Multivariate and Online Anomaly Detection and Localization for High-Dimensional Systems

Mahsa Mozaffari
Deparment of Electrical Engineering
University of South Florida
Tampa, FL 33620
Email: mmozaffari@mail.usf.edu

Yasin Yilmaz
Department of Electrical Engineering
University of South Florida
Tampa, FL 33620
Email: yasiny@usf.edu

***Abstract*—This paper considers the real-time and nonparametric detection and localization of anomalies in high-dimensional systems. The goal is to detect anomalies quickly and accurately such that the appropriate countermeasures could be taken before any possible harm is caused by the anomalous event. We propose a kNN-based sequential anomaly detection method in both semi-supervised and supervised settings, in conjuction with an effective method for localizing the anomalous data dimensions. We prove that the proposed method is asymptotically optimum in the minimax sense under certain conditions in terms of minimizing the average detection delay for a given false alarm constraint. The proposed method is shown to be capable of multivariate anomaly detection and also scalable to high-dimensional datasets. We further propose an online learning scheme that combines the desirable properties of our semi-supervised and supervised methods.**

I. Introduction

Anomaly detection is an important problem which deals with the identification of abnormal data patterns which do not conform to the normal behavior of a system. It has applications in a wide range of domains, such as cyber-security [1], quality control, medical health care [2], etc. The importance of anomaly detection lies in the fact that an anomaly in the observations is typically due to an unwanted behavior/event in the underlying system that needs to be dealt with by a field specialist. Due to the potential unpleasant and even catastrophic consequences of an undetected anomalous event in the system, it is crucial to detect the anomalies quickly and timely, so that the appropriate countermeasures could be taken in time. Moreover, certain applications require further information in addition to detection of the anomalies, explaining where the detected anomaly has occurred in the system.

Statistical anomaly detection approaches consider an anomaly as a change in the probability distribution of data, e.g., change in the mean, variance or correlation structure between individual data-streams. Multivariate anomaly detection has the potential to achieve better performance in comparison to univariate detection, especially in challenging settings. For instance, detecting the anomalous observations that appear to be normal (e.g., as a result of a malicious activity), and the detection of a change in the correlation structure of data [3] are two examples that highlight the importance of multivariate analysis and joint monitoring of data-streams, which in turn, leads to the high-dimensionality challenge. A practical multivariate anomaly detection method needs to scale well to high-dimensional data in real-time.

Parametric anomaly detection methods assume knowledge of the underlying probability distributions, hence they are not effectively applicable to high-dimensional real-world problems with complex distributions. Additionally, these methods are limited to the detection of certain types of anomalies that match the assumed distributions well. Nonparametric techniques, on the other hand, do not assume specific probability distributions for the data. Nonparametric anomaly detection methods based on k nearest neighbors (kNN) are proposed in several works, e.g., [4]–[7]. These geometric methods are based on the assumption that anomalous instances occur in the less concentrated regions of the nominal data space. Although the methods proposed in [4], [5] are effective in multivariate anomaly detection in high-dimensional data, they are not well suited for accurate detection in real-time systems as they do sample-by-sample detection without considering the sequential aspect of anomalies [8]. While [6] has a sequential nature, its computational complexity is note suitable for real-time applications.

Motivated by the aforementioned challenges, aiming at timely and accurate detection of anomalies in high-dimensional systems, in this paper we (i) prove the asymptotic optimality of the nonparametric sequential method proposed in [7] in the minimax sense, (ii) propose an extension for supervised settings with training data available for both nominal and anomalous cases, (iii) propose an anomaly localization approach based on the proposed detection methods, in order to identify the anomalous data dimensions, and (iv) introduce an online learning scheme by combining the advantages of both supervised and semi-supervised variants.

The rest of the paper is organized as follow. In Section II, we present the problem formulation and the related background information. In Section III, we present the semi-supervised, supervised, and unified variants of our anomaly detection method, as well as the localization technique. The experimental results on simulated data and a real dataset are provided in Section IV. Finally, the paper is concluded in Section V.

II. Problem Formulation

Suppose that a system is sequentially monitored through d-dimensional observations $\mathcal{X}_t = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t\}$ in time. Assuming an abrupt and persistent anomaly occurs at an unknown time τ in the observations, the objective is to detect the anomaly as soon as possible while satisfying a false alarm constraint. This problem is typically formulated as an online change detection problem:

$$f = f_0,\ t < \tau, \quad f = f_1 (\neq f_0),\ t \geq \tau, \tag{1}$$

where f is the true probability distribution of observations, and f_0 and f_1 are the nominal and anomalous probability distributions, respectively. The objective of the problem is to find the stopping time T that minimizes the average detection delay while satisfying a false alarm constraint, i.e.,

$$\inf_T \mathsf{E}_\tau[(T-\tau)^+] \ \text{ subject to } \ \mathsf{E}_\infty[T] \geq \beta, \tag{2}$$

where E_τ represents the expectation given that change occurs at τ, $(.)^+ = \max(., 0)$, and $\mathsf{E}_\infty[T]$ denotes the the expectation of false alarm period.

Lorden's minimax problem is a commonly used version of the above problem [9], in which the goal is to minimize the worst-case average detection delay subject to a false alarm constraint:

$$\inf_T \sup_\tau \operatorname*{ess\,sup}_{\mathcal{X}_\tau} E_\tau[(T-\tau)^+ | \mathcal{X}_\tau] \ \text{ s.t. } \ E_\infty[T] \geq \beta, \tag{3}$$

where "ess sup" denotes essential supremum that is equivalent to supremum in practice. In short, the minimax criterion minimizes the average detection delay for the least favorable change-point and the least favorable history of measurements up to the change-point while the average false alarm period is constrained by β.

The Cumulative Sum (CUSUM) detector provides the optimum solution to the minimax problem [10], given by (3),

$$\begin{aligned} S_t &= \max\{0, S_{t-1} + \ell_t\}, \\ T_c &= \inf\{t : S_t \geq h_c\}, \end{aligned} \tag{4}$$

where T_c is the stopping time, S_t is the test statistic, $\ell_t = \log \frac{f_1(\boldsymbol{x}_t)}{f_0(\boldsymbol{x}_t)}$ is the log-likelihood ratio at time t, $S_0 = 0$, and h_c is the predefined decision threshold. Considering ℓ_t as a statistical evidence for anomaly, the CUSUM algorithm accumulates the evidences over time, and stops when the cumulative evidence S_t is sufficiently high for reliable detection, where the level of "sufficiently high" is represented by h_c and chosen to satisfy the false alarm constraint β.

CUSUM requires the complete knowledge of the probability distributions f_0 and f_1, which are typically unknown in real-world applications. Generalized CUSUM (G-CUSUM) is a variation of CUSUM which knowing the distributions, estimates the parameters of f_0 and f_1 by maximum likelihood estimation and achieves asymptotic optimality. Moreover, CUSUM and in general parametric methods are limited to the detection of certain anomaly types whose true probability distribution matches the assumed f_1 well.

III. The proposed Method

A. Online Discrepancy Test (ODIT)

We have recently proposed a kNN-based sequential anomaly detection method, called Online Discrepancy Test (ODIT) [7] demonstrated its applications to cyberattack detection in smart grid [1] and intelligent transportation systems [11]. In this section, we present a modification for ODIT to prove its asymptotic optimality in the minimax sense under certain conditions. ODIT combines the sequential nature of CUSUM and the nonparametric nature of the Geometric Entropy Minimization (GEM) method [4] for multivariate and online anomaly detection. In a semi-supervised fashion, ODIT trains only on nominal data to learn a statistical description of normal system behavior, and tests the new observations in a sequential manner against the learned nominal model.We next describe the ODIT procedure with the proposed modification.

Considering a nominal training set $\mathcal{X}_N$ of size N, ODIT partitions $\mathcal{X}_N$ into two sets $\mathcal{X}_{N_1}$ and $\mathcal{X}_{N_2}$, where $N_1 + N_2 = N$, for computational efficiency as in the bipartite GEM algorithm [5]. Then, it computes the Euclidean distances between each point $\boldsymbol{x}_m \in \mathcal{X}_{N_1}$ and its k nearest neighbors in $\mathcal{X}_{N_2}$. The total kNN distance of $\boldsymbol{x}_m$ is defined as

$$L_m = \sum_{n=k-s+1}^{k} g_n(\boldsymbol{x}_m)^\gamma, \tag{5}$$

where $g_n(\boldsymbol{x}_m)$ is the Euclidean distance between point $\boldsymbol{x}_m \in \mathcal{X}_{N_1}$ and its nth nearest neighbor in $\mathcal{X}_{N_2}$, $s \in \{1, \ldots, k\}$ is a fixed number introduced for convenience, and $\gamma > 0$ is a weight also introduced for flexibility. Given a significance level $\alpha \in (0, 1)$, e.g., 0.05, the training phase is finished by choosing the $(1-\alpha)$th percentile of total kNN distances $\{L_m\}$. That is, ODIT selects the Kth smallest distance $L_{(K)}$, where $K = \lfloor N_1(1-\alpha) \rfloor$, as a baseline statistic for measuring the deviation of new observations from the nominal dataset in the test phase. In other words, in the training phase, ODIT practically learns the most compact region in the nominal data geometry.

During the test phase, for each observation $\boldsymbol{x}_t$, ODIT computes the total kNN distance L_t with respect to the nominal points in $\mathcal{X}_{N_2}$ using (5), and computes the anomaly evidence as

$$D_t = d(\log L_t - \log L_{(K)}), \tag{6}$$

where d is the dimensionality of the data. This equation is the modification that we propose for ODIT in this paper. In [7], D_t has the simpler form $D_t = L_t - L_{(K)}$. Although this simpler form of D_t and the form proposed in (6) have similar difference structures, and they perform quite similarly in practice, the new form given in (6) naturally appears while proving the asymptotic optimality of ODIT in the minimax sense, as shown in Theorem 1. D_t denotes a positive/negative evidence for anomaly. Positive D_t suggests that the observation lies outside the estimated most compact set of the nominal training set, hence it provides a positive evidence for anomaly. ODIT recursively updates a detection statistic Δ_t by accumulating

the anomaly evidences over time. The test continues until the first time Δ_t exceeds a predefined threshold h, suggesting that there is sufficient evidence supporting anomaly in the observations. The update and decision rule of ODIT are given as

$$\begin{aligned} \Delta_t &= \max\{\Delta_{t-1} + D_t, 0\}, \quad \Delta_0 = 0, \\ T &= \min\{t : \Delta_t \geq h\}, \end{aligned} \tag{7}$$

which is a CUSUM-like procedure (cf. (4)). The threshold h controls the trade-off between minimizing average detection delay and minimizing false alarm rate. Larger threshold would decrease the false alarm rate at the expense of larger detection delays, and smaller threshold would result in smaller detection delays and larger false alarm rates. Thus, h should be selected to strike a desired balance between false alarm rate and detection delay. The ODIT procedure is summarized in 1.

Algorithm 1 The proposed ODIT procedure

Input: $\mathcal{X}_N, k, s, \alpha, h$
Initialize: $\Delta \leftarrow 0, \mathrm{t} \leftarrow 1$
Training phase:
Partition $\mathcal{X}_N$ into two sets $\mathcal{X}_{N_1}$ and $\mathcal{X}_{N_2}$
For each $\boldsymbol{x}_m \in \mathcal{X}_{N_1}$ compute L_m as in (5)
Find $L_{(K)}$ by selecting the Kth smallest L_m
Test phase:
while $\Delta < h$ **do**
    Get new data $\boldsymbol{x}_t$ and compute D_t as in (6)
    $\Delta = \max\{\Delta + D_t, 0\}$
    $t \leftarrow t + 1$
Declare Anomaly

Theorem 1. *When the nominal distribution $f_0(\boldsymbol{x}_t)$ is finite and continuous, and the attack distribution $f_1(\boldsymbol{x}_t)$ is a uniform distribution, as the training set grows, the ODIT statistic D_t converges in probability to the log-likelihood ratio,*

$$D_t \xrightarrow{p} \log \frac{f_1(\boldsymbol{x}_t)}{f_0(\boldsymbol{x}_t)} \quad as \quad N \to \infty, \tag{8}$$

i.e., the ODIT converges to CUSUM, which is minimax optimum in minimizing expected detection delay while satisfying a false alarm constraint.

Proof: Consider a hypersphere $\mathcal{S}_t \in \mathbb{R}^d$ centered at $\boldsymbol{x}_t$ with radius $g_k(\boldsymbol{x}_t)$, the kNN distance of $\boldsymbol{x}_t$ with respect to the training set $\mathcal{X}_N$. The maximum likelihood estimate for the probability of a point being inside $\mathcal{S}_t$ under f_0 is given by k/N. It is known that, as the total number of points grow, this binomial probability estimate converges to the true probability mass in $\mathcal{S}_t$ in the mean square sense [12], i.e., $k/N \xrightarrow{L^2} \int_{\mathcal{S}_t} f_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$ as $N \to \infty$. Hence, the probability density estimate $\hat{f}_0(\boldsymbol{x}_t) = \frac{k/N}{V_d g_k(\boldsymbol{x}_t)^d}$, where $V_d g_k(\boldsymbol{x}_t)^d$ is the volume of $\mathcal{S}_t$ with the appropriate constant V_d, converges to the actual probability density function, $\hat{f}_0(\boldsymbol{x}_t) \xrightarrow{p} f_0(\boldsymbol{x}_t)$ as $N \to \infty$, since $\mathcal{S}_t$ shrinks and $g_k(\boldsymbol{x}_t) \to 0$. Similarly, considering a hypersphere $\mathcal{S}_{(K)} \in \mathbb{R}^d$ around $\boldsymbol{x}_{(K)}$ which includes k points with its radius $g_k(\boldsymbol{x}_{(K)})$, we see that as $N \to \infty$, $g_k(\boldsymbol{x}_{(K)}) \to 0$ and $\hat{f}_0(\boldsymbol{x}_{(K)}) = \frac{k/N}{V_d g_k(\boldsymbol{x}_{(K)})^d} \xrightarrow{p} f_0(\boldsymbol{x}_{(K)})$. Assuming a uniform distribution $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x}_{(K)}),\ \forall \boldsymbol{x}$, we conclude with $\log \frac{\frac{k/N}{V_d g_k(\boldsymbol{x}_{(K)})^d}}{\frac{k/N}{V_d g_k(\boldsymbol{x}_t)^d}} = d\left[\log g_k(\boldsymbol{x}_t) - \log g_k(\boldsymbol{x}_{(K)})\right] \xrightarrow{p} \log \frac{f_1(\boldsymbol{x}_t)}{f_0(\boldsymbol{x}_t)}$ as $N \to \infty$, where $L_t = g_k(\boldsymbol{x}_t)$ for $s = \gamma = 1$. For γ values different than 1, D_t converges to the log-likelihood ratio scaled by γ. ∎

Note that ODIT does not train on any anomalous data, i.e., does not use any knowledge of anomaly to be detected. While this generality is an attractive trait as it allows detection of any statistical anomaly, it also inevitably limits the performance for known anomaly types on which detectors can train. We will next extend ODIT to this case with available anomaly information. In Theorem 1, we show that in the lack of knowledge about anomalies, ODIT reasonably assumes an uninformative uniform likelihood for the anomaly case, and achieves asymptotic optimality under this assumption in the CUSUM-sense for certain parameter choices.

B. An Extension: ODIT-2

In this section we consider the case of having an anomaly training dataset in addition to the previously discussed nominal dataset. We extend the ODIT algorithm to take advantage of the anomaly dataset in order to improve the performance. Consider the nominal and anomalous datasets of $\mathcal{X}_N = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ and $\mathcal{X}'_M = \{\boldsymbol{x}'_1, \boldsymbol{x}'_2, ..., \boldsymbol{x}'_M\}$, respectively. In this case, the anomaly evidence for each observation instance can be computed by comparing the total distance L_t with respect to $\mathcal{X}_N$ to the total distance L'_t with respect to $\mathcal{X}'_M$. Hence, ODIT-2 doesn't require the borderline total distances of the train data to use as a baseline in testing (cf. 6). This implies that no training is needed for ODIT-2.

During the test phase, the anomaly evidence for each observation instance $\boldsymbol{x}_t$ is calculated by

$$D_t = d(\log L_t - \log L'_t) + \log(N/M), \tag{9}$$

where L_t and L'_t are the total kNN distances of $\boldsymbol{x}_t$ with respect to the points in $\mathcal{X}_N$ and $\mathcal{X}'_M$, respectively; and N and M are the number of points in the nominal and anomalous training sets. The statistic update and decision rule of ODIT-2 are computed in the same way as ODIT, given by (7).

The anomaly evidence D_t of ODIT-2 practically contrasts the new observation against the nominal and anomaly training data and computes a measure of how well $\boldsymbol{x}_t$ is aligned with the anomaly class as compared to the nominal class. The positive D_t suggests that the new observation is closer to the description of the anomaly class, compared to the nominal class. Due to the inherent difficulty in collecting anomaly samples, typically there is an imbalance between the nominal and anomaly training sets. The total kNN distances in a dense nominal set $\mathcal{X}_N$ are expected to be small as compared with the total kNN distances in a sparse anomaly dataset. The term $\log(N/M)$ is used as normalization factors to deal with such imbalance.

Corollary 1. *When the nominal distribution $f_0(\boldsymbol{x}_t)$ and anomalous distribution $f_1(\boldsymbol{x}_t)$ are finite and continuous, as the training sets grow, the ODIT-2 statistic D_t, given by* (9)*, converges in probability to the log-likelihood ratio,*

$$D_t \xrightarrow{p} \log \frac{f_1(\boldsymbol{x}_t)}{f_0(\boldsymbol{x}_t)} \quad as \quad M, N \to \infty, \tag{10}$$

i.e., ODIT-2 converges to CUSUM, which is minimax optimum in minimizing expected detection delay while satisfying a false alarm constraint.

Proof: From the proof of Theorem 1, we know that $\frac{k/N}{V_d g_k(\boldsymbol{x}_t)^d} \xrightarrow{p} f_0(\boldsymbol{x}_t)$ as $N \to \infty$. Similarly, we can show that $\frac{k/M}{V_d g'_k(\boldsymbol{x}_t)^d} \xrightarrow{p} f_1(\boldsymbol{x}_t)$ as $M \to \infty$, where $g'_k(\boldsymbol{x}_t)$ is the kNN distance of $\boldsymbol{x}_t$ in the anomalous training set $\mathcal{X}'_M$. Hence, we conclude with $\log \frac{\frac{k/M}{V_d g'_k(\boldsymbol{x}_t)^d}}{\frac{k/N}{V_d g_k(\boldsymbol{x}_t)^d}} = d\left[\log g_k(\boldsymbol{x}_t) - \log g'_k(\boldsymbol{x}_t)\right] + \log(N/M) \xrightarrow{p} \log \frac{f_1(\boldsymbol{x}_t)}{f_0(\boldsymbol{x}_t)}$ as $M, N \to \infty$, where $L_t = g_k(\boldsymbol{x}_t)$ and $L'_t = g'_k(\boldsymbol{x}_t)$ for $s = \gamma = 1$. ∎

It is also noteworthy that for challenging applications in which the nominal and anomaly datasets are very similar, a pre-processing step on the anomaly train set might be required to remove the data points that are similar to the nominal train set. This step is done by finding and removing the data points of $\mathcal{X}'_M$ which lie in the estimated most compact region of the nominal train set, i.e.,

$$\mathcal{X}'^{\text{clean}}_M = \mathcal{X}'_M \setminus \{\boldsymbol{x}'_m \in \mathcal{X}'_M : L_{\boldsymbol{x}'_m} \leq L_{(K_N)}\}, \tag{11}$$

where $L_{\boldsymbol{x}'_m}$ is the total distance of $\boldsymbol{x}'_m$ with respect to the nominal train set. If the cleaning process is performed on the anomaly training set, L'_t in (9) is computed with respect to $\mathcal{X}'^{\text{clean}}_M$.

C. Anomaly Localization

In this section, we propose a localization scheme to identify the data dimensions in which the anomaly has occurred, leading to the detection. An effective localization has a pivotal role in identifying the cause of anomaly and hence mitigating it in time. We approach this task by examining the contribution of each dimension individually to the increase in the detection statistic that resulted in detection. In ODIT detector, the increase in the detection statistic Δ_t, given by (7) and consequently the anomaly alarm, is caused by the increase in D_t, given by (6) which in turn is the result of an increase in the total distance L_t, given by (5). Let us assume x_t is the test observation at time t and $y_1, ..., y_k$ are its k nearest neighbors in the train set. We can rewrite the total kNN distance $L_t = \sum_{n=k-s+1}^{k} ||x_t - y_n||^\gamma$, for $\gamma = 2$ in terms of the sum of total distances along each dimension d:

$$L_t = \sum_{i=1}^{d} \delta_t^i, \quad \text{where} \quad \delta_t^i = \sum_{n=k-s+1}^{k} (x_t^i - y_n^i)^2, \tag{12}$$

where x_t^i and y_n^i are the ith dimension of the observation x_t and its nth nearest neighbor y_n, respectively, and δ_t^i is the contribution of the ith dimension of the observation x_t to Δ_t at time t. By analyzing the δ_t^i per each dimension, during a period of which Δ_t is increasing, we identify the anomalous data dimensions. To that end, we propose to use a history of $\mathcal{Q}_i = \{\delta_q^i : q = \hat{\tau}+1, ..., \hat{\tau}+S\}$ per each dimension i, where $\hat{\tau}$ is the estimated anomaly onset, is the most recent time that the detection statistic was zero. We perform a t-test on the S samples in $\mathcal{Q}_\rangle$ to decide whether the ith dimension is anomalous.

The localization procedure after an anomaly alarm is raised by ODIT at time T, is as follows:

1) Find $\hat{\tau} = \max\{t < T : \Delta_t = 0\}$
2) for each dimension i, compute the sample mean and sample standard deviation of $\mathcal{Q}_i$:

$$\overline{\delta_i} = \frac{1}{S} \sum_{t=\hat{\tau}+1}^{\hat{\tau}+S} \delta_t^i, \quad \eta_i = \sqrt{\frac{1}{S-1} \sum_{t=\hat{\tau}+1}^{\hat{\tau}+S} (\delta_t^i - \overline{\delta}_i)} \tag{13}$$

3) Identify the anomalous dimensions dimensions by t-test:

$$\text{dimension } i \text{ is anomalous, if } \quad \frac{\overline{\delta}_i - \mu_i}{\eta_i/\sqrt{S}} \geq \theta, \tag{14}$$

where μ_i is the sample mean of the contributions of dimension i of nominal training data i.e. $\{\delta_1^i, ..., \delta_{N_1}^i\}$, and θ is the $(1-\beta)$th percentile, given the significance level β, of Student's t-distribution with $S-1$ degree of freedom.

Significance level β makes the balance between sensitivity to anomalies and robustness to outliers. Given the β and S values, the threshold θ can be easily found according to the Student's t-distribution lookup table. Additionally, the number of the samples S should be at least 2, to ensure that the degree of freedom is at least 1.

Localization by ODIT-2 is slightly different. According to (9), the detection by ODIT-2 is caused by the increase in the $\log L_t - \log L'_t$. Similar to (12), we can write the L_t and L'_t in terms of the individual contributions, δ_t^i and $\delta_t^{'i}$, respectively. It is obvious that the increase in the $(\delta_t^i - \delta_t^{'i})$ for some dimensions i, causes the Statistic to increase. Therefore, the localization by ODIT-2 procedure is done by replacing δ_t^i with $(\delta_t^i - \delta_t^{'i})$ in (13) and (14).

D. The Unified Framework

Availability of labeled training data is a major limiting factor for improving the performance of anomaly detection techniques. While obtaining comprehensive and accurate labeled data for the anomaly class in several applications is very difficult, in most applications typically sufficient amount of labeled nominal data is available. Semi-supervised methods including ODIT, constitute a popular class of anomaly detection methods that build a model of normality only from the nominal training data, and perform anomaly detection by finding the data which deviates from this model. On the other hand, supervised techniques including ODIT-2, require both nominal and anomalous datasets to build models for

classifying data into nominal vs. anomaly classes. ODIT-2 outperforms the semi-supervised ODIT method, for the known anomaly types (as shown in Section IV). However, ODIT-2, and in general supervised anomaly detectors, have the drawback of achieving poor performance for detecting unknown anomaly types. Whereas, ODIT, and in general semi-supervised anomaly detection methods, are capable of detecting any anomaly type as long as it sufficiently deviates from the nominal model. This motivate us to combine the desirable properties of ODIT and ODIT-2, and propose an online learning scheme which is capable of detecting previously unseen anomalies and achieving better performance for the known anomalies.

In the unified framework, both ODIT and ODIT-2 run in parallel while a feedback loop includes the anomalous data points first detected by ODIT in the anomaly training set of ODIT-2 to empower the detection of similar anomaly types. The unified scheme, called ODIT-uni, monitors the detection statistics of ODIT and ODIT-2 in parallel, and stops the first time either one stops:

$$\Delta_t^{(1)} = \max\{\Delta_t^{(1)} + D_t^{(1)}, 0\},\ \Delta_t^{(2)} = \max\{\Delta_t^{(2)} + D_t^{(2)}, 0\}$$
$$T = \min\{t : \Delta_t^{(1)} \geq h_1 \text{ or } \Delta_t^{(2)} \geq h_2\}, \quad (15)$$

where $D_t^{(1)}$ and $D_t^{(2)}$ are the anomaly evidences given by (6) and (9), respectively, and h_1 and h_2 are the predefined thresholds of ODIT and ODIT-2. It is expected that for known anomaly types, $\Delta_t^{(2)} \geq h_2$ happens earlier. Whereas for unseen anomaly types, $\Delta_t^{(1)} \geq h_1$ is expected to detect the anomaly. If ODIT raises an alarm at time T, anomaly start time $\hat{\tau}$ is estimated as the last time before T that ODIT statistics was 0. Then, the feedback loop incorporates the data instances $\{\boldsymbol{x}_{\hat{\tau}+1}, \ldots, \boldsymbol{x}_T\}$ between $\hat{\tau}$ and T into the anomaly train set. The threshold h_1 needs to be selected sufficiently large to prevent false alarms by ODIT and consequently false inclusions of detected data instances into ODIT-2 training set, even though this causes an increase in the detection delay of unseen anomalies.

IV. Numerical Results

In this section we present numerical results to demonstrate the advantage of multivariate analysis by ODIT and ODIT-2 over the G-CUSUM detector, in a challenging case in which anomaly is defined as a change in the correlation between individual data-streams. We simulate a 100-dimensional system in which the nominal data is distributed according to a multivariate Gaussian distribution, with $\mu = 20$ and a diagonal covariance with standard deviation being $\sigma = 10$. The anomaly is defined as adding $\rho = 0.6$ correlation between 50% of the data-streams, while the mean and standard deviation are intact.

Fig. 1 compares the average performance of ODIT and ODIT-2 with the oracle CUSUM which has the complete knowledge of the underlying multivariate distributions. Since it is not tractable to estimate the high-dimensional multivariate distributions, G-CUSUM assumes independence among individual data-streams and combines the univariate analysis on each data-stream as in [13]. G-CUSUM fails to detect the anomalies in the observation as it is not able to monitor the correlations. The ODIT methods successfully detect the change in the covariance structure of observations by multivariate analysis while ODIT-2 outperforms ODIT as expected, and well-approximates the optimum CUSUM, which is not a practical detector.

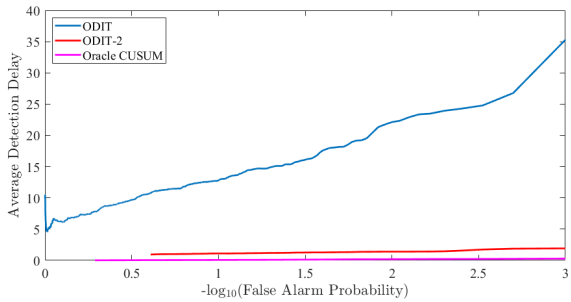
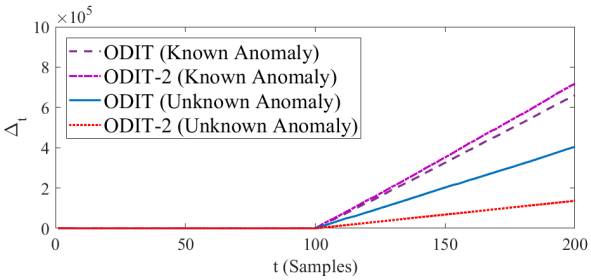
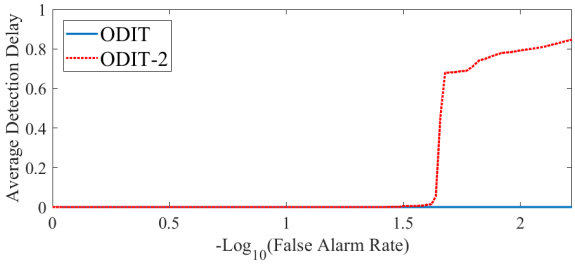


Fig. 1: Performance comparison of ODIT, ODIT-2 and CUSUM in the correlation monitoring example.



(a) Decision statistics of the proposed ODIT detectors in both scenarios (known and unknown types of anomalies).



(b) Performance comparison for the proposed ODIT detectors in the unknown anomaly type scenario.

Fig. 2: Experimental results on the N-BaIoT dataset.

We also applied the proposed ODIT detectors to a dataset of real traffic data for network-based detection of IoT botnet attacks (N-BaIoT dataset [14]) in order to compare ODIT detectors and demonstrate the unified framework presented in Section III-D. This dataset is gathered from 9 IoT devices under nominal operation and while infected by IoT-based botnets. In the experiments the dimensionality of the data is 1035. We assume that we have anomaly training data from past observations, where device 2 is acting maliciously. In order to fairly compare ODIT with ODIT-2, we test for two different scenarios: 1) device 2 is compromised (known anomaly type)

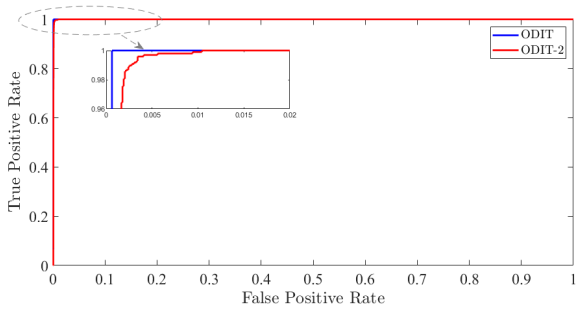


Fig. 3: ROC curve for anomaly localization using ODIT and ODIT-2 in the known attack scenario for the N-BaIoT dataset.

2) device 6 is compromised (unknown anomaly type).

In case of the first scenario, both ODIT and ODIT-2 are able to detect the anomaly with zero average detection delay. Fig. 2(b) compares the average performance of ODIT and ODIT-2 for the second scenario. Although ODIT-2 is still able to detect the anomaly, its performance degrades as compared to the known anomaly scenario. This is due to the mismatch between the type of anomaly in the observations and that of the anomaly train set. Fig. 2(a) shows the decision statistics of ODIT and ODIT-2 for both scenarios. Unlike the first scenario in which the decision statistic of ODIT-2 is stronger than that of the ODIT, in second scenario, the decision statistic of ODIT-2 becomes weaker than that of ODIT.

Next, Fig. 3 demonstrates the performance of localization techniques, in terms of the ROC curve (true positive rate vs. false positive rate) under the known anomaly type scenario. Both methods, identify the malicious device with very high accuracy, and very low false alarm rate.

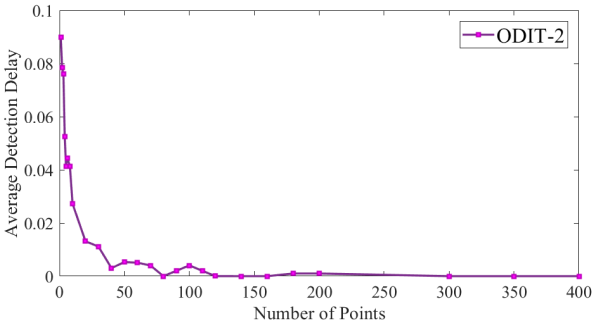


Fig. 4: The average detection delay of ODIT-2 for an unknown anomaly type, versus the number of the data instance of the unknown anomaly added to the anomaly train set. The false alarm probability is set to be $P(\textit{False Alarm rate}) = 0.01$

We employed the unified framework ODIT-uni to demonstrate the improvement of ODIT-2 performance for novel anomaly types, as the anomaly train set grows by the incorporation of new anomaly observations. Following the above experiment on N-BaIoT, we test for the performance of ODIT-2 in detecting the new anomaly type (scenario 2). Fig. 4 suggests that as the anomaly train set is enhanced by the new data instances of unknown anomaly type, ODIT-2 performance for detecting future observations of the same anomaly type improves and converges to zero for sufficiently enhanced training set.

V. Conclusion

In this paper, we proposed a multivariate and online anomaly detection and localization framework that is suitable for real-time and high-dimensional systems for both semi-supervised and supervised settings. We showed the asymptotic optimality of the proposed methods in the minimax sense, in terms of minimizing the average detection delay for a given false alarm constraint. The performance of the variations of proposed methods was evaluated in the challenging case of detecting a change in the covariance structure. Both ODIT and ODIT-2 successfully detect the change while ODIT-2 achieves a close performance to the oracle CUSUM detector, which is the minimax optimum detector but not tractable in practice. We also provided experiment results in the context of botnet detection on a real dataset (N-BaIoT). Combining the advantages of the variations of our method, we also proposed a unified ODIT scheme that can detect novel anomaly types, as well as improve its performance over time by enhancing its training set via the detected anomalous data instances. The experiments on the N-BaIoT dataset corroborated that the unified scheme efficiently learns to quickly and accurately detect similar anomalies in the future.

References

[1] Y. Yilmaz and S. Uludag, "Mitigating iot-based cyberattacks on the smart grid," in *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE, 2017, pp. 517–522.

[2] H. Zhang, J. Liu, and N. Kato, "Threshold tuning-based wearable sensor fault detection for reliable medical monitoring using bayesian network model," *IEEE Systems Journal*, vol. 12, no. 2, pp. 1886–1896, 2018.

[3] V. Avanesov, N. Buzun *et al.*, "Change-point detection in high-dimensional covariance structure," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 3254–3294, 2018.

[4] A. O. Hero, "Geometric entropy minimization (gem) for anomaly detection and localization," in *Advances in Neural Information Processing Systems*, 2007, pp. 585–592.

[5] K. Sricharan and A. O. Hero, "Efficient anomaly detection using bipartite k-nn graphs," in *Advances in Neural Information Processing Systems*, 2011, pp. 478–486.

[6] H. Chen, "Sequential change-point detection based on nearest neighbors," *arXiv preprint arXiv:1604.03611*, 2016.

[7] Y. Yilmaz, "Online nonparametric anomaly detection based on geometric entropy minimization," in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 3010–3014.

[8] M. Baker, "Statisticians issue warning over misuse of p values," *Nature News*, vol. 531, no. 7593, p. 151, 2016.

[9] G. Lorden *et al.*, "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[10] G. V. Moustakides *et al.*, "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[11] A. Haydari and Y. Yilmaz, "Real-time detection and mitigation of ddos attacks in intelligent transportation systems," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 157–163.

[12] A. Agresti, *An introduction to categorical data analysis*. Wiley, 2018.

[13] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

[14] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-baiot—network-based detection of iot botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018.