Feature Selection of Power System Transient Stability Assessment Based on Random Forest and Recursive Feature Elimination

Chun ZHANG
Yansong LI
Department of Electric Power Engineering
North China Electric Power University
Changping District, Beijing 102206, China
E-mail:chun_ai3@163.com

Zhihong YU
Fang TIAN
China Electric Power Research Institute
Haidian District, Beijing 100192, China
E-mail:zhhyu@epri.sgcc.com.cn

***Abstract*—A feature selection algorithm based on random forest and recursive feature elimination is proposed in this paper, which can extract the key feature subset of transient stability assessment. Firstly, the weighted random forest is adopted to pay special attention to the unstable samples, which constitute only a very small minority of the power system operation samples. Secondly, random forest combined with the recursive feature elimination strategy is used to determine the most reasonable threshold to select important features. The threshold that distinguishes a feature whether important or not is determined by the change of OOB error rate. Finally, the classification performance of the reserved feature subset is tested by support vector machine. Case study conducted on the New England 39-bus power system exhibits the validity of the proposed algorithm.**

***Index Terms*-- feature selection, machine learning, random forest, recursive feature elimination, transient stability assessment, power system**

I. INTRODUCTION

Transient stability assessment(TSA) is the basic means to ensure the safe and stable operation of power system. In recent years, with the rapid development of computer science and information technology, and cloud computing of big data, TSA based on machine learning and pattern recognition techniques has got much more attention and showed much promise[1].

The main tasks of TSA based on machine learning is establish mapping relationship between the characteristic features of system and system states, including feature extraction and selection, construction of training and testing sample set, classifier training, classification performance evaluation. The selection of input features is first step in the TSA implementation[2]. And feature selection for classification is a crucial task in TSA studies, where researchers try to select most relevant original features from the set of features that have predictive information, which can improve accuracy of the classifier, decrease the data dimensionality, and reduce training time

In view of the feature selection problem of power system transient stability assessment, the relevant research has been done by scholars at home and abroad. A new feature selection method based on an improved maximal relevance and minimal redundancy (mRMR) criterion was proposed for TSA[2]. In [3], a dual-stage feature selection method based on support vector machine was proposed. Ref.[4] proposed a embedded features selection method based on ACO and k-NN which adopted the clustering method firstly to cut out redundancy, then selected key features strongly correlated to stability. Feature importance of random forests is also used for the purpose of feature selection in Ref.[5]. However, the subset sizes of selected features have no effective methods to determine.

In this paper, we proposed a feature selection method based combination of random forest and recursive feature elimination (RF-RFE). Random forest (RF) is an ensemble classification algorithm with an internal measure of features importance[6]. Although random forest can calculate the weights of the features, and sort the features in descending order according to the weights, it tend to have unsatisfactory performances towards imbalanced data and cannot determine score threshold. Therefore, we present a new approach to solve those problems through weighted class and combining with recursive feature elimination strategy, which is a circulation procedure for eliminating features by a criterion. Experiments show that RF-RFE can overcome these defects and has a good performance.

The paper is organized as follows: In Section Ⅱ , we introduce the theory of random forest and application in feature selection. Section Ⅲ presents weighted random forest and the details of RF-RFE method. In Section Ⅳ , the experiment have been carried out, including original feature set, test system along with offline simulation is shown. And

Project Supported by the National Basic Research Program of China (973 Program) (2013CB228203) and Science and Technology Project of SGCC(XT71-15-001).

978-1-5090-5417-6/16/ $ 31.00 ©2016 IEEE

we present the performance of the proposed method and discussions. Finally, the conclusions of this paper are presented in section V.

II. THEORY OF RANDOM FOREST

A. Review of Random Forest

Random forest (RF) is a well-known machine learning algorithm, proposed by Breiman in 2001, that uses an ensemble of unpruned decision trees. The final result of an input sample is determined by the majority classification voting. In the process of model training, each decision tree is built on a bootstrap sample of the training data using a randomly selected subset of variables[7]. Therefore, some samples which are not used in the training process are called the "Out-of-Bag Samples (OOB)". They are used to evaluate the generalization performance of classifiers and the OOB error estimation provides an unbiased assessment of the accuracy as well as cross-valid. Further-more, compared with cross-valid, the OOB error is simpler and time saving. The diagram of random forest is shown in Figure1.

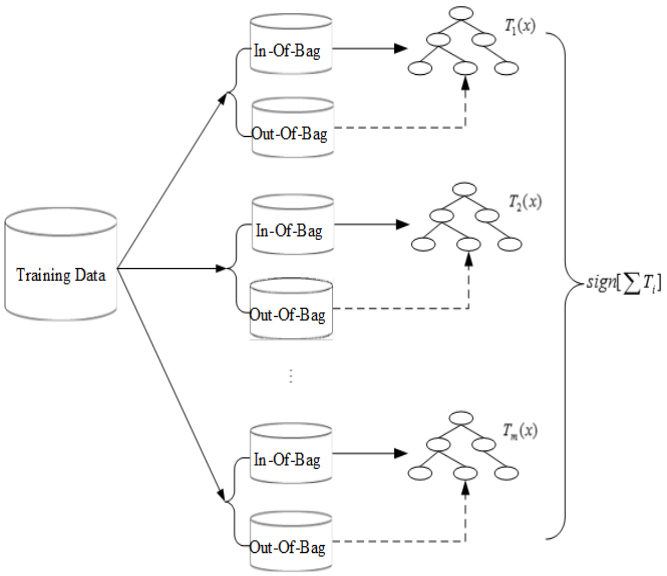

Figure 1. The flow diagram of random forest

B. Importance Scores Computation from Random Forest

In addition to classification and regression, RF provides an internal measure of feature importance by computing feature importance scores, which could be used to select important features. During constructing a random forest, each node of a decision tree is split into two children while a splitting criterion is reducing the impurity of a node that is measured by Gini importance[8]. In the process of a node splitting, ***i*** is the impurity of this node, its Gini importance is defined as:

$$i = 1 - \sum_{j} p^2(j) \qquad (1)$$

Where $p(j)$ is proportion of samples that is labeled j in this node. After spliting, the impurity of node is declined as:

$$\Delta i = i_{parent} - (p_{left} \cdot i_{left} + p_{right} \cdot i_{right}) \qquad (2)$$

Where p_{left}、p_{right} are sample proportion of left child node and left child node respectively, and i_{parent},i_{left},i_{right} are Gini importance of parent,left child and left child node respectively. For any one feature X_i, the sum of its impurity decrement in all decision trees is the Gini importance of X_i:

$$\alpha \Delta I = \sum_{k} \Delta i_k . \qquad (3)$$

This formula indicates the importance of every feature, and the greater value means that the feature is more important.

III. METHODOLOGY

A. Weighted Random Forest

Usually, random forest can be highly effective. However, the modern power system is pretty smart and strong. When a disturbance occurs in the power system, the case of rotor angle between two generators out-of-step is rare. Therefore, in the power system operation samples, unstable samples constitutes only a very small minority of the data. For such problems, the most commonly used random forest algorithm does not work well for such problems because it aims to minimize the overall error rate, rather than paying special attention to the "rare" class[9]. Therefore, a weighted random forest algorithm, which leans the interest towards correct classification of the rare class, is proposed to tackle the problem of imbalanced data in this paper. Stable samples are given smaller weight while unstable samples are given larger weight. In this paper, class weight of stable samples is set to 1 and the class weight of unstable samples is 3.

B. The RF-RFE Method

Random forest is one efficient filter feature ranking algorithm, which runs fast and can scale to very high dimensional datasets. It calculates the weights of the features, and sorts the features in descending order according to the weights. But the feature importance threshold cannot be calculated which decide how many features ranked ahead should be reserved. Since the threshold usually is determined by manual, it may affect the validity of feature selection. Hence we proposed a random forest recursive feature elimination algorithm(RF-RFE), which eliminates the feature having the smallest weight in each iteration[6]. This process will continue until the feature set is only one left. The final feature rank is got according to the eliminating order.

While removing one feature at a time from the previous set, the subset of features is evaluated by OOB error rate and OOB error rates are stored to obtain the curve relating feature sets and OOB error rates. From the curves of the graph, it is clear that OOB error rate is almost not changed at the start, because the excluded features are considered to make no contribution to classification ability. Once it rises sharply due to the elimination of important feature, the point is treated as the

threshold and the left subset of features should be reserved. Fig.2 shows the pseudo-code of the algorithm

Input:

Data set T

Set of p original features $F = \{f_1, \ldots, f_p\}$

Output:

Subset of features

Code:

Final ranking *R*

Repeat for *i* in $\{1 : p-1\}$

Rank set *F* using random forest

$f^* \leftarrow$ last ranked feature in *F*

$R(p-i+1) \leftarrow f^*$

$F \leftarrow F - f^*$

The size of feature subset is determined by the change of OOB error rates

Figure 2. Pseudo-code for the RF-RFE

IV. EXPERIMENTS

A. Test System

The New England 39-bus test system is used to demonstrate the usefulness of the proposed algorithm. This test power system model has been widely used in studies and documented in the literature for testing TSA algorithms, which involves 39 buses, 10 generation units, 19 loads, and 46 transmission lines. The reference power is 100MVA and the reference voltage is 345kV[10]. The generator is a 4th order model, and the load is the constant impedance model. The one line diagram of the test system is shown in Figure 3.

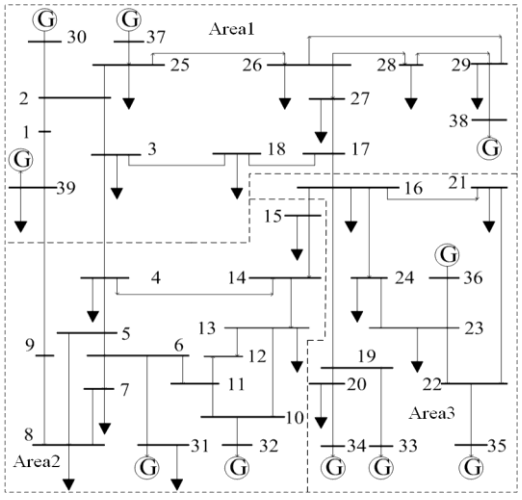

Figure 3. The New England 39-bus test system

B. Initial Input Feature Set

The choice of initial input feature set is first step of TSA. In order to help operators find key monitoring features and adjust the operation mode of power system promptly. Initial feature sets are obtained from steady variables of power flow. Based on the previous studies[11],[12],the initial features are presented as followed

TABLE I. TABLE TYPE STYLES

Feature Type	Number
the sum of all generators active and reactive output	2
the sum of system active and reactive loads	2
active and reactive shortage of system	2
the sum of area generators active and reactive output	6
the sum of area active and reactive loads	6
active and reactive shortage of area	6
branch active power flows	38
branch reactive power flows	38
branch current flows	38
active and reactive output of generators	20
active and reactive load of load bus	38
bus voltage magnitude	29
bus voltage phase angle differences	38

C. Power System Simulation

Data sets required for test are generated through offline dynamic simulation software package. And the numerical simulation is carried out by the commercial software PSD-BPA. In order to take into account all the possibilities, operating conditions of the test system are randomly changed. Considering 5 different kinds of load level (85%, 90%, 95%, 100%, 105%). Generator output is randomly changed, and the range of its fluctuation is 85% to 115%. Under each load condition, there are 400 different operation modes. Then, the power flow is solved with the changed load and generation. The power mismatch is assumed to be balanced by the swing generator on bus 39.If the power flow is converged[13]. A three-phase to ground fault at 50% of the length on transmission line between bus3 and bus4, and the fault line is cut off after 1s.The results calculated by the time-domain simulation are characterized as stable and unstable. The criterion for instability is whether the difference between any two generator angles exceeds 360-degree at the end of the simulation[2],[3],[14]. Finally, 2000 effective samples are generated totally and stored in the database, with 1790 stable samples and 210 unstable samples.

D. Feature Selection Result

The main parameter configurations of random forest are ntree and mtry. ntree is number of trees to build which is set to 300 and mtry denoting the number of features to be randomly selected for each tree which is set to $\sqrt{p}$ (p is the total number of features)[7].

Applying the RF-RFE algorithm proposed in this paper, the Figure4 is obtained. From the curves of the graph, it is clear that the elimination of important feature would lead to the sharp rise of OOB and 45 features ranked ahead are reserved at last.

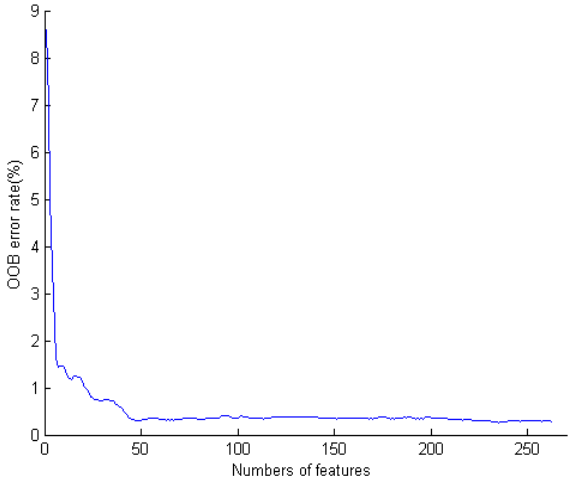

Figure 4. The number of features and OOB error rates

We performed an experiment to demonstrate the effectiveness of the proposed method by comparing it with the Boruta method. Boruta is a common algorithm that can evaluates variable importance by creating an ensemble of corresponding artificially added 'shadow' variables randomly sampled from the dataset and compute the importance scores of features[15]. Then, a linear Support Vector Machine (SVM) is used as the classifier to test the validity of selected feature. 1333 samples are randomly selected from all samples to train the model. The rest 667 samples compose the test sample set. The results are presented in Table Ⅱ .

TABLE II. THE RESULT OF FEATURE SELECTION

	Number of Features	False Positive	False Negative	Accuracy/ %
Original set	263	7/65	1/603	98.80
Boruta	132	7/70	2/598	98.65
RF-RFE	45	5/73	1/595	99.10

From the table2, we can conclude that RE-RFE can select more compact feature subsets compared to Boruta method, while preserving recognition accuracy.

At the same time, Sammon mapping algorithm is used to demonstrate the effectiveness of the proposed method from the perspective of data visualization. Sammon's projection is a nonlinear projection method to map a high dimensional space onto a space of lower dimensionality(usually 2D) and the metric relations of the data items are preserved as faithfully as possible. Sammon's projection is useful for preliminary analysis in all statistical pattern recognition, because a rough visualization of the class distribution can be obtained, especially the overlap of the classes[16],[17].

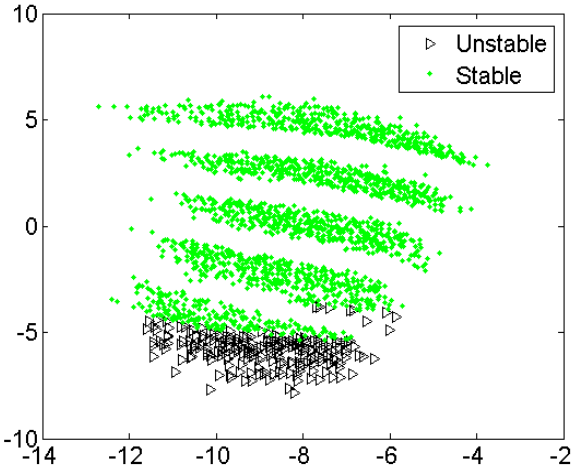

Figure 5. Mapping of original features

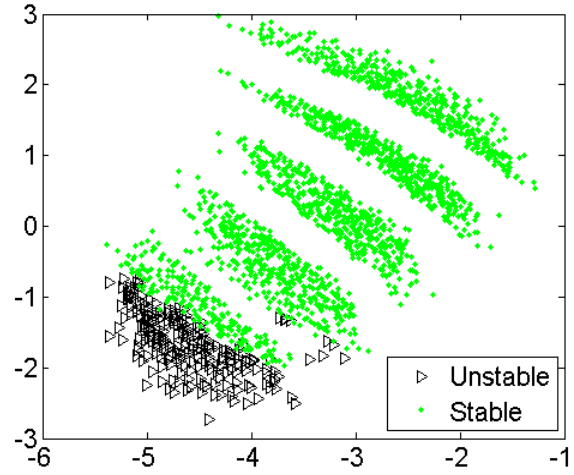

Figure 6. Mapping of feature subset selected by RF-RFE

The mapping result of classification without feature selection is shown in Figure5, and the result of 45 selected features is Figure6. From the above results, we can find that the information contents of original features are well preserved by selected feature because the mapping results of them almost maintain consistent, which proved the effectiveness of the proposed methodology from the other point of view.

V. CONCLUSION

This paper introduced a efficient feature selection method of power system transient stability assessment—RF-RFE. Random forest is an effective method for measuring feature importance but does not account for class imbalance. In addition, it is impossible to determine the threshold of feature importance. The proposed RF-RFE method can overcome such defects had been applied in the New England 39-bus system, comparing to well-known feature selection algorithm Boruta. The experimental results showed that the RF-RFE method had better performance than Boruta in the precision and feature space during the assessment.

REFERENCES

[1] Wang, Tong Wen, L. Guan, and Y. Zhang. "A Survey on Application of Artificial Intelligence Technology in Power System Stability Assessment." *Power System Technology* 33.12(2009):60-65.

[2] Yang, L. I., and G. U. Xueping. "Feature Selection for Transient Stability Assessment Based on Improved Maximal Relevance and Minimal Redundancy Criterion." *Zhongguo Dianji Gongcheng Xuebao/proceedings of the Chinese Society of Electrical Engineering* 33.34(2013):179-186.

[3] Shengyong, Y. E., et al. "Dual-stage Feature Selection for Transient Stability Assessment Based on Support Vector Machine." *Zhongguo Dianji Gongcheng Xuebao/proceedings of the Chinese Society of Electrical Engineering* 30.31(2010):28-34.

[4] Zhang, Xiao Qiang, and L. Guan. "Feature Selection for Transient Stability Assessment Based on ACO and k-NN." *Guangdong Electric Power* (2011).

[5] Yanhao H, Zhihong Y, Dongyu S, Xiaoxin Z. “Strategy of huge electric power system stability quick judgment based on massive historical online data.” *Zhongguo Dianji Gongcheng Xuebao/Proceedings of the Chinese Society of Electrical Engineering* (2016):596-603.

[6] Granitto, Pablo M., et al. "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products." *Chemometrics & Intelligent Laboratory Systems* 83.2(2006):83-90.

[7] Statnikov, A, and C. F. Aliferis. "Are random forests better than support vector machines for microarray-based cancer classification?." *AMIA. Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2007(2007):686-90.

[8] Breiman, Leo. "Random Forests." *Machine Learning* 45.1(2001):5-32.

[9] Chen, Chao, and L. Breiman. "Using Random Forest to Learn Imbalanced Data." unpublished. [Online]. Available: http://stat-www.berkeley.edu/tech-reports/666.pdf

[10] Wang, Bo, et al. "Power System Transient Stability Assessment Based on Big Data and the Core Vector Machine." *IEEE Transactions on Smart Grid* (2016):1-1.

[11] M. He, V. Vittal and J. Zhang, "Online dynamic security assessment with missing pmu measurements: A data mining approach," in *IEEE Transactions on Power Systems* 28.2(2013):1969-1977.

[12] He, Miao, J. Zhang, and V. Vittal. "Robust Online Dynamic Security Assessment Using Adaptive Ensemble Decision-Tree Learning." *IEEE Transactions on Power Systems* 28.4(2013):4089-4098.

[13] Geeganage, J., et al. "Application of Energy-Based Power System Features for Dynamic Security Assessment." *IEEE Transactions on Power Systems* 30.4(2015):1957-1965.

[14] Ye, Shengyong, et al. "Power System Transient Stability Assessment Based on Adaboost and Support Vector Machines." *Power and Energy Engineering Conference (APPEEC), 2012 Asia-Pacific* IEEE, 2012:1-4.

[15] Poona, N. K., and R. Ismail. "Reducing hyperspectral data dimensionality using random forest based wrappers." *IGARSS 2013 - 2013 IEEE International Geoscience and Remote Sensing Symposium* 2013:1470-1473.

[16] B. Lerner et al. "On the Initialisation of Sammon’s Nonlinear Mapping." *Pattern Analysis & Applications* 3.1(2000):61-68.

[17] Günther M. FOIDL. “Sammon Projection” unpublished. [Online]. Available:http://www.codeproject.com/articles/43123/sammon-projection