基于 Tsallis 熵的近似差分隐私 K-means 算法

杨舒丹[1,2], 李 男[1,2], 郑文娟[3], 杜启明[1]

[1]战略支援部队信息工程大学 网络空间安全学院 郑州 中国 450000
[2]数学工程与先进计算国家重点实验室 郑州 中国 450000
[3]32142 部队 保定 中国 071000

摘要 利用 K-means 算法对用户信息进行聚类时, 存在隐私泄露的风险。差分隐私保护技术可提供严格的隐私保护, 但目前大多数满足差分隐私的 K-means 算法在处理多维数据时, 存在随机选择质心和噪声添加不均衡的问题, 因而导致聚类结果不理想。为此, 本文提出一种基于 Tsallis 熵的近似差分隐私 K-means 算法。针对质心选择的随机性问题, 提出 Tsallis 熵对属性赋权的策略来优化对象间的欧氏距离, 然后对比各对象到唯一随机初始质心的赋权欧式距离来确定其余初始质心, 使算法在减少随机选择初始质心的同时, 提高模型准确率; 在此基础上, 针对噪声添加不均衡的问题, 提出一种能够平衡信噪比的隐私预算分配策略, 然后对迭代质心加入高斯扰动, 使算法在不增加计算复杂度的情况下满足 (ε,δ) – 差分隐私保护, 同时提升扰动结果的准确性; 最后在四个真实数据集上对算法进行有效性评价。实验结果表明, 所提出的算法能够在保证用户隐私安全的同时实现高效用的聚类。

关键词 近似差分隐私; 高斯机制; Tsallis 熵; K-means 聚类; 数据挖掘
中图法分类号 TP391 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2023.07.08

An approximate differential privacy K-means algorithm based on Tsallis entropy

YANG Shudan[1,2], LI Nan[1,2], ZHENG WenJuan[3], DU Qiming[1]

[1] Department of Cyberspace Security Academy, Strategic Support Force Information Engineering University, Zhengzhou 450000, China
[2] State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450000, China
[3] People's Liberation Army 32142 Unit, Baoding 071000, China

Abstract When using the K-means algorithm to cluster user information, there is a risk of privacy leakage. Differential privacy protection technology can provide strict privacy protection. However, most of the current K-means algorithms that meet differential privacy have problems of random selection of centroids and imbalanced noise addition when processing multi-dimensional data, which leads to unsatisfactory clustering results. To this end, this paper proposes an approximate differential privacy K-means algorithm based on Tsallis entropy. Aiming at the randomness problem of centroid selection, a strategy of attribute weighting by Tsallis entropy is proposed to optimize the Euclidean distance between objects, and then the weighted Euclidean distance between each object and the unique random initial centroid is compared to determine the remaining initial centroids. This allows the algorithm to reduce the random selection of the initial centroid while improving the accuracy of the model; on this basis, a privacy budget allocation strategy that can balance the signal-to-noise ratio is proposed for the problem of noise imbalance, and then add Gaussian perturbation to the iterative centroid, so that the algorithm satisfies (ε,δ) – differential privacy protection without increasing the time complexity, and at the same time improves the accuracy of the perturbation results; finally, the effectiveness of the algorithm is evaluated on four real data sets. Experimental results show that the proposed algorithm can achieve efficient clustering while ensuring user privacy.

Key words approximate differential privacy; Gaussian mechanism; Tsallis entropy; K-means clustering; data mining

1 引言

随着信息技术的快速发展, 人类社会正处于万物互联的大数据时代, 数据作为大数据时代的核心要素, 已成为技术创新的基石、经济繁荣的催化剂[1]。为了充分释放数据的价值, 以 K-means 聚类分析[2]为代表的数据挖掘[3]算法得到了广泛的应用。然而这些算法在发现知识的同时, 也在无时无刻地

通讯作者: 李男, 博士, 副教授, Email:279136411@qq.com。
本课题得到国家自然科学基金资助项目(No. 62472447)资助。
收稿日期: 2022-01-01; 修改日期: 2022-03-08; 定稿日期: 2023-04-20

收集着人们的个人信息, 如果在应用中不注重保护个人信息, 则攻击者很容易通过重构攻击[4-6]、成员推理攻击等手段获取这些信息, 进而造成个人隐私的泄露。因此, 在聚类过程中获得高质量数据分析的同时, 应该不断加强对个人隐私的保护, 只有这样才能促进大数据应用健康发展。

差分隐私[7]定义了一个与背景知识无关的量化分析模型, 并提供严格的数学证明, 具有强大的隐私保障水平, 已被广泛应用于数据挖掘领域[8-11]。已有文献提出满足差分隐私的 K-means 聚类算法: Blum 等人[12]最早提出了基于差分隐私的 K-means 算法, 并通过 SuLQ 平台[13]实现; Frank 等人[14]基于 PINQ 平台设计和实现了满足差分隐私的 K-means 聚类算法; Thanh 等人[15]通过直接关注数据点, 提出估计误差的隐私感知 K-means 算法, 以阻止对手推断未知数据点; Yu 等人[16]根据数据点的分布密度选择初始中心点, 提出剔除离群值的差分隐私 K-means 算法; 胡等人[17]以 K-means++的结果作为输入值, 通过交替的方式来确定初始质心, 提出了基于 Laplace 机制的 K-means 差分隐私算法; Dwork[18]等人提出了指数递减隐私预算分配策略的差分隐私 K-means 算法。

然而上述方法大多要求在传统的 K-means 算法的聚类过程添加 Laplace 噪声来实现差分隐私保护, 存在以下两方面局限:

(1) 传统的 K-means 算法由于忽略了数据对象各属性对聚类结果发挥的不同聚类作用, 同时随机选择初始质心, 导致质心的质量良莠不齐, 进而造成聚类效果不稳定的情况时常发生。

(2) 聚类数据集大多是多属性的数据样本, 采用 Laplace 机制处理多变量隐私问题时, 会产生过多的噪声[19], 同时随着迭代次数的不断增加, 质心趋于稳定, 等量的噪声会增加隐私预算的开销以及引起质心较大的波动, 进而影响聚类结果。

为了解决上述问题, 本文提出了一种基于 Tsallis 熵的 K-means 聚类近似差分隐私保护算法。针对问题 1, 利用能度量系统有序化程度且能对不同的数据集均有较好的适应性的 Tsallis 熵[20-22]对样本属性赋权, 并采用赋权欧氏距离作为相似性度量的依据, 然后随机选取一个初始质心, 计算其余数据样本到此质心的赋权欧氏距离以衡量其被选为初始质心的概率, 最后采用轮盘法依次选出其余初始质心。针对问题 2, 设计对迭代数越大的质心多分配隐私预算的策略, 相对降低噪声量, 提高信噪比。然后采用能更大程度地保留数据的可用性、更适合处理多维数据的高斯机制对迭代质心添加噪声, 使算法满足 (ε,δ) – 差分隐私保护。

2 背景知识

2.1 Tsallis 熵

信息熵是对系统有序化的一种度量, 但对于常见的幂律厚尾分布族[23]却不能通过信息熵来刻画[24], 而以 $\sum_{i=1}^{n} p(x_i)^q$ 为例的依赖于概率幂的熵度量可以做到。因此, Tsallis 熵为代表的泛化熵被提出。

定义 1. Tsallis 熵。Tsallis 熵通过一个可调节的参数 q 将其应用范围拓展到所谓的非拓展系统中[25], 具体定义形式如下:

$$S_q(X) = -\sum_{i=1}^{n} p(x_i)^q \ln_q p(x_i) \tag{1}$$

其中 $\ln_q(x) = (x^{1-q}-1)/(1-q)$, $q \neq 1, x \geqslant 0$ 是 $q-\log$ 函数[26], $X \in \{x_1,...,x_n\}$ 是一个随机变量 $p(x_i)$ 是取值为 x_i 的概率, $q \in \mathrm{R}$ 是一个可调节的参数, 当 $q \to 1$ 时, Tsallis 熵 $S_q(X)$ 退化为信息熵; 当 $q=2$ 时, Tsallis 熵 $S_q(X)$ 退化为基尼系数。

2.2 近似差分隐私

2.2.1 近似差分隐私的理论基础

定义 2. ε – 差分隐私。设随机算法 $M: X^n \to Y$。考虑任意两个只有一个条目不同的相邻数据集 $X, X' \in X^n$, 如果, 对于所有相邻的 X, X' 和所有的 $T \subseteq Y$, 满足:

$$pr[M(X) \in T] \leqslant e^{\varepsilon} pr[M(X') \in T] \tag{2}$$

则 M 满足 ε – 差分隐私(即满足纯差分隐私), 其中其中 M 的选择是随机的, ε 为隐私预算, 与隐私保护水平呈负相关, 当 ε 较小时, 隐私保护能力较强, 但此时噪声大, 容易导致数据失真, 反之亦然。在实际应用场景中, 当对高维向量、矩阵等复杂数据进行处理时, 为保证数据的可用性, 往往要求噪声不会过大, 因此 Dwork 等人提出了近似差分隐私[27]。

定义 3. (ε,δ) – 差分隐私。设随机算法 $M: X^n \to Y$。满足:

$$pr[M(X) \in T] \leqslant e^{\varepsilon} pr[M(X') \in T] + \delta \tag{3}$$

则 M 是 (ε,δ) – 差分隐私(即满足近似差分隐私), 其中 δ 为松弛因子, 且 $0 < \delta << 1$。当 δ 越小时, 隐私性损失量越微弱, 当 δ=0 时, M 的隐私保护能力退化为公式(2)。

定义 4. 敏感度。设 $f: X^n \to \mathrm{R}^n$。f 的 ℓ_2 敏感

性是:

$$\Delta_2 = \max_{X,X'} \|f(X) - f(X')\|_2 \tag{4}$$

其中 X, X' 是相邻数据集。

2.2.2 近似差分隐私的实现机制

高斯机制是实现近似差分隐私保护的主要方法, 其通过向数据中添加符合高斯分布的随机变量达到隐私保护的目的。高斯分布 $N(\mu, \sigma^2)$ 的密度函数为:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5}$$

其中均值和方差分别为 μ 和 σ^2, 该分布的可视化图像如图 1 所示:

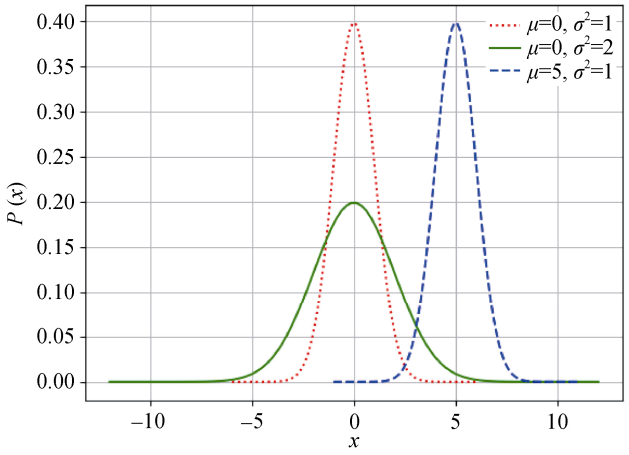

图 1 高斯分布

Figure 1 Gaussian distribution

定义 5. 高斯机制。设 $f: X^n \to \mathrm{R}^n$。高斯机制定义为:

$$M(X) = f(X) + (Y_1, ..., Y_k) \tag{6}$$

其中, Y_i 是独立的 $N(0, \sigma^2)$ 高斯分布函数, 而 $\sigma^2 \geqslant 2\ln(1.25/\delta)\Delta_2^2/\varepsilon^2$, 可见, 该噪声大小取决于查询函数的 ℓ_2 敏感性、给定的隐私预算 ε, 以及松弛因子 σ。在同样的隐私保护预算分配下, 高斯机制添加的噪声要小于拉普拉斯机制[27], 因此更大程度地保留了数据的可用性, 适于处理多维数据。在算法实现中, 松弛因子 σ 的大小也会大大影响模型的准确率。

2.2.3 近似差分隐私的性质

性质 1. 序列组合性。设 $M = (M_1, ..., M_k)$ 是一个算法序列, 其中 M_i 是 $(\varepsilon_i, \delta_i)$ – 差分隐私, 那么 M 满足 $(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i)$ – 差分隐私。即: 序列组合性允许对数据集进行多次查询操作, 最终总的隐私预算为各个查询操作所消耗的隐私预算总和。

3 基于 Tsallis 熵的 K-means 聚类差分隐私算法设计

本文建立了一个兼具隐私保护性和高可用性的差分隐私 K-means 聚类模型。图 2 是模型的概述, 其中数据挖掘器不能直接访问原始数据库, 但可以向差分隐私接口提交查询和相应的隐私预算, 差分隐私接口将返回一个保持统计特征的不敏感数据。然后, 数据挖掘者利用本文提出的基于 Tsallis 熵的 K-means 聚类近似差分隐私方案, 根据查询结果构建差分隐私模型, 并在不暴露原始数据集敏感信息的情况下发布该模型。

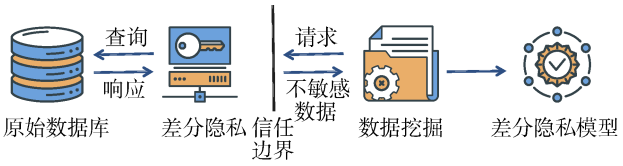

图 2 满足差分隐私的数据挖掘框架概述

Figure 2 A system overview of differential privacy data mining

本文提出的基于 Tsallis 熵的 K-means 聚类近似差分隐私方案将围绕高质量的聚类结果和可靠的隐私保护两个方面展开。针对 K-means 算法随机选择初始质心, 导致初始质心质量不稳定, 影响聚类的准确率这一问题, 本文设计了基于 Tsallis 熵对属性赋权的初始质心选择法对初始质心进行优化。在此基础上, 针对在聚类过程中, 攻击者恶意查询某个子类中用户信息这一隐私泄露现象, 本文依据迭代质心的特点设计隐私预算划分策略, 从信噪比的角度出发, 解决迭代数越大的质心其信噪比相对越小的问题, 然后向迭代质心添加高斯噪声, 以满足 (ε, δ) – 差分隐私, 从而实现对用户敏感信息的保护。文中假定所有的数据源可信, 但在进行数据挖掘过程中存在第三方攻击者的前提下, 设计了基于 Tsallis 熵的 K-means 聚类差分隐私算法, 目的是实现数据的可用且不可见, 为了方便和形式化描述, 将其命名为 DPTK-means。

DPTK-means 算法的基本框架如图 3 所示:

在图 3 中, 当攻击者查询某个子类中用户信息时, 反馈的查询结果不再是真实的用户聚类信息, 而是经过特定的近似差分隐私机制操作后返回的随机扰动结果。本文采用高斯机制实现隐私保护, 与不

考虑隐私保护的情况相比, 本文设计的 K-means 聚类模型是基于特定扰动的统计结果来构建。

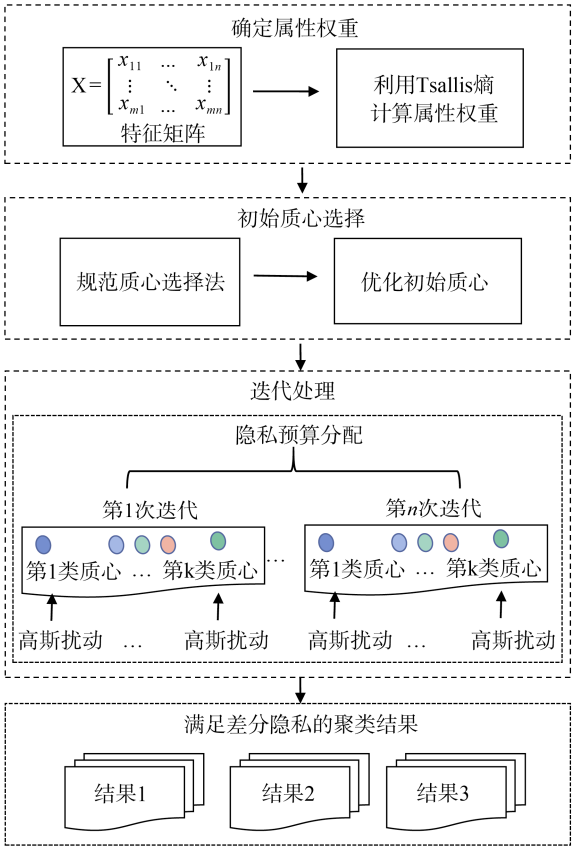

图 3 算法框架

Figure 3 The frame of algorithm

3.1 基于 Tsallis 熵的 K-means 算法

传统 K-means 算法选择初始质心具有随机性, 容易导致算法陷入局部最优, 难以获得稳定而精确的聚类结果。已有文献利用信息熵对聚类样本各属性赋权以提高聚类精度[28], 但并未考虑信息熵对不同数据集的适应性问题。Tsallis 熵具有度量系统有序化程度, 能适应不同类型数据等优点, 本文将其用于计算各属性权重, 通过对样本属性值进行适当缩放, 以获得较优的聚类结果。因此设计基于 Tsallis 熵的改进 K-means 算法, 重点考虑数据分布问题, 为了方便和形式化描述, 将其命名为 TK-means。在具体实现过程中, 围绕初始质心的选择这一问题进行展开: 首先采用 Tsallis 熵法计算各属性的权重, 再根据所设计初始质心算法得到优化后的 k 个初始质心, 具体的算法框架如图 3 部分所示。

3.1.1 确定属性权重

使用改进熵值法计算属性权重的目的是优化初始质心, 从而提高聚类准确度。在统计学中, 信息熵等指标的缺点在于没有一种指标总能在各种数据集上获得最好的效果, 缺乏对数据集的适应性。本文采用 Tsallis 熵统一常用的度量指标, 同时通过可调参数 q 来适应各种数据集, 这种思想第一次在文献[22]中被提出, 被用于决策树分类问题[20-21]。本文根据不同属性对聚类的作用不同, 采用 Tsallis 熵法计算各属性的权重, 以便为无序且不同规模的数据集聚类提供依据。

采用 Tsallis 熵法确定属性权重后, 本文将 TK-means 权值算法总结为如下:

Step1: 构造数据属性值矩阵。

$$X=\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \tag{7}$$

其中, n 为样本总数, m 为样本维数, x_{ij} 为第 $j\in\{1,...,m\}$ 维属性所对应的第 $i\in\{1,...,n\}$ 个数据样本的属性值。

Step2: 计算样本的属性值比重。

$$p_{ij}=x_{ij}/\sum_{i=1}^{n}x_{ij} \tag{8}$$

Step3: 计算第 j 维属性的 Tsallis 熵。

$$S_{q,j}=-\frac{1}{\ln n}\sum_{i=1}^{n}p_{ij}^{\,q}\ln_q p_{ij} \tag{9}$$

其中, $S_{q,j}$ 为衡量属性不纯度的指标, 因此对于给定的 j, $S_{q,j}$ 越小, 样本的属性值之间的差值越大, 该属性的聚类作用越大, 属性越重要, 反之成立。当属性值都相等时: $S_{q,j}=S_{q,\max}=1$, 该属性的聚类效果为零。

Step4: 计算第 j 维属性的权值。

$$w_j=\frac{1-S_{q,j}}{\sum_{j=1}^{m}(1-S_{q,j})} \tag{10}$$

Step5: 计算赋权欧氏距离。

$$d_w(x_a,x_b)=\sqrt{\sum_{j=1}^{m}w_j(x_{aj}-x_{bj})^2} \tag{11}$$

该方法的优点是利用第 j 维属性的权值适当放大或缩小相应的属性值, 使权值大的属性聚类作用更大, 权值小的属性聚类作用更小, 从而更能真实反映实际聚类时各属性发挥的作用。

3.1.2 初始质心选择法

为了避免 K-means 算法随机选择初始质心, 导

致每次聚类结果的差异较大。本文根据不同子类质心的相对距离与聚类效果成正比的原则, 设计了新的初始质心选择策略, 具体步骤如下:

Step1: 在数据集 X 中随机选择一个初始聚类质心 $c(T_i)$ [29]。

Step2: 计算数据集 X 中每个数据样本 x_i 与当前已有的初始质心的最短赋权欧氏距离聚类中心最短距离: $d_w(x_a,x_b)$。

Step3: 计算数据样本 x_i 被选为初始质心的概率: $p=d_w(x_a,c(T_i))^2/\sum_{x_b\in T_j} d_w(x_a,c(T_i))^2$, 并利用轮盘法(各个个体被选中的概率与其适应度大小成正比)选出下一个质心。

Step4: 重复 Step2, 3 直到选出 k 个初始质心。

以上通过改善初始质心选择方式和利用 Tsallis 熵权值法的预选初始质心方式, 目的是使得两两子类的质心的距离相对分散, 类中的数据样本呈现相似的效果。因本文针对中心化差分隐私问题进行展开, 即假定数据源可信, 没有敌手从数据输入端对聚类结果进行干扰, 隐私泄露只存在于数据挖掘过程中。因此算法的 worst-case 准确性将不会成为敌手攻击的缺口。此外, 初始质心的选择只与数据域相关, 与具体的数据无关, 因此对聚类过程中的迭代质心添加高斯噪声进行隐私保护并不产生影响。

3.2 满足近似差分隐私的 TK-means 算法

为保证聚类过程中, 用户信息的隐私安全, 引入差分隐私, 通过添加噪声对结果进行扰动, 使得攻击者在查询过程中无法区分从相邻数据集返回的结果。目前, 构建满足差分隐私保护的 K-means 聚类模型需要解决以下两个关键性问题。

(1) 如何设计预算策略, 以便于在将隐私保护水平进行量化表示的同时, 减少预算的浪费。

在聚类过程中, 随着迭代次数不断增加质心趋于稳定, 添加等量的噪声将引起质心较大的波动, 会影响聚类结果。因此隐私预算设置和划分至关重要。本文从信噪比平衡的角度出发, 依据迭代质心的特点设计隐私预算划分策略, 即: 当迭代次数增加时, 为查询分配更大的隐私预算, 以达到更小的扰动规模, 可以在很大程度上保护数据集的真实信息。因此, 本文在设计预算划分策略的重点就在于对迭代数越大的质心多分配隐私预算, 以相对降低噪声量, 提高信噪比。

(2) 如何实现对迭代质心的隐私保护, 在实现用户数据不可见的同时, 保证数据良好的可用性。

通常数据隐私性和可用性呈现明显的负相关关系, 而 $\varepsilon-$ 的差分隐私, 往往需要付出更多的隐私代价。因此本文采用 $(\varepsilon,\delta)-$ 差分隐私, 通过对迭代质心添加高斯噪声, 实现对个人隐私保护和整体统计趋势可用性之间的平衡。

3.2.1 建立隐私适量预算分配规则

如果将相同隐私预算产生的噪声加入到每次迭代生成的质心中, 那么在不同的迭代中, 扰动过程引起的信噪比明显会有相当大的不均衡性。因此, 本文根据迭代次数 H, 为第一次迭代分配预算份额为 $\frac{1}{H}$, 后续每次迭代的预算份额依次增大隐私预算的份额, 相应为: $\frac{1}{H-1},\frac{1}{H-2},...,1$, 以减少噪声的添加。在聚类过程中耗费的总隐私预算总份额为:

$$S=\frac{1}{H}+\frac{1}{H-1}+\frac{1}{H-2}+...+1 \tag{12}$$

其中, 迭代次数为 H。

单位隐私预算为:

$$\varepsilon_0=\frac{\varepsilon}{S} \tag{13}$$

其中, ε 为总隐私预算。

第 i 次迭代所需的隐私预算为:

$$\varepsilon_i=\frac{\varepsilon_0}{H-(i-1)} \tag{14}$$

3.2.2 迭代质心的隐私保护策略

差分隐私保护的 K-means 聚类模型通过添加噪声对结果进行扰动, 本文采用高斯机制实现对迭代质心的隐私保护。原因在于在同样的隐私保护预算分配下, 高斯机制对比拉普拉斯机制减少了噪声量。

某类别数据样本的质心为: $c(T_i)=\frac{1}{|T_i|}\sum_{x_b\in T} x_b$, 其中 $|T_i|$ 表示 T_i 中数据样本的个数。为了保护数据隐私, 需对每个中心点加入一定量的噪声, 加噪的方式是通过每一维度单独加噪的方式, 得到加噪后的中心点, 本文采用差分隐私添加噪声, 所添加的噪声为一个随机常量, 并未涉及乘除运算, 因此并不会增加算法的时间复杂度, 具体的证明过程参见 3.3.1 节。

$$c^*(T_i)=\frac{\sum_{x_b\in T_i} x_b+(Y_1,...,Y_d)}{|T_i|+Y_{k+1}} \tag{15}$$

其中 d 为样本维度。每一个 Y_i 服从独立同分布, 根据高斯机制可知, 所添加的噪声满足: $N(0,2\ln(1.25/\delta)/\Delta_2^2\varepsilon_i^2)$, 其中 Δ_2 为敏感度, 其具体

取值证明过程详见 3.3.3 节。

因此本文所提出的 DPTK-means 算法设计流程是:

首先根据所设计预选初始质心策略得到初始质心, 然后对这些质心进行迭代优化, 在迭代优化的过程中加入适应性的高斯噪声, 得到最终的聚类结果。具体步骤如下所示:

Step1: 采用 Tsallis 熵法计算聚类数据样本各属性的权值。

Step2: 根据预选初始质心算法得到 k 个初始质心。

Step3: 对所有数据对象进行扫描, 根据其与 k 个初始质心的欧氏距离将其分类到最近的聚类中。

Step4: 重新计算 k 个类的质心。

Step5: 根据迭代次数在每个质心上加入 $N(0,2\ln(1.25/\delta)\Delta_2^2/\varepsilon_i^2)$ 的高斯噪声, 其中 ε_i 表示第 i 次迭代所需的隐私预算。

Step6: 重复执行 Step3、4、5, 直到完成算法要求的迭代次数时, 循环终止。

3.3 DPTK-means 算法有效性及安全性分析

3.3.1 算法复杂度分析

在 DPTK-means 算法设计中, 有三个关键步骤, 包括选择最优质心、隐私保护预算分配和噪声计算。已知样本总数为 m, 迭代次数为 t。

(1) 在 K-means 聚类算法中, 为了选择最优质心, 需要根据样本到质心的距离进行最优簇划分, 从理论上来说该划分过程的计算复杂度为为 $O(mkt)$, 其中 k 为质心个数, 一般情况下 $k,t \ll m$。

(2) 隐私预算分配的复杂度是 $O(1)$, 分配中只需要对数据样本做常数操作, 而不涉及乘除运算。因此, 隐私预算的分配并不会增加算法复杂度。

(3) 与非隐私保护的 K-means 算法相比, 在计算迭代质心时, 噪声估计步骤带来了额外的计算量。事实上, 未添加噪声的迭代质心的计算也会产生大量的运算, 而在这基础上增加简单的常数计算, 并不会增加算法复杂度, 这也正是差分隐私区别于其他传统加密技术的优势所在。因此, 相对于非隐私保护的 K-means 聚类算法而言, DPTK-means 算法的效率不会下降。

综上分析, DPTK-means 算法的计算复杂度与非隐私保护 K-means 聚类算法大致相同, 即 $O(mkt)$, 因此, 本文提出的 DPTK-means 算法能保证与非隐私保护情况下相同量级的计算复杂度。

3.3.2 算法安全性分析

算法中的随机函数若能够提供满足高斯分布的噪声, 则可以为查询结果提供近似差分隐私保护。本文提出的 DPTK-means 算法通过实现高斯机制, 为聚类结果提供了近似差分隐私保护, 即在聚类过程中添加适当的满足高斯分布的噪声到质心。安全性证明过程如下:

对于每次迭代过程, 每个数据点只会影响 d 个求和查询与一个计数查询, 因此本文将一次迭代的输出看做 $d+1$ 维。由于对每个数据点做了 $[-r,r]$ 的范围限制, 因此全局敏感度为 $\Delta_2=\|d\cdot r+1\|_2$。

设单次迭代的查询函数为 $f:X^d\to X^d\times N$, $X,X'\in X^d$ 是待聚类数据中的一对相邻数据集, 这两个数据集分别被划分到不同的簇: $\{c^*(T_1),\ldots,c^*(T_k)\}$ 和 $\{c^{*'}(T_1),\ldots,c^{*'}(T_k)\}$, 其中存在 $c^*(T_j)\neq c^{*'}(T_j)$。设 $f(c^*(T_j))=f(c^{*'}(T_j))+v$。考虑噪声幅值 $x-N(0,\sigma^2)$, 则隐私损失随机变量[6]分布为:

$$\ln\left(\frac{pr\left[M\left(c^*\left(T_j\right)\right)=f\left(c^*\left(T_j\right)\right)+x\right]}{pr\left[M\left(c^{*'}\left(T_j\right)\right)=f\left(c^{*'}\left(T_j\right)\right)+x\right]}\right)=\ln\left(\frac{e^{-\frac{\|x\|_2^2}{2\sigma^2}}}{e^{-\frac{\|x+v\|_2^2}{2\sigma^2}}}\right)=\frac{x_jv_j}{\sigma^2}+\frac{v_j^2}{2\sigma^2} \tag{16}$$

因此在一次迭代过程中, 隐私损失随机变量的分布均值为 $\frac{v_j^2}{2\sigma^2}$, 方差为 $\frac{v_j^2}{\sigma^2}$。为了方便引入变量 $Z-N(0,1)$, 则隐私损失随机变量为: $\frac{\Delta_2}{\sigma}\cdot Z+\frac{\Delta_2^2}{2\sigma^2}$, 通过证明隐私损失随机变量的绝对值超过 ε_i 的概率最多为 δ_i 来证明 $(\varepsilon_i,\delta_i)-$差分隐私。因此, 在一次迭代过程中超过 ε_i 的概率重写为: $pr\left[|Z|\geqslant\frac{\varepsilon\sigma}{\Delta_2}-\frac{\Delta_2}{2\sigma}\right]$, 将其上界定为:

$$pr[|Z|\geqslant t-\frac{\varepsilon}{2t}] \tag{17}$$

其中 $\sigma=\frac{2\Delta t}{\varepsilon}$, 当 $pr[|Z|\geqslant t]$ 时, 也符合标准高斯尾部界限的, 由于 $\sigma^2=2\ln(1.25/\delta)\Delta_2^2/\varepsilon_i^2$, 因此, $pr[|Z|\geqslant t]\leqslant\delta_i$, 从而证明了每次迭代过程满足

(ε_i,δ_i) – 差分隐私。根据 (ε_i,δ_i) – 差分隐私的序列组合性可知, 整个聚类过程满足 $(\sum_{i=1}^{k}\varepsilon_i,\sum_{i=1}^{k}\delta_i)$ – 差分隐私。

4 实验及分析

本节通过实验对提出的 DPTK-means 算法进行验证分析, 实验采用 Python 进行编程实现, 物理环境为: 处理器 Intel(R) Core(TM) i7-9700 CPU @ 3.00GHz; 内存 64.0 GB; 操作系统 Windows 10。

4.1 实验数据集及评价指标

实验中采用 UCI 数据库[30]中的四个真实聚类数据集: 3D-roadNetwork、Adult、Wine-quality、Statlog。以 3D-roadNetwork 数据集为例, 包含 434,874 个实例, 其形式为{*ID*, 经度, 纬度, 海拔}。其余三个数据集的信息如表 1 所示。在以下数据集上的实验表明了本文所提出的算法在性能上的提升和在差分隐私约束下的有效性。

表 1 数据集

Table 1 Datasets

数据集	样本数	维数(属性数)
3D-roadNetwork	434874	4
Adult	48841	6
Wine-quality	4898	12
Statlog	946	18

在本文的实验中, 使用以下两个指标来评估性能: 标准化互信息(Normalized Mutual Information 简称 *NMI*)[31]和相对误差(Relative Error 简称 *RE*)[15]。*NMI* 是评价性能的常用指标。*NMI* 值越接近于 1, 意味着算法结果与标准结果之间的相似度越高, 聚类质量越好。*RE* 测量估计的质心相对于未实现隐私保护 K -means 的实际质心之间的误差。*RE* 值越小, 结果越相似, 差分隐私聚类的可用性越高。具体来说, 给定原始质心 $c(T_i)$ 和扰动质心 $c^*(T_i)$ 所在簇, 平均扰动计算为:

$$c^*(T_i)=\frac{1}{k}\frac{\left\|c^*(T_i)-c(T_i)\right\|}{|T_i|} \tag{18}$$

4.2 对比算法

为了评估本文所提出的 DPTK-means 算法的有效性, 将其性能与最常用的几个 K-means 聚类算法进行比较:

K-means: 传统 K-means 聚类算法。

DPK-means: 在传统 K-means 聚类算法的基础上, 为每次迭代均添加等量的拉普拉斯噪声。

EDPK-means[16]: 传统 K-means 聚类算法, 在每次迭代中添加拉普拉斯噪声, 并使用指数递减的隐私预算分配。

4.3 实验结果与分析

4.3.1 TK-means 算法聚类性能分析

实验 1: 各属性的 Tsallis 熵权值分布。

本次实验的目的是分析各属性 Tsallis 熵权值对聚类作用的影响程度。计算数据集各属性对应的 Tsallis 熵权值如图 4 所示, 为了简单方便, 当固定 q 时, 取 $q=2$。

在图 4 中, 横坐标分别代指四个数据集的属性集, 以 3D-roadNetwork 数据集为例, 其四个属性分别为: ID, 经度, 纬度, 海拔。通过图(a)中各属性的 Tsallis 熵权值可知, 属性 2 的权值为 0.288,而属性 4 的权值为 0.21, 同比下降了 7.8%。由此可以看出: 在聚类的过程中, 同一数据样本的不同属性的聚类作用存在显著差异, 因此在进行聚类时应加以区分, 这正是传统 K-means 算法聚类结果常常不理想的原因之一。

实验 2: TK-means 算法的有效性分析。

本次实验的目的是对比 TK-means 算法和 K-means 算法的聚类效果, 本文采用 *NMI* 评价指标对 TK-means 算法和 K-means 算法的聚类质量进行评估, 实验结果是取 $q=2$ 时, 在 10 次实验后统计聚类情况计算平均值得到。具体结果如图 5 所示:

图 5 展示了, 两种算法在不同数据集上的实验结果。在 Adult 数据集上聚类时, 当 $k=5$ 聚类效果达到最佳, 聚类评价指标 *NMI* 由 0.821 提高到 0.893; 在 Statlog 数据集上聚类时, 当 $k=4$ 聚类效果达到最佳, 聚类评价指标 *NMI* 由 0.644 提高到 0.705; 在 Wine-quality 数据集上聚类时, 当 $k=2$ 聚类效果达到最佳, 聚类评价指标 *NMI* 由 0.82 提高到 0.9; 在 3D-roadNetwork 数据集上观察 k 从 1 到 12 的聚类结果的变化情况, 不难发现 TK-means 算法对比传统的 K-means 算法聚类准确率在整体趋势上均有所提升, 当 $k=9$ 时 K-means 算法的聚类结果达到最优 $NMI=0.725$, 但此时基于 Tsallis 熵改进的 K-means 算法的 $NMI=0.778$, 同比提高了 5.3%, 具体的可视化分布如图 6 所示, 其他三个数据集可视化效果类似。当 $k=12$ 时基于 Tsallis 熵改进的 K-means 算法的聚类结果达到最优 $NMI=0.8$, 意味着算法结果与标准结果之间的相似度较高, 这充分说明了文中 TK-means 算法的有效性。

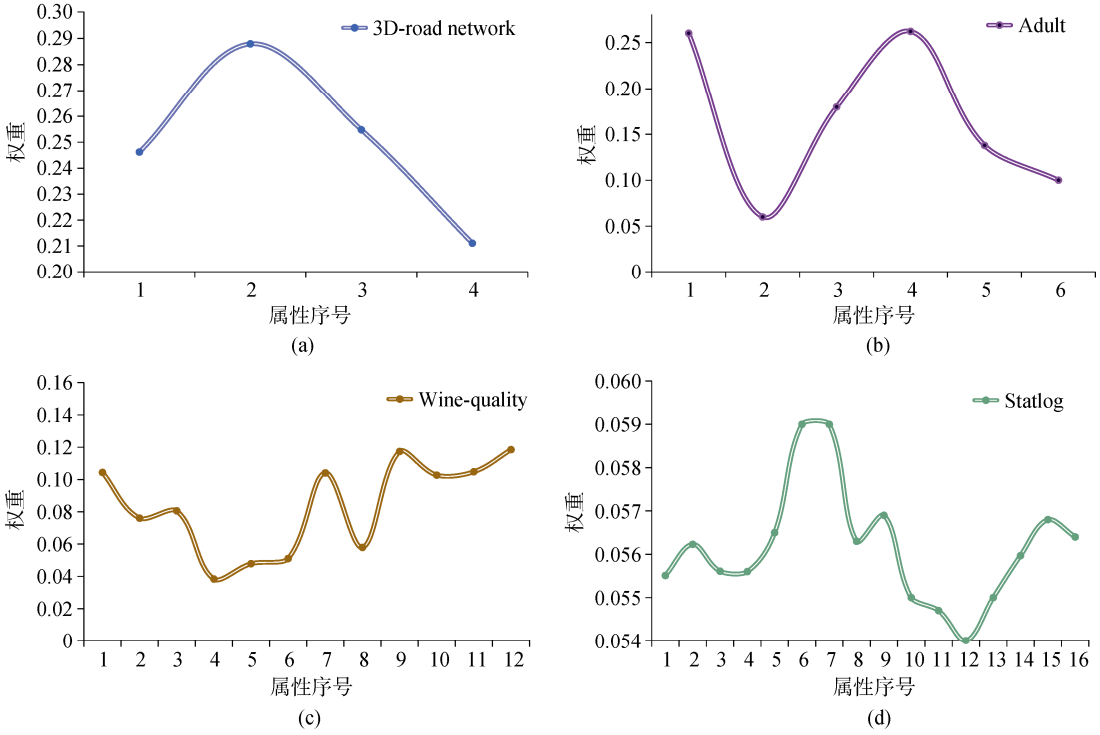

图 4　各属性 Tsallis 熵权值

Figure 4　Attribute Tsallis entropy weight

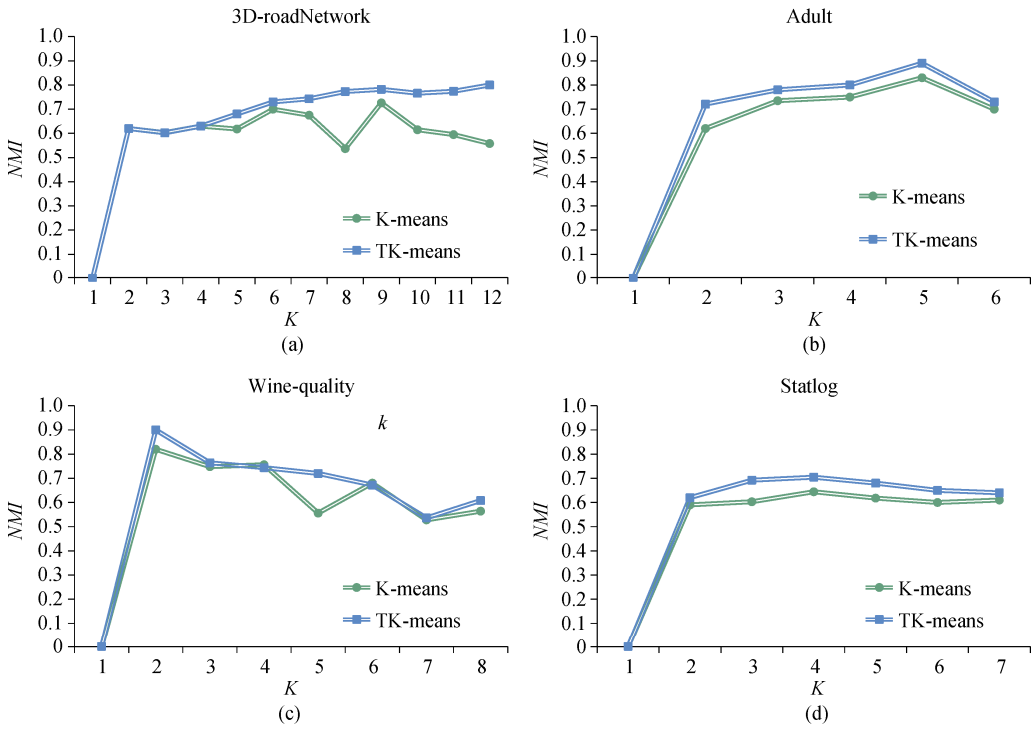

图 5　聚类结果

Figure 5　The clustering results

在图 6 中横纵坐标 $x1, x2$ 表示数据对象分布范围, 通过观察图 6 可知, 在 3D-roadNetwork 数据集上, 当 $k=9$ 时(图中红色圆点为质心 k 的数目), 在 K-means 算法进行聚类的结果中(图 6 左), 存在两个聚类质心有相距较近的问题, 这是由于选择初始质心的随机性所导致。而在 TK-means 算法进行聚类的结果中(图 6 右), 两两子类的质心的距离相对分散, 类内中的数据样本呈现均匀分布的效果。

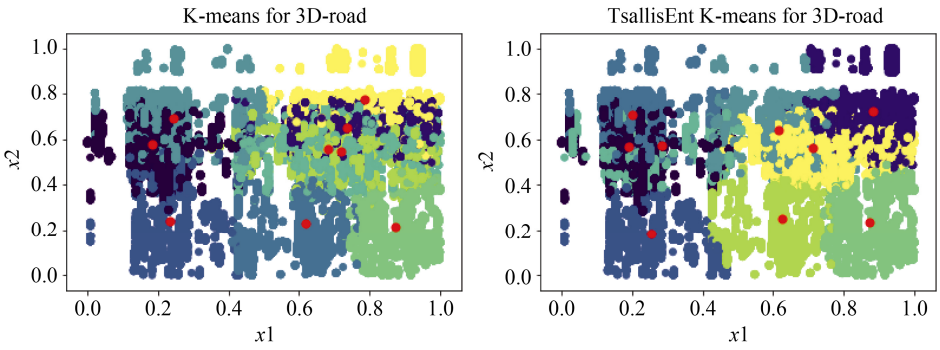

图 6　聚类结果可视化分布

Figure 6　Visual distribution

实验 3: TK-means 算法对不同数据集的适应性分析。

此次试验目的是为了全面评估 Tsallis 熵中参数 q 的影响, 本实验中 q 在 $[0.1, 10.0]$ 的范围内以 0.1 的步长遍历整个区间。对于每个选定的 q, 本文通过在 TK-means 算法上的聚类效果, 对其性能进行评估, 此处以 3D-roadNetwork 数据集在 $k=9$ 时, 不同值参数 q 对 Tsallis 熵分裂准则的影响的情况为例, 其实, 其他数据集也表现出类似的趋势。实验的迭代次数最大值为 100, 最终取 10 次实验的平均值, 图 7 展示了具体的聚类结果。

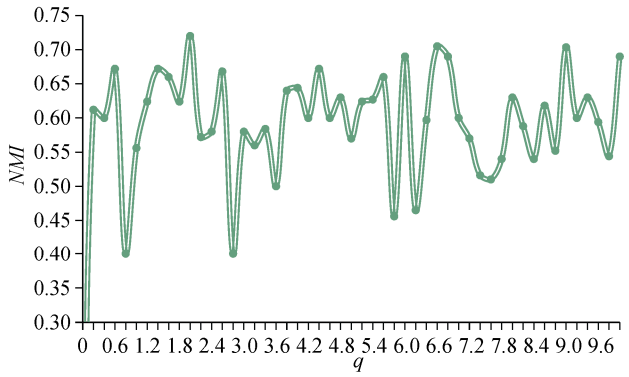

图 7　参数 q 的影响

Figure 7　The influence of parameter q

图 7 直观展示了 TK-means 算法在 3D- road-Network 数据集上选用不同参数 q 对聚类结果的影响。显然 *NMI* 对 q 的变化很敏感, 例如: 当 $q=1.8$ 时, 3D-roadNetwork 数据集的聚类质量最好, 并不是当 $q\to 1$ 时, Tsallis 熵 $S_q(X)$ 退化为信息熵, 也不是当 $q=2$ 时, Tsallis 熵 $S_q(X)$ 退化为基尼系数, 从而体现 Tsallis 熵通过可调参数 q 以适应各种数据集。本文通过选择不同的 q 实现不同的目标。总之, 实验结果表明参数 q 的确对聚类准确率有影响, 这也反映了 TK-means 算法的适应性和灵活性。

4.3.2　差分隐私约束下的有效性

对实验过程中选用的参数做简要介绍。首先, 四个数据集聚类 k 值的选择, 通过实验 2 可知四个数据集的最佳聚类 k 值, 具体信息如图表 2 所示:

表 2　最佳聚类数

Table 2　Optimal cluster number

数据集	k
3D-roadNetwork	12
Adult	5
Wine-quality	2
Statlog	4

其次针对不同的 ε, 由于噪声的随机性, 因此在以上三个算法对每个 ε 执行 10 次, 最终取评价指标的平均值。最后由于两个对比算法涉及松弛因子 σ, 因此本次实验中对于 DPTK-means 算法取 $\sigma=0.01$。

实验 4: 隐私保护预算对聚类结果的影响。

本次实验的目的是对比分析实现隐私保护后各个算法在数据集上的聚类结果的可用性和准确性。采用 *RE* 评价指标对 DPTK-means 算法和两个对比算法的质心偏离情况进行评估, 采用 *NMI* 评价指标对三个算法的聚类质量进行评估。通过观察随着隐私预算 ε 的变化, *RE* 值和 *NMI* 值的分布情况来判断噪声对算法聚类结果的影响程度, 实验结果分别如图 8、图 9 所示:

由图 8 可以看出, 在不同的数据集上, DPTK-means 算法的聚类结果的 *RE* 值要低于两个对比算法, 这意味着 DPTK-means 算法的聚类结果更优。通过文章第二节介绍差分隐私的定义可知, 当隐私预算

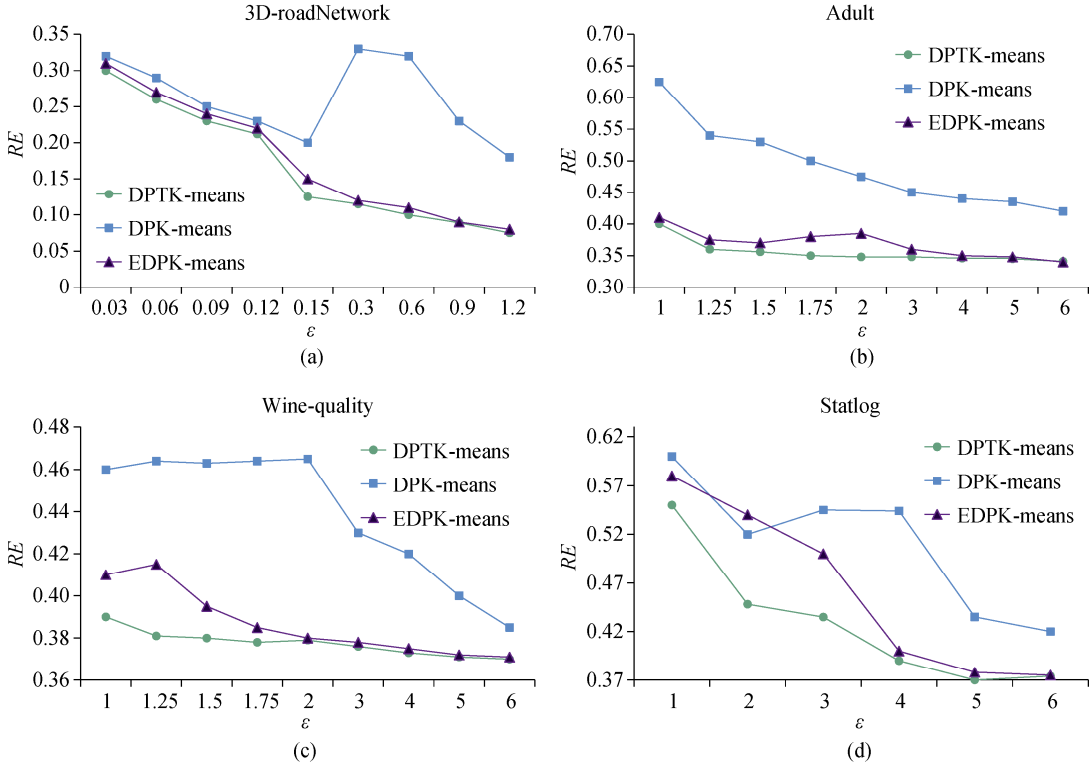

图 8　质心的偏离情况

Figure 8　Deviation of the center of mass

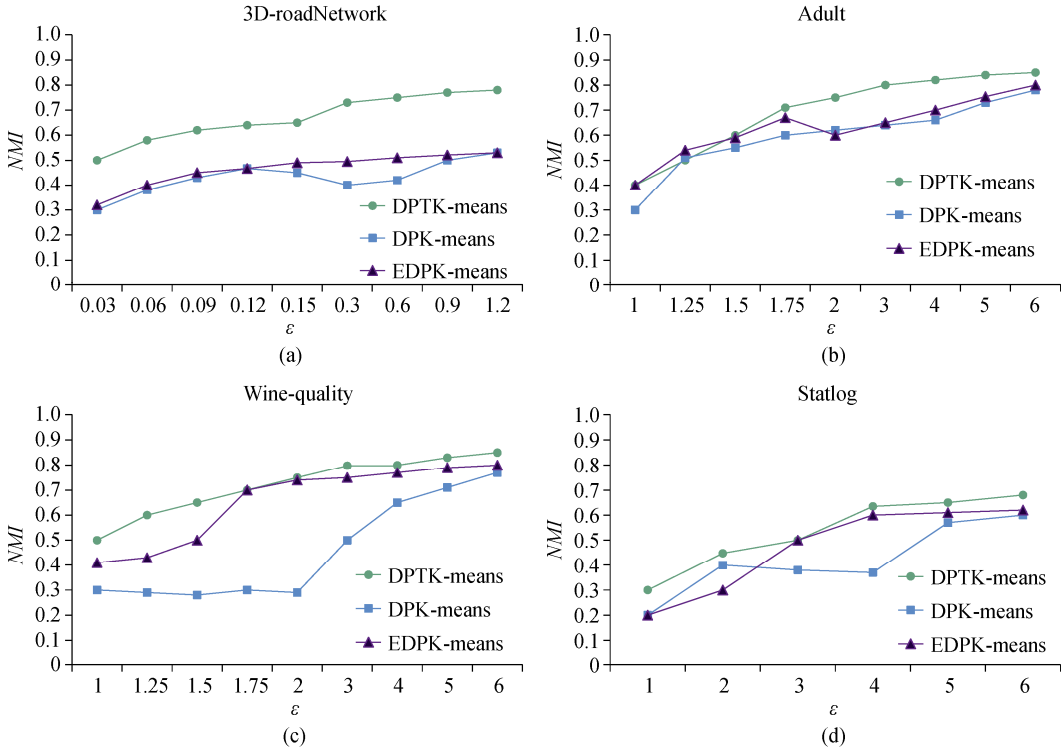

图 9　满足差分隐私条件下的聚类准确性

Figure 9　Clustering accuracy under the condition of differential privacy

越小时, 添加的噪声越多, 从而隐私保护的程度越高。随着隐私预算 ε 的增加, DPTK-means 的 *RE* 评价曲线得到了较明显的下降, 说明在较少的隐私预算下 DPTK-means 可以获得更高的可用性。与四种不同的数据集相比发现, DPTK-means 比两种对比算法更能适应高维数据。具体来说, 在 4 维数据集

3D-roadNetwork(图(a))和 6 维数据集 Adult 上(图(b)), DPTK-means 与 EDPK-means 具有相似的隐私保护性能。然而当隐私预算增加(添加的噪声减弱)时, DPK-means 算法的 *RE* 值仍然较高且在 3D-roadNetwork 数据集上波动较大, 通过分析可知, 是由于每一次迭代过程中加入等量的噪声导致算法性能较差; 在 10 维数据集 Wine-quality(图(c))和 18 维数据集 Statlog 上(图(d)), DPK-means 和 EDPK-means 的 *RE* 评价曲线整体均呈现较高的现象, 这表明其聚类质心发生较强的偏离。相比之下, DPTK-means 的 *RE* 值较低, 表明聚类质心只发生了较弱的偏离, 因此结果仍然具有较高的可用性上。通过上述分析可知, 由于差分隐私的引入, DPK-means 算法、EDPK-means 算法、DPTK-means 算法与不加入差分隐私的 K-means 算法($RE=0$)对比均呈现出聚类质心偏离, 即聚类效果变差的趋势, 因此这些算法的聚类结果可用性有所降低, 这表明要实现对数据隐私的保护需要牺牲一定的聚类精度。

由图 9 可知, 在四个数据集上, DPTK-means 算法的 *NMI* 评价值均高于两个对比算法, 这意味着 DPTK-means 算法的聚类质量更好。具体来说, 在数据量大, 维数少的 3D-roadNetwork 数据集上(图(a)), 较小的隐私预算就可保证聚类结果的准确性, 在这个数据集上, DPTK-means 算法的结果与对比算法相比有明显的提高, 随着 ε 的增大 DPTK-means 算法与两个对比算法在 *NMI* 上的差异比较明显。在数据量和维数中等的 Adult 数据集上(图(b)), 当 $\varepsilon<1.75$ 时, 三种隐私保护算法的 *NMI* 评价曲线相似, 当 $\varepsilon>1.75$ 时, DPTK-means 算法的 *NMI* 值与两个对比算法相比增长趋势平缓稳定。在数量中等, 维数较高的 Wine-quality 数据集上(图(c)), DPTK-means 和 EDPK-means 算法的效果比较接近, 均优于 DPK-means 算法。在高维数的 Statlog 数据集上(图(d)), 在相同的隐私保护水平下, DPTK-means 算法的聚类质量与两个对比算法相比呈现较好的趋势。通过上述分析可知, DPTK-means 算法的聚类准确性比 DPK-means、EDPK-means 算法较理想的原因在于, 首先此算法对初始质心进行了优化, 其次在聚类过程中加入比拉普拉斯机制能能保护数据可用性的高斯机制, 最后设计了适应于不同迭代质心的隐私预算分配策略, 减少了噪声分配的不均衡性, 从而提高了聚类准确性。

实验 5: 松弛因子对聚类结果的影响。

在 DPTK-means 算法高斯机制的构建中, 向迭代质心加入的噪声为 $N(0,2\ln(1.25/\delta)/\Delta_2^2\varepsilon_i^2)$。可见, 松弛因子 σ 影响噪声尺度, 进而影响聚类效果。因此, 本次实验观察隐私预算 ε 与松弛因子 σ 的互动对准确性的影响。通过随着隐私预算 ε 的变化, 观察不同松弛因子 σ 作用下 *NMI* 值的分布情况来判断其对算法聚类效果的影响程度, 结果如图 10 所示:

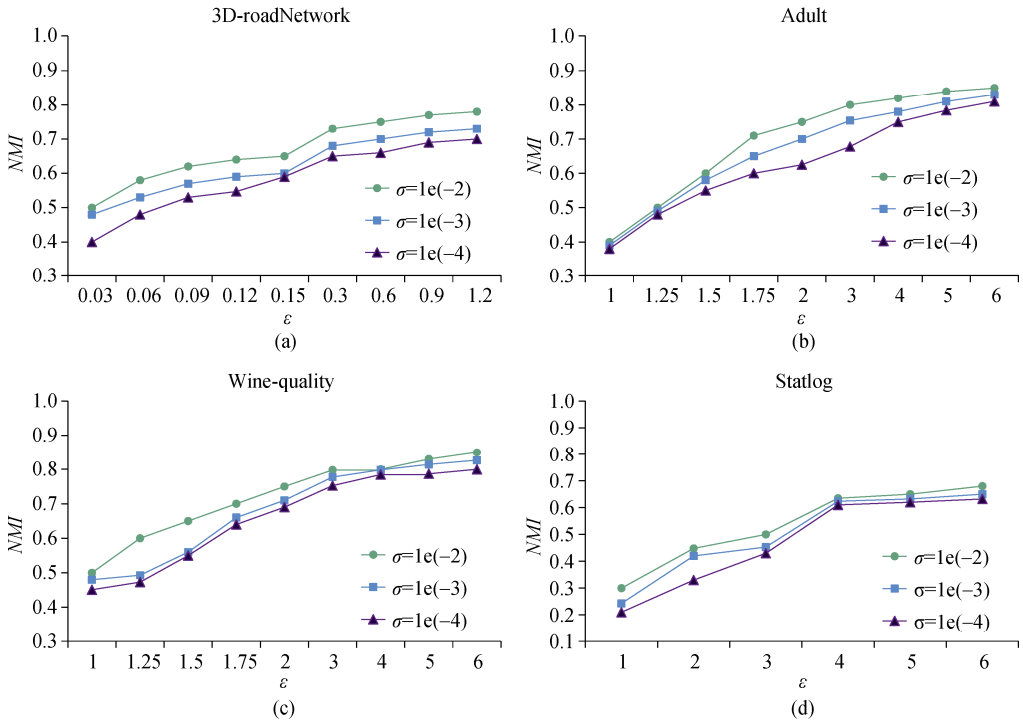

图 10 σ对聚类准确性的影响

Figure 10 The influence of σ on clustering accuracy

由图 10 可以看出, DPTK-means 算法的聚类质量在很大程度上受松弛因子 σ 的影响。在数据规模、维度不同的四个数据集上, 当 $\sigma = 0.01$ 时, 算法的聚类结果最好。当 $\sigma = 0.0001$ 时, 算法的 *NMI* 评价指标整体明显下降, 这意味着算法的聚类质量变差, 虽然随着隐私预算 ε 的增加, 算法聚类效果有所提升, 但仍然较低。

综上, 在相同隐私保护级别下, DPTK-means 算法在不同数据集上的性能均优于对比算法; 此外, 该算法受数据规模和维度的约束较小。这意味着本文所提出的算法在一定程度上达到了对数据的“可用且不可见”, 即实现了数据隐私保护和可用性之间的平衡。

5 结束语

本文首次提出了一种基于 Tsallis 熵的近似差分隐私 K-means 机制, 与标准的差分隐私聚类机制不同, 该机制在保护数据隐私时, 付出的隐私代价更小, 也更适合处理多维数据。通过对传统的 K-means 聚类算法进行改进, 增强了聚类作用, 然后在聚类质心优化过程的每次迭代中加入适应性的高斯扰动, 在不增加算法复杂度的同时, 保护了集群中间结果的数据隐私, 从而提高 DPTK-means 算法的性能, 实现数据隐私保护和可用性之间的平衡。实验结果表明, 该机制比传统的差分隐私方案在处理多维数据具有更好的性能。

参考文献

[1] Wang L, Xu J M. Application of Distributed K-Means Clustering Algorithm in Micro-Blog Hot Topic Discovery[J]. *Computer Simulation*, 2020, 37(8)121-125.
(王林, 许郡蒙. 分布式 K-means 聚类在微博热点主题发现的应用[J]. *计算机仿真*, 2020, 37(8)121-125).

[2] Ariosto Serna L, Alejandro Hernández K, Navarro González P. A K-Means Clustering Algorithm: Using the Chi-Square as a Distance[C]. *International Conference on Human Centered Computing*, 2019: 464-470.

[3] Kousika N, Premalatha K. An Improved Privacy-Preserving Data Mining Technique Using Singular Value Decomposition with Three-Dimensional Rotation Data Perturbation[J]. *The Journal of Supercomputing*, 2021, 77(9): 10003-10011.

[4] Garfinkel S, Abowd J M, Martindale C. Understanding Database Reconstruction Attacks on Public Data[J]. *Communications of the ACM*, 2019, 62(3): 46-53.

[5] Salem A, Bhattacharya A, Backes M, et al. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning[C]. *The 29th USENIX Conference on Security Symposium*, 2020: 1291-1308.

[6] Cohen A, Nissim K. Linear Program Reconstruction in Practice[J]. *Journal of Privacy and Confidentiality*, 2020, 10(1): 1-13.

[7] Dwork C. Differential privacy[C]. *Automata, Languages and Programming: 33rd International Colloquium ICALP 2006*, 2006: 1-12.

[8] Tang J, Korolova A, Bai X L, et al. Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12[EB/OL]. 2017: arXiv: 1709.02753. https://arxiv.org/abs/1709.02753.

[9] Ding B L, Kulkarni J, Yekhanin S. Collecting Telemetry Data Privately[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 3574-3583.

[10] Shen H, Liu Y J, Xia Z, et al. An Efficient Aggregation Scheme Resisting on Malicious Data Mining Attacks for Smart Grid[J]. *Information Sciences*, 2020, 526: 289-300.

[11] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response[C]. *The 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014: 1054-1067.

[12] Blum A, Dwork C, McSherry F, et al. Practical Privacy: The SuLQ Framework[C]. *The twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2005: 128-138.

[13] Menezes A. Topics in Cryptology - CT-RSA 2005: The Cryptographers' Track at the RSA Conference 2005, San Francisco, CA, USA, February 14-18, 2005. Proceedings[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[14] McSherry F D. Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis[C]. *The 2009 ACM SIGMOD International Conference on Management of data*, 2009: 19-30.

[15] Nguyen T D, Gupta S, Rana S, et al. Privacy aware k-means clustering with high utility[C]. *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference*, 2016: 388-400.

[16] Yu Q Y, Luo Y L, Chen C M, et al. Outlier-Eliminated *k*-Means Clustering Algorithm Based on Differential Privacy Preservation[J]. *Applied Intelligence*, 2016, 45(4): 1179-1191.

[17] Hu C, Yang G, Bai Y L. Clustering Algorithm in Differential Privacy Preserving[J]. *Computer Science*, 2019, 46(2): 120-126.
(胡闯, 杨庚, 白云璐. 面向差分隐私保护的聚类算法[J]. *计算机科学*, 2019, 46(2): 120-126.)

[18] Dwork C. A Firm Foundation for Private Data Analysis[J]. *Communications of the ACM*, 2011, 54(1): 86-95.

[19] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy[J]. *Foundations and Trends® in Theoretical Computer Science*, 2013, 9(3-4): 211-407.

[20] Wang Y S, Xia S T. Unifying Attribute Splitting Criteria of Decision Trees by Tsallis Entropy[C]. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017: 2507-2511.

[21] Wang Y S, Xia S T, Wu J. A Less-Greedy Two-Term Tsallis Entropy Information Metric Approach for Decision Tree Classification[J]. *Knowledge-Based Systems*, 2017, 120: 34-42.

[22] Tsallis C. Possible Generalization of Boltzmann-Gibbs Statistics[J]. *Journal of Statistical Physics*, 1988, 52(1): 479-487.

[23] Altmann E G, Gerlach M. Statistical Laws in Linguistics[EB/OL].

2015: arXiv: 1502.03296. https://arxiv.org/abs/1502.03296

[24] Maszczyk T, Duch W. Comparison of Shannon, Renyi and Tsallis entropy used in decision trees[C]. *Artificial Intelligence and Soft Computing-ICAISC 2008: 9th International Conference Zakopane*, 2008: 643-651.

[25] Tsallis C, Baldovin F, Cerbino R, et al. Introduction to Nonextensive Statistical Mechanics and Thermodynamics[EB/OL]. 2003: arXiv: cond-mat/0309093. https://arxiv.org/abs/cond-mat/0309093

[26] Umarov S, Tsallis C, Steinberg S. On a *q*-Central Limit Theorem Consistent with Nonextensive Statistical Mechanics[J]. *Milan Journal of Mathematics*, 2008, 76(1): 307-328.

[27] Dwork C, Kenthapadi K, McSherry F, et al. Our Data, Ourselves: Privacy via Distributed Noise Generation[C]. *The 24th annual international conference on The Theory and Applications of Cryptographic Techniques*, 2006: 486-503.

[28] Yuan F Y, Zhang X C, Luo S B. Accurate Property Weighted K-Means Clustering Algorithm Based on Information Entropy[J]. *Journal of Computer Applications*, 2011, 31(6)1675-1677
（原福永，张晓彩，罗思标. 基于信息熵的精确属性赋权 K-means 聚类算法[J]. *计算机应用*, 2011, 31(6)1675-1677）

[29] Xi Z, Sun Xiang'e. Research on Naive Bayes Ensemble Method Based on Kmeans++ Clustering[J]. *Computer Science*, 2019, 46(B06): 439-441, 451.
（钟熙, 孙祥娥. 基于 Kmeans++聚类的朴素贝叶斯集成方法研究[J]. *计算机科学*, 2019, 46(B06): 439-441, 451.）

[30] Lichman M. UCI machine learning repository[EB/OL]. University of California, Irvine, 2013.

[31] Manning C D, Raghavan P, Schutze H. Introduction to Information Retrieval IIR 19 : Web Search Basics[J]. 2008, 10.1017/CBO 9780511809071(19): 385-404.

杨舒丹 于 2020 年在电子科技大学成都学院网络工程专业获得学士学位。现在信息工程大学网络空间安全专业攻读硕士学位。研究领域为大数据与云计算安全。研究兴趣包括: 大数据安全、差分隐私。Email: 3189046684@qq.com

李男 于 2018 年在信息工程大学计算机科学与技术专业获得博士学位。现任信息工程大学副教授、硕士生导师。研究领域为大数据与云计算安全。研究兴趣包括高性能计算、大数据分析。Email: 279136411@qq.com

郑文娟 于 2012 年在原解放军外国语学院获得军事学硕士学位。现任 32142 部队软件助理工程师。研究领域为软硬件管理维护。研究兴趣包括: 数据库管理。Email: yiyi1987802@sina.com

杜启明 于 2019 年在信息工程大学计算机科学与技术专业获得学士学位。现在信息工程大学计算机可与与技术专业攻读硕士学位。研究领域为大数据挖掘。研究兴趣包括大数据分析、自然语言处理。Email: qimingducst@163.com