

PAPER • OPEN ACCESS

## On-Line Fault Diagnosis Method for Power Transformer Based on Missing Data Repair

To cite this article: Xiansi Lou *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **472** 012027

View the [article online](#) for updates and enhancements.

# On-Line Fault Diagnosis Method for Power Transformer Based on Missing Data Repair

**Xiansi Lou<sup>1</sup>, Weihan Liao<sup>1</sup>, Jianbo Xin<sup>2</sup>, Qiukuan Zhou<sup>2</sup>, Chen Kang<sup>2</sup>, Shiyong Ma<sup>3</sup> and Dunwen Song<sup>3</sup>**

1. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China  
Email: louxiansi@zju.edu.cn, 449057978@qq.com

2. Electric Power Research Institute of State Grid Jiangxi Electric Power Company, Nanchang 330006, China

Email: mandyzhuhai@163.com, zqk0791@163.com, kangchen-123@163.com

3. China Electric Power Research Institute, Beijing 100192, China

Email: abde@epri.sgcc.com.cn, 719528304@qq.com

**Abstract.** Data quality is an important factor affecting the accuracy of transformer fault diagnosis. In order to reduce the impact of missing data, an on-line fault diagnosis method using a loop iterations of improved k-Nearest Neighbour (kNN) and multi-class SVMs based on the missing data repair is proposed in this paper. In the kNN method, the improved Manhattan distance weighted by the negative exponent of the correlation coefficient is designed to measure the distance between samples. On one hand, the influence of the strong correlation indicators on the missing data can be highlighted to improve the accuracy of data repair. On the other hand, the improved Manhattan distance is suitable for an efficient search strategy based on the k-d tree which can achieve the fast search for massive historical data and meet the real-time demand of on-line diagnosis. Diagnosis test results show that the proposed method can keep the high diagnostic accuracy on the incomplete data and realize the efficient on-line fault diagnosis for transformers.

## 1. Introduction

The transformer is an important equipment and core component of the entire substation system and even the power grid which is responsible for the conversion and transmission of electrical energy. Once the transformer especially the oil-immersed main transformer fails, it will affect the safe and stable operation of the power grid and lead to severe economic losses. Dissolved gas analysis (DGA) is the widely recognized method to detect latent faults of transformers [1-3]. With the development of smart substations, oil chromatographic online monitoring devices have become widely applied in power systems [4]. However, due to the constraints of objective conditions such as network topology, geographical environment, and economic factors, the most transformer on-line monitoring systems still use low-quality communication modes like power distribution carrier, Zigbee wireless technology, and industrial distribution. These low-cost channels often occur broken codes under the attack of over-voltage, high-current and other strong electromagnetic interference resulting the loss of monitoring data [5]. The incompleteness of oil chromatographic data will seriously affect the accuracy of transformer fault diagnosis, and even lead to unavailability of traditional methods. Therefore, how to use the incomplete data to achieve accurate and fast online diagnosis for transformer faults has attracted wide attention in academic and industrial circles.



At present the fault diagnosis of transformers based on the incomplete data has two main ideas. The first idea is to clean the bad data and then use existing models such as neural network [6-8], Bayes classifiers [9-10] and support vector machines [11-14] to classify the failure type. There are some research results on the methods of cleaning transformer online monitoring data. The literature [15] treats the missing values as one type of outliers and a time-series analysis method is utilized to identify the data anomaly patterns. Then the autoregressive integrated moving average model (ARIMA) model is adopted to fit the data curve. Literature [5] proposes a method based on Pearson correlation coefficient and regression model to achieve the recovery of missing data. In literature [16] the wavelet neural network model is used to predict the missing data and correct the error data. And the effect of data cleaning can be improved by modifying the wavelet neural network parameters and combined prediction. The above data cleaning methods are all based on the time series analysis. When the missing data is in the middle of the time series, they will have satisfying performance. However, in the actual situation, the missing data is generally at the end of the time series. In addition, there are few literatures discussing the computational efficiency of the cleaning method under the large sample of oil chromatographic data.

Another idea is to select some flexible methods to deal with the missing data in the monitoring results. The rough set theory is a typical method. The literature [17] uses the rough set to reduce the transformer state information, and diagnoses transformer faults based on the reduction set. A transformer fault diagnosis method combining with Bayesian network classifier and rough set is proposed in [18]. This method has the fault-tolerance characteristics, and its performance is obviously better than using Bayesian classifier or rough set separately. The advantages of the rough set theory is that it can effectively analyse and deal with the inaccurate and incomplete data and extract implicit knowledge from a large amount of information. Nevertheless, when there is the lack of key characteristic indicators, the accuracy of such fault diagnosis method will rapidly decline.

Aiming at the deficiencies in the above studies, this paper proposes an online fault diagnosis method for the incomplete oil chromatographic data based on an improved k-Nearest Neighbour (kNN) and multi-class SVMs. This method which designs a loop iteration mechanism for data repair and fault diagnosis inherits the first idea. The reminder of this paper is structured as follow: section 2 introduces a data repair method based on improved kNN, fast search strategy is described in section 3, a cycle iterative diagnosis method is presented in section 4, case study on 309 sets of real oil chromatographic data is shown in section 5, followed by conclusions in section 6.

## 2. Data Repair Method based on Improved k-Nearest Neighbour

The k-Nearest Neighbour (kNN) algorithm is a supervised learning method for finding samples with the similar data pattern. This paper is based on a basic idea that the same transformer fault has a similar data pattern. An improved kNN method is proposed to repair the missing transformer oil chromatographic data. The working mechanism of this method is as follows: 1) input the incomplete test sample, 2) find the  $k$  nearest training samples in the history oil chromatographic database, 3) synthesize the  $k$  samples' information to estimate the missing data. In order to highlight the correlation between different oil chromatographic indicators, a modified Manhattan distance using the negative exponent of the correlation coefficient as the weight coefficient is adopted to measure the distance between two samples.

Assume that there are  $N$  sets of transformer oil chromatographic history data as training samples, and each sample  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$  has  $m$  characteristic gas indicators. The correlation coefficient  $\rho_{pq}$  between indicator  $p$  and  $q$  can be calculated as follow.

$$\rho_{pq} = \frac{\frac{1}{N-1} \sum_{i=1}^N (x_{ip} - \mu_p)(x_{iq} - \mu_q)}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_{ip} - \mu_p)^2} \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_{iq} - \mu_q)^2}} \quad (1)$$

Where  $\mu$  is the mean value of the corresponding characteristic gas indicators. The range of the correlation coefficient is  $[-1, 1]$  where  $-1$  ( $1$ ) denotes strong and negative (positive) correlations between the two characteristic gas indicators and  $0$  means no correlation. If two characteristic gas indicators have the close correlation, it is more credible to use one of them to estimate the other.

When the test sample  $x_j$  lacks the  $p$ th ( $1 \leq p \leq m$ ) gas indicator  $x_{jp}$ , the improved Manhattan distance between the test sample  $j$  and the training sample  $i$  is formulated as follow:

$$d_{ij|p} = \sum_{q=1}^m e^{-|\rho_{pq}|} \cdot |x_{iq} - x_{jq}| \quad q \neq p \quad (2)$$

By weighting the spatial distance of the gas indicators with the negative exponent of correlation coefficient, the correlation between the gas indicators can be reflected. Therefore, the test sample is more likely to approach the training sample which is closed on the strong correlation indicator. Select the  $k$  nearest training samples and calculate the mean of the corresponding gas indicators as the estimated value of the missing data:

$$\hat{x}_{jp} = \frac{\sum_{i=1}^k x_{ip}}{k} \quad (3)$$

Compared with the traditional stochastic filling or time series data repair methods, the improved kNN method considers the similarity of the oil chromatographic data pattern for the same transformer fault type. And due to breaking the temporal correlation of oil chromatographic indicators, it is suitable for the repair of missing data in the event of rapid changes during the development of faults.

### 3. Fast Sample Search Strategy Based on K-D tree

With the wide utilization of online monitoring devices for transformer oil chromatography, the volume of the monitoring data has increased dramatically. Taking a provincial power grid for example, the number of oil-immersed transformers above 220 kV is over 2,000. Millions of oil chromatographic data are produced every year, reaching the GB level of data volume. If the traditional linear scanning method is still used to find the  $k$  nearest training samples from the massive historical database, the calculation efficiency and operating speed will not meet the on-line requirements. The improved Manhattan distance presented in section 2 is suitable for a fast sample search strategy based on k-d tree.

#### 3.1. Construction of K-D tree

Each oil chromatographic sample usually contains 5-8 indicators, so an oil chromatographic sample set corresponds to a high-dimensional data space. The k-d tree, as a data structure for partitioning k-dimensional data space, can be applied to the search of key data in high-dimensional space. Figure 1 illustrates the structure and generation process of a k-d tree with  $k = 3$ .

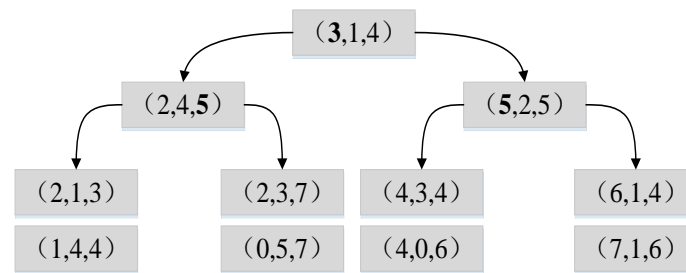
The generation process of a k-d tree includes the following three main steps:

Step1: Calculate variances of samples for each dimension. Select the dimension corresponding to the maximum value of the variance as the split field for effectively guaranteeing the balance of the tree.

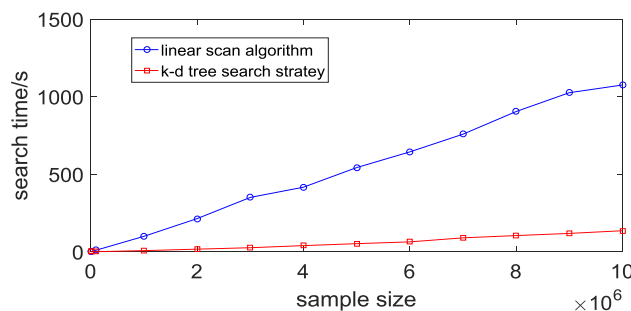
Step 2: Sort the sample according to the value of the split dimension, and the intermediate data point is selected as the pivot point which is represented in bold in Figure 1.

Step3: The samples are divided by the pivot point to form the left and right k-d subtrees. When the size of the subtree is smaller than the acceptable number of leaf nodes, the operation of generating the k-d tree is terminated otherwise return to Step 1. In Figure 1, the acceptable number of leaf nodes is 2.

When this method is applied to the structured storage for transformer oil chromatographic data, each sample is stored in one node of the k-d tree containing  $k$  indicators.



**Figure 1.** The structure of a three-dimensional k-d tree



**Figure 2.** The comparison of algorithm performances for different search strategy

### 3.2. Neighbour Search Strategy

The data structure based on k-d tree can achieve efficient neighbour search for oil chromatographic sample data. The literature [19] has proved that the complexity of neighbour search using k-d tree is  $O(\log n)$ , where  $n$  is the number of samples. When applied to the search of massive high-dimensional samples, it has significant advantages compared with the linear scan algorithm with the complexity of  $O(n)$ . The neighbour search strategy based on k-d tree structure contains the following three steps.

Step1: Find the leaf node where the test sample is located along the k-d tree, get the nearest sample, and add all of the nodes on the search path to a queue.

Step2: Make a hypersphere around the nearest sample with radius of the distance between the test sample and the nearest sample. Verify whether the sample in the queue is within this hypersphere. If not, remove the sample from the queue and repeat Step 2 until the queue is empty, otherwise go to Step 3.

Step3: Update the nearest sample, add its child node in the queue, and return to Step 2.

Figure 2 presents the relationship between search time and sample size of the oil chromatographic data containing five indicators using the linear scan algorithm and the k-d tree search strategy for the improved Manhattan distance defined in section 2. The test results show that when the k-d tree search strategy is used, the search time for neighbour samples increases slowly with the expansion of the sample size, which has a significant advantage comparing with the linear scan algorithm.

## 4. Cycle Iterative Diagnosis Method Based on kNN and SVMs

### 4.1. Design of Multi-class SVMs

The basic idea of the SVM algorithm is mapping the linearly inseparable sample data to high-dimensional feature space through a kernel function, which may be constructed as a linearly separable problem. By solving a convex optimization problem, the optimal hyper plane is obtained which can be utilized to achieve data classification. Since the selection of the optimal hyper plane only depends on support vectors, this method is robust and has a good generalization performance when applied to the transformer fault diagnosis. A single SVM can only solve a binary classification problem. In this paper, a multi-classifier containing 8 SVMs is designed to achieve accurate diagnosis for 6 main fault types. The structure of multi-class SVM classifier is shown in the figure 3.

#### 4.2. Cycle Iteration Fault Diagnosis Process

In the transformer fault diagnosis method based on incomplete data, the kNN method is used for the repair of missing data, and the multi-class SVMs classifier performs fault diagnosis based on the repaired sample. According to the results of the diagnosis, the sample space is reduced. And the missing data is re-estimated using the new  $k$  nearest neighbour samples. This iteration is cycled until the fault type of the neighbour sample is consistent with the fault diagnosis result of the test sample. The fault diagnosis process based on the missing data repair is shown in Figure 4, which specifically includes the following five steps:

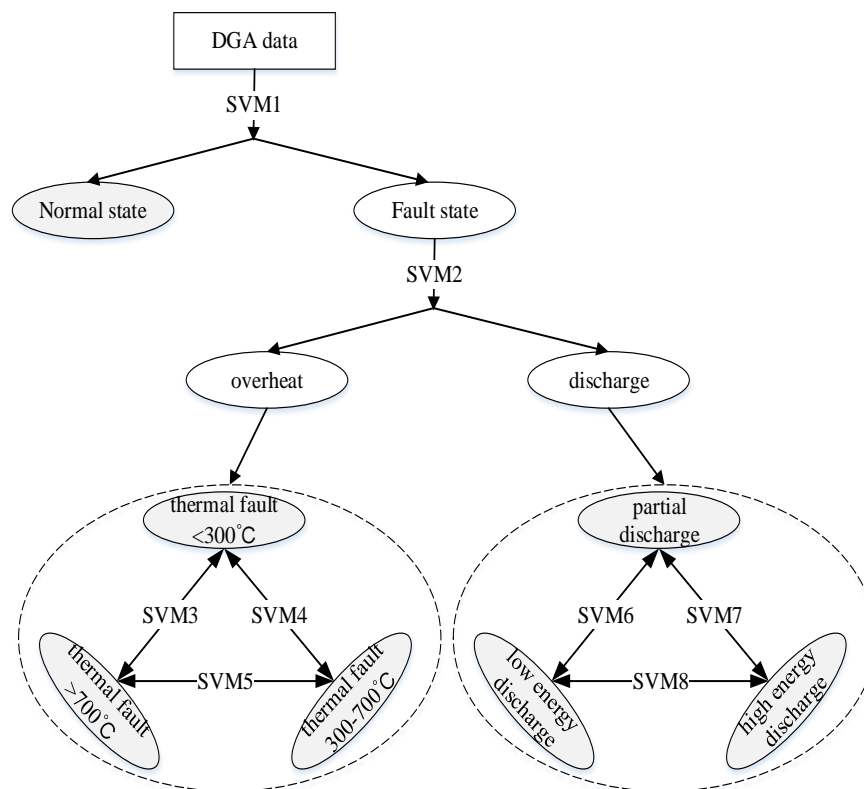
Step1: Input the oil chromatographic data. If there is a missing indicator, apply the kNN method to repair.

Step2: Use the multi-class SVMs model for fault diagnosis.

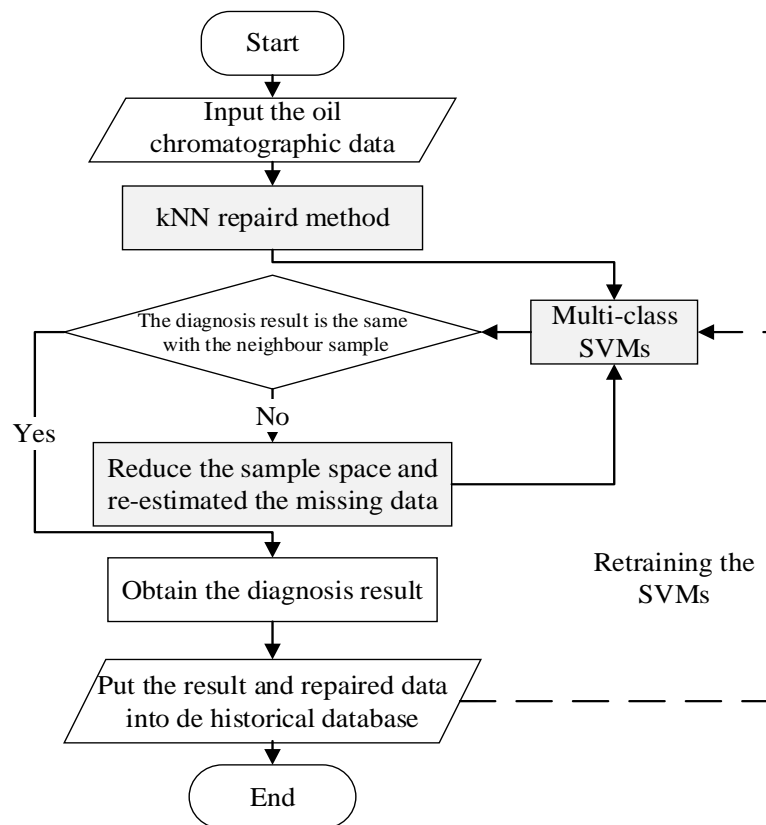
Step3: Judge whether the diagnosis result is consistent with the fault types of the  $k$  nearest training samples. If there is any inconsistency, go to Step 4, otherwise go to Step 5.

Step4: Select the historical samples which have the same fault type with the multi-class SVMs's diagnostic results as the new sample space. And re-use the kNN method to estimate the missing data. Go back to Step 2.

Step5: Obtain the fault diagnosis result, store the repaired data and diagnostic result in the historical database, and retrain the SVM multiple classifier periodically.



**Figure 3.** The fault diagnosis model for transformer based on multi-class SVMs



**Figure 4.** The fault diagnosis flow for transformer based on missing data repair

## 5. Case Study

In the case study, 309 sets of complete transformer oil chromatographic sample with the identified fault type, including thermal fault at  $<300^{\circ}\text{C}$  (T1), thermal fault at  $300^{\circ}\text{C}$ - $700^{\circ}\text{C}$  (T2), thermal fault at  $>700^{\circ}\text{C}$  (T3), partial discharge (PD), low energy discharge D1, and high energy discharge (D2), has been collected. And the incomplete data samples are constructed by random deletion. To facilitate results comparison, the leave-one method (LOOM) [20] is used to calculate the diagnostic accuracy rate for three different methods: kNN+ multi-class SVMs (proposed in this paper), ARIMA+ multi-class SVMs and rough set method.

**Table 1.** Comparison of fault diagnosis accuracy by different methods

Fault type	Number of training samples	Multi-class SVMs (complete data)	kNN+SVMs (incomplete data)	ARIMA+SVMs (incomplete data)	Rough set (complete data)	Rough set (incomplete data)
T1	47	85.11%	80.00%	74.47%	72.34%	59.58%
T2	39	87.18%	91.79%	82.05%	79.49%	68.21%
T3	51	76.47%	74.51%	70.59%	58.82%	44.71%
PD	47	74.47%	71.49%	68.09%	63.83%	62.13%
D1	49	87.76%	75.10%	77.55%	71.43%	55.92%
D2	34	79.41%	72.35%	58.82%	70.59%	72.35%
Normal	42	92.86%	87.14%	85.71%	71.43%	57.14%
total	309	83.17%	78.64%	74.11%	69.26%	59.09%

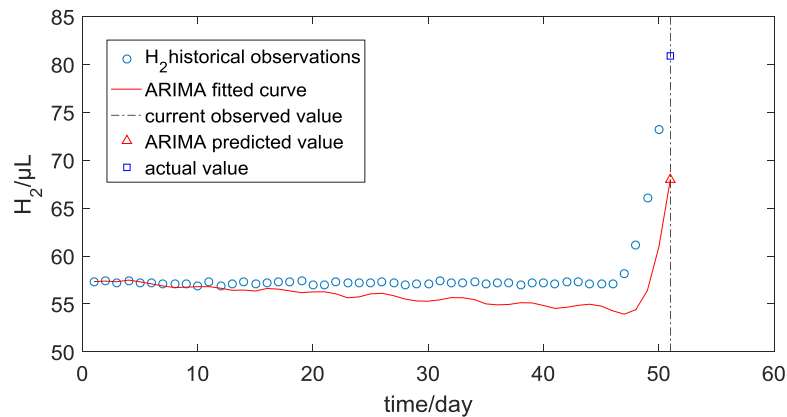
From Table 1, it can be seen that the diagnostic accuracy of the multi-class SVMs and the rough set method based on the complete oil chromatographic data is 83.17% and 69.26% respectively, which

has the equal classifier performance in the literature [21]. When the rough set method is applied to the incomplete data, the diagnostic accuracy is reduced to 59.09%. This is because the gas indicator is determinant attributes which seriously affect the accuracy of fault diagnosis.

Compared with the 3rd, 4th and 5th column data, the diagnostic accuracy dropped from 83.17% to 74.11% by using the ARIMA+ SVMs method. When using the kNN+ SVMs method proposed in this paper, the accuracy rate is 78.64% with only 4.53% reduced. Therefore, the high accuracy of the multi-class SVMs can still be maintained. A specific example is illustrated in the Table 2 and Figure 5 for the further analysis between the ARIMA+ SVMs and kNN+ SVMs.

**Table 2.** Missing data repair using kNN

Incomplete data sample	Neighbour samples	Actual value	Estimated value	Relative error
(*,12.85,3.65,2.6,0)	(83,13.05,2.3,1.65,0)	80.9	82.876	2.44%
	(85.34,11.05,2.82,0.79,0)			
	(84,14.17,0.98,0.85,0)			
	(87.41,10.13,2.3,0.16,0)			
	(74.63,13.73,7.58,4.06,0)			



**Figure 5.** Missing data predict using ARIMA

(\* , 12.85, 3.65, 2.6, 0) is an oil chromatographic sample with five indicators which corresponds to the gas of  $H_2$ ,  $CH_4$ ,  $C_2H_6$ ,  $C_2H_4$ , and  $C_2H_2$  respectively. \* denotes the missing data of  $H_2$ , and the actual value is 80.9 $\mu$ L. The final five closest data samples searched in the historical database by using the kNN method is shown in Table 2. The estimated value of  $H_2$  is 82.867 $\mu$ L with the relative error of 2.44%. The result of the fault diagnosis result using the repaired sample data is thermal fault at <300  $^{\circ}$ C (T1) which is consistent with the actual situation. When the ARIMA method is adopted to fit the history data of the  $H_2$ , the fitted curve is presented in Figure 5 and the predicted value is 68.02 $\mu$ L. Based on the predicted data, the result of the fault diagnosis is thermal fault at 300  $^{\circ}$ C-700  $^{\circ}$ C (T2) which is incorrect.

It can be seen from the Figure 5 that due to the rapid increase of  $H_2$  in the fault development course, the ARIMA method overfits the historical data under the normal condition leading to a conservative final predicted result. Compared to the predicted method based on the timing characteristics, the kNN method is a global related search strategy that breaks the mandatory association relationship of oil chromatographic data from the perspective of time. Therefore, it has an obvious advantage when applied to the rapid development of faults.

## 6. Conclusion

With the popularization and application of the transformer on-line monitoring system, the quality of monitoring data has gradually attracted people's attention. For the incompleteness of oil chromatographic monitoring data, the traditional transformer fault diagnosis method has a rapid



decline in accuracy or even cannot be applied. In this paper, an online transformer fault diagnosis method based on missing data repair is presented and the main innovations are as follows:

1) A fault diagnosis method based on the loop iteration of kNN and multi-class SVMs is proposed, which improves the accuracy of transformer fault diagnosis based on incomplete data.

2) An improved Manhattan distance with the negative exponent of the correlation coefficient as the weight is used to measure the distance between samples. Therefore, the influence of the strong correlation indicators can be highlighted.

3) The k-d tree-based data structure is applied to speed up the search in massive historical samples to meet the actual needs of transformer on-line fault diagnosis.

The results of the case study show that the proposed method can realize on-line transformer fault diagnosis based on incomplete data and adapt to the trend of power transmission equipment big data. Compared with the methods based on rough set theory or time series data prediction, the proposed method still has a higher diagnostic accuracy rate in the incompleteness of the key indicator or the situation of rapid development of faults. But there are also some limitations. This method is only applicable to the repair of missing data, but it does not have the ability to identify the wrong data. At the same time, more monitoring data and maintenance records should be used to evaluate the dynamic state of transformers which is also the direction of work for further research.

### Acknowledgement

Project Supported by Electric Power Research Institute of State Grid Jiangxi Electric Power Company (52182016001J) and China Electric Power Research Institute (XT71-16-053).

### References

- [1] GAO Jun, HE Junjia. Application of quantum genetic ANNs in transformer dissolved gas-in-oil analysis 2010 *J. Proceedings of the CSEE* **30(30)** pp 121-127.
- [2] ZHANG Dong-bo, XU Yu, WANG Yao-nan. Neural network ensemble method and its application in DGA fault diagnosis of power transformer on the basis of active diverse learning 2010 *J. Proceedings of the CSEE* **30(22)** pp 64-70.
- [3] Zheng H B, Liao R J, Grzybowski S, et al. Fault diagnosis of power transformers using multi-class least square support vector machines classifiers with particle swarm optimisation 2011 *J. IET Electr. Power Appl* **5(9)** pp 691-696.
- [4] LIANG Yongliang, LI Kejun, ZHAO Jianguo, et al. Research on the dynamic monitoring cycle adjustment strategy of transformer chromatography on-line monitoring devices 2014 *J. Proceedings of the CSEE* **34(9)** pp 1446-1453.
- [5] LIU Yuankun, LUAN Wenpeng, XU Yu, et al. Data cleaning method for distribution transformer 2017 *J. Power System Technology* **41(3)** pp1008-1014.
- [6] LIU Na, GAO Wensheng, TAN Kexiong. Fault diagnosis of power transformer using a combinatorial neural network 2003 *J. Transactions of China Electrotechnical Society* **18(12)** pp 11-16.
- [7] ZHANG Jianguang, ZHOU Hao, XIANG Canfang. Application of super SAB ANN model for transformer fault diagnosis 2004 *J. Transactions of China Electrotechnical Society* **19 (7)** pp 49-52.
- [8] ZANG Hongzhi, HU Yuhua, YU Xiaodong. Integrated ANN based on radial basis function applied in transformer fault diagnosis 2003 *J. Proceedings of the EPSA* **15(1)** pp 51-53.
- [9] WU Lizeng, ZHU Yongli, YUAN Jinsha. Novel method for transformer fault integrated diagnosis based on Bayesian network classifier 2005 *J. Transactions of China Electrotechnical Society* **20(4)** pp 45-51.
- [10] WANG Yongqiang, LU Langcheng, LI Heming. Intelligent fault diagnosis for power transformer based on bayesian network and DGA 2004 *J. High Voltage Engineering* **30(5)** pp 12-13.
- [11] TAO Xinmin, LI Zhen, LIU Furong, et al. Fault detection method for power transformer based on SVM using reduced vector set 2016 *J. High Voltage Engineering* **42(10)** pp 3199-3206.
- [12] WU Xiaohui, LIU Jiong, LIANG Yongchun, et al. Application of support vector machine in transformer fault diagnosis 2007 *J. Journal of Xi'an JiaoTong University* **41(6)** pp 457-457.

- [13] BI Jianquan, LU Mingming, GUO Chuangxin, et al. A transformer fault diagnosis method based on multi-classified probability output 2015 *J. Automation Electric Power System* **39(5)** pp 88-93.
- [14] GUO Chuangxin, ZHU Chengzhi, ZHANG Lin, et al. A fault diagnosis method for power transformer based on multiclass multiple-kernel learning support vector machine 2010 *J. Proceedings of the CSEE* **30(13)** pp 128-134.
- [15] YAN Yingjie, SHENG Gehao, CHEN Yufeng, et al. Cleaning method for big data of power transmission and transformation equipment state based on time sequence analysis 2015 *J. Automation Electric Power System* **7** pp 138-144.
- [16] LIN Jun, YAN Yingjie, SHENG Gehao, et al. Online monitoring data cleaning of transformer considering time series correlation 2017 *J. Power System Technology* **41(11)** pp 3733-3740.
- [17] MO Juan, WANG Xue, DONG Xing, et al. Diagnosis model of insulation faults in power equipment based on rough set theory 2004 *J. Proceedings of CSEE* **24(7)** pp 162-167.
- [18] ZHU Yongli, WU Lizeng, LI Xueyu. Synthesized diagnosis on transformer faults based on Bayesian classifier and rough set 2005 *J. Proceedings of the CSEE* **25(10)** pp 159-165.
- [19] Bentley, J. L. Multidimensional binary search trees used for associative searching 1975 *J. Communications of the ACM* **18(9)** pp 509-517.
- [20] Chang C S, Jin J, Chang C, et al. Online source recognition of partial discharge for gas insulated substations using independent component analysis 2006 *J. IEEE Transactions on Dielectrics and Electrical Insulation* **13(4)** pp 892-1002.
- [21] WANG Ke, LI Jinzhong, ZHANG Shuqi, et al. New features derived from dissolved gas analysis for fault diagnosis of power transformers 2016 *J. Proceedings of the CSEE* **36(23)** pp 6570-6579.