杨桂松, 温盼盼. 基于稀疏移动群智感知的时空数据推理方案[J]. 智能计算机与应用,2026,16(1):15-23. DOI:10.20169/j.issn.2095-2163.24032707

基于稀疏移动群智感知的时空数据推理方案

杨桂松, 温盼盼
(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘 要: 移动群众感知(MCS)是一种传感范式,可以实现大规模的智慧城市应用,如环境传感和交通监测。然而,由于收集数据的时空覆盖范围有限,传统的 MCS 经常出现性能下降问题。在这种情况下,提出了稀疏 MCS,可利用数据推理算法从用户收集的稀疏数据中恢复完整的数据。然而,现有的稀疏 MCS 方法往往因为数据稀疏度在恢复子区域感知数据时的累积误差,导致数据推理算法精度不够。累积误差使得算法的预测结果偏离实际情况,从而影响决策的准确性和可靠性,所以对数据推理算法带来了巨大的挑战。为了解决这一问题,本文提出了一种基于时空依赖模式补全的方法来用于稀疏 MCS 中的缺失数据推理。具体来说,首先提出时空模式层从非常有限的观测数据中获取感知数据复杂的时空相关性,然后利用缺失数据补全层自适应地挖掘序列中的缺失特征,并生成缺失值。本文在 2 个真实世界的城市感知数据集上进行了广泛的实验,实验结果表明了该方法的有效性。

关键词: 稀疏移动群智感知; 时空信息; 缺失数据推理; 注意力机制
中图分类号: TP309 **文献标志码**: A **文章编号**: 2095-2163(2026)01-0015-09

Spatio-temporal data inference scheme based on sparse Mobile Crowd Sensing

YANG Guisong, WEN Panpan
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Mobile Crowd Sensing (MCS) is a sensing paradigm that enables large-scale smart city applications such as environmental sensing and traffic monitoring. However, traditional MCS often suffers from performance degradation due to the limited spatial-temporal coverage of the collected data. In this context, sparse MCS is proposed, which utilizes data inference algorithms to recover complete data from sparse data collected by users. However, existing sparse MCS methods often result in insufficient accuracy of the data inference algorithm, due to the cumulative error of data sparsity in recovering subregion-aware data. The cumulative error makes the prediction results of the algorithm deviate from the actual situation, which affects the accuracy and reliability of the decision-making, so it poses a great challenge to the data inference algorithm. In order to solve this problem, this paper proposes a method based on spatiotemporal dependent pattern complementation for missing data inference in sparse MCS. Specifically, a spatial-temporal pattern layer is firstly proposed to obtain the complex spatial-temporal correlations of perceptual data from very limited observational data, and then a missing data complementation layer is utilized to adaptively mine the missing features in the sequences and generate the missing values. In this paper, extensive experiments are conducted on two real-world urban perceptual datasets, and the experimental results demonstrate the effectiveness of the method.

Key words: sparse Mobile Crowd Sensing; spatio-temporal information; missing data inference; attention mechanism

0 引 言

移动群智感知[1-3]通过与移动设备[4-5]相结合,成为一种首选的数据收集方法。移动用户参与从特定的传感区域收集数据,处理诸如环境监测和交通控制[6-7]等任务。精确地感知目标信号图区域对于移动群智感知应用程序是至关重要的,通常需要收集大量的高质量的数据。传统的移动群智感知系统通常依赖于招募大量的用户群来全面覆盖整个传感地图,以获得最佳结果。然而,考虑到用户的不确定的移动性和有限的感知预算等因素[8],在大多数情况下,移动群智感知获得的时空感知地图是稀疏的,

基金项目: 国家自然科学基金(61602305,61802257); 上海市自然科学基金(18ZR1426000,19ZR1477600)。

作者简介: 杨桂松(1982—),男,博士,副教授,硕士生导师,主要研究方向:物联网与普适计算等。Email:gs_yang@aliyun.com; 温盼盼(1999—),女,硕士研究生,主要研究方向:移动群智感知。

收稿日期: 2024-03-27

特别是面对大规模、细粒度的城市感知任务的实现过程中。

为了解决数据不足的挑战,研究人员提出了稀疏移动群智感知数据收集方法[9-11]。稀疏移动群智感知的目的是利用已经收集到的数据来推断感知分区的数据,从而完成完整的信号感知图。在稀疏移动群智感知中,感知数据单元内的时空相关性使数据重建成为可能。因其会优先考虑将总感知精度作为数据质量的度量标准,而不是仅仅关注覆盖范围。然而,在稀疏移动群智感知中仍然存在着需要解决的挑战。

首先,当稀疏移动群智感知检测到缺失区域时,KNN、GPR 和随机森林方法可用于缺失数据推断。高斯过程回归(GPR)[12]算法由于数据稀疏性,在学习足够的相关性来恢复信号感知数据上面临挑战。KNN-ST[13]结合时空信息使用线性回归模型对时空数据进行聚类分析,但对异常值非常敏感,导致预测结果不稳定。随机森林[14]通常确保在测量满足某些标准时的最佳性能,进行缺失数据推断。

其次,压缩感知和矩阵补全依赖于随机缺失数据的假设。压缩感知[15]可以通过预测不完整的数据来监测空气质量,预测未开发地区的污染水平。BCCS 技术[16]将移动群智感知与贝叶斯压缩感知相结合来重建信号图。如果不能满足这些基本条件,可能会影响性能,导致感知不均匀。

第三,稀疏移动群智感知获得的信号具有随机性特征,并且缺乏足够的参与导致大量不完整或低质量的信号图。这种质量较低的测量值会严重影响数据恢复的质量。虽然在数据不足或恢复性能差的情况下,矩阵补全可能面临挑战,但生成模型可能提供一种解决方案。深度矩阵分解方法[17]中,误差的累积极大影响了缺失数据恢复。矩阵补全和生成对抗网络相结合[18],用于时空数据的生成和预测。变分自动编码器(VAE)[19]被提出用于恢复缺失的交通数据。

为了解决上述问题,本文提出了一种基于时空依赖模式补全的缺失数据推理算法。GBTSIN 主要通过时空模式层对不同的时间步长和不同的节点进行多层次的训练和加权,有助于模型更灵活地适应不同序列的特征,提高决策性能并且降低了计算的复杂度。当任务发布者通过移动群智平台向参与者发布感知任务时,移动群智平台通过一定的激励机制积极引导工人获得更高奖励的感知区域数据。在工人上传信号后,移动群智平台利用 GBTSIN 模型对稀疏的感知区域数据进行训练补全,重建完整的信号图。上述过程一直持续到重建精度满足要求为止。

相较于现有的信号图重建算法,本文提出的方法具有显著的优势。本文的方法能够在无需任何关于信号图的假设信息(如信号图的矩阵秩和稀疏性)的情况下,高精度地重构信号图,从而提升信号图的感知数据质量。

1 系统模型

本节首先分析并提出了基于时空信息的城市感知信号图缺失数据重建问题,然后提出了稀疏移动群智感知推理框架。在城市感知活动中,常见的移动群智感知系统让多个工人从目标区域收集数据,这些数据将用于提供城市感知服务。在改进的城市感知场景中,工人被雇佣来收集特定子区域的数据,目标是利用数据推理算法来推断每个感知周期内的信号数据。

本文提出的 GBTSIN 模型框架如图 1 所示,主要由 2 个组件构成,分别是时空依赖层和缺失数据补全层。该模型使用包含时间和地理信息的数据集作为输入,对数据节点的邻域时空相关性和节点值进行建模。时空依赖层将交叉注意力机制引入时空相关性学习,以便从时间和空间维度提取相关特征,捕获节点序列的分布规律。同时自注意力机制用于学习自身节点序列的内部依赖关系,得到对应的状态向量。缺失数据补全层将每个状态向量应用于交叉注意力网络单元,使模型能够关注序列中的隐藏状态向量,生成注意力向量。最后,注意力向量和隐藏状态向量相加,得到输出向量,用于生成序列中的缺失值。

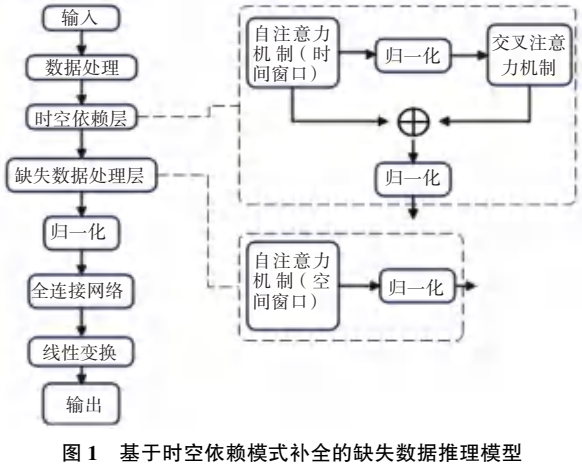


图 1 基于时空依赖模式补全的缺失数据推理模型

Fig. 1 Inference model for missing data based on spatio-temporal dependency pattern complementation

给定一个信号传感图 $G=\langle V,E\rangle$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 表示包含传感器节点数的集合，E 表示传感器节点所连接的边集 $\langle v_i,v_j\rangle \in V$。$\boldsymbol{W}_t \in R^{N_t\times N_t}$ 表示加权邻接矩阵，$W_t^{i,j}$ 表示第 i 个和第 j 个传感器节点之间的边的信息权重。传感器节点之间的属性特征矩阵记为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\} \in R^{N_t\times M_t}$，其中 m 表示属性特征的维数。

GBTSIN 模型利用图注意力机制来学习每个传感器节点在观察到的时间窗口 t 到 $t+1$ 内的信息表示。对于每个传感器节点 i，研究中考虑其邻居集合 $j \in N(i)$，其中 $X_{t:t+T}^j$ 是 $X_{t:t+T}^i$ 邻居传感器节点的观测数据。在信息传递的过程中，每个节点 $X_{t:t+T}$ 通过执行节点更新函数，将自身特征与相邻节点的信息相融合，从而更新自身特征。这样，每个节点都能够从周围的感知节点 $N(i)$ 收集信息，将其与本身的数据相结合并更新得到节点的推断值 x_t^i。为了考虑更远距离的信息交互，例如节点 k 对几跳距离之外的节点的影响，可以通过多次堆叠 k 层来实现。这样的层堆叠可以帮助捕捉更复杂的节点间关系，对表示向量 v_i 进行更新。

为了降低在图 G 上执行消息传递的计算和内存成本，GBTSIN 模型采用了一种有效的策略，即以 k 个虚拟节点为中心来传播信息。研究中将感知数据序列划分为多个窗口，通过识别这些窗口之间的相关性，来捕捉短期细粒度的感知序列结构特征。这有助于有效地处理较长的时间步长，减轻计算负担。

GBTSIN 模型的目标是尽可能最大限度地提高稀疏移动人群传感平台的效用。设计目标如下。

(1)感知数据合理性：从同一位置的多个用户收集的感知数据是真实一致的。

(2)感知预算最优化：在每个感知任务中，力求通过最少的感知单元来完成任务，以此降低预算开销，同时尽可能地最大化效用。

该研究问题可以表述如下：如何利用在随机时刻收集的不完整信号来创建和更新一个特定区域的信号图，同时最小化信号收集的成本。

2　基于时空依赖模式补全的缺失数据推理方法

本节介绍了一种新的策略，通过学习时空特征进行稀疏信号感知和缺失数据补全，从而有效地解决了数据高精度恢复的难题。本研究的目标是通过对给定序列 X_t 的分析以及学习到的时空维度信息，来补全输入信号图中的缺失值。

首先，将时空数据进行编码，并嵌入到图结构网络中，以便进行表示。在该网络中，每个节点都代表一个传感器设备上传的数据。为了捕获时间和地理位置信息，本模型通过注意力神经网络对节点进行学习。此外，节点之间的边则代表了节点之间的属性，如 $PM_{2.5}$、PM_{10}、NO_2 和距离等。

GBTSIN 模型主要由 2 个核心组件组成：时空依赖层和缺失数据补全层。首先，时空依赖层被用于学习时空图序列中的时空相关性。具体而言，交叉注意力网络用于挖掘时空图序列中的空间相关性和时间相关性。然后，缺失数据补全层用于自适应地学习空间序列中的缺失特征，并生成缺失值。最后，通过模型输出推理值和真实观测值的损失函数优化 GBTSIN 模型中的参数。

2.1　时空依赖层

GBTSIN 模型的时空依赖层是由注意力神经网络构成的。因此，本节首先简单介绍注意力网络的基本原理。在注意力网络中，目标节点会在空间维度上接收邻域节点信息以及自身的历史节点信息。

在注意力网络的基础上，进一步定义 GBTSIN 模型的时空依赖层。时空依赖层的核心思想在于整合交叉注意力机制，挖掘信号图序列间的时空相关性。在提出的时空依赖层中，需要更新的节点表示为 q，而起始节点用 k 表示。每个节点的值表示为 v。通过使用时空交叉注意力网络在时空节点之间传输并计算节点信息。

考虑到传统的注意力网络在处理长序列时可能会遇到困难，GBTSIN 在这里采用了掩码形式来替代传统的注意力操作。X_t^i 和 Y_t^i 分别为第 i 个节点的观测集和目标集。观测集合 X_t 包含了所有感知节点数据在时间 t 的属性信息，表示为 $X_t=\{\langle x_\tau^i,q_\tau^i\rangle m_\tau^i=1\}$，其中 q_τ^i 表示坐标点。同时，重建的信号感知图则构成目标集合 Y_t，其中包含了所有未感知节点数据 $\{q_\tau^i \mid m_\tau^i=1\}$。GBTSIN 致力于学习时空图中离散节点的表示。

从时间步长 s 到时间步长 τ 沿时间维度的传播由下标 $s\rightarrow\tau$ 表示。同样，从第 j 个节点发送到第 i 个节点的消息用上标 $j\rightarrow i$ 表示。从时间步 s 的第 j 个节点到时间步 τ 的第 i 节点的消息 $r_{s\rightarrow\tau}^{j\rightarrow i} \in R^d$ 计算为：

$$r_{s\rightarrow\tau}^{j\rightarrow i}=\mathrm{MLP}(h_s^j,h_\tau^i) \tag{1}$$

其中，MLP 为多层感知机，用于平衡自身节点信息和从其邻居节点聚合得到的信息。通过在节点

间传递并聚合信息,MLP 能够获取更全面、更准确的结构信息,从而学习到对节点的有效表示。

目标节点 q 表示为 h_τ^i, 起始节点 k 表示为 h_s^j, 节点值 v 定义为消息 $r_{s\to\tau}^{j\to i}$, 通过起始节点和目标节点来计算这个值。为了表示空间信息,研究中采用了一种节点序列间的交叉注意机制。对于每个邻居节点 $j \in N(i)$,h_τ^i 用于查询每个有效 $h_{t:t+T}^j$ 的编码表示,并收集消息集合 $R_\tau^{j\to i}$:

$$R_\tau^{j\to i} = \{ r_{s\to\tau}^{j\to i} \mid \langle x_s^j, q_\tau^i \rangle \} \tag{2}$$

然后使用可训练权值函数 w_α 对 $R_\tau^{j\to i}$ 中的消息进行线性变换,得到一个 Softmax 层的消息得分 $\alpha_{s\to\tau}^{j\to i}$:

$$\alpha_{s\to\tau}^{j\to i} = \frac{\exp(r_{s\to\tau}^{j\to i} w_\alpha)}{\sum_{r\in R_\tau^{j\to i}} \exp(r w_\alpha)} \tag{3}$$

其中, r 是分配给 $\alpha_{s\to\tau}^{j\to i}$ 的权衡标量,接着对生成的节点表示进行加权和组合,以获得时间上下文向量 c_τ^i, 公式定义如下:

$$c_\tau^i = \sum_{s:r_{s\to\tau}^i \in R_\tau^i} \alpha_{s\to\tau}^i \cdot r_{s\to\tau}^i \tag{4}$$

将第 j 个节点的时间消息聚合,获得第 j 个节点序列到第 i 个节点和时间步长 τ 上的边上下文向量 $e_\tau^{j\to i}$:

$$e_\tau^{j\to i} = \sum_{s:r_{s\to\tau}^{j\to i} \in R_\tau^{j\to i}} \alpha_{s\to\tau}^{j\to i} \cdot r_{s\to\tau}^{j\to i} \tag{5}$$

在得到向量 c_τ^i 和 $e_\tau^{j\to i}$ 后,对 h_τ^i 进行更新,得到最终的聚合步骤为:

$$h_\tau^{i,l+1} = \mathrm{MLP}(h_\tau^{i,l}, c_\tau^{i,l}, \sum_{j\in N(i)} e_\tau^{j\to i,l}) \tag{6}$$

在信息传递过程中,使用更新函数来更新连续层中的节点特征 h_{t+1}。 具体来说,将从消息函数得到的消息 r_{t+1} 与当前层节点的特征 h_t 进行合并。这可以通过以下公式表示:

$$h_t = u_t \otimes h_{t-1} + (1 - u_t) \odot r_t \tag{7}$$

其中, h_t 表示节点在时间 t 时的表示; u_t 表示更新函数;“⊙”表示哈达玛乘积。

给定在 $t-1$ 层的节点特征向量 h_{t-1}, 全局图的特征向量描述如下:

$$\hat{y} = \gamma_k(x_{t-1}^i, \rho_k(x_{t-1}^i, h_{t-1}^i, W_t)) \tag{8}$$

其中, γ_k 和 ρ_k 分别表示更新函数和信息函数 MLP。

实际上,在时空依赖模型中,每一层的每个节点都会将其状态作为消息发送给其他节点,节点的更新信息将通过特定的关系返回到目标节点,这些返回的数据包含结构信息,用于更新目标节点 h_τ^i。 节点的更新是基于层聚合函数的结果进行的,因此,在经过 l 层的处理后,节点的表示将被限制在其 l 跳的范围内。这个约束要求模型在重建目标输入时,需要考虑到数据点之间的空间关系对结果的影响。

在传递信息以进行更新时,现有模型的消息传递策略通常会忽略每个节点的重要性。为了解决这个问题,使模型加强对关键节点的关注,该模块引入了阈值高斯核函数来识别和关注在信息传递过程中起着关键作用的节点,有助于提升了信息传递的效果,对此可以表示为:

$$W^{i,j} = \begin{cases} \exp(-\dfrac{\mathrm{dist}(i,j)^2}{\gamma}), & \mathrm{dist}(i,j) \leqslant \delta \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

GBTSIN 框架的主要目标是学习那些将节点的显著性和全局特征结构融合在一起的连接特征。这些特征作为输入,可以帮助模型更好地理解和表示数据的结构。通过将结构特征(节点之间的连接方式)和属性特征(节点本身的特性)相结合,可以为结构训练提供更具体的标签有助于模型更准确地预测数据。然而,由于训练过程可能需要很长时间,因此捕获节点的长期依赖关系可能会很困难。这是因为随着时间的推移,节点的状态可能会发生很大的变化,而过程中将需要足够的时间和计算资源来捕获和理解这些变化。

GBTSIN 模型通过交叉注意力过程进行训练。在此过程中,每个表示都会聚合并更新所有从相邻节点获取的可用传感器数据。每个表示都代表一个特定的节点和时间步长,通过注意力因子对每个表示进行加权。然后,解码器对输出进行解码,从而得到最终重构的信号图。

实际上,在时空依赖模型中,每一层的每个节点都会将其状态作为消息发送给其他节点,节点的更新信息将通过特定的关系返回到目标节点,这些返回的数据包含结构信息,用于计算目标节点的更新。节点的更新是基于层聚合函数的结果进行的,因此,在经过 l 层的处理后,节点的表示将被限制在其 l 跳的范围内。 这个约束要求模型在重建目标输入时,需要考虑到数据点之间的空间关系对结果的影响。

GBTSIN 模型的时空依赖层是由图注意力神经网络构成的。因此,首先简单介绍图注意力网络的基本原理。在图注意力网络中,目标节点会在空间维度上接收邻域节点信息以及自身的历史节点信

息。通过整合交叉注意力机制，可以捕获图序列的时空模式。考虑到传统的注意力网络在处理长序列时可能会遇到困难，在这里采用了掩码形式来替代传统的图注意力操作。

在图注意力网络的基础上，进一步定义 GBTSIN 模型的时空依赖层。时空依赖层的核心思想在于挖掘信号图序列的时空相关性。

2.2　缺失数据补全层

本节提出了一种自适应学习历史信号图缺失数据特征的缺失数据补全层。该模块能够捕获存储在不同层中的信息，这些信息反映了不同层次的结构特征，从而影响推理目标序列的节点，并生成缺失值。

缺失数据补全层可以被理解为在时空图 $G'_{t:t+T}$ 的序列中上执行自注意力机制，其数学模型如下：

$$r_\tau^{j\to i} = \mathrm{MLP}(h_s^j, h_\tau^i) \tag{10}$$

$$R_\tau^i = \{ r_\tau^{j\to i} \mid \langle x_s^j, q_\tau^i \rangle \} \tag{11}$$

$$c_\tau^i = \sum_{s; r_\tau^{j\to i} \in R_\tau^i} \alpha_\tau^{j\to i} \cdot r_\tau^{j\to i} \tag{12}$$

该机制能够捕获感知数据序列在不同空间窗口之间的相关性。本文所提的方法首先剔除异常数据，然后利用时空图中的边属性实现更准确的数据推理。

具体来说，使用 h_s^j 和 h_τ^i 来计算注意力分数 $r_\tau^{j\to i}$。然后用 $r_\tau^{j\to i}$ 来权衡结果，计算得到消息集合 R_τ^i。接着再计算空间上下文向量 c_τ^i 以获得全局贡献度表示 G_t。h_s^j，h_τ^i 和 $r_\tau^{j\to i}$ 线性映射为 m 次，每次都使用不同的投影矩阵将其映射到 d_m 维空间下。例如，第 r 个映射的节点通过 h_s^j 和 h_τ^i 在第 t 个和第 j 个时间窗口之间产生一个 d_m 维得分计算得到的：

$$Z_{i,j}^r = \boldsymbol{W}_r \sigma(\boldsymbol{W}_r^q q_i + \boldsymbol{W}_r^k k_j + \boldsymbol{b}_r) \tag{13}$$

其中，$\boldsymbol{W}_r \in R^{d_q \times d_q}$ 表示权值矩阵；$\boldsymbol{W}_r^q \in R^{d_z}$ 以及 $\boldsymbol{W}_r^k \in R^{d_z}$ 分别表示 h_s^j 和 h_τ^i 的投影参数矩阵；$\boldsymbol{b}_r$ 表示偏置向量；$\sigma(\cdot)$ 表示元素激活函数；$r = 1, 2, \cdots, z$。

向量 $\boldsymbol{W} = [W_1, W_2, \cdots, W_n]^{\mathrm{T}}$ 通过迭代计算得到，其中 W_i 表示节点 i 的权值，反映了在图中的重要性。第 k 次迭代的计算方法如下：

$$W_k = \alpha \boldsymbol{M}^{\mathrm{T}} \boldsymbol{W}^{k-1} + (1-\alpha)\boldsymbol{e} \tag{14}$$

其中，α 表示权重参数；$\boldsymbol{M}$ 表示一个掩码矩阵，用于区分 X_t 中的缺失值和观测值。$M_{i,j} = \dfrac{1}{O_i}$ 表示节点在时间 t 上的观测值。相反地，如果 $M_{i,j} = 0$，则表示节点在时间 t 上的观测值缺失。这里的 O_i 表示节点 v_i 的邻居节点数量。$\boldsymbol{e}$ 是一个所有元素都为 1 的列向量。

对求和层进行标准化处理，将信息输入到 Softmax 层，得到标签：

$$\hat{Y}_G = \mathrm{Softmax}\Big(W\Big(\sum_{t=1}^{k} Z_i(h_i) \odot h_i\Big) + b\Big) \tag{15}$$

其中，W 和 b 分别表示权重和偏差参数。GBTSIN 模型不仅对节点特征进行了平均加权，还在图的表示 $\hat{Y}_G$ 上应用了最大池化函数，以提取最重要的信息和固定大小的输出。

最后，定义缺失数据的重建误差 T 为：

$$T(\hat{X}_t, X_t) = \sum_{h=1}^{t} \sum_{i=1}^{N_t} \frac{\langle m_h^i,\ l(\hat{x}_h^i, x_h^i) \rangle}{\langle \hat{\bar{m}}_h^i, m_h^i \rangle} \tag{16}$$

其中，$\hat{x}_h^i$ 表示重建的 x_h^i；$\bar{m}_h^i$ 表示 m_h^i 的逻辑二进制补码；$l(\cdot,\cdot)$ 表示逐元素误差函数；$\langle \cdot, \cdot \rangle$ 表示点积。在这个过程中，缺失数据重建误差的 T 值越小，结果就越理想。

3　实验设置及结果分析

本节将对评估设置和结果展开研究分析。首先介绍了评估指标、对比方法和数据集，然后在下一小节中评估了所提出的 GBTSIN 的性能。

3.1　实验环境设置

3.1.1　数据集

本文采用 2 种流行的移动群智感知数据集评估 GBTSIN 模型的性能，即交通流量数据[20]、U-AIR 空气质量数据[21]。其中，交通流量数据是来自洛杉矶县高速公路上的 207 个速度传感器的交通测量数据，每 5 min 记录一次数据。U-AIR 空气质量数据集包含来自中国 43 个城市监测站的各种空气污染物（PM2.5、PM10、SO_2、CO 等）的每小时 IAQI 值，该数据集包含约 25.7%的缺失数据。

3.1.2　对比方法

本文使用了 4 种流行的缺失值推理算法，即 KNN-ST[13]、贝叶斯压缩感知（BCS）[22-23]、深度矩阵分解（DMF）[17] 和 IGMC[24]，来评估 GBTSIN 的有效性。

（1）KNN-ST：该方法利用 KNN 算法对最近的 K 个感知周期进行平均，并设置加权参数，以利用感知数据的时空相关性进行缺失数据的推断。

（2）BCS：贝叶斯压缩感知利用信号的稀疏性，设计了基于不同信号自适应生成的观测矩阵，解决了矩阵确定性和存储问题，并结合向量机进行信号

重构。

(3)DMF:深度矩阵分解通过多层神经网络学习不同子区域之间的非线性关系,从而恢复稀疏矩阵映射。

(4)IGMC:基于归纳图的矩阵补全是一种利用图神经网络(GNN)进行数据分析的技术,用于预测推荐系统中的未知值。

3.1.3 评价指标

在进行时空数据推理时,评估数据推理模型性能的关键问题是如何选择合适的评估指标。本研究选择了平均绝对误差(Mean Absolute Error,MAE)和均方根误差(Root Mean Square Error, RMSE)作为定量指标,以验证本文提出模型的推理精度。具体来说,MAE 是推断值与真实值之间的绝对误差的平均值,而 RMSE 则是推断值和真实值之间差异的样本标准偏差。MAE 和 RMSE 的计算方法如下:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i, \hat{y}_i| \tag{17}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{18}$$

其中,n 表示传感单元的数量,y_i 和 $\hat{y}_i$ 分别表示未感知数据的实际值和推断值。

所有的数据推理方法都是通过迭代过程来执行的,这些指标被用于训练所需的参数。本文采用了统一的停止准则进行比较:

(1)如果连续迭代之间的恢复损失偏差小于预设的阈值,迭代就会停止。

(2)如果迭代次数达到了本研究设定的最大次数、即 240 次,迭代也会停止。

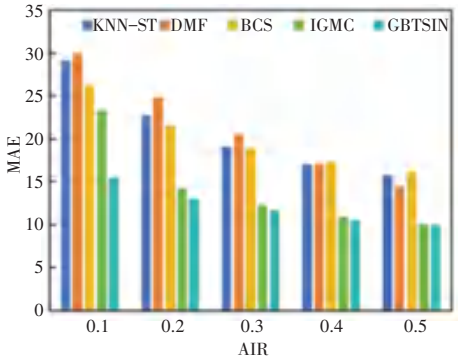
只要满足以上任一条件,迭代过程就会结束。

3.2 实验结果分析

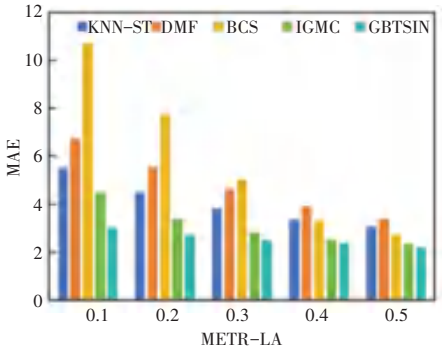
3.2.1 不同感知子区域覆盖度的推理精度对比

图 2 首先展示了 GBTSIN 模型和对比方法在不同感知子区域覆盖度的数据推理结果。研究时从数据集中随机选择一定比例的数据作为稀疏感知数据,并计算平均误差。随着感知数据的覆盖度增加(数据覆盖度从 10%到 50%),所有算法的 MAE 都相应降低,推理精度逐渐提高。

不同数据集的时空分布特征不同,IGMC 在数据高覆盖度情况下显示出优势,但其效果不如 GBTSIN 稳定。GBTSIN 方法的稀疏时空推理模型可以直接关注远距离的感知节点,减少了信息传播过程中的累积误差,同时降低计算复杂度。



(a) 数据集 1 上的结果对比



(b) 数据集 2 上的结果对比

图 2 不同感知子区域覆盖度性能对比

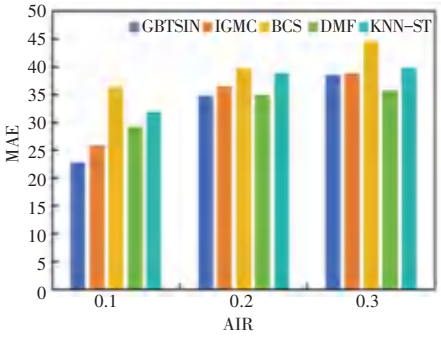
Fig. 2 Performance comparison of coverage in different sensing subregions

与 KNN-ST、BCS 和 DMF 相比,GBTSIN 可以从传感数据中提取复杂的特征,从而获得更好的数据推断精度。此外,GBTSIN 算法的推理误差较小,在实际应用场景中是可以接受的。根据图 2 中的实验结果,与其他常用的推理算法相比,本文的 GBTSIN 算法更适合补充稀疏数据。这可以归因于其良好的数据推理保真度和最小的数据推理误差。

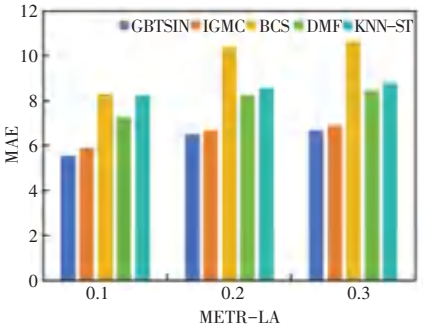
本文提出的 GBTSIN 算法在所有数据集中的推理精度误差都小于其他常用的方法,特别是在数据稀疏性较高的情况下(如感知数据覆盖度为信号图的 10%)。另一方面,由于真实数据受到整体数据分布的更大影响,分布特征具有非线性,整体效果较差。这进一步证明了 GBTSIN 模型在高缺失率场景下依旧具备很好的数据推理性能。

3.2.2 不同低质量数据比例下的的推理精度对比

通过在数据集中增加不相关数据,即噪声数据来模拟工人的感知数据集质量低的真实情况,控制低数据质量参与者占总体的比例来验证本文算法中感知数据集质量差异对任务模型训练的影响。不同低质量数据比例下的性能对比结果如图 3 所示。图 3 中显示了 5 种不同算法的 MAE 结果。



（a）数据集 1 上的结果对比



（b）数据集 2 上的结果对比

图 3　不同低质量数据比例下的性能对比

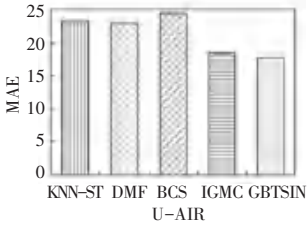
Fig. 3　Performance comparison of different low quality data ratios

本研究采用了根据不同的缺失数据分布来更新数据的实验方法，将干扰数据的概率设置为 10%～30%。这些异常数据的存在进一步增加了处理稀疏信号感知图的复杂性。对于 10% 比例的低数据质量，本文算法相比 3 种对比算法没有明显的性能下降，但是当低数据质量参与者比例继续加大时，由于噪声数据的影响，其他 3 种对比算法的准确率均开始变低。这是因为 KNN-ST、BCS 和 DMF 算法，更容易选择噪声数据，从而影响模型训练效果。综上分析表明 GBTSIN 不仅能够有效地处理高稀疏度数据，而且在处理异常数据时通过选取数据质量高的数据，减少低质量数据集参与训练，保持模型相对稳定。因此本文算法能够有效识别感知数据集中贡献度较高的数据，降低工人上传数据质量较低对模型训练的影响。

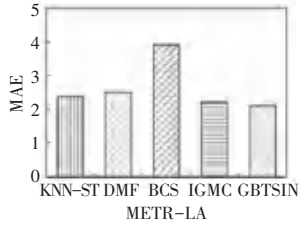
此外，本文模型对缺失数据分布的变化具有良好的鲁棒性。事实上，与基线方法相比，随着异常数据的增加，GBTSIN 的性能以较慢的速率下降，误差积累更少。进一步证明了该方法在处理复杂数据情况方面的优越性，对处理实际问题具有很高的价值。

3.2.3　模型性能对比

在本小节，柱状图被用于描述 GBTSIN 模型的数据推理性能。图 4 和图 5 直观地展示了 2 种数据集在特定缺失率下的推断值和真实值之间的差异。在 2 个数据集上，分析发现 BCS 的误差明显高于其他算法。这可能是由于 BCS 算法在处理这类数据时的一些固有的限制。相比之下，IGMC 算法的性能较好，其误差值与 GBTSIN 算法相差较小。由此说明在这些任务中，IGMC 和 GBTSIN 都可以提供相对准确的推理结果，即 GBTSIN 模型准确地学习并补全了稀疏时空数据信号图。



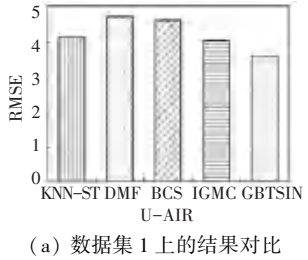
（a）数据集 1 上的结果对比



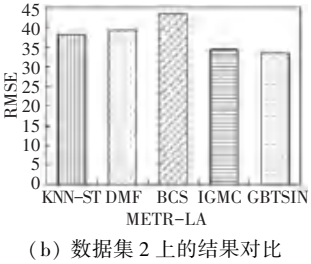
（b）数据集 2 上的结果对比

图 4　不同感知子区域覆盖度 MAE 对比

Fig. 4　Comparison of MAE for coverage of different perceptual subregions



（a）数据集 1 上的结果对比



（b）数据集 2 上的结果对比

图 5　不同感知子区域覆盖度 RMSE 对比

Fig. 5　Comparison of RMSE for coverage of different perceptual subregions

与对比试验相比，GBTSIN 模型具有最高的推理精度。GBTSIN 模型获得高精度的原因主要有 2 点。首先，时空模式层使得 GBTSIN 模型可以在时间和空间两个维度挖掘数据集中的时空模式。其次，数

据推理层使得 GBTSIN 模型具备了处理推断缺失值的能力。

3.2.4 不同因素的收敛过程分析

本小节还进一步分析了对 GBTSIN 模型的计算效率以及不同情况对其数据推理性能的影响。GBTSIN 模型是一种基于注意力机制的神经网络模型,本节在随机缺失模式、异常数据模式和块缺失模式下对其进行了评估,结果如图 6 所示。

研究中又对不同情况下的 GBTSIN 模型数据推理性能的收敛性进行了分析。随机缺失模式设置了 50%的随机缺失,用 R 来表示。同时考虑在数据集训练中设置块状缺失场景,用 B 来表示。异常数据模式设置为 10% 的干扰数据,用 M 来表示。GBTSIN 表示同时加入了时空依赖模式和缺失数据处理组件的正常数据集。

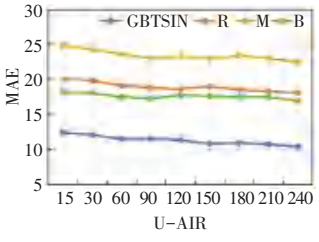


图 6 不同因素的收敛过程性能对比

Fig. 6 Performance comparison of convergence processes with different factors

结果表明,在随机缺失模式下,GBTSIN 模型的数据推理性能受到缺失率的影响较小。这是因为在随机缺失模式下,缺失值通常不会连续出现,即缺失值周围的数据仍然存在。GBTSIN 模型可以有效地捕捉缺失值的时空依赖关系,从而获得较高的数据推理精度。然而,在块状缺失模式下,GBTSIN 模型的数据推理性能受到缺失率的影响较大。随着迭代次数的增加,GBTSIN 模型的推理误差会逐渐降低。

块缺失模式下考虑块状的缺失数据与时空维度观测值的相关性下降,算法的推理精度低于 GBTSIN 算法。异常数据模式下选取 10%的干扰数据,这些噪声数据影响更新重构误差从而影响推理精度 MAE。相较而言,异常数据模式达到相应准确率所需要的聚合轮次更多,模型收敛较慢。

4 结束语

由于传感器采集技术的测量错误、设备故障、或者其他未知因素导致的数据偏差,无异常数据的情况在现实世界中很少存在。鉴于此,本文提出了一种新颖的考虑感知单元异常值的 GBTSIN 模型用于时空数据推理任务。GBTSIN 可以有效地利用收集到的稀疏感知信号图,学习其中复杂的时空特征,高精度地重构信号图,减少重构误差。

在实验部分,本文使用 2 种真实的时空数据集(交通流量数据集、空气质量指数数据集)验证了 GBTSIN 模型的推理性能。首先,本文将 GBTSIN 模型与存在的几种基线方法在 2 种数据集上进行了对比分析。实验结果表明,GBTSIN 模型在覆盖率 10%下仍然具有较好的推理性能,并在推理精度指标上优于基线方法。

最后,本文提出了一种综合评估工人上传数据质量的方法,并对模型的随机缺失和块缺失、及异常数据模式进行比较。实验结果进一步证实了本文提出的 GBTSIN 模型能有效地排除不相关数据的干扰,适用于具有缺失值的时空数据推理。

尽管本文提出的 GBTSIN 算法在信号图的重建误差上表现出了优势,但却并未考虑到实际应用中复杂场景的通信情况。此外,在处理大规模任务时,该算法的学习效率还有待提高。因此,还需要针对各种规模优化选择流程,以平衡算法训练时间和任务精度的指标要求。

参考文献

[1] TU Chunyu, YU Zhiyong, HAN Lei, et al. Adaptive budgeting for collaborative multi - task data collection in online sparse crowdsensing[J]. IEEE Transactions on Mobile Computing, 2023, 23(7): 7983-7998.

[2] LIU Y, KONG L, CHEN G. Data-oriented mobile crowdsensing: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(3): 2849-2885.

[3] WANG E, ZHANG M, YANG B, et al. Large - scale spatiotemporal fracture data completion in sparse crowdsensing[J]. IEEE Transactions on Mobile Computing, 2023, 23(7): 7585-7601.

[4] SUN Yuanhao, DING Weimin, SHU Lei, et al. On enabling mobile crowd sensing for data collection in smart agriculture: A vision[J]. IEEE Systems Journal, 2021, 16(1): 132-143.

[5] LIN Deyu, WANG Quan, MIN Weidong, et al. A survey on energy-efficient strategies in static wireless sensor networks[J]. ACM Transactions on Sensor Networks (TOSN), 2020, 17(1): 3.

[6] DINH T A N, NGUYEN A D, NGUYEN T T, et al. Spatial-temporal coverage maximization in vehicle - based mobile crowdsensing for air quality monitoring[C]//Proceedings of 2022 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway, NJ: IEEE, 2022: 1449-1454.

[7] CHEN H, GUO B, YU Z, et al. CrowdTracking: Real-time vehicle tracking through mobile crowdsensing[J]. IEEE Internet of Things Journal, 2019, 6(5): 7570-7583.

[8] LIU Wenbin, YANG Yongjian, WANG En, et al. Dynamic user recruitment with truthful pricing for mobile crowdsensing[C]//

Proceedings of the IEEE Conference on Computer Communications. Piscataway, NJ: IEEE, 2020: 1113-1122.
[9] WANG L, ZHANG D, WANG Y, et al. Sparse mobile crowdsensing: Challenges and opportunities [J]. IEEE Communications Magazine, 2016, 54(7): 161-167.
[10] WANG L, ZHANG D, YANG D, et al. SPACE-TA: Cost-effective task allocation exploitingintradata and interdata correlations in sparse crowdsensing [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2017, 9(2): 20.
[11] LIU Wenbin, WANG Leye, WANG En, et al. Reinforcement learning-based cell selection in sparse mobile crowdsensing [J]. Computer Networks, 2019, 161: 102-114.
[12] YAN Xiaobo, XIONG Weiqing, HU Liang, et al. Missing value imputation based on gaussian mixture model for the Internet of Things [J]. Mathematical Problems in Engineering, 2015, 2015: 548605.
[13] MARCHANG N, TRIPATHI R. KNN-ST: Exploiting spatio-temporal correlation for missing data inference in environmental crowd sensing [J]. IEEE Sensors Journal, 2020, 21(3): 3429-3436.
[14] JAIN N, GUPTA A, BOHARA V A. TS-MC: Two stage matrix completion algorithm for wireless sensor networks [C]// Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2019: 7590-7594.
[15] LIU Tong, ZHU Yanmin, YANG Yuanyuan, et al. Incentive design for air pollution monitoring based on compressive crowdsensing [C]//Proceedings of 2016 IEEE Global Communications Conference (GLOBECOM). Piscataway, NJ: IEEE, 2016: 1-6.
[16] YANG Bo, HE Suining, CHAN S H G. Updating wireless signal map with Bayesian compressive sensing [C]//Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems. New York: ACM, 2016: 310-317.
[17] WANG En, ZHANG Mijia, CHENG Xiaochun, et al. Deep learning-enabled sparse industrial crowdsensing and prediction [J]. IEEE Transactions on Industrial Informatics, 2020, 17(9): 6170-6181.
[18] WANG En, ZHANG M, XU Yuanbo, et al. Spatiotemporal fracture data inference in sparse urban crowdsensing [C]// Proceedings of the IEEE Conference on Computer Communications. Piscataway, NJ: IEEE, 2022: 1499-1508.
[19] BOQUET G, VICARIO J L, MORELL A, et al. Missing data in traffic estimation: A variational autoencoder imputation method [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2019: 2882-2886.
[20] WU Zonghan, PAN Shirui, LONG Guodong, et al. Graph wavenet for deep spatial-temporal graph modeling [C]// Proceedings of IJCAI International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019: 1907-1913.
[21] ZHENG Yu, LIU Furui, HSIEH H P. U-air: When urban air quality inference meets big data [C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1436-1444.
[22] HE Suining, SHIN K G. Steering crowdsourced signal map construction via Bayesian compressive sensing [C]//Proceedings of the IEEE Conference on Computer Communications. Piscataway, NJ: IEEE, 2018: 1016-1024.
[23] XIE Kun, LI Xiaocan, WANG Xin, et al. Active sparse mobile crowd sensing based on matrix completion [C]//Proceedings of the 2019 International Conference on Management of Data. New York: ACM, 2019: 195-210.
[24] ZHANG Muhan, CHEN Yixin. Inductive matrix completion based on graph neural networks [J]. arXiv preprint arXiv, 1904. 12058, 2019.