计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200008

# 基于差分隐私的 *K*-means 算法优化研究综述

孔钰婷　谭富祥　赵　鑫　张正航　白　璐　钱育蓉
新疆大学软件学院　乌鲁木齐 830000
新疆维吾尔自治区信号检测与处理重点实验室　乌鲁木齐 830046
新疆大学软件工程重点实验室　乌鲁木齐 830000
(1565066023@qq.com)

**摘　要**　差分隐私 *K*-means 算法(Differential Privacy *K*-means Algorithm,DP *K*-means)作为一种基于差分隐私技术的隐私保护数据挖掘(Privacy Preserving Data Mining,PPDM)模型,因简单高效且可保障数据的隐私而备受研究者的关注。文中首先阐述了差分隐私 *K*-means 算法的原理、隐私攻击模型,以分析算法的不足。然后从数据预处理、隐私预算分配、聚簇划分等 3 个角度讨论分析 DP *K*-means 算法改进研究的优缺点,并对研究中的相关数据集和通用评价指标进行了总结。最后指出 DP *K*-means 算法改进研究中亟待解决的挑战性问题,并展望了 DP *K*-means 算法的未来发展趋势。

**关键词**:差分隐私 *K*-means 算法;差分隐私;隐私保护;隐私保护数据挖掘
**中图法分类号**　TP309

## Review of *K*-means Algorithm Optimization Based on Differential Privacy

KONG Yu-ting,TAN Fu-xiang,ZHAO Xin,ZHANG Zheng-hang,BAI Lu and QIAN Yu-rong
College of Software,Xinjiang University,Urumqi 830000,China
Key Laboratory of Signal Detection & Processing in Xinjiang Autonomous Region,Xinjiang University,Urumqi 830046,China
Key Laboratory of Software Engineering,Xinjiang University,Xinjiang University,Urumqi 830000,China

**Abstract**　Differential privacy *K*-means algorithm (DP *K*-means),as a kind of privacy preserving data mining (PPDM) model based on differential privacy technology,has attracted much attention from researchers because of its simplicity,efficiency and ability to guarantee data privacy. Firstly,the principle and privacy attack model of differential privacy *K*-means Algorithm are described,and the shortcomings of the algorithm are analyzed. Then,the advantages and disadvantages of the improvement research of DP *K*-means algorithm are discussed and analyzed from three perspectives,including data preprocessing,privacy budget allocation and cluster partition,and the relevant data sets and common evaluation indexes in the research are summarized. At last,the challenging problems to be solved in the improvement research of DP *K*-means algorithm are pointed out,and the future development trend of DP *K*-means algorithm is prospected.

**Keywords**　Differential privacy *K*-means algorithm,Differential privacy,Privacy preservation,Privacy preserving data mining

## 1　引言

随着各行各业信息化的建设,信息系统产生并积累了大量的用户数据,各医疗机构、金融机构和企事业单位等常采用数据挖掘技术来发现数据中隐藏的知识,为商业决策和业务优化提供便利。但数据挖掘分析的过程存在泄露个人敏感信息的风险,此时可解决数据隐私泄露问题的隐私保护数据挖掘研究极具现实意义[1-4]。

现有的隐私保护技术可划分为基于数据加密、数据匿名和数据失真这 3 种技术[5]。基于数据加密技术的代表算法有对称加密算法、非对称加密算法和同态加密算法[6],这些算法通过将数据转换成密文的方式实现数据的安全存储、传输及计算,但存在加解密阶段效率过低且耗费过多计算资源或存储资源的问题。基于数据匿名技术的代表方法有 $k$-匿名[7]和 $l$-diversity 方法[8],通过对准标识符的泛化处理来实现数据的隐藏,但需要针对新型攻击不断完善模型。针对上述两种技术存在的理论性和经验局限性的问题,Dwork 提出差分隐私技术(Differential Privacy[9],DP),其优点在于:1)定义了严格的攻击模型,可抵御背景知识攻击和各种形式的新型攻击;2)为隐私泄露风险给出了严谨、定量化的表示和证明;

到稿日期:2020-12-01　返修日期:2021-03-29
基金项目:国家自然科学基金(61966035);自治区科技厅国际合作项目(2020E01023);自治区研究生科研创新项目(XJ2019G072)
This work was supported by the National Natural Science Foundation of China(61966035),International Cooperation Project of the Science and Technology Department of the Autonomous Region(2020E01023) and Autonomous Region Graduate Research Innovation Project(XJ2019G072).
通信作者:钱育蓉(qyr@xju.edu.cn)

3)技术实现机制为噪声机制,对数据集加入的噪声量与敏感度有关,与数据集的大小与维度无关,在实际应用中加入极少量的噪声可提供高级别的隐私保护。

由于差分隐私的各项优势,研究者们将其应用在数据挖掘领域,以保证算法或数据的隐私性[10-11]。差分隐私保护数据挖掘研究的实现模式可划分为接口模式(Interface)和完全访问模式(Fully Access)。差分隐私保护数据挖掘的方法分为差分隐私保护的分类算法、差分隐私保护的聚类算法、差分隐私保护的回归算法、差分隐私保护的频繁项集挖掘算法[12-13]。本文选择差分隐私保护数据挖掘中简单高效、应用广泛的差分隐私 $K$-means 聚类方法进行分析。在早期差分隐私概念提出后,就陆续出现了各种差分隐私 $K$-means 的应用模型,如基于接口的 SuLQ[14] 和 PINQ[15] 框架、基于样本聚合[16]的框架、基于核心集的快速聚类框架[17] 和数据挖掘平台 GUPT[18],这些框架从不同的角度实现了 $K$-means 算法的隐私保护,但存在实用性不强、准确率低等问题。因此,在 DP $K$-means 的研究与应用中需关注如何在提升聚类隐私性的同时保证数据的可用性[19-20]。

本文第 2 节对差分隐私 $K$-means 算法的隐私定义、算法思想、攻击模型与存在的不足进行了介绍和分析;第 3 节对改进的差分隐私 $K$-means 算法进行划分,并进行了讨论分析;第 4 节讨论了差分隐私 $K$-means 算法亟需解决的问题并展望了未来的发展趋势;最后总结全文。

## 2 差分隐私 K-means 算法

### 2.1 差分隐私

#### 2.1.1 相关定义

差分隐私技术是一种基于噪声机制的数据失真隐私保护技术,通过对算法的输入、中间参数、目标函数、输出等隐私泄露点加噪声干扰,来确保在数据集中插入或删除一条记录的操作不会影响算法的输出结果,使攻击者在拥有最大背景知识情况下仍无法获取到数据集中单条记录的信息,从而达到隐私保护的目的[21]。差分隐私的定义如下。

**定义 1**(ε-差分隐私,ε-differential privacy) 对于所有只相差一条记录的邻近数据集(adjacent dataset)$D$ 和 $D'$,给定隐私算法 $M$,$\Pr[E_s]$表示事件 $E_s$ 的隐私被披露的风险,$Range(M)$表示 $M$ 的取值范围。若算法 $M$ 满足 ε-差分隐私,则对 $Range(M)$的任一子集 $S_m$ 有:

$$\Pr[M(D)\in S_m]\leqslant e^{\varepsilon}\times\Pr[M(D')\in S_m] \tag{1}$$

其中,参数 ε 是决定隐私保护水平的可调整隐私预算参数。ε 越小,算法 $M$ 在相邻数据集上返回结果的概率分布(差值控制在 $e^{\varepsilon}$ 之内)越相似,越难区分相邻数据集中相差的那条数据记录,算法提供的隐私保护程度就越高。ε 的取值依据具体需求设定,以平衡输出结果的隐私性与可用性[22]。

差分隐私技术中除参数 ε 外还有一个敏感度参数,指删除数据集中任一记录对查询结果造成的最大改变,分为全局敏感度(global sensitivity[23])、局部敏感度(local sensitivity[24])、平滑敏感度(smooth sensitivity[25])。其中,全局敏感度独立于数据集,由函数本身决定,是影响随机噪声量的关键参数,且实际应用较多,其定义如下。

**定义 2**(全局敏感度) 设 $f$ 是将 $d$ 维数据集 $D$ 映射为实数空间内一个 $d$ 维向量的查询函数,即 $D\to R_d$,对于任意只相差一条记录的邻近数据集 $D$ 和 $D'$,函数 $f$ 的全局敏感度为:

$$GS_f=\Delta f=\max_{D,D'}\| f(D)-f(D')\|_1 \tag{2}$$

#### 2.1.2 实现机制

差分隐私技术通常通过 Laplace 机制(Laplace Mechanism[26],LM)和指数机制(Exponential Mechanism[27],EM)来对应实现数值型和非数值型数据的隐私保护[28],除这两种机制外还存在其他噪声机制,如高斯机制(Gaussian mechanism)[29]、几何机制(geometric mechanism)[30]、矩阵机制(matrix mechanism)[31]、函数机制(functional mechanism)[32]等。

Laplace 机制通过向查询结果中加入服从 Laplace 分布的随机噪声来实现 ε-差分隐私保护。在 Laplace 分布中,记位置参数为 0,尺度参数为 $b$(其中 $b=\Delta f/\varepsilon$),随机噪声为:

$$Lap(b)=\exp\left(-\frac{|x|}{b}\right)=\exp\left(-\frac{|x|\varepsilon}{\Delta f}\right) \tag{3}$$

其概率密度函数为:

$$p(x|b)=\frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)=\frac{\varepsilon}{2\Delta f}\exp\left(-\frac{|x|\varepsilon}{\Delta f}\right) \tag{4}$$

**定义 3**(Laplace 机制) 给定数据集 $D$,设有查询函数 $f$,其全局敏感度为 $\Delta f$,那么随机算法 $f(D)$提供 ε-差分隐私保护,其响应输出 $M(D)$为:

$$M(D)=f(D)+Lap(\Delta f/\varepsilon) \tag{5}$$

其中,$Lap(\Delta f/\varepsilon)$ 为随机噪声,服从尺度参数 $b=\Delta f/\varepsilon$ 的 Laplace 分布。由式(4)可知,加入的噪声与 $\Delta f$ 成反比,与 ε 成正比。对于同一查询函数 $f$,其敏感度 $\Delta f$ 相同;当函数 $f$ 的敏感度 $\Delta f$ 较小时,算法加入极少量的噪声即可保证函数的隐私性。通过调整 ε 的值可实现不同程度的隐私保护,即 ε 越小,加入的噪声越多,隐私保护级别就越高,但需注意若加入的噪声过多则会影响数据的可用性。

在实际应用中,若查询结果为实体对象,则需要用到指数机制,其定义如下。

**定义 4**(指数机制) 对于给定的数据集 $D$,设查询函数的输出域为 $R$,域中的任意值 $r\in R$ 为实体对象。其中,输出值 $r$ 的可用性函数可表示为:

$$q(D,r)\to R \tag{6}$$

其中,打分函数 $q(D,r)$可用来评估输出值 $r$ 的优劣程度。在打分函数中引入邻近数据集的概念,对于任意的邻近数据集 $D$ 和 $D'$,其敏感度可表示为:

$$\Delta q=\max_{D,D'}\| q(D,r)-q(D',r)\| \tag{7}$$

给定隐私算法 $M$,若从 $R$ 中选择并输出 $r$ 的概率满足式(8),则算法 $M$ 提供 ε-差分隐私保护。

$$M(D,q)\propto\exp\left(\frac{\varepsilon q(D,r)}{2\Delta q}\right) \tag{8}$$

#### 2.1.3 隐私性质

对于一个复杂的隐私保护问题,需多次应用差分隐私保护算法才能更好地保护数据的隐私[33]。为了将隐私预算 ε 合理地分配到整个隐私保护过程,需要用到隐私保护算法的两个组合性质:序列组合性与并行组合性[34]。

**性质 1**(序列组合性) 设有算法 $M_1,M_2,\cdots,M_n$,其隐私保护预算分别为 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$,那么对于同一数据集 $D$,由这些

算法构成的组合算法 $M(M_1(D), M_2(D), \cdots, M_n(D))$ 提供 $(\sum_{i=1}^{n}\varepsilon_i)$-差分隐私保护。

**性质 2**(并行组合性) 设有算法 $M_1, M_2, \cdots, M_n$，其隐私保护预算分别为 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$，那么对于不相交数据集 $D_1, D_2, \cdots, D_n$，由这些算法构成的组合算法 $M(M_1(D_1), M_2(D_2), \cdots, M_n(D_n))$ 提供 $(\max \varepsilon_i)$-差分隐私保护。

### 2.2 差分隐私 K-means 算法

#### 2.2.1 K-means 算法的流程

K-means 算法是基于划分的无监督学习方法，因其实用、简单、高效的特点被广泛应用于商业和科学研究[35]。K-means 的基本步骤如下：

Setp1 数据预处理，输入数据集 $D$，将数据归一化到 $[0,1]^d$ 空间中，得到 $n$ 个数据点 $p_1, p_2, \cdots, p_n$；

Step2 初始化，选定数据集需划分的簇数 $k$，从 $n$ 个数据点中随机选取 $k$ 个点 $c_1, c_2, \cdots, c_k$ 作为初始中心点(初始的聚类中心点)；

Step3 划分聚簇，计算每个数据点 $p_1, p_2, \cdots, p_n$ 与中心点 $c_1, c_2, \cdots, c_k$ 的欧氏距离 $dist$，将数据点划分到距离最近的中心点聚簇集合中 $S_1, S_2, \cdots, S_k$；数据对象 $p$(有 $m$ 维属性)与第 $i(1\leqslant i\leqslant k)$ 个聚类中心点的欧氏距离 $dist$ 的计算式为：

$$dist(c_i, p)=\sqrt{\sum_{j=1}^{m}(c_{ij}-p_j)^2} \tag{9}$$

Step4 更新聚类中心点，计算每个聚簇集合 $S_i(1\leqslant i\leqslant k)$ 中数据点的均值为：

$$c_i'=\frac{sum}{num}=\frac{\sum_{p\in S_i}p}{|S_i|} \tag{10}$$

将均值 $c_i'$ 更新为 K-means 算法新的聚类中心点，并计算各簇的误差平方和 SSE(Sum of Squares Due to Error)：

$$SSE=\sum_{i=1}^{k}\sum_{p\in S_i}(dist(c_i', p))^2 \tag{11}$$

Step5 算法结束判断，如果迭代次数达到设定的上限或者目标函数 SSE 收敛，则返回聚类中心点，输出聚类结果；否则返回 Step3—Step5。

#### 2.2.2 K-means 隐私攻击模型

通过对 K-means 算法流程的分析发现，其初始中心点和聚类中心点存在隐私泄露的问题，有两种常见的隐私攻击模型：基于聚类中心点的攻击和基于背景知识的攻击[36]。

(1)基于聚类中心点的攻击

在 K-means 聚类的迭代过程中，Step3 计算聚类中心点距离的操作会泄露隐私。假设攻击者获得了数据集 $D$ 中的一个数据 $p(x_1, x_2, \cdots, x_n)$ 及其在 $t$ 次迭代过程中与各聚簇中心点的距离 $dist(d_1, d_2, \cdots, d_k)$，以及发布的聚类结果 $C(c_1, c_2, \cdots, c_k)$。欧氏距离的计算式如式(9)所示，算法有 $k$ 个中心点并进行 $t$ 次迭代可以得到 $k\times t$ 个距离计算式，当 $k\times t\geqslant m$ 时，可计算出数据 $p$ 的 $m$ 个属性的准确值。在实际应用中，$k\times t\geqslant m$ 通常成立，迭代次数 $t$ 越大，样本属性维度 $m$ 越少，隐私暴露就越彻底。

如表 1 所列，假设攻击者获取到两次迭代的中心点 $\{c_1, c_2\}$ 和数据 $p(x_1, x_2, x_3)$ 到中心点的距离 $\{d_1, d_2\}$，可联立 4 个等式，求得 $x_1=1, x_2=1, x_3=2$，相当于数据 $p$ 的值已被攻击，从而暴露了隐私。

表 1 簇中心点与距离数据

Table 1 Cluster center point and distance data

| The number of iterations | Cluster center point $c_1$ | Distance $d_1$ | Cluster center point $c_2$ | Distance $d_2$ |
|---|---|---|---|---|
| 1 | (1,2,1) | 1.414 | (2,1,2) | 1 |
| 2 | (1.5,2,0.5) | 1.871 | (1.5,1,2) | 0.707 |

(2)基于背景知识的攻击

攻击者结合 K-means 聚类最终发布的聚类中心点和自己拥有的背景知识，可攻击出数据集的信息。假设有受保护的数据集 $D$，攻击者拥有最大背景知识 $D'$，$D'$ 与 $D$ 之间只相差一条数据 $p(x_1, x_2, \cdots, x_m)$(数据点有 $m$ 维属性)，攻击者还知晓数据 $p$ 属于以 $c(y_1, y_2, \cdots, y_m)$ 为中心点的聚簇 $S(p_1, p_2, \cdots, p_{n-1}, p)$(聚簇内有 $n$ 个数据点)，即知晓聚簇中除 $p$ 以外的所有数据点及聚类中心点。通过式(10)可推导出式(12)，等式中除 $p$ 的值未知外，其他数据点的值均已知，从而计算出数据 $p$ 的每个属性的准确值：

$$c=\frac{p+\sum_{j=1}^{n-1}p_j}{n} \tag{12}$$

#### 2.2.3 DP K-means 算法

通过 2.2.2 节可知，聚簇的中心点存在隐私泄露的风险，Dwork 等[37]指出对聚类中存在隐私风险的点添加少量的噪声即可避免泄露个人隐私信息。随后，Blum 等[14]提出用差分隐私技术保护 K-means 聚类中心点的方法——DP K-means，对 2.2.1 节中 K-means 算法的 Step4 进行改进，即对聚簇中数据的求和函数 $sum$ 及计数函数 $num$ 加少量噪声干扰，以实现簇聚类中心点的隐私保护。DP K-means 算法的基本步骤可在 K-means 的基础上更新 Step4。

Step4 更新聚类中心点，计算每个聚簇集合 $S_i(1\leqslant i\leqslant k)$ 中数据点的和 $sum$ 和数目 $num$ 分别为：

$$sum=\sum_{p\in S_i}p \tag{13}$$

$$num=|S_i| \tag{14}$$

分别添加拉普拉斯噪声 $Lap(b)$ 得到 $sum'$ 和 $num'$，聚类中心点可更新为：

$$c_i'=\frac{sum'}{num'}=\frac{\sum_{p\in S_i}p+Lap(\frac{\Delta f}{\varepsilon})}{|S_i|+Lap\left(\frac{\Delta f}{\varepsilon}\right)} \tag{15}$$

计算各簇的误差平方和为：

$$SSE=\sum_{i=1}^{k}\sum_{p\in S_i}(dist(c_i', p))^2 \tag{16}$$

#### 2.2.4 DP K-means 存在的不足

DP K-means 算法在 K-means 算法的基础上加入差分隐私技术，以保护聚类分析过程中的隐私信息，其存在 K-means 算法本身的局限性，如存在对初始值($k$ 值与初始中心点)和离群点敏感、易陷入局部最优解等问题。随着差分隐私技术的引入，聚类中心点被加入噪声，噪声的添加对聚类的准确性、收敛速度、数据的可用性均会造成影响。

针对 DP K-means 算法的不足，研究者们不断提出改进算法，并在聚类分析算法与差分隐私技术相结合的过程中寻找隐私性与可用性的平衡。目前，对 DP K-means 算法的改进方法可划分为以下 3 类：基于数据预处理的改进、基于隐私预算分配的改进和基于聚簇划分的改进。

## 3 改进的差分隐私 $K$-means 算法

### 3.1 基于数据预处理的改进

$K$-means 算法过度依赖初始中心点，DP $K$-means 基于 $K$-means 算法增加了差分隐私保护的知识，但其聚类的效果同样受初始中心点、聚类个数 $k$ 和离群点等的影响，输入数据或参数处理不当时，该算法易陷入局部最优解[38-40]，因此需从初始中心点与 $k$ 值选取、离群点处理等角度进行改进。

在初始中心点选取的改进上，Li 等[41]和 Ren 等[42]均从数据集划分入手，Li 等将数据集划分为 $k$ 个子集后，取各子集的均值为初始中心点，避免了随机选取初始中心点对聚类结果的影响，但取均值的方法易受离群点的影响，如图 1 所示，部分数据导致初始中心点发生了较大偏移。Ren 等提出了 DPLK-means 算法，通过对数据集划分出的每个子集执行 DP $K$-means 算法来改进初始中心点的选取，在处理大型数据集和多维数据集时其聚类效果较好，但不适用于数据量过小无法进行划分的小型数据集。

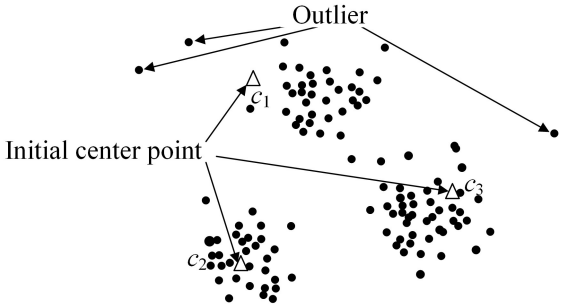



图 1 均值法取初始中心点

Fig. 1 Mean value method takes the initial center point

同样是优化初始中心点的选取，Canopy＋$K$-means 的优化思想得到了广泛的认可[43-44]。Li[45]、Yao[46]和 Shang 等[47]均采用 Canopy 算法对数据集进行预聚类或者粗聚类，输出的 $k$ 值与 Canopy 子集作为 DP $K$-means 的输入，避免了 $k$ 值与初始中心点选择的盲目性，减少了迭代次数，使算法快速收敛，同时可通过删除数据点数目较少的 Canopy 子集来消除数据集中的离群点，原理如图 2 所示。

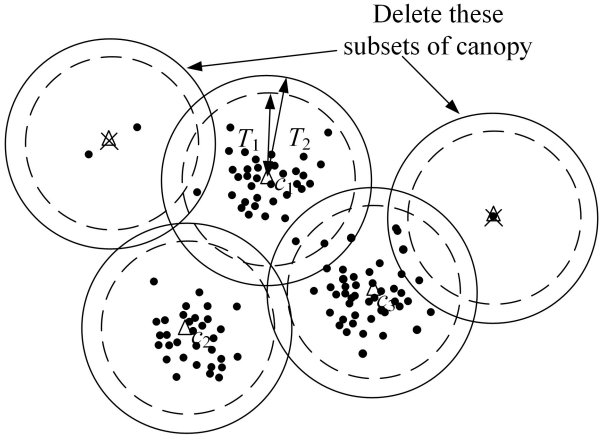



图 2 Canopy 预聚类示意图

Fig. 2 Canopy pre-clustering diagram

由图 2 可知，Yao 的算法选择在 MapReduce 框架上并行化实现，提升了算法的时效性；Shang 等在此基础上引入最大最小原则优化算法来优化 Canopy，提升了算法的稳定性。

另外，DP $K$-means 算法对离群值高度敏感，离群点的存在会导致聚簇更新中心点时发生极大的偏移，影响聚类结果[48]。考虑到离群点对 DP $K$-means 的负面影响，Yu 等[49]提出 OEDP k-means 算法，设计出离群点检测方法和剔除离群点的数据集分割方法 OEPT 对数据集 $DT=\{o_1,o_2,\cdots,o_n\}$ 进行预处理，将结果作为 DP $K$-means 的输入，以划分出 $k$ 个聚类。在 OEPT 方法中，首先设置 $r$ 最近邻区域（$r$-Nearest-Neighbour Area，rNNA），由 DT 中的任意数据点 $o$ 及其最邻近区域的 $r$ 个点组成，如图 3 所示，$r$-最近邻距离（r-nearest-neighbour distance）由 $o$ 与 $r$ 个点之间的欧氏距离 $dist(o,i)$ $(1\leqslant i\leqslant r)$ 计算得出，计算式为：

$$dist\_rNNA(o,DT)=\frac{\sum_{r}^{i=1}dist(o,i)}{r} \tag{17}$$

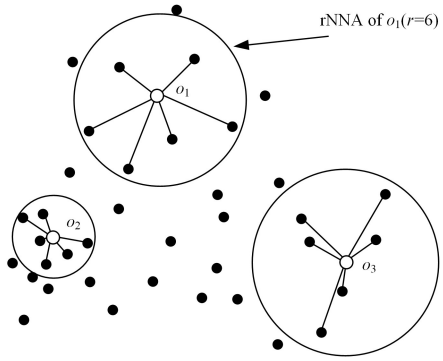



图 3 数据点 $o$ 的 rNNA($r$=6)

Fig. 3 rNNA ($r$=6) of data point $o$

数据点 $o$ 的 rNNA 密度（rNNA density）的计算式为：

$$dens\_rNNA(o,DT)=\frac{1}{dist\_rNNA(o,DT)} \tag{18}$$

数据点的 rNNA 密度越大，其 $r$-最近邻距离就越小。随后，标记数据集中第 $i$ 个对象与第 $j$ 个对象的距离为 $dist_{ij}$ $(i,j=1,2,\cdots,n)$，由其组成 $n\times n$ 的矩阵 $dist\_M$，$dist\_M$ 中的 $m$ 个最大值被记为 $dist_{ij}^{(m)}$，用 $top\_n$ 来表示要消除的离群点个数，其中 $top\_n=|DT|\times 0.05$，此时可计算出检测离群点的密度阈值 $\alpha$ 为：

$$d'=\{dist_{ij}\mid dist_{ij}\geqslant dist_{ij}^{(top\_n)},i,j=1,2,\cdots,n\} \tag{19}$$

$$\alpha=\frac{k}{mean(d')} \tag{20}$$

对于所有的点 $o_i$ $(1\leqslant i\leqslant n)$，若满足 $dens\_rNNA(o_i,DT)<\alpha$，则该点是离群点，从 DT 中将其删除。通过调整 $r$ 和 $\alpha$ 为适当的值，来检测 DT 中的离群值，识别出的离群值根据其到各聚类中心的距离被直接分配到最近的聚类中，不参与后期的聚类迭代分析，以保证数据的完整性。

Xiong 等[36]从数据密度入手，提出了 PADC k-means 算法，按密度对数据点排序，选取平均值作为初始中心点，通过数据点的密度检测并去除聚类过程中的离群值，以提高聚类结果的精度，在聚簇划分时使用相对距离并设置权重，有助于

准确划分数据点，降低离群点效应，提升聚类结果的精度。PADC k-means 算法中计算数据点密度值的计算式为：

$$density(x)=\frac{n}{\sum_{n}^{i=1}dist^2(x,y_i)} \tag{21}$$

其中，$x$ 代表数据点，$density(x)$ 代表数据点 $x$ 的密度值，$y_i$ 表示除该点以外的其他数据点，$n$ 表示数据点的数量。数据点的密度值越大，点周围就越紧凑，再引入离群点参数 $r$，接受数据点的数量为 $(n\cdot r)(r\in(0,1))$，其余点认定为离群点不参与中心点的计算，其中 $r$ 的具体取值需根据不同数据集的实验得出。算法采用数据点的方差来衡量聚簇的相似性，相似度较高时给予较大的权值。为了减小离群值对方差的影响，保留 90%离聚类中心点较近的数据点参与方差计算，即 $r=0.9$，方差计算式为：

$$s_i^2=\frac{\sum_{x\in c_i}\sum_{n\times 0.9}^{i-1}(x-c_i)^2}{n_{c_i}} \tag{22}$$

其中，$x$ 代表数据点，$c_i$ 代表聚类中心点，$n_{c_i}$ 代表 $c_i$ 所在聚簇的数据点的数量。根据方差 $s_i^2$ 可获得聚类的权值为：

$$w_i=\frac{1}{s_i^2} \tag{23}$$

在聚类过程中的距离计算中采用相对距离，计算式为：

$$dist^2(x,c_i)=w_i\cdot\sum_{d}^{i=1}(x_i-c_i)^2 \tag{24}$$

在 OEDP k-means 算法的基础上，Fan 等[50]基于 MapReduce 计算框架并行地计算各点的欧氏距离矩阵，并通过计算最近邻超球半径来导出离群点的判定阈值，选取的初始中心点更接近实际的中心点，减少了聚类过程的迭代次数。

Meng 等[51]通过引入群体智能算法为优化初始中心点的选取提供了新思路，提出了 DEDP K-means 算法，将 $K$-means 的聚类过程转换成种群进化求解，引入基于种群的差分进化(Differential Evolution，DE)启发式搜索算法来更新种群，通过比较选择出初始种群的适应值，再通过基于自适应对立的学习技术(adaptive opposition-based learning)来同时评估当前种群与反向种群，并在 Spark 平台上实现改进算法，提高了大规模数据集的处理效率。

### 3.2 基于隐私预算分配的改进

DP $K$-means 算法的实现是对聚类过程中的隐私泄漏点添加噪声。添加的噪声过多不仅会导致数据失真，还会导致最后的聚类结果发生偏移或算法无法收敛。因此，如何针对不同迭代过程及不同的聚簇分配适当的隐私预算是研究的难点。

Dwork[52]首先提出了统一分配法(uniform allocation)和二分法分配法(dichotomy allocation)的隐私预算 $\varepsilon$ 分配方法。假设数据集的维数为 $dim$，统一分配法指算法的迭代次数固定为 $N$，每次迭代需加入的噪声为：

$$Lap(b)=Lap((dim+1)N/\varepsilon) \tag{25}$$

统一分配法需要预定义迭代次数的参数 $N$，每次迭代中的隐私预算都相同，且需要额外的计算来确定最佳参数 $N$，因此聚类结果无法达到最佳。二分法分配的思想是对迭代次数 $N$ 未知的情况，整个算法的隐私保护预算为：

$$\varepsilon=\sum_{i=1}^{N}\varepsilon_i \tag{26}$$

后续的第 $i$ 次迭代中隐私预算递减为上一次迭代的一半，即第 $i$ 次迭代的隐私参数为 $\varepsilon_i=\frac{\varepsilon}{2^i}$，所添加的噪声为：

$$Lap(b)=Lap\left(\frac{dim+1}{\frac{\varepsilon}{2^i}}\right)=Lap\left(\frac{(dim+1)2^i}{\varepsilon}\right) \tag{27}$$

二分法分配隐私预算的缺陷在于消耗速度非常快，过小的隐私预算会产生较大的噪声，影响聚类中心点的更新。

还有一种序列和分配法(series sum allocation[53])，对于公式 $\sum_{i=1}^{\infty}1/(i(i+1))=1$，公式两边均乘上 $\varepsilon$，得公式 $\sum_{i=1}^{\infty}\varepsilon/(i(i+1))=\varepsilon$，即可得出每次迭代可分配的隐私预算为：

$$\varepsilon_i=\varepsilon/(i(i+1)) \tag{28}$$

序列和分配法的优势在于允许隐私预算的无限分配。为避免因迭代次数逐步增多，隐私预算随之减小，生成的噪声增大，从而导致数据失真的情况，Su 等[54]提出了最小隐私预算 $\varepsilon^m$ 的概念，假设有数据集 $D$，最小隐私预算 $\varepsilon^m$ 的计算式为：

$$\varepsilon^m=\left(\frac{500k^3}{N^2}(d+\sqrt[3]{4d\rho^2})^3\right)^{1/2} \tag{29}$$

其中，$N$ 代表记录数，$d$ 代表数据维数，$k$ 指需划分的聚簇个数，$\rho$ 指聚簇聚类中心点第 $i$ 维数据的归一化坐标 $\rho=sum/(2r\cdot num)$，$r$ 表示数据分布范围的参数，即将数据归一化到 $[r,-r]$ 范围内。

但上述 2 种隐私预算分配方法存在一些缺陷，未考虑到 $K$-means 的目标函数在前期迭代中收敛速度更快，隐私预算对聚类结果的影响更大的情况。Fu 等[55]从选取初始中心点和迭代更新中心点两个角度来动态分析隐私预算分配的不同对聚类效果的影响，证明了聚类效果在前期迭代时受隐私预算分配的影响更大。在此基础上，Fan 等[53]提出等差数列隐私预算分配方法(arithmetic progressions allocation)，将总的隐私预算分解为递减的等差序列，当隐私预算减小至阈值最小隐私预算 $\varepsilon^m$ 时，隐私预算分配方式退化为统一分配法。此分配方法使得算法在前几次迭代中能分配到更多的隐私预算，使算法快速收敛，避免了聚类中心点的更新被噪声所引导。

Zhang 等[56]在 MapReduce 框架上并行实现算法，考虑到在同一迭代中不同簇的数据特征不同，在每次划分好聚簇之后，根据聚簇内每个点的轮廓系数 $S(i)$ 来计算每个聚簇的平均轮廓系数 $S(k)$，其计算式分别为：

$$S(i)=(bi-ai)/\max(ai,bi) \tag{30}$$

$$S(k)=\sum_{num_k}^{i=1}S(i)/num_k \tag{31}$$

Zhang 等对 $k$ 个聚簇在 $t$ 次迭代中设置的隐私预算 $\varepsilon_k^t$ 和添加的噪声 $Noise_k^t$ 均不同，以减少噪声导致的聚类中心点偏差，其计算式分别为：

$$\varepsilon_k^t=\frac{\varepsilon}{2^t}[(1+S_k)/(1+\min S_k)] \tag{32}$$

$$Noise_k^t=Lap(\Delta f/\varepsilon_k^t) \tag{33}$$

Zhang 等改进算法的不足点在于，当簇过小时噪声会导致聚簇中出现空簇。

Mo 等[57]提出了基于距离聚类的自适应隐私预算参数分配机制，并设计评估函数 $F[U(\varepsilon),V(\varepsilon)]$来评估算法的可用性 $U(\varepsilon)$与隐私性 $V(\varepsilon)$，评估函数的计算式为：

$$F[U(\varepsilon),V(\varepsilon)]=a\cdot U(\varepsilon)+b\cdot V(\varepsilon) \tag{34}$$

其中，$a$ 和 $b$ 均为权重，根据算法的隐私性和可用性的重要程度来选择具体的值。算法的优化目标为在最小隐私约束下达到最大可用效应，可表示为：

$$\max_{U(\varepsilon)}\{\min_{V(\varepsilon)} F[U(\varepsilon),V(\varepsilon)]\} \tag{35}$$

此时隐私预算参数的计算式为：

$$\varepsilon^{*}=\arg_{\varepsilon}\{\max_{U(\varepsilon)}\{\min_{V(\varepsilon)} F[U(\varepsilon),V(\varepsilon)]\}\} \tag{36}$$

为了实现隐私预算参数自适应机制，Mo 等的改进算法使用 $\alpha-\beta$ 剪枝来寻找 $F[U(\varepsilon),V(\varepsilon)]$的最优值，以达到隐私性和可用性的平衡。

### 3.3 基于聚簇划分的改进

DP $K$-means 算法通过计算数据对象之间的距离来表示数据之间的相似度，并将与聚簇中心点相似度高的数据点划分到该聚簇中，迭代计算直至聚簇不发生变化，即完成数据的聚类过程。对隐私泄露点，即聚类中心点加噪声干扰后，能实现中心点的隐私保护，但聚类后期的迭代过程会受到影响，使算法出现不收敛、聚类中心点发生偏移、误差较大等问题。

在没有收敛保证的情况下，针对不同的数据集需要多次运行差分隐私 $K$-means 算法，以找到适合的迭代次数来终止算法的迭代，预定义迭代次数需消耗较大的计算成本；同时，在不收敛的情况下得到的结果与 $K$-means 的局部最优解有很大的差距，影响算法的效率与聚类质量。Dong 等[58]和 Zhang 等[59]在应用 LapDP(拉普拉斯机制)、ExpDP(指数机制)时就出现了算法不收敛的问题。Lu 等[60]在此基础上针对现有的交互式差分隐私聚类算法不收敛的问题，结合 LapDP、ExpDP 和样本聚合框架，通过向采样区中加入噪声来控制聚类中心点更新时的移动方向，以保证算法的收敛性、收敛速度和算法质量。Lu 等[61]根据 $t$ 次到 $t-1$ 次迭代的先验知识和 $t$ 次到 $t+1$ 次的后验知识两种策略来构建采样区，预测聚类中心点的移动方向，以多于 $K$-means 一倍的迭代次数和收敛速度为代价，来换取相同初始聚类中心点时算法可收敛到与 $K$-means 相同的解，保证了算法的收敛性与聚类质量。

DP $K$-means 中加入的噪声对聚类中心点更新会产生影响，Su 等首先提出基于网格的非交互式 EUGkM[54]算法，然后融合交互式 DPLloyd[14]算法提出 Hybrid[58]算法。Hybrid 算法的思想在于，通过 EUGkM 生成最接近真实聚类中心的近似初始聚类中心，再输入 DPLloyd 迭代精进聚类中心点；并提出分析噪声聚类中心点和真实聚类中心点之间的均方误差方法来优化迭代次数和隐私预算分配，均方误差越小，表明加入的噪声对聚类中心点更新的影响越小；并与 DPLloyd，GkM[18]和 PGkM[59]等多种方法进行对比分析，算法在聚类数 $k$ 和属性维度 $d$ 增加的情况下表现较为稳定，因此其思想可推广应用到其他具有迭代或增量算法结构的数据分析任务中[62]。

Lv 等[63]在 Su 等[54]和 Nguyen[64]研究的基础上提出了混合 $K$-means 与 $K$-mode 的混合算法 ODPCA，将数值属性的噪声聚类中心点与真实聚类中心点之间的均方误差(MSE)与分类属性因加噪引起的方差之和，定义为优化迭代次数的损失函数，以确保每次迭代的损失不超过阈值，并为每次迭代分配最小的隐私预算。

Ni 等[65]提出合并相邻簇以抵消噪声对聚类中心点影响的思想，并设计了 DP-KCCM 算法。初始聚类中心点选取方法遵循任意聚类中心点与域边界的距离至少为 $a$，且任意两个聚类中心点之间的距离至少为 $2a$ 的原则。Ni 等首先将需划分为 $k$ 类的数据集划分为 $n\times k$ 个聚类，然后在更新聚类中心点阶段根据迭代次数加入自适应噪声，在重复迭代至最大迭代次数后，将 $n\times k$ 个簇合并成 $k$ 个簇。在簇合并上有两种策略，如图 4 所示。

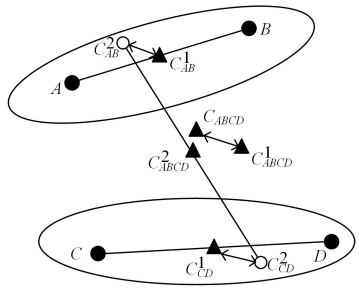



图 4 相邻簇合并策略

Fig. 4 Adjacent cluster merging strategy

假设 $n$ 取值为 4，图 4 中有数据点 $A,B,C,D$。第一种策略将 4 个数据点视为同一类别，加噪得到噪声聚类中心点为 $C^1_{ABCD}$；第二种策略将$\{A,B\}$和$\{C,D\}$视为相邻的两个簇，分别求得聚类中心点 $C^1_{AB}$与 $C^1_{CD}$，加噪分别得到噪声聚类中心点为 $C^2_{AB}$与 $C^2_{CD}$，合并这相邻两个簇的聚类中心点得到最终的聚类中心点 $C^2_{ABCD}$。从图 4 中可看出，最终的聚类中心点 $C^2_{ABCD}$与无噪声的聚类中心点 $C_{ABCD}$距离相近，在此情况下，加噪对聚类中心点的影响较小。但该算法未考虑离群值与样本不均衡的影响，且当数据量较小时，算法无法划分为 $n\times k$ 个聚簇。

Gao 等[66]采用群智能算法来改进聚类中心点的选取，提出了 DPHKMS 算法，在 Spark 环境下，混合粒子群优化和布谷鸟搜索等算法来优化差分隐私 $K$-means 算法的聚类中心点选择，充分利用启发式群智能算法的有效性与优势，改善了 DP $K$-means 算法的多样性。

### 3.4 讨论分析

#### 3.4.1 数据集

为了科学评价 DP $K$-means 及其改进算法的效用，使用公开数据集和评价指标对算法性能进行评估。本节对各学者研究并改进 DP $K$-means 算法时用到的数据集进行了总结，如表 2 所列，15 个数据集均来自 UCI Maching Learning Repository[67]，其中，高维和大型数据集的占比相对较小，且大部分数据集存在样本分布不均衡的问题，在 DP $K$-means 的研究中需注意数据预处理的工作。

表 2 DP K-means 研究数据集总结

Table 2 DP K-means research data set summary

| Name of dataset | Number of records | Number of Attributes | Number of Clusters | Attribute type | Size of classes |
|---|---|---|---|---|---|
| Iris | 150 | 4 | 3 | Real | 50,50,50 |
| Wine | 178 | 13 | 3 | Integer,Real | 59,71,48 |
| Haberman | 306 | 3 | 2 | Integer | 225,81 |
| Ecoli | 336 | 8 | 8 | Real | 143,77,2,2,259,20,5,52 |
| User Knowledge Modeling | 403 | 5 | 4 | Real | 50,129,122,130 |
| Climate | 540 | 18 | 2 | Real | 46,494 |
| Blood | 748 | 4 | 2 | Real | 570,178 |
| Wave | 5000 | 40 | 3 | Real | 1657,1647,1696 |
| Electrical | 10000 | 13 | 2 | Real | 3620,6380 |
| HTRU2 | 17898 | 9 | 2 | Real | 16258,1639 |
| MAGIC Gamma | 19020 | 11 | 2 | Real | 12332,6688 |
| Occupancy Detection | 20560 | 7 | 2 | Real | 15810,4750 |
| Credit-card | 30000 | 24 | 2 | Integer,Real | 23364,6636 |
| Adult | 48842 | 14 | 2 | Categorical,Integer | 37155,11687 |
| Covertype | 581012 | 54 | 7 | Categorical,Integer | 211840,283301,35754,2747,9493,17367,20510 |

3.4.2 评价指标

差分隐私 K-means 算法在实际应用中需要兼顾隐私性与可用性。从隐私性的角度可使用($\alpha$,$\beta$)-userfullness[68]来度量差分隐私算法,从聚类可用性的角度可使用的评价指标有误差平方和(SSE)、均方误差(MSE)、加速比、收敛速度、迭代次数、运行时间、F-measure[69]、CH(Calinski-Harabasz[42])、NICV(Normalized Intra-Cluster Variance[54])、归一化互信息 NMI(Normalized Mutual Information[70])、平均扰动 AP(Average Perturbation)、兰德系数 RI(Rand Index)、调整兰德系数 ARI(Adjusted Rand Index)与纯度(Purity)[71]等。

本节主要介绍 F-measure,CH,NICV 这 3 个较为通用的评价指标来评价 DP K-means 算法聚类可用性的指标。

(1)F-measure

在大部分的差分隐私 K-means 研究中,选择采用 F-measure 指标来衡量聚类的可用性,其是综合准确率和召回率的综合性指标。F-measure 一般是对两种算法在同一个数据集下处理得到的聚类结果进行比较,F-measure 值域是[0,1],其结果越大,两个聚类结果的相似度越大,聚类结果的可用性就越高。

F-measure 的计算过程如下:对于聚类数为 $k$ 的数据集,用 $CLUSTER$ 表示作为参考的未加隐私保护技术的 K-means 聚类结果或数据集的标准聚类结果,$CLUSTER'$ 表示加隐私保护技术的聚类结果,$U_i$ 为 $CLUSTER$ 中第 $i$ 个聚类集合($1\leqslant i\leqslant k$),$V_i$ 为相同标记下 $CLUSTER'$ 的第 $i$ 个聚类集合,$cover_i$ 为 $U_i$ 和 $V_i$ 相重合的记录数目,$|U_i|$ 和 $|V_i|$ 分别为 $U_i$ 和 $V_i$ 中的记录数目,记第 $i$ 个聚类集合的准确率为 $P_i$,其召回率为 $R_i$,则:

$$R_i=\frac{cover_i}{|U_i|} \tag{37}$$

$$P_i=\frac{cover_i}{|V_i|} \tag{38}$$

计算 $P_i$ 和 $R_i$ 的加权调和平均,记为 $F_i$,则:

$$F_i=\frac{2R_iP_i}{R_i+P_i} \tag{39}$$

再为各类聚类集合的 $F_i$ 进行加权平均,设 $N_{TOTAL}$ 为数据集的记录总数,此时,差分隐私 K-means 算法聚类结果的可用性度量公式为:

$$F\text{-}measure=\sum_{U_i\in CLUSTER}\frac{U_i}{N_{TOTAL}}F_i \tag{40}$$

(2)CH 指标

CH 指标是聚类有效性评估的内部指标,基于数据集本身的统计结果和算法的聚类结果,适用于实际类别信息未知的情况。CH 指标的计算式如下:

$$CH=\frac{\frac{1}{k-1}\sum_{i=1}^{k}T_id^2(c_i,C)}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{x\in c_i}d^2(x,c_i)} \tag{41}$$

其中,$k$ 指聚簇的数量,$n$ 指数据集中数据点的数量,$x$ 代表数据集中的任意点,$T_i$ 代表第 $i$ 个聚类点的数量,$C$ 代表原始数据集的中心点,$c_i$ 代表 $i$ 个聚簇的中心点,$d$ 代表两个点之间的距离。CH 指标由分离度与紧密度的比值得到,计算聚类集合中各点与聚类中心的距离平方和来度量类内的紧密度,计算各聚类中心点与数据集中心点的距离平方和来度量数据集的分离度。CH 越大代表类自身越紧密,类与类之间越分散,聚类结果更优。

(3)NICV 指标

归一化聚簇内方差 NICV 计算聚类结果中任意数据点到其最近的聚类中心点之间的平均平方距离,用于评估算法输出聚类中心点的质量。NICV 指标的值越小代表聚类的效果越好。假设算法对被归一化到 $d$ 维的数据集 $D=\{x_1,x_2,\cdots,x_N\}$ 进行聚类,其中 $x_i\in R^d$,将 $D$ 划分为 $k$ 个不相交子集 $(C_1^*,C_2^*,\cdots,C_k^*)$,对应的聚类中心为 $\{C_1,C_2,\cdots,C_k\}$,NICV 指标的计算式为:

$$NICV=\frac{1}{N}\sum_{j=1}^{k}\sum_{x_i\in C_j^*}\|x_i-C_j\|^2 \tag{42}$$

$$C_j^t=\frac{\sum_{x_i\in C_j^*}x_i^t}{|C_j^*|},\forall t\in\{1,\cdots,d\} \tag{43}$$

其中,$x_i^t$ 代表数据点 $x_i$ 的第 $t$ 维数据,$C_j$ 代表第 $j$ 个聚类中心的第 $t$ 维数据。

3.4.3 对比分析

3.1 节-3.3 节将 DP K-means 改进算法划分为 3 类分别进行分析阐述。表 3—表 5 分别从主要思想、主要优点、主要缺点、评价指标、算法误差等方面对算法进行对比分析。DP K-means 改进算法通常选取 F-measure 指标来评估算法可用性,本节用 1-F-measure 来表示算法误差。

表 3 基于数据预处理的 DP $K$-means 改进算法的对比分析

Table 3 Comparative analysis of DP $K$-means improved algorithm based on data preprocessing

| 文献 | 年份 | 算法名称 | 主要思想 | 主要优点 | 主要缺点 | 评价指标 | 算法误差 |
|---|---|---|---|---|---|---|---|
| 文献[41] | 2013 | IDP k-means | 数据集划分子集并取均值为初始中心点 | 避免初始中心点选择的随机性 | 易受离群点影响；不太适合处理小型数据集 | F-measure | (0,0.3) |
| 文献[45] | 2016 | DP Canopy K-means | Canopy 预处理数据集 | 避免初始值选取的盲目性；抗干扰性强；适合处理大型高维数据集 | 增加 Canopy 的 $T_1$，$T_2$ 参数的选取；删除部分数据点导致数据集不完整 | F-measure，准确度，迭代次数 | (0,0.2) |
| 文献[49] | 2016 | OEDP k-means | 根据数据密度选择初始中心点，利用 $r$ 最近邻区域检测与剔除离群点 | 消除离群点对算法的影响 | 不适合处理高维空间对称分布且密度均匀的数据集 | F-measure，执行时间 | (0.02,0.25) |
| 文献[42] | 2017 | DPLK-means | 数据集划分子集并执行 DP K-means | 算法适用于高维数据集和多维数据集 | 不适合处理小型数据集 | CH | - |
| 文献[46] | 2018 | IDP-Kmeans | Canopy 预处理数据，MapReduce 并行加速 | 解决 DP $K$-means 精度低和局部最优问题；适合处理大型高维数据集 | 增加了 Canopy 算法 $T_1$，$T_2$ 参数的选取 | F-measure，迭代次数 | (0,0.4) |
| 文献[36] | 2018 | PADC k-means | 根据数据密度检测异常值，用加权值的相对距离度量相似度 | 聚类划分准确 | 未考虑样本不均衡的问题 | F-measure | (0,0.6) |
| 文献[51] | 2018 | DEDP K-means | 差分进化算法和基于自适应对立学习技术，Spark 并行加速 | 避免陷入局部最优解，且收敛速度快，迭代次数少 | 未针对数据点的特征加入适应的噪声 | F-measure，收敛时间，迭代次数 | (0.05,0.3) |
| 文献[50] | 2019 | TripleP K-means | 计算数据点间的欧氏距离矩阵，结合最近邻超球半径判定离群点，MapReduce 并行加速 | 算法优于 OEDP k-means，IDP k-means，算法收敛速度较快，耗时短 | 剔除了离群点，聚类结果不完整 | F-measure，耗时 | (0,0.15) |

表 4 基于隐私预算分配的 DP $K$-means 改进算法的对比分析

Table 4 Comparative analysis of DP $K$-means improved algorithm based on privacy budget allocation

| 文献 | 年份 | 算法名称 | 主要思想 | 主要优点 | 主要缺点 | 评价指标 | 算法误差 |
|---|---|---|---|---|---|---|---|
| 文献[56] | 2018 | — | 根据轮廓系数评估聚类相似度并根据各聚簇的特征分配不同的隐私预算，MapReduce 并行加速 | 对小尺寸或大密度的数据集支持较好 | 未考虑离群点对聚簇平均轮廓系数的影响 | F-measure，加速比 | (0,0.8) |
| 文献[53] | 2019 | APDPk-means | 隐私预算按算术级数递减，设置最小隐私预算 | 能保证早期迭代的快速收敛 | 同一迭代各聚簇中加入噪声相同，未考虑数据特征 | F-measure，SSE | (0,0.75) |
| 文献[55] | 2019 | DPk-means++ | 动态分配中心点初始选取和迭代更新过程的隐私预算 | 分析不同的分配对聚类可用性的影响 | 未考虑到离群点对聚类可用性的影响 | RI，ARI | — |
| 文献[57] | 2019 | — | 基于距离聚类的自适应隐私预算参数分配机制 | 降低了输入参数对聚类结果的影响 | 需根据应用场景设定隐私性与可用性权重 | CH | — |

表 5 基于聚簇划分的 DP $K$-means 改进算法的对比分析

Table 5 Comparative analysis of DP $K$-means improved algorithm based on clustering division

| 文献 | 年份 | 算法名称 | 主要思想 | 主要优点 | 主要缺点 | 评价指标 | 算法误差 |
|---|---|---|---|---|---|---|---|
| 文献[66] | 2017 | DPHKMS | 利用粒子群优化和布谷鸟搜索等启发式群体智能算法优化聚类中心点的选取，MapReduce 并行化支持 | 可动态确定要划分数据集的最佳类数 | 未考虑到离群点对聚类可用性的影响 | 加速比，平均计算时间，F-measure | (0.1,0.65) |
| 文献[58] | 2017 | Hybrid | 结合非交互式算法 EUGkM 与交互式算法 DPLloyd 提出优化改进的混合方法 Hybrid | Hybrid 也适用于其他具有迭代或增量算法结构的数据分析任务 | 隐私预算分配较小时算法效果易受影响 | NICV | — |
| 文献[60] | 2019 | Prior | 向采样区中加入动态规划的噪声控制聚类中心点的移动 | 算法收敛，在交互式环境中聚类质量高 | 算法迭代次数较多导致加入的噪声较多 | Cost Gap，Iteration Ratio | — |
| 文献[63] | 2019 | ODPCA | 针对数值与分类属性的误差来设计损失函数，以优化迭代次数与隐私预算的分配 | 适合处理含有数值和分类数据的混合数据集 | 未考虑离群值对聚类的影响 | NICV | — |
| 文献[61] | 2020 | — | 建立结合先验知识与后验知识采样区，并控制质心收敛方向 | 算法可收敛且不会陷入局部最优解 | 收敛较慢导致加入的噪声过多 | Percentage of Matching Results，Cost Gap，Iteration Ratio | — |
| 文献[65] | 2020 | DP-KCCM | 合并相邻簇以抵消噪声对聚类中心点的影响，自适应噪声 | 算法抗噪声干扰能力强 | 未考虑离群值与样本不均衡的影响，且不太适合处理小型数据集 | NICV | — |

3.4.4 实验结果与分析

本节复现了部分算法，主要考察隐私保护预算参数 ε 取值对差分隐私 *K*-means 算法聚类结果可用性的影响。实验中作对比分析的 4 个差分隐私 *K*-means 算法为 DP *K*-means、OEDP k-means[49]、DPk-means ＋ ＋[55]、DP-KCCM[65]。实验数据集为来自 UCI Maching Learning Repository 的数据集 Iris(记录数 $n=150$，数据属性维度数 $dim=4$，聚类数 $k=3$)和 HTRU2(记录数 $n=17\,898$，数据属性维度数 $dim=9$，聚类数 $k=2$)。实验数据均归一化至[0,1]范围内，实验平台 CPU 2.30 GHz，8 GB 内存，Windows 10 操作系统，基于 Python 语言及相关库实现算法。

实验中对差分隐私 *K*-means 聚类算法可用性评估选取的评价指标为 F-measure，并选择数据集中标准的分类结果作为参考。F-measure 指标综合了准确率与召回率两个指标，其值越大，表明该算法的聚类结果与数据集真实分类的差距越小，算法的可用性越高。

在实验参数设置上，隐私保护预算参数 ε 在[0.1,1]中线性取值，聚类数 $k$ 设置为数据集标签的类别数，OEDP k-means 算法在计算数据点 $r$-最近邻距离时的参数 $r$ 的取值为 4，DP-KCCM 算法中初始聚类中心点与域边界的距离参数 $a$ 的取值为 0.02，算法计划在重复迭代至最大迭代次数 12 次后，将 $4\times k$ 个簇合并成 $k$ 个簇，即相邻簇的个数参数 $n=4$。由于差分隐私 *K*-means 算法中加入了符合 Laplace 分布的随机噪声，以保护数据隐私，聚类结果具有随机性，因此实验选取各算法 5 次运算结果的平均值。对于数据集 Iris 和 HTRU2，当隐私预算参数 ε 取值从 0.1 变化到 1 时，算法可用性度量指标 F-measure 值的变化情况如图 5 和图 6 所示。

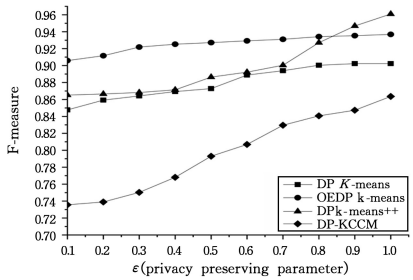



图 5 Iris 数据集上各算法的可用性

Fig. 5 Availability of algorithms on Iris dataset

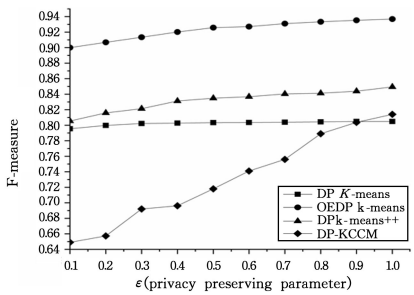



图 6 HTRU2 数据集上各算法的可用性

Fig. 6 Availability of algorithms on HTRU2 dataset

由图 5 和图 6 可知，随着隐私保护预算参数 ε 取值的增大，生成的随机噪声会减小，算法的隐私保护级别降低，F-measure 值会随之增大，即 4 个差分隐私 *K*-means 算法的可用性会逐步提升。针对数据集 Iris 和 HTRU2，算法 DPk-means＋＋和 DP-KCCM 受参数 ε 取值的影响较大，DP *K*-means 的表现较为稳定；OEDP k-means 算法优化了初始中心点选择策略并排除了离群点的影响，在 4 种算法中表现最优；其中，DP-KCCM 算法受实验参数设置及噪声方向不可控的影响，聚类结果波动较大。

## 4 亟待解决的问题及发展趋势

近年来，越来越多的研究者尝试研究隐私保护数据挖掘技术，差分隐私 *K*-means 算法自提出至今虽有很多改进，但大多处于研究阶段，且算法本身具有一定的局限性，实际应用较少。在前文综合介绍 DP *K*-means 算法研究现状的基础上，本节将讨论该算法亟待解决的问题与发展趋势。

### 4.1 亟待解决的问题

DP *K*-means 算法在实际的数据挖掘应用中，仍存在许多难题与挑战，主要有以下几个方面。

(1)数据特征的敏感性

DP *K*-means 算法挖掘分析数据的优势在于，在无标签的情况下可以快速高效地学习发现数据的内在规律，劣势在于，该算法继承了 *K*-means 在 $k$ 值与初始中心点的选取和离群点处理上存在局限性，另外因挖掘分析任务中提供数据的特征不同，差分隐私聚类算法为平衡隐私性与可用性，在设计上存在较大的差异。*K*-means 算法本身对噪声数据敏感，数据中的噪声点与离群点数据会影响聚类中心点的选择与优化，使聚类中心点发生较大的偏移。DP *K*-means 算法在 *K*-means 的基础上加入噪声，以保护聚类中心点隐私会带来新的难题，在聚簇更新阶段，不同尺寸的聚簇和多次迭代中加入的噪声应有所区别，同时对噪声的分配与控制往往可决定聚类中心点的移动方向、算法的收敛速度和准确率。现有的 DP *K*-means 改进研究算法大多数只能适用于具有特定特点的数据集，在真实的挖掘分析任务提供的数据集中，数据的值域、数据量规模、数据密度、数据形状、样本分布及均衡情况均存在差异，且需考虑混合数值型数据与非数值型数据的数据集的有效聚类与隐私保护。因此，如何基于数据的特征设计差分隐私 *K*-means 聚类算法，降低算法对数据集的敏感性，提升算法及数据的可用性是一个难题。

(2)隐私分配对聚簇更新、算法收敛与可用性的影响

DP *K*-means 算法中隐私预算的分配决定着算法的隐私保护程度，同时为提供隐私保护而加到每次迭代聚类中心点的噪声量对聚类过程会造成极大的影响。目前，常用的隐私预算分配方式为二分法、统一分配法，随后有研究者提出了基于最小隐私预算阈值的等差数列隐私预算分配方法及基于聚簇特征划分的自适应隐私预算分配方法，并结合数据的特征，以保证隐私预算的持续分配，但仍存在加入噪声过多，导致聚类中心点发生偏移、目标函数无法收敛的问题。因此，如何合理设计隐私预算分配方式，以保证在最小的隐私约束下达到最大化的聚类可用性也是研究的难点之一。

(3)高维、大型数据集隐私保护聚类的局限性

DP *K*-means 算法的研究大多数关注上述提到的两个难点，忽视了对高维、大型数据集的处理研究。在前文提到的

研究中只有 Zhang 等[56]在 MapReduce 框架下进行了大型高维数据集 Covertype 的聚类，但未对高维数据集做无关属性处理和降维操作，因此该算法不适用于处理高维稀疏数据集。在实际的数据挖掘任务中，大型高维数据居多，如何将算法拓展应用到大型高维数据集的聚类分析中是值得关注的研究热点。

### 4.2 发展趋势

根据 DP K-means 的研究现状与亟待解决的问题，主要有以下发展趋势。

(1)引入数据特征分析的相关理论

在海量数据的应用背景下，数据驱动的隐私保护挖掘技术需重点关注数据本身的特征，可引入数据特征分析的相关理论，从数据预处理、隐私预算分配等角度改进差分隐私 K-means 隐私保护模型。首先分析数据集的分布特征和分布类型，将数值型与非数值型数据划分开，然后需针对不同的数据类型采用不同的隐私保护机制，再通过数据密度、网格分布等方法实现数据集的离群值处理、初始聚类中心点划分和隐私预算的分配。针对维度较高的数据集，还需另外采用特征选择、降维、相关性分析等方法，对数据集进行降维处理[72-74]，找出相关性高且对聚类结果影响大的维度数据组成新的数据集，再输入差分隐私 K-means 模型进行聚类分析。

(2)融合其他机器学习或深度学习方法

针对 DP K-means 算法的局限性，可引入并融合其他机器学习或深度学习方法。目前引入机器学习或深度学习方法改进 DP K-means 算法的研究较少，导致该算法在实际应用中的适应性较低。可行的解决方案为：深入分析其他全局优化的机器学习或深度学习的思想及解决具体问题的调优策略，并将其融入到 DP K-means 算法中，以弥补算法的不足。

(3)结合同态加密、安全多方计算等技术增强隐私假设

差分隐私 K-means 算法在实际应用中受数据集特征与实际隐私需求的限制，设计出符合数据安全需求的交互式算法有一定难度，且较难保证所有存在数据挖掘需求的企业均有相应的数据安全挖掘和分析的技术人员。因此，基于云服务器的隐私保护数据挖掘外包服务将会是各企业的最优选择方案，以满足在无技术人员的情况下企业数据的安全挖掘和分析的需求。基于云服务器的差分隐私 K-means 聚类隐私保护模型已有相关研究，其未来的研究热点为与安全多方计算、同态加密等技术结合，提出实用且隐私假设更严格的数据隐私保护方案[75-77]。

**结束语** 本文关注隐私保护数据挖掘中差分隐私 K-means 聚类的研究，并对 DP K-means 算法原理、隐私攻击模型的不足进行了介绍，从 3 个角度对比分析了 DP K-means 的国内外研究现状，通过分析发现，该算法不仅对数据特征和隐私分配方式敏感，在处理大型高维数据集时，还存在一定的局限性，因此需要学者们进行进一步的分析研究。

## 参 考 文 献


[1] GAO Z, SUN Y, CUI X, et al. Privacy-Preserving Hybrid K-Means[J]. International Journal of Data Warehousing and Mining, 2018, 14(2): 1-17.

[2] NELSON B, OLOVSSON T. Security and privacy for big data: A systematic literature review[C]// IEEE International Conference on Big Data. IEEE, 2017.

[3] ZHANG X J, YANG H Y, LI Z, et al. Differentially Private Location Privacy-preserving Scheme with Semantic Location[J]. Computer Science, 2021, 48(8): 300-308.

[4] PENG C C, CHEN Y L, XUN Y M. k-modes Clustering Guaranteeing Local Differential Privacy [J]. Computer Science, 2021, 48(2): 105-113.

[5] ZHOU S G, LI F, TAO Y F, et al. Privacy Preservation in Database Applications: A Survey[J]. Chinese Journal of Computers, 2009, 32(5): 847-861.

[6] MIN Z, YANG G, SANGAIAH A K, et al. A privacy protection-oriented parallel fully homomorphic encryption algorithm in cyber physical systems[J]. EURASIP Journal on Wireless Communications and Networking, 2019, 2019(1): 15.

[7] LIU J, YIN S L, LI H, et al. A density-based clustering method for k-anonymity privacy protection[J]. Journal of Information Hiding and Multimedia Signal Processing, 2017, 8(1): 12-18.

[8] TEMUUJIN O, AHN J, IM D H. Efficient L-Diversity algorithm for preserving privacy of dynamically published datasets [J]. IEEE Access, 2019, 197: 122878-122888.

[9] DWORK C. Differential Privacy: A Survey of Results [C]// Theory and Applications of Models of Computation (TAMC 2008). Springer, 2008: 1-19.

[10] FANG Y, ZHU J, ZHOU W, et al. A Survey on Data Mining Privacy Protection Algorithms[J]. Netinfo Security, 2017, 2: 6-11.

[11] 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.

[12] INAN A, GURSOY M E, SAYGIN Y. Sensitivity analysis for non-interactive differential privacy: bounds and efficient algorithms[J]. IEEE Transactions on Dependable and Secure Computing, 2020, 17(1): 194-207.

[13] GAO Z Q, WANG Y T. Survey on differential privacy and its progress[J]. Journal on Communications, 2017, 38: 151-155.

[14] BLUM A, DWORK C, MCSHERRY F, et al. Practical Privacy: The SuLQ Framework[C]// Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems(PODS '05). 2016: 128-138.

[15] MCSHERRY F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[C]// Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2019: 19-30.

[16] NISSIM K, RASKHODNIKOVA S, SMITH A. Smooth sensitivity and sampling in private data analysis[C]// Thirty-ninth ACM Symposium on Theory of Computing. ACM, 2007.

[17] FELDMAN D, FIAT A, KAPLAN H, et al. Private coresets [C]// Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009. Bethesda, MD, USA: ACM, 2009.

[18] MOHAN P, THAKURTA A, SHI E, et al. GUPT: privacy preserving data analysis made easy[C]// Proceedings of the 2012

ACM SIGMOD International Conference on Management of Data. 2012:349-360.

[19] FLETCHER S, ISLAM M Z. Decision tree classification with differential privacy: A survey[J]. ACM Computing Surveys, 2019, 52(4):1-33.

[20] HAY M, MACHANAVAJJHALA A, MIKLAU G, et al. Principled Evaluation of Differentially Private Algorithms using DPBench[C]// ACM. 2016:139-154.

[21] YE Q Q, MENG X F, ZHU M J, et al. Survey on Local Differential Privacy[J]. Journal of Software, 2018, 29(7):1981-2005.

[22] HE X M, WANG X Y, CHEN H H. Study on choosing the parameter ε in differential privacy[J]. Journal on Communications, 2015, 36(12):124-130.

[23] CONSUL S, WILLIAMSON S. Differentially Private Median Forests for Regression and Classification[J]. arXiv: 2006.08795.

[24] ALVIM M, CHATZIKOKOLAKIS K, PALAMIDESSI C, et al. Invited Paper: Local Differential Privacy on Metric Spaces: Optimizing the Trade-Off with Utility[C]// 2018 IEEE 31st Computer Security Foundations Symposium (CSF). IEEE, 2018:262-267.

[25] BUN M, STEINKE T. Average-Case Averages: Private Algorithms for Smooth Sensitivity and Mean Estimation[J]. Advances in Neural Information Processing Systems, 2019, 17:181-191.

[26] WANG H, XU Z, XIONG L, et al. Conducting Correlated Laplace Mechanism for Differential Privacy[C]// International Conference on Cloud Computing and Security. Cham: Springer, 2017.

[27] DONG J, DURFEE D, ROGERS R. Optimal Differential Privacy Composition for Exponential Mechanisms and the Cost of Adaptivity[J]. arXiv: 1909.13830.

[28] YANG X, WANG T, REN X, et al. Survey on Improving Data Utility in Differentially Private Sequential Data Publishing[J]. IEEE Transactions on Big Data, 2021, 7(4):729-749.

[29] LIU F. Generalized Gaussian Mechanism for Differential Privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(4):747-756.

[30] TALWAR K, HARDT M A W. Geometric mechanism for privacy-preserving answers: U. S. Patent 8,661,047[P]. 2014-2-25.

[31] LI C, MIKLAU G, HAY M, et al. The matrix mechanism: optimizing linear counting queries under differential privacy[J]. VLDB Journal — the International Journal on Very Large Data Bases, 2015, 24(6):757-781.

[32] ZHAO J, CHEN Y, ZHANG W. Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions[J]. IEEE Access, 2019, 7:48901-48911.

[33] GONG M, XIE Y, PAN K, et al. A Survey on Differentially Private Machine Learning[J]. IEEE Computational Intelligence Magazine, 2020, 15(2):49-64.

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

[35] LIN W C, TSAI C F, KE S W, et al. Top 10 data mining techniques in business applications: a brief survey[J]. Kybernetes, 2017, 46(7):1158-1170.

[36] XIONG J, REN J, CHEN L, et al. Enhancing privacy and availability for data clustering in intelligent electrical service of IoT[J]. IEEE Internet of Things Journal, 2018, 6(2):1530-1540.

[37] DWORK C, ROTHBLUM G N, VADHAN S P. Boosting and Differential Privacy[C]// 2010 IEEE 51st Annual Symposium on Foundations of Computer Science. IEEE, 2010:51-60.

[38] YANG J C, ZHAO C. A Survey on K-Means Clustering Algorithm[J]. Computer Engineering and Applications, 2019, 55(23):7-14, 63.

[39] REN Y H. Survey of K-means algorithms on big data [J/OL]. Application Research of Computers. https://doi.org/10.19734/j.issn.1001-3695.2019.10.0581.

[40] XING Y N, QIAN Y R, NAN F Z, et al. Survey of optimization on K-means algorithm in Spark[J]. 2020, 37(3):641-647.

[41] LI Y, HAO Z F, WEN W, et al. Research on Differential Privacy Preserving k-means Clustering[J]. Computer Science, 2013, 3:287-290.

[42] REN J, XIONG J, YAO Z, et al. DPLK-Means: A Novel Differential Privacy K-Means Mechanism[C]// IEEE Second International Conference on Data Science in Cyberspace. IEEE, 2017.

[43] ZHANG G, ZHANG C, ZHANG H. Improved K-means Algorithm Based on Density Canopy[J]. Knowledge-Based Systems, 2018, 145:289-297.

[44] XIA D, NING F, HE W. Research on Parallel Adaptive Canopy-K-Means Clustering Algorithm for Big Data Mining Based on Cloud Platform[J]. Journal of Grid Computing, 2020, 18(2):263-273.

[45] LI L F. The analysis of K-means Clustering with Differential Privacy[D]. Sichuan: Southwest Jiaotong University, 2016.

[46] YAO S. An Improved Differential Privacy K-Means Algorithm Based on MapReduce[C]// 2018 11th International Symposium on Computational Intelligence and Design (ISCID). IEEE, 2018:141-145.

[47] SHANG T, ZHAO Z, GUAN Z, et al. A DP canopy k-means algorithm for privacy preservation of Hadoop Platform[C]// International Symposium on Cyberspace Safety and Security. Cham: Springer, 2017:189-198.

[48] SHAO R M, ZHANG L, LIU Y, et al. A-PAM Clustering Algorithm Based on Differential Privacy Preserving[C]// 2015 International Conference on Software, Multimedia and Communication Engineering. 2015.

[49] YU Q, LUO Y, CHEN C, et al. Outlier-eliminated k-means clustering algorithm based on differential privacy preservation[J]. Applied Intelligence, 2016, 45(4):1179-1191.

[50] FAN Y K, LIU J W. Parallel K-means algorithm with differential privacy preservation and outlier pruning[J]. Application Research of Computers, 2019, 6:1776-1781.

[51] MENG Q, ZHOU L. Research On Differential Privacy Preserving Clustering Algorithm Based On Spark Platform[J]. Journal of Computers (Taiwan), 2018, 29(1):47-62.

[52] DWORK C. Differential privacy [C]// International Colloquium on Automata, Languages, and Programming. Springer, 2006: 1-12.

[53] FAN Z, XU X. APDP k-means: A New Differential Privacy Clustering Algorithm Based on Arithmetic Progression Privacy Budget Allocation [C] // 2019 IEEE 21st International Conference on High Performance Computing and Communications. IEEE, 2019.

[54] SU D, CAO J, LI N, et al. Differentially Private K-Means Clustering[C]// ACM. 2016: 26-37.

[55] FU Y M, LI Z D. Research on k-means++ Clustering Algorithm Based on Laplace Mechanism for Differential Privacy Protection[J]. Netinfo Security, 2019, 218(2): 49-58.

[56] ZHANG Y, LIU N, WANG S, et al. A differential privacy protecting K-means clustering algorithm based on contour coefficients[J]. Plos One, 2018, 13(11): 1-15.

[57] MO R, LIU J, YU W, et al. A Differential Privacy-Based Protecting Data Preprocessing Method for Big Data Mining[C]// 2019 18th IEEE International Conference on Trust, Security And Privacy In Computing And Communications. IEEE, 2019.

[58] DONG S U, CAO J N E, NINGHUI L I, et al. Differentially Private K-Means Clustering and a Hybrid Approach to Private Optimization[J]. ACM Transaction on Information & System Security, 2017, 20(4): 16. 1-16. 33.

[59] ZHANG J, XIAO X, YANG Y, et al. PrivGene: differentially private model fitting using genetic algorithms[C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013: 665-676.

[60] LU Z, SHEN H. A convergent differentially private k-means clustering algorithm [C] // Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2019: 612-624.

[61] LU Z, SHEN H. Differentially Private k-Means Clustering with Guaranteed Convergence[J]. arXiv: 2002. 01043, 2020.

[62] ACS G, MELIS L, CASTELLUCCIA C, et al. Differentially private mixture of generative neural networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(6): 1109-1121.

[63] LV Z, WANG L, GUAN Z, et al. An Optimizing and Differentially Private Clustering Algorithm for Mixed Data in SDN based Smart Grid[J]. IEEE Access, 2019, 7: 45773-45782.

[64] NGUYEN H H. Privacy-Preserving Mechanisms for k-Modes Clustering[J]. Computers & Security, 2018, 78: 60-75.

[65] NI T, QIAO M, CHEN Z, et al. Utility-efficient Differentially Private K-means Clustering based on Cluster Merging-ScienceDirect[J]. Neurocomputing, 2021, 424: 205-214.

[66] GAO Z Q, ZHANG L J. DPHKMS: An Efficient Hybrid Clustering Preserving Differential Privacy in Spark[C]// International Conference on Emerging Internetworking. Cham: Springer, 2017.

[67] FRANK A, ASUNCION A. UCI machine learning repository [EB/OL]. [2022-01-13]. https://archive.ics.uci.edu/ml/index.php.

[68] PENG H L, JIN K Z, FU C C, et al. Private Time Series Pattern Mining with Sequential Lattice [J]. Acta Electronica Sinica, 2020, 48(1): 153-163.

[69] LI H C, WU X P, CHEN Y. k-means clustering method preserving differential privacy in MapReduce framework [J]. Journal on Communications, 2016, 37(2): 124-130.

[70] XIA C, HUA J, TONG W, et al. Distributed K-Means clustering guaranteeing local differential privacy[J]. Computers & Security, 2020, 90: 101699. 1-101699. 11.

[71] NGUYEN T D, GUPTA S, RANA S, et al. Privacy Aware K-Means Clustering with High Utility [C] // Pacific-Asia Conference on Knowledge Discovery & Data Mining. Springer International Publishing, 2016: 388-400.

[72] ZHANG T, ZHU T, XIONG P, et al. Correlated Differential Privacy: Feature Selection in Machine Learning[J]. IEEE Transactions on Industrial Informatics, 2019, 16(3): 2115-2124.

[73] RATHI M, RAJAVAT A. High Dimensional Data Processing in Privacy Preserving Data Mining [C] // 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT). IEEE, 2020.

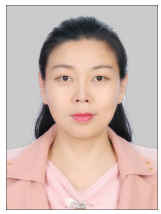
[74] ZHANG T, ZHU T, LIU R, et al. Correlated Data in Differential Privacy: Definition and Analysis[J]. Concurrency and Computation: Practice and Experience, 2020: e6015.

[75] TRAN H Y, HU J. Privacy-preserving big data analytics a comprehensive survey[J]. Journal of Parallel and Distributed Computing, 2019, 134: 207-218.

[76] GUAN Z, LV Z, DU X, et al. Achieving Data Utility-Privacy Tradeoff in Internet of Medical Things: A Machine Learning Approach[J]. Future Generation Computer Systems, 2019, 98: 60-68.

[77] SAKELLARIOU G, GOUNARIS A. Homomorphically encrypted k-means on cloud-hosted servers with low client-side load [J]. Computing, 2019, 101(12): 1813-1836.

**KONG Yu-ting**, born in 1997, postgraduate, is a member of China Computer Federation. Her main research interests include data mining and data security.



**QIAN Yu-rong**, born in 1980, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include network computing and remote sensing image processing.

（责任编辑：柯颖）