计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):225−235 [doi: 10.15888/j.cnki.csa.008733]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于属性分割的差分隐私异构多属性数据发布①

张小玉[1,2], 沈国华[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](南京大学 软件新技术与产业化协同创新中心, 南京 210093)
通信作者: 张小玉, E-mail: zxy4041@nuaa.edu.cn

摘　要: 针对现有多属性数据隐私发布方法无法兼顾属性的敏感性差异和计算效率低的问题, 提出了一种基于属性分割的差分隐私异构多属性数据发布方法 HMPrivBayes. 首先, 设计了满足差分隐私的谱聚类算法分割原始数据集, 其中相似矩阵的生成借助于属性最大信息系数. 其次, 借助属性信息, 该方法使用满足差分隐私的改进贝叶斯网络构建算法分别为每个数据子集构建贝叶斯网络. 最后, 以属性归一化风险熵为权重分配隐私预算, 对贝叶斯网络提取的属性联合分布添加异构噪声扰动, 实现了异构多属性数据保护. 实验结果表明, HMPrivBayes 可以在减少注入合成数据集中噪声量的同时, 提高合成数据计算效率.

关键词: 差分隐私; 异构多属性数据发布; 谱聚类; 属性分割; 贝叶斯网络; 隐私保护

引用格式: 张小玉,沈国华,杨阳.基于属性分割的差分隐私异构多属性数据发布.计算机系统应用,2022,31(10):225–235. http://www.c-s-a.org.cn/1003-3254/8733.html

Differentially Private Heterogeneous Multi-attribute Data Publication via Attribute Segmentation

ZHANG Xiao-Yu[1,2], SHEN Guo-Hua[1,2,3], 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 University, Nanjing 210093, China)

Abstract: Multi-attribute data privacy publication fails to balance the difference in attribute sensitivity and computational efficiency. For this reason, HMPrivBayes, a heterogeneous multi-attribute data publishing method with differential privacy based on attribute segmentation, is proposed. Firstly, the spectral clustering algorithm satisfying differential privacy is designed to segment the original data set, in which the similarity matrix is generated by the attribute maximum information coefficient. Secondly, with the help of attribute information, this method uses an improved Bayesian network construction algorithm to build Bayesian networks for each data subset. Finally, HMPrivBayes adds heterogeneous noise disturbance to the attribute joint distribution extracted from the Bayesian network to realize the protection of heterogeneous multi-attribute data, in which privacy budget is allocated based on the normalized risk entropy of attribute. The experimental results show that HMPrivBayes not only reduces the added noise but also improves the computational efficiency of synthetic data.

Key words: differential privacy; heterogeneous multi-attribute data publishing; spectral clustering; attribute segmentation; Bayesian network; privacy protection

① 基金项目: 国家自然科学基金 (61772270)
收稿时间: 2022-01-06; 修改时间: 2022-02-17; 采用时间: 2022-02-22; csa 在线出版时间: 2022-07-07

1 引言

近年来, 互联网应用的逐步普及, 带来便利的同时, 大量数据信息通过个人用户、企业单位、研究机构等源源不断地产生. 对产生的数据进行数据分析, 挖掘事物潜在的关联, 再反馈给互联网应用商, 可以为大众提供更好的服务. 然而, 这些数据中不乏个人隐私数据, 诸如医疗信息、金融信息、轨迹数据等, 直接向外界发布原始数据存在着巨大的隐私泄露风险. 因此, 对外发布数据时, 需要采取一些隐私保护措施来保障数据安全.

K-匿名[1] 及其相关扩展 L-Diversity[2]、T-closeness[3] 是隐私保护的重要方法, 其主要思想是将待发布数据划分为多个等价类, 等价类内的 k 个实体无法相互区分, 通过限制数据发布实现数据隐私保护. 然而, 一旦攻击者具有相当的背景知识, K-匿名就无法提供很好的隐私保护效果, 而且无法就其隐私保护程度进行定量分析. 差分隐私[4–9] 作为一种严格可证的新兴隐私保护模型, 其前提是假设数据攻击者具有的背景知识最大化. 通过向输出中添加随机扰动, 尽可能减小单个记录对输出的影响, 保证攻击者无法通过观察不同的输出结果重构真实的数据信息, 实现了单个记录是否参与数据集的不可区分性.

从已有的研究可以看出, 差分隐私在低维数据发布方面做了诸多努力, 但在实际场景中, 更为常见的却是多属性数据集, 即高维数据集. 如果将以往的低维数据发布算法直接应用于多属性数据发布, 存在查询敏感度大、隐私预算消耗过快等问题, 导致数据发布效用低.

针对高维数据发布场景, 常见方法有: 1) 根据树结构划分高维数据集, 通过平衡分区的近似误差和数据的扰动误差来提高数据发布精度; 2) 通过构建数据集的概率图模型计算属性间的条件分布, 通过噪声条件分布近似原始数据集的联合分布, 以此生成新的合成数据集, 避免直接对原始数据扰动产生巨大的扰动误差; 3) 利用数据降维技术, 将高维数据转化成近似的低维数据, 对转化后的低维数据添加噪声扰动, 减少扰动噪声量.

在利用树结构发布高维数据集的研究中, 文献 [10] 借助细粒度单元划分和 kd-树划分两种分区策略, 实现了差分隐私多维直方图发布. 文献 [11] 利用 kd-树和 quad-tree 设计混合数据结构实现多维数据划分, 从而减少了注入的噪声量. 文献 [12] 提出了差分隐私泛化数据发布算法 DiffGen, 通过构建决策树的过程完成数据的差分隐私保护. 基于概率图模型的研究也有所成果, 文献 [13] 提出数据发布方案 PrivBayes, 该方案通过构建原始数据集的贝叶斯网络模型, 分析原始数据属性之间的依赖关系, 学习提取近似噪声条件分布, 采样生成新的合成数据集. 文献 [14] 提出联合树算法 Jtree, 根据样本数据属性之间的关联构建依赖关系图, 结合联合树的推理基础, 从边缘关系图中学习一组噪声边缘表, 以此近似原始数据集的联合分布. 通过数据降维来近似高维数据发布的研究中, 文献 [15,16] 利用主成分分析技术识别高维数据的近似低维, 实现了高维数据向低维的转化, 并基于低维数据实现差分隐私保护. 文献 [17] 提出了差分隐私数据发布算法 DPPro, 该算法利用随机投影技术, 完成高维数据在低维上的映射, 保证了高维数据发布效用. 文献 [18] 提出 PriView 方法, 通过构建一组含噪声扰动的低维 View 视图, 提取α-边际分布, 生成合成数据集.

然而, 前文提到的各种方法对不同属性的隐私保护处理大多隐含地假设同构性, 即不考虑属性数据敏感性的异构, 不同属性的随机干扰程度均相同. 由于不同属性携带的信息量不同, 从而对攻击者推理目标对象隐私信息的贡献不同, 即隐私泄露风险与属性敏感性呈正相关. 例如, 性别属性只有两个属性值, 与年龄属性相比, 携带的信息量更少, 对目标对象的刻画程度较浅, 对攻击者推理敏感信息的贡献较少, 需要的隐私保护强度相对较低, 相应的属性敏感性也就较低. 不考虑属性信息量差异带来的敏感性异构, 会导致没有充分保护高敏感属性数据, 或者过度丢失低敏感属性数据的信息.

文献 [19] 提出了一种加权贝叶斯网络数据发布方法, 虽然该方法是根据属性的多样性分配权重, 但其目标是借助属性权重构建更好的贝叶斯网络. 文献 [20,21] 均在基于贝叶斯网络的数据发布算法中引入属性聚类, 以属性关联强弱分割属性集, 在减少隐私预算分割次数的基础上提高算法计算效率. 然而, 两者在隐私预算的分配上依然采取等分措施, 没有兼顾属性敏感性差异.

针对已有方法的不足之处, 本文提出了一种基于属性分割的差分隐私异构多属性数据发布算法 HMPriv-Bayes (heterogeneous multi-attribute private data

publishing via Bayesian networks), 主要贡献包括以下 3 个方面:

1) 根据属性之间的关联程度, 引入最大信息系数量化属性之间的相关关系, 设计满足差分隐私的谱聚类算法分割属性集, 能在一定程度上缩减贝叶斯网络结构空间, 提高算法计算效率.

2) 针对构建贝叶斯网络随机选取首个属性可能降低数据发布质量的问题, 以属性信息熵为权重选择属性, 尽可能提高数据发布质量, 保留数据实用价值.

3) 考虑到属性携带的信息量的不同, 基于贝叶斯网络提取属性条件分布时, 以属性归一化风险熵为权重分配隐私预算, 实现异构多属性保护.

2 基础知识

2.1 差分隐私

差分隐私 (differential privacy, DP) 保护模型保证在数据集中对一条记录进行增、删、改操作后, 查询返回的扰动输出没有明显差异, 以至于攻击者不能推断出目标记录是否在发布的数据集中, 即数据集中单个记录对输出的影响有限, 从而实现保护个体隐私.

定义 1. ε-差分隐私[4]. 设D_1和D_2为任意两个相邻数据集, 即它们彼此相差 1 条记录, $Range(M)$是随机算法M的取值范围, 若算法M在数据集D_1和D_2上的任意输出$S \in Range(M)$, 满足:

$$\frac{Pr(M(D_1) \in S)}{Pr(M(D_2) \in S)} \leqslant \exp(\varepsilon) \tag{1}$$

则称算法M满足ε-差分隐私. 其中, 隐私预算ε与隐私保护程度呈负相关, 与隐私泄露风险呈正相关. $Pr(M(D_1) \in S)$和$Pr(M(D_2) \in S)$表示算法M在数据集D_1和 D_2上输出为 S的概率.

任何满足定义 1 的噪音机制都可以视为差分隐私机制, 差分隐私中常见噪音机制有 Laplace 机制[22] 和指数机制[23], 其噪声扰动大小与随机函数的敏感度以及隐私预算参数密切相关.

定义 2. 全局敏感度[5]. 设查询函数$f: D \to \mathbb{R}^n$, f的全局敏感度定义如下:

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\| \tag{2}$$

其中, D_1和D_2表示任意两个相邻数据集, 全局敏感度Δf由查询函数f决定.

定义 3. Laplace 机制[22]. 对任意数据集D和查询函数$f: D \to \mathbb{R}^n$, 若算法M的输出结果满足:

$$M(D) = f(D) + Lap(\Delta f/\varepsilon) \tag{3}$$

则算法M实现了ε-差分隐私. 其中, $Lap(\Delta f/\varepsilon)$表示添加的噪声量, 噪声量与$\Delta f$成正比, 与$\varepsilon$成反比. Laplace 机制主要用于处理数值型数据.

定义 4. 指数机制[23]. 对任意数据集D, 给定评分函数$u: (D, r) \to R$, 若算法M满足:

$$M(D, u) = \left\{ r | Pr[r \in O] \propto \exp\left(\frac{\varepsilon u(D, r)}{2\Delta u} \right) \right\} \tag{4}$$

则算法M实现了ε-差分隐私. 其中, Δu为$u(D, r)$的全局敏感度, $Pr[r \in O]$为评分函数u输出r的概率, 评分越高, r被输出的概率越高. 指数机制通常用于处理非数值型数据.

对于一组满足差分隐私的随机独立算法, 差分隐私的串行组合性质和并行组合性质[24] 可以用于证明整体算法满足差分隐私保护.

性质 1. 串行组合[24]. 给定数据集D和D上 1 组随机独立算法$M_1(D), M_2(D), \cdots, M_m(D)$, 其中每个算法$M_i(D)$满足$\varepsilon_i$-差分隐私, 则$M = \{M_1, M_2, \cdots, M_m\}$满足$\sum_{i=1}^{m} \varepsilon_i$-差分隐私.

串行组合表明, 当一系列差分隐私机制应用于同一数据集时, 隐私预算和噪声数量呈线性累积.

性质 2. 并行组合[24]. 给定数据集D, 将D分割成m个互不相交的子集$D = \{D_1, D_2, \cdots, D_m\}$. 每个子集中的随机独立算法$M_1(D_1), \cdots, M_m(D_m)$均满足$\varepsilon_i$-差分隐私, 则组合算法$M = \{M_1, M_2, \cdots, M_m\}$满足$\max(\varepsilon_i)$-差分隐私.

并行组合表明, 当一系列差分隐私机制应用于数据集的不同子集时, 隐私保护的程度取决于ε_i的最大值.

2.2 谱聚类

谱聚类 (spectral clustering, SC)[25] 是利用图论中的图分割完成数据聚类的算法. 它的主要思想是先将数据集映射成一张带权无向图, 基于两点之间的距离计算顶点相似度, 通过最优划分准则分割图中结点, 最大化子图内部的相似度, 实现数据聚类. 标准的谱聚类算法实现流程如算法 1 所示.

算法 1. 谱聚类算法

输入: 数据集$D=\{x_1,x_2,\cdots,x_n\}$, 降维后的维度 k_1, 聚类方法, 簇数量 k
输出: 簇划分$C(c_1,c_2,\cdots,c_k)$

1. 计算 $n\times n$ 的相似矩阵 S: $s_{ij}=\|x_i-x_j\|_2$;
2. 根据相似矩阵 S 构建邻接矩阵 W: $w_{ij}=\exp\left(\frac{-s_{ij}}{2\sigma^2}\right)$
3. 计算 $n\times n$ 的度矩阵 B: $b_{ij}=\begin{cases}\sum_{k=1}^{d} w_{ik}, i=j\\ 0, i\neq j\end{cases}$;
4. 计算并标准化拉普拉斯矩阵$L=B^{-1/2}(B-W)B^{-1/2}$;
5. 计算 L 最小的 k_1 个特征值以及各自对应的特征向量$u_1, u_2, \cdots, u_{k_1}$;
6. 将特征向量组成矩阵并按行进行标准化, 最终组成 $n\times k_1$ 维的特征矩阵 U;
7. 将 U 中的每一行作为一个 k_1 维的样本, 按照输入的聚类方式进行聚类, 聚类维数为 k;
8. 返回簇划分$C(c_1, c_2, \cdots, c_k)$.

在本文中, 谱聚类主要用于属性聚类, 即根据属性之间的关联程度分割属性集. 本文选用最大信息系数 (maximal information coefficient, MIC) [26] 度量属性之间关联程度. 在此之前, 互信息是常用的属性关联度量手段, 然而, 互信息需要考虑连续属性的离散化, 而且不同数据集上的互信息计算结果无法相互比较. MIC 利用网格划分属性值域, 结合属性互信息, 可以更准确地识别出大数据集中属性间的线性或非线性关系, 以及非函数依赖关系.

定义 5. 最大信息系数[26]. 给定属性X, Y和有序对$<a,b>$, 样本数量为N, 将当前二维空间在x, y方向上划分为$a\times b$的网格G, 将属性数据点离散落入网格G中, 属性X和Y的最大信息系数计算如下:

$$MIC(X,Y)=\max_{a\times b<B}\frac{I(X,Y)}{\log_2\min(a,b)} \tag{5}$$

其中, B为网格划分$a\times b$的上限值, 通常取值$N^{0.6}$或$N^{0.55}$, $I(X, Y)$为属性 X 和 Y 的互信息值, 具体计算如下:

$$I(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}Pr(x_i,y_j)\log\frac{Pr(x_i,y_j)}{Pr(x_i)Pr(y_j)} \tag{6}$$

其中, $Pr(x_i,y_j)$表示属性X, Y取值x_i, y_j的联合分布.

2.3 贝叶斯网络

贝叶斯网络 (Bayesian network) 是常用的概率图模型之一, 用来表示一组随机变量 (属性) 之间的依赖关系. 形式上, 贝叶斯网络N可以表示为$N=(G, \theta)$. 其中, G表示有向无环图, 图中节点代表随机变量, 节点之间的有向边代表随机变量之间的依赖关系. 在G中, 若存在一条从X_j到X_i的有向边, 则称X_j为X_i的父节点. X_i的所有父节点构成X_i的父节点集合, 表示为$\prod_i$. θ表示N的参数集合, θ包含每个节点X_i的条件分布, 表示为$Pr[X_i|\prod_i]$. 对于一个包含属性集$X=\{X_1,X_2,\cdots,X_d\}$的贝叶斯网络N, N可以近似表示X的联合概率分布, 具体为$Pr_N[X]=\prod_{i=1}^{d}Pr[X_i|\prod_i]$. 图 1 表示包含属性节点$\{X_1,X_2,\cdots,X_5\}$的贝叶斯网络, 图中所有属性节点的联合概率分布为$Pr(X_1,X_2,X_3,X_4,X_5)=Pr(X_1)Pr(X_2|X_1)Pr(X_3|X_2)Pr(X_4|X_1,X_2)Pr(X_5|X_3,X_4)$.

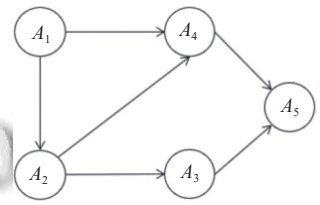


图 1 含 5 个节点的贝叶斯网络示例

3 基于属性分割的差分隐私异构多属性数据发布

3.1 HMPrivBayes 算法概述

HMPrivBayes 算法的整体流程如图 2 所示, 包括 4 个阶段: 属性聚类划分、构建差分隐私贝叶斯网络、生成属性加噪条件分布、生成待发布数据集. 为了方便描述, 本文假设所有属性均是二进制属性. 对于非二进制属性, 利用文献 [13] 提出的等宽法将非二进制属性转化成一组二进制属性. 由于等宽法的转化过程只依赖非二进制属性的值域, 且属性的值域是公开信息, 不涉及隐私数据. 因此, 属性转化过程不存在隐私泄露问题.

HMPrivBayes 算法的实现过程如算法 2 所示.

算法 2. HMPrivBayes 算法

输入: 数据集$D=\{x_1,x_2,\cdots,x_n\}$, 属性集$X=\{X_1,X_2,\cdots,X_d\}$, 属性簇个数 k, 隐私预算 ε
输出: 数据集 D 满足差分隐私的合成数据集 D'

1. 初始化 $D'=\varnothing$;
2. 分配隐私预算 $\varepsilon_1+\varepsilon_2+\varepsilon_3=\varepsilon$;
3. 执行算法 3, 返回聚类结果$C=\{C_1,C_2,\cdots,C_k\}$;
4. 依据 C 划分原始数据集$D=\{D_1,\cdots,D_k\}$;
5. for i=1 to k do
   执行算法 4, 求贝叶斯网络 N_i;
   基于P_i^*计算 N_i 的带噪联合分布, 采样得到合成数据集D_i', 并加入D';
   执行算法 5, 求带噪条件分布P_i^*;
   end for
6. 返回D'.

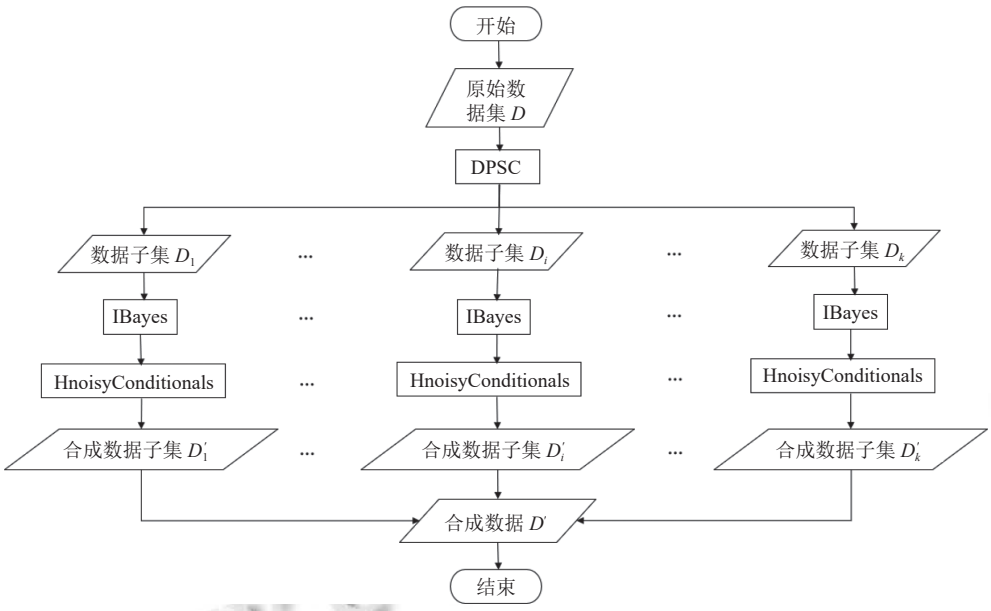


图 2 HMPrivBayes 算法流程图

3.2 差分隐私谱聚类算法

谱聚类算法是根据数据点之间相似性程度进行数据集聚类, 对于关联程度不高的属性对, 其相似值相应较小, 从而分割至不同的子集. 本文采用最大信息系数作为属性之间关联程度衡量指标, MIC 不仅能高效检测大数据集中不同类型属性之间关联关系, 而且不需要进行归一化处理. 选用 K-means++聚类算法实现特征矩阵U的聚类划分, K-means++算法选择初始的聚类中心时, 要求各聚类中心之间的相互距离要尽可能的远. 初始点选择的优化, 使得 K-means++算法在聚类结果准确性上有较大提升. 差分隐私谱聚类算法 DPSC 如算法 3 所示.

算法 3. DPSC 算法

输入: 数据集$D=\{x_1,x_2,\cdots,x_n\}$, 属性集$X=\{X_1,X_2,\cdots,X_d\}$, 尺度参数σ, 降维后的维度k_1, 聚类后的维度k, 隐私预算ε_1

输出: 属性集X聚类划分结果$C'=\{C_1,C_2,\cdots,C_k\}$

1. 计算$d\times d$的相似矩阵S: $s_{ij}=\begin{cases} MIC(X_i,X_j)+Lap\left(\frac{\Delta m\cdot d(d-1)}{\varepsilon_1\cdot(2\lfloor d/2\rfloor)}\right), i\neq j \\ 0, i=j \end{cases}$;

2. 执行算法 1 的步骤 2–步骤 6;

3. 使用 K-means++算法对新样本点$U=\{u_1,u_2,\cdots,u_{k_1}\}$进行聚类划分得到划分结果$C$;

4. 返回属性聚类结果$C'=\{C_1,C_2,\cdots,C_k\}$.

由于在计算属性关联程度时涉及到了原始数据集D的使用, 为了保护数据集中的数据隐私不被泄露, 采用 Laplace 机制对计算结果添加噪声.

首先计算属性间的最大信息系数的全局敏感度Δm, 由式 (5) 可知, MIC 的计算取决于属性间的互信息, 故计算相似矩阵S的全局敏感度$\Delta m=\Delta I$. 由文献 [13] 可知, $\Delta I=\frac{1}{n}\cdot\log n+\frac{n-1}{n}\cdot\log\frac{n}{n-1}$.

从算法 3 可以看出, Laplace 机制是通过多次扰动加噪过程实现的, 每次加噪都需要消耗部分隐私预算ε_1. 由定义可知, 相似矩阵S是一个对称矩阵, 且对角线元素不需要计算, 所以计算$d(d-1)/2$个不同元素的值就可以得到矩阵S. 每次访问数据集D可以计算$\lfloor d/2\rfloor$对属性对之间的 MIC, 因此, 对数据集D的访问次数, 即扰动总次数为$d(d-1)/(2\lfloor d/2\rfloor)$, 每次扰动添加的噪声量为$Lap((\Delta m\cdot d(d-1))/(\varepsilon_1\cdot(2\lfloor d/2\rfloor)))$. 由定义 3 和性质 1 可知, 算法 3 满足ε_1-差分隐私.

3.3 差分隐私贝叶斯网络算法

在基于贝叶斯网络构建数据发布模型的现有研究中, 大多以文献 [13] 中的 PrivBayes 为参照进一步改进贝叶斯网络的构建设计发布算法[19–21]. 由于 PrivBayes 在构建贝叶斯网络时, 随机选取首个属性加入父节点集合, 这样构建的贝叶斯网络不一定能真实反映原始数据集. 与此同时, 通过枚举的方式选取互信息最大的$(X_i,\prod_i)$对会造成大量无意义的计算, 影响贝叶斯网络的构建速度. 基于此, 本文提出改进的贝叶斯网络构建算法 IBayes, IBayes 引入信息熵作为选取首个属性的

依据, 同时缩减候选属性的$(X_i,\prod_i)$对, 以此达到快速构建“最合适”的贝叶斯网络的目的. IBayes 的实现过程如算法 4 所示.

算法 4. IBayes 算法

输入: 数据子集$D_i=\{x_1,x_2,\cdots,x_n\}$, 属性子集$C_i=\{X_1,X_2,\cdots,X_{|C_i|}\}$, 最大父节点数$l$, 隐私预算$\varepsilon_2$

输出: 数据子集D_i的贝叶斯网络N_i

1. 初始化$N_i=\varnothing$, $V=\varnothing$;
2. 使用指数机制从C_i中选取信息熵最大的属性X_j, 将$(X_j,\varnothing)$加入N_i, 将X_j加入V;
3. for t=2 to $|C_i|$ do
   初始化$\Omega=\varnothing$;
   for $X_s\in C_i\backslash V$
   令$V'=V$;
   If $S_i(X_s,X_c)$=0 且 $X_c\in V'$
   将X_c从V'中移除, 且$\Pi_s\in\binom{V'}{l}$;
   将$(X_s,\prod_s)$加入Ω;
   end for
   使用指数机制从Ω中选择互信息最大的(X_t,Π_t)加入N_i, 同时X_t加入V;
  end for
4. 返回N_i.

信息熵是对信息不确定性的度量, 与信息的不确定性呈正相关. 信息熵越大, 携带的信息量越多, 能更多地反映出数据的多样性, 增加数据的实用价值, 反之亦然. 因此, 引入信息熵作为选取首个属性的依据是合理的.

对于数据集D_i中的每个属性X_j, 其信息熵$H(X_j)$计算如下:

$$H(X_j)=-\sum_{j=1}^{n}p_j\log p_j \tag{7}$$

根据式 (7) 可知, 当取不同值的概率相等时信息熵最大, 即X_j中有n个不同的取值时$H(X_j)$取得最大值, 为$H(X_j)=-\sum_{s=1}^{n}1/n\cdot\log 1/n=\log n$.

通过图 3 的实例可以说明属性信息熵的最大差异.

	0	1
p_j	0	1

(a) 实例 1

	0	1
p_j	$\frac{1}{n}$	$\frac{n-1}{n}$

(b) 实例 2

图 3 信息熵差异实例

根据图 3(a) 取值概率分布计算$H(X_j)$的值为 0, 通过图 3(b) 计算$H(X_j)$的值为$\frac{1}{n}\cdot\log n+\frac{n-1}{n}\cdot\log\frac{n}{n-1}$. 在此基础上, 可以得出属性信息熵的全局敏感度$\Delta H$为:

$$\Delta H=\frac{1}{n}\cdot\log n+\frac{n-1}{n}\cdot\log\frac{n}{n-1} \tag{8}$$

PrivBayes 通过枚举选择互信息最大的属性对时, 候选$(X_t,\prod_t)$对越多, 对算法的时间复杂的影响越大. 因此, 本文在求解候选属性的父节点集时, 筛选关联程度弱的属性对, 减少$(X_t,\prod_t)$对互信息的计算量. 筛选操作通过关联矩阵S_i实现. 从相似矩阵S中提取出相关属性子集的最大信息系数, 根据式 (9) 得到S_i:

$$S_i(X_s,X_c)=\begin{cases}1, \text{ else}\\ 0, \begin{cases}S(X_s,X_c)\leqslant\eta\times\max S(X_s)\\ S(X_c,X_s)\leqslant\eta\times\max S(X_c)\end{cases}\text{ or}\end{cases} \tag{9}$$

IBayes 算法使用指数机制选取加入N_i中的首个属性节点, 同时从Ω中选择$(X_t,\prod_t)$对加入N_i, 评分函数分别采用属性信息熵和互信息. 由于$\Delta u=\Delta I=\Delta H$, 除选取加入$N_i$中的首个属性节点外, 选择$(X_t,\prod_t)$需要执行$|C_i|-1$次, 故将$\varepsilon_2$平均分为$|C_i|$份. 因此, 结合差分隐私指数机制, 选择$(X_t,\prod_t)$的概率表达式如式 (10):

$$P\left(X_t,\prod_t\right)=\frac{\exp\left(\frac{\varepsilon_2 u(X_t,\prod_t)}{2|C_i|\Delta u}\right)}{\sum_{(X_t,\prod_t)\in\Omega}\exp\left(\frac{\varepsilon_2 u\left(X_j,\prod_j\right)}{2|C_i|\Delta u}\right)} \tag{10}$$

IBayes 算法最终输出一个贝叶斯网络N_i, 除了使用指数机制选取首个属性节点和$(X_t,\prod_t)$对之外, 其他步骤均无需访问原始数据集D. 在构建贝叶斯网络之前, HMPrivBayes 算法已经通过算法 3 将原始数据集D划分为k个独立数据子集$D_1,\cdots,D_k$. 根据定义 4 和性质 1, 指数机制$\exp(\varepsilon_2 u(D_i,r)/(2|C_i|\Delta u))$满足$\varepsilon_2$-差分隐私, 因此 IBayes 算法满足$\varepsilon_2$-差分隐私.

3.4 差分隐私异构条件分布计算

通过贝叶斯网络计算属性的联合分布, 生成合成数据集, 可以近似原始数据集. 尽管贝叶斯网络的构建过程满足差分隐私保护, 但仍存在隐私泄露的可能. 因此, 需要对计算出来的属性联合分布添加噪声扰动, 以进一步保护隐私数据. 之前的研究均通过均分隐私预算实现联合分布噪声扰动, 这种方式没有考虑到属性敏感性差异. 属性携带的信息量越多, 帮助攻击者推出目标信息的贡献越大, 也就意味着它对外发布的敏感性越高. 基于此, 本文提出差分隐私异构条件分布计算算法 HnoisyConditionals, 具体实现过程如算法 5 所示.

算法 5. HnoisyConditionals 算法

输入: 数据集$D_i=\{x_1, x_2, \cdots, x_n\}$, 贝叶斯网络$N_i$, 隐私预算$\varepsilon_3$
输出: 根据N_i生成的带噪声的条件分布集P_i^*

1. 初始化$P_i^*=\varnothing$, 令$l'=\min(|C_i|-1,d)$;
2. for $j=l'+1$ to $|C_i|$ do
构建X_j的联合分布$Pr(X_j,\prod_j)$;
加入拉普拉斯噪声$Lap\left(\frac{2}{n\varepsilon_3}\times\frac{\sum_{j=l'+1}^{|C_i|}\exp(-OE_j)}{\exp(-OE_j)}\right)$, 实现差分隐私$Pr^*(X_j,\prod_j)$;
设$Pr^*(X_j,\prod_j)$中的负值归 0 并正常化;
从$Pr^*(X_j,\prod_j)$中提取$\Pr^*(X_j|\prod_j)$并加入P_i^*;
end for
3. for $j=1$ to l' do
从$Pr^*(X_{l'+1},\prod_{l'+1})$中提取$Pr^*(X_j|\prod_j)$加入$P_i^*$;
end for
4. 返回P_i^*.

根据式 (7), 归一化风险熵计算公式如下:

$$OE_j=-\frac{1}{\log n}\sum_{i=1}^{n}p_i\log p_i \tag{11}$$

在算法 5 中, 向联合分布注入噪声不再采取均分策略, 而是以归一化风险熵为权重分配隐私预算. 由文献 [13] 可知, 联合分布全局敏感度为$2/n$. 因此, 每次对联合分布$Pr(X_j,\prod_j)$添加的拉普拉斯噪声大小为:

$$Lap\left(\frac{2}{n\varepsilon_3}\times\frac{\sum_{j=l'+1}^{|C_i|}\exp\left(-OE_j\right)}{\exp\left(-OE_j\right)}\right) \tag{12}$$

其中, 属性归一化风险熵OE_j越大, 属性敏感性越高, 需要的隐私保护强度越大, 分配的隐私预算也就越小. 由定义 3 和性质 1 可知, 算法 5 满足ε_3-差分隐私.

3.5 隐私分析

定理 1. HMPrivBayes 算法满足ε-差分隐私.

证明: 在 HMPrivBayes 的整个执行过程中, 其中有 3 步涉及原始数据集的访问, 分别是属性分割、构建差分隐私贝叶斯网络和提取属性噪声条件分布, 下面分别讨论这 3 步的隐私保护强度.

对于属性分割, 需要重复访问数据集D计算属性对的 MIC, 每次访问数据集D可以计算$\lfloor d/2\rfloor$对属性对的 MIC, 相似矩阵S中有$d(d-1)/2$个不同的 MIC 值需要计算, 因此, 需访问$d(d-1)/(2\lfloor d/2\rfloor)$次数据集$D$. MIC 计算的敏感度为$\Delta m$, 故每次扰动添加的噪声量为$Lap((\Delta m\cdot d(d-1))/(\varepsilon_1\cdot(2\lfloor d/2\rfloor)))$, 进而属性分割满足$\varepsilon_1$-差分隐私.

在贝叶斯网络构建阶段, 需要借助指数机制$\exp\left(\varepsilon_2 H(X_j)\Big/(2|C_i|\Delta H)\right)$从数据子集$D_i$中选取加入$N_i$的首个属性节点. 指数机制$\exp(\varepsilon_2 u(X_t,\prod_t)/(2|C_i|\Delta u))$用于选择互信息最大的$(X_i,\prod_i)$对加入$N_i$, 执行$|C_i|-1$次, 因此贝叶斯网络构建满足$\varepsilon_2$-差分隐私.

从贝叶斯网络N_i中提取属性条件分布时, 需要往$|C_i|-l'$个属性联合分布中添加噪声扰动, 进一步保护数据隐私. 为了兼顾属性敏感性的差异, 隐私预算根据属性X_j的归一化风险熵OE_j异构分配, 添加的噪声量为$Lap\left(\left(2\times\sum_{j=l'+1}^{|C_i|}\exp(-OE_j)\right)\Big/\left(n\varepsilon_3\exp(-OE_j)\right)\right)$, 其中联合分布计算的敏感度为$2/n$. 因此, 属性条件分布计算满足$\varepsilon_3$-差分隐私.

综上, HMPrivBayes 算法满足ε-差分隐私, 其中$\varepsilon_1+\varepsilon_2+\varepsilon_3=\varepsilon$.

4 实验评估

4.1 实验设置

本次实验使用 Python 编程语言来实现所有的方法, 其中贝叶斯网络模型部分的实现参考了 Zhang 等人[13] 论文的实验相关代码. 实验的硬件环境为 AMD Ryzen7 4800H (2.90 GHz), 操作系统为 Windows 10 (64 位), 内存为 16 GB.

实验采用两个公开可用的数据集 Adult[27] 和 Big5[28]. 其中, Adult 为美国人口普查数据集, 包含 45 222 条个体信息记录, 每条记录包含 15 个属性. Big5是包含 19719 条人口普查记录的信息集合, 每条记录包含 57 个属性. 利用等宽法完成属性转化后, 二进制属性的数量分别为 52 和 175. 两个数据集的详细信息如表 1 所示.

表 1 数据集属性信息描述

数据集	记录数	属性数	转化后属性数
Adult	45 222	15	52
Big5	19 719	57	175

对于这两个数据集, 隐私预算分配策略为$\varepsilon_1=a_1/a\varepsilon$, $\varepsilon_2=a_2/a\varepsilon$, $\varepsilon_3=a_3/a\varepsilon$, 其中$a=a_1+a_2+a_3$, $a_1=\frac{d-1}{n}, a_2=\frac{2d}{nk}, a_3=\frac{2d}{nk}$. 属性分类簇个数$k$默认值为 3, 贝叶斯网络的最大入度数$l$默认值为 3.

4.2 评价指标

本文通过α-边际分布[29] 和 SVM (support vector machine) 分类[30] 来评估算法的性能. 实验选取 2-边际

分布和 3-边际分布作为α-边际分布评估的实例, 通过计算生成的噪声边际分布和原始边际分布的平均变量距离 (average variation distance, AVD)[31] 确性. L_1距离的一半:

$$Dist_{\mathrm{AVD}}(P,Q)=\frac{1}{2}\sum_{\omega\in\mho}|P(\omega)-Q(\omega)| \quad (13)$$

其中, $P(\omega)$和$Q(\omega)$分别表示加噪前后的边际分布, $\mho$表示边际分布集合.

实验通过度量 SVM 分类的准确性, 分析生成的噪声数据集的有效性. 本组实验在噪声数据集上同时训练 2 个分类器, 通过分析合成数据集的其他属性预测目标属性的分类结果. 对于 Adult 数据集, 选取 salary 作为分类实例, 预测个体每年收入是否超过 50k. 对于 Big5 数据集, 选取 age 作为分类实例, 预测个体年龄是否在 50 岁以下. 每个分类任务, 将合成数据集按照 8:2 分为训练数据集和测试数据集, 重复运行 50 次, 并记录实验结果的平均值.

4.3 实验结果

本文将通过 3 个不同的实验来分析 HMPrivBayes 算法的可用性和高效性. 为了更好地评估 HMPrivBayes 方案的性能, 本节采用的对比算法包括: PrivSCBN、PrivBayes、Jtree、NoPrivacy. 其中, PrivSCBN 算法[20]、PrivBayes 算法[13] 均是基于贝叶斯网络的数据发布算法, Jtree[14] 是基于联合树的数据发布算法, NoPrivacy 是 HMPrivBayes 不添加噪声扰动的数据发布情况.

第 1 个实验是为了验证 HMPrivBayes 算法的安全性, 比较在不同记录数量下, 敏感属性隐私泄露的可能性. 在 Adult 数据集原有的 45 222 条记录的基础上, 随机产生 54 778 条记录, 生成包含 100 000 条记录的数据集, 并将 salary 作为敏感属性. 由于隐私预算ε不是该实验关心的变量, 故将ε固定为 0.2. 实验结果如图 4 所示, 从图中可以看出, 随着记录数的增加隐私泄露概率逐渐下降, 即攻击者根据其他属性信息推断出 salary 属性值的概率逐渐下降, 而且 HMPrivBayes 算法安全性高于 PrivBayes 算法和 PrivSCBN 算法. 这是因为, HMPrivBayes 算法根据属性敏感性决定隐私预算分配大小, 能对多属性数据发布提供更安全的隐私保护.

第 2 个实验是对 HMPrivBayes 算法数据性能的分析, 图 5(a)–图 5(d) 分别对比了在不同ε取值下, 算法 HMPrivBayes、PrivSCBN、PrivBayes、Jtree 的合成数据集的边际分布与原始数据集边际分布之间 AVD 的变化. 图 5(e)–图 5(f) 显示了算法 HMPrivBayes、PrivSCBN、PrivBayes、Jtree 的合成数据集训练 2 个分类器的误分类率在不同ε取值下的变化趋势.

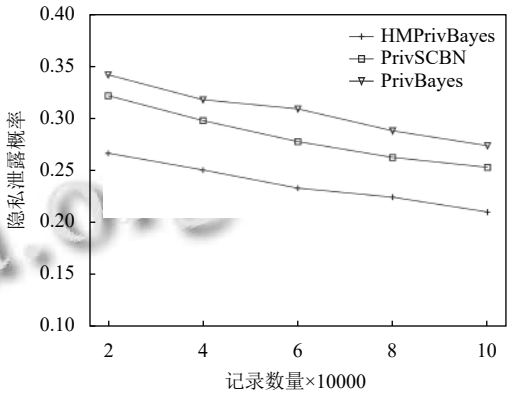
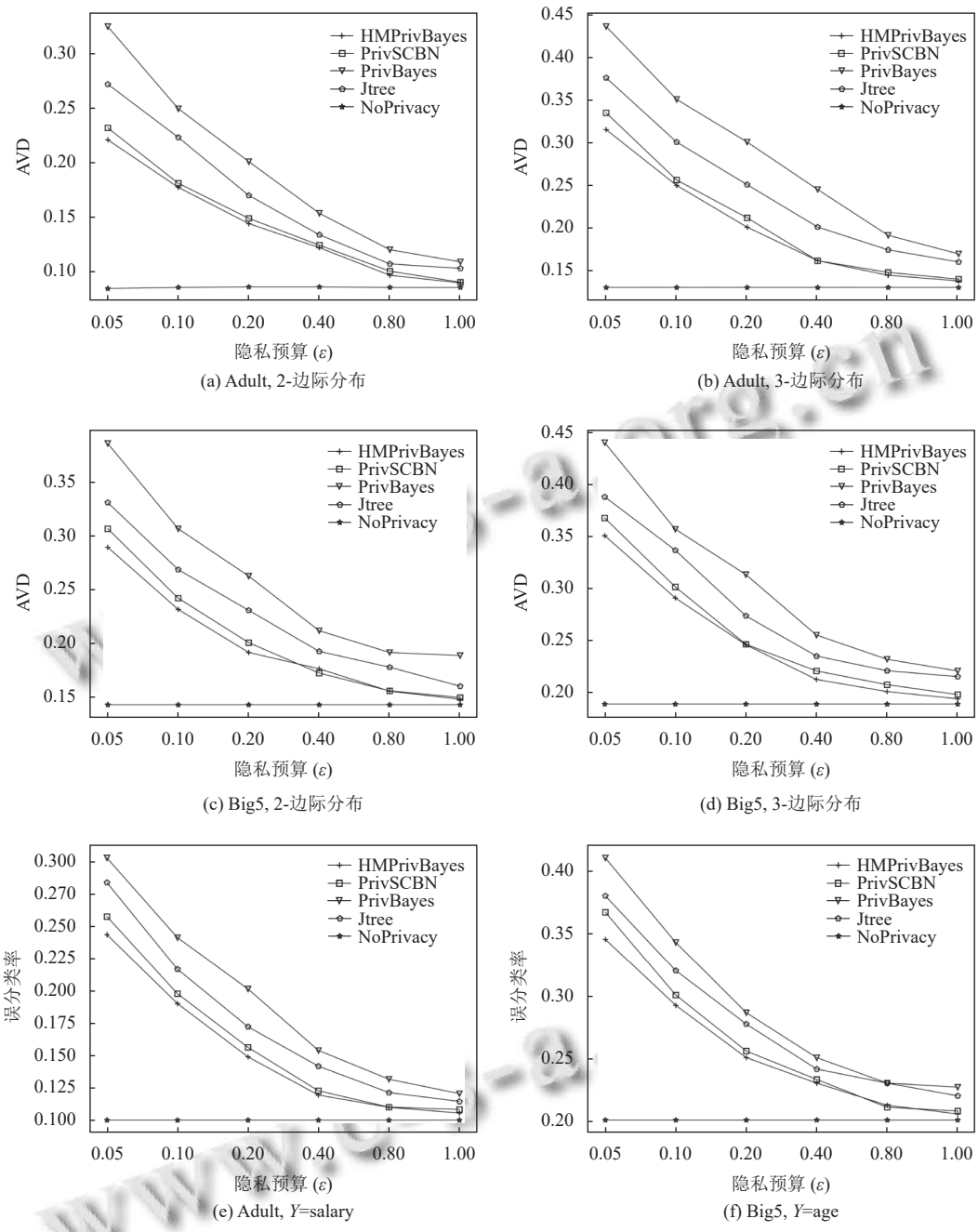


图 4 数据安全性分析

可以看出, 对于 NoPrivacy, 边际分布之间的 AVD 和分类器的误分类率均与隐私预算ε的取值无关, 且 NoPrivacy 的性能均优于其他方法的性能. 随着隐私预算ε取值的增大, 隐私保护强度下降, 合成数据的两项指标越来越接近 NoPrivacy 的情况, 这符合差分隐私的规律. 在多数情况下, HMPrivBayes 的性能均优于 PrivSCBN、PrivBayes、Jtree 的性能, 仅当隐私预算ε足够大时, HMPrivBayes 与 PrivSCBN 性能趋于接近. 这说明本文提出的各种贝叶斯网络改进策略以及异构属性发布策略是有效的.

第 3 个实验是对 HMPrivBayes 算法计算效率的分析, 比较在贝叶斯网络不同最大入度数l下, 算法的整体运行时间. 由于隐私预算ε不是该实验关心的变量, 故将ε固定为 0.2. 实验结果如图 6 所示, 随着最大入度数l的增加, 整体运行时间大幅上升. 比较图 6(a)–图 6(b) 可以看出, 随着属性数目的增加, 算法整体运行时间也大幅增加. 不过, HMPrivBayes 的运行时间远低于 PrivBayes 算法, 而且属性越多, 两者运行时间的差距越大, 这说明 HMPrivBayes 的效率得到了较好的提升. 效率的提升一方面在于, HMPrivBayes 利用聚类算法分割属性集, 缩减了单个贝叶斯网络的节点空间; 另一方面在于, 借助属性关联强度, 缩减了候选属性对集合.



(a) Adult, 2-边际分布 (b) Adult, 3-边际分布

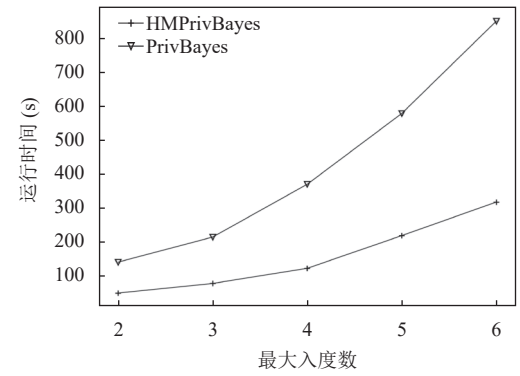
(c) Big5, 2-边际分布 (d) Big5, 3-边际分布

(e) Adult, *Y*=salary (f) Big5, *Y*=age

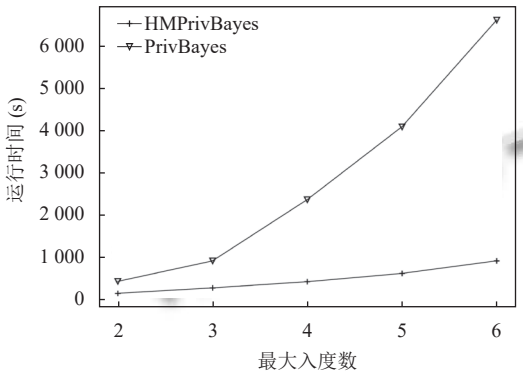
图 5 算法 HMPrivBayes、PrivSCBN、PrivBayes、Jtree 在不同隐私预算下的性能分析

5 结束语

针对差分隐私异构多属性数据发布问题, 本文提出了一种基于属性分割的异构多属性数据差分隐私发布方法 HMPrivBayes. 与已有多属性数据发布算法不同的是, HMPrivBayes 基于属性敏感性差异给予属性数据相对应的隐私保护强度, 避免了多属性数据隐私保护不均匀的问题, 进而提高数据可用性. 与此同时, 该方法通过引入属性分割、改进贝叶斯构建过程等都实现了算法整体计算效率的提升. 在不破坏属性关联的前提下, 以属性归一化风险熵为权重分配隐私预算, 实现异构多属性数据发布. 理论证明, HMPrivBayes 满足 ε-差分隐私. 实验结果表明, HMPrivBayes 方法在提升算法整体计算效率的基础上, 保证了数据发布的可用性在未来的研究中, 我们将考虑基于差分隐私的分布式异构多属性数据发布, 即如何在多方设置中实现异构多属性数据发布.



(a) Adult



(b) Big5

图 6 算法运行时间分析

参考文献

1 Sweeney L. K-Anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557–570. [doi: 10.1142/S0218488502001648]

2 Vel Kumar P M, Karthikeyan M. L diversity on K-anonymity with external database for improving privacy preserving data publishing. International Journal of Computer Applications, 2012, 54(14): 7–13. [doi: 10.5120/8632-2341]

3 Li NH, Li TC, Venkatasubramanian S. *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. 2007 IEEE 23rd International Conference on Data Engineering. Istanbul: IEEE, 2007. 106–115.

4 Dwork C. Differential privacy. 33rd International Colloquium on Automata, Languages and Programming. Venice: Springer, 2006. 1–12.

5 Dwork C, McSherry F, Nissim K, *et al*. Calibrating noise to sensitivity in private data analysis. 3rd Theory of Cryptography Conference on Theory of Cryptography Conference. New York: Springer, 2006. 265–284.

6 McSherry F, Talwar K. Mechanism design via differential privacy. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Providence: IEEE, 2007. 94–103.

7 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用. 计算机学报, 2014, 37(1): 101–122.

8 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护. 计算机学报, 2014, 37(4): 927–949.

9 Zhu TQ, Li G, Zhou WL, *et al*. Differentially private data publishing and analysis: A survey. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(8): 1619–1638. [doi: 10.1109/TKDE.2017.2697856]

10 Xiao YH, Xiong L, Yuan C. Differentially private data release through multidimensional partitioning. 7th VLDB Workshop on Secure Data Management. Singapore: Springer, 2010. 150–168.

11 Cormode G, Procopiuc C, Srivastava D, *et al*. Differentially private spatial decompositions. 2012 IEEE 28th International Conference on Data Engineering. Arlington: IEEE, 2012. 20–31.

12 Mohammed N, Chen R, Fung BCM, *et al*. Differentially private data release for data mining. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: ACM, 2011. 493–501.

13 Zhang J, Cormode G, Procopiuc CM, *et al*. PrivBayes: Private data release via Bayesian networks. ACM Transactions on Database Systems, 2017, 42(4): 25.

14 Chen R, Xiao Q, Zhang Y, *et al*. Differentially private high-dimensional data publication via sampling-based inference. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM, 2015. 129–138.

15 Chaudhuri K, Sarwate AD, Sinha K. A near-optimal algorithm for differentially-private principal components. The Journal of Machine Learning Research, 2013, 14(1): 2905–2943.

16 Wang S, Chang JM. Differentially private principal component analysis over horizontally partitioned data. 2018 IEEE Conference on Dependable and Secure Computing (DSC). Kaohsiung: IEEE, 2018. 1–8.

17 Xu CG, Ren J, Zhang YX, *et al*. DPPro: Differentially private high-dimensional data release via random projection. IEEE Transactions on Information Forensics and Security, 2017, 12(12): 3081–3093. [doi: 10.1109/TIFS.2017.2737966]

18 Qardaji WH, Yang WN, Li NH. PriView: Practical differentially private release of marginal contingency tables.

Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. Snowbird: ACM, 2014. 1435–1446.

19 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(10): 2342–2352.

20 洪金鑫, 吴英杰, 蔡剑平, 等. 基于属性分割的高维二值数据差分隐私发布. 计算机研究与发展, 2022, 59(1): 182–196. [doi: 10.7544/issn1000-1239.20200701]

21 陈恒恒, 倪志伟, 朱旭辉, 等. 基于聚类分析的差分隐私高维数据发布方法. 计算机应用, 2021, 41(9): 2578–2585. [doi: 10.11772/j.issn.1001-9081.2020111786]

22 Dwork C, Lei J. Differential privacy and robust statistics. Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda: Association for Computing Machinery, 2009. 371–380.

23 Dwork C, Naor M, Reingold O, *et al*. On the complexity of differentially private data release: Efficient algorithms and hardness results. Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda: Association for Computing Machinery, 2009. 381–390.

24 McSherry FD. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. Providence: Association for Computing Machinery, 2009. 19–30.

25 Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01). Vancouver: NIPS, 2001. 849–856.

26 Reshef DN, Reshef YA, Finucane HK, *et al*. Detecting novel associations in large data sets. Science, 2011, 334(6062): 1518–1524. [doi: 10.1126/science.1205438]

27 Adult Data Set. https://archive.ics.uci.edu/ml/datasets/adult. [2021-12-15].

28 Open-source Psychometrics Project. http://personality-testing.info/_rawdata/. [2021-12-15].

29 Barak B, Chaudhuri K, Dwork C, *et al*. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Beijing: ACM, 2007. 273–282.

30 周志华. 机器学习. 北京: 清华大学出版社, 2016.

31 Tsybakov AB. Introduction to Nonparametric Estimation. New York: Springer, 2009.

(校对责编: 牛欣悦)