 doi:10.20944/preprints202207.0407.v1

Anomaly Detection over Time Series Data

Zhu Zhenyi 20784183
HKUST
zzzhubh@connect.ust.hk

July 26, 2022

Abstract

The anomaly detection task is very important in computer science. And there are a lot of anomaly detection methods. Different from some thresholding methods, some unsupervised methods could make us get more accurate and faster result, which is the object of the project. In this paper, I tried to use EWMA and some other methods in two datasets: Webank time consuming indicators dataset and AIOps Challenge dataset. The paper consists nine parts: background of the project, related work, description of algorithms, implementation details, experimental setup and data sets used, experimental results and discussion, future directions, reference and meeting notes.

Key words: Anomaly Detection ; Time Series; EWMA

1. Background

Anomaly detection could generally understood to be the identification of rare events,observations, and items which deviate significantly from the majority of the data and do not conform to a well-defined notion of normal behavior. [1] Anomaly detection is applicable in a very large number and variety of domains, and is an important subarea of unsupervised machine learning. For example, it has many applications in fault detection, fraud information detection, event detection in sensor networks, cyber security intrusion detection, system health monitoring, detecting ecosystem disturbances, defect detection in images using machine vision, medical diagnosis and law enforcement. [2] WeBank is a Chinese private neobank, which was founded by Tencent, Baiyeyuan, Liye Group. Tencent company is now the single largest shareholder, with about 30 percent share ownership. WeBank's estimated valuation is almost 21 billion US dollars. David Ku is the company's CEO and Chairman.[3] So, for their online businesses, a system is needed to detect the anomaly of their transactions. In this project, we focus on time-consumption indicator and transaction time feature.

2. Related Work

2.1. Webank Anomaly Detection System

Based on information provided by WeBank's staff, their existing time series anomaly detection system mainly use some statistical methods such EWMA (Exponentially Weighted Moving Average) for detection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2022 doi:10.20944/preprints202207.0407.v1

2.2. ARIMA Anomaly Detection System

The ARIMA(AutoRegressive Integrated Moving Average) can be used as an anomaly detection model which has two parameters namely p and q. The p is for the AR (Auto Regression) and the q is for MA (Moving Average). The Auto Regression uses the previous lags to model the data and the Moving Average uses the previous forecast errors to model the data. [4]

2.3. LSTM Anomaly Detection System

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture[5] used in the field of machine learning and deep learning. LSTM networks are suitable for processing, classifying, and making the predictions based on time series data because there could exist lags of unknown duration between different events. And LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications. [6] So, this algorithm can be used to detection anomaly in KPI datasets.

3. Description of algorithms

The main algorithm I used in this anomaly detection task is EWMA. This algorithm is the core algorithm for Webank's detection system. EWMA is a first-order infinite impulse response filter that applies weighting factors which decrease exponentially. The weighting for each older datum decreases exponentially, never reaching zero. The following picture shows an example of the weight decrease.[7]

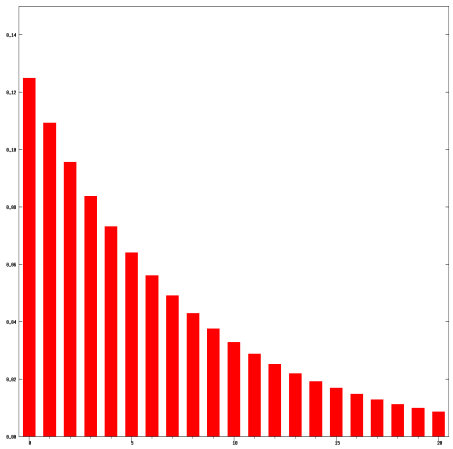


Figure 1: An example of the weight decrease

And the EWMA for a series Y can be calculated recursively as follows:

$$S_t = \begin{cases} Y_0, & t = 0 \\ Y_t + (1+\alpha) * S_{t-1}, & t > 0 \end{cases}$$

Where The coefficient α represents the degree of weighting decrease, a constant smoothing factor between 0 and 1. A higher α discounts older observations faster. Y_t is he value at a time period t. S_t is the value of the EMA at any time period t.[7]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2022 doi:10.20944/preprints202207.0407.v1

EWMA works well if we can make two assumptions about data, the values are Gaussian distributed around the mean and there is no seasonality.

Webank anomaly system divide their data into four different classes, so when used the algorithm to detect anomaly, I tried to use some methods such as DIFF-SKEW, LBTEST to judge the distribution of the data first. And if the distribution is a standard normal distribution, I would like to choose EWMA to find the anomaly. But if the distribution is not a normal distribution, I would like to choose other methods such as the 3-sigma criteria, CDF or difference to detect the anomaly in time series data.

4. Implementation details

4.1. Dataset One

This dataset is an open-source dataset comes from AIOps dataset. I choose different dataset group of KPI ID to perform anomaly detection. There are three columns in this dataset: timestamp, value and label. When the time series data is an anomaly point, the label equals to 1, and when the data is normal, then the label equals to 0. The label The basic structure of KPI ID da10a69f-d836-3baa-ad40-3e548ecf1fbd is as follows:

	timestamp	value	label
0	1476460800	0.012604	0
1	1476460860	0.017786	0
2	1476460920	0.012014	0
3	1476460980	0.017062	0
4	1476461040	0.023632	0
...	...	...	...
45459	1479193320	0.033835	0
45460	1479193380	0.030259	0
45461	1479193440	0.061780	0
45462	1479193500	0.021103	0
45463	1479193560	0.016680	0

45464 rows × 3 columns

Figure 2: Basic structure of AIOps dataset

4.1.1 Data analysis

Based on the dataset, I first try to do some data analysis to get some statistical description and distribution condition from the dataset. The following picture is the value distribution of the time series data. The horizontal axis represents timestamp and the vertical axis represents value.

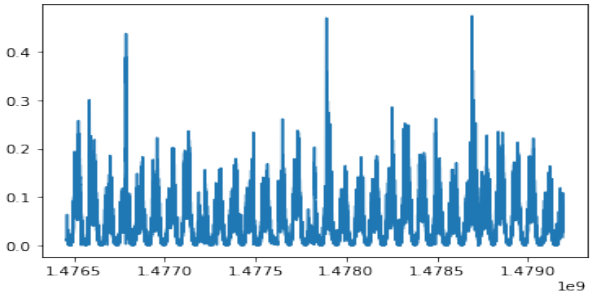


Figure 3: Value distribution

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2022 doi:10.20944/preprints202207.0407.v1

The following picture shows the original distribution of data in a more detailed scope. The picture contains timestamp before 5000. We could find that the data vary with time with some regularity.

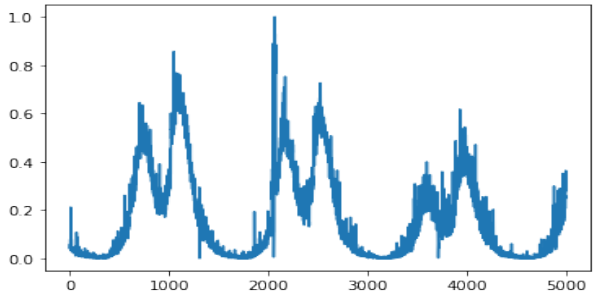


Figure 4: Detailed original value distribution

Then I prepared the normalized dataset and got similar distribution in Figure 4.

4.1.2 Skew and Lbtest

Before we use the EWMA model to detect the anomalies, we need to use some methods to determine the distribution of the data. And to help us tune parameters better, we need to use methods such as skew, lbtest and so on to help us choose the best parameter and standard.Here are some introduction to Skew and lbtest.

Diff-Skew
"Differential skew" refers to the time difference between the two single-ended signals in a differential pair. The operation of such links involves significant amounts of signal processing to recover clocks, reduce the effects of high-frequency losses, reduce ISI (intersymbol interference), and improve SNR. [8] It can be used to determine if the difference data distribution is normal distribution. If the diff-skewness is much greater than 1(positively skewed), or much less than -1(negatively skewed) , then we can consider the data are highly skewed.

LBTest
The Ljung–Box test is one of the statistical test of if any of a group of autocorrelations of a time series are different from zero. It can test overall randomness based on lags rather than test each distinct lag's randomness. And the statistical standard is very useful to help determine if the curve is stable or not. The LBTest would reject the independence of some values. If the p-value smaller than 0.05 in LBTest, then you can consider the values are dependent each other. And if p-value bigger than 0.05, which means you can't consider the dependence the values. [9]

Then I did some Skew and Lbtest experiment on the open dataset. After computing, we skew value equals to 1.9435 and lb-pvalue is 3.592352e-112. Which means that the distribution of 'difference' can be regarded as normal distribution, so we could use EWMA to find the anomaly.

4.1.3 EWMA and Anomaly Detection

The basic idea to detect anomalies is first use some tools introduced in 4.1.2 to determine if the distribution is normal distribution. If the distribution is normal distribution, we could use the EWMA model to fit our dataset. And then we can use some methods such as 3-sigma method, CDF (Cumulative Distribution Function) and so on to detect the outliers. However, if in step 2 we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2022 doi:10.20944/preprints202207.0407.v1

find that the distribution is not a standard normal distribution or even is not a normal distribution, we need to change the criteria in step 3. For example, we use more strict standard than 3-sigma method or we can just use difference to judge. And the following picture shows the result of EWMA for our original data, where I directly use the function EWM from Python. The anomaly then can be detected by use some standard to determine where the difference of original data and EWMA data is stable. More results about anomaly detection would be discussed in Part 6.

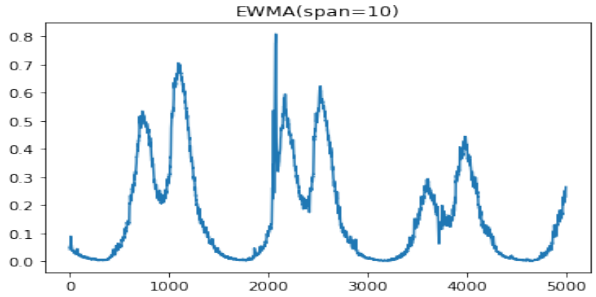


Figure 5: Original

4.2. Dataset Two

4.2.1 Data analysis

There are four main different kinds of classes for the time series data in Webank anomaly detection system. And the four classes are stable data, glitches, jagged and volatile data. The graph of the four different data are as follows. The first is the jagged graph.

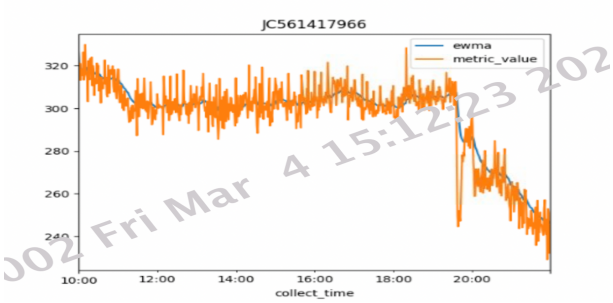


Figure 6: Jagged Graph

The following is volatile graph.

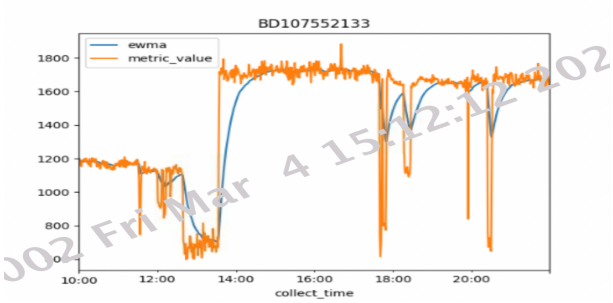


Figure 7: Volatile Graph

The following is stable graph.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2022 doi:10.20944/preprints202207.0407.v1

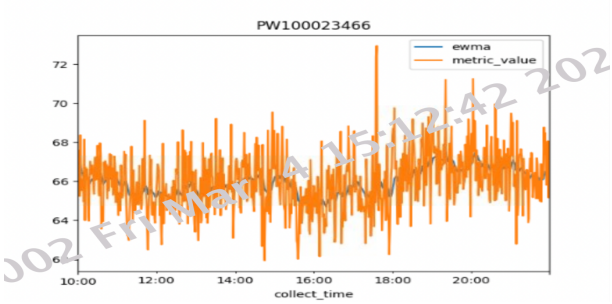


Figure 8: Stable Graph

The following graph shows the glitch curve.

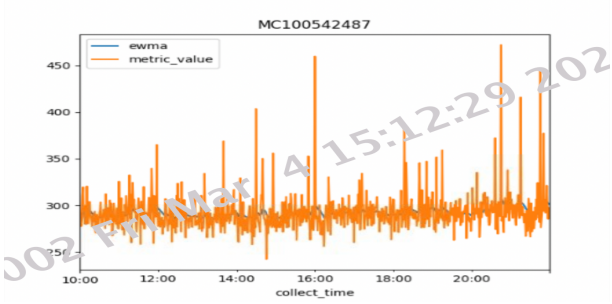


Figure 9: Glitches Graph

We need to change our threshold settings based on different conditions.

4.2.2 Skew and LBTest

Similar to 4.1, we need to use some methods to help us choose best parameters. I compute the skew value is 0.41 and lbtest p-value is close to 0 for the dataset. So the diff data is not too unstable, we could use EWMA to detect the anomalies.

4.2.3 EWMA and Anomaly Detection

Besides the ideas mentioned in 4.1.3, there are some conditions in Webank anomaly detection system because the dataset in Webank system is the real dataset for time-consumption indicator. When we do the anomaly detection, we only need to consider the start time and ignore the end time, we only need to consider the sudden increase as the anomaly condition. And when consider stable anomaly, the standard is if the time continues for more than three seconds, then we take the data as anomaly data. The accuracy of Webank's detection system is about 0.8. And the following picture shows the result of EWMA for Webank time-consumption data, where I directly use the function EWM from Python. More results about anomaly detection would be discussed in Part 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2022 doi:10.20944/preprints202207.0407.v1

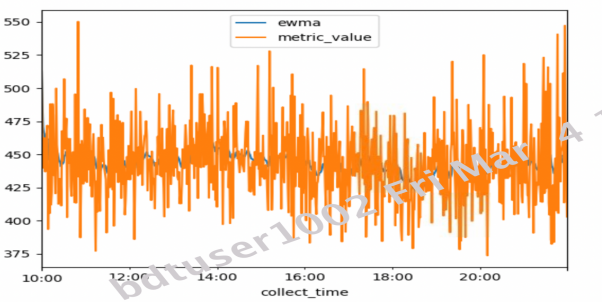


Figure 10: Jagged Graph

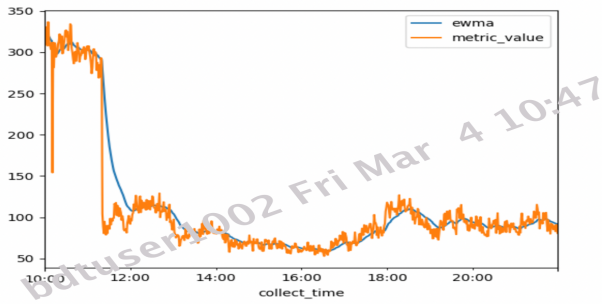


Figure 11: Jagged Graph

5. Experimental setup and data sets used

The first dataset comes from AIOps, and I choose some parts from the dataset to perform anomaly detection. And the platform is Jupyter notebook on my PC. The second dataset comes from Webank company, and due to the limitation from the company, I need to perform detection on the VDI (Huawei fusion access). And we need to add our IP to their whitelist to get access to the VDI. And the dataset from Webank is small and there are time-consumption data, label data and categories from the VDI. Because some irresistible conditions(online class, covid-19, quarantine and change houses, and Webank itself closed their network), my IP changes very frequently, I could not do enough work on Webank dataset.

6. Experimental results and discussion

6.1. Anomaly detection results for dataset one

First show the real anomaly graph

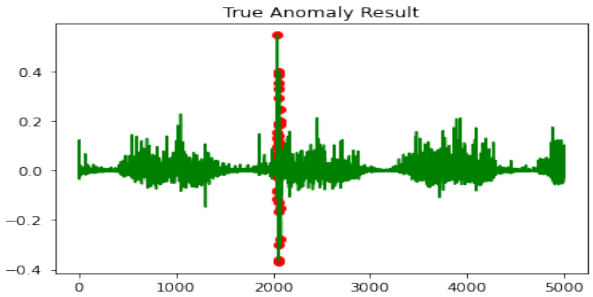


Figure 12: Real Anomaly Graph

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2022 doi:10.20944/preprints202207.0407.v1

The basic idea is discussed in part 4. And when detecting the anomalies, if the data is stable, then I set low threshold for stable dataset (such as 2, 2.5), high threshold for unstable dataset (such as 4, 5). I choose two different method to detect anomalies, the main method is n-sigma criteria. When choose n equals = 5.5, the result is as follows.

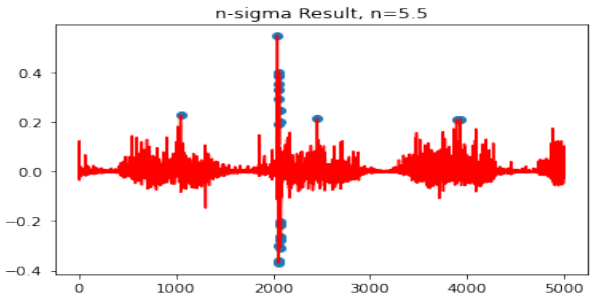


Figure 13: n-sigma Result Graph

There are 37 points that have been misclassification. So, the accuracy of the n-sigma detection is 0. 9926.And to present the results more clearly, I got the picture of anomaly detection result comparison as follows.

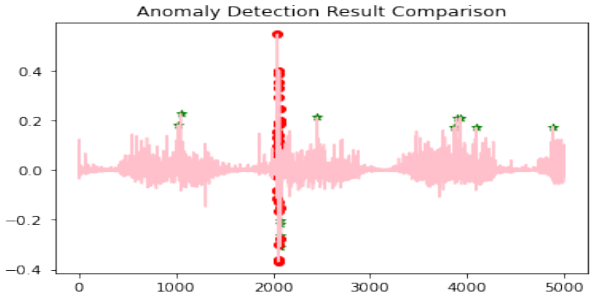
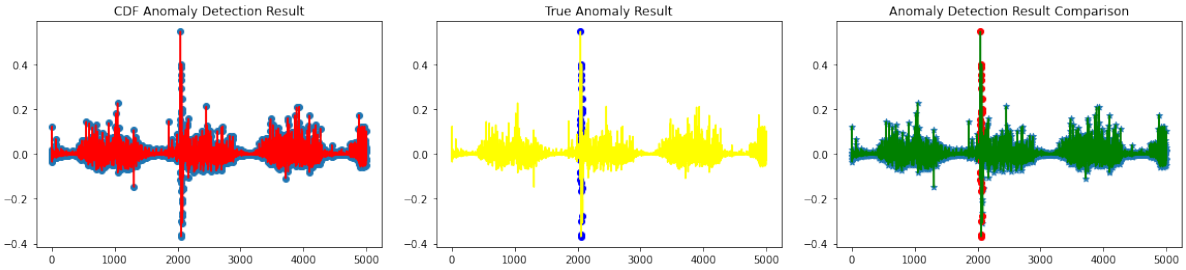


Figure 14: Anomaly Detection Comparison Result Graph

The method CDF is similar to n-sigma, so I would like to omit some details and just present the output result. We just need to change the threshold and some parameters, the following result(Figure 15.) shows the bad detection output because the wrong setting of threshold.



(a) (b) (c)

Figure 15: CDF Anomaly Detection Result

6.2. Anomaly detection results for dataset two

For different types of the curve, I tried different skew and LBTest to determine different curve and then change the threshold. Because some irresistible conditions I mentioned in part 5, I could

 doi:10.20944/preprints202207.0407.v1

not login in the VDI after March, so the results haven't been improved. On the stable data, the accuracy of the detection is about 0.8. But on some difficult data(Jagged), which has large skew and lbtest value, the accuracy is about 0.71.

6.3. Discussion

The dataset one is the transaction curve, while dataset two is time-consumption curve, the anomaly for second is to find the sudden increase, which may different from the first dataset in some details. When we use the open-source dataset to do anomaly detection, the result can be very good. However, when I try to do detection on Webank system, the accuracy is not too high. The best accuracy can only around 0.8. There are some reasons behind it, first is the volume of Webank system is not as much as the open-source dataset, the second and the most important reason I consider is that the data from Webank system comes from real world, while the open-source dataset may come from artificial construction, so the detection task is much harder in our Webank system.

7. Future directions

There are many methods to do anomaly detection for time series data. Based on the dataset provided by Webank and open-source platform, we can find that the statistical method EWMA is very useful and suitable for the detection because the accuracy of the method is high and the computing resource the method consume is not two much. But in the future, there could exist different kinds of time series dataset, so we should try different methods to fit the data. In the future, some statistical methods such as Holt-winters, ARIMA can be used for anomaly detection, and many methods in machine learning or natural language processing can also be used to detect the anomalies. For example, we could use LSTM (Long short-term memory), VAE (Variational autoencoder), or CNN models to do the detection.

8. References

[1] Chandola, V.; Banerjee, A.; Kumar, V. (2009). "Anomaly detection: A survey". ACM Computing Surveys. 41 (3): 1–58. doi:10.1145/1541880.1541882. S2CID 207172599.
[2] Aggarwal, Charu (2017). Outlier Analysis. Springer Publishing Company, Incorporated. ISBN 3319475770.
[3] "How Chinese FinTech helps serve the underserved?". USC Marshall. Retrieved 2019-10-06.
[4] [Online]. Available: https://www.analyticsvidhya.com/blog/2021/08/univariate-time-series-anomaly-detection-using-arima-model.
[5] [Online]. Available: https://en.wikipedia.org/wiki/Long short term memory.
[6] Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9(8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276. S2CID 1915014.
[7] [Online]. Available: https://en.wikipedia.org/wiki/Moving average.
[8] [Online]. Available: https://www.edn.com/handling-differential-skew-in-high-speed-serial-buses.
[9] [Online]. Available: https://stats.stackexchange.com/questions/64711/ljung-box-statistics-for-arima-residuals-in-r-confusing-test-results-comment124984-64711