International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 5, May - 2013

Abnormal Data Processing Of Heating Load Based On Density Estimation

Zhang Huaqiang, Meng Mengmeng, Ma Tong
Department of Electrical Engineering, Harbin Institute of Technology，Harbin 150001, China

Abstract

Due to the limitation of modern techniques and various interferences, there usually exist some abnormal data in Supervisory Control and Data Acquisition system. It will affect the accuracy of data analysis, load forecasting and even make serious mistakes in production scheduling only according to those data without processing. It is necessary to identify and correct anomalous data once again. However, there are some limitations in conventional data processing approaches, because they mainly focus on one dimension space. Due to horizontal and vertical continuities of heating load data, a novel abnormal data processing algorithm based on density estimation which identifies and processes data in two-dimension space is proposed. This method proves feasible by large numbers of examples and simulation results.

Keywords: *heating load; data pre-processing; two-dimension space; density estimation; abnormal data*

1. Introduction

Central heating system has become popular with the development of cities' heat-supply. But it brings about serious problems of energy shortage and environmental pollution. The modern heating systems require effective utilization of heating source and keeping pipeline balance, because it can help to prevent pollution and reduce consumption of energies [1][2]. The reliability of data in Supervisory Control and Data Acquisition (SCADA) system is not only the basis for heating load prediction and characteristic analysis, but also the foundation of regulating load network balance. The appearance of abnormal data, continuous missing data and fluctuation phenomenon caused by channel noise, impact load, sudden accident will seriously affect load forecasting results. High-quality data acquisition and prediction results are important for heating system production scheduling [3] [4]. Therefore, the identification and secondary correction of abnormal data in SCADA are of great significance.

A large quantity of work on detecting and modifying abnormal data has been studied by domestic and foreign scholars. A method to analyze abnormal load data by applying wavelet singularity detection principle is introduced in reference [5][6]. Reference [7] describes how to identify abnormal load data by taking advantage of redundant data in SCADA system. This method fits well into abnormal data caused by acquisition system fault, but it can do nothing with abnormal fluctuating loads. The ART2 artificial neural network model is utilized to identify and adjust anomalous data based on characteristic curves extracted from the classified data in advance[8]. However, these methods only take the horizontal or vertical continuity of load into consideration respectively. In other words, they deal with the data in one-dimensional space and have limitations to some extent. Given horizontal and vertical continuity of load simultaneously, a method to identify and correct abnormal data in two-dimension space based on data density estimation is presented. Firstly, convert the load data series into a two-dimensional data matrix with columns and rows corresponding to days and hours respectively[9]. Secondly, identify continuous missing data on the whole, eliminate and adjust the anomalous data. At last, adopt the actual data provided by a certain heating power plant for prediction, large numbers of examples and simulation results indicate that the proposed method is feasible and effective.

2. Identifying continuous missing data

Traditional identification methods are based on heating load viscous principle which refers to that adjacent load data will not mutate. Then set a threshold as upper limit of allowable variation among adjacent load data. When absolute value of difference between two adjacent points exceeds the set threshold, these

data are regarded as abnormal. The identification formula is:

$$\begin{cases} |L_{d,t} - L_{d,t-1}| > \alpha \\ |L_{d,t} - L_{d,t+1}| > \alpha \end{cases} \quad (1)$$

Where, $L_{d,t}$ is the load data at time t of the d day, α is the set threshold.

$L_{d,t}$ will be classified as abnormal data, when formula (1) is satisfied. But this approach has some problems in dealing with continuous missing data. In order to solve it, the change rate of adjacent load points is taken as abnormal data identification criteria. The improved identification formula can be written:

$$\begin{cases} \dfrac{L_{d,t} - L_{d,t-1}}{L_{d,t-1}} > \alpha, \quad t \neq 1 \\ \dfrac{L_{d,t} - L_{d-1,24}}{L_{d-1,24}} > \alpha, t = 1 \end{cases} \quad (2)$$

The whole process should be carried out in chronological sequence. Abnormal data should be corrected immediately, and be compared with next data once they are detected. The correction formula based on weighted average processing is:

$$L_{d,t} = \lambda_1 L_{d\text{-}1,t} + \lambda_2 L_{d\text{-}2,t} + ... + \lambda_m L_{d\text{-}m,t} \quad (3)$$

Where, $L_{d-m,t}$ is the load data at time *t* of *d*-*m* day; λ is weight coefficient which reflects the influence of $L_{d-m,t}$ on $L_{d,t}$, λ is defined as:

If $m \neq 1$, then,

$$\begin{cases} \lambda_j = \beta(1-\beta)^{j-1}, \beta \in (0,1), j = 1,2,...,m-1 \\ \sum_{j=1}^{m} \lambda_j = 1 \end{cases} \quad (4)$$

else $m = 1$, then,

$$\lambda_1 = \beta, \ \beta \in (0,1] \quad (5)$$

Where, β is smooth coefficient, $t = 1,2,...,24$.

Corrected $L_{d,t}$ is the sum of historical data at time t multiplied by different weight coefficients.

Compared with traditional threshold method, the improved method can not only effectively identify continuous missing data by formula (2), but also can avoid miscalculation caused by its reference value[10].

3. Main title

3.1. Method Description

The basic principle of data density estimation method is:

Suppose a collected two-dimensional data set *Z* which consists of *M* data points (the dots shown in Fig.1) and generate a seeds group *S* (the circles shown in Fig.1) which contains N seeds. The distance between seeds should be constant and the scope of seeds group must be large enough to cover the data set *Z*. Each data point z_j ($j \in \{1,2,...,M\}$) should be accompanied by a seed adsorption counter c_i whose initial value is zero. The seed adsorption counter is applied to sum up the absorbed seeds number. The seed adsorption counter value can be obtained by calculating the distances between data points and seeds.

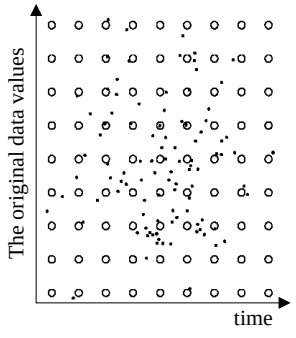


Figure 1. Simplified data density scheme

More specifically, calculate the distance between each seed s_i ($i \in \{1,2,...,N\}$) and each data point in set *Z* separately. Assume that z_k is the closest point of seed s_i, the sequence of neighbouring data point z_k is determined by the following formula:

$$k = \arg\min(\|s_i - z_j\|^2) \quad (6)$$

Where, $i \in \{1,2,...,N\}$, $j \in \{1,2,...,M\}$, $\|\cdot\|$ stands for the Euclidean distance, 'arg' is the abbreviation of argument. Formula (6) means the value of *j* is *k* when getting the minimum. Once z_k which is the closest point of seed s_i is detected according to formula (6), then the seed adsorption counter c_i that data point being attached to will be increased by one. If there are p data points with the same nearest distance to seed s_i, then the increment will be distributed equally among these data points. In other words, the seed adsorption counter of each data point is added by 1/p. The closest neighbour can be found through calculation with respect to formula (6) for each seed in the seeds group *S*. And keep updating seed adsorption counters in accordance with above rules until all the seeds are calculated.

The basic principle of detecting abnormal data is: a higher value of seed adsorption counter indicates that the corresponding data point attract more seeds. It

means there are few data points participating in competition with this specific data point around neighbourhood. Hence, it is a data with low density. Conversely, if the data has many other points nearby, there will be a fierce competition in adsorbing seeds among these points. It is clear that the value of seed adsorption counter attached to each data point becomes lower correspondingly. Therefore, the data whose seed adsorption counter value is higher than the set value can be classified as abnormal data, the set value is called seed absorption threshold [11].

3.2. Parameter Setting

The algorithm needs to determine two parameters which are seeds number N and seed adsorption threshold. In order to determine the seeds number, a simple and heuristic method is introduced:

1) Calculate the shortest distance between each data z_i and other data points:

$$d_i = \min(\|z_i - z_j\|^2), i, j \in \{1,2,...,M\} \text{and} i \neq j \quad (7)$$

2) Compute mean value of the shortest distance among data points according to equation (8), take it as the distance between neighbouring seeds:

$$\bar{d} = \frac{1}{M}\sum_{i=1}^{M} d_i \quad (8)$$

3) Determine the scope of seeds which can cover all data points. Assuming that one dimension of the data set ranges from $z_{\min}$ to $z_{\max}$, the upper boundary $s_{\max}$ and the lower boundary $s_{\min}$ of the seeds set in this dimension should meet:

$$\begin{cases} s_{\max} - z_{\max} > \bar{d} \\ z_{\min} - s_{\min} > \bar{d} \end{cases} \quad (9)$$

4) Calculate the seeds number after the determination of seeds scale and distance.

Seed absorption threshold can be determined according to overall distribution of seed absorption counter value. The steps to get seed absorption threshold are: Firstly, get the seed absorption counter value in ascending order. Secondly, set a seed adsorption threshold. If the counter value is higher than the threshold, the corresponding data will be modified according to equation (3). In order to obtain better results, the threshold can be adjusted flexibly depending on the specific circumstances [12].

3.3. Procedure Based On Density Estimation

The abnormal data processing approach based on density estimation can be summarized as:

1) Convert the load sequence into a two-dimensional data matrix;

2) Determine the seeds number N;

3) Generate a seeds group S with constant spacing;

4) The initial value c_k of the seed adsorption counter attached to data point z_k is zero;

5) Calculate the distance between the seed s_i and all data points, searching for the nearest data point z_k of the seed s_i, and update its seed adsorption counter c_k;

6) Repeat step 5 until the completion of all seeds processing;

7) Determine the seed absorption threshold;

8) Identify those abnormal data and revise them according to equation (3).

The processing flowchart is shown in Figure.2.

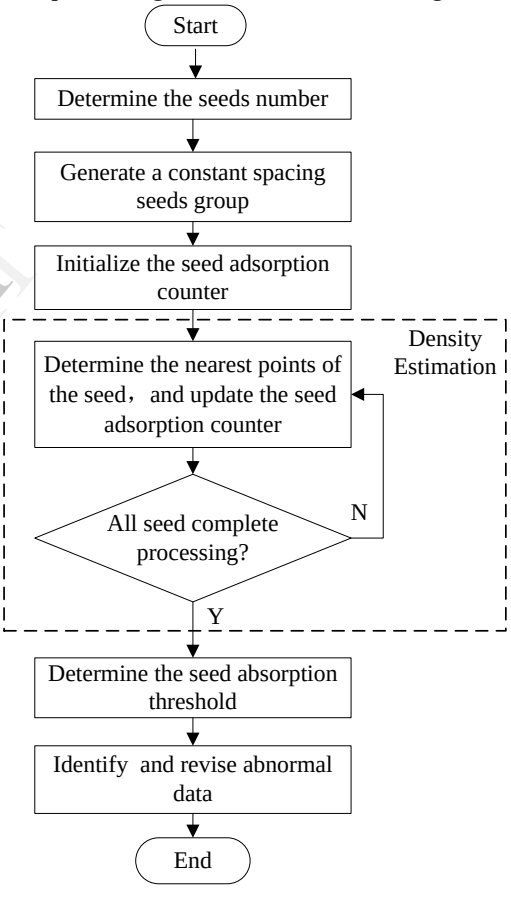


Figure 2. Flowchart of abnormal data processing

4. Simulation Analysis

The sample data are obtained from a heating power plant. From September 12, 2012 to December 11, 2012, take 91 days heating load data as example. According to the results of calculation, $\bar{d} \approx 0.00094$, $z_{\min} = 0.06$, $z_{\max} = 2.57$. Approximate

$\bar{d} = 0.001$. $s_{\max} = 2.571$ and $s_{\min} = 0.059$ are calculated by formula (9). The density estimation of 3D graph is shown in Fig.3, where the abscissa is the sampling time in one day with an interval of 30 minutes. The ordinate is the number of days, and the vertical coordinate represents the seed adsorption counter value corresponding to each data point.

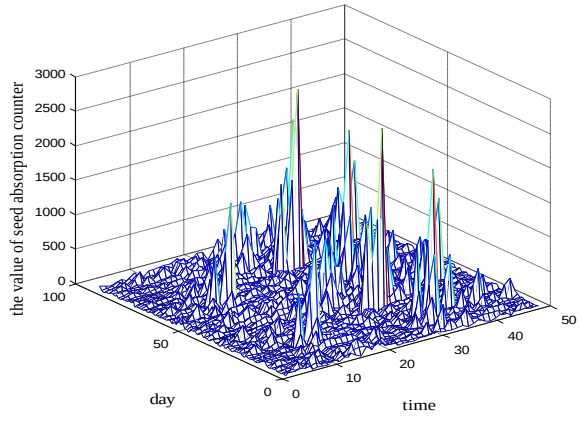
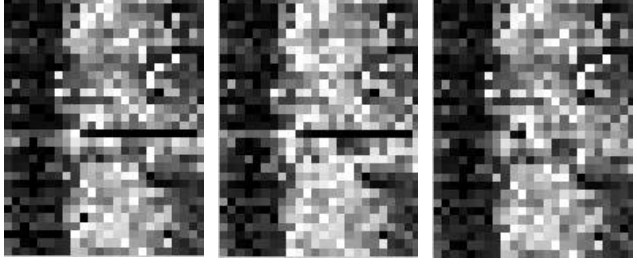


Figure 3. The density estimation of 3D graph

Because the ratio of abnormal data in short-term load forecasting does not exceed 3%, take the smooth coefficient β as 0.5. According to the "the closer, the bigger" principle and the constraint of λ , define λ_1=0.5 , λ_2=0.25 , λ_3=0.25. Combined with the improved methods, set the seed absorption threshold as 500, and 76 abnormal data points can be identified, accounting for about 1.74% of total data set. The following analysis will show the effect of abnormal data processing from two aspects.

4.1. Data Identification

Being normalized, the two-dimensional data matrix can be transformed into a new matrix in which elements are between 0 and 1. Because the normalized matrix is similar with grey image matrix, continuous missing data processing could be more intuitively characterized by grey image. Thus, convert the partial data which contains continuous missing data into grey graph, which is shown in Fig.4. A good identification result can be achieved and the noise in image is reduced obviously after effective treatment.



a) Grey image of unhandled data | b) Grey image of traditional method | c) Grey image of improved method

Figure 4. Comparison in grey image

Dealing with the heating load in horizontal or vertical direction respectively, traditional threshold method can identify 35 and 47 data correspondingly, and there appear 16 repeated identified data in both directions. Hence, the horizontal and vertical threshold method can detect 66 abnormal data in total. Compared with traditional threshold method, density estimation algorithm with the threshold of 500 can identify 76 abnormal points, and 55 of them are the same as traditional threshold method. Comparison result of the two methods is in tab.1. In case of the similar abnormal data rate, the method based on density estimation is more simple, feasible and effective.

Table 1. Comparison results of different methods

Algorithm	Traditional threshold method		Density estimation
	Horizontal processing only	Vertical processing only	
Threshold value	0.98	1.25	500
Number of abnormal data	35	47	76
The ratio of abnormal data/%	1.51		1.74
Overlapping	55		

4.2. Evaluation By Accuracy Rate Of Daily Load Prediction

Central heating system is a complicated multi-variable control system. Its characters of large heating area, many influence factors, strong internal relevance, long time delay and serious nonlinearity cause some difficulty in load forecasting. However, Radical Basis Function (RBF) neutral network which has been widely used in time series analysis and non-linear control can process any non-linear functions. Therefore, RBF neutral network is usually used to forecast the load data regardless of traditional and improved method. Taking the accuracy rate of daily load prediction as the evaluation index of prediction effect, it is defined as [15]:

$$A = (1 - \sqrt{\frac{1}{24}\sum_{i=1}^{24} E_i^2}) \times 100\% \qquad (10)$$

Where, E_i is the relative error at time i of the forecasting day, and A is accuracy rate of daily load prediction.

Load forecasting results in one week are shown in Tab.2("--" indicates the data of 0 value in the data set

which is unable to quantify the relative error). The proposed identification method based on data density estimation can well identify continuous missing data. The average prediction accuracy improved by 1.73%, its prediction effect is superior to traditional processing method based on single dimensional space.

Table 2. Comparison results of load forecasting accuracy

Time	Daily load accuracy/%		Accuracy improved by /%
	Traditional threshold method	Improved method	
Monday	97.55	98.49	0.94
Tuesday	96.36	98.72	2.36
Wednesday	--	99.47	--
Thursday	--	97.78	--
Friday	96.98	98.25	1.27
Saturday	95.95	98.42	2.47
Sunday	96.63	97.84	1.21
Mean value	96.694	98.424	1.73

5. Conclusion

Given the horizontal and vertical continuity characteristics of heating load， a novel algorithm which is capable of detecting and modifying anomalous data in two-dimensional space based on data density estimation is put forward. It can avoid the deficiencies of single dimensional space processing, eliminating the abnormal points from overall data and correcting them once again. The average prediction accuracy is improved by 1.73%. The results of example analysis and simulation results verify that the anomalous data identification method based on data density estimation is more feasible and effective than traditional methods.

6. References

[1] Wang Lei, Zhang Ruiqing and Sheng Wei. Regression forecast and abnormal data detection based on support vector regression, *Conference*, Proceedings of the CSEE, 2009, 29(8):92-96.

[2] Song Yongqi. Study on new measuring and control device in household metering heating system, *Dissertation*,. Harbin: Master's thesis of Harbin Institute of Technology, 2010: 1-18.

[3] Pang Qiang, Yuan Mingzhe, Zou Tao. On-line rectification method of flow measurement error in steam pipe network and its application, *Journal*, Chinese Journal of Scientific Instrument, 2013, 34(1): 46-50.

[4] Zhang Xiaolei, Zhang Yanyan, Tang Lixin. Steam allocation plan considering production and electricity generation, *Journal*, Control Engineering, 2012, 19(6): 997-1002.

[5] Niu Dongxiao. Echo state network with wavelet in load forecasting, *Journal*, Emerald Journal, 2012, 41(10): 1557-1570.

[6] Gao Shan, Shan Yunda. A new method of load data error-correction, *Conference*, Proceedings of the CSEE, 2001, 21(11): 105-108.

[7] Ye Feng, He Hua, Gu Quan. Bad data identification and correction for load forecasting in energy management system, *Journal*, Automation of Electric Power Systems, 2006, 30(15): 85-88.

[8] Gu Min, Ge Liangquan, Qin Jian. Identification and justification of dirty electric load data based on modified ART2 network, *Journal*, Automation of Electric Power Systems, 2007, 31(16): 70-74.

[9] Tong Shulin, Wen Fushuan, Chen Liang. A two-dimension wavelet threshold de-noising method for electric-load data processing, *Journal*, Automation of Electric Power Systems, 2012, 36(2): 101-103.

[10] Li Guangzhen, Liu Wenying, Yun Huizhou. A new data preprocessing method for bus load forecasting, *Journal*, Power System Technology, 2010, 2(34): 150-151.

[11] Wang Yang. A novel algorithm for outlier removal based on density, *Journal*, ACTA AUTOMATICA SINICA, 2010, 36(2): 333-346.

[12] Chen Liang, Wen Fushuan, Tong Shulin. A method to identify and correct the abnormal electric-load data based on density evaluation, *Journal*, Journal of South China University of Technology(Natural Science Edition), 2012, 40(2): 124-129.

[13] Zhan Tengxi, Guo Guanqi. Intelligent hybrid prediction method of the flue gas oxygen content in power plant, *Journal*, Chinese Journal of Scientific Instrument, 2010, 31(8): 1826-1833.

[14] Liu Yanwei. Research on the heating system load forecasting based on natural network, *Dissertation*, Tianjin: Master's thesis of Tianjin University, 2009: 16-25.

[15] Li Ruqi, Chu Jinshen, Xie Linfeng. Application of IAFSA-RBF neutral network to short-term load forecasting, *Conference*, Proceedings of the CSU-EPSA, 2011, 23(2): 142-147.