文章编号: 1001-0920 (2016) 01-0021-09 **DOI:** 10.13195/j.kzyjc.2014.1592

基于密度的划分式聚类过程参数选择算法

吴 杨, 王 韬, 李进东

(中国人民解放军军械工程学院 信息工程系，石家庄 050003)

摘 要: 为确定 K-means 等聚类算法的初始聚类中心, 首先由样本总量及其取值区间长度确定对应维上的样本密度统计区间数, 并将满足筛选条件的密度峰值所在区间内的样本均值作为候选初始聚类中心; 然后, 根据密度峰值区间在各维上的映射关系建立候选初始聚类中心关系树, 进一步采用最大最小距离算法获得初始聚类中心; 最后为确定最佳聚类数, 基于类内样本密度及类密度建立聚类有效性评估函数. 针对人工数据集及 UCI 数据集的实验结果表明了所提出算法的有效性.

关键词: 聚类算法；聚类中心；样本密度；关系树；最大最小距离

中图分类号: TP309 **文献标志码:** A

Clustering parameters selection algorithm based on density for divisional clustering process

WU Yang, *WANG Tao*, *LI Jin-dong*

(Department of Information Engineering，Ordnance Engineering College of PLA，Shijiazhuang 050003，China. Correspondent：WU Yang，E-mail：baiyanwy@163.com)

Abstract: In order to select the initial clustering centers for the divisional clustering algorithm such as the K-means algorithm, the sample density calculating regions number of each dimension is confirmed according to the samples number and their values, firstly. Then, the average value of the samples of the region with peak value satisfying the filtering conditions is taken as the candidate for the initial clustering center, and a relationship tree of the candidates is established on the mapping relations of the regions. Furthermore, the initial clustering centers are selected by using the maximal-minimal distance algorithm. To confirm the best number of the clusters, a clustering quality evaluation function is established according to the sample density and cluster density. Experiment results of the manual and UCI data sets show the effectiveness of the proposed algorithms.

Keywords: clustering algorithm；clustering center；sample density；relationship tree；maximal-minimal distance

0 引 言

聚类是数据分析领域最为重要的数据分析方法之一, 到目前为止, 尚无一种聚类算法可以普遍用于揭示各种多维数据的结构[1]. 传统聚类方法主要有: 基于划分的聚类[2]、基于层次的聚类[3]、基于网格的聚类[4]、基于密度的聚类[5]. 基于划分的聚类方法由于具有较低的实现复杂度, 是目前应用最为广泛的聚类方法之一. 基于划分的聚类算法往往需要预先指定聚类数目或聚类中心, 通过反复迭代运算, 逐步降低目标函数的误差值, 当目标函数值收敛时, 得到最终聚类结果[6], 而聚类数目和初始聚类中心的选择会对算法的执行结果产生较大的影响. 因此, 其参数选择自然地成为广大研究者关注的热点.

本文提出一种基于密度的划分式聚类过程参数选择算法, 该算法能够获得较好的初始聚类中心, 显著提高聚类算法的速度.

1 相关工作

基于划分的聚类算法其初始聚类中心的选择方法主要包括随机指定法、最大最小距离 (MMD) 算法[4]及其他改进算法. 随机指定法是以随机方式从样本中选择 k 个样本作为初始的聚类中心, 其特点是实现过程简单, 但往往无法获得全局最优解. MMD 算

收稿日期: 2014-10-19；**修回日期:** 2015-01-26.

基金项目: 国家自然科学基金项目(61173191)；军内科研项目(YJJXM12033).

作者简介: 吴杨(1985−), 男, 博士生, 从事网络协议识别技术的研究；王韬(1962−), 男, 教授, 博士生导师, 从事网络安全技术等研究.

法的基本思想是选择彼此间距离较大的样本作为初始聚类中心, 其结果虽优于随机指定法, 但易受到噪声数据的影响. Likas 等[7]针对随机指定初始聚类中心易使聚类算法收敛到局部极小的缺陷, 提出了一种全局的 K-means 算法, 算法通过迭代过程逐渐缩小初始聚类中心与真实聚类中心的距离, 但其初始参数选择的好坏会对算法的执行效率和结果产生较大的影响. 文献 [8] 为增加算法获得全局最优解的可能性, 基于可变阈值的思想, 选择较为分散的样本作为初始聚类中心, 其聚类结果优于传统的 K-means 算法. 文献 [9] 则从改进最大最小距离算法的角度出发, 将距离所有已初始化聚类中心距离最大的高密度点作为当前的聚类中心. 文献 [10] 针对 K-means 算法的初始聚类中心选择问题, 提出了采用递归调用方法寻找距离最大的样本作为初始聚类中心的方法, 在一定程度上提高了聚类的效率. Song 等[11]在确定 K-means 算法初始聚类中心的过程中, 将给定的点集看作是候选点集, 采用局部搜索算法对候选点集进行搜索. 文献 [2] 在 Song 方案的基础上, 从 k 个初始中心点的选取以及生成候选中心点集 2 个方面分别对 Song 的局部搜索算法提出了新的搜索策略, 有效降低了算法的时间复杂度, 且达到了更高的搜索精度, 而初始聚类中心的好坏会对其局部搜索算法的执行效率产生较大的影响.

最佳聚类数对获得高精度的聚类结果同样具有重要意义, 而最佳聚类数的确定主要围绕聚类有效性指标的建立而展开. 聚类有效性指标分为外部和内部有效性指标. 外部有效性指标主要是将聚类算法产生的结果与真实的划分结果进行比较分析, 以此来评估聚类的质量及算法的优劣. 常用的外部指标有: Rand、Adjusted Rand、Jaccard 和 Fowlkes-Mallows 指标[12]. 内部指标对聚类结果的评价主要依靠数据集本身和聚类结果的统计特征, 主要有: 基于数据集模糊划分的指标、基于数据集样本几何结构的指标、基于数据集统计信息的指标, 而最佳聚类数则常常根据内部指标予以确定. 常见的内部指标有: CH (Calinski-Harabasz) 指标[13]、DB (Davies-Bouldin) 指标[14]、KL (Krzanowski-Lai) 指标[15]、Wint (Weighted inter-intra) 指标[16]、IGP (In-Group Proportion)[17]等指标. CH 指标是基于全部样本的类内离差矩阵和类间离差矩阵的测度, 其最大值对应的类数即为最佳聚类数; DB 指标是基于样本的类内散度与各聚类中心间距的测度, 进行类数估计时其最小值对应的类数作为最佳聚类数; KL 指标是基于全部样本的类内离差矩阵的测度, 其最大值对应的类数作为最佳聚类数; Wint 指标的目标是最大化类内相似度和最小化类间相似度, 通常采用带罚项的指标进行类数估计, 其最大值对应的类数作为最佳聚类数; IGP 指标用来衡量在某一类中距离每个样本最近的样本是否在同一类中. 所有聚类的平均 IGP 指标越大表示聚类质量越好, 其最大值对应的类数即为最佳聚类数.

对于聚类结构难以判别的情况, 以上指标往往无法获得正确的聚类数. 针对已有聚类有效性评价指标的不足, 文献 [18] 提出了基于层次划分的聚类数评估方法 (COPS), 有效提高了确定聚类数的准确性, 其本质是度量所有数据点间的欧氏距离. 文献 [12] 在其基础上针对 K-means 算法提出了基于 BWP (Between-Within Proportion) 指标的最佳聚类数确定方法, 即通过计算某一类中的样本到其他每个类中样本平均距离的最小值及同一类中的样本到其他所有样本的平均距离来确定最佳聚类数. 以上方案的度量指标主要建立在度量样本或簇间欧氏距离的基础上, 而随着维数的增加, 样本间易出现距离趋近现象[19], 使得方法的有效性将降低.

2 背景知识

2.1 聚类的定义

聚类是将数据集 $X=(x_1,x_2,\cdots,x_3)$, 按照事先定义的规则划分到对应 k 个集合 $C_1,C_2,\cdots,C_k$ 中的过程, 而 $C_1,C_2,\cdots,C_k$ 应满足以下条件:

1) $C_i\neq\varnothing,\ i=1,2,\cdots,k$;

2) $\bigcup\limits_{i=1}^{k}C_i=X$;

3) $C_i\bigcap C_j=\varnothing,\ i,j=1,2,\cdots,k,\ i\neq j$.

2.2 最大最小距离算法

如前所述, MMD 算法的基本思想是选择彼此间距离较大的样本作为初始聚类中心. 该算法确定数据集 $X=(x_1,x_2,\cdots,x_N)$ 初始聚类中心的步骤如下.

Step 1: 以随机方式从 X 中选择一个样本作为第 1 个聚类中心 O_1.

Step 2: 从 X 中选择到 O_1 距离最大的样本作为第 2 个聚类中心 O_2.

Step 3: 计算剩余样本 x_i 分别到 O_1、O_2 的距离 d_{i1} 和 d_{i2}, 并获得它们的最小值 $d_i=\min(d_{i1},d_{i2}),\ i=1,2,\cdots,N$.

Step 4: 若 $D_t=\max\{d_i\}>a\|O_1-O_2\|$, 则将相应的样本 x_i 作为第 3 个聚类中心 O_3, 比例系数 a 用于控制聚类数量.

Step 5: 若已确定了 q 个聚类中心 ($4\leqslant q<k$), 则计算剩余样本到已有聚类中心的距离 d_{ij}. 若 $D_r=$

$\max\{\min(d_{i1}, d_{i2}, \cdots, d_{iq})\} > a\|O_1 - O_2\|$, 则将对应的样本 x_r 作为第 $q+1$ 个聚类中心.

Step 6: 重复相同的处理, 直到找到满足聚类数量的初始聚类中心.

在样本量为 N、样本维数为 h、聚类数为 k 的条件下, 最大最小距离算法在确定新的初始聚类中心时需进行 N 次距离计算, 算法的计算复杂度为 $O(Nhk)$.

2.3 最佳聚类数确定算法

根据聚类有效性指标确定最佳聚类数的算法步骤如下.

Step 1: 确定聚类数的取值范围 $[k_{\min}, k_{\max}]$.

Step 2: 依次选择 $[k_{\min}, k_{\max}]$ 中的元素作为聚类数 k, 采用聚类算法对数据集 X 进行聚类, 并更新算法的相关参数, 直至满足算法终止条件, 得到聚类输出 C^k. 由聚类指标函数 $Q(C^*)$ 计算 C^k 的聚类指标度量值 V_k.

Step 3: 根据聚类指标函数 $Q(C^*)$ 特点, 在 $(V_{k_{\min}}, V_{k_{\min}+1}, \cdots, V_{k_{\max}})$ 中寻找最佳的 V_{opt} 取值, 并由此获得最佳的聚类数 k_{opt}.

3 聚类参数选择算法

3.1 基于密度的初始聚类中心选择算法

下面将基于样本在其各维上不同样本密度统计区间内的密度, 给出一种聚类数为 k 时的初始聚类中心选择算法, 算法主要步骤如下.

Step 1: 统计并确定所有样本在各维上的取值范围 $[u_{j_{\min}}, u_{j_{\max}}]$, $1 \leqslant j \leqslant h$.

Step 2: 令 λ_1 为第 1 维上样本密度统计区间数, 统计样本在第 1 维特征参数下, 各密度统计区间内的样本密度 ($\Delta N_1(i)$ 为第 1 维上第 i 个密度统计区间内包含的样本量, Δu_1 为第 1 维上的样本密度统计区间长度), 即

$$\varphi_1(i) = \frac{\Delta N_1(i)}{\Delta u_1},\ 1 \leqslant i \leqslant \lambda_1. \tag{1}$$

Step 3: 按 $\varphi_1(i)$ 取值的大小由高到低遍历 $\varphi_1(i)$. 若 $\varphi_1(m)$ 为波峰, $\varphi_1(m)$ 不小于阈值 θ_1, $\varphi_1(m)$ 与前一密度峰间的样本密度统计区间数不小于 π_1, 则将第 m 个密度统计区间内的样本均值作为候选初始聚类中心 CoC_{1r_1}, 其中 r_1 为第 1 维上满足条件的候选初始聚类中心数.

Step 4: 若第 j 维上候选初始聚类中心已考查完毕 $(1 \leqslant j < h-1)$, 则针对第 j 维上筛选出的候选初始聚类中心, 令 λ_{j+1} 为第 $j+1$ 维上的样本密度统计区间数, 统计处于第 j 维上候选初始聚类中心所在样本密度统计区间内的样本在第 $j+1$ 维特征参数下, 各密度统计区间内的样本密度, 即

$$\varphi_{j+1}(i) = \frac{\Delta N_{j+1}(i)}{\Delta u_{j+1}},\ 1 \leqslant i \leqslant \lambda_{j+1}. \tag{2}$$

Step 5: 按 $\varphi_{j+1}(i)$ 取值的大小由高到低遍历 $\varphi_{j+1}(i)$. 若 $\varphi_{j+1}(m)$ 为波峰, $\varphi_{j+1}(m)$ 不小于阈值 θ_{j+1}, $\varphi_{j+1}(m)$ 与前一密度峰值间的密度统计区间数不小于 π_{j+1}, 则将第 m 个密度统计区间内的样本均值作为候选初始聚类中心 $\text{CoC}_{(j+1)r_{j+1}}$, 其中 r_{j+1} 为第 $j+1$ 维上满足条件的候选初始聚类中心数.

Step 6: 调整各参数, 重复执行 Step 4 和 Step 5, 直至特征向量的最大维数, 返回选定的候选初始聚类中心.

通过执行上述算法, 根据各度量值区间在各维上的映射关系, 可获得如图 1 所示的候选初始聚类中心关系树.

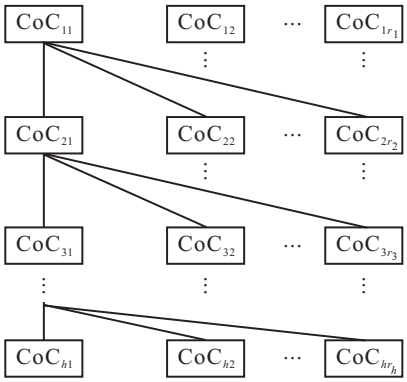

图 1　候选初始聚类中心关系树

初始聚类中心的筛选将基于最大最小距离算法的思想, 按照维数递增的方式进行分层筛选, 算法的主要步骤如下.

Step 1: 选择样本密度统计区间内密度最大的候选初始聚类中心作为第 1 个初始聚类中心 O_1;

Step 2: 从余下的候选初始聚类中心中选择距离 O_1 最远的候选初始聚类中心作为第 2 个初始聚类中心 O_2;

Step 3: 计算剩余候选值 O_i' 到已选择初始聚类中心 O_1 和 O_2 距离的最小值, 即 $d_i = \min(\|O_i' - O_1\|, \|O_i' - O_2\|)$, $i = 1, 2, \cdots, m$;

Step 4: 选择 $\max\{d_i\}$ 对应的候选初始聚类中心作为第 3 个初始聚类中心 O_3;

Step 5: 重复同样的操作, 直到找到 k 个初始聚类中心.

3.2 样本密度统计区间数的确定

鉴于类及类内样本在各维上密度的不同, 应使 λ_j 具有不同的取值, 以进一步降低方案的复杂度. 当第 j 维与 $j+1$ 维上样本的取值区间长度满足 $W_j >$

W_{j+1}时, 第j维上类的分布较为稀疏, 其样本密度统计区间数应适当地增加, 而第$j+1$维上的样本密度统计区间数则应适当地减少.

在实际的聚类过程中, k、N、h间往往满足$k \ll N$、$h \ll N$且$\sqrt[h]{N} \geqslant k$, 令$W_j \Big/ \sqrt[h]{\prod_{i=1}^{h} W_i}$为划分第$j$维上的样本密度统计区间数的微调参数, 定义第$j$维上的样本密度统计区间数为

$$\lambda_j = \frac{W_j}{\sqrt[h]{\prod_{i=1}^{h} W_i}} \sqrt[h]{N}. \tag{3}$$

3.3 候选初始聚类中心筛选参数的确定

为避免非候选初始聚类中心所在区间的密度峰值对候选初始聚类中心筛选结果的影响, 3.1 节中的算法在每一维j上设定了候选初始聚类中心筛选阈值θ_j. 在多数情况下, 类内样本在各维上的密度均大于所有样本在各维上的平均密度. 因此, 定义阈值θ_j为第j维上的样本平均密度, 即

$$\theta_j = \frac{N}{W_j}. \tag{4}$$

在实验过程中注意到: 当类间存在重叠结构时, 在对应的维上相邻的样本密度统计区间内可能会有多个密度峰值出现. 因此, 在选择密度峰值的过程中, 采用的策略是忽略处于前一密度峰值π_j个样本密度统计区间内的密度峰值, π_j的取值如下式所示:

$$\pi_j = \frac{\lambda_j - k}{2k}. \tag{5}$$

在式(5)的基础上, 给出如下定理.

定理 1 当任意两个类在第j维上不存在重叠结构时, π_j可以保证在第j维上筛选出的候选初始聚类中心包含在其各自的类中.

在给出定理 1 的证明之前, 先给出如图 2 所示的类分布示例. 其中: $y_{12} > y_{22}$, $y_{11} > y_{21}$. 定理 1 的证明过程如下.

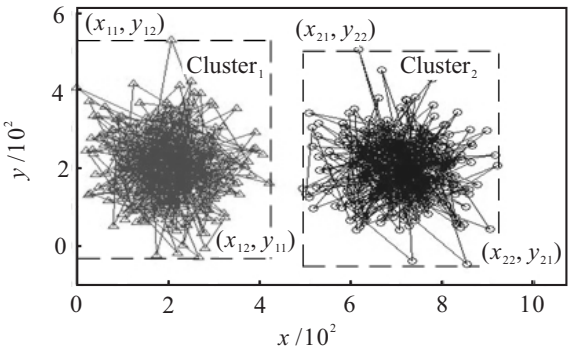

图 2 类分布示例

证明 由式(3)可得数据集在x方向上的样本密度统计区间数λ_x为

$$\lambda_x = (x_{22} - x_{11}) \sqrt{\frac{N}{(x_{22} - x_{11})(y_{12} - y_{21})}}. \tag{6}$$

在确定样本密度统计区间数λ_x后, x方向上单个样本密度统计区间的长度为

$$\Delta u_x = \frac{W_x}{\lambda_x} = \sqrt{\frac{(x_{22} - x_{11})(y_{12} - y_{21})}{N}}. \tag{7}$$

由式(5)可得

$$\pi_x = \frac{\lambda_x - 2}{4} = \frac{(x_{22} - x_{11}) \sqrt{\dfrac{N}{(x_{22} - x_{11})(y_{12} - y_{21})}} - 2}{4}. \tag{8}$$

因此, π_x对应的x方向上的样本统计区间长度Δ_x为

$$\Delta_x = \pi_x \Delta u_x = \frac{(x_{22} - x_{11}) - 2\Delta u_x}{4}. \tag{9}$$

不妨设定Cluster_1、Cluster_2在x方向上的密度峰值所在坐标O_{x1}、O_{x2}分别为

$$\begin{cases} O_{x1} = \dfrac{x_{12} + x_{11}}{2}, \\ O_{x2} = \dfrac{x_{22} + x_{21}}{2}. \end{cases} \tag{10}$$

类Cluster_1、Cluster_2的聚类中心间在x方向上的距离$\overline{W}_x$为

$$\overline{W}_x = O_{x2} - O_{x1} = \frac{x_{22} + x_{21} - x_{11} - x_{12}}{2}. \tag{11}$$

$\overline{W}_x$与Δx间的差值S_x为

$$S_x = \overline{W}_x - \Delta x = \frac{x_{22} - x_{11} + 2x_{21} - 2x_{12} + 2\Delta u_x}{4}. \tag{12}$$

当Cluster_1、Cluster_2在x方向不存在重叠部分时, 即$x_{22} > x_{11}$、$x_{21} > x_{12}$, 有$S_x > 0$. □

定理 2 当任意两个类在第j维上存在重叠的结构时, 由算法筛选出的第j维上的候选初始聚类中心均包含在对应的类中.

证明 以图 3 所示的x方向上的类Cluster_1、Cluster_2间的重叠结构分析为例(Cluster_1、Cluster_2非图 2 中的类)进行定理 2 的证明. 两类在x方向上重叠

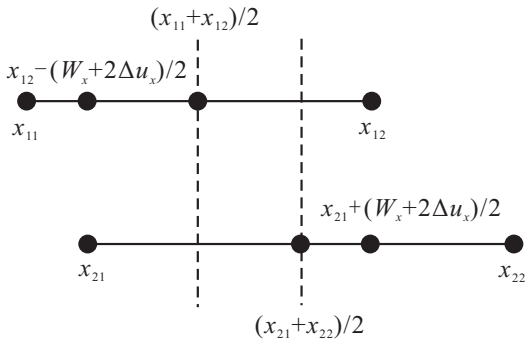

图 3 类在x方向上的重叠结构示例

部分的长度为 $(W_x + 2\Delta u_x)/2$. 因类内样本的密度围绕其聚类中心一般具有正态分布特点, 故出现密度峰值的区间一般为 $[(x_{11}+x_{12})/2, (x_{21}+x_{22})/2]$. 下面根据 Cluster$_1$、Cluster$_2$ 的聚类中心 $(x_{11}+x_{12})/2$、$(x_{21}+x_{22})/2$ 间的距离与 Δx 间的关系分几种情况进行说明.

1) 当 Cluster$_1$、Cluster$_2$ 聚类中心的距离 $|(x_{21}+x_{22})/2-(x_{11}+x_{12})/2| \leqslant \Delta x$ 时, 由算法获得初始聚类中心的个数为 1, 定理 2 成立.

2) 当 Cluster$_1$、Cluster$_2$ 聚类中心的距离 $|(x_{21}+x_{22})/2-(x_{11}+x_{12})/2| > \Delta x$ 时, 可能出现的候选初始聚类中心为:

① 候选初始聚类中心仅包含于两个类的重叠区域中, 此时候选初始聚类中心的个数为 1, 定理 2 成立;

② 候选初始聚类中心仅包含于非重叠区域中, 此时候选初始聚类中心的个数为 2, 定理 2 成立;

③ 候选初始聚类中心包含于重叠及非重叠区域中, 此时候选初始聚类中心的个数为 2 或 3, 定理 2 成立.

综上所述, 当任意两个类在第 j 维上存在重叠的结构时, 由算法筛选出的第 j 维上的候选初始聚类中心均包含在对应的类中. □

由定理 1 和定理 2 可知, 采用本文提出的候选初始聚类中心筛选算法, 可以保证筛选出的候选初始聚类中心包含在其各自的类中, 以确保进一步筛选出的初始聚类中心与真实聚类中心存在较高的相似性.

3.4　最佳聚类数的确定

鉴于目前多数聚类指标函数 $Q(C^*)$ 计算复杂度较高, 无法有效用于大规模及高维数据聚类的问题, 下面给出基于密度的用于评估 $C^k=(C_1, C_2, \cdots, C_k)$ 聚类有效性的指标 (CQED). 首先, 给出如下定义.

定义 1 (类内各维上的样本密度)　第 i 个类内的样本在第 j 维上的密度为

$$\sigma_{ij} = \frac{|C_i|}{W_{ij}}. \tag{13}$$

其中: $|C_i|$ 为类 i $(1 \leqslant i \leqslant k)$ 包含的样本量, W_{ij} 为类 i 中的样本在第 j 维上的取值长度.

定义 2 (类在各维上的类密度)　类在第 j 维上的类密度为

$$\rho_j = \frac{k}{\overline{W}_j}. \tag{14}$$

其中 $\overline{W}_j$ 为第 j 维上的类中心极值的差值, 有

$$\overline{W}_j = \begin{cases} O_{j_{\max}} - O_{j_{\min}}, & O_{j_{\min}} \neq O_{j_{\max}}; \\ 1, & O_{j_{\min}} = O_{j_{\max}}. \end{cases} \tag{15}$$

$O_{j_{\max}}$、$O_{j_{\min}}$ 为第 j 维上类中心的最大、最小取值.

定义 3 (各维上的样本密度均值)　类内的样本在第 j 维上的密度均值为

$$\widetilde{\sigma}_j = \frac{1}{k}\sum_{i=1}^{k}\sigma_{ij}. \tag{16}$$

根据最佳聚类数条件下, 类内样本密度较大和类密度较小的分布特征, 本文提出基于类内样本密度及类密度的有效性指标, 用于评价不同聚类数条件下, 数据集 X 的聚类质量. 在给定数据集 X 的聚类划分 $C^k=(C_1, C_2, \cdots, C_k)$ 之后, 用 $\widetilde{\sigma}$ 来度量类内样本在所有维上的密度分布情况, 用 $\widetilde{\rho}$ 来度量类在所有维上的密度分布情况, 其中

$$\widetilde{\sigma} = \frac{1}{hk}\sum_{j=1}^{h}\sum_{i=1}^{k}\sigma_{ij}, \tag{17}$$

$$\widetilde{\rho} = \frac{k}{h}\sum_{j=1}^{h}\frac{1}{\overline{W}_j}. \tag{18}$$

定义 4　S 为类内样本在所有维上的密度均值与类在所有维上的密度均值之差, 即

$$S = \widetilde{\sigma} - \widetilde{\rho}. \tag{19}$$

定义 5　A 为类内样本在所有维上的密度均值与类在所有维上的密度均值之和, 即

$$A = \widetilde{\sigma} + \widetilde{\rho}. \tag{20}$$

定义 6　有效性指标 $Q(C^*)$ 为 S 与 A 的比值, 即

$$Q(C^k) = \frac{S}{A} = \frac{\widetilde{\sigma}-\widetilde{\rho}}{\widetilde{\sigma}+\widetilde{\rho}}. \tag{21}$$

最佳聚类数 k_{opt} 则根据下式予以确定:

$$k_{\text{opt}} = \arg\max\{Q(C^k)\},\ k_{\min} \leqslant k \leqslant k_{\max}. \tag{22}$$

根据以上定义, 采用 2.3 节所示的最佳聚类数算法确定 k_{opt} 为最终的聚类数.

3.5　算法复杂度分析

3.5.1　初始聚类中心选择算法复杂度分析

初始聚类中心的选择主要分为候选初始聚类中心的选择和初始聚类中心的确定. 为便于分析, 假定各维上的样本密度统计区间数均为 λ. 统计样本在第 1 维特征参数下各样本密度统计区间内样本的密度时, 需进行 $2N\lambda$ 次运算操作. 在选择候选初始聚类中心的过程中, 假定密度峰值占样本密度统计区间总数的平均比例为 a, 由第 j 维参数确定第 $j+1$ 维参数的单步执行次数为 $\lambda a(N\lambda+\lambda)$, 样本的维数为 h 时, 算法的执行次数为

$$2N\lambda + \lambda a(N\lambda+\lambda) + \cdots + (\lambda a)^{h-1}(N\lambda+\lambda). \tag{23}$$

在采用最大最小距离算法确定最终的初始聚类

中心过程中, 算法的复杂度主要体现在度量候选初始聚类中心的距离, 鉴于候选初始聚类中心的数量较原始样本而言可以忽略不计, 因此, 初始聚类中心选择算法的计算复杂度可表示为$O(N(\lambda a)^{h-1})$. 在极端情况下, 当所有样本在各维上均匀分布且每一样本密度统计区间内仅有一个样本时, 由式(3)可知$\lambda^h \leqslant N$成立, 故

$$O(N(\lambda a)^{h-1}) \leqslant O(N(\lambda a)^h) \leqslant O(N\lambda^h) \leqslant O(N^2). \tag{24}$$

如上所述, 由于在选择初始聚类中心的过程中, 将密度峰值所在的区间的样本均值作为初始候选聚类中心, 而其他的样本密度统计区间则不会被考虑, 算法的实际运算复杂度将远小于$O(N^2)$.

3.5.2 最佳聚类数确定算法复杂度分析

最佳聚类数确定算法的计算复杂度主要是聚类算法执行过程的计算开销. 当$k = k_{\max}$时, 计算类内样本在所有维上的离散程度$\tilde{\sigma}$的迭代次数为$hk_{\max}$, 而计算簇在所有维上的离散程度$\tilde{\rho}$的迭代次数为h. 因此, 最佳聚类数确定算法指标CQED的计算复杂度为$O(hk_{\max})$. 由于$k_{\max}$为常数, 最佳聚类数确定算法为线性复杂度. 对于CH指标, 其计算全部样本的类内离差矩阵和类间离差矩阵测度的计算复杂度为$O(N)$; DB指标获得样本的类内散度与各聚类中心间距的测度的计算复杂度同样为$O(N)$; 具有相同计算复杂度的还有KL、COPS指标. 与上述指标相比, 对于BWP指标, 其需要计算某一类中的样本到其他每个类中样本平均距离的最小值以及同一类中的样本到其他所有样本的平均距离, 该指标的计算复杂度为$O(N^2)$; 而Wint、IGP指标与BWP指标具有相同的计算复杂度.

4 实验结果对比与分析

4.1 实验对象及相关设置

实验选择的主机为Windows XP操作系统, CPU主频为3.2 GHz (4核), 内存为4 GB. 实验分析对象为UCI数据集以及符合二维高斯分布的人工数据集MD_1. 采用的UCI数据集分别为Haberman's Survival (HS)、Dataset for ADL Recognition with Wrist-worn中的Accelerometer Descend Stairs (ADS)、IRIS以及Breast Cancer Wisconsin (BCW), 其相关参数如表1所示.

如图4所示, MD_1中包含5个类, 各类中的样本量分别为150、200、250、300及500, 且相邻类间在各维上存在着较多的重叠结构, 各类的中心分别为(200, 200)、(300, 500)、(500, 300)、(800, 150)及(1100, 400). 为检验算法的抗噪性能, 在二维空间内引入了占总样本量50%的随机噪声数据(700个噪声数据).

表1 UCI数据集信息

数据集	样本数	类数	维数
HS	306	2	3
ADS	594	3	3
IRIS	150	3	4
BCW	699	2	10

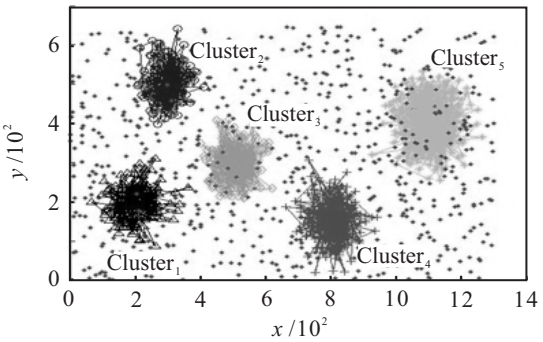

图4 人工数据集MD_1

4.2 初始聚类中心的选择

4.2.1 人工数据集上的初始聚类中心选择

采用本文提出的算法, 对MD_1数据集进行统计测试后, 得到如图5和图6所示的样本在x方向及y方向上样本的密度分布特征.

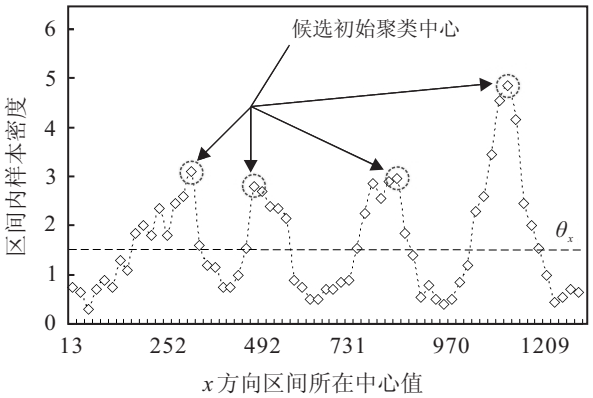

图5 MD_1在x方向上的样本密度分布特征

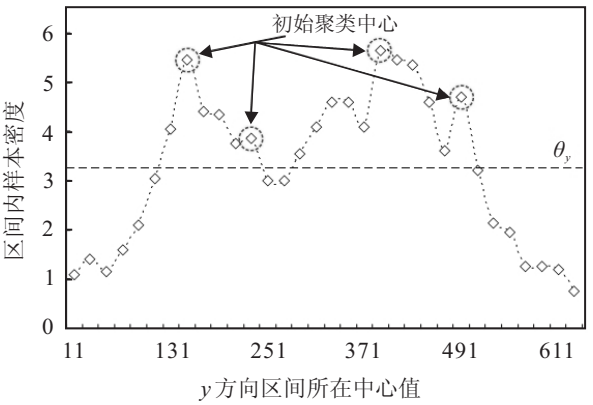

图6 MD_1在y方向上的样本密度分布特征

由式(3)计算获得的MD_1在x方向上的样本密度统计区间数为65, 而y方向上由于聚类间存在更大程度的重叠结构, 其样本密度统计区间数为32. 由式(4)可得x、y方向上的候选初始聚类中心筛选值分别为$\theta_x = 1.615$、$\theta_y = 3.230$. 由式(5)可得x、y方向上的密度峰值考察区间数分别为$\pi_x = 6$、$\pi_y = 3$. 因此, 在x方向上, 满足候选初始聚类中心筛选条件的取值分别为$x_1 = 332$、$x_2 = 492$、$x_3 = 850$、$x_4 = 1\,130$. 在y方向上, 满足候选初始聚类中心筛选条件的取值分别为$y_1 = 151$、$y_2 = 231$、$y_3 = 391$、$y_4 = 491$. 根据样本在各密度统计区间内的分布状况及最大最小距离算法, 确定以上各点的映射关系, 选定的初始聚类中心分别为(332, 231)、(332, 491)、(492, 391)、(850, 151)及(1 130, 391). 为说明本文提出的初始聚类中心选择方案的有效性, 在聚类数为最佳聚类数条件下, 定义初始聚类中心$U' = (u'_1, u'_2, \cdots, u'_k)$与真实聚类中心$U = (u_1, u_2, \cdots, u_k)$间的相似性度量值为[20]

$$\tau_i = \frac{4(u_i, u'_i)}{(|u_i| + |u'_i|)^2}. \tag{25}$$

由式(25)可得MD_1的初始聚类中心(332, 231)、(332, 491)、(492, 391)、(850, 151)、(1 130, 391)与真实聚类中心(200, 200)、(300, 500)、(500, 300)、(800, 150)、(1 100, 400)间的相似度分别为95.346 9%、99.846 8%、99.003 1%、99.908 8%、99.976 4%.

在利用MMD算法获取数据集MD_1的初始聚类中心的过程中, 由于数据集中50%噪声数据点的引入, 使得MMD算法获得的初始聚类中心与原始聚类中心相比存在较大的偏差, 通过1 000次实验获得的相似性度量值均值为62.156 2%, 其结果存在较大的随机性.

与本文算法相比, 文献[8]中的初始聚类中心仍以随机方式进行选择, 在选择过程中通过判断选择的样本到其他样本的距离是否大于给定的阈值来确定样本是否为初始聚类中心, 虽然缓解了算法陷入局部最优解的问题, 但初始聚类中心的选择仍具有较大的随机性, 通过1 000次实验获得的相似性度量值均值为70.079 1%.

文献[9]与文献[8]相比, 其初始聚类中心选择的随机性显著降低, 且不受噪声数据的干扰, 选择初始聚类中心的准确性有了一定的提高, 通过1 000次实验获得的相似性度量值均值为85.542 8%.

文献[10]采用逐一划分的方法寻找相距最大的数据对象作为初始聚类中心, 虽然可以避免算法陷入局部最优解的问题且初始聚类中心的选择结果也较为确定, 但算法易受噪声数据的干扰, 通过1 000次实验获得的相似性度量值均值为63.973 0%.

与MMD算法及文献[8-10]提出的算法相比, 本文提出的初始聚类中心选择算法具有更好的效果.

4.2.2　UCI数据集上的初始聚类中心选择

采用本文提出的算法、MMD算法及文献[8-10]中的算法, 对UCI数据集的聚类中心的相似度测试结果如表2所示.

表 2　聚类中心相似性度量结果

数据集	本文算法	MMD算法	文献[8]	文献[9]	文献[10]
HS	0.926 5	0.541 0	0.682 0	0.801 0	0.608 0
ADS	0.951 0	0.603 0	0.713 0	0.816 0	0.631 0
IRIS	0.956 0	0.594 0	0.707 5	0.867 0	0.645 0
BCW	0.916 5	0.555 0	0.691 5	0.832 0	0.626 0

由表2可以看出, 在1 000次重复实验中, 采用本文算法获得的初始聚类中心更接近真实的聚类中心, 与真实聚类中心的相似度可达到90%以上. 采用MMD算法获得的初始聚类中心与真实聚类中心的相似性较其他算法而言相对较低, 主要原因在于第1个初始聚类中心的选择为随机的且算法更易受到离群数据的影响. 相对于文献[8-10]提出的算法, 文献[9]算法获得的初始聚类中心与真实聚类中心之间具有相对较高的相似性. 与人工数据集上的测试结果类似, 文献[8]中算法的执行结果优于文献[10]中的算法. 以上结果的产生原因已在分析人工数据集结果过程中有所阐述, 这里不再赘述.

4.2.3　初始聚类中心对迭代次数的影响

为验证初始聚类中心对K-means算法迭代次数的影响, 实验过程中, 执行1 000次K-means算法的平均迭代次数如表3所示.

表 3　K-means算法迭代次数

数据集	本文算法	MMD算法	文献[8]	文献[9]	文献[10]
HS	2	5	4	3	4
ADS	4	8	6	5	7
IRIS	2	4	4	3	4
BCW	2	5	4	3	4
MD_1	2	6	5	4	5

由表3可以看出, 本文算法获得的初始聚类中心更加接近真实的聚类中心, 显著降低了K-means算法的迭代次数. MMD算法获得的初始聚类中心与真实聚类中心间存在较大的差异且算法运行结果存在随

机性, 导致 K-means 算法的平均迭代次数最大. 由于文献 [9] 中算法的初始聚类中心选择结果优于文献 [8] 和文献 [10] 中算法的选择结果, 使得 K-means 算法的平均迭代次数也小于文献 [8] 和文献 [10] 中的算法. 而文献 [8] 中的算法使得 K-means 算法的迭代次数较文献 [10] 中的算法有了一定程度的降低.

4.3 最佳聚类数的确定

为进一步对比分析 CQED 指标的有效性, 在相同初始聚类中心的条件下, 对比分析了采用 KL、Wint、IGP、COPS、BWP 指标的最佳聚类数测试结果, 如表 4 所示. 与 KL、Wint、IGP、COPS、BWP 指标相比, CQED 指标在所有数据集的测试过程中, 均得到了正确的聚类数.

表 4 不同指标下的聚类数分析结果

数据集	实际类数	实验获得类数					
		KL	Wint	IGP	COPS	BWP	CQED
HS	2	2	2	2	3	2	2
ADS	3	5	3	4	3	3	3
IRIS	3	3	3	3	5	3	3
BCW	2	2	4	2	2	3	2
MD_1	5	4	5	5	5	5	5

为进一步考察在最佳聚类数的条件下, 不同聚类度量指标的执行效率, 分别统计了执行 1 000 次 KL、Wint、IGP、COPS、BWP、CQED 指标的平均运算时间, 其在 HS、ADS、IRIS、BCW、MD_1 数据集上的运行结果如表 5 所示. 实验过程中, 计算 Wint、IGP、BWP 指标值的时间基本相同, 其原因在于 Wint、IGP、BWP 指标具有相似的计算复杂度. 同时, 由于 KL、COPS 指标间也具有相似的计算复杂度, 其运算时间也基本相近. 由于本文 CQED 指标的计算复杂度小于 KL、Wint、IGP、COPS、BWP 指标, 在确保聚类结果准确性的同时, 其运算效率均高于 KL、Wint、IGP、COPS、BWP 指标.

表 5 计算不同指标值的时间 s

数据集	t					
	KL	Wint	IGP	COPS	BWP	CQED
HS	0.017 5	0.232 3	0.216 0	0.018 2	0.217 9	0.016 6
ADS	0.032 9	0.461 0	0.438 4	0.033 6	0.483 1	0.031 0
IRIS	0.011 9	0.047 8	0.041 7	0.013 3	0.049 1	0.011 7
BCW	0.110 4	1.381 9	1.203 0	0.118 6	1.411 7	0.108 3
MD_1	0.090 5	1.857 3	1.793 3	0.093 3	1.871 7	0.089 2

5 结 论

初始聚类中心及最佳聚类数的选择, 直接影响到 K-means 等基于划分的聚类算法的执行效率及结果. 针对随机指定方法及最大最小距离等算法的不足, 本文基于类内样本密度较大的特点, 将密度峰值所在区间的中心值作为候选的初始聚类中心, 结合最大最小距离算法, 进一步从候选的初始聚类中心关系树筛选出初始聚类中心. 鉴于基于欧氏距离的聚类有效性度量函数无法适用于高维数据的度量, 且随着维数的增加存在距离趋近现象, 基于类内样本密度和类密度等定义, 建立了基于类内样本密度及类密度的最佳聚类数指标度量函数. 在人工数据集和 UCI 数据集上的测试结果表明: 采用本文提出的初始聚类中心选择算法获得的初始聚类中心与真实聚类中心间的相似性可达到 90% 以上; 提出的最佳聚类数确定算法在确保获得正确聚类数的同时, 具有更高的执行效率. 关于对聚类算法做进一步改进, 使其适用于不规则形状数据的聚类将是下一步的研究方向.

参考文献(References)

[1] Xu Rui. Survey of clustering algorithm[J]. IEEE Trans on Neural Networks, 2005, 16(3): 645-678.

[2] 王守强, 朱大铭. 基于最小聚类求解 k-means 问题算法[J]. 通信学报, 2010, 31(7): 46-52.
(Wang S Q, Zhu D M. Algorithm for the k-means clustering based on minimum cluster size[J]. J of Communications, 2010, 31(7): 46-52.)

[3] Gelbard R, Goldman O, Spiegler I. Investigating diversity of clustering methods: An empirical comparison[J]. Data & Knowledge Engineering, 2007, 63(1): 155-166.

[4] Pilevar A H, Sukumar M. GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases[J]. Pattern Recognition Letters, 2005, 26(7): 999-1010.

[5] 于勇前, 赵相国, 王国仁, 等. 一种基于密度单元的自扩展聚类算法[J]. 控制与决策, 2006, 21(9): 974-978.
(Yu Y Q, Zhao X G, Wang G R, et al. An self-expanded clustering algorithm based on density units[J]. Control and Decision, 2006, 21(9): 974-978.)

[6] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
(Sun J G, Liu J, Zhao L Y. Clustering algorithms research[J]. J of Software, 2008, 19(1): 48-61.)

[7] Likas A, Ulassis M, Uerbeek J. The global k-means clustering algorithm[J]. Pattern Recognition, 2003, 36(2): 451-461.

[8] 刘一鸣, 张化祥. 可变阈值的K-means初始中心选择方法[J]. 计算机工程与应用, 2011, 47(32): 56-58.
(Liu Y M, Zhang H X. Approach to selecting initial centers for K-means with variable threshold[J]. Computer Engineering and Applications, 2011, 47(32): 56-58.)

[9] 熊忠阳, 陈若田, 张玉芳. 一种有效的K-means聚类中心初始化方法[J]. 计算机应用研究, 2011, 28(11): 4188-4190.
(Xiong Z Y, Chen R T, Zhang Y F. Effctive method for cluster' initialization in K-means clustering[J]. Application Research of Computers, 2011, 28(11): 4188-4190.)

[10] 陈光平, 王文鹏, 黄俊. 一种改进初始聚类中心选择的K-means算法[J]. 小型微型计算机系统, 2012, 33(6): 1320-1323.
(Chen G P, Wang W P, Huang J. Improved initial clustering center selection method for K-means algorithm[J]. J of Chinese Computer Systems, 2012, 33(6): 1320-1323.)

[11] Song M J, Rajasekaran S. Fast k-means algorithms with con-stant approximation[C]. Proc of the 16th Annual Int Symposium on Algorithms and Computation. Sanya, 2005: 1029-1038.

[12] 周世兵. 聚类分析中的最佳聚类数确定方法研究及应用[D]. 无锡: 江南大学物联网工程学院, 2011.
(Zhou S B. Research and application on determining optimal number of clusters in cluster analysis[D]. Wuxi: School of IoT Engineering, Jiangnan University, 2011.)

[13] Calinski T, Harabasz J. A dendrite method for cluster analysis[J]. Communications in Statistics, 1974, 3(1): 1-27.

[14] Davies D L, Bouldin D W. A cluster separation measure[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227.

[15] Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset[J]. Genome Biology, 2002, 3(7): 1-21.

[16] Dimitriadou E, Dolnicar S, Weingessel A. An examination of indexes for determining the number of cluster in binary data sets[J]. Psychometrika, 2002, 67(1): 137-160.

[17] Kapp A V, Tibshirani R. Are clusters found in one dataset present in another dataset?[J]. Biostatistics, 2007, 8(1): 9-31.

[18] 陈黎飞. 高维数据的聚类方法研究与应用[D]. 厦门: 厦门大学计算机科学系, 2008.
(Chen L F. Research on clustering methods for high dimensional data and their applications[D]. Xiamen: Department of Computer Science, Xiamen University, 2008.)

[19] 倪巍伟, 陈耿, 吴英杰, 等. 一种基于局部密度的分布式聚类挖掘算法[J]. 软件学报, 2008, 19(9): 2339-2348.
(Ni W W, Chen G, Wu Y J, et al. Local density based distributed clustering algorithm[J]. J of Software, 2008, 19(9): 2339-2348.)

[20] 江依法, 周青, 陈伟燕. 一种改进的模板匹配算法及其在ECG波形识别中的应用[J]. 中国生物医学工程学报, 2012, 31(5): 775-780.
(Jiang Y F, Zhou Q, Chen W Y. An improved template-matching algorithm and its application in ECG waveform recognition[J]. Chinese J of Biomedical Engineering, 2012, 31(5): 775-780.)

(责任编辑：李君玲)