第35卷 第23期 计 算 机 工 程 2009年12月
Vol.35 No.23 Computer Engineering December 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)23—0181—03 文献标识码：A 中图分类号：TP18

基于免疫遗传算法的多维多层关联规则挖掘

朱 玉，张 虹，孔令东
(中国矿业大学环境与测绘学院，徐州 221008)

摘 要：提出一种基于免疫遗传算法的多维多层关联规则挖掘算法。免疫遗传算法具有很好的鲁棒性及全局搜索能力，能快速有效地进行全局优化搜索。针对现有多维多层关联规则挖掘中存在阈值定义不合理的缺点，依据多维和多层数据的共同特点，给出一种启发式的阈值自定义方式，结合免疫遗传算法提高挖掘效率和结果的准确性。结果表明，挖掘效率和质量有明显提高。

关键词：免疫遗传算法；多维多层；关联规则；数据挖掘

Multi-dimension and Multi-level Association Rule Mining Based on Immune Genetic Algorithm

ZHU Yu, ZHANG Hong, KONG Ling-dong
(School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221008)

【Abstract】 This paper proposes a method of mining multi-dimension and multi-level association rule based on the Immune Genetic Algorithm(IGA). The algorithm has good robustness and good whole global search capability. It searches more quickly and efficiently in the whole global optimization. It puts forward a heuristic user-defined threshold method, which based on the common characteristic of multi-dimensional and multi-level data, to overcome the drawbacks caused by the unreasonable method of defining threshold. Combined with the immune genetic algorithm, the precision and efficiency of mining association rules is improved. Results show that the efficiency and quality improved obviously.

【Key words】 Immune Genetic Algorithm(IGA); multi-dimension and multi-level; association rule; data mining

1 概述

关联规则的挖掘是数据挖掘领域中的一个重要研究方向，它通过挖掘数据库中的数据项集之间的某种潜在的关系，从而在大量数据中发现一些潜在和有趣的关联关系，以此为依据来帮助决策者做出合理、适当的决定[1]。关联规则根据不同的划分标准分为维内、多维规则以及单层、多层关联规则。对于维内关联规则的挖掘，已经研究得比较多。而对于多维多层关联规则的挖掘的研究相对要少得多。目前，多维多层挖掘算法普遍采用了人工定义不同层和不同维阈值的方式，并对经典的 Apriori 算法进行了改进[2]。当维和层较多的时候，需要人工定义大量的阈值，随意性较大，并且由于算法的复杂度增加，搜索空间、存储空间、运行时间也将大量增加[3]。

遗传算法[4](Genetic Algorithm, 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。目前，对标准的遗传算法已经做了很多的改进。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等问题。

近年来，受生物免疫理论启示的人工免疫系统方法研究成为热点。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[5]。

本文利用不同维和不同层取值多少这个先验知识，提出了一种启发式的阈值自定义方法，并将免疫遗传算法引入多维多层关联规则挖掘中，通过对多维多层关联规则挖掘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，有效克服了多维多层关联规则挖掘中的固有的缺点，收敛到全局最优解。通过实验证明了其良好的性能。

2 关联规则相关概念描述

2.1 关联规则

关联规则是数据项之间存在的规则，是在同一事件中出现的不同项之间的相关性。设 $I=\{i_1, i_2,\cdots, i_m\}$为项的集合，$DB$ 为事务集合，其中，每一个事务 T 都是项的集合，且有 $T\subseteq I$，每一个事务都有一个相关的标识符 TID 和它对应。设 X 为一个项集，当且仅当 $X\subseteq T$ 时，可以说事务 T 包含 X。关联规则是形如 $A\rightarrow B$ 的蕴涵式，其中 $A\subset I, B\subset I$，并且 $A\cap B=\varnothing$。一般用下面 2 个参数描述关联规则的属性：

(1)支持度 S(support)：事务集 DB 中同时包含事务 A 和 B 的百分比，称为规则 $A\Rightarrow B$ 具有支持度 S。

(2)置信度 C(confidence)：事务集 DB 中包含 A 的事务数与同时包含 B 的事务数的百分比，称为规则 $A\Rightarrow B$ 具有置信度 C。

支持度(用 Sup 表示)与置信度(用 $Conf$ 表示)的计算方法：

基金项目：江苏省自然科学基金资助项目(BK2005021)；江苏省普通高校研究生科研创新计划基金资助项目(CXB-1392)
作者简介：朱 玉(1977－)，男，博士研究生，主研方向：人工免疫，空间数据挖掘；张 虹，教授、博士生导师；孔令东，博士研究生
收稿日期：2009-07-20 **E-mail：**zhuyujs@139.com

$Sup(X \Rightarrow Y)$=(包含 X 和 Y 的事务数/事务总数)×100%;

$Conf(X \Rightarrow Y)$=(包含 X 和 Y 的事务数/包含 X 的事务数)×100%。

同时满足最小支持度(Sup_{min})和最小置信度($Conf_{min}$)的规则被称为强关联规则，即在关联规则挖掘中希望发现的关联规则[6]。

2.2 关联规则的挖掘过程

挖掘事物集合 DB 中所有关联规则的问题可以被划分为以下 2 个子问题:(1)找出所有具有最小支持度的项集(即频繁项集); (2)由频繁项集产生强关联规则，对于每一个频繁项集 l，找出其中所有的非空子集，然后对每一个这样的子集 a，如果 $Sup(l)$与 $Sup(a)$的比值大于最小置信度，则存在规则 $a \Rightarrow (l-a)$。

2.3 多维多层关联规则挖掘

关联规则中的多维关联规则是指各个属性维之间存在的关联规则，多层关联是指挖掘的规则内涉及不同的概念层次。目前多维多层关联规则挖掘算法多数采用了经典关联推荐算法。该算法主要采用自顶向下、逐层深入的方法进行挖掘，直到挖掘出最底层的强规则，在挖掘每一层上的关联规则时仍采用 Apriori 算法思想。多层关联规则的挖掘基本上采用“支持度-可信度”的框架，根据规则中涉及到的层次，采用递减最小支持度的方法，每个层次都有不同的支持度阈值，较低层次的最小支持度相对较小。而对于多维关联规则的挖掘则是在多维数据库之间更多有用的规则，同样涉及多阈值问题。而要确定算法中的多阈值通常有 2 种方法: (1)按比例增减阈值; (2)人工定义不同层和维的阈值。这 2 种方法的共同缺点是不同的阈值要完全凭借个人的经验来决定，而没有更加合理地参照依据，另外，当层和维比较多时，需要确定的阈值也较多。

3 免疫遗传算法

3.1 免疫遗传算法原理

免疫遗传算法(Immune Genetic Algorithm, IGA)[7]是基于生物免疫机制提出的一种改进的遗传算法，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。由生物免疫原理可知，生物免疫系统对入侵生命体的抗原通过细胞的分裂和分化作用，自动产生相应的抗体来抵御，这一过程被称为免疫应答。在免疫应答过程中，部分抗体作为记忆细胞保存下来，当同类抗原再次侵入时，记忆细胞被激活并迅速产生大量抗体，使再次应答比初次应答更快更强烈，体现了免疫系统的记忆功能。抗体与抗原结合后，会通过一系列的反应而破坏抗原。同时，抗体与抗体之间也相互促进和抑制，以维持抗体的多样性及免疫平衡，这种平衡是根据浓度机制进行的，即抗体的浓度越高，则越受抑制; 浓度越低，则越受促进，体现了免疫系统的自我调节功能。

免疫遗传算法与标准遗传算法相比，具有如下特点:

(1)产生多样抗体的能力: 通过细胞的分裂和分化作用，免疫系统可产生大量的抗体来抵御各种抗原，这对应于遗传算法中个体的多样性。这种机制可用于提高遗传算法的全局搜索能力而不陷于局部最优。

(2)自我调节机构: 免疫系统具有维持免疫平衡的机制，通过对抗体的抑制和促进作用，能自我调节产生适当数量的必要抗体。这对应于遗传算法中个体浓度的抑制和促进，利用这一功能可以提高遗传算法的局部搜索能力。

(3)免疫记忆功能: 产生抗体的部分细胞会作为记忆细胞而被保存下来，对于今后侵入的同类抗原，相应的记忆细胞会迅速激发而产生大量的抗体。如果遗传算法中能利用这种抗原记忆识别功能，则可以加快搜索速度，提高遗传算法的总体搜索能力。

3.2 算法流程

免疫遗传算法实现的流程如图 1 所示。

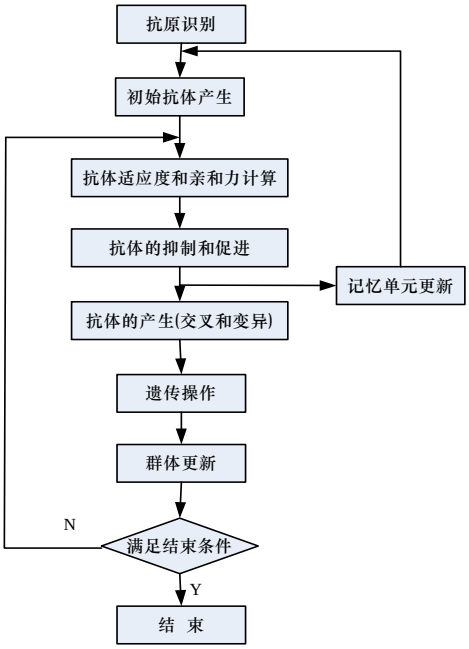

图 1 免疫遗传算法流程

对图 1 的说明如下:

(1)抗原识别。输入目标函数和各种约束作为免疫系统的抗原。

(2)初始抗体的产生，即生成初始解。抗原识别单元中，若系统求解过此类问题，则从记忆细胞库中搜寻该类问题的记忆抗体，从而生成初始抗体。不足的抗体由随机的方法在解空间中产生。

(3)亲和度计算。亲和度指两者的关联性。在生物体内，抗体与抗原都分别含有抗体和抗原决定基，它们均可以存在相互作用。故可以定义 2 类亲和度: 抗体与抗原之间的亲和度以及抗体与抗体之间的亲和度，前者相当于遗传算法中的适应度，后者体现了不同抗体之间的相似程度。

(4)记忆单元更新。将与抗原亲和度高的抗体加入到记忆存储单元中。

(5)抗体的抑制和促进。在免疫算法中，与抗原亲和度高的抗体自然受到促进，以较高的概率进入下一代，但这样往往会导致种群过于单一，易陷入局部最优，所以，要在算法中适当地采用抑制策略以保持种群中抗体的多样性，可以在构造抗体的选择概率时加入抗体浓度因素来实现。

(6)当前种群中抗体交叉、变异生成新一代抗体，进入下一代。算法通过综合考虑抗体适应度和其在种群中的浓度，构造选择概率对其进行选择，对选择出来的抗体群进行遗传操作(交叉、变异)，产生新一代抗体。既确保抗体群整体朝着适应度高的方向进化，又维持了种群中抗体的多样性。

4 基于免疫遗传算法的多维多层关联规则挖掘

4.1 算法基本思想

在多维多层关联规则挖掘中，不同维和层的不同取值个数是和阈值有密切关系的，不同取值个数和阈值大小呈反比，

取值个数越多阈值应越小，个数越少阈值应越高。因此，算法中只设置一个阈值，而所有维的最底层依据该维最底层的不同值个数多少作为先验知识，启发式的定义每个维中最底层的阈值。每一维的高概念层的阈值，由该层不同取值数和同一维中子层的不同取值数比较产生阈值，并且为了防止出现阈值大于1的情况，设置一个阈值上限。

则第 i 维最底层的最小支持度和最小可信度分别为

$$Sup_{\min}(X_{i1})=\frac{\sum_{i=1}^{n}d_{i1}}{nd_{i1}}Sup_{\min} \tag{1}$$

$$Conf_{\min}(X_{i1})=\frac{\sum_{i=1}^{n}d_{i1}}{nd_{i1}}Conf_{\min} \tag{2}$$

则第 i 维 j 层的阈值为

$$Sup_{\min}(X_{ij})=\frac{d_{ij-1}}{d_{ij}}Sup_{\min}(X_{ij-1}) \tag{3}$$

$$Conf_{\min}(X_{ij})=\frac{d_{ij-1}}{d_{ij}}Conf_{\min}(X_{ij}-1) \tag{4}$$

其中，n 为所有维的总数；d_{ij} 是第 i 维 j 层所拥有的不同取值个数；$Sup_{\min}$ 为最小支持度；$Conf_{\min}$ 为最小可信度。另外，$Sup_{\max}$ 为最小支持度上限；$Conf_{\max}$ 为最小置信度上限。

关联规则的挖掘通常分 2 步：(1)找出满足最小支持度的频繁规则；(2)由频繁规则产生强关联规则。第(1)步尤为重要，决定整个算法性能。本算法采用免疫遗传算法产生频繁规则。然后计算这些规则的各种组合是否满足最小可信度，从而产生强关联规则。

4.2 编码方案

用于关联规则发现的主要是事务型数据库，因此，在编码过程中抗体抗原主要采取搜索能力较强的二进制编码方式。在编码之前首先应对原始数据进行预处理，即对连续属性值进行离散化，经过离散化得到每个属性的若干个取值。将所有的决策属性和任务属性构造一个规则结构串，形如：$\{B_1, B_2,\cdots, B_m, A_1, A_2,\cdots, A_n\}$，其中 A_i 表示决策属性，B_j 表示任务属性，则根据规则结构串遗传编码为 $X=\{b_1b_2\cdots b_mx_1x_2\cdots x_n\}$，其中，$b_1b_2\cdots b_m$ 对应任务属性，$x_1x_2\cdots x_n$ 对应决策属性。每个属性可以采用合适长度的二进制编码，然后按顺序串接起来。

4.3 适应度函数

取适应度函数为

$$F(X)=w_c\times\frac{Conf(X)}{Conf_{\min}}+w_s\times\frac{Sup(X)}{Sup_{\min}} \tag{5}$$

其中，$w_c+w_s=1(w_c\geqslant 0, w_s\geqslant 0)$。

4.4 遗传操作的确定

遗传操作主要包括选择、交叉和变异。

(1)选择操作

采用基于免疫机制的选择策略。其方法为由个体的浓度概率 P_d 和适应度概率 P_f 决定选择概率 P：

$$P=\alpha P_f+(1-\alpha)P_d \qquad \alpha\text{为常数且}0<\alpha<1 \tag{6}$$

$$P_f=\frac{\text{群体中个体的适应值}}{\text{群体中所有个体适应值之和}}$$

$$P_d=\begin{cases}\dfrac{1}{M}(1-d) & \text{群体中浓度最大的个体}\\ \dfrac{1}{M}\left(1+\dfrac{d^2}{1-d}\right) & \text{群体中其他个体}\end{cases}$$

其中，M 为群体规模；d 为个体浓度，定义为

$$d=\frac{\text{群体中相同个体的数目}}{\text{群体规模}M}$$

(2)交叉操作

算法采用双点交叉法进行交叉。

设 $\boldsymbol{X}_1^l=\left[x_1^1, x_2^1, \cdots, x_n^1\right]$，$\boldsymbol{X}_2^l=\left[x_1^2, x_2^2, \cdots, x_n^2\right]$ 是 l 代的 2 个抗体，在第 i 个点和第 j 个点实施双点交叉，产生的下一代抗体是：

$$\boldsymbol{X}_1^{l+1}=\left[x_1^1x_2^1\cdots x_i^{'}\cdots x_j^{'}x_{j+1}^1\cdots x_n^1\right],\ \boldsymbol{X}_2^{l+1}=\left[x_1^2x_2^2\cdots x_i^{''}\cdots x_j^{''}x_{j+1}^2\cdots x_n^2\right]$$

其中，$x_k^{'}$ 和 $x_k^{''}$ $(i\leqslant k\leqslant j)$由下述线性组合产生：

$$x_k^{'}=\zeta x_k^1+(1-\zeta)x_k^2,\quad x_k^{''}=\zeta x_k^2+(1-\zeta)x_k^1 \tag{7}$$

其中，比例系数$\zeta\in[0, 1]$。

(3)变异操作

依据变异概率 $P_m(0.01\leqslant P_m\leqslant 0.1)$随机产生变异位，对变异位进行求反操作。

4.5 基于浓度的群体更新

基于浓度的群体更新总的目标是抑制浓度过高的抗体，同时保证适应度高的个体被选中的概率更大。抗体的浓度 C 定义为群体中具有最大适应度或近似最大适应度的抗体个数与群体中抗体总数的比率，调整个体的选择几率为 $p(i)$:

$$p(i)=\alpha C\left(1-\frac{F(i)}{\max Fitness}\right)+\beta\frac{F(i)}{\max Fitness} \tag{8}$$

其中，α, β 为 0 到 1 之间的可调参数；max$Fitness$ 为抗体的最大适应度或近似最大适应度；C 为抗体浓度。

从式(8)可以看出：当抗体浓度高时，适应度高的抗体被选中的几率就小；当抗体浓度不高时，适应度高的抗体被选中的几率就大。这样既保留了优秀个体，又可减少相似抗体的选择，确保了个体的多样性。

4.6 具体算法流程

具体算法流程如下：

Step1 输入算法参数(进化代数 T、群体规模 M 等)和数据集合。

Step2 初始化，t=0，随机生成初始群体。

Step3 计算个体支持度、置信度和适应度。

Step4 若 $t\geqslant T$ 或群体满足要求转 Step5，否则，转 Step6。

Step5 过滤出符合阈值的个体，输出最后群体并解释规则。

Step6 i=0，使用基于免疫机制的策略选择 2 个个体。

Step7 使用双点交叉对个体进行交叉，以变异概率 P_m 对新个体进行变异操作。

Step8 i=i+2，若 $i\geqslant M$，进行基于浓度的抗体选择，产生新一代群体并转 Step3，否则，转 Step6。

5 仿真实验

以某矿瓦斯突出监测数据库中的数据为例，数据库记录了某段时间瓦斯突出监测的 7 826 条记录，应用免疫遗传算法挖掘数据库中蕴涵的属性之间的关联性，取群体规模 M 为 200，进化代数 T 为 200 代，$Sup_{\min}$=0.01, $Conf_{\min}$=0.02, $Sup_{\max}$=0.8, $Conf_{\max}$=1.0, P_m=0.05，在式(5)中，w_c=0.6, w_s=0.4，在式(6)中 α=0.5，在式(7)中 ζ=0.6，在式(8)中 α=β=0.5。测试时挖掘出关联规则 292 条，挖掘时间为 21.7 s。

使用上述数据测试 Apriori 算法和 GA 算法的执行效率，实验结果比较如图 2 所示。可见，在多维多层关联规则挖掘中，免疫遗传算法的执行效率要比 Apriori 算法和 GA 算法有

明显的提高。 （下转第 186 页）