文章编号:1671-6833(2023)06-0019-06

改进的密度峰值聚类算法的差分隐私保护方案

葛丽娜[1,2,3], 陈园园[1], 王 捷[1,2], 王 哲[1]

(1.广西民族大学 人工智能学院,广西 南宁 530006;2.广西民族大学 网络通信工程重点实验室,广西 南宁 530006;3.广西民族大学 广西混杂计算与集成电路分析设计重点实验室,广西 南宁 530006)

摘 要:针对改进的密度峰值聚类(AdDPC)算法在计算局部密度时产生的隐私泄露问题以及算法的一次分配策略,提出一种改进的密度峰值聚类算法的差分隐私保护方案。该方案在算法计算局部密度的过程中添加 Laplace 随机噪声,使得即使攻击者拥有最大背景知识,也无法通过添加或者删除数据集中的某一点来获取相应的信息,从而利用差分攻击获取目标数据点的信息,达到保护隐私数据的目的,并且在分配非聚类中心点时引入可达定义改进 AdDPC 算法的分配策略,避免因为一次分配策略导致数据点分配错误的问题。实验对比了 DP-rcCFSFDP 算法、AdAPC-rDP 算法、IDP *K*-means 算法的 *F*-Measure 和 *ARI*,结果表明:当隐私预算大于 1.5 时,所提算法的 *F*-Measure 和 *ARI* 优于其他算法,所提算法能够在保护敏感数据的同时保证数据的可用性。

关键词:密度峰值; 差分隐私; 随机噪声; 聚类算法

中图分类号:TP391 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2023.03.010

聚类分析是数据挖掘中的关键技术之一,其主要思想是将数据集划分为不同的簇,使得同一簇内的数据相似度较高、不同簇间的数据相似度较低[1]。聚类分析的广泛应用[2-4]给数据拥有者带来了巨大的利益。但是,若这些数据被攻击者截获,并恶意利用,将对数据的提供者产生不利影响,甚至危及人身财产安全。因此,如何在聚类分析中保护用户的隐私安全成为了聚类分析的一个热门方向[5]。

差分隐私是 2006 年由 Dwork[6] 针对传统的隐私保护模型无法应对新型攻击方式而提出的一种具有严格的可证明性的隐私保护模型。该模型通过对原始数据或者数据的发布结果添加符合要求的随机噪声,使得攻击者即使已知除了目标用户的数据以外的其他所有数据,也无法获取目标用户的数据,达到保护用户的隐私数据的目的。

李杨等[7]为了提高 DP *K*-means 算法的可用性,降低算法中由于添加了随机噪声而对初始聚类中心点的选取造成的影响,先将数据集均分为 *m* 个子数据集,再对各子集添加噪声后的聚类中心点进行计算,并将这些点作为初始聚类中心点,提出了 IDP *K*-means 算法,但是算法中 *m* 的值需要人为选取。

密度峰值聚类(clustering by fast search and find density peaks, DPC)算法[8]是 2014 年提出的一种基于密度的聚类算法,该算法仅需输入 1 位参数、不存在迭代,且能够自动发现类簇中心,实现任意形状数据的高效聚类,因此广泛应用于各个领域。但是,DPC 算法也存在一些不足,如输入参数截断距离的值按照经验选取、聚类中心点需要人为选取以及非聚类中心点分配采取一步策略等,并且算法存在聚类时易导致敏感数据泄露的问题。

针对改进的 DPC 算法的非聚类中心点分配以及算法泄露隐私的问题,本文引入可达概念,对非聚类中心点的分配策略进行改进,并在算法计算数据点局部密度的步骤中添加符合要求的 Laplace 随机噪声,使得算法满足差分隐私保护。

1 相关知识

1.1 密度峰值聚类算法

密度峰值聚类(DPC)算法基于以下 2 个假设:①聚类中心被低密度邻居数据点包围;②聚类中心

收稿日期:2022-11-03;**修订日期**:2022-12-01

基金项目:国家自然科学基金资助项目(61862007);广西自然科学基金资助项目(2020GXNSFBA297103)

作者简介:葛丽娜(1969—),女,广西环江人,广西民族大学教授,博士,主要从事信息安全、物联网和智能计算研究,E-mail:66436539@qq.com。

引用本文:葛丽娜,陈园园,王捷,等. 改进的密度峰值聚类算法的差分隐私保护方案[J]. 郑州大学学报(工学版),2023,44(6):19-24. (GE L N, CHEN Y Y, WANG J, et al. Differential privacy protection scheme of adaptive clustering by fast search and find of density peaks[J]. Journal of Zhengzhou University (Engineering Science), 2023, 44(6): 19-24.)

与另一个密度更高的数据点之间的距离足够远。DPC算法的主要步骤包括:密度距离计算、聚类中心选取、剩余数据点分配。

Step 1 密度距离计算。DPC算法根据式(1)计算出数据点 x_i 的局部密度 ρ_i,当数据集规模较小时由式(2)计算 ρ_i。由式(3)计算其到更高密度数据点的距离 δ_i。

$$\rho_i = \sum_j \chi(d_{ij} - d_c),\text{其中}\chi(a) = \begin{cases} 1, a < 0; \\ 0, a \geqslant 0; \end{cases} \tag{1}$$

$$\rho_i = \sum_j \exp\left(-\frac{d_{ij}^2}{d_c^2}\right); \tag{2}$$

$$\delta_i = \begin{cases} \min\limits_j d_{ij}, \rho_j > \rho_i; \\ \max\limits_j d_{ij}, \text{其他}。 \end{cases} \tag{3}$$

式中:d_c 为截断距离;d_{ij} 为数据点 x_i 到数据点 x_j 之间的欧氏距离。d_c 的取值:d_c 的值使得数据点的平均近邻个数为整个数据集数据点总数的1%~2%[9]。

Step 2 聚类中心选取。聚类中心选取包括2种方法:①决策图法是算法根据局部密度和距离生成的以局部密度为横轴、距离为纵轴的决策图,再根据决策图人工选取最佳聚类中心的方法;②公式法[8]是根据式(4)生成决策度量 γ_i,再将决策度量进行降序排序,选取前 k 个值对应的数据点作为聚类中心的方法。

$$\gamma_i = \rho_i \times \delta_i。 \tag{4}$$

Step 3 剩余数据点分配。将剩余的非聚类中心点与距离该点最近并且拥有更高局部密度值的数据点归为一类。

改进的密度峰值聚类(adaptive clustering by fast search and find of density peaks, AdDPC)算法针对DPC算法选取聚类中心点需要人为参与的问题,将加权思想引入决策度量计算中,避免了人为参与选取聚类中心点。AdDPC算法的聚类步骤如下。

Step 1 计算出数据点间的欧氏距离矩阵。

Step 2 分别根据式(2)和式(3)计算出数据点 x_i 的局部密度 ρ_i 和距离 δ_i 的值,并对 ρ 和 δ 做归一化处理,生成决策图。

Step 3 根据式(4)计算出决策度量 γ 的值,归一化处理得到 γ^*,将 γ^* 降序排序。

Step 4 根据式(5)计算出 γ^* 的斜率变化率 kit_i,生成聚类中心点判别图:

$$kit_i = (i - \eta)k_i, i = 1,2,\cdots,50。 \tag{5}$$

式中:η 为加权因子。

Step 5 由式(6)计算出聚类中心点和非聚类中心点的分界点,并将 $\gamma_1^*, \gamma_2^*, \cdots, \gamma_m^*$ 对应的数据点作为聚类中心点,m 作为类簇数:

$$\gamma_m^* = \arg\max_i(kit_i)。 \tag{6}$$

Step 6 将剩余数据点分配到拥有更高局部密度且距离其最近的点所在的类簇中。

1.2 差分隐私

差分隐私[6]是一种拥有严格数学定义以及可证明性的信息保护技术,该模型的提出基于一种假设:数据攻击者具有最大背景知识,即攻击者知道除了需要保护的敏感信息以外的所有信息。

定义1 ε-差分隐私[6]。设随机算法 M,对于任意2个邻近数据集 D 和 D',若满足式(7),则称算法 M 满足 ε-差分隐私。

$$Pr[M(D) \in S] \leqslant e^{\varepsilon} \cdot Pr[M(D') \in S]。 \tag{7}$$

式中:$Pr[\cdot]$ 表示事件发生的概率;$S \in Range(M)$,$Range(M)$ 表示随机算法 M 的所有输出的集合;ε 为隐私预算,表示隐私保护程度,其值与隐私保护程度成反比。

定义2 全局敏感度[10]。设有查询函数 $f: D \to \mathbf{R}^d$,对于任意一对邻近数据集 D 和 D',函数 f 的全局敏感度 Δf 为

$$\Delta f = \max_{D,D'} \| f(D) - f(D') \|_1。 \tag{8}$$

式中:$\| \cdot \|_1$ 为1-范数距离。全局敏感度 Δf 表示函数 $f(\cdot)$ 在邻近数据集上查询结果的差异程度,差分隐私中所添加噪声的大小受全局敏感度的影响。

定义3 Laplace机制[6]。设有随机函数 $f(\cdot)$,$f(D)$ 为其对数据集 D 查询返回的输出结果,则Laplace机制定义为

$$M(D) = f(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right)。 \tag{9}$$

式中:Δf 表示敏感度;ε 为隐私预算;$M(D)$ 表示随机算法;$Lap\left(\frac{\Delta f}{\varepsilon}\right)$ 表示服从Laplace分布的随机噪声。

Laplace分布的概率密度函数的表达式为

$$Lap\left(x, \frac{\Delta f}{\varepsilon}\right) = \frac{1}{2\frac{\Delta f}{\varepsilon}} \exp\left(-\frac{|x|}{\frac{\Delta f}{\varepsilon}}\right)。 \tag{10}$$

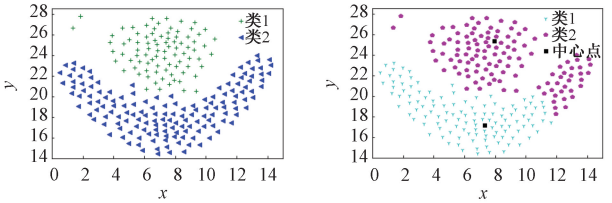
2 改进的密度峰值聚类算法的差分隐私保护方案

由AdDPC算法的步骤可以看出,该算法隐私泄露的关键是局部密度 ρ。若攻击者拥有除了目标数据点以外的所有数据点信息,则攻击者可以通过局部密度 ρ 的定义公式,计算出目标数据点与其他数据点的距离,进而推出目标数据点的相关信息。若

攻击者删除数据集中目标数据点,则其余数据点的局部密度 ρ 会相应减小,得到其余数据点在删除目标数据点前后的局部密度差,获取目标数据点与其余数据点间的距离,从而推断出目标数据点的真实信息。

由对 AdDPC 算法的隐私泄露问题分析可知,当数据拥有者发布算法聚类的局部密度 ρ 时,可能会导致隐私数据的泄露。而本文算法在局部密度的计算过程中添加 Laplace 随机噪声,使得该算法满足差分隐私保护的要求,进而达到保护隐私信息的目的。

AdDPC 算法的聚类原理是寻找具有较大局部密度,同时将距离其他更高局部密度点较远的数据点作为聚类中心,再将其余数据点划分到局部密度比该点大且相距较近的点所在的类簇中。当数据集分布较为均匀,或者数据集中某一类簇存在几个局部密度较大且距离其他局部密度更高的点较远的点,AdDPC 算法即使选取了正确的聚类中心点,也可能会出现部分非聚类中心点归类错误的问题。在 AdDPC 算法中,非聚类中心点的分配是采用一步分配策略,容易引起“多米诺效应”,即密度大的数据点若分配错误会导致密度较小的数据点归类错误的问题,从而影响算法的聚类性能,如图 1 所示。图 1(a)为 Flame 数据集的标准分类图,图 1(b)为 AdDPC 算法对 Flame 数据集的错误聚类结果图。



(a) Flame数据集的标准划分　　(b) AdDPC算法的错误聚类结果

图 1　Flame 数据集聚类

Figure 1　Clustering result of Flame dataset

对于 AdDPC 算法的非聚类中心点分配问题,本文引入 DP-rcCFSFDP 算法[11]中的可达定义,将其应用于非聚类中心点的归类,从而提高 AdDPC 算法的聚类效果,基于此,提出一种基于局部密度加噪的密度峰值聚类差分隐私保护算法(differential privacy preserving algorithm of clustering by fast search and find density peaks based on local density adding noise, DP-DPCL)。DP-rcCFSFDP 算法的相关定义如下。

定义 4　邻域点。给定数据点 p_i,在以 p_i 为圆心、E 为半径的邻域中的所有数据点称为 p_i 的邻域点。

定义 5　可达。给定一串数据点 $s_1, s_2, \cdots, s_n$,若每个数据点 s_{i+1} 为数据点 s_i 的邻域点时,则称数据点 s_n 到数据点 s_i 可达。

DP-DPCL 算法具体步骤如下。

Step 1　对输入数据集预处理,初始化数据点间的欧氏距离,根据式(2)计算出数据点的局部密度。

Step 2　选取合适的隐私预算 ε,根据局部密度 ρ 的敏感度,生成符合 Laplace 分布的噪声,并将随机噪声添加至局部密度 ρ 中,记为 l_ρ。

Step 3　将 l_ρ 降序排序,并按照式(3)计算距离 δ,根据 l_ρ 和 δ 生成决策图。

Step 4　根据式(4)计算决策度量 γ,归一化后得到 γ^*,将 γ^* 降序排序。

Step 5　根据式(5)计算 γ^* 的斜率变化率 kit_i,生成聚类中心点判别图。

Step 6　由式(6)计算聚类中心点和非聚类中心点的分界点,并将 $\gamma_1^*, \gamma_2^*, \cdots, \gamma_m^*$ 对应的数据点作为聚类中心点。

Step 7　将剩余数据点分配到与其距离更近且拥有更高局部密度的数据点所在类簇,生成初始聚类结果。

Step 8　对初始聚类结果进行遍历,若存在数据点对可达且不在一个类簇中,则将局部密度值小的数据点归类为局部密度值大的数据点所在类,得到最终的聚类结果。

在 DP-DPCL 算法中,局部密度每次所添加的噪声是随机生成的,而由于加噪导致输出的局部密度值不同,这样使得攻击者难以依靠获得的局部密度差值推导出目标数据点的真实信息。由于数据点到密度更高点的距离 δ 的值也会受到局部密度值变化的影响,从而进一步降低了隐私泄露的风险。

在 DP-DPCL 算法中使用的差分隐私保护技术中的 Laplace 机制是由算法的敏感度 Δf 以及隐私预算 ε 控制生成的满足 Laplace 分布的随机噪声。敏感度是指在数据集中增加或减少任意一个数据点对聚类结果的最大影响。对于任意一对邻近数据集,对 2 个数据集进行局部密度计算时,根据局部密度定义公式可知,数据点局部密度的最大差值为 1,因此,局部密度的敏感度 $\Delta f = 1$。

3　实验及结果分析

3.1　算法的差分隐私性证明

DP-DPCL 算法的差分隐私性证明如下:

假设数据集 D 和 D' 为一对邻近数据集,$R(D)$ 和 $R(D')$ 分别为 AdDPC 算法未添加噪声时作用于

数据集 D 和 D'的聚类结果,A 表示任意一种聚类结果。$S(D)$和 $S(D')$分别表示在 AdDPC 算法中添加 Laplace 噪声后的作用于数据集 D 和 D'的输出结果,B 表示添加噪声后的任意一种聚类结果。则由式(7)~(10)可得

$$\frac{Pr[S(D)\in B]}{Pr[S(D')\in B]}=\frac{\frac{1}{2b}\exp\left(-\frac{\varepsilon\mid A-R(D)\mid}{\Delta f}\right)}{\frac{1}{2b}\exp\left(-\frac{\varepsilon\mid A-R(D')\mid}{\Delta f}\right)}=$$

$$\exp\left(\frac{\varepsilon\{\mid A-R(D')\mid-\mid A-R(D)\mid\}}{\Delta f}\right)\leqslant$$

$$\exp\left(\frac{\varepsilon\mid R(D)-R(D')\mid}{\Delta f}\right)\leqslant$$

$$\exp\left(\frac{\varepsilon\parallel R(D)-R(D')\parallel_1}{\Delta f}\right)\leqslant\exp(\varepsilon)。$$

DP-DPCL 算法的差分隐私性得证。

3.2 实验结果分析

本实验由 Python 语言编程实现,实验环境为 Windows10 的 64 位操作系统、8 GB 内存、2.30 GHz 处理器。实验用于测试的数据集信息如表 1 所示。本实验中使用 F-Measure[12]和调整兰德指数 ARI[13]作为评价指标。用 $A=(A_1,A_2,\cdots,A_K)$表示数据集的真实划分,$B=(B_1,B_2,\cdots,B_{K'})$表示聚类算法对数据集的聚类结果。

表 1 数据集信息

Table 1 Dataset information

数据集	数据总数	属性维度	类簇数
Iris[14]	150	2	3
WDBC[15]	569	32	2
Aggregation[16]	788	2	7

F-Measure 值是同时考虑了精确率(*precision*)和召回率(*recall*)的一种评价指标。精确率用于衡量算法聚类结果的精确程度,召回率用于衡量聚类结果的完备程度,F-Measure 指标可以较为全面地区分算法的聚类能力。精确率、召回率和 F-Measure 分别按照式(11)、(12)、(13)来计算,F-Measure 取值为[0,1],值越大聚类效果越优。一般的聚类结果分布情况有 4 类:N_a 表示在真实划分 A 和聚类结果 B 中同属一个类簇的数据对的数目;N_b 表示在真实划分 A 中属于同一类簇,而在聚类结果 B 中不属于同一类簇的数据对的数目;N_c 表示在真实划分 A 中不属于同一类簇,而在聚类结果 B 中属于同一类簇的数据对的数目;N_d 表示在真实划分 A 和聚类结果 B 中均不属于同一类簇的数据对的数目。

$$precision=\frac{N_a}{N_a+N_c};\tag{11}$$

$$recall=\frac{N_a}{N_a+N_b};\tag{12}$$

$$F_\beta=(1+\beta^2)\frac{precision\cdot recall}{\beta^2\cdot precision+recall}。\tag{13}$$

若 $\beta=1$,则式(13)可以化为

$$F=\frac{2\cdot precision\cdot recall}{precision+recall}。\tag{14}$$

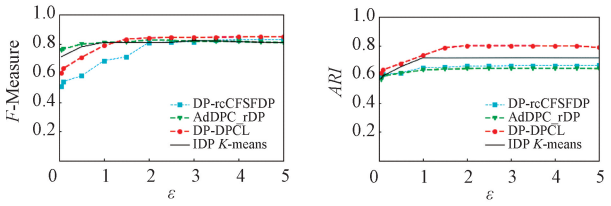
兰德指数(*RI*)是一种考虑在真实划分和聚类结果中均划分为同一类簇或均不是同一类簇,即 N_a 和 N_d 这 2 种情况的评价指标。这种指标评价比较片面,因此对兰德指数进行改进,提出调整兰德指数(*ARI*)。*ARI* 用于度量真实划分 A 和聚类结果 B 之间的相似度,其计算式如式(15)所示,取值为[-1,1],值越接近 1,聚类效果越好。

$$ARI(A,B)=\frac{RI(A,B)-E\{RI(A,B)\}}{\max(RI(A,B))-E\{RI(A,B)\}}。\tag{15}$$

式中:$RI(A,B)=\frac{N_a+N_d}{N_a+N_b+N_c+N_d}$;$E\{RI(A,B)\}$表示真实划分 A 和聚类结果 B 的期望兰德指数。

为对比算法在不同隐私预算下的聚类性能,将 DP-DPCL 算法与 AdDPC_rDP 算法、DP-rcCFSFDP 算法[16]、IDP K-means 算法[7]进行对比,通过算法的 F-Measure 和 *ARI* 指标值评估 DP-DPCL 算法的聚类性能。实验中 DP-DPCL 算法在 Iris、WDBC、Aggregation 数据集上的 d_c 取值分别为 2.5、1.5、2.5,邻域半径取值分别为 2.2%、0.8%、0.5%,加权因子设为 1.001。AdDPC_rDP 算法在 Iris、WDBC、Aggregation 数据集上的 d_c 取值分别为 2.5、3.0、2.5。DP-rcCFSFDP 算法的 d_c 取值分别为 2.5、2.5、4.0,根据文献[16],邻域半径分别为 1.8%、0.7%、0.4%。IDP K-means 算法中,根据文献[7],敏感度 $\Delta f=M+1$,M 为数据集的特征属性数,最大迭代次数 $N_{iter\_max}=1\,000$。

图 2 为 4 种算法在 Iris 数据集上聚类的输出结果。由图 2(a)可知,随着隐私预算 ε 的增加,各算法的 F-Measure 指标值逐渐增大并趋于稳定;当隐私预算 $\varepsilon>1.5$ 时,DP-DPCL 算法的 F-Measure 达到最优,此时,在 Iris 数据集上 DP-DPCL 算法聚类性能最佳。由图 2(b)可知,随着隐私预算的增加,各算法的 *ARI* 逐渐增加并趋于稳定;当隐私预算 $\varepsilon>1.5$ 时,DP-DPCL 算法的 *ARI* 达到最优。因此,在 Iris 数据集上,总体聚类性能最佳的是 DP-DPCL 算法。

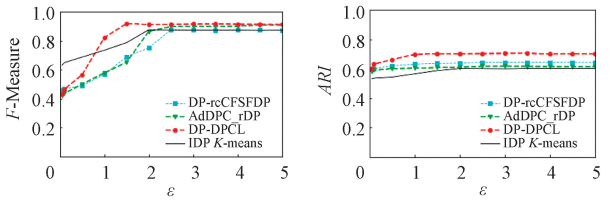


(a) *F*-Measure指标　　(b) *ARI*指标

图 2　Iris 数据集输出结果

Figure 2　Output results of Iris dataset

图 3 为 4 种算法在 WDBC 数据集上的输出结果。由图 3(a)可以看出,DP-DPCL 算法的 *F*-Measure 均优于 AdDPC_rDP 算法;当隐私预算 $\varepsilon>1.5$ 时,DP-DPCL 算法的 *F*-Measure 达到最优,其值均比其余 3 种算法高。由图 3(b)可知,DP-DPCL 算法的 *ARI* 均优于其余 3 种算法。

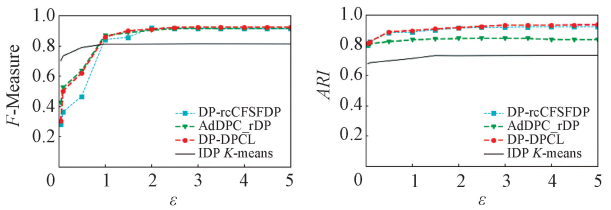


(a) *F*-Measure指标　　(b) *ARI*指标

图 3　WDBC 数据集输出结果

Figure 3　Output results of WDBC dataset

图 4 为 4 种算法在 Aggregation 数据集上的输出结果。由图 4(a)可知,当隐私预算 $\varepsilon>1.5$ 时,DP-DPCL 算法 *F*-Measure 均优于 AdDPC_rDP;当隐私预算 $\varepsilon=2.0$ 时,DP-rcCFSFDP 算法的 *F*-Measure 达到最优值,此时,DP-DPCL 算法的 *F*-Measure 较 DP-rcCFSFDP 算法小,其余情况下 DP-DPCL 算法均比 DP-rcCFSFDP 表现更佳。与 IDP *K*-means 算法的 *F*-Measure 比较可知,当隐私预算较小($\varepsilon<0.9$)时,IDP *K*-means 算法表现均比其余 3 种算法好,当隐私预算 $\varepsilon>0.9$ 时,IDP *K*-means 算法表现最差。图 4(b)结果显示,在 Aggregation 数据集上,DP-DPCL 算法的 *ARI* 最优。



(a) *F*-Measure指标　　(b) *ARI*指标

图 4　Aggregation 数据集输出结果

Figure 4　Output results of Aggregation dataset

实验结果表明,总体上看,DP-DPCL 算法的聚类性能与 AdDPC_rDP 算法、DP-rcCFSFDP 算法、IDP *K*-means 算法相比得到了提升,在不同的数据集上指标数值均优于其他 3 种算法。因此,DP-DPCL 算法可以在保护敏感数据的同时降低所添加的随机噪声对聚类可用性的影响。

4　结论

本文针对 AdDPC 算法的隐私泄露问题提出了 DP-DPCL 算法,并对算法的非聚类中心点分配策略进行了优化。首先对 AdDPC 算法存在的隐私泄露问题及非聚类中心点分配策略进行分析,在算法计算数据点局部密度的过程中添加 Laplace 噪声,并在算法分配非聚类中心点的过程中引入邻域和可达的概念,优化算法非聚类中心点分配的方法。实验结果表明,DP-DPCL 算法在隐私预算 $\varepsilon>1.5$ 时,聚类性能得到了提高,这是因为隐私预算过小时,添加的噪声过大,增加了其他算法对聚类中心选取的随机性,导致剩余数据点的分配可能比原始聚类算法的更好;而当预算增大时,DP-DPCL 算法对剩余数据点的分配方式进行了改进,使得更多的点分配到正确的类簇。但是,DP-DPCL 算法对不同的数据集聚类时,当最佳截断距离值不同,如何自适应选取最佳截断距离值需要进一步研究。

参考文献:

[1] SARLE W S. Algorithms for clustering data[J]. Technometrics, 1990, 32(2): 227-229.

[2] MA S H, GUO P K, YOU H R, et al. An image matching optimization algorithm based on pixel shift clustering RANSAC[J]. Information Sciences, 2021, 562: 452-474.

[3] DU Z J, LUO H Y, LIN X D, et al. A trust-similarity analysis-based clustering method for large-scale group decision-making under a social network[J]. Information Fusion, 2020, 63: 13-29.

[4] HASSAN B A, RASHID T A, HAMARASHID H K. A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star[J]. Computers in Biology and Medicine, 2021, 138: 104866.

[5] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1): 101-122.
XIONG P, ZHU T Q, WANG X F. A survey on differential privacy and applications[J]. Chinese Journal of Computers, 2014, 37(1): 101-122.

[6] DWORK C. Differential privacy[M]//Automata, Languages and Programming. Berlin: Springer, 2006: 1-12.

[7] 李杨, 郝志峰, 温雯, 等. 差分隐私保护 *k*-means 聚类方法研究[J]. 计算机科学, 2013, 40(3): 287-290.
LI Y, HAO Z F, WEN W, et al. Research on differential privacy preserving *k*-means clustering[J]. Computer

Science, 2013, 40(3): 287-290.

[8] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[9] 谢娟英, 高红超, 谢维信. *K* 近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学: 信息科学, 2016, 46(2): 258-280.
XIE J Y, GAO H C, XIE W X. *K*-nearest neighbors optimized clustering algorithm by fast search and finding the density peaks of a dataset[J]. Scientia Sinica (Informationis), 2016, 46(2): 258-280.

[10] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[M]//Theory of Cryptography. Berlin: Springer, 2006: 265-284.

[11] 陈韵. 基于差分隐私密度峰值聚类算法的研究和应用[D]. 南京: 南京邮电大学, 2020.
CHEN Y. Research of density peak clustering algorithm based on differential privacy preserving[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020.

[12] SASAKL Y. The truth of the *F*-Measure [J]. Teach tutor mater, 2007, 1(5): 1-5.

[13] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.

[14] DUA D, TANISKIDOU E K. UCI Machine learning repository[EB/OL]. (2017-02-13)[2022-10-11]. https://archive.ics.uci.edu/ml/index.php.

[15] STREET W N, WOLBERG W H, MANGASARIAN O L. Nuclear feature extraction for breast tumor diagnosis[C]// Proc SPIE 1905, Biomedical Image Processing and Biomedical Visualization. San Jose: SPIE, 1993: 861-870.

[16] GIONIS A, MANNILA H, TSAPARAS P. Clustering aggregation[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1-12.

Differential Privacy Protection Scheme of Adaptive Clustering by Fast Search and Find of Density Peaks

GE Lina[1,2,3], CHEN Yuanyuan[1], WANG Jie[1,2], WANG Zhe[1]

(1. School of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China; 2. Key Laboratory of Network Communication Engineering, Guangxi Minzu University, Nanning 530006, China; 3. Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi Minzu University, Nanning 530006, China)

Abstract: In order to solve the privacy leakage problem caused by adaptive clustering by fast search and find of density peaks(AdDPC) when calculating the local density and the primary allocation strategy, a differential privacy protection scheme of an improved density peak clustering algorithm was proposed. In this scheme, the Laplace random noise was added in the process of calculating the local density of the algorithm. In this way, even if the attacker had the maximum background knowledge, it could not obtain the corresponding information by adding or deleting a point in the dataset, thereby, differential attack was used to obtain the information of the target data point, and to achieve the purpose of protecting the privacy data. In addition, the reachability definition was introduced to improve the allocation strategy of AdDPC when assigning non-clustered center points, so as to avoid the problem of data point allocation error caused by the one-time allocation strategy. The experiment compared *F*-Measure and *ARI* values of DP-rcCFSFDP, AdAPC-rDP, IDP *K*-means, and results showed that: when the privacy budget was greater than 1.5, the *F*-Measure and *ARI* values of the proposed algorithm were better than those of other algorithms, and this algorithm could protect sensitive data and data availability at the same time.

Keywords: density peaks; differential privacy; random noise; clustering algorithm