

# 考虑属性排名的约简算法

朱晓钟<sup>1,2</sup>, 杨 勇<sup>1</sup>, 朱英丽<sup>1</sup>

(1. 西北师范大学 数学与信息科学学院, 甘肃 兰州 730070;

2. 河海大学 计算机与信息学院, 江苏 常州 213022)

**摘 要:**属性约简是粗糙集研究的重要内容之一。目前有多种计算约简集的方法,但计算效率普遍不高。杨萍等学者提出的基于二进制区分矩阵的启发式约简算法,考虑了属性的区分度和区分率,采用高效的逻辑运算获得约简集,提高了运算的效率。在该算法的基础上,首先指出其计算所得的约简集存在不确定性,然后给出一种考虑属性排名的改进的约简算法,消除了约简集的不确定性,并且可以迎合用户的需求。最后通过一个信息系统实例,验证该算法的可行性和有效性。

**关键词:**属性排名;粗糙集;属性约简;二进制区分矩阵

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2010)12-0082-04

## A Reduction Algorithm Considering Ranking Order of Attributes

ZHU Xiao-zhong<sup>1,2</sup>, YANG Yong<sup>1</sup>, ZHU Ying-li<sup>1</sup>

(1. College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China;

2. College of Computer and Information, Hohai University, Changzhou 213022, China)

**Abstract:** The attribute reduction is one of the major contents of rough set research. There are a variety of methods based on rough set theory to compute reduct set of an information system, but their computational efficiency is always not high. Yang Ping proposed a heuristic reduction algorithm based on binary-valued discernibility matrix, taking into account the degree and ratio of differentiation between condition attributes, adopting efficient logical operator computation to obtain reduct set. Based on the thesis of Yang Ping's algorithm, first of all points out the result of Yang Ping's algorithm is a reduct with some kind of uncertainty, then proposed an improved algorithm considering ranking of condition attributes. It eliminates the uncertainty of final reduct and meets the needs of user simultaneously. Finally, an information system example is used to show its feasibility and effectiveness.

**Key words:** ranking order of attributes; rough set; attribute reduction; binary discernibility matrix

## 0 引 言

粗糙集理论作为一种新的处理不确定性的有效数学工具,在计算机科学与技术领域中发挥了重要的作用。该理论由 Pawlak 于 1982 年首先提出<sup>[1]</sup>, 20 世纪 90 年代开始得到快速发展。经典粗糙集理论以等价关系(自反性、对称性、传递性)为基础,通过等价关系对论域进行划分,而知识即表现为等价关系对论域划分的结果,划分越细,知识越精确。为描述知识不确定性,粗糙集理论通过引入上、下近似运算来逼近论域中

的任一概念<sup>[2~4]</sup>。

属性约简是数据挖掘和粗糙集理论的核心内容之一,其目的是找到决策表的主要属性。大多数情况下,近似空间中的属性并不是同等重要的,甚至某些属性是冗余的。为此需要进行属性约简。现有的属性约简算法,如基于正区域的约简算法<sup>[5,6]</sup>、基于区分矩阵中属性频率的约简算法<sup>[7,8]</sup>以及基于信息熵的约简算法<sup>[9]</sup>等,都能得到合理的属性约简,但这些算法的复杂度依然较高。最近,杨萍等人<sup>[10]</sup>提出了利用二进制区分矩阵并考虑属性的区分度和区分率进行属性约简的新思路,实例证明了该算法的合理性和有效性,在属性约简中显示出其独特性。但遗憾的是,该算法在条件属性具有相同区分度和区分率时,采用了随机任意选择的策略,因此所得的属性约简集有随机性。文中基于文献<sup>[10]</sup>中的已有算法,通过引入用户对属性的排名<sup>[11]</sup>,消除了不确定性,进一步完善了该约简算法。

收稿日期:2010-03-18;修回日期:2010-06-19

基金项目:国家自然科学基金(10771171);兰州市科技计划项目(2008-1-34)

作者简介:朱晓钟(1978-),男,硕士,讲师,CCF 会员,研究方向为粗糙集理论及数据挖掘;杨 勇,博士,副教授,研究方向为粗糙集理论及其应用。

## 1 粗糙集基本概念

### 1.1 信息系统和近似集合

定义 1 四元组  $S = (U, A, V, f)$  是一个信息系统,其中: $U$  为对象的非空有限集合; $A$  为属性的非空有限集合; $V = \bigcup_{a \in A} V_a$ ,  $V_a$  是属性  $a$  的值域; $f: U \times A \rightarrow V$  是一个信息函数,它为每个对象的每个属性赋予一个信息值,即对任意  $a \in A, x \in U, f(x, a) \in V_a$ 。

定义 2 设  $P \subseteq A, X \subseteq U$ 。 $X$  关于  $P$  的下近似、上近似分别定义为:

(1)  $P_* X = \{x \in U \mid [x]_P \subseteq X\}$ ; (2)  $P^* X = \{x \in U \mid [x]_P \cap X \neq \emptyset\}$ 。其中  $[x]_P$  表示  $P$  划分下包含元素  $x \in U$  的等价类。一个等价类中所有对象之间具有不可区分关系。

### 1.2 区分矩阵及属性约简

定义 3 给定决策表  $S = (U, C \cup D, V, f)$ ,其中  $C$  是条件属性集,  $D$  是决策属性集,区分矩阵  $M_D = \{m_{ij}\}$  定义为:

$$m_{ij} = \begin{cases} \{a \in C; f(x_i, a) \neq f(x_j, a)\}, \\ \text{当 } f(x_i, D) \neq f(x_j, D) \text{ 时} \\ \emptyset, \text{其他} \end{cases}$$

定义 4 令  $R$  为一族等价关系,  $p \in R$ , 若  $\text{ind}(R) = \text{ind}(R - \{p\})$ , 则称  $p$  为  $R$  中不必要的, 否则称  $p$  为  $R$  中必要的。若对于每一个  $p \in R$  都是  $R$  中必要的, 则称  $R$  为独立的, 否则称  $R$  为依赖的。

定义 5 设  $P \subseteq R$ , 如果  $P$  是独立的, 且  $\text{ind}(P) = \text{ind}(R)$ , 则称  $P$  为  $R$  的一个约简。显然可知  $R$  有多种约简。 $R$  中所有必要属性组成的集合称为  $R$  的核 (core), 记作  $\text{core}(R)$ 。 $\text{core}(R) = \bigcap \text{red}(R)$ , 其中  $\text{red}(R)$  表示  $R$  的所有约简。需要说明的是, 一般属性约简不唯一, 其中包含关系最小的约简为最小约简, 而核是唯一的。

## 2 基于二进制区分矩阵属性约简基本概念

定义 6<sup>[12]</sup> 将对象间的不可区分关系用二进制形式矩阵来表达, 这种矩阵称为二进制区分矩阵。对象  $x_i$  和  $x_j$  关于条件属性  $c$  的区分元素用二进制形式来表示,  $M_B = \{m(i, j, c)\}$  定义如下:

$$m(i, j, c) = \begin{cases} 1 & \text{当 } c \in C \wedge f(x_i, c) \neq f(x_j, c) \wedge \\ & f(x_i, d) \neq f(x_j, d) \text{ 时} \\ 0 & \text{其他} \end{cases}$$

根据定义 6, 若决策表  $S = (U, C \cup D, V, f)$ , 其中  $U = \{x_1, x_2, \dots, x_{|U|}\}$ , 条件属性  $C = \{c_1, c_2, \dots, c_{|C|}\}$ , 决策属性  $D = \{d\}$ ,  $|I|$  表示集合的基数, 则  $S$  的二进制区分矩阵可由下面的算法获得:

STEP 1: 初始化,  $M_B = \text{null}$ ;

STEP 2: 将值填入  $M_B$ ;

for ( $i = 1; i \leq |U|; i++$ )

for ( $j = i + 1; j \leq |U|; j++$ )

{ if  $f(x_i, d) \neq f(x_j, d)$

for ( $k = 1; k \leq |C|; k++$ )

{ if ( $f(x_i, c_k) \neq f(x_j, c_k)$ )

$f(i, j, c_k) = 1$

else

$f(i, j, c_k) = 0$  }

}

STEP 3: 输出  $M_B$ , 程序结束。

二进制区分矩阵直接描述了论域中对象之间的分辨情况。若矩阵中某个元素为 1 或 0, 则说明所在的行属于不同决策的两个对象, 这两个对象在 1 或 0 所在的列属性下可分辨或不可分辨。此外, 若二进制区分矩阵中有全为 0 的行, 则说明相应的决策表是不协调的, 否则决策表是协调的; 若某一行的元素全为 1, 说明相应的两个对象在决策表  $S$  的任何一个条件属性下都可分辨, 此时去掉这一行不影响约简; 若某一行只有一个元素为 1, 其余元素均为 0, 则这个元素 1 对应的条件属性一定属于核属性。

定义 7<sup>[10]</sup> 属性的区分度定义为:  $F_1(c_k) = \sum_{i,j} m(i, j, c_k)$ , 其中  $i, j = 1, 2, \dots, |U|; k = 1, 2, \dots, |C|$ 。 $F_1(c_k)$  是二进制区分矩阵中第  $k$  列元素之和, 它表示属性  $c_k$  能区分的对象对的个数。区分度越强, 属性的重要性越大。

定义 8<sup>[10]</sup> 属性的区分率定义为:  $F_2(c_k) = \sum_{i,j} \frac{m(i, j, c_k)}{\sum_k m(i, j, c_k)}$   $i, j = 1, 2, \dots, |U|; k = 1, 2, \dots, |C|$ 。其中,  $\sum_k m(i, j, c_k)$  为二进制区分矩阵中, 对象对  $(x_i, x_j)$  所在行的元素之和, 它表示能区分  $(x_i, x_j)$  的条件属性的个数;  $\frac{m(i, j, c_k)}{\sum_k m(i, j, c_k)}$  表示在所有能区分  $(x_i, x_j)$  的属性中,  $c_k$  所占的比例, 即  $c_k$  对  $(x_i, x_j)$  的区分率。 $F_2(c_k)$  是  $c_k$  对所有对象对的区分率之和。区分率越大, 属性重要性越大。

## 3 考虑属性排名的启发式约简算法

### 3.1 算法的基本思想

因为用户对决策表的属性有偏好或对决策表中各属性关心程度不一样是客观存在的事实, 所以当属性约简有多种选择时, 考虑用户对属性的偏好是合理且

必要的。考虑用户的偏好后,改进的决策表属性约简算法的基本思想是:

(1) 在二进制区分矩阵中求出各条件属性的区分度  $F_1(c_i), i = 1, 2, \dots, |C|$ 。

(2) 以属性区分度作启发信息,依次选择区分度大的属性进入约简集  $R$ 。若两属性区分度相同则比较区分率,若区分率仍然相同,则查找用户设定的属性排名,取排名高的属性进入约简集  $R$ 。

(3) 约简集  $R$  的二进制形式和矩阵的每一行进行逻辑“与”运算,若运算结果中 1 的数目不为 0,说明当前的约简集已包含了将该对象对区分开的信息。

(4) 若约简集  $R$  能将矩阵中的所有对象对都区分开,则程序结束,返回的约简集  $R$  就是决策表的一个属性约简。

### 3.2 算法的描述

假设决策表  $S = (U, C \cup \{d\}, V, f)$ , 条件属性集  $C = \{c_1, c_2, \dots, c_k\}, k = |C|$ , 决策属性  $D = \{d\}$ ,  $R$  是  $C$  的一个约简,  $Q$  是  $R$  的二进制表示。

input: 决策表  $S$ , 属性排名表。

output: 条件属性集的约简集  $R$ 。

步骤:

step1: 初始化  $R = \emptyset, Q$  的所有位为 0。

step2: 生成二进制区分矩阵  $M_B, M_j$  表示矩阵的第  $j$  行。

step3: 计算属性的区分度  $F_1(c_i), i = 1, 2, \dots, k$ 。如果  $i = p$  时,  $F_1(c_p)$  取得最大值,则将  $Q$  的第  $P$  位置 1; 如果  $i = p$  和  $i = q$  时,  $F_1(c_p)$  与  $F_1(c_q)$  同时取得最大值,则比较  $F_2(c_p)$  和  $F_2(c_q)$ , 取值大的属性进入约简集  $R$ , 即  $Q$  的对应位置 1; 若仍相等,则查找属性排名表,假设  $c_q$  排名在  $c_p$  之前,则将  $Q$  的第  $q$  位置 1。

step4: 遍历  $M_B$  的所有行,如果  $M_j$  和  $Q$  进行逻辑“与”运算后,结果中 1 的数目不为 0,删除这一行。

step5: 如果  $M_B = \emptyset$ , 则将  $Q$  转换为  $R, R$  就是所求的约简集; 否则,转 step3。

## 4 实例分析

因为 Outlook 属性值只需眼睛即可确认。Windy 属性值也无需借助仪器。Temp 属性值需要温度计,但温度计较易获得。Humidity 属性值需要借助湿度计,该仪器较难获得。因此对用户来说,对决策系统中各属性的喜好或关心程度是不一样的。现假定用户依据获取属性值的难易程度确定的属性排名为  $a > d > b > c$ 。一个实际的天气预报决策表如表 1 所示。

(1) 计算出该决策表的二进制区分矩阵  $M_B$ , 见表 2。

表 1 一个实际的天气预报决策表

U	Outlook (a)	Temp (b)	Humidity (c)	Windy (d)	Decision (e)
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	Hot	Normal	False	P
6	Rain	Hot	Normal	True	N
7	Overcast	Hot	Normal	True	P
8	Sunny	Cool	Normal	False	P
9	Sunny	Mild	Normal	True	P
10	Overcast	Mild	High	True	P

表 2 决策表的二进制区分矩阵

	a	b	c	d
$x_1, x_3$	1	0	0	0
$x_1, x_4$	1	1	0	0
$x_1, x_5$	1	0	1	0
$x_1, x_7$	1	0	1	1
$x_1, x_8$	0	1	1	0
$x_1, x_9$	0	1	1	1
$x_1, x_{10}$	1	1	0	1
$x_2, x_3$	1	0	0	1
$x_2, x_4$	1	1	0	1
$x_2, x_5$	1	0	1	1
$x_2, x_7$	1	0	1	0
$x_2, x_8$	0	1	1	1
$x_2, x_9$	0	1	1	0
$x_2, x_{10}$	1	1	0	0
$x_3, x_6$	1	0	1	1
$x_4, x_6$	0	1	1	1
$x_5, x_6$	0	0	0	1
$x_6, x_7$	1	0	0	0
$x_6, x_8$	1	1	0	1
$x_6, x_9$	1	1	0	0
$x_6, x_{10}$	1	1	1	0

(2) 计算各属性区分度。 $F_1(a) = 15, F_1(b) = 12, F_1(c) = 11, F_1(d) = 11$ , 因  $a$  属性区分度最大, 所以  $a$  进入约简集  $R$ , 对应的  $Q = 1000$ , “与”运算后,  $M_B$  更新如表 3 所示。

(3) 重新计算区分度。由表 3 可知,  $F_1(a) = 0, F_1(b) = F_1(c) = 5, F_1(d) = 4$ 。由于  $b, c$  属性具有相同的区分度, 所以需要比较两者的区分率。计算可得  $F_2(b) = F_2(c) = 2$ , 可见  $b, c$  属性的区分率也相等。此时, 查找用户设定的属性排名, 得  $b > c$ , 所以选择属性  $b$  进入约简集  $R$ 。  $Q = 1100$ , “与”运算后,  $M_B$  更

新为表 4。

表 3 第一次更新后的二进制区分矩阵

	$a$	$b$	$c$	$d$
$x_1, x_8$	0	1	1	0
$x_1, x_9$	0	1	1	1
$x_2, x_8$	0	1	1	1
$x_2, x_9$	0	1	1	0
$x_4, x_6$	0	1	1	1
$x_5, x_6$	0	0	0	1

表 4 第二次更新后的二进制区分矩阵

	$a$	$b$	$c$	$d$
$x_5, x_6$	0	0	0	1

(4) 再次计算区分度。由表 4 可知,  $F_1(a) = F_1(b) = F_1(c) = 0, F_1(d) = 1$ 。选择属性  $d$  进入约简集  $R$ , 对应的  $Q = 1101$ , 执行“与”运算后,  $M = \emptyset$ , 程序终止。得出天气预报决策表的一个考虑了属性排名的约简集为  $\{a, b, d\}$ , 即  $\{\text{Outlook}, \text{Temp}, \text{Wind}\}$ 。

## 5 结束语

文献[10]提出了一种基于二进制区分矩阵以属性区分度和区分率作启发信息的约简算法, 但在区分度和区分率均相同时, 会得到有随机性的约简集。文中引入了用户设定的属性排名, 使得最终的约简集不但确定, 而且满足用户需求, 进一步完善了基于属性区分度和区分率的约简算法。

(上接第 81 页)

型构造所需解决的问题等方面研究了基于反馈的网格技术在高职共享资源建设中应用问题, 并就网格资源共享的系统的实现过程进行了讨论。

## 参考文献:

- [1] 张 炜, 丁振国. 基于信息网络的教育资源共享模型研究[J]. 微电子学与计算机, 2008(3): 120-126.
- [2] 都志辉, 陈 渝, 刘 鹏. 网格计算[M]. 北京: 清华大学出版社, 2002.
- [3] 柯和平, 李春林. 基于网格技术的区域性教育资源库共建共享机制研究[J]. 电化教育研究, 2008(1): 42-46.
- [4] 马常霞. 校园网格环境构建的关键技术研究[J]. 计算机技术与发展, 2006, 16(1): 48-50.
- [5] 程宏兵, 杨 庚. 基于网格技术的资源调度模型[J]. 计算机应用, 2006, 33(8): 2086-2090.
- [6] Intanagonwivat C, Govindan R, Estrin D. Directed diffusion for wireless sensor networking[J]. IEEE/ACM Trans. on

## 参考文献:

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982(11): 341-356.
- [2] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.
- [3] 刘 清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.
- [4] 王国胤, 姚一豫, 于 洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1245.
- [5] Li Zenquan. Suitability of fuzzy reasoning methods[J]. Fuzzy Sets and Systems, 1999, 108(3): 299-311.
- [6] 王小菊, 蒋 芸, 李永华. 基于依赖度之差的属性重要性评分[J]. 计算机技术与发展, 2009, 19(1): 67-70.
- [7] Smyth P, Goodman R M. An information theoretic approach to rule induction from databases[J]. IEEE Trans. Knowledge Discovery Data Mining, 1992(4): 301-316.
- [8] 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.
- [9] Wang G Y. Rough reduction in algebra view and information view[J]. International Journal of Intelligent System, 2003, 18(5): 679-688.
- [10] 杨 萍, 李济生, 黄永宣. 一种基于二进制区分矩阵的属性约简算法[J]. 信息与控制, 2009, 38(1): 70-74.
- [11] Han S Q, Wang J. Reduct and attribute order[J]. Journal of Computer Science and Technology, 2004, 19: 429-449.
- [12] Felix R, Ushio T. Rough sets-based machine learning using a binary discernibility matrix[C]// Proceeding of the Second International Conference on Intelligent Proceeding and Manufacturing of Materials. [s.l.]: [s.n.], 1999: 299-305.
- [7] 刘广帅. 网格技术在校资源中的应用[J]. 电脑与电信, 2007(4): 70-75.
- [8] 苟和平, 冯百明, 景永霞. 一种基于信息网格的多源信息集成方案[J]. 微电子学与计算机, 2008(11): 72-75.
- [9] 李学俭, 何文华. 基于 SOA 架构的高校数据资源整合研究[J]. 计算机技术与发展, 2010, 20(1): 78-81.
- [10] Foster I, Kesselman C, Nick J M, et al. Grid Services for Distributed System Integration[J]. Computer, 2002(6): 37-46.
- [11] 王小君, 何 庆. 资源网格中的一种资源检索机制[J]. 计算机技术与发展, 2010, 20(3): 63-66.
- [12] Buyya R, Abramson D, Giddy J. Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid[C]// In Proceedings of the HPC ASIA'2000, the 4th International Conference on High Performance Computing in Asia-Pacific Region. USA: IEEE Computer Society Press, 2000: 1-7.