

## Anomaly detection of online monitoring data of power equipment based on association rules and clustering algorithm

Yu-Xiang Cai<sup>†</sup> and Li-Jun Cai,

*State Grid Fujian Information & telecommunication Company,  
No.264, Wusi Road, Fuzhou, Fujian Province, China*

<sup>†</sup>E-mail: 150294189@qq.com  
www.fj.sgcc.com.cn

Zhou Lu

*Department of electrical engineering, Shanghai Jiaotong University  
No. 800 Dongchuan Road, Shanghai, China*

<sup>†</sup>E-mail: 672470719@qq.com

With the continuous research and development of smart grid and energy Internet, as well as the rapid construction of power transmission and transformation equipment in various places, the amount of data collected from the equipment is also increasing. To dig out the effective information must be to ensure the accuracy of the data. However, large data set must contain erroneous or abnormal data. The traditional method cannot handle the big data anomaly detection well. Therefore, this paper presents anomaly detection based on association rules and clustering algorithms. The association rules are used to find out the sequences with relevance in the dataset. Then the FCM algorithm are used to separate the abnormal data into a sensor abnormal that can be cleaned and a device abnormality that cannot be cleaned. For the correlation sequence, the sensor anomaly and the device abnormality are found by the method of association and clustering, then early warning and maintenance advice are given.

*Keywords:* Big data, Anomaly detection, Data cleaning, Association rules, FCM.

### 1. Introduction

With the development of big data mining and analysis in power system, the research methods involved should be based on the ideal and reliable data. However, due to the fact that the data which power system transmission and distribution equipment monitor are complex and diverse, there will be missing and inconsistent data problems and other issues [1-4]. In data processing, these data are generally referred to as "dirty data". The conclusions drawn from "dirty data" do not have any reference value. It can be seen that these "dirty data" need to be processed and cleaned before the data can be analyzed.

In the field of power system, due to the characteristics of the data which power system transmission and distribution equipment monitor, i.e., large amount of data and the diversity of data types, there must be a certain correlation between different data sets from the perspective of big data [5]. Therefore, the application of association rule in data cleaning is of great significance. In [6], the association rule is used in fault evaluation of power transformer. The innovation of this paper is introducing the gray evaluation method, and applying new dynamic value to the correlation coefficient in gray evaluation method. [7] is forecasting and analyzing cities' electric power load based on the association rule. The advantage of this paper is that it adapts to the periodic and stochastic characteristics of power system load variation, so that the analysis and arrangement of massive data are realized.

With the construction and development of smart grid, the quality and accuracy of power transmission and distribution equipment monitoring data are becoming more and more demanding. Also, data cleaning is becoming more and more important in the analysis and prediction of the power system's operating state[8]. In [9], center clustering method is used to detect and characterize the anomaly data in time series. It proposed using a spatial and temporal clustering to analyze the structure of each window, and considering the abnormal degree of each sub-sequence. In [10], anomaly data is detected and cleaned by sparse representation. The innovation lies in the ability to capture variable relationships and time-domain correlations without assuming that any model signals are generated. In [11], support vector machine is used to detect and process the data anomalies collected by the nuclear power plant. This method has the advantages of improving the detection rate of the multiple anomaly detection algorithm to the noise and reducing the false alarm rate.

In summary, anomaly detection of monitoring data has the following problems: it cannot find the exact exception point; there is a high false alarm rate; it cannot determine whether the abnormal points reflect the abnormal condition of power equipment.

This paper proposes a new method of anomaly data detection based on association rules and clustering algorithm. First, Apriori algorithm is used for calculating the association relationship of monitoring data. The variables with strong correlation are extracted. Then, fuzzy clustering method is used to detect the anomaly degree of each time series. At last, four types of abnormal data are classified which are in contribution for detecting the condition of power transformer. If the abnormal operation situation are reflected by the anomaly detected, early warning and maintenance advice are given.

## 2. Association Analysis of Time Series

### 2.1. Symbolic process of time series

Online monitoring data of power transmission equipment is of value type, but the Apriori algorithm for mining association rules is of Boolean type. In the need for make the data available for Apriori algorithm, the symbolic process like linear fitting and normalization of data has to be made. The trend value of data falls between  $[-1,1]$ , so that we can make the symbolic process after the standardization of trend value. Assume two time series  $x_1$  and  $x_2$ , the symbolic process is shown in Tab. 1.

Tab.1 The symbolic process of time series  $x_1$  and  $x_2$

The range of trend value	$x_1$	$x_2$
$[-1,-0.6]$	1a	2a
$[-0.6,-0.2]$	1b	2b
$[-0.2,0.2]$	1c	2c
$[0.2,0.6]$	1d	2d
$[0.6,1.0]$	1e	2e

Through this process, the input data can be transferred from value type to Boolean type. The basic items of two sub sequences at the same period can be combined to form a thing such as  $\{1b2a\}$ . All of things can make a set called  $I$ .

Apriori algorithm is used for the  $S$  to get the association rules between basic items so that the trend of multi-dimensional time series can be detected.

### 2.2. Association rules of time series

Association rules between two time series can be measured based on the coefficients like support value, confidence value and interest degree.

In section 2.1, we assume that  $I$  represents the set of items  $I = \{i_1, i_2, \dots, i_m\}$  and items  $A, B$  belongs to the items in the set  $I$  which means  $A \subset I$ ,  $B \subset I$ ,  $A \cap B = \emptyset$ . So the support value means the items  $A$  and  $B$  both occur in the event  $D$  at the same time. The support value can be calculated by

$$s = P(AB) \quad (1)$$

The confidence value means the probability of  $B$  are included in the event  $D$  which includes item  $A$ . The calculating format is

$$c = P(B | A) \quad (2)$$

The interest degree reflects the correlation degree of item  $A$  and  $B$  as follows.

$$i = \frac{P(AB)}{P(A) \cdot P(B)} \quad (3)$$

Based on the coefficients of items of sequences, the support value and confidence value of the whole time series X and Y can be calculated. Assume that  $A_i \rightarrow B_i$  is the rules which meet the conditions that interest degree above 1,  $n_i$  represents the occurrences of  $A_i \rightarrow B_i$ . So the support value and confidence value of two time series can be calculated by (4-5).

$$s(X \rightarrow Y) = \frac{\sum n_i \cdot s(A_i \rightarrow B_i)}{\sum n_i} \quad (4)$$

$$i(X \rightarrow Y) = \frac{\sum n_i \cdot i(A_i \rightarrow B_i)}{\sum n_i} \quad (5)$$

In this paper, the procedure to calculate the association rules for the time series are shown as fig.1.

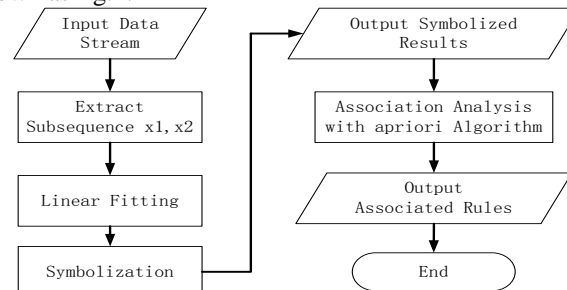


Fig. 1 The procedure for association rules of time series

Sliding window are used for the sub sequences of time series and then the symbolic process are made to get the items of trend values. Noise are removed from the original data and this procedure are available to many online monitoring data like oil gas, core grounding current and so on. Apriori algorithm is used for the items to get the support value and confidence value of the whole time series.

### 2.3. Case of two time series

The data of oil gas monitoring device in Shandong Province are used for the case study. The two time series of  $C_2H_4$  and oil temperature are shown in Fig.2a. We make several sliding windows of the whole time series and symbolic process are used in each subsequence in sliding windows. So the continuous points are transferred into segments which exactly show the trend of the subsequences. And the trend can be separated as rise, gentle and fall.

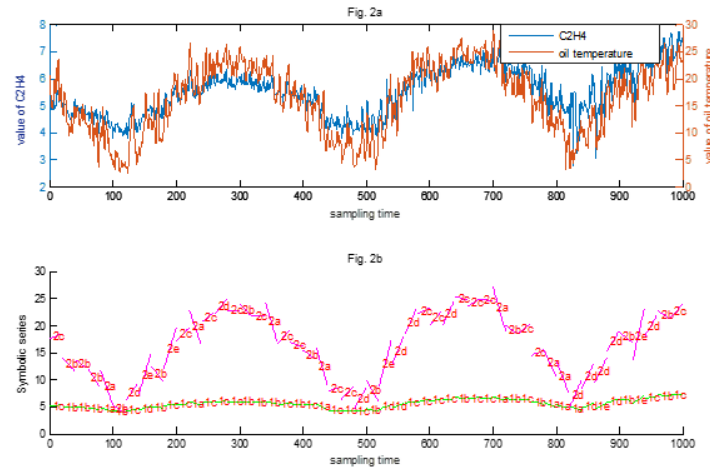


Fig. 2 Original data and linear fitting results of  $C_2H_4$  and oil temperature

The results and symbolic items of time series after linear fitting are shown in Fig. 2b. Based on the items, we can calculate the association rules of the time series of  $C_2H_4$  and oil temperature in Tab. 2.

Tab. 2 Several association rules (support value, confidence value and interest degree) of items of time series of  $C_2H_4$  and oil temperature

Items	Support value	Confidence value	Interest degree
1a->2a	0.8571	0.1200	6.1224
1b->2b	0.6667	0.2000	3.0303
1c->2c	0.6190	0.2600	1.8207

The strong correlation of the two time series of  $C_2H_4$  and oil temperature can be directly seen in Fig. 2. The support and confidence value of items calculated in Tab. 2 are in consist of the results in Fig.2. Based on the format (4-5), the total interest degree of the whole time series is 3.3710 that is definitely high. So we can get the conclusion that the two time series has strong correlation.

### 3. Clustering Method and the Anomaly Detection

#### 3.1. The clustering algorithm

We use furry clustering method (FCM) to detect the anomaly of time series. Assume that the time series  $X = x_1, x_2, \dots, x_p$  is with the length p. Let the sliding window with length q and step size r to separate the whole time series into n subsequences [12].

$$n = \frac{p-q}{r} + 1 \quad (6)$$

The FCM uses the objective function to get the membership degree of cluster center, in order to decide the category of values and achieve the goal of classifying. Assume there are  $c$  ( $c \leq n$ ) clustering centers  $V_1, V_2, \dots, V_c$  according to time series  $X_1, X_2, \dots, X_n$ . The objective function  $J$  is as follows.

$$J = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|V_i - X_k\|^2 \quad (7)$$

$u_{ik}$  is the fuzzy coefficient and is Euclidean distance. The objective function  $J$  is in response with the two function below.

$$V_i = \frac{\sum_{k=1}^n u_{ik}^m X_k}{\sum_{k=1}^n u_{ik}^m} \quad (8)$$

$$u_{ik} = \frac{1}{\sum_{j=1}^c \left( \frac{\|V_i - X_k\|}{\|V_j - X_k\|} \right)^{\frac{2}{m-1}}} \quad (9)$$

$U$  is membership matrix denoting the membership degree of  $X_1, X_2, \dots, X_n$  to  $V_1, V_2, \dots, V_c$ . The sequences can be reconstructed to  $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n$  based on the membership matrix  $U$  and clustering center  $V_1, V_2, \dots, V_c$ . The Euclidean distance of  $X_1, X_2, \dots, X_n$  and  $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n$  can be calculated below [13].

$$E_k = \|X_k - \hat{X}_k\| \quad (10)$$

The value  $E_1, E_2, \dots, E_n$  can be the anomaly degree of the original time series.

### 3.2. Procedure of anomaly detection

The anomaly situation of online monitoring data can be separated into five types: error data, missing values, abnormal points caused by interference and high frequency oscillation caused by communication noise [14]. The three types of the anomaly can reflect the abnormal operation situation of power equipment so that we should detect these anomaly. The procedure is as follows.

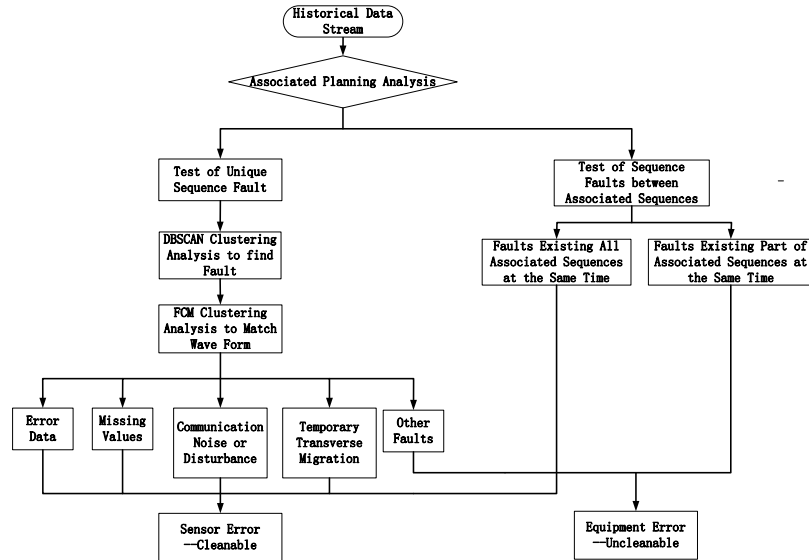


Fig. 3 The whole procedure of anomaly detection

- (1) The historical data of monitoring device are collected together for analysis. Then the association rules and clustering method are used for these data so that the time series which have strong correlation can be extracted.
- (2) The FCM are used for the monitoring data flow. If error data or missing values and high frequency oscillation occurs in single or a small amount of time series, we can consider that this anomaly may be caused by the device or communication. This phenomenon should lead us to inspect the monitoring device.
- (3) The time series with strong correlation has to be considered at the same time in the anomaly detection process. If abnormal points occur in most of the related time series, then we consider that it should be caused by the abnormal operation situations of the power equipment. This result makes an warning alarm and we should check the equipment in time.

#### 4. Case Study

The online monitoring data of Shandong Province are used for case study. The original data include oil gas and oil temperature of power transformers.

##### 4.1. Case one

We take the single dimensional time series of partial discharge for example. The anomaly degree of the original time series after clustering process are shown in Fig. 4. We can read from the figure that there are one anomaly degree with a

definitely high value. This means the anomaly may be the type of error data or missing data and the wave go back to normal situation after the data cleaning method.

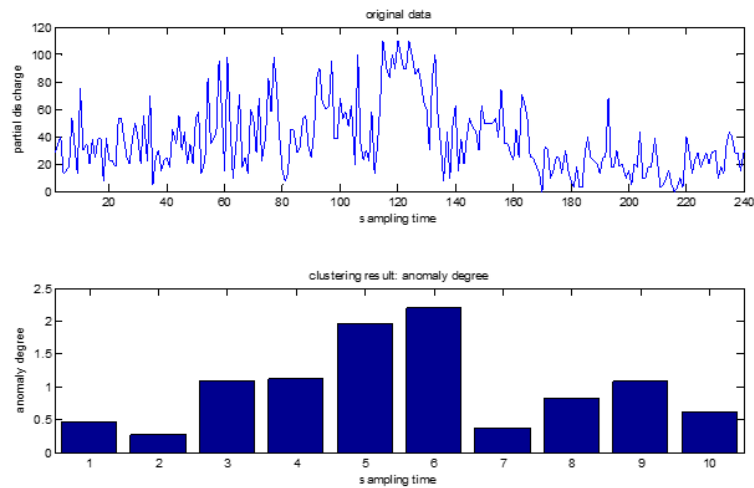


Fig. 4 Error or missing data of time series

The time series of ambient temperature is shown in Fig. 5. There are several high anomaly degrees. Considering the sensors of ambient temperature are installed outside the transformer. So the abnormal situation may be caused by the external interference or communication lose.

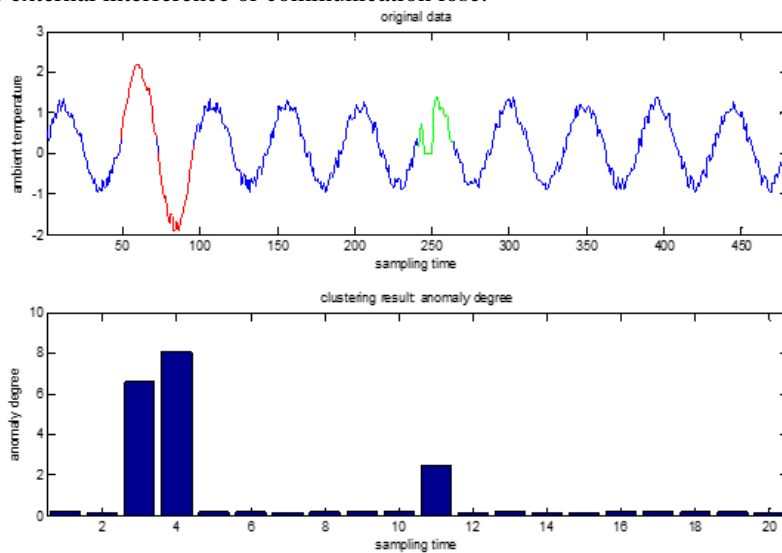


Fig. 5 Abnormal points caused by interference



#### 4.2. Case two

In section 2.3, We have conclusion that the time series of  $C_2H_4$  and oil temperature have strong correlation. So in this section the two time series are considered together for anomaly detection. In Fig. 6, we found two areas that abnormal situation occur.

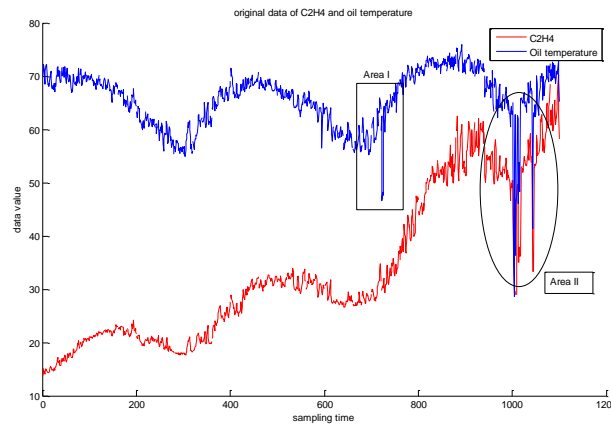


Fig. 6 Original data and the detected abnormal areas of  $C_2H_4$  and oil temperature

In Area II, there are a larger amount of abnormal points in both time series of  $C_2H_4$  and oil temperature. Due to the strong correlation of the two time series, the anomaly occurring at the same time should not be the coincidence. These may be caused by change of operation situations of power transformer. So we make the early warning of transformer's condition and give the advice that maintenance should be carried out in time.

In Area I, there are several abnormal points in just one time series of  $C_2H_4$  which meet the character of error data. So we get the conclusion that this situation don't correspond to the condition of transformer.

#### 5. Conclusions

Association rules in this paper are used to unearth the relationship of respective variables among complex data types. Then the links between time series of those variables are established which contribute to the final anomaly detection. According to the correlation between the time series and the anomaly degree, the abnormal data types are classified. Three abnormal types reflect the abnormal situation while one abnormal type is conducive to detect the condition of transformer. After the conclusion of anomaly detection, the maintenance advice can be given to the staff in substation.

The dimension of time series in this paper is not above five. To get higher dimensional time series, future work has to be researched.

## References

1. TuXinli, Liu Bo and Lin Weiwei, Survey of big data [J]. Application Research of Computers, 2014, 06:1612-1616+1623.
2. Li Guojie and Cheng Xueqi, Research Status and Scientific Thinking of Big Data [J]. Bulletin of the Chinese Academy of Sciences, 2012, 06:647-657.
3. Song Yaqi, Zhou Guoliang and Zhu Yongli, Present Status and Challenges of Big Data Processing in Smart Grid [J]. Power System Technology, 2013,37(4):927-935.
4. Zhang Wenliang, Liu Zhuangzhi and Wang Mingjun, Research Status and Development Trend of Smart Grid [J]. Power System Technology, 2009, 33(13):1-11.
5. Hu Yulong, Mining Dynamic Association Rules from Multiple Time-series data streams [D]. Harbin Institute of Technology, 2013.
6. Ma Gaofeng, The Study and Application of Association Rules in Assessment of Power Transformer Fault [D]. College of Computer Science of Chongqin University, 2013.
7. Geng Fang, The Recommendation of Urban Power Load Forecasting Model Based on Association Rules[D]. Tianjin University,2009.
8. Li Li, Zhang Deng, XieLongjun, et al. A Condition Assessment Method of Power Transformers Based on Association Rules and Variable Weight Coefficient[J]. Proceedings of the CSEE,2013, 33(24);152-159.
9. Sheng Gehao, Liu Yadong, Jiang Xiucheng, et al. Key Techniques and Development Trends of Power Transmission Equipment Intelligentization [J]. East China Electric Power, 2011, 39(9): 1379-1385.
10. M, Ester, H, -P, Kriegel, J, Sander, and, X, Xu. A density based algorithm for discovering clusters in large spatial databases with noise [J]. KDD-96 Proceedings , 1996: 226-231.
11. Chandresh, Kumar, Maurya, Durga, Toshniwal. Anomaly detection in nuclear power plant data using support vector data description[J]. Students' Technology Symposium, 2014: 82-86.
12. Pang Jingyue, Adaptive Anomaly Detection for Data Stream of Sequence-based Sliding Windows Model [D]. Harbin Institute of Technology, 2013.
13. Feng Shaorong and Xiao Wenjun, An Improved DBSCAN Clustering Algorithm [J]. Journal of China University of Mining & Technology, 2008, 37(1): 105-111.
14. She Chunhong, Methodological Research on Data Cleaning [J], Computer Applications, 2002, (12): 128-130.