Detecting anomalies in data center physical infrastructures using statistical approaches

Jianwen Huang, Zhen Chai, and Hailong Zhu*

The National Computer Network Emergency Response Technical Team/Coordination Center of China, 100029 Beijing, China

*Corresponding author e-mail: zhl@cert.org.cn

Abstract. Data center physical infrastructures including electrical system, cooling system, and other secondary systems support the developments of modern information technology. Early detection of the unexpected observations in physical infrastructures is of great significance to prevent the breakdown of the system and further losses. However, the state of the art technique for identifying anomalies in existing infrastructure monitoring platform mainly depends on fixed threshold method. An obvious drawback of the method is that it usually leads to a high misdetection rate. In this study, statistical anomaly detection approach is introduced to physical infrastructure monitoring. First, three important types of anomalies encountered in infrastructure monitoring platform are addressed, namely naïve point anomalies, contextual point anomalies, and level shifts. Then, a method based on Gaussian model is put forward to detect the above three anomalies. Because the proposed method can only effectively detect the naïve point anomalies; an improved approach combining the statistical test results on original and first-differenced monitoring data is provided. Performances of the proposed methods on a real data set are evaluated. Results show that the optimized anomaly detection approach has a good precision and can significantly lower the misdetection rate. In conclusion, this study will not only contribute to the improvement of existing monitoring platform but also benefit the preventive maintenance of data center physical infrastructures.

1. Introduction

A data center is a building used to house networked computers. Components of a data center can be divided into IT infrastructures and physical infrastructures. Function of physical infrastructures is to provide power and suitable environmental conditions for IT infrastructures. In a modern data center, physical infrastructures usually contain electrical system, cooling system, security and life safety system, as well as temperature and humidity sensors [1]. For the convenience of data center operation and management, a data center infrastructure management (DCIM) software that supports real-time monitoring is commonly adopted. To trigger alerts for unexpected observations, fixed thresholds are applied to the metrics being monitored. The main drawback of fixed threshold method is that it can’t directly identify contextual anomalies, collective anomalies, and level shifts [2, 3], which results in a high misdetection rate. The unawareness of some important anomalous observations may even cause the breakdown of physical infrastructures and huge losses. In addition to the main drawback, another disadvantage of thresholding method is that it depends on domain knowledge of electrical engineering
and refrigeration engineering. As a result, only a small part of the metrics being monitored has
thresholds.

To effectively detect the abnormal data points, anomaly detection techniques can be introduced as a
complementary to the thresholding method. Due to the critical role played by anomalies, anomaly
detection has been intensively investigated in many application domains, such as fraud detection [4, 5],
industrial damage detection [6, 7], medical condition monitoring, and network intrusion detection [8-
10]. Although a lot of surveys and reviews have been published on anomaly detection [2, 4, 11-15],
studies on identifying outliers in data center are much fewer. Xu et al. proposed a LSAP method based
on the link state database (LSDB) to detect the abnormal links and routings in data center network [16].
Li et al. developed a simple but effective algorithm to detect the outliers in data center traffic caused
by network attacks [17]. Besides anomaly detection in network, Rodrigo et al. and Kim et al.
investigated the techniques for detecting anomalous in clouds [18, 19]. The metrics used to
characterize cloud included network properties, CPU utilization, memory utilization, and so on. Wang
et al. put forward a novel statistical technique to diagnose anomalies in CPU usage and memory
utilization [20]. In contrast to the seldom studies on anomaly detection in IT infrastructures, there is
few investigations on detecting outliers in physical infrastructures. Lee et al. proposed a model-based
technique to detect the thermal anomaly in data center by comparing the expected and observed
thermal maps [21]. Disadvantages of this approach are that it relies on domain knowledge and it is
relatively complex to obtain the expected thermal maps by using heat generation and extraction
models. From the above literature review, it can be seen that anomaly detection for physical
infrastructures is relatively novel.

In consistent with the widespread applications of anomaly detection, techniques for detecting
anomalies are also diverse. According to the survey conducted by Chandola et al [2], different existing
anomaly detection techniques can be classified into six categories, which are classification based
approaches, nearest neighbor based approaches, clustering based approaches, statistical methods,
information theoretic methods, and spectral methods. Among the above six types of techniques,
statistical method is simple and the corresponding model can be built without labeled training data.
Moreover, statistical anomaly detection approach can be light-weight with a high computational
efficiency [2, 20]. As a result of the attractive advantages, statistical anomaly detection technique has
been widely used in a variety of areas [2, 5, 20, 22].

Besides the intrinsic properties of anomaly detection techniques, choosing a certain approach is
also associated with the characteristics of input data. For monitoring data in this study, the recorded
data points are related to each other and can be regarded as time series. Different types of anomalies in
time-series data have been discussed in literature. Fox proposed two types of outliers, namely additive
outliers (AO) and innovative outliers (IO) [23]. Additive outliers were associated with external factors,
such as measurement errors. Innovative outliers were related to endogenous changes, such as impulse
effects caused by internal factors. Based on the study of Fox, Tsay introduced another two types of
anomalies, namely level shift outliers (LS) and temporary change outliers (TC) [24]. More importantly,
Tasy put forward a unified equation to describe the above four types of outliers. Form a different point
of view, Chandola et al divided the anomalies into three categories, which include point anomalies,
contextual anomalies, and collective anomalies [2]. Overlaps can be often found among point
anomalies, AO, IO and TC. Due to the complexity and relatively smaller effect, collective anomalies
are not considered in this paper.

In this study, statistical anomaly detection technique is introduced to detect the outliers in
monitoring platform for data center physical infrastructures. Three types of anomalies including naive
point anomalies, contextual point anomalies, and level shifts are taken into account. A novel anomaly
detection approach based on Gaussian model is put forward. By combining the statistical test results
on original monitoring data and first-differenced monitoring data, the novel approach can effectively
detect the aforementioned three categories of anomalies. Experimental study on a real data set
indicates that the proposed method have a good precision and can significantly lower the misdetection
rate. In conclusion, this investigation will benefit the operation and preventive maintenance of
physical infrastructures in data centers.
2. Illustration of the interested types of anomalies

As the categories of anomalies greatly affect the effectiveness of anomaly detection algorithms, it is necessary to provide a brief explanation on the anomalies to be investigated in this paper. Although plenty types of anomalies may occur in the physical infrastructure monitoring platform, the major types of outliers that attract the attentions of data center administrators are shown in Figure 1.

It can be seen that data point at time $t_1$ is significantly lower than the rest observations. Data point at time $t_2$ is obviously lower than its neighbor observations. Therefore, data points at time $t_1$ and time $t_2$ can be considered as outliers. Interestingly, values of data points at time $t_2$ and time $t_4$ are the same. But data point at time $t_4$ can’t be regarded as an anomaly. This indicates that the given context leads to the data point at time $t_2$ to be an outlier. To distinguish the two different point anomalies, outliers at time $t_1$ and time $t_2$ are defined as naïve point outlier and contextual point outlier, respectively. For data center physical infrastructures, naïve point outliers are of great importance because they may indicate the faults of equipment. In Figure 1, data point at time $t_3$ is a change point suggesting the levels of time series before and after time $t_3$ are obviously different. In this study, it is defined as a level shift outlier. In practice, level shift outliers are often associated with the changes in operating modes. In conclusion, the three types of anomalies need to be diagnosed in this paper are naïve point outlier, contextual point outlier, and level shift outlier.

3. Anomaly Detection Algorithms

Among the numerous anomaly detection techniques, statistical method is adopted in this study due to its simplicity, effectiveness, and wide applications [2, 20]. To detect outliers with statistical method, two steps are suggested. First, a statistical model is built based on the input data. Second, a statistical test is used to determine whether the data point is anomalous or not. If the data points fall into the high probability regions of the obtained statistical model, the data points are considered to be normal. Otherwise, they are flagged as anomalies. Because of the central limit theorem, Gaussian distribution is usually considered to be the underlying distribution of a given time series. For example, the multivariate adaptive statistical filtering (MASF) for detecting anomalies in computer performances is relied on Gaussian law [25]. In this section, two methods based on Gaussian model are proposed to detect the outliers presented in Figure 1.

![Figure 1. Illustration of the three types of anomalies needs to be identified in this study. Anomaly at time $t_1$ is defined as naïve point outlier. Anomalies at time $t_2$ and time $t_3$ are defined as contextual point outlier and level shift outlier, respectively.](attachment:image1.png)
3.1. *The simple statistical approach (Approach I)*

Similar to the fixed threshold method, the simple statistical approach presented here also uses thresholds to identify outliers. But the thresholds obtained in statistical approach are based on statistical analysis rather than domain knowledge. The detailed description on the simple statistical approach (also named as Approach I) is shown in Algorithm 1.

**Algorithm 1: Detecting anomalies with Approach I**

**Input:** The time series \( X = x_1, \ldots, x_t, \ldots, x_T \)

**Output:** The set of outliers

1: calculate the mean value \( \mu \) of the input time series
2: calculate the standard deviation \( \sigma \) of the input time series
3: determine the parameter \( k \), and set the normal range to be \([\mu - 3\sigma, \mu + 3\sigma]\)

In this paper, \( k \) equals to 3

4: for all \( x_t \) in \( X \) do
5: if \( x_t \) is within the normal range, then
6: \( x_t \) is a normal data point
7: else
8: \( x_t \) is an outlier
9: end if
10: end for

Figure 2 shows that outliers in a current time series are correctly detected by Approach I, which indicates that Approach I can effectively identify naïve point outliers. If proper thresholds are applied to the current time series in Figure 2, the point outliers can also be diagnosed. To some extent, fixed threshold method has a similar performance to Approach I.

![Figure 2. Examples of a time series, the naïve point outliers are correctly detected by Approach I.](image)

**Figure 3.** Shortcomings of Approach I, level shift outliers and contextual point outliers in the temperature time series can’t be effectively identified.

However, for both Approach I and fixed threshold method, they can’t correctly detect the level shift outliers and contextual point anomalies. The temperature time series shown in Figure 3 is a good example.
3.2. The improved statistical approach (Approach II)

To overcome the drawbacks of Approach I, an improved statistical approach named as Approach II is put forward. Approach II contains two steps. Step one, Approach II use Algorithm 1 to detect naïve point outliers. Step two, first difference of the input time series is calculated, then Algorithm 1 is applied to the differenced series to identify level shift outliers and contextual point anomalies. Considering that a point outlier is identified twice in step two, the latter alarm is omitted. By combining the anomaly detection results obtained in the above two steps, Approach II can effectively detect the aforementioned three types of anomalies. The detailed processes of Approach II is given in Algorithm 2.

Algorithm 2: Detecting anomalies with Approach II

**Input:** The time series \(X = x_1, \ldots, x_t, \ldots, x_T\)

**Output:** The set of outliers

1: run Algorithm 1 to get the set of outliers, the set is denoted by \(U_1\), \(U_1 = x_{k1}, \ldots, x_{k_t}, \ldots, x_{kn}\) is smaller than \(T\)

2: calculate the first difference of the input time series, the differenced series is denoted by \(X'\)

3: apply Algorithm 1 to the differenced series \(X'\), the set of detected outliers is denoted by \(U_2\), \(U_2 = x'_{l1}, \ldots, x'_{lt}, \ldots, x'_{lm}\), \(lm\) is smaller than \(T\)

4: for all \(x'_{lt}\) in \(U_2\) do

5: replace \(x'_{lt}\) with \(x_{lt}\) based on the time index \(lt\)

6: end for

7: for all \(x_{lt}\) in \(U_2\) do

8: if \(lt > 1\) and \(lt - l(t-1) = 1\), then

9: remove \(x_{lt}\) from \(U_2\)

10: end if

11: end for

12: get the union set of set \(U_1\) and \(U_2\), denoted by \(U\). \(U\) is the collection of identified anomalies

Figure 4 proves that Approach II behaviors better than Approach I, the contextual point anomalies and level shift outliers with higher anomalous scores are successfully detected.

![Figure 4](image_url)

**Figure 4.** The undetected outliers in the temperature time series shown in Figure 3 are effectively identified by Approach II.

4. Experimental study

To further evaluate the performances of Approach I and Approach II, a real-world data set is utilized. Anomaly detection results are quantitatively characterized by precision, recall, and F1 score. A brief discussion is also provided based on the experimental results.

4.1. Description of the test data set

The test data set used in this section is made up of 60 time series obtained from the DCIM monitoring platform. Three types of time series are considered, namely voltage time series, current time series, and temperature time series. The real-time data is aggregated to day. Length of time series ranges from 55 to 181 days. Outliers in the time series are labeled by skilled operating engineers.
4.2. Evaluation metrics
Four kinds of results can be obtained in an anomaly detection process. If the data point is an outlier, and it is correctly detected, then the result is denoted by TP (true positive). Otherwise, the result is defined as FN (false negative). If the observation is normal, and it is identified as a normal point, then the result is denoted by TN (true negative). Otherwise, the result is defined as FP (false positive). Precision, recall, and F1 score are commonly used to evaluate performances of classifiers. Since anomaly detection can be also regarded as a special binary classification problem, the above three metrics are used in this study [20].

4.3. Results and discussion
Figure 5 shows the results of normalized confusion matrices. Recall of Approach I is 0.52, whereas that of Approach II is increased to 0.8. A higher recall value indicates that more outliers are detected. Therefore, Approach II have a better performance on detecting different types of anomalies. This result is in agreement with the comparative analysis between Figure 3 and Figure 4. Since types of outliers in a practical system are diverse, there is still a fraction of outliers can’t be identified by Approach II.

Table 1 demonstrates the anomaly detection results of Approach I and Approach II. It can be seen that both methods have a very high precision. Although Approach II presents a slightly lower precision than Approach I, recall of Approach II is dramatically enhanced. F1 score of Approach II is 87.8%, that of Approach I is 68.8%, indicating that Approach II has a better performance than Approach I.

Table 1. Performances of anomaly detection Approach I and Approach II.

| Method   | Precision | Recall | F1 score |
|----------|-----------|--------|----------|
| Approach I | 100%      | 52%    | 68.8%    |
| Approach II | 97%       | 80%    | 87.8%    |

As the fixed threshold method adopted by the existing DCIM monitoring platform is just comparable to Approach I. Approach II proposed in this study will contribute to the improvement of physical infrastructure maintenance.

5. Conclusion
In this study, anomaly detection technique is introduced to enhance the performance of existing monitoring platform. Three important types of outliers encountered in practical monitoring system are taken into account. Two statistical anomaly detection methods are proposed. Experimental study shows that Approach II can effectively detect level shift outliers and contextual point anomalies. As a result, Approach II has an obviously lower misdetection rate and a higher F1 score. In summary, this study can benefit the operation and preventive maintenance of physical infrastructures in data centers.

Acknowledgments
This work was financially supported by CNCERT/CC Science Foundation for Youths (NO. 2018Q02).
References

[1] M. Levy, J.O. Hallstrom, A new approach to data center infrastructure monitoring and management (DCIMM), Computing and Communication Workshop and Conference. (2017) 1-6.

[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Computing Surveys (CSUR). 41 (2009) 1-58.

[3] N. Laptev, S. Amizadeh, J. Flint, generic and scalable framework for automated time-series anomaly detection, In Proceedings of the 21st ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining. (2015) 1939-1947.

[4] M. Ahmed, A.N. Mahmood, M.R. Islam, A survey of anomaly detection techniques in financial domain, Future Generation Computer Systems. 55 (2016) 278-288.

[5] A. Sudjianto, S. Nair, M. Yuan, A. Zhang, D. Kern, F.C. Az, Statistical methods for fighting financial crimes, Technometrics. 52 (2010) 5-19.

[6] J. Rabatel, S. Bringay, P. Poncelet, Anomaly detection in monitoring sensor data for preventive maintenance, Expert Syst Appl. 38 (2011) 7003-7015.

[7] J. Rabatel, S. Bringay, P. Poncelet, Fuzzy anomaly detection in monitoring sensor data, IEEE International Conference on Fuzzy Systems. (2010) 1-8.

[8] P. Jongsuebsak, N. Wattanapongsakorn, C. Charmsripinyo, Network intrusion detection with fuzzy genetic algorithm for unknown attacks, International Conference on Information Networking. (2013) 1-5.

[9] W. Wang, X. Guan, X. Zhang, Processing of massive audit data streams for real-time anomaly intrusion detection, Comput Commun. 31 (2008) 58-72.

[10] J. Lin, E. Keogh, A. Fu, H.V. Herle, Approximations to magic: finding unusual medical time series, IEEE Symposium on Computer-Based Medical Systems. (2005) 329-334.

[11] A.A. Sodemann, M.P. Ross, B.J. Borghetti, A review of anomaly detection in automated surveillance, IEEE Transactions on Systems Man & Cybernetics Part C. 42 (2012) 1257-1272.

[12] M. Ahmed, A. Naser Mahmood, J. Hu, A survey of network anomaly detection techniques, Journal of Network & Computer Applications. 60 (2016) 19-31.

[13] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection for discrete sequences: a survey, IEEE Transactions on Knowledge & Data Engineering. 24 (2012) 823-839.

[14] F. Ricca, P. Tonella, Anomaly detection in web applications: a review of already conducted case studies, European Conference on Software Maintenance and Reengineering. (2005) 385-394.

[15] M. Xie, S. Han, B. Tian, S. Parvin, Anomaly detection in wireless sensor networks: a survey, Journal of Network & Computer Systems. 34 (2011) 1302-1325.

[16] G. Xu, Z. Wang, D. Zang, X. An, Anomaly detection algorithm of data center network based on LSDB, Journal of Computer Research & Development. (2018).

[17] A. Li, L. Gu, K. Xu, Fast anomaly detection for large data centers, Global Telecommunications Conference. (2015) 1-6.

[18] R.N. Calheiros, K. Ramamohanarao, R. Buyya, C. Leckie, S. Versteeg, On the effectiveness of isolation-based anomaly detection in cloud data centers, Concurrency & Computation Practice & Experience. 29 (2017).

[19] J. Kim, H.S. Kim, PBAD: Perception-based anomaly detection system for cloud datacenters, IEEE International Conference on Cloud Computing. (2015) 678-685.

[20] C. Wang, K. Viswanathan, L. Choudhr, V. Talwar, Statistical techniques for online anomaly detection in data centers, IEEE International Symposium on Integrated Network Management. (2011) 385-392.

[21] E.K. Lee, H. Viswanathan, D. Pompili, Model-based thermal anomaly detection in cloud datacenters, IEEE International Conference on Distributed Computing in Sensor Systems. (2013) 191-198.

[22] G. Fernandes, E.H.M. Pena, L.F. Carvalho, J.J.P.C. Rodrigues, Statistical, forecasting and metaheuristic techniques for network anomaly detection, Proceedings of the 30th Annual ACM Symposium on Applied Computing. (2015) 701-707.

[23] A.J. Fox, Outliers in time series, Journal of the Royal Statistical Society. 34 (1972) 350-363.

[24] R.S. Tsay, Outliers, level shifts, and variance changes in time series, Journal of Forecasting. 7 (2010) 1-20.

[25] B.J. P, S.A. W, MASF-multivariate adaptive statistical filtering, Int. CMG Conference. (1995) 1-10.