基于残差分析的混合属性数据聚类算法

邱保志 [1] 张瑞霖 [1] 李向丽 [1]

摘 要 针对混合属性数据聚类结果精度不高、聚类结果对参数敏感等问题, 提出了基于残差分析的混合属性数据聚类算法 (Clustering algorithm for mixed data based on residual analysis) RA-Clust. 算法以改进的熵权重混合属性相似性度量对象间的相似性, 以提出的基于 KNN 和 Parzen 窗的局部密度计算方法计算每个对象的密度, 通过线性回归和残差分析进行聚类中心预选取, 然后以提出的聚类中心目标优化模型确定真正的聚类中心, 最后将其他数据对象按照距离高密度对象的最小距离划分到相应的簇中, 形成最终聚类. 在合成数据集和 UCI 数据集上的实验结果验证了算法的有效性. 与同类算法相比, RA-Clust 具有较高的聚类精度.

关键词 聚类, 残差分析, 线性回归, 混合属性数据集, 聚类中心

引用格式 邱保志, 张瑞霖, 李向丽. 基于残差分析的混合属性数据聚类算法. 自动化学报, 2020, **46**(7): 1420−1432

DOI 10.16383/j.aas.2018.c180030

Clustering Algorithm for Mixed Data Based on Residual Analysis

QIU Bao-Zhi[1] ZHANG Rui-Lin[1] LI Xiang-Li[1]

Abstract For the existing mixed data clustering algorithm, there are some problems such as low clustering accuracy and parameters sensitive, a clustering algorithm for mixed data based on residual analysis (RA-Clust) is proposed. We use entropy weight to measure the similarity between objects with mixed attributes. Based on KNN and Parzen windows, we propose a method to calculate the local density of objects. Pre-selected cluster centers is conducted by linear regression and residual analysis. Then, the true cluster centers are selected according to objective optimization model proposed in this paper. Finally, the remaining objects are assigned into corresponding clusters according to the minimum distance from the high density objects. The experimental results on synthetic datasets and UCI datasets verify the effectiveness. Compared with similar algorithms, RA-Clust has a higher clustering accuracy.

Key words Clustering, residual analysis, linear regression, mixed data, cluster center

Citation Qiu Bao-Zhi, Zhang Rui-Lin, Li Xiang-Li. Clustering algorithm for mixed data based on residual analysis. *Acta Automatica Sinica*, 2020, **46**(7): 1420−1432

聚类分析在医学、图像分割、生物学、电子商务、互联网等领域得到了广泛应用[1−6]. 在实际应用环境中, 被聚类的数据通常含有数值属性和分类属性[7], 例如医学检测报告不仅有血压、脉搏等数值属性, 而且还存在性别、婚姻状况、疾病定性检验结果等分类属性[8]. 现有聚类算法大部分只能对数值属性数据聚类, 不能对混合属性数据聚类, 例如 K-means[9]、FCM[10]、DBSCAN[11]、DPC[12]、CLUB[13] 等. 为了解决混合属性数据聚类问题, 学者们提出了一些混合属性数据聚类算法, 例如 K-prototypes[14]、EKP[15]、IKP-MD[16]、FKP-MD[17]、DP-MD-FN[18]. 但是, 在没有先验知识的情况下, 这些算法难于确定聚类个数和选取合适的聚类中心. 造成聚类精度不高, 如何准确地确定聚类中心和聚类个数成为混合属性数据聚类领域亟待解决的问题.

以 K-means 为代表的基于划分的聚类算法以选取的初始聚类中心为基础, 依据相似性将其余的对象分配给相应的中心形成聚类, 通过反复计算新的聚类中心和再分配, 直至目标函数收敛为止, 形成最终聚类. 这种聚类中心计算方式和分配机制决定了这一类算法不能有效地处理非球形簇. 以 DBSCAN 为代表的基于密度的聚类是以任意一个核心点为中心, 将该核心密度可达的对象看作一个聚类, 由于聚类中心选取的随意性和核心点定义是基于邻域的原因, 决定了这一类算法不能有效处理高维数据和多密度聚类. DPC 算法以局部密度峰值点为中心, 中心周围低密度点的集合形成了以该峰值点为中心的一个聚类, 其结构简单、易于理解. 但是聚类中心的选取需要人为参与, 在缺少先验知识的情况下, 算法的参数难以确定. CLUB 算法通过 k

收稿日期 2018-01-12 录用日期 2018-04-16
Manuscript received January 12, 2018; accepted April 16, 2018
河南省基础与前沿技术研究项目 (152300410191) 资助
Supported by Basic and Advanced Technology Research Project of Henan Province (152300410191)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 郑州大学信息工程学院 郑州 450001
1. School of Information Engineering, Zhengzhou University, Zhengzhou 450001

近邻建立密度骨架, 并以密度骨架作为聚类的中心, 将未标记的数据对象划归到距离最近的高密度对象所在的簇中形成聚类, 可以有效地解决桥接和同一个簇中出现多个聚类中心而导致错误划分的问题. 但该算法只能处理数值属性数据集, 并不能对混合属性数据进行聚类.

为了解决混合属性数据聚类问题, K-prototypes、EKP、IKP-MD、FKP-MD 等算法以 K-means 聚类思想为基础, 通过重新定义相似性度量和改进聚类中心计算方式进行聚类, 解决了 K-means 不能对混合属性数据进行聚类的问题. DP-MD-FN 算法应用密度峰值技术对混合属性进行聚类, 解决了聚类中心的自动获取问题. 但是, 对混合属性数据聚类来说, 如何降低聚类中心提取的偏差和更真实地反映数据的分布情况是提高聚类精度的关键问题.

为了解决上述问题, 本文在改进的混合属性相似性度量方式[18] 的基础上, 提出了基于 K-nearest neighbor (KNN) 和 Parzen 窗[19] 的局部密度计算方法, 依据提出的目标优化模型自动选取正确的聚类中心, 通过划分形成聚类. 论文创新点如下:

1) 改进了混合属性数据的相似性度量;

2) 提出了一种基于 KNN 和 Parzen 窗的局部密度计算方法;

3) 提出了基于线性回归和残差分析的聚类中心预选取机制及聚类中心目标优化模型.

本文组织如下: 第 1 节给出了相似性度量方式、局部密度计算方法、聚类中心预选取机制和聚类中心目标优化模型, 给出了 RA-Clust 算法; 第 2 节给出了实验结果及其分析; 全文的总结在第 3 节给出.

1　RA-Clust 算法

1.1　相关定义

设混合属性数据集 D 含有 m 个属性, 其中 m_c、m_r 分别是分类属性和数值属性个数, 即 $m = m_c + m_r$. $Dom(A_i)$ 是分类属性 A_i 的取值集合, $Dom(A_i) = \{a_{i,1}, a_{i,2}, a_{i,3}, \cdots, a_{i,f}\}$, 表示分类属性 A_i 具有 f 个不同的取值.

定义 1 (支持度). 设 A_i 为分类属性, 属性值 $a_{i,j}$ ($a_{i,j} \in Dom(A_i)$) 关于属性 A_i 的支持度是数据集 D 中属性 A_i 取值等于 $a_{i,j}$ 的数据对象的个数, 即:

$$Sup(A_i | a_{i,j}) = |\{x | x \in D, x_i = a_{i,j}\}|, \quad 1 \le j \le f \;\; 1 \le i \le m_c \tag{1}$$

其中, x_i 表示数据对象 x 在分类属性 A_i 上的取值.

定义 2 (数值相似度). 设 $x, y \in D$, x 与 y 的数值属性的相似度定义为:

$$S_r(x,y) = \exp\left(-\frac{dist_r(x,y)^2}{2}\right) \tag{2}$$

其中, $dist_r(x,y)$ 表示 x 与 y 数值属性部分的欧氏距离. 数值属性相似度反映了对象之间在数值属性上的相似程度, 其取值区间为 [0,1].

定义 3 (分类相似度). 设 $x, y \in D$, x 与 y 的分类属性的相似度定义为:

$$S_c(x,y) = \sum_{i=1}^{m_c} W_i \theta(x_i, y_i), \quad \theta(x_i, y_i) = \begin{cases} 1, & x_i = y_i \\ 0, & x_i \ne y_i \end{cases} \tag{3}$$

其中, W_i 表示分类属性 A_i 所占的权重, 即: $W_i = H_c(A_i) / \sum_{i=1}^{m_c} H_c(A_i)$, $H_c(A_i)$ 表示分类属性 A_i 的熵, 即:

$$H_c(A_i) = -\sum_{\substack{j=1 \\ a_{i,j} \in Dom(A_i)}}^{f} p(a_{i,j}) \log_2(p(a_{i,j})) \tag{4}$$

其中, $P(a_{i,j})$ 表示分类属性 A_i 中的属性值 $a_{i,j}$ 的概率, 即: $P(a_{i,j}) = Sup(A_i | a_{i,j}) / |D|$

分类属性相似度反映了对象之间在分类属性上的相似程度, 其取值区间为 [0, 1].

不同数据集中的数值与分类属性个数是不同的, 本文将分类属性、数值属性占总属性的比例作为属性权重, 用来计算对象间的相似性.

定义 4 (对象相似性). 假设 $x, y \in D$, x 与 y 的相似性定义为:

$$S(x,y) = \frac{m_c}{m} S_c(x,y) + \frac{m_r}{m} S_r(x,y) \tag{5}$$

KNN 是一种有效度量对象周围分布的方法. 为了合理的表征对象的密度, 本文将 Parzen 窗密度估计与 KNN 相结合来度量一个对象的局部密度.

定义 5 (局部密度). 设 $x \in D$, 数据对象 x 的局部密度为 $den(x)$, 其定义如下:

$$den(x) = \sum_{y \in KNN(x)} \frac{\exp(-dist(x,y))}{k} \tag{6}$$

其中, $KNN(x)$ 为对象 x 的 k 近邻集合, $dist(x,y) = 1 - S(x,y)$ 为对象 x 与 y 的距离度量. 一个对象的 k 近邻距离越小, 该对象周围分布越稠密, 说明该对象的局部密度越大.

1.2　聚类中心的选取与优化

正确的聚类中心是提高聚类质量的关键, 错误地选取聚类中心可能造成错误的聚类结果或较差的

聚类精度[20−21]. 本文在文献 [22] 的启示下, 首先运用改进的残差分析和线性回归进行聚类中心预选取, 然后使用聚类中心目标优化模型确定聚类中心.

1) 聚类中心预选取: 假设随机变量 X 符合正态分布, $X \sim \mathrm{N}(\mu, \sigma^2)$, 利用统计学知识进行变换, 可以得出 $P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \le \mu \le \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}\right) = 1-\alpha$. 根据统计学知识可知: 残差是独立的, 对于任意残差 e_i, $\bar{X} = e_i$, $n = 1$ 在置信度为 $1-\alpha$ 的情况下, 置信区间为:

$$\left[e_i - \sigma \times Z_{\frac{\alpha}{2}}, e_i + \sigma \times Z_{\frac{\alpha}{2}}\right] \tag{7}$$

若某点的残差没有落在置信度为 $1-\alpha$ 的置信区间中, 视该点为奇异点, 即为聚类中心.

DC-MDACC 将密度的倒数与距离 δ[22] 进行回归分析, 然后将得出的回归模型进行残差分析, 选取出不在置信区间的数据对象, 将其作为聚类中心. 根据式 (7), 残差分析中显著性水平 α 的取值对聚类中心的获取十分敏感, α 取值偏大, 则置信区间变小, 导致选取的聚类中心集合包含了部分非聚类中心点. α 取值偏小, 则置信区间偏大, 只能提取部分聚类中心点, 严重影响了聚类的质量, 且 α 的取值没有任何先验知识. DC-MDACC 为解决此问题, 将 α 设置为 0.05, 再根据聚类结果去调节参数 α, 来获得更高质量的聚类, 此种方法并不具有普适性.

为了更加明显地区分聚类中心与非聚类中心对象, 增加残差分析的分离度, 我们将数据对象的密度权重作为回归分析的参数, 定义如下:

定义 6 (密度权重). 设 $x \in D$, 数据对象 x 的密度权重为 $\gamma(x)$, 其定义如下:

$$\gamma(x) = den(x) \times \delta(x) \tag{8}$$

其中, $\delta(x)$ 表示数据对象 x 与高密度数据对象之间的最小距离, 即:

$$\delta(x) = \min_{y:den(y)>den(x)} dist(x, y) \tag{9}$$

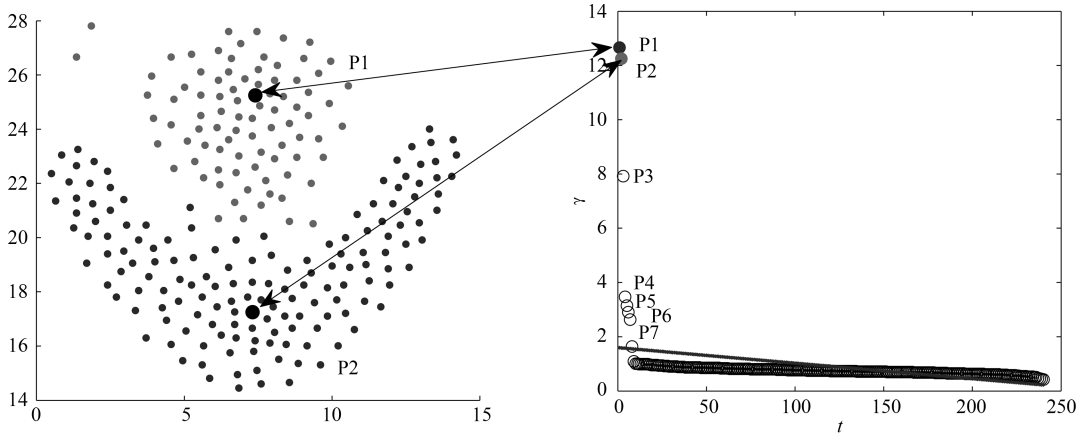
特别地, 对于局部密度最大的对象 x, 其距离为: $\delta(x) = \max_y dist\,(x, y)$.

聚类中心自身密度较大, 与其他密度更大的数据点之间存在较大的距离[12]. 根据式 (8) 可知, 聚类中心点的密度权重远比其他数据对象的密度权重值大. 则密度权重越大的数据对象, 作为聚类中心的可能性就越大. 依据密度权重进行降序排序后, 聚类中心存在于前半部分.

聚类中心拥有较大的密度权重, 对密度权重进行回归分析, 聚类中心将远离线性回归方程. 通过残差分析去除靠近线性回归方程的数据对象, 保留密度权重较大并且远离线性回归方程的数据对象, 形成聚类中心预选集 P_Set.

以合成数据集 Flame 为例, 计算数据集中每个数据对象的局部密度 den 和 δ 距离, 得到其密度权重. 图 1 (a) 是数据集的分布, 图 1 (b) 是数据对象的密度权重降序排序.

其中, P1、P2 是数据集 Flame 的聚类中心, 根据图 1 中的映射关系得知, P1 和 P2 具有最高的密度权重, 可以作为簇中心点. P3、P4、P5、P6、P7 的密度权重也较大, 其分布在 P1 与 P2 点的周围, 其他数据点的密度权重分布较为集中. 我们对每个数据对象的密度权重 γ 和递增的整型自变量 t 进行一元线性拟合. 线性回归方程的形式化定义如下: $T = a + bt + e$.



(a) 数据对象分布
(a) Distribution of data objects

(b) 密度权重的降序排序
(b) Descending order of density weights

图 1 Flame 数据集的密度权重分布 (降序)

Fig. 1 Density weight distribution of Flame (descending)

根据密度权重 γ 和递增的整型自变量 t 得出拟合曲线 $\gamma = f(t)\quad t \in \{1, 2, 3, \cdots, |D|\}$, 如图 1(b) 中直线所示. 对拟合出的函数关系 $\gamma = f(t)$ 进行残差分析 ($\alpha = 0.7$), 残差图如图 2.

图 2 中有 7 个数据对象 P1∼P7 没有落在残差的置信区间中, 在本算法中被视为奇异点, 形成聚类中心预选集 P_Set, 其中密度权重最高的两个点 P1、P2 是正确的簇中心. 为了得到正确的聚类个数并筛选出最具代表性的聚类中心, 我们将显著性水平 α 设置的尽可能大, 其设置为固定值 0.1[22], 确保置信区间足够窄, 得到更大的预选集.

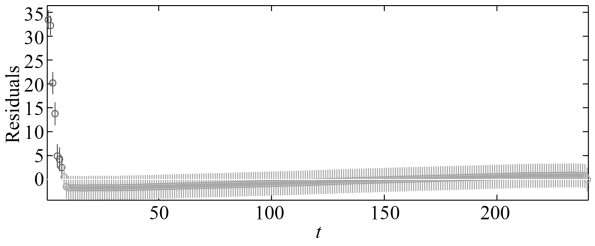


图 2　Flame 数据集的残差分布图 (降序)

Fig. 2　The residual distribution graph of Flame (descending)

2) 聚类中心优化: 聚类中心预选集 P_Set 包含了真实的聚类中心点和一些非聚类中心点 (聚类中心附近的点, 这些点同样拥有较大的密度权重), 下一步需要从预选集中选出最具代表性的聚类中心.

聚类结果评价指标是对聚类结果正确性的评价, 总体分为有标签评价指标和无标签评价指标, 有标签评价指标将聚类结果与真实的类标号进行对比, 从某个侧面进行评价, 如纯度 (Purity)[23]、标准化互信息 (Normalized mutual information, NMI)[18]、准确度 (Accuracy, ACC)[18], 无标签评价指标适用于缺少真实聚类信息的情况下, 对聚类结果进行度量, 如误差平方和 (Sum of the squared errors, SSE)[9]、邓恩指数 (Dunn validity index, DVI) 等[24]. 聚类是一个无监督的过程, 因此有标签评价往往用于最终聚类结果的评估. 在聚类过程中, 可以使用无标签评价指标来衡量聚类过程中的临时结果. 单从一个指标很难正确地评价聚类效果与实际的一致性. 例如 SSE 仅适用于评价球形聚类, DVI 对环状分布聚类测评较差. 为此, 本文定义一个将轮廓系数 (Silhouette coefficient, SC)、戴维森堡丁指数 (Davies-bouldin index, DBI)[24] 两个指标融合为一个综合指标 U 的聚类中心目标优化模型, 对在算法聚类过程中产生的聚类结果进行评价, 模型如下:

$$U^i = \frac{2 - SC^i + DBI^i}{2} \qquad 2 \le i \le |P\_Set| \tag{10}$$

其中, U^i 表示第 i 次迭代得到的目标优化模型的值, SC^i、DBI^i 分别表示第 i 次迭代时聚类结果的轮廓系数、戴维森堡丁指数.

聚类结果的轮廓系数为所有对象的轮廓系数的均值, 即 $\mathrm{SC} = \sum_{j \in D} SC(j) / |D|$. 对象 j 的轮廓系数计算公式如下:

$$SC(j) = \begin{cases} 1 - \frac{a(j)}{b(j)}, & a(j) < b(j) \\ 0, & a(j) = b(j) \\ \frac{b(j)}{a(j)} - 1, & a(j) > b(j) \end{cases} \qquad 2 \le j \le |P\_Set| \tag{11}$$

其中, $a(j)$ 表示对象 j 到其所属簇中其他对象距离的均值, $b(j)$ 表示对象 j 到其他簇中所有对象距离的均值的最小值. 轮廓系数的取值区间为 $[-1, 1]$, 结果越接近 1, 表示聚类结果越合理.

聚类结果的戴维森堡丁指数计算公式如下:

$$\mathrm{DBI} = \frac{1}{C\_no} \times \sum_{t=1}^{C\_no} \max_{j \ne t} \left[\frac{mean\,(C_t) + mean\,(C_j)}{dist(center\,(C_t), dist(center\,(C_i))} \right] \tag{12}$$

$mean(C_j)$ 表示类 j 中所有对象到其聚类中心距离的均值, $center(C_j)$ 表示类 j 的聚类中心, C_no 表示聚类个数, 该指数取值越小, 表示聚类结果越合理

由于 DBI 与 SC 指标取值大小意义不同, 为了其取值含义的一致性, 算法将轮廓系数指标归置为取值越小越好, 即使用轮廓系数最大值 2 减去轮廓系数.

真正的聚类中心选取过程是: 先对预选集 P_Set 中每个对象按照密度权重进行降序排序, 形成集合 P_Set_Desc. 然后从集合中第二个数据对象开始递进, 设置递进步长为 1, 在第 i 次迭代时, 将集合 P_Set_Desc 中前 i 个数据对象作为当前聚类中心, 计算当前目标优化模型的值, 直到迭代至集合中最后一个数据对象.

选择使目标优化模型达到最小值的迭代次数 I 作为最佳聚类中心个数, 并得出最佳的聚类中心. 即前 I 个数据对象作为聚类中心时, 聚类效果最好, 公式如下:

$$I = \arg\min_i (U^i) \qquad 2 \le i \le |P\_Set\_Desc| \tag{13}$$

表 1 是数据集 Flame 选取聚类中心点的过程,

表 1 目标优化模型的迭代计算过程

Table 1 Iterative calculation process of objective optimization model

	P1-P2	P1-P3	P1-P4	P1-P5	P1-P6	P1-P7	P1-P8	P1-P9	P1-P10	P1-P11	P1-P12	P1-P13
U	**1.0369**	1.2274	1.4342	1.4336	1.3781	1.3433	1.3957	1.2310	1.2207	1.1352	1.2556	1.2315
DBI	**0.4074**	0.4647	0.8507	0.6806	0.5671	0.5092	0.6260	0.4740	0.4140	0.3295	0.5415	0.4998
SC	**0.3335**	0.0099	−0.0178	−0.1866	−0.8192	−0.1775	−0.1655	0.0120	−0.0275	0.0588	0.0303	0.0367

其中, P1-P7 为中心预选集. 可以看出, 当选择 P1、P2 作为聚类中心时, 各项参数均为最优, 目标优化模型的取值最小, 符合真实的期望.

表 2 给出了在数据集 Flame 上本文的聚类中心选取策略与 DC-MDACC 的聚类中心选取策略在不同显著性水平 α 下得到的聚类中心个数的比较, 可以看出, 本算法的聚类中心选取过程不受显著性水平 α 的影响.

表 2 α 的取值与聚类中心个数

Table 2 The value of alpha and the number of cluster centers

α	0.6	0.5	0.1	0.05	0.02	0.01	0.001	0.0001
DC-MDACC	9	9	7	5	4	3	3	3
RA-Clust	2	2	2	2	2	2	2	2

1.3 算法步骤

RA-Clust 算法包含计算、聚类中心预选取、确定聚类中心和分派 4 个步骤. 算法首先计算每个数据对象的密度权重 γ, 其次使用线性回归和残差分析得出聚类中心的预选取集合 P_Set, 然后根据目标优化模型进行迭代计算, 确定最终的聚类中心, 最后将剩余的数据对象划分至对应的簇中. 算法描述如下.

算法 1. RA-Clust 算法

输入. 数据集 $data$, 近邻参数 k

输出. 数据集的聚类标签 $Label$

步骤 1. 计算:

根据式 (6)、(8)、(9) 计算数据对象的局部密度 den、距离 α、密度权重 γ;

步骤 2. 聚类中心预选取:

对密度权重 γ 进行线性回归与残差分析, 得出聚类中心预选集 P_Set;

将集合 P_Set 根据密度权重进行降序排序, 形成集合 P_Set_Desc;

步骤 3. 确定聚类中心:

for i= 2 → size(P_Set_Desc)

1) 取前 i 个对象作为当前聚类中心;

2) 根据目前的聚类中心, 将余下数据对象划分至距离最近的高密度对象所在的簇中;

3) 根据式 (10) 计算第 i 次迭代时目标优化模型的值 U^i;

End for

根据式 (13), 选择使目标优化模型达到最小值对应的数据对象作为最终的聚类中心;

步骤 4. 分派:

将剩余的数据对象划分至距离最近的高密度对象所在的簇中, 并返回聚类标签 $Label$.

2 实验结果分析

算法的实验环境: CPU 为 AMD Athlon X4750 Quad Core Processor 3.40 GHz, 内存为 4.00 GB, 操作系统为 Microsoft Windows 7, 算法编译环境为 Matlab R2014a.

实验数据集包括人工合成数据集和 UCI 机器学习数据集[25]. 详细信息见表 3, 其中 m 表示数据集的属性个数, m_r 表示数据集中数值属性的个数, m_c 表示其分类属性的个数, Class 表示数据集存在的聚类个数, Instance 表示数据集中对象个数. 其中编号 1～6 的数据集用来检测算法在数值属性上的聚类有效性, 编号 7～10 的数据集用来检测算法在分类属性上的聚类有效性, 编号 11～18 的数据集用来与现有混合属性聚类算法进行性能比较. 本文使用 ACC、Purity、NMI、RI、Fowlkes and mallows index (FMI)、Jaccard coefficient (JC)[26] 评价指标从多角度衡量聚类效果.

在对比实验中, 对于需要预先设置聚类个数的算法, 如 K-means、FCM、K-prototypes、EKP、IKP-MD 等, 将其参数设置为正确的聚类个数, 并运行 10 次, 取各聚类指标的均值作为最终的聚类效果. 对于其他算法, 设置参数的取值区间, 得到全部的聚类效果, 从中选择最优的聚类效果作为最终的结果.

实验数据集预处理方式如下:

1) 对混合属性数据集, 将数值属性放在分类属性的前面;

2) 删除数据集中含有缺失属性值的记录;

3) 将分类属性的属性值用整数进行替代, 例如对于性别属性, 1 代表男性, 2 代表女性.

2.1 数值属性数据集

本文采用的数值属性数据集包含 4 个人工合成数据集[13] 和两个 UCI 数据集: Wine、Seeds[25] 数据集. Flame 数据集中聚类呈半包围分布, 用来检测算法是否准确识别近距离的两个聚类. R15[13]

表 3 数据集的基本信息

Table 3 The basic information of the datasets

No.	Datasets	Data Sources	m	m_r	m_c	Class	Instance
1	Flame	Synthesis	2	2	0	2	240
2	R15	Synthesis	2	2	0	15	600
3	Spiral	Synthesis	2	2	0	3	312
4	Aggregation	Synthesis	2	2	0	7	788
5	Seeds	UCI	7	7	0	3	210
6	Wine	UCI	13	13	0	3	178
7	Soybean	UCI	35	0	35	4	47
8	SPECT Heart	UCI	22	0	22	2	267
9	Tic-tac-toe	UCI	10	0	10	2	958
10	Congressional Voting	UCI	16	0	16	2	435
11	Australian Credit Approval	UCI	14	6	8	2	690
12	Credit Approval	UCI	15	6	9	2	690
13	Heart Disease	UCI	13	6	7	2	303
14	German Credit	UCI	20	7	13	2	1 000
15	ZOO	UCI	16	1	15	7	101
16	Japanese Credit	UCI	15	6	9	2	690
17	Post Operative Patient	UCI	8	1	7	3	90
18	Hepatitis	UCI	19	6	13	2	155

数据集包含 15 个形状不同的聚类, 用来检测算法是否能完整识别数据集中所有聚类. Aggregation[27] 数据集共有 7 个大小不同的聚类, 并且存在聚类桥接现象, 用来检测算法能否处理存在桥接干扰的聚类. Spiral[13] 数据集共有三个流型簇, 用来检测算法是否可以识别任意形状的聚类. Wine 与 Seeds 数据集是高维数据集, 用来检测算法在真实数据下的聚类效果.

图 3 直观地显示了各算法在人工合成数据集下的聚类效果, 不同颜色代表产生的不同聚类. 表 4 给出了各算法在数值属性数据集上的聚类结果评价指标值, 表中加粗部分为该聚类指标最好的情况. 由于 FCM 算法引入模糊理论[28], 其聚类结果普遍好于 K-means. 但 FCM 算法采用的目标函数与 K-means 相似, 因此对于非球形簇, 如 Spiral 数据集, 聚类效果较差. RA-Clust、DBSCAN、CLUB 算法虽然采用不同的密度的度量方式, 但都是基于密度的聚类算法, 因此三者均可处理任意形状的聚类. 由于 DPC 和 DBSCAN 算法对输入参数的取值较为敏感, 算法的聚类结果并不稳定, 得到的聚类结果并不能代表其最佳水平. RA-Clust 算法在 6 个数据集中的 4 个上的各个评价指标都是最好的, 在 Aggregation 数据集的聚类结果与最佳聚类结果相差很小, 在 Wine 数据集上的聚类结果的 6 个评价指标中, 2 个指标最高, 说明 RA-Clust 算法在处理数值数据聚类时是有效地.

2.2 分类属性数据集

为了验证 RA-Clust 对分类属性数据聚类的有效性, 我们选取来自生物、医疗、游戏博弈、政治选举领域的数据集: Soybean、Tic-tac-toe、SPECT Heart、Congressional Voting[29−31] 进行实验, 并与其他聚类算法进行比较, 聚类结果评价指标见表 5.

由于 K-modes 算法采用了 K-means 中迭代计算的思想, 并且随机选取 k 个聚类中心, 容易造成局部最优解, 因此 K-modes 算法在各个数据集上的聚类效果并不稳定, 如 Tic-tac-toe 数据集, K-modes 的效果最差. EKP 算法融合了进化算法的思想, 使得算法可以达到全局最优, 聚类效果稳定, 在 Tic-tac-toe、Congressional Voting 数据集上效果较好. DP-MD-FN 算法对于参数较为敏感, 不同的参数导致其获取聚类中心的个数不同, SPECT Heart 数据集共有两个聚类中心, 但算法错误的获取了 3 个聚类中心, 导致其聚类精度不高. 本算法 (RA-Clust) 在各个数据集上拥有良好的聚类效果, 个别聚类指标虽未达到最佳, 但与最佳指标相差不大.

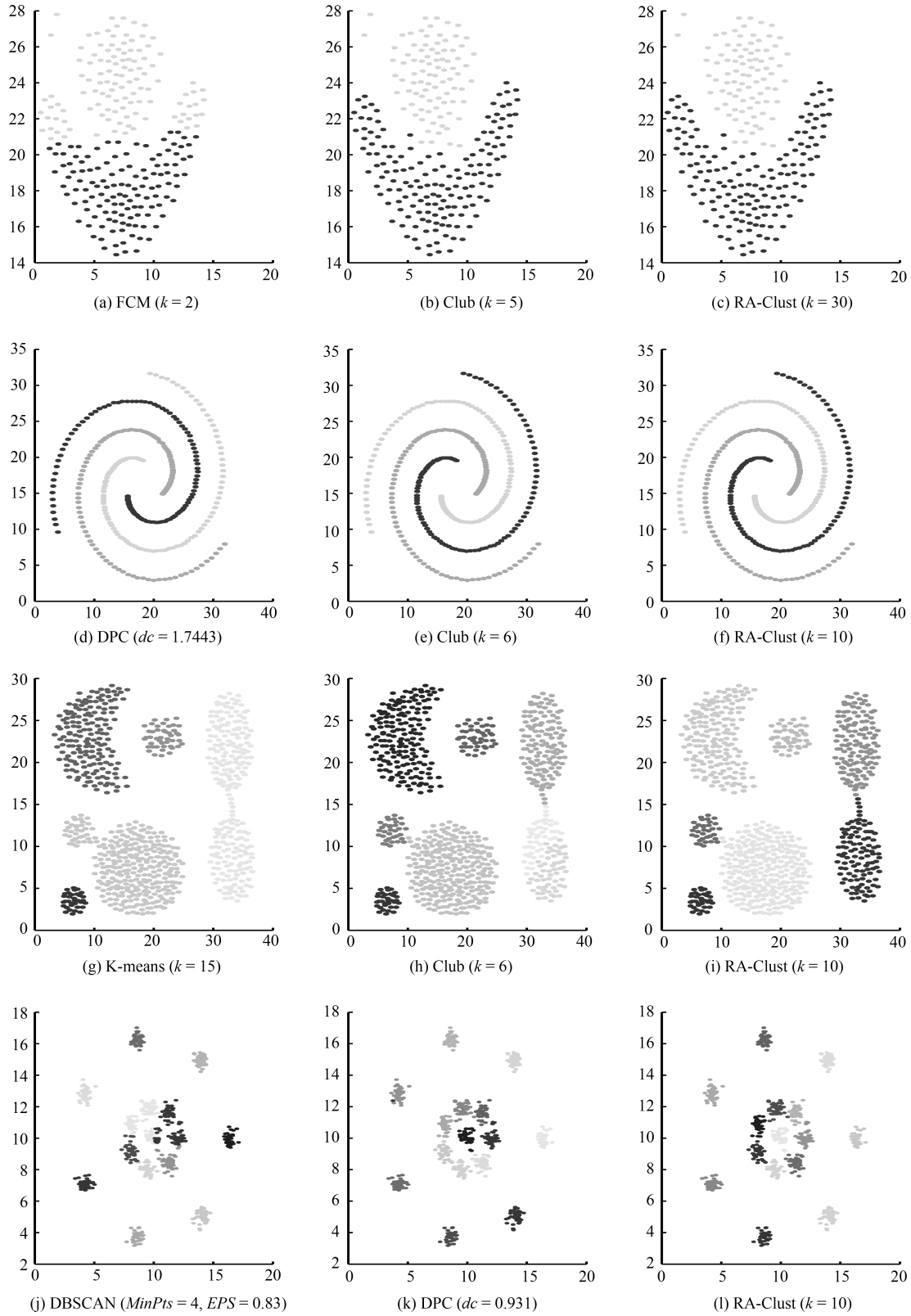


图 3 各算法在二维数据集上的聚类结果

Fig. 3 The clustering results of each algorithm on two-dimensional datasets

表 4　数值属性数据集上的聚类结果比较

Table 4　Comparison of clustering results on numerical attribute datasets

数据集	算法	参数	ACC (%)	NMI	Purity	JC	RI	FMI
Flame	K-means	$k=2$	82.9167	0.3939	0.8292	0.5684	0.7155	0.7253
	DPC	$dc=0.9301$	78.7500	0.4131	0.7875	0.5133	0.6639	0.6786
	DBSCAN	$MinPts=4$, $EPS=0.83$	94.1667	0.8448	0.9875	0.9144	0.9540	0.9561
	FCM	$k=2$	85	0.4420	0.8500	0.6032	0.7439	0.7530
	CLUB	$k=5$	**100**	**1**	**1**	**1**	**1**	**1**
	RA-Clust	$k=60$	**100**	**1**	**1**	**1**	**1**	**1**
Spiral	K-means	$k=3$	34.6154	0.00005	0.3494	0.1960	0.5540	0.3278
	DPC	$dc=1.7443$	**100**	**1**	**1**	**1**	**1**	**1**
	DBSCAN	$MinPts=10$, $EPS=1$	**100**	**1**	**1**	**1**	**1**	**1**
	FCM	$k=3$	33.9744	0.00002	0.3429	0.1956	0.5541	0.3272
	CLUB	$k=6$	**100**	**1**	**1**	**1**	**1**	**1**
	RA-Clust	$k=10$	**100**	**1**	**1**	**1**	**1**	**1**
Aggregation	K-means	$k=7$	73.3503	0.8036	0.8883	0.5676	0.8958	0.7321
	DPC	$dc=1$	94.0355	0.9705	0.9987	0.9591	0.9911	0.9793
	DBSCAN	$MinPts=4$, $EPS=0.83$	82.7411	0.8894	0.8274	**1**	**1**	**1**
	FCM	$k=7$	79.6954	0.8427	0.9315	0.6433	0.9187	0.7926
	CLUB	$k=6$	**100**	**1**	**1**	**1**	**1**	**1**
	RA-Clust	$k=12$	99.8731	0.9957	0.9987	0.9966	0.9993	0.9983
R15	K-means	$k=15$	79.5000	0.8989	0.7950	0.6075	0.9606	0.7704
	DPC	$dc=0.9500$	99.5000	0.9922	0.9950	0.9801	0.9987	0.9900
	DBSCAN	$MinPts=5$, $EPS=0.32$	78.1667	0.9121	0.7850	0.5927	0.9627	0.7642
	FCM	$k=15$	99.6667	0.9942	0.9967	0.9866	0.9991	0.9932
	CLUB	$k=7$	99.5000	0.9913	0.9950	0.9799	0.9987	0.9899
	RA-Clust	$k=10$	**100**	**1**	**1**	**1**	**1**	**1**
Seeds	K-means	$k=3$	55.2381	0.4924	0.6667	0.4430	0.7052	0.6198
	DPC	$dc=0.4$	62.06	0.6560	0.7340	0.6633	0.7125	0.7988
	DBSCAN	$MinPts=4$, $EPS=1.3$	34.2857	0.0183	0.3429	0.4964	0.7767	0.7046
	FCM	$k=3$	**89.5238**	**0.6744**	**0.8952**	0.6814	0.8743	0.8105
	CLUB	$k=24$	81.3412	0.6612	0.81314	0.6445	0.6122	0.7412
	RA-Clust	$k=9$	**89.5238**	**0.6744**	**0.8952**	**0.6815**	**0.8748**	**0.8106**
Wine	K-means	$k=3$	58.4270	0.3804	**0.7047**	0.3449	0.7032	0.5160
	DPC	$dc=0.3162$	58.43	0.2802	0.1794	0.5912	0.7016	0.6498
	DBSCAN	$MinPts=2$, $EPS=1.3$	38.2022	0.0268	0.3989	0.4864	0.7024	0.6888
	FCM	$k=3$	63.7303	0.4073	0.6373	**0.6957**	**0.9034**	**0.8206**
	CLUB	$k=24$	60.3321	0.4101	0.6033	0.6217	0.6234	0.7406
	RA-Clust	$k=25$	**64.6067**	**0.4277**	0.6461	0.6671	0.8904	0.8007

2.3　混合属性数据集

为了验证 RA-Clust 对混合属性数据聚类的有效性, 本文选取了医疗、生物、金融信贷领域产生的数据集: Heart Disease、ZOO、Post Operative Patient、Credit Approval、Australian Credit Approval、German Credit、Japanese Credit、Hepatitis[29−31] 进行实验, 并与其他算法进行比较, 聚类结果评价指标见表 6.

K-prototypes 和 EKP 算法中采用的相似性度量方式过于粗糙, 致使算法在各数据集上的聚类效果不佳. DP-MD-FN 算法采用阈值截断机制自动获取聚类中心, 其聚类结果对参数较为敏感, 由于参数的设置偏差, 可能导致算法达不到最优效果, 如在 Japanese Credit 数据集中, 真实聚类中心共有 2 个, 但算法截取了 5 个聚类中心点, 致使其聚类效果较差. FKP-MD 与 IKP-MD 算法采用了新的相似性度量和中心计算方式, 其聚类效果较好, 但算法参数较多且不易确定. 本算法在各数据集均能有效地聚

表 5 分类属性数据集上的聚类结果比较

Table 5 Comparison of clustering results on categorical attribute datasets

数据集	算法	参数	ACC (%)	NMI	Purity	JC	RI	FMI
Soybean	K-modes	$k=4$	**100**	**1**	**1**	**1**	**1**	**1**
	EKP	$k=4$, $Cp=0.8$, $Ip=0.5$	53.1915	0.2980	0.5745	0.2326	0.6947	0.3774
	FKP-MD	$Ite=100$, $k=4$, $m=1.1$	70.2128	0.7892	0.7872	0.5601	0.8205	0.7348
	IKP-MD	$Ite=100$, $k=4$, $\lambda=0.8$	**100**	**1**	**1**	**1**	**1**	**1**
	DP-MD-FN	$dc=6\,\%$, $t=5$	**100**	**1**	**1**	**1**	**1**	**1**
	RA-Clust	$k=30$	**100**	**1**	**1**	**1**	**1**	**1**
SPECT Heart	K-modes	$k=4$	60.9626	0.0697	**0.9198**	0.4963	0.5215	0.6768
	EKP	$k=2$, $Cp=0.8$, $Ip=0.5$	40.6417	0.0332	0.5241	0.4807	0.8137	0.6831
	FKP-MD	$Ite=200$, $k=2$, $m=1.4$	54.5455	0.0494	**0.9198**	0.4680	0.5015	0.6565
	IKP-MD	$Ite=200$, $k=2$, $\lambda=0.8$	67.3797	0.0568	**0.9198**	0.5398	0.5580	0.7094
	DP-MD-FN	$dc=1.5$, $t=3$	85.5615	**0.8549**	**0.9198**	0.7464	0.7491	0.8549
	RA-Clust	$k=65$	**90.3743**	0.0071	**0.9198**	**0.8245**	**0.8249**	**0.9056**
Tia-tac-toe	K-modes	$k=2$	54.6973	0.0005	0.6534	0.3669	0.5039	0.5369
	EKP	$k=4$, $Cp=0.8$, $Ip=0.5$	55.33	0.0075	0.6534	0.3560	0.5026	0.5256
	FKP-MD	$Ite=100$, $k=2$, $m=1.1$	57.0981	**0.0128**	0.6534	0.3623	0.5096	0.5324
	IKP-MD	$Ite=100$, $k=2$, $\lambda=0.8$	57.9332	0.0078	0.6534	0.3728	0.5121	0.5433
	DP-MD-FN	$dc=22.75\,\%$, $t=50$	64.3006	0.0066	0.6534	0.4883	0.5404	0.6674
	RA-Clust	$k=44$	**65.6576**	0.0067	**0.6566**	**0.5458**	**0.5486**	**0.7375**
Congressional Voting	K-modes	$k=2$	84.1379	0.4048	0.8414	0.5857	0.7325	0.7389
	EKP	$k=2$, $Cp=0.8$, $Ip=0.5$	83.6207	0.3602	0.8362	0.5678	0.7249	0.7244
	FKP-MD	$Ite=100$, $k=2$, $m=3.6$	84.9138	0.3962	0.8491	0.5902	**0.7427**	0.7423
	IKP-MD	$Ite=100$, $k=2$, $\lambda=2$	83.1897	0.3641	0.8319	0.5618	0.7191	0.7194
	DP-MD-FN	$dc=6.29\,\%$, $t=10$	80.6897	0.3802	0.8069	0.5343	0.6877	0.6966
	RA-Clust	$k=10$	**86.2069**	**0.4501**	**0.8621**	**0.7227**	0.7116	**0.7677**

类, 并且聚类指标大部分高于其他算法.

2.4 算法分析

2.4.1 复杂度分析

算法的时间复杂度是衡量算法效率的重要指标[32−33]. 本算法的计算开销主要有: 计算局部密度、计算 δ 距离、进行残差分析、迭代计算和对象划分. 本算法采用建立 KD 树[34] 的方法计算数据对象的局部密度, 时间复杂度为 $\mathrm{O}(n\log_2 n+n)$; 计算数据对象距高密度点最小距离的时间复杂度是 $\mathrm{O}(n\log_2 n+n)$, 残差分析的时间复杂度为 $\mathrm{O}(n)$[22]. 假设算法迭代次数为 $Item$, 聚类中心个数的取值范围是 $[1,\sqrt{n}]$[35], 因此 $Item$ 最坏情况下为 $\sqrt{n}$, 则算法迭代计算的时间复杂度为 $\mathrm{O}(n\sqrt{n})$. 在得到聚类中心时, 算法只需一次遍历就能完成对象划分, 时间复杂度为 $\mathrm{O}(n)$[36]. 根据上述分析, 算法的时间复杂度为 $\mathrm{O}(n\log_2 n+n+n\log_2 n+n+n\sqrt{n}+n)$, 由于 $n\sqrt{n}>n\log_2 n>n$, 所以算法的时间复杂度为 $\mathrm{O}(n\sqrt{n})$. 表 7 展示了本算法与对比算法的时间复杂度. 其中, 算法的时间复杂度优于 DBSCAN、DPC、DC-MDACC 和 DP-MD-FN 算法.

2.4.2 参数敏感性分析

RA-Clust 算法只有一个参数, 即近邻对象 k, 其参数个数少于多数聚类算法, 并易于确定. 设置参数 k 的取值区间为 $[1,100]$. 图 4 (a) 显示了算法在分类与数值属性数据集上参数 k 与聚类准确度的关系, 图 4 (b) 显示的是在混合属性数据集上参数 k 与聚类准确度的关系. 根据图 4 可知, 随着参数 k 取值的不断增加, 数据对象的采样空间不断扩展, 局部密度可以有效地表征数据的真实分布情况, 其聚类效果越来越好. 并且随着参数 k 值的增加, 聚类效果趋于平稳. 同时, 参数 k 对聚类的准确度并不敏感, 当参数 k 在 40 ∼ 80 之间时, 算法可以保持较好的聚类效果.

2.4.3 伸缩性分析

本文以 Flame 和 Soybean 数据集为基础, 将其扩充为高维度、大样本量的测试数据集, 用来检验算法的运行效率. 图 5 (a) 与图 5 (b) 显示了算法处理数值属性数据时运行时间与样本量、维度的关系. 图 5 (c) 与图 5 (d) 显示了算法处理分类属性数据时运行时间与样本量、维度的关系.

表 6 混合属性数据集上的聚类结果比较

Table 6 Comparison of clustering results on mixed datasets

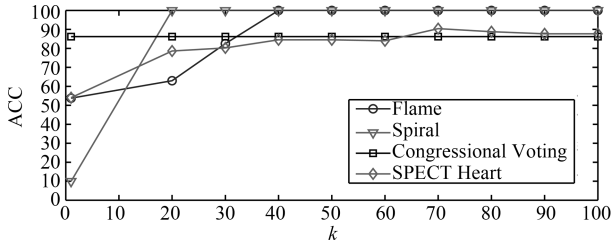
数据集	算法	参数	ACC (%)	NMI	Purity	JC	RI	FMI
Disease Heart	K-prototypes	$k=2$, $\lambda=0.4$	57.7558	0.0143	0.5776	0.3581	0.5101	0.5277
	EKP	$k=2$, $Cp=0.9$, $Ip=0.1$	52.4752	0.0065	0.5413	0.4820	0.4996	**0.6816**
	FKP-MD	$Ite=100$, $k=2$, $m=1.2$	52.4752	0	0.5413	0.3349	0.4984	0.5018
	DP-MD-FN	$dc=22\,\%$, $t=20$	75.9076	0.2018	0.7591	0.4639	0.6330	0.6338
	IKP-MD	$Ite=100$, $k=2$, $\lambda=0.8$	52.1452	0.0026	0.5413	0.3405	0.4993	0.5081
	RA-Clust	$k=30$	**77.5578**	**0.2291**	**0.7756**	**0.4858**	**0.6507**	0.6540
Credit Approval	K-prototypes	$k=2$, $\lambda=0.7$	55.2833	0.0134	0.5528	0.5015	0.5048	0.7062
	EKP	$k=2$, $Cp=0.8$, $Ip=0.5$	68.2609	0.1133	0.6826	0.4538	0.5661	0.6292
	FKP-MD	$Ite=100$, $k=2$, $m=1.3$	**83.7681**	0.3733	0.8377	0.5735	0.7277	0.7290
	DP-MD-FN	$dc=17\,\%$, $t=20$	82.2358	0.3742	0.8224	0.5522	0.7074	0.7115
	IKP-MD	$Ite=100$, $k=2$, $\lambda=0.8$	78.8406	0.2778	0.7884	0.5022	0.6659	0.6687
	RA-Clust	$k=70$	83.3078	**0.4013**	**0.8652**	**0.5827**	**0.7438**	**0.7368**
Australian Credit Approval	K-prototypes	$k=2$, $\lambda=0.4$	56.2319	0.0162	0.5623	0.5030	0.5071	0.7071
	EKP	$k=2$, $Cp=0.8$, $Ip=0.5$	55.9001	0.0048	0.5600	0.4566	0.5704	0.6317
	FKP-MD	$Ite=100$, $k=2$, $m=0.6$	55.6522	0.0034	0.5565	0.5049	0.5057	0.7101
	DP-MD-FN	$dc=18\,\%$, $t=20$	82.1739	0.3611	0.8217	0.5499	0.7066	0.7096
	IKP-MD	$Ite=100$, $k=2$, $\lambda=0.8$	81.7391	0.3105	0.8174	0.5469	0.7010	0.7072
	RA-Clust	$k=70$	**82.3188**	**0.3795**	**0.8652**	**0.5727**	**0.7400**	**0.7295**
German Credit	K-prototypes	$k=2$, $\lambda=0.15$	**67.0000**	0.0123	0.7000	0.4898	0.5580	0.6610
	EKP	$k=2$, $Cp=0.8$, $Ip=0.56$	54.1000	0.0014	0.7000	0.3865	0.5029	0.5578
	FKP-MD	$Ite=100$, $k=2$, $m=1.4$	**67.0000**	0.0096	0.7000	0.4942	0.5574	0.6658
	DP-MD-FN	$dc=1.5$, $t=3$	65.7000	0.0306	0.0716	**0.5121**	0.5704	**0.6831**
	IKP-MD	$Ite=130$, $k=2$, $\lambda=0.8$	29.0000	0.0169	0.7000	0.1860	0.4568	0.3542
	RA-Clust	$k=80$	66.3000	**0.0308**	**0.7240**	0.5050	**0.5717**	0.0752
ZOO	K-prototypes	$k=7$, $\lambda=0.6$	73.2673	0.7236	0.8416	0.5746	0.8798	0.7307
	EKP	$k=7$, $Cp=0.8$, $Ip=0.5$	61.3861	0.4641	0.7030	0.3780	0.8061	0.5504
	FKP-MD	$Ite=100$, $k=7$, $m=2.1$	83.1683	0.8689	0.4059	0.6488	0.9430	0.8055
	DP-MD-FN	$dc=7.94\,\%$, $t=11$	84.1584	0.8077	0.8416	0.8036	0.9523	0.8911
	IKP-MD	$Ite=100$, $k=7$, $\lambda=0.8$	87.1287	0.8778	**0.9307**	0.7749	0.9453	0.8760
	RA-Clust	$k=5$	**89.1089**	**0.8815**	0.8911	**0.9547**	**0.9897**	**0.9770**
Post Operative Patient	K-prototypes	$k=3$, $\lambda=0.7$	62.0690	0.0256	0.7241	0.4355	0.5354	0.6069
	EKP	$k=7$, $Cp=0.8$, $Ip=0.5$	67.7778	**0.0274**	0.7111	0.5398	0.5898	0.7131
	FKP-MD	$Ite=200$, $k=3$, $m=1.4$	53.3333	0.0231	0.7111	0.3516	0.4792	0.5210
	DP-MD-FN	$dc=81\,\%$, $t=3$	**70.1149**	0.0110	0.7126	**0.5800**	**0.5924**	**0.7572**
	IKP-MD	$Ite=150$, $k=3$, $\lambda=0.8$	41.1111	0.0228	0.7111	0.2641	0.4754	0.4340
	RA-Clust	$k=70$	**70.1149**	0.0110	**0.7326**	**0.5800**	**0.5924**	**0.7572**
Japanese Credit	K-prototypes	$k=2$, $\lambda=0.6$	55.2833	0.0134	0.5528	0.5015	0.5048	0.7062
	EKP	$k=2$, $Cp=0.8$, $Ip=0.5$	62.1746	0.0916	0.6738	0.3956	0.5594	0.5669
	FKP-MD	$Ite=100$, $k=7$, $m=2.1$	**83.3078**	0.3539	0.8331	0.5653	0.7215	0.7223
	DP-MD-FN	$dc=1.5$, $t=3$	56.9678	0.2184	0.7142	0.3657	0.5986	0.5430
	IKP-MD	$Ite=130$, $k=2$, $\lambda=0.8$	78.8668	0.2781	0.7887	0.5024	0.6661	0.6688
	RA-Clust	$k=80$	**83.3078**	**0.4013**	**0.8652**	**0.5827**	**0.7438**	**0.7368**
Hepatitis	K-prototypes	$k=2$, $\lambda=0.35$	65.0000	0.00003	0.8375	0.4794	0.5392	0.6518
	EKP	$k=2$, $Cp=0.8$, $Ip=0.5$	78.7500	0.0284	0.8375	0.6554	0.6611	0.7967
	FKP-MD	$Ite=100$, $k=7$, $m=1.3$	77.5000	0.2017	0.8375	0.5649	0.6465	0.7290
	DP-MD-FN	$dc=7.94\,\%$, $t=11$	78.7500	0.1794	0.8150	0.6541	0.7092	0.7916
	IKP-MD	$Ite=300$, $k=2$, $\lambda=0.8$	83.7500	0.2418	0.8375	0.6598	0.7244	0.7974
	RA-Clust	$k=10$	**86.2500**	**0.2847**	**0.8625**	**0.7019**	**0.7598**	**0.8262**

表 7 算法的时间复杂度分析

Table 7 The time complexity analysis of the algorithms

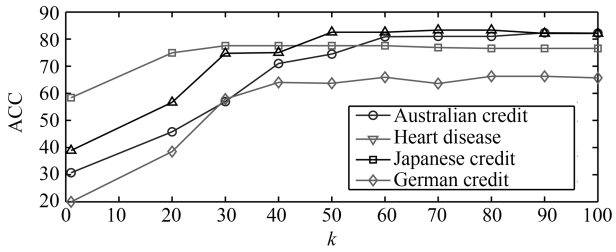
算法	时间复杂度
K-means	$O(Item \times n \times k)$[37]
FCM	$O(Item \times n \times k)$[10]
DBSCAN	$O(n^2)$[11]
DPC	$O(n^2)$[12]
CLUB	$O(n \log_2 n)$[13]
K-prototypes	$O((s+1) \times k \times n)$[17]
EKP	$O(T \times k \times n)$[15]
DC-MDACC	$O(iter \times m \times n^2)$[22]
DP-MD-FN	$O((r^2 m_c^2 + m_r^2)\, N^2)$[18]
IKP-MD	$O(k\,(m+p+Nm-Np)\, nl)$[16]
FKP-MD	$O(m^2 n + m^2 s^3 + k\,(m+p+Nm-Np)\, ns)$[17]
RA-Clust	$O(n\sqrt{n})$

从图 5 可以看出, 数据集维度的增长并没有造成运行时间的快速增加. 由于本算法采用的分类属性相似度度量方式涉及计算属性熵权, 计算部分较为复杂, 因此对于分类属性数据, 随着数据集样本量的增长, 运行时间会快速增加, 呈近指数趋势.



(a) 数值属性与分类属性数据集

(a) Numerical attribute and categorical attribute datasets

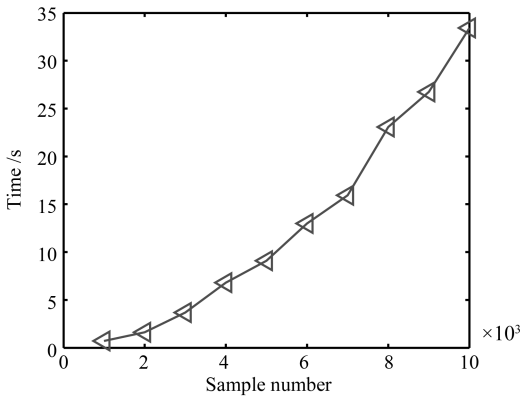


(b) 混合属性数据集

(b) Mixed datasets

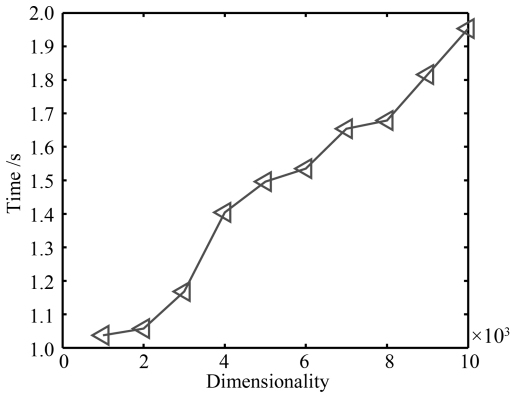
图 4 参数 k 与聚类准确度

Fig. 4 Clustering accuracy changes with parameter k



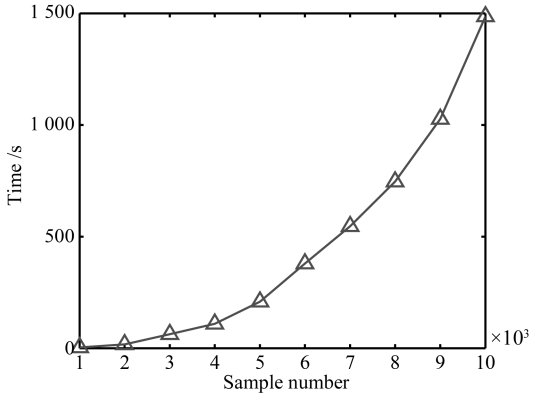
(a) 运行时间与样本量 (数值)

(a) Execution time and samples (numerical)



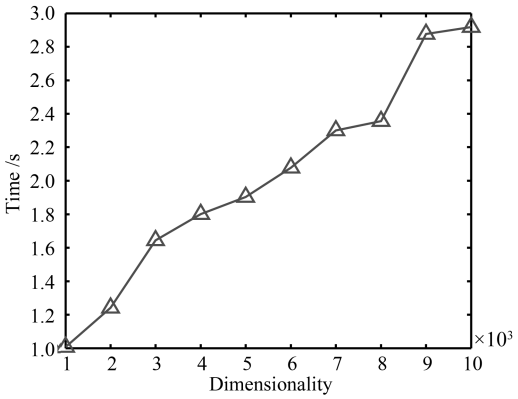
(b) 运行时间与维度 (数值)

(b) Execution time and dimensions (numerical)



(c) 运行时间与样本量 (分类)

(c) Execution time and samples (categorical)



(d) 运行时间与维度 (分类)

(d) Execution time and dimensions (categorical)

图 5 算法的运行时间与样本量、维度的关系

Fig. 5 Effect of the number of dimensions and samples on the execution time

3 结论

本文提出了基于残差分析的混合属性数据聚类算法 (RAClust). 算法改进了基于属性熵权重的混合属性相似度, 提出了基于 Parzen 窗与 KNN 的局部密度计算方法, 解决了全局密度度量失衡的问题. 通过线性回归与残差分析进行聚类中心预选取, 并依据提出的聚类中心目标优化模型对预选取的聚类中心进行优化, 解决了聚类中心获取偏差的问题. 算法只有一个参数, 易于确定, 可以对数值属性数据集、分类属性数据集和混合属性数据集聚类, 并具有较高的聚类精度.

References

1 Li X L, Han Q, Qiu B Z. A clustering algorithm using skewness-based boundary detection. *Neurocomputing*, 2018, **275**: 618−626

2 Han J W, Kamber M. *Data Mining: Concepts And Techniques*. New York: Morgan Kaufmann, 2006. 384

3 Wang Wei-Wei, Li Xaio-Ping, Feng Xiang-Chu, Wang Si-Qi. A survey on sparse subspace clustering. *Acta Automatica Sinica*, 2015, **41**(8): 1373−1384
(王卫卫, 李小平, 冯象初, 王斯琪. 稀疏子空间聚类综述. 自动化学报, 2015, **41**(8): 1373−1384)

4 Li X L, Geng P, Qiu B Z. A cluster boundary detection algorithm based on shadowed set. *Intelligent Data Analysis*, 2017, **20**(1): 29−45

5 Li Xiang-Li, Cao Xiao-Feng, Qiu Bao-Zhi. Clustering boundary pattern discovery for high dimensional space base on matrix model. *Acta Automatica Sinica*, 2017, **43**(11): 1962−1972
(李向丽, 曹晓锋, 邱保志. 基于矩阵模型的高维聚类边界模式发现. 自动化学报, 2017, **43**(11): 1962−1972)

6 Alswaitti M, Albughdadi M, Isa N A M. Density-based particle swarm optimization algorithm for data clustering. *Expert Systems with Applications*, 2018, **91**: 170−186

7 Wangchamhan T, Chiewchanwattana S, Sunat K. Efficient algorithms based on the k-means and chaotic league championship algorithm for numeric, categorical, and mixed-type data clustering. *Expert Systems with Applications*, 2017, **90**: 146−167

8 Qiu B Z, Cao X F. Clustering boundary detection for high dimensional space based on space inversion and Hopkins statistics. *Knowledge-Based Systems*, 2016, **98**: 216−225

9 Macqueen J. Some methods for classification and analysis of multiVariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: California Press, 1967. **1**(14): 281−297

10 Bezdek J C, Robert E, Full W. The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 1984, **10**(2): 191−203

11 Ester M, Kriegel H P, Xu X W, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon: Association for the Advancement of Artificial Intelligence, 1996. 226−231

12 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. *Science*, 2014, **344**(6191): 1492

13 Chen M, Li L J, Wang B, Chen J J, Pan L N, Chen X Y. Effectively clustering by finding density backbone based on knn. *Pattern Recognition*, 2016, **60**: 486−498

14 Huang Z X. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining & Knowledge Discovery*, 1998, **2**(3): 283−304

15 Zheng Z, Gong M G, Ma J J, Jiao L C, Wu Q D. Unsupervised evolutionary clustering algorithm for mixed type data Evolutionary Computation. In: Proceedings of evolutionary computation (CEC). Barcelona, Spain: IEEE, 2010. 1−8

16 Ji J C, Bai T, Zhou C G, Ma C, Wang Z. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, 2013, **120**: 590−596

17 Ji J C, Pang W, Zhou C G, Han X, Wang Z. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, 2012, **30**: 129−135

18 Ding S F, Du M J, Sun T F, Xu X, X Y. An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood. *Knowledge-Based Systems*, 2017, **133**: 294−313

19 Chen Hua, Zhang Jing, Zhang Xiao-Gang, Hu Yi-Han. A robust-elm approach based on parzen windiow's estimation for kiln sintering temperature detection. *Acta Automatica Sinica*, 2012, **38**(5): 841−849
(陈华, 章兢, 张小刚, 胡义函. 一种基于 Parzen 窗估计的鲁棒 ELM 烧结温度检测方法. 自动化学报, 2012, **38**(5): 841−849)

20 Bryant A C, Cios K J. A density-based clustering algorithm using reverse nearest neighbor density estimates. *IEEE Transactions on Knowledge & Data Engineering*, 2017, **PP**(99): 1−1

21 Carvalho F D A T D, Simões E C. Fuzzy clustering of interval-valued data with cityblock and hausdorff distances. *Neurocomputing*, 2017, **266**: 659−673

22 Chen Jin-Yin, He Hui-Hao. Research on density-based clustering algorithm for mixed data with determine cluster centers automatically. *Acta Automatica Sinica*, 2015, **41**(10): 1798−1813
(陈晋音, 何辉豪. 基于密度的聚类中心自动确定的混合属性数据聚类算法研究. 自动化学报, 2015, **41**(10): 1798−1813)

23 Aliguliyev R M. Performance evaluation of density-based clustering methods. *Information Sciences*, 2009, **179**(20): 3583−3602

24 Žalik K R, Žalik B. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, 2011, **32**(2): 221−234

25 UCI Machine Learning Repository [Online], available: http://archive.ics.uci.edu/ml/datasets.html, April 21, 2018

26 Yao H L, Zheng M M, Fang Y. Adaptive density peak clustering based on k-nearest neighbors with aggregating strategy. *Knowledge-Based Systems*, 2017, **133**: 208−220

27 Zhou Chen-Xi, Liang Xun, Qi Jin-Shan. A semi-supervised agglomerative hierarchical clustering method based on dynamically updating constraints. *Acta Automatica Sinica*, 2015, **41**(7): 1253−1263
(周晨曦, 梁循, 齐金山. 基于约束动态更新的半监督层次聚类算法. 自动化学报, 2015, **41**(7): 1253−1263)

28 Gao Jun, Sun Chang-Yin, Wang Shi-Tong. (2D)^{2}UFFCA: two-directional two-dimensional unsupervised feature extraction method with fuzzy clustering ability. *Acta Automatica Sinica*, 2012, **38**(4): 549−562
(皋军, 孙长银, 王士同. 具有模糊聚类功能的双向二维无监督特征提取方法. 自动化学报, 2012, **38**(4): 549−562)

29 Du M, Ding S, Xue Y. A novel density peaks clustering algorithm for mixed data. *Pattern Recognition Letters*, 2017, **97**: 46−53

30 Zhu S, Xu L. Many-objective fuzzy centroids clustering algorithm for categorical data. *Expert Systems with Applications*, 2018, **96**: 230−248

31 Chen J Y, He H H. A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data. *Information Sciences*, 2016, **345**: 271−293

32 Pan Z, Lei J, Zhang Y, Sun X, Kwong S. Fast motion estimation based on content property for low-complexity h.265/hevc encoder. *IEEE Transactions on Broadcasting*, 2016, **62**(3): 675−684

33 Pang Ning, Zhang Ji-Fu, Qing Xiao. A subspace clustering algorithm of categorical data using multiple attribute weights. *Acta Automatica Sinica*, 2018, **44**(3): 517−532
(庞宁, 张继福, 秦啸. 一种基于多属性权重的分类数据子空间聚类算法. 自动化学报, 2018, **44**(3): 517−532)

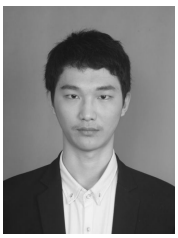
34 Redmond S J, Heneghan C. A method for initialising the k-means clustering algorithm using kd-trees. *Pattern Recognition Letters*, 2007, **28**(8): 965−973

35 Rezaee M R, Lelieveldt B P F, Reiber J H C. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters*, 1998, **19**(3-4): 237−246

36 Mehmood R, Zhang G, Bie R, Dawood H, Ahmad H. Clustering by fast search and find of density peaks via heat diffusion. *Neurocomputing*, 2016, **208**: 210−217

37 Yu S S, Chu S W, Wang C M, Chan Y K. Two improved k-means algorithms. *Applied Soft Computing*, 2017.

邱保志 郑州大学信息工程学院教授. 主要研究方向为数据库, 先进智能系统, 数据挖掘.
E-mail: iebzqiu@zzu.edu.cn
(**QIU Bao-Zhi** Professor at the School of Information Engineering, Zhengzhou University. His research interest covers database, advanced intelligent system, and data mining.)



张瑞霖 郑州大学信息工程学院硕士研究生. 主要研究方向为模式识别和数据挖掘. 本文通信作者.
E-mail: zzurlz@163.com
(**ZHANG Rui-Lin** Master student at the School of Information Engineering, Zhengzhou University. His research interest covers pattern recognition and data mining. Corresponding author of this paper.)

李向丽 郑州大学信息工程学院教授. 主要研究方向为计算机网络, 数据挖掘.
E-mail: iexlli@zzu.edu.cn
(**LI Xiang-Li** Professor at the School of Information Engineering, Zhengzhou University. Her research interest covers computer network and data mining.)