J Inf Process Syst, Vol.15, No.5, pp.1156~1170, October 2019
https://doi.org/10.3745/JIPS.02.0119

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

An Optimization Method for the Calculation of SCADA Main Grid's Theoretical Line Loss Based on DBSCAN

Hongyi Cao*, Qiaomu Ren*, Xiuguo Zou*, Shuaitang Zhang*, and Yan Qian*

Abstract

In recent years, the problem of data drifted of the smart grid due to manual operation has been widely studied by researchers in the related domain areas. It has become an important research topic to effectively and reliably find the reasonable data needed in the Supervisory Control and Data Acquisition (SCADA) system has become an important research topic. This paper analyzes the data composition of the smart grid, and explains the power model in two smart grid applications, followed by an analysis on the application of each parameter in density-based spatial clustering of applications with noise (DBSCAN) algorithm. Then a comparison is carried out for the processing effects of the boxplot method, probability weight analysis method and DBSCAN clustering algorithm on the big data driven power grid. According to the comparison results, the performance of the DBSCAN algorithm outperforming other methods in processing effect. The experimental verification shows that the DBSCAN clustering algorithm can effectively screen the power grid data, thereby significantly improving the accuracy and reliability of the calculation result of the main grid's theoretical line loss.

Keywords

Boxplot Method, DBSCAN Clustering Algorithm, Main Grid, SCADA, Theoretical Line Loss

1. Introduction

With the rapid development of the economy, culture, and technologies, the research on smart grid [1] has attracted considerable attention from the research institutions throughout the world [2]. The smart grid aims to construct big-data-based power grid which is capable of automatically monitoring the equipment, collecting user information and storing the power marketing data, etc. [3]. In the domain of power line transmission, the quality of power distribution network, which serves as the power grid terminal directly connecting to the users, has an impact on power quality [4].

Due to the wide application of cloud platform, big data are extensively used in various fields [5], and the artificial intelligence (AI) has become the hottest research direction in the academic and the industrial community at present [6]. The research on the transfer-learning-based AI tool had been used to diagnose the blinding retina diseases and pneumonia [7]. The AI team of the Intel released the open-source nGraph, which enabled the data scientists to focus on the research and development of data science instead of worrying about the problem of deploying deep neural network (DNN) model to different

Manuscript received May 21, 2018; first revision February 8, 2019; second revision March 12, 2019; accepted April 14, 2019.
Corresponding Author: Xiuguo Zou (xiuguozou@gmail.com)
* College of Engineering, Nanjing Agricultural University, Nanjing, China (chy1041928804@126.com, m18951155511@163.com, xiuguozou@gmail.com, zhangshuaitang2014@163.com, qianyan@njau.edu.cn)

equipment for efficient training and operation. The State Power grid Information & Telecommunication Co. Ltd. also founded big data teams to confront the challenges to the smart grid [8].

The stable and automated SCADA (Supervisory Control and Data Acquisition) has always been dependent on the data features for parameters of the power system in each dimension [9-11]. Nowadays, the theoretical line loss calculation of SCADA main grid mainly involves four types of data: line model, transformer model, topological data, and equipment running data. Wherein, the line model data mainly includes line model, length, the resistance of unit length, the susceptance of unit length, etc. [12]. The transformer model data mainly includes transformer model, nominal capacity, rated voltage, the percentage of no-load current, no-load loss, short line loss, the percentage of short-line voltage, etc. The topological data describes the connective relation among equipment in the main grid [13]. The equipment includes the switch, transformer winding, bus, capacitor and reactor, generator, load and line, etc. The running data of equipment includes the active state, the reactive state, and the current. The running data of the transformer includes the active state, the reactive state, the current and the tap position on the high voltage side, medium voltage side, and low voltage side [14]. The running data of switch includes the On-Off state. The running data of load is the active state and the reactive state. The running data of the bus is voltage. The theoretical line loss for the main grid is calculated by using the result of load flow calculation which based on the data of line and transformer models.

In the long-term collection and recording of high-density data [15], the visual fatigue, lowered mental concentration, and other physiological factors may affect the accuracy of data, and influences the correctness of evaluated steadiness data of the electric system [16]. The data collection problem of SCADA has existed for years, and there is still a lack of the appropriate solution. For the problem of data screening in the power system, the researchers pointed out that the data information collected via big data both includes the true data information and the false data information [17,18], which people hope to solve through the intelligent screening [19].

In recent years, the clustering method based on machine learning has been extensively applied to image recognition, video classification, semantics understanding, etc. The combination of the K-means clustering algorithm and the traditional Mahalanobis distance has been implemented to measure the upper and lower volume ratio for fruit trees in the orchards with significant effect [20]. Furthermore, the clustering method also exerts a satisfactory function in the ship noise data processing [21]. However, the research on the application of the clustering based on machine learning to the power data system has not been reported, which highlights the necessity of the machine learning based power data screening [8].

To cope with the low accuracy of feedback data in the statistic process of power grid, this paper applied the algorithms which based on data density features for data screening according to the correctness and reliability of big data probability distribution based on the theoretical basis of density-based spatial clustering of applications with noise (DBSCAN) [22], and compares them by means of the traditional methods such as boxplot method. The practical study proves that DBSCAN processing based on data density significantly outperforms other data processing methods.

In China, the line parameters (e.g., wire length and cross-sectional area) of the power grid model are influencing each other. The selection of line parameters has always been a difficult problem. The unsuitable selection of line parameters will lead to increase the line loss rate, further reducing the power quality. Nowadays, the compliance of a set of line parameters with the requirements for line loss depends entirely on the experience of electrical engineers. However, it is impossible to assign an experienced engineer during the process of collecting each piece of data in the power grid, which is extremely costly

and inefficient. Besides, the line parameters are completely determined by the engineer's experience, which is the lack of rationality. Therefore, it has become an important issue to determine the appropriate line parameters efficiently and cost-effectively and the combination of machine learning and engineers' experience has become a promising way to solve this issue.

2. Materials and Methods

2.1 Data Acquisition

The main data used in this research is from the smart grid data of Zhejiang Province in China. As the traditional manual collection of data, would result in the deviation of part of data in the dataset due to multiple factors such as long-term working, and the cluster screening is a proper choice to screen for the reasonable static power data according to the features of data itself [23].

This paper mainly studies the data analysis and cluster screening for static power data model under the main grid voltage classes of 35 kV, 50 kV, 100 V, 500 kV, 1,000 kV, respectively. The analysis result is compared using the traditional data screening methods (such as boxplot method and probability weight analysis method, etc.), and the comparison results are presented for the multidimensional data screening.

There are a total of 10,174 pieces of data in the line model. Table 1 shows part of the data, where 'cross-sectional_area' represents the cross-sectional area of the line, 'length' represents the length of the line, and 'voltage_class' represents the voltage level of the line.

Table 1. The raw data of the power line model

No.	Cross-sectional_area (mm^2)	Length (km)	Voltage_class (kV)
1	300	20.788	110
2	185	4.663	35
3	185	2.434	35
4	400	0.638	220
5	240	2.434	35
6	300	1.778	110
7	300	1.778	110
8	720	104.6	500
9	185	4.663	35
10	300	4.42	110
...		...	
10,174	240	3.439	110

As each voltage level has different requirements for the material of line, the lines are classified according to the voltage level. We have obtained 2,828 pieces of data from 35 kV line model, 4,447 pieces from 110 kV line model, 2,706 pieces from 220 kV line model, 183 pieces from 500 kV line model and 10 pieces from 1,000 kV line model. Some raw data of classified lines are presented in Table 2.

The big data based data processing method is not suitable to process the data from 500 kV and 1,000 kV line models due to its limited data size (183 pieces of data from 500 kV models; 10 pieces of data from 1,000 kV line models). Therefore, this paper only analyzes the line models under the three voltage levels of 35 kV, 110 kV, and 220 kV.

Table 2. Some raw data of three kinds of voltage classes

Classification	No.	Cross-sectional_area (mm^2)	Length (km)
Data of 35 kV voltage class	1	300	13.33
	2	240	7.06
	3	150	5
	4	150	3.08
	5	240	8
	6	240	10
	7	240	6.88
	...	...	
	2,828	300	7.2
Data of 110 kV voltage class	1	240	26.27
	2	300	4.56
	3	300	4.56
	4	300	4.56
	5	300	4.56
	6	300	2.54
	7	300	1.50
	...	...	
	4,447	240	15.49
Data of 220 kV voltage class	1	400	1.09
	2	400	1.09
	3	400	0.90
	4	400	0.87
	5	300	0.60
	6	400	0.39
	7	300	0.54
	...	...	
	2,706	240	0.44

2.2 Data Preprocessing

The monitoring and feedback data of the smart power system have some drawbacks such as huge data size, numerous data types, ineffective information, etc. [2]. To obtain the reasonable data fit for this research, the raw data is processed for the power monitoring and feedback which includes conductor length, cross-sectional area, resistance, circuit region, circuit model, etc. Considering the data extraction should conform to the practice of the extraction of cross-sectional area, in the case of one length of conductor with multiple cross-sectional area, the processing method is to divide one length of the conductor into multiple parts and to mark each part by corresponding cross-sectional area.

The missing data and nulls in the raw data were all weeded out in the data preprocessing to avoid influence on the experimental result. Besides, the general data were normalized with the consideration of the density level of cluster data.

2.3 Boxplot Method and Probability Weight

Boxplot is a graph used to display the dispersion of a set of data. It can show the maximum, minimum, median, and upper and lower quartiles of a set of data.

A boxplot is drawn according to the following steps:

1) Determine the median, upper quartile (Q3) and lower quartile (Q1) of the line loss data to be processed;

2) Determine the interquartile range of the line loss data from the line model IQR=Q3–Q1;

3) Two line segments are drawn at Q3+1.5*IQR and Q1–1.5*IQR, respectively, which are defined as the inner limits of the line loss data; two line segments are drawn at Q3+3*IQR and Q1–3*IQR, respectively, which are defined as the outer limits of line loss data. The data represented by the points outside the inner limit are all abnormal values, where those between the inner limit and the outer limit are mild ones, and those outside the outer limit are the extreme ones.

2.4 DBSCAN Method

DBSCAN is an algorithm depicting the compact degree of sample distribution based on a group of "neighborhood" parameters. For a given dataset D={$x_1,x_2,...,x_m$}, there are following concepts [24]:

1) ε-neighborhood: For $x_j \in$ D, its ε neighborhood includes the samples which keep distances from x_j no longer than ε in the sample set D, as shown by Eq. (1):

$$N_\varepsilon(x_j) = \{x_i \in D \mid dist(x_i, x_j) \le \varepsilon\} \tag{1}$$

where ε is the minimum radius, $dist(x_i,x_j)$ denotes the distance from x_i to x_j, i,j=1,2,...,m.

2) Core object: If x_j includes at least MinPts samples, i.e., $|N_\varepsilon(x_j)| \ge MinPts$, it is a core object.

3) Density-directly-reachable: If x_j is located in the ε-neighborhood of x_i, and x_j is the core object, x_j is called density-directly-reachable by x_i.

4) Density-reachable: For x_i and x_j, if there exists a sample sequence $p_1,p_2,...,p_n$, wherein, $p_1=x_i$, $p_n=x_j$ and p_{i+1} is density-directly-reachable by p_i, x_j is called density-reachable by x_i

5) Density-linked: For x_i and x_j, if x_k exists to make x_i and x_j density-reachable by x_k, x_i is called density-linked with x_j.

To research the relation between two neighborhood parameters, clustering analysis is carried out on the power line model data for an East China province, which the results are shown in Table 3. Let us denote the minimum radius as ε and the number of points within the minimum radius as MinPts.

Table 3. Clustering results of 110 kV experimental datasets

Sample set (kV)	Parameter (ε, MinPts)	Number of clusters	Outlier	Cross-sectional area (mm²)		Length (km)	
				S_max	S_min	L_max	L_min
110	(30, 40)	4	51	67.6	0.02	425	120
110	(30, 80)	3	145	67.6	0.02	425	185
110	(30, 120)	3	245	67.6	0.02	425	210
110	(30, 160)	2	397	67.6	0.028	340	210
110	(30, 200)	2	397	67.6	0.028	340	210

In Table 3, L_max and L_min represent the maximum and minimum of line length, respectively; S_max and S_min represent the maximum and minimum of cross-sectional area, respectively.

Observation of the above table finds that the influence of the two density parameters on the clustering result may have some intrinsic correlation. Based on the above observation and conjecture of data, the following experiments were conducted.

Experiment 1: Firstly, a dataset of operation model dataset with adequately large data size was selected, and each piece of data respectively included two data dimensions of conductor length and cross-sectional area of the conductor. The original data was put in the data model and density parameters were adjusted to obtain different clustering results. Then, the data size in the original dataset was doubled, and the data was put in the data model again. In the meanwhile, the parameter value of MinPts was doubled correspondingly, with the experimental result obtained as shown in Table 4.

Table 4. The clustering results of double times dataset

Sample set (kV)	Parameter (ε, MinPts)	Number of clusters	Outlier	Cross-sectional area (mm^2)		Length (km)	
				S_max	S_min	L_max	L_min
110	(30, 40)	5	72	67.6	0.02	425	70
110	(30, 80)	4	102	67.6	0.02	425	120
110	(30, 120)	4	104	67.6	0.02	425	120
110	(30, 160)	3	258	67.6	0.02	425	185
110	(30, 240)	3	490	67.6	0.02	425	210
110	(30, 320)	2	794	67.6	0.028	340	210
110	(30, 400)	2	794	67.6	0.028	340	210

Experiment 2: Firstly, a data set of an operation model dataset with an adequately large data size was selected, and each piece of data respectively included two data dimensions of conductor length and cross-sectional area. The original data was put in the data model and density parameters were adjusted to obtain different clustering results. Then, half of the data were randomly extracted from the raw data and put in the data model for operation. In the meanwhile, the MinPts has been reduced to half of the original value accordingly, with the result as shown in Table 5.

Table 5. The clustering results of half dataset

Sample set (kV)	Parameter (ε, MinPts)	Number of clusters	Outlier	Cross-sectional area (mm^2)		Length (km)	
				S_max	S_min	L_max	L_min
110	(30, 20)	6	12	67.6	0.0225	425	70
110	(30, 40)	4	49	67.6	0.0225	425	185
110	(30, 60)	4	50	67.6	0.0225	425	185
110	(30, 80)	3	122	67.6	0.0225	340	185
110	(30, 100)	2	204	67.6	0.0225	340	220

Comparison of the above two experiments can result in the following conclusions:

1) The clustering results for the dataset with different sizes under the same neighborhood parameters are different.

2) When the dataset is completely duplicated into twice that of the original dataset, the MinPts in the density parameters is also increased to twice that of the original value, and the clustering results are identical.

3) Increase the MinPts index to twice of the original value when the data size of the new dataset is twice that of the original dataset. Though the obtained clustering result is not identical to the result of the original dataset, however, they are roughly in compliance with each other.

In summary, the results obtained using DBSCAN clustering are related to the type of clustering parameters and the size of clustering data.

Based on the above conclusion, we attempt to find out the optimum density parameters under the dataset with a certain amount of data. When the new data is provided, just compare the new data size with that of the original clustering data, thereby adjust the parameter index of MinPts in the clustering algorithm according to the size of the new data. Then, the practical significance and functions of the two neighborhood parameters can be differentiated, where ε is the radius parameter, which determines the index of clustering density according to the concrete data variables. The MinPts is the data parameter, which determines the index of clustering density according to the data size of the clustering target.

3. Experiments and Results

Coupled with the development of the economy and productive forces like technologies, the quality of electric power has played an important role in people's life [25]. Currently, the power system mainly uses two models: the main grid line model and the main grid transformer model. The former mainly analyzes various attribute properties of the conductor with different voltage classes, and is mainly influenced by the length and cross-sectional area of conductors. The latter is a commonly used model judging the performance of the transformer and is mainly affected by line loss and capacity. For the above two common power data models, the paper analyzes the optimization method for related data and compares the optimization effect of the algorithms.

3.1 Boxplot Method and Probability Weight Analysis

As a mature and convenient means in traditional data screening and data preprocessing, the boxplot method has always been favored by a great many data researchers. For the traditional data screening of power related big data, the data processing method of boxplot targeting each variable is a reliable means.

The data for the main grid line model for an East China province is screened through the boxplot method and probability weight analysis, with the raw data is as shown in Table 6.

Table 6. The list of raw data statistics

Voltage_class (kV)	L_max (km)	L_min (km)	L_mean (km)	S_max (mm^2)	S_min (mm^2)	S_mean (mm^2)
35	25,555	0.01	22.25	500	50	203.3761
110	500	0.02	9.00	630	50	280.5564
220	116.69	0.032	9.90	675	70	383.9208
500	246.17	0.77	52.81	800	300	561.4583
1,000	275.28	118.2	180.47	630	630	630

Different boxplots are established to process the data of main grid line model with different voltage classes. The processed result is shown in Fig. 1.

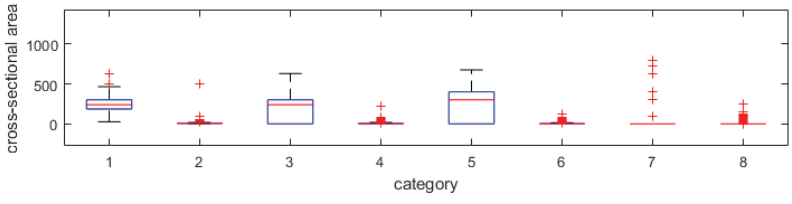


Fig. 1. The result of analysis by boxplot method.

For the length data in the data of main grid line, most of the frequencies by which the lengths corresponding to the voltage classes of 35 kV, 110 kV, 220 kV, 500 kV appear once, which makes it impossible to take scope using probability. The probability is uniform relatively, so the algorithms such as boxplot or something can be considered for data analysis. According to the characteristics of data structure and processing result, the values between the Q3+1.5*IQR and Q1−0.25*IQR are selected to constitute a box, while the rest ones are called abnormal values. Wherein, the interquartile range IQR = Q3 − Q1. For the data of main grid line model with the five voltage classes, the data respective partition is carried out using boxplot method, with the statistical result of the maximum and minimum as shown in Table 7.

Table 7. Result of the maximal and minimal lengths obtained by boxplot method

Voltage_class (kV)	L_max (km)	L_min (km)
35	24.3985	1.5048
110	22.7045	1.6643
220	32.9459	0
500	114.115	22.2375
1000	307.81	99.635

As the data is distributed discretely, it is advisable to process the cross-sectional area data in the line data of the main grid using boxplot method. Thus the probability weight analysis method is adopted.

The “big value” appearing for the first time within the minimal scope is taken as the lower limit of cross-sectional area data, and the “big value” appearing for the first time within the maximal scope is taken as the upper limit of sectional area data. The output interval is defined according to the two values and the reasonable sectional area data is screened for.

The data statistical result of the cross-sectional area in the experiment is as shown in Table 8.

It is found by dint of the related knowledge of probability statistics that the probability weight and frequency corresponding to the area of 70 are significantly higher than those corresponding to the areas of 50 and 60. Such is the situation that the data of 70 as the lower limit of statistics. Through the above methods, the maximum and minimum for the line model area with 5 voltage classes are obtained, as shown in Table 9.

3.2 Analysis via DBSCAN-based Clustering Algorithm

As the length and the cross-sectional area are two reference parameters which simultaneously act on the result of data screening, it is more reasonable to consider simultaneously these two variables for screening than to screen individual variables respectively, so the DBSCAN-based clustering algorithm is adopted for data analysis.

As raw data differ a lot in order of length and cross-sectional, while the DBSCAN algorithm requires the relatively average order of variables, the raw data is first subject to standardization by using the Eq. (2).

$$y = \log_{10} x \tag{2}$$

Python is used as a platform for big data processing. For the power grid data with the voltage of 35 kV, 110 kV, 220 kV, the clustering results are shown in Tables 10–12.

Table 8. The data statistical result of the cross-sectional area

Cross-sectional_area (mm^2)	Frequency	Probability
50.00	1	0.0468
65.00	1	0.0468
70.00	56	2.6180
82.50	3	0.1403
85.00	4	0.1870
90.00	2	0.0935
92.50	7	0.3273
95.00	59	2.7583
101.67	1	0.0468
105.00	6	0.2805
107.50	6	0.2805
109.00	3	0.1403
110.00	3	0.1403
113.33	1	0.0468
117.50	1	0.0468
120.00	118	5.5166
121.67	1	0.0468
122.50	3	0.1403
123.75	3	0.1403
125.00	2	0.0935
126.67	2	0.0935
127.50	13	0.6078
133.33	4	0.1870

Table 9. Result of the maximal and the minimal cross-sectional areas obtained by probability weight analysis method

Voltage_class (kV)	S_max (mm^2)	S_min (mm^2)
35	400	70
110	500	150
220	630	185
500	800	300
1000	630	630

Table 10. The clustering result of 35 kV voltage by DBSCAN algorithm

Parameter (ε, MinPts)	Number of clusters	Outlier	Cross-sectional area (mm^2)		Length (km)	
			S_max	S_min	L_max	L_min
(50, 80)	2	67	46.4	0.01	300	25
(30, 60)	4	67	46.4	0.01	300	25
(90, 70)	1	67	46.4	0.01	300	25
(50, 250)	2	67	46.4	0.01	300	25
(30, 40)	5	15	46.4	0.01	400	25
(50, 50)	3	15	46.4	0.01	400	25
(100, 40)	2	7	100	0.01	400	25
(10, 30)	8	53	36.5	0.01	400	70

Table 11. The clustering result of 110 kV voltage by DBSCAN algorithm

Parameter (ε, MinPts)	Number of clusters	Outlier	Cross-sectional area (mm²)		Length (km)	
			S_max	S_min	L_max	L_min
(60, 20)	3	23	100	0.02	465	35
(60, 50)	3	25	100	0.02	465	50
(40, 30)	3	43	79.088	0.02	426	95
(50, 50)	3	40	79.088	0.02	465	95
(80, 30)	1	9	100	0.02	500	35
(80, 50)	1	9	100	0.02	500	35
(100, 110)	1	9	100	0.02	500	35
(30, 40)	4	51	67.6	0.02	425	120

Table 12. The clustering result of 220 kV voltage by DBSCAN algorithm

Parameter (ε, MinPts)	Number of clusters	Outlier	Cross-sectional area (mm²)		Length (km)	
			S_max	S_min	L_max	L_min
(40, 30)	2	0	81.545	0.032	630	120
(80, 30)	2	0	115.126	0.032	675	70
(100, 110)	2	0	115.126	0.032	675	70
(100, 40)	2	0	115.126	0.032	675	70
(80, 50)	4	2	115.126	0.032	675	120
(60, 20)	2	2	115.126	0.032	675	120
(70, 50)	4	6	115.126	0.032	675	120
(30, 40)	5	60	81.545	0.032	630	220

The data volume is so limited, with only 183 and 10 pieces of data for voltage levels of 500 kV and 1,000 kV, respectively, that it is not suitable to cluster them using DBSCAN method. Therefore, the clustering will not be performed for the data of 500 kV and 1,000 kV.

The data results under each parameter are empirically judged by the electrical engineers from the State Grid Corporation of China. Tables 10–12 displays the line loss parameters under each voltage level. Table 13 presents the results obtained by the DBSCAN clustering method and the boxplot probability weighting algorithm.

Table 13. The result obtained using DBSCAN method

Voltage_class (kV)	S_max (mm²)	S_min (mm²)	L_max (km)	L_min (km)
35	400	25	46.4	0.01
110	465	95	79.088	0.02
220	630	220	81.545	0.032

3.3 Comparative Analysis of the Results

The results obtained using DBSCAN clustering method and boxline probability weighing algorithm are shown in Table 14.

Table 14. The results obtained using DBSCAN method and boxplot and probability weighing algorithm

Method	Voltage_class (kV)	S_max (mm^2)	S_min (mm^2)	L_max (km)	L_min (km)
Boxplot and probability weight analysis	35	400	70	24.3985	1.50475
	110	500	150	22.7045	1.66425
	220	630	185	32.9459	0
	500	800	300	114.115	22.2375
DBSCAN	35	400	25	46.4	0.01
	110	465	95	79.088	0.02
	220	630	220	81.545	0.032
	500	630	400	124.8	0.77

The comparison of the above two processing methods can reach the following conclusions:

1) In processing the data containing multiple variables, the DBSCAN algorithm can process and screen the multivariable data simultaneously, and the processing result is better than the result of only processing the data with a single variable. By contrast, in the boxplot and probability weighting method, a single variable takes a certain data close to the maximum and minimum values as the potentially suitable range of this parameter. The results obtained using the above methods, for not taking into account the connection of parameters, may cause inaccuracy. For example, for the 220 kV line model, the minimum line length obtained by the boxplot method is 0, which is impossible in actual situations.

2) DBSCAN has a higher data resolution, hence it can provide a more precise result for some practical problems without being influenced by some abnormal values. For example, for the power grid data in the East China province, some numerical values increased abnormally due to the misuse of the unit in the real process. If the traditional probability weight method had been used for analysis, the derived result would have had huge deviations.

4. Discussion

To demonstrate the advantages and disadvantages of the above two algorithms, the line data of a grid transformer is used to experimentally examine the above conclusions.

In contrast, the data processing results obtained by the boxplot method are shown in Table 7.

The main grid's transformer model is a data analysis model which is established based on the influence of the two variables of load and capacity on the result for the transformers with different windings under different voltage classes. Based on the above comparison between the clustering analysis and analysis via other data screening methods, the DBSCAN clustering algorithm is utilized to carry out a clustering analysis on the transformer model data, with clustering result shown in Table 15.

Wherein, the clustering for a certain parameter (e.g., $\varepsilon = 0.1$, MinPts = 20) obtains the result as shown in Fig. 2. Its ordinate represents the base 10 logarithm of the capacity.

It is observed from the Fig. 2 that the DBSCAN clustering algorithm obviously weeds out some unreasonable data with appropriate data size (take the above black data as example, the practical survey found that it is the waste data produced by users' incorrect filling of unit) and reserves the reasonable data conforming to the characteristics of power grid.

The thresholds of 110_10 kV transformer for boxplot method are presented in Table 16.

Table 15. The thresholds of 110_10 kV transformer in DBSCAN method in different parameters

Parameter (ε, MinPts)	Number of clusters	Outlier	Load		Capacity	
			Load_max (kW)	Load_min (kW)	C_max (kVA)	C_min (kVA)
(0.05, 30)	2	115	44.72	17.12	50,000	40,000
(0.05, 40)	2	117	44.72	17.12	50,000	40,000
(0.05, 50)	2	120	44.72	18.167	50,000	40,000
(0.05, 20)	4	62	44.72	17.12	50,000	31,500
(0.1, 20)	2	36	59.88	15	63,000	31,500
(0.1, 40)	2	56	52	15	63,000	31,500
(0.2, 30)	1	14	60	10	80,000	20,000
(0.2, 40)	1	14	60	10	80,000	20,000
(0.3, 30)	1	9	96.4	10	100,000	20,000
(0.3, 20)	1	9	96.4	10	100,000	20,000

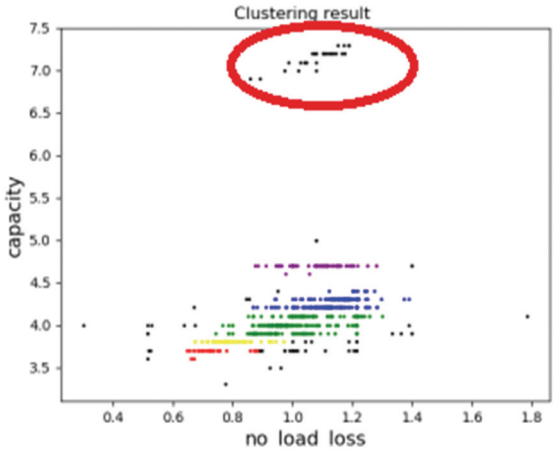


Fig. 2. The clustering result of the main grid's transformer model by DBSCAN algorithm.

Table 16. The thresholds of 110_10 kV transformer for boxplot method

Load_max (kW)	Load_min (kW)	C_max (kVA)	C_min (kVA)
200.4	10	100,000	20,000

By repeatedly searching and comparing the raw data of the transformer, it is found that a considerable part of the data presenting the power capacity is incorrectly counted in terms of the unit of their magnitude, resulting in the great offset of the overall data. For the data processing method based on the boxplot, the range of values of parameters can only be obtained by the extremum and median of the raw data, and the threshold offset caused by the magnitude error cannot be ignored. For example, the threshold range C_max=100,000 given in Table 16 is an obvious error caused by magnitude error. For the DBSCAN method, as long as the parameters are appropriately adjusted, the interference of the magnitude error can be eliminated by clustering, as indicated by the results listed in Table 15.

Finally, through the repeated confirmation with the engineers from the State Grid, the parameters of 110_10 kV transformer are set to (0.05, 50), and the thresholds are presented in Table 17.

Table 17. The final thresholds result of 110_10 kV transformer

Load_max (kW)	Load_min (kW)	C_max (kVA)	C_min (kVA)
44.72	18.167	50000	40,000

5. Conclusion

Currently, the proper data processing means are still absent in the statistics for smart grid data in the domain of power system. Although the traditional boxplot method and the probability weight method can screen the data, there still lacks the correlation between multiple dimensions of data, and the processing result is not perfect. The experimental verification indicates that the DBSCAN-based clustering method can effectively screen the monitoring and feedback data for the line transformer of the smart grid, and result in more satisfactory clustering effect in comparison with other methods. This method can significantly improve the accuracy of big data of the smart grid, thereby guaranteeing the reliability of the whole grid system.

It has been verified that through the above experiments the DBSCAN has a significantly better effect on obtaining the thresholds of multi-dimensional data than the boxplot method which can only analyze the one-dimensional data. Besides, the DBSCAN also solves the problem of dependence on the engineer's experience to completely determine the line parameters. The DBSCAN can determine the threshold ranges of the related parameters, e.g. line loss, inline model. Through further confirmation of the engineers from the State Grid, the threshold range of the relevant parameters of the line model is determined in this paper.

Acknowledgement

This paper is supported by the Fundamental Research Funds for the Central Universities of China (No. KYTZ201661), China Postdoctoral Science Foundation (No. 2015M571782), and Jiangsu Agricultural Machinery Foundation (No. GXZ14002), and University Student Entrepreneurship Training Program of Jiangsu Province (No. 201810307030T).

References

[1] N. Boumkheld, M. Ghogho, and M. El Koutbi, "Energy consumption scheduling in a smart grid including renewable energy," *Journal of Information Processing Systems*, vol. 11, no. 1, pp. 116-124, 2015.

[2] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid: the new and improved power grid: a survey," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 944-980, 2011.

[3] Y. Song, G. Zhou, and Y. Zhu, "Present status and challenges of big data processing in smart grid," *Power System Technology*, vol. 37, no. 4, pp. 927-935, 2013.

[4] H. Liu, "Safety problems and maintenance technology analysis of power line overhaul," *Science & Technology*, vol. 2018, no. 4, pp. 69-70, 2018.

[5] Y. Li and D. L. Shi, "Design and research of cloud platform of computer experiment center under OpenStack," *China Hi-tech Industrial Development Zone*, vol. 2018, no. 8, pp. 231-231, 2018.

[6] G. J. Li, "The scientific value of big data research," *Chinese Computer Society Newsletter*, vol. 8, no. 9, pp. 8-15, 2012.

[7] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131, 2018.

[8] State Grid Corporation of China, "ICT held big data to open a new era of smart grid discussion," [Online]. Available http://www.cet.sgcc.com.cn/html/sgit/col1180000040/2012-07/16/20120716164928056427252 9_1.html

[9] H. Wu, Z. Ma, Y. Jiang, L. Wang, H. Yang, Y. Li, P. Zuo, H. Jia, and W. Liu, "Direct observation of the carrier transport process in InGaN quantum wells with a pn-junction," *Chinese Physics B*, vol. 25, no. 11, article no. 117803, 2016.

[10] D. Jiang, X. Bu, B. Sun, G. Lin, H. Zhao, Y. Cai, and L. Fang, "Experimental study on ceramic membrane technology for onboard oxygen generation," *Chinese Journal of Aeronautics*, vol. 29, no. 4, pp. 863-873, 2016.

[11] D. Yang, W. Kong, B. Li, and X. Lian, "Intelligent vehicle electrical power supply system with central coordinated protection," *Chinese Journal of Mechanical Engineering*, vol. 29, no. 4, pp. 781-791, 2016.

[12] Y. Zhang, "Design and development of online calculation data interface for regional power grid theoretical line loss," *Guangxi University*, 2012.

[13] Y. Li, L. Liu, B. Li, J. Yi, Z. Wang, and S. Tian, "Calculation of line loss rate in transformer district based on improved k-means clustering algorithm and BP neural network," *Proceedings of the Chinese Society for Electrical Engineering*, vol. 36, no. 17, pp. 4543-4551, 2016.

[14] D. Z. Chen and Z. Z. Guo, "Distribution system theoretical line loss calculation based on load obtaining and matching power flow," *Power System Technology*, vol. 2015, no. 1, pp. 80-84, 2005.

[15] Q. Zhu, J. Dang, J. Chen, Y. Xu, Y. Li, and X. Duan, "Power system transient stability assessment method based on deep belief network," *Proceedings of the Chinese Society for Electrical Engineering*, vol. 38, no. 3, pp. 735-743, 2018.

[16] H. Xu, G. Jiang, M. Yu, T. Luo, Z. Peng, F. Shao, and H. Jiang, "3D visual discomfort predictor based on subjective perceived-constraint sparse representation in 3D display system," *Future Generation Computer Systems*, vol. 83, pp. 85-94, 2018.

[17] Q. M. Wu, "Research and implementation of data fault-tolerance technology in big data environment," School of Engineering Management and Information Technology, University of Chinese Academy of Sciences, 2016.

[18] H. X. Huang, "Research on data fault-tolerance technology in big data cloud storage," Fuzhou University, 2016.

[19] L. X. Zheng, "Innovative thinking of power data statistics based on data mining mode," *Hong Kong and Macao Economy*, vol. 2015, no. 29, pp. 118-118, 2015.

[20] L. Y. Min, Y. F. Cheng, H. Cheng, Z. L. Yang, Y. L. Wu, and L. Z. Ge, "Measure the volume ratio of upper and lower canopy layers of fruit trees based on M-K cluster," *Journal of Chinese Agricultural Machinery*, vol. 2018, no. 6, pp. 1-9, 2018.

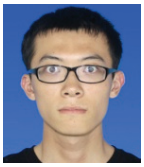
[21] H. Ding, "Automatic and rapid screening of data in ship noise database," *Ship Science and Technology*, vol. 40, no. 4, pp. 37-39, 2018.

[22] L. Bingzhan, L. Ziping, and L. Danbing, "Intelligent BIM operation and maintenance management system based on machine learning and maintenance application of BIM+MR: taking hospital construction as an example," *Civil Engineering Information Technology*, vol. 9, no. 6, pp. 22-27, 2017.

[23] S. Ye, X. Wang, Z. Liu, and Q. Qian, "Power system transient stability assessment based on support vector machine incremental learning method," *Dianli Xitong Zidonghua (Automation of Electric Power Systems)*, vol. 35, no. 11, pp. 15-19, 2011.

[24] Z. H. Zhou, *Machine Learning*. Beijing: Tsinghua University Press, 2015.

[25] X. Z. Gao, "Research on control strategy and power quality improvement of microgrid," Tianjin University, 2012.



Hongyi Cao https://orcid.org/0000-0003-2419-2396

He is an undergraduate student at Nanjing Agricultural University, majoring in automation. His current research interests include big data processing based on machine learning and deep learning.

Qiaomu Ren https://orcid.org/0000-0002-0843-9355

He is an undergraduate student at Nanjing Agricultural University, majoring in electronic information science and technology. His current research interests include big data processing based on machine learning and deep learning.

Xiuguo Zou https://orcid.org/0000-0002-8074-7555

He received the doctorate degree in Nanjing Agricultural University (China) in 2013. He currently works at Nanjing Agricultural University as an associate professor. His interests and research are focused on image processing and pattern recognition, and big data processing based on machine learning. He has authored over 10 technical journals in the area of the big data processing.

Shuaitang Zhang https://orcid.org/0000-0001-7547-5439

He received B.S. degrees in College of Engineering, Nanjing Agricultural University in 2015. Since September 2016, he is as a master student at Nanjing Agricultural University. His current research interests include image processing and pattern recognition, machine learning and deep learning.

Yan Qian https://orcid.org/0000-0001-8350-9352

She received the doctorate degree in Nanjing Agricultural University (China) in 2014. She currently works at Nanjing Agricultural University as an associate professor. Her interests and research are focused on machine learning and deep learning. She has authored over 10 technical journals in the area of machine learning and deep learning.