第33卷 第23期 计 算 机 工 程 2007年12月
Vol.33 No.23 Computer Engineering December 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)23—0194—03 文献标识码：A 中图分类号：TP301.6

基于动态粒度的并行人工免疫聚类算法

郝晓丽，谢克明
(太原理工大学计算机学院，太原 030024)

摘 要：从粒度的角度讨论了聚类结果和先验知识的协调度问题，提出了一种基于动态粒度的并行免疫聚类算法。鉴于并行人工免疫系统模型具有并行、随机搜索、反复进化和模式多样性等特点，将其与动态粒度模型相结合，在粒度变化过程中，通过对粒度粗化和细化的调整，选择合适粒度，保证了算法的聚类效率和聚类质量。实验证明，该算法在处理多样本、多属性、多类别问题时，是一种有效的方法。
关键词：动态粒度；人工免疫；聚类；协调度

Parallel Artificial Immune Clustering Algorithm Based on Dynamic Granulation

HAO Xiao-li, XIE Ke-ming
(School of Computer, Taiyuan Technology University, Taiyuan 030024)

【Abstract】 This paper discusses measure of harmony between clustering and transcendent knowledge, and a new clustering algorithm is proposed, which is parallel artificial immune clustering algorithm based on dynamic granulation. Artificial immune system model has the characteristics, such as parallel, random search, and maintains diversity. It is unified to dynamic granulation model. In the process of granulation changing, appropriate granulation can be made by adjusting, which can ensure clustering efficiency and quality of the new algorithm. Test results show that the algorithm is reasonable when the problem is handled, which has many samples, many attributions and many classifications.
【Key words】 dynamic granulation; artificial immune; clustering; harmony

在机器学习中，聚类是一个重要的研究课题[1]。聚类是将物理或抽象对象的集合分类分组成为由类似的对象组成的多个类的过程，这些对象与同一个簇中的对象彼此相似，与其他簇中对象相异。关于聚类分析有很多成功的方法，如划分聚类法、密度聚类法、层次聚类法、网格聚类法、模型聚类法、模糊聚类法等。

人工免疫算法是从体细胞理论和网络理论得到启发，模拟自然免疫系统功能的一种智能方法。本文提出的并行人工免疫算法则采用了并行机制，建立多个子种群，对其分别进行趋同操作，并且当每一代进化结束后，种群之间通过异化操作来交换优秀抗体所携带的优良信息，从而打破种群内的平衡状态。通过这样反复的进化操作，保持了模式的多样性，增强了个体的优势，最终获得最优个体。

在此基础上，本文将其与动态粒度模型相结合，从多种群并行搜索和信息粒度的角度来构造一种新的聚类算法，该算法不仅消除原有特征空间中聚类结果和先验知识之间的不协调性，而且可以准确揭示类中元素的“抱团”性质。应用此算法对相关数据进行聚类分析，并和其他算法进行比较，验证了基于动态粒度的并行人工免疫聚类算法的正确性和实效性。

1 基于动态粒度的聚类算法分析

1.1 统一粒度意义下的聚类协调性

聚类操作实质上是在样本点之间定义一种等价关系。属于同一类的任意两个样本点被看作是等价的，可以认为它们具有相近的性质，在当前的阈值尺度下是没有区别的，一个等价关系就定义了样本点集合的一个划分，它把样本点划分成一些子集，一个子集就对应着聚类形成的一类。由一簇粗细不等的等价关系便形成不同的聚类结果，这些等价关系形成一个偏序格结构。

从聚类的角度来看，先验知识中规定的某一类中的样本点，依照选定的特征空间和相似性测度，也应当聚为一类。然而在大多数情况下，都不会达到这种理想境界。常常是领域专家认为应当归为一类的点，往往在特征空间中距离特别远，而那些被认为分属不同类的点，却距离非常近，也就是说，聚类结果和先验知识之间存在某种不协调性。

聚类谱系图实际上是定义了一个粒度逐渐变细的等价关系序列，选择一个阈值实际上就是选定一个等价关系R，进而可以得到商集[2]，也就是知识体系U/R。如果由先验知识规定的类X能够使用现有的知识体系精确表达，就表示聚类结果和先验知识是协调的。而如果上近似和下近似不相同的话，就说明聚类结果和先验知识是不协调的。

1.2 基于动态粒度的聚类算法

通常在实际情况下，聚类的目的是寻找一个知识体系，这种知识体系一方面能够精确地表达先验知识规定的类X，同时又能表示构成类X诸元素的规律。采用一种动态粒度的知识体系，即对同一问题采用多个粒度来研究。在粒度变化过程中，有粒度粗化和粒度细化两种情况。当对问题刻画和描述过于粗糙，使得问题的某些性质被模糊，则需要采用粒

基金项目：国家自然科学基金资助项目(60374029)；国家高等学校博士学科点专项科研基金资助项目(20060112005)；山西省留学人员基金资助项目(2004-18)
作者简介：郝晓丽(1973－)，女，博士研究生，主研方向：人工智能，进化计算，故障检测；谢克明，教授、博士生导师
收稿日期：2006-12-13 **E-mail：**haoxiaoli2006@sina.com.cn

度细化操作；反之，由于刻画和描述得过于精细，每个样本自成一类，不能挖掘样本中的知识，丢失了一些对象的抱团性质，则采用粒度粗化操作。为了针对具体问题，寻求合适的粒度，采用如下两个等价划分的方法进行粒度合成。

定义 1 设 R_1 和 R_2 是论域 X 上的两个等价关系，若满足：

(1) $R_1 < R$ 且 $R_2 < R$

(2)存在 R'，使得 $R_1 < R'$, $R_2 < R'$，且 $R < R'$

则称 R 为 R_1 和 R_2 之积，记作 $R = R_1 \otimes R_2$。

定义 2 设 R_1 和 R_2 是论域 X 上的两个等价关系，若满足：

(1) $R < R_1$ 且 $R < R_2$

(2)存在 R'，使得 $R' < R_1$, $R' < R_2$，且 $R' < R$

则称 R 为 R_1 和 R_2 之和，记作 $R = R_1 \oplus R_2$。

对具体问题聚类时，首先预置一个等价关系 R_0 划分问题对应的集合(相应的粒度为 Δ_0)，得出初步结论 A_0。如满足需要，则聚类粒度合适。否则，分两种情况考虑：

(1)若与 Δ_0 比较粒度偏粗，则取一偏细等价关系 R_0'，令 $R_1 = R_0 \otimes R_0'$。再在 R_1 上进行分析，得出结论 A_1 和聚类粒度 Δ_1。如果 A_1 还粗，可以重复进行上述过程，将粒度继续细化。

(2)与 Δ_0 比较粒度偏细，则取一偏细等价关系 R_0'，令 $R_1 = R_0 \oplus R_0'$。再在 R_1 上进行分析，得出结论 A_1 和聚类粒度 Δ_1。如果 A_1 还细，可以重复进行上述过程，将粒度继续粗化。

由此，可以得到一个等价关系族 $P = \{R_n, R_{n-1}, \cdots, R_1\}$，其满足偏序关系 $R_n \le R_{n-1} \le \cdots \le R_1$，进而可以得到相应的商集序列，即知识体系族 $U/R_i (i = 1,2,\cdots,k)$。先验知识规定的类 X 可先由知识体系 U/R_1 来表示，令其边界中信息粒度集作为新的研究对象，再用知识体系 U/R_2 来表示，依次类推，直至达到目前知识体系所能表达的最大精细度为止。

定义 3 给定知识库 $K = (U, R)$，对 $\forall X \subseteq U$，给定一个具有偏序关系的等价关系族 $P = \{R_n, R_{n-1}, \cdots, R_1\}$，且满足 $R_n \le R_{n-1} \le \cdots \le R_1$，由此形成的聚类结果和 X 协调度为

$$H(P, X) = \frac{|P(X)|}{|X|} \tag{1}$$

其中，$|\ |$表示集合的基数。显然，协调度 $H(P, X) \in [0,1]$。当 $H(P, X) = 0$ 时，表示聚类结果和先验知识最不协调；当 $H(P, X) = 1$ 时，表示聚类结果和先验知识最协调，即现有的知识完全可以精确地描述先验知识。

1.3 算法的缺陷

基于动态粒度的聚类算法不仅消除了聚类结果和先验知识之间的主客观不协调性，而且有效提高了聚类的正确率。然而，该算法依然存在一些缺陷：(1)在很大程度上依赖于初始分类的选择，若初始分类严重地偏离全局最优分类时，用此算法可能陷入局部极小值，得到局部最优解。(2)特征选取的正确与否和维数的高低直接影响聚类结果的正确性。(3)不适合大规模数据挖掘。

2 基于动态粒度的并行免疫聚类算法

本文将并行免疫模型引入动态粒度的聚类算法，提出了基于动态粒度的并行免疫聚类算法。该算法不仅消除了原有聚类结果的不确定性，优化了算法性能，尤其在大规模数据挖掘中，更体现出新算法的并行效率。

2.1 基于并行免疫的聚类算法

将人工免疫算法应用到聚类分析中，可以解决传统聚类中运行效率低和初始化敏感度高等缺陷。运用标准的人工免疫算法[3]进行聚类时，首先采用聚类中心的编码方式将基因链构造为抗体，并在解空间中随机散布产生初始抗体群，再通过抗体对抗原亲和度的计算，分别对每个抗体进行评价，保留亲和度较高的抗体，而祛除亲和度较低的抗体，如此反复进行评价、选择和替代等操作，从而得到最优抗体。该算法虽然理论上可以搜索到全局最优，但依然存在两个严重的缺陷，即容易陷入局部最优的平衡态，以及进化后期搜索停滞不前，使得算法最终搜索的结果往往不是全局最优解，而是局部最优解。因此，本文提出了并行免疫模型，并将其引入了聚类算法。在此详细描述在新算法中应用到的算子操作。算法描述如下：

(1)编码。编码机制的选择对人工免疫算法的性能影响较大。本文依然采用聚类中心的编码方式。

(2)初始化种群的形成。当聚类个数 c 给定时，可以随机选取聚类中心，以及随机生成模糊矩阵。这样根据编码可以得到初始种群 $A(k)$，其规模为 N，并将其划分为若干个子种群，分别进行选择和趋同操作。

(3)抗体评价。为提高该算法的聚类效果，本文将簇内距和簇间距作为因子构造了适应度函数，抗体的适应值越大，则选择的概率越大，保证了种群中保留适应值大的抗体，加速算法的收敛。

将 n 个样本 $x_j (j = 1,2,\cdots,n)$ 划分为 c 个组 $G_i (i = 1,2,\cdots,c)$。当前种群中的每一个抗体都对应着 c 个中心，选择欧氏距离为组 j 中 x_k 与相应聚类中心 z_j 间的相似性指标时，价值函数可定义为

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} \left\| x_j - z_i \right\|^2 \tag{2}$$

其中，u_{ij} 是 x_j 属于组 G_i 的程度，取值为 0 或 1。

对于这 c 个中心，可将集合分为 c 簇。对于第 i 簇 G_i，其簇内的距离定义为

$$S_i = \frac{1}{|G_i|} \sum_{x \in c_i} \left| x - \overline{c_i} \right| \tag{3}$$

式中，$\overline{c_i}$ 是簇 G_i 的均值。定义簇 G_i 和簇 G_j 之间的簇间距离为 $d_{ij} = \left\| \overline{c_i} - \overline{c_j} \right\|$，并定义 $R_i = \max\{(S_i + S_j)\}/d_{ij}\}$，则聚类指标可定义为

$$D_B = \frac{1}{C} \sum_{i=1}^{C} R_i \tag{4}$$

抗体的适应度定义为

$$f(c) = \frac{1}{D_B} \tag{5}$$

(4)抗体的选择操作。在算法进化过程中，当有些抗体的规模达到一定程度后，而又不是最优解，就要对其进行限制，以防止过早收敛。本算法采用抗体浓度来抑制规模比较大又不是最优解的抗体。

抗体的浓度用于表示某个抗体以及与其很相似的抗体群的规模。抗体浓度定义为

$$Density(c_i) = \frac{1}{\rho(c_i)} = \frac{1}{\sum_{j=1}^{N} \left| f(c_i) - f(c_j) \right|} \tag{6}$$

则个体的选择概率为

$$P_s(c_i) = \frac{\rho(c_i)}{\sum_{i=1}^{n} \rho(c_i)} = \frac{\sum_{j=1}^{N} \left| f(c_i) - f(c_j) \right|}{\sum_{i=1}^{N} \sum_{j=1}^{N} \left| f(c_i) - f(c_j) \right|} \tag{7}$$

计算种群 $A(k)$ 中每个抗体的浓度，基于抗体浓度的概率式(6)选择其中的 m 个选择概率最高的抗体作为免疫记忆抗体保留，得到抗体群 $B(k)$。由此可以看出，抗体浓度越大，选择的概率越小，保证了进化种群中抗体的多样性，避免“早熟”。这使得含有有效进化基因的低适应度抗体也获得繁殖的机会。

(5)趋同操作。 $B(k)$ 中的抗体在趋同半径 R_Q 局部范围内进行趋同操作，产生的新抗体组成群体 $C(k)$。

(6)异化操作。单个子种群经过一定时间的进化容易陷入平衡状态，在该算法中，为打破这种状态，每进化一代，都需要选取每个子种群中通过趋同算法得到的最优抗体，与另一子种群的最优抗体进行竞争。通过异化算子，将交换两个子种群的最优抗体所携带的最优信息，以打破子种群内部的平衡态，由此产生种群 $D(k)$。

(7)抗体的传优操作。算法要求每经过 k 代，就执行传优操作，所以判断当前的代数是否为 k，如果不够，则不执行传优操作，让代数加 1；否则，通过传优算子执行传优操作。

(8)群体更新和终止条件判断。种群更新按照抗体的适应度函数 $f(c)=\dfrac{1}{D_B}$ 进行选择。若满足条件 $\left|J(t+1)-J(t)\right|<\varepsilon$，则算法终止。

2.2 基于动态粒度的免疫聚类算法

利用并行人工免疫聚类算法中的趋同、异化和传优算子，与动态粒度模型相结合，在粒度变化过程中，通过对粒度粗化和细化的调整，选择合适粒度，从并行优化和信息粒度的角度来构造一种新的聚类算法。

算法描述如下：

步骤 1 首先根据先验知识对样本集的所有样本点进行分类。

步骤 2 初始化聚类模型中参数：ε, N，R_Q，*population*。*population* 表示每次进行聚类操作的类重心数据点，随着聚类过程作动态调整。

步骤 3 按式(5)~式(7)进行基于浓度的抗体选择操作，并通过趋同、异化和传优算子，在当前种群中搜索到最优抗体。

步骤 4 根据编码和解码原理，可以将经过并行免疫优化算法寻优得到的最优抗体进行解码，得到新的聚类中心。

步骤 5 根据解码得到的新的聚类中心，调整 *population* 集合。若 *population* 集合中数据点个数等于 1，也就是类数为 1，则转入步骤 6；否则转入步骤 3。

步骤 6 聚类操作结束后，得到聚类谱系图和一系列由大到小排列的阈值，选取一组阈值相当于选定一个等价关系族。为了针对具体问题，采用定义 2 和定义 3 两种等价划分的方法进行粒度合成，寻求合适粒度，可以得到一个等价关系族 $P=\{R_n,R_{n-1},\cdots,R_1\}$，且满足偏序关系 $R_n\le R_{n-1}\le\cdots\le R_1$。

步骤 7 $i=1$。

步骤 8 计算此时的聚类结果和 X 的协调度

$$H_i(P,X)=\frac{\left|P_i(X)\right|}{\left|X\right|}$$

步骤 9 若 $H_i(P,X)=1$，那么输出聚类结果；否则 $i=i+1$，转至步骤 6。

3 实验和结果分析

本文采用标准免疫聚类算法(AI-artificial immune)，基于粒度原理的聚类算法(DG-dynamic granulation)，基于动态粒度的并行免疫聚类算法(DGAI)分别对二组数据进行试验。

(1)数据 1

选择Fisher的IRIS植物样本数据作为测试样本集[4]。它由分别属于 3 类不同植物的 150 个样本点组成，每个样本点均为 4 维模式向量，代表植物的 4 种特征数据。

对 IRIS 数据的聚类实验中，对各聚类算法的价值函数 J 最小值和聚类情况进行了比较，见表 1。

表 1 算法聚类结果比较

算法	J	错误分类数	分类正确率/(%)
AI	6 259.21	16	89.33
DG	6 147.53	12	92.00
DGAI	6 013.04	10	93.33

由表 1 可见，由于动态粒度原理和并行免疫算法的结合在聚类中的应用，使得 DGAI 聚类算法价值函数 J 比 AI 聚类算法和 DG 聚类算法的价值函数 J 要小。同时，DGAI 聚类算法的分类正确率也最高。

(2)数据 2

文献[5]中所用数据是某铁矿区阳起石的 11 个指标的 19 个样本。用标准免疫聚类算法进行多次聚类后，使得 $J=160.24$，但使用基于动态粒度的并行免疫聚类算法可以很快达到上述聚类和 J 值。如果单纯应用动态粒度聚类算法，则速度慢、计算量大，将并行免疫模型引入到动态粒度聚类算法中，可以很快达到全局最优分类，当样本数目多且类别数目较大时，上述的基于动态粒度的并行免疫聚类算法将比其他算法有更快的速度和更准确的结果。

4 结束语

本文将并行人工免疫算法与基于动态粒度的聚类算法相结合，为基于粒度的模拟进化聚类算法提供了新的思路。在算法中，对动态粒度进行了形式描述，给出了聚类粒度的调整方法，能够更快更好地选择合适粒度。同时，由于并行人工免疫算法的并行性和随机搜索性等特点，将其与动态粒度相结合，构造了新的聚类算法。实验证明，对大规模和完全随机分布的数据聚类问题，其优越性非常明显。

参考文献

1 Jain A K, Dubes R C. Algorithms for Clustering[M]. Englewood Cliffs: NJ Prentice Hall, 1988.

2 张燕平，张 铃，吴 涛. 不同粒度世界的描述法——商空间法[J]. 计算机学报, 2004, 27(3): 328.

3 葛 红，毛宗源. 免疫算法的改进[J]. 计算机工程与应用, 2002, 14(2): 47.

4 高 坚. 基于并行多种群自适应蚁群算法的聚类分析[J]. 计算机工程与应用, 2003, 39(25): 78.

5 张 维，潘福铮. 一种基于遗传算法的模糊聚类[J]. 湖北大学学报, 2002, 24(2): 101.