Pre-processing and audit of power consumption data based on composite mathematical statistics model

Ming Chen, State grid Shanghai municipal electric power company,Shanghai, China
Zengrui Huang, School of Computer Science, Fudan University; Qiang wu, Wei Xu, Boyue Xiong, State grid Shanghai municipal electric power company,Shanghai, China

Abstract: With the promotion of intelligent power consumption information acquisition system, power companies can easily obtain a large number of real and effective data sets from the acquisition system. These data sets can help power companies fully grasp the user's power consumption and then analyze user behaviors and power consumption characteristics. However, due to signal interference and other factors, in the currently running acquisition systems, some "dirty data" is inevitably mixed in. Due to the huge amount of data collected by the power system, the total amount of these contaminated data cannot be ignored, and the data extraction and analysis work caused a great deal of interference. Traditional outlier detection methods rely on the intuitive judgment of experienced staff. This paper attempts to design a power data preprocessing model based on mathematical statistics and using detection methods commonly used in mathematical statistics. This can help power grid workers improve the efficiency of auditing original abnormal data and optimizes the quality of the data sets.

Keywords: Smart grid; Composite mathematical statistics model; Data cleaning and preprocess

I. INTRODUCTION

The power system is the foundation of a country's industrialization. In recent years, the increasingly emerging concept of "smart grid" [1], is to ensure the stable and efficient operation of the grid. The power consumption data generated by the users in the smart grid operation is also a very valuable resource. With these data, the power company can mine various information, analyze the users' behavior habits, and obtain the user's full power consumption status, so that further analysis conclusions can be obtained [2]. Due to the large number of user groups, the differentiation of the smart meter collection environment, and the various uncontrollable factors such as equipment terminal failure and data transmission interference, in the actual operation process, the power data condition collected by the grid acquisition system is very complicated [3]. There are many cases of abnormal data in the data set, such as negative values, missing values (null value), abnormal zero values, extremely large values, relationship error data, redundant data, etc. [4]. However, for subsequent data analysis and decision-making work, many models and conclusions are based on idealized data set, and the abnormal data in the collected data sets will be misleading [5].

With the popularity of the concept of "Big Data" and the continuous development of data science in recent years, it has brought a new revolution to many traditional fields, and the power industry is no exception. Starts from the actual situation that the current data audit of the power company is inefficient, this paper proposes a model for efficient audit and data preprocessing based on the mathematical statistics methods. This model uses mathematical statistics to screen for extremely large, negative, and null values in the data set, and to identify certain special data anomaly categories. Through this model, data items for subsequent power behavior analysis can be extracted from a huge database system, and redundant data existing in the user power data set can be cleaned up, and the missing data can be filled according to the relationship between the data table items. This model can also correct the data of the disordered relationship, unifies the data format, and eliminates various influencing factors that may interfere with the later analysis work. In summary, the data preprocessed by this model can not only reduce the interference of noise data, optimize the data set, but also improve the efficiency of data audit and reduce unnecessary time waste.

II. SELECTION OF ANOMALY DATA IDENTIFICATION METHOD

2.1 Data quality overview

The electricity consumption data selected in the paper is more than 100 million pieces from 1.6 million users in Shanghai from June 1, 2017 to August 19, 2017. Due to factors such as power information collection equipment anomalies and data transmission signal interference, there are many kinds of data anomalies in the original data set, which can be roughly classified into the following categories. see table 1：

Error type	Detailed description
Null value	This data is not recorded in the database, and the daily electricity consumption graph is displayed as a breakpoint.
Negative value	The electricity data is recorded as a negative value
Extremely large value	The size of the electricity data far exceeds the normal power consumption value of the user.
Logical error	The items do not match the conversion relationship or the data is unreasonable.

978-1-5386-8549-5/18/$31.00 ©2018 IEEE

Table 1

2.2 Data preprocessing flowchart

The general flowchart of data preprocessing is as follows:

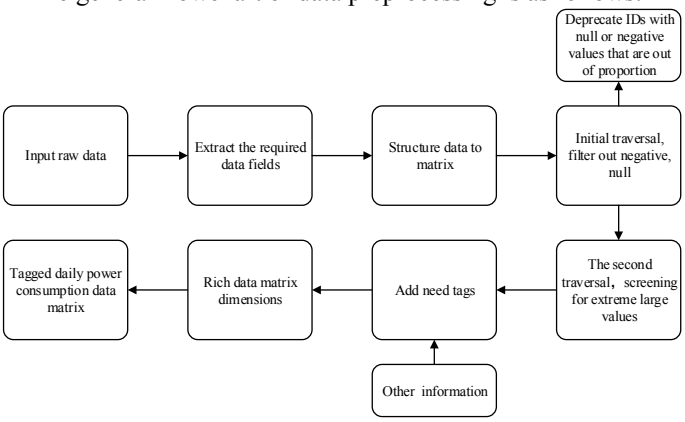

Figure 1 Data preprocessing flowchart

After got the original data set, we need to extract the required table items and data fields from the original data set and process them into the structured data we need. Then we check the power consumption data for negative and negative values. For the user data with a certain percentage of null and negative values, we need to discard it [6], then use our mathematical statistics method to filter out extreme values. The electricity data is used to generate an identifiable feature data matrix, and the matrix can be used for initial classification and identification by the user for further data analysis work. Some data visualization work is also performed on the data preprocessing. The results of the screening are output to the downstream.

2.3 Data screening method

The core issue of data screening is the principles and methods [7]. After extracting the required data from the original database, we give priority to the easy-to-discover anomaly data such as null and negative values to avoid interference with subsequent extremely large value screening. After the two types of abnormal data users are screened out, we need to divide the results into several cases. In the case of user data with a large proportion of null and negative values, choose to delete directly from the database, because these data have been missing a lot of useful information, which has little meaning for subsequent analysis; for users with small number of null and negative values, we choose to keep and do further processing.

In order to further screen the extremely large value data in the dataset, we chose two methods in statistics, the quartile detection method [8] and the Z-Score method [9].

The quartile method is used in statistics to describe the degree of dispersion of a data set. The process is to sort all the data from small to large, and then divide the data set into four equal parts. The upper limit of the first part is the first quartile Q1, and the upper limit of the second part is the second quartile Q2, which is the median, and the upper limit of the third part is the third quarter Q3, the most important data indicator of this method is the interquartile range IQR, which is the difference between the third quartile and the first quartile.:

$$IQR = Q_3 - Q_1$$

The third quartile Q3 plus K times the interquartile range IQR is used as the abnormality detection upper limit UL of the data set, and the first quartile Q1 is subtracted by K times the interquartile range IQR as the data set anomaly detection lower limit LL, they are calculated as follows:

$$UL = Q_3 + K \times IQR$$
$$LL = Q_1 - K \times IQR$$

The coefficient K can be adjusted according to the situation. When the general K value is set to 1.5, the abnormality detected is a general abnormality. When the K value is set to 3, the abnormality detected is extremely abnormal. Due to the complexity of the user's power consumption, the data tends to fluctuate greatly. In this paper, the K value is set to 3 considering the actual situation. The quartile method can usually be visualized as a boxplot[10]:

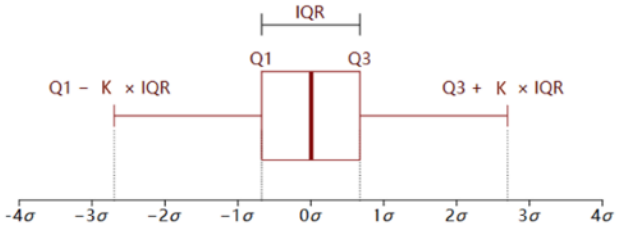

Figure 2 Boxplot

The Z-Score method, is a reduction method in statistics that analyzes the overall deviation of each data and data set by calculating the standard score Z-Score of each data in the data set. The essence of Z-Score is a dimensionless quantity, which is a pure digital mark of each data. It is converted from the original value of the data. The calculation method is as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - u)^2}$$
$$Z = \frac{x_i - u}{\sigma}$$

The u in the formula refers to the mean of the data set, and the σ refers to the standard deviation of the data set. It is generally considered that the Z-Score is a normal value at -3 to 3, and abnormal when the absolute value is greater than 3, and the threshold can be adjusted according to the actual situation [11].

Both the quartile detection and the Z-score method have their own one-sidedness, and they cannot fully synthesize the dataset in the case of individual use. For the quartile method, it is easy to see an overall overview of the data distribution, the calculation efficiency is also very high, but the exact values and details of the data distribution are not preserved [12]. For the Z-score method, it needs to calculate the Z-score value for each point which requires a lot of calculation and the efficiency of data processing will be affected. Therefore, we

combine these two methods and take the intersection of the data results after their screening as our final screening results.

III. Experiment Procedure

3.1 Experimental environment configuration

The experimental host configuration is as follows: The CPU is i7-7700, the main frequency is 3.6GHz, and the memory is 16GB. The selected development environment is Python 3.6.4 [MSC v.1900 64 bit (AMD64)] on win32, which integrates data analysis libraries such as numpy, pandas, plotly, etc.

3.2 Data extraction

The original experimental data is an Oracle database backup file, which contains dozens of data tables and thousands of data fields. We use the CX_Oracle library to extract the user ID, power use date, daily power consumption, and positive active power in the database. Several field such as the indications under the four rates are used as the analysis data. The format of each data is as follows:

ID	Date	Daily power consumption	Active power	Rate1	Rate2	Rate3	Rate4

Table 2

We firstly do a special treatment on the date and convert it to the corresponding number of the date in the year. Next, we did an interval analysis on daily electricity consumption. If one user's electricity consumption per day is regarded as one piece of data, after analyzing more than 100 million pieces of electricity consumption data, the interval distribution histogram is obtained, as shown in Fig. 3.

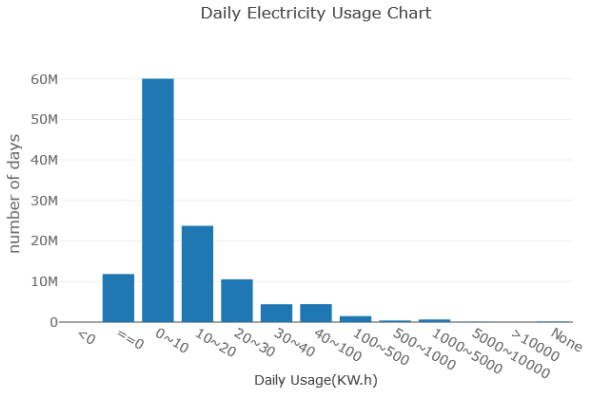

Figure 3 Electricity data histogram

As can be seen from the above figure, most of the electricity consumption data is concentrated between 0 and 100, especially from 0 to 30. In the typical anomaly data, there are more than 10,000 negative data, and nearly 100,000 null values, more than 50,000 extreme large-value data which is larger than 5000. These data are the main goals of our analysis and screening.

2.3 Data screening

For the above extracted data, we need to further simplify the processing for later analysis. We use the numpy library in Python to process the power consumption of each user into a two-dimensional matrix, where the first data of each matrix to the penultimate data is the power consumption data of the user's corresponding date, and the final data is the user ID of the user. Then use the code to screen all data for negative and null information, generate an abnormal user information table with negative and null values, as shown in Table 3.

[ID, Null number, Negative value number, Null date, Negative date]
[2405855, 5, 2, [158, 161, 174, 186, 204], [190, 210]]

Table 3

Through analysis, it is found that the number of users with null and negative values accounts for 2.37% of the total users, but most users only have one null value or one negative value. In general, if the number of null values of a user's power consumption data exceeds 20% of the total power consumption data, it is highly likely that the important information in the user behavior characteristics will be lost, so users with more than 20% of the null value data will be directly screened out. Specific to this experiment, if the null value number of user data is more than 16 days, this piece of data will be directly screened out. In the case of negative values, the situation is more complicated. The daily meter count is calculated based on the cumulative value of the positive active power. The positive active power difference between two adjacent days is the electricity consumption data. If the meter fails, and we replace it with a new one. the positive active power will be cleared to zero, so the difference between the day and the previous day will be negative, and the power consumption on this day will have a negative value. We can dig out the users who replace the failed meter from the users with negative values, but similar situations occur when the signal transmission is disturbed or the metering device fails. How do we distinguish between these two situations? Usually a user's meter will only be replaced once a month, so we treat two or three times a year as an anomaly.

After processing the negative and null values, we need to further analyze the abnormal large value data in the user's power consumption data. We chose quartile and standard score detection methods that do not require data set distribution. Firstly, we use quartile method to screen out the outliers and we record each user ID of the outliers. Because many home users may have very stable daily electricity consumption, once a reasonable amount of data fluctuations occur on a certain day, they will be detected by the system. Therefore, we need to use Z-score method to filter the result detected by the quartile detection method, and use the intersection of the two screening results as the final screening result, so that we get a composite mathematical statistics screening model. Its main body is based on the quartile detection method, but its screening result is partially corrected by the Z-score method. Finally, the screening result is processed into a matrix format for further analysis.

2.4 Typical anomaly preliminary classification

Based on the previous analysis of the measurement data, combined with the practice of power operation and maintenance, we can make some preliminary classifications for typical anomalies.

When the user has a negative value and its absolute value is extremely large, it is very likely that the user has changed the meter. The user can be screened out as a user classification for changing the meter behavior. Figure 4 shows the typical power consumption curve of the meter change user, because the absolute value of the extreme power consumption data is too large, so the ordinate is the logarithm of the user's power consumption.

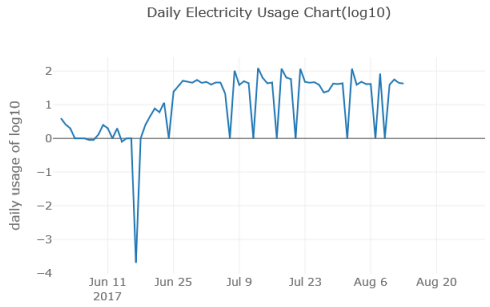

Figure 4 Meter change user electricity consumption curve

Due to smart meters are usually replaced only once in a few years, if there are two or more times in the data date of just over two months, we need to consider whether there is a problem with the signal transmission or facilities, which is classified as a meter failure. The typical power consumption diagram is shown in Figure 5.

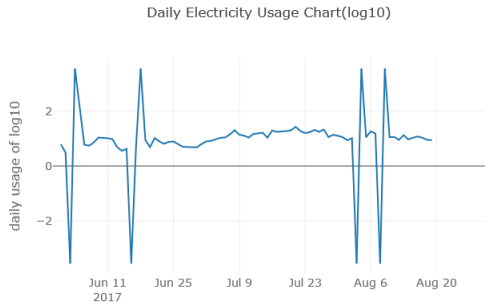

Figure 5 Metering device abnormal user electricity consumption curve

For the users who made the preliminary classification above, we manually mark the data, which is convenient for the subsequent use of machine learning algorithms to mine deeper information from the data set.

IV. Conclusion

The model of preprocessing and auditing electricity data based on the composite mathematical statistics model can efficiently and quickly extract the required data items from the original data set, screen out the abnormal and interference data, and structure the data. The data quality of the data set can be greatly optimized, and some typical feature data in the user data, such as meter change and meter failure, are initially classified and determined, which also facilitates further analysis and mining of the data information.

The future work is mainly to summarize more data features from the existing data set, so that more data can be classified in the data preprocessing stage, and more meaningful tags can be added to the data, which will be helpful to the optimization and improvement of the model.

V. References

[1] Grid Modernization and the Smart Grid [EB/OL]. https://www.energy.gov/oe/ grid-modernization-and-smart-grid
[2] Hongfei Sun, Jiaran Ni, Li Peng. The core technology of smart grid big data [J]. Popular Utilization of Electricity, 2017(2).27-28
[3] Hernandez, M.A., Stolfo, S.J. Real-World data is dirty: data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 1998,2(1):9-37.
[4] Xiaoxing Zhang, Qiyun Cheng, Quan Zhou. Dynamic intelligent cleaning of dirty data of power load based on data mining [J]. Automation of Electric Power Systems, 2005, 29(8):60-64.
[5] Jun Wu. Research on cleaning model of electric load data based on neural network [D]. Dalian University of Technology, 2010.
[6] Yuankun Liu, Wenpeng Luan, Yan Xu. Data cleaning method for distribution transformer[J]. Power System Technology, 2017, 41(3):1008-1014.
[7] Yaqi Song, Guoliang Zhou, Yongli Zhu. Current Status and Challenges of Smart Grid Big Data Processing Technology [J]. Power System Technology, 2013, 37(4):927-935. .
[8] Dieter Rasch．Mathematical Statistics[M]．Washington: Wiley Publish, 2012，37-40．
[9] Cheadle C, Vawter M P, Freed W J, et al. Analysis of microarray data using Z score transformation[J]. The Journal of molecular diagnostics, 2003, 5(2): 73-81.
[10] Abdi H, Williams L J. Newman-Keuls test and Tukey test [J]. Thousand Oaks, CA, 2010, 5-7 [11] *Mendenhall William, Sincich Terry. Statistics for Engineering and the Sciences (Fifth edit.) [M] London: Pearson 2006*
[12] Advantages & Disadvantages of a Box Plot[EB/OL]. https://sciencing.com/advantages-disadvantages-box-plot-12025269.html