第 38 卷 第 4 期
Vol.38 No.4

计 算 机 工 程
Computer Engineering

2012 年 2 月
February 2012

• 软件技术与数据库 •　　文章编号：1000—3428(2012)04—0046—03　　文献标识码：A　　中图分类号：TP311.12

一种完备的最小属性约简方法

于海燕，乔晓东

(中国科学技术信息研究所信息技术支持中心，北京 100038)

摘　要：为解决粗糙集中的属性约简问题，提出一种完备的最小属性约简方法。将差别矩阵中所有有关属性区分的信息都浓缩进一个差别向量组，计算每个属性在区分 2 个对象的属性集合中出现的概率，作为属性重要性的启发式信息，建立最小属性约简树，得到属性约简。分析结果表明，该方法可以获得所有的最小属性约简。

关键词：粗糙集；决策表；差别属性集；差别向量组；最小属性约简树；最小属性约简

Complete Minimal Attribute Reduction Method

YU Hai-yan, QIAO Xiao-dong

(Information Technology Support Center, Institute of Scientific and Technical Information of China, Beijing 100038, China)

【Abstract】Attribute reduction is the basic problem of rough sets theory. A method for minimal attributes reduction in consistent decision table is proposed in this paper. The discernible information in consistent decision tables is described with discernible vector array. A minimal attribute reduction tree is generated based on the probability of the attributes which discern two objects. All minimal attribute reductions are got from minimal attributes reduction tree. The result of the method is proved to be complete and minimal.

【Key words】rough set; decision table; discernible attribute set; discernible vector array; minimal attribute reduction tree; minimal attribute reduction

DOI: 10.3969/j.issn.1000-3428.2012.04.015

1　概述

粗糙集理论是一种新型的处理模糊和不确定知识的数学工具，属性约简是它的核心内容之一，目前国内外已经提出了许多属性约简方法。

基于差别矩阵[1]的属性约简是属性约简的方法之一，该方法将信息系统中所有有关属性的区分信息都浓缩进一个矩阵中，目前很多属性约简算法都是基于区分矩阵或在此基础上进行改进的[2-6]。文献[7]指出由 Hu 提出的基于差别矩阵的求核方法求出的核与基于正区域的属性约简的核是不等价的，也就是说这 2 种属性约简是不等价的，文献[8]进一步研究指出产生这个问题的根本原因：由于决策表信息系统的不相容性导致了 Hu 的方法出错，因此对于一致决策表基于差别矩阵的属性约简与基于正区域的属性约简是等价的。

对一个信息系统，一般而言求出所有约简与求最小约简都是 NP 难问题，因此，常借助于某种启发式信息来求近优解，但很多方法都是不完备的。本文根据差别矩阵的求核原理提出了一种完备的最小属性约简方法。

2　相关概念

定义 1 决策表定义为 $S=<U,R,V,f>$，$R=C\cup D$ 为属性集合，其中，U 是对象的有限集；C 是条件属性集；D 是决策属性集。对于每个属性 $a\in R$，V_a 为它的值域。每个属性有一个决策函数 $f:U\times R\to V$。

定义 2 把一个给定的决策表的属性区分的信息用一个差别向量组来表示[9]，即 $\boldsymbol{E}_D=(\boldsymbol{E}_{D_1},\boldsymbol{E}_{D_2},\cdots,\boldsymbol{E}_{D_n})$，它的任一项元素 $\boldsymbol{E}_{D_i}=(O_i,G_i,F_i)$ $(i=1,2,\cdots,n)$，其中，O 为可区分对象对；G 为差别属性集；n 为可区分对象的个数；$\boldsymbol{F}$ 为一个概率向量，$\boldsymbol{F}=(f(a_1),f(a_2),\cdots,f(a_m))$，它的每一项表示该属性出现的概率，$m$ 为属性个数，令 $j=1,2,\cdots,m$：

$$f(a_j)=\begin{cases}1/|G_i| & a_j\in G_i\\ 0 & a_j\notin G_i\end{cases}$$

定义 3 在一个决策表 $S=<U,R,V,f>$ 中，$R=C\cup D$，设 $\boldsymbol{M}=(m_{ij})$ 为差别矩阵，$\forall P\subseteq C$，若 P 满足：

(1) $\forall\varnothing\neq m_{ij}\in\boldsymbol{M}$，有 $P\cap m_{ij}\neq\varnothing$；

(2) $\forall a\in P$，$P'=P-\{a\}$ 均不满足(1)。

则称 P 是 C 关于 D 的一个属性约简。

定理 在一个概率向量中，如果其中的某一项元素取值为 1，则该元素所对应的属性为原信息表的核属性。

证明：令概率向量中的任意一项 $f(a_j)=1$，$j=1,2,\cdots,m$，因为 $\boldsymbol{F}=(f(a_1),f(a_2),\cdots,f(a_m))$，$f(a_j)=1/|C_i|$，所以 $|C_i|=1$，也就是在相应的区分对象的差别属性集只有该概率向量对应的一个属性，即在决策表所有属性中只有这个属性能够区分这 2 个对象，因此，这个属性就是原信息表的核属性。

3　最小属性约简算法

由定义 3 可知，差别矩阵中的每一个非空元素都必须有至少一个元素保留在属性约简中，用来区分 2 个相应的对象，所以，一个属性约简即能将所有对象区分开的属性的集合，一个最小属性约简即能将对象区分开的属性个数最少的属性约简。

本文讨论的方法是根据差别矩阵的求核原理，所以，只考虑一致决策表的情况。该方法在属性约简后决策表要能

基金项目：中国博士后科学基金资助项目“叙词表的自动集成及领域本体构建方法研究”(2011M500370)

作者简介：于海燕(1968－)，女，副教授、博士，主研方向：知识组织，数据挖掘，粒度计算；乔晓东，研究员、硕士

收稿日期：2011-08-26　**E-mail：**yuhyan68@163.com

100%地覆盖整个训练集，且对象之间都能区分开。

在差别矩阵中，属性组合数为 1 的元素表明除该属性外，其余条件属性无法将该元素对应 2 条决策属性不同的对象区分开，即该属性必须保留，与决策表中核属性的概念一致，所以，矩阵中所有属性组合数为 1 的属性均为决策表的核属性，所有包含核属性的元素所对应的对象都可以由核属性与其他对象区分开。

根据这个思路，首先将核属性加入约简，将包含核属性的元素删除，在剩余的元素中继续寻找能区分其余对象的属性。在剩余的元素中，每个元素包含的属性中至少要有一个属性来区分其对应的 2 个对象，即属性约简中至少要包含其中的一个属性，属性个数越少，被选中的概率越大，以此概率作为属性的重要性，每个属性被选中的概率为该元素包含的属性个数分之一，选择重要性大的属性加入约简，然后删除包含该属性的元素，重复这个过程，直到所有对象都被区分开。

将决策表中所有有关属性区分的信息浓缩进一个差别向量组，以每个属性在区分 2 个对象的属性集合中出现的概率作为属性重要性的启发式信息，建立最小属性约简树，得到最小属性约简，并证明该方法是完备的，且用该方法能得到所有最小属性约简。

最小属性约简算法的具体步骤如下：

输入 训练数据

输出 属性约简树

Step1 根据原始信息表建立差别向量组，并令 $stop=0$，$h=1$。

Step2 创建结点 N。

Step3 选择所有概率为 1 的属性即核属性。

Step4 如果没有核属性，则标记结点 N 为空属性标记 $\varnothing$，否则标记结点 N 为核属性集合，并删除包含核属性的对象对应的向量组，将剩余结果存入 V_{11}。

Step5 如果剩余差别属性集为空，则加上一个树叶，标记为“Y”，结束。

Step6 $h=h+1$，对每个分枝分别选择差别属性集中概率最大的属性，设该层中概率最大的属性共有 m 个，对每个属性做如下操作：

Step6.1 长出一个分枝结点，标记为 a_i，$i=1,2,\cdots,m$。

Step6.2 在上一层剩余元素的基础上删除差别属性集中包含属性 a_i 的项，将剩余结果存入 V_{hi}。

Step6.3 如果 V_{hi} 为空，则加上树叶，标记为“Y”，并令 $stop=1$。

Step7 如果 $stop=0$，则对深度为 h 的每一个分枝执行 Step6。

Step8 对每一个没有树叶的分枝加上树叶，标记为“N”，生成最小属性约简树。

每个树叶标记为“Y”的分枝所对应的属性集合都是一个最小属性约简。

4 算法分析

4.1 最小性证明

证明：设由该方法得到的任意一个约简为 $\{c_1,c_2,\cdots,c_n\}$，对应的概率为 $p_1,p_2,\cdots,p_n$，且 $p_1\geqslant p_2\geqslant\cdots\geqslant p_n$，若有任意属性 c_j (对应的概率为 p_j，且 $p_j<p_n$)将 c_i ($c_i\in\{c_1,c_2,\cdots,c_n\}$) 替换，则可能出现 3 种情况：

(1) c_j 只能将 c_i 对应的对象区分，则新属性集合为一个约简，且属性个数与 $\{c_1,c_2,\cdots,c_n\}$ 相等，原约简为最小属性约简。

(2) c_j 不能将 c_i 对应的对象区分，则新属性集合不是一个约简，需要另外选择属性区分 c_i，这时属性约简中的属性个数大于约简 $\{c_1,c_2,\cdots,c_n\}$ 中的属性个数，所以原属性约简仍为最小属性约简。

(3) c_j 不仅能区分 c_i，且能区分其他对象，设被区分的对象对为 O_k，则 c_j 必属于能将 O_k 区分开的属性子集，且 c_j 在这里的对应概率为 p_k，即 $p_j=p_k$，因为 $p_1\geqslant p_k\geqslant p_n$，对于概率大于等于 p_n 的属性都在最小属性约简树中处理过，所以不属于新的约简情况。

综上所述，该方法得到的属性约简是属性个数最少的，即为最小属性约简。

4.2 完备性证明

证明：根据 Pawlak 的定义，若属性集 $P\subset C$ 是给定决策表的条件属性集 C 的约简，则 P 应满足 2 个条件：

(1) $POS_P(D)=POS_C(D)$；

(2)对任意的 $a\in P$，$POS_{P-\{a\}}(D)\neq POS_C(D)$。

如果一个属性约简算法求得的属性集同时满足条件(1)和条件(2)，则称为属性约简的完备算法，如果只满足条件(1)，则称为属性约简非完备算法。

本文算法是根据差别矩阵的原理得到的最小属性约简，能将所有对象区分开，所以满足条件(1)是显然的。根据 4.1 节的最小性证明，该方法得到的最小属性约简中去掉任何一个属性都将有对象对不能区分开，即该约简将不再是一个约简，所以条件(2)满足。因此，该方法是完备的。

4.3 算法的复杂度

经典的基于分辨矩阵的属性约简算法，如 Hu 的差别矩阵方法求核属性的时间复杂度为 $O(n^2m)$，$n=card(u)$，$m=card(c)$；而本文算法只是通过对差别向量进行一次扫描就能将核属性识别出来，算法的时间复杂度最大仅为 $O(nm)$，对于每一层都是搜索一次，即找到属性出现概率最大的属性，假设约简后的属性个数为 r 个，则属性约简树最多为 r 层，所以，约简算法的最大时间复杂度为 $O(nm^r)$。

在实际搜索过程中由于搜索范围逐渐缩小，时间复杂度要小得多。

5 算例

给定一个决策表如表 1 所示，其中，a,b,c,d,e,f 为条件属性；g 为决策属性。

表 1 决策表

x	a	b	c	d	e	f	g
x_1	1	2	2	1	1	1	1
x_2	1	3	1	1	2	3	3
x_3	2	2	2	3	1	2	2
x_4	3	3	2	3	2	1	1
x_5	1	3	1	2	1	3	1
x_6	3	1	3	1	2	2	2
x_7	2	2	3	3	1	2	3
x_8	1	3	2	3	1	2	3

将属性区分的信息用差别向量组表示，如表 2 所示，为了描述清楚，将决策表的差别矩阵在此列出，可与差别向量

组进行比较:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	∅	*bcef*	*adf*	∅	∅	*abcef*	*acdf*	*bdf*
x_2		∅	*abcdef*	*acdf*	*de*	*abcf*	∅	∅
x_3			∅	*abef*	*abcdf*	∅	*c*	*ab*
x_4				∅	∅	*bcdf*	*abcef*	*aef*
x_5					∅	*abcdef*	*abcdf*	*cdf*
x_6						∅	*abde*	*abcde*
x_7							∅	∅
x_8								∅

表 2 差别向量组

O	*C*	*F*	*O*	*C*	*F*
(x_1,x_2)	(b,c,e,f)	(0,1/4,1/4,0,1/4,1/4)	(x_3,x_7)	(c)	(0,0,1,0,0,0)
(x_1,x_3)	(a,d,f)	(1/3,0,0,1/3,0,1/3)	(x_3,x_8)	(a,b)	(1/2,1/2,0,0,0,0)
(x_1,x_6)	(a,b,c,e,f)	(1/5,1/5,1/5,0,1/5,1/5)	(x_4,x_6)	(b,c,d,f)	(0,1/4,1/4,1/4,0,1/4)
(x_1,x_7)	(a,c,d,f)	(1/4,0,1/4,1/4,0,1/4)	(x_4,x_7)	(a,b,c,e,f)	(1/5,1/5,1/5,0,1/5,1/5)
(x_1,x_8)	(b,d,f)	(0,1/3,0,1/3,0,1/3)	(x_4,x_8)	(a,e,f)	(1/3,0,0,0,1/3,1/3)
(x_2,x_3)	(a,b,c,d,e,f)	(1/6,1/6,1/6,1/6,1/6,1/6)	(x_5,x_6)	(a,b,c,d,e,f)	(1/6,1/6,1/6,1/6,1/6,1/6)
(x_2,x_4)	(a,c,d,f)	(1/4,0,1/4,1/4,0,1/4)	(x_5,x_7)	(a,b,c,d,f)	(1/5,1/5,1/5,1/5,0,1/5)
(x_2,x_5)	(d,e)	(0,0,0,1/2,1/2,0)	(x_5,x_8)	(c,d,f)	(0,0,1/3,1/3,0,1/3)
(x_2,x_6)	(a,b,c,f)	(1/4,1/4,1/4,0,0,1/4)	(x_6,x_7)	(a,b,d,e)	(1/4,1/4,0,1/4,1/4,0)
(x_3,x_4)	(a,b,e,f)	(1/4,1/4,0,0,1/4,1/4)	(x_6,x_8)	(a,b,c,d,e)	(1/5,1/5,1/5,1/5,1/5,0)
(x_3,x_5)	(a,b,c,d,f)	(1/5,1/5,1/5,1/5/0,1/5)			

根据属性在差别属性集中出现的概率为 1 找到核属性，这里只有属性 c，将 c 作为最小属性约简树的根结点，并删除包含 c 的元素，剩余元素的差别属性集合为(a,b),(d,e),(a,d,f),(b,d,f),(a,e,f),(a,b,e,f),(a,b,d,e)，见表 3 的第 1 层。

表 3 建树过程分析表

层次	分枝属性	剩余属性子集
1	(c)	(a,b),(d,e),(b,d,f),(a,e,f),(a,b,e,f),(a,b,d,e)
2	(c,a)	(d,e),(b,d,f)
	(c,b)	(d,e),(a,d,f),(a,e,f)
	(c,d)	(a,b),(a,e,f),(a,b,e,f)
	(c,e)	(a,b),(a,d,f),(b,d,f)
3	(c,a,d)	∅
	(c,a,e)	(b,d,f)
	(c,b,d)	(a,e,f)
	(c,b,e)	(a,d,f)
	(c,d,a)	∅
	(c,d,b)	(a,e,f)
	(c,e,a)	(b,d,f)
	(c,e,b)	(a,d,f)

在剩余元素中选择概率最大的属性，即 a,b,d,e，它们的概率均为 1/2，为每个属性创建一个分枝，并在上一层剩余元素的基础上分别删除包含这些属性的元素，分别得到差别属性子集，见表 3 的第 2 层。由于每个分枝的剩余属性集合都不为空，因此继续对每个分枝分别选择概率最大的属性。为选出的每个属性创建一个分枝，并在本分枝上一层的剩余属性集合中删除包含该属性的元素，结果见表第 3 层，其中，分枝(c,a,d)、(c,d,a)的剩余差别属性子集为空，给这 2 个分枝加上树叶标记“Y”，由于出现了属性子集为空的情况，因此停止属性选择，给其他分枝加上树叶标记“N”，生成最小属性约简树，如图 1 所示。

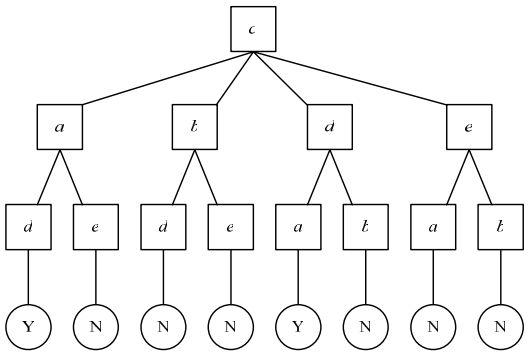

图 1 最小属性约简树

在最小属性约简树中，每一个标记为“Y”的分枝上的属性组合即为一个最小属性约简，这里有 2 个属性组合，即(c,a,d)和(c,d,a)，由于这 2 个分枝包含的属性相同，因此得到一个最小属性约简(a,c,d)。

6 结束语

本文提出一种完备的最小属性约简方法，根据差别矩阵的求核原理，以每个属性在 2 个对象属性集合中出现的概率作为属性重要性的启发式信息，逐渐缩小选择属性的范围，直到找到最小的属性约简。在处理过程中，由于分别处理的情况太多，因此需要建立一棵类似于决策树的最小属性约简树，以便从最小属性的约简树得到所有的最小属性约简。

参考文献

[1] Skowron A, Rauszer C. The Discernibility Matrics and Functions in Information System[M]. Dordrecht, Holland: Kluwer Academic Publishers, 1992.

[2] 王 珏, 王 任, 苗夺谦, 等. 基于 Rough Set 理论的“数据浓缩”[J]. 计算机学报, 1998, 21(5): 393-399.

[3] 潘 丹, 郑启伦. 属性约简自寻优算法[J]. 计算机研究与发展, 2001, 38(8): 904-910.

[4] Wang Jue, Wang Jiayang. Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489-504.

[5] 蒙 韧, 徐章艳, 杨炳镛. 基于 Skowron 差别矩阵属性约简的矩阵表示[J]. 计算机工程, 2010, 36(17): 54-56.

[6] 王加阳, 高 灿. 改进的基于差别矩阵的属性约简算法[J]. 计算机工程, 2009, 35(3): 66-68.

[7] 叶东毅, 陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报, 2002, 30(7): 1086-1088.

[8] 王国胤. 决策表核属性的计算方法[J]. 计算机学报, 2003, 26(5): 611-615.

[9] 鄂 旭, 高学东, 喻 斌. 基于扫描向量的属性约简方法[J]. 北京科技大学学报, 2006, 28(6): 604-608.

编辑 任吉慧