基于改进免疫遗传算法的聚类分析研究与应用*)

许文杰[1] 刘希玉[2]

(山东师范大学信息科学与工程学院 济南 250014)[1] (山东师范大学管理与经济学院 济南 250014)[2]

摘 要 本文分析了聚类的思想,将免疫原理引入到遗传算法并应用于聚类分析过程中,提出了改进的免疫遗传算法(MIGA)。该算法借鉴了免疫算法中有关浓度的定义,并使用了GA算法中交叉和变异的思想。通过实验显示该方法优于基本的遗传算法。

关键词 聚类分析,免疫原理,遗传算法

Clustering Analysis Based on Modified Immune Genetic Algorithm and its Application

XU Wen-Jie[1] LIU Xi-Yu[2]

(College of Information Science and Engineering, Shandong Normal University, Jinan 250014)[1]

(College of Management, Shandong Normal University, Jinan 250014)[2]

Abstract This paper analyzes the idea of clustering, which introduces immune principle to the genetic algorithm and applies it in the clustering analysis process. The paper also proposes the modified immune genetic algorithm (MIGA). The algorithm takes the definition of concentration in the immune algorithm and the idea of crossover and mutation from GA into consideration. The results of experiments demonstrate that this method is better than the simple genetic algorithm.

Keywords Clustering analysis, Immune principle, Genetic algorithm

1 引言

聚类是数据挖掘中的一类重要技术,是分析数据并从中发现有用信息的一种有效手段。聚类分析(Clustering Analysis)是一种无监督的模式识别方式,它将数据对象分组成为多个类或簇,使得在同一个簇中的对象之间具有较高的相似度,而不同簇中的对象差别很大。聚类分析在数据挖掘、图像分割、目标检测、特征提取和信号压缩等方面都有着广泛的应用。聚类算法大体可以划分为以下几类:划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法。基于划分的传统聚类方法,如C-均值算法等,具有较强的局部搜索能力,并以其简单、快速的特点而被广泛应用。

遗传算法(GA)是一种模拟生物群体遗传和进化机理的启发式优化算法。它是由美国的Holland教授等人提出的[1],从一个初始群体开始寻优的、具有较强的并行搜索能力,但容易出现"早熟"和局部搜索能力不足等问题。直接应用标准遗传算法来解决聚类问题,算法的性能得不到保证,必须尽可能地应用特定问题的领域知识来设计有效的遗传算法。目前,已有很多人对标准的遗传算法做了很多的改进:设计一种混合遗传算法[2],以K均值算子代替交叉算子,可以达到全局最优点;还有采取聚类中心的浮点编码方式[3],设计了交叉、变异算子,从而提高了搜索效率等。但是,由于进化算法固有的缺点,在进化过程中不可避免地产生了退化的可能,导致了进化后期的波动现象并且迭代次数过长和聚类准确率不高等问题[4]。

近年来,受生物免疫理论启示的人工免疫系统方法的研究成为热点。人工免疫系统在识别及优化问题上所具有的启发式搜索能力,使得很多应用领域都引入人工免疫的思想,将其与已有的智能方法结合,提高人工智能计算的整体性能。

本文将免疫算法引入遗传算法聚类中,通过对聚类问题的实际情况设计遗传编码,通过选择算子的合理设置,形成一种混合算法,既保证了搜索的全局性及搜索中学习的能力,又兼顾了收敛速度,有效克服了聚类中的固有的缺点,收敛到全局最优解。

2 聚类思想

聚类分析在许多领域特别是模式识别和图像处理中得到了广泛的应用。在各种聚类算法中,C-均值聚类算法[5]的应用最为广泛。C-均值聚类算法的基本原理如下:

设 $X=\{X_1,X_2,\cdots,X_n\}$ 为有限数据集,其中的 $X_i\in R^s$。各数据和聚类中心的关系可用隶属矩阵 M_{cn}(cn 为聚类中心的个数)来表示。M_{cn} 的定义如下:

$$M_{cn}=\{U\in R^{c\times n}\mid \sum_{i=1}^{c}U_{ik}=1,0<\sum_{k=1}^{n}U_{ik}<n\} \quad (1)$$

$$U_{ik}\in\{0,1\},1\leqslant i\leqslant c,1\leqslant k\leqslant n$$

式(1)中的 $U_{ik}=1$(或 $U_{ik}=0$)表示 X_k 属于(或不属于)第 i 个聚类中心 V_i。聚类的目标函数为:

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(U_{ik})D_{ik}^2(V_i,X_k) \quad (2)$$

式(2)中的 $D_{ik}^2(V_i,X_k)$ 为数据 X_k 到聚类中心 V_i 的欧拉距离。既然每一个待聚类的数据都被分配给了欧拉距离最

*)基金项目:国家自然科学基金项目、山东省自然科学基金重大项目(No. Z2004G02)、山东省中青年科学家奖励基金项目(No. 03BS003)、山东省教育厅科学技术项目(No. J05G01)、"泰山学者"建设工程专项经费资助。**许文杰** 硕士研究生,主要研究方向:进货计算、数据挖掘;**刘希玉** 教授,博士生导师,博士,主要研究方向:数据挖掘、人工神经网络、进化计算、实体造型技术等。

近的聚类中心，式（2）可进一步简化为

$$R(V)=\sum_{k=1}^{n}\min\{D_{1k}^2,D_{2k}^2,\cdots,D_{ck}^2\} \tag{3}$$

3 改进的免疫遗传聚类算法(MIGA)

在标准的遗传算法[6]聚类过程中，基本不利用外部信息，仅以适应度函数为依据，利用种群中每个个体的适应度来进行搜索，直接影响到遗传算法的收敛速度以及能否搜索到最优解。但是，如果适应度函数选择不当，会出现一系列的问题。比如，在遗传算法初期，通常会产生一些超常的个体，若按比例选择法，这些异常个体因竞争能力太突出而控制了选择的过程；或者在遗传算法后期，即算法接近收敛时，由于种群个体适应度差异较小时，继续优化的潜能降低，可能获得某个局部最优解。

3.1 抗原个体浓度的定义

根据免疫算法的原理，将待求解的问题视为抗原。提取数据样本特征的关键体现在对抗原的使用上，根据数据样本的某一区域的密度或其他相关特征来构造抗原抗免疫性。假设将抗原的个体浓度定义为 N_i，是样本 x_i 的半径为 R 的球形区域内 $H(x_i,R)=\{x_i^{'}:d(x_i,x_i^{'})\leqslant R\}$ 的相同样本 $x_i^{'}$ 数量。

$$P_d=\left(\frac{N_{\max}-N_i+a}{N_{\max}}\right)^{\beta} \tag{4}$$

式中 $N_{\max}$ 是所有样本中的最大浓度，$N_{\max}=\max\{N_i,i=1,\cdots,n\}$，$\beta$ 是抑制距离的指数控制参数，a 是距离为零的初始值。

3.2 改进的免疫遗传算聚类

本文将免疫原理进入到遗传算法聚类中算子上来，并其称之为改进免疫遗传算法（MIGA）。在 MIGA 中，为了降低适应度值过大或过小可能造成的不良影响，加快收敛速度，对各参数进行了合理的设置。与 GA 一样，MIGA 是一种基于自然选择和遗传变异等机制的全局性概率搜索算法，其交叉、变异算子是从 GA 借用过来的。通常情况下，免疫算法与 GA 虽然交叉和变异的概率可能各不相同，但是却采用相同的交叉和变异算子，选择方式也是基于选择概率进行个体选择。虽然，在表面上选择方式似乎没有什么差别，实际上隐藏着两种选择机制的本质差别，即 MIGA 的选择概率包含了解的适应度和浓度的信息，模拟了自然免疫系统的浓度调节机制，而 GA 的选择概率只包含了解的适应度的信息[7]。

本文将免疫的思想引入到遗传算法聚类中，使用基于免疫原理的选择操作，提高了全局收敛性和计算效率，其算法可描述为：

Step1：确定编码方式。

遗传算法中的进化过程是建立在编码机制基础上的，编码对于算法的性能，如搜索能力等影响很大。常用的编码方式有浮点数编码和二进制编码。相比之下，根据测试数据集的特征，我们采用符号编码，符合有意义积木块编码原则，便于在算法中利用所求解问题的专门知识。且设编码长度 $L=12$。

Step2：生成初始种群。

随机地产生 Psize 个初始个体，组成初始种群。个体表示聚类中心，随即产生的每个初始个体应处于所要聚类的向量所定义的空间之内。Psize 太小的话，易失去多样性；Psize 太大的话，时间消耗会很大。这里我们初始化种群数 Psize=80。

Step3：确定适应度函数。

适应度函数被用来描述个体的好坏程度。为了避免出现算法提前终止收敛过早和过慢的情况，首先对原始的适应值按递增排序后，根据得到的序号 Sorti 和调节因子 $\eta(0<\eta<1)$ 来计算适应度的值，这样可对过大的适应值有个缓冲的作用。

$$Fs_i=\eta(1-\eta)^{\sqrt{sort_i}} \tag{5}$$

令 $\eta=0.6$，然后根据 Fs_i 为选择依据，得

$$P_f=\frac{\text{群体中个体的适应度}}{\text{群体中总的适度之和}} \tag{6}$$

Step4：基于免疫原理的选择操作。

定义个体的选择概率 P，P 由适应度概率 P_f 和浓度概率 P_d 共同决定。这样不但保持了适应值对选择的促进作用，而且通过调节浓度，起抑制作用，保证了群体中个体的多样性。

$$P=\lambda P_f-(1-\lambda)P_d,\lambda\in(0,1) \tag{7}$$

产生随机数 P'。当 $P'<P$ 时，对应的个体被复制到下一代，直到生成 80 个中间群体。

Step5：交叉算子。

交叉操作通过交换两个父个体的一部分来产生新的子个体。交叉操作按照一定的交叉概率 P_c 进行。本文采用单点交叉，首先随机生成交叉点位置 $\text{pos}(1<\text{pos}<L)$，然后不重复地从两个父个体对交叉位后基因进行交叉运算，从而生成两个新的子代个体。

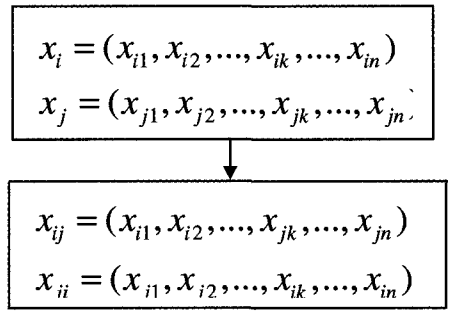

图 1 交叉操作示意图

Step5：变异操作。

交叉操作产生的子代个体除了继承父代个体的信息外，还会按一定的概率发生变异，这体现了生物遗传的多样性。本文使用固定的变异概率 $p_m(0.01<p_m<0.1)$ 对子代个体中的基因位进行变异操作，变异位置随机产生。

3.2 改进的免疫遗传聚类流程图

下面给出了用于聚类分析的免疫遗传算法的处理流程，如图 2。

4 实验

本实验使用 VS.NET 2003，在内存为 512MHz，CPU 为 Celeron 1.7GHz 的计算机上运行。针对数字图像，使用经典遗传算法（GA）和免疫遗传算法 MIGA 进行实验。

本次实验所采用算子的交叉概率为 0.6，变异概率为 0.03，由于实验数据量大小的关系，我们可以自行设定最大的迭代次数。分别使用 30 个二维平面上的图像信息进行识别，将图像序号和聚类后的所属的类号分别放置于图像的上、下部，使用图像的序号信息进行符号编码。为了比较，将每个算法

（下转第 210 页）

递模型，进一步分析了这种动态粗传递的特性，得到了动态粗传递信息不变定理、信息损失定理，及提高粗传递精确性的方法。最后给出了该模型的应用实例。通过研究得出：可以用双向 S-粗集来研究这种动态粗传递问题，动态粗传递是双向 S-粗集的一种新的应用。对于其它信息传递情况下的动态粗传递模型，限于篇幅，将另文讨论。

参 考 文 献

1 耿志强，朱群雄，李芳. 知识粗糙性的粒度原理及其约简[J]. 系统工程与电子技术，2004，26(8)：1112～1116

2 Mousavi A，Jabedar-Maralani P. Double-faced rough sets and rough communication [J]. Information Sciences，2002(148)：41～53

3 Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences，1982(11)：341～356

4 Pawlak Z. Rough sets：Theoretical Aspects of Reasoning about Data [M]. Boston，NA：Kluwer Academic Publishers，1991

5 王瑜，胡运发，张凯. 基于粗集理论的知识含量度量研究[J]. 计算机研究与发展，2004，41(9)：1500～1506

6 SHI Kaiquan. Two-direction S-rough sets [J]. International Journal of Fuzzy Mathematics，2005，13(2)：335～349

7 SHI Kaiquan. S-Rough sets and knowledge separation [J]. Journal of System Engineering and Electronics，2005(2)：32～37

8 SHI Kaiquan，CUI Yuquan. F—decomposition and $\overline{F}$— reduction of S-rough sets [J]. An International Journal of Advances in Systems Sciences and Applications，2004(4)：487～499

9 LU Changjing，SHI Kaiquan. Knowledge filter and its dependent reasoning discovery [J]. International Journal of Fuzzy Mathematics，2005，13(3)：613～626

10 Slizak D，Ziarko W. The investigation of the Bayesion rough set model [J]. International Journal of Approximate Reasoning，2005 (40)：81～91

(上接第 205 页)

运行 5 次，设定聚类的数目为 10，最大迭代次数为 80。

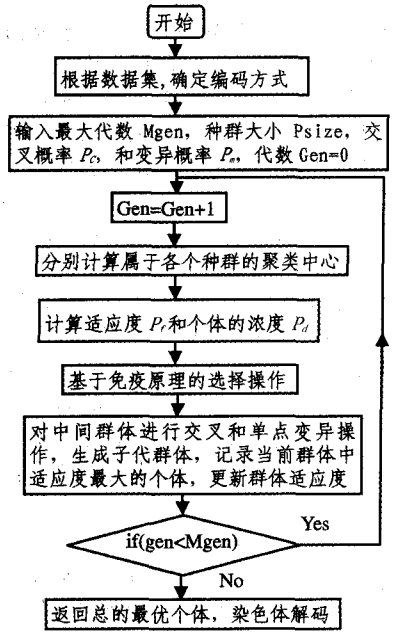

图 2 流程图

表 1 数据处理结果

运行次数	GA	MIGA
1	23	21
2	22	21
3	24	21
4	23	21
5	22	21

实验结果显示，本文提出的免疫遗传算法不论是在收敛速度上，还是聚类结果的稳定性上都优于传统的遗传算法。如果用该算法作用于较大的数据量，优越性更加明显。而将遗传算法与免疫算法相结合形成一种混合算法后，在保证准确性和提高执行效率的前提下，能更快地找到最有用的结果。

结论 本文将免疫原理引入遗传算法聚类中，通过对选择算子操作的优化，保证较高的准确率，又不至于陷入局部最优，克服了过早收敛的问题。MIGA 算法由于是降低了适应度的影响，因此会增加计算时间。如何通过对选择操作的参速更好的设置，以及提升计算速度，是进一步研究的方向。

参 考 文 献

1 Holland J H. Adaptation in Nature and Artificial System [M]. Cambridge：MIT Press，1992

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

3 Yang Kiduk，Jacob E. A hybrid approach to generating and utilizing faceted vocabulary for knowledge discovery on the web [C]. In：Proc of the 2004 Joint ACM/IEEE Conference on Digital Libraries. New York：ACM Press，2004. 395～395

4 高坚. 基于 C-均值和免疫遗传算法的聚类分析[J]. 计算机工程，2003，29(12)：65～66

5 李碧，雍正正. 一种改进的基于遗传算法的聚类分析方法[J]. 电路与系统学报，2002，7(3)：96～99

6 王小平，曹立明. 遗传算法[M]. 西安：西安交通大学出版社，2002. 1

7 郑日荣，毛宗源，等. 改进的人工免疫算法的分析研究[J]. 计算机工程与应用，2003，34：35～37

8 Han Jiawei，Kamber M. Data Mining：concepts and techniques [M]. Jim Gray：Series Editor Morgan Kaufmann Publishers，2000

9 Leclerc B. Minimum spanning trees for tree metrics：Abridgements and adjustments [J]. Journal of Classification，1995，12：207～241

10 Selim S Z，Alsultan K. A Simulated Annealing Algorithm for the Clustering Problem [J]. Pattern Recognition，1991，24 (10)：1003～1008

11 Maulik U，Bandy0padhyay S. Genetic Algorithm-based Clustering Technique [J]. Patten Recognition，2000，33 (9)：1455～1465

12 行小帅，潘进，焦李成. 基于免疫规划的 K-means 聚类算法[J]. 计算机学报，2003，30(4)：150～152

13 傅景广，许刚，王裕国. 基于遗传算法的聚类分析[J]. 计算机工程，2004，30(4)：122～124

14 Fayyad U，Reina C，Bradley P S. Initialization of Iterative Refinement Clustering Algorithms [R]：[Microsoft Research Technical Report. MSR-TR- 98-38]. June 1998. 1～5

15 王实，高文. 数据挖掘中的聚类算法[J]. 计算机科学，2000，27 (4)：42～45

16 杨淑莹. 图像模式识别[M]. 北京：清华大学出版社，北京交通大学出版社，2005

17 陈莉，焦李成. 基于自适应聚类的数据预处理算法 I [J]. 计算机应用与软件，2005，22(3)：28～29

18 李春华，朱燕飞，毛宗源. 一种新型的自适应人工免疫算法[J]. 计算机工程与应用，2004(22)：84～87

19 李洁，高新波，焦李成. 基于克隆算法的网络结构聚类新算法[J]. 电子学报，2004，32(7)：1195～1199