DOI：10.7652/xjtuxb201306010

采用不确定性度量的粗糙模糊 C 均值聚类参数获取方法

王学恩[1,2]，韩德强[1,2]，韩崇昭[1,2]
（1.西安交通大学智能网络与网络安全教育部重点实验室，710049，西安；
2.西安交通大学电子与信息工程学院，710049，西安）

摘要：针对粗糙模糊 C 均值聚类的阈值、权重选取问题，提出了一种基于不确定性度量的参数自适应获取方法。该方法将阈值选取归结为一个最优划分寻找问题，给出一种基于方差的划分优劣评价方法；利用信息熵来度量样本归属的模糊性，基于该模糊性度量和类簇的粗糙度，提出了一种权重参数自适应计算方法。将所提方法应用于粗糙模糊 C 均值聚类，并将分别基于所提方法与典型参数选取方法的粗糙模糊 C 均值聚类算法在人工数据集和真实数据集上进行实验比较。结果表明，基于所提参数确定方法的粗糙模糊 C 均值聚类能获得更好的聚类有效性和准确性。

关键词：聚类；粗糙模糊 C 均值；熵；粗糙度
中图分类号：TP18 **文献标志码**：A **文章编号**：253-987X(2013)06-0055-06

Selection Method for Parameters of Rough Fuzzy *C*-Means Clustering Based on Uncertainty Measurement

WANG Xue'en[1,2], HAN Deqiang[1,2], HAN Chongzhao[1,2]
(1. Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049 China; 2. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A selection method for parameters of the rough fuzzy *C*-means clustering based on uncertainty measurement is proposed. The selection of parameter threshold is converted into an optimal partition problem, and partitions are evaluated based on variances. The information entropy is employed to measure the fuzziness of the sample membership to clusters, then an adaptive method to calculate weights is presented based on the roughness of clusters and the fuzziness of samples. The adaptive parameter selection method is employed in rough fuzzy *C*-meaning algorithm. Experiments and comparisons with some existing typical parameters selection methods for rough fuzzy *C*-means clustering algorithm are performed on synthetic and real-world data sets, and the results show that the proposed algorithm achieves higher accuracy, and is effective.

Keywords: clustering; rough fuzzy *C*-means; entropy; roughness

聚类方法广泛应用于机器学习、数据挖掘等领域，它与分类方法不同，是一种无监督的模式识别方法。聚类的主要任务是根据某种相似性将样本集划分成若干个类或簇，使得同一个类中的样本之间具

收稿日期：2012-08-06。 作者简介：王学恩(1982—)，男，博士生；韩德强(通信作者)，男，副教授。 基金项目：国家自然科学基金群体创新基金资助项目(60921003)；国家自然科学基金资助项目(61074176，61104214)。
网络出版时间：2013-03-07 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20130307.0827.002.html

有较高的相似性，不属于同一个类的样本之间具有较高的差异性。现有的聚类方法有基于模型的方法、基于密度的方法、划分型方法、层次型方法等，其中 K 均值(KM)聚类方法是一种常用的方法[1]。

KM 聚类方法是一种硬划分的方法，数据空间被划分成 K 个互不重叠的区域，它对应的每个类都是精确的。现实中不同类别之间大多是有重合的，划分的边界并不是那么清晰。目前，一些软聚类方法能够应对这样的问题，其中最为流行的是模糊 C 均值(FCM)聚类方法，此外还有基于粗糙集的聚类方法[2-3]，其中粗糙 C 均值(RCM)聚类方法[2]是对 KM 方法的一种扩展。对于聚类之间的重叠区域，RCM 借用粗糙集中的上近似与下近似之间的边界区域来刻画，这样孤立点一般不会被分配到某个类的下近似中，而是分配到边界区域，从而降低了孤立点对聚类结果的影响。粗糙集与模糊集是两种不同的处理不确定信息的数学工具，它们之间有一定的互补性，很多研究将两者结合用于聚类分析中，其中包括粗糙模糊 C 均值(RFCM)聚类方法[4-5]。与 FCM 聚类相比，RFCM 能够更好地处理孤立点并减少聚类结果受噪声的影响；与 RCM 聚类相比较，RFCM 将模糊隶属度引入进来能够对边界区域和下近似区域进行更好的刻画，从而获得更优的聚类结果。

基于粗糙集的聚类方法目前已应用于图像分割、文本分类等领域[6]，然而在应用中首先需要解决参数(包括阈值和权重)的选择问题。文献[7]根据参数在聚类迭代前期和后期所起作用的变化趋势，给出一种阈值和权重随着迭代次数增加而自适应调整的方法；文献[8]给出利用遗传算法对初始参数进行优化的方法；文献[9]给出一种基于近似集合中的高斯距离比例的权重自适应选取方法；文献[10]给出一种基于阴影集的阈值选取方法[10]。本文认为，聚类中对于不同的数据集以及不同的类簇应该设定不一样的参数，文献[7]中方法没有考虑数据集的差异，文献[8]对所有的类簇使用相同的参数，文献[9-10]分别只研究了权重或阈值的获取方法。

考虑数据集和类簇的差异性，本文提出了一种基于不确定性度量的阈值和权重的自适应获取方法。首先将阈值选取问题归结为一个最优划分寻找问题，并给出一种基于方差的划分有效性评价函数，利用该评价函数为每一个类簇寻找对应的最优阈值；然后，基于每个样本点隶属于所有类的模糊性和聚类的粗糙度，给出了权值的自适应计算方法。最终，在人工数据集和 UCI(University of California Irvine)数据集[11]上进行聚类效果验证，结果表明所提方法能获得更好的有效性和准确性。

1 粗糙模糊 C 均值聚类方法

在 RFCM 聚类中使用上近似、下近似两个集合来描述一个聚类。令 x_i 表示一个数据点($i=1,\cdots,N$)，样本集记为 $X=\{x_1,\cdots,x_N\}$，C_k 表示第 k 个类簇，$\overline{C_k}$和$\underline{C_k}$分别表示类簇 k 的上、下近似集合，$C_k^B=\overline{C_k}-\underline{C_k}$表示边界集合，即上近似中不被下近似所包含的区域。$v_k$ 表示类簇 C_k 的中心，$d_{ki}=\|x_i-v_k\|$ 表示样本 x_i 与 v_k 之间的距离。

目前，主要有两种 RFCM 聚类方法，分别是由 Mitra 和 Maji 提出的[4-5]，其区别在于聚类中心的计算方法上。本文将 Mitra 等给出的粗糙模糊 C 均值聚类方法记为 RFCMⅠ，在聚类中每个类簇对应有一个模糊下近似区域和一个模糊边界区域，其类簇中心计算方法如下[4]

$$v_k^{\mathrm{RFCMⅠ}}=\begin{cases}A_1, & \text{if } C_k^B=\varnothing\wedge\underline{C_k}\neq\varnothing\\ B_1, & \text{if } C_k^B\neq\varnothing\wedge\underline{C_k}=\varnothing\\ w_{\mathrm{l}}A_1+w_{\mathrm{b}}B_1, & \text{if } C_k^B\neq\varnothing\wedge\underline{C_k}\neq\varnothing\end{cases}\tag{1}$$

$$A_1=\sum_{x_i\in\underline{C_k}}u_{ki}^m x_i/\sum_{x_j\in\underline{C_k}}u_{kj}^m\tag{2}$$

$$B_1=\sum_{x_i\in C_k^B}u_{ki}^m x_i/\sum_{x_j\in C_k^B}u_{kj}^m\tag{3}$$

式中：$w_{\mathrm{l}}+w_{\mathrm{b}}=1\wedge w_{\mathrm{l}}>w_{\mathrm{b}}$；隶属度 u_{ij} 与模糊 C 均值聚类中使用的隶属度一致。Maji 等提出位于下近似中的对象应该与其他类簇无关，并且它们对本类簇具有同等的贡献，此时类簇拥有一个清晰的下近似区域以及一个模糊边界区域，该方法记为 RFCMⅡ，其类簇中心计算方法如下[5]

$$v_k^{\mathrm{RFCMⅡ}}=\begin{cases}A_0, & \text{if } C_k^B=\varnothing\wedge\underline{C_k}\neq\varnothing\\ B_1, & \text{if } C_k^B\neq\varnothing\wedge\underline{C_k}=\varnothing\\ w_{\mathrm{l}}A_0+w_{\mathrm{b}}B_1, & \text{if } C_k^B\neq\varnothing\wedge\underline{C_k}\neq\varnothing\end{cases}\tag{4}$$

$$A_0=\sum_{x_i\in\underline{C_k}}x_i/|\underline{C_k}|\tag{5}$$

从式(1)～式(5)中可以看出，当 $C_k^B=\varnothing\wedge\underline{C_k}\neq\varnothing(\forall k)$时，ARFCMⅠ会退化成一个 FCM 聚类，而 RFCMⅡ则退化成 KM 聚类，当 $C_k^B\neq\varnothing\wedge\underline{C_k}=\varnothing(\forall k)$时，ARFCMⅠ和 RFCMⅡ均退化成 FCM 聚类方法。与 RCM 在确定上、下近似区域时使用

相对(或绝对)距离不同，在 RFCM 方法中使用模糊隶属度来确定上、下近似区域。令 u_{ki}、u_{li} 分别表示样本 x_i 到各个类簇隶属度的最大值和次大值，那么当 $u_{ki}-u_{li}\leqslant\tau$ 时，x_i 在类簇 C_k 和 C_l 的边界区域，否则 x_i 位于 C_k 的下近似区域。

粗糙集理论能够得到广泛的应用，其中一个重要的原因就是它无需人为干预就可以完成知识的获取。在 RFCM 聚类中，聚类质量直接受到阈值参数和权重参数的影响，而这些参数往往需要专家经验来人为设定，这就限制了该方法的推广及应用。如何自动获取合理的参数成为 RFCM 应用中一个亟需解决的问题。

2　参数获取方法

文献[7]认为在聚类迭代过程中，阈值应该是动态变化的，在聚类初期阈值较小，此时样本隶属于边界和下近似的关系不明确，聚类后期该隶属关系变得明确，样本隶属于下近似的可能性变大，其中包含的不确定性较少。本文认为在迭代过程中，阈值应该根据聚类中心点的位置以及样本分布动态的确定，选取使样本归属最为明确的阈值，然后根据选定阈值计算对应的权重参数。因此，聚类过程中参数会根据每步迭代的结果自适应地调整。

2.1　阈值选取方法

在 RFCM 聚类中，阈值的作用就是用来划分出聚类的边界区域和下近似区域。与文献[10]相同，本文根据每步迭代中得到的聚类结果动态地为每个类簇选定不同的阈值。首先获取每个聚类中心对应的最近邻点集 $N_k=\{x_i:d_{ki}=\min_j d_{ji},j=1,\cdots,K\}$，其对应于该聚类中心所属的清晰聚类集合。对类簇 C_k 每选定一个阈值 τ_k 对应有一个下近似 $\underline{C_k}=\{x_i:x_i\in N_k\wedge u_{ki}-u_{li}>\tau\}$ 和边界区域 $C_k^{B'}=\{x_i:x_i\in N_k\wedge u_{ki}-u_{li}\leqslant\tau\}$，其中 u_{li} 是 x_i 到各类簇隶属度的次大值，此时 $\underline{C_k}\bigcap C_k^{B'}=\varnothing$，并且 $\underline{C_k}\bigcup C_k^{B'}=N_k$，$\underline{C_k}$ 和 $C_k^{B'}$ 对 N_k 形成一个划分。一个好的阈值应该能够明确地分辨出样本所属的区域，对应得到的划分中样本归属的不确定性较小。基于此，可以通过对这个划分的评价来获得对阈值优劣的判断，阈值的选取问题转化成一个寻找最优划分的问题。在这样的一个划分中，两个区域之间的距离越大，各自划分区域内数据分布越集中，样本越能够清晰明确地划分到所属的区域。基于这样的考虑，使用两个区域中心之间的距离来度量两个划分区域的距离，用方差来度量样本的集中程度，定义如下目标函数来评价划分的优劣

$$J(\tau_k)=\frac{w_1\sum\limits_{x_i\in\underline{C_k}}(d_{ki}-d^{\mathrm{l}})^2+w_2\sum\limits_{x_i\in C_k^{B'}}(d_{ki}-d^{\mathrm{b}})^2}{|\,d^{\mathrm{b}}-d^{\mathrm{l}}\,|}\tag{6}$$

$$d^{\mathrm{l}}=\sum_{x_i\in\underline{C_k}}d_{ki}/|\,\underline{C_k}\,|\tag{7}$$

$$d^{\mathrm{b}}=\sum_{x_i\in C_k^B}d_{ki}/|\,C_k^B\,|\tag{8}$$

$$w_1=|\,\underline{C_k}\,|/|\,C_k\,|\tag{9}$$

$$w_2=|\,C_k^B\,|/|\,C_k\,|\tag{10}$$

式中：分别用 d^{l} 和 d^{b} 表示 $\underline{C_k}$ 和 $C_k^{B'}$ 中对象到聚类中心的平均距离。在此使用两个平均距离的差衡量 $\underline{C_k}$ 和 $C_k^{B'}$ 中心的距离，同时使用两个划分中对象到其中心距离分布方差的加权和来衡量整个划分内样本的集中程度，此处权重 w_1 和 w_2 代表了两个区域的大小，区域越大则贡献越大。显然，$J(\tau_k)$ 越小对应的划分越好，因而最优阈值对应最小评价指标 $\tau_k^*=\mathrm{argmin}_{\tau_k}J(\tau_k)$，其中 $\tau_k\in[0,u_k^{\max}]$，同时 $u_k^{\max}=\max_i(u_{ki})$。有很多搜索优化方法(如遗传算法、蚁群算法等)可以用来从 $[0,u_k^{\max}]$ 中找到全局最优的阈值，还有简单直接的枚举法[10]，即阈值从 0 以固定增量 λ 增加到 $u_k^{\max}$，相当于从可行空间中等距选取 $n_{\mathrm{s}}(n_{\mathrm{s}}=u_k^{\max}/\lambda)$ 个点，选取其中评价最好的点，该方法得到一个近似最优解，误差由每步的增量决定。

2.2　权重计算方法

权重的作用是衡量边界区域和下近似区域在聚类中心计算中的作用的大小。鉴于各个类簇使用不同的阈值，使得类簇对应的边界区域大小不同，每个类簇应该具有不同的权重。首先考虑样本点归属的模糊性，一个样本点到各个类簇的隶属度之间相差越大，该点的归属越清晰，反之则归属越模糊。在此用信息熵来刻画样本点隶属于各个类簇的模糊(或清晰)程度。对于样本点 x_i，给出样本点归属的模糊性度量如下

$$H(x_i)=-\log\frac{1}{K}+\sum_{k=1}^{K}u_{ki}\log(u_{ki})\tag{11}$$

$H(x_i)$ 取值越大则样本点的归属越清晰，模糊性也就越小。当 x_i 到各个类簇中心距离相等时，该点归属模糊性最大，$H(x_i)$ 取最小值 0。

在模糊粗糙集中粗糙度定义如下

$$r_k=\sum_{x_i\in C_k^B}u_{ki}\Big/\sum_{x_i\in\overline{C_k}}u_{ki}\tag{12}$$

由式(12)可见，边界区域越小类簇的粗糙度越小，此时边界区域在类簇中心计算中的作用也就越小。同

时考虑边界区域样本点归属的模糊性，当归属越清晰时，样本点在类簇中心计算中的作用越大。对于类簇 C_k，提出一种边界区域权重计算方法如下

$$w_{\mathrm{b}}^k = r_k E_{\mathrm{b}}/(E_{\mathrm{b}} + E_{\mathrm{l}}) \tag{13}$$

式中：E_{l} 和 E_{b} 分别是下近似区域和边界区域样本点归属的平均清晰程度，其计算如下

$$E_{\mathrm{l}} = \sum_{x_i \in \underline{C_k}} H(x_i)/\mid \underline{C_k} \mid \tag{14}$$

$$E_{\mathrm{b}} = \sum_{x_i \in C_k^B} H(x_i)/\mid C_k^B \mid \tag{15}$$

可见边界区域中样本点归属的模糊性越小，E_{b} 值越大，此时 w_{b}^k 也越大。当 $\underline{C_k} \neq \varnothing \wedge C_k^B \neq \varnothing$ 时，由于下近似中的样本点相对更接近聚类中心，所以 $E_{\mathrm{l}} > E_{\mathrm{b}}$，保证了 $w_{\mathrm{b}}^k < 0.5$。特别地，当 $\underline{C_k} = \varnothing$ 时，r_k 值为 1，令 w_{b}^k 为 1；当 $C_k^B = \varnothing$ 时，r_k 值为 0，对应令 w_{b}^k 取 0。最后，获得下近似区域权重 $w_{\mathrm{l}}^k = 1 - w_{\mathrm{b}}^k$。

2.3 基于参数自适应选取的 RFCM 聚类算法

基于前面两节给出的参数选取方法，本节给出了基于参数自适应获取的 RFCM Ⅰ 和 RFCM Ⅱ 聚类算法。在聚类迭代过程中首先使用 2.1 节中的方法获取划分阈值，然后使用 2.2 节中的方法计算权重参数，最后算法 ARFCM Ⅰ 及 ARFCM Ⅱ 分别使用式(1)和式(4)获取各个类簇的中心。为了减少初始点选取对结果的影响，在聚类过程中引入文献[12]给出的全局聚类策略。该方法将 K 个簇的聚类问题转化成一系列子聚类问题，是一种聚类个数递增的逐步寻优的方法。它的核心是在每一步根据代价函数(聚类误差)从样本集中选取一个点作为新的聚类中心加入到已有的聚类结果中，该方法能在一定程度上避免聚类进入局部极小点寻找到接近最优解的聚类结果。聚类流程描述如下。

Step 1：初始化。使用所有样本点平均值作为第 1 个聚类中心，即 $v_1(k) = \sum_{i=1}^{N} x_i/N$，$k := 1$。

Step 2：计算第 $k+1$ 个聚类中心。根据代价函数从样本集 X 中选取新的初始聚类中心，记为 x_i。以 $(v_1(k), \cdots, v_k(k), x_i)$ 为初始点，执行 1 次 RFCM 聚类。

Step 2.1：计算样本到第 $k+1$ 个聚类中心的隶属度。

Step 2.2：选择每个类簇对应的阈值 τ_k^* 并计算聚类的下近似区域和边界区域。

Step 2.3：计算权重 w_{l}^k、w_{b}^k。

Step 2.4：根据式(1)或式(4)更新聚类中心。

Step 2.5：此次聚类过程收敛则跳到下一步，否则返回 Step 2.1。

Step 3：$k := k+1$。

Step 4：终止判断。如果 $k = K$ 则程序终止，否则返回 Step 2。

3 实验研究

为了比较不同参数选取方法在聚类中的作用，实验对比了多种方法的聚类结果，包括 FCM 聚类方法、文献[7]给出的使用参数自适应调整方法获取参数的 RCM 聚类(实验中用 RCM 标记)、文献[10]给出的使用阴影集选取阈值的 RFCM Ⅰ 和 RFCM Ⅱ 方法(实验中用 SRFCM Ⅰ 和 SRFCM Ⅱ 标记)、本文给出的 ARFCM Ⅰ 和 ARFCM Ⅱ 聚类方法。对于 SRFCM Ⅰ 和 SRFCM Ⅱ 方法，选取权重 $w_l = 0.9$，同时在其阈值 α_k 的可行解空间 $[u_k^{\min}, (u_k^{\min} + u_k^{\max})/2]$ 中等距地选取 10 个点作为候选集，从中选取令 $V(\alpha_K)$ 最小的点作为最佳阈值。对于 ARFCM Ⅰ 和 ARFCM Ⅱ 方法，同样在其阈值 τ_k 的可行解空间 $[0, u_k^{\max}]$ 中等距地选取 10 个点作为候选集，选择令 $J(\tau_k)$ 最小的点作为最佳阈值。此外，考虑到聚类受初始点选取的影响，实验中对 ARFCM Ⅰ 和 ARFCM Ⅱ 之外的算法使用文献[13]给出的基于密度和距离的初始点选取方法。实验中采用人工数据集和 UCI 数据集分别对算法进行测试，并使用 DB[14](Davies-Bouldin)和 XB(Xie-Beni)指标[15]来考察算法的有效性，同时使用 Rand 指标[16]验证聚类结果的准确性。DB 和 XB 指标都是用来描述聚类的紧凑程度的，两种方法各有偏重，但 XB 和 DB 均是值越小聚类效果越好。Rand 指标计算两个划分之间的一致性(相似)程度，是在已知样本分类结果的情况下常用的一个聚类准确性的度量。实验环境为 CPU PIV 2.4，512M RAM，算法在 VC6.0 下实现。

3.1 人工数据集

实验采用文献[10]所提供的合成数据集，如图 1 所示，该数据集由 3 个聚类组成，每个聚类含有 50 个样本点。6 种聚类结果的评价指标比较如表 1 所示，从中可以看到，ARFCM Ⅰ 方法获得最有效的聚类结果(对应最小的 XB 和 DB 指标)，同时该方法所花费的时间也最长，另外 ARFCM Ⅱ 的结果优于其他方法。对于这个数据集 RCM 的聚类结果边界区域为空，此时 RCM 退化为 KM 聚类，因而其聚类结果较差。图 2 和图 3 分别直观地给出了 SRFCM Ⅰ 和 ARFCM Ⅰ 两个方法的聚类结果，其中基于阴

影集的阈值方法所选取的 3 个阈值为(0.349 175，0.299 759，0.300 241)，本文方法选取的 3 个阈值为(0.534 0，0.720 3，0.720 3)，对应计算得到的 3 个下近似区域权重 w_1 分别为(0.924 1，0.932 5，0.910 6)。对于本数据集，所有方法聚类结果的 Rand 指标均为 0.986 2，具有相同的准确率。

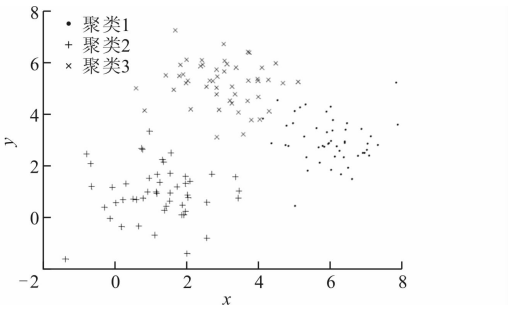

图 1　人工数据集

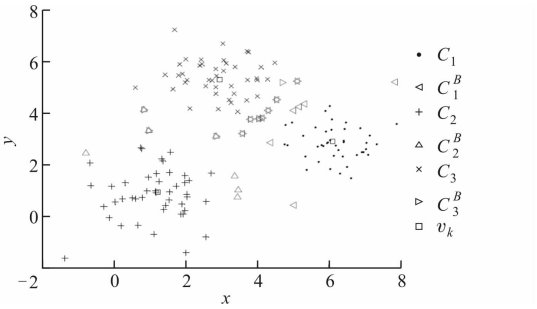

图 2　SRFCM Ⅰ聚类结果

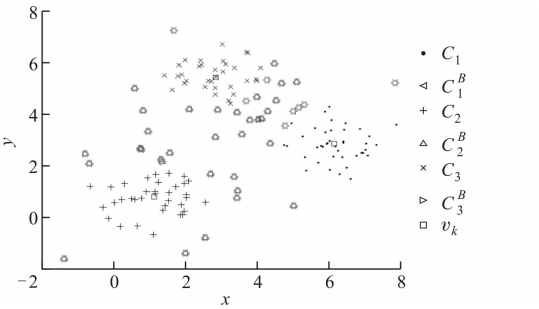

图 3　ARFCM Ⅰ聚类结果

表 1　人工数据集聚类结果

方法	XB 指标	DB 指标	时间/s
FCM	0.086 8	0.585 2	0.063
RCM	0.091 0	0.604 3	0.063
SRFCM Ⅰ	0.078 2	0.527 7	0.094
SRFCM Ⅱ	0.082 7	0.525 9	0.078
ARFCM Ⅰ	**0.074 8**	**0.358 6**	0.109
ARFCM Ⅱ	0.076 3	0.363 5	0.094

3.2　UCI 数据集

实验选用 5 个 UCI 数据集对算法进行测试，实验结果如表 2～表 6 所示，其中加粗数据是最好的聚类结果。结果表明，RCM 与 RFCM 方法能获得比 FCM 方法更有效的聚类，尤其对 Balance 数据集。ARFCM Ⅰ、ARFCM Ⅱ方法依聚类中数据的分布情况对每个类簇选定不同的阈值和权重参数，能够更好地描述数据的内在结构。结果表明，ARFCM Ⅰ在所有数据上获得了最有效的聚类结果，并且在除 Iris 和 Balance 数据集之外的数据集上获得最高的聚类准确性(对应最大 Rand 指标值)；ARFCM Ⅱ在除 Heat-Staglog 之外的数据集上获得最高的聚类准确性，同时该方法能获得较优的聚类有效性。由于 ARFCM(Ⅰ，Ⅱ)要执行 $k-1$ 次聚类，因而在时间消耗上劣于其他方法。

表 2　Iris 数据集聚类结果

方法	XB 指标	DB 指标	Rand 指标	时间/s
FCM	0.136 9	0.949 7	0.879 7	0.078
RCM	0.126 6	0.883 0	0.873 7	0.046
SRFCM Ⅰ	0.123 1	0.635 4	0.873 7	0.125
SRFCM Ⅱ	0.127 6	0.632 3	0.879 7	0.109
ARFCM Ⅰ	**0.112 3**	**0.585 8**	0.885 9	0.141
ARFCM Ⅱ	0.113 2	0.590 6	**0.892 3**	0.156

表 3　Wine 数据集聚类结果

方法	XB 指标	DB 指标	Rand 指标	时间/s
FCM	0.407 9	4.631 5	0.933 1	0.109
RCM	0.295 8	2.921 3	0.941 7	0.125
SRFCM Ⅰ	0.252 2	2.054 0	0.947 3	0.156
SRFCM Ⅱ	0.272 7	2.197 6	0.947 3	0.156
ARFCM Ⅰ	**0.240 2**	**1.761 2**	**0.962 0**	0.187
ARFCM Ⅱ	0.260 1	1.899 0	**0.962 0**	0.250

表 4　Balance 数据集聚类结果

方法	XB 指标	DB 指标	Rand 指标	时间/s
FCM	13.489 7	2.388 7	0.559 8	3.938
RCM	0.388 6	4.695 7	0.569 6	3.938
SRFCM Ⅰ	0.518 8	2.958 6	0.511 7	3.875
SRFCM Ⅱ	0.546 3	2.974 0	0.511 7	6.156
ARFCM Ⅰ	**0.268 8**	**2.335 9**	0.526 5	5.671
ARFCM Ⅱ	0.286 7	2.633 6	**0.609 6**	4.203

表 5 WDBC 数据集聚类结果

方法	XB 指标	DB 指标	Rand 指标	时间/s
FCM	0.312 5	2.020 3	0.866 0	2.938
RCM	0.252 9	1.577 7	0.866 0	2.907
SRFCM Ⅰ	0.261 6	1.690 9	0.869 1	3.094
SRFCM Ⅱ	0.236 8	1.519 4	0.869 1	3.047
ARFCM Ⅰ	**0.188 4**	**0.847 1**	**0.881 3**	8.265
ARFCM Ⅱ	0.191 2	0.862 9	**0.881 3**	7.297

表 6 Heat-Staglog 数据集聚类结果

方法	XB 指标	DB 指标	Rand 指标	时间/s
FCM	0.256 2	1.480 3	0.515 3	0.328
RCM	0.185 3	0.769 4	0.518 2	0.343
SRFCM Ⅰ	0.203 2	1.222 5	0.512 7	0.359
SRFCM Ⅱ	0.228 2	1.167 3	0.512 7	0.328
ARFCM Ⅰ	**0.172 8**	**0.753 0**	0.518 2	0.468
ARFCM Ⅱ	0.187 9	0.789 2	**0.522 9**	0.360

4 结 论

提出了一种基于方差、熵和粗糙度 3 种不确定性度量的 RFCM 阈值和权重自适应获取方法。所提方法在人工合成数据集和 UCI 数据集上进行了聚类有效性和准确性验证，并与已有的主要方法进行了性能对比。对比结果表明，本文方法具有更好的聚类有效性和准确性，减少了聚类过程中的人工干预，有利于 RFCM 的进一步推广。

参考文献：

[1] XU R, WUNSCH D. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[2] LINGRAS P, WEST C. Interval set clustering of web users with rough *k*-means [J]. Journal of Intelligent Information Systems, 2004, 23(1): 5-16.

[3] 何明，冯博琴，马兆丰，等. 基于熵和信息粒度的粗糙集聚类算法 [J]. 西安交通大学学报，2005，39(4)：342-345.
HE Ming, FENG Boqin, MA Zhaofeng, et al. Rough set clustering algorithm based on entropy and information granularity [J]. Journal of Xi'an Jiaotong University, 2005, 39(4): 342-345.

[4] 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.

[5] MAJI P, PAL 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.

[6] MAC PARTHALAIN N, SHEN Q. On rough sets, their recent extensions and applications [J]. Knowledge Engineering Review, 2010, 25(4): 365-395.

[7] ZHOU T, ZHANG Y N, YUAN H J, et al. Rough *K*-means cluster with adaptive parameters [C]// Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. Piscataway, NJ, USA: IEEE, 2007: 3063-3068.

[8] MITRA S. An evolutionary rough partitive clustering [J]. Pattern Recognition Letters, 2004, 25(12): 1439-1449.

[9] 周杨，苗夺谦，岳晓冬. 基于自适应权重的粗糙 *K* 均值聚类算法 [J]. 计算机科学，2011，38(6)：237-241.
ZHOU Yang, MIAO Duoqian, YUE Xiaodong. Rough *K*-means clustering based on self-adaptive weight [J]. Computer Science, 2011, 38(6): 237-241.

[10] ZHOU J, PEDRYCZ W, MIAO D Q. Shadowed sets in the characterization of rough-fuzzy clustering [J]. Pattern Recognition, 2011, 44(8): 1738-1749.

[11] FRANK A, ASUNCION A. UCI machine learning repository [EB/OL]. [2012-07-12]. http://archive.ics.uci.edu/ml.

[12] LIKAS A, VLASSIS N, VERBEEK J J. The global *K*-means clustering algorithm [J]. Pattern Recognition, 2003, 36(2): 451-461.

[13] 谢娟英，张琰，谢维信，等. 一种新的密度加权粗糙 *K*-均值聚类算法 [J]. 山东大学学报：理学版，2011，45(7)：1-6.
XIE Juanying, ZHANG Yan, XIE Weixin, et al. A novel rough *K*-means clustering algorithm based on the weight of density [J]. Journal of Shandong University: Natural Science, 2011, 45(7): 1-6.

[14] DAVIES D L, BOULDIN D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227.

[15] 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.

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

（编辑 杜秀杰 武红江）