引入 RNA 计算的遗传模糊 C 均值聚类算法

林 春,李安贵,刘钦圣
LIN Chun,LI An-gui,LIU Qin-sheng

北京科技大学 应用科学学院 数学力学系,北京 100083
School of Applied Science,University of Science and Technology Beijing,Beijing 100083,China

LIN Chun,LI An-gui,LIU Qin-sheng.Genetic Fuzzy C-means algorithm adding in computing of RNA.Computer Engineering and Applications,2009,45(24):50-52.

Abstract: The algorithm of FCM is applied extensively in fuzzy clustering analysis,but it has two disadvantages:The first one is that it can easily be trapped in a local optimum and also strongly depends on initialization,and the second one lies in its long time of computing a large number of data.The RNA computing which is based on the DNA computing is a new algorithm of intelligent optimum.To improve the ability of getting the global best solution and to increase the convergent speed,a genetic fuzzy cluster algorithm based on RNA computing(RNAGAFCM) is presented.The emulational experiment of RNAGAFCM shows that the new algorithm decreases the iterative times and increases the convergent speed.

Key words: Fuzzy C-means algorithm;RNA computing;genetic algorithm

摘 要: 模糊 C 均值算法(FCM)在聚类分析中是目前比较流行和应用比较广泛的一种算法。但它存在两个弱点:一是对初始化非常敏感,容易陷入局部极值点;二是处理大数据集时耗时太长。基于 RNA 的分子计算是近年来新兴的一种智能优化计算方法。提出了基于 RNA 计算的遗传模糊聚类算法(RNAGAFCM),来提高收敛速度和全局寻优能力。仿真实验表明新算法比现有的遗传模糊聚类算法减少了迭代次数,提高了收敛速度。

关键词: 模糊 C 均值算法;RNA 计算;遗传算法
DOI: 10.3778/j.issn.1002-8331.2009.24.016 **文章编号:** 1002-8331(2009)24-0050-03 **文献标识码:** A **中图分类号:** TP18

1 引言

模糊 C 均值算法(FCM)在模糊聚类分析中是一种应用最为广泛、最为灵敏的算法。但是它的一个致命弱点是对初始化非常敏感而容易陷入局部极小值[1]。针对这些问题,人们提出了将遗传算法与 FCM 算法结合,来解决对初始化敏感的问题。遗传算法是一种应用广泛的全局优化方法,它的主要优点是简单、通用、鲁棒性强和适合并行处理[2]。它比盲目的搜索效率要高得多,又比专门的针对特定问题的算法通用性强,是一种与问题无关的求解模式。因此把遗传算法与 FCM 结合起来,既能发挥遗传算法(GA)的全局寻优能力,又可以兼顾 FCM 的局部寻优能力。但是遗传算法仍然具有后期搜索效率低、局部寻优能力弱、易早熟等缺点。基于 RNA 的分子计算是近年来新兴的一种计算方法。基于 DNA 计算的 RNA 计算模型,以 DNA 计算为模板,根据互补配对原则,把 DNA 上携带的遗传信息传给 RNA。RNA 独特的单链结构和对基因信息腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和尿嘧啶(U)的垂直继承,使得 RNA 计算和遗传算法相结合成为可能。该文根据 RNA 计算与 GA 算法的特点提出一种基于 RNA 计算的遗传模糊 C 均值算法。

2 相关知识

2.1 模糊 C 均值算法(FCM)

FCM 聚类算法又称为基于目标函数的模糊聚类算法。算法的本质是要找到目标数据集的 c 个划分子集,再将目标数据集划分到每个子集中。它的基本思想是首先确定一个能够反映分类效果的目标函数,然后通过求解这个目标函数的极值来确定最佳的分类。目标函数为:

$$J_m(\boldsymbol{U},\boldsymbol{P})=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2 \tag{1}$$

m 即为加权指数,体现了模糊性。$\boldsymbol{U}$ 和 $\boldsymbol{P}$ 分别表示隶属度矩阵和聚类中心矩阵。$(d_{ik})^2$ 表示第 i 类中的样本与其类中心之间的距离一般表达式,其定义为:

$$(d_{ik})^2=\| x_k-p_i \|_A=(x_k-p_i)^{\mathrm{T}}\boldsymbol{A}(x_k-p_i) \tag{2}$$

当 $\boldsymbol{A}$ 取单位矩阵 $\boldsymbol{I}$ 时,上式对应欧几里德距离。聚类的准则为取 $J_m(\boldsymbol{U},\boldsymbol{P})$的极小值。用拉格朗日乘数法求解得到结果:

$$\begin{cases}\mu_{ik}=\dfrac{1}{\sum\limits_{j=1}^{c}\left(\dfrac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} & \text{当 } I_k=\varnothing \\ \mu_{ik}=0,\ \forall i\in\bar{I}_k,\text{以及}\sum\limits_{i\in I_k}\mu_{ik}=1 & \text{当 } I_k\neq\varnothing\end{cases} \tag{3}$$

作者简介: 林春(1983-),男,硕士研究生,研究方向:模糊数学及其应用,智能优化算法;李安贵(1949-),男,教授,研究方向:模糊数学及其应用;刘钦圣(1927-),男,教授。

收稿日期: 2008-05-13 **修回日期:** 2008-09-10

$$p_i=\frac{1}{\sum_{k=1}^{n}(\mu_{ik})^m}\sum_{k=1}^{n}(\mu_{ik})^m x_k \tag{4}$$

先随机确定 $\boldsymbol{P}$,再根据式(3)、式(4)反复修改聚类中心和隶属度,当算法收敛时理论上就得到了各个类的聚类中心。但是此算法却存在对初始化敏感,容易陷入局部最优的缺点。

2.2　基于 RNA 计算的遗传算法

遗传算法是建立在生物进化基础之上的算法,是一种基于自然选择和群体遗传机理的搜索算法[3]。

它模拟了自然选择和自然遗传过程中发生的繁殖、交配和突变现象。它将每个可能的解看做是群体(所有可能解)中的一个个体,并将每个个体编码成字符串形式,根据预定的目标函数对每个个体进行评价,给出一个适应度值。通过利用选择、交叉、变异三个遗传算子操作得到一群新的性状优良的个体,这样就逐步朝着更优解的方向进化。与传统算法相比,其能够从多个点构成的群体开始搜索,因此它不易陷入局部最优。将其应用到 C 均值聚类中可以利用其寻优能力来平衡聚类中心[4]。

基于 DNA 的 RNA 的计算是近年来新兴的一种计算方法。由于 RNA 是重要的基因物质,携带着丰富的遗传信息,从而能促进进一步模拟生物的遗传机理和基因调控机理。RNA 独特的单链结构和对基因信息腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和尿嘧啶(U)的垂直继承,使得基于 DNA 计算的 RNA 计算和遗传算法相结合成为可能。RNA 序列的解空间为 $E=\{A,U,G,C\}^L$,即 RNA 序列以 A、U、G、C 四个字母来对长度为 L, 并由尿嘧啶、胞嘧啶、腺嘌呤和鸟嘌呤四种碱基组成的序列进行编码。使用 0(00),1(01),2(10),3(11)四个数字对的四种碱基 A、U、G、C 进行编码[5],共有 $P_4^4=24$ 种可能的编码组合。如此编码,便于数学和逻辑操作,还体现了碱基互补的关系,如 C 和 G 碱基对是互补结合。该文提出的算法中所有讨论的均基于以上编码格式。

RNA 计算的独特方式还体现在它的操作算子上。转位算子、换位算子和置换算子是交叉操作的三种基本算子,在交叉过程中,综合使用以上三个算子,能够最大程度地增加染色体基因的多样性和保留基因中携带的遗传信息。

转位算子:将 RNA 序列中的一个子序列,转移至新的位置。

换位算子:将 RNA 序列中的两个或两段子序列互相交换位置。

置换算子:RNA 序列中的一个子序列被另一个子序列所替换。

3　基于 RNA 计算的遗传模糊聚类算法

3.1　模糊聚类的遗传算法

在编码方式上,针对聚类原型 P 编码而不针对划分矩阵 $\boldsymbol{U}$ 编码以减少搜索空间。同时采用 Gray 码编码方式,以提高编码精度。

编码:聚类中心为$[\boldsymbol{P}_{i-\min},\boldsymbol{P}_{i-\max}]$,其中最小值 $\boldsymbol{P}_{i-\min}=(p_{i1-\min},p_{i2-\min},\cdots,p_{ik-\min})^{\mathrm{T}}$,最大数值 $\boldsymbol{P}_{i-\max}=(p_{i1-\max},p_{i2-\max},\cdots,p_{ik-\max})^{\mathrm{T}}$。那么将 P_i 编码为$\{\beta_{i1},\beta_{i2},\cdots,\beta_{ik}\}$如下:

$$\beta_{ij}=round\left(\frac{p_{ij}-p_{ij-\min}}{p_{ij-\max}-p_{ij-\min}}\times 255\right) \tag{5}$$

解码:与编码互逆,若 β_{ij} 的 Gray 码解码后结果为 i,那么解码后结果为:

$$p_{ij}=p_{ij-\min}+\frac{i}{255}(p_{ij-\min}-p_{ij-\max}) \tag{6}$$

适应度函数:对于基于目标函数的模糊聚类,其目的是找到目标函数的极小值,即是目标函数越小,聚类效果越好,此时的适应度应该越大。因此可以利用目标函数来定义适应度函数 $f(\cdot)$:

$$\begin{cases} f(\boldsymbol{U},P)=[\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m D^2(x_k-p_i)+\zeta]^{-r} \\ D^2(x_k-p_i)=(x_k-p_i)^{\mathrm{T}}(x_k-p_i) \end{cases} \tag{7}$$

其中,ζ、r 为给定的常数,且 $r>0$,μ_{ik} 确定方式如下:

$$\begin{cases} I_k=\varnothing\Rightarrow\mu_{ik}=\left(\sum_{j=1}^{c}\left[\frac{D^2(x_k,p_i)}{D^2(x_k,p_j)}\right]^{\frac{1}{m-1}}\right)^{-1} \\ I_k\neq\varnothing\Rightarrow\mu_{ik}=0,\ \forall i\in\bar{I}_k,\ \sum_{i\in I_k}\mu_{ik}=1 \end{cases} \tag{8}$$

其中,$I_k=\{i|1\leqslant i\leqslant c,D(x_k,p_i)=0\}$,$\bar{I}_k=\{1,2,\cdots,c\}-I_k$。

3.2　基于 RNA 计算的遗传模糊聚类算法

在遗传算法中, 要确定交叉概率 P_c、遗传概率 P_m 两个参数。在 RNA 遗传模糊聚类中根据 RNA 计算的特点,确定两个参数如下:

$$P_c=\frac{(f_a-f_c)\cdot e^{a\cdot(f_c-f_c)}}{1+e^{a}} \tag{9}$$

其中 a 为一常数。

$$\begin{cases} P_{mi}=\dfrac{(f_a-f_m)\cdot e^{a\cdot\left(\frac{(i\bmod(l/2)+1)}{l/2}(1-b)+b\right)}}{1+e^{a}},\ f_a-f_m\geqslant c \\ P_{mi}=\dfrac{(f_a-f_m)\cdot e^{a\cdot\left(\frac{((l/2-i)\bmod l/2+1)}{l/2}(1-b)+b\right)}}{1+e^{a}},\ f_a-f_m<c \end{cases} \tag{10}$$

其中 a 同 P_c,b 为[0,1]之间的常数,与 a 同时起到平衡高位低位对变异概率的影响。有文献在研究 DNA 序列模型时指出,在同一个序列的不同位置,存在 hot spot 和 cold spot,位于 cold spot 的碱基其变异概率远远小于位于 hot spot 的碱基。c 为 hot spot 和 cold spot 的转折点,当 $f_a-f_m\geqslant c$,每 n 位编码从低位到高位的变异概率应该逐渐增大。当 $f_a-f_m<c$,每 n 位编码从低位到高位的变异概率应该逐渐减小。两个公式中 f_a、f_c 和 f_m 的确定参见文献[1]。

基于 RNA-GA 的模糊聚类算法步骤如下:

(1)设置最大的进化代数、染色体的编码长度以及群体个数,并计算每个个体的适应度值。

(2)执行选择操作,采用最优保留策略。

(3)执行交叉操作,以概率 1 执行置换操作(操作过程是随机生成[0,L]之间的一个整数 k,将一对父本从 k 位置到基因末尾的编码进行置换,保留父本中第一个个体)。再将生成的新个体以概率 P_c 分别执行转位和换位操作。生成随机数 r_1, 若 $r_1<P_c$,执行转位操作。再生成随机数 r_2,若 $r_2<P_c$,则执行换位操作,这样共生成 $N-1$ 个个体。

(4)生成 L 个[0,1]之间的随机数 r_i,对于每一位上,若 $r_i<P_{mi}$,执行变异操作时,根据适应度值的大小确定不同位上的变异概率。相应的位置执行变异操作:生成[0,3]之间的随机整数替换原来此位置上的值。

(5)对于步骤(4)产生的 $N-1$ 个新个体,采用最优保留策略,转到步骤(2),直至满足终止条件。

3.3 算法的收敛性分析及仿真实验

定理 1 采用最优保留策略且经过 RNA 选择、杂交、变异算子变换的种群序列$\{\boldsymbol{X}(n);n\geqslant 0\}$以概率 1 收敛到满意种群集$M^*$,且收敛过程中种群满意值是单调不减的(证明过程需证明序列为马尔可夫链,此处略)。

为了验证上述算法的有效性,选取了具有典型性的两组数据进行验证。分别是 hayes-roth 数据集和 wine 数据集。计算都是每种方法计算 10 次然后取平均值,与传统结合了 GA 的 FCM 算法进行比较,结果记录在表 1 和表 2 中,RNAGAFCM 算法的参数确定为 a=0.5,b=0.9,c=0.05。

表 1 hayes-roth **数据集两种聚类结果**

	GAFCM	RNAGAFCM
迭代次数	43	37
	37	43
	42	47
	51	43
	52	38
	42	44
	46	36
	46	42
	40	39
	39	38
平均迭代次数	43.8	40.7
目标函数值 J_2	145.030 6	145.030 6

表 2 wine **数据集两种聚类结果**

	GAFCM	RNAGAFCM
迭代次数	28	5
	22	13
	8	18
	16	6
	27	14
	26	14
	8	11
	14	15
	15	13
	13	15
平均迭代次数	17.7	12.4
目标函数值 J_2	1.796 1e+006	1.796 1e+006

从表 1 和表 2 中可以看出经过 10 次计算后,新算法 RNAGAFCM 对于这两组数据达到最优解的平均迭代次数 40.7、12.4,要比 GAFCM 算法的 43.8、17.7 有一定程度的减小。图 1 和图 2 是以上两次计算的图形表示。

观察图 1 和图 2 发现新算法不但较早收敛到最优值,计算 10 次操作中每一步迭代的平均值达到最优值的迭代次数,新算法(42,14)比传统算法(47,25)也有一定的优势。

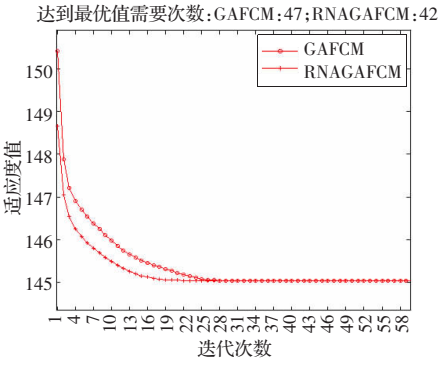

图 1 hayes-roth **数据集两种聚类结果**

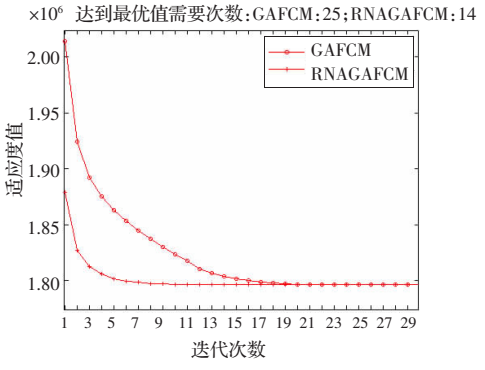

图 2 wine **数据集两种聚类结果**

对于算法中的各个参数,通过仿真实验确定(选定 iris 数据集)。图 3~图 5 为确定参数的实验结果。

从图 3~图 5 中可以看到,算法的收敛速率和寻优性能对参数 a、b 的设置比较敏感,而对参数 c 的设置不是太敏感。从图 3 看出对于参数 a,其值最好设置在 0.01~0.5 之间。从图 4 看出对于参数 b 的值最好设置在 0.9 左右,若过大,则基因的高、低位的差别显现不出来,若过小,则变异概率过大,容易成为随机搜索。从图 5 看对于参数 c 来说,改变其值对迭代次数没有多大影响,但是从时间来看,建议其值设置在 0.5~0.7 之间。

4 结语

提出了一种新的模糊聚类算法:基于 RNA 计算的遗传模糊聚类算法。实验表明该算法在解决聚类问题上有减少迭代次数和提高聚类性能的优势。算法本身实际上是将 RNA 计算与

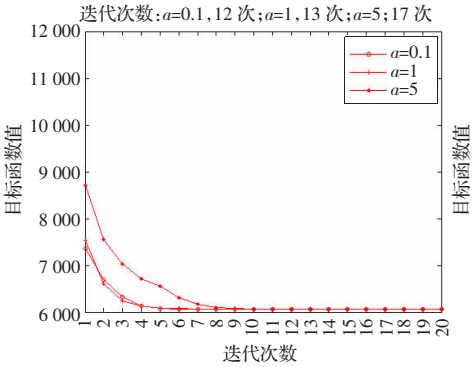

图 3 iris **数据集不同参数** a **聚类结果**

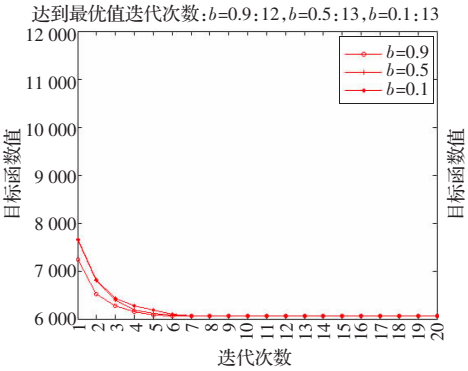

图 4 iris **数据集不同参数** b **聚类结果**

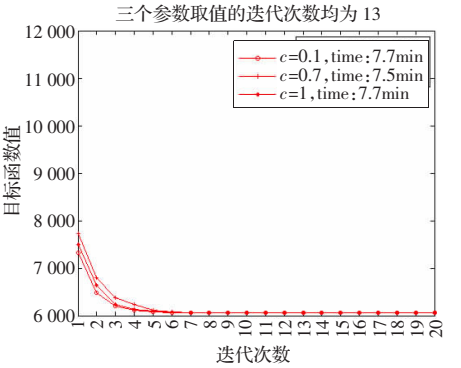

图 5 iris **数据集不同参数** c **聚类结果**

(下转 55 页)