本地差分隐私下面向离群点的真值发现算法研究

朱伊波[1],方贤进[1,2],张朋飞[1,3*],孙 笠[4],姜 茸[3]
(1. 安徽理工大学计算机科学与工程学院,安徽淮南 232001;2. 安徽理工大学煤炭无人化开采数智技术全国重点实验室,安徽淮南 232001;3. 云南财经大学云南省服务计算重点实验室,云南昆明 650221;4. 华北电力大学控制与计算机工程学院,北京 102206)

摘 要: 近年来,随着智能移动设备的普及和强大的传感能力,移动群智感知(Mobile CrowdSensing,MCS)已成为大规模感知城市动态的一种有潜力的技术.MCS中一个核心问题是如何从众多工人提交的嘈杂的感知数据中发现"真值".同时,真值发现过程中不可避免地面临隐私泄露问题.为应对这一挑战,研究者通常结合本地差分隐私(Local Differential Privacy,LDP)技术,通过对工人数据添加随机噪声实现隐私保护.然而,由于拉普拉斯分布的随机性和无界性,可能会注入大量噪声,从而产生离群点.此外,现有研究往往未能充分建模为满足LDP保护而注入的拉普拉斯噪声,导致求得的"真值"精度低,且现有的真值发现方法通常仅适用于离散值或无法严格满足LDP约束的问题.针对上述问题,本文提出一种基于LDP的面向离群点的真值发现算法LEADER.该算法首先对工人提交的数据添加拉普拉斯噪声,以确保工人隐私不被泄露.然后针对离群点问题,采用Huber损失函数作为度量距离,降低离群点对真值估计结果的影响.最后通过引入数据度量方法,优化工人和任务重要性权重分配,并根据提交值之间的相似性对工人进行分组,从而有效保护工人隐私的同时提高估计"真值"的精度.理论分析表明,LEADER算法在严格满足LDP约束的前提下,能够有效处理连续型数据,并实现高精度的真值发现.此外,与非隐私下的真值发现方法相比,LEADER算法在通信开销和计算开销方面保持相近.在两个真实数据集和一个合成数据集上的实验结果表明,LEADER算法的表现显著优于现有对比算法,噪声"真值"精度提升了至少18%.

关键词: 移动群智感知;真值发现;隐私保护;本地差分隐私
基金项目: 云南省服务计算重点实验室开放课题(No.YNSC24116);安徽省科技重大专项(No.18030901025);国家自然科学基金(No.61572034,No.62202164)
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112(2025)05-1541-18
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250021

A Study of Truth Discovery Algorithms for Forward Outliers Under Local Differential Privacy

ZHU Yi-bo[1], FANG Xian-jin[1,2], ZHANG Peng-fei[1,3*], SUN Li[4], JIANG Rong[3]
(1. *School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, Anhui* 232001, *China*; 2. *State Key Laboratory of Digital Intelligent Technology for Unmanned Coal Mining, Anhui University of Science and Technology, Huainan, Anhui* 232001, *China*; 3. *Yunnan Key Laboratory of Service Computing, Yunnan University of Finance and Economics, Kunming, Yunnan* 650221, *China*; 4. *School of Control and Computer Engineering, North China Electric Power University, Beijing* 102206, *China*)

Abstract: In recent years, with the widespread adoption of intelligent mobile devices and their powerful sensing capabilities, mobile crowdsensing (MCS) has emerged as a promising method for large-scale sensing of urban dynamics. A key challenge in MCS is discovering the truth from the noisy sensory data submitted by numerous workers. However, the process of truth discovery inevitably raises privacy concerns. To address these challenges, researchers frequently integrate local differential privacy (LDP) techniques by adding random noise to workers' data for privacy protection. Nonetheless, the randomness and unbounded nature of Laplace noise may inject excessive noise, resulting in outliers. Additionally, existing research often fails to adequately model the Laplace noise injected to satisfy LDP protection, resulting in low truth accuracy. Moreover, the current truth discovery methods are typically only applicable to discrete data, or cannot strictly satisfy the LDP constraints. To address the above issues, this paper proposes LEADER, an outlier-oriented truth discovery algorithm

收稿日期:2025-01-06;修回日期:2025-04-03;责任编辑:覃怀银
*通讯作者:张朋飞

under LDP. First, the algorithm adds Laplace noise to workers' data to ensure privacy protection. Second, it addresses outliers by adopting the Huber loss function to measure distances, mitigating their impact on truth estimation. Finally, through a data-driven metric approach, the algorithm optimizes the weight allocation for worker and task importance and groups workers based on the similarity of their submitted values. These enhancements enable LEADER to improve the accuracy of estimated truths while maintaining privacy protection. Theoretical analysis demonstrates that LEADER strictly satisfies LDP constraints, effectively handles continuous data, and achieves high-accuracy truth discovery. Furthermore, compared to non-private truth discovery methods, the LEADER algorithm maintains comparable communication and computational overhead. Experimental results on two real-world datasets and a synthetic dataset indicate that the LEADER algorithm significantly outperforms existing methods, achieving at least an 18% improvement in the accuracy of the noisy truth.

Key words: mobile crowdsensing; truth discovery; privacy protection; local differential privacy

Foundation Item(s): Foundation of Yunnan Key Laboratory of Service Computing (No.YNSC24116); Anhui Provincial Major Science and Technology Project (No.18030901025); National Natural Science Foundation of China (No.61572034, No.62202164)

1 引言

近年来,随着智能手机、平板电脑和汽车等搭载多个嵌入式传感器的智能移动设备的广泛普及,移动群智感知(Mobile CrowdSensing,MCS)得到了广泛研究和应用[1,2]. 在MCS中,由于主观(如恶意工人故意上传误导数据)或客观因素(如设备差异和环境影响),上传的感知数据可能不准确,导致数据噪声或矛盾,进而影响数据质量[3]. 因此,真值发现方法应运而生,旨在从质量参差不齐的感知数据中提取"真值",提高数据质量[4]. 其核心思想是通过评估工人提交的数据质量,并基于工人权重和感知数据的可信度进行迭代建模,从而获得更准确的聚合结果[5]. 具体来说,真值发现方法的核心思想是如果某工人提交的感知数据更接近"真值",则该工人的权重更高. 因此,高权重工人的数据在聚合过程中占更大的比重,从而有效过滤噪声和错误数据,提高MCS数据的可靠性和实用性.

尽管传统真值发现方法已被证明是有效的,但其迭代计算过程需要中央服务器收集每个工人的原始感知数据,这可能对MCS工人的隐私构成威胁. 隐私泄露不仅可能导致工人因担忧隐私风险而拒绝参与任务,还可能促使工人为规避隐私风险而上传虚假数据,从而严重影响数据质量,甚至对医疗领域的数据可靠性产生不利影响[6~8]. 因此,在MCS中保护工人隐私至关重要.

为应对这一问题,研究者通常采用本地差分隐私(Local Differential Privacy,LDP)技术来保护数据隐私. 其核心思想是工人在本地对感知数据进行扰动后再上传服务器,从而从源头上降低隐私泄露风险[9]. 为了实现严格且可证明的LDP,经典方法是在工人提交的原始数据的各个属性维度上注入服从拉普拉斯(Laplace)分布的随机噪声[10]. 该方法适用于多种数据场景,具备较强的灵活性和通用性. 因此,本文采用Laplace机制对工人提交的数据进行扰动,以有效增强隐私保护.

然而,由于Laplace分布的随机性和无界性,注入的噪声可能导致大量离群点的出现. 通常,为了减少离群点的影响,需要增加噪声的阈值范围,但这会导致敏感度参数增大,隐私预算减少,从而使注入的噪声大,进而降低隐私保护效果. 与此同时,现有的真值发现方法未能充分建模为满足LDP保护而注入的Laplace噪声,且大多仅适用于离散值数据或未能严格满足LDP约束,导致噪声"真值"的精度较低,难以满足实际应用需求.

为了解决上述问题,本文提出一种LDP下面向离群点的真值发现算法LEADER(truth discovery algorithms for forward outLiErs under locAl DifferEntial pRivacy). 该算法的核心思想在于通过添加Laplace噪声保护工人隐私信息,使用Huber损失衡量加噪感知值和噪声"真值"之间的距离,同时充分考虑到任务的重要性并对工人分组,进而形式化约束优化问题,再通过拉格朗日乘子法同时迭代计算求得工人的质量、任务的重要性和噪声"真值".

综上,本文的主要贡献如下:

(1)针对离群点问题,本文提出了一种严格满足LDP的真值发现算法LEADER,并对算法的隐私、效用和复杂度进行了理论分析. 理论分析表明,LEADER不仅严格满足LDP约束,而且求得的噪声"真值"精度高,且适用于处理连续型数据和稀疏数据场景. 与非隐私下真值发现算法相比,LEADER没有产生额外的通信开销,同时计算开销相当.

(2)从距离度量的角度出发,设计了一种基于Huber损失的距离度量方法,用于衡量工人在本地提交的经过Laplace加噪后的感知数据与估计"真值"之间的距离. 该方法在保护工人隐私数据的同时,降低随机加噪后产生的离群点影响,能够有效提高求得"真值"的精度.

(3)从数据度量的角度出发,设计了工人分组和任

务重要性建模优化问题. 具体来说,考虑到工人之间的选择性提交行为和相似性,本文对工人进行分组,并在此基础上同时考虑任务的重要性,设计了一种结合拉格朗日乘子法和坐标下降法的求解算法. 该算法通过优化工人的权重分配、分组和任务的重要性,有效提高了噪声“真值”发现精度、隐私保护强度和算法的鲁棒性.

(4)在两个真实数据集和一个合成数据集上的实验结果表明,相较于对比算法,LEADER算法在求得的噪声“真值”精度上至少提高了18%.

2　相关工作

2.1　真值发现

Li等人[11]提出了一种无需真实因果图就能评估因果发现的新方法,解决了合成数据集可能无法准确得到真值发现结果. Fang等人[12]提出了一种广义贝叶斯框架,该框架综合了多真值的特征,可实现准确、高效的多源数据整合,能够有效处理一个数据项有多个真值的情况. Bai等人[13]提出了一种无人机支持的智能真值发现方案,以低成本通信方式获取MCS的真值数据,能够在信息无验证的场景下MCS中进行数据真值发现. Wang等人[14]设计了一个真值发现和反向拍卖框架,能在群智感知应用中最大化系统效用和提高数据质量. Kang等人[15]提出了一种结合智能数据与投标的双重真值发现方案,以应对低质量工人及其欺骗性数据投标对MCS数据收集可信度的破坏,从而提升MCS服务的有效性.

2.2　离群点检测

离群值检测的经典方法主要包括统计方法、聚类方法和基于邻近度的方法. 统计方法通常假设数据服从某种已知分布(如正态分布),通过判断数据点是否位于低概率区域来识别离群值[16,17]. 但在数据分布未知或高维数据场景下表现不佳. 聚类方法的核心思路是将数据划分为不同簇,簇中样本数少的部分为离群值[18,19]. 但这类方法的效果依赖于所选聚类算法,且缺乏对离群点的专门优化.

基于邻近度的离群值检测方法包括基于距离、密度和深度三类方法,这些方法通过分析数据点之间的接近程度,将与大多数点接近程度较低的点判定为离群值[20]. 例如,Knorr等人[21]提出了一种利用不同的距离定义计算数据点之间的平均距离,通过全局参数来识别离群点. Ramaswamy等人[22]通过k近邻距离的离群值检测方法避免了全局参数的依赖. Schubert等人[23]提出了基于核密度估计方法能在多维数据中有效识别离群点. 基于深度的技术通过为每个点分配深度值,根据深度值筛选外部点作为离群点. Cárdenas-Montes[24]提出的基于粒子潜力井的离群点检测算法,将数据视为N维粒子,通过潜力井来区分离群点和簇中的对象. 该方法能够有效应对高维数据,但在处理大规模数据集时可能面临较高的计算成本,并且在某些复杂数据分布或高维场景下存在精度问题.

相比于基于密度和深度的方法,基于距离的离群值检测方法计算效率更高,适用性更广,特别适合处理高维数据. 因此,本文提出的基于LDP的离群点真值发现算法,在加噪过程中采用了基于距离的离群点检测方法,能够有效识别加噪后的离群点,从而减少噪声对真值发现精度的影响,提高算法的鲁棒性. 通过结合LDP保护机制和基于距离的离群点检测方法,本算法不仅能有效保护工人的隐私,还能在处理大规模和高维数据时保持较高的精度和效率.

2.3　基于LDP的真值发现方法

为了解决真值发现中的隐私泄露问题,本文基于LDP在工人端对数据添加噪声,使得上传前的数据已经经过隐私保护,从而有效防止隐私泄露.

基于LDP的隐私保护方法已成为真值发现研究的重要方向. 针对离散值,Sun等人[25]提出了一种通过收集工人上传的扰动数据推断真值的LDP真值发现方法. 为解决数据稀疏性问题,该方法设计了一种新型矩阵分解算法,在实现隐私保护的同时提升了数据有效性,并为每位工人提供统一的隐私级别. Li等人[26]提出了一种高效的LDP真值发现方法,允许工人在提交答案时添加个性化噪声,从而实现个性化的隐私保护. 该方法基于采样机制动态调整噪声强度,在隐私保护与真值推断准确性之间达成平衡. Huang等人[27]基于激励机制设计了适用于流数据的LDP隐私保护框架,有效应对直接应用LDP时长期隐私泄露和准确性下降的挑战. 针对现有激励机制未充分考虑工人隐私偏好和个性化报酬的问题以及连续值场景,Sun等人[28]提出了一种基于LDP的个性化隐私激励合约机制,提升了隐私保护的灵活性. Zhang等人[29]提出了一种不依赖独立性假设的LDP真值发现方法,通过严格满足LDP实现了有效的隐私保护. 进一步地,Zhang等人[30]将拉普拉斯噪声与固有高斯噪声结合进行联合概率估计,有效解决了在严格LDP条件下处理连续数据时的真值发现精度问题. Xiong等人[31]则提出了一种去中心化隐私保护框架,解决了群智感知数据隐私保护中的安全性和计算开销问题.

表1总结并对比了相关研究. 可以发现,当前尚无研究在LDP框架下深入探讨离群点与连续值场景的真值发现问题. 具体来说,现有方法主要关注离群点的检测与分析,但尚未提出针对离群点的有效真值发现方法. 此外,这些研究通常仅适用于离散值或连续值的单

一场景,要么无法严格满足LDP约束,要么未能在隐私保护与真值发现之间取得良好平衡.

表1 相关工作对比

相关工作	离散值	连续值	数据流	LDP	离群点
文献[11]	√				
文献[12]	√	√			
文献[13~15]		√			
文献[25,26]	√			√	
文献[27]			√	√	
文献[28~31]		√		√	
LEADER		√		√	√

3 预备知识

3.1 真值发现

真值发现最重要的特征是对数据源质量的估计,即可靠性估计. 为确定真值,它通过估计数据源的可靠性来聚合多源数据. 现有真值发现经典方法为异构数据冲突解决算法(Conflict Resolution on Heterogeneous data,CRH)[5]. 基本原理是如果工人提供的数据更接近真实值,则认为该工人更可靠,权重也相应较高,并且在聚合过程中为其分配更高的权重. 在涉及工人权重和任务"真值"这两组变量的优化问题中,CRH通过最小化目标函数,采用迭代优化方法,即在更新一组变量时固定另一组变量,直至收敛. 最终,工人权重和任务"真值"分别通过式(1)和式(2)进行迭代求解:

$$w_s=-\lg\left(\frac{\sum_{n=1}^{N}d\left(x_n^s,x_n^*\right)}{\sum_{s=1}^{M}\sum_{n=1}^{N}d\left(x_n^s,x_n^*\right)}\right) \tag{1}$$

$$x_n^*=\frac{\sum_{s=1}^{M}w_s\cdot x_n^s}{\sum_{s=1}^{M}w_s} \tag{2}$$

其中,M、N分别表示工人和任务总数,w_s表示第s个工人的权重(即工人质量),x_n^s表示第s个工人提交的第n个任务的感知值,x_n^*是计算得到的第n个任务的"真值". 此外,$d\left(x_n^s,x_n^*\right)$用于表示第$s$个工人提交的第$n$个任务的感知值与该任务估计"真值"之间的距离. 经典方法通常使用欧氏距离来衡量工人提交值与估计真值之间的差异[32].

3.2 本地差分隐私

LDP形式化定义如下:

定义1 LDP[9,10]. 假设x和x'是任意2个包含隐私的输入数据. 对于给定的$\varepsilon\in R^+$,随机化机制A满足ε-LDP,当且仅当对于所有的输入x、x'和输出y,以下公式成立:

$$\Pr\left(A(x)=y\right)\leqslant e^{\varepsilon}\times\Pr\left(A(x')=y\right)$$

其中,ε表示所需的隐私预算,更小的ε将提供更强的隐私保护. 攻击者无法区分随机化机制函数A的任何输出是来自x还是x',从而有效保护工人的隐私.

在满足LDP算法中,整体隐私预算遵循如下组合定理:

定理1 顺序组合定理[33]. 假设有满足LDP的一组扰动机制$A_i=\left\{A_1,A_2,\cdots,A_k\right\}$,每个扰动机制对应的隐私参数分别为$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_k$,则在对同一数据集使用这些扰动机制时,定义的LDP顺序组合定理满足$\sum_{i=1}^{k}\varepsilon_i$-LDP.

定理2 并行组合定理[34]. 假设有满足LDP的一组扰动机制$A_i=\left\{A_1,A_2,\cdots,A_k\right\}$,每个扰动机制对应的隐私参数分别为$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_k$,$P_1,P_2,\cdots,P_k$是数据集$P$的不相交子集,且$P=\bigcup_{i=1}^{k}P_i$,并满足$P_i\cap P_j=\varnothing$(对于任意$i\neq j$),$A(P_1)=f(P_1)+\mathrm{Lap}(\Delta f/\varepsilon_1)$满足$\varepsilon_1$-差分隐私. 令$A(P)=\bigcup_{i=1}^{k}A_i\ (P_1)$,并且对于每个$A_i$使用独立的随机性,则$f(P)=\bigcup_{i=1}^{k}f(\mathrm{P}_i)$. 此时$A(P)$满足$\max\left\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_k\right\}$-LDP.

定理3 后处理定理[34]. 假设有满足LDP的随机扰动机制A_1和在A_1输出上的任意扰动机制A_2,那么扰动机制A_2也具有相同的LDP隐私保护强度.

3.3 问题描述

在本文所提场景中,如图1所示,假设共有M个工人,需要对N个任务进行真值发现,且要求满足LDP下的真值发现,其中工人集合为$U=\left\{u_1,u_2,\cdots,u_M\right\}$,任务集合为$T=\left\{t_1,t_2,\cdots,t_N\right\}$,$\varepsilon$定义为每个工人的隐私预算,用来衡量隐私保护水平,较小的ε提供更强的隐私保护,工人s所做任务提交的感知数据值$\vec{x}^s=\left\{x_n^s\middle|n\in T_s\right\}$,其中$n$表示第$n$个任务,$T_s\subseteq T$表示工人$s$所做任务的集合,任务数量记为$\left|T_s\right|$. 该问题的核心是在严格满足LDP的条件下保护工人的感知数据值x_n^s,充分考虑实际场景中工人之间可能存在的社交关系或共同偏好,设计有效的工人分组方法,并联合优化求解工人权重和任务真值$\left\{\hat{x}_n^*\middle|n\in T\right\}$,以提高真值发现的精度.

表2对文中常用符号及其含义进行了总结和描述.

4 LEADER算法

4.1 算法描述

在LDP场景下,虽然注入Laplace噪声能满足隐私保护要求,但由于其随机性和无界性,容易引入大量噪声和离群点,导致精度下降和鲁棒性不足,且现有方法

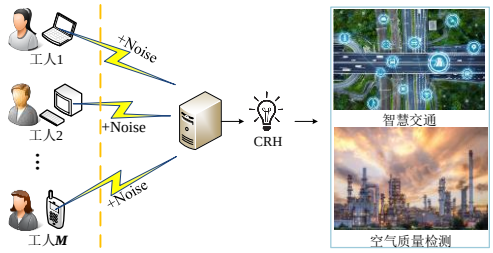


图1　本文问题场景

表2　本文常用符号

符号	符号定义
M	工人数
N	任务数
w_s	第s个工人的权重
y_n	第n个任务的重要性
z_o	第o个组的权重
x_n^s	第s个工人所做第n个任务的值
x_n^*	第n个任务的真值
U	工人集合
T	任务集合
x'^s_n	第s个工人做第n个任务的噪声值
$\hat{x}_n^{*o}$	第o组第n个任务的噪声真值
$\hat{x}_n^*$	第n个任务的噪声真值
c_{so}	第s个工人是否隶属于第o个组
τ_1,τ_2	迭代阈值

通常仅适用于离散数据或未严格满足LDP约束.

为了解决连续值场景中的噪声“真值”精度低的问题,本文提出一种LDP下面向离群点的真值发现算法LEADER,设计了距离度量和数据度量的优化策略. 在距离度量方面,基于Huber损失的距离度量方法通过结合平方误差和绝对误差,能在存在异常值时提供更鲁棒的结果,能够有效针对离群点. 在数据度量方面,观察到工人往往根据任务的重要性选择性地提交更为精确的数据. 因此,对于重要任务,工人提交的感知值通常更接近真实值,给任务赋予权重可以更好地估计真值. 同时,本文还发现工人提交的感知值之间常常存在相似性,这种相似性可能源于工人之间的社交关系,或是他们在能力、偏好以及行为模式上的相似性. 这些因素使得工人可能会提交相似的感知数据. 因此,基于相似性进行工人分组,能减少噪声和错误数据的影响,从而提高真值估计的精度. 通过任务权重和工人分组的双重优化,LEADER能够更准确地估计真值,并有效应对离群点的挑战. LEADER算法的主要流程如图2所示. 其主要由以下4个步骤组成:

步骤1　工人在本地对感知数据添加Laplace噪声后上传至MCS服务器.

步骤2　MCS服务器接收到加噪数据后,针对离群点,通过距离度量和数据度量分析,采用Huber损失函数并结合分组策略构建最小化目标函数.

步骤3　根据拉格朗日乘子法求解得到的迭代计算公式求得工人的质量、任务重要性、分组矩阵和每个组内任务的噪声“真值”直至收敛.

步骤4　根据步骤3求得的任务重要性和各个组内任务噪声“真值”调用CRH迭代计算求得每个任务的最终噪声“真值”直至收敛,最后再将噪声“真值”提交给MCS的发起者.

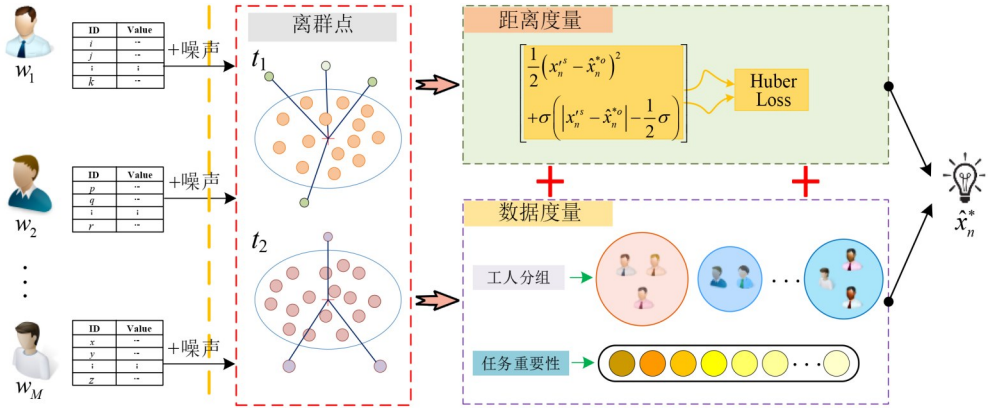


图2　LEADER流程图

4.2　算法实现

首先,工人在本地上传加噪后的感知任务噪声值x'^s_n,为了求得任务的真值$\hat{x}_n^*$,本文除了给工人赋予权重外,还考虑到任务的重要性对真值发现的负面影响,因为在实际场景中工人会选择对重要任务提交更精确的感知数据值,使得工人对重要任务的提交值更接近真实值. 本文用w_s表示第s个工人的质量,y_n表示第n个任务的重要性. 为了求得工人的质量、任务的重要性和噪声“真值”,存在一个基本的问题,如何对工人分组并根据分组求得真值,因此本文提出一个两阶段的真值

发现方法LEADER,具体建模过程如下.

在该算法中,第一阶段首先最小化工人上传噪的音值x'^{s}_{n}和待求解分组任务"真值"$\hat{x}^{*o}_{n}$之间的加权距离,因为本文采用工人分组策略,使得该阶段求得的任务真值为组内任务真值. 构建如式(3)所示的约束优化问题:

$$\min_{w_s,y_n,\hat{x}^{*o}_{n},c_{so}} \sum_{o=1}^{O}\sum_{s=1}^{M}\sum_{n=1}^{N} c_{so}w_s y_n d\left(x'^{s}_{n},\hat{x}^{*o}_{n}\right) \tag{3}$$

本文使用Huber损失来衡量工人提交的加噪感知数据值和待求解分组任务"真值"之间的距离,而不是使用典型的平方损失函数. 因为,Huber损失的距离度量方法通过结合平方误差和绝对误差,能在存在异常值时提供更鲁棒的结果,能够有效针对离群点,在处理带有噪声或异常值的数据时更稳定,更适合发现真值. 其损失函数定义如式(4)所示:

$$d\left(x'^{s}_{n},\hat{x}^{*o}_{n}\right)=\begin{cases}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2, & \left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|<\sigma\\ \sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right), & \text{otherwise}\end{cases} \tag{4}$$

其中,σ是一个超参数,用于控制损失函数的转折点,默认设置为1. 同时,本文观察到在实际场景中,感知数据上相似的工人会被分到同一组,因为这些工人在执行任务时工人之间存在社交关系或者有类似的能力、偏好或行为模式,即总体上工人之间存在一定的相似性. 因此,本文在考虑工人的质量和任务的重要性的同时采用分组策略. 此外,为了最小化目标函数,即最小化偏离"真值"的误差,如果求得的"真值"偏离高质量的工人和重要的任务,那么在目标函数中将给予其更大的惩罚. 工人分组通过最小化感知数据与任务真值的加权误差,影响优化约束,从而确保工人被分配到能最大程度减少加噪误差的组. 式(5)中的2个约束条件分别用来约束工人权重和任务重要性的分布,取值大于0小于1.

$$\min_{w_s,y_n,\hat{x}^{*o}_{n},c_{so}} \sum_{o=1}^{O}\sum_{s=1}^{M}\sum_{n=1}^{N} c_{so}w_s y_n\left[\begin{array}{l}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\ +\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right)\end{array}\right]$$

$$\text{s.t.}\begin{cases}\displaystyle\sum_{o=1}^{O} c_{so}=1, \quad c_{so}\in\{0,1\}, \quad 1\leqslant s\leqslant M\\ \displaystyle\sum_{s=1}^{M}\mathrm{e}^{-w_s}=1\\ \displaystyle\sum_{n=1}^{N}\mathrm{e}^{-y_n}=1\end{cases} \tag{5}$$

其中,c_{so}表示第s个工人是否隶属于第o个组,取值有2种可能性,若隶属于该组则$c_{so}=1$,否则$c_{so}=0$.

根据式(5)可知,本文的目标函数为

$$f(w_s,y_n,\hat{x}^{*o}_{n},c_{so})=\sum_{o=1}^{O}\sum_{s=1}^{M}\sum_{n=1}^{N} c_{so}w_s y_n\left[\begin{array}{l}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\ +\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right)\end{array}\right]$$

其中,$\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2$是一个凸的二次函数,绝对值函数$\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|$也是凸的,因此乘以正的常数$\sigma$后仍保持凸性. 由于目标函数是对这些凸函数的线性组合,且w_s、y_n是正的,c_{so}为二元选择变量(0或1),整体上保持了凸性. 因此,目标函数$f(w_s,y_n,\hat{x}^{*o}_{n},c_{so})$是一个凸函数,因此上述问题是凸优化问题. 可以结合拉格朗日乘子法来设计求解方案,具体步骤如下:

(1)首先得到式(5)对应的拉格朗日函数,如式(6)所示,其中μ、λ和γ是拉格朗日乘子.

$$\begin{aligned}&L\left(w_s,\hat{x}^{*}_{n},y_n,\mu,\lambda,\gamma_s\right)\\ &=\sum_{o=1}^{O}\sum_{s=1}^{M}\sum_{n=1}^{N} c_{so}w_s y_n\left[\begin{array}{l}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\ +\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right)\end{array}\right]\\ &+\mu\left(\sum_{s=1}^{M}\mathrm{e}^{-w_s}-1\right)+\lambda\left(\sum_{n=1}^{N}\mathrm{e}^{-y_n}-1\right)\\ &+\sum_{s=1}^{M}\gamma_s\left(\sum_{o=1}^{O}c_{so}-1\right)\end{aligned} \tag{6}$$

(2)接着求式(6)关于w_s的一阶导并令其等于0,可得式(7):

$$\mu\mathrm{e}^{-w_s}=\sum_{o=1}^{O}\sum_{n=1}^{N} c_{so}y_n\left[\begin{array}{l}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\ +\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right)\end{array}\right] \tag{7}$$

根据式(5)中的约束条件2,有

$$\mu\left(\mathrm{e}^{-w_1}+\mathrm{e}^{-w_2}+\cdots+\mathrm{e}^{-w_s}+\cdots+\mathrm{e}^{-w_M}\right)=\mu \tag{8}$$

(3)进而求得式(9):

$$\mu=\sum_{o=1}^{O}\sum_{s=1}^{M}\sum_{n=1}^{N} c_{so}y_n\left[\begin{array}{l}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\ +\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right)\end{array}\right] \tag{9}$$

将求得的λ回代入式(7),可以得到:

$$w_s=-\ln\frac{\displaystyle\sum_{o=1}^{O}\sum_{n=1}^{N} c_{so}y_n\left[\begin{array}{l}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\ +\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right)\end{array}\right]}{\displaystyle\sum_{o=1}^{O}\sum_{s=1}^{M}\sum_{n=1}^{N} c_{so}y_n\left[\begin{array}{l}\dfrac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\ +\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\dfrac{1}{2}\sigma\right)\end{array}\right]} \tag{10}$$

同理,求式(6)关于y_n的一阶导并令其为0,可得:

$$y_n=-\ln\frac{\sum_{o=1}^{O}\sum_{s=1}^{M}c_{so}w_s\left[\begin{array}{l}\frac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\+\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\frac{1}{2}\sigma\right)\end{array}\right]}{\sum_{o=1}^{O}\sum_{s=1}^{M}\sum_{n=1}^{N}c_{so}w_s\left[\begin{array}{l}\frac{1}{2}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)^2\\+\sigma\left(\left|x'^{s}_{n}-\hat{x}^{*o}_{n}\right|-\frac{1}{2}\sigma\right)\end{array}\right]} \tag{11}$$

进一步地,求式(6)关于$\hat{x}^{*o}_n$的一阶导并令其为0,可得:

$$\sum_{s=1}^{M}c_{so}w_s y_n\left[\begin{array}{l}-\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)\\+\sigma\cdot\mathrm{sgn}\left(x'^{s}_{n}-\hat{x}^{*o}_{n}\right)\end{array}\right]=0 \tag{12}$$

其中,sgn(·)是符号函数,如果内部值大于0则返回1,小于0则返回–1;如果等于0则返回0. 那么当$x'^{s}_{n}>\hat{x}^{*o}_{n}$时,求解可得:

$$\hat{x}^{*o}_{n}=\frac{\sum_{s=1}^{M}c_{so}w_s\left(x'^{s}_{n}-\sigma\right)}{\sum_{s=1}^{M}c_{so}w_s} \tag{13}$$

同理,当$x'^{s}_{n}\leqslant\hat{x}^{*o}_{n}$时,可得:

$$\hat{x}^{*o}_{n}=\frac{\sum_{s=1}^{M}c_{so}w_s\left(x'^{s}_{n}+\sigma\right)}{\sum_{s=1}^{M}c_{so}w_s} \tag{14}$$

综上所述可得:

$$\hat{x}^{*o}_{n}=\begin{cases}\dfrac{\sum_{s=1}^{M}c_{so}w_s\left(x'^{s}_{n}-\sigma\right)}{\sum_{s=1}^{M}c_{so}w_s}, & \text{if } x'^{s}_{n}>\hat{x}^{*o}_{n}\\[2ex]\dfrac{\sum_{s=1}^{M}c_{so}w_s\left(x'^{s}_{n}+\sigma\right)}{\sum_{s=1}^{M}c_{so}w_s}, & \text{if } x'^{s}_{n}\leqslant\hat{x}^{*o}_{n}\end{cases} \tag{15}$$

根据块坐标法,当x'^{s}_{n},w_s,y_n固定时,c_{so}使用下面定理中的公式更新.

定理4　固定w_s,y_n,可通过下面公式更新工人组指示符c_{so}:

$$c_{so}=\begin{cases}1, o^*=\arg\min\limits_{o}\sum\limits_{n=1}^{N}y_n d\left(x'^{s}_{n},\hat{x}^{*o}_{n}\right), 1\leqslant o^*\leqslant O\\0, o\neq o^*\end{cases} \tag{16}$$

证明　由于x'^{s}_{n},w_s,y_n都是固定的,式(5)只有一组未知变量是未知的,根据式(5)中的约束条件1,按照式(16)给c_{so}赋值时,使得工人s被分配到加噪后的感知数据与任务真值的任务重要性加权距离误差最小的组中,也就是将工人分配到与其提交的加噪感知数据最接近"真值"的组. 通过最小化总加权误差,确定工人应被分配到的组. 此时

$$\sum_{o=1}^{O}\sum_{n=1}^{N}c_{so}w_s y_n\left[\begin{array}{l}\frac{1}{2}\left(x'^{so}_{n}-\hat{x}^{*o}_{n}\right)^2\\+\sigma\left(\left|x'^{so}_{n}-\hat{x}^{*o}_{n}\right|-\frac{1}{2}\sigma\right)\end{array}\right]$$

这一项最小,此外由于感知数据是工人独立感知并提交的,因此工人彼此之间不存在直接的共享或依赖关系,则不同工人提交的感知数据值之间是独立不相关的,所以式(5)的目标函数最小.

第二阶段,根据真值发现经典算法CRH,本文通过给每个组分配一个权重z_o来迭代计算每个组的权重和各个任务的"真值",目标函数如式(17)所示:

$$\begin{aligned}&\min_{z_o,\hat{x}^{*}_{n}}\sum_{o=1}^{O}\sum_{n=1}^{N}z_o y_n\left[\begin{array}{l}\frac{1}{2}\left(\hat{x}^{*o}_{n}-\hat{x}^{*}_{n}\right)^2\\+\sigma\left(\left|\hat{x}^{*o}_{n}-\hat{x}^{*}_{n}\right|-\frac{1}{2}\sigma\right)\end{array}\right]\\&\text{s.t.}\quad\sum_{o=1}^{O}\mathrm{e}^{-z_o}=1\end{aligned} \tag{17}$$

其中,y_n根据式(11)可得. 因为,任务的重要性y_n是一种全局属性,在经过第一阶段迭代求解后已能较好地反映任务的特征. 相比之下,第二阶段的分组权重和任务"真值"是局部优化问题,重点在于组内数据的一致性和组间的差异性,因此不需要再次调整.

接着,通过拉格朗日乘子法求解,可得每个组的权重和任务的"真值"迭代计算公式,分别如式(18)和式(19)所示. 本文也用Huber损失来衡量分组中的任务"真值"和待求解任务"真值"$\hat{x}^{*}_{n}$之间的偏差.

$$z_o=-\ln\frac{\sum_{n=1}^{N}y_n\left[\begin{array}{l}\frac{1}{2}\left(\hat{x}^{*o}_{n}-\hat{x}^{*}_{n}\right)^2\\+\sigma\left(\left|\hat{x}^{*o}_{n}-\hat{x}^{*}_{n}\right|-\frac{1}{2}\sigma\right)\end{array}\right]}{\sum_{o=1}^{O}\sum_{n=1}^{N}y_n\left[\begin{array}{l}\frac{1}{2}\left(\hat{x}^{*o}_{n}-\hat{x}^{*}_{n}\right)^2\\+\sigma\left(\left|\hat{x}^{*o}_{n}-\hat{x}^{*}_{n}\right|-\frac{1}{2}\sigma\right)\end{array}\right]} \tag{18}$$

$$\hat{x}^{*}_{n}=\begin{cases}\dfrac{\sum_{o=1}^{O}z_o\left(\hat{x}^{*o}_{n}-\sigma\right)}{\sum_{o=1}^{O}z_o}, & \text{if } \hat{x}^{*o}_{n}>\hat{x}^{*}_{n}\\[2ex]\dfrac{\sum_{o=1}^{O}z_o\left(\hat{x}^{*o}_{n}+\sigma\right)}{\sum_{o=1}^{O}z_o}, & \text{if } \hat{x}^{*o}_{n}\leqslant\hat{x}^{*}_{n}\end{cases} \tag{19}$$

其中,任务的重要性和每个分组任务的"真值"分别由式(11)、式(15)可得.

LEADER算法流程如算法1所示.

算法1 LEADER算法

输入:工人的感知任务值x_n^s,工人分组数O,迭代阈值τ_1和τ_2

输出:工人的权重w_s,任务的重要性y_n,分组矩阵$\boldsymbol{C}$,组的权重z_o,任务真值$\hat{x}_n^*$

//本地执行

1. FOR s=1,2,⋯,M DO
2. FOR n=1,2,⋯,N DO
3. 工人调用Laplace机制对感知数据x_n^s加噪
4. 工人将加噪后的值x'^s_n上传给服务器
5. END FOR
6. END FOR

//服务器执行

7. 初始化w_s,y_n和$\hat{x}_n^*$
8. WHILE迭代次数或者满足迭代阈值τ_1
9. WHILE迭代次数或者满足迭代阈值τ_2
10. 根据式(10)更新w_s
11. 根据式(11)更新y_n
12. 根据式(15)更新$\hat{x}_n^{*o}$
13. 根据式(16)更新c_{so}
14. 根据式(18)更新z_o
15. 根据式(19)更新$\hat{x}_n^*$
16. RETURN w_s,y_n,$\boldsymbol{C}$,z_o和$\hat{x}_n^*$

在LEADER算法中,首先每个工人对其感知到的任务值进行Laplace加噪处理,并将加噪后的感知数据值x'^s_n上传到服务器(步骤1~6). 然后,服务器执行第一阶段真值发现迭代过程,首先初始化,然后根据式(10)更新工人的权重w_s,根据式(11)更新任务的重要性y_n,然后根据式(15)和式(16)更新第o组第n任务的真值$\hat{x}_n^{*o}$和工人被分配的组指示符c_{so},直到满足迭代次数或收敛阈值τ_1. 接下来,服务器开始第二阶段迭代,根据式(18)更新组的权重,并根据式(19)进一步更新任务的真值,直到满足迭代次数或收敛阈值τ_2. 最终返回工人的权重、任务的重要性、组的权重、分组矩阵和任务真值(步骤7~16). 其中分别用计算所有分组中所有任务的前后两次"真值"差距的平均值是否小于τ_1和所有任务前后两次"真值"差距的平均值是否小于τ_2来判断是否分别满足迭代阈值. 具体地讲,τ_1、τ_2的计算如式(20)、式(21)所示:

$$\tau_1=\frac{\sum_{n=1}^{N}\left|\hat{x}_{n,i}^{*o}-\hat{x}_{n,i+1}^{*o}\right|}{N} \tag{20}$$

$$\tau_2=\frac{\sum_{n=1}^{N}\left|\hat{x}_{n,k}^{*}-\hat{x}_{n,k+1}^{*}\right|}{N} \tag{21}$$

其中,$\hat{x}_{n,i}^{*o}$和$\hat{x}_{n,i+1}^{*o}$分别表示第o组中第n个任务的第i次和第i+1次得到的"真值". $\hat{x}_{n,k}^{*}$和$\hat{x}_{n,k+1}^{*}$分别表示第n个任务的第k次和第k+1次得到的"真值".

5 算法分析

在本节中,通过对LEADER的隐私、效用和复杂度分别进行全面分析,从理论上证明了其优越性.

5.1 隐私分析

首先为了保护工人隐私,LEADER只在本地对工人感知数据采用Laplace机制进行加噪,对于某一工人s,工人对任务n的感知值x_n^s添加Laplace噪声,生成加噪感知值x'^s_n,如式(22)所示:

$$x'^s_n=x_n^s+\mathrm{Lap}\left(\frac{\Delta f}{\varepsilon}\right) \tag{22}$$

其中,Lap(λ)是均值为0、尺度参数为$\lambda=\frac{\Delta f}{\varepsilon}$的拉普拉斯分布,$\Delta f$是感知值的全局灵敏度.

定理5 LEADER算法满足ε-LDP.

证明 工人在本地提交的初始感知值x_n^s经过Laplace加噪后输出x'^s_n的概率密度函数为

$$P\left(x'^s_n|x_n^s\right)=\frac{1}{2\lambda}\exp\left(-\frac{\left|x'^s_n-x_n^s\right|}{\lambda}\right) \tag{23}$$

因此,若证明该算法满足ε-LDP,首先也就是证明Laplace机制满足ε-LDP,对于任意同一任务的两个不同的感知值x_n^s和$x_n^{s'}$,满足以下不等式:

$$\frac{P\left(x'^s_n|x_n^s\right)}{P\left(x'^s_n|x_n^{s'}\right)}\leqslant \mathrm{e}^{\varepsilon} \tag{24}$$

根据Laplace分布的概率密度函数,代入式(22):

$$\begin{aligned}\frac{P\left(x'^s_n|x_n^s\right)}{P\left(x'^s_n|x_n^{s'}\right)}&=\frac{\frac{1}{2\lambda}\exp\left(-\frac{\left|x'^s_n-x_n^s\right|}{\lambda}\right)}{\frac{1}{2\lambda}\exp\left(-\frac{\left|x'^s_n-x_n^{s'}\right|}{\lambda}\right)}\\&=\exp\left(\frac{\left|x'^s_n-x_n^{s'}\right|-\left|x'^s_n-x_n^s\right|}{\lambda}\right)\end{aligned} \tag{25}$$

根据三角不等式,$|a-b|\leqslant|a-c|+|c-b|$,本文有

$$\left|x'^s_n-x_n^{s'}\right|-\left|x'^s_n-x_n^s\right|\leqslant\left|x_n^s-x_n^{s'}\right| \tag{26}$$

因此根据式(24),本文可以进一步得到:

$$\frac{P\left(x'^s_n|x_n^s\right)}{P\left(x'^s_n|x_n^{s'}\right)}\leqslant\exp\left(\frac{\left|x_n^s-x_n^{s'}\right|}{\lambda}\right) \tag{27}$$

由于$\lambda=\frac{\Delta f}{\varepsilon}$,而$\Delta f$是感知值的全局灵敏度,即任意2个可能输入$x_n^s$和$x_n^{s'}$的最大差值满足$\left|x_n^s-x_n^{s'}\right|\leqslant\Delta f$,进而可得:

$$\frac{P\left(x'^s_n|x_n^s\right)}{P\left(x'^s_n|x_n^{s'}\right)}\leqslant\exp\left(\frac{\Delta f}{\lambda}\right)=\exp(\varepsilon) \tag{28}$$

因此,LEADER算法中使用的Laplace噪声机制确保了每个工人的上传感知数据值是差分隐私保护的,且工人上传的加噪感知值x'^{s}_{n}满足ε-LDP.

此外,LEADER算法仅在步骤1中对包含工人隐私的初始感知数据进行加噪处理,并且遵循了Laplace噪声机制和定理1的顺序组合定理,严格满足ε-LDP. 算法后续步骤仅是对构建的约束优化问题进行求解过程,不影响整体隐私预算的分配. 因此,依据定理2的并行组合定理,总体上LEADER对所有工人的数据严格满足ε-LDP.

5.2　效用分析

本节通过对比不同算法求得的真值误差大小来评估LEADER算法的效用.

首先,用η_1表示3.2节中式(19)求得的任务真值,即

$$\eta_1=\hat{x}_n^*=\begin{cases}\dfrac{\sum\limits_{o=1}^{O}z_o\left(\hat{x}_n^{*o}-\sigma\right)}{\sum\limits_{o=1}^{O}z_o}, & \text{if } \hat{x}_n^{*o}>\hat{x}_n^*\\ \dfrac{\sum\limits_{o=1}^{O}z_o\left(\hat{x}_n^{*o}+\sigma\right)}{\sum\limits_{o=1}^{O}z_o}, & \text{if } \hat{x}_n^{*o}\leqslant\hat{x}_n^*\end{cases}\tag{29}$$

仅使用Laplace机制加噪未考虑任务重要性和工人分组的隐私保护下真值发现方法求得的真值$\hat{x}_n^{*\prime}$如式(30)所示:

$$\eta_2=\hat{x}_n^{*\prime}=\begin{cases}\dfrac{\sum\limits_{s=1}^{M}\left[w_s\left(\dfrac{1}{x'^{s}_{n}}+\dfrac{\sqrt{d}}{d}\right)\right]}{\sum\limits_{s=1}^{M}\left[w_s\dfrac{1}{\left(x'^{s}_{n}\right)^2}\right]}, & \text{if } x'^{s}_{n}>\hat{x}_n^{*\prime}\\ \dfrac{\sum\limits_{s=1}^{M}\left[w_s\left(\dfrac{1}{x'^{s}_{n}}-\dfrac{\sqrt{d}}{d}\right)\right]}{\sum\limits_{s=1}^{M}\left[w_s\dfrac{1}{\left(x'^{s}_{n}\right)^2}\right]}, & \text{if } x'^{s}_{n}\leqslant\hat{x}_n^{*\prime}\end{cases}\tag{30}$$

用η_0表示非隐私保护下经典真值发现方法CRH求得的真值,即

$$\eta_0=x_n^*=\frac{\sum\limits_{s=1}^{M}w_s x_n^s}{\sum\limits_{s=1}^{M}w_s}\tag{31}$$

进而,有定理6:

定理6　LEADER算法求得的任务真值更接近非隐私下的任务真值,即$|\eta_1-\eta_0|<|\eta_2-\eta_0|$.

证明　根据LEADER算法求得的真值,有

$$\begin{aligned}\mathrm{MAE}_{\mathrm{LEADER}}&=\frac{1}{N}\sum_{n=1}^{N}\left|x_n^*-\hat{x}_n^*\right|\\&=\frac{1}{N}\sum_{n=1}^{N}\left|\frac{\sum\limits_{s=1}^{M}w_s x_n^s}{\sum\limits_{s=1}^{M}w_s}-\frac{\sum\limits_{o=1}^{O}z_o\left(\hat{x}_n^{*o}\pm\sigma\right)}{\sum\limits_{o=1}^{O}z_o}\right|\end{aligned}\tag{32}$$

未考虑分组和任务重要性时,本文有

$$\begin{aligned}\mathrm{MAE}_{\mathrm{Laplace}}&=\frac{1}{N}\sum_{n=1}^{N}\left|x_n^*-\hat{x}_n^{*\prime}\right|\\&=\frac{1}{N}\sum_{n=1}^{N}\left|\frac{\sum\limits_{s=1}^{M}w_s x_n^s}{\sum\limits_{s=1}^{M}w_s}-\frac{\sum\limits_{s=1}^{M}\left[w_s\left(\dfrac{1}{x'^{s}_{n}}\pm\dfrac{\sqrt{d}}{d}\right)\right]}{\sum\limits_{s=1}^{M}\left[w_s\dfrac{1}{\left(x'^{s}_{n}\right)^2}\right]}\right|\end{aligned}\tag{33}$$

进而可得:

$$\begin{aligned}&\mathrm{MAE}_{\mathrm{LEADER}}-\mathrm{MAE}_{\mathrm{Laplace}}\\&=\frac{1}{N}\sum_{n=1}^{N}\left|x_n^*-\hat{x}_n^*\right|-\frac{1}{N}\sum_{n=1}^{N}\left|x_n^*-\hat{x}_n^{*\prime}\right|\end{aligned}\tag{34}$$

因为根据大数定律,随着样本数M_o的增加,组内的真值会趋近于最终任务真实值:

$$\lim_{M_o\to\infty}\hat{x}_n^{*o}=x_n^*\tag{35}$$

所以每组的估计值$\hat{x}_n^{*o}$和非隐私保护下真值发现方法CRH求得的真值x_n^*之间的误差随着组内观测数M_o增加趋近于0. 对于每个组o,随着组内工人数M_o增加,$\hat{x}_n^{*o}$趋近于x_n^*,从而$\mathrm{Var}(\hat{x}_n^{*o})\to 0$. 所以由于分组的任务真值是基于多个组内估计的任务真值,组内方差减小会导致整体估计误差MAE降低.

具体来说,由于每个组的均值逐渐接近真实值,本文可得:

$$\begin{aligned}\mathrm{MAE}_{\mathrm{LEADER}}&=\frac{1}{N}\sum_{n=1}^{N}\left|x_n^*-\frac{\sum\limits_{o=1}^{O}z_o\hat{x}_n^{*o}}{\sum\limits_{o=1}^{O}z_o}\right|\\&\approx\frac{1}{N}\sum_{n=1}^{N}\left|x_n^*-\frac{\sum\limits_{o=1}^{O}z_o x_n^*}{\sum\limits_{o=1}^{O}z_o}\right|=0\end{aligned}\tag{36}$$

这表明分组后的误差会随着样本数的增多逐渐减小,最终接近0. 进而$\mathrm{MAE}_{\mathrm{LEADER}}-\mathrm{MAE}_{\mathrm{Laplace}}<0$.

未分组时的样本x_n^s来自不同的观测值,这些观测值之间的差异较大,因此均值的估计误差较大. 即便随着样本数的增加,均值误差会逐渐减小,但相比于分组后的方法,由于每组内的样本相关性较低,整体的误差

会相对较大.

综上所述,LEADER算法求得的真值误差小于仅考虑Laplace加噪无分组和无任务重要性影响求得的真值误差,即 $\mathrm{Var}(\hat{x}_n^{*o})<\mathrm{Var}(\hat{x}_n^{*\prime})$,进而可得 $|\eta_1-\eta_0|<|\eta_2-\eta_0|$. 证毕.

5.3 复杂度分析

LEADER算法的复杂度分析如下:

(1)时间复杂度. 用户端:每个工人需要对每个任务的感知数据加噪并上传至服务器,共需执行 $M\times N$ 次操作,时间复杂度为 $O(MN)$. 服务器端:在第一阶段,服务器要重复多次迭代,每次迭代中,需要更新工人权重、任务重要性和分组矩阵. 对于每次迭代:更新工人权重 w_s 的复杂度是 $O(MN)$,更新任务重要性 y_n 的复杂度是 $O(NO)$,更新分组矩阵 c_{so} 和组内任务真值 $\hat{x}_n^{*o}$ 的复杂度为 $O(MO+NO)$. 本文定义第一阶段的总迭代次数为 I_1,因此第一阶段的时间复杂度为 $O\left(I_1(MN+NO+MO)\right)$. 在第二阶段,服务器根据式(18)和式(19)更新组权重 z_o 和最终的任务真值 $\hat{x}_n^*$,复杂度为 $O(NO)$ 和 $O(NO)$. 本文定义第二阶段总迭代次数为 I_2,因此第二阶段的时间复杂度为 $O\left(I_2NO\right)$. 因此,总体上LEADER的时间复杂度为 $O\left(I_1(MN+NO+MO)+I_2NO\right)$.

(2)通信复杂度. 该算法仅在工人上传服务器加噪的感知数据值时需要通信,共有 M 个工人且每个工人向服务器上传 N 个加噪后的感知值,总通信复杂度为 $O(MN)$.

综上所述,与经典的非隐私保护真值发现算法CRH相比,通信复杂度一致. 此外,由于LEADER算法引入了隐私保护和工人分组机制,其时间有所增加. 本文定义CRH的迭代阶段的迭代次数为 I_2,则相比于CRH,LEADER在时间复杂度上增加了 $O\left(I_1(NO+MO)+(I_1-I_3)(MN)+I_2NO\right)$. 但这些额外的计算开销换取了更高的隐私保护强度和"真值"精度.

6 实验评估

6.1 数据集

本文采用2个公开可用的数据集(Intel Berkeley、Weather)和一个合成数据集.

(1)Intel Berkeley数据集(http://www.kaggle.com/datasets/divyansh22/intel-berkeley-research-lab-sensor-data). 简称Int,该数据集是由Intel Berkeley实验室于2004年2月28日至4月5日部署的54个Mica2传感器对同一室内空间在36天内每隔30 s采样所得. 本文选取传感器第一天内不同节点采集到的温度、湿度和电压3个属性,共1 317条数据进行真值发现. 缺失数据通过该节点其他时间的采样数据填充,最终得到实验数据集.

(2)Weather数据集(http://paperswithcode.com/dataset/weather-ltsf). 简称Wea,该数据集为马克斯·普朗克生物地球化学研究所于2020年记录的每10 min一次的天气数据,包括21项气象指标. 本文选取1月31天内的142条数据,包含温度、湿度和风力3个任务作为本实验的数据集.

(3)合成数据集. 简称Syn,该数据集由1 200个工人对25个任务的数据组成,感知数据值范围为0~30,其中真值为15,为不失一般性,其中95%的感知数据集中在14~16范围内,剩下的5%数据分布在更广的0~30内.

6.2 评价指标和实验环境

本文采用MAE Change和KL-Divergence作为评估指标,以全面评估所提算法的精度与收敛速度. 具体而言,本文使用通用的MAE Change指标来衡量最终得到的噪声"真值"效用,其计算方法如式(37)所示:

$$\text{MAE Change}=\frac{\left|\sum_{k=1}^{K}|x_k^{*\prime}-x_k|-\sum_{k=1}^{K}|x_k^{*}-x_k|\right|}{|K|} \tag{37}$$

其中,$|K|$ 表示总的任务数量,$x_k^{*\prime}$ 表示隐私保护添加Laplace噪声后求得的第 k 个任务的噪声真值,x_k^* 表示非隐私保护下求得第 k 个任务的真值,x_k 则表示第 k 个任务的真实值. MAE Change表示在隐私保护前后,平均绝对误差(Mean Absolute Error,MAE)的变化量. 具体而言,MAE Change的值越小,说明该隐私保护的真值发现算法对任务效用的影响越小,从而表明该算法在保护隐私的同时有效地保留了数据的真实价值.

其次,本文采用KL散度(KL-Divergence)来衡量隐私保护前后工人权重的变化情况,其中 w 和 w^* 分别是隐私保护前后的权重,KL-Divergence具体计算方法如式(38)所示:

$$\mathrm{KL}\left(w\|w^*\right)=\sum p(w)\lg\frac{p(w)}{q(w^*)} \tag{38}$$

此外,本文通过计算不同算法在离群点与非离群点数据不平衡情况下的 F_1 分数(F_1-Socre),更全面地评估其在处理异常值时的鲁棒性. 相比于仅使用MAE Change作为误差指标,F_1-Socre结合了精准率(Precision)和召回率(Recall),能够更全面地评估算法在处理离群点时的性能,避免因单一误差指标带来的偏差. 其中,Precision、Recall和 F_1-Socre的计算方法分别如式(39)~式(41)所示:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{39}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{40}$$

$$F_1\text{-Score}=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{41}$$

其中,TP表示正确识别离群点的数量,FP表示误判为离群点的正常数据点,FN表示未能识别出的离群点.在本文的算法比较中,F_1-Socre更能全面反映算法对离群点的抑制能力,因此F_1-Socre越大越好,代表算法越有效.

本文的所有实验均在一台配备Intel i7-14700HX处理器、NVIDIA4070显卡、16 GB内存、20核28线程的笔记本电脑上进行,使用Python 3.11编写所有算法.所有实验隐私预算默认为$\varepsilon=0.5$. 为了确保算法稳定收敛,设置了默认迭代停止阈值$\tau_1=\tau_2=0.000\,1$,保证实验的高效执行和结果的可复现性. 针对不同数据集规模,实验环境进行了相应的调整. 对于较大数据集Int,在超参数方面,增大了阈值大小其值为$\tau_1=\tau_2=0.001$,以平衡精度和计算效率.

6.3　对比算法

在深入分析现有研究工作的过程中,本文发现现有方法难以直接解决本文所提出的由于Laplace加噪的无界性和随机性导致产生的离群点问题,以及基于欧氏距离的算法在此类噪声数据下表现出鲁棒性差的问题. 因此,本文从距离度量和数据度量的角度出发,设计了LEADER算法. 为了验证LEADER算法的有效性和精确度,本文将其与以下最新研究算法进行了对比实验,全面评估LEADER算法的性能表现.

(1)SampTD. 针对加噪后产生的离群点,本文设计了该对比方法,首先随机采样工人本地上传的Laplace加噪感知数据中的一半,然后通过调用非隐私保护下真值发现方法CRH求得任务的"真值",该方法严格满足LDP.

(2)PrunTD. 针对加噪后产生的离群点,本文设计了另一种对比方法,该方法首先通过将工人在本地上传Laplace加噪后的感知数据根据不同的隐私预算设计的剪枝范围进行剪枝,然后调用非隐私保护下真值发现方法CRH求得任务的"真值",该方法严格满足LDP.

(3)TESLA[30]. 此方法设计了一种噪声过滤处理机制,针对不同类型的注入噪声进行处理,然后再调用CRH算法进行真值发现. 该方法也严格满足LDP.

其中,上述对比算法均满足LDP,但是SampTD方法可能无法有效应对噪声对"真值"准确性的影响. PrunTD方法性能在低隐私预算情况下可能不够鲁棒. TESLA在面对Laplace加噪带来的离群点时,其噪声过滤机制可能导致信息损失,影响最终的"真值"精度.

6.4　对比实验

(1)隐私预算ε的影响

图3是在不同的隐私预算下,LEADER和对比算法在MAE Change和F_1-Socre两个评价指标上的效果对比. 图3(a)、图3(c)和图3(b)、图3(d)分别对应的是Int数据集和Wea数据集下不同隐私预算的对比效果. SampTD和PrunTD算法分别通过随机采样50%数据和剪枝策略后,调用CRH真值发现方法计算任务的真值,TESLA方法则是通过定制化噪声处理来减少噪声对结果的影响.

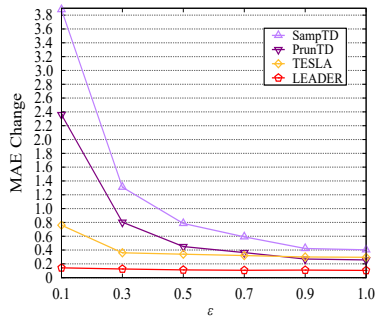
从实验结果来看,随着隐私预算的增加,所有算法的MAE Change都呈现下降趋势. 此外,本文观察到,首先PrunTD方法优于SampTD方法,原因是SampTD通过随机采样工人上传的加噪数据,导致大量有用信息丢失,并且由于Laplace分布的随机性和无界性,产生了大量离群点,真值计算的误差大幅增加. 而PrunTD剪枝策略有效限制了数据中的极端噪声(如超过一定范围的值),从而有效提高了数据质量,降低了加噪对真值计算的负面影响,因此MAE Change较小. 其次,TESLA方法优于前两种方法,原因是该方法针对不同类型的噪声设计定制的处理机制,减少了加噪数据对真值计算的干扰,使其在去噪方面表现较好. 最后,与最新的TESLA算法相比较,LEADER算法求得的真值精度至少提高了18%. 这是因为LEADER算法通过Huber损失函数度量加噪感知数据与估计真值之间的误差,其结合了均方误差和绝对误差的优点,对大误差进行了线性处理,从而有效抑制了离群点对真值迭代过程的干扰. 此外,LEADER通过赋予重要任务更高的权重并对工人进行分组,减少了低质量工人数据对真值计算的影响,同时通过将感知数据相似的工人分组,减小了Laplace噪声的随机性的影响,从而提高真值估计的精度.

在对比算法中尽管本文和研究者们采取了一些方法来提升"真值"的精度,然而这些方法均未能考虑到由于Laplace分布的随机性和无界性导致的大量离群点. 为满足LDP保护,通常需要使用较大的加噪范围,使得噪声量变大,进而降低了真值精度.

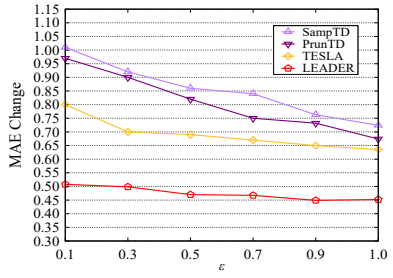
如图3(c)和图3(d)所示,F_1-Socre随着隐私预算ε的增加而上升,说明较高的隐私预算有助于提高任务的准确性. 其中,PrunTD相较于SampTD的F_1-Socre更高,表明裁剪策略比采样策略更有效. 本文提出的LEADER方法在所有实验条件下表现最佳,TESLA次之. 这是因为LEADER在计算真值时能更有效地抑制Laplace噪声引入的极端偏差. 通过优化策略调整数据,使因噪声偏移产生的离群点回归合理范围,从而减少其对真值计算的负面影响. 相比其他方法,LEADER能更好地将潜在离群点转化为合理值,提高整体数据稳定性. 因此,其F_1-Socre更高,表明LEADER在平衡隐私保护与真值发现精度方面更具优势,进一步验证了其优越性和鲁棒性.

(2)对工人权重w_s的影响

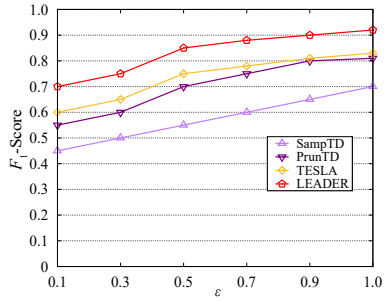
在真值发现问题中,权重更新是关键步骤,因为它直接影响算法对工人质量的估计,进而影响对任务真值的收敛性和准确性,优化真值估计过程. 为验证不同数据集上LEADER算法所求得的添加Laplace噪声后的权重分布与真实权重分布之间的差异,本文进行了相关实验,结果如图4所示. 图4(a)和图4(b)展示了在Int数据集和Wea数据集上,不同算法在KL散度指标上的对比情况. 图4(c)和图4(d)则展示了在相应数据集中,随机选取的8个工人估计权重值与真实权重值的比较结果,其中添加Laplace噪声后的估计权重值(Estimated Weight, EW)与非隐私下的真实权重值(True Weight, TW). 横坐标表示工人编号(ID),纵坐标表示相对应工人的权重值.



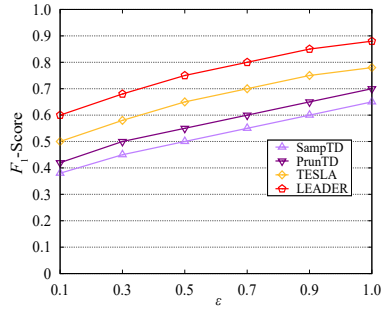
(a) Int: MAE Change



(b) Wea: MAE Change



(c) Int: F_1-Socre



(d) Wea: F_1-Socre

图3 隐私预算ε的影响

如图4(a)和图4(b)所示,在不同隐私预算下,SampTD和PrunTD算法由于在处理离群点时存在不足,导致工人权重估计误差较大,而相比之下,TESLA算法通过针对不同噪声类型的处理,能够有效降低离群点数据的影响. 与之相比,本文提出的LEADER算法在所有隐私预算设置下,计算的权重估计值与真实值之间的KL-Divergence分布都较小,这表明LEADER算法在隐私保护下的权重分布更接近真实分布. 其原因在于高质量工人提供的数据受到较小的噪声干扰,同时Huber损失度量能够有效抑制离群点带来的负面影响. 图4(c)和图4(d)的实验结果进一步验证了这一点,且在不同数据集以及任务类型下,LEADER算法的表现始终优于其他对比算法.

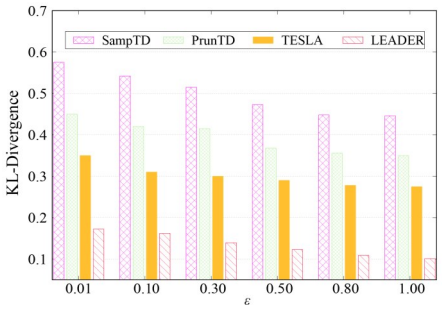
(3)平滑参数σ的影响

如图5(a)和图5(b)所示,本文通过在2个真实数据集上实验,验证了Huber损失中参数σ对LEADER真值发现算法的影响. 实验结果表明,当$\sigma=1$时,LEADER算法的性能最佳,能够有效抑制噪声和离群点的干扰,从而提高真值发现的精度. 其中,较小的σ值(如$\sigma=0.01$)对离群值过于敏感,放大了异常点的影响,导致MAE Change较高. 随着σ增大到合适范围(如$\sigma=1$),对离群值的鲁棒性增强,MAE Change达到最小,此时效用最高. 若σ过大(如$\sigma=3$),Huber损失趋近于均方误差(MSE),对异常值的处理能力下降,导致MAE Change再次增大. 此外,本文发现较大的σ性能比较小的σ差,这是因为较大的σ值会导致Huber损失函数趋近于MSE,从而减弱对离群点的抑制能力,降低模型的鲁棒性,导致性能下降.

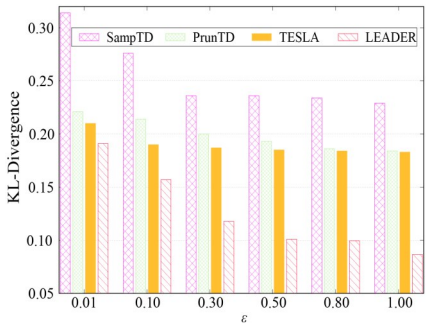
Huber损失能有效平衡绝对值误差L_1和均方误差L_2损失,其中参数σ决定了离群值的影响范围. 本文通过实验选择合适的$\sigma=1$,使得LEADER算法在鲁棒性与效用间找到最佳平衡点.

(4)工人分组数O的确定

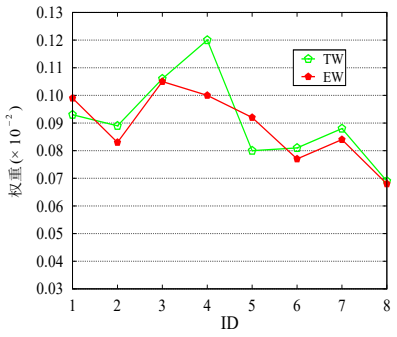
如图6所示,本文在2个真实数据集上,采用肘部法结合K-means聚类算法,确定了工人分组的最优分组数O. 具体而言,本文分别在Int数据集和Wea数据集上计算了不同分组数O下的误差平方和(SSE). 随着分组



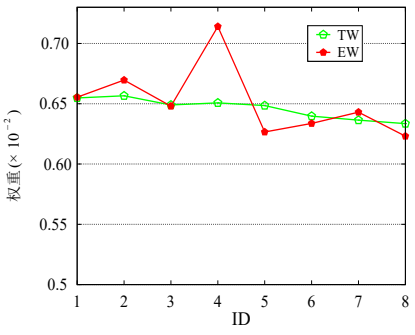
(*a*) Int: KL-Divergence对比



(*b*) Wea: KL-Divergence对比



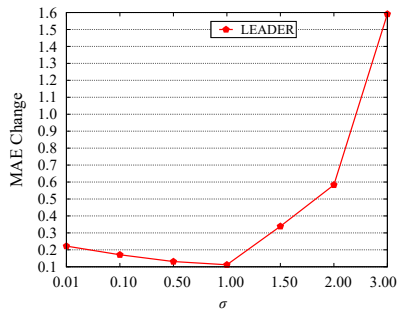
(*c*) Int: 权重对比



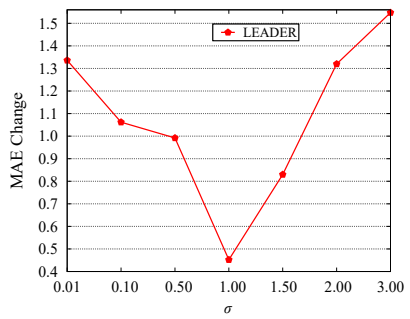
(*d*) Wea: 权重对比

图4　对工人权重分布的影响

数的增加,SSE逐渐减小,但在$O=3$时,SSE的下降趋势显著减缓,肘部特征明显. 因此,本文将工人分组数O确定为3.

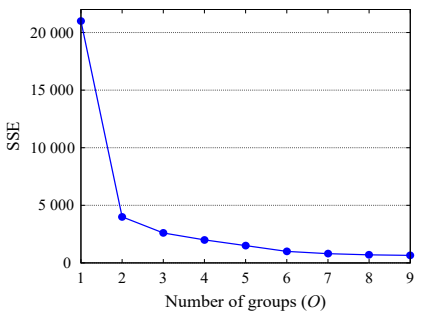


(*a*) Int数据集

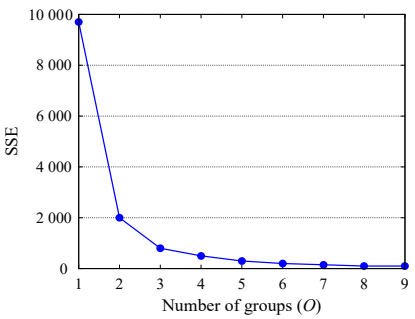


(*b*) Wea数据集

图5　平滑参数σ的影响



(*a*) Int数据集



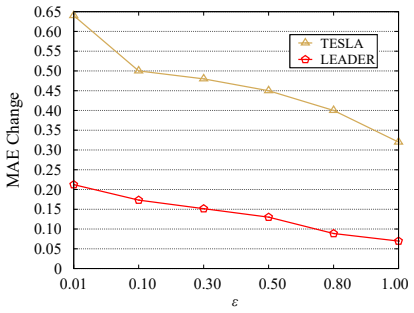
(*b*) Wea数据集

图6　工人分组数O的确定

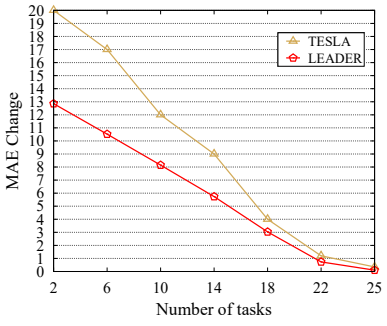
(5)在合成数据集Syn上的实验

图7(*a*)和图7(*b*)表明了LEADER算法在合成数据

集Syn上的有效性. 首先,如图7(*a*)所示,随着隐私预算 ε 的增大,LEADER算法始终优于TESLA,表明LEADER算法在合成数据集Syn上同样表现出了较好的效果,能够有效降低噪声对真值发现的影响. 其次,在图7(*b*)中,随着工人完成任务数量的增大,MAE Change逐渐变小,这表明任务完成数量越多,数据稠密度越高,算法的真值估计精度越高. 但在任务数量较少、数据较稀疏时,LEADER算法的MAE Change明显低于TESLA,表现出更好的鲁棒性. 这是因为LEADER能够精准建模任务的重要性,并对关键任务给予更高权重,从而减少了稀疏数据对真值估计结果精度的负面影响. 通过合理分配任务权重和工人分组策略,LEADER算法能够进一步抑制噪声和异常值(离群点)的影响,实现隐私保护与数据精度之间的平衡. 在隐私预算变化的情况下,LEADER也能有效降低噪声引入的影响,提高数据恢复精度从而展示了其在隐私保护与准确性之间的优越平衡.



(*a*) ε的变化



(*b*) 工人完成任务的数量

图7 Syn数据集上的性能比较

表3结果表明,高斯噪声(Gauss)下的MAE Change低于Laplace噪声,说明高斯噪声对真值发现的干扰较小. LEADER在2种噪声下均表现最佳,TESLA次之,而PrunTD和SampTD误差较大,尤其是SampTD,表明随机采样带来较大信息损失. Huber损失的引入使LEADER在不同噪声环境下均能有效降低误差,特别是在Laplace噪声下,其MAE Change远低于其他算法,表明其对高幅度噪声的抑制效果较好. 在高斯噪声下,LEADER误差进一步降低,验证了Huber损失的稳健性,能适应不同类型噪声并抑制离群点影响.

表3 Syn: 不同噪声类型的鲁棒性

噪声类型	对比算法	MAE Change
Laplace	SampTD	0.792 0
	PrunTD	0.560 6
	TESLA	0.453 0
	LEADER	0.130 1
Gauss	SampTD	0.682 1
	PrunTD	0.463 6
	TESLA	0.343 2
	LEADER	0.080 1

如表4所示,LEADER算法随着数据集规模的增加,运行时间呈线性增长,符合时间复杂度分析的预期,但MAE Change并未显著增加,整体保持稳定,说明算法在大规模数据集上依然能够有效抑制噪声影响,保持较高的精度. 进一步验证了LEADER算法在大规模MCS系统中的可扩展性和稳定性.

表4 不同规模Syn数据集上的性能

	(M=100, N=50)	(M=500, N=200)	(M=1 000, N=500)	(M=5 000, N=1 000)
MAE Change	0.649	0.568	0.652	0.635
t/s	1.60	27.2	111.2	567.4

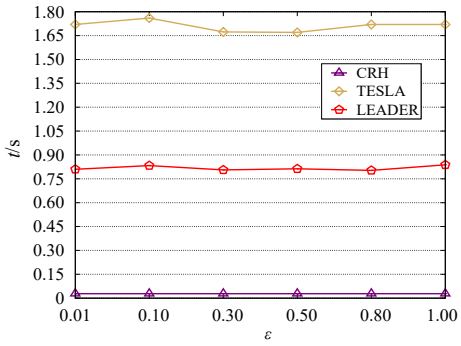
(6)运行时间

图8展示了在不同隐私预算ε下,非隐私保护下经典的真值发现算法CRH、对比算法TESLA和本文提出的LEADER算法的运行时间(以s为单位)的分布对比. 首先,如图8(*a*)与图8(*b*)所示,CRH算法的运行时间保持稳定,不受隐私预算变化的显著影响. 相比之下,LEADER算法在引入隐私保护机制的情况下,运行时间相对稳定,约0.79 s,这表明LEADER算法在不同隐私预算下保持一致的时间复杂度,具有良好的可扩展性. 相较于传统算法CRH,LEADER算法虽然增加了隐私保护,但其运行时间仍处于秒级范围,在可接受范围内,体现了其在隐私保护和计算效率之间的良好平衡. 相比于最新的对比算法TESLA,LEADER运行时间大大缩短. 因此,LEADER算法不仅能够提供有效的隐私保护,还能够在大规模数据处理场景中保持较高的运行效率,适应实际应用需求.

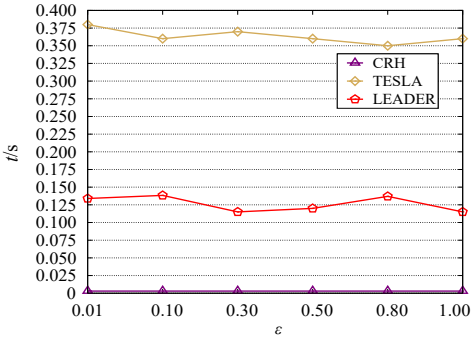
6.5 消融实验

(1)Huber损失的有效性

图9(*a*)和图9(*b*)说明了Huber损失距离度量方法的有效性. EucDist和ManDist方法是LEADER的变体,

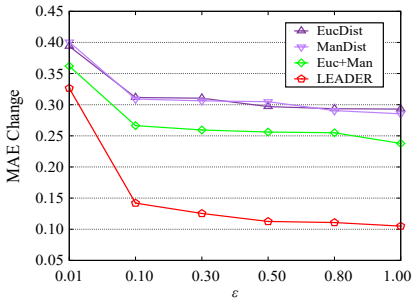


(a) Int数据集

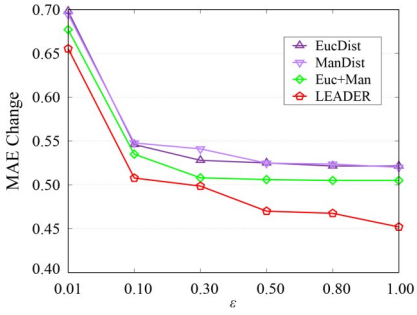


(b) Wea数据集

图8　运行时间



(a) Int数据集



(b) Wea数据集

图9　Huber损失的有效性

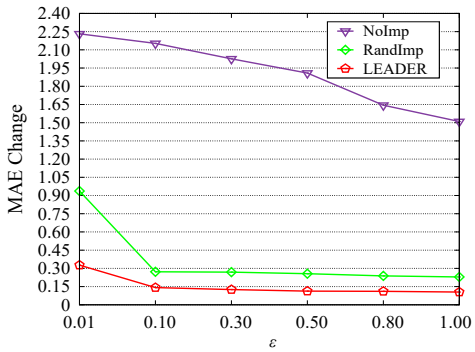
分别使用欧氏距离和曼哈顿距离来替换Huber损失度量加噪感知值和任务“真值”之间的距离,同时也充分考虑了工人分组和任务重要性. 首先,本文观察到,EucDist和ManDist方法效用相对较差且相当. 因为,由于Laplace噪声导致离群点的存在,而欧氏距离平方项对异常值特别敏感,因此会在加噪数据中放大离群点的影响. 而曼哈顿距离在处理高维数据时对所有维度的变化一视同仁,忽略了任务重要性维度的特性,因此在多任务场景下的表现较差. 此外,在高维稀疏数据中,两种方法可能面临样本区分能力不足和鲁棒性差的问题. 其次,Euc+Man距离度量方法优于EucDist和ManDist方法,因为它结合了两者的优点,有助于处理离群点和复杂数据分布,但由于无法确定二者的权重分布,数据集之间的差异可能影响整体表现. 第三,随着隐私预算ε的增大,基于Huber损失的LEADER方法性能总是最优的. 这是因为Huber损失在小误差时采用二次损失与欧氏距离相似,能够增强平滑性;大误差时切换为线性损失与曼哈顿距离相似,从而有效抑制离群点的影响. 相比单一度量和组合距离,Huber损失能够动态调整,更好地平衡鲁棒性与效用,尤其在不同数据集上能够自适应地应对不同噪声特性和数据分布.

(2)任务重要性y_n的有效性

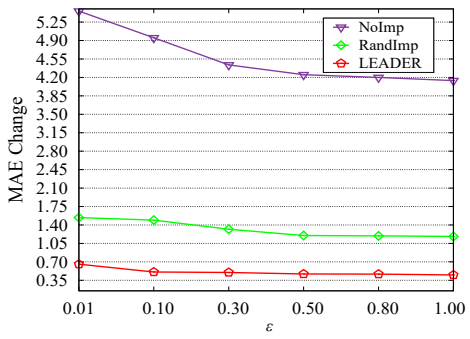
图10(a)和图10(b)说明了任务重要性在LEADER算法中有效性. NoImp方法和RandImp方法是LEADER的变体,分别是未考虑任务权重和任务权重随机分配的方法. 实验结果表明,随着隐私预算的增大,RandImp方法在性能上优于NoImp. 这是因为RandImp方法通过随机分配任务权重,考虑到了任务的重要性对真值发现的影响. 而NoImp方法则将所有任务被视为同等重要,导致算法无法区分关键任务和次要任务,从而降低了关键任务的真值精度. RandImp在一定程度上打破了任务权重固定分配的限制,从而提升了整体任务的真值精度. 其次,LEADER算法通过迭代优化来估计任务的重要性权重,相较于随机分配,显著提高了真值发现的精度. 这是因为LEADER算法基于数据驱动的方式迭代求解,在计算过程中考虑了各任务的重要性和数据分布之间的关系,从而更加准确地为任务分配权重. 特别是在任务数据质量差异较大的情况下,LEADER能更有效地分配资源,将更多的权重赋予关键任务,从而避免了噪声较大的任务对整体结果的负面影响. 因此,LEADER迭代优化方法显著优于随机分配和无任务重要性方法.

(3)工人分组的有效性

图11(a)和图11(b)说明了工人分组的有效性. NoGroup方法和FixGroup方法是LEADER的变体,分别是无工人分组和固定工人分组的方法. 实验结果表明,随着隐私预算ε的增大,首先FixGroup方法优于NoGroup,

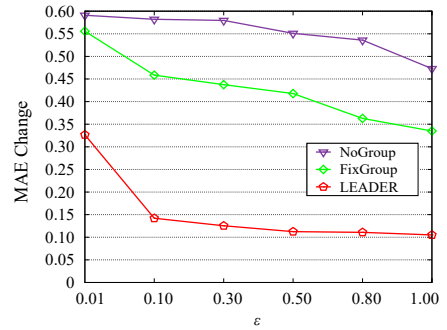


(*a*) Int数据集

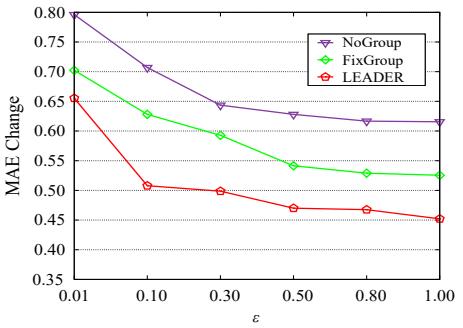


(*b*) Wea数据集

图10 任务重要性的有效性



(*a*) Int数据集



(*b*) Wea数据集

图11 工人分组的有效性

这是因为固定分组方法能够有效考虑工人之间的相似性,降低数据异质性对结果的负面影响,同时提高求得"真值"的精度. 工人被分为多个相似性较高的组,因此离群点工人很可能被分配到单独的组或一小部分组中,从而减少了这些工人对整体任务精度的影响. 分组后的组内数据更均匀,能够更有效地聚合高质量数据,有效地弱化了离群点的影响. 其次,LEADER算法的性能总是优于NoGroup方法和FixGroup方法,这是因为该算法动态地调整工人分组,通过优化分组与任务的重要性结合,组内的工人更加相似,从而更精准地匹配高质量工人与关键任务,聚合出更稳定、更准确的任务真值结果.

7 总结和展望

本文提出了一种面向离群点的LDP真值发现算法LEADER,用于在隐私保护与效用保证的条件下实现高精度的真值发现. 在LEADER中,为了应对Laplace加噪后产生离群点的影响,设计了距离度量和数据度量两种方法. 具体来说,为了度量加噪感知数据与估计真值之间的差异,设计了一种基于Huber损失的距离度量方法,有效提升了噪声"真值"的精度. 同时,从数据度量角度出发,为了优化工人的权重分配和任务的重要性,设计了一种基于工人分组和拉格朗日乘子法的优化算法,并结合坐标下降算法进行高效求解,从而提升了噪声"真值"的鲁棒性和算法的整体效能. 通过理论分析证明了LEADER在隐私、效用和复杂度上的优越性,并通过在两个真实数据集和一个合成数据集上的实验验证了LEADER的有效性. 事实上,该研究不仅为LDP条件下的真值发现问题提供了新的思路,还为高维稀疏数据集中的真值发现场景提供了有效的解决方案. 在未来的研究中,将进一步探索任务动态变化时的隐私保护真值发现方法.

参考文献

[1] 蒋伟进, 王海娟, 周为, 等. 基于自适应连续时间的群智感知轨迹隐私保护方案[J]. 电子学报, 2023, 51(10): 2894-2901.
JIANG W J, WANG H J, ZHOU W, et al. Track privacy protection scheme based on adaptive continuous time in crowdsensing[J]. Acta Electronica Sinica, 2023, 51(10): 2894-2901. (in Chinese)

[2] 童咏昕, 袁野, 成雨蓉, 等. 时空众包数据管理技术研究综述[J]. 软件学报, 2017, 28(1): 35-58.
TONG Y X, YUAN Y, CHENG Y R, et al. Survey on spatiotemporal crowdsourced data management techniques[J]. Journal of Software, 2017, 28(1): 35-58. (in Chinese)

[3] ZHANG P F, CHENG X, SU S, et al. Area coverage-

based worker recruitment under geo-indistinguishability[J]. Computer Networks, 2022, 217: 109340.

[4] WANG T C, LV C M, WANG C T, et al. A secure truth discovery for data aggregation in mobile crowd sensing[J]. Security and Communication Networks, 2021, 2021: 2296386.

[5] LI Q, LI Y L, GAO J, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2014: 1187-1198.

[6] AHMAD JALALI N, CHEN H S. Federated learning incentivize with privacy-preserving for IoT in edge computing in the context of B5G[J]. Cluster Computing, 2024, 28. DOI:10.1007/s10586-024-04788-7.

[7] 陈红松, 孟彩霞, 刘书雨. 基于经验小波变换的基因关联隐私保护实验研究[J]. 湖南大学学报(自然科学版), 2020, 47(2): 125-133.
CHEN H S, MENG C X, LIU S Y. Privacy protection experimental research on genes association ranking based on empirical wavelet transform[J]. Journal of Hunan University (Natural Sciences), 2020, 47(2): 125-133. (in Chinese)

[8] FENG J, WU Y F, SUN H, et al. Panther: Practical secure two-party neural network inference[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 1149-1162.

[9] 叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述[J]. 软件学报, 2018, 29(7): 1981-2005.
YE Q Q, MENG X F, ZHU M J, et al. Survey on local differential privacy[J]. Journal of Software, 2018, 29(7): 1981-2005. (in Chinese)

[10] HUANG W, ZHOU S J, ZHU T Q, et al. Improving Laplace mechanism of differential privacy by personalized sampling[C]//2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Piscataway: IEEE, 2020: 623-630.

[11] LI T P, WANG L, PENG D H, et al. Causal discovery evaluation framework in the absence of ground-truth causal graph[J]. IEEE Access, 2024, 12: 136502-136514.

[12] FANG X S, WANG X Z, SHENG Q Z, et al. Generalizing truth discovery by incorporating multi-truth features[J]. Computing, 2024, 106(5): 1557-1583.

[13] BAI J, GUI J S, HUANG G S, et al. UAV-supported intelligent truth discovery to achieve low-cost communications in mobile crowd sensing[J]. Digital Communications and Networks, 2024, 10(4): 837-852.

[14] WANG H, LIU A F, XIONG N N, et al. TVD-RA: A truthful data value discovery-based reverse auction incentive system for mobile crowdsensing[J]. IEEE Internet of Things Journal, 2024, 11(4): 5826-5839.

[15] KANG Y C, LIU A F, XIONG N N, et al. DTD: An intelligent data and bid dual truth discovery scheme for MCS in IIoT[J]. IEEE Internet of Things Journal, 2024, 11(2): 2507-2519.

[16] LOHRER A, KAZEMPOUR D, HÜNEMÖRDER M, et al. CoMadOut: A robust outlier detection algorithm based on CoMAD[J]. Machine Learning, 2024, 113(10): 8061-8135.

[17] RIAHI-MADVAR M, NASERSHARIF B, AZIRANI A A. Subspace outlier detection in high dimensional data using ensemble of PCA-based subspaces[C]//2021 26th International Computer Conference, Computer Society of Iran (CSICC). Piscataway: IEEE, 2021: 1-5.

[18] YANG P, WANG D, WEI Z J, et al. An outlier detection approach based on improved self-organizing feature map clustering algorithm[J]. IEEE Access, 2019, 7: 115914-115925.

[19] GAO X, YU J H, ZHA S, et al. An ensemble-based outlier detection method for clustered and local outliers with differential potential spread loss[J]. Knowledge-Based Systems, 2022, 258: 110003.

[20] 刘财辉, 刘地金. 离群点检测的邻近性方法综述[J]. 计算机工程与应用, 2022, 58(21): 1-12.
LIU C H, LIU D J. Survey of proximity methods for outlier detection[J]. Computer Engineering and Applications, 2022, 58(21): 1-12. (in Chinese)

[21] KNORR E M, NG R T, TUCAKOV V. Distance-based outliers: Algorithms and applications[J]. The VLDB Journal, 2000, 8(3): 237-253.

[22] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 427-438.

[23] SCHUBERT E, RENZ M, KRÖGER P. Generalized outlier detection with flexible kernel density estimates[C]//Proceedings of the 14th SIAM International Conference on Data Mining. Philadelphia: SIAM, 2014: 542-550.

[24] CÁRDENAS-MONTES M. Depth-based outlier detection algorithm[C]//Hybrid Artificial Intelligence Systems. Cham: Springer International Publishing, 2014: 122-132.

[25] SUN H P, DONG B X, WANG H, et al. Truth inference on sparse crowdsourcing data with local differential privacy[C]//2018 IEEE International Conference on Big Data (Big Data). Piscataway: IEEE, 2018: 488-497.

[26] LI Y L, MIAO C L, SU L, et al. An efficient two-layer mechanism for privacy-preserving truth discovery[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1705-1714.

[27] HUANG Y X, LIU F, ZHAO J C, et al. An incentive-based differential privacy-preserving truth discovery over

streaming data[C]//GLOBECOM 2022 IEEE Global Communications Conference. Piscataway: IEEE, 2022: 4848-4853.
[28] SUN P, WANG Z B, WU L T, et al. Towards personalized privacy-preserving incentive for truth discovery in mobile crowdsensing systems[J]. IEEE Transactions on Mobile Computing, 2022, 21(1): 352-365.
[29] ZHANG P F, CHENG X, SU S, et al. PrivTDSI: A local differentially private approach for truth discovery via sampling and inference[J]. IEEE Transactions on Big Data, 2023, 9(2): 471-484.
[30] ZHANG P F, CHENG X, SU S, et al. Effective truth discovery under local differential privacy by leveraging noise-aware probabilistic estimation and fusion[J]. Knowledge-Based Systems, 2023, 261. DOI: 1016/j.knosys.2022.110213.
[31] XIONG P, LI G R, LIU H Z, et al. Decentralized privacy-preserving truth discovery for crowd sensing[J]. Information Sciences, 2023, 632: 730-741.
[32] XU G W, LI H W, XU S M, et al. Catch you if you deceive me: Verifiable and privacy-aware truth discovery in crowdsensing systems[C]//Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. New York: ACM, 2020: 178-192.
[33] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Theory of Cryptography. Berlin: Springer, 2006: 265-284.
[34] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2013, 9(3/4): 211-487.

作者简介

朱伊波 男,1999年2月出生于安徽省蚌埠市. 现为安徽理工大学计算机科学与工程学院研究生. 主要研究方向为数据安全与隐私保护.
E-mail: 2023201103@aust.edu.cn

方贤进 男,1970年11月出生于安徽省六安市. 现为安徽理工大学计算机科学与工程学院教授. 主要研究方向为网络信息安全.
E-mail: xjfang@aust.edu.cn

张朋飞 男,1992年1月出生于河南省鄢陵县. 现为安徽理工大学计算机科学与工程学院讲师、研究生导师. 主要研究方向为数据隐私保护与可信人工智能.
E-mail: zpf.bupt@bupt.cn

孙 笠 男,1994年7月出生于河北省唐山市. 现为华北电力大学控制与计算机工程学院副教授、研究生导师. 主要研究方向为数据挖掘和机器学习.
E-mail: ccesunli@ncepu.edu.cn

姜 茸 男,1978年2月出生于云南省临沧市. 现为云南财经大学智能应用研究院副院长. 主要研究方向为数据安全与隐私保护、智能计算. 中国电子学会会员编号:E190035510M.
E-mail: jiangrong@ynufe.edu.cn