一种基于 Rough Set 理论的属性约简及规则提取方法*

常犁云 王国胤 吴 渝

(重庆邮电学院计算机科学与技术研究所 重庆 400065)

E-mail: netpopeye@263.net

摘要 该文针对 Rough Set 理论中属性约简和值约简这两个重要问题进行了研究,提出了一种借助于可辨识矩阵(discernibility matrix)和数学逻辑运算得到最佳属性约简的新方法.同时,借助该矩阵还可以方便地构造基于 Rough Set 理论的多变量决策树.另外,对目前广泛采用的一种值约简策略进行了改进,最终使得到的规则进一步简化.

关键词 Rough Set 理论,属性约简,值约简,多变量决策树.

中图法分类号 TP18

Rough Set 理论是由波兰科学家 Z. Pawlak 在 1982 年提出的一种处理含糊和不精确性问题的新型数学工具[1].这一理论从新的视角出发对知识进行了定义,它把知识看做是关于论域的划分,并引入代数学中的等价关系来讨论知识.经过十余年的发展,该理论已渗透到人工智能的各个分支,在模式识别、机器学习等方面也都已有成功的应用.

在 Rough Set 理论中,数据约简是非常重要的一个研究课题.研究人员发现,对许多大型系统,仅有部分数据库表属性必须保留,如果能将冗余属性删除,可大大提高系统潜在知识的清晰度.目前,国际上关于属性约简已有不少方法,本文提出的属性约简是基于 Rough Set 理论,利用 Skowron 提出的可辨识矩阵(discernibility matrix)来得到属性的最佳约简.

在对数据库进行分析时,我们经常关心蕴涵在数据当中的潜在知识.由于数据库系统一般数据量都较大,因此很有必要对信息表(也称决策表,是指将真实世界的信息以条件属性与决策属性构成的表的形式给出)进行值约简.目前,人们已在理论上证明了求取处理对象的所有值约简组合是一个 NP 完全问题,因此很难通过枚举法求出问题的最小值约简.本文提出的算法将文献[2]中所述的值约简策略进行了一定程度的改进,最终能够得到进一步优化的值约简组合.

本文分两节介绍属性约简算法及值约简算法.在介绍属性约简算法时,我们还提出了一种计算简单、语意明确的多变量决策树构造方法.文章最后以一个经典数据集为例,综合运用本文所提出的算法,最终得到较满意的规则.

1 最佳属性约简算法

在数据库系统中,经常存在大量的冗余属性.为减少冗余量,提高数据库中蕴涵知识的可理解程度,人们已在属性约简上作了许多工作,提出了一些比较有效的算法.例如,通过去除某属性后判断不可区分关系是否改变来决定是否应删除该属性[3],但该算法对于求取最佳约简是不完备的.事实上,所谓最佳属性约简应指出一个标准,即约简后得到的属性数最少,或最终得到的规则最简,或全部数据约简量最大,这因需要而定.本文介绍的算

---

* 本文研究得到国家自然科学基金和重庆市应用基础研究基金资助.作者常犁云,1974 年生,硕士,主要研究领域为 Rough Set 理论及应用.王国胤,1970 年生,博士,副教授,主要研究领域为神经网络,Rough Set 理论,知识获取,集成智能系统.吴渝,女,1970 年生,博士,讲师,主要研究领域为小波分析,Rough Set 理论.

本文通讯联系人:常犁云,广州 510620,广州市电信局科技处

本文 1998-10-20 收到原稿,1999-01-22 收到修改稿

法可根据具体需要得到在该要求下的最佳属性约简.

在已知关于 Rough Set 研究成果中，Skowron 提出的可辨识矩阵为我们求取最佳属性约简提供了很好的思路.该方法将信息表中所有有关属性区分信息都浓缩进一个矩阵当中，人们已发现可通过该矩阵方便地得到信息表的属性核[4](属性核是指信息表中不可删除的属性).本算法以可辨识矩阵为基础，重点研究矩阵中除属性核之外的其他属性组合，同时，利用一些简单的数学逻辑协助进行运算.

1.1　可辨识矩阵

可辨识矩阵由波兰华沙大学数学家 Skowron 提出，其定义为：令 $S=(U,A)$ 是一个信息系统，U 为论域且 $U=\{x_1,x_2,\ldots,x_n\}$，A 是条件属性集合，D 是决策属性，$a(x)$ 是记录 x 在属性 a 上的值，可辨识矩阵可表示为

$$(c_{ij})=\begin{cases}\{a\in A:a(x_i)\neq a(x_j)\} & D(x_i)\neq D(x_j)\\ 0 & D(x_i)=D(x_j)\\ -1 & a(x_i)=a(x_j)\quad D(x_i)\neq D(x_j)\end{cases}\qquad i,j=1,2,\ldots,n.$$

1.2　属性约简算法描述

令 M 是决策表 T 的可辨识矩阵，$A=\{a_1,a_2,\ldots,a_n\}$，是 T 中所有条件属性的集合.S 是 M 中所有属性组合的集合，且 S 中不包含重复项.令 S 中包含有 s 个属性组合，每个属性组合表示为 B_i，其公式化描述为 $B_i\in S$，$B_j\in S$，$B_i\neq B_j(i,j=1,2,\ldots,s)$.令 $Card(B_i)=m$，则 B_i 中每个条件属性表示为 $b_{i,k}\in B_i(k=1,2,\ldots,m)$.

由可辨识矩阵的定义我们知道：矩阵中属性组合数为 1 表明，除该属性外其余条件属性无法将信息表中决策不同的两条记录区分出来，即该属性必须保留，与决策表中核属性的概念一致.因此，矩阵中所有属性组合数为 1 的属性均为决策表的核属性(可能为空).令 C_0 是 M 中的核属性集，则有 $C_0\subset A$.

考虑到可辨识矩阵包含了决策表中的所有属性区分信息，因此，核属性外的其余有用属性应从属性组合数不为 1 的矩阵元素中分析取得.假设某信息表除 C_0 外剩余两个属性组合，分别表示为 $a_1a_2\ldots a_m$，$b_1b_2\ldots b_n$，为进行数学逻辑运算，将该属性组合以布尔值表示其中是否包含某个条件属性.例如，$a_1=0$ 表示不包含条件属性 a_1，而 $a_1=1$ 表示包含条件属性 a_1.根据可辨识矩阵可知，如果要识别所有决策不同的记录，则 $a_i(i=1,2,\ldots,m)$ 与 $b_j(j=1,2,\ldots n)$ 之中必然至少各需保留一条属性.构造表达式 $P=(a_1\vee a_2\vee\ldots\vee a_m)\wedge(b_1\vee b_2\vee\ldots\vee b_n)$，由以上分析得到 $P=1$.将 P 转化为析取范式形式，且令 P 中任意合取式项的值均等于 1，则该合取式代表的属性组合连同核属性即可将原决策表中的所有决策区分出来.由于析取范式由多个合取式构成，究竟采用哪组属性组合应根据需要而定，该属性组合与核属性一起构成在指定要求下的最佳属性约简.如信息表除 C_0 外还剩余 N 个属性组合，其处理方法可依此类推.

令 Redu 是决策表 T 属性约简后得到的属性集合，本约简算法的描述见算法 1.

算法 1.

第 1 步.将核属性列入属性约简后得到的属性集合，即 $\mathrm{Redu}=C_0$；

第 2 步.在可辨识矩阵中找出所有不包含核属性的属性组合 S，即

$$Q=\{B_i:B_i\cap\mathrm{Redu}\neq\varnothing,i=1,2,\ldots,s\},\quad S=S-Q;$$

第 3 步.将属性组合 S 表示为合取范式的形式，即 $P=\wedge\{\vee b_{i,k}:(i=1,2,\ldots,s;k=1,2,\ldots,m)\}$；

第 4 步.将 P 转化为析取范式形式；

第 5 步.根据需要选择满意的属性组合.如需属性数最少，可直接选择合取式中属性数最少的组合；如需规则最简或数据约简量最大，则需先进行属性值约简.

1.3　算法举例

现举一气象状况实例作为信息系统，如表 1 所示.

表 1 一个信息系统

	Condition Attribute(C)			Decision Attribute(D)
U	Outlook (a_1)	Temperature (a_2)	Humidity (a_3)	Windy (a_4)	Class
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

根据可辨识矩阵的定义我们知道，矩阵的对角线元素均为 0，且沿左对角线对称. 因此，我们只需计算上对角阵元素即可，如图 1 所示.

0	0	a_1	a_1a_2	$a_1a_2a_3$	0	$a_1a_2a_3a_4$	0	a_2a_3	$a_1a_2a_3$	$a_2a_3a_4$	$a_1a_2a_4$	a_1a_3	0
	0	a_1a_4	$a_1a_2a_4$	$a_1a_2a_3a_4$	0	$a_1a_2a_3$	0	$a_2a_3a_4$	$a_1a_2a_3a_4$	a_2a_3	a_1a_2	$a_1a_3a_4$	0
		0	0	0	$a_1a_2a_3a_4$	0	a_1a_2	0	0	0	0	0	$a_1a_2a_4$
			0	0	$a_2a_3a_4$	0	a_1a_2	0	0	0	0	0	a_4
				0	a_4	0	a_1a_3	0	0	0	0	0	$a_2a_3a_4$
					0	a_1	0	a_1a_4	a_2a_4	a_3a_2	$a_1a_2a_3$	$a_1a_2a_4$	0
						0	$a_1a_2a_3a_4$	0	0	0	0	0	$a_1a_2a_3$
							0	a_2a_3	a_1a_3	a_3a_4	a_1a_4	$a_1a_2a_3$	0
								0	0	0	0	0	$a_1a_2a_3a_4$
									0	0	0	0	a_3a_4
										0	0	0	a_1a_3
											0	0	a_1a_4
												0	$a_1a_2a_3a_4$
													0

图 1 可辨识矩阵

从上面的可辨识矩阵可知，核属性为$\{a_1,a_4\}$. 经观察，不包含核属性的属性组合只有 a_2a_3. 构造表达式 $P=a_2 \vee a_3$，由于 P 仅有一种属性组合，故无需进行形式变换. 这样，原决策表条件属性可约简为$\{a_1,a_2,a_4\}$或$\{a_1,a_3,a_4\}$. 若我们希望得到最简规则，将以上两种属性组合进行值约简后可知：以$\{a_1,a_2,a_4\}$为条件属性得到的规则有：

① $(a_1,sunny)$且(a_2,hot)，则 $Class=N$；

② $(a_1,overcast)$，则 $Class=P$；

③ $(a_1,rain)$且$(a_4,false)$，则 $Class=P$；

④ $(a_1,rain)$且$(a_4,true)$，则 $Class=N$；

⑤ $(a_1,sunny)$且$(a_2,mild)$且$(a_4,false)$，则 $Class=N$；

⑥ $(a_2,cool)$且$(a_4,false)$，则 $Class=P$；

⑦ $(a_2,mild)$且$(a_4,true)$，则 $Class=P$，

共 7 条规则. 而以$\{a_1,a_3,a_4\}$为条件属性得到的规则有：

① $(a_1,sunny)$且$(a_3,high)$，则 $Class=N$；

② $(a_1,overcast)$,则 $Class-P$;

③ $(a_1,rain)$且$(a_4,false)$,则 $Class=P$;

④ $(a_1,rain)$且$(a_4,true)$,则 $Class=N$;

⑤ $(a_1,sunny)$且$(a_3,normal)$,则 $Class=P$,

共 5 条规则. 原决策表的最佳属性约简应为$\{a_1,a_3,a_4\}$.

1.4 多变量决策树构造算法

目前,大多数决策树被限制在每个结点上只检验单个属性,这样的决策树被称为单变量决策树,如 ID3 系统. 由于这一限制,很多复杂概念的表达变得非常困难. 为了克服这一限制,人们开始利用各种方法构造多变量决策树. 苗夺谦、王珏等人以 Rough Set 理论为基础,提出了相对泛化的概念,并将它用于构造多变量检验[5],但该方法相对泛化的语义不易理解.

可辨识矩阵中元素项可分为 3 类. 第 1 类是核属性;第 2 类是除核属性之外还包含其他条件属性的属性组合;第 3 类是不包含核属性的属性组合. 其中第 3 类表明了仅由核属性无法区分其决策的记录,可将这些记录作为论域的一个划分. 在决策表保证一致性(无冲突记录)的前提下,除去无法由核属性区分其决策的记录之外,其余记录可根据决策的不同区分为 N 类(假设决策表有 N 个不同决策). 这样,整个决策表被划分为 $N+1$ 个不同部分. 该多变量决策树构造算法见算法 2.

算法 2.

第 1 步. 利用可辨识矩阵计算核属性集 C_0,若 $C_0=\varnothing$,转第 2 步;否则,转第 3 步;

第 2 步. 用 ID3 方法选择一个最佳属性作为该结点的检验;

第 3 步. 根据决策数 N,将决策表中的记录划分为 $N+1$ 类. 首先根据可辨识矩阵,统计出不包含核属性的所有属性组合所在的行与列,该行与列对应于决策表中的记录号,由这些记录可构成论域上的一个划分. 其余记录根据决策再将论域划分为 N 类,每一类中的记录都具有相同的决策.

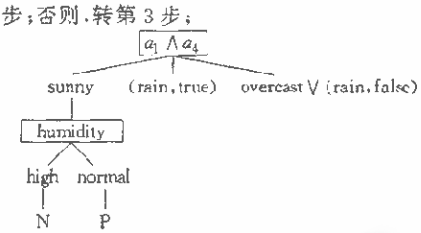

图2　多变量决策树

仍以表 1 为例来构造多变量决策树. 参照如图 1 所示的可辨识矩阵,可以得到 $C_0=\{a_1,a_4\}$. 不包含 a_1,a_4 的属性组合为 a_2a_3,且它们所在的行与列分别为[1,9],[2,11],[8,9]. 因此,记录{1,2,8,9,11}构成论域 U 上的一个划分. 除上述 5 条记录外,原信息表剩余记录中决策为 N 的记录有{6,14},决策为 P 的记录有{3,4,5,7,10,12,13},它们构成了论域 U 上的另外两个划分. 由此可得到本信息系统的多变量决策树,如图 2 所示.

2 值约简算法

Rough Set 理论还具有从信息表中抽取规则知识的能力,事实上,在 Rough Set 理论中抽取规则的过程正是对信息表进行值约简的过程.

分析最小值约简,我们可从值核入手. 所谓值核是指,在信息表的每条记录中寻找对得出决策影响最大的属性值. 关于信息表求值核,目前已有一些资料对其进行了介绍[2]. 本文介绍的值约简算法即建立在文献[2]的基础之上,针对文献[2]中未提出当删除某属性后,在剩余属性所构成的信息表中不存在重复记录时该属性对应值的处理方法提出了部分改进,使最终的值约简结果得到进一步简化.

改进后的值约简算法见算法 3.

算法 3.

第 1 步. 对信息表中条件属性进行逐列考察. 除去该列后,若产生冲突记录,则保留冲突记录的原该属性值;若未产生冲突但含有重复记录,则将重复记录的该属性值标为"*";对其他记录,将该属性值标为"?".

第 2 步. 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判断出决策,则将"?"标记为"*",否则,修改为原属性值;若某条记录的所有条件属性均被标记,则将标有"?"的属性项

修改为原属性值.

第3步. 删除所有条件属性均被标为"*"的记录及可能产生的重复记录.

第4步. 如果两条记录仅有一个条件属性值不同,且其中一条记录该属性被标为"*",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.

经过约简之后得到的新信息表,所有属性值均为该表的值核,所有记录均为该信息表的规则.

现举一例,如表2所示. 其中 a,b,c,d 为条件属性,e 为决策属性. 先由第1.2节所述方法对表2进行属性约简,得到表3. 再对表3进行值约简.

表2

U	a	b	c	d	e
1	1	0	0	1	1
2	1	0	0	0	1
3	0	0	0	0	0
4	1	1	0	1	0
5	1	1	0	2	2
6	2	1	0	2	2
7	2	2	2	2	2

表3

U	a	b	d	e
1	1	0	1	1
2	1	0	0	1
3	0	0	0	0
4	1	1	1	0
5	1	1	2	2
6	2	1	2	2
7	2	2	2	2

对于表3,以记录1为例,若删除属性 a,由于其余属性构成的信息表中不包含重复记录,因此该记录中属性 a 的值标为"?";若删除属性 b,由于 $a=1,d=1$ 与记录4的决策冲突,因此属性 b 保持原值;若删除属性 d,由于 $a=1,b=0$ 与记录2的决策相同,因此属性 d 的值标为"*". 逐条记录进行处理,得到表4.

表2中记录1与记录3属性 b 的值相同,但决策不同,应将表4中记录1的属性 a 改为原值;记录5仅由属性 d 的值即可判断出决策,应将该记录中属性 b 标为"*". 逐条记录进行处理,得到表5.

表5中记录6与记录7除属性 a 外其余属性值都对应相等,且由表2可知,仅由属性 d 即可判断出决策,因此删除记录7. 表6为值约简之后得到的信息表,该表中的记录反映出原信息表中的潜在规则,可总结为:

① $(a,1)$且$(b,0)$,则有$(e,1)$;

② $(a,0)$,则有$(e,0)$;

③ $(b,1)$且$(d,1)$,则有$(e,0)$;

④ $(d,2)$,则有$(e,2)$.

表4

U	a	b	d	e
1	?	0	*	1
2	1	?	*	1
3	0	?	?	0
4	?	1	1	0
5	*	?	2	2
6	*	*	?	2
7	?	*	?	2

表5

U	a	b	d	e
1	1	0	*	1
2	1	0	*	1
3	0	*	*	0
4	*	1	1	0
5	*	*	2	2
6	*	*	2	2
7	2	*	2	2

表6

U	a	b	d	e
1	1	0	*	1
2	0	*	*	0
3	*	1	1	0
4	*	*	2	2

3 实验数据验证

为验证本文提出的约简算法,我们采用了Fisher教授提出的著名的花朵识别数据集(Iris data set)进行测试. 该数据集共有150条被分为3类决策的记录,每条记录包含4个条件属性(petal-length,petal-width,septal-length,septal-width)和1个决策属性. 实验从所有数据集中任选120条记录作为学习样本,由于所有数据均为连续量值,我们利用波兰华沙大学与挪威科技大学联合开发的Rosetta软件中的Orthogoal Scaler方法对所有属性数据进行了离散化. 离散化后,采用本文介绍的方法对其进行属性约简,将特征属性数约简至3个(petal-length,petal-width,septal-width),值约简后获得10条规则,平均每条规则包含两个特征属性,总的数据约简量

约为 96.7%(评价方法参见文献[6]). 为证明其效果，我们用 Rosetta 软件作了对比测试，结果表明，采用针对本问题效果最佳的 Dynamic Reducts 方法进行属性约简后，也得到 3 条属性，结果与本实验一致；但经过值约简后，Rosetta 软件得到 13 条规则，且每条规则均由 3 条属性组成，总数据约简量为 93.5%.

为验证本算法所得规则的有效性，实验对剩余 30 个样本进行了测试. 结果有两个样本产生了误识，识别正确率为 93.3%；而利用 Rosetta 软件得到的规则对同样 30 个样本进行测试时，产生 3 个误识样本，识别正确率为 90%.

4 总　结

本文利用 Skowron 提出的可辨识矩阵，提出了一种计算信息表最佳属性约简的新方法. 该方法将数学逻辑运算运用于可辨识矩阵中核属性以外的其他属性组合，能够得到给定要求下的最佳属性约简. 在基于可辨识矩阵中包含有信息表中所有属性区分信息，本文还提出一种构造多变量决策树的直观方法. 最后，本文提出了一种改进的值约简方法，该方法可有效地简化最终得到的规则知识.

参考文献

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

2 王珏，苗夺谦，周育健. 关于 Rough Set 理论与应用的综述. 模式识别与人工智能，1996,9(4):337～344
(Wang Jue, Miao Duo-qian, Zhou Yu jian. Rough set theory and its application: a survey. Pattern Recognition and Aritificial Intelligence, 1996,9(4):337～344)

3 Pawlak Z, Grzymala-Busse J, Slowinski R *et al*. Rough sets. Communications of the ACM, 1995,38(11):89～95

4 Hu X,Cercone N. Learning in relational databases: a rough set approach. International Journal of Computational Intelligence, 1995,11(2):323～338

5 苗夺谦，王珏. 基于粗糙集的多变量决策树构造方法. 软件学报，1997,8(6):425～431
(Miao Duo-qian, Wang Jue. Rough sets based approach for multivariate decision tree construction. Journal of Software, 1997,8(6):425～431)

6 王珏，王任，苗夺谦等. 基于 Rough Set 理论的"数据浓缩". 计算机学报，1998,21(5):393～400
(Wang Jue, Wang Ren, Miao Duoqian *et al*. Data enriching based on rough set theory. Chinese Journal of Computers, 1998,21(5):393～400)

An Approach for Attribute Reduction and Rule Generation Based on Rough Set Theory

CHANG Li-yun　WANG Guo-yin　WU Yu

(*Institute of Computer Science and Technology　Chongqing University of Posts and Telecommunications　Chongqing*　400065)

Abstract　In this paper, the authors discuss two important issues in rough set research which are attribute reduction and value reduction. A new attribute reduction approach which can reach the best attribute reduction is presented based on discernibility matrix and logic computation. And a multivariate decision tree can be got with this method. Some improvements for a widely used value reduction method are also achieved in this paper. The complexity of acquired rule knowledge can be reduced effectively in this way.

Key words　Rough set theory, attribute reduction, value reduction, multivariate decision tree.