一种基于深度学习的异常数据清洗算法

匡俊搴[①②③] 赵 畅[①④] 杨 柳[⑤] 王海峰[⑤] 钱 骅*[①②③]

[①](中国科学院上海高等研究院 上海 201210)

[②](上海科技大学信息科学与技术学院 上海 201210)

[③](中国科学院大学 北京 100049)

[④](中国科学院大学微电子学院 北京 100049)

[⑤](中国科学院上海微系统与信息技术研究所 上海 200050)

摘 要：在物联网(IoT)中采用合适的异常数据清洗算法能极大地提升数据质量。许多研究人员采用统计学方法或分类聚类等方法对时-空相关数据进行清洗。但这些方法需要额外的先验知识，会给汇聚节点带来额外的计算开销。该文根据低秩-稀疏矩阵分解模型，提出一种基于深度神经网络的快速异常数据清洗算法，来解决物联网中时-空相关数据的清洗问题。结合感知数据的时-空相关性和异常值的稀疏性，将异常数据清洗问题转换为优化问题，并采用迭代阈值收缩算法(ISTA)求解该优化问题，再将ISTA算法展开成一个固定长度的深度神经网络。实际数据集的实验结果表明，该方法能够自动更新阈值，比传统的ISTA算法收敛速度更快，精度更高。

关键词：物联网；异常数据清洗；迭代阈值收缩算法；展开；深度神经网络

中图分类号：TN915; TP181 **文献标识码**：A **文章编号**：1009-5896(2022)02-0507-07

DOI: 10.11999/JEIT201097

An Outlier Cleaning Algorithm Based on Deep Learning

KUANG Junqian[①②③] ZHAO Chang[①④] YANG Liu[⑤]

WANG Haifeng[⑤] QIAN Hua[①②③]

[①](*Shanghai Advanced Research Institute*, *Chinese Academy of Sciences*, *Shanghai* 201210, *China*)

[②](*School of Information Science and Technology*, *ShanghaiTech University*, *Shanghai* 201210, *China*)

[③](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

[④](*School of Microelectronics*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

[⑤](*Shanghai Institute of Microsystem and Information Technology*, *Chinese Academy of Sciences*, *Shanghai* 200050, *China*)

Abstract: The use of appropriate abnormal data cleaning algorithms in the Internet of Things (IoT) can greatly improve data quality. Statistical methods or clustering methods are utilized to clean anomalies in Spatio-temporal data. However, these methods require additional prior knowledge, which will incur additional computational overhead for the sink node. In this paper, in line with the low-rank sparse matrix decomposition model, a fast anomaly cleaning algorithm based on a deep neural network is proposed to solve the Spatio-temporal data cleaning problem in IoT. Both the Spatio-temporal correlation of sensing data and the abnormal values' sparsity are considered in an optimization problem. The Iterative Shrinkage-Thresholding Algorithm (ISTA) is used to solve it. Then the ISTA is unfolded into a fixed-length deep neural network. The real-world dataset's experimental results show that the proposed method can automatically update the thresholds faster and more accurately than the traditional ISTA.

Key words: Internet of Things (IoT); Outlier cleaning; Iterative Shrinkage-Thresholding Algorithm (ISTA); Unfolding; Deep neural network

收稿日期：2020-12-30；改回日期：2021-07-21；网络出版：2021-11-09

*通信作者： 钱骅 qianh@sari.ac.cn

基金项目：国家自然科学基金(61971286)，国家重点研究发展计划(2020YFB2205603)，上海市科学技术委员会科技创新行动计划(19DZ1204300)

Foundation Items: The National Natural Science Foundation of China (61971286) , The National Key Research and Development Program of China (2020YFB2205603), The Science and Technology Commission Foundation of Shanghai (19DZ1204300)

1 引言

随着物联网应用在实际生活与生产中的普及，其以数据为中心的特点日益凸显。密集部署的传感器节点会产生大量的传感器数据，由于节点能量受限、监测环境较为复杂、节点容易遭受外界攻击等，经常出现异常值[1]。由于物联网系统的运行主要依赖传感器数据，数据冗余和数据异常值会大大降低物联网应用的有效性。因此，必须采用数据清洗技术来去除异常值对物联网系统的影响。数据清洗的研究内容包括：重复数据检测、异常数据检测、缺失数据处理、不一致数据处理、逻辑错误检测等，是从事后诊断角度提升和保证数据质量的主要手段[2,3]。设计物联网异常数据清洗算法，是物联网数据分析中的关键问题。

目前，在基于无线传感器网络的物联网领域中，基于异常值的数据清洗技术与异常点检测技术类似，可分为以下几类：第1类是基于统计学的方法[4]，假定数据集服从某种概率分布模型，把具有低概率的对象视为异常点，但实际情况不一定符合统计规律。第2类是基于聚类的方法[5]，如果某些聚类簇的数据样本量比其他簇少得多，而且这个簇里的数据的特征也与其他簇差异很大，则该簇里的大部分样本点可视为异常点，但该方法需要事先确定阈值，这对于不同的数据集往往是比较困难的。第3类是基于专门的异常点检测算法，包括一类支持向量机(One-Class SVM)[6]、孤立森林(Isolation forest)[7]等。One-Class SVM通过建立分类模型得到一组精确的异常值，技术难点在于计算复杂度高和选择合适的核函数，更适用于中小型数据集的原型分析；Isolation Forest具有线性时间复杂度，可以部署在大规模分布式系统上来加速运算，但不适用于特别高维的数据，在某些局部的异常点较多的时候可能不准确。

此外，基于递归主成分分析(Recursive Principal Component Analysis, R-PCA)的异常数据清洗算法也得到了很多应用。Zhou等人[8]提出稳定的主成分追踪的方法来解决有噪情况下R-PCA算法数据恢复准确性的问题，采用低秩-稀疏矩阵分解(Low-Rank and Sparse Matrix Decomposition, LRaSMD)的技术，将2维的测量矩阵分解为低秩矩阵、稀疏矩阵和噪声矩阵，然后设计算法进一步处理。该方法旨在从有稀疏干扰的数据中恢复出低秩矩阵，但需要准确估计正常模式的相关矩阵，计算量特别大。Xu等人[9]提出一种基于LRaSMD的字典重建和异常提取方法，利用完备字典和稀疏编码构造低秩矩阵达到清洗异常值的目的，不过字典学习的计算量巨大，恢复精度较低。

本文针对传统异常数据清洗算法需要先验统计知识、计算量大、精度低的弊病，在LRaSMD模型的基础上，提出了一种基于深度神经网络的迭代阈值收缩算法框架，对物联网中时-空相关数据进行快速清洗。迭代阈值收缩算法(Iterative Shrinkage-Thresholding Algorithm, ISTA)是梯度下降法的延伸，求解的是1范数稀疏性正则化约束下的反问题[10]。其在图像处理[11]、压缩感知[12]以及信号处理[13]等领域有着广泛的应用。由于LRaSMD可以转换为上述反问题，所以可以引入ISTA来进行异常数据清洗问题的求解。虽然1范数约束问题是凸的，使得ISTA具有全局收敛性，但是其也存在不足，比如收敛速度慢、对初始参数敏感等。为了解决这些问题，本文进一步将ISTA展开为定长的深度神经网络，以神经网络层数来代替迭代次数，从而构造出ISTA-Net框架。在实际数据集上对该框架进行了评估，结果表明，该数据清洗方法能够得到高质量的有效数据，算法收敛速度快，精度更高。

本文的其余部分组织如下。第2节介绍了系统模型和优化问题。第3节描述了所提出的基于深度神经网络的快速异常数据清洗算法框架。第4节使用真实数据集进行实验仿真，验证了算法的性能。第5节对全文进行了总结。

2 系统建模

2.1 问题描述

在本文中，将针对特定的无线传感器网络(Wireless Sensor Networks，WSNs)应用场景，利用数据的时-空相关性，设计适合多传感器数据的离线异常数据清洗算法。本场景中数据清洗的对象为多传感器数据，算法需满足3个条件：只利用测量数据的时-空相关性；只考虑存在异常点的情况；算法的输出为逼近真实的数据，达到较高的精度。

如图1所示，椭圆形代表监测区域，椭圆形内的若干个小圆圈代表传感节点，无线传感器网络主要由部署在监测区域内的大量传感器节点组成。这些传感节点负责采集环境数据，汇聚节点将周围若干个传感节点的数据集中起来，再经由基站以无线通信的方式传输给后台数据中心。

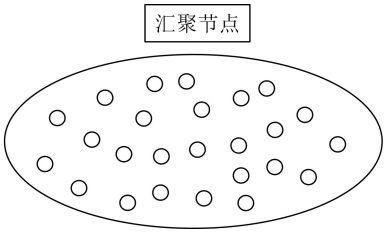


图 1 传感节点数据采集和传输示意图

由于传感器在某一时刻观测到的读数与在前一个时刻观测到的读数相似，并且相邻的多个传感节点的测量值相似，所以无线传感器的数据具有时-空相关性。当在某个域中表示时，信号有很多系数接近或等于零，因此假设时-空相关数据是低秩的[14]。

由于传感网中异常值的出现具有随机性和偶然性，数目非常少，并且数值任意大。而稀疏性模型是指信号在某些域中接近或等于零。所以可以用稀疏矩阵来近似表示异常数据。利用凸优化工具，尤其是l_1范数最小化，来实现稀疏表示[15]。

对于一个含有m个传感节点和一个固定汇聚节点组成的传感器网络，在时间序列$T=1,2,\cdots,n$内，可以将测量数据组合成测量矩阵$\boldsymbol{R}\in\mathbb{R}^{m\times n}$，矩阵$\boldsymbol{R}$的每一列都是$m$维的数据向量。应用LRaSMD模型[8,16]，如图2所示，将测量矩阵$\boldsymbol{R}$可以描述成实际数据$\boldsymbol{L}$、异常值$\boldsymbol{S}$和噪声$\boldsymbol{V}$的和，然后得到式(1)的模型

$$\boldsymbol{R}=\boldsymbol{L}+\boldsymbol{S}+\boldsymbol{V} \tag{1}$$

其中，实际数据$\boldsymbol{L}$为低秩矩阵，异常值$\boldsymbol{S}$为稀疏矩阵，随机噪声$\boldsymbol{V}$为零均值的高斯白噪声矩阵。

异常数据清洗问题的目的是利用测量数据求实际数据，即从$\boldsymbol{R}$中提取$\boldsymbol{L}$，并最小化误差的平方和。因此，LRaSMD模型可转化为以下的优化问题

$$\min_{\boldsymbol{L},\boldsymbol{S}}\frac{1}{2}\|\boldsymbol{R}-(\boldsymbol{L}+\boldsymbol{S})\|_{\mathrm{F}}^2+\lambda_1\|\boldsymbol{L}\|_*+\lambda_2\|\boldsymbol{S}\|_1 \tag{2}$$

其中，$\|\cdot\|_{\mathrm{F}}$简称F-范数，代表矩阵各项元素的绝对值平方的总和；$\|\cdot\|_*$代表核范数，是矩阵的秩的凸替代；$\|\cdot\|_1$代表矩阵$\boldsymbol{S}$的l_1范数，是l_0范数的凸替代；λ_1, λ_2用于权衡数据拟合误差和近似矩阵的秩。

2.2 迭代收缩阈值算法(ISTA)

式(2)是一个正则化的最小二乘问题，存在许多数值最小化求解算法。文献[17]采用ISTA算法，应用Moreau近端映射的定义，将矩阵$\boldsymbol{L}$和$\boldsymbol{S}$分离出来。

$$\begin{aligned}\mathrm{prox}(\boldsymbol{L})&=\underset{\boldsymbol{U}_1}{\mathrm{argmin}}\left\{\lambda_1\|\boldsymbol{L}\|_*+\frac{1}{2}\|\boldsymbol{U}_1-\boldsymbol{L}\|_{\mathrm{F}}^2\right\}\\&=\mathrm{SVT}_{\lambda_1}(\boldsymbol{L})\end{aligned} \tag{3}$$

$$\begin{aligned}\mathrm{prox}(\boldsymbol{S})&=\underset{\boldsymbol{U}_2}{\mathrm{argmin}}\left\{\lambda_2\|\boldsymbol{S}\|_1+\frac{1}{2}\|\boldsymbol{U}_2-\boldsymbol{S}\|_{\mathrm{F}}^2\right\}\\&=\mathcal{T}_{\lambda_2}(\boldsymbol{S})\end{aligned} \tag{4}$$

运算符SVT表示奇异值阈值算子，低秩矩阵$\boldsymbol{L}$的奇异值分解为$\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{\Sigma}$，即$\boldsymbol{L}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$，且

$$\mathrm{SVT}_{\lambda_1}(\boldsymbol{L})=\boldsymbol{U}\boldsymbol{\Lambda}_{\lambda_1}(\boldsymbol{\Sigma})\boldsymbol{V}^{\mathrm{T}} \tag{5}$$

$\lambda_1\in\mathbb{R}^+$定义的软阈值算子为

$$\Lambda_{\lambda_1}(l):=\mathrm{sign}(l)\max(|l|-\lambda_1,0) \tag{6}$$

l_1阈值算子为

$$\mathcal{T}_{\lambda_2}(s)=\mathrm{sign}(s)\max(|s|-\lambda_2,0) \tag{7}$$

将ISTA应用到最小化式(2)中的问题（$\boldsymbol{L}+\boldsymbol{S}$ ISTA），其解的迭代表达式为

$$\begin{aligned}\boldsymbol{L}^{k+1}&=\mathrm{SVT}_{\lambda_1/L_f}\left\{\left(1-\frac{1}{L_f}\right)\boldsymbol{L}^k-\frac{1}{L_f}\boldsymbol{S}^k+\frac{1}{L_f}\boldsymbol{R}\right\}\\\boldsymbol{S}^{k+1}&=\mathcal{T}_{\lambda_2/L_f}\left\{\left(1-\frac{1}{L_f}\right)\boldsymbol{S}^k-\frac{1}{L_f}\boldsymbol{L}^k+\frac{1}{L_f}\boldsymbol{R}\right\}\end{aligned} \tag{8}$$

其中，L_f是式(2)中2次项的利普希茨常数(本文中$L_f=2$)，k是迭代的次数，λ_1和λ_2是式(2)的正则化参数。

对于大规模问题，很难快速计算出每次的迭代

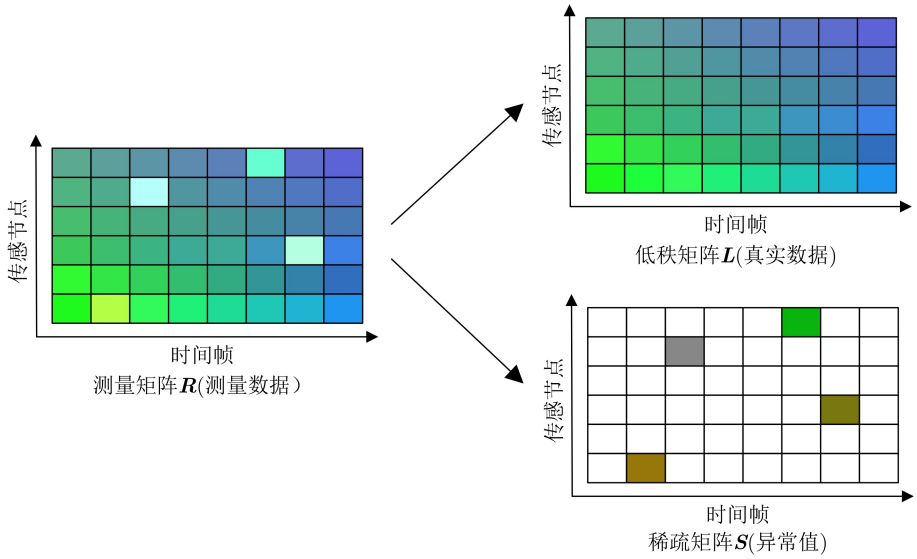


图 2 无噪情况下的低秩-稀疏模型

解。而式(8)中ISTA算法的迭代解依赖阈值参数λ_1和λ_2的选择。因此快速计算出合适的阈值参数是提高该算法收敛速度的关键。

3 算法描述

迭代展开(Unfolding)的概念于2012年被首次提出[18]，显著地改进了收敛性。一个迭代算法可以看作一个神经网络，其中的第k次迭代被视作第k层。迭代展开的方法利用了深度学习和基于模型框架的强大功能，在一些应用领域[19]中提高了算法性能，已经有研究人员对交替方向乘子法[20]、近似梯度法[21]等方法进行了展开。

ISTA与深度神经网络在结构上具有相似之处。深度学习其实是有着超过3层隐藏层的神经网络。将ISTA中的每一次迭代看作一个时间层，软阈值函数等价于激活函数，那么ISTA可以用神经网络展开。图3展示了ISTA的迭代解的数据流图，其中K代表深度神经网络的层数。

选择归一化均方误差(Normalized Mean Square Error, NMSE)作为训练过程的损失函数。对于给定数据集中的第i个数据帧，使用迭代阈值收缩算法(ISTA)分解$\boldsymbol{R}_i$，每次迭代过程中，需要学习的参数是λ_1和λ_2，最终得到该数据帧所对应的矩阵$\boldsymbol{L}^K$和$\boldsymbol{S}^K$，分别用$\hat{\boldsymbol{L}}_i$和$\hat{\boldsymbol{S}}_i$表示。网络输出值和真实值间的损失函数定义为

$$\begin{aligned}\mathrm{NMSE}=&\frac{1}{N}\sum_{i=1}^{N}\left(\alpha\frac{\left\|\boldsymbol{L}_i-\hat{\boldsymbol{L}}_i\right\|_{\mathrm{F}}^2}{\left\|\boldsymbol{L}_i\right\|_{\mathrm{F}}^2}\right.\\&\left.+(1-\alpha)\frac{\left\|\boldsymbol{S}_i-\hat{\boldsymbol{S}}_i\right\|_{\mathrm{F}}^2}{\left\|\boldsymbol{S}_i\right\|_{\mathrm{F}}^2}\right)\end{aligned}\tag{9}$$

其中，$\hat{\boldsymbol{L}}_i$和$\hat{\boldsymbol{S}}_i$是第i个数据帧所对应的神经网络输出，$\boldsymbol{L}_i$和$\boldsymbol{S}_i$是第i个数据帧所对应的真实值，N为数据帧数目，α是对实际数据矩阵估计误差以及异常值数据矩阵估计误差的均衡，$0<\alpha<1$，这里α取0.5。

为了获得最优参数，使用后向传播策略计算梯度或者参数。首先初始化网络权值和神经元的阈值。在前向传播中，使用已经更新的参数，按照式(8)计算隐藏层神经元和输出神经元的输入和输出，并计算NMSE。在反向传播中，根据式(9)中NMSE的定义，再利用二次方自适应学习率优化算法，来更新每个阶段中的每个阈值参数。

具体的ISTA-Net异常数据恢复算法如表1所示，其中步骤(5)—步骤(8)的目的是参照式(8)计算出第k层神经网络的输出$\boldsymbol{L}^k$和$\boldsymbol{S}^k$，之后这两个值将作为第$k+1$层神经网络的输入来参与运算。需要说明的是，每层神经网络的阈值参数λ_1和λ_2的更新是独立的。同时为了取得更好的训练效果[22]，实际中用于SVT和软阈值操作的阈值分别为$\sigma(\lambda_1^k)\cdot b_{\boldsymbol{L}}\cdot\max(\boldsymbol{L}^k)$和$\sigma(\lambda_2^k)\cdot b_{\boldsymbol{S}}\cdot\mathrm{mean}(\boldsymbol{S}^k)$，其中$\sigma(x)$是sigmoid函数，$b_{\boldsymbol{L}}$和$b_{\boldsymbol{S}}$是固定值，这里分别设置为0.1和1.5。

4 仿真结果

为了验证本文所采用的ISTA-Net算法在解决物联网异常数据清洗问题中的有效性，采用Intel Berkeley Research Lab[23]所测的温度数据进行仿真实验。该数据集包含54个传感器采集的14400条温度数据，每个传感器每隔30 s采集1次数据，每小

表 1 ISTA-Net异常数据恢复算法

已知：测量矩阵$\boldsymbol{R}$，深度神经网络层数K

(1) **初始化** $\boldsymbol{S}=\boldsymbol{L}=\boldsymbol{0}$，$\lambda_1>0$，$\lambda_2>0$
(2) **for** 数据集中的每个样本 **do**
(3) 初始化 $\boldsymbol{L}^0$，$\boldsymbol{S}^0$为全零矩阵，$k=0$
(4) **While** $k<K$ **do**
(5) $\boldsymbol{G}_{1_k}=\frac{1}{2}\boldsymbol{L}^k-\frac{1}{2}\boldsymbol{S}^k+\frac{1}{2}\boldsymbol{R}$
(6) $\boldsymbol{G}_{2_k}=\frac{1}{2}\boldsymbol{S}^k-\frac{1}{2}\boldsymbol{L}^k+\frac{1}{2}\boldsymbol{R}$
(7) $\boldsymbol{L}^{k+1}=\mathrm{SVT}_{\lambda_1/L_f}\{\boldsymbol{G}_{1_k}\}$
(8) $\boldsymbol{S}^{k+1}=\mathcal{T}_{\lambda_2/L_f}\{\boldsymbol{G}_{2_k}\}$
(9) $k\leftarrow k+1$
(10) **end while**
(11) 输出$\boldsymbol{L}^K$和$\boldsymbol{S}^K$，并计算归一化均方误差NMSE
(12) 执行会话
(13) **for** 隐藏层或输出层的每个神经元 **do**
(14) 更新网络中的每一个权值和偏差
(15) **end for**
(16) **end for**

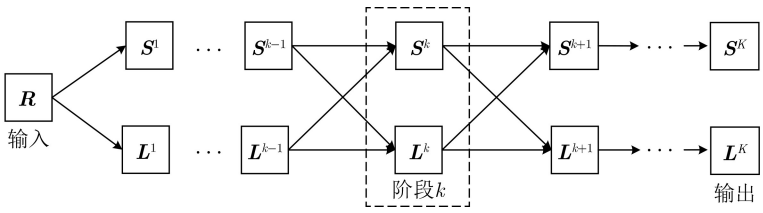


图 3 数据流图

时采集120次数据，温度范围在13.69～37.68°C。假定该数据集代表真实数据，采集过程中的噪声为高斯白噪声。选取其中49个传感器连续采集5天的温度数据，再加上随机异常值作为标注。将传感器1小时采集的数据看作1批数据，则5天总共有120批数据，选取的用于训练和测试的测量矩阵$\boldsymbol{R}$的维度为49×120。仿真实验中，前80批数据为训练数据集，后40批数据为测试数据集。

图4(a)比较了ISTA和ISTA-Net算法在测试阶段的损失函数随迭代次数(或神经网络层数)的变化情况。其中，设置初始阈值$\lambda_1=0.6$，$\lambda_2=0.1$，选取学习率$\eta=0.16$，异常值比例$\alpha=10\%$，异常值随机在10和−10两个值之间挑选，噪声均值为0，噪声方差$\sigma^2=0.01$。从图4(a)可以看出，两种算法的性能损失均是随着迭代次数(神经网络层数)的增加而逐渐下降，但是很明显ISTA-Net比ISTA算法的收敛速度更快。ISTA经过近600次迭代才收敛到10^{-2}以下，而ISTA-Net仅仅使用了25层神经网络就达到了相同的误差水平。

保证异常数据检测的准确性是解决异常数据清洗问题的前提。以F1分数为检测准确性的评估指标，其定义如下

$$\text{F1score}=\frac{2\times\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}} \tag{10}$$

其中，TP指的是确实包含异常值且被算法检测出的数据个数，FP指的是本身不包含异常值却被算法判定为异常的数据个数，FN指的是本身包含异常值但是算法却没有检测出的数据个数。从以上定义可以看出，F1分数越接近1，检测的正确率越高。

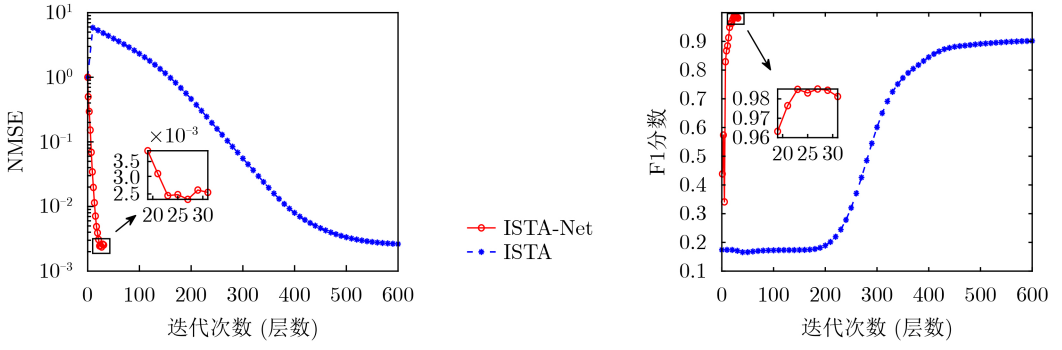
图4(b)描述了ISTA和ISTA-Net算法的F1分数随迭代次数(或神经网络层数)的变化关系。从图中可以看出两种算法的F1均是随着迭代次数的增加而逐渐增大，检测的准确率越来越高，收敛之后达到了0.9以上。不过，ISTA-Net的增加速度明显快于ISTA，并且收敛之后前者的F1还要大于后者。

在上述的对比中，ISTA-Net算法要优于ISTA的原因是ISTA算法本身是一个固定阈值的计算过程，算法的最终性能严重依赖分解软阈值时收缩阈值的初始值；而ISTA-Net算法受益于神经网络内部权重更新，对参数选择相对能够快速自动更新收缩阈值。因此ISTA-Net算法收敛更快，数据清洗的精度更高，性能得到了显著的提升。

需要指出的是，ISTA-Net前向传播的每一层的运算量和ISTA算法的1次迭代的运算量相同，通过网络训练得到最佳迭代参数后，ISTA-Net的计算过程与ISTA算法完全一致。此外，将ISTA算法展开为神经网络后，可以加快收敛速度，大大减少所需的迭代次数，这降低了整个算法的计算量。

在网络训练过程中，不同学习率下的ISTA-Net算法的损失随着训练数据批次数的变化如图5所示。其中，异常值、噪声以及初始阈值的设置均与上述相同，神经网络层数为25。由图5看出，不同学习率下的算法损失随着训练批次的增加而逐渐下降，同时学习率越大，损失函数收敛得越快，但是波动也越大。因此，在ISTA-Net算法的训练过程中，选择合适的学习率能够实现算法的收敛速度和恢复精度之间的平衡。

值得强调的是虽然对于Intel Berkeley Research Lab所测的温度数据集，ISTA-Net只需25层固定长度的深度神经网络就达到了优于传统ISTA算法的性能，但是对于不同的数据集，ISTA-Net需要的神经网络层数并不一定相同。不过尽管如此，ISTA-Net算法的收敛性仍远快于传统ISTA算法。这里选取由国家青藏高原科学数据中心提供的大纳伦河流域修正后的温度数据集[24]，相关参数设置与前述相同。图6将ISTA和ISTA-Net算法在该数据集中测试阶段的损失函数随迭代次数(或神经网络层数)的变化情况进行了对比。可以看出在大纳伦河流域数据集中ISTA-Net达到收敛所需的神经网络层数为60，不同于前一数据集的25层，但是相对于传统



(a) ISTA和ISTA-Net算法的NMSE比较　　(b) ISTA和ISTA-Net算法的F1分数比较

图 4 ISTA和ISTA-Net算法的性能对比

ISTA算法的近600次迭代才能收敛而言，性能的提升是比较显著的。

为了证实所提算法方案的优越性，本文将ISTA, ISTA-Net算法和孤立森林算法进行了对比。这里用到的性能指标仍是上面提到的F1分数，数据集使用Intel Berkeley Research Lab所测的温度数据集，数据中异常值比例将逐渐增加，其他的设置(比如噪声以及初始阈值的设置)均与上述相同，ISTA-Net用到的神经网络层数为25。图7描述了随着数据中异常值比例的增加，3种算法的F1分数的变化情况。可以看出，在异常值比例比较小时，ISTA和孤立森林算法的F1分数比较低。这是因为此时实际上被叠加了异常值的数据占的比重小，因此TP比较小，而此时两种算法的FP和FN相较于TP而言较高，进而导致F1分数比较低。但随着异常值比例的逐渐增加，TP的比重提升，ISTA和孤立森林算法的F1分数逐渐升高。值得强调的是，虽然在异常值比例较低时，孤立森林算法的F1分数略高于ISTA，但是随着异常值比例的提升，前者性能逐渐弱于后者，并且差距逐渐拉大。对于ISTA-Net算法而言，其F1分数也是随着异常值比例的增加而逐渐升高，并且其起始点要远高于另外两种算法，性能差距比较明显。

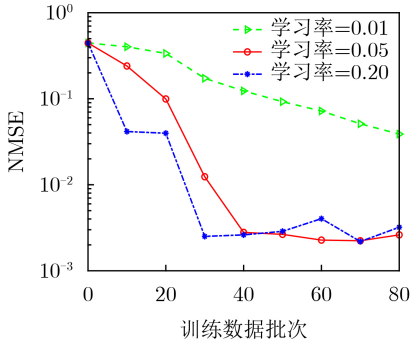


图 5 ISTA-Net损失随训练数据批次数的变化情况

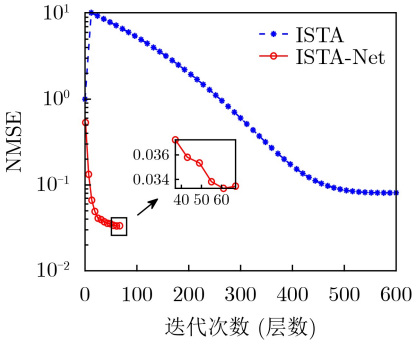


图 6 ISTA和ISTA-Net算法的NMSE比较

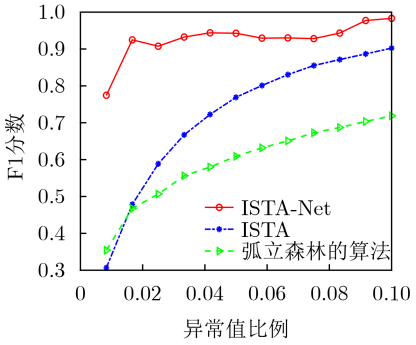


图 7 3种算法的F1分数随异常值比例的变化情况

由上述对比可以看出，ISTA-Net算法确实能够取得比ISTA以及孤立森林算法更好的性能，而且其对不同异常值比例的情况都表现得比较好，具有较好的鲁棒性。

5 结束语

本文针对传统异常数据清洗方法需要先验统计知识以及计算量大的问题，提出了一种基于神经网络的迭代阈值收缩算法，从而对物联网中时-空相关数据进行快速异常数据清洗。利用了感知数据的时-空相关性和异常值的稀疏性，根据低秩-稀疏矩阵分解模型，采用迭代收缩阈值算法(ISTA)求解优化问题，进一步将ISTA展开为定长的深度神经网络。在实际数据集上对算法进行了评估，仿真结果表明，该方法能够自动更新奇异值分解过程中的软阈值参数，克服了传统ISTA算法对初始参数敏感以及收敛速度慢的问题。在选择适当的阈值参数后，算法收敛速度更快，数据清洗的精度更高。

参 考 文 献

[1] 蒋俊正, 杨杰, 欧阳缮. 一种新的无线传感器网络中异常节点检测定位算法[J]. 电子与信息学报, 2018, 40(10): 2358–2364. doi: 10.11999/JEIT171207.
JIANG Junzheng, YANG Jie, and OUYANG Shan. Novel method for outlier nodes detection and localization in wireless sensor networks[J]. *Journal of Electronics & Information Technology*, 2018, 40(10): 2358–2364. doi: 10.11999/JEIT171207.

[2] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11): 2076–2082.
GUO Zhimao and ZHOU Aoying. Research on data quality and data cleaning: A survey[J]. *Journal of Software*, 2002, 13(11): 2076–2082.

[3] YU Tianqi, WANG Xianbin, and SHAMI A. Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems[J]. *IEEE Internet of Things Journal*, 2017, 4(6): 2207–2216. doi: 10.1109/JIOT.2017.2756025.

[4] KUMAR V and KHOSLA C. Data cleaning-a thorough analysis and survey on unstructured data[C]. The 8th International Conference on Cloud Computing, Data Science & Engineering, Noida, India, 2018: 305–309.

[5] DIAO Yinglong, LIU Keyan, MENG Xiaoli, *et al*. A big data online cleaning algorithm based on dynamic outlier detection[C]. 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Xi'an, China, 2015: 230–234.

[6] 田江, 顾宏. 孤立点一类支持向量机算法研究[J]. 电子与信息学报, 2010, 32(6): 1284–1288. doi: 10.3724/SP.J.1146.2009.00861.
TIAN Jiang and GU Hong. Outlier one class support vector machines[J]. *Journal of Electronics & Information Technology*, 2010, 32(6): 1284–1288. doi: 10.3724/SP.J.1146.2009.00861.

[7] ZOU Zhuping, XIE Yulai, HUANG Kai, *et al*. A docker container anomaly monitoring system based on optimized isolation forest[J]. *IEEE Transactions on Cloud Computing*, To be published. doi: 10.1109/TCC.2019.2935724.

[8] ZHOU Zihan, LI Xiaodong, WRIGHT J, *et al*. Stable principal component pursuit[C]. 2010 IEEE International Symposium on Information Theory, Austin, USA, 2010: 1518–1522.

[9] XU Yichu, DU Bo, ZHANG Liangpei, *et al*. A low-rank and sparse matrix decomposition-based dictionary reconstruction and anomaly extraction framework for hyperspectral anomaly detection[J]. *IEEE Geoscience and Remote Sensing Letters*, 2020, 17(7): 1248–1252. doi: 10.1109/LGRS.2019.2943861.

[10] DAUBECHIES I, DEFRISE M, and DE MOL C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint[J]. *Communications on Pure and Applied Mathematics*, 2004, 57(11): 1413–1457. doi: 10.1002/cpa.20042.

[11] BIOUCAS-DIAS J M and FIGUEIREDO M A T. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration[J]. *IEEE Transactions on Image Processing*, 2007, 16(12): 2992–3004. doi: 10.1109/TIP.2007.909319.

[12] CANDES E J, WAKIN M B, and BOYD S. Enhancing sparsity by reweighted l1 minimization[J]. *Journal of Fourier Analysis and Applications*, 2008, 14(5): 877–905.

[13] ELAD M. Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing[M]. New York: Springer, 2010: 185–200.

[14] CHENG Jie, YE Qiang, JIANG Hongbo, *et al*. STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks[J]. *IEEE Transactions on Wireless Communications*, 2013, 12(2): 850–861. doi: 10.1109/TWC.2012.121412.120148.

[15] 李鹏, 王建新, 曹建农. 无线传感器网络中基于压缩感知和GM(1, 1)的异常检测方案[J]. 电子与信息学报, 2015, 37(7): 1586–1590. doi: 10.11999/JEIT141219.
LI Peng, WANG Jianxin, and CAO Jiannong. Abnormal event detection scheme based on compressive sensing and GM (1, 1) in wireless sensor networks[J]. *Journal of Electronics & Information Technology*, 2015, 37(7): 1586–1590. doi: 10.11999/JEIT141219.

[16] LIU Jing and RAO B D. Robust PCA via ℓ_0-ℓ_1 regularization[J]. *IEEE Transactions on Signal Processing*, 2019, 67(2): 535–549. doi: 10.1109/TSP.2018.2883924.

[17] RAHMANI M and ATIA G K. High dimensional low rank plus sparse matrix decomposition[J]. *IEEE Transactions on Signal Processing*, 2017, 65(8): 2004–2019. doi: 10.1109/TSP.2017.2649482.

[18] ORTIZ-RODRIGUEZ J M and VEGA-CARRILLO H R. A neutron spectra unfolding code, based on iterative procedures, designed under LabVIEW environment[C]. 2012 IEEE Ninth Electronics, Robotics and Automotive Mechanics Conference, Cuernavaca, Mexico, 2012: 315–319.

[19] GIRYES R, ELDAR Y C, BRONSTEIN A M, *et al*. Tradeoffs between convergence speed and reconstruction accuracy in inverse problems[J]. *IEEE Transactions on Signal Processing*, 2018, 66(7): 1676–1690. doi: 10.1109/TSP.2018.2791945.

[20] YANG Yang, SUN Jian, LI Huibin, *et al*. ADMM-CSNet: A deep learning approach for image compressive sensing[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(3): 521–538. doi: 10.1109/TPAMI.2018.2883941.

[21] CHEN Yunjin, WEI Yu, and POCK T. On learning optimized reaction diffusion processes for effective image restoration[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 5261–5269.

[22] SOLOMON O, COHEN R, ZHANG Yi, *et al*. Deep unfolded robust PCA with application to clutter suppression in ultrasound[J]. *IEEE Transactions on Medical Imaging*, 2020, 39(4): 1051–1063. doi: 10.1109/TMI.2019.2941271.

[23] Intel Berkeley Research Lab. Intel lab data[EB/OL]. http://db.lcs.mit.edu/labdata/labdata.html, 2019.

[24] 苏凤阁. 大纳伦河流域修正后的温度和降水数据集(1951–2016)[R]. 国家青藏高原科学数据中心, 2019. doi: 10.11888/Hydro.tpdc.270216.
SU Fengge. Revised dataset of temperature and precipitation in the Greater Naren River Basin (1951–2016)[R]. National Tibetan Plateau Data Center, 2019. doi: 10.11888/Hydro.tpdc.270216.

匡俊搴：男，1998年生，博士生，研究方向为大数据信号处理.
赵　畅：女，1996年生，博士生，研究方向为无线传感器网络.
杨　柳：男，1993年生，博士，研究方向为无线传感器网络、分布式信号处理.
王海峰：男，1969年生，研究员，研究方向为移动通信、物联网.
钱　骅：男，1976年生，研究员，研究方向为无线通信、非线性信号处理、大数据信号处理.

责任编辑：马秀强