第 23 卷第 6 期
2006 年 12 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 23 No. 6
Dec. 2006

文章编号: 1000−8152(2006)06−0891−04

一种综合信息熵和遗传算法的知识约简方法

杨慧中, 王军霞, 丁　锋

(江南大学 控制科学与工程研究中心, 江苏 无锡 214122)

摘要: 针对粗糙集理论核心内容之一的知识约简问题, 本文结合信息论有关知识, 给出了粗糙集理论中一些概念和运算的信息表示, 并利用遗传算法作为约简工具, 提出了一种知识相对约简的方法. 为使所得约简相对最优, 将条件信息熵的重要性定义融入了适值函数中. 同时, 在适值函数的选取上引入了惩罚函数和罚系数, 从而保证所求的约简在包含最少而又非零个属性的基础上保持原有的分类能力. 通过实例分析可看出, 该算法对求解约简问题是快速有效的.

关键词: 粗糙集; 知识约简; 信息熵; 遗传算法

中图分类号: TP18　　**文献标识码**: A

Knowledge reduction approach based on information entropy and GA

YANG Hui-zhong, WANG Jun-xia, DING Feng

(Control Science and Engineering Research Center, Southern Yangtze University, Wuxi Jiangsu 214122, China)

Abstract: One essence of the rough-set theory is knowledge reduction. Computing the minimal knowledge reduction has been proved to be a NP hard problem. Firstly, an information representation of the concepts and operations of rough-set are presented. Secondly, a relative attribute reduction algorithm is developed via the information-entropy-based genetic algorithm (GA). Penalty function and coefficient are then used in fitness function to ensure fewer attributes while keeping consistency of knowledge-base classification in reduction. The significant definition of information entropy used in the fitness function can also make the reduction comparatively optimal. Finally, the experimental result shows that this approach can find the optimal relative attribute reduction effectively and rapidly.

Key words: rough set; knowledge reduction; information entropy; genetic algorithm

1　引言(Introduction)

粗糙集理论是由波兰学者Pawlak Z在1982年提出的[1], 是一种全新的刻划不完整性和不确定性问题的数学工具. 目前, 对粗糙集理论的研究多数是从粗糙集理论的代数观点出发的, 在代数表示下, 粗糙集理论的一些概念与运算的直观性较差, 因此一些学者提出了粗糙集理论的信息论观点[2].

“知识约简”被认为是粗糙集理论的精华, 其主要思想是, 在保持知识库分类能力不变的条件下, 删除其中不相关或不重要的知识, 导出问题的决策或分类规则. 在文献[3]中, 作者给出了约简的信息论观点包含代数观点这一证明, 并指出一个决策表在代数观点下的约简不一定能够保证约简后的信息熵不发生变化. 从信息论的角度讨论知识约简问题的方法大致有3种: 基于条件熵的CEBARKCC算法和CEBRKNC算法[3]以及基于互信息[4]的知识约简算法. 这3种算法在应用过程中会遇到待选属性的信息熵或互信息的值相等的情况, 这样选取哪个属性合适便影响了整个算法的速度. 因此, 针对上述问题, 本文提出了用遗传算法求解知识的最小约简. 将遗传算法用于粗糙集的信息论观点中, 既能避免在代数观点下直观性缺点, 也可以快速方便的搜索到最小约简.

2　粗糙集理论的信息论描述(Basic concepts of rough set information view)

设U为一个论域, 认为U上任一属性集合是定义在U上的子集组成的σ代数上的一个随机变量, 其概率分布可通过如下方法确定:

定义 1　设P, Q在U上导出的划分分别为

$$X = \{X_1, X_2, \cdots, X_n\}, Y = \{Y_1, Y_2, \cdots, Y_n\},$$

则P, Q在U上的子集组成的σ代数上的概率分布为

收稿日期: 2005−04−06; 收修改稿日期: 2006−03−07.
基金项目: 国家自然科学基金资助项目(60674092, 60574051).

$$\begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ p(X_1) & p(X_2) & \cdots & p(X_n) \end{bmatrix} = [X:p],$$

$$\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{bmatrix} = [Y:p].$$

其中: $p(X_i) = |X_i|/|U|, i = 1,2,\cdots,n; p(Y_j) = |Y_j|/|U|, j = 1,2,\cdots,m.$

定义 2 知识$Q(U/IND(Q) = \{Y_1, Y_2, \cdots, Y_m\})$相对于知识$P(U/IND(P) = \{X_1, X_2, \cdots, X_n\})$的条件熵$H(Q/P)$定义为

$$H(Q/P) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j/X_i) \log(p(Y_j/X_i)). \quad (1)$$

式中

$$\begin{cases} p(Y_j/X_i) = |Y_j \cap X_i|/|X_i|, \\ i = 1,2,\cdots,n,\ j = 1,2,\cdots,m. \end{cases} \quad (2)$$

定义 3 设$S = (U, A, V, f)$是一个信息系统, $A = C \cup D$, C为条件属性, D为决策属性, $P \subset C$,则对任意属性$a \in C - P$的相对决策属性D的重要性$SGF(a, P, D)$定义为

$$SGF(a, P, D) = H(D/P) - H(D/P \cup \{a\}). \quad (3)$$

定理 1 设U是一个论域, P是U的一个条件属性集合, D为决策属性, 若P中的一个属性r是P相对于决策属性D不必要的, 则

$$H(D/P) = H(D/P - \{r\}). \quad (4)$$

证 令

$$U/IND(P) = \{X_1, X_2, \cdots, X_n\},$$
$$U/IND(Q) = \{Y_1, Y_2, \cdots, Y_m\}.$$

不失一般性, 设

$$\text{pos}_p(D) = \{X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n\}.$$

由式(1)得

$$\begin{aligned} H(D/P) = & -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j/X_i) \log(p(Y_j/X_i)) = \\ & -\sum_{l=1}^{i-1} p(X_l) \sum_{j=1}^{m} p(Y_j/X_l) \log(p(Y_j/X_l)) - \\ & -p(X_i) \sum_{j=1}^{m} p(Y_j/X_i) \log(p(Y_j/X_i)) - \\ & -\sum_{l=i+1}^{n} p(X_l) \sum_{j=1}^{m} p(Y_j/X_l) \log(p(Y_j/X_l)) = \\ & -p(X_i) \sum_{j=1}^{m} p(Y_j/X_i) \log(p(Y_j/X_i)). \end{aligned}$$

若属性r是P相对于决策属性D不必要的, 则$\text{pos}_{p-\{r\}}(D) = \text{pos}_p(D)$, 故$H(D/P - \{r\}) = H(D/P)$.

值得注意的是该定理的逆不一定成立, 但当论域U是在P上相对于D一致时, 定理逆则存在.

定理 2 设U是一个论域, P是U的一个条件属性集合, D为决策属性, r为P中的一个属性, 若

$$H(D/P) \neq H(D/P - \{r\}), \quad (5)$$

则属性r是P相对于决策属性D必要的.

证 令

$$U/IND(P) = \{X_1, X_2, \cdots, X_n\},$$
$$U/IND(Q) = \{Y_1, Y_2, \cdots, Y_m\}.$$

不失一般性, 设

$$\text{pos}_p(D) = \{X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n\}.$$

因此

$$H(D/P) = -p(X_i) \sum_{j=1}^{m} p(Y_j/X_i) \log(p(Y_j/X_i)).$$

若$H(D/P) \neq H(D/P - \{r\})$, 则$\text{pos}_{p-\{r\}}(D) \neq \text{pos}_p(D)$必定成立, 故属性$r$是$P$相对于决策属性$D$必要的.

该定理的逆同样不存在, 当论域U是在P上相对于D一致时, 上定理逆则存在[3].

定理 3 设U是一个论域, P是U的一个条件属性集合, D为决策属性, 对于P中任意属性r, 若$H(D/P) \neq H(D/P - \{r\})$, 该条件属性$P$相对于决策属性$D$是独立的.

证 由定理2可得: P中的任意属性r相对于P是必要的, 因此, 条件属性P相对于决策属性D是独立的.

定理 4 设U是一个论域, P是U的一个条件属性集合, D为决策属性, 且论域U是在P上相对于D一致的, 属性$Q \subset P$, r是Q中的任一属性, 若属性$Q \subset P$是P相对于决策属性D的一个约简, 当且仅当以下条件满足:

$$H(D/P) = H(D/Q), \quad (6)$$
$$H(D/Q) \neq H(D/Q - \{r\}). \quad (7)$$

3 基于信息熵的遗传约简算法(GA reduction approach based on information entropy)

当决策表是一致的时候, 决策表约简计算的代数观点和信息论观点是等价的. 但是, 对于不一致决策表, 一个决策表在代数观点下的约简, 不一定能够保证约简后的信息熵不发生变化. 文献[3]给出了如果一个属性不能为另一个属性集合的分类增加任何信息就可将其进行约简这一证明, 因此定理4亦可用于

非一致决策表的约简计算.

3.1 适值函数的定义(Definition of fitness function)

适值函数是对个体位串的适应性进行评价的唯一确定指标, 从信息论的角度求解知识约简时, 就不得不考虑定义3中给出的重要性问题, 因此可以将适值函数定义为[5]

$$\text{Fit}=(\omega+\lambda)\cdot nnz(x)\cdot(1-\frac{\text{card}(x)}{n})/(1+\text{e}^{\tau(k_0-k)}). \tag{8}$$

其中: n表示条件属性的个数; $nnz(x)=\begin{cases}1,\ \text{card}(x)\neq 0\\ 0,\ \text{card}(x)=0\end{cases}$是防止个体(即条件属性个数)陷入全零的罚系数; card(x)表示每个个体x中1的个数, 即所选取的条件属性的个数, 可以看出个体中所含条件属性越少(但不能为零)适值越大; λ为某一较小给定值, 设$P\subset C$为相对核属性, 对于任意$a\in C-P$, 当$\max(SGF(a,P,D))$对应的属性a选中时(即对应的个体x中该位的值为1), λ取(0，0.5), 否则$\lambda=0$; $1/(1+\text{e}^{\varepsilon(k_0-k)})$为引入的惩罚函数, ω为惩罚因子, 一般取$\omega>1$; k_0取值为全部条件属性相对于决策属性的条件熵即$H(D/P)$, k为每个个体对应条件属性相对于决策属性的条件熵, 取$\tau>1$. 整个适值的定义目的就是使$k=k_0$, 而所含的条件属性数目尽可能的少, 同时在约简的过程中考虑到$SGF(a,P,D)$, 使得最终的约简能够达到最优.

3.2 遗传约简算法步骤(Steps of GA reduction approach)

步骤 1 计算所有条件属性C相对于决策属性D的条件熵k_0;

步骤 2 计算每个属性的重要性, 令$C'=\Phi$, 对于属性a, 若$H(D/C)\neq H(D/C-\{a\})$, 则$C'=C'\cup a$; 得到C'后, 计算C'相对于决策属性D的条件熵k', 若$k'=k_0$, 中止计算, 所得的C'即为相对最小约简; 若$k'\neq k_0$, 执行步骤3;

步骤 3 利用定义3计算出其他属性与相对核属性C'对决策属性的重要度;

步骤 4 随机产生m个长度为n(条件属性个数)的二进制串所代表的个体组成初始种群, 每个个体的每一位对应一个条件属性, 如取0表示不选择该属性, 取1则表示选择该属性; 对每个属性$a\in C$, 若$a\in C'$, 令该条件属性的对应位取值为1;

步骤 5 对每一个个体, 计算所含条件属性对决策属性D的条件熵k; 由适值公式计算出每个个体的适值, 找出适值最大的个体复制给下一代;

步骤 6 对于规模为m的种群, 计算每个个体的适值Fit$_i$及其被选择的概率$p_s(j)=\text{Fit}_j/\sum_{i=1}^{m}\text{Fit}_i, j=1,2,\cdots,m$, 从而计算出被选择的期望数$P_j=m^*p_s(j)$, 以轮盘赌的方式进行个体的选择, 构成新的种群;

步骤 7 根据交叉概率对种群进行交叉操作, 交叉运算采用一致交叉运算, 即染色体位串上的每一位按相同概率进行随机均匀交叉;

步骤 8 依据变异概率对种群进行变异操作, 变异运算是通过按变异概率随机反转某位等位基因的二进制字符值来实现;

步骤 9 对每一个个体, 计算所含条件属性对决策属性D的条件熵k; 依据适值公式计算出每个个体的适值;

步骤 10 判断遗传算法是否成熟(连续n代的最优个体适应值不再提高), 如果最优个体所对应的条件属性的条件熵$k=k_0$, 检验该属性集合的独立性, 若以上两条件皆具备, 则停止运算, 否则转至步骤6.

4 仿真实例(Case study)

文献[6]提出的信息系统为电力系统中的故障诊断模型, 该信息系统中论域U包含27个对象, 即27种故障类型, 决策属性:

$$D=\{\text{Fault}\},$$

$$\begin{aligned}U/D=\{&\{1,12,13\},\{2,10,15\},\{3,14\},\{4,6,7,16\},\\&\{5,8,9,11\},\{17\},\{18\},\{19\},\{20\},\{21\},\\&\{22\},\{23\},\{24\},\{25\},\{26\},\{27\}\};\end{aligned}$$

条件属性P为报警模式,

$$\begin{aligned}P=\{&CB_1,CB_2,CB_3,CB_4,A_m,B_m,C_m,L_1A_m,\\&L_1B_m,L_2B_m,L_2C_m,L_1A_p,L_1B_p,L_2B_p,\\&L_2C_p,L_1A_s,L_1B_s,L_2B_s,L_2C_s\},\end{aligned}$$

$$\begin{aligned}U/P=\{&\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\\&\{9\},\{10\},\{11\},\{12\},\{13\},\{14\},\{15\},\\&\{16\},\{17\},\{18\},\{19\},\{20\},\{21\},\{22\},\\&\{23\},\{24\},\{25\},\{26\},\{27\}\},\end{aligned}$$

$$H(D/P)=0.$$

表1是按照该约简算法求得的结果, 计算中各参数取值为: 种群规模$m=20$, 交叉概率$p_{\text{c}}=0.8$, 变异概率$p_{\text{m}}=0.05,\bar{\omega}=3,\tau=30$. 该信息系统的条件属性相对核为$\{A_m,C_m\}$. 结果显示了每一代的最优个体及适应值. 在本例中种群在第14代就出现最

优个体, 连续22代保持不变. 对应的约简属性为$Q = \{CB_1, CB_2, CB_3, CB_4, A_m, C_m, L_2B_m\}$, 经检验该属性相对于决策属性是独立的, 且$H(D/Q) = H(D/P) = 0$, 由定理6可知该约简为最小约简. 由此, 在保持分类能力不变的情况下, 寻找故障原因的范围由原来的19个减少到7个, 去除了12个冗余属性, 如果没有漏报现象, 产生各种故障的原因可以轻松排查, 故障诊断能力大为提高.

表 1 遗传约简计算结果

Table 1 Results of GA reduction approach

进化代数	最优个体	适应值
1	1110111101110000000	0.5803
2	1110101001101000000	0.8289
3	1110101001101000000	0.8289
4	1110101001101000000	0.8289
5	1110101001100000000	0.9119
6	1110101001100000000	0.9119
7	1110101001100000000	0.9119
8	1110101001100000000	0.9119
9	1110101001100000000	0.9119
10	1110101001100000000	0.9119
11	1110101001100000000	0.9119
12	1110101001100000000	0.9119
13	1111101001000000000	0.9646
14	1111101001000000000	0.9646
15	1111101001000000000	0.9646
16	1111101001000000000	0.9646
17	1111101001000000000	0.9646
18	1111101001000000000	0.9646
19	1111101001000000000	0.9646
20	1111101001000000000	0.9646
21	1111101001000000000	0.9646
22	1111101001000000000	0.9646
23	1111101001000000000	0.9646
24	1111101001000000000	0.9646
25	1111101001000000000	0.9646
26	1111101001000000000	0.9646
27	1111101001000000000	0.9646
28	1111101001000000000	0.9646
29	1111101001000000000	0.9646
30	1111101001000000000	0.9646
31	1111101001000000000	0.9646
32	1111101001000000000	0.9646
33	1111101001000000000	0.9646
34	1111101001000000000	0.9646
35	1111101001000000000	0.9646

5 结束语(Conclusion)

本文从粗糙集的信息论观点出发, 讨论了粗糙集理论中一些概念和运算的信息表示. 利用信息熵对优化计算所提供的启发信息以及遗传算法具有的全局优化和隐含并行性的优点, 给出了一种知识约简算法, 并为求解基于信息熵的粗糙集知识约简提供了理论依据, 通过实例分析可看出, 该算法在求解粗糙集理论知识约简的问题上是快速有效的. 但是, 该算法在求解知识约简问题只能达到相对的最优, 如何给出更有效的求解最优约简方法是下一步的工作目标.

参考文献(References):

[1] PAWLAK Z. Rough sets[J]. *Int J of Information and Computer Science*, 1982, 11(5): 341 – 356.

[2] 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 Duoqian, WANG Jue. An information representation of the concepts and operations in rough set theory[J]. *J of Software*, 1999，10(2): 113 – 116.)

[3] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759 – 766.
(WANG Huoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy[J]. *Chinese J of Computers*, 2002 ,25(7): 759 – 766.)

[4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681 – 684.
(MIAO Duoqian, HU Guirong. A heuristic algorithm for reduction of knowledge[J]. *J of Computer Research & Development*, 1999, 36(6): 681 – 684.)

[5] 陶志, 许宝栋, 汪定伟, 等. 基于遗传算法的粗糙集知识约简方法[J]. 系统工程, 2003，21(4): 117 – 122.
(TAO Zhi, XU Baodong, WANG Dingwei, et al. Rouch Set knowledge reduction approach based on GA[J]. *Systems Engineering*, 2003，21(4): 117 – 122.)

[6] ZHANG Qi, HAN Zhenxiang, WEN Fushuan. A new approach for fault diagnosis in power systems based on rough set theory[C] *// Proc of the 4th Int Conf on Advances in Power System Control, Operation and Management(APSCOM-97)*. Hong Kong: IEE Conference Publisher, 1997, 11: 597 – 602.

作者简介:

杨慧中 (1955—), 女, 工学博士, 教授，博士生导师, 主要从事过程建模与优化控制研究, E-mail: yhz_jn@163.com;

王军霞 (1979—), 女, 硕士研究生, 主要研究方向为粗糙集理论、数据融合, E-mail: amendaxia@yahoo.com.cn;

丁 锋 (1963—), 男, 教授, 工学博士, “太湖学者”特聘教授, 主要从事模型与辨识、过程控制、多率系统与自适应控制方面的研究.