基于本地差分隐私的K-modes聚类数据隐私保护方法

张少波[1,2],原刘杰[1],毛新军[2],朱更明[1]
(1. 湖南科技大学计算机科学与工程学院,湖南湘潭 411201; 2. 国防科技大学复杂系统软件工程重点实验室,湖南长沙 410073)

摘 要: 分类型数据聚类是数据挖掘的重要研究内容,聚类数据中通常包含用户一些敏感信息. 为保护聚类数据中的用户隐私,当前主要采用基于可信第三方隐私保护模型,但现实中第三方也存在隐私泄露风险. 针对此问题,该文引入本地差分隐私技术,提出一种去可信第三方的K-modes聚类数据隐私保护方法. 该方法首先利用随机采样技术对数据进行采样,然后使用本地差分隐私技术对采样数据进行扰动,最后通过聚类服务端与用户的交互迭代完成聚类. 在聚类过程中,无需可信第三方对数据进行隐私预处理,避免了第三方泄露用户隐私的风险. 理论分析证明了该方法的隐私性和可行性,实验结果表明该方法在满足本地差分隐私机制的前提下保证了聚类结果的质量.

关键词: 隐私保护; 本地差分隐私; 数据挖掘; K-modes聚类; 去可信第三方
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112(2022)09-2181-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20201374

Privacy Protection Method for K-modes Clustering Data with Local Differential Privacy

ZHANG Shao-bo[1,2], YUAN Liu-jie[1], MAO Xin-jun[2], ZHU Geng-ming[1]
(1. *School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan* 411201, *China*;
2. *Key Laboratory of Software Engineering for Complex Systems, National University of Defense Technology, Changsha, Hunan* 410073, *China*)

Abstract: Categorical data clustering is an important research content for data mining, and clustering data usually contains some sensitive information of user. In order to protect user privacy in clustering data, the privacy protection model based on trusted third-party is currently mainly adopted. However, in reality, the third-party also has the risk of privacy leakage. In this paper, we propose a privacy protection method for K-modes clustering data without trusted third-party by introducing local differential privacy technology. Our method first uses random sampling technology to sample the data, then perturbs the sampled data by using local differential privacy technology, and finally complete the clustering through the interaction between the server and the user. In the clustering process, our method does not require a trusted third-party to perform privacy preprocessing on the data, which avoids the risk of the third-party disclosing the user's privacy. Theoretical analysis proves the privacy and feasibility of our method. Experimental results show that our method guarantees the quality of the clustering results under the premise of satisfying the local differential privacy mechanism.

Key words: privacy protection; local differential privacy; data mining; k-modes clustering; no trusted third-party

1 引言

聚类是一种常用的数据挖掘方法,它按照某种特定标准将数据集分割为不同的簇,使得同一个簇中数据相似性较高[1]. K-means是聚类中的经典算法,其实现简单且聚类高效,它只适用于处理数值型数据集,而对分类型数据的聚类通常使用K-modes算法[2,3]. 目前,聚类分析在数据分析、服务推荐等多个领域发挥着重要的作用,但聚类数据中通常包含大量的个人敏感

收稿日期:2020-12-01;修回日期:2021-04-12;责任编辑:梅志强
基金项目:湖南省自然科学基金面上项目(No.2020JJ4317,No.2020JJ4250);湖南省教育厅科学研究重点项目(No.21A0318,No.19A275);湖南省研究生科研创新项目(No.CX20200999)

信息,如对客户数据进行聚类分析可为不同类型的客户提供个性化服务,但攻击者从这些信息中能推测出用户的兴趣爱好[4~6]. 因此在使用聚类对用户数据分析过程中,迫切需要保护用户的个人隐私.

差分隐私[7]是采用严格数学定义的隐私保护模型,它可以量化用户隐私保护程度,并且能抵御攻击者发起的背景知识攻击和合成攻击[8,9]. 目前差分隐私已广泛应用于数据挖掘、数据发布、位置服务等领域的隐私保护[10~12],而且它与随机扰动、数据交换等隐私保护技术相比[13,14],在聚类分析数据隐私保护方面具有明显的优势[15~17]. 差分隐私具有强健的隐私保护能力,但它需要一个可信第三方对数据进行处理,而由第三方造成的数据泄露事件却层出不穷,如谷歌,雅虎和微软等公司的数据意外泄漏事件,这不仅造成了用户隐私信息的泄露,还严重损害了公司声誉. 因此,现实中很难找到完全可信的第三方. 针对该问题,本地差分隐私(Local Different Privacy,LDP)在保证数据可用性的前提下,通过在用户端对数据进行扰动,实现了对用户隐私的去第三方保护,并且它与差分隐私同样采用了严格数学定义的隐私保护模型[18],因此它已成为隐私保护领域中解决此类问题的重要方法.

目前,国内外学者已提出一些本地差分隐私算法,其中RAPPOR算法是频数统计的经典算法[19],它误差小,数据可用性高,但它需要候选属性值已知. 针对RAPPOR的不足,文献[20]在其基础上对字符串进行映射,实现了无需候选属性值已知的频数统计. 针对集值数据的频繁项查询问题,文献[21]提出了包含两阶段机制的LDPMiner算法,文献[22]在其基础上进一步研究,提出了具有更高查询精度的SVIM(Set-Value Item Mining)算法. 不同于频率估计方法,文献[23]通过数据离散化操作实现了在[-1,1]区间中的均值估计,然而该方法的输出结果是两个固定值,这会导致估计值偏离[-1,1]区间. 针对该问题,文献[24]将[-1,1]区间中的任意值扰动到受约束区间[-C,C],然后在此区间内计算该扰动值的边界. 此外,本地差分隐私在空间范围查询,众包数据收集等领域也得到了广泛应用[25,26]. 虽然本地差分隐私可以有效应对第三方隐私泄露问题,但将它应用于聚类分析数据隐私保护时,仍然存在如下挑战:(1)如何降低噪声在聚类更新质心过程中的影响. 如果直接根据收集到的扰动数据对用户进行分簇,然后再基于扰动数据计算每个簇的质心,则会进一步放大噪声的影响. (2)如何以较小的噪声误差和通讯开销完成聚类. 如果用户将自身所有数据扰动并汇报,则所需的通讯开销以及聚类结果的噪声误差会较大. 针对上述挑战,本文提出一种基于本地差分隐私的K-modes聚类数据隐私保护方法(Local Different Privacy K-modes,LDPK). 该方法首先对数据随机采样,然后采用本地差分隐私技术在用户端对采样数据进行扰动,最后通过服务端与用户端的交互迭代完成聚类. 本文主要贡献如下:

(1) 构建了一个基于本地差分隐私的K-modes聚类数据隐私保护框架. 引入本地差分隐私技术在用户端对数据进行扰动,并通过服务端与用户端的交互迭代,实现了对聚类过程中用户数据的去第三方隐私保护.

(2) 为提高聚类结果的质量并降低通讯开销,在使用本地差分隐私技术扰动前,对数据进行随机采样,避免因隐私预算分割导致的聚类质量降低以及发送全部数据带来的通讯开销较高问题.

(3) 理论分析证明了方法的隐私性和可用性,真实数据集上的实验结果表明,该方法在满足本地差分隐私机制的前提下,有效保证了聚类结果质量.

2 本地差分隐私与K-modes聚类

2.1 本地差分隐私

针对第三方存在泄露用户隐私的风险,本地差分隐私直接在用户端对数据进行扰动,它能满足用户对个人隐私保护的更高要求. 本地差分隐私的形式化定义如下:

定义1 假设n名用户都至少拥有一条记录,其隐私保护算法为M、定义域为Dom、值域为Rnm. 如果任意两条记录$t(t\in \mathrm{Dom})$和$t'(t'\in \mathrm{Dom})$,经M处理后得到相同输出结果$t^*(t^*\in \mathrm{Rnm})$的概率满足式(1),则M满足$\varepsilon-\mathrm{LDP}$.

$$\Pr\left[(t)=t^*\right]\le \mathrm{e}^{\varepsilon}\times\Pr\left[(t')=t^*\right] \tag{1}$$

定义1中的ε是隐私预算,其值大于0. 它表示用户数据的隐私保护强度,ε越小,隐私保护程度越高,但相应的数据可用性就会降低,因此在具体应用中要从多个角度权衡ε的取值. 同时从定义1可看出,本地差分隐私通过控制任意两条记录输出结果的相似性来保护数据隐私. 经过本地差分隐私处理后,从输出结果逆推出输入数据是非常困难的.

2.2 K-modes聚类

K-modes算法由K-means扩展而来,主要应用于分类型数据的聚类. 它采用Hamming距离衡量两个点之间的间距[27],并通过计算属性值的众数来确定簇的质心. K-modes算法的具体步骤如下.

步骤1:确定需要划分的簇数k,从数据集中随机选择k个点作为起始质心.

步骤2:分别计算数据集中每个点与每个质心的距离,并将点划分给距离最近的质心.

步骤3:得到k个簇后,通过计算各个属性值的众数,然后确定每个簇的新质心.

步骤4:重复步骤2和步骤3,直到相邻两次的聚类结果不再发生变化.

3　K-modes聚类数据隐私保护方法

3.1　问题描述

假设存在用户集 $U=\{u_1,u_2,\cdots,u_n\}$ 和属性集 $M=\{A_1,A_2,\cdots,A_d\}$. 每个用户 $u_i(1\le i\le n)$ 都拥有一个 d 维属性元组 $m_i=\{a_1,a_2,\cdots,a_d\}$, $a_j(1\le j\le d)$ 是 A_j 的某个属性值. K-modes算法的目标是将用户划分为 k 个簇 $C=\{c_1,c_2,\cdots,c_k\}$. 在分簇过程中通常包含一些用户敏感信息,而采用可信第三方对聚类数据进行隐私保护的方法,却很难找到绝对可信的第三方来防止用户隐私数据泄露. 表1为本文中使用的重要符号.

表1　符号说明

符号	说明
U	用户集
M	属性集
V	质心集
m_i	用户 u_i 的数据
B_i	m_i 编码得到的比特字符串
b_j	从 B_i 中获得的采样数据
b_j'	b_j 扰动后的数据
c_y	用户所属的簇

3.2　具体实现方案

本文提出了一种基于本地差分隐私的K-modes聚类数据隐私保护方法,其整体框架如图1所示. 用户 u_i 向服务端发送数据时,为避免隐私预算分割并降低通讯开销,先对数据进行随机采样,然后采用本地差分隐私算法对采样数据进行扰动,最后将得到的扰动数据发送给服务端.

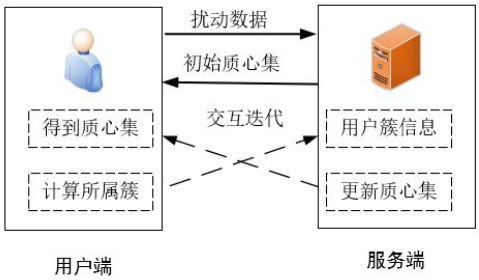


图1　基于LDP的K-modes聚类框架

上述过程无需第三方对数据进行隐私预处理,确保了用户的隐私不受不可信第三方的威胁. 服务端在收集到所有用户的扰动数据后,依据属性集信息确定初始质心集 $V=\{v_1,v_2,\cdots,v_k\}$,并将其发送给所有用户. 用户 u_i 从服务端接收到质心集后,根据自身真实数据 m_i 计算出距离最近的质心 $v_y(1\le y\le k)$,然后依据质心 v_y 确定所属的簇 c_y 并汇报. 服务端收到用户的簇信息后,结合扰动数据求解新的质心集. 最后重复上述的交互过程不断进行迭代,直到各个簇中的质心在相邻两次迭代中不再发生变化.

3.3　用户端数据处理

用户端的每个用户 u_i 都拥有一个真实数据集 $m_i=\{a_1,a_2,\cdots,a_d\}$,LDPK采用目前本地差分隐私技术中高精度的最优一元编码算法[28]对其进行扰动. 该算法在扰动前要对数据进行编码,以 a_1 为例,设它对应的属性 A_1 为"民族",则属性域大小 $|A_1|$ 为56,用一个长度为56的比特字符串 $b=\{0,0,\cdots,0\}$ 表示该属性域. 每个民族对应 b 中一个比特位,设 a_1 为汉族,而汉族对应第 x 个比特位,故将第 x 个比特位设为1,得到比特字符串 $b=\{0,\cdots,1,0\}$. 对 m_i 中每个属性值 a_j 都按照上述过程编码,以 x 表示属性值对应的比特位,以 $w(1\le w\le|b_j|)$ 表示 x 的取值范围,编码如式(2)所示:

$$b_j[w]=\begin{cases}1, & w=x\\ 0, & w\ne x\end{cases}\tag{2}$$

编码完成后得到比特字符串 $B_i=\{b_1,b_2,\cdots,b_j,\cdots,b_d\}$,然后对 B_i 进行随机采样,并对采样值 b_j 采用式(3)进行扰动.

$$\Pr[b_j'[w]=1]=\begin{cases}p, & \text{if } b_j[w]=1\\ q, & \text{if } b_j[w]=0\end{cases}\tag{3}$$

其中 $p=1/2$、$q=\dfrac{1}{e^{\varepsilon}+1}$,扰动完成后得到 b_j'. 数据扰动具体过程如算法1所示. 最后,用户将得到的扰动数据 (j,b_i') 发送给服务端. 因数据在用户端已经添加了噪声,故服务端无法从收集到的数据中获取用户的真实信息.

算法1　数据扰动

输入: 用户 u_i 的数据 $m_i=\{a_1,a_2,\cdots,a_d\}$, 隐私预算 ε;

输出: 扰动后的值 (j,b_j');

1. 将 $m_i=\{a_1,a_2,\cdots,a_d\}$ 编码为 $B_i=\{b_1,b_2,\cdots,b_d\}$;
2. 从 B_i 中随机采样 b_j, 同时设置 $b_j'=\{0,0,\cdots,0\}$,$|b_j'|=|b_j|$;
3. FOR $w=0$ to $|b_j|-1$ do
4. 　IF $b_j[w]=1$
5. 　　$\Pr[b_j'[w]=1]=1/2$
6. 　IF $b_j[w]=0$
7. 　　$\Pr[b_j'[w]=1]=\dfrac{1}{e^{\varepsilon}+1}$
8. Return (j,b_j')

在之后的迭代过程中，服务端先计算出质心集合$V=\{v_1,v_2,\cdots,v_k\}$，并发送给所有用户．然后每个用户u_i根据自身的真实数据m_i计算出距离最近的质心v_y．最后再依据质心v_y确定所属的簇c_y，并将其汇报给服务端．

3.4 服务端求解质心

服务端从所有用户收集到扰动数据$B'=\{b_1',b_2',\cdots,b_d'\}$后，首先从属性集中随机选取$k$个$d$维属性元组作为初始质心发送给用户，然后依据扰动数据和用户返回的簇信息计算出新的质心集$V=\{v_1,v_2,\cdots,v_k\}$．其具体步骤如下：

（1）根据每个用户汇报的c_y，将所有用户划分为k个簇$C=\{c_1,c_2,\cdots,c_k\}$．$|c_r|(1\le r\le k)$表示簇中的用户人数，因用户对数据进行了采样，故簇中每个属性对应的用户人数变为$|c_r|/d$．为便于计算，后续以$|c_r|'$表示$|c_r|/d$．

（2）计算每个簇中各个属性值的频率．以簇c_r中的属性A_j为例，统计对应扰动数据b_j'的每个比特位得到$S=\{s_1,s_2,\cdots,s_l\}$，$s_l\left(0\le l<|b_j'|\right)$表示$b_j'$中第$l$个比特位为1的个数，再结合其他数据计算出$A_j$每个属性值的估计频率$t'$．

（3）服务端在计算出各个簇中所有属性值的频率后，选取每个属性中频率最高的属性值，以它们的集合作为该簇的质心，最终得到新的质心集$V=\{v_1,v_2,\cdots,v_k\}$．服务端质心求解的具体过程如算法2所示．

算法2 计算迭代质心

输入：用户簇$C=\{c_1,c_2,\cdots,c_k\}$，每个簇的数据$B'=\{b_1',b_2',\cdots,b_d'\}$，$p$，$q$；

输出：新的质心集合$V=\{v_1,v_2,\cdots,v_k\}$；

```
FOR C from r=1 to k do
    From b_1' to b_d' do
        统计每个比特位得到S={s_1,s_2,…,s_l}
        FOR i=0 to l-1 do
            t_i'=(s_i-|c_r|'×q)/(|c_r|'×(p-q))
        Return a=Max(t')对应的属性值
    Return v_r={a_1,a_2,…,a_d}
Return V={v_1,v_2,…,v_k}
```

服务端最后向用户发送新的质心集，并根据从用户收集到的簇信息不断重复上述过程，直到各个簇中的质心在相邻两次迭代中不再发生变化．

3.5 隐私性和可用性分析

本文提出的LDPK方法中只有算法1需要隐私预算，因此只要算法1满足本地差分隐私的定义，则LDPK方法同样满足该定义．

引理1 算法1满足本地差分隐私的定义．

证明 设存在属性A_j的两个属性值x_1和x_2，由它们的扰动结果b_j'可得：

$$\frac{\Pr\left[b_j'|x_1\right]}{\Pr\left[b_j'|x_2\right]}=\frac{\prod_{i\in|b_j'|}\Pr\left[b_j'[i]|x_1\right]}{\prod_{i\in|b_j'|}\Pr\left[b_j'[i]|x_2\right]} \tag{4}$$

因b_j'中每个比特位都是独立扰动，故式(4)只在x_1和x_2处不同，可以得出：

$$\frac{\Pr\left[b_j'|x_1\right]}{\Pr\left[b_j'|x_2\right]}=\frac{\Pr\left[b_j'[x_1]|x_1\right]\Pr\left[b_j'[x_2]|x_1\right]}{\Pr\left[b_j'[x_1]|x_2\right]\Pr\left[b_j'[x_2]|x_2\right]} \tag{5}$$

对于式(5)，当b_j'中的x_1位置为1，x_2位置为0时，它右侧的比值达到最大：

$$\begin{aligned}\frac{\Pr\left[b_j'|x_1\right]}{\Pr\left[b_j'|x_2\right]}&\le\frac{\Pr\left[b_j'[x_1]=1|x_1\right]\Pr\left[b_j'[x_2]=0|x_1\right]}{\Pr\left[b_j'[x_1]=1|x_2\right]\Pr\left[b_j'[x_2]=0|x_2\right]}\\&=\frac{p}{q}\times\frac{1-q}{1-p}\end{aligned} \tag{6}$$

又因$p=1/2$，$q=1/\left(\mathrm{e}^{\varepsilon}+1\right)$，故将它们代入式(6)的右侧后可得：

$$\frac{p}{q}\times\frac{1-q}{1-p}=\mathrm{e}^{\varepsilon} \tag{7}$$

因此，算法1满足本地差分隐私的定义．

服务端在更新质心时，由于没有收集用户的真实数据，它不能计算出每个属性值的真实频率t，只能得出估计频率t'．为了降低噪声的影响，我们希望计算出的t'满足无偏性．因此，需要通过如算法2的步骤5所示来计算t'．

引理2 通过算法2的步骤5计算出的估计频率t'满足无偏性．

证明 假设t与t'分别为簇c_r中某属性值a的真实频率与估计频率，g与g'分别为a的真实频数和估计频数．设a'为a对应的比特位，s是扰动数据中a'为1的个数．

由于服务端无法获得a的真实频数g，为了求解t'，需要计算估计频数g'．因用户以两种概率对每个比特位进行响应，故$|c_r|'$个用户对a'的响应结果构成了满足二项分布的$|c_r|'$个0/1序列．根据该二项分布，构造相应的似然函数：

$$\begin{aligned}L(g)=&\left[\frac{g}{|c_r|'}\times p+\left(1-\frac{g}{|c_r|'}\right)\times q\right]^{s}\\&\times\left[\frac{g}{|c_r|'}\times(1-p)+\left(1-\frac{g}{|c_r|'}\right)\times(1-q)\right]^{|c_r|'-s}\end{aligned} \tag{8}$$

对式(8)两侧取对数并对g求导即可求出它的极大似然估计g'：

$$g'=\frac{s-|c_r|'\times q}{p-q} \tag{9}$$

对于求解出的g'可以证明其满足无偏性:

$$E\left[g'\right]=E\left[\frac{s-|c_r|'\times q}{p-q}\right]=\frac{\left(g\times p+\left(|c_r|'-g\right)\times q\right)-|c_r|'\times q}{p-q}=g \tag{10}$$

因g'满足无偏性,故可求出无偏估计频率t':

$$t'=\frac{s-|c_r|'\times q}{|c_r|'\times\left(p-q\right)} \tag{11}$$

对用户数据直接扰动存在的聚类质量降低和通讯开销较高问题,LDPK方法通过对数据采样,使用户在发送扰动数据时只需发送采样值,大大降低了通讯开销. 但为了降低扰动数据中噪声的影响,需要较大的数据量,而采样使任意属性值对应的用户总数变为实际值的1/d,因此需要对采用这两种方式得到的扰动数据的可用性进行分析.

引理3　对用户数据采样相比于分割隐私预算可以提高扰动数据的可用性.

证明　设有n名用户的数据是d维属性元组,隐私预算为ε,其某个属性值的估计频数为g',f为它对应的比特位,s是扰动数据中f为1的个数,真实频率为t. 若不采样且该属性值获得全部ε,则由式(9)可得g'的方差为:

$$\mathrm{Var}\left[g'\right]=\mathrm{Var}\left[\frac{s-n\times q}{p-q}\right]=\frac{n\times t\times p\times\left(1-p\right)+n\times\left(1-t\right)\times q\times\left(1-q\right)}{\left(p-q\right)^2} \tag{12}$$

将$p=1/2$,$q=1/\left(\mathrm{e}^{\varepsilon}+1\right)$代入式(12)可以得到如下的方差表达式:

$$\mathrm{Var}\left[g'\right]=\frac{n\times 4\mathrm{e}^{\varepsilon}}{\left(\mathrm{e}^{\varepsilon}-1\right)^2}+n\times t \tag{13}$$

式(13)中$n\times t$为属性值的真实频数,它是一个常数,为了便于计算将其省略. 同时又因两种方法中属性值对应的用户数目不同,导致无法直接比较它们的方差,故对式(13)进行如下转换:

$$\mathrm{Var}\left[g'/n\right]=\frac{4\mathrm{e}^{\varepsilon}}{n\times\left(\mathrm{e}^{\varepsilon}-1\right)^2} \tag{14}$$

对d维属性随机采样使得属性值对应的用户数变为n/d,其方差以η_1表示:

$$\eta_1=\frac{4d\times\mathrm{e}^{\varepsilon}}{n\times(\mathrm{e}^{\varepsilon}-1)^2} \tag{15}$$

对隐私预算分割使得每个属性获得的隐私预算变为ε/d,其方差以η_2表示:

$$\eta_2=\frac{4\mathrm{e}^{\varepsilon/d}}{n\times(\mathrm{e}^{\varepsilon/d}-1)^2} \tag{16}$$

如果采用随机采样得到的扰动数据可用性优于隐私预算分割,那么η_2和η_1应当满足$\eta_1<\eta_2$,而它们之间的大小关系如下:

$$\begin{aligned}\eta_1-\eta_2&=\frac{4}{n}\times\left(\frac{\mathrm{e}^{\varepsilon/d}}{\left(\mathrm{e}^{\varepsilon/d}-1\right)^2}-\frac{d\times\mathrm{e}^{\varepsilon}}{\left(\mathrm{e}^{\varepsilon}-1\right)^2}\right)\\&=\frac{4\mathrm{e}^{\varepsilon/d}}{n\times\left(\mathrm{e}^{\varepsilon/d}-1\right)^2\times\left(\mathrm{e}^{\varepsilon}-1\right)^2}\\&\quad\times\left(\left(\mathrm{e}^{\varepsilon}-1\right)^2-d\times\mathrm{e}^{\varepsilon-\varepsilon/d}\times\left(\mathrm{e}^{\varepsilon/d}-1\right)^2\right)\end{aligned} \tag{17}$$

由上文可知隐私预算大于0. 因此,对于式(17)结果的左侧部分可以得出:

$$\frac{4\mathrm{e}^{\varepsilon/d}}{n\times\left(\mathrm{e}^{\varepsilon/d}-1\right)^2\times\left(\mathrm{e}^{\varepsilon}-1\right)^2}>0 \tag{18}$$

定义$y=\mathrm{e}^{\varepsilon/d}$,将它代入式(17)的结果部分. 同时又因式(18)大于0,故将式(17)结果的左侧部分舍去以简化运算,则式(17)可化简为:

$$\begin{aligned}\eta_1-\eta_2&=\left(y^d-1\right)^2-d\times y^{d-1}\times\left(y-1\right)^2\\&=\left(y-1\right)^2\times\left[\left(y^{d-1}+y^{d-2}+\cdots+1\right)^2-d\times y^{d-1}\right]\\&>0\end{aligned} \tag{19}$$

由式(19)可知$\eta_1<\eta_2$,因此对用户数据随机采样相比于分割隐私预算,可以提高扰动数据可用性.

4　实验结果分析

实验主要采用两个真实数据集,一个数据集来自IPUMS网站的公开数据,从中选取了USA的5万条普查数据,如表2所示,每条记录包含5个分类属性. 另一个是隐私保护研究领域常用的UCI数据库中Adult数据集,经过删除其中的无效记录后共有30162条记录,如表3所示,每条记录分别选取6个分类属性.

表2　USA普查数据集属性

属性	属性域大小
FAMSIZE	15
SEX	2
RACE	9
SCHOOL	3
EMPSTAT	4

实验的硬件环境为:Intel(R) Core(TM) i5-7300HQ CPU @2.50 GHz 2.50 GHz, 8.00 GB内存. 软件环境为:Microsoft Windows 10. 采用PyCharm开发平台,以Python编程语言实现.

表3 Adult数据集属性

属性	属性域大小
WORKCLASS	7
EDUCATION	16
MARITAL	7
RELATIONSHIP	6
RACE	5
SEX	2

实验采用准确率AC(Accuracy)和熵E(Entropy)作为聚类质量评价指标,以无隐私保护下的K-modes聚类结果作为真实值. 评价指标如式(20)和式(21)所示:

$$\mathrm{AC}=\sum_{j=1}^{k}h_j\Big/N \tag{20}$$

$$E=\sum_{j=1}^{k}\frac{c_j}{N}\sum_{i=1}^{k}-\frac{|c_j\cap t_i|}{|c_j|}\times\ln\left(\frac{|c_j\cap t_i|}{|c_j|}\right) \tag{21}$$

其中k是聚类簇数,h_j是采用隐私保护算法得到的聚类簇中正确聚类的数据个数,N是数据集的大小,t_i是无隐私保护下得到的聚类簇,c_j是采用隐私保护算法得到的聚类簇. 实验研究相关参数对LDPK算法性能的影响并与现有差分隐私保护下的K-modes算法[17](Different Privacy K-modes,DPK)进行对比. 聚类簇数k取3,同时为了降低初始质心选择和扰动数据中噪声对结果的影响,每个实验进行300次,结果取平均值.

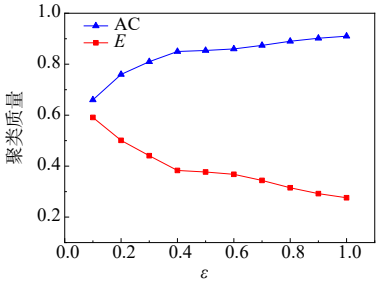
4.1 参数变化对LDPK性能的影响

改变隐私预算ε,其他参数保持不变,分析隐私预算大小对LDPK性能的影响. 由图2可知随着ε的增大,聚类结果的准确率随之提升,熵随之降低,其原因在于扰动数据中的噪声添加量取决于ε的值. ε越小,扰动数据中添加的噪声越多,则聚类结果质量越低. ε越大,扰动数据中添加的噪声越少,则聚类结果质量越高.

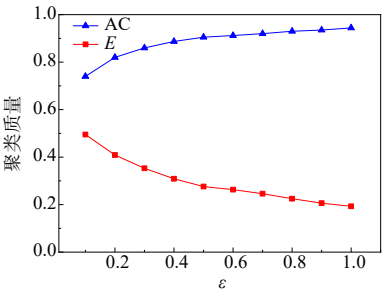
将隐私预算ε设置为1.0,其他参数保持不变. 从数据集中随机抽取不同数目的记录,分析数据集大小N对LDPK性能的影响. 由表4和表5可知随着数据集的增大,聚类结果的准确率随之增加,熵随之降低,其原因在于本地差分隐私中每个用户以一定的概率汇报真实值,而根据大数定理,同等条件下实验重复次数越多,随机事件的结果越接近其真实频率. 因此数据量越大,响应随机性的影响越小,聚类结果的质量也就越高.

4.2 算法性能对比

改变隐私预算ε,其他参数保持不变,对比LDPK与DPK的聚类结果质量. 由图3可知,DPK的聚类结果质量略优于LDPK,其原因在于DPK依靠第三方对真实数据进行隐私处理,可以更好的控制噪声添加,所以聚类



(a) Adult



(b) USA

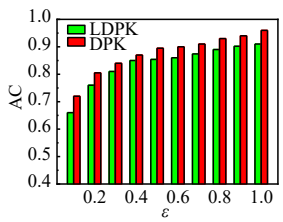
图2 隐私预算ε对算法性能的影响

表4 Adult大小对AC和E的影响

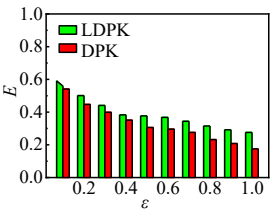
N(万)	1.0	1.5	2.0	2.5	3.0
AC	0.78	0.82	0.84	0.89	0.92
E	0.48	0.44	0.40	0.31	0.26

表5 USA大小对AC和E影响

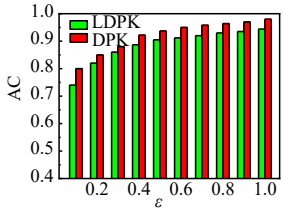
N(万)	1.0	1.5	2.0	2.5	3.0
AC	0.73	0.76	0.81	0.84	0.86
E	0.50	0.47	0.42	0.38	0.35



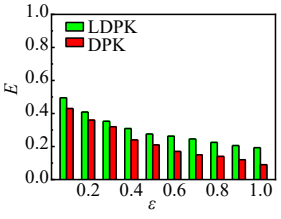
(a) Adult: AC



(b) Adult: E



(c) USA: AC



(d) USA: E

图3 性能对比

结果的质量较高. 但LDPK的聚类结果质量与DPK相比差距不大,这表明通过对用户数据的随机采样以及服务端与用户端的交互迭代,LDPK有效保证了聚类结果的质量. 同时与DPK相比,LDPK不需要任何第三方对真实数据进行隐私预处理,避免了第三方泄露用户隐私的风险,提高了用户数据隐私的保护程度.

5 结语

针对K-modes聚类数据中用户敏感信息的隐私保护问题,当前主要依靠基于可信第三方的隐私保护方法,但实际应用中该第三方也存在隐私泄露风险. 本文提出了一种本地差分隐私下的K-modes聚类数据隐私保护方法. 该方法结合本地差分隐私和随机采样技术在用户端对数据进行扰动,使得整个聚类过程中服务端都无法获得用户的真实信息,同时它基于去第三方思想,避免了第三方泄露用户隐私的风险. 在真实数据集上的实验结果表明,该方法在满足本地差分隐私机制的前提下,有效保证了聚类结果的质量.

本文提出的方法使用户所有数据受到同等程度的隐私保护,但有些数据不需要很强的隐私保护度. 因此在未来工作中,我们将尝试对用户数据敏感度分级,允许服务端直接收集敏感度较低的数据,而对敏感度较高的数据采用本地差分隐私进行保护,以满足用户对不同敏感度数据的个性化隐私保护需求.

参考文献

[1] COELHO A L V, SANDES N C. Data clustering via cooperative games: A novel approach and comparative study[J]. Information Sciences, 2021, 545: 791-812.

[2] SAROJ K. Review: study on simple k mean and modified K mean clustering technique[J]. International Journal of Computer Science Engineering and Technology, 2016, 6 (7): 279-281.

[3] XIAO Y Y, HUANG C H, HUANG J Y, et al. Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering[J]. Pattern Recognition, 2019, 90: 183-195.

[4] ZHANG S B, MAO X J, CHOO K K R, et al. A trajectory privacy-preserving scheme based on a dual-K mechanism for continuous location-based services[J]. Information Sciences, 2020, 527: 406-419.

[5] ZHANG S B, WANG G J, BHUIYAN M Z A, et al. A dual privacy preserving scheme in continuous location-based services[J]. IEEE Internet of Things Journal, 2018, 5(5): 4191-4200.

[6] CHAVES A, MOURA I, BERNARDINO J, et al. The privacy paradigm: An overview of privacy in Business Analytics and Big Data[C]//2020 15th Iberian Conference on Information Systems and Technologies(CISTI). Piscataway: IEEE, 2020: 1-6.

[7] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2013, 9(3/4): 211-407.

[8] DEWRI R, THURIMELLA R. Exploiting service similarity for privacy in location-based search queries[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25 (2): 374-383.

[9] MEDKOVÁ J. Composition attack against social network data[J]. Computers & Security, 2018, 74: 115-129.

[10] 彭慧丽, 金凯忠, 付聪聪, 等. 基于序列格的隐私时序模式挖掘方法[J]. 电子学报, 2020, 48(1): 153-163.
PENG H L, JIN K Z, FU C C, et al. Private time series pattern mining with sequential lattice[J]. Acta Electronica Sinica, 2020, 48(1): 153-163. (in Chinese)

[11] 陈思, 付安民, 柯海峰, 等. MCDP: 基于神经网络的多集群分布式差分隐私数据发布方法[J]. 电子学报, 2020, 48(12): 2297-2303.
CHEN S, FU A M, KE H F, et al. MCDP: multi-cluster differential privacy data publishing method based on neural network[J]. Acta Electronica Sinica, 2020, 48(12): 2297-2303. (in Chinese)

[12] 郑孝遥, 罗永龙, 汪祥舜, 等. 基于位置服务的分布式差分隐私推荐方法研究[J]. 电子学报, 2021, 49(1): 99-110.
ZHENG X Y, LUO Y L, WANG X S, et al. Research on location-based distributed differential privacy recommendation method[J]. Acta Electronica Sinica, 2021, 49(1): 99-110. (in Chinese)

[13] XIAO X K, TAO Y F, CHEN M H. Optimal random perturbation at multiple privacy levels[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 814-825.

[14] KIFER D. On estimating the swapping rate for categorical data[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2015: 557-566.

[15] SU D, CAO J N, LI N H, et al. Differentially private K-means clustering and a hybrid approach to private optimization[J]. ACM Transactions on Privacy and Security, 2017, 20(4): 1-33.

[16] NGUYEN T D, GUPTA S, RANA S T, et al. Privacy Aware K-Means Clustering with High Utility[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2016: 388-400.

[17] NGUYEN H H. Privacy-preserving mechanisms for k-modes clustering[J]. Computers & Security, 2018, 78: 60-75.

[18] 叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述[J]. 软件学报, 2018, 29(7): 1981-2005.
YE Q Q, MENG X F, ZHU M J, et al. Survey on local differential privacy[J]. Journal of Software, 2018, 29(7): 1981-2005. (in Chinese)

[19] ERLINGSSON Ú, PIHUR V, KOROLOVA A. RAPPOR: randomized aggregatable privacy-preserving ordinal response[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM, 2014: 1054-1067.

[20] KAIROUZ P, BONAWITZ K, RAMAGE D. Discrete distribution estimation under local privacy[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, USA: ACM, 2016: 2436-2444.

[21] QIN Z, YANG Y, YU T, et al. Heavy hitter estimation over set-valued data with local differential privacy[C]// CCS' 16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM, 2016: 192-203.

[22] WANG T H, LI N H, JHA S. Locally differentially private frequent itemset mining[C]//2018 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2018: 127-143.

[23] 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. Piscataway: IEEE, 2013: 429-438.

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

[25] KULKARNI T. Answering range queries under local differential privacy[C]//Proceedings of the 2019 International Conference on Management of Data. New York, USA: ACM, 2019: 1832-1834.

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

[27] AMIR A, AMIT M, LANDAU G M, et al. Period recovery of strings over the Hamming and edit distances[J]. Theoretical Computer Science, 2018, 710: 2-18.

[28] WANG T H, BLOCKI J, LI N H, et al. Locally differentially private protocols for frequency estimation[C]//Proceedings of the 26th USENIX Conference on Security Symposium. Berkeley, CA, USA: USENIX Association, 2017: 729-745.

作者简介

张少波 男, 博士, 1979 年生于湖南邵东. 现为湖南科技大学副教授、硕士生导师. 主要研究方向为移动社交网络、大数据、人工智能、区块链的安全和隐私保护等.
E-mail: shaobozhang@hnust.edu.cn

原刘杰 男, 1995 年生于河南开封. 现为湖南科技大学硕士研究生. 主要研究方向为大数据隐私保护.
E-mail: ljyuan@mail.hnust.edu.cn

毛新军 男, 1970 年生于浙江江山. 博士, CCF 杰出会员, 现为国防科学技术大学计算机学院教授, 博士生导师. 主要研究领域为智能软件技术, 多智能体系统等.

朱更明 男, 1963 年生于湖南邵阳. 现为湖南科技大学教授、硕士生导师. 主要研究方向为信息安全、智能设备机器视觉及控制等.
E-mail: zhu.gm@163.com