计算机科学
COMPUTER SCIENCE

基于多维稀疏表示的空气质量指数数据补全

蔡启铨，卢举鸿，於志勇，黄昉菀

引用本文

蔡启铨，卢举鸿，於志勇，黄昉菀.基于多维稀疏表示的空气质量指数数据补全[J]．计算机科学, 2023, 50(8): 52-57.

CAI Qiquan, LU Juhong, YU Zhiyong, HUANG Fangwan. Data Completion of Air Quality Index Based on Multi-dimensional Sparse Representation [J]. Computer Science, 2023, 50(8): 52-57.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500277

基于多维稀疏表示的空气质量指数数据补全

蔡启铨[1] 卢举鸿[1] 於志勇[1,2] 黄昉菀[1,2]
1 福州大学计算机与大数据学院 福州 350108
2 福建省网络计算与智能信息处理重点实验室(福州大学) 福州 350108
(2292639773@qq.com)

摘 要 近年来,日益严重的空气污染正成为影响人们身体健康的危险因素之一。空气质量指数数据可以为政府提供大气环境变化的规律,也可以用于对大气污染的控制和管理。但该数据在采集的过程中不可避免地存在缺失,导致了对其进行数据挖掘的难度升高。为了更加充分地利用已经搜集到的数据,对缺失数据进行补全是非常必要的。然而,现有的补全方法往往在高缺失率情况下表现不佳。基于此提出将缺失矩阵补全问题转换为稀疏矩阵重构问题,并设计了一种基于多维稀疏表示的数据补全方法。该方法首先利用训练数据模拟各种随机缺失情况并用于过完备字典的学习,然后利用学习后字典的上半部分获得具有缺失值的矩阵的稀疏表示,最后将该稀疏表示与字典的下半部分相结合得到重构后的估计矩阵。实验结果表明,所提方法在多维时序空气质量指数数据补全问题上优于传统的矩阵补全方法,尤其是在数据缺失比较严重的情况下具有明显的优势。

关键词:空气质量指数;缺失数据;矩阵补全;字典学习;多维稀疏表示
中图法分类号 TP391

Data Completion of Air Quality Index Based on Multi-dimensional Sparse Representation

CAI Qiquan[1],LU Juhong[1],YU Zhiyong[1,2] and HUANG Fangwan[1,2]
1 College of Computer and Data Science,Fuzhou University,Fuzhou 350108,China
2 Fujian Key Laboratory of Network Computing and Intelligent Information Processing(Fuzhou University),Fuzhou 350108,China

Abstract In recent years,air pollution has become increasingly serious and become one of the risk factors affecting people's health. The air quality index(AQI) can provide the government with the laws of atmospheric environment changes,and can also be used for air pollution control. However,the data is inevitably missing in the process of collection,which leads to the difficulty of data mining. However,given the poor performance of existing completion methods under a high miss rate,this paper transforms the missing-matrix-completion problem into a sparse-matrix-reconstruction problem and designs a data completion method based on multi-dimensional sparse representation. The method first uses the training data to simulate various random missing cases for over-complete dictionary learning. Then,the sparse representation of the matrix with missing values is obtained by using the upper part of the learned dictionary. Finally,the sparse representation is combined with the lower part of the dictionary to obtain the reconstructed estimation matrix. Experimental results show that the proposed algorithm is superior to the traditional matrix method in the completion of multi-dimensional time series of AQI,especially in the case of serious missing.

Keywords Air quality index,Missing data,Matrix completion,Dictionary learning,Multi-dimensional sparse representation

1 引言

大气污染已经严重影响到社会的可持续发展,对人们的生命健康造成威胁[1]。世界卫生组织2019年发表的报告表示空气污染已经成为最大的健康威胁问题,由空气污染引起的癌症、中风以及心脏和大脑疾病每年会导致700万人的死亡[2]。因此各国越来越重视对空气质量的监测,并提出了各种用于评价空气质量的指标。其中,空气质量指数(Air Quality Index,AQI)是国际上普遍采用的判定空气质量的重要指标。它是定量描述空气质量状况的非线性无量纲指数,其数值越大,代表空气污染越严重,对人体的危害越大。根据2012年中华人民共和国生态环境部科技标准司发布的《环境空气质量指数(AQI)技术规定(试行)(HJ 633-2012)》[3],参与评价的大气污染物包括PM10、PM2.5、二氧化硫(SO2)、二氧化氮(NO2)、一氧化碳(CO)和臭氧(O3)。AQI数据可以为政府提供大气环境变化的记录,发现其中规律,从而用于对

到稿日期:2022-05-30 返修日期:2022-11-04
基金项目:国家自然科学基金(61772136);福建省引导性项目(2020H0008);福建省中青年教师教育科研项目(JAT210007)
This work was supported by the National Natural Science Foundation of China(61772136),Fujian Provincial Guiding Project(2020H0008) and Educational Research Project for Young and Middle-aged Teachers in Fujian Province(JAT210007).
通信作者:黄昉菀(hfw@fzu.edu.cn)

大气污染的控制和管理。

我国已经建成了以标准监测站为主、微型监测站为辅的大气环境监测体系。在监测站进行数据采集的过程中，一些不可控的因素所导致的数据缺失是不可避免的，而这些数据的缺失对分析和挖掘这些监测数据造成了一定的影响。此外，受限于成本，部署的监测站点往往也不够密集，这也会导致 AQI 数据在空间上的稀疏性。因此，在空气质量指数数据的分析与利用中，为了更加充分地利用已经搜集到的数据，对缺失数据进行补全是非常必要的。

目前，已有大量的研究致力于解决该问题。但绝大多数工作在高缺失率下表现不佳（详见第 2 节）。基于此，本文提出将缺失矩阵补全问题转换为稀疏矩阵重构问题，并设计了一种基于多维稀疏表示的数据补全方法（Multidimensional Sparse Representation Completion，MSRC）。实验结果表明，在高缺失率的多维时序 AQI 数据补全问题上，该方法的补全精度优于传统的矩阵补全方法。

2 相关工作

由于数据缺失问题的广泛存在，因此已有大量的研究致力于缺失数据的补全。这些方法针对的数据维数或者类型都不尽相同，也有各自的优缺点。

文献[4]提出了一种环境时空改进压缩感知算法（Environmental Space Time Improved Compressive Sensing，ESTICS），该方法虽然在很多数据集上得到使用，但其在缺失率较高时的补全效果不容乐观，这是压缩感知算法本身对数据结构的高要求所导致的。文献[5]迭代地使用从软阈值奇异值分解中获得的估计值来替换缺失值，可用于解决大规模低秩矩阵补全问题，但该方法没有针对高缺失率的应用场景进行研究。类似地，文献[6]提出了一个时间正则化矩阵分解框架（Temporal Regularized Matrix Factorization，TRMF），该框架利用数据之间的时间依赖进行建模，可以处理具有缺失值的高维时间序列数据挖掘任务，但并未验证其在高缺失率下的可行性。文献[7]提出了基于克罗内克压缩感知的补全算法，利用多变量时间序列的时空特性构造稀疏表示基，并根据缺失数据的位置设计度量矩阵，但该方法仅将缺失数据预测问题建模为稀疏向量恢复问题而非稀疏矩阵恢复问题。文献[8]提出将空气污染物空间矩阵分解为表示空间相关性的低秩矩阵和处理异常值的稀疏矩阵，稳健地填充空气污染物数据集中未观测到的值，但该方法要求污染物空间矩阵必须是低秩矩阵。为了提高缺失数据恢复的准确性，文献[9]利用每个数据序列的内部模式和跨源的相关信息来约束矩阵分解的目标函数，但该方法需要数据集额外增加不同来源的相关信息，限制了其应用场景。

近年来，各种深度学习数据补全方法也受到了越来越多的关注。如文献[10]使用长短期记忆神经网络（Long Short-Term Memory，LSTM）对 PM2.5 数据进行补全，相较于基本的插补算法，其能够使补全的数据更好地应用于预测中。但基于深度学习的方法在数据量较小时的性能较差，而当数据量较大时，其计算代价又很高。此外，也有研究将稀疏表示应用于缺失数据补全，但目前仅局限于处理一维向量数据的补全。例如文献[11]提出的采用单测量向量（Single Measurement Vector，SMV）的稀疏表示模型在数据缺失率高达 95% 的情况下仍然可以有效地补全缺失数据。该方法虽然也能用于矩阵补全，但是会丢失矩阵的潜在结构信息。

综上所述，传统矩阵分解模型存在的普遍问题是对矩阵缺失位置较为敏感，对于存在整行或整列缺失的矩阵，即使在低缺失率下补全效果也通常不尽人意。其次，随着缺失率的不断升高，部分模型的补全效果下降显著，有的模型甚至没有探究高缺失率下的补全精度。最后，近年来备受关注的深度学习模型对训练数据量和计算成本都有很高的要求，这显然不适合缺失大量数据或实时性要求较高的应用。

3 问题定义与系统框架

本文所研究的 AQI 数据缺失补全问题可形式化描述为：假设数据集 $\boldsymbol{Y}=\{\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k,\cdots,\boldsymbol{Y}_t\}$，包含共计 t 天的某个地区多个站点的监测数据。其中，第 k 天的 AQI 数据 $\boldsymbol{Y}_k\in\boldsymbol{R}^{m\times n}$，$n$ 表示站点个数，m 表示一天的观测时刻数，例如每小时观测一次，则 $m=24$。假设该数据集中有 l 天的完整数据，可以用其训练一个补全模型，对剩余 $t-l$ 天存在缺失的 AQI 数据进行补全。显然，该问题可被视为一个矩阵补全问题。受到稀疏表示在高缺失率向量补全中的应用的启发，本文提出了基于多维稀疏表示的数据补全方法 MSRC，其框架如图 1 所示。

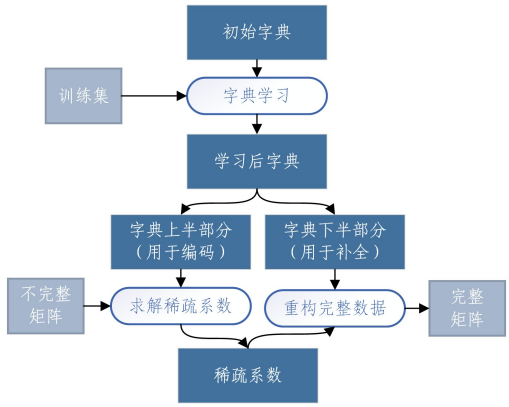

图 1 MSRC 框架图

Fig. 1 Frame diagram of MSRC

该方法的原理是利用训练数据学习一个合适的过完备字典，该字典的上半部分承担缺失矩阵编码功能，用于提取缺失矩阵的稀疏表示；字典的下半部分则承担稀疏矩阵重构功能，用于恢复完整矩阵，实现对缺失值的补全。

4 基于多维稀疏表示的数据补全方法

在过去的十年里，寻找目标信号的稀疏表示已经成为一项重要的研究工作。稀疏表示的概念表明，一个信号可以被分解成几个基本信号的线性组合。因此，仅需少数的非零元素就能够代表目标信号所传递的大部分信息。目前，稀疏表示已成为许多应用中非常重要的工具，如信号重构和特征选择等[12-13]。稀疏表示主要有两种不同的模型：采用单测量向量（SMV）模型来表示一维原始信号（可视为向量）和采用多测量向量（Multiple Measurement Vector，MMV）模型来表示多维原始信号（可视为矩阵）。由于本文的研究对象是 n 个

站点 m 个观测时刻的 AQI 数据，因此应使用 MMV 模型进行稀疏表示。下面首先简要介绍 MMV 模型，然后介绍本文提出的基于多维稀疏表示的数据补全方法。

4.1 多维稀疏表示简介

多测量向量模型 MMV 可利用一个过完备字典 $\boldsymbol{D}\in\boldsymbol{R}^{m\times p}(p\gg m)$处理多维原始信号$\boldsymbol{Y}_k\in\boldsymbol{R}^{m\times n}$并得到其稀疏系数矩阵$\boldsymbol{X}_k\in\boldsymbol{R}^{p\times n}$[14]。$\boldsymbol{X}_k$的特点是所有列具有相同的稀疏性结构，即每一列的非零项都出现在同一行。其示意图如图 2 所示，字典 $\boldsymbol{D}$ 中的虚线框出的列与稀疏系数矩阵 $\boldsymbol{X}$ 中的非零行相关联。图中灰色块表示该位置的元素值非零，白色块则表示元素值为零。

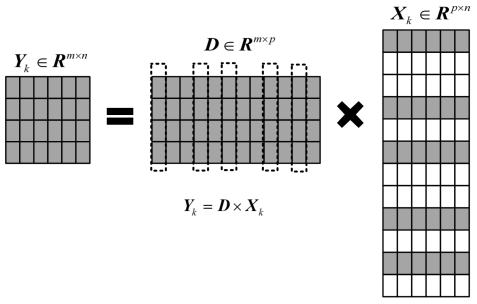

图 2 MMV 模型的示意图

Fig. 2 Schematic diagram of MMV model

MMV 的关键在于构建合适的过完备字典 $\boldsymbol{D}$ 并求解基于 $\boldsymbol{D}$ 能够重构出原始信号矩阵 $\boldsymbol{Y}$ 的最稀疏的系数矩阵$\boldsymbol{X}_k$，其目标函数为：

$$\min S(\boldsymbol{X}_k)\quad \text{s.t.}\ \boldsymbol{Y}_k=\boldsymbol{D}\boldsymbol{X}_k \tag{1}$$

其中，$S(\boldsymbol{X}_k)$为$\boldsymbol{X}_k$的稀疏度。考虑到稀疏表示存在一定的误差，可将式(1)的约束条件放松为：

$$\min S(\boldsymbol{X}_k)\quad \text{s.t.}\ \|\boldsymbol{Y}_k-\boldsymbol{D}\boldsymbol{X}_k\|_2^2\leqslant\delta \tag{2}$$

其中，δ 是一个预先给定的正实数，可将其视为重构误差的上界。通过调换优化目标和约束条件，式(2)还可转换为：

$$\min\ \|\boldsymbol{Y}_k-\boldsymbol{D}\boldsymbol{X}_k\|_2^2\leqslant\delta\quad \text{s.t.}\ S(\boldsymbol{X}_k)\leqslant d \tag{3}$$

其中，d 是一个正整数，可将其视为稀疏度的上界。最后，上述带约束条件的优化问题还可以通过拉格朗日乘子定理转换为无约束条件的优化问题：

$$\min\ \|\boldsymbol{Y}_k-\boldsymbol{D}\boldsymbol{X}_k\|_2^2+\lambda S(\boldsymbol{X}_k) \tag{4}$$

其中，λ 为正则化因子($0<\lambda<1$)，用于权衡重构误差和稀疏度之间的相对重要性。

由于$\boldsymbol{X}_k$的每一列都具有相同数量的非零项，因此$\boldsymbol{X}_k$的稀疏性$S(\boldsymbol{X}_k)$可以等价于其第 j 列向量$\boldsymbol{x}_k^j$的稀疏度($j=1,2,\cdots,n$)，即令 $S(\boldsymbol{X}_k)=\|\boldsymbol{x}_k^j\|_0$。$\|\cdot\|_0$表示向量的$l_0$范数，指向量中的非零项个数。需要说明的是，解决$l_0$范数最小化问题需要枚举所有行中可能的情况，属于 NP-hard 问题[15]。为了解决该问题，Chen 等[16]提出在 MMV 模型中，可以将l_0范数最小化问题转化为l_1范数最小化问题，即令：

$$S(\boldsymbol{X}_k)=\max\{\|\boldsymbol{x}_k^1\|_1,\cdots,\|\boldsymbol{x}_k^j\|_1,\cdots,\|\boldsymbol{x}_k^n\|_1\} \tag{5}$$

其中，$\|\cdot\|_1$为向量的l_1范数，指向量中所有非零项的绝对值之和。

4.2 基于多维稀疏表示的数据补全方法 MSRC

基于 MMV 的数据补全方法 MSRC 的思路是将过完备字典分割成上下两个部分，其中上半部分用于提取不完整矩阵(含有缺失值)的主要特征，而下半部分则承担获得估计矩阵(缺失位置填充了估计值)的作用。在上述过程中，字典的好坏将直接影响缺失值的估计精度。为了学习到一个合适的字典，本文首先将数据集 $\boldsymbol{Y}$ 划分成两部分：不存在缺失值的训练集$\boldsymbol{Y}^A$(共计 l 天)和存在缺失值的测试集 $\boldsymbol{Y}^B$(共计 r 天)。训练阶段的具体步骤如下：

步骤 1 将每一天的训练数据$\boldsymbol{Y}_u^A\in\boldsymbol{R}^{m\times n}$($u=1,2,\cdots,l$)先进行随机缺失，再用同一列的最近邻对缺失值进行暂时性预补，得到矩阵$\boldsymbol{Y}'^A_u$。

步骤 2 将$\boldsymbol{Y}_u^A$和$\boldsymbol{Y}'^A_u$进行上下拼接得到矩阵$\boldsymbol{W}_u^A=\begin{bmatrix}\boldsymbol{Y}'^A_u\\ \boldsymbol{Y}_u^A\end{bmatrix}\in\boldsymbol{R}^{2m\times n}$。

步骤 3 构建一个初始的过完备字典$\boldsymbol{D}^0\in\boldsymbol{R}^{2m\times p}(p\gg 2m)$，并将其上半部分命名为$\boldsymbol{D}^0_{\text{up}}\in\boldsymbol{R}^{m\times p}$，下半部分命名为$\boldsymbol{D}^0_{\text{lw}}\in\boldsymbol{R}^{m\times p}$。

步骤 4 利用所有的$\boldsymbol{W}_u^A$($u=1,2,\cdots,l$)对字典 $\boldsymbol{D}$ 进行共计 i 轮的迭代学习，得到合适的字典$\boldsymbol{D}^i\in\boldsymbol{R}^{2m\times p}$。该字典被认为对所有的$\boldsymbol{W}_u^A$($u=1,2,\cdots,l$)具有很好的重构能力。

鉴于此，本文将该字典运用于测试集缺失数据的补全。其过程如图 3 所示，图中斜线块表示该位置的数值存在缺失，但已使用同一列的未缺失最近邻进行预补。

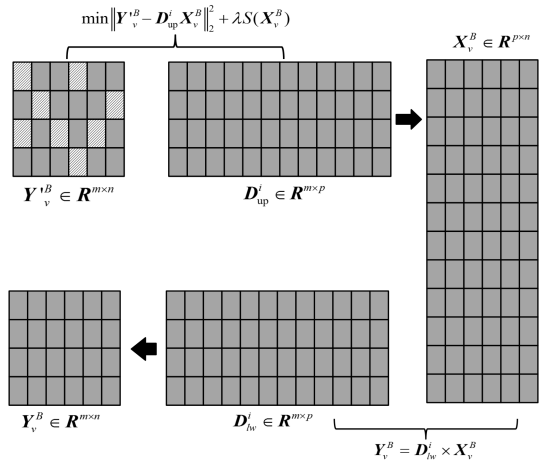

图 3 MSRC 模型用于矩阵补全的示意图

Fig. 3 Schematic diagram of MSRC model used for matrix completion

其具体步骤如下：

步骤 1 对于测试集中每一个不完整矩阵$\boldsymbol{Y}_v^B\in\boldsymbol{R}^{m\times n}$($v=1,2,\cdots,r$)，同样利用同一列的最近邻对缺失值进行暂时性预补，得到矩阵$\boldsymbol{Y}'^B_v$($v=1,2,\cdots,r$)。

步骤 2 利用字典$\boldsymbol{D}^i$的上半部分$\boldsymbol{D}^i_{\text{up}}\in\boldsymbol{R}^{m\times p}$求解每一个$\boldsymbol{Y}'^B_v$的稀疏系数矩阵$\boldsymbol{X}_v^B\in\boldsymbol{R}^{p\times n}$。

步骤 3 将字典$\boldsymbol{D}^i$的下半部分$\boldsymbol{D}^i_{\text{lw}}\in\boldsymbol{R}^{m\times p}$与$\boldsymbol{X}_v^B$相乘，得到估计矩阵$\widetilde{\boldsymbol{Y}}_v^B\in\boldsymbol{R}^{m\times n}$。最后将$\widetilde{\boldsymbol{Y}}_v^B$中对应于$\boldsymbol{Y}_v^B$缺失位置的元素值填充到$\boldsymbol{Y}_v^B$中。

4.3 字典构造与学习

在 MSRC 方法的训练过程中，过完备字典的构造与学习是极其重要的环节，下面介绍其具体实现方法。

4.3.1 *初始字典构造*

由于存在字典学习环节，因此初始字典可以采用一些常见的数学方法直接生成，如离散傅里叶变换（Discrete Fourier Transform，DFT）、离散余弦变换（Discrete Cosine Transform，DCT）[17]等。本文使用 Matlab 中的二维离散余弦变换函数来生成一个初始字典$\boldsymbol{D}^0$。

4.3.2 *字典学习环节*

鉴于初始字典在重构信号时存在较大的误差，需要利用字典学习对其进行改进。字典学习的目的是让字典能够更好地表达出原始信号，这是一个不断减小重构误差的过程。在向量稀疏表示中，K-SVD 算法是最常见的字典学习算法[18]。由于该算法是逐列对字典进行更新，因此会产生以下两个问题。其一，因为每次只更新一列，所以计算复杂度很高。其二，该算法只适用于 SMV 模型，并不适用于本文的 MMV 模型。针对上述不足，本文在 MSRC 的字典学习环节采用了 multi-KSVD 算法，利用对误差矩阵的奇异值分解同时更新稀疏系数矩阵中的所有非零项和字典中对应的所有列向量[15]。字典迭代更新（共 i 轮）过程如图 4 所示。

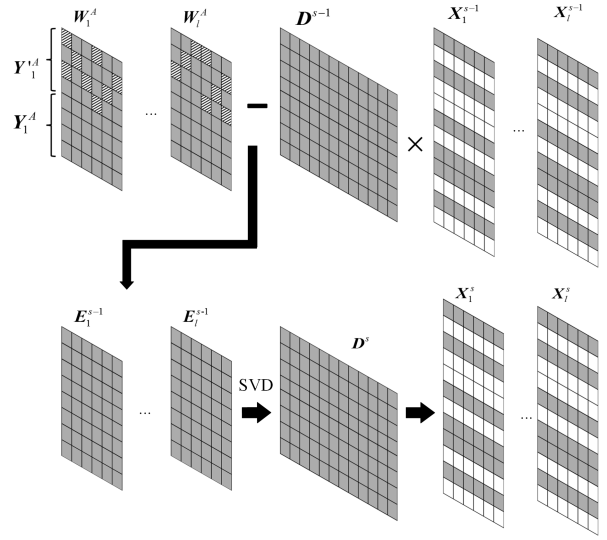

图 4 multi-KSVD 算法更新字典的示意图

Fig. 4 Schematic diagram of multi-KSVD algorithm updating dictionary

具体步骤如下：

步骤 1 对所有的$\boldsymbol{W}_u^A$（$u=1,2,\cdots,l$），利用上一轮得到的字典$\boldsymbol{D}^{s-1}$（$1\leqslant s\leqslant i$）求解它们的稀疏系数矩阵$\boldsymbol{X}_u^{s-1}$（$u=1,2,\cdots,l$）。

步骤 2 利用$\boldsymbol{W}_u^A-\boldsymbol{D}^{s-1}\boldsymbol{X}_u^{s-1}$得到重构误差矩阵$\boldsymbol{E}_u^{s-1}$（$u=1,2,\cdots,l$）。

步骤 3 对$\boldsymbol{E}^{s-1}=[\boldsymbol{E}_1^{s-1},\cdots,\boldsymbol{E}_l^{s-1}]$进行矩阵的奇异值分解得到更新后的字典$\boldsymbol{D}^s$。

步骤 4 重复步骤 1－步骤 3，直至完成 i 轮迭代。

4.4 求解稀疏系数矩阵

在 1.2 节训练阶段的步骤 4（字典学习）和补全阶段的步骤 2 都涉及求解矩阵基于过完备字典的稀疏表示（即稀疏系数矩阵）。本文采用多维稀疏贝叶斯学习算法（Multidimensional Sparse Bayesian Learning，M-SBL）进行求解，这是因为该算法对噪声具有较好的鲁棒性，其性能优于混合范数方法或子空间贪婪算法[19-20]。M-SBL 算法采用了一种利用自动相关性的经验贝叶斯策略，可以认为是标准稀疏贝叶斯学习算法 SBL 从求解向量稀疏表示到求解矩阵稀疏表示的拓展。M-SBL 算法产生的稀疏系数矩阵符合 MSRC 的需要，其每列共享相同的稀疏性结构。

5 实验设计及结果分析

5.1 实验设计说明

本文使用北京市环境保护检测中心网站提供的 2020 年 3 月份北京市 AQI 数据进行实验，旨在验证 MSRC 在矩阵缺失补全方面的有效性[21]。该数据集包含北京市 34 个站点的 AQI 数据，采集频率为每小时 1 次，因此每天的 AQI 数据可视为一个矩阵$\boldsymbol{Y}_k\in\boldsymbol{R}^{24\times 34}$。本文将前 25 天的数据作为训练集，将最后 5 天的数据作为测试集。测试集中第 30 天的部分数据如表 1 所列（缺失率为 55%），其中有两个值的方格表示该处数字被模拟缺失，括号内为 MSRC 的估计值。

表 1 3 月 30 日部分数据展示

Table 1 Partial data display on March 30

时间	站点 30	站点 31	站点 32	站点 33	站点 34
19	69	89	108(111)	94(97)	88(93)
20	73	94	113(118)	95	89(98)
21	78	94(99)	119(121)	103	90(100)
22	79	91(101)	121(122)	110(107)	94

本文将 MSRC 与均值填充（MEAN）、软奇异值迭代补全算法（softImpute）、环境时空改进压缩感知算法（ESTICS）以及无字典学习环节的 MSRC-NL 算法（Multidimensional Sparse Representation Completion-No Learning）进行补全性能的比较。上述基准方法的具体说明如下：

（1）均值填充使用同一列上未缺失数据的均值作为该列上缺失值的估计值。

（2）softImpute 通过核范数正则化将低秩矩阵近似拟合到缺失值矩阵。该方法用当前的猜测值填充缺失值，然后用软阈值奇异值分解算法在完整矩阵上求解优化问题，并进行多次迭代。

（3）ESTICS 算法在传统矩阵分解的基础上增加了时间平滑和空间相关两种约束。最后利用得到的两个分解矩阵 $\boldsymbol{L}$ 和 $\boldsymbol{R}$ 计算估计矩阵$\widetilde{\boldsymbol{Y}}_v^B=\boldsymbol{L}\boldsymbol{R}^{\mathrm{T}}$。

（4）MSRC-NL 指直接使用生成的初始字典进行补全，也就是跳过了 MSRC 的字典学习阶段。添加该实验是为了验证字典学习在 MSRC 算法中的必要性。

为了全面评估各种方法在不同缺失率下的表现，本文将测试集进行随机缺失，缺失率由 5%逐步增加到 95%，每次增加 10%。需要说明的是，当缺失率较高时，在测试矩阵中有可能出现整列缺失的情况，这会导致部分基准方法效果非常差，甚至出现执行错误的问题。因此，当缺失矩阵出现整列缺失时，本文在该列的第一行填充整个矩阵已有观测值的均值。

5.2 实验结果分析

实验主要比较 AQI 数据在不同方法下的补全效果，可通过比较真实值y_i和估计值$\widetilde{y}_i$的差异得到。本文采用了缺失

补全领域普遍采用的3个评价指标：平均绝对百分比误差(Mean Absolute Percentage Error，MAPE)、平均绝对误差(Mean Absolute Error，MAE)、确定系数R^2(R-square)。它们的计算公式如下：

$$MAPE=\frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{\tilde{y}_i-y_i}{y_i}\right| \tag{6}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|\tilde{y}_i-y_i| \tag{7}$$

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\tilde{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2} \tag{8}$$

其中，N 为测试集中缺失值的个数，$\bar{y}_i$ 为 N 个 y_i 的均值。上述3个指标中，除了R^2是越接近1越好之外，其余2个指标均是越小越好。

表2－表4分别列出了MSRC与基准方法在不同评价指标上的结果。从图5－图7可以更直观地看出在不同缺失率下各方法补全效果的变化。

表2 不同缺失率下各方法的*MAPE*值

Table 2 *MAPE* of each method with different missing rates

缺失率	MEAN	SoftImpute	ESTICS	MSRC-NL	MSRC
5%	31.8888	13.4564	28.0511	78.3208	18.602
15%	29.2116	12.5381	27.1109	64.5203	18.4477
25%	29.9631	22.8538	29.6654	61.5751	19.7468
35%	31.3593	36.8330	29.6630	72.2707	22.8757
45%	30.8525	53.4077	31.1720	59.3358	23.0027
55%	29.2046	67.6408	31.6926	58.1292	23.7284
65%	29.4178	80.3238	32.9810	55.9783	24.3205
75%	32.9402	89.1771	37.7528	63.3705	26.3449
85%	33.0022	95.9358	50.0285	59.9838	28.0522
95%	51.6686	99.3569	68.3143	60.6335	29.5409

表3 不同缺失率下各方法的*MAE*值

Table 3 *MAE* of each method with different missing rates

缺失率	MEAN	SoftImpute	ESTICS	MSRC-NL	MSRC
5%	10.6451	4.2760	11.4682	32.8861	6.6758
15%	10.9722	5.3106	12.5998	29.3067	6.9541
25%	11.4922	11.7509	14.2875	30.3959	7.5635
35%	11.8215	20.4862	14.4616	34.1312	8.1715
45%	11.9494	29.2157	15.9824	30.4869	9.1969
55%	11.9748	36.4183	17.2145	28.3854	9.8612
65%	12.0873	42.0059	18.3907	26.4690	9.4900
75%	12.5946	45.5688	20.5443	30.3593	10.5305
85%	13.4142	48.6146	26.8637	28.6506	12.0482
95%	24.5357	50.0966	35.4429	30.9607	13.1837

表4 不同缺失率下各方法的*RMSE*值

Table 4 *RMSE* of each method with different missing rates

缺失率	MEAN	SoftImpute	ESTICS	MSRC-NL	MSRC
5%	14.4949	6.3112	16.2009	45.3099	9.3905
15%	14.9807	8.0687	18.3224	37.8712	9.9519
25%	16.6855	17.5773	20.8580	39.7263	10.8283
35%	16.4630	28.3937	20.8007	47.9982	11.8295
45%	16.8534	37.5317	23.2416	40.2896	14.9728
55%	17.6193	45.1637	24.8788	38.0111	15.0954
65%	17.6011	50.6647	26.1560	38.2901	13.9052
75%	18.0508	54.4047	28.3060	40.0642	16.1077
85%	19.7471	57.5166	35.2912	38.2582	18.5631
95%	36.9970	58.9430	45.3546	40.6442	20.0081

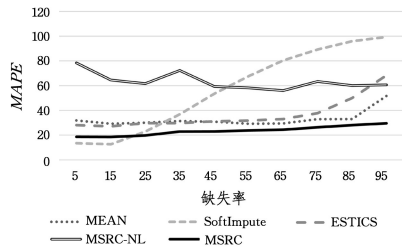

图5 不同缺失率下各方法的*MAPE*值变化

Fig. 5 Changes in *MAPE* of each method with different missing rates

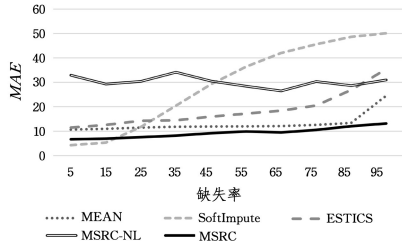

图6 不同缺失率下各方法的*MAE*值变化

Fig. 6 Changes in *MAE* of each method with different missing rates

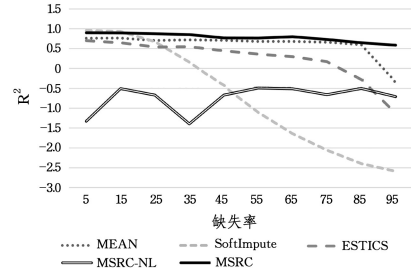

图7 不同缺失率下各方法的R^2值变化

Fig. 7 Changes in R^2 of each method with different missing rates

通过对比表2－表4的实验结果可以发现，在低缺失率情况下(5%～15%)，softImpute表现最优，MSRC紧随其后。虽然ESTICS与SoftImpute，MSRC一样都是矩阵分解算法，但ESTICS的表现最差，原因是ESTICS增加了时间平滑和空间相关两种约束，使其对降低补全误差的关注度有所减少。但是随着缺失率的不断升高，softImpute的性能急剧下降。当缺失率超过15%后，其表现不如MSRC；当缺失率超过25%后，其表现也不如ESTICS。出现该现象的原因是随着缺失率的升高，矩阵中的列或行出现数据完全缺失的概率不断增大，这对softImpute的性能影响最大，而添加了时间和空间约束的ESTICS受其影响则较小。MSRC由于利用完整的历史数据模拟各种缺失情况对字典进行了充分学习，因此补全效果优于上述两种方法。最后，由图4－图7可以看出，除了MSRC-NL，其余所有方法在数据缺失率不断提高的情况下补全精度都不断降低。但是本文提出的MSRC补全精度的降幅最小，充分说明了本文方法对不同缺失率的鲁棒性。通过比较MSRC与MSRC-NL的实验结果可以看出，字典学习对于MSRC方法来说非常重要。没有学习的MSRC-NL难以提取数据的特征，无论在何种缺失率下，其补全效果都不佳。

综上所述，MSRC方法在多维时序AQI数据补全问题上优于传统的矩阵补全方法，尤其是在数据缺失比较严重的

情况下具有明显的优势。

结束语 针对多维时序 AQI 数据补全问题，本文提出了一种基于多维稀疏表示的数据补全方法 MSRC。该方法最大的特色在于利用字典学习技术实现缺失矩阵的稀疏表示，并将其用于生成缺失值的估计。在 2020 年 3 月北京市空气质量指数数据集上的实验表明，即使在数据大量缺失的情况下，所提方法也能有效地恢复出整个矩阵的缺失值。当缺失情况较为严重时，该方法相比传统的矩阵补全方法具有明显的优势。该项研究不仅能够拓展稀疏表示的更多应用，并且能够对大气污染治理与预防有所贡献。在未来的工作中，将致力于将该算法拓展到其他应用领域，并研究其与稀疏群智感知技术的结合。

参 考 文 献

[1] WU R，SONG X，BAI Y，et al. Are current Chinese national ambient air quality standards on 24-hour averages for particulate matter sufficient to protect public health? [J]. Journal of Environmental Sciences，2018，71：67-75.

[2] World Health Organization. World health statistics 2019：monitoring health for the SDGs，sustainable development goals[M]. World Health Organization，2019.

[3] Ministry of Environmental Protection. Technical Regulation on Ambient Air Quality Index (on trial)：HJ 633-2012[S]. Beijing：China Environmental Science Press，2012.

[4] KONG L，XIA M，LIU X Y，et al. Data loss and reconstruction in sensor networks[C]// 2013 Proceedings IEEE INFOCOM. IEEE，2013：1654-1662.

[5] MAZUMDER R，HASTIE T，TIBSHIRANI R. Spectral regularization algorithms for learning large incomplete matrices[J]. The Journal of Machine Learning Research，2010，11：2287-2322.

[6] 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，2016，29：847-855.

[7] GUO Y，SONG X，LI N，et al. An efficient missing data prediction method based on Kronecker compressive sensing in multivariable time series[J]. IEEE Access，2018，6：57239-57248.

[8] LIU X，WANG X，ZOU L，et al. Spatial imputation for air pollutants data sets via low rank matrix completion algorithm[J]. Environment International，2020，139：105713.

[9] SONG X，GUO Y，LI N，et al. A Novel Approach Based on Matrix Factorization for Recovering Missing Time Series Sensor Data[J]. IEEE Sensors Journal，2020，20(22)：13491-13500.

[10] YUAN H，XU G，YAO Z，et al. Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks[C]// Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 2018：1293-1300.

[11] LI P G，YU Z Y，HUANG F W. Power load data completion based on sparse representation[J]. Computer Science，2021，48(2)：128-133.

[12] RAJ S，RAY K C. Sparse representation of ECG signals for automated recognition of cardiac arrhythmias[J]. Expert Systems with Applications，2018，105：49-64.

[13] HE G，DING K，LIN H. Fault feature extraction of rolling element bearings using sparse representation[J]. Journal of Sound and Vibration，2016，366：514-527.

[14] DU X，CHENG L，CHENG G. A heuristic search algorithm for the multiple measurement vectors problem[J]. Signal Processing，2014，100：1-8.

[15] YANG J，MA J. Feed-forward neural network training using sparse representation[J]. Expert Systems with Applications，2019，116：255-264.

[16] CHEN J，HUO X. Theoretical results on sparse representations of multiple-measurement vectors[J]. IEEE Transactions on Signal Processing，2006，54(12)：4634-4643.

[17] RAO K R，YIP P. Discrete cosine transform：algorithms，advantages，applications[M]. Academic Press，2014.

[18] AHARON M，ELAD M，BRUCKSTEIN A. K-SVD：an algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing，2006，54(11)：4311-4322.

[19] YE J C，KIM J M，BRESLER Y. Improving M-SBL for joint sparse recovery using a subspace penalty[J]. IEEE Transactions on Signal Processing，2015，63(24)：6595-6605.

[20] WIPF D P，RAO B D. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem [J]. IEEE Transactions on Signal Processing，2007，55(7)：3704-3716.

[21] Beijing Municipal Ecological and Environmental Monitoring Center [DB/OL]. http：//www. bjmemc. com. cn/.

CAI Qiquan，born in 1997，undergraduate. His main research interests include data completion and so on.

HUANG Fangwan，born in 1980，postgraduate，senior lecturer，is a member of China Computer Federation. Her main research interests include computational intelligence，machine learning and big data analysis.

（责任编辑：何杨）