doi: 10.7690/bgzdh.2023.12.009

基于隐私保护的改进 *K*-means 算法

王彩鑫，王丽丽，杨洪勇
（鲁东大学信息与电气工程学院，山东 烟台 264025）

摘要：针对传统 *K*-means 聚类算法聚类过程以及聚类结果公示时可能出现隐私泄露的问题，提出具有差分隐私保护的改进 *K*-means 算法。在原有 *K*-means 基础上引入密度度量，提高簇类的类内相似性，保证选取的中心处于相对密集区域；引入距离度量，降低簇类的类间相似性，保证不同类聚中心排斥性较高；引入类间平均最大相似度，动态规划最佳聚类个数 *K* 和最佳初始类内中心；引入了隐私保护拉普拉斯噪声，保护信息的安全性。实验结果表明，该算法比传统算法具有更高的聚类可用性和数据可靠性。

关键词：差分隐私；*K*-means 聚类；动态规划
中图分类号：TP393 **文献标志码：**A

Improved *K*-means Algorithm Based on Privacy Protection

Wang Caixin, Wang Lili, Yang Hongyong
(*School of Information and Electrical Engineering*, *Ludong University*, *Yantai* 264025, *China*)

Abstract: Aiming at the problem of privacy disclosure in the clustering process of traditional *K*-means clustering algorithm and the publicity of clustering results, an improved *K*-means algorithm with differential privacy protection was proposed. On the basis of the original *K*-means, density measurement is introduced to improve the in-class similarity of clusters and ensure that the selected centers are in relatively dense areas. The distance measure is introduced to reduce the similarity between clusters and ensure the high repulsion of different cluster centers. The average maximum similarity between classes is introduced, and the optimal number of clusters K and the optimal initial intra-class center are dynamically programmed. Privacy protection Laplacian noise is introduced to protect information security. Experimental results show that this algorithm has higher cluster availability and data reliability than traditional algorithms.

Keywords: differential privacy; *K*-means clustering; dynamic programming

0 引言

随着互联网在电子商务、金融、财政、军事等领域的大规模应用[1-4]，许多组织、公司都存储了大量数据信息。数据收集会导致隐私泄露，攻击者可以借助数据分析提取这些信息中的私人数据。近年来，数据安全以及隐私泄露的事件不断发生，如何保证数据隐私安全已成为大数据应用领域的研究热点，引起了海内外学者关注。文献[5]应用理性密码学与多方安全计算算法提出隐私保护框架；文献[6]对联邦学习中隐私泄露原因及防范措施进行叙述；文献[7]对近年的数据泄露通信内容进行分析；文献[8]提出差分隐私保护模型来解决数据库的信息泄露问题，该模型下单个数据在数据集中与该数据不在数据集中最终结果相差极小，攻击者无法获得真实原始数据。

由于其高效率及高速度的特点，*K*-means 聚类算法在许多领域得到广泛应用[9-11]。*K*-means 聚类算法仍有一些缺陷，例如以初始的随机数据作为聚类分析法中心、人工确定类簇个数以及聚类过程会出现隐私泄露问题。针对以上问题，文献[12]选取高密度数据集中的数据作为初始聚类中心，对初始聚类中心点进行改进并提高了聚类的准确性；文献[13]引入少量标记算法，在一定程度下解决了随机选择初始聚类中心的问题；文献[14]将最近邻算法与 *K*-mean 算法相结合，能有效提高算法准确性；文献[15]提出了基于 Canopy 算法的改进算法，能自动确定类簇个数并选择初始聚类中心；文献[16]提出了基于峰度检验的改进算法，提高了算法对复杂数据集的适应性；文献[17]将密度参数与 *K*-means 算法相结合，确定最优类簇个数；文献[18]提出了基于权重密度的改进算法，提高了聚类的精确性。

笔者在对已有优化算法学习的基础上，提出基于隐私保护的改进 *K*-means 算法，在原有 *K*-means 基础上引入密度度量，保证簇类的类内相似度较高；引入距离度量，保证簇类的类间相似度较低；引入

收稿日期：2023-08-13；**修回日期：**2023-09-07
基金项目：国家自然科学基金(61673200)；山东省重大基础研究项目(ZR2018ZC0438)
第一作者：王彩鑫(1999—)，女，山东人。

类间平均最大相似度，动态确定最佳类簇个数并能得到最佳初始类内中心；引入差分隐私保护机制，保护信息的安全性。

1　差分隐私保护技术

1.1　隐私保护定义

随着信息技术的深入普及，大量数据被发布共享。由于很多数据集中包含了许多私密数据，会导致敏感数据在整体数据被共享发布时发生泄露。虽然整体数据在共享之前已经删除其标识符属性，但是大数据分析攻击案例表明此种简单步骤不能保护隐私信息。数据隐私保护就是在满足数据隐私安全的基础上做到数据可用性最大[19]。

常用的隐私保护技术包含联邦学习、安全多方计算、同态加密以及差分隐私等。联邦机器学习可使多个个体在自身敏感信息不被泄露的前提下实现强化学习；多方安全计算可在第三方不可信情况下，保证数据隐私的同时得到既定的目标函数；同态加密能确保对敏感数据进行先计算后加密与先加密后计算 2 种操作得到的结果等价；差分隐私保护技术是基于强大的数学理论，可以对隐私保护提供量化评价的方法[20]。

1.2　差分隐私保护定义

对于有限域 X，$x \in X$ 是 X 中的元素，集合 A 是由 x 构成的具有样本个数 n、属性维度 d 的数据集。对数据集 A 采用不同映射方式的过程为查询，$F=\{f_1, f_2, f_3 \cdots\}$ 即为一组查询，利用算法 Q 计算 F，使该结果符合隐私保护，这就是隐私保护机制[21]。

设 A 数据集以及 A' 数据集之间存在同一属性结构，它们的对称差为 $A\Delta A'$，其中 $|A\Delta A'|$ 为 $A\Delta A'$ 的记录数目，如果 $|A\Delta A'|=1$，则 A 与 A' 互为邻近数据集。

设有算法 Q，对任何 2 个互为邻近数据集的集合 A 与 A'，C 为任意一种聚类划分方式，ε 被定义为隐私保护预算，如果算法 Q 满足：

$$\Pr[Q(A) \in C] \leqslant e^{\varepsilon} * \Pr[Q(A') \in C]。 \tag{1}$$

则称算法 Q 提供 ε。

1.3　差分隐私保护相关概念

1.3.1　隐私保护预算

隐私保护预算 ε 是衡量算法 Q 在邻近数据集上获取近乎一致的输出结果之间的可能性概率。它与算法 Q 能够达到的隐私保护强度密切相关，ε 的大小与隐私保护强度成反比。ε 的取值在一定程度上体现了数据的安全性。

1.3.2　敏感度

差分隐私保护中有全局敏感度以及局部敏感度。设 f 是关于数据集 A 和实数向量的一种映射函数，对邻近数据集 A 与 A'，f 的全局敏感度为：

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

全局敏感度与数据集反应的个体差异程度有关。Δf 越大，掩盖数据集中全部对象所需噪声就越大。

设 f 是关于数据集 A 和实数向量的一种映射函数，对邻近数据集 A 与 A'，f 的局部敏感度为：

$$LS_f(D) = \max_{A} \| f(A) - f(A') \|_1。 \tag{3}$$

1.3.3　实现机制

差分隐私主要是通过加入 Laplace 噪音干扰机制来实现的。Laplace 机制主要面向数值型数据，在查询得到的结果中添加拉普拉斯噪声进而做出查询响应。笔者应用 Laplace 机制来进行隐私保护，Laplace 机制的定义为：

数据集 A 在查询函数 f 下，全局敏感度为 Δf，若算法 $Q(A)$ 满足

$$Q(A) = f(A) + \mathrm{Lap}(\Delta f / \varepsilon)。 \tag{4}$$

则称算法 $Q(A)$ 提供差分隐私保护。对于均值为 0，且方差为 b 的 Laplace 分布记为 Lap(b)。Lap(b) 的概率密度函数为：

$$p(x) = \exp\left(-|x|/b\right)/2b。 \tag{5}$$

2　*K*-means 算法

在聚类算法中，最常用的算法是 *K*-means 算法。*K*-means 算法的基本思路：先从原始数据集中自由选择 K 个数据作为最初聚类中心，再将剩下数据按照它们与每个聚类中心的距离大小分配到相应的各个簇类之中，然后重新统计推算所有类的最新聚类中心。其中聚类中心更新公式为：

$$z_g = \sum_{x \in Z_j} x \Big/ \left|Z_j\right|。 \tag{6}$$

式中：$\left|Z_j\right|$ 为数据集中类 Z_j 中的数据个数；x 为类 Z_j 内的数据。

重复上述过程，直至该算法的评价函数达到收敛或者迭代次数到达最高频次。评价函数使用的是类内距离平方和：

$$\mathrm{Dis}=\sum_{i=1}^{K}\sum_{x_g\in Z_j}\|x_g-z_j\|^2。\tag{7}$$

式中：K 为总的聚类个数；x_g 为类 Z_j 内的数据；z_j 为类 Z_j 的类中心向量。

标准 K-means 算法(算法 1)执行步骤如下所述：

输入：数据集 X，初始类簇数据个数 K，最大迭代次数。

输出：K 个簇、评价函数值 Dis。

算法流程：

1) 从已有的数据集 X 中，自由选择 K 个数据作为最初聚类中心。

2) 对数据集 X 剩下的数据按照它们与每个聚类中心的距离大小分配到相应的各个簇类中。

3) 按照聚类更新公式计算最新聚类中心。

4) 重复步骤 2)以及步骤 3)至评价函数达到收敛或者迭代次数到达最高频次。

若能得到相同类簇的数据相似度高且不同类簇的数据相似度低的聚类结果，则该算法聚类效果较好。常见的 K-means 聚类算法存在 2 大问题：

1) 初始自由选择数据作为聚类中心。每次操作所选取的初始聚类中心都是随机的，导致每次中心选取可能与其他此操作选取的中心不同，可能导致每次聚类结果存在不同差异。

2) 初始类簇数量 K 由人工输入。人工确定类簇数量 K 可能导致输出的聚类结果不符合最佳聚类结果。

3 基于隐私保护的改进 K-means 算法

3.1 基于动态分配聚类中心的 K-means 算法

提出基于动态分配聚类中心的 K-means 算法。该算法引入密度度量，从全体数据中选取点密度最大的 N 个数据作为备选中心集合，再从此集合中选取中心点，可以保证选取的中心处于相对密集区域，进而保证簇类的类内相似性较高。同时引入距离度量，当选定新的聚类中心时，通常选取距离当前中心较远的点，这样可以保证不同类聚中心排斥性较高，进而保证簇类的类间相似性较低。算法引入类间平均最大相似度，进而动态确定最佳算法簇类数量 K 并能得到最优初始簇类中心。该算法解决了初始随机选择数据作为聚类中心以及初始类簇数据个数 K 由人工输入问题。

点密度使用的是点 x 在邻域 R 范围内的数据个数：

$$\mathrm{Dis}(x)=\{q\in z\mid \mathrm{dist}(x,q)\leqslant R\}。\tag{8}$$

式中：x 为数据集中的数据点；R 为输入的邻域半径。

类间距离是指不同簇类的聚类中心的相距距离：

$$d_{ij}=\|z_i-z_j\|^2。\tag{9}$$

式中 z_i 为第 i 类的聚类中心。

类内距离是指该类簇中所有点到该类簇聚类中心的距离平方的均值：

$$s_i=\frac{1}{|Z_i|}\sum_{x\in Z_i}\|x-z_i\|。\tag{10}$$

平均类间最大相似度是指每个簇类与剩余其他簇类的最大相似度的平均值：

$$AM=\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}\max\left\{\frac{s_i+s_j}{d_{ij}}\right\}。\tag{11}$$

AM 值越大，意味着类与类的相似度大，此时的聚类效果差；反之，AM 值越小，意味着类与类的相似度小，此时的聚类效果好。当 AM 值不增加时，说明此时类与类的相似度最小，此时的聚类效果亦是最好，当前的 K 值符合最佳聚类结果的要求。

基于动态分配聚类中心的 K-means 算法(算法 2)执行步骤如下所述：

输入：数据集 X，邻域半径 R，高密度点的个数 N，初始 AM 值，最大迭代次数。

输出：K 个簇、评价函数值 Dis。

算法流程：

1) 计算给定的数据集 X 的各个数据的点密度，选取其中 N 个最大点密度的数据，将其放进备选中心点集 B 中。

2) 从备选中心点集 B 选取 2 个最大点密度的数据作为初步初始聚类中心，并将该两点从备选中心点集 B 中移除。

3) 从备选中心点集 B 中选取一个数据使其距离目前已有初始聚类中心最远，将其设置为下一个初始聚类中心点，并将该点从备选中心点集 B 中移除。

4) 将数据集 X 中的所有数据按照以上聚类中心进行迭代并划分到相应簇类之中。

5) 计算此时的 AM 值。

6) 如果当前计算的 AM 值比之前的 AM 值小，

则可以继续进行算法，进行步骤 7)。否则，真正初始聚类中心选定为 *AM* 值在最小时的聚类中心，进行步骤 8)。

7) 按照聚类更新公式计算最新聚类中心，从备选中心点集 *B* 中选取一个数据使其距离最新聚类中心的距离最小值取最大值，将其设置为下一个初始聚类中心点，并将该点从备选中心点集 *B* 中移除。

8) 计算数据集 *X* 剩余数据到各聚类中心的距离，按照它们与各聚类中心的距离大小分配到相应簇类中。

9) 按照聚类更新公式重新计算聚类中心。

10) 重复步骤 8)以及步骤 9)至评价函数达到收敛或者迭代次数到达最高频次。

此算法通过 *K*-1 次动态分配，最终得到了 *K* 个最佳的聚类中心。此聚类中心并不是随意选取的，而是按照数据的具体特征计算得到的，所以它可以更加有效地增强整个聚类过程的抗波动性，较好地增强聚类过程的准确性。

但是 *K*-means 算法以及改进后的算法存在 2 个隐私泄露方面的问题：

1) 在聚类迭代的过程中，计算数据与各聚类中心点的距离时可能会发生隐私泄露。在聚类迭代划分类簇时，攻击者可以不断获取数据与各聚类中心点的距离，通过这些数据可以推断出原数据集的部分属性。隐私泄露大小与迭代次数成正比。

2) *K*-means 算法最终结果显示为聚类中心，攻击者可能根据结果获取数据集的分布以及部分特性，从而出现隐私泄露。

K-means 算法的数据集可能包含用户的个人非公开信息，在进行聚类的过程以及最终聚类结果发布之后都可能出现隐私泄露。

3.2　基于隐私保护的改进 *K*-means 算法

在 *K*-means 算法的执行过程中，聚类中心的确定过程是关键部分。计算每个数据所属类簇会出现隐私泄露问题，输出最终类聚中心结果也可能会出现隐私泄露问题；但是，在计算更新中心点的过程中仅暴露中心点的近似基数，即中心点近似值，而并不是中心点本身，即可做到不出现隐私泄露问题的同时，不对聚类结果产生较大准确性影响。在基于动态分配聚类中心的 *K*-means 算法的基础上引入差分隐私算法。

每次迭代的隐私预算 ε' 与数据维度 *d*、迭代次数 *t*、数据集范围 *r*、初始隐私预算值 ε 有关：

$$\varepsilon' = \varepsilon/(dr+1)t \text{。} \tag{12}$$

引入拉普斯噪声，具体公式为：

$$f=\mathrm{Lap}(b)\text{。} \tag{13}$$

式中 $b=\Delta f/\varepsilon'$，其概率密度函数为：

$$p(x)=\frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)=\frac{\varepsilon'}{2\Delta f}\exp\left(-\frac{\varepsilon'*|x|}{\Delta f}\right)\text{。} \tag{14}$$

基于隐私保护的改进 *K*-means 算法（算法 3）执行步骤如下所述：

输入：数据集 *X*，邻域半径 *R*，高密度点的个数 *N*，初始 *AM* 值，数据维度的 *d*，隐私预算 ε，查询函数敏感度 Δf，最大迭代次数。

输出：*K* 个簇、评价函数值 Dis。

算法流程：

1) 计算给定数据集 *X* 各个数据的点密度，选取其中 *N* 个最大点密度的数据，将其放进备选中心点集 *B* 中。

2) 从备选中心点集 *B* 选取 2 个最大点密度的数据作为初步初始聚类中心，并将该 2 点从备选中心点集 *B* 中移除。

3) 从备选中心点集 *B* 中选取一数据使其距离目前已有初始聚类中心最远，将其设置为下一个初始聚类中心点，并将该点从备选中心点集 *B* 中移除。

4) 将数据集 *X* 中的所有数据按照以上聚类中心进行迭代并划分到相应簇类之中。

5) 计算此时的 *AM* 值。

6) 如果当前计算的 *AM* 值比之前的 *AM* 值小，则可以继续进行算法，进行步骤 7)。否则，如果当前计算的 *AM* 值比之前的 *AM* 值大，则真正初始聚类中心选定为 *AM* 值在最小时的聚类中心，进行步骤 8)。

7) 按照聚类更新公式计算最新聚类中心，从备选中心点集 *B* 中选取一数据使其距离最新聚类中心的距离最小值取最大值，将其设置为下一个初始聚类中心点，并将该点从备选中心点集 *B* 中移除。

8) 计算数据集 *X* 剩余数据到各聚类中心的距离，按照它们与各聚类中心的距离大小分配到相应簇类中。

9) 按照聚类更新公式重新计算聚类中心，计算当前拉普斯噪声，并对当前聚类中心添加拉普斯噪声。

10) 重复步骤 8)以及步骤 9)至评价函数达到收敛或者迭代次数到达最高频次。

在执行 K-means 算法的过程引入了差分隐私保护，当攻击者获取某一信息所在的类簇中心点信息以及其他数据信息时，仍然无法获得目标隐私信息，该算法较好地保护了信息的安全性。

3.3 聚类安全性分析

设存在邻近数据集 A 与 A'，$Q(A)$以及 $Q(A')$表示邻近数据集 A 与 A'利用改进 K-means 算法获得的具有隐私保护的输出结果，C 为聚类划分方式，$\theta(x)$为添加噪声后的聚类划分结果，$r(A, x)$和 $r(A', x)$为对数据集 A 和 A'的真实聚类划分结果，安全性证明如下：

1) 由式(1)可得算法 Q 提供隐私保护的方法是对输出结果进行随机化过程，所以：

$$\Pr[Q(A)\in C]=\Pr[\mathrm{Lap}(b)=\theta(x)-r(A,x)]。\tag{15}$$

式中：$b=\Delta f/\varepsilon'$；$\varepsilon'=\varepsilon/(dr+1)t$。

2) 代入 Lap(b)的概率密度函数可得：

$$\Pr[\mathrm{Lap}(b)=\theta(x)-r(A,x)]=\exp\left(-|\theta(x)-r(A,x)|/b\right)/2b。\tag{16}$$

3) 结合步骤 1)以及步骤 2)可得：

$$\Pr[Q(A)\in C]=\exp\left(-\varepsilon'|\theta(x)-r(A,x)|/\Delta f\right)/2b。\tag{17}$$

同理可得：

$$\Pr[Q(A')\in C]=\exp\left(-\varepsilon'|\theta(x)-r(A',x)|/\Delta f\right)/2b。\tag{18}$$

4) 由全局敏感度的定义可得：

$$\|r(A,x)-r(A',x)\|_1\leqslant\Delta f。\tag{19}$$

5) 由 $\varepsilon'=\varepsilon/(dr+1)t$ 并结合步骤 3)可得：

$$\begin{aligned}&\Pr[Q(A)\in C]/\Pr[Q(A')\in C]=\\&\exp\left(-\varepsilon'(|\theta(x)-r(A',x)|-|\theta(x)-r(A,x)|)/\Delta f\right)\leqslant\\&\exp\left(-\varepsilon(|\theta(x)-r(A',x)|-|\theta(x)-r(A,x)|)/\Delta f\right)\leqslant\\&\exp\left(-\varepsilon(|r(A,x)-r(A',x)|)/\Delta f\right)=\\&\exp\left(-\varepsilon(\|r(A,x)-r(A',x)\|_1)/\Delta f\right)=\exp(\varepsilon)。\end{aligned}\tag{20}$$

6) 最终可得：

$$\Pr[Q(A)\in C]/\Pr[Q(A')\in C]\leqslant\exp(\varepsilon)。\tag{21}$$

因此，可得基于隐私保护的改进 K-means 算法满足 ε-差分隐私。

4 实验验证

4.1 实验设计

4.1.1 实验环境

实验软件环境是 Win10 操作系统下的 Anaconda python3.7，硬件环境是 Intel(R), Core(TM), i7-7700HQU CPU@1.70 GHz。

4.1.2 实验数据

实验所用数据集来自 UCI 数据库中的 Iris 数据集、Raisin_Dataset 数据集、Wine 数据集以及文献[22]的 S-sets 数据集，其中 S-sets 数据集由 4 个子数据集 S_1～S_4 组成，各数据集具体数据如表 1 所示。

表 1 数据信息

数据集	样本数	数据维度
Iris	150	4
Raisin_Dataset	900	7
Wine	178	13
S1—S4	5 000	2

由于数据集数据较大，笔者对以上数据集数值分别进行归一化处理，将数据集的数值都归一到[0, 1]之间。

4.1.3 评价指标

4.1.3.1 聚类性能分析

笔者采用类内距离平方和(Dis)用来评估算法聚类的性能。具体计算可见式(7)。评价函数可以评判聚类结果的优劣，评价函数值与聚类结果成反比。

4.1.3.2 聚类可用性分析

笔者采用 F 分数对算法进行聚类可用性分析。F 分数是一种度量聚类可用性的衡量方法，F 分数与信息检索的准确率以及召回率有关。设 x 为数据集 X 的大小，a 为该数据集 X 中正确分类的标签，x_a 为类 a 中数据点个数，x_b 为簇 Z_b 数据点个数，x_{ab} 为类 a 以及簇 Z_b 的数据点交集部分的数量，准确率(Prec)以及召回率(Reca)的计算公式为：

$$\mathrm{Prec}(a,b)=\max_{a,b}\{x_{ab}/x_b\}；\tag{22}$$

$$\mathrm{Reca}(a,b)=\max_{a,b}\{x_{ab}/x_a\}。\tag{23}$$

另外对于类 a 以及簇 Z_b 的 F 分数计算公式如下：

$$F(a,b)=\frac{(\alpha^2+1)\mathrm{Prec}(a,b)\bullet\mathrm{Reca}(a,b)}{\alpha^2\mathrm{Prec}(a,b)+\mathrm{Reca}(a,b)}。\tag{24}$$

令 Prec(a, b)和 Reca(a, b)具有相同权重，即令 α=1，所以对于数据集 X，总 F 分数值的计算公式如下：

$$F=\sum_a\frac{x_a}{x}\max_j\{F(a,b)\}。\tag{25}$$

对于数据集 X，总 F 分数值越大意味着聚类可用性越好。

4.2　实验结果

4.2.1　聚类性能结果

为评价基于隐私保护的改进 *K*-means 算法的性能，本实验用标准 *K*-means 算法、基于动态分配聚类中心的 *K*-means 算法以及基于隐私保护的改进 *K*-means 算法分别对以上数据集进行实验。为避免单次实验出现误差，对上述 3 种算法取 50 次实验结果的平均值视为最终结果。表 2 为这 3 种算法在各数据集下的评价函数值，即类内距离平方和值。

表 2　3 种算法评价函数值

数据集	标准 *K*-means 算法	基于动态分配聚类中心的 *K*-means 算法	基于隐私保护的改进 *K*-means 算法
Iris	0.574 615	0.429 593	0.425 721
Raisin_Dataset	0.241 687	0.237 440	0.204 162
Wine	0.216 091	0.215 672	0.206 860
S_1	92.069 043	73.767 559	71.239 380
S_2	97.736 398	95.647 008	94.821 958
S_3	62.980 461	62.140 359	52.252 677
S_4	45.296 839	41.372 232	41.348 386

评价函数被用来判断聚类结果的好坏，相同数据集的评价函数值越小，代表该算法下类中数据对象越集中，该算法的聚类效果越好。由上表可以看出，在相同数据集下，基于隐私保护的改进 *K*-means 算法的评价函数值低于其他 2 种算法，故该算法一定情况下优化了聚类性能。

4.2.2　聚类可用性结果

为评价基于隐私保护的改进 *K*-means 算法的可用性，本实验用标准 *K*-means 算法、基于动态分配聚类中心的 *K*-means 算法以及基于隐私保护的改进 *K*-means 算法分别对以上数据集进行 *F* 分数值计算。为避免单次实验出现误差，对上述 3 种算法取 50 次实验结果的平均值作为最终结果。表 3 为这 3 种算法在各数据集下的 *F* 分数值。

表 3　3 种算法 *F* 分数值

数据集	标准 *K*-means 算法	基于动态分配聚类中心的 *K*-means 算法	基于隐私保护的改进 *K*-means 算法
Iris	0.818 491	0.823 502	0.824 265
Raisin_Dataset	0.735 281	0.737 003	0.749 616
Wine	0.656 955	0.663 610	0.669 257
S1	0.542 755	0.564 290	0.545 370
S2	0.476 839	0.477 783	0.473 717
S3	0.519 322	0.556 601	0.532 997
S4	0.491 677	0.503 664	0.499 563

总 *F* 分数值越大意味着聚类可用性越好。由表中可以看出：基于动态分配聚类中心的 *K*-means 算法以及基于隐私保护的改进 *K*-means 算法比标准 *K*-means 算法的聚类可用性强，另外基于隐私保护的改进 *K*-means 算法聚类可用性略低于基于动态分配聚类中心的 *K*-means 算法，这是由于在基于动态分配聚类中心的 *K*-means 算法基础上添加噪声，添加噪声也必然会导致聚类可用性略微降低。另外，通过观察可得基于隐私保护的改进 *K*-means 算法聚类可用性与基于动态分配聚类中心的 *K*-means 算法的可用性相差较小，由此也可看出基于隐私保护的改进 *K*-means 算法的聚类可用性较好。

4.2.3　聚类结果分析

由于 *S*-sets 数据集为 2 维数据集可以进行可视化操作，故基于 S_1 子数据集可视化比较标准 *K*-means 算法、基于动态分配聚类中心的 *K*-means 算法以及基于隐私保护的改进 *K*-means 算法 3 种算法的聚类效果，如图 1—3 所示。

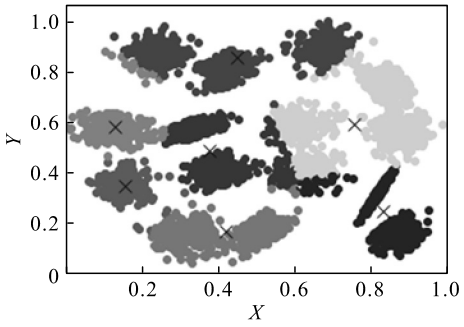

图 1　标准 *K*-means 算法聚类效果

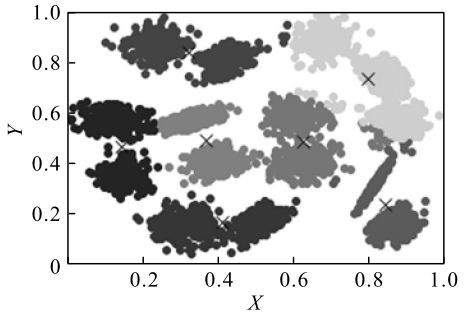

图 2　基于动态分配聚类中心的 *K*-means 算法聚类效果

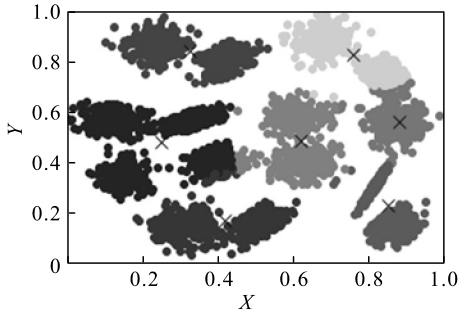

图 3　基于隐私保护的改进 *K*-means 算法聚类效果

图 1 描述的是标准 *K*-means 算法的聚类效果，图 2 和 3 分别对应着基于动态分配聚类中心的 *K*-means 算法的聚类效果以及基于隐私保护的改进

K–means 算法的聚类效果。图中横坐标表示 2 维坐标轴的 *X* 轴坐标，纵坐标表示 2 维坐标轴的 *Y* 轴坐标；图中采用不同明度的点表示不同类簇中的点的位置，中间十字表示相应类簇的中心点位置。综合以上实验结果，可直观地看出笔者提出的优化算法在 S_1 数据集上的聚类结果优于标准 *K*–means 算法的聚类结果。

上面的实验结果采用了 S_1 子数据集，为验证所提出算法的普遍适用性，笔者又选取了 S_2、S_3、S_4 子数据集进行验证。实验结果均可直观地看出笔者提出的优化算法在该数据集上的聚类效果好于标准 *K*–means 算法的聚类效果。

基于子数据集 S_2、S_3、S_4 可视化比较标准 *K*–means 算法、基于动态分配聚类中心的 *K*–means 算法以及基于隐私保护的改进 *K*–means 算法 3 种算法的聚类效果，如图 4—12 所示。

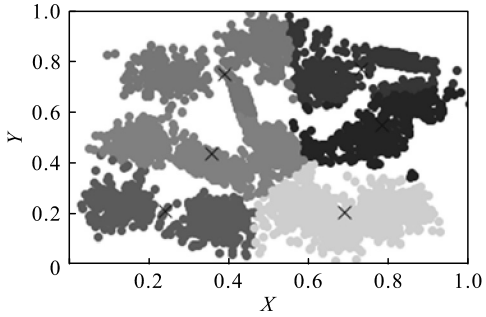

图 4 数据集 S_2 下标准 *K*–means 算法聚类效果

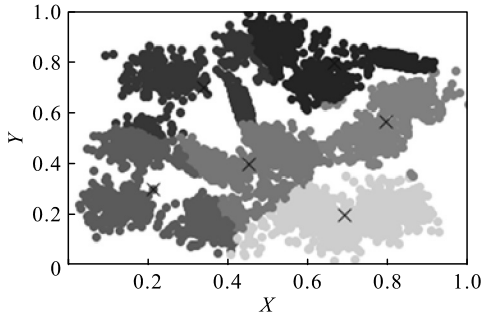

图 5 数据集 S_2 下基于动态分配聚类中心的 *K*–means 算法聚类效果

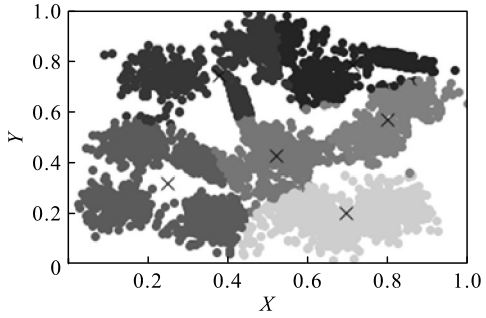

图 6 数据集 S_2 下基于隐私保护的改进 *K*–means 算法聚类效果

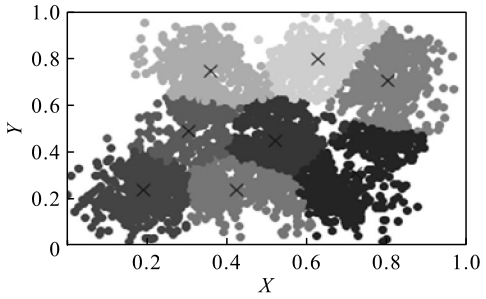

图 7 数据集 S_3 下标准 *K*–means 算法聚类效果

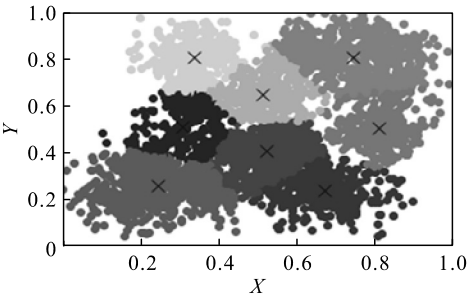

图 8 数据集 S_3 下基于动态分配聚类中心的 *K*–means 算法聚类效果

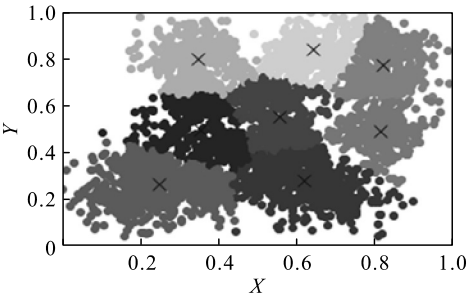

图 9 数据集 S_3 下基于隐私保护的改进 *K*–means 算法聚类效果

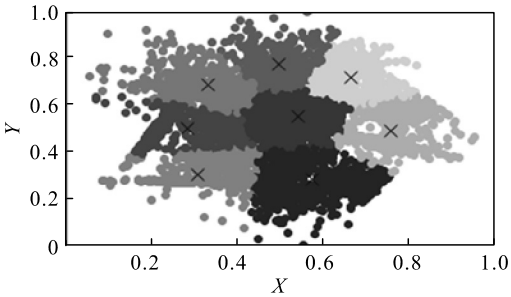

图 10 数据集 S_4 下标准 *K*–means 算法聚类效果

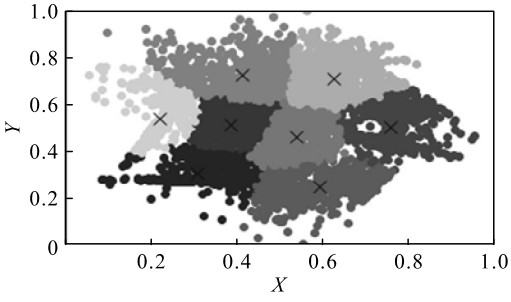

图 11 数据集 S_4 下基于动态分配聚类中心的 *K*–means 算法聚类效果

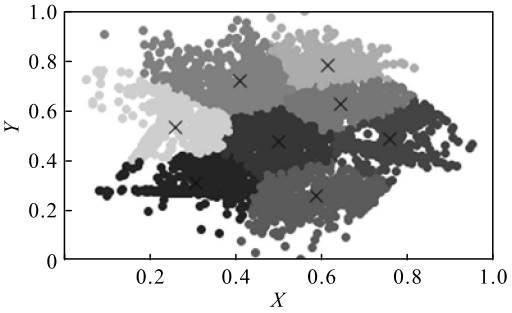

图 12　数据集 S_4 基于隐私保护的改进 *K*-means 算法聚类效果

5　结论

笔者提出基于隐私保护的改进 *K*-means 算法，在原有 *K*-means 基础上引入密度度量、距离度量以及类间平均最大相似度，可动态确定最佳聚类个数 *K* 以及最佳初始聚类中心。同时，引入了差分隐私保护拉普拉斯噪声，较好地保护了信息的隐私性。通过实验验证，该算法具有较好的聚类性能以及数据可靠性，但在动态规划的过程中需计算所有数据的点密度，当数据量过于庞大时，会存在时间复杂度较高的问题。下一步，笔者将对降低算法的时间复杂度进行研究。

参考文献：

[1] 陈强. 非结构化智能金融投研平台的开发与行业应用[J]. 计算机系统应用, 2022, 31(2): 78−87.

[2] YAN N N, ZHANG Y P, XU X, et al. Online finance with dual channels and bidirectional free-riding effect[J]. International Journal of Production Economics, 2021, 231: 107834.

[3] 张建根. 支撑财政转移支付的区块链技术方案研究[J]. 计算机工程与应用, 2021, 57(19): 150−155.

[4] 张晗, 宋志华, 王彤, 等. 基于军事背景的大学计算机课程案例教学探索与实践[J]. 计算机工程与科学, 2019, 41(S1): 210−212.

[5] 程小刚, 郭韧, 周长利. 基于理性密码学的分布式隐私保护数据挖掘框架[J/OL]. 计算机工程与科学: 1−8[2022−04−06]. http: //kns.cnki.net/kcms/detail/43.1258.TP.20220401.1645.002.html.

[6] 王腾, 霍峥, 黄亚鑫, 等. 联邦学习中的隐私保护技术研究综述[J/OL]. 计算机应用: 1−15[2022−05−04]. http://kns.cnki.net/kcms/detail/51.1307.TP.20220425.1937.008.html.

[7] LOUISE T, IQBAL G, TAIWO O, et al. A framework for data privacy and security accountability in data breach communications[J]. Computers & Security, 2022, 116: 102657.

[8] DWORKC. Differentialprivacy[M]. Berlin Heidelberg: Springer-Verla, 2006: 1−12.

[9] 贺军义, 吴梦翔, 宋成, 等. 基于 UWB 的密集行人三维协同定位算法[J]. 计算机应用研究, 2022, 39(3): 790−796.

[10] 刘向举, 路小宝, 方贤进, 等. 软件定义网络环境下的低速率拒绝服务攻击检测方法[J/OL]. 计算机应用: 1−8[2022−05−04]. http: //kns.cnki.net/kcms/detail/51.1307.TP.20211014.1706.023.html.

[11] 张峻豪, 吴飞, 朱海. 基于 GD−Kmeans 和菲涅尔理论的 WiFi 手势识别方法[J]. 计算机工程与应用, 2020, 56(19): 126−131.

[12] 凌玉龙, 张晓, 李霞, 等. 改进 kmeans 算法在学生消费画像中的应用[J]. 计算机技术与发展, 2021, 31(10): 122−127.

[13] SHI Y G, YAN S Y, HE M B, et al. Hybrid Data Mining Method of Telecom Customer Based on Improved Kmeans and XGBoost[C]. Temple Circus: IOP Publishing Ltd, 2021.

[14] 林涛, 赵璨. 最近邻优化的 *k*-means 聚类算法[J]. 计算机科学, 2019, 46(S2): 216−219.

[15] 汪丽娟. Flink 下的 *K*−Means 优化并行与任务调度研究[D]. 乌鲁木齐: 新疆大学, 2019.

[16] WANG T X, GAO J Y. An Improved K-Means Algorithm Based on Kurtosis Test[C]. Xi'an: Proceedings of 2019 3rd International Conference on Artificial Intelligence, Automation and Control Technologies (AIACT 2019), 2019.

[17] 张亚迪, 孙悦, 刘锋, 等. 结合密度参数与中心替换的改进 *K*−means 算法及新聚类有效性指标研究[J]. 计算机科学, 2022, 49(1): 121−132.

[18] 周浩天. 混合数据聚类个数与初始类中心确定算法及实现[D]. 太原: 山西大学, 2020.

[19] 陈儒玉, 戴欢, 高玉建, 等. 基于区块链的电子学位证照数据保护共享方法[J]. 计算机工程, 2022, 48(4): 50−60, 80.

[20] 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.

[21] 胡闯. 面向差分隐私保护的聚类算法研究[D]. 南京: 南京邮电大学, 2019.

[22] FRÄNTI P S S. *K*-means properties on six clustering benchmark datasets[J]. ApplIntell, 2018, 48: 4743−4759.