Data Cleaning Based on Multi-sensor Spatiotemporal Correlation

Baozhu Shao[1], Chunhe Song[2,3(✉)], Zhongfeng Wang[2,3], Zhexi Li[4], Shimao Yu[2,3], and Peng Zeng[2,3]

[1] Liaoning Electric Power Research Institute, State Grid Liaoning Electric Power Co., Ltd., Shenyang 110000, People's Republic of China
[2] Key Laboratory of Networked Control Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, People's Republic of China
songchunhe@sia.cn
[3] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China
[4] Shenyang Power Supply Company, State Grid Liaoning Electric Power Co., Ltd., Shenyang 110000, People's Republic of China

Abstract. Sensor-based condition monitoring systems are becoming an important part of modern industry. However, the data collected from sensor nodes are usually unreliable and inaccurate. It is very critical to clean the sensor data before using them to detect actual events occurred in the physical world. Popular data cleaning methods, such as moving average and stacked denoise autoencoder, cannot meet the requirements of accuracy, energy efficiency or computation limitation in many sensor related applications. In this paper, we propose a data cleaning method based on multi-sensor spatiotemporal correlation. Specifically, we find out and repair the abnormal data according to the correlation of sensor data in adjacent time and adjacent space. Real data based simulation shows the effectiveness of our proposed method.

Keywords: Data cleaning · Spatiotemporal correlation · Sensor networks

1 Introduction

The development of modern network technology, especially the development of the Internet of Things (IoT), has made tremendous progress in industrial modernization. In particular, the sensor-based device condition monitoring system is becoming an important part of modern industry. In this kind of application, real time data mining of sensor data to promptly make intelligent decisions is essential [1–3]. However, the data collected from sensor nodes are usually unreliable and inaccurate due to the complex environments, hardware limitations, wireless interferences, etc., which further influence quality of raw data and aggregated results. Thus, it is extremely important to ensure the reliability and accuracy of sensor data before the decision-making process.

Many data cleaning approaches have been proposed, such as supervised neural network methods [5], unsupervised LOF algorithms [6–8], clustering algorithms [9, 10]

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 235–243, 2019.
https://doi.org/10.1007/978-3-030-32388-2_20

and moving average [11]. Supervised algorithms usually require high computation capability and large storage space, which are not suitable for data cleaning of low-storage and low-power sensors. The unsupervised algorithms are designed to find a continuous period of abnormal data, which is not suitable to find the isolated abnormal point in the time series. Some time series based abnormal data detection algorithms also have high time and space complexity, which are not suitable for sensor network applications.

In many sensor-based applications, sensors are densely deployed and the sampling frequency of each sensor is high. The data of individual sensor usually have high temporal correlation, and the data of closed sensors usually have high spatial correlation. In this paper, we propose a multi-sensor based data cleaning approach based on the spatiotemporal correlation between sensor data, which can find and repair abnormal data efficiently. The time and space complexity of the proposed method are very low, so that the abnormal data can be found and repaired efficiently.

The remainder of this paper is organized as follows. In Sect. 2, we present the models and assumptions of this work. We analysis the problem and propose our algorithm in Sect. 3. Section 4 shows experimental results and analysis. Section 5 concludes this paper.

2 Models and Assumptions

We assume there are n sensors densely deployed in a surveillance region, each sensor reports the data in a small time-slot cycle and all sensors are time synchronous. In this case, the data reported from all sensors have temporal correlation and spatial correlation [12].

Let $x_{i,t}$ be a report from the sensor i at time t. Spatial correlation means that the data series of two sensors have similar trends if the two sensors are geographically closed to each other. For example, let $\{x_{1,t}, \ldots, x_{m,t}\}$ and $\{x_{1,s}, \ldots, x_{m,s}\}$ be normal data sets of m adjacent sensors at time t and s respectively. If the m adjacent sensors are closed enough, then there exists a parameter L and a small threshold σ, such that

$$L - \sigma \leq \{|x_{i,t} - x_{i,s}|\}_{1 \leq i \leq m} \leq L + \sigma. \tag{1}$$

Time correlation means that the data in a short period are usually similar with each other. For example, let $\{x_{m,t+1}, x_{m,t+2}, \ldots, x_{m,t+\Delta t}\}$ be a normal data set of the m-th sensor in Δt time slots. If Δt is small, then there exists a small threshold δ, such that

$$\max_{1 \leq i,j \leq \Delta t} \{|x_{m,i} - x_{m,j}|\} \leq \delta. \tag{2}$$

Sensors can produce abnormal data when working in unideal conditions. For example, high volatility, characterized by a sudden rise of variance in the data, can be caused by hardware failure or a weakening in battery supply. Single spikes, occasional unusually high or low readings occurred in a series of otherwise normal reading, can be

caused by battery failure. Intense single spikes that occur with high frequency may indicate hardware malfunction [15].

Since the abnormal data is mainly generated by each sensor node itself, the abnormal data generated by different sensors have no correlation, thus we can detect the abnormal data according to the spatiotemporal correlation of the multiple sensors.

3 Problem Analysis and Algorithm Design

In this section, we discuss how to detect the abnormal data by spatiotemporal correlation and how to repair the abnormal data.

Let $X = \{X_1, X_2, \ldots, X_n\}$ be the collected data from n sensors in T time-slots, where $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\}^{\mathrm{T}}$ is the data set collected by sensor i and $\{\cdot\}^{\mathrm{T}}$ means the transpose of $\{\cdot\}$. The sequence of sensors is sorted according to the position of sensors, such that sensors with close serial numbers are also close to each other. Then temporal correlation refers to the relationship of the data in the column, while spatial correlation refers to the relationship of the data in the line.

Let

$$\Delta X = \{\Delta x_1, \Delta x_2, \ldots, \Delta x_{n-1}\},$$

where $\Delta x_i = \{\Delta x_{i1}, \Delta x_{i2}, \ldots, \Delta x_{iT}\}^{\mathrm{T}}$ and $\Delta x_{it} = |x_{i,t} - x_{i+1,t}|$. According to formula (1), if both $x_{i,t}$ and $x_{i+1,t}$ are normal data, then there exists a parameter L and a small threshold σ, such that $L - \sigma \le \Delta x_{it} \le L + \sigma$. Therefore, if Δx_{it} is not in the region $[L - \sigma, L + \sigma]$, then $x_{i,t}$ or $x_{i+1,t}$ may be abnormal data.

The parameter L is critical in this process. A general method to get L is to let it be the mean of Δx_i. However, when there are abnormal data, the average of Δx_i can be far away from the real gap between X_i and X_{i+1}. Then normal Δx_{it} may be not in $[L - \sigma, L + \sigma]$. To avoid this problem, we use the median of Δx_i instead of mean, since the abnormal data may cause large changes to the mean while have no impact to the median.

If $|x_{i,t} - x_{i+1,t}|$ is not in the region $[L - \sigma, L + \sigma]$, we need to determine which one of $x_{i,t}$ and $x_{i+1,t}$ is abnormal. According to formula (2), for adjacent timeslot t' of t, if $|x_{i,t} - x_{i,t'}|$ is bigger than the small threshold δ, $x_{i,t}$ may be abnormal data. However, if $x_{i,t'}$ is also abnormal, it will cause a false positive. To address this problem, we take a set of data with adjacent timeslots of t, say $x_{i,\Delta t}$, and compute the median $\overline{x_{i,\Delta t}}$ of $x_{i,\Delta t}$. Then if $|x_{i,t} - \overline{x_{i,\Delta t}}| > \delta$, we say $x_{i,t}$ is abnormal. When the abnormal data is detected, we can require the abnormal data by replacing the abnormal data with $\overline{x_{i,\Delta t}}$.

The pseudo code for the proposed method is shown in Algorithm 1. Firstly, the data set X of N sensors in T time slots are divided into several groups, where each group contains data set of N sensors in m time. Then the entire data set is divided into $num = \lceil T/m \rceil$ groups, and each group is a $m \times N$ matrix. For each $m \times N$ matrix, we calculate the differences between all adjacent columns to get the difference $m \times (N - 1)$ matrix D_k, and each D_k is a difference matrix of the X_k matrix. Secondly, for each column of D_k, we find its median L_{kj} and specify a threshold σ. For each matrix D_k, if

$|D_k[i][j] - L_{kj}| < \sigma$, the difference of $X_k[i,j]$ and $X_k[i,j+1]$ are normal, let $A_{ij} = 1$; if not, one of $X_k[i,j]$ and $X_k[i,j+1]$ are abnormal, let $A_{ij} = 0$. Thirdly, for the value of $A_{ij} = 1$, let $M_{ij} = \{X_k[i,j], X_k[i,j+1], \ldots, X_k[i,j+\lambda]\}$, and $m_{ij} = median(M_{ij})$. Compare $|X_k[i][j] - m_{ij}|$ with $|X_k[i+1][j] - m_{i+1,j}|$, if the former is greater than the latter, $X_k[i][j]$ is the abnormal value; otherwise, $X_k[i][j+1]$ is the abnormal value. Finally, for the data $X_k[i][j]$ judged to be abnormal values, the corresponding median m_{ij} is assigned to the value, and the repair is completed.

Algorithm 1 : Data cleaning algorithm

Input: data set $X = \{X_1, X_2, ..., X_n\}$, $X_i = \{x_{i1}, x_{i2}, ..., x_{iT}\}^T$;Time series length of the matrix m; Threshold σ ; The length of the time series used to calculate the median λ

S1: Building matrix D ,let $D[i][j] = X[i][j+1] - X[i][j]$

S2: Let $num = \lceil T/m \rceil$, divide the matrix D into num small matrices $D_k, k = 1, 2, ..., num$. At the same time, the X matrix is re-divided into num matrices, and each D_k matrix is a difference matrix of the X_k matrix.

S3: **for** each matrix D_k , calculate the median L_{kj} of each column

if $|D_k[i][j] - L_{kj}| < \sigma$,**return** $A[i][j] = 1$

S4: **for** each matrix D_k ,

if $A[i][j] = 1$

let $M_{ij} = \{X_k[i,j], X_k[i,j+1], ..., X_k[i,j+\lambda]\}$, $m_{ij} = median(M_{ij})$

Compare $|X_k[i][j] - m_{ij}|$ with $|X_k[i+1][j] - m_{i+1,j}|$, if the former is greater than the latter, $X_k[i][j]$ is the abnormal value; otherwise, $X_k[i][j+1]$ is the abnormal value.

S5: For all the data $X_k[i][j]$ judged to be abnormal values, the corresponding median m_{ij} is assigned to the value, and the repair is completed.

We can see that the entire algorithm does not involve any complicated calculations, just some addition and subtraction of a matrix, so the time complexity of the algorithm is $O(n)$.

4 Experiment Analysis

We use the sensor data from Intel Labs to conduct experiments [14]. The data set contains temperature data collected by 53 sensors deployed at the Intel Berkeley Research Laboratory from February 28 to April 5, 2004. The sensor distribution is shown in Fig. 1. The sensor records data twice per second and collects time-stamped topology information every 31 s.

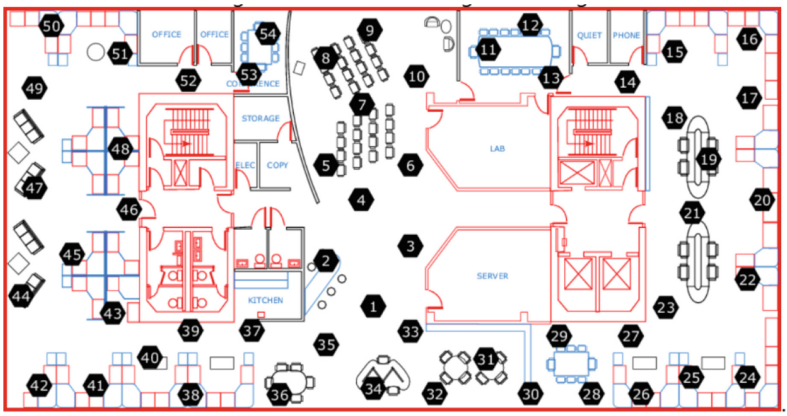


Fig. 1. The location of 54 sensors in the laboratory [14].

As shown in Fig. 2(a), from the normal data of 53 sensors in one day, we can see that they have similar trends. We take the sensor serial number as the x-axis and the temperature value as y-axis, and randomly take the data of 50 adjacent moments. As shown in Fig. 2(b), we can see that they have extremely similar trends. This shows that our algorithm based on spatiotemporal correlation is applicable. In the experiment, we let $m = 15$, $\sigma = 0.1$.

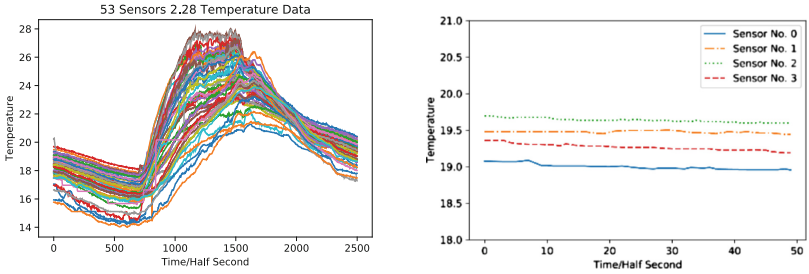
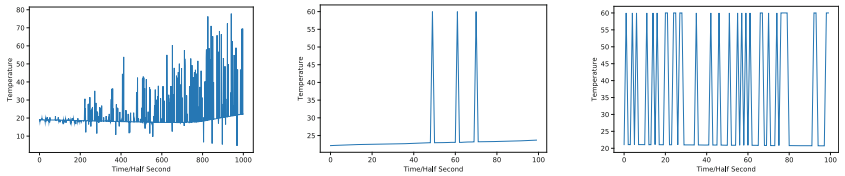


Fig. 2. Raw data of 53 sensors. (a) Spacial correlation (b) Temporal correlation

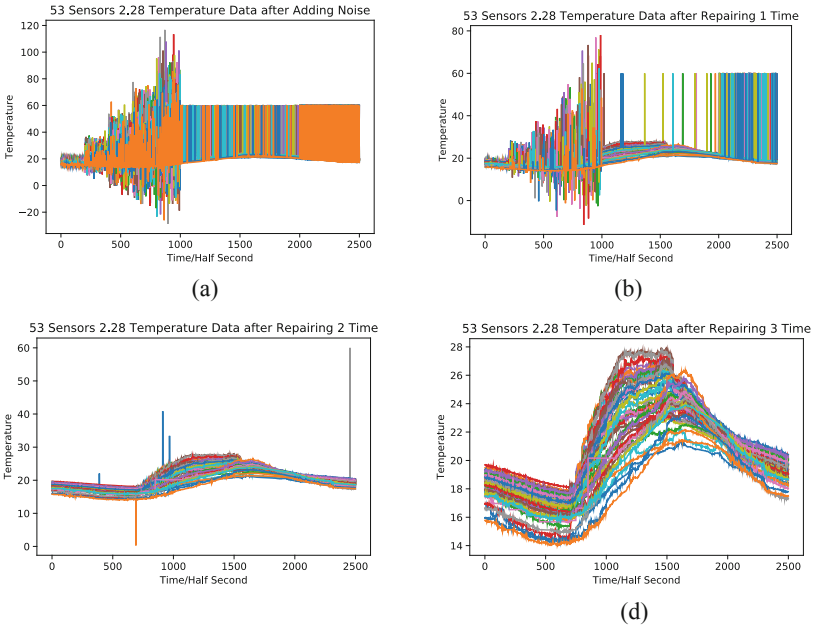
As discussed in Sect. 2, there are usually 3 types of abnormal data, as shown in Fig. 3. We add them to the raw data set as follows: 25% of data in 0–1000 time slots are replaced by the first type of abnormal data; 1% of data in 1000–2000 time slots are replaced by the second type of abnormal data; 20% of data in 2000–2500 time slots are replaced by the third type of abnormal data. The data set with abnormal data is shown in Fig. 4(a). After the first round repair, the result is shown in Fig. 4(b). After the

second round repair, the result is shown in Fig. 4(c). After the third round repair, the result is shown in Fig. 4(d). We can see that after three rounds repair, most of the abnormal data are repaired.



(a) (b) (c)

Fig. 3. 3 abnormal data types. (a) High volatility; (b) Single spikes; (c) Intense single spikes.



(a) (b) (d)

Fig. 4. The sensor data set with abnormal data and the three round repaired data. (a) The data set of 53 sensors after adding abnormal data; (b) The data set after the first round repair; (c) The data set after the second round repair; (d) The data set after the third round repair.

Figure 5 shows the raw data of the No. 1 sensor and the data after three repairs. We can see the repair effect of the algorithm more clearly.

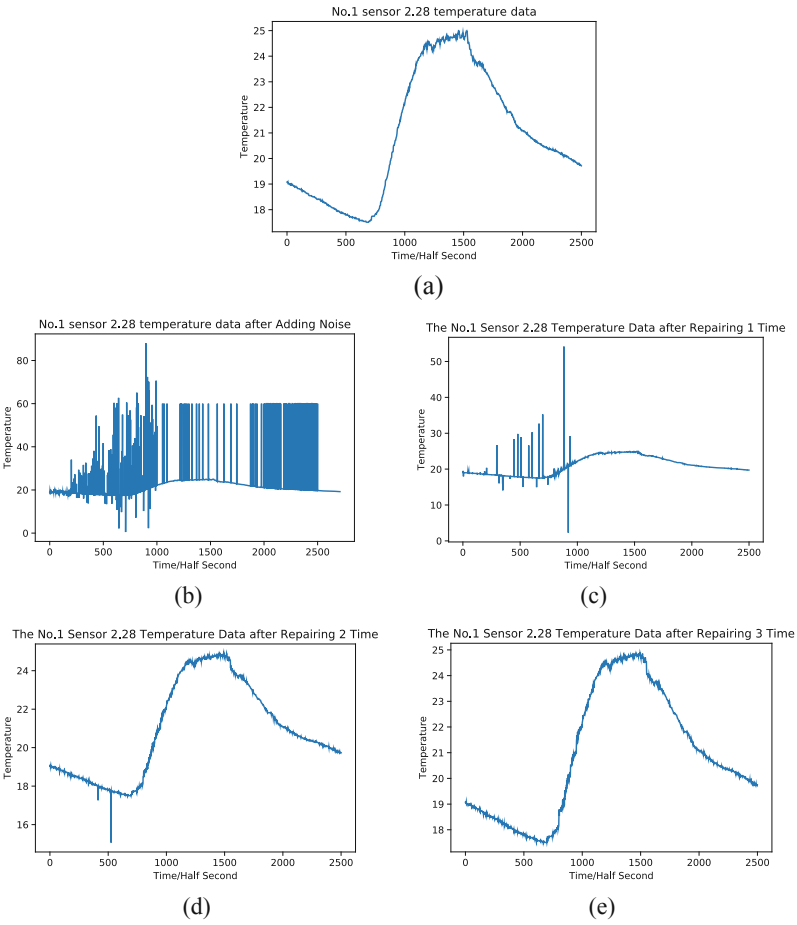


Fig. 5. The raw data of the No. 1 sensor and the data after three round repairs. (a) The raw sensor data; (b) Sensor data with abnormal data; (c) The sensor data after the first round repair; (d) The sensor data after the second round repair; (e) The sensor data after the third round repair.

We use the mean absolute error (MAE) to measure the accuracy of the algorithm. The smaller the MAE, the higher accuracy of the algorithm.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}(\bar{x}_t - x_t), \tag{2}$$

where x_t is the raw dataset data and $\bar{x}_t$ is the repaired dataset data.

We compare our method (MSC) with moving smoothing [16] and stacked denoise autoencoder (SDAE) [17]. The repair results of the three kinds of noise are shown in Fig. 6. we can see that our method always has the smallest MAE.

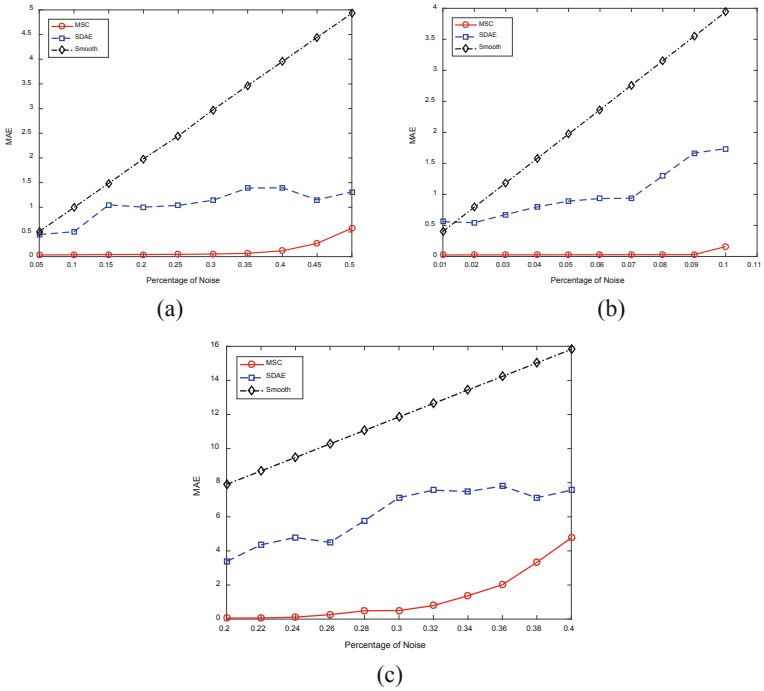


Fig. 6. (a) MAE with the first type of abnormal data; (b) MAE with the second type of abnormal data; (c) MAE with the third type of abnormal data.

5 Conclusion

This paper proposes a multi-sensor abnormal data detection method based on temporal and spatial correlation, which can indicate and repair the abnormal data according to the correlation of sensor data in adjacent time and adjacent space. We then conduct experiments with sensor data from Intel Labs to verify the effectiveness of our approach.

Acknowledgments. This work was supported by the State Grid Corporation Science and Technology Project (Contract No.: SG2NK00DWJS1800123).

References

1. Pan, S.K., Jiang, J.A., Chen, C.P.: Conductor temperature estimation using the hadoop mapreduce framework for smart grid applications. In: 2014 IEEE International Conference of High Performance Computing and Communications (2014)
2. Ganyun, L.V., Haozhong, C., Haibao, Z., Lixin, D.: Fault diagnosis of power transformer based on multi-layer SVM classifier. Electr. Power Syst. Res. **74**(1), 1–7 (2005)

3. Shi, W., Zhu, Y., Zhang, J., et al.: Improving power grid monitoring data quality: an efficient machine learning framework for missing data prediction. In: IEEE Computer Society, pp. 417–422 (2015)
4. Wang, Q., Kundur, D., Yuan, H., Liu, Y., Lu, J., Ma, Z.: Noise suppression of corona current measurement from HVdc transmission lines. IEEE Trans. Instrum. Meas. **65**(2), 264–275 (2016)
5. M. Yang and J. Ma.: Data completing of missing wind power data based on adaptive BP neural network. In:Proc. PMAPS pp. 1–6. Oct, Location (2016)
6. Breunig, M.M., Kriegel, H.P., Ng, R.T.: LOF: identifying density-based local outliers. In: ACM Sigmod International Conference on Management of Data, pp. 93–104 (2000)
7. Kriegel, H.P., Schubert, E., Zimek, A.: LoOP: local outlier probabilities, pp. 1649–1652 (2009)
8. Salehi, M., et al.: Fast memory efficient local outlier detection in data streams. In: IEEE Transactions on Knowledge and Data Engineering, pp. 1–1 (2016)
9. Christopher, T., Divya, M.T.: A comparative analysis of hierarchical and partitioning clustering algorithms for outlier detection in data streams. Int. J. Adv. Res. Comput. Commun. Eng. (2015)
10. Gurav, R.B., Rangdale, S.: Hybrid approach for outlier detection in high dimensional dataset. Int. J. Sci. Res. **3** (2014)
11. Chen, J., Li, W., Lau, A., Cao, J., Wang, K.: Automated load curve data cleansing in power systems. IEEE Trans. Smart Grid **1**(2), 213–221 (2010)
12. He, W., Qiao, P.L., Zhou, Z.J., et al.: A new belief-rule-based method for fault diagnosis of wireless sensor network. IEEE Access. **6**, 9404–9419 (2018)
13. Chen, P.Y., Yang, S., Mccann, J.A.: Distributed real-time anomaly detection in networked industrial sensing systems. IEEE Trans. Ind. Electr. **65**(6), 3832–3842 (2015)
14. Intel Lab Data. http://db.lcs.mit.edu/labdata/labdata.html
15. Ni, K., et al.: Sensor network data fault types. ACM Trans. Sens. Netw. **5**(3), 25:1–25:29 (2009)
16. Jeffery, S.R., Alonso, G., Franklin, M.J., Hong, W., Widom, J.: Declarative support for sensor data cleaning. In: Proceedings of PerCom (2006)
17. Dai, J., Song, H., Sheng, G., et al.: Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders. IEEE Access. **PP**(99):1 (2017)