K-Modes 聚类数据收集和发布过程中的混洗差分隐私保护方法

蒋伟进[1,2,3]，陈艺琳[1,3]，韩裕清[1,3]，吴玉庭[1,3]，周为[1,3]，王海娟[3,4]

（1. 湖南工商大学计算机学院，湖南 长沙 410205；2. 武汉理工大学计算机与人工智能学院，湖北 武汉 430070；
3. 湘江实验室，湖南 长沙 410205；4. 湖南工商大学前沿交叉学院，湖南 长沙 410205）

摘 要：针对目前聚类数据收集与发布安全性不足的问题，为保护聚类数据中的用户隐私并提高数据质量，基于混洗差分隐私模型，提出一种去可信第三方的 K-Modes 聚类数据收集和发布的隐私保护方法。首先，使用 K-Modes 聚类数据收集算法对用户数据进行采样并加噪，再通过填补取值域随机排列发布算法打乱采样数据的初始顺序，使恶意攻击者不能根据用户与数据之间的关系识别出目标用户。然后，尽可能减小噪声的干扰，利用循环迭代的方式计算出新的质心完成聚类。最后，从理论层面上分析了以上 3 种方法的隐私性、可行性和复杂度，并利用 3 个真实数据集和近年来具有权威性的同类算法 KM、DPLM、LDPKM 等进行准确率、熵值的对比，验证所提方法的有效性。实验结果表明，所提方法的隐私保护和发布数据质量均优于当前同类算法。

关键词：混洗差分隐私；K-Modes 聚类；隐私保护；数据收集；数据发布
中图分类号：TP309
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2024004

Shuffled differential privacy protection method for K-Modes clustering data collection and publication

JIANG Weijin[1,2,3], CHEN Yilin[1,3], HAN Yuqing[1,3], WU Yuting[1,3], ZHOU Wei[1,3], WANG Haijuan[3,4]

1. School of Computer Science, Hunan University of Technology and Business, Changsha 410205, China
2. School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China
3. Xiangjiang Laboratory, Changsha 410205, China
4. College of Frontier Intersection, Hunan University of Technology and Business, Changsha 410205, China

Abstract: Aiming at the current problem of insufficient security in clustering data collection and publication, in order to protect user privacy and improve data quality in clustering data, a privacy protection method for K-Modes clustering data collection and publication was proposed without trusted third parties based on the shuffled differential privacy model. K-Modes clustering data collection algorithm was used to sample the user data and add noise, and then the initial order of the sampled data was disturbed by filling in the value domain random arrangement publishing algorithm. The malicious attacker couldn't identify the target user according to the relationship between the user and the data, and then to reduce the interference of noise as much as possible a new centroid was calculated by cyclic iteration to complete the clustering. Finally, the privacy, feasibility and complexity of the above three methods were analyzed from the theoretical level, and the accuracy and entropy of the three real data sets were compared with the authoritative similar algorithms KM, DPLM and LDPKM in recent years to verify the effectiveness of the proposed model. The experimental results show that the privacy protection and data quality of the proposed method are superior to the current similar algorithms.

Keywords: shuffled differential privacy, K-Modes clustering, privacy protection, data collection, data publication

收稿日期：2023−08−07；**修回日期：**2023−11−08

通信作者：陈艺琳，chen1986746095@163.com

基金项目：国家自然科学基金资助项目（No.72088101, No.61772196）；湖南省自然科学基金重点资助项目（No.2020JJ4249）；新零售虚拟现实技术湖南省重点实验室基金资助项目（No.2017TP1026）；湖南省教育厅科学研究重点基金资助项目（No.21A0374）；湖南省学位与研究生教学改革基金资助项目（No.2022JGYB194）

Foundation Items: The National Natural Science Foundation of China (No.72088101, No.61772196), The Key Project of the Natural Science Foundation of Hunan Province (No.2020JJ4249), Key Laboratory of Hunan Province for New Retail Virtual Reality Technology (No.2017TP1026), Key Scientific Research Project of Hunan Provincial Department of Education (No.21A0374), Hunan Provincial Degree and Graduate Teaching Reform Project (No.2022JGYB194)

0 引言

聚类是通过计算不同对象之间的相似度来对其进行分类的过程，可将相似性较高的对象归为同一类。大多数聚类数据包含用户隐私信息，如果数据拥有者不经过保护处理便收集或公开发布数据，必然会造成众多用户的隐私信息泄露。

目前，数据的收集和发布领域中，保护数据隐私主要依赖于传统的差分隐私（DP, differential privacy）模型，包括中心化差分隐私（CDP, central differential privacy）[1-2]和本地化差分隐私（LDP, local differential privacy）[3-4]。中心化差分隐私无法保证数据收集者是完全可信的，而本地化差分隐私对数据进行加噪处理，导致数据的可用性变差。考虑到上述 2 种模型的不足，混洗差分隐私（SDP, shuffled differential privacy）模型[5-7]在数据和数据中心之间引入随机排列者（RA, random arranger）[8]，主要负责对数据中心的加噪数据进行重新排序。随机排列者混淆用户和数据之间的对应关系，迷惑恶意攻击者对目标对象的身份鉴别，从而实现用户数据隐私保护。

随着聚类数据的广泛应用与分析，研究其隐私保护成为国内外学者的关注点。Zhang 等[9]对随机采样得到的数据进行扰动，并通过服务端与用户端的交互迭代完成聚类。Sassi 等[10]用 K-Modes 算法聚类动作，将聚合后的动作用于关联算法，进而实现一个端到端的竞争情报解决方案。Coelho 等[11]提出一种新颖的分区聚类算法，它能够完全恢复基因表达的 2 个数据集的真实聚类结构。Xiao 等[12]提出一种用于 K-Modes 聚类的整数线性规划方法，其能够为小规模数据集提供最优聚类结果，直接对该线性规划方法优化求解。Duan 等[13]融合局部和全局信息以进行一步多视图聚类，将子空间表示学习、多视图信息融合和聚类结合到联合优化模型中，实现一步聚类。Zhang 等[14]提出了混洗应答机制以线性分解的方式扰动用户数据，使用用户消息均匀随机排列算法保护了用户数据。Balcer 等[15]在混合模型中提出差分隐私直方图协议，其误差不依赖于论域大小。方晨等[16]设计多种优化策略来调整隐私预算分配，减小总体噪声规模。Liu 等[17]根据谱聚类生成初始分区，采用频繁项集和贪心搜索相结合的策略对初始分区进行优化，达到减小解空间的目的。Zhang 等[18]通过边缘计算服务器和云服务器两层聚合，提高聚合效率，降低通信开销。Liang 等[19]设计一种兼顾可用性和发布效率的面向数据流频繁模式发布的差分隐私保护方案。Wang 等[20]提出了一种满足差分隐私数据流关键模式挖掘算法，平衡了隐私和数据效用。Chen 等[21]通过神经网络算法预测隐私参数，明显提高了预测的精度和效率。Tian 等[22]根据轨迹数据分布特征生成位置聚簇，利用抽样机制和指数机制选择出各个代表元，从而泛化原始数据形成待发布轨迹数据。张东月等[23]通过构建网格划分评分指标以及循环反馈机制，在服务端提出自适应网格聚合方法。陆佳炜等[24]通过基于密度信息的聚类中心检测方法，筛选出最合适的 k 个语义特征向量作为 k-means 算法的初始中心，进行聚类划分。

综上所述，针对传统差分隐私在 K-Modes 聚类数据的收集和发布方面存在的诸多挑战以及现有方案的不足，本文提出了基于混洗差分隐私的 K-Modes 聚类数据收集与发布（SDPKM, shuffled differential privacy protection for K-Modes clustering data collection and publication）方法。该方法首先使用 K-Modes 聚类数据收集（CDC-KM, K-Modes clustering data collection）算法收集数据并加噪扰动，然后随机排列者对数据重新排序并发布，最后对数据进行迭代式求解质心。本文的主要贡献如下。

1) 构建一个面向 K-Modes 数据收集和发布过程的混洗差分隐私保护模型，有效提高了聚类数据的隐私安全性。

2) 设计了一种取值域填补随机排列发布（FRV-RPP, fill in range of value random permutation pulication）算法，其中，随机排列者对扰动后的数据重新排列，混淆用户与数据之间的对应关系。

3) 提出了一种 K-Modes 聚类数据收集算法，其核心思想是利用随机采样和添加噪声项收集并扰动用户数据。

4) 从隐私性方面证明提出的方案满足 SDP，并用理论分析验证其可行性，通过真实数据集进行实验验证，结果表明，该模型具有较高的可行性和有效性。

1 本文方法描述

在 K-Modes 聚类数据收集和发布过程中的混洗差分隐私方法包含恶意攻击者、用户、数据中

心和随机排列者四类实体。用户为了获得更好的服务必须把个人信息提交到服务器中，其中包含个人的隐私信息，混洗差分隐私方法在数据发布之前添加随机混洗步骤，消除任何潜在的数据模式；恶意攻击者是外部的威胁，会在不同环节以各种形式、方法发起攻击以获取用户隐私数据，进而通过分析数据来推断出敏感信息；数据中心是数据处理的核心部分，负责收集用户数据，使用抽样函数随机选取并添加噪声项，根据数据的取值域大小，计算、确定隐私预算，用虚拟值填补取值域以保持一致，利用 K-Modes 聚类算法更新迭代出新的质心，随机扰动数据，并将带噪数据发送给随机排列者；随机排列者为了防止恶意攻击者通过分析数据顺序来推断出敏感信息，重新排序噪声数据，传递给数据中心便于发布数据，服务用户。

1.1　问题描述

假设存在一组用户集 $U=\{u_1,u_2,u_3,\cdots,u_k\}$ 和相关的属性集 $M=\{M_1,M_2,M_3,\cdots,M_d\}$，每个用户 u_i $(1\leqslant i\leqslant n)$ 拥有一个 d 维的属性元组 $\mathrm{ma}_i=\{m_1,m_2,m_3,\cdots,m_d\}$，$m_j$ $(1\leqslant j\leqslant d)$ 是用户 u_i 的属性中的任意一个。K-Modes 算法按照距离相似度把用户划分成 k 个簇 $C=\{c_1,c_2,c_3,\cdots,c_k\}$，在这一过程中，通常含有用户的隐私信息，存在泄露敏感信息的风险，本文采用混洗差分隐私模型混淆用户和信息之间的对应关系，从而保护用户信息安全。

1.2　FRV-RPP 算法

1.2.1　FRV-RPP 算法概述

本文用 a_i 表示用户的属性 M_i 的取值域大小，$a_{i_{\max}}$ 表示每个用户的属性 M_i 的最大取值。FRV-RP 算法采用填补取值域的方法来消除属性间的异构性，分别对每个属性维度中添加 $a_{\max}-a_{i_{\max}}$ 个取值，使用户的属性取值域统一化，所有的属性取值为 $a_{\max}$。

填充后的值域使用 RR 机制，RR 机制是一种常用的扰动机制，当用户的原始数据和数据集 D_s 中某个随机取值相等时，用户发布真实值的概率为 $p=\dfrac{\mathrm{e}^{\varepsilon}}{\mathrm{e}^{\varepsilon}+d-1}$；否则，发布非真实值的概率为 $q=\dfrac{1}{\mathrm{e}^{\varepsilon}+d-1}$，其中 ε 为隐私预算。接着，随机排列者将扰动后的用户数据 $([\langle M_i,y_i\rangle]_{i\in[n]})$ 重新排列为 $([\langle \tilde{M}_i,\tilde{y}_i\rangle]_{i\in[n]})$，计算属性 M_i 取值为 v 的频率为

$$\tilde{f}_v=\frac{d\sum\limits_{i\in[n]} l_{\{\tilde{M}=i\wedge y_i=v\}}-nq}{n(p-q)}$$

最后，消除填补位置。具体过程如算法 1 所示。

算法 1　FRV-RPP

输入　n 个用户的数据（第 i 个用户数据为 $v_i=(v_{i1},v_{i2},v_{i3},\cdots,v_{id})$），属性维度 d 及相应的取值域大小 $\{m_1,m_2,m_3,\cdots,m_d\}$，最小隐私预算 ε_c，隐私泄露概率 $\delta\in(0,1]$

输出　发布结果 F

1) $a_{\max}=\max\{m_1,m_2,m_3,\cdots,m_d\}$;
2) $\varepsilon_l=\mathrm{Calculate}(\varepsilon_c,\delta)$; //计算放大后的隐私预算 ε_l
3) for 属性 $i=1$ to d do
4) 　　$\mathrm{Putting}(a_i,a_{\max})$; //用虚拟值填充每个维度的取值域
5) end for
6) for 用户 $i=1$ to n do
7) 　　$M_i=\mathrm{Random}(0,d)$; //随机选择一维属性
8) 　　$y_i=\mathrm{RR}(v_{i\theta_i},a_{\max},\varepsilon)$; //在 $a_{\max}$ 内扰动
9) 　　u_i send M_i,y_i to Shuffler S; //第 i 个用户把带噪数据 M_i 和 y_i 发送给随机排列者
10) end for
11) Shuffler S //打乱所有带噪数据，形成发布结果 F，并将其发送给收集方
12) return F

1.2.2　FRV-RPP 的隐私性能分析

定理 1　FRV-RPP 满足 (ε_c,δ)-混洗差分隐私，$\varepsilon_c\leqslant\sqrt{\dfrac{14\ln\left(\dfrac{2}{\delta}\right)(\mathrm{e}^{\varepsilon_l}+a_{\max}-1)}{n-1}}$，$a_{\max}$ 是用户数据属性的最大取值。

证明　用 A 表示填补取值域的随机排列发布算法，M_i 表示第 i 个用户随机选取的要发布数据的属性，$([\langle M_i,y_i\rangle]_{i\in[n]})$ 表示随机排列者扰乱前的用户输出数据，用 R 表示随机排列者扰乱后的用户输出数据 $([\langle \tilde{M}_i,\tilde{y}_i\rangle]_{i\in[n]})$，$\prod$ 表示 n 个用户的一种排列方式。接下来证明 $\mathrm{Pr}_{R\sim A(D)}\left[\dfrac{\Pr[A(D)=R]}{\Pr[A(D')=R]}\geqslant\mathrm{e}^{\varepsilon_c}\right]\leqslant\delta$。

这里用 T 表示前 $n-1$ 个用户中发送真实值的一类，R_T 表示前 $n-1$ 个用户中发布随机值的一类，前提是第 n 个用户发布真实值，$\Pr\left[A(D_s)=R\,|\,(T,R_T)\right]$ 表示为

$$\Pr\left[A(D_s)=R\,|\,(T,R_T)\right]=\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}l_{\{\Theta_{\pi(j)}=\tilde{\theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}\prod_{j\in R_T}\frac{1}{d}\frac{1}{k_{\max}}\frac{1}{d}l_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\}\tag{1}$$

根据式(1)可得

$$\frac{\Pr\left[A(D)=R|(T,R_T)\right]}{\Pr\left[A(D')=R|(T,R_T)\right]}=\frac{\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}l_{\{\Theta_{\pi(j)}=\tilde{\theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}\prod_{j\in R_T}\frac{1}{d}\frac{1}{k_{\max}}\frac{1}{d}l_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}\right\}}{\sum_{\pi}\frac{1}{n!}\left\{\prod_{j\in T}\frac{1}{d}l_{\{\Theta_{\pi(j)}=\tilde{\theta}_j\wedge y_{\pi(j)}=\tilde{y}_j\}}\prod_{j\in R_T}\frac{1}{d}\frac{1}{k_{\max}}\frac{1}{d}l_{\{\Theta_{\pi(n)}(v'_n)=y_{\pi(n)}\}}\right\}}=\frac{c\sum_{\pi\in P}l_{\{\Theta_{\pi(n)}(v_n)=y_{\pi(n)}\}}}{c\sum_{\pi\in P}l_{\{\Theta_{\pi(n)}(v'_n)=y_{\pi(n)}\}}}=\frac{\sum_{j\in[n]\setminus T}\sum_{\pi\in P_j}l_{\{\tilde{\theta}_j(v_n)=\tilde{y}_j\}}}{\sum_{j\in[n]\setminus T}\sum_{\pi\in P_j}l_{\{\tilde{\theta}_j(v'_n)=\tilde{y}_j\}}}=\frac{\sum_{j\in[n]\setminus T}l_{\{\tilde{\theta}_j(v_n)=\tilde{y}_j\}}}{\sum_{j\in[n]\setminus T}l_{\{\tilde{\theta}_j(v'_n)=\tilde{y}_j\}}}=\frac{N_{R,T,R_T}}{N'_{R,T,R_T}}\tag{2}$$

其中，π 表示随机排列结果，P 表示排列所有用户发布真实值（$i\in T$）对应的 $l_{\{M_{\pi(i)}=\tilde{M}_i\wedge y_{\pi(i)}=\tilde{y}_i\}}\neq 0$ 的用户，这些用户可以被划分为 $n-|T|$ 个相同大小的子集合。在 $i\in[n]\setminus T$ 的基础上，P_i 是第 n 个用户的扰动值对应第 i 个位置的数据，剩余的 $[n-1]\setminus T$ 个用户不确定，因此，$P_i=l_{\{\tilde{\Theta}_i(v_n)=\tilde{y}_i\}}(n-1-|T|)!$。

N_{R,T,R_T} 和 N'_{R,T,R_T} 服从伯努利分布，即

$$N_{R,T,R_T}\sim\mathrm{Bin}\left(n-1,\frac{1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}\right)+1$$

$$N'_{R,T,R_T}\sim\mathrm{Bin}\left(n-1,\frac{1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}\right)$$

$$\Pr\left[\frac{N_{R,T_a,R_T}}{N'_{R,T_a,R_T}}\geqslant\mathrm{e}^{\varepsilon_c}\right]\leqslant\exp\left(-\frac{\theta}{3}\left(\mathrm{e}^{\frac{\varepsilon_c}{2}}-1-\frac{1}{\theta}\right)^2\right)+\exp\left(-\frac{\theta}{2}\left(1-\mathrm{e}^{\frac{-\varepsilon_c}{2}}\right)^2\right)$$

其中，$\theta=\mathrm{E}[N'_{R,T,R_T}]=\dfrac{n-1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}$。

令 $\theta=\dfrac{n-1}{\mathrm{e}^{\varepsilon_l}+\sum k_i-1}=\dfrac{14\ln\left(\frac{2}{\delta}\right)}{\varepsilon_c^2}$，可得 A 满足 $\left(\sqrt{\dfrac{14\ln\left(\frac{2}{\delta}\right)(e^{\varepsilon_l}+a_{\max}-1)}{n-1}},\delta\right)$-混洗差分隐私。证毕。

1.2.3 FRV-RPP 的可用性分析

定理 2 $\mathrm{E}[\tilde{f}_v]=f_v$ 成立，其中，f_v 表示 v 的真实频率，$\tilde{f}_v$ 表示 v 的估计频率。

证明

$$\mathrm{E}[\tilde{f}_v]=\mathrm{E}\left[\frac{d\left(\sum_{j\in[n]}l_{\{\tilde{\theta}_j=i\wedge\tilde{y}_j=v\}}-nq\right)}{n(p-q)}\right]=\frac{d\mathrm{E}\left[\sum_{j\in[n]}l_{\{\tilde{\Theta}_j=i\wedge y_j=v\}}\right]-nq}{n(p-q)}=\frac{d\left[nf_v\frac{p}{d}+n(1-f_v)\frac{q}{d}\right]-nq}{n(p-q)}=f_v$$

证毕。

定理 3 方差 $\mathrm{Var}[\tilde{f}_v]$ 满足

$$\mathrm{Var}[\tilde{f}_v]\approx\frac{d^2\left[\frac{\varepsilon_c^{\,2}(n-1)}{14\ln\left(\frac{2}{\delta}\right)}-1\right]}{n\left[\frac{\varepsilon_c^{\,2}(n-1)}{14\ln\left(\frac{2}{\delta}\right)}-k_{\max}\right]^2}$$

其中，$\tilde{f}_v$ 表示 v 的估计频率，$k_{\max}$ 表示用户数据属性取值最大。

证明

$$\mathrm{Var}\left[\tilde{f}_v\right]=\mathrm{Var}\left[\frac{d\left(\sum_{j\in[n]}l_{\{\tilde{\theta}_j=i\wedge\tilde{y}_j=v\}}-nq\right)}{n(p-q)}\right]=\mathrm{Var}\left[\frac{d\sum_{j\in[n]}l_{\{\tilde{\theta}_j=i\wedge\tilde{y}_j=v\}}}{n(p-q)}\right]=\frac{q(d-p)}{n(p-q)^2}+$$

$$\frac{f_v(d-p-q)}{n(p-q)} \approx \frac{\dfrac{d\varepsilon_c^2(n-1)}{14\ln\left(\dfrac{2}{\delta}\right)}-1}{n\left[\dfrac{\varepsilon_c^2(n-1)}{14\ln(\dfrac{2}{\delta})}-k_{\max}\right]^2}$$

证毕。

1.2.4　FRV-RPP 的效率和通信开销分析

在 FRV-RPP 算法中，每个用户仅调用一次 RR 机制，其时间复杂度和空间复杂度均为 $O(1)$。收集方接收到随机排列者发来的扰动数据，一共需要统计 $da_{\max}$ 个数据，由于每个用户发送一个值，随机排列者需要进行 n 次累加，该算法的时间复杂度为 $O(n)$，空间复杂度为 $O(da_{\max})$。用户发布数据的维度及扰动结果使一个用户和随机排列者之间的通信开销为 $O(\mathrm{lb}d+\mathrm{lb}a_{\max})$，随机排列者发送重新排序后的数据所需要的通信开销为 $O(n(\mathrm{lb}d+\mathrm{lb}a_{\max}))$。

该算法在满足 (ε_c,δ) -混洗差分隐私的基础上实现 K-Modes 聚类数据的发布，从数据发布结果精度和定理 3 来看，FRV-RPP 算法发生误差概率为 d，参加随机排列的用户数量较多。从通信开销方面来看，该算法使用一个随机排列者，用填补取值域方法统一维度，需要在加噪过程中获得所需的维度以及扰动信息。

综上分析，FRV-RPP 在时间复杂度和空间复杂度方面具有较好的性能，处于可接受的范围中。

1.3　CDC-KM 算法

1.3.1　CDC-KM 算法概述

CDC-KM 算法基于混洗差分隐私模型，输入用户的数据 t，数据大小 d，数据属性集 M，输出数据长度 k，隐私参数 α；输出干扰后获得的数据 t'，其长度为 k。该算法首先处理用户数据 t 得到 $\hat{t}$，其长度为 l，$|\hat{t}|=l$；其次从所有的用户数据属性集中随机选择一个 t'，选中 t' 的概率为

$$\exp\frac{\left(\dfrac{-\alpha\,\mathrm{dist}(t,t')}{2}\right)}{\Omega} \tag{3}$$

其中，Ω 是概率泛化因子，从算法 2 中的定义式可以看出，概率泛化因子由两部分构成，前一部分是相似度为 0 的子空间，后一部分是相似度大于 0 的子空间。具体过程如算法 2 所示。

算法 2　CDC-KM 算法

输入　用户数据 t，属性维度 d，$d \leqslant l$，属性集 $M=\{m_1,m_2,m_3,\cdots,m_d\}$，输出数据长度 k，隐私参数 α

输出　收集并扰动的用户数据 t'

1) $t'=\varnothing$
2) 等长处理 t
3) $\hat{t}=t, M=\{m_1,m_2,m_3,\cdots,m_d\}\cup\{m_{d+1},\cdots,m_{d+l}\}$ //原有的属性域 $m_1,m_2,\cdots,m_d$，填补的噪声项为 $m_{d+1},\cdots,m_{d+l}$
4) for $i=0;i<l-d;i++$ do
5)　　$\hat{t}=t\cup m_{d+i+1}$
6) end for
7) 随机化处理 $\hat{t}$
8) $\Omega=\mathrm{e}^{\frac{-\alpha}{2}0}\mathrm{C}_d^k+\sum_{\mathrm{inter}=1}^{k}\left(\mathrm{e}^{\frac{-\alpha}{2}(k-\mathrm{inter})}\left(\mathrm{C}_m^{\mathrm{inter}}\mathrm{C}_d^{k-\mathrm{inter}}\right)\right)$
9) $r=\mathrm{uniform\_random}(0.0,1.0)$
10) $\mathrm{inter}=0$
11) $p=\mathrm{e}^{\frac{-\alpha}{2}0}\dfrac{\mathrm{C}_d^k}{\Omega}$
12) while $p<r$ do
13)　　$\mathrm{inter}+=1$
14)　　$p=p+\dfrac{\mathrm{e}^{\frac{-\alpha}{2}(k-\mathrm{inter})}\mathrm{C}_m^{\mathrm{inter}}\mathrm{C}_{d-1}^{k-1-\mathrm{inter}}}{\Omega}$
15) end while
16) $t'=t'\cup\mathrm{sample}(\hat{t},\mathrm{inter})$ //从 $\hat{t}$ 中以等概率不放回的方式随机选取 inter 项，目的是保留真实数据
17) $t'=t'\cup\mathrm{sample}(M-\hat{t},k-\mathrm{inter})$ //从 $M-\hat{t}$ 中以等概率不放回的方式随机选取 $k-\mathrm{inter}$ 项，目的是引入噪声数据
18) return t'

在算法 2 中，第 3)步～第 6)步是对用户数据做等长处理，使所有的用户数据 t 的长度为 l；第 9)步基于随机函数生成概率 r；第 12)步～第 15)步确定子样本空间；第 16)步～第 17)步用 sample 抽样函数从 $\hat{t}$ 中随机选取 inter 项，从剩余数据中选取 $k-\mathrm{inter}$ 项，并拼接。CDC-KM 算法的时间复杂度是 $O(d)$，与数据属性大小成正比。

1.3.2　CDC-KM 算法的隐私性分析

定理 4　基于指数机制的 CDC-KM 算法满足混洗差分隐私约束。

证明　CDC-KM 算法本质上是指数机制的具体实现，对初始数据 t，以 $p=\dfrac{\mathrm{score}(t')}{\sum_{s\in\mathrm{CandI}}\mathrm{score}(s)}$ 的概

率随机选取一个属性集 t'。令该算法中的随机机制为 $\Re$，由 t 得到 t' 的概率为 $\Pr[\Re(t)=t']=\dfrac{\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t,t')}{2}}}{\sum\limits_{z\in \mathrm{CandI}}\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t,z)}{2}}}$。

下面证明 $\dfrac{\Pr[\Re(t_1)=t']}{\Pr[\Re(t_2)=t']}\leqslant \mathrm{e}^{\frac{\alpha \mathrm{dis}(t_1,t_2)}{2}}$ 成立，即证明 CDC-KM 算法满足混洗差分隐私约束。

$$\frac{\Pr[\Re(t_1)=t']}{\Pr[\Re(t_2)=t']}\leqslant \frac{\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_1,t')}{2}}}{\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_2,t')}{2}}}\frac{\sum\limits_{z\in \mathrm{CandI}}\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_2,z)}{2}}}{\sum\limits_{z\in \mathrm{CandI}}\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_2,z)}{2}}}=$$
$$\mathrm{e}^{\frac{\alpha \mathrm{dis}(t_2,t')-\mathrm{dis}(t_1,t')}{2}}\frac{\sum\limits_{z\in \mathrm{CandI}}\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_2,z)+\alpha \mathrm{dis}(t_1,z)-\alpha \mathrm{dis}(t_1,z)}{2}}}{\sum\limits_{z\in \mathrm{CandI}}\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_1,z)}{2}}} \tag{4}$$

由于 $\alpha \mathrm{dis}(t_1,t_2)$ 与 z 无关，可以在式(4)中将其提到求和运算之前，根据 dis 求距离满足 $\mathrm{dis}(t_2,t')-\mathrm{dis}(t_1,t')\leqslant \mathrm{dis}(t_1,t_2)$ 可得

$$\frac{\Pr[\Re(t_1)=t']}{\Pr[\Re(t_2)=t']}\leqslant \mathrm{e}^{\frac{\alpha \mathrm{dis}(t_2,t_1)}{2}}\frac{\mathrm{e}^{\frac{\alpha \mathrm{dis}(t_2,t_1)}{2}}\sum\limits_{z\in \mathrm{CandI}}\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_1,z)}{2}}}{\sum\limits_{z\in \mathrm{CandI}}\mathrm{e}^{\frac{-\alpha \mathrm{dis}(t_1,z)}{2}}}\leqslant$$
$$\mathrm{e}^{\frac{\alpha \mathrm{dis}(t_2,t_1)}{2}}\mathrm{e}^{\frac{\alpha \mathrm{dis}(t_2,t_1)}{2}}=\mathrm{e}^{\alpha \mathrm{dis}(t_2,t_1)}$$

证毕。

1.4 求解质心方法

1.4.1 迭代法计算质心

数据中心收集到用户数据后，从属性集中任意选取 k 个 d 维属性元组作为初态质心发送给用户，随后根据用户数据和返回的簇信息计算出新的质心集 $V=\{v_1,v_2,v_3,\cdots,v_k\}$，具体过程如算法 3 所示。

算法 3 迭代式计算质心

输入 用户簇 $U=\{u_1,u_2,u_3,\cdots,u_k\}$，每个簇中的数据 $B'=\{b'_1,b'_2,b'_3,\cdots,b'_d\}$，采样时的扰动参数 p 和 q

输出 新的质心集 $V=\{v_1,v_2,v_3,\cdots,v_k\}$

1) for U from $r=1$ to k do
2) from b'_1 to b'_d do
3) 计算每个比特位获得 $S=\{s_1,s_2,s_3,\cdots,s_l\}$
4) for $i=0$ to $l-1$ do
5) $t'_i=\dfrac{s_i-|u_r|'q}{|u_r|'(p-q)}$
6) $a=\max(t')$ 对应的属性
7) end for
8) $v_r=\{a_1,a_2,a_3,\cdots,a_d\}$
9) $V=\{v_1,v_2,v_3,\cdots,v_k\}$
10) end for

根据收集到的每个用户的 u_i，将用户划分为 k 个簇 $U=\{u_1,u_2,u_3,\cdots,u_k\}$，用户采样数据，因此，簇中每个属性对应的用户数为 $\dfrac{|u_r|}{d}$。然后，以簇 u_r 中属性 a_j 为例，统计该属性对应的扰动数据 b'_j 的每位得到 $S=\{s_1,s_2,s_3,\cdots,s_l\}(0\leqslant l\leqslant|b'_j|)$，$s_l$ 表示 b'_j 中第 l 位是 1 的个数，结合采样时的扰动参数计算出 a_j 的所有属性对应的估计频率 t'。选取每个属性中频率最高的数据值，将其作为该簇的质心，得到新的质心集 $V=\{v_1,v_2,v_3,\cdots,v_k\}$。将新的质心集发送给用户，并不断从用户处搜集簇信息，重复算法 3，直到每个簇的质心在相邻两次迭代中不变。

1.4.2 分析求解质心方法中的隐私性和可用性

定理 5 求解质心中的计算得到的估计频率 $t'_i=\dfrac{s_i-|u_r|'q}{|u_r|'(p-q)}$ 满足无偏性。

证明 在更新过程中，因没有收集用户的真实数据，不能计算出所有属性的真实频率 t，只能得到估计频率 t'。下面证明 t' 满足无偏性。

这里，令 t 是簇 u_r 中一个属性 a 的真实频率，而 t' 是这个属性的估计频率，g 是 a 的真实频率，g' 是 a 的估计频率。a' 是 a 的位，s 是扰动数据中 a' 是 1 的数目。由于不能获得 a 的真实频率 g，但要求出 t'，就必须计算出 a 的估计频率 g'。用户以 2 个不同概率响应每位，因此，$|u_r|'$ 个用户对 a' 的响应结果构成了 $|u_r|'$ 个 0/1 序列，并且该序列满足二项分布。根据二项分布的性质，构造其似然函数为

$$L(g)=\left[\frac{g}{|u_r|'}p+\left(1-\frac{g}{|u_r|'}\right)q\right]^{s}\cdot$$
$$\left[\frac{g}{|u_r|'}(1-p)+\left(1-\frac{g}{|u_r|'}\right)(1-q)\right]^{|u_r|'-s} \tag{5}$$

$$g'=\frac{s-|u_r|'q}{p-q} \tag{6}$$

接下来证明 g' 的无偏性，如式(7)所示

$$\mathrm{E}[g'] = \mathrm{E}\left[\frac{s-|u_r|'q}{p-q}\right] = \frac{(gp+(|u_r|'-g)q)-|u_r|'q}{p-q} = g \tag{7}$$

由此可得 g' 满足无偏性，可求出无偏估计频率 $t' = \dfrac{s-|u_r|'q}{|u_r|'(p-q)}$。证毕。

定理 6　采样用户数据可以有效提高扰动数据的可行性。

证明　假设 n 名用户的数据有 d 维属性，隐私预算为 ε，用户数据中的某一属性的估计频率为 g'，f 是该维属性对应的位，s 是扰动数据中 f 为 1 的数目，真实频率为 t。如果不对用户数据进行采样，并且该属性获得全部隐私预算 ε，那么可以得到 g' 的方差为

$$\mathrm{Var}[g'] = \mathrm{Var}\left[\frac{s-nq}{p-q}\right] = \frac{ntp(1-p)+n(1-t)q(1-q)}{(p-q)^2} \tag{8}$$

将 $p=\dfrac{1}{2}, q=\dfrac{1}{e^{\varepsilon}-1}$ 代入得到 g' 的方差，即

$$\mathrm{Var}[g'] = \frac{n4\mathrm{e}^{\varepsilon}}{(\mathrm{e}^{\varepsilon}-1)^2} + nt \tag{9}$$

为了方便计算，可以省略 nt，它是属性的真实频率，是一个常数。而属性对应的用户个数不一致，不能直接比较方差，因此将式(9)变换为

$$\mathrm{Var}\left[\frac{g'}{n}\right] = \frac{4\mathrm{e}^{\varepsilon}}{n(\mathrm{e}^{\varepsilon}-1)^2} \tag{10}$$

下面比较随机采样数据和分割隐私预算这 2 种方法。随机采样 d 维属性，使属性对应的用户数目为 $\dfrac{n}{d}$，其方差为

$$v_1 = \frac{4d\mathrm{e}^{\varepsilon}}{n(\mathrm{e}^{\varepsilon}-1)^2} \tag{11}$$

分割隐私预算让每个属性获得的隐私参数为 $\dfrac{\varepsilon}{d}$，其方差为

$$v_2 = \frac{4\mathrm{e}^{\frac{\varepsilon}{d}}}{n(\mathrm{e}^{\frac{\varepsilon}{d}}-1)^2} \tag{12}$$

通过比较 v_1 和 v_2 大小来证明随机采样和分割隐私参数之间的可用性，即

$$v_1 - v_2 = \frac{4}{n}\left(\frac{\mathrm{e}^{\frac{\varepsilon}{d}}}{\left(\mathrm{e}^{\frac{\varepsilon}{d}}-1\right)^2} - \frac{d\mathrm{e}^{\varepsilon}}{(\mathrm{e}^{\varepsilon}-1)^2}\right) = \frac{4\mathrm{e}^{\frac{\varepsilon}{d}}}{n(\mathrm{e}^{\frac{\varepsilon}{d}}-1)^2(\mathrm{e}^{\varepsilon}-1)^2}\left((\mathrm{e}^{\varepsilon}-1)^2 - d\mathrm{e}^{\varepsilon-\frac{\varepsilon}{d}}(\mathrm{e}^{\frac{\varepsilon}{d}}-1)^2\right) \tag{13}$$

由 $\varepsilon > 0$ 可得

$$\frac{4\mathrm{e}^{\frac{\varepsilon}{d}}}{n\left(\mathrm{e}^{\frac{\varepsilon}{d}}-1\right)^2(\mathrm{e}^{\varepsilon}-1)^2} > 0 \tag{14}$$

令 $y = \mathrm{e}^{\frac{\varepsilon}{d}}$，并代入式(14)，由于式(14)的不等号左侧大于 0，舍去这部分以方便计算，化简为

$$v_1 - v_2 = (y^d-1)^2 - dy^{d-1}(y-1)^2 = (y-1)^2\left[(y^{d-1}+y^{d-2}+\cdots+1)^2 - dy^{d-1}\right] > 0 \tag{15}$$

由式(15)可得 $v_1 > v_2$，证毕。

2　具体实施

为了验证本文提出的 SDPKM 方法的有效性，本节在实验环境中通过对比已有算法—LDPKV 方法[27]、KM 算法[25]、DPLM 算法[26]、LDPKM 方法[9]和 AdaPub 机制[28]，评价 SDPKM 在收集和发布数据质量方面的性能。LDPKV 方法是基于本地化差分隐私的键值数据收集方法；KM 算法是未使用隐私保护的初始聚类收集和发布算法；DPLM 算法是受中心化差分隐私保护的聚类收集和发布算法；LDPKM 方法为本地化差分隐私聚类收集和发布方法；AdaPub 机制研究聚类数据流的隐私保护发布问题。

本文实验使用 3 个数据集，包括 Adult 数据集[29]、IPUMS 数据集[14]和 Kosarak 数据集[15]。其中，Adult 是隐私保护研究领域中常用的数据集，属于加州大学欧文分校（UCI）数据库，包含 48 842 条数据记录，14 个属性，本文从中选取 3 万条数据记录，每条记录有 5 个属性；IPUMS 包含约 25 万条数据，61 个属性，具有平衡性和代表性，可以反映总体的真实情况，本文从中选取 3 万条数据，4 个属性；Kosarak 是网站点击流量的数据集，用于评估推荐系统性能，判断用户是否对商品感兴

趣，包含约 100 万条记录，本文从中选取 3 万条数据，2 个属性，作为用户数据。3 个数据集的详细信息如表 1 所示。

表 1 数据集详细信息

数据集	用户数	属性名称	属性域大小
Adult	30 000	Workclass	7
		Education	16
		Relationship	6
		Sex	2
		Race	5
IPUMS	30 000	School	3
		Famsize	15
		Sex	2
		Race	8
Kosarak	30 625	User	6
		Item	10

结合上述数据集，实验采用准确率 A[30]和熵值 E[9]作为 K-Modes 聚类质量评价标准，如式(16)和式(17)所示。A 和 E 是聚类正确性比值和质量优劣中常用的度量指标，$A\in[0,1]$，A 越大，聚类质量越高；熵值则相反，E 越小，聚类质量越高。在噪声的干扰下，聚类质量越好，聚类的隐私安全性越能够得到保证。

$$A=\frac{\sum_{j=1}^{k}h_j}{N} \tag{16}$$

$$E=\sum_{j=1}^{k}\frac{c_j}{N}\sum_{i=1}^{k}-\frac{|c_j\bigcap t_i|}{|c_j|}\ln\left(\frac{|c_j\bigcap t_i|}{|c_j|}\right) \tag{17}$$

其中，k 是聚类簇数，h_j 是隐私保护算法下得到的聚类簇中正确聚类的数据数目，N 是数据集大小，t_i 是无隐私保护下的聚类簇，c_j 是隐私保护算法下的聚类簇。

本文分别在 Adult、IPUMS 和 Kosarak 这 3 个数据集上，通过确定 δ、ε=0.1～1.0，运行 LDPKV 方法、KM 算法、DPLM 算法、LDPKM 方法、AdaPub 机制以及本文方法 SDPKM，比较准确率及熵值。聚类簇数 k 取 3，每个实验进行 100 次，结果取平均值。

2.1 数据集样本数量对准确率和熵值的影响

当 ε 固定为不同值时，Adult、IPUMS 和 Kosarak 数据集样本数量 N 对 SDPKM 方法准确率的影响如图 1 所示。由图 1 可以发现，当 ε 固定为不同值时，对于不同数据集，数据集样本数量越多，SDPKM 方法的准确率越高。

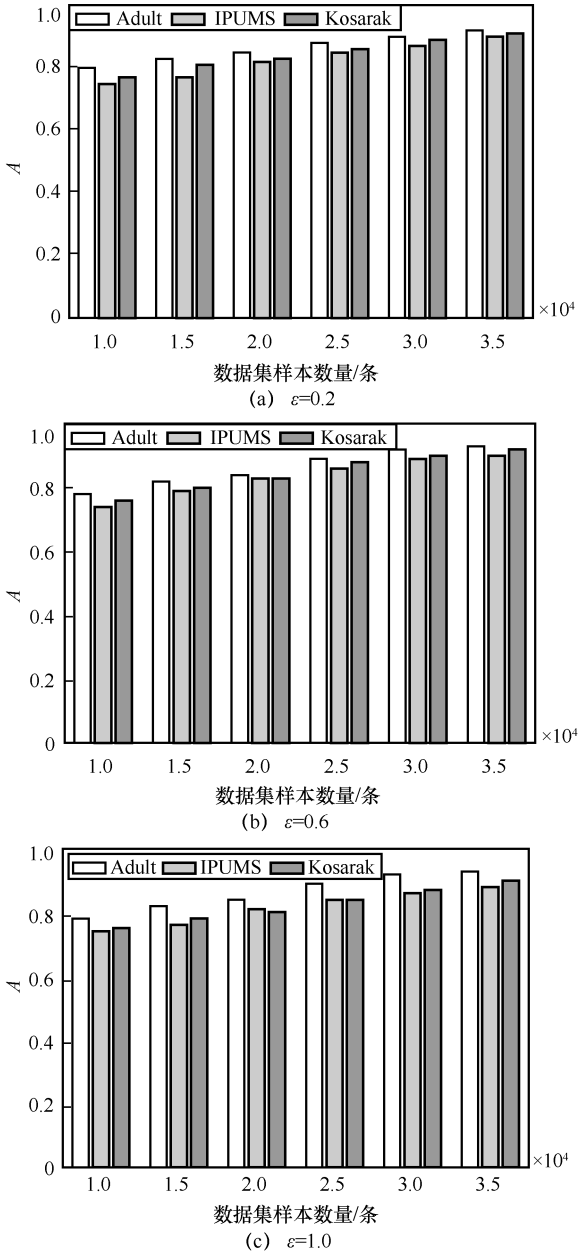


图 1 数据集样本数量对 SDPKM 方法准确率的影响

当 ε 固定为不同值时，Adult、IPUMS 和 Kosarak 数据集样本数量对 SDPKM 方法熵值的

影响如图 2 所示。从图 2 可以看出，当ε固定为不同值时，对于不同数据集，数据集样本数量越多，SDPKM 方法的熵值越小。其原因是在混洗差分隐私模型中，每个用户以一定概率上传真实值，再由大数定律可知，在相同条件下，实验数据越多，算法结果的可信性就越大，数据量越少则可信性就越小。这就说明了数据集样本数量越多，对相应随机性的影响越小，聚类数据质量就越高。

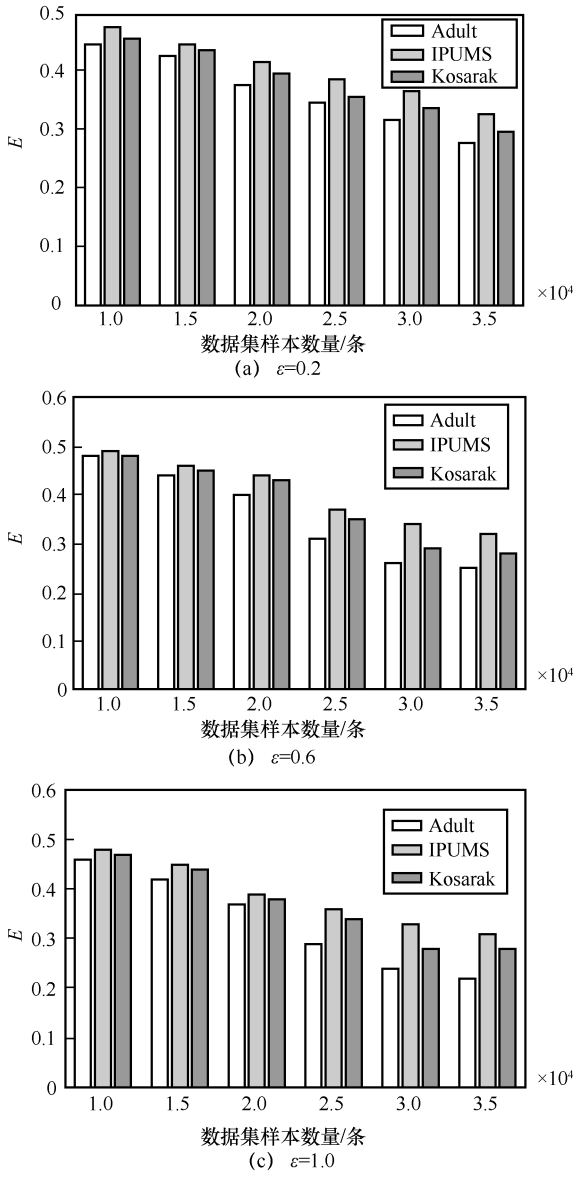


图 2　数据集样本数量对 SDPKM 方法熵值的影响

2.2　算法性能对比

当δ固定为不同值时，不同算法在 Adult 数据集上的准确率对比如图 3 所示。从图 3 可知，当δ固定为不同值时，随着ε的增大，SDPKM 方法的准确率提高。具体地，如图 3(b)所示，当$\delta=10^{-6}$时，SDPKM 的准确率明显优于其他 5 种算法，当ε从 0.1 增加到 0.5 时，SDPKM 的准确率在 Adult 数据集上比 KM 高了 2 个数量级，其原因是 SDPKM 利用填补取值域随机排列发布算法对加噪后的数据进行刷新重新排列，可以更好地保护隐私数据。

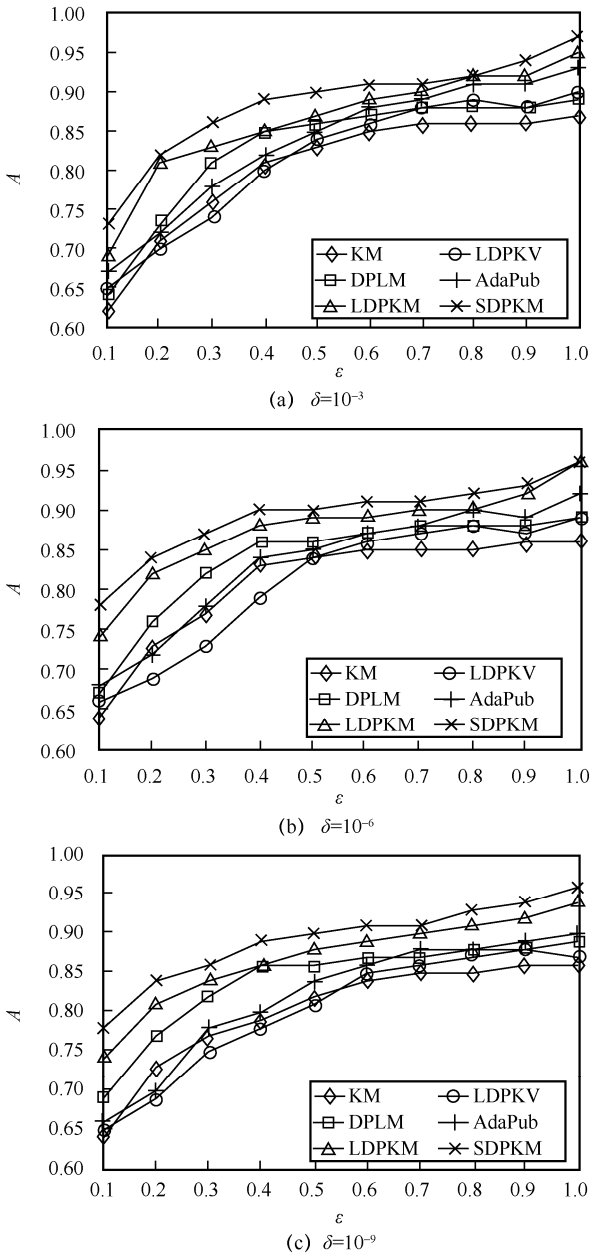
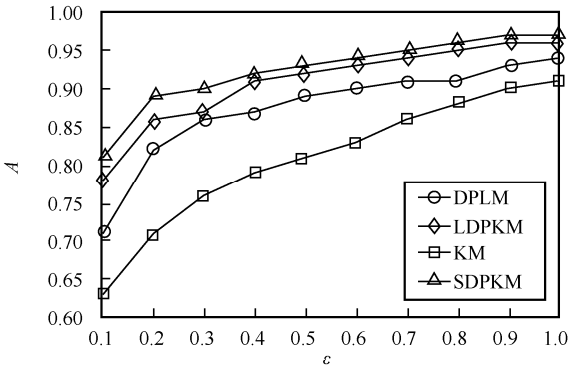


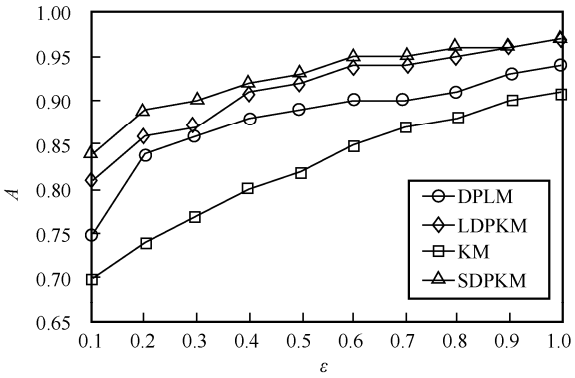
图 3　不同算法在 Adult 数据集上的准确率对比

当δ固定为不同值时，不同算法在 IPUMS 数据集上的准确率对比如图 4 所示。从图 4 可知，当

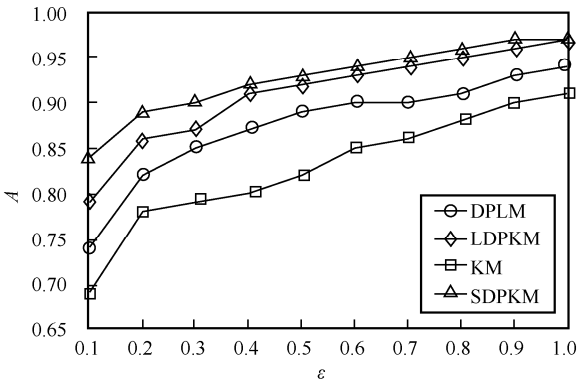
δ 固定为不同值时，随着 ε 的增大，不同算法的准确率都提高。具体地，如图 4(c)所示，当 $\delta=10^{-9}$、$\varepsilon=0.4$ 时，SDPKM 准确率是 KM 准确率的 2 倍，达到了 DPLM 准确率的 96%，与 LDPKM 准确率值相当，其原因是 KM 没有对数据进行隐私处理，存在极大的泄露敏感数据风险，因此，KM 的准确率较低。



(a) $\delta=10^{-3}$



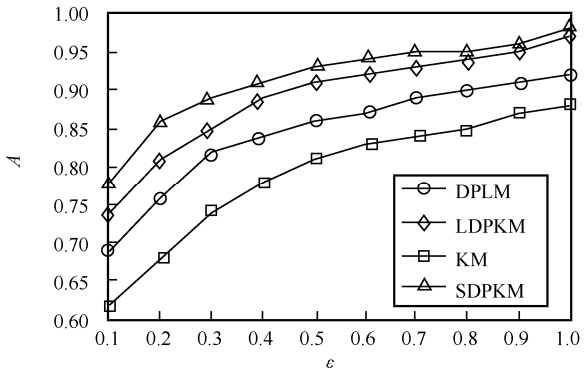
(b) $\delta=10^{-6}$



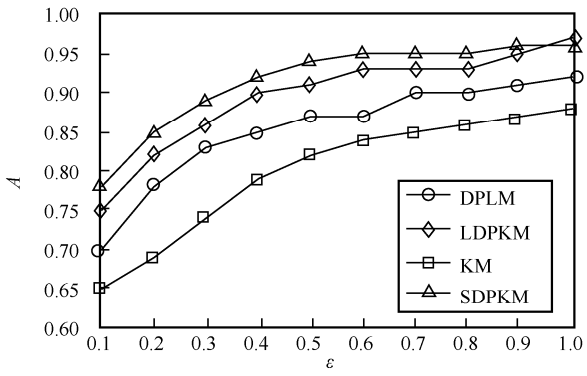
(c) $\delta=10^{-9}$

图 4 不同算法在 IPUMS 数据集上的准确率对比

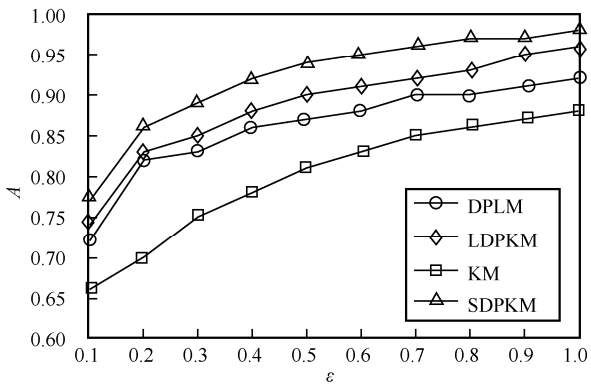
当 δ 固定为不同值时，不同算法在 Kosarak 数据集上的准确率对比如图 5 所示。从图 5 可知，当 δ 固定为不同值时，随着 ε 的增大，4 种算法的准确率都提高，由图 5(b)可知，当 $\delta=10^{-6}$、$\varepsilon=0.5$ 时，SDPKM 方法的准确率是 DPLM 方法的 90%，其原因是 DPLM 使用传统的差分隐私对数据进行加噪，一定程度上保护了数据，但相比于本文所提 SDPLKM 使用的混洗差分隐私模型，依然存在攻击者识别出目标用户，进而造成数据泄露的风险。



(a) $\delta=10^{-3}$



(b) $\delta=10^{-6}$

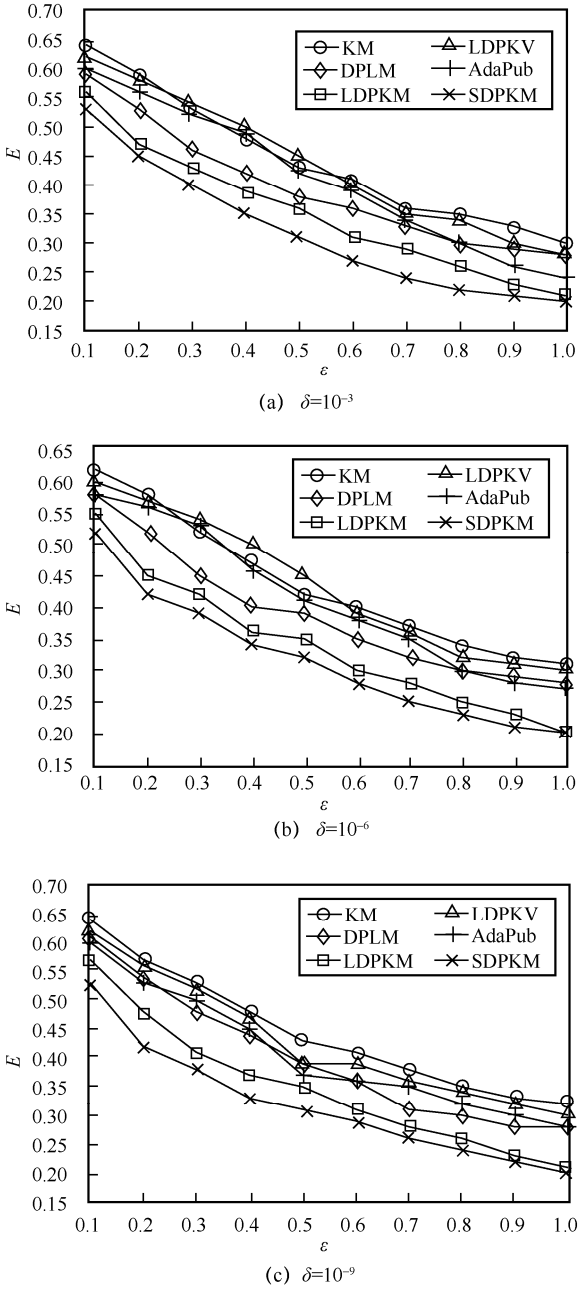


(c) $\delta=10^{-9}$

图 5 不同算法在 Kosarak 数据集上的准确率对比

当 δ 固定为不同值时，不同算法在 Adult 数据集上的熵值对比如图 6 所示。从图 6 可知，随着 ε 的增大，熵值呈现出下降的趋势，其原因是噪声的添

加量取决于ε，ε越大，噪声添加量越少，聚类结果质量越高；反之，ε越小，噪声添加量越多，聚类结果质量越低。



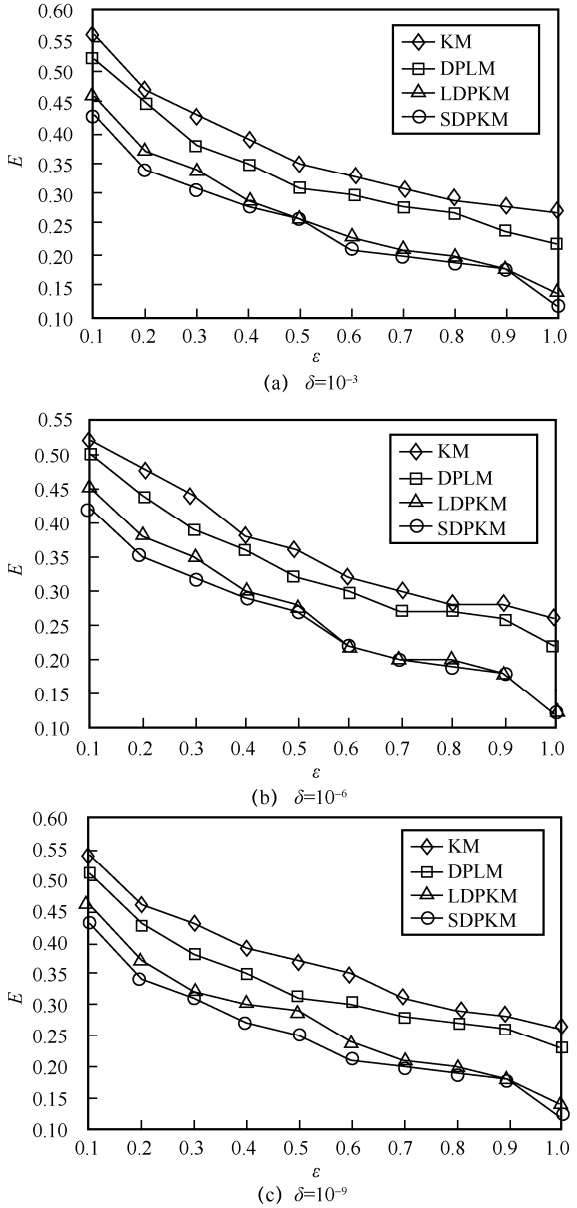
(a) $\delta=10^{-3}$

(b) $\delta=10^{-6}$

(c) $\delta=10^{-9}$

图 6　不同算法在 Adult 数据集上的熵值对比

当δ固定为不同值时，不同算法在 IPUMS 数据集上的熵值对比如图 7 所示。从图 7 可知，随着ε逐渐增大，熵值呈现出下降的趋势。当$\varepsilon \geqslant 0.6$ 时，SDPKM 和 LDPKM 的熵值相当，并且优于 KM 和 DPLM 算法，这得益于可信第三方的数据收集。4 种算法中，SDPKM 熵值最小，即聚类效果最好，在一些情况下，SDPKM 的熵值能够比 KM 小一个数量级。



(a) $\delta=10^{-3}$

(b) $\delta=10^{-6}$

(c) $\delta=10^{-9}$

图 7　不同算法在 IPUMS 数据集上的熵值对比

当δ固定为不同值时，不同算法在 Kosarak 数据集上的熵值对比如图 8 所示。从图 8 可知，随着ε逐渐增大，熵值逐渐减小，即熵值与隐私预算成反比。

综上所述，SDPKM 的准确率和熵值性能均优于其他 5 种算法，在ε变化的整个区间中，SDPKM 的准确率甚至达到了 KM 算法的 82%。证明了本文方法的有效性，即对用户数据进行收集扰动、迭代计算质心以及填充取值域随机排列发布，保证了用户数据的安全性，避免了用户隐私泄露的风险，提高了对用户数据的保护程度，同时保证了数据质量。

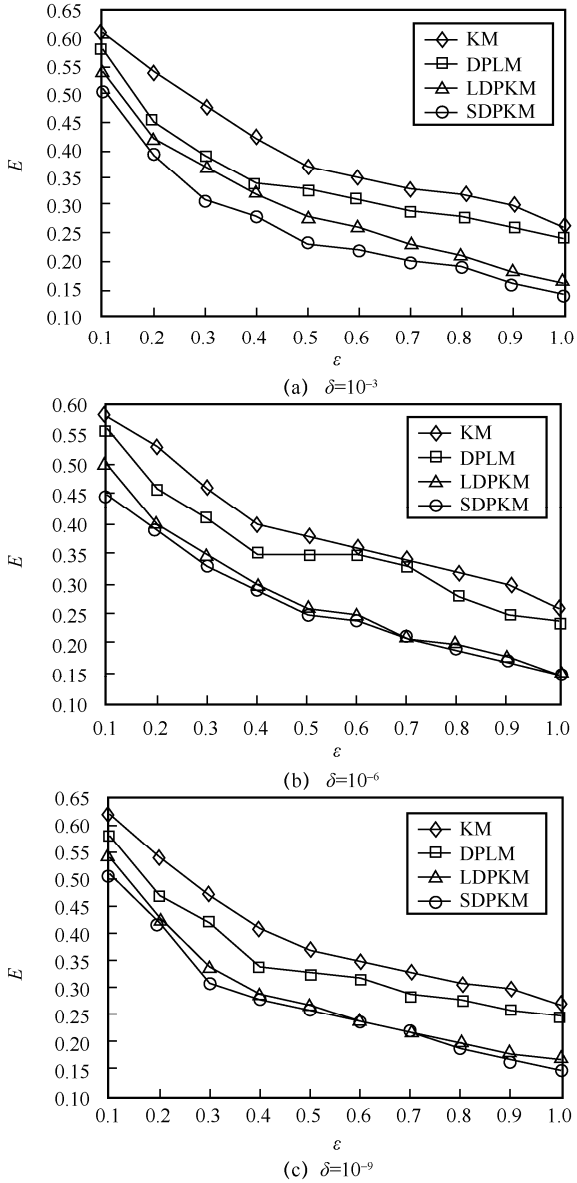


图 8　不同算法在 Kosarak 数据集上的熵值对比

3　结束语

本文针对 K-Modes 聚类数据中的用户敏感信息的隐私保护问题，基于混洗差分隐私模型，提出了一种 K-Modes 聚类数据隐私保护方法。该方法包含 K-Modes 聚类数据收集扰动算法、循环迭代计算质心完成聚类以及取值域填补随机排列发布算法，保证整个过程中恶意攻击者无法获取用户的真实信息。通过 3 个真实数据集与现有的 LDPKV、KM、DPLM、LDPKM 算法和 AdaPub 机制进行准确率和熵值的对比分析，实验结果表明，所提 SDPKM 方法在满足混洗差分隐私的基础上，有效地保证了聚类数据的质量与隐私安全性。今后的研究方向如下：1) 如何将本文方法应用到轨迹数据在收集与发布过程中的隐私保护；2) 如何将用户数据进行敏感度分类，实施不同程度的隐私保护，满足用户对数据个性化隐私保护的需求；3) 如何将该模型思想应用到实际案例中，如社区规划等；4) 如何在初始属性域上扰动数据并分析隐私效果。

参考文献：

[1] XU S Z, CHENG X, SU S, et al. Differentially private frequent sequence mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 2910-2926.

[2] WANG N, XIAO X K, YANG Y, et al. PrivSuper: a superset-first approach to frequent itemset mining under differential privacy[C]//Proceedings of 2017 IEEE 33rd International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2017: 809-820.

[3] 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.

[4] WANG T H, LI N H, JHA S. Locally differentially private frequent itemset mining[C]//Proceedings of 2018 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2018: 127-143.

[5] BALLE B, BELL J, GASCÓN A, et al. The privacy blanket of the shuffle model[C]//Proceedings of Annual International Cryptology Conference. Cham: Springer, 2019: 638-667.

[6] ERLINGSSON Ú, FELDMAN V, MIRONOV I, et al. Amplification by shuffling: from local to central differential privacy via anonymity[C]//Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms. New York: ACM Press, 2019: 2468-2479.

[7] WANG T, DING B, XU M, et al. Improving utility and security of the shuffler-based differential privacy[J]. arXiv Preprint, arXiv: 1908.11515, 2019.

[8] LIU Y F, WANG N, WANG Z G, et al. Collecting and analyzing multidimensional categorical data under shuffled differential privacy[J]. Journal of Software, 2022, 33(3): 1093-1110.

[9] ZHANG S B, YUAN L J, MAO X J, et al. Privacy protection method for K-Modes clustering data with local differential privacy[J]. Acta Electronica Sinica, 2022, 50(9): 2181-2188.

[10] SASSI D B, FRINI A, CHAIEB M, et al. A rough set-based competitive intelligence approach for anticipating competitor's action[J]. Expert Systems With Applications, 2022, 204: 117523.

[11] 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.

[12] 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.

[13] DUAN Y Q, YUAN H L, LAI C S, et al. Fusing local and global information for one-step multi-view subspace clustering[J]. Applied Sciences, 2022, 12(10): 5094.

[14] ZHANG X J, XU Y X, XIA Q R. Histogram publication under shuffled differential privacy[J]. Journal of Software, 2022, 33(6): 2348-2363.

[15] BALCER V, CHEU A. Separating local & shuffled differential privacy

via histograms[J]. arXiv Preprint, arXiv: 1911.06879, 2019.

[16] 方晨, 郭渊博, 王娜, 等. 基于生成对抗网络的差分隐私数据发布方法[J]. 电子学报, 2020, 48(10): 1983-1992.
FANG C, GUO Y B, WANG N, et al. Differential private data publishing method based on generative adversarial network[J]. Acta Electronica Sinica, 2020, 48(10): 1983-1992.

[17] LIU P J, LI H Y, WANG T Y, et al. Multi-stage method for online vertical data partitioning based on spectral clustering[J]. Journal of Software, 2022, 34(6): 2804-2832.

[18] ZHANG X J, ZHANG J W, HUANG C, et al. Verifiable encrypted medical data aggregation and statistical analysis scheme[J]. Journal of Software, 2022, 33(11): 4285-4304.

[19] LIANG W J, CHEN H, ZHAO S Y, et al. A differentially private scheme for top-k frequent itemsets mining over data streams[J]. Chinese Journal of Computers, 2021, 44(4): 741-760.

[20] WANG J Y, LIU C, FU X C, et al. Crucial patterns mining with differential privacy over data streams[J]. Journal of Software, 2019, 30(3): 648-666.

[21] 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.

[22] TIAN F, WU Z Q, LU L F, et al. Personalized differential privacy protection mechanism for trajectory data publishing[J]. Chinese Journal of Computer, 2021, 44(4): 709-723.

[23] 张东月, 倪巍伟, 张森, 等. 一种基于本地化差分隐私的网格聚类方法[J]. 计算机学报, 2023, 46(2): 422-435.
ZHANG D Y, NI W W, ZHANG S, et al. A local differential privacy based privacy-preserving grid clustering method[J]. Chinese Journal of Computers, 2023, 46(2): 422-435.

[24] 陆佳炜, 吴涵, 张元鸣, 等. 融合功能语义关联计算与密度峰值检测的 Mashup 服务聚类方法[J]. 计算机学报, 2021, 44(7): 1501-1515.
LU J W, WU H, ZHANG Y M, et al. Mashup service clustering method via integrating functional semantic association calculation and density peak detection[J]. Chinese Journal of Computers, 2021, 44(7): 1501-1515.

[25] LU S Y, WANG G H, QIU Z H, et al. Differentially private algorithm for graphical bandits[J]. Journal of Software, 2022, 33(9). 3223-3235.

[26] BALAKRISHNAN S, SURESH KUMAR K, BALASUBRAMANIAN M, et al. Opinion mining for breast cancer disease using apriori and k-modes clustering algorithm[C]//Rising Threats in Expert Applications and Solutions. Berlin: Springer, 2022: 43-51.

[27] 张啸剑, 付楠, 孟小峰. 基于本地差分隐私的键-值数据精确收集方法[J]. 计算机学报, 2020, 43(8): 1479-1492.
ZHANG X J, FU N, MENG X F. Key-value data accurate collection under local differential privacy[J]. Chinese Journal of Computers, 2020, 43(8): 1479-1492.

[28] TENG W, YANG X Y, REN X B, et al. Data-adaptive privacy-preserving mechanism for data stream publishing in real-time[J]. 2021, doi: 10.1360/SSI-2020-0076.

[29] OUYANG J, YIN J, XIAO Z H, et al. Transaction data collection for itemset mining under local differential privacy[J]. Journal of Software, 2021, 32(11): 3541-3562.

[30] MANCHINI C, OSPINA R, LEIVA V, et al. A new approach to data differential privacy based on regression models under heteroscedasticity with applications to machine learning repository data[J]. Information Sciences, 2023, 627: 280-300.

[作者简介]

蒋伟进（1964− ），男，湖南益阳人，博士，湖南工商大学教授、硕士生导师，主要研究方向为信息安全、网络安全和群智感知。

陈艺琳（2000− ），女，河南许昌人，湖南工商大学硕士生，主要研究方向为信息安全和差分隐私。

韩裕清（2000− ），男，湖南长沙人，湖南工商大学硕士生，主要研究方向为信息安全和联邦学习。

吴玉庭（1998− ），女，湖南益阳人，湖南工商大学硕士生，主要研究方向为信息安全和群智感知。

周为（2000− ），男，湖南益阳人，湖南工商大学硕士生，主要研究方向为信息安全和群智感知。

王海娟（2000− ），女，江西九江人，湖南工商大学硕士生，主要研究方向为差分隐私和群智感知。