ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 5,Sep. 2017,pp. 939-947
DOI:10.16337/j.1004-9037.2017.05.011

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

基于结构化噪声矩阵补全的 WSNs 收集数据重建方法

陈正宇[1]　陈　蕾[2,3]　胡国兵[1]　戴　华[3]

(1.金陵科技学院电子信息工程学院，南京，211169；2.江苏省无线传感网高技术研究重点实验室，南京，210003；3.南京邮电大学计算机学院，南京，210003)

摘　要：许多科学研究都需要对环境数据进行分析，这些环境数据通常是通过部署在研究区域内的无线传感器网络(Wireless sensor networks，WSNs)来收集的。收集数据的完整性和准确性决定了科研结果的可靠性。然而，在数据收集过程中普遍存在的数据丢失和错误影响了收集数据的可用性，为此需要利用收集到的数据重建完整的环境数据。基于环境数据低秩特性，将数据重建问题建模为 $L2,1$ 范数正则化矩阵补全模型，提出一种基于结构化噪声矩阵补全的 WSNs 收集数据重建方法(Data reconstruction approach via matrix completion with structural noise，DRMCSN)。真实数据集上的实验结果表明，该方法性能优于现有算法，不仅能以较高的精度恢复缺失的环境数据，而且能辨识出收集到错误数据的传感器节点。

关键词：无线传感器网络；数据收集；矩阵补全；数据重建

中图分类号：TP311　　**文献标志码**：A

Data Reconstruction in WSNs via Matrix Completion with Structural Noise

Chen Zhengyu[1], Chen Lei[2,3], Hu Guobing[1], Dai Hua[3]

(1. School of Electronic and Information Engineering, Jinling Institute of Technology, Nanjing, 211169, China; 2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, 210003, China; 3. College of Computer Science & Technology, Nanjing University of Posts & Telecommunications, Nanjing, 210003, China)

Abstract: Many scientific work needs to analyze the environmental data which are usually collected by wireless sensor networks(WSNs)deployed in research areas. The integrity and accuracy of the collected data determine the reliability of the research results. However, data loss and error usually occur during the process of data collection, which affect the availability of collected data. Therefore, it is necessary to reconstruct the environmental data from the incomplete and erroneous sensory data. Based on the low-rank feature of the environmental data, an efficient data reconstruction approach via matrix completion with structural noise (DRMCSN) is proposed by formulating data reconstruction problem as a $L2,1$-norm regularized matrix completion model. Finally, experimental results on a real dataset demonstrate that the proposed approach can not only effectively reconstruct the environmental data, but also recognize the sen-

基金项目：江苏省自然科学基金(BK20130096，BK20161516，BK20161104)资助项目；国家自然科学基金(61300240，61572263)资助项目；江苏省高校自然科学基金(15KJB520027)资助项目；中国博士后科学基金(2015M581794)资助项目；江苏省博士后科研资助计划(1501023C)资助项目；安徽省自然科学基金(1608085MF127)资助项目；金陵科技学院高层次人才工作启动(JIT-201527)资助项目。

收稿日期：2017-07-26；**修订日期**：2017-09-21

sor nodes that collect erroneous data.

Key words: wireless sensor networks; data collection; matrix completion; data reconstruction

引 言

无线传感器网络(Wireless sensor networks, WSNs)在环境资源监测、节点定位、基础设施监测和生物习性监测等科学研究场景中得到了广泛的应用[1,2],而数据收集通常是实现这些应用的基础环节。但由于受到传感器节点硬件特性、资源条件、无线通信和部署环境等因素的制约,在数据收集过程中通常会出现数据丢失和数据错误等问题[3,4]。例如,在 Intel 室内实验项目中,研究人员通过 3 个星期的观察发现有近 40%的数据丢失和 8%的数据错误[4]。数据丢失和错误给相关应用的准确性和可靠性带来了巨大的挑战[5]。因此,利用 WSNs 收集到的含有错误元素的不完整数据集来重建原始环境数据具有十分重要的意义。

近年来,矩阵补全技术越来越受到研究者的普遍关注。矩阵补全技术能通过存在元素缺失的不完整数据矩阵重构出原始的完整低秩矩阵,目前,已在推荐系统、图像处理、标记分类、网络监测和 WSNs 等领域中得到广泛的应用[6]。在 WSNs 领域中,数据采集、节点定位等应用场景已利用矩阵补全技术来恢复缺失的数据信息。在数据收集应用场景中,为了减少 WSNs 数据收集能耗,文献[7]提出一种有效数据收集方法(Efficient data collection approach, EDCA),该方法只收集网络中部分节点采集的环境数据,从而得到一个不完整的环境数据矩阵,然后利用环境数据矩阵的低秩特性将数据恢复问题建模成缺失数据矩阵的补全问题,并设计优化算法解决该问题。由于该方法仅利用矩阵的低秩特性,因此在数据丢失严重的情况下,重建误差仍然较大。文献[8]利用收集数据矩阵的低秩和瞬时稳定性特征,提出空时压缩数据收集方法(Spatiotemporal compressive data collection, STCDC),瞬时稳定性在求解矩阵补全问题的最优解时提供稳定性约束,从而降低了环境数据的重建误差。文献[9]分析了 WSNs 收集数据丢失模式,并利用收集数据的低秩、时间和空间相关性等特征,提出基于时空提升的环境数据重建方法(Environmental space-time improved compressive sensing, ESTI-CS)算法,结合秩最小估计以及时空特征恢复丢失数据。文献[10]提出联合矩阵补全和稀疏约束的数据恢复方法(Data recover method with joint matrix completion and sparsity constraints, DRMCSC),利用数据的稀疏特征和矩阵的低秩特性,将稀疏约束和矩阵补全结合在同一优化问题中,并设计交替最小化算法实现数据重建。以上算法仅考虑恢复丢失数据的问题,均没有考虑数据错误问题以及其对数据恢复精度的影响。而已有的研究通常利用数据相关性和统计特性来检测错误数据,比如文献[11,12]分别提出基于直方图和基于收集数据统计特性的 Outlier 数据检测方法;文献[13]提出利用序列检测方法识别传感器网络中的错误数据;文献[14]研究异常收集数据和链路中断的存在对数据收集精确度的影响,并基于压缩感知理论识别和纠正异常收集数据。然而,上述错误数据检测算法都没有考虑数据缺失对于错误数据检测性能的影响。

对于不完整低秩数据矩阵的重建问题,文献[15]提出当不完整的低秩数据矩阵中的某些行受到损坏时,传统矩阵补全算法恢复性能往往显得不够稳定,为此将不完整矩阵中受到损坏的行看作是数据行受到结构化噪声的污染,并将存在这类噪声情形的矩阵补全问题称为结构化噪声矩阵补全问题。文献[16]将结构化噪声矩阵补全问题应用于 Web 服务的 QoS 预测中。受到文献[15,16]的启发,本文针对 WSNs 数据采集过程中同时存在的数据丢失和数据错误等问题,提出一种基于结构化噪声矩阵补全的 WSNs 收集数据重建方法(Data reconstruction via matrix completion with structural noise, DRMCSN)。该方法将错误数据建模为收集数据受到行结构化噪声的污染,利用环境数据矩阵的低秩特性,将含有错误收集数据情况下的环境数据重建问题建模为结构化噪声矩阵补全模型,并利用文献[15]提出的矩阵空间交替线性分裂 Bregman 迭代算法实现该问题的求解。仿真实验表明,DRMCSN 能以较高的精度恢复缺失数据,同时能辨识出收集到错误数据的传感器节点。

1 预备知识

1.1 基本定义

假设矩阵 $\boldsymbol{X}\in\mathbf{R}^{n_1\times n_2}$ 的奇异值分解为 $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$,其中,$\boldsymbol{\Sigma}=\mathrm{diag}\{\sigma_i\mid 1\leqslant i\leqslant \min(n_1,n_2)$ 且 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_{\min(n_1,n_2)}>0\}$,则有如下定义[15]

(1)矩阵 $\boldsymbol{X}$ 的 Frobenius 范数:$\|\boldsymbol{X}\|_{\mathrm{F}}=\sqrt{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}|X_{ij}|^2}$;(2)矩阵 $\boldsymbol{X}$ 的核范数:$\|\boldsymbol{X}\|_*=\sum_{i=1}^{\min(n_1,n_2)}|\sigma_i|$;(3)矩阵 $\boldsymbol{X}$ 的 $L2,1$ 范数:$\|\boldsymbol{X}\|_{2,1}=\sum_{i=1}^{n_1}(\sum_{j=1}^{n_2}X_{ij}^2)^{1/2}$;(4)对任意 $\gamma\geqslant 0$,则其对应的奇异值阈值算子为 $D_\gamma(\boldsymbol{X})=\boldsymbol{U}S_\gamma(\boldsymbol{\Sigma})\boldsymbol{V}^{\mathrm{T}}$。式中:$S_\gamma(\boldsymbol{\Sigma})=\mathrm{diag}\{\max(0,\sigma_i-\gamma)\mid i=1,2,\cdots,\min(n_1,n_2)\}$[17]。

1.2 相关定理

定理 1[17] 对任意 $\tau,\mu>0$,$\boldsymbol{Z}\in\mathbf{R}^{n_1\times n_2}$,有 $D_{\frac{\tau}{\mu}}(\boldsymbol{Z})=\arg\min_{\boldsymbol{X}\in\mathbf{R}^{n_1\times n_2}}\left\{\tau\|\boldsymbol{X}\|_*+\frac{\mu}{2}\|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}}^2\right\}$。

定理 2[18] 近邻前向后向分裂(Proximal forward backward splitting, PFBS)。对于无约束优化问题

$$\min_{\boldsymbol{X}\in\mathbf{R}^{n_1\times n_2}}J(\boldsymbol{X})+H(\boldsymbol{X}) \tag{1}$$

假定 $J(\boldsymbol{X})$ 是 $\mathbf{R}^{n_1\times n_2}$ 上具有下半连续性质的凸函数,$H(\boldsymbol{X})$ 是 $\mathbf{R}^{n_1\times n_2}$ 上凸光滑下半连续函数且具有 β-Lipschitz 连续导数,即 $\|\nabla H(\boldsymbol{X}_1)-\nabla H(\boldsymbol{X}_2)\|_{\mathrm{F}}\leqslant\beta\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\mathrm{F}}$,$\forall\boldsymbol{X}_1,\boldsymbol{X}_2\in\mathbf{R}^{n_1\times n_2}$,则对任意的初始值 $\boldsymbol{X}^0$ 及 $0<\delta<\frac{2}{\beta}$,用如式(2)生成的迭代序列 $\boldsymbol{X}^{k+1}$ 收敛到无约束优化问题(1)的唯一解为

$$\boldsymbol{X}^{k+1}=\mathrm{prox}_{\delta J(\boldsymbol{X})}(\boldsymbol{X}^k-\delta\nabla H(\boldsymbol{X}^k))=\arg\min_{\boldsymbol{X}\in\mathbf{R}^{n_1\times n_2}}\left\{\delta J(\boldsymbol{X})+\frac{1}{2}\|\boldsymbol{X}-(\boldsymbol{X}^k-\delta\nabla H(\boldsymbol{X}^k))\|_{\mathrm{F}}^2\right\} \tag{2}$$

定理 3[16] 对任意 $\tau,\mu>0$,$\boldsymbol{W}\in\mathbf{R}^{n_1\times n_2}$,函数 $H(\boldsymbol{X})=\tau\|\boldsymbol{X}\|_{2,1}+\frac{\mu}{2}\|\boldsymbol{X}-\boldsymbol{W}\|_{\mathrm{F}}^2$ 存在全局最小点 $\boldsymbol{X}^*$:$(\boldsymbol{X}^*)^{(i)}=\max\left\{\|W^{(i)}\|_2-\frac{\tau}{\mu},0\right\}\cdot\frac{\boldsymbol{W}^{(i)}}{\|\boldsymbol{W}^{(i)}\|_2}$,$i=1,2,\cdots,n_1$,其中,$\boldsymbol{X}^{(i)}$ 表示矩阵 $\boldsymbol{X}$ 的第 i 行,$\|\cdot\|_2$ 表示向量 $L2$ 范数。

定理 1,2 和 3 的证明可分别详见文献[16-18]。

2 系统模型与问题描述

2.1 系统模型

设监测区域内部署 N 个传感器节点 $v_1,v_2,\cdots,v_N$,周期性地收集环境数据。将每轮收集时间间隔称为一个时隙。设收集总时间为 T 个时隙,则对于某一类环境数据,总数据量为 $N\times T$。该环境数据可用矩阵 $\boldsymbol{X}$ 表示为

$$\boldsymbol{X}=\begin{bmatrix}X(1,1) & X(1,2) & X(1,3) & \cdots & X(1,T)\\ X(2,1) & X(2,2) & X(2,3) & \cdots & X(2,T)\\ \vdots & \vdots & \vdots & & \vdots\\ X(N,1) & X(N,2) & X(N,3) & \cdots & X(N,T)\end{bmatrix}\in\mathbf{R}^{N\times T} \tag{3}$$

式中:$X(i,j)$ 表示节点 v_i 对应于时隙 j 的原始环境数据。然而,由于收集过程中存在数据丢失,Sink 节点得到的是一个有很多元素丢失的不完整矩阵,也称为采样矩阵,用 $\boldsymbol{S}$ 表示;将收集到的数据称为采样数据;将采样数据占总数据量的比例称为数据采样率。

定义 $\Omega\subseteq[N]\times[T]$($[N]=\{1,\cdots,N\}$,$T=\{1,\cdots,T\}$)为采样数据在采样矩阵中的下标索引集合。$P_\Omega(\cdot)$ 为正交投影算子,表示当 $(i,j)\in\Omega$ 时,$S(i,j)$ 为采样元素,即有

$$[P_{\Omega}(\boldsymbol{S})]_{ij}=\begin{cases}S(i,j) & (i,j)\in\Omega\\ 0 & \text{其他}\end{cases} \tag{4}$$

由于存在数据错误,采样数据可能有两种情况,即原始环境数据 $X(i,j)$和错误数据 $F(i,j)$。采样数据 $S(i,j)$可表示为

$$S(i,j)=\begin{cases}X(i,j) & \text{节点 } v_i \text{ 在时隙 } j \text{ 收集到的原始环境数据}\\ F(i,j) & \text{节点 } v_i \text{ 在时隙 } j \text{ 收集到的错误数据}\end{cases} \tag{5}$$

错误数据 $F(i,j)$可表示为原始环境数据与噪声值的叠加,即

$$F(i,j)=X(i,j)+Z(i,j)$$

式中:$Z(i,j)$为噪声值。将收集到的错误数据的传感器节点称为故障节点,将故障节点所占的比例称为节点故障率。在实际应用中,某些传感器节点容易成为故障节点,这些节点在采样矩阵中所对应的数据行含有错误元素。对于这类行元素的错误问题,可视为采样矩阵受到行形式的结构化噪声的污染,进一步可将采样矩阵表示为

$$P_{\Omega}(\boldsymbol{S})=P_{\Omega}(\boldsymbol{X}+\boldsymbol{Z}) \tag{6}$$

式中:$\boldsymbol{Z}=(Z(i,j))_{N\times T}$为行形式的结构化噪声矩阵。在矩阵 $\boldsymbol{Z}$ 中,如果节点 v_i 在时隙 j 收集到错误数据,则 $Z(i,j)\neq 0$,否则 $Z(i,j)=0$。

2.2 问题描述

WSNs 收集数据的重建问题就是利用 Sink 节点收集到的采样矩阵 $\boldsymbol{S}$ 来重建原始环境数据矩阵 $\boldsymbol{X}$。利用环境数据矩阵的低秩特性,可以将数据重建问题建模为矩阵补全问题[7,8]。在求解矩阵补全问题时,为了有效地平滑结构化噪声,将噪声矩阵 $\boldsymbol{Z}$ 的 $L2,1$ 范数正则化项引入到标准矩阵补全问题中[16],从而将含有错误数据的无线传感器网络收集数据重建问题建模为基于 $L2,1$ 范数正则化的结构化噪声矩阵补全模型,即有

$$\min_{\boldsymbol{X},\boldsymbol{Z}\in\mathbf{R}^{n\times n_t}}\|\boldsymbol{X}\|_*+\lambda\|\boldsymbol{Z}\|_{2,1}\quad \text{s.t.}\quad P_{\Omega}(\boldsymbol{S})=P_{\Omega}(\boldsymbol{X}+\boldsymbol{Z}) \tag{7}$$

式中:$\boldsymbol{S}$ 为采样矩阵,$\boldsymbol{X}$ 为优化矩阵,$\boldsymbol{Z}$ 为结构化噪声矩阵,λ 为一个用来平衡结构化噪声和矩阵低秩程度的可调参数。求解矩阵补全问题(7)得到的最优解 $\boldsymbol{X}_{\text{opt}}$和 $\boldsymbol{Z}_{\text{opt}}$可以用来重建环境数据矩阵 $\boldsymbol{X}_{\text{rec}}$。

3 基于结构化噪声矩阵补全的数据重建方法

基于第 2 节提出的系统模型,本节将给出 DRMCSN 算法的具体实现方法。首先,将问题(7)松弛为无约束优化问题,即

$$\min_{\boldsymbol{X},\boldsymbol{Z}\in\mathbf{R}^{N\times T}}\tau(\|\boldsymbol{X}\|_*+\lambda\|\boldsymbol{Z}\|_{2,1})+\frac{1}{2}\|P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}-\boldsymbol{Z}\|)_{\text{F}}^2 \tag{8}$$

问题(8)不存在解析解。基于文献[15]提出的矩阵空间交替线性分裂 Bregman 迭代算法,可将问题(8)转化为求解 2 个子问题,即

$$\text{子问题 1:}\begin{cases}\boldsymbol{X}^{k+1}=\arg\min_{X\in\mathbf{R}^{N\times T}}\tau\|\boldsymbol{X}\|_*-\tau\langle\boldsymbol{G}_{\boldsymbol{X}}^k,\boldsymbol{X}\rangle+\frac{1}{2}\|P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}-\boldsymbol{Z}^k)\|_{\text{F}}^2\\ \boldsymbol{G}_{\boldsymbol{X}}^{k+1}=\boldsymbol{G}_{\boldsymbol{X}}^k+\frac{1}{\tau}P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^{k+1})\end{cases}$$

其中,$\boldsymbol{G}_{\boldsymbol{X}}^k$ 为次微分$\partial\|\boldsymbol{X}^k\|_*$的一个次梯度,$\langle\cdot,\cdot\rangle$表示矩阵的内积运算。

$$\text{子问题 2:}\begin{cases}\boldsymbol{Z}^{k+1}=\arg\min_{\boldsymbol{Z}\in\mathbf{R}^{N\times T}}\tau\lambda\|\boldsymbol{Z}\|_{2,1}-\tau\lambda\langle\boldsymbol{G}_{\boldsymbol{Z}}^k,\boldsymbol{Z}\rangle+\frac{1}{2}\|P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}\|_{\text{F}}^2\\ \boldsymbol{G}_{\boldsymbol{Z}}^{k+1}=\boldsymbol{G}_{\boldsymbol{Z}}^k+\frac{1}{\tau\lambda}P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^{k+1})\end{cases}$$

其中,$\boldsymbol{G}_{\boldsymbol{Z}}^k$ 为次微分$\partial\|\boldsymbol{Z}^k\|_{2,1}$的一个次梯度。

(1)求解子问题 1

在子问题 1 中,令 $J(\boldsymbol{X})=\tau\|\boldsymbol{X}\|_*-\tau\langle\boldsymbol{G}_{\boldsymbol{X}}^k,\boldsymbol{X}\rangle$,$H(\boldsymbol{X})=\frac{1}{2}\|P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}-\boldsymbol{Z}^k)\|_{\mathrm{F}}^2$,易知函数 $J(\boldsymbol{X})$和 $H(\boldsymbol{X})$满足定理 2 的基本条件,因此子问题 1 存在唯一解,依据 PFBS 算法,按式(9)方法迭代生成序列 $\boldsymbol{X}^{k+1}$收敛到该唯一解,即

$$\boldsymbol{X}^{k+1}=\mathrm{prox}_{\delta_X J(\boldsymbol{X})}(\boldsymbol{X}^k-\delta_X\nabla H(\boldsymbol{X}^k))=$$
$$\arg\min_{\boldsymbol{X}\in\mathbf{R}^{N\times T}}\left\{\delta_X\tau\|\boldsymbol{X}\|_*+\frac{1}{2}\|\boldsymbol{X}-(\boldsymbol{X}^k+\varpi\delta_X\boldsymbol{G}_{\boldsymbol{X}}^k+\delta_X P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^k-\boldsymbol{Z}^k))\|_{\mathrm{F}}^2\right\}\tag{9}$$

且应有 $\boldsymbol{G}_{\boldsymbol{X}}^{k+1}=\boldsymbol{G}_{\boldsymbol{X}}^k-\frac{1}{\tau\delta_{\boldsymbol{X}}}(\boldsymbol{X}^{k+1}-\boldsymbol{X}^k-\delta_X P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^k-\boldsymbol{Z}^k))$,类似于文献[15]的推导过程,并令 $\boldsymbol{V}^k=\boldsymbol{V}^{k-1}+\delta_X P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^k-\boldsymbol{Z}^k)$,则式(9)可简化为

$$\boldsymbol{X}^{k+1}=\arg\min_{\boldsymbol{X}\in\mathbf{R}^{N\times T}}\left\{\delta_X\tau\|\boldsymbol{X}\|_*+\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{V}^k\|_{\mathrm{F}}^2\right\}\tag{10}$$

根据定理 1 可得

$$\boldsymbol{X}^{k+1}=D_{\varpi\delta_X}(\boldsymbol{V}^k)\tag{11}$$

因此,子问题 1 可按式(12)迭代求解

$$\begin{cases}\boldsymbol{V}^k=\boldsymbol{V}^{k-1}+\delta_X P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^k-\boldsymbol{Z}^k)\\ \boldsymbol{X}^{k+1}=D_{\varpi\delta_X}(\boldsymbol{V}^k)\end{cases}\tag{12}$$

式中参数 δ_X 可作如式(13)的估算,即

$$\|\nabla H(\boldsymbol{X}_1)-\nabla H(\boldsymbol{X}_2)\|_{\mathrm{F}}=\|P_{\Omega}(\boldsymbol{X}_1+\boldsymbol{Z}^k-\boldsymbol{S})-P_{\Omega}(\boldsymbol{X}_2+\boldsymbol{Z}^k-\boldsymbol{S})\|_{\mathrm{F}}=$$
$$\|P_{\Omega}(\boldsymbol{X}_1-\boldsymbol{X}_2)\|_{\mathrm{F}}\leqslant 1\cdot\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\mathrm{F}}\tag{13}$$

令函数$\nabla H(\boldsymbol{X})$的 Lipschitz 常数 β 为 1,进一步,根据定理 2,可将参数 δ_X 设置为$\frac{1}{\beta}=1$。

(2)求解子问题 2

类似于子问题 1 的求解过程,可得

$$\boldsymbol{Z}^{k+1}=\arg\min_{\boldsymbol{Z}\in\mathbf{R}^{N\times T}}\left\{\tau\lambda\delta_Z\|\boldsymbol{Z}\|_{2,1}+\frac{1}{2}\|\boldsymbol{Z}-(\boldsymbol{Z}^k+\lambda\tau\delta_Z\boldsymbol{G}_{\boldsymbol{Z}}^k+\delta_Z P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^k)\|_{\mathrm{F}}^2)\right\}\tag{14}$$

式中,$\boldsymbol{G}_{\boldsymbol{Z}}^{k+1}=\boldsymbol{G}_{\boldsymbol{Z}}^k-\frac{1}{\tau\lambda\delta_Z}(\boldsymbol{Z}^{k+1}-\boldsymbol{Z}^k-\delta_Z P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^k))$。类似于子问题 1 的方法,参数 δ_Z 取值为 1。参考文献[15]的推导步骤,并令 $\boldsymbol{W}^k=\boldsymbol{W}^{k-1}+\delta_Z P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^k)$,则

$$\boldsymbol{Z}^{k+1}=\arg\min_{\boldsymbol{Z}\in\mathbf{R}^{N\times T}}\left\{\tau\lambda\delta_Z\|\boldsymbol{Z}\|_{2,1}+\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{W}^k\|_{\mathrm{F}}^2\right\}\tag{15}$$

根据定理 3 可得

$$(\boldsymbol{Z}^{k+1})^{(i)}=\max\{(\|\boldsymbol{W}^k)^{(i)}\|_2-\tau\lambda\delta_Z,0\}\frac{(\boldsymbol{W}^k)^{(i)}}{\|(\boldsymbol{W}^k)^{(i)}\|_2}\quad i=1,2,\cdots,N\tag{16}$$

因此,子问题 2 可求解为

$$\begin{cases}\boldsymbol{W}^k=\boldsymbol{W}^{k-1}+\delta_Z P_{\Omega}(\boldsymbol{R}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^k)\\ (\boldsymbol{Z}^{k+1})^{(i)}=\max\{\|(\boldsymbol{W}^k)^{(i)}\|_2-\tau\lambda\delta_Z,0\}\frac{(\boldsymbol{W}^k)^{(i)}}{\|(\boldsymbol{W}^k)^{(i)}\|_2}\quad i=1,2,\cdots,N\end{cases}\tag{17}$$

基于子问题 1 和子问题 2 的求解方法,在确定算法最大迭代次数等参数后,可以得到矩阵补全问题(7)的最优解,即恢复的数据矩阵 $\boldsymbol{X}_{\mathrm{opt}}$和恢复的噪声矩阵 $\boldsymbol{Z}_{\mathrm{opt}}$。利用矩阵 $\boldsymbol{X}_{\mathrm{opt}}$和 $\boldsymbol{Z}_{\mathrm{opt}}$可以重建环境矩阵 $\boldsymbol{X}_{\mathrm{rec}}$,具体方法包括以下两个步骤:

(1)用恢复的数据矩阵 $\boldsymbol{X}_{\mathrm{opt}}$中的对应元素 $X_{\mathrm{opt}}(i,j)$来填充采样矩阵 $\boldsymbol{S}$ 中丢失的元素,即重建环境矩阵 $\boldsymbol{X}_{\mathrm{rec}}$满足

$$X_{\mathrm{rec}}(i,j)=\begin{cases}S(i,j) & (i,j)\in\Omega\\ X_{\mathrm{opt}}(i,j) & \text{其他}\end{cases}\tag{18}$$

(2)通过恢复的噪声矩阵 $\boldsymbol{Z}_{\text{opt}}$ 识别故障节点。在 $\boldsymbol{Z}_{\text{opt}}$ 中,含有非零元素的行所对应的传感器节点为故障传感器节点;所有元素为 0 的行所对应的传感器节点为正常传感器节点。在识别出故障器节点后,可将重建环境矩阵 $\boldsymbol{X}_{\text{rec}}$ 中含错误数据的行用恢复数据矩阵 $\boldsymbol{X}_{\text{opt}}$ 中所对应的行替代,即

$$\boldsymbol{X}_{\text{rec}}^{(i)}=\begin{cases}\boldsymbol{X}_{\text{opt}}^{(i)} & \boldsymbol{Z}_{\text{opt}}^{(i)}\ \text{含有非零元素}\\ \boldsymbol{X}_{\text{rec}}^{(i)} & \text{其他}\end{cases} \tag{19}$$

式中:$\boldsymbol{X}_{\text{rec}}^{(i)}$ 和 $\boldsymbol{X}_{\text{opt}}^{(i)}$ 分别表示矩阵 $\boldsymbol{X}_{\text{rec}}$ 和 $\boldsymbol{X}_{\text{opt}}$ 第 i 行数据。

综上,可以给出基于结构化噪声矩阵补全的数据重建方法(DRMCSN)如算法 1 所示,其中输入为无线传感器网络采样数据矩阵 $P_{\Omega}(\boldsymbol{S})$,最大迭代次数 Max 以及各类参数,如 λ,τ 等;输出为重建数据矩阵 $\boldsymbol{X}_{\text{rec}}$。算法首先设置初值矩阵为零矩阵(第 1 行),第 3~4 行求解子问题 1,5~8 行实现子问题 2 的求解,11~22 行实现环境数据的重建。

算法 1 基于结构化噪声矩阵补全的数据重建方法(DRMCSN)

输入:数据收集矩阵 $\boldsymbol{P}_{\Omega}(\boldsymbol{S})$,$\lambda,\tau,\delta_{\boldsymbol{X}}$ 和 $\delta_{\boldsymbol{Z}}$,最大迭代次数 Max;

输出:$\boldsymbol{X}_{\text{rec}}$;

初始化 $\boldsymbol{X}^{0}=0$, $\boldsymbol{Z}^{0}=0$, $\boldsymbol{V}^{-1}=0$,$\boldsymbol{W}^{-1}=0$;

FOR $k=0$ to MAX

$\boldsymbol{V}^{k}=\boldsymbol{V}^{k-1}+\delta_{\boldsymbol{X}}P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k}-\boldsymbol{Z}^{k})$;

$\boldsymbol{X}^{k+1}=D_{\omega_{\boldsymbol{X}}}(\boldsymbol{V}^{k})$;

$\boldsymbol{W}^{k}=\boldsymbol{W}^{k-1}+\delta_{\boldsymbol{Z}}P_{\Omega}(\boldsymbol{S}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^{k})$;

FOR $i=1$ to N

$(\boldsymbol{Z}^{k+1})^{(i)}=\max\{\|(\boldsymbol{W}^{k})^{(i)}\|_{2}-\tau\lambda\delta_{\boldsymbol{Z}},0\}\dfrac{(\boldsymbol{W}^{k})^{(i)}}{\|(\boldsymbol{W}^{k})^{(i)}\|_{2}}$;

END FOR

END FOR

$\boldsymbol{X}_{\text{opt}}=\boldsymbol{X}^{\text{Max}+1}$;$\boldsymbol{Z}_{\text{opt}}=\boldsymbol{Z}^{\text{Max}+1}$;

FOR $i=1$ to N

FOR $j=1$ to T

IF$(i,j)\in\Omega$

$X_{\text{rec}}(i,j)=\text{S}(i,j)$;

ELSE

$X_{\text{rec}}(i,j)=X_{\text{opt}}(i,j)$; /* 填充丢失的元素 */

END IF

END FOR

IF $\boldsymbol{Z}_{\text{opt}}^{(i)}\neq 0$

$\boldsymbol{X}_{\text{rec}}^{(i)}=\boldsymbol{X}_{\text{opt}}^{(i)}$; /* 替换故障传感器节点收集的数据行 */

END IF

END FOR

RETURN $\boldsymbol{X}_{\text{rec}}$

4 仿真实验

4.1 实验条件

为了评估算法性能,采用 Intel 室内项目中采集的温度数据作为实验数据。选取 52 个传感器节点在连续 300 个时隙中采集的温度数据构建原始环境数据矩阵 $\boldsymbol{X}_{N\times T}$,其中 $N=52$,$T=300$。利用 $\boldsymbol{X}_{N\times T}$ 合成得到用于实验的采样矩阵 $\boldsymbol{S}_{N\times T}$,合成步骤为

(1)根据数据采样率,产生随机的采样数据下标索引集合 Ω。依据 Ω 从原始环境数据矩阵 $\boldsymbol{X}_{N\times T}$ 中

采样元素,得到合成数据矩阵 $\boldsymbol{S}_{N\times T}$,满足

$$S_{ij}=\begin{cases}X_{ij} & (i,j)\in\Omega\\ 0 & \text{其他}\end{cases} \tag{20}$$

(2)基于传感器节点故障率,确定故障节点数 m,从 $\boldsymbol{S}_{N\times T}$ 中随机选取 m 行,并将这 m 行中 50%的非零元素叠加随机噪声 Z_{ij}。假设错误数据的下标索引集合为 υ。最终得到合成的采样矩阵 $\boldsymbol{S}_{N\times T}$,满足

$$S_{ij}=\begin{cases}S_{ij}+Z_{ij} & (i,j)\in\upsilon\\ S_{ij} & \text{其他}\end{cases} \tag{21}$$

式中:Z_{ij} 是均值为 0,方差为 δ^2 的正态分布随机变量,即 $Z_{ij}\sim N(0,\delta^2)$。

执行算法 1 后,通过重建数据矩阵 $\boldsymbol{X}_{rec}$ 和原始环境数据矩阵 $\boldsymbol{X}_{N\times T}$ 的对比来衡量算法性能。对于 λ 和 τ 等可调参数自适应设置的理论研究还没有展开,本文依据所处理问题的先验知识对其进行交叉验证。

4.2 性能参数定义

(1)丢失数据恢复误差 ε_m,表示恢复丢失环境数据的精确程度。ε_m 可以表示为

$$\varepsilon_m=\frac{\sqrt{\sum_{i,j;(i,j)\in\Omega'}(X(i,j)-X_{rec}(i,j))^2}}{\sqrt{\sum_{i,j;(i,j)\in\Omega'}(X(i,j))^2}} \tag{22}$$

式中:Ω' 表示丢失数据下标索引集。

(2)数据重建误差 ε_{rec},表示重建环境数据矩阵的精确程度,同时能反映恢复丢失数据的精确程度和重建错误数据行的精确程度。其可以表示为

$$\varepsilon_{rec}=\frac{\sqrt{\sum_{i,j;1\leqslant i\leqslant N,1\leqslant j\leqslant T}(X(i,j)-X_{rec}(i,j))^2}}{\sqrt{\sum_{i,j;i\in E'\cap(i,j)\in\Omega'}(X(i,j))^2+\sum_{i,j;i\in E,1\leqslant j\leqslant T}(X(i,j))^2}} \tag{23}$$

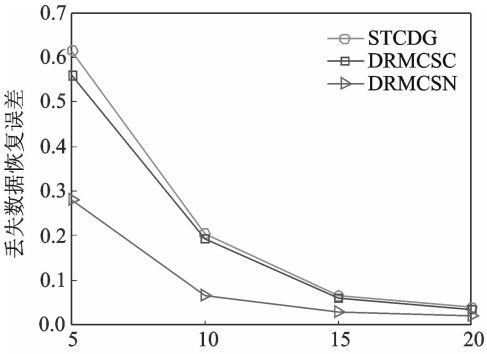
式中:E 为故障节点集合,E' 为正常传感器节点集合。$X(i,j)$ 满足:$i,j:i\in E'\cap(i,j)\in\Omega'$,表示正常传感器节点丢失的数据;$X(i,j)$ 满足:$i,j:i\in E,1\leqslant j\leqslant T$ 表示故障传感器节点对应的数据行。

4.3 仿真结果和分析

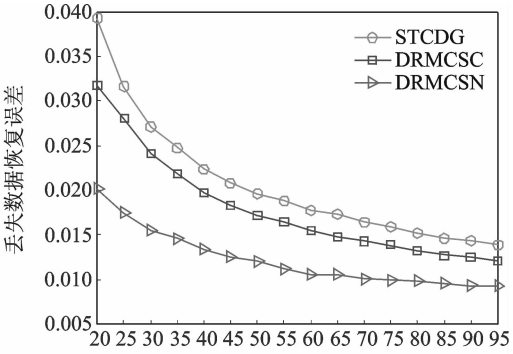
将 DRMCSN 算法与 STCDG[8] 和 DRMCSC[10] 进行对比分析,用于对比的数据为 15 次随机实验结果的平均值。图 1 给出了在不同数据采样率的情况下,丢失数据恢复误差的性能对比。在实验中,设置节点故障率为 40%,数据采样率在 5%到 95%之间。为便于观察,将恢复误差通过图 1(a,b)两张图片呈现。如图 1 所示,数据采样率越低,所有算法对丢失数据恢复误差越大。在数据采样率为 5%时,STCDG 和 DRMCSC 算法的恢复误差接近 60%,而 DRMCSN 算法不到 30%。可见在数据采样比较低的情况下,DRMCSN 算法明显优于其他两个算法,而且采样率越低,DRMCSN 算法的优势越明显。随着采样率的升高,所有算法的恢复误差都逐渐降低。在数据采样率比较高的情况下,几种算法恢复误差都较小,但是 DRMCSN 算法性能还是明显优于其他算法。DRMCSN 算法减少了数据错误对丢失数据恢复误差的影响,从而带来了性能上的优势。

图 2 给出了不同节点故障率的条件下,丢失数据恢复误差的性能对比,其中设置数据采样率为 50%,传感器节点故障率从 10%递增到 90%。随着故障传感器节点数量的增加,所有算法对缺失元素的恢复误差随之增加,但 DRMCSN 始终优于其他算法,并且当传感器节点故障率较高时,DRMCSN 的优越性更加明显。图 2 反映了错误数据的存在对丢失数据恢复误差的影响程度。STCDG 和 DRMCSC 算法由于没有考虑错误数据的存在,因此恢复误差受到的影响较大,而 DRMCSN 算法中引入的噪声矩阵的 L2,1 范数平滑了错误数据的影响,从而丢失数据恢复误差相对较小。

在不同故障率的条件下,数据重建误差性能对比如图 3 所示,其中数据采样率为 50%,传感器节点故障率从 10%到 90%。从图 3 可见 DRMCSN 算法性能明显优于其他算法,其主要原因为:(1)由图 2 可知 DRMCSN 算法由于减少了错误数据的影响,从而对丢失数据的恢复误差优于其他算法;(2)由于



(a) 数据采样率5%~20%
(a) Data sampling rate from 5% to 20%



(b) 数据采样率20%~95%
(b) Data sampling rate from 20% to 95%

图 1 不同数据采样率下的丢失数据恢复误差

Fig. 1 Recovery error of missing data under different sampling rate

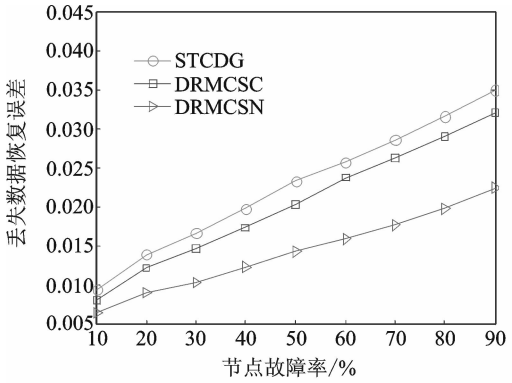


图 2 不同故障率下丢失数据恢复误差

Fig. 2 Recovery error of missing data under different faulty rates

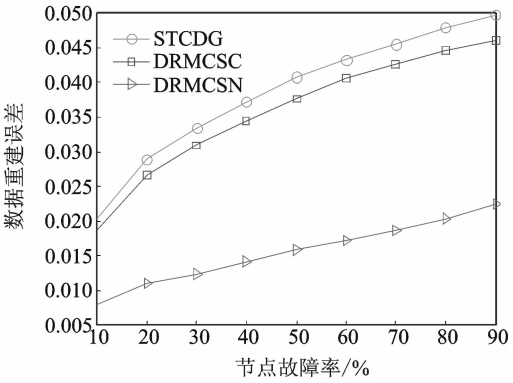


图 3 不同故障率下数据重建误差

Fig. 3 Reconstruction error under different faulty rates

STCDG 和 DRMCSC 算法不能识别故障传感器节点，不能对错误数据行进行替换，因此对应的数据重建误差大，而 DRMCSN 能够识别故障节点，并能对错误数据行进行替换，从而对错误数据行的重建性能优于其他算法。

5 结束语

无线传感器网络数据收集过程数据丢失和错误等问题普遍存在。数据的丢失和错误影响了收集数据的可用性和准确性，因此本文提出一种基于结构化噪声矩阵补全的无线传感器网络收集数据重建方法。该方法旨在从含有错误数据元素的不完整收集数据中重建环境数据。将数据错误建模为原始环境数据受到结构化噪声的污染，进一步利用环境数据矩阵低秩特性，从而将无线传感器网络收集数据重建问题建模为结构化噪声矩阵补全问题，并基于矩阵空间交替线性分裂 Bregman 迭代算法实现该问题的求解。实验结果表明，该方法可以显著提高环境数据的重建精度，并可以识别采集到错误数据的传感器节点。本文的后续工作将包括利用收集数据矩阵的时间和空间相关性进一步提升数据恢复精度等。

参考文献：

[1] 李正周，缪鹏飞，刘勇，等. 基于无线传感器网络的大型场所火情检测与定位算法[J]. 数据采集与处理，2014，29(6)：964-969.
Li Zhengzhou, Mou Pengfei, Liu Yong, et al. Fire detection and position algorithm for large room based on wireless sensor

network[J]. Journal of Data Acquisition and Processing, 2014, 29(6):964-969.
[2] 朱志宇, 苏岭东. 基于分布式粒子滤波的二进制无线传感器网络目标跟踪[J]. 数据采集与处理, 2015,30(3):564-570.
Zhu Zhiyu, Su Lingdong. Target tracking based on distributed particle filtering in binary wireless sensor network[J]. Journal of Data Acquisition and Processing, 2015,30(3):564-570.
[3] Xie K, Ning X, Wang X, et al. Recover corrupted data in sensor networks: A matrix completion solution[J]. IEEE Transactions on Mobile Computing, 2017,16(5):1434-1448.
[4] Kamal A R M, Bleakley C, Dobson S. Packet-level attestation (PLA): A framework for in-network sensor data reliability [J]. ACM Transaction on Sensor Networks, 2013, 9(2): 1-19.
[5] Koushanfar F, Potkonjak M. Markov chain-based models for missing and faulty data in MICA2 sensor motes[C]//The 4th IEEE conference on Sensors. Irvine, California, USA:IEEE, 2005:1430-1434.
[6] 陈蕾,陈松灿.矩阵补全模型及其算法研究综述[J].软件学报,2017,28(6):1547-1564.
Chen Lei, Chen Songcan. Survey on matrix completion models and algorithms[J]. Journal of Software, 2017,28(6):1547-1564.
[7] Cheng Jie, Jiang Hongbo, Ma Xiaoqiang, et al. Efficient data collection with sampling in WSNs: Making use of matrix completion techniques [C]//2010 IEEE Global Communications Conference. Miami, Florida, USA:IEEE, 2010: 1-5.
[8] Cheng Jie, Ye Qiang, Jiang Hongbo, et al. STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks[J]. IEEE Transaction on Wireless Communication, 2013, 12(2):850-861.
[9] Kong Linghe, Xia Mingyuan, Liu Xiaoyang, et al. Data loss and reconstruction in sensor networks[C]//International Conference on Computer Communiations. Turin, Italy:IEEE, 2013: 1654-1662.
[10] He Jingfei, Sun Guiling, Zhang Ying, et al. Data recovery in wireless sensor networks with joint matrix completion and sparsity constraints [J]. IEEE Communications Letters, 2015, 19(12):2230-2233.
[11] Sheng Bo, Li Qun, Mao Weizhen et al. Outlier detection in sensor networks[C]//The 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc). Montreal, Quebec, Canada: ACM, 2007:219-228.
[12] Ding Min, Cheng Xiuzhen. Robust event boundary detection in sensor networks—A mixture model based approach[C]//International Conference on Computer Communications. Rio de Janeiro, Brazil:IEEE, 2013: 2991-2995.
[13] Guo Shuo, Zhong Ziguo, Chen Jiming, et al. Detecting faulty nodes with data errors for wireless sensor networks[J]. ACM Transactions on Sensor Networks, 2014, 10(3):1-27.
[14] Tang Yu, Zhang Bowu, Jing Tao, et al. Robust compressive data gathering in wireless sensor networks[J]. IEEE Transactions on Wireless Communications, 2013, 12(6): 2754-2761.
[15] 陈蕾, 杨庚, 陈正宇,等. 基于线性 Bregman 迭代的结构化噪声矩阵补全算法[J]. 计算机学报, 2015(7):1357-1371.
Chen Lei, Yang Geng, Chen Zhengyu, et al. Linearized bregman iteration algorithm for matrix completion with structural noise[J]. Chinese Journal of Computer, 2015, 38(7):1357-1371.
[16] 陈蕾, 杨庚, 陈正宇, 等. 基于结构化噪声矩阵补全的 Web 服务 QoS 预测[J]. 通信学报, 2015, 36(6):49-59.
Chen Lei, Yang Geng, Chen Zhengyu, et al. Web services QoS prediction via matrix completion with structural noise[J]. Journal on Communications, 2015, 36(6):49-59.
[17] Cai Jianfeng, Candes E J, Shen Z. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal of Optimization, 2010, 20(4): 1956-1982.
[18] Combettes P L, Wajs V R. Signal recovery by proximal forward-backward splitting[J]. Multiscale Modeling and Simulation, 2005, 4(4): 1168-1200.

作者简介:

陈正宇(1978-),男,博士,副教授,研究方向:无线传感器网络、机器学习和信号与信息处理,E-mail: zych@jit.edu.cn。

陈蕾(1975-),男,博士,副教授,研究方向:机器学习、服务计算和信息安全。

胡国兵(1978-),男,博士,副教授,研究方向:数字信号处理、认知无线电等。

戴华(1982-),男,博士,副教授,研究方向:数据管理与安全、数据库安全等。