计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700172

支持本地化差分隐私保护的 k-modes 聚类方法

彭春春　陈燕俐　荀艳梅
南京邮电大学计算机学院、软件学院、网络空间安全学院　南京 210003
(1219043729@njupt.edu.cn)

摘　要　如何在保护数据隐私的同时进行可用性的数据挖掘已成为热点问题。鉴于在很多实际应用场景中,很难找到一个真正可信的第三方对用户的敏感数据进行处理,文中首次提出了一种支持本地化差分隐私技术的聚类方案——LDPK-modes(Local Differential Privacy K-modes)。与传统的基于中心化差分隐私的聚类算法相比,其不再需要一个可信的第三方对数据进行收集和处理,而由用户担任数据隐私化的工作,极大地降低了第三方窃取用户隐私的可能性。用户使用满足本地 d-隐私(带有距离度量的本地差分隐私技术)定义的随机响应机制对敏感数据进行扰动,第三方收集到用户扰动数据后,恢复其统计特征,生成合成数据集,并进行 k-modes 聚类。在聚类过程中,将数据集上频繁出现的特征分配给初始聚类中心点,进一步提高了聚类结果的可用性。理论分析和实验结果表明了 LDPK-modes 的隐私性和聚类可用性。

关键词:本地化差分隐私;k-modes;d-隐私;聚类;隐私保护
中图法分类号　TP309

k-modes Clustering Guaranteeing Local Differential Privacy

PENG Chun-chun,CHEN Yan-li and XUN Yan-mei
College of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210003,China

Abstract　How to conduct usability data mining while protecting data privacy has become a hot issue. In many practical scenarios,it is difficult to find a trusted third party to process the sensitive data. This paper proposes the first locally differentially private k-modes mechanism(LDPK-modes) under this distributed scenario. Differing from standard differentially private clustering mechanisms,the proposed mechanism doesn't need any trusted third party to collect and preprocess users data. Users disturb their data using a random response mechanism that satisfies the definition of local d-privacy (local differential privacy with distance metric). When the third party collects the user's disturbed data,it restores its statistical features and generates a synthetic data set. The frequent attributes on the data set are assigned to the initial cluster center and then start k-modes clustering. Theoretical analysis shows that the proposed algorithm satisfies local d-privacy. Experimental results show that our proposal can well preserve the quality of clustering results without a trusted third-party data collector.

Keywords　Local differential privacy,k-modes,d-privacy,Clustering,Privacy preserving

1　引言

大数据时代,信息技术为人类社会带来便捷的同时,也产生了数据安全与用户隐私的问题。如何在保护用户数据隐私的同时,满足服务提供方合理的数据使用需求成为了大数据时代的一个巨大挑战,受到了学术界的广泛关注。为了平衡隐私保护程度和数据可用性,需要引入形式化定义对隐私保护进行量化。2006 年,微软研究院的科学家 Dwork 提出的差分隐私保护(Centralized Differential Privacy,DP)技术[1],作为一种可严格证明的隐私保护模型,严格定义了隐私保护的强度,任意一条记录的添加或删除都不会影响最终的查询结果。同时,该模型定义了极为严格的攻击模型,不关心攻击者具有的背景知识,即使攻击者掌握除源数据外的任何辅助信息,差分隐私仍能使源数据的隐私泄漏风险控制在一个可接受的范围。相比已有的隐私保护模型,如 k-匿名模型[2]、l-多样性模型[3]等需要特殊攻击假设和背景知识的方法,差分隐私因其独特的优势,成为了当前学术界的研究热点。

传统的差分隐私技术(中心化差分隐私)需要先将源数据集中到一个数据中心,再进行隐私保护,因此始终基于一个前提假设:第三方数据收集者是可信的,即保证第三方数据收集者不会窃取或泄露用户的敏感信息。然而,在实际应用[4]中,即使第三方数据收集者宣称不会窃取和泄露用户的敏感信

到稿日期:2020-07-25　返修日期:2020-09-01　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:国家自然科学基金(61572263,61272084)
This work was supported by the National Natural Science Foundation of China(61572263,61272084).
通信作者:陈燕俐(chenyl@njupt.edu.cn)

息，但是用户的隐私依旧得不到保障。2013 年，专家提出的本地化差分隐私(Local Differential Privacy，LDP)技术[5-6]，将数据的隐私化处理过程从第三方数据收集者转移到每个用户上，可以在不需要信任第三方数据收集者的情况下，在本地环境下对个人敏感数据加噪以实现隐私保护，同时从宏观角度保证数据收集者可正确地推断出群体统计信息。这既保证了群体统计信息的相对准确性，又保护了个人的原始数据，解决了用户隐私数据被不可信第三方外包管理的问题，因此 LDP 适用场景更广，实用程度更高。

数据挖掘[7]是一门从海量数据中获取目标信息的技术，但是这些数据可能包含大量的用户隐私信息，如行为轨迹数据、身份数据、医疗数据等。一旦不法分子获取到这些数据，用户的人身财产安全可能受到损害。因此，对数据进行隐私保护的同时进行可用性的数据挖掘日益成为重要的研究领域[8]。聚类分析是数据挖掘的主要任务之一，可以将无标签的数据划分为若干类别或簇，进而获得数据的分布状况，在聚类分析中引入差分隐私保护技术也逐渐成为研究的热点。k-means 算法由于其高效性和实用性成为了广泛应用于数据挖掘的聚类方法，然而传统的 k-means 算法只适用于连续属性的数据集(数值型数据)。k-modes 算法作为 k-means 的一种变种，使用汉明距离(Hamming distance)[9]计算点之间的距离并用众数代替均值计算簇的质心，适用于离散属性的数据集(分类型数据)。但目前的差分隐私保护 k-modes 算法均采用中心化技术，受到基于可信的第三方数据收集者的假设限制。因此，本文首次提出了一个支持本地化差分隐私技术的隐私保护 k-modes 聚类方案——LDPK-modes。LDPK-modes 使用随机响应机制扰动用户数据，并在不暴露某个具体用户隐私的前提下，使用合成数据集进行聚类，在保证隐私性的同时保证了聚类可用性。

本文的贡献如下：

(1)设计了一种满足本地差分隐私的频数预测算法——LDP-FO，并证明了该算法满足本地 d-隐私[10]的定义。为保证聚类过程中用户数据的隐私性，算法通过随机响应机制扰动用户数据，并由服务器聚合数据尽可能地恢复其原始信息。

(2)将 LDP-FO 算法与 k-modes 聚类算法相结合，提出一种支持本地差分隐私的聚类方案——LDPK-modes，与中心化差分隐私聚类算法相比，其不需要信任第三方数据收集者，能更好地保障用户的隐私，并且在多次迭代聚类中无须划分隐私预算，减少了噪声的引入，从而提高了聚类可用性。

本文第 2 节介绍了相关工作；第 3 节描述了相关定义与理论基础；第 4 节提出了一种本地差分隐私下的频数预测算法，然后将其应用于 k-modes 算法并同时优化了初始聚类中心点的选取；第 5 节进行了仿真实验，分析与评估了本文提出的新算法；最后总结全文并展望未来。

2 相关工作

差分隐私作为新兴的隐私保护技术，结合其应用场景及针对数据处理和收集方式的不同，主要存在中心化和本地化两种模型。目前，大部分研究成果集中在基于可信的第三方数据收集者的中心化差分隐私保护模型，存在着较大的隐私安全泄露风险。本地化模型中，用户先在本地加入满足差分隐私的噪声扰动，然后数据收集者根据收集到的噪声数据，从统计学的角度近似估计出用户群体的统计特征。

本地化差分隐私技术的研究方向包括扰动机制的研究以及统计数据的发布。基于随机响应技术[11]的扰动机制是当前的研究热点，其相关工作主要集中在满足本地化差分隐私保护的算法研究上。统计数据的发布，根据算法的查询类型进行分类，主要包括离散型数据的频数统计查询发布和连续型数据的均值统计查询发布。Erlingsson 等[12]提出了一种众包环境下的本地化差分隐私数据收集技术。在此基础上，他们进一步提出了一种新型的频数发布方法 RAPPOR，提高了数据收集的精度。RAPPOR 采用随机应答策略和 BloomFilter[13]保证在研究用户群体数据时不能窥探到个体的信息，实现了针对客户端群体的类别、频数、直方图和字符串类型统计数据的隐私保护分析。为了解决 RAPPOR 通信代价高这一缺点，Bassily 等[14]提出了一种简洁直方图(SH)机制，该机制下每个用户仅报告 SH 中随机选择的一位数据，但此方法仅限于简单的数字或类别属性，不适用于更复杂的数据挖掘任务。Duchi 等[15-16]提出 MeanEst 方法用于数据均值发布，其主要思想是对数据添加正向和负向的噪声并统计大量数据抵消噪声，使统计结果具有一定实用性。Nguyên 等[17]通过采样技术降低了 MeanEst 中数据的传输代价，同时保证了相似的发布精度。

针对应用于聚类分析中的差分隐私保护，Blum 等[18]提出了差分隐私 k-means 聚类算法，在获取聚类结果的同时保护数据隐私，但其聚类性能受噪声和数据分布的影响。随后不断有改进的差分隐私保护的 k-means 算法[19-21]被提出。在本地环境中，Xia 等[22]提出了本地化差分隐私保护的 k-means 聚类算法，在不需要可信的第三方数据收集者的前提下能够达到和中心化差分隐私 k-means 算法相似的聚类性能表现。2018 年，Nguyen 等[23]使用差分隐私技术解决了 k-modes 聚类中出现的隐私泄露问题，首先分析了相对于 k-means 算法，实现 k-modes 算法中心化差分隐私保护所要面临的问题，接着分别提出了交互式和非交互式环境下的差分隐私方案。2019 年，Lv 等[24]实现了混合数据上的中心化差分隐私保护聚类算法，将 k-means 和 k-modes 算法相结合以处理更为复杂的数据类型的聚类分析。

3 定义与理论基础

3.1 本地化差分隐私

在本地差分隐私中，每个用户都使用机制 M 扰动输入 x(原始数据)，并将 $M(x)$ 上传到服务器进行数据统计特征分析。

定义 1[5-6](本地差分隐私(LDP)) 对于给定的 $\varepsilon\in\mathbb{R}^+$，一个随机扰动机制 $M(x)$ 满足 ε-LDP，当且仅当任意输入对 x,x' 和输出 $y\in Range(M)$ 满足：

$$\frac{\Pr(M(x)=y)}{\Pr(M(x')=y)}\leqslant e^{\varepsilon} \tag{1}$$

其中，$Range(M)$是机制 M 的所有可能输出的集合。ε(隐私预算)值越小，对隐私的保护程度越强，相应的数据实用性就会越小。

定义 2[10](本地 d-隐私(Local d-Privacy)) 对于给定的 $\varepsilon\in\mathbb{R}^+$，一个随机扰动机制 $M(x)$满足 d,ε-LDP，当且仅当任意输入对 x,x'和输出 $y\in Range(M)$满足：

$$\frac{\Pr(M(x)=y)}{\Pr(M(x')=y)}\leqslant e^{\varepsilon\cdot d(x\cdot x')} \tag{2}$$

其中，$d(\cdot)$是距离度量，对于任意输入 x,x',x''满足 3 个属性，即 $d(x\cdot x)=0$，$d(x\cdot x')=d(x'\cdot x)$，以及三角不等式：

$$d(x\cdot x')+d(x\cdot x'')\leqslant d(x'\cdot x'') \tag{3}$$

本地 d-隐私是 LDP 的广义版本，它引入了一种距离度量标准，该距离度量标准使用输入对之间的距离(相似程度)缩放隐私，并在 $d(x\cdot x')>1$ 时放宽隐私约束，从而提供更好的数据实用性，尤其是对于多个特征的扰动。当数据集中的任意两项数据的距离都为 1 时，本地 d-隐私就等于 LDP。

3.2 汉明距离

在信息论中，两个等长字符串之间的汉明距离是两个字符串对应位置的不同字符的个数(将一个字符串变换成另外一个字符串所需要替换的字符个数)。例如："11110000"与"11011000"之间的汉明距离是 2。本文采用汉明距离作为距离度量标准，用户 u_a,u_b的特征向量$\boldsymbol{X}_a,\boldsymbol{X}_b$之间的汉明距离为 $L=d(\boldsymbol{X}_a\cdot\boldsymbol{X}_b)=\sum_{j=1}^{m}x_{a,j}\oplus x_{b,j}$，其中 $0\leqslant L\leqslant m$。

3.3 DP-mode-Lloyd 聚类

Nguyen 等[23]首次提出了 k-modes 算法的中心化差分隐私版本，称为 DP-modes-Lloyd 算法。DP-modes-Lloyd 算法的主要思想如下：

(1)首先输入所有用户特征向量集合 X、聚类个数 K、迭代次数 T、隐私预算 ε，随机选取 K 个特征向量$\boldsymbol{C}^0=\{\boldsymbol{C}_1^0,\boldsymbol{C}_2^0,\cdots,\boldsymbol{C}_K^0\}$作为初始聚类中心点。

(2)将 X 中的特征向量划分到距离它汉明距离最近的中心点处，形成 K 个不相交的子集。对于第 g 个子集的特征向量的第 j 个特征，计算该特征所有取值的计数 $sum=\{n_{gj}(1),n_{gj}(2),\cdots,n_{gj}(k)\}$，$1\leqslant g\leqslant K$，$1\leqslant j\leqslant m$ 。

(3)对 sum 添加满足差分隐私模型的拉普拉斯噪声，得到$sum'=sum+Laplace(e^{-\varepsilon/(m+T)})$，将第 g 个子集的聚类中心点的第 j 个特征更新为sum'中计数最大的特征。

(4)不停重复步骤(2)、步骤(3)直到达到最大迭代次数或者聚类中心不再变化，返回聚类中心$\boldsymbol{C}^T$。

4 LDPK-modes 聚类方案

Nguyen 等提出的 DP-modes-Lloyd 聚类算法是基于中心化的差分隐私保护技术，建立在收集用户信息和更新聚类中心点的第三方数据收集者不会泄露隐私的假设的基础上。另外，由于初始聚类中心点是随机选取的，可能会导致算法陷入局部最优解，运行结果不稳定。针对上述问题，本节首先提出一个本地 d-隐私下的频数预测(Frequency Oracle，FO)算法——LDP-FO，该算法可以在保护用户信息的同时估计拥有某个特征向量的人数，接着将 FO 算法与 k-modes 聚类算法相结合，首次提出了一个满足本地化差分隐私的 k-modes 聚类方案，即 LDPK-modes。

4.1 系统概况

(1)系统模型。本文系统模型包含一个第三方服务器和 n 个用户，用户集 $U=\{u_1,u_2,\cdots,u_n\}$。每个用户拥有 m 个特征$\{v_1,v_2,\cdots,v_m\}$，不失一般性，每个特征$|v_j|=k$，$1\leqslant j\leqslant m$。假设用户u_i拥有特征向量$\boldsymbol{X}_i=\{x_{i,1},x_{i,2},\cdots,x_{i,m}\}$，$x_{i,j}\in v_j$。第三方服务器旨在根据用户上传的特征信息将用户划分为 K 个簇并返回聚类中心集合 $\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_K\}$，其中$\boldsymbol{C}_g$是由 m 个特征组成的特征向量，$1\leqslant g\leqslant K$。如图 1 所示，在将数据上传给第三方服务器之前，每个用户独立地通过随机响应机制对其特征向量进行扰动以保护其隐私，其中用户 u_i 上传的扰动特征向量为 $\boldsymbol{Y}_i=\{y_{i,1},y_{i,2},\cdots,y_{i,m}\}$，$y_{i,j}\in v_j$。在收集完成用户上传的扰动数据之后，服务器则尽量恢复其原始信息，同时分析出用户原始数据的聚类结果。

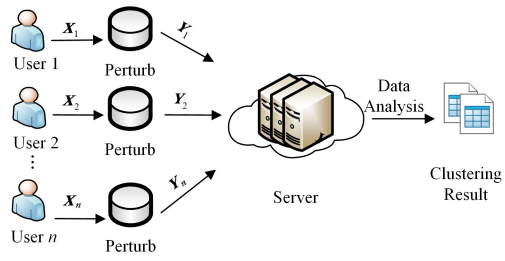


图 1 系统模型

Fig. 1 System model

(2)威胁模型。假设第三方服务器是不受信任的，并且每个用户仅信任自己。未经授权的数据共享或由于黑客活动导致的数据泄露可能造成隐私泄漏。因此，假设对手知道用户的上传数据(扰动数据)，且知道用户采用的扰动机制。本文假设所有用户都是诚实地遵循这种扰动机制，因此不考虑某些用户恶意上传不良数据以欺骗服务器的情况。

4.2 LDP-FO 频数预测算法

LDP-FO 频数预测算法共分为编码(Encode)、扰动(Perturb)、聚合(Aggregate)3 个步骤。用户执行前两个步骤，先将其特征向量编码成数字，然后用随机响应机制对编码后的信息进行扰动，并将扰动后的数据上传给服务器。服务器则执行第三个步骤，使用扰动之后的信息来估计拥有某特征向量的用户个数。

(1)编码。用户首先将其特征信息编码成整数。具体来说，对每个特征v_j的所有可能值进行排序，将用户的该特征置为对应的序号，即$v_j\in\{1,2,\cdots,k\}$。将原始信息转化为整数可减少通信代价并降低算法耦合性。

(2)扰动(Ψ_{perturb})。对于一个用户的特征向量$\boldsymbol{X}_i=\{x_{i,1},x_{i,2},\cdots,x_{i,m}\}$，扰动机制$\Psi_{\text{perturb}}(\boldsymbol{X}_i)$输出$\boldsymbol{Y}_i=\{y_{i,1},y_{i,2},\cdots,y_{i,m}\}$。由汉明距离定义可知，$d(\boldsymbol{X}_i\cdot\boldsymbol{Y}_i)\in\{0,1,\cdots,m\}$，即原始数据经过扰动之后有 $m+1$ 类可能。对于传统的 LDP 扰动机制[25]，当输出不等于输入时，通常输出为任意一个不为原始数据的扰动数据，即输出任意距离的扰动数据概率相等。

而本文采用随机响应扰动机制，即让扰动数据的输出概率与原始数据之间的距离相关，即$\boldsymbol{X}_i$经扰动过后与原始数据之间的汉明距离为z的概率为：

$$\Pr(\boldsymbol{Y}_i^z \mid \boldsymbol{X}_i)=\frac{e^{\varepsilon(m-z)}\binom{m}{z}(k-1)^z}{\sum_{j=0}^{m} e^{\varepsilon j}\binom{m}{j}(k-1)^{m-j}} \tag{4}$$

其中，m和k为用户拥有特征数和每个特征域的大小，ε为隐私预算，$\boldsymbol{Y}_i^z$为与原始数据距离为z的扰动特征向量，$z\in(0,1,2,\cdots,m)$。

从式(4)可以看出，当扰动数据和原始数据的汉明距离$z=0$，即扰动数据等于原始数据时，概率的分子为$e^{\varepsilon m}$；当$z=1$时，扰动机制输出一个与原始向量距离为1的扰动向量，即扰动数据与原始数据有一个特征不同，则此时与原始数据距离为1的扰动输出共有$\binom{m}{1}(k-1)^1$种可能，概率的分子为$e^{\varepsilon(m-1)}\binom{m}{1}(k-1)^1$；当$z=m$时，即用户的扰动数据的所有特征都与原始数据不一样，此时共有$(k-1)^m$种可能，概率的分子为$(k-1)^m$。例如，用户拥有两个特征，每个特征域的大小为2(即$m=2,k=2$)，用户在编码后共有4种不同的值，即11,12,21和22，则用户u_i的扰动机制($\varepsilon=1$)如表1所列。

表1 随机响应的输出概率示例

Table 1 Example of output probability of a random response

X_i	Y_i			
	⟨1,1⟩	⟨1,2⟩	⟨2,1⟩	⟨2,2⟩
⟨1,1⟩	$\frac{e^2}{e^2+2e+1}$	$\frac{e}{e^2+2e+1}$	$\frac{e}{e^2+2e+1}$	$\frac{1}{e^2+2e+1}$
⟨1,2⟩	$\frac{e}{e^2+2e+1}$	$\frac{e^2}{e^2+2e+1}$	$\frac{1}{e^2+2e+1}$	$\frac{e}{e^2+2e+1}$
⟨2,1⟩	$\frac{e}{e^2+2e+1}$	$\frac{1}{e^2+2e+1}$	$\frac{e^2}{e^2+2e+1}$	$\frac{e}{e^2+2e+1}$
⟨2,2⟩	$\frac{1}{e^2+2e+1}$	$\frac{e}{e^2+2e+1}$	$\frac{e}{e^2+2e+1}$	$\frac{e^2}{e^2+2e+1}$

定义3 本文提出的扰动机制Ψ_{gperturb}满足本地d-隐私。

证明：假设有两个用户u_a，u_b，它们的特征向量分别为$\boldsymbol{X}_a=\{x_{a,1},x_{a,2},\cdots,x_{a,m}\}$，$\boldsymbol{X}_b=\{x_{b,1},x_{b,2},\cdots,x_{b,m}\}$。由扰动机制$\Psi_{\text{perturb}}$的定义可知，它们输出$\boldsymbol{Y}_c$的概率为：

$$\begin{aligned}\frac{\Pr(\boldsymbol{Y}_c\mid\boldsymbol{X}_a)}{\Pr(\boldsymbol{Y}_c\mid\boldsymbol{X}_b)}&=\frac{e^{\varepsilon(m-\sum_{j=1}^{m}x_{a,j}\oplus y_{c,j})}}{\sum_{j=0}^{m}e^{\varepsilon j}\binom{m}{j}(k-1)^{m-j}}*\frac{\sum_{j=0}^{m}e^{\varepsilon j}\binom{m}{j}(k-1)^{m-j}}{e^{\varepsilon(m-\sum_{j=1}^{m}x_{b,j}\oplus y_{c,j})}}\\&=\frac{e^{\varepsilon m-\varepsilon(\sum_{j=1}^{m}x_{a,j}\oplus y_{c,j})}}{e^{\varepsilon m-\varepsilon(\sum_{j=1}^{m}x_{b,j}\oplus y_{c,j})}}\\&=e^{\varepsilon(\sum_{j=1}^{m}x_{b,j}\oplus y_{c,j}-\sum_{j=1}^{m}x_{a,j}\oplus y_{c,j})}\\&\leqslant e^{\varepsilon(\sum_{j=1}^{m}x_{a,j}\oplus x_{b,j})}\\&=e^{\varepsilon\cdot d(\boldsymbol{X}_a,\boldsymbol{X}_b)}\end{aligned} \tag{5}$$

其中，$\boldsymbol{Y}_c$为Ψ_{perturb}的任意输出值，式(5)中的不等式由三角不等式得来。由本地d-隐私定义可知，机制Ψ_{perturb}满足d,ε-LDP。

综上所述，任意两个用户输出相同信息的概率不仅由隐私预算控制，还受它们的原始数据之间的汉明距离的影响。即对于两个原始数据相差很大的用户，他们输出不同结果的概率会很大。同时，随着用户数据维度的增加，该机制更倾向于将更多特征的关联关系保留下来。而传统的LDP中，两个任意输入输出相同结果的概率仅受隐私预算控制(两组汉明距离相差很大和相差很小的输入对在扰动过后输出相同结果的概率都小于或等于e^{ε})，本文的机制在扰动过程中保留了更多的有用信息，提高了数据实用性。

每个用户u_i通过扰动机制得到输出$\boldsymbol{Y}_i$后，将其上传给服务器端。算法1给出了用户扰动机制实现的具体过程。

算法1 Ψ_{perturb}

输入：用户u_i的原始特征向量$\mathbf{X}_i$，汉明距离的度量d(·)；用户拥有的特征数m；每个特征域的大小k；隐私预算ε

输出：用户u_i的扰动特征向量$\mathbf{Y}_i$

```
Y_i = ∅, list = [ ], distance = 0
for z ← 0 to m do:
list[z] = e^{ε(m-z)}·C(m,z)·(k-1)^z / Σ_{j=0}^{m} e^{εj}·C(m,j)·(k-1)^{m-j}  /* 划分 m+1 个区间，其中第 z 个区间代表 d(X_i · Y_i)=z 时的概率 */
end for
r ← 从(0,1)中随机取一个值 /* 用户 u_i 执行一次随机扰动 */
for x ← 0 to m do:
    if Σ_{i=0}^{x} list[i] ≥ r:
        distance = x /* 依据随机值 r 判断用户该次扰动后的数据和原始数据的汉明距离 */
        break
    end if
end for
X_i' ← 从X_i中随机选取 m−distance 个特征 /* X_i' 为该次扰动中未被改变的特征 */
for each x_ij in X_i' do:
    y_ij = x_ij
end for
for each x_ij in (X_i \ X_i') do: /* (X_i \ X_i') 为该次扰动中改变的特征 */
    x_ij ← 随机从 ([1,k] \ x_ij) 中取一个值
    y_ij = x_ij
end for
return Y_i
```

当所有用户信息上传之后，服务器开始执行聚合，可得到所有原始特征向量的无偏频数预测值。

(3)聚合($\Phi_{aggregation}$)。c_l和c_l^*分别为所有用户扰动前后的第l种特征向量的频数，$1\leqslant l\leqslant k^m$。

$$c_l=\sum_{i=1}^{n}1_l(\boldsymbol{X}_i),\ c_l^*=\sum_{i=1}^{n}1_l(\boldsymbol{Y}_i) \tag{6}$$

其中，$1_l(\cdot)$为指标函数，用于判断一个特征向量的所属类别。对于特征向量$\boldsymbol{X}_i=\{x_{i,1},x_{i,2},\cdots,x_{i,m}\}$，当$\sum_{j=1}^{m}k^{m-j-1}(x_{i,j}-$

$1)+1=l$ 时，$\mathbb{1}_l(\boldsymbol{X}_i)=1$，否则为 0。例如，用户拥有 3 个特征（每个特征域的大小为 4），即 $l\in 4^3$。假设用户 u_a 的特征向量 $\boldsymbol{X}_a=\{1,2,4\}$，则该用户所属的类别为 $0*4^2+1*4^1+3*1+1=8$，$\mathbb{1}_8(\boldsymbol{X}_a)=1$，$\mathbb{1}_{\text{else}}(\boldsymbol{X}_a)=0$。

c_l 和 c_l^* 的关系可以由式(4)得出：

$$E[c_l^*]=\sum_{j=1}^{k^m}c_j*\frac{e^{\varepsilon(m-d(c_j\cdot c_l))}}{\sum_{w=0}^{m}e^{\varepsilon w}\binom{m}{w}(k-1)^{m-w}} \tag{7}$$

其中，$E[c_l^*]$ 为 c_l^* 的期望，$d(c_j\cdot c_l)$ 表示所属类别为 j,l 的两个特征向量之间的汉明距离。

在实际情况中，服务器只可以得到 c_l^*（用户的扰动信息）的值，因此可用转换矩阵 $\boldsymbol{P}$ 来求解 c_l 的估计值。矩阵 $\boldsymbol{P}$ 是一个实对称矩阵，因此对 c_l 的计算很简单。对于 $p_{ij}\in\boldsymbol{P}$，由式(4)可知：

$$p_{ij}=\frac{e^{\varepsilon(m-d(c_i\cdot c_j))}}{\sum_{w=0}^{m}e^{\varepsilon w}\binom{m}{w}(k-1)^{m-w}} \tag{8}$$

其中，p_{ij} 为特征向量扰动前后所属类别分别为 i,j 的概率，$1\leqslant i,j\leqslant k^m$。

综上所述，有：

$$\begin{aligned}&\boldsymbol{P}\cdot[c_1,c_2,\cdots,c_{k^m}]^{\mathrm{T}}\\&=\begin{pmatrix}p_{11}&p_{12}&\cdots&p_{1k^m}\\p_{21}&p_{22}&\cdots&p_{2k^m}\\\vdots&\vdots&\ddots&\vdots\\p_{k^m1}&p_{k^m2}&\cdots&p_{k^mk^m}\end{pmatrix}\begin{pmatrix}c_1\\c_2\\\vdots\\c_{k^m}\end{pmatrix}\\&=E[c_1^*,c_2^*,\cdots,c_{k^m}^*]\end{aligned} \tag{9}$$

由矩阵 $\boldsymbol{P}$ 和 c_l^* 的期望的线性性质可知，通过向量 $[c_1^*,c_2^*,\cdots,c_{k^m}^*]$ 可以估计原始向量 $[c_1,c_2,\cdots,c_{k^m}]^{\mathrm{T}}$ 为：

$$\begin{aligned}\hat{\boldsymbol{c}}&=[\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_{k^m}]^{\mathrm{T}}\\&=E[c_1,c_2,\cdots,c_{k^m}]^{\mathrm{T}}\\&=\boldsymbol{P}^{-1}\cdot[c_1^*,c_2^*,\cdots,c_{k^m}^*]\end{aligned} \tag{10}$$

当矩阵 $\boldsymbol{P}$ 可逆时，显然该聚合机制对 c_l 的估计值是无偏的。聚合机制的详细步骤如算法 2 所示。

算法 2 $\Phi_{\text{aggregation}}$

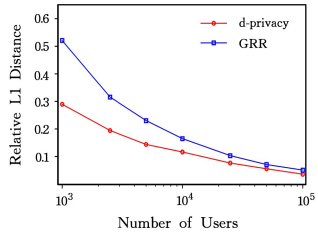
输入：所有用户的扰动特征向量 $\mathbf{Y}$；汉明距离的度量 d(·)；用户拥有特征数 m；每个特征域的大小 k

输出：对所有特征向量频数的无偏预测值 $\hat{\mathbf{c}}=[\hat{c}_1,\hat{c}_2,\cdots,\hat{c}_{k^m}]$

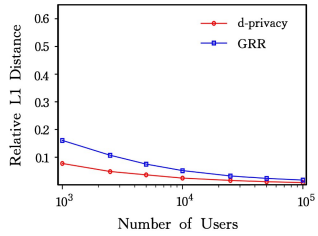
1. $\mathbf{c}^*=[c_1^*=0,c_2^*=0,\cdots,c_{k^m}^*=0]$，$\mathbf{P}=\varnothing$
2. for each y in **Y** do:
3. $\quad l=\sum_{j=1}^{m}k^{m-j-1}(y_{i,j}-1)+1$ /* 判断每个用户扰动向量 **y** 的所属类别 */
4. $\quad c_l^*=c_l^*+1$
5. end for
6. for i ← 1 to k^m do:
7. $\quad$ for j ← 1 to k^m do:
8. $\quad p_{ij}=\dfrac{e^{\varepsilon(m-d(c_i\cdot c_j))}}{\sum_{w=0}^{m}e^{\varepsilon w}\binom{m}{w}(k-1)^{m-w}}$ /* 求扰动矩阵 */
9. $\quad$ end for
10. end for
11. $\hat{\mathbf{c}}=\mathbf{P}^{-1}\cdot\mathbf{c}^*$ /* 预测原始数据的无偏估计值 */
12. for each c in $\hat{\mathbf{c}}$ do:
13. $\quad$ c=0 if c<0/* 预测为负的值调整为 0 */
14. end for
15. for each c in $\hat{\mathbf{c}}$ do:
16. $\quad c=\left\lceil\dfrac{c}{\text{sum}(\hat{\mathbf{c}})}*|\mathbf{Y}|\right\rceil$ /* 归一化 */
17. end for
18. return $\hat{\mathbf{c}}$

算法 2 中，步骤 11—步骤 18 是为了修正频数预测中的一些错误值，当隐私预算较小时，对某些特征向量的频数预测可能为负值，将这些值置 0，并对其他特征向量的频数预测做归一化处理。

为了验证该频数预测算法的实用性，设计了一个频数预测实验，比较了本文频数预测算法与传统 LDP 频数预测算法（Generalized Randomized Response，GRR）[25] 的预测结果准确率，并展示了本地化差分隐私算法的实用性如何受限于参与算法的用户数。实验中，从标准差 $\sigma=12$ 的高斯分布中采样每个用户的特征向量（用户有 3 个特征，每个特征域的大小为 2），并将其扰动后上传给数据收集器，数据收集器则根据聚合机制来估计每个特征向量的真实频数。对于频数预测算法的准确率，通过特征向量的真实频数和预测频数之间的 L1 距离与参与算法人数的比值来衡量，该值越小，预测结果就越准确，为 0 则代表预测结果与真实结果一致。针对 1 000 至 100 000 之间的不同数量的用户运行此实验，数据收集器进行的频数预测的误差如图 2 所示。



(a)ε=1



(b)ε=2

图 2 频数预测误差对比

Fig. 2 Comparison of frequency prediction error

从图 2 中可以看到，两种算法的预测结果的准确率都随着隐私预算和用户数量的增加而提高，前者是因为减少了噪声，后者则可以用大数定律[26]来解释，即重复进行多次相同的实验，其随机事件的频数近似于它的概率。从图 2 中还可

以看出，本文提出的本地 d-privacy 下的频数预测算法预测的准确率高于 GRR 算法，尤其当用户数量较少时，优势更加突出，估计误差约为 GRR 误差的一半，因此可以很好地处理小用户群。

4.3 LDPK-modes 聚类方案

本节提出的 LDPK-modes 聚类方案的框架如图 3 所示，由用户端和第三方服务器端组成。由于第三方服务器被认为不可信，因此数据扰动的功能由用户端完成，服务器端则实现对扰动数据的聚合与聚类。用户端采用 3.2 节中的方法先将其特征向量编码成整数，然后对编码后的信息用随机响应机制进行扰动，最后将扰动后的数据上传给服务器，具体过程如算法 1 所示。服务器收到扰动数据后，首先按照算法 2 进行聚合，预测所有特征向量的无偏预测值，然后服务器根据预测值生成与用户原始信息分布相似的合成数据集，具体过程如算法 3 所示，最后在生成的数据集上选择初始聚类中心，并进行 k-modes 聚类，具体过程如算法 4 所示。

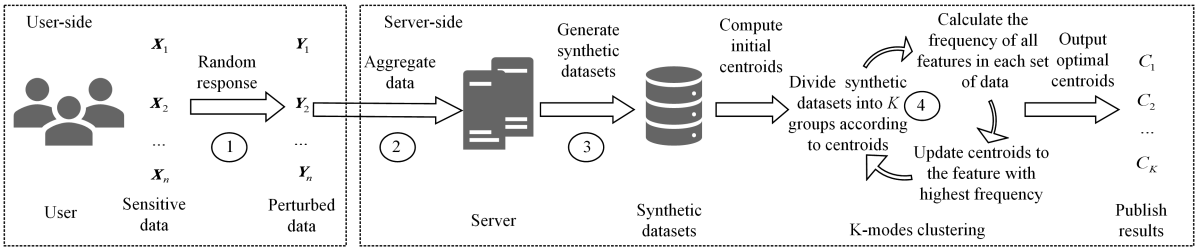


图 3 LDPK-modes 方案的框架

Fig. 3 Framework of LDPK-modes

算法 3 生成合成数据集

输入：所有特征向量无偏预测值 $\hat{\mathbf{c}}$；聚类个数 K；用户拥有特征数 m；每个特征域的大小 k；用户总数 n

输出：合成数据集 $\hat{\mathbf{X}}$

$\hat{\mathbf{X}}=\emptyset$，i=0 /* 初始化合成数据集，其中 $\hat{X}_i\in\hat{\mathbf{X}}$，$\hat{X}_i=\{\hat{x}_{i,1},\hat{x}_{i,2},\cdots,\hat{x}_{i,m}\}$，$1\leqslant i\leqslant n$ */

for x ← 0 to (k^m-1) do:

  for y ← 0 to ($\hat{\mathbf{c}}$[x]−1) do: /* 生成 $\hat{\mathbf{c}}$[x]个所属类别为 x 的合成数据 */

    temp=x

    for z←1 to m do: /* 生成所属类别为 x 的数据的 m 个特征 */

      $g=\left\lfloor\frac{temp}{k^{m-z}}\right\rfloor+1$

      $\hat{x}_{i,z}=g$

      temp=temp−g+1

    end for

    i=i+1

  end for

end for

return $\hat{\mathbf{X}}$

算法 3 中，服务器根据预测值生成一个用于聚类分析的合成数据集。合成数据集可得到用户数据的真实分布情况的无偏预测，并且无须曝光具体某个用户的实际特征向量，很好地保护了用户的隐私。为了得到最优的聚类中心集合，需要多次迭代聚类算法。中心化差分隐私下的聚类算法每次迭代访问原始数据集中的数据都会划分隐私预算，产生更多的噪声。而 LDPK-modes 使用合成数据集进行聚类迭代，整个过程中服务器只需在生成合成数据集阶段与用户数据交互一次，用户在该阶段使用全部的隐私预算扰动数据。同时，合成数据集还可用于更多的场景，如初始聚类中心点的计算、范围查询、Top-k 查询等。

聚类过程如算法 4 所示。

算法 4 基于合成数据集进行 k-modes 聚类

输入：合成数据集 $\hat{\mathbf{X}}$；聚类个数 K；迭代次数 T；用户拥有特征数 m；每个特征域的大小 k；用户总数 n

输出：聚类中心 $C=\{C_1,C_2,\cdots,C_K\}$

f←$\lfloor(\sqrt[m]{K})\rfloor$，$C^0=\emptyset$，candidate=$\emptyset$ /* candidate 为候选初始中心点 */

for x←1 to m do:

  for y←1 to f do:

    v_x^y←$\hat{\mathbf{X}}$中第 x 个特征出现频数第 y 大的值

  end for

end for

while ($|C^0|\leqslant K$):

  for j ← 1 to m do:

    r←随机从[1,f]选取一个整数

    candidate 的第 j 个分量←v_j^r

  end for

  if (candidate$\notin C^0$):

    C^0.append (candidate) /* 生成 K 个不同的初始聚类中心点 */

  end if

end while

/* 迭代聚类 */

for x ← 1 to T do:

  $\hat{\mathbf{X}}_1,\hat{\mathbf{X}}_2,\cdots,\hat{\mathbf{X}}_K$←将$\hat{\mathbf{X}}$中数据划分到离 C^{x-1} 中最近的聚类中心

  for y←1 to K do:

    for z←1 to m do:

      for g←1 to k do:

        $n_{yz}(g)$←$\hat{\mathbf{X}}_y$中第 z 个特征第 g 个值的计数

      end for

      C_y^x第 z 个分量← $\underset{g}{\mathrm{argmax}}(n_{yz}(g))$

    end for

  end for

  if ($C^{x-1}==C^x$):

    return C^x

  end if

end for

return C^T

算法 4 中，步骤 1－步骤 15 将合成数据集中最频繁的前 f 个特征均匀地分配给初始 K 个不同的聚类中心，相比随机选取初始中心可能陷入局部最优解，该方法有助于 k-modes 算法快速收敛到簇内总距离低（聚类效果好）的中心点。在得到初始聚类中心点后，步骤 16－步骤 30 执行 k-modes 聚类算法，并在达到迭代次数或者聚类中心稳定后，返回最终的聚类结果。

5 实验与结果分析

5.1 实验环境与数据

本文实验环境为 Intel(R) Core(TM) i5-9400F CPU @ 2.90 GHz，16 GB 内存，操作系统为 Windows10 64 位，用 Python 语言进行仿真实验。为了验证本文算法的有效性，使用来自 UCI 数据集中的 Auto Mpg，Car Evaluation，Nursery 和 CoverType 4 个真实数据集来进行分析对比。Auto-Mpg 记录了 398 辆汽车的行驶记录，每条记录由汽缸、行驶距离、马力、重量、加速度、车型、年份和产地这 8 个属性描述。Car Evaluation 记录了 1 728 名顾客的汽车购买信息，每条信息由汽车售价、保养价格、车门数量、乘载人数、容量和舒适度这 6 个方面构成。Nursery 最初是为托儿所的申请排名而开发的分层决策模型，包含由 8 个分类型数据构成的 12 960 条记录。CoverType 则为 581 012 条森林有关信息，每条信息由 10 个数值型数据和 2 个分类型数据构成。为了增强聚类效果，本文对各数据集进行了相应的预处理，处理后的数据形式如表 2 所列。

表 2 预处理后的数据集

Table 2 Datasets after preprocessing

Data Set	Alias	Records	K	$\lvert \boldsymbol{X}_i \rvert$	$\{\lvert v_j \rvert\}$
Auto Mpg	D1	398	3	3	{323}
Car Evaluation	D2	1 728	4	6	{222222}
Nursery	D3	12 960	5	8	{22222222}
CoverType	D4	581 012	7	2	{47}

5.2 评价标准

首先采用规范化簇内方差（Normalized Intra-Cluster Variance，NIVC）[27] 来衡量聚类效果，其计算公式如下：

$$NIVC=\frac{1}{N}\sum_{i=1}^{N}d(\boldsymbol{X}_i \cdot \underset{C_g,C_g\in C}{\operatorname{argmin}}\, d(\boldsymbol{X}_i \cdot C_g)) \tag{11}$$

其中，$\boldsymbol{X}_i$ 是第 i 个用户的特征向量，C 是聚类中心集合，N 为数据集大小。$NIVC$ 值越小代表聚类中心与簇中数据距离越近，聚类效果越好。反之，说明聚类效果越差。

其次，由于噪声的引入，差分隐私保护聚类算法的结果实用与否尤其重要。因此，采用 F-measure[28] 值来衡量聚类结果的可用性。F-measure 综合考虑了信息检索与挖掘中的准确率（Accuracy，AC）和召回率（Recall，RE）。相比其他评价指标，F-measure 的结果更具有针对性。给定一个数据集 X，K 为聚类个数，$\mathbb{C}=\{\mathbb{C}_1,\mathbb{C}_2,\cdots,\mathbb{C}_K\}$ 为差分隐私保护聚类算法对 X 进行聚类的结果，$P=\{P_1,P_2,\cdots,P_K\}$ 是真实标签，上述衡量指标的定义如下：

$$AC=\frac{1}{N}\max_{j_1,j_2,\cdots,j_K\in S}\sum_{i=1}^{K}n_{ij_i} \tag{12}$$

$$RE=\frac{1}{K}\sum_{i=1}^{K}\frac{n_{ij_i^*}}{n_{\mathbb{C}_i}} \tag{13}$$

$$F\text{-}Measure=\frac{(\alpha^2+1)AC*RE}{\alpha^2(AC+RE)} \tag{14}$$

其中，N 为数据集大小，$n_{ij}=|\mathbb{C}_i\cap P_j|$，$n_{ij_i^*}$ 表示正确分到 i 类的对象个数，$n_{\mathbb{C}_i}$ 表示 $\mathbb{C}_i$ 中对象的个数。

为使准确率和召回率获得相等的权重，设置 $\alpha=1$。*F-measure* 值的范围为[0，1]，其值越大说明聚类结果的可用性越高。

5.3 实验结果与分析

本文分别在 4 个数据集上运行 K-modes 算法、DP-modes-Lloyd 算法和本文提出的 LDPK-modes 方法。其中，K-modes 算法为未受任何隐私保护的原始聚类算法，DP-modes-Lloyd 算法为中心化差分隐私保护的 K-modes 聚类算法。首先，在各个数据集上运行 1 000 次 K-modes 算法，并取得在聚类效果最好情况下的 NIVC 指标，并将其作为 DP-modes-Lloyd 和 LDPK-modes 性能评估的参照物。然后，在 4 个数据集上分别运行 50 次 DP-modes-Lloyd 和 LDPK-modes，迭代次数设置为 $T=1$ 和 $T=5$，并逐步将隐私预算 ε 从 0.1 提升至 2。

图 4 和图 5 分别给出了迭代次数 $T=1$ 和 $T=5$ 时，在数据集 D1－D4 上执行两种聚类算法得到的 $NIVC$ 结果。

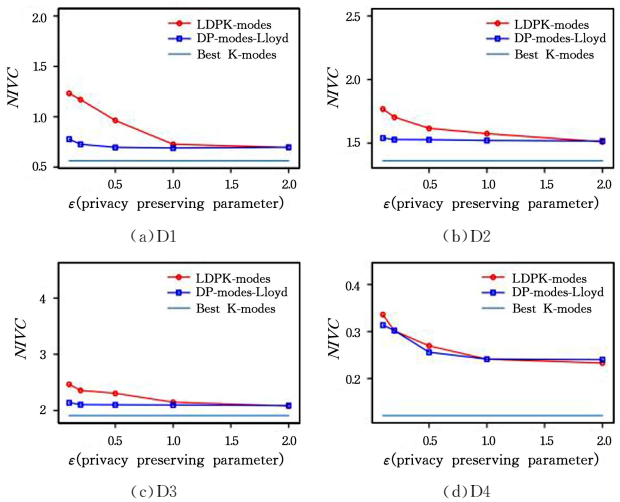


图 4 $T=1$ 时 $NIVC$ 值的比较

Fig. 4 Comparison of $NIVC$ value when $T=1$

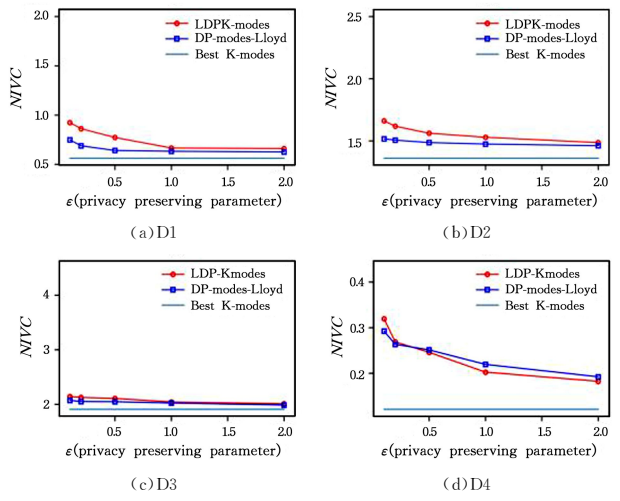


图 5 $T=5$ 时 $NIVC$ 值的比较

Fig. 5 Comparison of $NIVC$ value when $T=5$

图 6 给出了在迭代次数 $T=5$ 时，4 个数据集上两种算法在不同隐私预算大小下 *F-measure* 值的比较。

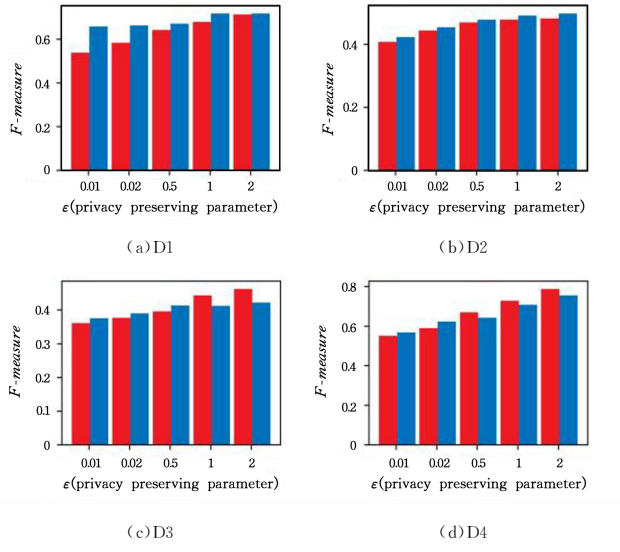


图 6 $T=5$ 时 *F-measure* 值的比较

Fig. 6 Comparison of *F-measure* value when $T=5$

(1)隐私预算 ε 对 *NIVC* 值的影响

如图 4、图 5 所示，在隐私预算 ε 较小时，DP-modes-Lloyd 比 LDPK-modes 有着更好的聚类效果，但随着隐私预算的增加，这种优势逐渐减小，LDPK-modes 在隐私预算 ε 较大时，聚类性能优于 DP-modes-Lloyd。这是因为本地化差分隐私技术没有一个可信的数据处理第三方，在较小的隐私预算的情况下，为了达到相似的隐私保护程度需要引入更大的噪声来保护原始数据。随着隐私预算的增加，本地d-privacy可以以更高的概率将原始数据的相关信息保留下来，而以较小的概率将原始数据扰动成与其相差很大的数据。

(2)迭代次数对 *NIVC* 值的影响

随着迭代次数的增加，图 5 中 LDPK-modes 和 DP-modes-Lloyd 的 *NIVC* 值相比图 4 更加切合，且 LDPK-modes 的 *NIVC* 值比 DP-modes-Lloyd 收敛于最优 *NIVC* 值更加迅速。这是因为，两种算法都因迭代次数的增加而获得了更优的聚类结果，但在 DP-modes-Lloyd 算法中，随着迭代次数的增加，每轮迭代都要访问一次原始数据，每次访问都需要消耗一部分隐私预算，而在总隐私预算固定的情况下，每轮迭代获得的隐私预算越小，在每次更新聚类中心点时引入的噪声就越多。而 LDPK-modes 采用合成数据集进行聚类，服务器只需在生成数据集阶段与所有用户进行一次交互，因此在每轮迭代中无需划分隐私预算，减少了噪声的引入。

(3)数据集大小对 *NIVC* 值的影响

图 4 中，LDPK-modes 和 DP-modes-Lloyd 的 *NIVC* 值随着数据集中数据量的增加更加切合。这是因为在本地化差分隐私保护的算法中，算法使用的数据数目决定了数据的实用性能。数据数目越大，算法结果的实用性就越高。数据数目越小则实用性越低，这一特性可由大数定律来解释。本地化差分隐私中，每个用户分别以相同的随机响应机制来扰动其数据，然后服务器基于公共的概率参数对扰动结果进行统计分析。由于用户的随机响应机制是随机事件，用户的实际扰动信息往往不同于理论扰动信息。而服务器则跟据实际的扰动信息来分析数据统计信息。随着数据数目增大，随机事件的频数与概率之间的相对差值减小，本地化差分隐私的分析结果则更加接近真实结果。

(4)隐私预算 ε 对 *F-measure* 值的影响

由图 6 可知，在相同的 ε 下，LDPK-modes 和 DP-modes-Lloyd 具有相似的 *F-measure* 值，因此本文算法和 DP-modes-Lloyd 算法有着相似的聚类可用性。并且，随着 ε 的增加，*F-measure* 值也逐渐增加，这是因为隐私保护程度降低之后，添加的噪声量也会降低，聚类效果会得到提高。但是，本文提出的 LDPK-modes 方案与 DP-modes-Lloyd 相比，在整个聚类过程中，只有用户自己可以接触到原始数据，从而不需要担忧第三方数据收集者是否可信，因此隐私保护性能更优，适用场景更广，实用程度更高。

结束语 本文提出了一种带有距离度量的本地化差分隐私聚类方案，该方案根据数据之间的汉明距离缩放隐私预算大小，以提高频数预测精度，改进聚类中心点的选择，使 k-modes 算法快速输出 *NIVC* 值低的聚类结果。在不需要可信第三方的前提下，本文算法与中心化差分隐私 k-modes 算法有着相似的聚类性能表现。受 k-modes 算法的影响，本文算法只能处理分类型数据，今后将研究支持更为复杂的数据类型的本地化差分隐私聚类方法。

参 考 文 献

[1] DWORK C. Differential privacy: A survey of results[C]// International Conference on Theory and Applications of Models of Computation. Springer, Berlin, Heidelberg, 2008: 1-19.

[2] SWEENEY L. k-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[3] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. l-diversity: Privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 3.

[4] 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.

[5] 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. IEEE, 2013: 429-438.

[6] KASIVISWANATHAN S P, LEE H K, NISSIM K, et al. What can we learn privately[C]// Proc. of the 49th Annual IEEE Symp. on Foundations of Computer Science (FOCS). IEEE, 2008: 531-540.

[7] HOPE T, CHAN J, KITTUR A, et al. Accelerating innovation through analogy mining[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 235-243.

[8] MENDES R, VILELA J P. Privacy-preserving data mining:

methods,metrics,and applications[J]. IEEE Access,2017,5:10562-10582.

[9] HAMMING R W. Error detecting and error correcting codes [J]. The Bell System Technical Journal,1950,29(2):147-160.

[10] GU X,LI M,CAO Y,et al. Supporting both range queries and frequency estimation with local differential privacy[C]// 2019 IEEE Conference on Communications and Network Security (CNS). IEEE,2019:124-132.

[11] WARNER S L. Randomized response:A survey technique for eliminating evasive answer bias[J]. Journal of the American Statistical Association,1965,60(309):63-69.

[12] 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. 2014:1054-1067.

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

[14] BASSILY R,SMITH A. Local,private,efficient protocols for succinct histograms[C]// Proc. of the 47th Annual ACM on Symp. on Theory of Computing. ACM,2015:127-135.

[15] DUCHI J C,JORDAN M I,WAINWRIGHT M J. Local privacy,data processing inequalities,and statistical minimax rates [J]. arXiv:1302.3203,2013.

[16] WAINWRIGHT M J,JORDAN M I,DUCHI J C. Privacy aware learning[C]// Advances in Neural Information Processing Systems. 2012:1430-1438.

[17] NGUYÊN T T,XIAO X,YANG Y,et al. Collecting and analyzing data from smart device users with local differential privacy [J]. arXiv:1606.05053,2016.

[18] BLUM A,DWORK C,MCSHERRY F,et al. Practical privacy:the SuLQ framework[C]// Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2005:128-138.

[19] REN J,XIONG J,YAO Z,et al. DPLK-means:A novel Differential Privacy K-means Mechanism[C]// 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC). IEEE,2017:133-139.

[20] FU Y M,LI Z Z. Research on *k*-means++ Clustering Algorithm Based on Laplace Mechanism for Differential Privacy Protection[J]. Netinfo Security,2019,19(2):43-52.

[21] HU C,YANG G,BAI Y L. Clustering Algorithm in Differential Privacy Preserving[J]. Computer Science,2019,46(2):120-126.

[22] XIA C,HUA J,TONG W,et al. Distributed K-Means clustering guaranteeing local differential privacy[J]. Computers & Security,2020,90:1-11.

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

[24] LYU Z,WANG L,GUAN Z,et al. An optimizing and differentially private clustering algorithm for mixed data in SDN-based smart grid[J]. IEEE Access,2019,7:45773-45782.

[25] WANG T,BLOCKI J,LI N,et al. Locally differentially private protocols for frequency estimation[C]// 26th Security Symposium (Security 17). 2017:729-745.

[26] NEWEY K W,MCFADDEN D. Large sample estimation and hypothesis[J]. Handbook of Econometrics,1994,4:2111-2245.

[27] NISSIM K,RASKHODNIKOVA S,SMITH A. Smooth sensitivity and sampling in private data analysis[C]// Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing. 2007:75-84.

[28] JIANG H,YI S,LI J,et al. Ant clustering algorithm with K-harmonic means clustering[J]. Expert Systems with Applications,2010,37(12):8679-8684.

PENG Chun-chun, born in 1996, postgraduate. His main research interests include privacy preserving and data mining.

CHEN Yan-li, born in 1969, Ph.D, professor. Her main research interests include network security and computer architecture.