文章编号: 1002-0411(2008)02-0166-05

粗糙集理论的双重学习方法研究

陈 果
(南京航空航天大学民航学院, 江苏 南京 210016)

摘 要: 提出了一种新的粗糙集双重学习方法,该方法能用遗传算法实现外层学习,用规则提取方法进行内层学习. 其基本思想是:首先引入遗传算法,将属性编码,并针对不同的属性组合进行规则提取;然后用测试样本对规则集进行检验,并基于所得到的识别率建立适应度函数;最后在合适的遗传算子下获取最佳的属性组合及相应的知识规则. 与其他方法相比,本文所提粗糙集双重学习方法集属性约简和规则提取于一体,整个过程具有很强的自适应能力. 最后,用算例对本文方法进行了验证.

关键词: 粗糙集理论; 遗传算法; 机器学习; 属性约简; 值约简; 规则提取; 双重

中图分类号: TP18 **文献标识码**: A

On Double-Layer Learning Method of Rough Set Theory

CHEN Guo
(College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: A new double-layer learning method of rough set is put forward, which can carry out the outer layer learning by genetic algorithms and the inner layer learning by rule extraction. Firstly, the genetic algorithm is introduced, the attributes are coded into binary codes, and the rules are extracted under various attribute combinations. Secondly, the test samples are employed to test the obtained rules, and the fitness function is constructed on the basis of the obtained recognition rate. Finally, the optimum attribute combinations and the corresponding knowledge rules are obtained with proper genetic operators. In comparison with other methods, the presented method combines attribute reduction with rule extraction, and possesses a stronger self-adaptive ability. In the end, examples are given to verify the proposed method.

Keywords: rough set theory; genetic algorithm (GA); machine learning; attribute reduction; value reduction; rule extraction; double-layer

1 引言 (Introduction)

近年来,粗糙集理论[1,2]已经成为人工智能领域一个新的学术热点,在模式识别、机器学习、知识获取、知识发现和决策分析等领域得到了广泛的研究和应用,日益受到了国内外专家和科研人员的关注和重视. 粗糙集理论作为一种新的分析和处理不精确、不一致、不完整信息和知识的数学工具,为智能信息处理提供了有效的处理技术. 它通常包括数据补齐、连续属性离散、属性约简、值约简及规则生成等步骤. 其中属性约简是关键步骤,它能从数据中去除对分类没有任何用处的条件属性,为后续知识规则的提取提供了基础. 但是属性约简的结果往往有很多,究竟用哪个结果提取出的知识泛化能力最强? 分类能力最好? 属性约简方法并没有给出指导性建议.

本文将属性约简、值约简 (规则提取)及规则推理融为一体来考虑,提出用机器双重学习的方法来实现粗糙集理论的知识获取. 其基本思路是将值约简 (规则提取)作为内层学习,将属性约简作为外层学习,将样本随机分为训练样本和测试样本集,用测试样本按一定的推理机制检验通过内层学习得到的知识规则,根据得到的识别率构造遗传算法的适应度函数,用遗传算法实现外层属性约简的学习. 通过

收稿日期: 2006-11-21
基金项目: 国家自然科学基金资助项目 (50705042); 航空科学基金资助项目 (2007ZB52022)

外层和内层的不断学习,最后得到具有最优分类能力和最强泛化能力的知识规则.

2　粗糙集理论的机器双重学习策略(Machine double-layer learning policy of rough set theory)

本文构造了粗糙集理论的机器双重学习方法.算法的基本思想是:在样本集中,随机选取一半作为训练样本,其余作为测试样本.首先,在给定初始属性组合的情况下,利用粗糙集理论的值约简(规则提取)算法获取知识规则,并以此对测试样本进行分类测试,计算识别率,形成遗传算法的适应度函数.然后运用遗传算法的学习机制,自动调节模型的属性组合,在新的属性组合下,应用粗糙集理论的知识规则提取方法得到知识规则.最后输出具有最优分类能力和最强泛化能力的知识规则及属性组合,整个参数选择过程均为自动完成,基本上不需要人工干预.因此,该学习算法具有很强的自适应能力.

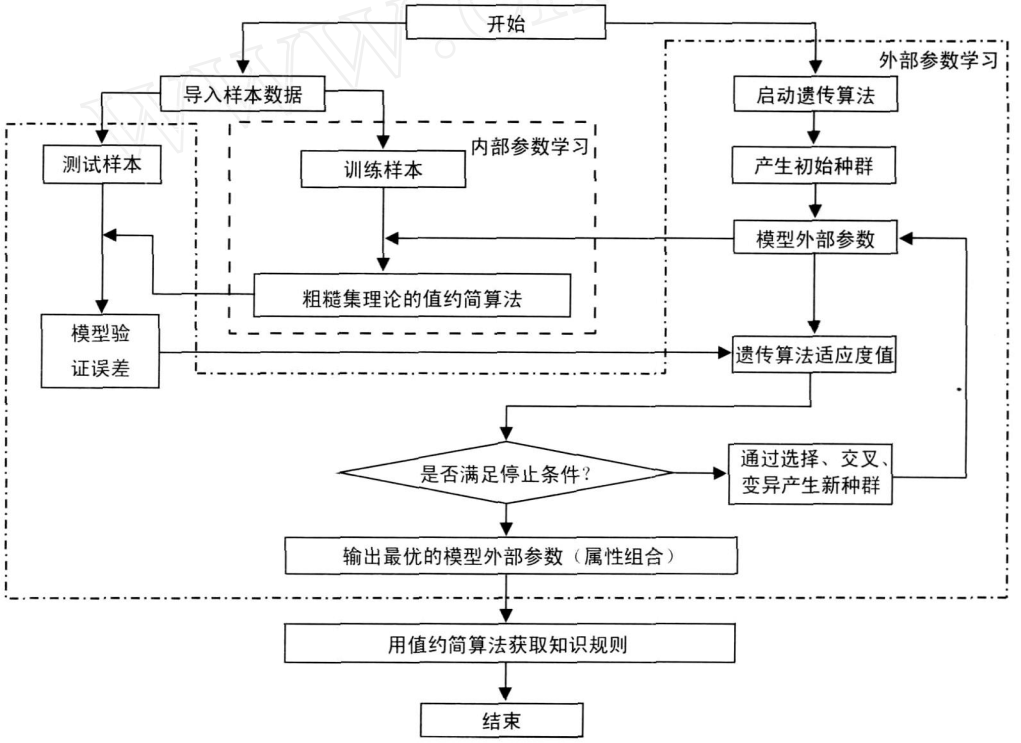

图 1　粗糙集理论的机器双重学习流程

Fig 1　Machine double-layer learning of rough set theory

粗糙集理论的机器双重学习的学习流程如图 1所示,该学习过程包括内部参数(知识规则)的学习和外部参数(属性组合)的学习,在图 1中,虚线框为内部参数的学习过程,点划线框中为外部参数的学习.下面将详细阐述该方法的学习过程.

(1) 算法开始后,读取样本数据,同时启动遗传算法,对模型外部参数进行编码,产生初始种群,对种群解码后得到模型的外部参数.

(2) 随机选取一半样本数据作为训练样本,其余为测试样本.在获取了模型的初始外部参数(属性组合)和训练样本数据后,将启动模型的学习算法,通过对训练样本进行值约简(内层学习)得到知识规则.对所得到的知识规则,用测试样本数据进行验证,以衡量知识的泛化能力.通常,知识规则对测试样本的识别率在一定程度上反映了模型的泛化能力,因此可以转化为遗传算法的适应度函数值.

(3) 依据种群中各染色体的适应度值,对种群中的个体进行选择、交叉和变异以获得新一代的种群,再对新一代的种群中的染色体解码,获取模型新的外部参数(属性组合),然后启动值约简算法,通过对训练样本的学习得到新的知识规则.再运用测试样本对知识规则进行验证,并对外部参数(属性组合)进行调整,进行下轮的外部参数(属性组合)学习,直到达到遗传停止条件.

(4) 最后输出模型最优的外部参数(属性组合),同时利用粗糙集的值约简算法,通过对给定训练样本的学习,得到最优的知识规则.

3 粗糙集理论的基本概念(Basic concepts of rough set theory)

定义 1 决策表:决策表是一个通过信息表来进行知识表达的系统,其列为属性,其行为实例对象.一般来讲,决策表由四部分构成,$S=\langle U, R, V, f\rangle$,其中 U 是论域,$R=C\cup D$ 是属性集合,子集 C 和 D 分别称为条件属性和决策属性. V 为属性值域,$f: U\times R\rightarrow V$ 为一个信息函数,指定了 U 中每一个对象的属性值.

定义 2 不可分辨关系:对于每个属性子集 $B\subseteq R$,定义不可分辨二元关系 $IND(B)$,即:$IND(B)=\{(x, y)|\ (x, y)\in U^2,\ \forall b\in B\ (b(x)=b(y))\}$. 在粗糙集理论中,不可分辨关系是定义其他概念的基础.

定义 3 基本集:由论域中相互间不可分辨的对象组成的集合,是组成论域知识的颗粒.

定义 4 属性约简:如果 $B\subseteq A$,属性集 A 和属性集 B 相对于决策属性的分类一致,也就是具有相同的分类能力,这就称 B 为 A 的相对约简. 对于论域 U, P 和 Q 为定义在 U 上的两个等价关系簇且 $Q\subseteq P$. 如果:①$IND(Q)=IND(P)$,②Q 是独立的,则称 Q 是 P 的一个绝对约简. 对于论域 U, P 和 Q 为定义在 U 上的两个等价关系簇,P 的所有 Q 不可省略的原始关系簇称为 P 的 Q 核,记为 $CORE_Q(P)$. 如果记 P 的所有 Q 约简关系簇为 $RED_Q(P)$,则有 $CORE_Q(P)=\cap RED_Q(P)$.

定义 5 值约简:对于属性约简后的决策表,仍然含有部分冗余信息,对于规则集合中的每条规则的任意条件属性,如果去掉该条件属性,该规则不和规则集中的其他规则冲突,则可以从该规则中去掉该条件属性. 经过这样处理的规则集合中所有规则均不含有冗余条件属性,这一过程即为值约简.

定义 6 决策规则:对于决策表 $S=\langle U, R, V, f\rangle$,其中 U 是论域,$R=C\cup\{d\}$ 是属性集合,子集 C 和 $\{d\}$ 分别称为条件和决策属性集合,$\{a_1, a_2, \cdots, a_n\}\in P$,则公式 $(a_1, v_1)\wedge(a_2, v_2)\wedge\cdots\wedge(a_n, v_n)$ 称为 P 的基本公式,如果 A 是 P 的基本公式且 $B=(d, d_i)$,则 $A\rightarrow B$ 为基本决策规则.

定义 7 规则绝对覆盖度及可信度:对于决策表 $S=\langle U, R, V, f\rangle$,其中 U 是论域,$R=C\cup\{d\}$ 是属性集合,子集 C 和 $\{d\}$ 分别称为条件和决策属性集,决策规则 $A\rightarrow B$ 的不确定性可以用参数对 (α, β) 来表示,则规则可以表示为:

$$A\rightarrow B|\ (\alpha, \beta) \tag{1}$$

其中,

$$\alpha=|X\cap Y| \tag{2}$$

$$\beta=|X| \tag{3}$$

$$X=\{x|\ x\in U\wedge A_x\} \tag{4}$$

$$Y=\{x|\ x\in U\wedge B_x\} \tag{5}$$

A_x 表示条件属性值满足公式 A 的实例集,B_x 表示决策属性值满足公式 B 的实例集,这里,参数 β 表示了该规则在决策表中的绝对覆盖度,α/β 就是该规则的可信度.

定义 8 多数优先的规则推理方法:多数优先的规则选择策略就是认为覆盖多数样本的规则(即根据多个样本得到的规则)具有更大的适应性,具有得到合适结论的更高的概率. 其基本思想是:假设有两条不一致的规则 R_1 和 R_2 同时与一个待识别样本匹配,则:

(1) 若 $\alpha_1/\beta_1=\alpha_2/\beta_2$,则 $\gamma=\gamma_i, \beta_i=\max\{\beta_i|\ i=1, 2\}$.

(2) 若 $\alpha_1/\beta_1=\alpha_2/\beta_2$,则

1) 若 $\beta_1=\beta_2$,则 $\gamma=\gamma_i, \alpha_i=\max\{\alpha_i|\ i=1, 2\}$,即在频度一样的情况下选择可信度较大的那条规则的结论;

2) 若 $\beta_1\neq\beta_2$,则:①若 $\beta_1>\beta_2, \alpha_1/\beta_1>\alpha_2/\beta_2$(出现频度大的规则的可信度高),则 $\gamma=\gamma_1$;②若 $\beta_1>\beta_2, \alpha_1/\beta_1<\alpha_2/\beta_2$(出现频度大的规则的可信度低),则 $\gamma=\gamma_i, \alpha^2/\beta_i=\max\{\alpha_i^2/\beta_i|\ i=1, 2\}$.

4 粗糙集理论双重学习的内层规则提取方法(Inner layer rule extracting method of double-layer learning for rough set theory)

本文引用文[3]的启发式值约简方法作为机器双重学习的内层学习即值约简(即实现规则提取).

现将该算法简述如下:

算法输入:决策表 S(假定系统有 n 条记录,$m-1$ 个条件属性,1个决策属性).

算法输出:决策规则集合.

STEP1: 对决策表中条件属性进行逐列考察. 若删除该列后产生冲突记录,则保留冲突记录的原属性值;对于其他记录,将属性值标记为"?".

STEP2: 删除可能产生的重复记录,并考察每条含有标记"?"的记录. 若仅由未被标记的属性值即可以判断出决策,则将标记"?"改为"*";否则,将

标记“?”改为原属性值;若某条记录的所有条件属性均被标记,则标记“?”修改为原来属性值.

STEP3: 删除所有条件属性均被标记为“*”的记录及可能产生的重复记录.

STEP4: 如果两条记录仅有一个条件属性值不同,且其中一条记录的属性被标记为“*”,那么,对该记录如果可由未被标记的属性值判断出决策,则删除另外一条记录;否则,删除本记录.

经过上述值约简之后得到的新决策表,所有属性值均为该表的值核,所有记录均对应为一条决策规则.

5　粗糙集理论双重学习的外层属性组合优化 (Combinatorial optimization of outer layer attributes of double-layer learning for rough set theory)

本文粗糙集理论双重学习的外层属性组合优化的遗传算法流程为:

(1) 对所有条件属性进行二进制编码得到染色体,并随机产生种群. 染色体的每位对应每个条件属性,其状态“1”和“0”分别表示该条件属性的“取”和“舍”.

(2) 设定种群数目 n,种群数目太小,遗传算法的性能将变得很差或根本找不出问题的解;种群数目太大,则会增加计算量,使收敛时间增长,种群数目一般取为 30～100个.

(3) 对种群中的染色体解码,得到每条染色体代表的条件属性组合,由启发式值约简方法得到知识规则,并用该知识规则对测试样本进行分类,按多数优先[2]的原则进行规则推理,得到规则集对测试样本的识别率

$$r = \frac{\text{被正确识别的样本数}}{\text{总的测试样本数}} \tag{6}$$

群中,然后对父代种群进行选择、交叉和变异等遗传算子运算,从而繁殖出下一代新种群其它 $n-1$个基因串. 本文采用转轮法作为选取方法,适应度大的基因串选择的机会大,从而被遗传到下一代的机会就大;相反,适应度小的基因串选择的机会小,从而被淘汰的机率也大. 对选择出的种群进行交叉和变异操作,本文采用文[4]的均匀交叉法和文[5]的基本位变异法,交叉率一般取为 0.5～0.9. 变异率太大,将使遗传算法变为随机搜索,太小则不会产生新个体,一般取为 0.01～0.1.

(5) 如果达到设定的繁衍代数,返回最好的基因串,并以此获取最佳属性组合,再由值约简算法得到最优的知识规则,算法结束. 否则,回到 (3)继续下一代的繁衍.

6　算例 (Computational examples)

为了说明本文粗糙集理论双重学习方法的有效性,进行了规则知识获取实验,选择本文方法、文[6]的辨识矩阵属性约简法和文[7]的归纳法属性约简法,比较了它们得到的属性组合以及在所获得的知识规则下测试样本的识别率. 实验中采用了 5组取自 UCI机器学习数据库的不同实验数据,在每组数据中随机选一半用于学习,利用所得到的规则对其余数据进行测试. 在计算中,连续属性的离散方法采用 Nguyen和 Skowron提出的贪心算法[2]. 计算结果如表 1所示.

表 1　不同方法的比较结果

Tab. 1　Comparison results of various methods

数据集	样本数	文[6]的辨识矩阵属性约简法		文[7]的归纳法属性约简法		本文双重学习方法	
		属性组合	识别率	属性组合	识别率	属性组合	识别率
Iris	150	{a1 a3 a4}	0.949	{a1 a3 a4}	0.949	{a3 a4}	0.974
Ecoli	336	{a1 a2 a5 a6 a7}	0.511	{a1 a2 a5 a6 a7}	0.511	{a1 a6}	0.642
Glass	214	{a1 a2 a3 a4 a5 a6 a7}	0.427	{a1 a2 a3 a4 a5 a6 a7}	0.427	{a2 a3 a4 a5 a6 a7}	0.564
HSV	122	{a1 a2 a3 a4 a6 a7 a9 a11}	0.300	{a1 a2 a3 a4 a6 a7 a9 a11}	0.300	{a6}	0.700
Pima	768	{a1 a2 a3 a4 a5 a6 a7 a8}	0.379	{a1 a2 a3 a4 a5 a6 a7 a8}	0.379	{a1 a6}	0.690

从表 1中可以看出,文 [6]的辨识矩阵属性约简法和文 [7]的归纳属性约简法所得到的结果完全一样,道理很简单,因为它们的思路基本一样,均是基于属性核,然后逐渐增加属性以获得属性约简形式.然而,并没有考虑该属性约简形式到底对测试样本是否最好.而本文的粗糙集理论双重学习方法,不是从属性核出发,它通过内层和外层学习,充分考虑学习样本和测试样本,因此能够得到最佳的属性组合,从而解决了目前属性约简方法不能有效获取最佳属性约简形式的问题.从表 1中不难看出,用本文方法得到的结果不仅属性个数要少得多而且识别率也明显提高.

7 结论 (Conclusions)

本文针对目前的属性约简方法缺乏对属性约简形式进行选取的问题,将属性约简、值约简 (规则提取)融为一体来考虑,提出了粗糙集理论的双重学习方法.首先,将样本随机分为训练样本和测试样本;然后,用值约简算法作为内层学习算法,获得知识规则,用知识规则对测试样本进行测试得到识别率,并根据识别率构造遗传算法的适应度函数;最后,用遗传算法实现属性的组合优化,进行外层学习.该算法的优点在于整个计算过程自动完成,自动求出最好的属性组合形式.本文方法与根据属性核求取属性约简的方法不一样,因此得到的属性组合形式与属性约简形式不完全一致,但它能够保证所获取的知识规则具有最好的分类能力和最强的泛化能力.最后的比较实验充分验证了本文的粗糙集双重学习算法的有效性和优势.

参 考 文 献 (References)

[1] Pawlak Z. Rough set [J]. International Journal of Information and Computer Science, 1982, 11(5): 341～356.

[2] 王国胤. Rough集理论与知识获取 [M]. 西安: 西安交通大学出版社, 2001.

[3] 常犁云, 王国胤, 吴渝. 一种基于 Rough Set理论的属性约简及规则提取方法 [J]. 软件学报, 1999, 10(11): 1206～1211.

[4] Syswerda D. Uniform crossover in genetic algorithms [A]. Proceedings of the Third International Conference on Genetic Algorithms [C]. San Francisco, CA, USA: Morgan Kaufmann Publisher, 1989. 2～9.

[5] Goldberg D E. Genetic Algorithms in Search, Optimization, and Machine Learning [M]. Boston, MA, USA: Addison-Wesley Publishing Company, 1989.

[6] 顾军华, 周艳聪, 宋洁, 等. 一种新的求解属性值约简算法 [J]. 南开大学学报 (自然科学版), 2003, 36(4): 38～42.

[7] 吴福保, 李奇, 宋文忠. 基于粗集理论知识表达系统的一种归纳学习方法 [J]. 控制与决策, 1999, 14(3): 206～211.

作者简介

陈 果 (1972 -),男,副教授.研究领域为航空发动机状态监测与故障诊断,智能诊断与专家系统,机器学习与知识获取,图像处理及模式识别,非线性转子动力学等.