Method for Detecting Anomaly Data of WAMS System Based on GA-iForest

Hongwen Yan¹, Jiawei Li*, Jian Zuo², Jihong Tang²
¹ College of Computer and Communication Engineering, Changsha University of Science & Technology, Changsha, Hunan ,410004, China
² State Grid Hunan Electric Power Company, Changsha, Hunan ,410007, China
*Corresponding author’s e-mail: csust_ljw@163.com

Abstract. Wide area measurement system problems such as anomalies, data missing, and real-time response deeply affect the operation and maintenance of the power grid. Considering the accuracy requirements of the system for anomaly detection, the detection accuracy and stability of traditional isolation forest are poor, so this paper proposes a new data anomaly detection method named GA-iForest. This method uses genetic algorithms to select isolated trees with high accuracy and obvious differences to optimize the structure of isolated forests. The new data anomaly detection method GA-iForest improves the accuracy and stability of detection, and mean while maintaining the high efficiency. The experiment uses standard simulation data sets and data from one province's wide-area measurement system for experimental simulation. The result shows that the GA-iForest method has significantly improved accuracy and stability compared with the traditional isolation forest, LOF, and K-means methods.

1. Introduction
The Wide Area Measurement System (WAMS) achieves high-speed data collection, but a variety of data problems are hidden in the rapidly increasing data volume, which caused a huge impact on operation and maintenance of the power grid [1]. The main purpose of the WAMS system to detect anomaly data is to improve the reliability of the collected data and reduce non-technical losses in the planning and operation of the power grid. The research on the data anomaly detection method has great significance to ensure the safe and reliable operation of the power grid system and improve the service level of the power grid [2-3]. In practical applications, high responsiveness of the WAMS system makes it difficult to obtain sufficient labelled training data despite the large amount of data exists, so "massive data and lack of information" is still an important issue for planning and operation of the power grid.

In the field of unsupervised learning, anomaly detection methods are mainly focused on some aspects, such as: distance-based [4], density-based [5], and model-based [6-7]. These methods distinguish outliers by segmenting outliers and other samples using the "outlier characteristics" of anomalous objects. However, With the huge amount of data in WAMS, the application of distance-based and density-based methods are greatly limited due to the complexity of calculation and have been gradually replaced by other methods.

In model-based anomaly detection methods which often uses a decision tree to distinguish anomalous data. Liu et al. [9] proposed an Isolation Forest (iForest) algorithm based on the structure
of the binary tree, by comparing with other algorithms, the model achieved good performance with linear time complexity. Although the characteristics of the isolated forest are suitable for the anomaly detection of the WAMS system, the artificial random factor brings the problem of insufficient accuracy and instability. Yu et al. [11] combined iForest with Local Outlier Factor (LOF), their method generated anomaly candidates by scanning the data set and then obtained anomaly results from the candidate dataset. The result shows the accuracy and stability have been improved, but the steps of the method are complicated and the execution efficiency is low. Aryal S et al. [12] proposed a method to improving iForest based on Relative Mass. This method calculates the path length of the samples to globally and locally sort the samples, which solves the problem of local cover-up, but increased the computational complexity and expending.

Isolation Forest avoids complex calculation processes, so it is more suitable for anomaly detection in WAMS systems. However, Isolation Forest still has some shortcomings:

1) The detection accuracy is related to each iTee in the Isolation Forest, a large number of uneven iTrees lead to insufficient detection accuracy of the model;

2) The introduction of random factors causes the detection stability of the forest to be affected by the randomly constructed iTee, which leads to the problem of poor detection stability of the model.

Aiming at the above shortcomings, this paper proposes a new anomaly detection method called GA-iForest. Specifically, the low time complexity of the iForest model is suitable for the large data volume and high rate requirements of the WAMS system. The problem of insufficient accuracy and instability of the iForest model are solved by optimize the forest structure through a newly defined genetic algorithm. The fitness function of the genetic algorithm are redefined combined with the characteristics of Isolation Forest. Based on the GA-iForest methods, this paper implements the detection of anomaly data in the WAMS system and provides a reliable foundation for advanced applications of big data in the power grid.

2. Anomaly detection method based on GA-iForest

2.1Anomaly data in WAMS

The data set used in this paper is derived from the circuit state information which collected by the PMU elements in each line in the WAMS system. These data include current, voltage, power, phase angle, and power angle generated during the operation of the power grid. As the physical side and the information side of the power grid is increasingly coupled with each other, any physical interference in the power grid system will cause disturbances in the state information of the circuit, resulting in fluctuations in the collected data. When there is a line or component failure in the operation of the power grid, a large data fluctuation change is displayed on the information side. This change is regarded as a signal that the system is anomaly, and the data with sharp fluctuations is regarded as anomaly data in the WAMS system.

When anomaly signals are generated, it is particularly important to identify anomaly data in the system. The system can quickly locate faulty lines or even locations by detecting anomaly data, so anomaly detection plays a very important role in power grid maintenance. The anomaly data in the WAMS system is characterized by anomaly fluctuations of the data. On the whole, the anomaly data of the WAMS system has a unique "outlier" characteristic. This characteristic can be used by the system to distinguish anomaly data.

2.2Isolation Forest algorithm

The core of the iForest algorithm is to use the characteristics of anomaly data: (1) the proportion of anomaly data is small; (2) the anomaly data has outlier characteristics, which is significantly different from normal data [8]. When performing anomaly detection on numerical data sets, iForest constructs a recursive binary tree iTee to spatially divide the sample set until iTee divides each sample into corresponding leaf nodes. Due to the outlier characteristic, the anomaly data is divided earlier and
closer to the root node, while normal data is farther away from the root node and has a larger path length.

Isolation Forest is composed of a large number of iTree decision trees, and realizes outlier detection by calculating the path length of each node. The decision tree T (iTree) is a binary tree structure, N is a node of T, if N is a leaf node, it is called an external node, if N is a non-leaf node, it is called an internal node; all samples are classified to the leaf node N. The length of the path is equal to the number of edges experienced by the root node to N, denoted as $h(d)$. Because iTree is based on a binary tree structure, the expected value of the sample search failure path length of iTree and binary tree are equal. The calculation formula of the Isolation Forest search failure path length is defined as:

$$H(i) = \ln(i) + \gamma, \quad \gamma \text{ is Euler's constant; } n \text{ is the number of leaf nodes; when it is determined, } C(n) \text{ is the average value of } h(d), \text{ which is used for standardization.}$$

After obtaining the path length of the data $d$ in each tree, calculate the anomaly score of $d$ according to its path length to determine whether the data is abnormal. The calculation formula of the anomaly score $S(d, n)$ is as follows:

$$S(d, n) = 2^{\frac{-E(h(d))}{C(n)}}$$

$E(h(d))$ is the average of the node depths of all iTree sets. It can be known from the formula that when $E(h(d))$ approaches $C(n)$, $S(d, n) \rightarrow 0.5$, which means that there is no obvious anomaly value in the data; when $E(h(d))$ approaches 0, $S(d, n) \rightarrow 1$, that is, the data is an abnormal value; when $E(h(d))$ approaches to $(n-1)$, $S(d, n) \rightarrow 0$. At this time, the data is likely to be normal.

2.3 Improvement of Isolation Forest based on genetic algorithm

2.3.1 Key points of using genetic algorithm to improve iForest

Genetic algorithms are widely used in the field of evolutionary calculus. The most common applications are in the problem of finding optimal solutions such as TSP, JSP, and combinatorial optimization problems [14-15]. In the search strategy using the genetic algorithm to optimize the model, the calculation does not rely on other auxiliary information. The key point is to determine the fitness value function of the search direction. The structure of the Isolation Forest model is similar to ensemble learning, the detection performance and generalization ability of the model will have "diversity and accuracy" evolutionary constraints when optimized by genetic algorithms.

Based on genetic algorithm, this paper proposes GA-iForest model, which can improve the accuracy and stability. This method realized the optimization of Isolation Forest by determining the appropriate fitness function.

2.3.2 Fitness function based on diversity and accuracy

The most important indicator of the genetic algorithm iteration process is the fitness value of different base classifier. The genetic process of the genetic algorithm is based on the fitness value of the single base classifier to evaluate the genetic probability, thereby determining the genetic order of each base classifier.

When the genetic algorithm is used to optimize the structure of iForest, we calculate two indicators of the iForest at the same time: diversity and accuracy. Diversity is the degree of inconsistency in the structure of different iTrees in the forest. Maintaining diversity is to avoid overfitting due to a large number of similar iTrees; Accuracy refers to the forest's detection ability, because different structure of the forest brings different detection results, and the iTree with high accuracy is retained to improve the overall detection accuracy of the forest. Diversity is the measurement of difference between different individuals, and common methods of measuring differences are based on statistical measures; accuracy is determined by evaluating the accuracy of the
overall model. Generally speaking, cross-validation can be used to obtain evaluate results, this paper calculates the individual difference and precision based on statistical methods. The determination process of the fitness function is as follows:

(1) First, the diversity of different iTrees is obtained by the k-statistic method, and the difference value is calculated. Given the training set \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \), for the task of anomaly detection two classification, \( y_i \in \{-1, +1\} \), the detection results of the trees \( T_i \) and \( T_j \) are shown in Table 1.

Table 1. Test results of \( T_i \) and \( T_j \).

| \( T_i = +1 \) | \( T_i = -1 \) |
|-----------------|-----------------|
| \( T_j = +1 \)  | \( a \)          | \( c \)          |
| \( T_j = -1 \)  | \( b \)          | \( d \)          |

In Table 1, \( a \) represents the number of samples that are positive \( T_i \) and \( T_j \), \( b \) represents positive \( T_i \) and negative \( T_j \), \( c \) represents positive \( T_j \) and negative \( T_i \), \( d \) represents negative \( T_i \) and \( T_j \). \( a + b + c + d = m \). Based on Table 1, the differences between iTrees are given:

\[
k = \frac{p_1 - p_2}{1 - p_2} \tag{3}
\]

In formula (3), \( p_1 \) is the probability that the classification results of the two trees are the same, and \( p_2 \) is the probability that the classification results of the two trees include inconsistent and consistent both, which can be obtained from Table 1:

\[
p_1 = \frac{a + d}{m} \tag{4}
\]

\[
p_2 = \frac{(a + b)(a + c) + (c + d)(b + d)}{m^2} \tag{5}
\]

The difference values \( k_{i,j} \) of \( T_i \) and \( T_j \) form a difference value matrix:

\[
K = \begin{bmatrix}
  k_{1,1} & k_{1,2} & \cdots & k_{1,L} \\
  k_{2,1} & k_{2,2} & \cdots & k_{2,L} \\
  \vdots & \vdots & \ddots & \vdots \\
  k_{L,1} & k_{L,2} & \cdots & k_{L,L}
\end{bmatrix}
\]

The \( k \) value at different positions indicates the degree of difference between different trees. A larger \( k \) value indicates a smaller difference between two trees and \( k \leq 1 \).

After obtaining the difference value matrix, we can get the difference measure of the entire forest:

\[
Q = \frac{(1-k_{i,1}) + (1-k_{i,2}) + \cdots + (1-k_{i,L}) + (1-k_{j,1}) + \cdots + (1-k_{j,L})}{IL - I} \tag{6}
\]

where \( Q \) is the difference between iForest, \( k_{i,j} \) is the difference between \( T_i \) and \( T_j \), and \((1-k_{i,j})\) is the degree of difference between the two trees. When the two trees are completely similar, the value is \( k_{i,j} = 1 \), so a larger \( Q \) indicates the inside of the forest. The greater the difference, the diversity of individuals is defined by the obtained difference metric \( Q \).

(2) Secondly, the accuracy value of each tree iTree is obtained through cross-validation. Cross-validation can avoid overfitting and ensure the generalization ability of the model. First divide the data set \( D \) into \( N \) mutually exclusive subsets of similar size. Each time use \( N-1 \) subsets for training, use 1 subset for testing, and use \( N \) subsets as training sets for training and testing one by
one. In this way, \( N \) test results can be obtained. \( N \) is usually taken as 10, and the average value of \( N \) metric values is taken as the precision value \( X_i \) of a single iTree.

After obtaining the accuracy value of a single iTree, you can get the accuracy measure of the entire forest:

\[
P = \frac{X_1 + X_2 + \cdots + X_n}{n}
\]

(7)

\( P \) represents the accuracy value of iForest, \( n \) represents the number of trees iTree in the forest, and the larger \( P \) the better accuracy of the forest.

(3) After obtaining the precision value and difference value of a single forest, genetic algorithm was used to calculate the fitness value of each forest model based on the "accuracy and diversity" principle from iForest. Its fitness function is as follows:

\[
F(f_i) = \frac{1}{W_1Q + W_2P}
\]

(8)

\( F(f_i) \) represents the fitness function of the forest \( f_i \), \( P \) represents the accuracy measure of the forest, and \( Q \) represents the \( f_i \) difference measure of the forest. \( W_1 \) and \( W_2 \) are weights determined according to experience. In this paper, \( W_1 = 0.8 \) and \( W_2 = 0.2 \). The fitness value of the individual is determined by the above formula. The smaller the fitness value is, the better the forest detection effect is.

2.4 Data anomaly detection based on GA-iForest

2.4.1 Basic Genetic algorithm process

Genetic algorithm is a classic evolutionary algorithm. The optimal individual is selected from the group through selection, crossover, and mutation steps [16]. The process of the standard genetic algorithm is as follows:

Step1: Initialize the group;
Step2: Calculate the fitness value of each initial individual in the population;
Step3: Perform hybridization operation according to probability \( cp \);
Step4: Perform mutation operation according to probability \( mp \);
Step5: Select operation based on individual fitness value and a preset rule to select the next generation of hereditary individuals;
Step6: If the termination condition is met, go to Step7, otherwise go to Step2;
Step7: Output the individual with the best fitness value in the population as a satisfactory solution to the problem.

2.4.2 GA-based iForest build process

The optimization of iForest is considered from the basic steps of GA. The most important ones include initialization, crossover, mutation and selection. Specific steps are as follows:

1) Construct a single individual iForest. The forest construction process is as described in Section 2.2. The iTree and iForest are constructed by continuously dividing the dataset space;
2) Group initialization. A single forest is considered as an individual. Each individual contains multiple iTrees as the "gene coding" of each individual. Each forest is a linked list. List \([i]\) represents the corresponding iTree. Constructing an initial group of \( N \) individuals \((N = 1,2,3,4)\), that is, the initial group is List1, List2, List3, and List4;
3) Calculate the fitness value of each individual, and calculate the fitness value of each List in the initial group according to formula (8);
4) Crossing. The \( N \) linked lists in the initial group cross each other according to the probability \( cp \), and the front and back half of each individual linked list are interchanged. Specifically, List1 \([1 \ldots i]\)
and List2 \([i + 1, \ldots, n]\) interchange, the same way List3 \([i]\) and List4 \([i]\); after crossing, the individual \(N\) in the initial group, from the initial \(N = 4\) to \(N = 8\) (List1, List2, List3, \ldots, List8);

5) Mutation. Mutually generated individual lists are mutated according to the probability \(m_p\), where List \([i]\) is an iTree binary tree corresponding to the forest, and a tree List \([i]\) is randomly selected for mutation operation. Specifically, the mutation The operation is to reconstruct the selected iTree;

6) Selection. The Forest List is selected according to the fitness value. The fitness function consists of precision and degree of difference. The selection method refers to the random ordering rule \([13]\). The original individual is selected, and the next generation of \(N = 4\) individuals is selected. If the termination condition is reached, the output is returned to step 3 or the next iteration is performed;

7) Termination condition and output. The termination condition is that when the fitness value of individuals in successive generations in the GA optimization process is higher than the traditional Isolation Forest model, the iteration is stopped or the maximum number of iterations is reached, and the output is the optimal individual in the current group List is the iForest with better fitness value.

3. Experimental results
In this paper, the standard simulation datasets breastw, shuttle, and the dataset wams from a wide area measurement system of one province are selected as the test dataset. Table 2 lists the statistical characteristics of each data set. The experimental simulation was run on a computer with 8GB of memory and a CUP of 2.3GHz. The default parameters for constructing the forest were set to: \(L = 100\), \(n = 256\), that is, 100 iTrees were constructed per forest, and the subsample size of the training iTree was 256.

| Data set | Number of samples | Number of features | Number of outliers |
|----------|-------------------|-------------------|-------------------|
| breastw  | 683               | 9                 | 239               |
| shuttle  | 49097             | 9                 | 3437              |
| wams     | 1000              | 14                | 350               |

The AUC value of each model can be used to draw its own ROC curves. The AUC value represents the probability that the predicted positive example is ranked before the negative example. The ROC curve can easily measure the performance of different models. Larger AUC values indicate better model performance \([17-18]\). The ROC curves of different data sets are shown in Figures 1-3.

Figure 1. Test results of the breastw dataset
Figure 2. Test results of the shuttle dataset

Figure 3. Test results of the wams dataset

Table 3. AUC values of different algorithms.

| Dataset | GA-iForest | Isolation Forest | LOF | K-means |
|---------|------------|------------------|-----|---------|
| breastw| 0.9937     | 0.9881           | 0.1309 | 0.7832 |
| shuttle| 0.9983     | 0.9971           | 0.5145 | 0.4958 |
| wams   | 0.9553     | 0.9233           | 0.5979 | 0.5477 |

It can be known from Figures 1 to 3 that the GA-iForest model is feasible and has better detection results than Isolation Forest, LOF, and K-means. It can be known from Table 3 that under the same data set, the optimized Isolation Forest has a higher AUC value and better model performance than the traditional Isolation Forest.

In order to compare the improvement of the accuracy of GA-iForest and Isolation Forest, we iterate 10 detection operations on the data set wams, and obtain the detection accuracy measures of the two models for the data set. The iterative detection results are shown in Table 4.

Table 4. Accuracy values for 10 iterations of dataset wams

| Iteration | GA-iForest | Isolation Forest |
|-----------|------------|------------------|
| round=1   | 0.9693     | 0.9444           |
| round=2   | 0.9678     | 0.9605           |
| round=3   | 0.9693     | 0.9546           |
It can be known from Table 4 that under the same conditions of the data set, the average accuracy of the isolated forest optimized based on genetic algorithm is improved by about 1.2% compared to traditional Isolation Forest, and the highest accuracy is improved by 2.1%, indicating that the genetic algorithm optimization has achieved accuracy improvement.

In order to compare the stability, a line chart is used to indicate the change trend and the degree of change. The trend graph of the 10 experiments results is shown in Figure 4.

![Figure 4. Comparison of the stability of the wams](image)

As shown in Figure 4, the degree of change of GA-iForest is smaller than that of Isolation Forest, which indicates that the fluctuation of the optimized model becomes smaller and the model stability improves. Repeated experiments were performed on the wams data set, the degree of change was recorded every 10 iterations, and the standard deviation was recorded every 10 times. The results are shown in Table 5.

| Round | GA-iForest | Isolation Forest |
|-------|------------|------------------|
| round=4 | 0.9634 | 0.9488 |
| round=5 | 0.9678 | 0.9634 |
| round=6 | 0.9678 | 0.9575 |
| round=7 | 0.9663 | 0.9605 |
| round=8 | 0.9722 | 0.9531 |
| round=9 | 0.9693 | 0.9619 |
| round=10 | 0.9707 | 0.9590 |
| Average | 0.9684 | 0.9564 |

From the standard deviation recorded in every 10 iterations in Table 5, we can see that GA-iForest has a lower standard deviation than iForest, so the stability of the model has improved.

4. Discussion
Aiming at the problem of insufficient accuracy and instability of Isolation Forest in WAMS system anomaly detection, this paper proposes a GA-iForest method. GA-iForest uses genetic algorithms to
optimize the structure of Isolation Forest, which improves the detection accuracy and stability of Isolation Forest. GA-iForest uses k-statistics and cross-validation to define the fitness function in the optimization model, and uses random ordering rules to constrain the evolutionary selection process. The experimental results show that the accuracy and stability of the GA-iForest algorithm are better than Isolation Forest, LOF, and K-means algorithms under the same experimental data set and experimental times.

4.1 Effectiveness analysis

As shown in the ROC curve of Figure 1-3, compared with other algorithms, GA-iForest has a larger area enclosed by the curve and the coordinate axis. In Table 3, the AUC value of GA-iForest is greater than Isolation Forest, LOF, and K-means. By comparing the ROC curve and the AUC value, we know that GA-iForest's model has the best performance. This is because GA-iForest and Isolation Forest have no difference in the forest topology. Neither of them requires complicated distance and density calculations, so the detection performance is better than LOF and K-means. At the same time, GA-iForest uses genetic algorithms to compare and screen iTrees in the forest. Through continuous crossover and mutation, the quality of iTrees is improved, and the model detection effect is improved.

From the ROC curve, we can see that for the dataset breastw, shuttle, and wams, the performance improvement of GA-iForest produced some different results. In these three data sets, the proportion of anomaly data in dataset breastw and wams is 35%, and 7% in dataset shuttle; breastw and shuttle had a dimension of 9, and wasm data had a dimension of 14.

It can be seen from Figure 1-3 that the improvement degree of the dataset breastw and wams is higher than that of the shuttle dataset. From the perspective of the abnormal distribution of the data set, the data set breastw and wams have a larger proportion of abnormal points than shuttle, which means that the abnormal distribution in the data set sample is more concentrated, showing a state of local and continuous abnormal aggregation. The traditional Isolation Forest is not sensitive to local anomalies. From the AUC value in Table 3, it can be seen that the detection performance of Isolation Forest for shuttle is higher than breastw and wams. GA-iForest solves this problem. Comparing the detection results of breastw and shuttle, the data dimensions of the two are the same, and the improved breastw detection effect is very close to shuttle. The reason why the Breastw and shuttle detection effects are close is that the optimization of the genetic algorithm makes the isolated forest better detect local and continuous anomalies. The mutation process of which genetic algorithm changes the structure of the iTree so that the abnormal objects in the local concentration are in the binary tree. It has a shorter search path and becomes easier to detect. From the perspective of the data dimension, the traditional Isolation Forest has a lower detection performance for high-dimensional data. As shown in Table 3, the AUC value of wams is much lower than breastw and shuttle. GA-iForest alleviates the performance degradation of Isolation Forest when processing high-dimensional data. The abnormal proportions of both wams and breastw are similar, but the performance improvement of wams is much higher than breastw, which means that GA-iForest detects high-dimensional anomalies. With a higher performance improvement, it alleviates the low performance of isolated forests for high-dimensional detection. The reason why the performance improvement of wams is much higher than that of breastw is that the constraint of GA-iForest's fitness function on iTree diversity makes the forest insensitive to dimensions. Therefore, each time when building an iTree, the increase of the data dimension will not decrease the performance of iTree.

Based on Figures 1-3 and Table 3, Isolation Forest has a good detection effect on the large data set shuttle which indicating that isolated forests are suitable for WAMS systems under big data condition. The GA-iForest proposed in this paper, while maintaining the original characteristics of Isolation Forest, solves the problem that Isolation Forest is not sensitive to local anomalies, and alleviates the decrease in the effect of the original model on high-dimensional data processing. The anomalous data generated by the WAMS system has the characteristics of high dimensionality and local continuity, and GA-iForest solves these problems well, and is more suitable for anomaly detection of the WAMS system.
4.2 Accuracy and stability analysis

Tables 4 and 5 record the detection accuracy values and standard deviation of GA-iForest and Isolation Forest. As shown in Table 4 and Table 5, under the condition of the same data set, the accuracy value obtained by GA-iForest through multiple iterations is higher than Isolation Forest, and the accuracy performance of GA-iForest is better than Isolation Forest. Take the average value of 10 iterations as one standard deviation which can be used as the metrics of stability of models. The results shown in Table 5 presents that the standard deviation of GA-iForest in multiple experiments is lower than Isolation Forest. The degree of curve fluctuation of the two models in Figure 4 also showed the same result, and GA-iForest showed better stability.

The accuracy and stability of GA-iForest perform well because the genetic algorithm fitness value function uses k-statistics and cross-validation to measure the diversity and accuracy of the forest, respectively. The k-statistics uses statistical methods to determine the diversity between different individuals and calculate the difference $Q$; cross-validation uses the leave-one-out method to calculate the accuracy of the individuals to obtain the precision value $P$. Weighting $Q$ and $P$, the obtained fitness value function constrains the search direction of Isolation Forest very well. Through continuous experiments to obtain the weight value, the determined fitness value function is finally obtained. The fitness function that combines diversity and accuracy on the one hand retains the diversity of iTrees and ensures the generalization ability of the model. On the other hand, the accuracy value of the model is eliminated by the elimination of iTrees with low accuracy.

The experiments verify the standard simulation data set and one province’s WAMS system data set and compare with the Isolation Forest, LOF, and Kmeans algorithms. The results show that GA-iForest's anomaly detection stability and detection accuracy rate are the best among all algorithms. This is because the genetic algorithm optimizes the internal structure of the forest, which enables the forest to obtain a "accuracy and diversity" performance based on the idea of ensemble learning, making the detection results of each iTree more excellent and compact.

5. Conclusion

This paper proposed a new algorithm GA-iForest to solve the inaccuracy and instability of Isolation Forest in WAMS system anomaly detection. The experiment simulated the anomaly detection on the standard simulation data set and one province’s wams data set, by comparing with the Isolation Forest, LOF, and Kmeans algorithms, the simulation results show that under the same conditions and data sets, The GA-iForest algorithm works best in terms of accuracy and stability of anomaly detection.

However, the GA-iForest model also has certain shortcomings that is the complexity of calculating the fitness value of the genetic algorithm. It takes a lot of time to calculate the fitness value of each iTree during the training process of GA-iForest, which affects the search time and efficiency of the final optimization result. Based on the independent characteristics of iTrees in the Isolation Forest, in future work, researchers will combine the framework of big data parallel calculations, put the fitness value calculation process of each iTree into a parallel framework, and obtain the fitness values of different iTrees and then performing selection. The optimization of the algorithm can greatly reduce the training time. After combining the big data parallel computing framework, the detection efficiency of finding the best individual will be further improved, which is also the direction and focus of our future research.

Author Contribution :
Investigation, Jian Zuo; Methodology, Jiawei Li ; Supervision,Hongwen Yan; Writing, Jiawei Li; Project Management,Jihong Tang.

Funding :
This work is funded by National Natural Science Foundation of China (No.51677007) and Technology project of state grid hunan power co., LTD. (No.5216A5170012)
Acknowledgments:
The author expresses his appreciation to the Hunan Electric Power Research Institute of the State Grid and thanks to the experts of the Electric Power Research Institute for their guidance.

References
[1] Gore, R., & Valsan, S. P. (2016, January). Big Data challenges in smart Grid IoT (WAMS) deployment. In:2016 8th International Conference on Communication Systems and Networks (COMSNETS). Bangalore, India. pp: 1-6. IEEE.
[2] ZHANG Dong-xia, MIAO Xin, LIU Li-ping, et al. (2015) Research on development strategy for smart grid big data. Proceedings of the CSEE, 35(1): 2-12.
[3] PENG Xiao-sheng, DENG Di-yuan, CHENG Shi-jie, et al. (2015) Key technologies of electric power big data and its application prospects in smart grid. Proceedings of the CSEE, 35(3): 503-511.
[4] Zoulikun, W. . (2005). Research on algorithms for mining distance-based outliers. Chinese Journal of Electronics, 14(3), 485-490.
[5] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In:Proceedings of the 2000 ACM SIGMOD international conference on Management of data. Dallas, Texas, USA. pp: 93-104.
[6] He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10), 1641-1650.
[7] Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), 212-223.
[8] Fei Tony Liu,Kai Ming Ting,Zhi-Hua Zhou. (2008) Isolation Forest. In:2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, pp: 413-422.
[9] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 1-39.
[10] Leandro Aparecido Passos Júnior, Cato César Oba Ramos, Rodrigues, D. , Pereira, D. R. , André Nunes de Souza, & Costa, K. A. P. D. , et al. (2016). Unsupervised non-technical losses identification through optimum-path forest. Electric Power Systems Research, 140, 413-423.
[11] Yu, X. , Tang, L. A. , & Han, J. . (2009). Filtering and Refinement: A Two-Stage Approach for Efficient and Effective Anomaly Detection. In:The Ninth IEEE International Conference on Data Mining. Miami, Florida, USA, pp:90-104
[12] Aryal, S., Ting, K. M., Wells, J. R., & Washio, T. (2014, May). Improving iforest with relative mass. In:Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taiwan. pp: 510-521.
[13] Runarsson T P , Yao X . (2002) Stochastic ranking for constrained evolutionary optimization. IEEE Transactions on Evolutionary Computation, 4(3): 284-294
[14] Sharma, P. , Wadhwa, A. , & Komal, K. . (2014). Analysis of selection schemes for solving an optimization problem in genetic algorithm. International Journal of Computer Applications, 93(11), 1-3
[15] Yu, Z., & Zhao, C. (2017, April). A Combination Forecasting Model of Extreme Learning Machine Based on Genetic Algorithm Optimization. In:2017 International Conference on Computing Intelligence and Information System (CIIS). Nanjing, China. pp: 29-32.
[16] Abu-Zitar, R. , & Nuseirat, A. F. . (2002). Performance evaluation of genetic algorithms and evolutionary programming in optimization and machine learning. Cybernetics and Systems, 33(3), 203-223
[17] Bradley A P. (1997) The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition, 30(7): 1145-1159.
[18] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
[19] Liu, J. M., Tian, J., Cai, Z. X., Zhou, Y., Luo, R. H., & Wang, R. R. (2017, July). A hybrid semi-supervised approach for financial fraud detection. In 2017 International Conference on Machine Learning and Cybernetics (ICMLC). Ningbo, China. pp: 217-222.

[20] Beghi, A., & McLoone, S. (2017, May). Anomaly detection through on-line isolation Forest: An application to plasma etching. In 2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC). Saratoga Springs, NY, USA, USA. pp: 89-94.