model and the feature extraction method PCA is synchronously conducted for comparing the performance of proposed GSA-RF method.

2. Methodology

2.1. Random forest

Random forest (RF) as an ensemble learning method has been recently applied to engineering practice [21–23]. Two powerful ML techniques, bootstrap aggregating [24] and random subspace [25] are integrated into RF method. The generalization of this method is thus excellent because it aggregates results of numerous decision trees, but the computational cost is large. In this algorithm, n bootstrap sets with arbitrary samples and features are extracted from the training set using the bagging method. The bootstrap sets are employed to train decision trees. Each node tests a particular feature, and the leaves of the tree represent the output labels, as shown in Fig. 2. The final result is obtained by aggregating the outputs from all leaves [26], which can be expressed as:

$$y = \frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} y_i(x) \tag{1}$$

where y = average output of a total amount of n_{tree}; $y_i(x)$ = individual prediction of a tree for an input vector x.

2.2. Global sensitivity analysis

2.2.1. Variance-based global sensitivity analysis

Variance-based sensitivity indices represent the contribution to the output due to variation of an input variable. Eq. (2) represents a model with k features, and the output y is a scalar.

$$\boldsymbol{Y} = f(\mathbf{X}) \tag{2}$$

where $\mathbf{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_k]$. $f(\mathbf{X})$ = the output of RF-based prediction model in this study. The process of variance-based GSA starts from analysis of variance (ANOVA) functional decomposition [27]:

$$Y = f_0 + \sum_{i=1}^{k} f_i + \sum_{i=1}^{k} \sum_{i<j} f_{ij} + \ldots + f_{12\ldots k} \tag{3}$$

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

where $f_i = f_i(\boldsymbol{X}_i)$, $f_{ij} = f_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ and so on. In this condition, the total variance can be defined as:

$$V(\boldsymbol{Y}) = \sum_{i=1}^{k} V_i(\boldsymbol{Y}) + \sum_{i=1}^{k}\sum_{i<j} V_{ij}(\boldsymbol{Y}) + \ldots + V_{ij\cdots k}(\boldsymbol{Y}) \tag{4}$$

where $V_i = V(f_i(\boldsymbol{X}_i))$, $V_{ij} = V(f_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j))$ and so on.

The *Sobol*' sensitivity index is defined as the ratio of partial variance to total variance [19]. The first and total sensitivity index are give by:

$$S_i = \frac{V_{Xi}(E_{X\sim i}(\boldsymbol{Y}|\boldsymbol{X}_i))}{V(\boldsymbol{Y})} \tag{5}$$

$$S_{Ti} = 1 - \frac{V(E(\boldsymbol{Y}|\boldsymbol{X}_{\sim i}))}{V(\boldsymbol{Y})} = \frac{E(V(\boldsymbol{Y}|\boldsymbol{X}_{\sim i}))}{V(\boldsymbol{Y})} \tag{6}$$

where $\boldsymbol{X}_i$ = the ith feature; $\mathbf{X}_{\sim i}$ = matrix of all features without including $\boldsymbol{X}_i$; $\boldsymbol{Y}/\boldsymbol{X}_i = \boldsymbol{Y}$ values under a given $\boldsymbol{X}_i$; $\boldsymbol{Y}/\mathbf{X}_{\sim i} = \boldsymbol{Y}$ values under $\mathbf{X}_{\sim i}$ matrix. S_i = first-order index, measuring the effect of parameter X_i on the total variance. S_{Ti} = total-order index, measuring the effect of parameter $\boldsymbol{X}_i$ as well as the interaction of $\boldsymbol{X}_i$ and other variables.

Numerous estimators for S_i and S_{Ti} have been proposed [27,28]. In this study, first-order index proposed by Saltelli, et al. [29] and total-order index developed by Jansen [30] are utilized, as shown in Eqs. (7)–(9). The superiority of such two estimators has been demonstrated by Saltelli, et al. [29].

$$V_{Xi}(E_{\sim i}(Y|X_i)) = \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{B})_j \left(f\left(\mathbf{A}_{\mathrm{B}}^{(i)}\right)_j - f(\mathbf{A})_j \right) \tag{7}$$

$$E(V(Y|X_{\sim i})) = \frac{1}{2N}\sum_{j=1}^{N} \left(f(\mathbf{A})_j - f\left(\mathbf{A}_{\mathrm{B}}^{(i)}\right)_j \right)^2 \tag{8}$$

$$V(Y) = \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A})_j^2 - \left(\frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A})_j \right)^2 \tag{9}$$

where $\mathbf{A}$, $\mathbf{B}$ = two independent sampling matrices with $N/2$ (N is the total number of data in the database) sets of data. $\mathbf{A}_{\boldsymbol{B}}^{(i)}$ = all columns are from $\mathbf{A}$ except the ith column which is from $\boldsymbol{B}$. That is, a model with k features means it has k different $\mathbf{A}_{\boldsymbol{B}}$ matrices. The composition of matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A}_{\boldsymbol{B}}^{(i)}$ are shown in follows:

$$\mathbf{A} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots x_i^{(1)} & \cdots x_k^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots x_i^{(2)} & \cdots x_k^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(N/2)} & x_2^{(N/2)} & \cdots x_i^{(N/2)} & \cdots x_k^{(N/2)} \end{bmatrix} \tag{10}$$

$$\mathbf{B} = \begin{bmatrix} x_1^{(N/2+1)} & x_2^{(N/2+1)} & \cdots x_i^{(N/2+1)} & \cdots x_k^{(N/2+1)} \\ x_1^{(N/2+2)} & x_2^{(N/2+2)} & \cdots x_i^{(N/2+2)} & \cdots x_k^{(N/2+2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(N)} & x_2^{(N)} & \cdots x_i^{(N)} & \cdots x_k^{(N)} \end{bmatrix} \tag{11}$$

$$\mathbf{A}_{\mathrm{B}}^{(i)} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots x_i^{(N/2+1)} & \cdots x_k^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots x_i^{(N/2+2)} & \cdots x_k^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(N/2)} & x_2^{(N/2)} & \cdots x_i^{(N)} & \cdots x_k^{(N/2)} \end{bmatrix} \tag{12}$$

2.2.2. Density-based global sensitivity analysis

Liu, et al. [31] pointed out that variance-based sensitivity indices fail to rank the input variables of a model with a highly skewed distribution. To this end, tons of research works have conducted to develop moment-independent sensitivity indices, that is, density-based method. This approach evaluates the significance of input variables to output variables based on the probability density function (PDF) of the model output. Due to the massive computational costs of PDF, Pianosi and Wagener [20] proposed a novel method which characterizes the distribution of output by its cumulative distribution function (CDF). The advantage is that the appropriation of an empirical CDF from model output comes at no computational costs and does not require to tune any parameter. In this study, CDF is used to characterize the conditional and unconditional distributions of the model output.

The determination of a model is consistent with Eq. (2). The unconditional CDF of model output y is defined by $F_Y(\boldsymbol{Y})$ and the conditional CDF is defined by $F_{Y|X_i}(\boldsymbol{Y})$ as x_i is fixed. $F_{Y|X_i}(\boldsymbol{Y})$ explains the variation of the distribution of model output due to the change of x_i value. The distance between $F_Y(\boldsymbol{Y})$ and $F_{Y|X_i}(\boldsymbol{Y})$ is able to reflect the effect of x_i on the model output, which is quantified by Kolmogorov–Smirnov statistic [32,33]:

$$KS(X_i) = \max \left| F_Y(Y) - F_{Y|X_i}(Y) \right| \tag{13}$$

Each input variable $\boldsymbol{X}_i$ has a corresponding maximum KS value, which is used to determine feature prioritization and select feature. The detailed steps for the implementation of the CDF are presented in Fig. 3 and the meaning behind each step is shown as following:

(e) Derive the unconditional CDF of the $\boldsymbol{Y}$;
(f) Generate n random samples to be used as conditioning values of $\boldsymbol{X}_i$;
(g) Compute conditional CDF of the settlement based on conditioning values of $\boldsymbol{X}_i$;
(h) Compute the maximum KS statistic between the conditional and unconditional CDFs.

2.3. Principle component analysis

Principle component analysis (PCA) reduces the dimension of input variables by using an orthogonal transformation. PCA creates a set of new variables called principle components (PCs), which are the linear combination of original variables. This method starts by the standardization of input variables matrix:

$$\mathbf{M}_i = \frac{\left(\boldsymbol{X}_i - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{X}_i^{(j)} \right)}{\sqrt{\frac{1}{N-1}\left(\sum_{j=1}^{N} \left(\boldsymbol{X}_i - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{X}_i^{(j)} \right)^2 \right)}} \tag{14}$$

where the reason behind using $N-1$ instead of N to calculate the covariance is Bessel's correction. $\mathbf{M}$ = the standardized input variables matrix ($N \times k$). Based on this, the covariance matrix $\mathbf{C}$ is calculated by:

$$\mathbf{C} = \frac{1}{N-1}\mathbf{M}_i \bullet \mathbf{M}_i^T \tag{15}$$

Then the eigenvectors $\mathbf{V}_v$ ($k \times k$) and the eigenvalues $\mathbf{V}_s$ ($k \times 1$) of the correlation matrix $\mathbf{C}$ is obtained:

$$\mathbf{C} \bullet diag(\mathbf{C}) \bullet \mathbf{C} \bullet \mathbf{V}_s = \mathbf{V}_s \bullet \mathbf{V}_v \tag{16}$$

where $diag(\mathbf{C})$ = the diagonal elements of the covariance matrix $\mathbf{C}$. The columns of the eigenvector matrix $\mathbf{V}_v$ and eigenvalue matrix $\mathbf{V}_s$ are sorted in order of decreasing eigenvalue. The eigenvalues represent the distribution of the source data's contribution, and the contribution rate for each eigenvector is defined as:

$$\lambda_i = \frac{\mathbf{V}_{si}}{\sum_{i=1}^{k} \mathbf{V}_{si}} \tag{17}$$

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

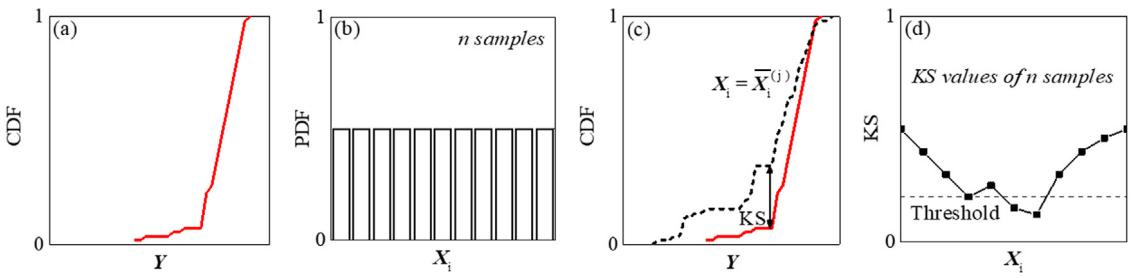

Fig. 3. Flowchart of sensitivity analysis based on CDF method.

where $\lambda_1 + \lambda_2 + \cdots + \lambda_k = 1$. The objective of PCA method is to find p ($p<k$) eigenvalues with the cumulative contribution rate exceeding the threshold (85% generally). The corresponding eigenvectors are thus extracted, forming a new eigenvector matrix $\mathbf{V}_p$ ($k\times p$). Ultimately, the projection of all input data is obtained:

$$\mathbf{P} = \mathbf{M} \bullet \mathbf{V}_p \tag{18}$$

where $\mathbf{P}$ = the principal component matrix ($N\times p$).

2.4. Proposed feature selection method

Fig. 4 presents the flowchart of proposed FS method, which consists of two phases. The flowchart starts from establishing a model to predict tunnelling-induced settlement. As mentioned before, RF method is utilized in this research. RF model is iteratively trained until it reaches the maximum iterations (termination criteria). Herein, the optimum RF model is determined by the grid search method [34] with the number of tree increasing at one interval. At the second phase, GSA method is utilized to search the optimum feature subset.

This research develops a novel FS method based on GSA method. The ranking of input variables can be determined by each GSA method. The sum of ranking from different GSA methods is ultimately used to evaluate the significance of input variables. After ranking input features, similar to the filter-wrapper method, RF models with different number of features are developed to predict tunnelling-induced settlement (GSA-RF). Herein, the first case is that the number of features increases from one to twelve in an increasing order, and the second case is that the number of features increases in a descending order. Based on this, the optimum feature subset can be determined and validated. Meanwhile, the feature extraction method PCA is also utilized to explore the optimum variable space, considering integration of PCA and ML methods has already been applied to predict tunnelling-induced settlement [18]. RF models with different number of principle components are established to predict tunnelling-induced settlement (PCA-RF). Ultimately, the performance GSA-RF and PCA-RF is compared.

3. Model construction

3.1. Data source

The proposed FS method is applied to a practical tunnel engineering. To mitigate increasing traffic issues arisen from urbanization in China, numerous cities have devoted to develop metro system [35,36]. Tunnelling-induced settlement occasionally cause the damage to surrounding structures and infrastructures, especially in densely populated city area [37]. Meanwhile, tunnelling-induced settlement is related to numerous extrinsic and intrinsic factors. Therefore, current research works focus on developing ML-based models for predicting tunnelling-induced settlement [38,39]. The datasets used in this study is from Changsha Metro Line 4 project, as attached in Appendix.

The factors associated with the tunnelling-induced settlement are generally classified into four categories: tunnel geometry, geological conditions and shield operation parameters and anomalous condition [23]. To this end, twelve input variables are selected including five shield operational parameters (thrust *Th*, torque *To*, grout filling *Gf*, penetration rate *Pr*, face pressure *Fp*), five geological parameters (modified blow counts of standard penetration test *MSPT* and dynamic penetration test *MDPT* of soil layers, modified uniaxial compressive strength of weathered rocks *MUCS*, groundwater table *W* and the geological conditions *Gc* at the cutterhead face), one geometry parameter (cover depth of the tunnel *C*) and one anomalous condition stoppage *St*. Due to the consistent tunnel specification in this project, the cover depth of tunnel is the unique geometry factor. The geological conditions at the cutterhead face are categorized into four types: soil, gravel, rock and mixed-face ground, and they are marked as 1, 2, 3 and 4, respectively. In terms of shield stoppage, 1 denotes the stoppage and 0 denotes the continuous advancement. Output variable is the ground surface settlement. Finally, a total number of 294 sets of field records with twelve input variables and one output variable is established. Table 1 presents the range of these input and output variables. The datasets and the reason for selecting these parameters can refer to Zhang, et al. [23]. Herein, 80% of data randomly sampled out of database is selected as the training set, and the retained 20% is used to test model.

The database is mapped to the interval (−1, 1) using a data normalization algorithm before training the ML model for saving the computational cost and eliminating the scale of different parameters. For a parameter x, the normalized value x is computed by

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} (\bar{x}_{\max} - \bar{x}_{\min}) + \bar{x}_{\min} \tag{19}$$

where $x_{\max}$ and $x_{\min}$ = the maximum and minimum value of the variable x; $\bar{x}_{\max}$ and $\bar{x}_{\min}$ = the maximum and minimum values of the variable x after normalization. The final outputs need to be transformed into the original vector space.

3.2. Performance indicators

Performance assessment is conducted to calculate the prediction error, which is an indispensable part of establishing a ML model. Two performance indicators: mean absolute error (MAE), root mean square error (RMSE) are selected to evaluate the prediction performance of model in this research. The definition of MAE and RMSE is as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{1}^{n} |r_i - p_i| \tag{20}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{1}^{n} (r_i - p_i)^2} \tag{21}$$

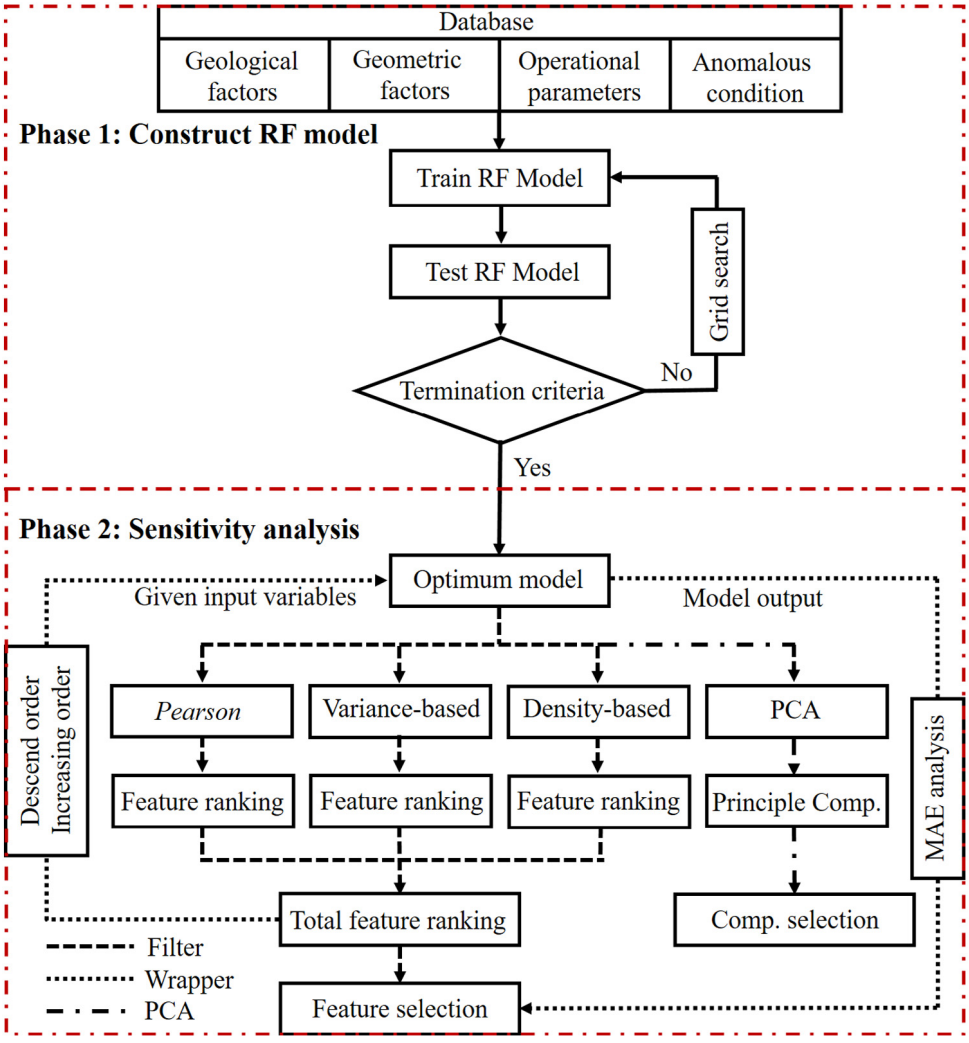

Fig. 4. The flowchart of proposed FS method.

Table 1
Range of parameters in database.

Variable (unit)	Parameter type	Data (236)			Unit
		Min.	Max.	Ave.	
Torque (*To*)	Input	0.29	4.70	2.40	MN m
Penetration rate (*Pr*)	Input	2.41	45.8	23.01	mm/rev
Thrust (*Th*)	Input	7.00	24.20	13.11	MN
Face pressure (*Fp*)	Input	0.00	2.50	1.09	bar
Grout filling (*Gf*)	Input	4.00	13.10	6.02	m^3
Overburden (*C*)	Input	9.10	26.35	17.45	m
Tunnel depth below the water table (*W*)	Input	0.00	20.30	11.80	m
Modified standard penetration test (*MSPT*)	Input	0.00	38.72	7.01	/
Modified dynamic penetration test (*MDPT*)	Input	0.00	12.44	0.38	/
Modified uniaxial compressive strength (*MUCS*)	Input	0.00	36.30	6.89	MPa
Ground condition (*Gc*)	Input	1	4	0.45	/
Stoppage (*St*)	Input	0	1	0.29	/
Maximum settlement (*S*)	Output	−46.48	3.05	−4.40	mm

where r = actual settlement; p = predicted settlement; n = total number of events considered. MAE reflects the average magnitude of error between predicted and measured value, while RMSE describes the standard deviation of differences between them.

4. Results

4.1. Random forest based settlement prediction model

In the grid search method, the maximum number of trees is set as 200 considering the tradeoff between model prediction performance and computational cost. MAE and RMSE values of

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

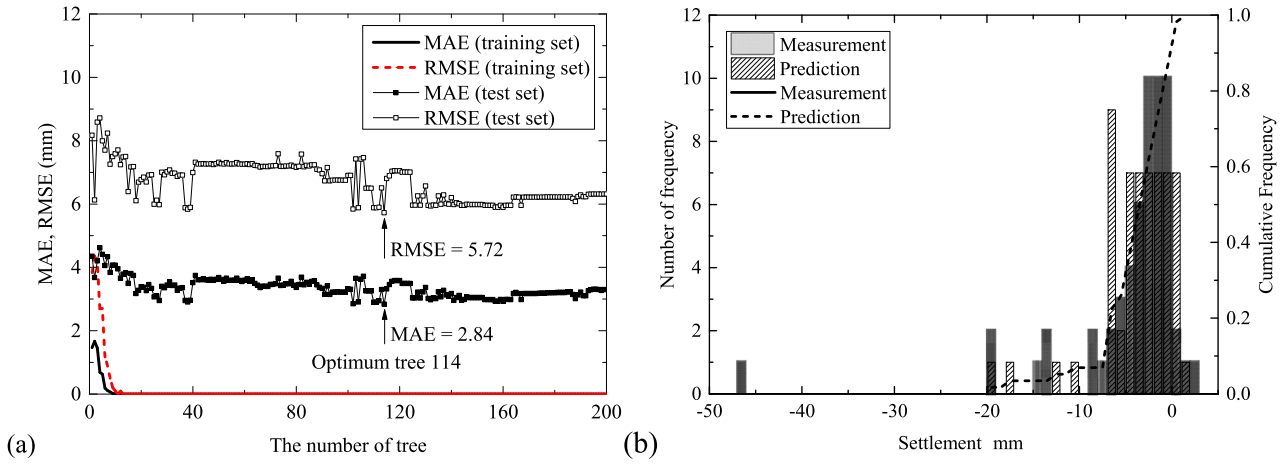

Fig. 5. Predicted settlement by RF-based model: (a) indicator values for training and test sets; (b) distribution of predicted and measured settlement.

the training and test sets are plotted against the number of trees in Fig. 5(a). For the training set, MAE and RMSE values decrease dramatically with the increase in the number of trees, and the errors are equal to zero when the number of trees exceeds ten. For the test set, note that for the low number of trees, MAE and RMSE values decline with the increasing number of trees, which refers to the mitigation of underfitting. However, beyond 150 trees, MAE and RMSE values saturate. Any further increase in the number of trees cannot enhance model prediction performance. The optimum number of trees in RF model is here identical to be 114. The corresponding MAE and RMSE values of the test set are 2.84 and 5.72, respectively.

Fig. 5(b) illustrates the distribution of predicted and measured settlement in the test set. It can be observed that the measured settlement primarily ranges from −19.5 mm to 2.5 mm while the range of predicted settlement is from −19.5 mm to 1.5 mm. The cumulative frequency of predicted settlements shows great agreement with measured results, especially for the settlement less than 7.5 mm. It indicates the accuracy of RF model is acceptable. The histogram of the settlement explains that the large error is primarily derived from the prediction of large settlements, since the data source of the training set is scarce at that range. The cumulative frequency of the predicted settlement is regarded as the baseline for the next GSA, meanwhile RF model with 114 trees is used for the further analysis.

4.2. Sensitivity analysis

The scatter plot of twelve input variables with respect to the settlement is presented in Fig. 6 and *Person* correlation coefficient R calculated by Eq. (22) is also presented in the graph. It can be observed that all input parameters have a relatively poor correlation (R<0.5) with the settlement, and the data is widely distributed. Thrust, face pressure and grout filling show high correlation with the settlement among five operational parameters. Thrust and face pressure have a positive correlation with the settlement, which means that the increase in two parameters can effectively reduce settlement. In contrast, there is a negative correlation between the grout filling and the settlement. Fig. 6 indicates the settlement is positively related to the *MUCS* value while it is negatively related to the *SPT* value. It means that tunnelling-induced settlement reduces as the rocks appeared around the tunnel. It is noteworthy that the ground conditions at the cutterhead face and shield stoppage are the most influential factors to the settlement with the correlation coefficient R reaching 0.4. Overall, settlement weakly correlates to the twelve input variables, and a simple linear relationship between variables does not exist.

$$R = \frac{\sum_{i=1}^{n} (r_i - \bar{r})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n} (r_i - \bar{r})^2 \sum_{i=1}^{n} (p_i - \bar{p})^2}} \tag{22}$$

where $\bar{r}$ = average value of measured settlements; $\bar{p}$ = average value of predicted settlements; the meaning of r_i and p_i is similar to that in Eq. (20).

The first-order and total-order index of twelve input features based on the *Sobol'* method are presented in Fig. 7. In regard to the first-order sensitivity index, four most relevant input features in order are *W*, *St*, *Gc*, *To* and meanwhile four most non-significant features in order are *MDPT*, *MSPT*, *Pr* and *Gf*. This result shows great agreement with the ranking of input features based on the *Pearson* method. It can be observed that total-order sensitivity index values of shield operational parameters *To*, *Pr*, *Fp* and *Gf* increase dramatically compared with the first-order sensitivity index. Meanwhile the total-order sensitivity index values of geometric, geological and anomalous condition factors virtually maintain unchangeable. These phenomena are consistent with the actual conditions because total-order sensitivity index takes the interaction among input variables into consideration. Tunnel geometric characteristics, geological condition and anomalous condition are the objective factors to the tunnelling-induced settlement. Values of shield operational parameters are affected by these objective factors. This is the reason why the interaction among shield operational parameters is more obvious than that in other parameters.

Twelve conditional CDFs calculated by PAWN approach are shown in Fig. 8. The red line is the baseline of the unconditional CDF of predicted settlement as mentioned in Section 2.2.2. The grey and black lines represent the conditional CDF of predicted settlement. Regarding the values of input parameters for obtaining the conditional CDF, note that the value parameter being studied increases from −1 to 1 at 0.2 interval because all variables are mapped to the range of −1 to 1, meanwhile the values of remaining parameters are consistent with actual values. The black lines represent that the fixed value of parameter being studied is non-negative and the grey lines represent negative values. The Kolomogorov–Smirnov statistic of each feature is presented under the corresponding CDF.

Visual analysis of the CDFs illustrates that *Th* and *Gf* are the most influential factors among five shield operational parameters and *MUCS* is the most significant factor among the geological factors, because conditional CDFs of such parameters exist large

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

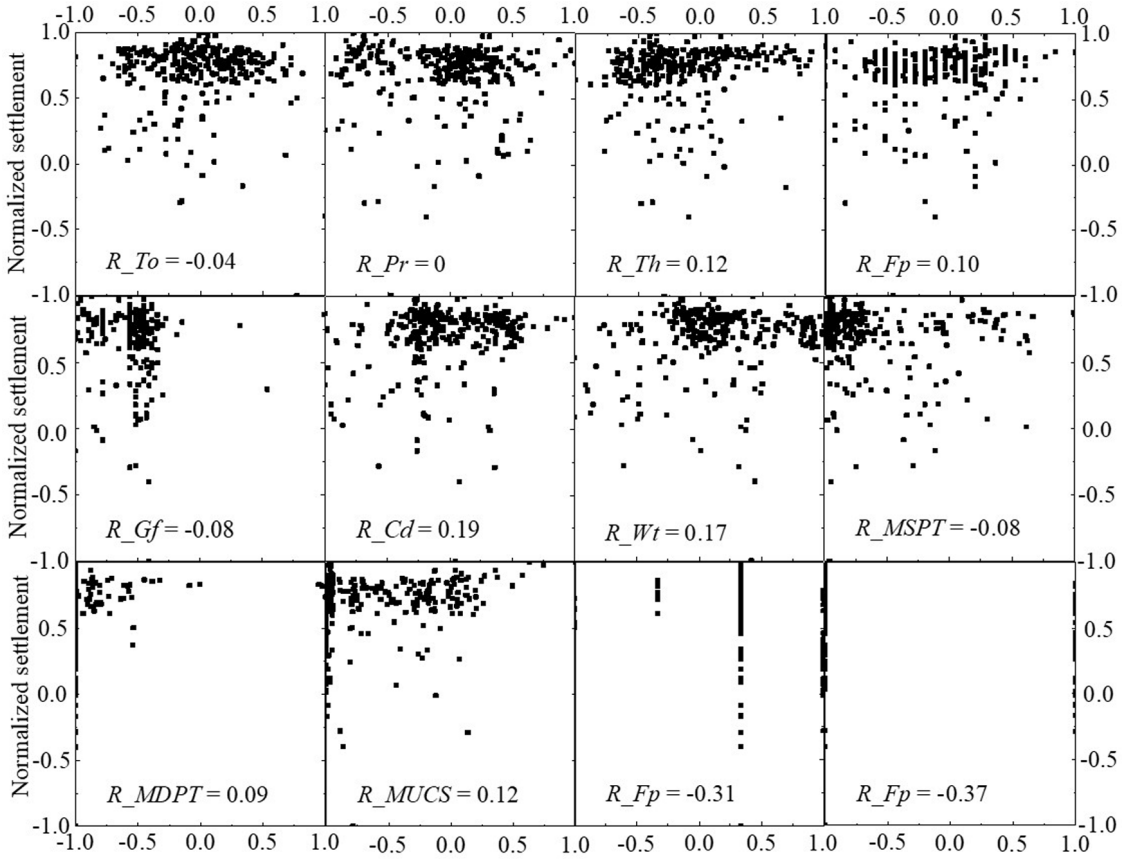

Fig. 6. Scatter plot of input variables with respect to settlement together with *Pearson* correlation coefficient.

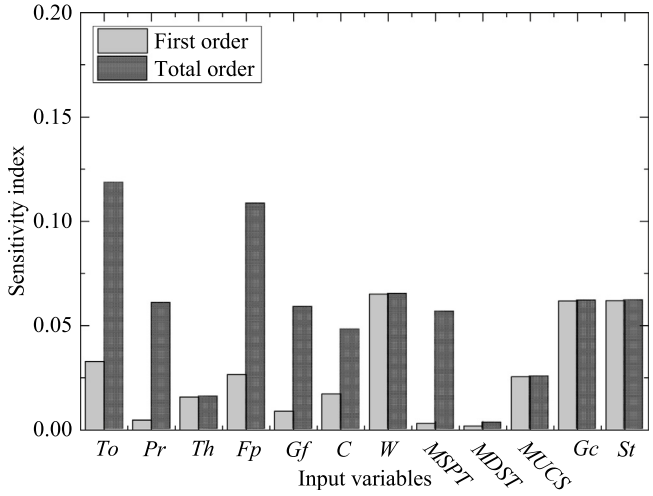

Fig. 7. Results of first-order and total-order index for input variables.

difference with the unconditional CDF. The anomalous condition shield stoppage also has great impact on the settlement. The geometric factor *C* seems to be insignificant, because CDFs are totally overlapped. Similar condition can be seen in the *Gc* factor. In quantitative terms, Kolomogorov–Smirnov value 0.2 is set as the threshold value. Those features are insignificant if the Kolomogorov–Smirnov values are always below the threshold. The final ranking of features are determined by the average Kolomogorov–Smirnov value.

Table 2 presents the ranking of input variables based on the results of three GSA methods. The ranking of input variables

Table 2
Scores of three sensitivity analysis.

Variable	Index			Ranking			Overall ranking
	Pearson	*Sobol'*	PAWN	*Pearson*	*Sobol'*	PAWN	
To	−0.04	0.03	0.24	9	3	4	16(10)
Pr	0	0	0.15	10	6	6	22(12)
Th	0.12	0.02	0.25	5	4	4	13(7)
Fp	0.10	0.03	0.13	6	3	7	16(9)
Gf	−0.08	0.01	0.27	8	5	3	16(8)
C	0.19	0.02	0.15	3	4	6	13(6)
W	0.17	0.07	0.15	4	1	6	11(3)
MSPT	−0.08	0	0.18	8	6	5	19(11)
MDPT	0.09	0	0.34	7	6	1	14(5)
MUCS	0.12	0.03	0.28	5	3	2	10(1)
Gc	−0.31	0.06	0.10	2	2	8	12(4)
St	−0.37	0.06	0.13	1	2	7	10(2)

are virtually identical for *Pearson* and *Sobol'* methods while the results of PAWN method is obviously different. The ultimate ranking of input variables are determined by the overall ranking, which is defined as the sum of ranking calculated by three different GSA methods. Note that these variables with same scores rank randomly without discernable impact on the model performance. The rank of input variables in a decreasing trend are *MUCS*, *St*, *W*, *Gc*, *MDPT*, *C*, *Th*, *Gf*, *Fp*, *To*, *SPT*, *Pr* in this study, which provides a basis for establishing RF-based model.

4.3. GSA-RF model

In order to verify the rationality of ultimate features ranking based on GSA, RF models with different combinations of features are developed. In Fig. 9, the solid line represents the performance of model which is trained by the features (from one to twelve)

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

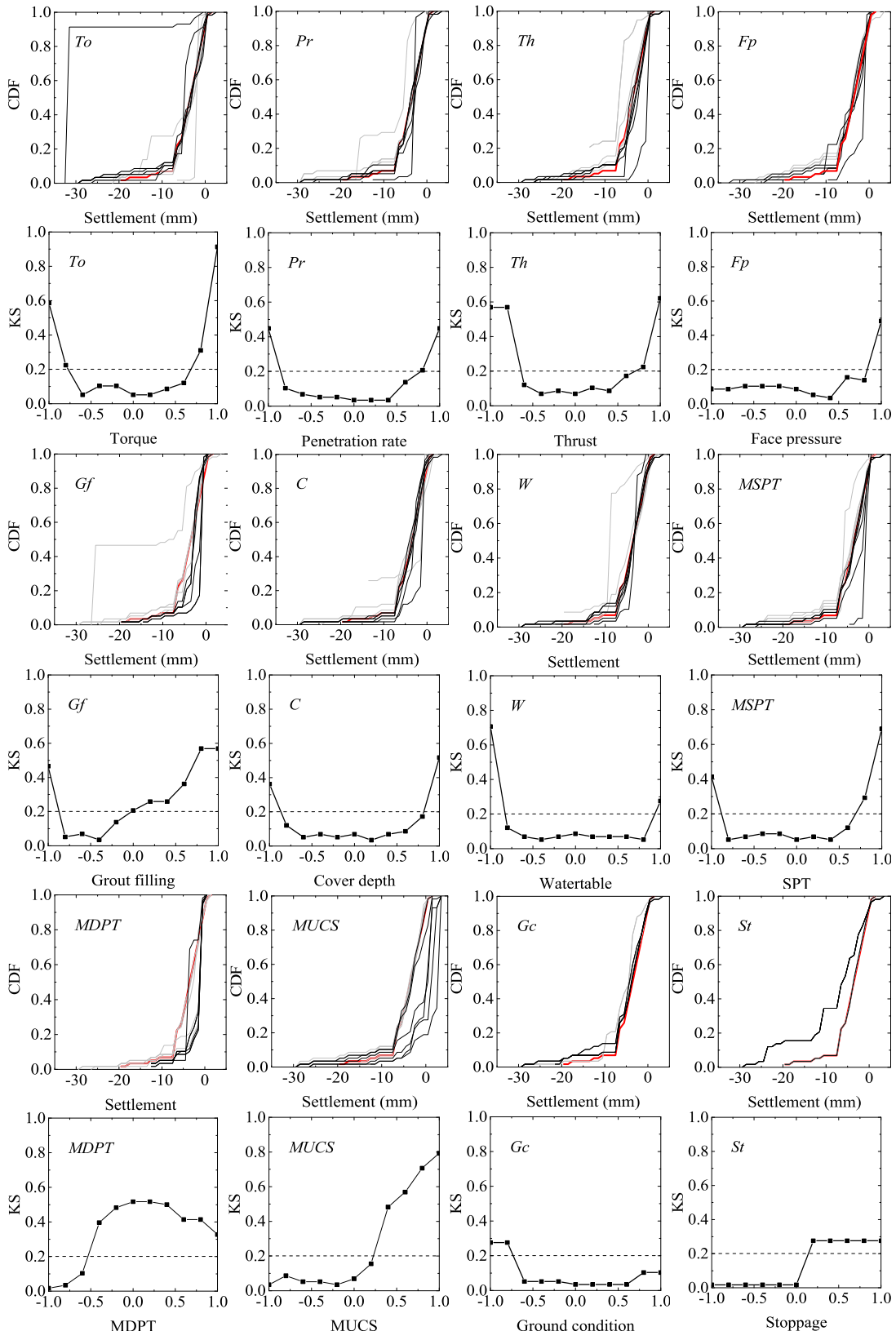

Fig. 8. Unconditional model output distribution (red line), conditional model output distribution (grey and black line) and *KS* values at input variables.

selected from the front of the list (descending order). The dash line represents the performance of model which is trained by the features selected from the rear of the list (increasing order). Accordingly, in each order, twelve RF-based settlement prediction models are established, that is, the number of input variables increases from one to twelve with an increasing or decreasing order. Each model is trained 200 times in order to identify the optimum number of tree in RF model at each case. In the end,

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

Table 3
Optimum number of trees in models with various features.

Feature	Descending order	Tree	Increasing order	Tree
1	*MUCS*	2	Pr	1
2	*MUCS, St*	25	*Pr, MSPT*	6
3	*MUCS, St, W*	26	*Pr, MSPT, Gf*	8
4	*MUCS, St, W,Gc*	3	*Pr, MSPT, Gf, Fp*	78
5	*MUCS, St, W, Gc, MDPT*	82	*Pr, MSPT, Gf, Fp, To*	190
6	*MUCS, St, W, Gc, MDPT, C*	28	*Pr, MSPT, Gf, Fp, To, Th*	138
7	*MUCS, St, W, Gc, MDPT, C, Th*	4	*Pr, MSPT, Gf, Fp, To, Th, C*	171
8	*MUCS, St, W, Gc, MDPT, C, Th, To*	8	*Pr, MSPT, Gf, Fp, To, Th, C, MDPT*	55
9	*MUCS, St, W, Gc, MDPT, C, Th, To, Fp*	31	*Pr, MSPT, Gf, Fp, To, Th, C, MDPT, Gc*	123
10	*MUCS, St, W, Gc, MDPT, C, Th, To, Fp, Gf*	13	*Pr, MSPT, Gf, Fp, To, Th, C, MDPT, Gc, W*	68
11	***MUCS, St, W, Gc, MDPT, C, Th, To, Fp, Gf, MSPT***	**55**	*Pr, MSPT, Gf, Fp, To, Th, C, MDPT, Gc, W, St*	61
12	*MUCS, St, W, Gc, MDPT, C, Th, To, Fp, Gf, MSPT, Pr*	114	*Pr, MSPT, Gf, Fp, To, Th, C, MDPT, Gc, W, St, MUCS*	114

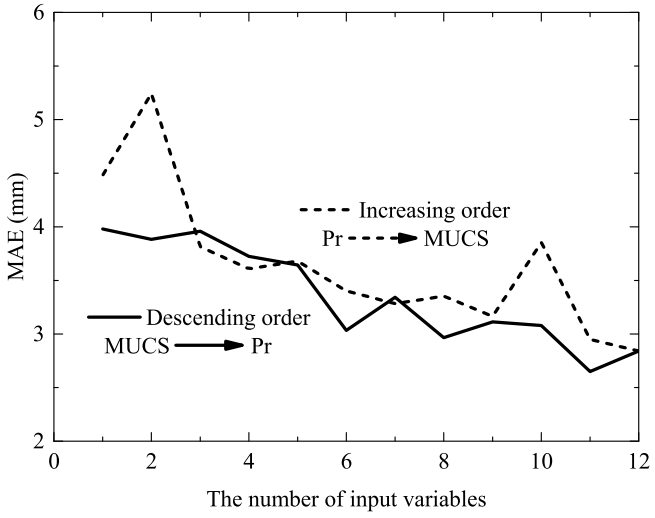

Fig. 9. Average MAE for the test set generated by optimum RF models with 24 combinations of input variables.

a total number of $2\times12\times200$ computational times are implemented. It can be seen from Fig. 9 that RF model with features in a descending order virtually outperforms RF model with features in an increasing order at each case, especially when the number of input variables is less than two. It is reasonable because in the context of same number of input features, RF model with more relevant features broadly presents better prediction accuracy. MAE value of RF model is identical as the number of features reaches the twelve at two orders. It is also noteworthy that the variation of MAE values is not smooth, because the variables with same score rank randomly as mentioned above. The general trend is that MAE value progressively decreases with the increase in the number of features.

From the perspective of Fig. 9, the optimum features subset is also identified. Table 3 presents the optimum number of trees in all RF models. Herein, RF model with 55 trees and eleven features excluding *Pr* shows the best performance, and the corresponding MAE value is 2.65 mm. All of these results indicate the ranking of features based on the total GSA scores is reasonable. GSA-based FS method is beneficial to determine the optimum combination of input variables in ML-based models.

4.4. PCA-RF model

PCA method is employed to reduce complexity of parameter space on the basis of eliminating the interaction between input variables. The twelve input variables used in the RF-based settlement prediction model are processed by the PCA method.

Table 4
Contribution rate and the cumulative contribution rate of twelve principal components.

Principal component	Eigenvalues	Contribution rate	Cumulative contribution rate
Comp. 1	3.11	25.92%	25.92%
Comp. 2	1.94	16.20%	42.12%
Comp. 3	1.38	11.52%	53.64%
Comp. 4	1.14	9.51%	63.15%
Comp. 5	0.96	8.00%	71.15%
Comp. 6	0.85	7.10%	78.25%
Comp. 7	0.80	6.68%	84.93%
Comp. 8	0.61	5.13%	90.06%
Comp. 9	0.43	3.62%	93.68%
Comp. 10	**0.38**	**3.17%**	**96.85%**
Comp. 11	0.26	2.20%	99.05%
Comp. 12	0.11	0.95%	100.00%

The eigenvalues, contribution rate and cumulative contribution rate of twelve principal components are listed in Table 4. It can be observed that the first seven principal components with a cumulative contribution rate of about 85% and the cumulative contribution rate of first ten principal components reaches 95%.

In order to determine the performance of principal components, RF models with different combinations of principal components are established. The principal components with high contribution rate are selected from the front of the list, thereby a total number of 12 RF-based models are developed. The corresponding MAE values of twelve optimum RF model are presented in Fig. 10. The optimum number of trees in these models are 8, 4, 116, 26, 15, 2, 20, 68, 10, 2, 154 and 13, respectively. The error of RF model decrease with the number of principal component increasing from the one to seven. The corresponding MAE value deceases from 5.37 mm to 3.45 mm. Then MAE value rises slightly as the number of principal components increase from eight to nine. RF model with ten principal components presents the best performance with MAE = 3.30 mm, in which the corresponding cumulative contribution rate of ten input principal components reaches 95%.

5. Discussion

In order to compare the performance of GSA-RF and PCA-RF settlement prediction models, Fig. 11 presents the scatter plot of predicted settlements using three optimum RF models, that is, RF model trained by the original datasets with twelve input variables (see Fig. 11(a)), GSA-RF model determined in Section 4.2 (see Fig. 11(b)) and PCA-RF model determined in Section 4.3 (see Fig. 11(c)), respectively. In regard to the training set, the predicted settlements using RF and GSA-RF show perfect agreement with the measured settlements. The absence of agreement between the measured and predicted values is observed in PCA-RF model. In terms of the test set, it can be observed that the predictions

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

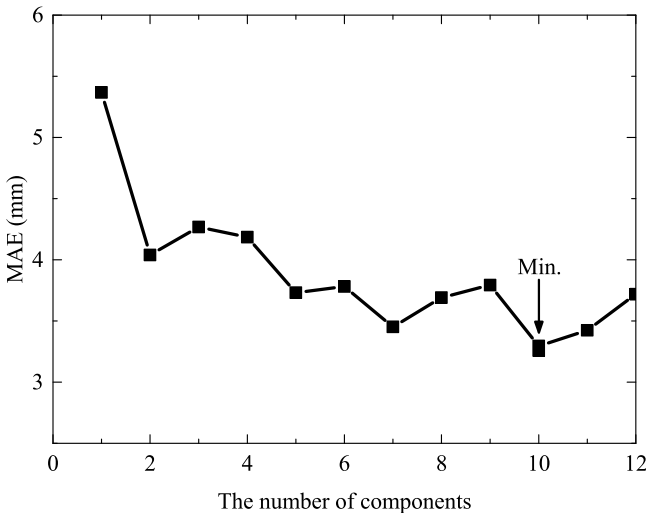

Fig. 10. MAE values for the test set generated by optimum RF models with 12 combinations of principal components.

yield lowest error with MAE = 2.65 and RMSE = 5.70 mm in GSA-RF model. Meanwhile the majority of points closely lie in a straight line with slope of 1 (P = M line). The largest error is observed in the PCA-RF model with the MAE = 3.30 and RMSE = 5.35 mm, respectively. In contrast to RF model developed on the original database, GSA-RF model broadly improves the prediction accuracy while the PCA-RF model increases the prediction error. Although both of FS and PCA can reduce the dimensionality of input variables, thereby reduces computational cost, GSA-RF model obviously outperform the PCA-RF model, which is attributed to difference of the principle of two dimensionality reduction methods. GSA based FS method directly selects feature subsets that have great impacts on the output variables from the original features space, therefore, such FS method is beneficial to improve model performance and understand the interaction mechanism between input and output variables. PCA based FS method creates new features by the combination of original features. Such FS method sacrifices the contribution of some input variables to reduce dimensionality, undoubtedly resulting in the decrease of model prediction performance. Meanwhile these new combined features have no physical meaning, which means that PCA is not suitable to develop model with the emphasis on the readability, interpretability, and transparency.

Fig. 12 shows the evolution of ground settlements predicted using three models, compared with measured settlements. In contrast to PCA-RF method, predicted settlements using RF model developed with original dataset and dataset processed by GSA

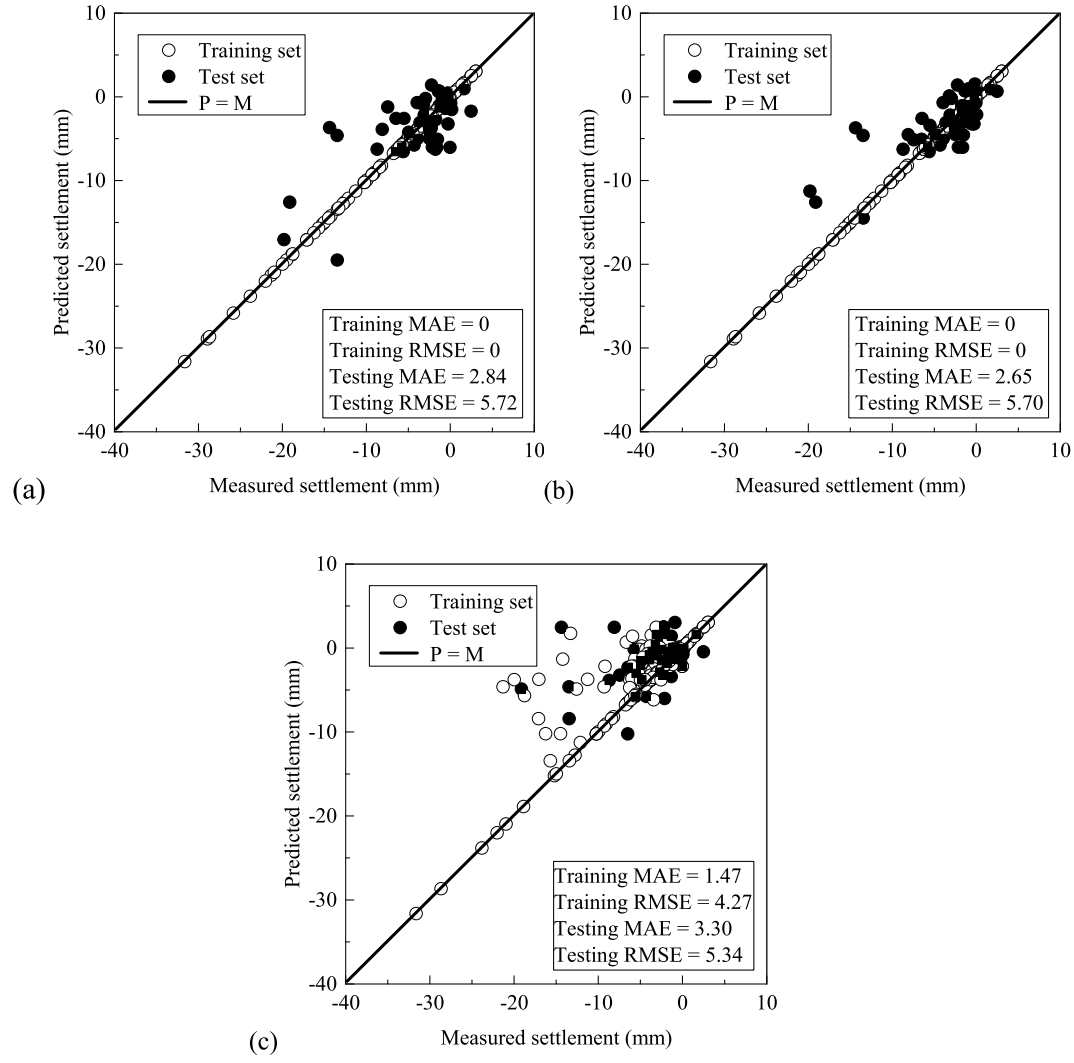

Fig. 11. Predicted settlement using RF model developed by: (a) the original dataset; (b) eleven features (GSA-RF); (c) ten principal components (PCA-RF).

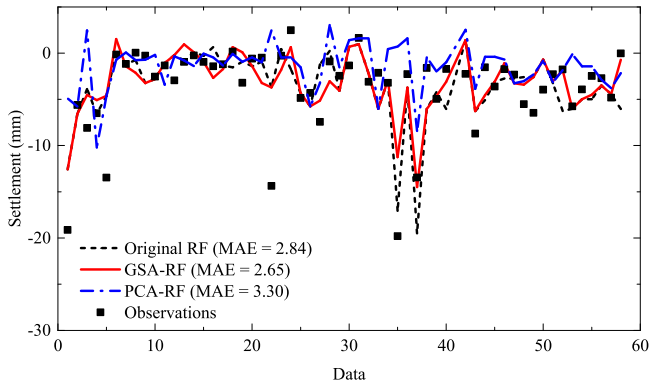

Fig. 12. Evolution of predicted settlement using three optimum RF-based models.

method show great agreement with the evolution of actual observations. Compared with RF model developed on the basis of original dataset, GSA-RF model enhance the model prediction performance and obviously decrease prediction error at many monitoring points such as the first and the last monitoring points. Because GSA method ranks the significance of input variables to the settlement, the insignificant parameter *Pr* has been identified by the filter-wrapper method, these factors remarkably improve the prediction capability of RF model. However, the predicted values using PCA-RF method vary remarkably, resulting in losing fidelity at some points. These factors indicate the proposed GSA-based FS method is able to effectively determine the optimum features subsets, thereby is beneficial to develop ML-based models.

6. Conclusions

In this research, a novel feature selection (FS) method on the basis of global sensitivity analysis (GSA) is proposed. GSA methods *Pearson*, *Sobol*' and PAWN index are used to investigate the significance of input variables to the output variables. The new feature ranking criteria is developed upon the sum of scores of input variables calculated by three GSA methods for reducing the uncertainty of GSA. This novel FS method is integrated with machine learning algorithm random forest (RF) for determining optimum feature subsets, thereby improves ML-based models prediction performance (GSA-RF). GSA-RF is applied to a practical tunnel engineering to predict tunnelling-induced settlement for validating the reliability of proposed method. GSA-based FS method is able to effectively detect the insignificant features to settlement and obviously improve model prediction performance by excluding these insignificant features.

The integration of dimensionality reduction method PCA and RF (PCA-RF) was also developed to compare the performance of GSA-RF in predicting tunnelling-induced settlement. The proposed GSA-based FS method effectively identifies the optimum parameters that have great impact on the output variables, which is beneficial to improve model performance and understand the interaction mechanism between input and output variables. PCA based FS method created new features by the combination of original features. Such FS method sacrifices the contribution of some input variables to reduce dimensionality, resulting in the decrease of model prediction performance. Meanwhile PCA based FS method is not suitable to develop model with the emphasis on the readability, interpretability, and transparency, because these new combined features have no physical meaning. Although both of FS and PCA can reduce the dimensionality of input variables, proposed GSA-based prediction model obviously outperforms the PCA-RF prediction model.

Proposed GSA-RF model accurately captured tunnelling-induced settlement and determine the significant of input variables to the settlement, which provides construction guideline and predict risks in engineering practice. Therefore, it is recommended to similar engineering.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.105859.

Appendix

The datasets used in this study are available for download at: https://www.researchgate.net/publication/336208927_Database_for_maximum_settlement_collected_from_Changsha_Metro_Line_4_Liugoulong_to_Fubuhe_station.

References

[1] L.M. Zhang, X.G. Wu, W.Y. Ji, S.M. AbouRizk, Intelligent approach to estimation of tunnel-induced ground settlement using wavelet packet and support vector machines, J. Comput. Civ. Eng. 31 (2017) 04016053.
[2] R.P. Chen, P. Zhang, H.N. Wu, Z.T. Wang, Z.Q. Zhong, Prediction of shield tunneling-induced ground settlement using machine learning techniques, Front. Struct. Civ. Eng. (2019) http://dx.doi.org/10.1007/s11709-019-0561-3, in press.
[3] Z. Tao, L. Huiling, W. Wenwen, Y. Xia, GA-SVM based feature selection and parameter optimization in hospitalization expense modeling, Appl. Soft Comput. 75 (2019) 323–332.
[4] W.F. Gao, L. Hu, P. Zhang, J.L. He, Feature selection considering the composition of feature relevancy, Pattern Recognit. Lett. 112 (2018) 70–74.
[5] E.S. Hosseini, M.H. Moattar, Evolutionary feature subsets selection based on interaction information for high dimensional imbalanced data classification, Appl. Soft Comput. 82 (2019).
[6] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.
[7] X. Zeng, Z. Zhen, J. He, L. Han, A feature selection approach based on sensitivity of RBFNNs, Neurocomputing 275 (2018) 2200–2208.
[8] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[9] A. Saltelli, I.M. Sobol, About the use of rank transformation in sensitivity analysis of model output, Reliab. Eng. Syst. Saf. 50 (1995) 225–239.
[10] T. Hong, S.T. Purucker, Spatiotemporal sensitivity analysis of vertical transport of pesticides in soil, Environ. Model. Softw. 105 (2018) 24–38.
[11] H.C. Frey, S.R. Patil, Identification and review of sensitivity analysis methods, Risk Anal. 22 (2010) 553–578.
[12] N.A. Stiber, Pantazidou M., M.J. Small, Expert system methodology for evaluating reductive dechlorination at TCE sites, Environ. Sci. Technol. 33 (1999) 3012–3020.
[13] C.Y. Zhao, A.A. Lavasan, R. Hölter, T. Schanz, Mechanized tunneling induced building settlements and design of optimal monitoring strategies based on sensitivity field, Comput. Geotech. 97 (2018) 246–260.
[14] K.M. Hamdia, H. Ghasemi, X.Y. Zhuang, N. Alajlan, T. Rabczuk, Sensitivity and uncertainty analysis for flexoelectric nanostructures, Comput. Methods Appl. Mech. Eng. 337 (2018) 95–109.
[15] J.J. Pérez-Barea, F. Fernández-Navarro, M.J. Montero-Simó, R. Araque-Padilla, A socially responsible consumption index based on non-linear dimensionality reduction and global sensitivity analysis, Appl. Soft Comput. 69 (2018) 599–609.
[16] L.M. Zhang, X.G. Wu, H.P. Zhu, S.M. AbouRizk, Performing global uncertainty and sensitivity analysis from given data in tunnel construction, J. Comput. Civ. Eng. 31 (2017) 04017065.
[17] W. Liu, X. Wu, L. Zhang, Y. Wang, J. Teng, Sensitivity analysis of structural health risk in operational tunnels, Automat. Constr. 94 (2018) 135–153.
[18] 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, Tunn. Undergr. Space Technol. 68 (2017) 142–152.
[19] I.M. Sobol', Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Math. Comput. Simulation 55 (2001) 271–280.

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.

[20] F. Pianosi, T. Wagener, A simple and efficient method for global sensitivity analysis based on cumulative distribution functions, Environ. Model. Softw. 67 (2015) 1–11.
[21] J. Zhou, X. Shi, K. Du, X.Y. Qiu, X.B. Li, H.S. Mitri, Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel, Int. J. Geomech. 17 (2016) 04016129.
[22] Y. Zhou, S.Q. Li, C. Zhou, H.B. Luo, Intelligent approach based on random forest for safety risk prediction of deep foundation pit in subway stations, J. Comput. Civ. Eng. 33 (2019) 05018004.
[23] P. Zhang, R.-P. Chen, H.-N. Wu, Real-time analysis and regulation of EPB shield steering using random forest, Automat. Constr. 106 (2019) 102860.
[24] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[25] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. 20 (1998) 832–844.
[26] A. Liaw, M. Wiener, Classification and regression by random forest, R News 23 (2002) 18–21.
[27] I.M. Sobol', Sensitivity estimates for nonlinear mathematical models, Math. Model. Comput. Exp. 1 (1993) 407–414.
[28] I.M. Sobol', Global sensitivity analysis indices for the investigation of nonlinear mathematical models, Mat. Mod. 19 (2007) 23–24.
[29] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, S. Tarantola, Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index, Comput. Phys. Comm. 181 (2010) 259–270.
[30] M.J.W. Jansen, Analysis of variance designs for model output, Comput. Phys. Comm. 117 (1999) 35–43.
[31] H. Liu, W. Chen, A. Sudjianto, Relative entropy based method for probabilistic sensitivity analysis in engineering design, J. Mech. Des. 128 (2006).
[32] A.L. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione, G. Ist. Degli Attuari 4 (1933) 83–91.
[33] N. Smirnoff, On the estimation of the discrepancy between empirical curves of distribution for two independent samples, Bull. Math. Univ. Mosc. 2 (1939).
[34] I.-C. Yeh, Application of neural networks to automatic soil pressure balance control for shield tunneling, Automat. Constr. 5 (1997) 421–426.
[35] M. Jiang, Z.Y. Yin, Influence of soil conditioning on ground deformation during longitudinal tunneling, C. R. Mec. 342 (2014) 189–197.
[36] P. Zhang, Y. Liu, X. Kang, K. Zhong, R.P. Chen, Application of horizontal MJS piles in tunneling beneath existing twin tunnels, in: The 2nd International Symposium on Asia Urban GeoEngineering, Changsha, 2018, pp. 323–331.
[37] R.-P. Chen, X.-T. Lin, X. Kang, Z.-Q. Zhong, Y. Liu, P. Zhang, H.-N. Wu, Deformation and stress characteristics of existing twin tunnels induced by close-distance EPBS under-crossing, Tunn. Undergr. Space Technol. 82 (2018) 468–481.
[38] R.P. Chen, P. Zhang, X. Kang, Z.Q. Zhong, Y. Liu, H.N. Wu, Prediction of maximum surface settlement caused by EPB shield tunneling with ANN methods, Soils Found. 59 (2019) 284–295.
[39] M. Hasanipanah, M. Monjezi, A. Shahnazar, D. Jahed Armaghani, A. Farazmand, Feasibility of indirect determination of blast induced ground vibration based on support vector machine, Measurement 75 (2015) 289–297.

Please cite this article as: P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal (2019) 105859, https://doi.org/10.1016/j.asoc.2019.105859.