一种基于随机投影的本地差分隐私高维数值型数据收集算法

孙慧中，杨健宇，程祥，苏森
北京邮电大学网络与交换技术国家重点实验室，北京 100876

摘要
对满足本地差分隐私的高维数值型数据收集问题进行了研究。设计了一种基于随机投影技术的满足本地差分隐私的高维数值型数据收集算法Multi-RPHM，在满足本地差分隐私的条件下，该算法处理维度较高的数据时能够保证所收集的数据的高效用。从理论上证明了该算法满足ε-本地差分隐私的要求。在合成数据集上进行的实验结果验证了该算法的有效性。

关键词
高维数值型数据；隐私保护；本地差分隐私；随机投影

中图分类号：TP309 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020001

A high-dimensional numeric data collection algorithm for local difference privacy based on random projection

SUN Huizhong, YANG Jianyu, CHENG Xiang, SU Sen
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract
The problem of high-dimensional data collection satisfying local differential privacy was studied. A new locally differentially private algorithm called Multi-RPHM was proposed based on the random projection technology, which achieved the high utility of the collected high-dimensional numeric data while satisfying the local differential privacy. The algorithm was formally proved to meet ε-local differential privacy. The effectiveness of the algorithm was comfirmed through experiments on synthetic datasets.

Key words
high-dimensional numeric data, privacy protection, local differential privacy, random projection

1 引言

随着互联网和云计算等信息技术的发展，各种智能设备日益普及，用户的高维数值型数据被许多服务提供商（如谷歌等互联网公司）收集。通过收集用户的高维数值型数据，服务提供商能够分析和挖掘这些数据的价值，以提供更好的用户体验，并增加收益。例如，在推荐系统中，用户的商品评分数据就是一种典型的高维数值型数据，通过收集用户的商品评分数据，服务提供商能够分析商品流行趋势，从而更有效地为用户推荐商品，并且更合理地投放广告，以增加营业额。然而，用户的高维数值型数据中往往包含大量的敏感信息（如兴趣偏好等），如果没有隐私保护，直接对这些数据进行收集可能导致严重的用户隐私泄露问题，进而阻碍商业运营。因此，用户高维数值型数据收集中的隐私问题亟待解决。

隐私保护的数据收集技术为解决数据收集带来的个人隐私泄露问题提供了一种可行的方案。近年来提出的差分隐私（differential privacy，DP）技术[1-3]是目前比较先进的隐私保护技术。与传统的基于匿名的隐私保护技术（例如，k-匿名[4]和L-多样性[5]）不同，差分隐私技术提供了一种严格的、可证明的隐私保护手段，并且其提供的隐私保护强度并不依赖于攻击者掌握的背景知识。本地差分隐私技术（local differential privacy，LDP）[6-7]是一种专门解决数据收集导致个人隐私泄露问题的技术，该技术已被应用于众多现实应用软件之中，如Google公司的Chrome浏览器[8]等。该技术的主要思想是每个用户在将自己的真实数据发给数据收集者之前就对其进行加噪处理。由于用户的真实数据始终存储在用户本地，本地差分隐私技术可以有效地避免不可信收集者的恶意攻击，从根本上为用户提供隐私保护。

当前，本地差分隐私技术已被应用于一维或多维分类型数据收集[8-13]以及多维数值型数据收集[14-15]中。其中，一种可以用于处理这些问题的简单方案是数据收集者直接调用Multi-HM算法[14]。该算法是当前先进的、满足本地差分隐私的多维数据收集算法，该算法的基本思路是每个用户从属性集合中随机选取几个属性，并进行加噪处理，然后将加噪后的属性信息发送给数据收集者。然而，运用该算法收集到的数据的准确性受维度高低（即属性个数大小）影响明显，在处理具有较高维度的用户数据时，会导致收集的数据中包含大量的噪声，因此不适用于用户高维数值型数据的收集。为此，本文提出了一种基于随机投影技术的本地差分隐私数据收集算法——Multi-RPHM算法。在该算法中，首先用户基于随机投影技术对自身原始高维数据进行降维，然后数据收集者对降维后的数据进行收集并进行维度还原。直观上，由于数据收集者只需收集低维数据，因此Multi-RPHM算法能有效降低收集到的数据中包含的噪声，获得较高的数据效用。

2 预备知识与问题定义

2.1 高维数值型数据

用户的高维数值型数据是一种典型的个人数据，由多个数值型属性构成，每个属性反映用户不同方面的信息。特别地，给定一个属性集合$\boldsymbol{A}=\{A_1, A_2, \cdots, A_d\}$，其

中，d表示属性数量，A_j代表第j个属性，并且每个属性的取值均为实数。据此，本文将一个用户的高维数值型数据表示为一个元组$\boldsymbol{t}=[t[A_1],t[A_2],\cdots,t[A_d]]$，其中$t[A_j]$代表元组$\boldsymbol{t}$中第$j$个属性的取值。本文假定所有属性的取值范围均为$[-1,1]$，即$t[A_j]\in[-1,1]$（$1\leqslant j\leqslant d$）。

2.2 本地差分隐私

本地差分隐私[6-7]的定义如下。

ε-本地差分隐私：给定一个隐私参数ε，对于一个随机算法M，当且仅当任意两个输入值v、v'和任意一个可能的输出值$O\in\mathrm{Range}(M)$满足计算式（1），则称算法M满足ε-本地差分隐私。

$$\Pr[M(v)=O]\leqslant \mathrm{e}^{\varepsilon}\times\Pr[M(v')=O] \quad (1)$$

特别地，对于一系列本地差分隐私算法，整体隐私保护强度满足如下串行机制[16]。

串行机制：给定r个本地差分隐私算法$M_i(1\leqslant i\leqslant r)$，其中第$i$个算法$M_i$满足$\varepsilon_i$-本地差分隐私，则算法序列$M_i(v)$满足$\left(\sum_{i=1}^{r}\varepsilon_i\right)$-本地差分隐私。

2.3 问题定义

给定n个用户$u_i(1\leqslant i\leqslant n)$，其中$u_i$代表第$i$个用户。每个用户$u_i$拥有的高维数值型数据用元组$\boldsymbol{t}_i$来表示。本文的目标是设计一个满足本地差分隐私的算法，使一个不可信的数据收集者收集到的用户高维数值型数据集$\{t_i^*|1\leqslant i\leqslant n\}$与用户的原始数据集$\{t_i|1\leqslant i\leqslant n\}$具有相同的统计特征。为了便于分析，本文假定所有用户均采用相同的隐私参数ε。

文中常用的符号及说明见**表1**。

表1 符号列表

符号	说明
$\boldsymbol{A}$	属性集合
d	$\boldsymbol{A}$中的属性个数
A_j	第j个属性
n	用户的数量
u_i	第i个用户
t_i	u_i的元组
$t_i[A_j]$	t_i中A_j的取值
$z[A_j]$	A_j的真实均值
$z^*[A_j]$	A_j的估计均值
x_i	u_i的低维数据元组
$R_{d\times q}$	投影维度为q的投影矩阵

3 Multi-HM算法

目前，可以处理该问题的方法共有3种：Laplace加噪算法[2]、MeanEST算法[15]和Multi-HM算法。在Laplace加噪算法中，每个用户向其原始数据的各个属性维度注入满足Laplace分布的随机噪声，然后将加噪后的数据发送给数据收集者。在MeanEST算法中，首先，用户根据其原始数据产生2个集合，每个集合中包含多个数据元组；然后，用户依照特定概率选择一个集合；最后，用户在该集合中随机选取一个数据元组，并将其作为扰动结果发送给数据收集者。但是，这2种方法均存在一定的缺陷：Laplace加噪算法引入的随机噪声是无界的，即噪声的取值可能无穷大或无穷小，会导致收集的数据噪声较大、效用较差；而对于MeanEST算法，用户返回的扰动元组始终落在原始数据域之外，也会导致收集的数据的效用较差。为了解决上述2种方法存在的缺陷，参考文献[14]中提出了一种新的满足本地差分隐私的多维数值型数据收集算法，即Multi-HM算法。

具体地，对于每个用户u_i，Multi-HM算法（算法1）如下。其输入是用户数据$t_i \in [-1,1]^d$、隐私参数ε，输出是扰动结果$t_i^* \in [-C \cdot d, C \cdot d]^d$。在该算法中，用户$u_i$首先初始化扰动结果$t_i^*$，计算参数$k$、$\varepsilon^*$、$C$和$\alpha$。然后，在$\boldsymbol{A}$中随机选取$k$个属性构成集合$\boldsymbol{S}$；接着，对每个属性$A_j \in \boldsymbol{S}$，用户$u_i$计算$t_i^*[A_j]$。

算法1 Multi-HM算法

输入：用户数据$t_i \in [-1,1]^d$、隐私参数ε。

输出：扰动结果$t_i^* \in [-C \cdot d, C \cdot d]^d$。

初始化$t_i^* = [0,0,\cdots,0]$

令$k = \max\left\{1, \min\left\{d, \frac{\varepsilon}{2.5}\right\}\right\}$，$\varepsilon^* = \varepsilon / k$，$C = \frac{\exp(\varepsilon^*/2)+1}{\exp(\varepsilon^*/2)-1} \cdot d / k$，$\alpha = \begin{cases} 1-\mathrm{e}^{-\varepsilon^*/2}, \varepsilon^* > 0.61 \\ 0, \varepsilon^* \leqslant 0.61 \end{cases}$

在$\boldsymbol{A}$中随机选取k个属性构成属性集合$\boldsymbol{S}$；

for $A_j \in \boldsymbol{S}$ do

在[0,1]内选取随机数f；

if $f > \alpha$ then

$$t_i^*[A_j] = \begin{cases} C \cdot d / k, \Pr = (\mathrm{e}^{\varepsilon^*}-1)/(2\mathrm{e}^{\varepsilon^*}+2) \cdot t_i[A_j] + \frac{1}{2} \\ -C \cdot d / k, \Pr = -(\mathrm{e}^{\varepsilon^*}-1)/(2\mathrm{e}^{\varepsilon^*}+2) \cdot t_i[A_j] + \frac{1}{2} \end{cases}$$

else

令$l(t_i[A_j]) = \left(\frac{C+1}{2} \cdot t_i[A_j] - \frac{C-1}{2}\right) \cdot d / k$，

$r(t_i[A_j]) = (l(t_i[A_j]) + C - 1) \cdot d$；

在[0,1]内选取随机数x；

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

随机选取$t_i^*[A_j] \in [l(t_i[A_j]), r(t_i[A_j])]$；

else

随机选取$t_i^*[A_j] \in [-C, l(t_i[A_j])) \cup (r(t_i[A_j]), C]$；

返回t_i^*

参考文献[4]证明了Multi-HM算法满足ε-本地差分隐私，并且对数据收集者运用该算法收集到的数据集进行了效用分析。然而，根据其分析结果，本文发现运用Multi-HM算法收集到的数据的效用受属性个数d的影响明显。随着d的增大，数据效用逐渐变差。对于较大的d（如d>200），Multi-HM算法会产生较大的误差，难以满足实际应用需求。

4 Multi-RPHM算法

4.1 算法设计

为解决Multi-HM算法在处理较大属性个数d时误差较大的问题，本文提出了Multi-RPHM算法。用户首先通过随机投影对自身原始高维数据进行降维，然后数据收集者收集用户降维后的数据并进行维度还原。随机投影（random projection，RP）[17]是一种有效的降维技术，在投影维度选择合理时，降维后的低维数据能够保留原始数据的特征信息。

Multi-RPHM算法的整体框架如**图1**所示，共包括4个步骤。

- 数据收集者生成一个随机投影矩阵，并将其广播给所有用户。
- 每个用户利用该投影矩阵对其原始高维数据进行降维处理，分别得到对应的低维数据元组。
- 数据收集者利用Multi-HM算法对所有用户降维后的数据进行收集。
- 数据收集者利用投影矩阵的广义逆矩阵对收集到的低维扰动数据进行维度恢复，得到与原始维度相同的高维数据。

Multi-RPHM算法（算法2）如下。其输入为所有用户原始高维数据$\{t_i \in [-1,1]^d | 1 \leqslant j \leqslant n\}$、隐私参数$\varepsilon$、投影维度$q$，输出为高维数据矩阵$\boldsymbol{T}^*$。在该算法中，数据收集者首先生成一个$d \times q$维的随机投影矩阵$\boldsymbol{R}_{d \times q}$，并将其广播给所有用户。具体地，$\boldsymbol{R}_{d \times q}$采用经典的正交矩阵方法构造，即首先生成一个$d \times q$维的随机矩阵，该

矩阵中每个元素均由均值为0、标准差为 $1/\sqrt{k}$ 的高斯分布采样生成，然后将该矩阵进行Gran-Schmidt正交化，使得矩阵的每一列都是标准正交的，最后对矩阵的每一列进行归一化处理。对于每个用户u_i，首先计算q维的向量$\boldsymbol{x}_i = t_i R_{d\times q}$，然后将$\boldsymbol{x}_i$中的所有值映射到[-1,1]。具体地，对于 $\forall s \in [1,2,\cdots,q]$，如果 $x_i[s] < -1$，则令 $x_i[s] = -1$；如果 $x_i[s] > 1$，则令 $x_i[s] = 1$。接着，用户u_i执行Multi-HM算法对$\boldsymbol{x}_i$进行隐私处理，得到扰动后的低维数据元组$\boldsymbol{x}_i^*$，并将其发送给数据收集者。在收集到所有用户的扰动结果集合$\{x_i^* | 1 \leqslant i \leqslant n\}$后，数据收集者构建$n\times q$维的矩阵$\boldsymbol{X}^*$，该矩阵的第$i$行为$\boldsymbol{x}_i^*$。最后，数据收集者通过广义逆矩阵$\boldsymbol{R}^+=(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}\boldsymbol{R}^{\mathrm{T}}$对$\boldsymbol{X}^*$进行维度恢复，得到原始属性维度的数据集$\boldsymbol{T}=\boldsymbol{X}^*\boldsymbol{R}^+$，其中，$\boldsymbol{T}^*$的第$i$行$(\boldsymbol{T}^*)_i=\boldsymbol{t}_i^*$对应用户$u_i$的具有隐私保护的高维数据。

算法2 Multi-RPHM

输入：用户原始高维数据$\{t_i \in [-1,1]^d | 1 \leqslant j \leqslant n\}$、隐私参数$\varepsilon$、投影维度$q$。

输出：估计高维数据矩阵$\boldsymbol{T}^*$。

数据收集者生成 $d\times q$ 维的随机投影矩阵 $R_{d\times q}$，并广播给所有用户；

for 每个用户u_i do

计算 $x_i = t_i R = [x_i[1], x_i[2], \mathrm{L}\ , x_i[q]]$；

for $s \in [1,2,\mathrm{L}\ ,q]$ do

如果 $x_i[s] \notin [-1,1]$，将$x_i[s]$映射到[-1,1]；

end for

将x_i和ε作为参数执行Multi-HM算法，生成扰动结果x_i^*，并发送给数据收集者；

end for

数据收集者根据集合$\{x_i^* | 1 \leqslant i \leqslant n\}$构建矩阵$\boldsymbol{X}^*$；

数据收集者计算矩阵 $T^* = X^* \cdot R$ ；

返回$\boldsymbol{T}^*$；

为了说明Multi-RPHM算法的有效性，本文对其误差进行了讨论分析。Multi-RPHM算法的误差来源主要包括两个方面：一是数据降维与维度恢复；二是低维数据的隐私化处理。当原始数据维度较大时（例如 $d > 200$），Multi-RPHM算法通过降维能够有效地减少隐私化处理引入的误差，并且保证维度恢复产生的误差相对较小，从而降低总体误差。因此，Multi-

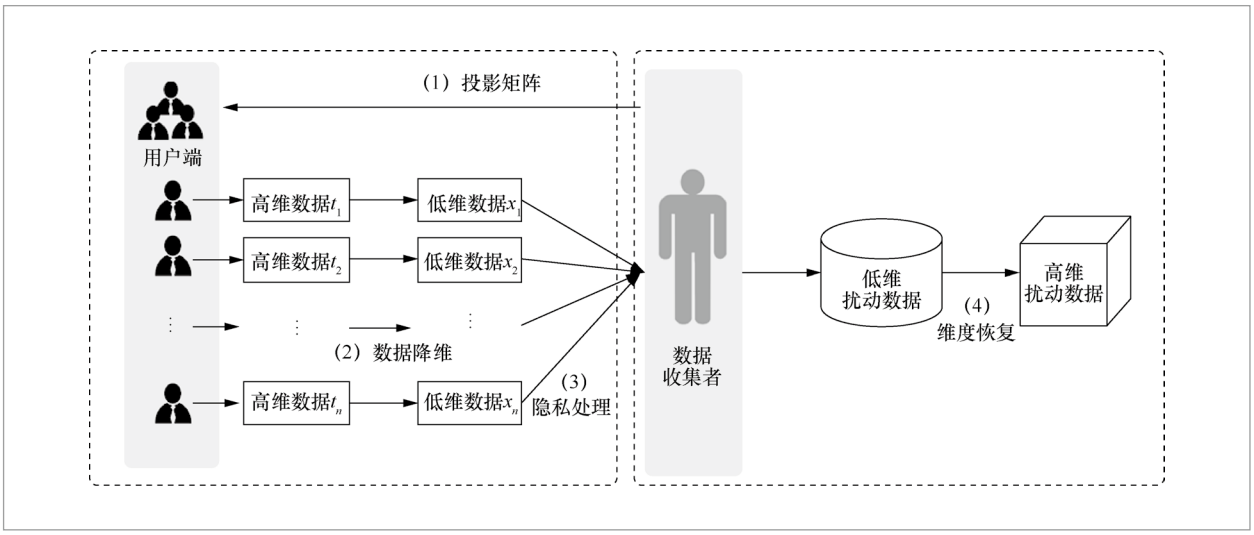

图1 Multi-RPHM 算法的整体框架

RPHM算法能够在保证数据隐私的同时，降低误差，提高数据效用，具有一定的有效性。

4.2 隐私分析

对于任意用户 $u_i\left(i=1,2,\cdots,n\right)$ 及隐私预算ε，Multi-RPHM算法满足ε-本地差分隐私，具体证明过程如下。

对于任意两个不同的高维数据元组 $t_1,t_2\in[-1,1]^d$，用户端的隐私化处理流程如**图2**所示，即用户首先利用投影矩阵$\boldsymbol{R}$将其降维为$\boldsymbol{x}_1,\boldsymbol{x}_2$，然后通过Multi-HM算法对其进行扰动，最后将扰动后的低维数据元组 $\boldsymbol{x}_1^*,\boldsymbol{x}_2^*$ 传送给数据收集者。

由本地差分隐私的定义可知，对于上述高维数据元组 $t_1,t_2\in[-1,1]^d$ 以及数据收集者收集的任意低维扰动数据$\boldsymbol{x}^*$，要证Multi-RPHM算法满足ε-本地差分隐私，即证：

$$\frac{\Pr(\boldsymbol{x}_1^*=\boldsymbol{x}^*|t_1)}{\Pr(\boldsymbol{x}_2^*=\boldsymbol{x}^*|t_2)}\leqslant \mathrm{e}^{\varepsilon} \tag{2}$$

由于Multi-HM算法满足ε-本地差分隐私，即对于任意的$\boldsymbol{x}_1,\boldsymbol{x}_2\in[-1,1]^q$ 以及任意的输出$\boldsymbol{x}^*$，始终有：

$$\frac{\Pr(\boldsymbol{x}_1^*=\boldsymbol{x}^*|\boldsymbol{x}_1)}{\Pr(\boldsymbol{x}_2^*=\boldsymbol{x}^*|\boldsymbol{x}_2)}\leqslant \mathrm{e}^{\varepsilon} \tag{3}$$

另外，由于随机投影矩阵$\boldsymbol{R}_{d\times q}$的生成不依赖于用户的数据，因而$\boldsymbol{R}_{d\times q}$不泄露隐私（$\boldsymbol{R}^+$同理）。因此，由给定的$\boldsymbol{R}_{d\times q}$以及式(2)可得到：

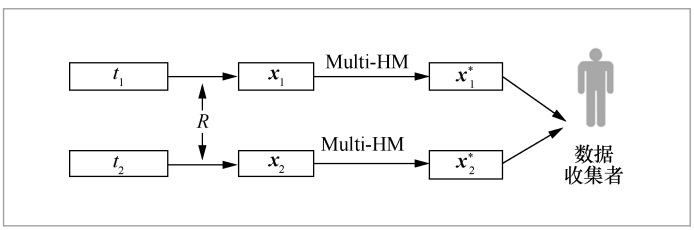

图 2 t_1、t_2 的算法处理流程

$$\frac{\Pr(\boldsymbol{x}_1^*=\boldsymbol{x}^*|t_1)}{\Pr(\boldsymbol{x}_2^*=\boldsymbol{x}^*|t_2)}=\frac{\Pr(\boldsymbol{x}_1^*=\boldsymbol{x}^*|t_1\boldsymbol{R})}{\Pr(\boldsymbol{x}_2^*=\boldsymbol{x}^*|t_2\boldsymbol{R})}$$
$$=\frac{\Pr(\boldsymbol{x}_1^*=\boldsymbol{x}^*|\boldsymbol{x}_1)}{\Pr(\boldsymbol{x}_2^*=\boldsymbol{x}^*|\boldsymbol{x}_2)}\leqslant \mathrm{e}^{\varepsilon} \tag{4}$$

综上所述，$\frac{\Pr(\boldsymbol{x}_1^*=\boldsymbol{x}^*|t_1)}{\Pr(\boldsymbol{x}_2^*=\boldsymbol{x}^*|t_2)}\leqslant \mathrm{e}^{\varepsilon}$ 成立。

因此，Multi-RPHM算法满足$\boldsymbol{\varepsilon}$-本地差分隐私。

5 实验分析

5.1 实验设置

本文在多个合成数据集上对Multi-RPHM算法进行了测试。每个合成数据集包含10 000个用户的高维数据记录，维度（即属性个数）分别为200、300、400、500、600。特别地，根据参考文献[14]，本文设置这些数据集均由均值 $\mu=1/3$ 、标准差$\sigma=1/4$ 的高斯分布采样生成，并且数据取值在$[-1,1]$内。

为了评估Multi-RPHM算法的性能，参照参考文献[4]的评估方法，本文选取均方误差（mean square error，MSE）作为评测指标，对收集到的高维数据集的效用进行评估。具体地，假设原属性均值为$\boldsymbol{Z}=[Z[A_1],\cdots,Z[A_d]]$，估计均值为$\boldsymbol{Z}^*=[Z^*[A_1],\cdots,Z^*[A_d]]$，其中$d$代表属性个数，则$\mathrm{MSE}(\boldsymbol{Z}^*)$的计算式如下：

$$\mathrm{MSE}(\boldsymbol{Z}^*)=$$
$$E[(\boldsymbol{Z}^*-\boldsymbol{Z})^2]=\frac{1}{d}\sum_{j=1}^{d}(Z^*[A_j]-Z[A_j])^2 \tag{5}$$

由于Multi-HM算法是目前比较先进的满足本地差分隐私的多维数据收集算法，因此，为了验证Multi-RPHM算法的有效性，本文选取Multi-HM算法作为实验中的对比算法。

在实验中，本文设置隐私参数$\varepsilon \in \{0.6, 0.8, 1.0, 1.2, 1.4\}$。对于Multi-RPHM算法，本文设置参数$q = 0.3d$。本文所有实验均在内存为8 GB、处理器为Intel Core i5 2.9 GHz的计算机上进行。实验结果均为各种算法运行10次的平均结果。

5.2 结果分析

首先，本文固定隐私参数ε=1.0，投影维度q=0.3d，用户数n=10 000，在属性维度不同的合成数据集上对Multi-HM算法和Multi-RPHM算法进行测试，实验结果如**图3**所示。可以看出，随着属性维度d增加，Multi-HM算法的效果明显变差，而Multi-RPHM算法受维度影响不大，算法性能稳定。这是因为Multi-HM算法受维度d影响较大，而Multi-RPHM算法通过随机投影技术将高维数据降至低维，且仅收集低维数据，降低了隐私化处理引入的扰动误差。当维度d较大时（如400、500、600），Multi-RPHM算法的优势变得更加明显，这说明其具有良好的可扩展性，更适用于高维数据的收集。

其次，为了测试隐私保护强度对算法准确性的影响，本文固定属性维度d=400，投影维度q=0.3d，以评估Multi-HM算法和Multi-RPHM算法在不同隐私参数下的性能，实验结果如**图4**所示。值得注意的是，**图4**中的RP代表只进行数据降维及维度恢复操作，不添加隐私保护时的均值误差。可以看出，随着ε变大，两种算法的MSE均逐渐减小，Multi-RPHM算法的误差逐渐趋近RP，且始终低于Multi-HM算法。特别地，当隐私参数ε较小时（如0.6、0.8、1.0），Multi-RPHM算法明显优于对比算法Multi-HM。正如第4.1节中分析的，这是因为Multi-RPHM算法的总体误差由两个因素引起，一个是数据的降维与维度恢复，另一个是低维数据的隐私化处理。当ε较小时，用户采用Multi-HM算法直接对原始的高维数据隐私化处理会引入大量噪声，而Multi-RPHM算法通过降维有效地降低了这部分误差的引入，从而总体误差较小，优势更加明显。

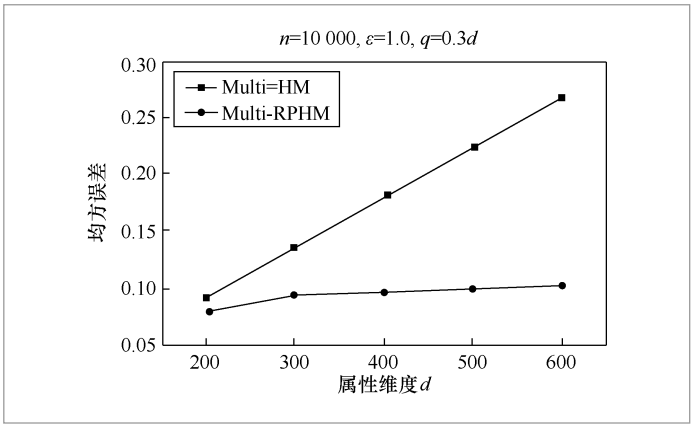

图 3 属性维度改变时 Multi-RPHM 算法与 Multi-HM 算法对比

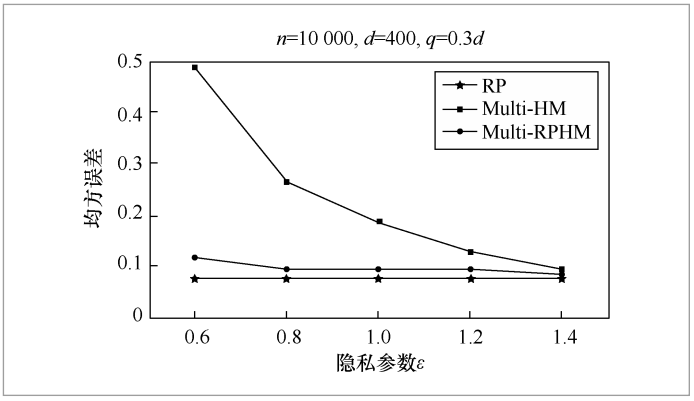

图 4 不同隐私保护强度下 Multi-RPHM 算法与 Multi-HM 算法对比

6 结束语

针对满足本地差分隐私的高维数值型数据收集问题，本文提出了一种基于随机投影的隐私数据收集算法，即Multi-RPHM算法。本文在理论上证明了Multi-RPHM算法满足ε-本地差分隐私。同时，

实验结果验证了Multi-RPHM算法的有效性。本文提出的Multi-RPHM算法适用于多种真实场景，具有较强的实际应用价值。此外，需要指出的是，高维数据一般包括数值型、分类型、混合型3种类型，本文主要聚焦在高维数值型数据收集的问题上，而高维分类型和混合型数据的收集具有新的问题场景和挑战，解决这两个问题具有很高的研究价值与现实意义，能够丰富高维数据收集问题的理论体系。本文提出的基于数据降维的解决思路，能够为解决上述问题提供一定的借鉴意义。

参考文献:

[1] DWORK C. Differential privacy[C]// ICALP, July 10-14, 2006, Venice, Italy. Heidelberg: Springer, 2006: 1-12.

[2] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis [C]//TCC, March 4-7, 2006, New York, USA. Heidelberg: Springer, 2006: 265-284.

[3] DWORK C. Differential privacy: a survey of results[C]//TAMC, April 25-29, 2008, Xi'an, China. Heidelberg: Springer, 2008: 1-19.

[4] SWEENY L. K-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 2012, 10(5): 557-570.

[5] MACHANAVAJJHALA A, KIFERD, GEHRKE J, et al. L-diversity: privacy beyond K-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 24.

[6] KASIVISWANATHAN P S, LEEHK, NISSIM K, et al. What can we learn privately[C]//FOCS, October 25-28, 2008, Philadelphia, USA. Piscataway: IEEE Press, 2008: 531-540.

[7] 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.

[8] ERLINGSSON Ú, PIHUR V, KOROLOVA A. RAPPOR: randomized aggregatable privacy-preserving ordinal response [C]//CCS, November 3-7, 2014, Scottsdale, USA. New York: ACM Press, 2014: 1054-1067.

[9] KAIROUZ P, BONAWITZ K, RAMAGE D. Discrete distribution estimation under local privacy[J]. Proceedings on Privacy Enhancing Technologies, 2016: 84-104.

[10] BASSILY R, SMITH A. Local, private, efficient protocols for succinct histograms[C]// STOC, June 14-17, 2015, Portland, USA. New York: ACM Press, 2015: 127-135.

[11] WARNER S L. Randomized response: a survey technique for eliminating evasive answer bias[J]. Journal of the American Statistical Association, 1965, 60(309): 63-69.

[12] NGUYÊN T T, XIAO X K, YANG Y, et al. Collecting and analyzing data from smart device users with local differential privacy[J]. Computer Science, 2016, arXiv: 1606.05053.

[13] QIN Z, YANG Y, YU T, et al. Heavy hitter estimation over set-valued data with local differential privacy[C]//CCS, October 24-28, 2016, Vienna, Austria. New York: ACM Press, 2016: 192-203.

[14] WANG N, XIAO X K, YANG Y, et al. Collecting and analyzing multidimensional data with local differential privacy [C]// ICDE, April 8-11, 2019, Macao, China. Piscataway: IEEE Press, 2019: 638-649.

[15] DUCHI J C, WAINWRIGHT M J, JORDAN M I, et al. Minimax optimal procedures for locally private estimation[J]. American Statistical Association, 2018, 113(521): 182-201.
[16] MCSHERRY F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[C]//SIGMOD, June 29-July 2, 2009, Providence, USA. New York: ACM Press, 2009: 19-30.
[17] JOHNSON W B, LINDENSTRAUSS J. Extensions of Lipschitz mappings into a Hilbert space[J]. Contemporary Mathematics, 1984, 26(189-206): 1.

作者简介

孙慧中（1998- ），女，北京邮电大学网络与交换技术国家重点实验室硕士生，主要研究方向为隐私保护和机器学习。

杨健宇（1994- ），男，北京邮电大学网络与交换技术国家重点实验室博士生，主要研究方向为隐私保护。

程祥（1984- ），男，北京邮电大学副教授、博士生导师，主要研究方向为数据挖掘、知识工程、隐私保护等。其研究成果已发表在包括IEEE ICDE、IEEE ICDM、AAAI、IJCAI、EMNLP、IEEE TKDE、IEEE TDSC、*IEEE TSC*等在内的国际会议和期刊上。主持与大数据隐私保护相关的国家自然科学基金青年科学基金项目1项、国家自然科学基金项目面上项目1项，并作为科研骨干参与多项国家级和部级科研项目。

苏森（1971- ），男，北京邮电大学教授，计算机学院执行院长，中国计算机学会理事，服务计算专业委员会秘书长，“数字中国产业发展联盟”副理事长。2005年入选教育部“新世纪优秀人才支持计划”，2017年入选国家“万人计划”科技创新领军人才。目前主要研究方向为智能数据服务、数据隐私保护、社交网络分析。获国家科技进步奖二等奖1次，中国通信学会科技进步奖一等奖1次，教育部科技进步奖二等奖1次。

收稿日期： 2019-07-31
基金项目： 国家自然科学基金资助项目（No.61872045）
Foundation Item: The National Natural Science Foundation of China (No.61872045)