计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000116

城市空气质量数据的时空主动采样与联合推测

粮奥奇[1,2]　黄伟杰[1,2]　於志勇[1,2,3]　黄昉菀[1,2,3]
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 305108
3 大数据智能教育部工程研究中心　福州 350108
(1977881391@qq.com)

摘　要　当前,城市中的环境数据仍以固定站点作为主流采样方式,但高昂的全采样成本使其难以大规模扩展。在此背景下,通过局部采样并结合推测算法来推断其余未采样数据的方法成为了当前研究的热点。现有的研究通常使用两种不同的模型分别进行主动采样和缺失推测,存在计算成本高和误差易累积等不足。基于此,提出了一种时空主动采样与联合推测一体化模型(Spatiotemporal Active-sampling and Joint Inference,SAJI)。该模型不仅能选择带来高推测精度的采样站点,还可以确定其主动采样时刻,最后利用多测量向量(Multiple Measurement Vector,MMV)恢复算法联合推测出所有站点的缺失值。实验结果表明,相比于基线算法,SAJI可以充分利用时空相关性使得未采样站点获得有价值的预补值,并利用后续的联合推测算法在低采样率下获得最高的推测精度。

关键词:时空主动采样;时空相关性;遗传算法;压缩感知;联合推测
中图分类号　TP391

Spatiotemporal Active-sampling and Joint Inference of Urban Air Quality Data

LANG Aoqi[1,2],HUANG Weijie[1,2],YU Zhiyong[1,2,3] and HUANG Fangwan[1,2,3]
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
3 Engineering Research Center of Big Data Intelligence,Ministry of Education,Fuzhou 350108,China

Abstract　Currently,environmental data in cities are still sampled by fixed stations as the mainstream sampling method,but the high cost of full sampling makes it difficult to be scaled up on a large scale. In this context,the method of extrapolating the remaining unsampled data through local sampling and inference algorithm has become a hot topic in current research. Existing studies usually use two different models for active sampling and missing inference,respectively,which suffer from the shortcomings of high computational cost and easy accumulation of errors. Based on this,this paper proposes a spatiotemporal active-sampling and joint inference(SAJI) integration model. The model can not only select the sampling sites with high prediction accuracy,but also determine their own active sampling time. Finally,the missing values of all sites can be inferred jointly by using Multiple Measurement Vector(MMV) recovery algorithm. The experimental results show that compared with the baseline algorithms,SAJI can make full use of spatiotemporal correlation to obtain valuable prefilled values for the unsampled sites and achieve the highest inference accuracy using the subsequent joint inference algorithm at low sampling rates.

Keywords　Spatiotemporal active sampling,Spatiotemporal correlation,Genetic algorithm,Compressed sensing,Joint inference

1　引言

随着经济社会的发展,工业生产过程中产生的废气以及汽车普及产生的尾气加剧了城市空气质量的恶化。空气污染由于对环境、经济和人类健康的不利影响,已成为一项重大的全球挑战[1]。随着城市空气污染的加剧,对空气质量进行细粒度持续监测的需求愈发迫切[2]。

当前,文献[3-4]等通过设计合理分析时空相关性的模型来进行时空克里金任务,即在给定的时间段内,通过已采样位置的数据推测未采样位置的数据。文献[5-6]则通过输入历史时空数据,训练深度学习模型学习时空数据内在规律和模式,利用该模型预测未来时刻的时空数据。然而,这些方法都有一个共同点,即它们都是基于固定的监测点获取数据,通过获取到的数据来实现相应的推测。但是,一个值得深思的问

基金项目:国家自然科学基金(62332014);福建省促进海洋与渔业产业高质量发展专项资金(FJHYF-ZH-2023-02);福厦泉国家自主创新示范区协同创新平台项目(2022FX5)
This work was supported by the National Natural Science Foundation of China(62332014),Fujian Provincial Special Fund for Promoting High-Quality Development of Marine and Fishery Industry(FJHYF-ZH-2023-02) and Fuzhou-Xiamen-Quanzhou National Independent Innovation Demonstration Zone Collaborative Innovation Platform(2022FX5).
通信作者:黄昉菀(hfw@fzu.edu.cn)

题是，监测站的建设受到土地可用性和高昂维护成本的极大限制，导致固定监测点数量往往非常有限，并且分布不平衡，大多集中在人员密集的区域，难以获得大城市细粒度空气污染数据[7]。以北京市为例，其 36 个空气质量监测站点的分布如图 1 所示，可以看出，在中心城区站点分布十分集中，而在市区北部则只有零零散散几个站点。

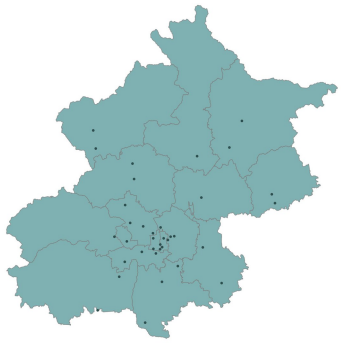

图 1 北京市空气监测站点分布图

Fig. 1 Distribution of air monitoring stations in Beijing

移动群智感知(Mobile Crowd Sensing，MCS)的兴起为解决这一问题提供了新的思路。MCS 将携带感知设备的个体作为数据感知的核心节点，这些个体在日常生活中的移动中自然地参与感知任务，可以有效补充甚至替代传统的固定式监测站点。这种模式为大规模区域(如社区乃至整个城市)的传感数据收集带来了前所未有的便捷性和效率提升。随着配备丰富传感器和强大计算能力的智能设备的广泛应用，MCS 已应用于各种城市规模的环境监测应用中，如空气质量监测[8-9]和噪声监测[10]。

感知数据的质量和成本是 MCS 的关键指标。为了监测城市的细粒度空气污染，需要收集不同地点的大量环境数据。感知数据的过程中使用的人力、设备和时间会造成大量的平台开销，这和平台的低成本设计目标并不符合。稀疏群智感知(Sparse MCS)的提出打开了新的方式，由于城市数据(如空气质量数据)的时空相关性，可以只选择几个最值得采样的位置/时刻收集数据，然后据此推断其余位置/时刻的数据，这样既降低了采集成本又能保证数据质量。此时，采样位置和采样时间的选择以及推测算法就显得尤为重要。

现有的研究通常使用两种不同的模型分别进行主动采样和缺失推测[11-12]。该做法存在两点不足：一是计算成本高，难以满足实时性任务要求；二是两个模型存在误差累积且无法保证优化目标的一致性。文献[13]提出了一种基于压缩感知的采样与推测一体化算法，该算法采用自适应测量矩阵(Adaptive Measurement Matrix，AMM)构建方法，通过字典学习技术和列选择方法来选择单个监测站点中最值得采样的时刻，特别适用于低采样率情况。该算法可在线更新，因此在多轮采样中仍能保持采样的有效性和推测的精度。遗憾的是，AMM 仅考虑了单一站点的采样时刻选择，并不适用于整个城市细粒度的全面监测。基于此，本文提出了一种时空主动采样与联合推断一体化模型(SAJI)，该模型不仅能选择带来高推测精度的若干个采样站点，还可以确定每个采样站点的主动采样时刻。通过采样位置和时刻的联合优化，在低采样率下可以获得较高的推测精度。本文的主要贡献包括：

(1)提出了一种局部站点主动采样、其余站点被动预补的策略。该策略首先通过元启发式算法和 AMM 算法选择能带来高推测精度的采样站点及其主动采样时刻，然后利用站点间的相关性，选择最相关的若干个站点中采样数量多的时刻进行未采样站点相应时刻的预补。

(2)将多站点的缺失值推测问题转化为压缩感知中的多测量向量(Multiple Measurement Vector，MMV)重构问题。根据所有站点的主动采样值或预补值，利用联合稀疏重加权算法进行所有站点的未采样/预补时刻的联合推测。

(3)北京空气质量数据集的实验结果表明，在低采样率下，本文模型 SAJI 可以充分利用时空相关性使得未采样站点获得有价值的预补值，并利用后续的联合推测算法获得最高的推测精度。

2 相关工作

站点选择是稀疏 MCS 应用中的基本问题。不同 MCS 系统的数据可能涉及不同的时空相关性，因此设计适当的站点选择策略是一项非常重要的任务。合适的站点选择策略和后续数据的推断有着直接关系。文献[14]中提出了一种基于联合训练的半监督学习方法，该方法能够通过少量的站点数据结合城市其他数据集(如路网、POI 等)找寻站点间时间空间上的关系从而推测其他站点的空气质量。文献[15]提出了一种基于时空多视点学习的方法，该方法致力于找寻站点间的空间关系和各个站点自身的时间关系，从时空和全局-局部的角度同时对一组地理传感读数中的缺失值进行集体填充。上述两种方法更关注的是站点的时空关系从而进行数据推测，而我们更关注的是采样站点的最优选择。文献[11]基于主动学习方案，在每次迭代中选择一定数量最有价值的位置收集数据，并为这些位置设置不同的激励，同时使用压缩感知推断未选择位置的数据，通过激励的方式，实现满足数据质量要求下的成本最小化。文献[16]介绍了一种基于成本的排序选择算法，该算法巧妙地利用历史采样数据构建矩阵，并通过对矩阵进行分解，构建了一个融合时间和位置的二分图模型。借助模型该算法直接选择下一时刻成本最低的位置进行采样，最后借助矩阵补全技术，算法能够恢复出完整的数据。这两种方法和本文方法的共同点是都更关注成本而非推测精度。文献[17-18]提出了一个框架，初始给定一个周期数据推断误差上界和要求达到误差上界的周期数 m，每一个周期迭代地选择多个位置进行采样，持续进行此过程直至当前周期内的数据满足预定的推断误差要求，然后进行下一周期的采样，最终确保至少 m 个周期的推理误差低于预定的界。这种方法旨在通过推断算法使得选择的站点数最少，但是无法准确知道初始的误差上界和周期数，并且每个周期迭代选择和误差评估效率过低。文献[19]提出了一种位置混淆机制，通过不确定性数据推断算法，在隐私保护的情况下选择推断数据方差最大的站点，本质上就是选择不确定性较高的站点，但并不保证这是全局最优的选择，可能会出现过度依赖于不确定性而陷入局部最优的问题，同时该方法更多地关注隐私保护问题。

当前，使用比较广泛效果较好的是基于训练的强化学习方法。文献[20-22]研究了在预先给定的推理误差约束下，如何使选择的站点数量最小化问题，提出了一种基于深度强化学习的最佳站点选择策略，该策略可以在充分的数据训练下

近似全局最优策略。文献[23]研究了如何在站点数约束下最小化推理误差，提出了一种基于用户轨迹预测和强化学习的用户招募算法，该算法在站点约束下选择最优用户采集数据从而提高数据推测精度。总而言之，这些基于训练的强化学习方法都是为了训练一个近似最优的站点选择模型。在数据充足的情况下，强化学习等模型能够表现出良好的性能，但对于缺少大量历史数据的应用场景，现有的方法适应度较差。并且这类模型对于计算成本要求较高，对于一些实时性的任务、有多轮执行周期的任务，难以保证结果的可靠性。

3 问题描述

本文以多站点空气质量指数(Air Quality Index，AQI)时间序列为例，将采样和推测问题形式化如下：假设有 P 个待采样站点，每个站点的待采样序列为 n，则总共需要采样的数据为 $n\times P$，已知采样率为 $R(0<R<1)$，则应当主动采样 $M=\lfloor n\times P\times R\rfloor$ 个数据，然后推测剩余的 $n\times P-M$ 个数据。以图 2 为例，总共 5 个待采样的站点在 4 个时刻共计需记录 20 个数据，若采样率为 40%，则应当采样的数据个数为 8 个。图中纵轴表示站点编号，横轴表示站点待采样的时刻。左图白色的圆圈代表不被选中的时刻，蓝色代表需要采样的时刻，而右图粉色代表利用已采样的蓝色时刻进行推测的结果。本问题旨在给定总体采样率的前提下，确定最优的采样策略(包括站点和时刻的选择)，以最小化对剩余数据的推测误差。总体采样率的计算式如式(1)所示：

$$M=\sum_{i=1}^{P}m_i,\ N=n\times P$$
$$\text{s.t. } R=\frac{M}{N} \tag{1}$$

其中，m_i 表示站点 i 的采样个数，各个站点的 m_i 可以是不同的；R 代表总体采样率；M 代表所有待采样站点采样个数的和；N 代表站点个数和采样时长的乘积。

在总体采样率 R 固定的情况下，主动采样方案可以有两种：一种是每个站点都参与采样；另一种是选择部分站点参与采样。考虑到 AQI 数据存在的时空相关性和中心城区密集的站点分布，本文更倾向于中心城区的部分密集站点可以不参与采样，而是利用相邻站点推测而得。这样做的好处是在总成本不变的情况下，其余站点的采样率得以提升，这对于偏远地区站点的数据推断是有帮助的。因为偏远地区站点与其他站点的时空相关性较弱，只能通过提升自身的采样率来帮助其未采样时刻的数据推测。

基于此，本文将多站点 AQI 时间序列的主动采样问题分为两阶段，第一阶段是进行采样站点的主动选择，第二阶段是对这些站点分别求解其主动采样时刻，目标是实现全局推测精度的最大化。

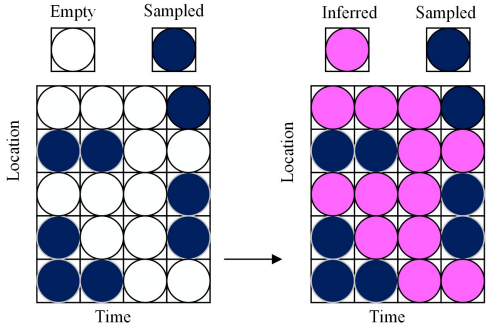

图 2 多维时序数据采样和推断图

Fig. 2 Sampling and inference diagram of multi-dimensional time series data

4 方法与实现

由于需要同时确定采样位置和采样时刻，因此本文设计了一种时空主动采样与联合推断一体化模型 SAJI，整体框架如图 3 所示。

首先是对采样站点的主动选择，该步骤使用元启发式算法探索哪些站点进行主动采样能够带来全局最优的推测准确率，接着使用 AMM 算法[13]为每个主动采样站点选择需要主动采样的时刻。然后利用站点间的时空相关性得到未采样站点需要预补的时刻，并进行预补。最后，根据所有站点的主动采样值或预补值采用多测量向量 MMV 联合推测方法[24]进行所有站点的未采样/预补时刻的推测。

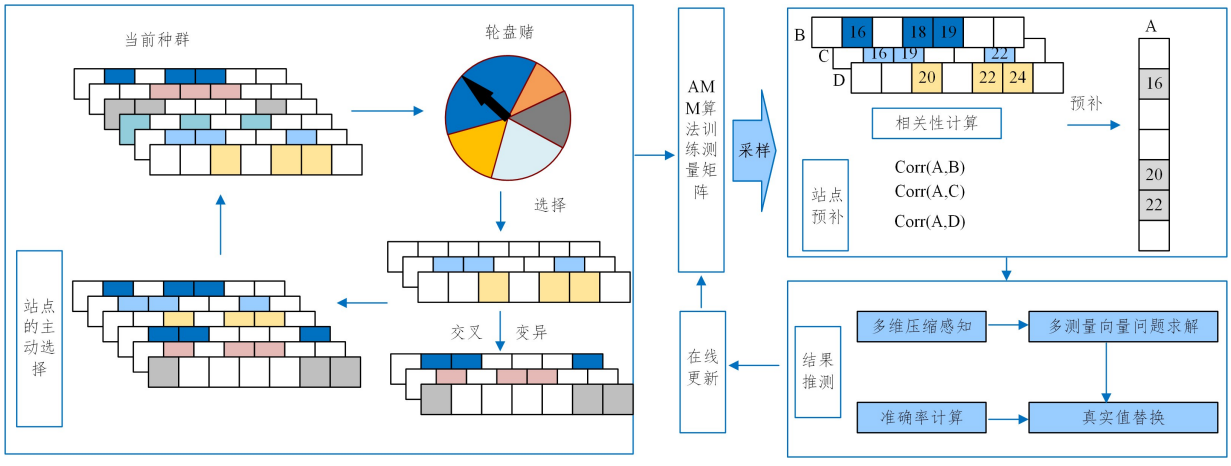

图 3 SAJI 的整体框架图

Fig. 3 Overall frame diagram of SAJI

4.1 站点的主动选择与结果推测

本节将阐述 SAJI 模型中的核心步骤，即如何选择采样的站点。若给定主动采样站点的个数 P_t，那么对于剩余的 $P-P_t$ 个不采样站点，总共有 $C_P^{P-P_t}$ 种组合，即有 $C_P^{P-P_t}$ 种解。单纯地计算各个解的站点组成并不需要耗时太多，但是对于其中的每一个解，均需要执行全局推测算法来评价解的好坏，计算代

价过高。此外，最佳采样站点的个数P_t也需要进行多次实验得出。该种暴力求解中积少成多的耗时将不足以支持实时性较强的主动采样任务。

元启发式算法如遗传算法对解决此类组合问题有着很强的适应性。在时间方面，遗传算法在选出一组候选解时可以并行地处理，大幅度减少了计算时间。在求解方面，由于算法能够在广泛的解空间中寻找最优解，所以遗传算法在本文的目标问题中有着不俗的解决能力，故本文选用该方法进行目标解的求解。具体步骤如下所示：

Step1 染色体的表示。若有 P 个待采样站点，遗传算法中的染色体即可使用长度为 P 的多进制字符串表示每个站点的状态。由于本文只关心站点的选择与不选择，因此只有两个状态，使用二进制即可，“1”表示选择采样，“0”表示不选择。

Step2 原始种群的选择。遗传算法原始祖先种群的选择可以基于完全随机，但是选择的方式会影响到算法收敛的速度，也会影响到种群的多样性，继而影响到解的质量。为此，本文将基于先验知识进行初始解的生成。具体而言，如果选择与其他站点高度相关的站点作为采样站点，利用该站点的采样值进行未采样站点的预补，将有效减小预补值与真实值之间的差异。此外，对于与其他站点无关联或关联较小的站点，应直接采样其真实值，而非使用误差较大的预补值。基于此，本文利用分组随机进行原始种群的初始化。分组过程如下：

假设已知各站点间的相关性，对第 i 个站点来说，能够得到其与其余站点的相关性数值之和Sum_i，计算方法如式(2)所示：

$$Sum_i=\sum_{j=1}^{P}Corr(i,j)\quad \text{s.t. } i\neq j \tag{2}$$

其中，$Corr(i,j)$为站点 i 和 j 的相关性度量值(计算方法详见第 4.2 小节)。利用以上计算方式，可以将所有站点的 Sum 值按从大到小排序，并且按照站点数量等分为 3 组：高相关组、中相关组、低相关组。初始种群中的每个解均在高相关和低相关组中分别随机选取 $P/6=P/3\times1/2$ 个站点组成。

Step3 种群适应度计算。对于上一步骤得到的种群结果，计算其中每一个解的适应度。由于本问题的目标为推测精度的最大化，因此遗传算法的适应度即为推测精度。适应度计算过程如下：对于种群中的一个解，从二进制字符串中可知哪些站点需要主动采样；接着利用 AMM 算法得到每个需要采样站点的主动采样时刻并进行采样；然后根据 4.3 节的预补策略得到所有未采样站点的预补值；最后利用 4.4 节的重构算法计算该解的推测精度，即可得到该解的适应度。

Step4 种群更新策略。假设当前种群的规模为$\mathbb{Q}$，在计算完当前种群所有解的适应度后，本文采取了一系列策略来保持种群的多样性和进化能力。

首先，本文使用轮盘赌保留上一代种群中的$\mathbb{Q}/2$个优质解。具体操作如下：根据每个解的适应度将其转化为一个轮盘区间，其中区间的长度与适应度成正比，即假设前 i 个解的适应度之和为$\mathfrak{Y}_i$，则第 i 个解所处的轮盘区间为$[\mathfrak{Y}_i/\mathfrak{Y}_{\mathbb{Q}},\mathfrak{Y}_{i+1}/\mathfrak{Y}_{\mathbb{Q}}]$，在每一次选择解时，均随机生成一个在$[0,1)$之间的浮点数，浮点数命中的区间即为本次选择的解。经过循环$\mathbb{Q}/2$后即可选出上一代种群中的优质解。

其次，为了引入新的多样性，另一半$\mathbb{Q}/2$个解则使用交叉和变异进行生成。在已使用轮盘赌选择的$\mathbb{Q}/2$解中，随机挑选两个解进行单点交叉和简单反转，生成一个后代解并将其加入当前种群。重复此过程$\mathbb{Q}/2$次，即可确保当前种群的规模仍然为$\mathbb{Q}$个解。

Step5 循环迭代 Step3－Step4，直至达到迭代次数上限。

Step6 在遗传算法计算结束后，即可得到最后一代种群，选择其中适应度最高的一个解作为最终解。根据该最终解，即可得知有多少数量的站点需要进行主动采样，而最终解的适应度即为 SAJI 的最终推测精度。

4.2 站点间的相关性计算

站点的主动选择需要依赖站点间的相关性，本节将介绍其计算方法。常用的相关性计算方法有两种，一种是基于站点间的距离，一种是基于数据相似度。

若采用基于站点间的距离，由于可以将地球近似为一个球体，对于球面上的两个点，可以使用半正矢距离进行计算。然而，出于地理或人为因素，某些站点物理意义上(即经纬度)的相近并不会使得数据分布大致相同。图 4 为北京市 36 个站点中 4 个站点 3 天 PM2.5 的真实采样数据，右下角为站点的相对位置，可以看出站点 14 即使与 9，10，20 邻近，但是部分数据分布却相差甚远，若简单地使用半正矢距离作为衡量相似度的标准，则误差会加大。

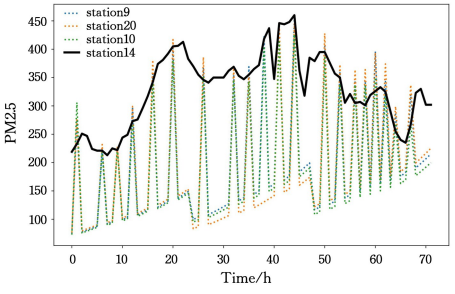

图 4 部分站点分布和数据分布图

Fig. 4 elected site distribution and data distribution maps

因此，本文采用另一种方法，利用站点的历史采样数据分布相似度。此类方法有很多种，如相关系数，其中皮尔逊相关系数[25]、斯皮尔曼相关系数[26]是最常用的方法。考虑到 AQI 数据并不能严格符合正态分布，不建议使用皮尔逊相关系数，建议采用斯皮尔曼相关系数 ρ 进行计算：

$$\rho=1-\frac{6\sum D^2}{n(n^2-1)} \tag{3}$$

其中，n 为数据的长度，D 表示两组数据对应位置上的数据的等级差，某个数的等级就是将所在的数组从小到大排列后，这个数所在的位置。斯皮尔曼相关系数越大越相关。

4.3 站点的预补策略

对于已采样的站点，从压缩感知的角度来说，测量矩阵需符合 AMM 算法[13]中所述形式，即测量矩阵 $\boldsymbol{\Phi}$ 的行数为采样数，列数为时刻数，其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，矩阵中非零的列下标即为主动采样的时刻点。

为了使用 MMV 方法进行联合推测，需要得到未采样站点部分时刻的预补值，最终得到与采样站点一样大小的测量向量。因此，本文提出了预补策略，具体步骤如下。

Step 1　假设待采样时刻为 10,采样率为 30%,则第 i 个站点采样到的数据为 $X_i \in \mathbb{R}^3$。进一步假设某个不采样的站点为 A,在已采样的站点中,可以找到与其相关性排名前 $k=3$ 名的站点为 B,C,D。

Step 2　单个站点采样时刻的选择可用 AMM 算法得到,此处不再赘述。假设已利用 AMM 算法得到站点 B,C,D 的测量矩阵,从测量矩阵即可看出哪些时刻被选中。即假设站点 B 的采样时刻分别为[1,2,5],站点 C 的采样时刻分别为[2,5,6],站点 D 的采样时刻分别为[2,6,9]。在 10 个待采样时刻中,每个时刻被采样的次数如表 1 所列。

表 1　每个时刻被采样次数

Table 1　Number of times sampled per moment

采样时刻	采样次数
时刻 1	1
时刻 2	3
时刻 5	2
时刻 6	2
时刻 9	1

Step 3　从多个分布相似的数据上来看,某个时刻被采样的次数多少可以反映该时刻所包含的信息比重,所以未采样的站点优先选择采样次数较多的时刻进行预补。从表 1 可得知,以上例子有 5 个时刻被选中采样,按次数多少进行从大到小排序即可得到 2,5,6,9,1。根据需采样个数为 3,则站点 A 的预补时刻取前 3 个时刻为[2,5,6],并且可以使用该信息构建站点 A 的测量矩阵。除此之外,站点 A 的采样值可以根据所对应的预补站点进行预补。举例来说,站点 A 在第 5 时刻需要进行预补。实际上,在第 5 时刻进行了真实采样的站点为 B 和 C。那么,站点 A 在第 5 时刻的预补值可以通过将站点 B 和 C 在该时刻的采样值根据相似度大小进行加权求和来计算,具体计算方式如下:

$$V_A' = \frac{Corr(A,B)}{Corr(A,B)+Corr(A,C)} \times V_B + \frac{Corr(A,C)}{Corr(A,B)+Corr(A,C)} \times V_C \tag{4}$$

其中,V_A' 为站点 A 的预补值,V_B 为站点 B 的采样值,$Corr(A,B)$ 为站点 A 和站点 B 的相关性数值。详细过程如图 5 所示。

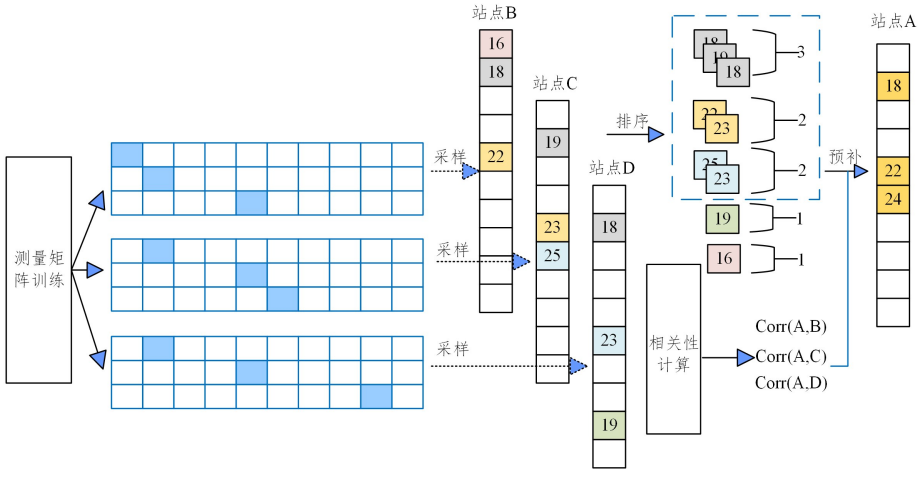

图 5　站点预补原理图

Fig. 5　Schematic diagram of pre-compensation for stations

4.4　MMV 重构算法

基于所有站点的采样值/预补值和测量矩阵,即可使用压缩感知进行缺失值的填补。压缩感知多维数据推测实际上为多测量向量 MMV 重构问题,如式(5)所示:

$$\boldsymbol{Y}=\boldsymbol{AS} \tag{5}$$

其中,$\boldsymbol{Y}=[y_1,\cdots,y_P]\in\mathbb{R}^{m\times P}$ 是从 P 个站点得到的部分采样值或预补值,$\boldsymbol{A}\in\mathbb{R}^{m\times n}$ 称为传感矩阵,而矩阵 $\boldsymbol{S}=[s_1,\cdots,s_P]\in\mathbb{R}^{n\times P}$ 表示重构的稀疏信号。

重构算法的实现方式不是本文的关注点,现有的方法如联合稀疏重加权算法(Joint Sparse Reweighted l_2/l_P,JS-Reweighted l_2/l_P)[24]已可满足使用,具体的算法流程如算法 1 所示。

算法 1　JS-Reweighted l_2/l_p

输入:采样矩阵 $\mathbf{Y}$,传感矩阵 $\mathbf{A}$

输出:信号的推测结果 $\mathbf{S}$

1. 初始化 k=1,$\mathbf{S}_k=\mathbf{A}^{\dagger}\mathbf{Y}$,q,ρ

2. 计算权重 $\mathbf{W}_{k+1}=\mathbf{I}_{P^{t-1}\times P_t}\otimes \mathrm{diag}(c_k[1]^{1-q/2},\cdots,c_k[m]^{1-q/2})$,其中,$c_k[i]=(\|\mathbb{S}[i]\|_2^2+\varepsilon)^{1/2}$

3. 更新稀疏信号,令 $\mathbf{B}_{k+1}=\boldsymbol{\Phi}\mathbf{W}_{k+1}$,$\mathbf{S}_{k+1}=\mathbf{W}_{k+1}\mathbf{B}_{k+1}^{H}(\mathbf{B}_{k+1}\mathbf{B}_{k+1}^{H}+\lambda\mathbf{I})^{-1}\mathbf{Y}$

4. 若 $\|\mathbf{S}_{k+1}-\mathbf{S}_k\|_2/\|\mathbf{S}_k\|_2<\rho$ 结束算法,否则 k=k+1,回到步骤 2。

在该联合稀疏重加权推测算法的输入方面,求解模型与式(5)相同,但是维度发生变化。具体来说,压缩矩阵变为 $\boldsymbol{Y}=[y_1^{\mathrm{T}},\cdots,y_P^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{mP\times 1}$(将矩阵转化成一个列向量),稀疏信号变为 $\boldsymbol{S}=[s_1^{\mathrm{T}},\cdots,s_P^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{nP\times 1}$,传感矩阵为 $\boldsymbol{A}=diag\{A_1,\cdots,A_P\}\in\mathbb{R}^{mP\times nP}$,其中,每个 $A_{P_i}=\Phi_{P_i}\boldsymbol{\Psi}$,$A_{P_i}\in\mathbb{R}^{m\times n}$,$\boldsymbol{\Phi}_{P_i}$ 为第 i 个站点的测量矩阵,$\boldsymbol{\Psi}$ 为稀疏基,和 AMM 算法[13]中使用的稀疏基相同。由此引出稀疏惩罚项如式(6)所示:

$$J^{(q,\epsilon)}(S)=\sum_{i=1}^{m}(\|\mathbb{S}[i]\|_2^2+\varepsilon)^{q/2},0\leqslant q\leqslant 1 \tag{6}$$

其中,$\mathbb{S}[\boldsymbol{i}]=[s_1(i),s_2(i),\cdots,s_P(i)]$ 表示 $s_1,s_2,\cdots,s_P$ 第 i 个元素的向量,$\|\mathbb{S}[i]\|_2$ 为式(7)所述形式:

$$\|\mathbb{S}[i]\|_2=(\sum_{p=1}^{P}|s_p(i)|^2)^{1/2} \tag{7}$$

ε 是一个很小的数字,目的是使惩罚式(6)更加有效。最后问题转换为求解式(8):

$$\min_{S}\|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2+\lambda J^{(q,\epsilon)}(S) \tag{8}$$

其中,λ 是与噪声相关的正则化参数,当处于无噪声时,λ 值为

一个极小值。

经过该算法计算后，即可得到每个站点的稀疏信号 $\boldsymbol{S}_p$（$p\in\{1,\cdots,P\}$），再经过式(9)计算，即可得到重构后的信号。

$$\boldsymbol{X}_p'=\boldsymbol{\Psi}\boldsymbol{S}_p \tag{9}$$

上式的稀疏基的逆矩阵 $\boldsymbol{\Psi}^{-1}$ 采用文献[13]的设置，如式(10)所示：

$$\boldsymbol{\Psi}^{-1}=\begin{bmatrix} 2 & -1 & 0 & & 1 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 1 & 0 \\ 0 & -1 & 2 & & 0 & 0 & 1 \\ & \vdots & & \ddots & & \vdots & \\ 1 & 0 & 0 & & 2 & -1 & 0 \\ 0 & 1 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 1 & & 0 & -1 & 2 \end{bmatrix} \tag{10}$$

4.5 在线更新

SAJI的在线更新与AMM算法[13]的在线更新方法一致，只在线更新被选中站点的主动采样时刻，不对主动采样站点进行重新选择。这样做的原因在于站点的相关性系数需要基于完整历史数据计算。当采样成本有限时，站点是无法做到全采样的。因此，本文在完成首次站点的主动选择后，只更新其主动采样时刻。

5 实验

本章在北京市的空气质量数据集上验证所提模型SAJI的有效性。

5.1 数据集

本文选用了2014年3月－5月北京36个监测站点的PM2.5数据集[27-28]。该数据集以小时为步长采集数据，使用的数据段原始缺失率为3.375%。对于初始值缺失的时刻，采用24小时内的均值预补。同样地，考虑到PM2.5存在一定的周期性，本文实验将数据维度设置为 $n=7$ 天$\times 24$ 小时$=168$。训练集选用数据段的前4周，剩余8周作为测试集。

5.2 评价标准

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)。计算式如下：

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i}\right| \tag{11}$$

$$\text{s.t. } i\notin\Gamma$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2} \tag{12}$$

$$\text{s.t. } i\notin\Gamma$$

其中，n 表示数据维度，x 表示真实值，$\hat{x}$ 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，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，但RMSE采用误差的平方，会将绝对误差放大，所以对于极端值更加敏感。

5.3 基线方法

在基线方法方面，为说明本文所采用的站点选择策略的有效性，将与以下方法进行对比：

(1)自适应测量矩阵方法(AMM)[13]：将采样个数平分给所有站点，即所有站点都参与采样。每个站点均使用AMM方法进行主动选择采样时刻，并对每个站点单独推测。

(2)自适应测量矩阵联合恢复方法(Multi-AMM)：在站点的时刻选择方面与AMM方法一致。在推测阶段，使用联合稀疏重加权算法[24]对所有站点联合推测。

(3)边缘保持分段随机采样联合恢复方法(Multi-ERandom)：在站点的时刻选择方面使用文献[28]的边缘保持分段随机采样方法(Epp-Random)，该方法在每轮采样时，将所有采样点 n 分为 $K+2$ 段，其中，首尾两段进行全采样，中间 K 段进行随机采样。在推测阶段，使用联合稀疏重加权算法[24]对所有站点联合推测。

(4)随机采样方法(Random)：在SAJI方法中的站点选择步骤中，不使用遗传算法进行站点选择，而是随机选择站点进行主动采样。其余步骤和SAJI方法保持一致。

(5)贪心采样方法(Greedy)：在SAJI方法中的站点选择步骤中，不使用遗传算法进行站点选择，而是选择与其余站点相关性较高的站点进行主动采样。其余步骤和SAJI方法保持一致。

(6)基于遗传算法的主动采样方法(GA)：对于所有站点，利用遗传算法基于完整的历史采样数据进行训练，旨在得到一个二维矩阵，该矩阵将决定每个站点的主动采样时刻。该方法和SAJI的不同之处在于SAJI只有部分站点参与采样，而GA是所有站点都参与采样。GA的具体流程为：

Step 1 利用大小为站点数量×待采样时刻数量的二维矩阵进行染色体的表示(在当前实验设置中矩阵大小为36×168)，矩阵内的数值为0或者1，代表该时刻是否进行主动采样。

Step 2 随机生成一组染色体作为初始解，由于每个站点的采样个数均相同，因此染色体矩阵中的每一行“1”的个数均相等，采样个数取决于当前采样率大小。

Step 3 对当前所有解进行计算适应度。具体来说，根据染色体的矩阵信息，即可得知每个站点需要采样的时刻；然后经过主动采样后，即可利用联合稀疏重加权算法进行未采样时刻的推测；最后利用推测误差作为该解的适应度。

Step 4 对计算适应度后的解集合进行选择、交叉和变异生成下一组解。

Step 5 重复Step 3－Step 4直至达到迭代次数上限。

Step 6 在训练结束后，利用以上步骤得到的最佳解进行后续的多轮采样并且计算误差。

需要说明的是，为避免推测算法带来的差异性，选用了单独推测的AMM算法和联合推测的AMM算法进行比较。与此同时，选用了文献[13]中表现较好的Epp-Random方法进行综合比较。另外，本文提出的SAJI方法和GA方法采用相同设置，种群规模设置为12，变异率和交叉率都为50%，迭代次数为1000次。

5.4 实验结果分析

将采样率设置在10%～30%，以5%为步长进行多次实验。由于测试的采样轮数为8次，因此在给定的采样率下，每种采样方法均进行了8轮连续采样。此外，为尽可能消除随机性带来的结果偏差，每种采样率下的每一种采样方法，均进

行了 10 次的独立实验后取平均值作为结果。表 2 为每种采样方法的 MAPE 结果，表 3 为每种采样方法的 RMSE 结果。实验结果分析如下：

(1)在低采样率下，与其他方法相比，SAJI 方法能够带来最高的推测精度。这说明，在采样成本有限的情况下，与其让每个站点都采集少量的值，不如将采样成本集中在若干个能带来高推测精度的站点上。这可以使得在相同的采样率下，主动采样的部分站点有着更多的采样次数，因此在低采样率时，可以获得比所有基线方法更高的推测精度。

(2)比较 AMM 算法和 Multi-AMM 算法的结果可以看出，在推测算法方面，当采样率在 15%～30%的时候联合推测算法的性能优于单独推测。这是因为联合推测算法能够更充分地利用多个站点的信息，并且能够减少可能存在的测量过程噪声干扰。即使某个测量向量受到干扰，其他测量向量仍能提供有用信息，增强整个推测系统的鲁棒性。但是在固定采样率为 10%的时候我们发现一个问题，即 AMM 的性能要优于联合推测算法，这主要是在采样率过低的情况下，将采样个数平分给所有站点，致使每个站点可采样的时刻数量很少，此时进行联合推测算法反而会引入一定的噪声，对最终结果产生影响。在这种情况下，AMM 只对单一站点的数据进行独立推测，虽然信息量有限，但避免了联合过程中可能引入的额外误差。

(3)Greedy 算法结果最差的原因在于该算法将相关性最强的一组站点作为最终解进行主动采样，而相关性较差的站点由于数据分布和其余站点相差较大，导致预补结果变差，直接使得最终推测结果劣化。

(4)GA 算法能够得到相对较优的结果，得益于其作为启发式算法能够在该主动采样问题巨大的解空间中进行多次探索，得到基于当前数据分布最佳的采样点，但是大量的求解导致该算法无法满足实时性任务的要求。并且，在多轮采样中，由于采样得到的数据并非完整数据，GA 算法也难以进行实时更新，只能主动采样相同的时刻，所以对异常情况难以处理。

表 2　不同采样算法基于北京 PM2.5 数据集的 MAPE

Table 2　MAPE of different sampling algorithms based on Beijing PM2.5 dataset

采样率/%	AMM	Multi-AMM	Multi-ERandom	Random	Greedy	GA	SAJI
10	44.04	51.80	50.70	58.60	65.41	49.28	**36.43**
15	39.57	36.67	37.23	47.75	48.94	38.52	**29.93**
20	34.19	30.92	30.75	40.79	42.84	29.89	**26.33**
25	31.94	25.96	23.16	35.28	36.41	25.57	**22.32**
30	32.14	21.91	27.25	32.46	32.03	23.56	**20.44**

表 3　不同采样算法基于北京 PM2.5 数据集的 RMSE

Table 3　RMSE of different sampling algorithms based on Beijing PM2.5 dataset

采样率/%	AMM	Multi-AMM	Multi-ERandom	Random	Greedy	GA	SAJI
10	28.57	30.61	27.92	34.45	39.98	27.97	**17.37**
15	27.01	22.16	24.42	29.29	29.71	22.15	**14.33**
20	23.61	21.26	22.98	25.98	26.70	19.17	**13.40**
25	19.62	15.07	17.08	22.79	23.70	17.49	**12.17**
30	18.82	16.18	25.58	21.26	20.92	15.51	**11.31**

5.5　消融实验

为了验证本文提出的基于遗传算法的站点选择策略的优势，我们将站点选择策略替换为贪心策略和随机策略，前者选择与其余站点相关性较高的站点进行主动采样，后者随机选择站点进行采样。实验对比如图 6 和图 7 所示，可以看出，在 10%～30%的不同采样率范围内，无论是 RMSE 还是 MAPE，本研究提出的基于遗传算法的站点选择策略均展现出最高的精度。当替换为贪心策略和随机策略时，预测精度均出现显著下降。值得注意的是，在 10%～15%采样率下，与 30%采样率相比，贪心策略和随机策略的精度大幅降低，相比之下本文方法降低幅度较小。由此可知本文方法在所有采样率下均表现出较低的误差且性能更为稳定。

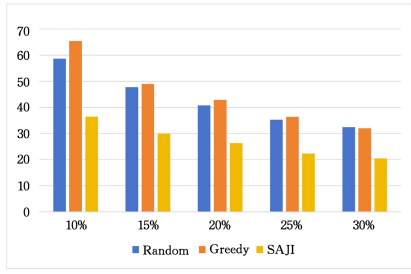

图 6　不同站点选择策略的 MAPE 结果对比

Fig. 6　Comparison of MAPE results for different site selection strategies

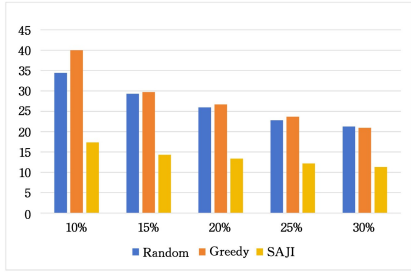

图 7　不同站点选择策略的 RMSE 结果对比

Fig. 7　Comparison of RMSE results for different site selection strategies

5.6　可视化分析

为了研究进行主动采样的站点的地理分布以及斯皮尔曼相关系数对于遗传算法的影响，本文做了两个可视化分析。图 8 展示了以斯皮尔曼相关系数作为站点间相关性度量所得到的遗传算法最终解的站点分布，无序号标识的黑点代表主动选择的站点，其余颜色代表未采样站点，站点间的连线代表预补的来源站点。在这个解中，共有 20 个主动采样站点和 16 个未采样站点。

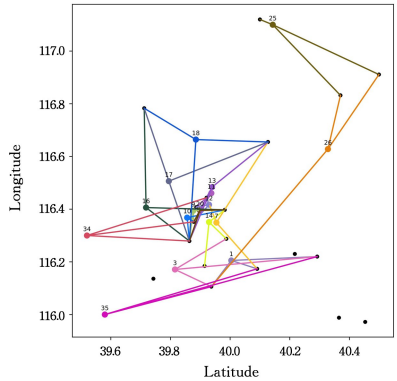

图 8　遗传算法解图示

Fig. 8　Diagram of genetic algorithm solution

需要注意的是，个别站点的预补并不一定与其最邻近的站点有关。通过使用相关系数捕捉站点数据分布间的相似性，能够更好地得到高精度推测值。通过图 8 可观察到，本文算法针对右下角的边缘站点采取了积极的采样策略，这一举措主要归因于这些站点与其他站点相关性较低，难以通过其他站点进行推测。因此，对这些站点进行主动采样对于提升数据推断的整体准确性显得尤为重要。相比之下，对于与其他站点具有较高相关性的站点，本文算法则实施了部分采样，并利用这些与其高度相关的站点进行预补。这种策略在确保整体采样率保持稳定的同时，实现了数据主动采集的高效性，从而进一步优化了数据采集的过程。

本文还可视化了斯皮尔曼相关系数对于遗传算法收敛速度的影响，如图 9 所示。图中横轴表示迭代次数，纵轴表示该次迭代所对应的种群中最优个体的适应度。在迭代早期，斯皮尔曼相关系数显示最优个体的适应度已经接近整体最优水平。这一发现不仅彰显了遗传算法在解决复杂优化问题时的卓越效率与强大能力，同时也揭示了斯皮尔曼相关系数作为一种先进的统计工具，在优化算法收敛速度、性能趋势方面所展现出的独特优势与巨大潜力。

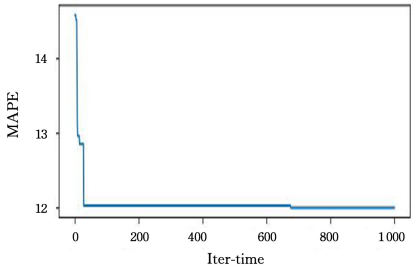

图 9 遗传算法收敛图

Fig. 9 Convergence diagram of the genetic algorithm

结束语 本文提出了一种应用于多维时间序列场景的时空主动采样与联合推测一体化模型 SAJI。该模型首先进行时空主动采样，既选择了需要主动采样的站点，又确定了这些站点的主动采样时刻；然后利用时空相关性对未采样站点的关键时刻进行预补；最后利用 MMV 重构算法进行全局联合推测。北京空气质量数据集的实验结果表明 SAJI 在具有时空相关性的数据上表现出良好的性能，这要归功于站点之间呈现的数据分布关联性使得未采样站点能够获得有价值的预补值，在后续的联合推测步骤中可以获得较高的推测精度。在未来的工作中，除了可以进一步优化联合推测算法之外，还可以考虑如何实时更新站点间的相关性。

参 考 文 献

[1] FENG T, SUN Y, SHI Y, et al. Air pollution control policies and impacts: A review[J]. Renewable and Sustainable Energy Reviews, 2024, 191: 114071.

[2] SOKHI R S, MOUSSIOPOULOS N, BAKLANOV A, et al. Advances in air quality research-current and emerging challenges [J]. Atmospheric Chemistry and Physics Discussions, 2021, 2021: 1-89.

[3] HU J, LIANG Y, FAN Z, et al. Graph Neural Processes for Spatio-Temporal Extrapolation[C]// Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023: 752-763.

[4] WU Y, ZHUANG D, LABBE A, et al. Inductive graph neural networks for spatiotemporal kriging[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 4478-4485.

[5] WU Z, PAN S, LONG G, et al. Graph wavenet for deep spatial-temporal graph modeling[J]. arXiv: 1906. 00121, 2019.

[6] ROTH A, LIEBIG T. Forecasting unobserved node states with spatio-temporal graph neural networks[C]// 2022 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2022: 740-747.

[7] TIAN Y, JIANG Y, LIU Q, et al. Temporal and spatial trends in air quality in Beijing[J]. Landscape and urban planning, 2019, 185: 35-43.

[8] LIU T, ZHU Y, YANG Y, et al. Incentive design for air pollution monitoring based on compressive crowdsensing[C]// 2016 IEEE Global Communications Conference (GLOBECOM). IEEE, 2016: 1-6.

[9] PAN Z, YU H, MIAO C, et al. Crowdsensing air quality with camera-enabled mobile devices[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017: 4728-4733.

[10] XU Y, ZHU Y, QIN Z. Urban noise mapping with a crowd sensing system[J]. Wireless networks, 2019, 25: 2351-2364.

[11] LIU T, ZHU Y, YANG Y, et al. ALC2: When active learning meets compressive crowdsensing for urban air pollution monitoring[J]. IEEE Internet of Things Journal, 2019, 6(6): 9427-9438.

[12] ZHU K, ZHANG A, NIYATO D. Cost-effective active sparse urban sensing: Adversarial autoencoder approach[J]. IEEE Internet of Things Journal, 2021, 8(15): 12064-12078.

[13] HUANG W J, GUO X W, YU Z Y, et al. Active Sampling of Air Quality Based on Compressed Sensing Adaptive Measurement Matrix [J]. Computer science, 2024, 51(7): 116-123.

[14] ZHENG Y, LIU F, HSIEHH P. U-air: When urban air quality inference meets big data[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 1436-1444.

[15] YI X, ZHENG Y, ZHANG J, et al. ST-MVL: Filling missing values in geo-sensory time series data[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016.

[16] XIE K, LI X, WANG X, et al. Active sparse mobile crowd sensing based on matrix completion[C]// Proceedings of the 2019 Inernational Conference on Management of Data. 2019: 195-210.

[17] WANG L, ZHANG D, YANG D, et al. SPACE-TA: Cost-effective task allocation exploiting intradata and interdata correlations in sparse crowdsensing[J]. ACM Transactions on Intelligent Systems and Technology(TIST), 2017, 9(2): 1-28.

[18] 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.

[19] WANG L, ZHANG D, YANG D, et al. Differential location privacy for sparse mobile crowdsensing[C]// 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016: 1257-1262.

[20] LIU W, WANG L, WANG E, et al. Reinforcement learning-

based cell selection in sparse mobile crowdsensing[J]. Computer Networks,2019,161:102-114.

[21] LIU W,YANG Y,WANG E,et al. Multi-dimensional urban sensing in sparse mobile crowdsensing[J]. IEEE Access,2019,7:82066-82079.

[22] WANG L,LIU W,ZHANG D,et al. Cell selection with deep reinforcement learning in sparse mobile crowdsensing[C]// 2018 IEEE 38th International Conference on Distributed Computing Systems(ICDCS). IEEE,2018:1543-1546.

[23] LIU W,YANG Y,WANG E,et al. User recruitment for enhancing data inference accuracy in sparse mobile crowdsensing[J]. IEEE Internet of Things Journal,2019,7(3):1802-1814.

[24] DING Y,RAO B D. Joint dictionary learning and recovery algorithms in a jointly sparse framework[C]// 2015 49th Asilomar Conference on Signals, Systems and Computers. IEEE, 2015: 1482-1486.

[26] BENESTY J,BENESTY J. Speech Enhancement Via Correlation Coefficients[J]. Fundamentals of Speech Enhancement, 2018:45-64.

[27] SCHOBER P,BOER C,SCHWARTEL A. Correlation coefficients:appropriate use and interpretation[J]. Anesthesia & Analgesia,2018,126(5):1763-1768.

[28] ZHENG Y,YI X,LI M,et al. Forecasting fine-grained air quality based on big data[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015:2267-2276.

LANG Aoqi, born in 2001, undergraduate, is a member of CCF (No. U8575G). 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.