文章编号: 1001-0920(2008)05-0497-06

基于疫苗提取及免疫优化的粗糙集属性约简

徐雪松[1], 章 兢[1], 贺 庆[2], 何昭晖[1], 王炼红[1]
(1. 湖南大学 电气与信息工程学院, 长沙 410082; 2. 中南大学 信息科学与工程学院, 长沙 410083)

摘 要: 针对约简属性组合的爆炸问题,将 RS 属性核参数作为先验信息的免疫疫苗引入抗体编码,概率性对种群接种疫苗. 将属性集合的分类近似标准作为抗体适应度,通过在免疫克隆选择过程中引入聚类竞争机制,提高抗体群分布的多样性及亲和力成熟,从而获得多个属性约简及最小约简的平衡. 实验结果表明,这种粗糙集属性约简方法对于多维条件属性集是快速且有效的.

关键词: 克隆选择; 粗糙集; 属性约简; 数据挖掘
中图分类号: TP182 **文献标识码**: A

Attribute reduction method of rough sets based on bacterin extraction and immune optimum

XU Xue-song[1], *ZHANG Jing*[1], *HE Qing*[2], *HE Zhao-hui*[1], *WANG Lian-hong*[1]
(1. College of Electrical and Information Engineering, Hu'nan University, Changsha 410082, China; 2. College of Information Sciences and Engineering, Central South University, Changsha 410083, China. Correspondent: XU Xue-song, E-mail: daniel613@126.com)

Abstract: Aiming at the problem of large scales of attribute reduction, a prior knowledge of attribute kernel as bacterins is introduced to antibody coding and the population is vaccinated in a stochastic way. The classification approximation quality is taken as the antibody affinity, then a cluster and competition mechanism is applied in clonal selection process to enhance the diversity of antibody and affinity maturation. Minimum reductions and more reductions of the rough sets are obtained. Experimental results show that the approach is effective and quick in solving attribute reduction, and has a remarkable quality of the global convergence reliability and convergence velocity.

Key words: Clonal selection; Rough sets; Attribute reduction; Data mining

1 引 言

现实数据挖掘过程中数据结构和形式多种多样,如何从大量的属性中选择有用的属性子集,提高分类规则及有效信息获取的准确性,使得属性约简及规则提取成为数据挖掘研究的一个重要内容. Wong[1]已证明,找出决策表的最小约简是一个 NP 难题,其原因是属性多元化及属性组合的爆炸问题. Kryszkiewicz[2]在经典 Rough 集基础上构建了相似模型,对不确定信息系统进行推理,发现决策规则. 文献[3]将粗糙集理论与模糊数学相结合探讨属性的约简问题. 文献[4]从信息角度对决策系统中的属性重要度进行度量,提出一种基于信道容量的知识相对约简算法来减小知识约简过程中的搜索空间. 文献[5]以属性的重要性作为启发式信息减小约简过程中的搜索空间,提出了基于互信息的知识相对约简算法. 还有学者采用进化算法方法对其进行优化求解,如文献[6]结合遗传算法优化原理求取决策表中一个最小的相对约简集合,文献[7]采用免疫算法对故障诊断属性集合进行全局优化,均取得了一定效果.

本文根据克隆选择算法优秀的全局及局部寻优能力,将粗糙集属性集合的核属性提取作为先验信息的免疫疫苗引入抗体编码. 同时在克隆选择过程中引入聚类竞争机制,使各个小局域中相对较为优秀的抗体能迅速被选中,进入克隆扩增过程,实现抗体亲和力的成熟,并提高抗体群的多样性,加快收敛

收稿日期: 2007-01-28; **修回日期**: 2007-05-22.
基金项目: 国家自然科学基金重点项目(60634020); 教育部博士点基金项目(20060532026).
作者简介: 徐雪松(1978—),男,湖南郴州人,博士生,从事机器学习及数据挖掘、进化计算的研究; 章兢(1957—),男,湖南湘潭人,教授,博士生导师,从事智能控制、复杂系统控制等研究.

速度.通过对决策表中的条件属性进行简化,实现从多属性约简集合寻求最优选择,并提高粗糙集理论的实际应用能力.

2 粗糙集的约简与核

粗糙集理论是 1982 年由 Pawlak[8] 提出的一种描述不完整性和不确定性的数学理论.它从新的角度对知识进行了定义,将知识看作是关于论域的划分,从而认为知识是有粒度的.知识的粒度性是造成使用已有知识不能精确地表示某些概念的原因.

一个决策表的信息系统 S 可表示为 $S = (U, R, V, f)$.其中:U 是有限非空集合,称为论域;$R = C \cup D, C \cap D = \varnothing$, R 为属性集合,C 和 D 分别为条件属性和决策属性;$V = \bigcup_{r \in R} V_r$ 为属性值集合;$f: U \times R \rightarrow V$ 是一个信息函数,它指定 U 中每个对象的属性值.对 $\forall X \subseteq U$ 和 $\mathrm{Ind}(B)$,X 的上近似集 $\overline{B}X$ 和下近似集 $\underline{B}X$ 为

$$\overline{B}X = \{x \mid (x \in U \wedge [x]_B \subseteq X)\},$$

$$\underline{B}X = \{x \mid (x \in U \wedge [x]_B \cap X \neq \varnothing)\}. \quad (1)$$

集合 $BN_B(X) = \overline{B}(X) - \underline{B}(X)$ 称为 X 的 B 边界域;$\mathrm{POS}_B(X) = \underline{B}(X)$ 称为 X 的 B 正域.

属性约简是粗糙集理论的一个核心内容,粗糙集理论在知识表达系统的基础上定义了约简与核这两个非常重要的概念,进而提供了分析多余属性的方法,对知识的处理是通过对决策表中的属性值的处理实现的.

令 R 为一等价关系簇,并且 $r \in R$,如果 $\mathrm{Ind}(R) = \mathrm{Ind}(R - \{r\})$,称 r 为 R 中可省略的;否则 r 为 R 中不可省略的.对于属性子集 $P \in R$,若存在 $Q = P - r, Q \in P$,使 $\mathrm{Ind}(Q) = \mathrm{Ind}(P)$,且 Q 为最小子集,则 Q 称为 P 的约简,表示为 $\mathrm{Red}(P)$.P 中所有约简属性集中都包含的不可省略关系的集合,即约简集 $\mathrm{Red}(P)$ 的交集称为 P 的核,它是表达知识必不可少的重要属性集.表示为 $\mathrm{Core}(P)$,即

$$\mathrm{Core}(P) = \cap \mathrm{Red}(P). \quad (2)$$

核属性是描述对象不可缺少的属性.实际上,一般产生约简的方法是逐个向核中添加可省略的属性并进行检查.由于可省略属性关系集合的幂集基数是多少就有多少种添加的方式,随着规则表的不断增大,规则约简的复杂性呈指数增长.符合条件的属性集合非常多,其中最简集合通常也不唯一,这样的集合约简可归纳为多模态优化模型,计算所有约简与求取最佳约简成为 NP 难题.

3 属性约简的免疫优化

1959 年 Burnet[9] 提出了克隆选择学说.Castro[10] 在其研究中描述了免疫克隆优化算法的实现,指出了传统进化算法强调自然选择中的个体竞争而一般保存种群规模不变.克隆操作一方面通过抗体-抗原亲和度实现个体间的竞争;另一方面利用亲和度调节或抑制过度竞争,以保存抗体群的多样性,并通过个体增生为某一抗体同时采用多种变异和重组策略提供了可能.同时由于免疫进化为分布无中心控制的并行操作,使得在抗体群体中,每个抗体在进化过程中都能独立进行优化.而新生成的抗体可在更广泛的搜索空间中进行寻优,从而实现了多模态的局部和全局寻优,因此可有效应用于粗糙集属性约简的多集合寻优.

3.1 疫苗提取及初始抗体种群

在实际的免疫优化计算中,初始群体若是接近问题解,将缩短求解时间,提高算法效率.根据核属性定义可知,任何决策表的相对核具有唯一性.所以以属性核为启发式信息,将其作为疫苗注入抗体编码,对初始群体的选取进行优化.在进化过程中从各代种群中选出优良个体并提取免疫疫苗,概率地对后代种群的个体接种疫苗.接种疫苗是利用疫苗确定位上的等位基替代个体相应位上等位基因的操作.接种疫苗可加速优良模式的繁殖,修复被交叉和变异破坏的优良模式.通过种群与疫苗库相互作用、协同进化,极大地提高了其收敛速度.

假设有 10 个条件属性 $\{a_1, a_2, \cdots, a_{10}\}$ 的决策表属性核为 $\{a_3, a_6, a_8\}$,在初始群体选择时将其编码为“ * *1 * *1 *1 * * ”,并作为先验信息参数引入抗体编码.令 $A = a_1, a_2, \cdots, a_l$ 是二进制的抗体编码,记为 $A = e(X)$.抗体 A 中,a_i 被视为抗体基因,将抗体位串分为 m 段,每段长为 l_i.对其进行接种疫苗就是按照先验知识修改抗体基因某些位上的分量,使所得个体以较大的概率获取更高的适应度.疫苗来源于属性核的先验知识,它所包含的信息量及其准确性对算法的性能起到了重要作用.

3.2 适应度函数

适应度函数即为优化的目标函数.由粗糙集理论,假设 X_i 为论域上的一个子集.条件属性集 $P \subseteq C$,决策属性集为 D,Y_i 为决策属性的类别.可定义分类近似质量为

$$f_B(U, D) = \sum_{i=1}^{n} |\underline{B} - A_i| / |U|. \quad (3)$$

由式(1)可知,$\underline{B}(X) = \cup\{Y_i \in U/\mathrm{Ind}(B): Y_i \subseteq X\}$ 为集合 X 的下近似,由此可设定优化的适应度函数为

$$F(B) = f_B(U, D) + \left[\frac{N - m}{N}\right]. \quad (4)$$

3.3　免疫进化操作

不失一般性, 定义规模为 N 的初始抗体群 $A(k)$, 每个抗体的编码长度为 L, 形态空间为 S, 那么 $A(k) \in S^{N\times L}$. 从而可构建初始聚类竞争算法, 其流程如图 1 所示.

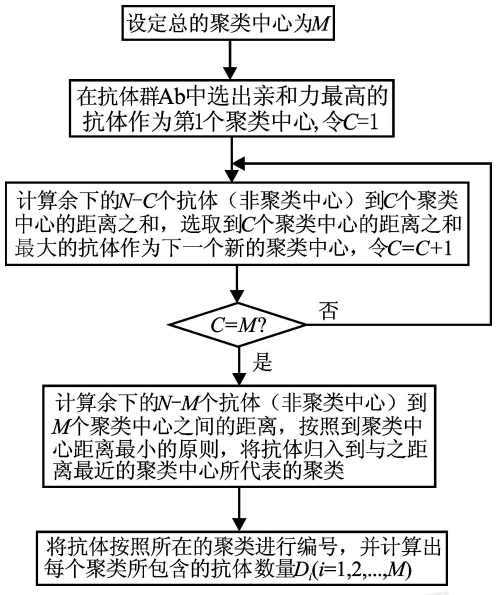

图 1　免疫抗体聚类竞争流程

聚类方式完成后, 得到规模为 N 的聚类抗体群 $A(k)$, 这 N 个抗体分别属于预先设定的 M 个聚类, 其 M 值可根据种群规模进行设定, 一般取 10 ～ 20. 相似的抗体经过聚类竞争后将会归属于同一个聚类, 其中只有亲和力最高的抗体和代表聚类中心的抗体才能被选中并进入克隆扩增及亲和力成熟过程. 引入竞争克隆机制, 促使抗体群中亲和力高且浓度低的抗体能进入克隆选择扩增的过程, 在每个聚类族的小局域中实行免疫克隆扩增、免疫基因及选择操作, 其算子操作定义如下.

定义 1　免疫克隆操作 T_c^C

$$T_c^C(A(k)) = [T_c^C(A_1(k)), T_c^C(A_2(k)), \cdots, T_c^C(A_m(k))]^{\mathrm{T}}. \quad (5)$$

其中: $T_c^C(A_i(k)) = I_i \times A_i(k)$, $i = 1,2,\cdots,n$, I_i 为元素 1 的 q_i 维行向量, 称 A_i 的 q_i 克隆. 对抗体群按下式进行克隆扩增:

$$N_i(k) = \mathrm{round}\left[N \times \frac{\mathrm{aff}(A_i(k))}{\sum_j \mathrm{aff}(A_j(k))}\right]_T. \quad (6)$$

其中: $N_i(i = 1,2,\cdots,T)$ 为第 i 个抗体的克隆规模; round(·) 为取整函数. 亲和力越高的抗体, 克隆规模越大, 克隆复制出的相同抗体就越多. 克隆过后, 种群变为

$$\overline{A}(k) = \{\overline{A}_1(k), \overline{A}_2(k), \cdots, \overline{A}_n(k)\}. \quad (7)$$

定义 2　免疫基因操作 T_g^C

免疫基因操作依据概率 p_m^i 对克隆后的群体进行变异操作.

$$P_g(A_{ij}(k) \to A_{ij}''(k)) = \begin{cases} (p_m^i)^{H(A_{ij}''(k),A_{ij}(k))}(1-p_m^i)^{1-H(A_{ij}''(k),A_{ij}(k))}, \\ \quad A_{ij}(k) \in \overline{A_i(k)}; \\ 0,\ A_{ij}(k) \notin \overline{A_i(k)}. \end{cases} \quad (8)$$

其中 $H(\cdot)$ 为海明距离.

定义 3　免疫选择操作 T_s^C

$\forall i = 1,2,\cdots,n$, 若

$$B_i(k) = \{A_{ij}''(k) \mid \max(f(A_{ij}''(k)), j = 1,2,\cdots,n\},$$

则有

$$p_s^k(A_i \to B) = \begin{cases} 1,\ f(A_i(k)) < f(B_i(k)); \\ \exp\left(-\dfrac{f(A_i(k)) - f(B_i(k))}{a}\right), \\ \quad f(A_i(k)) \geq f(B_i(k)),\ A_i(k) \text{ 不是最优抗体}; \\ 0,\ f(A_i(k)) \geq f(B_i(k)),\ A_i(k) \text{ 是最优抗体}. \end{cases} \quad (9)$$

其中 $a > 0$ 为抗体多样性阈值. 一般 a 越大, 多样性越好; 反之多样性越差.

定义 4　克隆删除操作 T_d^C

抗体 $A(k)$ 经过重组或突变后得到抗体 $A'(k)$, 如果 $A(k)$ 的亲和力低于重组或突变前的父代抗体 $A'(k)$, 即 $\mathrm{aff}(A'(k)) < \mathrm{aff}(A(k))$, 则删除 $A'(k)$, 用其父代抗体 $A(k)$ 代替.

克隆删除算子防止算法运行中抗体出现退化, 减缓收敛速度, 降低收敛的全局可靠性. 因此, 在这些免疫操作中, 免疫克隆操作 T_c^C 是一个 $I^n \to I^{N_c(k)}$ 的确定性映射, 实现空间的扩张. 免疫基因操作 T_g^C 在单一抗体周围产生一个变异解的群体, 利用局部搜索增加提高抗体和抗原亲和度的可能性, 这是一般进化算法所不具备的机理. 免疫选择操作 T_s^C 是一个 $I^{N_c(k)+n} \to I^n$ 的映射, 通过局部择优实现了种群的压缩. 免疫克隆运算通过空间的扩张和压缩, 将局部搜索和全局搜索结合起来实现问题的求解.

4　属性约简算法流程

4.1　算法执行过程

输入: 一个决策表 $S = (U, R, V, f)$, $R = C \cup D$, R 为属性集合, C 和 D 分别为条件属性和决策属性, $P \in R$.

输出: 决策表的一个属性约简 R.

Step1: 引入属性核 Core(P) 参数, 作为免疫疫苗初始化抗体群 Ab, 随机产生 N 个抗体, 并从中产生 $(N-1)/2$ 个疫苗, 组成疫苗库 $V(n)$.

Step2: 计算出决策属性对每个个体所含条件属性的个体适应值,选择符合式(4)的各自适应度最高的抗体,并计算该克隆群体的平均适应度.

Step3: 对抗体实施聚类及克隆、选择、变异等操作.

Step4: 计算其亲和力,若平均适应度显著变化,则返回 Step2 继续优化;否则将亲和力相近的抗体进行压缩,并剔除结构相同的抗体.

Step5: 随机产生规模为 N_r 的抗体群,从疫苗库 $V(n)$ 中随机选取疫苗抗体对种群 N_r 进行交叉接种操作,并将生成的新抗体作为补充抗体.

Step6: 选出亲和力最高的 N_s 个个体加入到抗体群中,若得到的抗体同时满足适应度阀值条件,输出此抗体为记忆库 M,得到优化的属性约简 R 集合.

Step7: 适应值不再提高,则终止计算,输出记忆库集合 M 即为优化的属性约简 R 集合.

4.2 算法可行性及收敛性分析

文献[11]表明,免疫克隆选择算法形同进化算法,属于有限齐次马尔可夫链,并证明了其收敛性.同样,基于聚类竞争的免疫克隆选择优化算法的整个聚类竞争及免疫操作过程状态变化均在有限空间中进行,种群序列$\{A(n), n \geq 0\}$是有限的. 由于

$$A(k+1) = T(A(k)) = T_c^C(A(k))\ T_s^C(A(k))\ T_g^C(A(k))\ T_d^C(A(k))$$

均与 n 无关,仅相邻状态 $A(k+1)$ 与 $A(k)$ 有关,可知$\{A(n), n \geq 1\}$是有限齐次马尔可夫链.

在上述算法中初始种群的规模为 n,初始种群中的全部近似解看成是状态空间 S 个个体,$s_i \in S$ 表示 S 中的第 i 个状态,V_k^i 表示随机变量 V 在第 k 代时所处的状态 s_i. 另外 f 是 X 上的适应度函数,表示为 $s' = \{x \in X \mid f(x) = \max f(x)\}$,则可定义算法的收敛性为

$$\lim_{k\to\infty} \sum_{s_i} p\{A_k^i\} = 1. \tag{10}$$

该定义表明,当算法迭代到足够多的次数后,群体中包含全局最佳个体的概率接近于 1,称之为算法收敛.

证明 设随机过程$\{A(k)\}$的转移概率为 $p_{ij}(k)$,且 $p_{ij}(k) = p\{A_{k+1}^j / A_k^i\} \geq 0$,记 $p\{A_k^i\}$ 为 p_i^k, $p_k = \sum_i p_i(k)$.

由马尔可夫链的性质可知

$$p_{k+1} = \sum_{s_i} \sum_j p_i(k)\ p_{ij}(k) = \sum_{i\in T} \sum_{j\in T} p_i(k)\ p_{ij}(k) + \sum_{i\notin T} \sum_{j\notin T} p_i(k)\ p_{ij}(k).$$

因为

$$\sum_{i\in T} \sum_{j\in T} p_i(k)\ p_{ij}(k) + \sum_{i\notin T} \sum_{j\in T} p_i(k)\ p_{ij}(k) = \sum_{i\in T} p_i(k) = p_k,$$

所以

$$\sum_{i\notin T} \sum_{j\notin T} p_i(k)\ p_{ij}(k) = p_k - \sum_{i\notin T} \sum_{j\in T} p_i(k)\ p_{ij}(k),$$

故 $0 \leq p_{k+1} \leq p_k - \sum_{i\notin T} \sum_{j\in T} p_i(k)\ p_{ij}(k) \leq p_k \leq 1$.

又 $\lim_{k\to\infty} p_k = 0$,则

$$1 \geq \lim_{k\to\infty} \sum_{s_i} p_i(k) \geq \lim_{k\to\infty} \sum_{i\in T} p_i(k) = 1 - \lim_{k\to\infty} p_k = 1.$$

可证明式(10)以概率 1 收敛. □

由以上证明可知,文中以属性核相对稳定信息作为免疫疫苗引入抗体编码,其采用的适应值函数可控制抗体核属性的免疫克隆选择优化朝着最小约简的方向进化收敛,得到的 R 为问题的最优解.

5 仿真实验

以工业环境下采集到的振动、声学征兆的滚动轴承技术诊断数据为例,其中:论域 $U = \{1, 2, \cdots, 16\}$,条件属性集 $C = \{s_1, s_2, \cdots, s_{12}\}$,决策属性 $D = \{0 \mid 1\}$. 表 1 为条件属性符号说明;表 2 为数据集,对其进行离散化处理时采用本文方法进行属性约简.

表 1 条件属性符号说明

s_1:	频率在 500 ～ 2 000 Hz,用 dB 表示的噪音值;
s_2:	频率在 4 ～ 16 kHz,用 dB 表示的噪音值;
s_3:	s_1 转换为用 mP 表示的值;
s_4:	s_2 转换为用 mP 表示的值;
s_5:	频率在 0.1 ～ 1 kHz,轴向上测量的振动加速度;
s_6:	频率在 1 ～ 11 kHz,轴向上测量的振动加速度;
s_7:	频率在 0.1 ～ 1 kHz,径向垂直方向上测量的振动加速度;
s_8:	频率在 1 ～ 11 kHz,径向垂直方向上测量的振动加速度;
s_9:	频率在 0.1 ～ 1 kHz,径向水平方向上测量的振动加速度;
s_{10}:	频率在 1 ～ 11 kHz,径向水平方向上测量的振动加速度;
s_{11}:	频率在 0.1 ～ 1 kHz 合成的振动加速度为 $\sqrt{s_5^2 + s_7^2 + s_9^2}$;
s_{12}:	频率在 1 ～ 11 kHz 合成的振动加速度为 $\sqrt{s_6^2 + s_8^2 + s_{10}^2}$;
D_0:	0 表示轴承的技术状态好;1 表示轴承的技术状态差.

免疫优化中选取进化代数 100,初始抗体群规模 $N = 100$,聚类中心数 $M = 15$,候选抗体群数量 $N_r = 100$,重组概率 P_i, $P_S = 0.9$,变异概率 $P_m = 0.1$,运行 20 次取平均值,得到实验结果如表 3 所示.

为验证数据集中属性约简所提取的完备性,算法中的属性约简利用波兰大学和挪威科技大学联合开发的 Rosetta 软件完成. 在实验中可知,随着属性的增加,需搜索的属性集数目也迅速增加,而遍历这些可能的属性集寻找最小约简,是一个十分费时的过程.

表 2　滚动轴承技术信息数据表

U	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	D
1	97.2	83.1	1.4	0.3	11	22	25	21	14	15	30.7	33.9	0
2	109.5	97.5	6	1.5	33	42	165	115	58	55	178	134.2	0
3	92.9	82.6	0.9	0.3	5.5	7.5	17	12.5	12.5	12.5	21.8	19.2	0
4	97.7	86.3	1.5	0.4	8.5	14.5	22.5	17	15	13	28.3	25.9	0
5	107.7	97.7	4.9	1.5	23	31	95	75	46	43	108	91.8	1
6	94.2	79.4	1	0.2	6.3	2.8	14.5	5.2	15.5	4.5	22.1	7.4	0
7	94.4	75.8	1	0.1	6.5	3.1	11.5	5.7	18	4.2	22.3	7.7	0
8	94.9	77.6	1.1	0.2	6	4.5	12.5	6.8	20.5	5.5	24.7	9.8	0
9	95.3	78.8	1.2	0.2	6.3	3.7	15	4.8	21	5.5	26.6	8.2	0
10	93	71.7	0.9	0.1	8.5	1.1	17	3.5	22	1.8	29.1	4.1	0
11	94.3	76.8	1	0.1	6.3	2.8	11	5.5	19.5	4.8	23.3	7.8	0
12	95.1	76.6	1.1	0.1	6.5	2.5	12	5.5	11.5	4.5	17.8	7.5	0
13	102.4	102.1	2.6	2.5	5	17	17	25.5	9.5	26.5	20.1	40.5	0
14	107.7	97.3	4.9	1.5	27	33	90	81	47	46	105.1	98.8	1
15	100	89	2	0.6	13	19	56	34	30	27	64.8	47.4	1
16	103	104	2.8	3.2	20	27	76	76	38	52	87.3	96	1

表 3　属性约简优化计算结果

迭代次数	最优个体	个体适应度	最优个体出现代数
1	100000111100	0.653 1	2
3	010001111000	0.742 6	3
6	000001111100	0.793 0	2
8	110011000000	0.822 2	4
12	110010000000	0.845 6	87

表 4　属性约简完备性比较

本文方法属性项	Rosetta 属性项
s_1 s_7 s_8 s_9 s_{10}	s_1 s_2 s_5 s_6 s_7 s_8 s_9 s_{10} s_{11} s_{12}
s_6 s_7 s_8 s_9 s_{10}	s_1 s_2 s_5 s_6 s_7 s_8 s_9 s_{10}
s_2 s_6 s_7 s_8 s_9	s_2 s_6 s_7 s_8 s_9
s_1 s_2 s_5 s_6	s_6 s_7 s_8 s_9 s_{10}
s_1 s_2 s_5	s_1 s_2 s_5

由表 3 和表 4 可知,基因免疫优化的粗糙集属性约简中,由于具有较快的全局收敛速度,在第 1 代中找到具有 5 个条件属性的约简,经过进一步迭代,系统可在第 12 代中找到最小约简$\{s_1\ \ s_2\ \ s_5\}$,个体最佳适应度值为 0.845 6. 所得最小约简结果与通过 Rosetta 软件计算一致,并在其他所得条件属性中更为直观且简单. 图 2 显示了属性约简免疫优化的种群变化曲线及个体适应度的变化情况.

为进一步检验算法的有效性,采用发表于“PopularScience”上的 CTR 数据集进行约简,并与文献[6,12]的典型值进行比较. 算法参数为:二进制编码,最大进化代数 50,初始抗体群规模 $N = 50$,

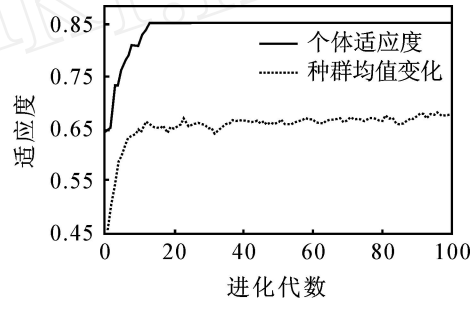

图 2　RS 约简免疫优化进化求解

聚类中心数 $M = 10$,重组概率 P_i, $P_S = 0.8$,变异概率 $P_m = 0.01$,运行 50 次取平均值得到实验结果如表 5 所示.

表 5　CTR 数据实验结果种群代数

方法	最优个体	平均代数
文献[6]	110110001	30
文献[12]	110110001	3
本文	110110001	2

文献[6]的遗传算法优化将近 30 代之后才出现最优种群;文献[12]采用可行域概念的遗传约简方法在约简速度上有了明显提高,平均 3 代便可求得最小约简,但其完备性不及本文方法. 实验结果表明,本文算法与基于遗传算法为代表的进化约简算法相比,对于大规模决策表属性约简而言,具有收敛速度快,属性约简完备,兼顾全局及局部搜索等特点.

6　结　　论

数据挖掘中的属性约简是粗糙集理论研究的核

心内容之一,由于属性约简中条件属性繁多、相容性及不唯一性,使得找出一个决策表的最小约简是NP难题.本文从免疫优化的角度出发,根据RS理论中核属性相对稳定的参数信息作为免疫疫苗优化抗体编码,并由条件属性和决策属性的近似分类质量给出了免疫适应度优化目标,提出了一种在优化初始群体基础上压缩属性的免疫克隆竞争选择算法.通过引入聚类竞争机制,加速了抗体亲和力的成熟,提高了全局搜索能力,同时也提高了抗体群的多样性.在加强局部及全局搜索能力的同时,保持了该算法快速收敛特性.仿真实验分析表明,该方法能有效地对多维条件规则集进行约简,在获得最简约简的同时保持了较好的约简完备性.但本文未能对约简前后的数据分析及挖掘质量展开进一步实验比较和性能评价,这是后续研究工作中的一项重要内容.

参考文献(References)

[1] Wong S K M, Ziarko W. On optimal decision rules in decision tables [J]. Bulletin of Polish Academy of Sciences, 1985, 32(11/12): 693-696.

[2] Kryszkiewicz M. Rough set approach to incomplete information systems [J]. Information Sciences, 1998, 112(1): 39-49.

[3] Tsang G C Y, Chen De-gang, Tsang E C C, et al. On attributes reduction with fuzzy rough sets[C]. IEEE Int Conf on Systems, Man and Cybernetics. Pasadena, 2005: 775-780.

[4] 王亚英, 张春慨, 邵惠鹤. 启发式知识约简算法的研究与应用[J]. 控制与决策, 2001, 16(6): 886-889.
(Wang Ya-ying, Zhang Chun-kai, Shao Hui-he. Research and application of heuristic algorithm for reduction of knowledge[J]. Control and Decision, 2001, 16(6): 886-889.)

[5] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.
(Miao Duo-qian, Wang Jue. An information representation of the concepts and operations in rough Set theory[J]. J of Software, 1999, 10(2): 113-116.)

[6] Dai Jian-hua, Li Yuan-xiang. Heuristic genetic algorithm for minimal reduction decision system based on rough set theory[C]. Int Conf on Machine Learning and Cybernetics. Beijing, 2002: 4-5.

[7] 梁霖, 徐光华. 基于克隆选择的粗糙集属性约简方法[J]. 西安交通大学学报, 2005, 39(11): 1231-1235.
(Liang Lin, Xu Guang-hua. Reduction of rough set attribute based on immune clone selection[J]. J of Xi'an Jiaotong University, 2005, 39(11):1231-1235.)

[8] Pawlak Z. Rough sets [J]. Int J of Computer Information Science, 1982, 11(5): 341-356.

[9] Burnet F M. The clonal selection theory of accquired immunity [M]. Cambridge: Cambridge University Press, 1959.

[10] De Castro L N, Von Zuben F J. Learning and optimization using the clonal selection principle [J]. IEEE Trans on Evolutionary Computation, Special Issue on Artificial Immune Systems, 2002, 6(3): 239-251.

[11] 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.
(Jiao Li-cheng, Du Hai-feng, Liu Fang, et al. Immunological computation for optimization, Learning and Recognition [M]. Beijing: Science Publisher, 2006.)

[12] 李订芳, 章文, 李贵斌, 等. 基于可行域的遗传约简算法[J]. 小型微型计算机系统, 2006, 27(2): 312-314.
(Li Ding-fang, Zhang Wen, Li Gui-bin, et al. Genetic reduction algorithm based on feasible region[J]. Mini-Micro Systems, 2006, 27(2): 312-314.)

[13] 张国云, 章兢, 向文江. 滚动轴承技术故障诊断的支持向量机方法研究[J]. 计算机工程与应用, 2005, 16(4): 227-229.
(Zhang Guo-yun, Zhang Jing, Xiang Wen-jiang. A novel SVM approach to the technique state diagnosis of the trundle bearing [J]. Computer Engineering and Applications, 2005, 16(4): 227-229.)

(上接第 496 页)

[9] Basile F, Chiacchio P, Giua A. Petri net monitor design with control and observation costs[C]. Proc IEEE the 39th Int Conf on Decision and Control. Sidney, 2000: 424-429.

[10] Basile F, Chiacchio P, Giua A. On the choice of suboptimal monitors for supervisory control of Petri nets[C]. Proc of the 1998 IEEE Int Conf on Systems, Man and Cybernetics. San Diego, 1998: 752-757.

[11] Basile F, Chiacchio P, Giua A. Suboptimal supervisory control of Petri nets in presence of uncontrollable transitions via monitor places[J]. Automatica, 2006, 42(6): 995-1004.