Anomaly detection in multi-class time series

Weihong Wang¹, Zhuolin Wu²*, Xuan Liu³, Lei Jia⁴ and Xiaoguang Wang⁵

¹ College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang, 310000, China
² College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang, 310000, China
³ Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, B3H 4R2, Canada
⁴ Research and Development Center of Agricultural Bank of China, Hangzhou, Zhejiang, 310000, China
⁵ Beijing Broada Software Ltd, Hangzhou, Zhejiang, 310000, China
*Corresponding author’s e-mail: 2112012132@zjut.edu.cn

Abstract. For modern operation and maintenance systems, they are usually required to monitor multiple types and large quantities of machine’s key performance indicators (KPIs) at the same time with limited resources. In this paper, to tackle these problems, we propose a highly compatible time series anomaly detection model based on K-means clustering algorithm with a new Wavelet Feature Distance (WFD). Our work is inspired by some ideas from image processing and signal processing domain. Our model detects abnormalities in the time series datasets which are first clustered by K-means to boost the accuracy. Our experiments show significant accuracy improvements compared with traditional algorithms, and excellent compatibilities and operating efficiencies compared with algorithms based on deep learning.

1. Introduction

Huge amounts of data are generated daily for modern enterprises like financial securities companies. Generally, these data are arranged in strict accordance with the time sequence based on a statistical index. Based on our extensive experiences in real businesses, we summarize the following several challenges facing current time series anomaly detection services:

- Wide variety of data types.
  When monitoring tens of thousands of machine metrics simultaneously, different types of anomalies are associated with different tasks. For instance, innovational outlier, additive outlier, level shift, temporary change. Many models in the past have good detection accuracy only for a certain type of data, like DONUT [4] model.
- Limited computing resources.
  There are some good approaches for time series anomaly detection based on neural network like LSTM-based VAE-GAN [5]. However, these methods generally consume a lot of computing resources.
- Category change of data.
This is a point that many studies have missed. In reality, the type of data produced by a machine is not always the same, it changes with the services provided by the machine. Therefore, a successful model should possess the ability to detect these changes in time and respond accordingly.

With these challenges in mind, we decided to design a more competitive model which combines accuracy, time efficiency, and compatibility all together. Figure 1 shows the difference between our solution and traditional solutions. Our main design is to cluster the data first, then use different detectors to handle different types of data.

![Old solution vs Our solution](image)

Figure 1. Our solution compared to traditional solutions.

Our contributions for this paper are as follows:
- We designed four high-efficiency and high-accuracy anomaly detection algorithms named R detector, T detector, SR detector, SG detector. R and T detector are completely based on our own theory.
- We propose a self-selection algorithm model based on k-means clustering algorithm. We define the WFD distance which was adopted to measure the similarity between the sequences.
- We created and open sourced a well-labeled time series dataset: Robin. This data comes from the actual production data of an anonymous commercial bank.

2. Theory

Below we explain the basic theories used in our algorithm.

2.1. Periodic analysis

Regarding whether a sequence has periodicity, we can use contour plot of the square of wavelet coefficients’ modulus to describe it more vividly, or use the wavelet variance curve to visually determine its periodic intensity. The wavelet variance can be deduced as:

\[
w(a,b) = \int \mathcal{F}(t) \left| a^{\frac{1}{2}} \overline{\varphi(\frac{t-b}{a})} \right| dx
\]

\[
V(a) = \int_{-\infty}^{\infty} w(a,b)^2 db
\]

where \(w(a,b)\) represents wavelet coefficients, \(a\) is the scale factor, \(b\) is the translation factor, \(|a^{\frac{1}{2}} \overline{\varphi(\frac{t-b}{a})}|\) represents subwavelet generated by basis wavelet function \(\varphi(t)\), \(\overline{\varphi}\) is the complex conjugate function of \(\varphi\). \(f(t)\) requires to be a sequence exists in a square integrable real number space.
$V(a)$ represents the wavelet variance, which integrates the wavelet coefficients in the $b$ domain. The change curve of $V$ with $a$, which is what we call the wavelet variance curve.

![Wavelet Coefficients](image1)

![Wavelet Variance](image2)

Figure 2. It is clearly that the periodicity of this sequence is relatively strong at about 300 on the whole, but the periodicity is relatively weak near 3000 on the time scale, due to the fact that there is a descending anomaly near this position in the original sequence.

2.2. SR

Spectral Residual is an algorithm that uses Fast Fourier Transform (FFT) to calculate the salient part of an image. [6] has proved that SR also has great efficiency in time series anomaly detection. It uses the log amplitude spectrum of the input to subtract the average log amplitude spectrum to get the salient part. Given a sequence $x = (f_0, f_1, \ldots, f_n)$:

$$A(f) = \text{Amplitude}(\mathcal{F}(x))$$ (3)

$$P(f) = \text{Phrase}(\mathcal{F}(x))$$ (4)

$$L(f) = \log(A(f))$$ (5)

$$AL(f) = h \cdot f \cdot L(f)$$ (6)

$$R(f) = L(f) - AL(f)$$ (7)

$$S(x) = \| \mathcal{F}^{-1}(\exp(R(f) + iP(f))) \|$$ (8)

Where $\mathcal{F}$ denote Fourier Transform, $\mathcal{F}^{-1}$ denote Inverse Fourier Transform. The $S(x)$ we calculated above is called Saliency Map.

2.3. SG

Savitzky-Golay filter is widely used in data stream smoothing and denoising. It is a filtering method based on local polynomial least square fitting in time domain. When filtering the signal, the low-frequency components are fitted and the high-frequency components are smoothed out.
3. Method
In this section, we elaborate on our proposed model and framework.

3.1. Algorithms
Based on the theoretical basis of section 2, we now develop our own WFD distance and four detection algorithms.

3.1.1. WFD.
In our model, we mapped each sequence to a point \((x, y, z)\) in the three-dimensional space. Among them, \(x\) reflecting the relative strength of the overall fluctuation of the data. The coordinate \(y\) represents the standard deviation of the sequence, and \(z\) represents the periodic intensity of the sequence. Given a sequence \(f = (f_0, f_1, \ldots, f_n)\), this can be calculated as:

\[
x = 15 + \tan^{-1}\left(\frac{\overline{A}}{5} - 14\right)
\]

\[
y = \text{std}(f) = \sqrt{\frac{\sum_{i=0}^{n} (f_i - \overline{f})^2}{n}}
\]

\[
z = \max\left(V(a)\right), \ a \in (1 \ldots K)
\]

\(\overline{A}\) represents the mean value of the approximate coefficients after 5th order wavelet decomposition of the sequence, \(D_i\) represents the detail coefficients obtained by decomposition of each order. The value of \(K\) depends on the specific situation. \(V(a)\) is in (2). When the three coordinate values of a sequence are obtained, we use the Euclidean distance of the three-dimensional space coordinate points to determine the finally distance between the time series.

3.1.2. T detector.
This detection algorithm is mainly used to detect periodic abnormalities. Given a sequence \(f = (f_0, f_1, \ldots, f_n)\), it will be deduced as follows:

\[
y_{\text{half}} = \text{Half}\left(\frac{\mathcal{F}(f)}{\text{len}(f)}\right)
\]

\[
\text{peak} = \text{Maxpeak}(y_{\text{half}})
\]

\[
T = \frac{\text{len}(f)}{\text{peak}}
\]

where \(\text{Half}(a)\) means the first half of the sequence \(a\), \(\mathcal{F}\) denote Fourier Transform, \(\text{Maxpeak}(a)\) means the maximum crest of the sequence \(a\), \(\text{len}(a)\) means the length of sequence \(a\). \(T\) will be considered as the main oscillation period of the input sequence. Then, for each point \(f_i\):

\[
G_1 = \{i - T - 10 \ldots i - T + 10\}
\]

\[
G_2 = \{i - 2 \cdot T - 10 \ldots i - 2 \cdot T + 10\}
\]

\[
K_1 = \min\left(\{f_i - f_j\} | j \in G_1\right)
\]

\[
K_2 = \min\left(\{f_i - f_j\} | j \in G_2\right)
\]
The value of \( K \) depends on the specific situation. Finally, \( \text{Label}(i) = 1 \) means that \( f_i \) was detected as an anomaly point.

### 3.1.3. R detector

A very simple and effective way to locate mean drift abnormalities is to perform \( r \) transformation on the sequence \( f = (f_0, f_1, \ldots, f_n) \), which can be formulated as:

\[
    r_i = \frac{\sum_{j=1}^{i} f_j}{w_i} - \frac{\sum_{j=1}^{i} f_j}{\sum_{j=1}^{i} w_j}
\]

\[
    \text{Label}(i) = \begin{cases} 
        1, & \text{if } r_i - \bar{r} > 3 \times \text{std}(r) \\
        0, & \text{else}
    \end{cases}
\]

Where \( \text{std}(r) \) denotes the standard deviation of \( r \). \( \text{Label}(i) = 1 \) means that a mean shift abnormal occurred at \( f_i \), \( w \) depends on the specific situation. As shown in figure 3, in the \( r \)-space, the abnormal data can be observed clearly.

![Figure 3. Example of r transform.](image)

### 3.1.4. SR-\( \theta \) detector

After getting the Saliency Map, we defined the following rules to locate and identify outliers:

\[
    \text{Label}(i) = \begin{cases} 
        1, & \text{if } S_i - \bar{S} > \theta \times \text{std}(S) \\
        0, & \text{else}
    \end{cases}
\]

Where \( S \) denotes \( S(x) \) in equation (8). \( \theta \) is a parameter that depends on the actual situation. \( \text{std}(S) \) denotes the standard deviation of \( S \). \( \text{Label}(i) = 1 \) means that \( f_i \) was considered to be an anomaly point.

### 3.1.5. SG detector

Given a sequence \( f = (f_0, f_1, \ldots, f_n) \):

\[
    \text{Label}(i) = \begin{cases} 
        1, & \text{if } f_i > (1+r) \times \text{smooth}(i) \\
        1, & \text{if } f_i < (1-r) \times \text{smooth}(i) \\
        0, & \text{else}
    \end{cases}
\]
Where smooth denotes the smoothed value. \( r \) is a parameter depends on the actual situation. \( Label(i) = 1 \) means that \( f_i \) was considered to be an anomaly point.

3.2. Framework

Our model structure is mainly divided into two parts: the offline computing module responsible for time series clustering and the online detection module responsible for real-time detection. Different detectors will be saved in our detectors pool. The offline module automatically reads data periodically to select the cluster centers and finds a best detector for each cluster center, then writes it into the correspondence table to help provide better detection accuracy for the online detection module. When a new sequence to be detected reaches the detection module through the queue or pipeline, the program will quickly calculate the similarity between this sequence and the cluster centers. Then it will look up the correspondence table to obtain the corresponding detector.

![Model structure](image)

Figure 4. Model structure.

4. EXPERIMENTS

We contributed a dataset named Robin time-series-dataset to the open-source community. Data provided by an anonymous commercial bank, published at https://github.com/Voce-lin/Robin-time-series-dataset.

4.1. Datasets

- **Our Robin**: The dataset contains a total of 22 time series, each time series contains 2400 sampling points, the sampling interval is 5 minutes.
- **AIOps’ KPI-final**: This dataset is released by the 1st match for AIOps (2018AIOps). It contains 58 curves, 5922913 data points and 134114 tagged abnormal points.
- **Yahoo’s Webscope S5**: This dataset is provided by Yahoo to benchmark anomaly detection algorithms. The dataset consists of real time series and artificial time series with marked outliers. It contains 58 curves, 572966 data points and 3915 tagged abnormal points.
- **Numenta’s NAB**: It is a novel benchmark for evaluating algorithms for anomaly detection in streaming, real-time applications. It is composed of over 50 labeled real-world and artificial timeseries data files. 365558 data points and 33495 tagged abnormal points.
4.2. Metrics.
In some anomaly detection work like [4][6], researchers aggregate multiple consecutive abnormal points in a time series into an abnormal event. As long as the algorithm detects one point in an event, it is deemed to have correctly detected the abnormal event. This measurement strategy is relatively loose and is suitable for tasks that do not require high detection accuracy. In our work, we focus more on the accuracy, so we adopt a more rigorous measurement strategy, that is, when the algorithm detects an abnormal point, only this single point will be considered as abnormal. To determine an abnormal event, all abnormal points have to be detected. In our measurement strategy, we use 4 indicators, namely F1, recall, precision and CPU time.

4.3. Comparison of results.
For the experiment of this article, we used seven operators in the detectors pool with different parameters according to the mentioned four kinds of detection algorithm to meet our actual needs. We compared our model MCD with DONUT [4], LSTMAD[7], Banpei’s Hotelling[1] based on Hotelling's theory, luminol[2] and iqr_ad[3]. Among them, Hotelling, luminol, and iqr_ad do not require training data, so when comparing these three algorithms, we treat all the data as both the test set and the training set. For DONUT and LSTMAD, they are based on deep learning and require additional training data sets. Therefore, our strategy is to treat the first half of each series as the training set and the second half as the test set. In our algorithm model, this strategy is also adopted.

| Table 1. Results of cold-start comparison. |
|------------------------------------------|
| KPI | Yahoo | NAB | ABC |
|-----|-------|-----|-----|
|     | F1    | recall | precision | time | F1    | recall | precision | time | F1    | recall | precision | time | F1    | recall | precision | time |
| Hotelling | 0.36 | 0.40 | 0.33 | 19.44 | 0.38 | 0.52 | 0.30 | 12.82 | 0.26 | 0.19 | 0.40 | 4.59 | 0.16 | 0.40 | 0.10 | 0.20 |
| luminol | 0.12 | 0.23 | 0.08 | 932.82 | 0.46 | 0.66 | 0.36 | 149.65 | 0.31 | 0.29 | 0.34 | 15.25 | 0.18 | 0.45 | 0.11 | 1.01 |
| iqr_ad | 0.38 | 0.31 | 0.49 | 15.40 | 0.45 | 0.31 | 0.78 | 10.43 | 0.31 | 0.25 | 0.39 | 2.24 | 0.37 | 0.31 | 0.46 | 0.23 |
| MCD | 0.51 | 0.48 | 0.54 | 9.62 | 0.71 | 0.79 | 0.65 | 62.71 | 0.32 | 0.21 | 0.64 | 3.78 | 0.62 | 0.73 | 0.54 | 0.26 |

| Table 2. Results of train-test split comparison |
|-----------------------------------------------|
| KPI | Yahoo | NAB | ABC |
|-----|-------|-----|-----|
|     | F1    | recall | precision | time | F1    | recall | precision | time | F1    | recall | precision | time |
| LSTMAD | 0.52 | 0.51 | 0.54 | 1042 | 0.47 | 0.68 | 0.36 | 84.47 | 0.36 | 0.79 | 0.23 | 33.18 | 0.31 | 0.70 | 0.20 | 6.60 |
| DONUT | 0.44 | 0.42 | 0.47 | 7600 | 0.38 | 0.77 | 0.18 | 2431 | 0.30 | 0.78 | 0.18 | 406.15 | 0.57 | 0.54 | 0.60 | 52.80 |
| MCD | 0.50 | 0.51 | 0.49 | 84.83 | 0.73 | 0.77 | 0.60 | 156.09 | 0.31 | 0.35 | 0.42 | 53.89 | 0.74 | 0.73 | 0.74 | 12.21 |

It can be seen from table 1 that compared with these three traditional algorithms, the accuracy of our proposed algorithm model is the best. In terms of time-consuming performance, our performance is relatively stable. It can be seen from table 2 that the overall effect of the algorithm based on deep learning is higher than that of the traditional algorithm. The performance of our algorithm model still surpasses them in the Yahoo dataset and the Robin dataset. In addition, the detection time of these two deep learning algorithms rises sharply with the length of the time series (the length of the KPI data set is the highest), while our algorithm model does not have this phenomenon.

5. RELATED WORKS
STL [8], Based on loess, the data at a certain time is decomposed into trend component, seasonal component and residual. [9] Focus on the detection of time series up and down translation. In 2018, [4] proposed DONUT, they implement unsupervised anomaly detection using VAE, [6] Combined SR with
CNN neural network. [10] proposed a robust anomaly detection algorithm (MEDIFF) to monitor online business metrics in real time, using robust statistical metric--median--of the time series to decouple the trend and seasonal components. The above methods can achieve satisfied detection results on specific types of time series, but our model can deal with more types of anomalies at the same time.

6. CONCLUSIONS & FUTURE PLAN
Time series anomaly detection is still an extremely challenging topic. According to the experimental data obtained from our proposed model, what we can foresee is that pre-classifying a large number of time series and then detecting them is a path worth exploring. Our experiments showed significant improvements. Compared with both traditional algorithms and algorithms based on deep learning, our F1-score has improved by nearly 40% on average, our CPU-time consumption are relatively stable on different datasets, showing excellent efficiency and compatibility. Our model provided well satisfied intelligent operation and maintenance experience for our customers in the actual business.

In the future, we will work closely with engineers to find better feature engineering methods and get better clustering effect. We will also consider using decision tree in the model and some lightweight deep learning methods based on neural network to deploy high-precision anomaly detection for core services and services with higher priority. And we will provide fast and real-time detection services for massive low priority business simultaneously.

References
[1] GitHub. (2020) Banpei. https://github.com/tsruebee/banpei.
[2] GitHub. (2018) luminol. https://github.com/linkedin/luminol.
[3] GitHub. (2020) Anomaly Detection Toolkit (ADTK). https://github.com/arundo/adtk.
[4] Xu, H., Feng, Y., Chen, J., Wang, Z., Qiao, H., and Chen, W., et al. (2018). Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. 187-196.
[5] Niu, Z., Yu, K., and Wu, X. (2020). LSTM-based VAE-GAN for time-series anomaly detection. Sensors, 20(13), 3738. doi:10.3390/s20133738
[6] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, and Xiaoyu Kou, et al. 2019. Time-series anomaly detection service at Microsoft. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19). 3009–3017. DOI:https://doi.org/10.1145/3292500.3330680
[7] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. 2015. Long short term memory networks for anomaly detection in time series. In 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015.
[8] Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: a seasonal-trend decomposition procedure based on Loess. J Off Stat 6: 3–73.
[9] Ma, M., Zhang, S., Dan, P., Xin, H., and Dai, H. (2018). Robust and Rapid Adaption for Concept Drift in Software System Anomaly Detection. 2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE). IEEE.
[10] Li, T., Geng, Y., and H Jiang. (2020). Anomaly detection on seasonal metrics via robust time series decomposition.