2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) | 978-1-6654-3892-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICMTMA52658.2021.00087

Research on network intrusion detection method of power system based on random forest algorithm

Guowei ZHU*, Hui YUAN, Yan ZHUANG, Yue GUO, Xianfei ZHANG, Shuang QIU
State Grid Information&Communication Branch of Hubei Electric Power Co., Ltd, Wuhan, 430077, China
E-mail address: guoweizhu@whu.edu.cn,
328699780@qq.com,13986270090@139.com,992941480@qq.com,729774042@qq.com,296227500@qq.com

Abstract: Aiming at the problem of low detection accuracy in traditional power system network intrusion detection methods, in order to improve the performance of power system network intrusion detection, a power system network intrusion detection method based on random forest algorithm is proposed. Firstly, the power system network intrusion sub sample is selected to construct the random forest decision tree. The random forest model is optimized by using the edge function. The accuracy of the vector is judged by the minimum state vector of the power system network, and the measurement residual of the power system network attack is calculated. Finally, the power system network intrusion data set is clustered by Gaussian mixture clustering Through the design of power system network intrusion detection process, the power system network intrusion detection is realized. The experimental results show that the power system network intrusion detection method based on random forest algorithm has high network intrusion detection performance.

Key word: Random forest algorithm; Power system; Network intrusion; Intrusion detection;

I. INTRODUCTION

Firstly, the power system network intrusion sub sample is selected to construct the random forest decision tree. The random forest model is optimized by using the edge function. The accuracy of the vector is judged by the minimum state vector of the power system network, and the measurement residual of the power system network attack is calculated. The power system network intrusion data set is clustered by Gaussian mixture clustering .Finally,through the design of power system network intrusion detection process, the power system network intrusion detection is realized. The experimental results show that the power system network intrusion detection method based on random forest algorithm has high network intrusion detection performance.

Power system network is a typical physical information system, which mainly realizes the security and reliability of power system network operation by integrating power transmission and network communication of physical information. However, due to the high dependence ratio of power system network on power data communication, it is vulnerable to physical attacks from external network, so network intrusion detection is produced [1]. Network intrusion detection is to identify the illegal use weight of users through the analysis results of relevant information in the network, and shield the users who are authorized by the network but still have authority abuse. Network intrusion detection method based on the power system software performance and hardware performance has a low impact, real-time monitoring of users logging into the power system malicious operation behavior [2]. The power system network can be aware of network intrusion mode, functional structure configuration, and software in low version for a long time. The user behavior monitoring function of network intrusion detection method can effectively monitor the target objects logged into the power system [3].

Dai Yuanfei et al,[4] aiming at the problem of noise features or redundancy in traditional network intrusion detection methods, which leads to the reduction of the accuracy of intrusion detection model and the long training time, the feature selection algorithm is applied to network intrusion detection, and a differential optimal feature subset is generated by using feature selection algorithm and different discretization, and it is normalized The results show that this method can improve the accuracy of network intrusion detection and effectively reduce the training time of detection model Xin et al,[5] aimed at the problem that network intrusion detection is facing with low detection efficiency, sequence mining was applied to the design of network intrusion detection method. Firstly, based on the inherent rules of normal network connection sequence, an attribute selection algorithm was designed to realize the automation of attribute selection, and the multidimensional sequence of network intrusion was compressed into one-dimensional sequence, and the sequence mining was used for training Network, through the connection to get a normal rule base, and finally determine whether the network connection is normal, through the experimental verification, the intrusion detection method is more efficient.

Based on the above research background, this paper applies random forest algorithm to the design of power system network intrusion detection method, so as to improve the performance of power system network intrusion detection.

II. DESIGN OF NETWORK INTRUSION DETECTION METHOD IN POWER SYSTEM

2.1 Building random forest

On the basis of random forest, the construction of combined multi classifier has certain advantages, and there is no over fitting phenomenon in the operation process [6]. However, when the redundancy characteristics of network intrusion data of power system are obvious or the imbalance of data set is high, the final classification effect will be different from the expected effect. First of all, it is necessary to select the

978-1-6654-3892-6/21/$31.00 ©2021 IEEE
DOI 10.1109/ICMTMA52658.2021.00087

corresponding power system network intrusion sub samples and build a decision tree to complete the modeling. In the process of combining multiple classifiers, cart algorithm is used to prune the classifiers, and the majority voting is used to complete the combination processing of classifiers [7].The random forest model is shown in Figure 1.

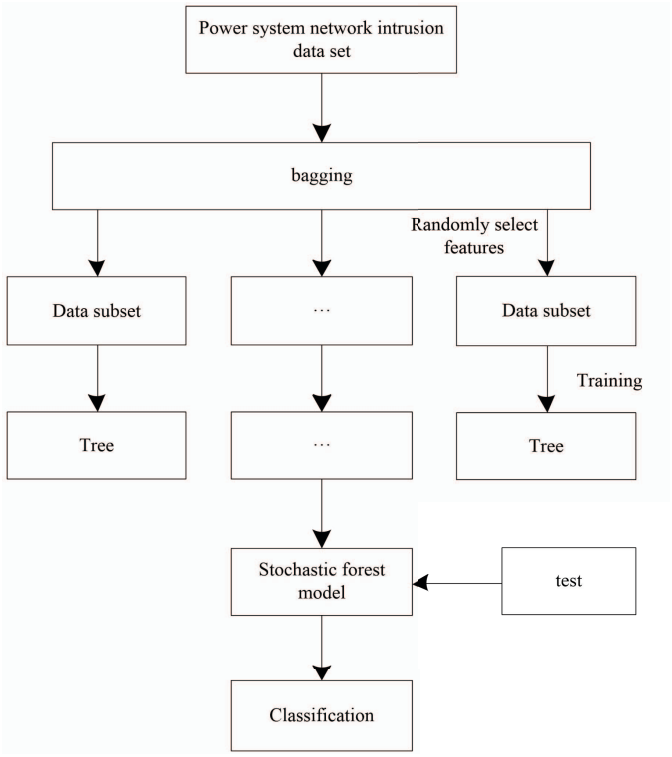

Figure 1 random forest model

In the process of continuous formation of random forest decision tree, multiple classifiers are transformed from single state to combined state. If there is close correlation between single tree and forest, the classification result of random forest will become worse. In the process of precision test, some key conditions should be controlled under the condition of low precision as far as possible to ensure that the trees are growing too much The depth in the process is maximized. When optimizing the random forest model, the edge function can be used to detect the model [8],Then the edge function is expressed as:

$$\mathrm{mg}(X,Y)=\mathrm{av}_k\left(I\left(h_k(X)=Y\right)\right)-\max\mathrm{av}_k\left(I\left(h_k(X)=j\right)\right) \quad (1)$$

$I(\cdot)$ represents the corresponding indicative function, Y represents the correct classification vector, j represents the wrong classification vector, and av_k represents the average value. In the process of decision tree operation, if the correlation between trees is relatively small, it can be considered that the classification ability of a single tree will be stronger.

2.2 Establishment of power system network attack model

Stochastic forest model can provide reliable power system data and estimate network intrusion state in real time.The measurement vector $Z=\left[Z_1,Z_2,\cdots,Z_m\right]^{\mathrm{T}}\in\mathbf{R}_m$ is used to represent the measurement data collected by random forest model, including node voltage, current and branch power of power system [9].The observable model of formula (2) can be used to estimate the state of power system network.

$$Z=H(Y)+E \quad (2)$$

$Y=\left[Y_1,Y_2,\cdots,Y_n\right]^{\mathrm{T}}\in R_n$ Represents the state vector of the power system, and $E=\left[E_1,E_2,\cdots,E_m\right]^{\mathrm{T}}\in R_m$ represents the measurement error vector, $E\sim N\left(0,\sigma^2\right)$ it usually obeys Gaussian distribution,and $H(Y)=\left[H_1(Y),H_2(Y),\cdots,H_m(Y]\right]^{\mathrm{T}}\in\mathrm{R}_{\mathrm{m}\times\mathrm{n}}$ Jacob ian matrix representing network topology of power system.

The criterion of power system network state estimation is to solve the minimum power system state

vector $\hat{Y}$, using the least square method to find and observe the best state vector of the model vector $\hat{Y}$, the minimum state vector of power system network can be expressed as follows:

$$\hat{Y} = \left(H^{\mathrm{T}} W H\right)^{-1} H^{\mathrm{T}} W Z \qquad (3)$$

W represents the weight matrix ,replacing the calculated state vector $\hat{Y}$ into the formula (2) , calculating the measurement residuals $r = Z - H\left(\hat{Y}\right)$, that is to say, the vector deviation between the measurement vector and the estimated measurement vector, so as to judge the accuracy of the vector.

Assuming that the external network attack the Jacobian matrix of the power system network topology , and it has certain control ability, then the network attacker will design a network attack vector $F = \left[F_1, F_2, \cdots, F_m\right]^{\mathrm{T}} \in R_m$, and in this case, the observation model of formula (4) can be used to estimate the state of power system network,

$$Z_F = H(Y) + E + F \qquad (4)$$

After the state estimation of power system network, it will get a wrong state vector $\hat{Y}_F = \hat{Y} + V$ At this time, the measurement residual of network attack can be calculated：

$$r_F = Z_F - H\left(\hat{Y}_F\right) = Z - H(\hat{Y}) + F - H(V) \qquad (5)$$

When F equals to $H(V)$, $r_F = r$, Then the state estimation of power system network can not detect the external network attack vector *F*. Before the state estimation of power system network, the collected measurement data of power system network should be detected to realize the establishment of power system network attack model.

2.3 Clustering the process of power system network intrusion data set

Gaussian mixture clustering is used to cluster data sets of power system network intrusion. Gaussian mixture clustering model is a probabilistic clustering method. Gaussian distribution is used as the parameter model of clustering processing of intrusion data sets [10]. The probability density function of Gaussian distribution is：

$$f(x|\ \mu) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad (6)$$

μ represents the mean value of Gaussian distribution , and σ representing the standard deviation of Gaussian distribution. , so for a Gaussian mixture model,there is *k* Gaussian distribution：

$$P\left(x_j\right) = \sum_{i=1}^{K} w_i \cdot P\left(x_j|\ \mu_i, \Sigma_i\right) \qquad (7)$$

x_j represents that in j network intrusion feature vector samples of dimensional power system , the $P\left(x_j|\ \mu_i, \Sigma_i\right)$ means that the mean value is μ_i .The covariance matrix is Σ_i and means the probability density function , representing the weight of Gaussian distribution in the model w_i . The probability density representation of Gaussian distribution can be calculated by using formula (6),

$$P\left(x_j|\ \mu_i, \Sigma_i\right) = \frac{1}{(2\pi)^{\frac{j}{2}} \left|\Sigma_i\right|^{\frac{1}{2}}} \cdot e\left\{-\frac{1}{2}\left(x_j - \mu_i\right)\sum_i^{-1} x_j - \mu_i\right\} \qquad (8)$$

According to the formula(8), it can be seen that the Gaussian mixture model belongs to a density function of parameter probability type, which can be modeled by the weighting of Gaussian components and continuous distribution of arbitrary precision [11]. The sample space generation of Gaussian mixture model is based on $w_1 \cdot w_2, \cdots + w_K$ probability and generates corresponding network intrusion data samples according to probability density function.

Assuming that there is N power system network intrusion data sample $C_N = \left(C_1; C_2; \cdots C_N\right)$, and for each power system network intrusion data sample, the probability density function of power system network intrusion data sample can be obtained as follows:

$$P(X, C) = \prod_{j}^{N} P\left(x_j|\ C_j\right) \qquad (9)$$

The clustering process of Gaussian mixture model is the inverse process of generating power system network intrusion data samples [12]. Firstly, given the number of clustering clusters of power system network intrusion, the mean vector *K*, covariance matrix and weight of each mixed component can be obtained through a specific power system network intrusion data set. Each component will correspond to a cluster, and a posterior density characteristic is usually used each cluster is represented by the eigenvector.

After clustering power system network intrusion data by Gaussian mixture model, the power system network intrusion data samples will be divided into different types of clusters. Because the Gaussian mixture model itself has the clustering method of probability density function, the correlation between each cluster is relatively small, which is convenient for random forest algorithm to analyze the network intrusion data of power system and the training of samples has good expected effect [13].

2.4 The implementation of network intrusion detection in

power system

The random forest algorithm is used to detect the network intrusion of power system. The algorithm can be regarded as a set of many decision trees, which does not need prior knowledge and is easy to explain. However, if the random forest algorithm is used alone, it can not guarantee the minimum correlation of the power system network intrusion data samples, resulting in the greatly reduced classification accuracy [14]. Therefore, before training, the random forest classifier first clusters the network intrusion data of power system, divides the network intrusion data set into several clusters, and trains a corresponding random forest classifier for each cluster,the specific process is as follows:

Firstly,defining $h(V_i, \Theta_i)$ representing a posterior density feature vector obtained by Gaussian mixture model clustering V_i and k independent distribution random vector $(\Theta_1, \Theta_1, \cdots, \Theta_k)$ having been produced the resulting decision tree.Power system network intrusion detection process is shown in Figure 2.

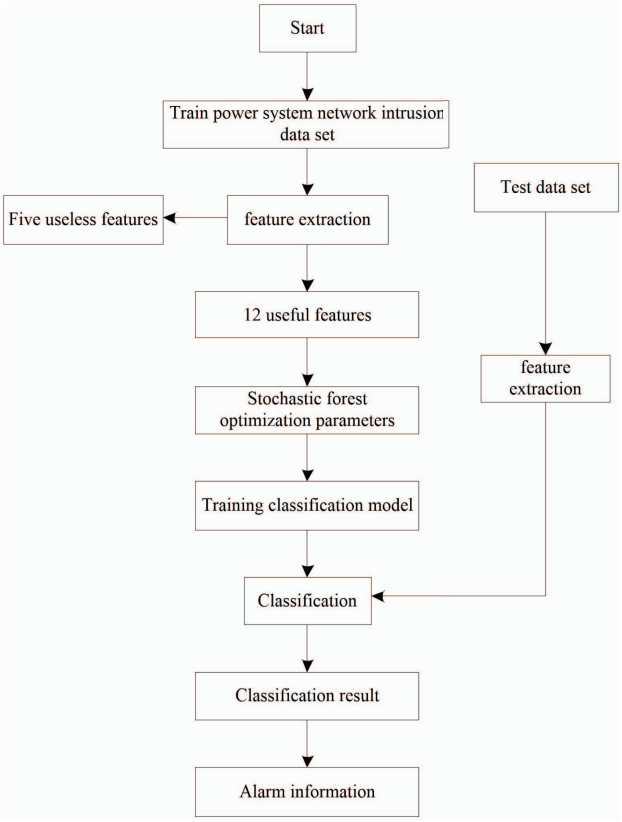

Figure 2 :network intrusion detection flow of power system

Power system network intrusion detection can be divided into four stages, the specific steps are as follows:

Step 1: extract network intrusion features of power system

As the information in the power system network is relatively complex, it is necessary to carefully select the characteristics of the network intrusion data of the power system, select some features that can be used, reveal the essential characteristics of the network intrusion data of the power system, reduce the dimension of the intrusion data, make the detection speed faster, and improve the efficiency of network intrusion detection [15];

Step 2: optimize the detection parameters

The random forest algorithm is used to construct decision tree continuously to find the optimal detection parameters and improve the classification effect;

Step 3: train the classification model

The power system network intrusion data is trained in the random forest algorithm to get the classification model of power system network intrusion detection;

Step 4: classification and recognition

Using the test set of power system network intrusion data for classification and identification, if there is power system network attack behavior, it will send out alarm information.

III. EXPERIMENTAL ANALYSIS

3.1 Describing the power system network intrusion data

Using the power system network intrusion detection database as training data set and test data set, the power system network intrusion detection database can not only reflect the flow structure of the power system network

and the external invasion situation, but also can modify and expand the network at any time. There are Normal、Probe、R2L、U2R and Dos five types of data, including 125793 training data and 22544 test data.

3.2 Processing network intrusion data in power system

At present, many researches divide the power system network intrusion detection database into 20% test data set and 80% training data set, but the power system network intrusion detection method based on random forest algorithm does not need to be split, because the method uses random forest classifier to train the power system network intrusion data, which can complete the decomposition of the intrusion data while training the internal data The intrusion data is classified into normal data and attack data. The attack data includes DOS, probe, r2l and u2r. The comparison of intrusion times of the two tags is shown in Table 1.

Table 1 Comparison of intrusion times of two tags

		Invasion times
Label 1	Normal data	67343
	Intrusion data	58630
Label 2	Dos	45927
	Probe	11656
	R2L	995
	U2R	52

Because there are character type and constants in the network intrusion detection database of power system, when extracting the features of network intrusion data of power system, the data features are normalized first,the formula is：

$$ual_{new} = \frac{val - V_{\min}}{V_{\max} - V_{\min}} \qquad (10)$$

ual represents the original value of network intrusion data in power system, and ual_{new} representing the data values after normalization, and $V_{\max}$ representing the maximum value in the network intrusion data of this kind of power system,and $V_{\min}$ representing the minimum value in the network intrusion data of this kind of power system.

3.3. Experimental process

Step 1: firstly, the network intrusion detection database of power system is imported and divided into 80% and 20% cross validation sets;

Step 2: the training set of network intrusion data of power system is extracted and divided into normal mode and intrusion mode. In the intrusion mode, it is subdivided into four types of intrusion modes, and each type of data set is converted into numerical intrusion data, and then normalized;

Step 3: Gaussian mixture clustering is used to process the training set of power system network intrusion data, and the maximum iteration times and values are set. After clustering, different types of clusters are obtained;

Step 4: train random forest classifiers with different types of clusters;

Step 5: import the test set of power system network intrusion data into the random forest trainer trained in step 4 to detect the network intrusion of power system.

3.4. Analysis of experimental results

Comparing the network intrusion detection method based on random forest algorithm with the network intrusion detection method in reference [4] and network intrusion detection method in reference [5], the comparison results of the accuracy rate and false alarm rate of the power system network intrusion detection are obtained, as shown in Table 2 and table 3.

Table 2 comparison results of power system network intrusion detection accuracy

Number of tests	Accuracy of network intrusion detection in power system %		
	Document [4] Network Intrusion Detection Methods	Document [5] Network Intrusion Detection Method	Network Intrusion Detection Method of Power System Based on Random Forest Algorithm
1	83.64	93.01	99.69
2	84.15	94.12	98.78
3	82.31	92.54	99.21
4	84.67	91.67	99.67
5	83.67	92.53	98.97
6	82.46	93.46	99.14
7	82.73	92.76	99.67
8	83.14	91.67	98.74
9	84.56	92.34	99.65
10	85.21	93.65	98.99

It can be seen from the test results in Table 2 that the power system network intrusion detection method based on random forest algorithm has the highest accuracy in detecting network intrusion of power system, followed by the network intrusion detection method in literature [5], and network intrusion detection method in document [4] has increased the experimental steps and reduced the power system due to the splitting of network intrusion detection database of power system System network intrusion detection accuracy.

Table 3 comparison results of false alarm rate of power system network intrusion detection

Number of tests	False alarm rate of network intrusion detection in power system%		
	Document [4] Network Intrusion Detection Methods	Document [5] Network Intrusion Detection Method	Network Intrusion Detection Method of Power System Based on Random Forest Algorithm
1	1.34	0.83	0.034
2	1.23	0.74	0.026
3	1.41	0.85	0.017
4	1.45	0.75	0.028
5	1.34	0.72	0.032
6	1.35	0.71	0.044
7	1.24	0.84	0.057
8	1.36	0.82	0.063
9	1.27	0.76	0.027
10	1.54	0.81	0.042

It can be seen from the results in Table 3 that the power system network intrusion detection method based on random forest algorithm has the greatest advantage in terms of false alarm rate. After calculation, the average false alarm rate in the test process is 0.037%, which is suitable for the intrusion detection of power system network.

IV. CONCLUSION

In this paper, the power system network intrusion detection method based on random forest algorithm is proposed. The random forest model is constructed by using the random forest decision tree. Before the state estimation of the power system network, the collected power system network measurement data is detected to realize the establishment of the power system network attack model Finally, through the experimental analysis, the feasibility of this method is verified.

REFERENCE

[1] Zong W , Chow Y W , Susilo W . Interactive three-dimensional visualization of network intrusion detection data for machine learning[J]. Future Generation Computer Systems, 2020, 102:292-306.

[2] Bhuyan M H , Bhattacharyya D K , Kalita J K . Towards Generating Real-life Datasets for Network Intrusion Detection[J]. International Journal of Network Security, 2015, 17(6):683-701.

[3] Tao M , Fen W , Jianjun C , et al. A Hybrid Spectral Clustering and Deep Neural Network Ensemble Algorithm for Intrusion Detection in Sensor Networks[J]. Sensors, 2016, 16(10):1701.

[4] Dai Yuanfei, Chen Xing, Chen Hong,et al. Feature selection based approach to network intrusion detection[J]. Application Research of Computers, 2017, 34(8): 2429-2433.

[5] ZHAO Xin, YE Mao, ZHU Ying-ying,et al. Novel network intrusion detection algorithm based on sequence data mining[J]. COMPUTER ENGINEERING AND APPLICATIONS, 2010, 46(5): 89-92,153.

[6] Sedjelmaci H , Senouci S M , Abu-Rgheff M A . An Efficient and Lightweight Intrusion Detection Mechanism for Service-Oriented Vehicular Networks[J]. IEEE Internet of Things Journal, 2017, 1(6):570-577.

[7] Ha T , Kim S , An N , et al. Suspicious Traffic Sampling for Intrusion Detection in Software-Defined Networks[J]. Computer Networks, 2016, 109(nov.9):172-182.

[8] Chen M H , Chang P C , Wu J L . A population-based incremental learning approach with artificial immune system for network intrusion detection[J]. Engineering Applications of Artificial Intelligence, 2016, 51(may):171-181.

[9] Zheng K , Cai Z , Zhang X , et al. Algorithms to speedup pattern matching for network intrusion detection systems[J]. Computer Communications, 2015, 62(may 15):47-58.

[10] Wang W , Sheng Y , Wang J , et al. HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection[J]. IEEE Access, 2018, 6(99):1792-1806.

[11] Riecker M , Biedermann S , El Bansarkhani R , et al. Lightweight Intrusion Detection in Wireless Sensor Networks[J]. International Journal of Information Security, 2015, 14(2):155-167.

[12] Aziz A S A , El-Ola Hanafi S , Hassanien A E . Comparison of classification techniques applied for network intrusion detection and classification[J]. Journal of Applied Logic, 2017, 24(pt.a):109-118.

[13] Saravanan K , Senthilkumar A . Security Enhancement in Distributed Networks Using Link-Based Mapping Scheme for Network Intrusion Detection with Enhanced Bloom Filter[J]. Wireless Personal Communications, 2015, 84(2):821-839.

[14] Zha Y , Li J . CMA: A Reconfigurable Complex Matching Accelerator for Wire-speed Network Intrusion Detection[J]. IEEE Computer Architecture Letters, 2018, PP(99):1-1.

[15] Zhang B C , Hu G Y , Zhou Z J , et al. Network Intrusion Detection Based on Directed Acyclic Graph and Belief Rule Base[J]. Etri Journal, 2017, 39(4):592-604.