ISSN 2616-5775 Vol. 5, Issue 4: 46-51, DOI: 10.25236/AJCIS.2022.050408

Anomaly Detection and Trend Prediction in Intelligent Operations Based on Prophet and S-ESD

Yihong Zhang[1, a], Zenan Ji[2]

[1]School of Science, North China Institute of Science & Technology, Langfang, 065201, China
[2]School of Data Science, Zhejiang University of Finance & Economics, Hangzhou, 310018, China
[a]375681870@qq.com

Abstract: *Anomaly detection and trend prediction based on KPI business indicators play a key role in intelligent operation and maintenance in various fields. The types of indicators analyzed vary greatly for different scenarios of operation and maintenance, but all have time-series characteristics. In this paper, three core indicators are selected based on time series to build anomaly detection and trend prediction models. After preprocessing the index data, a statistical-based S-ESD time series anomaly detection method is firstly implemented. Anomaly detection is performed on three core indexes, and the detected anomalies are extracted and corrected. Use the revised KPI indicator data to further build a Prophet trend prediction model based on decomposable (trend + season + holiday), select MAPE as the evaluation index, perform model training on the indicator data, predict the next three days by step size, and get the fitting Curve and trend for the next three days. After parameter tuning, the MAPE of the Prophet prediction model is less than 0.1, the running time is shorter, and the trend prediction accuracy is higher, which can be practically applied to the field of intelligent operation and maintenance.*

Keywords: *Intelligent O&M; S-ESD Anomaly Detection; Prophet; Grid Search*

1. Introduction

Anomaly detection and trend prediction are the first problems to be solved in intelligent operation and maintenance. This kind of problem is analyzed and diagnosed through KPI business indicators directly related to business, system, and product. The indicators mainly include key performance indicators such as service performance and server hardware health status. Design and application-related effective anomaly detection and prediction models can make operation and maintenance. Personnel can more accurately grasp network performance, reduce false positives and abnormal false positives, and improve the efficiency and quality of network optimization.

At present, there are roughly three types of anomaly detection algorithms based on time series:1) Based on unsupervised clustering methods, Wu Rui, Zhang Anqin and others proposed an improved K-means anomaly detection algorithm [1], Zhong Cheng, Luo Cheng et al. The kernel function is combined with the RA clustering detection algorithm [2]. This method avoids the dependence on the training data set, but is sensitive to the parameter setting. If the setting is unreasonable, the clustering result will be unsatisfactory. 2) Based on the supervised classification method, Xu Qinzhen, Yang Luxi and others adopted the local decision hierarchical support vector machine anomaly detection method [3], and obtained relatively stable results, but in many applications, normal samples are easy to obtain, while abnormal samples are easy to obtain. The collection cost is high, and there are still difficulties in practical application.3) Anomaly detection method based on statistics. Cui Weilan, Yin Xunwei and others proposed an anomaly detection method based on adaptive AR model [4]. This method has fast convergence, high accuracy, and can reduce the false alarm rate, so that anomaly detection can be achieved. more acurrate. n this paper, an S-ESD detection algorithm open sourced by Twitter is also used to detect anomalies in KPI indicators [5]. In the trend prediction of KPI performance index data, the time series prediction model based on deep learning is more popular at present. Wang Xin, Wu Ji and others adopted the fault time series prediction method based on LSTM recurrent neural network [6], and obtained high accuracy., but due to the weak interpretability of neural network time series prediction and the large framework, this paper adopts a Facebook open source Prophet prediction model to predict the trend of KPI performance indicators [7-8], and uses the operator base station KPI Empirical analysis of the indicator dataset.

ISSN 2616-5775 Vol. 5, Issue 4: 46-51, DOI: 10.25236/AJCIS.2022.050408

2. Data processing

The intelligent operation and maintenance research data comes from the KPI performance indicators of operators' base stations provided by the Big Data Challenge. The data is from 0:00 on August 28, 2021 to 23:00 on September 25, 29 days, covering 58 cells with 5 base stations The corresponding 67 KPI indicators, the time interval is one hour. According to different base stations, three core indicators are selected for data preprocessing, one of which is the total throughput (bits) of the downlink data sent by the PDCP layer of the cell and the total throughput (bits) of the uplink data received by the PDCP layer of the cell. bits) are constructed by summing the two indicators. Use python to view the lack of data and standardize the data, and use matplotlib to visualize the time series fluctuation diagram and data distribution diagram of the indicator data, so as to train the data for abnormal detection and trend prediction later.

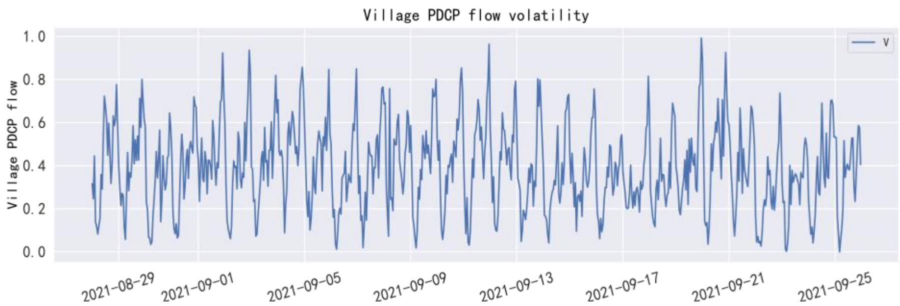

Figure 1: Change of village PDCP flows

3. Methodology

3.1. Core Indicators

The average number of users in a cell represents the average number of users who are online via mobile phones in a cell covered by a base station within a certain period of time. The data set has given specific values.

The PDCP traffic of a cell refers to the sum of the uplink and downlink traffic of cells covered by a base station in a certain period of time.

$$Total\ pdcp\ flows = downstream\ flows + upstream\ flows \tag{1}$$

The average number of active users indicates the average number of people who have registered with the wireless network in a cell covered by a certain base station within a certain period of time.

3.2. Time series anomaly detection based on S-ESD algorithm

3.2.1. S-ESD algorithm

In view of the periodicity and trend of time series data, anomaly detection cannot be treated as an isolated sample point, so Twitter engineers proposed Seasonal ESD and Seasonal Hybrid ESD Algorithm [1], which can extend the ESD algorithm to time series. Firstly, the time series data is decomposed into trend components, periodic components and residual components using STL. Applying ESD to the residual components after STL decomposition can detect abnormal points in time series. However, we will find that there are some spurious anomalies in the residual component.

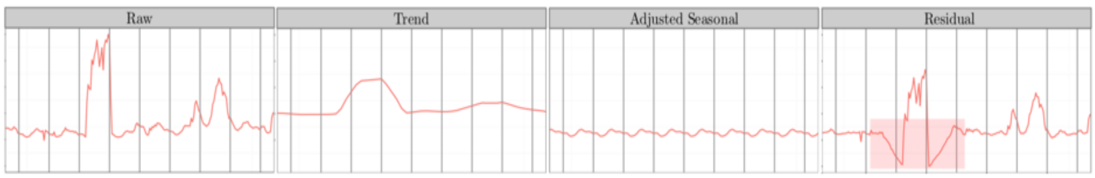

Figure 2: STL with Trend Removal

In the red rectangular box, the downward protrusion is falsely reported as an outlier. In order to solve the problem of reducing the accuracy of false positives, the S-ESD algorithm replaces the trend

ISSN 2616-5775 Vol. 5, Issue 4: 46-51, DOI: 10.25236/AJCIS.2022.050408

component with the median, and the calculation formula of the remainder is as follows:

$$R_x = X - S_x - \tilde{X} \tag{2}$$

Among them, X is the original time series data, S_x is the periodic component after STL decomposition, $\tilde{X}$ is the median of X.

3.2.2. Anomaly detection model based on S-ESD

First, take the logarithm of the preprocessed data of the three indicators and visualize them. From the drawn distribution diagram, it can be seen that the three indicators basically obey the normal distribution, which is in line with the basic assumption of the S-ESD algorithm. For the implementation of the S-ESD anomaly detection algorithm, the pyculiarity of python can be used to detect the anomaly points of the three core indicators respectively. Pyculiarity is a python implementation version of twitter time series data anomaly detection, by checking whether the deviation of the maximum value and the minimum value from the mean is abnormal. The result obtained is shown in Figure 3:

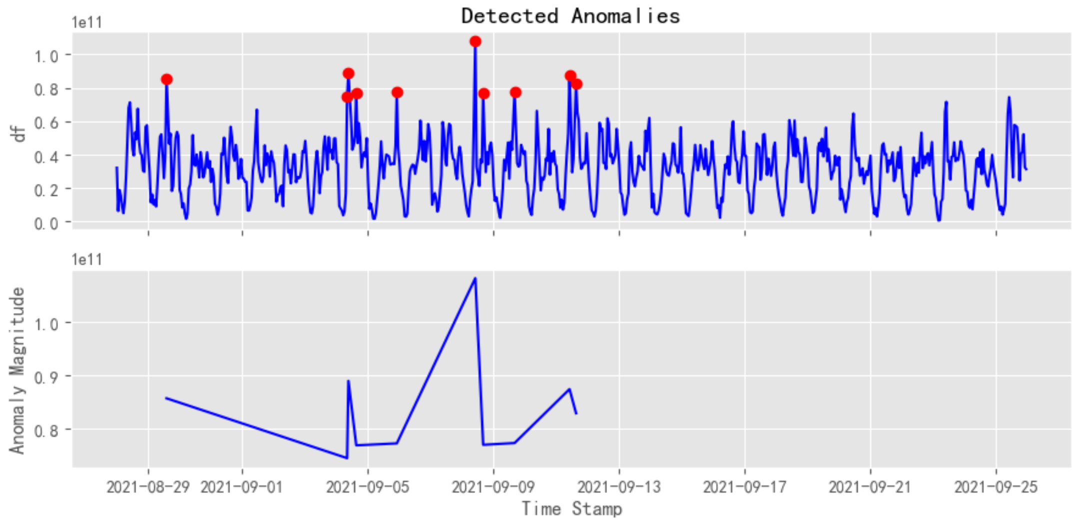

Figure 3: Anomaly detection based on S - ESD

Some of the above abnormal points are detected by S-ESD, combined with the periodicity and trend of the indicators, the time points that are outliers are extracted, and the average value at the same time point as the outliers is selected to replace the outliers to correct the time series of the indicators data.

3.3. Time series trend forecast based on Prophet

3.3.1. Prophet principle

Prophet, Facebook's open-source Python and R-based data forecasting tool, has gained a lot of attention with a simpler, more flexible way of forecasting and the ability to obtain forecasts comparable to experienced analysts. The overall construction of the model is as follows:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{3}$$

Among them, the model as a whole consists of three parts: growth trend, seasonal trend and the impact of holidays on the predicted value. Where $g(t)$ represents the growth function, which is used to fit the non-periodic changes of the predicted value in the time series;$s(t)$ is used to represent the cyclical changes, such as weekly, seasonal in each year, etc;$h(t)$ indicates the impact of potential holidays with a non-fixed period on the forecast value in the time series. Finally ε_t is the noise term, which represents the fluctuation not predicted by the model. Here, it is assumed that ε_t is Gaussian distributed.

3.3.2. Predictive model building based on Prohet

First, the time series display of the corrected data of the three indicators is carried out. It is found that the series still has some undetected noise. It can be resampled every day to obtain a new series to reduce the noise. detection. The specific model steps are as follows:

Prophet objects are created, holiday parameters are specified, and then trained on past data. Next, build a prediction framework, specify data granularity and prediction step, and then make predictions for future dates by step and draw prediction effect and component trend graphs. Finally, the data within the time period is detected, and a reasonable range of data fluctuation is obtained by adding or subtracting 3σ from the upper and lower bounds of Prophet prediction, and when the actual data exceeds this range, it is labeled as abnormal data.

For the time-series data of the three indicators, a time-series data fitting model is obtained to fit the data trends and fluctuations, and the upper and lower bounds of the predicted values are included in the prediction results of the model. The obtained results are shown in Figure 4.

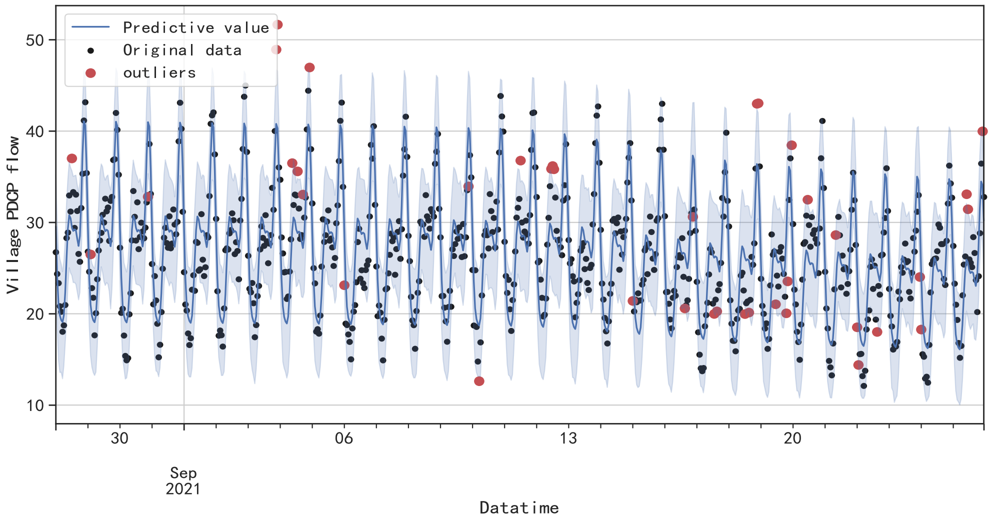

Figure 4: Prediction results of the model

Define a function outlier_detection function that uses the upper and lower bounds of the confidence interval of the model's predicted values to determine whether the sample is an outlier, and a total of 27 outliers are found from the sequence.

4. Based on the parameter tuning of Prohet prediction model

4.1. Grid Search and cross-validation

Grid search is a method of parameter tuning, generally tuning hyperparameters. It utilizes a simple brute force exhaustive search. Among all the candidate parameter selections, by looping through each possibility, the best performing parameter is the final result. Cross-validation is a method of evaluating statistical analysis, the generalization ability of machine learning algorithms to datasets independent of training data, and can avoid overfitting problems.

4.2. Model evaluation index MAPE

The accuracy of predicting continuous data is generally referred to as 1-MAPE.

$$MAPE = \sum_{t=1}^{n} \left| \frac{observed_t - predicted_t}{observed_t} \right| \times \frac{100}{n} \tag{4}$$

4.3. Grid parameter tuning based on Prophet model

In the constructed Prohet trend prediction model, changepoint_prior_scale may be the most influential parameter, which determines the flexibility of the trend, especially the degree of change of the trend at the trend change point. If it is too small, the trend will underfit and the variance that should be modeled using the trend change will end up being handled by the noise term. In addition, the seasonality_prior_scale parameter controls the flexibility of seasonality, a larger value allows the seasonality to accommodate larger fluctuations, and a smaller value reduces the magnitude of the seasonality. The holidays_prior_scal parameter controls flexibility to accommodate holiday effects. Some common parameters and their meanings are shown in Table 1:

ISSN 2616-5775 Vol. 5, Issue 4: 46-51, DOI: 10.25236/AJCIS.2022.050408

Table 1: Parameters and meanings of Prohet Prediction model

parameter	meaning
growth	Growth model,Divided into linear and nonlinear
Change Points	Time sequence of the change point, default value=None
n_changepoints	The number of potential changepoint
changepoint_prior_scale	The growth model of flexibility
seasonality_prior_scale	Adjust the strength of the seasonal component
holidays_prior_scale	Adjust the strength of the model components during holiday
seasonality_mode	business may be multiply seasonal time sequence
interval_width	The extent of the future trend of the time change
uncertainty_samples	Growth trend between the simulation drawing number

There are many parameters that can be adjusted based on the Prophet prediction model. Select the most important parameters for grid search combined with cross-validation to optimize the model, automatically traverse the possibility of all parameter combinations, according to the set evaluation function, automatically Choose the optimal parameter combination.

5. Results

In this paper, the statistical S-ESD anomaly detection algorithm is first used to detect anomalies in the three core indicators, and the detected anomalies are extracted and corrected. For the obvious anomalies that have not been detected, continue to build a Prophet trend prediction model on the indicator data, and make fitted predictions and visualizations of the existing dates. Labeled as anomalous data when the actual data exceeds the upper and lower confidence intervals of the predicted value. The model training of the three indicator data is carried out separately, and the next three days are predicted according to the step size, and the forecast data and trend of the next three days are obtained, as shown in Figure 5:

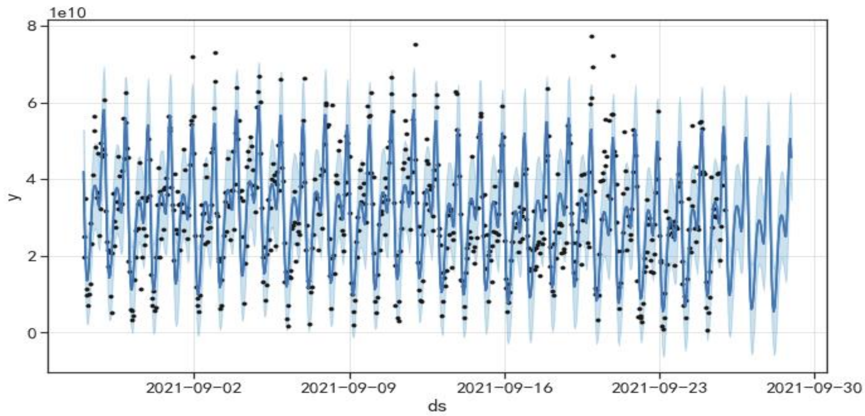

Figure 5: Final trend prediction results

We choose to use grid search combined with cross-validation, and use MAPE as the evaluation index to find the optimal parameter combination. Based on the efficiency of grid parameter tuning, two parameters, changepoint_prior_scale and seasonality_prior_scale, which can most affect the model are selected for combined parameter tuning. The optimal parameter combination is: changepoint_prior_ scale=0.1, seasonality_prior_scale=0.01, the MAPE value of the Prophet prediction model is 0.0783, and the prediction The error is less than 0.1, as shown in Table 2:

Table 2: Results of the grid search for tuning references

changepoint_prior_scale	seasonality_prior_scale	MAPE
0.001	0.01	0.143443
0.001	0.1	0.174466
0.001	1	0.10178
0.001	10	0.095428
0.01	0.01	0.093767
0.01	0.1	0.08023
0.01	1	0.080167
0.01	10	0.08011
0.1	0.01	0.078339

ISSN 2616-5775 Vol. 5, Issue 4: 46-51, DOI: 10.25236/AJCIS.2022.050408

6. Conclusion

In this paper, we use Twitter's open source S-ESD time series anomaly detection algorithm [5], which uses STL to decompose the time series, and then ESD test to detect outliers, plot the time series and label the outliers to get more reasonable results. Since the statistical-based anomaly detection method has a complete and mature theory, it is more interpretable. The anomaly detection in this paper does not need the label value of anomaly, but determines the outliers based on the historical change pattern of data, which requires less data quality and is more suitable for wide application than the other two types of anomaly detection methods.

In this paper, we take the trend prediction aspect of Facebook's open source Prophet model [7-8], and combine the anomaly detection method of Prophet to further improve the accuracy of the prediction, and reduce the MAPE of Prophet model evaluation through parameter tuning to achieve the same effect as Yang Zhen, Nie Yanwu et al. take the prediction based on time series seasonal models such as Prophet [9] with the same effect. The KPI performance index data based on intelligent operation and maintenance are often large-scale, and the time series prediction using Prophet model can guarantee the large-scale or fine-grained data prediction without losing the requirement of prediction accuracy, and also has the function of mutation point identification, which can be widely used in the field of intelligent operation and maintenance. However, the Prophet model in this paper can be further improved based on the combined Prophet-LSTM model-based time series prediction approach [10] adopted by Peng Pei, Liu Min et al. which can obtain higher prediction accuracy while keeping the model more interpretable.

References

[1] WU Rui, ZHANG Anqin, TIAN Xiuxia, et al. Anomaly Detection Algorithm for Power Data Based on Improved K-means [J]. Journal of East China Normal University:Natural Science Edition, 2020(4): 9.
[2] ZHONG Cheng, LUO Cheng. Clustering and Sequence Analysis Methods for Unsupervised Anomaly Detection [J]. Journal of Computer Research and Development, 2008(z1): 6.
[3] XU Qinzhen, YANG Luxi. An Anomaly Detection Method Based on Hierarchical Support Vector Machine for Supervised Local Decision Making [J]. Journal of Electronics and Information Technology, 2010(10): 5.
[4] CUI Weilan, YIN Xunwei, CHENG Yongqiang. A Network Traffic Anomaly Detection Method Based on Statistics [J]. Network Security Technology and Application, 2008(1): 2.
[5] Hochenbaum, Jordan, Owen S. Vallis, and Arun Kejariwal. "Automatic Anomaly Detection in the Cloud Via Statistical Learning." arXiv preprint arXiv:1704.07706 (2017).
[6] WANG Xin, WU Ji, LIU Chao, et al. Fault Time Series Prediction Based on LSTM Recurrent Neural Network [J]. Journal of Beihang University, 2018, 44(4): 13.
[7] Taylor SJ, Letham B. 2017. - Scale forecasting. PeerJ 5 Preprint: e3190v2
[8] Gorbach AN, Tseitlin NA Shopping behavior: spontaneous series analysis and regression models in market research. - Kiev: Osvita Ukrainy, 2011. -- 298 pages.
[9] YANG Zhen, NIE Yanwu, SUN Yahong, et al. Prediction and Comparative Analysis of Tuberculosis Incidence Based on Prophet et al. Time Series Seasonal Models [J]. Modern Preventive Medicine, 2021, 48(21): 7.
[10] PENG Pan, LIU Min. Short-term Load Forecasting Method Based on Prophet-LSTM Combined Model [J]. Proceedings of the CSU-EPSA, 2021, 33(11): 6.