面向差分隐私保护的聚类算法

胡 闯[1,2] 杨 庚[1,2] 白云璐[1,3]

(南京邮电大学计算机学院 南京 210003)[1] (江苏省大数据安全与智能处理重点实验室 南京 210023)[2]

(南京中医药大学信息技术学院 南京 210023)[3]

摘 要 大数据时代的数据挖掘技术在研究和应用等领域取得了较大发展,但大量敏感信息披露给用户带来了众多威胁和损失。因此,在聚类分析过程中如何保护数据隐私成为数据挖掘和数据隐私保护领域的热点问题。传统差分隐私保护k-means算法对其初始中心点的选择较为敏感,而且在聚簇个数k值的选择上存在一定的盲目性,降低了聚类结果的可用性。为了进一步提高差分隐私k-means聚类方法聚类结果的可用性,研究并提出一种新的基于差分隐私的DPk-means-up聚类算法,同时进行了理论分析和比较实验。理论分析表明,该算法满足ε-差分隐私,可适用于不同规模和不同维度的数据集。此外,实验结果表明,在相同隐私保护级别下,与其他差分隐私k-means聚类方法相比,所提算法有效提高了聚类的可用性。

关键词 差分隐私,k-均值,聚类算法,隐私保护

中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.02.019

Clustering Algorithm in Differential Privacy Preserving

HU Chuang[1,2] YANG Geng[1,2] BAI Yun-lu[1,3]

(College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[1]

(Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing 210023, China)[2]

(College of Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China)[3]

Abstract Data mining has made great progress in the field of research and application of big data, but sensitive information disclosure could bring users many threats and losses. Therefore, how to protect data privacy in clustering analysis has become a hot issue in data mining and data privacy protection. Traditional differential privacy k-means is sensitive to the selection of its initial centers, and it has a certain blindness in the selection of cluster number k, which reduces the availability of clustering results. To improve the availability of clustering results of differential privacy k-means clustering, this paper presented a new DPk-means-up clustering algorithm based on differential privacy and carried out theoretical analysis and comparison experiment. Theoretical analysis shows that the algorithm satisfies ε-differential privacy, and can be applied to data sets with different sizes and dimensions. In addition, experimental results indicate that the proposed algorithm improves clustering availability than other differential privacy k-means clustering methods at the same level of privacy preserve.

Keywords Differential privacy, k-means, Clustering algorithms, Privacy preserving

1 引言

随着云计算和大数据的快速发展,数据挖掘技术在一些深入的研究和应用中取得了长足的进步[1]。作为数据挖掘的重要方法之一,聚类算法可以挖掘隐含的、未知的知识和规则,并且在大量相关数据的业务决策中具有重要潜在价值。例如,对海量数据进行挖掘并对动态数据进行聚类以开展科学研究、人口普查、市场预测、广告推广等。因此,聚类算法在统计数据分析、模式识别、图像处理、生物学和营销等领域具有广泛的应用前景。

但与此同时,大量敏感信息的披露将给用户带来无法估量的威胁和损失。因此,在聚类分析的同时引入隐私保护技术成为数据挖掘和数据隐私保护领域的热点问题。早期提出的一些隐私保护模型需要不断改进以抵御新的攻击,如背景知识攻击[2]和合成攻击[3]。而且其中一些模型未能平衡数据可用性和保密性。作为最常用的聚类方法之一,k-means算法的实现较简单,同时提供了高速聚类。一些文献已经涉及面向差分隐私保护的k-means聚类,但传统差分隐私保护

到稿日期:2018-01-29 返修日期:2018-04-19 本文受国家自然科学基金项目(61572263),江苏省自然科学基金政策引导类计划——前瞻性联合研究项目(2016ZS04)资助。

胡 闯(1994—),男,硕士生,主要研究领域为差分隐私保护;杨 庚(1961—),男,博士,教授,主要研究领域为网络与信息安全、分布式与并行计算、大数据隐私保护,E-mail:yangg@njupt.edu.cn(通信作者);白云璐(1988—),女,博士生,主要研究领域为信息安全与隐私保护。

k-means 算法对其初始中心点的选择较为敏感，而且在聚簇个数 k 值的选择上存在一定的盲目性，降低了聚类结果的可用性。例如，Blum 等[4]提出了一个差分隐私 k-means 方法，但是其聚类结果的可用性对于噪声不稳健。Li 等[5]提出了另一种基于初始中心的差分隐私 k 均值方法，利用 k 均值聚类来促进差分隐私，然而其选择初始中心时并未考虑聚类过程中异常值的负面影响。Yu 等[6]提出首先消除数据集中的异常点，再根据数据点的密度分布选择初始聚类中心点，并添加噪声到原始数据中的差分隐私 k 均值方法，但尚未证明该方法适用于大规模数据集聚类的隐私保护。Ren 等[7]提出在满足差分隐私保护的前提条件下，使用将数据集随机划分成多个子集后获得初始中心点的方式改进传统的 DPk-means；其不足之处在于数据集划分子集的个数较难确定，从而影响了聚类结果的稳定性。因此，本文提出了一种面向差分隐私的 DPk-means-up 均值隐私保护聚类算法。DPk-means-up 算法在 k 均值聚类算法的迭代过程中增加了满足特定分布的适当的随机噪声，使得聚类结果在一定程度上失真，达到了隐私保护的目的，同时保证了数据的可用性。

本文的贡献如下：

1）为保证 k-means 聚类算法的安全性，通过在 k-means 算法的中心点添加适当的噪声，设计了基于差分隐私的 DPk-means-up 算法，并证明了该算法满足差分隐私条件。

2）与现有的差分隐私 k 均值算法相比，所提算法以 k-means＋＋的结果作为输入值，然后通过交替进行一系列非局部“跳跃”并执行传统的 k-means 算法，改善了初始中心点的选择。它可以有效避免 k 值盲目性和初始点敏感性，并且能减少其迭代次数，从而提高聚类的可用性，同时保护隐私。

本文第 2 节介绍相关工作；第 3 节描述相关定义与理论基础；第 4 节先简要叙述传统的差分隐私 k-means 算法和差分隐私 k-means＋＋算法，接着介绍提出的一种新的差分隐私保护聚类算法；第 5 节进行仿真实验，分析与评估了第 4 节所提的新算法；最后总结研究内容并展望未来的研究方向。

2　相关工作

近年来，国内外掀起了一股差分隐私保护技术的研究热潮，尤其在数据挖掘领域，但大多处于理论研究阶段。国内差分隐私保护相关的研究相对少见。Dwork 自 2006 年第一次提出差分隐私（Differential Privacy，DP）[8]的概念后，研究者[9-12]持续补充和完善了差分隐私理论，其中文献[9]提出的 Laplace 实现机制是其实现的理论基础，并且针对于差分隐私保护中存在的流数据和连续观测现象问题，其提出了泛隐私（Pan-Privacy）的概念，相对于差分隐私，泛隐私的隐私保护强度更高。Li 等[13]构建了一种新模型，即将 k-匿名算法与差分隐私保护相结合，并将该模型应用于微数据的发布。除此以外，Li 等[14]提出了矩阵机制概念，将查询行为当作基本计数操作的线性组合，通过分析查询之间是否存在某种关联获得查询关系矩阵，为相应查询结果集添加噪声，降低噪声添加量。Xiao 等[15]针对计数查询，将 Haar 小波变换成小波树后添加噪声以实现差分隐私保护，同时高了查询结果的正确性，并同 Hay 等[16]研究出用柱状图来回答数据集任意范围查询的方法。Hardt 等[17]基于高维空间凸多面体均匀采样提出了 k-模算法，该算法给出了差分隐私保护模型噪声尺度的部分下界，但由于其特殊的取样条件，导致在现实中应用时效率较低。同时，Hardt 等[18]基于权重加乘的方法缩小在线查询时系统响应的误差边界。

针对应用于聚类中的差分隐私保护，Blum 等[4]于 2005 年提出在 SuLQ 平台上实现差分隐私 k-means 算法（Differential Privacy k-means Algorithm，DPk-means），但其查询函数敏感度较大且并未给出迭代过程中如何设置隐私预算，降低了聚类结果的可用性。Dwork[4]在 Blum 的基础上详细分析了差分隐私 k-means 算法中每个查询函数的敏感度计算方法，提出了两种情况下隐私预算的不同分配方法，并且给出了整个查询序列的总敏感度。Nissim 等[19]提出的 Pk-means 方法使 k-means 聚簇结果的中心满足差分隐私保护，同时给出了计算查询函数敏感度的方法和误差下界。Yu 等[6]提出的 OEDPk-means 方法根据数据点的密度分布选择初始中心点并且添加噪声到原始数据上，提高了聚类结果的可用性和聚类效率。Ren 等[7]为了提高 DPk-means 方法聚类结果的可用性，提出在满足差分隐私保护的前提条件下，使用将数据集随机划分为多个子集后获得初始中心点的方式改进传统的 DPk-means。Su 等[20]基于对现有的交互式和非交互式下的经验误差行为分析，提出了一个差分隐私 k-means 聚类概要的 EUGkM 非交互式方法。

3　定义与理论基础

3.1　差分隐私

差分隐私基于数据失真保护技术，通过添加随机噪声达到数据失真的效果，同时保持某些数据的属性或者数据在统计方面的性质不变。差分隐私保护技术确保了在数据集中任意添加或者删除一条记录不会影响最终的查询结果。

定义 1[8]　设有随机算法 M，如果数据集 D 和 D' 最多相差一条记录，$S\subseteq Range(M)$，若算法 M 满足：

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

则称算法 M 提供 ε-差分隐私保护，其中参数 ε 称为隐私保护预算。

定义 2[21]　设有查询函数 $f:D\to D^d$，对于任意的邻近数据集 D 和 D'，函数 f 的敏感度为：

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

其中，$\| f(D)-f(D')\|$ 是 $f(D)$ 和 $f(D')$ 之间的 1-阶范数距离。

差分隐私的主要实现机制是噪音扰动机制，Laplace 机制和指数机制是最常用的两种机制。其中 Laplace 机制针对数值型数据进行处理，指数机制则主要处理非数值型数据。噪音机制受全局敏感性和隐私预算制约。本文使用 Laplace 机制来实现差分隐私保护。

定理 1[7,19]（Laplace 机制）　给定数据集 D，设有函数 $f:D\to D^d$，其敏感度为 Δf，那么随机算法 $M(D)$：

$$M(D)=f(D)+Lap\left(\frac{\Delta f}{\varepsilon}\right) \tag{3}$$

提供 ε-差分隐私保护，服从尺度参数为 $\Delta f/\varepsilon$ 的 Laplace 分布。

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。记位置参数为 0、尺度参数为 b 的 Laplace 分布为 $Lap(b)$，其概率密度函数为：

$$\Pr[\eta=x]=\frac{1}{2b}\exp(-\frac{|x|}{b}) \tag{4}$$

另外，一个复杂的隐私保护问题通常需要多次应用差分隐私保护算法才能得到解决。在这种情况下，为了保证整个过程的隐私保护水平控制在给定的预算 ε 之内，需要合理地将全部预算分配到整个算法的各个步骤中。

性质 1[22] 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。

3.2 差分隐私保护 k-means 聚类

差分隐私聚类算法旨在确保当数据集的任何一条记录被删除时，簇中心的变化不会导致隐私被泄露。Blum 等[4]首次提出了 Lloyd 算法的差分隐私版本，我们称之为差分隐私 k-means 算法(DPk-means)。

DPk-means 算法的伪代码详见算法 1。在 DPk-means 中，每次迭代都以满足差分隐私保护的方式来更新簇的质心。给定一个簇，DPk-means 会计算簇集中的数据点的总个数，并将簇集中数据点的每个维度的坐标相加求和；然后通过添加拉普拉斯噪声来扰乱计数和求和查询。DPk-means 在经过多次迭代达到收敛条件后输出聚类后的簇集质心。

算法 1 DPk-means 算法

输入：数据集 D，数据维度 d，数据集范围[−r，r]，簇的数量 k，迭代次数 t，初始中心点集 IC，隐私预算 ε

输出：k 个聚类结果中心点$\{o^1,o^2,\cdots,o^k\}$

1. if IC is empty then
2. 随机选取 k 个点$\{o^1,o^2,\cdots,o^k\}$作为初始中心点；
3. else
4. $\{o^1,o^2,\cdots,o^k\}\leftarrow$IC；
5. end
6. $\varepsilon'\leftarrow\frac{\varepsilon}{(dr+1)t}$；
7. while iterate until t times do
8. for each j (j=1,2,…,k) do
9. Cluster $C^j\leftarrow\{\|x^l-o^j\|\leqslant\|x^l-o^i\|,x^l\in D,\forall 1\leqslant i\leqslant k\}$；
10. $\langle o_1^j,o_2^j,\cdots,o_d^j\rangle\leftarrow$NoisyCentroidUpdate(d,$C^j$,ε′)
11. end
12. end
13. return $\{o^1,o^2,\cdots,o^k\}$；
14. Function NoisyCentroidUpdate(d,C,ε)
15. Define $\prod_{[-r,r]}x=\begin{cases}-r, & \text{if } x<-r\\ x, & \text{if } -r\leqslant x\leqslant r\\ r, & \text{ir } x>r\end{cases}$
16. num←|C|+Lap($\frac{1}{\varepsilon}$)；
17. for each dimension i(i=1,2,…,d) do
18. $sum_i\leftarrow\sum_{x^l\in C}x_i^l+Lap(\frac{1}{\varepsilon})$；
19. $o_i\leftarrow\prod_{[-r,r]}(\frac{sum_i}{num})$；
20. end
21. return Cluster centroids $\langle o_1,o_2,\cdots,o_d\rangle$；

作为传统 k-means 算法的一种改进，k-means++算法由 Arthur 等[23]提出，并且在当今被视为执行 k-means 的一种标准方法(一个主要表现是用于科学计算的非常流行的 scikit-learn 软件包(Pedregosa 和 Varoquaux，2011)中 k-means 的默认实现方法就是 k-means++)。k-means++算法与 k-means 算法的主要不同之处在于在初始点的选取中，k-means++更偏向于选择远离现有中心的下一个中心点。k-mean++的差分隐私版本被称为差分隐私 k-means++算法(DPk-means++)，算法的具体细节详见算法 2。

算法 2 DPk-means++算法

1. Initialization：C←$\{c_1\}$，从样本集 X 中随机选择一个数据点作为 c_1；
2. i←1；
3. while i<k do /* 选取下一个中心点 */
4. i←i+1；
5. C←C∪{x} with x drawn at random from X with probability；
6. $P(x)=\frac{D(x)^2}{\sum_{x\in X}D(x)^2}$；
7. whereby $D(x)=\min_{c\in C}\|c-x\|$；
8. end
9. Perform DPk-means using C as the initial set of centers；
10. return resulting value of C.

4 差分隐私保护 DPk-means-up 聚类算法

为了减少 DPk-means 算法中聚类初始点的随机选择导致的聚类可用性较差，以及进一步提高 DPk-means++算法的效率，我们考虑改进初始中心点的选取方法，提出一种新的差分隐私 k-means-up(k-means with utility plus)算法，称之为 DPk-means-up 算法。

4.1 DPk-means-up 算法介绍

本节首先给出符号和相应的描述，如表 1 所列。

表 1 参数表示

Table 1 Parameter representation

符号	含义
X	数据集
λ	簇内最“无用”的中心点
μ	簇内距离平方和最大的中心点
C	中心点集
$\phi(C,X)$	给定的 X 和 C 下的误差平方和
ϕ_{best}	存放更新的最小误差
C_{best}	存放更新的最优中心点集
C_i	簇 i 的质心点
dim	数据集维度
sum	一个簇的点的总和
num	一个簇内的点的个数
sum'	sum 加噪后的结果
num'	num 加噪后的结果
ε	隐私保护预算
$dist(x,y)$	x 和 y 两点间的距离
E	误差平方和
$Lap(b)$	尺度参数为 b 的拉普拉斯噪声
$retry$	重试次数

算法的主要步骤如下所述。

第一阶段的具体步骤如下：

1)将 k-means++算法在数据集上运行得到的误差 $\phi(C,X)$ 和聚类中心点集 C 分别存储到 ϕ_{best} 和 C_{best} 中。

2)确定中心 μ 和 λ,并将簇内最"无用"的中心 μ 移动到具有最大簇内误差的中心的位置(加上一个小的随机数 o,也适用于 μ 本身,但符号相反)。

3)使用当前的 C 作为初始中心点执行 k-means 算法。

4)判断 $\phi(C,X)$ 与 ϕ_{best} 的大小。如果 $\phi(C,X)$ 小于 ϕ_{best},则转到步骤 1),否则进行下一步。

5)循环执行第一阶段描述中的步骤 2)－步骤 4),直到大于给定的重试次数最大值。

6)返回最优的中心点 C_{best}。

第二阶段的具体步骤如下:

7)遍历数据集 X 中的每个点,计算它到每个中心点之间的距离并将其分配到最近的中心点所在的簇,形成 k 个簇。

8)计算每个簇内数据点的总和及点的数量,分别添加噪声 $Lap(b)$,得到 $sum'=sum+Lap(b)$ 和 $num'=num+Lap(b)$。更新簇的质心为 sum'/num'。

9)重复第二阶段中的步骤 7)－步骤 8),直到误差平方和(sum of the squared errors)收敛。

两点之间的距离计算使用欧氏距离计算公式,计算方法如式(5)所示:

$$dist(x,y)=\sqrt{\sum_{i=0}^{dim}(x_i-y_i)^2} \tag{5}$$

误差平方和使用式(6)计算,并且值越小,聚类结果越独立和紧凑。

$$E=\sum_{i=1}^{n}\sum_{x\in C_i}\|x-C_i\|^2 \tag{6}$$

算法 3　选取聚类的中心点集

输入:结果中心点集 k-means++,样本数据集 X,簇的数量 k,偏移向量 ε,查询函数敏感度 Δf,隐私预算 ε

输出:差分隐私保护聚类结果

第一阶段:

```
1. Seeding:C←(result of k-means++)
2. φ_best←φ(C,X)
3. C_best←C
4. retry_max←n / * n∈{0,1,2,…} * /
5. retry←0 / * initialize retry counter * /
6. while retry≤retry_max do
7.     while true do
8.         λ←arg min_{c_i∈C} φ(C\{c_i},X) / * find least useful center * /
9.         μ←arg min_{C_i∈C} Σ_{x∈C_i} ‖x−C_i‖² / * find center with max local error * /
10.        u←random vector from d-dimensional unit hypersphere
11.        d_μ←√(1/|C_μ| Σ_{x∈C_μ} ‖x−μ‖²)
12.    o←∈d_μu / * offset vector,∈=0.01 * /
13.    λ←μ+o
14.    μ←μ−o
15.    Perform k-means using the current C as initial set of centers
16.    if φ(C,X)<φ_best then
17.        φ_best←φ(C,X)
18.        C_best←C
19.        retry←0/ * improvement! reset retry counter * /
20.    else
21.        break
22.    end
23.  end while
24.  retry←retry+1
25.  C_best←C
26. end while
27. return C_best
```

第二阶段:

```
28. double a[X.length]
29. double b=Δf/ε
30. while the sum of the squared errors is converges do
31.   for l←0 to X.length do
32.     for x←0 to X.length do
33.       a[k]=dist[X[x],initial[y]]
34.     end for
35.     find the minimum value of a
36.     categorize X[x] to the nearest center point
37.   end for
38.   for y←0 to k do
39.     sum=the sum of data points of the y cluster
40.     num=the number of data points of the y cluster
41.     sum′=sum+Lap(b)
42.     num′=sum+Lap(b)
43.     centerpoint=sum′/num′
44.   end for
45. end while
```

4.2　参数设置

实验中需要用到两个参数:ε(隐私预算)和 k(聚类的簇的个数)。

1)ε:选择合理的预算分配策略以使 ε 的生命周期尽可能长。流行的分配策略包括线性分布、均匀分布、指数分布、手动分配和混合分布[24]。通常,ε 倾向于设定在(0.01,0.1)的范围内,或者在一些情况下设定为 ln2 或 ln3[25]。因此,[0,1]区间内呈线性分布的分配方式一般用于实验中 ε 的分配。

设两个数据集 D_1 和 D_2 仅相差一条记录,计算 k 个中心点的过程就相当于对空间$[0,1]^d$ 进行划分的直方图查询,分母 $count$ 的敏感度为 1。对于分子 sum,由于数据集已被归一化分布在$[0,1]^d$ 中,因此分子的敏感度最大为 d。在 d 维空间$[0,1]^d$ 的数据点集中添加或删除一个点对于每一维的和,其敏感度为 1。因此,整个查询序列的敏感度为 $d+1$。

在聚类算法中,不同的数据集通常会执行不同的迭代次数才能达到收敛条件,Dwork 在文献[11]中提出了两种方法来设置隐私预算 ε:

①若迭代次数 N 固定,则每次迭代消耗的隐私预算为 ε/N,根据定理 1,每次可添加大小为 $Lap((d+1)N/\varepsilon)$ 的噪声来获取 ε-差分隐私保护;

②若迭代次数 N 未知，要在迭代过程中不断地调整隐私预算 ε 的值。

根据以往经验，前期迭代对聚类结果的影响要大于后期迭代。因此本文实验选择在聚类过程中逐步增加隐私预算，第一次分配的预算为 $\varepsilon/2$，噪声大小为 $Lap((d+1)N/\varepsilon)$，之后每次迭代消耗的预算是前一次的一半，直到最后一次迭代完成为止。

2)k：由于本文的目的是将差分隐私技术应用于 k-means 方法，而不是简单的聚类应用，并且我们优化了初始中心点的选取，因此，我们根据实验过程中使用的数据集所提供的参考分类的数量来选择 k 值。

4.3 安全性分析

算法中的随机函数若能够提供满足拉普拉斯分布的噪声，则可以为查询结果提供差分隐私保护。本文所提算法通过实现拉普拉斯机制，为聚类结果提供了差分隐私保护，即在聚类过程中添加适当的满足拉普拉斯分布的噪声到中心点。

假设 D_1 和 D_2 是仅相差一条记录的两个数据集，令 $M(D_1)$ 和 $M(D_2)$ 分别表示算法 DPk-means-up 在邻近数据集 D_1 和 D_2 上的输出结果，S 是任意一种聚类划分方式，其中 $\varphi(x)$ 表示加噪后的聚类查询结果，$r(D_1,x)$ 和 $r(D_2,x)$ 表示在数据集 D_1 和 D_2 的真实聚类查询，Δf 是查询函数的全局敏感度。安全性证明过程如下：

1)由定理 1 可知算法 M 通过对输出结果的随机化来提供隐私保护，因此

$$\Pr[M(D_1)\in S]=\Pr[Lap(b)=\varphi(x)-r(D_1,x)] \tag{7}$$

其中，$b=\Delta f/\varepsilon$。

2)由式(4)可得：

$$\Pr[Lap(b)=\varphi(x)-r(D_1,x)]=\frac{1}{2b}\exp\left(-\frac{|\varphi(x)-r(D_1,x)|}{b}\right) \tag{8}$$

其中，$b=\Delta f/\varepsilon$。

3)由步骤 1)和步骤 2)，可得：

$$\Pr[M(D_1)\in S]=\frac{1}{2b}\exp\left(-\frac{\varepsilon|\varphi(x)-r(D_1,x)|}{\Delta f}\right) \tag{9}$$

同理可得：

$$\Pr[M(D_2)\in S]=\frac{1}{2b}\exp\left(-\frac{\varepsilon|\varphi(x)-r(D_2,x)|}{\Delta f}\right) \tag{10}$$

4)由定义 2 可知：

$$\|r(D_1,x)-r(D_2,x)\|_1\leqslant\Delta f \tag{11}$$

再由步骤 3)可以得到：

$$\begin{aligned}\frac{\Pr[M(D_1)\in S]}{\Pr[M(D_2)\in S]}&=\frac{\exp\left(-\frac{\varepsilon|\varphi(x)-r(D_1,x)|}{\Delta f}\right)}{\exp\left(-\frac{\varepsilon|\varphi(x)-r(D_2,x)|}{\Delta f}\right)}\\&=\exp\left(\frac{\varepsilon(|\varphi(x)-r(D_2,x)|-|\varphi(x)-r(D_1,x)|)}{\Delta f}\right)\\&\leqslant\exp\left(\frac{\varepsilon|r(D_1,x)-r(D_2,x)|}{\Delta f}\right)\\&\leqslant\exp\left(\frac{\varepsilon\|r(D_1,x)-r(D_2,x)\|_1}{\Delta f}\right)\\&\leqslant\exp(\varepsilon)\end{aligned} \tag{12}$$

因此由定义 1 可知，DPk-means-up 满足 ε-差分隐私。

5 实验分析

5.1 实验数据和环境

本文使用 python 语言进行模拟仿真实验。实验环境为 Intel(R) Core(TM) i5-3317U CPU @1.70GHz，12GB 内存，Windows10 64 位操作系统。算法中所使用的数据集来自于文献[26]，具体信息如表 2 所列，其中 house8 数据集表示一组房屋的 RGB 值，其中每个颜色分量用一个八进制数表示；unbalance 数据集表示由 6500 个向量和 8 个高斯聚类合成的二维数据集；ConfLongDemo 数据集包含 8 个属性并且其中只有 3 个为数字属性；birch3 数据集表示一个随机位置和随机大小的簇组成的二维数据集。

表 2 数据信息

Table 2 Data information

数据集	别名	样本数	簇的数量	属性数	属性类型
house8	DS1	34113	3	3	Real
unbalance	DS2	6500	8	2	Real
ConfLongDemo	DS3	164860	11	11	Real
birch3	DS4	100000	100	2	Real

5.2 评价标准

首先，噪声对数据可用性的影响对于隐私保护来说非常重要。通常，数据可用性可以通过两种方式进行评估：理论分析和实践验证。对于前者，(β,γ)-可用性[27]通常用于衡量差分隐私算法的可用性；对于后者，流行的可用性度量包括相对误差、绝对误差、欧拉函数和 F-measure[28]。合适的指标的选择取决于使用的具体数据。

在本文中，由于参考类已经由选定的数据集提供，因此使用 F-measure 来评估聚类结果可用性。F-measure(也称为 F 分数)是与信息检索中准确率和召回率有关的聚类可用性的标准度量方法。与其他评价指标相比，F-measure 的结果更具有针对性。假设 n 代表给定的数据集 a 的大小，i 代表数据集的正确分类的类标签，n_i 和 n_j 分别代表类 i 和簇 C_j 中的数据点的数量，n_{ij} 代表类 i 和簇 C_j 的交集部分中的数据点的数量，则准确率和召回率的计算公式定义如下：

$$precision(i,j)=\max_{i,j}\left\{\frac{n_{ij}}{n_j}\right\} \tag{13}$$

$$recall(i,j)=\max_{i,j}\left\{\frac{n_{ij}}{n_i}\right\} \tag{14}$$

对于给定的类 i 和簇 C_j，F-measure 计算公式的定义如下：

$$F\text{-}measure(i,j)=\frac{(\beta^2+1)\cdot precision(i,j)\cdot recall(i,j)}{\beta^2\cdot precision(i,j)+recall(i,j)} \tag{15}$$

设置 $\beta=1$ 以使 $precision(i,j)$ 和 $recall(i,j)$ 获得相同的权重。对于一个大小为 n 的数据集，整个 $F\text{-}measure$ 值的计算如下：

$$F=\sum_i \frac{n_i}{n}\max_j\{F\text{-}measure(i,j)\} \tag{16}$$

F-measure 值的范围为[0,1]，该值越大意味着算法有更好的聚类可用性。

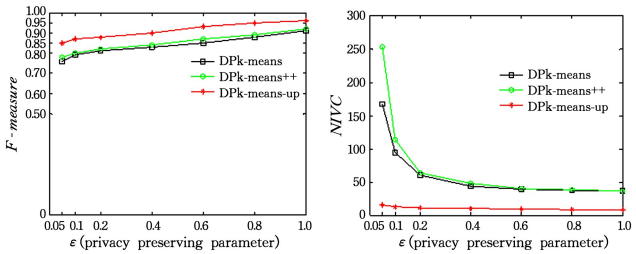
其次，我们使用规范化簇内方差（Normalized Intracluster Variance，NICV）[19] 来评估聚类的性能，*NICV* 的计算公式如下：

$$NICV=\frac{1}{N}\sum_{i=1}^{n}\sum_{x\in C_i}\|x-C_i\|^2 \tag{17}$$

其中，C_i 是第 i 个质心，N 为数据集大小，x 为样本数据集。

5.3　实验结果与分析

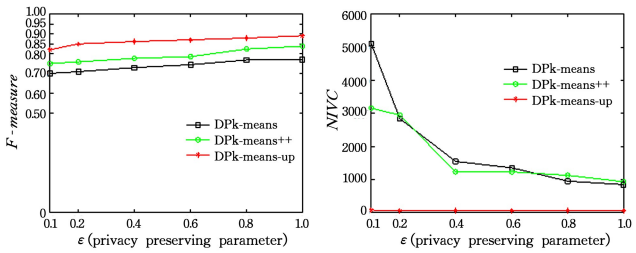
我们在 4 个包含分类标签的数据集上运行 DPk-means，DPk-means＋＋和 DPk-means-up 算法。首先对 DS1－DS4 4 个数据集进行预处理，然后将每个属性的取值归一化到[0,1]上，再对 4 个数据集分别进行差分隐私 k-means 聚类和 IDP k-means 聚类，并逐步将 ε 的值从 0 调高到 1.0。实验结果显示的是对应的每个 ε 值，在 4 个数据集上分别调用 3 个算法聚类 50 次后得到的 *F-measure* 和 *NICV* 的平均值。图 1(a)－图 4(a)和图 1(b)－图 4(b)分别展示了在数据集 DS1－DS4 上执行 3 种算法得到的 *F-measure* 值和 *NICV* 值的比较。



(a) *F-measure*　　(b) *NICV*

图 1　DS1 上运行的结果

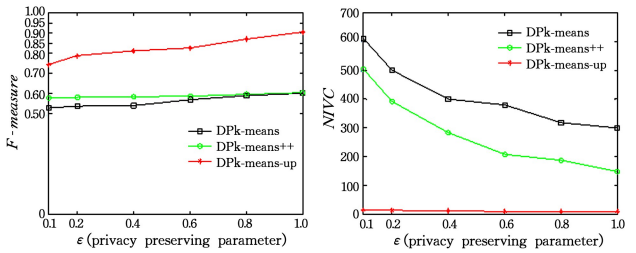
Fig. 1　Results on DS1



(a) *F-measure*　　(b) *NICV*

图 2　DS2 上运行的结果

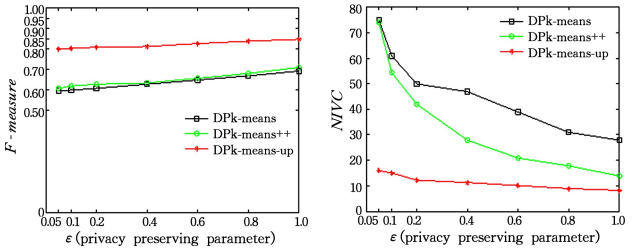
Fig. 2　Results on DS2



(a) *F-measure*　　(b) *NICV*

图 3　DS3 上运行的结果

Fig. 3　Results on DS3



(a) *F-measure*　　(b) *NICV*

图 4　DS4 上运行的结果

Fig. 4　Results on DS4

F-measure 值越大，加入噪声前后的聚类结果越相似，算法的可用性越好。同时，一个更小的簇内方差（*NICV*）提高了算法的性能。

1）聚类可用性分析

如图 1(a)－图 4(a)所示，在相同的 ε 值下，DPk-means-up 算法相比于其他算法具有更高的 *F-measure* 值，因此所提算法的聚类结果更接近于原始数据，更好地保证了聚类的可用性。随着差分隐私预算值的增加，*F-measure* 值也随之增加，这表明聚类结果随着隐私水平的降低而提高。

2）聚类性能分析

如图 1(b)－图 4(b)所示，在相同的 ε 值下，DPk-means-up 算法的 *NICV* 值明显小于其他两种算法。而且 DPk-means 和 DPk-means＋＋算法的 NICV 曲线随着值的增大波动明显，而本文算法则相对稳定，这些结果表明所提算法优于其他两种算法，这主要是由于 DPk-means-up 算法对初始中心点的优化选择改善了聚类的效果。

3）其他分析

由表 2 可知，DS4 数据集的属性数量大于另外 3 个数据集。然而，从图 4 可以看出，实验结果并未由于此因素而出现不稳定的情况。类似地，与 DS1 和 DS2 数据集相比，DS3 数据集不仅具有更多的属性，也包含大量的记录，而且我们可以看到，图 3 中的实验结果仍然是稳定的，这些结果表明了 DPk-means-up 均值算法也适用于大规模数据集和多维数据集。

因此，与 DPk-means 聚类和 DPk-means＋＋聚类相比，所提算法具有更高的聚类可用性和聚类性能。

结束语　为了解决在现有的差分隐私 k-means 算法中的聚类结果效率不高的问题，本文提出了一种新的 DPk-means-up 算法。它通过改进初始聚类中心的选择有效提高了差分隐私聚类的效率和聚类结果的可用性，同时更好地保护了隐私。在今后的工作中，我们计划使用不同的隐私预算分配策略来提高提议的安全性，通过优化本文算法进一步提高聚类结果的准确性并且将进一步探索 DPk-means-up 方法的应用。

参 考 文 献

[1] MADDEN S，FRANKLIN M J，HELLERSTEIN J M. TAG：a tiny aggregation service for ad-hoc sensor networks[C]// Proceedings of the 5th Symposium on Operating Systems Design and Implementation. New York，USA，2002：131-146.

[2] MACHANAVAJJHALA A，GEHRKE J，KIFER D，et al. l-di-

versity:Privacy beyond k-anonymity[C]// Proceedings of the 22nd International Conference on Data Engineering. IEEE, 2006:24-24.

[3] GANTA S R,KASIVISWANATHAN S P,SMITH A. Composition attacks and auxiliary information in data privacy[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 265-273.

[4] BLUM A,DWORK C,MCSHERRY F,et al. Practical privacy: the SuLQ framework[C]// Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM,2005:128-138.

[5] LI Y,HAO Z,WEN W,et al. Research on differential privacy preserving *k*-means clustering [J]. Computer Science,2013, 40(3):287-290.

[6] 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.

[7] REN J,XIONG J,YAO Z,et al. DPLK-means:a novel differential privacy *k*-means mechanism[C]// 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC). IEEE,2017:133-139.

[8] DWORK C. Differential privacy[C]// Proceedings of the 33rd International Conference on Automata,Languages and Programming-Volume Part II. Springer,Berlin,Heidelberg,2006:1-19.

[9] DWORK C,MCSHERRY F,NISSIM K,et al. Calibrating noise to sensitivity in private data analysis[C]// Theory of Cryptography Conference(TCC). 2006:265-284.

[10] DWORK C,LEI J. Differential privacy and robust statistics [C]// Proceedings of the 41st annual ACM Symposium on Theory of Computing. ACM,2009:371-380.

[11] DWORK C. A firm foundation for private data analysis[J]. Communications of the ACM,2011,54(1):86-95.

[12] DWORK C. Differential privacy[M]// Encyclopedia of Cryptography and Security. Springer US,2011:338-340.

[13] LI N,QARDAJI W,SU D. Provably Private Data Anonymization:Or,k-Anonymity Meets Differential Privacy[J/OL]. Corr, 2010,abs/1101.2604:32-33.

[14] LI C,HAY M,RASTOGI V,et al. Optimizing linear counting queries under differential privacy[C]// Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM,2010:123-134.

[15] XIAO X,WANG G,GEHRKE J. Differential privacy via wavelet transforms[J]. IEEE Transactions on Knowledge and Data Engineering,2011,23(8):1200-1214.

[16] HAY M,RASTOGI V,MIKLAU G,et al. Boosting the accuracy of differentially private histograms through consistency[J]. Proceedings of the VLDB Endowment, 2010, 3 (1-2): 1021-1032.

[17] HARDT M,TALWAR K. On the geometry of differential privacy[C]// Proceedings of the 42nd ACM Symposium on Theory of Computing. ACM,2010:705-714.

[18] HARDT M,ROTHBLUM G N. A multiplicative weights mechanism for privacy-preserving data analysis[C]// Proceedings of the 51st Annual IEEE Symposium. IEEE,2010:61-70.

[19] NISSIM K,RASKHODNIKOVA S,SMITH A. Smooth sensitivity and sampling in private data analysis[C]// Proceedings of the 39th annual ACM Symposium on Theory of Computing. ACM,2007:75-84.

[20] SU D,CAO J,LI N,et al. Differentially private *k*-means clustering and a hybrid approach to private optimization[J]. ACM Transactions on Privacy and Security (TOPS), 2017,20(4): 1-33.

[21] DWORK C. Differential privacy:A survey of results[C]// International Conference on Theory and Applications of Models of Computation. Springer,Berlin,Heidelberg,2008:1-19.

[22] 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,2009:19-30.

[23] ARTHUR D,VASSILVITSKII S. *k*-means++:The advantages of careful seeding[C]// Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics,2007:1027-1035.

[24] CHEN R,ACS G,CASTELLUCCIA C. Differentially private sequential data publication via variable-length n-grams[C]// Proceedings of the 2012 ACM Conference on Computer and Communications Security. ACM,2012:638-649.

[25] DWORK C. Differential privacy in new settings [C]// Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathema-tics,2010:174-183.

[26] FRÄNTI P. Clustering datasets[OL]. http://cs. joensuu. fi/sipu/datasets.

[27] NGUYEN H H,KIM J,KIM Y. Differential privacy in practice [J]. Journal of Computing Science and Engineering,2013,7(3): 177-186.

[28] JIANG H,YI S,LI J,et al. Ant clustering algorithm with K-harmonic means clustering[J]. Expert Systems with Applications,2010,37(12):8679-8684.

[29] FANG Y J,ZHU J Z,ZHOU W,et al. A survey on data mining privacy protection algorithms[J]. Netinfo Security,2017(2): 6-11. (in Chinese)
方跃坚,朱锦钟,周文,等. 数据挖掘隐私保护算法研究综述[J]. 信息网络安全,2017(2):6-11.

[30] ZHANG F X,JIANG C H. Research on query on privacy anonymity algorithm based on grid clustering[J]. Netinfo Security, 2015(8):53-58. (in Chinese)
张付霞,蒋朝惠. 一种基于网格聚类的查询隐私匿名算法研究[J]. 信息网络安全,2015(8):53-58.