文章编号:1001 - 2486(2011)03 - 0145 - 06

粗糙模糊 C 均值融合聚类*

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 均值融合聚类算法,该算法通过粗糙集上、下近似的引入改变了模糊 C 均值算法中隶属度函数的分布情况,修正了类心的更新公式和模糊隶属度计算公式,降低了计算复杂度,在改变模糊隶属度分布的同时,通过使得每一类总的隶属度变化保持最小,进一步提出了边界调节参数的自适应选择算法,实验结果表明,粗糙模糊 C 均值融合算法具有较好的效果。

关键词:模糊 C 均值聚类;粗糙集;粗糙模糊 C 均值聚类
中图分类号:O159 **文献标识码**:A

Rough Fuzzy C-Means Combination Clustering

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

Abstract: A deep-seated rough fuzzy C-Means combined (RFCMC) clustering algorithm is proposed. The algorithm alters the distribution of fuzzy membership function by combining the lower approximation and upper approximation. Accordingly, the computation of class centroid and fuzzy membership is modified. Moreover, a self-adaptive adjusting edge parameter algorithm is prepresented. The results from experiments prove the improved effects.

Key words: fuzzy C-Means clustering; rough sets; rough fuzzy C-Means clustering

众所周知,C 均值聚类是聚类算法中最经典的算法。1981 年,Bezdek 改进了经典 C 均值算法,提出模糊 C - 均值(Fuzzy C-Means)聚类算法(FCM)[2],此算法以最小类内平方误差和为聚类准则,与 C 均值聚类算法不同之处在于不将样本分成分明子集,而是计算每个样本属于各模糊子集(聚类)的隶属度,通过目标函数极小化来获得最优的聚类。该算法提出后在图像的分割、压缩、识别等领域得到了广泛应用。2002 年,Lingras 等人提出了粗糙 C 均值算法[3],其基本思想是把粗糙集理论中上、下近似引进到了 C 均值算法中,在一个类的下近似中的对象肯定是属于这个类的,而位于边界(上近似与下近似的差集)的对象,在粗糙 C 均值算法中认为由于信息的缺乏而不能明确判断,即在粗糙 C 均值算法中承认了类之间存在重叠。在模糊 C 均值算法中,模糊隶属度函数的引入实际上也承认了一定程度的重叠,但在处理这部分重叠时,模糊 C 均值是用精确的方式来刻画模糊性。2006 年,Mitra 进一步将模糊 C 均值算法与粗糙 C 均值聚类算法结合起来,提出了粗糙模糊 C 均值算法[4],该算法通过在类心更新公式中对属于上、下近似中对象分别赋予不同权重来获得新的类心,2007 年,王丹也将模糊 C 均值算法结合粗糙集进行了改进[5],将类心更新公式中对所有对象的计算改进为只对上近似中对象的计算,从而获得更加准确的类心,但两种算法都分别从不同角度引入了边界调节参数,而此参数的设置成为算法应用的瓶颈。从本质上来说,两种算法都是希望将模糊 C 均值中精确的模糊性刻画方式和粗糙 C 均值中边界刻画模糊性的方式结合起来进行聚类,进而提高算法的性能,Mitra 在这方面做了不少工作[6-7]。本文结合两种粗糙模糊 C 均值算法的设计不足,提出新的粗糙模糊 C 均值聚类算法,新的方法构造了边界调节参数 ε 的自适应选择算法。

本文组织如下:第 1 节介绍与本文相关聚类算法,包括模糊 C 均值算法、粗糙 C 均值算法和粗糙模糊 C 均值算法,第 2 节针对现有粗糙模糊 C 均值算法设计的不足,提出新的粗糙模糊 C 均值算法,并构造参数自适应的优化算法,在第 4 节

* 收稿日期:2010 - 10 - 18
基金项目:国家自然科学基金资助项目(60872152)
作者简介:王丹(1981—),男,讲师,博士生。

对算法性能进行实验,并对比了各种算法。

1 预备知识

1.1 模糊 C 均值算法(FCM)

令 $X=\{x_1,x_2,\cdots x_N\}\subset R^p$ 为待分类对象,$V=\{v_1,v_2,\cdots v_c\}$,v_i 为第 i 类的聚类中心,u_{ij} 为第 j 个对象属于第 i 个类的隶属度,权重 $m\in(1,\infty)$ 为模糊因子,$d_{ij}^2=\|x_j-v_i\|^2=\sum_{k=1}^{p}(x_{jk}-v_{ik})^2$ 为对象与类心的距离。模糊 C 均值算法通过不断调整类心和隶属度函数来进行聚类,其 v_i、u_{ij} 的迭代计算公式:

$$v_i=\sum_{j=1}^{N}u_{ij}^m x_j/\sum_{j=1}^{N}u_{ij}^m,\ i=1,2,\cdots,c \tag{1}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{m-1}}} \tag{2}$$

1.2 粗糙 C 均值算法(RCM)

在粗糙 C 均值算法中,上、下近似集被引入聚类中,聚类中的每个对象遵循以下原则:

(1)每个对象最多属于一个类的下近似;

(2)如果一个对象属于某一个类的下近似,一定属于这个类的上近似;

(3)如果一个对象不属于任一个类的下近似,那么一定属于至少两个类的上近似;

从上面原则可以看出粗糙 C 均值算法允许存在一定的聚类重叠。当然上面的 3 个条件并不是独立的。若再从类的角度来看这个问题,对于聚类后的每一个类,同样存在三种情况(1)只有下近似元素,(2)只有上近似元素,(3)既有下近似元素,也有上近似元素。根据这些原则,记 U_i 表示第 i 个聚类,$\underline{B}U_i$,$\overline{B}U_i$ 分别表示类 U_i 的下近似和上近似,边界则为 $\overline{B}U_i-\underline{B}U_i$,粗糙 C 均值算法的类心计算可调整为

$$v_i=\begin{cases}w_l\dfrac{\sum\limits_{x_j\in\underline{B}U_i}x_j}{|\underline{B}U_i|}+w_u\dfrac{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}x_j}{|\overline{B}U_i-\underline{B}U_i|} & \underline{B}U_i\neq\phi,\overline{B}U_i-\underline{B}U_i\neq\phi\\ \dfrac{\sum\limits_{x_j\in\underline{B}U_i}x_j}{|\underline{B}U_i|} & \underline{B}U_i\neq\phi,\overline{B}U_i-\underline{B}U_i=\phi\\ \dfrac{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}x_j}{|\overline{B}U_i-\underline{B}U_i|} & \underline{B}U_i=\phi,\overline{B}U_i-\underline{B}U_i\neq\phi\end{cases} \tag{3}$$

上式中 $|\cdot|$ 表示集合的基数(即集合中元素的个数),w_l,w_u 分别为下近似、上近似所占的权重,且 $w_l+w_u=1$,一般情况下 $0.5\leqslant w_l<1$,即在决定类的类心时下近似起的作用应该更大一些。粗糙 C 均值算法的步骤如下:

(1)选定 c 个初始聚类中心 v_i,$i=1,2,\cdots,c$;

(2)决策每一个对象 x_j 所属的分类,决策规则如下:

(a)计算 d_{kj},$k=1,2,\cdots,c$,若 $d_{ij}=\min\limits_{1\leqslant k\leqslant c}d_{kj}$;

(b)若存在类 U_l,使得 d_{lj} 满足:$|d_{lj}-d_{ij}|<\varepsilon$,则 $x_j\in\overline{B}U_i$ 且 $x_j\in\overline{B}U_l$ 否则 $x_j\in\underline{B}U_i$;

(3)按照公式(3)更新类心;

(4)重复(2)~(3)步骤,直到收敛,这里收敛的条件可以取类心不再发生变化或取某个目标达到最优,比如构造类内距离与类间距离的某个函数。

从式(3)可以看出,若 $\underline{B}U_i=\overline{B}U_i$,粗糙 C 均值算法则转化为经典 C 均值算法。注意到,在粗糙 C 均值算法聚类过程中,有 3 个参数 w_l,w_u,ε 要先给出来。Mitra 指出参数的选择是粗糙 C 均值算法最大的挑战。

1.3 粗糙模糊 C 均值算法(RFCM)

粗糙模糊 C 均值算法将上、下近似的概念引入模糊 C 均值算法中,在模糊 C 均值的类心更新公式中,通过加权的方式对下近似和上近似中的对象分别进行模糊化计算。粗糙模糊 C 均值算法的迭代公式如下:

$$v_i=\begin{cases}w_l\dfrac{\sum\limits_{x_j\in\underline{B}U_i}u_{ij}^m x_j}{\sum\limits_{x_j\in\underline{B}U_i}u_{ij}^m}+w_u\dfrac{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m x_j}{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m} & \underline{B}U_i\neq\phi,\overline{B}U_i-\underline{B}U_i\neq\phi\\ \dfrac{\sum\limits_{x_j\in\underline{B}U_i}u_{ij}^m x_j}{\sum\limits_{x_j\in\underline{B}U_i}u_{ij}^m} & \underline{B}U_i\neq\phi,\overline{B}U_i-\underline{B}U_i=\phi\\ \dfrac{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m x_j}{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m} & \underline{B}U_i=\phi,\overline{B}U_i-\underline{B}U_i\neq\phi\end{cases} \tag{4}$$

隶属程度 u_{ij} 更新同式(2)。

粗糙模糊 C 均值算法的步骤如下:

(1)选定初始隶属度 u_{ij},$i=1,2,\cdots,c$,$j=1,2,\cdots,N$;

(2)决策每一个对象 x_j 所属的分类,决策规则如下:

(a)对于 $u_{kj}(k=1,2,\cdots,c)$,若 $u_{ij}=\min\limits_{1\leqslant k\leqslant c}u_{kj}$;

(b)若存在类 U_l,使得 d_{lj} 满足:$|u_{lj}-u_{ij}|<\varepsilon$,则 $x_j\in\overline{B}U_i$ 且 $x_j\in\overline{B}U_l$ 否则 $x_j\in\underline{B}U_i$;

(3)按照式(3)更新类心;

(4)按照式(2)更新 u_{ij};

(5)重复(2)~(4)步骤直到收敛,这里收敛的条件可以取隶属度函数不再发生变化或变化很小。

2　粗糙模糊 C 均值融合算法(RFCMC)

Mitra 提出的粗糙模糊 C 均值算法中通过引入参数 w_l,w_u 来刻画聚类过程中下近似和上近似的重要程度,这实际上是粗糙 C 均值算法的简单套用。粗糙模糊 C 均值算法没有从本质上把粗糙集的精髓用进来。从粗糙集的本质来说,下近似中的对象的意义是肯定属于某个类的对象,在粗糙模糊 C 均值算法更新类心公式中,下近似对象的计算依然乘以隶属度因子 u_{ij}^m 也不符合粗糙集的思想,所以 Mitra 的粗糙模糊 C 均值算法并没有从本质上融合模糊 C 均值算法和粗糙 C 均值算法。

沿用上面的记号,本文中,将粗糙模糊 C 均值算法的类心和隶属函数更新公式分别修改为式(5)和(6):

$$v_i=\begin{cases}w_l\dfrac{\sum\limits_{x_j\in\underline{B}U_i}x_j}{|\underline{B}U_i|}+w_u\dfrac{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m x_j}{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m} & \underline{B}U_i\neq\phi,\overline{B}U_i-\underline{B}U_i\neq\phi\\ \dfrac{\sum\limits_{x_j\in\underline{B}U_i}x_j}{|\underline{B}U_i|} & \underline{B}U_i\neq\phi,\overline{B}U_i-\underline{B}U_i=\phi\\ \dfrac{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m x_j}{\sum\limits_{x_j\in\overline{B}U_i-\underline{B}U_i}u_{ij}^m} & \underline{B}U_i=\phi,\overline{B}U_i-\underline{B}U_i\neq\phi\end{cases}\tag{5}$$

$$u_{ij}=\begin{cases}1 & x_j\in\underline{B}U\\ \dfrac{1}{\sum\limits_{k=1}^{c}\left(\dfrac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{m-1}}} & x_j\in\overline{B}U_i-\underline{B}U_i\\ 0 & x_j\notin\overline{B}U_i\end{cases}\tag{6}$$

上面的改进公式实际上是相当于在模糊 C 均值的基础上对隶属度函数进行了调整。在类的边界部分的对象依然采用模糊 C 均值算法中精确的隶属程度来刻画,而对于完全属于类的对象,采用粗糙集的下近似刻画方法,对于完全不属于类的对象,采用粗糙集的负域(上近似的补集)的刻画方法。这实际上是通过粗糙集的融合改变了模糊隶属度的分布情况。即将下近似中的对象的隶属度提高到 1,而将负域中的对象的隶属度降低到 0。对于这两个动作,若希望对于每个类的总的隶属程度能够保持不变,即对于某个类 U_i,通过提高 $\underline{B}U_i$ 中对象的隶属度到 1,降低 $U-\overline{B}U_i$ 中对象隶属度到 0 后,使得 $\sum\limits_{j=1}^{N}u_{ij}$ 保持变化最小,即保持了模糊 C 均值算法中每个类总的隶属程度变化不大,但通过这种方式的处理,已经使得模糊隶属度函数具有更清晰的边界了。这里通过 ε 的优化来完成这个思想。对于类 U_i:

- 总的提升的隶属度为 $\sum\limits_{x_j\in\underline{B}U_i}(1-u_{ij})$;
- 总的降低的隶属度为 $\sum\limits_{x_j\notin\overline{B}U_i}u_{ij}$;

所以应该选择 ε 使得

$$\delta(\varepsilon)=\min\left|\sum_{x_j\in\underline{B}U_i}(1-u_{ij})-\sum_{x_j\notin\overline{B}U_i}u_{ij}\right|\tag{7}$$

粗糙模糊 C 均值融合聚类算法流程如下:

(1)选定初始隶属度 u_{ij},$i=1,2,\cdots,c$,$j=1,2,\cdots,N$;

(2)对于 $\varepsilon\in[\varepsilon_{\min},\varepsilon_{\max}]$;

决策每一个对象 x_j 所属的分类,决策规则如下:

(a)对于 $u_{kj}(k=1,2,\cdots,c)$,若 $u_{ij}=\min\limits_{1\leqslant k\leqslant c}u_{kj}$;

(b)若存在类 U_l,使得 u_{lj} 满足:$|u_{lj}-u_{ij}|<\varepsilon$,则 $x_j\in\overline{B}U_i$ 且 $x_j\in\overline{B}U_l$ 否则 $x_j\in\underline{B}U_i$;

(3)根据式(7)计算 $\delta(\varepsilon)$,若 $\varepsilon_{op}=\underset{\varepsilon}{\operatorname{argmin}}\delta(\varepsilon)$(这里 $\underset{\varepsilon}{\operatorname{argmin}}\delta(\varepsilon)$ 表示的不是 $\delta(\varepsilon)$ 的最小值,而是使得 $\delta(\varepsilon)$ 达到最小的参数 ε 的值),则转步骤(4),否则转步骤(2);

(4)按照式(5)更新类心,按照式(6)更新 u_{ij};

(5)重复(2)~(4)步骤直到收敛,这里收敛的条件可以取隶属度函数不再发生变化或变化很小。

本文提出的粗糙模糊 C 均值融合聚类算法,从本质上改变了类心的计算方式,通过隶属度函数分布的调整,降低了算法运行的计算量,即使得原计算下近似类心时不再与隶属度函数作乘法运算,同时在隶属度函数的更新中对于非上近似的对象的隶属程度直接赋予 0,而不再进行计算。当然这个计算量的降低主要针对的是类心和隶属度函数的计算,实际上在寻找最优参数 ε 时,由于

要进行多次对象归属的判断,以此计算 $\delta(\varepsilon)$,这部分是增大了运算量的,但这对于自适应参数选择也是必然的,这部分的计算量主要跟最大、最小 ε 值有关,这部分增大的运算量约为 $O(KcN)$,这里 K 为由最大、最小 ε 值的宽度带来的一个总搜索次数(注意到 $\varepsilon\in[0,1]$,此宽度引起的步长不会带来计算量的剧增)。另一方面下近似中的对象不再包含隶属度的乘积运算,加快了算法的收敛速度,新的类心更新公式有利于算法更快地收敛,在融合粗糙集核心思想进入算法中的同时对 ε 构造了自适应调整算法,降低了算法运行时 ε 选择的困难。在 Mitra 的粗糙模糊 C 均值算法中 ε 的选择在实际操作时是很难把握的,本文算法改进了这方面的问题。

3 聚类效果指标

我们采用 Davies-Bouldin[4] 指标和 β 指标刻画聚类的效果。好的聚类一般具有较小的类内距离,较大的类间距离,即要求同一类的对象具有最大的相似性,不同类的对象具有最大的相异性。Davies-Bouldin 指数以此刻画聚类效果,β 指数通过各类对象关于数据中心的分布与类内距离比值来衡量聚类的效果,仍沿用上面的符号,DB 指数和 β 指数的计算如下:

$$\mathrm{DB}=\frac{1}{c}\sum_{k=1}^{c}\max_{k\neq j}\left\{\frac{S(U_k)+S(U_j)}{d(U_k,U_j)}\right\} \tag{8}$$

$$\beta=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2} \tag{9}$$

(8)式中,$S(U_i)=\frac{\sum_{x_j\in U_i}\|x_j-v_i\|}{|U_i|}$ 表示类内平均距离,此值越小越好,$d(U_k,U_j)$ 表示类间距离,即 $d(v_k,v_j)$,此值越大越好。DB 指数越小,聚类效果越好。(9) 式中 c 为类的个数,x_{ij} 表示第 j 个对象属于第 i 类。$\bar{x}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}x_{ij}}{N}$ 为所有对象的中心,N 为对象的总个数,n_i 为第 i 类的对象个数,$\bar{x}_i=\frac{\sum_{j=1}^{n_i}x_{ij}}{n_i}$ 为第 i 类的类心。β 指数越大,聚类效果越好。

4 实验结果

本文中用两个实验来对算法性能进行比较。实验 1 采用人工生成的数据,实验 2 将算法应用在实际图像聚类中进行结果比较。

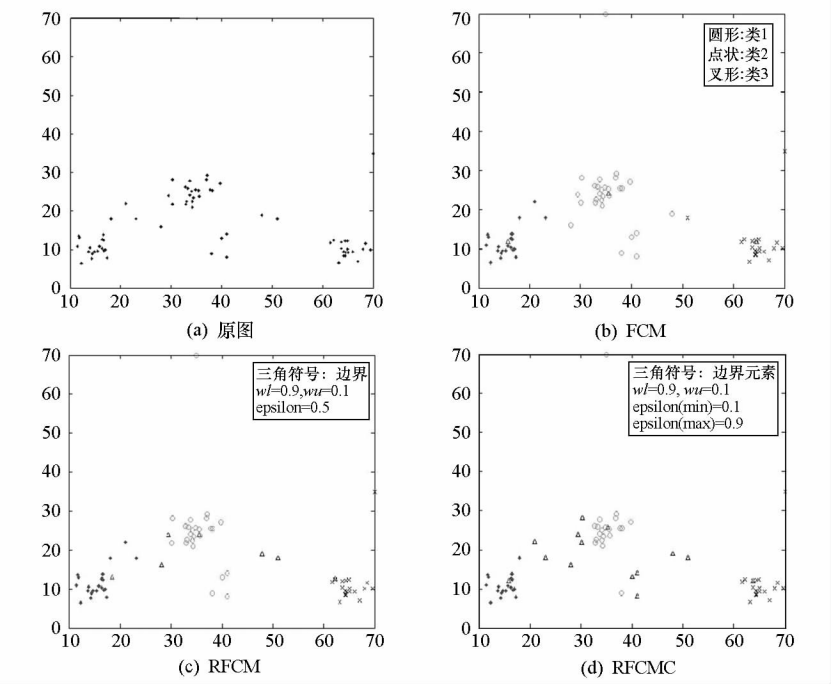

图 1 人工数据聚类结果

Fig.1 Results of clustering

图 1 显示人工数据聚类结果,从聚类结果来看,本文算法较好地分辨出了边界信息。聚类效果指标情况和算法运行情况如表 1 所示。

表 1　算法聚类效果

Tab.1　Clusterin effects of algorithms

算法	DB 指数	β 指数	迭代次数
FCM	0.235	28.689	71
RFCM	0.295	31.202	达到最大迭代次数
RFCMC	0.157	81.162	6

对于 Mitra 提出的粗糙模糊 C 均值算法,ε 的选择是不容易的事,图 2 中显示了 DB 指数和 β 指数随 ε 变化的情况,显然 ε 的选择对结果影响较大。在本文算法中,ε 的选择是自适应的,图 3 显示本文算法中 DB 指数和 β 指数随权重系数 w_l 的变化情况。

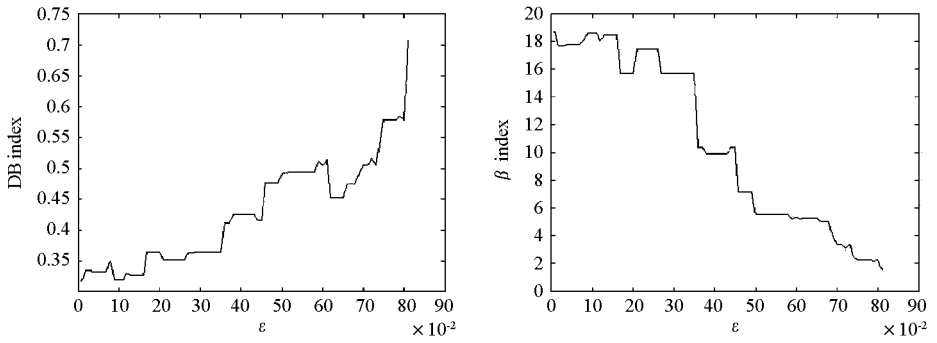

图 2　DB 指数和 β 指数关于 ε 变化曲线

Fig.2　Variation of DB index and β index by ε

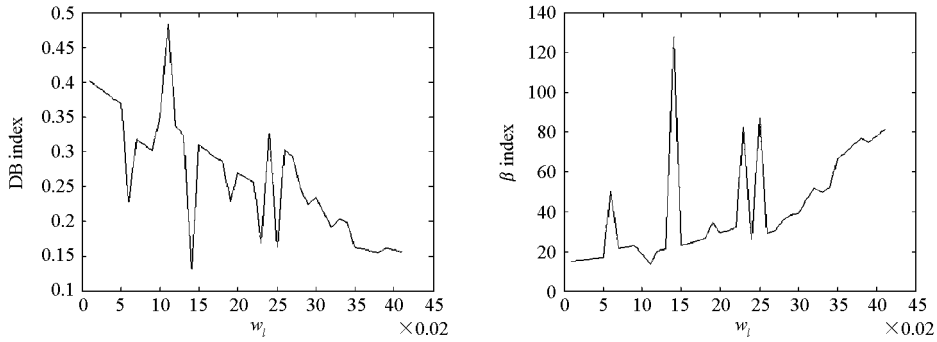

图 3　DB 指数和 β 指数关于 w_l 变化曲线

Fig.3　Variation of DB index and β index by w_l

从结果来看,选择较大的 w_l 将取得较好的效果。这主要是由于本文算法在改变模糊隶属度分布时实际上潜在地对下近似进行了强调,从这个层面上来说,本文算法也算是给出了参数 w_l 选择的一个指导方向。

将模糊 C 均值算法、粗糙模糊 C 均值算法和本文方法用在图像聚类上结果见图 4。

从结果来看,对于有噪声污染的图像,模糊 C 均值算法出现了误判,但其在抑制噪点上具有一定的优势,粗糙模糊 C 均值降低了误判的比例,本文方法能较准确判断分类,但也正由于其对边界的强调,噪点抑制不是很好,若再加上一些其它的信息能进行较好的判断,如对图像再平滑一次,即可消除这些脉冲噪点。

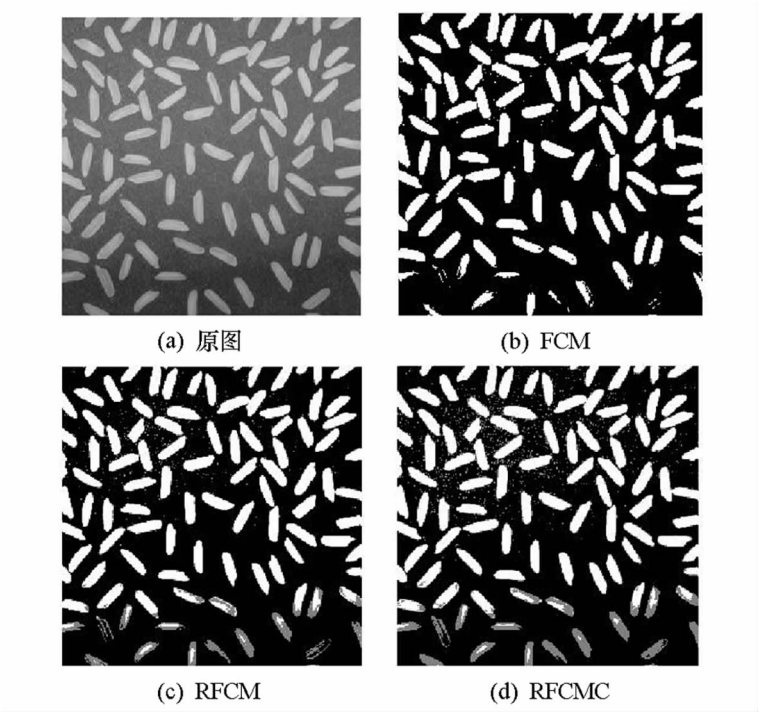

图 4 图像聚类结果比较
Fig.4 Effect of clustering

5 结 论

本文针对粗糙 C 均值算法和模糊 C 均值算法更深层次的融合,提出了新的粗糙模糊 C 均值融合算法,通过粗糙集核心思想的引入,该算法修正了类心的更新公式和模糊隶属度的计算公式,利用粗糙集上、下近似的引入改变了模糊隶属度函数的分布情况,降低了计算复杂度,在改变模糊隶属度分布的同时,通过使得每一类总的隶属度保持变化最小,进一步优化了参数 ε 的选择,由于算法对下近似的强调使得参数 w_l 应该选择较大的数值。

参 考 文 献:

[1] Pawlak Z. Rough Sets. International Journal of Computer and Information science[J]. 1982(11):241 – 356.

[2] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms [M]. New York: Plenum, 1981.

[3] Lingras P, West C. Interval Set Clustering of Web Users with Rough K-Means [R]. Technical Report 2002 – 002, Department of Mathematics and Computer Science, St. Mary's University, Halifax, Canada.2002.

[4] Mitra S, Banka H, Pedrycz W. Rough-Fuzzy Collaborative[J]. IEEE Transactions on Systems, Man, and Cybernetics-part B: Cybernetics, 2006(36), 4: 795 – 805.

[5] 王丹,吴孟达.粗糙模糊 C 均值算法及其在图像聚类中的应用[J].国防科技大学学报, 2007(2): 76 – 80.

[6] Mitra S, Barman B. Rough-Fuzzy Clustering: An Application to Medical Imagery[C]//RSKT, 2008:300 – 307.

[7] Mitra S, Banka H. Application of Rough Sets in Pattern Recognition[C]// Transactions on Rough Sets Ⅶ, 2007:151 – 169.