An OPTICS Clustering-based Anomalous Data Filtering Algorithm for Condition Monitoring of Power Equipment

Qiang Zhang[1], XuwenWang[2], XiaojieWang[3]

1State Grid Electric Power Research Institute, Beijing 100192, China
2Institute of Medical Information, CAMS & PUMC, Beijing 100020, China
3Beijing University of Posts and Telecommunications, Beijing 100876, China
zhangqiang7@sgepri.sgcc.com.cn
wang.xuwen@imicams.ac.cn
xjwang@bupt.edu.cn

Abstract. In allusion to the widespread anomalous data in substation primary equipment condition monitoring, this paper proposes an OPTICS (Ordering Points To Identify the Clustering Structure) clustering-based condition monitoring anomalous data filtering algorithm. Through the characteristic analysis of historical primary equipment condition monitoring data, an anomalous data filtering mechanism was built based on density clustering. The effectiveness of detecting anomalous data was verified through the experiments on one 110kV substation equipment transformer oil chromatography and the GIS (Gas Insulated Substation) SF6 density micro water. Compared with traditional anomalous data detection algorithms, the OPTICS Clustering-based algorithm has shown significant performance in identifying the features of anomalous data as well as filtering condition monitoring anomalous data. Noises were reduced effectively and the overall reliability of condition monitoring data was also improved.

Keywords: Anomalous Data; OPTICS Clustering; Condition Monitoring; Data Mining

1 Introduction

In recent years, along with the development of electric power equipment intelligent communication, the acquisition of highly reliable data of substation primary equipment is required in smart grid. However, the anomalous data of equipment condition monitoring may affect the reliability of data.

At present, the main causes of anomalous data in condition monitoring included: 1) the breakdown of signal acquisition part occurs frequently, such as failure of sensor. 2) the high failure rate of communication in the substation. 3) the poor anti-interference ability of measurement system and 4) the failure of transmission or processing of data. The large number of anomalous data may bring negative effect to the condition monitoring system, fault diagnosis system and the reliability of data analysis. Therefore, filtering

the anomalous data effectively becomes an important research problem in the smart grid.

The traditional methods are mainly based on threshold [1] and the statistic of measurement of outliers, such as three sigma criteria (Pauta Rule). These principles were based on the scope of a certain accuracy to detect outliers [2]. However, the outlier data distribution characteristic and regularities in electric power equipment field was different from statistical methods in the distribution of the default assumption. In the real application, these methods can filter out some part of the normal data or even real failure data of equipment, which lead to the accuracy degradation of acquisition data and reduce the performance of fault diagnosis.

In this paper, the clustering technique was first applied in the analysis of anomalous data of condition monitoring, we proposed an OPTICS clustering-based anomalous data filtering algorithm [3]. First, we utilized OPTICS density clustering algorithm for mining the distribution characteristics of acquisition data. Then, we designed an anomalous data filtering strategy, combining with electric power equipment state standards and threshold decision rules, etc. Compared with traditional methods, our method excavated potential characteristics distribution of condition monitoring data of substation equipment, and filtered more anomalous data effectively.

2 Related work

Data processing of power transmission and transformation equipment condition monitoring was divided into monitoring data upload, data storage, and data alarm analysis [4]. Anomalous data was processed before data storage, as shown in figure 1:

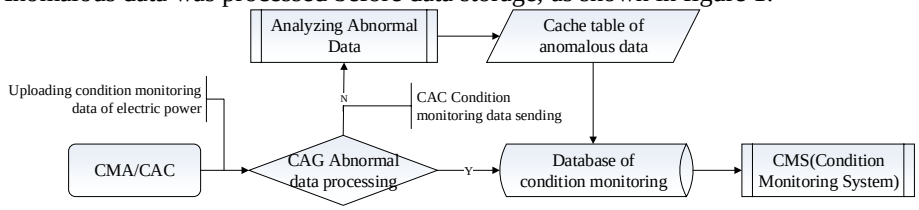


Fig. 1. The work flow of processing electric equipment condition monitoring anomalous data

Power equipment monitoring devices in substation, which access to the CAG (Condition Acquisition Gateway) using CMA (Condition monitoring agent) and CAC (Condition Acquisition Controller) devices, achieve the condition information of equipment. The main station system integrated condition monitoring data and performed the secondary processing of monitoring data, including the anomalous data filtering, threshold calculation of monitoring data [5] and data storage.

2.1 Threshold method

The traditional method of condition monitoring data acquisition uploaded data to data servers directly through the IEC61850 communication protocols without the data pre-processing step. However, data processing was increasingly important. Recently, some preprocessing schemes showed effectiveness in reducing noises, such as the Pauta Rule [6], see as formula (1). $\overline{x}$ was the sample mean, δ was the sample variance.

$$\left| x_d - \overline{x} \right| > 3\delta \tag{1}$$

2.2 Statistic method

Statistical methods for the anomalous data detection, such as the Grubbs test [7], the Dixon test [8] and the t–test, often assumed that the sample data fit the independent normal distribution, and filtered anomalous data according to the level of significance test and confidence interval. The main difference between them was the test statistic distribution and the critical value table.

Taking the Grubbs test formula as an example for anomalous data detection, see as follows:

$$G_{(n)} = \left(X_{(n)} - \overline{x} \right) \Big/ s \tag{2}$$

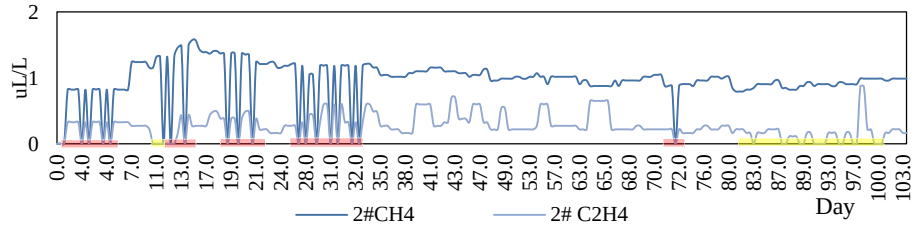
$$G'_{(n)} = \left(\overline{x} - X_{(1)} \right) \Big/ s \tag{3}$$

$G_{(n)}$: function that $x_{(i)} > \overline{x}$. $G'_{(n)}$: function that $x_{(i)} < \overline{x}$. $\overline{x}$: Sample mean. S: The standard deviation.

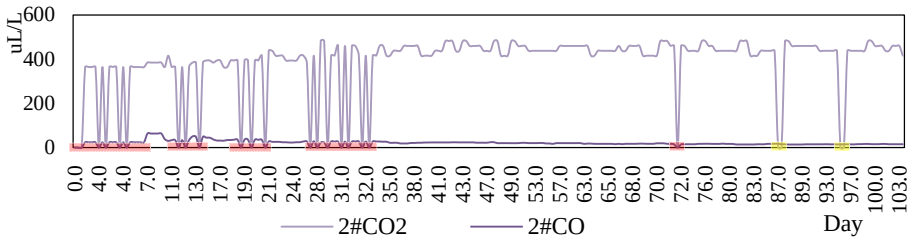
3 Our Method

3.1 Anomalous data

Electric power equipment data was not a scalar. It was necessary to evaluate the outliers of a data object collection, which had multiple dimension parameters. There were multiple causes of anomalous data of the device object, such as the sensor fault, communication failure, collecting device failure and so on.



(a) Anomalous data distribution of 2# transformer oil chromatography methane and ethylene parameter



(b) Anomalous data distribution of 2# transformer carbon monoxide, carbon dioxide parameter

Fig. 2. The historical sample data of a substation transformer oil chromatography equipment condition, the red and yellow areas were two kinds of anomalous data.

As shown in figure 2, taking the equipment operating data of 2# substation oil chromatography part parameters as an example, the red areas of real-time data represented a typical anomalous condition, since all of their parameters had a zero value. The yellow areas in figure 2(a) showed another type of anomalous condition, which has abnormal C_2H_4 parameter (zero value) and normal CH_4 value. It was difficult to judge the current condition of the equipment by one or two parameters.

Observing the occurrence of anomalous data, which had high frequency at the initial operating stage, we made a definition among different parameters for anomalous data of power equipment condition monitoring:

Definition 1: electric power equipment condition anomalous data was the collection of observed object, which have significant differences or inconsistent performance with the whole data set of the equipment [9].

Definition 2: electric power data outlier degree referred to the degree of deviation between the object data and the whole or local data set.

3.2 An OPTICS Clustering-based anomalous data filtering algorithm

Unlike partitioning cluster, such as k-means algorithm, the OPTICS[12] algorithm was not sensitive to dirty data or anomalous data. It was also less dependent on the neighborhood radius of initial parameter and the min points of neighborhood. OPTICS algorithm generated an order of augmented cluster for clustering analysis. The order represented the density-based cluster structure of various samples. In other words, it can get DBSCAN clustering results based on the any parameter and MinPts from the generated sequence. Like DBSCAN, OPTICS requires two parameters: ε, which describes the maximum distance (radius) to consider, and MinPts, describing the number of points required to form a cluster. A point P is a core point if at least MinPts points are found within its ε-neighborhood $N_\varepsilon(p)$. Contrary to DBSCAN, OPTICS also considers

points that are part of a more densely packed cluster, so each point is assigned a core distance that describes the distance to the MinPts-th closest point:

$$\text{core}-\text{dist}_{\varepsilon,\text{minPts}}(p)=\begin{cases}\text{Undefined} & \text{if } |N_{\varepsilon}(p)| < \text{MinPts}\\ \text{MinPts - th smallest distance to } N_{\varepsilon}(p) & \text{otherwise}\end{cases} \quad (4)$$

The reachability-distance of another point o from a point p is the distance between o and p, or the core distance of p:

$$\text{reachablity}-\text{dist}_{\varepsilon,\text{minPts}}(o,p)=\begin{cases}\text{Undefined} & \text{if } |N_{\varepsilon}(o)| < \text{MinPts}\\ \max(\text{core - dist}_{\varepsilon,\text{MinPts}}(p), dist(p,o)) & \text{otherwise}\end{cases} \quad (5)$$

If p and o are nearest neighbors, this is the $\varepsilon' < \varepsilon$ we need to assume in order to have p and o belong to the same cluster. Both the core-distance and the reachability-distance are undefined if no sufficiently dense cluster (w.r.t. ε) is available. Given a sufficiently large ε, this will never happen, but then every ε-neighborhood query will return the entire database, resulting in $O(n^2)$ runtime. Hence, the ε parameter is required to cut off the density of clusters that is no longer considered to be interesting and to speed up the algorithm this way.

Pseudo code:

```
OPTICS(datasets, eps, MinPts)
    for each point p : datapoints
       p->reachability-distance = UNDEFINED
    for each unprocessed point p of datasets
       N = getNeighbors(p, eps)
       mark p as processed
       output p to the ordered queue
       if (core-distance(p, eps, Minpts) != UNDEFINED)
          Seeds = empty priority queue
          update(N, p, Seeds, eps, Minpts)
          for each next q in Seeds
             N' = getNeighbors(q, eps)
             mark q as processed
             output q to the ordered list
             if (core-distance(q, eps, Minpts) !=
UNDEFINED)
update(N', q, Seeds, eps, Minpts)
```

The flow chart of OPTICS clustering-based condition monitoring anomalous data filtering algorithm was shown in figure 3:

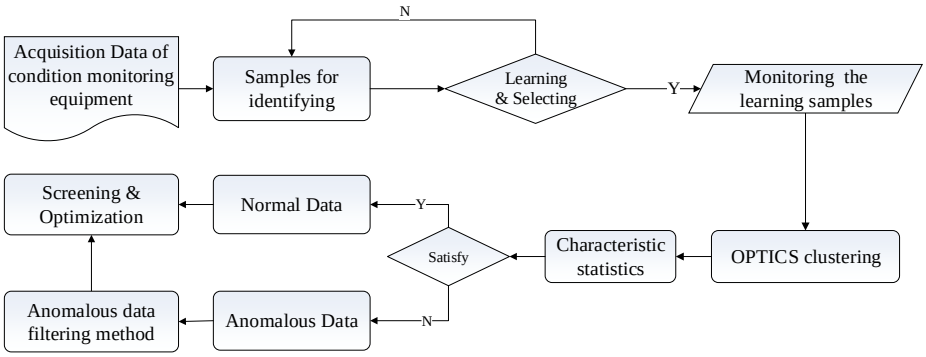


Fig. 3. Flow chart of condition monitoring anomalous data filtering algorithm

The main steps:

- Obtain condition acquisition data object as learning samples.
- Sort the data object by OPTICS clustering algorithm and output the clusters by density.
- Calculate the characteristic probability of the sorted clusters results: according to the OPTICS clustering results of ordered sample object id, calculate the number of core object clusters, the cluster distances and the similarity between clusters.
- According to the sample quantity within clusters and cluster similarity, sorting clusters from high to low order, it can distinguishes the anomalous cluster from other clusters.
- Filter out the anomalous clusters, optimize the learning samples.

4 Experiment

4.1 Transformer oil chromatographic anomalous data analysis

Data.

The Experiment chose nearly 3 years transformer 2# DGA (Dissolved Gas-in-oil Analysis) of an 110kV step-down substation as sample data. The amount was nearly 4000. The anomalous data was generated in several periods. In the early operation, the cause of anomalous data was instability. The anomalous rate of data was higher and displayed irregular distribution. After that, the anomalous data displayed a periodic distribution.

The main measure standard of transformer condition monitoring parameters of anomalous data was as follows.

- Monitoring parameters value of the object were all zero;
- The whole or local parameter value of object data deviated from the mean value.
- Local monitoring parameters (such as CO CO_2) of object data were zero.

Based on the above definition of anomalous data, we can judge the anomalous condition of equipment.

Clustering Analysis.

We obtained the object set of sorted clusters by OPTICS, setting ε =2, MinPts =5. Core objects were sorted based on density. Table 1 shows the sample of cluster results. Each row represents a core object vector. The second column represents the number of sorted sample objects in each cluster. Items in Column 3 to 10 are normalized values of parameters in the core object vector, such as CH_4, C_2H_6, C_2H_4, etc. It was verified by professional maintainers that the more objects the cluster contained, the more possible it was a normal cluster.

Table 1. The OPTICS cluster density core objects of transformer oil chromatographic sample

Cluster id	Quantity of objects	CH_4	C_2H_6	C_2H_4	C_2H_2	H_2	CO	CO_2	total hydrocarbon
0	42	0.0068	0	0.00067	0	0.03007	0.11486	0	0.00753
1	358	0.0022	0	0.00062	0	0.01046	0.04952	0.93	0.00289

Verification of algorithm.

(1) Recognition accuracy P_a .

The anomalous data accuracy was defined as follows:

$$P_a = \frac{N_{a,t}}{N_{x,t}} \tag{6}$$

$N_{a,t}$: The number of normal acquisition data under timing constants

$N_{x,t}$: All samples under the same timing constants.

The recognition accuracy P_a is a ratio between normal samples and the total data.

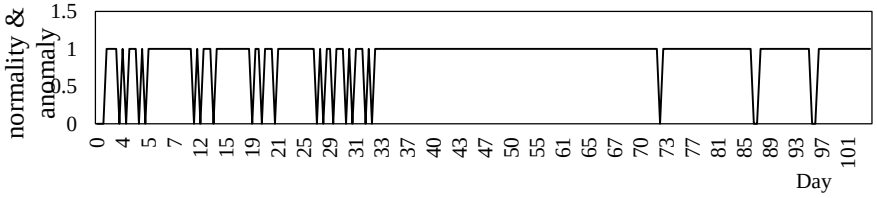
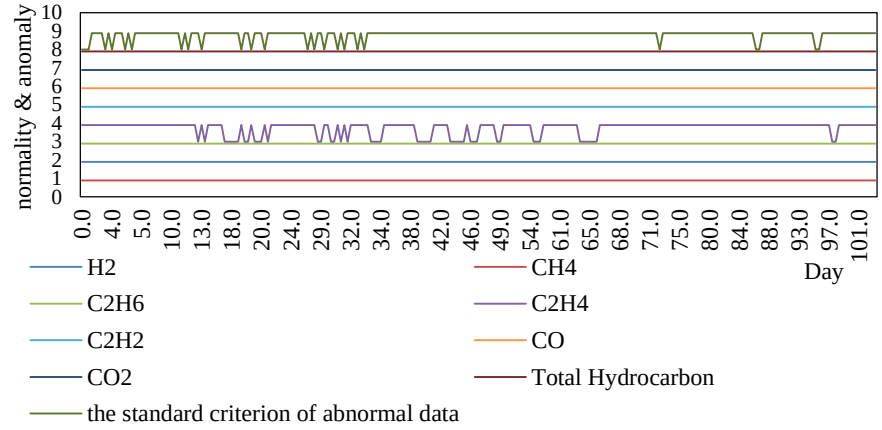


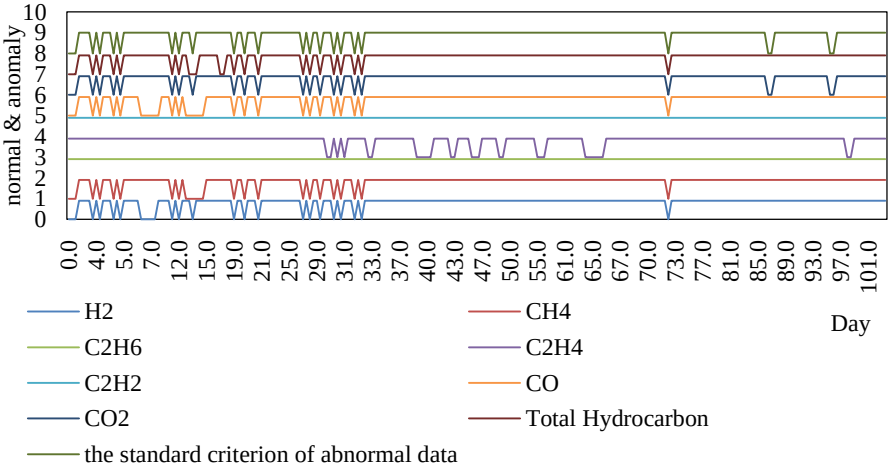
Fig. 4. The data result of OPTICS Clustering, 1: normal object 0: anomalous object

As shown in figure 4, according to the sorting result of OPTICS clustering algorithm, we obtained the condition statistics table of equipment anomalous data distribution.

Value “1” represented the normal condition of equipment and “0” represented the anomalous condition.



(a) The status contrast chart of data in Pauta Rule



(b) The status contrast chart of data in Grubbs test method

Fig. 5. Statistical comparison of data parameters in Pauta and Grubbs test method, n=3 α=0.01

As shown in figure 5, the Pauta Rule and the Grubbs test gave the contrast curves of normal and anomalous data. The x axis represented sampling time series. The y axis represented each monitoring parameter anomalies. Each color curve in high position was in the normal condition. The number in y axis were used to separate each parameter in positions. The comparison of figure 5 (a) and (b) showed the in-conformity of anomalous distribution. For example, when the parameter CH_4 was anomalous, other parameters should also be anomalous. If the anomalous data object were filtered out based on partial attributes, some normal data may also be eliminated.

The transformer oil chromatography was given as an example to verify the OPTICS. Four methods were used to calculate the recognition accuracy: 1) our method; 2) traditional methods (including Pauta Rule, Grubbs and Dixon), as shown in table 2. Our method improved the recognition accuracy rate by 30%.

Table 2. The comparison of transformer oil chromatogram condition monitoring anomalous data recognition accuracy

Time	Pauta Rule	Grubbs	Dixon	Our method
4 months	0.65	0.5875	0.7208	**0.9833**
8 months	0.6577	0.5676	0.7336	**0.9712**
1 year	0.4315	0.5030	0.6278	**0.9818**

(2) Internal measurement [10].

The structure of power equipment condition monitoring data sets was unknown. The evaluation of clustering results depended on the characteristic and value of the data set. Under the situation, the measurement of anomalous data clustering analysis was referred to the tightness and separation degree. In addition, the size of a single cluster should also be considered. The minimization of inner-cluster distance was achieved by variance within the cluster. Based on the anomalous data filtering algorithm, the historical data curve of the parameters showed better smoothness and stability in figure 6.

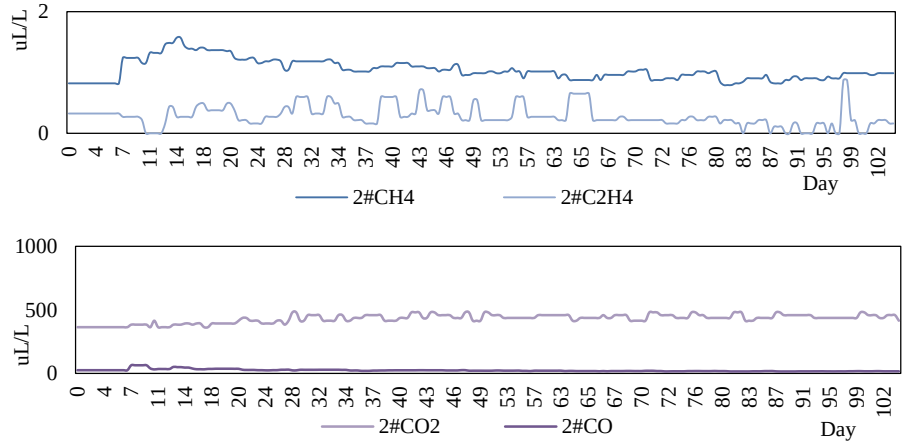


Fig. 6. Historical data curve of 2# substation transformer oil chromatogram monitoring parameters based on OPTICS anomalous data filtering algorithm

(3) Special case of anomalous analysis.

A group of typical anomalous data objects were taken for further analysis of the OPTICS anomalous data filtering algorithm. Checking by the professional electric power equipment maintainers, the anomalous data points of one-year oil chromatogram condition were listed in table 3.

Table 3. Typical anomalous samples of transformer oil chromatogram gas component condition monitoring data

Data Object Id	H_2 (uL/L)	CH_4 (uL/L)	C_2H_6 (uL/L)	C_2H_4 (uL/L)	C_2H_2 (uL/L)	CO (uL/L)	CO_2 (uL/L)	Total Hydro-carbon (uL/L)
396	4.375	0.959	0	0.217	0	17.259	0	1.176
399	3.858	0.847	0	0	0	14.811	0	0.847
439	3.858	0.931	0	0	0	14.405	0	0.931
440	3.898	0.903	0	0.162	0	14.675	0	1.065

In theory, the value of H2, CO, CO_2 in table 3, which were real data in operation period, should not be zero. However, the acquisition data value was zero. The traditional methods such as threshold value and the statistical magnitude failed to recognize the zero value as outliers. Our method used OPTICS clustering to sort and obtain the density of the certain data object: {0.0283, 0, 0.0061, 0, 0.1012, 0.5212, 0, 0.0344}.

We compared the sample data object with the anomalous object {0,0,0,0,0,0,0,0} and the normal object {0.0021,0,0.0006,0.000001,0.0088,0.0437,0.9419,0.0026} generated in formal clustering. The similar distance between the sample object and the anomalous object was 0.5329, while the similar distance between the sample object and the normal object was 1.1255. It was obvious that the sample was closer to the anomalous core object, so the sample data would be classified into the anomalous core cluster.

4.2 GIS density and moisture anomalous data analysis

The excessive moisture content and the gradual decline of density may influence the electrical performance of the high voltage electrical equipment, such as GIS. Meanwhile, the working condition of gas insulated bus-bar was directly related with inner-temperature [11]. So the moisture, gas density, temperature and pressure were the main condition monitoring parameters of GIS. The data structure of GIS was similar with transformer oil chromatography. We also use the OPTICS, setting ε =2, MinPts =5, to analyze the anomalous condition of GIS equipment.

Table 4. GIS SF6 density micro water anomaly data recognition accuracy rate

Time	Pauta Rule	Grubbs	Dixon	Our Method
4 months	1	0.519 2	0.798 0	0.663 4
8 months	0.915 5	0.8	0.853 3	0.844 4
1 year	0.997 0	0.714 2	0.746 3	0.831 3

Table 5. Samples of GIF SF6 moisture data normalized OPTICS core objects

Cluster id	Quantity of objects	Temperature	Density	Pressure under 20° C	Dew point	Moisture
0	69	0.068	0.082	0.013	-0.43	0.378
1	163	0.063	0.103	0.002	-0.6	0.227
2	168	0.045	0.125	0.002	-0.74	0.089

As shown in table 4 and table 5, the anomalous data of moisture based on OPTICS algorithm, compared with the Pauta Rule, Grubbs test, and Dixon test, was different from transformer oil chromatogram results. The recognition accuracy decreased at first, and then increased. It was found that the temperature sensor of GIS SF6 contained anomalous data filtering mechanism. In first four months, the data collected from sensors were in the normal range. After 8 months, the growth of anomalous data was caused by communication failure.

5 Conclusion

This paper proposed the OPTICS-based anomalous data filtering algorithm for condition monitoring of electric power equipment. Our method was compared with threshold method and statistical method in real-data experiments.

Experimental results showed that our method has effectively filtered out the anomalous data of power equipment condition monitoring. The average accuracy of anomalous recognition was 87%.

Some anomalous data samples, which were too difficult to be identified by traditional methods, were catched by our method. So the proposed algorithm showed better effect than traditional methods. It was believed to be more suitable for multidimensional condition monitoring data collection.

In addition, the parameter ε in OPTICS was often set by experience. Normally, the sample data contained both normal data and anomalous data. However, if there was no anomalous data in the sample data, the OPTICS algorithm may produce an average distribution of the ordered density core objects.

In further work, we will keep on optimizing the abnormal data filtering strategy of the power condition monitoring equipment, and carry out further comparison experiments with more clustering algorithms.

References

1. Production and Technology Department of State Grid. Q/GDW169-2008 Guide for condition evaluation of oil-immersed power transformers(reactors) [S]. Beijing: China Electric Power Press,2008(in Chinese).

2. State Economic and Trade Commission. DL/T722—2000 Guide to the analysis and the diagnosis of gases dissolved in transformer oil[S]. Beijing：China Electric Power Press,2001(in Chinese).
3. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). Evangelos Simoudis, Jiawei Han, Usama M. Fayyad, eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9
4. ZHANG Sheng-zhen, YAO Jing-qi. Research on Abnormal Condition Monitoring Data Filtering and Alarm Mechanism [J]. Electric Power Information and Communication Technology, 2013, (1).
5. CHEN Yong-qiang, LIU Kai-pei. A Condition Monitoring Method of Generators Based on RBF Dynamic Threshold Model[J]. Proceedings of the CSEE, 2007, (26):96-101.
6. LI Jiu-long,ZHOU Ling-ke. Detecting and Identifying Gross Errors Based on“3σ Rule”[J]. Computer and Modernization, 2012, (1).
7. Osorio, F., G.A. Paula and M. Galea, On estimation and influence diagnostics for the Grubbs' model under heavy-tailed distributions. 2009. p. 1249-1263.
8. McDaniel W C, Dixon W J, Walls J T. Examination Of Dixon's Up And Down Method For Small Samples In The Estimation Of The ED50 For Defibrillation[C]. //Engineering in Medicine and Biology Society, Vol, 13:, Proceedings of the Annual International Conference of the IEEE. IEEE, 1991:760 - 761.
9. ZHANG Xuan, CHEN Mingxi, XIAO Fengping. Origin Used in Comparison the Methods of Eliminating the Excrescent Data [J]. Experiment Science & Technology, 2012, (1):74-76.
10. ZHANG Weijiao, LIU Chunhuang, LI Fangyu. Method of Quality Evaluation for Clustering [J]. Computer Engineering, 2005, (20):10-12.
11. Zhuang Jianhuang, Ke Min, Qin Wei. Research on SF6 Gas Density and Moisture Online Monitoring for High-voltage Apparatus [C]. //Computer Measurement & Control. 2013.
12. Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). OPTICS: Ordering Points To Identify the Clustering Structure. ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.