Universe Scientific Publishing

Anomaly Detection and Prediction Based on Holt-Winters Method

Zhuofan Zhong, Huiqi Fan, Sijia Fan, Zihan Wang, Yuchen Wang
Hangzhou Normal University, Hangzhou 310000, China.

Abstract: Outliers, detection and prediction of abnormal periods, and prediction of trends in a given time base on existing data are the first problems to be solved in intelligent operation and maintenance. This paper took KPI performance index of 58 cells covered by 5 base stations from August 28 to September 25, 2021 as research data, chose 3 indexes (average number of users, PDCP traffic and average activation number), established a set of outlier prediction system based on Holt-Winters method.

Keywords: Anomaly Detection; Holt-Winters Method; Boxplot Box Diagram; Anomaly Prophet Prediction.

1. Introduction

This paper took the KPI performance index of 58 cells covered by 5 base stations from August 28 to September 25, 2021 as research data, aimed to solve the following two main problems:

(1) Anomaly detection. The index data provided by the operator base station is used to detect the abnormal values of all the cells in the above three key indicators, the abnormal values include two cases: abnormal isolated points and abnormal periods.

(2) Anomaly prophet prediction. On the basis of anomaly detection, a prediction model is established based on the data before the abnormal value to predict whether the abnormal value will occur in the future.

2. Subject of the Research

2.1 Analysis of Problem One

For problem 1, we first study the concept of outliers, that is, a group of measured values whose deviation from the average value is more than twice the standard deviation, then how to reasonably find out the standard to judge the anomaly is the key step. Then carry on the visual analysis of the three KPI indicators, use python to import the data and draw the image to judge whether the KPI value of the three core indicators has the trend and periodicity. To better preserve the trend and periodicity of the time series, we choose the cubic exponential smoothing method to predict the data of each time point according to the existing data, and then record the error between the observed value and the predicted value as $\in$.[1] Finally, we use the Boxplot method to find out the time point corresponding to the error value above the upper edge and below the lower edge, and we define it as an abnormal isolated point in the original image. Secondly, we define the time period as 24 hours, and our time series is divided according to this standard. If there are two or more abnormal points in a period, we call it an abnormal period.

2.2 Analysis of Problem Two

For problem 2, we need to detect the abnormal value of problem 1, and use the data before the abnormal value to predict whether the abnormal value will occur in the future. We believe that this problem should be based on problem 1, so we establish a set of outlier prediction system based on Holt-Winters based on problem 1. The trend data of the forecasting

system represents the overall change trend of the time series. If we look at the three KPI index data in a period of ten days or dozens of days, there is no obvious trend, so we can remove the trend data. For each data that is tested as an outlier, we calculate a threshold as a criterion based on the data previously marked as normal. According to this standard, we regard the error exceeding this threshold in the data in the next three days as abnormal data.[2]

3. Model Establishment and Solution

3.1 Modeling of Problem One

To calculate the error between the observed value and the predicted value, we need to establish a prediction model based on cubic exponential smoothing algorithm, through which the predicted value of each point is calculated and stored. Because of the large number of cells, we only take the average number of users of 26019030 cells as an example to judge the abnormal outliers.

We first consider the recursive relation of the first smoothing algorithm:[3]

$$S_i = \alpha x_i + (1-\alpha)S_i - 1$$

Where α is the smoothing parameter, S_i is the smoothing value of the first i data. The closer α is to 1, the closer the smoothed value is to the data value of the current time, the less smooth the data is; the closer 9 is to 0, the closer the smoothed value is to the smooth value of the previous I data, the smoother the data is.

The prediction formula of the first smoothing algorithm is $X_{i+h} = S_i$. Therefore, the predicted time series is a straight line, which can not reflect the trend and periodicity of the time series.

On this basis, a new variable T is added to represent the trend after smoothing, that is, quadratic exponential smoothing:

$$S_i = \alpha x_i + (1+\alpha)S_{i-1} + T_{i-1}$$
$$\overline{I}_i = \beta(S_i - S_{i-1}) + (1-\beta)T_{i-1}$$

Where α is the horizontal smoothing parameter, β is the trend smoothing parameter, S_i is the smoothing value of the first I data. So we can get the prediction formula of quadratic exponential smoothing is $X_{i+h} = S_i + hT_i$, and the prediction result is an oblique straight line.

But this data set has a strong periodicity. So a new parameter P needs to be added to represent the trend after smoothing. In the cubic exponential smoothing, we use the accumulation method to derive it and get the following recursive formula:

$$S_i = \alpha(x_i - P_{i-k}) + (1-\alpha)(S_{i-1} + T_{i-1})$$
$$T_i = \beta(S_i - S_{i-1}) + (1-\beta)T_{i-1}$$
$$P_i = \gamma(x_i - S_i) + (1-\gamma)P_{i-k}$$

Where k is the period and γ is the seasonal smoothing parameter.

Based on the above derivation, we obtain the prediction formula of cumulative cubic exponential smoothing as follows:

$$X_{i+h} = S_i + hT + P_i + k + (h \bmod k)$$

After establishing the prediction model based on cubic exponential smoothing, we record $Y = |\, y_i - y_i^*|$, in which y_i and y_i^* represent the real value and the predicted value respectively. We summarize the y value of the average number of users of the cell into the box chart, and the time points that exceed the upper and lower limits are recorded as outliers.

After getting the outliers, we determine that the standard of the time period is 24 hours, because after continuous division and testing, we find that the 24-hour period can ensure the continuity of the abnormal period. We stipulate that if there are two or more abnormal outliers in each time period, the period is called an abnormal period.

3.2 Solution of Problem-model

To calculate the error between the observed value and the predicted value, we need to establish a prediction model

based on cubic exponential smoothing algorithm, through which the predicted value of each point is calculated and stored. Because of the large number of cells, we only take the average number of users of 26019007 cells as an example to judge the abnormal outliers.

We import the data through Python, and after debugging, we determine the parameters of Holt-Winters, which are $\alpha = 0.8, \beta = 0.2, \gamma = 0.5, S_0 = x_0, T_0 = x_1 - x_0, P = 0$ mentioned in the previous section.[4] Because of the large number of cells, only the anomaly detection results corresponding to the average number of users in 26019030 cells are put into the paper:

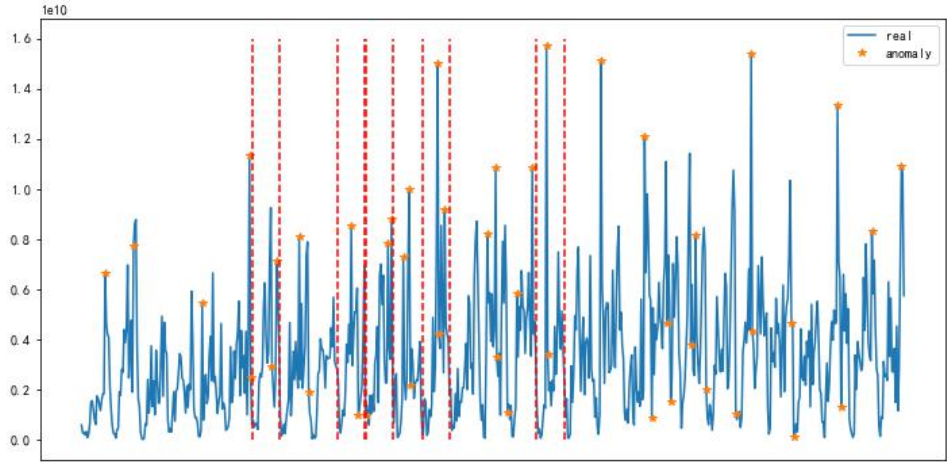

Figure 1: The Average Number of Users in 26019030 Cells

As shown in Figure 1, the position of the orange point in the figure is the abnormal isolated point. After the abnormal isolated point is drawn, we divide the 29-day time series according to 24 hours, and each day is a time period. We can see that there are two or more abnormal isolated points in the time period between the two red dotted lines in the figure, which is called the abnormal period. According to the above similar method, we summarize the three core KPI of 58 cells, and get the following results:

Table 1: Summary Table of Abnormal Data

	Time Period Selection Criteria	The number of Abnormal Outliers	Number of Abnormal periods
Average Number of Users in the Cell	24h	2610	542
Cell PDCP Traffic	24h	2203	430
Average Number of Active Users	24h	2907	639

3.3 Modeling of Problem Two

The trend data of the prediction system represents the overall change trend of time series. If we view the three KPI indicator data in ten days or dozens of days, there is no obvious trend. Therefore, we can remove the trend data part, and the iterative formula of each part can be simplified as:

$$a[t] = a(Y(t) - s[t-k]) + (1-a)a[t-1]$$
$$s[t] = \gamma(Y(t) - a[t]) + (1-\gamma)s[t-k]$$

The formula for future prediction based on existing data is as follows:

$$Y^*[t+h] = a[t] + s[t-k+1+(h-1)\,mod\,k]$$

Where, it represents the span between the future prediction time and the current time. Obviously, the choice will affect the accuracy of prediction, and the larger the prediction, the larger the error will be.

Holt-winters needs to input a large number of data sequences for the calculation of periodic data $S(t)$ for each prediction. $A(t)$ and $S(t)$ are interrelated, which is easy to see from the perspective of periodicity. To calculate $a[t]$ in the current period, $s[t-k]$ calculated in the previous period is used, and $S(t)$ in the current period will also be used in the calculation of $a(t+k)$ in the next period.[5]

Therefore, the Holt-Winters forecast data is mainly divided into two parts: 1. The cycle number of time series is calculated at a regular time, $s(t)$; 2. Real-time prediction of residuals.

The period number of time series does not need to be calculated in real time. Here, we choose a day and update $S(t)$ once every 24 hours. The method can be calculated using the simplified formula above. In order to ensure the accuracy of prediction, it is necessary to detect whether outliers appear in the 24 hours before each outlier. If so, it is necessary to delete outliers so as not to use these outliers in the calculation of periodic data and thus affect the prediction of the next cycle.

After the number of cycles is calculated, the difference between the actual data and the periodic data is calculated to the residual data, and the residual data is then carried out a moving average to predict the residual at the next moment. Then the residual of the next moment is added to the periodic data to obtain the predicted value of the moment.

3.4 Solution of Problem-model

For each data that is tested as an outlier, we calculate a threshold value based on the previously marked normal data as the evaluation standard, and the threshold value is calculated as follows:

For each outlier point, we take the error sequence of 24 hours before the outlier point to form an error window, and divide the error window into two sequences missing only the mansa data and the first data e_1,e_2 of the window. Calculation threshold ϵ :

$$\epsilon = \frac{|\bar{e_1} - \bar{e_2}|}{Var(e_1)}$$

Then, the residual error of the next time to be predicted is calculated according to the existing normal data, and the error window is moved to the right for one hour. The new error ϵ' is calculated according to the above method, and the size of the two is compared. If $\epsilon' > \epsilon$, the anomaly occurs at this time.

Summary

This paper takes the KPI performance index of 58 cells covered by 5 base stations from August 28 to September 25, 2021 as the research data, selects three core indicators (average number of users, cell PDCP traffic, average activation number) for analysis, and establishes a set of outlier detection and prediction system based on Holt-Winters method, which clarifies the criteria for judging outliers and abnormal periods. It lays a solid foundation for the establishment of Prophet forecasting model to forecast the overall trend in the later stage.

References

[1] "Natural Gas load forecasting based on Holt-Winters Model", Technology and Market, No. 07, 2020, Hu Kai 23.

[2] "Analysis and Application of time Series Prediction based on Holt-Winters in big data Monitoring." computer and Modernization, No.11, 2019, Wang Yufei, du Tiancang, p. 5.

[3] Prediction of hospitalization based on Holt-winters exponential smoothing method. Chinese Medical Records, No. 09, 2018, Zhou Mengjun, pp. 52-53.

[4] BlackEyes_SGC, outlier detection algorithm for data https://blog.csdn.net/u011204487/article/details/105731065, October 30th, 2021.