Recognition method of wide area measurement abnormal data based on GRA-IForest

Xing Zhou¹, Zhi Yong², Liu Wenfei², Li Jiawei³, Qing Can³, Ma Rui³

¹State Grid Gansu Electric Power Company, Lanzhou 730030, Gansu, China.
²State Grid Gansu Electric Power Corporation Research Institute, Lanzhou 730070, Gansu, China.
³Electrical and Informational Engineering College, Changsha University of Science and Technology, Changsha 410114, Hunan, China

Abstract: Aiming at the anomaly recognition method of large area measurement system (WAMS), a method based on Grey Relational Analysis and Isolation Forest algorithm is proposed, considering the attributes of power data and the big data characteristics of large area measurement system. In this method, the relational matrix of power data attribute sequence is obtained through grey relational analysis, and the dependent attributes are selected according to the relational matrix for attribute sequence merging. Then, the data set is partitioned by cascaded isolated forest model and abnormal data are detected. Taking the data of a province's wide-area measurement system as an example for experimental simulation, compared with LOF, Kmeans and IForest, under the same inspection time, the GRA-IForest model has achieved good inspection results and achieved a higher overall detection accuracy and lower standard deviation.

1. Introduction
In recent years, with the development of Internet of things, smart grid[1-2] and energy Internet[3-4], the coupling of physical side and information side in power system is getting closer, as a result, the system is extremely vulnerable to the adverse effects of abnormal data on the information side, however, some of the original anomaly detection methods[5-6] have some disadvantages, such as low accuracy, poor flexibility and complex calculation. At present, the mainstream power data anomaly detection algorithms KNN, K-Means, DBSCAN and LOF are not ideal for high-dimensional data[7-8] and real-time power big data, and their online application capabilities are not strong. Liu and his colleagues[9-10] proposed an isolated forest (Isolation Forest) algorithm based on the characteristics of random partitioning. The time complexity of the algorithm is very low, and the detection effect is good. In literature [11] and [12], IForestASD algorithm is proposed, which can effectively detect the abnormal cases of flow data by using the frame of sliding window to process flow data. Yu and his colleagues[13-15] combined LOF and IForest to generate a set of abnormal candidate sets through the word scanning data set, and then detect the candidate sets to get the abnormal results.

However, in solving the problem of power data, a single IForest has some limitations, which leads to the problem of poor stability, and there is also room for improvement of accuracy, which is mainly reflected in the characteristics of power data attribute dependence.

Under the premise of fully considering the characteristics of power data, this paper adopts a new anomaly detection method GRA-IForest. This method uses Grey Relational Analysis (Grey Mrrelation...
Aru) to obtain the dependency between power data attributes. By considering the dependency relationship to optimize the forest construction process of IForest, outliers can be divided faster, making the algorithm more stable, the detection accuracy and performance has been improved.

2. Isolation Forest algorithm and improvement

2.1. Isolation Forest algorithm
The core problem of IForest is the method of separating data (see Fig.1), the specific process is as follows:

1) Select sample set: Randomly select $\varphi$ sample data points from a province's wide area measurement system data training set as the training data of a single iTree to construct an iTree;
2) Select partition attributes and points: In the process of constructing iTree, randomly select a dimension $w$ from the corresponding training data, generate a cut point in this dimension $d \in (\min(w), \max(w))$;
3) Divide left and right subtrees: Use $d$ as a comparison value to divide the node data of all dimensions. Samples smaller than $d$ are divided into left child nodes, and samples larger than or equal to $d$ are divided into right child nodes;
4) Repeat steps 2) and 3), until the node is indivisible (only one data item or the maximum number of cuts is reached).

![Fig.1 The process of partitioning data in iTree](image)

It can be seen from Fig.1 that d4 is most likely to be anomalous because it is divided very early. By constructing N iTrees with the structure shown in Fig.1, the forest structure of IForest anomaly detection is finally formed, the feature values in the data set calculate the corresponding anomaly score by traversing the forest, the calculation formula is as follows:

$$C(n) = 2H(n-1) - (2(n-1)/n)$$  \hspace{1cm} (1)

Among them, $H(i) = \ln(i) + \xi$, $\xi$ is Euler constant; n is the number of leaf nodes.

$$S(x, n) = 2 \frac{E(h(x))}{C(n)}$$  \hspace{1cm} (2)

The obtained result $S(x, n)$ is the abnormal score of each feature data, and it is determined whether it is abnormal according to the area where the abnormal score is located.

However, the traditional isolated forest randomly selects an attribute to divide the sample space at a time. In the face of highly correlated data between attributes, there are problems of poor stability and slow efficiency, see Fig.2 for details:
Fig. 2 IForest single property selection process

As shown in Fig. 2, assume that the attributes a3, a4, a5 are the attribute sequence where the abnormal data is located, the traditional IForest randomly selects an attribute to divide the sample space each time, and the number of steps and positions to complete the division are polarized. However, when we consider merging highly correlated attributes and dividing them at the same time, we will avoid this extreme nature:

Fig. 3 IForest merge property selection process

As shown in Fig. 3, when we merge highly related attributes, each time one of the attributes is selected, other related attributes are also selected, and then divided at the same time. The polarization that this division avoids, makes the model more stable, it also reduces the division steps, ability to find the location of abnormal data in fewer steps.

Compared with the traditional IForest, after considering attribute dependence, IForest can better reduce IForest's inherent "differentiation" during the process of attribute selection and division process, while reducing attribute selection and division steps after selection.

2.2. GRA-Isolation Forest algorithm

There is a dependency relationship between the attributes of power data. When processing power data, the concept of "multiple" is introduced. It is intended to treat multiple attributes with high dependencies as one operation object, that is, multiple attributes. The other attributes are regarded as independent attributes. In the process of constructing iTree, if the selected attribute is an independent group, the division method is carried out according to the original method; if the attribute belongs to multiple reorganization, all attributes of the multiple reorganization are treated as one attribute. However, considering the strength and the physical significance of the attribute correlation, merging too many attributes will result in unsatisfactory model effects. Therefore, the attributes suitable for merging can be obtained by obtaining the gray correlation degree between the attributes.

2.2.1. Attribute merging based on gray correlation

1) Get attribute relevance
Use the method of gray correlation analysis to obtain the correlation degree hidden among the attributes of power data, mainly through the similarity of the curves, the calculation steps are as follows:

① Construct a sequence matrix. Based on all attribute characteristic data of historically correct data set, create the initial attribute sequence matrix $D = [D_1, D_2, \ldots, D_m]$, where $m$ is the number of attributes and $N$ is the number of historical data collection points. In this paper, the data of 1000 collection points is taken.

$$D = \begin{bmatrix} D_1(1) & \cdots & D_1(m) \\ D_2(1) & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ D_N(1) & \cdots & D_N(m) \end{bmatrix}$$  \hspace{1cm} (3)

② Nondimensionalization. Use the mean method to eliminate the dimensional effects of different dimensions. Using formula (4), a dimensionless matrix can be obtained, record as

$$D'(k) = D_i(k) / m(\text{sum}_k), \quad i = 1 \cdots N, k = 1 \cdots m$$  \hspace{1cm} (4)

$m(\text{sum}_k)$ is the average of column $K$.

③ Calculate the correlation coefficient.

$$\xi_{pq}(i) = \frac{\min\min[D(p) - D(q)] + 0\max\max[D(p) - D(q)]}{|D(p) - D(q)| + 0\max\max[D(p) - D(q)]}$$  \hspace{1cm} (5)

among them: $p$ and $q$ are the serial numbers of attributes; $\theta$ is the resolution coefficient, generally $0.5$; $D(p)$ and $D(q)$ take the features of column $p$ and column $q$ of dimensionless row $i$, respectively; $i = 1, 2, \cdots, N$; $\xi_{pq}(i)$ is the correlation coefficient of columns $p$ and $q$ in row $i$.

④ Calculate the degree of association.

$$r_{pq} = \frac{1}{N-1} \sum_{i=1}^{N} \xi_{pq}(i), i = 1, 2, \cdots, N$$  \hspace{1cm} (6)

The $r_{pq}$ is the correlation degree of column $p$ and column $q$.

⑤ Relevance ranking.

After obtaining the correlation matrix between the attributes, the order of the correlation degrees is used to describe the degree of association between the attributes. The $m$ attributes are used as the parent sequence, and the other $m-1$ attributes are used as the child sequence. The same parent sequence is arranged according to the degree of relevance, forming the relevance order, and recorded as $\{d\}$. Correlation order reflects the "good and bad" relationship of other attributes for a certain attribute. If $r_{so} > r_{ot}$, then it means that for the same parent sequence $s$, the subsequence $D_N(s)$ is better than the subsequence $D_N(t)$, in which the attribute $s$ is in front of the attribute $t$.

2) Multiplicity selection based on gray correlation

After getting the correlation matrix, we can determine the independent attribute group of multiple attribute combination. In the relevance matrix, when the relevance ranking of one attribute and another attribute is the first, if the relevance ranking of another attribute is also the first, then these two attributes are considered as multiple attribute groups and can be combined. Otherwise, we will continue to find the corresponding sorting position of the next attribute in sequence. For consideration of computational complexity, the maximum multiplicity of this paper is triple, that is, there are at most three attributes in the combined attribute group.
2.2.2. Isolation Forest based on multiple attributes

After Isolation Forest obtains the data set of merged attributes, it randomly selects one attribute to divide, and if it is an independent group, it divides the data space according to the original way; If it's a multiple combination, two merged attributes A and B, Select to divide the cutting point A = a, B = b, if the trend of characteristic curve of A and B is positively correlated, the data space on the same side of a and B (the same size) will be cut to the left subtree, and the rest to the right subtree; If A is negatively correlated with B, the same is true. When the merged group is a triple attribute, the judgment method is the same, that is, the cutting point is A= A, B= B, C= C, and the trend correlation between A, B and C is judged at the same time, and the data space is divided according to the feature size comparison of the three attributes.

Through the above steps to divide iTree, you can divide the data that the attribute depends on into the same space when you divide the space, so as to reduce the steps to separate the abnormal data.

3. GRA-IForest algorithm flow

The specific steps of the method of combining gray correlation analysis with isolated forest are as follows:

1) Select reference sequence $c_k$ and comparison sequence $c_p$: select one attribute $D_i'(p)$ as $c_k$ from the sequence matrix, and put the rest attributes $D_i'(q)$ into $c_p$;

2) Calculate the correlation degree $r_{pq}$;

3) Repeat steps 1 and 2 until all attributes are used as reference objects to obtain the correlation matrix of the attribute sequence set;

4) Select the attributes to be merged according to the correlation matrix;

5) Establish iTree and build anomaly detection model;

Construct the forest in the manner proposed in section 1.2.2, based on which the anomaly detection model based on GRA-IForest can be established, the flow chart is shown in Fig.4:

![Fig.4 Flow chart for GRA-IForest](image-url)
4. Example analysis

In this section, we first obtain the attribute correlation matrix of the data set through the GRA experiment, select multiple attributes, and select the correlation degree of the combined attributes from the correlation matrix, as shown in Table 1.

Secondly, the accuracy and stability of GRA-Iforest, Isolation Forest, LOF and K-means algorithms are compared through experiments.

Through the analysis and test of Iforest, we can know that the calculation time is in direct proportion to the number of isolation tree iTree construction, and the accuracy of classification increases gradually after iTree exceeds a certain threshold, which is close to 0. After the test, the value of iTree is 100, which is the best. At the same time, the size of training subset also has a threshold to improve the speed and accuracy of IForest. Considering the characteristics of binary tree and computing time, 256 samples of training subset are the best. The ROC curve of the test data set is shown in Fig.5 to Fig.7. Obtain the AUC value of each model through the ROC curve. The AUC values of different algorithms in the data set are shown in Table 2.

| Degree of correlation | atr1 | atr7 | atr9 | atr10 | atr11 |
|-----------------------|------|------|------|-------|-------|
| atr1                  | \    | 0.995| \    | \     | \     |
| atr7                  | 0.995| \    | \    | \     | \     |
| atr9                  | \    | \    | \    | 0.998 | 0.999 |
| atr10                 | \    | \    | 0.998| \     | 0.999 |
| atr11                 | \    | \    | \    | \     | 0.999 |

According to the table, atr1 and atr7 are multiple reorganizations, atr9, atr10, and atr11 are another multiple reorganizations, the correlation between other attributes is relatively low or according to the multiple of gray correlation in Section 1.2.1, treated as an independent group.

![Fig.5 Test result r1 for dataset](image)

AUC
Fig. 6 Test result r2 for dataset

Fig. 7 Test result r3 for dataset

Table 2 Different algorithms detect corresponding AUC values on data sets

| Results | GRA-IForest | IForest | LOF | Kmeans |
|---------|-------------|---------|-----|--------|
| r1      | 0.889       | 0.881   | 0.597 | 0.465 |
| r2      | 0.895       | 0.890   | 0.598 | 0.465 |
| r3      | 0.897       | 0.892   | 0.613 | 0.473 |

From Fig. 5 to Fig. 7 and Table 2, under the same data set, the overall results of GRA-IForest showed better detection accuracy, this shows that GRA-IForest shows better anomaly detection effect than traditional IForest under the condition of considering attribute dependence.

Fig. 8 Stability comparison of test results of data set
Table 3 Standard deviation for GRA-IForest and IForest

|     | GRA-IForest | IForest |
|-----|-------------|---------|
| r1  | 0.017       | 0.017   |
| r2  | 0.013       | 0.016   |
| r3  | 0.0107      | 0.0129  |

Figure 8 shows the smoothness of the curve and the standard deviation of the results of 10 iterations, which shows that GRA-IForest has better stability.

From the experimental results, we can know that the overall results of GRA-IForest are the best in stability and detection accuracy, and the structure is simple to implement when detecting large data anomaly in power system wide-area measurement. Considering the attribute dependence between power data, GRA-IForest is better adapted to the data characteristics of the power system.

5. Conclusion
This paper discusses various data anomaly detection methods in the power system. Considering the characteristics of large-scale measurement of large data in the power system, a new anomaly detection method GRA-IForest is adopted. Through the experimental analysis of the wide-area measurement data set of the power system, the overall performance of the GRA-IForest algorithm in terms of accuracy and stability is better, and it is more suitable for the characteristics of power data.

Of course, the model in this paper also has certain limitations. Due to the randomness of the forest itself, there will be cases where the model performs poorly, mainly because the choice of attributes is random. Secondly, the structure of IForest also has some room for improvement, mainly manifested in the problem of uneven accuracy and high memory consumption caused by the difference of a large number of iTree in IForest. Reducing the number of iTree and improving the quality of iTree is the focus of future research.

Acknowledgment
This work was supported by the Science and Technology Project of State Grid Corporation (Grant No. 52272218000X).

REFERENCES
[1] Zhang Dongxia, Miao Xin, Liu Liping, Zhang Yan, Liu Scientific Research, et al. Research on the development of smart grid big data technology [J]. Chinese Journal of Electrical Engineering, 2015, 35 (01): 2-12.
[2] Peng Xiaosheng, Deng Diyuan, Cheng Shijie, Wen Jinyu, Li Zhaohui, Niu Lin, et al. Key technologies of big data for smart grid applications [J]. Chinese Journal of Electrical Engineering, 2015, 35 (03): 503-511.
[3] Li Junnan, Li Wei, Li Huijun, He Xinming, Zhang Shilin. Research on Power Energy Big Data Collection and Application Based on Big Data Cloud Platform [J]. Electric Measurement & Instrumentation, 2019, 56 (12): 104-109.
[4] Zhao Wei, Sun Xiao. The photovoltaic power plant data acquisition system[J]. New Energy, 2018, 21(9): 31-34.
[5] Yan Yingjie, Sheng Gezhen, Chen Yufeng, Jiang Xiuchan, Guo Zhihong, Du Xiuming, et al. Anomaly detection method for state data of transmission and transformation equipment based on big data analysis [J]. Chinese Journal of Electrical Engineering, 2015, 35 01): 52-59.
[6] Chai Pengfei, Sun Yajuan, Shen Ke, Dong Jing. Research on comprehensive evaluation method of power quality based on data envelope analysis [J]. Electric Measurement & Instrumentation, 2016, 53 (04): 124-128.
[7] Zhuang Chijie, Zhang Bin, Hu Jun, et al. Detection of abnormal power consumption patterns of power users based on unsupervised learning [J]. Proceedings of the CSEE, 2016, 36 (2): 379-387.

[8] Wei Daqian, Wang Bo, Liu Dichen, et al. Modeling of measurement big data and abnormal data detection based on high-dimensional random matrix description [J]. Proceedings of the CSEE, 2015, 35 (S1): 59-66.

[9] DONG XU, YANJUN WANG, YULONG MENG, ZIYING ZHANG. An Improved Data Anomaly Detection Method Based on Isolation Forest[J]. 2017 10th International Symposium on Computational Intelligence and Design (ISCID), 2017:287-291.

[10] LIU F T, TING K M, ZHOU Z H. Isolation-based Anomaly Detection [J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6 (1): 1556-4681.

[11] Yang Mao, Yang Qiongqiong, et al. Research on abnormal data identification of wind turbines based on cloud segmented optimal entropy algorithm [J]. Chinese Journal of Electrical Engineering, 2018, 38 (08): 2294-2301 + 2539.

[12] Cheng Zhiyou, Zhu Weiwei, Tao Qing, Wang Miao, Hu Zhengyang. Evaluation method of power quality of power distribution system based on improved radar chart [J]. Electrical Measurement & Instrumentation, 2019, 56 (14): 34-39 + 67.

[13] Zhao Teng, Wang Lintong, Zhang Yan, et al. Identification of user power related factors and power consumption prediction method using mutual information and random forest algorithm [J]. Chinese Journal of Electrical Engineering, 2016 (3): 604-614.

[14] ARYAL S, TING K M, WELLS J R, et al. Improving IForest with Relative Mass[J]. 2014.

[15] YU Xiao, TANG LU AN, HAN JIAWEI. Filtering and refinement: a two-stage approach for efficient and effective anomaly detection[C]//Proc Int Conf on Data Mining ICDM 09. Piscataway, NJ, USA: IEEE, 2009:90-104.