Research on Abnormal Data Detection Method of Power Measurement Automation System

Ming-fei Qu(✉) and Nan Chen

College of Mechatronic Engineering, Beijing Polytechnic, Beijing 100176, China
qmf4528@163.com

Abstract. Aiming at the problems of long time consuming and low accuracy in traditional methods of abnormal data detection in power measurement automation system, this paper studies the methods of abnormal data detection in power measurement automation system. Design the data storage structure table of the electric power metering automation system database, and repair the missing data and denoise the data in the data table. Perform PAA calculation on the data to get the data feature sequence. After the P clustering algorithm pre-clusters the data, the iForest model is used to detect abnormal data to complete the research on the method. The experimental results show that the proposed detection method has the advantages of short detection time and high precision of 91.26–95.67%.

Keywords: Power measurement automation system · Abnormal data · Detection method · Iforest

1 Introduction

Electric power metering automation system refers to a system that can collect, monitor and analyze electric data on power generation side, power supply side, power distribution side and power sale side of power plants, substations, public transformers, special transformers and low-voltage customers, including metering automation master station, communication channel and metering automation terminal. The metering automation system realizes remote automatic real-time meter reading, abnormal alarm of electricity consumption information, voltage and power quality monitoring, line loss analysis, and prepayment by collecting, monitoring, analyzing and processing information such as data, voltage, current, load and other information of each monitoring terminal Functions such as fee management, electricity consumption inspection, load management and control, and power outage statistics provide data support for grid operation and management [1]. But before the realization of these advanced application research, the most basic premise is to ensure the timeliness, integrity and reliability of the data collected by the system (Fig. 1).

At present, the data content involved in the measurement automation system mainly includes various indicators of the terminal such as online rate, automatic meter reading rate, etc., collected data such as table codes, electricity, and power factor, and these basic data are subjected to secondary calculations such as voltage divider advanced application data such as loss, line loss, line loss in sub-stations, etc. Faced with such a

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 217–228, 2021.
https://doi.org/10.1007/978-3-030-67871-5_20

large amount of data, relying on manual or traditional database software tools for daily data quality checks has been unable to effectively guarantee the reliability of data quality. The traditional method of outlier data detection is to use RBF classifier to detect outlier data, and the time cost of this detection method for large system data detection is too high, and the error of detection results is also large [2]. Therefore, based on the above analysis, this paper studies the abnormal data detection method of power measurement automation system.

2 Abnormal Data Detection Method of Electric Power Measurement Automation System

2.1 Establish Data Structure Table of Electric Power Measurement Automation System

The data used to detect abnormal data in the power metering automation system is data in tables such as the communication flow table of the system terminal and the field operation and maintenance record table. Therefore, the system data structure table needs to be established in the terminal database.

The communication flow table is used to record the communication status between each terminal and the master station, mainly including communication flow, number of reconnections and online time, etc. the specific definition is shown in Table 1. The data in this table is collected and counted by the master station at zero every day, reflecting the previous day's communication. If the terminal fails, the master station will not be able to collect current data, so there is no corresponding record in the database. The sending and receiving bytes in the table refer to the master station. The sending bytes represent the number of bytes sent from the master station to the terminal, and the received bytes represent the number of bytes received by the master station [3]. The data flow in the table indicates the flow used to transmit electric energy data among all flows. In addition to the above flow, the table also includes other control information of the terminal, including the number of reconnections, alarm flow and heartbeat flow. The online time indicates the number of heartbeat signals received by the master station, and the terminal transmits once per minute. If the value is 1440, it indicates that the terminal is online for 24 h (Table 2).

Table 1. Communication flow table

Project	Database field name	Data
Terminal code	rtuid	——
Data date	datatime	——
Send (downstream) byte	sendbytes	——
Receive (uplink) bytes	recvbytes	——
Reconnect times	logintimes	——
Data flow	databytes	——
Alarm flow	alarmbytes	——
Heartbeat flow	linkbytes	——
online time	onlinetimes	——

The on-site operation and maintenance table records the maintenance and inspection results of the operation and maintenance personnel on the fault end, including the terminal information and fault information. The table is generated by the system to generate terminal related information, including all kinds of numbers, data time and location from the master station and fault related information, which are filled in by the operation and maintenance personnel. The fault type and fault description are manually filled in by the operation and maintenance personnel and then input into the system. Therefore, for the same fault, the names filled in by different operation and maintenance personnel may be different, which needs further processing.

Line loss refers to the loss of electrical energy during power transmission and transformation. The line loss table is used to record the input and output power and line loss rate of each line, and contains basic information related to line loss for each line daily, as shown in the following table:

Table 2. Line loss table

Numbering	Project	Database field name
1	Line number	LINEID
2	Line name	DISC
3	Date of data	DATATIME
4	Input power	ENERGY_IN
5	Output power	ENERGY_OUT
6	Line loss rate	LINELOSS_RATE
7	Date data sent	SENDDATE

After the above data table is established in the database of the electric power metering automation system, the data collected and stored in the database by the system is processed.

2.2 Processing Data of Power Metering Automation System

Due to various reasons, the data will be incomplete and inconsistent. These data are called error data, which has a great impact on subsequent anomaly detection. Therefore, data cleaning is very important for abnormal data detection.

Data cleaning must first delete the redundant data in the data set. Redundant data is the only characteristic that destroys every record in the data set. When multiple identical records appear, the redundant data must be deleted. Every user must have electricity readings for every hour of the day, and the serious absence is defined as:

1) 20% of the reading points are missing from the curve;

2) The curve is continuously missing more than 2 consecutive reading points. If the data is missing to a serious extent, the user is excluded from the research scope, and the multi-level Lagrangian interpolation method is used to repair the missing value. The missing value repair formula is as follows [4]:

$$P_t = \frac{\sum_{k=1}^{m_1} P_{t-k} + \sum_{i=1}^{m_2} P_{t+i}}{m_1 + m_2} \tag{1}$$

In formula (1): m_1 is the number of forward periods, m_2 is the number of backward periods, t is the time when the system data is missing, P_t is the missing value after repair, P_{t-k} is the system data at time k before t System data, k is the at time i after time t. After the data is patched, the data is denoised by smoothing the system timing relationship curve.

Set the total operation time T of the system for a period of time t' has been in the abnormal state recorded by the observation equipment, the starting point of the abnormal state is recorded as P_{start}, the ending point of the abnormal state is recorded as P_{end}, the starting point of the abnormal state is recorded as $time_{Pstart}$, and the ending point of the abnormal state is recorded as $time_{Pend}$. Suppose that the data set $P_{t'}$ collected by the power metering automation system in t' is expressed as $P_{t'} = \{P_{start}, P_2, \cdots, P_{n-1}, P_{end}\}$, and the data point $P_{i'}$ in t' after processing is recorded as [5, 6]:

$$P_{i'} = \frac{P_{i'-1} \pm (P_{start} - P_{end})}{time_{Pend} - time_{Pstart}} \tag{2}$$

After processing, the data points in the time series relation curve can only be divided into two types, normal data and abnormal data. After processing the data of electric power measurement automation system, the abnormal data features are extracted.

2.3 Feature Extraction of Abnormal Data

In the process of abnormal data detection in the electric power measurement automation system, the time series of each data in the network database is first obtained, which takes the average value of the time series as the element. The specific steps are as follows:

Suppose that $Q = \{q_1, q_2, \cdots, q_m\}$ and $C = \{c_1, c_2, \cdots, c_n\}$ represent the two data time series, and w_q and w_c represent the time series of the two data time series. Use the following formula to calculate the feature average of all data elements in each time series [7]:

$$q_i = \frac{1}{w}(m)\frac{Q}{C} \bullet \frac{\{q_1, q_2, \cdots, q_m\}}{\{c_1, c_2, \cdots, c_n\}} \bullet \frac{[w_q, w_c]}{Q} \tag{3}$$

In formula (4), w represents the eigenvalue to form a new data sequence, m represents the piecewise aggregation approximation, and w_q and w_c represent the mean value of data elements. Assuming that $d(i,j)$ represents the dynamic time bending distance, ξ represents the eigenvalue to form a new data series, and N represents the data change form of the data time series, then use Eq. (4) to obtain the data characteristic series with the average value of the time series as the element [8, 9].

$$r(i,j) = \frac{L_{DTW} \times D_{w_q \times w_c} \bullet d(i,j)}{[\xi \bullet N] \bullet q_i}[\theta] \tag{4}$$

In formula (4), L_{DTW} represents the process in which the two time series are first converted into feature sequences, $D_{w_q \times w_c}$ represents the original time series data, and $[\theta]$ represents the distance accumulation matrix. After extracting the abnormal data features of the power metering automation system, the abnormal data is detected.

2.4 Implementation of Abnormal Data Detection

The clustering analysis of the AP algorithm uses an abnormal data feature similarity matrix, and the similarity between data points is expressed by the square of the negative Euclidean distance. If there are n data, then these data points constitute the similarity matrix S of $n \times n$, $S(i,j)$ represents the similarity between data points i and j, the calculation formula is as follows:

$$\begin{cases} S(i,j) = -\left\|x_i - x_j\right\|^2 \\ S(i,j) \in (-\infty, 0] \end{cases} \tag{5}$$

The element value $S(i,j)$ on the diagonal of the similarity matrix is used to judge the cluster center. $S(i,j)$ is called the preference parameter, which indicates the suitability of data point i as the cluster center of the class. If its value is larger, it means that the point is more suitable to be the cluster center. Set the mean value of similarity matrix as *preference*:

$$preference = \frac{\sum_{i,j=1,i\neq j}^{n} S(i,j)}{n \times (n-1)} \tag{6}$$

The AP algorithm continuously updates the attraction matrix R and the attribution matrix A during the iteration process. The attraction information $R(i,k)$ is the information sent by the sample point i to the possible clustering center k, which indicates the degree of attraction of the sample point i to k. If the value is larger, it indicates that k is more likely to become the center of i; the attribution degree information $A(i,k)$ is potentially. The information sent by the clustering center k to the sample data i expresses the degree of attribution of k as the center of the sample point i. If the value is larger, i is more likely to belong to the cluster with the center k. In the iterative process,

the above two information matrices depend on each other and are updated alternately. The update process is as follows [10, 11]:

then,

$$R(i,k) \leftarrow S(i,k) - \max\{A(i,k') + S(i,k')\} \tag{7}$$

When $i = k$,

$$R(k,k) \leftarrow S(k,k) - \max\{A(k,k') + S(k,k')\} \tag{8}$$

When the number of iterations of attraction matrix R and attribution matrix A exceeds the maximum number of iterations given in advance or the change of $R(i,k) + A(i,k)$ is lower than a given threshold, the iterative process will stop. After the AP clustering algorithm clusters the data sets, iforest detects the outliers of the clustered data.

Assume that the whole data set after AP algorithm classification is Ψ, $\Psi = (\psi_1, \psi_2, \cdots, \psi_n)$, where ψ_q represents the q-th cluster.

iForests is composed of f iTree isolated trees, each iTree is a binary tree structure. An iTree training procedure is as follows [12, 13]:

1) Put the m values of dataset $D = \{x_1, x_2, \cdots, x_m\}$ into the root node of the tree.
2) Randomly specify an attribute r, and randomly generate a split value p in the current data. The size of the split value p is the number between the maximum and minimum values of the specified attribute r in the current data.
3) The partition value p divides the current data space into two sub spaces, divides the data less than p in the specified attribute r into the left sub tree, and divides the data greater than or equal to p into the right sub tree.
4) Recursion steps 2 and 3 generate new nodes until there is only one data in the node.

The following figure is a schematic diagram of the construction of iTree.

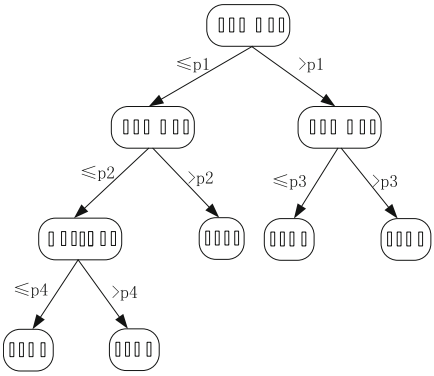


Fig. 1. Schematic diagram of iTree construction

The training steps for f iTree are as follows:

Randomly select ς samples from cluster ψ_q, and use these ς samples to establish an iTree according to the above process. Perform the above steps f times to get f isolated trees, which form an iTree set. When the itrees collection is built, the next step is the exception evaluation phase of the data.

First, calculate the path degree $h(x)$ of the data x to be detected. Path degree $h(x)$ refers to the number of iterations from the root node to the end of the leaf node. For an itrees tree, data x is moved down the partition condition corresponding to the creation until the leaf node is accessed, and the path length $h(x)$ is recorded. Because the structure of itrees is the same as that of the binary search tree, the path length of the leaf node containing data x is the same as the path length of the failed query in the binary search tree, that is, from the root node to the middle node, reach the leaf node, the number of edges traversed. Traverse the itrees set and calculate the exception score for data x. According to the abnormal score, judge the abnormal data. So far, the research on abnormal data detection method of power measurement automation system has been completed.

3 Experimental Research

In order to verify the effectiveness of this method, the following design of comparative experiments.

3.1 Experimental Content

This experiment is a simulation experiment. The experimental group is the abnormal data detection method of the electric power metering automation system studied in this paper, and the experimental group is the traditional abnormal data detection method of the electric power metering automation system. A total of two sets of experiments were conducted in the experiment. The first set of experiments compared the iForest model of the experimental group and the RBF model of the comparative group when detecting data sets containing the same abnormal data; the second group of experimental comparison indicators were abnormalities of the experimental group and the comparative group Data detection method When the abnormal data is detected in the experimental data set, the recall rate and precision rate of the method.

3.2 Experiment Preparation

The power metering data selected in this experiment includes the power consumption information of some users. The details of the data source are as follows (Table 3):

Table 3. Experimental data set details

Serial number	Features	Type	Explain
1	CJ_MP_ID	Integer	Measurement point identification
2	DATA_DATE	Datetime	Data timescale
3	DATA_SORCE	Integer	data sources
4	WRITE_DATE	Datetime	Write date
5	PZ	Float	Total active power
6	PA	Float	Phase A active power
7	PB	Float	Phase B active power
8	PC	Float	Phase C active power
9	QZ	Float	Total reactive power
10	QA	Float	Phase A reactive power
11	QB	Float	Phase B reactive power
12	QC	Float	Phase C reactive power
13	MID_I	Float	Zero sequence current

The calculation formula of the accuracy rate of the abnormal data detection method is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

The calculation formula of the recall rate of the abnormal data detection method is as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

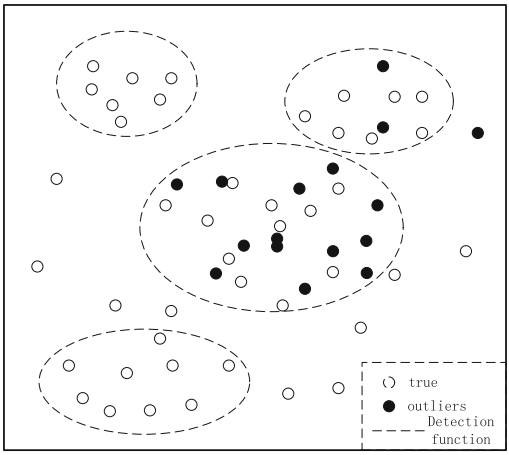
In formulas (9) and (10), TP is normal data judged as positive by the detection method. FP is the data judged as positive by the detection method, but is actually abnormal, and FN is the data judged as abnormal by the detection method, but is actually normal data.

Different data sets carry out anomaly data detection one by one, record the precision and recall, take the precision as the ordinate and recall as the abscissa, and draw the P-R curve. If the P-R curve of one detection method is included by another, the latter is better than the former. If the P-R curves of the two classifiers cross, we can choose the balance point, which is the value when p = R. the larger the balance point BEP, the better the performance of the detection method.

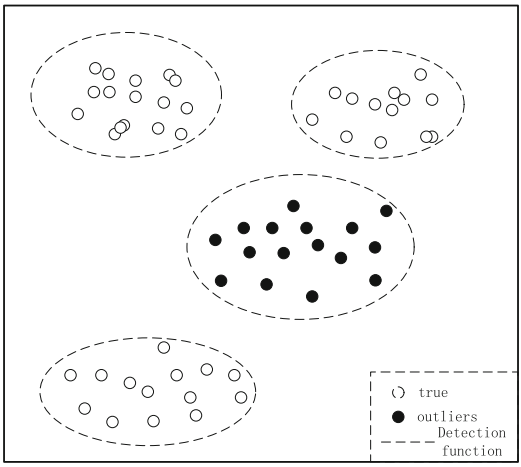
Record the experimental data of the two experiments, process and analyze the experimental data, and draw the corresponding experimental conclusion.

3.3 Experimental Results

In order to verify the effectiveness of this method, the abnormal data of iforest model and contrast group RBF model are detected, and the detection results are shown in the figure below.



(A) RBF detection effect



(B) iForest detection effect

Fig. 2. Comparison of detection effect

Analysis of the data in Fig. 2 shows that iforest can basically detect abnormal data when the dataset containing outliers is used for detection, while the abnormal data detected by RBF is far less than that detected by iforest, indicating that the abnormal data detection effect of iforest model in power metering automation system is better.

The abnormal data detection time of iforest model and contrast group RBF model are compared and analyzed. The comparison results are shown in Table 4.

Table 4. Comparison of detection operation time of two methods/MS

The amount of data	iForest	RBF
5000	521	763
10000	772	995
15000	1064	1571
20000	1474	2346
25000	1838	2892
30000	2345	3487
35000	2760	4109
40000	3596	6007

It can be seen from the analysis of the above figure that when the data set with outliers is used for detection, iforest can basically detect the outliers, while the outliers detected by RBF are far less than those detected by iforest, and several normal data in the RBF detection results. According to Table 4, the detection time of iforest is short and the detection accuracy is high. It shows that in this detection experiment, the detection effect of iforest on abnormal data is better.

However, iForest cannot detect local outliers, so on the basis of the first experiment, the second experiment tests the abnormal data detection method combining iForest and clustering algorithms studied in this paper from a data perspective.

The experimental results of the second experiment are shown in the table below, and the experimental conclusions are obtained by analyzing the data in the table (Table 5).

Table 5. Comparison of test results between the experimental group and the comparison group

Data set number	Experimental group method		Contrast group method	
	Recall rate/%	Accuracy rate/%	Recall rate/%	Accuracy rate/%
1	100	95.45	80.81	74.35
2	100	94.13	81.93	73.98
3	100	95.67	75.97	74.11
4	100	91.84	80.30	73.27
5	100	92.63	78.97	74.65
6	100	92.01	79.23	74.00
7	100	94.75	75.25	73.93
8	100	94.75	82.04	73.88
9	100	91.26	77.71	73.33
10	100	94.42	77.66	74.69
11	100	94.69	79.42	73.69
12	100	95.11	78.45	73.83

According to the above table, the recall rate of the test methods in the experimental group is 100%, and the recall rate interval of the test methods in the comparative group is seventy-four point three five ~ 82.04% The results show that the experimental group can accurately judge the normal data. The precision interval of the test method in the experimental group is ninety-one point two six ~ 95.67%. The precision interval of the control group method is seventy-three point two seven ~ 74.69%, which is far lower than the detection method of experimental group, indicating that the accuracy of abnormal data detection of experimental group is high and the number of false detection is small.

In summary, the automatic abnormal data detection method of the power metering system studied in this paper takes less time, has higher detection accuracy, and has higher practicability.

4 Conclusion

In the power metering automation system, the detection of abnormal data may obtain more useful value than the analysis of normal data. This paper proposes to use a combination of clustering algorithm and isolated forest as an anomaly detection method. Through comparison experiments with traditional methods, it proves that the research method takes less time to detect, and the detection results are more reliable, and the method is feasible.Fund projects

Design and implementation of new energy inverter based on MCU control (CJGX2016-KY-YZK034)

References

1. Yang, X., Qu, Y., Pang, H., et al.: Power metering pipeline fault warning technology based on deep learning algorithm. Electron. Des. Eng. **28**(04), 153–157 (2020)
2. Gao, S., Li, C.: An improved spectral clustering algorithm for anomaly detection of power data. Comput. Simul. **36**(11), 239–242 + 304 (2019)
3. Yang, J., Zeng, X., Yao, L., et al.: Research on abnormal electricity monitoring based on large data mining. Autom. Instrum. **08**, 219–222 (2019)
4. Tong, X., Yu, S.: Fault detection algorithm for transmission lines based on random matrix spectrum analysis. Autom. Electr. Power Syst. **43**(10), 101–115 (2019)
5. Xu, G., Ning, B., Zhong, Y.: Automatic matching of voltage blackout events in metering automation system. Electron. Test **04**, 111–112 (2019)
6. Huang, J., Dai, B., Zhang, L., et al.: Study on dynamic identification of abnormal data of electric energy measurement device. Guangxi Electr. Power **41**(04), 53–55 + 64 (2018)
7. Chen, Q., Zheng, K., Kang, C., et al.: Detection methods of abnormal electricity consumption behaviors: review and prospect. Autom. Electri. Power Syst. **42**(17), 189–199 (2018)
8. Liu, S., Glowatz, M., Zappatore, M., et al. (eds.): e-Learning, e-Education, and Online Training, pp. 1–374. Springer, Heidelberg (2018)

9. Zhang, J., Chen, F., Li, B., et al.: Research on energy metering abnormity and fault analysis technology based on metrological automation system. Yunnan Electr. Power **46**(02), 63–65 (2018)
10. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy. **21**(9), 902 (2019)
11. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)
12. Ding, M., Li., C., Li, H., et al.: Fault detection method of photovoltaic inverter based on massive data mining. Electr. Autom. **40**(03), 30–32 (2018)
13. Shuai, L., Gelan, Y.: Advanced Hybrid Information Processing, pp. 1–594. Springer, Heidelberg