计算机科学
COMPUTER SCIENCE

基于压缩感知自适应测量矩阵的空气质量主动采样

黄伟杰, 郭贤伟, 於志勇, 黄昉菀

引用本文

黄伟杰, 郭贤伟, 於志勇, 黄昉菀. 基于压缩感知自适应测量矩阵的空气质量主动采样[J]. 计算机科学, 2024, 51(7): 116-123.

HUANG Weijie, GUO Xianwei, YU Zhiyong, HUANG Fangwan. Active Sampling of Air Quality Based on Compressed Sensing Adaptive Measurement Matrix [J]. Computer Science, 2024, 51(7): 116-123.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400111

基于压缩感知自适应测量矩阵的空气质量主动采样

黄伟杰[1] 郭贤伟[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
(211027099@fzu.edu.cn)

摘 要 随着城市化进程的不断加快,工业发展、人口聚集使得空气质量问题日益严峻。出于对采集成本的考虑,对空气质量的主动采样正受到越来越多的关注。但现有模型要么只能迭代选择采样位置,要么难以实时更新采样算法。基于此,提出了一种基于压缩感知自适应测量矩阵的空气质量主动采样方法,将采样位置的选择问题转化为矩阵的列子集选择问题。该方法首先利用历史完整数据进行字典学习,然后将学习后的字典经过列子集选择后得到能够指导批量采样的自适应测量矩阵,最后结合利用空气质量数据特性构建的稀疏基矩阵恢复出未采样的数据。该方法使用压缩感知模型一体化实现采样和推断,避免了使用多个模型的不足。此外,考虑到空气质量的时序变动问题,在每一次的主动采样后,字典还会利用最新数据进行在线更新以指导下一次的采样。两个真实数据集上的实验结果表明,经过字典学习后得到的自适应测量矩阵在低于20%的多个采样率下,恢复性能优于所有基线。

关键词:群智感知;压缩感知;自适应测量矩阵;字典学习;主动采样
中图分类号 TP391

Active Sampling of Air Quality Based on Compressed Sensing Adaptive Measurement Matrix

HUANG Weijie[1],GUO Xianwei[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 With the continuous acceleration of urbanization,industrial development and population agglomeration make the problem of air quality increasingly serious. Due to the cost of sampling,more and more attention is paid to active sampling of air quality. However,the existing models can either only select the sampling location iteratively or hardly update the sampling algorithm in real time. Motivated by this,an active sampling method of air quality based on compressed sensing adaptive measurement matrix is proposed in this paper. The problem of sampling location selection is transformed into the column subset selection problem of the matrix. Firstly,the historical complete data is used for dictionary learning. After column subset selection of the learned dictionary,an adaptive measurement matrix that can guide batch sampling is obtained. Finally,the unsampled data is recovered by using the sparse basis matrix constructed by the data characteristics of air quality. This method uses a compressed sensing model to realize sampling and inference integrally,which avoids the shortcoming of using multiple models. In addition,considering the timing variation of air quality,after each active sampling,the dictionary is updated online with the latest data to guide the next sampling. Experimental results on two real datasets show that the adaptive measurement matrix obtained after dictionary learning has better recovery performance than all baselines at multiple sampling rates less than 20%.

Keywords Crowd sensing,Compressed sensing,Adaptive measurement matrix,Dictionary learning,Active sampling

1 引言

当今世界,人类环境不断恶化,空气质量是人们必须重视的问题之一。世界卫生组织(WHO)估计,全球有90%以上的人口暴露在不符合世卫组织指南的空气质量中[1]。由于城市空气污染严重,空气质量监测的服务需求不断上升[2]。城市空气质量数据主要通过部署空气质量监测站来获取。以北京市为例,其大气环境监测网络包括36个监测站,且绝大部分分布在主城区[3]。在人口众多的重要城市中只部署少量监测站的现状反映出该监测方式存在着建设周期长、运行

到稿日期:2023-04-16 返修日期:2023-09-08
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)

成本高、占地面积大等缺点。因此，如何以少量的成本投入获得城市中多个位置的环境数据成为亟需解决的问题。移动群智感知(Mobile Crowd Sensing，MCS)的提出为此提供了很好的解决方案。MCS 以携带感知设备的人群作为感知数据的节点，他们在移动过程中参与执行相应的感知任务，以补充/替代固定式的社会环境数据监测站点。由于个人移动设备的普及和感知功能的增加，MCS 呈现出感知类别多、无需专门配置设备、位置可灵活变动等优点[4]。目前，基于 MCS 范式已开发了许多以环境为中心的应用平台，如噪声收集统计[5]、城市垃圾统计[6]等。在这些平台中，用户被称作参与者，其接收任务后，将基于自身位置的感知结果发送至平台，平台即可进行汇总处理。城市中的空气质量同样可以通过平台发布任务进行收集，如图 1 所示。

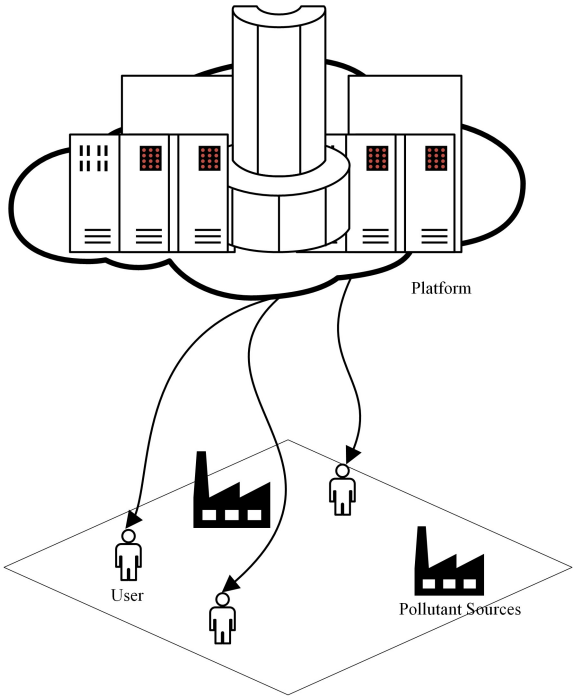

图 1 城市空气质量移动群智感知平台

Fig. 1 Mobile crowd sensing platform for urban air quality

感知数据的质量和成本是 MCS 应用中的两个重要关注点，如何权衡二者之间的关系，即以相对少的成本去感知高质量数据，是移动群智感知任务分配所追求的目标。为了获得多个位置在既定时间范围内的高质量感知结果，最简单的方法是在每个时刻对所有位置进行全采样[7]。但这将导致感知成本增加，因为参与者的时间成本、设备的使用与通信成本等都与采集数据量成正比。由于环境数据(如温/湿度、污染物浓度)大多存在着时空相关性，因此，另一种方法是只选择少量几个位置进行数据采集，再通过机器学习的方法推测其余位置的数据，以达到质量和成本之间的平衡[8]。虽然这种方法已被广泛使用，但仍有以下问题需要深入研究，如采样位置如何选择，数据推测算法如何选择，两者的选择是否相关等。目前，数据推测算法的研究工作较为成熟，例如基于压缩感知的推测算法[9]。该算法首先根据已有的采样数据设计出测量矩阵，然后利用时间序列的时域平滑特性选择与测量矩阵具有低相干性的稀疏基，最后将缺失数据推测问题转化成稀疏向量恢复问题。虽然该算法可以在高缺失率下达到很高的推测精度，但没有涉及采样位置的选择。事实上，根据主动学习的思想，采样位置的选择同样会影响到推测精度[10]。基于此，本文的主要工作将围绕采样位置的主动选择展开，主要贡献包括：

(1)提出了一种自适应测量矩阵(Adaptive Measurement Matrix，AMM)的构建方法，利用已有的少量历史数据，将主动采样问题转化为字典学习后的列子集选择问题，实现了采样与推测一体化的压缩感知算法。

(2)提出了一种在线更新的方式，使每一次主动采样都可学习前一次采样序列的特征，为持续性多轮主动采样提供了更高的推测精度。

2 相关工作

自 2004 年 Donoho 提出压缩感知(Compressed Sensing，CS)后[11]，CS 成为了信号领域使用最广泛的信号压缩与恢复的方法。压缩感知存在的主要挑战包括：通过各类约束条件设计重构信号的高效率算法；构造稀疏基使得原始信号在该变换下更稀疏；设计满足约束等距性(Restricted Isometry Property，RIP)条件的测量矩阵。本文的主要关注点是设计测量矩阵，并用于实现主动采样。

压缩感知测量矩阵的常见构建方法包括随机生成和基于学习这两种。随机测量矩阵具有随机生成的元素，如均值为 0、方差为 $\frac{1}{m}$(m 为矩阵的行数)的高斯随机矩阵[12]，元素服从独立分布的特普利兹矩阵[13]，元素间独立不相关且服从伯努利分布的随机伯努利矩阵[14]。这些随机矩阵主要应用于数据的压缩存储，虽然在重构数据时有着良好的性能，但若用于指导采样，则只能实现随机采样，不能从历史数据中学习主动采样策略，因此并不适合本文所研究的问题。

此外，也有一些工作提出了基于学习的方法来构建测量矩阵。Hegde 等[15]通过学习保持成对距离的近等距嵌入来从训练数据中学习测量矩阵。文献[16-17]通过训练数据寻求最佳约束等距性质从而构建测量矩阵。近年来，深度学习的流行也促使部分研究利用该技术学习测量矩阵。文献[18]在深度学习中使用自动编码器设计测量矩阵支持复杂稀疏信号的恢复。文献[19]提出的深度概率子采样模型能够联合优化任务的自适应子采样以构建表现良好的测量矩阵。文献[20]提出了一种基于非均匀压缩感知的采样-恢复方法，其中利用了隐马尔可夫树(Hidden Markov Tree，HMT)来考虑采样和重建阶段之间稀疏系数的相关性，大大提升了 CS 恢复的性能。文献[21]提出了一种自适应压缩感知算法，构建的自适应测量矩阵所需要的计算量低于标准 CS。这类基于数据驱动的方法相较于随机矩阵在重构时性能更优，但是仍然不能指导主动采样，且需要的训练样本数量过多，计算代价很高。

在主动采样方面，文献[22]在稀疏群智感知框架下，以人群位置作为采样点，利用历史采样数据构建的矩阵进行矩阵分解以构建时间-位置二分图，按成本多少进行排序，选择下一时刻成本最低的一个位置进行采样，最后通过矩阵补全以恢复完整数据。与本文方法不同的是，这种方法更多的是

关注采样的成本，而非推测数据的质量。文献[8]提出了一种框架，在给定推测误差的约束下，迭代地选择最值得采样的位置点，当评估推测的数据质量满足约束时则停止采样。在理想情况下，这类方法能以较少的采样点得到高质量数据，但是在现实采样情况下，很难在每一次迭代时都得到准确的推测误差，且迭代选择的方式计算复杂度较高。文献[23]在地震数据采样中使用了图像处理领域中广泛使用的泊松碟采样方法。该方法会随机选择一些具有一定半径的圆形区域，在每个区域内只采样一个点，相邻采样区域之间没有重叠部分，这有助于确保采样点的分布均匀，且保持了圆形区域内的随机采样。文献[24]在此基础上进行改进，以压缩感知推测精度为目标，提出了边缘保持分段随机采样方法。该方法首先将目标采样序列分为若干段，在每一段中进行随机采样，这样可以有效控制最大采样间隔；并且对采样序列的首尾两段进行全采样，可使推测精度有效提升。但是这两种方法均保留了随机采样，推测精度存在随机性。在批量选择方面，文献[25-26]利用强化学习从历史数据中挖掘时空相关性，度量下一采样时刻哪些位置点更值得采样。但是这类方法需要大量的训练集进行充分的离线训练后，才可进行多轮采样，不适用于仅有少量历史数据的情况。

综上所述，在测量矩阵的研究工作中，目前主要围绕如何更好地重构数据展开，并没有关注如何用于主动采样。而在主动采样的研究工作中，尚未有基于压缩感知测量矩阵实现的工作。鉴于此，本文将如何构建主动采样的测量矩阵作为研究目标。该矩阵不仅可以批量选择采样位置，而且可以在每一次采样后进行实时更新以适应时序变动的环境数据。

3 用于主动采样的测量矩阵构建

压缩感知采样与推测一体化模型的整体框架图如图2所示。

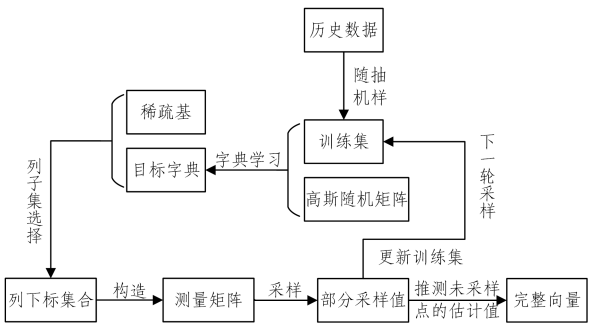

图2 整体框架图

Fig. 2 Overall framework

3.1 压缩感知简介

压缩感知采用亚采样将 n 维向量 $\boldsymbol{X}=[x_1,\cdots,x_n]^{\mathrm{T}}$ 利用测量矩阵 $\boldsymbol{\Phi}\in\mathbb{R}^{m\times n}$ 投影成 m 维向量 $\boldsymbol{Y}=[y_1,\cdots,y_m]^{\mathrm{T}}$（$m<n$）。其前提是，要求 $\boldsymbol{X}$ 是 k-稀疏的（即 $\boldsymbol{X}$ 中只有 k 个非零元素）或者在某种变换下是 k-稀疏（k-sparse）的。实际应用中的信号多属于第二种情况，即 $\boldsymbol{X}=\boldsymbol{\psi S}$，其中 $\boldsymbol{\psi}\in\mathbb{R}^{n\times n}$ 称为稀疏基，$\boldsymbol{S}\in\mathbb{R}^n$ 是 k-稀疏向量。压缩感知的原理如图3所示，其公式化描述如下：

$$\boldsymbol{Y}=\boldsymbol{\Phi X}=\boldsymbol{\Phi\psi S}=\boldsymbol{AS} \tag{1}$$

其中，$\boldsymbol{A}=\boldsymbol{\Phi\psi}\in\mathbb{R}^{m\times n}$ 被称为传感矩阵。

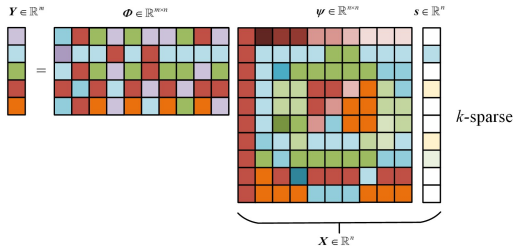

图3 压缩感知原理图

Fig. 3 Schematic diagram of compressed sensing

对于式(1)而言，若已知 $\boldsymbol{Y}$ 和 $\boldsymbol{A}$，则求解稀疏向量 $\boldsymbol{S}$ 时通常采用贪婪追踪算法或凸松弛算法[27]。其中贪婪追踪算法适用于对 $\boldsymbol{S}$ 采用 l_0 范数约束的情况。

$$\begin{aligned}&\hat{\boldsymbol{S}}=\arg\min\|\boldsymbol{S}\|_0\\&\text{s.t.}\ \|\boldsymbol{Y}-\boldsymbol{AS}\|_2^2\leqslant\varepsilon\end{aligned} \tag{2}$$

其中，$\hat{\boldsymbol{S}}$ 为 $\boldsymbol{S}$ 的估计；ε 为噪声或误差；$\|\cdot\|_0$ 表示向量中非零元素的个数，也称为“稀疏度”。贪婪追踪算法的代表性算法是正交匹配追踪（Orthogonal Matching Pursuit，OMP）算法，由于其具有简单性和近似解的求解效率，因此得到了广泛的应用[28]。此外，由于 l_0 范数最小化问题是一个 NP 问题，因此也可以将 l_0 范数松弛为 l_1 范数。

$$\hat{\boldsymbol{S}}=\arg\min\|\boldsymbol{S}\|_1\quad\text{s.t.}\ \|\boldsymbol{Y}-\boldsymbol{AS}\|_2^2\leqslant\varepsilon \tag{3}$$

其中，$\|\cdot\|_1$ 表示向量中非零元素的绝对值之和。l_1 范数最小化问题常采用凸松弛算法求解，常见的有基追踪（Basis Pursuit，BP）算法[29]、迭代收缩阈值算法（Iterative Shrinkage-Thresholding Algorithm，ISTA）[30]等。

3.2 采样与推测

对于待采样向量 $\boldsymbol{X}=[x_1,\cdots,x_n]^{\mathrm{T}}$ 而言，已知采样率为 $r(0<r<1)$，本文研究的问题是在 $\boldsymbol{X}$ 中找到 m 个关键位置（$m=n\times r$ 的向下取整值）进行采样，并利用它们推测出其余元素值，推测误差越小越好。由于不关注参与者任务分配问题，因此首先假设当采样位置和时刻确定后，默认一定有参与者进行响应并采样。其次，由于只关注主动采样对推测误差的影响，因此本文进一步假设参与者的采样结果是真实可信的，即十分接近或等于实际环境的数值，并不会出现需在空旷场地采样而在室内采样导致结果与室外不同，以及传感器本身错误或使用错误等问题。

为了基于压缩感知实现采样与推测一体化的算法，本文的研究目标是利用测量矩阵 $\boldsymbol{\Phi}$ 进行主动采样。该矩阵的特点是每一行只有一个元素为1（其余元素均为0）且1均分布在不同列。此时，具有非零列的下标即为主动采样的位置点。举例来说，假设 $\boldsymbol{\Phi}\in\mathbb{R}^{3\times6}$，说明 $\boldsymbol{X}\in\mathbb{R}^6$，采样率 $r=50\%$。若经过训练后得到：

$$\boldsymbol{\Phi}=\begin{bmatrix}0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&1&0\end{bmatrix} \tag{4}$$

说明应主动采样的元素为 x_2，x_3 和 x_5，可利用一个掩码向量 $\boldsymbol{mask}=[0,1,1,0,1,0]^{\mathrm{T}}$ 进行表示。当参与者完成采集任务后，将这3个元素构成采样向量 $\boldsymbol{Y}=[y_1,y_2,y_3]^{\mathrm{T}}$，其中 $y_1=x_2$，$y_2=x_3$，$y_3=x_5$。最后，利用 $\boldsymbol{Y}$ 和 $\boldsymbol{A}=\boldsymbol{\Phi\psi}$（$\boldsymbol{\psi}$ 是事先

选定的稀疏基），根据式(2)或式(3)先得到稀疏向量$\hat{\boldsymbol{S}}$，再利用$\hat{\boldsymbol{X}}=\boldsymbol{\psi}\hat{\boldsymbol{S}}$得到未采样的$x_1$，$x_4$和$x_6$的估计值。

3.3 测量矩阵的学习

为了得到可以用于主动采样的测量矩阵Φ，需要充分挖掘训练数据中各采样位置的重要性。具体做法如下：

Step 1 首先，从历史数据集中取出t条完整采样的训练数据，按列排放构成训练矩阵$\mathbb{X}^R=[X_1^R,\cdots,X_t^R]\in\mathbb{R}^{n\times t}$。然后，对$\mathbb{X}^R$中的每一列均随机抽样$m$个采样值，得到压缩矩阵$\mathbb{Y}^R=[Y_1^R,\cdots,Y_t^R]\in\mathbb{R}^{m\times t}$。

Step 2 生成一个初始字典$\boldsymbol{D}=[d_1,\cdots,d_n]\in\mathbb{R}^{m\times n}$，将$\mathbb{Y}^R$中的每一列$\boldsymbol{Y}_i^R$（$1\leqslant i\leqslant t$）根据式(5)求解其重构向量$\boldsymbol{T}_i^R$：

$$\begin{aligned}&\min_{T_i^R}\|\boldsymbol{Y}_i^R-\boldsymbol{D}\boldsymbol{T}_i^R\|_2^2\\&\text{s.t. }\|\boldsymbol{T}_i^R\|_0\leqslant k,1\leqslant i\leqslant t\end{aligned}\tag{5}$$

其中，k表示稀疏度上限。

Step3 利用字典学习技术对初始字典进行更新，即求解以下优化模型：

$$\begin{aligned}&\min_{D,T_i^R}\frac{1}{t}\sum_{i=1}^{t}\|\boldsymbol{Y}_i^R-\boldsymbol{D}\boldsymbol{T}_i^R\|_2^2\\&\text{s.t. }\|\boldsymbol{T}_i^R\|_0\leqslant k,1\leqslant i\leqslant t\end{aligned}\tag{6}$$

式(6)有两个待优化目标$\boldsymbol{D}$和$\mathbb{T}^R=[T_1^R,\cdots,T_t^R]\in\mathbb{R}^{n\times t}$。本文选取K-SVD算法进行更新优化[31]。

Step 4 经过字典学习后的$\boldsymbol{D}$与选定的稀疏基$\boldsymbol{\psi}\in\mathbb{R}^{n\times n}$逆相乘，可得矩阵$\boldsymbol{M}\in\mathbb{R}^{m\times n}$（即$\boldsymbol{M}=\boldsymbol{D}\boldsymbol{\psi}^{-1}$）。对$\boldsymbol{M}$进行列子集选择操作得到最重要的$m$个列（详见3.4节），将其列下标加入集合$\Gamma$。

Step 5 根据集合Γ得到用于主动采样的掩码向量$\boldsymbol{mask}=[m_1,\cdots,m_n]^{\mathrm{T}}$，即当$j\in\Gamma$时，$m_j=1$，否则$m_j=0$（$1\leqslant j\leqslant n$）。

Step 6 将掩码向量$\boldsymbol{mask}$转换为自适应测量矩阵$\boldsymbol{\Phi}\in\mathbb{R}^{m\times n}$。转换的方式是：首先将$\boldsymbol{\Phi}$初始化为全零矩阵；然后依次取出$\boldsymbol{mask}$中$m$个非零元素，假设第$i$个非零元素（$1\leqslant i\leqslant m$）的下标为$j$（$1\leqslant j\leqslant n$），则将$\boldsymbol{\Phi}$中第$i$行第$j$列的元素值$\boldsymbol{\Phi}(i,j)$置为1。

3.4 矩阵的列子集选择

本节将详细阐释3.3节的Step 4中列子集选择的原理。首先给出矩阵的列子集选择问题（Column Subset Selection Problem，CSSP）的定义[32]。给定一个矩阵$\boldsymbol{M}\in\mathbb{R}^{m\times n}$与一个正整数$c$（$c$小于等于矩阵$\boldsymbol{M}$的秩$\rho$），在矩阵$\boldsymbol{M}$中选择$\boldsymbol{c}$列构成一个新矩阵$\boldsymbol{C}\in\mathbb{R}^{m\times c}$，使得两者的残差最小化。

$$\min_{C}\|\boldsymbol{M}-\boldsymbol{C}\boldsymbol{C}^{+}\boldsymbol{M}\|_{\mathrm{F}}^2\tag{7}$$

其中，$\|\cdot\|_{\mathrm{F}}$表示Frobenius范数，是矩阵各元素的平方和的开方值；$\boldsymbol{C}^{+}$代表矩阵$\boldsymbol{C}$的伪逆，$\boldsymbol{C}\boldsymbol{C}^{+}$表示矩阵$\boldsymbol{M}$的列空间投影矩阵；$\boldsymbol{M}-\boldsymbol{C}\boldsymbol{C}^{+}\boldsymbol{M}$表示求解$\boldsymbol{M}$中的每个向量在$\boldsymbol{M}$的列空间上的投影$\boldsymbol{C}\boldsymbol{C}^{+}\boldsymbol{M}$，再将其从$\boldsymbol{M}$中减去，得到的矩阵即为$\boldsymbol{M}$投影到其列空间的正交补空间上的部分。在最小化该公式的Frobenius范数后，可以保证选择的列向量$\boldsymbol{C}$能够最好地代表原矩阵中的信息。CSSP是组合优化任务，适用于从矩阵中选择一个小的但有代表性的列向量样本集。在$O(n^c)$的时间复杂度下，可以生成所有情况的矩阵$\boldsymbol{C}$，从而找到式(7)的最优解。但在实际应用场景中，该方案需要花费大量的时间，是不切实际的。对于该问题，现有的成果已可以在多项式时间内找到该问题的最佳逼近解[32-34]。由于对该算法的计算优化问题不是本文的关注点，故直接采用文献[32]中的方法进行求解，相关证明不再赘述。

具体来说，对于矩阵$\boldsymbol{M}\in\mathbb{R}^{m\times n}$，经过奇异值分解（Singular Value Decomposition，SVD）后可表示为：

$$\begin{aligned}\boldsymbol{M}&=\boldsymbol{U}_{\boldsymbol{M}}\boldsymbol{\Sigma}_{\boldsymbol{M}}\boldsymbol{V}\boldsymbol{M}^{\mathrm{T}}{}_{\boldsymbol{M}}\\&=(\boldsymbol{U}_c,\boldsymbol{U}_{\rho-c})\begin{pmatrix}\boldsymbol{\Sigma}_c&\boldsymbol{0}\\\boldsymbol{0}&\boldsymbol{\Sigma}_{\rho-c}\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_c^{\mathrm{T}}\\\boldsymbol{V}_{\rho-c}^{\mathrm{T}}\end{pmatrix}\\&=\sum_{i=1}^{\rho}\boldsymbol{\sigma}_i(\boldsymbol{M})\boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}}\end{aligned}\tag{8}$$

其中，矩阵$\boldsymbol{M}$的秩$\rho=rank(\boldsymbol{M})\leqslant\min(m,n)$；$\boldsymbol{U}_c\in\mathbb{R}^{m\times c}$和$\boldsymbol{U}_{\rho-c}\in\mathbb{R}^{m\times(\rho-c)}$包含左奇异向量$\boldsymbol{u}_i$（$1\leqslant i\leqslant\rho$）；$\boldsymbol{V}_c\in\mathbb{R}^{n\times c}$和$\boldsymbol{V}_{\rho-c}\in\mathbb{R}^{n\times(\rho-c)}$包含右奇异向量$\boldsymbol{v}_i$（$1\leqslant i\leqslant\rho$）；$\boldsymbol{\Sigma}\in\mathbb{R}^{\rho\times\rho}$是一个对角矩阵，包含了从大到小排列的奇异值$\sigma_1(\boldsymbol{M})\geqslant\cdots\geqslant\sigma_\rho(\boldsymbol{M})>0$。为了选择出最重要的$c$个列，可先通过式(9)计算矩阵$\boldsymbol{M}$中的所有列集合$\{1,\cdots,n\}$上的概率分布$P=\{p_1,\cdots,p_n\}$，$p_i\geqslant0$，$\sum_{i=1}^{n}p_i=1$，再提取概率最大的前$c$列为最终结果。

$$p_i=\frac{\frac{1}{2}\|(V_c)_{(i)}\|_2^2+\frac{1}{2}\|(\Sigma_{\rho-c}V_{\rho-c}^{\mathrm{T}})^{(i)}\|_2^2}{\sum_{j=1}^{n}\|(V_c)_{(j)}\|_2^2+\sum_{j=1}^{n}\|\Sigma_{\rho-c}V_{\rho-c}^{\mathrm{T}})^{(j)}\|_2^2}\tag{9}$$

举例来说，若求得的概率分布$P=\{0.21,0.32,0.01,0.11,0.09,0.26\}$，则当$c=3$时将选择$\boldsymbol{M}$中第一、二、六列。

3.5 稀疏基的选择

由于空气质量数据存在着时空相关性，同一站点时间尺度上又存在着平滑性，故使用以下稀疏基$\boldsymbol{\psi}$的逆，将非稀疏向量$\boldsymbol{X}$转换为稀疏向量$\boldsymbol{S}$。

$$\begin{aligned}\boldsymbol{S}=\boldsymbol{\psi}^{-1}\boldsymbol{X}&=\begin{bmatrix}2&-1&0&&0&0&0\\-1&2&-1&\cdots&0&0&0\\0&-1&2&&0&0&0\\&\vdots&&\ddots&&\vdots&\\0&0&0&&2&-1&0\\0&0&0&\cdots&-1&2&-1\\0&0&0&&0&-1&2\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\\vdots\\x_{n-2}\\x_{n-1}\\x_n\end{bmatrix}\\&=\begin{bmatrix}2x_1-x_2\\-x_1+2x_2-x_3\\-x_2+2x_3-x_4\\\vdots\\\vdots\\-x_{n-2}+2x_{n-1}-x_n\\-x_{n-1}+2x_n\end{bmatrix}\end{aligned}\tag{10}$$

其意义为：若向量$\boldsymbol{X}$中的某一个位置的采样值与其最相邻的两个位置采样值的均值相差不大，则向量$\boldsymbol{X}$经过变换后的向量$\boldsymbol{S}$中的元素基本为0或接近0。采用上述稀疏基的另一个原因是，文献[9]已证明该稀疏基与采用式(4)形式的测量矩阵具有较低的相关性。

3.6 自适应测量矩阵的在线更新

根据3.3节的步骤得到自适应测量矩阵$\boldsymbol{\Phi}$后，即可用于后续时刻的主动采样。但考虑到环境数据会随着时间的推移

发生变动，当前时刻使用的主动采样位置未必适合下一时刻，因此本文认为自适应测量矩阵应进行在线更新。具体做法为：当完成一次主动采样后，即将新采样向量 $\boldsymbol{Y}=[y_1,\cdots,y_m]^{\mathrm{T}}$ 加入训练集 $\mathbb{Y}^R$。新采样向量由真实值构成，并不存在任何推测误差。同时，将离当前时刻最远的训练数据删除（即以新换旧），以保证训练集大小固定不变。更新后的训练集将用于新一轮的字典学习，使得主动采样位置随着数据的时序变动进行自适应调整。

4 实验设计及结果分析

本文在中国两个城市的空气质量数据集上验证了所提出的自适应测量矩阵采样算法 AMM 的有效性。

4.1 数据集简介

本文所使用的数据集均来自 UCI（University of California at Irvine）机器学习库[35]。根据原始数据集的缺失情况与数据平稳特征，选用了两个相同时期不同地点仅包含少量缺失数据的子集进行实验。一个是 2014 年 6 月－ 8 月的上海美国领事馆提供的 PM2.5 数据子集，缺失率为 2.033%；另一个是 2014 年 6 月－8 月的北京东四站点的 PM2.5 数据子集，缺失率为 0.992%。两个数据子集均以小时为单位采集数据，对于极少的缺失值统一采用最近邻进行预补。考虑到空气质量数据存在一定的周周期性，本文实验将数据维度设置为 $n=7$ 天 $\times 24$ 小时 $=168$。最后，将数据集中前 4 周的数据作为训练集，剩余 8 周的数据作为测试集。

4.2 评价指标

本文实验通过计算未采样位置的估计值与其真实值之间的误差来衡量推测结果的优劣，如无真实值则不参与计算。评价指标为平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）和均方根误差（Root Mean Square Error，RMSE），计算公式如下：

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i}\right| \quad \text{s.t. } i\notin\Gamma \tag{11}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2} \quad \text{s.t. } i\notin\Gamma \tag{12}$$

其中，n 表示数据维度，x_i 表示真实值，$\hat{x}_i$ 表示经过推测算法得到的估计值，集合 Γ 是已采样点的下标集合。*MAPE* 衡量的是推测值和实际值偏离的相对大小情况，不容易受极端值的影响，*RMSE* 衡量的是推测值和实际值偏离的绝对大小情况。*RMSE* 采用误差的平方，会将绝对误差放大，所以对极端值更加敏感。

4.3 基准方法

（1）完全随机采样方法（Random）：根据给定的采样数量 m，在每一次采样时随机选择位置进行采样。由于是随机采样，该方法可保证每一条测试数据的采样位置是不同的。

（2）边缘保持分段随机采样方法（Epp-Random）[24]：每轮采样时，将所有采样点 n 分为 $K+2$ 段，其中，向量的首尾两段长度为 G_1 且为全采样，中间 K 段每段长为 K_i，对每段 K_i 进行随机采样 G_{2i} 个点，即总的采样个数为 $m=2G_1+\sum_{i=1}^{K}G_{2i}$。

（3）逐个贪心采样方法（Greedy）：利用训练数据经过 m 轮迭代逐个选择采样点。首先设置采样点的下标集合 $\Gamma=\varnothing$。在每轮迭代中，选择能够与之前采样点合作带来最小推测误差的点下标加入 Γ。该方法的特点是采样点下标集合 Γ 固定不变，即测试集的每条数据均采用相同的 m 个位置进行采样，不存在在线更新过程。

（4）委员会查询（Query by Committee，QBC）[36]方法：它维护一个预测模型集合（集合被称为委员会），通过选择预测差异最大的若干个数据点作为下一时刻的采样对象。具体步骤如下：

Step 1 选取两种时间序列预测模型，如自回归模型（Autoregressive model，AR）、移动平均模型（Moving Average，MA）来构建委员会。

Step 2 每个模型都基于预测算法训练集预测出下一时刻的完整序列。

Step 3 根据所有模型的预测结果，计算每个位置预测值的方差。

Step 4 取方差最大的 m 个位置作为下一次采样的位置点。

Step 5 利用采样值推测出完整序列。

Step 6 将推测后的序列加入预测算法训练集中。

Step 7 重复 Step 2－Step 6。

上述 4 种对比算法和本文提出的 AMM 算法均属于采样算法，为了保证比较的公平性，未采样位置的推测算法统一采用压缩感知算法，即每种方法在得到某条测试数据的采样点下标集合 Γ 后，均需先执行 3.3 节的 Step5 和 Step6 得到测量矩阵 $\boldsymbol{\Phi}\in\mathbb{R}^{m\times n}$。然后，根据 Γ 得到测试数据的采样向量 $\boldsymbol{Y}$，再利用 $\boldsymbol{Y}$ 和 $\boldsymbol{A}=\boldsymbol{\Phi\psi}$（$\boldsymbol{\psi}$ 的生成详见 3.5 节），根据式（2）利用 OMP 算法得到稀疏向量 $\hat{\boldsymbol{S}}$。最后，利用 $\hat{\boldsymbol{X}}=\boldsymbol{\psi}\hat{\boldsymbol{S}}$ 得到未采样位置的估计值。

4.4 实验结果与分析

为了降低采样成本，采样率越低越好。因此，为了衡量低采样率（低于 20%）下的采样效果，本文将采样率设置为 3%～17%，以 2%为步长进行多次实验。由于测试集为连续 8 周的数据，因此在给定的采样率下，每种方法均进行了 8 轮采样，每轮会给出一周 m 个需要采样的位置（m 为数据维度 168 * 采样率的向下取整值）。推测误差为 8 轮推测误差的平均值。为了实现公平比较，在给定的采样率下分别进行了 10 次独立实验以避免结果出现偏差，最后取平均值进行排名。表 1 和表 2 列出了上海 PM2.5 数据子集的实验结果，表 3 和表 4 列出了北京 PM2.5 数据子集的实验结果。

表 1 不同采样算法基于上海 PM2.5 数据子集的 MAPE

Table 1 MAPE of different sampling algorithms based on Shanghai PM2.5 data subset

采样率	Random	Epp-Random	Greedy	QBC	AMM
3%	58.922	60.804	57.847	68.033	**51.003**
5%	50.762	53.950	51.174	65.389	**44.293**
7%	48.675	45.665	48.643	56.957	**39.578**
9%	44.204	41.095	49.335	51.312	**37.211**
11%	41.932	37.888	47.514	50.604	**33.351**
13%	39.919	34.474	44.273	47.106	**31.989**
15%	38.499	33.449	44.384	45.768	**29.528**
17%	36.109	31.385	44.945	42.011	**28.089**

表 2 不同采样算法基于上海 PM2.5 数据子集的 RMSE

Table 2 RMSE of different sampling algorithms based on Shanghai PM2.5 data subset

采样率	Random	Epp-Random	Greedy	QBC	AMM
3%	0.313	0.233	0.253	0.289	**0.231**
5%	0.266	0.221	0.241	0.255	**0.218**
7%	0.239	0.196	0.229	0.235	**0.195**
9%	0.232	0.187	0.212	0.217	**0.184**
11%	0.196	0.162	0.209	0.198	**0.161**
13%	0.182	0.155	0.187	0.195	**0.151**
15%	0.178	**0.136**	0.187	0.196	0.148
17%	0.166	**0.133**	0.188	0.207	0.141

表 3 不同采样算法基于北京 PM2.5 数据子集的 MAPE

Table 3 MAPE of different sampling algorithms based on Beijing PM2.5 data subset

采样率	Random	Epp-Random	Greedy	QBC	AMM
3%	71.648	83.486	52.084	66.884	**51.730**
5%	65.881	71.068	49.839	59.045	**46.623**
7%	61.568	63.890	45.329	52.566	**38.815**
9%	55.864	55.755	45.305	50.123	**36.928**
11%	54.938	49.415	36.748	49.941	**33.262**
13%	53.141	43.043	40.214	47.525	**31.572**
15%	48.518	39.571	36.746	39.634	**32.383**
17%	47.309	37.503	33.28	41.985	**27.882**

表 4 不同采样算法基于北京 PM2.5 数据子集的 RMSE

Table 4 RMSE of different sampling algorithms based on Beijing PM2.5 data subset

采样率	Random	Epp-Random	Greedy	QBC	AMM
3%	0.269	0.251	0.267	0.251	**0.235**
5%	0.232	0.199	0.238	0.240	**0.189**
7%	0.211	**0.181**	0.187	0.217	0.185
9%	0.191	**0.162**	0.187	0.204	0.178
11%	0.178	**0.158**	0.168	0.219	0.160
13%	0.167	0.145	0.162	0.182	**0.140**
15%	0.156	**0.123**	0.149	0.164	0.132
17%	0.152	0.127	0.141	0.142	**0.126**

表 5 不同采样算法的排名统计

Table 5 Ranking statistics of different sampling algorithms

采样率	上海子集的 MAPE	上海子集的 RMSE	北京子集的 MAPE	北京子集的 RMSE
Random	3	4	5	4
Epp-Random	2	**1**	4	2
Greedy	4	3	2	3
QBC	5	5	3	5
AMM	**1**	2	**1**	**1**

根据表 1－表 4 的实验结果，可得到表 5 的排名统计。实验结果分析如下：

(1)在各种低采样率下，利用 AMM 方法进行采样再进行推测的平均绝对百分比误差要小于其他采样方法。由于推测算法相同，可知 AMM 方法的采样位置最好，带来的推测误差最小。排名第二的 Epp-Random 方法进行采样点选择的思想与 AMM 相当，均是保留原始采样信息的最大化，即基于推测精度最大化的采样位置点选择。但 Epp-Random 在分段后仍是基于随机进行选择，推测的精度仍会在一个区间内浮动，存在随机性。

(2)Epp-Random 在两个数据子集的综合排名位列第二，其优势在于：与 Random 不同，Epp-Random 固定了边缘采样个数$G_1=2$，同时进行的分段随机采样可以控制所有采样点间的最大采样间隔。而 Random 是完全随机采样，可能出现部分时间段密集采样，使得最大采样间隔较大，导致信息丢失。另一方面，边缘点的采样也会影响推测精度，Random 对于边缘点的采样不可控，导致性能差距较大。但需要指出的是，在 3%～7%的低采样率下，Epp-Random 的表现却不如 Random。出现该现象的原因在于：低采样率下的采样点个数非常少，Epp-Random 需要消耗掉边缘固定的 $2G_1$ 个采样点，因此中间段的采样个数非常少，使得其最大采样间隔大于 Random，导致性能下降。

(3)Greedy 算法在上海子集上的 MAPE 指标排名第四，不如 Random 方法；但从 RMSE 指标上看，则排名第三。Greedy 算法在北京子集上，无论是从 MAPE 指标还是 RMSE 指标上看，均优于 Random 和 QBC。这说明虽然 Greedy 算法在每轮采样中的采样位置是相同的，但是这些采样位置对于训练数据而言是能带来最小推测误差的最佳位置。因此，只要测试数据和训练数据的数据规律差异不大，其推测误差也会较小；但如果测试数据和训练数据的数据规律差异较大，Greedy 算法的推测误差就会显著上升。图 4 展示了上海数据子集的 PM2.5 随时间变化的情况，实线和虚线分别为训练集和测试集的时间序列。不难发现，上海数据子集的训练集数据变化趋势与测试集差异较大，Greedy 方法从训练数据中得到的最佳采样位置并不适用于测试数据，因此其在 MAPE 指标上不如 Random 方法。

(4)在多轮采样中，Random 方法、Epp-Random 方法、QBC 方法和 AMM 方法在每轮采样中的采样位置均不相同。QBC 方法虽然采用主动采样策略，但其推测误差是最大的。造成其性能低下的原因是，在进行多轮采样时，只有第一轮的预测模型是基于真实的历史数据训练得到的，从第二轮开始，都利用上一轮的采样值推测出完整序列加入预测算法训练集用于更新预测模型。由于推测值本身存在误差，因此，后续多轮采样的预测误差逐渐增大，这严重破坏了利用预测方差不确定性进行采样位置选择的有效性。所以 QBC 方法并不适用于多轮采样。而反观 AMM 方法，其测量矩阵在线更新过程中，加入训练集的是上一轮的采样真实值，无须进行未采样值的推测，所以不会造成推测误差对采样位置选择的干扰。

那么，在多轮采样过程中，是否有必要在线更新采样位置呢？为了验证采样位置在线更新的必要性，本文在两个数据子集中基于 AMM 方法进行了测量矩阵有/无在线更新的性能比较。若无在线更新，则每一轮均会使用相同的采样位置进行采样；若有在线更新，则每一轮的采样位置都是基于新一轮的字典学习和列选择得到的，采样位置是可变的。实验结果如图 5 和图 6 所示，可以看出，在多轮采样中，不同采样率下，采样位置更新的推测误差均比选择相同位置进行采样的推测误差小。这充分说明了采样位置在线更新的必要性。由于环境数据会随着时间的推移发生规律变化，当前轮使用的主动采样位置未必适合下一轮，因此必须对自适应测量矩阵进行在线更新以保证 AMM 算法的优越性。

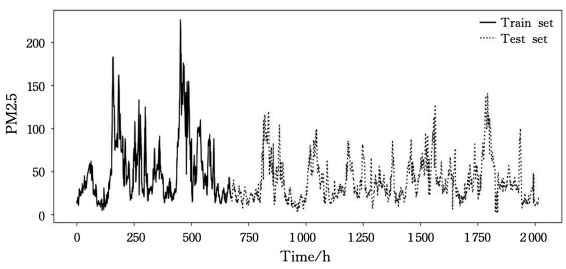

图 4 上海数据子集中训练集和测试集的数据变化趋势图

Fig. 4 Data change trend of training set and test set in Shanghai data subset

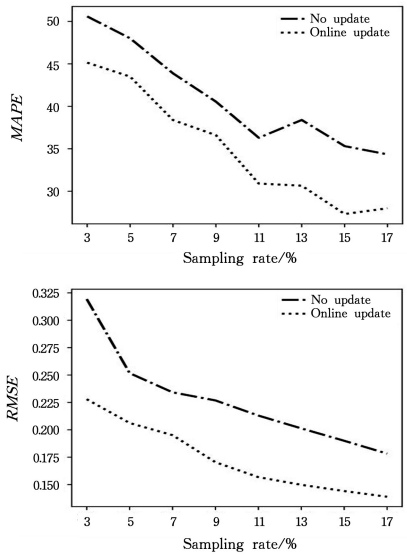

图 5 上海数据子集中 AMM 方法在多轮采样时采样位置有/无更新的比较

Fig. 5 Comparison of AMM methods in Shanghai data subset with/without update of sampling position in multi-round sampling

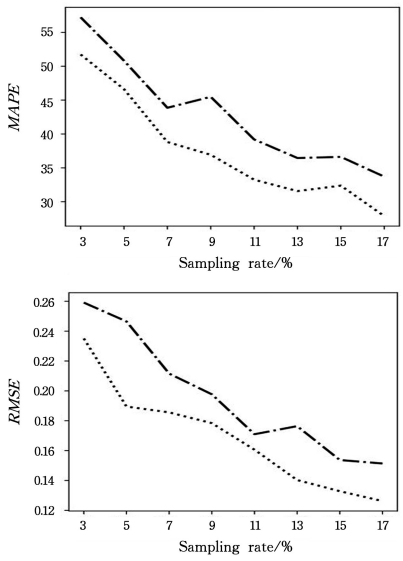

图 6 北京数据子集中 AMM 方法在多轮采样时采样位置有/无更新的比较

Fig. 6 Comparison of AMM methods in Beijing data subset with/without update of sampling position in multi-round sampling

结束语 本文主要研究了对城市中空气质量数据进行部分采集后推测其余未采样值的问题。在现有研究中，采样和推测一般采用不同算法实现，而本文提出了一种基于压缩感知一体化实现采样与推测的算法。该方法构建的测量矩阵不仅可以用于采样位置的选择，而且可以用于未采样数据的推测。将该方法与几种基线进行比较，实验结果表明，该方法选择的采样位置优于其他基线方法，可以获得最小的推测误差。在未来的工作中，我们将对该方法进行两个方面的研究。一是对时间尺度不平稳的数据进行稀疏基学习，以获得更好的推测效果；二是将该方法从向量数据推广到矩阵数据，通过捕捉时空信息，进一步提升城市空气质量推测效果。

参 考 文 献

[1] SHADDICK G, THOMAS M L, MUDU P, et al. Half the world's population are exposed to increasing air pollution[J]. NPJ Climate and Atmospheric Science, 2020, 3(1): 1-5.

[2] SOKHI R S, MOUSSIOPOULOS N, BAKLANOV A, et al. Advances in air quality research-current and emerging challenges [J]. Atmospheric Chemistry and Physics, 2022, 22(7): 4615-4703.

[3] YANG X, ZHANG Z. An attention-based domain spatial-temporal meta-learning(ADST-ML) approach for PM2.5 concentration dynamics prediction[J]. Urban Climate, 2023, 47: 101363.

[4] GUO B, WANG Z, YU Z, et al. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm[J]. ACM Computing Surveys(CSUR), 2015, 48(1): 1-31.

[5] JEZDOVIĆ I, POPOVIĆ S, RADENKOVIĆ M, et al. A crowdsensing platform forreal-time monitoring and analysis of noise pollution in smart cities[J]. Sustainable Computing: Informatics and Systems, 2021, 31: 100588.

[6] BALLATORE A, VERHAGEN T J, LI Z, et al. This city is not a bin: crowd mapping the distribution of urban litter[J]. Journal of Industrial Ecology, 2022, 26(1): 197-212.

[7] SHENG X, TANG J, ZHANG W. Energy-efficient collaborative sensing with mobile phones[C]// Proceedings of IEEE Conference on Computer Communications(INFOCOM). IEEE, 2012: 1916-1924.

[8] WANG L, ZHANG D, PATHAK A, et al. CCS-TA: Quality-guaranteed online task allocation in compressive crowdsensing [C]// Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015: 683-694.

[9] SONG X, GUO Y, LI N, et al. A novel approach for missing data prediction in coevolving time series [J]. Computing, 2019, 101(11): 1565-1584.

[10] BUDD S, ROBINSON E C, KAINZ B. A survey on active learning and human-in-the-loop deep learning for medical image analysis[J]. Medical Image Analysis, 2021, 71: 102062.

[11] DONOHO D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[12] TSAIG Y, DONOHO D L. Extensions of compressed sensing [J]. Signal Processing, 2006, 86(3): 549-571.

[13] CAI C, BAI E, JIANG X Q, et al. Simultaneous Audio Encryption and Compression Using Parallel Compressive Sensing and Modified Toeplitz Measurement Matrix[J]. Electronics, 2021,

10(23):2902.
[14] MENDELSON S,PAJOR A,TOMCZAK-JAEGERMANN N. Uniform uncertainty principle for Bernoulli and subgaussian ensembles[J]. Constructive Approximation,2008,28(3):277-289.
[15] HEGDE C,SANKARANARAYANAN A C,YIN W,et al. Numax:A convex approach for learning near-isometric linear embeddings[J]. IEEE Transactions on Signal Processing,2015,63(22):6109-6121.
[16] TROPP J A. A mathematical introduction to compressive sensing [J]. Bulletin of the American Mathematical Society,2017,54(1):151-165.
[17] XU K,LI Y,REN F. A data-driven compressive sensing framework tailored for energy-efficient wearable sensing[C]// Proceedings of 2017 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2017:861-865.
[18] LI S,ZHANG W,CUI Y,et al. Joint design of measurement matrix and sparse support recovery method via deep auto-encoder [J]. IEEE Signal Processing Letters,2019,26(12):1778-1782.
[19] HUIJBEN I A M,VEELING B S,VAN SLOUN R J G. Deep probabilistic subsampling for task-adaptive compressed sensing [C]// Proceedings of 8th International Conference on Learning Representations. ICLR,2020.
[20] SHAHRASBI B,RAHNAVARD N. Model-based nonuniform compressive sampling and recovery of natural images utilizing a wavelet-domain universal hidden Markov model [J]. IEEE Transactions on Signal Processing,2016,65(1):95-104.
[21] MALLOY M L,NOWAK R D. Near-optimal adaptive compressed sensing[J]. IEEE Transactions on Information Theory,2014,60(7):4001-4012.
[22] XIE K,LI X,WANG X,et al. Active sparse mobile crowd sensing based on matrix completion[C]// Proceedings of the 2019 International Conference on Management of Data. 2019:195-210.
[23] TANG G. Seismic data reconstruction and denoising based on compressive sensing and sparse representation [D]. Beijing:Tsinghua University,2010.
[24] CAO J J,XIAO J M,ZHU Y F,et al. Efficient shallow seismic-acquisition method based on compressed sensing theory[J]. Progress in Geophysics,2022,37(5):1920-1932.
[25] LIU W,WANG L,WANG E,et al. Reinforcement learning-based cell selection in sparse mobile crowdsensing[J]. Computer Networks,2019,161:102-114.
[26] WANG L,LIU W,ZHANG D,et al. Cell selection with deep reinforcement learning in sparse mobile crowdsensing[C]// Proceedings of 2018 IEEE 38th International Conference on Distributed Computing Systems(ICDCS). IEEE,2018:1543-1546.
[27] ZHANG Z,XU Y,YANG J,et al. A survey of sparse representation:algorithms and applications[J]. IEEE Access,2015,3:490-530..
[28] PATI Y C,REZAIIFAR R,KRISHNAPRASAD P S. Orthogonal matching pursuit:Recursive function approximation with applications to wavelet decomposition[C]// Proceedings of 27th Asilomar Conference on Signals,Systems and Computers. IEEE,1993:40-44.
[29] CHEN S S,DONOHO D L,SAUNDERS M A. Atomic decomposition by basis pursuit[J]. SIAM Review,2001,43(1):129-159.
[30] BLUMENSATH T,DAVIES M E. Iterative hard thresholding for compressed sensing[J]. Applied and Computational Harmonic Analysis,2009,27(3):265-274.
[31] 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.
[32] BOUTSIDIS C,MAHONEY M W,DRINEAS P. An improved approximation algorithm for the column subset selection problem[C]// Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete algorithms. Society for Industrial and Applied Mathematics,2009:968-977.
[33] PAN C T. On the existence and computation of rank-revealing LU factorizations [J]. Linear Algebra and Its Applications,2000,316(1/2/3):199-222.
[34] BOUTSIDIS C,DRINEAS P,MAGDON-ISMAIL M. Near-optimal column-based matrix reconstruction[J]. SIAM Journal on Computing,2014,43(2):687-717.
[35] LIANG X,LI S,ZHANG S,et al. PM2.5 data reliability,consistency,and air quality assessment in five Chinesecities[J]. Journal of Geophysical Research:Atmospheres,2016,121(17):10220-10236.
[36] BURBIDGE R,ROWLAND J J,KING R D. Active learning for regression based on query by committee[J]. Lecture Notes in Computer Science,2007,4881:209-218.

HUANG Weijie, born in 1999, postgraduate. His main research interests include machine learning and so on.

HUANG Fangwan, born in 1980, Ph.D, senior lecturer, is a member of CCF (No. D3015M). Her main research interests include computational intelligence, machine learning and big data analysis.

(责任编辑:何杨)