Trend Analysis of Small Sample Operation and Maintenance Data Based on Function Fitting

Yang Bo¹,a, Wei Jun¹,b, Yan Runzhen¹,c, Li Ce¹,d, Zhang Yintie²,e

¹State Grid Gansu Electric Power Company Information and Communication Company. Lanzhou, China
²Information System Integration Branch of NARI Technology Co., Ltd. Nanjing, China

¹yangbo@gs.sgcc.com.cn, b13893137904@163.com, cyanrz_xt@gs.sgcc.com.cn, dlice_xt@gs.sgcc.com.cn, ezhangyintie@sgepri.sgcc.com.cn

Abstract. After years of informatization construction, the State Grid Corporation's data centers have reached a very large scale in the headquarters and provincial company’s network companies, providing good support for the deployment of various business systems. Compared with the rapid development of the data center infrastructure construction, the data center-related operation and maintenance management work concept is relatively lagging behind, the data center generates a large amount of operation and maintenance data every day, and relevant analysis and processing work of the data was ignored. In response to the overall requirements of the State Grid Corporation to build a ubiquitous power Internet of Things, the operation and maintenance of the computer room has been upgraded to a new level, and the operation and maintenance work will be transformed from the traditional lag problem solving problem to the forward-looking work of predicting the operation and maintenance risk in advance. Through the analysis and processing of operation and maintenance data, this paper establishes the functional relationship between operation and maintenance indicator data and time, so as to predict the events that may cause system downtime in the future.

1. Introduction

1.1. Data Center Operation and Maintenance Data Trend Analysis Status and

Data center construction is at the forefront of information chemical operation lines. Hosts, networks, storage, and security devices are no longer simple stacks, but form a complex system with stable and efficient system performance [10]. It provides a quality infrastructure for the deployment of relevant business systems. Compared with the data center infrastructure that keeps pace with the times, the data center's backward operation and maintenance management technology has seriously delayed the need for various industries to use data centers to accelerate business development [7]. Excessive allocation of resources leads to waste of resources, while, insufficient resource allocation leads to system running and increases the risk of downtime. The imbalance in data center resource allocation delays the innovation and development of enterprises.
The development of contemporary enterprises relies on enterprise data centers to provide information support and provide decision support for business leaders. The current business operations, user profiles, development plans, etc. all depend on the robustness of the data center operations. The criminals cannot only steal data center information through technical means; but also can cause the server to crash, resulting in service interruption. The impact of any simple failure in the data center is catastrophic for enterprises. The huge economic losses are still secondary, the loss of customers caused by system downtime, the failure of decision-making, and the missed development opportunities are all difficult to compensate [1]. Due to the huge losses caused by the current data center failure, all enterprises are paying equal attention to the operation and maintenance of the data center. Enterprises hope to change the current status of operation and maintenance, and will continuously improve the technical level of operation and maintenance personnel. The method of shortening the system failure recovery time will be changed to analyze the system nodes where the failure may occur in advance, and avoid the failure as much as possible. Even if the failure is not predicted in advance, it can provide a corresponding direction for the operation and maintenance personnel to solve the problem [4, 5, 6, and 8]. Therefore, the data trend analysis of data center operation and maintenance will inevitably become an important research direction in the operation and maintenance industry.

1.2. Function Fitting and Machine Learning for Predictive Analysis

Data is the basis for machine learning, and the quality of data will directly affect the prediction results of the system's operational state prediction model [9]. The amount of data that people generally understand generally refers to the number of entries in the data, but it lacks accuracy. The training model is not as much as the amount of data, we also need to consider the resource consumption caused by training [11]. To find a balance between model accuracy and system resource overhead, we introduce the variable $K$, which is the number of sample data divided by the number of sample features. The larger the $K$ value, the more suitable for function fitting. On the contrary, the smaller the $K$ value, the more suitable for model generation using machine learning algorithms. For example, there are now 100 pieces of data, and each piece of data has 2 characteristic values. The fitting function is calculated by the scatter points on the 100 two-dimensional planes, and the amount of data is still sufficient. While, the same amount of data is modeled by machine learning algorithms. Through our prior knowledge, the amount of data is insufficient. If there are 100 pieces of data, each piece of data has 200 features. If the data with such characteristics is fitted with a function, the fitting function must have an over-fitting risk. If machine learning is used to perform predictive model calculation, it will be a very easy problem to solve.

By analyzing the characteristics of the data in data center operation and maintenance, we divide a day into 1440 time nodes, and the training data is 1440 two-dimensional feature vectors. Such data volume is obviously not enough for machine learning. First, the number of the number is only 1,440, and the second feature vector dimension is only two-dimensional. In the same situation, the 1440 points are fitted to a function curve, and the amount of data is very abundant. If such a functional expression can be found, all data points can be represented within the allowable range of deviations. Then we think that the function can represent the development trend of the data. How to select and improve the appropriate fit function model is the focus of the discussion below.

2. Research review

2.1. Common function fitting method

A simple understanding of the fit is to connect a series of points on a plane with a smooth curve. The process of calculating a curve expression is function fitting. Function fitting can select curve fitting function according to the shape of curve and prior knowledge. The over-fitting phenomenon is easy to occur in function fitting, which is caused by the existence of noise data in data sets. In this paper, we use the normal distribution rule to eliminate some possible abnormal data for function fitting. How to get a good fitting function, we need to focus on the following four questions:
1. Model selection based on prior knowledge. Linear model, polynomial model, exponential model or other models are selected according to the distribution characteristics of scattered features.

2. Fitting the data according to one or more selected model types to determine the optimal solution.

3. Evaluate the fitting model with test data.

4. Predict according to the collected data and evaluate the prediction results.

The interpolation method and the least square method [2] are commonly used in function fitting. The interpolation method is based on knowing the function value (exact value) of a series of points \( x_i \) on the interval \([a,b]\) without knowing the analytic formula of function \( f(x) \). The interpolation function \( P(x) \) is constructed to replace \( f(x) \) in order to find the function value of non-interpolation points. To construct the fitting function by interpolation method, it is necessary to pass through the interpolation points.

The least square method is to find the functional relationship \( f(x) \) between independent variable and dependent variable in a series of known experimental data \((x_i,y_i)\) (uncertain value). The fitting curve \( f(x) \) is used to approximate the experimental data. The least square method does not require the function curve to pass through a specific point \((x_i,y_i)\).

The advantage of the least square method is that the error between the fitting curve and the real value is the smallest [3, 12]. The operating parameters of each index are the fluctuating intervals when the system runs stably. There are some errors between the parameters of the system prediction index and the actual values, but the errors should float in a smaller interval. This idea is consistent with the least square method, so the least square method is chosen as the method of calculating the fitting function.

2.2. Value determination of fitting function model

The purpose of data center operation and maintenance is to ensure the stable operation of business system. The usage curves of CPU, memory, IO, etc. and the monthly consumption curves of storage space of a single business system are similar in general. For different application scenarios, the method of extracting fitting curve has similarity. We take the utilization rate of CPU in information system every day as an example to introduce the acquisition of function model value.

In order to accurately calculate the prediction model, the CPU utilization rate was \([x_1, x_2, x_3, \ldots, x_{30}]\) at 9:00 a.m. every day for 30 days. This set of data is basically equal in theory, and the noise caused by the abnormal operation of the system is eliminated. In order to obtain the model value to represent the utilization rate of CPU at 9 o'clock under the condition of stable operation of the system, a daily CPU utilization curve is constructed. We need to find a value that minimizes the sum of the value and the absolute values \(x_1, x_2, x_3, \ldots, x_{30}\).

Assuming that the model value of CPU utilization rate at 9 o'clock is \( x \), the conditions that \( x \) needs to satisfy are described in mathematical language as follows: 
\[
f(x) = \left( \sum_{i=1}^{30} (x - x_i)^2 \right),
\]
When function \( f(x) \) gets the minimum value, \( x \) is the model value CPU utilization at 9 o'clock. For the derivative of 
\[
f'(x) = 2 \sum_{i=1}^{30} (x - x_i) = 0
\]
is solved to obtain the model value at 9 o'clock. In the same way, the model values are obtained every minute of the whole day.

3. Computing Method of Fitting Function

We get the utilization model value of CPU every minute of the day, a total of 1440 discrete points. From zero o'clock, we select a series of scatters from the 1440 scattered points every 30 minutes, and connect the adjacent scatters with a smooth curve to get a daily CPU utilization curve, as shown in
Next we will discuss how to determine functional expressions to represent CPU utilization curves.

![Graph of CPU utilization over time](image)

**Fig. 1 CPU utilization**

### 3.1. Principle of Least Square Method

The least square method is used to determine the fitting function, assuming that the CPU utilization curve function is $f(x) = ax + b$ (a, b is any real number). Assume that $Y_i$ is the actual value and $Y_j$ is the function predicted value. The minimum absolute value of the difference between $Y_i$ and $Y_j$ is the criterion for optimizing the fitting function.

Let $\varphi = \sum (Y_i - Y_j)^2$, substituting $f(x)$ for $\varphi$ to get $\varphi = \sum (Y_i - aX_i - b)^2$. In order to find the minimum value, we can find the partial derivatives of $Y$ with respect to $a$ and $b$ respectively, and make the partial derivatives equal to 0 to obtain the following variance groups:

1. $\sum 2(b + aX_i - Y_i) = 0$  
2. $\sum 2X_i(b + aX_i - Y_i) = 0$

By solving the equations, the following conclusions can be obtained:

1. $b = \frac{\sum Y_i}{n} - a \frac{\sum X_i}{n}$  
2. $a = \frac{n \sum X_iY_i - \sum X_i \sum Y_i}{n \sum X_i^2 - \sum X_i \sum X_i}$

Where, the correlation coefficient $R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$. The closer the correlation coefficient is to 1, the closer the function fitting is to the real value.

### 4. Verification and conclusion of scenario application

According to the curve performance characteristics and the curve alignment characteristics of various fitting functions, we find that the CPU model curve and polynomial function curve alignment are similar. So we choose to use polynomial function to fit CPU model curve. We use quadratic polynomial, cubic polynomial and even hexagonal polynomial to fit scattered points. We found that the higher the number of polynomials, the better the fitting effect of the model. Figure 2 shows the fitting of scatter points.
points by different fitting polynomials. The expressions of fitting functions and related coefficients are listed in the Table I.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Maximum term of function} & \text{Function expression} & \text{Correlation coefficient R}^2 & \text{Solution time} \\
\hline
2 & y = -6E-05x^2 + 0.0907x + 9.3481 & 0.6851 & 0.02s \\
3 & y = -5E-08x^3 + 6E-05x^2 + 0.0236x + 16.98 & 0.7609 & 0.23s \\
4 & y = 2E-10x^4 - 7E-07x^3 + 0.0006x^2 - 0.1485x + 28.218 & 0.9115 & 1.3s \\
5 & y = 3E-13x^5 - 8E-10x^4 + 6E-07x^3 - 6E-05x^2 - 0.0163x + 22.782 & 0.9769 & 1.5s \\
6 & y = -5E-16x^6 + 2E-12x^5 - 4E-09x^4 + 3E-06x^3 - 0.001x^2 + 0.1051x + 19.465 & 0.9811 & 8.43s \\
\hline
\end{array}
\]

Table I shows that when the number of functions is 5, the correlation of fitting function has reached 0.9769. It takes only 1.5 seconds to calculate the fitting function. When the number of functions is increased to 6 times, although the correlation coefficient of the fitting function is higher, the calculation time consumed has reached 8.43s. We think that the resource consumption caused by the increase of such a small correlation coefficient is considerable. In the actual production environment, we can choose to give up the accuracy of function fitting appropriately in order to improve the efficiency of the system. So in this scenario, we choose a polynomial function of 5 times as the CPU utilization curve of the system in one day.

The binary classifier is trained by using the model values and CPU utilization at the time of system failure.

![Fig. 2 CPU Utilization](image)

**Tab. 1 The expression of fitting function and correlation coefficient**
Figure 3 shows that the point inside the two red lines is the state point without causing system failure, and the point outside is the fault point. If the return value is greater than 0, the parameters are brought into the classifier function, and the inner point of the red line is the normal point. Conversely, if the return value is less than or equal to 0, the fault point is on the red line or outside.

In order to simplify the calculation, we select some of the 1440 model-valued nodes to calculate the fitting function. Now we use the nodes which are not used to calculate the fitting function to verify the validity of the fitting function.

### Tab. 2 Value returned from the classifier

| Serial number | Calculated value | Actual value | Deviation | Value returned from the classifier | Interval of calculated values |
|---------------|-----------------|--------------|-----------|-----------------------------------|-------------------------------|
| 1             | 23.5            | 24           | -0.5      | 0.8,0.85                          | [22,25]                       |
| 2             | 32              | 33           | -1        | 1.1,1.2                           | [30,34]                       |
| 3             | 23.8            | 25           | -1.2      | 1.2,1.7                           | [22,26]                       |
| 4             | 40              | 37           | 3         | 3.2,4                             | [37,42]                       |
| 5             | 52.3            | 53           | -0.7      | 2.2,2.23                          | [50,55]                       |
| 6             | 55.9            | 54           | 1.9       | 1.9,1                             | [53,57]                       |
| 7             | 33              | 33           | 0         | 0,8,0.8                           | [31,35]                       |
| 8             | 30.1            | 35           | -4.9      | 2,2,3.2                           | [29,35]                       |

The calculated and actual values of the model are brought into the classifier for calculation, and the return state values are the same. Prove that it returns in the same state. All of them are in normal operation state, and the predicted value reflects the actual operation state of the system. It shows that the CPU utilization rate in this area does not cause system failure.

As shown in Table II, by validating the model values, we find that there is a certain deviation between the actual value and the calculated value. The node deviation is within a reasonable range, which can better reflect the system's operation status at that time. Although all the predicted values are within the error range, we also find that there are some large deviations, which may lead to missed reports of abnormal conditions and misinformation of the normal state of the system. In the follow-up work, we need to improve the fitting function model and classification model, optimize the accuracy of the system, and make the absolute error between the predicted value and the actual test value as small as possible.

### 5. Conclusion

In this paper, we introduce a method of system state prediction by function fitting. Firstly, we calculate the model value of the fitting function by collecting the system running state data in many days. The model value satisfies the absolute error and minimum of the system running value every day. Then we calculate the fitting function of the model value in one day. According to the distribution state of the discrete points, we choose the polynomial function as the fitting function model. The higher the number of polynomial functions, the better the fitting effect of the scatter points. However, the increasing computational complexity of the number of polynomial functions is also very prominent. By verifying the accuracy of the fitting function, we find that the function can effectively represent the stable operation of the system throughout the day within the allowable range of system errors.

In order to predict the running state of the system, we analyze the running data of the first 30 days in real time and dynamically, and establish the corresponding running state curve. If there is a big difference between the real-time state curve and the prediction curve of the system operation on that day, then the system operation and maintenance personnel need to check the corresponding state items and locate the root cause of the abnormal performance of the system in time.
Acknowledgments
This work was financially supported by the State Grid Corporation of Science and Technology (WBS number: 522723180006).

References
[1] The Best Choice of Wavelet Basis Function in Fault Diagnosis [J]. Chen Zexin. Mechanical Science and Technology. 2005 (02)
[2] Research on Surface Fitting Algorithms Based on Least Squares [J]. Li Ertao, Zhang Guoxuan, Zeng Hong. Journal of Hangzhou University of Electronic Science and Technology. 2009 (02)
[3] Simulation of least squares curve fitting based on MATLAB [J]. Chen Lanfeng, Yang Jingyu, Cui Song, Pan Qingchao, Li Liu. Journal of Shenyang Normal University (Natural Science Edition). 2014 (01)
[4] Research on the Construction Strategy of Data Mining Application Platform Based on Cloud Computing [J]. Bi Xu, Chen Fanli. Digital Technology and Application. 2016 (03)
[5] Cloud Computing and Large Data Intelligent Operation and Maintenance [J]. Zhao Zhiyuan, Ji Ying, Zhang Jigang. Network Security and Informatization. 2017 (12)
[6] Exploration and Practice of Intelligent Operation and Maintenance [J]. Liu Shifa, Bi Yongjun. Electronic Finance. 2017 (08)
[7] Data Center Automation Operation and Maintenance Management and Platform Construction [J]. Guo Wei, Chen Xiuqian. Electronic Technology and Software Engineering. 2017 (16)
[8] Scenario Analysis of Operation and Maintenance Automation Application in Bank Data Center [J]. Zheng Youping. China Financial Computer. 2017 (01)
[9] Data analysis should have personality [J]. Huang Kun. Chinese computer users. 2007 (20)
[10] China-Japan High Level Seminar on Standard Development of Data Centers [J]. Standardization of Engineering Construction. 2018 (05)
[11] Characteristic function of uncertain distribution and validity test of fitting. Journal of Harbin University of Technology, 2019 (04). Meng Xiangfei, Wang Ying, Li Chao, Lu Maolong.
[12] LFM signal parameter estimation based on moving least squares curve fitting [J]. Qilin, Zhangfang, Chen Enqing. Journal of Zhengzhou University (Engineering Edition). 2011 (03)