文章编号：2095-2163(2023)05-0126-06 中图分类号：TP391.9 文献标志码：A

基于密度权重的优化差分隐私 K-medoids 聚类算法

王圣节，巫朝霞
（新疆财经大学 统计与数据科学学院，乌鲁木齐 830012）

摘 要：K-medoids 算法作为数据挖掘中重要的一种聚类算法，与差分隐私保护的结合有助于信息数据的安全，原有的基于差分隐私保护的 K-medoids 聚类算法在初始中心点的选择上仍然具有盲目性和随机性，在一定程度上降低了聚类效果。本文针对这一问题提出一种基于密度权重的优化差分隐私 K-medoids（DWDPK-medoids）聚类算法，通过引入数据密度权重知识，确定算法的初始中心点和聚类数，以提高聚类效果和稳定性。安全性分析表明，算法满足 ε -差分隐私保护；通过对 UCI 真实数据集的仿真实验表明，相同隐私预算下该算法比 DPK-medoids 具有更好的聚类效果和稳定性。

关键词：数据挖掘；差分隐私；K-medoids 算法；密度权重

Optimal differential privacy K-medoids clustering algorithm based on density weights

WANG Shengjie, WU Zhaoxia
(College of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumqi 830012, China)

【Abstract】 As an important kind of clustering algorithm in data mining, the combination of K-medoids algorithm and differential privacy protection helps the security of information data. However, the original K-medoids clustering algorithm based on differential privacy protection is still blind and random in the selection of initial centroids, which reduces the clustering effect to some extent. To address this problem, an optimal differential privacy K-medoids (DWDPK-medoids) clustering algorithm based on density weights is proposed to determine the initial centroids and the number of clusters of the algorithm by introducing the knowledge of data density weights to improve the clustering effect and stability. The security analysis shows that the algorithm satisfies -differential privacy protection; the simulation experimental results on real UCI datasets show that the algorithm has better clustering effect and stability than DPK-medoids under the same privacy budget.

【Key words】 data mining; differential privacy; K-medoids algorithm; density weighting

0 引 言

大数据技术和人工智能飞速发展，海量数据的收集、存储、发布和分析越来越容易[1]。聚类分析作为数据挖掘的主要任务之一，已在许多领域内得到了广泛的应用，如：社交领域、外卖平台、电商网购。然而这些领域内的数据大都包含用户的隐私信息，容易遭受不法分子的攻击进而造成隐私泄露。因此，在聚类分析过程中，如何有效保护用户数据隐私，是当下研究热点，具有现实意义。

针对隐私泄露问题，早期学者提出 k-anonymity 及其一系列扩展模型，通过对准标识符的泛化处理来实现数据的隐藏，但该类模型容易受到一致性攻击和背景知识攻击，需要针对新型攻击不断完善模型[2]。2006 年，微软研究院的 Dwork[3] 提出了差分隐私技术（Difference Privacy，DP），不关心攻击者拥有的背景知识。差分隐私建立在严格的数学证明基础之上，通过在原始的查询结果（数值或离散型数值）中添加干扰数据（即噪声），再返回给第三方；加入干扰后，可以在不影响统计分析的前提下，无法定位到具体数据，从而防止个人隐私数据泄露，进而提供了强大的隐私保护。

Avrim 等人[4]最早将差分隐私保护与聚类技术相结合，设计了一种差分隐私 K-means（DPK-means）算法，布置于 SulQ 框架平台，通过发布添加合理噪声的相似值来更新查询值；李扬、郝志峰[5]

基金项目：国家自然科学基金（61941205）。

作者简介：王圣节（1997-），男，硕士研究生，主要研究方向：差分隐私保护；巫朝霞（1975-），女，博士，教授，硕士生导师，主要研究方向：信息安全。
通讯作者：巫朝霞 Emial：wuzhaoxia828@163.com

收稿日期：2022-12-12

等提出了满足 ε -差分隐私保护的 IDPK-means 算法，数据集平均划分成 p 个子集，计算子集加扰后的中心点作为初始中心点，以此提升最终聚类效果；吴伟民[6]等提出了一种基于差分隐私保护的 DP-DBScan 聚类算法，在添加少量噪声的情况下，保证了隐私安全与聚类效果；傅彦铭[7]等提出一种基于拉普拉斯机制的差分隐私保护 k-means++聚类算法（DPk-means++聚类算法），确保算法在不同维度数据集的情况下的聚类可用性和准确性。针对传统聚类算法存在的隐私泄露的风险，郑孝遥、陈冬梅[8]等提出一种基于差分隐私保护的谱聚类算法，以实现社交网络聚类效果和隐私的平衡；高瑜[9]等则针对数据集中噪声点和离群点地对聚类的影响，将差分隐私与 K-medoids 算法相结合，提出了一种满足 ε -差分隐私保护的聚类算法（DPk-medoids），该算法使用拉普拉斯机制在每次发布真实的中心点之前向中心点添加噪声，然后发布具有隐私保护性质的中心点，在一定程度上保证了个人隐私的安全性和聚类的有效性。

1　预备知识

1.1　差分隐私（Difference Privacy，DP）

差分隐私（Difference Privacy，DP）具有严格的数学定义，是一种通过任何模型、算法添加合理噪声的数据失真隐私保护技术，对可能被恶意攻击造成隐私泄露的关键部分添加干扰噪声，例如模型、算法的输入输出、梯度参数、权重参数、目标函数等，以保证模型、算法的隐私。

定义 1（差分隐私[3]）　假设存在任意一个随机算法 M，对于相邻数据集 D 和 D'，$Pr[E_S]$ 表示事件 E_S 的隐私披露风险，$Range(\mathrm{M})$ 表示随机算法 M 的取值范围，P_m 为输出结果的所有可能值的集合，S_m 为 P_m 的任意子集，如果算法 M 满足式(1)：

$$P_r[\mathrm{M}(D) \in S] \leqslant \exp(\varepsilon) \times P_r[\mathrm{M}(D') \in S] \quad (1)$$

则称算法 M 满足 ε -差分隐私。其中数据集 D 和 D' 为至多相差一条数据的相邻数据集。ε 为隐私预算，ε 越小，则算法 M 在 D 和 D' 输出分布越接近；反之，输出分布相差越大。

定义 2（全局敏感度[10]）　对于任何查询函数 $f: D \to R^d$，R 表示映射的实空间，d 表示函数 f 的查询维数，输入是一个数据集，输出是 d 维查询结果。对于任意相邻数据集，则查询函数的全局敏感度，式(2)：

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

定义 3（Laplace 机制[10]）　已知任意查询函数 $f: D \to R^d$，给定数据集 D，敏感度为 Δf，令 A 表示隐私聚类算法，若 $A(D)$ 满足式(3)：

$$A(D) = f(D) + \left(Lap_1\left(\frac{\Delta f}{\varepsilon}\right), Lap_2\left(\frac{\Delta f}{\varepsilon}\right), \cdots, Lap_d\left(\frac{\Delta f}{\varepsilon}\right) \right)^{\mathrm{T}} \quad (3)$$

则称隐私聚类算法 A 满足 ε -差分隐私保护，其中 $Lap\left(\frac{\Delta f}{\varepsilon}\right)$ 是服从尺度参数为 $\frac{\Delta f}{\varepsilon}$ 的 Laplace 分布函数。

1.2　差分隐私 K-medoids 算法

原有的差分隐私 K-medoids 算法在 K-medoids 聚类过程中对中心点添加少量合适的噪声，使中心点的披露风险满足差分隐私定义，以此为最终的聚类结果提供差分隐私保护。现有应用在用户数据基于差分隐私的 K-medoids 聚类算法（DP K-medoids）的具体步骤如下：

Step 1　随机选择 p 个点作为初始聚类中心点，并在每个中心点上加入拉普拉斯噪声；

Step 2　计算数据集中每个点与中心点之间的距离，并将该点从其最近的中心点分配给簇，形成 k 个簇；

Step 3　在每个集群中，依次选择一个点来计算用该点替换原始中心点所产生的消耗，并选择一个消耗少于原始中心点的点作为一个新的中心点，并对该点添加噪声；

Step 4　重复步骤 2 和步骤 3，直到迭代次数达到阈值。

2　基于密度权重的差分隐私保护 K-medoids 算法

2.1　相关概念

假设存在数据集 $D = \{x_1, x_2, \cdots, x_n\}$，是包含 n 个样本对象的数据集合，每个样本对象存在 d 维属性特征；数据集被划分为 k 个簇类，C 为簇集合：$C = \{c_1, c_2, \cdots, c_k\}$。

定义 4　数据集 $D = \{x_1, x_2, \cdots, x_n\}$ 中，任意两样本对象 x_i 和 x_j 之间的距离用欧氏距离 $d(x_i, x_j)$，表示简记为 d_{ij}，公式(4)：

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{p} (x_{ip} - x_{jp})^2} \quad (4)$$

其中：$i = 1, 2, \cdots, n$；$j = 1, 2, \cdots, n$；$p = 1, 2, \cdots, d$；x_{ip} 表示第 i 个数据对象的第 p 维属性。

定义 5　数据集 $D = \{x_1, x_2, \cdots, x_n\}$ 中，所有样

本对象的平均距离,公式(5):

$$AverDis(D) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d(x_i, x_j) \quad (5)$$

其中,n 是数据集 D 中样本点的个数,d_{ij} 表示数据集中样本点 x_i 到样本点 x_j 之间的欧氏距离。

定义 6 样本点 x_i 的密度参数是以数据集中任意样本对象 x_i 为中心,$AverDis(D)\rho(i)$ 为半径的区域样本对象的总数,公式(6):

$$\rho(i) = \sum_{j=1}^{n} S(d(x_i, x_j) - AverDis(D)) \quad (6)$$

当 $x < 0$ 时,$S(x) = 1$,$x \geqslant 0$ 时,$S(x) = 0$。

样本点密度图如图 1 所示。

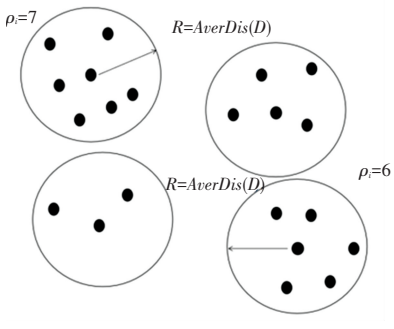


图 1 样本点密度

Fig. 1 Sample point density

定义 7 设 u_i 是以样本对象 x_i 为中心,$AverDis(D)$ 为半径的区域内所有样本对象集,$d(x_i, x_j) < AverDis(D)$,则式(7):

$$u_i = \frac{2}{\rho(i)[\rho(i)-1]} \sum_{i=1}^{\rho(i)} \sum_{j=i+1}^{\rho(i)} d(x_i, x_j) \quad (7)$$

定义 8 类簇之间距离 s_i 代表样本点 x_i 与另一局部密度较高的样本点 x_j 之间的距离。如果样本点 x_i 的局部密度是最大值,s_i 则定义为 $\max(d_{ij})$;否则 s_i 被定义为 $\min(d_{ij})$,公式(8):

$$S_i = \begin{cases} \min\limits_{x_j:\rho(j)>\rho(i)}(d_{ij}), & \text{若 } \exists j, \text{则} \rho(j) > \rho(i) \\ \max\limits_{x_j \in D}(d_{ij}), & \text{若 } \forall j, \text{则} \rho(j) \leqslant \rho(i) \end{cases} \quad (8)$$

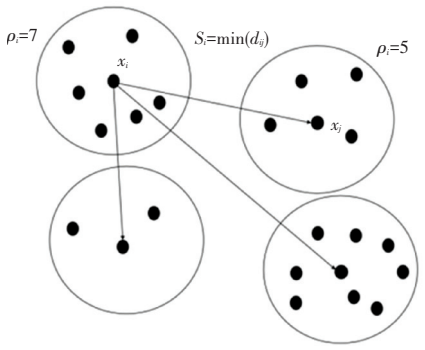
样本点类簇距离的计算原则如图 2 所示。

定义 9 假设数据集 D 被划分为 k 个聚类,聚类 $C_j(j \leqslant k)$ 的中心点为 c_j,聚类结果的误差平方和是每个聚类的样本与其聚类中心之间的距离的平方和,即式(9):

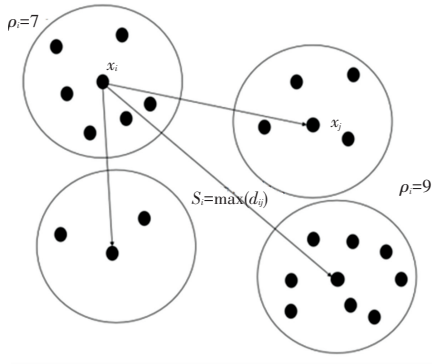
$$E = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_i - c_j), x_i \in c_j \quad (9)$$

定义 10 定义选取中心点的密度权重如式(10):

$$w_i = \rho(i) \times \frac{1}{u_i} \times s_i \quad (10)$$



(a) 情况 1:若存在 $\rho_j > \rho_i$



(b) 情况 2:若不存在 $\rho_j > \rho_i$

图 2 样本点类簇距离计算原则

Fig. 2 Calculation principle of cluster distance of sample points

通过公式(10)可知,样本点 x_i 的密度参数 ρ_i 越大,代表点 x_i 周围的样本点越密集,则密度权重 w_i 越大;u_i 越小表明以样本对象 x_i 为中心,$AverDis(D)$ 为半径的区域内样本对象相似程度越高,则密度权重 w_i 越小;类簇之间距离 s_i 越大,即代表两个类簇之间距离远,簇中样本点相似程度低,则密度权重 w_i 越大。

2.2 算法思想

为减少 DP K-medoids 算法在初始中心点选择的随机性,影响最终聚类结果,引入密度权重这一概念。选取权重最大的样本点作为聚类中心,以此来减少随机选取中心点带来的不确定性,提高聚类效果。算法步骤如下:

第一步,根据公式(6)计算数据集 D 中所有样本点的密度大小,取密度最高的样本点 c_1 设置成第一个聚类中心,并将该聚类中心 c_1 添加到中心集合 C 中,此时中心集合 $C = \{c_1\}$;然后,将满足定义 7 中样本点到初始簇中心之间的距离小于 $AverDis(D)$ 条件的所有样本点添加到当前聚类簇中,并将这些样本点从数据集 D 中移除;

第二步，根据公式(6)～公式(8)计算剩余样本点的密度 $\rho(i)$、簇间平均距离 u_i 和类簇距离 s_i，根据公式(10)计算剩余样本点的密度权重，选择密度权重系数最大的样本点 c_2 作为第二个聚类中心，将中心点 c_2 添加到中心集合 C 中，此时中心集合中 $C=\{c_1, c_2\}$；同样，从数据集中删除距离小于 $AverDis(D)$ 的剩余样本点与该初始聚类中心之间的所有样本点；

第三步，重复第二步至数据集 D 被清空。此时聚类中心集合 $C=\{c_1,c_2,\cdots,c_k\}$，数据集被划分成 k 个类簇，将得到的聚类数和聚类中心作为输入，对数据集进行 DP K-medoids 运算。

2.3　基于密度权重的优化 DP K-medoids 算法流程

DP K-medoids 在传统 K-medoids 算法的基础上添加了差分隐私保护技术，但也容易受初始中心点、聚类个数等因素影响其聚类效果。当输入数据或参数处理不当时，算法容易陷入局部最优解[11]。因此，本文通过对原有算法在初始中心点及聚类个数的选择上进行改进，选择密度权重较大的样本点作为初始中心点，再将结果产生的中心点与簇类数作为差分隐私的 K-medoids 算法的输入参数，避免初始中心点和 k 值选择的随机性，进而产生具有隐私保护能力的良好聚类结果。

算法流程如下：

输入　初始样本数据集 $D=\{x_1,x_2,\cdots,x_n\}$

输出　带有差分隐私保护的聚类结果

Step 1　输入原始数据集 $D=\{x_1,x_2,\cdots,x_n\}$；

Step 2　根据定义 7 计算数据集中所有样本点的密度大小，选取密度最大的样本点 c_1 作为第一个聚类中心，添加到中心集合中，$C=\{c_1\}$；

Step 3　计算剩余样本点的密度 $\rho(i)$、簇间距离 S_i、簇内样本点距离和 u_i；

Step 4　确定第二个聚类中心的条件：根据定义 11 计算得到剩余样本点的密度权重最大的样本点，将其添加到中心集合中，$C=\{c_1,c_2\}$；

Step 5　重复 Step3～Step4，直至数据集清空，此时 $C=\{c_1,c_2,\cdots,c_k\}$；

Step 6　将上述步骤得到的聚类中心和聚类数作为输入；

Step 7　对中心集合 $C=\{c_1,c_2,\cdots,c_k\}$ 添加拉普拉斯噪声，式(11)：

$$Lap(b)=e^{(-|x|/b)} \tag{11}$$

其中，$b=\Delta f/\varepsilon$，返回 $C'=\{c_1',c_2',\cdots,c_k'\}$；

Step 8　计算剩余样本点与中心点 $C'=\{c_1',c_2',\cdots,c_k'\}$ 的距离，按最近原则分配到中心点，形成簇；

Step 9　对于产生的每个初始簇，计算其簇中每个样本点与簇中其他样本点的距离和，选择距离和最小的样本点，更新为该簇的新中心点，并在新的中心点添加拉普拉斯噪声；

Step 10　重复 Step9～Step10，当中心点稳定不再发生改变或是达到预设的迭代次数，终止循环；

Step 11　输出带有差分隐私保护的聚类。

2.4　安全性分析

在 DWDPK-medoids 算法中，通过添加适量服从拉普拉斯分布的噪声对中心点进行隐私保护，使最终的聚类结果满足差分隐私定义。假设 $M(D)$ 和 $M(D')$ 代表经过 DWDPK-medoids 算法的在数据集 D 和 D' 的输出结果，D_1 和 D_2 是两个只相差一个记录的相邻数据集，S 代表一种随机的聚类划分方法，$\varphi(x)$ 为添加噪声之后的聚类结果，则有式(12)：

$$\frac{P_r[M(D)\in S]}{P_r[M(D')\in S]}=\frac{\exp\left(-\frac{\varepsilon\,|\,\varphi(x)-r(D,x)\,|}{\Delta f}\right)}{\exp\left(-\frac{\varepsilon\,|\,\varphi(x)-r(D',x)\,|}{\Delta f}\right)}=$$

$$\exp\left(\frac{\varepsilon\,|\,\varphi(x)-r(D',x)\,|-\varepsilon\,|\varphi(x)-r(D,x)\,|}{\Delta f}\right)\leqslant$$

$$\exp\left(\frac{\varepsilon\,\|\,r(D,x)-r(D',x)\,\|}{\Delta f}\right)\leqslant\exp(\varepsilon) \tag{12}$$

式(12)符合差分隐私的定义。因此，DWDPK-medoids 算法可以提供 ε -差分隐私保护。

3　实验结果与分析

3.1　实验环境及数据

通过实验对 DWDPK-medoids 算法的可用性进行分析。实验环境为 Windows 10 操作系统，Intel(R) Core(TM) i5-6200U CPU @ 2.40 GHz，12 GB 机带 RAM，采用 Python3.6 编程语言，通过 Pycharm 专业版编辑器进行仿真实验。数据来源于 UCI Knowledge Discovery Archive database 官网的真实数据集，见表 1。

表 1　实验数据集信息

Tabl. 1　Experimental data set information

数据集	样本数	属性类型	属性数
Iris	150	Real	4
Blood	210	Real	5
WiFi	2 000	Real	7

3.2　实验评价指标

本文采用戴维森堡丁指数(Davies-Bouldin

Index,DBI)指数作为聚类评价有用性指标。通过计算簇与簇之间相似度,再通过计算所有相似度的平均值,衡量整个聚类结果的优劣。如果簇与簇之间的相似度越高(*DBI* 指数高),说明簇与簇之间的距离越小,此时聚类效果就越差,反之越好。

R_{ij} 表示簇 C_i 与簇 C_j 的相似度,公式(13):

$$R_{ij} = \frac{S_i + S_j}{S_{ij}} \tag{13}$$

其中,S_i 代表簇 C_i 内所有样本点到中心点 c_i 的平均距离,S_{ij} 表示第 i 个簇与第 j 个簇之间的距离(即两个簇中心之间的距离)。

$$DBI = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}(R_{ij}) \tag{14}$$

其中,N 是聚类簇数。

3.3 实验结果分析

考虑到不同数据集维度不同,数据大小指标不一样,对数据集做 0-1 归一化处理。分别在 3 个数据集上运行 DPK-medoids 算法以及本文提出的 DWDPK-medoids 算法。由于添加服从拉普拉斯分布噪声的随机性,对每个算法进行 10 次实验,取其平均值。*DB* 值越小,证明聚类有效性越高。仿真实验结果如图 3~图 5 所示,横坐标代表隐私预算 ε,纵坐标代表聚类效果评价指标 *DB* 指数。

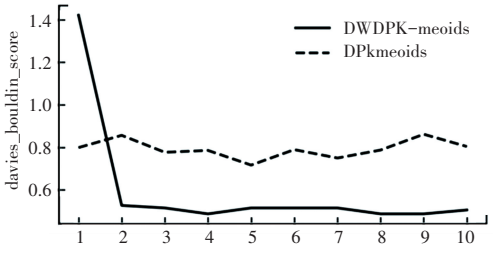


图 3 Iris 数据集上的 *DB* 指数图

Fig. 3 *DB* exponential graph on Iris dataset

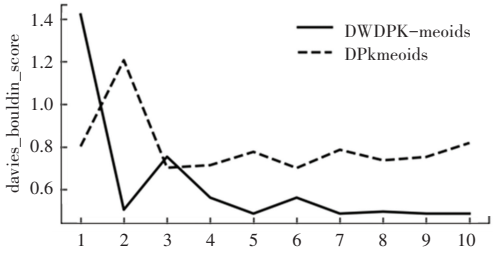


图 4 Blood 数据集上的 *DB* 指数图

Fig. 4 *DB* exponential graph on blood dataset

图 3 和图 4 为小样本数据集下的算法效果,可见在相同隐私预算情况下,DWDPK-medoids 算法的 *DB* 指数低于 DPK-medoids 算法,聚类效果更加好;在隐私预算 $\varepsilon = 2$ 的情况下,聚类效果开始趋于稳定,此时达到数据可用性和隐私保护平衡的状态。图 4 和图 5 是 Blood 数据集和 WiFi 数据集的对比图,数据集越大,*DB* 指数越大,最终的聚类效果越低。图 5 的 DWDPK-medoids 算法在 $\varepsilon = 1$ 时,整体聚类效果趋于稳定,而 DPK-medoids 算法在隐私预算 $\varepsilon = 2$ 时,*DB* 指数开始下降,随着 ε 的增大,逐渐与 DWDPK-medoids 算法 *DB* 指数相趋近,表明 DWDPK-medoids 算法能够消耗较小的隐私预算达到数据可用性与隐私保护的平衡。

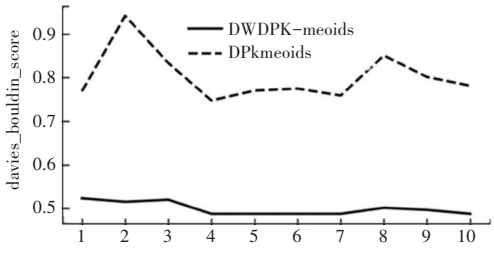


图 5 Wifi 数据集上的 *DB* 指数图

Fig. 5 *DB* exponential graph on wifi dataset

从实验结果可知,DWDPK-medoids 算法在一定程度上的聚类效果优于原有的 DPK-medoids 算法。DB 指数越低,则最终聚类结果所分的簇与簇之间距离越大,代表聚类效果越好。

4 结束语

差分隐私与 K-medoids 算法相结合在保证聚类效果的前提下,能够有效保护相关数据的安全性。针对 DPK-medo-ids 算法在选取初始中心点和聚类个数的随机性和盲目性,本文提出一种基于密度权重的优化差分隐私 K-medoids 算法(DWDPK-medoids 算法),通过引入密度权重的相关概念,确定初始聚类中心点以及聚类个数,提高聚类效果。仿真实验结果表明,DWDPK-medoids 算法在多样本和多维度的数据集上具有更高的聚类效果。在今后的研究中,应当针对算法复杂度进行优化,在保障数据隐私的前提下添加少量噪声获得更好的聚类效果。

参考文献

[1] 颜飞,张兴,李畅,等. 基于差分隐私的海量数据发布方法研究[J]. 计算机应用与软件,2018,35(11):314-320.

[2] Latanya Sweeney.k-anonymity:a model for protecting privacy[J]. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems,2002, 10(5):557-570.

[3] DWORK C. Differential privacy[C]//Proceedings of the 33rd international colloquium on automata, Languages and Programming. Berlin: Springer, 2006:1-12.

(下转第 139 页)