

基于自适应权重的 RFCM 聚类算法

张 强,王红卫,陈 游,王文哲

(空军工程大学 航空航天工程学院,陕西 西安 710038)

摘 要: 提出了基于自适应权重的 RFCM 聚类算法,主要思路是在每一次迭代中,根据每个对象与聚类中心欧式距离不同,选择改进的正半轴反余切函数对欧式距离进行重新分布后确定均衡因子,通过均衡因子动态地调节固定权重获得自适应权重,进而运行 RFCM 算法.最后,基于人工和 UCI 数据集的仿真验证了所提算法的有效性,并应用于雷达信号分选,验证了所提算法的实用性.

关键词: 聚类;自适应;权重;粗糙集;模糊集;均衡因子

中图分类号: TP18

文献标识码: A

文章编号: 1000—7180(2016)12—0080—05

Rough Fuzzy C-Means Clustering Algorithm Based on Self-adaptive Weights

ZHANG Qiang, WANG Hong-wei, CHEN You, WANG Weng-zhe

(Aeronautics and Astronautics Engineering College, Air Force Engineering University,X'ian 710038, China)

Abstract: This paper proposed RFCM clustering algorithm based on self-adaptive weights. According to different distance between every data object and clustering center, using improved arc cotangent function to redistribute the distance, acquiring adaptive weights through the equivalence factor in each iteration, then carrying out RFCM clustering algorithm. The simulations are performed on synthetic and UCI data sets, and the results show the validity of the proposed algorithm. Furthermore, the proposed algorithm is applied in sorting radar signal, and the results show the practicability of the proposed algorithm.

Key words: clustering; self-adaptive; weights; rough set; fuzzy set; equivalence factor

1 引言

聚类分析是数据挖掘的一种重要方法,广泛应用于图像识别、文本聚类、雷达信号分选等领域.聚类是指按照某种相似性将一组没有类别标记的对象划分成若干类^[1],并使得同一类内的数据对象是相似的,而不同类中的数据对象是不相似的^[2].

粗糙集和模糊集能够很好地处理不确定信息,诸多研究将两者用于改进传统聚类算法,并应用于聚类分析中.其中粗糙模糊 C 均值(RFCM)聚类算法^[3-4]等. RFCM 相比其他算法更加适合对复杂、非均匀分布的数据进行聚类分析,可以有效地处理聚

类边界模糊问题,而且能够更好地处理孤立点并减少聚类结果受噪声的影响,进而优化聚类结果. RFCM 聚类算法存在的问题是忽略了不同数据对象分布的不均衡性,对所有对象统一地赋予固定权重.比如在下近似集中,每个对象与聚类中心的距离不同,表明对象与中心的紧密程度不同^[5],若对所有对象均赋予相同的固定权重,会使聚类中心点在迭代的过程中有较大程度上的变化,严重影响聚类性能.

针对 RFCM 聚类算法的问题,提出一种基于自适应权重的 RFCM 聚类算法,根据同一类的数据对象对计算聚类中心影响不同,设计自适应的权重计算策略对每个对象计算权重.

2 RFCM 聚类算法

Lingrus^[6]提出了 RCM 聚类算法,主要思路是利用粗糙集的上近似集和下近似集的概念来改进 C 均值聚类算法. RFCM 聚类算法在 RCM 聚类算法的基础上加入了模糊隶属度权值,并以隶属度作为相异度判断准则. 目前, Mitra^[3]和 Maji^[4]分别提出两种聚类算法, Maji 提出的算法考虑了下近似集中的对象应与其他类无关,并调整了隶属度函数的分布,从而改变了聚类中心的更新方式和降低了算法的时间复杂度,故采用 Maji 提出的 RFCM 聚类算法.

令 $X = \{x_1, \dots, x_N\}$ 为样本数据集, x_j 表示一个数据对象, C_i 表示第 i 个类, $\overline{C}_i$ 和 $\underline{C}_i$ 分别表示类 i 的上近似集和下近似集; $V = \{v_1, \dots, v_k\}$, v_i 表示类 C_i 的聚类中心; u_{ij} 表示第 j 个对象属于第 i 个类的隶属度. RFCM 算法聚类中心计算公式如下:

$$v_i = \begin{cases} w_l A + w_u B, & C_i \neq \emptyset \wedge (\overline{C}_i - \underline{C}_i) \neq \emptyset \\ A, & C_i \neq \emptyset \wedge (\overline{C}_i - \underline{C}_i) = \emptyset \\ B, & C_i = \emptyset \wedge (\overline{C}_i - \underline{C}_i) \neq \emptyset \end{cases} \quad (1)$$

$$A = \frac{\sum_{x_j \in \underline{C}_i} x_j}{|\underline{C}_i|} \quad (2)$$

$$B = \frac{\sum_{x_j \in (\overline{C}_i - \underline{C}_i)} u_{ij}^m x_j}{\sum_{x_j \in (\overline{C}_i - \underline{C}_i)} u_{ij}^m} \quad (3)$$

式中, w_l 和 w_u 分别表示下近似集和上近似集的权重, 且有 $w_l > w_u$, $w_l + w_u = 1$, w_l 和 w_u 都是预先设定好的固定权重.

模糊隶属度函数更新公式如下:

$$u_{ij} = \begin{cases} 1, & x_j \in \underline{C}_i \\ u, & x_j \in \overline{C}_i - \underline{C}_i \\ 0, & x_j \notin \overline{C}_i \end{cases} \quad (4)$$

$$u = \frac{1}{\sum_{z=1}^k \left(\frac{d_{ij}}{d_{zj}} \right)^{\frac{2}{m-1}}} \quad (5)$$

式中, $d_{ij} = ||x_j - v_i||$ 和 $d_{zj} = ||x_j - v_z||$ 分别表示数据对象 x_j 与聚类中心 v_i 和 v_z 之间的欧式距离.

RFCM 聚类算法的具体步骤如下所示.

Step1 初始化. 选定总数为 N 的数据集, 设定目标聚类个数 k 、上近似集权值 w_u 、下近似集权值

w_l 、隶属度判断阈值 τ 、隶属度收敛阈值 ϵ 、模糊指数 m 和初始隶属度 u_{ij} , 计算初始聚类中心 v_i ;

Step2 对于数据对象 x_j , 计算其到各个类隶属度的最大值 u_{ij} ($1 \leq i \leq k$) 和次大值 u_{lj} ($1 \leq l \leq k$), 当 $u_{ij} - u_{lj} \leq \tau$ 时, 则 $x_j \in \overline{C}_i$, 且 $x_j \in \underline{C}_i$, 否则 $x_j \in \underline{C}_i$;

Step3 按照式(1) ~ (3) 更新聚类中心, 按照式(4)、(5) 更新模糊隶属度函数;

Step4 重复 Step2 和 Step3, 直至 $u_{ij} < \epsilon$ 满足收敛条件.

3 基于自适应权重的 RFCM 聚类算法

RFCM 算法更新聚类中心时采用由以往经验而定的固定权重, 但固定权重通常会忽略数据分布的不平衡性, 并无法考虑迭代过程中聚类中心位置的变化, 进而影响聚类算法性能. 因此基于数据对象的分布, 提出了自适应的权重的 RFCM 聚类算法: 在聚类过程中, 同一个类中的样本与聚类中心之间的距离并不相同, 即对类心计算的影响也有所差异, 根据当前的数据分布状态确定均衡因子, 在每一次迭代过程中通过均衡因子动态地调整权重, 从而降低了 RFCM 算法对初始权重的依赖.

3.1 均衡因子

均衡因子用来描述不同数据对象在所属类分布的均衡程度, 均衡因子值越小, 均衡程度越小, 即不均衡程度越大. 距离是数据对象在类内是否均衡分布最显著的度量方式, 因此选择距离计算平衡因子: 当数据对象距离聚类中心越远时, 说明均衡程度越小, 因此均衡因子越小; 当数据对象距离聚类中心越近时, 说明均衡程度越大, 因此均衡因子越大. 但不能直接利用距离进行计算, 这是因为距离值通过归一化处理后差异变得很小, 且呈线性分布, 无法很好地体现出不同数据对象之间的差异. 为了更好地刻画数据分布的均衡程度, 需要对距离值进行重新分布后再计算均衡因子.

考虑到不同数据对象与聚类中心的欧式距离取值范围为 $[0, +\infty)$, 且欧式距离大小与均衡因子呈反比, 数据对象与聚类中心之间的欧式距离越大, 均衡因子越小. 所以需要寻找自变量取值范围包括 $[0, +\infty)$, 且在此取值范围上单调递减的函数. 改进的反余切函数 $y = 2/\pi \times \operatorname{arccot} x$ 正半轴部分, x 轴取值范围为 $[0, +\infty)$, y 轴取值范围为 $(0, 1]$, 在取值范围上单调递减, 能够很好地满足要求, 且计算简单. 所以采用改进的正半轴反余切

函数对距离值进行重新分布后计算均衡因子. 函数曲线如图 1 所示.

依据重新分布后距离计算均衡因子公式如下

$$M_{ij} = \frac{2/\pi \times \operatorname{arccot} d_{ij}^2}{\sum_{x_n \in C_i} (2/\pi \times \operatorname{arccot} d_{in}^2)}$$

(6)

式中, M_{ij} 为数据对象 x_j 的均衡因子, x_j 和 x_n 是类 C_i 内的数据对象, $d_{in} = \|x_n - v_i\|$ 和 $d_{ij} = \|x_j - v_i\|$ 分别表示数据对象 x_n 和 x_j 与所在聚类中心 v_i 之间的欧式距离.

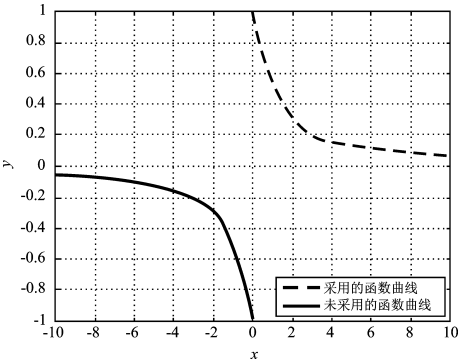


图 1 改进的反余切函数曲线

采用均衡因子能够动态地调整固定权重获得自适应权重. 当数据对象距离聚类中心越远时, 均衡因子越小, 采用均衡因子调整固定权重, 可以获得小于固定权重的自适应权重; 当数据对象距离聚类中心越近时, 均衡因子越大, 采用均衡因子调整固定权重, 可以获得大于固定权重的自适应权重. 自适应权重的大小, 反映了数据对象和聚类中心的紧密程度.

3.2 基于自适应权重的 RFCM 聚类算法

基于自适应权重的 RFCM 聚类算法聚类中心迭代公式如下:

$$v_i = \begin{cases} M_{ij}w_lA + M_{ij}w_uB, & \underline{C_i} \neq \varnothing \wedge (\overline{C_i} - \underline{C_i}) \neq \varnothing \\ A, & \underline{C_i} \neq \varnothing \wedge (\overline{C_i} - \underline{C_i}) = \varnothing \\ B, & \underline{C_i} = \varnothing \wedge (\overline{C_i} - \underline{C_i}) \neq \varnothing \end{cases}$$

(7)

式中, 当 $x_j \in \underline{C_i}$, 均衡因子为 M_{ij} ; 当 $x_j \in (\overline{C_i} - \underline{C_i})$, 均衡因子为 M'_{ij} . M_{ij} 和 M'_{ij} 计算公式相

同, 只是分别选取各自区域数据对象进行计算, 所以采用两种表示进行区别.

基于自适应权重的 RFCM 聚类算法的步骤如下.

Step1: 初始化. 选定总数为 N 的数据集, 设定目标聚类个数 k 、上近似集权重 w_u 、下近似集权重 w_l 、隶属度判断阈值 τ 、隶属度收敛阈值 ϵ 、模糊指数 m 和初始隶属度 u_{ij} , 计算初始聚类中心 v_i ;

Step2: 对于数据对象 x_j , 计算其到各个类隶属度的最大值 $u_{ij} (1 \leq i \leq k)$ 和次大值 $u_{lj} (1 \leq l \leq k)$, 当 $u_{ij} - u_{lj} \leq \tau$ 时, 则 $x_j \in \underline{C_i}$, 且 $x_j \in \overline{C_l}$, 否则 $x_j \in \overline{C_i}$;

Step3: 对于数据对象 x_j , 按照式(6) 计算均衡因子;

Step4: 按照式(2)、(3)、(7) 更新聚类中心, 式(4)、(5) 更新模糊隶属度函数;

Step5: 重复 Step2、Step3 和 Step4, 直至 $u_{ij} < \epsilon$ 满足收敛条件.

4 仿真实验分析

实验中分别采用人工和 UCI 数据集测试算法性能, 采用 DB^[7] 和 XB^[8] 指标评价算法的有效性, 采用 Rand^[9] 指标评价结果的准确性. DB 和 XB 越小表示聚类结果越好; Rand 越大表示聚类准确率越高.

采用 RCM 聚类算法、RFCM 聚类算法和基于自适应权重的 RCFM(实验中用 SRCFM 标记) 聚类算法进行聚类实验. 3 种算法在实验中所采用的聚类参数如表 1 所示.

4.1 人工数据集聚类实验

实验中采用人工合成的数据集, 该数据集由线性分布的 3 个聚类组成, 每个聚类含有 40 个样本点, 如图 2 所示.

分别采用 3 种算法在人工数据集上进行聚类实验, 人工数据集聚类结果如表 2 所示, 可见 SRFCM 聚类算法 DB 指标和 XB 指标最小, Rand 指标最大, 聚类算法性能最好, 同时运行时间也最长, 这是因为确定均衡因子需要一定的时间.

表 1 聚类参数

算法	聚类个数	模糊指数	上近似集权重	下近似集权重	隶属度判断阈值	距离判断阈值	隶属度收敛阈值	距离收敛阈值
RCM	3		0.1	0.9		0.2		0.01
RFCM	3	2	0.1	0.9	0.2		0.01	
SRFCM	3	2	0.1	0.9	0.2		0.01	

表 2 人工数据集聚类结果				
算法	XB 指标	DB 指标	Rand 指标	运行时间/s
RCM	0.092 5	0.615 2	0.825	0.063
RFCM	0.079 7	0.524 5	0.875	0.087
SRFCM	0.075 3	0.463 5	0.908	0.112

图 3~5 分别直观地给出了 3 种算法的聚类结果.可以看出,SRFCM 聚类算法错误聚类点最小,而且使得靠近聚类中心的对象更加紧凑,类间区分更明显.

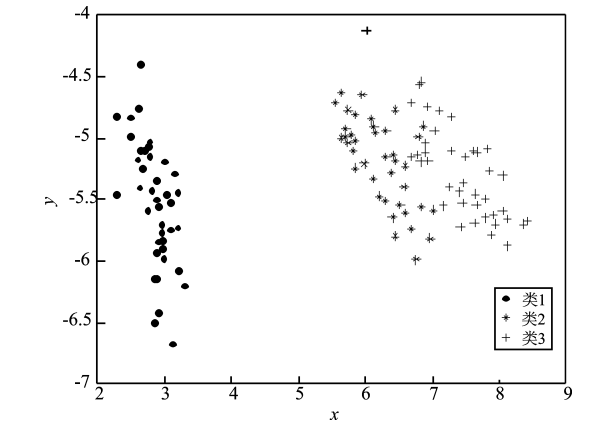


图 2 人工数据集

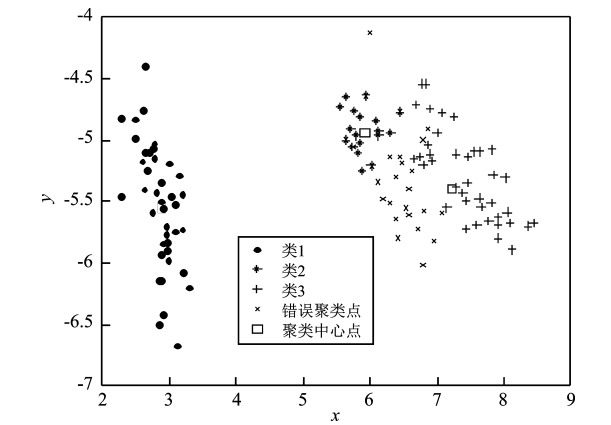


图 3 RCM 算法聚类结果

4.2 UCI 数据集聚类实验

实验中采用 UCI 数据集,分别是数据集 Iris 和 Wine^[17],具体如表 3 所示.

表 3 UCI 数据集基本信息			
数据集	类数	维数	样本点个数
Iris	3	4	150
Wine	3	13	178

分别采用 3 种聚类算法每个 UCI 数据集进行聚类实验,聚类结果如表 4 和表 5 所示. SRFCM 聚类算法取得了比 RCM 和 RFCM 更有效的聚类结

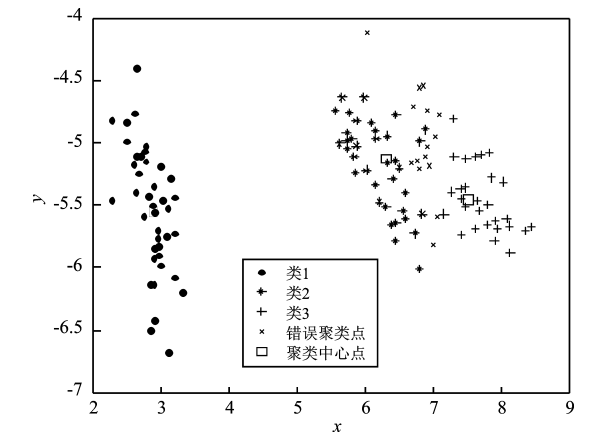


图 4 RFCM 算法聚类结果

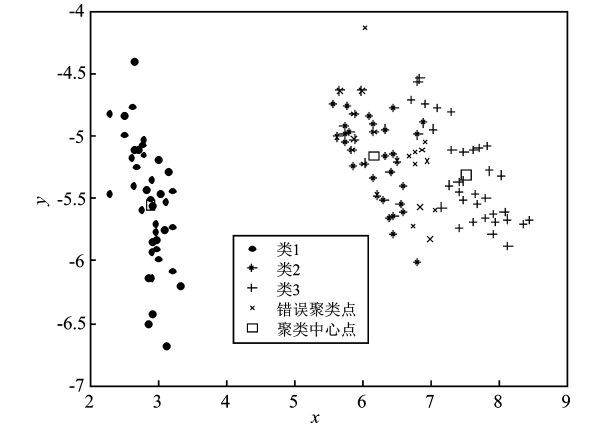


图 5 SRFCM 算法聚类结果

果,运行时间虽有增长,但并没有太大的牺牲.

表 4 Iris 数据集聚类结果				
算法	XB 指标	DB 指标	Rand 指标	运行时间/s
RCM	0.127 5	0.826 4	0.872	0.047
RFCM	0.121 8	0.635 3	0.879	0.119
SRFCM	0.114 1	0.598 5	0.883	0.145

表 5 Wine 数据集聚类结果				
算法	XB 指标	DB 指标	Rand 指标	运行时间/s
RCM	0.294 7	2.931 0	0.941	0.127
RFCM	0.261 8	2.145 0	0.947	0.156
SRFCM	0.252 3	1.865 5	0.958	0.189

4.3 雷达信号聚类分选实验

随着战场的电磁环境越加密集复杂,面对高密度、分离性差、噪声众多的雷达信号,传统的聚类算法比如 k-means 无法进行有效的聚类分选. RFCM 聚类算法能够很好地对密集复杂雷达信号进行分选,因此针对 RFCM 聚类算法存在的问题加以改进,并应用于雷达信号分选,使其更好地对雷达信号进行分选显得十分必要.

从文献[10]选取用于信号分选的 4 部 3cm 的机载脉冲多普勒雷达仿真数据,如表 6 所示,其中 PRI 是指脉冲重复间隔。

表 6 雷达脉冲参数仿真数据

雷达信号	PRI/ μ s	载频/ GHz	脉宽/ μ s	到达角/ $^{\circ}$	脉冲数/个
1	3~4 PRI 组变	8.90~9.30 频率捷变	0.3~0.4	38~41	300
2	30~60 PRI 组变	9.10~9.30 频率跳变	0.8~0.9	36~39	26
3	4~5 PRI 组变	9.40~9.80 频率捷变	0.4~0.5	40~43	228
4	40~70 PRI 组变	9.60~9.80 频率跳变	0.9~1.0	42~45	18

分别采用 RFCM 聚类算法和 SRFCM 聚类算法对表 3 的雷达仿真数据进行聚类分选,均采用 50 次 Monte-Carlo 实验. 2 种聚类算法中设定聚类数目为 4,初始聚类中心随机选择. 隶属度判断阈值 τ 的变化对聚类结果有较大的影响,设定 τ 取值范围为[0,1],为了研究 τ 的变化对聚类结果的影响,分别选用不同的 τ 值进行雷达信号聚类分选实验,其他参数如表 1 不变。

选用分选准确率评价聚类分选结果. 分选准确率是衡量雷达信号分选方法优劣的主要指标,其定义为被准确分选的总脉冲数所占脉冲总数的百分比,结果中对 50 次实验的分选准确率进行平均. 聚类分选结果如图 6 所示。

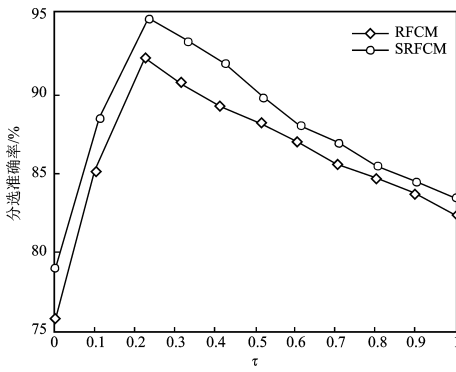


图 6 聚类分选结果

由图 6 可见,在 τ 的各个取值上,SRFCM 聚类算法相比 RFCM 聚类算法取得了更高的分选准确率. 当 τ 增大到一定值时,分选准确率开始下降,故选用 SRFCM 聚类算法对雷达信号进行分选时,必须设定一个合适的 τ 值。

5 结束语

本文针对 RFCM 聚类算法中存在的对不同数

据对象统一赋予固定权重的问题,提出了一种基于自适应权重的 RFCM 聚类算法. 通过选用改进的反余切函数重新分布后的距离计算均衡因子,再使用均衡因子动态地调节固定权重获得自适应权重,进而运行聚类运算,直至结果稳定. 仿真实验表明,基于自适应权重的 RFCM 聚类算法在人工数据集和 UCI 数据集上取得了更有效的聚类结果,应用于雷达信号分选,取得了更高的分选准确率. 另外,所提算法会受到主观参数的影响,比如文中讨论的隶属度判断阈值,还有其他的主观参数比如:聚类个数,上近似集权重和下近似集权重等,如何改善这些主观参数对聚类结果的影响是下一步的研究方向。

参考文献:

[1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.

[2] 郭军华. 数据挖掘中聚类分析的研究[D]. 武汉: 武汉理工大学,2003.

[3] Mitra S, Banka H, Pedrycz W. Rough-fuzzy collaborative clustering[J]. IEEE Transactions on Systems, Man, and Cybernetics; Part B, 2006, 36(4):795-805.

[4] Maji P, Pai S K. Rough set based generalized fuzzy c-means algorithm and quantitative indices[J]. IEEE Transactions on Systems, Man, and Cybernetics; Part B, 2007, 37(6):1529-1540.

[5] 刘兵,夏士雄,周勇. 基于样本加权的可能性模糊聚类算法[J]. 电子学报,2012,40(2): 371-375.

[6] Panwan Lingras. Interval set clustering of web users with rough k-means[J]. Journal of Intelligent Information System, 2004,23(11): 5-16.

[7] Dames D L, Bouldin D W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine intelligence, 1979, 1(2): 224-227.

[8] XIE X L, BENI, G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine intelligence, 1991, 13(8): 841-847.

[9] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(2/3): 193-218.

[10] 张腾飞,陈龙,李云. 基于簇内不平衡度量的粗糙 K-means 聚类算法[J]. 控制与决策,2013,28(10): 1479-1484.

作者简介:

张 强 男,(1991-),硕士研究生. 研究方向为电子对抗理论与技术. E-mail: zjslwdyx@163.com.