

基于RS理论的快速属性约简求核方法

代广珍^{1,3}, 徐超²

- (1. 安徽工程大学 电气工程学院, 安徽 芜湖 241000;
2. 安徽大学 电子科学与技术学院, 安徽 合肥 230039;
3. 安徽工程大学 电气传动与控制省级重点实验室, 安徽 芜湖 241000)

摘要:粗糙集是用来处理不确定、不完备数据的重要工具之一。属性约简是粗糙集理论研究的一个重要内容,属性核则是属性约简所涉及的一个重要概念,对简化属性约简具有不可替代的重要性。文中指出属性约简的作用,及其涉及到的一个重要概念——属性核的概念和重要性。分析了目前常用求属性核方法,大都建立在内存中,需要构造差别矩阵,存在时空复杂度较大的不足。提出了一种无需建立差别矩阵的简单求属性核方法,并通过实例验证了正确性。

关键词:属性约简;属性核;差别矩阵

中图分类号:TP18

文献标识码:A

文章编号:1673-629X(2011)04-0133-03

A Rapid Approach to Compute Attribute Core of Attribute Reduction Based on RS

DAI Guang-zhen^{1,3}, XU Chao²

- (1. School of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China;
2. School of Electronic Science and Technology, Anhui University, Hefei 230039, China;
3. Anhui Provincial Key Laboratory of Electrical Transmission and Control, Wuhu 241000, China)

Abstract: Rough set theory is a new mathematical tool to research imprecise and incomplete data. Attribute reduction refers to delete unrelated or unimportant knowledge with keeping the ability of knowledge classification. Attribute core is an important concept of attribute reduction and has irreplaceable significance to simplify attribute reduction. Attempts to point out the role of attribute reduction and analyze an important conception concerned and its importance—attribute core. Methods used now almost stay in memory, and need to create discernibility matrix, so that it may lead to some demerits as time and space complex. Based on these defaults, an easier way is given to compute attribute core without creating discernibility matrix and verified by an example.

Key words: attribute reduction; attribute core; discernibility matrix

0 引言

信息化带来的数据爆炸,使得数据不确定性更加显著。如何从这些模糊的、不精确的、不完整的大量信息中获取有价值的知识对智能信息处理提出了严峻挑战。粗糙集理论应运而生,它不仅能够处理复杂系统中的数据和信息,也可处理模糊的和不确定信息^[1]。在粗糙集理论中,求核及属性约简都是非常关键的步骤,也是粗糙集理论和应用研究的焦点问题之一^[2,3]。属性约简是指删除知识库中不相关或不重要的知识而不影响知识库的分类能力。在属性约简的过程中,涉及到的一个重要概念是属性核(core),核的重要性体现

在:核是所有属性约简的交集,是指在属性约简中不能被删除的知识库特征属性的集合。目前,求取属性核的方法有很多,常用的有:基于差别矩阵的求核方法和基于信息系统理论的条件熵求核方法^[4-10,12,13]。这些方法一般都需要占用大量的内存空间或需要经过繁琐的数学运算。文中在以前的研究者研究基础上,提出了一种不需要占用大量的内存空间或繁琐的数学运算,而只需经过简单的比较判断就可以快速求取属性核的方法,并通过实例验证了求解的正确性。

1 属性核快速求取方法

1.1 基本概念

差别矩阵。

波兰数学家 Skowron 提出差别矩阵^[6]定义如下:

定义1:决策信息系统 $S = (U, A)$, 其中 $U = \{x_1,$

收稿日期:2010-09-09;修回日期:2010-12-10

作者简介:代广珍(1975-),男,安徽巢湖人,讲师,研究方向为数据挖掘、粗糙理论;徐超,教授,研究方向为网络与智能信息系统嵌入式系统。

x_2, x_3, \dots, x_n 称为论域, 表示讨论的对象非空有限集合; $A = C \cup D$ 为对象的属性集合, C 为对象的条件属性的非空有限集合, D 为对象的决策属性的非空有限集合, $C \cap D = \emptyset$ 时, S 又称为决策系统或决策表, $a(x)$ 是对象 x 在属性 a 上的取值, 差别矩阵定义为:

$$(C_{ij}) = \begin{cases} 1 & \{a \in A \mid a(x_i) \neq a(x_j)\} \quad D(x_i) \neq D(x_j) \\ 0 & D(x_i) = D(x_j) \\ -1 & a(x_i) = a(x_j), D(x_i) \neq D(x_j) \end{cases}$$

运用差别矩阵进行属性约简时, 首先要根据决策表构建差别矩阵, 而构建差别矩阵时空复杂性较大, 且无法利用数据库查询语言的优势; 文献[5]分析提出属性约简必须从属性核开始的结论; 文献[11]分析指出没有必要求出属性核, 只需要使得所设计的约简算法能够确保约简结果中包含属性核。另外, 在构建差别矩阵时, 由于针对决策属性属于同一类的对象, 差别矩阵中相应的元素为空(用 0 表示), 也就是说, 只有不属于同一类的记录相应的元素才不为 0。因此, 可以通过决策属性来将 U 分为正例 P 和反例 N 两类, 同一类中各对象的关系相对于差别矩阵中为 0 的元素; 差别矩阵中不为 0 的元素是可以用来区分两类中的相应对象的属性。

由此, 文中给出以下定义。

定义 2: 决策信息系统 $S = (U, A, V, F)$, 其中 $U = \{x_1, x_2, x_3, \dots, x_n\}$ 称为论域, 表示讨论的对象非空有限集合; $A = C \cup D$ 为对象的属性集合, C 为对象的条件属性的非空有限集合, D 为对象的决策属性的非空有限集合, 且 $C \cap D = \emptyset$; $V = \bigcup_{p \in A} V_p$, V_p 为对象的某个属性 p 取值的集合; F 表示论域中对象的属性与属性值的对应关系, 即 $\forall p, x, p \in A, x \in U$, 有 $f(x, p) \in V_p$; 属性集 $P \subseteq A$ 在 U 上导出的划分 $U/\text{IND}(P) = \{X_1, X_2, X_3, \dots, X_m\}$, 称为一个不可分辨关系族, 其中的任何一个元素 $[x]_P = \{x_i \mid \forall a \in P, f(x_i, a) = f(x, a), i \neq j\}$ 称为一个等价关系或等价类。

定义 3: 对于决策信息系统 S , 假设有 $U/\text{IND}(D) = \{P, N\}$, 其中 P 称为正例集, N 称为反例集, 则能够区分 P 和 N 中对象的属性集为:

$$Q_{ij} = \{a_k \mid f(x_i, a_k) \neq f(y_j, a_k), x_i \in P, y_j \in N, a_k \in C\}$$

其中, $k = 1, 2, \dots, |C|$; $i = 1, 2, \dots, |P|$; $j = 1, 2, \dots, |N|$, $|C|$ 表示条件属性数, $|P|$ 表示正例中对象数, $|N|$ 表示反例中对象数, $f(x_i \text{ or } y_j, a_k) \in V_{a_k}$ 表示正例或反例中的某个对象在第 k 个属性的取值。

1.2 求核方法分析

文献[9]在分析了 HU 算法后, 指出其利用改进的

差别矩阵来确定核方法的缺陷, 并举出反例运用 P -近似精度来计算说明决策表 1 中的单属性 $C2$ 为一个属性约简。

表 1 决策表

对象	属性			
	C1	C2	C3	D
x1	1	0	1	1
x2	1	0	1	0
x3	0	0	1	1
x4	0	0	1	0
x5	1	1	1	1

P -近似精度公式如下:

$$\gamma_p = \sum_{i=1}^k \text{card}(P_{X_i}) / \text{card}(U)$$

其中, X_i 表示由 D 导出的等价类构成 U 的一个划分: $\{X_1, X_2, \dots, X_k\}$ 的一个子集, P_{X_i} 表示 X_i ($X_i \subseteq U$) 关于 P ($P \subseteq C$) 的下近似: $P_{X_i} = \{x \in U \mid [x]_P \subseteq X_i\}$, $\text{card}(\cdot)$ 表示集合的基数。将一个近似精度等于条件属性 C 的近似精度的单属性看是一个属性约简的说法, 与属性核的定义, 即属性集的所有约简的交集称为属性核不一致。

另外, 文献[9]分析了反例中相对决策属性存在的不相容性可能导致 HU 法的欠缺, 于是删除反例中的记录 x_1 , 作者后来指出尽管存在不相容性, 但求核方法是对的, 对于不存在不相容性的情况, 作者并没有给出详细的分析, 而按照作者的分析方法, 删除不相容记录的情况也有多种。例如, 删除记录 x_2 后得差别矩阵与删除 x_1 后所得的差别矩阵(为了简洁起见, 矩阵中只列出满足条件 $\{a \in A \mid a(x_i) \neq a(x_j)\}$, 且 $D(x_i) \neq D(x_j)$ 的元素)对照如下:

删除记录 x_2 的差别矩阵

	x1	x3	x4
x1	\hat{C}		\hat{D}
x3	\hat{C}		\hat{D}
x4	\hat{C}		\hat{D}

删除记录 x_1 的差别矩阵

	x2	x3	x4
x2	\hat{C}		\hat{D}
x3	\hat{C}		\hat{D}
x4	\hat{C}		\hat{D}

显然, 由表 2 可知, 删除记录 x_2 后的决策表的属性核为 $\text{CORE}(C) = \{C1\}$, 而新的决策表中仍然存在不相容性, 这与文献[9]指出的“尽管存在不相容性, 但求核方法是对的”不一致。而完全消除不相容性的情况更多, 如下所示:

删除记录 x_1 、 x_3 的差别矩阵

	x_1	x_2	x_3
x_1	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_2	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_3	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	(C_1, C_2)	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

删除记录 x_1 、 x_4 的差别矩阵

	x_1	x_2	x_3
x_1	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_2	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	(C_1)	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_3	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	(C_2)	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

删除记录 x_2 、 x_3 的差别矩阵

	x_1	x_2	x_3
x_1	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_2	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	(C_2)	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_3	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	(C_1, C_2)	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

删除记录 x_2 、 x_4 的差别矩阵

	x_1	x_2	x_3
x_1	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_2	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
x_3	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

由以上各矩阵可以看出,不相容性对求取属性核是存在影响的。

为了避免陷入数据不相容性对求取属性核和属性约简影响的困惑,以及通常的求核方法都建立在内存中构造不差别矩阵时空复杂度较大或需要大量的计算的情况,提出了一种简单的求核方法,具体算法描述见1.3节,运用该法不需要构造不差别矩阵,也不需要进行大量的计算。由 Skowron 提出的差别矩阵可知,矩阵中属性组合数为1的元素唯一区分该元素所对应的两个对象,该属性在属性约简中必须保留,即为 S 的核属性。由此,给出下面的定理^[6,9]。

定理:对于决策信息系统 S ,如果 $\exists a \in Q_{ij} \cap Q_{i'j'}$, 且 $|Q_{ij} \cap Q_{i'j'}| = 1$, 则 $a \in \text{CORE}_D(C)$ 。其中, $\text{CORE}_D(C)$ 为 S 的属性核, $|Q|$ 表示属性集合 Q 中元素的个数。

推论:假设存在属性 Q_{ij} , 使得 $|Q_{ij}| = 1$, 则有 $Q_{ij} \subseteq \text{CORE}_D(C)$ 。

显然,由差别矩阵求核方法可以得证。

1.3 算法描述

算法的伪代码描述如下:

INPUT: $S = (U, C \cup D)$

OUTPUT: $\text{CORE}_D(C)$

Step1 if $D(x_i) = p$ (p 为正例对象的决策属性值) then $P = P \cup \{x_i\}$

else if $D(x_i) = n$ (n 为反例对象的决策属性值) then $N = N \cup \{x_i\}$

其中 $i = 1, 2, \dots, |U|$;

Step2 if $x_i \in P, x_j \in P$ and $f(x_i, a_k) = f(x_j, a_k)$ then

$P = P - \{x_j\}$

其中 $i, j = 1, 2, \dots, |P|, k = 1, 2, \dots, |C|$;

else if $x_i \in N, x_j \in N$ and $f(x_i, a_k) = f(x_j, a_k)$ then

$N = N - \{x_j\}$

其中 $i, j = 1, 2, \dots, |N|, k = 1, 2, \dots, |C|$;

Step3 if $x_i \in P, y_j \in N$ and $f(x_i, a_k) \neq f(y_j, a_k)$

then $Q_{ij} = Q_{ij} \cup \{a_k\}$

其中 $k = 1, 2, \dots, |C|, i = 1, 2, \dots, |P|, j = 1, 2, \dots, |N|$;

Step4 if $|Q_{ij} \cap Q_{i'j'}| = 1$ or $|Q_{ij}| = 1, i$ and $i' = 1, 2, \dots, |P|, j$ and $j' = 1, 2, \dots, |N|$

then $\text{CORE}(C) = \text{CORE}(C) \cup ((Q_{ij} \cap Q_{i'j'}) \cup Q_{ij})$

2 实例分析

仍然以上面的决策表1为例,通过运用上述算法来计算属性核以验证算法的正确性。

由表1求得正例集 P 和反例集 N 分别为:

表2 正例集 $P(D=1)$

对象	属性		
	C_1	C_2	C_3
x_1	1	0	1
x_3	0	0	1
x_5	1	1	1

表3 反例集 $N(D=0)$

对象	属性		
	C_1	C_2	C_3
x_2	1	0	1
x_4	0	0	1

显然,能够区分正例集和反例集中的各对象的属性集分别为:

$Q_{11} = \Phi, Q_{12} = \{C_1\}, Q_{21} = \{C_1\}, Q_{22} = \Phi, Q_{31} = \{C_2\}, Q_{32} = \{C_1, C_2\}$;

最后得到决策表1的属性核为: $\text{CORE}(C) = \{C_1, C_2\}$ 。可见,求得的结果与叶方法和杨方法的结果是一致的,说明算法是有效的。

3 结束语

属性约简能够去除冗余的属性,从而为保持知识库分类能力不变的情况下,尽量为减小时空复杂性和计算量提供了可能。由于属性核是约简后所得的属性集必然包含的,在去除属性核后的属性中,寻找辨识决策属性导出的不同等价类中的对象,问题规模相对变小,因此操作也变得相对简单。文中通过文献[6]中的实例1验证了提出的求核方法,并运用在旅游电子商务平台的属性约简及分类决策中,取得了许多有价

(下转第140页)

时,合理设置临时数组的大小,提高了系统性能。例如:格构建系统中用于存放中间元素属性关系的二维数组 pro_all_k ,列数 n 即属性个数,行数 m 即元素的个数, m 的理论最小值为 ele_num ,理论最大值 $C_{ele_num}^2 + ele_num$ 即 $(1/8) * (ele_num^2 - ele_num)$ 。

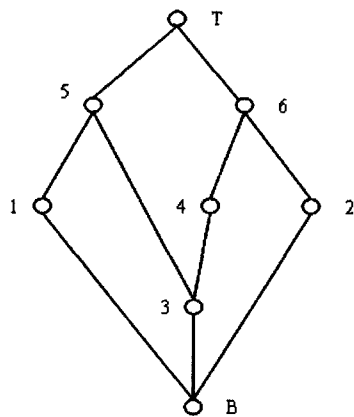


图 6 概念格图

3、数据类型。系统中的数组都采用 int 类型。

4、临时计数器。在系统中设置临时计数器降低了系统的时间复杂度。

3 结束语

文中借助概念格这一数学理论的指导,在 Eclipse 平台下设计一个概念格构造算法的插件,半自动地实现并生成概念格图。为以后各种运用形式概念分析方法的应用^[10-12]提供了一种较为直观的基础保障。

(上接第 135 页)

值的信息。对文中提出的求核算法,虽然快捷有效,但在属性和数据对象增加及其对该算法的影响方面还有待研究,后期工作将继续增强算法的普适性和健壮性。

参考文献:

- [1] 苗夺谦,李道国.粗糙集理论、方法与应用[M].北京:清华大学出版社,2008.
- [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] 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.
- [4] Hu Xiaohua, Cercone N. Learning in relational data-bases: a rough set approach[J]. Computational Intelligence, 1995, 11(2):323-337.
- [5] Wong S K M, Ziarko W. On Optimal Decision Rules in Decision Tables[J]. Bulletin of Polish Academy of Science, 1985, 33:693-696.
- [6] Skowron A, Rauszer C. The Discernibility Matrices and functions in Information System[M]//Intelligent Decision Support

参考文献:

- [1] Wille R Restructuring lattice theory: an approach base on hierarchies of concepts[M]. Ordered Sets. Reidel, Dordrecht-Boston; [s. n.], 1982.
- [2] Carpineto C, Romano G. A lattice conceptual clustering system and its application to browsing retrieval[J]. Machine Learning, 1996, 24(2):95-122.
- [3] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M]. Berlin: [s. n.], 1999.
- [4] 刘亚滨,杨红.精通 Eclipse[M].北京:电子工业出版社,2004:2-5.
- [5] 刘甫迎,谢春,徐虹. Java 程序设计实用教程[M].北京:科学出版社,2005:85-280.
- [6] 段友祥,郭辉,林桂华.一种通过交互界面实现用例抽取算法的研究与应用[J].计算机技术,2006,1(2):27-29.
- [7] Lindig C. Fast Concept Analysis[EB/OL]. 2002. <http://www.st.cs.uni-lindig.pdf>.
- [8] Stumme G, Maedche A. FCA-merge: bottom-up merging of ontologies[C]//17th Intel. Conf. on Artificial Intelligence(IJCAI'01). Germany:Springer,2001:225-230.
- [9] 陈玲.浅谈算法的复杂性和常用算法[J].教育现代化,2004,3(2):66-67.
- [10] 蒋平,任胜兵,林娟.形式概念分析在软件工程中的应用[J].计算机技术与发展,2008,18(4):127-129.
- [11] 张涛,周爱武,谢荣传.基于概念格和关联规则 Web 个人化系统[J].计算机技术与发展,2008,18(2):139-142.
- [12] 胡可云,陆玉昌,石纯一.基于概念格的分类和关联规则的集成挖掘方法[J].软件学报,2000,11(4):1478-1484.

Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publishers, 1992:331-362.

- [7] 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,1999,11(11):1206-1211.
- [8] 李佩,刘玉树,王蕾.一种粗糙集属性约简算法[J].计算机工程与应用,2002(5):15-16.
- [9] 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.
- [10] 杨明,孙志挥.改进的差别矩阵及其求核方法[J].复旦学报:自然科学版,2004,43(5):865-868.
- [11] 乔梅,韩文秀.基于 Rough 集和数据库技术的属性约简算法[J].计算机工程,2005(6):18-19.
- [12] 杨飞,代广珍.属性约简在高校就业决策分析中的应用[J].计算机技术与发展,2007,17(7):223-225.
- [13] 赖桃桃,冯少荣,张东站.基于改进差别矩阵的核增量式更新算法[J].计算机应用,2009(9):2477-2480.