Policy-Based Reinforcement Learning for Time Series Anomaly Detection

Mengran Yu, Shiliang Sun*

School of Computer Science and Technology, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, P.R. China

Abstract

Time series anomaly detection has become a crucial and challenging task driven by the rapid increase of streaming data with the arrival of the Internet of Things. Existing methods are either domain-specific or require strong assumptions that cannot be met in realistic datasets. Reinforcement learning (RL), as an incremental self-learning approach, could avoid the two issues well. However, the current investigation is far from comprehensive. In this paper, we propose a generic policy-based RL framework to address the time series anomaly detection problem. The policy-based time series anomaly detector (PTAD) is progressively learned from the interactions with time-series data in the absence of constraints. Experimental results show that it outperforms the value-based temporal anomaly detector and other state-of-the-art detection methods whether training and test datasets come from the same source or not. Furthermore, the tradeoff between precision and recall is well respected by the PTAD, which is beneficial to fulfill various industrial requirements.

Keywords: Time Series Anomaly Detection; Reinforcement Learning; Policy-Based Methods

*Corresponding Author.

Email addresses: mengranyu97@gmail.com (Mengran Yu), shiliangsun@gmail.com (Shiliang Sun)

1. Introduction

Anomaly detection has been a hot spot since critical systems need to detect anomalies as early as possible to avoid big troubles [1]. An anomaly is considered as a point or a change that differs from other data. With the arrival of the Internet of Things, innumerable sensors are installed to collect amounts of data which change over time and are called time-series data. Since there exist complicated abnormal patterns which may be spatial or temporal depending on whether they are contextual or not in time-series data and this type of data often occur with a periodic or seasonal mode, it is a challenging issue to detect anomalies precisely in these data.

Time series anomaly detection has been investigated with numerous applications, including intrusion detection, credit card fraud, and medical diagnoses. Currently, not only the supervised learning methods [2, 3] but also the semi-supervised learning and unsupervised learning approaches [4, 5, 6] have been proposed. Existing approaches are generally designed for specific applications and particular datasets [7, 8, 9, 10, 11, 12]. A specialized anomaly detector is often applied to only one use-case. It is troublesome to construct an anomaly detector for multiple use-cases. There are also a few generic methods which achieve great performance on time-series benchmark datasets. Unfortunately, these approaches require especially analyzing the characteristic of the full data or making strong assumptions, e.g., the training data is anomaly-free which is yet unrealistic in most real datasets [13, 14].

Reinforcement learning (RL) [15] conforms to the learning process of human beings. The agent receives positive feedback when useful information is learned and acquires a negative signal when learning useless or harmful information to instruct the following behaviors. It follows the incremental self-learning process that the agent autonomously learns a generic framework from interactions with the environment without any assumption and constraint. Hence, it provides a novel way of solving the anomaly detection problem. Due to the consecutive characteristic of the time-series data, it is natural to adapt the time series anomaly detection process into the framework of RL.

However, explorations of this field are extremely insufficient in the current literature. Huang et al. [16] attempted a value-based deep reinforcement learning (DRL) time series anomaly detector which adopted the Deep Q-Function Network (DQN) algorithm [17], however, just with a brief instruc-

tion. Their experimental results demonstrated the possibilities of detecting abnormal behaviors with RL methods. As the other type of methods, policy-based DRL algorithms, also show excellent performance in sequential prediction and robot control since they directly operate in the policy space [18, 19, 20]. Hence, we wonder how well the policy-based DRL detection methods perform when compared with the value-based DRL detection methods and whether RL is an effective solution to the time series anomaly detection problem.

This paper adapts the policy-based RL framework into the time series anomaly detection problem and proposes a general anomaly detector. Experimental results demonstrate the effectiveness of the proposed detector on homologous and heterologous datasets when compared with the value-based RL time series anomaly detector and other advanced detection methods. The contributions of this paper are listed as follows. Firstly, we propose a novel policy-based DRL time series anomaly detector (PTAD) based on the asynchronous advantage actor-critic (A3C) algorithm [21], which is one of the most advanced DRL algorithms to address the anomaly detection problem. Secondly, compared with the value-based DRL detector and other state-of-the-art anomaly detection techniques, the proposed PTAD performs superiorly whether on the same or the different source and target datasets. Furthermore, the optimal stochastic detection policy acquired from the PTAD allows adjusting the criterion of distinguishing normal and abnormal behaviors, which controls the tradeoff between precision and recall to meet some particular demands in various applications, i.e., avoiding incorrect anomaly detections which result in unnecessary interruptions.

The remainder of this paper is organized as follows. Section 2 describes various anomaly detection strategies and algorithms. The basic concepts of RL and the formalization of RL based time series anomaly detection problem are illustrated in Section 3. Section 4 introduces the proposed PTAD in detail and compares the traits between the value-based and policy-based DRL detection approaches. Descriptions of datasets, experimental settings and results are shown in Section 5. Section 6 concludes the work of this paper and sketches directions for possible future work.

2. Related Work

There are numerous algorithms and strategies about anomaly detection over the recent years, which could be roughly classified into two categoriza-

tions: statistical based methods and machine learning based approaches.

Statistical based methods construct a statistical model from the given data and operate a statistical inference test to justify whether new data fit the model. Non-parametric models are generally histogram based [22] and kernel function based [23], which learn the underlying distribution of normal behaviors from the given data directly. Gaussian model [24], regression model [25], mixture model of parametric distributions [26, 27] are classical parametric statistical models, which instead assume that the underlying distribution of normal data matches the presupposed distribution. However, these methods depend on the assumption that normal behaviors suit the predefined distribution, which is not often true in realistic datasets.

Machine learning based approaches learn a model from the labeled training data and distinguish new data between normal class and abnormal class with the model. There are two ways to learn the model where one is classification and the other is clustering. Bayesian networks [28], support vector machines [29], rule based [30] and neural networks [31] are common machine learning algorithms to build the anomaly detection classifiers. Clustering anomaly detections are mainly based on k-nearest-neighbours algorithm [32], which is expanded by local outlier factor (LOF) [33] and connectivity based outlier factor (COF) algorithm [34]. Because new anomalous behaviors might break out in nature and relate to the contextual information, machine learning based approaches are suitable for domain-specific applications where informative training data are available.

Due to the complexity of time-series data, e.g., the pattern of data is continuously changing and temporal dependencies are contained in them, some specific techniques and algorithms are explored. Skyline [35] is a real-time anomaly detection system developed by Esty Inc in 2014. Twitter Inc. released its package to detect anomalies which is robust, from a statistical standpoint, in the presence of seasonality and an underlying trend [36]. ContextOSE [37] is based on contextual anomaly detection, which captures the local rather than global information. Numenta and Numenta TM [4] are expanded from the hierarchical temporal memory (HTM), which is a detailed computational theory of the neocortex and the core is storing and recalling spatial and temporal patterns. With the development of deep learning, RNN or LSTM based time series anomaly detectors were proposed [38] where they learn a predictive model from the normal training time stamps and mark normal or abnormal depending on the error between the predictive values and true values. There are also some variants based on the autoencoder [39, 40].

Recently, RL is taken into consideration for solving the time series anomaly detection problem because of its generic framework and incremental self-learning property. Bourdonnaye et al. [41] learned binocular fixations with informative reward requiring little supervised information where the reward computation was based on an anomaly detection mechanism which used convolutional autoencoders. They just regarded anomaly detection as an auxiliary technique for generating the feedback signals. Huang et al. [16] attempted a value-based DRL time series anomaly detector with the DQN algorithm, which built a bridge between RL and anomaly detection. However, the acquired deterministic policy was unsatisfactory to dynamically modulate the threshold of justifying an anomaly for several requirements in different applications. Therefore, we propose the PTAD based on the A3C algorithm, which is not only dynamically adjustable for controlling the trade-off between precision and recall in various circumstances but also improves the F1 scores of detecting abnormal behaviors.

3. Preliminaries

Time series anomaly detection can be modeled as a sequential decision process, which is formulated to the Markov decision process (MDP) in RL. Hence, the MDP bridges the gap between time series anomaly detection and RL. The significant feature of RL is that the agent interacts with the environment. That is, an action taken by the agent at the current time step t will affect the state at time step $t+1$. Then the following actions will be influenced by the reward received from the environment. In this section, we firstly introduce basic conceptions in RL and then illustrate the formalization of RL based time series anomaly detection problem in detail.

3.1. Background: Reinforcement Learning

A reinforcement learning problem is usually represented as an MDP whose specific form is a tuple of five elements: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$. At time step t, assume that the agent is at the state $s_t \in \mathcal{S}$ and selects the action $a_t \in \mathcal{A}$ according to the policy π, where π is a mapping from a state s_t to an action a_t. The agent will receive an immediate reward r_t where $r_t = R(s_t, a_t)$ and obtain the next state s_{t+1} according to the state transition probabilities function $P(s_{t+1}|S = s_t, A = a_t)$. These interactions with environment $\mathcal{E}$ come into being a trajectory τ until the agent reaches a terminal state. The goal of the agent is to maximize the expected return $\mathbb{E}[R_t]$ from each state s_t

where the return is $R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$. γ represents the discount factor which ranges from 0 to 1 and measures the importance of current rewards on future rewards. The agent finally obtains an optimal policy π^* via learning from experiences.

There are two methods to evaluate the policy π which are the state-value function $V(s)$ and the action-value function $Q(s, a)$.

$$V^{\pi}(s) = \mathbb{E}[R_t | s_t = s], \tag{1}$$

$$Q^{\pi}(s, a) = \mathbb{E}[R_t | s_t = s, a_t = a]. \tag{2}$$

By these definitions, the state-value function $V(s)$ is the expected return under the state s and the action-value function $Q(s, a)$ is the expected return for selecting a specific action a under the state s. The optimal state-value function and action-value function are defined as

$$V^*(s) = \max_{\pi} V^{\pi}(s), \tag{3}$$

$$Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a). \tag{4}$$

In the value-based RL methods, the agent optimizes the target policy indirectly by maximizing the corresponding value function where the action-value function $Q(s, a)$ is usually chosen. We now consider the off-policy value-based learning that means the behavior policy is for sampling and the target policy is for optimizing, whose typical algorithm is called Q-learning. The target policy π is greedy, that is,

$$\pi(s_{t+1}) = \operatorname{argmax}_{a'} Q(s_{t+1}, a'). \tag{5}$$

The behavior policy μ is ϵ-greedy where the agent chooses the greedy action with probability $1 - \epsilon$ and randomly chooses an action with probability ϵ, that is,

$$\mu(s_{t+1}) = \begin{cases} \frac{\epsilon}{|\mathcal{A}|} + 1 - \epsilon & \text{if } a^* = \operatorname{argmax}_{a \in \mathcal{A}} Q(s, a) \\ \frac{\epsilon}{|\mathcal{A}|} & \text{otherwise} \end{cases}, \tag{6}$$

where $|\mathcal{A}|$ is the cardinality of the action space. The iterative formula in training process is given as

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right). \tag{7}$$

In the policy-based RL methods, the agent directly parameterizes and optimizes the target policy $\pi(a|s; \theta)$ with the parameters θ. Compared with the

value-based algorithms, the policy-based approaches are effective for high-dimensional or continuous action spaces and can learn stochastic policies which are more practical than deterministic policies. The most famous framework is the actor-critic where an actor network $\pi(a|s;\theta)$ with parameters θ estimates the target policy π to take an action under a specific state and a critic network $V(s;\upsilon)$ with parameters υ approximates the state-value function $V(s)$ to evaluate the current policy. The critic uses the least-squares policy evaluation where the minimum loss function is written as

$$L(\upsilon) = \mathbb{E}\left[\left(r + \gamma V^{\pi_\theta}(s';\upsilon) - V^{\pi_\theta}(s;\upsilon)\right)^2\right]. \tag{8}$$

The actor updates policy parameters θ with the policy gradient theorem which instructs to iterate in directions of suggestions of the critic. Furthermore, it is beneficial to consider the entropy of the policy for improving exploration. The final maximum loss function including the entropy regularization term is shown as below

$$L(\theta) = \mathbb{E}\left[\log\pi(a|s;\theta)\delta_v + \beta H(\pi(a|s;\theta))\right], \tag{9}$$

where $\delta_v = r+\gamma V^{\pi_\theta}(s';\upsilon)-V^{\pi_\theta}(s;\upsilon)$ is the advantage function, $H(\pi(a|s;\theta)) = -\pi(a|s;\theta)\log\pi(a|s;\theta)$ is the entropy of the policy $\pi(a|s;\theta)$.

3.2. Formalization of RL Based Time Series Anomaly Detection Problem

Time series anomaly detection could be considered as an MDP because the decision of normal or abnormal at the current time step will change the environment by whether it triggers an anomaly detection or not. And the next decision will be influenced by the changing environment. Hence, it is natural to adapt the temporal anomaly detection into the framework of RL [16]. Next, we instruct the concrete time series anomaly detection formulation.

State Since the next action taken by the agent is affected by the changing environment which is comprised of the previous decisions and the current time series, the state includes two parts where one is the sequence of the previous actions, i.e., $s_{action} = \langle a_{t-m+1}, a_{t-m+2}, \ldots, a_t\rangle$ and the other is the current time series, i.e., $s_{time} = \langle s_{t-m+1}, s_{t-m+2}, \ldots, s_t\rangle$. We want to know the action a_{t+1} with the previous m actions and m time stamps. The state space $\mathcal{S}$ is regarded as infinite because the real time series have a variety of alterations.

Action It is simple to define the action space $\mathcal{A} = \{0, 1\}$ where 0 represents the normal behavior and 1 means an anomaly is detected.

Reward Designing a proper reward function is important for the agent to learn an effective policy. There are several types of temporal data, such as the labeled, the semi-labeled and the unlabeled, which correspond to supervised learning, semi-supervised learning and unsupervised learning separately in conventional machine learning. The RL agent needs relatively correct instructions for learning an effective detection policy. Therefore, we consider the labeled training data and construct the reward function with labels.

We utilize the confusion matrix of the prediction problem in traditional machine learning to design the reward function $R(s, a)$. The confusion matrix is shown as Table 1 where the positive means detecting an anomaly and the negative represents normal behaviors. The reward function is designed below where A, B, C, and D are the positive number and can be set different values according to the characteristic of the training time series data and realistic demands. For example, if a wrong anomaly detection is forbidden in some real applications, we can fix the B slightly larger to give apparent negative feedback.

$$R(s, a) = \begin{cases} A & \text{if the action is a TP} \\ -B & \text{if the action is a FP} \\ -C & \text{if the action is a FN} \\ D & \text{if the action is a TN} \end{cases} \tag{10}$$

Table 1: Confusion matrix.

Prediction /True Value	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Policy The target policy is represented by the anomaly detector which has two forms. One is for the deterministic policy acquired from the value-based detector. It gives the explicit action under the current state and is written as

$$\pi(s) = a, \quad s \in \mathcal{S}, a \in \mathcal{A}. \tag{11}$$

The other is for the stochastic policy obtained from the policy-based detector. It provides the probability of each action under the present state that

means the criterion of determining an action is adjustable. It is formulated as

$$\pi(s, a) = p(a|s), \quad s \in \mathcal{S}, a \in \mathcal{A}. \tag{12}$$

Value Function The state-value function of a detection policy is the same as the standard form which we specialize into a particular formula as follow

$$V_\pi = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | s_t = s \right]. \tag{13}$$

And the action-value function is defined as

$$Q_\pi = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | s_t = s, a_t = a \right]. \tag{14}$$

Optimal policy The optimal policy is our ideal anomaly detector maximizing the expected reward whose forms are following:

$$\pi^* = \text{agrmax}_\pi V_\pi \quad or \quad \pi^* = \text{agrmax}_\pi Q_\pi. \tag{15}$$

4. Policy-Based Time Series Anomaly Detector (PTAD)

In this section, we present a policy-based DRL time series anomaly detector, called PTAD, for discovering abnormal behaviors in time-series data. Furthermore, we compare the traits between the policy-based and value-based DRL time series anomaly detectors.

For the RL based time series anomaly detection setting, the environment is a time series repository which contains a large population of labeled time-series data. With these data, the environment is able to generate specific states for training the agent and determine the goodness of the actions taken by the agent. Another essential component of the setting is the agent that simulates how the time series anomaly detector operates and optimizes. It takes the current n time stamps and previous n decisions as input and outputs a new decision for the next time stamp. Since the whole optimal process is not related to any assumptions and constraints, the agent could be applied in similar time series anomaly detection tasks.

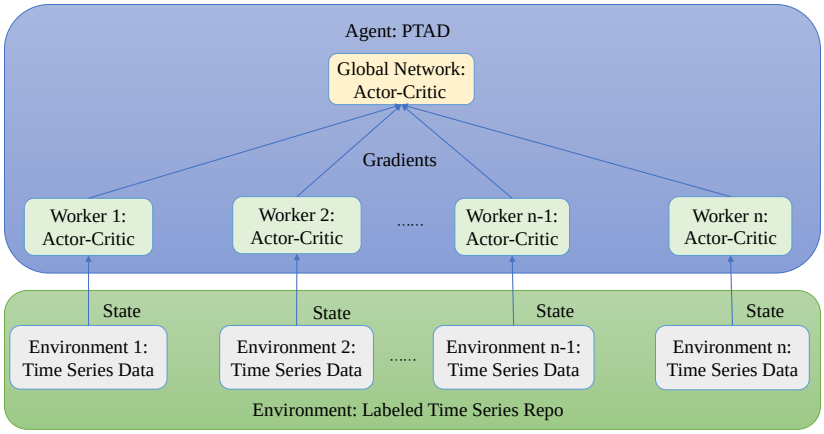


Figure 1: The asynchronous interactions between the PTAD (the agent) and the labeled time series repository (the environment).

4.1. The Proposed Approach

We construct the PTAD with the A3C algorithm, which adopts an asynchronous mechanism for decreasing correlations between the successive examples. Figure 1 illustrates the overall asynchronous interactions between the PTAD (the agent) and the labeled time series repository (the environment). The below box marks the environment. There are n independent environments which contain the whole labeled time-series data but rank these sequences inconsistently. Each environment provides time stamps of distinct time series as states for its worker and change itself by the received actions. The upper box indicates the PTAD which possesses a global network and n local network, also called workers. All networks take the actor-critic framework. Every worker explores an individual environment and calculates the gradients with the rewards sent by its corresponding environment. In our experiment, every agent owns a different initial environment that could improve the anomaly detection performance because different agents learn from different time series at the same time in order to avoid overfitting some specific abnormal patterns. The global network collects the gradients from the workers and optimizes the targeted policy.

Figure 2 shows the internal structure of the PTAD, which maintains three

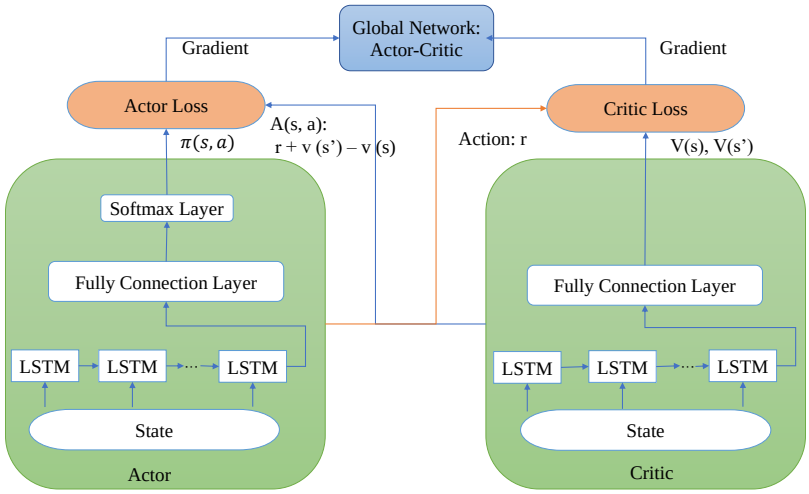


Figure 2: The internal structure of the PTAD.

major components. A Recurrent Neural Network (RNN) implemented by the Long-Short Term Memory (LSTM) is used to extract the sequential information within the state and output the encoded features to the next unit. The inputs of states are processed similarly by the actor network which estimates the policy and the critic network which approximates the state-value function. Taking the outputs from the RNN as inputs, the fully connection layer yields two values, i.e., $P(a = 0|s)$ and $P(a = 0|s)$, indicating the possibilities of two actions in the actor network and outputs a value of the current state in the critic network. There exists a softmax layer to normalize the outputs of the fully connection layer into the range $[0, 1]$ and give the final decision of the action for the current states by probabilities in the actor network.

These workers do not update the target policy but collect samples to compute gradients. The losses and the gradients in the actor network are calculated with the policy gradient theorem that is involved with the policy π and the advantage function $A(s, a)$ given by the critic network. The losses are the differences between $r + v(s')$ and $v(s)$ where r is obtained after the action taken by the actor network under the given state s in the critic network. Once the global network updates its parameters, it should pass them to local networks to keep consistent.

4.2. Trait Comparison

Compared with the value-based DRL anomaly detector, the proposed PTAD has several potential advantages. The value-based detector generates a deterministic policy which takes each action under specific states unchangeably. However, the PTAD could modulate the threshold of judging whether the current state is an anomaly to alter the decision because it yields a stochastic policy. It operates an advantageous tradeoff between precision and recall for some certain demands, e.g., assuring higher precision in cloud operation anomaly detection to relieve heavy work burdens of operation engineers or greater recall is essential in anomaly detection for healthcare because missing anomalies may lead to increases in health risks of patients.

Furthermore, our PTAD is available for the anomaly detection task which needs to output the degrees of anomalies, e.g., ranging from 0 to 1. It means that the action space is continuous. However, the value-based temporal anomaly detector fails because it is troublesome to enumerate values of all actions and select the maximum during each update.

5. Experiments

In this section, we will demonstrate the effectiveness of PTAD by performing experiments on two classical time series datasets compared to the state-of-the-art anomaly detection methods.

5.1. Datasets

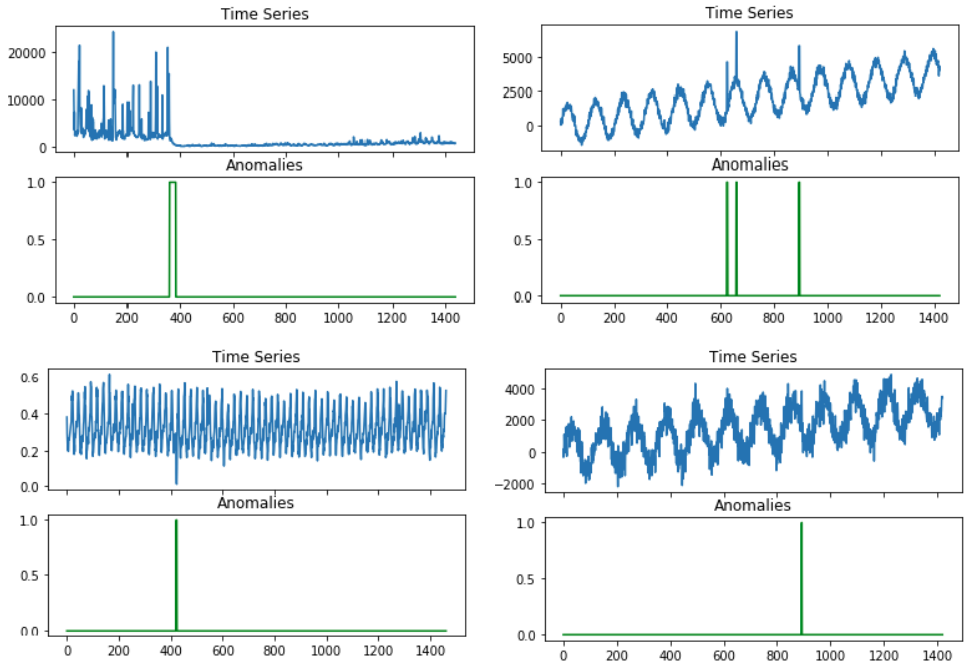
We experiment on two classical temporal anomaly detection datasets which are Yahoo benchmark dataset and Numenta Anomaly Detection (NAB) dataset.

Yahoo Benchmark dataset [1] The dataset consists of real and synthetic tagged anomaly time series and contains four subsets, called A1, A2, A3, and A4. In our experiment, Yahoo A1 and A2 benchmark datasets are selected for testing the capability of different anomaly detectors on the same source and target datasets. There are 67 time series and 100 time series in A1 and A2, respectively. A1 benchmark dataset contains real Yahoo membership login data and has complex temporal patterns. Each time series has various anomaly types, even has different lengths. Compared with the

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=a

A1 benchmark dataset, it is relatively easy to detect the anomalies because there are almost point anomalies and all time series have the same lengths in the A2 benchmark dataset.

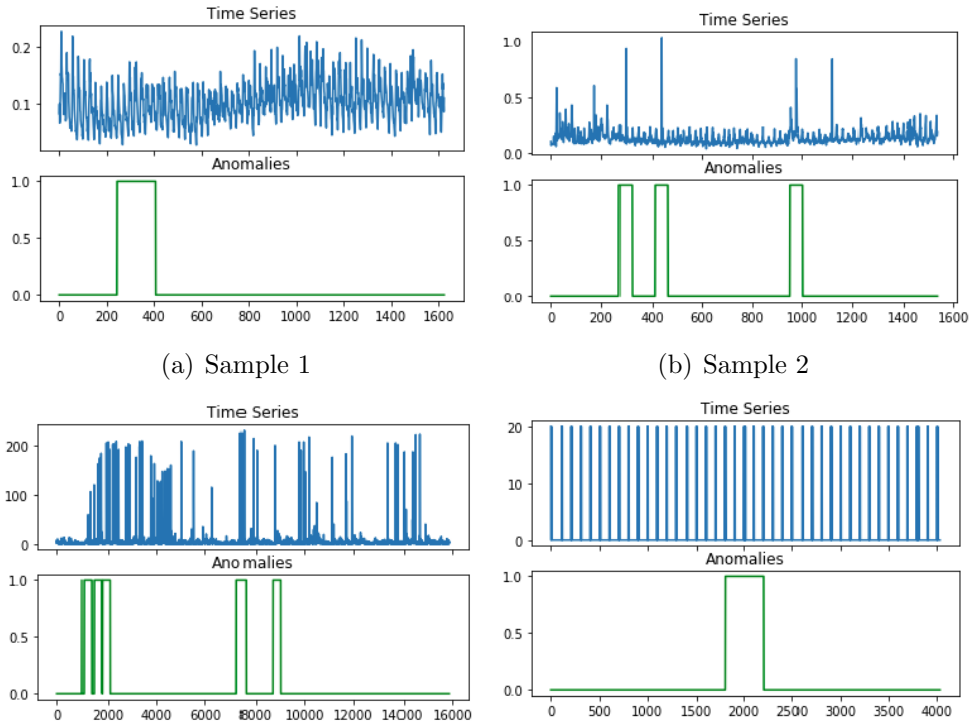
Figure 3 shows some examples of the two datasets. The blue lines mean the original time series data and the green lines indicate the current state where 0 represents normal and 1 marks abnormal. Figure 3(a) indicates that the A1 Benchmark time series data have no obvious anomaly behaviors through artificial recognition and relatively Figure 3(b) shows the distinct anomaly patterns even though there exist some disturbances on the A2 benchmark dataset. When comparing the performance of various anomaly detectors on different source and target datasets, we select the whole Yahoo benchmark dataset which comprises 367 time series as the training set.



(a) A1 samples (b) A2 samples

Figure 3: Yahoo benchmark samples. The blue lines mean the original time series and the green lines indicate the current state where 0 represents normal and 1 marks abnormal.

NAB dataset [2] NAB dataset is usually used for evaluating anomaly detection algorithms in streaming, real-time applications. It includes 58 labeled real-world or synthetic time series, each with 1000 - 22000 time stamps. For example, realTraffic subset contains real-time traffic data from the Twin Cities Metro area in Minnesota and artificialNoAnomaly subset is artificial without anomalies. The anomaly patterns are also complicated and there is no single anomaly detection method that can distinguish all anomalies. We draw some representative examples as Figure 4. It shows that some of these time series are periodic and whether a sharp increase is an anomaly or not is related to the contextual information.



(a) Sample 1 (b) Sample 2

(c) Sample 3 (d) Sample 4

Figure 4: Numenta benchmark samples. The anomaly patterns are difficult to grasp and whether a sharp increase is an anomaly or not is related to the contextual information.

[2] https://github.com/numenta/NAB

5.2. Evaluation Metric

We employ the widely used F_1 scores to measure the quality of different anomaly detection models. The metric F_1 is defined as follows

$$F_1 = \frac{2 * precision * recall}{precision + recall}, \tag{16}$$

where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$. From the definition, we can see that a larger F_1 scores indicates a better performance.

5.3. Comparing methods and Experimental Setups

We consider the following methods to compare:

Skyline [35]: This method calculates time series anomaly scores by voting from different expert detectors.

Twitter [36]: It is based on seasonal hybrid extreme studentized deviate algorithm and performs excellent in seasonal univariate time series.

ContextOSE [37]: It is based on contextual anomaly detection, which captures the local rather than global information. The procedure is selecting a subset of time series, calculating the centroid of the selected time series and then predicting the value of time series with the centroid and other features.

Numenta and Numenta TM [4]: These detection methods are based on Hierarchical Temporal Memory (HTM). At the core of HTM are time-based continuous learning algorithms that store and recall spatial and temporal patterns.

RNN-TAD [38]: This method is a neural network based anomaly detector. It learns a predictive model from the normal training time stamps based on the LSTM and marks normal or abnormal depending on the error between the predictive values and true values.

VTAD [16]: It is the value-based DRL time series anomaly detector which adopts the DQN algorithm.

The model parameters of these methods are set as consistent as the references. Notably, our implement of the VTAD is a standard Q-learning process with the scalar reward, which is slightly different from what Huang et al. [16] do where they calculate the loss with all actions

$$\mathbb{E}_{a\in\mathcal{A},batch}\left[\left(r + \gamma \text{max}_{a'} Q(s', a'|\theta_{i-1}) - Q(s, a|\theta_i)\right)^2\right],$$

not the action a^* selected by the target policy

$$\mathbb{E}_{batch}\left[\left(r + \gamma \text{max}_{a'} Q(s', a'|\theta_{i-1}) - Q(s, a^*|\theta_i)\right)^2\right],$$

because they adopt a vector reward function.

The PTAD is trained with a multi-core CPU of 8 threads without the GPU. The local network delivers the gradients to the global network every 5 steps and the learning rates of actor network and critic network are 0.001 and 0.0001, respectively. The total number of training episodes is 20000. The parameters in reward function is set as $A = C = 5, B = D = 1$.

5.4. Comparison Results

The comparisons among different time series anomaly detectors are not only on the same but also on different training and test datasets.

5.4.1. Performance of the Same Source and Target Datasets

We compare the performance of Twitter anomaly detection, RNN-TAD, VTAD and PTAD on the test part of the Yahoo A1 and A2 benchmark datasets. For the Twitter anomaly detector, it just analyzes statistical characteristics on each time series without training. For the other three detectors, all data in each benchmark dataset are originally divided into training and test parts by a ratio of 8:2. Hence, there are 13 test time series in the A1 benchmark dataset and 20 test time series in the A2 benchmark dataset. We eliminate abnormal time stamps in training set for the RNN-TAD to construct a normal predictor.

Table 2 shows the comparison results of test examples on the Yahoo A1 benchmark dataset. Although the PTAD is not the best in every test time series, it outperforms other methods averagely. The minimal standard variation is also achieved by the proposed PTAD. At the same time, the detection results of Twitter anomaly detector, RNN-TAD, and VTAD contain only 6, 7, 7 over 0.5 F_1 scores in the 13 test time series, which illustrates that they can not generalize well in the A1 benchmark detection task. The PTAD almost gets F_1 scores over 0.5 except the 10th and 11th examples, which shows more stable anomaly detection performance. Furthermore, the two RL based detectors achieve better results compared with Twitter anomaly detector and RNN-TAD, which illustrates that RL is an effective tool for the time series anomaly detection problem.

Table 3 shows the comparison results of test examples on the Yahoo A2 benchmark dataset. Since the data in Yahoo A2 benchmark dataset have relatively simple anomaly patterns which almost are the point anomalies, the RNN-TAD, VTAD, and PTAD get perfect results on all test time series.

Table 2: Performance comparisons among Twitter anomaly detection, RNN-TAD, VTAD and PTAD on the test part of the Yahoo A1 benchmark datasets. The bold indicates the best.

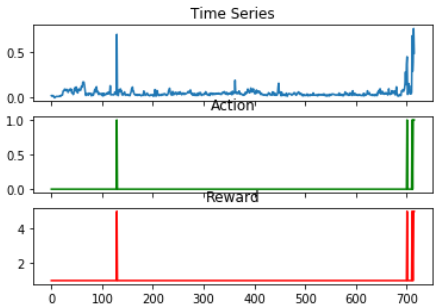
Index of Examples	Twitter	RNN-TAD	VTAD	PTAD
1	0.07	**1.00**	0.50	0.67
2	0.80	0.93	**0.95**	0.86
3	0.02	0.56	**0.81**	0.56
4	0.50	**0.95**	0.81	0.79
5	**0.63**	0.40	0.50	0.50
6	0.55	0.49	**0.80**	0.70
7	0.04	0.09	0.08	**0.63**
8	0.35	0.73	**1.00**	**1.00**
9	**0.62**	0.35	0.35	0.60
10	**0.25**	0.00	0.08	0.06
11	**0.34**	0.29	0.21	0.29
12	0.67	0.78	**0.90**	**0.90**
13	0.79	0.70	**1.00**	**1.00**
Arithmetic Mean	0.43	0.56	0.61	**0.66**
Standard Variation	0.27	0.31	0.33	**0.26**
Numbers of $F_1 > 0.5$	6	7	7	**11**

However, the Twitter anomaly detector receives a poor result in this detecting task.

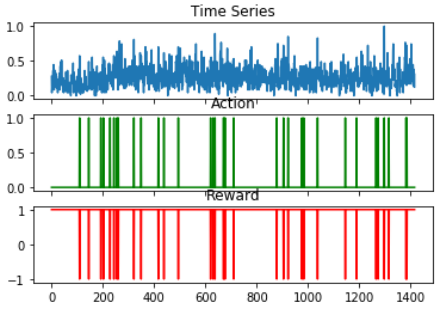
Table 3: Performance comparisons among Twitter anomaly detector, RNN-TAD, VTAD and PTAD on the test part of the Yahoo A2 benchmark datasets. The bold indicates the best.

Index of Examples	Twitter	RNN-TAD	VTAD	PTAD
1	0.33	**1.00**	**1.00**	**1.00**
2	0.67	**1.00**	**1.00**	**1.00**
3	0.33	**1.00**	**1.00**	**1.00**
4	0.70	**1.00**	**1.00**	**1.00**
5	0.67	**1.00**	**1.00**	**1.00**
6	0.33	**1.00**	**1.00**	**1.00**
7	0.18	**1.00**	**1.00**	**1.00**
8	0.67	**1.00**	**1.00**	**1.00**
9	0.33	**1.00**	**1.00**	**1.00**
10	0.70	**1.00**	**1.00**	**1.00**
11	0.67	**1.00**	**1.00**	**1.00**
12	0.60	**1.00**	**1.00**	**1.00**
13	0.18	**1.00**	**1.00**	**1.00**
14	0.50	**1.00**	**1.00**	**1.00**
15	0.33	**1.00**	**1.00**	**1.00**
16	0.59	**1.00**	**1.00**	**1.00**
17	0.67	**1.00**	**1.00**	**1.00**
18	0.60	**1.00**	**1.00**	**1.00**
19	0.18	**1.00**	**1.00**	**1.00**
20	0.67	**1.00**	**1.00**	**1.00**
Arithmetic Mean	0.50	**1.00**	**1.00**	**1.00**
Standard Variation	0.19	**0.00**	**0.00**	**0.00**
Numbers of $F_1 > 0.5$	12	**20**	**20**	**20**

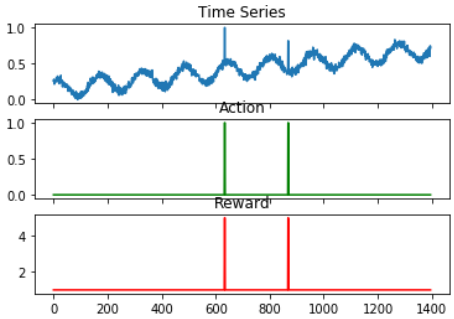
Figure 5 shows several detection results of PTAD in Yahoo A1 and A2 benchmark datasets. For each subfigure, the first row is the raw time series data which are preprocessed with the min-max normalization and marked with the blue line. The green line in the middle row indicates the decisions of the PTAD and the red line in the last row evaluates judgments via rewards. The reward is 5 if the detector correctly distinguishes an anomaly, otherwise



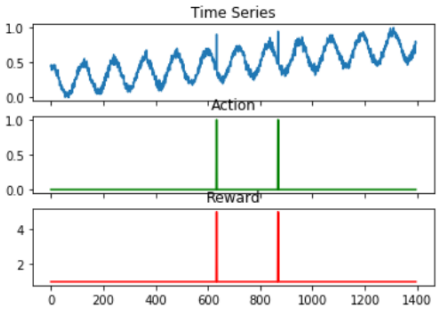
(a) Satisfactory result in A1 dataset



(b) Unsatisfactory result in A1 dataset



(c) Satisfactory result in A2 dataset



(d) Satisfactory result in A2 dataset

Figure 5: Detection results of PTAD in the Yahoo A1 and A2 benchmark datasets. The raw time series are marked as blue lines and the green lines mean decisions of the PTAD and the red lines evaluate judgments via rewards. The reward is 5 if the detector correctly distinguishes an anomaly, otherwise -1. The reward is 1 if the detector correctly judges a normal, otherwise -5.

-1. The reward is 1 if the detector correctly judges a normal, otherwise -5.

For Figure 5(a), 5(c), 5(d), they show satisfactory results in Yahoo A1 and A2 datasets where the PTAD precisely finds out all the anomalies. There are also unsatisfactory results shown in Figure 5(b). Since the raw time series seems too complicated and has no obvious anomaly patterns, it is troublesome to justify whether a sharp increase is an anomaly.

5.4.2. Performance of Different Source and Target Datasets

We present comparison results of the Twitter, Skyline, Numenta, Numenta TM, contextOSE, RNN-TAD, VTAD and PTAD on source the Yahoo benchmark dataset and target the NAB dataset. For Twitter, Skyline, Numenta, Numenta TM, and contextOSE anomaly detector, they predict

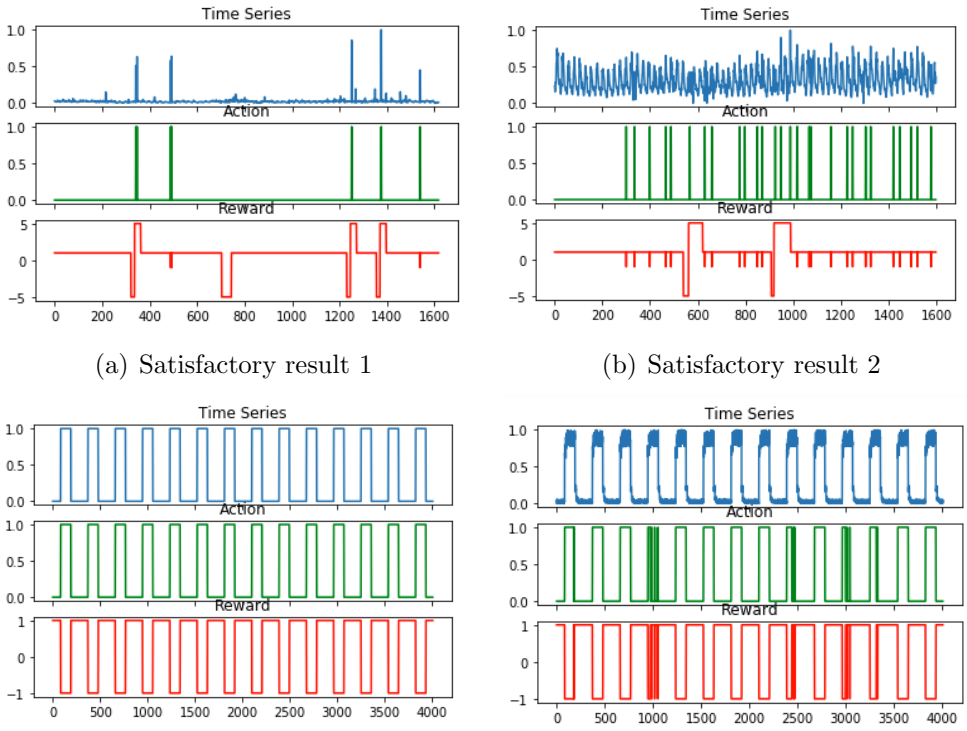
Table 4: Performance comparisons among Twitter, Skyline, Numenta, Numenta TM, contextOSE, RNN-TAD, VTAD and PTAD where the last three methods are trained on Yahoo Benchmark dataset and are tested on Numenta dataset. The bold indicates the best.

Name of Subsets	Twitter	Skyline	Numenta	Numenta TM	contextOSE	RNN-TAD	VTAD	PTAD
artificialNoAnomaly	0.72	**1.00**	0.80	**1.00**	0.87	**1.00**	0.21	0.20
artificialWithAnomaly	0.00	0.05	0.02	0.02	0.01	0.19	**0.73**	0.59
realAdExchange	0.01	0.02	0.05	0.05	0.03	0.16	**0.74**	0.71
realAWSCloudwatch	0.06	0.12	0.05	0.03	0.04	0.24	0.59	**0.60**
realKnownCause	0.02	0.01	0.02	0.02	0.01	0.25	0.33	**0.56**
realTraffic	0.03	0.10	0.04	0.05	0.03	0.25	0.71	**0.73**
realTweets	0.00	0.04	0.01	0.01	0.01	0.09	0.62	**0.67**
Arithmetic Mean	0.12	0.19	0.14	0.17	0.14	0.31	0.56	**0.58**
Standard Variation	0.25	0.33	0.27	0.34	0.30	0.29	**0.19**	**0.19**
Numbers of $F_1 > 0.5$	1	1	1	1	1	1	5	**6**

whether next time stamp is abnormal by modeling the current sequences in a given time series, which means that there is no need to train with the Yahoo dataset. For RNN-TAD, VTAD and PTAD, the whole Yahoo benchmark dataset is used for training and similarly, we remove the abnormal behaviors for the RNN-VTAD to learn a normal predictor. Table 4 shows the average detection results of seven subsets and the best performance is marked as the bold.

PTAD outperforms other detectors including the VTAD in most subsets of NAB datasets. Compared with these state-of-the-art open-source detection techniques, deep learning based methods including RNN-TAD, VTAD, and PTAD discover time series anomalies more precisely and completely. Other than the artificialNoAnomaly subset, the RL based time series anomaly detectors achieve superior performance than other non-RL methods. For the artificialNoAnomaly subset, the first six techniques, especially Skyline, Numenta TM and RNN-TAD, successfully make a decision that there exists no anomaly, while VTAD and PTAD fail to give a correct judgment. It is probably because these RL based detectors do not learn the periodicity from the whole Yahoo benchmark which lacks periodic data and regard every periodic increase as anomaly mistakenly.

We show some detecting results of PTAD on the NAB datasets as Figure 6. Figure 6(a) and 6(b) indicate satisfactory judgments and Figure 6(c) and 6(d) show unsatisfactory results which come from the artificialNoAnomaly subset where the data are periodic. Figure 6(a) illustrates whether a sharp increase is an anomaly depends on the context and practical information since



(a) Satisfactory result 1

(b) Satisfactory result 2

(c) Unsatisfactory result 1

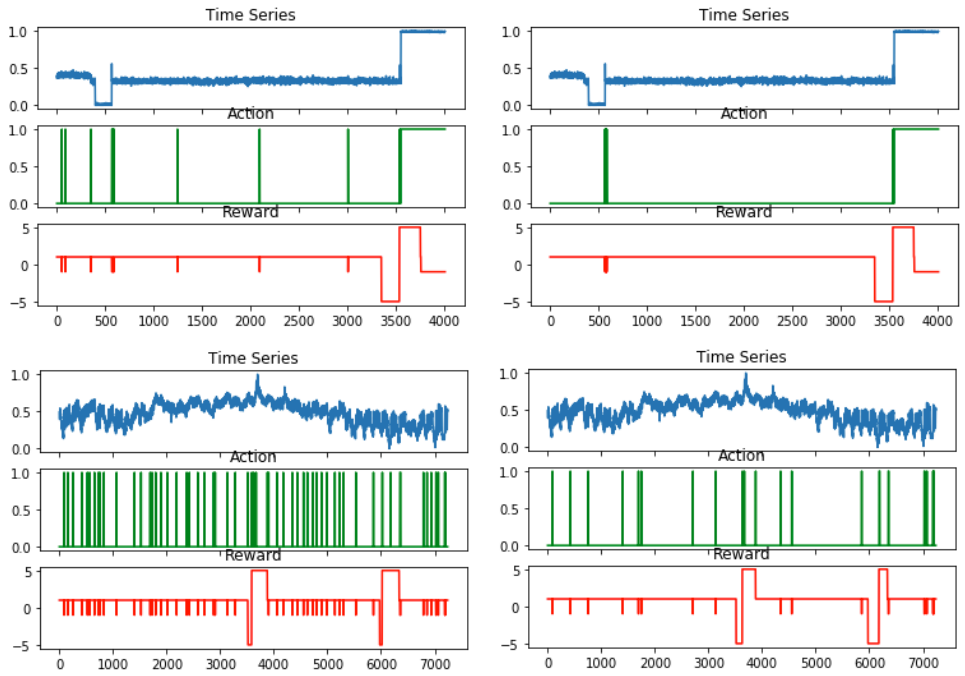
(d) Unsatisfactory result 2

Figure 6: Detection results of PTAD in NAB datasets. The raw time series are marked as blue lines and the green lines mean decisions of the PTAD and the red lines evaluate judgments via rewards. The reward is 5 if the detector correctly distinguishes an anomaly, otherwise -1. The reward is 1 if the detector correctly judges a normal, otherwise -5.

the first rise in 200-400 time stamp is labeled as an anomaly, while the second climb in 400-600 time stamp is not. Figure 6(b) exposes the complicated data patterns and unseen anomalies. The feature of periodic increases is not learned by the PTAD and as Figure 6(c), 6(d) show, the predicting results seem a bit bad. Preprocessing this periodic data to eliminate the periodicity and remain valuable information may be an effective method.

5.5. Adjustability of the PTAD

We also find that adjusting the threshold of determining an anomaly improves the performance slightly by experiment. Figure 7 shows some improved results obtained by the improved detectors where we determine to give a chance of reporting an anomaly when the probability of justifying an



(a) Original Results (b) Improved Results

Figure 7: Improved NAB results with the PTAD. The raw time series are marked as blue lines and the green lines mean decisions of the PTAD and the red lines evaluate judgments via rewards. The reward is 5 if the detector correctly distinguishes an anomaly, otherwise -1. The reward is 1 if the detector correctly judges a normal, otherwise -5. The improved detector eliminates numerous uncertain detections, which significantly raises the precision.

anomaly exceeds 0.9. This means the PTAD has more confidence for the judgment.

Figure 7(a) shows the original detection results and Figure 7(b) illustrates the improved detection results. The original time series are marked as blue lines as well and the green lines mean the decisions of anomaly detectors at any time steps and the red lines evaluate whether the detector gives right judgments or not. It is obvious that the improved detector eliminates numerous uncertain detections, which significantly raises the precision. Conversely, if higher recall is required, the threshold of judging an anomaly can be set lower. Hence, we claim that our PTAD can control the tradeoff between the precision and the recall to meet realistic requirements in different applications. However, the previous VTAD does not possess this merit.

6. Conclusion

We have proposed a general policy-based RL framework to settle the time series anomaly detection problem. Compared with the value-based RL time series anomaly detector and other time series anomaly techniques, our detector achieves the best performance not only on the same but also on different source and target datasets. Furthermore, our detector generates a stochastic policy which slightly improves the detection performance and can explore the tradeoff between the precision and the recall for meeting practical requirements.

As future work, we will investigate how to integrate more information, e.g., periodicity, into the design of anomaly detectors. Furthermore, it is interesting to model a normal RL predictor which outputs the value of the next time stamp in the training phase and gives the final detection results on each test time series according to the error between predictive state and true state. It could reduce the dependency on labels and then extend to the more common semi-supervised and unsupervised settings.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Project 61673179, and Shanghai Knowledge Service Platform Project (No. ZF1213).

References

[1] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Computing Surveys 41 (2009) 15:1–15:58.

[2] S. Chauhan, L. Vig, Anomaly detection in ECG time signals via deep long short-term memory networks, in: IEEE International Conference on Data Science and Advanced Analytics, 2015, pp. 1–7.

[3] A. R. Tuor, R. Baerwolf, N. Knowles, B. Hutchinson, N. Nichols, R. Jasper, Recurrent neural network language models for open vocabulary event-level cyber anomaly detection, in: Workshops at the AAAI Conference on Artificial Intelligence, 2018, pp. 285–293.

[4] S. Ahmad, A. Lavin, S. Purdy, Z. Agha, Unsupervised real-time anomaly detection for streaming data, Neurocomputing 262 (2017) 134–147.

[5] O. Gorokhov, M. Petrovskiy, I. Mashechkin, Convolutional neural networks for unsupervised anomaly detection in text data, in: International Conference on Intelligent Data Engineering and Automated Learning, 2017, pp. 500–507.

[6] K. Ghasedi Dizaji, X. Wang, H. Huang, Semi-supervised generative adversarial network for gene expression inference, in: International Conference on Knowledge Discovery and Data Mining, 2018, pp. 1435–1444.

[7] L. Zhu, N. Laptev, Deep and confident prediction for time series at Uber, in: International Conference on Data Mining Workshops, 2017, pp. 103–110.

[8] D. Oh, I. Yun, Residual error based anomaly detection using autoencoder in SMD machine sound, Sensors 18 (2018) 1308.

[9] D. Park, Y. Hoshi, C. C. Kemp, A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder, Robotics and Automation Letters 3 (2018) 1544–1551.

[10] J. Wei, J. Zhao, Y. Zhao, Z. Zhao, Unsupervised anomaly detection for traffic surveillance based on background modeling, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 129–136.

[11] T. Nolle, A. Seeliger, M. Mühlhäuser, Binet: multivariate business process anomaly detection using deep learning, in: International Conference on Business Process Management, 2018, pp. 271–287.

[12] R. Chalapathy, S. Chawla, Deep learning for anomaly detection: A survey, Arxiv Preprint arxiv 1901.03407 (2019) 1–50.

[13] N. Laptev, S. Amizadeh, I. Flint, Generic and scalable framework for automated time-series anomaly detection, in: International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1939–1947.

[14] S. Venkataraman, J. Caballero, D. Song, A. Blum, J. Yates, Black box anomaly detection: is it utopian?, HotNets (2006) 127.

[15] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT Press, 2018.

[16] C. Huang, Y. Wu, Y. Zuo, K. Pei, G. Min, Towards experienced anomaly detector through reinforcement learning, in: AAAI Conference on Artificial Intelligence, 2018, pp. 8087–8088.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533.

[18] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, Y. Bengio, An actor-critic algorithm for sequence prediction, ArXiv Preprint arXiv:1607.07086 (2016) 1–10.

[19] Y. P. Pane, S. P. Nageshrao, R. Babuška, Actor-critic reinforcement learning for tracking control in robotics, in: Conference on Decision and Control, 2016, pp. 5819–5826.

[20] J. Wang, X. Ding, M. Lahijanian, I. C. Paschalidis, C. A. Belta, Temporal logic motion control using actor–critic methods, The International Journal of Robotics Research 34 (2015) 1329–1344.

[21] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: International Conference on Machine Learning, 2016, pp. 1928–1937.

[22] K. Yamanishi, J.-I. Takeuchi, G. Williams, P. Milne, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Mining and Knowledge Discovery 8 (2004) 275–300.

[23] G. R. Kumar, N. Mangathayaru, G. Narsimha, An approach for intrusion detection using novel Gaussian based kernel function., Journal of Universal Computer Science 22 (2016) 589–604.

[24] S. Shekhar, C.-T. Lu, P. Zhang, Detecting graph-based spatial outliers: algorithms and applications (a summary of results), in: International Conference on Knowledge Discovery and Data Mining, 2001, pp. 371–376.

[25] A. M. Bianco, M. Garcia Ben, E. Martinez, V. J. Yohai, Outlier detection in regression models with ARIMA errors using robust estimates, Journal of Forecasting 20 (2001) 565–579.

[26] D. Agarwal, An empirical Bayes approach to detect anomalies in dynamic multidimensional arrays, in: International Conference on Data Mining, 2005, pp. 8–16.

[27] D. Agarwal, Detecting anomalies in cross-classified streams: a Bayesian approach, Knowledge and Information Systems 11 (2007) 29–44.

[28] K. Das, J. Schneider, Detecting anomalous records in categorical datasets, in: International Conference on Knowledge Discovery and Data Mining, 2007, pp. 220–229.

[29] J. Ma, S. Perkins, Time-series novelty detection using one-class support vector machines, in: International Joint Conference on Neural Networks, volume 3, 2003, pp. 1741–1745.

[30] G. Tandon, P. K. Chan, Weighting versus pruning in rule validation for detecting network and host anomalies, in: International Conference on Knowledge Discovery and Data Mining, 2007, pp. 697–706.

[31] S. Mukkamala, G. Janoski, A. Sung, Intrusion detection using neural networks and support vector machines, in: International Joint Conference on Neural Networks, volume 2, 2002, pp. 1702–1707.

[32] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, in: ACM Sigmod Record, volume 29, 2000, pp. 427–438.

[33] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, LOF: identifying density-based local outliers, in: ACM sigmod record, volume 29, 2000, pp. 93–104.

[34] J. Tang, Z. Chen, A. W.-C. Fu, D. W. Cheung, Enhancing effectiveness of outlier detections for low density patterns, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2002, pp. 535–548.

[35] Esty, Skyline, `https://github.com/etsy/skyline`, 2014.

[36] Twitter, Twitter anomaly detection, https://github.com/twitter/AnomalyDetection/releases, 2015.

[37] S. Mikhail, Contextual anomaly detection, https://github.com/smirmik/CAD, 2015.

[38] P. Malhotra, L. Vig, G. Shroff, P. Agarwal, Long short term memory networks for anomaly detection in time series, in: European Symposium on Artificial Neural Networks, 2015, pp. 89–94.

[39] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, G. Shroff, LSTM-based encoder-decoder for multi-sensor anomaly detection, Arxiv Preprint Arxiv:1607.00148 (2016) 1–5.

[40] T. Amarbayasgalan, B. Jargalsaikhan, K. Ryu, Unsupervised novelty detection using deep autoencoders with density based clustering, Applied Sciences 8 (2018) 1468.

[41] F. Bourdonnaye, C. Teulière, T. Chateau, J. Triesch, Learning of binocular fixations using anomaly detection with deep reinforcement learning, in: International Joint Conference on Neural Networks, 2017, pp. 760–767.