基于模糊 K－Modes 和免疫遗传算法的聚类分析

曹文婷，邹 海，段凤玲
（安徽大学 计算智能与信号处理教育部重点实验室，安徽 合肥 230039）

摘 要：为了克服传统的模糊 K－Modes 算法分类正确率低、收敛速度慢的缺点，文中将免疫遗传算法应用到聚类分析中，提出了一种基于模糊 K－Modes 和免疫遗传算法的聚类算法。通过引入免疫算子，不仅提高了收敛速度，而且避免了陷于局部极小，从而能较快地收敛到全局最优解。免疫算子包括抽取疫苗、接种疫苗和选择疫苗。实验结果证明，此算法具有较好的聚类效果，且稳定性强。

关键词：模糊聚类；K－Modes；免疫遗传算法；优化计算
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1673－629X(2009)02－0151－03

Cluster Analysis Based on Fuzzy K－Modes and Immune Genetic Algorithm

CAO Wen-ting, ZOU Hai, DUAN Feng-ling
(Ministry of Education Key Laboratory of Intelligence Computing and Signal Processing, Anhui University, Hefei 230039, China)

Abstract: To overcome the shortcomings of the low clustering correctness and slow convergent speed in the basic fuzzy K－Modes clustering algorithm, the immune genetic algorithm is introduced into the cluster analysis. Cluster analysis based on fuzzy K－Modes and immune genetic algorithm is proposed. Both the analytical and experimental studies indicate that this method is faster and more efficient to converge upon the optimal value. Immune operators including vaccine extraction, vaccination and vaccine selection. The experimental result shows that this improved algorithm is effective and steady contrast with basic fuzzy K－Modes.

Key words: fuzzy clustering; K－Modes; immune genetic algorithm; optimization computation

0 引 言

聚类分析在数据挖掘中广泛应用[1]，其目的是以事物间的相似性作为类属划分的准则，将一个数据集划分为若干聚类，属于无监督分类的范畴。模糊聚类算法 K－Mode[2]是对具有分类属性的数据进行聚类的一种有效的算法。它是在 Huang 提出的 K－Modes[3]算法的基础上，对模糊的一种改进。

1 模糊 K－Modes 聚类算法

模糊 K－Modes 算法是对模糊 K－Means 算法的扩展。用 Modes 代替了 Means，并且对每个聚类中的对象分派隶属度，克服了模糊 K－Means 算法只能处理数值型数据的缺点，模糊 K－Modes 算法通过交替更新聚类中心和划分矩阵，使得目标函数值达到最小[4,5]。

假定 $X=[x_1,x_2,\cdots,x_n]$ 是一组数据元组，其中 $x_i=[x_{i1},x_{i2},\cdots,x_{im}]$，表示第 i 个样本的 m 个属性值。聚类个数为 k。用以下代价函数最小作为聚类的准则：

$$E(\boldsymbol{U},\boldsymbol{Q})=\sum_{l=1}^{k}\sum_{i=1}^{n}u_{li}^{\alpha}d(\boldsymbol{Q}_l,X_i) \tag{1}$$

式中 $\boldsymbol{Q}_l=[q_{l1},q_{l2},\cdots,q_{lm}]$ 代表聚类 l 的原型模式，即 mode，$u_{li}\in[0,1]$ 表示模糊划分矩阵 $\boldsymbol{U}_{k\times n}$ 的一个元素，它表示对象 X_i 划分到聚类 l 总的隶属度；$\sum_{l=1}^{k}u_{li}=1$；其中 d 表示差异测度；$\alpha>1$ 是加权指数。对于具有分类属性的数据，定义差异测度为：

$$d(Q_l,X_i)=\sum_{j=1}^{m}\delta(q_{lj},x_{ij}) \tag{2}$$

其中，$\delta(q_{lj},x_{ij})=\begin{cases}0,(q_{lj}=x_{ij})\\1,(q_{lj}\neq x_{ij})\end{cases}$，$q_{lj}$ 和 x_{ij} 表示在第 j 个分类属性上的取值；m 表示属性的个数。划分矩阵

收稿日期：2008－06－25
基金项目：安徽省自然科学研究项目(2005kj001)
作者简介：曹文婷(1983－)，女，硕士研究生，研究方向为数据挖掘、人工智能；邹 海，硕士生导师，高级工程师，研究方向为数据挖掘、中间件技术。

的更新方法如下：

$$u_{li} = 1, \text{if } X_i = Q_l$$

$$u_{li} = 0, \text{if } X_i = Q_h, h \neq l$$

$$u_{li} = \frac{1}{\sum_{h=1}^{l}\left[\frac{d(Q_l, X_i)}{d(Q_h, X_i)}\right]^{1/(a-1)}}, \text{if } X_i = Q_l, X_i = Q_h \quad h = 1,2,\cdots,k \tag{3}$$

聚类中心的更新方法如下。

设 A_j 是属性 j 上的值集合：

$$A_j = \{a_j^{(1)}, a_j^{(2)}, \cdots, a_j^{(n_j)}\}$$

n_j 是第 j 个分类属性的可取值的个数；聚类中心 $Q_l = [q_{l1}, q_{l2}, \cdots, q_{lm}], l = 1,2,\cdots,k$。有如下定理：式(1)取得最小，当且仅当 $q_{lj} = a_j^{(r)} \in A_j, j = 1,2,\cdots, m$，其中 $a_j^{(r)}$ 满足

$$\sum_{i, x_{i,j} = a_j^{(r)}} u_{li}^a \geqslant \sum_{i, x_{i,j} = a_j^{(t)}} u_{li}^a, 1 \leqslant t \leqslant n_j$$

上式中 $a_j^{(r)}$ 表示在属性 j 的不同取值的数据集上的不同。

2 基于模糊 K－Modes 和免疫遗传算法的聚类算法

2.1 免疫遗传算法

免疫遗传算法是遗传算法和生物免疫思想的结合，既能够保留遗传算法全局搜索特性，又能避免未成熟的收敛[6]。它将求解问题的目标函数对应为入侵生命体的抗原，而问题的解对应为免疫系统产生的抗体，整个免疫机制由免疫算子(抽取疫苗，接种疫苗，选择疫苗)组成[7~9]。免疫遗传算法的流程图见图 1。

2.2 算法设计

2.2.1 编码

在模糊 K－Modes 聚类算法实现中，要寻求使代价函数最小的模糊划分 U 和原型模式 Q。对 $Q_l = [q_{l1}, q_{l2}, \cdots, q_{lm}]^T$ 矩阵编码，相比较对 U 矩阵编码，可以降低搜索效率。将原型中的 k 组特征连接起来，根据各自的取值范围，将其量化值编码成二进制的基因串。参数集依据每个原型 Q_l 取值。

2.2.2 适应度函数

聚类目标函数越小，则适应度函数越大，此时聚类效果越好。构造适应度函数如下：

$$f = \frac{1}{1 + E(U, Q)} \tag{4}$$

2.2.3 免疫算子

(1) 抽取疫苗：将前 k 代中的两个不同的最优个体 x_1 和 x_2 的基因位上的共同特点信息作为疫苗。

(2) 接种疫苗：设个体为 $X = [x_1, x_2, \cdots, x_n]$，从当前种群中按一定比例选取 N_e 个个体，按照先验知识，修改 X 的某些分量，使所得新个体以较大的概率适应最优解的取值范围，它须满足：若 X 已经是最佳个体，则 X 以概率 1 转移为 X；若 X 是最差个体，即 X 以概率为 0 转移为 X。

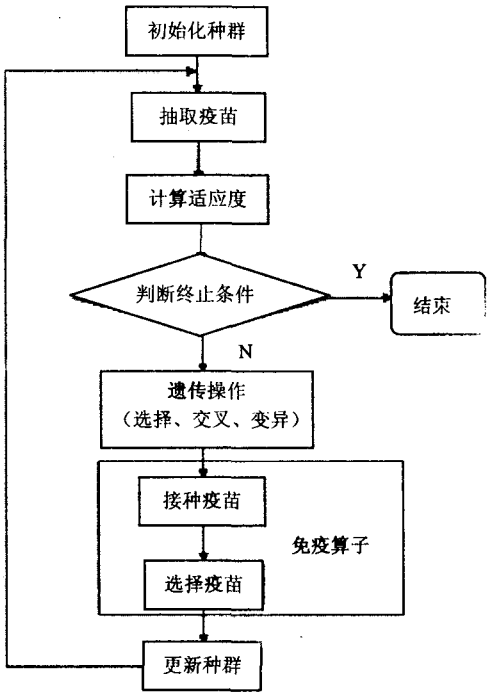

图 1 免疫遗传算法的流程图

(3) 选择疫苗：首先对接种了疫苗的个体进行检测。若其适应度优于父代，则该个体将进入下一代种群。若其适应度不如父代，说明它出现严重的退化，则该个体将被父代中相应的个体所替代。

文中，个体选择概率为：$p = \alpha p_f + (1 - \alpha) p_d$，其中 $0 < \alpha, p_f, p_d < 1$，p_f 为适应度概率，p_d 为浓度概率，α 为亲和系数。计算公式表示如下：

$$p_f = J / \sum_{i=1}^{n} J_i \tag{5}$$

其中，n 为种群规模，J 为个体适应度，$J = \exp(-f/T_k)$，T_k 是当前 k 代的温度，它是单调递减趋向于 0 的退火温度控制序列，其作用是控制个体的多样性变化速度。设定为 $T_k = T_0 \times 0.99^{k-1}$，$T_0$ 是初始温度。下面定义几个名词：

(1) 多样度：为了有效地保持种群进化个体的多样性，必须要度量和评价个体之间的差异。差异度量的精细程度，制约了免疫遗传算法个体多样性的水平。这里，个体之间的差异性由平均信息熵 $H(N)$ 表述。

设有 N 个，每个抗体长度为 M，采用的符号集大小为 S，则抗体基因座 j 的信息熵 $H_j(N)$ 可定义为：

$$H_j(N) = \sum_{i=1}^{s} - p_{ij} \log p_{ij} \tag{6}$$

p_{ij} 为第 i 个符号出现在基因座 j 上的概率,可表示为:

$$p_{ij} = \frac{\text{在基因座 } j \text{ 上出现第 } i \text{ 个符号的总个数}}{N} \tag{7}$$

由信息熵可得平均熵:

$$H(N) = \frac{1}{M}\sum_{j=1}^{M} H_j(N) \tag{8}$$

(2) 相似度:相似度 $A_{i,j}$ 是两个抗体 i 和 j 之间相似的程度,$A_{i,j}$ 越大,表示两个抗体越相似。

$$A_{i,j} = \frac{1}{1 + H(2)} \tag{9}$$

(3) 抗体浓度:抗体浓度 C_i,表示群体中相似抗体所占的比重,即:

$$C_i = \frac{\text{与抗体 } i \text{ 相似度大于 } \lambda \text{ 的抗体数}}{\text{抗体总数 } N} \tag{10}$$

其中 λ 为相似度常数,一般取为 $0.9 \leqslant \lambda \leqslant 1$。

再由计算所得的抗体浓度,找出浓度最大的个体 $1,2,\cdots,t$,则定义这 t 个个体的浓度概率 P_d 为 $\frac{1}{N}\left(1 - \frac{t}{N}\right)$,其他 $N - t$ 个个体的浓度概率 P_d 为 $\frac{1}{N}\left(1 + \frac{t^2}{N^2 - N_t}\right)$。

2.3 算法执行步骤

Step1:产生初始种群 A_K;

Step2:根据先验知识抽取疫苗,对当前种群 A_K 中的个体解码并进行评价,从当前种群中提取疫苗,并记录当前种群中的最优个体 x;

Step3:若当前种群中已包含最优解,或已大于指定的迭代次数,则算法结束,否则,继续;

Step4:对当前的第 k 代种群 A_K 进行交叉操作,得到新的种群 B_K;

Step5:对种群 B_K 进行变异操作,得到新的种群 C_K;

Step6:在种群 C_K 中选取 N_e 个个体接种疫苗,得到种群 D_K。计算 D_K 种群中个体的适应度、相似度、浓度,得出每个个体的适应度概率和浓度概率,进而求出每个个体的选择概率,采用赌轮法得到新一代种群 A_{k+1},若发现 A_{k+1} 中最优个体没有最优个体 x 优秀,则将 A_{k+1} 中的最差个体由 x 来替代,转 Step2。

3 仿真试验

为了验证本算法的性能,选用 uci 数据集的 IRIS 数据作为测试样本,来比较传统的模糊 K－Modes 聚类算法与基于免疫遗传算法的聚类算法的性能。IRIS 数据由 150 个四维向量组成,每一个样本的 4 个分量分别表示 IRIS 的 PetalLength,PetalWidth,SepalLength 和 SepalWidth。整个样本集包含 3 个 IRIS 种类 Setosa,Versicolor 和 Virginica,每类各有 50 个样本。

免疫遗传算法的参数设置如下:种群大小为 50,交叉概率为 0.35,变异概率为 0.005,最大运行次数为 100。初始温度 T_0 为 $2e + 5$,接受疫苗注射的个体数 N_e 为 10,相似度常数 λ 为 0.95。每次试验运行 5 次。仿真结果见表 1。

表 1　两种算法的正确率比较

正确率	模糊 K－Modes 算法	免疫遗传 模糊 K－Modes 算法
第 1 次	47.5%	90.8%
第 2 次	65.8%	92.3%
第 3 次	78.4%	93.1%
第 4 次	92.3%	94.3%
第 5 次	84.2%	94.3%

由表 1 可以看出,文中的算法比传统算法的正确率高,都在 90% 以上,并且十分稳定。在收敛速度和分类性能两方面都获得了很好的效果。

4 结束语

文中提出的基于模糊 K－Modes 和免疫遗传算法的聚类算法,克服了传统的基于模糊 K－Modes 聚类算法分类正确率低,收敛速度慢的缺点,引入免疫遗传算子,免疫算法中的个体多样性保持机制能避免 GA 的未成熟收敛现象。提高了全局和局部收敛速度,试验结果取得了较好的聚类效果。

参考文献:

[1] 姜　园,张朝阳,仇佩亮,等.用于数据挖掘的聚类算法[J].电子与信息学报,2005,27(4):655－662.

[2] Huang Zhexue, Ng M K. A Fuzzy K－Modes Algorithm for Clustering Categorical Data[J]. IEEE Transactions on Fuzzy Systems,1999,7(4):446－452.

[3] Huang Zhexue. Extensions to the K－means Algorithms for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery,1998,2:283－304.

[4] 赵　恒,杨万海.基于属性加权的模糊 K－Modes 聚类算法[J].2003,25(10):1299－1302.

[5] 赵　恒,杨万海.模糊聚类精确度分析 K－Modes[J].计算机工程,2003,29(12):27－28.

[6] 焦李成,杜海峰,刘　芳,等.免疫优化计算、学习与识别[M].北京:科学出版社,2006.

[7] 焦李成,杜海峰.人工免疫系统进展与展望[J].电子学报,2003,31(10):1540－1548.

[8] 王文卓,张　巧,吴春国,等.遗传算子对免疫算法性能影响的分析[J].小型微型计算机系统,2007,28(8):1448－1451.

[9] 葛　红,毛宗源.免疫算法的实现[J].计算机工程,2003,29(5):62－64.