

一个基于 d 阶张量奇异值分解的互联网流量数据恢复方法^{*}

段宇轩, 刘金杰

(重庆师范大学 重庆国家应用数学中心, 重庆 401331)

摘要:精准恢复互联网流量数据能降低不完全数据对网络任务过程的损害,因此,针对互联网流量数据的相似性和周期性这一时空特性,基于 d 阶张量奇异值分解(d 阶 T-SVD),结合时空正则化策略,对具有四阶张量结构性质的互联网流量数据构建了恢复模型。这一模型的核心在于能够深入挖掘数据的同时保留了内部复杂的结构特性,从而实现更高质量的数据恢复。采用交替极小化方法,设计了一种高效且稳定的算法,以实现模型的精确求解。最后选取了 2 个真实的互联网流量数据集,模拟随机性缺失和结构性缺失的数据场景,对提出的方法进行了全面验证。实验结果表明,该方法在数据恢复性能上相较于现有方法,展现出明显的优势。

关键词:互联网流量数据恢复; d 阶张量奇异值分解;时空正则化;交替极小化;张量填充

中图分类号:O221.2

文献标志码:A

文章编号:1672-6693(2024)04-0084-10

互联网流量数据,作为一类蕴含鲜明时空特性的结构数据,核心在于记录从网络起始点到接收点(即 OD 对)在特定时间内的信息传输量。一个完整的流量数据在互联网工程设计、管理中至关重要^[1-3]。然而,现实情况中,由于多种不可控因素的存在,获取的流量数据往往是不完整或受污染的,这严重阻碍了数据的有效利用。因此,如何从不完整的流量数据中尽可能准确恢复原有数据具有重要意义。

众所周知,基于数据低秩特性的优化模型是恢复缺失数据的有效手段。自 Candès 等人^[4]于 2009 年开创性提出低秩矩阵恢复模型以来,这一领域的研究不断深入。与此同时,作为矩阵的高阶扩展,张量以多维特性为数据恢复提供了新的视角。然而,与矩阵低秩模型不同的是,张量秩的定义依赖于具体的分解方法。常见的分解方法有 CANDECOMP/PARAFAC(CP)分解^[5-6]、Tucker 分解^[7-8]、张量火车(TT)分解^[9]以及张量奇异值分解(T-SVD)^[10]。除此之外,基于张量分解建立的数据恢复模型能够在一定程度上保留数据结构信息,已被广泛应用于图像处理、计算机视觉等工程领域^[11-13]。每种方法各有千秋,但也存在局限性。特别针对互联网流量数据拥有的特殊时空特点(如网络节点分布、流量数据周期性),直接照搬通常的恢复方法时,效果并不理想。

近些年来,研究者们通过探索张量分解结合时空正则化的数据恢复模型,以更好地适应互联网流量数据的特性。Zhang 等人^[14]通过矩阵的稀疏正则化分解,首次结合时空正则化,建立了对流量数据恢复的 SRMF 模型。Acar 等人^[15]基于张量 CP 分解,提出了 CP 加权最优化(CPWOPT)方法;Zhou 等人^[16]结合 CPWOPT,融合时空正则化信息,创新性提出了 STTC 方法。此外,基于 TT 分解,Zhang 等人^[17]结合时空正则化,利用张量的全局相关性解决流量数据缺失问题。Ling 等人^[18]通过三阶张量 T-积分分解,结合时空正则化建立了流量数据恢复的 TCTF2R 模型。需要指出的是,当前流量数据恢复的研究主要基于三阶张量进行模型构建。虽然这些研究取得了一定的进展,但在恢复具有结构性缺失特点的数据(如在特定时间段缺失)时,效果有待改进。事实上,在结构性缺失的情况下,恢复模型需更加充分考虑数据内部间的结构特性,如何有效恢复流量数据仍是一个挑战。

值得注意的是,2021 年,Qin 等人^[19]将三阶张量奇异值分解推广到 d 阶张量奇异值分解(d 阶 T-SVD),对 d 阶张量进行可逆变换,更加充分地融合了数据内部之间的结构特征。鉴于互联网流量数据的复杂性和多维性,流量数据可以通过四阶张量形式进行组织,以更全面地捕捉数据的内部结构特征,即流量数据 $\mathcal{X} \in$

* 收稿日期:2024-05-21 网络出版时间:2024-10-11T10:23

资助项目:重庆英才计划“包干制项目”(No. cstc2022ycjh-bgzxm0040);重庆市教育委员会科学技术研究项目(No. KJZD-K202200506);重庆师范大学校级基金项目(No. 22XLB005);重庆国家应用数学中心科研项目(No. ncamc2022-msxm02)

第一作者简介:段宇轩,男,研究方向为运筹学与控制论,E-mail:duanyuxuan@stu.cqnu.edu.cn;通信作者:刘金杰,女,副教授,博士,E-mail:jinnie.liu@cqnu.edu.cn

网络出版地址:https://link.cnki.net/urlid/50.1165.N.20241010.1945.013

$\mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$, 其中 n_1, n_2 代表数据观测的始发点和接收点, n_3 表示一天内测量的次数, n_4 代表测量天数。同时, 互联网流量数据具有相似性以及周期性。受此启发, 本文基于 d 阶张量奇异值分解, 结合时空正则化策略, 建立了对四阶互联网流量数据恢复的优化模型。通过在张量的第 3 个以及第 4 个维度上进行可逆变换, 充分融合数据内部之间的结构特性, 结合时空正则化, 提供了一种有效方法来恢复缺失数据, 特别是在处理结构性缺失数据时展现出明显优势。本文还设计了基于交替极小化方法的求解算法, 并通过在 2 个真实互联网流量数据集上的实验验证了所提方法的有效性和优越性。

1 符号说明与预备知识

本文用小写字母 $a, b, c \dots$ 表示标量, 加粗小写字母 $\mathbf{a}, \mathbf{b}, \mathbf{c} \dots$ 和加粗大写字母 $\mathbf{A}, \mathbf{B}, \mathbf{C} \dots$ 分别表示向量和矩阵, 大写花体字母 $\mathcal{A}, \mathcal{B}, \mathcal{C} \dots$ 表示张量。实数域和复数域分别记为 $\mathbf{R}$ 和 $\mathbf{C}$, 单位矩阵记为 $\mathbf{I}$, 零张量记为 $\mathcal{O}$ 。对于任意的正整数 n , 定义 $[n] := \{1, 2, \dots, n\}$ 。对于 d 阶实张量 $\mathcal{A} = (a_{i_1 i_2 \dots i_d}) \in \mathbf{R}^{n_1 \times n_2 \times \dots \times n_d}$, 相应的 Frobenius 范数定义为 $\|\mathcal{A}\|_F = \left(\sum_{i_1 \dots i_d} |a_{i_1 \dots i_d}|^2 \right)^{1/2}$ 。记 $\mathcal{A}(i_1, \dots, i_{p-1}, :, i_{p+1}, \dots, i_{q-1}, :, i_{q+1}, \dots, i_d)$ 为 $\mathcal{A}$ 沿模- p 和模- q 的

矩阵切片。特别地, 张量 $\mathcal{A}$ 沿前 2 个模态的矩阵切片记为 $\mathbf{A}^j = \mathcal{A}(:, :, i_3, \dots, i_d) \in \mathbf{R}^{n_1 \times n_2}$, 其中:

$$j = (i_d - 1)n_3 \dots n_{d-1} + \dots + (i_4 - 1)n_3 + i_3, i_d \in \{1, \dots, n_d\}.$$

$\text{bdiag}(\cdot)$ 算子表达式为:

$$\text{bdiag}(\mathcal{A}) = \text{diag}(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^{n_3 n_4 \dots n_d - 1}, \mathbf{A}^{n_3 n_4 \dots n_d}) \in \mathbf{R}^{n_1 n_3 \dots n_d \times n_2 n_3 \dots n_d}.$$

$\mathcal{A} \triangle \mathcal{B}$ 表示 d 阶张量 $\mathcal{A} \in \mathbf{R}^{n_1 \times \dots \times n_d}$ 和 $\mathcal{B} \in \mathbf{R}^{n_1 \times \dots \times n_d}$ 沿前 2 个模态矩阵切片相乘, 即:

$$\mathcal{C} = \mathcal{A} \triangle \mathcal{B} \Leftrightarrow \text{bdiag}(\mathcal{C}) = \text{bdiag}(\mathcal{A}) \cdot \text{bdiag}(\mathcal{B}).$$

矩阵 $\mathbf{A}_{(k)} \in \mathbf{R}^{n_k \times \prod_{i \neq k} n_i}$ 表示张量 $\mathcal{A}$ 的模- k 展开, 张量 $\mathcal{A}$ 中第 $(i_1, i_2, \dots, i_d)$ 元素映射为 $\mathbf{A}_{(k)}$ 中第 (i_k, j) 元素, 其中:

$$j = 1 + \prod_{l=1, l \neq k}^d (i_l - 1) J_l, J_l = \prod_{t=1, t \neq k}^{l-1} n_t.$$

记 $\mathcal{A}(i_1, \dots, i_{l-1}, :, i_{l+1}, \dots, i_d) \in \mathbf{R}^{n_l}$ 为张量 $\mathcal{A}$ 的模- l 纤维。张量 $\mathcal{A}$ 在矩阵 $\mathbf{U}$ 下的模- n 积表示为 $\mathcal{A} \times_n \mathbf{U}$, 对于 $\mathcal{B} = \mathcal{A} \times_n \mathbf{U}$ 有 $\mathbf{B}_{(n)} = \mathbf{U} \cdot \mathbf{A}_{(n)}$ 。对于任意可逆矩阵 $\mathbf{U}_{n_i} \in \mathbf{C}^{n_i \times n_i}, i = \{3, 4, \dots, d\}$, 定义 d 阶张量 $\mathcal{A} \in \mathbf{C}^{n_1 \times \dots \times n_d}$ 的任意可逆变换 $L(\cdot): \mathbf{C}^{n_1 \times \dots \times n_d} \rightarrow \mathbf{C}^{n_1 \times \dots \times n_d}$ 为 $\mathcal{A}_L \triangleq L(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{U}_{n_3} \times_4 \mathbf{U}_{n_4} \times_5 \dots \times_d \mathbf{U}_{n_d}$, 且在 L 下的逆变换定义为 $L^{-1}(\mathcal{A}) = \mathcal{A} \times_d \mathbf{U}_{n_d}^{-1} \times_{d-1} \mathbf{U}_{n_{d-1}}^{-1} \times_{d-2} \dots \times_3 \mathbf{U}_{n_3}^{-1}$, 特别地, $L^{-1}(L(\mathcal{A})) = \mathcal{A}$ 。

下面给出关于 d 阶张量 T-SVD 相关定义。

定义 1 (d 阶 T-积^[19]) 对于 d 阶张量 $\mathcal{A} \in \mathbf{R}^{n_1 \times l \times n_3 \times \dots \times n_d}$ 和 $\mathcal{B} \in \mathbf{R}^{l \times n_2 \times n_3 \times \dots \times n_d}$, 在任意可逆变换 L 下的 T-积定义为 $\mathcal{C} = \mathcal{A} \otimes_L \mathcal{B} = L^{-1}(\mathcal{A}_L \triangle \mathcal{B}_L) \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times \dots \times n_d}$ 。

定义 2 (d 阶共轭转置^[19]) 对于 d 阶张量 $\mathcal{A} \in \mathbf{C}^{n_1 \times n_2 \times n_3 \times \dots \times n_d}$, 如果对于所有 $i_j \in [n_j], j \in \{3, \dots, d\}$ 满足 $\mathcal{A}_L^*(:, :, i_3, \dots, i_d) = (\mathcal{A}_L(:, :, i_3, \dots, i_d))^*$, 则称 $\mathcal{A}^* \in \mathbf{C}^{n_2 \times n_1 \times n_3 \times \dots \times n_d}$ 为 $\mathcal{A}$ 的共轭转置。

定义 3 (d 阶正交张量^[19]) 对于 d 阶张量 $\mathcal{Q} \in \mathbf{R}^{n \times n \times n_3 \times \dots \times n_d}$, 如果满足 $\mathcal{Q}^* \otimes_L \mathcal{Q} = \mathcal{Q} \otimes_L \mathcal{Q}^* = \mathcal{I}$, 则称 $\mathcal{Q}$ 为正交张量。其中 $\mathcal{I} \in \mathbf{R}^{n \times n \times n_3 \times \dots \times n_d}$ 为单位张量, 即 $\mathcal{I}$ 的所有沿前 2 个模态的矩阵切片均为单位矩阵。

定理 1 (d 阶 T-SVD^[19]) d 阶张量 $\mathcal{A}$ 为 $n_1 \times n_2 \times \dots \times n_d$ 大小的实张量, 那么它能够被分解为:

$$\mathcal{A} = \mathcal{U} \otimes_L \mathcal{S} \otimes_L \mathcal{V}^*, \quad (1)$$

其中: $\mathcal{U} \in \mathbf{R}^{n_1 \times n_1 \times n_3 \times \dots \times n_d}$ 和 $\mathcal{V} \in \mathbf{R}^{n_2 \times n_2 \times n_3 \times \dots \times n_d}$ 是正交的, $\mathcal{S} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times \dots \times n_d}$ 是 f -对角张量, 即 $\mathcal{S}$ 的所有沿前 2 个模态的矩阵切片是对角矩阵。

下面的算法 1 给出了关于式(1)计算 d 阶张量 $\mathcal{A}$ 的 T-SVD 算法。

算法 1 d 阶张量奇异值分解^[20]

输入: $\mathcal{A} \in \mathbf{R}^{n_1 \times \dots \times n_d}$ 以及线性变换 L 对应的变换矩阵 $\{\mathbf{U}_{n_i}\}_{i=3}^d$ 。

输出: $\mathcal{A}$ 的 T-SVD 成份: $\mathcal{U}, \mathcal{S}, \mathcal{V}$ 。

1. 对张量 $\mathcal{A}$ 进行线性变换: $L(\mathcal{A}) \rightarrow \mathcal{A}_L$;
2. 计算 $\mathcal{U}_L, \mathcal{S}_L$ 和 $\mathcal{V}_L$ 的各个矩阵切片:
 for $i_3 \in \{1, \dots, n_3\}, \dots, i_n \in \{1, \dots, n_d\}$ do
 $[U, S, V] = \text{svd}(\mathcal{A}_L(:, :, i_3, \dots, i_d))$,
 $\mathcal{U}_L(:, :, i_3, \dots, i_d) = U, \mathcal{S}_L(:, :, i_3, \dots, i_d) = S, \mathcal{V}_L(:, :, i_3, \dots, i_d) = V$;
 end for
3. 对 $\mathcal{U}_L, \mathcal{S}_L$ 和 $\mathcal{V}_L$ 分别进行逆线性变换: $L^{-1}(\mathcal{U}_L) \rightarrow \mathcal{U}, L^{-1}(\mathcal{S}_L) \rightarrow \mathcal{S}$ 和 $L^{-1}(\mathcal{V}_L) \rightarrow \mathcal{V}$.

2 模型与算法

针对观测到四阶不完全的互联网流量数据 $\mathcal{M} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$, 基于 d 阶 T-SVD 结合时空正则化, 构建下面的数据恢复模型:

$$\begin{aligned} \min_{\mathcal{X}, \mathcal{U}, \mathcal{S}, \mathcal{V}} & \frac{1}{2} \|\mathcal{X} - \mathcal{U} \otimes_L \mathcal{S} \otimes_L \mathcal{V}^*\|_F^2 + \frac{\alpha_1}{2} \|\mathbf{H} \cdot \text{Mat}_3(\mathcal{X})\|_F^2 + \frac{\alpha_2}{2} \|\mathbf{K} \cdot \text{Mat}_4(\mathcal{X})\|_F^2, \\ \text{s. t. } & P_\Omega(\mathcal{X}) = P_\Omega(\mathcal{M}), \mathcal{X} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}, \mathcal{U} \in \mathbf{R}^{n_1 \times n_1 \times n_3 \times n_4}, \\ & \mathcal{S} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}, \mathcal{V} \in \mathbf{R}^{n_2 \times n_2 \times n_3 \times n_4}. \end{aligned} \quad (2)$$

其中: α_1, α_2 均为大于 0 的正则化参数, Ω 为不完全张量 $\mathcal{M}$ 观察到元素的指示集, $P_\Omega(\cdot)$ 为线性算子, $\mathcal{X}$ 是需要恢复张量, $\mathcal{U}, \mathcal{V}$ 是正交的, $\mathcal{S}$ 是 f -对角张量, 并且:

$$\text{Mat}_3(\mathcal{X}) = [\mathcal{X}(1, 1, :, 1), \dots, \mathcal{X}(n_1, 1, :, 1), \dots, \mathcal{X}(1, n_2, :, n_4), \dots, \mathcal{X}(n_1, n_2, :, n_4)] \in \mathbf{R}^{n_3 \times n_1 n_2 n_4},$$

以及

$$\text{Mat}_4(\mathcal{X}) = [\mathcal{X}(1, 1, 1, :), \dots, \mathcal{X}(n_1, 1, 1, :), \dots, \mathcal{X}(1, n_2, n_3, :), \dots, \mathcal{X}(n_1, n_2, n_3, :)] \in \mathbf{R}^{n_4 \times n_1 n_2 n_3}.$$

为了更好地捕捉流量数据之间的时间稳定性特征, 即相邻时间段的数据相似, 选择约束矩阵 $\mathbf{H}$ 为 $(n_3 - 1) \times n_3$ 大小的 Toeplitz 矩阵, 即:

$$\mathbf{H} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}_{(n_3-1) \times n_3}.$$

同样地, 根据 $\text{Mat}_4(\mathcal{X})$ 每一列的含义, 选择约束矩阵 $\mathbf{K}$ 为 $(n_4 - 1) \times n_4$ 大小的 Toeplitz 矩阵来捕获数据的周期性。另外, 通过引入辅助变量 $\mathcal{Y}, \mathcal{Z} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$ 完全分离目标函数从而简化问题的求解。因此, 结合罚方法, 问题(2)松弛为:

$$\begin{aligned} \min_{\mathcal{X}, \mathcal{Y}, \mathcal{S}, \mathcal{Z}, \mathcal{U}, \mathcal{V}} & f(\mathcal{X}, \mathcal{U}, \mathcal{S}, \mathcal{V}, \mathcal{Y}, \mathcal{Z}) = \frac{1}{2} \|\mathcal{X} - \mathcal{U} \otimes_L \mathcal{S} \otimes_L \mathcal{V}^*\|_F^2 + \frac{\alpha_1}{2} \|\mathbf{H} \cdot \text{Mat}_3(\mathcal{Y})\|_F^2 + \\ & \frac{\alpha_2}{2} \|\mathbf{K} \cdot \text{Mat}_4(\mathcal{Z})\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X} - \mathcal{Z}\|_F^2, \\ \text{s. t. } & P_\Omega(\mathcal{X}) = P_\Omega(\mathcal{M}). \end{aligned} \quad (3)$$

其中: β_1, β_2 为罚参数。

采用交替极小化算法求解模型(3)。对于 $q = 0, 1, 2, \dots$, 令 $\Theta^q = (\mathcal{X}^q, \mathcal{U}^q, \mathcal{S}^q, \mathcal{V}^q, \mathcal{Y}^q, \mathcal{Z}^q)$ 为当前迭代点, 通过如下几个子问题的求解, 得到第 $q+1$ 个迭代点 Θ^{q+1} 。

1) $\mathcal{X}$ 子问题求解。求解 $\mathcal{X}^{q+1}$ 即求解优化问题(4):

$$\begin{aligned} \min_{\mathcal{X} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}} & \frac{1}{2} \|\mathcal{X} - \mathcal{U}^q \otimes_L \mathcal{S}^q \otimes_L \mathcal{V}^{q*}\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X} - \mathcal{Y}^q\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X} - \mathcal{Z}^q\|_F^2, \\ \text{s. t. } & P_\Omega(\mathcal{X}) = P_\Omega(\mathcal{M}). \end{aligned} \quad (4)$$

因此该问题的解为:

$$\mathcal{X}_{ijkl}^{q+1} = \begin{cases} \frac{1}{1 + \beta_1 + \beta_2} (\mathcal{U}^q \otimes_L \mathcal{S}^q \otimes_L \mathcal{V}^{q*} + \beta_1 \mathcal{Y}^q + \beta_2 \mathcal{Z}^q)_{ijkl}, & (i, j, k, l) \notin \Omega \\ \mathcal{M}_{ijkl}, & \text{其它。} \end{cases} \quad (5)$$

2) $\mathcal{U}, \mathcal{S}$ 和 $\mathcal{V}$ 子问题求解。求解 $\mathcal{U}^{q+1}, \mathcal{S}^{q+1}$ 和 $\mathcal{V}^{q+1}$ 即求解优化问题(6):

$$\min_{\substack{\mathcal{U} \in \mathbf{R}^{n_1 \times n_1 \times n_3 \times n_4}, \mathcal{S} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}, \mathcal{V} \in \mathbf{R}^{n_2 \times n_2 \times n_3 \times n_4}}} \frac{1}{2} \|\mathcal{X}^{q+1} - \mathcal{U} \otimes_L \mathcal{S} \otimes_L \mathcal{V}^*\|_F^2. \quad (6)$$

由于 $\mathcal{U}, \mathcal{V}$ 是正交的, $\mathcal{S}$ 是 f -对角张量, 由 d 阶张量 T-SVD 最佳低秩逼近^[19]可知, 通过算法 1 对 $\mathcal{X}^{q+1}$ 进行 d 阶 T-SVD 分解得到 $\mathcal{U}^{q+1}, \mathcal{S}^{q+1}$ 和 $\mathcal{V}^{q+1}$ 。

3) $\mathcal{Y}$ 子问题求解。求解 $\mathcal{Y}^{q+1}$ 即求解优化问题(7):

$$\min_{\mathcal{Y} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}} \frac{\alpha_1}{2} \|\mathbf{H} \cdot \text{Mat}_3(\mathcal{Y})\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X}^{q+1} - \mathcal{Y}\|_F^2, \quad (7)$$

等价于求解:

$$\min_{\text{Mat}_3(\mathcal{Y}) \in \mathbf{R}^{n_3 \times n_1 n_2 n_4}} \frac{\alpha_1}{2} \|\mathbf{H} \cdot \text{Mat}_3(\mathcal{Y})\|_F^2 + \frac{\beta_1}{2} \|\text{Mat}_3(\mathcal{X}^{q+1}) - \text{Mat}_3(\mathcal{Y})\|_F^2.$$

所以该问题的解为:

$$\text{Mat}_3(\mathcal{Y}^{q+1}) = (\alpha_1 \mathbf{H}^T \mathbf{H} + \beta_1 \mathbf{I})^{-1} (\beta_1 \text{Mat}_3(\mathcal{X}^{q+1})),$$

因此有:

$$\mathcal{Y}^{q+1} = \text{IMat}_3((\alpha_1 \mathbf{H}^T \mathbf{H} + \beta_1 \mathbf{I})^{-1} (\beta_1 \text{Mat}_3(\mathcal{X}^{q+1}))), \quad (8)$$

其中: $\text{IMat}_3(\cdot)$ 是 $\text{Mat}_3(\cdot)$ 的逆。

4) $\mathcal{Z}$ 子问题求解。求解 $\mathcal{Z}^{q+1}$ 即求解优化问题(9):

$$\min_{\mathcal{Z} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}} \frac{\alpha_2}{2} \|\mathbf{K} \cdot \text{Mat}_4(\mathcal{Z})\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X}^{q+1} - \mathcal{Z}\|_F^2. \quad (9)$$

与 $\mathcal{Y}$ 子问题类似, 式(9) 的最优解为:

$$\mathcal{Z}^{q+1} = \text{IMat}_4((\alpha_2 \mathbf{K}^T \mathbf{K} + \beta_2 \mathbf{I})^{-1} (\beta_2 \text{Mat}_4(\mathcal{X}^{q+1}))), \quad (10)$$

其中: $\text{IMat}_4(\cdot)$ 是 $\text{Mat}_4(\cdot)$ 的逆。

综上所述, 模型(2)的算法框架如下。

算法 2 基于 d 阶 T-SVD 和时空正则化的张量填充算法 (Tensor completion with d -order T-SVD and spatio-temporal regularization, SRHSVD)

输入: 参数 $\alpha_1, \alpha_2, \beta_1, \beta_2, \epsilon > 0$, 不完全张量 $\mathcal{M}$, 指示集 Ω 以及初始点 Θ^0 。

输出: 恢复的张量 $\mathcal{X} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$;

```

for  $q=0, 1, 2, \dots$  do
    通过式(5)更新  $\mathcal{X}^{q+1}$ ;
    通过算法 1 更新  $\mathcal{U}^{q+1}, \mathcal{S}^{q+1}$  以及  $\mathcal{V}^{q+1}$ ;
    通过式(8)更新  $\mathcal{Y}^{q+1}$ ;
    通过式(10)更新  $\mathcal{Z}^{q+1}$ ;
    If  $\|\mathcal{X}^{q+1} - \mathcal{X}^q\|_F / \|\mathcal{X}^q\|_F < \epsilon$ , then
        end if
    end for

```

3 收敛性分析

本节将对算法进行收敛性分析。分别从迭代产生的点列 $\{\Theta^q\}_{q=0}^\infty$ 的充分下降性、相对绝对误差以及连续性等方面说明算法的全局收敛性。定义 $\mathcal{A}^q = \mathcal{U}^q \otimes_L \mathcal{S}^q \otimes_L \mathcal{V}^{q*}$ 。

定理 2 定义 $\xi_1 = \alpha_1 \lambda_{\min}(\mathbf{H}^T \mathbf{H}) + \beta_1, \xi_2 = \alpha_2 \lambda_{\min}(\mathbf{K}^T \mathbf{K}) + \beta_2$ 。对于迭代产生的序列 $\{\Theta^q\}_{q=0}^\infty$, 有 $f(\Theta^{q+1}) + \xi \|\Theta^{q+1} - \Theta^q\|_F^2 \leq f(\Theta^q)$, 其中 $\xi = \frac{1}{2} \min\{1 + \beta_1 + \beta_2, \xi_1, \xi_2\}$ 。

证明 先证明 $\mathcal{X}$ 子问题的下降性。

对于任意整数 $q \geq 0$, 令 $f_1(\mathcal{X}) = \frac{1}{2} \|\mathcal{X} - \mathcal{A}^q\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X} - \mathcal{Y}^q\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X} - \mathcal{Z}^q\|_F^2$, 则对于任意的 $\mathcal{X}^q, \mathcal{X}^{q+1}$ 以及

$\lambda \in [0, 1]$, 有:

$$f_1(\lambda \mathcal{X}^q + (1-\lambda) \mathcal{X}^{q+1}) \leq \lambda f_1(\mathcal{X}^q) + (1-\lambda) f_1(\mathcal{X}^{q+1}) - \frac{1+\beta_1+\beta_2}{2} \lambda (1-\lambda) \|\mathcal{X}^{q+1} - \mathcal{X}^q\|_F^2$$

成立, 所以 f_1 是 $1+\beta_1+\beta_2$ -强凸的。因此对于任意的 $\mathcal{X}^q$ 和 $\mathcal{X}^{q+1}$, 有:

$$f_1(\mathcal{X}^{q+1}) + \langle \nabla f_1(\mathcal{X}^{q+1}), \mathcal{X}^{q+1} - \mathcal{X}^q \rangle + \frac{1+\beta_1+\beta_2}{2} \|\mathcal{X}^{q+1} - \mathcal{X}^q\|_F^2 \leq f_1(\mathcal{X}^q).$$

当 $\emptyset \subsetneq \Omega \subsetneq [n_1] \times [n_2] \times [n_3] \times [n_4]$ 时, 问题(4)等价于:

$$\min_{\mathcal{X}_{\Omega^c}} f_{1\Omega^c}(\mathcal{X}_{\Omega^c}) = \frac{1}{2} \|\mathcal{X}_{\Omega^c} - \mathcal{A}_{\Omega^c}^q\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X}_{\Omega^c} - \mathcal{Y}_{\Omega^c}^q\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X}_{\Omega^c} - \mathcal{Z}_{\Omega^c}^q\|_F^2,$$

所以 $f_{1\Omega^c}$ 是关于 $\mathcal{X}_{\Omega^c}$ 的 $1+\beta_1+\beta_2$ -强凸函数。由于 $\mathcal{X}^{q+1}$ 是问题(4)的解, 所以 $\mathcal{X}_{\Omega^c}^{q+1}$ 是 $f_{1\Omega^c}$ 的最优点, 因此 $\nabla f_{1\Omega^c}(\mathcal{X}_{\Omega^c}^{q+1}) = \mathcal{O}$ 。所以有:

$$f_{1\Omega^c}(\mathcal{X}_{\Omega^c}^{q+1}) + \frac{1+\beta_1+\beta_2}{2} \|\mathcal{X}_{\Omega^c}^{q+1} - \mathcal{X}_{\Omega^c}^q\|_F^2 \leq f_{1\Omega^c}(\mathcal{X}_{\Omega^c}^q). \quad (11)$$

由于 $\mathcal{X}_{\Omega^c}^{q+1} = \mathcal{X}_{\Omega}^q = \mathcal{M}_{\Omega}$, 所以结合式(11)可以得到:

$$f(\mathcal{X}^{q+1}, \mathcal{U}^q, \mathcal{S}^q, \mathcal{V}^q, \mathcal{Y}^q, \mathcal{Z}^q) + \frac{1+\beta_1+\beta_2}{2} \|\mathcal{X}^{q+1} - \mathcal{X}^q\|_F^2 \leq f(\mathcal{X}^q, \mathcal{U}^q, \mathcal{S}^q, \mathcal{V}^q, \mathcal{Y}^q, \mathcal{Z}^q). \quad (12)$$

另外, 当 $\Omega = \emptyset$ 时, 问题(4)为无约束优化问题, 式(12)成立。当 $\Omega = [n_1] \times [n_2] \times [n_3] \times [n_4]$ 时, 因为 $\mathcal{X}^{q+1} = \mathcal{X}^q = \mathcal{M}$, 因此式(12)仍然成立。

与 $\mathcal{X}$ 子问题相似, 可以得到对于 $\mathcal{U}, \mathcal{S}$ 和 $\mathcal{V}$ 子问题有:

$$f(\mathcal{X}^{q+1}, \mathcal{U}^{q+1}, \mathcal{S}^{q+1}, \mathcal{V}^{q+1}, \mathcal{Y}^q, \mathcal{Z}^q) + \frac{1}{2} \|\mathcal{A}^{q+1} - \mathcal{A}^q\|_F^2 \leq f(\mathcal{X}^{q+1}, \mathcal{U}^q, \mathcal{S}^q, \mathcal{V}^q, \mathcal{Y}^q, \mathcal{Z}^q), \quad (13)$$

以及对于 $\mathcal{Y}$ 和 $\mathcal{Z}$ 子问题分别有:

$$f(\mathcal{X}^{q+1}, \mathcal{U}^{q+1}, \mathcal{S}^{q+1}, \mathcal{V}^{q+1}, \mathcal{Y}^{q+1}, \mathcal{Z}^q) + \frac{\alpha_1 \lambda_{\min}(\mathbf{H}^T \mathbf{H}) + \beta_1}{2} \|\mathcal{Y}^{q+1} - \mathcal{Y}^q\|_F^2 \leq f(\mathcal{X}^{q+1}, \mathcal{U}^{q+1}, \mathcal{S}^{q+1}, \mathcal{V}^{q+1}, \mathcal{Y}^q, \mathcal{Z}^q) \quad (14)$$

和

$$f(\mathcal{X}^{q+1}, \mathcal{U}^{q+1}, \mathcal{S}^{q+1}, \mathcal{V}^{q+1}, \mathcal{Y}^{q+1}, \mathcal{Z}^{q+1}) + \frac{\alpha_2 \lambda_{\min}(\mathbf{K}^T \mathbf{K}) + \beta_2}{2} \|\mathcal{Z}^{q+1} - \mathcal{Z}^q\|_F^2 \leq f(\mathcal{X}^{q+1}, \mathcal{U}^{q+1}, \mathcal{S}^{q+1}, \mathcal{V}^{q+1}, \mathcal{Y}^{q+1}, \mathcal{Z}^q), \quad (15)$$

其中: $\lambda_{\min}(\cdot)$ 指最小特征值。所以对于给定的 ξ_1 与 ξ_2 , 当 $\xi = \frac{1}{2} \min\{1+\beta_1+\beta_2, \xi_1, \xi_2\}$ 时, 结合式(12)~(15)有:

$$\begin{aligned} f(\Theta^q) - f(\Theta^{q+1}) &\geq \\ \xi(\|\mathcal{X}^{q+1} - \mathcal{X}^q\|_F^2 + \|\mathcal{A}^{q+1} - \mathcal{A}^q\|_F^2 + \|\mathcal{Y}^{q+1} - \mathcal{Y}^q\|_F^2 + \|\mathcal{Z}^{q+1} - \mathcal{Z}^q\|_F^2) &\geq \xi \|\Theta^{q+1} - \Theta^q\|_F^2. \end{aligned}$$

因此结论成立。

证毕

定理 2 说明了算法产生的迭代序列在目标函数值上是单调下降的, 确保了充分下降性。下面定理 3 将说明算法迭代产生序列的相对误差性。

定理 3 设 $\{\Theta^q\}_{q=0}^{\infty}$ 为迭代产生的序列, 那么对于 $\nabla f(\Theta^{q+1})$ 有 $\|\nabla f(\Theta^{q+1})\|_F^2 \leq \gamma \|\Theta^{q+1} - \Theta^q\|_F^2$, 其中 $\gamma = \sqrt{3} \max\{1, \beta_1, \beta_2\}$ 。

证明 先考虑 $\mathcal{X}$ 子问题的相对误差。

对于任意的 $k \geq 0$, 定义 $g_k(\mathcal{X}) = \frac{1}{2} \|\mathcal{X} - \mathcal{A}^k\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X} - \mathcal{Y}^k\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X} - \mathcal{Z}^k\|_F^2$ 。对于任意的 $\mathcal{X} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$, 有 $P_{\Omega}(\mathcal{X}) = P_{\Omega}(\mathcal{M})$, 因此可以得到:

$$g_k(\mathcal{X}) = \frac{1}{2} \|\mathcal{X}_{\Omega^c} - \mathcal{A}_{\Omega^c}^k\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X}_{\Omega^c} - \mathcal{Y}_{\Omega^c}^k\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X}_{\Omega^c} - \mathcal{Z}_{\Omega^c}^k\|_F^2 + h_k,$$

其中: $h_k = \frac{1}{2} \|\mathcal{X}_{\Omega} - \mathcal{A}_{\Omega}^k\|_F^2 + \frac{\beta_1}{2} \|\mathcal{X}_{\Omega} - \mathcal{Y}_{\Omega}^k\|_F^2 + \frac{\beta_2}{2} \|\mathcal{X}_{\Omega} - \mathcal{Z}_{\Omega}^k\|_F^2$ 。所以, 对于任意 $k \geq 0$, 有:

$$\nabla_{\mathcal{X}_{\Omega^c}} g_k(\mathcal{X}) = (1 + \beta_1 + \beta_2) \mathcal{X}_{\Omega^c} - \mathcal{A}_{\Omega^c}^k - \beta_1 \mathcal{Y}_{\Omega^c}^k - \beta_2 \mathcal{Z}_{\Omega^c}^k.$$

又 $\mathcal{X}^{q+1}$ 是问题(4)的最优解,即 $\nabla_{\mathcal{X}_{\Omega^c}} g_q(\mathcal{X}^{q+1}) = \mathcal{O}$, 且 $\nabla_{\mathcal{X}_{\Omega^c}} f(\Theta^{q+1}) = \nabla_{\mathcal{X}_{\Omega^c}} g_{q+1}(\mathcal{X}^{q+1})$, 因此有:

$$\begin{aligned} \|\nabla_{\mathcal{X}_{\Omega^c}} f(\Theta^{q+1})\|_F^2 &= \|\nabla_{\mathcal{X}_{\Omega^c}} g_{q+1}(\mathcal{X}^{q+1}) - \nabla_{\mathcal{X}_{\Omega^c}} g_q(\mathcal{X}^{q+1})\|_F^2 = \\ &= \|\mathcal{A}_{\Omega^c}^{q+1} - \mathcal{A}_{\Omega^c}^q + \beta_1(\mathcal{Y}_{\Omega^c}^{q+1} - \mathcal{Y}_{\Omega^c}^q) + \beta_2(\mathcal{Z}_{\Omega^c}^{q+1} - \mathcal{Z}_{\Omega^c}^q)\|_F^2 \leq \\ &= 3(\|\mathcal{A}_{\Omega^c}^{q+1} - \mathcal{A}_{\Omega^c}^q\|_F^2 + \beta_1\|\mathcal{Y}_{\Omega^c}^{q+1} - \mathcal{Y}_{\Omega^c}^q\|_F^2 + \beta_2\|\mathcal{Z}_{\Omega^c}^{q+1} - \mathcal{Z}_{\Omega^c}^q\|_F^2) \leq \gamma^2 \|\Theta^{q+1} - \Theta^q\|_F^2, \end{aligned}$$

其中: $\gamma = \sqrt{3} \max\{1, \beta_1, \beta_2\}$. 故 $\|\nabla_{\mathcal{X}_{\Omega^c}} f(\Theta^{q+1})\|_F \leq \gamma \|\Theta^{q+1} - \Theta^q\|_F$.

对于 $\mathcal{U}, \mathcal{S}$ 和 $\mathcal{V}$ 子问题, 定义函数 $t(\mathcal{X}, \mathcal{A}) = \frac{1}{2} \|\mathcal{X} - \mathcal{A}\|_F^2$, 则有:

$$\|\nabla_{\mathcal{A}} f(\Theta^{q+1})\|_F^2 = \|\nabla_{\mathcal{A}} t(\mathcal{X}^{q+1}, \mathcal{A}^{q+1})\|_F^2.$$

由于 $\mathcal{A}^{q+1}$ 是问题(6)的最优解, 因此 $\nabla_{\mathcal{A}} t(\mathcal{X}^{q+1}, \mathcal{A}^{q+1}) = \mathcal{A}^{q+1} - \mathcal{X}^{q+1} = \mathcal{O}$. 所以 $\|\nabla_{\mathcal{A}} f(\Theta^{q+1})\|_F = \mathcal{O}$.

同理可得 $\|\nabla_{\mathcal{Y}} f(\Theta^{q+1})\|_F = \mathcal{O}$ 和 $\|\nabla_{\mathcal{Z}} f(\Theta^{q+1})\|_F = \mathcal{O}$.

综上所述, 对于 $\nabla f(\Theta^{q+1})$ 有:

$$\begin{aligned} \|\nabla f(\Theta^{q+1})\|_F^2 &= \|\nabla_{\mathcal{X}_{\Omega^c}} f(\Theta^{q+1})\|_F^2 + \|\nabla_{\mathcal{A}} f(\Theta^{q+1})\|_F^2 + \|\nabla_{\mathcal{Y}} f(\Theta^{q+1})\|_F^2 + \\ &= \|\nabla_{\mathcal{Z}} f(\Theta^{q+1})\|_F^2 \leq \gamma^2 \|\Theta^{q+1} - \Theta^q\|_F^2, \end{aligned}$$

其中: $\gamma = \sqrt{3} \max\{1, \beta_1, \beta_2\}$. 故结论得证. 证毕

定理 4 对于迭代产生的序列 $\{\Theta^q\}_{q=0}^\infty$, 存在一个子列 $\{\Theta^{q_j}\}_{j=0}^\infty$ 以及点 $\hat{\Theta}$, 使得 $\lim_{j \rightarrow \infty} \Theta^{q_j} = \hat{\Theta}$ 且 $\lim_{j \rightarrow \infty} f(\Theta^{q_j}) = f(\hat{\Theta})$.

证明 定义 $\mathbf{E} = \mathbf{E}_1 \times \mathbf{E}_2 \times \mathbf{E}_3 \times \mathbf{E}_4 \times \mathbf{E}_5 \times \mathbf{E}_6$, 其中 $\mathbf{E}_1 = \mathbf{E}_3 = \mathbf{E}_5 = \mathbf{E}_6 = \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$, $\mathbf{E}_2 = \mathbf{R}^{n_1 \times n_1 \times n_3 \times n_4}$, $\mathbf{E}_4 = \mathbf{R}^{n_2 \times n_2 \times n_3 \times n_4}$. 由于 $\{\Theta^q\}_{q=0}^\infty \subset \mathbf{E}$ 有界且相对紧, 所以存在子列 $\{\Theta^{q_j}\}_{j=0}^\infty$ 以及 $\hat{\Theta} \in \mathbf{E}$ 使得 $\lim_{j \rightarrow \infty} \Theta^{q_j} = \hat{\Theta}$.

另外, 由于 $\{\mathcal{X}^q\}_{q=0}^\infty \subset \overline{\mathbf{E}}_1$, 其中 $\overline{\mathbf{E}}_1 = \{\mathcal{X} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4} \mid \mathcal{X}_{ijkl} = \mathcal{M}_{ijkl}, (i, j, k, l) \in \Omega\}$, 所以 $\{\Theta^q\}_{q=0}^\infty \subset \overline{\mathbf{E}} = \overline{\mathbf{E}}_1 \times \mathbf{E}_2 \times \mathbf{E}_3 \times \mathbf{E}_4 \times \mathbf{E}_5 \times \mathbf{E}_6$. 由于 $\overline{\mathbf{E}}$ 为闭集, 因此 $\hat{\Theta} \in \overline{\mathbf{E}}$. 又由于 f 在 $\overline{\mathbf{E}}$ 上是连续的, 所以 $\lim_{j \rightarrow \infty} f(\Theta^{q_j}) = f(\hat{\Theta})$. 证毕

定理 4 说明了算法产生迭代序列的有界性以及连续性. 由文献[21]中 KL 函数的定义可知, $f(\Theta^q)$ 是 KL 函数. 另外, 由于定理 2~定理 4 满足文献[21]中的 H1、H2 以及 H3 条件, 所以根据文献[21]的定理 2.3 可知, 点列 $\{\Theta^q\}_{q=0}^\infty$ 收敛于 f 的临界点.

下面给出算法 2 的计算时间复杂度.

在每次迭代过程中, $\mathcal{X}$ 子问题计算时间复杂度为 $O(2n_1 n_2 n_3 n_4 (n_1 + n_2))$, $\mathcal{U}, \mathcal{S}$ 和 $\mathcal{V}$ 子问题的计算时间复杂度为 $O(2n_2^2 n_3 n_4 (7n_1 + 4n_2))$, $\mathcal{Y}$ 和 $\mathcal{Z}$ 子问题的计算时间复杂度分别为 $O(2n_1 n_2 n_3^2 n_4)$ 和 $O(2n_1 n_2 n_3 n_4^2)$. 因此总的计算时间复杂度为 $O(2n_2 n_3 n_4 (n_1 (n_1 + 8n_2 + n_3 + n_4) + 4n_2^2))$.

4 数值实验

在本节中, 将基于 d 阶 T-SVD 和时空正则化策略的互联网流量数据恢复方法(SRHSVD)应用在 2 个真实的互联网流量数据, 分别为 GÉANT 数据集^[22]和 Abilene 数据集(<http://abilene.internet2.edu/observatory/data-collections.html>)上. 其中 GÉANT 数据集是每隔 15 min 记录一次 23 个出发节点到 23 个接收节点的流量数据, Abilene 数据集是每隔 5 min 记录一次 11 个出发节点到 11 个接收节点的流量数据. 对于 2 个数据集均选 7 d 的数据, 因此 2 个数据集按照四阶张量形式排列的大小分别为 $23 \times 23 \times 96 \times 7$ 和 $11 \times 11 \times 288 \times 7$. 值得注意的是, 所有的实验结果均是随机进行 5 次实验后取平均值, 所有数值实验均在 64 位 Windows 操作系统的笔记本电脑(32 核 13th Gen Intel(R) Core(TM) i9-13900HX CPU @2.20 GHz)上进行.

在数值实验中, 选择归一化标准平均绝对误差(σ_{NMAE})来评估方法对数据的恢复效果, 即:

$$\sigma_{\text{NMAE}} := \frac{\sum_{(i,j,k,l) \notin \Omega} |(\mathcal{X}_{\text{true}})_{ijkl} - (\mathcal{X}_{\text{rec}})_{ijkl}|}{\sum_{(i,j,k,l) \notin \Omega} |(\mathcal{X}_{\text{true}})_{ijkl}|},$$

其中: X_{true} 和 X_{rec} 分别代表真实数据和恢复数据。显然, σ_{NMAE} 值越低, 说明算法对于数据的恢复效果越好, 反之, 说明效果越差。在参数设置中, 通过不断调整参数的大小以达到最大的恢复效果, 值为 $\alpha_1=1, \alpha_2=0.005, \beta_1=1, \beta_2=1, \epsilon=1e-4$, 其中 α_2 对于实验结果的影响较大。同时可逆变换矩阵 $\mathbf{L}$ 为离散 Fourier 变换矩阵。选取 5 个互联网流量数据恢复的方法与 SRHSVD 方法进行比较, 其中第 1 个是矩阵填充方法, 其余 4 个是张量填充方法:

- 1) SRMF: 基于矩阵分解, 并且结合时空正则项的恢复方法^[14];
- 2) CPWOPT: 基于张量 CP 分解的张量填充方法^[15];
- 3) TCTF2R: 基于张量 T-积分分解, 并结合时空正则项的交通流量恢复方法^[18];
- 4) SRTNN: 基于张量核范数, 并结合时空正则项的交通流量恢复方法^[23];
- 5) BGCP: 基于贝叶斯高斯 CP 分解的数据恢复方法^[24]。

首先考虑数据随机缺失时的恢复效果。图 1 展示了在随机采样率为 10%~90% 下, 各种方法在 2 个数据集上恢复的 σ_{NMAE} 值。由图 1 可以看出, 在样本率只有 10% 时, 在 GÉANT 数据集中, TCTF2R 方法以及 SRTNN 方法的恢复效果比 SRHSVD 方法更好一点; Abilene 数据集中, SRHSVD 方法的恢复效果不及 TCTF2R, 究其原因, 可能是在缺失率非常高的情况下, SRHSVD 方法未能充分融合不完全数据的低秩性, 这也是下一步值得研究的一个课题。但是当样本率高于 20% 时, 无论在哪个数据集, SRHSVD 方法对于缺失数据的恢复效果更佳。因此, 当样本率高于 20% 时, SRHSVD 方法对数据恢复的效果具有一定的竞争优势。

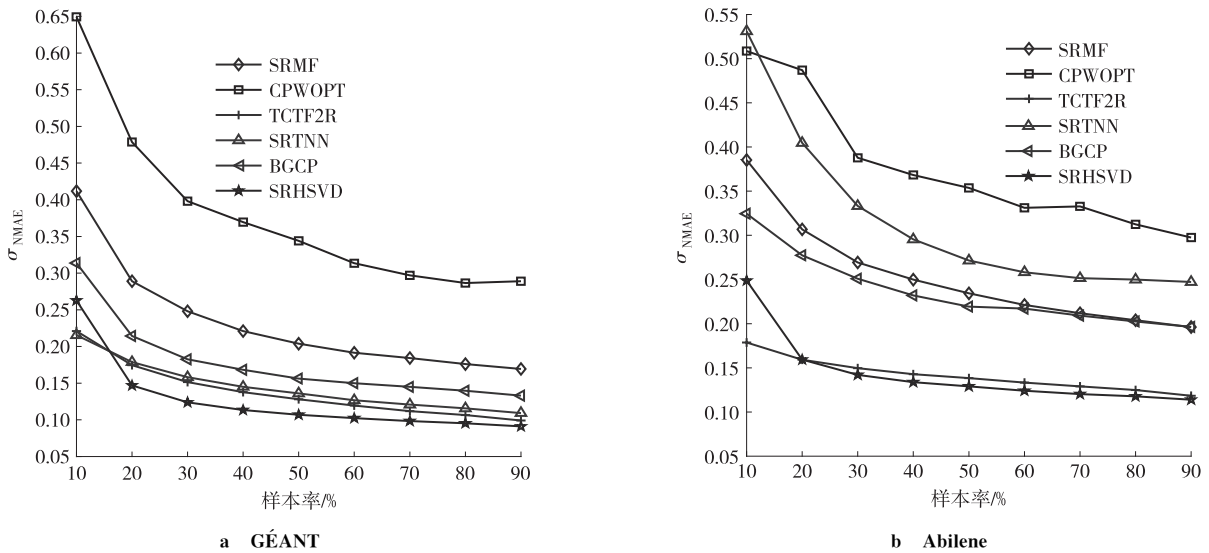


图 1 不同随机样本率下各种算法对数据恢复的 σ_{NMAE} 值

Fig. 1 σ_{NMAE} for recovering datasets with respect to diverse random sample ratios

另一方面, 为了体现 SRHSVD 方法充分利用数据内部的结构特性, 接下来考虑 2 种结构性缺失^[5,23]:

1) xxTimeRandLoss: 该情景模拟了在某些特定因素(如设备监控过载)作用下, 数据于特定时间的结构化缺失。这种情况下, 特定时间点的数据按照一定随机比例缺失。即随机选取 $xx\%$ 的互联网流量张量 $\mathcal{M} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$ 所有沿前 2 个模态的矩阵切片, 然后对所选每个矩阵随机选取一定比例(q)的元素进行删除, 得到不完整的张量 $\mathcal{M}$ 。

2) xxElemRandLoss: 该情景模拟了在某些特定因素(如使用不可靠的数据传输设备)作用下, 数据于特定发出与接收节点的结构化缺失。这种情况下, 特定发出与接收节点的数据按照一定随机比例缺失。即随机选取 $xx\%$ 的互联网流量张量 $\mathcal{M} \in \mathbf{R}^{n_1 \times n_2 \times n_3 \times n_4}$ 所有沿模-3 和模-4 的矩阵切片, 然后对所选每个矩阵随机选取一定比例(q)的元素进行删除, 得到不完整的张量 $\mathcal{M}$ 。

对于 2 种结构性缺失, 均考虑下列 9 种缺失情况:

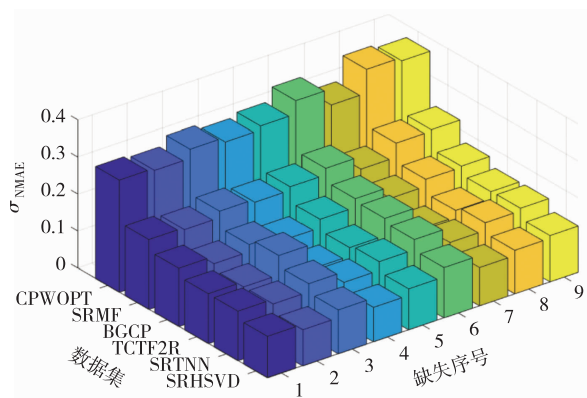
- 1) Missing Id 1-3: 令 $xx=25$, 然后分别对应令 $q=30\%$ 、 50% 以及 70% 。
- 2) Missing Id 4-6: 令 $xx=50$, 然后分别对应令 $q=30\%$ 、 50% 以及 70% 。
- 3) Missing Id 7-9: 令 $xx=75$, 然后分别对应令 $q=30\%$ 、 50% 以及 70% 。

表 1 列出了在 xxTimeRandLoss 结构性缺失下的各种算法对 GÉANT 和 Abilene 数据集恢复的 σ_{NMAE} 值, 图 2 展示了在 xxElemRandLoss 结构性缺失下的各种算法对 GÉANT 和 Abilene 数据集恢复的 σ_{NMAE} 值。通过表 1 以及图 2, 观察到在 2 种结构性缺失下, SRHSVD 方法在数据恢复的精度上展现出明显的优势, 说明在结构性缺失情况下, SRHSVD 方法仍能更好地恢复数据。图 3 通过对数据恢复效果的可视化, 直观地观察到 SRHSVD 方法对数据结构性缺失的恢复具有独特的优势。

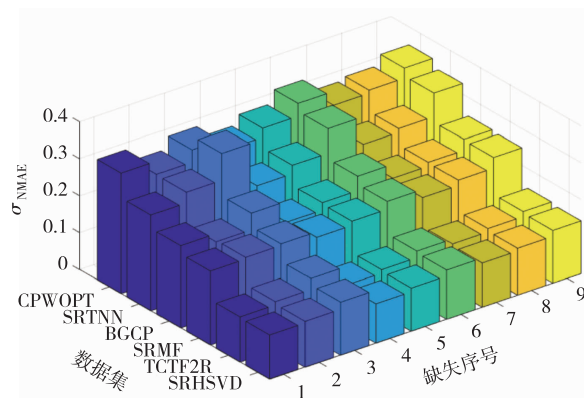
表 1 xxTimeRandLoss 结构性缺失下对 GÉANT 和 Abilene 数据集恢复的 σ_{NMAE} 值

Tab. 1 σ_{NMAE} for recovering GÉANT and Abilene datasets with respect to diverse Missing Id in xxTimeRandLoss

缺失序号	SRMF		CPWOPT		TCTF2R		SRTNN		BGCP		SRHSVD	
	GÉANT	Abilene	GÉANT	Abilene	GÉANT	Abilene	GÉANT	Abilene	GÉANT	Abilene	GÉANT	Abilene
1	0.178	0.202	0.295	0.298	0.099	0.118	0.109	0.246	0.137	0.192	0.091	0.115
2	0.192	0.213	0.316	0.308	0.103	0.121	0.112	0.248	0.140	0.193	0.094	0.116
3	0.223	0.237	0.343	0.335	0.104	0.124	0.116	0.256	0.142	0.201	0.094	0.118
4	0.180	0.203	0.303	0.313	0.106	0.123	0.114	0.246	0.139	0.201	0.095	0.117
5	0.194	0.216	0.320	0.316	0.110	0.126	0.119	0.253	0.145	0.201	0.098	0.118
6	0.221	0.237	0.360	0.352	0.116	0.131	0.124	0.262	0.150	0.208	0.100	0.122
7	0.182	0.208	0.301	0.307	0.108	0.127	0.117	0.252	0.143	0.202	0.096	0.119
8	0.197	0.218	0.316	0.332	0.118	0.132	0.125	0.257	0.149	0.212	0.101	0.122
9	0.226	0.244	0.369	0.366	0.131	0.139	0.138	0.273	0.162	0.225	0.109	0.130



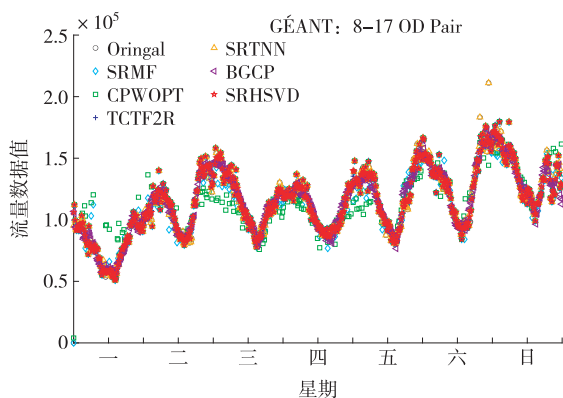
a GÉANT



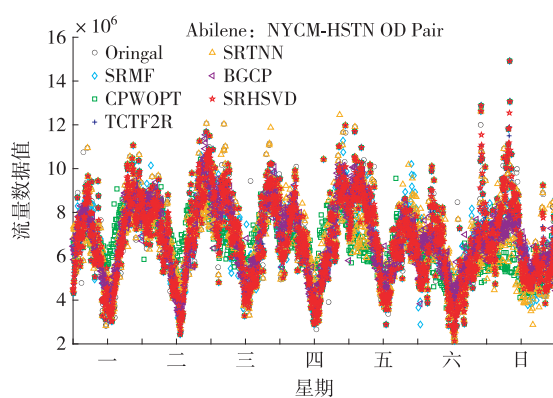
b Abilene

图 2 xxElemRandLoss 结构性缺失下不同数据集恢复的 σ_{NMAE} 值

Fig. 2 σ_{NMAE} for recovering GÉANT and Abilene datasets with respect to diverse Missing Id in xxElemRandLoss



a xxTimeRandLoss



b xxElemRandLoss

图 3 不同情形下 Missing Id 5 缺失下对数据恢复的可视化

Fig. 3 Visualization of data recovery under Missing Id 5 in different situations

5 总结

本文提出了一种基于 d 阶张量奇异值分解的互联网流量数据恢复模型。通过对四阶张量进行可逆变换充分融合数据之间的内部结构特性,同时加入时空正则项刻画数据的相似性和周期性。在求解模型中通过引入辅助变量,采用交替极小化方法进行求解。数值实验证明了本文所提出的 SRHSVD 方法在互联网流量数据恢复效果上具有明显优势,尤其是在结构性缺失情况下,SRHSVD 方法对于互联网流量数据的恢复效果更加明显。另外,由于 d 阶张量奇异值分解中的 f -对角张量能够表达出张量秩的特性,所以结合低秩性对模型改进从而提升数据恢复效果值得进一步研究。

参考文献:

- [1] ROUGHAN M, THORUP M, ZHANG Y. Traffic engineering with estimated traffic matrices [C]//CROVELLA M. Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement, New York: Association for Computing Machinery, 2003: 248-258.
- [2] CAHN R S. Wide area network design: concepts and tools for optimization [M]. San Francisco: Morgan Kaufmann Publishers Inc, 1998.
- [3] ZHANG Y, GE Z H, GREENBERG A, et al. Network anomography [C]//Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, October 19-21, 2005, Berkeley CA USA. Berkeley: USENIX Association, 2005: 30-30.
- [4] CANDÈS E J, RECHT B. Exact matrix completion via convex optimization [J]. Foundations of Computational Mathematics, 2009, 9: 717-772.
- [5] CARROLL J D, CHANG J J. Analysis of individual differences in multidimensional scaling via an N -way generalization of "Eckart-Young" decomposition [J]. Psychometrika, 1970, 35(3): 283-319.
- [6] HARSHMAN R A. Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis [R]. Michigan: UCLA Working Papers in Phonetics, 1970, 16: 84.
- [7] TUCKER L R. Implications of factor analysis of three-way matrices for measurement of change [J]. Problems in Measuring Change, 1963, 15(3): 122-137.
- [8] TUCKER L R. The extension of factor analysis to three-dimensional matrices [M]//NORMAN F. Contributions to Mathematical Psychology. New York: Holt, Rinehart and Winston, 1964: 110-119.
- [9] OSELEDETS I V. Tensor-train decomposition [J]. SIAM Journal on Scientific Computing, 2011, 33(5): 2295-2317.
- [10] KILMER M E, MARTIN C D. Factorization strategies for third-order tensors [J]. Linear Algebra and its Applications, 2011, 435(3): 641-658.
- [11] KERNFELD E, KILMER M, AERON S. Tensor-tensor products with invertible linear transforms [J]. Linear Algebra and Its Applications, 2015, 485: 545-570.
- [12] XU W H, ZHAO X L, NG M. A fast algorithm for cosine transform based tensor singular value decomposition [EB/OL]. (2019-02-08) [2024-05-21]. <http://arxiv.org/pdf/1902.03070>.
- [13] SONG G J, NG M K, ZHANG X J. Robust tensor completion using transformed tensor singular value decomposition [J]. Numerical Linear Algebra with Applications, 2020, 27(3): e2299.
- [14] ZHANG Y, ROUGHAN M, WILLINGER W, et al. Spatio-temporal compressive sensing and internet traffic matrices [J]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 267-278.
- [15] ACAR E, DUNLAVY D M, KOLDA T G, et al. Scalable tensor factorizations for incomplete data [J]. Chemometrics and Intelligent Laboratory Systems, 2011, 106(1): 41-56.
- [16] ZHOU H B, ZHANG D F, XIE K, et al. Spatio-temporal tensor completion for imputing missing internet traffic data [C]//2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC), December 14-16, Nanjing China, Piscataway: IEEE, 2015: 1-7.
- [17] ZHANG Z Y, LING C, HE H J, et al. A tensor train approach for internet traffic data completion [J]. Annals of Operations Research, 2021, 339: 1461-1479.
- [18] LING C, YU G H, QI L, et al. T -product factorization method for internet traffic data completion with spatio-temporal regularization [J]. Computational Optimization and Applications, 2021, 80: 883-913.
- [19] QIN W J, WANG H L, ZHANG F, et al. Low-rank high-order tensor completion with applications in visual data [J]. IEEE

- Transactions on Image Processing, 2022, 31: 2433-2448.
- [20] 覃文金. 基于奇异值分解的低秩高阶张量恢复算法与应用研究[D]. 重庆: 西南大学, 2022.
- QIN W J. Research on algorithm and application for low-rank high-order tensor recovery based on singular value decomposition [D]. Chongqing: Southwest University, 2022.
- [21] ATTOUCH H, BOLTE J, SVAITER B F. Convergence of descent methods for semi-algebraic and tame problems; proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods[J]. Mathematical Programming, 2013, 137(1): 91-129.
- [22] UHLIG S, QUOTIN B, LEPROPRE J, et al. Providing public intradomain traffic matrices to the research community[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(1): 83-86.
- [23] LI C, CHEN Y N, LI D H. Internet traffic tensor completion with tensor nuclear norm[J]. Computational Optimization and Applications, 2024, 87(3): 1033-1057.
- [24] 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.

Operations Research and Cybernetics

An Internet Traffic Data Completion Strategy with d th-Order Tensor Singular Value Decomposition

DUAN Yuxuan, LIU Jinjie

(National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 401331, China)

Abstract: The aim is to accurately recover Internet traffic data, which could reduce the negative impact caused by incomplete data on the network tasks. Since the traffic data could be represented by a fourth order tensor, and considering its spatiotemporal characteristics, an optimal recovery model was proposed based on d th-order tensor singular value decomposition (d th-order T-SVD) combined with spatiotemporal regularization strategies. The key feature of this model lies in its ability to deeply explore the data while preserving its internal complex structural properties, thereby achieving higher-quality data recovery. An efficient and stable algorithm is developed to solve this model accurately by utilizing the alternating minimization method. Finally, the proposed method was comprehensively validated by simulating both random and structural data loss scenarios on two real internet traffic datasets. Experimental results demonstrate that this method exhibits significant advantages in data recovery performance compared to existing methods.

Keywords: internet traffic data recovery; d th-order tensor singular value decomposition; spatio-temporal regularization; alternating minimization; tensor completion

(责任编辑 黄 颖)