文章编号：2095-2163(2021)12-0082-05 中图分类号：TP301.6 文献标志码：A

基于改进密度的簇内均值最小距离聚类算法

段桂芹
（广东松山职业技术学院 计算机与信息工程学院，广东 韶关 512126）

摘 要：针对密度聚类算法在聚类过程中存在的参数设置敏感、收敛时间长等问题，提出了一种改进密度聚类算法。首先使用自定义密度公式计算样本密度，得出候选代表点集合；再选取与其它候选代表点距离之和最小对象为首个初始聚类中心，使用最大乘积法完成初始中心选择；在簇中心更新环节，将与簇内均值最小距离的对象作为该簇的临时中心，使用最小距离法划分样本至所属簇中；重复该环节，直到收敛。在 UCI 数据集上的测试结果表明，改进密度算法相对 K-means 算法和其它两种改进算法具有更好的稳定性、更高的聚类准确率和更少的聚类耗时。

关键词：聚类；密度聚类；簇内均值最近点；候选代表点

Minimum distance clustering algorithm based on improved density

DUAN Guiqin
（Computer And the Information Engineering Institute, Guangdong Songshan Polytechnic College, Shaoguan 512126, China）

【Abstract】 Aiming at the problems of parameter setting sensitivity and long convergence in the clustering process of density clustering algorithm, an improved density clustering algorithm is proposed. Firstly, the sample density is calculated by using the user-defined density formula to get the set of candidate representative points. Then the object with the minimum distance from other candidate representative points is selected as the first initial clustering center with the maximum product method . In the process of updating the cluster center, the object with the minimum distance from the average value in the cluster is taken as the temporary center of the cluster, and the sample is divided into the cluster by the minimum distance method. The process is repeated until convergence. The test results on UCI dataset show that the improved density algorithm has better stability, higher clustering accuracy and less clustering time than k-means algorithm and the other two improved algorithms.

【Key words】 clustering; density clustering; nearest point of mean in cluster; candidate representative points

0 引 言

聚类的目标是将同一簇中样本相似度最大化，不同簇间样本相似度最小化。根据聚类过程的不同，通常将聚类分成基于密度、划分、模型、网格、层次 5 种聚类模型[1-2]。作为经典的基于划分模型的聚类算法，K-means 具有计算便捷、易于理解等特点[3]，由于 K-means 算法的初始聚类中心选取是随机的[4]，因此极易出现局部最优，导致聚类结果不稳定。此外，当样本中存在噪声或离群点时，聚类中心与真实中心存在较大偏差，易生成较差的聚类结果[5-6]。

1 相关算法研究现状

为了解决聚类分析中的问题，研究者们提出了多种改进算法。Zhai 等人[7]根据相距最远的数据对象不会划分至同一簇的基本思想，使用最大距离选取初始聚类中心，解决了原始算法更新次数多、耗时长等问题。在此基础上，段桂芹[8]将最大距离乘积法与 K-means 算法相结合，用相对分散的数据对象构造簇中心集合，优化了初始聚类中心，该方法在一定程度上解决了聚类结果不稳定等问题，但依然存在迭代次数多、运算耗时长等现象。邹臣嵩[9]针对文献[8]的这一问题，提出了一种协同 K 聚类算法。该算法通过样本的分布情况统计其密度参数，选取与样本集中心点最远的高密度对象为首个聚类中心，再通过最大乘积法求得其余聚类中心。徐红艳等[10]提出基于网格划分的快速确定全局阈值算法。该方法通过网格划分思想，将数据自适应地划分到多个网格空间中，再利用密度估计来计算密度，进而实现快速查找类簇峰值和低谷，最终完成数据集的有效识别。虽然文献[7-10]在一定程度上优

基金项目：广东省普通高校特色创新项目(2021KTSCX227)；韶关市科技计划项目(200811224533986)；韶关市科技计划项目(210718114531595)；广东省普通高校重点领域专项(2021ZDZX1124)。

作者简介：段桂芹(1979-)，女，硕士，讲师，主要研究方向：数据挖掘、计算机教育。

收稿日期：2021-09-06

化了 K-means 算法的初始聚类中心选择过程，但对噪声的影响有所忽略。潘品臣[11]根据密度参数理论，对数据集中的高密度点和非离群点区域进行了计算与识别，规避了选取噪音点为初始聚类中心以及中心分布不理想等问题。卜秋瑾[12]针对密度峰值聚类算法处理多密度峰值数据集时，人工选择聚类中心易造成簇的误划分问题，提出一种结合遗传 k 均值改进的密度峰值聚类算法。该算法能找到准确簇个数同时避免算法陷入局部最优，在提高聚类质量的同时，保证了聚类质量的稳定性。陈奕延[13]将样本涵盖的经典信息拓展到了模糊集上，利用寻找密度峰值的方法对模糊样本进行聚类，提出了误差更小的针对连续型模糊集与离散型模糊集的改进型欧氏距离，使改进后的欧氏距离具有更小的误差。

在对上述文献分析研究的基础上，本文从密度聚类算法入手，从两个方面对密度聚类算法进行了改进：一是根据高密度样本被低密度样本紧密围绕这一思想，提出了新的密度计算方法。目的是提高聚类中心的代表性，解决密度参数选取敏感等问题。二是在簇更新阶段，将与簇内均值距离最近的样本点作为该簇的临时中心，减少了迭代次数，降低运算耗时。

2　密度聚类算法改进

改进的密度聚类算法，由初始中心选择和簇中心迭代计算两部分构成。

(1)初始中心选择：首先使用自定义密度公式获取各样本密度，将高密度样本作为聚类中心候选代表点，并生成候选代表点集合；在集合中选取与其他候选代表点距离和最小者作为首个初始聚类中心，再使用乘积最大法完成初始聚类中心选择，得到初始聚类中心集合，即 $Z=\{z_1,z_2,\cdots,z_k\}$。

(2)簇中心迭代计算：根据集合 Z 完成初次聚类，计算簇内各样本到簇均值中心的距离矩阵。为了降低均值法所得簇中心和实际簇中心位置间的偏差，将与簇内均值距离最近的样本作为该簇的临时中心，生成临时簇中心集合，即 $Z=\{z_1',z_2',\cdots,z_k'\}$，再通过最小距离法将相关样本划分至所属簇中。重复簇中心迭代计算过程，直至准则函数收敛，完成聚类。

2.1　基本概念与公式

设 X 为含有 n 个样本的数据集，$X=\{x_1,x_2,\cdots,x_n\}$，样本的属性个数为 p，$x_i=\{x_i^1,x_i^2,\cdots,x_i^p\}$。$X$ 划分为 k 个簇，$X=\{C_1,C_2,\cdots,C_k\}$，$|C_i|$ 表示第 i 簇所含样本个数，z_k 表示第 k 簇的中心，多个簇中心所构成的集合为 Z，即 $Z=\{z_1,z_2,\cdots,z_k\}$。

定义 1　任意两样本间的欧氏距离为：

$$d(x_i,x_j)=\sqrt{\sum_{l=1}^{p}(x_i^l-x_j^l)^2} \tag{1}$$

其中：$i=1,2,\cdots,n$；$j=1,2,\cdots,n$；$l=1,2,\cdots,p$

定义 2　任意样本 x_i 的距离和定义为：该样本到数据集各样本的距离之和。

$$distSum(x_i)=\sum_{j=1}^{n}d(x_i,x_j) \tag{2}$$

定义 3　样本 x_i 的密度为：

$$density(x_i)=\sum_{j=1,x_i\neq x_j}^{n}\frac{distSum(x_j)}{d(x_i,x_j)} \tag{3}$$

本文密度的定义遵循的思路：从位置关系来看，当某个样本 x_i 被其它样本紧密围绕时，说明该样本与其它样本间的距离和相对较小；反之，当样本 x_i 和其它样本的位置关系较为分散时，说明该样本与其它样本间的距离和相对较大。密度表达式用样本 x_i 和 x_j 之间的距离做分母，用 x_j 到全部样本的距离之和做分子，用二者距离比值的累加和表示样本 x_i 被其它样本围绕的紧密程度，即 x_i 的密度。

以样本 x_i 为例。当式(3)累加和中的分子较大时，意味着除 x_i 外其它样本的累加距离和也较大；当分母较小时，意味着 x_i 到其它样本距离的累加和较小。因此，当分子越大且分母越小时，表达式的值越大，说明 x_i 被比自身密度低的样本所围绕的密集程度越大，即 x_i 的相对密度越大，其作为簇中心的代表性越强。

定义 4　样本集的平均密度定义为：

$$avgDensity(X)=\frac{\sum_{i=1}^{n}density(x_i)}{n} \tag{4}$$

定义 5　候选代表点集合定义为：密度高于样本集平均密度 α 倍的样本集合

$$H=\{h_i\} \tag{5}$$

其中：$x_i,\ x_j\in C_t,t=1,2,\cdots,k$。

定义 6　候选代表点间的距离矩阵定义为

$$\boldsymbol{HDist}=\begin{bmatrix}0 & d(h_1,h_2) & \cdots & d(h_1,h_j)\\ d(h_2,h_1) & 0 & \cdots & d(h_2,h_j)\\ \cdots & \cdots & 0 & \cdots\\ d(h_j,h_1) & d(h_j,h_2) & \cdots & 0\end{bmatrix} \tag{6}$$

式中，j 表示集合 H 所含元素个数。

定义 7 簇内样本与本簇均值中心间的距离矩阵定义为：

$$\boldsymbol{distMean}(m)=\begin{bmatrix} d(x_1,mean(C_m)) \\ d(x_2,mean(C_m)) \\ \cdots \\ d(x_{|C_m|},mean(C_m)) \end{bmatrix} \quad (7)$$

式中，$m=1,2,\cdots,k$，C_m 表示第 m 簇的样本集合。

定义 8 簇更新后，将与簇内均值距离最近的样本 x_i 作为该簇的中心。x_i 满足以下条件：

$$d(x_i,mean(C_m))=\min(distMean(m)) \quad (8)$$

定义 9 聚类误差平方和 E 的定义为

$$E=\sum_{i=1}^{k}\sum_{j=1}^{m}|x_{ij}-z_i|^2 \quad (9)$$

式中，x_{ij} 为第 i 簇中第 j 个样本，z_i 为第 i 簇的簇中心。

2.2 算法描述

步骤 1 使用式(1)~式(3)计算各样本的密度；

步骤 2 使用式(4)、式(5)得到候选代表点集合 H，其中参数 α 为 1.0；

步骤 3 用式(1)、式(6)计算候选代表点间的距离矩阵，在 H 中选择与其它候选代表点距离和最小者作为首个聚类中心 z_1 存储至集合 Z 中；

步骤 4 在集合 H 中选择与 z_1 距离最远的候选代表点 z_2 存储至集合 Z 中；

步骤 5 从集合 H 中选择满足条件：$\max(d(hi,z_1)\times d(hi,z_2))$ 的代表点，作为 z_3 存储至集合 Z 中；

步骤 6 重复运行步骤 5，直至 $|Z|=k$；

步骤 7 使用式(1)计算 X 中各样本与集合 Z 中各候选点的距离，并划分至距离最小的簇中；

步骤 8 使用式(7)计算簇内各样本到簇均值中心的距离矩阵，根据式(8)将距离簇内均值最近的样本作为该簇的新中心；

步骤 9 重复步骤 7、8，更新簇中心集合 Z；

步骤 10 将 X 中的样本按距离划分至最近的簇中，使用式(9)计算并判断 E 是否收敛，若收敛，则算法终止；若未能收敛，将跳转至步骤 7，再次更新簇中心。

2.3 算法复杂度分析

本文算法的时间复杂度为 $O(n^2+nkt)$，在初始聚类中心选择过程中，本文算法首先计算了各样本的密度，进而得出候选代表点集合，再使用距离乘积最大法对候选点从空间分布的角度进行了二次筛选。虽然计算量有所增加，但各中心的代表性得到增强，初步反映了样本集的几何结构，为簇更新次数的降低提供支撑。在簇中心更新过程中，本文算法选取与簇内均值距离最近的样本作为该簇的临时中心，生成了临时簇中心集合，避免了均值法所得簇中心和实际簇中心位置存在偏差的隐患，相比均值算法，本文算法可以降低噪音的干扰，减少更新次数，降低计算开销，提高运算效率。

3 实验仿真与分析

实验运行环境：CPU Intel Core i7-2670 2.20 GHz，硬盘 1T，内存 8G；操作系统 Win10-64 位；仿真软件采用 Matlab 2011b。在有效性验证方面，采用聚类准确率、聚类各阶段开销、Rand 指数、Jaccard 系数等指标，将 K-means 算法、文献[8]中算法、文献[9]中算法与本文算法进行了比较。实验过程中，K-means 算法共运行 200 次，取其平均值作为该算法的实验结果。实验数据集详见表 1。

表 1 UCI 数据集

Tab. 1 UCI dataset

样本集名称	样本总数	属性个数	正确聚类数
iris	150	4	3
balance-scale	625	4	3
new-thyroid	215	5	3
haberman	306	3	2
wdbc	569	30	2

3.1 算法性能对比与分析

图 1~图 5 是 K-means 算法、文献[8-9]算法以及本文算法的聚类准确率、簇中心各阶段计算开销、簇更新次数等实验的对比结果。由图 1 可知，在 iris 和 wdbc 数据上，本文算法的聚类准确率明显高于 K-means算法，略高于文献[8-9]算法，在 balance、thyroid 和 haberman 数据集上的聚类准确率优于其它 3 种算法。

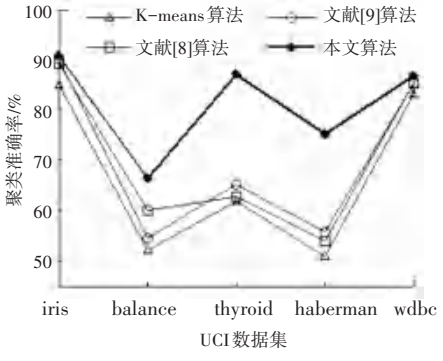

图 1 聚类准确率

Fig. 1 Clustering accuracy

图 2 可知，K-means 的初始中心从样本集中随机选择，耗时较少，而文献[8-9]对初始聚类中心的选择过程进行了优化，一定程度上增加了计算开销，故耗时相对较多。与文献[8-9]相比，本文算法先对各样本的密度进行计算，再对高密度代表点进行了二次筛选，以确定初始聚类中心集合。因此，在计算量方面开销较大。除 thyroid 数据集的耗时低于文献[9]外，其他数据集上的耗时均略高于文献[8-9]。

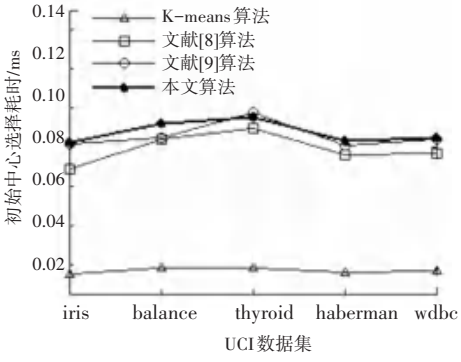

图 2　初始中心选择耗时

Fig. 2　Initial center selection time

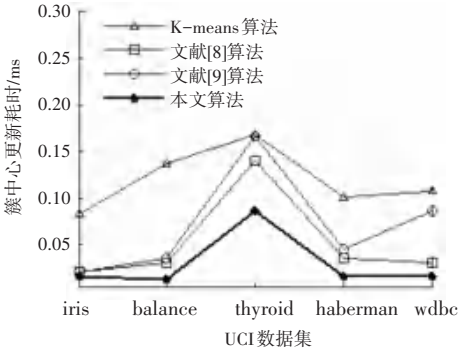

图 3　簇中心更新耗时

Fig. 3　Cluster center update time consuming

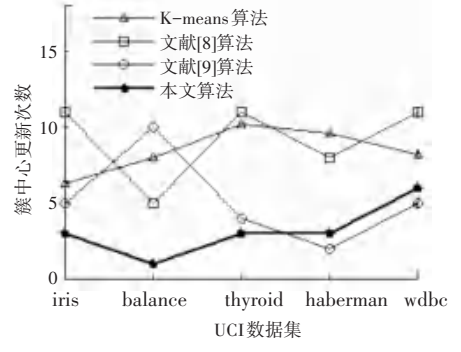

图 4　簇中心更新次数

Fig. 4　Number of cluster center updates

从图 3 可见，本文算法的簇中心更新耗时小于 K-means 和文献[8-9]算法。主要原因在于本文算法将与簇内均值距离最近的样本点作为该簇的临时中心，使得簇中心的存在更加具体。每一次更新后，簇中心的位置和样本的分布情况会愈加明了，簇中心的代表性得到增强，从而降低了簇更新耗时。

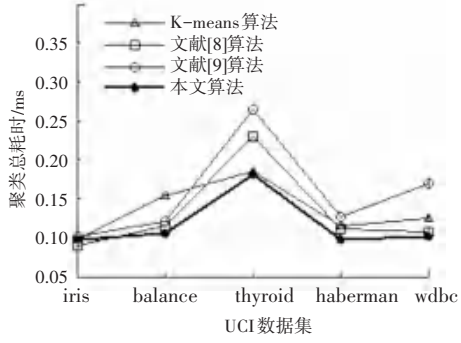

图 5　聚类总耗时

Fig. 5　Total clustering time

从图 4、图 5 可知，本文算法在中心点迭代次数、总耗时上总体上优于 K-means 算法和文献[8-9]算法。这是因为 K-means 算法随机选择了初始聚类中心，使得准则函数容易收敛到局部最优，且簇更新次数不稳定。此外，文献[8-9]依然沿用均值中心算法完成簇更新，并未对该阶段进行优化，未能更好地体现临时中心点在当前簇中的代表性。

综上，本文在初始中心选择阶段所提出的密度计算方法，使得初始中心空间分布合理，具有较强的代表性。簇更新后，簇中心的存在清晰明了，在整体上能够减少簇更新次数，降低运算耗时。

3.2　其它外部评价指标对比

在聚类结果评价方面，除使用上述常用指标外，还使用 Rand、Jaccard、Adjusted Rand Index 3 个评价指标[14-17]对 5 种样本的聚类结果加以测试对比。观察表 2 ~ 表 4，在 Rand 指数对比结果中，除 Balance 数据集本文算法略低于文献[8]算法外，其它 4 个数据集的 Rand 指数均优于其它 3 种算法；在 Jaccard 系数和 Adjusted Rand Index 参数对比结果中，本文算法全部优于其它 3 种算法。从几种常见聚类指标对比实验结果中可以看出：本文提出的改进聚类算法稳定性更强，聚类质量更高。

表 2　Rand 指数比较

Tab. 2　Comparison of rand index

	K-means	文献[8]	文献[9]	本文
iris	0.82	0.873 7	0.879 7	0.885 9
balance	0.58	0.667 8	0.605 5	0.664 6
thyroid	0.54	0.579 8	0.602 4	0.790 7
haberman	0.45	0.498 3	0.499 1	0.650 0
wdbc	0.72	0.750 3	0.750 3	0.770 7

表 3 Jaccard 系数比较

Tab. 3 Comparison of Jaccard coefficient

	K-means	文献[8]	文献[9]	本文
iris	0.66	0.681 9	0.695 5	0.711 6
balance	0.29	0.400 6	0.319 1	0.410 7
thyroid	0.39	0.402 9	0.416 6	0.697 9
haberman	0.36	0.391 8	0.391 5	0.616 4
wdbc	0.62	0.649 9	0.649 9	0.670 3

表 4 Adjusted Rand Index 参数比较

Tab. 4 Comparison of Adjusted Rand index parameter

	K-means	文献[8]	文献[9]	本文
iris	0.69	0.716 1	0.730 0	0.745 3
balance	0.15	0.301 4	0.172 6	0.304 5
thyroid	0.152	0.164 0	0.211 1	0.573 8
haberman	0	-0.003 5	-0.000 5	0.132 4
wdbc	0.45	0.491 4	0.491 4	0.533 7

4 结束语

本文用改进密度算法对 K-means 聚类算法进行了优化,解决了密度聚类算法的参数设置敏感、收敛时间长等问题。新算法的初始聚类中心相对分散且具有代表性,能够在聚类初期反映出样本的大致分布;在簇更新阶段,选取了与簇内均值距离最近的样本点作为该簇的临时中心,使得簇中心的位置更加准确,减少了迭代次数和计算开销。对比测试表明,新算法能够快速准确逼近全局最优解。

参考文献

[1] SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research [J]. Journal of Software, 2008, 19(1):48-61.

[2] HAN J, KAMBER M. Data mining concept and techniques[M]. Beijing : China Machine Press. 2001: 443-496.

[3] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proc, of, Berkeley Symposium on Mathematical Statistics and Probability, 1967:281-297.

[4] YANG S L, LI Y S, HU XX, et al. Optimization Study on k Value of Kmeans Algorithm[J]. Systems Engineering-Theory & Practice, 2006.

[5] YUAN Fang, ZHOU Zhiyong, SONG Xin. K-means clustering algorithm with meliorated initial center[J]. Computer Engineering, 2007, 33(3):65-66.

[6] NTOUTSI I, ZIMEK A, PALPANAS T, et al. Density-based Projected Clustering over High Dimensional Data Streams[C]// Proceedings of the 2012 SIAM International Conference on Data Mining, 2012:987-998.

[7] ZHAI Donghai, YU Jiang, GAO Fei, et al. K-means text clustering algorithm based on initial cluster centers selection according to maximum distance[J]. Application Research of Computers, 2014, 31(3):713-715.

[8] 段桂芹. 基于均值与最大距离乘积的初始聚类中心优化算法[J]. 计算机与数字工程, 2015, 43(3):379-382.

[9] 邹臣嵩, 杨宇. 基于最大距离积与最小距离和协同聚类算法[J]. 计算机应用与软件, 2018, 35(5):297-301.

[10] 徐红艳, 普蓉, 黄法欣,等. 基于网格和密度比的 DBSCAN 聚类算法研究[J]. 计算机与数字工程, 2020, 48(6):1269-1274.

[11] 潘品臣, 姜合, 吕奕锟. 一种非独立同分布下 K-means 算法的初始中心优化方法[J]. 小型微型计算机系统, 2019, 40(6):1254-1259.

[12] 卜秋瑾, 段隆振, 段文影. 结合遗传均值改进的密度峰值聚类算法[J]. 计算机工程与设计, 2020, 41(4):1012-1016.

[13] 陈奕延, 李晔, 李存金. 一种基于密度峰值的针对模糊混合数据的聚类算法[J]. 计算机工程与科学, 2020, 42(2):317-324.

[14] William M, Rand. Objective Criteria for the Evaluation of Clustering Methods [J]. Journal of the American Statistical Association, 1971, 66(336):846-850.

[15] CALINSKI T, HARABASZ J. A dendrite method for cluster analysis[J]. Communications in Statistics, 1974, 3(1):1-27.

[16] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(1):193-218.

[17] Paul E Green, Jonathan Kim, Frank J. Carmone. A preliminary study of optimal variable weighting in K-means clustering[J]. Journal of Classification, 1990, 7(2):271-285.

(上接第 81 页)

[11] 徐鹏. 铁路线路轨道动态不平顺变化特征研究[D]. 北京: 北京交通大学, 2009.

[12] 中华人民共和国行业标准. [2009]674 号 高速铁路无砟轨道工程施工精调作业指南[S]. 北京:中国铁道出版社, 2009.

[13] 杨帆,方成刚,洪荣晶,等. 改进遗传算法在车间调度问题中的应用[J/OL]. 南京工业大学学报(自然科学版):1-8[2021-03-29].

[14] 李敏强. 遗传算法的基本理论与应用[M]. 北京: 科学出版社, 2002:17-26.

[15] 于蒙,刘德汉. 改进 PSO-GA 算法求解混合流水车间调度问题[J/OL]. 武汉理工大学学报(交通科学与工程版):1-13[2021-03-29].

[16] 陈宇奇,陈家琪. 基于 PSO 优化的车辆稳定性研究[J]. 电子科技,2017,30(1):83-86.