

Based on Advanced Genetic Algorithm and Rough Set of Attribute Reduction

ZHANG Yangzi, MU Xiaodong, ZHANG Jinwei

The Second Artillery Engineering College, Xi'an, China

Abstract: To availablely acquire an attribute of minimum opposite reduction, an algorithm based on advanced genetic algorithm and rough set of attribute reduction is projected. The selection operator of the algorithm put in the optimal retention policy, raise the evolution speed; Joined an adaptive control to the genetic and metamorphosis operator, keep the varied of the solution, improved the overall searching ability. Finally the arithmetic is application to select the evaluation radiation dosage characteristic of the special personnel, the simulation result shows: the arithmetic can effective solve an attribute of minimum opposite reduction.

Keywords: rough set; attribute reduction; advanced genetic algorithm

基于改进遗传算法和粗糙集的属性约简

张阳子, 慕晓冬, 张金巍

第二炮兵工程学院 指挥自动化系, 西安, 中国, 710025

【摘要】为了有效地获得属性的最小相对约简, 提出了一种基于改进遗传算法 粗糙集的属性约简算法。该算法对选择算子加入了最优保存策略, 提高了进化速度; 在遗传算子和变异算子中加入了自适应控制, 保存解的多样性, 提高了全局搜索能力。最后将本算法应用于特种作业人员受照剂量评价特征属性的选取, 仿真结果表明: 该算法能够有效地解决属性的最小相对约简。

【关键词】粗糙集; 属性约简; 改进遗传算法

1 引言

粗糙集理论是波兰数学家 Z.Pawlak 教授于 1982 年提出的一种处理不精确与不完全数据的数学理论, 其主要特点是不需要预先给定某些特征或属性的数量描述, 而是直接从给定问题的分类知识出发, 通过不可分辨关系确定给定问题的知识约简。目前, 粗糙集理论同神经网络、模糊理论、遗传算法等结合, 已被广泛应用于人工智能、模式识别和数据挖潜等各种应用领域。属性约简是粗糙集理论中的一个重要研究课题。属性约简就是保持知识库分类或决策能力不变的条件下, 删除不相关或不重要的属性【1、2】。

遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种全局优化搜索算法。其主要特点是群体搜索策略和群体中个体之间的信息交换, 搜索不依赖梯度信息, 是一种有效的全局优化算法, 因此遗传算法有很强的全局搜索能力, 能够在很短的时间内找到全局最优解【4、5】。

本文设计出一种基于改进遗传算法的粗糙集约简算法。在进化过程中每代最优个体不进行交叉变异而是直接进入下一代, 又通过对交叉概率和变异概率的自适应设计, 进而获得遗传约简算法, 该算法可大大降低约简时间, 提高计算

的效率。

2 粗糙集基本概念

定义 1 称四元有序组 $S = (U, A, V, f)$ 为信息系统, 其中 U 为所考虑对象的非空有限集合, 称为论域; A 为属性的非空有限集合; V_a 为属性 a 的值域; $f: U \times A \rightarrow V$ 是一个信息函数, $\forall x \in U, a \in A, f(x, a) \in V_a$, 对于给定对象 x , $f(x, a)$ 赋予对象 x 在属性 a 下的属性值。信息系统也可简记为 $S = (U, A)$ 。

定义 2 设 $S = (U, A, V, f)$ 为信息系统, 若 A 可划分为条件属性集 C 和决策属性 D , 即 $C \cup D = A$, $C \cap D = \emptyset$, 则称信息系统为决策表。

定义 3 设 R 为一个等价关系簇, $P \in R$, 如果 $IND(R) = IND(R - \{P\})$, 则称 P 为等价关系族 R 中不必要的; 否则称 P 为 R 中必要的。如果 R 中任意一个等价关系 P 都是必要的, 则称 R 为独立的, 否则称 R 为依赖的。

定义 4 设 $Q \subseteq P$, 若 Q 是独立的, 且 $IND(Q) = IND(P)$, 则称 Q 是等价关系族 P 的一个约简 (Reduct)。 P 中所有必要关系的集合称为等价关系族 P 的核 (Core), 记作 $CORE(P)$ 。

定义 5 信息系统 $S = (U, A, V, f)$, $A = C \cup D$, $C \cap D = \emptyset$, $B \subset C$, 则

$$k(B, D) = \frac{|Pos_B(U|_{ind(D)})|}{|U|}$$

表示条件属性对决策属性的依赖度。

定义 6 信息系统 $S = (U, A, V, f)$, $A = C \cup D$, $C \cap D = \emptyset$, 则任意属性 a 在属性集 C 中的重要性定义为:

$$SGF(a, C) = k(C, D) - K(C - \{a\}, D)$$

3 基于改进遗传算法的粗糙集属性约简法

3.1 编码方法

由于遗传算法不能直接处理解空间的解数据, 因此必需通过编码将它们表示成遗传空间的基因型串结构数据。本文使用固定长度的二进制符号串来表示群体中的个体, 其等位基因由二值符号集 $\{0, 1\}$ 组成, 其中每一位对应一个条件属性, 若某位取值为 1, 则表示选择其对应的条件属性; 若某位取值为 0, 则表示不选择其对应的条件属性。初始种群中各个体的基因可用均匀分布的随机数生成。

3.2 适应度函数的选取

遗传算法在求解过程中是以适应度函数为依据的, 利用种群中每个个体的适应度值来进行搜索。因此, 适应度函数的选取至关重要, 它直接影响到遗传算法的收敛速度, 以及最终能否得到问题的最优解。

定义适应度函数【6】如下:

$$F(x) = (1 - \text{card}(x)) / n + k$$

其中: $\text{card}(x)$ 为染色体中 1 的个数, 即染色体所含条件属性的个数; n 为染色体的长度, 即条件属性的个数; k 为决策属性对染色体所含的条件属性的依赖度。该函数可以控制染色体朝着最小约简的方向进化: k 越大, 说明决策属性 D 依赖条件属性 C 越强, 当 $k = 1$ 时, 决策信息完全由条件信息确定; 通过 $\text{card}(x)$ 来控制染色体所含条件属性的个数。通过这两方面构成的适应度函数, 可以在保持决策属性对整体条件属性依赖度不变的情况下, 找到所含条件属性最少的约简。

3.3 遗传操作的设计

3.3.1 选择算子

选择算子是指在上一代群体中按照某些指标挑选出的、参与繁殖下一代群体的一定数量的个体的一种机制。本文采用适应度比例法和最优保存策略相结合的方法选择。

适应度比例法也称为轮盘赌的形式选择。其具体操作过程如下:

(1) 计算出群体中所有个体的适应度总和;

(2) 计算出每个个体的相对适应度, 即各个个体被遗传到下一代群体中的概率:

(3) 使用模拟赌盘操作 (即做 0~1 之间的随机数) 来确定各个个体被选中的次数。

最优保存策略: 在遗传算法的运行过程中, 会产生一些优良个体。如果按照传统的选择方法, 他们的优良模式有可能被后面的遗传操作破坏, 就会降低群体的平均适应度, 这样对于进化而言是不好的, 所以我们就把适应度最高的个体保留到下一代群体中, 不使其参与遗传操作, 并且用它来代替本代群体中经过遗传操作后所产生的适应度最低的个体【4】。其具体操作过程为:

(1) 找出当前群体中适应度最高和最低的个体的集合。

(2) 若当代群体中的个体的适应度比迄今为止最好个体的适应度高, 则以当代群体中的最佳个体作为新的迄今为止的最好个体。

(3) 用迄今为止的最好个体替换掉当代群体中的最差个体。

最优保存策略可视为选择操作的一部分, 是遗传算法收敛性的一个重要保证条件。但若此最优个体是局部最优的就会使算法的全局搜索能力不强, 因此本文采用最优保存策略与适应度比例法配合使用。

另外, 推广之, 即在每一代中保留多个较优个体不参与遗传操作而直接将它们放到下一代群体中, 这种选择方法也称为稳态复制。本文就结合这种策略进行改进。

3.3.2 交叉算子

交叉算子是指对选择后的父代个体进行基因模式的重组而产生后代个体的繁殖机制。在个体繁殖过程中, 交叉能引起基因模式的重组, 从而有可能产生含优良性能的基因模式的个体。目前有多种交叉方式, 其中较为简单的两种为一点交叉和两点交叉。本文采用两点交叉, 其方法如下:

对群体中的个体进行两两随机配对, 对每一对相互配对的个体, 随机设置某两个不同基因座为交叉点, 对每一对相互配对的个体, 依交叉概率 P_c 相互交换两个个体在两个交叉点之间的部分染色体, 从而产生出两个新的个体。

3.3.3 变异算子

变异是指模拟生物在自然的遗传环境中由于某种偶然因素引起的基因模式突变的个体繁殖方式。在变异算子中, 以变异概率 P_m 在群体中选取个体, 随机选择个体的某个基因进行基因取反从而产生新的个体。在遗传算法中, 变异算子增加了群体中基因模式的多样性, 从而增加了群体进化过程中自然选择的作用, 避免早熟现象的出现。

3.4 交叉概率和变异概率的自适应确定

3.4.1 交叉概率的自适应确定

交叉算子在遗传操作中起核心作用，主要用来产生新个体，实现算法的全局搜索能力。从群体整体进化过程来看，交叉概率应该能随进化过程逐渐变小，到最后趋于某一稳定值，以避免对算法后期的稳定性造成冲击而导致算法不能收敛，或收敛过程加长；而从产生新个体的角度来看，群体中的所有个体在交叉操作上应该具有同等地位，相同概率，从而使GA在搜索空间具有各个方向的匀性。本文为此设计与进化代数相关的交叉概率计算公式：

$$P_c = \frac{1}{1 + e^{\alpha G}} + \beta$$

其中， G 为进化代数， α, β 为定常系数， α 代表交叉概率的变化曲率， β 代表交叉概率的收敛极限。

3.4.2 变异概率的自适应确定

变异算子在遗传操作中起辅助作用，主要用来维持群体多样性，防止出现未成熟收敛。在算法早期，群体中个体多样性丰富，此时的变异概率应该小些，以提高算法的运行速度；而随着进化的进行，个体越来越向适应度高的个体靠近，致使个体越来越单一，此时的变异概率就应该大些，以维持群体的多样性。同样的原因，同一代群体中个体的变异概率应该随个体的优劣而变化，即加大优质个体变异概率。为此设计了如下的与遗传进化代数和个体适应度相关的自适应变异概率：

$$P_m = \begin{cases} \frac{k_1}{1 + e^{-\alpha G}} - \frac{f - \bar{f}}{f_{\max} - \bar{f}}, & f > \bar{f} \\ k_2, & f \leq \bar{f} \end{cases}$$

其中， G 为进化代数， α, k_1, k_2 为定常系数。 α 代表变异概率的变化速度； k_1 与具体问题有关，是为保证遗传算法不退化为随机搜索， P_m 所能取到的最大值； k_2 为一个比较小的变异概率，一般取 0.01。

4 遗传算法属性约简过程

由于粗糙集属性约简没有固定的结束条件，即不存在理想理解的模型，故无法给出具体的终止函数。但可以设想如果连续几代的种群的平均适应度的差值不超过一个给定的阈值，即可认为已经得到了一定程度的属性约简，结束遗传算法。

通过上述分析，可以得到遗传算法在粗糙集属性约简中的工作流程，如图一所示：

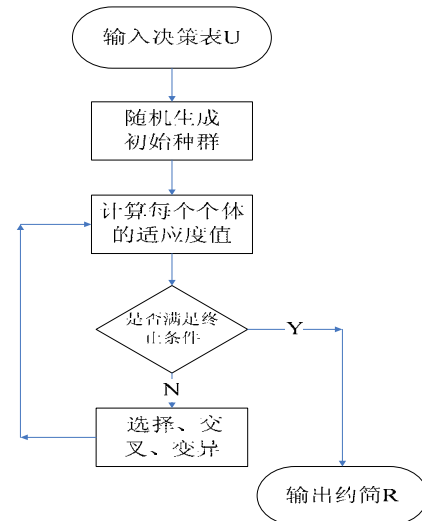


图 1. 基于遗传算法和粗糙集的属性约简工作流程

5 算例

为了考察本算法的有效性，使用本文算法对特种作业人员受照剂量评价特征属性选取进行了研究。采用基于遗传算法和粗糙集的属性约简算法对条件属性集进行属性约简，最后得到特征属性的集合。

在本试验中，条件属性集的选取来自于某单位2007年特种作业人员的受照剂量(内照射、外照射)监测数据表、个人基本资料表和任务基本情况表中的属性，根据经验和已获取的数据选取可能对作业人员的受照剂量有影响的11个属性作为条件属性集 C ，即 $C = \{a, b, c, d, e, f, g, h, i, j, k\}$ 。根据对特种作业人员受照状况的评价作为决策属性集 D 构建初始决策表。

初始决策表经过预处理后形成最终决策表，最终决策表如表1所示。其中： a 表示执行核武器操作任务的年限， b 表示人员级别， c 表示人员类别， d 表示任务描述， e 表示本次任务中的累计受照时间， f 表示核素种类， g 表示核素摄入方式， h 表示尿氚测量结果， i 表示尿钍测量结果， j 表示尿钚测量结果， k 表示外照射个人剂量监测结果。

根据上文所述基于改进遗传算法和粗糙集的属性约简算法，利用VC++6.0编程实践。实验结果表明：经过23代就能达到最优结果，以后均不再变化；用标准遗传算法要经过86代才能达到最优，最优结果一样。最后得出的约简集 R 为， $\{c, h, I, k, f, g\}$ 所以 R 是条件属性集 $C = \{a, b, c, d, e, f, g, h, i, j, k\}$ 的最小约简。从结果可以看出，在对特种作业人员的受照状况进行评价时，可以将作业人员的人员类别、尿氚测量结果、尿钍测量结果、外照射监测结果、核素种类和核素摄入方式六个属性作为主要的特征属性，通过这六个属性的属性值对作业人员的受照状况作出评价。

6 结论

属性约简是粗糙集理论研究的核心内容之一，但是找出一个决策表的最小约简是 NP 难问题。本文提出了一种基

U	a	b	c	d	e	f	g	h	i	j	k	des
1	1	2	1	*	1	*	*	*	1	*	2	0
2	1	2	*	1	1	1	1	1	*	*	2	0
3	1	*	1	*	1	*	*	*	1	*	3	0
4	1	3	*	1		1	1	1	*	*	3	0
5	1	2	1	*	1	*	*	*	1	*	1	0
6	1	2	*	1	1	1	1	1	*	*	1	0
7	2	2	*	*	1	*	1	1	1	*	1	0
8	1	*	2	*	1	*	1	2	*	*	1	1
9	1	2	2	3	1	2	1	*	1	*	1	1
10	1	3	2	*	1	*	1	2	*	*	3	1
11	2	3	2	2	1	1	1	2	*	*	2	1
12	2	2	3	2	*	2	*	*	1	*	2	1
13	2	2	3	2	*	3	*	*	1	*	2	1
14	1	3	2	3	1	2	1	*	1	*	3	1
15	1	2	2	*	1	*	1	2	*	*	2	1
16	1	2	2	3	1	2	1	*	1	*	2	1
17	2	*	3	2	*	4	*	2	*	2	1	2
18	1	2	*	2	*	2	3	*	2	*	1	2
19	3	2	3	2	1	2	1	*	2	*	1	2
20	2	*	3	2	*	4	*	1	*	2	3	2
21	2	2	3	2	*	3	*	*	2	*	2	2
22	2	3	3	2	1	2	1	*	2	*	3	2
23	2	*	3	2	*	4	*	1	*	2	2	2
24	2	2	3	2	1	2	1	*	2	*	2	2

表 1. 最终决策表

于改进遗传算法的粗糙集属性约简算法,并在选择算子中加入最优保存策略,加快了进化速度;在交叉算子和变异算子中对交叉概率和变异概率进行了改进,提高了全局搜索能力,并保证算法收敛。最后将本算法应用于特种作业人员受照剂量评价特征属性选取,实验结果表明了该算法的有效性。

通过和标准遗传算法相比可以看出,该算法能够有效地对决策表进行约简,无论在约简的准确性上,还是平均运行代数上都取得了较好的结果,特别是数据规模较大时更加节省计算时间。

Reference (参考文献)

- [1] Pawlak Z. Rough sets: Theoretical aspects of reasoning about data. Boston: Kluwer Academic Publishers, 1991.
- [2] 张文修, 吴伟志, 梁吉业, 李德玉. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001: 1-2.
- [3] 邹艳. 粗糙集理论在不完备信息系统知识获取中的应用研究[重庆师范大学硕士学位论文]. 重庆: 重庆师范大学, 2006.3.
- [4] 赵云珍. 遗传算法及其改进[昆明理工大学硕士学位论文]. 昆明: 昆明理工大学, 2005.1.
- [5] 王杨. 基于小生境遗传算法的粗糙集属性约简方法. 计算机工程, 2008.3.
- [6] 潘伟, 王云峰, 伞冶. 基于自适应遗传算法的粗糙集知识约简算法. 计算机工程与应用, 2008, 44 (15) .