计算机科学
COMPUTER SCIENCE

基于时空注意力机制的多元时间序列异常检测

梁李芳, 关东海, 张吉, 袁伟伟

引用本文

梁李芳, 关东海, 张吉, 袁伟伟. 基于时空注意力机制的多元时间序列异常检测[J]. 计算机科学, 2023, 50(11A): 230300022-8.

LIANG Lifang, GUAN Donghai, ZHANG Ji, YUAN Weiwei. Spatial-Temporal Attention Mechanism Based Anomaly Detection for Multivariate Times Series [J]. Computer Science, 2023, 50(11A): 230300022-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300022

基于时空注意力机制的多元时间序列异常检测

梁李芳[1]　关东海[1]　张　吉[2]　袁伟伟[1]
1 南京航空航天大学计算机科学与技术学院　南京 211106
2 澳大利亚南昆士兰大学　昆士兰 图文巴 4350
(2249945728@qq.com)

摘　要　物联网系统被广泛应用于各种基础设施,系统中涉及许多相互连接的传感器,这些传感器产生大量的多元时间序列数据。由于物联网系统容易遭受网络攻击,多元时间序列异常检测方法被用于及时监测系统中发生的异常,这对于保障系统安全至关重要。然而,由于高维传感器数据关系复杂,现有的大多数异常检测方法难以明确学习多元时间序列的相关性,导致异常检测的准确率较低。因此,提出一种基于时空注意力机制的多元时间序列异常检测方法(STA)。首先,以图形结构的形式学习传感器间的关系,再使用多跳图注意力网络为图中每个传感器节点的多跳邻居节点分配不同的注意力权重,用于捕捉序列的空间相关性。其次,采用基于长短时间记忆网络的时间注意力机制自适应地选择相应的时间序列,用于学习序列的时间相关性。在4个真实世界传感器数据集上的实验结果表明,STA可以比基线方法更准确地检验时间序列中的异常,其F_1分数分别优于最佳基线31.03%,14.29%,15.91%和21.74%。此外,消融实验和灵敏度分析验证了模型中的关键组件的有效性。总的来说,STA可以有效捕捉多元时间序列中的空间和时间相关性,提高模型的异常检测性能。

关键词:多元时间序列;注意力机制;图注意力网络;长短时间记忆网络;时间相关性;空间相关性;异常检测
中图法分类号　TP391

Spatial-Temporal Attention Mechanism Based Anomaly Detection for Multivariate Times Series

LIANG Lifang[1],GUAN Donghai[1],ZHANG Ji[2] and YUAN Weiwei[1]
1 School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China
2 University of Southern Queensland,Toowoomba,Queensland 4350,Australia

Abstract　Internet of Things systems are widely used in a variety of infrastructure,involving many interconnected sensors that generate large amounts of multivariate time series data. Since the Internet of Things systems are vulnerable to network attacks,multivariate time series anomaly detection methods are used to timely monitor anomalies occurring in the system,which is crucial for securing the system. However,due to the complex relationships of high-dimensional sensor data,most existing anomaly detection methods have difficulty in learning the correlation of multivariate time series explicitly,resulting in low accuracy of anomaly detection. Therefore,a multivariate time series anomaly detection method(STA) based on spatial and temporal attention mechanism is proposed. STA first learns the relationship between sensors in the form of a graph structure and then uses a multi-hop graph attention network to assign different attention weights to the multi-hop neighbor nodes of each sensor node in the graph for capturing the spatial correlation of the sequence. Secondly,STA use a temporal attention mechanism-based long short-term memory network to adaptively select the corresponding time sequences to study the temporal correlation of sequences. Experimental results on four real-world sensor datasets show that STA can detect anomalies in time series more accurately than the baseline approach,with its F_1 score outperforms the optimal baseline by 31.03%,14.29%,15.91% and 21.74%,respectively. In addition,ablation experiments and sensitivity analysis validate the effectiveness of the key components in the model. In general,STA can effectively capture the spatial and temporal correlations in multivariate time series and improve the anomaly detection performance of the model.

Keywords　Multivariate times series,Attention mechanism,Graph attention network,Long short-term memory network,Temporal correlation,Spatial correlation,Anomaly detection

1　引言

物联网作为一种新型网络系统,被广泛应用于工业控制系统、航天探测器、互联网服务器等各个领域,其通过配置许多相互关联的传感器监控系统的工作情况,这些传感器持续产生大量的多元时间序列数据。当系统受到攻击时,多元时间序列数据就会发生异常。然而,物联网系统经常受到各种安全威胁。为了能及时察觉系统故障,避免造成严重的经济

基金项目:国防基础科研计划(JCKY2020204C009)
This work was supported by the National Defense Basic Research Program(JCKY2020204C009).
通信作者:关东海(dhguan@nuaa.edu.cn)

损失，多元时间序列的异常检测受到了学术界和工业界的广泛关注。

随着物联网基础设施的快速发展，传感器数据变得庞大而复杂，手动监测数据中的异常已经远远无法满足需求。针对时间序列的自动异常检测方法应运而生。该方法可以帮助工作人员快速检测出高维数据中的异常，从而及时采取相应的措施，避免造成潜在的系统灾难。

由于大多数训练数据中很少甚至没有异常标签，人工标记异常标签将会导致巨大的时间和财力投入，异常检测算法通常采用无监督学习算法进行训练。与有监督学习相比，无监督学习还有一个潜在的优势，就是能够发现未曾出现过的异常行为。

传感器数据之间非线性相关，彼此相互影响。例如，由于工业控制系统的工作流程特性，安全水处理(Secure Water Treatment，SWaT)[1]数据集的所有传感器之间存在复杂的拓扑结构关系。正如，打开阀门会导致水流入或流出水箱，造成水箱中液位传感器的数据变化；同时，水箱中液位传感器等传感器也能够通过监控系统的状态，决定何时打开和关闭阀门。

由于传感器之间的复杂关系难以捕捉，现有的大多数方法[2-3]在对传感器中潜在的相互依赖关系进行建模时，无法明确哪些传感器之间相互影响。为了充分利用传感器的空间相关性，图形神经网络被用来建模传感器之间的依赖关系，其中包括图形卷积网络[4]和图注意力网络[5]。给定传感器节点间的拓扑图结构，图形卷积网络可以有效地提取拓扑图的空间特征，但无法处理动态图问题。此外，传感器间的相互关系应是非对称的，即其拓扑结构是有向图，图形卷积网络在处理有向图时存在瓶颈。图注意力网络的出现不仅解决了上述问题，还通过使用注意力系数分配不同的学习权重给不同的邻居节点，从而将节点特征之间的相关性更好地融入模型。图注意力网络及其变体已经在时间序列建模方面取得进展，如被用于道路网络交通流量预测[6-7]、时间序列异常检测[8]和知识推荐[9]等。

图注意力网络在时间序列异常检测方面的一个典型应用是图偏差网络(Graph Deviation Network，GDN)[8]。该方法学习传感器数据之间的关系图，并通过图注意力网络构建维度间的依赖关系图，能够有效学习序列的空间相关性，具有优异的异常检测性能。但该方法只考虑了传感器间的空间相关性，却忽略了多元时间序列的时间依赖性；且只关注传感器节点的单跳邻居节点，不考虑未与传感器节点直接相连但可造成间接影响的多跳邻居节点。

为了解决上述问题，本文提出的基于时空注意力机制的多元时间序列异常检测(Spatial-Temporal Attention Mechanism Based Anomaly Detection for Multivariate Times Series，STA)方法在图注意力网络中引入多跳邻居节点，考虑图结构中未直接相连节点的信息传播，并加入基于 LSTM[10]的时间注意力机制分析序列的时间依赖性。

本文的主要贡献如下：

1)提出一种基于时空注意力机制的无监督异常检测方法 STA，通过使用多跳图注意力网络捕捉序列间的复杂关系，并引入时间注意力机制建模序列的时间依赖性，有效提高了时间序列的异常检测性能。

2)在 4 个来自现代网络物理系统的数据集上评估 STA 方法。结果表明，STA 在 4 个数据集上的 F_1 分数分别优于最佳基线 31.03%，14.29%，15.91%和 21.74%。消融实验进一步证明了 STA 中不同组件的有效性。

3)灵敏度分析实验研究了时间序列窗口大小对 STA 异常检测性能的影响。结果表明，关键组件的引入能够提高 STA 异常检测的性能和稳定性。

2 相关工作

异常检测是时间序列数据挖掘的主要任务之一，用于从数据中确定异常值[11]。现有的深度学习时间序列异常检测方法大致可分为两类：基于重构的异常检测方法和基于预测的异常检测方法。

2.1 基于重构的异常检测方法

基于重构的异常检测方法通过对输入数据进行有损重建来检测异常。自动编码器(Autoencoder，AE)[12]是一种无监督的深度神经网络，基于反向传播算法和最优化方法训练模型，使用重建误差作为异常分数。其包括两个主要部分：编码器和解码器。编码器将输入数据编码为低维的隐变量来学习最有信息量的特征，解码器把隐藏层的隐变量还原到初始维度。自编码器是用于多元时间序列异常检测的许多基于深度学习的模型的基础[13]。LSTM-AE[14]是一种基于长短时间记忆网络的自动编码器模型，主要对时间依赖性进行建模。LSTM 是一种特殊的递归神经网络，能够学习长期依赖性。在自编码器中加入 LSTM，可以更好地学习序列的时间依赖性。但是该方法在同一时间只能对单个时间数据进行处理，导致模型训练速度慢。主成分分析[15]通过捕捉数据中的大部分方差对数据集进行降维，将原数据映射到低维特征空间，用低维特征重构原始数据，通过重构误差进行异常检测。该方法很难对时间序列的综合信息进行编码，缺乏对空间和时间依赖性的充分考虑。OmniAnomaly[16]是一种先验驱动随机模型，通过应用变分自编码器将时间序列信号建模为随机表示。其直接返回多元时间序列输入的后验重建概率[13]，如果给定输入的重建可能性低于阈值，则将其预测为异常[17]。该方法更加关注特征间的异常，缺乏对时间依赖性的关注。

2.2 基于预测的异常检测方法

基于预测的异常检测方法通过给定的历史时间序列预测下一个(或几个)时间点的数据，使用预测误差判断异常。LSTM 和无参数动态阈值方法[18]使用长短期记忆递归神经网络学习时间序列的长期时间依赖性，提高预测性能，使用无参数的阈值方法评估残差。TCNAE[19]是一种基于时间卷积网络(Temporal Convolutional Network)的自编码器模型。TCN 通过引入膨胀因果卷积，对时间序列进行特征提取，相较于 LSTM 和 GRU 等通用递归架构更有优势。使用 TCN 不仅可以避免递归网络的梯度消失等问题，还可以并行计算输出。但该方法主要对时间信息进行建模，难以捕捉特征间的复杂依赖关系；且由于卷积操作导致感受野受限，对全局数据关联性的捕获不足。GDN[8]使用图注意力网络构建传感器之间的依赖关系，能够有效学习多元时间序列的空间相关性，但是 GDN 不善于从时间序列中获取时间特征。

序列之间的复杂关系和每个序列时间依赖性都会影响多元时间序列的数值。基于重构的方法难以同时从特征和时间两个维度重构多维序列，本文采用基于预测的方法，由于复杂

多元时间序列的不可预测性，引入多跳图注意力网络和时间注意力机制学习时间序列的时空相关性。

3 STA

3.1 问题说明

训练数据表示为$x_{\text{train}}=[x^1,\cdots,x^{T_{\text{train}}}]$，$T_{\text{train}}$表示训练集的长度，$x^t\in R^N$表示任意时刻$t$处$N$维时间序列的值。测试数据表示为$x_{\text{test}}=[x^1,\cdots,x^{T_{\text{test}}}]$。训练数据中不存在异常，模型通过训练数据建模多元时间序列的构建过程，对测试数据进行异常检测。$y(t)\in\{0,1\}$用于指示每个测试时间刻度是否存在异常，$y(t)=1$表示t时刻发生异常。

采用基于预测的方法，在时刻t，根据具有特定窗口大小w的历史时间序列数据定义模型的输入：

$$I^t:=[x^{t-w},x^{t-w+1},\cdots,x^{t-1}] \tag{1}$$

$$I^t\in R^{N\times w} \tag{2}$$

模型的目标输出是t时刻每个传感器的预期数值，即$\tilde{x}^t$。

3.2 方法架构

本文提出一种基于时空注意力机制的多元时间序列异常检测方法(STA)，该方法的架构如图1所示。STA的架构包含3个部分：图结构学习、多跳图注意力网络和基于LSTM的时间注意力机制。t时刻的输入为大小为w的历史时间序列数据I^t，图结构学习通过计算传感器节点之间的相似性学习传感器节点之间的有向图结构，使用邻接矩阵A进行表示。多跳图注意力网络能够学习传感器间的依赖关系图，得到多跳注意力矩阵M，用于更新节点特征。对于更新后的节点数据，STA采用基于长短时间记忆网络的时间注意力机制捕捉序列的时间相关性。STA通过预测下一时间点的数据值$\tilde{x}^t$对多元时间序列进行异常检测。

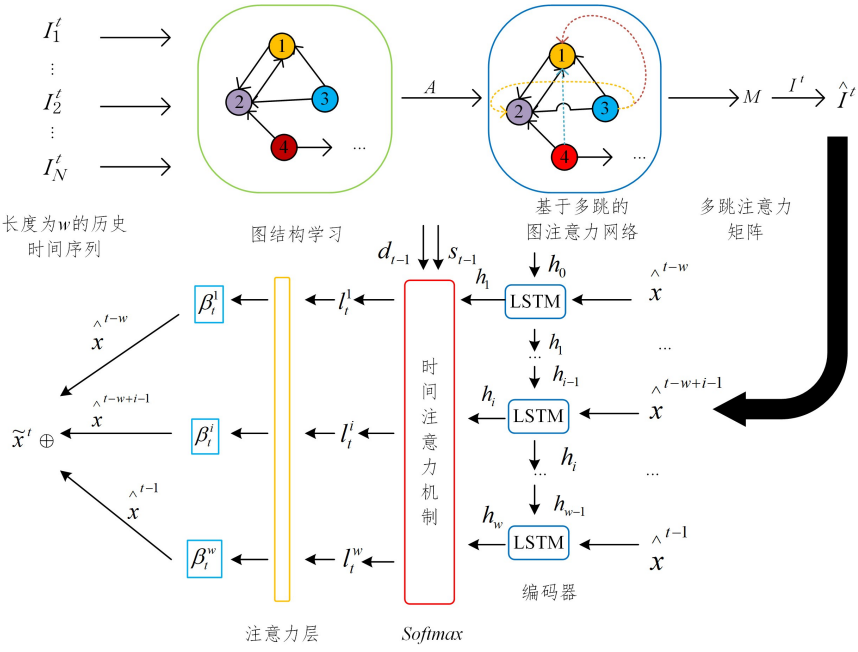


图1 STA架构的可视化

Fig. 1 Visualization of STA architecture

3.3 图形结构学习

STA需要用到传感器数据之间的关系图。通常来说，传感器数据之间的关系图是没有预先定义的，本文通过余弦相似性学习传感器间的关系。考虑到传感器之间的相互关系是非对称的，本文使用有向图G表征传感器之间的关系，图中的节点表示传感器，从一个节点指向另一个节点的有向边指明第一个节点作用于第二个节点。采用邻接矩阵A表示有向图G，$A_{ij}=1$表示节点i到节点j的有向边存在，反之$A_{ij}=0$则表示不存在。

在没有任何先验知识的情况下，传感器i的相邻候选集是除其本身之外的所有传感器，表示为：

$$\mathscr{C}_i\subseteq\{1,2,\cdots,N\}\backslash\{i\} \tag{3}$$

其中，“\”表示集合$\{1,2,\cdots,N\}$中不包含元素i。

为每个传感器节点引入嵌入向量，该向量不仅可以表示该节点的特征，还可以通过计算嵌入向量之间的相似性表示传感器之间的依赖性。定义$v_i\in R^d$表示第i个传感器节点的嵌入向量，其中$i\in\{1,2,\cdots,N\}$，N表示传感器节点的个数，d表示嵌入向量的维度。传感器节点i与其候选节点j之间的相似性为两节点之间的余弦相似度，表示为：

$$e_{ij}=\cos[v_i,v_j]=\frac{v_i\cdot v_j}{\|v_i\|\times\|v_j\|} \tag{4}$$

从节点i的候选集中选择排列前k的相似节点搭建有向边，即

$$A_{ij}=1\{j\in TopK(\{e_{ik}\mid k\in\mathscr{C}_i\})\} \tag{5}$$

用于构造有向图，该有向图即为传感器之间的图形结构。

3.4 多跳图注意力网络

为了学习多元时间序列之间的空间相关性，STA使用多跳图注意力网络捕捉传感器之间复杂的依赖关系。

对于3.3节中构建的有向图结构，STA首先使用图注意力网络学习相邻传感器节点之间的注意力权重，得到单跳注意力矩阵$M^1\in R^{N\times N}$。典型的图神经网络假设一个节点的表示受其在图形结构中相邻节点的影响，并通过单跳邻居节点的表征来更新节点特征。图形注意力网络(Graph Attention Network，GAT)在图神经网络中加入了注意力机制，可以根据传感器节点之间的依赖关系给每个单跳邻居节点分配不同的注意力权重。

为此，在 t 时刻，图注意力网络可以计算节点 i 与其邻居节点之间的注意力系数 $\alpha_{i,j}$，如图 2 所示，具体计算过程如下：

$$g_i^t = v_i \oplus WI_i^t \tag{6}$$

$$\pi(i,j) = LeakyReLU(a^{\mathrm{T}}(g_i^t \oplus g_j^t)) \tag{7}$$

$$\alpha_{i,j} = \frac{\exp(\pi(i,j))}{\sum_{k \in \mathcal{N}_i \cup \{i\}} \exp(\pi(i,k))} \tag{8}$$

$$j \in \mathcal{N}_i \tag{9}$$

其中，$I_i^t \in R^w$ 是 t 时刻传感器节点 i 的输入特征，$\mathcal{N}_i = \{j \mid A_{ij} > 0\}$ 是从邻接矩阵 A 中学习到的节点 i 的相邻节点集。共享参数 $W \in R^{d \times w}$ 是对每个节点应用线性变换的可训练权重矩阵，其用于对节点的特征进行增维。v_i 是传感器节点 i 的嵌入向量，其表征了不同传感器节点的不同特征。$\oplus$ 表示拼接，g_i^t 将传感器嵌入 v_i 和相应的变换特征 WI_i^t 连接起来。a 是注意力机制的学习系数向量，用于把拼接后的高维特征映射到一个实数上。本文选用 $LeakyReLU$ 作为非线性激活函数计算注意力系数，并使用 softmax 函数归一化注意力系数。通过 $\alpha_{i,j}$ 可得到单跳注意力矩阵 M^1。

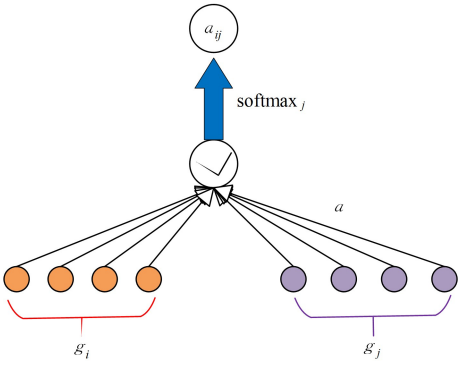


图 2 注意力系数的计算过程

Fig. 2 Calculation process of attention coefficient

由于图形结构中未直接相连的节点之间也可提供重要的网络信息，本文引入多跳邻居节点，通过注意力扩散过程计算多跳注意力权重矩阵 M，该过程基于单跳注意力矩阵 M^1 的幂。根据第 4.8 节中对多跳图注意网路的跳数分析，本文最多只考虑三跳邻居节点，具体计算过程如下：

$$M = \sum_{i=0}^{3} \theta_i M^i \tag{10}$$

$$\sum_{i=0}^{3} \theta_i = 1 \tag{11}$$

$$1 \geqslant \theta_i > 0 \tag{12}$$

其中 $M^0 = I$，I 是单位矩阵。θ_i 是注意力衰减因子，且 $\theta_i > \theta_{i+1}$，因为在消息聚合中，距离更远的节点的权重应该更小。由于不同数据集中传感器之间的依赖关系不同，θ_i 经过对比实验结果进行取值。

使用多跳注意力矩阵 M 可以将输入特征向量 I^t 映射到聚合表示 $\hat{I}^t$，利用这些注意力权重，可以自适应地聚合传感器节点的空间特征，如图 3 所示。传感器节点 i 的聚合表示计算如下：

$$\hat{I}_i^t = \sum_{j \in [1,N]} M_{i,j} I_j^t \tag{13}$$

其中，$M_{i,j}$ 为节点 j 对节点 i 的注意力权重，且：

$$\sum_{j=1}^{N} M_{ij} = 1 \tag{14}$$

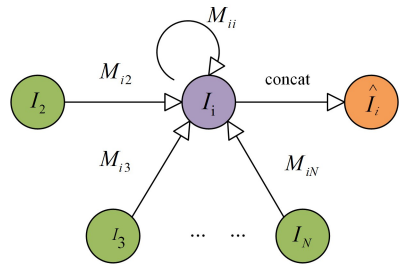


图 3 传感器节点 i 的聚合表示

Fig. 3 Aggregate representation of sensor node i

3.5 基于长短时间记忆的时间注意力机制

多跳图注意力网络只考虑多元序列的维度间依赖性，而基于 LSTM 的时间注意力机制用于学习序列的时间信息。与其他采用 LSTM 的方法[18,20]不同，STA 只将 LSTM 单元作为编码器，学习输入序列的隐藏状态，然后通过学习每个隐藏状态的注意力权重选择相应的时间分量对下一时间步的时间序列进行预测。具体过程如下：

t 时刻的输入：

$$\hat{I}^t := [\hat{x}^{t-w}, \hat{x}^{t-w+1}, \cdots, \hat{x}^{t-1}] \tag{15}$$

用于预测 $\tilde{x}^t$。首先将输入序列编码为机器学习中的隐藏状态，将 w 个隐藏状态定义为 $\{h_1, h_2, \cdots, h_w\}$，则第 i 个隐藏状态被更新为：

$$h_i = f_1(h_{i-1}, \hat{x}^{t-w+i-1}) \tag{16}$$

其中，f_1 为 LSTM 单元。

然后参考编码器 LSTM 单元中的先前隐藏状态 $d_{t-1} \in R^p$ 和单元状态 $s_{t-1} \in R^p$，可以计算每个编码器隐藏状态的注意力权重。具体计算如下：

$$l_t^i = v_d^{\mathrm{T}} \tanh(W_d[d_{t-1}; s_{t-1}] + U_d h_i) \tag{17}$$

$$1 \leqslant i \leqslant w \tag{18}$$

$$\beta_t^i = \frac{\exp(l_t^i)}{\sum_{j=1}^{w} \exp(l_t^j)} \tag{19}$$

其中，$v_d \in R^m$，$W_d \in R^{m \times 2p}$，$U_d \in R^{m \times n}$ 是需要学习的参数，β_t^i 表示 t 时刻第 i 个编码器隐藏状态对预测的重要性。

由于每个编码器隐藏状态被映射为输入的时间分量，可以将该注意力权重用于对应的时间分量，从而跨所有时间步长自适应地选择对应的输入序列。使用注意力权重 β_t^i 计算 $\hat{I}^t$ 的加权平均和可预测 $\tilde{x}^t$，如图 4 所示。具体计算如下：

$$\tilde{x}^t = \sum_{i=1}^{w} \beta_t^i \hat{x}^{t-w+i-1} \tag{20}$$

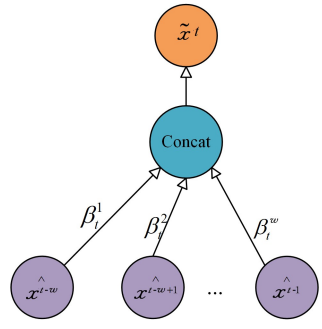


图 4 $\tilde{x}^t$ 的计算过程

Fig. 4 Calculation process of $\tilde{x}^t$

3.6 异常诊断

STA 使用预测输出 $\tilde{x}^t$ 与观测数据 x^t 之间的均方误差作为

最小化损失函数：

$$L_{\mathrm{MSE}}=\frac{1}{T_{\mathrm{train}}-w}\sum_{t=w+1}^{T_{\mathrm{train}}}\|\tilde{x}^t-x^t\|_2^2 \tag{21}$$

训练过程如算法 1 所示。

算法 1 STA

输入：训练数据 $x_{train}=[x^1,\cdots,x^{T_{train}}]$；滑动窗口大小 w

训练：

1. for each epoch do
2. for t=w+1,…,T_{train} do
3. 输入数据 $I^t:=[x^{t-w},x^{t-w+1},\cdots,x^{t-1}]$
4. 根据每个传感器节点的嵌入向量 v_i，i∈{1,2,…,N}构造有向图 G，并用邻接矩阵 A 表示
5. 按照式(6)－式(12)计算多跳注意力权重矩阵 M，通过式(13)计算得出传感器节点的聚合表示 $\hat{I}^t$
6. 根据式(16)－式(20)得到 t 时刻的预测值 $\tilde{x}^t$
7. end for
8. 通过 Adam 最小化式(21)
9. end for

异常分数是将时刻 t 的预测值与观测值进行比较，t 时刻第 i 个传感器的误差值计算为 $err_i(t)=|\tilde{x}_i^t-x_i^t|$。由于不同的传感器可能具有非常不同的特性，它们的偏差值大小可能会有差异性。为了保证结果的可靠性，对传感器的误差值进行标准化处理：

$$a_i(t)=\frac{err_i(t)-u_i}{\sigma_i} \tag{22}$$

其中，u_i 和 σ_i 分别是 $err_i(t)$ 值的中位和四分位间距。

使用 max 函数对每个传感器的异常得分进行聚合，计算时刻 t 的总体异常：

$$A(t)=\max_i a_i(t) \tag{23}$$

为了防止数值的急剧变化对异常检测造成影响，采用简单移动平均法（SMA）对 $A(t)$ 进行平滑处理，将平滑分数 $A_s(t)$ 作为异常分数。将阈值设定为验证集上的最大异常分数，在测试时，当 t 时刻的异常得分大于阈值时，将该时刻标记为异常。

4 实验

本节在 4 个数据集上与 8 种异常检测方法进行对比，实验对比结果验证了 STA 方法异常检测性能的优越性。此外，还从消融研究、灵敏度分析和多跳图注意网路的跳数分析 3 个角度论证了关键组件的有效性。

4.1 数据集

使用 4 个来自现代网络物理系统的公开可用的多元时间序列数据集评估所提方法，如表 1 所列。

表 1 实验中使用的 4 个数据集的统计数据

Table 1 Statistics of four datasets used in experiments

Datasets	Features	Train	Test	Anomalies/%
SWaT	51	49680	44991	12.21
SMD	38	28479	28479	9.46
WADI	127	118800	17280	5.82
DMDS	32	50760	17280	1.56

1）SWaT（安全水处理，Secure Water Treatment）数据集[1]是从包含 51 个传感器的水处理实验台收集的，由 7 天的正常运行数据和 4 天的模拟攻击数据组成，分别作为训练集和测试集。由于系统工作流程的特点，所有传感节点之间存在自然的拓扑结构关系。

2）SMD（服务器机器，Server Machine Dataset）[21]数据集是在 5 周内从一家大型互联网公司收集的，由每分钟定期采样的 28 台具有 38 个传感器的服务器数据组成，其训练集和测试集被按照 1∶1 的比例划分。本文选用其中的一个服务器数据。

3）WADI（配水，Water Distribution）数据集[22]从包含 127 个传感器的水分布实验台获取，其是 SWaT 实验台的扩展。它包含两周的正常数据作为训练集，两天的攻击数据作为测试集。

4）DMDS（工业控制系统中执行器诊断方法的开发与应用，Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems）基准[23]由卢布林糖厂的真实工艺数据以及生成人工故障的模拟器组成，在此只使用工业系统中具有诱发故障的真实公开可用数据集。使用其中 6 天的正常数据作为训练集，3 天的人为攻击数据作为测试集。

本文为了加快训练速度，通过取中值将 SWaT，WADI 和 DMDS 的原始数据样本每 10 秒降采样一次。

4.2 基线方法

为了验证 STA 方法异常检测性能的优越性，将其与 8 种主流的异常检测方法进行比较。

1）RawSignalBaseline：将任何信号重构为“0”。

2）PCA[15]：主成分分析（Principal Component Analysis），将高维表示映射到低维表示，并将映射后的重建误差用于计算异常分数。

3）UnivarAutoEncoder：单变量全连接自动编码器（Univariate Fully-Connected Auto-Encoder），为每个通道训练单独的自动编码器。

4）AutoEncoder[12]：自动编码器，是一种无监督的深度神经网络，包括编码器和解码器。

5）LSTM-AE[14]：长短时间记忆自动编码器，是一种基于长短时间记忆网络的自动编码器方法。

6）TcnAE[19]：时间卷积网络自动编码器，是一种基于时间卷积网络（Temporal Convolutional Network）的自编码器方法。

7）OmniAnomaly[16]：采用随机递归神经网络进行时间序列异常检测，并利用重建概率来解释检测到的异常。

8）GDN[8]：图形偏差网络（Graph Deviation Network），是一种无监督异常检测方法，其将结构学习方法与图注意力网络相结合来学习多变量时间序列之间的复杂关系，并使用注意力权重解释检测到的异常。

4.3 评价指标

使用精确度（Prec）、召回率（Rec）和 F_1 分数评估各方法的性能。

$$F_1=\frac{2\times Prec\times Rec}{Prec+Rec} \tag{24}$$

$$Prec=\frac{TP}{TP+FP} \tag{25}$$

$$Rec=\frac{TP}{TP+FN} \tag{26}$$

其中，TP 表示被模型预测为正常的正样本数量，TN 表示被模型预测为异常的负样本数量，FP 表示被模型预测为正常

的负样本数量，FN 表示被模型预测为异常的正样本数量。

4.4 实验设置

使用 CUDA 12.0.89 和 PyTorch 几何库在 PyTorch 1.5.1 版本上实现我们的方法及其变体，在带有 Intel(R) Core(TM) i9-10900 CPU @ 2.80GHz 2.81GHz 和 NVIDIA GeForce RTX 2080Ti 显卡的服务器上训练模型。使用学习率为 0.001 的 Adam 优化器优化模型。滑动窗口大小设置为 15，滑动窗口的步长为 5。在模型训练阶段，设置样本批次训练的数据大小为 128，采用早起停止策略在 100 个迭代轮次内训练，patience 设为 10。将式(11)中的 k 分别设为 15(SWaT)，10(SMD)，30(WADI)和 15(DMDS)。选用的 4 个数据集都包括训练集和测试集，每次实验还需要从训练集中随机划分 10%的数据作为验证集。对每个数据集进行 10 次重复实验，取 10 次重复实验的平均值作为实验结果。

4.5 实验结果

表 2 列出了在 SWaT，SMD，WADI 和 DMDS 数据集上文中方法和基线方法的精确度、召回率和 F_1 分数性能指标值。

表 2 异常检测模型对比结果

Table 2 Comparison results of anomaly detection model

方法	SWaT			SMD			WADI			DMDS		
	Prec	*Rec*	F_1	*Prec*	*Rec*	F_1	*Prec*	*Rec*	F_1	*Prec*	*Rec*	F_1
RawSignalBaseline	21.26	2.23	0.04	96.53	9.31	0.16	55.24	15.77	0.24	**72.60**	18.91	0.30
PCA[15]	95.40	13.89	0.24	94.30	8.61	0.15	58.36	10.86	0.18	48.14	15.71	0.23
UnivarAutoEncoder	42.15	3.41	0.06	76.88	10.98	0.19	77.82	17.85	0.29	58.43	18.09	0.27
AutoEncoder[12]	87.76	18.04	0.29	**98.98**	7.23	0.13	59.42	16.12	0.25	53.47	33.34	0.41
LSTM-AE[14]	95.29	13.56	0.23	94.09	8.27	0.15	69.21	13.85	0.23	46.13	27.22	0.34
TcnAE[19]	67.09	15.56	0.25	98.06	7.53	0.13	65.89	17.42	0.27	71.26	24.57	0.36
OmniAnomaly[16]	78.08	24.77	0.37	97.47	7.16	0.13	**99.54**	15.32	0.26	58.00	15.40	0.24
GDN[8]	60.53	**65.23**	0.58	54.83	**80.56**	0.63	59.11	35.48	0.44	58.72	37.41	0.46
STA	**95.69**	63.40	**0.76**	71.32	73.56	**0.72**	62.17	**42.64**	**0.51**	72.09	**45.92**	**0.56**

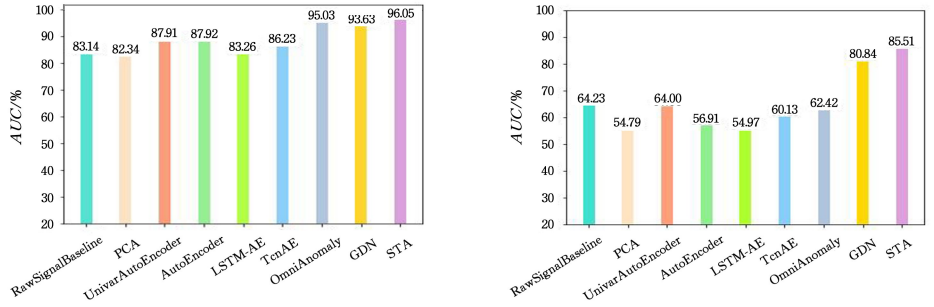
注：最好的结果用粗体突出显示。

结果表明，STA 在 4 个数据集上的 F_1 分数始终优于所有基线方法。具体来说，STA 在 SWaT，SMD，WADI 和 DMDS 上的 F_1 分数分别优于最佳基线 31.03%，14.29%，15.91%和 21.74%。此外，STA 在 SWaT 上的精确度优于基线方法，且召回率仅次于 GDN；STA 在 SMD 上的召回率仅次于 GDN；STA 在 WADI 和 DMDS 上的召回率优于基线方法。这些实验结果证明 STA 的双重注意力机制，能帮助模型更好地利用多元时间序列中的时间和空间信息，从而提高模型的异常检测性能。

STA 方法采用多跳图注意力网络和基于 LSTM 的时间注意力机制，能够有效捕捉多元时间序列中复杂的时空相关性。与其他深度学习方法相比，传统方法(RawSignalBaseline 和 PCA)表现较差，因为它们难以对多元时间序列的相关性进行编码。对比发现，UnivarAutoEncoder 的性能比 AutoEncoder 差，因为前者对每个特征训练单独的自动编码器，没有考虑时间序列之间的复杂关系。此外，GDN 取得了比其他基线更好的性能，但其性能始终差于 STA 方法。这是因为 GDN 同样采用了图注意网络对特征间关系进行构建，然而其没有考虑多跳节点之间的信息传播且对时间依赖性的关注度不够。这再次表明 STA 方法引入多跳邻居节点和时间注意力机制的可行性。

通过对比 4 个数据集的实验结果可以发现，SWaT 数据集在 F_1 分数性能方面的改进最令人深刻。主要原因有两点：1)SWaT 和 WADI 的所有传感节点之间存在自然的拓扑结构关系，特征之间存在较强的依赖性；2)如表 1 所列，SWaT 比 WADI 更加平衡且特征维度较低。因此，本文提出的方法更适用于维度较低且特征之间强相关的数据集。

进一步将 AUC 作为 SWaT 和 SMD 数据集的性能指标，如图 5 所示。对于这两个数据集，STA 方法的精确度始终优于其他基线方法。我们将这种优势归因于 STA 方法对时间和空间信息的有效利用。结果表明，STA 采用时空注意力机制有助于捕获多尺度信息依赖，提高异常检测的真阴性率。



(a)SWaT (b)SMD

图 5 SWaT 和 SMD 数据集的 *AUC*

Fig. 5 *AUC* results of SWaT and SMD datasets

4.6 消融实验

为了验证关键组件对于提高模型性能的有效性，在 SMD 数据集上进行消融实验，性能指标选用精确度、召回率、F_1 分数和 AUC。模型结构分别为：

1)STA 模型。本文提出的原始模型，在图注意力网络中引入多跳邻居节点，并增加了基于 LSTM 的时间注意力机制模块。

2)STA w/o(without) multi-hop 模型。在 STA 原始模型中去除图注意力网络中的多跳机制，只考虑与每个传感器节点直接相连的邻居节点，保留基于 LSTM 的时间注意力机制模块。

3)STA w/o(without) att-temp 模型。在 STA 原始模型中移除基于 LSTM 的时间注意力机制模块，为所有时间步长分配相等权重，保留图注意力网络中的多跳机制。

4)STA w/o(without) mulhop-temp 模型。在 STA 原始模型中同时去除图注意力网络中的多跳策略和时间注意力机制，只依靠图注意力网络捕捉传感器之间的复杂关系。

实验结果如表 3 所列。结果表明，引入时间注意力机制的 STA w/o(without) multi-hop 模型相较于 STA w/o(without) mulhop-temp 模型，在 SMD 数据集上的 F_1 分数提高了 9.23%，表明引入时间注意力机制能够提高异常检测的性能。考虑多跳邻居节点的 STA w/o(without) att-temp 模型相较于 STA w/o(without) mulhop-temp 模型，在 SMD 数据集上的 F_1 分数提高了 3.08%，证明考虑多跳节点可以更好地捕捉传感器间的相互关系，但是缺乏对序列时间依赖性的考虑，导致模型的提升效果一般。

表 3　消融实验

Table 3　Ablation study

Method	*Prec*	*Rec*	F_1	AUC
STA	**71.32**	73.76	**0.72**	**0.96**
STA w/o(without) multi-hop	69.66	74.66	0.71	0.95
STA w/o(without) att-temp	60.15	79.42	0.67	0.94
STA w/o(without) mulhop-temp	57.83	**79.64**	0.65	0.94

注：最好的结果用粗体突出显示。

STA 模型相较于只保留时间注意力机制模块的 STA w/o(without) hop-att_temp 模型，在 SMD 数据集上的 F_1 分数提升了 1.41%；相较于只保留图注意力网络中多跳机制的 STA w/o(without) att-temp 模型，在 SMD 数据集上的 F_1 分数提升了 7.47%；相较于 STA w/o(without) mulhop-temp 模型，在 SMD 数据集上的 F_1 分数提升了 10.9%。消融实验结果分析说明，STA 引入多跳邻居节点和时间注意力机制都有助于提高模型的异常检测性能。这也说明了文中方法优于同样采用图注意力网络的最优基线(GDN 方法)的优势所在，同时为其优于其他基线方法提供了解释。

4.7　灵敏度分析

在 SMD 数据集上分析了时间窗口大小对 STA 模型以及 3 个消融模型的异常检测性能的影响，实验结果如图 6 所示。

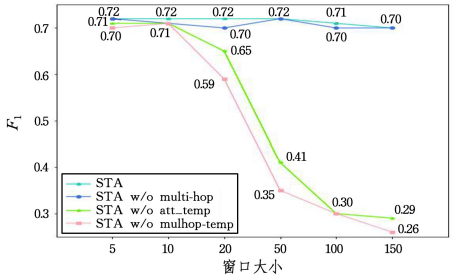


图 6　F_1 分数随窗口大小的变化

Fig. 6　Variation of F_1 score with window size

通过对比 STA 模型与 STA w/o(without) multi-hop 模型的折线图，可以发现随着窗口大小的变化，前者的 F_1 分数始终大于或等于后者的 F_1 分数，表明引入多跳节点可以有效提高模型的异常检测性能；同理，对比 STA w/o(without) att-temp 模型与 STA w/o(without) mulhop-temp 模型，也可以得出上述结论。

通过对比 STA 模型与 STA w/o(without) att-temp 模型的折线图，可以发现随着窗口的增大，前者的 F_1 分数明显更为稳定且始终大于后者的 F_1 分数，说明引入时间注意力机制不仅能更好地建模多元时间序列的时空信息，还能使模型更加稳定。同理，对比 STA w/o(without) multi-hop 模型与 STA w/o(without) mulhop-temp 模型，也可以得出该结论。

4.8　多跳图注意力网络的跳数分析

为了研究图注意力网络中多跳邻居节点的跳数大小对 STA 方法异常检测性能的影响，以 SWaT 数据集为例进行了相应的分析，性能指标选用精确度、召回率、F_1 分数和 AUC，实验结果如图 7 所示。

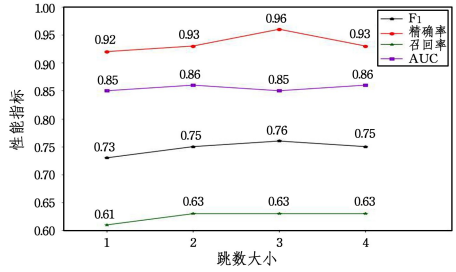


图 7　性能指标随跳数大小的变化

Fig. 7　Variation of performance indicators with the number of hops

通过对比可以发现，当节点跳数为 3 时，STA 的综合性能指标最好，其精确度、召回率和 F_1 分数最大，且 *AUC* 仅比最优值小 0.01；当节点跳数为 2 和 4 时，STA 的性能几乎一样，且两者的 F_1 分数和精确度比节点跳数为 3 时的数值小；当节点跳数为 1 时，STA 的 4 个性能指标最差。因此，本文将多跳图注意力网络的最大节点跳数设为 3。

结束语　文中提出了基于时空注意力机制的多元时间序列异常检测方法，称为 STA。STA 采用多跳图注意力网络和基于 LSTM 的时间注意力机制模块同时捕捉多元时间序列空间相关性和时间相关性。STA 在 4 个真实世界数据集上的异常检测性能优于目前主流的异常检测方法，相较于这些方法，STA 不仅能够更好地学习多元序列间的复杂关系，还能动态捕捉每个序列的时间依赖性。此外，消融实验和灵敏度分析证明使用 STA 中的关键组件能提高异常检测性能和稳定性。

通过对比 STA 方法在 4 个数据集上的结果，可以发现该方法更适用于数据特征维度较低且特征之间相关性强的数据集。在未来工作中将针对该问题进行改进，提高 STA 在其他数据集上的性能。

参 考 文 献

[1]　MATHUR A P. TIPPENHAUER N O. SWaT: a water treatment testbed for research and training on ICS security[C]// 2016 International Workshop on Cyber-physical Systems for Smart Water Networks(CySWater). IEEE, 2016: 31-36.

[2]　ZHANG C, SONG D, CHEN Y, et al. A deep neural network for

unsupervised anomaly detection and diagnosis in multivariate time series data[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:1409-1416.

[3] LIN S,CLARK R,BIRKE R,et al. Anomaly detection for time series using vae-lstm hybrid model[C]// 2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP 2020). IEEE,2020:4322-4326.

[4] SCHLICHTKRULL M,KIPF T N,BLOEM P,et al. Modeling relational data with graph convolutional networks[C]// The Semantic Web:15th International Conference(ESWC 2018). Heraklion,Crete,Greece,Springer International Publishing,2018:593-607.

[5] VELIČKOVIĆ P,CUCURULL G,CASANOVA A,et al. Graph attention networks[J]. arXiv:1710.10903,2017.

[6] ZHANG Z,LI M,LIN X,et al. Multistep speed prediction on traffic networks:A deep learning approach considering spatio-temporaldependencies[J]. Transportation Research Part C:Emerging Technologies,2019,105:297-322.

[7] ZHENG C,FAN X,WANG C,et al. GMAN:A Graph Multi-Attention Network for Traffic Prediction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:1234-1241.

[8] DENG A,HOOI B. Graph neural network-based anomaly detection in multivariate time series[C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence. 2021.

[9] WANG X,HE X,CAOY,et al. Kgat:Knowledge graph attention network for recommendation[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019:950-958.

[10] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural computation,1997,9(8):1735-1780.

[11] PANG G,SHEN C,CAO L,et al. Deep learning for anomaly detection:A review[J]. ACM Computing Surveys(CSUR),2021,54(2):1-38.

[12] HAWKINS S,HE H,WILLIAMS G,et al. Outlier detection using replicator neural networks[C]// Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg,2002:170-180.

[13] GARG A,ZHANG W,SAMARANJ,et al. An evaluation of anomaly detection and diagnosis in multivariate time series[J]. IEEE Transactions on Neural Networks and Learning Systems,2021,33(6):2508-2517.

[14] MALHOTRA P,RAMAKRISHNAN A,ANANDG,et al. LSTM-based encoder-decoder for multi-sensor anomaly detection[J]. arXiv:1607.00148,2016.

[15] SHYU M L,CHEN S C,SARINNAPAKORN K,et al. A novel anomaly detection scheme based on principal component classifier[R]. Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering,2003.

[16] SU Y,ZHAO Y,NIU C,et al. Robust anomaly detection for multivariate time series through stochastic recurrent neural network[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019:2828-2837.

[17] KIM S,CHOI K,CHOI H S,et al. Towards a rigorous evaluation of time-series anomaly detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022:7194-7201.

[18] HUNDMAN K,CONSTANTINOU V,LAPORTE C,et al. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018:387-395.

[19] BAI S,KOLTER J Z,KOLTUNV. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv:1803.01271,2018.

[20] PARK D,HOSHI Y,KEMP C C. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder[J]. IEEE Robotics and Automation Letters,2018,3(3):1544-1551.

[21] SU Y,ZHAO Y,NIU C,et al. Robust anomaly detection for multivariate time series through stochastic recurrent neural network[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019:2828-2837.

[22] AHMED C M,PALLETI V R,MATHUR A P. WADI:a water distribution testbed for research in the design of secure cyber physical systems[C]// Proceedings of the 3rd International Workshop on Cyber-physical Systems for Smart Water Networks. 2017:25-28.

[23] BARTYŚ M,PATTON R,SYFERT M,et al. Introduction to the DAMADICS actuator FDI benchmark study[J]. Control Engineering Practice,2006,14(6):577-596.

LIANG Lifang, born in 1999, postgraduate. Her main research interests include anomaly detection for time series and so on.

GUAN Donghai, born in 1981, Ph.D, associate professor, postgraduate supervisor. His main research interests include data mining, knowledge inference and so on.