可穿戴装置个性化本地差分隐私保护方案

卢 岑[1],沈苏彬[2]
(1. 南京邮电大学 物联网学院,江苏 南京 210003;
2. 南京邮电大学 计算机学院,江苏 南京 210023)

摘 要:本地差分隐私(local differential privacy,LDP)可以对可穿戴装置(wearable devices)采集到的数据进行隐私保护,每个用户都会在本地扰乱自己的数据,并且将扰动后的数据发送给数据汇聚服务器,以保护用户免受私人信息泄漏的影响。可穿戴装置采集到的数据是多维的,但是现有的针对可穿戴装置多维数据的个性化本地差分隐私保护研究比较少而且不完善。针对现有个性化本地隐私方案存在的最坏情况下噪声方差大的问题,采用结合机制,结合随机响应机制和分段机制,对数值型数据进行扰动,提出了一种处理数值型数据的个性化本地差分隐私保护方案,并将该方案应用到多维数值型数据,通过随机采样提高数据可用性。此外,分别从理论分析和仿真验证的角度对提出的本地差分隐私方案与现有解决方案进行了对比分析和实验。实验结果表明,提出的方案在最坏情况下的噪声方差方面优于现有解决方案,并且具有更好的数据可用性。

关键词:本地差分隐私;个性化;多维数据;可穿戴装置;随机响应
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2022)02-0107-07
doi:10.3969/j.issn.1673-629X.2022.02.017

Personalized Local Differential Privacy Protection Scheme for Wearable Devices

LU Cen[1],SHEN Su-bin[2]
(1. School of Internet of Things,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2. School of Computer,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:Local differential privacy (LDP) can protect the privacy of data collected by wearable devices. Each user will disturb his own data locally and send the disturbed data to the data aggregation server,to protect users from the impact of private information leakage. The data collected by the wearable device is multi-dimensional,but the existing research on personalized local differential privacy protection for the multi-dimensional data of the wearable device is relatively few and incomplete. Aiming at the problem of large noise variance in the worst case of the existing personalized local privacy schemes,the combination mechanism,combined with the random response mechanism and the piecewise mechanism,is used to disturb the numerical data,and a personalized local differential privacy protection scheme for processing numerical data is proposed. The scheme is applied to multi-dimensional numerical data,and data availability is improved through random sampling. In addition,comparative analysis and experiments are carried out on the proposed local differential privacy scheme and the existing solutions from the perspectives of theoretical analysis and simulation verification. Experimental results show that the proposed scheme is better than existing solutions in terms of noise variance in the worst case,and has better data availability.

Key words:local differential privacy;personalization;multidimensional data;wearable devices;random response

0 引 言

随着可穿戴技术的发展,可穿戴装置成为了人体卫生和保健的数据源,能不断监测和传递用户的生命体征数据,例如血压、心率、体脂等数据,同时还能测量运动过程中的卡路里消耗、步伐、心率和速度等[1]。医疗机构和健康机构通过收集并分析这些数据来为用户提供更好的服务。但是,就设备安全性和公众的隐私接受度而言,可穿戴装置还不成熟。2016 年,欧盟通

收稿日期:2021-03-14 **修回日期**:2021-07-15
基金项目:国家自然科学基金(61502246);南京邮电大学科研项目(NY220202)
作者简介:卢 岑(1997-),女,硕士,研究方向为物联网隐私保护;沈苏彬,博导,研究员,CCF 高级会员(E2000054B2s),研究方向为物联网及其应用、未来网络及其应用。

过了《一般数据法案》(general data protection regulation,GDPR),该法规规定了个人数据保护跨越国界,明确了用户的知情权以及个人数据隐私的保护[2]。然而,可穿戴装置中的嵌入式传感器通常可在未征得用户同意的情况下采集和获取个人以及周围环境的数据,这种情况会侵犯用户的隐私并违反相关法规。

针对隐私量化和隐私保护的需求,研究者提出了差分隐私技术[3-5],根据第三方数据汇聚服务器是否可信,差分隐私可分为中心化差分隐私和本地差分隐私。中心化差分隐私假设第三方是可信的,每个用户将自己的真实数据发送给数据汇聚服务器,然后数据汇聚服务器通过满足差分隐私的扰动算法对数据进行处理。然而,并不是所有的第三方都是可信的。针对第三方不可信的情况,本地差分隐私通过在用户端对真实数据进行扰动,然后将扰动后的数据汇聚到数据服务器中保护用户的数据隐私安全[6]。

但是本地差分隐私为所有个人提供了相同级别的隐私保护,每个用户对于其数据可接受的隐私级别的期望却不相同,这可能导致某些用户的隐私保护不足,而其他用户则受到过度保护。因此,在用户本地对数据进行数据扰动时,应该允许用户个性化地设置自己的隐私偏好,实现个性化的隐私保护。目前的个性化本地差分隐私存在两个问题,第一,大部分个性化差分隐私都是针对一维数值型数据的,而可穿戴装置收集的数据存在多个数值型属性,是多维的。第二,现有的个性化差分隐私都是通过随机响应机制或者添加噪声(主要是拉普拉斯噪声)实现的,将其应用于可穿戴装置中会产生隐私保护程度低和数据可用性低等问题。

在现有本地差分隐私保护方法的基础上,该文提出了一种可穿戴装置个性化本地差分隐私保护方案,允许用户设置自己的隐私偏好,实现对可穿戴装置多维数值型数据的个性化本地差分隐私。同时采用结合机制,结合随机响应机制和分段机制,解决随机响应机制最坏情况下噪声方差大的问题,提高可穿戴装置对用户多维数值型数据的隐私保护,并且提高数值型数据的数据可用性,通过理论验证和实验仿真证明可穿戴装置个性化本地差分隐私保护方案的有效性。

1 相关工作

随着可穿戴技术的发展,可穿戴装置中的数据隐私问题受到越来越多的关注。对可穿戴装置数据的攻击可分为被动攻击或主动攻击两种,被动攻击的基本目标是访问网络中共享的一定数量的私有数据或从公共数据集中推断出任何关键信息[7]。为了克服隐私量化和背景攻击等隐私问题,2006 年引入了一种重要的隐私方法,称为差分隐私。差分隐私通过添加所需的噪声量并在隐私和准确性之间保持健康的平衡来保护统计数据或实时数据。而对于不可信的第三方数据收集者,许多学者提出了本地差分隐私(LDP)[8-9],本地差分隐私防止了数据管理者对确切的私人数据的收集。

LDP 可以通过传统的随机响应技术实现,Erlingsson 等[10]提出了 RAPPOR 框架,该框架基于发布二进制属性的随机响应机制,他们将这种机制与 Bloom 过滤器结合使用,Bloom 过滤器直观地增加了另一级的保护,并增加了对手推断私人数据的难度。后续论文[11]将 RAPPOR 扩展到更复杂的统计数据,例如联合分布和关联测试以及包含大量潜在值的分类属性。但是 RAPPOR 通信开销大,不适合用在可穿戴装置中。Wang 等[12]研究了相同的问题,并提出了不同的方法,他们将 k 个可能的值转换为具有 k 个元素的噪声向量,并将后者发送给数据收集者。Bassily 和 Smith[13]提出了一个渐进最优解,用于在 LDP 下建立大分类域上的频率分布直方图。但是,上述所有方法都集中在单个分类属性上,与文中多维数值型数据研究工作不同。Ren 等[14]研究了发布多维属性的问题,并采用了 k-size 向量的思想(类似于文献[12]),但是这种方法在数据收集者和用户之间需要相当高的通信成本,因为它涉及多个 k 大小矢量的传输。Kairouz 等[15]提出了极值机制,这是离散输入数据的 LDP 机制,即每个输入域 X 包含有限数量的可能值,这些机制的输出分布具有关键属性。

因为 LDP 能很好地保护用户数据的隐私,故在室内定位数据的收集[16]、移动感知的推理控制[17]以及众包数据的发布[18]等应用中都有考虑。可穿戴装置本地差分隐私应用方面,马方方等[19]提出了可穿戴装置多维数值型数据个性化隐私保护方案(personalized local privacy scheme,PLPS),使用安全域对敏感数据进行规范化,最后使用伯努利分布对分组的多维数据进行扰动,并使用属性安全域恢复干扰结果。马方方等提出的方法比 Harmony 算法具有更低的最大相对误差,但是当 ε 值大于 2 时,噪声方差会趋于 1,不会随着 ε 的增大而减小。涂子璇[20]针对可穿戴装置的数值型流数据均值发布,为防止用户的隐私信息泄露提出一种基于自适应采样的可穿戴装置差分隐私均值发布方法。

在个性化差分隐私方面,Mousumi Akter[21]提出了一种新颖的方法,即数字聚合的私有估计(private estimation of numeric aggregates,PENA),在确保个性化的本地差分隐私的同时计算数字数据的聚合,但是该方法只适用于一维数值型数据。Datong Wu[22]根据

LDP 和用户的个性化要求提供了新颖的隐私定义,并展示了机制的最佳效用和隐私保证,但是提出的机制只适用于空间数据,也就是说只针对于位置的隐私保护。

2　问题的分析与描述

可穿戴装置的数据收集模型如图 1 所示。可穿戴装置首先通过传感器收集用户的各种数据,然后通过蓝牙与移动设备相连,将数据传输到移动设备中,最后第三方数据汇聚服务器收集各个移动设备的数据。

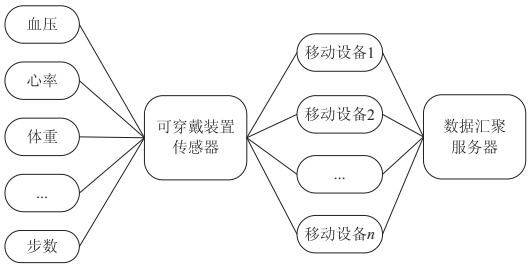

图 1　可穿戴装置数据收集模型

本地差分隐私是基于中心化差分隐私提出的数据收集框架,不同于中心化差分隐私对于可信第三方的假设,其针对的是不可信的第三方数据收集者(也就是图 1 中的数据汇聚服务器),本地差分隐私定义如下:

定义 1:本地差分隐私。给定 n 个用户,每个用户对应一条记录,给定一个隐私算法 M 及其定义域 Dom(M)和值域 Rom(M)。若算法在任意两条记录 t 和 $t'(t' \subseteq \mathrm{Rom}(M))$ 上得到相同输出结果 $t^*(t^* \subseteq \mathrm{Rom}(M))$ 并满足下列不等式,则 M 满足 ε -本地差分隐私。

$$\Pr[M(t) = t^*] \leqslant e^{\varepsilon} \times \Pr[M(t') = t^*]$$

同时,个性化本地差分隐私定义如下:

定义 2:个性化本地差分隐私。给定 n 个用户,用户 u_i 的隐私设置偏好为 ε_i,对于任意两个输入 t 和 t' 和任意的输出结果 t^* 满足下列不等式,则 M 满足个性化本地差分隐私。

$$\Pr[M(t) = t^*] \leqslant \mathrm{MAX}(e^{\varepsilon_i}) \times \Pr[M(t') = t^*]$$

目前关于可穿戴装置个性化数值型数据的研究比较少,马方方等提出的可穿戴装置个性化本地差分隐私方案(personalized local privacy scheme,PLPS)采用随机响应机制(local random response,LRR)对数据进行扰动,但因为随机响应机制总返回 $t_i^* = \dfrac{e^{\varepsilon}+1}{e^{\varepsilon}-1}$ 或 $t_i^* = -\dfrac{e^{\varepsilon}+1}{e^{\varepsilon}-1}$,输出的噪声值 t_i^* 总是具有绝对值 $\left|\dfrac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\right| > 1$,无论隐私预算多大,$t_i^*$ 在最坏情况下的噪声方差总是大于 1。PLPS 的方差如下式所示:

$$D(t_i^*) = E[(t_i^*)^2] - (E[t_i^*])^2 = \left(\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\right)^2 - t_i^*$$

当 $t_i = 0$ 时得到最坏情况下的噪声方差为 $D_w(t_i^*) = \left(\dfrac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\right)^2$,最坏情况下的噪声方差曲线如图 2 所示,PLPS 在最坏情况下的噪声方差总是大于 1。而噪声方差的大小会直接影响扰动数据的方差大小,扰动数据的方差越小差分隐私保护越成功。

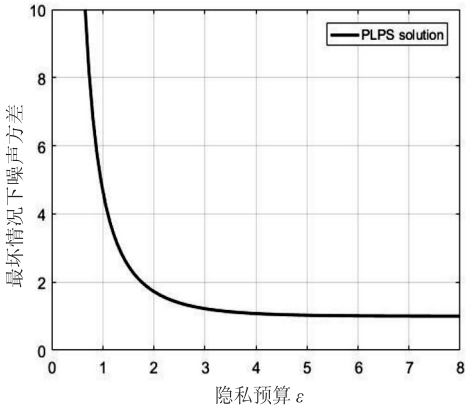

图 2　PLPS 在最坏情况下的噪声方差

3　可穿戴装置个性化本地差分隐私保护方案

3.1　方案设计

设计方案的隐私保护目标:对可穿戴设备多维数值型数据进行个性化的隐私保护,在保护用户隐私的前提下,减小最坏情况下的噪声方差,同时保证数据均值估计的可用性。

针对 PLPS 中存在的最坏情况下噪声方差大的问题,采用结合机制解决,结合机制结合随机响应机制和分段机制,具体描述如下:

结合机制
输入:原始元组 $t_i \in [-1,1]$ 和隐私预算 ε
输出:扰动后的元组 t_i^*
(1) if $\varepsilon < 0.6$ then
(2)　选择随机响应机制对数据进行扰动
(3) else
(4)　从[0,1]中随机取样得到 x
(5)　if $x < e^{-\varepsilon/2}$ then
(6)　选择随机响应机制对数据进行扰动
(7)　else
(8)　选择分段机制对数据进行扰动
(9) return t_i^*

当 $\varepsilon < 0.6$ 时,选择随机响应机制对数据进行扰动,否则从[0,1]中随机取样 x,当 $x < \mathrm{e}^{-\varepsilon/2}$ 时,选择随机响应机制对数据进行扰动,否则选择分段机制对数据进行扰动,分段机制描述如下:

分段机制

输入:原始元组 $t_i \in [-1,1]$ 和隐私预算 ε

输出:扰动后的元组 t_i^*

(1)从[0,1]中随机取样得到 x

(2)if $x < \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}$ then

(3)从 $[l(t_i), r(t_i)]$ 中随机采样得到 t_i^*

(4)else

(5)从 $[-C, l(t_i)) \cup (r(t_i), C]$ 中随机采样得到 t_i^*

(6)return t_i^*

结合机制在最坏情况下的噪声方差为:

$$D_{结}(t_i^*) = \begin{cases} \frac{3+e^{\varepsilon/2}}{3e^{\varepsilon/2}(e^{\varepsilon/2}-1)} + \frac{(e^{\varepsilon}+1)^2}{e^{\varepsilon/2}(e^{\varepsilon}-1)^2} & \varepsilon > 0.6 \\ \left(\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\right)^2 & \varepsilon \leqslant 0.6 \end{cases}$$

采用结合机制扰动数据和 PLPS 最坏情况下噪声方差的对比如图 3 所示。由对比图可以看到,无论 ε 怎么变化,采用结合机制扰动数据后在最坏情况下的噪声方差比 PLPS 小,也就是说,隐私保护程度比 PLPS 更好。

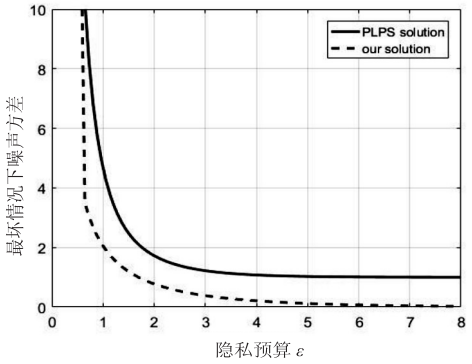

图 3 采用结合机制扰动数据和 PLPS 最坏情况下的噪声方差

大部分可穿戴装置差分隐私方法都是使用差分隐私的组合特性对单一属性进行直接扰动,而可穿戴装置通常收集的数值型数据是多维的,将结合机制直接应用到多维数值型数据中,会降低数据的可用性。为了提高数据的可用性,采用随机采样,每个用户提交 k 个属性,随机采样将每个属性的隐私预算从 ε/d 增加到 ε/k,从而减少了产生的噪声方差。但是从 d 个属性中采样 k 个会带来额外的估计误差,可以通过将 k 设置为适当的值来平衡,其中 $k = \max\{1, \min\{d, \lfloor \frac{\varepsilon}{2.5} \rfloor\}\}$。

3.2 方案描述

可穿戴装置个性化本地差分隐私保护方案描述如下:不可信数据汇聚服务器向可穿戴装置用户端发送整体隐私预算 ε,用来约束用户的隐私预算 ε_i。在可穿戴装置用户端,用户设置各自的隐私预算 ε_i($\varepsilon/\sqrt{d} \leqslant \varepsilon_i \leqslant \varepsilon$),因为均值估计算法的最大绝对误差渐进边界是 $O(\sqrt{d\log(\frac{d}{\beta})}/(\min \varepsilon_i \sqrt{n}))$,当 $\min \varepsilon_i = \varepsilon/\sqrt{d}$ 时达到最大绝对误差渐进边界,其中 ε 是全部隐私预算,对多维数值型数据设置属性域,使用属性域对数据进行归一化处理,把数值归一化到[-1,1]区间。然后,采用结合机制对数据进行扰动,为了提高数据的可用性,每个用户提交 k 个属性,不可信第三方服务器获取扰动后的数据,再对数据进行均值估计,最后进行归一化还原操作。

可穿戴装置个性化差分隐私保护方案

输入:用户集 $U = \{u_1, u_2, \cdots, u_i\}$,用户的数据集 $t_i = \{t_i[A_1], t_i[A_2], \cdots, t_i[A_d]\}$,$1 \leqslant i \leqslant n$,总隐私预算 ε,用户 u_i 的隐私预算 $\varepsilon_i = w_i \cdot \varepsilon$,$1/\sqrt{d} \leqslant w_i \leqslant 1$

输出:均值 z_j,$1 \leqslant j \leqslant d$

(1)for $i = 0$ to n do

(2) for $j = 0$ to d do

(3) $t_i^*[A_j]$ =结合机制($t_i[A_j]$, ε_i)

(4) end

(5)发送 $t_i^* = \{t_i^*[A_1], t_i^*[A_2], \cdots, t_i^*[A_k]\}$ 到服务器

// $k = \max\{1, \min\{d, \lfloor \frac{\varepsilon}{2.5} \rfloor\}\}$

(6)end

(7)服务器 Server:汇聚数据,求均值 $z_j = \frac{1}{n}\sum_{i=1}^{n} t_i^*[A_j]$,$1 \leqslant j \leqslant d$

3.3 方案隐私性和可用性分析

3.3.1 隐私性分析

用户设置自己的隐私预算为 ε_i,根据个性化本地差分隐私的定义,需要证明:

$$\frac{\Pr[M(t_i[A_j], \varepsilon_i) = t_i^*]}{\Pr[M(t_i^*[A_j], \varepsilon_i) = t_i^*]} \leqslant \mathrm{MAX}(e^{\varepsilon_i})。$$

(1)随机响应机制隐私性分析。

$$\frac{\Pr[\mathrm{LRR}(t_i[A_j], \varepsilon_i) = t_i^*]}{\Pr[\mathrm{LRR}(t_i^*[A_j], \varepsilon_i) = t_i^*]} \leqslant \frac{(\max t_i[A_j] \cdot (e^{\varepsilon_i}-1) + e^{\varepsilon_i} + 1)}{(\min t_i[A_j] \cdot (e^{\varepsilon_i}-1) + e^{\varepsilon_i} + 1)} = \frac{(1 \cdot (e^{\varepsilon_i}-1) + e^{\varepsilon_i} + 1)}{((-1) \cdot (e^{\varepsilon_i}-1) + e^{\varepsilon_i} + 1)} = e^{\varepsilon_i} \leqslant e^{\varepsilon}$$

因为 $e^{\varepsilon_i} \leqslant \mathrm{MAX}(e^{\varepsilon_i})$,所以随机响应机制满足个性化差分隐私。

(2)分段机制隐私性分析。

分段机制的概率密度函数为:

$$\mathrm{pdf}(t_i^* = x \mid t_i) = \begin{cases} p, & x \in [l(t_i), r(t_i)] \\ \frac{p}{e^{\varepsilon}}, & x \in [-C, l(t_i)) \cup (r(t_i), C] \end{cases}$$

其中,$p=\frac{e^{\varepsilon}-e^{\varepsilon/2}}{2e^{\varepsilon/2}+2}$。

由此可以得出:

$$\Pr[\mathrm{PM}(t_i[A_j],\varepsilon_i)=t_i^*]=p(C-1)+\frac{p}{e^{\varepsilon_i}}(C+1)$$

也就是说,无论如何,$\Pr[\mathrm{PM}(t_i[A_j],\varepsilon_i)=t_i^*]$ 都为一个常数。

$$\frac{\Pr[\mathrm{PM}(t_i[A_j],\varepsilon_i)=t_i^*]}{\Pr[\mathrm{PM}(t_i^*[A_j],\varepsilon_i)=t_i^*]}=1=e^0\leqslant \mathrm{MAX}(e^{\varepsilon_i})$$

所以根据个性化差分隐私的定义,分段机制同样满足个性化本地差分隐私。

由以上分析可以看出,无论是随机响应机制还是分段机制都满足个性化本地差分隐私,而文中的方案结合了随机响应机制和分段机制,根据差分隐私并行组合特性,文中的方案也满足个性化本地差分隐私。

3.3.2　可用性分析

由文献[18]可知,PLPS 的均值估计最大绝对误差为 $O(\sqrt{d\log(d)}/(\min\varepsilon_i\sqrt{n}))$,当 $\min\varepsilon_i=\varepsilon/\sqrt{d}$ 时,PLPS 的最大绝对误差渐进边界为 $O(\sqrt{d\log(d)}/(\varepsilon\sqrt{n}))$,而文中的方案最大绝对误差为 $O(\sqrt{d\log(\frac{d}{\beta})}/(\varepsilon\sqrt{n}))$(其中 $\beta=\frac{\max\{1,\min\{d,\lfloor\frac{\varepsilon}{2.5}\rfloor\}\}}{d}\geqslant 1$),也就是说最大绝对误差渐进边界为 $O(\sqrt{d\log(\frac{d}{\beta})}/(\varepsilon\sqrt{n}))\leqslant O(\sqrt{d\log d})/\varepsilon\sqrt{n}$,误差越小,则表明数据可用性越高。

通过上面的分析可以看到,文中的方案既满足个性化本地差分隐私,而且最大绝对误差小于 PLPS,在数据可用性方面优于 PLPS 方案。

4　实验结果和分析

实验研究属性个数 d、用户数 n 和隐私预算 ε 对数值型数据均值估计可用性的影响,并与 PLPS 进行了定量对比。实验采用的评价标准是最大绝对误差 MAE(maximum absolute error),$\mathrm{MAE}=\max_{1\leqslant j\leqslant d}|Z[A_j]-X[A_j]|$,$X[A_j]$ 是原始统计均值,$Z[A_j]$ 是第三方汇聚后计算获得的均值估计值。接下来,使用控制变量法(单一变量法)分别对隐私预算 ε、数据属性个数 d 及用户数 n 对 MAE 的影响进行实验和分析。

(1)隐私预算 ε 对 MAE 的影响。

为了研究隐私预算对可用性的影响,随机生成虚拟数据集,ε 取值为[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15],考虑数据属性个数 $d=10/20$ 和用户数 $n=100/10\ 000$ 的情况,不同隐私预算对 MAE 的影响,如图 4 所示。总体上看,MAE 随着隐私预算的增大而减小。这是因为,隐私预算本质上代表着用户对隐私的保护程度,隐私预算越大,代表用户想要对隐私保护的程度就越小,因此第三方收集者得到的用户数据就越准确,自然地,第三方收集者对原始数据的估计也就越准确,因此最大绝对误差也就会相应的更小。也就是说,如果 $\varepsilon\to\infty$,那么 $\mathrm{MAE}\to 0$。另一方面,从图 4 中可以明显看出,对于不同的隐私预算,文中的方案效果均优于 PLPS。当第三方收集者拿到扰动后的数据时,对于原始数据的均值估计,使用文中的方案更加准确。

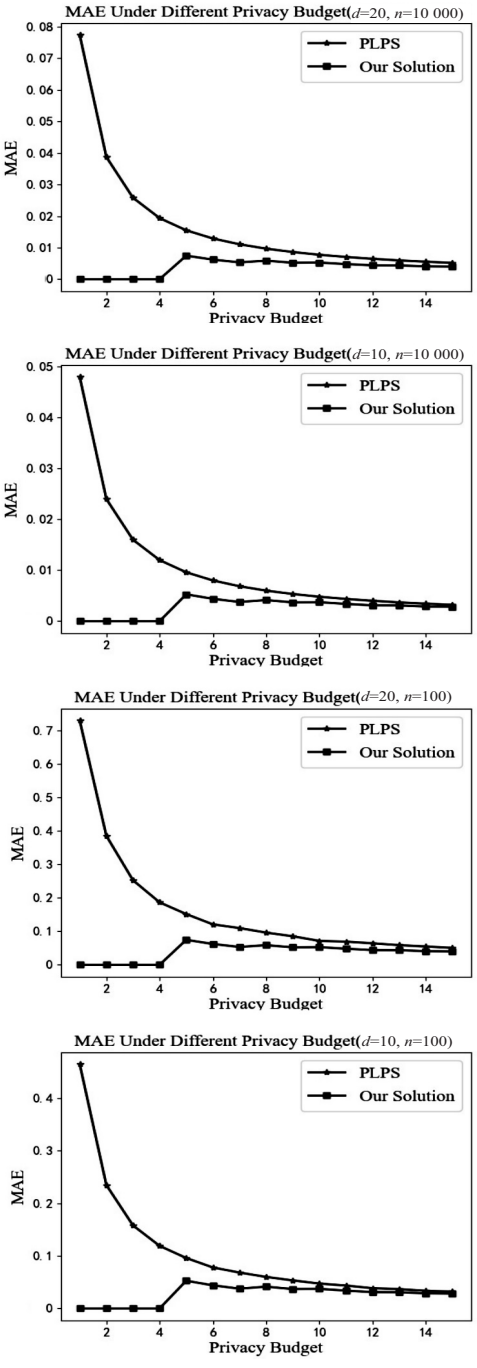

图 4　隐私预算对 MAE 的影响

(2)属性个数 d 对于 MAE 的影响。

为了研究属性个数对可用性的影响,随机生成虚拟数据集, d 取值为[5,10,15,20,25,30],考虑数据隐私预算 $\varepsilon = 5/0.5$ 和用户数 $n = 100/10\ 000$ 的情况,不同属性个数对 MAE 的影响,如图 5 所示。

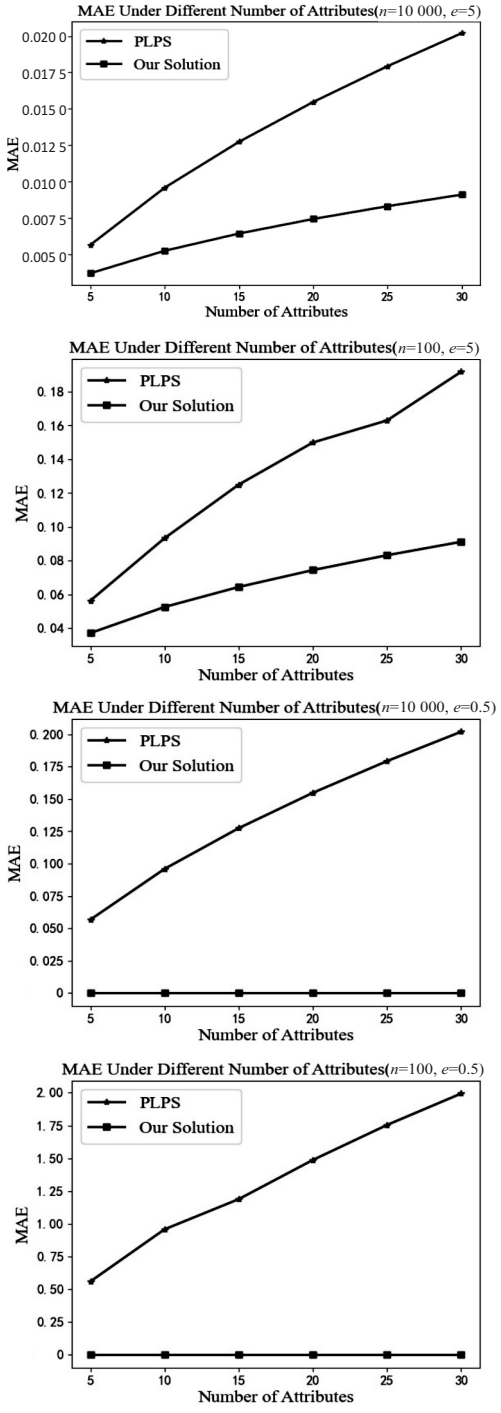

图 5 属性个数对 MAE 的影响

MAE 与属性个数呈正相关,即属性个数的增多会导致 MAE 增大,这本质上体现了数据维度的增加对于第三方数据收集者对原始数据整体估计值误差的积累过程。横向来看,文中的方案效果依然大幅度优于 PLPS。

(3)用户数 n 对 MAE 的影响。

为了研究用户数对可用性的影响,随机生成虚拟数据集, n 取值为[5 000,10 000,15 000,20 000,25 000,30 000,35 000,40 000,45 000,50 000,55 000,60 000],考虑数据隐私预算 $\varepsilon = 5/0.5$ 和属性个数 $d = 20/200$ 的情况,不同用户数对 MAE 的影响,如图 6 所示。

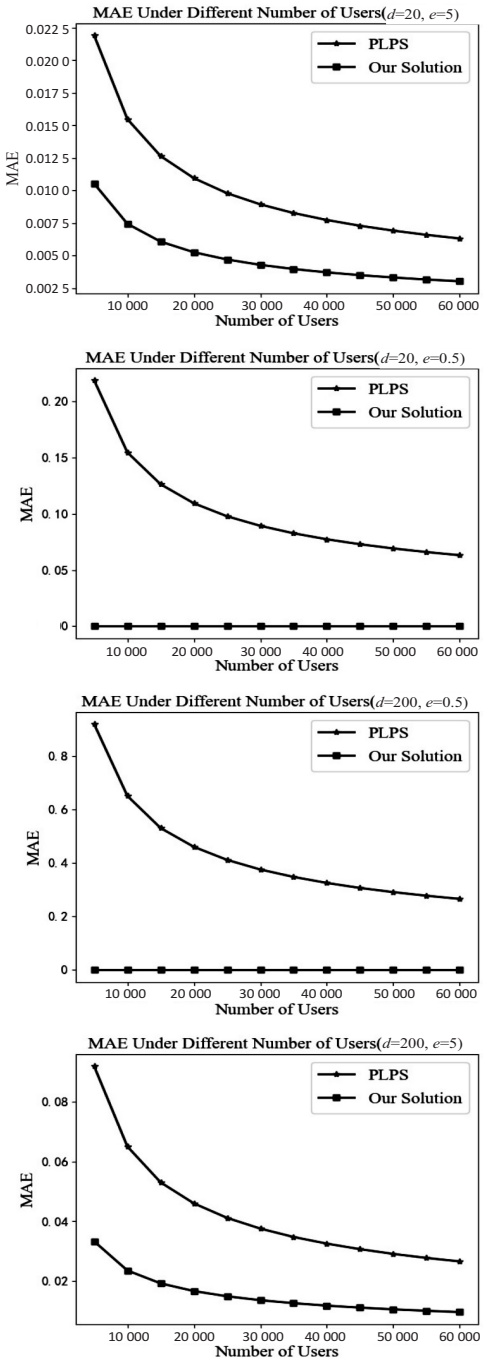

图 6 用户数对 MAE 的影响

图 6 展示了 MAE 随用户数量的变化规律。纵向来看,随着用户数量的增加 MAE 逐渐减小,因为 MAE 与用户数 n 的 1/2 次方呈反比例关系,本质上是由于用户对隐私预算的分摊。从另外一个角度也可以理解

为,随着用户数量的增加,第三方数据收集者能够获取的数据样本也就越多。因为无论是 PLPS 还是文中的方案,第三方数据收集者均可以对原始数据进行宏观统计量的无偏估计,因此数据量越多,宏观量的估计也就越精确。横向上看,文中的方案效果依然比 PLPS 好。

5　结束语

为了防止可穿戴装置用户隐私泄露,文中通过采用结合机制对数值型数据进行扰动,结合随机响应机制和分段机制减少最坏情况下的噪声方差,通过随机采样提高多维数据的数据可用性,并且针对不同用户的隐私需求提出了可穿戴装置个性化本地差分隐私保护方案。理论证明,文中方案满足了个性化本地差分隐私保护需求。仿真实验结果表明,采用文中方案对可穿戴装置多维数值型数据进行隐私保护,不仅能减小最坏情况下的噪声方差,而且拥有更高的数据可用性。但是文中方案的个性化是针对每个用户的所有属性相同保护程度,针对不同属性的个性化还需要进一步的研究。

参考文献:

[1] EHRANI K,ANDREW M. Wearable technology and wearable devices:everything you need to know[EB/OL]. (2015-9-18). http://www. wearabledevices. com/what-is-a-wearable-device/.

[2] 张　彭. 大数据安全背景下欧盟《通用数据保护条例(GDPR)》研究[D]. 上海:华东师范大学,2020.

[3] DWORK C. Differential privacy:a survey of results[C]// International conference on theory and applications of models of computation. Xi'an,China:Springer,2008:1-19.

[4] DWORK C,LEI J. Differential privacy and robust statistics [C]//Proceedings of the 41st annual ACM symposium on theory of computing. Bethesda,MD,USA:ACM,2009:371-380.

[5] SMITH A. Privacy-preserving statistical estimation with optimal convergence rates[C]//Proceedings of the annual ACM symposium on theory of computing. San Jose,California,USA:ACM,2011:813-822.

[6] 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.

[7] GIRALDO J, SARKAR E, CARDENAS A,et al. Security and privacy in cyber-physical systems:a survey of surveys [J]. IEEE Design & Test,2017,34(4):7-17.

[8] DUCHI J C,JORDAN M I,WAINWRIGHT M J. Local privacy and statistical minimax rates[C]//2013 IEEE 54th annual symposium on foundations of computer science. Berkeley,CA,USA:IEEE,2013:429-438.

[9] KASIVISWANATHAN S P, LEE H K, NISSIM K,et al. What can we learn privately? [C]//2008 49th annual IEEE symposium on foundations of computer science. Philadelphia,PA,USA:IEEE,2008:793-826.

[10] ERLINGSSON Ú,PIHUR V,KOROLOVA A. Rappor:randomized aggregatable privacy-preserving ordinal response [C]//Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. USA:Internet Society,2014:1054-1067.

[11] FANTI G,PIHUR V,ERLINGSSON L. Building a RAPPOR with the unknown:privacy-preserving learning of associations and data dictionaries[C]//Proceedings on privacy enhancing technologies. [s. l.]:[s. n.],2016:41-61.

[12] WANG T,LOPUHAA-ZWAKENBERG M,LI Z,et al. Locally differentially private frequency estimation with consistency[C]//Network and distributed system security symposium. USA:Internet Society,2020.

[13] BASSILY R,SMITH A. Local,private,efficient protocols for succinct histograms[C]//Proceedings of the forty-seventh annual ACM symposium on theory of computing. Portland,Oregon,USA:ACM,2015:127-135.

[14] REN X,YU C M,YU W,et al. LoPub:high-dimensional crowdsourced data publication with local differential privacy [J]. IEEE Transactions on Information Forensics & Security,2018,13(9):2151-2166.

[15] KAIROUZ P,OH S,VISWANATH P. Extremal mechanisms for local differential privacy[J]. The Journal of Machine Learning Research,2016,17(1):492-542.

[16] KIM J W,KIM D H,JANG B. Application of local differential privacy to collection of indoor positioning data[J]. IEEE Access,2018,6(99):4276-4286.

[17] LIU C,CHAKRABORTY S,MITTAL P. Deeprotect:enabling inference-based access control on mobile sensing applications[J]. arXiv:1702.06159,2017.

[18] 霍　峥,张　坤,贺　萍,等. 满足本地化差分隐私的众包位置数据采集[J]. 计算机应用,2019,39(3):763-768.

[19] 马方方,刘树波,熊星星,等. 可穿戴设备数值型敏感数据本地差分隐私保护[J]. 计算机应用,2019,39(7):1985-1990.

[20] 涂子璇,刘树波,熊星星,等. 可穿戴设备的数值型流数据差分隐私均值发布[J]. 计算机应用,2020,40(6):1692-1697.

[21] AKTER M,HASHEM T. Computing aggregates over numeric data with personalized local differential privacy[C]// Australasian conference on information security and privacy. Auckland,New Zealand:Springer,2017:249-260.

[22] WU D,WU X,GAO J,et al. A personalized preservation mechanism satisfying local differential privacy in location-based services[C]//International conference on security and privacy in digital economy. Quzhou,China:Springer,2020:161-175.