文章编号：2096-1618(2020)04-0406-06

基于2型模糊集的粗糙模糊C-means算法

鲍杨婉莹，蒋　瑜，李　冬
（成都信息工程大学软件工程学院，四川 成都 610225）

摘要：聚类算法在图像处理、模式识别等领域有广泛应用，粗糙模糊C-means算法是近年来研究较多的聚类算法。在面对聚类结构不同的样本时，传统的粗糙模糊C-means算法存在聚类簇心偏向性和隶属度选取的问题，使聚类结果不理想。提出一种基于2型模糊集的粗糙模糊C-means算法，算法采用2型模糊集理论，计算样本的次隶属度，从而描述样本的深层信息，根据样本最大隶属度和次大隶属度之间的差别，将样本划分到类簇的上下近似集中，根据上下近似集的权重，迭代并重新计算簇心，直到达到设定阈值或者满足算法终止条件。将改进的粗糙模糊C-means算法在人工数据集和UCI数据集上进行实验对比，结果表明改进的粗糙模糊C-means算法具有良好的性能。

关　键　词：聚类；粗糙集；2型模糊集；粗糙模糊C-means
中图分类号：TP301.6　　**文献标志码**：A
doi：10.16836/j.cnki.jcuit.2020.04.007

0　引言

聚类是从无类别的样本中选择相似样本聚合并且放大样本之间差异的过程，通过聚类过程得到的结果是相同属性的类簇并且该聚类结果能够与其他属性的类簇明显区分开[1]。常见的聚类算法根据划分方式的不同可分为两类，一类为硬聚类算法，一类为软聚类算法[2]。传统的聚类算法多为硬聚类算法，例如硬C-means算法，该算法将样本空间划分为不重叠的类簇，每个类簇都是精确的，即硬聚类算法得到的样本的隶属度通常为0或者1，一个样本只能完全属于或者完全不属于某一个类簇。现实情况下，不同类簇间多有重叠，类簇间的边界并不是十分清晰，利用硬划分的方法去衡量样本隶属度得到的结果往往不尽人意，在处理聚类边界问题时，很难在类簇周围设定清晰边界。因此，常常使用模糊集和粗糙集来处理样本集的模糊性和确定性。

模糊集是一种处理模糊性问题的方法，利用模糊集思想将数据样本隶属度设定为0到1的区间，能够体现样本的模糊性。Lingras等[3]提出模糊C-means算法（Fuzzy C-means，FCM）。但是，FCM算法对于多个类簇的边界样本分辨能力不足，使簇心的迭代出现偏向性。随后，Lingras等[4]利用粗糙集上下近似集思想提出粗糙C-means算法（Rough C-means，RCM）。RCM算法将那些确定归属于类簇的样本划分到类簇下近似区域，不确定归属的样本则划分到边界区域。

Mitra等[5]利用RCM算法迭代时发现类簇边界区样本相对于簇心的归属存在模糊性，然而却没有相应的值衡量，使簇心迭代结果不够准确，但模糊集能够详细刻画样本隶属度问题，于是提出RFCM算法（Rough Fuzzy C-means，RFCM）。随后，Maji等[6]认为在RFCM算法迭代过程中下近似集样本隶属度计算过于复杂。因为样本被分配到类簇下近似集就一定属于该类簇，所以将簇类下近似集合样本的隶属度设置为1，从而一定程度上提升了算法的性能。文献[7]、[8]对RFCM的算法步骤做了改进，在图像分割上取得较好的实验结果。

上述模糊集算法的隶属度多采用0到1区间上的具体数值衡量，即Type-1型模糊度量。这种度量方式虽然考虑了样本的空间分布，但是一些样本隶属度表现出的模糊性很难用一个精确的数值衡量，人为地利用具体数值衡量隶属度使样本部分信息缺失，不能客观全面反映样本的实际情况[9]。针对以上问题提出一种改进的RFCM算法，引入Type-2型模糊度量方式，对模糊隶属度进一步模糊化，从而增强隶属度的模糊程度，使其更加符合样本的深层信息，在人工数据集和UCI数据集上实验并采用多种衡量指标对比分析，证实了改进算法的有效性。

1　模糊集在C-means算法中的应用

1.1　FCM算法

FCM算法是模糊聚类算法中应用最为广泛的一

收稿日期：2019-11-08
基金项目：四川省教育厅重点资助项目（2017Z032）

种算法,它采用迭代优化目标函数得到样本对类中心的隶属度,从而达到对数据集的模糊分类的目的[10]。相较于硬 C-means 算法,FCM 算法聚类结果更加符合样本实际情况。FCM 通过计算数据样本的隶属度衡量样本类别从属关系,其标准目标函数为:

$$J_{FCM}=\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{m}\parallel x_i-C_j\parallel^2 \tag{1}$$

$$\sum_{j=1}^{c}\mu_{ij}=1,0\le\mu_{ij}\le1,\quad i=1,2,\cdots,n \tag{2}$$

其中 C 表示类簇中心,μ_{ij} 表示样本 x_i 属于 C_j 的隶属度,m 为模糊指数,其中 d 表示样本之间的欧式距离,例如 $d_{ij}=\parallel x_i-C_j\parallel$ 表示样本 x_i 与 C_j 的欧式距离,其中隶属函数和聚类中心迭代公式如下:

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}},\quad i=1,2,\cdots,n,\quad j=1,2,\cdots,c \tag{3}$$

$$C_j=\frac{\sum_{i=1}^{n}\mu_{ij}^{m}x_i}{\sum_{l=1}^{n}\mu_{jl}^{m}},j=1,2,\cdots,c \tag{4}$$

为保证算法的收敛性,通常将模糊指标取值范围设定为 $m>1$。通过迭代式(3) 和式(4),直至满足收敛条件,达到最优解。

1.2　RFCM 算法

RFCM 算法利用下近似集和边界集概念,将确切属于当前簇心的聚类对象归入该类簇的下近似集,将相关不确定的聚类对象归入该类簇的边界集,重新计算边界域中的聚类对象与簇心的隶属度,将隶属度作为参数加入聚类簇心公式迭代,其中 w_l 为下近似集权重,w_b 为边界集权重。则 RFCM 簇心迭代公式如下:

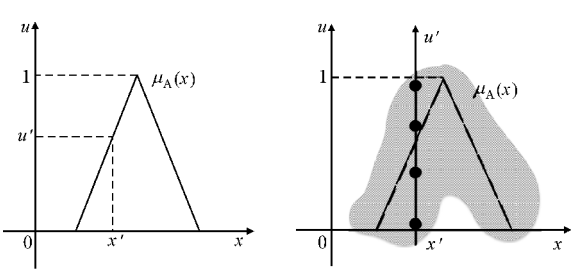
$$v_i=\begin{cases}w_l\dfrac{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}X_j}{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}}+w_b\dfrac{\sum_{x_j\in\hat{C}_i}\mu_{ij}^{m}X_j}{\sum_{x_j\in\hat{C}_i}\mu_{ij}^{m}}, & \underline{C_i}\neq\varnothing\wedge\hat{C}_i\neq\varnothing\\ \dfrac{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}X_j}{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}}, & \underline{C_i}\neq\varnothing\wedge\hat{C}_i=\varnothing\\ \dfrac{\sum_{x_j\in\hat{C}_i}\mu_{ij}^{m}X_j}{\sum_{x_j\in\hat{C}_i}\mu_{ij}^{m}}, & \underline{C_i}=\varnothing\wedge\hat{C}_i\neq\varnothing\end{cases} \tag{5}$$

Sahil 等对 RFCM 算法进行修改,改进方法首先计算样本与每个类簇的隶属度,选取所有隶属度中最大值和次大值,通过计算这两者之间的差值,决定样本的分配。若差值大于设定阈值 λ,则将样本分配到隶属度最大值所在类簇的下近似集;否则,分配到隶属度最大值和次大值所在类簇的边界集[7]。该算法簇心迭代公式如下:

$$v_i=\begin{cases}w_l\dfrac{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}X_j}{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}}+w_b\dfrac{\sum_{x_j\in\hat{C}_i}\mu_{ij}^{m}X_j}{\sum_{x_j\in\hat{C}_i}\mu_{ij}^{m}}, & \underline{C_i}\neq\varnothing\wedge\hat{C}_i\neq\varnothing\\ \dfrac{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}X_j}{\sum_{x_j\in\underline{C_i}}\mu_{ij}^{m}}, & \text{otherwise}\end{cases} \tag{6}$$

2　Type-2 型模糊度量对模糊聚类算法的影响

Zadeh 提出的模糊集合论,对经典集合中的元素引入隶属度的概念使其模糊化,从而达到描述集合不确定性的目的,即为 Type-1 型模糊集。随后,Zadeh 又提出 Type-2 型模糊集。Type-2 型模糊集可以理解为“不同个体对同一事物的不同解释”,通过将隶属度进一步模糊化,即求解次隶属度,增强对集合模糊性的刻画能力,即 Type-2 型模糊度量值是对 Type-1 型模糊度量值的进一步模糊化,将“精确”的 Type-1 型隶属度值替换为“模糊”的隶属度值的范围[11],如图 1 所示。图 1(a) 为 Type-1 型模糊隶属度函数分布,任意选取样本 x' 其相对于类簇 A 的隶属度为“精确”的数值 μ'。图 1(b) 为 Type-2 型模糊隶属度函数分布,任意选取 x' 其相对于类簇 A 的隶属度为一个模糊范围,以 x' 点向 X 轴做垂线,与模糊区域相交的直线上所有点都可以取值,因此样本 x' 相对于类簇 A 的隶属度为一个集合。Type-2 型模糊隶属度函数为三维函数提供更加广泛的自由度,当样本属于某种类型的“程度”不确定时,使用 Type-2 型模糊集能够更好适用于这种情况。

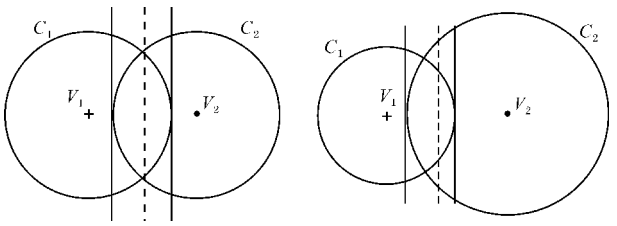


(a)Type-1型模糊隶属度函数　(b)Type-2型模糊隶属度函数

图 1　Tpye-1 型与 Type-2 型模糊隶属度函数区别

在聚类中常常遇到类簇间重合部分样本划分的问题,而重合部分的样本划分主要分为类簇的聚类结构相同和不同两种情况,如图 2 所示,设两个类簇 C_1 和 C_2 以及簇心 V_1 和 V_2。在图 2(a) 中,当类簇聚类结构相同时,根据样本模糊隶属度直接划分即可,在图 2(b) 中,当类簇聚类结构不同时,直接计算重合部分样本的隶属度会使簇心的选取存在偏向性。因此,要考虑类簇聚类的结构,计算样本隶属度的模糊性,再

确定簇心。使用 Type-2 型模糊集理论,考虑到聚类结构的不同,对主隶属度进行平滑调节。



(a)类簇聚类结构相同 (b)类簇聚类结构不同

图 2 类簇聚类结构图

由于之前计算机硬件性能的局限性,Type-2 型模糊集理论近年才得到广泛应用。利用 Type-2 型模糊集能够描述更加深层次的不确定性信息,特别是在处理类簇样本数量不一致或者需要描述样本深层次信息的问题时,能够得到更好的效果。

3 改进的 RFCM 算法

基于 Sahil 改进的 RFCM 算法,对其进行了扩展和优化。由上所述,RFCM 算法均采用模糊隶属度函数为 Type-1 型模糊集度量。Type-1 型模糊集中元素的隶属度都是用一个“精确”的数值表示,其隶属度函数是二维的。虽然该度量在一定程度上考虑了类簇内部数据样本的空间分布,但其具体量化也限制了数据本身潜在的更深层次信息的描述。而 Type-2 型模糊集中元素的隶属度为分布在[0,1]的一个模糊集合,隶属度函数为三维的,因此Type-2型模糊集在刻画和处理不确定关系时就多了一个新的自由度表达不确定关系,使描述的关系模糊性增强,进而能够更好地处理复杂的模糊环境和不确定的模糊隶属度关系,因此Type-2型模糊集可以被理解为广义的模糊集[11]。对次隶属度降型得到 Type-2 型模糊隶属度计算方法如公式(7)所示[11],其中 μ 为式(3)计算得到的主隶属度。

$$\gamma_{ij}=\mu_{ij}-\frac{1-\mu_{ij}}{2} \tag{7}$$

$$v_i=\begin{cases} w_l\dfrac{\sum\limits_{x_j\in \underline{C_i}}\gamma_{ij}^m X_j}{\sum\limits_{x_j\in \underline{C_i}}\gamma_{ij}^m}+w_b\dfrac{\sum\limits_{x_j\in \hat{C}_i}\gamma_{ij}^m X_j}{\sum\limits_{x_j\in \hat{C}_i}\gamma_{ij}^m}, \underline{C_i}\neq\varnothing\wedge\hat{C}_i\neq\varnothing \\ \dfrac{\sum\limits_{x_j\in \underline{C_i}}\gamma_{ij}^m X_j}{\sum\limits_{x_j\in \underline{C_i}}\gamma_{ij}^m}, \underline{C_i}\neq\varnothing\wedge\hat{C}_i=\varnothing \\ \dfrac{\sum\limits_{x_j\in \hat{C}_i}\gamma_{ij}^m X_j}{\sum\limits_{x_j\in \hat{C}_i}\gamma_{ij}^m}, \underline{C_i}=\varnothing\wedge\hat{C}_i\neq\varnothing \end{cases} \tag{8}$$

改进后的算法首先在样本集中随机选取簇心进行初始化。然后,计算样本与簇心之间的欧式距离,该距离用来计算样本属于各个类簇的隶属度,计算各个类簇隶属度中最大值与次大值相减的结果。若其结果大于设定阈值 λ ,则将样本添加到最大隶属度所在类簇的下近似集;若为其他情况,则将样本添加到最大值与次大值所在类簇的边界集。对所有样本重复以上步骤并完成分配,随后用改进的粗糙聚类算法重新计算簇心,重复迭代直到两次连续迭代的类簇中各个样本的隶属度差异小于设定阈值β 或者达到设定迭代次数,算法终止。隶属度计算公式和簇心迭代如式(8)所示。

综上所述,改进 RFCM 算法具体步骤如下:

改进 RFCM 算法步骤

Step1:在给定数据集中选取一些随机点为初始聚类簇心;

Step2:计算 n 维空间中样本 X_k 与簇心之间的距离;

Step3:利用步骤 2 中的距离和式(7)计算样本的隶属度;

Step4:假定 γ_{ik} 为最大隶属度, γ_{jk} 为次大隶属度,若 $\gamma_{ik}-\gamma_{jk}>\lambda$;将 X_k 添加到 $\overline{C_i}$;否则将 X_k 添加到 $\hat{C}_i$ 和 $\hat{C}_j$;

Step5:用式(8)重新计算簇心;

Step6:若第 t 次迭代与 $t-1$ 次迭代的聚类中心各数据点的隶属度值之差小于设定阈值β 或者达到设定迭代次数,算法终止;否则重复 Step2 ~ Step6.

因为算法初始化时簇心的选取是随机的,在不同迭代时结果往往会不同,上下近似集的权重根据近似集的重要性选取,而 Step4 中的 λ 是根据数据集的分布不同而变换,若 λ 的值越大,则类簇边界集中的数据越多。β 是控制停止条件的阈值参数,该阈值根据实际需求调节。改进一方面是针对传统算法收敛速度较慢,在面向不均衡数据在隶属度计算存在偏向性的问题时,引入了 Type-2 型模糊隶属度对其改进。

4 实验及结果分析

将改进的 RFCM 算法在人工数据集和 UCI 数据集与传统的 RFCM 和 RFCM(Sahil)进行实验比较。

4.1 相关衡量指标

实验主要用以下 4 个指标衡量聚类效果。

(1)OK 表示位于类簇下近似集并且正确的聚类样本个数。

(2)πOK 表示位于类簇下近似集但错误的聚类样本个数。

(3)DBI 表示类簇内部紧密程度,其值的含义为任意两个类簇的类内平均距离之和与两个类簇簇心距离比值的最大值[12]。具体计算方法如式(9)所示,其中分子为类簇内所有点到该簇心的平均距离之和,分母为两个类簇簇心之间的距离,n 为类簇数量。

(4)DVI 为任意两个类簇之间最短距离与任意类簇的类内最大距离比值[13]。具体计算方法如式(10)所示,其中分子为两个类簇之间最短距离,分母为簇类内部最大距离。

DBI 值越小类簇内部越紧密,不同类簇间距越远,DVI 值越大类簇间距越大,则簇类内部越紧密。

$$\mathrm{DBI}=\frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \tag{9}$$

$$\mathrm{DVI}=\frac{\min_{1\le i<j\le n}d(i,j)}{\max_{1\le k\le n}d'(k)} \tag{10}$$

4.2　人工数据集实验结果

为验证改进算法的有效性,实验采用人工合成的数据集,该数据集由线性分布的 8 个类,34 个样本组成,模拟某地区不同物种的地域分布,如图 3 所示。为了保证实验的公平性,算法的下近似集权重系数设为 0.75,边界集权重系数为0.25,模糊系数 m 设置为 2,最大迭代次数为 100 次,其中改进算法阈值 λ 设置为 0.03。按照上述的 4 个指标衡量聚类算法性能,具体实验结果如表 1 所示。

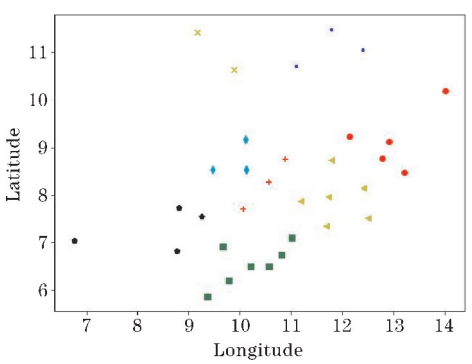
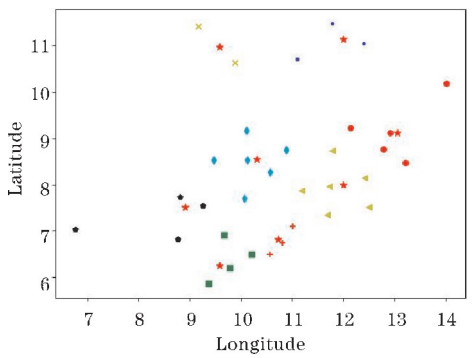


图 3　人工数据集分布

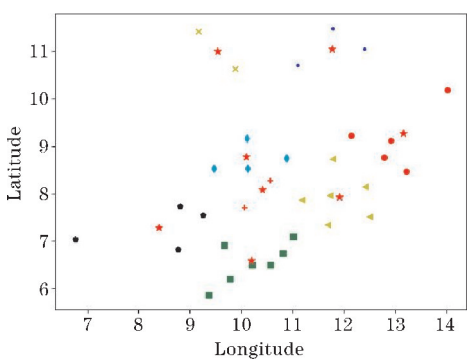
表 1　相关算法在 UCI 数据集上实验结果

指标	RFCM	RFCM(Sahil)	改进的 RFCM
OK	28	32	33
πOK	4	2	1
DBI	0.693	0.716	0.651
DVI	0.65	0.6654	0.672

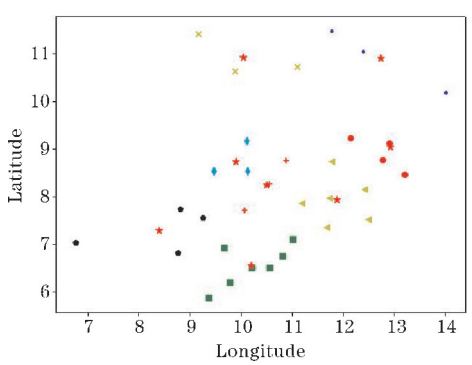
图 4 为三种算法在人工数据集上的聚类结果,其中‘★’为聚类簇心,每个类簇用不同的符号表示。从实验结果可以看出,3 种算法都没有把左下角的样本单独聚为一类。RFCM 算法左下角的簇心位置受到右边类簇的影响向右偏移,对边界区样本类别判定不准确。RFCM(Sahil)算法和改进的 RFCM 算法左下角的簇心位置没有过于受到右侧类簇的影响,但是 RFCM(Sahil)算法对于上方两个类簇的模糊隶属度量不能真实反映样本的模糊程度,最终在样本类别判定存在误差,而改进的 RFCM 算法效果较好。从实验结果可以看出,RFCM 算法和改进的 RFCM 算法类簇内部结构较为紧密,在簇心选择方面,考虑了样本隶属度更加深层次的信息,从而使聚类效果具有较好的表现。



(a) RFCM



(b) RFCM(Sahil)



(c) 改进的 RFCM

图 4　人工数据实验结果

4.3　UCI 数据集实验结果

为保证实验的公平性,算法的参数选取相同的数值,具体算法相关参数设置如表 2 所示,算法最大迭代次数为 100 次,其中改进算法阈值 λ 设置为0.03。

表 2 相关算法参数设置

数据集	RFCM			RFCM(Sahil)			改进的 RFCM		
	m	w_b	w_l	m	w_b	w_l	m	w_b	w_l
Iris	2	0.10	0.90	2	0.10	0.90	2	0.10	0.90
Soybean(Small)	1.1	0.25	0.75	1.1	0.25	0.75	1.1	0.25	0.75
Zoo	2	0.35	0.65	2	0.35	0.65	2	0.35	0.65

为实验结果方便对比,在 UCI 数据集上选取 Iris、Soybean(Small)、Zoo 3 个数据集进行实验。其中,Iris 数据集记录 3 种类别的样本集,共有 150 个样本数据,每个类别均有 50 个样本,每个样本均有 4 种属性值;Soybean(Small)数据集记录了 4 种类别的样本集,共有 47 个样本数据,类别样本个数为 10,10,10,17,每个样本都有 35 种属性值;Zoo 数据集大致分为 7 种类别的样本集,共有 101 个样本数据,每个样本都有 17 种属性。相关算法在 UCI 数据集上实验结果如表 3 所示。

表 3 相关算法在 UCI 数据集上实验结果

数据集	指标	RFCM	RFCM(Sahil)	改进的 RFCM
Iris	OK	131	134	133
	πOK	2	1	1
	DBI	0.83	0.6082	0.583
	DVI	2.18	2.74	2.865
Soybean(Small)	OK	34	32	35
	πOK	0	0	0
	DBI	2.37	1.8317	1.807
	DVI	0.663	0.6345	0.656
Zoo	OK	80	79	82
	πOK	8	7	6
	DBI	1.92	1.209	1.2316
	DVI	0.8122	1.257	1.4711

由表 3,在 Iris 数据集上,改进的 RFCM 虽然在精度上稍微低于 RFCM(Sahil)算法,但是在 DBI 和 DVI 指标上则要好于 RFCM(Sahil)算法,改进的 RFCM 算法表现良好的聚类效果。在 Soybean(Small)数据集上,改进算法和传统的 RFCM 算法均有较好的准确率,改进的 RFCM 算法相较于 RFCM(Sahil)算法的类簇内部更加紧密。在 Zoo 数据集上,改进算法也具有较好的表现。并且,根据簇心个数选择的不同,DBI 和 DVI 的值在簇心个数为 7 的时候效果最好,这也和 UCI 数据集划分相一致,进一步说明改进的算法能够更好地选取簇心,使类簇内部距离更紧密,类簇间区别更大。在聚类结构相同的 Iris 数据集上,改进的 RFCM 算法簇心选取更加合理,在聚类结构不同的 Soybean(Small)数据集和 Zoo 数据集上,改进的 RFCM 算法在精度及 DBI 或 DVI 上均有较好的表现。

5 结论

在数据挖掘中,C-means 聚类算法常被用来做异常信号检测、图形分析等实际应用。主要提出一种改进的 Type-2 型 RFCM 算法,Type-1 型的 RFCM 算法多采用固定数值的模糊隶属度值,对样本隶属度描述与实际情况存在误差,不能描述样本深层次信息。另外,传统 C-means 算法步骤在簇心选择不够优化、存在偏向性等问题。在 RFCM 算法基础上增加 Type-2 型模糊集思想,提出改进的 Type-2 型 RFCM 算法,模糊化了样本隶属度,通过在 UCI 数据集上的实验分析对比,验证改进算法的有效性。而如何恰当选取聚类的类簇个数,则是下一步的研究方向。

致谢:感谢成都信息工程大学青年学术带头人科研基金项目(J201609)对本文的资助

参考文献:

[1] 贺玲,吴玲达,蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究,2007,24(1):10-13.

[2] 王学恩,韩德强,韩崇昭. 采用不确定性度量的粗糙模糊 C 均值聚类参数获取方法[J]. 西安交通大学学报,2013,47(6):55-60.

[3] Lingras P, Yan R, Jain A. Clustering of Web Users:K-Means vs. Fuzzy C-Means[C]. Proceedings of the 1st Indian International Conference on Artificial Intelligence, IICAI 2003, Hyderabad, India, December 18-20,2003. DBLP,2003.

[4] Lingras P, West C. Comparison of Conventional and Rough K-Means Clustering [C]. International Workshop on Rough Sets. 2003.

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

[6] MajiP, Pal S K. RFCM:A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets[J]. 2007.

[7] Tripathy B K, Sobti S, Shah V. A Refined Rough

Fuzzy Clustering Algorithm[C]. IEEE International Conference on Computational Intelligence & Computing Research. IEEE,2015.

[8] Jiao S, Yu L, Ying Z, et al. Enhanced rough-fuzzy C-means algorithm with strict rough sets properties [J]. Applied Soft Computing,2016,46:827-850.

[9] Sukhveer, SINGH, Harish, et al. Comments on "Some new distance measures for type-2 fuzzy sets and distance measure based ranking for group decision making problems"[J]. Frontiers of Computer Science,2018,12(2):396-400.

[10] Zhou K, Yang S. Effect of cluster size distribution on clustering: a comparative study of K-means and fuzzy C-means clustering[J]. Pattern Analysis and Applications,2019:1-12.

[11] Begum SA, Devi O M, Begum S A, et al. A Rough Type-2 Fuzzy Clustering Algorithm for MR Image Segmentation[J]. International Journal of Computer Applications,2013,54(4):4-11.

[12] Davies D L, Bouldin D W. A Cluster Separation Measure[J]. IEEE Trans Pattern Anal MachIntell,1979(2):224-227.

[13] Dunn J C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-separated Clusters[J]. Journal of Cybernetics, 1973,3(3):32-57.

Type-2 Fuzzy Set based Rough Fuzzy C-means Clustering Algorithm

BAO Yangwanying,　JIANG Yu,　LI Dong

(College of Software Engineering, CUIT, ChengDu 610225, China)

Abstract: Clustering algorithm is widely used in image processing, pattern recognition and other fields. The RFCM algorithm is a clustering algorithm that has been studied more in recent years. When the clustering structure of sample is different, the traditional RFCM algorithm has the problem of cluster center bias and membership selection, which makes the clustering result worse. This paper proposes a RFCM C-means algorithm based on type-2 fuzzy set. The refined RFCM algorithm uses the type-2 fuzzy set theory to describe the deep information of the sample by calculating the sub-degree of membership of the sample. The sample is divided into the upper and lower approximation sets of the cluster based on the difference between the maximum membership degree and the second-largest membership degree, and according to the weights of the upper and lower approximation sets, the cluster center is iterated and recalculated until the set threshold is reached or the algorithm termination condition is met. , The performances of improved RFCM C-means algorithm experimented on the artificial datasets and the UCI datasets are compared, the results show that the improved RFCM algorithm has good performance.

Keywords: clustering; rough set; type-2 fuzzy set; rough fuzzy C-means