一种基于 CSA 的模糊聚类新算法

李 洁　　高新波　　焦李成
（西安电子科技大学电子工程学院　西安　710071）

摘　要：在聚类分析中，模糊 k 均值算法是目前应用最为广泛的方法之一，然而该算法对初始化敏感，容易陷入局部极值点。为此，该文提出一种基于克隆选择的模糊聚类新算法以实现全局优化处理。在新算法中，由于克隆算子能够将进化搜索与随机搜索、全局搜索和局部搜索相结合，因而通过对候选解进行克隆算子操作，能够快速得到全局最优解。用人造数据和 IRIS 实际数据所做测试结果表明了新算法的有效性。

关键词：聚类分析，克隆选择算法，模糊 k 均值算法，遗传算法
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2005)02-0302-04

A CSA-Based New Fuzzy Clustering Algorithm

Li Jie　　Gao Xin-bo　　Jiao Li-cheng
(*School of Electronic Engineering, Xidian Univ., Xi'an* 710071, *China*)

Abstract　In cluster analysis, Fuzzy K-Means (FKM) algorithm is one of the most widely used methods. However, FKM algorithm is much more sensitive to the initialization, and easy to fall into local optimum. For this purpose, this paper presents a clonal selection based new algorithm for fuzzy clustering analysis, for global optimization. Since the clonal operator can combine the evolutionary search and random search, and incorporate the global search with local search, by the clonal operation on candidate solutions, the new algorithm can quickly obtain the global optimum. The experimental results with synthetic data and IRIS real data illustrate the effectiveness of the new algorithm.

Key words　Cluster analysis, Clonal selection algorithm, Fuzzy k-means algorithm, Genetic Algorithm(GA)

1　引言

聚类分析是多元统计分析的方法之一，也是统计模式识别中非监督模式分类的一个重要分支[1]。在传统的聚类方法中，基于目标函数的聚类算法由于把聚类问题归结为一个优化问题，具有深厚的泛函基础，从而成为聚类算法研究的主流，k-均值算法就是其中最典型的一种[2]。但由于 k-均值聚类目标函数是高度非线性和多峰的函数，因此，用梯度法优化目标函数时，搜索方向总是沿着能量减小的方向，使算法很容易陷入局部极值点，只有当初始化较好时算法才能收敛到全局最优解。为此，随着遗传算法(Genetic Algorithm, GA)的出现，人们提出了基于 GA 的聚类方法，尽管该方法能以较高的概率收敛到全局最优点，但收敛速度较慢，而且还容易出现早熟[2]。

为解决这些问题，本文提出一种基于克隆选择算法(Clonal Selection Algorithm, CSA)的模糊聚类新方法。CSA 是一种新兴的人工免疫系统方法[3-5]，它借助于生物学免疫系统的抗体克隆选择机理，构造适用于人工智能的克隆算子。由于基于克隆算子的克隆选择算法是群体搜索策略，本质上固有并行性和搜索变化的随机性，在搜索中不易陷入局部极小值，最终能以较大的概率获得问题的全局最优解，且具有较快的收敛速度。因此，基于 CSA 的模糊聚类新算法将会具有较高的效率，并且也适合于大数据集的聚类分析。

本文的安排如下：第2节简单介绍模糊聚类算法，第3节讨论CSA以及基于CSA的模糊聚类算法，第4节为实验结果，并将本文提出的新方法与遗传算法和模糊k-均值算法进行了性能比较。最后总结全文，并指出进一步的研究方向。

2　模糊聚类算法

设 $X=\{x_1,x_2,\cdots,x_n\}$ 是待聚类分析的数据集，其中 $x_i=[x_{i1},x_{i2},\cdots,x_{im}]^{\mathrm{T}}$ 表示第 i 个样本的 m 个特征值。聚类分析就是将 X 划分成 k 个不相交的子集 $X_1,X_2,\cdots,X_k$，要求满足下列条件：

$$X_1\cup X_2\cup\cdots\cup X_k=X;\quad X_i\cap X_j=\varnothing,\quad 1\le i\ne j\le k \tag{1}$$

样本 $x_l(1\le l\le n)$ 对子集（类） $X_i(1\le i\le k)$ 的隶属关系可用隶属函数表示为

$$\mu_{X_i}(x_l)=\mu_{il}=\begin{cases}1, & x_l\in X_i\\ 0, & x_l\notin X_i\end{cases} \tag{2}$$

2003-09-18 收到，2004-02-25 改回
国家自然科学基金和 863 计划资助课题

其中，隶属函数必须满足条件 $\mu_{il} \in M_h$，

$$M_h = \left\{ \mu_{il} \middle| \mu_{il} \in \{0,1\};\ \sum_{i=1}^{k} \mu_{il} = 1, \forall l;\ 0 < \sum_{l=1}^{n} \mu_{il} < n, \forall i \right\} \quad (3)$$

在模糊聚类中，样本集 X 被划分成 k 个模糊子集 $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_k$，有 $\bigcup_{i=1}^{k} \mathrm{supp}(\tilde{X}_i) = X$，其中 supp 表示取模糊集合的支撑集[2]，而且样本的隶属函数 μ_{il} 从{0,1}二值扩展到[0,1]区间，满足条件 $\mu_{il} \in M_f$：

$$M_f = \left\{ \mu_{il} \middle| \mu_{il} \in [0,1];\ \sum_{i=1}^{k} \mu_{il} = 1, \forall l;\ 0 < \sum_{l=1}^{n} \mu_{il} < n, \forall i \right\} \quad (4)$$

对于一组给定的样本集，模糊聚类分析可以很容易获得它的一个模糊k-划分：$\boldsymbol{U} = \left\{ \mu_{il} \middle| 1 \le i \le k; 1 \le l \le n \right\}$。但是，要保证划分有意义，则需要依据问题的需要定义合适的划分准则 $D(\cdot)$。假定每个模糊子集 $\tilde{X}_i (1 \le i \le k)$ 有一个聚类原型 $\boldsymbol{p}_i$，则样本 $\boldsymbol{x}_l$ 与模糊子集 $\tilde{X}_i$ 的相似性可以通过它与聚类原型 $\boldsymbol{p}_i$ 间的失真度 $d_{il} = D(\boldsymbol{x}_l, \boldsymbol{p}_i)$ 来度量，一般情况下，$D(\cdot)$ 是定义为欧几里德距离的相异性测度。对于具有实特征的数据集，即 $X \subset R^m$，则有

$$D^2(\boldsymbol{x}_l, \boldsymbol{p}_i) = (\boldsymbol{x}_l - \boldsymbol{p}_i)^{\mathrm{T}} (\boldsymbol{x}_l - \boldsymbol{p}_i) \quad (5)$$

为了保证聚类的结果能使“物以类聚”，就以极小化的每类样本与该类模式间的失真度构造模糊聚类的目标函数，然后通过优化该目标函数获得样本集的最佳模糊 k-划分 $\boldsymbol{U}^* = [\mu_{il}]$ 和每类的聚类原型 $P^* = \{\boldsymbol{p}_i, 1 \le i \le k\}$。常见的模糊聚类目标函数形式如下：

$$C(\boldsymbol{U}, P) = \sum_{i=1}^{k} \sum_{l=1}^{n} \mu_{il}^{\alpha} \left(D(\boldsymbol{x}_l, \boldsymbol{p}_i) \right)^2, \quad \mu_{il} \in [0,1] \quad (6)$$

其中 α 是模糊度控制参数。模糊 k 均值(FKM)算法就是通过迭代优化法极小化目标函数而得到最终分类结果，但 FKM 聚类算法极易陷入局部极值点，而得不到最优划分。

3　基于CSA的模糊聚类算法

为了克服传统的 FKM 算法对初始化敏感的缺点，我们采用 CSA 来优化其目标函数。

3.1 克隆选择算法(CSA)

1958 年 Burnet 提出著名的克隆选择学说[6]：抗体以受体的形式存在于细胞表面，抗原可与之选择性地反应。该反应可导致细胞克隆性增殖。克隆性在细胞水平上表现出 T 细胞和 B 细胞抗原受体结构的极端多样性，因而直接导致了抗体网络的多样性、记忆性和特异性。CSA 正是基于抗体克隆选择这一生物特性而形成的一种新的人工免疫系统方法。

与进化计算一样，人工免疫系统方法也能解决函数优化问题。假设所需优化的函数为 $\varphi: \prod_{i=1}^{m} [d_i, u_i] \to R(d_i < u_i)$，其中 m 是优化变量的个数，变量 $x_i \in [d_i, u_i]$，则抗原就是被优化的函数 φ，抗体群 $\bar{\boldsymbol{A}} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_N\}$ 为抗体 $\boldsymbol{A}$ 的 N 元组，抗体 $\boldsymbol{A}_i$ 是解空间 S^m 中的一个点。

抗体–抗原亲和度函数 f 一般是 φ 的函数，抗体–抗体亲合力定义为

$$W_{ij} = \left\| \boldsymbol{A}_i - \boldsymbol{A}_j \right\|, \qquad i, j = 1, 2, \cdots, N \quad (7)$$

$\|\cdot\|$ 为任意范数，$\boldsymbol{W} = (W_{ij})_{N \times N}$ 为抗体–抗体亲和力矩阵，它反应了种群的多样性。

克隆选择算法可简单的描述为：

步骤 1　$l = 0$，初始化抗体群落 $\bar{\boldsymbol{A}}(0)$，设定算法参数，计算初始种群的亲和度；

步骤 2　依据亲和度和设定的抗体克隆规模，进行克隆算子操作[6]，获得新的抗体群 $\bar{\boldsymbol{A}}(l+1)$；

步骤 3　$l = l + 1$，若满足终止条件，停止计算；否则，返回步骤 2。

假设 $M = \{\bar{\boldsymbol{A}} \mid \max(f(\bar{\boldsymbol{A}})) = f^*, \forall \bar{\boldsymbol{A}} \in S^N\}$ 为满意种群，即满意种群集 M 中的任意抗体群 $\bar{\boldsymbol{A}}$ 中至少包含一个最优解。可以证明克隆选择算法生成的抗体种群序列 $\{\bar{\boldsymbol{A}}(l), l \ge 0\}$ 是有限非齐次可约马尔可夫链，它以概率 1 收敛到满意种群 M [5]。

3.2 基于 CSA 的模糊聚类算法

为了利用 CSA 求解数据集的模糊聚类问题，首先需要解决以下 3 个问题：(1) 如何将聚类问题的解编码到抗体中；(2) 如何构造抗体-抗原亲和度函数来度量每个抗体对聚类问题的响应程度；(3) 如何选择克隆算子，确定操作参数的取值，以确保快速收敛到最优解。

3.2.1 *编码方案*　我们由式 (6) 定义的聚类目标函数可知，聚类的目的就是要获得数据集 X 的一个模糊划分矩阵 $\boldsymbol{U}$ 和聚类的原型 P。而 $\boldsymbol{U}$ 和 P 是相关的，即已知其一则可求得另一个的解，所以，在基于 CSA 的模糊聚类算法中，我们可令一组聚类原型 P 就是一个抗体，这样把原型中的 k 组特征连接起来，根据各自的取值范围，就可以将其量化值编码成一个抗体。

$$\begin{aligned} \bar{\boldsymbol{A}} &= \left\{ \underbrace{\zeta_1, \zeta_2, \cdots, \zeta_m}_{\mathrm{Encode}(p_1) = A_1(0)}, \cdots, \xi_i, \cdots, \underbrace{\zeta_{(k-1)m+1}, \zeta_{(k-1)m+2}, \cdots, \zeta_{km}}_{\mathrm{Encode}(p_k) = A_k(0)} \right\} \\ &= [\boldsymbol{A}_1(0), \boldsymbol{A}_2(0), \cdots, \boldsymbol{A}_k(0)] \end{aligned} \quad (8)$$

式中参数集依据每个原型 $\boldsymbol{p}_i$ 取值。

3.2.2 *抗体–抗原亲合度函数构造*　由聚类目标函数定义可知，目标函数越小，则聚类效果越好，而此时抗体–抗原亲和度应该越大。因此我们借助目标函数来构造抗体–抗原亲和度函数如下式：

$$f(\boldsymbol{A}) = \frac{1}{1 + C(\boldsymbol{U}, P)} = \frac{1}{1 + \sum_{i=1}^{k} \sum_{l=1}^{n} \mu_{il}^{\alpha} \left(D(\boldsymbol{x}_l, \boldsymbol{p}_i) \right)^2} \quad (9)$$

3.2.3 *克隆算子选择*　在基于 CSA 的模糊聚类算法中，克隆算子包括克隆操作、免疫基因操作（克隆重组和克隆变异）以及克隆选择和克隆死亡 4 个部分。

(1) 克隆操作 T_c^C　定义克隆操作为

$$T_c^C(\bar{A}(l))=[T_c^C(A_1(l))\,T_c^C(A_2(l))\cdots T_c^C(A_k(l))] \tag{10}$$

其中 $T_c^C(A_i)=A_i\otimes I_i$，$i=1,2,\cdots,k$；$I_i$ 为 q_i 维全 1 行向量，称它为抗体 A_i 的 q_i 克隆。q_i 的大小取决于 $f(A_i)$ 和 N_c，$N_c>N$ 是与克隆规模有关的设定值。一般情况下：

$$q_i=\mathrm{Int}\left(N_c\times f(A_i)\Big/\sum_{j=1}^{k}f(A_j)\right),\qquad i=1,2,\cdots,k \tag{11}$$

Int(x) 表示对 x 上取整。经克隆操作后，抗体种群变为

$$\bar{A}'(l)=[\bar{A}(l)\ A_1'(l)\ A_2'(l)\cdots A_k'(l)] \tag{12}$$

其中

$$A_i'(l)=[A_{i1}(l)\,A_{i2}(l)\cdots A_{iq_i-1}(l)],\quad A_{ij}=A_i,\quad j=1,2,\cdots,q_i-1 \tag{13}$$

(2) 免疫基因操作 T_g^C　(a)克隆变异依据概率 p_m^i 对克隆后的抗体群进行变异操作，为了保留抗体原始种群的信息，变异算子并不作用于 $\bar{A}(l)\in\bar{A}'(l)$。(b)克隆重组克隆变异之后，对新的抗体群依据概率 p_c^i 进行克隆重组操作，同样为了保留抗体原始种群的信息，克隆重组算子并不作用于 $\bar{A}(l)\in\bar{A}'(l)$。

(3) 克隆选择操作 T_s^C　对于 $\forall i=1,2,\cdots,k$，存在新抗体 $B=\{A_{ij}'(l)\,|\,\max f(A_{ij}'(l)),\ j=1,2,\cdots,q_i-1\}$，则 B 取代 $A_i(l)\in\bar{A}(l)$ 的概率为

$$p_s(A_i\to B)=\begin{cases}1, & f(A_i)\ge f(B)\\ \exp\{[-f(A_i)-f(B)/\beta]\}, & f(A_i)\ge f(B)\\ & \text{且}A_i\text{不是目前种群的最优解}\\ 0, & f(A_i)\ge f(B)\\ & \text{且}A_i\text{是目前种群的最优解}\end{cases} \tag{14}$$

式中 $\beta>0$ 是一个与抗体种群多样性有关的参数，一般 β 取值越大，多样性越好。

(4) 克隆死亡操作 T_d^C　经克隆选择后，新的抗体群为

$$\bar{A}(l+1)=[A_1(l+1),\ A_2(l+1),\ \cdots,A_k(l+1)] \tag{15}$$

若在 $\bar{A}(l+1)$ 中存在 $A_i(l+1)$ 和 $A_j(l+1)$，满足 $f(A_i(l+1))=f(A_j(l+1))=\max f(\bar{A}(l+1))\quad i\ne j$，则可随机产生一个新的抗体，以概率 p_d 任意死亡 $A_i(l+1)$ 和 $A_j(l+1)$ 中的一个。

除了上述算子以外，我们在该聚类算法中又定义了一个新的算子：一步迭代算子。对于每个抗体，将其解码到聚类原型 P 之后，再进行一步迭代算子。该算子包含以下两个步骤：

$$\mu_{ij}=\left(\sum_{l=1}^{k}\left(D(\boldsymbol{x}_j,\boldsymbol{p}_l)\right)^{-2}\right)^{-1}\Big/\left(D(\boldsymbol{x}_j,\boldsymbol{p}_i)\right)^2,\ \forall i,j \tag{16}$$

$$\boldsymbol{p}_i=\sum_{j=1}^{n}\mu_{ij}^{\alpha}\boldsymbol{x}_j\Big/\sum_{j=1}^{n}\mu_{ij}^{\alpha} \tag{17}$$

在根据式(17)获得了新的聚类原型后，再将其编码到抗体中，重新进行上述克隆算子的操作，直到聚类原型收敛到最优解。

本文提出的基于 CSA 的模糊聚类新算法的流程可用下图表示。

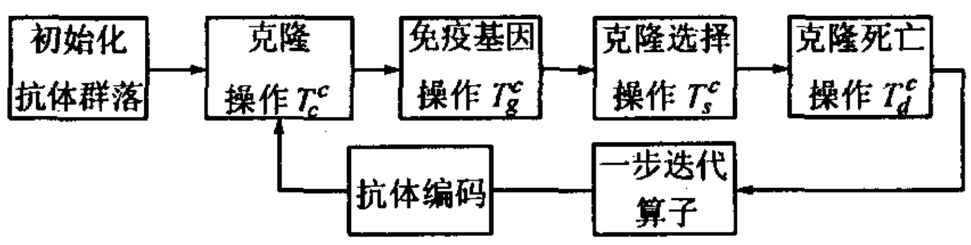


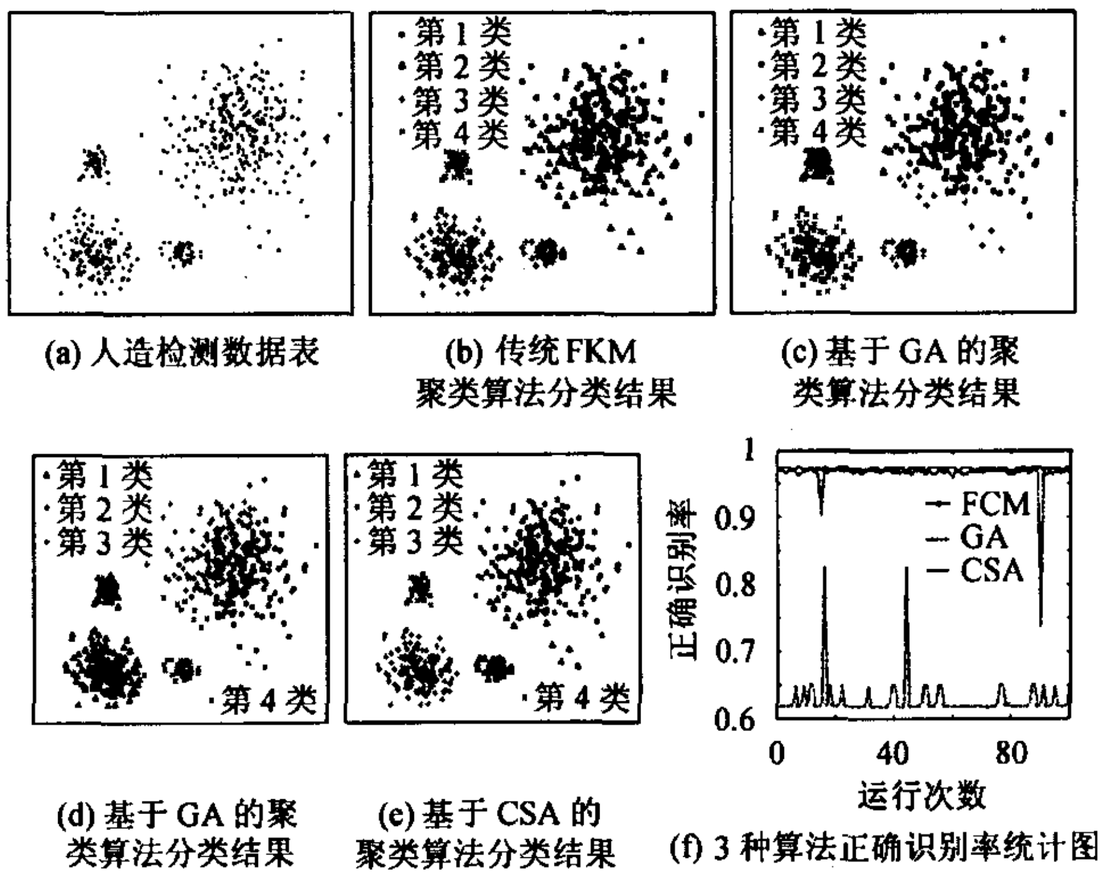
图 1　基于 CSA 的聚类新算法的流程框图

4　实验结果

为了测试基于CSA的聚类算法的有效性，我们给出一些初步的实验结果，对基于CSA的FKM聚类算法、基于遗传算法（GA）的FKM算法以及传统的FKM算法进行了比较实验。从分类性能及收敛速度两方面进行了比较分析，并测试了一步迭代算子对基于CSA的FKM算法的收敛性能的影响。人造和实际两类数据的测试实验结果均显示出新算法的优良性能。

4.1 新算法分类性能检验

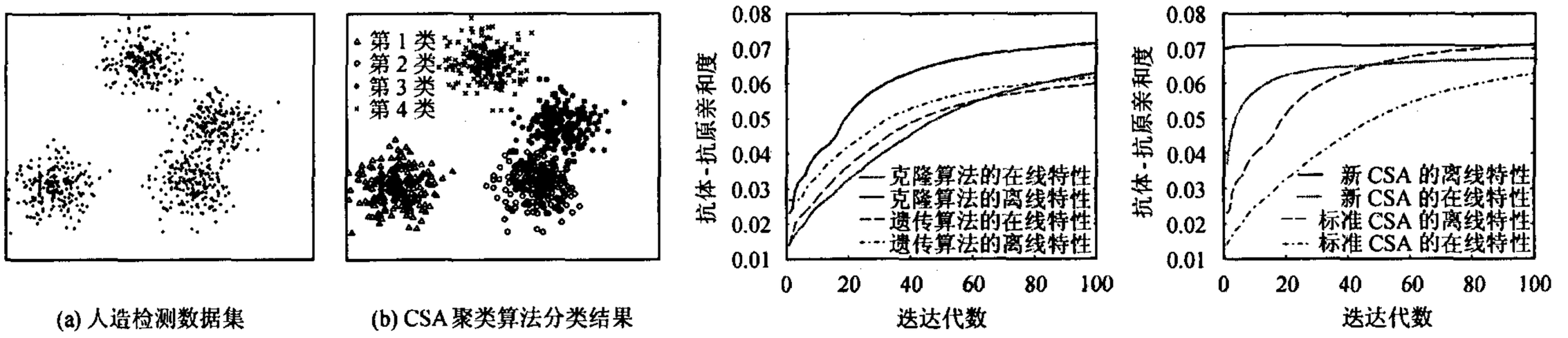
为了便于直观显示，我们构造的数据样本仅具有两个特征，是由 4 组不同方差的正态分布的二维数据点合成的，共包含 550 个样本，如图 2(a)所示。图 2(b)是用传统 FKM 算法 ($\alpha=2$) 分类结果；可以看出由于样本集中，几类数据的分布方差不同，传统 FKM 算法陷入局部极值点。图 2(c)和 2(d)均为基于 GA 的 FKM 的聚类结果（人口数 $N=100$，交叉率、变异率自动获得，选择概率 0.6），其中，图 2(c)搜索到全局最优解，而图 2(d)却产生早熟。图 2(e)是本文新算法聚类的结果（人口数 $N=100$，$N_C=200$，交叉率 0.8,变异率 0.01，死亡率 0.4），图中显示基于 CSA 的聚类算法得到的是该数据集的最优划分。



(a) 人造检测数据表　(b) 传统FKM聚类算法分类结果　(c) 基于 GA 的聚类算法分类结果

(d) 基于 GA 的聚类算法分类结果　(e) 基于 CSA 的聚类算法分类结果　(f) 3 种算法正确识别率统计图

图 2　测试实验结果(一)



(a) 人造检测数据集　(b) CSA 聚类算法分类结果　(c) CSA 与 GA 的离、在线特性　(d) 本文算法的收敛曲线

图 3　测试实验结果(二)

我们在同一数据集上，采用以上 3 种方法分别独立运行 100 次，统计出各自的正确识别率，如图 2(f)所示。从图中我们看到，在 100 次运行结果中，基于 CSA 算法得到的总是全局最优解，从而验证了基于 CSA 的算法能以概率 1 收敛到全局最优；而基于 GA 的算法有 2 次发生早熟，没有得到最优划分；同样由于传统的 FKM 对初始化敏感，因此其正确识别率较低，只有 2 次识别率高于 80%。在 100 次实验中，基于 CSA 的算法平均正确识别率为 96.79%，基于 GA 的算法平均正确识别率为 96.47%， 而 FKM 算法的平均正确识别率只有 63.13%。

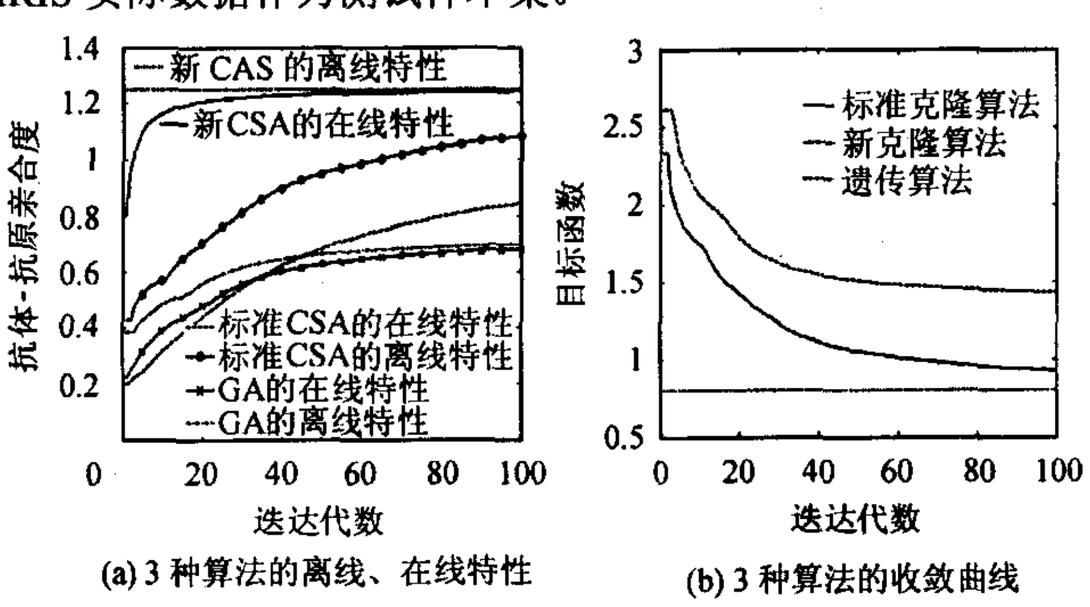
4.2　新算法收敛性检验

为了检验新算法的收敛性，我们分别用人造和实际数据进行了测试实验。

4.2.1 *人造数据集测试实验*　测试数据集为包含了 4 组正态分布的二维点的合成数据，共包含 800 个样本，如图 3(a)所示。图 3(b)是用本文新方法得到的聚类结果。

图 3(c)分别给出了基于标准克隆算法和遗传算法的模糊 k-均值聚类算法的在线和离线特性。从中我们可以看到：在每一代中，基于克隆算法得到的最优解的抗体-抗原亲和度比遗传算法的要好，而且收敛速度更快。图 3(d)是我们提出的增加了一步迭代算子的 CSA 聚类新算法以及标准 CSA 聚类算法的离线和在线特性，该图清楚地显示出：在引入一步迭代算子的操作之后，我们新算法的收敛速度有了明显的提高。

4.2.2 *实际数据集测试实验*　在该实验中，我们采用著名的 IRIS 实际数据作为测试样本集。



(a) 3 种算法的离线、在线特性　(b) 3 种算法的收敛曲线

图 4　测试实验结果(三)

图 4(a)分别显示了上述 3 种算法的离线和在线特性，我们看到，基于 GA 的聚类算法收敛速度最慢，基于标准 CSA 的算法比 GA 算法收敛速度快，而且种群的多样性保持较好，而本文的新算法收敛速度最快，而且迅速收敛到全局最优解。图 4(b)给出了 3 种算法在每次迭代中，最优解的目标函数值。

5　结束语

本文提出将克隆选择算法与模糊k-均值算法相结合用于分析大数据集聚类问题。并在人造和实际两类数据集上对其聚类性能进行了综合评估。实验结果表明新方法能够有效地发现数据中聚类结构，而且基于CSA的新算法收敛速度快，不依赖于算法的初始化，以概率1收敛到全局最优解。

本文的重点放在如何利用克隆选择算法解决聚类问题。然而，在运用该方法解决实际的聚类问题时，还要面对这样一个问题：数据集应该分为几类才合适？即聚类有效性问题，这将是我们进一步研究的重要方向。

参 考 文 献

[1] 何清. 模糊聚类分析理论与应用研究进展. 模糊系统与数学, 1998, 12(2): 89 – 94.

[2] 高新波. 模糊聚类算法的优化及应用研究. [博士论文], 西安: 西安电子科技大学, 1999 年.

[3] De Castro L N, Von Zuben F J. The clonal selection algorithm with engineering applications. Proc. of GECCO'00, Workshop on Artificial Immune Systems and Their Applications, Las Vegas, USA, 2000: 36 – 37.

[4] Kim Jungwon, Bentley P J. Towards an artificial immune system for network intrusion detection: An investigation of clonal selection with a negative selection operator. Proc. of the 2001 Congress on Evolutionary Computation, Seoul, Korea, 2001, 2: 1244 – 1252.

[5] Du Haifeng, Jiao Licheng. Clonal operator and antibody clonal algorithm. Proc. of the First International Conference on Machine Learning and Cybernetics, Beijing, 4 – 5 Novermber, 2002: 506 – 510.

[6] 周光炎.免疫学原理.上海:上海科学技术出版社, 2000: 31 – 32.

李　洁：女, 1972年生，讲师, 硕士, 主要研究方向为人工智能、模式识别.

高新波：男, 1972年生，教授, 博士生导师, 主要研究方向为模式识别、图象处理、人工智能等.

焦李成：男, 1959年生，教授，博士生导师, 主要研究方向为模式识别、图象处理、人工智能等.