Transient Stability Assessment of Power System based on Improved Maximum Correlation Minimum Redundancy Random Forest Algorithm

Xudong Zhang, Xihuai Wang

School of Shanghai, Maritime University, Shanghai 201306, China

Abstract

In order to improve the accuracy of power system transient stability assessment, this paper proposes a feature selection method based on the combination of maximum correlation minimum redundancy (mRMR) and random forest out-of-bag error (RFOE). method (mRMR-RFOE), and use the random forest algorithm to build a power system steady-state evaluation model, and use mRMR and RFOE to rank the features respectively. The final ranking result is obtained by combining the two kinds of feature sorting, and the optimal feature subset required by the random forest algorithm is selected according to the optimal number of features obtained from the experiment. Finally, the random forest classification model is trained using the optimal feature subset. The experimental results show that compared with other classification methods, this method can improve the accuracy of power system steady-state assessment, which is of practical significance.

Keywords

Power System; Transient Stability Assessment; Maximum Correlation Minimum Redundancy; Feature Selection; Random Forest Out-of-bag Error.

1. Introduction

Power system transient stability assessment (TSA) [1] is an important research content for the safe operation of power systems.With the continuous improvement of power system load, interconnection level and the access of new energy sources, the scale of the system is becoming more and more complex, and its operation is getting closer to the stable limit. In recent years, the frequent occurrence of power system collapse accidents has had a huge impact on the economy. When the power system encounters faults such as short-circuit and disconnection, it is of great significance to the safe and stable operation of the power system if it can be predicted in advance whether the disturbance will cause the instability of the power system and take corresponding measures in time. Therefore, how to conduct steady-state assessment of the power system becomes particularly important.

At present, the transient stability assessment of power system includes time domain simulation method, direct method and artificial intelligence method. The time domain simulation method mainly models the components of the power system and constructs nonlinear differential equations. The model can accurately reflect the electromechanical transient process during the operation of the power system, but the calculation amount and calculation time are very large. stability, short time and fast speed. However, due to the corresponding simplification and assumption of the system when constructing the transient energy function, the calculation results are too conservative and not accurate enough.

With the continuous development of the intelligent construction of the power system, a large number of power electronic devices such as new energy and high-voltage DC modules have been introduced, which makes the power system model more complex, and it is difficult for the traditional time domain simulation method and direct method to accurately evaluate the system online. The artificial intelligence method relies on the wide-area measurement system (WAMS) and the phasor measurement unit (PMU) for real-time high-speed acquisition of the grid synchronization phase angle and main data. Taking the characteristic quantities of the power system during the steady state and transient period as the input, through the offline learning algorithm model, the transient stability of the power system can be quickly judged.Artificial intelligence algorithms for transient evaluation, including decision trees, random forests, support vector machines and deep learning, etc., among which input feature selection and dimensionality reduction are the most important issues. According to pattern recognition theory, the input features of classifiers are too much will increase the computational cost of knowledge discovery, reduce the accuracy of the training model, and even lead to the " curse of dimensionality " problem [2] . At the same time, it has been proved that the optimal feature subset selection is an NP-hard problem [3] , and the high-dimensionality of power system is an important problem in theoretical research and engineering practice [4]. Therefore, feature selection is an urgent problem to be solved in the research and application of power system steady-state assessment.

The feature selection problem of transient stability assessment [5-9] , scholars at home and abroad have carried out many explorations. Reference 5 uses grey relational clustering for feature selection, considering the degree of association between features, between features and class attributes, and compares and sorts them. It can effectively improve the classification performance on the AdaBoost algorithm integrated with the naive bayesian classifier as the base classifier . Reference 6 first used principal component analysis to reduce the dimension, and then used the sequence floating backward algorithm to eliminate the redundancy of features to improve the classification effect, and achieved good robustness and accuracy. Reference 7 introduces the recursive feature elimination method to sort the importance of features and adjust the ratio of stable and unstable samples to reduce the phenomenon of sample imbalance. Reference 8 uses variational auto-encoder (VAE) for feature extraction, combined with convolutional neural network (CNN) to filter noise, which has the characteristics of high evaluation accuracy, strong anti-noise interference ability, and low evaluation error rate for unstable samples. Reference 9 performs feature compression through the maximum correlation and minimum redundancy feature selection method (mRMR).

Based on reference 9 , this paper proposes to improve mRMR and apply it to feature selection for transient stability assessment. Considering the post-fault measured information that the PMU can provide, a new feature selection method for transient stability assessment, mRMR-RFOE-based feature selection, is proposed on the basis of constructing the original feature set composed of system features. First, the importance of the features is sorted by calculating the error value of the data outside the bag of RF , and then the mutual information between the features and between the features and the class variables is calculated by the filtering algorithm mRMR to sort the features again, and test different feature pairs. The effect of model accuracy to find the optimal number of features k . After the above two features are screened, the feature ranking is selected before compared with the unscreened feature set, the optimal subset greatly reduces the irrelevant redundant features and improves the accuracy of the steady-state evaluation of the power system significantly [10] . In the model classification stage, the random forest algorithm with higher classification accuracy is used as the classification model, thereby further improving the accuracy of the steady state assessment of the power system.

2. Basic Theory

2.1 Maximum Correlation Minimum Redundancy Criterion

Given two random variables x and y , their probability densities are p(x) and p(y) , and the joint probability density is p(x,y) , then the mutual information between x and y is defined as:

$$I(x;y) = \iint P(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \quad (1)$$

The measures of maximum correlation and minimum redundancy are defined as:

$$maxD(S,c), D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \quad (2)$$

$$maxR(S), R = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i;x_j) \quad (3)$$

In the formula, S and |S| are the feature set and the number of features included; c is the target category; $I(x_i;c)$is the mutual information between feature i and target category c ; $I(x_i;x_j)$ is the mutual information between feature i and feature j ; D is the mean value of mutual information between each feature x_i and category c in the feature set S, indicating the correlation between the feature set and the corresponding category; R is the size of the mutual information between the features in S , which represents the redundancy between the features.

The goal of feature selection is to expect the selected feature subset to have the highest classification performance and as few feature dimensions as possible, which requires the largest correlation between feature sets and categories and the smallest redundancy between features. Considering the above two metrics comprehensively, the maximum correlation minimum redundancy criterion is obtained as follows:

$$max\emptyset(D,R), \qquad \emptyset = D - R \quad (4)$$

2.2 Incremental Search Algorithm

In practical applications, an incremental search algorithm can be used to select approximately optimal features defined by $\emptyset(.)$. Assuming that the original feature set is X, and the feature subset containing m-1 features has been selected as S_{m-1}, the task of feature selection is to select from the remaining feature set {X- S_{m-1}} the m-th feature maximizes $\emptyset(.)$ in Eq. (4). This feature should satisfy:

$$max_{x_j \in X - S_{m-1}}[I(x_i;c) - \frac{1}{m-1}\sum_{x_i \in S_{m-1}} I(x_i;x_j)] \quad (5)$$

2.3 Random Forest Out-of-Bag Error

Random Forest [11] is an ensemble learning algorithm composed of multiple decision trees, and each decision tree is assigned an independent subspace and is allowed to grow freely. Finally, a simple majority vote is used to designate the category with the most votes as the final classification result . The training set of each decision tree in the random forest is composed of sample data that is repeated and randomly selected equal to N from the original data set N. This method is called Bootstrap Sampling , and the obtained sample set is called for the bootstrap set [12] . When using the self-service sampling method to collect data, if there are n original data , the probability that each data will not be sampled is $P = (1 - \frac{1}{n})^n$, when n tends to infinity, P≈36.8% , indicating that there are about 36.8 % of the original data. It will not appear in the training set. These data that do not appear in the training set are out-of-bag data (out-of-bag data , OOB data) . Random forest can calculate the importance of features and rank them . Common importance calculation methods are mainly divided into three types: frequency of statistical features as segmentation features, Gini index and OOB data calculation error value . In this paper , OOB data is selected to calculate the error value and sort the features. The specific steps are as follows:

Step 1: Use k groups of out-of-bag data (OOB data) to calculate the error value of each decision tree separately, denoted as$Err_{OOB}1, Err_{OOB}2, \cdots, Err_{OOB}k$.

Step 2: Randomly rearrange the i-th feature of the k groups of out-of-bag data and keep other features unchanged, and then recalculate the error value as $Err_i 1, Err_i 2, \cdots, Err_i k$.

Step 3: The formula for calculating feature importance as follows:

$$import_x = \frac{1}{k}\sum_{m=1}^{k}(Err_i m - Err_{OOB} m) \tag{6}$$

Step 4 : Features are sorted based on importance, and the top m features are selected based on the resulting optimal number of features m.

3. Transient Stability Feature Construction

Table 1. Input features of the dataset

number	Input feature quantity
1	Mean value of system mechanical power at time t_0
2 -3	The maximum/minimum value of the active power shock to the generator at time t_1
4	Average value of generator acceleration power at time t_1
5	The relative rotor angle of the generator with the largest acceleration at time t_1
6	Maximum relative rotor angle at time t_1
7-8	Average/maximum value of rotor acceleration at time t_1
9-10	Variance of generator rotor acceleration/acceleration power at time t_1
11	Average value of generator acceleration power at time t_2
12 -13	Maximum value/average value of generator rotor kinetic energy at time t_2
14	The impact on the system at time t_2
15	The relative rotor angle of the generator with the largest rotor kinetic energy at time t_2
16	Maximum relative rotor angle at time t_2
17	The maximum relative rotor angular velocity at time t_2
18	Generator rotor angular velocity range value at time t_2
19	Maximum relative rotor acceleration at time t_2
20	Minimum value of initial rotor acceleration at time t_2
21 -22	Rotor acceleration/rotor kinetic energy range value at time t_2
23	Sum of generator active power at time t_2
24	The difference between the rotor angles of the leading machine and the rear machine at time t_2
25	Generator acceleration power variance at time t_2
26	The variance of the initial acceleration of the generator at time t_2
27-28	Generator rotor angular velocity/mechanical power average value at time t_2
29	Rotor kinetic energy of generator with maximum rotor angle at time t_2
30	Total system energy adjustment before and after failure
31	Relative rotor angle range at t_1, t_2
32	Relative rotor angular velocity range value at t_1, t_2
33	Relative rotor acceleration range value at t_1, t_2

The original features of transient stability assessment can be classified from the following perspectives: from time, it can be divided into static features and dynamic features; from space, it can be divided into grid parameter features and generator parameter features [13] ; from the change of system scale , which can be divided into stand-alone features and system features [14] . In recent years, the introduction of PMUs has made it possible to obtain synchronized post-fault real-time information, thus providing new input features for TSA . The original feature construction in this paper adopts the following principles:

1) Mainstream principle: Through in-depth analysis of the essential characteristics of the physical process of the transient stability of the power system , select the feature quantities that are strongly related to the transient stability of the system.

2) Real-time principle: After the disturbance occurs, the feature quantity reflecting the operating state of the system can be quickly obtained.

3) Systematic principle: Select system features instead of stand-alone features, and the dimension of the input feature quantity does not increase with the increase of the system scale, so as to avoid the problem of dimensionality disaster.

According to the above principles, on the basis of synthesizing the existing research literature, through a large number of simulation analysis, this paper constructs a set of original feature sets consisting of 33 system features, as shown in Table 1. The selected features do not contain stand-alone features, so the feature dimension is independent of the system size.

Among them, t0 is the stable operation time of the system, t1 is the time when the fault occurs, and t2 is the time when the fault is removed. The selected original feature set includes the system features of each stage of the fault, which can comprehensively reflect the stability of the system.

4. mRMR-RFOE Feature Selection in Power System Transient Stability Assessment

4.1 Based on mRMR-RFOE Power System Transient Stability Evaluation Process

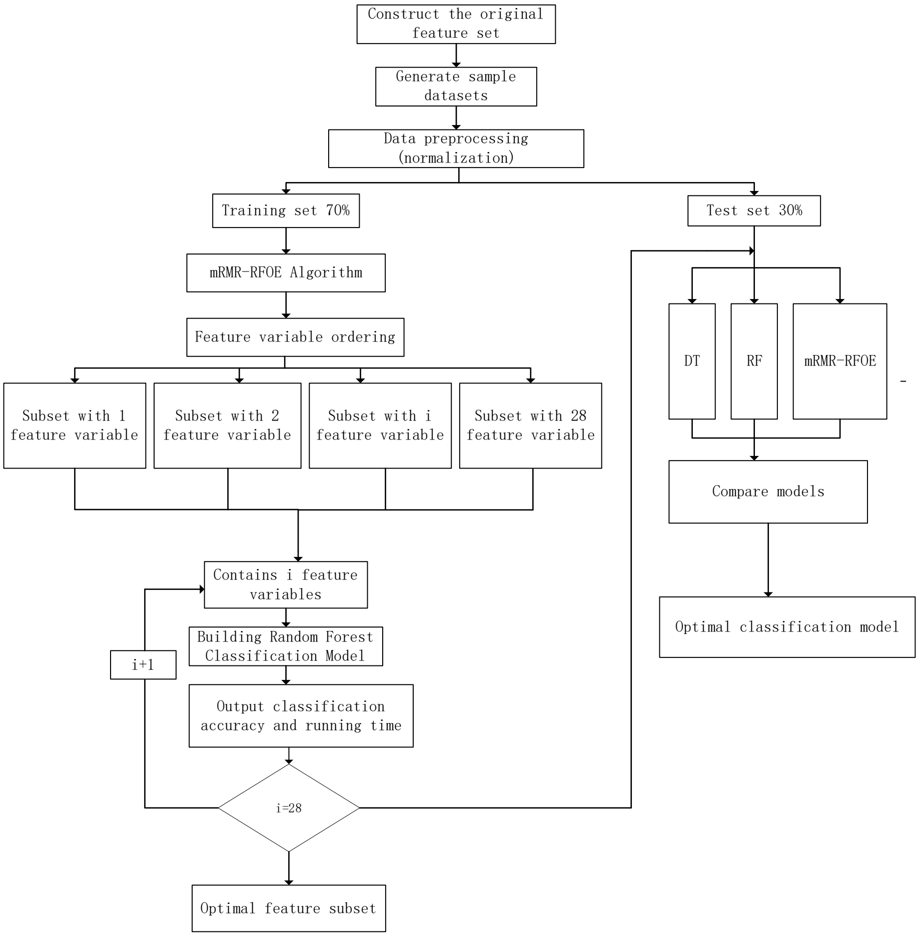
Construct the original feature set
Generate sample datasets
Data preprocessing (normalization)
Training set 70%
Test set 30%
mRMR-RFOE Algorithm
Feature variable ordering
Subset with 1 feature variable
Subset with 2 feature variable
Subset with i feature variable
Subset with 28 feature variable
DT
RF
mRMR-RFOE
Compare models
Optimal classification model
Contains i feature variables
Building Random Forest Classification Model
i+1
Output classification accuracy and running time
i=28
Optimal feature subset

Figure 1. Based on flow chart of power system transient stability assessment for mRMR-RFOE

The transient stability evaluation process of power system based on mRMR-RFOE is shown in Figure 1. , which mainly includes the following steps.

4.2 Example Introduction

The New England 10-machine 39-node system consists of 10 generators, 39 buses and 46 lines, representing a 345 kV power network in New England, USA, and the generator connected to the 39th bus is equivalent to the external grid machine. The system topology is shown in Figure 2.

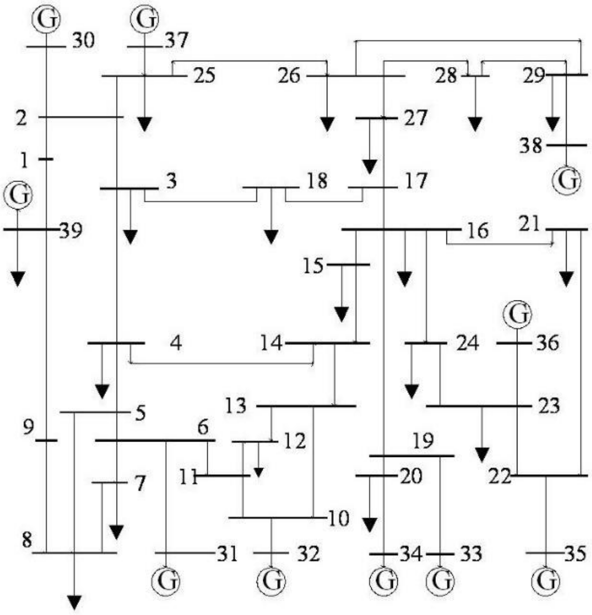

Figure 2. IEEE 39 node system

4.3 Construction of the Sample Set

The synchronous generator of the 39-node system uses the classic model, and its load is constant impedance. In terms of system load, 10 different load levels are set, ranging from 80% to 125%, with an increment of every 5%. At the same time, under each load condition, the generator makes corresponding output changes. The fault location is randomly selected in each operating environment, and the fault type is a three-phase short-circuit fault. The fault is set to occur at 0.1s, and the fault is cleared at 0.2s, 0.3s, 0.4s, and 0.5s, respectively. After the fault is cleared, the system circuit is closed again, and the system topology does not change. The overall simulation time is set to 5s, and the transient stable state of the system is judged by whether the power angle difference of any two generators in the system exceeds 180° at the end of the simulation. If it exceeds 180°, the system is judged to be transiently unstable, and the corresponding category label is marked as 0; otherwise, the system is judged to be stable, and the category label is marked as 1. The above system simulation runs are implemented in Matlab/Simulink. The model simulation generated a total of 1560 sets of sample data, including 940 sets of transiently stable samples and 620 sets of transiently unstable samples. The Min-max normalization method is used to normalize all sample data to eliminate the influence of the difference between attribute dimensions on the algorithm.

4.4 Evaluation Indicators

The transient stability evaluation is a two-class problem, and the classification evaluation indicators used in this paper are as follows.

Table 2. Confusion matrix

Evaluation result	Real results	
	Stablize	Unstable
Stablize	TP	FP
Unstable	FN	TN

TP represents the number of samples that are correctly classified as stable; TN represents the number of samples that are correctly classified as unstable; FN represents the number of samples that are misclassified as unstable; FP represents the number of samples that are incorrectly classified as stable.

Accuracy:

$$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Kappa coefficient:

For consistency check, classification accuracy measurement.

$$Kappa = \frac{p_0-p_e}{1-p_e} \tag{8}$$

p0 is the total classification accuracy is Accurary , which pe is the calculation method. In the binary classification problem, if the number of real samples in each class is A1 , A2 , the actual number of predicted samples of each type is B1 , B2; the total number of samples is N,

$$p_e = \frac{A_1 \times B_1 + A_2 \times B_2}{N \times N} \tag{9}$$

In this paper, the accuracy and Kappa coefficient are used as the final evaluation indicators. The larger the two indicators, the better the model performance.

5. Analysis of Results

5.1 Feature Selection based on mRMR-RFOE

The data set is randomly divided into 70% training set and 30% test set using stratified sampling. First, the importance of 33 features is sorted according to the random forest out-of-bag error RFOE , and 28 relatively important features are selected, as shown in the Fig. 3 , and then use mrmr feature selection to compare the classification accuracy and kappa coefficient of different feature numbers, and the results are shown in Fig. 4 . Finally, 8 important feature variables with low correlation coefficients were selected, and the overall classification accuracy rate has reached 98.29% , achieving an ideal diagnostic effect in a relatively short time. selected 8 The characteristic variables are [13, 15, 17, 18, 25, 27, 32, 33] in sequence.

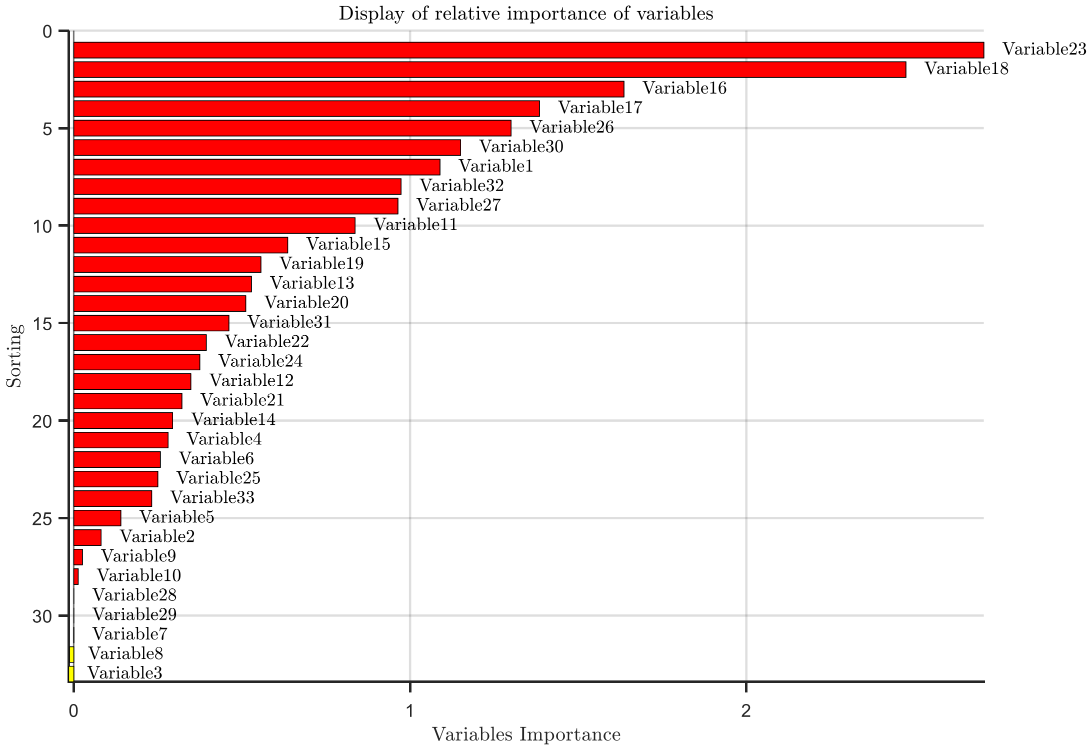

Figure 3. RFOE feature importance ranking

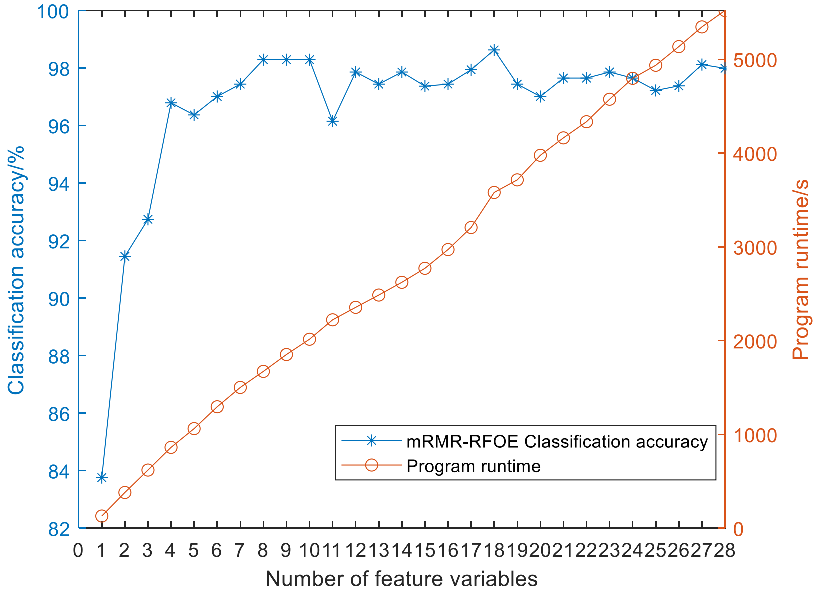

Figure 4. Changes in classification accuracy with the number of feature variables

Table 3 shows the correlation coefficient matrix between the filtered feature variables. Some of the original 28 feature variables are highly correlated. For example, the correlation between feature 28 and feature 17 is as high as 1. For the two variables with high correlation, one of them can be eliminated, which will not affect the model. Generalization ability has a big impact. The correlation among the 8 variables after screening is small, which further proves the rationality of the remaining 8 feature variables after feature selection.

Table 3. Correlation coefficient matrix between filtered feature variables

V	13	15	17	18	25	27	32	33
13	1	0.24	0.17	-0.01	-0.62	-0.19	0.17	-0.01
15	0.24	1	0.09	0.06	-0.19	-0.06	0.09	0.02
17	0.17	0.09	1	0.11	0.27	0.64	0.19	0.34
18	-0.01	0.06	0.11	1	0.38	0. 19	0.20	0.11
25	-0.62	-0.19	0.27	0.38	1	0.29	0.27	0.15
27	-0.19	-0.06	0.64	0.1 9	0. 29	1	0.63	-0.09
32	0.17	0.09	0.19	0.20	0.27	0.63	1	0.34
33	-0.01	0.02	0.34	0.1 1	0.15	-0.09	0.34	1

5.2 Overall Model Performance Comparison

5.2.1 Model Comparison

The decision tree (DT) and random forest (RF) with the algorithm proposed in this paper, the data set is divided according to the fault clearing time, 70% of each sample set is used to train the model, and 30% of the model is used to verify the performance of each model. The evaluation results are shown in Figure 5 and Figure 6; Take 70% of the total sample set as the training set and 30% as the test set. The evaluation results are shown in Table 4 . As shown in the figure, mRMR-RFOE has a higher evaluation index value than other classification algorithms on the sample sets with different fault removal times, can adapt to different data sets, and is less affected by the fault removal time in the transient process , indicating that the proposed model has good robustness and strong generalization ability.

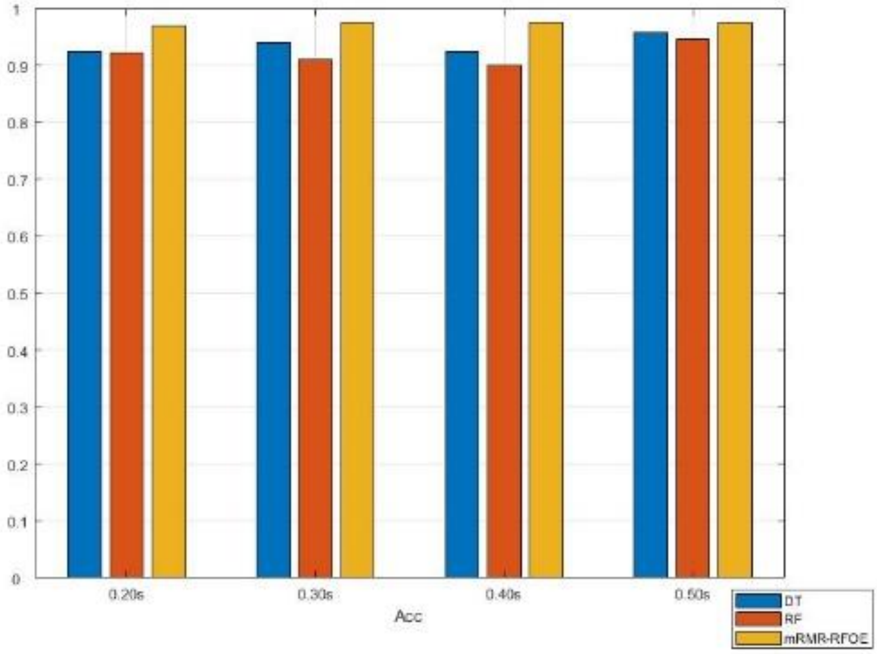

Figure 5. Model Acc comparison on different sample sets

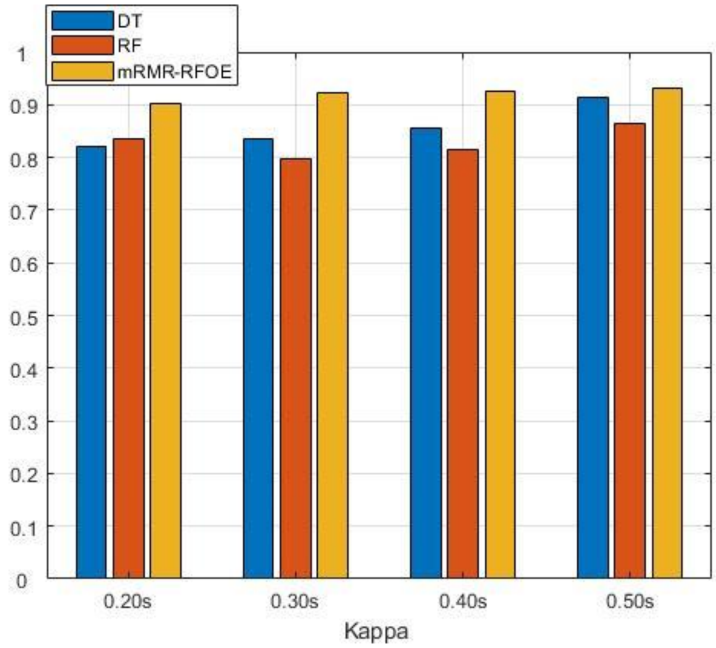

Figure 6. Model Acc comparison on different sample sets

With the accuracy rate and Kappa coefficient in Table 4 , mRMR-RFOE has a classification accuracy rate of 0.9829 and a Kappa coefficient value of 0.9657 , which is better than other classification algorithms. sex.

Table 4. Overall evaluation performance comparison

Algorithm	Accurary	Kappa	Program running time
DT	0.9530	0.9082	774s
RF	0.9786	0.9573	803s
mRMR - RFOE	0.9829	0.9654	1672s

5.2.2 Using Different Scale Test Sets

The models are trained with different sizes of training sets, the proportions of the training sets are 0.2 , 0.4 , 0.6 and 0.8 , and the rest are used as test sets, and then the evaluation ability of the model is verified as shown in Tables 5 and 6 . As shown in the table, when using different scales of training samples to train the model, mRMR-RFOE can maintain high accuracy even when there are few training samples, and when only 20% of the training set is used, the classification accuracy only drops to 0.9623 , while other classification algorithms, DT The most serious drop to 0.8619 ; DT dropped to 0.8 3 73 in terms of Kappa coefficient , compared to other trained models mRMR-RFOE It has strong learning ability, excellent performance on small data sets, strong model generalization ability, and can overcome the shortage of training samples in transient evaluation.

Table 5. Model accuracy Acc comparison

Training samples	0.8	0.6	0.4	0.2
Test sample	0.2	0.4	0.6	0.8
DT	0.9271	0.9407	0.9003	0. 86 19
RF	0.9676	0.9679	0.9541	0.9239
m RMR - RF0E	0.9840	0.9712	0.9679	0.9623

Table 6. Model Kappa Coefficient Comparison

Training samples	0.8	0.6	0.4	0.2
Test sample	0.2	0.4	0.6	0.8
DT	0.8591	0.8828	0.8546	0.8373
RF	0.9346	0.9340	0.9083	0.8533
mRMR - RFOE	0.9678	0.9405	0.9355	0. 9268

6. Conclusion

In this paper, the basic principle and training process of the mRMR-RFOE algorithm used in the transient evaluation of the power system are described in detail, and a transient evaluation model of the power system based on mRMR-RFOE is established. The simulation results are as follows:

1) The mRMR-RFOE algorithm can extract effective feature information and improve the accuracy of transient evaluation. This method has higher accuracy than conventional classification algorithms such as DT and RF;

2) The mRMR-RFOE algorithm has limited training samples, that is, it performs well on small sample data sets, and has a higher evaluation accuracy advantage than other classification algorithms;

3) The model framework of this paper can better abstract features and is less affected by changes in fault conditions . It has better evaluation performance on sample sets with different fault removal times, and the model has better robustness and generalization ability.

References

[1] Tang Yi , Cui Han , Li Feng , et al . A review of the application of artificial intelligence in power system transient problems [J]. Chinese Journal of Electrical Engineering , 2019 , 39(1): 4-15, 317.

[2] Jain A k, Duin R P W, Mao J C.Statistical pattern recognition: a review[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37.

[3] Guyon I, Elisseeff A.An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003(3): 1157-1182.

[4] Wang Hao, Sun Hongbin, Zhang Boming, etc. Feature selection method based on hybrid mutual information and its application in static voltage stability evaluation [J] . Chinese Journal of Electrical Engineering, 2006 , 26(7) : 77-81.

[5] Lu Jinling, Li Hongwei, Liu Haijun. Research on transient stability assessment method based on integrated Bayesian classifier [J]. Journal of North China Electric Power University (Natural Science Edition), 2010,37(03):14-20.

[6] Li Xiangwei , Liu Siyan , Gao Kunlun . Evaluation method for transient stability of power system based on differential evolution extreme learning machine [J]. Science Technology and Engineering , 2020,20(01): 213-217.

[7] Zhang Linlin , Hu Xiongwei , Li Peng , Shi Fang , Yu Zhihong . Transient Stability Evaluation Method of Power System Based on Extreme Learning Machine [J].Journal of Shanghai Jiaotong University ,2019, 53(06):749-756.DOI :10.16183/j.cnki.jsjtu.2019.06.017.

[8] Zhou Yue , Tan Bendong , Li Miao , Yang Xuan , Zhou Qiangming , Zhang Zhenxing , Tan Min , Yang Jun . Transient stability assessment method of power system based on deep learning [J]. Electric Power Construction , 2018,39(02):103-108.

[9] Li Yang , Gu Xueping . Feature Selection for Transient Stability Assessment Based on Improved Maximum Correlation Minimum Redundancy Criterion [J]. Chinese Journal of Electrical Engineering , 2013, 33(34): 179-186+27. DOI: 10.13334/j.0258 -8013.pcsee.2013.34.024.

[10] Rajab K D. New hybrid features selection method: a case study on websites phishing[J]. Security & Communication Networks, 2017(2) : 1-10.

[11] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[12] Li Xian, Wang Yan, Luo Yong, etc. Nasopharyngeal tumor segmentation based on random forest feature selection algorithm[J]. Computer Applications, 2019 , 39(5) : 1485-1489.

[13] Ye Shengyong , Wang Xiaoru , Liu Zhigang , et al . Two-stage feature selection for transient stability assessment based on support vector machine [J]. Chinese Journal of Electrical Engineering , 2010, 30(31): 28-34.

[14] Gu Xueping, Zhang Wenchao. Input Feature Selection of Transient Stable Classification Neural Network Based on Tabu Search Technology [J]. Chinese Journal of Electrical Engineering, 2002, 22(7): 66-70.