基于 LRTC-TNN 的瞬时水流量数据连续插值方法

赵金伟[1,2],刘杰东[1,2],邱万力[1,2],黑新宏[1,2*]
(1. 西安理工大学 计算机科学与工程学院,陕西 西安 710048;
2. 网络计算与安全技术陕西省重点实验室,陕西 西安 710048)

摘　要:瞬时水流量数据在采集、整理、存储过程中均存在不同程度的数据缺失问题,不但会造成数据分析上的偏差,还会影响后期决策,尤其是连续水流量缺失问题。国内外关于水流量数据缺失值插补的研究方法很多,然而针对相邻时间存在连续缺失值的插补问题还没有完备的解决方案。因此,基于瞬时水流量数据集的低秩假设,提出一种基于非凸低秩张量补全模型(A Nonconvex Low-Rank Tensor Completion Model-Truncated Nuclear Norm,LRTC-TNN)的瞬时水流量缺失值插补方法。通过乘子交替方向法(Alternating Direction Method of Multipliers,ADMM)求解最优的 LRTC-TNN 模型。利用通用速率参数自动确定张量模态的截断,运用张量补全的策略对连续缺失值进行预测。将该方法用于某地水厂管道瞬时水流量数据插值实验中并与其它最新的和传统的方法进行对比,取得了非常好的效果。

关键词:时间序列;水流量;缺失值插补;张量补全;低秩张量;截断核范数
中图分类号:TP18;TV737　　**文献标识码**:A　　**文章编号**:1673-629X(2023)05-0035-07
doi:10.3969/j.issn.1673-629X.2023.05.006

Continuous Imputation Method of Instantaneous Water Flow Data Based on LRTC-TNN

ZHAO Jin-wei[1,2],LIU Jie-dong[1,2],QIU Wan-li[1,2],HEI Xin-hong[1,2,*]
(1. School of Computer Science and Engineering,Xi'an University of Technology,Xi'an 710048,China;
2. Shaanxi Key Laboratory of Network Computing and Security Technology,Xi'an 710048,China)

Abstract:In the process of collection,sorting and storing instantaneous water flow data,there are different degrees of data missing,which will not only cause deviation in data analysis,but also affect later decision-making,especially the problem of continuous water flow missing. In the research on missing value imputation of water flow data at home and abroad,there is no complete solution to the imputation problem of continuous missing values in adjacent time. Therefore,based on the low rank assumption of the instantaneous water flow data set,we propose a nonconvex low rank tensor completion model-truncated nuclear norm (LRTC-TNN) to impute the missing values in the instantaneous water flow time series data. The optimal LRTC-TNN model is solved by the alternating direction method of multipliers (ADMM),and the truncation of tensor modes is automatically determined by using the general rate parameters. The missing values are predicted by using the tensor completion strategy. The proposed method is applied to the imputation experiment of instantaneous water flow data in a water plant and compared with other latest and traditional methods,which is efficient,especially for the imputation of continuous missing values.

Key words:time series;water flow;missing value imputation;tensor completion;low-rank tensor;truncated nuclear norm

0　引　言

对水厂管道瞬时水流量数据的合理利用和分析是构建城市科学供水系统的基础,是城市用水生产调度的科学依据,是解决城市用水供求关系的关键。然而,在瞬时水流量数据的采集、存储、整理等阶段容易引入缺失值。这将直接影响到水流量分析和预测等下游任务的准确性、有效性、科学性。城市供水系统、水网系统的搭建也会失去来自于数据的有效参考。国内外在

收稿日期:2023-01-07　　**修回日期**:2023-03-09
基金项目:国家自然科学基金(62176210,U20B2050,61672027);陕西省教育厅重点实验室项目(18JS076)
作者简介:赵金伟(1974-),男,博士,副教授,CCF 会员(32027M),研究方向为可解释深度神经网络、智慧水利;通信作者:黑新宏(1976-),男,博士,教授,研究方向为轨道交通、智慧水利。

时间序列数据缺失值插补问题的研究非常多,但专门针对上游任务中水流量数据连续缺失值的插补问题的研究工作相对较少。因此,急需更为有效的水流量连续缺失值的解决方案,为水流量分析与挖掘等下游任务和工业应用提供准确、完整的数据。

从数据采集的过程可知,瞬时水流量数据具有较强的时间序列特性,但同时它也与地域、人口密度、风土人情、气候特点有很大关系。如果只考虑时间特征,会发现它具有非线性和随机波动的特点,这就为瞬时水流量缺失值的插补带来了困难[1]。近年来,Kabir 等人[2]提出用单一均值插补、多重插补等方法插补水流量数据中的缺失值。王志良等人[3]提出采用各种时间序列插补方法对水文站流量缺失数据进行补全。吴若景等人[4]从时空分布的角度构建网络模型插补水文监测站径流量数据的缺失值。对瞬时水流量数据缺失值处理方法的研究,大致可分为三类。第一类,直接采用零值、均值或中位数代替瞬时水流量数据中的缺失值。第二类,采用经典的机器学习或深度学习方法插补缺失值。第三类,利用基于时间序列的经典算法、机器学习和深度学习方法对瞬时水流量数据中的缺失值进行插补。然而,第一类方法相对其它两类简单粗糙,容易造成统计特征和分布特征的偏差,其结果往往易于陷入局部最优[5-6]。第二类方法未考虑瞬时水流量数据时间序列特性,将其看作常规缺失值进行处理,无法有效利用数据本身的特点,导致插补结果不准,甚至欠拟合或过拟合。第三类方法虽然充分考虑到了瞬时水流量数据的时间序列特性,强化了数据之间的时间依赖性,但弱化了特征之间的其它相关性。

基于瞬时水流量数据集的低秩假设,该文提出一种基于非凸低秩张量补全模型(Nonconvex Low-Rank Tensor Completion Model - Truncated Nuclear Norm, LRTC-TNN)的瞬时水流量连续缺失值插补方法。创新如下:第一,在低秩张量补全模型(Low - Rank Tensor Completion Model,LRTC)的基础上引入了适用于瞬时水流量数据集的截断核范数(Truncated Nuclear Norm,TNN),该引入不仅能使 LRTC 很好地保留重要信息,也不会导致次要信息的丢失,这样 LRTC 便能更有效地提取到空间信息,掌握数据的特征相关性。第二,重新定义了瞬时水流量数据的时空关系,并利用提出的 LRTC-TNN 模型对含有连续缺失值的瞬时水流量数据进行处理。这些改进不仅能使 LRTC-TNN 完美地兼容瞬时水流量数据集,而且能充分发挥 LRTC-TNN 方法高维特性的优势,很好地利用时间序列特性把握数据的时间依赖性,同时可以有效地提取数据的特征相关性,一定程度上解决了第三类方法中存在的问题,为处理含有连续缺失值的瞬时水流量数据提供了有效方案。

1 基本原理

1.1 符号定义与标注

该文将继续沿用文献[7]中的符号定义与标注,用 $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ 和 $x_{i_1 \cdots i_d}$ 分别表示 d 阶张量及其对应的项,$\boldsymbol{X} \in \mathbb{R}^{m \times n}$ 表示矩阵,$x_i \in \mathbb{R}^n$ 表示向量,x_{ij} 表示标量。对于矩阵和张量的 *Frobenius* 范数分别用 $\| \boldsymbol{X} \|_{\mathrm{F}} = \sqrt{\sum_{ij} x_{ij}^2}$ 和 $\| \mathcal{X} \|_F = \sqrt{\sum_{i_1, i_2, \cdots, i_d} x_{i_1, i_2, \cdots, i_d}}$ 表示。用 $\mathcal{X}_{(k)} \in \mathbb{R}^{n_k \times (\prod_{i \neq k} n_i)}$ 表示张量按 $k = 1, \cdots, d$ 模式展开,$\mathrm{fold}_k(\cdot)$ 表示在 k 模式下将矩阵转换为高阶张量的折叠算子,因此有 $\mathrm{fold}_k(\mathcal{X}_{(k)}) = \mathcal{X}$。

1.2 非凸低秩张量补全模型

非凸低秩张量补全模型(Nonconvex Low-Rank Tensor Completion Model,LRTC)与低秩矩阵补全的原理类似,LRTC 本质上是建立在输入张量的低秩假设上,是一种张量处理模型。对于有部分观测值的三阶张量 $\mathcal{Y} \in \mathbb{R}^{M \times N \times T}$,LRTC 模型表示为:

$$\begin{aligned} &\min_{\mathcal{X}} \operatorname{rank}(\mathcal{X}) \\ &\text{s.t.} \quad P_\Omega(\mathcal{X}) = P_\Omega(\mathcal{Y}) \end{aligned} \tag{1}$$

其中,$\mathcal{X} \in \mathbb{R}^{M \times N \times T}$ 是希望恢复的张量,Ω 是所观测到的项的索引集,$\operatorname{rank}(\cdot)$ 表示代数秩,运算符 $P_\Omega: \mathbb{R}^{M \times N \times T} \to \mathbb{R}^{M \times N \times T}$ 表示在 Ω 上的正交投影,$P_\Omega^\perp: \mathbb{R}^{M \times N \times T} \to \mathbb{R}^{M \times N \times T}$ 表示关于 Ω 互补集的投影,且有 $P_\Omega(\mathcal{X}) + P_\Omega^\perp(\mathcal{X}) = \mathcal{X}$。在张量补全技术中,式(1)张量秩最小化问题是个 NP-hard 问题且难以计算[7]。为了解决式(1)中张量最小化问题,文献[8]参考了文献[7]中用核范数(Nuclear Norm)替换秩函数的方法,用核范数代替式(1)中的张量秩,得到新的最小化问题。

$$\begin{aligned} &\min_{\mathcal{X}} \sum_{k=1}^{3} \alpha_k \| \mathcal{X}_{(k)} \|_* \\ &\text{s.t.} \ P_\Omega(\mathcal{X}) = P_\Omega(\mathcal{Y}) \end{aligned} \tag{2}$$

其中,α_k 是权重参数且满足 $\alpha_k \geqslant 0 (k = 1, 2, 3)$。对于任意矩阵 $\boldsymbol{X}$,核范数 $\| \boldsymbol{X} \|_* = \sum_i \sigma_i(\boldsymbol{X})$,其中 $\sigma_i(\boldsymbol{X})$ 表示矩阵 $\boldsymbol{X}$ 中第 i 个最大奇异值。

1.3 LRTC-TNN

虽然核范数最小化在矩阵和张量数据的插补任务中很有效,但研究发现,通过基于奇异值的某些非凸函数有更好的效果[9-11]。由于截断核范数(Truncated Nuclear Norm,TNN)[10,12]具有更高的灵活性,不但能保留大的奇异值还会引入较小的奇异值,避免信息丢失,所以引入 TNN 来替换式(2)中的核范数,得到式(3)。

$$\min_{\mathcal{M},\mathcal{X}_1,\mathcal{X}_2,\mathcal{X}_3} \sum_{k=1}^{3} \alpha_k \| \mathcal{X}_{k(k)} \|_{r_k,*}$$
$$\text{s.t.} \begin{cases} \mathcal{X}_k = \mathcal{M}, k = 1,2,3 \\ P_\Omega(\mathcal{M}) = P_\Omega(\mathcal{Y}) \end{cases} \tag{3}$$

其中，$\| \bullet \|_{r_k,*}$ 表示每个张量在 k 模式下的截断核范数，$r_k = \lceil \theta \cdot \min\{n_k, \prod_{h \neq k} n_k\} \rceil, \forall k \in \{1,2,\cdots,d\}$，$\lceil \cdot \rceil$ 表示向上取整，θ 是一个通用速率参数，用来控制张量 $\mathcal{X}$ 按 d 模式展开的整个截断且应保证 $1 \leqslant r_k \leqslant \min\{n_k, \prod_{h \neq k} n_k\}$。适当地设置速率参数 θ，每个张量模式的截断将自动分配。辅助张量 $\mathcal{M}$ 和额外约束 $\mathcal{X}_k = \mathcal{M}, k = 1,2,3$ 将保证目标函数中变量的依赖关系，$\mathcal{M}$ 用来保留观察信息，然后将这些信息广播到张量 $\mathcal{X}_k$ 中。式(3) 中张量 $\mathcal{X}_k$ 与 TNN 相关，$\mathcal{M}$ 建立与 $\mathcal{Y}$ 的关系。

1.4　算法求解

利用乘子交替方向法(Alternating Direction Method of Multipliers, ADMM)[7,10] 求解式(3)。ADMM 以迭代的方式将张量补全优化原始问题转换为式(4)、式(5)、式(6) 三个子问题，遵循 $\mathcal{X}_1^{l+1} \rightarrow \cdots \rightarrow \mathcal{X}_3^{l+1} \rightarrow \mathcal{M}^{l+1} \rightarrow \mathcal{T}^{l+1}$ 的求解顺序。

$$\mathcal{X}_k^{l+1} = \text{fold}_k(U\text{diag}(\sigma(\mathcal{X}_{(k)}))V^{\mathrm{T}})$$
$$\sigma_i(\mathcal{X}_{(k)}) = \begin{cases} [\sigma_i \mathcal{M}_{(k)}^l - \frac{1}{\rho_k}\mathcal{T}_{k(k)}^l - \frac{\alpha_k}{\rho_k}]_+, \text{if } i > r_k \\ \sigma_i(\mathcal{M}_{(k)}^l - \frac{1}{\rho_k}\mathcal{T}_{k(k)}^l), \text{otherwise} \end{cases} \tag{4}$$

$$\mathcal{M}^{l+1} = \frac{1}{\sum_{k=1}^{3}\rho_k} \sum_{k=1}^{3} (\rho_k \mathcal{X}_k^{l+1} + \mathcal{T}_k^l) \tag{5}$$

$$\mathcal{T}_k^{l+1} := \mathcal{T}_k^l + \rho_k(\mathcal{X}_k^{l+1} - \mathcal{M}^{l+1}) \tag{6}$$

特别的，当 $\rho_1 = \rho_2 = \rho_3 = \rho$，$\mathcal{T}_k^{l+1}$ 被表示为式(7)。

$$\mathcal{T}^{l+1} := \mathcal{T}^l + \rho(\tilde{\mathcal{X}}^{l+1} - \tilde{\mathcal{M}}^{l+1}) \tag{7}$$

1.5　瞬时水流量数据时空关系

该文将二维瞬时水流量数据转换成满足“日期块 × 日期 × 时间”关系的三阶张量数据，用 $\mathcal{Y}$ 表示转换后含有缺失值的瞬时水流量张量数据，$\mathcal{X}$ 表示最终填补缺失值后的张量。将 $\mathcal{Y}$ 作为输入，通过张量补全任务式(3) 求解 $\mathcal{X}$，其本质是张量补全过程。利用算法 1 中的 ADMM 将张量补全任务式(3) 分解成迭代求解 $\mathcal{X}_k$，$\mathcal{M}$，$\mathcal{T}$ 的三个子问题。其中 $\mathcal{X}_k$ 是 $\mathcal{X}_{(k)}$ 经过基于 TNN 的奇异值分解处理后重构的张量，k 可以取 1,2,3，意味着能从不同的方向上提取 $\mathcal{X}_k$ 的各种特征。$\mathcal{M}$ 用来保留观察信息，然后将这些信息广播到张量 $\mathcal{X}_k$ 中，并对 $\mathcal{X}_k$ 进行迭代更新。$\mathcal{T}$ 在 ADMM 中用来进行对偶更新。每次迭代都会通过解 $\mathcal{X}_k$ 得到 $\mathcal{X}$，再判断 $\mathcal{X}$ 是否达到收敛条件，若达到则输出，若未达到，则不断迭代更新 $\mathcal{X}$。

1.6　基于 LRTC – TNN 的瞬时水流量数据插值算法

正如图 1 所示，为了利用 LRTC – TNN 模型将带有缺失值的瞬时水流量数据插补成完整张量，给出以下预处理步骤：

第一步，依据 1.5 节的方法通过张量展开和折叠结合奇异值分解的方式将原始瞬时水流量数据调整为具有特定规则的张量结构；

第二步，基于 LRTC – TNN 模型将张量补全任务式(3) 分解为 3 个子问题，通过迭代收敛求解这些子问题预测缺失水流量数据；

第三步，进行缺失值填补。

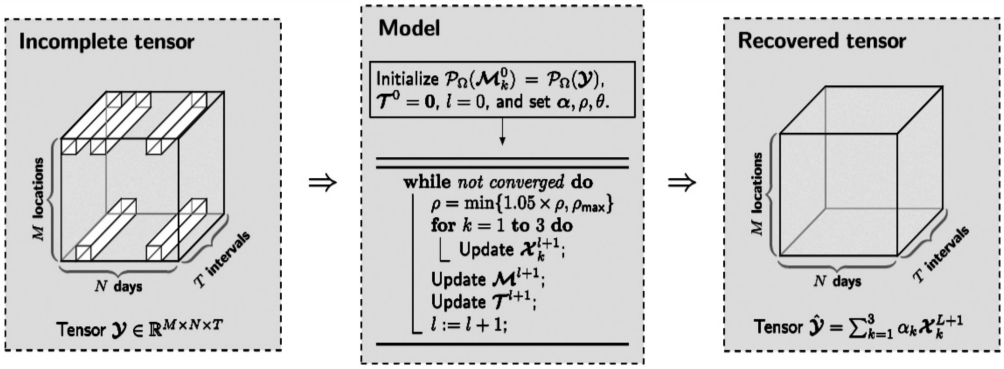

图 1　LRTC-TNN 模型进行插补的主要过程

式(3) 的求解方案正如图 1 的中间面板所示，先初始化一些必要变量和参数，再按照式(4)、式(5) 和式(7) 迭代求取 $\mathcal{X}_k^{l+1}$、$\mathcal{M}^{l+1}$ 和 $\mathcal{T}^{l+1}$，最终可求解式(3)。

该算法的伪码如下：

算法 1：基于 LRTC-TNN 的瞬时水流量数据插值算法。

说明：$\mathcal{Y}_{D \times T}$：含缺失值的矩阵　$\mathcal{X}_{D \times T}$：缺失值被填充后的矩阵

$\mathcal{Y}_{M \times N \times T}$：含缺失值的张量　$\mathcal{X}_{M \times N \times T}$：最后补全的张量

输入：y

数据关系转换：$y \rightarrow \mathcal{Y}$

初始化：$\alpha, \rho, \theta, \varepsilon$, max_inter

$\tilde{\mathcal{X}}_{3 \times M \times N \times T} = 0, \tilde{\mathcal{T}}_{3 \times M \times N \times T} = 0, \mathcal{M} = \mathcal{Y}$

It=0

While true:

$\rho = \min\{\rho \times 1.05, \rho_{\max}\}$

for k in (1,2,3):

由 $\mathcal{M}$ 和 $\tilde{\mathcal{T}}$ 更新 $\tilde{\mathcal{X}}_{[k]}$

由 $\tilde{\mathcal{X}}$ 和 $\tilde{\mathcal{T}}$ 更新 $\mathcal{M}$

由 $\tilde{\mathcal{X}}$ 和 $\mathcal{M}$ 更新 $\tilde{\mathcal{T}}$

$\mathcal{X}$ = einsum($\tilde{\mathcal{X}}$)

由 $\mathcal{Y}$ 和 $\mathcal{X}$ 计算 tol

it++

if(tol< ε) or(it ≥ max_iter):

break

return $\mathcal{X}$

数据关系转换: $\mathcal{X} \rightarrow x$

输出: x

2 实 验

2.1 实验概述

在本节中,将传统的均值插补、常用的两种机器学习方法(K 近邻和 XGBoost)、近期新提出的深度学习方法(GAIN 和 HyperImpute)以及 BTMF(Bayesian Temporal Matrix Factorization, BTMF)[13]、TRMF(Temporal Regularized Matrix Factorization,TRMF)[14]和该文提出的方法应用于瞬时水流量数据缺失值插补问题中。最后利用 MAPE(Mean Absolute Percentage Error,MAPE)和 RMSE(Root Mean Square Error,RMSE)分数对这些算法进行评估。

2.2 数据集说明

实验中采用的数据集采集于某水厂管道 183 天从第 0 时刻开始至第 22 时刻结束每隔 15 分钟记录一次的瞬时水流量。每天共记录 88 个数据项,组成一个样本,即每个样本中有 88 个特征变量,数据集中 0 代表缺失值。由于一些不可控因素,采集的数据中有多个样本中存在连续缺失值或独立缺失值。从采集过程可知,该数据具有极强的时间序列数据特性。表 1 对该数据集进行了简要表示。

表 1 瞬时水流量数据集

数据集	$N \times D$	数据缺失模式	缺失率
内容	183×88	RM 和 CM	0.001 7

在实验中,人为地对数据设置缺失率,分别为 0.3、0.6、0.8。特别需要说明的是,Jafraste 等人在研究[15]中将数据缺失模式分为三类:完全随机缺失(Missing Completely At Random, MCAR)、随机缺失(Missing At Random,MAR)和非随机缺失(Missing Not At Random, MNAR)。目前,大多研究是针对 MCAR[16-17]和 MAR[18-19]两种模式。该文将数据缺失模式总结为两类,将 MCAR 和 MAR 归为一类并记作 RM(Random Missing),用以泛指随机出现的独立缺失值,取 MNAR 中连续缺失形式作为特例,并记作 CM(Continuous Missing),表示连续缺失值。

2.3 基线模型

将引入以下基线模型与 LRTC-TNN 模型进行对比:

· 贝叶斯时态矩阵分解(Bayesian Temporal Matrix Factorization,BTMF[13]):2021 年 Chen X 等人将向量自回归(vector autoregressive,VAR)合并到传统的贝叶斯矩阵分解(Bayesian Matrix Factorization,BMF)模型中提出了 BTMF 模型,通过处理 VAR 中的系数矩阵识别和解释不同时间因素之间的因果关系,因此 BTMF 能有效处理时间序列中的缺失值。

· 时态正则化矩阵分解(Temporal Regularized Matrix Factorization,TRMF[14]):2016 年 Rao N 等人运用多重自回归过程对潜在的时间因素进行建模,用一个动态模型增强时间平滑性。BTMF 和 TRMF 都是由贝叶斯时间张量分解模型[20]推广得到。

· 均值插补:该方法将缺失值用该时刻下其它采集到的可观测数据的平均值进行填充,数据集规模大小对插补性能有直接影响。

· 中值插补:该方法将缺失值用该时刻下其它采集到的可观测数据的中值进行填充,插补性能同样受到来自数据集规模大小的直接影响。

· 最频繁数值插补:该方法将缺失值用该时刻下其它采集到的可观测数据中出现次数最多的数进行填充。

· K 近邻(K Nearest Neighbor,KNN):该方法会参考缺失值附近的可观测数据的 k 个值,并用这 k 个值的均值填充缺失值。

· XGBoost(eXtreme Gradient Boosting):对于含有缺失值的特征,通过枚举所有缺失值在当前节点是进入左子树还是右子树来决定缺失值的处理方式。

· HyperImpute[21]:2022 年 Jarrett 等人提出了一种通用的迭代插补框架,用于自适应和自动配置模型及其超参数。基于线性模型、树、XGBoost、CatBoost 和神经网络的回归和分类方法对缺失值进行迭代插补,其本质是一种集成学习方法。该方法在大量公共数据集上进行缺失值插补实验并取得了 SOTA 效果。

· 生成对抗插值网络(Generative Adversarial Imputation Networks,GAIN[22]):2018 年 Yoon 等人提出了一种基于 GAN 模型,通过生成器对缺失值进行估算,利用判别器确定实际观察到哪些成分以及哪些成分被估算,并以提示向量的形式向判别器提供额外信息的缺失值插补模型。

该文对基线模型的选择做出了如下考虑,一方面

选取遵循矩阵结构的方法,该类方法多是基于低秩时间矩阵分解的模型,它们的插值估算能充分利用时间信息和数据的低秩性,如 BTMF、TRMF。另一方面选取遵循张量结构的方法,此类方法能利用更多维度的信息进行插值估算,如文中方法。最后,还对比了传统方法、基于机器学习的方法和基于神经网络的方法。

2.4　超参数设置

参考文献[8]并结合文中数据集情况,对模型中的超参数进行合理设置。将 α_1、α_2、α_3 均设置为 1/3,用以捕捉模型处理过程中三个关键张量展开的权重。设置学习率参数 $\rho = 10^{-4}$,ρ 的大小会影响模型的收敛速度,ρ 过大会增加学习时的迭代次数从而减慢模型的收敛速度,可能会导致过度学习,ρ 过小又会因为迭代次数过少使模型收敛速度过快,可能导致学习不足。设置 $\theta = 0.3$,θ 是一个通用速率参数,又称核范数截断率参数,用于控制输入在按某模式张量展开时的整个截断。θ 和 ρ 的设置直接决定了模型性能。设置 $\varepsilon = 10^{-4}$,ε 是用于求解模型的迭代收敛条件。设置模型最大迭代次数为 200。在以上超参均不变的情况下,分别设置缺失率 0.3、0.6、0.8 用于对模型性能进行评估,同时也可作为对照实验。将 183×88 的数据集转换成 3×61×88 的输入张量,即是将二阶张量数据集进行抽象的升维形成三阶张量,使原数据集由 1 个水厂 183 天的瞬时水流量记录抽象成 3 个日期块 61 天的瞬时水流量记录,虚拟出了一个“日期块”维度,使之符合 LRTC-TNN 模型的输入层要求。对于 BTMF 和 TRMF 中的参数,分别参照了文献[13-14]中实验进行合理设置。将 K 近邻法中的 K 值设置为 3,XGBoost 中参数依照数据集情况合理设置。GAIN 和 HyperImpute 中的参数均由文献[21]中框架自动配置。以上实验在同一随机数种子下进行。

2.5　实验结果分析

由于数据集中的缺失值是完全随机出现的,有些缺失值只单独出现,称为独立缺失值,有些缺失值则相邻出现,称为连续缺失值。鉴于文中方法插值效果在整个数据集上的良好表现且独立缺失值在瞬时水流量数据集中只占 0.000 3,该文只对不同缺失率下连续缺失值的样本进行实验,用于说明连续缺失值插补情况,具体参阅图 2 ~ 图 4,横轴表示第几次采集,纵轴表示瞬时水流量,左边面板代表连续缺失值出现在中间时刻样本的插补情况,右边面板代表连续缺失值出现在采集开始时刻样本的插补情况,图中曲线中平行于横轴的线段即为缺失值。可知,随着数据集中缺失率的提高,插值曲线与真实曲线的重叠越来越少,且波动越来越剧烈,意味着随着可用信息的减少,文中方法的插值效果也会变差。其次,还能看到在缺失率低于 0.6 的情况下,文中方法对连续缺失值的插补效果都比较稳定,无剧烈波动,当缺失率达到0.8的时候,插值曲

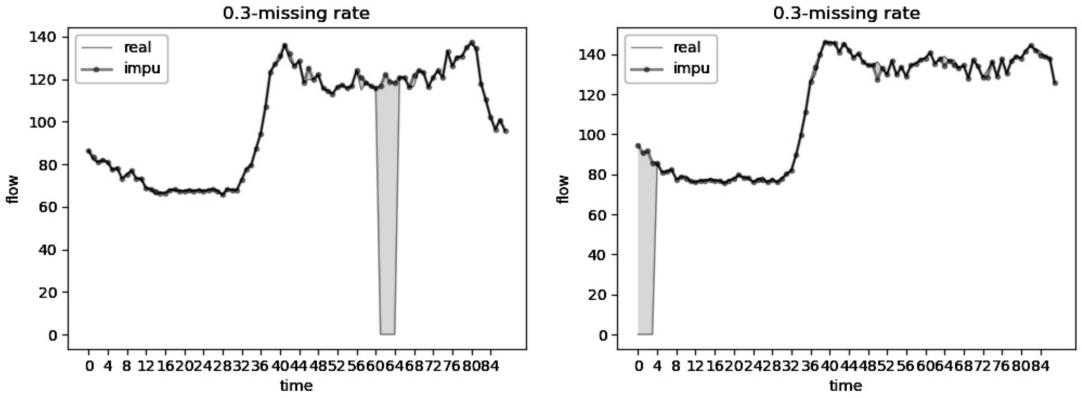

图 2　0.3 数据缺失率

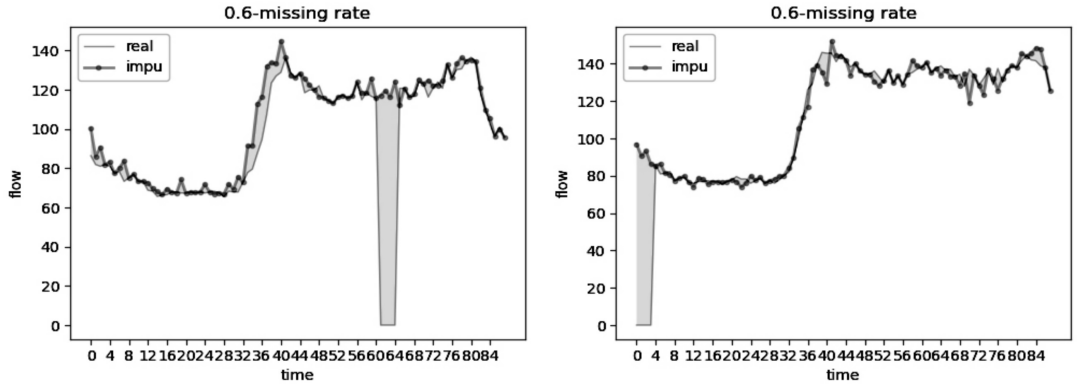

图 3　0.6 数据缺失率

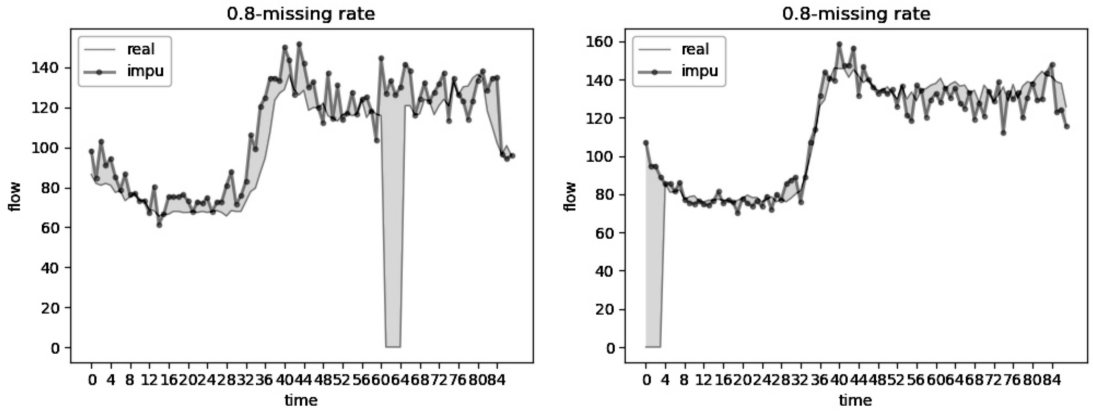

图 4 0.8 数据缺失率

线在真实曲线附近上下波动剧烈,此时的插值效果已经差强人意。除此之外,右边面板插值曲线与真实曲线的重叠情况在不同的缺失率情况下都要比左边面板好,间接说明了文中方法并不完全依赖之前时间步这一单独因素对缺失值进行插补,还会综合其它维度信息对缺失值进行插补。

表 2 展示了不同缺失率下各算法的 MAPE 和 RMSE 值,得分的右上角标代表了性能排名,1 为最优。除了传统的均值、中值、最频繁插值方法外,可以了解到其它所有算法的缺失率越高,MAPE 和 RMSE 值也越高,揭示了模型的误差也越大。在缺失率为 0.3 的时候,可见 KNN 的 MAPE 和 RMSE 值都要比文中方法和基于神经网络方法的低,有更好的性能,但随着缺失率的增大,KNN 的各项也急剧变大,性能明显变差,这也符合 KNN 方法的特点。在不同缺失率下,文中方法的插值性能都能进入前三且接近甚至优于一些基于神经网络方法的 SOTA 模型。此外,发现传统的缺失值插补方法在不同缺失率下,MAPE 和 RMSE 值都远高于其它方法,且几乎无变化,这也暗示了此类插补方法并不适合瞬时水流量数据的连续缺失值插补任务,下文将不再对其进行讨论。

表 2 各模型的表现比较(MAPE/RMSE)

方法\缺失率		0.3	0.6	0.8
传统方法	MEAN	0.147/24.19	0.149/24.17	0.149/24.32
	MEDIAN	0.125/25.72	0.125/25.50	0.126/25.76
	MostFrequent	0.137/29.75	0.135/29.08	0.134/28.88
机器学习方法	KNN	$0.029^{1}/5.36^{1}$	0.044/8.59	0.138/20.03
	XGBoost	0.033/6.85	0.048/9.32	0.106/18.72
神经网络方法	HyperImpute	$0.033^{3}/6.06^{3}$	$0.032^{1}/5.32^{1}$	$0.035^{1}/6.61^{1}$
	GAIN	0.038/6.33	$0.033^{2}/5.85^{2}$	$0.062^{3}/11.19^{3}$
最优化方法	BTMF	0.033/7.19	0.045/14.77	0.069/19.17
	TRMF	0.039/7.55	0.067/12.76	0.121/20.54
	Ours	$0.032^{2}/5.93^{2}$	$0.043^{3}/7.92^{3}$	$0.061^{2}/10.50^{2}$

图 5 中揭示了各插补方法在不同缺失率下 MAPE 和 RMSE 值的变化趋势。显然,在不同缺失率下文中方法都有接近基于神经网络方法的先进表现,尤其是在高缺失率下,MAPE 和 RMSE 值都是平稳变化,表现出了与 HyperImpute 方法同样的稳定性。而 TRMF、KNN 和 XGBoost 的各项值会随着缺失率的提高出现剧烈的变化,当缺失率达到 0.6 的时候,斜率明显增大,各项值的变化幅度急剧变大,在高缺失率下,性能明显变差。与文中方法类似,虽然 BTMF 在不同缺失率下各项值的变化都较为平稳,但整体而言,MAPE 和 RMSE 值都高于文中方法,性能差于文中方法。另外值得一提的是,文中方法在缺失率为 0.3 和 0.8 的时候,各项值都低于 GAIN 方法,性能略优于 GAIN 方法,其次,GAIN 方法随着缺失率的增加,其斜率明显变大,而文中方法斜率变化却非常小,基本接近 SOTA 模型 HyperImpute 方法,这也说明了文中方法具有较强的稳定性。从本质上进行分析,HyperImpute 方法是一种网络式的集成学习方法,主要依靠数据驱动,且需人为设置的超参以及网络训练过程中需要更新的参数较多,极易受到超参和数据量大小的影响,而提出的 LRTC-

TNN 方法是一种基于一定数学先验知识构建数学模型解决问题的单一算法,对数据和超参的依赖并不如 HyperImpute 那般明显。通过实验可知,LRTC-TNN 方法性能接近 HyperImpute 模型性能,也反映了所构建的数学模型的合理性和科学性。

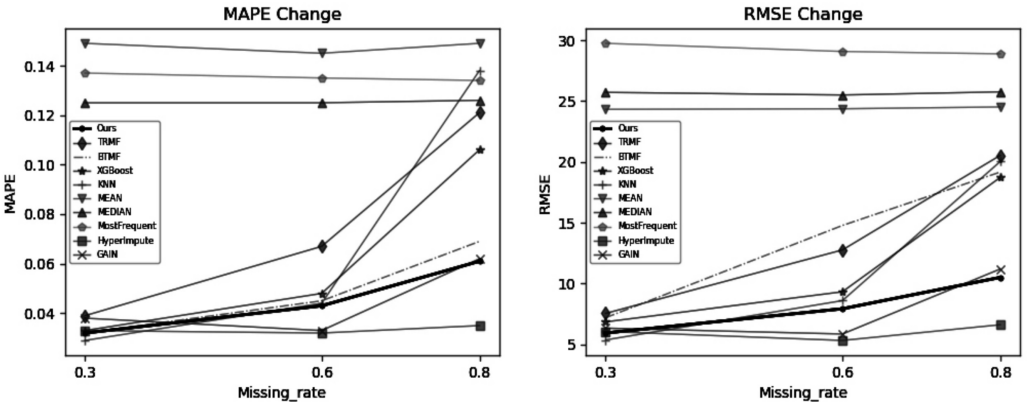

图 5　MAPE 和 RMSE 值变化

3　结束语

通过实验展示了含有连续缺失值的瞬时水流量数据在不同缺失率情况下的插补情况,论证了基于 LRTC-TNN 的瞬时水流量数据插值方法在数据集低缺失率情况下的良好性能,在较高缺失率情况下仍有较强的稳定性,体现在 MAPE 和 RMSE 分数随缺失率的提高没有出现强烈波动。基于 LRTC-TNN 的瞬时水流量数据插值方法在瞬时水流量数据缺失值插补任务中的良好表现也间接说明了对瞬时水流量数据集的低秩假设是合理的且能高效捕捉数据的时间依赖性和特征相关性,进而对缺失值进行有效估算。该方法取得了较好的插值效果,一方面说明了该方法对低维时间序列数据具有很好的兼容性,另一方面也为处理瞬时水流量数据的连续缺失值问题提供了有效方案。

参考文献:

[1] ZHU X,CHEN J. Urban water consumption forecast based on PQPSO-LSSVM[C]//2013 ninth international conference on natural computation (ICNC). Shenyang: IEEE, 2013: 834-837.

[2] KABIR G, TESFAMARIAM S, HEMSING J, et al. Handling incomplete and missing data in water network database using imputation methods[J]. Sustainable and Resilient Infrastructure,2020,5(6):365-377.

[3] 王志良,黄　珊,陈海涛. 黄河流域水文数据插补方法比较及应用[J]. 人民黄河,2020,42(7):14-18.

[4] 吴若景,杨　勇,鲍振鑫. 黄淮海流域河流空间网络构建及径流量时空插值[J]. 华北水利水电大学学报:自然科学版,2019,40(6):7-14.

[5] BEAULIEU-JONES B K,MOORE J H. Missing data imputation in the electronic health record using deeply learned autoencoders[J]. Pac Symp Biocomput,2016,22:207-218.

[6] RYU S,KIM M,KIM H. Denoising autoencoder-based missing value imputation for smart meters[J]. IEEE Access,2020,8: 40656-40666.

[7] LIU J,MUSIALSKI P,WONKA P,et al. Tensor completion for estimating missing values in visual data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):208-220.

[8] CHEN X, YANG J, SUN L. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation [J]. Transportation Research Part C: Emerging Technologies, 2020,117:102673.

[9] YAO Q,KWOK J T,WANG T,et al. Large-scale low-rank matrix learning with nonconvex regularizers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(11):2628-2643.

[10] HU Y,ZHANG D,YE J,et al. Fast and accurate matrix completion via truncated nuclear norm regularization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012,35(9):2117-2130.

[11] GU S,LEI Z,ZUO W,et al. Weighted nuclear norm minimization with application to image denoising[C]//2014 IEEE conference on computer vision and pattern recognition (CVPR). Columbus:IEEE,2014:2862-2869.

[12] ZHANG D,HU Y,YE J,et al. Matrix completion by truncated nuclear norm regularization[C]//2012 IEEE conference on computer vision and pattern recognition. Providence: IEEE, 2012:2192-2199.

[13] CHEN X,SUN L. Bayesian temporal factorization for multidimensional time series prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,44(9):4659-4673.

[14] YU H F,RAO N,DHILLON I S. Temporal regularized matrix factorization for high-dimensional time series prediction[J]. Advances in Neural Information Processing Systems,2016,29

(下转第 87 页)