Nominality Score Conditioned Time Series Anomaly Detection by Point/Sequential Reconstruction

Chih-Yu Lai*
Department of EECS, MIT
Cambridge, MA 02139
chihyul@mit.edu

Fan-Keng Sun
Department of EECS, MIT
Cambridge, MA 02139
fankeng@mit.edu

Zhengqi Gao
Department of EECS, MIT
Cambridge, MA 02139
zhengqi@mit.edu

Jeffrey H. Lang
Department of EECS, MIT
Cambridge, MA 02139
lang@mit.edu

Duane S. Boning
Department of EECS, MIT
Cambridge, MA 02139
boning@mtl.mit.edu

目录

摘要

Abstract

Time series anomaly detection is challenging due to the complexity and variety of patterns that can occur. One major difficulty arises from modeling time-dependent relationships to find contextual anomalies while maintaining detection accuracy for point anomalies. In this paper, we propose a framework for unsupervised time series anomaly detection that utilizes point-based and sequence-based reconstruction models. The point-based model attempts to quantify point anomalies, and the sequence-based model attempts to quantify both point and contextual anomalies. Under the formulation that the observed time point is a two-stage deviated value from a nominal time point, we introduce a *nominality score* calculated from the ratio of a combined value of the reconstruction errors. We derive an *induced anomaly score* by further integrating the nominality score and anomaly score, then theoretically prove the superiority of the induced anomaly score over the original anomaly score under certain conditions. Extensive studies conducted on several public datasets show that the proposed framework outperforms most state-of-the-art baselines for time series anomaly detection.

问题：对时间相关关系进行建模以发现上下文异常，同时保持点异常检测的准确性。

解决方法：提出了一种利用基于点和基于序列的重建模型的无监督时间序列异常检测框架，并根据重建误差的组合值的比率计算的名义分数，由名义分数和异常分数来得出诱导异常分数。

效果：理论上证明在一定条件下诱导异常得分相对于原始异常得分的优越性。对多个公共数据集进行的广泛研究表明，所提出的框架优于大多数最先进的时间序列异常检测基线。

引言

问题描述:

时间异常包括点异常和上下文异常。点异常是单个数据点显着偏离时间序列预期行为的情况,而上下文异常是指在特定上下文或条件下偏离预期行为的数据点。

在模型试图学习时间相关关系以检测上下文异常的同时，可能会失去对点异常的精度或准确度，这涉及到一个重要的权衡问题。在高维数据中，时间关系的建模变得更加困难，因此使用基于序列的重建模型可能会面临点异常和上下文异常检测之间的权衡，导致重建结果嘈杂且性能欠佳。

引言

解决方法

本文首先尝试了基于点的重建方法，但发现它们表现出较低的方差，因为这些方法不需要对时间相关关系进行建模。尽管缺乏时间信息，基于点的重建模型的异常分数已经展现出有竞争力的性能。

为了弥合基于点的模型和基于序列的模型之间的差距，我们引入了一种名义分数，通过该分数和原始异常分数计算得出诱导异常分数。发现诱导的异常分数可以优于原始的异常分数。通过点/顺序重建（NPSR）创造了名义分数条件时间序列异常检测的方法。

方法

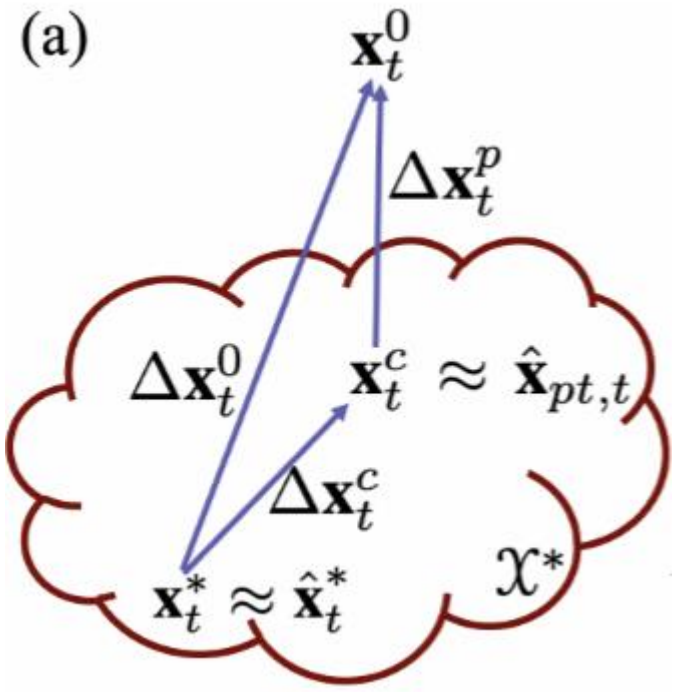
3.1 Problem Formulation

令 $\mathrm{X} = \{x_1, \cdots, x_T\}$表示 $x_t \in R^D$ 的多元时间序列，其中 T 是时间长度，D 是维度或通道数。存在一组对应的标签 $\mathrm{y} = \{y_1, \cdots, y_T\}$，$y_t \in \{0,1\}$ 指示时间点是正常$(y_t = 0)$还是异常$(y_t = 1)$。

对于给定的 X，目标是产生所有时间点的异常分数 $\mathrm{a} = \{a_1, \cdots, a_T\}$，在 $a_t \in R$ 和相应的阈值 θ_a 处，使得预测标签 $\hat{y} = \{\hat{y}_1, \cdots, \hat{y}_T\}$，其中 $\hat{y}_t \triangleq 1_{a_t \geq \theta_a}$，尽可能匹配 y。

为了量化 $\hat{y}$ 和 y 的匹配程度，或者 a 是否有可能产生良好的 $\hat{y}$，这项工作主要关注没有点调整的最佳 F1 分数（F1*），也称为逐点 F1 分数，其定义为考虑所有阈值的最大可能 F1 分数。

方法 3.2 Nominal Time Series and Two-stage Deviation：



将观察到的数据表示为 $X^0 = \{x_1^0, \ldots, x_T^0\}$，其中 $x_t^0 \in R^D$。假设对于每个 X^0，存在相应的基础名义时间序列数据 $X^* = \{x_1^*, \ldots, x_T^*\}$，该数据来自名义时间相关过程 $x_t^* = f^*(t): N \to R^D$。t 处相应的总偏差 (Δx_t^0)定义为 $\Delta x_t^0 \triangleq x_t^0 - x_t^*$。

将 χ^*表示为所有 $t \in \{1, \cdots, T\}$ 的所有可能的 x_t^* 的集合。Δx_t^0 可以分为两个加性因子 Δx_t^{c}和 Δx_t^{p}，因此，根据定义，$= x_t^* + \Delta x_t^{\mathrm{c}}$ 和 $x_t^0 = x_t^{\mathrm{c}} + \Delta x_t^{\mathrm{p}}$ 我们将 Δx_t^{c}定义为分布内偏差，其中 $x_t^{\mathrm{c}} \in \chi^*$；$\Delta x_t^{\mathrm{p}}$为分布外偏差，当且仅当 $x_t^0 \notin \chi^*$ 时，它才非零。

Δx_t^{p}可以是量化点异常，并且 Δx_t^{c} 可以是量化上下文异常。因为无论 Δx_t^{c} 有多大，我们仍然有 $x_t^{\mathrm{c}} \in \chi^*$，即分布内偏差值 x_t^{c} 仍然在所有可能的名义时间点数据的集合中，并且无法使用基于点的模型进行检测。另一方面，学习了 χ^*后，基于点的模型可能可以抵消由 Δx_t^{p} 引起的偏差值。

方法 3.2 Nominal Time Series and Two-stage Deviation：

例子

假设我们从 2D 位置传感器的流数据中获得一个数据集，其中 $x_t^*, x_t^c, x_t^0 \in R^2$，并且我们了解到名义时间序列是一个点围绕原点以一定角度的圆周运动速度 ω 和半径 r，其中

$R_{\min} \le \mathrm{r} \le R_{\max}$。因此，我们可以推导出

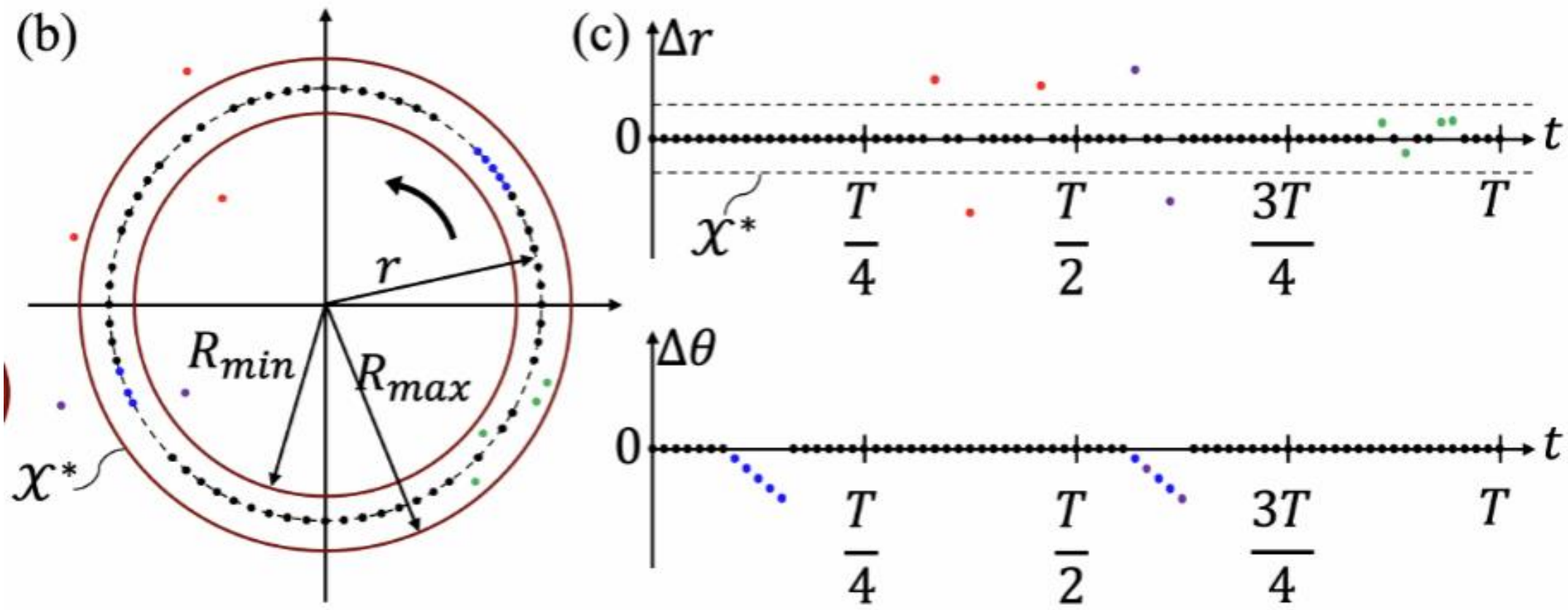
$$x^* = \left\{[xy]^T \middle| R_{\min}^2 \le x^2 + y^2 \le R_{\max}^2\right\} \text{ 和 } x_t^* = [r\cos\omega t \;\; r\sin\omega t]^T \text{。}$$

设角速度的意外变化作为上下文异常的一个可能原因。例如，系统故障可能导致 t1 和 t2 之间的圆周运动减慢，即 $\Delta x_t^c = [r(\cos\omega' t - \cos\omega t) \;\; r(\sin\omega' t - \sin\omega t)]^T$ 对于 $t \in \{t_1, \cdots, t_2\}$ 且其它地方 $\Delta x_t^c = 0$ 。

各个时间点的噪声测量可能会导致观察到的时间序列中的点异常，即 $\Delta x_t^P = \left[\omega_{x,t} \; \omega_{y,t}\right]^T$，其中$\omega_{x,t}$ 或 $\omega_{y,t}$ 不为零，使得对于一些 t，$x_t^0 \notin \chi^*$ 。

方法 3.2 Nominal Time Series and Two-stage Deviation：

黑点是 $x_t^0 = x_t^*$ 的时间点（无异常）。蓝点是表现出减速的时间点（上下文异常）。红点是具有噪声测量值的时间点（点异常）。紫色点是具有减速和噪声测量的时间点（点和上下文异常）。绿点是具有噪声测量的时间点，但 $x_t^0 \in \chi^*$，因此仍然是上下文异常，因为它们的偏差无法通过观察单个时间点来检测。



方法

3.3 The Nominality Score

名义分数 $N(\cdot)$ 表示某个时间点的正常程度。

如果对于每个可能的正常阈值 $\theta_N > 0$，对于所有的 $t \in \{1, \cdots, T\}$ 有

$$P(N(t) > \theta_N | y_t = 0) > P(N(t) > \theta_N | y_t = 1)。$$

名义分数大于 θ_N 的正常点部分严格大于名义分数大于 θ_N 的异常点部分。

在本文，将 N(t) 定义为 x_t^c 和 x_t^0 之间的 L2 范数平方之比。

$$N(t) \triangleq \frac{\|\Delta \mathbf{x}_t^c\|_2^2}{\|\Delta \mathbf{x}_t^0\|_2^2} = \frac{\|\Delta \mathbf{x}_t^c\|_2^2}{\|\Delta \mathbf{x}_t^c + \Delta \mathbf{x}_t^p\|_2^2} = \frac{\|\mathbf{x}_t^c - \mathbf{x}_t^*\|_2^2}{\|\mathbf{x}_t^0 - \mathbf{x}_t^*\|_2^2}$$

方法

3.3 The Nominality Score

示例：在此推导中，我们在下标中添加 n 和 a 分别表示仅与正常点和异常点相关的变量。考虑一个玩具数据集，其中 $\Delta x^c_{t,n}, \Delta x^p_{t,n}, \Delta x^c_{t,a}, \Delta x^P_{t,a}$ 已定义：

$$\Delta \mathbf{x}^c_{t,n} \sim \mathcal{N}(0, I_D), \quad \Delta \mathbf{x}^p_{t,n} \sim \mathcal{N}(0, I_D), \quad \Delta \mathbf{x}^c_{t,a} \sim \mathcal{N}(0, I_D), \quad \Delta \mathbf{x}^p_{t,a} \sim \mathcal{N}(0, \alpha^2 I_D)$$

$$2N_n(t) = 2\frac{\|\Delta \mathbf{x}^c_{t,n}\|_2^2}{\|\Delta \mathbf{x}^c_{t,n} + \Delta \mathbf{x}^p_{t,n}\|_2^2} \sim \mathrm{F}(D, D)$$

$$(1+\alpha^2)N_a(t) = (1+\alpha^2)\frac{\|\Delta \mathbf{x}^c_{t,a}\|_2^2}{\|\Delta \mathbf{x}^c_{t,a} + \Delta \mathbf{x}^p_{t,a}\|_2^2} \sim \mathrm{F}(D, D)$$

$2N_n(t) = 2\frac{\|\Delta \mathbf{x}^c_{t,n}\|_2^2}{\|\Delta \mathbf{x}^c_{t,n} + \Delta \mathbf{x}^p_{t,n}\|_2^2} \sim \mathrm{F}(D, D)$

$\mathbb{P}(N(t) > \theta_N | y_t = 0)$

$2N(t) > 2\theta_N$

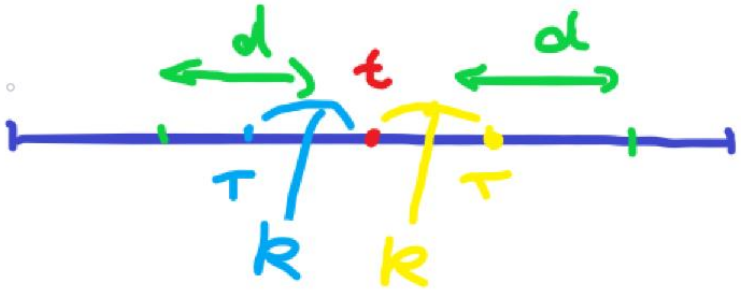
$\sim F(D,D) > 2\theta_N$

其中 F 是具有 D 和 D 自由度的 F 分布。如果 α > 1，则 $N(\cdot)$ 成为适当的名义分数，因为

$$\mathbb{P}(N(t) > \theta_N | y_t = 0) = \int_{2\theta_N}^{\infty} f(x; D, D)dx > \int_{(1+\alpha^2)\theta_N}^{\infty} f(x; D, D)dx = \mathbb{P}(N(t) > \theta_N | y_t = 1)$$

其中 $f(\cdot; D, D)$ 是自由度为 D 和 D 的 F 分布的概率密度函数。事实上，可以合理地假设 $\Delta x^P_{t,a}$ 的方差大于 $\Delta x^p_{t,n}$。

方法



3.4 The Induced Anomaly Score

用于整合任何给定的 $N(\cdot)$ 和异常分数 $A(\cdot)$，以产生诱导异常分数 $\hat{A}(\cdot)$，其中性能将优于使用 $A(\cdot)$或平滑的 $A(\cdot)$。对于两个相近的时间点 t 和 τ，假设 t 异常的可能性受到 τ 的影响。将这种效应量化为 $A(t;\tau)$，它是 t 由于 τ 引起的异常得分。通过对 t 周围的 τ 范围求和，我们得到 t 处的诱发异常分数：

$$\hat{A}(t) \triangleq \sum_{\tau=\max(1,t-d)}^{\min(T,t+d)} A(t;\tau)$$

其中 d 是感应长度。$A(t;\tau)$ 定义为 $A(\tau)$ 的门控值，它由从 τ 到 t 的名义分数控制：

$$A(t;\tau) \triangleq A(\tau) \prod_{k=\min(\tau+1,t)}^{\max(t-\mathbb{1}_{t=\tau},\tau-1)} g_{\theta_N}(N(k)) = \begin{cases} A(\tau)g_{\theta_N}(N(\tau+1))...g_{\theta_N}(N(t)) & t>\tau \\ A(\tau) & t=\tau \\ A(\tau)g_{\theta_N}(N(\tau-1))...g_{\theta_N}(N(t)) & t<\tau \end{cases}$$

其中门函数 $g_{\theta_N}(N)$ 是 N 的某个变换函数，以阈值 θ_N 为条件。

方法

3.4 The Induced Anomaly Score

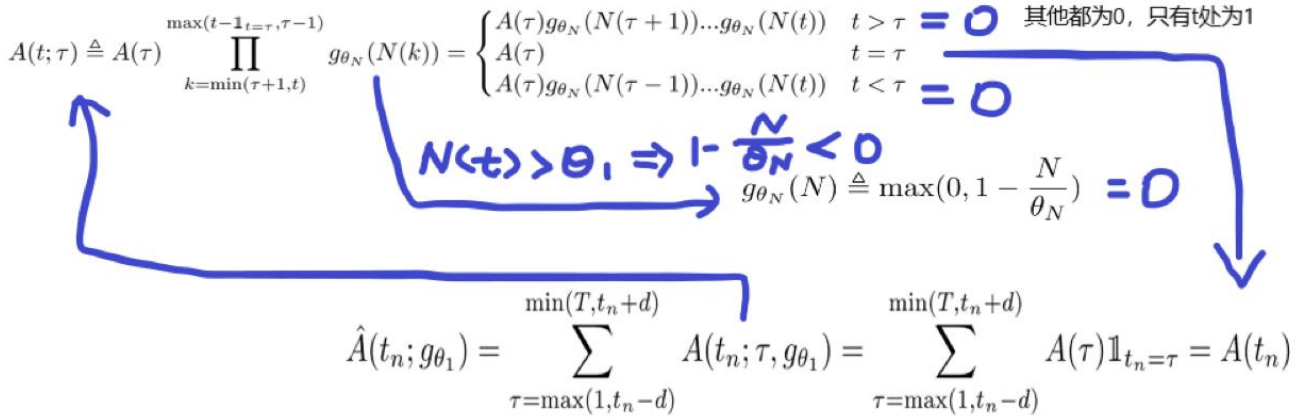
我们可以使用 $\hat{A}(t) = \hat{A}(t; g_{\theta_N})$和 $A(t;\tau) = A(t;\tau, g_{\theta_N})$ 来表示这些值以 g_{θ_N} 为条件。$\hat{A}(\cdot)$ 可以被认为是 $A(\cdot)$ 的一些（非标准化）加权平滑值，其中权重是某个范围内 $g_{\theta_N}(N(\cdot))$ 的乘积。我们考虑以下两种情况：

Claim 1 Using a soft gate function

$$g_{\theta_N}(N) \triangleq \max(0, 1 - \frac{N}{\theta_N})$$

如果对于所有正常点 $(y_t = 0)$ 存在 θ_1 使得 $N(t) \geq \theta_1$，则 $F1^*(\hat{A}(\cdot; g_{\theta_1}); y) \geq F1^*(A(\cdot); y)$，即使用以 g_{θ_1}作为门函数的诱导异常得分的最佳 F1 得分大于或等于使用原始异常得分的最佳 F1 得分。

方法



3.4 The Induced Anomaly Score Proof 1

对于任何正常时间点 t_n，我们有

$$\hat{A}(t_n; g_{\theta_1}) = \sum_{\tau=\max(1,t_n-d)}^{\min(T,t_n+d)} A(t_n; \tau, g_{\theta_1}) = \sum_{\tau=\max(1,t_n-d)}^{\min(T,t_n+d)} A(\tau)\mathbb{1}_{t_n=\tau} = A(t_n)$$

对于任意异常点t_a，我们有

$$\hat{A}(t_a; g_{\theta_1}) = \sum_{\tau=\max(1,t_a-d)}^{\min(T,t_a+d)} A(t_a; \tau, g_{\theta_1}) \geq A(t_a)$$

对于异常点，通过潜在地使 Aˆ (t) 高于 A(t)，我们得到潜在更高的 F1* 。

方法

3.4 The Induced Anomaly Score

Claim 2 Using a hard gate function $g_{\theta_N}(N) \triangleq \mathbb{1}_{N<\theta_N}$

如果 d = 1，并且存在两个阈值:

(i) θ_2，使得所有异常时间点 $(y_t = 1)$的 $N(t) < \theta_2$

(ii) $\theta_\infty = \infty$，则 $F1^*\left(\hat{A}\left(\cdot; g_{\theta_2}\right); y\right) \geq F1^*\left(\hat{A}\left(\cdot; g_{\theta_\infty}\right); y\right)$，即以 g_{θ_2} 作为门函数的诱导异常得分的最佳 F1 得分大于或等于使用以 g_{θ_∞} 作为门函数的诱发异常得分的最佳 F1 得分。

方法

3.4 The Induced Anomaly Score

Proof 2

对于任意一个异常时间点t_a，我们有

$$\hat{A}(t_a; g_{\theta_2}) = \sum_{\tau=\max(1,t_a-1)}^{\min(T,t_a+1)} A(\tau) = A(t_a-1)\mathbb{1}_{t_a>1} + A(t_a) + A(t_a+1)\mathbb{1}_{t_a<T}$$

其中第一个等式源自 $g_{\theta_2}\big(N(t_a)\big) = 1$。

对于任何正常时间点 t_n，我们有

$$\hat{A}(t_n; g_{\theta_2}) = \sum_{\tau=\max(1,t_n-1)}^{\min(T,t_n+1)} A(t_n; \tau, g_{\theta_2}) \leq A(t_n-1)\mathbb{1}_{t_n>1} + A(t_n) + A(t_n+1)\mathbb{1}_{t_n<T}$$

因为 $N(t_n)$可能大于 θ_2，因此 $g_{\theta_2}\big(N(t_n)\big) \leq 1$，但是，我们有

$$\hat{A}(t; g_{\theta_\infty}) = \sum_{\tau=\max(1,t-1)}^{\min(T,t+1)} A(\tau) = A(t-1)\mathbb{1}_{t>1} + A(t) + A(t+1)\mathbb{1}_{t<T}$$

方法

3.4 The Induced Anomaly Score

Proof 2

对于任何点，我们有

$$\hat{A}(t; g_{\theta_\infty}) = \sum_{\tau=\max(1,t-1)}^{\min(T,t+1)} A(\tau) = A(t-1)\mathbb{1}_{t>1} + A(t) + A(t+1)\mathbb{1}_{t<T}$$

无论正常点还是异常点，因为对于任何 t，$N(t) < \theta_\infty$。因此，由于 $\hat{A}(t_a; g_{\theta_2}) = \hat{A}(t_a; g_{\theta_\infty})$ 且 $\hat{A}(t_n; g_{\theta_2}) \leq \hat{A}(t_n; g_{\theta_\infty})$，因此与使用 θ_∞ 相比，使用 θ_2 时我们得到的 F1* 可能更高。

$\hat{A}(\cdot; g_{\theta_\infty})$可以被视为 $A(\cdot)$上的平滑值（或平移简单移动平均），周期为 2d + 1。

Claim 2 意味着通过对 $N(\cdot)$ 进行调节并计算 $\hat{A}(t)$，与使用 $A(\cdot)$的简单平滑值相比，可以提高性能。在实践中，我们可以放宽 d 的约束，并使用其他门函数来产生更灵活的架构。

方法

3.5 Point-based Reconstruction Models

考虑一些模型 M_{pt}，它逐点重建每个时间点 t：$\hat{x}_{pt,t} \triangleq M_{pt}\left(x_t^0\right)$。使用 M_{pt} 时，产生异常分数的方法是使用基于点的重建均方误差，定义为$a_{Pt} = \left\{a_{Pt,1}, \cdots, a_{Pt,\mathrm{T}}\right\}$，其中 $a_{pt,t} \triangleq \left\|\hat{x}_{pt,t} - x_t^0\right\|_2^2$。由于 M_{pt} 学习捕获所有法线点数据的分布，因此我们假设 $\hat{x}_{pt,t} \in \chi^*$ 或非常接近。此外，由于 $\hat{x}_{pt,t}$可以抵消点异常，因此我们假设 $x_t^c \approx \hat{x}_{pt,t}$。

$$\mathbf{X}^c = \{\mathbf{x}_1^c, ..., \mathbf{x}_T^c\} \approx \hat{\mathbf{X}}^c = \{\hat{\mathbf{x}}_1^c, ..., \hat{\mathbf{x}}_T^c\}, \quad \hat{\mathbf{x}}_t^c = \hat{\mathbf{x}}_{pt,t} = \mathcal{M}_{pt}(\mathbf{x}_t^0)$$

M_{pt} 可以是任何能够逐点重建 x^0的模型。这里通过使用一个简单的基于 Performer 的自动编码器可以实现最佳性能，该自动编码器实际上具有发现时间信息的潜力。

由于基于 Performer 的自动编码器尝试在一批时间点上进行优化，这减少了过度拟合的影响，并允许模型更好地泛化到未见过的数据。

方法

3.6 Sequence-based Reconstruction Models

与x_t^c相反，x_t^*不仅应该在 χ^*中，而且还服从时间相关关系。因此，逼近x_t^*的模型（M_{seq}）有必要以时间点序列作为输入。

$$\mathbf{X}^* \approx \hat{\mathbf{X}}^* = \{\hat{\mathbf{x}}_1^*, \ldots, \hat{\mathbf{x}}_T^*\} \triangleq \mathcal{M}_{seq}(\mathbf{X}^0)$$

使用基于 Performer 的堆叠编码器作为 M_{seq}，它从周围的 2γ 点预测中间 δ 点，以强制学习时间相关关系。连接 M_{seq} 输出的所有预测时间点来构造 $\hat{x}^*$。

方法

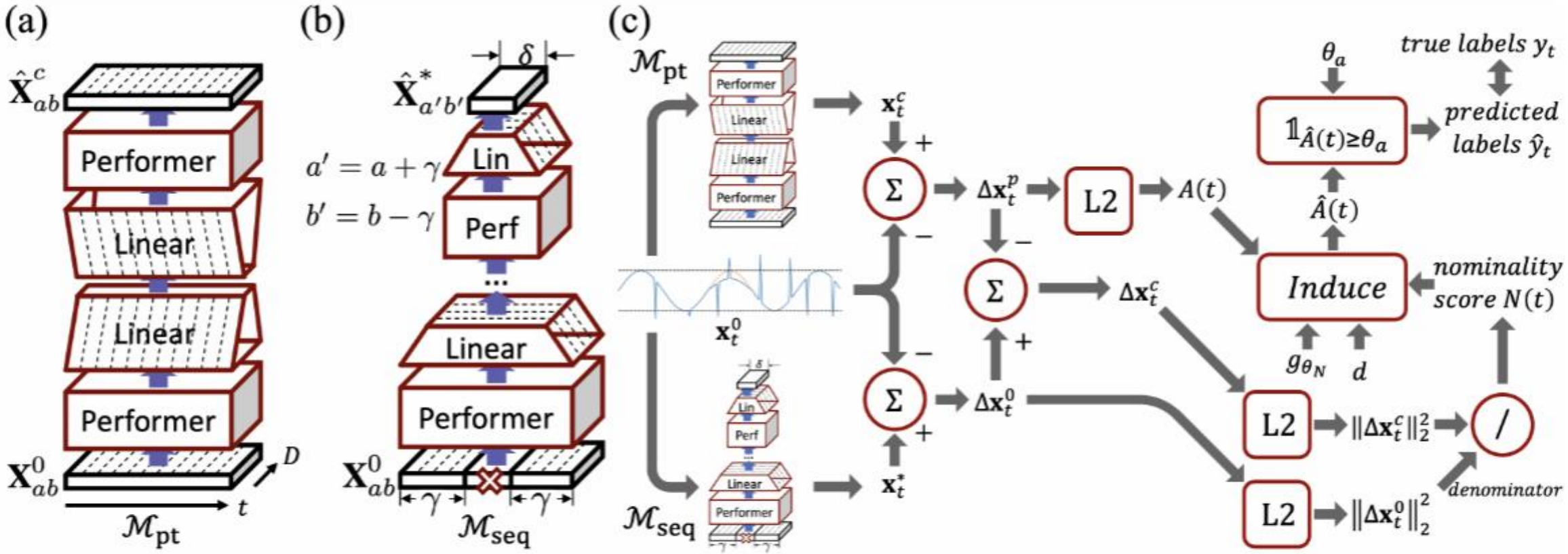


Figure 3: (a) Performer-based autoencoder $\mathcal{M}_{pt}$, (b) Performer-based stacked encoder $\mathcal{M}_{seq}$, and (c) main scheme for NPSR. GELUs are used as the activation function for each layer.

方法

Algorithm 1 NPSR $\mathrm{F1}^*$ Evaluation (soft gate function)

function NPSR($\mathcal{M}_{pt}, \mathcal{M}_{seq}, \mathbf{X}^0 = \{\mathbf{x}_1^0, ..., \mathbf{x}_T^0\}, \mathbf{y} = \{y_1, ..., y_T\}, \theta_N, d$)

Construct $\hat{\mathbf{X}}^c = \{\hat{\mathbf{x}}_1^c, ..., \hat{\mathbf{x}}_T^c\}$ with $\hat{\mathbf{x}}_t^c \leftarrow \mathcal{M}_{pt}(\mathbf{x}_t^0)$ ▷ (15)

Construct $\hat{\mathbf{X}}^* = \{\hat{\mathbf{x}}_1^*, ..., \hat{\mathbf{x}}_T^*\} \leftarrow \mathcal{M}_{seq}(\mathbf{X}^0)$ ▷ (16)

Construct $A(\cdot)$ with $A(t) \leftarrow \|\hat{\mathbf{x}}_t^c - \mathbf{x}_t^0\|_2^2$ ▷ section 3.5

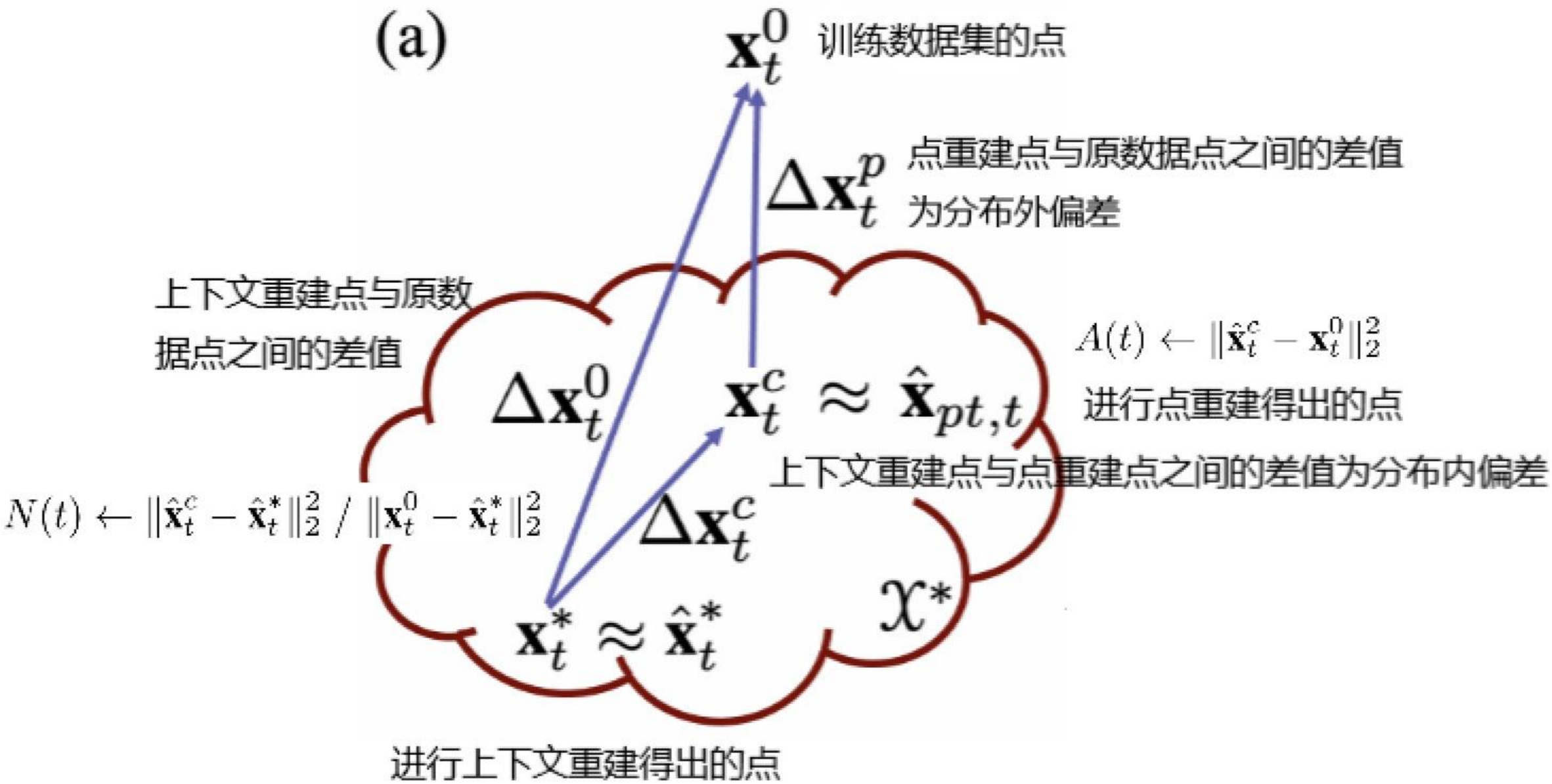
Construct $N(\cdot)$ with $N(t) \leftarrow \|\hat{\mathbf{x}}_t^c - \hat{\mathbf{x}}_t^*\|_2^2 \ / \ \|\mathbf{x}_t^0 - \hat{\mathbf{x}}_t^*\|_2^2$ ▷ (1)

Construct $g_{\theta_N}(N(\cdot))$ with $g_{\theta_N}(N(t)) \leftarrow \max(0, 1 - N(t)/\theta_N)$ ▷ (8)

Construct $A(\cdot;\cdot)$ with $A(t;\tau) \leftarrow A(\tau) \prod_{k=\min(\tau+1,t)}^{\max(t-\mathbb{1}_{t=\tau},\tau-1)} g_{\theta_N}(N(k))$ ▷ (7)

Construct $\hat{A}(\cdot)$ with $\hat{A}(t) \leftarrow \sum_{\tau=\max(1,t-d)}^{\min(T,t+d)} A(t;\tau)$ ▷ (6)

return $\mathrm{F1}^* \leftarrow \max_{\theta_a} \mathrm{F1}(\hat{\mathbf{y}}(\hat{A}(\cdot), \theta_a); \mathbf{y})$


(a)
$\mathbf{x}_t^0$ 训练数据集的点
$\Delta\mathbf{x}_t^p$ 点重建点与原数据点之间的差值为分布外偏差
上下文重建点与原数据点之间的差值
$\Delta\mathbf{x}_t^0$
$\mathbf{x}_t^c \approx \hat{\mathbf{x}}_{pt,t}$
$A(t) \leftarrow \|\hat{\mathbf{x}}_t^c - \mathbf{x}_t^0\|_2^2$
进行点重建得出的点
上下文重建点与点重建点之间的差值为分布内偏差
$N(t) \leftarrow \|\hat{\mathbf{x}}_t^c - \hat{\mathbf{x}}_t^*\|_2^2 \,/\, \|\mathbf{x}_t^0 - \hat{\mathbf{x}}_t^*\|_2^2$
$\Delta\mathbf{x}_t^c$
$\mathcal{X}^*$
$\mathbf{x}_t^* \approx \hat{\mathbf{x}}_t^*$
进行上下文重建得出的点

实验

数据

SWaT（安全水处理）：SWaT 数据集是在 11 天内从具有 51 个传感器的小型水处理测试台收集的。在过去 4 天内，使用不同的攻击方法注入了 41 个异常，而在前 7 天内仅生成正常数据。

WADI（水分配测试台）：WADI 数据集是从一个精简的城市供水系统获取的，该系统有 123 个传感器和执行器，运行了 16 天。前 14 天仅包含正常数据，其余两天有 15 个异常段。

PSM（池化服务器指标）：PSM 数据集是从 eBay 的多个应用程序服务器节点内部收集的。有 13 周的训练数据和 8 周的测试数据。

MSL（火星科学实验室）和 SMAP（土壤湿度主动被动）：MSL 和 SMAP 数据集是 NASA 收集的公共数据集，包含来自航天器监测系统事件意外异常（ISA）报告的遥测异常数据。数据集分别有 55 和 25 维。训练集包含未标记的异常。

SMD（服务器机器数据集）：SMD是从一家大型互联网公司收集的，包含来自28台服务器机器和38个传感器的5周的数据。前 5 天仅包含正常数据，最后 5 天间歇性注入异常数据。

rimSyn（修剪合成数据集）：原始合成数据集是使用三角函数和高斯噪声生成的。我们从获取数据集并修剪测试数据集，使得仅存在一段异常。

实验

我们使用 F1* 对照多种深度学习算法和简单启发式方法评估 NPSR 的性能。

NPSR 利用 M_{pt} 精确捕获点异常，误报率较低（给定最佳阈值）。此外，它通过$\hat{A}(\cdot)$的计算合并M_{seq}，获得了检测上下文异常的能力，同时又不影响其检测点异常的能力。

Table 2: Best F1 score (F1*) results on several datasets, with bold text denoting the highest and underlined text denoting the second highest value. The deep learning methods are sorted with older methods at the top and newer ones at the bottom.

Algorithm \ Dataset	SWaT	WADI	PSM	MSL	SMAP	SMD	trimSyn
Simple Heuristic [11, 30, 31]	0.789	0.353	0.509	0.239	0.229	0.494	0.093
DAGMM [26]	0.750	0.121	0.483	0.199	0.333	0.238	0.326
LSTM-VAE [22]	0.776	0.227	0.455	0.212	0.235	0.435	0.061
MSCRED [24]	0.757	0.046	0.556	0.250	0.170	0.382	0.340
OmniAnomaly [9]	0.782	0.223	0.452	0.207	0.227	0.474	0.314
MAD-GAN [23]	0.770	0.370	0.471	0.267	0.175	0.220	0.331
MTAD-GAT [27]	0.784	0.437	0.571	0.275	0.296	0.400	<u>0.372</u>
USAD [28]	0.792	0.233	0.479	0.211	0.228	0.426	0.326
THOC [18]	0.612	0.130	-	0.190	0.240	0.168	-
UAE [11]	0.453	0.354	0.427	<u>0.451</u>	0.390	0.435	0.094
GDN [12]	<u>0.810</u>	<u>0.570</u>	0.552	0.217	0.252	<u>0.529</u>	0.284
GTA [41]	0.761	0.531	0.542	0.218	0.231	0.351	0.256
Anomaly Transformer [40]	0.220	0.108	0.434	0.191	0.227	0.080	0.049
TranAD [25]	0.669	0.415	**0.649**	0.251	0.247	0.310	0.282
NPSR (combined)	-	-	-	0.261	**0.511**	0.227	-
NPSR	**0.839**	**0.642**	<u>0.648</u>	**0.551**	<u>0.505</u>	**0.535**	**0.481**

实验

对于多实体数据集，我们观察到标准方法（每个实体训练一个基于点和基于序列的模型）优于 MSL 和 SMD 数据集的组合方法。鉴于实体之间的差异可能很大，这并不奇怪。然而，对于 SMAP 数据集，我们观察到组合方法表现更好。我们将这样的结果归因于 SMAP 航天器是常规的，因此实体之间的遥测结果可以具有相似的基础分布。这有助于从增加的训练数据中进行附加学习。

实验

$$g_{\theta_N}(N) \triangleq \max(0, 1 - \frac{N}{\theta_N}) \qquad g_{\theta_N}(N) \triangleq \mathbb{1}_{N<\theta_N}$$

4.4 Ablation Study

比较了产生不同异常分数（$A(\cdot)$ 或 $\hat{A}(\cdot)$）的五种方法的性能。对于前两种方法，分别使用Mpt和Mseq的重构误差。由于 Mpt 的表现大多优于 Mseq，因此我们使用基于点的重建误差 $A(\cdot)$ 来计算后三种方法的$\hat{A}(\cdot)$ 。第三种方法对应于$A(\cdot)$ 的非归一化简单平滑值。第四种和第五种方法分别使用门函数 (11) 和 (8)，但相同的 θ_N 对应于训练数据 ($N_{\text{tr}\,n}$) 中名义分数的 98.5 百分位。

Table 3: AUC and F1* for different methods and datasets, with bold text denoting the highest and underlined text denoting the second highest value. The mean (μ_d) and standard deviation (σ_d) of the performance metrics evaluated across $d = 1, 2, 4, 8, 16, 32, 64, 128, 256$ are shown.

Dataset		SWaT		WADI		PSM		MSL		SMAP		SMD		trimSyn	
Method		AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*
$\mathcal{M}_{pt}$ ($\|\hat{\mathbf{x}}_t^c - \mathbf{x}_t^0\|_2^2$)		0.908	**0.839**	0.819	0.629	<u>0.790</u>	<u>0.626</u>	0.640	0.366	0.647	0.329	0.820	0.485	0.721	0.100
$\mathcal{M}_{seq}$ ($\|\hat{\mathbf{x}}_t^* - \mathbf{x}_t^0\|_2^2$)		0.899	0.755	0.843	0.559	0.766	0.576	0.621	0.351	0.611	0.292	0.820	0.482	<u>0.832</u>	<u>0.345</u>
$\mathcal{M}_{pt}$ + Hard (11)	μ_d	**0.912**	0.813	0.827	<u>0.630</u>	0.775	0.621	<u>0.708</u>	0.451	**0.665**	**0.389**	<u>0.835</u>	<u>0.492</u>	0.785	0.144
($\theta_N = \infty$)	σ_d	0.005	0.034	0.007	0.037	0.023	0.020	0.032	0.038	0.010	0.036	0.025	0.052	0.037	0.021
$\mathcal{M}_{pt}$ + Hard (11)	μ_d	**0.912**	0.820	<u>0.844</u>	0.625	0.779	0.624	**0.718**	**0.467**	<u>0.659</u>	0.386	0.833	0.495	0.791	0.292
($\theta_N = 98.5\%N_{trn}$)	σ_d	0.005	0.024	0.007	0.023	0.017	0.015	0.041	0.051	0.012	0.034	0.024	0.050	0.069	0.121
$\mathcal{M}_{pt}$ + Soft (8)	μ_d	0.909	<u>0.837</u>	**0.856**	**0.639**	**0.804**	**0.636**	0.698	<u>0.465</u>	0.656	<u>0.388</u>	**0.840**	**0.525**	**0.862**	**0.434**
($\theta_N = 98.5\%N_{trn}$)	σ_d	0.000	0.001	0.011	0.008	0.005	0.004	0.031	0.061	0.005	0.039	0.003	0.011	0.063	0.099

实验

4.4 Ablation Study

首先，我们观察到 Mpt 优于 Mseq，并且尽管没有对时间相关关系进行建模，但它自己也取得了有竞争力的结果。其次，通过平滑 A(·)，大多数数据集的 AUC 和 F1* 都会增加。这意味着平滑通常是提高性能的有效方法。此外，我们的实验表明，软门函数和适当的 θN 在 F1* 方面平均表现最好。这表明名义分数的分布主要是重叠的，并且软门函数将更适合防止正常时间点异常分数的过度积累，从而减少误报。该方法在很宽的 d 范围内也具有总体稳定的 AUC 和 F1*（低 σd）。这是有道理的 - 当时间点 τ 距离时间点 t 较远时，更多的门函数输出乘到 A(τ) 上，因此 A(t, τ) → 0。然而，结果还表明门的最佳选择函数和 θN 可能取决于手头的具体数据集。

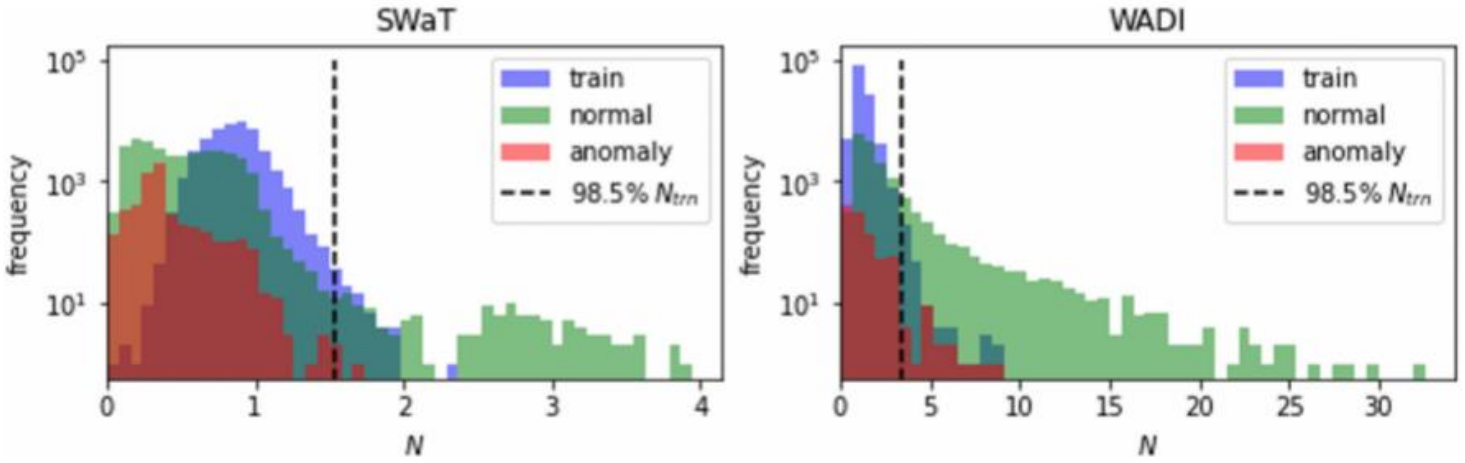


Figure 4: Histograms of the nominality scores for the SWaT and WADI dataset.

实验

4.5 Detection Trade-off Between Point and Contextual Anomalies

我们详细阐述了检测点和上下文异常之间的权衡，并将它们与 Mpt 和 Mseq 的性能联系起来。从时间序列中查找时间信息会带来更好的性能，这似乎很直观。然而，建模复杂性随着考虑的时间点数量的增加而增加，这导致难以专注于单个时间点的重建。图显示了使用 Mpt 或 Mseq 计算的 $A(\cdot)$，以及使用 WADI 数据集通过 NPSR 计算的 $\hat{A}(\cdot)$。考虑到如果 A(t) ⩾ θa，则 t 点被预测为异常，Mseq 由于尖峰而具有较高的误报率。相比之下，Mpt 以逐点方式更准确地检测异常。这凸显了 Mpt 在此数据集上相对于 Mseq 的优越性。此外，NPSR计算出的诱发异常分数具有非常低的误报率，有时甚至可以学会识别Mpt或Mseq未检测到的异常。这表明该数据集的 $\hat{A}(\cdot)$优于原始 $A(\cdot)$。然而，Mpt 的性能可能并不总是优于 Mseq。假阴性可以在 t = 14800 和 t = 14900 之间可视化，其中 Mpt 难以识别异常，但 Mseq 可以有效检测异常的时间依赖性关系。这表明异常段包含比点异常相对更多的上下文。由于Mpt的重构误差被用作 $A(\cdot)$，我们失去了有效利用Mseq的重构误差的优势。这导致 $\hat{A}(\cdot)$不够高，无法在该段内达到 θa。未来的一个重要方向是探索如何在多个模型中适当地选择$A(\cdot)$。

实验

4.5 Detection Trade-off Between Point and Contextual Anomalies

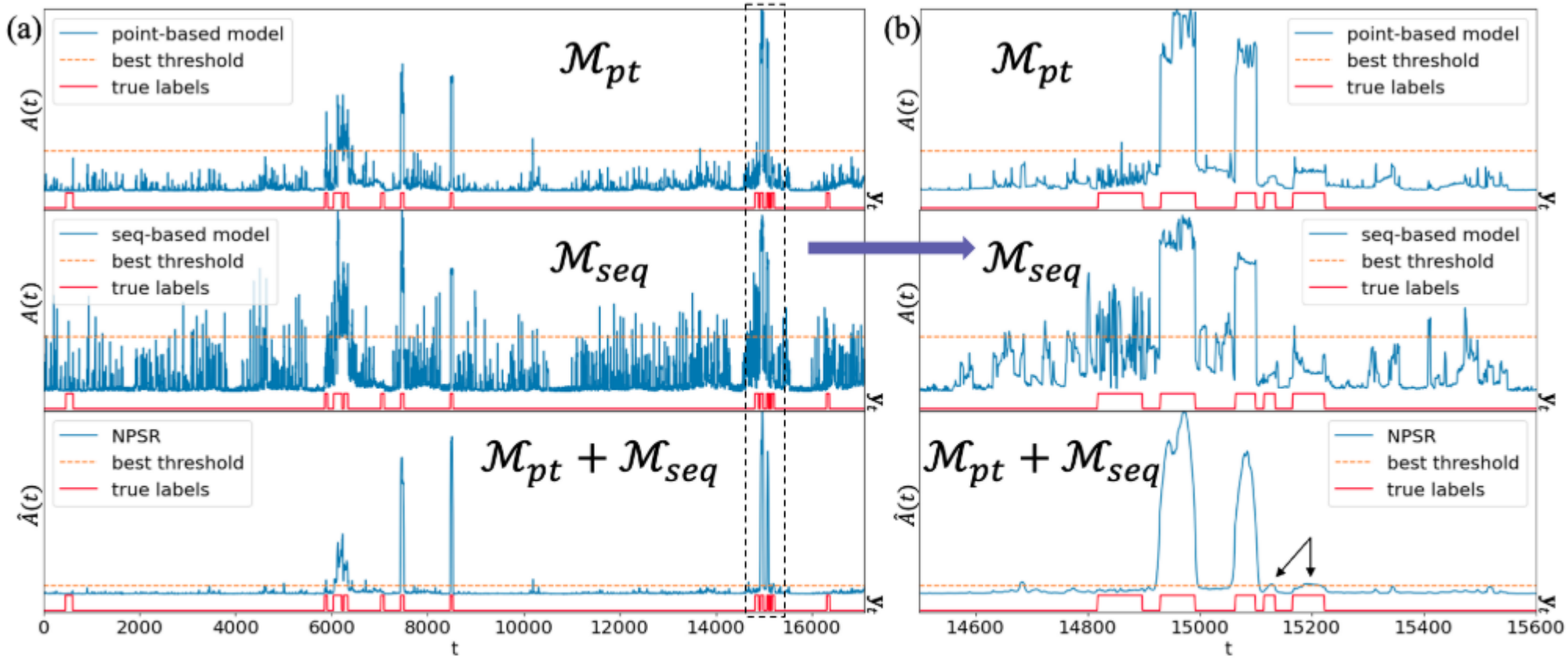


Figure 5: (a) Anomaly scores using $\mathcal{M}_{pt}$, $\mathcal{M}_{seq}$ and NPSR (soft gate function, $\theta_N = 99.85\% N_{trn}$, and $d = 16$), and the true labels of the WADI dataset. (b) Magnification for $t \in \{14500, ..., 15600\}$.

总结

引入了一种用于无监督时间序列异常检测的改进框架。指定点和上下文异常之间的关系，并得出名义分数和诱导异常分数，以提供具有可证明优越性的基于理论的算法。 NPSR 可捕获点异常和上下文异常，从而实现较高的组合精度和召回率。

结果表明，NPSR 表现出高性能、广泛适用且训练过程相对简单。它有可能减少故障监控的劳动力需求，相应地加快决策速度，还可以通过防止能源浪费或系统故障来促进人工智能的可持续性。

总结

限制

训练中使用的基于点的模型不包含时间信息，这使得有效重建低维数据集具有挑战性。

对于单变量时间序列来说尤其具有挑战性，因为原始输入不适用于基于点的模型。

该问题的一种可能的解决方案是通过聚合多个时间点来增加维度。然而，这种方法的有效性还有待证实。

NPSR 的另一个限制是缺乏自动阈值 (θa) 查找方法，这使得在部署模型时很难确定合适的阈值。为了解决这个问题，可以定义目标误报率，并使用验证集估计实现该目标率的阈值，因为只需要正常数据。

异常检测中如果某时间部分包含比点异常相对更高比例的上下文异常。当使用 Mpt 时，无法有效检测异常，但当使用 Mseq 时，可以获得了更高的异常分数。当前的方法利用Mpt的重建误差作为异常分数计算的基础，从而忽略了Mseq生成的重建误差的有效性。