文章编号:1001 - 2486(2007)02 - 0076 - 05

粗糙模糊 C - 均值算法及其在图像聚类中的应用*

王 丹,吴孟达
(国防科技大学 理学院,湖南 长沙 410073)

摘 要:提出一种新的粗糙模糊 C 均值算法(RFCM),该算法基于粗糙集的上、下近似的概念改进了 FCM 的目标函数,从而改变了隶属度函数的分布,使得隶属度函数的分布更加合理,同时 RFCM 的时间复杂性比 FCM 更低。将 RFCM 用于图像的聚类,相对于 FCM 算法,图像的边缘更光滑,同时对初始隶属度矩阵敏感度更低。该算法具有较好的稳定性,是一种实用的算法。

关键词:粗糙集;模糊 C - 均值算法;粗糙模糊 C - 均值算法
中图分类号:O159 文献标识码:A

Rough Fuzzy C-Means Algorithm and Its Application to Image Clustering

WANG Dan, WU Meng-da
(College of Science, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Based on the rough set model proposed by Pawlak, a new fuzzy C-means algorithm-rough fuzzy C-means algorithm (RFCM) is presented. The algorithm employs a new objective function which incorporates the concepts of the upper approximation and the lower approximation in rough sets, and which produces better results than Fuzzy C-mean algorithm at time complexity, clustering precision, the sensitivity to initial degree of membership matrix. The better effect can be testified by many experiments.

Key words: rough sets; fuzzy C-mean algorithm; rough fuzzy C-means algorithm

模糊 C - 均值(Fuzzy C-Means)算法[1]是由 Bezdek 提出的一种模糊聚类算法,该算法提出后在图像的分割、压缩、识别等领域得到了广泛了应用。但存在着如下问题:(1)收敛到局部极值;(2)算法性能依赖于初始聚类中心和初始隶属度矩阵;(3)须事先确定聚类数目;(4)计算量大;(5)对处于不同类边界处的元素分辨能力不高。针对这些问题,很多学者进行了研究。徐月芳[8]等学者利用遗传算法对目标函数进行寻优,来解决收敛到局部极值和算法对初始中心和隶属矩阵的依赖性,高新波、裴继红、于剑等[7,9]学者对 FCM 算法中平滑因子 m 和确定类的数目和加快算法计算速度进行了研究,Doğan zdemir[3]等学者改进了目标函数以获得更好的聚类效果。对于上述问题,本质上是由于迭代过程中生成的隶属度矩阵分布不合理所造成的。本文提出了一种新的模糊 C - 均值算法——粗糙模糊 C - 均值算法(Rough Fuzzy C-Means),基于粗糙集理论对目标函数进行改进,从而改变隶属度矩阵的分布。实验表明:算法的收敛性较好,对初始聚类中心和初始隶属度矩阵的敏感度较低,计算量较 FCM 算法小,对边界元素的分辨率较高。

1 预备知识

定义 1[6] 给定信息系统 $S=(U,C\cup\{d\},D)$,$[x]_B$ 为 U 上由等价关系构成的与等价的元素构成的集合,设 $X\subseteq U$ 是一组对象,$B\subseteq C$ 是一组属性,X 相对于 B 的下近似、上近似、边界和负域分别定义如下:

* 收稿日期:2006 - 11 - 09
基金项目:国防科技大学资助项目(JC03 - 02 - 003)
作者简介:王丹(1981—),男,助教,硕士。

$\underline{R}X = \{x \in U : [x]_B \subseteq X\}$,$\overline{R}X = \{x \in U : [x]_B \cap X \neq \phi\}$

$Bn(X) = \overline{R}X - \underline{R}X$,$Neg(X) = \{x \in U : x \notin \overline{R}X\}$

下近似表示肯定属于某个集合的对象构成的集合,上近似表示可能属于某个集合的对象构成的集合,负域表示一定不属于某个集合的对象构成的集合。

令 $X = \{x_1, x_2, \cdots, x_n\} \subset R^p$ 是模式空间 R^p 中的一个有限数据集,$x_k = (x_{k1}, x_{k2}, \cdots, x_{kp}) \in R^p$ 称为特征矢量或模式矢量(或简单称为对象),x_{kj}为模式矢量 x_k 的第 j 个特征(属性),c 为类的数目,u_{ij}表示第 j 个对象对于第 i 个类的隶属程度,模糊 C – 均值算法的迭代计算公式为[1]:

$$v_i = \sum_{j=1}^{N} u_{ij}^m x_j \Big/ \sum_{j=1}^{N} u_{ij}^m, \quad i = 1,2,\cdots,c, u_{ij} = 1 \Big/ \sum_{k=1}^{c} \left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{m-1}}$$

从上述迭代公式可以看出,$x_j \in w_i$ 的隶属程度与所有的类(质心)都相关,这与实际情况似乎有点不相符,有些对象可能完全不属于某些类,那么计算这些对象的隶属程度时,这些类就不应该起作用。FCM 算法的目标函数是计算所有对象对每一个类的距离和的极值,实际情况中,对于任何一个类不是每一个对象都是属于它的,有些对象完全不属于这个类,在 FCM 算法中,这些对象对算法起了干扰作用,同时这也是 FCM 算法对初始聚类中心,初始隶属矩阵、参数 m,类数目敏感的原因。FCM 算法的隶属度矩阵分布遍及到整个对象空间,使得计算的复杂度也增大了。

2　粗糙模糊 C – 均值算法

定义 2　设 $X = \{x_1, x_2, \cdots, x_n\}$ 为待分类对象集,对于第 i 类 w_i,其质心为 v_i,类的数目为 c,定义 $\underline{R}w_i = \{x_j \mid x_j \in w_i\}$,$\overline{R}w_i = \{x_j \mid \| x_j - v_i \| \leqslant A_i, A_i > 0\}$,则

(1)若 $x_j \in \underline{R}w_i$,则 $\forall k \in \{1,2,\cdots,c\}$,$k \neq j$,$x_j \in \overline{R}w_i$,$x_j \notin \overline{R}w_k$;

(2)若 $x_j \in \overline{R}w_i$,则至少存在 $k \in \{1,2,\cdots,c\}$,使得 $x_j \in \overline{R}w_k$。

其中 A_i 称为上近似限,上近似限刻画了所有可能属于第类的对象的边界,若某个对象不属于上近似限所界定的范围,则它属于这个类的负域,即完全不属于这个类。

定义 3　粗糙模糊 C – 均值算法的目标函数为:

$$J_m(U,V) = \sum_{j=1}^{N} \sum_{i=1}^{c} \sum_{x_j \in \overline{R}w_i} u_{ij}^m d_{ij}^2$$

约束条件为:　$u_{ij} \in [0,1]$　$0 < \sum_{j=1}^{n} u_{ij} < N$　$\sum_{i=1}^{c} \sum_{x_j \in \overline{R}w_i} u_{ij} = 1$

同样可以得到粗糙模糊 C 均值算法的迭代公式:

$$u_{ij} = 1 \Big/ \sum_{k=1}^{c} \sum_{x_j \in \overline{R}w_k} \left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{m-1}} \tag{1}$$

质心计算公式不变:

$$v_i = \sum_{j=1}^{N} u_{ij}^m x_j \Big/ \sum_{j=1}^{N} u_{ij}^m \tag{2}$$

从 RFCM 算法很容易得到以下两个性质:

(1)若 $x_j \in \underline{R}w_i \Leftrightarrow u_{ij} = 1$;(2) $u_{ij}^{\mathrm{FCM}} \leqslant u_{ij}^{\mathrm{RFCM}}$。

下面具体给出 RFCM 算法的步骤。

(1)确定类数 $c(2 \leqslant c \leqslant N)$,参数 m,初始矩阵,类的上近似边界 A_i 和一个适当小数,$s = 0$

(2)按式(2)计算质心 $\{v_i^{(s)}\}$

(3)若 $x_j \notin \overline{R}w_i$ 则 $u_{ij} = 0$,否则按式(1)更新 $u_{ij}^{(s)}$

(4)若 $\| U^{(s)} - U^{(s+1)} \| < \varepsilon$,则停止,否则,$s = s + 1$,转(2)

RFCM 算法的主要思想是把属于某个类的对象分成了肯定的、可能的和否定的三个集合,以所有可能的对象的最小类内平方误差和为聚类准则进行聚类。RFCM 算法和 FCM 算法最大的不同在于,它认为 x_j 属于 w_i 的隶属度 u_{ij} 的计算只与上近似中包含 x_j 的类有关,若某个类 w_k 的上近似中不包含 x_j,则这

个类对 x_j 的隶属度是没有任何贡献的。

Doğan zdemir 在文献[3]给出了一个目标函数 $J_m = \sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}^m d_{ij}^2 - \alpha \sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}^m$，给出了 u_{ij} 的更新公式为：$u_{ij} = 1 \Big/ \sum_{k=1}^{c} \left(\frac{d_{ij}^2 - \alpha}{d_{kj}^2 - \alpha}\right)^{\frac{1}{m-1}}$，根据此公式，以 v_i 为中心半径为 α 的超球体内的 u_{ij} 为1。Doğan zdemir 所做的工作实际上和本文的思想很相近，相当于给出了一个下近似限，肯定了一些对象，以此改进隶属度矩阵的分布，但这种改进没有涉及 u_{ij} 的本质。同时，由于各个类分布的不同，对所有的类`使用相同的下近似限 α 也存在着不合理的地方，Doğan 在文中也没有给出 α 的几何意义以及怎么选取。

3 关于上近似限 A_i 的讨论

FCM 算法还有一个缺点就是没有考虑到对象的客观分布情况，而对象或图像的分布情况恰恰可以由对象的频率分布图(直方图)清晰的表示出来，在聚类算法中考虑频率分布图的分布情况能更好的进行聚类。在 RFCM 算法中，利用频率分布图的信息是通过上近似限 A_i 表示的，上近似限的几何意义是各个聚类在频率分布图中横轴(灰度级)上跨越的宽度(简称跨度)。当上近似限取得足够大时，RFCM 算法退化为 FCM 算法或改进的 FCM 算法。由于不同图像有不同的频率分布图，不同数据集的分布不同，得到 A_i 理论上的计算公式是困难的一件事情，只能采取一些经验的方法来确定上近似限。通过分析数据或图像的频率直方图，对每个聚类寻找适当的上近似限，上近似限取值宜大不宜过小，上近似限取得过小，使得聚类错误率过高。同时不同类之间的上近似限应该尽量的有差别，上近似限不同才能使得不同的类分开的可能性增大。这样做可以减小对初始隶属度矩阵的依赖。另外，在聚类过程中，定义一种新的粗糙度的概念来对上近似限进行调整。

定义4 设数据集 X，数据集中的元素为 $x_i(i=1,2,\cdots,n)$，则关于数据集的粗糙度定义如下：

$$\gamma_X = \frac{\sum_{1\leqslant i,j\leqslant n} s_{ij}}{n(n-1)/2}$$

其中，$s_{ij} = \begin{cases} 1, & |x_i - x_j| \leqslant A_x \\ 0, & |x_i - x_j| > A_x \end{cases}$，$A_x$ 为数据集 X 的上近似限。

显然，γ_X 的意义就是对于数据集 X 中任意两点，计算其距离，若 γ_X 大于一定的比例，比如95%，则可以认为 A_x 比较合适，否则应该调整，若 γ_X 过小，说明 A_x 比较小，应该增大，若 γ_X 过大，则说明 A_x 还可以继续减小。在聚类过程中，根据粗糙度的值来动态调整 A_i 的大小。在这里，我们给出了一个动态调整 A_i 的 RFCM 算法，主要思想就是根据 RFCM 算法的每次迭代过程中，计算每个类的粗糙度的大小，若达到1，则上近似限以一定的步长减小，若粗糙度小于90%，则上近似限以一定步长增大。

(1)确定类数 $c(2\leqslant c\leqslant N)$，参数 m，初始矩阵，类的上近似边界 A_i、上近似限变化步长 δ 和一个适当小数，$s=0$

(2)按式(2)计算质心 $\{\bar{v}_i^{(s)}\}$

(3)若 $x_j \notin \bar{R}w_i$，则 $u_{ij}=0$，否则按式(1)更新 $u_{ij}^{(s)}$

(4)根据 $u_{ij}^{(s)}$ 和式(3)计算每个类的粗糙度，若对于类 c_i，若 $\gamma_x = 1$，则 $A_i = A_i - \delta$，若 $\gamma_x < 90\%$，则 $A_i = A_i + \delta$，转(2)，否则转(5)

(5)若 $\| U^{(s)} - U^{(s+1)} \| \leqslant \varepsilon$，则停止，否则，$s = s+1$，转(2)

4 实验及结果分析

下面我们将 RFCM 算法与 FCM 算法和 Doğan zdemir 改进的 FCM 算法应用到图像聚类中进行比较，同时考虑参数的变化对聚类的影响。

(1) RFCM 算法、FCM 算法、Doğan 改进的 FCM 算法的质量比较

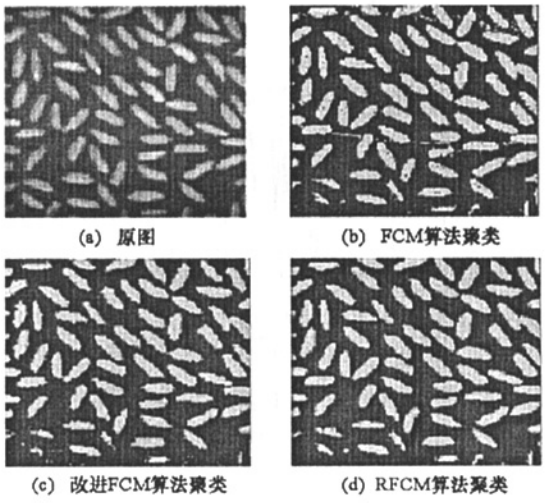

图 1　聚类效果比较

Fig. 1　Effect of clustering

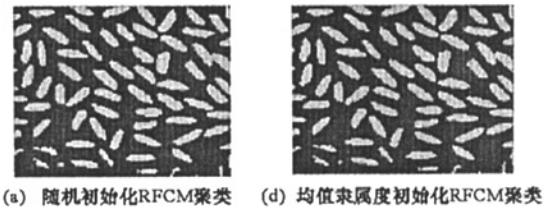

图 2　初始聚类中心影响

Fig. 2　Influence of initial clustering center

分别用 FCM 算法、Doğan 改进的 FCM 算法和 RFCM 算法对图像进行聚类,聚类结果如图 1(a) ~ (d)所示。可见,FCM 算法聚类的效果较差,分类错误率较高,Doğan 改进的 FCM 算法聚类效果较 FCM 算法好,但注意在不同类的边缘分辨率不高,RFCM 算法改进了上述缺点,类边缘平滑性较好,分类清楚,是一种较好的聚类算法。

(2) 聚类的时间复杂度比较

从理论上来说,由于 RFCM 算法需要计算的对象数目较 FCM 算法算法少,所以需要的计算时间也相应的少一些。表 1 显示了在同一台机器、同样的环境下,FCM 算法、Dofian 改进的 FCM 算法、RFCM 算法对上面图像进行聚类所需要的时间。

表 1　三种算法的计算时间比较

Tab. 1　Running time of 3 algorithm

	FCM 算法	Doğan 改进的 FCM 算法	RFCM 算法
计算时间(s)	19.334	23.4530	18.797
单次迭代的平均时间(s)	6.448	5.8632	6.2657

(3) 初始聚类中心和初始隶属度矩阵的影响

为了检验算法对聚类中心和初始隶属度矩阵的敏感度,采用随机生成初始隶属度矩阵和取均值隶属度矩阵为 $u_{ij}=\frac{1}{c}, i=1,2,\cdots,c, j=1,2,\cdots,N$ 对算法进行检验,图 2 为采用上述两种初始方法后,RFCM 的聚类结果。通过均值隶属度初值化方法和 50 次的随机初始化方法,用 RFCM 算法对图像进行聚类发现,算法稳定性较好,均能达到较好的聚类效果。图 2(a)、(b)的聚类效果与图 1(d)(对初始聚类中心进行了刻意的取定)进行比较发现,RFCM 算法对初始聚类中心和隶属度矩阵的要求非常低,但聚类效果却比较好。FCM 算法和 Doğan 改进的 FCM 算法对初始聚类中心和初始隶属度矩阵要求较高。

(4) 关于参数 m 的敏感性

从图 3(a)、(b)可以看出 FCM 算法随 m 的变化产生的效果不同,当 $m=9$ 或 $m=10$ 时 FCM 算法陷入了局部最优点,而 RFCM 算法能保持比较好的稳定性。而且对于不同 m 的值,可以通过调节上近似限达到相同的效果。

(5) 对于高维数据进行聚类的算法性能实验

对于探察 RFCM 算法在高维聚类上的表现,设计两种实验来考察:一种是任意给定初始隶属度矩阵,考察动态调整上近似限情况下 RFCM 算法的性能和 FCM 算法的性能比较;一种是固定调整好的上近似限,我们来看 RFCM 算法随初始隶属度变化的情况。在实验过程中为了更好地考察方法的优劣性,就取基于 L_p 范数的一般距离度量来衡量对象之间的相似程度,下面是实验结果。

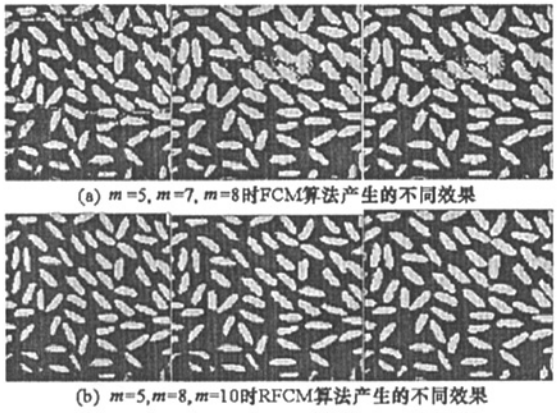

图3 参数 m 敏感性实验

Fig. 3 Sensibility of m

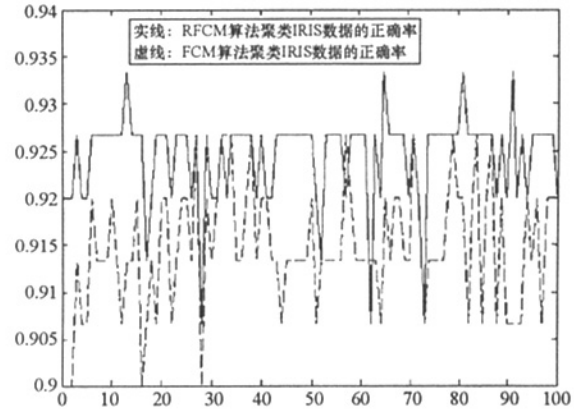

图4 高维数据 FCM 与 RFCM 比较

Fig. 4 High-dimension data clustering of FCM and RFCM

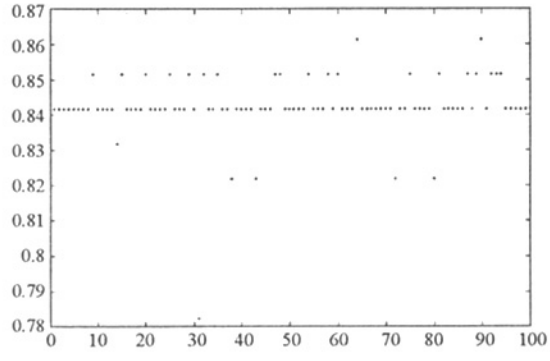

图5 高维数据下 FCM 聚类正确率

ig. 5 Accuracy of high-dimension data clustering of FCM

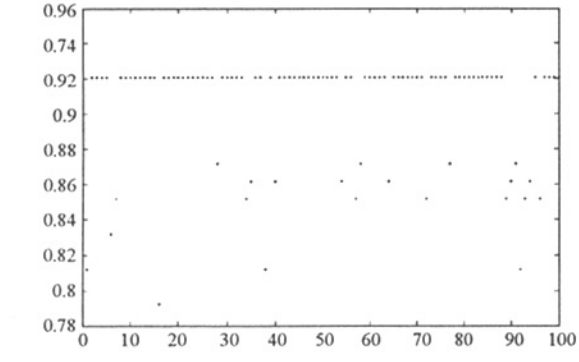

图6 高维数据下 RFCM 聚类正确率

Fig. 6 Accuracy of high-dimension data clustering of RFCM

图4是对任意给定的初始隶属度矩阵考察 RFCM 算法和 FCM 算法的性能比较，这里采用的数据集是 IRIS 数据集，从图中可以明显看出 RFCM 算法的聚类性能优于 FCM 算法。图5是 FCM 算法对于随机生成 100 个隶属度矩阵的聚类正确率，图6是 RFCM 算法取定上近似限的情况下对随机生成 100 组初始隶属度矩阵的聚类正确率，比较图5和图6可以发现，RFCM 算法的聚类性能比较稳定，平均情况下比 FCM 算法要好，在对 RFCM 算法的实验中，我们取定 7 个类的上近似限分别为：4.7，2.2，4，1.8，1.45，1.6，8.23，这是首先通过上近似限的动态调整得到的一个结果。

5 结论

本文基于粗糙集的思想研究模糊 C－均值算法，粗糙模糊 C－均值算法结合粗糙集上、下近似的概念对模糊 C－均值算法进行了质的改进，通过实验，发现该算法相对于 FCM 算法和 Doğan 改进的 FCM 算法，具有时间复杂性低、精度高、对初始聚类中心和初始隶属度矩阵敏感度低的特点。但 RFCM 算法引进了称为上近似限的参数，若上近似限取得过窄，会使得聚类结果中类过小，上近似限太宽，算法演变为 FCM 算法和改进的 FCM 算法，在理论上，对上近似限的取法如何进行研究是进一步研究的方向。

参考文献：

[1] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[J]. New York：Plenum，1981.
[2] Bezdek J C. Convergence Theory for Fuzzy c-Means：Counterexamples and Repairs[J]. IEEE Transactions on Systems，man，and cybernetics，1987，SMC－17(5).
[3] zdemir D，Akarun L. A Fuzzy Algorithm for Color Quantization of Images[J]. Pattern Recognition，2002，35：1785－1791.
[4] Tucker W T. Counterexamples to the Convergence Theorem for Fuzzy ISODATA Clustering Algorithms[M]. In the Analysis of Fuzzy Information，J. C. Bezdek，Ed. Boca Raton，FL：CRC Press，1987，3，CH 7.
[5] Duda R，Hart P. Pattern Classification and Scene Analysis[M]. New York：Wiley，1973.
[6] Pawlak Z. Rough Sets[J]. International Journal of Computer and Information science，1982(11)：241－356.
[7] 高新波，裴继红，谢维信. 模糊 c 均值聚类算法中加权指数 m 的研究[J]. 电子学报，2000，28(4).
[8] 徐月芳. 基于遗传模糊 C－均值聚类算法得图像分割[J]. 西北工业大学学报，2000，20(4).
[9] 于剑，程乾生. 关于 FCM 算法中的权重指数 m 的一点注记[J]. 电子学报，2003，31(3)