第35卷 第22期 计 算 机 工 程 2009年11月
Vol.35 No.22 Computer Engineering November 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)22—0216—02 文献标识码：A 中图分类号：TP311.12

基于混合遗传克隆算法的关联规则挖掘

符保龙
(柳州职业技术学院信息工程系，柳州 545006)

摘 要：针对在数据挖掘应用中关联规则挖掘的问题，给出一种基于混合遗传克隆算法的关联规则挖掘方法，该算法将遗传算法和克隆算法优点相结合，通过克隆操作来产生一组新的个体，独立地对所产生的各个体进行变异，交叉操作，同时采用自适应方式动态选取交叉和变异概率，有效地克服了遗传算法容易陷入局部最优的缺点，从而求得问题的最优解。实验结果表明，该方法能高效地解决关联规则挖掘问题。

关键词：数据挖掘；关联规则；遗传算法；克隆算法

Association Rule Mining Based on Hybrid Genetic Clonal Algorithm

FU Bao-long
(Department of Information Engineer, Liuzhou Vocational Technological College, Liuzhou 545006)

【Abstract】 Aiming at the problem of association rules mining in the application of data mining, this paper proposes a method of mining association rules based on Hybrid Genetic Clonal Algorithm(HGCA). This algorithm combines with the Clonal Algorithm(CA) and Genetic Algorithm(GA) to fully exert respective advantages. It generates a new group of individuals through clonal operation, makes mutation and crossover independently all the generated individuals respectively, uses adaptive crossover probability and mutation probability so as to restrain premature convergence. Experimental results demonstrate that this method can solve association rule mining effectively.

【Key words】 data mining; association rule; Genetic Algorithm(GA); Clonal Algorithm(CA)

1 概述

数据挖掘(data mining)是从大量的无规律的繁杂数据中抽取出潜在的、不为人知的有用信息、模式和趋势的过程[1]。当今信息社会中数据和数据库的爆炸式增长，使人类分析数据并从中提取有用信息的能力远远不能满足实际需要。关联规则挖掘技术为解决这类问题提供了行之有效的途径，它能够从繁杂信息中自动发现隐藏在数据中的模式信息，了解用户的行为模式，从而做出预测性分析。

近年来，人们研究了许多挖掘算法。Apriori 算法是关联规则挖掘的经典算法，但计算复杂度高，不能满足对大规模数据库的实时挖掘要求；Park 等人提出 DHP 算法，可是修整和剪枝属性在许多实际应用中是不切实际的；分割算法减少了 I/O 消耗，但在处理高维项目集事件中存在问题；FP-growth 算法[2]创建了一种关系紧密的树结构，有助于候选项目集的产生，但需要遍历整个数据库，计算量大，对大规模数据库而言，效率非常低。文献[3-4]中提到了基于遗传算法(Genetic Algorithm, GA)的关联规则挖掘方法。虽然 GA 具有很强的随机性、鲁棒性和隐含并行性，能快速、有效地进行全局优化搜索是处理大规模数据项目集的有效方法，但是它在选择进化的过程中，并未对优秀个体的信息予以充分利用，而只是对这些个体给予简单机械的重复保留；另外 GA 在交叉和变异进化中的盲目性和随机性，导致搜索效率不高。

针对上述算法的缺点，本文在 GA 基础上，引入免疫克隆思想，提出了一种基于混合遗传克隆算法(Hybrid Genetic Clonal Algorithm, HGCA)的挖掘方法。

2 问题的定义

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。为便于分析，作以下形式化定义：

定义 1(数据项集 I) $I=\{i_1,i_2,\cdots,i_n\}$，其中 i_k $(1\le k\le n)$ 是数据项。

定义 2(事务数据集 T) $T=\{T_1,T_2,\cdots,T_m\}$，T_k $(1\le k\le m)$ 是事务数据集 T 的数据项，也是数据项集 I 中的数据项，并且 $T\subseteq I$ 一种关联规则是形如 $X\Rightarrow Y$ 的蕴涵关系，其中，$X\subset I$，$Y\subset I$，并且 $X\cap Y=\varnothing$

定义 3 支持度 $S(X\Rightarrow Y)$：

$$S(X\Rightarrow Y)=\frac{|\{D:X\cup Y\subseteq T,D\in T\}|}{|T|}$$

如果 T 中由 S% 的记录支持子集 X，称支持度为 S。它反映了该规则 $X\Rightarrow Y$ 在 T 中所占的比例，说明了 $X\Rightarrow Y$ 在事务集 T 出现的普遍程度。

定义 4 可信度 $C(X\Rightarrow Y)$：

$$C(X\Rightarrow Y)=\frac{|\{D:X\cup Y\subseteq T,D\in T\}|}{|D:X\subseteq T,D\in T|}$$

可信度 $C(X\Rightarrow Y)$ 说明 $X\Rightarrow Y$ 成立的必然程度。它表明在 T 中支持 X 的事务中，有 C% 的事务同时也支持 Y。

基金项目：广西省教育厅科研基金资助项目(200808LX242)
作者简介：符保龙(1978－)，男，讲师、硕士，主研方向：数据挖掘，演化计算
收稿日期：2009-03-10 **E-mail：**gxsony@163.com

3 HGCA 算法的应用

3.1 HGCA 算法模型

克隆算法(Clonal Algorithm, CA)是模拟生物免疫系统的多克隆机理，不仅采用变异，交叉等操作实现抗体间的信息交换，而且还要充分利用抗体在变化过程中已经获得的对抗原反应的特性，进一步增加克隆的多样性[5]。为提高算法的运行效率，有效控制抗体规模，本文对遗传算法融入免疫克隆思想，设计的 HGCA 算法模型定义如下：

定义 5(HGCA 算法模型) $\mathrm{HGCA}=(\varphi, f, A, N, \Gamma, \Psi, \Phi, \Xi)$ 是本文设计的挖掘算法。其中，φ 是抗体的编码；f 是抗体的亲和度评价函数；A 是初始抗体群；N 是群体规模；Γ 是克隆算子；Ψ 是交叉算子；Φ 是变异算子；Ξ 是终止条件。

3.2 抗体编码

在 HGCA 算法编码中，字符串代表抗体，它是遗传信息传递的载体，抗体中的每个位置的元素代表遗传因子。采用实数数组的方法进行编码，实数数组的元素个数与事务数据库中的字段的个数相对应，元素值代表了字段的属性值。

用一个长度为 n 的数组来表示事务数据库的个体编码，$A[1]$ 表示字段 1，$A[n]$ 表字段 n；用数值 k $(1 \le k \le n)$ 表示属性值 k，则 $A[k]$ 表示相对应的字段的属性值。用 0 值表示此属性与其他的属性无关联。HGCA 算法中的抗体定义如下：

定义 6(抗体编码 φ) $\varphi = A[1]A[2]\cdots A[i]\cdots A[n]$，其中，$A[i]$ 表示抗体基因；n 为抗体长度，$i=1,2,\cdots,n$。

3.3 亲和度函数

定义 7 亲和度函数 $f(X \Rightarrow Y)$：

$$f(X \Rightarrow Y)=S(X \Rightarrow Y)+C(X \Rightarrow Y)$$

$f(X \Rightarrow Y)$ 表明在各种规则的竞争中，只有支持度和可信度都高才有可能生存下来。

3.4 克隆

在 HGCA 算法中，克隆算子能有效扩大群体的规模。每个抗体与抗原的亲和度越大，抗体的克隆规模也就越大。抗体群 A 中每一个抗体 a_i 按规模 $\ln(\alpha \times N / i)$ 克隆到新的抗体群，α 为克隆系数，$\ln(*)$ 为自然对数函数。

3.5 交叉

交叉是指把 2 个父代抗体的部分结构加以替换重组而生成新抗体的操作。交叉概率 P_c 直接影响算法的收敛性。P_c 越大最佳抗体的遗传模式被破坏的可能性越大；P_c 过小会使算法搜索过程缓慢。本文定义的交叉算子如下：

(1)计算交叉概率 P_c

$$P_c=\begin{cases} 0.9-\dfrac{0.01\times(f(\mathrm{X})-\overline{f(\mathrm{X})})}{f_{\max}(\mathrm{X})-f_{\min}(\mathrm{X})} & f(\mathrm{X}) \ge \overline{f(\mathrm{X})} \\ 0.9 & f(\mathrm{X}) < \overline{f(\mathrm{X})} \end{cases}$$

其中，$f_{\max}(\mathrm{X})$ 为抗体中最大的亲和度值；$f_{\min}(\mathrm{X})$ 为抗体中最小的亲和度值；$\overline{f(\mathrm{X})}$ 为每代抗体群的平均亲和度值。

(2) $V_1^{'}=P_c \times V_1+(1-P_c)\times V_2$，$V_2^{'}=P_c \times V_2+(1-P_c)\times V_1$。$V_1, V_2$ 分别是父抗体，$V_1^{'}, V_2^{'}$ 分别为子抗体。

3.6 变异

变异可以提高群体中抗体的多样性，扩大搜索范围，用来寻找更优秀的抗体。变异概率 P_m 的选择是影响算法行为和性能的关键所在。P_m 过小不易产生新的个体结构；若 P_m 过大，则遗传算法变成纯粹的随机搜索。本文定义的变异算子如下：

(1)计算变异概率 P_m

$$P_{\mathrm{m}}=\begin{cases} 0.01-\dfrac{0.09\times(f(\mathrm{X})-\overline{f(\mathrm{X})})}{f_{\max}(\mathrm{X})-f_{\min}(\mathrm{X})} & f(\mathrm{X}) \ge \overline{f(\mathrm{X})} \\ 0.01 & f(\mathrm{X}) < \overline{f(\mathrm{X})} \end{cases}$$

其中，$f_{\max}(\mathrm{X})$ 为抗体中最大的亲和度；$f_{\min}(\mathrm{X})$ 为抗体中最小的亲和度；$\overline{f(\mathrm{X})}$ 为抗体的平均亲和度。

(2) $V^{'}=V+P_m \times \exp(-f(*))\times N(0,1)$，$V$ 和 $V^{'}$ 分别是父抗体和子抗体；$N(0, 1)$ 是均值为 0，方差 $\sigma=1$ 的高斯变量；P_m 是变异概率；$f(*)$ 是 V 的亲和度。

3.7 HGCA 算法框架

HGCA 算法流程如下：

(1)随机在解空间产生初始抗体群 $A=\{a_1, a_2, \cdots, a_N\}$，种群规模 N，最大进化代数 $MaxGen$，并进行编码。

(2)将抗体群 A 中的抗体按照亲和度由大到小降序排列，得到 $A(k)=\{b_1, b_2, \cdots, b_N\}$，且 $f(b_i) \ge f(b_{i+1})$，$f(*)$ 为亲和度函数，$k=0$。

(3)计算抗体支持度、可信度；将结果添加到有趣规则表，并将大于最小阈值的规则添加到关联规则表中。

(4)计算抗体群中每个抗体 b_i 的亲和度 $f(*)$，$i=1, 2, \cdots, N$。

(5)分别对抗体进行克隆，交叉，变异操作得到下一代抗体。

(6)当 $k \ge maxGen$ 时，算法结束，从关联规则表中提取关联规则；否则，$k=k+1$，返回到步骤(2)。

4 实验及结果分析

本文将算法 HGCA 应用到关联规则挖掘实验中，算法执行中用到的一些参数设置如下：种群规模为 100，最大进化代数 $maxGen$ 为 200。在人工数据集上分别运行 HGCA, GA, CA 3 种算法，该数据集包含 600 个事务和 20 个属性。算法运行了 50 次。得到的平均结果显示如图 1 和图 2 所示。每次算法收敛，关联规则的最大数目绝大多数都出现在 100 次迭代。

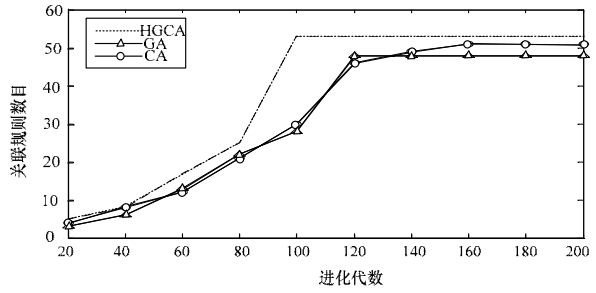

图 1 3 种算法挖掘的关联规则数目

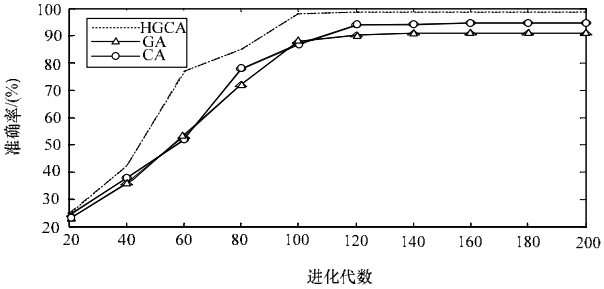

图 2 3 种算法挖掘产生的规则准确率

从图 1 中可以看出，HGCA 算法能有效地挖掘关联规则，并且有较好地收敛速度。由图 2 可见，HGCA 算法挖掘出来

(下转第 220 页)