面向关联属性的差分隐私信息熵度量方法

吴宁博[1,2],彭长根[1,3],牟其林[2]
(1. 贵州大学计算机科学与技术学院,贵州贵阳 550025; 2. 中电科大数据研究院有限公司,贵州贵阳 550022; 3. 公共大数据国家重点实验室,贵州贵阳 550025)

摘 要: 针对差分隐私非交互式多属性关联的合成数据集发布问题,基于信息熵、汉明失真提出了发布数据集隐私度、数据效用、隐私泄露风险的量化方法. 首先,利用互信息量分析属性相关度,并以关联依赖图模型表达属性关联. 其次,基于图中关键隐私泄露路径构建马尔可夫隐私泄露链,并结合信息熵提出一种关联属性隐私度量模型及方法,可以有效的度量由关联属性引起的隐私泄露量. 最后,通过具体实例验证了模型与方法的有效性,并对比分析了该方法的优势.

关键词: 差分隐私; 信息熵; 隐私度量; 关联属性
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112 (2019)11-2337-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.11.015

Information Entropy Metric Methods of Association Attributes for Differential Privacy

WU Ning-bo[1,2], PENG Chang-gen[1,3], MOU Qi-lin[2]
(1. *College of Computer Science & Technology, Guizhou University, Guiyang, Guizhou* 550025, *China*;
2. *CETC Big Data Research Institute Co., Ltd., Guiyang, Guizhou* 550022, *China*;
3. *National Key Laboratory of Public Big Data, Guiyang, Guizhou* 550025, *China*)

Abstract: Privacy leakage and utility measurement are widely concerned issues in multi-attribute datasets by non-interactive differential privacy publishing. In this paper, we have proposed several quantification methods by using information entropy and hamming distortion to quantify the privacy of published dataset, utility of dataset and risk of privacy leakage. First, we have tailored the existing mutual information concept to analyze the relationship among associated attributes and constructed an associated dependency graph model to analyze their correlations among multi-attribute. After that, we have developed a privacy quantification method based on information entropy and privacy leakage Markov chain, which is generated based on the graph of privacy leakage path that has a valid efficiency measurement of the privacy leakage leading by associated attributes. Finally, to justify the efficiency of the proposed model, we have included an illustrative example and demonstrated the advantage of our method by comparing with other methods.

Key words: differential privacy; information entropy; privacy metric; association attributes

1 引言

Dwork 等人[1,2]提出的差分隐私(DP)保护框架具有强背景知识假设、忽视隐私攻击者的计算能力和严格数学理论支撑等特性,迅速成为隐私保护领域研究的热点. 目前,差分隐私主要有交互式和非交互式两种工作模式[1],在差分隐私非交互式工作模式中,数据发布者通过差分隐私噪声扰动发布原始数据集的近似副本,供数据使用者进行查询分析使用. 其中,衡量数据发布中的隐私泄露程度和数据效用势必涉及到隐私度量. 隐私度量问题是研究平衡隐私泄露与发布数据效用的关键,同时也是隐私泄露风险分析及量化评估的理论基础. 由此,研究

收稿日期:2018-10-08;修回日期:2019-03-22;责任编辑:李勇锋
基金项目:国家自然科学基金(No. U1836205, No. 61662009, No. 61772008, No. 11761020);贵州省科技重大专项计划(No. 20183001);贵州省自然科学基金(黔科合基础[2017]1045);贵州省研究生科研基金立项课题(No. KYJJ2017005);"十三五"国家密码发展基金(No. MMJJ20170129);贵州省科技计划课题(黔科合重大专项字[2017]3002,黔科合平台人才[2017]5788,黔科合重大专项字[2018]3007)

隐私度量问题具有重要的理论和实际应用价值.

发布数据的隐私泄露与数据效用是一对极大极小的对偶问题,长期以来学术界都致力于研究平衡隐私泄露与数据效用的有效方法[3,4]. 其中,隐私保护的隐私泄露与数据效用度量是基础. 量化信息流作为一种基于信息论广泛被接受的量化方法,受到了研究者的关注[5]. 从信息熵的角度研究差分隐私机制隐私泄露风险与数据效用度量问题,对于隐私泄露风险量化分析和最优化机制设计具有重要的理论及实践意义.

近年来,基于信息熵的隐私量化已经受到了研究者的广泛关注. Sankar 等[3]构建数据库概率模型,基于信息熵理论提出了数据集类别属性与相关联属性的隐私度量方法. Barthe 等[6]利用最小熵原理量化了差分隐私保护的二进制数据库最大隐私信息泄露量. Calmon 和 Fawaz[7]提出 ε-信息隐私满足 2ε-差分隐私,进一步指出数据集与扰动响应输出结果的互信息隐私泄露上界. Darakhshan[8]通过定义失真函数,研究建立了差分隐私与失真度量的联系,提出以互信息度量差分隐私的隐私泄露量. 随后,Alvim 等[9]使用最小熵原理定义最小熵泄露风险函数,量化了差分隐私机制隐私泄露风险与数据效用. 2016 年,彭长根等[10]研究了隐私信息传播基本通信模型,并运用信息熵、联合信息熵、条件熵、平均互信息量等对隐私信息进行度量. 同年,Wang 等[11]提出一种可识别隐私量化方法,并基于可识别、差分隐私和互信息隐私三个不同隐私概念,研究了三种隐私量化的基本联系,同时证明了互信息最佳机制满足 ε-差分隐私. 2018 年,Kalantari 等[4]将隐私信源概率分布划分为均匀分布、有序递减分布和其它分布,并针对三种不同的信源分布,结合信息熵理论研究了汉明失真下隐私泄露与数据效用的度量方法,最后针对不同信源分布研究了平衡隐私泄露与数据效用的最优化差分隐私信道问题. 此外,文献[12,13]也都从不同角度对基于信息熵的隐私泄露风险、数据效用度量开展了相关研究工作.

鉴于上述分析,构建隐私信息传播通信模型使用信息熵度量隐私信息泄露具有理论可行性. 然而,目前相关研究工作中多以记录元组为离散随机变量构建离散信源模型且多以二进制交织信道为研究对象,较少考虑元组属性关联关系且缺少多类型属性组合方面的研究.

2 基础知识

2.1 失真度量

由于信道存在的噪声和干扰,香农离散无记忆信道[14]是信源字母表 $\mathcal{X}$ 到再生字母表 $\hat{\mathcal{X}}$ 的概率映射. 对于固定信源概率分布,信道传输信息率依赖于信道失真特性. 为定量描述信道传输失真,定义失真测量函数.

定义 1 单符号失真度量是一个非负函数 $d:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}^+$ 映射给定字母表 $\mathcal{X}$ 和 $\hat{\mathcal{X}}$ 的乘积空间到一个非负实数.

多数情况下,通常假设 $\mathcal{X}$ 和 $\hat{\mathcal{X}}$ 是离散有限集合,且 $\mathcal{X}=\hat{\mathcal{X}}$. 对于汉明失真测量具有以下定义形式

$$d(x,\hat{x})=\begin{cases}0, & x=\hat{x}\\ 1, & x\neq\hat{x}\end{cases} \tag{1}$$

对所有的 $x\in\mathcal{X}$ 和 $\hat{x}\in\hat{\mathcal{X}}$. 汉明失真 $d(x,\hat{x})$ 测量信源输入符号与信宿再生符号之间的误差距离.

2.2 差分隐私

从信息论通信的视角,差分隐私噪声扰动机制可抽象为原始数据输入到噪声扰动输出的离散无记忆噪声信道. 因此,差分隐私的定义可以表述为信道条件概率分布的特性.

定义 2[4,13] 数据集元素取值于离散的有限字母表 $\mathcal{X}$,如果一个随机化机制 $P_{\hat{X}|X}(\hat{X}|X)$ 称之为 ε-差分隐私,当且仅当对于任意的相邻输入元素 x 和 x',在输出字母表空间 $\hat{\mathcal{X}}$ 中满足

$$P(\hat{x}|x)\leqslant e^{\varepsilon}P(\hat{x}|x') \tag{2}$$

其中 $x,x'\in\mathcal{X}$ 和 $\hat{x}\in\hat{\mathcal{X}}$.

3 关联属性隐私度量问题

数据集是由元组和列属性构成的二维关系表,其中的元组表示特定个体实体,列是实体属性. 例如,健康医疗数据集通常包含姓名、性别、出生日期、血压、体重、疾病等属性描述信息. 然而,这些属性集合中通常包含有个体敏感的隐私属性. 考虑可信数据管理者存储有关个体隐私信息的原始数据集 $\boldsymbol{D}$,每条记录是有关特定个体 k 个维度的属性描述信息. 抽象数据集单属性为离散随机变量 $X_i(1\leqslant i\leqslant k)$,$k$ 维度属性构成序列长度为 k 的一个随机向量 $\boldsymbol{X}$. 其中 $\boldsymbol{X}$ 表示多属性随机变量集合,记作 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_i,\cdots,X_k\}$. 属性 X_i 的值域基数,记作 $m_i=|\mathcal{X}_i|$,属性域全集 $\mathcal{X}=\prod_{i=1}^{k}\mathcal{X}_i$,基数 $m=\prod_{i=1}^{k}m_i$. 数据集 $\boldsymbol{D}=(d_1,\cdots,d_n)$ 表示 n 个体数据信息,是取值于数据全集 $\mathcal{X}^n$ 的多重集. 由此,数据集 $\boldsymbol{D}$ 可视为随机向量 $\boldsymbol{X}$ 的 n 次独立随机观测样本,记作 $\boldsymbol{X}^n$.

差分隐私非交互式数据发布场景中,数据管理者旨在发布原始数据集 $\boldsymbol{X}^n$ 的近似副本 $\hat{\boldsymbol{X}}^n$,自然的考虑差分隐私输出近似合成数据集 $\hat{\boldsymbol{X}}^n$ 和原始数据集 $\boldsymbol{X}^n$ 具有相同的取值域 $\mathcal{X}^n$. 假设数据集元组记录独立抽样于一个分布 $P(\boldsymbol{X})$,元组多属性之间存在关联. 从信息论的视角可以

将差分隐私非交互式数据发布抽象为信息论经典点到点的通信模型,差分隐私随机化机制接收原始数据集$\boldsymbol{X}^n$输入,并输出近似数据集$\hat{\boldsymbol{X}}^n$,形成噪声信道$P(\hat{\boldsymbol{x}}|\boldsymbol{x})$,信道条件转移概率的比值决定差分隐私保护强度.

差分隐私已有的隐私度量研究工作中较少考虑属性之间存在的关联性依赖. 然而,实际应用场景中,有关个体信息的多维属性之间极少存在完全独立的情况,多维属性之间的关联可能造成个体隐私泄露问题. 例如医疗数据集中,血压与体重的关联、出生日期与社会保障号的关联等,极大可能会泄露个人的健康状况(敏感信息). 关联属性识别是寻找隐私泄露途径的关键,但是,真实数据集中通常混合数值型属性和类别型属性,为关联分析带来难度. 此外,基于隐私泄露关键途径,细粒度考虑多属性间关联的隐私度量需要新的解决方法.

4　关联属性隐私度量模型及方法

4.1　关联属性识别

数据集$\boldsymbol{D}$中个体记录d由k维属性集合$\boldsymbol{X}=\{X_1,X_2,\cdots,X_i,\cdots,X_k\}(1\leqslant i\leqslant k)$表示,如果存在$l(l<k)$个属性与属性$X_i$关联,则称之为一个关联属性组,记作$\boldsymbol{R}_i=\{X_i,X_j\in\boldsymbol{X}|$所有与$X_i$相关联的属性$X_j\}$.

针对属性集合$\boldsymbol{X}$依次分析属性对$(X_i;X_j)$且$X_i\neq X_j$之间的相关性. m_i,m_j分别为属性基数,z_{ij}表示频率矩阵满足X_i属性为i类且X_j属性为j类的频率统计值. 则由此可以得到一个$m_i\times m_j$的二维属性联合频率表,如表1所示.

表1　二维属性联合频率表

	$X_j(1)$	$X_j(2)$	…	$X_j(m_j)$
$X_i(1)$	z_{11}	z_{12}	…	z_{1m_j}
$X_i(2)$	z_{21}	z_{22}	…	z_{2m_j}
…	…	…	…	…
$X_i(m_i)$	z_{m_i1}	z_{m_i2}	…	$z_{m_im_j}$

二维联合概率分布p_{ij},记作p_{ij}为属性$X_i=i$(X_i属性取值域中第i个值和$X_j=j$(X_j属性取值域中第j个值的元组频率,$p_{i\cdot}$和$p_{\cdot j}$分别表示边缘概率分布. 其中$p_{i\cdot}=\sum_j p_{ij}$和$p_{\cdot j}=\sum_i p_{ij}$. 由于互信息量作为相关性分析的度量,相比与其它统计相关性识别方法具有敏感度小的优势,能够克服线性与非线性关系的局限性[15,16]. 因此本文的研究中采用互信息量方法识别属性集合元素之间的关联性,据此定义属性X_i,X_j之间的相关度为互信息量

$$I(X_i;X_j)=\sum_{i=1}^{m_i}\sum_{j=1}^{m_j}P_{ij}\log_2\frac{P_{ij}}{P_i\cdot P_j}\tag{3}$$

4.2　关联属性图模型

数据集$\boldsymbol{D}$中属性对的相关性反映元组属性的近似依赖强度,通过计算属性之间的互信息量可以得到属性之间的相关度.

定理1　对于彼此相关联的属性对$(X_i;X_j)$,互信息量$I(X_i;X_j)$表示为相关度θ_{ij},$I(X_j;X_i)$表示为相关度θ_{ji},则有$\theta_{ij},\theta_{ji}\geqslant 0$且$\theta_{ij}=\theta_{ji}$.

定理1证明为互信息性质,在此不再赘述.

推论1　如果属性X_i,X_j相关度$\theta_{ij}=\theta_{ji}=0$,则属性对$(X_i;X_j)$相互独立,即是属性间条件概率$P(X_i|X_j)=P(X_j|X_i)=0$.

针对数据集$\boldsymbol{D}$中的元组属性集合$\boldsymbol{X}$,采用互信息的相关性分析,生成数据集属性关联依赖无向图,图中边具有相关度权值形成带权无向图G,属性构成顶点集合V,属性之间关联依赖构成边集合. 属性的关联依赖强度θ_{ij}表示带权无向图中顶点X_i到X_j的边$(X_i;X_j)$的权值.

数据集属性关联依赖图模型采用邻接矩阵的形式描述,以下给出带权无向图的邻接矩阵定义.

定义3　数据集属性关联带权无向图邻接矩阵定义为

$$\boldsymbol{\Theta}[i][j]=\begin{cases}0, & \theta_{ij}<\delta\\ \theta_{ij}, & \theta_{ij}\geqslant\delta\end{cases},\quad 1\leqslant i,j\leqslant k\tag{4}$$

其中,θ_{ij}为属性对相关度$I(X_i;X_j)$,δ为过滤伪相关设置的阈值,矩阵相应值设置0表示无相关的独立属性. 基于数据集相关属性分析,可以将属性集合中所有相关属性表示为相关度邻接矩阵$\boldsymbol{\Theta}$.

由于互信息量具有对称性、非负性、凸函数性等特点[14]. 易知相关度邻接矩阵$\boldsymbol{\Theta}$具有如下的性质:(1)相关度邻接矩阵$\boldsymbol{\Theta}$是对称矩阵,$\theta_{ij}=\theta_{ji}$;(2)相关度邻接矩阵$\boldsymbol{\Theta}$对角线元素均为0.

数据集关联属性识别生成带权无向图邻接矩阵$\boldsymbol{\Theta}$算法,初始化设置带权无向图G的邻接矩阵元素为0,默认初始状态数据集属性完全独立. 依次计算属性对之间相关互信息量,并根据阈值过滤属性边集合,设置边权值θ_{ij}. 最后,输出属性关联带权无向图邻接矩阵$\boldsymbol{\Theta}$.

算法　生成属性相关度邻接矩阵$\boldsymbol{\Theta}$

输入:数据集$\boldsymbol{D}$,属性集合$\boldsymbol{X}=\{X_1,X_2,\cdots,X_i,\cdots,X_k\}$,$k$,$\delta$
输出:邻接矩阵$\boldsymbol{\Theta}$

```
1.  初始化 Θ
2.    for i = 1,…,k do
3.      for j = 1,…,i do
4.        Set θ_ij = θ_ji = 0
5.      end for
6.    end for
7.    for i = 1,…,k do
8.      Set vector V = X\{X_i}
9.      for each attribute in vector X_j ← V
10.   Compute θ = I(X_i;X_j) with dataset D
```

11. If ($\theta \geq \delta$)
12. Set $\theta_{ij} = \theta_{ji} = \theta$
13. end if
14. end for
15. end for
16. Ouput $\boldsymbol{\Theta}$

基于互信息的关联属性相关度计算复杂度是数据集属性问题规模 k 的函数,上述算法计算输出属性关联带权无向图邻接矩阵 $\boldsymbol{\Theta}$ 的计算时间复杂度为 $O(k^2)$,是一个多项式时间算法. 此外,由于相关度邻接矩阵 $\boldsymbol{\Theta}$ 满足主对角线元素为 0 的 k 阶对称矩阵,由此可基于矩阵的压缩存储将矩阵 $\boldsymbol{\Theta}$ 存储到 $k(k-1)/2$ 个存储单位空间.

4.3 关联属性隐私度量方法

本节针对差分隐私非交互式数据发布问题,基于属性关联依赖图,从隐私信源熵、隐私度与隐私泄露风险、数据效用三个方面展开隐私度量,提出关联属性的隐私信息熵度量方法.

4.3.1 隐私信源熵

多属性隐私信源熵表示为属性的联合信源熵,数据集每一行记录的 k 个维度属性构成序列长度为 k 的一个随机向量 $\boldsymbol{X}$. 概率空间为数据集域值空间的 m 种所有可能组合,$\boldsymbol{x}_i$ 表示第 i 个可能的属性组合,则随机向量 $\boldsymbol{X}$ 的概率模型可以表示为属性联合概率分布. 随机向量 $\boldsymbol{X}$ 的信源熵表示为多属性联合信源熵 $H(\boldsymbol{X}) = H(X_1X_2\cdots X_k)$. 特别的,数据集属性相互独立是多属性关联的一种特殊情况,此时属性联合概率 $P(\boldsymbol{X}) = \prod_{1\leqslant i\leqslant k} X_i$ 成立,元组属性 $\boldsymbol{X} = \{X_1, X_2, \cdots, X_i, \cdots, X_k\}$ 的联合信源熵表述为

$$H(\boldsymbol{X}) = \sum_{1\leqslant i\leqslant k} H(X_i) \tag{5}$$

联合信源熵刻画个体隐私信息不确定度等于各离散属性不确定度的叠加,能够从属性整体上对个体隐私信息进行度量. 假设数据集记录独立同分布(i. i. d),独立取值于属性的所有可能组合,则可以将其抽象为离散平稳无记忆信源的 n 次扩展构成的一个新隐私信源,记作信源 $\boldsymbol{X}^n$. 由此,其联合信源熵可表述为

$$H(\boldsymbol{X}^n) = nH(\boldsymbol{X}) \tag{6}$$

4.3.2 隐私熵与隐私泄露风险

差分隐私非交互式数据发布中,原始数据集的隐私熵,即是隐私攻击者观察噪声信道输出合成数据集后对原始数据集仍具有的不确定度. 采用条件熵度量所有的隐私变量的不确定度,定义平均隐私熵 E 为

$$E = \frac{1}{n} H(\boldsymbol{X}^n \mid \hat{\boldsymbol{X}}^n) \tag{7}$$

假设数据集抽样 n 元组服从先验概率分布 $P(\boldsymbol{X})$,互信息量度量差分隐私合成数据集 $\hat{\boldsymbol{X}}^n$ 包含有关信源 $\boldsymbol{X}^n$ 的信息量,自然可以利用 $I(\boldsymbol{X}^n; \hat{\boldsymbol{X}}^n)$ 从数据管理者角度定义隐私泄露风险函数 $L(\cdot)$ 量化平均互信息泄露量

$$L(\cdot) = \frac{1}{n} I(\boldsymbol{X}^n; \hat{\boldsymbol{X}}^n) = \frac{1}{n}[H(\boldsymbol{X}^n) - H(\boldsymbol{X}^n \mid \hat{\boldsymbol{X}}^n)] \tag{8}$$

由式(8)可知,固定数据集信源分布,隐私泄露风险与隐私熵负相关. 互信息量是关于信源分布和差分隐私噪声信道转移条件概率的二元函数. 一种自然的考虑,基于互信息量的隐私泄露风险度量可为隐私泄露风险量化评估,差分隐私保护机制评价提供有力支撑.

4.3.3 数据效用

数据效用度量差分隐私噪声信道输出合成数据集副本与原始数据集的失真距离,非负的失真函数 $d(\boldsymbol{X}^n, \hat{\boldsymbol{X}}^n) \to \mathbb{R}^+$ 度量合成数据集与原始输入数据集序列的失真程度,衡量发布数据集的整体可用性. 由于数据集混合数值型与类别型属性,通常总是假设 $\mathcal{X} = \hat{\mathcal{X}}$.

汉明失真度量输入与输出序列对应位置不同符号数目,是一种极端的失真度量方式. 其优势在于无论差分隐私机制输入与输出符号改变多小,基于汉明失真的效用度量能够维持较高的灵敏度,尤其对于类别型数据很有意义. 此外,平均意义的汉明失真度量发布合成数据集与原始数据集的平均失真程度,定义发布合成数据集的期望汉明失真函数

$$\mathrm{E}[d(\boldsymbol{X}, \hat{\boldsymbol{X}})] = \Pr\{\boldsymbol{X} \neq \hat{\boldsymbol{X}}\} \tag{9}$$

其中 $\Pr\{\boldsymbol{X} \neq \hat{\boldsymbol{X}}\}$ 是 $\boldsymbol{X}$ 与 $\hat{\boldsymbol{X}}$ 之间的误差概率. 利用期望汉明失真,可以从隐私信源、差分隐私噪声信道、合成扰动数据集等方面度量差分隐私保护系统整体上平均信息失真程度,进而有效的对发布合成数据集副本与原始数据集效用进行度量.

4.3.4 关联属性的隐私度量

关联属性图模型是一种有效表达多属性关联的方式,在图 G 中考虑与敏感属性(记作随机变量 X_h,基数 m_h)直接关联的属性组 $\boldsymbol{R}_h$. 假设关联属性组 $\boldsymbol{R}_h$ 为数据发布者和用户已有的知识,考虑敏感属性 X_h 编码经过差分隐私信道输出 $\hat{X}_h$ 的情况,则有信道特性 X_h 与 $\hat{X}_h$ 构成条件概率依赖关系. 由于差分隐私信道输出的条件分布仅依赖敏感属性 X_h 的分布,而与关联属性组 $\boldsymbol{R}_h$ 条件独立,故有:

定理 2 无向图 $G = (V, E)$ 中,敏感属性顶点 X_h 及其直接关联属性组 $\boldsymbol{R}_h$,与差分隐私信道输出 $\hat{X}_h$ 的概率依赖构成马尔可夫隐私链 $\boldsymbol{R}_h \to X_h \to \hat{X}_h$ 关系.

马尔可夫隐私链关系中,关联属性组集合$\boldsymbol{R}_h$的联合概率视为初始状态,条件概率$P(X_h/\boldsymbol{R}_h)$表述关联属性组联合概率分布条件下敏感属性的一步状态转移概率矩阵MP,进一步利用条件熵$H(X_h/\boldsymbol{R}_h)$度量隐私关联的不确定度. 此外,考虑差分隐私机制$P_{\hat{X}|X}(\hat{x}|x)$信息论噪声信道的条件概率转移矩阵具有$P(\hat{X}_h|X_h)$的形式. 其中,信道条件转移概率$p(\hat{x}_h^j|x_h^i)$表示隐私属性$\mathscr{X}_h$取值空间第i个值转移输出空间$\hat{\mathscr{X}}_h$第j个值的概率. 由此,差分隐私机制$P_{\hat{X}|X}(\hat{x}|x)$满足ε_{DP}^*差分隐私,则有

$$\varepsilon_{DP}^* = \min_{p(\hat{x}_h/x_h)} \log[p(\hat{x}_h^j|x_h^i)|p(\hat{x}_h^j|x_h^t)], \forall x_h^i \neq x_h^t \quad (10)$$

由式(8),隐私泄露风险是攻击者观察合成数据集输出概率分布$P(\hat{X}_h)$获得有关敏感信息的互信息量$I(\hat{X}_h;X_h)$. 依据数据处理不等式,则有关联属性组与敏感属性之间的互信息量满足$I(\boldsymbol{R}_h;\hat{X}_h) \leqslant I(\boldsymbol{R}_h;X_h)$,即是隐私关联属性组包含的隐私信息上界. 当$X_h$与$\hat{X}_h$独立时,类似于差分隐私信道通信中断,$I(\boldsymbol{R}_h;\hat{X}_h)$达到下限值0. 此外,对于满足失真度$D=\Pr(X_h \neq \hat{X}_h)$的差分隐私试验信道,互信息隐私泄露量满足

$$I(X_h;\hat{X}_h) = H(X_h) - H(X_h|\hat{X}_h) \quad (11a)$$

$$\geqslant H(X_h) - H(D) - D\log_2(m_h - 1) \quad (11b)$$

当$\mathscr{X}_h = \hat{\mathscr{X}}_h$时,依据费诺不等式易证式(11b),互信息的隐私泄露度量严格依赖于差分隐私噪声信道转移概率$P(\hat{X}_h|X_h)$与数据原始概率分布$P(X_h)$. 此时,依据失真理论的数据效用度量式(9)改变为$D=\Pr(X_h \neq \hat{X}_h)$,量化差分隐私噪声信道输出数据$\hat{X}_h$与原始数据$X_h$的期望汉明失真度. 由此易知,数据效用度量依赖于差分隐私信道转移概率矩阵的误差概率,是差分隐私信道统计特性.

5　实例分析

实验中采用机器学习 Adult① 公开数据集,选取数据集中 Age、Workclass、Education、Marital-status、Occupation、Race、Sex 七个属性,记作$X_1, X_2, \cdots, X_7$. 原始数据集包含数值型属性和类别属性,含有 30718 条数据记录,数据集域空间$\mathscr{X}$基数$m=7902720$. 基于4.1节方法统计样本数据二维属性对(X_i, X_j)频数,除以样本数据总和n得到频率意义的列联表. 依据大数定律,当n趋于无穷大时,二维属性联合频率近似于二维联合概率. 采用互信息量计算属性的相关度得到属性相关度矩阵$\boldsymbol{\Theta}$如下表2所示.

设置门限阈值参数$\delta=0.05$过滤属性伪相关现象,根据属性相关度矩阵,生成属性依赖图,如图1所示,图中属性为顶点,属性关联以无向图边表示. 从图1可见,属性关联图模型表达的属性关联信息是有效的.

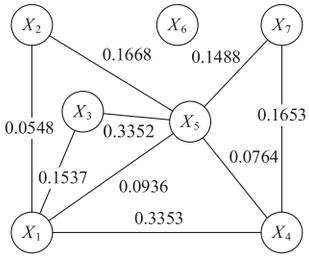

图1　属性关联依赖图

现假设X_4属性为敏感属性,则图中到X_4顶点存在路径则构成隐私链,与属性X_4邻接相关联属性组$\boldsymbol{R}_h=\{X_1, X_5, X_7\}$. 敏感属性$X_4$记作随机变量$X_h$,与敏感属性相关的关联属性组构成$\boldsymbol{R}_h$. 首先,统计样本观测数据集中关联属性组$\boldsymbol{R}_h$联合概率,则可以计算关联属性组联合信息熵$H(\boldsymbol{R}_h)=9.6712$. 其次,统计敏感属性与关联属性组的联合概率,并据此计算联合信息熵$H(\boldsymbol{R}_h X_h)=10.8469$. 由于原始数据集中敏感属性为类别型变量,且敏感属性域取值空间为七个不同的类别型数据,构成信源字母表空间. 依据其概率分布计算信息熵$H(X_h)=1.82$. 基于熵与互信息之间的关系,则可计算互信息量$I(\boldsymbol{R}_h;X_h)=0.6442$,即是由关联属性引起的敏感属性信息熵泄露量. 此外,条件熵$H(X_h|\boldsymbol{R}_h)$度量X_h尚存隐私不确定度,即$H(X_h)-I(\boldsymbol{R}_h;X_h)=1.1758$.

从信息论的角度分析马尔可夫模型的隐私泄露链$\boldsymbol{R}_h \to X_h \to \hat{X}_h$关系,则信息论差分隐私信道转移矩阵$P(\hat{x}_h|x_h)$构成马尔可夫状态转移矩阵. 依据数据处理不等式可得$I(\boldsymbol{R}_h;\hat{X}_h) \leqslant 0.6442$,即为隐私关联属性组泄露差分隐私扰动数据信息的上界. 此外,依据式(11b)费诺不等式度量互信息泄露风险量满足$I(X_h;\hat{X}_h) \geq 1.82 - H(D) - D\log_2 6$,其中$D=\Pr(X_h \neq \hat{X}_h)$.

针对敏感属性变量X_h,考虑对称离散信道情形[13],敏感属性字母表空间每个符号正确传递的概率为1-D,错误传输的概率为$D/6$,则此时差分隐私噪声信道转移概率矩阵为(7×7)阶对称矩阵. 依据式(10)计算信息论信道满足差分隐私参数,图2给出了差分隐私ε_{DP}^*与失真度的变化关系,与经典差分隐私定义保持一致性. 另一方面,对于已知的信源分布和特定的信道机制,利用式(11a)可以验证互信息泄露(MI)与失真度D的变化关系,图3表明随着失真度$D \to 0$时,互信息隐私泄露量$I(X_h;\hat{X}_h) \to H(X_h)$,验证了信息熵度量方法的有效性.

① http://archive.ics.uci.edu/ml/

表 2 属性相关度矩阵

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0	0.0548	0.1537	0.3353	0.0936	0.0097	0.0119
X_2	0.0548	0	0.0429	0.0272	0.1668	0.0102	0.0168
X_3	0.1537	0.0429	0	0.0308	0.3352	0.0147	0.0063
X_4	0.3353	0.0272	0.0308	0	0.0764	0.0185	0.1653
X_5	0.0936	0.1668	0.3352	0.0764	0	0.019	0.1488
X_6	0.0097	0.0102	0.0147	0.0185	0.019	0	0.0095
X_7	0.0119	0.0168	0.0063	0.1653	0.1488	0.0095	0

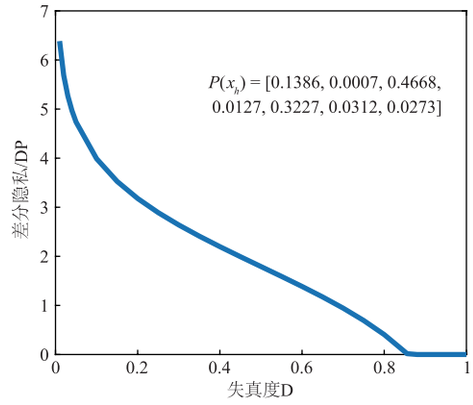

图2 差分隐私参数ε_{DP}^{*}与失真度关系

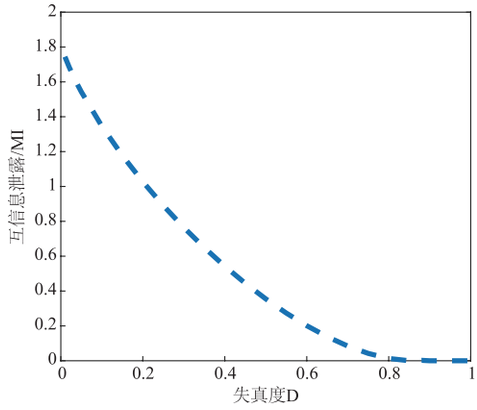

图3 互信息隐私泄露与失真度关系

文献[3]中考虑依据联合概率分布表达类别型属性之间具有的相关性,而对数值型属性且服从均值为 0 的高斯分布(如两相关联数值型随机变量 X 和 Y),方差 σ_X^2 和 σ_Y^2,则采用经典的皮尔逊相关系数 $\rho_{XY}=E[XY]/(\sigma_X\sigma_Y)$ 度量两属性变量之间的相关度. 然而,由于皮尔逊相关系数方法适用于两两变量的线性相关性且变量总体满足或接近高斯分布的特点,在刻画非线性关系和非数值文本相关性中存在不足. 对比文献中的方法,本文中采用的方法具有以下优势:(1)本文的研究面向关系数据集发布,针对更具体的差分隐私保护机制;(2)基于互信息方法分析多维属性关联的相关度,是数值型变量线性相关分析的推广与延伸,能够克服应用于数据集混合数值型和类别型属性的局限性,表达更复杂的相关关系;(3)基于样本观测数据计算属性相关度,设置门限阈值消除属性伪相关影响,并据此得到属性关联依赖图,进一步将其划分隐私关联属性组和隐私敏感属性,相比于主观划分更有科学理论依据.

6 结束语

本文针对差分隐私非交互式数据发布场景,立足 Shannon 信息论提出了面向数据集关联属性的隐私度量模型及方法. 通过以属性为离散随机变量,元组为随机向量,构建了一种属性关联、元组记录独立同分布的离散无记忆信源的 n 次扩展信源. 围绕隐私传播通信模型分别从隐私信源熵、隐私度与隐私泄露风险、数据效用的角度给出了具体的信息熵量化方法. 最后,以具体实例验证了提出模型方法的有效性,并进一步基于马尔可夫链分析了关联属性导致的隐私泄露量,对比分析了该方法的优势.

下一步工作中拟基于失真度、互信息隐私泄露研究保真度准则下差分隐私效用与隐私平衡的最佳试验信道机制问题,力图为平衡隐私泄露风险与数据效用提供一种基于信息论的理论支撑.

参考文献

[1] Dwork C. Differential privacy[A]. International Colloquium on Automata, Languages, and Programming[C]. Berlin, Heidelberg: Springer, 2006. 1 – 12.

[2] Dwork C, Mcsherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[J]. Lecture Notes in Computer Science, 2006, 3876(8): 265 – 284.

[3] Sankar L, Rajagopalan S R, Poor H V. Utility-privacy tradeoffs in databases: An information-theoretic approach[J]. IEEE Transactions on Information Forensics & Security, 2013, 8(6): 838 – 852.

[4] Kalantari K, Sankar L, Sarwate A D. Robust privacy-utility tradeoffs under differential privacy and hamming distortion[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2816 – 2830.

[5] Alvim MS, Chatzikokolakis K, Degano P, Palamidessi C. Differential privacy versus quantitative information flow[R]. Technical Report, 2010.

[6] Barthe G, Kopf B. Information-theoretic bounds for differentially private mechanisms[A]. Computer Security Foundations Symposium[C]. IEEE, 2011. 191 – 204.

[7] Pin Calmon F D, Fawaz N. Privacy against statistical inference[A]. Proceedings of 50th Annual Allerton Conference on Communication, Control, and Computing[C]. Monticello IL, USA, 2012. 1401 – 1408.

[8] Darakhshan J Mir. Information-Theoretic Foundations of

Differential Privacy[M]. Foundations and Practice of Security. Berlin Heidelberg:Springer,2013. 374 – 381.

[9] Alvim M S, Andrés M E, Chatzikokolakis K, et al. On the information leakage of differentially-private mechanisms [J]. Journal of Computer Security, 2015, 23(4): 427 – 469.

[10] 彭长根,丁红发,朱义杰,田有亮,符祖峰. 隐私保护的信息熵模型及其度量方法[J]. 软件学报,2016,27(8):1891 – 1903.
Peng C G, Ding H F, ZHU Y J, Tian Y L, Fu Z F. Information entropy models and privacy metrics methods for privacy protection[J]. Journal of Software, 2016, 27(8): 1891 – 1903. (in Chinese)

[11] Wang W, Ying L, Zhang J. On therelation between identifiability, differential privacy, and mutual-information privacy[J]. IEEE Transactions on Information Theory, 2016, 62(9):5018 – 5029.

[12] Cuff P, Yu L. Differentialprivacy as a mutual information constraint[A]. ACM SIGSAC Conference on Computer and Communications Security[C]. New York, NY, USA, 2016. 43 – 54.

[13] Alvim M S, Andrés M E, Chatzikokolakis K, et al. Differentialprivacy: On the trade-off between utility and information leakage[A]. International Workshop on Formal Aspects in Security and Trust [C]. Berlin, Heidelberg: Springer, 2012. 39 – 54.

[14] Shannon C E. A mathematical theory of communication [J]. Bell System Technical Journal, 1948, 27(3): 379 – 423.

[15] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets[J]. Science, 2011, 334(6062):1518 – 1524.

[16] 梁吉业,冯晨娇,宋鹏. 大数据相关分析综述[J]. 计算机学报,2016,39(1):1 – 18.
Liang Ji-ye, Feng Chen-jiao, Song Peng. Asurvey on correlation analysis of big data[J]. Chinese Journal of Computers, 2016, 39(1): 1 – 18. (in Chinese)

作者简介

吴宁博　男,1989 年生,河南驻马店人,博士研究生,主要研究方向为数据安全、隐私保护.
E-mail:hn_dragon@163.com

彭长根(通信作者)　男,1963 年生,贵州锦屏人,博士、教授、博士生导师,CCF 会员,主要研究方向为密码学、信息安全、大数据隐私保护等.
E-mail:peng_stud@163.com