



基于相异性度量选取初始聚类中心改进的K-means聚类算法

廖纪勇, 吴晟, 刘爱莲

引用本文:

廖纪勇, 吴晟, 刘爱莲. 基于相异性度量选取初始聚类中心改进的K-means聚类算法[J]. 控制与决策, 2021, 36(12): 3083–3090.

在线阅读 View online: <https://doi.org/10.13195/j.kzyjc.2020.0554>

您可能感兴趣的其他文章

Articles you may be interested in

基于波段影像统计信息量加权K-means聚类的高光谱影像分类

Algorithm based on band statistical information weighted K-means for hyperspectral image classification

控制与决策. 2021, 36(5): 1119–1126 <https://doi.org/10.13195/j.kzyjc.2019.1516>

基于相互邻近度的密度峰值聚类算法

Density peaks clustering based on mutual neighbor degree

控制与决策. 2021, 36(3): 543–552 <https://doi.org/10.13195/j.kzyjc.2019.0795>

基于边缘峰度度量的特征缩减模糊聚类算法

Feature-reduction fuzzy clustering algorithm based on marginal kurtosis measure

控制与决策. 2021, 36(11): 2665–2673 <https://doi.org/10.13195/j.kzyjc.2020.0220>

基于时空聚类求解带容积约束的选址-路径问题

Time-space cluster based location-routing problem with capacitate constraints

控制与决策. 2021, 36(10): 2504–2510 <https://doi.org/10.13195/j.kzyjc.2020.0073>

基于搜索空间划分与Canopy K-means聚类的种群初始化方法

Population initialization based on search space partition and Canopy K-means clustering

控制与决策. 2020, 35(11): 2767–2772 <https://doi.org/10.13195/j.kzyjc.2019.0358>

基于相异性度量选取初始聚类中心 改进的 K -means 聚类算法

廖纪勇, 吴 晟[†], 刘爱莲

(昆明理工大学 信息工程与自动化学院, 昆明 650500)

摘 要: 选取合理的初始聚类中心是正确聚类的前提, 针对现有的 K -means 算法随机选取聚类中心和无法处理离群点等问题, 提出一种基于相异性度量选取初始聚类中心改进的 K -means 聚类算法. 算法根据各数据对象之间的相异性构造相异性矩阵, 定义了均值相异性和总体相异性两种度量准则; 然后据此准则来确定初始聚类中心, 并利用各簇中数据点的中位数代替均值以进行后续聚类中心的迭代, 消除离群点对聚类准确率的影响. 此外, 所提出的算法每次运行结果保持一致, 在初始化和处理离群点方面具有较好的鲁棒性. 最后, 在人工合成数据集和 UCI 数据集上进行实验, 与 3 种经典聚类算法和两种优化初始聚类中心改进的 K -means 算法相比, 所提出的算法具有较好的聚类性能.

关键词: 聚类分析; K -means 算法; 初始聚类中心; 离群点; 相异性度量; 鲁棒性

中图分类号: TP391

文献标志码: A

DOI: 10.13195/j.kzyjc.2020.0554

开放科学(资源服务)标识码(OSID):



引用格式: 廖纪勇, 吴晟, 刘爱莲. 基于相异性度量选取初始聚类中心改进的 K -means 聚类算法[J]. 控制与决策, 2021, 36(12): 3083-3090.

Improved K -means clustering algorithm for selecting initial clustering centers based on dissimilarity measure

LIAO Ji-yong, WU Sheng[†], LIU Ai-lian

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Selecting a reasonable initial clustering center is the premise of correct clustering. Most of the existing K -means algorithms have some shortcomings, such as randomly selecting clustering centers and unable to deal with outliers, an improved K -means clustering algorithm for selecting initial clustering centers based on dissimilarity measure is proposed. According to the dissimilarity of each data object, the dissimilarity matrix is constructed, and two measures of mean dissimilarity and total dissimilarity are defined. Then the initial clustering center is determined according to the criteria, and the median of data points in each cluster is used to replace the mean value for the subsequent iteration of clustering center, so as to eliminate the effect of outliers on clustering accuracy. In addition, the proposed algorithm maintains consistent results every time, and has better robustness in initializing and handling outliers. Finally, experiments are performed on the synthetic datasets and UCI datasets. Compared with three classical clustering algorithms and two improved K -means algorithms, the proposed algorithm has better clustering performance.

Keywords: clustering analysis; K -means algorithm; initial clustering center; off-group points; dissimilarity measure; robustness

0 引 言

聚类分析根据“物以类聚”原则将相似度高的对象聚为一类, 而簇间对象彼此相似度尽量低^[1]. 针对各领域的不同应用, 经过数十年的深入研究, 国内外学者提出了大量的聚类算法, 并广泛应用于生物学、模式识别、图像分析等领域^[2-4]. 目前已有的聚类算

法一般可分为以下几类: 基于划分的聚类算法, 如 K -means^[5]、 K -medoids^[6]、AP^[7] 等; 基于密度的聚类算法, 如 DBSCAN^[8]、OPTICS^[9] 等; 基于网格的聚类算法, 如 WaveCluster^[10]、STING^[11] 等. 然而, 传统的聚类算法都存在缺点, 如 K -means 算法对初始聚类中心的选择极其敏感, 容易陷入局部最优解, 聚类结果不稳

收稿日期: 2020-05-11; 修回日期: 2020-08-05.

责任编辑: 薛建儒.

[†]通讯作者. E-mail: ws8146@163.com.

定.

近年来,针对 K -means 聚类算法易受离群点的影响和随机选取初始聚类中心点等问题,研究者提出了相应的改进方法. 其中 K -medoids 算法是最著名的改进算法,通过最小化平方误差和的方法寻找局部最优解,选取簇中距离最小的点作为聚类中心,解决了离群点对聚类结果的影响,但初始聚类中心点的选取是随机的,对结果的稳定性产生较大影响. 文献[12]提出了一种基于离群因子和最大最小算法优化初始聚类中心的 K -means 算法,消除了离群点对聚类结果的影响,但第1个初始聚类中心点的选取是随机的. 文献[13]选取数据集中平均差异度较大的点作为初始聚类中心,消除了随机性对聚类准确性的影响,但无法处理异常值对聚类结果的影响. 文献[14]根据 Pearson 相关系数计算各数据的交互关系并形成一个对应的新数据集,利用最小最大值的方法选取初始聚类中心点,解决了初始聚类中心点敏感的问题. 但是,改进算法没有同时解决随机选取初始聚类中心和离群点影响的问题.

针对上述两类问题,本文提出一种基于相异性度量选取初始聚类中心改进的 K -means 聚类算法(improved K -means clustering algorithm for selecting initial clustering centers based on dissimilarity measure, IK-DM). 算法在进行迭代之前,首先构造相异性矩阵,计算数据点的均值相异性,选取数据点中相异性最大的数据点作为初始聚类中心;然后,将各簇中数据点的中位数代替均值进行后续聚类中心点的选取,消除离群点对聚类中心的影响. 只需输入聚类数,就能够自动确定聚类中心点,并且算法每次运行所得结果完全一致,保证了聚类结果的稳定性.

1 IK-DM算法

1.1 相关定义

给定数据集 $X_{n \times m} = \{x_1, x_2, \dots, x_n\}$. 其中: $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]$, n 为数据集样本总数, m 为样本维度. 本文算法涉及到一些新的定义: 数据点的相异性 dis , 相异性矩阵 $\text{dis}M$, 均值相异性 Adis , 总体相异性 Tdis .

定义1 用数据点的欧氏距离定义它们的相异性,计算公式为

$$\text{dis}(x_i, x_j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}, 1 \leq i, j \leq n. \quad (1)$$

定义2 相异性矩阵存储 n 个对象两两之间的临近度,是一个对称矩阵,即

$$\text{dis}M = \begin{bmatrix} 0 & \text{dis}(x_1, x_2) & \dots & \text{dis}(x_1, x_n) \\ \text{dis}(x_2, x_1) & 0 & \dots & \text{dis}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{dis}(x_n, x_1) & \text{dis}(x_n, x_2) & \dots & 0 \end{bmatrix}.$$

定义3 均值相异性是数据点 x_i 与数据集中每个对象的距离平均值,记为 $\text{Adis}(x_i)$,有

$$\text{Adis}(x_i) = \frac{1}{n} \sum_{j=1}^n \text{dis}(x_i, x_j), \quad (2)$$

其中 $\text{Adis}(x_i)$ 反映数据点 x_i 在整个数据集的位置情况. 其值越大,说明 x_i 周围数据分布越稀疏且与其他点远离程度越高;反之越稠密.

定义4 数据集的总体相异性定义如下:

$$\text{Tdis} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{dis}(x_i, x_j). \quad (3)$$

分析式(3)可知,数据集的总体相异性与全体数据的分布有关,体现了数据集的稀疏程度.

1.2 IK-DM算法描述

本文算法首先对数据点进行相异性计算,构造相异性矩阵,并计算出各数据点的均值相异性,选取均值相异性最大的数据点作为第1个初始聚类中心点;然后计算数据集的总体相异性,选取除已被选为中心点以外的均值相异性最大的数据点作为第2个聚类中心点,并计算该数据点与已选定的各个聚类中心点的相异性,若大于数据集的总体相异性,则该样本可以作为第2个聚类中心,否则选取均值相异性第2大的样本进行判断,循环该操作直至选出所有的初始聚类中心. IK-DM 算法利用相异性尽可能大的数据点作为初始聚类中心点,可以避免聚类中心点过于邻近的情况,能够降低算法的迭代次数. 算法主要步骤如下.

算法1 IK-DM算法.

输入: 数据集 X , 聚类簇数 k ;

输出: 聚类结果 C .

step 1: 计算. 根据式(1)~(3)计算数据对象的相异性 dis 、均值相异性 Adis 、总体相异性 Tdis , $k = 1$.

step 2: 初始聚类中心选取.

step 2.1: 选取第1个初始聚类中心点,即 $\mu_1 = \arg \max_{i \in \{1, 2, \dots, n\}} [\text{Adis}(x_i)]$, 用0代替该样本点的均值相异性并更新, $k = k + 1$;

step 2.2: 计算均值相异性最大的样本与已选聚类中心的相异性,找出满足以下条件的样本 x_i 作为第 k 个聚类中心: $\text{dis}(x_i, \mu_j | j = 1, 2, \dots, k-1) \geq \text{Tdis}$, 用0代替该样本点的均值相异性并更新, $k = k + 1$;

step 2.3: 判断聚类中心点个数是否等于设定的聚类簇数, 若相等, 则聚类中心点为 $\{\mu_1, \mu_2, \dots, \mu_k\}$, 否则转 step 2.2.

step 3: 改进的 K -means 聚类算法.

step 3.1: 根据距离最近原则确定各数据点 $x_j (1 \leq j \leq n)$ 所属簇, 标记为 $\lambda_j = \arg \max_{i \in \{1, 2, \dots, k\}} \{\|x_j - \mu_i\|_2\}$, 将数据点 x_i 划入相对应的簇 $C_{\lambda_j} = C_{\lambda_j} \cup x_j$.

step 3.2: 用中位数代替均值计算新的聚类中心, 将每个簇的数据进行排序, 计算新的聚类中心, 即

$$\mu'_i = \begin{cases} x_{\lfloor \frac{|C_i|+1}{2} \rfloor}, & |C_i| \text{ 为奇数;} \\ \frac{1}{2}(x_{\lfloor \frac{|C_i|}{2} \rfloor} + x_{\lfloor \frac{|C_i|}{2} \rfloor + 1}), & |C_i| \text{ 为偶数.} \end{cases}$$

step 3.3: 如果 $\mu_i \neq \mu'_i$, 则用新的聚类中心 μ'_i 替代 μ_i 并转 step 3.1, 否则聚类结束.

1.3 算法分析

1.3.1 计算复杂度

IK-DM 算法首先需要构造相异性矩阵, 时间开销为 $O(n^2)$; 其次, 选取初始聚类中心的过程中, 后续 $(k-1)$ 个聚类中心点与已选取的聚类中心点之间的均值相异性均需大于总体相异性, 所以至多扫描 n 个数据点, 于是在最坏的情况下, 时间复杂度为 $O(kn)$; 最后, 利用改进的 K -means 聚类算法进行迭代, 所需的计算开销为 $O(tkn)$, 其中 t 是算法收敛需要迭代的次数. 因此, IK-DM 算法的时间复杂度为 $O(n^2 + (t+1)kn)$.

1.3.2 鲁棒性与收敛性分析

根据文献[14], 从以下3个方面分析鲁棒性: 1) 初始聚类中心和聚类数的鲁棒性, IK-DM 算法利用启发式算法自动选取初始聚类中心, 且每次选取的中心点保持一致; 2) 检测不同数量簇的能力, IK-DM 算法能够处理大型数据集, 对于不同形状的簇具有一定的处理能力; 3) 处理离群点和噪声的能力, IK-DM 算法利用簇中数据点的中位数代替均值选取新的聚类中心点, 能够将离群点排除在候选聚类中心点之外, 降低异常值对聚类结果的影响, 因此, 采用该算法处理离群点具有很好的鲁棒性.

由于 IK-DM 算法是 K -means 的泛化版本, 收敛性依然与 K -means 算法保持一致. 下面定性地描述 IK-DM 算法的收敛性. IK-DM 算法的目标损失函数为

$$J(\mu_1, \mu_2, \dots, \mu_k) = \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^k (x_j - \mu_i)^2. \quad (4)$$

IK-DM 算法的目标是使 J 的值最小化. 假设当前 J 没有达到最小值, 则可以固定参数 μ_k , 通过调整

每个样本点的所属类别使 J 函数减小. 同理, 固定样本点的分配, 更新参数 μ_k , 调整每个类的聚类中心也可以使得 J 函数减小. 运行 IK-DM 算法可以单调优化目标损失函数, 保证算法的收敛性.

2 实验结果与分析

为了验证本文改进算法的聚类性能, 利用数据集对 IK-DM 算法进行分析与评估. 在对比实验中, 将本文提出的 IK-DM 算法与 K -means^[5]、 K -medioids^[6]、AP^[7]、OFMMK-means^[12] 以及 IIEFA^[15] 算法进行性能比较, 并对所得结果进行适当的分析.

实验环境: 硬件平台为 Intel I5-6200U, 2.3 GHz, 8 GB RAM; 软件环境为 Windows 10, 64 位, Matlab R2014b.

2.1 算法参数设置

为了保证对比实验的有效性, 对每种算法都进行参数调优.

1) 由于 K -means、 K -medioids、OFMMK-means 算法每次聚类结果都会有波动, 为了增加实验结果的可靠性, 在每个数据集上运行算法 30 次, 取各聚类评价指标的平均值作为最终的聚类结果.

2) AP 和 IIEFA 算法的最大迭代次数均设为 $t_{\max} = 500$, 相似度的计算利用负欧氏距离, 偏向参数 p 取中位数, 阻尼因子 $\lambda \in [0, 1]$. IIEFA 算法中光强吸引系数取 $\gamma = 0.1$, 吸引度系数 $\beta = 1$, 步长因子 $\alpha \in [0, 1]$.

AP 和 IIEFA 算法以 0.01 为步长调整 λ 与 α 的取值. 当聚类结果达到最优时, 将各数据集上的最优参数作为最终的聚类输入值.

2.2 实验数据集

为了充分验证本文算法的聚类效果, 采用不同特征的数据集进行实验, 主要包括人工合成数据集和 UCI 真实数据集, 详细信息见表 1.

表 1 数据集的基本特征

编号	数据集	数据来源	样本个数	维度	类别个数
1	DS1	文献[16]	2 000	2	5
2	Flame	文献[17]	240	2	2
3	Aggregation	文献[17]	788	2	7
4	R15	文献[17]	600	2	15
5	Iris	文献[18]	150	4	3
6	Wine	文献[19]	178	13	3
7	Seeds	文献[16]	210	7	3
8	Yeast	文献[17]	1 484	7	10
9	Glass	文献[17]	214	9	6
10	Ecoli	文献[19]	336	6	8
11	Waveform	文献[18]	5 000	40	3

编号 1~4 为人工合成数据集,包含 4 个任意形状、不同密度、不同数据量的二维数据集,并含有桥接噪声或随机噪声.

2.3 评价指标

1) 准确度 (clustering accuracy, ACC)^[20],定义如下:

$$ACC = \sum_{i=1}^k y_i / n, \tag{5}$$

其中 y_i 表示簇 i 中聚类正确的数据点个数. ACC 的值越大,表明聚类结果越准确.

2) 平均聚类纯度 purity ^[21],定义如下:

$$\text{purity} = \frac{1}{k} \sum_{i=1}^k \frac{y_i}{y_i^s}, \tag{6}$$

其中 y_i^s 表示簇 i 中数据点的个数. purity 的值越大,说明每个簇的质量越高且分类越准确.

3) 标准化互信息 (normalized mutual information, 简称 NMI)^[22],定义如下:

$$NMI = \frac{\sum_{i=1}^k \sum_{j=1}^m n_{i,j}^{a,b} \log \frac{n \cdot n_{i,j}^{a,b}}{n_i^a \cdot n_j^b}}{\sqrt{\left(\sum_{i=1}^k n_i^a \log \frac{n_i^a}{n}\right) \cdot \left(\sum_{j=1}^m n_j^b \log \frac{n_j^b}{n}\right)}}. \tag{7}$$

其中: m 表示数据集本质聚类类别数, $n_{i,j}^{a,b}$ 表示数据点属于真实标签类 j 但被划分到聚类结果簇 i 中的个数, n_i^a 表示聚类结果簇 i 中数据点的个数, n_j^b 表示真实标签类 j 中数据点的个数.

2.4 实验结果

2.4.1 人工合成数据集

表 2 展示了各种算法在人工合成数据集上的聚类评价指标值,其中加粗的数据表示聚类指标最好

的实验结果. 为了更直观地显示 IK-DM 算法与其余 5 种算法的聚类情况,将聚类对比效果进行可视化处理. 使用不同颜色分辨不同的簇,结果如图 1 所示.

通过对实验结果的分析,可得以下结论:

1) 从表 2 的 3 个有效性指标来看, K -means 和 K -medioids 算法聚类效果最差. AP 算法的聚类效果整体优于 K -means、 K -medioids、OFMMK-means 算法,但该算法聚类结果的聚类数往往多于真实数据集的类别数. IIEFA 算法利用启发式的算法来确定初始聚类中心,在多数情况下的聚类效果较好,且对于簇间密度差异较大的 Flame 数据集, IIEFA 算法优于其他 5 种算法. 在处理簇间密度相对均匀且含有异常值的 DS1、Aggregation、R15 数据集时, K -means 及其改进的 OFMMK-means、 K -medioids、AP 算法的聚类效果不佳;而 IK-DM 算法能够准确处理有桥接干扰或离群点的数据,并且所得到的聚类簇数与真实类别数一致,所发掘的聚类代表点大部分都是数据集中的真实数据点,本质上更具有代表性.

2) 图 1 是各算法在合成数据集上的聚类效果,从图 1(a)、1(b)和 1(e)中可以看出, K -means、 K -medioids、OFMMK-means 算法聚类性能明显不佳. 造成这种现象的主要原因是,算法选取聚类中心具有随机性,从而增加了有效聚类的难度,且聚类中心容易偏离实际聚类代表点,导致聚类效果变差. AP 算法所发掘的聚类中心点虽然都是数据集中的真实数据点,但从图 1(c)中可以看出,该算法得到的聚类中心数远远大于真实聚类中心点数目. 从图 1(f)可以看出,本文算法在处理密度分布较为均匀的数据集时都能够取得理想的聚类效果,这是因为 IK-DM 算法既考虑了初始聚类中心点的选取,又利用中位数代替均值来降低离群点的影响,保证了算法的稳定性.

表 2 各种算法在人工合成数据集上的性能比较

数据集	评价指标	K -means	K -medioids	AP	IIEFA	OFMMK-means	IK-DM
DS1	ACC	0.752 8 (0.205 3)	0.746 5 (0.142 1)	0.975 7 (0.000 0)	0.996 0 (0.000 0)	0.974 5 (0.001 7)	0.996 0 (0.000 0)
	purity	0.751 9 (0.219 4)	0.756 3 (0.0995)	0.994 3 (0.000 0)	0.967 9 (0.000 0)	0.970 2 (0.003 3)	0.994 3 (0.000 0)
	NMI	0.794 0 (0.054 7)	0.853 8 (0.055 2)	0.980 4 (0.000 0)	0.960 7 (0.000 0)	0.966 1 (0.000 4)	0.986 1 (0.000 0)
Flame	ACC	0.798 5 (0.025 3)	0.813 5 (0.039 8)	0.865 5 (0.000 0)	0.865 5 (0.000 0)	0.803 8 (0.054 7)	0.837 5 (0.000 0)
	purity	0.803 2 (0.010 2)	0.811 5 (0.003 7)	0.856 6 (0.000 0)	0.894 1 (0.000 0)	0.827 8 (0.018 1)	0.831 5 (0.000 0)
	NMI	0.398 9 (0.001 2)	0.566 5 (0.014 0)	0.441 2 (0.000 0)	0.566 5 (0.000 0)	0.452 1 (0.012 5)	0.398 9 (0.000 0)
Aggregation	ACC	0.769 1 (0.103 1)	0.835 4 (0.035 2)	0.866 5 (0.000 0)	0.953 5 (0.000 0)	0.901 2 (0.034 1)	0.960 3 (0.000 0)
	purity	0.749 5 (0.123 6)	0.813 6 (0.074 9)	0.833 4 (0.000 0)	0.948 7 (0.000 0)	0.894 5 (0.041 2)	0.956 5 (0.000 0)
	NMI	0.838 3 (0.068 5)	0.789 9 (0.023 5)	0.775 0 (0.000 0)	0.925 7 (0.000 0)	0.925 7 (0.000 6)	0.925 7 (0.000 0)
R15	ACC	0.811 7 (0.125 4)	0.813 3 (0.132 8)	0.921 8 (0.000 0)	0.992 5 (0.000 0)	0.908 7 (0.022 5)	0.992 5 (0.000 0)
	purity	0.783 3 (0.105 2)	0.785 3 (0.116 6)	0.921 8 (0.000 0)	0.986 5 (0.000 0)	0.891 7 (0.012 0)	0.986 5 (0.000 0)
	NMI	0.923 3 (0.002 1)	0.931 9 (0.001 5)	0.935 4 (0.000 0)	0.985 7 (0.000 0)	0.985 7 (0.000 2)	0.985 7 (0.000 0)

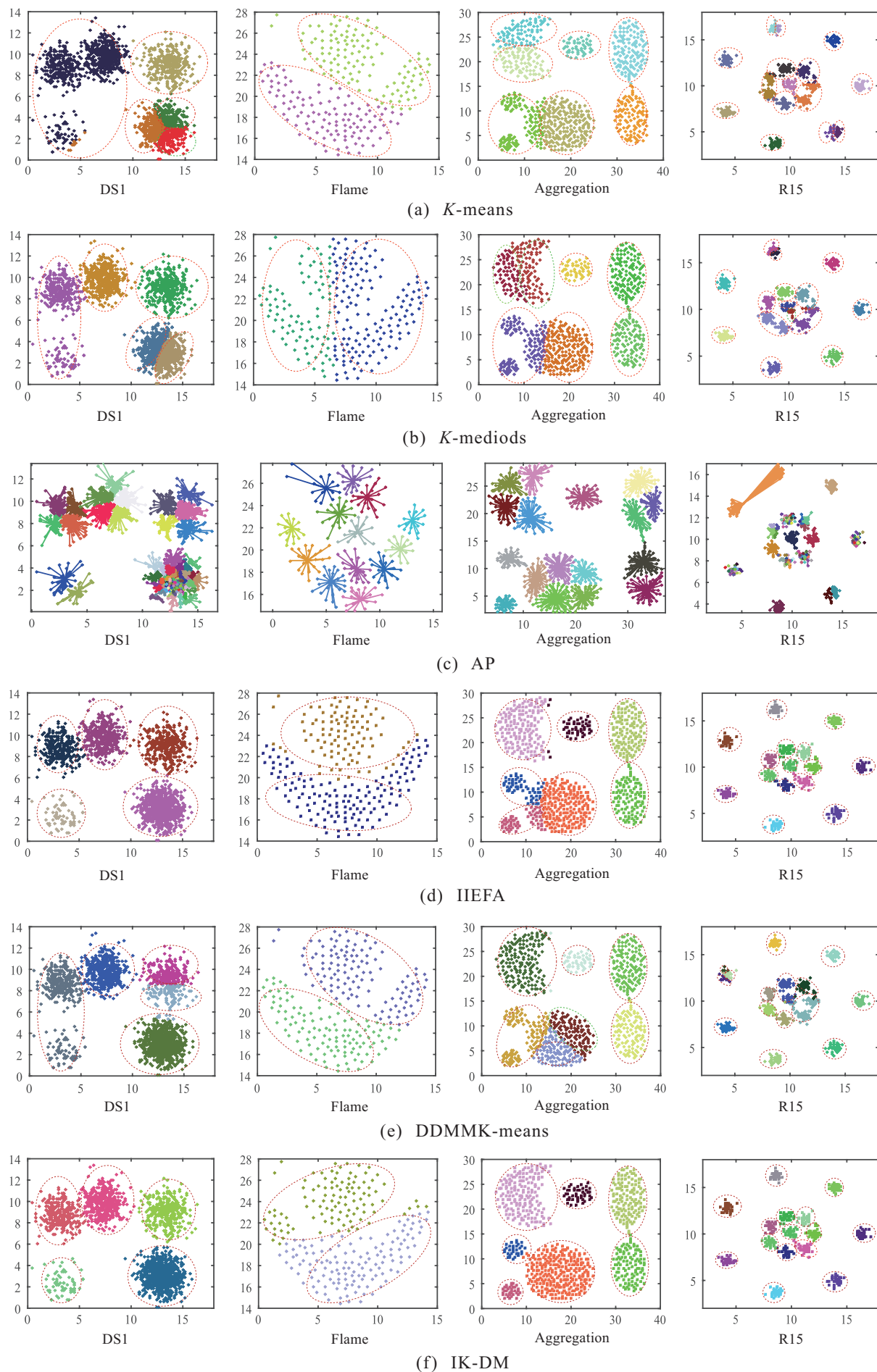


图1 聚类效果比较

3) IK-DM 算法在处理簇间密度差别较大的 Flame 数据集时,算法有一定的局限性. 主要是因为 IK-DM 算法是基于划分的聚类算法,聚类过程根据数据点与聚类中心的相似性原则将数据对象分配给相应的聚类中心形成簇,这一机制决定了 IK-DM 算法处理非球形簇时存在一定的局限性.

综上,本文所提出的 IK-DM 算法在4个人工合成

的数据集中有3个均达到最佳聚类效果,并且 IK-DM 算法可以通过自身的判断机制自动获取聚类中心,更具有使用价值.

2.4.2 真实数据集

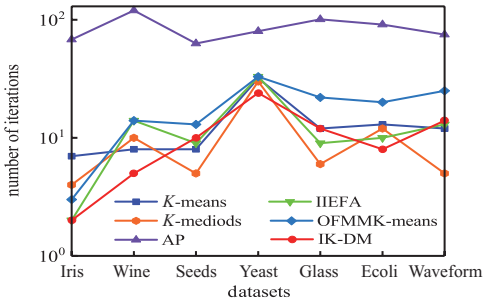
为了对 IK-DK 算法的聚类性能及实际使用价值作进一步的探讨分析,选取 UCI 中 7 个真实数据集进行验证,表 3 给出了各种算法的聚类评价指标值.

表 3 各种算法在 UCI 数据集上的性能比较

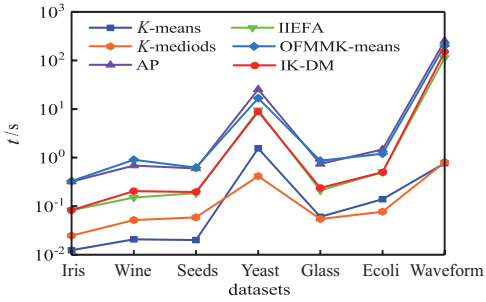
数据集	评价指标	<i>K</i> -means	<i>K</i> -medioids	AP	IIEFA	OFMMK-means	IK-DM
Iris	ACC	0.718 1 (0.105 6)	0.821 2 (0.068 2)	0.893 3 (0.000 0)	0.881 8 (0.000 0)	0.849 2 (0.044 1)	0.900 0 (0.000 0)
	purity	0.720 1 (0.107 3)	0.850 9 (0.038 0)	0.887 5 (0.000 0)	0.897 8 (0.000 0)	0.840 5 (0.051 2)	0.911 9 (0.000 0)
	NMI	0.743 1 (0.016 5)	0.758 2 (0.003 9)	0.742 7 (0.000 0)	0.741 9 (0.000 0)	0.723 5 (0.032 6)	0.766 1 (0.000 0)
Wine	ACC	0.578 7 (0.043 2)	0.719 1 (0.005 4)	0.652 1 (0.000 0)	0.731 2 (0.000 0)	0.623 8 (0.054 7)	0.707 9 (0.000 0)
	purity	0.616 4 (0.009 8)	0.725 8 (0.004 1)	0.679 8 (0.000 0)	0.707 3 (0.000 0)	0.657 8 (0.018 1)	0.725 8 (0.000 0)
	NMI	0.414 0 (0.074 5)	0.418 2 (0.015 8)	0.395 0 (0.000 0)	0.424 1 (0.000 0)	0.435 2 (0.012 5)	0.435 2 (0.000 0)
Seeds	ACC	0.882 6 (0.011 5)	0.887 5 (0.006 9)	0.893 9 (0.000 0)	0.895 2 (0.000 0)	0.887 7 (0.002 5)	0.895 2 (0.000 0)
	purity	0.887 9 (0.006 7)	0.893 5 (0.002 7)	0.895 5 (0.000 0)	0.896 0 (0.000 0)	0.892 8 (0.023 0)	0.900 1 (0.000 0)
	NMI	0.673 9 (0.010 5)	0.681 4 (0.012 4)	0.636 8 (0.000 0)	0.710 1 (0.000 0)	0.681 2 (0.013 1)	0.694 9 (0.000 0)
Yeast	ACC	0.347 7 (0.033 4)	0.326 8 (0.044 5)	0.384 2 (0.000 0)	0.357 1 (0.000 0)	0.362 5 (0.022 5)	0.396 9 (0.000 0)
	purity	0.293 2 (0.026 5)	0.299 4 (0.022 5)	0.365 8 (0.000 0)	0.314 9 (0.000 0)	0.337 4 (0.026 4)	0.378 9 (0.000 0)
	NMI	0.374 5 (0.022 9)	0.357 0 (0.036 5)	0.464 5 (0.000 0)	0.305 1 (0.000 0)	0.452 3 (0.002 2)	0.418 9 (0.000 0)
Glass	ACC	0.490 7 (0.022 1)	0.406 5 (0.085 6)	0.532 7 (0.000 0)	0.514 0 (0.000 0)	0.496 8 (0.038 2)	0.542 1 (0.000 0)
	purity	0.403 7 (0.001 2)	0.407 0 (0.006 5)	0.406 5 (0.000 0)	0.481 2 (0.000 0)	0.400 5 (0.006 5)	0.417 5 (0.000 0)
	NMI	0.286 2 (0.036 7)	0.327 4 (0.001 7)	0.512 8 (0.000 0)	0.451 2 (0.000 0)	0.512 8 (0.001 0)	0.487 9 (0.000 0)
Ecoli	ACC	0.470 2 (0.100 1)	0.506 5 (0.045 6)	0.571 5 (0.000 0)	0.586 1 (0.000 0)	0.523 6 (0.000 9)	0.574 4 (0.000 0)
	purity	0.448 3 (0.101 3)	0.517 1 (0.007 1)	0.549 8 (0.000 0)	0.555 7 (0.000 0)	0.526 5 (0.001 5)	0.557 9 (0.000 0)
	NMI	0.543 6 (0.074 5)	0.571 2 (0.029 6)	0.703 5 (0.000 0)	0.724 5 (0.000 0)	0.663 6 (0.015 7)	0.738 6 (0.000 0)
Waveform	ACC	0.504 6 (0.000 5)	0.506 4 (0.000 8)	0.510 5 (0.000 0)	0.512 8 (0.000 0)	0.496 8 (0.001 7)	0.512 8 (0.000 0)
	purity	0.499 5 (0.035 0)	0.504 5 (0.000 4)	0.511 0 (0.000 0)	0.512 6 (0.000 0)	0.503 8 (0.003 3)	0.512 6 (0.000 0)
	NMI	0.316 2 (0.019 8)	0.346 8 (0.027 4)	0.360 5 (0.000 0)	0.366 3 (0.000 0)	0.354 6 (0.007 4)	0.365 4 (0.000 0)

从表 3 的结果可以看出, *K*-mean、*K*-medioids 和 OFMMK-means 算法聚类结果波动较大,聚类效果不佳,这是因为 3 种算法在聚类之前对初始聚类中心的选取是随机的,且不考虑数据的分布情况. AP 算法在 Yeast 和 Glass 数据集上的 NMI 值最优,且不需要事先确定聚类类别数,但该算法聚类结果的聚类数往往多于真实数据集的类别数. IIEFA 算法在 Seeds 和 Waveform 数据集上表现较好. 从总体上来看,本文算法在各个数据集上的评价指标值大部分都优于与之比较的算法,其余指标均与最佳指标相差不大,如处理 Iris 和 Waveform 数据集时均优于其他算法. 因此,结合实验结果,可以验证本文算法是有效可行的,主要原因如下:

- 1) 本文算法通过启发式算法寻找最优初始聚类中心,且得到的聚类数与数据集真实的类别数相同,每次聚类结果一致,算法性能稳定;
- 2) 利用每个簇中的中位数代替 *K*-means 算法的均值进行后续聚类中心点的选取,消除了离群点对聚



(a) 运行时间(log)



(b) 迭代次数(log)

图 2 UCI 数据集上的性能对比

类结果的影响.

为了进一步比较算法的运行效率,图2给出了各算法在UCI数据集上运行30次的平均时间和平均迭代次数,纵坐标为取对数后的值.

从图2可以看出,各算法的运行时间与数据集的数量和维度成正相关,迭代次数与算法的初始聚类中心相关.在数据量较大时,AP算法的运行时间最长、迭代次数最多.这是因为AP算法需要构造相似度矩阵,并不断迭代更新吸引度和归属度的值以确定聚类中心点,迭代所消耗的时间取决于算法的迭代次数,整个算法时间复杂度达到 $O(n^2\log^n)$.OFMMK-means算法利用最大最小算法来选取初始中心点,不必计算每个数据点的相似度,但是,OFMMK-means算法的聚类时间主要与样本数和迭代次数有关,其变化趋势与迭代次数的变化趋势趋于一致.IIEFA算法基于萤火虫算法来优化初始聚类中心,如果待处理的数据量较大,则依然需要花费大量时间.而随机选取

初始聚类中心的 K -means和 K -mediods算法,运行速度较快,但聚类结果不稳定,准确率较低,选取的聚类中心较差,会增加算法的迭代次数和运行时间.相比之下,IK-DM算法根据均值相异性最大值来确定初始聚类中心,在多数情况下,算法容易收敛且迭代次数最少.但是,在选取初始聚类中心的过程中需要构造相异性矩阵,随着数据量的增大,矩阵规模快速增加,因此,运行时间比传统的 K -means和 K -mediods算法高.

2.5 鲁棒性分析

由于AP算法和IIEFA算法的聚类结果具有稳定性,这里只给出 K -means、 K -mediods、OFMMK-means和IK-DM算法在Iris和Waveform数据集上的详细情况.如表4所示,4种算法分别运行6次,给出了算法每次运行时所选取的初始聚类中心和初始中心所对应的真实类别,以此来检验算法的鲁棒性.为了便于表示,用数据集中样本点的序号表示初始聚类中心.

表4 4种算法选取初始聚类中心的结果对比

数据集	序号	K -means		K -mediods		OFMMK-means		IK-DM	
		初始中心	真实类别	初始中心	真实类别	初始中心	真实类别	初始中心	真实类别
Iris	1	(10, 65, 123)	(1, 2, 3)	(60, 92, 122)	(2, 2, 3)	(5, 47, 134)	(1, 1, 3)	(14, 61, 119)	(1, 2, 3)
	2	(39, 88, 134)	(1, 2, 3)	(24, 78, 83)	(1, 2, 2)	(47, 99, 105)	(1, 2, 3)	(14, 61, 119)	(1, 2, 3)
	3	(76, 92, 122)	(2, 2, 3)	(64, 85, 104)	(2, 2, 3)	(5, 77, 135)	(1, 2, 3)	(14, 61, 119)	(1, 2, 3)
	4	(31, 68, 80)	(1, 2, 2)	(54, 109, 126)	(2, 3, 3)	(11, 24, 107)	(1, 1, 3)	(14, 61, 119)	(1, 2, 3)
	5	(65, 92, 144)	(2, 2, 3)	(58, 69, 115)	(2, 2, 3)	(87, 91, 125)	(2, 2, 3)	(14, 61, 119)	(1, 2, 3)
	6	(52, 105, 108)	(2, 3, 3)	(105, 117, 142)	(3, 3, 3)	(34, 55, 142)	(1, 2, 3)	(14, 61, 119)	(1, 2, 3)
Waveform	1	(1 950, 547, 2 954)	(1, 3, 3)	(2 601, 4 319, 489)	(1, 1, 2)	(6, 593, 3 619)	(2, 2, 3)	(254, 2 214, 3 819)	(1, 2, 3)
	2	(79, 4 317, 4 171)	(1, 2, 3)	(4 541, 540, 2 584)	(1, 1, 2)	(20, 521, 873)	(1, 2, 3)	(254, 2 214, 3 819)	(1, 2, 3)
	3	(3 345, 391, 2 501)	(1, 2, 2)	(23, 716, 2 797)	(1, 2, 2)	(7, 2 887, 4 138)	(1, 1, 3)	(254, 2 214, 3 819)	(1, 2, 3)
	4	(611, 2 858, 1 090)	(1, 2, 3)	(504, 2 927, 2 539)	(1, 1, 2)	(23, 636, 1 108)	(1, 2, 3)	(254, 2 214, 3 819)	(1, 2, 3)
	5	(3 356, 2 998, 280)	(2, 3, 3)	(3 815, 415, 3 307)	(1, 1, 3)	(586, 32, 3 307)	(2, 2, 3)	(254, 2 214, 3 819)	(1, 2, 3)
	6	(99, 282, 763)	(1, 1, 1)	(2 953, 2 203, 10)	(1, 2, 3)	(4 996, 2 285, 1)	(1, 2, 3)	(254, 2 214, 3 819)	(1, 2, 3)

从表4的结果可以看出, K -means与 K -mediods算法都是随机选取初始聚类中心,在大多数情况下,选取的聚类中心可能位于同一个簇,导致聚类中心分布不均匀,算法迭代次数增加.OFMMK-means算法是从第2个初始聚类中心的选取才开始使用最大最小值方法,而第1个聚类中心的选取依然是随机的,导致每次选取的聚类中心不一致,且第2个聚类中心点与第1个聚类中心点可能属于同一个簇,导致聚类结果不稳定.而IK-DM算法在两个数据集上每次运行所选取的初始聚类中心都完全一致,并且每个聚类中心点与真实类别的每个簇相对应,从而降低了算法的迭代次数,因此,算法具有很好的鲁棒性.

3 结 论

针对 K -means算法的聚类结果受离群点影响较大,且无法正确选取初始聚类中心的问题,本文提出了一种自行确定初始聚类中心改进的 K -means算法,引入数据点的相异性信息,根据总体相异性衡量各数据点是否可以作为初始聚类中心,并利用簇中数据点的中位数代替均值进行后续聚类中心的迭代.在各个数据集上的实验结果表明,改进后的算法能够改善离群点对聚类结果的影响,算法的稳定性和准确率得到显著提高,迭代次数减少,对真实数据集的聚类性能优异.但是,本文算法在运行过程中所消耗的时间略微大于 K -means和 K -mediods算法,如何提升算法

的执行速度是今后研究的重要工作.

参考文献(References)

- [1] Kacprzyk J, Pedrycz W. Springer handbook of computational intelligence[M]. Berlin: Springer Publishing Company, 2015: 578-600.
- [2] Li X L, Han Q, Qiu B Z. A clustering algorithm using skewness-based boundary detection[J]. Neurocomputing, 2018, 275: 618-626.
- [3] Chen H Z, Wang W W, Feng X C, et al. Discriminative and coherent subspace clustering[J]. Neurocomputing, 2018, 284: 177-186.
- [4] Wu J J, Liu H F, Xiong H, et al. K -means-based consensus clustering: A unified view[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(1): 155-169.
- [5] MacQueen J B. Some methods for classification and analysis of multi-variate observations[C]. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281-297.
- [6] Zhou J, Pan Y Q, Chen C L P, et al. K -medoids method based on divergence for uncertain data clustering[C]. Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics. Budapest: IEEE, 2016: 2671-2674.
- [7] Fery B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.
- [8] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996, 96(34): 226-231.
- [9] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure[C]. Proceedings of the ACM SIGMOD International Conference on Management of Data. Pennsylvania: ACM Press, 1999, 28(2): 49-60.
- [10] Sheikholeslami G, Chatterjee S, Zhang A. Wavecluster: A multi-resolution clustering approach for very large spatial databases[C]. Proceedings of the 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann Publishers Inc, 1998, 98: 428-439.
- [11] Wang W, Yang J, Muntz R R. STING: A statistical information grid approach to spatial data mining[C]. Proceedings of the 23rd International Conference on Very Large Data Bases. Athens: Morgan Kaufmann Publishers Inc, 1997, 97: 186-195.
- [12] 唐东凯, 王红梅, 胡明, 等. 优化初始聚类中心的改进 K -means 算法[J]. 小型微型计算机系统, 2018, 39(8): 1819-1823.
(Tang D K, Wang H M, Hu M, et al. Optimizing initial cluster center of improved K -means algorithm[J]. Journal of Chinese Computer Systems, 2018, 39(8): 1819-1823.)
- [13] 李武, 赵娇燕, 严太山. 基于平均差异度优选初始聚类中心的改进 K -均值聚类算法[J]. 控制与决策, 2017, 32(4): 759-762.
(Li W, Zhao J Y, Yan T S. Improved K -means clustering algorithm optimizing initial clustering centers based on average difference degree[J]. Control and Decision, 2017, 32(4): 759-762.)
- [14] Yang M S, Wu K L. A similarity-based robust clustering method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(4): 434-448.
- [15] Xie H L, Zhang L, Lim C P, et al. Improving K -means clustering with enhanced firefly algorithms[J]. Applied Soft Computing, 2019, 84: 105763.
- [16] 何熊熊, 管俊轶, 叶宣佐, 等. 一种基于密度和网格的簇心可确定聚类算法[J]. 控制与决策, 2017, 32(5): 913-919.
(He X X, Guan J Y, Ye X Z, et al. A density-based and grid-based cluster centers determination clustering algorithm[J]. Control and Decision, 2017, 32(5): 913-919.)
- [17] 于彦伟, 贾召飞, 曹磊, 等. 面向位置大数据的快速密度聚类算法[J]. 软件学报, 2018, 29(8): 2470-2484.
(Yu Y W, Jia Z F, Cao L, et al. Fast density-based clustering algorithm for location big data[J]. Journal of Software, 2018, 29(8): 2470-2484.)
- [18] Lichman M. UCI machine learning repository[EB/OL]. (1998-08-16)[2020-08-05]. <http://archive.ics.uci.edu/ml/datasets>.
- [19] Nie F P, Wang C L, Li X L. K -multiple-means: A multiple-means clustering method with specified K clusters[C]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage: ACM Press, 2019: 959-967.
- [20] Chen W Y, Song Y Q, Bai H J, et al. Parallel spectral clustering in distributed systems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 568-586.
- [21] Zhang X L, Wang W, Nørnvåg K, et al. K -AP: Generating specified K clusters by efficient affinity propagation[C]. Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney: IEEE Press, 2010: 1187-1192.
- [22] Wang Y T, Chen L H. K -MEAP: Multiple exemplars affinity propagation with specified K clusters[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(12): 2670-2682.

作者简介

廖纪勇(1996—), 男, 硕士生, 从事数据挖掘及机器学习的研究, E-mail: 1255380612@qq.com;

吴晟(1960—), 男, 教授, 从事计算机软件技术及算法设计等研究, E-mail: ws8146@163.com;

刘爱莲(1969—), 女, 副教授, 从事光纤通信技术及宽带通信技术等研究, E-mail: 1004039241@qq.com.

(责任编辑: 李君玲)