免疫规划+K均值混合聚类算法*)

行小帅[1,2] 潘 进[3] 焦李成[1]

(西安电子科技大学雷达信号处理国家重点实验室 西安710071)[1]

(山西师范大学物理与信息工程学院 临汾041004)[2] (西安通信学院计算机与信息工程系 西安710106)[3]

A Novel Hybrid Clustering Algorithm Incorporating K-means into Canonical Immune Programming Algorithm

XING Xiao-Shuai[1,2] PAN Jin[3] JIAO Li-Cheng[1]

(Key Lab for Radar Signal Processing, Xidian University, Xi'an 710071, China)[1]

(Colloge of Physics and Information Engineering, Shanxi Teacher's University, Linfen, Shan'xi 041004, China)[2]

(Dep of Computer and Information Engineering, Xi'an Communications Institute, Xi'an 710106, China)[3]

Abstract A novel Hybrid Clustering Algorithm (HCA) that incorporates the K-means into the canonical immune programming algorithm is proposed after analyzing the advantages and disadvantages of the classical k-means clustering algorithm in the paper. The theory analyse and experimental results show, the algorithm not only avoids the local optima and is robust to initialization, but also increases the convergent speed, meanwhile evidently restrains the degenerating phenomenon during the evolutionary process.

Keywords Immune programming, K-means clustering, Adept degree, Clustering algorithm

1. 引言

聚类分析(Clustering Analysis)是一种无监督的模式识别方法。聚类产生的每一组数据称为一个簇,簇中的每一数据称为一个对象。聚类的目的是使同一簇中对象的特性尽可能地相似,而不同簇对象间的特性差异尽可能地大。聚类的任务是把一个未标记的模式按某种准则划分成若干子集,要求相似的样本尽量归为同一类,而不相似的样本归为不同的类,故又称无监督分类。目前,各种聚类方法已广泛应用于数据挖掘、模式识别、图像处理、雷达目标检测、生物工程、空间遥感技术等诸多领域[1~3]。

K均值聚类算法也称硬C均值聚类算法,1967年由MacQuen首先提出[4,5]。该算法是解决聚类问题的一种经典算法,具有算法简单、快速而且能有效地处理大数据库的优点。然而这种算法对不同的初始值可能会导致不同的聚类结果;在线K均值算法的执行结果对输入顺序有关。其次,这种算法是基于目标函数的聚类算法,一般都采用梯度法求解极值,由于梯度法的搜索方向总是沿着能量减小的方向进行,因此易陷入局部极小值。这两大缺陷较大地限制了它的应用范围。为了克服上述缺点,近几年来人们提出了对目标函数进行优化的各种算法,1989年文[6]提出了模拟退火法对硬分类矩阵U进行退火处理的C-均值算法;1994年文[7]提出了利用进化策略对目标函数进行聚类的方法;1999年文[8]根据遗传算法的原理,提出了以K均值算子代替遗传算法中的交叉算子,设计了一种混合遗传算法,并根据Gunter引入的有限状态齐次马尔可夫链方法[9]证明了该算法是以概率为1收敛于全局最优的。2000年文[10]采用了聚类中心的浮点编码方式,设计了浮点数交叉和变异算子,从而提高了遗传算法的搜索效率。但是,实验表明,当样本数目、类别和维数较多时,这两种基于遗传算法的聚类方法常会产生早熟现象。2001年文[11]为了提高收敛速度,减小早熟现象,采用遗传搜索与K均值局部优化相结合的混合遗传算法,并结合聚类问题的实际情况,设计了一种有效的基于最近基因匹配的交叉算子,使得交叉过程能不断地产生有意义的新个体,保证了群体的多样性,在一定程度上减小了早熟现象。然而,由于进化算法在进化过程中不可避免地会产生退化现象,因此将导致进化后期的波动现象以及迭代次数过长和聚类准确率不高的现象。

免疫规划[12]是一种有效的全局寻优算法,它具有两个显著的特点。第一,继承了进化规划的"随机进化,适者生存"的特点,因此具有良好的全局寻优性。第二,引入了对不良基因组合模式的"免疫"功能,可以利用人脑对问题的认识避免完全随机算法对无解区域的搜索,从而减少了计算,提高了速度。本文正是将免疫规划算法的全局搜索能力与传统的K均值聚类算法的局部优化能力相结合,提出了免疫规划+K均值混合聚类算法(Hybrid Clustering Algorithm, HCA)。理论分析和仿真结果表明,该算法不仅有效地避免了局部极小值问题和初始化选值敏感性的缺点,而且可以克服进化后期的波动现象并且有较快的收敛速度。

2. K均值聚类算法

该算法将 n 个向量 $x_j(1,2,\cdots,n)$ 划分为 c 个组 $G_i(i=1,2,\cdots,c)$,并求每组的聚类中心,使得非相似性(或距离)指标的价值函数(目标函数)达到最小。当选择欧氏距离为组 j 中向量 x_k 与相应聚类中心 c_i 间的相似性指标时,价值函数可定

*)本文受国家自然科学基金(60133010)和(60073053)、山西师大校科研基金资助。行小帅 副教授,访问学者,主要研究领域包括:数据挖掘、进化算法、人工神经网络与智能信息处理等。潘 进 教授,博士,主要研究领域包括:进化算法与多用户检测、子波理论与应用等。焦李成 教授,博士生导师,主要研究领域包括:非线性理论、人工神经网络、子波理论与应用、进化算法、数据挖掘与多用户检测等。

义为：

$$J=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\left(\sum_{k,x_k\in G_i}\|x_k-c_i\|^2\right) \tag{1}$$

式中 $J_i=\sum_{k,x_k\in G_i}\|x_k-c_i\|^2$ 是组 i 内的价值函数。并且 J_i 的值取决于 G_i 的几何特性和 c_i 的位置。

划分过的组可用一个 $c\times n$ 的二维隶属矩阵 U 来定义。如果第 j 个数据点属于 i 组，则 U 中的元素 u_{ij} 为1；否则，该元素取0。聚类中心 c_i 一旦确定，则可导出如下使式(1)最小 u_{ij}：

$$u_{ij}=\begin{cases}1 & for\ k\neq i, if\ \|x_j-c_i\|^2\leqslant\|x_j-c_k\|^2,\\ 0 & other\end{cases} \tag{2}$$

应当注意，如果 c_i 是 x_j 的最近的聚类中心，则 x_j 属于 i 组。由于一个给定的数据点只能属于一个组，因此隶属矩阵 U 具有如下性质：

$$\sum_{i=1}^{c}u_{ij}=1,\ \forall\ j=1,\cdots,n\quad 且\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}=n$$

另一方面，如果固定 u_{ij}，则使式(1)最小的最佳的聚类中心就是 i 组中所有向量的均值：

$$c_i=\frac{1}{|G_i|}\sum_{k,x_k\in G_i}x_k \tag{3}$$

式中 $|G_i|$ 是 G_i 的模或 $|G_i|=\sum_{j=1}^{n}u_{ij}$。该算法的主要步骤如下：

算法1：K 均值聚类算法

(1)初始化聚类中心 $c_i, i=1,\cdots,c$；一般的做法是从所有的数据中任取 c 个点；

(2)确定隶属矩阵 U。根据式(2)进行；

(3)计算价值函数。根据式(1)进行，如果其值小于某个确定的阈值，或它相对于上次价值数值的改变量小于某个阈值，则停止运行并输出结果。否则，继续；

(4)修正聚类中心。根据式(3)进行，转向第二步。

K 均值聚类具有算法简单，收敛速度较快的优点，但这种算法存在对初始化选值敏感性的问题和易陷入局部极小值，因此较大地限制了它的应用范围。为了保持 K 均值算法快速收敛的特点，同时又能避免算法收敛易陷入局部极小值和对初始选值敏感性的缺点，我们将免疫规划所具有的全局寻优和对不良基因组合模式免疫功能的特点与 K 均值算法所具有的局部搜索能力的优点相结合，提出了如下的聚类算法。

3. 混合聚类算法

免疫规划是在进化规划的基础上提出来的，它在进化规划中加入了免疫操作，在一定程度上抑制了进化规划完全随机寻优过程中的退化现象。可以证明，免疫规划是以概率为1收敛于全局最优个体的。由于本文的重点在于免疫规划在 K 均值算法中的应用，故这里只给出算法的步骤，而不做深入的理论分析。算法的基本框架如图1所示。

算法2：免疫规划＋K 均值混合聚类算法

设集合中的对象已抽象为空间 R^n 中的点，其相似程度已抽象为 R^n 中的距离。为了将集合中的点按照其相互间距离分为 K 类，可采取如下的步骤。

(1)产生初始群体及其编码：在集合中随机选取 N 组对象，每组 K 个点，代表着聚类的 K 个中心。将每一组对象编码为一个 $n\times K$ 维的种群个体 c，随机选取的 N 组对象构成了初始种群 C_0。也就是说，对搜索空间进行 $n\times K$ 维的实数编码。初始群体的规模为 N；

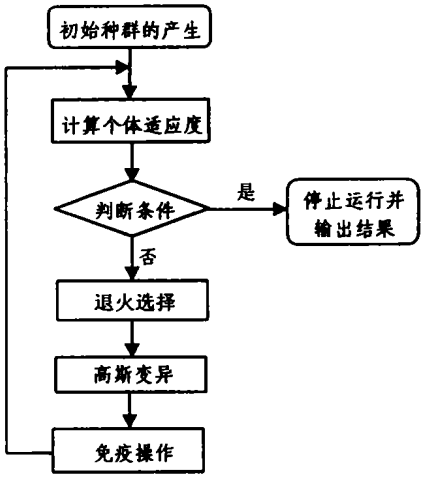

图1　免疫规划＋K 均值混合聚类算法

(2)计算适应度：当前群体中的每一个体 c，都对应着 K 个中心，对于这 K 个中心，按照 K 均值算法将集合分为 K 簇。对于第 i 簇 C_i，其簇内的距离定义为：

$$S_i=\frac{1}{|C_i|}\sum_{x\in C_i}\|x-z_i\| \tag{4}$$

式中 z_i 是簇 C_i 的均值。定义簇 C_i 和簇 C_j 之间的簇间距离为 $d_{ij}=\|z_i-z_j\|$，并定义 $R_i=\max_{j,j\neq i}\{(S_i+S_j)/d_{ij}\}$，则 Davies-Bouldin 聚类指标定义为：

$$DB=\frac{1}{K}\sum_{i=1}^{K}R_i \tag{5}$$

个体 c 的适应度定义为 $f(c)=1/DB$。在计算完一个个体的适应度之后，用所产生的 K 个均值 $\{z_i|i=1,\cdots,K\}$ 构成新的个体，并以此代替当前的个体 c。最后，进行停机条件的判断。若条件满足，则停止运行并输出结果。否则，继续；

(3)退火选择：是指在目前的子代群体 $C_{k-1}=(c^1,c^2,\cdots,c^N)$ 中以概率

$$P(c^i)=\frac{f(c^i)\exp(f(c^i)/T_k)}{\sum_{i=1}^{N}f(c^i)\exp(f(c^i)/T_k)} \tag{6}$$

选择个体 c^i 进入新的父代群体 A_k，其中 T_k 是单调递减趋于0的退火温度控制序列，其作用是控制个体多样性的变化速度；而且 T_0 的取值应充分大，以保证在搜索初期所有的个体都有较大的进化机会。随着 T_k 的减小，个体的多样性将逐步地受到限制。

(4)高斯变异：对当前的父代群体 A_k 进行高斯变异操作，产生新的子代群体 B_k。设

$$A_k=(a^1,a^2,\cdots,a^N),a^i\in\Omega\subset R^n,a^i=(a_1^i,\cdots,a_n^i),i=1,\cdots,N, \tag{7}$$

对个体 a^i 进行高斯变异，可按下式产生新的个体 b^i，

$$b_j^i=a_j^i+\eta_j^i,j=1,\cdots,n, \tag{8}$$

式中 η_j^i 服从高斯分布 $N(0,\sigma_{ij}^2)$，方差 σ_{ij}^2 根据下式计算，

$$\sigma_{ij}=\frac{T_kM_j+m_j}{T_k+f(a^i)} \tag{9}$$

对群体 A_k 进行高斯变异是指按照一定变异概率对群体中的部分个体进行高斯变异；

(5)免疫操作：由接种疫苗和免疫检测两部分构成。对当前子代群体 B_k 进行免疫操作，得到经免疫操作后的子代群体 C_k。

接种疫苗：设有个体 $b\in R^n$，$b=(b_1,\cdots,b_n)$。对 b 接种疫苗指的是按照人脑对问题的先验知识修改 b 的某些分量。这些先验知识可以是最优解的某些分量的大概取值范围，也可以是一些分量之间的某种关系。接种疫苗操作应该满足如下条件：若 b 已经是最优个体，则 b 以概率1转移为 b。设有群体 $B=(b_1,\cdots,b^N)$，对 B 进行接种疫苗操作指的是在 B 中按比例 α 随机抽取 $N_a:=\alpha N$ 个个体接种疫苗。疫苗是从对问题的先验知识中提出来的。疫苗中所含的信息量对算法的性能起着重要的作用；

免疫检测：是指对已接种疫苗的个体进行检测。若其适应度不如其父代，则说明在变异的过程中已发生了严重的退化。这时由其父代个体代替其参加选择竞争；

(6)返回第2步：对当前的子代适应度进行计算，并进行停机检测；

4. 仿真实验

第一个实验中分别采用 K 均值算法和本文提出的 HCA 算法对图2所示的500个点的2维点集进行了聚类。试验重复进行了1000次，表1给出了两种算法的性能比较，图2、3、4给出了其中随机抽取的一次试验的结果。图2是聚类的结果；图3的横坐标是迭代次数，纵坐标是进化过程中保留最佳个体时的最大适应度值。图4是进化过程中不保留最佳个体时的平均适应度随迭代次数的变化情况。

第二个试验中分别采用以上两种方法对图5所示的500个点的3维点集进行了聚类。试验进行了1000次，表2给出了两种算法的性能比较，图5、6、7给出了其中随机抽取的一次试验的结果。图5是聚类的结果；图6的横坐标是迭代次数，纵坐标是进化过程中保留最佳个体时的最大适应度值。图7是进化过程中不保留最佳个体时的平均适应度随迭代次数的变化情况。

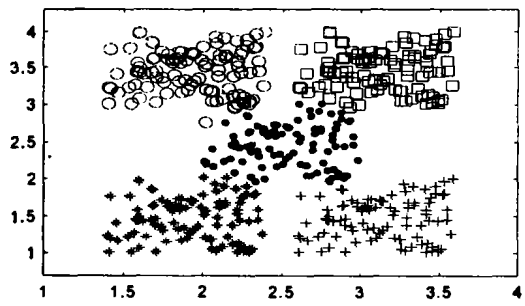

图2　实验1中对500个点的2维点集进行分类的结果

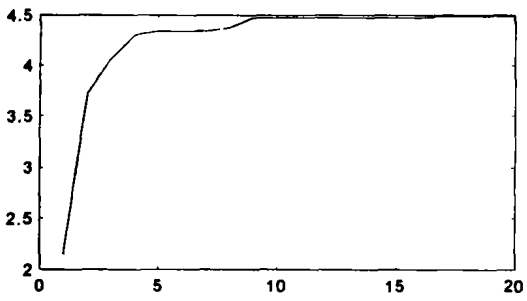

图3　实验1的进化过程中保留最佳个体时的最大适应度随迭代次数的变化

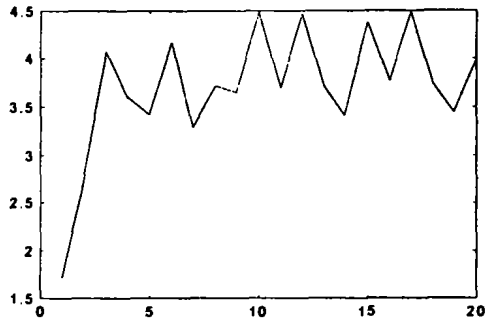

图4　实验1的进化过程中不保留最佳个体时的平均适应度随迭代次数的变化情况

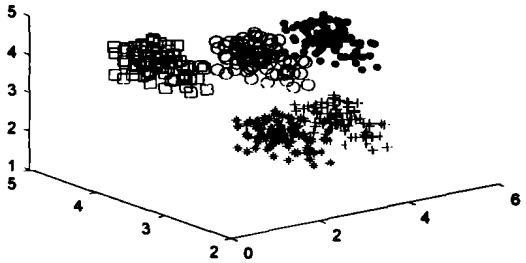

图5　实验2中对500个点的3维点集进行分类的结果

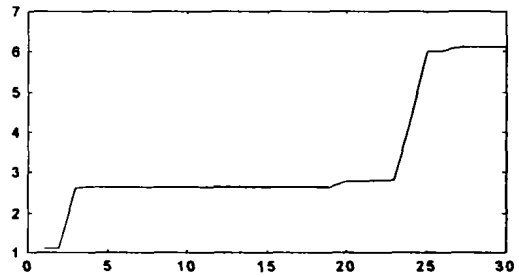

图6　实验2的进化过程中保留最佳个体时的最大适应度随迭代次数的变化情况

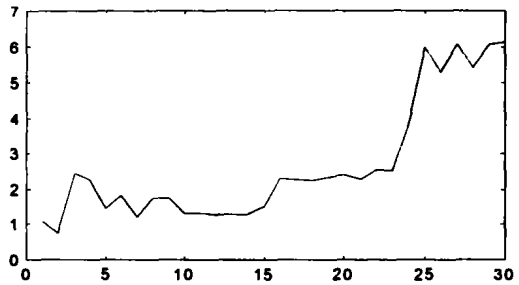

图7　实验2的进化过程中不保留最佳个体时的平均适应度随迭代次数的变化情况

在以上两个试验中，对原始的数据(如图2和图5中点的分布)是有所了解的，所以我们可以从中提取出一些“疫苗”，比如个体相邻基因位的大小之差等。由于这些疫苗提高了群体的适应度，减小了算法在无解区域内进行搜索的概率，从而提高了搜索效率。这使得算法既能有效地克服局部极值和对初始化取值敏感性问题，同时也减轻了在进化后期的波动现象又具有较快的收敛速度。另外，在实验中取 T_k 恒等于1。因为实验中的迭代次数很少，温度控制序列的作用并不明显，所以在实验中没有使用。

(下转第158页)

证毕。

结合定理5.1，我们给出下列说法。

定义5.1 对任一 $A\in R(\Phi)$，称 $P_-(A)$为事件 A 的概率下估计值，$P^-(A)$为 A 的概率上估计值，而称$[P_-(A),P^-(A)]$为 A 的粗糙概率区间估计。

定理5.2 $P_-(\Phi)=P^-(\Phi)=0$。

证明：$\because \Phi_-=\Phi=\Phi^-$，$\therefore P_-(\Phi)=P(\Phi_-)=P(\Phi)=0$，$P^-(\Phi)=P(\Phi^-)=P(\Phi)=0$。证毕。

定理5.3 若 $A_i\in R(F)(i=1,2,\cdots,n)$，$A_i\cap A_j=\Phi(i\neq j)$，则：

$$P_-(\mathop{Y}_{i=1}^{n}A_i)=\sum_{i=1}^{n}P_-(A_i),P^-(\mathop{Y}_{i=1}^{n}A_i)=\sum_{i=1}^{n}P^-(A_i)。$$

证明：由定理4.2.1之(3)，令 $A_{n+1}=A_{n+2}=\cdots=\Phi$，应用 $P_-(\Phi)=0$这一结果，得：

$$P_-(\mathop{Y}_{i=1}^{n}A_i)=P_-(\mathop{Y}_{i=1}^{\infty}A_i)=\sum_{i=1}^{\infty}P_-(A_i)=\sum_{i=1}^{n}P_-(A_i)。$$

同理可得 $P^-(\mathop{Y}_{i=1}^{n}A_i)=\sum_{i=1}^{n}P^-(A_i)$。证毕。

定理5.4 对任一 $A\in R(F)$，有 $P_-(\overline{A})=1-P_-(A)$，$P^-(\overline{A})=1-P^-(A)$。

证明：$\because A\cup\overline{A}=\Omega$，$A\cap\overline{A}=\Phi$，$\therefore$ 由定理5.3得 $P_-(A)+P_-(\overline{A})=1$，即 $P_-(\overline{A})=1-P_-(A)$。

同理可证 $P^-(\overline{A})=1-P^-(A)$。证毕。

定理5.5 对 $A、B\in R(F)$，若 $A\supseteq B$，则：

(1)$P_-(A-B)=P_-(A)-P_-(B)$，且 $P_-(A)\geqslant P_-(B)$；

(2)$P^-(A-B)=P^-(A)-P^-(B)$，且 $P^-(A)\geqslant P^-(B)$。

证明：我们只给出(1)的证明，对(2)可做类似讨论。

$\because A=(A-B)\cup B$，$(A-B)\cap B=\Phi$，$\therefore$ 由定理5.3得

$P_-(A)=P_-(A-B)+P_-(B)$，

即 $P_-(A-B)=P_-(A)-P_-(B)$。

又 $\because P_-(A-B)\geqslant 0$，

$\therefore P_-(A)\geqslant P_-(B)$。证毕。

小结 我们以粗糙随机信息系统 S 的论域 Ω 上的等价关系 R 为基础，逐步引进了粗糙样本空间 $R(\Omega)$和粗糙 σ-代数 $R(F)$，从而形成了粗糙可测空间$(\Omega,R(\Omega),R(F))$。为了在$(\Omega,R(\Omega),R(F))$上引进概率测度，我们构造了$(\Omega,R(\Omega),R(F))$的近似可测空间(Ω,F_R)。然后利用(Ω,F_R)上的概率测度 P 为$(\Omega,R(\Omega),R(F))$引进了下近似概率 P_-和上近似概率 P^-，这样形成了粗糙概率空间$(\Omega,R(\Omega),R(F),P_-,P^-)$，从而解决了粗糙随机信息系统 S 的可测性问题。最后我们还讨论了上、下近似概率的若干性质。有关粗糙随机信息系统的进一步研究，我们还在进行中。

参 考 文 献

1 易树鸿. 粗糙随机事件及其概率估计. 计算机科学，2002，29(9.专刊)：126～134

2 中山大学数学力学系. 概率论及数理统计. 人民教育出版社，1980

3 袁震东. 近代概率引论——测度、鞅和随机微分方程. 科学出版社，1991

4 Pawlak Z. Rough set. International Journal of Information and Computer Science，1982(11)：341～356

5 Pawlak Z. Rough set — theoretical aspects of reasoning about data. Dordrecht：Kluwer Academic Publishers，1991

6 刘清. Rough 集及 Rough 推理. 科学出版社，2001

(上接第152页)

表1 第一个实验两种算法性能的比较

实验方法	样本点数	迭代次数	实验次数	正确聚类次数	聚类正确率
K均值方法	500	20	1000	723	72.3%
混合聚类方法	500	20	1000	989	98.9%

表2 第二个实验两种算法性能的比较

实验方法	样本点数	迭代次数	实验次数	正确聚类次数	聚类正确率
K均值算法	500	30	1000	714	71.4%
混合聚类算法	500	30	1000	978	97.8%

结论 HCA是在传统的聚类算法中结合了免疫进化的机理，将全局寻优与局部搜索能力相集成，并引入了接种疫苗和免疫检测，从而提高了聚类算法的性能。理论分析和实验结果表明，本文提出的HCA算法明显地克服了K均值算法中的局部极值问题和对初始值敏感性的缺点，并且具有较快的收敛速度，同时也减轻了进化后期的波动现象。

参 考 文 献

1 Jain A K, Murty M N, Flynn P J. Data clustering: A survey. ACM Compute. Surv., 1999, 31:264～323

2 Jain A K, Dubes R C. Algorithms for clustering data. Englewood Cliffs, NJ: prentice Hall, 1988

3 高新波，谢维新. 模糊聚类理论发展及应用的研究进展. 科学通报，1999，44(21)：2241～2251

4 MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist, Prob, 1967, 1: 281～297

5 Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2: 283～304

6 许雷. 一种聚类新算法：模拟退火[J]. 模式识别与人工智能，1989，1：1～16

7 Babu G P, Murty M N. Clustering with evolution strategies [J]. Pattem Recognition, 1994, 2(27): 321～329

8 Krishma K, Murty M N. Genetic K-Means Algorithm. IEEE Trans on System, Man, and Cybernetics, Part B, 1999, 29(3): 433～439

9 Rudolph G. Convergence of Canonical Genetic Algorithm. IEEE Trans on Neural Networks, 1994, 5(1): 96～100

10 Mali U, Bandyopadhyay S. Genetic Algorithm-Based Clustering Technique. Patten Recognition, 2000, 33(9): 1455～1465

11 胡玉锁，陈宗海. 基于混合遗传算法的聚类分析. 模式识别与人工智能，2001，14(3)：352～335

12 王磊，潘进，焦李成. 免疫规划. 计算机学报，2000，23(8)：806～812