Nonparametric Sequential Change Detection for High-Dimensional Problems

Yasin Yılmaz
Electrical Engineering, University of South Florida

Allerton 2017

Outline

Introduction

Anomaly Detection

- Objective: identify patterns that deviate from a nominal behavior
- Applications: cybersecurity, quality control, fraud detection, fault detection, health care, ...

Anomaly Detection

- Objective: identify patterns that deviate from a nominal behavior
- Applications: cybersecurity, quality control, fraud detection, fault detection, health care, . . .

In literature typically

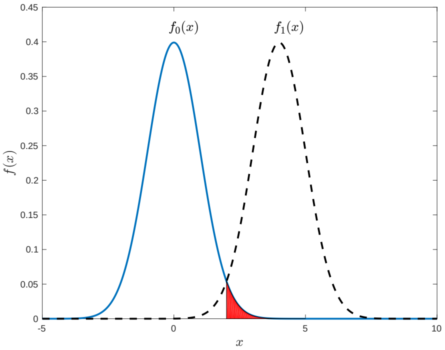
statistical outlier detection
=
anomaly detection

However an outlier could be

- nominal tail event

  or

- real anomalous event (e.g., mean shift)



Problem Formulation

Instead of *anomaly* = *outlier*, consider also temporal dimension

Proposed Model

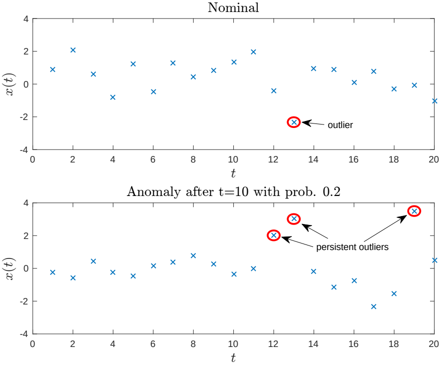
anomaly = *persistent outliers*

Objective

Timely and accurate detection of anomalies in high-dimensional datasets

Approach

Sequential & Nonparametric anomaly detection



Motivating Facts: IoT Security, Smart Grid, ...

- IoT devices: 8.4B in 2017 and expected to hit 20B by 2020 [1]
- IoT systems: highly vulnerable – needs scalable security solutions [2]
- Mirai IoT botnet: largest recorded DDoS attack with at least 1.1 Tbps bandwidth (Oct. 2016) [2]
- Persirai IoT botnet targets at least 120,000 IP cams (May 2017) [3]
- A plausible cyberattack against the US grid: 100M people may be left without power with up to $1 trillion of monetary loss [4]

[1] R. Minerva, A. Biru, and D. Rotondi, "Towards a definition of the Internet of Things (IoT)," IEEE Internet Initiative, no. 1, 2015.

[2] E. Bertino and N. Islam, "Botnets and Internet of Things Security," Computer, vol. 50, no. 2, pp. 76-79, Feb. 2017.

[3] Trend Micro, "Persirai: New Internet of Things (IoT) Botnet Targets IP Cameras", May 9 , 2017, available online

[4] Trevor Maynard and Nick Beecroft, "Business Blackout," Lloyd's Emerging Risk Report, p. 60, May 2015.

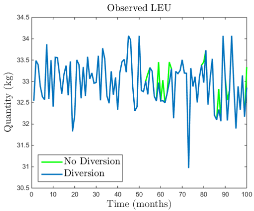
Motivating Facts: IoT Security, Smart Grid, . . .

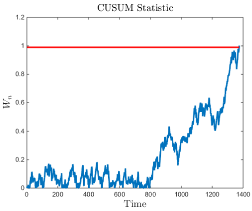
Challenges:

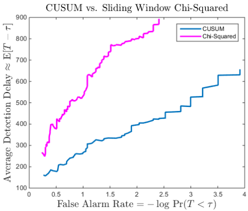
- Unknown anomalous distribution: parametric methods, as well as signature-based methods (e.g., antivirus) are not feasible
- High-dimensional problems: even nominal distribution is difficult to know
- Nonparametric methods are needed
- Timely and accurate detection is critical

Background

Sequential Change Detection - CUSUM







$$\inf_{T} \sup_{\tau} \sup_{\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\tau\}} E_\tau[T-\tau \mid T \geq \tau] \quad \text{s.t.} \quad E_\infty[T] \geq \beta$$

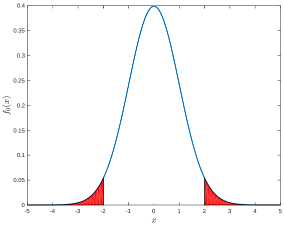
$$W_t = \max\left\{ W_{t-1} + \log \frac{f_1(\boldsymbol{x}_t)}{f_0(\boldsymbol{x}_t)}, 0 \right\}$$

$$T = \min\{t : W_t \geq h\}$$

Statistical Outlier Detection

- Needs to know a statistical description f_0 of the nominal (e.g., no attack) behavior (baseline)
- Determines instances that significantly deviate from the baseline
- With f_0 completely known, x is outlier if $\int_x^\infty f_0(y)dy < \alpha$ (p-value)
- Equivalently, if $x \notin$ most compact set of data points under f_0 (minimum volume set)

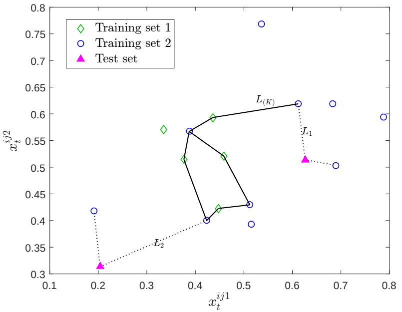
$$\Omega_\alpha = \arg\min_{\mathcal{A}} \int_{\mathcal{A}} dy \quad \text{subject to} \quad \int_{\mathcal{A}} f_0(y)dy \geq 1 - \alpha$$



- Uniformly most powerful test when anomalous distribution is a linear mixture of f_0 and the uniform distribution
- Coincides with minimum entropy set which minimizes the Rényi entropy while satisfying the same false alarm constraint

Geometric Entropy Minimization (GEM)

- High-dimensional datasets: even if f_0 is known, very computationally expensive (if not impossible) to determine Ω_α
- Various methods for learning Ω_α
- GEM is very effective with high-dimensional datasets while asymptotically achieving Ω_α for $\lim_{K,N\to\infty} K/N \to 1-\alpha$



- Training: Randomly partitions training set into two and forms K-kNN graph [5]

$$\bar{\mathcal{X}}_K^{N_1} = \arg\min_{\mathcal{X}_K^{N_1}} \mathcal{L}_k(\mathcal{X}_K^{N_1}, \mathcal{X}^{N_2}) = \sum_{i=1}^{K} \sum_{l=k^*}^{k} |e_{i(l)}|^\gamma$$

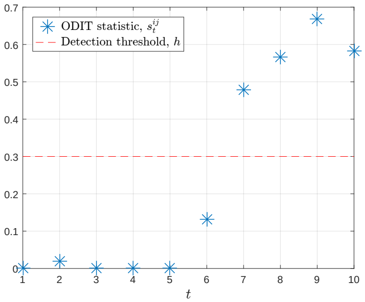
- Test: new point $\boldsymbol{x}_t \in \mathbb{R}^d$ outlier if $\boldsymbol{x}_t \notin \bar{\mathcal{X}}_K^{N_1+1}$, equivalently if $L_t = \sum_{l=k^*}^{k} |e_{t(l)}|^\gamma > L_{(K)}$

[5] A. O. Hero III, "Geometric entropy minimization (GEM) for anomaly detection and localization", NIPS, pp. 585-592, 2006

ODIT: Online Discrepancy Test

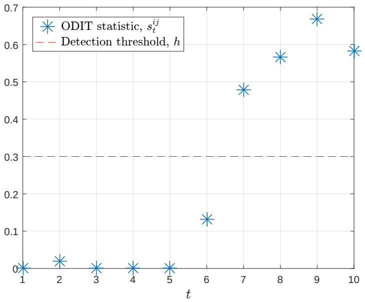
Online Discrepancy Test (ODIT)

- GEM lacks the temporal aspect
- In GEM, $\boldsymbol{x}_t$ is outlier if $L_t = \sum_{l=k^*}^{k} |e_{i(l)}|^\gamma > L_{(K)}$
- In ODIT, $D_t = L_t - L_{(K)}$ is treated as some positive/negative evidence for anomaly
- D_t approximates $\ell_t = \log \frac{p(r(\boldsymbol{X}_t)|\mathrm{H}_1)}{p(r(\boldsymbol{X}_t)|\mathrm{H}_0)}$ between H_1 claiming $\boldsymbol{x}_t$ is anomalous and H_0 claiming $\boldsymbol{x}_t$ is nominal



Online Discrepancy Test (ODIT)

- GEM lacks the temporal aspect
- In GEM, $\boldsymbol{x}_t$ is outlier if $L_t = \sum_{l=k^*}^{k} |e_{i(l)}|^\gamma > L_{(K)}$
- In ODIT, $D_t = L_t - L_{(K)}$ is treated as some positive/negative evidence for anomaly
- D_t approximates $\ell_t = \log \frac{p(r(\boldsymbol{X}_t)|\mathrm{H}_1)}{p(r(\boldsymbol{X}_t)|\mathrm{H}_0)}$ between H_1 claiming $\boldsymbol{x}_t$ is anomalous and H_0 claiming $\boldsymbol{x}_t$ is nominal



  - Assuming independence, $\sum_{t=1}^{T} D_t$ gives aggregate anomaly evidence until time T (as $\sum_{t=1}^{T} \ell_t$, sufficient statistic for optimum detection)
  - Similar to CUSUM (optimum minimax sequential change detector), ODIT decides using

$$T_d = \min\{t : s_t \geq h\}, \qquad s_t = \max\{s_{t-1} + D_t, 0\}$$

Theoretical Justification - Asymptotic

Asymptotic Optimality - Scalarized problem

As training set grows ($N_2 \to \infty$) ODIT is asymptotically optimum for

$$\mathrm{H}_0 : r(\boldsymbol{x}_t) \sim f_0^k, \forall t$$
$$\mathrm{H}_1 : r(\boldsymbol{x}_t) \sim f_0^k, t < \tau, \quad \text{and} \quad r(\boldsymbol{x}_t) \sim f_{uni}^k, t \geq \tau$$

- $\{\boldsymbol{x}_t\}$ independent
- $r(\boldsymbol{x}_t)$ kNN distance
- $f_0(\boldsymbol{x}_t) > 0$ Lebesgue continuous
- f_0^k and f_{uni}^k distributions of kNN distance under f_0 and uniform distr. on a d-dimensional grid with spacing r_α where $\int_{r_\alpha}^{\infty} f_0^k(r)\mathrm{d}r = \alpha$

Sketch of the Proof

- For independent $\{\boldsymbol{x}_t\}$, continuous $f_0 > 0$ defines a non-homogeneous Poisson point process with continuous rate $\lambda(\boldsymbol{x}) > 0$.
- Obtain a homogeneous Poisson point process with rate k by defining a d-dimensional non-homogeneous grid with volume $k/\lambda(\boldsymbol{x})$ [6]
- For this homogeneous Poisson point process, nearest neighbor function is given by

$$D_{\boldsymbol{X}}(r^d) = k \frac{\mathrm{d}v_d(\boldsymbol{x}, r)}{\mathrm{d}r^d} e^{-kv_d(\boldsymbol{X}, r)}$$

- Under H_0, $r(\boldsymbol{x}_t) = r_t$ comes from f_0^k which can be computed using training set as L_t.
- Under H_1, $r(\boldsymbol{x}_t) = r_\alpha$ comes from f_{uni}^k which has a single atom at r_α, computed as $L_{(K)}$.
- As training set grows, $L_t \to r_t$ and $L_{(K)} \to r_\alpha$
- The optimum CUSUM test computes $\log \frac{D_{\boldsymbol{X}}(r_\alpha)}{D_{\boldsymbol{X}}(r_t)} = kc(r_t^d - r_\alpha^d)$

[6] Robert Gallager. 6.262 Discrete Stochastic Processes, Chapter 2. Spring 2011. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Theoretical Justification - Nonasymptotic

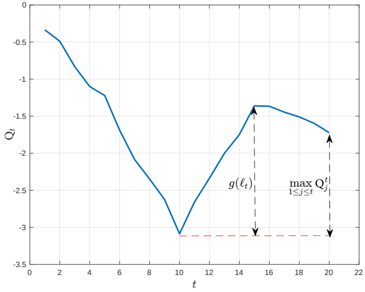
- CUSUM procedure can be expressed in terms of a general discrepancy metric, applicable to any number sequence
  - stop when discrepancy $g(\ell_t)$ [7] of observations with respect to f_0 is large enough

Discrepancy and CUSUM

$$T_c = \min\{t : g(\ell_t) \geq h_c\},$$

$$\ell_t = \left[\log \frac{f_1(\boldsymbol{x}_1)}{f_0(\boldsymbol{x}_1)} \dots \log \frac{f_1(\boldsymbol{x}_t)}{f_0(\boldsymbol{x}_t)}\right],$$

$$g(\ell_t) = \max_{1 \leq n_1 \leq n_2 \leq t} \sum_{i=n_1}^{n_2} \ell_t^i,$$

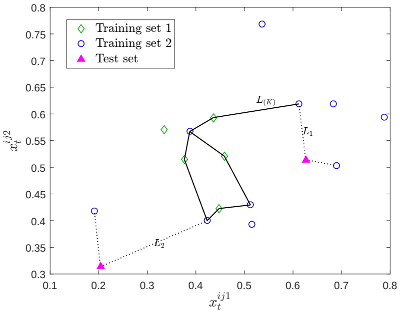


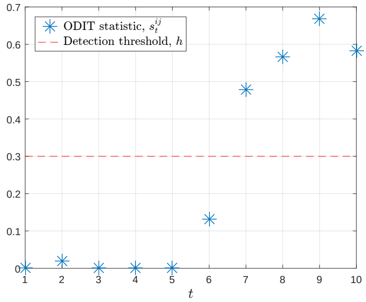
$$Q_t = \sum_{i=1}^{t} \ell_t$$

[7] B. A. Moser et al., "On stability of distance measures for event sequences induced by level-crossing sampling", IEEE Trans. Signal Process., vol. 62, no. 8, pp. 1987–1999, 2014.

ODIT Algorithm

- Initialize: $s \leftarrow 0$, $t \leftarrow 1$
- Partition training set into $\mathcal{X}^{N_1}$ and $\mathcal{X}^{N_2}$
- Determine $L_{(K)}$ from K-kNN graph $\bar{\mathcal{X}}_K^{N_1}$
- While $s < h$
  - Get new data $\boldsymbol{x}_t$ and compute $D_t = L_t - L_{(K)}$
  - $s = \max\{s + D_t, 0\}$
  - $t \leftarrow t + 1$
- Declare anomaly

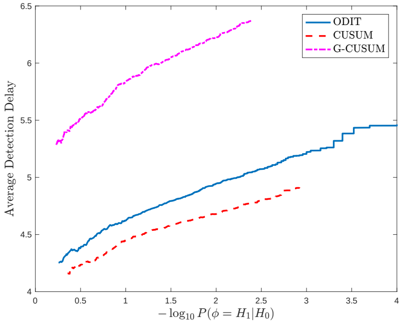




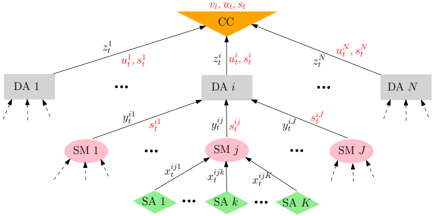
Numerical Results

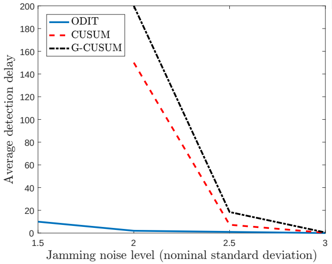
Simulations

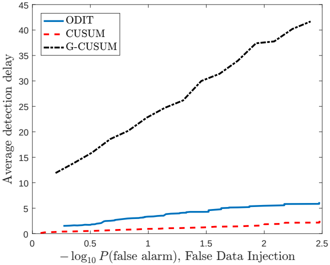
- f_0 is a 2D independent Gaussian with zero mean and $\sigma = 0.1$
- $f_1 = 0.8 f_0 + 0.2 U[0, 1]$
- Training set $10,000$ points ($N_1 = 1000$, $N_2 = 9000$)
- $\alpha = 0.05$, $k = 1$, $K = \alpha N_1$
- Parametric clairvoyant CUSUM knows both f_0 and f_1 exactly
- Generalized CUSUM exactly knows f_0, but estimates the uniform distribution upper bound as 0.9



Cybersecurity in Smart Grid





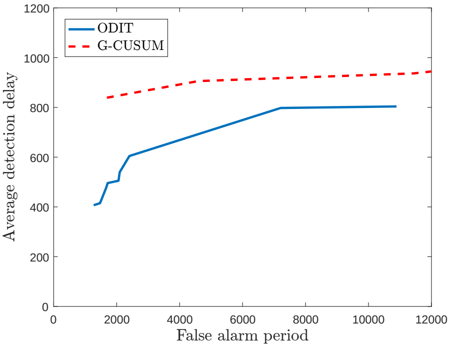


- Control center, 10 data aggregators, $1,000$ smart meters, $10,000$ smart appliances
- 3% of the HANs are attacked. In each attacked HAN, each smart appliance is attacked with prob. 0.5
- Baseline iid $\sim \mathcal{N}(0.5, 0.1^2)$
- Attack data either
  $\sim \mathcal{N}(0.5, (0.1\eta)^2), \quad \eta > 1$ (Jamming) or
  $\sim \mathcal{N}(0.5+\Delta, 0.1^2), \quad \Delta \in \mathbb{R}$ (False Data Injection)
- Even a small mismatch between the actual and assumed parameter values degrade the performance of CUSUM

Human Activity Recognition

- Online monitoring of a dynamic system using "Heterogeneity Human Activity Recognition Dataset" [8] obtained from the UCI Machine Learning Repository

- Smartwatch accelerometer data: 3.5M data points with 5 numeric features
- 6 activities: biking, sitting, standing, walking, stair up, and stair down
- Focusing on activity transitions we tested online detection performance
- G-CUSUM fits multivariate Gaussian models to baseline and anomalous dist.
- Re-train after detecting a change in the activity ($N_1 = 10$, $N_2 = 20$)



[8]A. Stisen et al., "Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition," SenSys, 2015.

Conclusion

Conclusions

- With the proliferation of IoT devices, and the ease of triggering DoS attacks even from unsophisticated malicious parties, there is an increasing need for developing scalable and effective solutions.
- A novel anomaly detection framework
  - Scalable: applicable to high-dimensional datasets (big data problems)
  - Nonparametric: agnostic to data-type and protocol
  - Online system monitoring
  - Asymptotically optimum for testing against uniformly distributed anomalies
- Outperforms sequential change detector CUSUM that estimates parameters from data
- Outperforms even clairvoyant CUSUM in case of a small to moderate variance increase (e.g., Jamming attack)

Questions?

Thank you!