Towards a Rigorous Evaluation of Time-series Anomaly Detection

Siwon Kim, Kukjin Choi, et al.

Kyunghee Kim

Department of Statistics
Sungkyunkwan University

January 16, 2022

Overview

1. Introduction
2. Experimental results
3. Discussion
4. Further discussion

Introduction

Anomaly detection; AD

As Industry 4.0 accelerates system automation, consequences of system failures can have significant social impact(Lee 2008). To prevent this failure, the detection of anomalous is more important.

Anomalies

Synonym: outlier observation, anomaly, discrodant observation, discords, contaminants

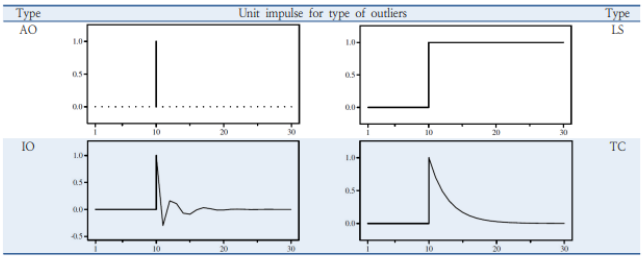
- Unwanted Data: noise, erroneous
- Event of interest: Analyze the outlier itself

Introduction

Types of anomaly

- Point anomaly
  1. Additional Outlier(AO)
  2. Innovational Outlier(IO)
- Contextural anomaly
  1. Level shift(LS)
  2. Temporal Change(TC)

Table 3. Type of outliers (López, 2016)



Time series Anomaly Detection; TAD

Unsupervised TAD

Normal data are accesible during the training time and assign different anomaly scores to inputs depending on the degree of their abnormality.

i.e., low anomaly scores: normal inputs, high anomaly scores: abnormal inputs

- Reconstruction-based AD (autoencoder, GAN)

  Minimize the distance between a normal input and its reconstruction
- Forecasting-based AD (LSTM, gated recurrent unit)

  Distance between the predicted and ground truth signal is an anomaly score.
- Others (GNNs: Graph neural networks, RADM: Real Time anomaly detection in multivariate time series)

TAD

TAD datasets
SWaT
WADI
SMD
MSL and SMAP

Table: benchmark datasets

TAD methods
USAD(Unsupervise anomaly detection)
DAGMM(Deep Autoencoding Gaussian mixture model)
LSTM-VAE(LSTM-based variational autoencoder)
OmniAnomaly
MSCRED(Multi-scale convolutional recurrent encoder-decoder)
THOC(temporal hierarchical one-class network)

Table: TAD method examples

Point adjustment; PA

For some datasets, the reported F1 scores exceed 0.9, giving an encouraging impression of today's TAD capabilities. However, we need to consider most of current TAD methods measure the F1 score after **Point adjustment(PA)** proposed by(Xu et al. 2018)(Su et al. 2019; Audibert et al. 2020; Shen, Li, and Kwok 2020).

PA works as follows:

If at least one moment in a contiguous anomaly segment is detected as an anomaly, the entire segment is then considered to be correctly detected.

- $F1$: Computed without PA
- $F1_{PA}$: Computed with PA

TAD formulation1

$\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_T}\}$ is normalized and split into a series of windows $\mathbf{W} = \{\mathbf{w_1}, \ldots, \mathbf{W_{T-\tau+1}}\}$, where $w_t = \{x_t, \ldots, x_{t+\tau-1}\}$ and τ is the window size.

Anomaly label is $y_t \in \{0, 1\}$ and labels are obtained by comparing anomaly score $A(w_t)$ with TAD threshold δ

$$\hat{y_t} = \begin{cases} 1, & A(w_t) > \delta \\ 0, & otherwise. \end{cases} \tag{1}$$

where, $A(w_t) = MSE(w_t, \hat{w_t}) = \frac{1}{\tau}||w_t - \hat{w_t}||_2$. After find $\hat{y_t}$ compute precision(P), recall(R), and F1 score.

TAD formulation2

Multiple anomaly segments.

S: set of M anomaly segments $\mathbf{S} = \{S_1, \ldots, S_M\}$, where $S_m = \{t_s^m, \ldots, t_e^m\}$. t_s^m, t_e^m: start and end times of S_m.

$$PA\ adjusted\ \hat{y}_t = \begin{cases} 1, & A(w_t) > \delta\ or\ t \in S_m\ and \quad \exists_{t' \in S_m} A(w_{t'}) > \delta \\ 0, & otherwise. \end{cases} \tag{4}$$

According to equation (4), after PA, the P, R, F1 can only increase.

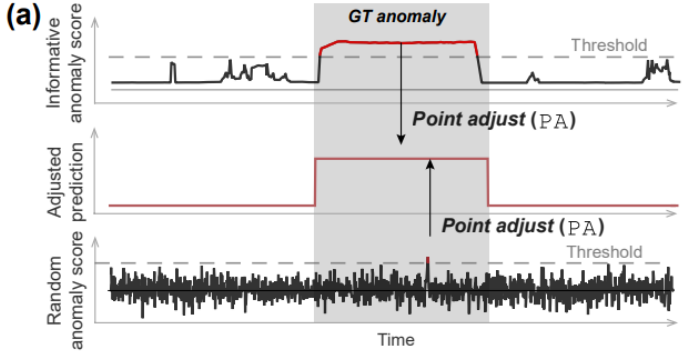
Problems with PA

1. High possibility of overestimating the model performance.
2. Without PA, existing methods exhibit no improvement over the baseline.

In this paper, raise the question of whether the current TAD methods are being properly evaluated and suggest directions for rigorous evaluation of TAD.

Problems with PA

1. Overestimate the model performance



Despite their disparity, the predictions afterr PA become indistinguishable.

It is difficult to conclude that a model with a higher $F1_{PA}$ performs better than the others.

Restate Recall

Where $\gamma = Pr(t \in S)$ is a test dataset anomaly ratio and $Pr(A(w_{t'}) < \delta^{'}) = \int_0^{\delta^{'}} U(0,1) = \delta^{'}$.

$$
\begin{aligned}
R =& Pr(\hat{y}_t = 1 | y_t = 1) \\
=& Pr(\hat{y}_t = 1 | t \in S) \\
=& 1 - Pr(\hat{y}_t = 0 | t \in S) \\
=& 1 - Pr(\forall t^{'} \in S, A(w_{t'}) < \delta^{'} | t \in S) \\
=& 1 - \Pi_{t' \in S} Pr(A(w_{t'}) < \delta^{'} | t \in S) \\
=& 1 - \frac{1}{\gamma} \Pi_{t' \in S} Pr(A(w_{t'}) < \delta^{'}) \\
=& 1 - \delta^{'(t_e - t_s)}
\end{aligned}
\tag{6}
$$

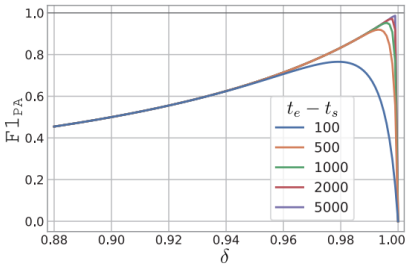
Restate Precision

The anomaly ratio $\gamma \in [0,0.2]$ and $t_e - t_s$ ranges from 100 to 5,000.

$$
\begin{aligned}
P =& Pr(y_t = 1 | \hat{y}_t = 1) \\
=& \frac{Pr(\hat{y}_t = 1 | y_t = 1) Pr(y_t = 1)}{Pr(\hat{y}_t = 1)} \\
=& R \times \frac{Pr(y_t = 1)}{Pr(\hat{y}_t = 1)} \\
=& R \times \frac{\gamma}{Pr(\hat{y}_t = 1, y_t = 1) + Pr(y_t \hat{=} 1, y_t = 0)} \\
=& (1 - \delta'^{(t_e - t_s)}) \frac{\gamma}{(\gamma - \delta'^{(t_e - t_s)}) + (1 - \gamma)(1 - \delta')}.
\end{aligned}
\tag{7}
$$

Untrained model with comparably high F1

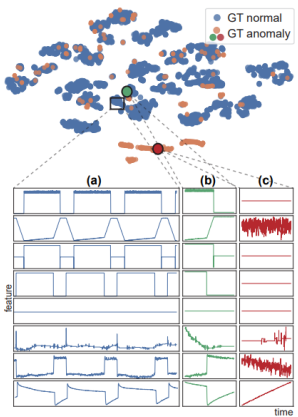
Varying with δ', $t_e - t_s$, we can always obtain the $F1_{PA}$ close to 1 **by changing** δ'.



Without training, the outputs are likely to close to zero. However the effect of PA, obtained from an untrained model look informative.

Also it is shown that $F1_{PA}$ increases even more when the window size gets longer.

New protocol for TAD; PA%K



Test Dataset: SWaT(TAD benchmark dataset)

Blue: Normal samples

Orange: Abnormal samples

GT anomalies: (b), (c)

(b) shared more pattern with normal data(a) than (c).

$\rightarrow$ Due to incomplete test set labeling such as (b), F1 can underesti

Therefore, the paper propose a new evaluation protocol PA%K.

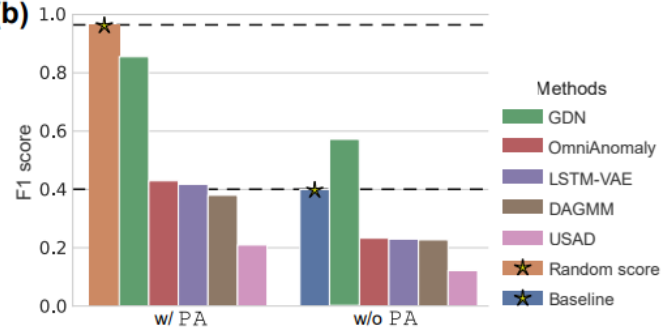
PA%K

- Effects
  1. mitigate the overestimation effect of $F1_{PA}$
  2. mitigate the possibility of underestimation of F1.
- The idea
  1. threshold: K ($0 \geq K \geq 100$)
  2. modify (4), $\hat{y}_t = \begin{cases} 1, & F1_{PA} > \delta \text{ or } t \in S_m \text{ and } \frac{|\{t' | t' \in S_m, A(w_{t'}) > \delta\}|}{|S_m|} > K \\ 0, & otherwise. \end{cases}$

     The idea of PA%K is to apply PA to Sm only if the ratio of the number of correctly detected anomalies in Sm to its length exceeds the threshold K.

Problems with PA

2. Without PA, existing TAD have no improvement over the baseline



Mostly, TAD methods do not seem to have obtained a significant improvement over the baseline that this paper proposes. Furthermore, several methods fail to exceed it.

New baseline for TAD

Classification baseline

Case1: Defined as a random guess

New baseline

$$A(w_t) = ||w_t - \eta||_2 \simeq ||w_t||_2 \tag{8}$$

where $\eta = f_\theta(w_t)$ and $\theta \sim \mathcal{N}(0, \sigma^2)$,

Case2: input itself($\eta = 0$), Case3: $\eta \neq 0$ but most of them have a value of zero.

Setting

Datasets: Benchmark TAD datasets on table1

Methods: on table2

Baseline

1. Random anomaly score: $\mathcal{A}(w_t) \sim \mathcal{U}(0, 1)$
2. Input itself an anomaly score(New): $\mathcal{A}(w_t) = ||w_t||_2 \leftrightarrow \eta = 0$, extreme case of Eq. 8.
3. Anomaly score from the randomized model(New): Eq. 8.

Comparison results1: Correlation between $F1_{PA}$ and F1

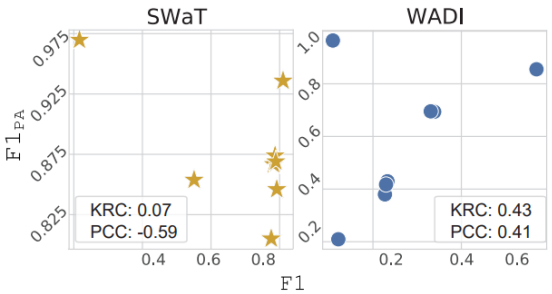


Figure 4: Correlation between $F1_{PA}$ and F1 of the existing methods on SWaT and WADI dataset. The Kendall rank correlation (KRC) and Pearson correlation coefficient (PCC) are indicated in the figure.

However, these numbers are insufficient to assure the existence of correlation and inference using only $F1_{PA}$ may have the risk of improper evaluation of the detection performance.

Comparison results2

Compare the results of the AD methods with case 1-3. (Use the best numbers reported in the original papers and officially reproduced results(Choi et al. 2021); if there were no available scores, reproduced them referring to the officially provided codes.)

- The reproduced results are marked as †.
- Bold and underlined indicate the best and second best scores.
- The uparrow(↑) means
  1. $F1_{PA}$ is higher than Case 1
  2. F1 is higher than Case 2 or 3

Comparison results2 (Con't)

	SWaT		WADI		MSL		SMAP		SMD	
	$F1_{PA}$	F1	$F1_{PA}$	F1	$F1_{PA}$	F1	$F1_{PA}$	F1	$F1_{PA}$	F1
USAD	0.846 (↓)	0.791 (↑)	0.429 (↓)	0.232 (↓)	0.927 (↓)	0.211† (↓)	0.818 (↓)	0.228† (↓)	0.938(↑)	0.426† (↓)
DAGMM	0.853 (↓)	0.550 (↓)	0.209 (↓)	0.121 (↓)	0.701 (↓)	0.199† (↓)	0.712 (↓)	**0.333**† (↑)	0.723 (↓)	0.238† (↓)
LSTM-VAE	0.805 (↓)	0.775 (↓)	0.380 (↓)	0.227 (↓)	0.678 (↓)	0.212† (↓)	0.756 (↓)	0.235† (↑)	0.808 (↑)	0.435† (↓)
OmniAnomaly	0.866 (↓)	0.782 (↓)	0.417 (↓)	0.223 (↓)	0.899 (↓)	0.207† (↓)	0.805 (↓)	0.227† (↓)	**0.944** (↑)	0.474† (↓)
MSCRED	0.868 (↓)	0.662† (↓)	0.346 (↓)	0.087† (↓)	0.775 (↓)	0.199† (↓)	0.942† (↓)	0.232† (↑)	0.389† (↓)	0.097† (↓)
THOC	0.880 (↓)	0.612† (↓)	0.506 (↓)	0.130† (↓)	0.891 (↓)	0.190† (↓)	0.781† (↓)	0.240† (↑)	0.541† (↓)	0.168† (↓)
GDN	0.935 (↓)	**0.81** (↑)	0.855 (↓)	**0.57** (↑)	0.903 (↓)	0.217† (↓)	0.708† (↓)	0.252† (↑)	0.716† (↓)	**0.529**† (↑)
Case 1	**0.969**	0.216	**0.965**	0.109	**0.931**	0.190	**0.961**	0.227	0.804	0.080
Case 2	0.873	0.781	0.694	0.353	0.812	0.239	0.675	0.229	0.896	0.494
Case 3	0.869	0.789	0.695	0.331	0.427	**0.236**	0.699	0.229	0.893	0.466

Table 2: F1 score for various methods. † indicates the reproduced results. Bottom three rows represent the followings: **Case 1**. Random anomaly score, **Case 2**. Input itself as a anomaly score, **Case 3**. Anomaly score from a randomized model. Please refer to the manuscript for the detailed explanations. Bold and underlined cases indicate the best and the second best, respectively. ↑ is marked in the following cases: (1) $F1_{PA}$ is higher than **Case 1**, (2) F1 is higher than **Case 2** or **3**.

Comparison results (Con't)

Case 1

Expectation: Case1 does not reflects abnormality in an input $\rightarrow$ not able to detect anomalies.

Result: $F1_{PA}$ seems to perform well. (Overestimation effect of PA)

According to restated R and P, $F1_{PA}$ depends on γ, $t_e - t_s$, δ.

1. The anomaly ratio(γ) of SMD was low. cf. table 1.
2. The anomaly segment length($t_e - t_s$) is short.

   $\rightarrow$ **shorter anomaly segments**, it's less affected.

Comparison results (Con't)

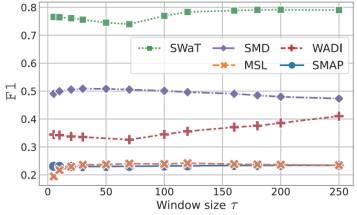
Case 2 and 3

Depends on the length of the input window.

1. longer window, the F1 baseline becomes even larger.

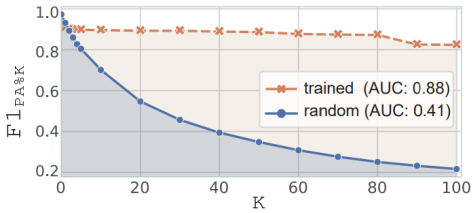
   A longer winow is more likely to contain more point anomalies.

2. τ becomes too large, F1 saturated or degrades because the window that used to contain only normal signals unexpectedly contain anomalies in it.



Effect of PA%K protocol

If K=0, it is equal to $F1_{PA}$, K=100, it is equal to F1.



Setting: SWaT dataset from Case1

Result: Well-trained model show constant result regardless of value K. $F1_{PA\%K}$ of Case1(blue) rapidly decreased when K increased. It demonstrates that PA%K distinguished the formal from the latter regardless of K.

Summary

Summary

Current evaluation of TAD has pitfalls in two respects

1. PA overestimates the detection performance.
2. Results have been compared only with existing methods not against the baseline.

Suggestion

1. To mitigate overestimation PA, the paper propose $F1_{PA\%K}$.
2. Suggest new baselines(Case2 and 3) and carefully determine the window size.

Further discussion

Outlier detection for multivariate long memory process

Based on Tsay(2000), we detect 4 types of Outlier(AO, IO, LS, TC) by VHAR (Vector heterogeneous autoregressive)(Corsi, 2009).

T: number of observations Y_t: observed time series, $\{\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{kt})'\} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$

$$Y_t^{(d)} = \beta_0 + \Phi^{(d)} Y_{t-1d}^{(d)} + \Phi^{(w)} Y_{t-1d}^{(w)} + \Phi^{(m)} Y_{t-1d}^{(m)} + \epsilon_t, t = 1, 2, ..., T,$$

$$Y_{t-1}^{(w)} = \frac{1}{5} \sum_{j=0}^{4} Y_{t-jd}^{(d)}, Y_{t-1}^{(m)} = \frac{1}{22} \sum_{j=0}^{21} Y_{t-jd}^{(d)}$$

Further discussion(Con't)

$x_t = (x_{1t}, \ldots, x_{kt})'$: k dimensional multivariate time series which follows VHAR process.

$y_t = (y_{1t}, \ldots, y_{kt})'$: Observed time series, $\omega = (\omega_1, \ldots, \omega_k)'$: impact size.

$$\Phi(B)x_t = c + \epsilon_t, t = 1, \ldots, T \text{ where } \Phi(B) = I - \Phi_1 B - \cdots - \Phi_{22} B^{22}.$$

Time series with outlier: $y_t = x_t + \alpha(B)\omega\zeta_t^{(h)}, \zeta_h^{(h)} = \begin{cases} 1, t = h \\ 0, t \neq h. \end{cases} \quad (*)$

Let, $\alpha(B) = \Psi(B)$ for IO, I for AO, $(1 - \delta B)^{-1}$ for TC and $(1 - B)^{-1}$ for LS.

Restates (*) by multiply $\Pi(B)$ and subtract $c_0 \rightarrow a_t = \epsilon_t + \Pi(B)\alpha(B)\omega\zeta_t^{(h)}$.

Further discussion(Con't)

Statistics

1. $J_{i,h} = \hat{w_{i,h}}' \Sigma_{i,h}^{-1} \hat{w_{i,h}}$ where $\Sigma_{i,h} = (\sum_{i=0}^{n-h} \hat{\Pi_i}' \Sigma^{-1} \hat{\Pi_i})^{-1}$
2. $C_{i,h} = max_{1 \le j \le k} \frac{|\hat{w_{j,i,h}}|}{\sqrt{\sigma_{j,i,h}}}$.

Therefore, We need to find $\hat{a_t}$ and $\hat{\omega_{i,h}}$ for each outlier by use the under equation.

$$\hat{a_t} = \epsilon_t + \Pi(B)\alpha(B)w\zeta_t^h, \quad \hat{\omega}_{i,h} = -\Big(\sum_{i=0}^{n-h} \hat{\Pi}_i^* \Sigma^{-1} \hat{P}i^*\Big)^{-1} \sum_{i=0}^{n-h} \hat{\Pi}_i^{*\,T} \Sigma^{-1} \hat{a}_{h+i}$$

$$\Pi(B)\alpha(B) = \Pi(B)^* = I - \sum_{i=1}^{\infty} \Pi_i^* B^i = (I - \sum_{i=1}^{\infty} \Pi_i B^i)(1 - \delta B)^{-1},$$

(AO:δ=0, TC:$0 < \delta < 1$, LS:$\delta = 1$).

Further discussion(Con't)

1. Not using rolling window (window size=1)
2. Not using PA, we empirically get Critical values to find out outliers.
3. Find both outliers, point(AO,IO) and subsequence(LS, TC).

Further discussion(Con't)

VFARIMA, k=3, n=300, iterate=1000

VARFIMA,n=300	50%	90%	95%	97.5%	99%
$J_{max}(I, h_I)$	14.32	18.13	19.53	20.65	22.27
$J_{max}(A, h_A)$	14.44	18.29	19.56	20.89	22.56
$J_{max}(L, h_L)$	7.62	11.90	13.29	15.21	17.16
$J_{max}(T, h_T)$	13.64	17.60	19.23	20.82	21.95

VARFIMA,n=300	50%	90%	95%	97.5%	99%
$C_{max}(I, h_I^*)$	3.36	3.83	4.01	4.13	4.26
$C_{max}(A, h_A^*)$	3.36	3.80	4.03	4.17	4.30
$C_{max}(L, h_L^*)$	2.37	3.04	3.24	3.44	3.67
$C_{max}(T, h_T^*)$	3.28	3.76	3.91	4.08	4.29

TAD metric

Xu, H.; Chen, W. et al. 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications.

'It is acceptable for an algorithm to trigger an alert for any point in a contiguous anomaly segment, if the delay is not too long. Some metrics for anomaly detection have been proposed to accommodate this preference, e.g., [22], but most are not widely accepted, likely because they are too complicated. We instead use a simple strategy.'

[22] Alexander Lavin and Subutai Ahmad. 2015. Evaluating Real-Time Anomaly Detection Algorithms–The Numenta Anomaly Benchmark. In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, 38– 44.

Thank you