一种基于本地化差分隐私的网格聚类方法

张东月 倪巍伟 张 森 付 楠 侯立贺

(东南大学计算机科学与工程学院 南京 211189)

(东南大学计算机网络和信息集成教育部重点实验室 南京 211189)

摘 要 随着移动互联网应用的不断深入,产生了大量个体数据,采集分布在不同终端上的数据进行聚类可以发现人群行为模式,支撑应用服务的深入开展.然而这些数据往往包含个体敏感信息,在缺少可信数据采集者的情况下,直接采集数据进行聚类存在泄露个体数据隐私的风险.近年来,本地化差分隐私(Local Differential Privacy,LDP)以其严谨的数学理论基础得到隐私保护领域研究者的持续关注.现有基于 LDP 的聚类研究多数采用基于划分的聚类方法,存在仅适用凸状分布数据以及聚类质量损失较大问题.针对该问题,聚焦网格聚类,提出基于 LDP 的隐私保护网格聚类方法.首先,设计网格划分评估指标,通过调节网格划分粒度调控网格密度估算误差和簇边缘信息损失,指导网格结构选取;然后,在服务器与终端间构建循环反馈机制,利用数据分布信息迭代优化扰动粒度,降低差分噪声注入量,在保护终端数据隐私安全的前提下,提升网格密度估算精度;最后,在服务器端,提出基于网格结构的自适应网格聚合方法,提升隐私保护聚类质量.理论分析和实验结果表明,所提方法在兼顾各终端个体数据隐私的同时,对不同分布数据有良好的聚类效果.

关键词 隐私保护;本地化差分隐私;网格聚类;网格划分评估指标;循环反馈机制

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2023.00422

A Local Differential Privacy Based Privacy-Preserving Grid Clustering Method

ZHANG Dong-Yue NI Wei-Wei ZHANG Sen FU Nan HOU Li-He

(*Department of Computer Science and Engineering, Southeast University, Nanjing* 211189)

(*Key Laboratory of Computer Network and Information Integration in Southeast University, Ministry of Education, Nanjing* 211189)

Abstract With the continuous deepening of mobile Internet applications, a large number of individual data have been produced. Collecting data distributed on different terminals for clustering can find the behavior patterns of people and support the in-depth development of application services. However, these data often contain individual sensitive information. Directly collecting data for clustering has the risk of revealing individual data privacy in the case of a lack of trusted data collectors. In recent years, localized differential Privacy (local differential privacy, LDP) has been continuously concerned by researchers in privacy protection because of its rigorous mathematical theory. Most of the existing LDP-based clustering methods use partition-based clustering methods, which are only suitable for convex distribution data and have the problem of large clustering quality loss. To tackle these problems, we focus on grid clustering and propose a local differential privacy based privacy-preserving grid clustering method. Firstly, we design an evaluation index of grid division, which adjusts the grid density estimation error and the loss of cluster edge

收稿日期:2022-01-17;在线发布日期:2022-10-08.本课题得到国家自然科学基金(61772131,62072156)资助.**张东月**,博士研究生,主要研究方向为数据挖掘、数据隐私安全保护.E-mail:dongyue@seu.edu.cn.**倪巍伟**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为数据挖掘、数据隐私安全保护.E-mail:wni@seu.edu.cn.**张 森**,博士研究生,主要研究方向为数据挖掘、数据隐私安全保护.**付 楠**,博士研究生,主要研究方向为数据挖掘、数据隐私安全保护.**候立贺**,博士研究生,主要研究方向为数据挖掘、数据隐私安全保护.

《计算机学报》

information to guide the selection of grid structure. Then, we construct a cyclic feedback mechanism between the server and the terminal, which uses data distribution information to iteratively optimize the disturbance granularity, reduce the amount of differential noise injection, and improve the accuracy of grid density estimation. Finally, we propose an adaptive grid aggregation method based on grid structure to improve the accuracy of privacy protection clustering on the server-side. Theoretical analysis and experimental results show that the proposed method considers the privacy of each terminal's individual data and has a good clustering effect on different distributed data.

Keywords　privacy protection; local differential privacy; grid clustering; evaluation index of grid division; circular feedback mechanism

1　引　言

聚类是数据挖掘、机器学习的基本功能[1-3]. 随着移动应用的快速发展，越来越多的个体数据分散在不同移动终端上，采集这些数据进行聚类，可以提取用户行为模式，支撑业务决策. 然而，终端上的个体数据往往涉及个体社交关系、位置、财务支出等敏感信息. 现实世界难以找到可信的数据采集者，面临保护个体数据隐私和获取全局数据聚簇模式的矛盾. 如何在不依赖可信第三方的前提下，实现兼顾分布终端个体数据隐私的聚类挖掘成为隐私保护领域迫切需要解决的问题.

本地化差分隐私[4]作为一种有严谨数学理论基础的隐私模型，通过在终端向个体数据中添加随机噪声实现隐私保护，具有对攻击者背景知识依赖弱、无需可信第三方数据收集者即可获取数据统计信息的优点. 目前该模型已广泛应用于非可信环境下多终端数据的安全采集与分析场景中[5-10]. 例如，谷歌公司将 LDP 应用在 Chrome 浏览器以统计用户行为数据[9]、苹果公司使用该技术采集 IOS 10 系统中用户常用表情符号[10]等. 在基于 LDP 的隐私保护聚类方面，现有研究工作多数采用基于划分的聚类方法(例如 K-means 等)[11-13]，通过共享初始簇中心，在终端实现个体数据的划分. 然后终端将划分信息扰动后提交给服务器，支撑服务器对聚簇中心的迭代更新，实现个体数据隐私保护与聚类挖掘的兼顾. 然而基于划分的聚类对数据分布的固有限制，导致其无法有效应用于非凸状分布数据，差分噪声的注入更是使得非凸状分布数据的隐私保护聚类质量无法保障.

针对上述问题，聚焦网格聚类，提出基于 LDP 的隐私保护网格聚类方法-AGCluster. 设计网格划分评估指标，通过调节网格划分粒度调控网格密度估算误差和簇边缘信息损失，指导网格结构选取；在服务器与终端间构建循环反馈机制，利用数据分布信息迭代优化扰动粒度，降低差分噪声注入量，在保护终端数据隐私安全的前提下，提升网格密度估算精度；最后，在服务器端，提出自适应网格聚合方法，根据网格结构自动调整网格中心，并设置动态距离阈值判断密度相连，提高网格相似性度量质量，降低聚合误差. 论文主要贡献包括：

(1) 提出基于本地化差分隐私的网格聚类方法-AGCluster. 采用解析影响聚类效果成因，优化网格结构和扰动粒度的思想，调控差分噪声规模，实现隐私保护与聚类质量的兼顾.

(2) 设计网格划分评估指标，量化网格划分粒度对网格密度估算误差和簇边缘信息损失的影响，指导网格结构选取. 在服务器与终端间构建循环反馈机制，利用数据分布信息迭代优化扰动粒度，降低差分噪声注入量，提升网格密度估算精度.

(3) 提出基于网格结构的自适应网格聚合方法，通过自动调整网格中心和设置动态距离阈值判断密度相连，提高网格间相似性度量精度，降低聚合误差.

(4) 分析所提方案误差上界. 在 6 个实验数据上对比 7 种聚类方法，验证 AGCluster 方法在兼顾终端数据隐私前提下对各类数据分布的聚簇效果.

本文第 2 节对隐私保护聚类相关工作进行概述；第 3 节介绍网格和 LDP 预备知识，并分析基于 LDP 的隐私保护网格聚类所面临的问题；第 4 节设计网格划分评估指标，给出网格结构选取方法. 构建循环反馈机制优化网格粒度并证明其有效性. 在此基础上，提出基于网格结构的自适应网格聚合方法；第 5 节设计实验验证 AGCluster 算法的有效性；最后，总结全文并展望后续研究工作.

2 相关工作

近年来隐私保护聚类得到研究者的持续关注. 文献[14]提出一种基于差分隐私的隐私保护 K-means 算法,通过将隐私分析修改为实值函数,从而极大的降低了差分噪声注入量. 然而,该算法查询函数的敏感度较大,降低了聚类结果的可用性. 文献[15]提出一种基于差分隐私的孤立点剔除 K-means 算法. 在数据预处理阶段,根据样本密度剔除离群点,并在选择聚类中心时注入 Laplace 噪声. 文献[16]提出了一种 IDP-K-means 算法. 通过将数据集划分为多个子集,并以子集中的中心作为初始簇中心,提高聚类性能. 文献[17]提出一种基于差分隐私的多核 DBSCAN 算法,重新定义初始簇中心的选择过程,并在距离计算中注入 Laplace 噪声,从而实现隐私保护. 但基于差分隐私的隐私保护聚类方法需要假设数据收集者是可信的.

针对现实世界难以找到可信第三方的问题,一些学者开始关注基于 LDP 的隐私保护聚类方法研究工作. 文献[18]提出最小包围球方法,利用局部哈希敏感函数限制簇中心选取范围,降低差分噪声对更新簇中心的影响,减少隐私保护 K-means 迭代次数. 但该方法无法发现小型簇,聚类质量较低. 文献[19]设计权重分配策略优化隐私保护 K-means 簇中心选取. 以附近个体数据数量作为候选簇中心的权重,进而通过后置处理方式在候选中心中选取簇中心,降低服务器与终端间交互次数. 但为保护个体数据隐私,该方法对权重注入差分噪声,直接影响最优簇中心选取,降低聚类质量. 文献[20]提出预置候选簇中心方法,将迭代次数限制在候选簇中心数量范围内,明确给出基于 LDP 的隐私保护 K-means 误差上界. 但该方法利用拉普拉斯噪声保护个体数据隐私,噪声注入量较大. 因此,文献[21]采用 RAPPOR 机制保护个体数据隐私,通过对位向量分配不同的隐私预算,降低差分噪声注入量. 但该方法同样采用基于划分的聚类,存在仅适用于凸状分布数据的问题.

为解决非凸状分布数据的隐私保护聚类问题,文献[22]提出一种非交互式聚类框架,通过 Bit Vector 将欧几里德距离投影到 Hamming 空间,然后扰动 Bit Vector 实现个体数据隐私的保护. 该方法优点在于,采用 Bit Vector 有效维持个体数据间距离,提升对各类数据分布的适用性. 但存在聚类结果对 Bit Vector 生成策略比较敏感,导致聚类结果不稳定的问题.

综上,现有基于 LDP 的隐私保护聚类方法多采用基于划分的聚类方法,存在仅适用凸状分布数据问题. 针对该问题,考虑聚焦网格聚类,提出基于 LDP 的隐私保护网格聚类方法,利用网格聚类对各类分布数据的良好适用性,实现隐私保护与聚类质量的兼顾.

3 预备知识及问题描述

3.1 网格相关定义

假设各终端上的个体数据包含 d 个属性,所有个体数据构成 d 维数据空间 S. 将空间每一维 $s_i(i=1,2,\cdots,d)$ 均匀划分为 p_i 份,可将空间 S 划分为 $N=\prod_{i=1}^{d} p_i$ 个网格. 每个网格简化表示为 $g=(j_1,j_2,\cdots,j_d)$.

定义 1. 网格密度[23]. 给定网格 g,以 $V(g)$ 表示其体积,$E(g)$ 为映射到该网格的个体数据数量,则其网格密度为

$$D_g=\frac{E(g)}{V(g)}.$$

网格密度反应了网格内个体数据数量,不同网格包含的个体数据量不同,本文引入稠密网格对此进行区分,定义如下.

定义 2. 稠密网格[23]. 给定密度阈值 D,如果 $D_g\geq D$,则称网格 g 为稠密网格. 反之,$D_g<D$,则称网格 g 为稀疏网格.

将 N 条个体数据均匀的映射到 N_g 个网格中,则密度阈值 D 可表示为 $N/N_gV(g)$.

定义 3. 兄弟网格[24]. 将网格 g 每一维均匀划分,得到较小粒度网格集合 G. 对于任意两个网格 $g_i,g_j\in G$,则网格 g_i 与 g_j 互称兄弟网格. g_i 为 g 的孩子网格,g 为 g_i 的父母网格.

3.2 本地化差分隐私模型

本地化差分隐私定义如下.

定义 4. ε-本地化差分隐私. 对于任意两个个体数据 x 和 y,经过本地扰动算法 M 处理后输出 s 且满足以下公式,则说明扰动算法 M 满足 ε-LDP.

$$\Pr[M(x)\in s]\leq e^{\varepsilon}\Pr[M(y)\in s],$$

其中,s 属于扰动算法 M 的值域;ε 为隐私预算,值越小表示 M 的保护程度越强.

性质 1. 序列组合性质[25]. 给定个体数据集和 n 个扰动算法 $M_1,M_2,\cdots,M_n$,且 $M_i(1\leq i\leq n)$ 满足 ε_i-本地化差分隐私,则在该数据集上的序列组合满足 $\sum_{i=1}^{n}\varepsilon_i$-本地化差分隐私.

随机应答机制[26-27]是实现本地化差分隐私的常用技术. 该机制思想是：用户在响应敏感布尔问题时，以概率 p 真实应答，以 $1-p$ 给出相反应答. 考虑采用随机应答机制对个体数据所处网格的编码进行扰动，以保护个体数据隐私. 具体网格编码扰动步骤如下：

(1) 编码. 将数据空间划分为 N_g 个网格，用户 u 的个体数据映射在网格 $g_i(1\leq i\leq N_g)$，则其编码可表示为 $Encode(u)=g_i$.

(2) 扰动. 为保护个体数据隐私，终端将编码扰动后再发送给服务器，扰动规则如下：

$$\Pr[Encode(\tilde{u})=g_j]=\begin{cases}p=\dfrac{e^{\varepsilon}}{e^{\varepsilon}+N_g-1}, & j=i\\ q=\dfrac{1}{e^{\varepsilon}+N_g-1}, & j\neq i\end{cases}.$$

(3) 聚合. 服务器收到所有编码后，统计每个网格中数据量，估算校正网格密度为

$$\hat{D}_g=\left(\sum_{u=1}^{N}Encode(\tilde{u})-\frac{N}{e^{\varepsilon}+N_g-1}\right)\cdot\frac{e^{\varepsilon}+N_g-1}{V(g)(e^{\varepsilon}-1)}.$$

通常以均方误差度量估算误差，而 LDP 中频数估算通常是无偏的，所以网格内数据量估算误差即为方差，$var(\hat{D}_g)=\dfrac{N_g-2+e^{\varepsilon}}{N(e^{\varepsilon}-1)^2}$.

3.3　问题描述

基于 LDP 的隐私保护网格聚类的质量损失主要源于以下两方面：一是网格密度估算存在误差. 在基于 LDP 的隐私保护网格聚类中通过将邻近稠密网格相连形成簇. 而为了满足 LDP，终端将网格编码扰动后发送给服务器，导致服务器对网格密度的估算存在偏差，稠密网格筛选不准确. 在此基础上进行聚簇，直接影响聚类质量；二是隐私保护网格聚类易丢失簇边缘信息. 数据在空间中分布不均. 通常处于簇内部的网格，其密度较高；处于簇边缘上的网格，其密度较低，容易被当做稀疏网格处理. 尤其在 LDP 保护下，网格内的数据以一定概率被随机映射到其它网格，可能导致空间中产生更多的稀疏网格，加重簇边缘信息的损失.

以图 1 为例，服务器统计数据分布结果. 图中网格内元组(1,0)表示该网格编码为 1，内部数据量为 0. 将个体数据映射到网格，终端上传不加差分噪声的网格编码到服务器，统计结果如图 1(a)所示. 各网格内数据量分别为：0、14、1、9、38、2、1、0、0. 其中网格 2、4、5、6 为稠密网格，可参与网格聚簇. 设置隐私预算 $\varepsilon=1$，各终端将扰动后的网格编码上传到服务器，统计结果如图 1(b)所示. 各网格内数据量分别为：0、18、0、6、36、0、6、−6、−6. 其中网格 2、4、5、7 为稠密网格. 因为网格密度的估算存在偏差，导致图 1(b)与图 1(a)中稠密网格分布有差异，在此基础上进行聚簇，影响最终聚簇质量. 经 LDP 扰动后易丢失簇边缘信息，如图 1(a)处于簇边缘上的网格内只有 2 条数据，扰动后网格内数据以 0.56 的概率被全部随机到其它网格，造成簇边缘信息丢失.

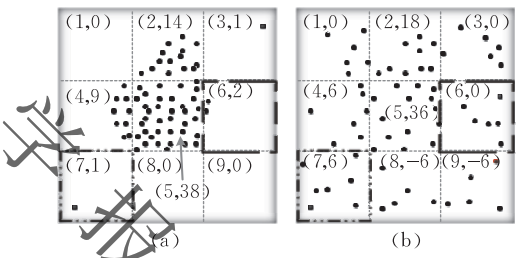

图 1　基于 LDP 的隐私保护网格聚类方法面临问题

针对上述问题，考虑采用优化网格结构选择和扰动粒度的思想，调控差分噪声规模，实现隐私保护与聚类质量的兼顾.

具体思路如图 2 所示：算法主要由网格划分评

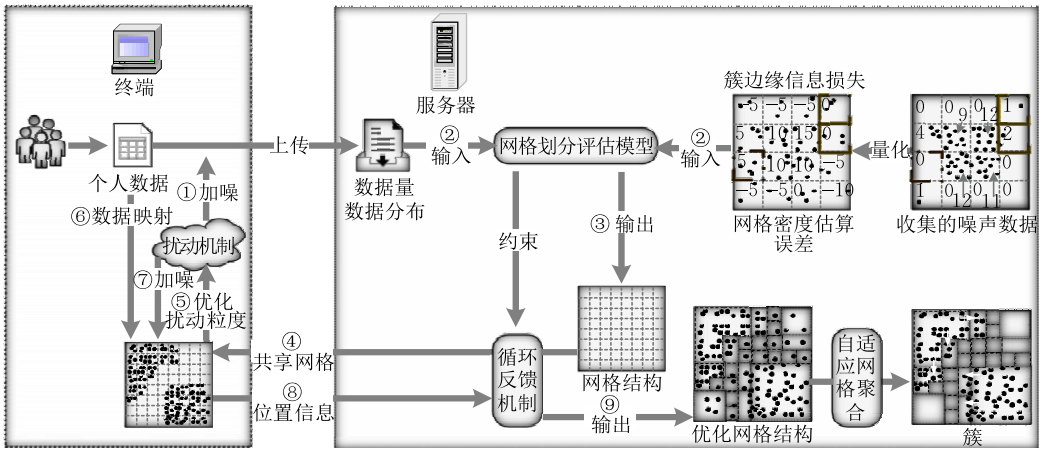

图 2　AGCluster 方法思路图

估指标设计、循环反馈机制构建以及自适应网格聚合三部分组成.第一部分,设计网格划分评估指标,指导网格结构获取(步骤①～③).服务器收集扰动后的个体数据,进行数据量和数据分布估算.将估算信息输入网格划分评估指标,并调节网格划分粒度参数调控网格密度估算误差和簇边缘信息损失,指导网格结构获取;第二部分,构建循环反馈机制,优化网格结构和扰动粒度(步骤④～⑨).在服务器与终端间构建循环反馈机制,服务器利用数据分布信息迭代优化网格结构,在此过程中以网格划分评估指标约束网格结构的调整,保证网格密度估算误差和簇边缘信息损失逐步降低.终端根据网格结构优化 LDP 扰动粒度,降低差分噪声注入量;第三部分,提出自适应网格聚合方法进行聚簇.经循环反馈机制优化后,空间中网格单位形状不规则.传统的网格聚合方法难以支撑对此类网格相似性的度量,考虑根据网格结构自动调整网格中心,设置动态距离阈值判断密度相连,实现对不规则网格相似性的度量,提高聚类质量.

4 基于LDP的隐私保护网格聚类算法-AGCluster

4.1 网格划分评估指标

网格密度估算误差.在基于 LDP 的隐私保护网格聚类中,个体数据被映射到网格,为保护个体数据隐私各终端将网格编码扰动后上传到服务器.服务器收集数据并统计网格密度,筛选稠密网格用于聚类.为降低差分噪声注入对稠密网格分布的影响,服务器需要对每个网格密度进行估算,估算值与无噪声注入的网格密度真值间存在误差,即为网格密度估算误差.

由于每个网格内估算误差相对独立,空间中任意两个网格具有相同估算误差.网格密度估算误差量化如下.

定义 5. 网格密度估算误差量化函数.在体积为 $V(S)$ 的 d 维数据空间中,设置每个网格划分粒度为 m.终端扰动个体数据所处网格的编码并将结果上传到服务器.则网格密度估算误差量化函数为

$$MDEE=\frac{V(S)}{N(e^{\varepsilon_1}-1)^2m^d},$$

其中,ε_1 为隐私预算;N 为个体数据总量.当数据空间体积一定时,网格划分粒度越小,网格密度估算误差越大.

影响聚类质量的另一要素是差分噪声加注后易丢失簇边缘信息.个体数据在空间内分布不均,通常处于簇边缘上的网格中数据量较少,经 LDP 扰动后,这类网格内数据存在较大概率被全部随机到其它网格,导致该类网格被误判为稀疏噪声网格,进而影响聚簇精度.

簇边缘信息损失与处于簇边缘的网格中数据量有关,如果被误判的数据量越多,则簇边缘信息损失越大.具体量化如下.

定义 6. 簇边缘信息损失量化函数.假设处于簇边缘上的网格被误判数量为 ne,且每个网格内数据量为 $E(g_{edge})$,则簇边缘信息损失函数为

$$CIL=\sum_{edge=1}^{ne}E(g_{edge}).$$

将 d 维个体数据映射到划分粒度为 m 的网格中,数据分布所占空间大小为 r,此时根据超方体体积与表面积关系可估算 $ne=2dr^{d-1/d}/m^{d-1}$;服务器收集个体数据总量为 N,假设数据均匀分布在每个网格,则 $E(g_{edge})\leq Nm^d/r$.此时,簇边缘信息损失量化函数可表示为 $2dmNr^{-1/d}$.

当数据空间体积一定时,网格划分粒度越小,簇边缘信息损失越低.

分析易知,网格密度估算误差和簇边缘信息损失与网格划分粒度相关,而这两个因素又影响最终聚簇质量,网格密度估算误差和簇边缘信息损失越小,则服务器得到的稠密网格分布越与无差分噪声注入时接近,在此基础上进行网格聚合,聚类质量越高.考虑基于 $MDEE$ 和 CIL 设计网格划分评估指标,通过调节网格划分粒度调控网格密度估算误差和簇边缘信息损失,指导网格结构获取.

定义 7. 网格划分评估指标.网格划分评估指标为网格密度估算误差量化函数与簇边缘信息损失量化函数之和,表示如下:

$$MSM=\frac{1}{N(e^{\varepsilon_1}-1)^2m^d}+2dmNr^{-1/d}.$$

MSM 构建网格划分粒度与聚类质量间的关联.MSM 越小,表示基于 LDP 的隐私保护网格聚类以当前网格结构收集个体数据时,聚簇质量越高.

定理 1. 对 N 条 d 维个体数据,服务器基于 MSM 得到最优网格划分粒度为

$$m_{opt}=\frac{1}{\sqrt[d+1]{2N^2(e^{\varepsilon_1}-1)^2r^{-1/d}}},$$

其中,ε_1 为隐私预算,$r\in(0,1]$ 为个体数据分布所占空间大小,当个体数据维度较大时,$r^{-1/d(d+1)}$ 接

《计算机学报》

近于 1.

证明.　网格划分评估指标的二阶导数为 $d(d+1)/N(e^{\varepsilon_1}-1)^2 m^{d+2}$，在(0,1)区间上恒大于 0，所以 MSM 为严格凸函数，可通过求其极小值取得最小值，此时 $m_{opt}=1/\sqrt[d+1]{2N^2(e^{\varepsilon_1}-1)^2 r^{-1/d}}$.　　证毕.

AGCluster 获取初始网格结构详见算法 1.

算法 1.　获取初始网格结构 $Mopt(\varepsilon_1, m, D)$.

输入：隐私预算 ε_1，网格划分粒度 m，网格密度阈值 D

输出：初始网格结构

1. FOR each u user DO
2. u reports $\leftarrow DE(Encode(u), \varepsilon_1)$;
3. END FOR
4. Generate $girdList$ with grid side length m
5. FOR each $grid$ in $girdList$ DO
6. Collector aggregates $E(g_i)$;
7. Collector calibrates D_g as;
8. //估算网格密度
9. $\hat{D}_g \leftarrow \left(\sum_{u=1}^{N} Encode(\tilde{u}) - \frac{N}{e^{\varepsilon_1}+N_g-1}\right)\cdot\frac{e^{\varepsilon_1}+N_g-1}{V(g_i)(e^{\varepsilon_1}-1)}$;
10. END FOR
11. //根据定理 1 获取最优网格划分粒度
12. $r \leftarrow \frac{sum(D_g > D)}{sum(gridList)}$; $m_{opt} \leftarrow \frac{1}{\sqrt[d+1]{2N^2(e^{\varepsilon_1}-1)^2 r^{-1/d}}}$.
13. Return m_{opt}

在算法 1 中，首先终端将个体发送服务器之前，对个体数据进行扰动，从而实现隐私保护的目的(步骤 1～3)；服务器收集扰动后的个体数据，统计每个网格内的数据量，估算校正网格密度(步骤 4～10)；然后，以稠密网格数量与网格总数的比值近似估算 r，并根据定理 1 获取最优网格划分粒度 m_{opt}(步骤 11～12)；最后，将数据空间划分为有限个粒度为 m_{opt} 的网格(步骤 13).

4.2　循环反馈机制

当数据空间中所有网格的划分粒度相同时，可将网格划分粒度统一设置为 m_{opt} 以降低网格密度估算误差和簇边缘信息损失，提高聚类质量. 进一步分析发现，网格划分粒度与网格密度估算误差成反比，与簇边缘信息损失成正比，且簇边缘信息损失产生于稀疏网格中. 如果服务器能够收集终端数据，差异处理稠密网格和稀疏网格，保证在不产生簇边缘信息损失的情况下，增大稠密网格划分粒度，降低网格密度估算误差；在不增大网格密度估算误差的情况下，减小处于簇边缘上的稀疏网格划分粒度，降低簇边缘信息损失，则可以进一步提高聚簇质量.

基于此思想，考虑在服务器与终端间构建循环反馈机制. 服务器利用数据分布的稠密状态迭代优化网格结构. 终端根据网格结构优化扰动粒度，调控差分噪声的注入量，在保护终端数据隐私安全的前提下，提升网格密度估算精度.

基于迭代的循环反馈机制如图 3 所示. 服务器将生成的初始网格结构共享给终端，然后收集终端扰动后的数据，估算网格密度. 对于稠密网格，如果其兄弟网格也稠密，将其合并为一个大粒度的网格后仍然稠密，不产生簇边缘信息损失；对于稀疏网格，如果其兄弟网格中存在稠密网格，说明该网格可能处于簇边缘上，则将其划分为多个较小粒度的网格，可降低簇边缘信息损失. 最后将调整好的网格结构共享给终端，支撑扰动粒度优化. 重复上述过程，直到所有网格划分粒度趋于稳定.

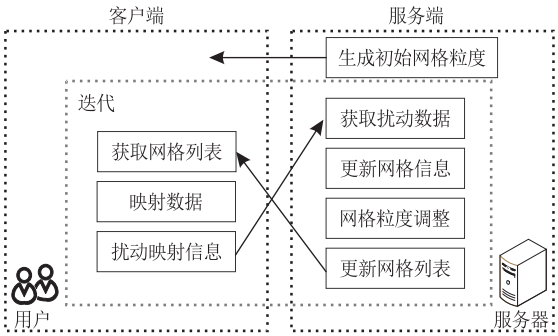

图 3　基于迭代的循环反馈机制思路图

为保证网格结构经基于迭代的循环反馈机制调整后得到优化，以网格划分评估指标约束网格调整，详细见定理 2. 此外，保存历次迭代中网格的状态信息，辅助判定当次迭代中每个网格的稠密程度，从而降低密度估算误差的影响，加快网格结构优化.

定理 2.　基于迭代的循环反馈机制在条件 $c_1 \le \frac{\varepsilon_2}{\ln(1+\varepsilon_2\sqrt{(1+c_2 d\ln 2)(1-r)})} - c_2$ 约束下调整网格，可保证网格密度估算误差和簇边缘信息损失降低.

其中 c_1 为网格粒度扩大最小次数；c_2 为网格粒度划分最大次数.

证明.　基于迭代的循环反馈机制需要多次迭代优化网格结构，每次收集个体数据需要分配隐私预算保护隐私. 假设所有网格中粒度扩大最小次数为 c_1，粒度划分最大次数为 c_2，则每个网格的密度估算误差上界为 $1/N(e^{\varepsilon_2/c_1+c_2}-1)^2$. 以 m_{opt} 作为初始网格结构中每个网格的划分粒度，且通过将每一维均匀分为 2 份来进行网格划分，则基于迭代的循环反馈机制中网格密度估算误差和簇边缘信息损失可

计算机学报

示为

$$MSM(I)\leq\frac{r}{N(e^{\varepsilon_2/c_1+c_2}-1)^2(2^{c_1}m_{opt})^d}+\frac{1-r}{N(e^{\varepsilon_2/c_1+c_2}-1)^2(m_{opt}/2^{c_2})^d}+\frac{2dm_{opt}Nr^{-1/d}}{2^{c_2}}$$
$$=\frac{1}{Nm_{opt}^d(e^{\varepsilon_2/c_1+c_2}-1)^2}\times\left[\frac{r}{2^{c_1d}}+(1-r)2^{c_2d}\right]+\frac{2dm_{opt}Nr^{-1/d}}{2^{c_2}}.$$

分析 MSM(I)发现,在基于迭代的循环反馈机制中簇边缘信息损失减小为原来的 $1/2^{c_2}$ 倍,如果网格密度估算误差也减小,则网格结构经基于迭代的循环反馈机制调整后得到优化.为此需要满足如下不等式:

$$\frac{1}{Nm_{opt}^d(e^{\varepsilon_2/c_1+c_2}-1)^2}\times\left[\frac{r}{2^{c_1d}}+(1-r)2^{c_2d}\right]\leq\frac{1}{N(e^{\varepsilon_2}-1)^2m_{opt}^d}.$$

由于 $\varepsilon_2>0$,化简得到:

$$\frac{r}{2^{c_1d}}+(1-r)2^{c_2d}\leq\frac{(e^{\varepsilon_2/c_1+c_2}-1)^2}{(e^{\varepsilon_2}-1)^2}\leq\frac{(e^{\varepsilon_2/c_1+c_2}-1)^2}{\varepsilon_2^2}.$$

因为 $2^{c_1d}\gg1$,由此可得到:

$$\ln 2c_2d\leq\ln\frac{(e^{\varepsilon_2/c_1+c_2}-1)^2}{\varepsilon_2^2(1-r)}.$$

由 $\frac{(e^{\varepsilon_2/c_1+c_2}-1)^2}{\varepsilon_2^2(1-r)}>0$,可得到:

$$(1+\ln 2c_2d)\varepsilon_2^2(1-r)\leq(e^{\varepsilon_2/c_1+c_2}-1)^2.$$

进一步化简得到:

$$c_1\leq\frac{\varepsilon_2}{\ln(1+\varepsilon_2\sqrt{(1+c_2d\ln 2)(1-r)})}-c_2.$$ 证毕.

基于迭代的循环反馈机制细节见算法 2.

算法 2. 基于迭代的循环反馈机制算法 *IterAdaptiveM*(ε_2,D).

输入:隐私预算 ε_2,网格密度阈值 D

输出:网格列表 *girdList*(*Adaptive*)

```
1. Generate girdList with grid side length m_opt;
2. WHILE flag>0
3.    FOR each u user DO
4.       u reports←DE(Encode(u),ε2);
5.    END FOR
6.    FOR each grid in girdList DO
7.       Update grid feature;
8.       IF ∀ brother grids satisfy D_g>D THEN
9.          //将该网格与其兄弟网格合并
10.         MergeGrid(brothergrids);
11.         flag←flag+1;
12.      IF D_g<D
13.         //将该网格分割为粒度更小的网格
14.         SplitGrid(grid);
15.         flag←flag+1;
16.   END FOR
17. Return girdList
```

在算法 2 中,服务器生成初始网格结构并共享给终端(步骤 1);终端将个体数据映射到网格,根据网格结构调整扰动粒度,对个体数据所处网格的编码进行扰动并将结果发送到服务器(步骤 3～5);服务器接收数据,估算校正网格密度,然后根据稠密程度对网格划分粒度进行调整(步骤 6～16).重复这一过程,当所有网格划分粒度均趋于稳定,即得到优化的网格结构.

当数据分布较密集时,采样部分数据也能获取较准确的数据分布信息.据此进一步构建第二种循环反馈机制——基于网格树的循环反馈机制.该机制将个体数据随机映射到不同划分粒度的网格结构中,可达到比算法 2 更小的误差界.证明见定理 4.

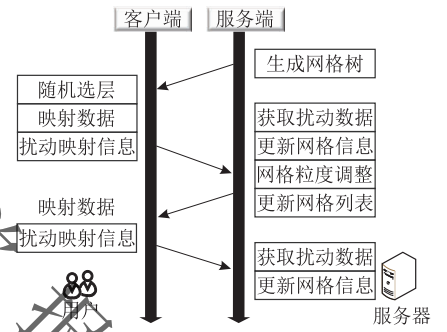

图 4 基于网格树的循环反馈机制思路图

基于网格树的循环反馈机制如图 4 所示.服务器生成网格树.该树根结点为整个数据空间,从根到叶结点网格划分粒度逐渐减小.终端将自身数据随机映射到树中某一层的网格内.由于上层网格结构中网格划分粒度较大,网格密度估算更精确,因此服务器接收数据后从上到下调整网格.如果当前层某网格为稀疏,表示该网格存在簇边缘信息损失问题,则进一步判断其孩子网格稀稠状态.如果当前层某网格为稠密,则将其确定为该区域的最优网格结构.

基于网格树的循环反馈机制中网格树高度 H 应适当.H 太大则需要大量空间存储网格树,H 太小则限制网格划分粒度调整范围,使得部分区域无法获取最优网格结构.根据循环反馈机制调整后,网格结构应得到优化的思想,给出 H 参考范围.详见定理 3.

定理 3. 初始网格结构中每个网格的划分粒度为 m_{opt},个体数据维度为 d,以基于网格树的循环

反馈机制优化网格结构，则网格树高度设置范围为 $\log_2\frac{1}{m_{opt}}\leq H\leq\frac{1+\ln(1/m_{opt})}{1-r+d\ln 2}$.

证明.　以基于网格树的循环反馈机制优化网格结构，为保证初始网格结构中每块区域都有进一步优化的空间，应使得叶子节点的网格划分粒度小于 m_{opt}，则：

$$\log_2(1/m_{opt})\leq H.$$

假设所有网格中粒度扩大最小次数为 c_1，粒度划分最大次数为 c_2，则基于网格树的循环反馈机制中网格密度估算误差和簇边缘信息损失可表示为

$$\mathrm{MSM(T)}\leq\frac{r}{N/H(e^{\varepsilon_2}-1)^2(2^{c_1}m_{opt})^d}+\frac{1-r}{N/H(e^{\varepsilon_2}-1)^2(m_{opt}/2^{H-c_1})^d}+\frac{2dm_{opt}Nr^{-1/d}}{2^{H-c_1}}.$$

调整后网格结构应得到优化，因此：

$$c_1\geq\frac{\ln H(1-r)}{d\ln 2}+H.$$

根据所有网格中粒度扩大最小次数应满足 $c_1\leq\log_2(1/m_{opt})$，可得：

$$H\leq\frac{1+\ln(1/m_{opt})}{1-r+d\ln 2}.$$

综上，可得到网格树的高度参考范围：$\log_2\frac{1}{m_{opt}}\leq H\leq\frac{1+\ln(1/m_{opt})}{1-r+d\ln 2}$.　证毕.

基于网格树的循环反馈机制详见算法 3.

算法 3.　基于网格树的循环反馈机制算法 *TreeAdaptiveM*(ε_2, D, H).

输入：隐私预算 ε_2，网格密度阈值 D，树高 H

输出：自适应网格列表 *girdList*

1. Generate a *girdTree* at height H;
2. FOR each u user DO
3.　Randomly partition all users into groups H;
4.　u reports←$DE(Encode(u),\varepsilon_2)$;
5. END FOR
6. FOR each *grid* in *girdTree* DO
7.　Update grid feature;
8. FOR each *grid* in *girdTree* DO
9.　IF *grid* satisfy $D_g>D$ THEN
10.　//将该网格添加进网格列表
11.　*AddTogirdList*(*grid*);
12.　ELSE
13.　*JudgmentSubgrid*(*grid*);
14. END FOR
15. Return *girdList*

在基于网格树的循环反馈机制中，服务器首先生成高度为 H 的网格树并共享给终端（步骤 1）. 终端将个体数据随机映射到某一层网格结构中，然后发送扰动后的网格编码到服务器（步骤 2～5）. 服务器接收数据，校正网格密度. 从上到下为每块区域选择合适的网格结构（步骤 6～14）.

定理 4.　估算网格内数据量时，基于网格树的循环反馈机制比基于迭代的循环反馈机制拥有更小的误差界.

证明.　对于任意一条个体数据 u_i，以 $g(u_i)=\{1,0\}$ 表示该数据是否映射在网格 g，$\tilde{g}(u_i)$ 表示对 $g(u_i)$ 扰动后的结果. LDP 扰动规则为：以 $p=e^{\varepsilon_2}/(e^{\varepsilon_2}+1)$ 的概率保持 $g(u_i)$ 值不变，以 $q=1/(e^{\varepsilon_2}+1)$ 的概率取相反值，则

$$|g(u_i)-\tilde{g}(u_i)|\leq 1,\ \left|\frac{g(u_i)-\tilde{g}(u_i)}{p-q}\right|\leq\frac{e^{\varepsilon_2}+1}{e^{\varepsilon_2}-1},$$

$g(u_i)-\tilde{g}(u_i)$ 的方差可表示为

$$Var(g(u_i)-\tilde{g}(u_i))=O(1/\varepsilon_2^2).$$

设网格树高度为 H，则对于算法 2 来说每层网格结构所映射数据总量为 N/H. $E(g)$ 为无差分噪声注入时映射到网格 g 的数据量，$\hat{E}(g)$ 为注入差分噪声时服务器对该网格内数据量的估算值，则伯恩斯坦不等式可表示为

$$\begin{aligned}&\Pr(|\hat{E}(g)-E(g)|\geq\lambda_t)\\&\leq 2\cdot\exp\left(-\frac{\lambda_t^2N/H}{\frac{2H}{N}\sum_{i=1}^{N/H}Var(g(u_i)-\tilde{g}(u_i))+\frac{2}{3}\lambda_t\cdot\frac{e^{\varepsilon_2}+1}{e^{\varepsilon_2}-1}}\right)\\&=2\cdot\exp\left(-\frac{\lambda_t^2N/H}{O(1/\varepsilon_2^2)+O(1/\varepsilon_2)}\right)=\beta.\end{aligned}$$

因此存在 $\lambda_t=O(\sqrt{(H\ln 1/\beta)/N\varepsilon_2^2})$ 使得算法 3 以 $1-\beta$ 的概率保证 $|\hat{E}(g)-E(g)|<\lambda_t$. 设基于迭代的循环反馈机制迭代次数为 H，同理可证存在 $\lambda_i=O(H\sqrt{\ln(1/\beta)/N}/\varepsilon_2)$，使得算法 2 以 $1-\beta$ 的概率保证 $|\hat{E}(g)-E(g)|<\lambda_i$. 由于 $\lambda_t<\lambda_i$，所以基于网格树的循环反馈机制具有更小的网格密度估算误差界.　证毕.

4.3　自适应网格聚合方法

本节基于网格结构设计自适应网格聚合方法. 引入核心网格概念，快速定位处于簇中心的网格；然后自动调控网格中心并设置动态距离阈值判断其余网格与核心网格间的相似性，提高聚簇质量.

通常处于簇中心上的网格，其周围网格的密度较高；反之，处于簇边缘上的网格，其周围网格的密度较低. 据此引入核心网格概念.

计算机学报

定义 8. 核心网格. 与网格 g 中心间距离小于 d_l 的网格集合为 $L(g)$，当 $L(g)$ 中网格密度之和大于阈值 θ 时，则称网格 g 为核心网格. 定义如下：

$$Den(g)=\left|\left\{\sum_{g_i\in L(g)}D(g_i)\,|\,|g,g_i|<d_l\right\}\right|>\theta,$$

其中，$|g,g_i|$ 表示网格 g 与网格 g_i 中心间的欧式距离，d_l 为距离阈值.

循环反馈机制中网格与周围稠密网格合并为较大划分粒度的网格. 为了减少算法参数，对于循环反馈机制优化后的网格结构，考虑根据网格划分粒度快速定位核心网格.

经循环反馈机制优化后，空间中网格形状不规则，固定的距离计算方法和距离阈值难以准确度量这类网格间的相似性. 所以本文根据网格结构自动调整网格中心，设置动态距离阈值判断密度相连，实现对形状不规则网格间相似性的度量，提高聚类质量.

定义 9. 密度相连. 对于任意网格 g_i 与 g_j，如果 g_i 与 g_j 之间距离小于给定阈值 $m_{\min}$，则认为这两个网格密度相连.

其中 $m_{\min}$ 为两个网格中划分粒度的最小值. 当两个网格的划分粒度不同时，网格间距离为较小划分粒度网格的中心与较大划分粒度网格中孩子网格中心间的长度. 孩子网格中心计算如下：

对于任意两个网格 g_i 和 g_j，其网格中心分别为 c_i 和 c_j，网格划分粒度分别为 m_i 和 m_j（$m_i>m_j$）. 令 $K=m_i/m_j$，$k=(-1)^f\lceil|c_i-c_j|/m_j\rceil$，当 $c_i-c_j>0$ 时，f 为奇数；当 $c_i-c_j<0$ 时，f 为偶数. 则网格 m_i 的孩子网格中心为

$$c_i'=\begin{cases}c_i+\dfrac{1-K}{2}m_j, & k<-\dfrac{K}{2}\\ c_i+\dfrac{(1+2k)m_j}{2}, & k<0 \text{ and } |k|\le\dfrac{K}{2}\\ c_i+\dfrac{(2k-1)m_j}{2}, & k>0 \text{ and } |k|\le\dfrac{K}{2}\\ c_i+\dfrac{K-1}{2}m_j, & k>\dfrac{K}{2}\end{cases}.$$

自适应网格聚合方法详见算法 4.

算法 4. 基于网格结构的自适应网格聚合方法 $AGAggregation(\varepsilon_1,\varepsilon_2,\varepsilon_3,D,H)$.

输入：隐私预算 $\varepsilon_1,\varepsilon_2,\varepsilon_3$，网格密度阈值 D，树高 H

输出：簇 Clusters

1. Collector calculate $m_{opt}\leftarrow Mopt(\varepsilon_1,m,D)$;
2. Collector optimize grid structure:
   $girdList\leftarrow TreeAdaptiveM(\varepsilon_2,D,H)$ OR
   $IterAdaptiveM(\varepsilon_2,D)$;
3. FOR each u user DO
4. u reports$\leftarrow DE(Encode(u),\varepsilon_3)$;
5. END FOR
6. FOR each $grid$ in $girdList$ DO
7. Update grid feature;
8. IF $D_g<D$ THEN
9. //筛选稠密网格
10. $girdList_{remove}(grid)$;
11. END FOR
12. //根据网格边长定位核心网格
13. $Clusters\leftarrow InitializationCluster(girdList)$;
14. FOR each $grid$ in $girdList$ DO
15. //搜索密度相连网格完成聚簇
16. $Clusters\leftarrow AssignGird(grid,Clusters)$;
17. Return $Clusters$

在自适应网格聚合方法中，首先服务器获取网格结构（步骤 1）；然后基于循环反馈机制优化网格结构和扰动粒度（步骤 2）；终端将个体数据映射到优化网格结构中，扰动个体数据所处网格的编码并将其上传到服务器，服务器收集数据后估算网格密度，筛选稠密网格（步骤 3～11）；查找核心网格作为初始簇，通过搜索与核心网格密度相连的网格，完成聚簇（步骤 12～16）.

5 实验分析

本节对 AGCluster 方法的聚类质量及聚类效率进行实验，验证所提方法隐私保护聚类效果. 实验环境为 Intel(R) Xeon(R) E-2134 CPU(3.5 GHz)，64 GB 内存，Windows 10 操作系统.

5.1 实验准备

(1) 实验数据

实验选用 2 类实验数据，人工合成数据 circles、moons、aniso 以及 AData，真实数据 3D Road Network、CarGPS 以及 Credit card clients. 详细实验数据信息见表 1.

表 1 实验数据信息表

数据集	属性数	数据量	簇数
circles	2	15000	2
moons	2	15000	2
aniso	2	15000	3
AData	2	150000	3
3D Road Network	4	434874	2
CarGPS	4	8669500	2
Credit card clients	24	300000	7

表 1 中真实数据 3D Road Network[28] 为丹麦北日德兰半岛的道路网络信息，其形式为{ID，经度，维度，海拔}. CarGPS[29-30] 是由 Microsoft Research

收集的真实轨迹数据. 从中随机抽取一部分组成 CarGPS 数据集. Credit card client 为信用卡客户数据集，包含个人信用额度、过去付款历史、账单金额等信息. 由于真实数据没有真实类别标签，考虑采用 Calinski-Harabaz 指数[31] 度量其最佳簇数，并以 DBSCAN 聚簇结果作为数据类别标签，方便后续实验中对比方法的参数设置和聚类结果评估.

(2) 评估指标

为充分衡量各隐私保护聚类方法的聚类质量，采用 *Accuracy*、*Purity* 和 *Fowlkes-Mallows Index* 作为聚类评估指标.

Accuracy 代表聚类正确的百分比，是最流行的聚类评价指标. *Accuracy* 定义如下：

$$AC(p,r)=\frac{\sum_{i=1}^{n}\delta(r_i,map(p_i))}{n},$$

其中，p_i 和 r_i 分别表示数据所对应的簇标签和真实划分标签，n 为数据总量，δ 表示函数如下：

$$\delta(a,b)=\begin{cases}1, & a=b\\ 0, & \text{其他}\end{cases}.$$

$AC\in[0,1]$，值越大，表明聚类质量越高.

Purity 为纯度，是一种简单透明的评估指标. 其定义如下：

$$Purity(\Omega,C)=\frac{1}{N}\sum_{k}\max_{j}|w_k\cap c_j|,$$

其中，N 表示样本总数，$\Omega=\{w_1,w_2,\cdots,w_k\}$ 表示聚类结果，$C=\{c_1,c_2,\cdots,c_j\}$ 表示真实划分类别. $Purity\in[0,1]$，值越接近 1，表明聚类质量越高.

Fowlkes-Mallows Index，简称 *FMI*. 被定义成 *Recall* 和 *Precision* 的几何平均值. 形式如下：

$$FMI=\sqrt{Recall\times Precision},$$

其中，*Recall* 为召回率，*Precision* 为精度. 公式为 $Recall=\frac{TP}{TP+FN}$，$Precision=\frac{TP}{TP+FP}$. TP 表示在真实类别与聚类结果中均属于相同簇的数据量，FP 表示在真实类别为相同簇而在聚类结果中为不同簇的数据量，FN 表示在真实类别中为不同簇而在聚类结果中为相同簇的数据量.

(3) 对比方法

以 Kmeans、CLIQUE、LDPKmeans[21]、kCluster[22]、UGC、AGC1 及 AGC2 作为对比方法. 其中，Kmeans 为非隐私保护划分式聚类方法，CLIQUE 为非隐私保护网格聚类方法，LDPKmeans 和 kCluster 为基于 LDP 的隐私保护划分式聚类方法. UGC、AGC1 和 AGC2 为本文所提基于 LDP 的隐私保护网格聚类方案，均以自适应网格聚合方法完成聚簇，区别是 UGC 算法没有使用循环反馈机制对网格结构和扰动粒度进一步优化. AGC1 和 AGC2 算法分别采用基于迭代的循环反馈机制和基于网格树的循环反馈机制，优化扰动粒度.

5.2　聚类质量评估

在各类分布实验数据上，对 AGCluster 方法的聚类质量进行实验分析.

实验 1.　在实验数据 Credit card clients 上，验证网格划分评估指标的有效性. 当隐私预算 ε 为 1 时，根据网格划分评估指标获得该实验数据的最优网格划分粒度 m_{opt} 为 0.014. 在 m_{opt} 附近取多个网格划分粒度，并对比 UGC 方法的聚类质量，结果如表 2 所示. 当网格划分粒度为 m_{opt} 时，聚类评估指标取得最大值，因为网格划分评估指标能够较好的平衡网格密度估算误差和簇信息损失. 当网格划分粒度增大时，网格密度估算误差降低，但簇信息损失增大. 当网格划分粒度减小时，簇信息损失降低，但网格密度估算误差增大. 因此当网格划分粒度不等于 m_{opt} 时，聚类评估指标值出现不同程度的下降.

表 2　不同网格划分粒度下聚类评估指标对比

数据集	Credit card clients($\varepsilon=1$)				
	$m=0.01$	$m=0.012$	$m_{opt}=0.014$	$m=0.02$	$m=0.03$
Accuracy	0.542	0.615	0.698	0.599	0.497
Purity	0.657	0.691	0.747	0.671	0.646
FMI	0.492	0.508	0.511	0.496	0.488

实验 2.　在实验数据 circles、moons 以及 aniso 上，验证 AGCluster 方法对各类分布数据的适用性. 为了尽可能维持簇的原始轮廓，以便直观展示各隐私保护聚类方法对不同分布数据的适用性，实验中各方法的隐私预算 ε 统一设置为 5.

实验结果如图 5 所示. Kmeans、LDPKmeans 以及 kCluster 方法将邻近簇中的数据聚到一起，聚簇效果较差. 因为这 3 种基于划分的聚类方法不能有效处理非凸状分布数据. 相比与上述方法，CLIQUE、UGC、AGC1 和 AGC2 方法则能够将邻近数据聚到同簇，而非邻近数据聚到不同簇. 因为这 4 种是网格聚类方法，通过搜索密度相连的网格完成聚簇，适用于各类分布实验数据. 分析方法在实验数据 aniso 聚簇结果发现，AGC1 和 AGC2 方法中各簇边缘有少量信息损失，而 UGC 方法中各簇边缘信息损失较多，说明经循环反馈机制优化扰动粒度后，能够更好维护簇边缘信息，提高聚簇质量.

计算机学报

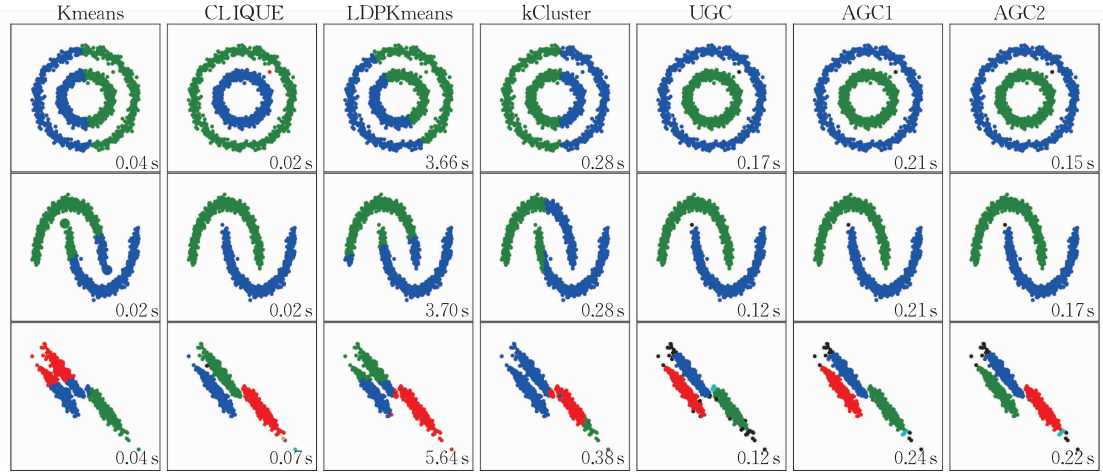

图 5 circles、moons 以及 aniso 实验数据上聚类效果对比

对比各方法完成聚簇所需时间发现，基于 LDP 的划分式隐私保护聚类方法平均运行时间是 Kmeans 的 70 倍，而本文所提方法平均运行时间是 CLIQUE 的 5 倍. 因为基于划分的聚类方法簇中心对差分噪声敏感，导致需要更多次迭代才能获取稳定的簇中心.

实验 3. 在实验数据 AData、3D Road Network、CarGPS 以及 Credit card client 上，对比各方法的 *Accuracy*、*Purity* 和 *Fowlkes-Mallows Index* 指标. 实验中各隐私保护聚类方法的隐私预算 ε 分别取 0.25，0.5，1，2，4，8，10. 为充分评估差分噪声影响下的聚类质量，对各方法运行 50 次，取评估指标均值作对比.

实验结果如图 6～图 9 所示. 随着 ε 增大，差分噪声注入量逐渐减小，各方法聚类质量逐渐提高. 当 $\varepsilon>2$ 时，聚类质量评估指标值趋于稳定，呈现轻微涨幅. 从评估指标角度分析发现，Kmeans 和 QLIQUE 聚类质量分别高于基于 LDP 的划分式聚类和网格聚类，说明注入差分噪声会降低算法聚类质量；UGC 与 LDPKmeans、kCluster 聚类质量相近，说明网格划分评估指标能够有效指导服务器获取合适网格结构，平衡网格密度估算误差和簇边缘信息损失，支撑隐私保护网格聚类获取较好的聚簇效果；UGC 聚类质量低于 AGC1 和 AGC2，说明循环反馈机制能够有效优化扰动粒度，降低差分噪声的注入量，在保护终端数据隐私安全的前提下，提升方法聚类质量.

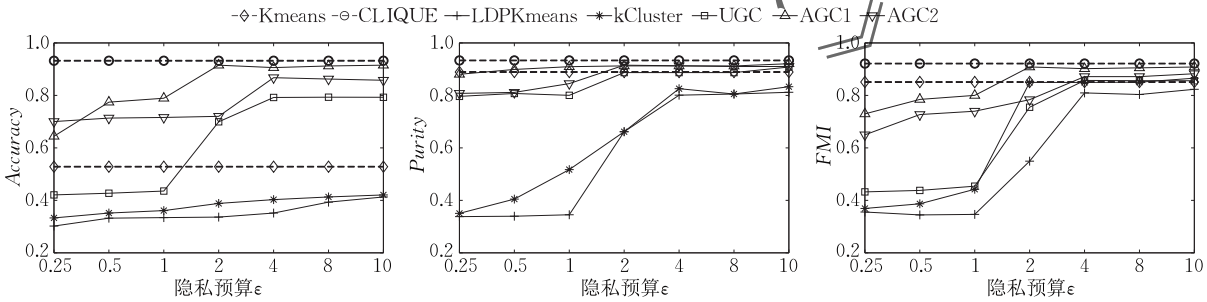

图 6 AData 实验数据上聚类质量对比

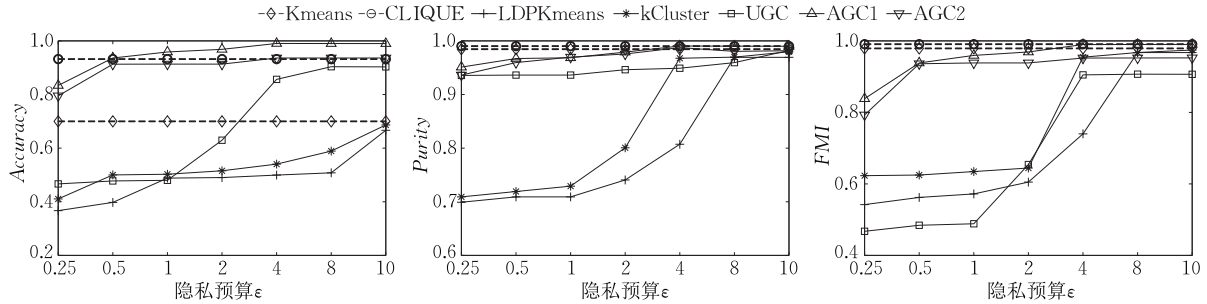

图 7 3D Road Network 实验数据上聚类质量对比

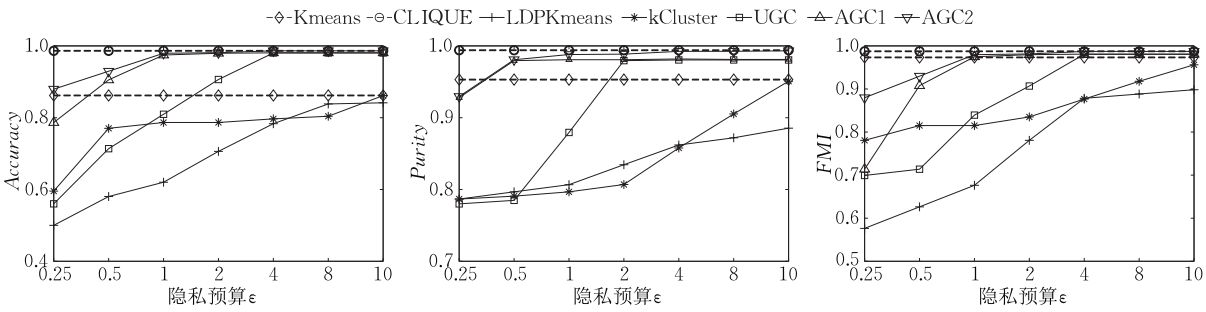

图 8　CarGPS 实验数据上聚类质量对比

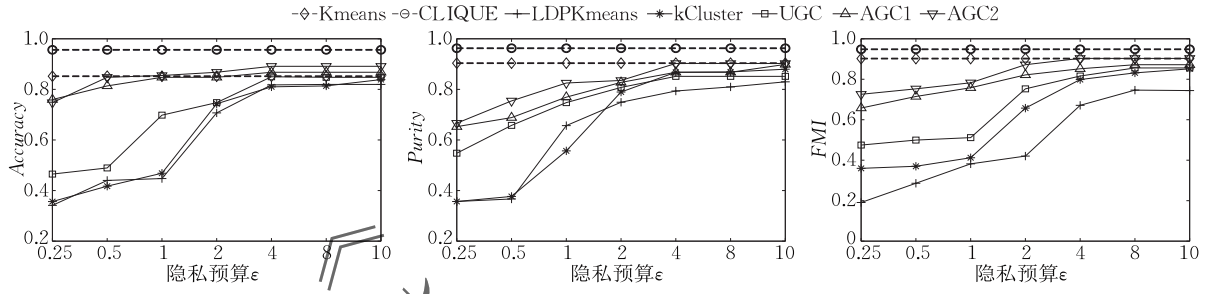

图 9　Credit card clients 实验数据上聚类质量

从评估指标变化幅度分析发现，本文所提隐私保护网格聚类方法具有较高聚类质量的同时，评估指标变化幅度较小，而 LDPKmeans 和 kCluster 随着 ε 增大，聚类质量变化幅度较大. 说明循环反馈机制能够有效降低网格密度估算误差和簇边缘信息损失，保证较高和较稳定的聚类质量.

从实验数据角度分析. 当隐私预算相同时，实验数据量越大，UGC、AGC1 和 AGC2 评估指标值越高. 因为网格密度估算误差受个体数据总量影响，数据量越多，网格密度估算误差越低，隐私保护网格聚类方法的聚类质量越高. 当隐私预算相同时，实验数据维度越大，隐私保护聚类方法的评估指标值越低. 因为隐私保护聚类方法需要为每一维分配隐私预算，实验数据维度越大，则每一维分配的隐私预算越小，差分噪声的注入量越大. AGC1 和 AGC2 方法聚类质量比较. 图 6 和图 7 显示 AGC1 方法的聚类质量较优，而图 8 和图 9 显示 AGC2 方法的聚类质量较优. 因为，在 AData 和 3D Road Network 实验数据中，数据分布较稀疏、簇密度较低，将这类实验数据分组会增加簇内部网格与稀疏网格的区分难度，增大簇信息损失. 相反，CarGPS 和 Credit card clients 实验数据中，数据分布较稠密、簇密度较高，采样部分数据也能较准确的判断网格稀稠状态，不容易造成簇信息损失.

5.3　聚类效率评估

在实验数据 AData、3D Road Network、CarGPS 以及 Credit card clients 上，对比各方法的聚簇效率. 实验中隐私预算统一设置为 0.1. 实验结果如图 10 所示，UGC、AGC1 和 AGC2 聚簇效率高于 LDPKmeans 和 kCluster. 因为本文所提隐私保护网格聚类方法，通过将聚簇对象从个体数据转为网格单元，大大减少聚簇单位数量，提高了聚簇效率. 同时，在循环反馈机制中以网格状态历史信息辅助服务器判断当前迭代中网格的稠密程度，降低差分噪声对网格状态判断的影响，加快网格结构优化. 而 LDPKmeans 和 kCluster 方法采用基于划分的聚簇方式，需要多次迭代才可能获取收敛的簇中心. 尤其在 LDP 下，数据被注入差分噪声，给算法迭代次数带来了不确定性.

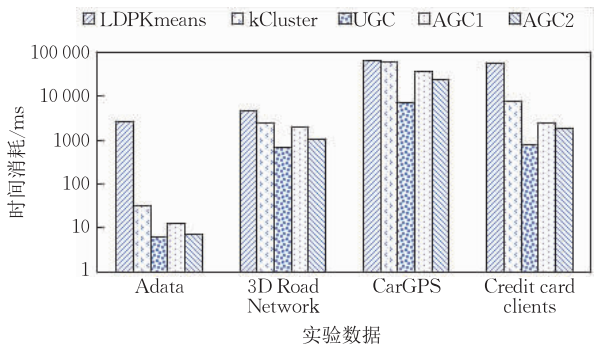

图 10　各实验数据上聚簇效率对比

6　总　结

现有基于 LDP 的隐私保护聚类研究多数采用基于划分的聚类方法，存在仅适用凸状分布数据以及聚类质量损失较大问题. 针对该问题，聚焦网格聚类，

提出基于LDP的隐私保护网格聚类方法AGCluster.采用解析影响聚类效果成因,优化网格结构选择和扰动粒度的思想,在不降低隐私保护强度的同时,减小网格密度估算误差和簇边缘信息损失,提升隐私保护聚类质量.通过设计网格划分评估指标,调节网格划分参数调控网格密度估算误差和簇边缘信息损失,指导网格结构的选取;在服务器与终端间构建循环反馈机制,利用数据分布信息迭代优化扰动粒度,降低差分噪声的注入量,在保护终端数据隐私安全的前提下,提升网格密度估算精度.最后,在服务器端,提出自适应网格聚合方法,根据网格结构自动选择相似性计算方式,提升隐私保护聚类质量.理论分析和实验验证了AGCluster方法的有效性.

下一步,将对隐私保护数据降维方法进行研究,提升高维数据上网格聚类的聚类质量与效率.

参 考 文 献

[1] Min E, Guo X, Liu Q, et al. A survey of clustering with deep learning: From the perspective of network architecture. IEEE Access, 2018, 6: 39501-39514

[2] Zhang H, Lin Z, Zhang C, et al. Robust latent low rank representation for subspace clustering. Neurocomputing, 2014, 145: 369-373

[3] Huang Mao-Feng, Ni Wei-Wei, Wang Jia-Jun, et al. A logarithmic spiral based data perturbation method for clustering. Chinese Journal of Computers, 2012, 35(11): 2275-2282(in Chinese)
(黄茂峰, 倪巍伟, 王佳俊等. 一种面向聚类的对数螺线数据扰动方法. 计算机学报, 2012, 35(11): 2275-2282)

[4] Kasiviswanathan S P, Lee H K, Nissim K, et al. What can we learn privately? SIAM Journal on Computing, 2011, 40(3): 793-826

[5] Ding B, Kulkarni J, Yekhanin S. Collecting telemetry data privately//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA, 2017: 5430-5439

[6] Johnson N, Near J P, Song D. Towards practical differential privacy for SQL queries. Proceedings of the VLDB Endowment, 2018, 11(5): 526-539

[7] Ouyang Jia, Yin Jian, Xiao Zheng-Hong, et al. Transaction data collection for itemset mining under local differential privacy. Journal of Software, 2021, 32(11): 3541-3562(in Chinese)
(欧阳佳, 印鉴, 肖政宏等. 面向频繁项集挖掘的本地差分隐私事务数据收集方法. 软件学报, 2021, 32(11): 3541-3562)

[8] Zhang Xiao-Jian, Fu Nan, Meng Xiao-Feng. Key-value data accurate collection under local differential privacy. Journal of Computers, 2020, 43(8): 1479-1492(in Chinese)
(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. 计算机学报, 2020, 43(8): 1479-1492)

[9] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale, USA, 2014: 1054-1067

[10] Cormode G, Jha S, Kulkarni T, et al. Privacy at scale: Local differential privacy in practice//Proceedings of the 2018 International Conference on Management of Data. Shanghai, China, 2018: 1655-1658

[11] Wang T, Zheng Z, Rehmani M H, et al. Privacy preservation in big data from the communication perspective—A survey. IEEE Communications Surveys Tutorials, 2019, 21(1): 753-778

[12] Wang T, Zhang X, Feng J, et al. A comprehensive survey on local differential privacy toward data statistics and analysis. Sensors, 2020, 20(24): 7030

[13] Xiong X, Liu S, Li D, et al. A comprehensive survey on local differential privacy. Security and Communication Networks, 2020, 2020: 1-29

[14] Blum A, Dwork C, McSherry F, et al. Practical privacy: The SuLQ framework//Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Baltimore, USA, 2005: 128-138

[15] Yu Q, Luo Y, Chen C, et al. Outlier-eliminated k-means clustering algorithm based on differential privacy preservation. Applied Intelligence, 2016, 45(4): 1179-1191

[16] Yang L I. Research on differential privacy preserving k-means clustering. Computer Science, 2013, 59(1): 1-34

[17] Ni L, Li C, Wang X, et al. DP-MCDBSCAN: Differential privacy preserving multi-core DBSCAN clustering for network user data. IEEE Access, 2018, 6: 21053-21063

[18] Nissim K, Stemmer U. Clustering algorithms for the centralized and local models//Proceedings of the Algorithmic Learning Theory. Lanzarote, Spain, 2018: 619-653

[19] Stemmer U, Kaplan H. Differentially private k-means with constant multiplicative error//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York, USA, 2018: 5436-5446

[20] Stemmer U. Locally private k-means clustering//Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Baltimore, USA, 2020: 548-559

[21] Xia C, Hua J, Tong W, et al. Distributed k-means clustering guaranteeing local differential privacy. Computers and Security, 2020, 90: 101699

[22] Sun L, Zhao J, Ye X. Distributed clustering in the anonymized space with local differential privacy. arXiv preprint arXiv:1906.11441, 2019

[23] Dong S, Liu J, Liu Y, et al. Clustering based on grid and local density with priority-based expansion for multi-density data. Information Sciences, 2018, 468: 103-116

[24] Wan L, Ng W K, Dang X H, et al. Density-based clustering

of data streams at multiple resolutions. ACM Transactions on Knowledge Discovery from Data, 2009, 3(3): 14:1-14:28

[25] McSherry F D. Privacy integrated queries: An extensible platform for privacy-preserving data analysis//Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. Providence, USA, 2009: 19-30

[26] Wang T, Blocki J, Li N, et al. Locally differentially private protocols for frequency estimation//Proceedings of the 26th USENIX Security Symposium. Vancouver, Canada, 2017: 729-745

[27] Ye Qing-Qing, Meng Xiao-Feng, Zhu Min-Jie, et al. Survey on local differential privacy. Journal of Software, 2017, 29(7): 1981-2005(in Chinese)
(叶青青, 孟小峰, 朱敏杰等. 本地化差分隐私研究综述. 软件学报, 2017, 29(7): 1981-2005)

[28] Kaul M, Yang B, Jensen C S. Building accurate 3D spatial networks to enable next generation intelligent transportation systems//Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management. Milan, Italy, 2013, 1: 137-146

[29] Yuan J, Zheng Y, Zhang C, et al. T-drive: Driving directions based on taxi trajectories//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose, USA, 2010: 99-108

[30] Yuan J, Zheng Y, Xie X, et al. Driving with knowledge from the physical world//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: Association for Computing Machinery, 2011: 316-324

[31] Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-Theory and Methods, 1974, 3(1): 1-27

ZHANG Dong-Yue, Ph. D. candidate. His research interests include data mining and data privacy protection.

NI Wei-Wei, Ph. D., professor, Ph. D. supervisor. His current research interests include data mining and data privacy protection.

ZHANG Sen, Ph. D. candidate. His research interests include data mining and data privacy protection.

FU Nan, Ph. D. candidate. His research interests include data mining and data privacy protection.

HOU Li-He, Ph. D. candidate. His research interests include data mining and data privacy protection.

计算机学报

Background

In recent years, privacy-protected clustering based on local differential privacy has been concerned by researchers. Most of the existing research work adopts partition-based clustering methods (such as k-means, etc.), which realizes the division of individual data in the terminal by sharing the initial cluster center. Then the terminal will disturb the partition information and submit it to the server to support the iterative update of the clustering center by the server to achieve the balance between individual data privacy protection and clustering mining. However, due to the inherent limitation of partition-based clustering on data distribution, it can not be effectively applied to non-convex distribution data, and the privacy protection clustering quality of non-convex distribution data can not be guaranteed by the injection of differential noise. To solve this problem, we focus on grid clustering, propose a local differential privacy based privacy-preserving grid clustering method, AGCluster. Individual data is mapped to a grid, and LDP is used to disturb grid number to protect individual data privacy. The idea of analyzing the cause of the clustering effect, optimizing the selection of grid structure and the size of disturbance, and adjusting the scale of differential noise are adopted to achieve the balance between privacy protection and clustering quality. First, the grid partition evaluation index is designed, and the grid density estimation error and cluster edge information loss are controlled by adjusting the grid partition granularity to guide the selection of grid structure; Then, a circular feedback mechanism is constructed between the server and the terminal. The data distribution information is used to optimize the disturbance granularity and reduce differential noise injection iteratively. Under the premise of protecting the data privacy of the terminal, the accuracy of the grid density estimation is improved; Finally, an adaptive grid aggregation method based on grid structure is proposed to improve privacy-preserving clustering accuracy. Theoretical analysis and experimental results show that the proposed method has a good clustering effect for different distributed data while considering individual privacy. Theoretical analysis and experimental results show that the AGCluster algorithm has high clustering accuracy for all kinds of distributed data.

This work is mainly supported by the National Natural Science Foundation of China (Nos. 61772131 and 62072156).