计算机科学
COMPUTER SCIENCE

基于压缩感知的相关性数据填补方法

任兵, 郭艳, 李宁, 刘存涛

引用本文

任兵, 郭艳, 李宁, 刘存涛.基于压缩感知的相关性数据填补方法[J]. 计算机科学, 2023, 50(7): 82-88.

REN Bing, GUO Yan, LI Ning, LIU Cuntao. Method for Correlation Data Imputation Based on Compressed Sensing [J]. Computer Science, 2023, 50(7): 82-88.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600209

基于压缩感知的相关性数据填补方法

任 兵 郭 艳 李 宁 刘存涛
陆军工程大学通信工程学院 南京 210007
(renbing2872@sina.com)

摘 要 数据缺失现象在数据的采集和传输过程中经常发生,而对数据集中缺失数据的不当填补,会对后续的数据挖掘工作产生不利的影响。为了更有效地对缺失数据集进行填补,针对相关性数据,提出了一种基于压缩感知的缺失数据填补方法。首先,将缺失数据填补问题转化为压缩感知框架下的稀疏向量恢复问题;其次,针对数据的相关性特点构造了专门的稀疏表示基,从而能够更好地实现数据的稀疏化;最后,提出了一种快速迭代加权阈值算法,在传统的快速迭代收缩阈值算法的基础上引入了一种新的加权因子及重启动策略,提高了算法的收敛性能和数据的重构精度。仿真结果表明,所提算法能够高效地填补缺失数据,与传统的快速迭代收缩阈值算法相比,重构成功率和重构速度都得到了提升。同时,在数据稀疏变换效果较差的情况下,所提算法仍然能够完成对缺失数据集的填补,具有更好的鲁棒性。

关键词:压缩感知;数据填补;相关性数据;正交特征向量基;迭代加权阈值法
中图法分类号 TN911.7

Method for Correlation Data Imputation Based on Compressed Sensing

REN Bing,GUO Yan,LI Ning and LIU Cuntao
College of Communication Engineering,Army Engineering University of PLA,Nanjing 210007,China

Abstract The phenomenon of missing data occurs frequently during the acquisition and transfer of data,and improper handling of missing data sets can adversely affect subsequent data mining efforts. In order to fill the missing data set more effectively,a method for data imputation based on compressed sensing is proposed for correlation data. First,the problem of missing data imputation is transformed into a sparse vector recovery problem under the compressed sensing framework. Second,a specialized sparse representation base is constructed for correlation data,so the data sparsity can be better realized. Finally,the fast iterative weighted thresholding algorithm(FIWTA) is proposed,which is refined based on the fast iterative shrinkage-thresholding algorithm (FISTA). The proposed algorithm adopts a new iterative weighted method and introduces a restart strategy,which greatly improves the convergence of the algorithm and the reconstruction accuracy of the data. Simulation results show that the proposed algorithm is able to fill the missing data efficiently,and both the reconstruction success rate and the reconstruction speed are improved compared with the traditional fast iterative shrinkage-thresholding algorithm. Meanwhile,even when the sparse transformation of the data is less effective,imputation of missing data sets can still be accomplished with better robustness.

Keywords Compressed sensing,Data imputation,Correlation data,Orthonormal eigenvector basis,Iterative weighted thresholding algorithm

1 引言

随着科学技术的发展,数据的采集与使用日益成为构建现代社会的基石。特别是近年来,由于传感器的广泛使用,在海量数据的基础上诞生了一个全新的数字社会,数据在生产生活中的重要性愈加凸显[1-2]。传感器大多结构简单、功耗很低,因此仅被用于测量温度、压力、湿度等环境参数,测量值需要被传送到数据中心进行进一步处理。在数据的传输过程中,由于硬件故障、信道衰落、信道冲突和线路阻断等,数据丢失的情况非常普遍。

对缺失数据集进行预处理是数据分析前的一个重要工作,如果不能准确地填补数据集中的缺失值,很多现有的数据分析方法将无法使用。当数据集的缺失值较少时,常常采用直接删除的方法完成处理工作。但是,当缺失值较多时,直接删除会导致大量信息的丢失,特别是当数据量本身较小时,直接删除就很不合时宜。对缺失值进行填补是重构不完整数据

到稿日期:2022-06-23 返修日期:2022-07-20
基金项目:国家自然科学基金(61871400);江苏省自然科学基金(BK20211227)
This work was supported by the National Natural Science Foundation of China(61871400) and Natural Science Foundation of Jiangsu Province,China(BK20211227).
通信作者:郭艳(guoyan_1029@sina.com)

集的另一种思路，实践中通常采用统计学方法或机器学习方法，用计算得出的估计值来替换缺失值，以完成对缺失数据集的填补[3]。

目前研究者们已经提出了大量的方法来完成缺失值的估计，不同的填补方法有各自的优势和缺点。线性插值是最简单的数据填补方法之一，但在连续数据丢失的情况下，算法的精度往往不佳[4]；K 近邻法（KNN）不需要提前预估缺失值的定量或定性属性，并且该方法可以直接处理多个缺失值，但在大数据集上运行时，所需的时间较长[5]；基于粗糙集理论的数据填补方法是处理不确定性问题的有效工具，但其无法完成参数优化和缺失数据分类，导致数据填补的精度较低[6]；随机森林（RF）可以处理高维度数据且准确率较高，但是在处理大量数据时计算开销很大[7]；支持向量机（SVM）对异常值不敏感，鲁棒性强，但算法的计算复杂度较高[8]；神经网络的性能优越，但训练需要庞大的数据集且容易产生过拟合的问题[9]；张量补全方法在高维数据填补中具有优势，其充分地利用了不同维度数据之间的隐含信息，但也存在着计算复杂的问题[10-11]。上述算法在各自情境下都能够取得较好的效果，但是普遍无法在有大量数据缺失时有效地恢复数据，而基于压缩感知的重构算法能够很好地解决此类问题。

压缩感知是一种新兴的采样理论，能够在较小采样数的基础上完成对原始信号的重构。根据奈奎斯特采样定理，想让采样之后的数字信号完整保留原始信号中的信息，采样频率必须高于原始信号最高频率的 2 倍。压缩感知的方法突破了采样原理的限制，能够以远低于最高频率 2 倍的采样频率完成信号采样，极大地“压缩”了采样的数量。根据信号的稀疏性质，只需采样少量的数据值，即可通过求解一个非线性的优化问题完成对原始信号的恢复。也就是说，采集到的数据是一个稀疏的向量，通过重构算法即可完成稀疏向量的重构。那么，对于存在大量缺失数据的不完整数据集，我们也可以将其看作是一个稀疏的向量，利用压缩感知的原理，依托少量已知数据完成对数据集的补全。

在压缩感知体系中，稀疏矩阵构造、测量矩阵设计和重构算法选择是数据重构的关键，其中重构算法在很大程度上决定了重构的效率。基于凸优化类的算法是一类常用的重构算法，迭代收缩阈值算法（Iterative Shrinkage-thresholding Algorithm，ISTA）是对梯度下降方法的一次重要改进，但是 ISTA 的收敛速度并不令人满意。为了解决这个问题，快速迭代收缩阈值算法（Fast Iterative Shrinkage-thresholding Algorithm，FISTA）引入了 Nesterov 加速步骤，取得了很好的效果[12]。于是 FISTA 成为一种经典的重构算法，也出现了大量对 FISTA 算法的改进研究。文献[13]提出了一种 ASFISTA 算法，其对 FISTA 算法进行了自适应步长的改进，从而提高了算法的收敛性能；文献[14]总结出一种针对 FISTA 算法的通用回溯策略，有效降低了迭代的成本；文献[15]将顺序子空间优化（SESOP）应用于 FISTA 算法，进一步加快了算法的收敛速度；文献[16]使用了一种逐元素自适应阈值的 LISTA 算法，极大地提升了算法收敛速度；文献[17]设计了一种无特征值迭代收缩阈值算法（EFISTA），避免了大规模问题中特征值计算困难的问题。

根据以上分析，本文提出了一种基于压缩感知的缺失数据填补方法。首先，将缺失数据填充问题转化为稀疏信号的恢复问题。其次，基于 FISTA 算法进行改进，采用一种新的迭代加权方法，并引入重启动策略，提出了一种快速迭代加权阈值算法。该算法极大地提高了 FISTA 算法的速度和精度，能够在稀疏度未知的情况下完成对缺失数据集的重构。仿真结果表明，相比其他算法，该方法能够更高效地填补缺失数据，恢复速度更快且鲁棒性更好。

2　压缩感知理论

已知一个 $N\times1$ 维信号 $\boldsymbol{x}$，它的 N 个元素中只有 K 个是非零的，其中 $K\ll N$，那么我们称这个向量是严格稀疏的。当信号是稀疏的，就可以将信号压缩表示为：

$$\boldsymbol{y}=\boldsymbol{\Phi x} \tag{1}$$

其中，$\boldsymbol{y}$ 是 $M\times1$ 维测量信号，$\boldsymbol{\Phi}$ 是 $M\times N$ 维的测量矩阵，$\boldsymbol{x}$ 是采集的原始 $N\times1$ 维信号。

现实中的信号并不总是稀疏的。我们可以将信号变换到另一组基底上，使得信号在该组基底上是稀疏的，这就是信号的稀疏表示：

$$\boldsymbol{x}=\boldsymbol{\Psi\theta} \tag{2}$$

其中，$\boldsymbol{\Psi}$ 为 $N\times N$ 维的稀疏矩阵，$\boldsymbol{\theta}$ 为满足稀疏条件的 $N\times1$ 维信号。

一般来说，如果除了 K 个值以外的其他值都很小，我们也认为这个向量是 K 稀疏的，那么有：

$$\boldsymbol{y}=\boldsymbol{\Phi x}=\boldsymbol{\Phi\Psi\theta}=\boldsymbol{A\theta} \tag{3}$$

其中，$\boldsymbol{A}=\boldsymbol{\Phi\Psi}$ 是 $M\times N$ 维的测量矩阵。

接收端在收到信号 $\boldsymbol{y}$ 后，需要完成信号 $\boldsymbol{x}$ 的重构。文献[18]中引入了约束等距特性（Restricted Isometry Property，RIP），假设 $\boldsymbol{x}$ 已经满足稀疏特性，只要测量矩阵 $\boldsymbol{\Phi}$ 满足以下约束等距特性（RIP），那么即便在有噪声的情况下也能成功重构出原始信号。

$$(1-\delta_K)\|\boldsymbol{x}\|_2^2\leqslant\|\boldsymbol{\Phi x}\|_2\leqslant(1+\delta_K)\|\boldsymbol{x}\|_2^2 \tag{4}$$

其中，δ_k 被称为约束等距常数，且 $\delta_k\in(0,1)$。满足以上条件后，通过求解一个优化问题，就可以得到稀疏信号 $\hat{\boldsymbol{\theta}}$，从而恢复出 N 维原始信号 $\hat{\boldsymbol{x}}$。

3　数据填补模型

3.1　稀疏表示基设计

根据压缩感知的原理，要实现信号的成功重构，待恢复的数据必须具有较好的稀疏性能。但是，在实际的数据集中，采样后的数据并不能很好地达到所要求的稀疏性能。当前，已经有几种常见的稀疏表示基能够很好地实现数据的稀疏功能，常见的稀疏表示基有离散余弦变换（Discrete Cosine Transform，DCT）、离散小波变换（Discrete Wavelet Transform，DWT）、离散傅里叶变换（Direct Fouriet Transform，DFT）。

我们日常收集到的传感器数据大部分都具有很强的相关性。由于协方差矩阵能够很好地去除数据的相关性，因此可以尝试利用协方差矩阵的这个性质来完成数据的稀疏化[19]。

定义协方差矩阵为：

$$E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} \tag{5}$$

其中，$\boldsymbol{U}$ 是正交特征向量基（Orthonormal Eigenvector Basis，OEB），$\boldsymbol{\Lambda}$ 是对角线为特征值的对角矩阵。我们将 $\boldsymbol{U}$ 作为稀疏表示基，那么稀疏过程可以表示为：

$$\boldsymbol{x}=\boldsymbol{\Psi}\boldsymbol{\theta}=\boldsymbol{U}\boldsymbol{\theta} \tag{6}$$

根据压缩感知理论，只要信号除了 K 个值以外的其他值都很小，我们也认为这个向量是 K 稀疏的。通过上面的分析，只需讨论协方差矩阵的特征值是否符合稀疏性条件，即可判断经稀疏后数据的实际稀疏性。

Electricity 数据集包含了 2012－2014 年的 321 个客户的用电量。原始数据每 15 min 记录一次，我们将数据转换为每小时的消耗量，并选取其中 256 个客户的数据作为最终数据集。

Traffic 数据集描述了旧金山湾区高速公路上不同传感器测量的道路占用率，包含加州交通部 48 个月（2015－2016 年）每小时收集的数据的集合，选取其中 256 列数据作为最终数据集。

图 1 和图 2 分别给出了这两个数据集中前 256 组数据的协方差矩阵的特征值。由图中可以看出大多数特征值几乎接近零，换句话说，其能量集中在几个相对较大的元素上。因此，本文提出的稀疏表示基具有较强的稀疏能力。

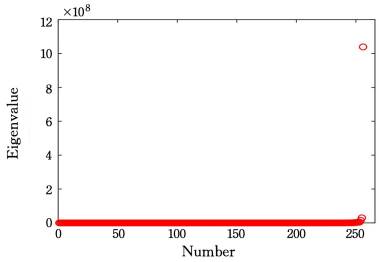

图 1 Electricity 中协方差矩阵的特征值

Fig. 1 Eigenvalues of covariance matrix in Electricity

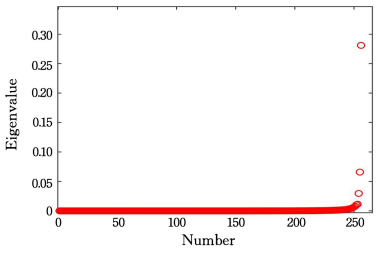

图 2 Traffic 中协方差矩阵的特征值

Fig. 2 Eigenvalues of covariance matrix in Traffic

假设协方差矩阵的特征值主要集中在 d 个特征值，其余的特征值的大小趋于零，并且有 $\lambda_1>\lambda_2>\cdots>\lambda_d\cdots>\lambda_N$。信号 $\boldsymbol{\theta}=[\theta_1,\theta_1,\cdots,\theta_N]^{\mathrm{T}}$ 的估计值为 $\hat{\boldsymbol{\theta}}=[\theta_1,\theta_1,\cdots\theta_d,0,\cdots,0]^{\mathrm{T}}$，则估计的误差如下：

$$\begin{aligned}E(\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|)&=E(\sum_{i=d+1}^{N}\boldsymbol{\theta}_i^2)\\&=E(\sum_{i=d+1}^{N}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{U}_i)\\&=\sum_{i=d+1}^{N}(\boldsymbol{U}_i^{\mathrm{T}}E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\boldsymbol{U}_i)\\&=\sum_{i=d+1}^{N}(\boldsymbol{U}_i^{\mathrm{T}}\lambda_i\boldsymbol{U}_i)\\&=\sum_{i=d+1}^{N}\lambda_i\to 0\end{aligned} \tag{7}$$

特征值的数值主要集中在 d 个元素中，而 d 远小于 N。因此，信号 $\boldsymbol{\theta}$ 是稀疏表示基 $\boldsymbol{U}$ 转换下的稀疏信号。

3.2 测量矩阵设计

缺失数据集存在部分采样数据的缺失，对于此类数据集的填补问题，实质上就是利用现存数据恢复出原始的数据集。针对缺失数据集的填补问题，我们设计了专门的测量矩阵，并根据测量值恢复出原始数据。

$$\boldsymbol{\Phi}=\begin{bmatrix}1&0&0&0&\cdots\\0&0&1&0&\cdots\\0&0&0&0&\cdots\\\cdots&\cdots&\cdots&\cdots&\cdots\end{bmatrix}$$

这里测量矩阵的作用是将未缺失的数据从原始数据集中筛选出来。对于一个数据总量为 N、未缺失数据个数为 M 的数据集，其测量矩阵 $\boldsymbol{\Phi}$ 为一个 $M\times N$ 维的矩阵。未缺失数据中的第 m 个值对应着原始数据集中的第 n 个值，因此它的相对位置信息可以用 (m,n) 表示。在测量矩阵 $\boldsymbol{\Phi}$ 中，将每个未缺失数据的相对位置信息，即 (m,n) 处的值设置为 1，矩阵中其余的值均设置为 0。

3.3 相关性分析

要判定测量矩阵是否满足 RIP 条件是一个相当困难的问题。因此，为了衡量 CS 恢复阶段的有效性，文献[20]中提出，稀疏表示基与测量矩阵之间只要满足相关性足够小的条件，稀疏信号就能够较好地完成重构。为了进一步降低计算的复杂度，可以通过计算两个矩阵的非相关性来间接衡量其相关性性质。

首先将 $\boldsymbol{\Phi}$ 的每一行投影到由 $\boldsymbol{\Psi}$ 的所有列生成的空间中。在此之后，我们将在该空间中获得的最稀疏投影作为衡量非相关性的指标。形式上，投影过程可以表示为：

$$\boldsymbol{\zeta}_j=(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\varphi}_j^{\mathrm{T}} \tag{8}$$

其中，$\boldsymbol{\varphi}_j$ 代表测量矩阵 $\boldsymbol{\Phi}$ 的第 j 行，$\boldsymbol{\zeta}_j$ 是 $\boldsymbol{\varphi}_j$ 在 $\boldsymbol{\Psi}$ 各列生成的空间上的投影向量，因此可以得到非相关性的一个度量值：

$$I(\boldsymbol{\Phi},\boldsymbol{\Psi})=\min_{j=1,\cdots,N}\|\boldsymbol{\zeta}_j\|_0 \tag{9}$$

$I(\boldsymbol{\Phi},\boldsymbol{\Psi})$ 越大，表明稀疏表示基与测量矩阵的非相关性越强，即相关性越小。但是，该方法以最值作为标准，获得的非相关性特征并不全面。文献[21]提出了一种平均互相关性的定义，能够很好地弥补这方面的问题。对于 $\boldsymbol{A}=\boldsymbol{\Phi}\boldsymbol{\Psi}$，令 $\boldsymbol{G}=\tilde{\boldsymbol{A}}^{\mathrm{T}}\tilde{\boldsymbol{A}}$，其中 $\tilde{\boldsymbol{A}}$ 为 $\boldsymbol{A}$ 的标准化矩阵，那么其平均互相关 $\mu_t(\boldsymbol{A})$ 可以定义为：

$$\mu_t(\boldsymbol{A})=\frac{\sum_{1\leqslant m,n\leqslant N,m\neq n}(|g_{mn}|\geqslant t)\cdot|g_{mn}|}{\sum_{1\leqslant m,n\leqslant N,m\neq n}(|g_{mn}|\geqslant t)} \tag{10}$$

其中，g_{mn} 为矩阵 $\boldsymbol{G}$ 的第 m 行第 n 列元素。

$$(|g_{mn}|\geqslant t)=\begin{cases}1, & \text{if } |g_{mn}|\geqslant t\\0, & \text{else}\end{cases} \tag{11}$$

其中，t 为阈值。常见的数据缺失问题主要为随机缺失和均匀缺失，为了全面地了解测量矩阵和稀疏表示基的相关性性质，分别对这两种情况进行考察。

取 $t=0.1$，在不同测量值下，稀疏表示基与测量矩阵之间的平均互相关变化规律如表 1、表 2 所列。

表 1　数据随机丢失下，稀疏表示基与测量矩阵之间的平均互相关

Table 1　Average cross-correlation when data is randomly missing

M	40	80	120	160
OEB	0.1946	0.1556	0.1382	0.1293
DCT	0.1856	0.1459	0.1327	0.1155
DFT	0.1712	0.1329	0.1249	0.1016
DWT	0.4755	0.3758	0.3147	0.2762

表 2　数据均匀丢失下，稀疏表示基与测量矩阵之间的平均互相关

Table 2　Average cross-correlation when data is evenly missing

M	40	80	120	160
OEB	0.1932	0.1553	0.1388	0.1287
DCT	0.2472	0.2349	0.2129	0.1899
DFT	0.3914	0.3608	0.3405	0.3131
DWT	0.4548	0.3699	0.2998	0.2549

根据表 1 和表 2 的数据可知，4 种稀疏表示基的平均相关大小有明显的规律。在数据随机丢失的情境下，正交特征向量基仅略差于 DCT 和 DFT 稀疏表示基。但是在数据均匀丢失的情况下，该稀疏表示基相比其他几种经典的稀疏表示基有很大的优势。综合以上分析，本文提出的稀疏表示基与测量矩阵具有较好的平均互相关性质，选取其作为稀疏表示基是合适的。

4　快速迭代加权阈值算法

文献[22]提出了一种基于同伦思想的迭代加权阈值方法。该算法在迭代部分采用迭代加权阈值方法，通过在迭代中同时对权重 $\boldsymbol{w}$ 和解 $\boldsymbol{x}$ 进行优化，从而获得了比常规的交替优化更好的恢复精度。本文借鉴了该算法在迭代部分同时优化的设计思路，将其融入 FISTA 算法中，并引入重启动策略，提出了一种快速迭代加权阈值算法。

4.1　迭代加权阈值方法

压缩感知的信号重构问题，可以通过求解欠定方程的最稀疏解来实现。也就是求解以下方程：

$$\min_{x\in R^n} \|\boldsymbol{x}\|_0 \quad \text{s.t.}\ \ \boldsymbol{y}=\boldsymbol{\Phi x} \tag{12}$$

但是，上述 l_0 范数最小化问题是一个 NP 难问题，因此不能在实际中直接进行求解[23]。为了解决这一问题，我们可以考虑将其转换为 l_1 范数最小化问题来进行求解，即求解以下问题：

$$\min_{x\in R^n} \|\boldsymbol{w}\circ\boldsymbol{x}\|_1 \quad \text{s.t.}\ \ \boldsymbol{y}=\boldsymbol{\Phi x} \tag{13}$$

其中，$\boldsymbol{w}\circ\boldsymbol{x}=(w_1x_1,w_2x_2,\cdots,w_nx_n)^{\mathrm{T}}$。

于是，问题可以重新表示为以下等效最小化问题：

$$\min_{x,w} \sum_{i=1}^{n} w_i(|x_i|-\varepsilon) \quad \text{s.t.}\ w_i\in\{0,1\} \tag{14}$$

其中 ε 是较小的正数。接下来，对该问题添加约束条件，则可以转化为：

$$\min_{x,w} F_{\lambda,\varepsilon}(\boldsymbol{x},\boldsymbol{\omega})=\min_{x,w} J_{\varepsilon}(\boldsymbol{x},\boldsymbol{\omega}) + \lambda f(\boldsymbol{x}) \tag{15}$$

其中，等式右边 $J_{\varepsilon}(\boldsymbol{x},\boldsymbol{\omega})=\sum_{i=1}^{n} w_i(|x_i|-\varepsilon)$，$f(\boldsymbol{x})=\|\boldsymbol{Ax}-\boldsymbol{y}\|_2^2/2$，根据近端梯度下降法，当满足 L-Lipschitz 条件时，存在 Lipschitz 常数 $L>0$，使得 $\|\nabla f(\boldsymbol{x}')-\nabla f(\boldsymbol{x})\|\leqslant L\|\boldsymbol{x}'-\boldsymbol{x}\|$。那么，在 $\boldsymbol{x}'$ 处，$f(\boldsymbol{x})$ 可以展开为：

$$\begin{aligned} g_L(\boldsymbol{x}',\boldsymbol{x}) &= f(\boldsymbol{x}')+\langle\nabla f(\boldsymbol{x}'),\boldsymbol{x}-\boldsymbol{x}'\rangle+\frac{L}{2}\|\boldsymbol{x}-\boldsymbol{x}'\|^2 \\ &= \frac{L}{2}\left\|\boldsymbol{x}-\left(\boldsymbol{x}'-\frac{1}{L}\nabla f(\boldsymbol{x}')\right)\right\|^2+Const \end{aligned} \tag{16}$$

综上，原式的分量形式可以表示为：

$$\sum_{i=1}^{n}\min_{x_i,\omega_i} w_i(|x_i|-\varepsilon)+\frac{\lambda L}{2}(x_i-z_i)^2 \tag{17}$$

其中，$\boldsymbol{z}=\boldsymbol{x}'-\frac{1}{L}\nabla f(\boldsymbol{x}')=\boldsymbol{x}'-\frac{1}{L}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{Ax}'-\boldsymbol{y})$，由此我们可以得到这个问题的解为以下形式。

(1) 当 $\varepsilon\geqslant\frac{1}{2\lambda L}$

$$(x_i,\omega_i)=H_{L,\lambda,\varepsilon}(x_i')=\begin{cases}(z_i,0), & \text{if } |z_i|\geqslant\varepsilon+\frac{1}{2\lambda L}\\ (soft(z_i),1), & \text{if } |z_i|<\varepsilon+\frac{1}{2\lambda L}\end{cases} \tag{18}$$

其中，$soft(z)=\mathrm{sign}(\boldsymbol{z})\max(|\boldsymbol{z}|-\frac{1}{\mu L},0)$。

(b) 当 $\varepsilon<\frac{1}{2\lambda L}$

$$(x_i,\omega_i)=H_{L,\lambda,\varepsilon}(x_i')=\begin{cases}(z_i,0), & \text{if}|z_i|\geqslant\sqrt{\frac{2\varepsilon}{\lambda L}}\\ (0,1), & \text{if}|z_i|<\sqrt{\frac{2\varepsilon}{\lambda L}}\end{cases} \tag{19}$$

4.2　重启动方案

文献[24]提出了一种重启动方案，并证明了使用该方案后算法将会加速收敛。即满足以下条件时，算法将会重新启动：

$$\nabla f(\boldsymbol{u}^k)^{\mathrm{T}}(\boldsymbol{x}^{k+1}-\boldsymbol{x}^k)>0 \tag{20}$$

在执行 FIWTA 算法的过程中，我们并不计算梯度，但可以将其视为广义梯度的形式。在该方案中，我们采用如下过程作为广义梯度迭代的过程：

$$\boldsymbol{x}^{k+1}=H_{L,\lambda,\varepsilon}(\boldsymbol{u}^k):=\boldsymbol{u}^k-sG(\boldsymbol{u}^k)^{\mathrm{T}} \tag{21}$$

其中，$G(\boldsymbol{u}^k)$ 是 $\boldsymbol{u}^k$ 处的广义梯度。因此，梯度重启方案相当于在满足以下情况时进行重启：

$$G(\boldsymbol{u}^k)^{\mathrm{T}}(\boldsymbol{x}^{k+1}-\boldsymbol{x}^k)>0 \tag{22}$$

即：

$$(\boldsymbol{u}^k-\boldsymbol{x}^{k+1})^{\mathrm{T}}(\boldsymbol{x}^{k+1}-\boldsymbol{x}^k)>0 \tag{23}$$

4.3　快速迭代重加权算法

FISTA 是一种经典的重构算法，主要是将加速度应用于 ISTA 算法中。该算法在 ISTA 算法的基础上加入了一个提升收敛速率的优化步骤，从而显著缩短了迭代所需时间。

本文提出的 FIWTA 算法以 FISTA 为框架，从精度和速度两个方面分别进行改进。为了改善 FISTA 算法的精度，本文用迭代加权阈值算法的迭代部分替换了原本 FISTA 框架中基于 ISTA 的迭代步骤，从而改善了算法的单次迭代效果。

为了加快算法的迭代速度，本文使用了自适应重启动技术。该技术立足于收敛函数在迭代过程中表现出的周期性震荡现象，并就震荡带来的迭代浪费问题进行了改进。当观察到目标函数值在震荡波段的底部时，我们就会重新启动算法，令 $t^{k+1}=1, \boldsymbol{u}^{k+1}=\boldsymbol{x}^{k+1}$。这样的操作使得 t^k 从 1 开始重新迭代，从而防止收敛函数震荡上升，进而大大加快算法的收敛速度，有效减少迭代次数。

快速迭代重加权算法的流程如算法 1 所示。

算法 1 FIWTA 算法

Input: $(\mathbf{A}, \mathbf{y}, \mathbf{x}^0, \lambda, \varepsilon, \zeta)$

Output: $(\mathbf{x}^{k+1})$

1. Initialization: $\mathbf{u}^0=\mathbf{x}^0, t^0=1, k=0$;
2. Repeat
3. $(\mathbf{x}^{k+1}, \omega^{k+1}) = H_{L,\lambda,\varepsilon}(\mathbf{u}^k)$;
4. $t^{k+1}=\dfrac{1+\sqrt{1+4t^k}}{2}$;
5. $\mathbf{u}^{k+1}=\mathbf{x}^{k+1}+\left(\dfrac{t^k-1}{t^{k+1}}\right)(\mathbf{x}^{k+1}-\mathbf{x}^k)$;
6. If $(\mathbf{u}^k-\mathbf{x}^{k+1})^T(\mathbf{x}^{k+1}-\mathbf{x}^k)>0$ then
7. $t^{k+1}=1$;
8. $\mathbf{u}^{k+1}=\mathbf{x}^{k+1}$;
9. end if
10. $k=k+1$;
11. Until $\|\mathbf{x}^{k+1}-\mathbf{x}^k\|<\zeta$.

如果去掉 FIWTA 中的重启动步骤，会得到一个迭代加权阈值算法(Iterative Weighted Thresholding Algorithm，IWTA)。IWTA 仅迭代部分与 FISTA 不同，其他步骤都相同。根据对式(18)和式(19)的分析，在单次迭代中，IWTA 的收敛速度在最坏情况下与 FISTA 相同，因此 IWTA 的收敛速度会比 FISTA 更快。文献[12]已经证明 FISTA 的收敛速度为 $O(1/k^2)$，由于本文提出的 FIWTA 采用了重启动机制，这将极大地降低迭代的计算复杂度，算法的收敛速度也会得到进一步的提升，因此 FIWTA 的收敛速度最终会小于 $O(1/k^2)$。

图 3 是 IWTA，FISTA 和 FIWTA 算法在不同稀疏比下的迭代次数变化规律。可以看出，随着稀疏比的增加，IWTA 算法所需的迭代次数总是略少于 FISTA 算法，这是基础迭代的精度改进带来的效果。同时，由于引入了重启动机制，FISTA 算法与 FIWTA 算法在迭代次数上的差距更加明显，本文提出的算法始终在迭代次数上保持较大优势。

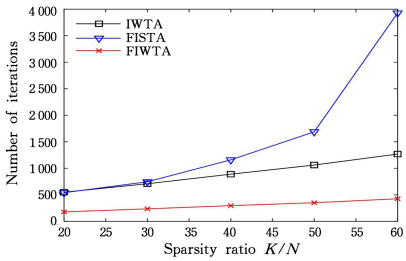

图 3 当 $N=400, M=200$ 时迭代次数与稀疏比的关系

Fig. 3 Number of iterations vs. sparsity ratio K/N when $N=400, M=200$

本文算法以 FISTA 为基本框架，采用同时优化权重 $\boldsymbol{w}$ 和解 $\boldsymbol{x}$ 的迭代设计思路，改善了每次迭代的精度。而重新启动机制的加入，大大加快了算法的收敛速度，从而实现了更好的收敛效果。

5 仿真结果及分析

5.1 重构成功率

本文通过与 ISTA，FISTA，Greedy FISTA[25]等算法进行比较来验证 FIWTA 算法的性能。构造一个稀疏度 $K=50$ 的 N 维原始数据($N=400$)，从中随机抽取 M 个元素，我们定义 M/N 的数值为测量比，即已知数据占所有数据的比例。在重构实验中，若原始数据 $\boldsymbol{x}$ 和估计数据 $\hat{\boldsymbol{x}}$ 满足 $\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2<10^{-6}$，则视为重构成功。在第一个实验中，我们固定稀疏度 K 的值，更改测量值 M，并观察算法在不同测量比 M/N 下的重构概率，如图 4 所示。

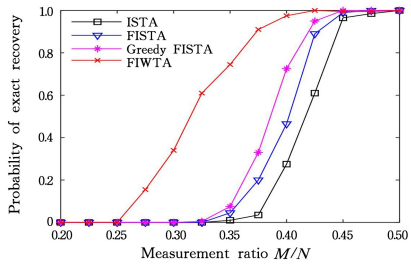

图 4 当 $N=400, K=50$ 时重构成功概率与观测比的关系

Fig. 4 Probability of exact recovery vs. measurement ratio M/N when $N=400, K=50$

由图 4 可知，算法的重构概率随测量比的增大而增大，但测量比对不同算法的影响不同。相比其他算法，FIWTA 算法能够在更低的测量比下完成对缺失数据集的重构。特别是在测量比小于 0.325 时，仅有 FIWTA 算法能够实现重构。因此在较低测量比时，算法具有更大的优势。这是由于 FIWTA 算法在迭代步骤中采用对权重 $\boldsymbol{w}$ 和解 $\boldsymbol{x}$ 同时优化的方法，提高了迭代过程中的算法精度。

与实验室不同，实践中数据的稀疏度往往是未知的，在部分情况下，稀疏矩阵并不能很好地完成数据的稀疏表示。因此，对缺失数据重构的性能评价，还需要比较在相同测量数的情况下，重构算法对不同稀疏度 K 的适应性。在这个实验中，我们固定测量数 M 的值，改变稀疏度 K 的大小，观察各种算法在不同的稀疏比 K/N 下的重构概率，如图 5 所示。

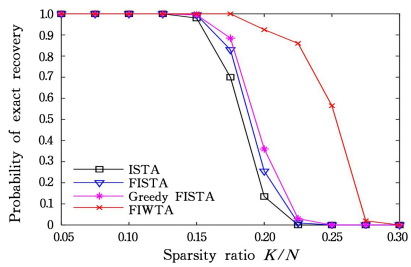

图 5 当 $N=400, M=200$ 时重构成功概率与稀疏比的关系

Fig. 5 Probability of exact recovery vs. sparsity ratio K/N when $N=400, M=200$

由图 5 可知，其他算法在稀疏比接近 0.2 时，重构概率先后降至零点，即不能完成重构。当稀疏比大于 0.25 时，仅有

FIWTA 算法能够完成缺失数据的重构。可见在固定测量数的情况下，FIWTA 算法对于稀疏度的宽容程度更高。在相同的测量数下，本文提出的算法能够在更大的稀疏度范围内完成缺失数据的重构，具有更好的鲁棒性。

5.2　重构时间

下面比较不同算法的计算复杂度。固定测量数 M 的值，计算各算法在不同的稀疏度条件下的平均重构时间，仿真结果如图 6 所示。

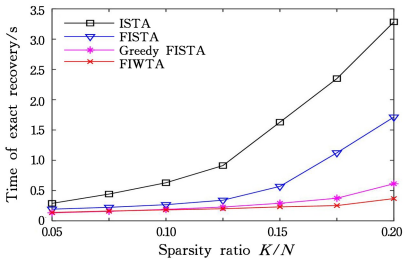

图 6　当 $N=400$，$M=200$ 重构时间与稀疏比的关系

Fig. 6　Time of exact recovery vs. sparsity ratio K/N when $N=400$, $M=200$

参考图 5 中的结论，其他算法在稀疏比较高时无法完成缺失数据集的重构。为了进行重构时间的比较，我们仅截取稀疏比小于 0.2 时的重构时间图像来进行分析。

如图 6 所示，FIWTA 算法在稀疏度逐渐增大的情况下，重构时间并没有呈现出陡峭的上升趋势，时间曲线总体较为平缓。这是由于 FIWTA 算法加入了重启动机制，减少了迭代次数，所需的重构时间极大缩短。

综上所述，FIWTA 算法具有更快的收敛速度和更好的重构精度。下面在真实数据集中对本文提出的数据填补体系进行验证，考察其实际填补效果。

5.3　实例测试

Traffic 数据集描述了旧金山湾区高速公路上不同传感器测量的道路占用率，包含加州交通部 48 个月（2015－2016 年）每小时收集的数据的集合，选取其中 256 列数据作为最终数据集。我们分别将 DCT，DWT，DFT 和 OEB 作为稀疏表示基，同时采用 $M\times N$ 维随机测量阵作为测量矩阵进行测试，数据维度 N 为 256，M 的取值范围是[100，180]。我们选取 256 个时刻的数据来构建正交特征向量基，重构算法采用本文提出的 FIWTA 算法，不同稀疏基下的均方根误差和平均运行时间随测量数的变化如图 7 和图 8 所示。

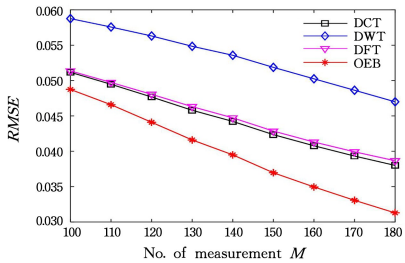

图 7　均方根误差与测量数 M 的关系

Fig. 7　RMSE vs. number of measurement M

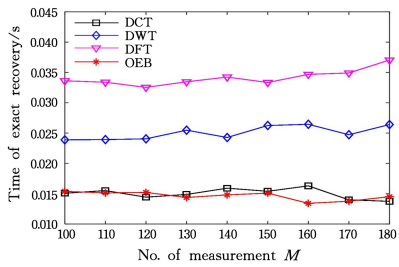

图 8　运行时间与测量数 M 的关系

Fig. 8　Runing time vs. number of measurement M

根据图 7 的变化趋势可知，在正交特征向量基的变换下，算法能够得到更为精准的重构值，在 4 种稀疏表示基中具有很大的优势。这是由于稀疏表示基在设计时，就已经很大程度上提取了数据集的一些先验信息。从图 8 可以看出，正交特征向量基所需的恢复时间在 4 种稀疏表示基中是最少的。实例证明，本文提出的数据填充方法效果出色。

结束语　本文提出了一种基于压缩感知的数据填补方法。针对相关性的数据集，构造了专门的正交特征向量基来进行稀疏化，将缺失数据填充问题转换为稀疏向量恢复问题。其次，本文提出了一种改进的迭代加权阈值算法。采用一种新的迭代加权方法，并引入重启动策略，对 FISTA 算法进行了改进。仿真结果表明，所提算法在具有更好鲁棒性的同时，也拥有更高的重构精度和较大的运行速度优势。本文算法虽然可以完成对多维数据的填补，但是需要根据部分已知数据来构造稀疏表示基。下一步将会针对没有先验数据的多维缺失数据填补问题进行研究。

参 考 文 献

[1]　QIU X G，CHEN B，ZHANG P. Emergency Management Oriented Artificial Society Construction and Computational Experiments[M]. Beijing：Science Press，2017：32-59.

[2]　HE M. Introduction to big data-big data thinking and innovative applications[M]. Beijing：Publishing House of Electronics Industry，2019：2-10.

[3]　LIN W C，TSAI C F. Missing value imputation：a review and analysis of the literature(2006－2017)[J]. Artificial Intelligence Review，2020，53(2)：1487-1509.

[4]　HUANG G L. Missing data filling method based on linear interpolation and lightgbm[J]. Journal of Physics：Conference Series，2021，1754(1)：012187.

[5]　SANJAR K，BEKHZOD O，KIM J，et al. Missing Data Imputation for Geolocation-based Price Prediction Using KNN-MCF Method[J]. ISPRS International Journal of Geo-Information，2020，9(4)：227.

[6]　PRIETO-CUBIDES J，ARGOTY C. Dealing with Missing Data using a Selection Algorithm on Rough Sets[J]. International Journal of Computational Intelligence Systems，2018，11(1)：1307-1321.

[7]　XIAO J Y，BULUT O. Evaluating the Performances of Missing Data Handling Methods in Ability Estimation From Sparse Data[J]. Educational and Psychological Measurement，2020，80(5)：932-954.

[8] KHALDY M A, KAMBHAMPATI C. Performance Analysis of Various Missing Value Imputation Methods on Heart Failure Dataset[C]// IntelliSys. Proceedings of SAI Intelligent Systems Conference. Berlin: Springer, 2016: 415-425.

[9] SAROJ A J, GUIN A, HUNTER M. Deep LSTM Recurrent Neural Networks for Arterial Traffic Volume Data Imputation [J]. Journal of Big Data Analytics in Transportation, 2021, 3(2): 95-108.

[10] 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.

[11] LU W Q, ZHOU T, LI L H, et al. An improved tucker decomposition-based imputation method for recovering lane-level missing values in traffic data[J]. IET Intelligent Transport Systems, 2022, 16(3): 363-379.

[12] BECK A, TEBOULLE M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems[J]. Siam J Imaging Sciences, 2009, 2(1): 183-202.

[13] PAN S, YAN K, LAN H, et al. Adaptive step-size fast iterative shrinkage-thresholding algorithm and sparse-spike deconvolution[J]. Computers & Geosciences, 2020, 134: 104343.

[14] CALATRONI L, CHAMBOLLE A. Backtracking strategies for accelerated descent methods with smooth composite objectives [J]. SIAM Journal on Optimization, 2017, 29(3): 1-25.

[15] ZHU T. Accelerating monotone fast iterative shrinkage-thresholding algorithm with sequential subspace optimization for sparse recovery[J]. Signal Image and Video Processing, 2020, 14(1): 1-10.

[16] KIM D, PARK D. Element-Wise Adaptive Thresholds for Learned Iterative Shrinkage Thresholding Algorithms[J]. IEEE Access, 2020, 4: 45874-45886.

[17] TONG C, TENG Y, YAO Y, et al. Eigenvalue-free iterative shrinkage-thresholding algorithm for solving the linear inverse problems[J]. Inverse Problems, 2021, 37(6): 5867-5877.

[18] CANDES E J, TAO T. Decoding by linear programming[J]. IEEE Transactions Information Theory, 2005, 51(12): 4203-4215.

[19] WU X, XIONG Y, YANG P, et al. Sparsest Random Scheduling for Compressive Data Gathering in Wireless Sensor Networks [J]. IEEE Transactions on Wireless Communications, 2014, 13(10): 5867-5877.

[20] QUER G, MASIERO R, MUNARETTO D, et al. On the interplay between routing and signal representation for Compressive Sensing in wireless sensor networks[C]// Information Theory & Applications Workshop. 2009: 206-215.

[21] ELAD M. Optimized projections for compressed sensing[J]. IEEE Transactions on Signal Processing, 2007, 55(12): 5695-5702.

[22] ZHU W X, HUANG Z L, CHEN J L, et al. Iteratively weighted thresholding homotopy method for the sparse solution of underdetermined linear equations[J]. Science China Mathematics, 2021, 64(3): 639-664.

[23] LI J J, JIANG Y, QIU T, et al. The Estimation Algorithm of OFDM Sparse Channel Based on Compressed Sensing[J]. Journal of Chongqing University of Technology (Natural Science), 2021, 35(4): 117-122.

[24] DONOGHUE B, CANDES E. Adaptive restart for accelerated gradient schemes[J]. Foundations of Computational Mathematics, 2015, 15(3): 715-732.

[25] YANG L, LI H, LI P, et al. Sparse Representation for SAR Ground Moving Target Imaging Based on Greedy FISTA[J]. Journal of Signal Processing, 2020, 35(11): 1844-1852.

REN Bing, born in 1993, postgraduate. His main research interests include compressed sensing and big data.

GUO Yan, born in 1971, Ph. D, professor. Her main research interests include unmanned intelligent system, compressed sensing and localization.

(责任编辑:何杨)