Available online at www.sciencedirect.com
ScienceDirect
Energy Reports 9 (2023) 698–711
www.elsevier.com/locate/egyr

2022 International Conference on Frontiers of Energy and Environment Engineering, CFEEE 2022, 16–18 December, 2022, Beihai, China

Outlier detection and data filling based on KNN and LOF for power transformer operation data classification

Dexu Zou[a], Yongjian Xiang[b], Tao Zhou[b,*], Qingjun Peng[a], Weiju Dai[a], Zhihu Hong[a], Yong Shi[c], Shan Wang[a], Jianhua Yin[d], Hao Quan[b]

[a] *Electric Power Research Institute, China Southern Power Grid Yunnan Power Grid Co., Ltd., Kunming 650217, China*
[b] *School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China*
[c] *China Southern Power Grid Yunnan Power Grid Co., Ltd., Kunming 650217, China*
[d] *School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China*

Received 30 March 2023; accepted 8 April 2023
Available online 20 April 2023

Abstract

The missing and abnormal data in power transformer operation and monitoring greatly affect the accuracy of fault diagnosis and thus threaten the stable operation of power systems. To conduct outlier detection and improve data quality for safety warning, this paper proposes a transformer operation data preprocessing method based on KNN (K-nearest neighbor) and LOF (local outlier factor) for power transformer operation data classification. Firstly, this paper analyzes the characteristics of transformer operation data. Secondly, the local reachable density of the input data is calculated by LOF algorithm. The local outlier factor score of the data is derived according to the local reachable density, and the abnormal data is output according to the abnormal score. Then, KNN algorithm is utilized to classify the relevant data around the abnormal value and missing value of the transformer. The data are filled or corrected according to the classification results. Thirdly, the elbow method is used to determine the optimal K value and cluster operation data by K-Means algorithm. Finally, the proposed method is applied and verified with real transformer operation data in case study. The results show the method can effectively detect and correct the abnormal and missing data, conduct transformer data cleaning and preprocessing and provide accurate and effective data samples for transformer fault diagnosis.

Peer-review under responsibility of the scientific committee of the 2022 International Conference on Frontiers of Energy and Environment Engineering, CFEEE, 2022.

Keywords: Power transformer; Outlier detection; Data sufficiency; LOF; KNN

1. Introduction

Power transformer is one of the most important equipment in the power systems. If the transformer fails to be diagnosed and treated in time, it will bring huge property losses and safety hazards [1,2]. With the continuous

* Corresponding author.
E-mail address: zhoutaonjust@njust.edu.cn (T. Zhou).

https://doi.org/10.1016/j.egyr.2023.04.094

Peer-review under responsibility of the scientific committee of the 2022 International Conference on Frontiers of Energy and Environment Engineering, CFEEE, 2022.

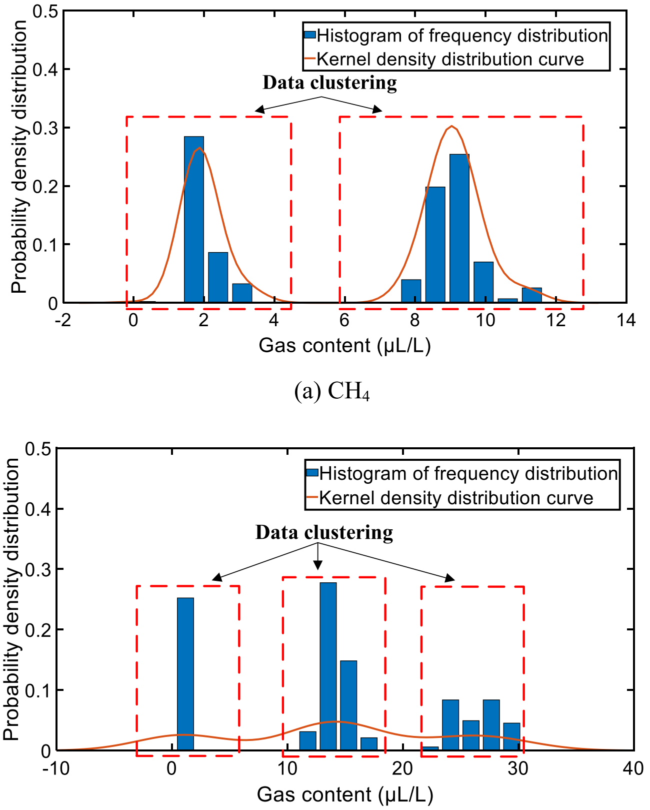
development and enrichment of power equipment sensors, communication devices and storage devices, the detected data becomes more complex and high-dimensional [3]. And because of the transformer running state, different environment, communication line loss and other reasons, the data obtained by the sensor has some missing and mixed with noise. This will lead to large errors in transformer fault diagnosis. Cleaning transformer data, detecting abnormal data, filling abnormal data and missing data are very important for the construction of smart grid and the safe and stable operation of power equipment [4,5].

In terms of transformer data filling, the traditional method uses special values such as mode, average value, and maximum value to fill missing values. This method has large errors and low accuracy [6,7]. In order to improve the accuracy of data cleaning and ensure the accuracy of fault diagnosis, some scholars have introduced machine learning into the fault diagnosis and data management of power equipment [8,9]. Some scholars have introduced the isolated forest method into power transformer fault diagnosis and data management, but when the transformer abnormal data is less, the effect of this method is not ideal, and the error is large [10]. With the development of intelligent algorithms, some scholars use multiple difference compensation method to supplement the missing value of transformer. Since multiple imputation is to first fill in missing values randomly in the missing data, and then based on the filled data to other column data feature iteration, but in the case of high latitude amounts of data, the effect of this scheme is not the best [11–13]. Some scholars have introduced GRNN into the filling of transformer data, which improves the accuracy of filling. But this method requires a complete data set to learn [14]. In the detection of abnormal values of transformer data, the detection of abnormal data such as dissolved gas in transformer oil and winding state based on frequency response is judged by directly setting a threshold. This method is simple but has a large error [15]. The traditional method uses partial discharge detection technology to detect partial discharge in transformers. This method has low intelligence, relies on experience, and has low reliability. With the development of intelligent technology, outlier detection based on linear model, PCA (principal component analysis) and other technologies are used in outlier detection, but this method has a large error in dealing with nonlinear data [16–18]. Some scholars propose to use k-means clustering algorithm to detect outliers, but this method cannot accurately determine outliers according to specific data types [19]. Using normal data and images to train YOLOv3 model to detect abnormal parts of power transformers, but this method does not preprocess the data well and the error is relatively large [20]. In terms of transformer data clustering, fuzzy clustering analysis is applied to transformer fault diagnosis, but the method is simple and its accuracy is relatively low [21,22]. In the transformer data preprocessing, the traditional method of data filling error is large, and after detecting the data directly to delete abnormal data, the destruction of data continuity and integrity.

To conduct outlier detection and improve data quality for safety warning, this paper proposes a transformer operation data preprocessing method based on KNN and LOF. Firstly, the local reachable density of the input data is calculated by LOF algorithm. The local outlier factor score of the data is derived according to the local reachable density, and the abnormal data is output according to the abnormal score. Secondly, KNN algorithm is utilized to classify the relevant data around the abnormal value and missing value of the transformer. The data are filled or corrected according to the classification results. Finally, the elbow method is used to determine the optimal K value and cluster operation data by K-Means algorithm. KNN algorithm can identify the type of outliers and consider its correlation with the surrounding data, so it has high accuracy in filling data. LOF algorithm can be used to quantify the abnormal data, and the local correlation of transformer data is considered. By using the combination of LOF algorithm and KNN algorithm, the time series characteristics of transformer data are considered, the continuity and integrity of data are ensured, and accurate and effective data samples are provided for transformer data clustering and fault diagnosis.

2. Transformer data characteristics analysis

The time series of most state quantities of transformers in normal operation state satisfy the stationarity assumption. However, transformer operation data will be affected by sensor fault, electromagnetic interference and other factors in the process of collection, transmission and storage. The original data will inevitably produce concept drift and invalid abnormal data. The main factors affecting the quality of transformer operation monitoring data include data drift, data anomaly and data missing, etc. The low-quality data will affect status detection and fault identification. In this section, the transformer operation data and Dissolved Gas Analysis (DGA) data are selected to analyze the characteristics of transformer status.



(a) CH_4

(b) C_2H_2

Fig. 1. Frequency distribution histogram and kernel density estimation curve of transformer oil chromatographic data.

2.1. Data drift

In the dynamic and non-stationary operation environment, the probability distribution of the operation data will change unforeseeably with the passage of time, which will cause the change of the mapping relationship between the input data and the targets. This phenomenon is defined as concept drift. The concept drift will make the state quantity present the complex probability distribution and produce many kinds of normal data pattern.

Taking the DGA online monitoring data of two transformers as an example, 430 CH_4 and 991 C_2H_2 gas content samples were collected. Fig. 1(a) and (b) show the frequency distribution histogram and kernel density distribution function curve distribution of the two gases. There are two clusters in the probability distribution maps of both gases, indicating that there may be two or three types of patterns in the data.

Incremental drift and abrupt drift are two main types of data drift. (1) Incremental drift occurs when transformer status changes, which is characterized by continuous abnormal growth or decline of state monitoring value over time. (2) Abrupt drift often occurs under sensor failure, replacement or calibration, which is characterized by the change of the state detection value in a short period of time in the shape of a step function.

Take the monitoring data of O_2 in the oil chromatography data of a transformer as an example. Fig. 2 shows the measured value of oxygen gas content in a certain period of time. In the red dashed box, there is a sudden change in the oxygen volume fraction at a certain moment, but the data change before and after that moment is relatively stable. Mutational drift can make normal monitoring data appear multimodal characteristics, which may cause the failure of statistically based anomaly detection methods, because the measurement value of the new model may be outside the confidence interval determined by the old model data.

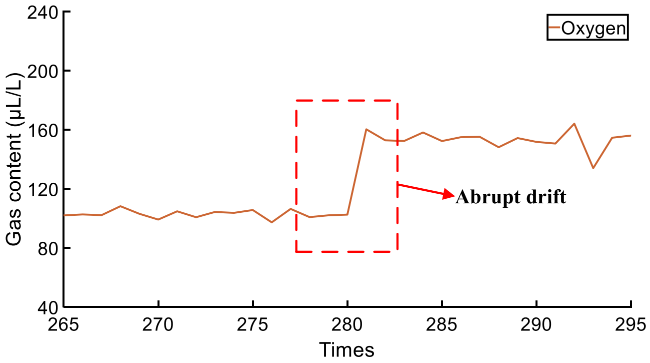


Fig. 2. Abrupt drift.

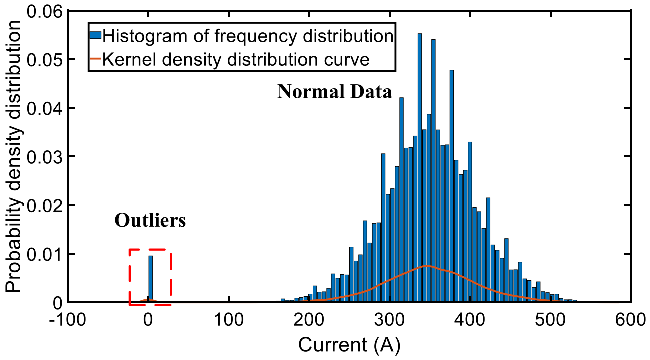


Fig. 3. Data anomaly.

2.2. *Data anomaly*

The definition of outliers is as follows: outliers are different from other regular data in the data set, and these data are not random deviations, but are generated for other completely different reasons. The task of outlier detection is to find data patterns that are inconsistent with the expected behaviors in the running data set, such as outliers, inconsistent observations, deviations and singular values.

Fig. 3 shows the monitoring data of the middle and high voltage side current amplitude of a transformer, a total of 96 308. As can be seen from the figure, the data distribution can be divided into two chunks, and the sample value on the left is significantly far from the normal data, which can be defined as outliers.

2.3. *Data missing*

The missing values of condition monitoring data usually have complex changing rules. According to the characteristics and degree of missing values, the missing values can be divided into three categories: (1) Isolated missing values: in the data segment, a measurement value of a state quantity is missing, but its adjacent test values are intact; (2) Continuously missing variables: in a period of time, the measured values of a certain state quantity are continuously missing, while the test values of other state quantities are relatively complete; (3) Continuous missing samples: in a period of time, multiple classes of state quantities appear continuous missing, which destroys the continuity of the time series and divides the original time series into data segments of different lengths.

Missing values in operation data are very common, which are usually caused by the following reasons: (1) natural factors such as sensor failure, signal transmission interruption and data upload failure; (2) Human factors including invalid data and error value of abnormal data removal. Missing values destroy the continuity and integrity of time series data, and most fault identification methods cannot accept input data with missing values, which

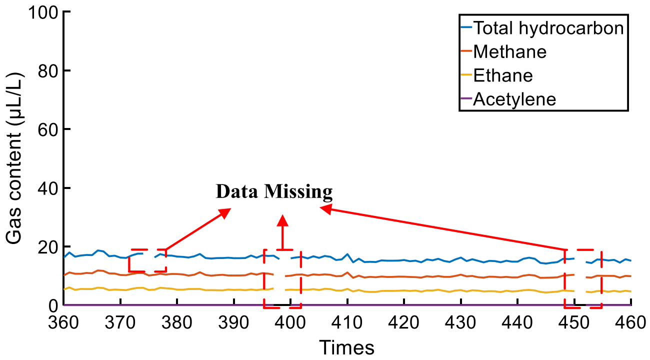


Fig. 4. Data missing.

seriously affects the efficiency of transformer fault identification. Taking the oil chromatographic data of a certain transformer as an example, as shown in Fig. 4, among the monitored state quantities, the state parameter values of total hydrocarbon, methane (CH_4), ethane (C_2H_6) and acetylene (C_2H_2) were missing for a total of 7 times during the monitoring period of about 100 days.

3. Methodology

The quality of transformer data exerts an important impact on the accuracy of transformer fault diagnosis. At present, the data of transformer operation has the characteristics of large volume, high dimension and diversity of data structure [23,24]. These characteristics cause great difficulties for sensors to obtain power equipment data stably and efficiently. The data quality will be affected by the environment during transmission, which will be mixed with a lot of noise and abnormal data [25–27]. Therefore, it is urgent to carry out transformer operation data preprocessing and cleaning. Traditional data filling has mode filling, average filling, multiple interpolation, etc. Each filling method is suitable for specific application scenarios. With the development of artificial intelligence (AI), domestic and foreign scholars introduce machine learning into data cleaning. Therefore, this paper proposes a transformer data preprocessing and cleaning method based on KNN and LOF which can take advantage of AI technology and is universal and suitable for diverse scenarios with a higher accuracy and efficiency.

3.1. KNN algorithm

KNN algorithm is a supervised learning classification algorithm in machine learning. KNN algorithm divides the data set to be filled into complete data set and missing data set. For each missing data, find its nearest K samples in the full sample set. Determine specific types of missing value data based on K sample data types. Finally, according to the distance between each sample and the missing value data to allocate their respective contribution. Because of the strong correlation between adjacent data of transformer, KNN algorithm has a good effect in filling transformer data. The algorithm flow chart is shown in Fig. 5.

The specific implementation process of the algorithm is as follows:

(1) Input transformer data to establish transformer data matrix.

$$X = \{x_1, x_2, \ldots, x_m\} \tag{1}$$

(2) Euclidean distance is used to calculate the distance between missing values and other samples.

$$d = \sqrt{\sum_{r=1}^{n} (x_{ir} - x_{jr})^2} \tag{2}$$

where, x_{ir} is the position of the target missing value, sand x_{jr} is the position of other samples.

(3) In the dataset, find K points closest to the target missing value x_{ir} based on the distance d, and then find the same type of data as the missing data in K points.

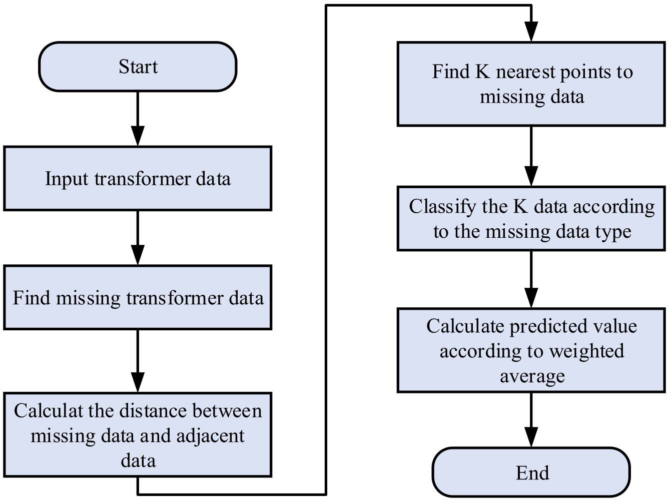

Start
Input transformer data
Find missing transformer data
Calculat the distance between missing data and adjacent data
Find K nearest points to missing data
Classify the K data according to the missing data type
Calculate predicted value according to weighted average
End

Fig. 5. Flow chart of KNN algorithm.

(4) Calculate the value of filling the missing value.

$$x_0 = \sum_{i=1}^{k} \omega_{\mathrm{ir}} x_{ir} \tag{3}$$

where, ω_{ir} is the weight of K nearest neighbor data, which is the reciprocal of distance.

The correlation between adjacent data of transformer is relatively large, KNN algorithm can define its weight according to the distance between the missing value and the surrounding data. Through the above steps, the missing data of the transformer can be well filled.

3.2. LOF algorithm

LOF algorithm is an unsupervised outlier detection method and a high-precision outlier detection algorithm based on data density. It calculates the local reachable density of the data, and calculates the local outlier factor score according to the local reachable density. The LOF algorithm for outlier detection is not determined by the absolute value, but by the density of domain points. Finally, according to different types of data, appropriate thresholds are selected to detect outliers. The algorithm flow chart is shown in Fig. 6.

The specific steps of this algorithm are as follows:

(1) Calculate the distance between each transformer data.

(2) According to the distance between the data, it is arranged in descending order.

(3) For each data point, the nearest K point is found according to the adjacent distance.

(4) Calculate the local reachable density of each point.

$$d_k(P, O) = \max\{d_k(P), d(P, O)\} \tag{4}$$

where, $d_k(P, O)$ is the reachable distance between point P and point O, $d_k(P)$ is the Kth distance of point P.

$$N_k(P) = \{O' \subset D\backslash\{P\} | d(P, O) \leq d_k(P)\} \tag{5}$$

where, $N_k(P)$ is the transformer data point in the K-distance of point P.

$$\rho_k(P) = \frac{|N_k(P)|}{\sum_{O \subset N_k(P)} d_k(P, O)} \tag{6}$$

where, $d_k(P, O)$ is the reachable distance between point P and point O, $\rho_k(P)$ is the local reachable density of point P.

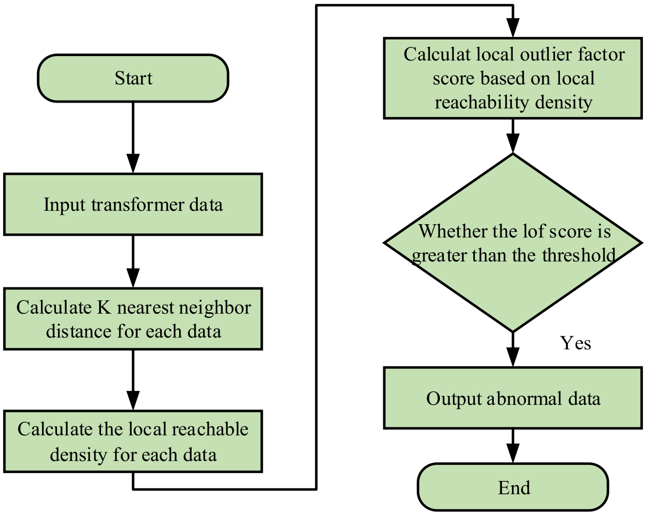

Start
Input transformer data
Calculate K nearest neighbor distance for each data
Calculate the local reachable density for each data
Calculat local outlier factor score based on local reachability density
Whether the lof score is greater than the threshold
Yes
Output abnormal data
End

Fig. 6. LOF algorithm flow chart.

(5) The LOF score of local outlier factor of each point is obtained.

$$LOF_K(P) = \frac{\sum_{O \subset N_k(P)} \frac{\rho_k(O)}{\rho_k(P)}}{|N_k(P)|} \tag{7}$$

where, $LOF_K(P)$ is the local outlier factor score of point P.

If the LOF score is closer to 1, it means that the density of point p is close to its neighbors; If the value is less than 1, it means that the density at point p is greater than the density of neighboring points; If the value is greater than 1, it means that the density at point p is less than the density of neighboring points. The higher the value, the greater the possibility that point p is an outlier.

3.3. K-means clustering algorithm

The K-means algorithm is a clustering algorithm. Its main idea is to assign each point to the cluster represented by the nearest cluster center point, given the K value and K initial cluster center points. After all points are allocated, the center point of a class cluster is recalculated according to all points in the class cluster, and then the steps of allocating points and updating the center point of the class cluster are iterated until the change of the center point of the class cluster is very small or the specified number of iterations is reached. The algorithm flow chart is shown in Fig. 7.

The specific steps of K-means algorithm are as follows:

(1) Select X initialized samples as the initial cluster center. Each sample contains n objects, of which each object has m dimension attributes;

(2) For each sample x_i in the dataset, calculate its distance to k cluster centers and divide it into the class corresponding to the cluster center with the smallest distance. The Euclidean distance from each object to each cluster center is calculated, as shown in the following formula:

$$dis(X_i, C_j) = \sqrt{\sum_{t=1}^{m}(X_{it} - C_{jt})^2} \tag{8}$$

where, x_i represents the ith object, c_j represents the jth cluster center, x_{it} represents the tth attribute of the ith object, c_{it} represents the tth attribute of the ith cluster center.

(3) The distance from each object to each cluster center is compared in turn, and the object is assigned to the cluster of the nearest cluster center to obtain k clusters. The new clustering center is calculated according to the

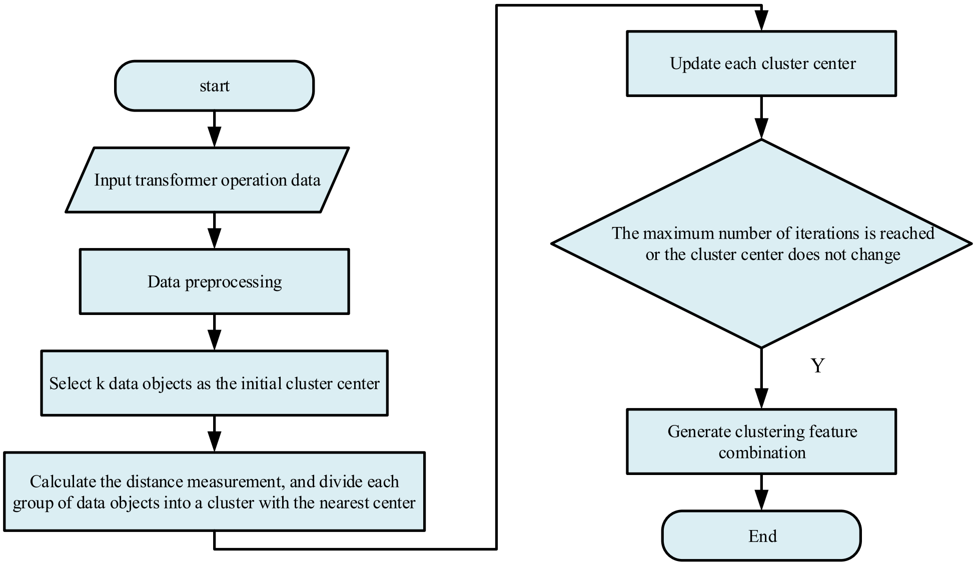

start
Input transformer operation data
Data preprocessing
Select k data objects as the initial cluster center
Calculate the distance measurement, and divide each group of data objects into a cluster with the nearest center
Update each cluster center
The maximum number of iterations is reached or the cluster center does not change
Y
Generate clustering feature combination
End

Fig. 7. K-means algorithm flow chart.

objects in the cluster, that is, the mean value of all objects in each dimension.

$$C_t = \frac{\sum_{X_i \in S_l} X_i}{|S_l|} \tag{9}$$

where, c_t represents the center of the first cluster, s_l represents the number of objects in the first cluster, x_i represents the ith object in the lth cluster.

(4) Repeat steps (2) and (3) until the number of iterations or minimum error changes is reached.

3.4. Overall framework

Firstly, the local reachable density of the input data is calculated by LOF algorithm. The local outlier factor score of the data is derived according to the local reachable density, and the abnormal data is output according to the abnormal score. Then, KNN algorithm is utilized to classify the relevant data around the abnormal value and missing value of the transformer. The data are filled or corrected according to the classification results. The KNN algorithm is used to fill and correct the outliers, and the data type of the outliers is identified. The correlation between the outliers and the surrounding data is considered. The LOF algorithm can quantify the abnormal situation of the data, and the local correlation of the transformer data is considered. By using the combination of LOF algorithm and KNN algorithm, the time series characteristics of transformer data are considered, the continuity and integrity of data are ensured, and accurate and effective data samples are provided for transformer data clustering and fault diagnosis. The implementation process is shown in Fig. 8.

The framework can conduct transformer operation data cleaning and provide more accurate data for transformer data clustering and transformer fault diagnosis.

4. Case study and analysis

The proposed method is applied to the cleaning and preprocessing of practical operation data of a regional power grid. The oil chromatography data and transformer operating current data is taken for example to show the feasibility of the method. The oil chromatographic data include hydrogen, methane, ethane, ethylene, acetylene, total hydrocarbon, carbon monoxide and carbon dioxide. This data set contains a large amount of missing data and abnormal data. This paper uses this data to verify the effectiveness of this strategy.

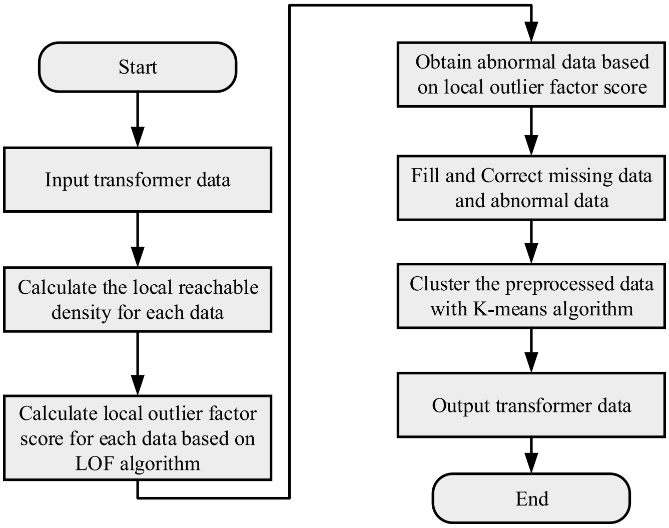

Start
Input transformer data
Calculate the local reachable density for each data
Calculate local outlier factor score for each data based on LOF algorithm
Obtain abnormal data based on local outlier factor score
Fill and Correct missing data and abnormal data
Cluster the preprocessed data with K-means algorithm
Output transformer data
End

Fig. 8. Data preprocessing flow chart.

Table 1. Current data processing results.

A-phase current value	B-phase current value	C-phase current value
134.343 (190.125)	135.711	132.78
135.125 134.538	136.004 135.418	133.073 131.9
138.642	**138.349** (202.137)	136.004
144.504	145.97	142.745
154.47 (90.234)	155.349	153.004
170.591	171.177	167.953
190.523	190.816	**189.058** (112.127)
214.265	215.437	211.041

4.1. Outlier detection

This paper uses the oil chromatographic data and transformer operating current data of a transformer as the sample data set of this paper. The following is a part of the dirty data of the data set, which has abnormal data. The results processed using this policy are in Table 1. The red data is the result of detecting and filling the abnormal data, and the parenthesis is the original abnormal data.

It can be seen from Tables 1 and 2 that the strategy in this paper can eliminate abnormal data according to the set LOF threshold, and use KNN algorithm to supplement abnormal values and missing values. From the specific values in Table 2, we can see that this strategy has a good effect on outlier detection. And its filling value conforms to the change law of transformer data, ensuring the accuracy of data. Fig. 9 shows the score of LOF of CH_4 in the dataset used in this paper.

It can be seen from Fig. 9 that the LOF value of the transformer data usually fluctuates around 1, but the LOF of the outliers will be large. LOF algorithm can select appropriate LOF threshold according to different data types to

Table 2. Oil chromatographic data processing results.

H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	Total hydrocarbon	CO	CO_2
114.34	11.06	3.05	0.31	0.0	14.43	**244.73** (299.59)	3129.93
110.48	11.50	3.45	0.37	0.0	15.32	268.58	**3205.83** (3485.69)
103.19 (132.45)	10.49	3.10	0.31	0.0	13.90	240.27	3152.9
104.46 102.131	10.64 10.41	3.27 3.20	0.35 0.36	0.0 0.0	14.26 14.05	243.24 238.45	3173.03 3153.98
95.66	**10.00** (13.24)	2.96	0.28	0.0	12.93	227.35	2905.12
105.53	10.24	3.06	**0.32** (0.61)	0.0	13.61	238.58	3143.15
101.54	10.37	**3.21** (7.40)	0.32	0.0	13.76	234.62	3217.93

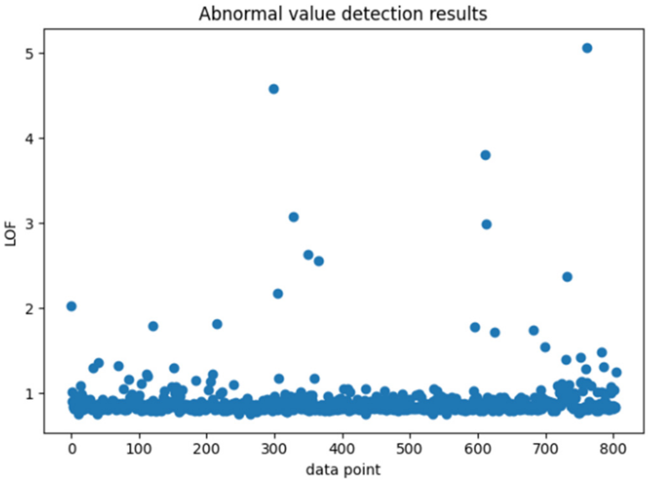


Fig. 9. LOF score of transformer data.

detect abnormal data. When the LOF value is 2, there are 11 abnormal data. When LOF is 3, there are 4 abnormal data.

4.2. Data filling

The real data of transformer oil chromatogram is used as the verification data set to verify the filling of missing values of transformer. Table 2 shows the chromatographic data of transformer oil in a certain area, in which 10% of the data are selected as missing data. This paper uses KNN, random forest and multiple imputation to fill in the random missing data. The red part is the data randomly selected as the missing value.

Table 3 shows the filling effect using the strategy in this paper. The red part is the filling data.

Table 4 shows the accuracy comparison of missing value filling between this strategy, random forest algorithm and multiple filling algorithm.

It can be seen from Table 5 that the KNN algorithm is the best for data filling, with an error rate of only 0.68%, the random forest algorithm is 2.2%, and multiple interpolation is 0.99%. Due to the strong correlation between adjacent data of transformer, it is effective to use KNN algorithm to fill data. This strategy completes transformer data cleaning and preprocessing, and improves the accuracy of data.

Table 3. Oil chromatogram sample data.

H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	Total hydrocarbon	CO	CO_2
10.83	76.93	24.35	82.61	3.87	187.76	481.49	2764.72
10.89	77.34	24.73	84.35	3.94	190.36	477.28	2815.53
10.90	76.81	24.58	**83.68**	3.90	**188.98**	477.35	2796.41
10.99	77.98	24.35	83.34	3.91	189.58	492.99	2785.98
10.81	77.16	24.57	83.52	3.92	189.17	484.40	2785.16
10.41	73.62	23.44	79.69	3.70	180.45	457.16	2680.41
10.77	76.20	23.90	81.17	3.77	185.03	482.79	2717.82
10.79	**76.36**	24.43	82.87	**3.89**	187.55	**477.77**	**2766.02**
10.78	76.03	**23.84**	81.17	3.79	184.83	483.28	2714.71

Table 4. Results of missing values filled by the strategy in this paper.

H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	Total hydrocarbon	CO	CO_2
10.83	76.93	24.35	82.61	3.87	187.76	481.49	2764.72
10.89	77.34	24.73	84.35	3.94	190.36	477.28	2815.53
10.90	76.81	24.58	**83.81**	3.90	**189.23**	477.35	2796.41
10.99	77.98	24.35	83.34	3.91	189.58	492.99	2785.98
10.81	77.16	24.57	83.52	3.92	189.17	484.40	2785.16
10.41	73.62	23.44	79.69	3.70	180.45	457.16	2680.41
10.77	76.20	23.90	81.17	3.77	185.03	482.79	2717.82
10.79	**76.44**	24.43	82.87	**3.90**	187.55	**480.37**	**2760.62**
10.78	76.03	**23.54**	81.17	3.79	184.83	483.28	2714.71

Table 5. Comparison of filling error rates of different algorithms.

	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	Total hydrocarbon	CO	CO_2	Total error rate
KNN	0.08%	0.11%	1.23%	2.2%	0.25%	0.13%	0.54%	0.91%	0.68%
Random forest algorithm	0.73%	2.3%	0.86%	0.98%	5.65%	0.65%	0.75%	5.87%	2.22%
Multiple interpolation algorithm	0.36%	1.43%	1.39%	1.34%	1.54%	0.85%	0.4%	0.61%	0.99%

4.3. Data clustering

Case data were selected from the operation monitoring data on the high-voltage side of a 220 kV main transformer from January 1, 2017 to June 30, 2022. Including high voltage side A phase, B phase, C phase current amplitude, a total of 96 326, data resolution of 30 min, the data has been the above method pretreatment.

The selection of K has a great impact on k-means algorithm, which is one of the difficulties and shortcomings of K-means algorithm. Common methods to select K include elbow method, Gap Statistic method, etc. This example selects elbow method to determine K.

The core index of elbow method is SSE (Sum of the Squared Errors), which represents the clustering errors of all samples and the quality of clustering effect. The formula is as follows:

$$SSE = \sum_{i}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{10}$$

where, c_i is the ith cluster; p is the sample point in c_i, m_i is the center of mass of c_i.

As the number of clusters K increases, the sample division will be more refined, and the degree of aggregation of each cluster will gradually increase, so the error square and SSE will naturally gradually decrease. Moreover, when K is less than the real number of clusters, the increase of K will greatly increase the aggregation degree of each cluster, so the decrease of SSE will be large. When K reaches the real number of clusters, the return of aggregation degree obtained by increasing K will decrease rapidly, so the decrease of SSE will decrease sharply and then tend to be gentle with the continuous increase of K value. The relationship between SSE and K presents an elbow shape, and the corresponding K value of the elbow is the true clustering number of the data.

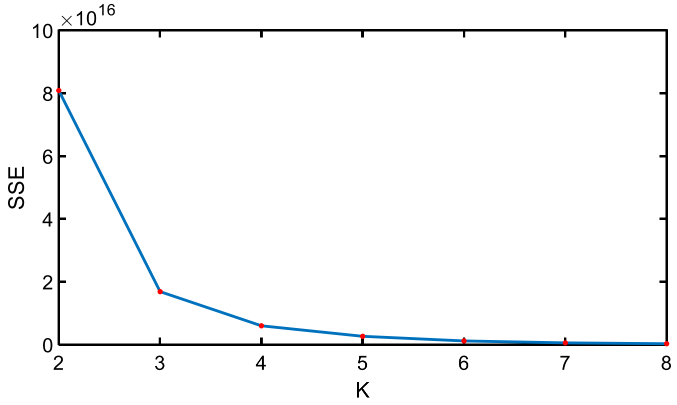


Fig. 10. K values by the elbow method.

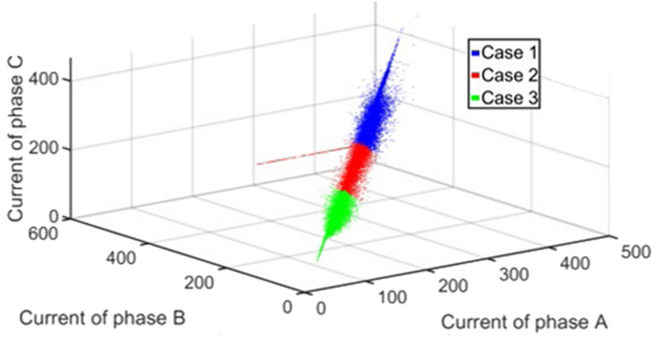


Fig. 11. Case data classification when K = 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Set K as 1 to 8, and use k-means algorithm to get the corresponding SSE value. Fig. 10 shows the corresponding relationship between SSE and K. It can be seen from the figure that when K is less than 3, the curve drops rapidly. When K is greater than 3, the curve tends to be stable. Combined with elbow method, 3 can be considered as the best classification value of case data. Fig. 11 shows the classification results of case data when K equals 3. It can be seen from the figure that K-means algorithm divides the original data into red, blue and green data. It can be seen from Fig. 11 that the K-means algorithm divides the original data into red, blue and green data and has a good clustering effect.

5. Conclusion

This paper proposes a transformer operation data preprocessing method based on KNN and LOF for operation data classification, which can conduct the detection and correction of transformer abnormal data and supplement of missing values. The method is applied and verified with practical operation data which ensures the continuity and integrity of transformer data. Some conclusions can be derived as follows:

- LOF algorithm can effectively detect the transformer data, taking into account the local correlation of the transformer data, and quantify the abnormal conditions of the data. It has a strong pertinence to detect outliers and can select different thresholds according to different data types to detect outliers.
- KNN algorithm can classify relevant data around abnormal values and missing values of transformers, and correct data according to their types. The method considers the correlation between adjacent data of transformer data and improves the filling accuracy of outliers and missing values.
- The elbow method is used to determine the value of K, which improves the accuracy of transformer data clustering and provides more accurate data for later fault diagnosis.

- The proposed method based on LOF and KNN algorithm is very suitable for handling the characteristics of the time series of transformer data. The detection and filling of transformer data ensures the continuity and integrity of the data and provides accurate and effective data samples for transformer data clustering and fault diagnosis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work was supported by the Electric Power Research Institute of Yunnan Power Grid Co., Ltd, China (No. YNKJXM20220009).

References

[1] Liao Ruijin, Wang Youyuan, Liu Hang, et al. Research status of condition assessment methods for power transmission and transformation equipment. High Volt Technol 2018;44(11):3454–64.

[2] Li Gang, Zhang Bo, Zhao Wenqing, et al. Data science issues in power equipment condition assessment: challenges and prospects. Power Syst Autom 2018;42(21):10–21.

[3] Jiang Xiuchen, Sheng Gehao. Research and application of big data analysis for power equipment status. High Volt Technol 2018;44(4):1041–50.

[4] Qiu Jian. Research on power chinese text data mining technology and its application in reliability. Hangzhou: Zhejiang University; 2016.

[5] Zhao Wenqing, Zhu Yongli. Review of power transformer condition assessment. Transformer 2007;44(11):9–12.

[6] Guo Zhimao, Zhou Aoying. A survey of data quality and data cleaning. J Softw 2002;13(11).

[7] Ye Sha. A review of data missing and its processing methods. Electron Test 2017;18.

[8] Yan Yingjie, Sheng Gehao, Chen Yufeng, et al. A big data cleaning method for power transmission and transformation equipment status based on time series analysis. Power Syst Autom 2015a;39(7):138–44.

[9] Yan Yingjie, Sheng Gehao, Chen Yufeng, et al. Anomaly detection method for state data of power transmission and transformation equipment based on big data analysis. Chin J Electr Eng 2015b;35(1):52–9.

[10] Hen LinS, Hu Guoqing, Chen Lizhang, et al. Application of missing forest algorithm in missing value filling. China Health Stat 2014;31(5):774–6.

[11] Li Chunlin, Gao Yupeng, Li Shengyu. Bootstrap variance estimation for multiple interpolation of incomplete data. Stat Decis Mak 2017;18(18):74–6.

[12] Skinner C. Multiple-imputation inferences with uncongenial sources of input. Statist Sci 1994;9(4):561–3.

[13] Abma R, Kabir N. Comparisons of interpolation methods. Lead Edge 2005;24(10):984–9.

[14] Islam MM, Lee G, Hettiwatte SN. Missing measurement estimation of power transformers using a GRNN. In: Australasian universities power engineering conference. AUPEC, 2017, p. 1–5.

[15] Shen Ming, Yin Yi, Wu Jiandong, et al. On line monitoring and inspection of transformer winding deformation. J Electron Tech Technol 2014;29(11):184–90.

[16] Wenzhi Li, Juan ZHU. Cause analysis and treatment of abnormal oil chromatographic data of transformer. Electr Eng 2020;1:115.

[17] Yao He, Hongchi LIANG, Hongsong LIAN, et al. Outlier detection of power transformer oil chromatographic data based on algorithm sliding windows and multivariate Gaussian distribution. High Volt Appar 2020;56(1):203.

[18] Jingmin Fan, Chenyang FU, Hao YIN, et al. Power transformer condition assessment based on online monitor with SOFC chromatographic detector. Int J Electr Power Energy Syst 2020;118:105805.

[19] Lin Jun, Sheng Gehao, et al. Online monitoring data cleaning of transformer considering time series correlation. In: 2018 IEEE/PES transmission and distribution conference and exposition (T & D). 2018, p. 1–9.

[20] Li Xun. Design of infrared anomaly detection for power equipment based on YOLOv3. In: 2019 IEEE 3rd conference on energy internet and energy system integration (EI2). 2019, p. 2291–4.

[21] Song Bin, Yu Dongping, Liao Dongmei, et al. Fuzzy cluster analysis of dissolved gases in transformer fault diagnosis. High Volt Technol 2001;27(3):69–71.

[22] Wu Junjun. Transformer fault classification analysis based on fuzzy clustering. Mod Manuf Technol Equip 2019;03:179–80.

[23] Liu Hang, Wang Youyuan, Ling Xuanhong, et al. Prediction method of the dissolved gas volume fraction in transformer oil based on multi factors. High Volt Eng 2018;44(04):1114–21.

[24] Wang Yang, Tao Yueyue, Wei Tianyi, et al. The study of methods of power transformer state assessment. Hubei Electr Power 2017;41(03):25–31.

[25] Wang Xinzui, Guo Lihong, Xiao Yongpeng, et al. Face recognition algorithm based on BDPCA and KNN. J Wuhan Univ Technol 2009;31(19):130–3.

[26] Wang Fuzhong, Shi Xiuli. Tap-changer fault diagnosis of transformer based on improved PSO-SVM. Electron Meas Technol 2016;39(11):190–4.

[27] Khalyasmaa Alexandra, et al. Machine learning algorithms for power transformers technical state assessment. In: International multi-conference on engineering. Computer and Information Sciences (SIBIRCON); 2019.