

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2023.0300

基于图神经网络的航班地面保障数据插补算法

邢志伟¹, 孙恪^{1,2}, 罗谦^{2,*}, 刘畅², 张涛², 乔迪³

(1. 中国民航大学 电子信息与自动化学院, 天津 300300; 2. 中国民航局第二研究所 工程技术研究中心, 成都 610041;

3. 四川省机场集团有限公司成都天府国际机场分公司, 成都 610041)

摘 要: 针对航班地面保障数据缺失问题, 提出一种基于图神经网络的数据插补算法。通过降噪编码器降低原始数据中噪声对训练的影响, 增强提取特征的可靠性; 建立一种图表示学习框架, 使用聚合函数聚合采样区间内节点的特征, 实现神经网络节点的状态更新, 得到第1次嵌入特征; 应用长短时记忆网络对航班的时序信息进行第2次嵌入得到隐藏层的状态空间; 通过反卷积神经网络进行特征还原, 提出一种损失函数实现网络的迭代, 在迭代多次后得到最终的航班地面保障数据插补结果。使用西南某机场2018年4~6月份的航班地面保障数据对所提算法进行测试, 结果表明: 相比于其他算法, 所提算法在低缺失率时, 插补误差平均降低了约74%; 在较高缺失率时, 插补误差平均降低了约68%; 所提算法迭代次数约在100次, 正则化系数约为0.5时, 插补误差达到最低。

关 键 词: 航班地面保障; 图嵌入; 缺失值插补; 递归神经网络; 自动编码器; 多元时间序列

中图分类号: V352; TP311.13

文献标志码: A

文章编号: 1001-5965(2025)05-1528-11

航班地面保障过程作为过站航班的关键环节, 能够反应机场运行效能, 为机场协同决策提供重要参考^[1]。在航班地面保障实际运行过程中, 保障数据采集大多依靠人工采集, 这些数据往往在采集、传送和存储过程中出现缺失^[2]。缺失的数据不仅不足以正确的表征机场运行态势, 而且较差的数据质量对下游的挖掘也会造成较大影响。数据插补方面的研究由来已久, 而对航班地面保障数据缺失问题的研究较少, 需要进一步探究。

传统的处理缺失值的方法是采用均值插补^[3]或删除该样本^[4], 而这种方法可能会潜在改变样本的分布, 出现过拟合等情况^[5], 在样本缺失率较高时, 偏差可能会在下游处理任务被放大。因此, 如何实现航班地面保障缺失数据精确插补是机场高效运行的关键问题之一。Rubin^[6]提出针对数据缺失

可能原因的讨论, 部分研究者提出期望最大化算法、最邻近算法及链式方程多元插补等基于统计的方法^[7-11]。简单的统计方法在应对样本量充足、缺失率较低的情景下会带来较大的方便, 而在某些场景中可能并不适用。

数据插补的本质是依据历史数据建模从而通过模型反馈对数据缺失值进行填补, 考虑神经网络天生对变量间的非线性关系映射的优势, 越来越多的研究者尝试将其应用到数据插补上来^[12-14]。Wang等^[15]提出改进的编码器算法, 通过对特征层正交分解, 实现了高效的特征提取。图神经网络由于整合了数据图结构特征的神经网络, 相比于传统的网络结构能够进一步把握变量间的关联, Cui等^[16]定义网络间流量的转化为一种马尔可夫过程, 较好的捕捉了网络的时空效率, 提升了预测效率并节省了计

收稿日期: 2023-05-30; 录用日期: 2023-10-16; 网络出版时间: 2023-11-21 12:59

网络出版地址: link.cnki.net/urlid/11.2625.V.20231121.1044.001

基金项目: 国家重点研发计划项目(2018YFB1601200); 国家自然科学基金青年科学基金(62203452); 中国民航大学研究生科研创新资助项目(2022YJS105)

*通信作者. E-mail: cauc_xzw@163.com

引用格式: 邢志伟, 孙恪, 罗谦, 等. 基于图神经网络的航班地面保障数据插补算法[J]. 北京航空航天大学学报, 2025, 51(5): 1528-1538.
XING Z W, SUN K, LUO Q, et al. Imputation algorithm for flight ground support data based on graph neural network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2025, 51(5): 1528-1538 (in Chinese).

算资源; Yao 等^[17] 提出一种空间交互图卷积网络模型, 将图形卷积与映射函数结合从图表示学习的角度预测数据流量; Yang 等^[18] 提出一种在线学习的框架能够对实时采集的数据进行插补, 通过将数据分解成时间和空间特征, 再分别进行处理, 但其研究的对象仍然是静态下的; Cini 等^[19] 提出图递归插补网络 (graph recurrent imputation networks, GRIN) 模型专门用于对时间序列数据的插补; You 等^[20] 通过标签预测的方法尝试使用含有缺失值的数据集直接进行预测, 并提出一种基于图表示学习的插补框架 (Grape)。国内也有部分研究者试图由低秩补全理论解决交通和气象等领域的缺失值问题^[21-23], 而这些研究缺少对数据时序的讨论。

为解决时序数据的缺失问题, Che 等^[24] 提出一种基于门控循环单元 (gated recurrent unit, GRU) 的用于插补的递归神经网络 (recurrent neural networks, RNN) (简称 GRU-D), GRU-D 将缺失机制融入到深度网络中从而能够捕捉到时序的长时依赖性, 并且实现了较精准的预测结果, Tian 等^[25] 采用多尺度平滑来推断丢失数据, 以克服采集到的交通数据长度不同、采样不规则等问题, 并提出一种基于长短期记忆 (long short term memory, LSTM) 网络的新方法 (简称 LSTM-M), 学习变量间的时间关联信息, 以实现缺失数据的插补。还有部分针对多元时间序列 (multivariate time series, MTS) 的插补模型, Luo 等^[26] 提出用于 MTS 分布建模的多元时间序列生成对抗网络 (multivariate time series-generative adversarial networks, MTS-GAN), 并对 MTS 生成任务制定约束将其应用于 MTS 插补; Spinelli 等^[27] 将编码器网络引入标准 GAN 架构中, 该架构避免了对抗阶段中对数据的预处理。而基于 GAN 的插补网络往往需要一个额外的阶段来对模型进行训

练^[28], 实际情况中数据的缺失满足训练对质量的要求, 现有方法大多是在假设数据集完整的基础上进行模型的验证, 缺少针对实际业务的考虑。

本文在上述研究的基础上, 基于 LSTM 网络提出基于图的时间序列归纳式插补 (简称 IGITS-LSTM) 算法。首先, 运用降噪自动编码器网络结构, 在编码器阶段加入随机噪声以保证提取到可靠的特征, 规避了在训练阶段对数据的预处理过程; 其次, 结合实际业务流程, 通过保障节点间信息的传递, 聚合采样区间内的节点信息, 更新当前节点的特征嵌入, 在门控神经网络中引入衰退率, 控制节点间传递的信息量; 然后, 应用 LSTM 网络捕捉多航班间的时间相似度, 得到隐藏层的特征向量; 最后, 在反卷积神经网络上将提取到的特征向量还原为原本的维度, 通过提出的损失函数对网络参数多次迭代后, 得到最终插补结果。

1 问题描述

航班地面保障过程是一个存在并行工作流程的动态混杂过程, 根据实际业务流程可大致分为航油补充、客舱服务、货仓装卸、机务巡检等主要并行流程, 各个流程之间不仅存在先后关系, 还有一定的逻辑推理次序。保障资源必须按照机场地面保障运行标准的进离场时间规定按时到位并提供保障服务。图 1 为机场航班地面保障一般作业流程。

本文针对航班地面保障过程中各个保障节点中存在的数据缺失问题, 通过算法模型对历史数据进行分析, 推断网络节点中的缺失数据, 从而提升航班保障数据的质量, 为下游任务提供充足的参考。

假设航班地面保障数据一天采集 n 个航班量, 一个航班包含 D 个保障节点。将航班地面保障数据组织为矩阵形式:

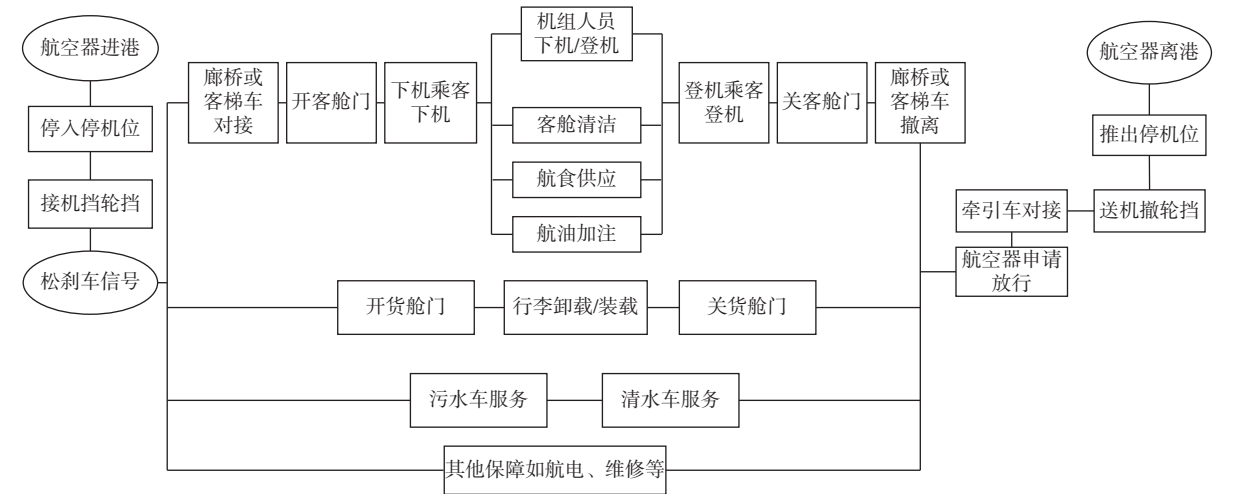


图 1 航班地面保障流程

Fig. 1 Flight ground operation process

$$X = \begin{bmatrix} x_{t_1}^1 & \cdots & x_{t_i}^1 & \cdots & x_{t_n}^1 \\ \text{null} & \cdots & \text{null} & \cdots & x_{t_n}^2 \\ \vdots & & \vdots & & \vdots \\ x_{t_1}^D & \cdots & x_{t_i}^D & \cdots & x_{t_n}^D \end{bmatrix} \quad (1)$$

式中： X 为采集到的航班地面保障流程原始数据矩阵；null 为缺失值； t_i 为第 i 架航班的到达时刻，由于不同航班之间的采样间隔并不是连续不变的，因此，通过航班到达时刻来标记一次保障过程； d 为一次保障流程中的第 d 个保障节点； $x_{t_i}^d$ 表示 t_i 时刻到达航班的保障过程中的第 d 个保障节点。

本文针对数据缺失常见的原因进行分析，一方面航班地面保障数据通常由人工收集^[2]，地服保障人员在完成保障工作后可能出现人为的失误，忘记记录时间节点导致该保障节点的数据缺失，另一方面机场数据库积累了大量的保障数据，信号干扰或网络攻击也可能造成数据的大量缺失，而且为节省存储空间对数据进行主动压缩和数据在传输过程中出现的意外缺失都是不可避免的。定义一种遮蔽矩阵 M 来表示航班地面保障数据中存在缺失的部分：

$$M = [m_{t_1}, m_{t_2}, \cdots, m_{t_n}] = \begin{bmatrix} m_{t_1}^1 & \cdots & m_{t_i}^1 & \cdots & m_{t_n}^1 \\ m_{t_1}^2 & \cdots & m_{t_i}^2 & \cdots & m_{t_n}^2 \\ \vdots & & \vdots & & \vdots \\ m_{t_1}^D & \cdots & m_{t_i}^D & \cdots & m_{t_n}^D \end{bmatrix} \quad (2)$$

其中： $m_{t_i}^d$ 定义如下：

$$m_{t_i}^d = \begin{cases} 1 & x_{t_i}^d \neq \text{null} \\ 0 & \text{其他} \end{cases} \quad (3)$$

采集的航班保障流程数据缺失率具体定义为

$$\alpha = \frac{\sum_1^n \sum_{d=1}^D (1 - m_{t_i}^d)}{nD} \times 100\% \quad (4)$$

定义 X' 为缺失恢复后的航班地面保障数据矩阵：

$$X' = \begin{bmatrix} x_{t_1}^1 & \cdots & x_{t_i}^1 & \cdots & x_{t_n}^1 \\ x_{t_1}^2 & \cdots & x_{t_i}^2 & \cdots & x_{t_n}^2 \\ \vdots & & \vdots & & \vdots \\ x_{t_1}^D & \cdots & x_{t_i}^D & \cdots & x_{t_n}^D \end{bmatrix} \quad (5)$$

因此，航班地面保障数据插补问题可以看作学习一个函数来恢复缺失的保障数据，具体可以描述为

$$F(X, M) = X' \quad (6)$$

2 基于图嵌入的航班地面保障数据插补算法

本文以图自动编码器框架为基础，分析多航班保障业务间的关联关系，提出 IGITS-LSTM 算法。首先，考虑不同航班保障过程节点间的特征相似度构建航班保障过程网络模型并通过全连接神经网络对节点的特征维度进行扩充；然后，利用嵌入式门控循环单元 (gated recurrent unit-aggregate, GRU-A) 对扩充后的张量模型分割后进行采样聚合，得到第 1 次嵌入的特征张量；其次，考虑航班间的时间关联关系，用 LSTM 网络对特征张量进行第 2 次特征提取，得到最终的中间层表示；最后，在解码层的帮助下将中间层还原为初始维度，并与初始数据进行整合，获得一个完整的数据集。整个数据插补流程示意如图 2 所示，图中， H 为特征提取过程中的中间向量。

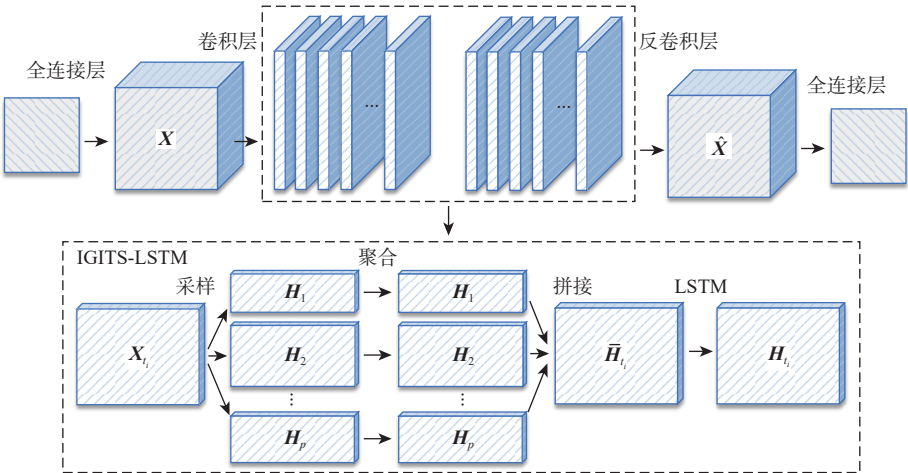


图2 航班地面保障数据插补流程示意图

Fig. 2 Schematic diagram of flight ground operation data imputation process

2.1 GRU-A 嵌入

为构建航班地面保障网络结构模型，需要先对

保障信息中的图结构信息进行提取，在航班地面保障网络拓扑图中，将每个保障环节看作是图中一个

节点, 拓扑图的邻接矩阵 $\mathbf{A}$ 由特征向量的相似矩阵 $\mathbf{S}$ 得到, 根据节点特征属性计算两两节点间的欧氏距离得到相似矩阵 $\mathbf{S}$, 对于缺失值采用局部平均的计算方式进行初始化:

$$S_i = \text{distance}(x_{t_i}^m, x_{t_i}^n) \tag{7}$$

式中: S_i 为在 t_i 下的航班保障节点 m 和 n 特征向量的欧氏距离。参考大规模流形正则化算法, 设定距离参考阈值, 当节点间的度量高于该阈值时则进行关联, 得到的邻接矩阵随后用于图嵌入过程。

图嵌入过程与基于空间域的卷积方法(简称 GraphSAGE)相似, 从航班保障流程时序图中选取具有相似特征结构的节点进行采样。与传统的知识图谱网络由某个节点的发散状拓扑结构不同,

航班地面保障网络呈现的是一张存在唯一起讫点, 从起点到终点为通路状态的拓扑结构, 符合保障过程中节点随时序向前发展的实际业务场景, 并且满足节点的排列不变性。由此, 本文提出一种全新的采样的方法, 如图 3 所示, 其中, p 和 f 分别为不同的时序通道和航班保障节点数据不同的特征维度, 根据网络拓扑结构的特点, 确定航班保障流程的初始节点(上轮挡)和结束节点(撤轮挡), 这 2 个节点往往是唯一且确定的, 依据航班地面保障时序图的邻接矩阵 $\mathbf{A}$, 采用路径搜索算法, 得到该时序图下从起始保障节点到终止节点的所有通路, 并依据采样到的路径对保障节点进行划分, 确定采样范围。

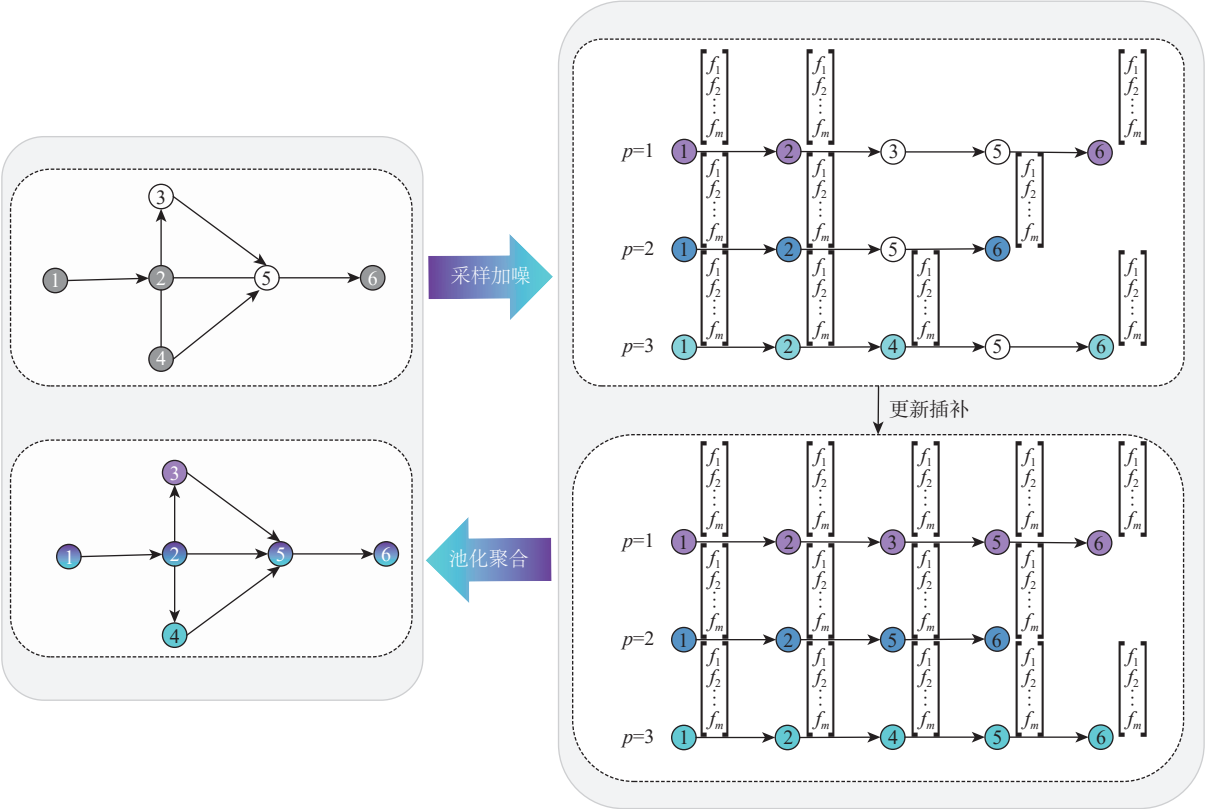


图 3 GRU-A 采样与聚合示意图

Fig. 3 Schematic diagram of GRU-A sample and aggregate

航班保障流程中节点间信息的交互方式依存于时间信息中, 本文提出用于时序插补的 GRU-A 模块, 先引入时滞向量 δ 以反映当前值与先前时序的时间滞后属性, 定义在当前值缺失状态下时间的滞后程度, 初始状态下该值默认为 0, 具体计算式为

$$\delta_d = \begin{cases} x_t^d - x_{t-1}^{d-1} & m_{t_i}^d = 1, i > 1 \\ \delta_{d-1} + x_t^d - x_{t-1}^{d-1} & m_{t_i}^d = 0, i > 1 \\ 0 & t = 1 \end{cases} \tag{8}$$

式中: δ_d 为一次保障流程中第 d 个节点的时滞;

x_t^d 为第 d 个保障节点的时间属性。

为进一步对缺失值进行合理的插补, 引入衰退率 γ 对传统的 GRU 网络进行改进, 通过控制上一个时序信息传入当前时序的比例, 实现缺失值的精确插补, 第 d 个保障节点更新时的计算衰退率 γ_d 为

$$\gamma_d = \exp(-\max(0, W_\gamma \delta_d + b_\gamma)) \tag{9}$$

式中: W_γ 和 b_γ 为带训练的参数, 指数函数负幂是使得衰退率 γ_d 在 (0,1) 范围内单调递减。

具体衰退率 γ 在 GRU-A 模块中的应用存在 2 处, 如式(10)和式(11)所示, 上一序列的数值在经

过全连接层的处理后通过衰退率 γ 对原始数据中缺失的部分进行插补。

$$\hat{\mathbf{x}}_d = \mathbf{w}_x \bar{\mathbf{h}}_{d-1} + \mathbf{b}_x \quad (10)$$

$$\bar{\mathbf{x}}_d = \mathbf{m}_{t_i} \odot \mathbf{x}_d + (\mathbf{I} - \mathbf{m}_{t_i}) \gamma_d \odot \hat{\mathbf{x}}_d \quad (11)$$

式中: $\bar{\mathbf{h}}_{d-1}$ 为上一个节点的隐藏状态; $\hat{\mathbf{x}}_d$ 为用于插补缺失值的状态估计; $\mathbf{I}$ 为单位矩阵; $\mathbf{m}_{t_i}$ 为值全为 0 或 1 的向量; $\bar{\mathbf{x}}_d$ 为保障流程中节点 d 插补后完整的特征向量; $\mathbf{w}_x$ 和 $\mathbf{b}_x$ 都是带训练的参数向量; $\odot$ 表示 Hadamard 积, 其含义表示为

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \odot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{bmatrix}$$

衰退率 γ 还作用于网络中的隐藏层, 较远间隔的数据对当前节点的嵌入参与度小于关联度较大的采样点附近的数据。GRU-A 模块的具体聚合过程如下:

$$\mathbf{h}'_d = \gamma_d \bar{\mathbf{h}}_{d-1} \quad (12)$$

$$\mathbf{r}_d = \sigma(\mathbf{W}_r \bar{\mathbf{x}}_d + \mathbf{U}_r \mathbf{h}'_{d-1} + \mathbf{V}_r \mathbf{m}_{t_i} + \mathbf{b}_r) \quad (13)$$

$$\mathbf{z}_d = \sigma(\mathbf{W}_z \bar{\mathbf{x}}_d + \mathbf{U}_z \mathbf{h}'_{d-1} + \mathbf{V}_z \mathbf{m}_{t_i} + \mathbf{b}_z) \quad (14)$$

$$\tilde{\mathbf{h}}_d = \tanh(\mathbf{W} \bar{\mathbf{x}}_d + \mathbf{U}(\mathbf{r}_d \odot \mathbf{h}'_{d-1}) + \mathbf{V} \mathbf{m}_{t_i} + \mathbf{b}) \quad (15)$$

$$\bar{\mathbf{h}}_d = (\mathbf{I} - \mathbf{z}_d) \odot \mathbf{h}'_{d-1} + \mathbf{z}_d \odot \tilde{\mathbf{h}}_d \quad (16)$$

式中: $\mathbf{h}'_d$ 为被衰退算子处理后的隐藏状态向量; $\mathbf{r}_d$ 为重置门; $\mathbf{z}_d$ 为更新门; $\tilde{\mathbf{h}}_d$ 为候选隐藏状态; $\mathbf{W}_r$ 、 $\mathbf{W}_z$ 、 $\mathbf{W}$ 、 $\mathbf{U}_r$ 、 $\mathbf{U}_z$ 、 $\mathbf{U}$ 、 $\mathbf{V}_r$ 、 $\mathbf{V}_z$ 、 $\mathbf{V}$ 为待训练的参数矩阵; $\mathbf{b}_r$ 、 $\mathbf{b}_z$ 、 $\mathbf{b}$ 为待训练的参数向量。

航班保障网络数据经过 GRU-A 模块特征提取后依然是一个子块, 且每个子块之间存在相同的节点, 本文采用均值池化的方式对一次航班保障过程的子块进行聚合, 对重合的节点特征向量取平均作为聚合后节点的特征表示, 如下:

$$\bar{\mathbf{H}}_{t_i} = \sigma(\mathbf{w} \cdot \text{MEAN}(\{\bar{\mathbf{H}}_p, p = \{1, 2, \dots, P\}\})) \quad (17)$$

式中: $\bar{\mathbf{H}}_{t_i}$ 为 t_i 时刻下航班地面保障数据隐藏状态表征矩阵; $\bar{\mathbf{H}}_p$ 为单航班地面保障网络采样后分割的子块; $\mathbf{w}$ 为待训练的参数矩阵; $\sigma(\cdot)$ 为激活函数, MEAN 为均值函数。

实际航班保障过程是一个针对多航班问题的研究, 不同航班由于不同的任务需求, 机型大小、乘客数量等都会存在较大区别, 因此, 多航班地面保障流程图的拓扑网络就是一个随时序变化的动态网络, 通过 GRU-A 对单个时序内的网络节点进行填补, 考虑到不同航班之间可能存在时间上的特征关联, 额外通过一种基于 LSTM 网络的时序插补模

型对不同航班之间时间维度上可能存在的特征关联进行提取, LSTM 网络通过“门控”机制来调节信息流。

经过嵌入后的特征向量 $\mathbf{h}_d$ 拼接后形成特征矩阵 $\mathbf{H}_t$, 再将单个航班保障数据矩阵拼接形成张量数据 $\mathbf{H}$, 用反卷积神经网络对其维度进行还原得到插补后完整的矩阵 $\hat{\mathbf{X}}$, 具体过程为

$$\mathbf{H}_{t_i} = \text{concat}(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_D) \quad (18)$$

$$\mathbf{H} = \text{concat}(\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_D) \quad (19)$$

$$\hat{\mathbf{X}} = \text{sigmoid}(\mathbf{LH}\boldsymbol{\theta}) \quad (20)$$

式中: $\mathbf{L}$ 和 $\boldsymbol{\theta}$ 为待训练的参数矩阵, sigmoid 为激活函数。基于 IGITS-LSTM 算法的航班地面保障数据插补框架如图 4 所示, 图中, $\mathbf{G}$ 表示数据图结构, 在 LSTM 模块中 $\mathbf{C}_{t-1}$ 表示上一步的记忆状态, $\hat{\mathbf{C}}_t$ 表示当前时刻下备选状态, $\mathbf{h}_t$ 表示隐藏状态, $\mathbf{f}_t$ 表示遗忘门的状态输出, $\mathbf{j}_t$ 表示输入门的状态输出, $\mathbf{o}_t$ 表示输出门的状态输出。

2.2 损失函数

为使解码器中的 $\hat{\mathbf{X}}$ 更接近于原始矩阵 $\mathbf{X}$, 需要在隐藏层中找出最好的中间向量 $\mathbf{H}$ 。而如何以更合理的数值替代矩阵中的缺失值是一个值得考虑的问题。受文献 [29] 的启发, 本文提出一种评估插补值精确度的方法。首先, 定义两部分损失函数, 损失函数的一部分是信息损失, 数据插补结果 $\hat{\mathbf{X}}$ 应该更加接近初始不完整的时间序列 $\mathbf{X}$, 另一部分是对抗损失, 迫使参数训练得到的嵌入表达更加准确; 其次, 具体描述两部分损失函数的构造细节。

信息损失由原始样本 $\mathbf{X}$ 和插补结果 $\hat{\mathbf{X}}$ 的平方差定义。由于插补出的数据本身可能受到训练次数的影响, 因此, 为减小这种不确定因素带来的影响, 本文仅仅考虑未缺失部分的方差:

$$L_{\text{info}} = \|\mathbf{X} \odot \mathbf{M} - \hat{\mathbf{X}} \odot \mathbf{M}\|_2 \quad (21)$$

为在无监督环境中学习有用的预测表示, 引入一个基于图表示学习的损失函数, 如式 (22) 所示, 通过随机梯度下降算法调节矩阵权重和聚合函数中的参数, 图嵌入中间层向量 $\mathbf{h}$ 将会学习到图更多有用的特征, 对抗损失函数会使得采样序列中离节点较近的节点有更加相似的表示, 而远离节点的表示则更加不同。

$$L_G = -\lg(\sigma(\mathbf{h}^T \mathbf{h}_{\text{adj}})) - QE_{h_{\text{adj}} \sim P(h_{\text{adj}})} \lg(\sigma(-\mathbf{h}^T \mathbf{h}_{\text{adj}})) \quad (22)$$

式中: $\mathbf{h}_{\text{adj}}$ 为固定长度上与 $\mathbf{h}$ 相邻的节点; $P(\cdot)$ 为负采样分布, E 表示负采样分布下的期望; Q 为负采样的数量。

通过损失函数来不断优化插补效果, 最后, 数

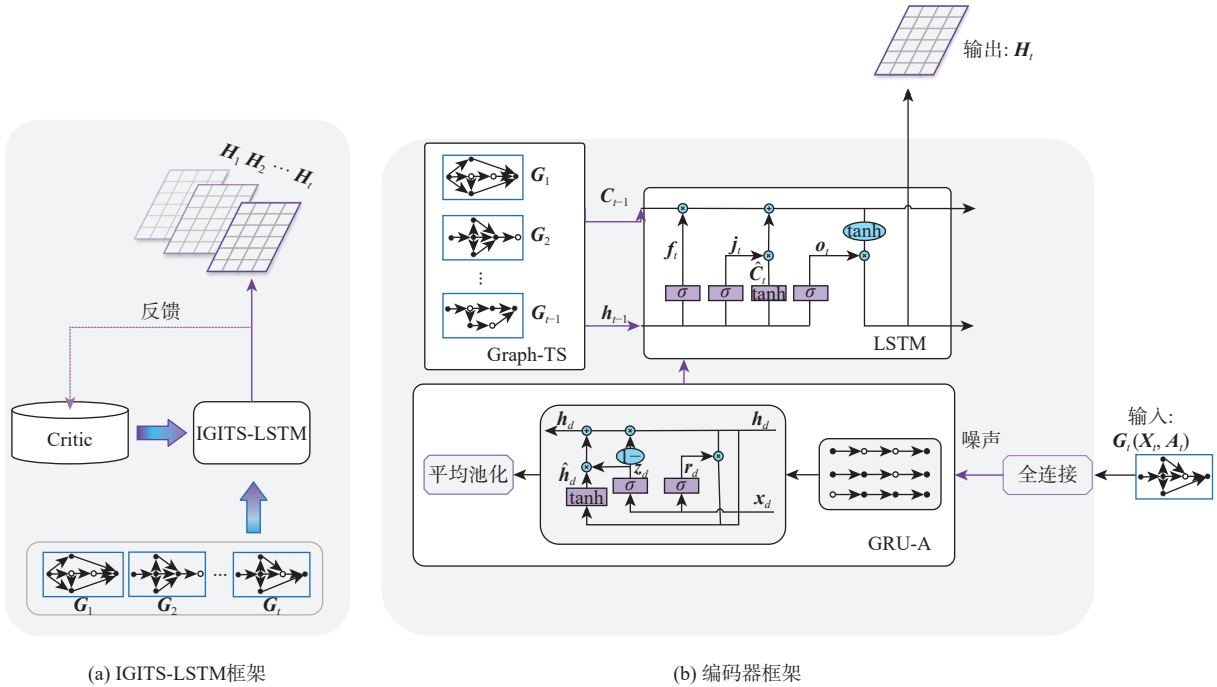


图 4 航班保障数据插补框架

Fig. 4 Flight ground support data imputation framework

据插补损失函数由信息损失和对抗损失两部分组成:

$$L = (1 - \lambda)L_{\text{info}} + \lambda L_G \quad (23)$$

其中: λ 为正则化系数, 在训练中由人工调节, 负责调节信息损失和对抗损失二者的比例。

IGITS-LSTM 算法的伪代码如下所示, 对于每一个初始的时间序列 $\mathbf{X}$, 通过一个随机遮蔽矩阵对其中的部分数据置零, 将人工处理缺失后的矩阵输入 GRU-A 模块, 得到第 1 层的嵌入向量; 再通过 LSTM 模块对数据的时间关联特征进行进一步提取, 得到隐藏层的最终向量表示; 最后, 通过解码器对中间向量处理回原本的维度 $\hat{\mathbf{X}}$, 通过损失函数 L 将结果反馈到神经网络中, 当数据插补误差不断收敛到一个可以接受的阈值内, 则可以将插补结果替换初始序列中的缺失部分如下:

$$\mathbf{X}' = \mathbf{X} \cdot \mathbf{M} + (\mathbf{I} - \mathbf{M}) \cdot \hat{\mathbf{X}} \quad (24)$$

IGITS-LSTM 算法的伪代码如下:

1. 输入 $\mathbf{X}$: 原始航班地面保障数据; $\mathbf{A}$: 航班地面保障网络邻接矩阵; epoch: 插补值最大更新迭代次数; $\mathbf{M}$: 遮蔽矩阵。
2. 输出 $\mathbf{X}'$: 插补后的航班地面保障数据;
3. $\mathbf{X}_t^d \leftarrow \text{fcn}(\{\mathbf{x}_t^d, \forall i \in (1, 2, \dots, n), d \in (1, 2, \dots, D)\})$
//扩充维度
4. for $v \leftarrow 1$ to epoch do
5. for $i \leftarrow 1$ to n do
6. $\mathbf{X}_{t_i} \leftarrow \{\mathbf{Y}_{t_i}^1, \mathbf{Y}_{t_i}^2, \dots, \mathbf{Y}_{t_i}^P\}$ //依据时序网络图

中存在的通路采样出数据子块

7. for $p \leftarrow 1$ to P do
8. $\tilde{\mathbf{Y}}_t^p \leftarrow \text{noise}(\mathbf{Y}_t^p)$ //对子块添加噪声
9. $\tilde{\mathbf{H}}_p \leftarrow \text{aggregator}(\tilde{\mathbf{Y}}_t^p)$ //对采样到的子块进行第 1 次特征嵌入
10. end for
11. $\tilde{\mathbf{H}}_t \leftarrow \text{MeanPool}(\tilde{\mathbf{H}}_1, \tilde{\mathbf{H}}_2, \dots, \tilde{\mathbf{H}}_P)$ //通过均值池化对子块进行聚合
12. $\mathbf{H}_{t_i} \leftarrow \text{LSTM1}(\mathbf{w}_1(\mathbf{H}_{t-1}, \tilde{\mathbf{H}}_t, \mathbf{C}_{t-1}) + \mathbf{b}_1)$
//对特征向量进行第 2 次嵌入
13. $\mathbf{C}_{t_i} \leftarrow \text{LSTM2}(\mathbf{w}_2(\mathbf{H}_{t-1}, \tilde{\mathbf{H}}_t, \mathbf{C}_{t-1}) + \mathbf{b}_2)$
14. end for
15. $\mathbf{H} \leftarrow \text{concat}(\{\mathbf{H}_{t_i}, i \in (1, 2, \dots, n)\})$
16. $\hat{\mathbf{X}} \leftarrow \text{sigmoid}(\mathbf{LH}\boldsymbol{\Theta})$ //对特征向量进行反卷积还原回初始维度
17. end for
18. $\mathbf{X}' \leftarrow \mathbf{X} \cdot \mathbf{M} + (\mathbf{I} - \mathbf{M}) \odot \hat{\mathbf{X}}$

3 实验分析

3.1 数据来源与评价指标

为验证本文 IGITS-LSTM 算法的有效性, 选取西南某机场 2018 年 4~6 月的航班地面保障数据进行缺失值插补实验。其 2 个月间的航班数量约为 25 400 个, 保障节点的数据缺失率在 20% 左右。航班保障过程通常包含下客、卸载行李、清洁、燃油、餐食等 42 个保障节点, 但是由于航班机型, 旅客等因素的不同, 其实际保障过程会省略其中部分

节点,因此,在航班保障网络建图过程中以实际记录数据为准。

数据插补的准确性由 3 个指标进行度量,包括平均绝对误差(mean absolute error, MAE),平均绝对百分比误差(mean absolute percentage error, MAPE)和均方根误差(root mean square error, RMSE)。

3.2 插补误差分析

在实验数据上,应用了 7 种数据插补算法进行插补,分别是平均值插补(mean imputation, MI)、K 最近邻插补(K-nearest neighbor, KNN)、矩阵分解

(matrix factorization, MF)、链式方程多重插补(multiple imputations by chained equations, MICE)及 3 种基于 RNN 改进的插补算法 GRU-D、LSTM-M、IGITS-LSTM。通过比较 MAE、MAPE 和 RMSE 来评估不同算法的插补性能。在模拟随机缺失模式下,分别设定 2 类缺失率,低缺失率为 10% 和 20%,高缺失率为 50% 和 60%,在每种缺失率下重复实验 10 次以消除随机性带来的影响,分别计算 10 次 3 种指标并取平均值。其最终实验结果如表 1 和表 2 所示,并得到如下结论。

表 1 低缺失率下每种算法的插补表现

Table 1 Imputation performance of each algorithm with low missing rates

算法	MAE/s		MAPE/%		RMSE/s	
	缺失率为 10%	缺失率为 20%	缺失率为 10%	缺失率为 20%	缺失率为 10%	缺失率为 20%
MI ^[3]	122.34	130.11	6.38	6.51	176.23	208.18
KNN ^[9]	63.49	75.32	3.17	3.59	101.33	128.04
MF ^[11]	55.38	63.91	2.51	3.12	88.60	106.09
MICE ^[10]	52.32	57.88	2.49	2.89	83.21	92.61
GRU-D ^[24]	23.42	27.31	1.17	1.14	38.64	45.06
LSTM-M ^[25]	39.24	42.55	1.96	2.12	66.38	68.93
IGITS-LSTM	11.32	12.71	0.55	0.64	23.74	25.97

表 2 高缺失率下每种算法的插补表现

Table 2 Imputation performance of each algorithm with high missing rates

算法	MAE/s		MAPE/%		RMSE/s	
	缺失率为 50%	缺失率为 60%	缺失率为 50%	缺失率为 60%	缺失率为 50%	缺失率为 60%
MI ^[3]	135.72	141.12	6.78	7.05	217.12	225.81
KNN ^[9]	88.32	100.33	4.44	4.78	143.05	167.01
MF ^[11]	68.42	75.86	3.42	3.79	113.54	125.17
MICE ^[10]	67.71	78.91	3.38	3.76	115.11	128.62
GRU-D ^[24]	35.62	50.39	1.74	2.52	65.12	83.65
LSTM-M ^[25]	53.37	67.91	2.71	3.26	90.61	111.37
IGITS-LSTM	15.22	23.21	0.73	1.77	25.42	59.86

1) MI、KNN、MF、MICE 这 4 种算法的插补准确度在高低 2 种类型的缺失率下,插补精度明显低于另外 3 种算法,其中,MI 算法精度更是远远低于 3 种常规插补算法,这可能是由于均值插补不考虑节点之间的内在关联,以固定的值插补必然会带来较大的误差,尽管 KNN、MICE 和 MF 算法考虑到了部分节点的关联属性,插补精度明显提升,但只是通过笼统的聚类规则对节点特征进行了关联,因此,提升效果有限。

2) 基于 RNN 的改进算法明显优于传统的插补算法,GRU-D、LSTM-M 和 IGITS-LSTM 这 3 种算法由于其考虑到了数据的时间属性,其对节点特征的嵌入往往能建立更多的关联,捕获到长时序中的信息传递规律,在对时序数据进行插补时能够带

来较好的效果。

3) 本文 IGITS-LSTM 算法明显优于其他算法,表中加粗部分的值表示最优结果,并且在高缺失率下这种差距变得更加明显,在 10% 和 20% 缺失率的情况下,与较先进的算法相比,IGITS-LSTM 算法在插补精度上提升了约 74%,而在 50% 和 60% 缺失率情况下,插补精度提升了约 68%。这可能是本文算法中的 GUR-A 模块能够根据实际航班地面保障的业务流程进行建图,并通过两段图嵌入算法对保障节点信息进行深度提取,还原出的数据更能接近业务的真实情况。

4) 尽管 MI 算法的数据插补性能不如其他几种算法,但是该算法在缺失率发生较大变化时,波动率却是最小的,不仅如此,基于 RNN 的算法在面对

缺失率从 20%~50% 的突变时波动变化均大于传统的数据插补算法, 而且在缺失率达到 60% 时, 所有算法插补精准度均有较大的下降, 这可能是数据量缺失较大对节点间特征关联的嵌入方法带来了一定影响。

3.3 基于插补数据的预测精度分析

为确认插补后的完整数据是否适合下游任务, 在 3.2 节 7 种算法插补后的数据集上训练了不同的回归模型, 包括决策树 (decision tree, DT) 回归器、线性回归器 (liner regressor, LR)、随机森林 (random forest, RF) 回归器、支持向量机 (support vector regression regressor, SVR) 回归器和 RNN 回归器。上述回归器均通过 scikit-learn 实现, 并将航班地面保障数据的缺失率设置在 30%, 在上述 7 种数据插补算法对缺失的数据进行补充后, 使用 MAE 对回归器在不同算法插补结果进行评价, 结果如图 5 所示, 并可以得到如下结论。

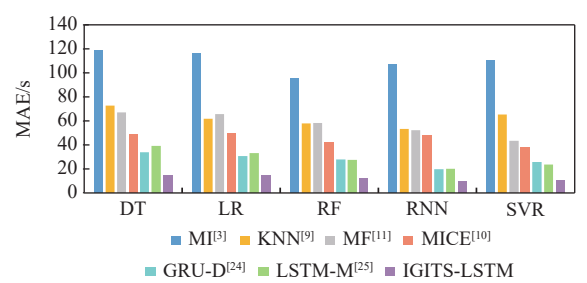


图 5 不同插补数据下不同回归模型训练得到的航班地面保障数据预测 MAE

Fig. 5 MAE of flight ground operation data prediction by different regression models trained on different imputed datasets

1) 基于 RNN 的回归模型预测指标明显整体略低于其他 4 种算法。尽管相同数据下不同预测方法的预测指标变化并不明显, 但说明了 RNN 在相同数据下预测精准度更高, 更适合航班地面保障这种时间属性的数据类型。

2) 在每种回归模型下, 插补精度较高数据大多数情况下会使得预测精度也有所提升。通过插补误差分析实验可以发现, 本文算法模型在大多数情况下能使航班保障数据的插补精度有一定的提升, 图 5 表明插补精度越高的数据下, 相同的回归模型执行预测任务效果整体呈现下降趋势, 这说明提升数据质量在大多数情况下能够对下游任务带来正向的增益。

3) 较高的数据质量, 较好的回归模型在某些情况并不一定能够带来更精准的预测结果。由图 5 可知, 在 SVR 回归模型下, 30% 的数据缺失率下, LSTM-M 插补算法带来了比 GRU-D 算法更精准的插补结果, 因此, 数据预测结果也更加精准, 但是在

DT 回归模型下结果却恰恰相反, 这说明较好的数据质量不一定能够带来较好的数据预测结果, 而在同样采取 MI 算法插补的数据上看 RNN 的预测精度甚至不如 RF, 这说明在较差的数据质量下较好的回归模型也不一定能够带来较精准的结果。出现这种现象的原因可能是在数据插补阶段的图嵌入与在预测阶段的特征提取存在一定的重合, 这种重合可能导致了模型的冗余, 使得模型预测效果进一步降低。

3.4 训练迭代次数的影响性分析

在训练阶段, 训练迭代的次数 (epoch) 总是会影响训练的结果, 因此, 需要对不同迭代次数下算法的插补精度进行实验。在训练开始前先对数据集进行处理, 分成 20%、30% 和 50% 缺失率下 3 种数据集, 再通过不同迭代次数下的模型对其进行数据插补, 观察精度的变化。如图 6 所示, 数据插补精度确实很大程度上收到了迭代次数的影响。在 20% 和 30% 缺失率下的数据集, 模型大约在 100 次迭代后达到最优精度, 在 50% 缺失率下, 模型大约在 110 次迭代后才达到最优精度, 而在较低缺失率下的模式训练已经出现过拟合; 另外在 50% 缺失下模型的训练在达到最优后呈现上下波动的趋势, 并不像在较低缺失下逐渐偏向过拟合。深度神经网络总是试图先学习数据中的低频信息然后在拟合像噪声一样的高频信息, 因此, 需要在训练阶段相对较早的停止神经网络的训练过程, 实际训练中需要尝试各种不同的 epoch 以获得最佳的训练结果。

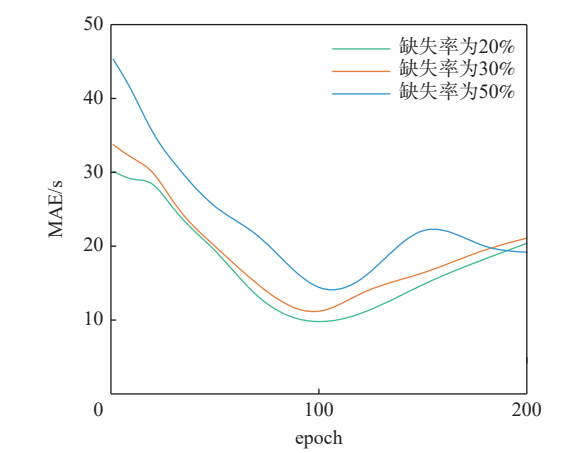


图 6 训练迭代次数的影响

Fig. 6 Effect of influence of training epoch

3.5 正则化系数影响性分析

正则化系数同样是影响模型训练的重要参数之一, 通过尝试不同的参数值来比较数据插补的结果。在 20%、30% 和 50% 数据缺失率下观察数据插补精度随超参数变化而变化的规律。如图 7 所示, 在不同缺失率下, 模型的插补精度均在 0.5 时达

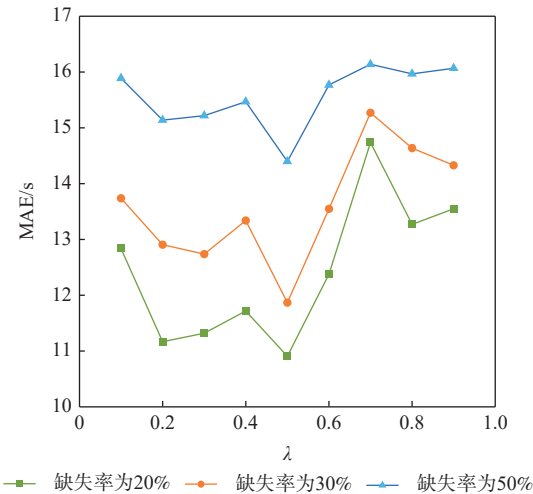


图7 正则化系数λ的影响

Fig. 7 Influence of regularization coefficient λ

到最优结果,而且随着值的增加,模型的插补精度变的越来越差。 λ 是用于平衡信息损失函数和对抗损失函数比率的参数,损失函数是用于更新模型参数的关键函数,因此,损失函数的偏差可能对训练结果带来巨大的影响,增加的值就相当于在损失函数中考虑了更多关于样本的对抗,而忽视了插补值本身的偏差,这样填补出来的数据也会逐渐偏离实际。

4 结 论

1) 本文设计了一种基于降噪自动编码器框架的航班地面保障数据插补框架,通过在训练过程中对原始数据加入随机噪声可以提升隐藏层提取特征的稳定性,同时避免了训练插补模型前对数据的预处理阶段,符合机场航班地面保障数据缺失率较大的场景。

2) 综合考虑航班地面保障实际业务过程,提出一种基于图嵌入的数据插补算法,根据航班地面保障网络中节点间的消息传递实现对缺失节点数据值的补充,同时结合多航班时间关联属性,进一步提取了特征向量,应用反卷积神经网络还原特征向量,在迭代多次模型后,可以得到相对准确的完整的航班地面保障数据。

3) 使用西南地区某机场 2018 年 4~6 月份的航班地面保障数据进行仿真实验,模拟多种缺失率,进行结果对比,仿真结果表明,本文提出的缺失值插补算法相对于较新的算法,在 10%~20% 缺失率下的插补误差平均降低了约 74%,在 50%~60% 缺失率下的插补误差平均降低了约 68%,在模型预测方面,经过该算法插补后的数据在同类回归模型中也表现较好。合适的 epoch 和 λ 能够显著提

升模型的插补效果。

由于本文模型仅仅考虑了航班地面保障数据,其数据的特征提取与插补机制源于航班保障流程的图结构特征,其适用性存在一定的局限,在未来的研究中,需要考虑如何将模型泛化到更多具有类似 MTS 属性的数据集中;保障模型在较高缺失率下也不会出现较大的插补偏差,并且由于本文只讨论了随机缺失模式下的插补效果,并没有对另外 2 种缺失模式进行验证,设计模型也没有考虑到缺失数据自身的属性,因此,挖掘数据缺失过程中的机理模型,可能有助于进一步提升模型的插补精度。

参考文献 (References)

[1] SCHMIDT M. A review of aircraft turnaround operations and simulations[J]. Progress in Aerospace Sciences, 2017, 92: 25-38.

[2] DU J H, HU M H, ZHANG W N. Missing data problem in the monitoring system: a review[J]. IEEE Sensors Journal, 2020, 20(23): 13984-13998.

[3] JUNNINEN H, NISKA H, TUPPURAINEN K, et al. Methods for imputation of missing values in air quality data sets[J]. Atmospheric Environment, 2004, 38(18): 2895-2907.

[4] HAWTHORNE G, HAWTHORNE G, ELLIOTT P. Imputing cross-sectional missing data: comparison of common techniques[J]. Australian and New Zealand Journal of Psychiatry, 2005, 39(7): 583-590.

[5] BERTSIMAS D, PAWLOWSKI C, ZHUO Y D. From predictive methods to missing data imputation: an optimization approach[J]. The Journal of Machine Learning Research, 2017, 18(1): 7133-7171.

[6] RUBIN D B. An overview of multiple imputation[C]//Proceedings of the Survey Research Methods Section of the American Statistical Association. Princeton: Citeseer, 1988.

[7] BELLO A L. Imputation techniques in regression analysis: looking closely at their implementation[J]. Computational Statistics & Data Analysis, 1995, 20(1): 45-57.

[8] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22.

[9] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[10] HUGHES R A, HERON J, STERNE J A C, et al. Accounting for missing data in statistical analyses: multiple imputation is not always the answer[J]. International Journal of Epidemiology, 2019, 48(4): 1294-1304.

[11] CHEN X Y, HE Z C, SUN L J. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation[J]. Transportation Research Part C: Emerging Technologies, 2019, 98: 73-84.

[12] 曾小华, 宋美洁, 宋大风, 等. 基于车联网信息的公交客车行驶工况数据处理方法[J]. 吉林大学学报(工学版), 2021, 51(5): 1692-1699.

ZENG X H, SONG M J, SONG D F, et al. Data processing method of bus driving cycle based on vehicular network information[J]. Journal of Jilin University (Engineering and Technology Edition),

2021, 51(5): 1692-1699(in Chinese).

[13] KUPPANNAGARI S R, FU Y, CHUENG C M, et al. Spatio-temporal missing data imputation for smart power grids[C]//Proceedings of the Twelfth ACM International Conference on Future Energy Systems. New York: ACM, 2021: 458-465.

[14] 余嘉茵, 何玉林, 崔来中, 等. 针对大规模数据的分布一致缺失值插补算法[J]. 清华大学学报 (自然科学版), 2023, 63(5): 740-753.

YU J Y, HE Y L, CUI L Z, et al. Distribution consistency-based missing value imputation algorithm for large-scale data sets[J]. Journal of Tsinghua University (Science and Technology), 2023, 63(5): 740-753(in Chinese).

[15] WANG Y X, LI K, GAN S J, et al. Missing data imputation with OLS-based autoencoder for intelligent manufacturing[J]. IEEE Transactions on Industry Applications, 2019, 55(6): 7219-7229.

[16] CUI Z Y, LIN L F, PU Z Y, et al. Graph Markov network for traffic forecasting with missing data[J]. Transportation Research Part C: Emerging Technologies, 2020, 117: 102671.

[17] YAO X, GAO Y, ZHU D, et al. Spatial origin-destination flow imputation using graph convolutional networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(12): 7474-7484.

[18] YANG J M, PENG Z R, LIN L. Real-time spatiotemporal prediction and imputation of traffic status based on LSTM and Graph Laplacian regularized matrix factorization[J]. Transportation Research Part C: Emerging Technologies, 2021, 129: 103228.

[19] CINI A, MARISCA I, ALIPPI C. Filling the g_ap_s: multivariate time series imputation by graph neural networks[EB/OL]. (2022-02-10)[2023-05-08]. <https://arxiv.org/abs/2108.00298v3>.

[20] YOU J X, MA X B, DING D Y, et al. Handling missing data with graph representation learning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 19075-19087.

[21] 陈小波, 陈程, 陈蕾, 等. 基于改进低秩矩阵补全的交通量数据缺失值插补方法[J]. 交通运输工程学报, 2019, 19(5): 180-190.

CHEN X B, CHEN C, CHEN L, et al. Interpolation method of traffic volume missing data based on improved low-rank matrix completion[J]. Journal of Traffic and Transportation Engineering, 2019, 19(5): 180-190(in Chinese).

[22] 张波, 宋国君. 大规模空气质量监测数据缺失处理方法实证研究[J]. 中国环境科学, 2022, 42(5): 2078-2087.

ZHANG B, SONG G J. Research on the missing value methods for large-scale online air quality monitoring data[J]. China Environmental Science, 2022, 42(5): 2078-2087(in Chinese).

[23] 陈帅, 赵明, 郭栋, 等. 基于 SVD-KDR 算法的工业监测数据插补技术[J]. 机械工程学报, 2021, 57(2): 30-38.

CHEN S, ZHAO M, GUO D, et al. Missing data imputation using SVD-KDR algorithm in industrial monitoring data[J]. Journal of Mechanical Engineering, 2021, 57(2): 30-38(in Chinese).

[24] CHE Z P, PURUSHOTHAM S, CHO K, et al. Recurrent neural networks for multivariate time series with missing values[J]. Scientific Reports, 2018, 8(1): 6085.

[25] TIAN Y, ZHANG K L, LI J Y, et al. LSTM-based traffic flow prediction with missing data[J]. Neurocomputing, 2018, 318: 297-305.

[26] LUO Y H, CAI X R, ZHANG Y, et al. Multivariate time series imputation with generative adversarial networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 1603-1614.

[27] SPINELLI I, SCARDAPANE S, UNCINI A. Missing data imputation with adversarially-trained graph convolutional networks[J]. Neural Networks, 2020, 129: 249-260.

[28] GUO Z J, WAN Y M, YE H. A data imputation method for multivariate time series based on generative adversarial network[J]. Neurocomputing, 2019, 360: 185-197.

[29] CHEN Y Y, LV Y S, WANG F Y. Traffic flow imputation using parallel data and generative adversarial networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(4): 1624-1630.

Imputation algorithm for flight ground support data based on graph neural network

XING Zhiwei¹, SUN Ke^{1, 2}, LUO Qian^{2, *}, LIU Chang², ZHANG Tao², QIAO Di³

- (1. Electronic Information and Automation Institute, Civil Aviation University of China, Tianjin 300300, China;
- 2. Engineering Technology Research Center, The Second Institute of Civil Aviation Administration of China, Chengdu 610041, China;
- 3. Chengdu Tianfu International Airport Branch, Sichuan Airport Group Co., Ltd., Chengdu 610041, China)

Abstract: A data imputation algorithm based on a graph neural network is proposed to address the issue of missing flight ground support data. Firstly, to reduce the impact of noise in the original data on training denoising autoencoder is applied to enhance the reliability of feature extraction. Secondly, a graph representation learning framework is established to get the first embedding, using aggregation functions to aggregate the features of nodes within the sampling interval to achieve state updating. Furthermore, a long and short-term memory neural network is constructed to embed the temporal feature of flights to obtain the final state space of the hidden layer. Lastly, a loss function is suggested to iterate the deconvolution neural network, which is employed for feature restoration. The final flight ground operation data imputation result was acquired after numerous iterations, and the technique was evaluated using ground operation data from a specific airport in Southwest China from April to June 2018. The results showed that compared to other algorithms, the proposed algorithm imputation error decreased by an average of about 74% at low missing rates. At higher missing rates, the imputation proposed algorithm error decreased by an average of about 68%. When the number of iterations of the proposed algorithm is about 100 and the regularization coefficient is about 0.5, the imputation error reaches the lowest.

Keywords: flight ground operation; graph embedding; missing value imputation; recurrent neural network; autocoder; multivariate time series

Received: 2023-05-30; **Accepted:** 2023-10-16; **Published Online:** 2023-11-21 12:59
URL: link.cnki.net/urlid/11.2625.V.20231121.1044.001
Foundation items: National Key Research and Development Program of China (2018YFB1601200); Young Scientists Fund of National Natural Science Foundation of China (62203452); Civil Aviation University of China Research Innovation Project for Postgraduate Students (2022YJS105)
*** Corresponding author.** E-mail: cauc_xzw@163.com