计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800191

基于密度峰值聚类的高斯混合模型算法

王卫东　徐金慧　张志峰　杨习贝
江苏科技大学计算机学院　江苏 镇江 212100

摘　要　由于存在大量服从高斯分布的样本数据,采用高斯混合模型(Gaussian Mixture Models,GMM)对这些样本数据进行聚类分析,可以得到比较准确的聚类结果。通常采用EM算法(Expectation Maximization Algorithm)对GMM的参数进行迭代式估计。但传统EM算法存在两点不足:对初始聚类中心的取值比较敏感;迭代式参数估计的迭代终止条件是相邻两次估计参数的距离小于给定的阈值,这不能保证算法收敛于参数的最优值。为了弥补上述不足,提出采用密度峰值聚类(Density Peaks Clustering,DPC)来初始化EM算法,以提高算法的鲁棒性,采用相对熵作为EM算法的迭代终止条件,实现对GMM算法参数值的优化选取。在人工数据集及UCI数据集上的对比实验表明,所提算法不但提高了EM算法的鲁棒性,而且其聚类结果优于传统算法。尤其在服从高斯分布的数据集上的实验结果显示,所提算法大幅提高了聚类精度。

关键词:密度峰值聚类;相对熵;高斯混合模型;EM算法;聚类算法
中图法分类号　TP391.4

Gaussian Mixture Models Algorithm Based on Density Peaks Clustering

WANG Wei-dong,XU Jin-hui,ZHANG Zhi-feng and YANG Xi-bei
College of Computer Science,Jiangsu University of Science and Technology,Zhenjiang,Jiangsu 212100,China

Abstract　Due to the existence of a large number of sample data which obey the Gaussian distribution,GMM (Gaussian mixture models) is used to cluster these sample data and get more accurate clustering results. In general,EM algorithm(expectation maximization algorithm) is used to estimate the parameters of GMM iteratively. However,the traditional EM algorithm has two shortcomings:it is sensitive to the initial clustering center;the itera-tive termination condition of iterative parameter estimation is to judge that the distance between two adjacent estimated parameters is less than a given threshold,which can't guarantee that the algorithm converges to the optimal value of the parameters. In order to overcome the above shortcomings,density peaks clustering (DPC) is proposed to initialize EM algorithm to improve the robustness of the algorithm. The relative entropy is used as the iteration termination condition of the EM algorithm to optimize the parameters of GMM algorithm. The comparative experiments on artificial datasets and UCI datasets show that the new algorithm not only improves the robustness of EM algorithm,but also outperforms the traditional clustering algorithm. On the datasets which obey Gaussian distribution,the new algorithm greatly improves the clustering accuracy.

Keywords　Density peaks clustering,Relative entropy,Gaussian mixture models,Expectation maximization algorithm,Clustering algorithm

1　引言

聚类分析是进行数据挖掘[1-2]、实现商业智能的重要工具,被广泛应用于医疗诊断、图像处理[3]、信息检索[4]和生物信息学等领域。聚类分析是将一组未知分布的样本数据分类到不同的类或者簇中的一个过程。对于聚类结果的评价,要求同一个簇中的对象有很大的相似性,而不同簇间的对象有很大的相异性。聚类分析算法主要包括以下4类[1-9]:基于划分的聚类方法、基于层次的聚类方法、基于网格的聚类方法和基于集成技术的聚类方法。

由于来自工业企业及服务行业的数据急速增长,如何从这些大数据中抽取出有价值的信息成为了目前的研究热点,而聚类分析是解决上述问题的有力工具。文献[1]分析了如何运用不同的聚类算法来满足稀疏工业数据集的各种聚类需求。文献[2]针对不同的聚类算法在处理不确定数据、多媒体数据、图形数据、生物数据、流数据、文本数据、时间序列数据等数据类型时的特点,系统地讨论了数据聚类算法的应用。

在各种聚类方法中,基于划分的聚类方法因其简单易用等特点而得到了广泛的应用,如著名的k-means[5]及其改进算法。文献[6]运用k-means聚类算法对教学满意度的调查数据进行挖掘和分析,分析过程快速、高效、合理,从而为采取有效措施改进教学策略和教学设计提供了可靠、有价值的参

到稿日期:2020-08-28　返修日期:2020-11-30
基金项目:国家自然科学基金(61572242)
This work was supported by the National Natural Science Foundation of China(61572242).
通信作者:王卫东(78653221@qq.com)

考依据。为了解决 k-means 算法随机选取初始类别中心的问题，文献[7]将谱聚类算法与 k-means 算法结合，优化了 k-means 的初始类别中心点坐标。文献[8]提出了一种动态 k-means 聚类算法，该算法在聚类的过程中可以动态地修改 k 值，实现了对聚类类别数的优化。

但是 k-means 算法的缺点也很明显，如对聚类中心均值的简单使用，且假设数据点是呈圆形分布的。而有限混合模型是分析复杂现象的一种灵活而强有力的建模工具，它提供了用简单结构模拟复杂密度的有效方法。其中，基于高斯分布的高斯混合模型 GMM[10-11]用于解决同一集合中的数据包含多个不同的高斯分布的情况。GMM 采用有限个高斯分布的概率密度函数，按照一定的权重进行混合，来量化数据的分布，从而达到聚类的目的。

对于高斯混合模型 GMM 的参数，通常采用 EM[10-11]（期望最大化）算法进行估计。EM 算法的优点是简单稳定，可以在不知道待估计参数先验信息和观测数据不完备的情况下，通过迭代来计算参数的最大似然估计。但 EM 算法对输入的初始值比较敏感，采用传统的随机初始值方法时，聚类结果往往波动较大，即算法的鲁棒性很差。而聚类结果的大幅波动极大地限制了 GMM 算法的应用。EM 算法的另一个问题是，在采用迭代式参数估计时，其迭代终止条件并不能保证所估计的参数值是最优值。

对于 EM 算法的初始值敏感问题，主要有两类解决方案：1)对随机初始值进行优化；2)对初始值进行优化选取。如采用 k-means 算法对随机选取的初始值进行迭代式优化，但这种方法的结果同样不稳定，也会有较大的偏差。在优化选取初始值方法中，有代表性的是文献[12]中引入的 binning 方法，即装箱法。该方法将数据空间在各维上划分成一个一个的箱子，再把数据点投射到对应的箱子中，属于密度估计方法。使用该方法的难点在于每一维上最优或者近似最优的 bin 宽度的选取。目前，尚没有针对迭代终止条件的优化算法。

2014 年 6 月，Science 上发表了一篇题为“采用密度峰值点快速搜索的聚类算法”的文章，其提出的算法简称为 DPC[9]算法。DPC 定义了两个评价指标：局部密度 ρ_i 和距离 δ_i。通过这两个指标，该算法可以较准确地刻画出聚类中心的特征。因此，对于长期困扰聚类分析的两个基本问题（如何预估类簇数和初始类簇中心），DPC 提出了行之有效的解决方案。但由于该算法将样本判别为离其最近且密度比它大的样本所在的类，当类间样本相互重叠时，聚类的准确率就会大大降低，故该算法的聚类准确率并不高。

本文提出采用密度峰值聚类算法 DPC 来初始化 EM 算法，解决了 EM 算法的初始值敏感问题。另一方面，本文提出采用相对熵（也称为 KL 散度）作为 EM 算法的迭代终止条件，实现了对 GMM 算法参数估计值的优化选取。

2 高斯混合模型

高斯混合模型将聚类的数据看作是来自于 K 个高斯分布的混合概率分布，这些高斯分布分别代表不同的类。由中心极限定理可知，在样本数量足够多的情况下，每个 K 所代表的区域就可以用高斯分布来描述。高斯函数的另一个优点是具有良好的计算性能，高斯混合模型如下：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k f(y|\theta_k) \tag{1}$$

其中，α_k 是系数，$\alpha_k\geqslant 0$，且 $\sum_{k=1}^{K}\alpha_k=1$，$f(y|\theta_k)$ 是高斯密度函数，参数 $\theta_k=(u,\sigma_k^2)$。

通常采用 EM 算法对 GMM 中的参数 α_k，θ_k 进行估计。EM 的输入数据为观测数据 $y_1,y_2,\cdots,y_N$，算法的输出是高斯混合模型的参数值。EM 算法分为两步：E 步是求解目标函数期望，通常是求解目标函数取对数之后的期望值；M 步是期望最大化，采用极大似然估计、拉格朗日乘数法对参数求偏导，最终确定新的参数。EM 算法的步骤如下：

(1)随机选取参数 α_k，θ_k 的初值，开始迭代；

(2)E 步：依据当前模型的参数，计算分模型 k 对观测数据 y_j 的响应度；

$$\hat{r}_{jk}=\frac{\alpha_k f(y|\theta_k)}{\sum_{k=1}^{K}\alpha_k f(y|\theta_k)} \tag{2}$$

其中，$j=1,2,\cdots,N$；$k=1,\cdots,K$。

(3)M 步：计算新一轮迭代的模型参数；

$$\hat{\mu}_k=\frac{\sum_{j=1}^{N}\hat{r}_{jk}y_j}{\sum_{j=1}^{N}\hat{r}_{jk}},k=1,2,\cdots,K \tag{3}$$

$$\delta_k^2=\frac{\sum_{j=1}^{N}\hat{r}_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\hat{r}_{jk}},k=1,2,\cdots,K \tag{4}$$

$$\hat{\alpha}_k=\frac{\sum_{j=1}^{N}\hat{r}_{jk}}{N},k=1,2,\cdots,K \tag{5}$$

(4)当满足下述公式时，迭代结束。

$$\|\hat{\alpha}_k-\hat{\alpha}_{k-1}\|\leqslant\varepsilon \tag{6}$$

其中，ε 为给定的阈值，当不满足式(6)时，算法转步骤(2)，继续迭代。

EM 算法实现简单，数值计算稳定，因此得到了广泛的应用。但 EM 算法也存在明显的不足：算法采用随机的方式选取初始值，导致初始值敏感，从而使 GMM 算法的聚类结果大幅波动；迭代的终止条件是相邻两次迭代参数 α_k 的距离小于给定的阈值，但该条件并不能保证估计的参数是最优的。

3 DPC 算法

DPC 算法可以有效地预估数据集的类簇数和初始聚类中心。但使用该算法进行聚类时，将样本判别为离其最近且密度比它大的样本所在的类，易发生所谓的“多米诺骨牌效应”[13]。一旦某一个样本分类错误，就会带来一连串的样本分类错误，因此 DPC 算法的聚类准确率并不高。本文仅采用 DPC 算法预估数据集的类簇数和初始聚类中心，目的是解决 EM 算法的初始值敏感问题。

为了刻画聚类中心，DPC 定义了两个评价指标：局部密度 ρ_i 和距离 δ_i。其基本思想是：聚类中心是其局部密度大于围绕它的邻居的局部密度；同时，不同类中心之间的距离相对较远。因此，聚类中心就是 ρ_i 和 δ_i 同时较大的数据点。局部密度 ρ_i 是通过截断函数 Cut-off kernel 来定义的，如式(7)所示：

$$\rho_i=\sum_{j\in I_s}X(d_{ij}-d_c) \tag{7}$$

其中，d_{ij} 表示样本 i 和 j 的距离，d_c 为截断距离，函数 $X(x)$ 的定义如下：

$$X(x)=\begin{cases}1, & x<0\\ 0, & x\geqslant 0\end{cases} \tag{8}$$

从式(7)、式(8)可以看出，局部密度 ρ_i 描述的是围绕样本点的样本密度，相当于落入以 d_c 为半径的超球体内的样本数量。显然，d_c 取不同的值，ρ_i 的值也将发生相应的改变，因此 ρ_i 的值依赖于 d_c 的取值。

另一个参数是距离 δ_i，它描述的是数据点之间的距离，定义如下：

$$\delta_i=\begin{cases}\min\limits_{j\in I_s^i}\{d_{ij}\}, & I_s^i\neq\varnothing\\ \max\limits_{j\in I_s}\{d_{ij}\}, & I_s^i=\varnothing\end{cases} \tag{9}$$

其中，指标集为：

$$I_s^i=\{k\in I_s:\rho_k>\rho_i\} \tag{10}$$

从式(9)、式(10)可以看出，当 x_i 具有最大局部密度时，δ_i 表示 S 中与 x_i 距离最大的数据点与 x_i 之间的距离；否则，δ_i 表示在所有局部密度大于 x_i 的数据点中，与 x_i 距离最小的那个数据点与 x_i 之间的距离。

DPC 算法将每个数据点的 ρ_i 值和 δ_i 值表示在一个二维决策图上。用户根据决策图的分布情况，对聚类中心点进行选择，这是一个人工操作，无法自动完成。

针对 DPC 算法的不足，文献[14]提出了一种根据不同的数据集自动计算出 d_c 最优值的方法，简称 DF-DPC 算法。但是，该算法并没有很好地解决 DPC 算法聚类准确率低的问题。文献[15]提出了一种扩展的 E-CFSFDP 算法，该算法先调用 DPC 算法，然后执行一个子类的合并步骤，其目的是解决类簇中存在多密度峰值(或称为无密度峰值)的情况，但该算法的聚类准确率也没有明显提高。文献[16]针对 DPC 算法效率较低的问题，提出了一种基于相对邻域和剪枝策略的密度峰值快速搜索聚类算法 RP-DPC，该算法属于时间性能优化算法，其聚类精度与 DPC 算法相同。

4 基于相对熵的迭代优化

EM 算法的迭代终止条件是：当相邻两次迭代的参数值 α_k 的距离小于给定的阈值时，迭代终止。但是，一方面阈值的取值并没有明确的标准；另一方面也不能保证参数 α_k，θ_k 取得最优值。本文研究了高斯混合模型的理论基础，基于最小错误的贝叶斯决策，如图 1 所示。

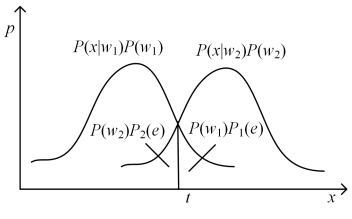


图 1 贝叶斯决策

Fig. 1 Bayesian decision making

图 1 中，$P(\omega_1)P_1(e)$ 为样本属于类别 1，但被划分为类别 2 的情况。同样，$P(\omega_2)P_2(e)$ 为样本属于类别 2，但被划分为类别 1 的情况。设总体的平均错误率为 $P(e)$，则 $P(e)$ 为：

$$P(e)=P(\omega_2)P_2(e)+P(\omega_1)P_1(e)$$

$$P_1(e)=\int_t^{\infty}p(x|\omega_1)P(\omega_1)\mathrm{d}x$$

$$P_2(e)=\int_{-\infty}^{t}p(x|\omega_2)P(\omega_2)\mathrm{d}x$$

当采用 EM 算法估计的参数 α_k，θ_k 取得最优值时，贝叶斯决策的平均错误率 $P(e)$ 达到最小值。因此，EM 算法的迭代终止条件应该为：当平均错误率 $P(e)$ 最小时，迭代结束。

如何判断平均错误率 $P(e)$ 最小？对于某个样本，将其判别为两类的概率相差较小时，其被错分的概率较大，反之被错分的概率较小。本文引入信息论中的相对熵(relative entropy)[17]概念，相对熵是对两个概率分布间差异的非对称性度量，等价于两个概率分布的信息熵[17](shannon entropy)的差值。以两类的分类问题为例，当某个样本被判别为两类的概率差异越大，则相对熵越大，该样本被错分的概率就越小；否则，相对熵越小，被判别为两类的概率差异越小，该样本被错分的概率就越大。

当第 i 个样本的两类分类概率为 $P(i)$ 和 $Q(i)$ 时，其相对熵的计算式如下：

$$D_i(p\parallel Q)=\sum_i P(i)\log_a\frac{P(i)}{Q(i)} \tag{11}$$

式(11)表示求 P 与 Q 之间的对数差在 P 上的期望值。当 $P(i)=Q(i)$ 时，相对熵为 0，其他情况下相对熵大于 0。以相对熵为评价样本分类的度量指标，其值越大，聚类结果的可信度就越高，否则可信度越低。

本文判断相对熵小于某个阈值的样本数，当该样本数在某次迭代时取得极小值时认为 EM 算法在本次迭代中所估计的参数 α_k，θ_k 达到最优值。对于多类问题，由于概率较大的两类发生错分的可能性较大，因此取两个分类概率最大的概率值来计算其相对熵。

对于样本 y_i，概率最大的两个概率值分别是 $P(y_i)$ 和 $Q(y_i)$。利用式(11)计算相对熵 $D_i(p\parallel Q)$。设相对熵的阈值为 δ，采用下式判断相对熵小于阈值 δ 的样本数。

$$f_j(y_i)=\begin{cases}1, & D_i(p\parallel Q)<\delta\\ 0, & D_i(p\parallel Q)\geqslant\delta\end{cases} \tag{12}$$

其中，$i=1,\cdots,N$ 为样本数，$j=1,\cdots,T$ 为 EM 算法的迭代次数。采用式(13)计算某次迭代相对熵小于阈值的样本总数。

$$Sum_j=\sum_{i=1}^{N}f_j(y_i) \tag{13}$$

采用式(14)作为 EM 算法的迭代终止条件。

$$Sum_{t-1}\geqslant Sum_t\leqslant Sum_{t+1} \tag{14}$$

式(14)说明，当某次迭代的相对熵小于阈值 δ 的样本数与相邻两次迭代对比为最小值时，迭代终止。最后，取该次迭代的参数值作为 EM 算法的估计值输出。

5 实验结果及分析

5.1 实验数据集及评价指标

本文实验采用人工数据集[18-20]上的 5 个数据子集(见表 1)和 UCI 机器学习数据库[21]中的 4 个数据集(见表 2)。这些数据集都是测试聚类算法的经典数据集。人工数据集是二维数据集，可以直观地展示不同的聚类初始值选取方法对聚类结果的影响。所有实验的实验环境均为 Win10 64 bit 操作系统，Matlab R2018a 软件，8 GB 内存，CPU Intel i7-9750H，GPU RTX2060。

表 1 人工数据集

Table 1 Artificial datasets

数据集	样本数	特征数	类簇数
flame	240	2	2
Aggregation	788	2	7
Pathbased	300	2	3
Spiral	312	2	3
Jain	373	2	2

表 2 UCI 数据集

Table 2 UCI datasets

数据集	样本数	特征数	类簇数
iris	150	4	3
wine	178	13	3
WDBC	569	30	2
waveform	5000	21	3

UCI 数据集中许多数据是真实的测量数据，iris 数据集也称为鸢尾花卉数据集，wine 数据集的 13 个属性是葡萄酒的 13 种化学成分，WDBC 是乳腺癌数据集，waveform 是噪声波形数据集，这些数据集在样本规模、特征数、类簇数等方面差别很大。

为了消除实验数据不同量纲对实验结果的影响，将其转化为无量纲的纯数值，以便于不同单位或量级的特征数据能够进行比较。本文采用 min-max 进行标准化处理。

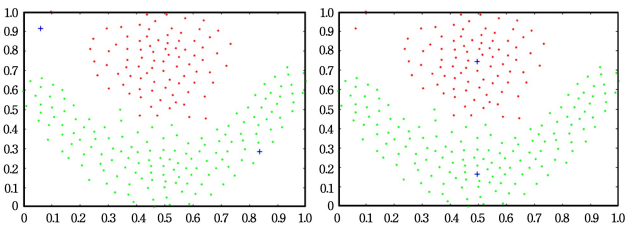
$$x_{ij}^{*}=\frac{x_{ij}-\min_j}{\max_j-\min_j} \tag{15}$$

其中，x_{ij} 表示第 i 个样本的第 j 个特征值，$\min_j$ 和 $\max_j$ 分别表示该特征的最小值和最大值。这样将特征数据归一化到[0,1]区间。

采用两种度量指标对实验结果进行评价：聚类准确率(Accuracy，Acc)和调整兰德指数(Adjusted Rand Index，ARI)[22]。两种指标的取值上界为 1，取值越大表示聚类结果越好。Acc 是评价聚类结果的精度指标，需要计算正确聚类的样本数占样本总数的比例。ARI 基于样本对计数，其中，RI 计算样本预测值与真实值之间的相似度，RI 的取值范围是[0,1]。对于随机结果，ARI 具有更高区分度，取值范围是[−1,1]，值越大表示聚类结果和真实情况越吻合。

5.2 参数 $\boldsymbol{\alpha_k}$，$\boldsymbol{\theta_k}$ 的初始值

α_k 的初始值为各类相等的均值。对于参数 θ_k 的初始值，EM 采用随机法选取初始类别中心。本文采用 DPC 算法得到初始类别中心。图 2 给出了随机法和 DPC 算法在 flame 数据集上选取的初始中心点。



(a)随机初始中心点 (b)DPC 初始中心点

图 2 在 flame 数据集上两种初始中心选取方法的对比

Fig. 2 Comparison of two initial center selection methods on flame dataset

从图 2 可以看出，随机选取的初始中心与真实的类别中心偏差较大，而利用 DPC 算法得到的初始类别中心明显优于随机类别中心。图 3 给出了采用随机初始聚类中心方法、运行 3 次得到的 3 个随机初始聚类中心与 DPC 初始中心的聚类准确率对比。先采用 k-means 对 3 个随机初始中心进行优化，再与 DPC 进行对比。

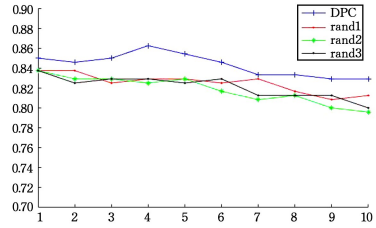


图 3 迭代 10 次的聚类准确率

Fig. 3 Clustering accuracy of 10 iterations

图 3 给出了将 EM 算法迭代 10 次的各次聚类准确率。图 3 中，采用 DPC 选取初始中心后，GMM 算法的准确率明显高于采用随机中心的传统算法。而 3 个随机初始中心的聚类准确率指标相差不大。图 3 还表明，将聚类准确率的均值及方差作为衡量算法的鲁棒性指标，DPC 算法的优势明显。采用 DPC 算法后，EM 估计的参数波动更小，提高了算法的鲁棒性。

随着迭代次数的增加，聚类准确率反而缓慢下降，原因是 GMM 算法的参数存在最优值，在达到最优值后继续迭代，反而会使错分的样本数增加，从而使得聚类的准确率下降。

5.3 采用相对熵的迭代优化

对于 EM 算法的迭代终止条件，本文不再采用相邻两次估计参数的距离，而是以相对熵小于某个阈值的样本数量作为衡量标准。当类别数大于两类时，则选取样本所有聚类概率中最大的两个概率，计算其相对熵。表 3 列出了在 UCI 的 iris 数据集上本文算法的聚类结果。

表 3 iris 数据集上本文算法的聚类结果

Table 3 Clustering results of our algorithm on iris dataset

阈值	1	2	3	4	5	6	7	8	9	10
$\delta=0.5$	14	9	9	8	4	1	4	3	1	1
$\delta=0.6$	17	10	10	8	5	2	4	3	3	2
$\delta=0.7$	18	10	10	8	5	2	4	3	2	2
$\delta=0.8$	20	13	11	9	5	7	4	3	3	3
Dist	0.009	0.007	0.016	0.006	0.008	0.012	0.013	0.012	0.010	0.010
Acc/%	89.33	93.33	96.67	97.33	99.33	98.67	98.00	98.00	96.67	97.33

实验将相对熵的阈值 δ 设置为 0.5～0.8，考察选取不同的阈值对本文算法的影响，即本文算法对阈值参数 δ 的鲁棒性。用 $Dist=\|\hat{\alpha}_k-\hat{\alpha}_{k-1}\|$ 表示传统 EM 算法估计的相邻两个参数的距离，聚类结果采用聚类准确率指标 Acc。实验中将 EM 算法迭代 10 次。

EM 算法的初始参数是通过 DPC 算法得到的，本文算法的聚类准确率很高，最高达到了 99.33%。这是目前在 iris 数据集上取得的最佳准确率指标，说明在服从高斯分布的数据集上，采用 DPC+GMM 可以获得极佳的聚类效果。

但并不是迭代的次数越多聚类的准确率就越高，最优值出现在第 5 次迭代中。实验计算了相邻两次迭代参数 α_k 的欧氏距离，可以看出该距离波动不大，最小值与其他距离

值相差很小。第 4 次迭代对应的聚类准确率并不是最优值，甚至不是次优值。因此，距离测度不能保证估计的参数值是最优的。

本文判断相对熵小于给定阈值 δ 的样本数。当相对熵阈值 δ 取值为 0.5～0.8 时，在第 6 次迭代时有 3 次取得了最小值，第 5 次迭代有 1 次得到了最小值。因此，采用本文的相对熵作为迭代终止条件存在明显的最小值，其所对应的聚类准确率指标 3 次是次优的 98.67%，1 次对应最优的99.33%。实验结果表明，本文的基于相对熵的迭代终止条件优于传统的基于距离的迭代终止条件，且算法对阈值 δ 的取值并不敏感，算法的鲁棒性较好。

5.4 在人工数据集上的对比实验

本文选取了 5 个二维人工数据集，这些数据集有的服从高斯分布，有的并不服从高斯分布。图 4 给出了采用 DPC 算法获得的初始聚类中心。

flame 数据集的 DPC 选取结果见图 2。图 4 给出了其余 4 个数据集的 DPC 选取结果。可以看出，在 5 个数据集上都得到了较准确的初始类别数与初始聚类中心点坐标。

从图 2 可以看出，flame 数据集近似服从高斯分布。图 4 中，在 Pathbased 数据集的 3 个子类中，有 2 个子类服从高斯分布，Spiral 数据集不服从高斯分布，Jain 数据集近似服从高斯分布，而 Aggregation 数据集服从高斯分布。

可以预计，本文算法在 Aggregation 数据集上会取得较好的聚类结果，在其他 3 个数据集上的聚类结果为次优，而在 Spiral 数据集上的聚类准确率较低。

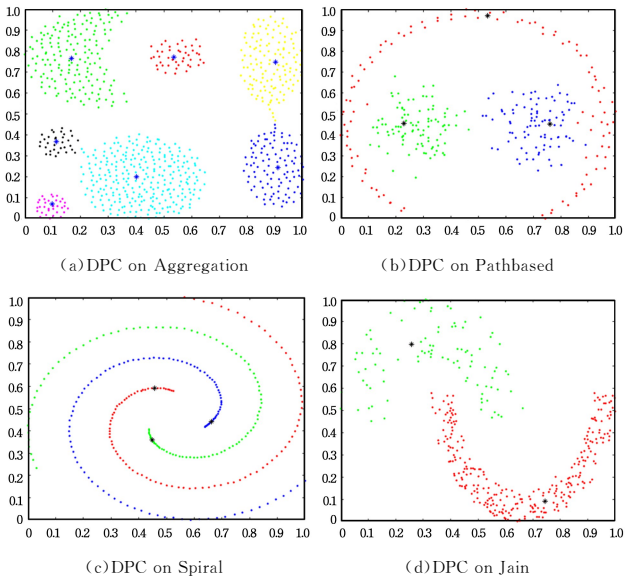


图 4 DPC 在 4 个人工数据集上得到的聚类中心

Fig. 4 Clustering centers obtained by DPC on four artificial datasets

从表 4 可以看出，当本文的相对熵阈值 δ 取值为 0.5～0.8 时，本文算法在 5 个人工数据集上的 ARI 和 Acc 指标波动很小，说明本文算法对阈值参数 δ 的取值并不敏感，算法的鲁棒性较好。本文算法在 5 个数据集上的 ARI 和 Acc 指标都好于传统的 GMM 算法。同时，本文算法在 Aggregation 数据集上取得了极高的聚类准确率，在 Spiral 数据集上的准确率最低，而在其他 3 个数据集上的结果为次优。

表 4 人工数据集上的对比实验

Table 4 Comparative experiment on artificial datasets

算法阈值 数据集	δ=0.5		δ=0.6		δ=0.7		δ=0.8		GMM	
	ARI	*Acc*	*ARI*	*Acc*	*ARI*	*Acc*	*ARI*	*Acc*	*ARI*	*Acc*
flame	0.762	0.863	0.749	0.854	0.749	0.854	0.749	0.854	0.710	0.825
Aggregation	0.997	0.996	0.997	0.996	0.997	0.996	0.997	0.996	0.987	0.983
Pathbased	0.720	0.690	0.720	0.690	0.720	0.690	0.720	0.690	0.712	0.670
Spiral	0.496	0.349	0.496	0.349	0.496	0.349	0.496	0.349	0.331	0.323
Jain	0.771	0.868	0.771	0.868	0.759	0.860	0.759	0.860	0.751	0.845

5.5 在 UCI 数据集上的对比实验

在 UCI 的 4 个数据集上，将本文算法与相关算法进行对比，包括 DPC 算法和 DF-DPC 算法[14]、E-DPC 算法[15]、传统的 GMM 算法及经典的基于划分的聚类算法 *k*-means。实验采用 Acc 和 ARI 两个评价指标，实验结果如表 5 所列。

表 5 UCI 数据集上的对比实验

Table 5 Comparative experiment on UCI datasets

数据集	DPC		DF-DPC		E-DPC		GMM		k-means		本文算法	
	ARI	*Acc*	*ARI*	*Acc*	*ARI*	*Acc*	*ARI*	*Acc*	*ARI*	*Acc*	*ARI*	*Acc*
iris	0.720	0.887	0.561	0.727	0.862	0.933	0.856	0.866	0.660	0.825	0.982	0.986
wine	0.672	0.882	0.756	0.916	0.672	0.882	0.941	0.946	0.830	0.932	0.954	0.960
WDBC	0.471	0.845	0.451	0.838	0.471	0.845	0.875	0.933	0.826	0.928	0.884	0.938
Waveform	0.268	0.586	0.276	0.492	0.338	0.716	0.453	0.506	0.254	0.501	0.468	0.533

从表 5 可以看出，本文算法在前 3 个数据集上的聚类精度都是最高的，尤其是与 DPC 相关的 3 种算法相比，聚类准确率 Acc 及 ARI 指标的绝对值平均提高了近 10%。同时，在这 3 个数据集上本文算法也显著优于传统 GMM 算法及 *k*-means 算法。

在 Waveform 数据集上，本文算法与对比算法的差别不大，在 Acc 指标上甚至明显低于 E-DCP 算法。由于 Waveform 数据集的样本数据是噪声波形数据，这些数据显然并不服从高斯分布，因此本文算法不具备任何优势。

对于本文算法的算法复杂度，由于采用 DPC 进行初始中心的选取，而 DPC 算法首先需要计算数据集中任意两个样本间的欧氏距离，其时间复杂度为 $O(m*n^2)$[16]，其中 m 为样

本特征个数，n 为数据集样本个数。当处理海量高维数据时，大量的高维欧氏距离计算会带来极大的时间开销，可以采用文献[16]提出的相对邻域和剪枝策略(称为 RP-DPC 算法)。RP-DPC 算法具有与 DPC 相同的聚类效果，并且时间性能显著优于 DPC 及其改进算法。

结束语 本文针对 EM 算法在估计高斯混合模型算法的参数时，存在对参数的初始值敏感和迭代终止条件不能保证取得最优值的问题，提出了采用密度峰值聚类算法 DPC 来初始化 EM 算法，解决了 EM 算法的初始值敏感性问题。本文还提出采用相对熵小于某个阈值的样本数量作为 EM 算法的迭代终止条件，实现了对参数估计值的优化选取。

在人工数据集和 UCI 数据集上的实验表明，本文算法明显优于传统 GMM 算法。尤其是在数据集服从高斯分布的情况下，本文算法取得了极佳的聚类效果。但当样本数据不服从高斯分布时，本文算法的优势并不明显，说明在具有不同分布特征的数据集上，应采用不同的聚类策略。因此，如何预估样本数据各个子类的分布特征成为进一步提高聚类准确率的关键。

参 考 文 献

[1] BENABDELLAH A C, BENGHABRIT A, BOUHADDOU I. A survey of clustering algorithms for an industrial context[J]. Procedia Computer Science, 2019, 148: 291-302.

[2] OYELADE J, ISEWON I, OLADIPUPO O, et al. Data Clustering: Algorithms and Its Applications[C]// 2019 19th International Conference on Computational Science and Its Applications (ICCSA). Saint Petersburg, Russia, 2019: 71-81.

[3] CHAI W Y, YANG F, YUAN S F, et al. Multi-class Gaussian Mixture Model and Neighborhood Information Based Gaussian Mixture Model for Image Segmentation[J]. Computer Science, 2018, 45(11): 272-277.

[4] ZOU C M, CHEN D. Unsupervised Anomaly Detection Method for High-dimensional Big Data Analysis[J]. Computer Science, 2021, 48(2): 121-127.

[5] MCQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Los Angeles: University of California, 1967: 281-297.

[6] SHEN H, DUAN Z. Application Research of Clustering Algorithm Based on K-Means in Data Mining[C]// 2020 International Conference on Computer Information and Big Data Applications (CIBDA). Guiyang, China, 2020: 66-69.

[7] SAPKOTA N, ALSADOON A, PRASAD P W C, et al. Data Summarization Using Clustering and Classification: Spectral Clustering Combined with k-Means Using NFPH[C]// 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). Faridabad, India, 2019: 146-151.

[8] ZAKIR H, NASIM A, AHMADR B, et al. A dynamic K-means clustering for data mining[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2019, 13(2): 521-526.

[9] ALEX R, ALESSANDRO L. Clustering by fast search and find of density peaks[J]. Science, 2014, 344: 1492-1496.

[10] DEMPSTERA P, LAIRDN M, RUBIND B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. Journal of the Royal Statistical Society, 1997, 39(1): 1-38.

[11] YANG M S, LAI C Y, LIN C Y. A robust EM clustering algorithm for Gaussian mixture models[J]. Pattern Recognition, 2012, 45(11): 3950-3961.

[12] YUE J, WANG S T. Algorithm EM and Its Initialization in Gaussian Mixture Model Based Clustering[J]. Microcomputer Information, 2006, 11(22): 244-247.

[13] XIE J Y, GAO H C, XIE W X. K-nearest neighbors optimized clustering algorithm by fast search and finding the density peaks of a data set[J]. Scientian Sinica Informationis, 2016, 46(2): 258-280.

[14] WANG S L, WANG D K, LI C Y, et al. Clustering by fast search and find of density peaks with data field[J]. Chinese Journal of Electronics, 2016, 3(25): 397-402.

[15] ZHANG W K, LI J. Extended fast search clustering algorithm: widely density clusters, no density peaks[J]. arXiv: 1505.05160, 2015.

[16] JI X, YAO S, ZHAO P. Relative Neighborhood and Pruning Strategy Optimized Density Peaks Clustering Algorithm[J]. ACTA Automatica Sinica, 2020, 46(3): 562-575.

[17] YANG W, CAI L, WU F. Image segmentation based on gray level and local relative entropy two dimensional histogram[J]. PLOS ONE, 2020, 15(3): 1-9.

[18] GIONIS A, MANNILA H, TSAPARAS P. clustering aggregation[C]// Proceedings of ACM Transactions on Knowledge Discovery from Data. 2007, 1(1): 1-30.

[19] LIMIN F, ENZO M. Flame, a novel fuzzy clustering method for the analysis of DNA microarray data[J]. BMC Bioininformatics, 2007, 8(1): 3-17.

[20] CHANG H, YEUNG D Y. Robust path-based spectral clustering[J]. Pattern Recognition, 2008, 41(1): 191-203.

[21] LI C M. UCI machine learning repository [EB/OL]. http://archive.ics.uci.edu/ml.

[22] VINH N X, EPPS J, BAILEY J. Information theoretic measures for clustering comparison: is a correction for chance necessary? [C]// Proceedings of ICML'09. Montreal, 2009: 1073-1080.

WANG Wei-dong, Ph. D, associate professor. His main research interests include pattern recognition and intelligent information processing.