DOI:10.15923/j.cnki.cn22-1382/t.2019.1.12

遗传算法下的粗糙集属性约简算法及其有效性分析

郑文彬[1,2]， 胡敏杰[1,2*]， 何秋红[1,2]

(1.闽南师范大学 计算机学院，福建 漳州 363000；
2.闽南师范大学 福建省粒计算及其应用重点实验室，福建 漳州 363000)

摘 要：将核引入遗传算法初始群体，根据决策属性对条件属性的依赖度来加强局部搜索能力，并保证全局寻优，得到最佳搜索效果。

关键词：遗传算法；粗糙集属性；约简算法；有效性

中图分类号：TP 312 **文献标志码**：A **文章编号**：1674-1374(2019)01-0066-06

Rough set attribute reduction algorithm based on genetic algorithm and its validity analysis

ZHENG Wenbin[1,2]， HU Minjie[1,2*]， HE Qiuhong[1,2]

(1.School of Computer Science, Minnan Normal University, Zhangzhou 363000, China;
2.Lab of Granular Computing, Minnan Normal University, Zhangzhou 363000, China)

Abstract: A kernel is introduced into the initial group of genetic algorithm. The dependence of the decision attribute on conditional attribute is used to strengthen the local search ability to ensure the global optimization and obtain the optimal search results.

Key words: genetic algorithm; rough set attribute; reduction algorithm; validity.

0 引 言

粗糙集理论属于数据挖掘方法中的高效方法，也是全新强有力对不确定性信息数学工具处理的方法。不确定性信息处理指的是对于不完整、模糊及不精准信息和组合信息实现处理，其被广泛应用到机器学习、人工智能、故障诊断、模式识别及数据分析挖掘中。使用粗糙集约简，能够选择条件属性集，将条件属性和决策不相关删除，使用条件属性约简集替代原本属性集。对数据量较大、属性维度较高的信息系统，在人们可接受时间及具有有效资源背景下，根据遍历及枚举的方法无法得到最小属性约简，人们一般只能够得到属性近似约简，粗糙集属性约简算法为目前高效

收稿日期：2018-07-25
基金项目：福建省教育厅科技项目(JAT170347，JAT170350)
作者简介：郑文彬(1971—)，男，汉族，福建仙游人，闽南师范大学高级讲师，硕士，主要从事数据挖掘、粗糙集及其应用方向研究，E-mail:361172697@qq.com. *通讯作者：胡敏杰(1979—)，女，汉族，湖北武汉人，闽南师范大学副教授，硕士，主要从事数据挖掘方向研究，E-mail:396484776@qq.com.

方法，所以使属性约简算法效率提高，对粗糙集属性约简来说尤为重要。

1　粗糙集的基本理论

粗糙集理论指的是对不确定性处理的数学工具，也是全新软计算方法。目前，粗糙集备受人们的重视，其有效性已经被证实，属于现代国际中人工智能理论和应用领域中的研究热点。在多种实际系统中都具有不同程度的不确定性因素，收集数据常常具有不完整、不确定及噪声，所以需要对其进行处理。粗糙集理论使知识理解划分成为数据，每个被划分的集合就是概念。粗糙集理论思想就是使用已知知识库，使不确定及不精准知识使用已知知识库刻画，此理论和其他处理不精准及不确定问题理论的主要区别就是其不需要提供问题需要的数据集合之外信息，所以在处理问题不确定描述过程中较为客观，因为此理论没有处理不确定及不精准原始数据的机制，那么此理论及概率论等处理具有一定的互补性[1]。粗糙集理论的主要定义为：

1)决策表。决策表 S 指的是四元组 $S\leqslant U, R, V, F>$，其中的 U 指的是非空有限对象集，称之为论域；$R=C\cup D$ 的属性集合，C 指的是条件属性集，D 指的是结果属性集，V 指的是属性值集合。F 指的是信息函数，属于 U 中每个对象 x 的属性值。

2)不可分辨关系。在决策表 S 中，对每个属性自己定义成为不可分辨关系，也就是：

$$\mathrm{IND}(B)=\{(x,y)\mid (x,y)\in U2,\ \forall b\in B(b(x)=b(y))\}$$

3)正域。假设 U 属于论域，P 与 Q 指的是 U 中的两个等价关系簇。

4)下近似集。对于每个概念中的 X 与不可分辨关系，包括在 X 中的最大可定义集都是以 B 进行确定的[2]。

2　遗传算法分析

遗传算法是将达尔文所提出的生物进化论和孟德尔提出的遗传学理论，对自然界生物从低级到高级进行模拟的高级净化过程，将初始种群作为起点，使用适者生存自然法则对个体进行选择，并且使用变异、交叉等策略产生下一代种群，逐渐进化到满足期望条件。遗传算法将净化思想作为基础，常用来对复杂优化问题进行解决。在遗传算法不断完善及发展的过程中，算法效率在不断的提高。并且遗传算法自身具备开放性，能够使其和其他算法相互融合，以此提高算法的效率[3]。

2.1　种群和个体

种群为遗传算法中求解问题解空间子集，种群中全部元素都属于个体，其在迭代过程中在不断的发生变化，但是种群个体数量不会发生变化。

2.2　编码

编码指的是将需要优化的问题朝着遗传算法容易处理的方式进行转变，遗传算法性能和编码方式具有一定的联系，所以选择合适求解问题编码方式是算法设计的主要内容，常见编码方式包括树型编码、二进制编码、自适应编码及实数编码。

2.3　选择

选择指的是以个体适应度值的优劣程度对种群个体进行选择，也就是以一定概率 Pr 从上一代种群中实现个体选择，之后进行操作。一般选择方法包括随机便利、轮盘赌和排序选择等。遗传算法中最早的选择策略就是轮盘赌，此方法使种群中全部个体适应度的和作为轮盘，每个个体和轮盘中的某个区域进行对应，个体适应度越高，那么占比就会越高[4]。

2.4　适应度函数

适应度指的是种群个体对于环境适应程度，此指标主要是对种群中个体优劣程度进行描述。适应度函数主要指的是个体在进化计算过程中的最优解程度，遗传算法在搜索过程中不利用外部信息评价，以此导致适应度成为种群个体评价的主要标准，所以选择适应度函数和设计对遗传算法具有一定的影响。

2.5　交叉

交叉操作指的是以一定概率 Pc 从种群种选择种群个体构成配对，之后将其基因串某部分实现交叉，以此产生全新种群个体过程。交叉操作不仅保持原本种群优良个体特点，并且还使算法能够对全新基因空间进行搜索，使全新种群个体具备多样性。二进制编码大部分都是利用单点交叉策略。

假设每个个体都具有八位二进制表达，其中的两个个体分别为 F1 = 11100111，F2 = 10011010，假如交叉位置为 3，那么个体低三位交换，得到全新个体：R1 = 11100010，R2 =

10011111。在进行个体交叉操作的过程中，对每次个体交叉操作概率进行控制[5]。

3 基于遗传算法的粗糙集属性约简算法

3.1 遗传约简算法

在决策问题的过程中，寻找最小相对约简具有重要的作用，结合遗传算法和粗糙集，效果良好。

3.1.1 编码方式

因为遗传算法无法对空间解数据进行直接处理，所以利用编码使其转变成为遗传空间基因型串结构数据。利用固定长度二进制符号对群体个体进行表示，等位基因通过二值符号集{0,1}构成。初始群体个体基因使用均匀发布随机数表示，比如 100111001000011100 就为个体，此个体染色体长度为 $n=18$，其中的每位都与条件属性相互对应。如果取值为 1，那么其指的是选择某个对应条件属性，如果取值为 0，那么就表示不对相应条件属性进行表示[6]。

3.1.2 个体适应度评价

将适应度函数定义成为：

$$F(x)=\frac{1-\mathrm{card}(x)}{n}+k$$

式中：card(x)——染色体中 1 的数量，也就是染色体中条件属性的数量；

n ——染色体长度，也就是条件属性数量；

k ——决策属性对于染色体条件属性的依赖度。此函数能够对染色体控制最小约简方向：k 越大，表示决策属性 D 对于属性 C 依赖程度就会越强；在 k 为 1 的时候，决策信息通过条件信息进行确定。

利用 card 对染色体中条件属性长度进行控制，以此所创建的适应度函数不仅能够保证决策属性对于整体条件属性依赖度不改变，还能够寻找具有条件属性小的约简。

3.1.3 选择操作

利用适应度比例选择方法，从目前群体中选择优良个体，将其到下一代群体中进行复制。具体流程为：

1)对群体中全部个体适应度总和进行计算；

2)对个体相对适应度大小进行计算，也就是个体到下一代群体遗传的概率；

3)利用赌盘操作模拟对个体被选中数量进行确定[7]。

3.1.4 交叉操作

使用单点交叉算子进行执行。对群体个体实现两两随机配对，对每对相互配对个体随机实现交叉点的设置，对每对配对个体根据假设的交叉概率 Pc 在交叉点中相互交换个体部分染色体，以此产生全新个体。

3.1.5 变异操作

利用变异算法使个体基因根据编译概率指定变异点，使每个指定变异点中的属性不变异，对其基因值进行取反运算，以此产生全新个体。

3.1.6 最优保存

在得到全新个体以后，假如最坏个体适应值比上一代最好个体适应值要小，那么上一代最好个体替代最新最坏个体，此方法能够保证算法收敛。

3.2 基于遗传算法粗糙集属性约简算法

因为遗传算法无法实现理解空间解数据直接处理，所以就要利用编码使其转变成为遗传空间中基因型串结构数据。文中利用固定长度二进制符号串表示群体个体，等位基因通过二值符号集构成。初始群体中的个体基因值使用均值分布随机数生成，比如 1001100 就是个体，其中的每位对应条件属性。如果值为 1，那么对相应条件属性选择；假如值为 0，那么表示不对相应条件属性选择，以上个体相应属性为{$c1$,$c4$,$c5$}[8]。

3.2.1 基于区分矩阵粗糙集约简算法

因为二进制区分矩阵核和基于正区域核两者定义不同，其不仅能够在决策表中使用，也能够应用到不相容决策表中。

对给定信息系统 $S=(U,A,V,F)$ 定义成为区分矩阵 $\boldsymbol{M}=\{M_{ij}\}$，其中 M_{ij} 表示为：

$$M_{ij}=\begin{cases}a\in C:f(x_i a)\neq f(x_j,a)\\ f(x_i,D)=f(x_j,D)\\ \min\{|d(x_i)|\},\{d(x_j)|\}=1\\ \text{其他}\end{cases}$$

式中：$d(x_i)$——U 中全部和 x_i 在关系中的等价元素相应决策属性值创建的集合基数。

在简化分区矩阵中使关系 IND(C)值得出，求解方法复杂度为 $0(m*n*\log n)$，其中 m 指的是属性集数量，n 指的是 U 元素数量，但是并不理想。算法为：

1)输入决策表 $S-(U,C,D,V,F,d)$，$U=\{x_1,x_2,\cdots,xn\}$，$C=\{c_1,c_2,\cdots,c_r\}$。

2)实现 IND(C)的输出。

3)对每个 $c_i(i=1,2,\cdots,r)$得到 $f(x_j)(j=1,2,\cdots,n)$的最小值和最大值，最大值为 M_i，最小值为 m_i。

4)通过静态链表对对象进行存储，分别为 $x_1,x_2,\cdots,x_n$，使表头指针为 x_1。

5)For($i=1;i<r+1;i++$)；第 i 次分配创建 M_i-m_i+1 空队列，使链表中的对象 $x\in U$ 根据链表中的顺序到 $f(x,c_i)-m_i$ 队列中分配。第 i 次收集中的表头指针对第一个非空队列头指针指向，对每个非空对列尾指针进行修改，使其对下个非空队列对头对象指向，以此使 M_i-m_i+1 个队列重新构成链表。

6)根据上一步所得出的链表对象序列，使 $t=1,B_1=\{x_1^{'}\}$；for($j=2;j<n+1;j++$)。

7)如果 B_i 中的全部对象在决策属性中的值相同，那么使 B_i 中的第一个对象融入到 U 中。

文中是遗传算法和粗糙集结合实现约简，免疫遗传算法能够结合生物免疫系统自适应识别及排除侵入机体抗原性异物功能，在遗传算法中融入生物免疫系统记忆、学习、识别及多样性的特点，免疫遗传约简算法的思路为：

1)主要步骤就是选择适应度函数，之后创建适应度函数，表示为：

$$F(r)=\left(1-\frac{\mathrm{card}(r)}{n}\right)+\frac{\mathrm{card}(W)}{\mathrm{card}(M)}$$

式中：r ——染色体中的解；

n ——染色体长度；

$\mathrm{card}(M)$——$\boldsymbol{M}$ 区分矩阵中的非空元素数量，满足 $\boldsymbol{M}$。

适应度函数前部分中的驱使算法朝着属性数最小方向进行搜索，后部分保证 R 为约简[9]。

2)促进产生抗体。要想能够促进抗体高适应度，就要抑制高浓度抗体。抗体相似性利用抗体编码欧几里得距离进行表示，两个抗体之间的欧几里得距离表示为：

$$d=\sqrt{\sum_{1\leqslant i\leqslant n}(a_i-b_i)^2}$$

式中，d 值在不断的增加，那么两者相似度就会越低。假如 d 为 0，那么表示抗体一致。

3)选择操作。使群体抗体利用轮盘赌选择方式从目前抗体群中将优良个体进行选择，将其到下一代群体中进行复制。利用相似性矢量矩作为选择概率，将其定义为：

$$PS(x_i)=a\frac{\rho(x_i)}{\sum_{i=1}^{N}\rho(x_i)}+(1-a)\frac{1}{N}e^{-\frac{C_i}{\beta}}$$

式中：α、β ——常数调节因子；

$F(x)$——适应度函数。

通过以上公式可以看出，此选择概率不仅和抗体适应度具有密切的关系，还和抗体相似度具有密切的关系。此种抗体群体的选择能够避免出现抗体陷入局部最优解问题，使抗体多样性进行保证。

4)在算法中使用基本位变异算子及单点交叉算法，利用最优保存策略对算法收敛进行保证。

①使用基数排序算法对矩阵进行简化区分，假如决策属性 D 和条件属性 C 依赖度和属性核进行确定，两者相等，那么结束运算；如果不相等，进行以下操作。

②通过随机产生 m 个长度 n 二进制串代表个体构成初始抗体群，对核中的属性相应值为 1，否则值为 0，对初始抗体群中的个体适应度进行计算。

③以上一步所计算的适应度，刺激适应度比较大的个体，对抗体浓度进行计算，删除大浓度个体。

④对选择个体概率进行计算，利用轮盘赌方法对个体进行选择。

⑤根据交叉及变异概率产生全新个体，变异的时候保证核属性相应基因位不出现变异。

⑥使用最优保存策略取出父代个体中高适应度个体，将其到下一代个体中进行复制。

⑦如果使用十代最优个体适应度不提高，那么计算终止。如果提高，转到③继续计算。

文中使用的算法在对等价关系进行计算的过程中使用基数排序思想，使等价关系计算时间复杂度表示为 $O(m*n)$，空间复杂度表示为 $O(n)$。在运算算法的过程中融合决策属性对于条件属性依赖度，并且和抗体浓度相互结合，对净化过程个体多样性进行维持，使搜索能力得到提高，避免出现局部最优。

3.2.2　算法可行性分析

文中所提出的基于遗传算法的粗糙集属性约简算法设计过程中，使用 Sigmoid 函数使连续优化问题算法和粗糙属性约简问题相互联系，以此对文中算法可操作性进行验证。之后所分析的最优更新操作都能够提高文中算法约简效果，并且

使用适应度函数能够对种群中接近约简、属性数较少粒子保存进行保证，所以通过多次迭代之后还能够寻找最小约简，以对文中算法有效性进行保证[10]。

4 算法实验

为了对文中算法有效性及可行性进行分析，列举简单实例对算法运算步骤进行说明，之后利用大数据约简对算法有效性进行验证。

其中的实例对象集合为 $U=\{U_1,U_2,\cdots,U_{10}\}$，条件属性集表示为 $C=\{a,b,c,d,e\}$，决策属性表示为 $\{D\}$。

利用可辨识矩阵方法对属性约简结果进行该计算，将其作为对比信息，以下为属性约简步骤。假如粒子维数表示为5，种群数目表示为3，最大迭代次数表示为 $T=10$。

首先，以属性集 $C=\{a,b,c,d,e\}$ 对属性依赖度值进行计算；之后，通过初始化粒子群和属性，将其使用二进制方式进行表示：$P_1=10101$，$P_2=00100$，$P_3=00111$。以粒子适应度值计算，设置全局最优粒子为 $P_1=10101$。

决策信息系统实例见表1。

表1 决策信息系统实例

U	a	b	c	d	e	D
U_1	2	1	3	3	1	1
U_2	3	2	1	1	2	2
U_3	2	1	3	3	2	1
U_4	1	2	3	1	2	4
U_5	1	1	4	2	1	3
U_6	1	1	2	2	1	5
U_7	3	2	1	1	2	2
U_8	1	1	4	2	1	3
U_9	2	1	3	3	1	1
U_{10}	3	2	1	1	2	2

最后，实现遗传算法循环迭代，因为迭代计算过程复杂，计算步骤结果描述见表2。

表2 计算步骤结果的描述

迭代	当前个体	个体更新	适应度值	平均适应度值	最优位置	gbest	选择个体	交叉点	变异点
第一次	$P_1=10101$	00110	0.94	0.84	00110	00110	P_1		
	$P_2=00100$	11101	0.78		11101			4	4
	$P_3=00111$	10111	0.78		00111			4	
第二次	$P_1=00110$	10111	0.78	0.84	00110	00110		3	
	$P_2=11101$	01110	0.86		01110		2		
	$P_3=10101$	11110	0.78		11110			3	2
第三次	$P_1=10110$	00110	0.94	0.89	00110	00110	P_1		
	$P_2=01110$	00111	0.86		00111			3	
	$P_3=10111$	10110	0.86		10110			3	1
第四次	$P_1=00110$	01110	0.86	0.89	00110	00110			
	$P_2=00110$	00111	0.86		00111				
	$P_3=00111$	00110	0.94		00110		P_3		

因为全局最优粒子 gbest 连续四次迭代都没有出现变化，所以运算终止，全局最优位置为00110，也就是此决策表中的最小相对约简属性表示为 cd。文中所设计的算法和相关研究学者算法对比表明，此算法能够有效节约运算时间。

5 结 语

粗糙集理论属于全新处理不确定性、含糊性问题的数学工具，其主要优势就是不需要相关数据预备及其他信息，所以粗糙集理论因为自身的优势备受人们重视，其和遗传算法相互结合成为

研究的重点内容。属性约简为粗糙集理论中的主要研究内容，但是寻找决策表最小约简为较难的问题。文中所提出的基于遗传算法粗糙集约简方法，不仅能够提高算法的局部搜索能力，还能够保证此算法全局寻优特点。实验结果表明，此算法不仅能够实现决策表约简，在对大规模数据计算的时候节约时间。在今后工作过程中，要对遗传约简算法继续进行完善。

参考文献：

[1] 肖厚国.一种基于遗传算法的粗糙集约简方法[J].江苏第二师范学院学报，2016(6)：1-3.

[2] 孙宇航，常晋义，谢从华.一种启发信息遗传算法的粗糙集属性约简算法[J].电脑知识与技术，2015，11(7)：281-285.

[3] 黄伟，赵寅邦，陈乔生，等.基于改进遗传算法和粗糙集的变压器故障诊断[C]// 全国智能电网用户端能源管理学术年会，2015.

[4] 段会龙，吕旭东，尹梓名.一种基于遗传算法和粗糙集的属性约简方法及精神状态评估方法[P].CN 104298873 A. 2015.

[5] 时光，智军，陈军.基于免疫遗传算法的粗糙集属性约简算法[J].电子技术与软件工程，2015(11)：85-86.

[6] 张炜，范年柏，汪文佳.基于自适应遗传算法的股票预测模型研究[J].计算机工程与应用，2015，51(4)：254-259.

[7] 宫丽娜.粗糙集模糊神经网络软件质量预测[J].长春工业大学学报：自然科学版，2014，35(1)：82-85.

[8] 戴上平，刘素军，郑素菲.基于 GA-PSO 的粗糙集属性约简算法[J].计算机工程与科学，2015，37(2)：397-401.

[9] 王丹，周连喆.改进遗传算法柔性作业车间调度[J].长春工业大学学报，2017，38(4)：361-370.

[10] 吴尚智，罗艺纯，翟敬鹏.基于遗传粒子群和粗糙集的最小属性约简算法[J].计算机工程与科学，2016，38(5)：1007-1013.