Anomaly detection of power grid dispatching platform based on Isolation Forest and K-means fusion algorithm

Liang Shouyu\textsuperscript{1}, Zhang Kun\textsuperscript{1}, Fang Wenchong\textsuperscript{1}, Zhou Zhifeng\textsuperscript{1}, Hu Rong\textsuperscript{1}, Zhu Wen\textsuperscript{1}, Li Yingchen\textsuperscript{1}, Wang Yichang\textsuperscript{2}, Hou Jian\textsuperscript{3}

\textsuperscript{1}China Southern Power Grid Dispatching Power Dispatching Control Center
\textsuperscript{2}Zhaotong Power Supply Bureau, Yunnan Power Grid
\textsuperscript{3}Nanning Power Supply Bureau, Guangxi Power Grid
shcorh@163.com

Abstract. It is very important to detect abnormal Access IP of power grid dispatching platform accurately and rapidly to ensure the safety of power production. An Isolation Forest and K-means fusion algorithm is proposed, which not only solves the deficiency of Isolation Forest only being able to detect anomalies in binary classification, but also solves the defect of Isolation Forest threshold setting based on artificial experience or prior assumptions by designing a threshold setting strategy for Isolation Forest anomalies. The Access IP data of the southern power grid dispatching platform is taken as an example to evaluate the proposed algorithm and model. Through control experiments, the effectiveness and advancedness of the algorithm are illustrated in terms of AUC-ROC curve, accuracy, recall, and F1 score.

1. Introduction
Power grid dispatching is an effective guarantee to ensure that all kinds of power production work are carried out in an orderly manner and to provide users with reliable power supply. It plays a decisive role in the safe operation of the power grid. The power grid dispatching platform is the basis for modernizing the automatic management of power grid dispatching. It is not only an important means to ensure the safety, stability, and economic operation of the power grid, but also an important infrastructure for the power system [1]. The combined operation of electricity and other components plays an important role in ensuring the safe, economic, stable and reliable operation of the power grid [2].

It is a great challenge for power grid dispatching platform to attack the server by making a large number of malicious requests to the server. Therefore, it is particularly important to detect potential intrusion and attack threats, realize visible and sensible business security, solve data leakage caused by network attacks, and track the identity of hackers through traceability services. The effective detection of abnormal access is an important guarantee for the safe operation of the power grid dispatching platform.

The current mainstream anomaly detection technologies mainly include: (1) Anomaly detection of boxplots based on the distribution of graphic positions, by obtaining the upper and lower quartiles of the boxplots [3], and then calculating the abnormal points, this method does not require artificial setting The upper threshold value avoids the problem of high-frequency modification, but lacks Accuracy and fewer abnormal points are identified in most cases; (2) Anomaly detection based on distance and density, including K-means clustering algorithm [4] and LOF density Location algorithm
These anomaly detection algorithms are simple in principle and easy to implement. However, they are often computationally inefficient and the accuracy of the detection effect is average. For example, in the K-means algorithm, the choice of the initial clustering center often has a great influence on the clustering effect of the final outlier, and the algorithm is stable. (3) At present, the mainstream of anomaly detection algorithm based on unsupervised model is the isolated forest algorithm [6], but the isolated forest algorithm can only detect whether the sample is abnormal, that is, it can only carry out binary classification processing, and to determine the threshold value of abnormal points often needs to be obtained through artificial experience or based on a priori assumption, that is, to get the estimated proportion of abnormal samples in the samples to be predicted. Also, there are many others algorithm, such as based on Time Series, LSTM and Fourier [7] [8] [9].

2. Isolation Forest algorithm analysis

Isolation Forest belongs to non-parametric and unsupervised methods, that is, no mathematical model is defined and no marked training is required. Isolation Forest is an integrated algorithm composed of multiple I-tree. Each Isolation Tree is a binary Tree structure because it is divided into dichotomies every time [6]. Isolation Tree makes use of the characteristics of small number and large difference of anomalies to compare the abnormal point with the normal point and isolate it near the root of the tree. This unique feature enables the algorithm to construct an effective model with only a small proportion of training data. \( h(\mathbf{x}) \) is the path length of sample \( \mathbf{x} \), specifically the number of edges of sample \( \mathbf{x} \) traversing an iTree until the traversal ends at the leaf node. Given a data set containing \( n \) samples, the average path length of the tree \( c(n) \) is:

\[ c(n) = 2H(n-1) - (2(n-1)/n) \]  

(1)

Where \( H(i) \) is a function, \( \ln(i) + 0.5772156649 \).

Let the abnormal score of sample \( \mathbf{x} \) be \( s(\mathbf{x}, n) \), then:

\[ s(\mathbf{x}, n) = 2^{-\frac{E(h(\mathbf{x}))}{c(n)}} \]  

(2)

Where \( E(h(\mathbf{x})) \) is the average value of \( h(\mathbf{x}) \) of all iTree traversed by sample \( \mathbf{x} \).

- when \( E(h(\mathbf{x})) \rightarrow c(n) \), \( s \rightarrow 0.5 \)
- when \( E(h(\mathbf{x})) \rightarrow 0 \), \( s \rightarrow 1 \)
- when \( E(h(\mathbf{x})) \rightarrow n-1 \), \( s \rightarrow 0 \)

The performance of Isolation Forest in AUC and execution time is better than LOF and RF, which are based on time complexity. In addition, Isolation Forest has a fast convergence rate and a small integration scale, which can effectively detect anomalies. Therefore, Isolation Forest has linear time complexity and low memory requirements, which is an ideal algorithm for prediction of large data sets. However, compared with the anomaly detection algorithm based on distance and density, although Isolation Forest has a great improvement in AUC and detection speed, this algorithm still has the following two problems: (1) it can only judge whether the sample is an abnormal sample, that is, it is only valid for dichotomy; (2) it is difficult to determine the threshold of outliers.

3. Isolation Forest and K-means fusion algorithm

Aiming at the above problems, an Isolation Forest and K-means fusion algorithm is proposed for the power grid dispatching platform. The specific work is as follows:

According to the data types and business characteristics of the power grid platform, IP access types are set as ‘Normal’, ‘Suspected’ and ‘Abnormal’. Firstly, the Isolation Forest algorithm was used to calculate the abnormal scores of each sample, and the corresponding abnormal scores were saved as new input data and entered into the next task stage for calculation.

Secondly, this paper uses the k-means algorithm to cluster the abnormal scores, and uses the euclidean distance as the similarity evaluation standard. The formula is as follows.

\[ d(\nu(x), c) = \sqrt{(\nu(x)_i - c_i)^2 + (\nu(x)_j - c_j)^2} \]  

(3)
Where \( d(v(x), c) \) is euclidean distances between the abnormal score \( v(x) \) of sample \( x \) and the center \( c \) of the cluster; \( v(x)_i \) and \( v(x)_j \) are respectively the values of the abnormal score \( v(x) \) in the \( i \)-th and \( j \)-th dimensions; \( c_i \) and \( c_j \) are respectively the values of the center \( c \) in the \( i \)-th and \( j \)-th dimensions.

Finally, in order to avoid the subjectivity of the artificial threshold and adopting the tactics of anomaly score according to the samples distribution, setting thresholds for specific design strategy is: if the exception score distribution is right skewed distribution, because the right skewed distribution has a longer tail on the right side of the mean and the median is greater than the number of features, the abnormal scores in logarithm, the anomaly threshold initial clustering center is the maximum, minimum and average respectively; If the distribution of fractions is left skewed, the distribution of left skewed has the opposite characteristic to the right skewed distribution, so the cube root of the anomaly scores can be taken. At this time, the initial clustering center of the anomaly is the maximum, the minimum and the average respectively.

Fig 1. A fig with distribution of right skewed, symmetrical and left skewed.

### Isolation Forest and K-means fusion algorithm

| Input: | Abnormal scores for all samples of Isolation Forest |
|--------|----------------------------------------------------|
| Process: | Abnormal scores for all samples of Isolation Forest |
| 1 Step1: | If abnormal scores is right skewed distribution: |
| 2 | \( \text{new abnormal scores} = \text{cube root (abnormal scores)} \) |
| 3 | center0 = max(new abnormal scores) |
| 4 | center1 = mean(new abnormal scores) |
| 5 | center2 = min(new abnormal scores) |
| 6 | If abnormal scores is Symetric distribution: |
| 7 | center0 = max(abnormal scores) |
| 8 | center1 = mean(abnormal scores) |
| 9 | center2 = min(abnormal scores) |
| 10 | If abnormal scores is left skewed distribution: |
| 11 | \( \text{new abnormal scores} = \text{cube root (abnormal scores)} \) |
| 12 | center0 = max(new abnormal scores) |
| 13 | center1 = mean(new abnormal scores) |
| 14 | center2 = min(new abnormal scores) |
| 15 Step2: | While K-means not converging: |
| 16 | For abnormal scores of all samples - 3: |
| 17 | calculate which class each point belongs to. |
| 18 | For center0,center1,center2: |
| 19 | Find out all abnormal scores points belongng to this category |
| 20 | Modify center0,center1,center2 coordinates to the center coordinates of these data points |
| 21 End | End |
| Output: | All sample labels |

The proposed algorithm architecture is shown in the following figure.
4. Experimental results and analysis
This section mainly studies the detection effect of the improved Isolation Forest algorithm on multiple types of abnormal access in the power grid dispatching platform. The data set is composed of the daily request IP access information data collected by the power grid dispatching platform. Specifically, the original data is firstly preprocessed to eliminate the features with the proportion of missing value higher than a certain threshold, and the features with strong correlation are deleted by calculating the correlation coefficient between the features. Secondly, the features are further processed, including constructing time-series difference features by using sliding Windows, constructing discrete features corresponding to several continuous features by dividing boxes, and constructing statistical features such as maximum value, minimum value, mean value, variance and frequency. There are 40190 pieces of available training data processed by the above data, with 104 features, including 103 input features and 1 output feature, including normal access, suspected abnormal access and abnormal access. The processed data is shown in the following table.

| Sample type       | Normal | Suspected | Anomaly |
|-------------------|--------|-----------|---------|
| Sample size       | 29710  | 6420      | 4060    |

The confusion matrix can be used to measure the classification effect of the model. TP is the real example, FP is the false positive example, and FN is the false negative example. The composition of the confusion matrix is shown in table 2 below:

| Prediction | Positive | Negative |
|------------|----------|----------|
| Truth      | TP       | FP       |
| Negative   | FN       | TN       |

To verify the performance of the algorithm, the following performance indicators are selected: Precision, the Accuracy rate indicates how much the model predicted correctly in all the classifications. The calculation formula is as follows:

\[
\text{Precision} = \frac{TP}{TP+FP} \tag{4}
\]

Recall. Recall represents how many of the positive example classes are correctly predicted by the model. The calculation formula is as follows:

\[
\text{Recall} = \frac{TP}{TP+FN} \tag{5}
\]

F1_score, The index of F1_score measures both the Accuracy and Recall, so the index of F1_score is also measured. The calculation formula of F1_score is expressed as:

\[
F1\_score = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{6}
\]

AUC-ROC. The AUC-ROC curve is a performance measure of classification problems under different threshold Settings. ROC is a probability curve, and AUC represents the degree or measure of separability. The full name of AUC is area under the curve, that is, the area under the ROC curve. It shows how well the model can distinguish between categories. The higher the AUC, the higher the ability of the model to distinguish the types of abnormal access. Specifically, the closer the AUC is to 1, which means that it has good separability, and the closer the AUC is to 0, which means that it has the worst separability. When the AUC is 0.5, it means that the model has no class separation ability. Compared with the evaluation indexes that depend on the judgment threshold, such as the precision
rate, Recall and F1_score, AUC has no such problem. AUC is also insensitive to the proportion of positive and negative samples. ROC curve is drawn by TPR and FPR, where TPR is on the Y-axis and FPR is on the X-axis. The calculation formulas are as follows:

\[
FPR = \frac{FP}{TN+FP} \tag{7}
\]

\[
TPR = Recall = \frac{TP}{TP+TN} \tag{8}
\]

4.1. Experimental design

In order to verify the effectiveness of the algorithm proposed in this paper, a comparison test is designed here with the k-means anomaly detection algorithm for verification. Since the ROC curve is usually used to measure the classification effect of the classifier in dichotomies, in order to expand the ROC curve and the ROC area to multiple categories, the output needs to be binarized and a ROC curve is drawn for each category. In addition, by giving equal weight to each category, a comprehensive macro average ROC curve can be drawn for all categories. Meanwhile, the Precision, Recall and F1-score of the two algorithms in each category were calculated and verified by comparison.

4.2. Experimental results

In order to verify the effectiveness of the algorithm proposed in this paper, a comparison test is designed here with the k-means anomaly detection algorithm for verification. Firstly, the Isolation Forest algorithm was used to score the abnormal samples for the sample data. Secondly, observe the abnormal score distribution of all samples. If it is positive skewed distribution or negative skewed distribution, calculate the cube root of the score and modify it to conform to the normal distribution. Finally, the maximum, mean and minimum values of the abnormal score data after the modified distribution were taken as the initial clustering center of the three clusters, and the abnormal score data were clustered, so as to classify the normal access, suspected abnormal access and abnormal Access IP of the dispatching platform, and complete the algorithm identification process.

![Fig 3. The ROC curve of the Isolation Forest and K-means fusion algorithm.](image)

![Fig 4. The ROC curve of the K-means algorithm](image)

Specifically, the corresponding evaluation indexes of the two algorithms for accessing IP desensitization data sets on the dispatching platform of China southern power grid are shown in the following table:
Table 3. Isolation Forest and K-means fusion algorithm is compared with the evaluation indexes of k-means algorithm.

| Metric   | Class   | Isolation Forest and K-means fusion algorithm | K-means |
|----------|---------|-----------------------------------------------|---------|
|          | Normal  | 0.97                                          | 0.87    |
|          | Suspected| 0.94                                          | 0.76    |
|          | Anomaly | 0.83                                          | 0.81    |
| AUC      | Normal  | 0.99                                          | 0.98    |
|          | Suspected| 0.78                                          | 0.38    |
|          | Anomaly | 1.00                                          | 0.69    |
| Precision| Normal  | 0.99                                          | 0.79    |
|          | Suspected| 0.94                                          | 0.74    |
|          | Anomaly | 0.67                                          | 0.65    |
| Recall   | Normal  | 0.99                                          | 0.87    |
|          | Suspected| 0.85                                          | 0.51    |
|          | Anomaly | 0.80                                          | 0.67    |

It can be observed from the ROC curve that the ROC curve of the Isolation Forest and K-means fusion algorithm in three categories is higher than the performance of k-means algorithm. Specifically, the performance is most obvious in categories 'Normal' and 'Suspected', while the effect is still improved by 0.02 units in the category 'Anomaly'. On the other indicators, the improved algorithm based on Isolation Forest is also completely superior to the k-means algorithm.

According to the above experimental results, compared with k-means algorithm, the Isolation Forest and K-means fusion algorithm has the best classification effect on abnormal access detection of the dispatching platform of China southern power grid.

5. Conclusion
In this paper, based on the China Southern Power Grid dispatching platform, based on the isolated forest, an improved algorithm for detecting abnormal access to the isolated forest is proposed. At the same time, as a starting point, the threshold setting strategy of anomalous samples is optimized so that it does not depend on human experience, which improves the accuracy and stability of the anomaly detection algorithm.

References
[1] W. Sun et al., Extended application of online security and stability analysis system in Liaoning power grid, 2016 China International Conference on Electricity Distribution (CICED), Xi'an, 2016, pp. 1-5.
[2] C. Wang and C. Wang, The Study of the Existing Problems and Processing Measures Based on Power Grid Dispatching and Safe Operation[C]/ 2018 Chinese Control And Decision Conference (CCDC), Shenyang, 2018, pp. 5194-5198.
[3] G. Shevlyakov, K. Andrea, L. Choudur, P. Smirnov, A. Ulanov and N. Vassilieva, Robust versions of the Tukey boxplot with their application to detection of outliers[C]/ 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 6506-6510.
[4] S. N. Lighari and D. M. A. Hussain, Testing of algorithms for anomaly detection in Big data using apache spark[C]/ 2017 9th International Conference on Computational Intelligence and Communication Networks (CICIN), Gime, 2017, pp. 97-100.
[5] Z. Gan and X. Zhou, Abnormal Network Traffic Detection Based on Improved LOF Algorithm, [C]/2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 2018, pp. 142-145.

[6] F. T. Liu, K. M. Ting and Z. Zhou, Isolation Forest[C]/ 2008 Eighth IEEE International Conference on Data Mining, Pisa, 2008, pp. 413-422.

[7] L. Zhang, X. Shen, F. Zhang, M. Ren, B. Ge and B. Li, Anomaly Detection for Power Grid Based on Time Series Model[C]/ 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), New York, NY, USA, 2019, pp. 188-192.

[8] X. Wang, T. Zhao, H. Liu and R. He, Power Consumption Predicting and Anomaly Detection Based on Long Short-Term Memory Neural Network[C]/ 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2019, pp. 487-491.

[9] H. Zhao et al., A Fourier Series-Based Anomaly Extraction Approach to Access Network Traffic in Power Telecommunications[C]/ 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), Dalian, 2017, pp. 550-553.