软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(3):1093−1110 [doi: 10.13328/j.cnki.jos.006450]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

混洗差分隐私下的多维类别数据的收集与分析*

刘艺菲[1], 王 宁[1], 王志刚[1], 谷 峪[3], 魏志强[1], 张啸剑[2], 于 戈[3]

[1](中国海洋大学 信息科学与工程学部, 山东 青岛 266100)

[2](河南财经政法大学 计算机与信息工程学院, 河南 郑州 450046)

[3](东北大学 计算机科学与工程学院, 辽宁 沈阳 110819)

通信作者: 王宁, E-mail:wangning8687@ouc.edu.cn

摘 要: 随着大数据时代的到来, 如何在保护用户隐私的前提下完成多维类别数据上的频率分布估计问题成为研究热点. 已有的工作主要是基于中心化差分隐私模型或本地化差分隐私模型完成安全算法的设计. 鉴于上述两种模型在隐私保护程度或发布结果可用性方面的弊端, 基于新兴的混洗差分隐私模型, 设计用户数据收集策略, 进而提供高安全、高可用的频率分布估计服务. 考虑到多维类别属性的多维特征以及不同属性上取值域大小不等的异构特点, 从扰动算法以及洗牌方式等角度出发, 设计了基于单洗牌者以及多洗牌者的数据发布方案 ARR-SS 和 SRR-MS. 此外, 结合上述两种方案的优势, 通过填补技术消除属性间异构问题, 提出了基于取值域填补的单洗牌者数据发布方案 PSRR-SS. 从理论上分析了 3 种策略的隐私保护程度以及误差级别, 并利用 4 个真实数据集验证所提出方案在频率估计问题上的有效性. 此外, 将所提方案作为带噪数据库生成技术的加噪组件, 评估随机梯度下降算法在生成带噪数据上的训练结果的可用性. 实验结果展现了所提方案优于当前同类算法.

关键词: 混洗差分隐私; 隐私保护; 多维类别数据; 频率估计

中图法分类号: TP311

中文引用格式: 刘艺菲, 王宁, 王志刚, 谷峪, 魏志强, 张啸剑, 于戈. 混洗差分隐私下的多维类别数据的收集与分析. 软件学报, 2022, 33(3): 1093−1110. http://www.jos.org.cn/1000-9825/6450.htm

英文引用格式: Liu YF, Wang N, Wang ZG, Gu Y, Wei ZQ, Zhang XJ, Yu G. Collecting and Analyzing Multidimensional Categorical Data Under Shuffled Differential Privacy. Ruan Jian Xue Bao/Journal of Software, 2022, 33(3): 1093−1110 (in Chinese). http://www.jos.org.cn/1000-9825/6450.htm

Collecting and Analyzing Multidimensional Categorical Data Under Shuffled Differential Privacy

LIU Yi-Fei[1], WANG Ning[1], WANG Zhi-Gang[1], GU Yu[3], WEI Zhi-Qiang[1], ZHANG Xiao-Jian[2], YU Ge[3]

[1](Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China)

[2](College of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou 450046, China)

[3](School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract: The big era is coming with the ever-growing demands on frequency estimation based on sensitive multi-dimensional categorical data. The existing works are devoted to designing privacy protection algorithms based on centralized differential privacy or local differential privacy. However, the above models provide either the weak level of privacy protection or low accuracy of published results. Therefore, standing on the emerging shuffled differential privacy which remedies the above modes, the data collection mechanisms are designed, providing frequency distribution estimation service with high security and high availability. Considering the multi-dimensional

* 基金项目: 国家自然科学基金(61902365, 61902366, 62072156); 中央高校基本科研业务费(202042008); 中国博士后基金(2019M652473, 2019M652474, 2020T130623); 青岛市自主创新重点研发(20-3-2-12-xx); 青岛市博士后应用项目

本文由“数据库系统新型技术”专题特约编辑李国良教授、于戈教授、杨俊教授和范举教授推荐.

收稿时间: 2021-06-30; 修改时间: 2021-07-31; 采用时间: 2021-09-13; jos 在线出版时间: 2021-10-21

characteristics of data and the heterogeneous characteristics existed in different attributes, the mechanisms including SRR-MS with multiple shufflers and ARR-SS with one shuffler are firstly proposed. And then in order to combine the advantages of the above two mechanisms, PSRR-SS with one single shuffler, is proposed to eliminate the heterogeneity among attributes by means of padding dummy values technology to the attribute domains. This study detailedly analyzes the degree of privacy protection and the error level of three strategies theoretically, and evaluates the performance of the proposed mechanisms on frequency estimation by using four real datasets. Besides, the proposals are used as the perturbing component of the techniques generating synthetic data and the training results of stochastic gradient descent are evaluated based on synthetic data. The experimental results show that the proposed method outperforms the existing algorithms.

Key words: shuffled differential privacy; privacy protection; multidimensional categorical data; frequency estimation

大数据时代, 各企业机构都在充分地运用“数据红利”, 然而随之而来的是频频见诸报端的隐私泄露事件. 随着用户隐私保护意识逐步增强以及数据安全立法日趋完善, 获取大规模个体用户的敏感数据已不现实, 这为数据库系统的构建提出了极大的挑战. 因此, 如何在保证个体敏感信息不被泄露的前提下完成用户与数据收集者之间的信息共享, 成为当前新型数据库系统技术研究的重要分支. 另一方面, 类别数据(categorical data)上的频率估计, 即直方图发布是在隐私保护领域受到广泛关注的经典基础问题. 其可作为其他复杂学习挖掘技术的核心组件, 例如频繁模式挖掘中的项集频率的估计、随机梯度下降技术梯度向量的离散化后的分布估计等. 可见, 如何在满足隐私保护的前提下从个体用户收集数据, 获取高可用的频率分布, 是亟待解决的问题.

传统的差分隐私模型包括中心化差分隐私模型(centralized differential privacy, CDP)[1−3]以及本地化差分隐私模型(local differential privacy, LDP)[4−7], 其中,

- CDP 要求一个完全可信的第三方对数据统计结果加噪, 而现实中第三方很难达到完全可信, 因此 CDP 可提供的安全保证程度较弱.
- LDP 不存在任何可信的第三方, 需要将本身数据扰动后再完成与其他方数据的共享. 由于该模型下每一个用户都会对原始数据加噪, 导致统计结果可用性差.

鉴于上述两种模型的缺点, 混洗差分隐私模型(shuffled differential privacy, SDP)[8−12]应运而生. 该模型在用户与数据收集者之间引入洗牌者 Shuffler, 用户将加噪的数据传给 Shuffler, Shuffler 负责对所有用户的加噪数据进行混洗, 并将洗牌后的结果发给收集者分析. 洗牌操作切断了用户和数据之间的关联关系, 带来了隐私收益. SDP 模型兼顾了 CDP 下发布结果高准确性和 LDP 下隐私保护水平高的两大优点, 因此, 本研究基于 SDP 模型, 设计多维类别数据的频率估计技术.

对于多维类别属性而言, 其最重要的特征是多维性, 从单个用户收集每个维度的数据会造成收集信息量过大, 隐私预算损耗更多, 影响发布结果; 此外, 用户发布所有维度的数据也会增加通信开销. 另一方面, 多维类别属性每一维度取值域大小不同, 造成不同属性间存在异构特征, 而这种异构性特征会直接导致每一个维度数据所调用的 LDP 扰动机制的参数不同. 在 SDP 模型下, 在分析洗牌操作所带来的隐私收益过程中, 需获取 m 个用户发布维度 d_i 上取值为 c_i 的概率以及 $m-1$ 个用户发布维度 d_i' 上取值为 c_i' 的概率, 并讨论两种概率的比值. 明显地: 若 d_i 的取值域大小不同于 d_i', 将使 LDP 扰动机制的参数不同, 为两种概率比值的讨论带来挑战, 进而给应用 SDP 模型解决多维类别属性的频率估计问题带来困难. SDP 模型下的研究工作还处于探索阶段, 大部分工作是针对单维数据的发布策略进行讨论. 目前为止, 还没有一个满足混洗差分隐私解决多维类别属性发布问题的有效方法. 另一方面, LDP 模型下的众多工作都针对多维数据进行设计, 并验证采样技术即每个用户采一个维度数据并利用满足 LDP 约束的发布机制(以随机应答机制 Randomized Response 为代表)发布相应值的方式对于提高发布结果精度有显著效果[5,13−15]. 由于 SDP 模型下从用户收集数据的组件仍需满足 LDP 约束, 所以本研究延续 LDP 下提高发布结果精度的技术路线, 即采样技术, 完成用户数据的发布. 然而多维类别属性存在异构特征, 即每个属性取值域可能不同, 在利用 Randomized Response(RR)机制扰动真实数据时, 真实值被发布的概率不同, 即需使用不同参数化 RR 机制, 这为 SDP 模型中洗牌操作引入的隐私收益的分析带来极大的挑战.

为应对上述挑战, 本文从同参数化 RR 机制角度出发, 即不同维度上调用 RR 机制发布真实值的概率相同, 并从洗牌方式及加噪扰动方式入手, 设计 SDP 下多维度类别数据的频率估计方法, 具体贡献点如下:

(1) 提出 SDP 模型下, 多维类别数据频率分布估计框架.

(2) 设计基于多洗牌者的单维扰动发布算法(SRR-MS). 设置多个洗牌者, 每个洗牌者负责对同参数化 RR 扰动结果的洗牌置乱, 即仅负责发布采样到同一维度用户的洗牌. 该方案本质上是对用户按维度分组, 用户调用 RR 机制扰动相对应的维度发布, 相同组用户通过同一洗牌者洗牌, 便于洗牌操作的隐私收益分析.

(3) 设计基于单洗牌者的多维扰动发布算法(ARR-SS). 不同于 SRR-MS 方案, 该方案设置单个洗牌者, 负责对所有用户的带噪数据进行洗牌. 为了使参与洗牌数据满足经同参数化 RR 处理的要求, 该方案将 RR 输出域由单个类别属性上的值域扩展为全部属性值域的集合.

(4) 为了满足同参数化要求, SRR-MS 导致洗牌数量减少, 间接导致隐私收益降低, ARR-SS 导致 RR 扰动域增大, 收集精度降低. 本文提出了基于取值域填补的单洗牌者发布算法 PSRR-SS, 充分利用上述两方案优势, 克服其不足. 其利用虚拟取值将每个属性取值域填补至最大值, 利用 RR 机制在新的取值域上对真实数据加噪, 单个洗牌者对所有用户的带噪数据洗牌. PSRR-SS 是 SRR-MS 与 ARR-SS 之间的平衡方案.

(5) 理论分析了 3 种方法满足(ε_c,δ)-混洗差分隐私以及频率估计的误差边界, 并讨论 3 种方法的通信开销以及算法的时间复杂度. 通过真实数据实验分析, PSRR-SS 方法具有较好的可用性.

1 相关工作

在 CDP 模型中, 指数机制被广泛应用于在保证用户隐私的前提下进行类别数据的频率估计[16]. 指数机制的思想是: 对查询以一定的概率输出结果, 以此实现差分隐私, 这个概率值由打分函数确定. LDP 模型下估计类别数据频率的技术相对成熟, 谷歌 Chrome 浏览器使用的 RAPPOR[4]方法实现用户浏览数据的隐私保护. 文献[6]讨论了传统 RR 机制在基于 LDP 模型的单维类别数据上频率估计问题的理论误差, 并提出了 OUE 及 OLH 方案. 其中, RR 方法在属性取值域较低时性能较好; OUE 和 OLH 分别利用一元编码和本地哈希提高精度, 适用于取值域较大的情况[6].

此外, 存在大量工作基于 LDP 模型完成多维数据的发布[5,13−15,17−21]. 文献[13,14,17]针对单个用户具有多个项数据的场景, 设计了满足 LDP 的项频数估计方案. 文献[15]设计了如何从用户的多个键值数据中收集信息, 进而完成键上的频数以及值上的均值估计. 文献[5]结合 LDP 模型研究了多维属性的发布问题, 应用 k 长度向量的思想, 将输入转换为带噪的向量发给收集者. Wang 等人[18]提出了分段机制(PM)和混合机制(HM), 能够处理同时包含数值属性和分类属性的多维数据. Yang 等人[19]提出了 HDG 技术, 通过对单维或者任意两维的数据信息的统计, 完成多维数据上的范围查询. Calm[20]和 FT[21]设计 LDP 发布方案用户完成多维数据上 Marginal 信息的统计. 上述两种方案借助发布某些低敏感的中间信息完成, FT 发布 Marginal 对应的 Hadamard Matrix 中的系数, Calm 发布指定数量的短 Marginal. 上述多维工作中, 一直采用采样技术解决敏感度大、发布结果精度低的问题. SDP 下暂无多维数据发布的工作. 鉴于采样技术的有效性, 本文同样采用采样技术应对多维数据带来的挑战. 但是, 由于在 SDP 模型中需要对采样到的不同维度的数据进行统一的洗牌处理, 并分析洗牌操作的隐私收益, 不同维度的异构特征为 SDP 下采样技术的应用分析带来新挑战.

SDP 模型是由 Bittau 等人[22]提出的, 他们设计了 ESA 框架, 包括编码(encoder)、洗牌(shuffler)和分析(analyzer)这 3 部分. ESA 通过洗牌操作打乱了用户和带噪数据之间的关联关系, 以此提高了隐私保护程度. 文献[9]讨论了洗牌操作带来的隐私收益水平. 其中, Balle 等人[9]证明了使用隐私预算为ε的 RR 机制时, 隐私保护程度可提高至 $\sqrt{14\ln(2/\delta)(\mathrm{e}^{\varepsilon}+k-1)/(n-1)}$, 其中, k 为 RR 机制的扰动域大小.

上述工作都是在理论上分析洗牌带来的隐私收益, 进行实际验证的研究相对较少. 文献[8,12]设计了多信息通信机制, 让每个用户发送多条消息. 文献[11]提出了结合 SDP 模型、利用 OUE 机制发布多维数据的方法,

并理论分析了在 SDP 模型下使用 OUE 机制的隐私收益. 文献[12]提出了 pureDUMP 和 mixDump 方法, 让用户除了发布扰动值之外, 还发送一些没有意义的随机值. 这些方法都通过增加洗牌量放大隐私收益, 但引入了较大的通信开销. 多维类别属性由于维度大、取值域异构, 给应用 SDP 模型带来困难, 目前还未出现一种基于 SDP 模型估计多维类别数据分布的有效方法. 基于上述分析, 本文提出了一种基于混洗差分隐私模型的多维类别数据频率发布算法 PSRR-SS. 该方法通过填补不同维度取值域, 结合采样机制与随机响应机制, 完成高精度的频率发布.

2 基础知识与问题描述

2.1 中心化差分隐私

定义 1((*ε*,*δ*)-中心化差分隐私). 给定邻居数据集 D 和 D', 扰动算法 S, 输出 y'. D 和 D'中仅一条记录不同. 如果 S 作用在数据集上得到任意 y'的概率满足下列不等式, 那么 S 满足(ε,δ)-中心化差分隐私:

$$\Pr[S(D)\in y']\leqslant e^{\varepsilon}\times\Pr[S(D')\in y']+\delta,$$

其中, ε为隐私预算, 它衡量隐私保护程度; $\delta(\delta\in(0,1])$为隐私泄露概率.

引理 1(并行化特征[23]). 给定一个由$\{S_1,\ldots,S_i,\ldots,S_m\}$组成的扰动算法 S, 如果 $S_i(1\leqslant i\leqslant m)$应用于输入数据集一个不相交的子集并且满足$\varepsilon_i$-差分隐私, 那么 S 满足ε-差分隐私, 其中, $\varepsilon=\max_i\varepsilon_i$.

2.2 本地化差分隐私

定义 2((*ε*,*δ*)-本地化差分隐私). v 和 v'为任意两个用户的真实数据, 给定扰动算法 S, 输出 y'. 如果 S 作用在 v 和 v'上得到任意 y'的概率满足下列不等式, 那么 S 满足(ε,δ)-本地化差分隐私:

$$\Pr[S(v)\in y']\leqslant e^{\varepsilon}\times\Pr[S(v')\in y']+\delta.$$

RR 机制是 LDP 下用户在本地对自身数据扰动的一个常用机制, 其基本思想是: 让用户以 p 的概率发送真实值, 以 q 的概率随机选取值域内另一个值. 给定值域 $D=\{1,2,\ldots,d\}$, RR 机制满足:

$$\Pr[RR(v)=y]=\begin{cases}p=\dfrac{e^{\varepsilon}}{e^{\varepsilon}+d-1}, & v=y\\ q=\dfrac{1}{e^{\varepsilon}+d-1}, & v\neq y\end{cases}\tag{1}$$

其中, v 表示用户的原始数据, y 表示值域 D 内某个不确定的取值.

2.3 混洗差分隐私

用户 u_i 的数据 $t_i\in V$. 每个用户 u_i 在本地使用满足ε_l 的本地化差分隐私算法 $R:V\to Y$ 扰动 $v_i:y_i=R(v_i)$. 令 $O=\{y_1,y_2,\ldots,y_n\}$为 n 个用户的扰动结果, 其中, y_i 表示第 i 个用户的扰动结果. $S:Y^n\to Y^n$ 为洗牌者对 n 个用户的输出结果进行随机洗牌操作. 设机制 $M=R\circ S$ 为上述两个步骤的结合, M 满足(ε_c,δ)-混洗差分隐私, 当且仅当对于任意邻居数据集 D 和 D'(n 个用户中仅有一个用户数据不同), 任意输出集合 $y'\subset Y^n$, 存在:

$$\Pr[M(D)\in y']\leqslant e^{\varepsilon_c}\times\Pr[M(D')\in y']+\delta.$$

在混洗差分隐私模型中, 用户对查询给出的随机答复被称作隐私毯子. 基于 LDP 模型产生的分布可以分解为两部分, 一部分依赖于真实值的分布, 另一部分是独立随机的分布(隐私毯子), 此过程被称为隐私毯子分解[9]. 公式(1)中给出的 RR 的输出分布可以被分解为

$$\Pr[RR(v)=y]=(1-\gamma)\Pr^{v}[y]+\gamma\Pr[Uni(D)=y]\tag{2}$$

其中, $\Pr^{v}[y]$表示依赖于 v 的真实值形成的分布, $Uni(D)$是均匀随机分布, 且 $\Pr[Uni(D)=y]=1/d$. 应用 RR 机制时, 通过令 $\Pr^{v}[y]=l_{\{v=y\}}$让γ最大化$\left(\gamma=\dfrac{d}{e^{\varepsilon_l}+d-1}\right)$, $\Pr^{v}[y]=l_{\{v=y\}}$, 只有在发布值与真实值相同时, 该概率值为 1, 其余为 0. 输出以 $1-\gamma$的概率依赖于真实值, 以γ的概率随机发布. 给定 n 个用户, 除第 n 个用户外, 其余 $n-1$ 个用户的输出可以看作包含一些均匀噪音, 这些噪音使输出具有不确定性. 对于 $v\in D$, 噪音服从 $Bin(n-1,\gamma/d)$,

即服从 $Bin\left(n-1,\frac{1}{e^{\varepsilon_l}+d-1}\right)$. 文献[10]指出: 使用满足$\varepsilon_l$-LDP 的 RR 算法扰动本地数据时, 洗牌后结果满足(ε_c,δ)-SDP, 其中, $\varepsilon_c=\sqrt{14\ln(2/\delta)\cdot\frac{e^{\varepsilon_l}+d-1}{n-1}}$.

2.4 问题描述

给定 n 个用户 $U=\{u_1,u_2,\ldots,u_n\}$, 多维类别属性 $D=\{A_1,A_2,\ldots,A_d\}$, 每个类别属性的取值域大小分别对应$\{k_1,k_2,\ldots,k_d\}$. 每个用户 u_i 拥有一个记录 $v_i=\{v_{i1},v_{i2},\ldots,v_{id}\}$, 对应 d 个属性的取值. 用户在本地使用扰动机制对t_i 扰动, 得到 $\tilde{t}_i$. 混洗方 S 对所有扰动数据洗牌, 即 $S(\tilde{v}_1,\tilde{v}_2,...,\tilde{v}_n)$. 收集者基于洗牌后的数据进行频率估计. 采用误差平方和 SSE(sum square error)度量发布结果精度. 本文要设计一个满足(ε_c,δ)-混洗差分隐私的多维类别数据频率估计方法, 在避免高隐私消耗的同时, 尽可能地使误差较小. 误差平方和表达式为

$$SSE(F,\tilde{F})=\sum_{j=1}^{d}\sum_{v=1}^{k_j}(f_v,\tilde{f}_v)^2,$$

其中, F 和 $\tilde{F}$ 分别表示原始类别数据频率分布和估计的类别数据频率分布, f_v 和 $\tilde{f}_v$ 分别表示某一个类别数据上某一种取值的真实频率和估计频率.

3 基本方法

多维类别数据维度高和取值域异构等特点增加了发布过程的难度, 本节提出了多维类别数据频率估计的两个基本方法, 分别从不同维度数据的洗牌方式和采样数据的加噪扰动方式两个角度来消除异构性. 本节介绍了这两个基本方法的过程, 然后进行了隐私性和可用性分析.

3.1 基于多洗牌者的单维扰动发布算法(SRR-MS)

LDP 模型下, 已经有许多工作[13,14]证明划分隐私预算不如对用户分组效果好. 本文借鉴了这一经验, 提出了 SRR-MS 算法. SRR-MS 设置多个洗牌者, 每个洗牌者负责来自同一扰动域的带噪数据的洗牌. 具体来说, 将所有数据打乱后, 按属性个数将用户分组, 每个用户只发布所在组对应属性的值. 每一组用户利用ε_c 求得在本地实际使用的预算ε_l 后, 均使用 RR 机制扰动相应属性的值, 扰动范围是相应类别属性的取值域, 并将扰动值传给相应 Shuffler 洗牌. 以图 1 为例, 将用户按属性个数随机分为 4 组, 第 i 组用户发布第 i 个属性值, 使用 RR 机制在第 i 个属性取值范围内扰动, 即以 $p_i=e^{\varepsilon_{li}}/(e^{\varepsilon_{li}}+k_i-1)$ 的概率发布第 i 个属性取值域的真实值, 以 $q_i=1/(e^{\varepsilon_{li}}+k_i-1)$ 的概率发布该取值域上的非真实值, 并将扰动之后的值发给 Shuffler S_i 扰动. 其余各组用户发布方式与此类似. 所有洗牌者将数据发送给收集方统计频率分布 $\tilde{f}_v=\frac{d\sum_{j\in[n/d]}l_{\{y_j=v\}}-nq}{n(p-q)}$, 当 $y_j=v$ 成立时, $l_{\{y_j=v\}}$ 值为 1; 否则为 0. SRR-MS 方法的具体过程如算法 1 所示.

算法 1. SRR-MS.

输入: n 个用户的数据(第 i 个用户数据为 $v_i=(v_{i1},v_{i2},\ldots,v_{id})$), 属性维度 d 及相应的取值域大小$\{k_1,k_2,\ldots,k_d\}$, 隐私预算ε_c, δ.

输出: 多维类别数据的频率估计分布 $\tilde{F}$.

1. 对 n 个用户的数据洗牌并划分为 d 组;
2. **for** groups i=1 to d **do**
3. $\varepsilon_{li}=Calculate(\varepsilon_c,\delta,i)$; //为每组计算放大后的隐私预算$\varepsilon_{li}$
4. **for** user j in group i **do**
5. $y_j=RR(v_{ji},k_i,\varepsilon_{li})$; //在第 i 个属性的取值域内扰动
6. User j send y_j to Shuffler S_i; //第 j 个用户把带噪数据 y_j 发送给第 i 个洗牌器

7. **end for**
8. Shuffler S_i 打乱所有带噪数据，并将其发送给收集方；
9. **end for**
10. 收集方根据带噪数据计算频率、进行无偏校正并发布带噪分布 $\tilde{F}$；
11. **return** $\tilde{F}$.

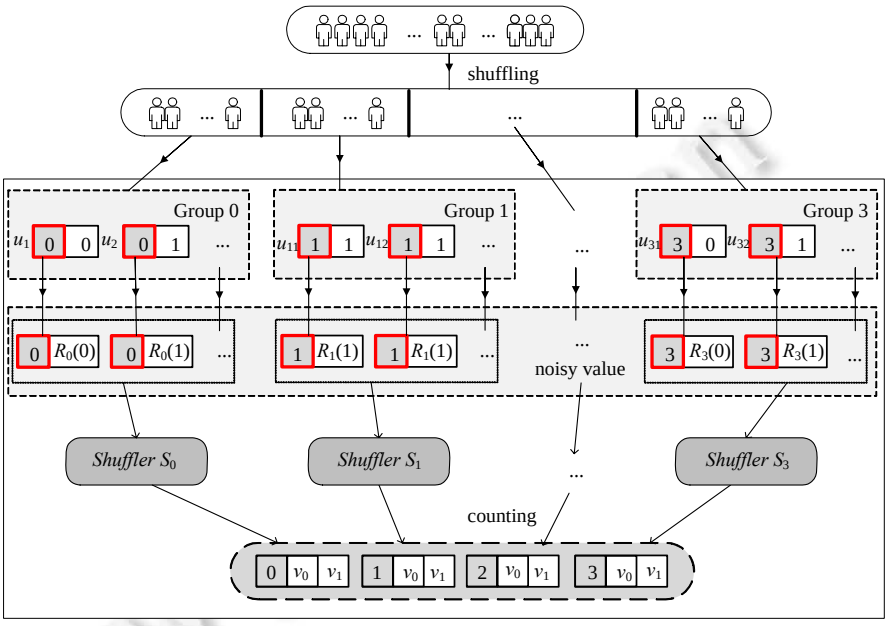


图 1 应用 SRR-MS 算法的多维类别数据发布框架

3.1.1 SRR-MS 算法的隐私效用分析

SRR-MS 算法将用户按照类别属性个数分组扰动并使用对应的 Shuffler 洗牌，根据差分隐私的并行化特征可知：由于不同组的用户之间不存在交叠，每组确保的隐私级别与全部用户确保的隐私级别相同. 即若整个算法满足(ε_c,δ)-SDP，则在本方案用户分割方式下，每组用户满足(ε_c,δ)-SDP.

定理 1. SRR-MS 算法满足(ε_c,δ)-混洗差分隐私，其中，$\varepsilon_c \leqslant \sqrt{\dfrac{14\ln(2/\delta)(\mathrm{e}^{\varepsilon_l}+k_i-1)}{n/d-1}}$.

证明：根据文献[9], n 个用户使用 RR 机制在大小为 d 的取值域上扰动时，满足 $\varepsilon_c=\sqrt{14\ln(2/\delta)\cdot\dfrac{\mathrm{e}^{\varepsilon_l}+d-1}{n-1}}$. 此处，第 i 组参与洗牌的用户数量为 n/d，在第 i 个类别属性的取值域 k_i 上扰动，所有用户将扰动后的数据发送给相应的 Shuffler 洗牌，因此定理 1 成立. □

3.1.2 SRR-MS 算法的可用性分析

定理 2. 假设 f_v 和 $\tilde{f}_v$ 分别表示 v 的真实频率和估计频率，则 $E[\tilde{f}_v]=f_v$ 成立.

证明：$E[\tilde{f}_v]=E\left[\dfrac{d\sum_{j\in[n/d]}I_{\{y_j=v\}}-nq}{n(p-q)}\right]=\dfrac{d\cdot E[\sum_{j\in[n/d]}I_{\{y_j=v\}}]-nq}{n(p-q)}=\dfrac{d\left[\frac{n}{d}\cdot f_v\cdot p+\frac{n}{d}\cdot(1-f_v)\cdot q\right]-nq}{n(p-q)}=f_v.$ □

定理 3. 假设 $\tilde{f}_v$ 表示 v 的估计频率，k_i 表示第 i 个类别属性的取值域大小，则方差 $Var[\tilde{f}_v]$ 满足:

$$Var[\tilde{f}_v] \leqslant \frac{d\left(\dfrac{\varepsilon_c^2(n/d-1)}{14\ln(2/\delta)}-1\right)}{n\left[\dfrac{\varepsilon_c^2(n/d-1)}{14\ln(2/\delta)}-k_i\right]^2}.$$

证明: $Var[\tilde{f}_v]=Var\left[\dfrac{d\sum_{j\in[n/d]}l_{\{y_j=v\}}-nq}{n(p-q)}\right]=Var\left[\dfrac{d\sum_{j\in[n/d]}l_{\{y_j=v\}}}{n(p-q)}\right]=\dfrac{d^2}{n^2(p-q)^2}\sum_{j\in[n/d]}Var[l_{\{y_j=v\}}]$.

对于第 i 组用户而言, 存在 $\dfrac{nf_v}{d}$ 个用户发布真实值 v, 对应 $\sum_{j\in[n/d]}Var[l_{\{y_j=v\}}]=\dfrac{nf_v}{d}p(1-p)$; 存在 $n(1-f_v)$个用户真实值不是 v, 对应 $\sum_{j\in[n/d]}Var[l_{\{y_j=v\}}]=\dfrac{n(1-f_v)}{d}q(1-q)$. 于是可知:

$$Var[\tilde{f}_v]=\frac{d^2}{n^2(p-q)^2}\left[\frac{nf_v}{d}p(1-p)+\frac{n(1-f_v)}{d}q(1-q)\right]=\frac{dq(1-q)}{n(p-q)^2}+\frac{df_v(1-p-q)}{n(p-q)}.$$

按照文献[6]中的处理, f_v 通常为较小的值, 可得:

$$Var[\tilde{f}_v]\approx\frac{dq(1-q)}{n(p-q)^2}=\frac{d(\mathrm{e}^{\varepsilon_l}+k_i-2)}{n(\mathrm{e}^{\varepsilon_l}-1)^2}=\frac{d\left(\dfrac{\varepsilon_c^2(n/d-1)}{14\ln(2/\delta)}-1\right)}{n\left[\dfrac{\varepsilon_c^2(n/d-1)}{14\ln(2/\delta)}-k_i\right]^2} \tag{3}$$

证毕. □

3.1.3 SRR-MS 通信代价及复杂度分析

在 SRR-MS 算法中, 每个用户端仅调用一次 RR 算法发布某个类别属性上的值, RR 扰动算法的时间复杂度为 O(1), 空间复杂度为 O(1), 所以 SRR-MS 算法中, 用户端的时间、空间复杂度与 RR 相同都为 O(1). 对于 Shuffler S_i 发过来的 n/d 个用户的扰动结果, 数据收集端需要统计 k_i 个取值上的频率估计. 由于每个用户仅发布一个值, 所以需进行 n/d 次累加, 利用 k_i 个计数器记录每种取值的累加频率. 所以为了统计第 i 个类别属性上的频率信息, SRR-MS 的时间复杂度为 O(n/d), 空间复杂度为 O(k_i). 由于需要统计 d 个 Shuffler 发送的结果, SRR-MS 算法在收集者端的时间复杂度为 O(n), 空间复杂度为 O$\left(\sum k_i\right)$.

由于每个用户仅发布某个属性上的取值给洗牌者, 所以位于第 i 个分组内单个用户端与洗牌者之间的通信开销为 O($\log_2 k_i$), n 个用户的平均通信开销为 O$\left(\sum_{i=1}^{d}\log_2 k_i/d\right)$. 洗牌者将洗牌结果传给收集者, 传输第 i 组用户的洗牌结果的通信开销为 O($n\log_2 k_i/d$), 传输 d 组洗牌结果的通信开销为 O$\left(\sum_{i=1}^{d}n\log_2 k_i/d\right)$.

3.2 基于单洗牌者的多维扰动发布算法(ARR-SS)

每一维取值域大小不同, 为隐私收益分析带来挑战. 本文提出了 ARR-SS 算法, 将 RR 机制的扰动域设为所有维度的所有取值的集合, 完成同参数化处理. 具体来说, 每个用户随机选择一个属性值发布, 在所有属性范围上进行扰动. 如图 2 所示, 数据库中具有 4 个属性, 此处以每个属性只有两种取值为例. 每个用户随机选择之后的值都定位到一个长度为 8 的数组中, 例如用户 1 选择了编号为 1 的属性, 值为 1, 因此它在数组的第 4 个位置. 每个用户随机选择的属性值定位到相应的数组中, 使用 RR 机制扰动, 即以 $p=\mathrm{e}^{\varepsilon_l}/\left(\mathrm{e}^{\varepsilon_l}+\sum k_i-1\right)$ 的概率发布第 i 个属性取值域上的真实值, 以 $q=1/\left(\mathrm{e}^{\varepsilon_l}+\sum k_i-1\right)$ 的概率发布该取值域上的非真实取值. 得到扰动结果后, 进行反定位, 获取扰动结果位于的维度 Θ_i 以及在该维度上的取值 y_i. 经过反定位后的数据发给唯一的 Shuffler, 打乱后的数据($[\langle\tilde{\Theta}_j,\tilde{y}_j\rangle]_{j\in[n]}$)发送给收集者统计频率分布 $\tilde{F}$. 具体来说, 收集者对带噪数据按

维度编号进行分组，计算第 i 个属性上取值为 v 的频率信息 $\tilde{f}_v=\dfrac{d\left(\sum_{j\in[n]}I_{\{\tilde{\Theta}_j=i\wedge\tilde{y}_j=v\}}-nq\right)}{n(p-q)}$. ARR-SS 只设置一个 Shuffler 进行洗牌，参与洗牌的用户数比 SRR-MS 多，隐私收益高. 然而扰动域较大，因此降低了发布结果的准确性. ARR-SS 算法的具体过程如算法 2 所示.

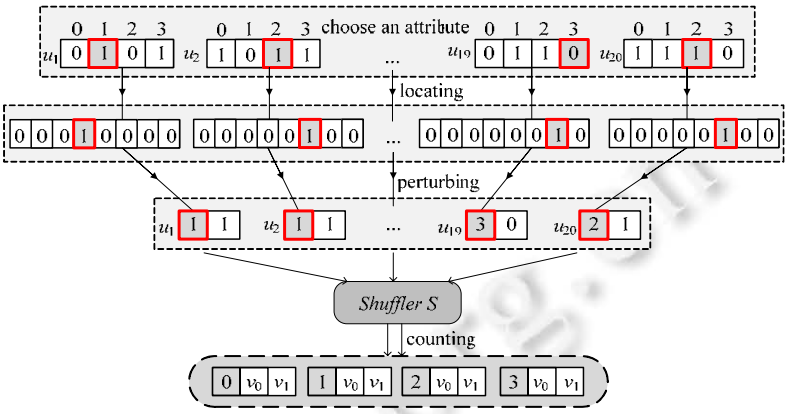


图 2　应用 ARR-SS 算法的多维类别数据发布框架

算法 2. ARR-SS.

输入: n 个用户的数据(第 i 个用户数据为 $v_i=(v_{i1},v_{i2},\ldots,v_{id})$)，属性维度 d 及相应取值域大小$\{k_1,k_2,\ldots,k_d\}$;

输出: 多维类别数据的频率估计分布 $\tilde{F}$.

1. $total\_domain=Sum(k_1,k_2,\ldots,k_d)$;
2. $\varepsilon_l=Calculate(\varepsilon_c,\delta)$;　//计算放大后的隐私预算$\varepsilon_l$
3. **for** users i=1 to n **do**
4. $\Theta_i'=Random(0,d)$;　//随机选一维属性
5. $index=Locate(v_{i\Theta_i'})$;　//获取 $v_{i\Theta'}$在所有属性范围内的索引
6. $x_i=RR(index,total\_domain,\varepsilon_l)$;　//在所有属性范围内扰动
7. $\Theta_i, y_i=ReLocate(x_i)$;　//反定位
8. User i send Θ_i, y_i to Shuffler S;　//第 i 个用户把带噪数据 Θ_i, y_i 发送给洗牌器
9. **end for**
10. Shuffler S 打乱所有带噪数据，并将其发送给收集方;
11. 收集方根据带噪数据计算频率、进行无偏校正并发布带噪分布 $\tilde{F}$;
12. **return** $\tilde{F}$.

3.2.1　ARR-SS 算法的隐私效用分析

定理 4. ARR-SS 算法满足(ε_c,δ)-混洗差分隐私，其中，$\varepsilon_c\leqslant\sqrt{\dfrac{14\ln(2/\delta)(e^{\varepsilon_l}+\sum k_i-1)}{n-1}}$.

证明: 根据第 2.3 节中隐私毯子分解可得:

$$\Pr[RR(v)=y]=(1-\gamma)\Pr^v[y]+\gamma\Pr[Uni(\Psi)=y],$$

其中，$\gamma=\sum k_i/(e^{\varepsilon_l}+\sum k_i-1)$, $\Pr^v[y]$是发布真实值 v 的用户形成的分布，$Uni(\Psi)$表示以 $\Pr[Uni(\Psi)=y]=1/\sum k_i$ 为概率的随机选择. 用 A 表示 ARR-SS 算法，$\sum k_i$ 是所有数据取值域大小之和，Θ_j 表示第 j 个用户随机选到发布数据的属性，$[\langle\Theta_j,y_j\rangle]_{j\in[n]}$表示洗牌之前的用户输出，$R=[\langle\tilde{\Theta}_j,\tilde{y}_j\rangle]_{j\in[n]}$ 表示洗牌之后的输出. Π表示 n 个用户的某种排列顺序.

下面证明 $\Pr_{R\sim A(D)}\left[\frac{\Pr[A(D)=R]}{\Pr[A(D')=R]}\geqslant \mathrm{e}^{\varepsilon_c}\right]\leqslant\delta$. 假设第 n 个用户发布真实值, 用 T 表示前 n–1 个用户中发布真实值的部分, R_T表示前 n–1 个用户中发布随机值的部分, $\Pr[A(D)=R|(T,R_T)]$可以表示为:

$$\begin{aligned}&\Pr[A(D)=R\,|\,(T,R_T)]\\&=\sum_{\pi}\Pr[\pi]\left\{\prod_{j\in T}\Pr[\Theta_{\pi(j)}]l_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}\cdot\prod_{j\in R_T}\Pr[\Theta_{\pi(j)}]\frac{1}{\sum k_i}\cdot\Pr[\Theta_{\pi(n)}]l_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\}\\&=\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}l_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}\cdot\prod_{j\in R_T}\frac{1}{d}\frac{1}{\sum k_i}\cdot\frac{1}{d}l_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\},\\&\frac{\Pr[A(D)=R\,|\,(T,R_T)]}{\Pr[A(D')=R\,|\,(T,R_T)]}=\frac{\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}l_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}\prod_{j\in R_T}\frac{1}{d}\frac{1}{\sum k_i}\cdot\frac{1}{d}l_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\}}{\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}l_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}\prod_{j\in R_T}\frac{1}{d}\frac{1}{\sum k_i}\cdot\frac{1}{d}l_{\{\Theta_{\pi(n)}(v_n')=y_{\pi(n)}\}}\right\}}\\&=\frac{c\sum_{\pi\in P}l_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}}{c\sum_{\pi\in P}l_{\{\Theta_{\pi(n)}(v_n')=y_{\pi(n)}\}}}=\frac{\sum_{j\in[n]\backslash T}\sum_{\pi\in P_j}l_{\{\tilde{\Theta}_j(v_n)=\tilde{y}_j\}}}{\sum_{j\in[n]\backslash T}\sum_{\pi\in P_j}l_{\{\tilde{\Theta}_j(v_n')=\tilde{y}_j\}}}\\&=\frac{\sum_{j\in[n]\backslash T}l_{\{\tilde{\Theta}_j(v_n)=\tilde{y}_j\}}}{\sum_{j\in[n]\backslash T}l_{\{\tilde{\Theta}_j(v_n')=\tilde{y}_j\}}}=\frac{N_{R,T,R_T}}{N'_{R,T,R_T}},\end{aligned}$$

其中, P 表示所有令发布真实值的用户($j\in T$)对应的 $l_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}$ 不为 0 的用户排列, 可以被划分为 n–$|T|$个相同大小的子集. 对于 $j\in[n]\backslash T$, P_j 表示第 n 个用户的扰动值对应第 j 个位置的数据, 而其余$[n-1]\backslash T$ 个用户是不确定的, 因此, P_j 的大小为 $l_{\{\tilde{\Theta}_j(v_n)=\tilde{y}_j\}}\cdot(n-1-|T|)!$.

N_{R,T,R_T} 服从伯努利分布 $N_{R,T,R_T}\sim Bin\left(n-1,\frac{1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}\right)+1$, 而 $N'_{R,T,R_T}\sim Bin\left(n-1,\frac{1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}\right)$.

令 $\theta=E[N'_{R,T,R_T}]=\frac{n-1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}$, 可得 $\Pr\left[\frac{N_{R,T,R_T}}{N'_{R,T,R_T}}\geqslant\mathrm{e}^{\varepsilon_c}\right]\leqslant\exp\left(-\frac{\theta}{3}\left(\mathrm{e}^{\varepsilon_c/2}-1-\frac{1}{\theta}\right)^2\right)+\exp\left(-\frac{\theta}{2}(1-\mathrm{e}^{-\varepsilon_c/2})^2\right)$.

令 $\theta=\frac{n-1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}=\frac{14\ln(2/\delta)}{\varepsilon_c^2}$, 可知 A 满足 $\left(\sqrt{\frac{14\ln(2/\delta)(\mathrm{e}^{\varepsilon_l}+\sum k_i-1)}{n-1}},\delta\right)$-混洗差分隐私. □

3.2.2 ARR-SS 算法的可用性分析

定理 5. 假设 f_v 和 $\tilde{f}_v$ 分别表示 v 的真实频率和估计频率, 则 $E[\tilde{f}_v]=f_v$ 成立.

证明: $E[\tilde{f}_v]=E\left[\frac{d(\sum_{j\in[n]}l_{\{\tilde{\Theta}_j=i\wedge\tilde{y}_j=v\}}-nq)}{n(p-q)}\right]=\frac{d\left[nf_v\left(\frac{1}{d}\cdot p+\frac{d-1}{d}\cdot q\right)+n(1-f_v)q\right]-dnq}{n(p-q)}=f_v$. □

定理 6. 假设 $\tilde{f}_v$ 表示 v 的估计频率, $\sum k_i$ 表示所有属性取值域大小之和, 则方差 $Var[\tilde{f}_v]$ 满足:

$$Var[\tilde{f}_v]\approx\frac{d^2\left[\frac{\varepsilon_c^2(n-1)}{14\ln(2/\delta)}-1\right]}{n\left[\frac{\varepsilon_c^2(n-1)}{14\ln(2/\delta)}-\sum k_i\right]^2}.$$

证明: $Var[\tilde{f}_v]=Var\left[\frac{d(\sum_{j\in[n]}l_{\{\tilde{\Theta}_j=i\wedge\tilde{y}_j=v\}}-nq)}{n(p-q)}\right]=Var\left[\frac{d\sum_{j\in[n]}l_{\{\tilde{\Theta}_j=i\wedge\tilde{y}_j=v\}}}{n(p-q)}\right]=\frac{d^2}{n^2(p-q)^2}\sum_{j\in[n]}Var[l_{\{\tilde{\Theta}_j=i\wedge\tilde{y}_j=v\}}]$.

有 nf_v 个用户发布真实值 v, 对应 $\sum_{j\in[n]} Var[l_{\{\tilde{\Theta}_j=i\wedge\tilde{y}_j=v\}}]=nf_v\left[\frac{p}{d}+\frac{q(d-1)}{d}\right]\left[1-\frac{p}{d}-\frac{q(d-1)}{d}\right]$; 存在 $n(1-f_v)$个用户真实值不是 v, 对应 $\sum_{j\in[n]} Var[l_{\{\tilde{\Theta}_j=i\wedge\tilde{y}_j=v\}}]=n(1-f_v)q(1-q)$, 于是可知:

$$Var[\tilde{f}_v]=\frac{d^2}{n^2(p-q)^2}\left\{nf_v\left[\frac{p}{d}+\frac{q(d-1)}{d}\right]\left[1-\frac{p}{d}-\frac{q(d-1)}{d}\right]+n(1-f_v)q(1-q)\right\}\approx\frac{d^2\left[\frac{\varepsilon_c^2(n-1)}{14\ln(2/\delta)}-1\right]}{n\left[\frac{\varepsilon_c^2(n-1)}{14\ln(2/\delta)}-\sum k_i\right]^2} \tag{4}$$

证毕. □

3.2.3 ARR-SS 通信代价及复杂度分析

在 ARR-SS 算法中, 每个用户端调用一次 RR 算法发布其真实数据定位到的位置编号, 然后根据发布的编号进行反定位, 进而完成发布. 由于 d 个属性上每个属性取值域大小可能不同, 所以需遍历每个维度, 探测扰动数据是否位于此维度, 该部分的时间复杂度为 O(d). ARR-SS 算法中, 用户端的时间复杂度为 O(d), 空间复杂度为 O(1). 对于 Shuffler 发过来的 n 个用户的扰动结果, 需进行 n 次累加, 统计 $\sum k_i$ 个取值上的频率估计. ARR-SS 在数据收集端的时间复杂度为 O(n), 空间复杂度为 $O\left(\sum k_i\right)$.

由于每个用户需将反定位数据(包括所选维度信息及扰动结果)发送给洗牌者, 所以位于单个用户端与洗牌者之间的通信开销为 $O\left(\log_2 d+\log_2\sum k_i\right)$, 其中, k_i 表示发布维度的取值域大小. 洗牌者需要将洗牌结果发送给收据收集者, 传输 n 个用户洗牌结果的通信开销为 $O\left(n\left(\log_2 d+\log_2\sum k_i\right)\right)$.

4 基于取值域填补的单洗牌者发布算法(PSRR-SS)

通过方差分析可看出, SRR-MS 和 ARR-SS 算法存在缺陷: SRR-MS 算法为用户分组使得洗牌的用户数量减小, 导致隐私收益降低; ARR-SS 算法中用户在所有属性范围内扰动, 误差较大. 为了更好地应用 SDP 模型, 本文提出了基于取值域填补的单洗牌者发布算法, 该方法直接对取值域处理消除不同类别属性的异构.

4.1 PSRR-SS算法

现用 k_i 表示第 i 个属性的取值域大小, $k_{\max}$ 为所有取值域的最大值. 为消除属性间的异构, 一种最直观的方式是向每个维度的取值域添加 $k_{\max}-k_i$ 个虚拟的取值, 使所有属性取值域为 $k_{\max}$, 将填补后的域作为 RR 的扰动域. 如图 3 所示, 数据库维度是 3, 每一维取值范围依次是 2, 3, 5. 这样, 每一维度需要填充到的最大值是 5. 以用户 1 为例, 随机选到编号为 0 的属性, 真实值为 1, 因此定位到 0 属性的第 2 个位置. 接下来, 将扰动范围填充到 5 个值, 使用 RR 机制, 以 $p=e^{\varepsilon_l}/(e^{\varepsilon_l}+k_{\max}-1)$ 的概率发布真实值, 以 $q=1/(e^{\varepsilon_l}+k_{\max}-1)$ 的概率发布其余值. 所有用户将扰动后的数据($[\langle\Theta_j,y_j\rangle]_{j\in[n]}$)发送给唯一的 Shuffler 打乱, 打乱后的数据($[\langle\tilde{\Theta}_j,\tilde{y}_j\rangle]_{j\in[n]}$)发送给收集者统计频率分布 $\tilde{F}$. 使用 $\tilde{f}_v=\frac{d\sum_{j\in[n]}l_{\{\tilde{\Theta}_j=i\wedge y_j=v\}}-nq}{n(p-q)}$ 计算第 i 个属性上的取值为 v 的频率, 最后将填充位置消除. 用户在本地使用 RR 机制的扰动范围为属性取值域最大值, 发布结果比 ARR-SS 更准确. 此外, PSRR-SS 设置一个 Shuffler, 隐私收益高, 具体过程见算法 3.

算法 3. PSRR-SS.

输入: n 个用户的数据(第 i 个用户数据为 $v_i=(v_{i1},v_{i2},\ldots,v_{id})$), 属性维度 d 及相应的取值域大小$\{k_1,k_2,\ldots,k_d\}$, 隐私预算ε_c, δ.

输出: 多维类别数据的频率估计分布 $\tilde{F}$.

1. $k_{\max}=\mathrm{Max}(k_1,k_2,\ldots,k_d)$;
2. $\varepsilon_l=Calculate(\varepsilon_c,\delta)$; //计算放大后的隐私预算$\varepsilon_l$
3. **for** attributes i=1 to d **do**

4. *Padding*(k_i,$k_{\max}$); //用虚拟值填充每一个维度的取值域
5. **end for**
6. **for** users i=1 to n **do**
7. Θ_i=*Random*(0,d); //随机选一维属性
8. $y_i = RR(v_{i\Theta_i}, k_{\max}, \varepsilon)$; 在 $k_{\max}$ 内扰动
9. User i send Θ_i, y_i to Shuffler S; //第 i 个用户把带噪数据Θ_i, y_i 发送给洗牌器
10. **end for**
11. Shuffler S 打乱所有带噪数据, 并将其发送给收集方;
12. 收集方计算频率、进行无偏校正、消除填充的无效值并发布带噪分布 $\tilde{F}$;
13. **return** $\tilde{F}$.

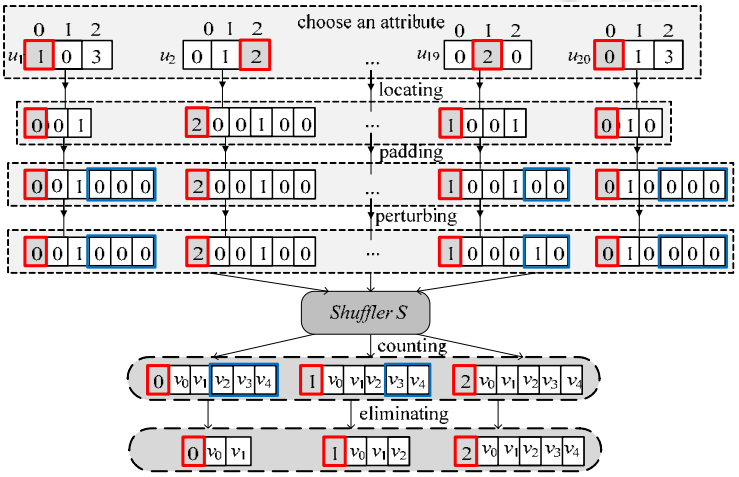


图 3 应用 PSRR-SS 方法的多维类别数据发布框架

4.2 PSRR-SS算法的隐私效用分析

定理 7. PSRR-SS 算法满足(ε_c,δ)-混洗差分隐私, 其中, $\varepsilon_c \leqslant \sqrt{\dfrac{14\ln(2/\delta)(\mathrm{e}^{\varepsilon_l}+k_{\max}-1)}{n-1}}$, $k_{\max}$ 是类别属性域的最大取值.

证明: 与定理 4 描述类似, 用 A 表示 PSRR-SS 算法, 下面证明 $\Pr_{R\sim A(D)}\left[\dfrac{\Pr[A(D)=R]}{\Pr[A(D')=R]} \geqslant \mathrm{e}^{\varepsilon_c}\right] \leqslant \delta$. 用 T 表示前 n–1 个用户中发布真实值的部分, R_T 表示前 n–1 个用户发布随机值的部分, 则:

$$\Pr[A(D)=R\,|\,(T,R_T)] = \sum_{\pi}\Pr[\pi]\left\{\prod_{j\in T}\Pr[\Theta_{\pi(j)}]I_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j \wedge y_{\pi(j)}=\tilde{y}_j\}}\cdot\prod_{j\in R_T}\Pr[\Theta_{\pi(j)}]\frac{1}{k_{\max}}\cdot\Pr[\Theta_{\pi(n)}]I_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\}$$

$$= \sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}I_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j \wedge y_{\pi(j)}=\tilde{y}_j\}}\cdot\prod_{j\in R_T}\frac{1}{d}\frac{1}{k_{\max}}\cdot\frac{1}{d}I_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\},$$

$$\frac{\Pr[A(D)=R\,|\,(T,R_T)]}{\Pr[A(D')=R\,|\,(T,R_T)]} = \frac{\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}I_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j \wedge y_{\pi(j)}=\tilde{y}_j\}}\cdot\prod_{j\in R_T}\frac{1}{d}\frac{1}{k_{\max}}\cdot\frac{1}{d}I_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\}}{\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}I_{\{\Theta_{\pi(j)}=\tilde{\Theta}_j \wedge y_{\pi(j)}=\tilde{y}_j\}}\cdot\prod_{j\in R_T}\frac{1}{d}\frac{1}{k_{\max}}\cdot\frac{1}{d}I_{\{\Theta_{\pi(n)}(v'_n)=y_{\pi(n)}\}}\right\}} = \frac{\sum_{j\in[n]\setminus T}I_{\{\tilde{\Theta}_j(v_n)=\tilde{y}_j\}}}{\sum_{j\in[n]\setminus T}I_{\{\tilde{\Theta}_j(v'_n)=\tilde{y}_j\}}} = \frac{N_{R,T,R_T}}{N'_{R,T,R_T}}.$$

N_{R,T,R_T} 服从伯努利分布 $N_{R,T,R_T} \sim Bin\left(n-1, \frac{1}{\mathrm{e}^{\varepsilon_l}+k_{\max}-1}\right)+1$, 而 $N'_{R,T,R_T} \sim Bin\left(n-1, \frac{1}{\mathrm{e}^{\varepsilon_l}+k_{\max}-1}\right)$.

令 $\theta = E[N'_{R,T,R_T}] = \frac{n-1}{\mathrm{e}^{\varepsilon_l}+k_{\max}-1} = \frac{14\ln(2/\delta)}{\varepsilon_c^2}$, 可知 A 满足 $\left(\sqrt{\frac{14\ln(2/\delta)(\mathrm{e}^{\varepsilon_l}+k_{\max}-1)}{n-1}}, \delta\right)$-混洗差分隐私. □

4.3 PSRR-SS算法的可用性分析

定理 8. 假设 f_v 和 $\tilde{f}_v$ 分别表示 v 的真实频率和估计频率, 则 $E[\tilde{f}_v] = f_v$ 成立.

证明: $E[\tilde{f}_v] = E\left[\frac{d\sum_{j\in[n]} l_{\{\tilde{\Theta}_j=i \wedge y_j=v\}} - nq}{n(p-q)}\right] = \frac{dE[\sum_{j\in[n]} l_{\{\tilde{\Theta}_j=i \wedge y_j=v\}}] - nq}{n(p-q)} = \frac{d\left[nf_v\frac{p}{d} + n(1-f_v)\frac{q}{d}\right] - nq}{n(p-q)} = f_v$. □

定理 9. 假设 $\tilde{f}_v$ 表示 v 的估计频率, $k_{\max}$ 表示类别数据的最大取值域, 则方差 $Var[\tilde{f}_v]$ 满足:

$$Var[\tilde{f}_v] \approx \frac{\frac{d\varepsilon_c^2(n-1)}{14\ln(2/\delta)} - 1}{n\left[\frac{\varepsilon_c^2(n-1)}{14\ln(2/\delta)} - k_{\max}\right]^2}.$$

证明:

$$\begin{aligned}
Var[\tilde{f}_v] &= Var\left[\frac{d\sum_{j\in[n]} l_{\{\tilde{\Theta}_j=i \wedge \tilde{y}_j=v\}} - nq}{n(p-q)}\right] \\
&= Var\left[\frac{d\sum_{j\in[n]} l_{\{\tilde{\Theta}_j=i \wedge \tilde{y}_j=v\}}}{n(p-q)}\right] \\
&= \frac{d^2}{n^2(p-q)^2}\sum_{j\in[n]} Var[l_{\{\tilde{\Theta}_j=i \wedge \tilde{y}_j=v\}}] \\
&= \frac{d^2}{n^2(p-q)^2}\left[nf_v\frac{p}{d}\left(1-\frac{p}{d}\right) + n(1-f_v)\frac{q}{d}\left(1-\frac{q}{d}\right)\right] \\
&= \frac{q(d-q)}{n(p-q)^2} + \frac{f_v(d-p-q)}{n(p-q)} \\
&\approx \frac{\frac{d\varepsilon_c^2(n-1)}{14\ln(2/\delta)} - 1}{n\left[\frac{\varepsilon_c^2(n-1)}{14\ln(2/\delta)} - k_{\max}\right]^2}
\end{aligned} \tag{5}$$

证毕. □

4.4 PSRR-SS通信代价及复杂度分析

在 PSRR-SS 算法中, 每个用户端仅调用一次 RR 算法, 在 PSRR-SS 算法中, 用户端的时间、空间复杂度与 RR 相同为 O(1). 对于 Shuffler 发来的扰动结果, 需统计 $dk_{\max}$ 个取值上的频率估计. 由于每个用户仅发布一个值, 所以需进行 n 次累加, 记录每种取值的累加频率, 所以 PSRR-SS 的时间复杂度为 O(n), 空间复杂度为 O($dk_{\max}$). 由于每个用户需发布所选维度信息及扰动结果, 所以单个用户端与洗牌者之间的通信开销为 O($\log_2 d + \log_2 k_{\max}$). 洗牌者发送洗牌结果的通信开销为 O($n(\log_2 d + \log_2 k_{\max})$).

- SRR-MS, ARR-SS 和 PSRR-SS 性能理论分析比较.

3 种算法都是在满足(ε_c,δ)-SDP 的前提下完成数据的发布, 所以 3 种策略的隐私保护程度相同. 在发布结果精度方面, 观察 3 种方案的误差公式(3)–公式(5)发现: SRR-MS 和 ARR-SS 的理论误差级别正比于 d^2; 而 PSRR-SS 的误差级别为 d, 且 PSRR-SS 参与洗牌的用户数量多. 明显地, 从发布结果精度方面, PSRR-SS 有较

大的优势. 在通信开销方面, 由于 SRR-MS 使用多个洗牌者, 每个洗牌者发布收集固定维度的用户扰动值, 所以不需传输维度信息, 通信代价仅为 $O\left(\log_2 \sum k_i\right)$; 而另外两种方案都是使用一个洗牌者, 需要在加噪过程中动态地选择发布维度, 所以需要传递维度信息以及维度上的扰动之信息, 通信代价为 $O\left(\log_2 d+\log_2 \sum k_i\right)$ 或 $O(\log_2 d+\log_2 k_{\max})$. 可见, 方案 SRR-MS 具有较低的通信开销. 另外, 由于 PSRR-SS 使用取值域填充机制, 所以为了传递每个维度上的扰动值, 需使用 $\log_2 k_{\max}$ 个 bit 编码, 传输开销略大于 SRR-MS. 3 种方案都使用基本的 RR 对数据进行扰动, 其中只有 ARR-SS 需对扰动数据进行反定位处理, 所以用户端时间复杂度方面, SRR-MS 和 PSRR-SS 相近, ARR-SS 会差一些. 综上分析, 可见 PSRR-SS 时间复杂度以及发布结果可用性方面都具有较好的性能, 在通信方面与 SRR-MS 相比需额外传递 $\log_2 d$ 个 bit, 在可接受范围内.

5 实验结果与分析

为验证本文所提的多维数据处理技术的有效性, 本文从单个维度上的频率分布(1-marginal)、多个维度上的频率分布(k-marginal)以及随机梯度下降模型学习结果发布这 3 种发布任务出发, 通过对比已有技术, 评估本文所提方法 SRR-MS、ARR-SS 以及 PSRR-SS 在结果发布精度以及通信方面的性能. 实验使用了 4 个真实数据集, 包括二进制数据集 Kosarak 和非二进制数据集 Bank、Adult、Mexico, 数据集具体信息见表 1. 所有实验通过固定δ, 变化ε_c 展示算法性能, δ的取值为数据集规模的倒数. 如无特殊说明, 展示的所有结果为实验运行 20 次, 取平均值的结果. 实验机内存为 16 GB, CPU 为 Intel(R) Core(TM) i5-10500 CPU@3.10 GHz, 操作系统为 Win10.

表 1 数据集

名称	用户数	属性个数	最大取值域
Kosarak	65 536	8	2
Bank	45 211	17	12
Adult	45 222	15	41
Mexico	200 000	13	9

5.1 多维数据单个维度上的频率分布(1-marginal)发布效果比较

本节利用本文所提方案收集所有维度上的频率分布, 即 1-marginal 查询. 由于当前并不存在 SDP 下的多维数据发布技术, 所以并不存在直接可用的比较算法. 因此, 我们借助于方案 SRR-MS 中处理多维数据的框架, 即将用户分为 d 组, 每组用户用于完成单维上的频率分布的统计. 文献[11]给出了混洗差分隐私模型下单维上频率发布技术, SGRR、SUE 以及 SOLH. SUE 与 SOLH 的性能相近, 因此, 我们将 SGRR 与 SUE 作为代表, 将其与 SRR-MS 框架结合对比方案. 与 SGRR 的结合本质为我们所提方案 SRR-MS, 将与 SUE 的结合称为 SUE-MS. 此外, 将 LDP 下的采样技术与 OUE 技术的结合作为对比方案, 展现 SDP 技术在发布结果精度方面的优势. 将所有维度上频率分布的误差平方和作为测试指标, 具体定义见第 2.4 节.

图 4(a)描述了 OUE、SUE-MS、SRR-MS、ARR-SS 和 PSRR-SS 这 5 种算法在 Kosarak 数据集上频率发布的 SSE 结果比较. 在ε_c 变化过程中, 5 种算法的 SSE 均呈下降趋势. 因为ε_c 越大, ε_l 越大, 加入噪音越少, SSE 越小. 当$\varepsilon_c \geqslant 0.4$ 时, SUE-MS 算法存在ε_l. 当$\varepsilon_c=0.4$ 时, ε_l 值较小, 发布效果与 OUE 相近; 当$\varepsilon_c>0.4$ 后, SUE-MS 算法由于洗牌带来隐私收益, 表现优于 OUE. 给定ε_c, 使用 SRR-MS 算法求得的ε_l 最小, 但其扰动域小; ARR-SS 算法求得的ε_l 约为 SRR-MS 算法求得的两倍, 但扰动域大. 综合以上特点, SRR-MS 和 ARR-SS 算法的效果接近. 给定ε_c, PSRR-SS 算法下的ε_l 比 ARR-SS 算法的ε_l 稍大, 隐私收益高, 且扰动域小, 因此表现最好. 图 4(b)为 5 种算法在 Mexico 数据集上的 SSE 结果比较. 当ε_c 小于 0.3 时, ARR-SS 算法下ε_l 为负值, 因此图中没有结果. 显然, 在类别属性较多且取值域异构的前提下, PSRR-SS 算法效果突出, OUE 由于没有放大预算, 表现最差.

图 4(c)和图 4(d)分别是 5 种算法在 Bank 和 Adult 数据集上的 SSE 结果. 由于 SRR-MS 算法只在ε_c 取值较大时才能求得ε_l, 为了将 5 种算法进行对比, 在特定ε_c 下, 只计算 SRR-MS 算法下存在ε_l 的类别属性的 SSE, 并

将 5 种算法进行比较. 图 4(c)中, SRR-MS 算法折线在ε_c从 0.6 至 0.8 过程中呈上升趋势. 这是因为ε_c变大导致存在ε_l的类别属性增多, 而新增属性的ε_l值较小(0.2–0.5), 导致 SSE 变大. 因此, 图 4(c)和图 4(d)只需关注纵向对比结果. 在图 4(c)中, ε_c=0.6 时, ARR-SS 方案下用户扰动数据使用的ε_l数值小, 且扰动域较大, 导致其发布效果不如 OUE. 除此之外, 其余方案得益于隐私放大, 均优于 OUE. 图 4(d)中, ε_c 值为 0.4 时, SRR-MS 比 PSRR-SS 效果好. 这是因为此时只有两个取值域为 2 的类别属性参与 SSE 计算, 而 PSRR-SS 需要填充至 $k_{\max}$=41, 导致 SSE 较大. 除此之外, PSRR-SS 表现一致优于 SRR-MS 和 ARR-SS. SUE-MS 在$\varepsilon_c \geqslant 0.7$ 时能求得ε_l, 对ε_c 取值有一定限制, 并且纵向对比发现该方案表现不如 SRR-MS. 从图 4 可以看出, PSRR-SS 通常在ε_c 取较小值时也能够放大预算, 并且表现始终优于 SUE-MS.

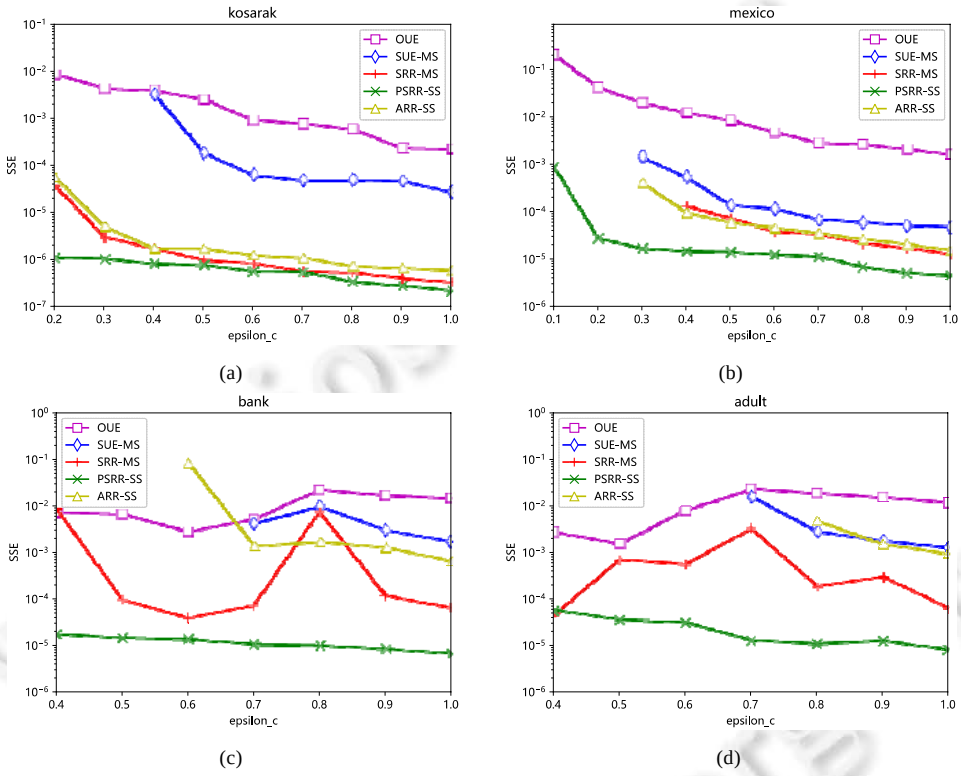


图 4 多维数据单个维度上的频率分布发布效果比较

表 2 和表 3 分别展示了在 Kosarak 和 Mexico 数据集上, PSRR-SS 算法与其他算法的效果相差百分比, 可见, 当ε_c取值较大、性能稳定时, PSRR-SS 在 Kosarak 数据集上领先其他算法 30%左右, 在 Mexico 数据上领先其他算法 60%左右. 表 4 为 4 个数据集上, 所提出算法在个体用户端的通信花销. 其中, SRR-MS 算法通信开销最小, ARR-SS 与 PSRR-SS 除了需发布扰动值还需发布维度信息, 因此两种方案的通信开销大于 SRR-MS. 此外, 以 Mexico 数据集为代表, 本文所提算法的用户端执行时间均可达到 0.005 ms, 具有较高的发布效率.

表 2 方法效果相差百分比(Kosarak)

方法	ε_c								
	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SRR-MS	97.2	65.8	52.7	20.1	34.0	1.9	34.9	31.6	33.5
ARR-SS	98.2	80.0	53.4	56.0	55.4	50.6	54.4	59.1	63.6
SUE-MS	–	–	99.9	99.6	99.2	98.9	99.3	99.4	99.2
OUE	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9

表 3 方法效果相差百分比(Mexico)

方法	ε_c						
	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SRR-MS	88.7	80.2	67.2	66.9	67.6	70.4	64.1
ARR-SS	84.3	75.6	72.2	67.2	72.9	75.0	69.8
SUE-MS	97.2	89.4	89.0	83.2	87.9	89.9	90.1
OUE	99.8	99.8	99.7	99.6	99.7	99.8	99.7

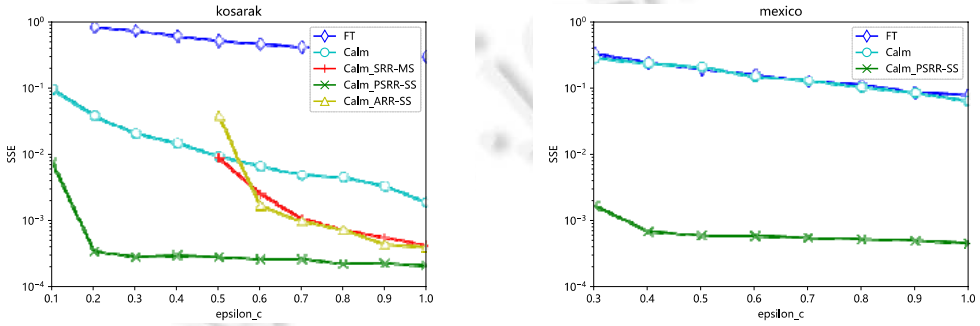
表 4 通信开销

方法	数据集			
	Kosarak	Mexico	Bank	Adult
SRR-MS	1.00	2.69	2.77	3.40
ARR-SS	4.00	11.69	11.77	7.40
PSRR-SS	4.00	13.00	13.00	10.00

5.2 多维数据多个维度上的联合频率分布(*k*-marginal)发布效果比较

指定维度个数 k, 本节所提方案收集所有 k 个维度组合上的联合频率分布信息, 即 k-marginal 查询. d 为数据的维度, 共有 C_d^k 种 k 长度的属性组合, 直接收集所有 marginal 将引入较大的误差. 在 SDP 模型下, 还未存在 k-marginal 发布问题的研究. 在 LDP 模型下, 解决上述问题的两种先进技术为 Calm[20]和 FT[21], 两种方案的思路都是通过发布满足 LDP 约束的可构建 k-marginal 查询结果的中间信息完成最终发布任务. FT 通过扰动傅里叶系数的方式完成 LDP 下 k-marginal 的发布, Calm 通过扰动 m 个长度为 w 的 marginal 完成 k-marginal 的发布. 但上述两种方案都是基于 LDP 设计, 并不适用于 SDP 场景. 由于文献[20]指出其性能优于 FT, 为解决直接发布 C_d^k 个 marginal 误差较大的问题, 本节借助 Calm 的处理框架, 完成 k-marginal 的发布. 具体来说, 将发布 m 个长度为 w 的 marginal 看作为发布 m 个维度的数据, 然后调用 SRR-MS、ARR-SS 和 PSRR-SS 完成 SDP 下 m 个长度为 w 的 marginal 的发布, 形成方案 Calm_SRR-MS、Calm_ARR-SS、Calm_PSRR-SS. 由于在 SDP 模型下未存在 k-marginal 问题的相关研究, 本节将 LDP 下的 Calm 以及 FT 作为对比方案. 参照文献[20]中的评测, 本节展示 3-marginal 上的 SSE 实验结果对比.

Bank 与 Adult 数据集下可计算的ε_l太少, 此处在 Kosarak 和 Mexico 数据集下比较 Marginal 发布效果. 图 5(a)描述了 FT、CALM、Calm_SRR-MS、Calm_ARR-SS 和 Calm_PSRR-SS 这 5 种算法在 Kosarak 数据集上发布 Marginal 的 SSE 结果. Calm_SRR-MS 和 Calm_ARR-SS 算法只在$\varepsilon_c \geqslant 0.5$ 时才能使用, 在$\varepsilon_c=0.5$ 时, Calm_SRR- MS 的ε_l小, 扰动域较大, 因此准确性较低. $\varepsilon_c \geqslant 0.6$ 时, 两算法发布 Marginal 效果相当, 并且效果好于 Calm 与 FT 算法, 这得益于洗牌带来的隐私放大. $\varepsilon_c=0.1$ 时, FT 算法误差超过 1, 因此在图中只展示$\varepsilon_c \geqslant 0.2$ 时的效果. 5 种方法中, Calm_PSRR-SS 效果最好, 在一些情况下, SSE 能够比 Calm 算法提高两个数量级, 比 FT 算法提高 3 个数量级. Calm_PSRR-SS 隐私放大的程度最高, 效果好于 Calm_SRR-MS 和 Calm_ARR-SS, 并且在ε_c 较小时也适用, 因此在 Mexico 数据集上只与 Calm 和 FT 算法对比即可, 效果如图 5(b)所示. 显然, Calm_PSRR-SS 依然表现突出, 将发布 Marginal 的 SSE 降低两个数量级.



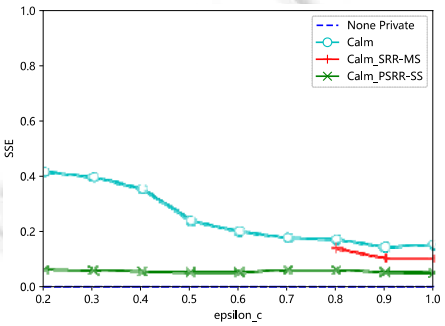
(a) (b)

图 5 基于 kosarak 和 mexico 比较 3-Marginal 发布效果

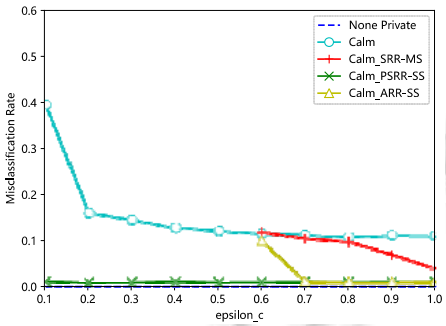
5.3 随机梯度下降技术学习结果比较

文献[20]中给出了基于 m 个 w-marginal 生成合成数据集的方案. 本节利用上述 4 种方案 Calm、SRR-MS、ARR-SS 和 PSRR-SS 发布 w-marginal, 并借助文献[20]完成合成数据集的生成, 使用随机梯度下降技术在合成数据集上完成线性回归、逻辑回归以及支持向量机模型的学习. 随机梯度下降算法中的学习步长 $\eta=1/\sqrt{t}$, t 为迭代次数. 使用具有代表性的非二进制数据集 Mexico 进行此节的测评, 由于合成数据集生成算法的效率较慢, 选择 Mexico 中的部分属性参与此部分评估. 具体来说, 保留 Mexico 中的收入属性作为预测标签, 在线性回归中进行等间距离散化处理, 离散成 10 种取值作为预测标签, 在逻辑回归或支持向量机中, 将收入大于均值的数据的标签设置为 1, 否则设置为–1. 从剩余属性中选择 5 个, 每一个属性转化为多个维度, 每一维度对应一个二进制取值域(例如某个维度的有 c_i 种取值,则转化为 c_i–1 个维度), 转化后的维度为特征属性. 将转化后数据的 80%的用户用于合成数据集的构建, 剩余 20%的用户作为测试数据集. 用误差平方和衡量线性回归方法下训练效果, 用误判率衡量逻辑回归与支持向量机方法下训练效果.

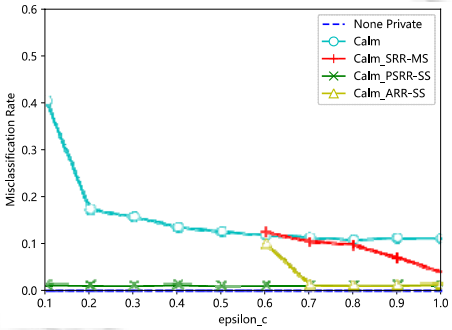
图 6(b)、图 6(c)分别展示了 Calm、Calm_SRR-MS、Calm_ARR-SS 和 Calm_PSRR-SS 这 4 种方案在逻辑回归以及支持向量机模型下的效果对比, 其中, non-private 表示不带噪的原始数据集. 其中, Calm_SRR-MS 和 Calm_ARR-SS 只在$\varepsilon_c\geqslant 0.6$ 时才能使用. 在$\varepsilon_c>0.8$ 时, 应用 SRR-MS 和 ARR-SS 算法得出的带噪数据集训练结果比 CALM 好, 并且 ARR-SS 优于 SRR-MS. 这是因为在$\varepsilon_c=0.7$ 时, 根据定理 1 和定理 4, ARR-SS 计算得到的隐私预算放大值ε_l 超过 SRR-MS 下的 2 倍. 线性回归下收入属性取值域较大, 导致 Calm_ARR-SS 方案的扰动域太大, 无法求得ε_l. 图 6(a)展示在线性回归下, 除 Calm_SRR-MS 以外其余方案的效果对比, Calm_SRR-MS 方案在$\varepsilon_c\geqslant 0.8$ 时能够使用. Calm_PSRR-SS 得出的数据集在 3 种模型中效果良好, 由于隐私放大程度最高, 因此合成数据集能够较好地反映真实数据集的特征, 使学习的模型预测准确性高.



(a) Linear



(b) Log



(c) SVM

图 6 基于 mexico 评估机器学习效果

6 总 结

本文针对多维类别数据的频率估计问题, 基于新兴的混洗差分隐私模型提出了 3 种算法: SRR-MS, ARR-SS 和 PSRR-SS. 前两种为基本方法, 分别从洗牌方式及加噪扰动方式解决异构问题. PSRR-SS 算法基于取值域填补技术、随机响应和单个洗牌者来进行频率收集. 在不分割隐私预算的前提下, 每个用户随机选择一维数据, 在大小一致的取值域上扰动并发送给 Shuffler 洗牌, 不仅破坏了用户和数据间的关联, 还使用户和收集者之间的通信代价较小. 本文从理论角度分析了 3 种算法的隐私性和可用性, 通过 4 个数据集进行误差平方和的对比分析, 并嵌入 Calm 方法合成带噪数据集用于随机梯度下降算法的训练, 结果表明, Calm_PSRR-SS 算法具有较好的效果. 今后的研究考虑如下两个方面: (1) 如何在原始取值域上扰动数据并分析隐私收益; (2) 如何利用多个洗牌者实现高精度的频率估计.

References:

[1] Xu SZ, Su S, Cheng X, Li Z, Xiong L. Differentially private frequent sequence mining via sampling-based candidate pruning. In: Proc. of the ICDE. IEEE Computer Society, 2015. 1035−1046.

[2] Xu SZ, Cheng X, Su S, Xiao K, Xiong L. Differentially private frequent sequence mining. IEEE Trans. on Knowledge and Data Engineering, 2016, 28(11): 2910−2926.

[3] Wang N, Xiao XK, Yang Y, Zhang ZJ, Gu Y, Yu G. Privsuper: A superset-first approach to frequent itemset mining under differential privacy. In: Proc. of the ICDE. IEEE Computer Society, 2018. 809−820.

[4] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proc. of the CCS. ACM, 2014. 1054−1067.

[5] Ren XB, Yu CM, Yu W, Yang S, Yang XY, McCann JA, Yu PS. LoPub: High-dimensional crowdsourced data publication with local differential privacy. IEEE Trans. on Information Forensics and Security, 2018, 13(9): 2151−2166.

[6] Wang TH, Blocki J, Li N, Jha S. Locally differentially private protocols for frequency estimation. In: Proc. of the SP. USENIX Association, 2017. 729−745.

[7] Wang TH, Li N, Jha S. Locally differentially private frequent itemset mining. In: Proc. of the SP. 2018. IEEE Computer Society, 127−143.

[8] Balcer V, Cheu A. Separating local & shuffled differential privacy via histograms. In: Proc. of the CoRR. 2019.

[9] Balle B, Bell J, Gasc´on A, Nissim K. The privacy blanket of the shuffle model. In: Proc. of the CRYPTO. Springer, 2019. 638−667.

[10] Erlingsson U´, Feldman V, Mironov I, Raghunathan A, Talwar K, Thakurta A. Amplification by shuffling: From local to central differential privacy via anonymity. In: Proc. of the SODA. 2019. 2468−2479.

[11] Wang TH, Xu M, Ding B, Zhou JR, Hong C, Huang ZC, Li N, Jha S. Improving utility and security of the shuffler-based differential privacy. Proc. of the VLDB Endowment, 2020, 13(13): 3545−3558.

[12] Li XC, Liu WR, Chen ZY, Huang KZ, Qin Z, Zhang L, Ren K. DUMP: A dummy-point-based framework for histogram estimation in shuffle model. In: Proc. of the CoRR. 2020. 967−984.

[13] Wang TH, Li NH, Jha S. Locally differentially private heavy hitter identification. IEEE Trans. on Dependable and Secure Computing, 2021, 18(2): 982−993.

[14] Wang N, Xiao XK, Yang Y, Hoang TD, Shin H, Shin J, Yu G. PrivTrie: Effective frequent term discovery under local differential privacy. In: Proc. of the ICDE. IEEE Computer Society, 2018. 821−832.

[15] Gu XL, Li M, Cheng YQ, Xiong L, Cao Y. PCKV: Locally differentially private correlated key-value data collection with optimized utility. In: Proc. of the CoRR. 2020. 967−984.

[16] McSherry F, Talwar K. Mechanism design via differential privacy. In: Proc. of the FOCS. IEEE Computer Society, 2007. 94−103.

[17] Qin Z, Yang Y, Yu T, Khalil I, Xiao XK, Ren K. Heavy hitter estimation over set-valued data with local differential privacy. In: Proc. of the SIGSAC. ACM, 2016. 192−203.

[18] Wang N, Xiao XK, Yang Y, Zhao J, Hui SC, Shin H, Shin J, Yu G. Collecting and analyzing multidimensional data with local differential privacy. In: Proc. of the ICDE. IEEE, 2019. 638−649.

[19] Yang JY, Wang TH, Li NH, Cheng X, Su S. Answering multi-dimensional range queries under local differential privacy. Proc. of the VLDB Endowment, 2020, 14(3): 378−390.

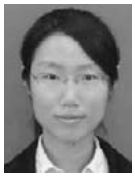
[20] Zhang ZK, Wang TH, Li NH, He SB, Chen JM. CALM: Consistent adaptive local marginal for marginal release under local differential privacy. In: Proc. of the CCS. ACM, 2018. 212−229.

[21] Cormode G, Kulkarni T, Srivastava D. Marginal release under local differential privacy. In: Proc. of the SIGMOD. ACM, 2018. 131−146.

[22] Bittau A, Erlingsson Ú, Maniatis P, Mironov I, Raghunathan A, Lie D, Rudominer M, Kode U, Tinnes J, Seefeld B. Prochlo: Strong privacy for analytics in the crowd. In: Proc. of the 26th Symp. on Operating Systems Principles. ACM, 2017. 441−459.

[23] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In: Proc. of the SIGMOD. ACM, 2009. 19−30.

刘艺菲(1998－), 女, 硕士生, 主要研究领域为数据隐私保护.

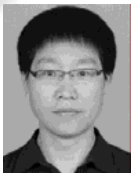


王宁(1988－), 女, 博士, 讲师, CCF 专业会员, 主要研究领域为数据隐私保护, 数据管理.

王志刚(1987－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为云计算, 图数据挖掘.

谷峪(1981－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为图、空间数据管理.

魏志强(1969－), 男, 博士, 教授, CCF 高级会员, 主要研究领域为智能信息处理, 社交媒体以及大数据分析.



张啸剑(1980－), 男, 博士, 讲师, CCF 学生会员, 主要研究领域为隐私保护, 数据挖掘, 图数据管理.

于戈(1962－), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为数据库系统, 数据科学, 大数据技术, 区块链技术.