本地差分隐私下的高维数据发布方法

蔡梦男, 沈国华, 黄志球, 杨阳

引用本文

蔡梦男, 沈国华, 黄志球, 杨阳. 本地差分隐私下的高维数据发布方法[J]. 计算机科学, 2024, 51(2): 322-332.

CAI Mengnan, SHEN Guohua, HUANG Zhiqiu, YANG Yang. High-dimensional Data Publication Under Local Differential Privacy [J]. Computer Science, 2024, 51(2): 322-332.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600142

本地差分隐私下的高维数据发布方法

蔡梦男[1,2]　沈国华[1,2,3]　黄志球[1,2,3]　杨　阳[1,2]
1 南京航空航天大学计算机科学与技术学院　南京 211106
2 南京航空航天大学高安全系统的软件开发与验证技术工业和信息化部重点实验室　南京 211106
3 软件新技术与产业化协同创新中心　南京 210093
(cai_mengnan@nuaa.edu.cn)

摘　要　从众多用户收集的高维数据可用性越来越高,庞大的高维数据涉及用户个人隐私,如何在使用高维数据的同时保护用户的隐私极具挑战性。文中主要关注本地差分隐私下的高维数据发布问题。现有的解决方案首先构建概率图模型,生成输入数据的一组带噪声的低维边缘分布,然后使用它们近似输入数据集的联合分布以生成合成数据集。然而,现有方法在计算大量属性对的边缘分布构建概率图模型,以及计算概率图模型中规模较大的属性子集的联合分布时存在局限性。基于此,提出了一种本地差分隐私下的高维数据发布方法 PrivHDP(High-dimensional Data Publication Under Local Differential Privacy)。首先,该方法使用随机采样响应代替传统的隐私预算分割策略扰动用户数据,提出自适应边缘分布计算方法计算成对属性的边缘分布构建 Markov 网。其次,使用新的方法代替互信息度量成对属性间的相关性,引入了基于高通滤波的阈值过滤技术缩减概率图构建过程的搜索空间,结合充分三角化操作和联合树算法获得一组属性子集。最后,基于联合分布分解和冗余消除,计算属性子集上的联合分布。在 4 个真实数据集上进行实验,结果表明,PrivHDP 算法在 k-way 查询和 SVM 分类精度方面优于同类算法,验证了所提方法的可用性与高效性。

关键词:本地差分隐私;高维数据;数据发布;边缘分布;联合分布
中图分类号　TP309

High-dimensional Data Publication Under Local Differential Privacy

CAI Mengnan[1,2],SHEN Guohua[1,2,3],HUANG Zhiqiu[1,2,3] and YANG Yang[1,2]
1 College of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China
2 Key Laboratory of Safety-Critical Software,Ministry of Industry and Information Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China
3 Collaborative Innovation Center of Novel Software Technology and Industrialization,Nanjing 210093,China

Abstract　With the increasing availability of high-dimensional data collected from numerous users,preserving user privacy while utilizing high-dimensional data poses significant challenges. This paper focuses on the problem of high-dimensional data publication under local differential privacy. State-of-the-art solutions first construct probabilistic graphical models to generate a set of noisy low-dimensional marginal distributions of the input data,and then use them to approximate the joint distribution of the input dataset for generating synthetic datasets. However,existing methods have limitations in computing marginal distributions for a large number of attribute pairs to construct probabilistic graphical models,as well as in calculating joint distributions for attribute subsets within the probabilistic graphical models. To address these limitations,this paper proposes a method PrivHDP(high-dimensional data publication under local differential privacy) for high-dimensional data publication under local differential privacy. Firstly,it uses random sampling response instead of the traditional privacy budget splitting strategy to perturb user data. It proposes an adaptive marginal distribution computation method to compute the marginal distributions of pairwise attributes and construct a Markov network. Secondly,it employs a novel method to measure the correlation between pairwise attributes,replacing mutual information. This method introduces a threshold technique based on high-pass filtering to reduce the search space during the construction of the probabilistic graphical model. It combines sufficient triangulation operations and a joint tree algorithm to obtain a set of attribute subsets. Finally,based on joint distribution decomposition and redundancy elimination,the pro-

到稿日期:2023-06-16　返修日期:2023-11-15
基金项目:国家自然科学基金(U2241216,61772270);民航应急科学与技术重点实验室开放基金(NJ2022022)
This work was supported by the National Natural Science Foundation of China(U2241216,61772270) and Opening Fund of Key Laboratory of Civil Aviation Emergency Science&Technology(CAAC)(NJ2022022).
通信作者:沈国华(ghshen@nuaa.edu.cn)

posed method computes the joint distribution over attribute subsets. Experimental results on four real datasets demonstrate that the PrivHDP algorithm outperforms similar algorithms in terms of k-way query and SVM classification accuracy, validating its effectiveness and efficiency.

Keywords Local differential privacy, High-dimensional data, Data publication, Marginal distribution, Joint distribution

1 引言

随着互联网技术和大数据产业的发展，日常生活中有越来越多的数据被采集。集成传感器和传感系统的普及使得能够从多个来源收集和分析数据，更好地了解人群的相关知识。相关机构发布了一些规模庞大的数据集，使科研人员能够挖掘其潜在的丰富知识。人们能够从高维数据（具有多个属性的数据）中提取潜在的信息和模式，为群体和个人提供准确可靠的预测。例如，捕捉个人电力消费和社区电力消费之间的关系可以实现智能电网中的实时定价[1]。同样，收集和挖掘历史医疗记录和遗传信息可以帮助医院工作人员诊断和监测患者的健康状况[2]。来自智能手机用户的环境监测数据可以改善城市规划并提高人们的生活质量[3]。直接使用这些数据集会带来个人隐私泄露的风险。

差分隐私（Differential Privacy，DP）[4-5]被提出作为隐私保护的事实标准。然而，大多数基于差分隐私的隐私保护方案只针对集中式数据集，并假设服务器是可信的。为解决这些问题，本地差分隐私（Local Differential Privacy，LDP）[6]被提出，用于保护分布式参与者的个人数据隐私。在本地差分隐私设置中，存在多个用户和一个服务器，每个用户将扰动后的数据发送给服务器，服务器无法看到每个用户的私有数据，并根据扰动后的数据推断数据分布。本地差分隐私技术能够在保护每个用户隐私的同时收集统计数据，无需依赖对第三方服务器的信任。

引入本地差分隐私后，Ren 等[7]提出了基于本地差分隐私的高维数据集发布方法 LoPub。该方法使用 RAPPOR 扰动用户数据来提供本地隐私保护。获得扰动数据后，聚合器计算成对属性的边缘分布构建概率图模型，基于联合树得到一组属性子集。LoPub 使用一组属性子集上的低维联合分布近似原始数据集的分布生成合成数据集。然而，LoPub 存在两个局限：1）使用期望最大化（Expectation Maximization，EM）方法推断边缘分布时分割隐私预算导致误差较大，基于互信息计算成对属性相关性构建概率图模型，导致依赖图中存在一些大的属性簇，降低了合成数据集的质量；2）计算大规模的属性子集的联合分布时表现不佳，仍然面临高维度。

为弥补上述不足，本文提出了一种高效的本地差分隐私下高维数据发布方法 PrivHDP。本文的主要贡献包括 4 个方面：

1）提出了自适应边缘分布估计方法，该方法接受来自用户的扰动数据，并计算成对属性的边缘分布，以构建 Markov 网。使用基于抽样的随机响应代替隐私预算分割策略收集用户的扰动数据。

2）使用新的度量方法代替互信息度量成对属性间的相关性，引入基于高通滤波的阈值过滤技术，缩减了概率图构建时的搜索空间，结合充分三角化操作和联合树算法获得一组属性子集。

3）基于联合分布分解和冗余消除，计算大规模的属性子集的联合分布。

4）在 4 个真实的数据集上进行实验，实验结果验证了所提方法的可用性和高效性。

2 相关工作

目前，高维数据的隐私保护工作主要分为两类：基于差分隐私的高维数据发布和基于本地差分隐私的高维数据发布。

差分隐私下最先进的解决方案是利用概率图模型解决这个问题。Zhang 等[8]提出了 PrivBayes，通过构建一个满足差分隐私的贝叶斯网络来近似所有属性的联合分布，最后基于这个分布生成合成数据。Chen 等[9]提出了 JTree，该方法通过识别所有属性的成对独立性构建 Markov 网和联合树，通过联合树算法和拉普拉斯机制生成联合分布，最终生成合成数据集。Chen 等[10]采用朴素贝叶斯的条件独立假设来计算原数据集的联合分布，然后采用指数机制生成发布的数据集。Zhang 等[11]提出了 PrivSyn，该方法首先选择一大组低阶边缘分布并生成一个随机数据集，其中每个属性与数据集中的一维边缘分布匹配。逐渐更新合成的数据集，以与集合中所有边缘分布保持一致。Zhang 等[12]提出了一种基于联合树的隐私高维数据发布方法。Li 等[13]使用 Copula 函数对高维数据的联合分布进行建模。这种方法在属性域较小时效率不高，限制了其在域值较小的属性上的应用。Cai 等[14]通过选择合适的边缘分布构建马尔可夫随机场，随后使用马尔可夫随机场进行数据合成。同时，差分隐私下也存在基于替代技术的其他数据发布方法[15]。然而，与基于概率图的方法相比，其计算效率或性能并不令人满意。

在本地差分隐私的设置中，一个基本问题是所有数据都在用户本地，没有用户拥有全局统计信息，这对推断多维数据属性的分布提出了更大的挑战。大多数现有的 LDP 研究主要集中在单个类别属性上的频率估计。当每个用户都有多个属性时，需要计算多个属性的联合分布。Fanti 等[16]提出了一种基于 EM 的方法，用于估计 LDP 下任意两个属性的联合分布，但是随着属性维数的增大，数据稀疏性会导致巨大的效用损失，同时导致时间复杂度极高。Ren 等[17]通过结合 EM 算法和 LASSO 回归的方法计算多个属性的联合分布。这种方法运行缓慢，并且其为每个属性分割隐私预算，进而导致精度降低。Kulkarni 等[18]将基于傅里叶变换的方法应用于本地差分隐私。该方法用于在集中式差分隐私下计算属性的边缘分布，提高了精度并降低了通信成本，但仅适用于二进制属性集。在基于本地差分隐私的高维数据发布方面，Ren 等[7]提出了 LoPub，用于应对 LDP 下高维数据发布的挑战。然而，LoPub 在计算多维联合分布时应用期望最大化方法导致性能不佳。Wang 等[19]试图通过 copula 函数提高性能，但该方法仍然存在同样的局限。还有一些工作专注于本地差分隐私下的特定任务，例如范围查询和均值估计。文献[20-21]提供了 LDP 下的隐私保护范围查询。其他文献

[22-23]专注于 LDP 下均值估计的具体任务。

3 相关定义

3.1 本地差分隐私

传统差分隐私没有解决因为不可信服务器导致的隐私泄露问题，因而本地差分隐私被提出，用于保护用户端的个人隐私。

定义 1(ε-本地差分隐私) 随机算法 A 满足 ε-本地差分隐私(ε-LDP)，当且仅当对于任何一对输入 x_1 和 x_2 以及任何可能的输出 $T\in Range(A)$，有：

$$Pr[A(x_1)\in T]\leqslant e^{\varepsilon}\times Pr[A(x_2)\in T] \tag{1}$$

其中，$Range(A)$ 表示算法 A 所有可能输出的集合。输出的概率由算法 A 和隐私预算 ε 决定。隐私预算 ε 决定了隐私级别，较小的 ε 意味着更强的隐私保证，反之亦然。由于用户不向聚合器报告其真实输入 x，而只报告 $A(x)$，因此即使聚合器是恶意的，用户的隐私仍然能得到保护。

3.2 LDP 扰动机制

随机响应技术是本地差分隐私的主流扰动机制。本节主要介绍随机响应机制及其变体，包括两个主要步骤：扰动和校正。这些步骤由一对算法指定：每个用户使用 Ψ 扰动输入值，聚合器使用 Φ 校正扰动后的值。

3.2.1 随机响应

1965 年，Warner[24] 首次提出随机响应(Randomized Response，RR)，并将其作为隐私调查访谈中获取敏感或非法行为信息的研究方法。其主要思想是通过对敏感信息(例如犯罪行为)进行合理的否认来保护个人隐私。在随机响应中，每个参与者都持有自己的数据，并以差异隐私的方式回答二进制问题。随机响应机制让用户 i 分别以概率 p 和概率 $1-p$ 报告其真实值 b_i 和其非真实值。

定理 1 随机响应机制满足 ε-LDP，$e^{\varepsilon}=\dfrac{p}{1-p}$。

3.2.2 广义随机响应

随机响应机制用于用户数据为二值数据的情况。Kairouz等[25]提出了一种阶梯机制 K-RR，其设计用于数据为多值数据的情况。Wang 等[26]将其归纳为广义随机响应(Generalized Randomized Response，GRR)，形式化表示如下：d 表示用户属性候选集的大小，v 表示用户的私有真实值，扰动函数为：

$$Pr[\Psi_{GRR}(v)=z]=\begin{cases}p=\dfrac{e^{\varepsilon}}{e^{\varepsilon}+d-1}, & \text{if } z=v\\ q=\dfrac{1}{e^{\varepsilon}+d-1}, & \text{if } z\neq v\end{cases}$$

定理 2 广义随机响应机制满足 ε-LDP，$e^{\varepsilon}=\dfrac{p}{q}$。

为了推断真实值 v 的频率，将用户报告 v 的次数表示为 $C(v)$，用户总数表示为 n，然后计算

$$\Phi_{GRR}(v):=\frac{C(v)/n-q}{p-q} \tag{2}$$

该频率估计的方差为：

$$Var[\Phi_{GRR}(v)]=\frac{d-2+e^{\varepsilon}}{(e^{\varepsilon}-1)^2\times n} \tag{3}$$

3.2.3 优化的一元编码

用户的输入 v 被编码为长度为 d 的位向量，v 对应的位被设置为 1，其余位置为 0。$Encode(v)=[0,\cdots,0,1,0,\cdots,0]$。扰动机制的两个关键参数是 p(1 被扰动后仍为 1 的概率)和 q(0 被扰动后为 1 的概率)。通过最小化方差，优化的一元编码(Optimized Unary Encoding，OUE)[26]选择了合适的参数 p 和 q。扰动函数如下：

$$Pr[B'[i]=1]=\begin{cases}p=\dfrac{1}{2}, & \text{if } B[i]=1\\ q=\dfrac{1}{e^{\varepsilon}+1}, & \text{if } B[i]=0\end{cases}$$

定理 3 优化的一元编码机制满足 ε-LDP，$e^{\varepsilon}=\dfrac{p}{1-p}*\dfrac{1-q}{q}$。

聚合器计算与 v 对应的位置为 1 的报告数量，表示为 $C(v)$，根据 $C(v)$ 得到频率的无偏估计

$$\Phi_{OUE}(v):=\frac{C(v)/n-q}{p-q} \tag{4}$$

该无偏估计的方差为：

$$Var[\Phi_{OUE}(v)]=\frac{4e^{\varepsilon}}{(e^{\varepsilon}-1)^2\times n} \tag{5}$$

3.2.4 隐私预算分割和随机抽样响应

当每个用户有 k 条信息时，有两种方法可以实现隐私保护。第一种方法是分割隐私预算，其中每个用户在发送 k 条信息时满足 ε/k-LDP。本地差分隐私的组合原理确保整个扰动过程满足 ε-LDP。另一种方法是将用户划分成不同的组，每个组的用户报告 $1/k$ 条信息。每个用户采样 k 条信息中的一条信息，扰动时满足 ε-LDP。通常情况下，划分用户比分割隐私预算表现好，分割隐私预算后，低隐私预算会导致噪音过大，对准确性造成一定影响。使用 $p_1=1/k$ 对 k 条信息进行采样，并使用随机响应扰动采样得到的值。

4 体系结构和问题描述

4.1 体系结构

本地差分隐私下高维数据收集和发布的体系结构如图 1 所示。大量用户从各种传感设备和系统中生成多维数据，经过 LDP 处理后，将扰动后的数据发送到服务器。服务器负责聚合数据并进行必要的处理，发布近似原始数据的合成数据集，并将其提供给第三方研究人员进行数据分析。从高维数据集中挖掘和提取的信息可以提供丰富的社会知识。

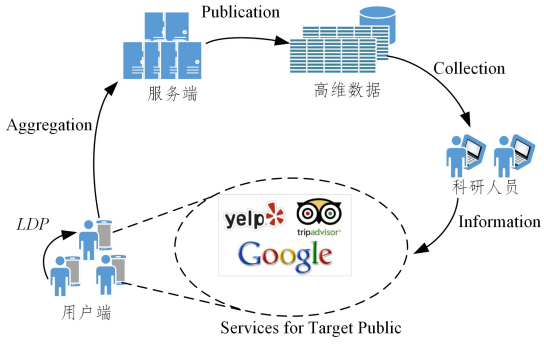


图 1 本地差分隐私下高维数据收集和发布的体系结构

Fig. 1 LDP-based high-dimensional data collection and publication architecture

4.2 问题描述

每个用户都有 d 维属性的数据，经过 LDP 处理后发送给服务器，服务器收集所有用户扰动后的数据，发布一个合成数据集，该数据集具有和原始数据集近似的所有属性的联合分布。N 为用户总个数即数据记录总数，$V=\{V^1,V^2,\cdots,V^N\}$ 表示原始用户的数据集，V^i 表示第 i 个用户的多维数据。假设用户数据有 d 个属性，属性集为 $A=\{A_1,A_2,\cdots,A_d\}$。第 i 个用户的多维数据表示为 $V^i=[v_1^i,v_2^i,\cdots,v_d^i]$，其中 v_j^i 表示第 i 个用户数据的第 j 个属性。对于数据集的属性 A_j（$j=1,2,\cdots,d$），将 A_j 的属性候选集表示为 $\Omega_j=\{\omega_j^1,\omega_j^2,\cdots,\omega_j^{|\Omega_j|}\}$，$\omega_j^i$ 是第 j 个属性的第 i 个可能的取值，$|\Omega_j|$ 是第 j 个属性的候选集大小。表 1 列出了相关符号说明。

表 1 符号说明

Table 1 Notation description

符号	说明
V	原始高维数据集
$\hat{V}$	合成高维数据集
N	用户个数
d	属性个数
V^i	第 i 个用户的数据
A_j	数据集的第 j 个属性
v_j^i	V^i 的第 j 个属性
Ω_j	第 j 个属性候选集
ω_j	Ω_j 的候选值
M_{A_j}	第 j 个属性的边缘分布表
P_A	属性集的联合分布

使用上述符号，问题的定义如下。每个用户独立拥有 d 维的数据，所有用户的数据是一个 $N\times d$ 维度的数据集 V，目标是发布一个和数据集 V 近似的数据集 $\hat{V}$，满足：

$$P_V(A_1\cdots A_d)\approx P_{\hat{V}}(A_1\cdots A_d) \tag{6}$$

其中，$P_V(A_1\cdots A_d)$ 表示 $P_V(v_1^i=\omega_1,\cdots,v_d^i=\omega_d)$，$i=1,\cdots,N$，$\omega_1\in\Omega_1,\cdots,\omega_d\in\Omega_d$。$P_V(v_1^i=\omega_1,\cdots,v_d^i=\omega_d)$ 被定义为数据集 V 上的 d 维属性联合分布。

5 PrivHDP 算法

本文基于现有工作的不足，提出了一个新的解决方案，用于解决基于本地差分隐私的高维数据发布问题。5.1 节介绍了 PrivHDP 算法的总体思想，之后介绍了算法各个组成部分的实现细节。

5.1 PrivHDP 算法概述

算法的目标是基于本地差分隐私发布一个合成的数据集，该数据集与原始的高维数据集具有统计相似性（满足式(6)的统计分布）。

用户首先需要对其个人数据进行本地扰动，以保护其个人信息。一旦扰动后的数据被发送到中央服务器，服务器必须计算原始用户数据的分布，以合成一个保护隐私的数据集。有两种方案可用于计算原始用户数据的分布。一种朴素的方法是直接独立计算每个属性的一维分布。然而，这种方法得到的合成数据集效用较低，因为它未能捕捉属性之间的相关性。另一种方法是服务器可以计算所有属性的 d 维联合分布，从而得到更接近原始数据集的合成数据集。对于一个有 d 维属性的数据集来说，数据输入域值为 $\Omega_1\times\Omega_2\times\cdots\times\Omega_d$，总域值的基数为 $\prod_{j=1}^{d}|\Omega_j|$。域值基数随着维度数量的增加呈指数增长，从而导致低信噪比和不可扩展性。因此，现存的技术挑战是如何在考虑属性相关性的同时尽可能地降低维度。

本文提出的基于本地差分隐私的高维数据发布算法 PrivHDP 分为以下 5 步：

1）本地数据扰动。首先在用户本地对数据进行扰动，扰动机制有 OUE 和 GRR 等，从而保证每个用户的输出满足本地差分隐私，在 5.2 节将介绍具体细节。将扰动后的数据发送到服务器。

2）自适应边缘分布计算。使用两种分布估计方案计算成对属性的边缘分布。计算边缘分布时，基于关键参数对估计误差的影响，形式化定义两种方案的估计方差。在实际计算中自适应地选择边缘分布估计方法和相应的扰动机制，在 5.2 节和 5.4 节将介绍具体细节。

3）属性降维。不同于现有方案使用互信息计算成对属性的相关性，考虑到误差和计算复杂度，使用一种新的相关性度量方法。引入基于高通滤波的阈值过滤技术构建 Markov 网，结合充分三角化操作和联合树算法达到属性分割的目的，生成一组属性子集。具体的降维算法将在 5.3 节中介绍。

4）联合分布计算。将大规模属性子集的联合分布分解为条件概率的乘积，对条件概率中包含的属性进行冗余消除，消除其中的条件独立性。最后通过冗余消除后的条件概率得到更精确的联合分布。

5）数据合成。根据属性子集的联合分布和子集间的相关性，对低维属性数据集进行采样，不断更新数据，得到近似原始数据的合成数据集。

5.2 自适应边缘分布计算

使用傅里叶变换（Fourier Transform，FT）在本地差分隐私下计算成对属性的边缘分布。对于非二值型数据进行编码转换使用傅里叶变换方法时，属性的候选集大小会影响估计的准确性，傅里叶变换方法的效率和可扩展性表现不佳。随后有学者扩展了 PriView 方法[27]，此方法用于在传统的集中式差分隐私下计算边缘分布。基于关键参数对估计误差的影响，形式化定义两种方法的估计方差，在实际计算中自适应地选择估计方案。

5.2.1 傅里叶变换

傅里叶变换计算边缘分布的基本思想是计算 k 维边缘只需要傅里叶域中的系数，用户不用提交边缘分布值，而是提交扰动后的傅里叶系数。

定义 2（傅里叶变换） 用户的输入向量 $\boldsymbol{v}$ 的傅里叶变换形式是 $\boldsymbol{\theta}=\boldsymbol{\varphi v}$，其中 $\boldsymbol{\varphi}$ 是正交对称的 $2^d\times 2^d$ 矩阵，$\varphi_{i,j}=2^{-d/2}(-1)^{\langle i,j\rangle}$。

给定任意向量 $\boldsymbol{v}$，在傅里叶变换下的表示为傅里叶系数 $\boldsymbol{\theta}=\boldsymbol{\varphi v}$。$\boldsymbol{\varphi}$ 是一个傅里叶矩阵，$\boldsymbol{\varphi}$ 中的每一行或列由形式为 $\pm\frac{1}{2^{d/2}}$ 的元素组成，符号由 i,j 的内积决定。$\boldsymbol{v}$ 是经过一元编码处理过的用户输入，只有一个位置为 1，其余位置为 0。

定理 4 傅里叶系数 $H_k=\{\theta_\alpha:|\alpha|\leqslant k\}$ 可以计算任意 k 维边缘分布 m。

$$C^m(v)=\sum_{\alpha\leqslant m}\langle\varphi_\alpha,v\rangle\sum_{\beta;\beta\wedge m=\alpha}\varphi_{\alpha,\beta}=\sum_{\alpha\leqslant m}\theta_\alpha\left(\sum_{\beta;\beta\wedge m=\alpha}\varphi_{\alpha,\beta}\right)\tag{7}$$

其中，C^m 计算由 m 编码的属性值的所有组合的总频率，例如 $m=1001$（编码第一个和第四个属性）。使用$\leqslant$符号，$\alpha\leqslant m$ 当且仅当 $\alpha\wedge m=\alpha$。为了计算 $m=1001$ 编码的边缘，只需要 4 个傅里叶系数，分别是 θ_{0000}，θ_{1000}，θ_{0001}，θ_{1001}。为了计算 k 维边缘，需要 $\sum_{s=1}^{k}\binom{d}{s}$ 个系数。

每个用户对计算边缘分布所需的所有傅里叶系数进行随机采样，根据其输入计算傅里叶变换，随后使用随机响应机制扰动傅里叶系数。只需要 T 个系数即可重构 k 维边缘分布，$|T|=\sum_{s=1}^{k}\binom{d}{s}$。该算法主要分为以下两步。

1）扰动：每个用户采样一个傅里叶系数索引 $i\in T$，基于本地的数据计算第 i 个傅里叶系数 $\theta=(-1)^{\langle j,i\rangle}$，通过随机响应扰动该系数并将其发送给服务器。

2）聚合：服务器端对 $\hat{\theta}$ 进行无偏估计。这些系数通过式(7)被用来重构目标边缘分布 m。

算法 1 和算法 2 详细说明了用户和服务器端的细节。每个用户可以使用 1 位来描述 $RR(\theta)$，最多使用 d 位来描述采样的傅里叶系数索引，降低通信成本。

算法 1 用户端扰动算法

输入：用户数据位置为 1 的索引 j；傅里叶系数 H_k；隐私预算 ε

输出：扰动的傅里叶系数 $\hat{\theta}$ 和采样的系数索引 i

1. 随机采样一个傅里叶系数，索引 $i\in H_k$
2. 计算相应的傅里叶系数 $\theta=(-1)^{\langle j,i\rangle}$
3. 随机响应机制扰动系数 $\hat{\theta}\leftarrow RR(\theta)$
4. return($\hat{\theta}$,i)

算法 2 服务端聚合算法

输入：所有用户的扰动数据($\hat{\theta}$,i)

输出：目标边缘分布表 C^m

1. 根据扰动数据($\hat{\theta}$,i)，$H[i]\leftarrow\hat{\theta}$
2. for all $j\in T$ do
3. $\Theta^*[j]\leftarrow 2^{-d}\frac{\sum_{i=1}^{N}H_{i^*[j]/(2p-1)}}{N_j}$，$N_j$ 是 j 的计数
4. return $C^m\leftarrow$式(7)

以上描述算法是针对二值型数据的情况，考虑将这种方法应用于更一般的输入数据，即属性的候选集大于 2 的情况。假设属性的域值基数为 $|\Omega_1|$，$|\Omega_2|$，…，$|\Omega_d|$，计算 k 维边缘。每个属性域值经过编码处理，得到新的二进制属性，$d_2=\sum_{i=1}^{d}\lceil\log_2|\Omega_i|\rceil$，$k_2=\sum_{i=1}^{k}\lceil\log_2|\Omega_i|\rceil$。

5.2.2 LoView

LoView 的基本思想是策略性地选择一组属性子集，在获得属性子集的边缘分布表后，可以重构任意的 k 维边缘分布。

通过在一个具有 8 个属性的数据集上计算 3 维边缘分布的示例，来说明算法步骤，主要分为以下 4 步。

1）选择属性子集。首先需要选择 n 组属性子集，目的是使得每个 2 维或 3 维边缘分布都被包含在所选的属性子集中。例如，以下 6 组属性子集覆盖了所有的 2 维边缘分布。

$\{A_1,A_2,A_3,A_4\}$，$\{A_1,A_5,A_6,A_7\}$，$\{A_2,A_3,A_5,A_8\}$
$\{A_4,A_6,A_7,A_8\}$，$\{A_2,A_3,A_6,A_7\}$，$\{A_1,A_4,A_5,A_8\}$

2）扰动后的边缘分布。在集中设置中采用的方法是将拉普拉斯噪声添加到真实的边缘分布，计算有噪声的边缘分布。在 LDP 环境下，使用 GRR 和 OUE 扰动本地数据。不同于分割隐私预算，与随机抽样响应的思想一致，将用户群体分成 n 组，每个组中的用户报告 n 个属性子集中的一个边缘分布。

3）一致性步骤。得到带噪声的边缘分布后，可以直接获得部分 3 维边缘分布。例如从 $\{A_1,A_2,A_3,A_4\}$ 的边缘表中边缘化 A_4，获得 $\{A_1,A_2,A_3\}$ 的 3 维边缘分布。计算没有被覆盖的 3 维边缘分布时，需要依赖 n 组属性子集中提供的信息。例如边缘 $\{A_2,A_3\}$ 可以从 $\{A_1,A_2,A_3,A_4\}$ 和 $\{A_2,A_3,A_5,A_8\}$ 中计算得到。由于这两个属性子集是独立扰动后的结果，因此计算 $\{A_2,A_3\}$ 的两个不同路径可能会有不同的结果。对有噪声的边缘分布执行约束性推断，以确保属性子集的边缘分布都是非负的且相互一致。

4）计算 k 维边缘分布。从 n 组一致性处理后的属性子集中，可以重建任何 k 维边缘分布。当给定一组属性时，若给定的属性都被覆盖在一个属性子集中，则可以通过边缘化其他属性直接计算边缘分布。当没有属性子集包含给定属性时，使用最大熵估计方法计算边缘分布。例如，当给定 $\{A_1,A_3\}$ $\{A_1,A_6\}$ $\{A_3,A_6\}$ 的边缘分布时，最大熵估计法在 $\{A_1,A_3,A_6\}$ 所有可能的边缘分布中找到与给定 3 个边缘分布一致的边缘分布，即具有最大熵的边缘分布。

值得特别说明的是扰动机制的选择。每个属性子集的域值基数不同，当出现较大的域值基数时，需要消除对域值基数 $|D|$ 的依赖，因此，GRR 和 OUE 这两种扰动机制的选择非常重要。根据 GRR 和 OUE 的估计方差，$Var[\Phi_{GRR(\varepsilon)}(v)]=\frac{|D|-2+e^\varepsilon}{(e^\varepsilon-1)^2\times n}$，$Var[\Phi_{OUE(\varepsilon)}(v)]=\frac{4e^\varepsilon}{(e^\varepsilon-1)^2\times n}$。比较两个公式，可以看出 GRR 的估计方差依赖域值基数 $|D|$，而 OUE 的方差则是把 $|D|-2+e^\varepsilon$ 替换成 $4e^\varepsilon$。对于 $|D|$ 较小的情况，可以选择 GRR 作为扰动方法。具体来说，$|D|-2+e^\varepsilon<4e^\varepsilon$，即 $|D|<3e^\varepsilon+2$ 的情况下选择 GRR 作为扰动方法更好，否则选择 OUE 作为扰动方法更好。

5.2.3 边缘分布计算

对于二值型的数据集以及属性域值的基数较小的数据集，傅里叶变换方法计算边缘分布能在通信开销较低的情况下获得良好的效用。对于多值型的数据集或者属性域值的基数较大的数据集，LoView 计算边缘分布的误差较小。本文形式化定义了这两种边缘分布计算方法的估计值方差，以期望在实际情况中根据不同的参数选择合适的估计方案。

傅里叶变换方法估计方差：分析计算一维边缘分布的方差，$Var[\Phi_{RRS(\varepsilon)}]=\frac{e^\varepsilon}{(e^\varepsilon-1)^2}$。对于每一个 k 维边缘分布，需要 $T=\sum_{s=1}^{k}\binom{d}{s}$ 系数的估计值。对于每一个傅里叶系数索引，有 N/T 用户报告每个索引。估计值的方差与用于估计它的用户个数成反比。因此，

$$Var_1=\frac{e^\varepsilon}{(e^\varepsilon-1)^2}\times\frac{T^2}{N}\tag{8}$$

计算 k 维边缘分布要枚举所有 α，系数方差乘以 2^k。对于 k 维边缘分布，属性的边缘信息从 $\binom{d}{k}\times d$ 边缘分布中构建。因此，单属性边缘分布的方差是：

$$Var_{\mathrm{FT}}=\frac{Var_1}{\binom{d}{k}\times d}\times 2^k=\frac{e^{\varepsilon}}{(e^{\varepsilon}-1)^2}\times\frac{T^2}{N}\times\frac{2^k}{\binom{d}{k}\times d} \tag{9}$$

LoView 估计方差：对于 n 组属性子集，有 N/n 用户报告其中每个边缘分布。边缘分布值的方差与用于估计它的用户个数成反比。

$$Var_2=\frac{L-2+e^{\varepsilon}}{(e^{\varepsilon}-1)^2}\times\frac{n}{N} \tag{10}$$

L 是一个边缘表中的单元数。当属性域值不同时，L 是 n 中一个属性子集边缘表的单元数。单属性边缘分布通过边缘化边缘表中的其他属性得到，因此 $Var_3=Var_2\times L$。每个属性的边缘分布由 $\frac{n\times k'}{d}$ 边缘信息构建，k' 是一个属性子集中包含的属性个数。

$$Var_{\mathrm{LoView}}=\frac{Var_3}{\frac{n\times k'}{d}}=\frac{L-2+e^{\varepsilon}}{(e^{\varepsilon}-1)^2}\times\frac{L}{k'}\times\frac{d}{N} \tag{11}$$

计算 k 维边缘分布时，边缘分布的估计值受 k 中每个属性的估计误差影响，将 $k\times Var_{\mathrm{LoView}}$ 作为估计方差。

5.3 属性降维

属性维度的增加导致属性的域值基数呈指数增长，因此在数据合成之前降低属性的维度至关重要。在高维数据隐私发布中的一个挑战是高维数据属性之间的关系复杂，简单的线性处理不能反应高维属性间的本质关系。基于概率图模型，属性降维主要分为以下 3 步。

1）计算成对属性相关性。现有的解决方案大多使用互信息度量成对属性间的相关性，计算复杂度高，导致结果精度降低。在此基础上，使用独立差异（Independent Difference，InDif）度量成对属性的相关性。对于任意两个属性 A_1 和 A_2，InDif 计算二维边缘分布 M_{A_1,A_2} 和假设两个属性独立相乘得到的二维边缘分布 $M_{A_1}\times M_{A_2}$ 的 L1 距离。M_A 是属性 A 的边缘分布表，显示属性候选值的频率。

$$InDif_{A_1,A_2}=|M_{A_1,A_2}-M_{A_1}\times M_{A_2}| \tag{12}$$

图 2 给出了 InDif 的计算过程。图 2(a)和图 2(b)是性别和色盲属性的一维边缘分布。图 2(c)通过假设两个属性独立直接计算得到。图 2(d)是真实的二维边缘分布。根据式(12)，$InDif=0.0908$。

gender	M_{gender}
male	0.6
female	0.4

(a)marginal for gender

color vision	$M_{\text{color vision}}$
normal	0.917
color blind	0.083

(b)marginal for vision

⟨gender, color vision⟩	$M_{\text{(gender,color vision)}}$
⟨male,normal⟩	0.550 2
⟨male,color blind⟩	0.049 8
⟨female,normal⟩	0.336 8
⟨female,color blind⟩	0.033 2

(c)2-dimensional marginal assume indepent

⟨gender, color vision⟩	$M_{\text{(gender,color vision)}}$
⟨male,normal⟩	0.535
⟨male,color blind⟩	0.065
⟨female,normal⟩	0.382
⟨female,color blind⟩	0.018

(d)actual 2-dimensional marginal

图 2 InDif 计算示例

Fig. 2 Example of InDif calculation

2）构建概率图模型。Markov 网用来表示属性间的相关性。每个属性 A_j 是关系图中一个节点，两个节点 A_j 和 A_i 之间存在一条边表示这两个属性相关。基于任意两个属性间的相关性，可以构造所有属性的概率图模型。首先，初始化邻接矩阵 $C_{d\times d}$ 中所有数值为 0。然后选择所有属性对 (A_j,A_i) 计算属性间的相关性 $InDif_{A_i,A_j}$，和阈值 $\mu_{i,j}$ 进行比较。

$$\mu_{i,j}=\min(|\Omega_i|-1,|\Omega_j|-1)\times\phi^2/2 \tag{13}$$

其中，$\phi(0\leqslant\phi\leqslant1)$ 是确定相关性的参数，实验中设置 ϕ 为 0.6。如果 $InDif_{A_i,A_j}\leqslant\mu_{i,j}$，对于较小的 $\mu_{i,j}$，认为 A_j 和 A_i 是相互独立的。如果 $InDif_{A_i,A_j}>\mu_{i,j}$，则设置 $C_{i,j}$ 和 $C_{j,i}$ 的值为 1。

3）分割属性集。通过三角化操作，依赖关系图 $C_{d\times d}$ 可以被转化为联合树。通过联合树算法，分割属性集后得到一组属性子集 $(C_1,C_2,\cdots,C_m)$，每个属性子集包含相关度较高的几个属性。属性子集内的相关度高，子集间相关性低。由于属性子集间的低相关性，低维的属性子集可以单独处理。整个过程如算法 3 所示。

算法 3 属性降维算法

输入：属性 j 的候选集 $\Omega_j(1\leqslant j\leqslant d)$；属性 j 的边缘分布表 $M_{A_j}(1\leqslant j\leqslant d)$；成对属性 i 和 j 的边缘分布表 $M_{A_i,A_j}(1\leqslant i\leqslant d,1\leqslant j\leqslant d)$；相关性参数 ϕ

输出：一组属性子集 $(C_1,C_2,\cdots,C_m)$

1. 初始化 $C_{d\times d}\leftarrow0$
2. for i=1 to d−1 do
3. for j=i+1 to d do
4. $\mathrm{InDif}_{A_i,A_j}\leftarrow|M_{A_i,A_j}-M_{A_i}\times M_{A_j}|$
5. $\mu_{i,j}\leftarrow\min(|\Omega_i|-1,|\Omega_j|-1)\times\phi^2/2$
6. if $\mathrm{InDif}_{A_i,A_j}>\mu_{i,j}$ then
7. $C_{i,j},C_{j,i}\leftarrow1$
8. endif
9. endfor
10. endfor
11. 根据 $C_{d\times d}$ 构建 Markov 网
12. 三角化操作和联合树算法得到一组属性子集
13. return $C_1,C_2,\cdots,C_m$

5.4 联合分布计算

属性降维后得到一组属性子集，合成数据集需要这些属性子集上的联合分布。小规模的属性子集的联合分布计算相对简单，大规模的属性子集的联合分布计算仍然存在很大问题。

为解决这个问题，基于联合分布的分解和冗余消除计算大规模的属性子集上的联合分布。给定一个属性子集 C 包含 d 个属性 $\{A_1,\cdots,A_d\}$，首先将联合分布分解为多个条件概率的乘积。

$$Pr(A_1,\cdots,A_d)=\prod_{i=1}^{d}Pr(A_i|A_1,\cdots,A_{i-1})$$

其中，$A_1,\cdots,A_{i-1}$ 表示 A_i 的条件集。A_i 的条件集中可能存在冗余，即给定条件集的一些属性，条件集中的其他属性和 A_i 是条件独立的。因此，消除条件集中的冗余可以得到一个新的规模较小的条件集。具体过程如算法 4 所示。

首先初始化属性候选集 Q 为 $\{A_1,\cdots,A_d\}$，集合 S 和 M 为空。S 和 M 分别用来存储因子分解的结果和计算联合

分布用到的因子。采用前向选择策略找到近似的分解结果，即首先确定因子分解的最后一项 $Pr(A_d|A_1,\cdots,A_{d-1})$。该策略共有 d 轮迭代，根据属性集合域值大小 $|Q|=\prod_{A_j\in Q}|\Omega_{A_j}|$，每轮迭代分为两种情况。

1) $|Q|\geqslant\sigma$，将属性集合中的每个属性 A_j 看作目标属性，使用最小冗余最大相关性特征选择方法消除冗余，冗余消除结果表示为 E_{A_j}。最后得到 $|Q|$ 项结果 $\{E_{A_1},\cdots,E_{A_{|Q|}}\}$。选择使估计方差最小的属性 A_h 和其对应的冗余消除结果 E_{A_h}。

2) $|Q|<\sigma$，随机选择属性集合中的一个属性 A_h，$E_{A_h}=Q\backslash A_h$。

将 (A_h,E_{A_h}) 添加进集合 S 和 M，从属性候选集 Q 中删除 A_h。重复上述过程，直到属性候选集 Q 为空。最后计算集合 M 中的分布，乘以 S 中的因子分解结果，得到整个属性子集上的联合分布。

算法 4 联合分布估计算法

输入：属性子集 C={A1,…,Ad}；属性相关性矩阵 InDif；隐私预算 ε；属性子集大小阈值 σ

输出：属性子集 C 的联合分布 Pr(C)

```
1. 初始化 Q={A1,…,Ad},S←∅,M←∅
2. for i=d to 1 do
3.     if |Q|<σ
4.         从 Q 中随机选择一个属性 Ah,E_Ah=Q\Ah
5.     else
6.         for Aj∈Q do
7.             E_Aj←RedundancyEliminate(Aj,Q\Ah,InDif)
8.             Var(Aj)←|E_Aj ∪ Aj|
9.         选择估计方差最小的属性 Ah
10.        将(Ah,E_Ah)添加到 S 和 M 中
11.        从 Q 中删除 Ah
12. 估计 M 中的分布
13. Pr(C)=∏_{(Ah,E_Ah)∈S} Pr(Ah|E_Ah),Pr(Ah|E_Ah)从 Pr(Ah,E_Ah)计算得到
14. return Pr(C)
```

5.5 数据合成

基于属性子集和属性子集的联合分布生成合成数据集。首先初始化集合 A，以保存所有属性的索引。随机选择一个属性子集 C_i 和子集的联合分布，从属性子集中的属性 A_j，$A_j\in C_i$ 采样新数据 $\hat{V}_C$。然后从属性子集的集合中将属性子集 C_i 删除，并将 C_i 中的属性包含到集合 A 中，找到属性子集 C_i 的相邻属性子集 D，即相交的属性子集。在相邻属性的属性子集中，每个属性子集的遍历和采样如下。首先获得属性子集 D 上的联合分布和条件分布 $P(A_{D-A}|A_{D\cap A})$，通过该条件分布和采样数据 $\hat{V}_{D\cap A}$ 对 $\hat{V}_{D-A}$ 进行采样。遍历 D 后，对第一个相邻属性子集的属性进行采样。然后在剩余的属性子集的集合中随机选择一个属性子集，并对第二个相邻属性子集中的属性进行采样，直到集合为空。最后，根据原始数据的关联性和联合分布，生成一个新的数据集 $\hat{V}$。

5.6 整体工作流程

PrivHDP 算法的整体工作流程如算法 5 所示。在第一阶段使用分配系数 ω 将用户分为 n_1 和 n_2 两组，n_1 用于计算边缘分布构造概率图模型，n_2 用来估计属性被分割后所有属性子集的联合分布，实验中设置 $\omega=0.5$。第二阶段通过算法 3 构造 Markov 网，生成由多个属性子集组成的联合树。第三阶段分别估计每个属性子集的联合分布。最后，根据联合树的结构和所有属性子集的联合分布，生成合成数据集。

算法 5 PrivHDP 算法

输入：用户个数 N；隐私预算 ε；分配系数 ω

输出：合成数据集 $\hat{V}$

```
1. n1←ω·N,n1←(1−ω)·N
2. 自适应边缘分布估计算法通过 n1 用户
3. 一组属性子集 C(C1,C2,…,Cm)←属性降维算法
4. for small Ci∈C do
5.     Pr(Ci)←边缘分布计算结果
6. for large Ci∈C do
7.     Pr(Ci)←联合分布估计算法通过 n2 用户
8. 生成合成数据集 V̂ 通过数据合成算法
9. return V̂
```

5.7 隐私保证和复杂性分析

5.7.1 隐私保证

算法流程中，每个用户只发送一次本地数据且满足 ε-LDP，因此整个解决方案满足 ε-LDP。

定理 5 本地差分隐私下的高维数据发布算法满足 ε-LDP。

5.7.2 复杂性分析

通过讨论解决方案中关键步骤的时间复杂性来分析整体解决方案的时间复杂性。首先假设：1)用于构造概率图模型的用户数为 n_1；2)用于估计属性子集联合分布的用户数为 n_2；3)属性域值 $|\Omega_{A_j}|$ 的平均大小表示为 r；4)联合分布计算时属性子集中的属性个数为 $\hat{d}$。

定理 6 自适应边缘分布计算最坏情况下的时间复杂性为 $\mathrm{Max}(O(n_1+\binom{d}{k}\cdot 2^{2k}),O(n_1+m\cdot 2^L))$。

证明：傅里叶变换方法使用随机响应扰动用户数据，处理时间是 n_1 的数量级，但是计算式(9)枚举所有 α 花费 $O(2^k)$。构造 $\binom{d}{k}$ 边缘分布表，每个条目为 2^k，总计算时间为 $O(n_1+\binom{d}{k}\cdot 2^{2k})$。LoView 主要处理用户的报告，每次处理时间为 $O(2^L)$。每个用户的报告是一个值，只需要聚合扰动数据，运行时间为 $O(n_1+m\cdot 2^L)$。

定理 7 属性降维最坏情况下的时间复杂性为 $O(d^2\cdot r^2)$。

证明：根据算法 3，每次循环的时间开销主要由两部分组成：1)计算成对属性的相关性；2)计算属性相关性阈值。对应的时间开销分别是 $O(r^2)$ 和 $O(1)$。算法最多循环 $d\cdot(d-1)/2$ 次，最坏情况下的时间复杂度为 $O(d^2\cdot r^2)$。

定理 8 计算属性子集的联合分布时间复杂性为 $O(\hat{d}^2+n_2\cdot\sum_{i=1}^{\hat{d}}r^i+\hat{d}r^{\hat{d}})$。

证明：根据算法 4，估计属性子集的联合分布的时间开销

主要由 3 部分组成：1）确定因子分解结果；2）估计因子分解结果中每个因子的分布；3）计算属性子集的联合分布。对应的时间开销分别是 $O(\hat{d}^2)$，$O(n_2 \cdot \sum_{i=1}^{\hat{d}} r^i)$ 和 $O(\hat{d}r^{\hat{d}})$。最坏情况下联合分布计算的时间复杂性为 $O(\hat{d}^2 + n_2 \cdot \sum_{i=1}^{\hat{d}} r^i + \hat{d}r^{\hat{d}})$。

实际计算中，通过采用冗余消除方法，因子分解结果的大小可以显著减小，复杂性将显著降低。

6 实验评估

6.1 实验环境

实验运行环境为 Windows 10 64 位操作系统，Intel(R) Core(TM) i7-7500U CPU @2.70 GHz 2.90GHz，24.0 GB 内存，Python 3.7。

6.2 数据集

实验所采用的 4 个数据集 POS，Adult，US Census，TPC-E 均被广泛用于高维数据发布。表 2 列出了每个数据集的具体细节。

POS 是包含 50 万用户的商户交易数据集的一部分。Adult 是从 1994 年美国人口普查中提取的，包含个人信息，如性别、工资和教育水平等属性。US Census 包含 1990 年人口普查的个人记录。我们将数据集预处理为 30 个分类属性。TPC-E 是一个混合类型数据集，来源于 TPC 的在线事务处理程序，其中包含 TPC-E 基准中“交易类型”“安全”和“安全状态”等信息的交易记录。Adult 和 TPC-E 数据集都包含连续和分类属性，预处理中将连续属性离散化为分类属性。对于每个连续属性，将其属性域划分为固定数量的等宽区间，对一些连续域进行合并简化处理。

表 2 4 种数据集的信息描述

Table 2 Descriptions of four datasets

Datasets	Datatype	Records	Attributes	Domain size
POS	Binary	515 597	16	2^{16}
Adult	Integer	45 222	15	2^{52}
US Census	Integer	185 788	30	2^{45}
TPC-E	Mixed	40 000	24	2^{77}

6.3 评价指标

为了评估不同方法的性能，针对两个目标，分别使用 k-way 查询和 SVM 分类度量 PrivHDP 算法发布高维数据的可用性与准确性。k-way 查询中 k 的取值在属性个数不同时分别取 2 和 6，使用误差平方和（Sum of Squared Errors，SSE）度量结果的准确性。其次，在合成数据集上训练 SVM 分类模型并进行预测。使用正确分类率度量分类结果的准确性。对于每项任务，重复实验 20 次并报告平均值。

6.4 竞争算法

列联表法（Full Contingency Table Method，FC）：一种简单的方法，直接构建完整的列联表。一旦完整的列联表可用，就可以计算任何 k 维边缘分布。然而，由于必须查询所有属性的每个值在整个属性域中的频率，因此时间复杂度和空间复杂度随着 d 的增加呈指数增长。

最大期望算法（Expectation Maximization Method，EM）：该方法允许每个用户单独上传每个属性的值，并分割隐私预算。聚合器执行最大期望算法来重建边缘分布表。

基准算法（UNI）：绘制了基准算法的误差平方和，该算法返回所有边缘分布的平均分布。如果某一方法表现不如基准方法，则说明该方法构造的边缘分布可用性较低。

LoPub：基于最大期望算法和 Lasso 回归计算多维联合分布，属性降维时使用互信息计算成对属性的相关性。

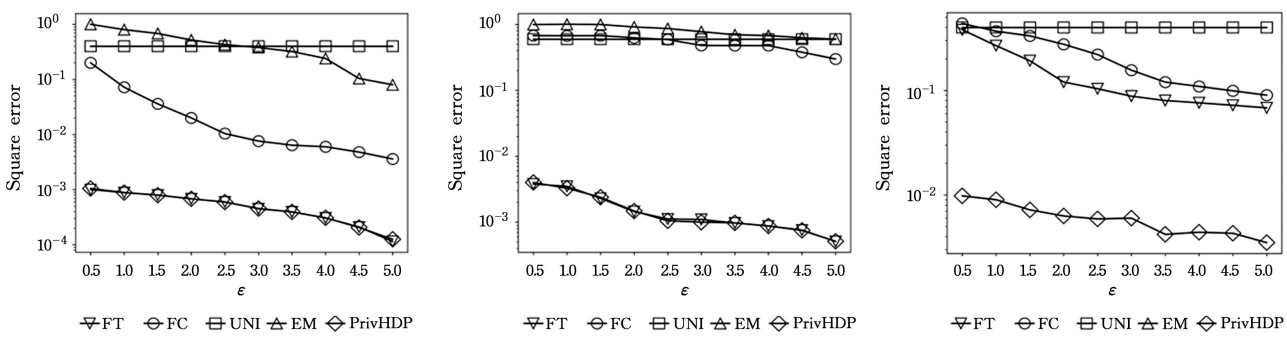
LoPub with InDif：在 LoPub 算法中使用 InDif 代替互信息衡量属性对之间的相关性，有助于了解使用 InDif 的效果，是 LoPub 的改进版本。

6.5 实验结果与分析

6.5.1 二值数据集 k-way 查询

在第一组实验中，在二进制数据集上比较了 PrivHDP 算法与 FT，FC 和 EM 等现有方法的性能。

从图 3 可以看出，在所有设置中，PrivHDP 算法总体上优于现有方法。并且对于较大的 d 和 k 值，自适应边缘分布计算具有明显优势。在 $d=8, k=2$ 和 $d=16, k=2$ 的情况下，所提算法的性能与 FT 算法几乎相同，对于二进制数据集，自适应边缘分布计算使用 FT 计算边缘分布。然而，随着 d 和 k 值的增大，傅里叶变换算法的估计方差增加，PrivHDP 方法与 FT 和 FC 方法之间的误差相差 1～2 个数量级。当隐私预算较小时，所有算法的 SSE 接近 UNI，这表明当隐私预算较小时扰动较大，所有算法提供的信息都较少。尽管如此，PrivHDP 方法在这种情况下仍然表现最佳。



(a) POS，$d=8, k=2$ (b) POS，$d=16, k=2$ (c) POS，$d=16, k=6$

图 3 POS 数据集上 k-way 查询误差分析

Fig. 3 Error analysis of k-way query on POS dataset

与其他方法相比，EM 表现不佳。其中一个原因是 EM 要求用户报告其所有属性的信息，这导致数据可用性降低，因为需要为每个属性分割隐私预算。此外，当 k 较大时，EM 算法的推断时间过长，降低了其运行效率。因此，我们在 $k=6$ 的实验中不考虑 EM 算法的性能。

在 $d=8, k=2$ 的情况下对这几种算法进行理论分析。为了进行比较，粗略地假设 V_0 为每个估计值的方差。在这种设定下，FT 和 FC 的估计方差为 $\sum_{s=1}^{k}\binom{d}{s}\times V_0$ 和 $2^d\times V_0$。$d=8$，$k=2$ 时，两种算法的估计方差分别为 $36V_0$ 和 $256V_0$。图中的实验结果与理论分析相符。$d=16$ 时，其他方法的误差明显增大。其中 FC 方法的误差大小只取决于 d 的大小，因为它构建了一个完整的列联表。当 d 更大时，现有的算法都无法以很好的效率扩展。

6.5.2 多值数据集 k-way 查询

在第二组实验中，在非二进制数据集上比较了 PrivHDP 算法与现有方法的性能。

从图 4 中两个数据集的不同设置可以看出，PrivHDP 方法的表现要明显优于其他方法。实验设置中，为了减小计算复杂度，对连续属性进行预处理。从图中可以看到，在二进制数据集上，FT 算法的表现明显优于 FC 算法。而在非二值型数据集上，FT 的表现较差，因为属性的编码过程不仅增加了计算复杂度，也让计算边缘分布所需的傅里叶系数大大增加。当 $d=8, k=2$ 时，假设每个属性都拥有 4 个取值的情况，实际上，真实数据集中属性的取值要远大于 4。经过编码之后，$d=16, k=4$ 时，通过前面的方差分析，估计方差为 $2516V_0$，这比 $36V_0$ 要大得多。当 $d=16, k=6$ 时，FT 算法时间花费太多，且结果已经不具有可用性，在此情况下我们不绘制 FT 算法。

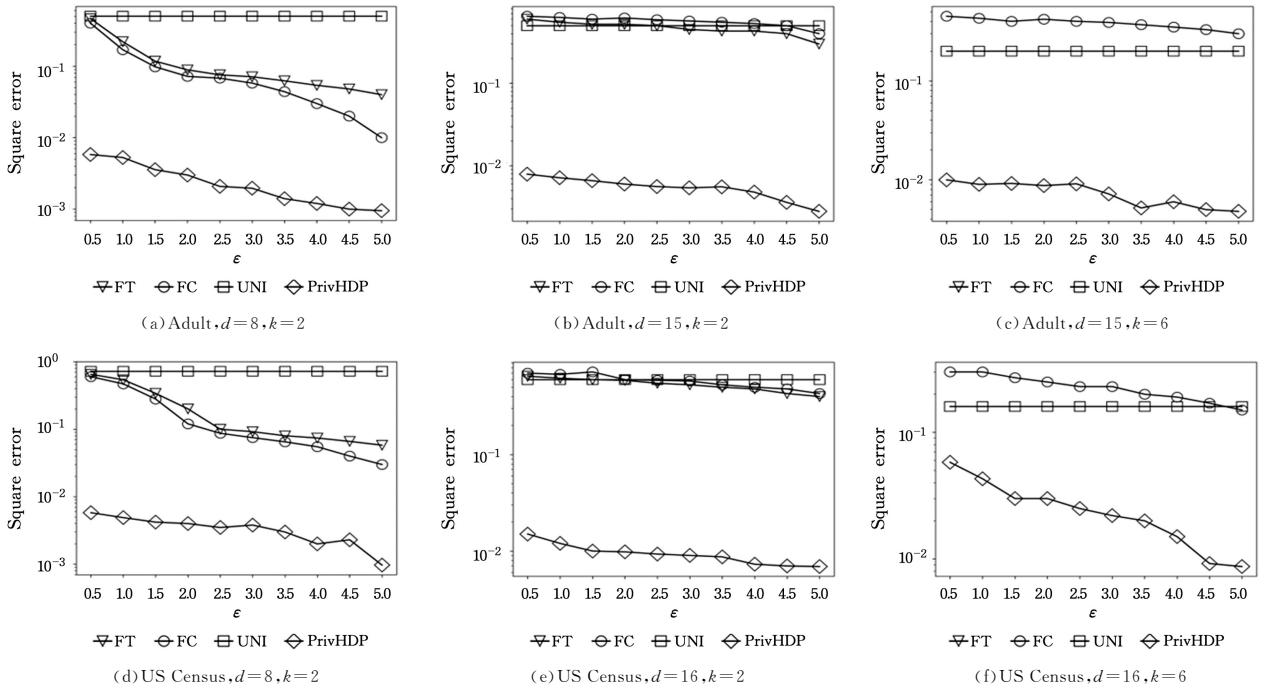


图 4 Adult 和 US Census 数据集上 k-way 查询误差分析

Fig. 4 Error analysis of k-way query on Adult and US Census datasets

6.5.3 SVM 分类结果

在第三组实验中，使用 SVM 分类率在 3 个数据集上比较了 PrivHDP 算法与现有方法发布的高维数据集的可用性。

为了评估算法的整体性能，选择一个典型的下游任务来衡量发布的高维数据集的效用。首先从 3 个原始数据集中获取训练集和测试集。在 3 个原始数据集上进行实验，获得合成数据集，然后在合成的数据集上训练 SVM 分类器模型。计算数据集中所有二进制属性的平均 SVM 分类准确度。NoNoise 方法表示不执行本地差分隐私的方法，在原始数据集上训练 SVM 分类模型进行预测，这是最佳目标(使用相同的属性集)。其中 80%的数据为训练集，20%为测试集，通过测试集上的正确分类率来评估其效用。同时，为了更好地评估方法的性能，我们通过从数据集的真实分布中进行采样来增加数据集的大小。

图 5 给出了 PrivHDP 算法和现有算法在 3 个不同的数据集上的表现情况。总体而言，算法在这 3 个数据集上的表现优于现有方法。随着隐私预算 ε 的增加，算法的性能提高，这与实验结果一致。通过从真实分布中进行采样来增加数据集的大小，可以看出更多的数据可以提高结果的准确性。此外，该算法在二值型数据集上的表现优于非二值型数据集，因为非二进制属性有更多的取值，域值基数大，其属性可能值的分布比二值型数据集的分布更稀疏，联合分布估计可能存在更大的偏差。PrivHDP 算法与 LoPub 算法在 TPC-E 数据集上存在明显差异，因为混合型数据集 TPC-E 存在属性间的相关性不是特别明显的情况，导致生成了规模较大的属性子集。PrivHDP 算法的降维处理过程可以有效减少误差，生成更紧密的属性子集。面对大规模属性子集时，PrivHDP 算法中的联合分布计算可以得到更准确的联合分布。

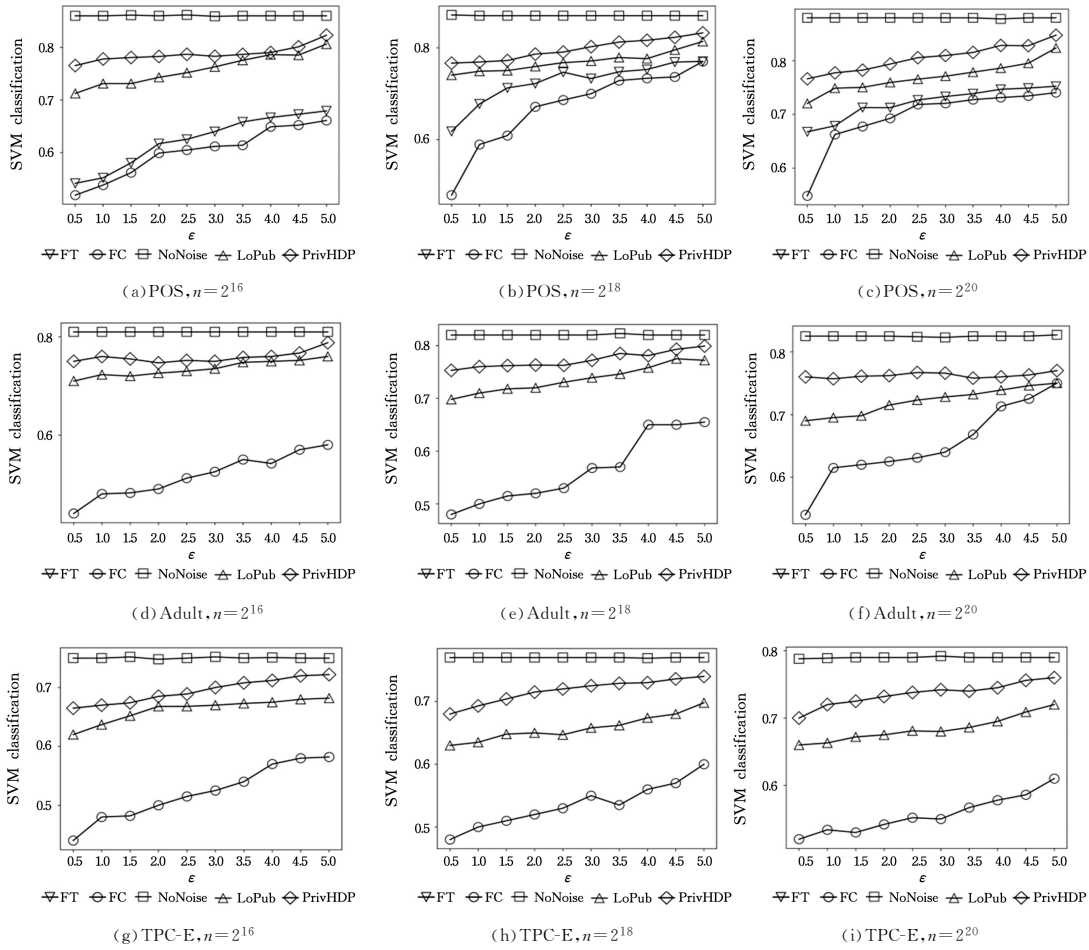


图 5 3 种数据集下 SVM 的分类结果

Fig. 5 SVM classification results on three datasets

6.5.4 时间开销

在最后一组实验中，评估了 PrivHDP 算法与 LoPub 和 LoPub with InDif 在 4 个数据集上的运行时间。

图 6 给出了算法与 LoPub 方法及改进的 LoPub 方法在 4 个数据集上的运行时间，固定隐私预算 $\varepsilon=4$。LoPub 使用互信息衡量成对属性的相关性，导致在构建的依赖图中存在一些大的属性簇，增加了计算成本，且后续估计属性簇的联合分布时仍然面临高维的问题。为实现合成数据集的可用性，在使用互信息计算时，设置相关性参数 $\phi=0.3$。使用 InDif 计算时，设置相关性参数 $\phi=0.6$。实验结果表明，使用 InDif 作为度量方法可以有效缩短计算时间。PrivHDP 算法与 LoPub 相比，在处理高维数据时更加高效和可扩展。

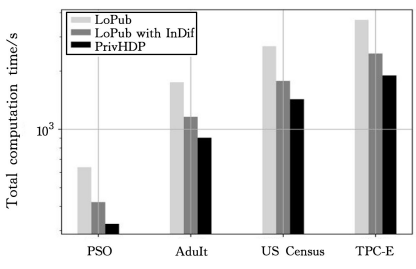


图 6 4 种数据集上的时间开销

Fig. 6 Time overhead on four datasets

结束语 针对本地差分隐私下高维数据发布问题，首先对现有解决方案进行分析，针对现有方法的不足，提出了一种高效的本地差分隐私下高维数据发布方法。使用随机采样响应代替传统的隐私预算分割策略扰动用户数据，提出自适应边缘分布计算方法。使用独立差异代替互信息度量成对属性间的相关性，引入基于高通滤波的阈值过滤技术缩减概率图构建过程的搜索空间，结合充分的三角化操作获得属性子集。最后基于联合分布分解和冗余消除，计算属性子集的联合分布。实验对比分析结果验证了 PrivHDP 算法发布高维数据的可用性与高效性。

未来计划进一步探索两个方面。首先，考虑多个属性组合中低频值的问题。大的域值基数导致信噪比低，组合频率与属性的域值基数成反比，使得估计低频值的频率更加困难。考虑的解决方案是使用没有缺失值的元组推断缺失值，并使用填充的元组基于学习的先验知识获得更准确的分布信息。其次，计划设计更有效的属性降维方法，以提高发布的准确性。未来的工作将集中在生成更紧密的属性子集和高效的数据合成过程上。考虑到属性的重要程度可能会有先后，还未有对属性重要程度的统一的度量标准，在降维处理时考虑属性的特性是未来需要思考的工作。此外，工作需要处理每个属性的候选集为数值类型。将方案扩展到其他属性类型也是未来重要的工作。

参 考 文 献

[1] SHEN H, ZHANG M W, SHEN J. Efficient privacy-preserving cube-data aggregation scheme for smart grids[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(6): 1369-1381.

[2] NAVEED M, AYDAY E, CLAYTON E W, etal. Privacy in the genomic era [J]. ACM Computing Surveys (CSUR), 2015, 48(1): 1-44.

[3] KHAN F, REHMAN A U, ZHENG J, et al. Mobile crowdsensing: A survey on privacy-preservation, task management, assignment models, and incentives mechanisms[J]. Future Generation Computer Systems, 2019, 100: 456-472.

[4] DWORK C. Differential privacy[C]// International colloquium on automata, languages, and programming. Berlin: Springer, 2006: 1-12.

[5] DWORK C, ROTH A. The algorithmic foundations ofdifferential privacy[J]. Foundations and Trends © in Theoretical Computer Science, 2014, 9(3/4): 211-407.

[6] 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. IEEE, 2013: 429-438.

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

[8] ZHANG J, CORMODE G, PROCOPIUC C M, et al. Privbayes: Private data release via bayesian networks[J]. ACM Transactions on Database Systems(TODS), 2017, 42(4): 1-41.

[9] CHEN R, XIAO Q, ZHANG Y, et al. Differentially private high-dimensional data publication via sampling-based inference[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 129-138.

[10] CHEN X, LIU J, FENG X Q, et al. Differentially Private Synthetic Dataset Releasing Algorithm Based on Naive Bayes[J]. Computer Science, 2015, 42(1): 236-238.

[11] ZHANG Z K, WANG T H, LI N H, et al. PrivSyn: Differentially Private Data Synthesis[C]// 30th USENIX Security Symposium (USENIX Security 21). 2021: 929-946.

[12] ZHANG X J, CHEN L, JIN K Z, et al. Private High-Dimensional Data Publication with Junction Tree [J]. Computer Research and Development, 2018, 55(12): 2794-2809.

[13] LI H, XIONG L, JIANG X. Differentially private synthesization of multi-dimensional data using copula functions[C]// Advances in Database Technology: Proceedings. International Conference on Extending Database Technology. NIH Public Access, 2014.

[14] CAI K T, LEI X Y, WEI J X, et al. Data Synthesis via Differentially Private Markov Random Fields[C]// Proceedings of the VLDB Endowment. New York: ACM, 2021: 2190-2202.

[15] GIUSEPPE V, GRACE T, MARK B, et al. New oracle-efficient algorithms for private synthetic data release[C]// International Conference on Machine Learning(ICML). Vienna: ACM, 2020: 9765-9774.

[16] 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. 2016: 41-61.

[17] REN X B, YU C M, YU W, et al. High-dimensional crowdsourced data distribution estimation with local privacy[C]// 2016 IEEE International Conference on Computer and Information Technology(CIT). IEEE, 2016: 226-233.

[18] CORMODE G, KULKARNI T, SRIVASTAVA D. Marginal release under local differential privacy[C]// Proceedings of the 2018 International Conference on Management of Data. 2018: 131-146.

[19] WANG T, YANG X Y, REN X B, et al. Locally private high-dimensional crowdsourced data release based on copula functions [J]. IEEE Transactions on Services Computing, 2019, 15(2): 778-792.

[20] DU L K, ZHANG Z K, BAI S J, et al. AHEAD: adaptive hierarchical decomposition for range query under local differential privacy[C]// Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021: 1266-1288.

[21] KULKARNI T. Answering range queries under local differential privacy[C]// Proceedings of the 2019 International Conference on Management of Data. 2019: 1832-1834.

[22] LIU L K, ZHOU C L. RCP: Mean Value Protection Technology Under Local Differential Privacy[J]. Computer Science, 2023, 50(2): 333-345.

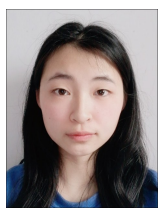
[23] WANG N, XIAO X K, YANG Y, et al. Collecting and analyzing multidimensional data with local differential privacy[C]// 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, 2019: 638-649.

[24] 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.

[25] KAIROUZ P, BONAWITZ K, RAMAGE D. Discrete distribution estimation under local privacy[C]// International Conference on Machine Learning. PMLR, 2016: 2436-2444.

[26] WANG T H, BLOCKI J, LI N H, et al. Locally differentially private protocols for frequency estimation[C]// 26th USENIX Security Symposium(USENIX Security 17). 2017: 729-745.

[27] QARDAJI W, YANG W N, LI N H. Priview: practical differentially private release of marginal contingency tables[C]// Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. 2014: 1435-1446.



CAI Mengnan, born in 2000, postgraduate. Her main research interests include privacy protection and deep learning.

SHEN Guohua, born in 1976, Ph.D, associate professor. His main research interests include software engineering, privacy protection, semantic web and description logic.

(责任编辑:喻藜)