文章编号：2095-2163(2023)02-0155-07 中图分类号：TP309.7 文献标志码：A

基于差分隐私保护的二分 k 均值聚类算法研究

马文博，巫朝霞
（新疆财经大学 统计与数据科学学院，乌鲁木齐 830011）

摘 要：针对差分隐私保护 k 均值聚类算法(DP k-means)随机选取初始点，导致算法往往收敛于局部最优，进而影响聚类效果的问题，本文结合差分隐私的相关理论以及层次聚类的思想提出了一种基于差分隐私保护的二分 k 均值聚类算法(DP Bi-k-means)。首先，以得到全局最优为目标，将随机选取初始点的过程进行改进，由上至下对目标数据集进行二分；其次，在迭代过程实现基于拉普拉斯机制的差分隐私保护。经安全性分析以及实验结果证明：该算法与传统差分隐私保护 k 均值算法(DP k-means)相比，可以避免聚类结果受初始点的影响陷入局部最优解，从而优化聚类效果，并为聚类分析提供了有效的隐私保护能力。

关键词：差分隐私；二分 k 均值聚类算法；拉普拉斯机制

Research on bisecting k-means clustering algorithm based on differential privacy protection

MA Wenbo, WU Zhaoxia
(School of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumqi 830011, China)

【Abstract】 For the traditional differential privacy protection k-means clustering algorithm (DP k-means), the randomly selected initial point can result in that the algorithm often convergence to the local optimal instead of the global optimal. The problem in turn affects the effectiveness of clustering. This paper combines the theory of differential privacy with the idea of hierarchical clustering and proposes the bisecting k-means algorithm based on differential privacy protection (DP Bi-k-means). Firstly, the process of random selection of the initial point is improved to obtain the global optimality, and the target dataset is bisected from top to bottom. Then, differential privacy protection based on the Laplace mechanism is implemented in the iterative process. Finally, through security analysis and experimental results, it can be proved that compared with the traditional differential privacy-preserving k-means algorithm (DP k-means), the proposed algorithm can avoid the clustering results from falling into the local optimal solution affected by the initial point. This optimizes clustering results and provides effective privacy protection capabilities for cluster analysis.

【Key words】 differential privacy; bisecting k-means; Laplace mechanism

0 引 言

当下数据科学的发展一日千里，所产生的数据也呈爆炸式增长，结合传统统计学的理论基础与现代计算机科学的计算优势所产生的衍生学科在处理复杂问题的能力上也有了质的飞跃。机器学习、数据挖掘、人工智能等新的理论和技术也应运而生，为大数据的信息挖掘与应用带来了新的路径和视野，新理论、新技术的应用极大提升了社会服务的效率、增强了商业运营的能力、促进了相关科学研究的发展，但对大数据的滥用及相关的隐私泄漏问题也日渐显现。用户举手投足间产生的海量信息中往往包含着大量的敏感信息，如：金融信息、医疗信息、行为特点信息等，而对这些敏感信息的滥用将会切实危害到信息提供者的隐私安全，不利于大数据行业的健康发展，大数据的利用与隐私保护已经成为了一对尖锐的问题，保护数据隐私的情况下高效的利用数据是当下研究的重要关注点[1-2]。

隐私保护的发展也决定了数据科学的应用范围与发展方向。目前隐私保护方法可以分为密码学方法、信息隐藏方法以及数据处理方法。密码学方法主要是研究数据的加解密方案，以及不同的密钥分

作者简介：马文博(1995-)，男，硕士研究生，主要研究方向：大数据隐私安全；巫朝霞(1975-)，女，博士，教授，硕士生导师，主要研究方向：信息安全研究。
通讯作者：巫朝霞 Email：wuzhaoxia828@163.com

收稿日期：2022-12-05

配管理机制,包括同态加密、对称加密等[3-4];信息隐藏方法则通过对原始数据的形态变换,将隐私信息隐写与公开信息再进行传输进而保护原始信息,如数字水印等[5];数据处理方法是减少隐私数据之间存在的关联性,通过添加数据扰动、数据匿名化等方法实现隐私保护。防止复杂网络分析、深度学习等大数据计算的过程中泄露敏感信息。Dwork 在数据扰动的思想下于 2006 年提出了差分隐私保护技术,是可由严密数学逻辑证明的数据扰动隐私保护技术[6]。通过对所要处理的数据添加符合隐私预算 ε 的噪声,从而实现在假定攻击者拥有最大知识背景的情况下依然能对数据进行有效保护,使数据在保证其可用性的同时不会泄露敏感信息。

在大数据处理分析中,聚类分析是数据挖掘的核心问题之一[7]。k-means 作为基于划分聚类的经典算法,有着原理直观、可解释性强、算法复杂度低、收敛速度快、对于处理大数据具有良好的伸缩性等诸多优势。但传统 k-means 算法对初始点的选取非常敏感,聚类性能易受初始节点的影响,在算法迭代过程中存在隐私泄露的安全隐患[8]。傅彦铭等[9]人提出了一种基于拉普拉斯机制的差分隐私保护 k-means++算法,保证了在不同安全级别下算法的可用性;李洪成[10]等人针对传统隐私保护方法无法应对任意背景知识下恶意分析的问题,提出了分布式环境下满足差分隐私的 k-means 算法;马银方[11]等人提出了基于差分隐私保护的 KDCK-medoids 动态聚类算法,解决 k-medoids 算法不能对动态数据进行聚类的问题;Dwork[12]等人提出了差分隐私保护 k-menas 算法中隐私预算的分配方法。

本文针对 k-means 算法可能存在的隐私泄露以及易受初始点选取影响的缺陷,提出了一种基于拉普拉斯噪声机制的差分隐私保护二分 k-means 算法(DP Bi-k-means),在聚类的过程中实现了差分隐私保护机制。实验结果表明,该算法与传统差分隐私保护 k-means 算法相比,可以避免聚类结果受初始点的影响陷入局部最优解,从而优化聚类效果,并为聚类分析提供了严格的隐私保护。

1 差分隐私相关理论

1.1 差分隐私定义

差分隐私是基于不同机制对不同类型的数据进行失真的隐私保护方法。通过对原始数据、算法参数、以及输出结果等关键信息添加服从特定分布的噪音,使在全体数据集中任意添加或删除一条数据并不显著改变该数据集的信息熵,使攻击者在拥有最大背景知识时也无法准确判断某条数据是否在该数据集中,从而保证了所有隐私信息的安全[6]。满足差分隐私的数据集能够抵抗对隐私数据的分析,具有信息论意义上的安全性。

定义 1 ε-差分隐私

假设 D 和 D' 是两个具有相同数据结构且差别为一条数据记录的相邻数据集,函数 M 为随机函数,其取值范围表示为 $Range(M)$,存在集合 S,且 $S \in Range(M)$,E 为查询函数,$Pr(E)$ 表示隐私泄露的风险。

若随机算法 M 对数据集 D 和数据集 D' 进行计算,得到的输出结果 $M(D)$ 和 $M(D')$ 使 $M(D) \in range(M)$,$M(D') \in range(M)$ 成立,且满足式(1):

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

则称算法 M 满足隐私参数为 ε 的差分隐私[6]。在 ε 确定的情况下,差分隐私的严格数学定义保证了算法 M 在计算时其输出结果的概率分布是随机相似的。即使遭到最大背景知识的差分攻击,也能对数据集中任意一条的数据进行保护。由公式(1)可知隐私参数 ε 的值越小,$M(D)$ 与 $M(D')$ 的概率分布越相似,随机算法 M 的隐私保护能力越强。

风险泄露曲线如图 1 所示,可知在位置参数相同的情况下两曲线的差为 $|e^{\varepsilon}|$。即满足任意随机算法在相邻数据集上输出同一个结果的概率的比值 $Z \in [e^{-\varepsilon}, e^{\varepsilon}]$。

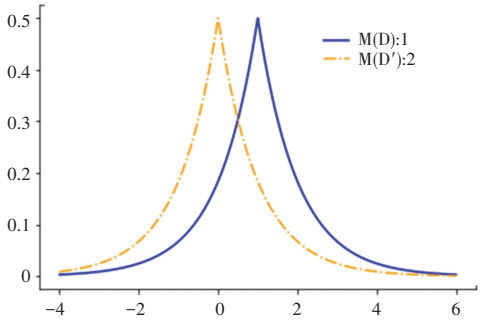

图 1 隐私泄露风险曲线

Fig. 1 Risk curve of privacy leakage

定义 2 全局敏感度

敏感度是差分隐私保护中的一个重要参数,全局敏感度是指对数据集中任意一个数据进行修改时对查询结果造成的最大影响。全局敏感度与查询函数性质相关,与数据集无关。

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，对于任意只相差一条数据的邻近数据集 D 和 D'，函数 f 的全局敏感度，式(2)：

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

对于全局灵敏度较小的函数，只需要添加少量的噪声，就可以使在更改一条数据时对查询结果的影响具有不可分辨性。然而，当全局灵敏度较大时，需要向输出添加大量的噪声，以满足 ε 差分隐私。为了避免因添加过量噪音导致数据可用性变差，针对不同的问题常常引入不同的噪声机制。

定义 3　拉普拉斯机制

差分隐私技术通常通过拉普拉斯机制(Laplace Mechanism)实现对数值型数据的隐私保护。在此机制下可以对原数据或随机函数的查询结果添加服从拉普拉斯分布的随机噪声，来保证满足 ε 的差分隐私。当噪声函数的概率密度为公式(3)时，则此噪声函数服从拉普拉斯分布记为 $x \backsim Lap(\Delta f/\varepsilon)$

$$\int(x \mid u,b) = \frac{1}{2b}\exp\left(-\frac{|x-u|}{b}\right) \tag{3}$$

其中，位置参数为 u，尺度参数为 $b(b > 0)$，式(4)：

$$b = \frac{\Delta f}{\varepsilon} \tag{4}$$

给定数据集 D 时，设有查询函数 F，全局敏感度为 Δf，根据差分隐私定义可得经 ε 的差分隐私保护的查询函数 M，式(5)：

$$M(D) = F(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right) \tag{5}$$

由公式(5)可知在敏感度已知的情况下，在 ε 的差分隐私保护的查询函数 M 中，隐私参数 ε 越小，添加的噪声越多，隐私保护程度越高。

1.2　差分隐私的特性

隐私保护问题往往是复杂的系统工程。对于多层次的数据结构，多样的查询需求，需要多次使用差分隐私保护算法才能保证数据的隐私安全。为了平衡算法的可用性与算法隐私保护能力，需要借助差分隐私保护算法的两个组合性质合理的设置隐私参数 ε。

性质 1　序列组合性

如果给定一个组合函数 $S(s_1(c), s_2(c), s_3(c)\cdots s_n(c))$，对于给定的数据集 C 进行差分隐私保护，每一个独立的函数 $s_1, s_2, s_3\cdots s_n$ 都满足差分隐私，且差分隐私预算分别为 $\varepsilon_1, \varepsilon_2, \varepsilon_3\cdots \varepsilon_n$ 则对于整体的组合函数都满足 ε 差分隐私，式(6)：

$$\varepsilon = \sum_{i=1}^{n} \varepsilon_i \tag{6}$$

性质 2　并行组合性

如果给定一个组合函数 $S(s_1(C_1), s_2(C_2), s_3(C_3)\cdots s_n(C_n))$，对给定的不相交的数据集 $C_1, C_2, C_3\cdots C_n$ 进行差分隐私保护，每一个独立的函数 $s_1, s_2, s_3\cdots s_n$ 都满足差分隐私，且差分隐私预算分别为 $\varepsilon_1, \varepsilon_2, \varepsilon_3\cdots \varepsilon_n$，则对于整体的组合函数都满足 ε 差分隐私，式(7)：

$$\varepsilon = \max_{1\leqslant i\leqslant n} \varepsilon_i \tag{7}$$

2　基于拉普拉斯机制保护的二分 k-均值聚类算法

二分 k 均值聚类算法(Bisecting k-means)是传统 k-means 聚类算法结合层次聚类算法思想中分裂策略的改进算法。首先使所有对象初始化为一个簇，自上而下递归地进行分裂，将原始簇一分为二；再选择其中一个新的簇进行以上操作。因为聚类的误差平方和(SSE)能够衡量聚类性能，该值越小表示数据点越接近于其质心，聚类效果就越好；选择哪一个簇进行划分取决于对其划分是否可以最大程度降低 SSE 的值；重复该过程直到分出了指定的 k 个簇。相对于 k-means 聚类算法，该算法最后优化时采用的质心是多次二分产生的，避免了因随机产生质心而得到局部最优化结果。

2.1　k-means 聚类算法收敛性证明

给定观测点集 $D = \{x_1, x_2, \cdots, x_n\}$，k-means 算法针对聚类所得到的簇划分 $C = \{C_1, C_2, \cdots, C_n\}$ 最小化平方误差，式(8)：

$$E = \sum_{i=1}^{k} \sum_{x\in c_i} \| x - \mu_i \|_2^2 \tag{8}$$

假设簇数 k 已确定，使用误差平方和(SSE)作为目标函数，可以变形获得畸变函数，式(9)：

$$J(\mu, c) = \sum_{i=1}^{n} \| x^{(i)} - \mu_c^{(i)} \|^2 \tag{9}$$

其中，$x^{(i)}$ 为观测集中第 i 个观测，c 是观测对象属于的类。

基于最短距离判断可知，$c = \text{argmin} \| x - \mu_j \|^2$，$\mu_c$ 为聚类簇的中心点，对畸变函数求偏导 $\partial \mu_j (j = 1, 2, \cdots, k)$，式(10)：

$$\frac{\partial J(\mu, c)}{\partial \mu_j} = -2\sum_{i=1}^{k} (x^{(i)} - \mu_c^{(i)}) \frac{\partial \mu_c^i}{\partial \mu_j} = -2\sum_{i=1}^{k} (x^{(i)} - \mu_j) 1\{c^{(i)} = j\} \tag{10}$$

令畸变函数偏导为 0，可以求得极值点，式(11)：

$$\mu_j = \frac{\sum_{i=1}^{n} 1\{c^{(i)} = j\}\, x^{(i)}}{\sum_{i=1}^{n} 1\{c^{(i)} = j\}} \tag{11}$$

证得收敛。

从 k-means 的算法可以发现,误差平方和函数(SSE)是一个严格的坐标下降过程。每一次朝一个变量 C_i 寻找最优解,式(12):

$$C_i = \frac{1}{m}\sum x \tag{12}$$

其中, m 是C_i 所在簇的元素的个数。

即当前聚类的均值就是当前方向的最优解,这与k-means 的每一次迭代过程一样,保证误差平方和函数(SSE)每一次迭代时,都会减小,最终收敛。

由于误差平方和函数(SSE)是一个非凸函数,所以不能保证收敛于全局最优解。

2.2 二分 k 均值算法隐私泄露风险

二分 k-means 聚类算法是基于距离的聚类方法,其更新迭代聚类中心点的过程与传统 k-means 相同,其中迭代参数 sum 为分类过程中簇内的点到簇中心点的距离之和,num 为分类过程中簇内点的个数, C_k 为聚类过程中的簇中心,式(13):

$$C_k = \frac{\text{sum}}{\text{num}} \tag{13}$$

由公式(13)可知,二分 k 均值算法在基于背景知识对聚类中心点的攻击下会泄露整体数据集从而发生隐私泄露[8]。

2.3 差分隐私保护的二分 k 均值(DP Bi-k-means)算法设计

针对二分 k 均值聚类算法在聚类过程中存在数据泄露的问题,ε-差分隐私保护二分 k 均值聚类算法通过在算法迭代中心点添加噪音,达到对中心点数据的保护,算法的步骤如下:

步骤 1 预设聚类个数 k;

步骤 2 将整体数据划分为一个簇,计算质心及总误差平方和;

步骤 3 使用 k-means 算法将这个簇二分为两个簇;

步骤 4 计算该簇一分为二之后的总误差平方和;

步骤 5 选择满足条件的可以分解的簇进行划分;

步骤 6 重复步骤 3-步骤 5 操作,直到达到指定的簇数 k 为止。

迭代过程中分别计算各簇迭代参数 sum 与 num 并分别添加拉普拉斯噪音,从而得到新的聚类中心,式(14):

$$C'_k = \frac{\text{sum} + lap(\frac{\Delta f}{\varepsilon})}{\text{num} + lap(\frac{\Delta f}{\varepsilon})} \tag{14}$$

通常在步骤 5 时会选择可以使总误差平方和最小的簇进行划分,本文选用划分后簇内误差平方和最大的簇进行划分。实验证明:在保证其聚类效果优于传统 k-means 算法的情况下,相较于通常的二分 k-means 算法提高了算法运行效率。

2.4 差分隐私保护的二分 k 均值(DP Bi-k-means)算法安全性证明

DP Bi-k-means 算法通过向迭代过程添加服从拉普拉斯的噪音实现对算法敏感参数的保护,进而保护整体数据集,其中隐私保护强度由全局敏感度 Δf、隐私参数 ε 决定。由定义 2 可知:在算法迭代中心点时,敏感度参数 $\Delta f = 1$。同一簇内计算距离之和的敏感度 $\Delta f \leqslant M$,M 为数据集的特征个数,所以 DP Bi-k-means 算法的全局敏感度为 $M + 1$。

假设随机函数整体的隐私预算为 ε, 则选取更新中心点所消耗的隐私预算为 $\frac{\varepsilon'}{k}$,迭代次数为 k。

根据定义 1,对随机算法 M 更新的迭代中心点进行计算可得式(15);

$$Pr[M(D) \in S] \leqslant e^{\varepsilon} \times Pr[M(D') \in S]$$
$$\frac{Pr[M(C'_1) \in S]}{Pr[M(C'_2) \in S]} \leqslant e^{\varepsilon} \tag{15}$$

根据性质 1,差分隐私的序列组可得式(16):

$$\varepsilon = \sum \frac{\varepsilon'}{k} \tag{16}$$

即 DP Bi-k-means 算法满足 ε -差分隐私。

3 实验结果及分析

通过 3 个方面对差分隐私保护的二分 k 均值算法(DP Bi-k-means)进行实验评价:

(1)DP Bi-k-means 与 DP k-means 在隐私参数动态变化的情况下对比聚类效果,分析新算法的算法特征;

(2)分析在相同隐私参数条件下新算法相比传统算法的抗局部最优解的能力;

(3)通过使用不同的分类策略提高新算法的运

行效率。

实验平台为 windows10 操作系统，cpu3.2 GHZ，8 GB内存，采用 python 语言及相关库进行模拟实验。数据集来自 UCI 机器学习实验室的 IRIS 数据集和 Wine 数据集，见表 1。

表 1　数据集信息

Tab. 1　Information of the data set

数据集	样本数	属性数	预分类数
IRIS	150	4	3
Wine	178	13	3

3.1　实验评价指标

为了对 DP Bi-k-means 算法的聚类效果进行综合评价，本文以 DP k-means 作为对比算法，分别从内部评估法中选取轮廓系数（Silhouette Coefficient）、DB 指数（Davies-Bouldin score）、CH 指数（Calinski-Harabasz Index）、外部评估方法中选取调兰德指数（Adjusted Rand Index）来对两种差分隐私聚类算法进行多方面比较。

通过对 DP Bi-k-means 与 DP k-means 的误差平方和（SSE）进行比较，来分析新算法的聚类性能与鲁棒性，并对算法改进后选用划分后簇内误差平方和最大的簇进行划分的运行效率进行实验分析。

3.2　实验结果

3.2.1　聚类效果分析

实验中通过控制隐私预算参数 ε，进而在不同的隐私保护水平下对算法进行聚类效果的比较实验。

实验在指定分类个数 $k=3$ 的预设下，在隐私参数 ε 逐步增加的情况下对试验指标进行测度。由于在 k 均值聚类算法中超参数只有聚类个数 k，所以实验结果完全受隐私预算参数 ε、算法本身以及随机误差决定。通过直接度量聚类效果的角度进行实验分析，统计算法的内部指标轮廓系数、DB 指数和 CH 指数。经实验得到轮廓系数图如图 2 所示，可知在 ε 变动的情况下，相比 DP k-means 算法本文提出的 DP Bi-kmeans 算法轮廓系数整体更接近于 1，分类目标与所划分类别更加匹配，且波动程度较小聚类效果更稳定；在相同实验条件下得到两种聚类算法的 CH 系数图如图 3 所示，可知 DP Bi-k-means 算法的 CH 系数更大，各分类内部越紧密，各分类之间越疏松，聚类能力较 DP k-means 算法有一定优势；实验分析 DB 指标得到 DB 指标图如图 4 所示，DP Bi-k-means 算法的 DB 指数相较于与 DP k-means 算法的 DB 指数更小，各分类之间距离更大，分类结果更好。

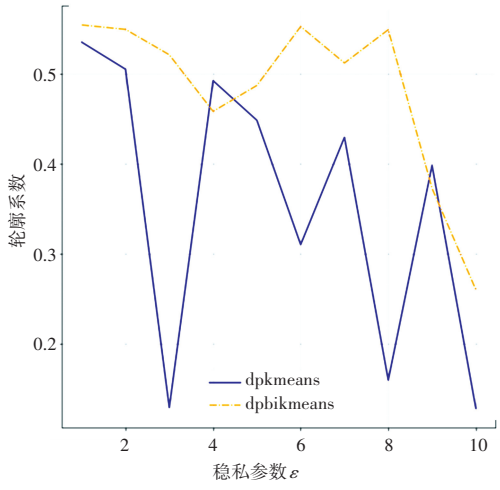

图 2　轮廓系数图

Fig. 2　Silhouette coefficient index chart

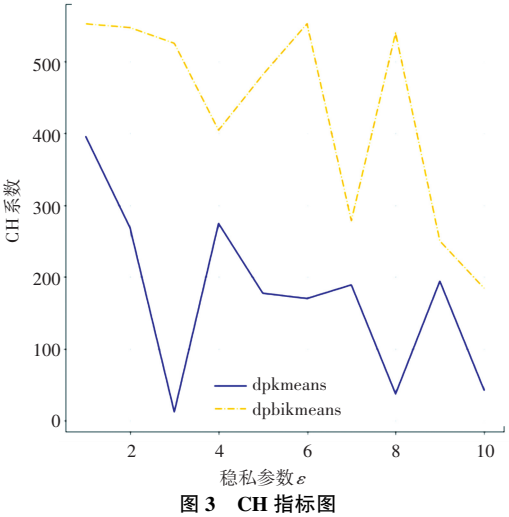

图 3　CH 指标图

Fig. 3　CH indicator chart

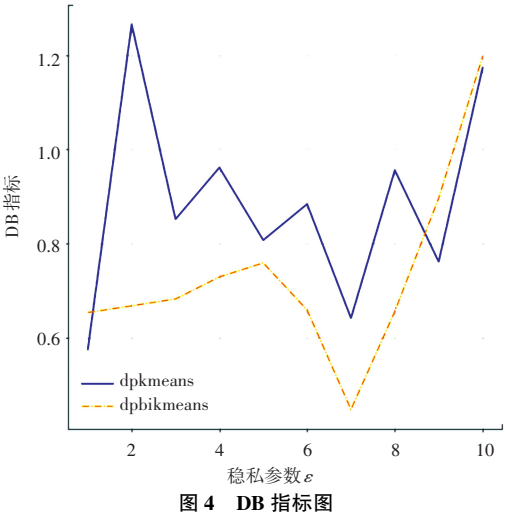

图 4　DB 指标图

Fig. 4　DB indicator chart

在将原始 IRIS 数据分类结果作为外部参考，将两种算法的调兰德指数（ARI）作为外部评估指标，

实验得出的 ARI 指数图如图 5 所示,可以观察到在隐私参数逐步增加,隐私保护水平在逐渐减小的过程中 DP Bi-k-means 算法的 ARI 指数相比 DP k-means 的 ARI 指数始终更大,DP Bi-k-means 算法的分类之间聚类的重叠程度更低。

通过对两种算法进行内部、外部评估,得知 DP Bi-k-means 相较于 DP k-means 有着更好的聚类性能。

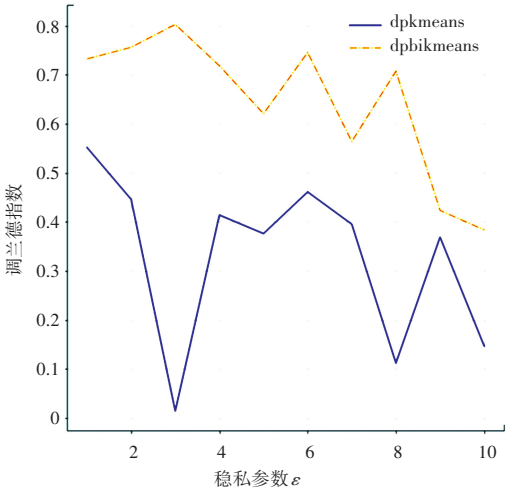

图 5 调整兰德指数图

Fig.5 Adjusted Rand index chart

3.2.2 鲁棒性分析

在对 IRIS 数据集聚类过程中进行两种算法的鲁棒性分析,在数据集聚类簇数 $k = 3$ 时进行实验,动态变化的隐私参数 ε 对 DP Bi-k-means 算法与 DP k-means算法的总误差平方和(SSE)的影响情况如图 6 所示,可以观察到 DP Bi-k-means 算法相较于 DP k-means 算法,平均误差平方和降低 61.15,且波动幅度更小,所以 DP Bi-k-means 算法有着更好的鲁棒性。

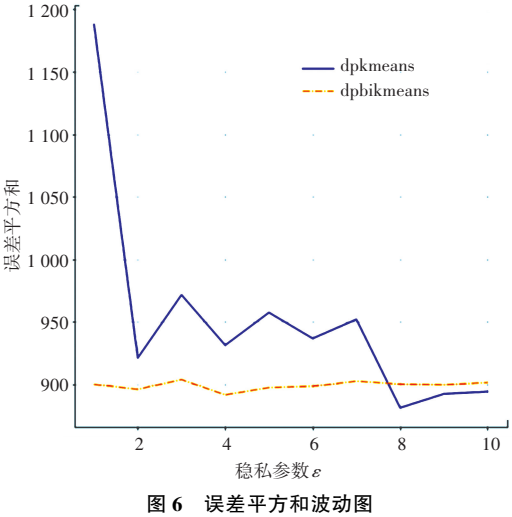

图 6 误差平方和波动图

Fig. 6 Sum of the squared errors chart

3.3.2 聚类效率分析

算法运行效率与相同条件下算法运行时间有直接联系,本文对比了两种划分方法在相同条件下的运行时间,运行时间图如图 7 所示 ,DP Bi-k-means 1 是选用划分后簇内误差平方和最大的簇进行划分,DP Bi-k-means 2 则为使用总误差平方和最小的簇进行划分,本文算法选择划分后簇内误差平方和最大的簇进行划分。由图 7 可知 DP Bi-k-means 1 比 DP Bi-k-means 2 运行时间更短,证明在保证其聚类效果优于传统 k-means 算法并且满足差分隐私保护的情况下,相较于一般二分 k 均值算法的划分方式提高了算法的运行效率。

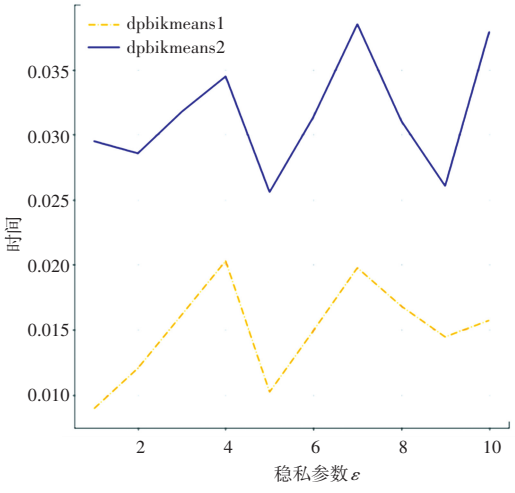

图 7 运行时间图

Fig. 7 Runtime chart

4 结束语

本文针对差分隐私保护的 k 均值聚类算法易受初始点影响,导致算法整体陷入局部最优解的问题,结合分层聚类的思想提出了一种差分隐私保护二分 k 均值聚类算法。在动态改变隐私参数的实验中与传统算法进行对比,由实验结果可知,新算法在保证其隐私保护能力的前提下,提高了聚类效果,增强了聚类算法的鲁棒性,并通过选取合适的簇划分规则提高了算法的运行效率,并证明新算法在不同隐私保护水平的情况下都有较好的性能,下一步将继续研究基于此算法的融合算法。

参考文献

[1] LU Tianliang, WANG Qiao, LIU Yingqing. Problems of User's Privacy Leakage During Insecure Communication[J]. Net Info Security, 2015, 15(9): 119-123.

(下转第 164 页)