计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200122

基于本地化差分隐私的键值数据关联分析

孙 林 平国楼 叶晓俊
清华大学软件学院 北京 100084
(sunl16@mails.tsinghua.edu.cn)

摘 要 在群智感知系统中,从分布式数据源中持续收集和分析数据可以为先进的数据挖掘模型提供决策支持。由于数据中可能包含个人相关的信息,数据的采集和分析过程中通常伴随着隐私泄露的风险。本地化差分隐私作为先进的隐私保护方案可在用户的隐私性和数据的可用性之间提供较好的权衡。当前,键值数据作为异构类型数据,其同时含有分类数据和数值数据,基于本地化差分隐私在多维度下对键值数据进行关联分析面临着一定的挑战。针对隐私保护前提下键值数据的发布和关联分析问题,首先定义了键值数据的频率关联和均值关联问题,然后提出了适用于键值对的索引独热编码,为键值数据提供本地化差分隐私保护,最后在扰动的数据上对键值数据进行关联分析。基于仿真数据集和真实数据集的实验和理论分析验证了所提方案的有效性。

关键词:本地化差分隐私;键值数据;关联分析;均值估计;频率估计
中图法分类号 TP391

Correlation Analysis for Key-Value Data with Local Differential Privacy

SUN Lin,PING Guo-lou and YE Xiao-jun
School of Software,Tsinghua University,Beijing 100084,China

Abstract Crowdsourced data from distributed sources are routinely collected and analyzed to produce effective data-mining models in crowdsensing systems. Data usually contains personal information,which leads to possible privacy leakage in data collection and analysis. The local differential privacy (LDP) has been deemed as the de facto measure for trade-off between privacy guarantee and data utility. Currently,the key-value data is a kind of heterogeneous data types in which the key is categorical data and the value is numerical data. Achieving LDP for key-value data is challenging. This paper focuses on key-value data publishing and correlation analysis under the framework of LDP. Firstly,the frequency correlation and mean correlation in key-value data are defined. Then the indexing one-hot perturbation mechanism is proposed to provide LDP guarantees. At last,the correlation results can be estimated in the perturbed space. Theoretical analysis and experimental results on both real-word and synthetic dataset validate the effectiveness of proposed mechanism.

Keywords Local differential privacy,Key-value data,Correlation analysis,Mean estimation,frequency estimation

1 引言

在群智感知环境下,广泛的数据收集与分析可以提供业务支持和智能决策。然而,数据的发布通常伴随着潜在的隐私泄露风险[1-2]。近年来,差分隐私(Differential Privacy,DP[3])概念被视为隐私保护下数据收集与分析的标准方法,因此被广泛用于深度学习系统[4-5]和统计分析[6]中。然而,DP的部署需要一个可信的数据收集者,这在分布式环境下通常难以实现。在这样的背景下,本地化差分隐私(Local Differential Privacy,LDP[7])被提出,用于分布式环境下数据的收集与分析任务。通过添加噪声以及扰动的方式,LDP保证不同的数据可以以接近相同的概率编码到同一个数据,因此保护了数据隐私。

通常来说,数据收集与分析场景中包含了n个分布式用户和一个不可信的数据分析者。LDP协议可分为编码、扰动、汇聚分析3个步骤。首先每个用户将本地数据编码到指定类型(如布隆过滤器[8]),然后用户采用满足LDP的算法对编码的数据进行扰动,并将扰动之后的数据发送给数据收集者,最后数据收集者对用户扰动之后的数据进行汇聚,并进行统计信息分析。

已有的LDP算法通常关注不同的数据类型或不同的数据分析任务。当前,LDP已经在分类数据[9]、数值数据[10-11]和集合数据[12-13]等数据类型上被用于频率/均值估计[12,14]和边缘概率发布[15]等任务。随着非关系型数据库的兴起,键值

到稿日期:2020-12-13 返修日期:2021-05-13 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划项目(2019QY1402)
This work was supported by the National Key Research and Development Program of China(2019QY1402).
通信作者:叶晓俊(yexj@tsinghua.edu.cn)

数据(key-value data[16-17])已经成为数据存储和分析的基本数据类型。基于 LDP 的键值数据分析受到了广泛关注。键值数据为异构类型数据,在键值对中,键是分类数据,值是数值数据,并且值的存在依赖于键。因此,为键值数据提供 LDP 的保护需要考虑键值之间的依赖关系。在此背景下,PrivKV[16]定义了键值数据中的频率分析和均值分析目标,并提出了一种考虑键值关系的扰动方法,以提供 LDP 保证。

然而,当前基于 LDP 的键值数据分析方法仅支持单个键值的数据分析。随着数据挖掘技术的不断发展,不同属性之间的数据(如边缘概率[18]、列联表[19])分析也成为了数据驱动应用的重要组成部分。在键值数据中,PrivKV 及其衍生方法(PCKV)无法提供对不同键之间的数据分析。为了应对键值数据分析中的此类挑战,本文提出了键值数据中的关联分析。图 1 以医疗数据为例给出了键值数据中的关联分析。

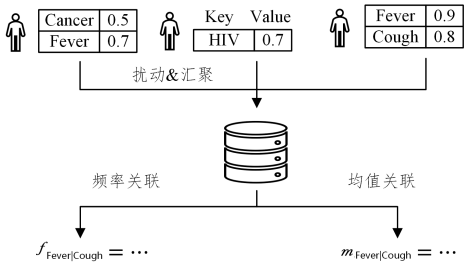


图 1　键值数据的关联分析

Fig. 1　Correlation analysis for key-value data

用键值数据来表示每个用户的数据,其中键表示是否患有特定症状,而值表示对应症状的严重程度。已有的键值数据分析仅支持单个键的频率分析和均值分析。而本文提出的关联分析支持不同键之间的统计信息发现,包含频率关联和均值关联两类。频率关联支持不同键之间的条件概率分析,均值关联支持给定键条件下的均值估计。以咳嗽和发热为例,频率关联支持分析在咳嗽情况下发热的可能性,均值关联支持分析在咳嗽情况下发热的平均严重程度。二者可以为先进的诊断系统提供更深层次的决策支持。

键值数据中的关联分析主要有以下几方面的问题。首先,在一个键值对$\langle k,v\rangle$中,值 v 的存在性依赖于键 k 的存在性,并且键值的数据类型是不同的。其次,关联分析涉及不同的键和值的对应关系,而已有的支持关联分析的方法仅支持二值数据[18]。为了解决此类问题,本文首先提出了状态编码,用于对单个键值对的状态进行表示,然后严格定义了两种键值数据的关联分析,即频率关联和均值关联,最后提出了索引独热(Indexing One Hot,IOH)编码,用于 LDP 框架下的键值数据关联分析。基于仿真数据集和真实数据集的实验结果和理论分析表明了本文方案的有效性。

2　研究背景与问题定义

2.1　本地化差分隐私

在信息共享的过程中,数据中的隐私是难以量化的。本地化差分隐私的思想来源于一个很朴素的观察:对于一个输出空间为 Range(M)的编码机制 M 来说,如果任何不同的输入编码到同一个输出的可能性是相近的,那么数据观察者就无法根据接收到的数据判断原始数据,因此数据的隐私得到了保护。基于这个思想,本地化差分隐私的定义如定义 1 所示。

定义 1(ε-LDP[14,17,20])　一个扰动机制 M 是满足 ε-LDP 的,当且仅当对于任意的输入t_1和t_2以及输出 $y\in\text{Range}(M)$,有:

$$\Pr[M(t_1)=y]\leqslant e^{\varepsilon}\cdot\Pr[M(t_2)=y]$$

其中,ε 为隐私预算,ε 越接近于 0,扰动机制 M 提供的隐私保护程度就越高。

随机响应[21]为常用的达到 LDP 的方法,其核心思想为对于二元数据以一定的概率发送真实数据,以一定概率发送与真实数据相对的值。由于随机响应只适用于二元数据,一元编码(Unary Encoding,UE[22])被提出,用于在输入空间为多个值的情况下进行扰动。对于独热编码之后的向量 $\boldsymbol{B}$,UE 的扰动过程为:

$$\Pr[\boldsymbol{B}'[i]=1]=\begin{cases}p, & \text{if } \boldsymbol{B}[i]=1\\ q, & \text{if } \boldsymbol{B}[i]=0\end{cases}$$

根据 LDP 的定义,UE 提供 $\ln\frac{p\cdot(1-q)}{q\cdot(1-p)}$-LDP。根据参数 p 和 q 设置的不同,UE 可分为对称一元编码(Symmetric UE,SUE)和最优一元编码(Optimal UE,OUE)。其中,在 SUE 中,$p=\frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}$,$q=\frac{1}{e^{\varepsilon/2}+1}$;在 OUE 中,$p=\frac{1}{2}$,$q=\frac{1}{e^{\varepsilon}+1}$。在频率估计任务中,OUE 方法具有最优的最小估计方差。

2.2　问题定义

在基于 LDP 的键值数据收集和分析框架中,我们假定用户集为$U=\{u_1,u_2,\cdots,u_n\}$,键空间为$K=\{k_1,k_2,\cdots,k_d\}$。用户 u_i拥有一个或多个键值对,记为S_i。其中,每个键值对都用$\langle k,v\rangle$的形式表示。用户 u_i的第 j 个键值对记为$\langle k_{i,j},v_{i,j}\rangle$,当用户不具有对应的键值对时,对应的键值对可表示为$\langle 0,-\rangle$。不失一般性,已有的方案中均假设键值对中值的范围为$v_i\in[-1,1]$。给定条件 C 表示对键的条件约束(其定义详见第 3 节),U^C表示满足条件 C 的用户集合。本文考虑以下两类键值数据的关联分析。

(1)频率关联:计算条件 C 下键 k 对应的键值对的频率,其定义为:

$$f_{k|C}=\frac{|\{u_i\mid u_i\in U^C\wedge\exists\langle k,v\rangle\in S_i\}|}{|U^C|}$$

(2)均值关联:计算条件 C 下键 k 对应的键值对中值的均值,其定义为:

$$m_{k|C}=\frac{\sum v\mid\exists\langle k,v\rangle\in S_i\wedge u_i\in U^C}{|\{u_i\mid u_i\in U^C\wedge\exists\langle k,v\rangle\in S_i\}|}$$

3　键值数据的关联分析

在数据挖掘算法中,不同属性之间的数据分析对数据挖掘模型的构建必不可少。本节引入键值数据中的关联分析,首先定义了键值数据中的频率关联和均值关联。在此基础上,我们提出了索引独热 IOH 算法用于 LDP 下的键值数据编码,然后从分析者的角度基于扰动的数据进行关联分析。

3.1 键值数据的状态编码

基于 PrivKV 的编码方案首先对键值对$\langle k,v\rangle$中的键和值分别采用随机响应的方法进行编码。由于在键值数据中，键不存在时，值没有含义，因此 PrivKV 方法在编码值的过程中需要考虑键的编码结果。通过状态化步骤，我们将键值对中的键和值进行绑定，从而在扰动的过程中保留单个键值对中值和键的依赖关系。

如图 2 所示，对于键 k，若用户 u_i 的数据满足 $k\in S_i$，则该键值对记为$\langle 1,v^*\rangle$，否则记为$\langle 0,-\rangle$。当 k 存在时，由于值的空间是连续的，首先需要对值进行离散化操作；当 k 不存在时，离散化的结果记为$\langle 0,0\rangle$。为了保证无偏性，将键值对中值的离散化过程表示为：

$$Discretization(v)=\begin{cases}1, & \text{w. p. } (1-v)/2\\ -1, & \text{w. p. } (1+v)/2\end{cases}$$

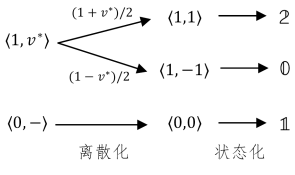


图 2 键值对的离散化和状态化

Fig. 2 Discretization and state for key-value data

对于 k 存在的情况，首先根据值的情况对其进行离散化以保证无偏性。经过离散化，可能的键值对状态为$\langle 0,0\rangle$，$\langle 1,-1\rangle$，$\langle 1,1\rangle$ 3 种情况。为了方便计算机编码以及后文中利用状态进行索引，我们将这 3 种状态记为 1，0，2。因此，离散化和状态化的过程可以简述为：

$$State(\langle k,v\rangle)=k\cdot Discretization(v)+1$$

表 1 列出了一个医疗系统的案例，其中用户为 $U=\{$用户 A，用户 B，用户 $C\}$，键的空间为 $K=\{$癌症，发热，咳嗽$\}$，对应的值表示病人此种疾病的严重程度(为了方便描述，假设键值对中的值已经经过离散化)。以用户 C 为例，其对应的键值对为$\langle 0,0\rangle$，$\langle 1,-1\rangle$，$\langle 1,-1\rangle$，表示该用户不患有癌症但患有发热和咳嗽，并且发热和咳嗽的程度皆为-1。

表 1 键值数据案例表

Table 1 Example of key-value data

用户	$\langle$癌症，$v\rangle$	$\langle$发热，$v\rangle$	$\langle$咳嗽，$v\rangle$
用户 A	$\langle 1,1\rangle$	$\langle 0,0\rangle$	$\langle 1,-1\rangle$
用户 B	$\langle 1,-1\rangle$	$\langle 1,1\rangle$	$\langle 1,1\rangle$
用户 C	$\langle 0,0\rangle$	$\langle 1,-1\rangle$	$\langle 1,-1\rangle$

在关联分析中，不同属性(即不同键)之间的关联关系可以为决策提供数据支撑。如在临床诊断中，对于到医院就诊的人来说，发热的可能性大致为 1%。然而，如果知道某个病人已经咳嗽，则其统计意义上发热的概率就可能高达 15%。发热和咳嗽之间的相关关系可以帮助医生更好地进行针对性的治疗。在键值数据中，本文主要关心以下两类关联分析。

(1)频率关联：描述不同键之间的关联关系。如在给定咳嗽的情况下，患者发热的概率是多大？该问题可形式化为：$f_{\text{发热}|\text{咳嗽}=1}$，其中下标“发热|咳嗽$=1$”表示在已知咳嗽的情况下求发热的概率。

(2)均值关联：描述值与不同键的关联关系。如在给定咳嗽但并未患有癌症的情况下，患者发热的平均严重程度是多大？该问题可形式化为：$m_{\text{发热}|\text{咳嗽},\text{癌症}=1,0}$。其中，下标“发热|咳嗽，癌症$=1,0$”表示在患有咳嗽且不患癌症的情况下发热的均值。

本文用以下形式来表示条件 C：$C=\{ck_1,\cdots,ck_L=c_1,\cdots,c_L\}$，其中 ck_i 表示某个键，c_i 表示键 ck_i 是否存在。如“咳嗽，癌症$=1,0$”中给定了两个条件，因此 $L=2$，对应的条件中，键分别为 $ck_1=$咳嗽，$ck_2=$癌症。同时，此条件指定了 $c_1=1$，$c_2=0$，即表明患有咳嗽并且不患有癌症。

由于给定的关联条件是无序的，如“咳嗽，癌症$=1,0$”与“癌症，咳嗽$=0,1$”所表示的含义一致，本文首先定义(α,β)-条件用于对关联分析中的条件进行描述。

定义 2 (α,β)-条件表示：关联分析中，条件 C 可等价于一个(α,β)元组，其中，α 表明哪些键是存在于条件 C 中，β 表明了对应键的存在性。对于键空间 $K=\{k_1,k_2,\cdots,k_{|K|}\}$ 来说，有 $\alpha_{i;k_i\in ck_1,ck_2,\cdots,ck_L}=1$，且 $\beta_i=c_{j;\exists ck_j=k_i}$。

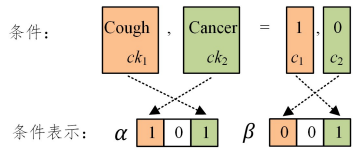


图 3 键值数据的条件表示

Fig. 3 Conditional representation for key-value data

条件表示可以使得无序的条件按照有序的方式进行组合。如图 3 所示，在表 1 中，键空间为 $K=\{$癌症，发热，咳嗽$\}$。“咳嗽，癌症$=1,0$”与“癌症，咳嗽$=0,1$”用(α,β)-条件表示均为：(101,001)。在关联分析中，条件表示可以对满足给定条件的用户进行筛选。在非交互式的分布式环境下，用户端对数据编码时无法准确地洞悉数据分析者对数据的分析要求。因此，用户需要将本地的键值对都进行编码，为可能的关联分析提供数据支持。为了保证用户信息的隐私性，我们需要对编码之后的数据进行扰动。为了提高数据的有效性，不同用户在进行编码之后需要在一定程度上满足相似性。基于以上两个原则，本文设计了适用于键值数据的索引独热 IOH 编码。利用 IOH 对键值数据编码的过程如图 4 所示。

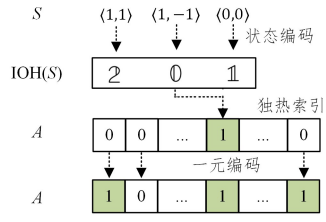


图 4 索引独热编码

Fig. 4 Indexing one-hot mechanism

在编码的过程中，用户先将本地的键值数据进行离散化并以此对每个键值对进行状态化。状态化之后的键值数据为一个$|K|$维向量。基于状态化之后的向量，用户本地采用独热编码的方式对键值集合 S_i 进行编码。即用户新建向量 $\boldsymbol{A}_i=\{0\}^{3^d}$，并令 $\boldsymbol{A}_i[\text{IOH}(S_i)]=1$。其中，索引值 IOH 为：

$$IOH(S_i)=\sum_{j=1}^{d}3^{d-j}\cdot State(\langle k_{i,j},v_{i,j}\rangle)$$

索引 IOH 是根据用户本地所有的键值对计算得到的，因此编码的向量 $\boldsymbol{A}$ 包含了所有键值对的信息，故此编码方案可

用于关联分析。根据独热编码的性质，对于不同用户 u_i，u_j，其独热编码的结果 $\boldsymbol{A}_i$，$\boldsymbol{A}_j$ 最多只有两个比特位不同，因此可以用 UE 方法对键值对进行保护以实现 LDP。即每个用户 u_i 利用参数 p，q 对 $\boldsymbol{A}_i$ 进行扰动，并将扰动之后的向量发送给数据收集者。用 IOH 对用户键值对编码的过程如算法 1 所示。

算法 1　索引独热编码(IOH)

输入：键值对列表 S，概率参数 p，q

输出：比特串 **A**

1. 计算 IOH(S)

2. 初始化空向量：$\mathbf{A}=[0,0,\cdots,0]$，$|\mathbf{A}|=3^d$

3. 令 $\mathbf{A}[\mathrm{IOH}(S)]=1$

4. for i←1 to 3^d do：

$$\Pr[\mathbf{A}[i]=1]=\begin{cases}p, & \text{if } \mathbf{A}[i]=1\\ q, & \text{if } \mathbf{A}[i]=0\end{cases}$$

5. 将扰动结果 **A** 共享给数据收集者。

在聚合了所有用户的扰动结果 $\boldsymbol{A}$ 之后，数据分析者对其进行求和以及无偏矫正，即：

$$\boldsymbol{A}_s[j]=\frac{\sum_i \boldsymbol{A}_i[j]-nq}{p-q}$$

其中，$\boldsymbol{A}_i[j]$表示用户 u_i 扰动后的向量 $\boldsymbol{A}_i$ 的第 j 位，并且 p，q 为 SUE 或 OUE 参数；$\boldsymbol{A}_s$ 表示求和向量，是对未扰动求和向量的无偏估计。下文将基于求和向量 $\boldsymbol{A}_s$ 求解键值数据的频率关联和均值关联。

3.2　键值数据的频率关联

由于在用户编码后的数据 $\boldsymbol{A}_i$ 和求和向量 $\boldsymbol{A}_s$ 中，每一位的值都包含了所有键值对的信息，在频率关联分析中需要计算出满足 (α,β)-条件的用户在 $\boldsymbol{A}_i$ 中的索引。频率关联分析的是在给定条件下满足某个键存在的情况，此类问题可转化为求满足某种条件的用户的数量并利用数量的比值表示频率关联。因此，我们定义频率计数器的操作，如定义 4 所示。

定义 4(频率计数器)　给定求和向量 $\boldsymbol{A}_s$ 和 (α,β)-条件，频率计数器 $F_\beta^\alpha[\boldsymbol{A}_s]$ 从 $\boldsymbol{A}_s$ 中统计满足条件 (α,β) 的用户数量。

以分析咳嗽和发热的关系为例，即 $f_{\text{发热}|\text{咳嗽}=1}$，其可等价于计算满足条件“发热，咳嗽＝1，1”的用户数量和满足条件“咳嗽＝1”的用户数量的比值。其中，条件“发热，咳嗽＝1，1”等价于条件(011，011)，条件“咳嗽＝1”等价于(001，001)。因此有 $f_{\text{发热}|\text{咳嗽}=1}=F_{011}^{011}[\boldsymbol{A}_s]/F_{001}^{001}[\boldsymbol{A}_s]$。不失一般性，在给定条件 (α,β) 下，键 k 的频率关联可表示为(其中 $\alpha\vee\alpha[k]=1$ 表示将 α 的第 k 位置 1，β 同理)：

$$f_{k|C=(\alpha,\beta)}=\frac{f_{\beta\vee\beta[k]=1}^{\alpha\vee\alpha[k]=1}[\boldsymbol{A}_s]}{f_\beta^\alpha[\boldsymbol{A}_s]}$$

上式表明，频率关联可通过不同参数的两个频率计数器进行求解。因此，我们关注形如 $F_\beta^\alpha[\boldsymbol{A}_s]$ 的频率计数器的计算。条件 α 中，若对于键 i 有 $\alpha[i]=0$，则说明条件 C 中未指定键 i 的存在情况，因此我们需要对第 i 个键进行分类讨论。如在条件(011，011)中，第一个键对应的 α 为 0，因此可分为患有癌症和不患有癌症两种情况，即 $F_{011}^{011}[\boldsymbol{A}_s]=F_{011}^{111}[\boldsymbol{A}_s]+F_{111}^{111}[\boldsymbol{A}_s]$。简单起见，当 $F_\beta^\alpha[\boldsymbol{A}_s]$ 中 α 全为 1 时，记作 $f_\beta[\boldsymbol{A}_s]$。因此，$F_\beta^\alpha[\boldsymbol{A}_s]$ 可以按照以下方式展开：

$$F_\beta^\alpha[\boldsymbol{A}_s]=\sum_{\gamma,\gamma\wedge\alpha=\beta}F_\gamma[\boldsymbol{A}_s]$$

频率计数的展开中，我们将 (α,β)-条件中未指定的键转换成了指定的键 $F_\gamma[\boldsymbol{A}_s]$ 进行计算。γ 中的第 i 项表示了第 i 个键值对中值的状态。在键值对状态化的过程中，键存在时，对应的键值对状态为{0，2}；当键不存在时，对应的键值对状态为 1。因此，在计算 $F_\gamma[\boldsymbol{A}_s]$ 时，对于每一个键的存在与否需要我们进行展开。d 维状态向量 $\boldsymbol{\gamma}$ 的索引的集合可由公式 $I(\boldsymbol{\gamma})$ 表示。$I(\boldsymbol{\gamma})$ 对应的 3 进制展开式为 $I(\boldsymbol{\gamma})=I(\gamma_1)|I(\gamma_2)|\cdots|I(\gamma_d)$。其中，“|”表示拼接操作，并且：

$$I(\boldsymbol{\gamma})=\begin{cases}\{0,2\}, & \text{if } \gamma_i=1\\ \{1\}, & \text{if } \gamma_i=0\end{cases}$$

如 $I(101)$ 的计算过程为：$I(101)=I(1)|I(0)|I(1)=\{3,5,21,23\}$。对于 F_γ，有：

$$F_\gamma[\boldsymbol{A}_s]=\sum_{i\in I(\boldsymbol{\gamma})}\boldsymbol{A}_s[i]$$

3.3　键值数据的均值关联

类似于频率关联，条件 C 下键 k 的均值关联 $m_{k|C=(\alpha,\beta)}$ 可等价于满足条件 $(\alpha\vee\alpha[k]=1,\beta\vee\beta[k]=1)$ 的用户数据的均值估计。因此，均值关联可表示为：

$$m_{k|C=(\alpha,\beta)}=\frac{S_{k|C}[\boldsymbol{A}_s]}{f_{\beta\vee\beta[k]=1}^{\alpha\vee\alpha[k]=1}[\boldsymbol{A}_s]}$$

其中，$S_{k|C}$ 为对满足条件 C 且键为 k 的键值对中值的求和。在状态统计中，键值对中有数值含义的值被编码到了 0 和 2 两种状态，分别表示键值对〈1，−1〉和〈1，1〉。因此，$S_{k|C}$ 的求和问题可转换为统计条件 C 中键值对为〈1，1〉和〈1，−1〉的个数，分别记为 $S_{k|C}^+$ 和 $S_{k|C}^-$。继而 $S_{k|C}$ 可表示为 $S_{k|C}[\boldsymbol{A}_s]=S_{k|C}^+[\boldsymbol{A}_s]-S_{k|C}^-[\boldsymbol{A}_s]$。类似于 $f_\beta^\alpha[\boldsymbol{A}_s]$ 对 (α,β) 的条件展开过程，对于 $S_{k|C}^+$ 和 $S_{k|C}^-$，有：

$$S_{k|C}^+=\sum_{\boldsymbol{\gamma},\boldsymbol{\gamma}\wedge\alpha=\beta}S_{k,\boldsymbol{\gamma}}^+[\boldsymbol{A}_s],S_{k|C}^-=\sum_{\boldsymbol{\gamma},\boldsymbol{\gamma}\wedge\alpha=\beta}S_{k,\boldsymbol{\gamma}}^-[\boldsymbol{A}_s]$$

对于 d 维状态向量 $\boldsymbol{\gamma}$，其对应 $\boldsymbol{A}_s$ 的索引的集合可分别由公式 $I^+(\boldsymbol{\gamma})$ 和 $I^-(\boldsymbol{\gamma})$ 表示，其 3 进制展开式分别为：

$$I^+(\boldsymbol{\gamma})=I(\gamma_1)|\cdots|I(\gamma_{k-1})|I^+(\gamma_k)|I(\gamma_{k+1})|\cdots|I(\gamma_d)$$

$$I^+(\boldsymbol{\gamma})=I(\gamma_1)|\cdots|I(\gamma_{k-1})|I^-(\gamma_k)|I(\gamma_{k+1})|\cdots|I(\gamma_d)$$

其中，$I^+(\gamma_k)$ 和 $I^-(\gamma_k)$ 为：

$$I^+(\gamma_k)=\begin{cases}\{2\}, & \text{if } \gamma_i=1\\ \{1\}, & \text{if } \gamma_i=0\end{cases}$$

$$I^-(\gamma_k)=\begin{cases}\{0\}, & \text{if } \gamma_i=1\\ \{1\}, & \text{if } \gamma_i=0\end{cases}$$

结合上文的内容，我们便可以计算出条件 C 下的键 k 的均值关联。频率关联和均值关联的主要区别为：对于键 k，频率关联只关注键是否存在并对键存在的用户进行计数，而均值关联中需要分别统计出值为 1 和 −1 的键值对的个数并进行求均值的操作。

总的来说，本节介绍了键值数据的条件分析。受到独热编码和 UE 机制的启发，本节提出了 IOH 扰动方案以对用户键值数据实现 ε-LDP，并支持编码结果中对频率关联和均值关联进行估计。

4　实验结果分析

本节通过一系列实验在 LDP 框架下对键值数据的关联分析效果进行衡量。

4.1　数据集及衡量方法

(1)数据集描述。本文同时采用了真实数据集 MovieLens 和仿真数据集 GAUSS 用于评价 IOH 方案的有效性。MovieLens 包含了 13.8 万个用户对 2.7 万部影片的评分，每

个用户至少含 20 个有效评分。我们提取评分数量最多的前 10 部电影作为键空间。同时每部电影的评分用$\langle k,v\rangle$数据表示，其中 k 表示电影 id，v 表示归一化到区间$[-1,1]$的评分。GAUSS 数据集中，键空间大小分别设置为 2,4,8，用户数量分别为 10^5 和 10^6，并且键和值均服从高斯分布。

(2)评价指标。我们采用平均绝对误差(MAE)作为频率关联和均值关联的估计误差，其定义为：

$$MAE=\sum_{T}|\theta-\hat{\theta}|^2$$

其中，θ 和 $\hat{\theta}$ 分别表示实验的真实和观测指标；T 表示实验次数，每一次实验结果对应 $T=50$ 次实验。

4.2 键值对的频率估计与均值估计

在键值数据的状态化中，我们将离散化的值与键进行了绑定，因此状态中同时包含了键和值的信息。基于此原理，我们首先利用广义的随机响应来实现单个键值数据的状态扰动(称为状态编码，state encoding)。对于状态 $t\in 1,0,2$，令 $p=e^{\varepsilon}/(e^{\varepsilon}+2)$，其核心扰动方程为：

$$\Pr[Perturb(t)]=\begin{cases}t, & \text{w. p. } p\\ \text{otherwise}, & \text{w. p. } (1-p)/2\end{cases}$$

数据收集者在收到扰动的状态后，对每个状态的键值对数量进行统计，记为 M_0，M_1，M_2，则频率 f 和均值 m 可以估计为：

$$\hat{f}=\frac{M_0+M_2}{n},\hat{m}=\frac{M_2-M_0}{M_0+M_2}$$

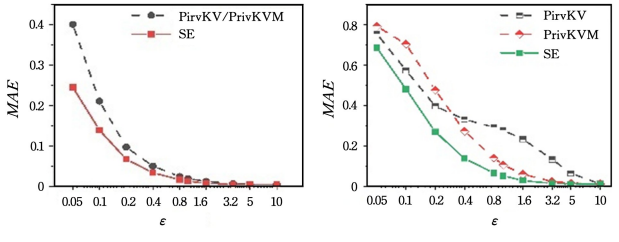
用于键值数据的不同方法的带宽如表 2 所列。其中，PrivKVM 为迭代的方法，需要交互地收集数据，在 SE 方法中，由于编码结果中只有 3 种状态，因此含有最低带宽。

表 2 PrivKV，PrivKVM，SE 方法的比较

Table 2 Comparison between PrivKV, PrivKVM and SE

Algorithm	Interactive	Bandwidth
PrivKV	No	$\log d+2$
PrivKVM(t rounds)	Yes	$t(\log d+2)$
SE	No	$\log 3d$

图 5 给出了不同方法在 MovieLens 数据集上的估计误差。我们尝试的隐私预算范围为 0.05～10。由于 PrivKV 与 PrivKVM 均采用随机响应去解决键值数据的频率估计，因此其含有同样的误差。随着隐私预算的提高，扰动中随机性下降，不同方法的误差均出现了明显的降低。对于图 5(a)的频率估计和图 5(b)的均值估计，SE 的误差均低于基于 PrivKV 的方法。



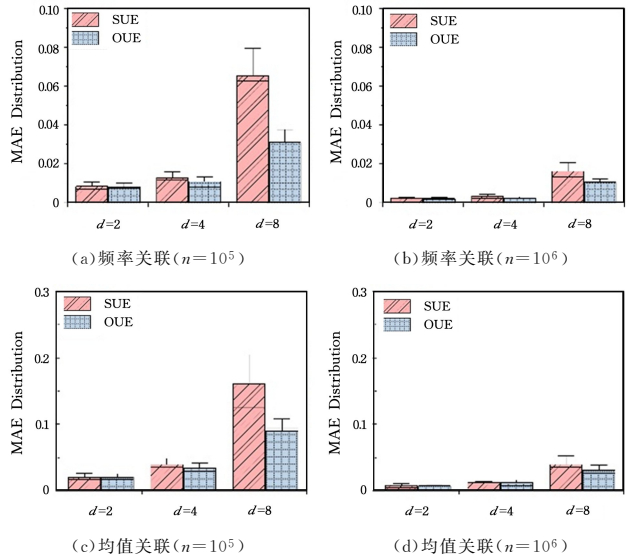
(a)频率估计 (b)均值估计

图 5 MovieLens 数据集上的频率/均值估计

Fig. 5 Frequency/mean estimation on MovieLens dataset

4.3 键值数据的关联分析

图 6 给出了不同数据量以及不同键空间大小下的关联分析误差。我们在 IOH 编码过程中分别采用了 SUE 和 OUE 对键值数据来实现 LDP。对比图 6(a)和图 6(b)以及图 6(c)和图 6(d)，随着数据量的增大，频率关联及均值关联的误差都有较明显的降低。对比图 6(a)和图 6(c)以及图 6(b)和图 6(d)，在同等数据量下，均值关联的误差相对于频率估计的误差较高。其本质原因为，在均值关联中，有值含义的键值对数量远远小于用户数。我们观察到在频率关联和均值关联的估计中，OUE 也优于 SUE。



(a)频率关联($n=10^5$) (b)频率关联($n=10^6$)

(c)均值关联($n=10^5$) (d)均值关联($n=10^6$)

图 6 GAUSS 数据集上的关联分析

Fig. 6 Correlation analysis on GAUSS dataset

结束语 本文在本地化差分隐私框架下，考虑数据收集与分析场景下的键值数据关联分析。本文首次定义了键值数据中的频率关联和均值关联分析。同时，本文提出了键值数据的状态化和 IOH 算法用于实现键值数据的扰动。基于真实数据集和仿真数据集的实验验证了本文方案的有效性。

IOH 扰动方法需要编码用户本地所有键值对，因此具备较高的空间复杂度。在未来的工作中，我们打算使用采样的方法对本地键值对进行筛选，以降低关联分析的时间和空间复杂度。

参 考 文 献

[1] YANG G M, YANG J, ZHANG J P. Research on Data Publishing of Privacy Preserving[J]. Computer Science, 2011(9): 17-23.

[2] WANG B, YANG J. Research on Anonymity Technique for Personalization Privacy-preserving Data Publishing[J]. Computer Science, 2012, 39(4): 168-171.

[3] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Theory of Cryptography Conference. Heidelberg: Springer, 2006: 265-284.

[4] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna: ACM, 2016: 308-318.

[5] PHAN N H, WANG Y, WU X, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior

prediction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Arizona:AAAI Press,2016:1309-1315.

[6] 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. Sydeny: ACM,2015:129-138.

[7] 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. Washington: IEEE,2013:429-438.

[8] BLOOM B H. Space/time trade-offs i n hash coding with allowable errors[J]. Communications of the ACM,1970,13(7):422-426.

[9] ERLINGSSON U,PIHUR V,KOROLOVA A. RAPPOR:Randomized aggregatable privacy-preserving ordinal response[C]// Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Arizona: ACM, 2014: 1054-1067.

[10] NGUYEN T T,XIAO X X,YANG Y,et al. Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy[EB/OL]. https://arxiv. org/abs/1606. 05053.

[11] WANG N,XIAO X K,YANG Y,et al. Collecting and Analyzing Multidimensional Data with Local Differential Privacy [C]// IEEE 35th International Conference on Data Engineering. Macau:IEEE Press,2019:638-649.

[12] QIN Z,YANG Y,YU T,et al. Heavy hitter estimation over set-valued data with local differential privacy[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna:ACM,2016:192-203.

[13] SWANG T H,LI N H,JHA S. Locally differentially private frequent itemset mining[C]// 2018 IEEE Symposium on Security and Privacy. San Francisco:IEEE Press,2018:127-143.

[14] BASSILY R,NISSIM K,STEMMER U,et al. Practical locally private heavy hitters[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York:Curran Associates Inc,2017:2285-2293.

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

[16] YE Q Q,HU H B,MENG X F. PrivKV:Key-Value Data Collection with Local Differential Privacy[C]// 2019 IEEE Symposium on Security and Privacy (SP). San Francisco:IEEE Press, 2019:317-331.

[17] GU X L,LI M,CHENG Y Q,et al. PCKV:Locally Differentially Private Correlated Key-Value Data Collection with Optimized Utility[C]// 29th USENIX Security Symposium. USENIX Association,2020:967-984.

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

[19] BARAK B,CHAUDHURI K,DWORK C,et al. Privacy,accuracy,and consistency too:a holistic solution to contingency table release[C]// Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. New York:ACM,2007:273-282.

[20] DING B L,KULKARNI J,YEKHANIN S. Collecting telemetry data privately[C]// Proceedings of the 31st International Conference on Neural Information Processing. California: ACM, 2017:3574-3583.

[21] DUCHI J C,JORDAN M I. Minimax Optimal Procedures for Locally Private Estimation[J]. Journal of the American Statistical Association,2018,113(521):182-201.

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

SUN Lin, born in 1993, Ph.D. His main research interests include privacy protection and data mining.

YE Xiao-jun, born in 1964, professor. His main research interests include cloud data management, data security and privacy, and database system testing.