一种基于分辨矩阵的新的属性约简算法

汪小燕，杨思春
（安徽工业大学 计算机学院，安徽 马鞍山 243002）

摘 要：属性约简是粗糙集理论中的重要研究内容之一，但属性约简是一个NP难题，需要通过启发式知识实现。文中提出利用分辨矩阵求不同的条件属性组合相对于决策属性的正域的方法，并给出新的求核属性的方法。在此基础上，提出了一种利用分辨矩阵实现属性约简的新算法，该算法能快速求最少属性且实现简单，并实现了属性约简与规则提取的同步，最后通过实例证明了其正确性。

关键词：粗糙集；分辨矩阵；属性约简；决策树
中图分类号：TP311；TP301.6 **文献标识码**：A **文章编号**：1673-629X(2008)02-0077-03

A New Algorithm for Attribute Reduction Based on Discernible Matrix

WANG Xiao-yan, YANG Si-chun
(School of Computer Science, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: Attribution reduction is one of the important topics in the research on rough set theory. But the attribution reduction is a NP problem, and it needs to be realized by knowledge of elicitation method. Proposes the method that can calculate the positive region of different condition attribution combination relatively decision attribution. A new approach to compute the core attribute is also put forward in this paper. On the basic of this, the new algorithm for attribute reduction based on discernible matrix is proposed. The algorithm can get the smallest attributes quickly and be realized easily. It is realized that attribute reduction is step with rules collection. And it is proved to be workable in the practice.

Key words: rough set; discernible matrix; attribute reduction; decision tree

0 引 言

粗糙集理论是Pawlak等学者在1982年提出的处理不确定、不精确和不完全数据的一种新的数学工具，主要用于知识的简化及知识依赖性的分析[1]。属性约简是粗糙集挖掘知识的核心内容之一，它描述了信息系统属性集中的每个属性是否都是必要的以及如何删除不必要的知识。约简的关键是求出决策表的最小条件属性集，去掉冗余属性，得到简化的决策表，并且约简后的子集和整个属性集的分类能力是一样的。文中根据粗集理论中分辨矩阵的相关概念，提出计算不同的条件属性组合相对于决策属性的正域的方法，并给出计算核属性的新方法。

利用分辨矩阵正域作为启发式知识，提出了一种基于分辨矩阵的核属性和非核属性同步选择的属性约简的算法，在该属性约简的基础上，能够简单地生成一棵决策树，进行规则提取。最后通过实例验证了该算法。

收稿日期：2007-05-24
基金项目：安徽省自然科学基金项目(KJ2007B245)
作者简介：汪小燕(1974-)，女，安徽桐城人，讲师，硕士，研究方向为数据挖掘、粗糙集理论。

1 相关理论

定义1 分辨矩阵由Skowron提出，其定义为[2]：

令 $S=(U,R,V,f)$ 是一个信息系统，U 为论域且 $U=\{x_1, x_2, \cdots, x_n\}$，$R=C\cup D$ 是属性集合，子集 C 和 D 分别是条件属性集和决策属性集，$V=\bigcup Vr$ 是属性值的集合，Vr 表示属性 $r\in R$ 的属性值范围，即属性 r 的值域，$f: U\times R\rightarrow V$ 是一个信息函数，它指定 U 中每一个对象 x 的属性值。$r(x)$ 是对象 x 在属性 r 上的值，$D(x)$ 是记录 x 在 D 上的值，则分辨矩阵记为：

$$(c_{ij})_{n\times n}=\begin{cases} r\in C: r(x_i)\neq r(x_j) & D(x_i)\neq D(x_j) \\ 0 & D(x_i)=D(x_j) \\ -1\ \forall r, \exists r(x_i)=r(x_j) & D(x_i)\neq D(x_j) \end{cases}$$

显然，分辨矩阵是一个按主对角线对称的矩阵，在考虑分辨矩阵的时候，只需要考虑其上三角(或下三角)部分就可以了。

定义 2 设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，Q 的 P 正区域为[2]：

$$POS_P(Q)=\bigcup_{x\in U/Q}P_-(x)$$

定义 3 设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，若 $POS_P(Q)=POS_{(P\setminus\{r\})}(Q)$，则称 r 为 P 中相对于 Q 可省略的(不必要的)，简称 P 中 Q 可省略的；否则，称 r 为 P 中相对于 Q 不可省略的(必要的)[2]。

定义 4 设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，若 P 的 Q 独立子集 $S\subset P$，有 $POS_S(Q)=POS_P(Q)$，则称 S 为 P 的 Q 约简[2]。

2 基于分辨矩阵的新的属性约简算法

定理 1 对任意决策表信息系统 $S=(U,C\cup\{d\},V,f)$，其分辨矩阵中，令所有为 -1 的元素所对应的行列元素的并集为 JH_C，从 U 中除去 JH_C，构成条件属性相对于决策属性的正域，即 $POS_C(D)=U\setminus JH_C$。

证明：设 m_{ij} 为决策表中任意元素，若 $m_{ij}=-1$，则 m_{ij} 所对应的两个不同的实例对象 x_i,x_j 满足：对于 $\forall c\in C, f(x_i,c)=f(x_j,c)$，且 $f(x_i,D)\neq f(x_j,D)$，即实例对象 x_i,x_j 条件属性完全相同，决策属性不相同，这两个对象，根据现有的条件属性，无法确定归入哪一个决策类。而正域 $POS_C(D)$ 中的对象是能够确定分类的对象，则从 U 中除去所有为 -1 的元素所对应的行列元素的并集(设为 JH_C)，构成条件属性相对于决策属性的正域，即 $POS_C(D)=U\setminus JH_C$。

由定理 1 可以得到：

推论 1 设 $S=(U,C\cup D)$ 是一个信息决策系统，C 是条件属性集，$A\subset C$，$D=\{d\}$ 是决策属性集，则分辨矩阵中，某一条件属性组合 A 相对于决策属性的正域 $POS_A(D)$ 为从 U 中除去分辨矩阵中所有不为 0 的元素项中不包含 A 中属性的元素项所对应的行列元素的并集(设为 JH_A)，即 $POS_A(D)=U\setminus JH_A$。

引理 1 在信息决策系统分辨矩阵中，若 $JH_{C\setminus c_i}\neq JH_C$($JH_{C\setminus c_i}$ 表示分辨矩阵中所有不为 0 的元素项中不包含 $\{C\setminus c_i\}$ 中属性的元素项所对应的行列元素的并集)，则属性 c_i 为核属性，否则属性 c_i 为非核属性。

证明：由定理 1 及推论 1 知：$POS_{C-c_i}(D)=U-JH_{C\setminus c_i}$，$POS_c(D)=U-JH_C$，若 $JH_{C\setminus ci}\neq JH_C$，则 $POS_{c-c_i}(D)\neq POS_c(D)$，由定义 3 知属性 c_i 为核属性。同理，若 $JH_{C\setminus c_i}=JH_C$，则 $POS_{c-c_i}(D)=POS_c(D)$，所以属性 c_i 为非核属性。

在建立分辨矩阵的基础上，根据定理 1 和推论 1，很容易计算不同的条件属性组合相对于决策属性的正域，通过判断 $JH_{C\setminus c_i}$ 与 JH_C 是否相等来决定 c_i 是否为核属性。该方法计算核属性简单并且适用于任何决策系统(相容决策系统和不相容决策系统)。

引理 2 设 $S=(U,C\cup D)$ 是一个信息决策系统，C 是条件属性集，$D=\{d\}$ 是决策属性集，$RED(R)\subset C$，若 $JH_{RED(R)}=JH_C$($JH_{RED(R)}$ 表示分辨矩阵中所有不为 0 的元素项中不包含 $RED(R)$ 中属性的元素项所对应的行列元素的并集)，则称 $RED(R)$ 为该信息决策系统的一个约简。

证明：由定理 1 及推论 1 知，$POS_{RED(R)}(D)=U-JH_{RED(R)}$，$POS_c(D)=U-JH_C$，若 $JH_{RED(R)}=JH_C$，则 $POS_{RED(R)}(D)=POS_c(D)$，由定义 4 知 $RED(R)$ 为该信息决策系统的一个约简。

目前，大部分的属性约简算法[3~5]通常是从核开始的，文中提出的属性约简算法对核属性和非核属性的选取是同时进行的。利用定理 1 和推论 1 计算相对正域的大小来选择属性加入到约简集中，以 $JH_{RED}(R)$ 与 JH_C 是否相等为条件判断约简是否结束。

新的属性约简算法描述如下：

输入：决策表 $S=(U,R,V,f)$，$R=C\cup D(C\cap D=\varnothing)$，$C$ 是条件属性集，$D=\{d\}$ 是决策属性集。

输出：约简集 RED(R)

步骤：

(1)RED(R)=∅

(2)for i=1 to m //m 为条件属性的个数

(3)计算 $POS_{c_i}(D)$

(4)next i

(5)取 c_1 满足：$|POS_{c_1}(D)|=\max\{|POS_{c_i}(D)|\}(c_i\in C)$

(6)置 $RED(R)=RED(R)\cup\{c_1\}$；

(7)if $JH_{RED(R)}=JH_C$ then go (10) else go (8)；

(8)计算所有 $c_i\in C-RED(R)$ 的值 $POS_{RED(R)\cup\{c_i\}}(D)$，取 c_2 满足：

$|POS_{RED(R)\cup\{c_2\}}(D)|=\max\{|POS_{RED(R)\cup\{c_i\}}(D)|\}(c_i\in C-RED(R))$

(9)$RED(R)\leftarrow RED(R)\cup\{c_2\}$，go (7)；

(10)输出最小约简 RED(R)。

下面通过例子说明算法的正确性。

例 某一信息决策系统见表 1，其中条件属性集 $C=\{a,b,c\}$，决策属性集 $D=\{d\}$，其分辨矩阵见表 2。

表 1　某一信息决策系统

U	a	b	c	e	d
1	0	0	0	0	0
2	1	0	1	1	1
3	1	1	0	0	0
4	0	2	0	1	1
5	1	2	0	0	1
6	1	0	0	0	0
7	1	2	1	1	1
8	0	0	1	1	1

表 2　表 1 的分辨矩阵

U	1	2	3	4	5	6	7	8
1	0	*ace*	*bce*	*be*	*ab*	0	*abce*	*ce*
2		0	*bce*	0	0	*ce*	0	0
3			0	*abe*	*b*	0	*bce*	*abce*
4				0	0	*abe*	0	0
5					0	*b*	0	0
6						0	*bce*	*ace*
7							0	0
8								0

① 计算。

$JH_a = \{1,2,3,4,5,6,7,8\}$,则 $POS_a(D) = \varnothing$

同理可得:

$POS_b(D) = \{3,4,5,7\}$

$POS_c(D) = \{2,7,8\}$

$POS_e(D) = \{2,4,7,8\}$

选择 b 或 e 加入到 $RED(R)$ 中,设 $RED(R) = \{e\}$

② 由于 $JH_{RED(R)} = \{1,3,5,6\}$,而 $JH_C = \varnothing$,$JH_{RED(R)} \neq JH_C$,故需要增加其它属性,以构成最小约简。

$POS_{RED(R)\cup\{a\}}(D) = \{1,2,4,5,7,8\}$

$POS_{RED(R)\cup\{b\}}(D) = \{1,2,3,4,5,6,7,8\}$

$POS_{RED(R)\cup\{c\}}(D) = \{2,4,7,8\}$

选择 b 加入到 $RED(R) = \{e,b\}$,此时 $JH_{RED(R)} = JH_C$

该决策表的一最小属性约简为 $\{e,b\}$。

3　基于新的属性约简的规则提取

文中提出的新的属性约简算法,通过计算相对正域的大小来选择属性加入到约简集中,相对正域越大,根据该条件属性组合(或单个属性)能够确定分类的对象就越多。故每次选择的属性可以作为决策树的结点,第一次选择的属性可以作为根结点,其后选择的属性可以作为叶子结点。根据新的属性约简算法对表 1 构建的决策树,如图 1 所示。

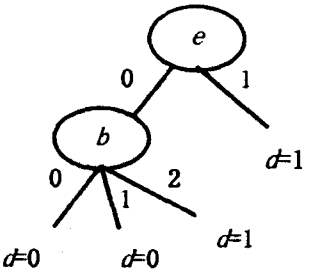

图 1　表 1 的决策树

由图 1 得出表 1 的 4 条规则如下:

$e = 0 \wedge b = 0 \rightarrow d = 0$; $e = 0 \wedge b = 1 \rightarrow d = 0$;

$e = 0 \wedge b = 2 \rightarrow d = 1$; $e = 1 \rightarrow d = 1$

4　结束语

文中提出计算不同的条件属性组合相对于决策属性的正域的方法,并给出计算核属性的新方法。利用分辨矩阵正域作为启发式知识,提出了一种基于分辨矩阵的核属性和非核属性同步选择的属性约简的算法,在该属性约简的基础上,能够简单地生成一棵决策树,进行规则提取。该算法的主要特点有:利用分辨矩阵的正域作为启发式知识;判断约简条件简单。最后通过例子分析,表明该算法是有效的。

参考文献:

[1] Pawlak Z. Vagueness and uncertainty: A Rough Set Prospective[J]. Inter J of Computer Intelligence, 1995, 11(2): 37 - 41.

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

[3] 李秀红,史开泉. 一种基于知识粒度的属性约简算法[J]. 计算机应用,2006,26(6):76 - 77.

[4] 李　珊,肖怀铁,付　强. 改进的粗集属性约简的启发式算法[J]. 电光与控制,2006,13(8):46 - 48.

[5] 王亚英,张春慨,邵惠鹤. 启发式知识约简算法的研究与应用[J]. 控制与决策,2001,16(6):886 - 889.

(上接第 76 页)

新方法[J]. 北方工业大学学报,2002,14(1):1 - 7.

[4] 邹建成. 基于原根的数字图像置乱技术[J]. 北方工业大学学报,2001,13(3):14 - 16.

[5] 李国富. 基于正交拉丁方的数字图像置乱方法[J]. 北方工业大学学报,2001,13(1):14 - 17.

[6] 齐东旭. 分形及其计算机生成[M]. 北京:科学出版社,1997.

[7] 柏　森,曹长修. 一类基于行列式计算思想的图像置乱加密算法[J]. 计算机工程与应用,2002,38(8):37 - 39.

[8] 邹建成,李国富,齐东旭. 广义 Gray 码及其在数字图像置乱中的应用[J]. 高校应用数学学报 A 辑,2002,17(3):363 - 370.