EasyChair Preprint

№ 8445

# Univariate Time Series Anomaly Detection Based on Variational AutoEncoder

Leehter Yao and Youwei Chang



July 10, 2022

# Univariate Time Series Anomaly Detection Based on Variational AutoEncoder

1st Leehter Yao
*dept. of Electrical Engineering*
*National Taipei University of Technology*
1, Sec.3, Chung Hsiao E. Rd, Taipei 106, Taiwan
ltyao@ntut.edu.tw

2nd You-Wei Chang
*Graduate Institute of Automation Technology*
*National Taipei University of Technology*
1, Sec.3, Chung Hsiao E. Rd, Taipei 106, Taiwan
t109618041@ntut.edu.tw

***Abstract*—In the field of anomaly detection, the boundaries of anomalies are always blurred, and professional knowledge is required to define them, which consumes a lot of manpower and time to mark what anomalies are. In this paper, a Variational Auto-Encoder(VAE) neural network model is used, and an unsupervised learning anomaly detection model that considers both temporal dependencies and reconstructed features. In the calculus of marking outliers, we propose a two-dimensional sliding window with a clustering algorithm to solve the traditional method of judging outliers using a single threshold. Experimental results based on Yahoo Webscope dataset show that the performance can be ameliorated by the proposed method.**

***Index Terms*—Anomaly detection, Variational Auto-Encoder, two-dimensional sliding window**

## I. Introduction

Anomaly detection has been one of the areas of machine learning research for a long time due to the wide range of applications. In everyday life, the anomalies we observe are the focus of our attention. When something deviates significantly from the rest of the distribution, it is marked as an anomaly or outlier. Anomaly is a well-defined normal behavior in the data that is different from the behavior pattern, also known as discordant observations in different fields or exceptions , etc.

Anomaly detection algorithms usually output their results for use and verification. There are two common ways of outputting results [1]. One is anomaly scores, anomaly detection algorithms score each piece of data. The degree of abnormality of the data to the normal situation. The second is abnormal mark, marking each piece of data as a binary mark of abnormal or normal. The anomaly score is more flexible in application. In actual use, the anomaly score can be combined with the threshold to generate the anomaly label. The anomaly label can be used to determine the index of the algorithm, and the result of the model is good or bad, which is more clear in application.

Common anomaly detection algorithms include statistical-based anomaly detection. Statistical anomaly detection algorithms can generally be divided into two categories, parametric methods and nonparametric methods. Parametric methods, such as Gaussian Model and Regression Model, use the Gaussian distribution of the data to find values other than three standard deviations from the mean value as abnormal data [2], and the non-parametric method has Histogram-Based algorithm [3] [4], this method first optimizes the algorithm model by establishing a histogram, if the test data can be divided into a certain histogram , it is normal, otherwise it is abnormal. According to the set size of the histogram, the abnormal and normal classification will be affected. If the histogram is set too small, the normal data may be outside the histogram, otherwise, the abnormal data may be within the histogram.

A novel VAE based anomaly detector is proposed in this paper. A generative model, VAE, will be used to extract the high level features and reconstruction errors.

## II. Methodology

### A. Architecture of Variational Auto-Encoder

The framework of the time series anomaly detection system developed by this research can be divided into two frameworks. One is the time series reconstructor, which reconstructs the data to the original through data preprocessing and Variational Autoencoder. the second , as an anomaly detector, according to the reconstruction error, the two-dimensional diagonal sliding window proposed by us and the clustering algorithm are used to detect the cluster to which the anomalous data belongs.

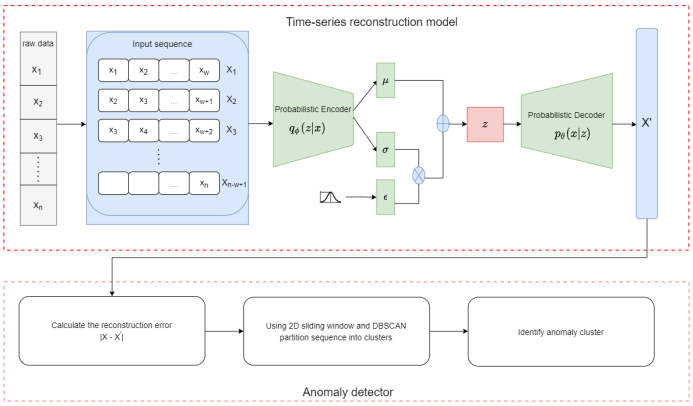


Fig. 1. Time Series Anomaly Detection Architecture

VAE is a generative model. The model can learn a distribution $P(X)$ that is very close to the real distribution of the data, and then generate real data through $P(X)$, but this is

not easy to get, since there are so many factors that affect the distribution of the real data, we cannot be exhaustive about it, so introduce a random variable $z$ in the generative model to help the modeling.Suppose there is a dataset $X = \{x^{(i)}\}_{i=1}^{m}$ ,$m$ is the feature.The function is shown as follow:

$$P_\theta(X^{(i)}) = \int P_\theta(x^{(i)}|z)P_\theta(z)dz \tag{1}$$

$$\theta^* = \arg\max_\theta \prod_{i=1}^{n} P_\theta(x^{(i)}) \tag{2}$$

Unfortunately, it is not easy to compute $P_\theta(X^{(i)})$ in this way, as it is very expensive to check all the possible values of $z$ and sum them up. To narrow down the value space to facilitate faster search, we would like to introduce a new approximation function to output what is a likely code given an input $x$, $q_\phi(z|x)$, parametrized by $\phi$.

### B. Loss Function

The estimated posterior $q_\phi(z|x)$ should be very close to the real one $P_\theta(z|x)$. We can use Kullback-Leibler divergence to quantify the distance between these two distributions. KL divergence $D_{KL}(X||Y)$ measures how much information is lost if the distribution $Y$ is used to represent $X$. In our case we want to minimize $D_{KL}(q_\phi(z|x)||p_\theta(z|x))$ with respect to $\phi$. So we have:

$$\begin{aligned} D_{KL}(q_\phi(z|x)||p_\theta(z|x)) = log(P_\theta(x)) \\ +D_{KL}(q_\phi(z|x)||p_\theta(z)) - E_{z\ q_\phi(z|x)} log(p_(x|z)) \end{aligned} \tag{3}$$

we want to maximize the log-likelihood of generating real data (that is $log(p_(x)$ and also minimize the difference between the real and estimated posterior distributions (the term $D_{KL}$, works like a regularizer). Note that $P_\theta(x)$ is fixed with respect to $q_\phi$. The loss function is shown as follow:

$$\begin{aligned} L_{VAE}(\theta,\phi) &= -log(P_\theta(x)) + D_{KL}(q_\phi(z|x)||P_\theta(Z|x)) \\ &= -E_{z\ q_\phi(z|x)} log(p_(x|z)) + D_{KL}(q_\phi(z|x)||p_\theta(z)) \end{aligned} \tag{4}$$

$$\theta^*, \phi* = \arg\min_{\theta,\phi} L_{VAE} \tag{5}$$

In Variational Bayesian methods, this loss function is known as the variational lower bound, or evidence lower bound. The "lower bound" part in the name comes from the fact that KL divergence is always non-negative and thus $-L_{VAE}$ is the lower bound of $log(P_\theta(x))$.

$$-L_{VAE} = log(P_\theta(x)) - D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \leq log(P_\theta(x)) \tag{6}$$

### C. Training

Ideally, we could train the VAE with the entire time series as input ,and the model output the entire reconstruction. However, in many practical applications, time series are often very long.

To solve this problem, we use sliding windows technique(7),dividing the long time series into short chunks. The sliding windows is controlled by one parameters: window size $w$ . Specifically, for each dataset $x \in \mathbb{R}^T$ , we have:

$$\mathcal{X} = \begin{pmatrix} x_1 & x_2 & \cdots & x_w \\ x_2 & x_3 & \cdots & x_{w+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_T \end{pmatrix} \tag{7}$$

$$L = T - w + 1 \tag{8}$$

### D. Anomaly Score

The next step is to compute anomaly score for each data point $x$, the input data of VAE model is a chunk of time series $\mathcal{X}$, and the reconstruction output of the model $\mathcal{X}'$. We choose absolute reconstruction error as error $e$, so we have :

$$e = |\mathcal{X} - \mathcal{X}'| \tag{9}$$

$$e = \begin{pmatrix} x_1 - x'_1 & x_2 - x'_2 & \cdots & x_w - x'_w \\ x_2 - x'_2 & x_3 - x'_3 & \cdots & x_{w+1} - x'_{w+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L - x'_L & x_{L+1} - x'_{L+1} & \cdots & x_T - x'_T \end{pmatrix} \tag{10}$$

$$E = \begin{pmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,w} \\ e_{2,2} & e_{2,3} & \cdots & e_{2,w+1} \\ \vdots & \vdots & \ddots & \vdots \\ e_{L,L} & e_{L,L+1} & \cdots & e_{L,T} \end{pmatrix} \tag{11}$$

We take the anti-diagonal of the error matrix(11)slide the $w/timesw$ window size. We call this technique two-dimension sliding window , and we have diagonal error $DE$ :

$$DE = \begin{pmatrix} e_{1,w} & e_{2,w} & \cdots e_{w,w} & \\ e_{2,w+1} & e_{3,w+1} & \cdots e_{w+1,w+1} & \\ \vdots & \vdots & \ddots & \vdots \\ e_{(L-w+1),L} & e_{(L-w+2),L} & \cdots e_{L,L} & \end{pmatrix} \tag{12}$$

## III. EXPERIMENTAL RESULTS

### A. DATA SET DESCRIPTION

Yahoo Webscope data set is a publicly available data set released by Yahoo Labs. This data set consists of 367 real and synthetic time series with point anomaly labels. Each time series contains 1, 420 - 1, 680 instances. This anomaly detection benchmark is further divided into four sub-benchmarks namely A1 Benchmark, A2 Benchmark, A3 Benchmark, and A4 Benchmark.

A1 Benchmark contains real Yahoo membership login data, which tracks the aggregate status of logins on Yahoo network , whereas, other three sub-benchmarks contain synthetic data. A2Benchmark and A3Benchmark contain only outliers, while A4Benchmark also contains change-point anomalies. In synthetic data, outliers are present on random positions. In each data file, there is a Boolean attribute – label – indicating if the

value at a particular time stamp is considered as anomalous or normal. In addition to value and label, A3Benchmark and A4Benchmark contain additional fields such as change-point, trend, noise, and seasonality. However, we are discarding all the additional attributes and only using value attribute for all the experiments.

### *B. EVALUATION METRIC*

F-score is most commonly used singleton metric which serves as an indicator of the model's performance. Therefore, we employed F-score (13) as the evaluation metric for our models. All the anomaly detection methods in this experimental setting are applied on each time series separately. Average F-scores per sub-benchmark are reported for each method.

$$F - score = 2.\frac{Precision.Recall}{Precision + Recall} \tag{13}$$

## REFERENCES


[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Comput. Surv, vol. 41, Jul. 1,2009, doi: 10.1145/1541880.1541882.

[2] W. A. Shewhart, Economic control of quality of manufactured product. Macmillan and Co Ltd, London, 1931.

[3] D. Dasgupta and F. Nino, "A comparison of negative and positive selection algorithmsin novel pattern detection," in Smc 2000 conference proceedings. 2000 ieeeinternational conference on systems, man and cybernetics.'cybernetics evolving tosystems, humans, organizations, and their complex interactions'(cat. no. 0, 2000, vol.1: IEEE, pp. 125-130.

[4] P. Helman and J. Bhangoo, "A statistically based system for prioritizing information exploration under uncertainty," Ieee transactions on systems, man, and cybernetics-parta: Systems and humans, vol. 27, no. 4, pp. 449-466, 1997.