Conformal Anomaly Detection in Event Sequences

AMSS Academy of Mathematics and Systems Science, CAS

Shuai Zhang Chuan Zhou Yang Liu Peng Zhang Xixun Lin Shirui Pan

ICML International Conference On Machine Learning

TL; DR

We propose a novel conformal anomaly detection method for event sequences, which combines two newly designed non-conformity scores with provably valid p-values for hypothesis testing.

Problem Statement

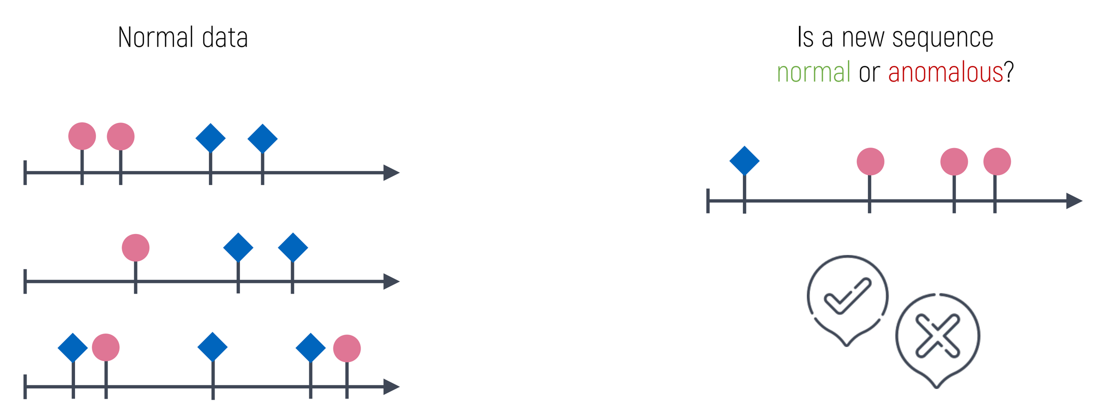

Figure 1. An illustration of anomaly detection in event sequences. Source: [1].

Anomaly detection in event sequences is a crucial task in safety-critical applications. For example, rapid information spreading may signal rumors, abnormal transaction activities may reveal fraud, and irregular patient health records may suggest rare medical conditions.

This task can be formulated as a hypothesis testing for temporal point processes (TPPs):

$$H_0: \ X_{\text{test}} \sim P_X \qquad \text{vs.} \qquad H_1: \ X_{\text{test}} \nsim P_X,$$

where P_X denotes some unknown data-generating TPP.

Preliminary

Theorem 1 (Time-rescaling theorem [2]). Let $X = \{(t_i, m_i)\}_{i=1}^{N}$ be a sequence of random event points on the interval $[0, T]$ corresponding to a TPP $\{N_m(t)\}_{m=1}^{M}$ with conditional intensity functions (CIFs) $\{\lambda_m^*(t)\}_{m=1}^{M}$. For each $m \in \mathcal{M} = \{1, \ldots, M\}$, denote the events of type-m from $N_m(t)$ as $X^{(m)} = (t_1^{(m)}, \ldots, t_{N_m(T)}^{(m)})$, where the number of events satisfies $\sum_{m=1}^{M} N_m(T) = N$. If each $\lambda_m^*(t)$ is positive on $[0, T]$ and $\Lambda_m^*(T) = \int_0^T \lambda_m^*(s)\,\mathrm{d}s < \infty$ almost surely, then for each $m \in \mathcal{M}$, the transformed sequence

$$Z^{(m)} = \left(\Lambda_m^*(t_1^{(m)}), \ldots, \Lambda_m^*(t_{N_m(T)}^{(m)})\right)$$

forms a standard Poisson process (SPP, i.e., the Poisson process with unit rate) on $[0, \Lambda_m^*(T)]$. Moreover, the sequences $\{Z^{(m)}\}_{m=1}^{M}$ are independent.

Proposition 1 [3]. For the SPP, conditionally on the event count in $[0, V]$ being equal to N, the normalized arrival times $\tau_1/V, \ldots, \tau_N/V$ are independently and uniformly distributed on $[0, 1]$.

Proposition 2 [3]. The inter-event times $w_i = \tau_i - \tau_{i-1}$ in the SPP are independent and follow an Exponential(1) distribution.

References

[1] O. Shchur, "Modeling continuous-time event data with temporal point processes." `https://shchur.github.io/assets/pdf/2021-10-12-btsa-presentation.pdf`, 2021.

[2] D. J. Daley, D. Vere-Jones, *et al.*, "An introduction to the theory of point processes: volume i: elementary theory and methods," Springer, 2003.

[3] P. A. Lewis, "Some results on tests for poisson processes," *Biometrika*, 1965.

[4] V. Vovk, A. Gammerman, and G. Shafer, "Algorithmic learning in a random world," Springer, 2005.

Proposed Method: CADES

Non-Conformity Scores for Event Sequences

Based on the above time-rescaling theorem and two properties of the SPP, we propose two non-conformity scores for event sequences:

$$s_{\text{arr}}(X) := D_{\text{KL}}(f_{\text{arr}} \| \hat{f}_{\text{arr}}) = \int_{-\infty}^{\infty} f_{\text{arr}}(x) \log\left(\frac{f_{\text{arr}}(x)}{\hat{f}_{\text{arr}}(x)}\right) \mathrm{d}x,$$

where $f_{\text{arr}}(x) = \mathbb{1}_{[0,1]}(x)$ is the uniform PDF, and $\hat{f}_{\text{arr}}(x) = \frac{1}{h_1 N} \sum_{i=1}^{N} \phi(\frac{x - \tau_i/V}{h_1})$ is the kernel density estimation (KDE) of the normalized arrival times τ_i/V of Z, obtained by applying time-rescaling and concatenation to X.

$$s_{\text{int}}(X) := D_{\text{KL}}(f_{\text{int}} \| \hat{f}_{\text{int}}) = \int_{-\infty}^{\infty} f_{\text{int}}(x) \log\left(\frac{f_{\text{int}}(x)}{\hat{f}_{\text{int}}(x)}\right) \mathrm{d}x,$$

where $f_{\text{int}}(x) = \mathrm{e}^{-x} \mathbb{1}_{[0,\infty)}(x)$ is the exponential PDF, and $\hat{f}_{\text{int}}(x) = \frac{1}{h_2(N+1)} \sum_{i=1}^{N+1} \phi(\frac{x - w_i}{h_2})$ is the KDE of the inter-event times $w_i = \tau_i - \tau_{i-1}$ of Z.

Test Procedure with Bonferroni Correction

For the proposed score s_{arr}, the classical conformal p-value [4] of X_{test} is computed as:

$$p_{\text{arr}}^r(X_{\text{test}}) = \frac{|\{X_{\text{cal}} \in \mathcal{D}_{\text{cal}} : s_{\text{arr}}(X_{\text{test}}) \le s_{\text{arr}}(X_{\text{cal}})\}| + 1}{n_{\text{cal}} + 1}.$$

As we show that both small and large values of the proposed score can indicate OOD sequences, we utilize the two-sided p-value:

$$p_{\text{arr}}(X_{\text{test}}) = 2 \min\{p_{\text{arr}}^l(X_{\text{test}}),\ p_{\text{arr}}^r(X_{\text{test}})\}.$$

To leverage the complementary sensitivities of s_{arr} and s_{int} to different abnormal patterns, we combine them for OOD detection using the Bonferroni corrected p-value:

$$p_{\text{cor}}(X_{\text{test}}) = \min\{2(1+\varepsilon)p_{\text{arr}}(X_{\text{test}}), 2(1+\varepsilon)p_{\text{int}}(X_{\text{test}})\}.$$

Algorithm 1: CADES: Conformal Anomaly Detection in Event Sequences

Input: Clean dataset $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}}$, test sequence X_{test}, target level $\alpha \in (0, 1)$.

1: Train a neural TPP with CIFs $\{\lambda_m^*(t)\}_{m=1}^{M}$ on $\mathcal{D}_{\text{train}}$;

2: Apply time-rescaling and concatenation for $\mathcal{D}_{\text{cal}}$ and X_{test};

3: Calculate the scores s_{arr} and s_{int} for $\mathcal{D}_{\text{cal}}$ and X_{test};

4: Compute the Bonferroni corrected p-value $p_{\text{cor}}(X_{\text{test}})$.

Output: Declare X_{test} as OOD or abnormal if $p_{\text{cor}}(X_{\text{test}}) \le \alpha$, otherwise as ID or normal.

Theoretical Guarantees

Proposition 3 (Marginal false positive rate (FPR) control). Suppose the test sequence X_{test} and the dataset $\mathcal{D}$ are i.i.d. (or, more generally, exchangeable), then the p-value $p_{\text{cor}}(X_{\text{test}})$ is valid, i.e., for every $\alpha \in (0, 1)$, $\mathbb{P}_{H_0}(p_{\text{cor}}(X_{\text{test}}) \le \alpha) \le \alpha$.

Theorem 2 (Calibration-conditional FPR control). Let $\alpha, \delta \in (0, 1)$ and $\varepsilon \ge 0$. Let $\mathcal{D}_{\text{cal}}$ be a calibration set of size n_{cal}, $a = \lfloor (n_{\text{cal}} + 1) \frac{\alpha}{4(1+\varepsilon)} \rfloor$, $b = n_{\text{cal}} + 1 - a$, and $\mu = \frac{a}{a+b}$. For a given $\delta > 0$, let n_{cal} be such that

$$I_{(1+\varepsilon)\mu}(a, b) \ge 1 - \frac{\delta}{4},$$

where $I_x(a, b)$ denotes the CDF of the Beta(a, b) distribution. If $s_{\text{arr}}(X)$ and $s_{\text{int}}(X)$ are continuously distributed, then for a new sequence X_{test}, the probability of incorrectly identifying X_{test} as OOD conditioned on $\mathcal{D}_{\text{cal}}$ while using Algorithm 1 is bounded by α with probability $1 - \delta$, i.e.,

$$\mathbb{P}[\mathbb{P}_{H_0}(\text{declare OOD} \mid \mathcal{D}_{\text{cal}}) \le \alpha] \ge 1 - \delta.$$

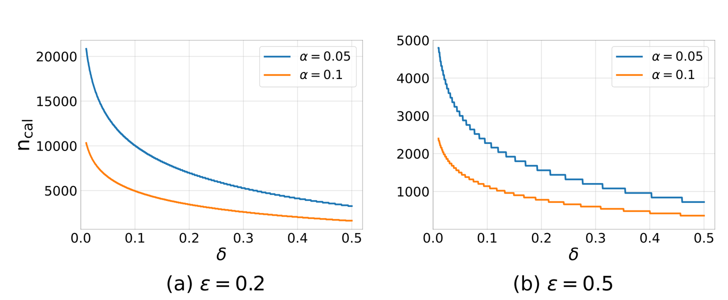

Figure 2. Calibration set size n_{cal} that guarantees the calibration-conditional FPR is bounded by α with probability $1 - \delta$.

Experimental Results

Goodness-of-Fit (GOF) Test for SPP

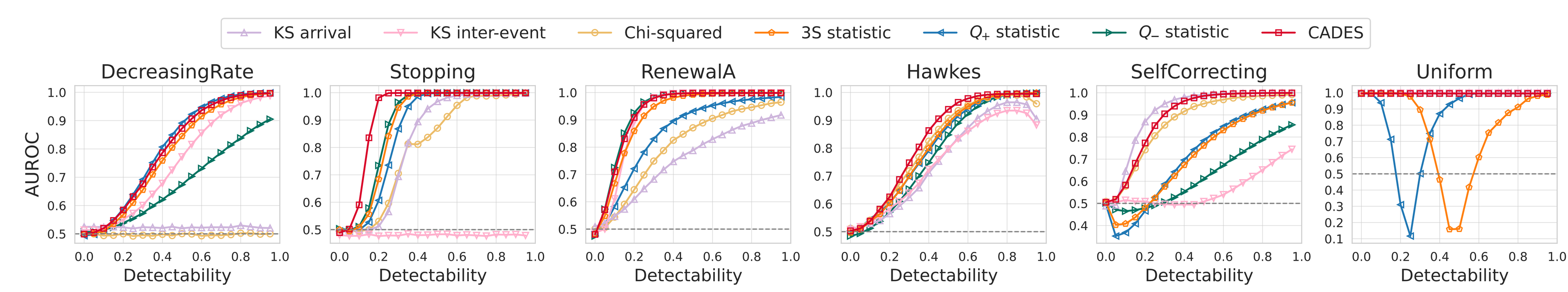

Figure 3. Performance of GOF test for the SPP measured by AUROC (higher is better).

Detecting Anomalies in Synthetic Data

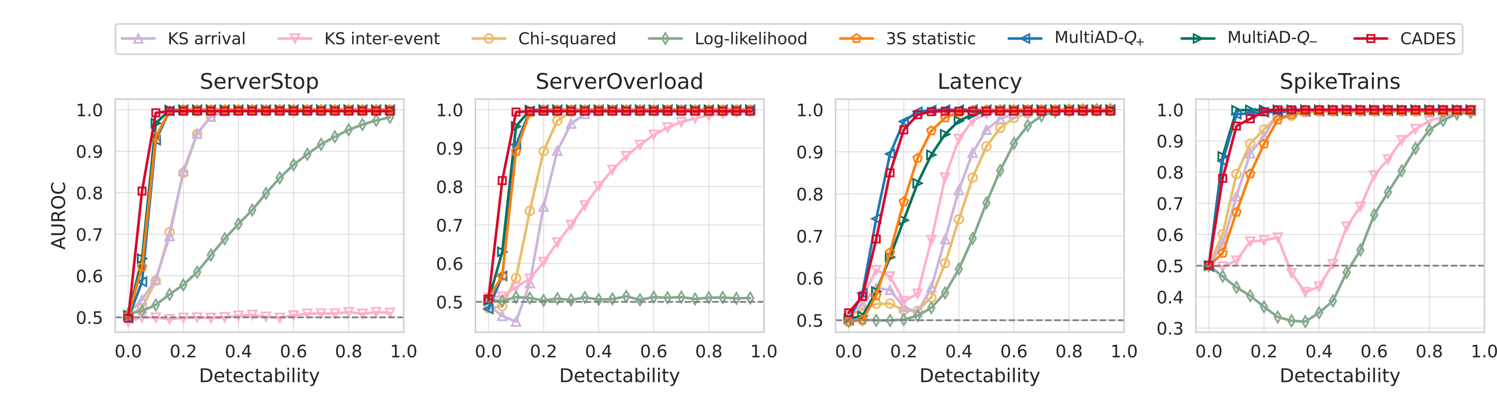

Figure 4. Performance of OOD detection on synthetic datasets measured by AUROC.

Detecting Anomalies in Real-World Data

Table 1. AUROC (%) for OOD detection on real-world datasets. Best results are in **bold** and second best are <u>underlined</u>.

Dataset	KS arrival	KS inter-event	Chi-squared	Log-likelihood	3S statistic	MultiAD-Q_+	MultiAD-Q_-	CADES (ours)
LOGS - Packet corruption (1%)	47.24	71.80	67.27	90.92	95.03	92.44	**96.61**	<u>96.48</u>
LOGS - Packet corruption (10%)	64.96	98.72	49.35	98.98	99.30	99.31	**99.53**	<u>99.48</u>
LOGS - Packet duplication (1%)	61.88	79.59	21.26	81.97	<u>91.46</u>	91.24	78.15	**92.88**
LOGS - Packet delay (frontend)	90.31	47.46	95.70	**99.55**	96.10	97.97	95.27	<u>98.15</u>
LOGS - Packet delay (all services)	95.13	96.60	94.35	96.30	99.16	**99.59**	99.31	<u>99.33</u>
STEAD - Anchorage, AK	62.31	78.44	70.75	88.16	91.73	84.00	<u>99.16</u>	**99.31**
STEAD - Aleutian Islands, AK	53.37	86.48	64.17	97.08	99.80	<u>99.86</u>	99.84	**99.95**
STEAD - Helmet, CA	61.94	98.83	73.62	96.96	93.82	70.71	<u>99.13</u>	**99.30**
Average Rank	7.50	5.63	7.00	4.25	3.63	3.50	<u>3.00</u>	**1.50**

FPR Control

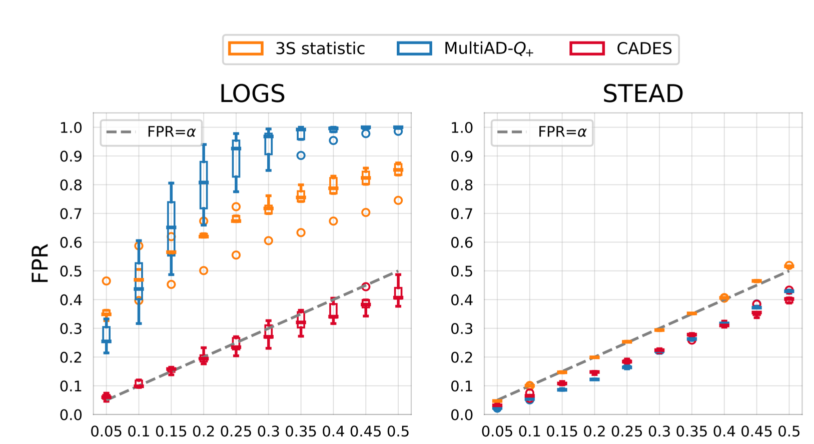

Figure 5. Boxplots of FPR for OOD detection on two real-world datasets under different target levels α.

Table 2. TPR (%) for OOD detection on the STEAD dataset under the target level $\alpha = 0.05$.

Dataset	3S statistic	MultiAD-Q_+	CADES
STEAD - Anchorage, AK	74.30	67.14	**95.46**
STEAD - Aleutian Islands, AK	**100**	**100**	**100**
STEAD - Helmet, CA	69.20	6.50	**98.38**

Email: zhangshuai2021@amss.ac.cn