ISSN: 2472-3703 DOI: 10.6911/WSRJ.202002_6(2).0011

Faulty Electrical Components Data Analysis and Modeling Based on Statistics

Guanying Jiang[1, a, *]

[1]College of Economics, Jinan University, Guangzhou 510632, China

[a]Corresponding author e-mail: 543308671@qq.com

Abstract

It is essential for electrical components to maintain a stable working condition in the industrial production process. The detection of various gas concentrations can assist in identifying the faulty status. To this end, in this paper, we propose a novel detection scheme to improve the efficiency and accuracy of faulty electrical components identification based on a gas concentrations dataset.

Keywords

Ensemble learning, Gas concentration, Faulty electrical components, Data preprocessing.

1. INTRODUCTION

Missing, null, abnormal negative and zero values of explanatory variables are summarized. Data for H_2 CH_4 C_2H_6 C_2H_4 respectively contained 14, 28, 81, and 83 zero values, accounting for 2.29%, 4.58%, 13.24%, 13.56 %, which is relatively small and will not have much impact on subsequent analysis. However, C_2H_2 contains 434 zero values, accounting for 70.92% of the total observations. Based on it, the normalization for C_2H_2 is meaningless and may even lead to biased results. Other than this, it is found that 89.89% of C_2H_2 non-zero observations are faulty components, indicating that C_2H_2 non-zero records may be the main feature of faulty components. However, C_2H_2 cannot act as the only indicator for faulty detection since 6.22% of C_2H_2 zero records belongs to faulty components.

2. DISTRIBUTION CHARACTERISTICS

To understand the distribution characteristics of explanatory variables, the Shapiro-Wilk normality test, the normal QQ chart, the robust statistics, and the boxplots will be conducted in this section. An unbalanced sample distribution can be found while correlation matrix is used to explore the dependencies among all variables.

2.1. Normality Test for Explanatory Variables

The data can be considered to be approximately normal only when the test statistics are close to or equal to 1 in the Shapiro-Wilk test. The test results of the six explanatory variables are nearly 0.1 and the p-values are all far less than 0.05, except for the NO_2 above 0.9. Therefore, the null hypothesis is rejected, and the explanatory variable are not regarded as normally distributed data. However, referring to the normal Q-Q chart, NO_2 data presents the most approximately normal shape, so that a further analysis for it is required.

2.2. Quantile Characteristics of Explanatory Variables

The descriptive statistics can be calculated for each explanatory variable using R software. It is observed that there are big gaps between the upper quartile and the maximum value for variables except NO_2, in which the minimum gap is a hundred times than the latter. The main

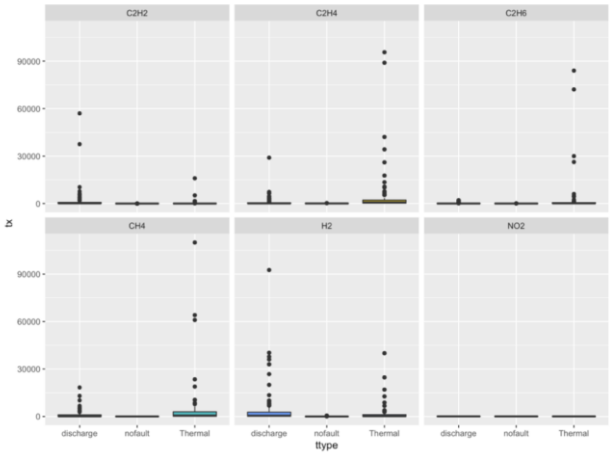
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202002_6(2).0011

reason is that there are too many extreme outliers above the upper quartile and that the mean and standard deviation statistics are not robust but extremely sensitive to outliers. Therefore, the location and scale parameters may be overestimated and thus affect the model's prediction. We recalculate robust statistics as follows:

Median(x) ± 1.58 * IQR / sqrt (n),

Where IQR is the difference between the upper and lower quartiles, and n is the sample size. Given the recalculated maximum and minimum values, we mark the data points that are not in the range as outliers. The number of abnormal observations of the gas concentration data are 127, 83, 66, 78, 0, and 121 respectively.

As shown in boxplots, the six variables except NO_2 present similar distribution patterns. There are evident abnormalities under the faulty states such as discharge or thermal while no abnormal records exist under the normal status for electrical components. Due to the dimensional problem, the concentration of NO_2 does not show significant differences under various states. Yet, it can be found that its mean level increases as the degree of faulty becomes worsen. Under this circumstance, it is still unclear whether NO_2 is helpful in predictive models so that we leave it as a candidate.



2.3. Correlation Analysis Between Variables

The correlation analysis for each variable pair is conducted, in which not only the dependencies of the explanatory variables but also the linear interpretation of explanatory variables to the response one can be revealed. Variables with strong correlation are supposed to be transformed or removed with the aim of avoiding multicollinearity. According to the correlation matrix, it can be seen that the correlation coefficients between variables mainly fall into the range [0.5, 0.9] and that the association with the response variable is still weak. The results based on Kendall τ correlation coefficient are similar to that of Spearman.

3. ANOMALY PROCESSING

3.1. Detection Stage I

In the abnormal dataset, the above identified abnormal data points are tagged and removed the duplicates. 475 data points are tagged in total in the explanatory variable matrix, accounting for 12.94% (= 475 / (612 * 6)) of the total number of observations. Among them, the variable containing the most abnormal data points is H_2, with totally 127 anomalies. NO_2 is the only variable that contains no exceptions. In fact, the number of cases with at least an abnormal variable is 187, accounting for 30.56% (= 187/612) of the total number of observations. The

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202002_6(2).0011

proportion is no more than half of the observation, and it can be still assumed that the remaining data can represent the overall characteristics.

The normal and abnormal datasets are then split. Given the two new datasets, we present the frequency analysis of the response variables. There are only 18 faulty-free observations (accounting for 9.626%) in the anomaly dataset, remaining 90.37% faulty components. As for the normal dataset, 407 components are faulty-free cases (95.765%) while the rest 18 ones are faulty (4.235%). The actual fault observations total 187 (115 discharges, 72 Thermal) whereas the abnormal one accounts for 169 (108 discharges, 61 Thermal), approximate to 90.37%, indicating that the rest 9.63% are in the normal dataset.

Through a basic anomaly identification process, we can effectively detect around 90% of the faulty components, which will save a lot of manpower and resources in practice. Thus, it is sensible that the model can only detect the anomalies from the abnormal dataset since they have significant differences with those in normal dataset. However, the rest faulty cases have trivial distinction with the actual normal ones, which are hard to extract for these models. Therefore, aiming for the rest 9.63% faulty components, we have to train a more robust identification model. Apart from these, in production activities, the cost of failing to report a faulty component is much higher than that of falsely reporting a normal component, so that the 9.63% missing observations cannot be ignored due to the high recognition rate of 90.37% of the basic method. In other words, modeling process is supposed to focus on reducing the false negative rate but not only the overall false alarm rate.

3.2. Detection Stage II

In the secondary abnormality identification stage, the extracted position for anomaly and its statistics are summarized as follows. It is found that 16 of the 39 newly identified observations are faulty components, accounting for 41.03% (= 16/39). Though the precision is not so significant as in the first stage, the two-stage identification achieves 98.93% (=(169+16)/187) coverage of faulty components, in which only 2 cases left out. For this sample data, the two-stage method can solve the problem of fault detection efficiently. However, it may not suit for other datasets. Thus, the subsequent modeling is still based on the results of the first detection stage.

3.3. Alleviate Abnormal Data

Given that the anomalous data accounts for about 1/3 of the original dataset, data information may be lost if it is directly removed. To make the abnormal data also applicable, we displace the abnormal data points with the robust maximum values estimated in the previous stage. The distribution of the processed dataset is not changed theoretically, with the compressed dimensions mainly.

4. EXTENSIONS

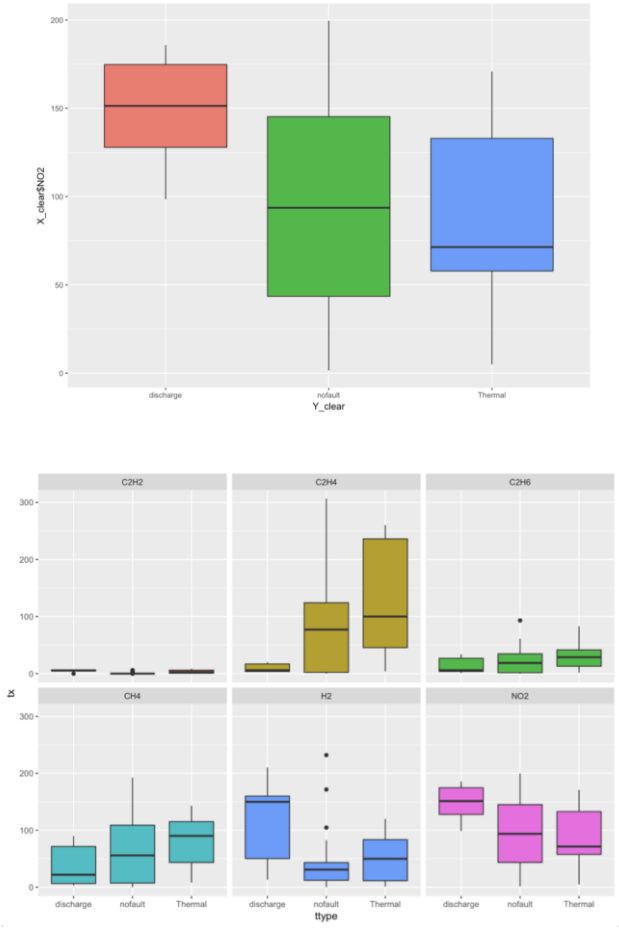
Three extensive discussions are shown as follows. Whether the seemingly useless NO_2 needs to be removed, whether there exist leading indicators, and the process control chart for anomaly detection will all be revealed in this section.

4.1. Delete the Explanatory Variable NO_2 or Not?

In the above boxplots, the mean levels of NO_2 have no significant variations as other factors. Yet, in the following re-drawn boxplots, the distribution characteristics are evidently different from the previous one. It is found that the distribution of NO_2 has the most significant changes, in which the boxplot of the discharge status present higher mean level than the rest two statuses. That is, in the newly obtained dataset, NO_2 will be an effective indicator to identify the discharge status and thus we will not eliminate it in the preprocessing stages in this research.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202002_6(2).0011

The distributions of the remaining explanatory variables also vary in different degrees. The C_2H_2 for nonfaulty data are all zero and that for thermal or discharge fall in the interval (0, 8.5). In particular, C_2H_4 and H_2 can help to distinguish discharge status from the other two. The patterns for thermal and nonfaulty are so close that we ought to deal with them through modeling.

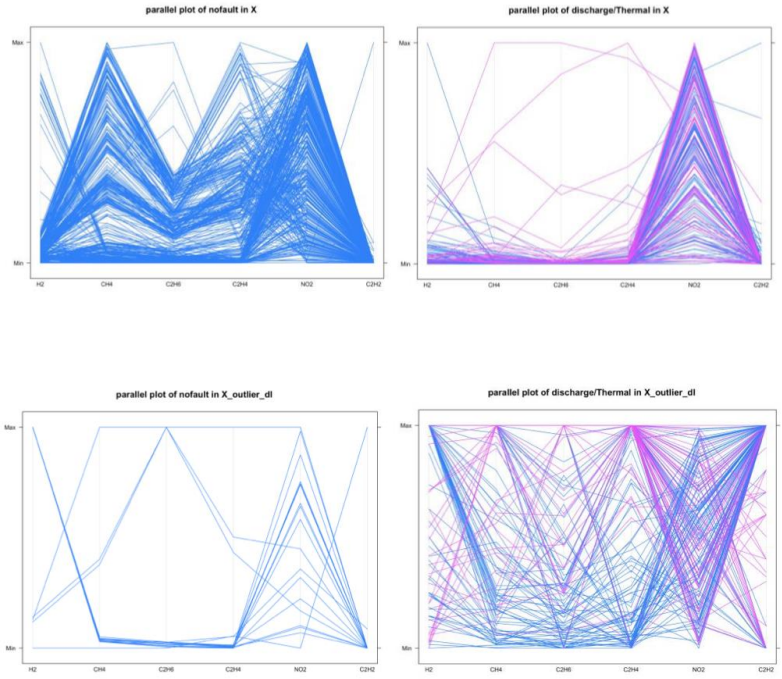


4.2. Are There Early Warning Indicators?

Parallel coordinate charts, also called contour maps, can plot the values of all variables for observations on the same chart in order to intuitively understand the distribution characteristics of each cases and how the overall pattern of the sample dataset. There are still many faulty observations in the secondary anomaly detection stage, indicating that indicators have different priorities during the fault generation process.

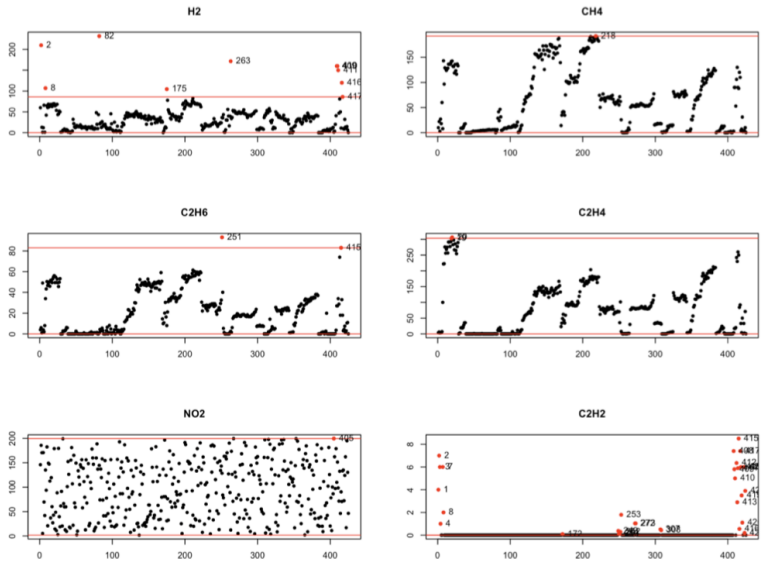
The fault-free graph patterns of the abnormal dataset and the processed abnormal one is the same. Besides, they are also similar to that in original dataset, which explains why these normal components are misjudged as abnormal data. However, we can also regard them as potential failure components. Even though the failure has not yet occurred, it may be at that edge. In practice, troubleshooting of these components can be considered as a precautionary measure.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202002_6(2).0011



4.3. Constructing Anomaly Detection Control Charts

Apart from the statistics, statistical process control chart can help achieve real-time anomaly monitoring and present visualization results. The single-indicator control charts are shown as follows. The red control limits, corresponding to the maximum and minimum robust statistics, work as warning lines. Once new components are monitored, the control chart can detect the anomaly in time. The control chart can be utilized in the first detection stage, especially for non-statistical professionals.



5. ENSEMBLE LEARNING METHODS

Several classification and prediction models are selected for comparison and analysis, including Back-Propagation Neural Network (BP), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors Algorithm (KNN), and Gradient Boost Ensemble Learning method (GB).

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202002_6(2).0011

Evaluation index of model effect includes confusion matrix, accuracy and consistency index Kappa, fault identification coverage (the proportion of correctly identified faults to the actual number of faults), and the rate of false alarms of normal components (the proportion of components that are identified as faults but actually not).

The results indicate that RF, KNN and GB present the best prediction precision. These models show 100% faults coverage for dataset whose abnormal data is removed directly. For the raw dataset and the preprocessed dataset, RF and GB also have 100% faults identification rate but the false negative rate fluctuates around 1.5%. That is, when we do not process abnormal data, RF and GB are preferable methods whereas the KNN would be better if the anomaly is processed. Generally, the best model for fault electrical components identification is the Gradient Boost method, followed by the Random Forest. Other models will have different applicability under various circumstances, especially the preprocessing.

	Raw dataset			Dataset without anomaly			Preprocessed dataset		
	Accuracy	Kappa	Missed	Accuracy	Kappa	Missed	Accuracy	Kappa	Missed
DT	96.59%	94.87%	2	98.98%	98.47%	0	84.62%	76.93%	4
RF	99.38%	99.07%	0	100.00%	100.00%	0	95.80%	93.71%	0
KNN	98.45%	97.67%	2	100.00%	100.00%	0	99.65%	99.48%	0
GB	99.69%	99.53%	0	100.00%	100.00%	0	99.65%	99.48%	0

6. CONCLUSION

The detection of abnormal electrical components has attracted numerous research interests in recent years [1-5]. In this research, we propose a detection scheme for faulty electrical components through multi-step preprocessing and ensemble learning modeling. Preprocessing a large number of zero, negative and abnormal values is the key to ensuring effective detection. Anomaly diagnosis, as the first step in troubleshooting, significantly improves the efficiency of faulty component identification and helps to distinguish the leading indicators. The ensemble learning methods present the highest prediction precision among all competitive models.

REFERENCES

[1] Koushanfar, F., & Potkonjak, M. (2005). Markov chain-based models for missing and faulty data in MICA2 sensor motes. In IEEE Sensors, 2005. (pp. 576–579).

[2] Ge, W., & Fang, C.-Z. (1988). Detection of faulty components via robust observation. International Journal of Control, 47(2), 581–599.

[3] Ardini, J. L., & Allison, R. J. (1988). Apparatus for physically locating faulty electrical components.

[4] Xinru, X., Xueliang, S., Shuliang, F., Jinliang, Z., Ke, L., & Jiang, S. (2016). Device for detecting electrical components with pins and detection control method.

[5] Tsuboi, Y., Hada, J., & Morimoto, M. (1998). Component detection method.