中国科学: 信息科学 2019 年 第 49 卷 第 12 期: 1586–1605
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
信息系统在复杂应用中的科学与技术专刊 · 论文

基于 Bayes 网络的高维感知数据本地隐私保护发布

任雪斌[1], 徐静怡[1], 杨新宇[1*], 杨树森[2]

1. 西安交通大学计算机科学与技术学院, 西安 710049
2. 西安交通大学数学与统计学院, 西安 710049
* 通信作者. E-mail: yxyphd@mail.xjtu.edu.cn

收稿日期: 2019–06–04; 修回日期: 2019–07–29; 接受日期: 2019–08–26; 网络出版日期: 2019–12–16

国家自然科学基金 (批准号: 61572398, 61772410, 61802298, U1811461, 11690011)、国家重点研发计划 (批准号: 2017YFB 1010004)、中央高校基本科研业务费 (批准号: xjj2018237) 和中国博士后基金 (批准号: 2017M623177) 资助项目

摘要 群智感知系统通过对高维感知数据的发布和分析为人们带来巨大数据价值的同时, 也给参与者的隐私带来了极大的隐患. 目前, 各种基于差分隐私的隐私保护方法被提出, 但大部分方法不能同时解决高维感知数据间复杂的属性关联问题和来自不可信服务器的隐私威胁问题. 基于此, 本文提出了基于 Bayes 网络的高维感知数据本地隐私保护发布机制. 该机制实现了用户端的本地数据保护, 杜绝了其他方直接访问用户原始数据的可能, 根本上保护了用户的数据隐私. 感知服务器端在接收到用户本地隐私保护的数据后, 基于 Bayes 网络方法对高维数据的维度相关性进行识别, 将高维数据属性集划分为多个相对独立的低维属性集, 进而依次合成新的数据集, 可以有效地保留原始感知数据的属性维度相关性, 保证合成数据集与原始数据集具有尽可能相似的统计特性. 通过大量仿真实验验证了该方法的有效性, 实验结果表明该方法在有效的本地隐私保护下的合成数据具有较高的数据效用性.

关键词 群智感知系统, 感知数据, 高维数据, 本地差分隐私, Bayes 网络

1 引言

随着设备制造工艺、通信技术、数据处理、算法设计等诸多方面的快速发展, 物联网技术逐渐兴起, 群智感知 [1,2] 作为实现物理世界感知到信息价值服务的一种重要桥梁, 也应运而生. 如图 1 所示, 群智感知系统通过散布于广泛空间内的用户所携带的各种智能设备, 实现跨时空的物理世界感知和数字化, 以达成大规模群智数据的获取 [3]. 群智感知数据然后经感知服务器发布给第三方用户进行各类分析挖掘和机器学习实现数据价值的获取, 最终对社会生产生活提供精确的信息反馈和决策指导 [4]. 群智感知数据除规模大之外, 由海量异质感知设备得到的感知数据往往还具有多维 (多个属性维度)

引用格式: 任雪斌, 徐静怡, 杨新宇, 等. 基于 Bayes 网络的高维感知数据本地隐私保护发布. 中国科学: 信息科学, 2019, 49: 1586–1605, doi: 10.1360/SSI-2019-0119
Ren X B, Xu J Y, Yang X Y, et al. Bayesian network-based high-dimensional crowdsourced data publication with local differential privacy (in Chinese). Sci Sin Inform, 2019, 49: 1586–1605, doi: 10.1360/SSI-2019-0119

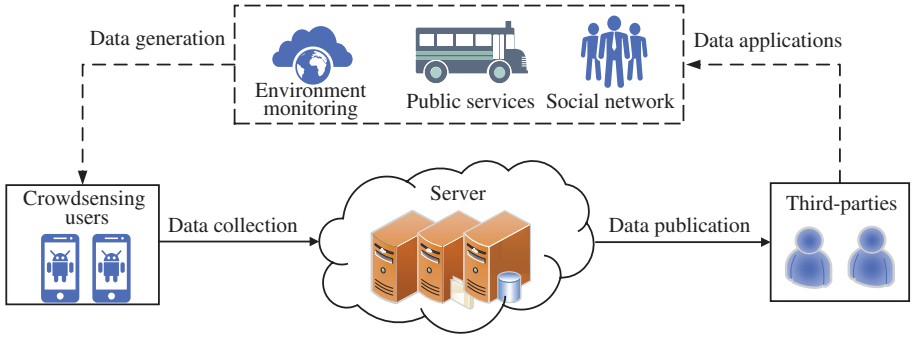


图 1 (网络版彩图) 群智感知系统示意图

Figure 1 (Color online) A concept figure of crowd sensing systems

甚至高维的特性. 对其中属性维度间的相关性挖掘是实现群智感知数据价值的重要方式. 例如, 通过对患者健康记录中不同身体特征的相关性分析, 可以发现或预测其潜在的患病风险 [5]; 通过对手机用户的购物、浏览行为进行关联分析, 可以实现个性化的智能推荐系统 [6].

群智感知数据中通常包含了感知用户自身及其环境信息 (如 GPS 数据) 和日常行为信息 (如计步数据) 等敏感信息. 这些敏感信息如果被超出感知目的地滥用或发布, 或者在数据产生到消亡的生命周期内无法得到有效保护, 都可能会造成感知用户隐私的暴露 [7~9], 给他们带来广告滋扰, 经济利益受损, 甚至是人身安全的威胁. 因此, 对群智感知数据的隐私保护尤为重要, 并已受到业界和学界的广泛关注 [10]. 目前, 常用的隐私保护技术包括基于匿名的方法 [11] (如 K- 匿名, L- 多样性和 T- 邻近等) 和基于加密的方法 [12,13] (如同态加密, 秘密共享, 安全多方计算等). 然而, 基于匿名的方法通常缺乏严格的隐私安全保证, 仅适用于小规模数据的隐私保护 [14]. 基于加密的方法虽然具有较好的安全保证, 但是加密操作会带来较大的计算开销, 难以适用于资源受限的感知设备 [15]. 近年来, 差分隐私的概念被提出 [16], 并由于其严谨的数学定义和灵活的组合特性, 受到学界和业界的广泛关注, 已成为隐私保护的一种事实标准. 由于其轻量性, 差分隐私特别适合于群智感知数据等大数据的处理和分析场景.

然而, 将差分隐私应用于群智感知系统中的高维感知数据发布, 仍然面临以下两个研究挑战.

挑战一: 非本地隐私保护. 现有的隐私保护研究大多着眼于对已收集的数据的处理, 而不考虑数据获取过程中的隐私暴露风险, 而且暗含着数据服务器是隐私安全的这一强假设. 实际上, 虽然现有端到端加密技术可以保证通信过程感知数据不被窃取, 中心化差分隐私技术可以防止第三方用户从发布数据中进行差异攻击和推测攻击获取原始感知数据, 但是, 最终存放在感知服务器上的仍然是原始感知数据, 很容易遭受内部攻击 [15,17] (如数据库泄露和服务器管理人员不当操作等). 因此, 有效的隐私保护应当是在感知设备端实现对原始感知数据的本地隐私保护.

挑战二: 维度灾难. 群智感知数据的一个特性在于感知数据的属性维度高, 并且属性维度间存在着复杂的关联性 [18], 从而无法直接对每维数据进行独立的隐私保护 [19]. 而对高维数据进行直接的隐私保护, 在相同隐私保证下, 不仅隐私保护后的感知数据效用性较低, 而且还会导致较大的计算开销 [20,21]. 因此, 在保证原始数据属性相关性的条件下对数据进行隐私保护具有很大的挑战.

针对以上的挑战, 不同的隐私保护机制被相继提出. 一方面, 这些机制中有的可以为分布式系统提供一定程度的本地隐私保护, 但因为效用性不足或计算复杂度较高而难以适用于高维数据上; 另一方面, 有的机制研究重点解决高维数据的中心化隐私保护, 通过对高维数据进行降维后实现 “分而治之” 的一般低维隐私保护, 取得了不错效果, 但是却不能为分布式系统提供有效的本地隐私保证. 为了解

决群智感知系统中现有隐私保护机制在本地隐私和数据高维之间的兼容性难的问题, 本文提出了一种基于 Bayes 网络的高维感知数据本地隐私保护发布机制, 具体贡献如下所示:

(1) 本文针对群智感知系统, 提出了一个满足本地差分隐私的高维感知数据汇聚和发布机制. 不但可以为群智感知用户提供本地隐私保证, 而且可以对高维感知数据的统计特性进行近似估计并发布具有相似分布的合成数据, 达到本地差分隐私与高维数据效用性之间的良好折衷.

(2) 本文针对 Bayes 网络的构建方法提出了基于熵估计的启发式改进, 可以使得构建的 Bayes 网络较好地保留属性间的相关性, 并减小构建过程中的计算量, 在一定程度上提高算法的效率与稳定性.

最后, 本文在多个真实数据集上对所提出的机制进行了大量的仿真实验. 实验结果表明, 本文所提出的机制很好地保留了高维数据的属性相关性, 使得合成数据集在统计查询任务和分析任务上均有较好的准确性.

2 相关工作

本文主要研究群智感知系统中高维感知数据发布过程中本地差分隐私保护的问题. 因此, 主要从高维数据发布的隐私保护和本地端差分隐私保护研究两方面对研究现状进行分析和总结.

统计发布中, 差分隐私最早被提出通过给发布结果添加适量的随机噪声以防单条数据记录被推测. 例如, 典型的数据发布隐私保护 [22~24] 是对数据值域上的直方图添加敏感度 (直方图敏感度为 1) 校准后的拉普拉斯 (Laplace) 噪声. 然而, 随着数据维度的增多, 一方面统计高维直方图的计算量随维度成指数增长; 另一方面, 高维直方图中大部分数据桶上频数为零, 呈现出很大的稀疏性, 原有的隐私保护噪声将导致极低的信噪比, 从而丧失数据的效用性. 目前, 针对高维数据的隐私发布问题, 大多研究主要是利用属性划分的思想, 将高维数据划分为多个低维数据簇, 再进行隐私发布处理. PriView [25] 先通过选择构建 k 个低维属性集合的边缘分布, 进而估计出高维的联合分布, 然而该方法假设所有属性是相互独立的, 且均等地处理所有属性对, 这并不符合移动群智感知系统中属性关联的实情. 大部分相关的研究则是利用属性间相关性作为划分依据, 如 PrivBayes [20] 利用 Bayes 网络来表示属性间相关性, 通过属性间相关性对数据进行划分, 然而由于该方法采样指数机制选取相关属性对, 当属性对太多时, 指数机制的选择精度会大大降低. 文献 [26] 对 Bayes 方法进行了加权改进. 又如, Chen 等 [21] 利用依赖图和联合树的方法来表示数据的维度关联, 此种方法计算了所有属性中任意两个属性间的相关性, 虽可能找出尽可能多的相关性, 但算法的复杂度很大. 还有一些文献, 主要用马尔科夫 (Markov) 链来表示数据的相关性, 其对时间相关的数据的应用更为有效. PrivHD [27] 通过马尔科夫网及网络分割形成联合树来进行高维数据的降维分割, 同时其引入了满足差分隐私的高通滤波技术来缩减指数机制的搜索空间. 然而, 这些方法都是集中式的处理, 并不适用于群智感知系统分布式环境.

而本地端差分隐私 [28, 29] 是适用于分布式环境的一种隐私保护方法, 它是在差分隐私基础上提出的一种隐私保护概念, 属于一个相对较新的研究领域 [30~34]. 有文献提出了基于压缩输入域的扰动机制 [35], 基于信息扭曲的扰动机制 [36] 和随机化应答技术 [37] 的本地差分隐私实现方法, 其中随机化应答技术是本地化差分隐私的主流扰动机制. RAPPOR [29] 基于随机化应答技术实现了本地保护, 但该方法对低维数据的统计查询比较有效, 随着数据维度的增加, 其通信代价呈指数倍的增长. 在 RAPPOR 机制的基础上, Kairouz 等 [38] 基于属性变量取值未知的情形提出了 O-RAPPOR 方法. O-RAPPOR 在 RAPPOR 中编码和解码方法的基础上, 引入了哈希 (Hash) 映射和分组 (cohort) 操作, 其目的是减小属性取值本身对随机化应答处理的影响. K-RR [39] 是另一种单值频数发布下的经典方法, 不同于 RAPPOR 对每种取值编码后进行随机化应答处理, K-RR 方法直接在变量的多个取值之间进行随机

化应答. 类似地, Kairouz 等[38] 在 K-RR 的基础上, 基于属性变量取值未知的情形提出了 O-RR 方法, 同样引入哈希映射和分组操作, 得到映射的哈希结果后, 再利用 K-RR 中的扰动方法进行隐私保护处理. 后来, 针对一对多扰动的情形, 文献 [40] 提出了 k-Subset 方法, 该方法将扰动输出扩展至集合的形式, 即对于指定的单个输入, 其可能会有多种输出结果. 此外, 近年来还有关于本地差分隐私应用于各类数据类型的研究, 如图数据[41], 集值数据[42], 键值数据[33] 等. 也有其他分布式环境下的差分隐私保护研究, 如文献 [43] 提出分布式环境下满足差分隐私的逻辑回归模型.

3 系统模型

本文考虑一个群智感知系统, 该系统由大量感知用户与一个中心服务器连接组成. 感知用户首先感知和采集各自具有多个属性维度的数据记录, 并对数据记录进行本地隐私保护, 然后将隐私保护后的多维数据记录发送给中心服务器. 服务器收集到所有本地隐私保护后的数据记录后对高维数据的统计分布进行估计分析, 然后, 基于该统计分布合成一个近似分布的全新数据集并发布给第三方用户进行公开查询和挖掘. 本文的关注重点为数据隐私, 因此不考虑具体的网络模型.

问题描述. 假定系统中有 N 个用户, 每个用户记录包含 d 个属性, 数据发布的任务旨在由中心服务器发布一个与原始数据集大小相同且具有相似分布的合成数据集. 设原始数据集为 $X = \{X^1, X^2, \ldots, X^N\}$, 其中 X^i 表示第 i 个用户的数据记录. 数据集的属性集合为 $A = \{A_1, A_2, \ldots, A_d\}$, 用 x_j 表示对应属性 A_j 的取值, 则单个用户的数据表示为 $X^i = \{x_1^i, x_2^i, \ldots, x_d^i\}$, x_j^i 是用户 i 的第 j 个属性的取值. 属性值域为 $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_d\}$, 其中 $\Omega_j = \{\omega_j^1, \omega_j^2, \ldots, \omega_j^{|\Omega_j|}\}$ 表示属性 A_j 的值域, 其中 ω_j^i 是属性 A_j 的第 i 种取值, $|\Omega_j|$ 是值域的模. 中心服务器端接收到所有用户数据后通过一系列处理, 最终发布一个与原始数据集 X 属性集合 A 值域相同的且具有 N 条记录的近似合成数据集 X^*, 使得 X^* 上关于属性集合 A 的联合概率分布满足

$$P_{X^*}(A_1, A_2, \ldots, A_d) \approx P_X(A_1, A_2, \ldots, A_d), \tag{1}$$

其中 $P_X(A_1, A_2, \ldots, A_d) \triangleq P_X(x_1 = \omega_1, \ldots, x_d = \omega_d)$ 是原始数据集 X 在属性集合 A 上 d 维的联合概率分布, x_i 表示第 i 个属性变量, $\omega_i \in \Omega_i$.

4 基础知识

4.1 本地差分隐私

差分隐私 (differential privacy, DP)[16] 是一种通过向原始数据添加适量的随机噪声来掩盖真实数据的隐私保护技术, 可大规模应用且有良好的数学理论基础. 基于差分隐私技术的保护机制常作用于集中式数据库, 假设数据已经安全获取, 而且收集者是可信的. 事实上, 数据库服务器可能并不总是隐私安全可信的. 因此, 本地差分隐私 (local differential privacy, LDP)[39,44] 的概念被提出, 其强调将数据扰动的功能从中心服务器端移至用户端本地进行, 使用户可以独立地处理自身的敏感信息.

本地化差分隐私保护模型充分考虑到数据汇聚过程中数据收集者窃取或泄露参与者 (用户) 隐私的可能, 在本地化差分隐私模型中, 每个参与者 (用户) 先对自身所持数据进行隐私化处理, 然后再将处理后的数据发送给中心服务器 (即数据收集者), 中心服务器对收集到的数据进行统计分析, 获得分析结果的同时, 又保证个体的隐私信息不被泄露. 本地差分隐私的定义如下.

定义1 (本地差分隐私 [44]) 给定 N 个用户, 每个用户对应一条记录. 给定隐私保护算法 M 及其定义域 $\mathrm{Dom}(M)$ 和值域 $\mathrm{Ran}(M)$, 若算法 M 在任意两条记录 X^i 和 $\hat{X}^i$ ($X^i, \hat{X}^i \in \mathrm{Dom}(M)$) 上得到相同输出结果 X^* 的概率满足

$$P(M(X^i) = X^*) \leqslant e^{\varepsilon} P(M(\hat{X}^i) = X^*), \tag{2}$$

则认为机制 M 满足 ε- 本地差分隐私.

由上述定义可知, 本地差分隐私技术是通过控制任意两条记录输出结果的相似性, 来确保机制 M 满足 ε- 本地差分隐私的 [44].

随机化应答方法 [37] 是目前最常见的本地隐私保护技术, 该方法主要是利用响应的不确定性来对原始数据进行隐私保护. 随机化应答技术最早应用于社会学研究中, 当用户在对隐私问题进行回答时, 在 "是" 和 "否" 两个答案中进行随机选择. 其中, 被调查者在一定的概率 p 下给出其真实答案, 同时可在概率 $1-p$ 下随机给出答案. 被调查者的真实响应无法确定而使得被调查者的隐私得以保护, 而当存在大量调查者的响应时, 又可通过概率推测出较为真实的结果, 从而保证数据的效用.

4.2 Bayes 网络

Bayes 网络是一种概率图模型, 常用于变量间依赖关系的处理 [21, 26], 在形式上, 它是一个有向无环图 (directed acyclic graph, DAG). 假设 A 是数据集 D 上的属性集, 其维度大小为 d, Bayes 网络将 A 中的每个属性 A_i 表示为一个节点, 并且用一条有向边连接某两个节点来表示它们之间相关, 假设节点 A_i 直接影响节点 A_j, 则用一条从 A_i 指向 A_j 的有向弧 (A_i, A_j) 来表示两者有因果关系或者非条件独立, 即 $A_i \to A_j$. Bayes 网络的核心是条件概率, 本质上是利用先验知识确立一个随机变量 (属性) 间的关联关系.

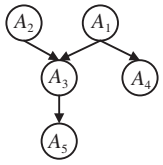
Bayes 网络可看作 d 个属性对 (AP 对, attribute parent) 的集合, 每个属性对包含一个节点与其所有父节点的集合 Π_i, 属性对表示为 (A_i, Π_i). 令 N 表示一个 Bayes 网络图, A 表示网络中所有节点的集合, $A = (A_1, A_2, \ldots, A_d)$, 则所有属性的联合概率分布可表示为

$$P(A) = \prod_{i=1}^{d} P(A_i|\Pi_i) = P(A_d|A_1, \ldots, A_{d-1}) \cdots P(A_2|A_1)P(A_1). \tag{3}$$

图 2 是 Bayes 网络的一个示例. 图中展示了通过 Bayes 网络将 5 个联合节点分解为 5 个 AP 对, 也即低维属性簇的一个分组情况, 图中所有节点 $A_1, A_2, \ldots, A_5$ 的联合概率可计算为 $P(A_1, A_2, \ldots, A_5) = P(A_1)P(A_2)P(A_3|A_1A_2)P(A_4|A_1)P(A_5|A_3)$.

5 高维感知数据本地差分隐私保护发布算法

基于以上的系统模型、问题定义和相关背景知识, 本文提出了一种基于 Bayes 网络的高维感知数据本地差分隐私保护发布机制, 可以实现对用户本地隐私保护后的高维感知数据在中心服务器端进行高效用性的数据发布. 图 3 展示了本文整个方案的基本框架, 主要包含 3 个模块: 本地端隐私保护, 基于 Bayes 网络的高维感知数据降维, 采样合成数据集, 其中本地隐私保护机制是在感知用户本地端进行的, 而高维感知数据降维处理和采样合成数据均是在中心服务器端进行.



(a)

Attribute	Attribute-parent pairs
A_1	$(A_1,\varnothing)$
A_2	$(A_2,\varnothing)$
A_3	$(A_3,\{A_1,A_2\})$
A_4	$(A_4,\{A_1\})$
A_5	$(A_5,\{A_3\})$

(b)

图 2 Bayes 网络示例

Figure 2 An example of Bayesian network. (a) A Bayesian network with five attributes; (b) corresponding attribute-parent pairs

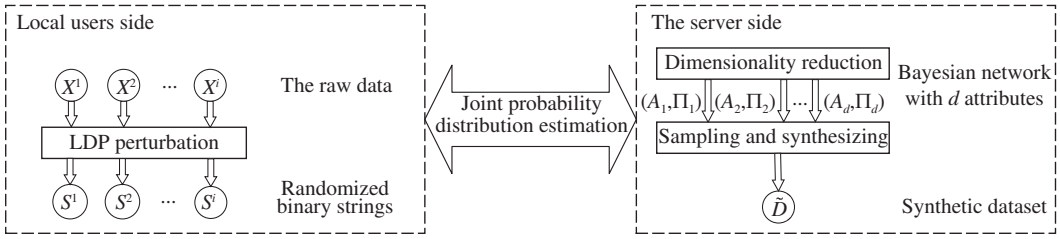


图 3 高维数据本地隐私保护发布机制框架

Figure 3 The framework of our proposed mechanism

5.1 本地端隐私保护机制

随机化应答技术可以实现本地隐私保护, 但是其仅可对包含两种取值的离散型数据进行扰动处理, 对于多值的情况并不适用. 要处理多值数据, 本文借鉴 RAPPOR [29] 中基本的对变量进行二进制化的方法将用户数据 x_j^i 用二进制串 s_j^i 的形式表示1). 本文主要根据属性的值域及属性取值在值域中的位置来构成二进制串: 本地用户根据属性变量 A_j 的值域大小 $|\Omega_j|$ 来确定二进制串的长度, 每种取值 ω_j^i 对应二进制串中的一位, 记 loc_i 为取值 ω_j^i 所在的位, 在转换数据时, 将二进制串中第 loc_i 位置 1, 其他位置 0, 即可得到该数据的唯一二进制串 s_j^i. 同时可知, 该变量中所有取值的特征二进制串是相互独立的, 故所有属性取值可唯一表示. 如图 4 所示, 即为一个属性二进制化的示例图, 下方为属性的值域及对应取值的特征二进制串.

5.1.1 本地隐私保护算法过程

本地端差分隐私保护具体处理过程如算法 1 所示, 主要包含以下 3 个步骤.

Step 1. 二进制化处理. 对于用户 i, 假设有一条 d 个属性的原始记录 $X^i=\{x_1^i,x_2^i,\ldots,x_d^i\}$, 其中 x_j^i 表示用户 i 第 j 个属性的取值. 对于每个属性 A_j, 其值域大小为 $|\Omega_j|$, 通过对比原始数据 x_j^i 与属性值域集合 Ω_j 确定数值 x_j^i 的位置 loc, 将长度为 $|\Omega_j|$ 二进制串 s_j^i 的第 loc 位置为 1.

1) 对值域较大的情况亦可使用布隆过滤器方法将数据进行二进制编码.

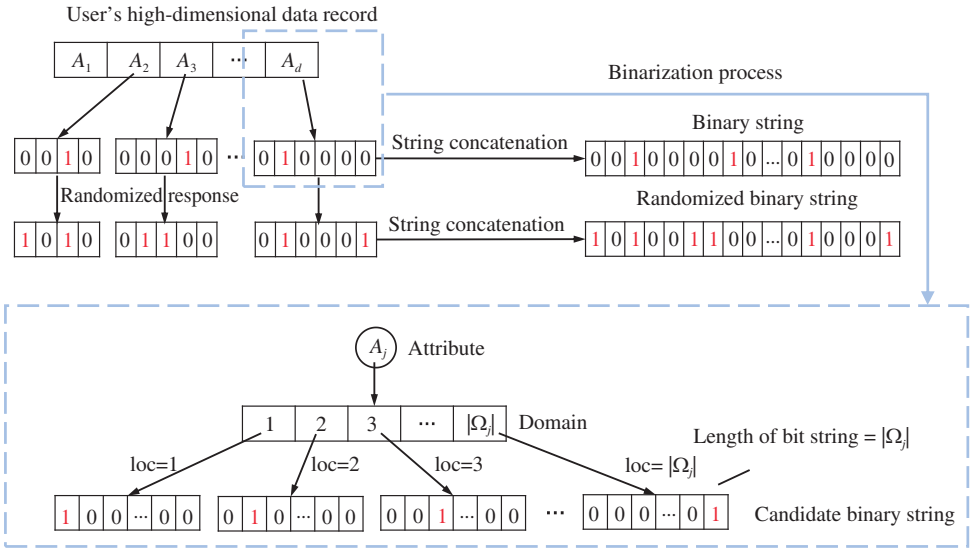


图 4 (网络版彩图) 数据二进制化示意图

Figure 4 (Color online) Illustration of data binarization process

Algorithm 1 Data transformation with local differential privacy

Input: User's data record $\{x_j^i | j = 1, 2, \ldots, d\}$, attribute set $A = \{A_1, A_2, \ldots, A_d\}$, random flipping probability f;

Output: Randomized binary string $\hat{s}^i$ of raw data X^i;

1: **for** $1 \leqslant j \leqslant d$ **do**

2: Each user i transform each attribute jth value into a binary string s_j^i;

3: Randomly flip each bit of s_j^i to obtain a randomized binary string $\hat{s}_j^i$;

4: **end for**

5: Concatenate randomized binary strings for all d attributes to obtain $\hat{s}^i$.

Return: $\hat{s}^i$.

Step 2. 比特位随机翻转. 二进制串 s_j^i 的每个比特位按照如下公式进行随机性赋值.

$$\hat{s}_j^i[b] = \begin{cases} s_j^i[b], & \text{with probability of } 1-f, \\ 1, & \text{with probability of } f/2, \\ 0, & \text{with probability of } f/2. \end{cases} \tag{4}$$

Step 3. 二进制串连接. 将用户 i 所有属性的二进制串 $\hat{s}_j^i$ $(j = 1, 2, \ldots, d)$ 连接起来得到一个具有 $\sum_{j=1}^{d} |\Omega_j|$ 位的比特向量 $\hat{s}^i$, 并将其发送到中心服务器, 此时认为该比特向量已经具有本地隐私保护.

5.1.2 隐私性分析

定理1 假设数据记录有 d 个属性, 其在用户本地端进行随机化应答的翻转概率为 f, 则用户端的本地差分隐私级别 [29] 为

$$\varepsilon = 2d \ln \left(\frac{1 - \frac{1}{2}f}{\frac{1}{2}f} \right).$$

证明 本地隐私保护机制首先在本地端将用户数据进行隐私扰动, 只有用户自身持有原始数据, 在数据发送后, 杜绝了其他参与者和攻击者获取用户原始信息的机会, 故此, 用户的隐私信息得以保

护. 另外, 中心服务器在获取用户数据后, 并未对收集的数据进行噪声添加处理, 故本文的隐私保证主要来源于本地端处理.

令 T 表示用户的原始二进制串, T' 表示翻转后的二进制串, T 和 T' 分别为两个用户的不同记录, 则条件概率比值 $\frac{P(T'=T^*|T=T_1)}{P(T'=T^*|T=T_2)}$ 即与隐私级别 ε 相关, 记作 RR. 由于用户原始数据中单个属性的二进制串中只有一位为 1 (代表该属性取值的位置 1), 故用户原始二进制串乱序可得 $T=\{t_1=1,\ldots,t_d=1,t_{d+1}=0,\ldots\}$, 由转换公式可知, 位数值不变的概率为 $1-\frac{1}{2}f$, 位数值改变的概率为 $\frac{1}{2}f$. 据参考文献 [29] 中的分析, 可计算

$$\begin{aligned}\mathrm{RR}=\frac{P(T'=T^*|T=T_1)}{P(T'=T^*|T=T_2)}&\leqslant\max_{T_1,T_2,T^*\in\mathrm{Ran}(M)}\frac{P(T'=T^*|T=T_1)}{P(T'=T^*|T=T_2)}\\&\leqslant\max_{T^*\in\mathrm{Ran}(M)}\left(1-\frac{1}{2}f\right)^{2\left(t'_1+\cdots+t'_d-t'_{d+1}-\cdots-t'_{2d}\right)}\times\left(\frac{1}{2}f\right)^{2\left(t'_{d+1}+\cdots+t'_{2d}-t'_1-\cdots-t'_d\right)},\end{aligned}$$

当且仅当 $t'_i=1-t_i$ 时, 比值最大, $\mathrm{RR}_{\max}=(\frac{1-\frac{1}{2}f}{\frac{1}{2}f})^{2d}$, 得 $\varepsilon=2d\ln(\frac{1-\frac{1}{2}f}{\frac{1}{2}f})$.

5.2 本地隐私保护后的高维数据合成与发布

5.2.1 基本思想

本地隐私保护后的高维数据发布目标旨在发布一个与原始数据集统计特性相似 (如概率分布, 以式 (2) 所定义为依据或目标) 的新数据集. 为此, 有两种直观的方法: 一是单独地估计每一维上的概率分布然后一维一维采样合成高维数据, 然而不考虑维度之间的相关性将导致最终合成数据无法进行多维联合查询和相关性分析, 丢失高维数据的价值; 另一种是将所有属性维度同时进行概率分布估计并基于估计的概率分布采样合成新数据集, 然而, 以下将会分析, 由于组合特性, 所有维度组成的完全属性值域空间将随着维度数以指数方式递增, 从而导致极大的运算复杂度和极低的估计准确性. 可见, 隐私保护的高维数据发布的核心在于选择合适的方案来降低维度, 将高维数据分解为多个低维数据, 但同时尽可能保持原有属性维度上的相关性, 使得数据仍然可以满足多维联合查询和相关性分析的作用.

本文采用 Bayes 网络来构建高维数据中属性维度间的相关性, 并利用其多维联合概率分布的组合特性, 实现对高维联合概率分布的估计. 特别地, 中心服务器端接收到所有用户本地隐私保护后的数据后, 通过低维下可行的联合概率分布估计算法计算属性间的相关性, 然后构建 Bayes 网络, 以合成新的数据集并发布. 考虑到本文处理的感知数据为异构数据, 本文用互信息来度量属性间的相关性, 而互信息计算的关键在于从本地保护后的感知数据中求解计算双方属性联合概率分布. 在本文 Bayes 网络构建过程中, 需求解单个属性 A_j 与其父节点集合 Π_j 间的互信息, 故需求解多维属性间的联合概率分布. 接下来将先介绍 m 维联合概率分布估计算法, 然后阐述 Bayes 网络的构建步骤.

5.2.2 多维联合概率分布估计

本文主要扩展期望最大化的估计算法 (EM 算法) [45, 46] 来计算任意多维 (如 m 维) 属性间的联合概率分布. 根据自定义的收敛精度 δ, 通过不断迭代, 得到概率分布的期望值, 即为所求, 具体过程说明如下. 记 m 维属性集合为 $A=\{A_1,A_2,\ldots,A_m\}$, 索引集合为 $C=\{1,2,\ldots,m\}$, 用 ω_j 表示属性 A_j 的取值, 则 m 维属性联合概率分布可简单表示为 $P(\omega_C)$ 或 $P(\omega_1\omega_2\cdots\omega_m)$, 共有 N 个用户. 如算法 2 所示, 多维联合概率分布估计算法包含以下环节.

Step 1. 参数初始化. 设初始联合概率为 $P_0(\omega_1\omega_2\cdots\omega_m)=\frac{1}{(\prod_{j=1}^{m}|\Omega_j|)}$ (算法 2 第 1 行).

Algorithm 2 m-dimensional joint probability distribution estimation algorithm

Input: Attributes index set $C=\{1,2,\ldots,m\}$, randomized binary string $\hat{s}_j^i$ $(1\leqslant j\leqslant m)$, flipping probability f, convergence accuracy δ, attribute set $A=\{A_1,A_2,\ldots,A_m\}$, attribute domain size $\Omega=\{\Omega_1,\Omega_2,\ldots,\Omega_m\}$;

Output: m-dimensional joint probability distribution $P(\omega_C)$;

1: Initialize $P_0(\omega_C)=\frac{1}{(\prod_{j=1}^m|\Omega_j|)}$;
2: **for** each user $i=1,\ldots,N$ **do**
3: **for** each attribute $j\in C$ **do**
4: Compute $P(\hat{s}_j^i|\omega_j)=\prod_{b=1}^{|\Omega_j|}(\frac{f}{2})^{s_j^i[b]}(1-\frac{f}{2})^{1-s_j^i[b]}$;
5: **end for**
6: Compute $P(\hat{s}_C^i|\omega_C)=\prod_{j=1}^m P(\hat{s}_j^i|\omega_j)$;
7: **end for**
8: Set $t=1$;
9: **repeat**
10: **for** each user $i=1,\ldots,N$ **do**
11: Compute $P_t\left(\omega_C|\hat{s}_C^i\right)=\frac{P_{t-1}(\omega_C)P(\hat{s}_C^i|\omega_C)}{\sum_{\omega_C}P_{t-1}(\omega_C)P(\hat{s}_C^i|\omega_C)}$ $(\omega_C\in\Omega_1\times\Omega_2\times\cdots\times\Omega_m)$;
12: **end for**
13: Compute $P_t(\omega_C)=\frac{1}{N}\sum_{i=1}^N P(\omega_C|\hat{s}_C^i)$;
14: $t=t+1$;
15: **until** $\max_{\omega_C}P_t(\omega_C)-\max_{\omega_C}P_{t-1}(\omega_C)\leqslant\delta$;

Return: $P(\omega_C)=P_t(\omega_C)$.

Step 2. 条件概率计算. 计算每个用户的 m 维数据的条件概率, 即 $P(\hat{s}_1^i\hat{s}_2^i\cdots\hat{s}_m^i|\omega_1\omega_2\cdots\omega_m)$. 由于用户的二进制串中每个比特位的含义是不同的, 且比特位的翻转是相互独立的, 可知其 m 维属性联合条件概率即为每个比特位条件概率的乘积, 即 $P(\hat{s}_1^i\hat{s}_2^i\cdots\hat{s}_m^i|\omega_1\omega_2\cdots\omega_m)=\prod_{b=1}^{|\Omega|}(\frac{f}{2})^{s_C[b]}(1-\frac{f}{2})^{1-s_C[b]}$ (算法 2 第 2∼7 行).

Step 3. 期望最大化估计. 初始化迭代次数 $t=1$ (算法 2 第 8 行), 算法进入期望最大化算法估计的迭代过程, 包含两个步骤.

- E 步: 后验概率计算. 已知所有用户二进制串的条件概率, 由 Bayes 概率公式可得

$$P_t\left(\omega_1\omega_2\cdots\omega_m|\hat{s}_1^i\hat{s}_2^i\cdots\hat{s}_m^i\right)=\frac{P_{t-1}(\omega_1\omega_2\cdots\omega_m)P(\hat{s}_1^i\hat{s}_2^i\cdots\hat{s}_m^i|\omega_1\omega_2\cdots\omega_m)}{\sum_{\omega_1}\cdots\sum_{\omega_m}P_{t-1}(\omega_1\omega_2\cdots\omega_m)P(\hat{s}_1^i\hat{s}_2^i\cdots\hat{s}_m^i|\omega_1\omega_2\cdots\omega_m)},$$

其中 $P_{t-1}(\omega_1\omega_2\cdots\omega_m)$ 是 $t-1$ 次迭代时的值 (算法第 10∼12 行).

- M 步: 迭代更新参数 $P_t(\omega_C)$. 将 N 个用户的后验概率求平均 $P_t(\omega_C)=\frac{1}{N}\sum_{i=1}^N P(\omega_C|\hat{s}_C^i)$ 替换上一轮的先验概率, 作为新的 k 维联合概率分布 (算法 2 第 13 行), 然后返回 E 步.

以上步骤不断迭代直至两轮联合概率的差值小于收敛精度 δ, 即 $\max_{\omega_C}P_t(\omega_C)-\max_{\omega_C}P_{t-1}(\omega_C)\leqslant\delta$, δ 可根据精度要求自行定义 (算法 2 第 15 行).

一般来说, 如果初始值选择合适的话, 基于 EM 算法的多维联合概率分布估计算法经过一定迭代次数后可以收敛到一个较好的估计值. 但是, 随着维度 m 的增多, 多维组合后的状态空间大小为 $\prod_{j=1}^m|\Omega_j|$, 具有指数增长的趋势. 因此, 带来算法复杂度的急遽增大. 而且, 随着状态空间的增加, 很多状态实际真实值根本不存在 (也即稀疏特性), 然而 EM 算法仍然可能会为这些稀疏状态进行概率分布估计, 带来极大的估计误差, 从而最终表现出极大的效用性丢失.

5.2.3 Bayes 网络构建

得到任意 m 维属性的联合概率分布, 即可求解 m 维属性间的互信息, 互信息 I 的取值越大, 则

说明两者越相关. 假设我们要基于数据集 D 构建一个最大入度数为 k (即每个节点的最大父节点个数为 k) 的 Bayes 网络 N, 将 $A=\{A_1, A_2, \ldots, A_d\}$ 中每个属性作为 Bayes 网络的一个节点, 网络是从属性集中逐个选取节点构建而成的. 算法 3 详细说明了 Bayes 网络的构建过程, 如下所示.

Algorithm 3 Bayesian network construction algorithm

Input: Dataset D, maximal degree of bayesian network k, attribute set A;
Output: Bayesian network N;
1: Initialize $N=\emptyset$, $S=\emptyset$;
2: Randomly pick a node from A as X_1 and add it into S, add $(X_1, \emptyset)$ into N;
3: **for** $i=2$ **to** d **do**
4: For $\forall X \in A/S$ and $\forall \Pi \in C_S^k$, add (X, Π) into Ω and compute $I(X, \Pi)$;
5: Choose the attribute-parent pairs (X_i, Π_i) with the maximal I;
6: Add X_i into S and add (X_i, Π_i) into N;
7: **end for**
Return: $N=\{(X_1, \emptyset), (X_2, \Pi_2), \ldots, (X_d, \Pi_d)\}$.

假设 S 集合保存已添加的属性, 初始设 $S=\emptyset$, k 为父节点的最大个数 (算法 3 第 1 行). 首先, 从 d 个属性中随机选取一个属性作为 Bayes 网络的初始节点 X_1, 并将其父节点集合置为空, 即 $\Pi_1=\emptyset$ (算法 3 第 2 行). 同时将 X_1 添加到集合 S 中, 将 $(X_1, \emptyset)$ 添加至 N. 其次, 将 S 中的节点按 k 个取值, 得到 $C_{|S|}^k$ 个所有可能父节点集合 (S 中不足 k 个节点时, 将其整体看作一个父节点集合, 目的是为了保证父节点个数不超过 k), 分别与 A/S 中所有节点组成 AP 对 (X, Π) 存入 Ω 内, 求 Ω 内所有 AP 对的互信息 I (算法 3 第 4 行, 互信息的计算方法在算法后详述). 然后, 选取使得 I 值最大的属性对添加至 Bayes 网络中, 同时将该点移至集合 S (算法 3 第 5, 6 行). 重复上述步骤, 至 $S=\{1, 2, \ldots, d\}$, 则 Bayes 网络构建完成.

算法 3 中第 4 行提到的相关性计算, 计算公式为

$$I(X, \Pi)=\sum_{x \in \mathrm{Ran}(X)} \sum_{y \in \mathrm{Ran}(\Pi)} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \tag{5}$$

其中 $\mathrm{Ran}(X)$ 和 $\mathrm{Ran}(\Pi)$ 分别表示属性节点 X 和属性集合 Π 的值域, $p(x, y)$ 是 (X, Π) 取值为 (x, y) 时的联合概率, 可用第 5.2.1 小节中的算法 2 求解, 此时属性维度 $m=2$, $p(x)$ 和 $p(y)$ 分别表示 X 和 Π 取值为 x, y 时的先验概率, 根据联合概率与边缘概率的关系, $p(x)$ 和 $p(y)$ 可直接从 $p(x, y)$ 联合概率处得到.

5.2.4 Bayes 网络改进

以上 Bayes 网络构建可以达成对高维属性的分解, 有效降低计算负载并提升本地隐私下的数据发布效用性, 然而, 观察仍存在两个缺陷: 一方面, 算法 3 中随机选取初始节点 (第 2 行), 每次构建的 Bayes 网络具有很大的不确定性, 相关属性节点选取不定, Bayes 网络近似的联合概率分布随机误差较大. 另一方面, 算法 3 中 $I_{\max}$ 的选取需要计算 Ω 内所有属性对的互信息, 然而, 在迭代过程中, 每次仅从 V 中选取一个节点, 相关性较弱的属性对会在之后的互信息计算中重复出现, 这些重复的计算既浪费了内存, 又增加了计算量. 针对这两点缺陷, 本文还提出了一种 Bayes 网络构建的改进算法, 如算法 4 所示.

考虑到式 (5) 中互信息可以改写为信息熵的形式, 也即

$$I(X, \Pi)=H(X)+H(\Pi)-H(X, \Pi), \tag{6}$$

Algorithm 4 An enhanced algorithm for Bayesian network construction

Input: Dataset D, degree of Bayesian network k, attribute set A;
Output: Bayesian Network N;
1: Initialize $N=\emptyset$, S=$\emptyset$ and $V=A$;
2: Compute the entropy H for each attribute in A, and choose the attribute with the maximal entropy as X_1, add it into S and $(X_1,\emptyset)$ into N;
3: **for** $i=2$ **to** d **do**
4: $\Omega=\emptyset$;
5: **if** $|S|>k$ **then**
6: For $\forall X\in A/S$ and $\forall\{\Pi\in C_S^k|X_{\text{pick}}\in\Pi\}$, add (X,Π) and (X_j,Π_j) into Ω, then compute $I(X,\Pi)$;
7: **else**
8: For $\forall X\in A/S$ and $\forall\Pi\in C_S^k$, add (X,Π) into Ω, then compute $I(X,\Pi)$;
9: **end if**
10: Choose the attribute-parent pair with the maximal I and denote as $(X_{\text{pick}},\Pi_{\text{pick}})$;
11: Choose the attribute-parent pair with the maximal value of $I'(X_j,\Pi_j)$ $(X_j\neq X_{\text{pick}})$ and denote as $(X_{\text{pick}},\Pi_{\text{pick}})$;
12: Add X_{pick} into S and $(X_{\text{pick}},\Pi_{\text{pick}})$ into N;
13: **end for**
Return: $N=\{(X_1,\emptyset),(X_2,\Pi_2),\ldots,(X_d,\Pi_d)\}$.

其中 $H(x)$ 表示变量的信息熵, 描述一个变量的不确定性, 熵值越大, 则表明该节点的不确定性越大. $H(X,\Pi)$ 表示变量 X 与 Π 之间的交叉熵. 受该公式启发, 寻找互信息值 I 最大的属性对时, 一般信息熵值 $H(X)$ 越大的属性 X 越可能被挑中. 因此, 本文考虑启发式地选取信息熵最大的属性作为初始节点构建 Bayes 网络, 从而概率上增大所构建的 Bayes 网络中的属性间相关性, 更好地保持高维数据间的联合查询准确性. 由此, 可以通过熵值比较来进行初始节点的选择, 具体见算法 4 中第 2 行. 其中, 信息熵的计算公式如下:

$$H(x)=-\sum p(x_i)\log p(x_i),\quad i=1,2,\ldots,|\Omega|,$$

其中 x_i 是 x 属性的第 i 个可能取值, $p(x_i)$ 为 x_i 的边缘概率.

针对属性互信息计算过程的冗余重复过程, 本文考虑对算法 3 的第 4∼6 行进行修改, 减少 Ω 内属性对的个数, 从而降低计算量. 首先, 当 $|S|>k$ 时, Ω 内属性对大于 1, 计算所有属性对的互信息 I, 选取取得最大互信息的属性对 $(X_{\text{pick}},\Pi_{\text{pick}})$ 和剩余其他所有属性 $X_j\neq X_{\text{pick}}$ 中互信息 I 值最大的属性对 (X_j,Π_j) (算法 4 中第 9 和 10 行); X_{pick} 即为要添加至 Bayes 网络的节点, 分别将 X_{pick} 和 $(X_{\text{pick}},\Pi_{\text{pick}})$ 添至 S 与 N 中 (算法 4 中第 11 行). 然后, 在下一轮迭代时, 只将包含上轮迭代中 X_{pick} 的 Π 与 $\forall X\in A\backslash S$ 组成属性对并存入 Π 中, 并将上轮迭代中选取的 (X_j,Π_j) 添至 Π 中 (算法 4 第 5 行). 最后, 再次计算互信息, 选择节点, 至 $V=\emptyset$ 则结束. 整个改进后的细节如算法 4 所示.

5.3 合成数据集

由式 (3) 可知, 可以根据 Bayes 网络的条件概率分布独立地采样生成每个属性上的数据, 从而合成新的数据集. 具体步骤如下:

步骤 1, 选取所构成的 Bayes 网络中父节点集合为 $\emptyset$ 的节点 (属性) 作为采样初始节点 X_1, 根据 5.2.1 小节中计算的边缘概率分布采样该节点数据, 其中各属性数据采样个数为记录个数 N.

步骤 2, 从未采样节点中，随机选取一个其父节点集合 Π_i 已被采样的节点 X_i 作为本轮的采样节点, 根据 5.2.1 小节中得到的联合概率分布来计算条件概率分布 $P(X_i|\Pi_i)$, 以条件概率分布为依据来采样节点.

表 1 数据集信息描述

Table 1 Details of datasets

Dataset	Data type	Dataset size	Dimension	Domain size (processed)
NLTCS	Binary	25174	16	2^{16}
Adult	Non-Binary	45222	15	$\approx 2^{26}$
TPC-E	Non-Binary	40000	24	$\approx 2^{38}$

不断重复步骤 2, 直至所有属性节点采样完毕, 所有节点的采样数据构成一个新的 $N \times d$ 的合成数据集, 该合成数据集即一定程度上保留与原始数据集统计概率分布特性相似的特性. 由于以上计算过程均是基于本地隐私保护后的用户数据处理实现的, 所以该算法过程整体上仍然保证了感知用户的本地端隐私.

6 实验评估

本节对文中所提出的机制在真实数据集上进行仿真实现并对机制性能就 Bayes 网络构建准确性、合成数据集多维概率分布准确性和合成数据集的分类任务准确度等多方面进行评估和分析.

6.1 实验设置

数据集. 在仿真实验中本文使用了 3 个真实世界数据集: (1) NLTCS, 是一个美国护理调查中心的数据集, 记录了 21574 名残疾人不同时间段的日常活动; (2) Adult, 1994 年美国人口普查中抽取出的 45222 个居民的部分信息; (3) TPC-E, 来自于 TPC 开发的在线事务处理程序, 记录了包括交易、交易类型、安全性、安全状态等 40000 条信息数据. 实验中, 简单起见, 对非二进制数据集 Adult 和 TPC-E 进行了采样处理, 并对属性值域进行合并压缩, 处理后 3 个数据集的具体信息如表 1 所示.

实验方法. 所有的仿真实验都是利用 Python2.7 进行实现的, 实验运行的硬件配置为 Intel i5-3470 内核, CPU 频率 3.20 GHz, 内存 8 GB, Windows 10 操作系统. 群智感知系统数据发布流程是通过以下步骤进行仿真的: 首先, 用户节点从数据集中依次读取数据, 并在本地进行隐私保护的数据处理, 并最终生成隐私保护的比特串. 然后, 这些比特串被发送到中心服务器端进行学习, 构建贝叶斯网络模型, 基于 Bayes 网络采样合成并最终发布一个全新的数据集进行任意查询.

实验参数. 仿真过程中, 所有数据集本地端隐私保护处理时的翻转概率 f 取值为 0.1 到 0.9 之间. 二值数据集 NLTCS 构建 Bayes 网络时, 最大入度 k 共有 1, 2, 3, 4 四种取值情况. 非二值数据集构建 Bayes 网络时最大入度 k 考虑了 1 和 2 两种取值情况.

评价指标. 实验目的为评估具有本地隐私保护的数据发布方法所合成数据的效用性, 主要从 3 个方面来评测合成数据效用性: 首先, 比较本地差分隐私保护后的合成数据集与原始数据集在构成 Bayes 网络方面的相关性识别差距, 用以衡量本地隐私保护下高维数据中维度相关性的丢失程度. 其次, 比较合成数据集中对多维属性上的边缘概率准确性, 通过比较合成数据集边缘概率分布和原始数据集边缘概率分布间的均方差距离来衡量其准确性, 从而衡量统计查询时的可靠性. 最后, 比较合成数据集与原始数据集在数据分析任务上的差别, 如 SVM 分类, 进一步衡量本地隐私保护下高维数据整体效用性的程度.

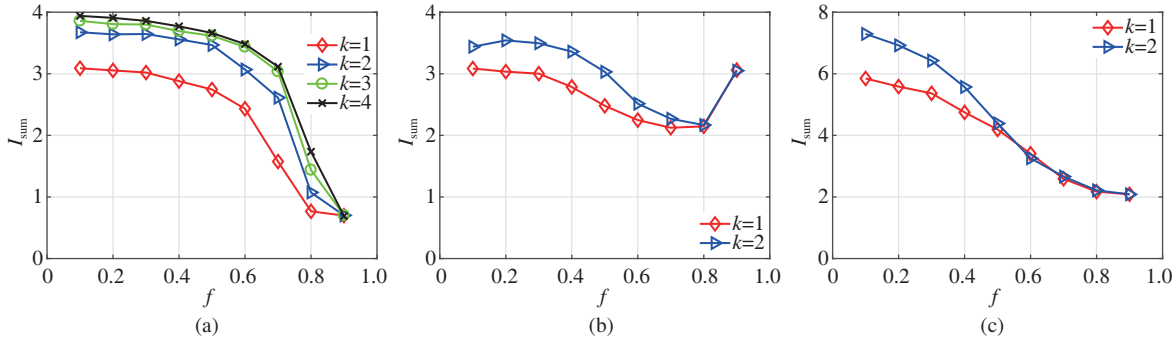


图 5 (网络版彩图) 合成数据集 I_{sum} vs. f

Figure 5 (Color online) I_{sum} (synthetic dataset) vs. f. (a) NLTCS; (b) Adult; (c) TPC-E

6.2 实验结果

6.2.1 Bayes 网络构建与属性相关性准确度

本小节通过实验研究 k 和 f 的取值对属性相关性识别准确度的影响, 以及基于熵的启发式初始节点选取方法对构建 Bayes 网络的影响. 本文用 Bayes 网络来建模高维数据中属性维度之间的关联性, Bayes 网络中的节点表示一个属性维度. 而 Bayes 网络中节点的最大入度 k 直接影响到单个节点的父节点个数, 从而影响数据相关属性对的计算. 此外, 翻转概率 f 在本地端隐私保护时, 决定用户传输数据的扰动概率, 进而影响 Bayes 网络构建的准确度.

图 5 展示了不同取值 k 和 f 下构建的 Bayes 网络中所有属性对的互信息之和 $I_{\text{sum}} = \sum_{i=1}^{d} I(A_i, \Pi_i)$. 互信息是衡量属性间相关程度的信息度量, 互信息越大则说明属性间相关性越高, 因此, 较大的互信息和 I_{sum} 一定程度上反映了高维数据中属性维度间相关性保留程度较好. 由图 5 可知, 整体而言, 随 k 值的增加, 各数据集的 I_{sum} 呈增长趋势. 直观地, k 值越大, 构建的 Bayes 网络越近似于数据集全概率分布 $\Pr[A]$. 然而, 图 5 中显示, 当 k 到达一定值时, I_{sum} 随 k 值的增长趋势十分缓慢, 此时则表明, 更大的 k 值对属性间相关性的开采没有更大的帮助, 即具有相关性的属性对已基本选择完全. 特别地, 对于不同数据集, 当 f 大到一定值时, Bayes 网络中属性相关性的识别准确度基本不受 k 值的影响. 然而, I_{sum} 随着 f 的变化趋势却又不尽相同, 这是因为互信息的计算与属性的边缘概率分布和联合概率分布均有关. 不同翻转概率时, 即使相同的 Bayes 网络, 所有属性 AP 对的互信息之和 I_{sum} 也会不同, 但至于 I_{sum} 变大或减小则视数据情况而定.

图 6 特别比较了随机选取 (random) 和基于信息熵 (entropy) 的启发式选取初始节点所构建出的 Bayes 网络中所有属性间的互信息之和 I_{sum}, 其中 $k = 2$. 从图中可以看出, 整体上, 基于信息熵的启发式初始节点选取方法所构建的 Bayes 网络中的所有属性间的互信息之和均不低于随机选取初始节点所构建的 Bayes 网络的属性间互信息和. 这说明了基于信息熵的启发式初始节点选取方法可以比原始随机选取方法更好地保持高维属性间的相关性, 从而保证合成数据集上联合查询的准确性. 就单个数据集而言, 在二值数据集 NLTCS 上, 由于二值属性分布的稀疏性并不明显, 属性间相关性本身比较强, 因而随机选取和基于信息熵的选取之间差异并不特别明显; 在非二值的 Adult 数据集和 TPC-E 数据集上, 一般地, 在 f 值较小的情况下, 基于熵的启发式选取方法比随机选取方法的互信息要高得比较多, 而在 f 值较大的情况下差异亦不明显. 这是因为在 f 较小的情况下, 概率分布估计准确性高, 基于熵的启发式方法有着明显的优势; 而随着 f 增大, 本身联合概率分布估计误差增大, Bayes 网络构建中选取随机性增大, 基于熵的启发式方法优势并不明显.

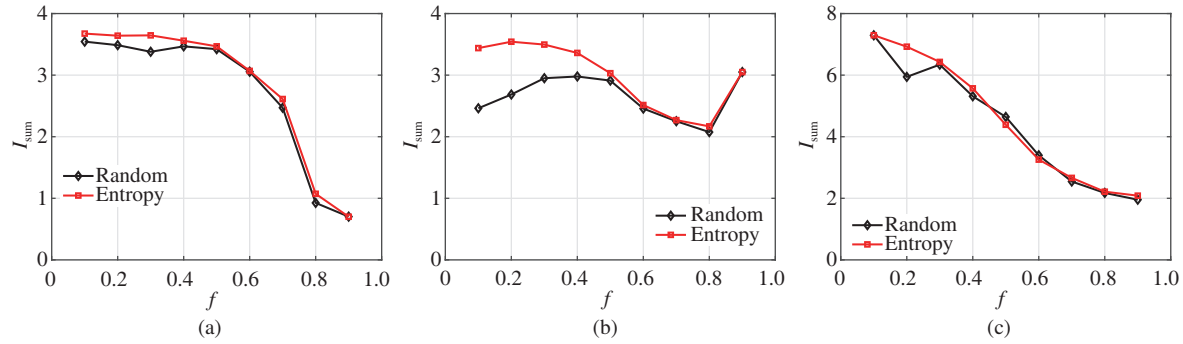


图 6 (网络版彩图) 互信息和 I_{sum} vs. f ($k = 2$)

Figure 6 (Color online) I_{sum} (synthetic dataset) vs. f ($k = 2$). (a) NLTCS; (b) Adult; (c) TPC-E

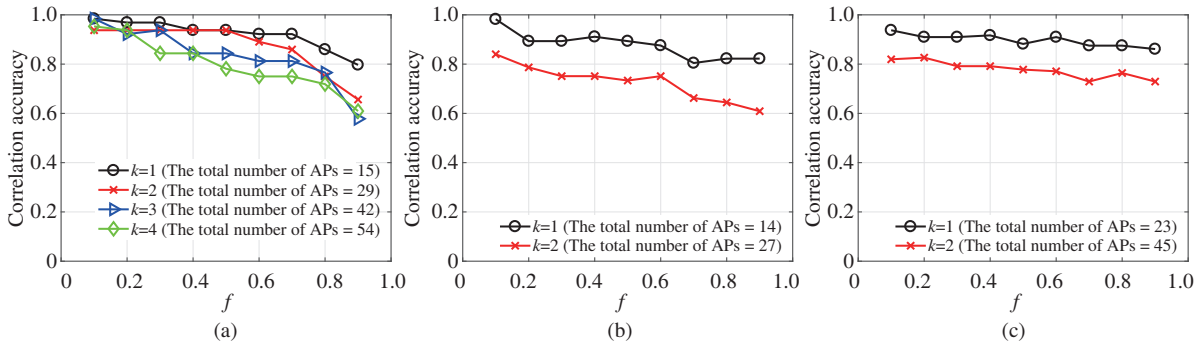


图 7 (网络版彩图) 相关性识别准确率

Figure 7 (Color online) Accuracy of correlation identification. (a) NLTCS; (b) Adult; (c) TPC-E

图 7 展示了不同 f 取值下构建的 Bayes 网络相比原始数据构建的 Bayes 网络中边连接的相对准确度. Bayes 网络中两两属性间是否存在边也直观地反映了相关属性的判断, 通过比较隐私保护后模型识别的相关属性与原始数据集上相关属性的识别正确率可以有效地反应本文机制对属性相关性识别的准确度. 其中, 需要强调的是, 由于不同的节点度 k 下原始数据构建的 Bayes 网络也是不同的, 不同 k 值之间无法直接对比, 因此, 这里只看 f 对构建准确度的影响. 从图 7 的实验结果可看出, 给定不同节点度 k, 随 f 值的增加, 属性间相关性识别的准确率整体降低, 这是因为隐私保护程度增加使得构建的 Bayes 网络的构建精度会带来一定程度损失.

6.2.2 统计查询精确度

本小节实验评估合成数据集的统计查询精度. 通过比较合成数据集与原始数据集上给定 a 维属性联合概率分布 (通过直接在合成后数据集上统计得到) 的误差来衡量合成数据集的统计查询精度. 用 Q_a 表示所有 a 维属性的联合. 其中误差采用了文献 [20, 21, 47] 中使用的平均偏差距离 (average variation distance) 进行衡量, 定义如下:

$$\text{AVD}(Q_a, \hat{Q}_a) = \frac{1}{2} \sum_{\omega \subset \Omega} \left| Q_a(\omega) - \hat{Q}_a(\omega) \right|,$$

其中 Ω 是 a 维属性联合的值域, Q_a 为真实数据集得到的联合概率, $\hat{Q}_a$ 为合成数据集得到的联合概率. 此外, 还同时使用了文献 [48] 中使用的 KL 散度进行度量.

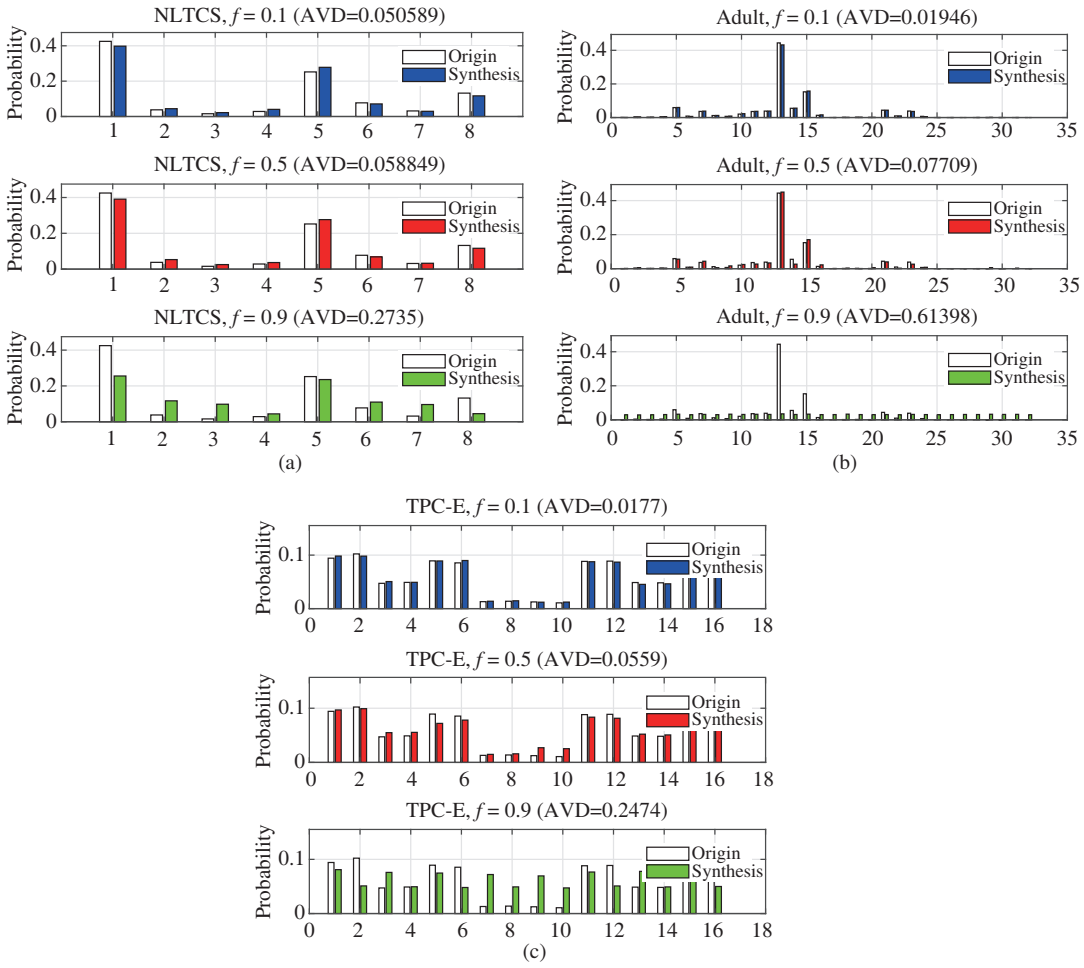


图 8　(网络版彩图) 不同值域 $|\Omega|$ 下的概率分布估计示意图

Figure 8　(Color online) Distribution estimation under different domain size $|\Omega|$. (a) NLTCS ($|\Omega| = 8$); (b) Adult ($|\Omega| = 32$); (c) TPC-E ($|\Omega| = 16$)

图 8 首先展示了在 NLTCS, Adult 和 TPC-E 3 个合成数据集上不同值域大小 ($|\Omega|$ 分别为 $8, 16, 32$) 的联合概率分布估计与原始联合概率分布对比的效果示意, 及对应分布差异的 AVD 值大小. 其中, 当 f 值较小为 0.1 时, 也即隐私保护比较弱的时候, 可以看出在不同值域上, 合成数据集上联合概率分布与原始数据集上的联合概率分布误差都很小. 当 f 值较大取 0.9 时, 即隐私保护很强时, 在小值域 $|\Omega| = 8$ 的情况下合成数据集上的联合概率分布仍可以大致显示原有分布趋势, AVD 值也较小, 效用性表现仍然可以接受; 而在值域较大的时候 ($|\Omega| = 16$ 或 32) 合成数据集的联合概率分布估计的偏差比较大, 对应 AVD 值也比较大. 而当 f 取值适中为 0.5 时, 隐私保护中等, 此时, 合成数据集的联合概率分布在不同值域上均体现出较好的原始分布特性, AVD 值也较小, 反映了较好的数据效用性.

图 9 分别展示了在 NLTCS, Adult 和 TPC-E 3 个数据集上, 以节点入度 $k = 2$ 为标准构建 Bayes 网络, 选取不同 f 得到的合成数据集与原始数据集进行 a 维联合概率分布查询的误差. 其中, 图 9(a)~(c) 显示了基于平均偏离误差 AVD (average variation distance) 的误差, 图 9(d)~(f) 显示了对应的 KL 散度值变化. 在此, 我们主要比较了 2 维至 5 维联合概率, 用 Q_a 表示 a 维联合概率. 由实验结果可看出, 整体来说, 在每个查询 Q_a 下, 随着 f 取值的增大, 平均偏差误差与 KL 散度均是呈上升趋势的, 且 f

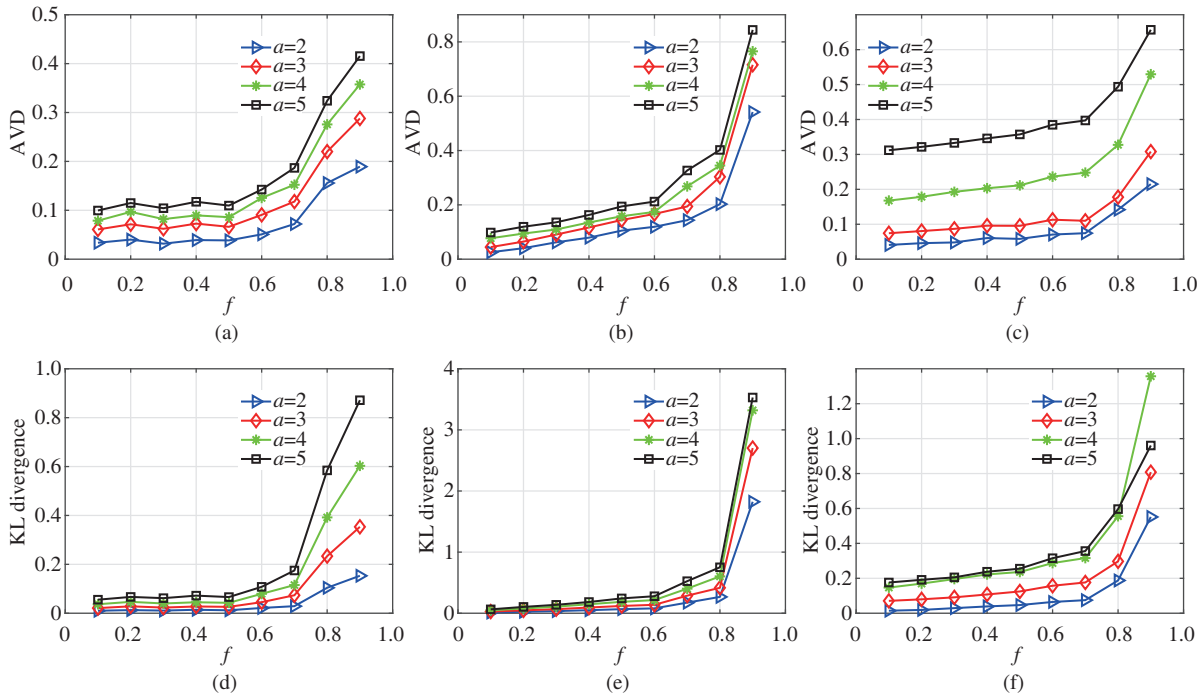


图 9 (网络版彩图) a 维联合概率分布估计误差

Figure 9 (Color online) Error of a-way joint probability estimation. (a) and (d) NLTCS ($k = 2$); (b) and (e) Adult ($k = 2$); (c) and (f) TPC-E ($k = 2$)

较大时, 会急剧增加. 这是因为, f 的取值越大, 用户隐私保护程度越高, 以越大的概率发送模糊的数据, 与原始数据的分布误差越大, 反映了隐私保护与数据效用性之间的折衷关系. 同时, 可以看出, 相同的隐私保护程度 f 下, 随着查询维度 a 从 2 维增加到 5 维, Q_a 对应的平均偏离误差及 KL 散度也会明显增大, 这是因为随着查询维度增大, 对应多维组合的状态空间变得越来越稀疏, 多维联合概率分布估计的误差也会随之增大, 数据效用性丢失明显. 这也解释了为保证本地隐私保护后仍然能恢复较好的数据效用性, 有必要对高维数据进行降维处理的原因, 这与文中的分析结论相一致.

6.2.3 分类准确度

本小节实验评估在合成数据集上进行多维数据分析的准确度. 分别在 3 个原始数据集上训练多个 SVM 分类器并得其平均测试准确率, 然后针对每个数据集上的合成数据集, 本文同样训练相同数目的 SVM 分类器, 比较两者分类预测的准确度. 在此实验中, 将各个数据集中 80% 的记录取作训练集, 另 20% 的记录作为测试集, 依次将数据集的每个二值属性作为分类标签, 训练多个分类器, 其中每个分类任务仿真 5 次, 计算平均分类准确率.

图 10 描述了在 NLTCS, Adult 和 TPC-E 3 个数据集上分别进行 SVM 分类的平均准确度. 可以看出, 随着 f 的增大, 分类准确度呈现下降趋势, 这也反映出隐私保护带来的数据效用性的折衷. 当 f 比较小 ($f < 0.5$) 的时候, 分类准确度比较高且接近原始数据的准确度; 而当隐私保护程度适当 (如 $f = 0.5$) 时, 合成数据集上的 SVM 分类准确度依然相对比较高且比较接近无隐私保护的情况, 这是因为 SVM 仅仅考虑二值属性, 而二值属性一般不太稀疏所以其概率统计准确性高且相关性不易丢失. 从整体上而言, 本文中基于本地隐私保护后的高维数据在一定程度下保留了较好的数据效用性.

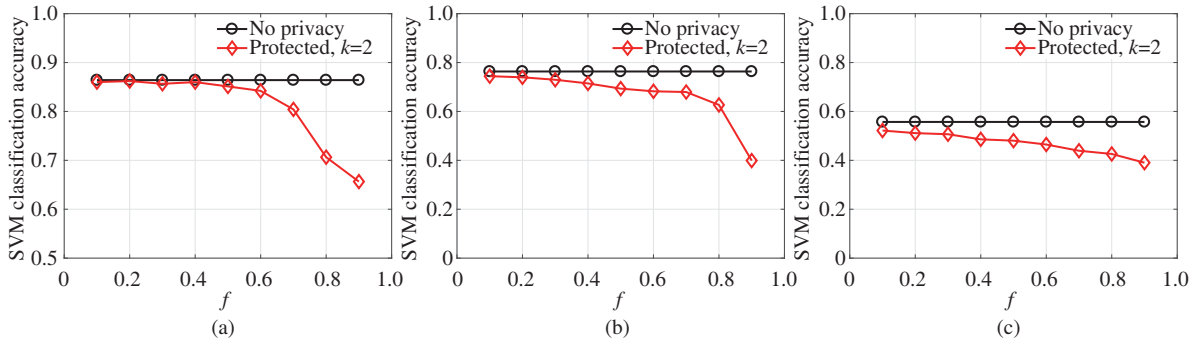


图 10 (网络版彩图) SVM 分类准确率

Figure 10 (Color online) SVM classification accuracy. (a) NLTCS; (b) Adult; (c) TPC-E

7 总结

在群智感知系统中, 为了实现具有本地差分隐私保护的高维感知数据发布, 本文首先讨论了已有的本地隐私保护技术和高维数据隐私保护方法, 并在此基础上提出了基于 Bayes 网络的高维感知数据本地隐私保护发布机制. 在该机制下, 一方面在用户端对每个用户数据进行本地差分隐私保护保证隐私; 另一方面, 感知服务器首先接收和汇集各个用户隐私保护后的数据, 其次根据保护后的数据进行低维联合概率分布估计和互信息计算以构建属性维度关联关系的 Bayes 网络, 然后将感知数据按照属性维度进行降维分块, 最后结合 Bayes 网络利用低维概率分布估计算法对本地隐私保护后的数据进行采样并合成新的数据集. 大量仿真实验验证, 该机制在满足隐私保护的同时可以实现高效用的数据发布, 特别地, 在合成数据集上进行多维联合概率分布查询和数据分类任务均具有接近原始数据的准确率.

参考文献

1 Guo B, Wang Z, Yu Z, et al. Mobile crowd sensing and computing: the review of an emerging human-powered sensing paradigm. ACM Comput Surv, 2015, 48: 1–31

2 Yurur O, Liu C H, Sheng Z, et al. Context-awareness for mobile sensing: a survey and future directions. IEEE Commun Surv Tut, 2016, 18: 68–93

3 Li G L, Wang J N, Zheng Y D, et al. Crowdsourced data management: a survey. IEEE Trans Knowl Data Eng, 2016, 28: 2296–2319

4 Mohammed N, Chen R, Fung B, et al. Differentially private data release for data mining. In: Proceedings of ACM SIGKDD. New York: ACM, 2011. 493–501

5 Naveed M, Ayday E, Clayton E W, et al. Privacy in the Genomic Era. ACM Comput Surv, 2015, 48: 1–44

6 Kohavi R, Provost F. Applications of data mining to electronic commerce. Data Min Knowl Discov, 2001, 5: 5–10

7 Clarke R. What's 'privacy'. 2006. http://www.rogerclarke.com/DV/Privacy.html

8 Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name. 2013. ArXiv: 1304.7605

9 Zhou X, Demetriou S, He D, et al. Identity, location, disease and more: Inferring your secrets from android public resources. In: Proceedings of ACM CCS. New York: ACM, 2013. 1017–1028

10 Jin H M, Su L, Ding B L, et al. Enabling privacy-preserving incentives for mobile crowd sensing systems. In: Proceedings of IEEE ICDCS, Nara, 2016. 344–353

11 Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst, 2002, 10: 557–570

12 Lu R X, Liang X H, Li X, et al. EPPA: an efficient and privacy-preserving aggregation scheme for secure smart grid communications. IEEE Trans Parallel Distrib Syst, 2012, 23: 1621–1631

13 Mármol F, Sorge C, Ugus O, et al. Do not snoop my habits: preserving privacy in the smart grid. IEEE Commun Mag, 2012, 50: 166–172
14 Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proceedings of IEEE S&P, Washington, 2008. 111–125
15 Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT Theor Comput Sci, 2014, 9: 211–407
16 Dwork C. Differential privacy. In: Proceedings of ICALP. Berlin: Springer, 2006. 1–12
17 Acs G, Castelluccia C. I have a dream! (differentially private smart metering). In: International Workshop on Information Hiding. Berlin: Springer, 2011. 118–132
18 Zhu T Q, Xiong P, Li G, et al. Correlated differential privacy: hiding information in non-IID data set. IEEE Trans Inform Forensic Secur, 2015, 10: 229–242
19 McSherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of ACM SIGMOD. New York: ACM, 2009. 19–30
20 Zhang J, Cormode G, Procopiuc C M, et al. PrivBayes: private data release via Bayesian networks. In: Proceedings of ACM SIGMOD. New York: ACM, 2014. 1423–1434
21 Chen R, Xiao Q, Zhang Y, et al. Differentially private high-dimensional data publication via sampling-based inference. In: Proceedings of ACM SIGKDD. New York: ACM, 2015. 129–138
22 Wu Y J, Ge C, Zhang L Q, et al. An algorithm for differential privacy streaming data publication based on matrix mechanism under exponential decay mode. Sci Sin Inform, 2017, 47: 1493–1509 [吴英杰, 葛晨, 张立群, 等. 指数衰减模式下基于矩阵机制的差分隐私流数据发布算法. 中国科学: 信息科学, 2017, 47: 1493–1509]
23 Su D, Cao J N, Li N H, et al. PrivPfC: differentially private data publication for classification. VLDB J, 2018, 27: 201–223
24 Zhu T Q, Li G, Zhou W L, et al. Differentially private data publishing and analysis: a survey. IEEE Trans Knowl Data Eng, 2017, 29: 1619–1638
25 Qardaji W, Yang W N, Li N H. PriView: practical differentially private release of marginal contingency tables. In: Proceedings of ACM SIGMOD. New York: ACM, 2014. 1435–1446
26 Wang L, Wang W P, Meng D. Privacy preserving data publishing via weighted Bayesian networks. J Comput Res Develop, 2016, 53: 2343–2353 [王良, 王伟平, 孟丹. 基于加权贝叶斯网络的隐私数据发布方法. 计算机研究与发展, 2016, 53: 2343–2353]
27 Zhang X J, Chen L, Jin K Z, et al. Private high-dimensional data publication with junction tree. J Comput Res Develop, 2018, 55: 2794–2809
28 Ye Q Q, Meng X F, Zhu M J, et al. Survey on local differential privacy. J Soft, 2018, 29: 159–183 [叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述. 软件学报, 2018, 29: 159–183]
29 Erlingsson Ú, Korolova A, Pihur V. Rappor: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of ACM CCS. New York: ACM, 2014. 1054–1067
30 Chen R, Li H R, Qin A K, et al. Private spatial data aggregation in the local setting. In: Proceedings of IEEE ICDE, Piscataway, 2016. 289–300
31 Cormode G, Kulkarni T, Srivastava D. Marginal release under local differential privacy. In: Proceedings of ACM SIGMOD. New York: ACM, 2018. 131–146
32 Wang N, Xiao X K, Yang Y, et al. Privtrie: effective frequent term discovery under local differential privacy. In: Proceedings of IEEE ICDE, Piscataway, 2018. 821–832
33 Ye Q Q, Hu H B, Meng X F, et al. Privkv: key-value data collection with local differential privacy. In: Proceedings of IEEE S&P, Piscataway, 2019. 1–15
34 Zhang Z K, Wang T H, Li N H, et al. CALM: consistent adaptive local marginal for marginal release under local differential privacy. In: Proceedings of ACM CCS. New York: ACM, 2018. 212–229
35 Xiong S J, Sarwate A D, Mandayam N B. Randomized requantization with local differential privacy. In: Proceedings of IEEE ICASSP, Shanghai, 2016. 2189–2193
36 Sarwate A D, Sankar L. A rate-disortion perspective on local differential privacy. In: Proceedings of IEEE Allerton

Conference on Communication, Control, and Computing, Monticello, 2015. 903–908
37 Warner S L. Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc, 1965, 60: 63–69
38 Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy. In: Proceedings of ICML, 2016. 2436–2444
39 Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy. In: Proceedings of NIPS. Cambridge: MIT Press, 2014. 2879–2887
40 Wang S W, Huang L S, Wang P Z, et al. Mutual information optimally local private discrete distribution estimation. 2016. ArXiv: 1607.08025
41 Qin Z, Yu T, Yang Y, et al. Generating synthetic decentralized social graphs with local differential privacy. In: Proceedings of ACM CCS. New York: ACM, 2017. 425–438
42 Qin Z, Yang Y, Yu T, et al. Heavy hitter estimation over set-valued data with local differential privacy. In: Proceedings of ACM CCS. New York: ACM, 2016. 192–203
43 Wang P Y, Zhang H. Distributed Logistic regression with differential privacy. Sci Sin Inform, 2019. doi: 10.1360/N112018-00214 [王璞玉, 张海. 分布式隐私保护 -Logistic 回归. 中国科学: 信息科学, 2019. doi: 10.1360/N112018-00214]
44 Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates. In: Proceedings of IEEE FOCS, Washington, 2013. 429–438
45 Fanti G, Pihur V, Erlingsson Ú. Building a RAPPOR with the unknown: privacy-preserving learning of associations and data dictionaries. Proc Privacy Enhancing Technol, 2016, 2016: 41–61
46 Ren X B, Yu C-M, Yu W R, et al. Lopub: high-dimensional crowdsourced data publication with local differential privacy. IEEE Trans Inform Forensic Secur, 2018, 13: 2151–2166
47 Zhang J, Cormode G, Procopiuc C M, et al. PrivBayes: private data release via Bayesian networks. ACM Trans Database Syst, 2017, 42: 25
48 Acs G, Castelluccia C, Chen R. Differentially private histogram publishing through lossy compression. In: Proceedings of IEEE ICDE, Washington, 2012. 1–10

Bayesian network-based high-dimensional crowdsourced data publication with local differential privacy

Xuebin REN[1], Jingyi XU[1], Xinyu YANG[1*] & Shusen YANG[2]

1. *School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an* 710049*, China*;
2. *School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an* 710049*, China*
* Corresponding author. E-mail: yxyphd@mail.xjtu.edu.cn

Abstract High-dimensional crowdsourced data is pervasive in crowdsensing systems and with the development of IoTs it can produce rich knowledge for the society. However, it also creates serious privacy threats for crowdsourcing participants. To mitigate the privacy concerns in crowdsensing systems, local differential privacy has been derived from the de facto standard of differential privacy in order to achieve strong privacy guaranteed in distributed systems. However, directly achieving local differential privacy on high-dimensional crowdsourced data may lead not only to a prohibitive computational burden but also low data utility. Therefore, in this paper, we propose a local private high-dimensional data publication scheme for crowdsensing systems. In particular, on the participants' side, high-dimensional records are locally perturbed to protect privacy, while on the server's side, the probability distribution of original data is recovered by taking advantage of both the expectation maximization algorithm and the theory of the Bayesian network. Extensive experiments on real-world datasets demonstrated the effectiveness of the proposed scheme that can synthesize approximate datasets with local differential privacy.

Keywords crowd sensing system, crowdsourced data, high-dimensional data, local differential privacy, Bayesian network

Xuebin REN was born in 1989. He received his Ph.D. degree from the Department of Computer Science and Technology from Xi'an Jiaotong University (XJTU), China, in 2017. He is currently a lecturer at the School of Computer Science and Technology in XJTU. He was a visiting Ph.D. student at the Department of Computing in Imperial College London from 2016 to 2017. He is currently looking at the security and privacy issues of big data analysis and machine learning in distributed and edge computing systems such as the Internet of Things and cyber-physical systems.

Xinyu YANG was born in 1973. He received his diploma in Computer Science and Technology from Xi'an Jiaotong University (XJTU), China, in 2001 and his B.S., M.S., and Ph.D. degrees from XJTU in 1995, 1997, and 2001, respectively. He is currently a professor at the School of Computer Science and Technology in XJTU. His research interests include wireless communication, mobile ad hoc networks, and network security and privacy.

Jingyi XU was born in 1996. She received her B.S. degree from the Department of Computer Science and Technology, Xi'an Jiaotong University (XJTU), China, in 2017. She is currently an M.S. candidate at the School of Computer Science and Technology in XJTU. Her research interests include privacy-preserving algorithm design and differential privacy.

Shusen YANG was born in 1984. He received his Ph.D. in computing from Imperial College London in 2013. He is currently a professor at the Institute of Information and System Science at Xi'an Jiaotong University (XJTU). Before joining XJTU, he worked as a lecturer (assistant professor) at the University of Liverpool from 2015 to 2016, and a research associate at Imperial College and Intel Collaborative Research Institute for sustainable connected cities from 2013 to 2014.