一种改进的基于差别矩阵的属性约简方法

赵荣利,崔志明,陈建明
(苏州大学 智能信息处理及应用研究所,江苏 苏州 215006)

摘 要:粗糙集理论作为一种新型的软计算方法,在数据挖掘方面的应用越来越被人们所重视。利用粗糙集理论进行数据挖掘,得到知识规则,最重要的一点就是基于粗糙集的属性约简。文中在区分矩阵的基础上,改进了计算信息系统属性约简的方法,使属性约简计算量大幅度减小,可以快速得到给定要求下的属性约简。

关键词:粗糙集;属性约简;区分矩阵
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2006)11-0032-02

An Improved Reduction Method of Attribution Based on Discernable Matrix

ZHAO Rong-li, CUI Zhi-ming, CHEN Jian-ming
(The Institute of Intelligent Information Processing and Application, Suzhou University, Suzhou 215006, China)

Abstract: Rough set is a new method of soft calculation. It is being recognized gradually to use in date mining. Utilizing rough set executes data mining and obtains knowledge, attribution reduction is very important. In this paper, an improved reduction method of attribution for information system is proposed. It is based on discernable matrix. It largely makes reduction simply and convenient, and can obtain the attribution reduction fastly.

Key words: rough set; attribution reduction; discernable matrix

0 引 言

粗糙集理论是波兰数学家 Z. Pawlak 于 1982 年首先提出来的一种分析数据的数学理论。该理论是一种处理具有信息不确定、不精确、不完善系统的新数学工具,它提供了一整套方法,从数学上严格地处理数据分类问题[1]。粗糙集理论是目前使用较多的一种归纳学习方法,已应用于机器学习、知识发现、数据挖掘、决策支持与分析、专家系统、归纳推理和模式识别等许多科学和工程领域[2]。

数据挖掘是从大量数据中提取或"挖掘"知识。而这正是粗糙集中约简所要达到的目的。知识约简,就是在保持知识库分类能力不变的条件下,删除其中不相关的或不重要的知识[3]。所以利用粗糙集理论进行数据挖掘,得到知识规则,最重要的一点就是基于粗糙集的属性约简(知识约简)。目前已提出了若干个求属性约简的算法[4,5]。

文中提出的属性约简基于 Rough Set 理论,引入叶东毅等提出的一个新的差别矩阵,使属性约简计算量大幅度减小,快速得到给定要求下的属性约简。

收稿日期:2006-03-01
基金项目:苏州市 2004 年度科技攻关项目(SG0406)
作者简介:赵荣利(1981-),女,河南洛阳人,硕士研究生,研究方向为数据挖掘、智能化信息处理;崔志明,教授,博士生导师,研究方向为智能化信息处理、计算机网络应用与数据库应用;陈建明,副教授,研究方向为计算机网络应用和数据库应用。

1 张文修等属性约简方法的不足

决策表可定义为 $S=(U,A,V,f)$,其中 U:对象的非空有限集合,称为论域,$A=C\cup D$,$C\cap D=\varnothing$,C 称为条件属性集,D 称为决策属性集;V:属性值的集合;f:$U\times A\to V$ 是一个信息函数。

区分矩阵:令 $\boldsymbol{S}=(U,A,V,f)$ 是一个决策表,$|U|=n$,$\boldsymbol{S}$ 的区分矩阵是一个 $n\times n$ 矩阵,其任一元素为 $a(x,y)=\{a\in A\mid f(x,a)\neq f(y,a)\}$,$a(x,y)$ 就是区别记录 x 和 y 的所有属性的集合。

区分函数:引入一个布尔函数,称其为区分函数,用 $\triangle$ 表示,对每个属性 $a\in A$,指定一个布尔变量 a,若 $a(x,y)=\{a_1,a_2,\cdots,a_k\}\neq\varnothing$,则指定布尔函数 $a_1\vee a_2\vee\cdots\vee a_k$,用 $\sum a(x,y)$ 来表示;若 $a(x,y)=\varnothing$,则指定布尔变量为 1。(布尔)区分函数 $\triangle$ 可定义为:$\triangle=\prod_{(x,y)\in U\times U} a(x,y)$。

区分函数 $\triangle$ 有如下性质:函数 $\triangle$ 的极小析取范式中的所有合取式是属性集 A 的所有约简。

文献[3]提出的方法基于区分矩阵和逻辑运算,该属性约简算法可以得到信息系统的所有可能的属性约简结果,它把数学的逻辑运算公式运用到属性组合情况的搜索中,简化了属性约简,但可以明显看到,因区分矩阵中

$a(x,y) \neq \varnothing$ 的项很多,从而得到的区分函数中析取表达式很多,导致了化简这些逻辑公式变得异常复杂,计算量很大。但可以发现 $a(x,y) \neq \varnothing$ 的项有很多是重复的,这引发笔者考虑能否采取一定的变通方法把区分矩阵进行等价约简,简化约简过程,从而改进此种算法的不足。

在以后的阅读中,其中叶东毅的一篇"一个新的差别矩阵及其求核方法"[6] 使笔者想到能否把求核结合到这一化简过程中。由粗糙集理论知道,任何决策表的相对核是唯一的,并且它包含在所有的相对约简中[3],所以把相对核作为属性约简的起点是正确的。也通过实例验证了此方法的正确性和简便高效性。

2　改进的基于差别矩阵的属性约简方法

文中方法把张文修和叶东毅提出的思想有效结合,极大简化了约简过程,使属性约简变的简便高效。方法基本思路:先利用叶东毅等提出的方法求取决策系统的相对属性核,以核作为约简的起点,使区分矩阵得到极大化简,从而得到新的化简了的区分矩阵,再进一步执行约简运算。

2.1　叶东毅提出的新的差别矩阵求核方法

(1) Hu 算法的缺陷。

Hu 等学者提出了简洁的利用差别矩阵来确定核的方法,其中改进的差别矩阵 $\boldsymbol{M} = \{m_{ij}\}$ 定义为:

$$m_{ij} = \begin{cases} a \in C: f(x_i,a) \neq f(x_j,a), \\ \quad \text{当 } f(x_i,D) \neq f(x_j,D) \text{ 时} \\ \text{空集,当其它情况时} \end{cases} \tag{1}$$

由以上定义的差别矩阵,Hu 文献中未加证明的给出下面结论:当且仅当某个 m_{ij} 为单个属性时,该属性属于核 core(C),该结论仅在某些情况下是正确的,文献[6] 给出了正反例。

针对 Hu 方法的 缺陷,文献[6] 提出了新的差别矩阵定义并给出了求核的方法。

(2) 新的差别矩阵及求核方法。

定义:对于给定的信息系统(I),定义差别矩阵 $\boldsymbol{MS} = \{m'_{ij}\}$ 为(其中 m_{ij} 如公式(1) 所定义):

$$m'_{ij} = \begin{cases} m_{ij}, \min\{|d(x_i)|, |d(x_j)|\} = 1 \\ \text{空集,当其它情况时} \end{cases} \tag{2}$$

文献[6] 给出并证明了如下结论:定理:对于给定的信息系统(I),若记 $\boldsymbol{SM}(C) = \{m'_{ij}, m'_{ij}$ 为单个属性$\}$,则有 $\boldsymbol{SM}(C)$ = core(C),即当且仅当某个 m'_{ij} 为单个属性时,该属性属于核 core(C)。

2.2　改进的属性约简方法的具体步骤

设 $\boldsymbol{M}$ 是决策表 T 的区分矩阵,$C = \{c_1, c_2, \cdots, c_k\}$ 是 T 中所有条件属性的集合,N 是 M 中所有属性组合的集合,设 N 中包含有 n 个属性组合,每个属性组合表示为 N_i。

设 Card(N_i) = m,则 N_i 中每个条件属性表示为:$n(i,k) \in N_i (k = 1,2,\cdots,m)$。

以相对核为起点,基于区分矩阵的属性约简算法分为以下几步:

第一步:写出决策表的区分矩阵 Tab1。

第二步:依据上面提出的公式(2) 计算出决策表的相对核,记为 C_0 = core(C)。

第三步:在区分矩阵中找出所有包含核属性 core(C) 的属性组合 N,用数学公式描述为:$N = \{N_i, N_i \cap C_0 \neq \varnothing, i = 1,2,\cdots,n\}$,在区分矩阵中,把所有这些包含核属性的属性组合 N_i 的值都修改为 0。

第四步:选取区分矩阵中的所有非空属性组合的元素 $N_i (N_i \neq 0, N_i \neq \varnothing)$,表示成相应的析取逻辑表达式 $L_{ij} = \sum n(i,k)\ (n(i,k) \in N_i)$。

第五步:把所有的析取逻辑表达式 L_{ij} 进行合取运算,得到一个合取逻辑表达式 L 即:$L = \wedge L_{ij}$。

第六步:将合取逻辑表达式 L 转化为析取逻辑表达式,即得到属性约简的结果。析取逻辑表达式中的每一个合取项就对应一个属性约简。

3　实例比较

设有一决策表如表 1 所示。

表 1　决策表

U/A	x_1	x_2	x_3	x_4	x_5	y
1	c	6	y	E	h	m
2	c	6	n	E	h	h
3	c	6	n	E	lo	m
4	c	4	y	E	h	h
5	s	4	n	B	lo	h
6	c	4	n	E	lo	m
7	c	4	n	B	m	m
8	s	4	n	E	h	h
9	s	4	n	E	m	h
10	c	4	n	B	h	m

由上表知本决策表:$A = C \cup D$,条件属性 $C = \{x_1, x_2, \cdots, x_5\}$,决策属性 $D = \{y\}$。

对应上表的区分矩阵如表 2 所示。

表 2　对应表 1 的区分矩阵

1	2	3	4	5	6	7	8	9	10
1									
2 $x3$									
3	$x5$								
4 $x2$		$x2x3x5$							
5 $x1x2x3x4x5$		$x1x2x4$							
6	$x2x5$		$x3x5$	$x1x4$					
7	$x2x4x5$		$x3x4x5$	$x1x5$					
8 $x1x2x3x5$		$x1x2$			$x1x5$	$x1x4$			
9 $x1x2x3x5$		$x1x2x5$			$x1x5$	$x1x4$			
10	$x2x4$		$x3x4$	$x1x5$			$x1x4$	$x1x4$	$x5$

利用文献[3] 的属性约简方法:

$\triangle = x_3 \wedge x_5 \wedge x_2 \wedge (x_2 \vee x_3 \vee x_5) \wedge \cdots \wedge (x_1 \vee x_4) \wedge (x_1 \vee x_4 \vee x_5)$

此处约简计算将异常复杂。

(下转第 37 页)

行时间为(8,16) - (4,8) = (4,8)。如果对变迁序列 $\delta = t_1, t_4, t_6, t_5, t_7, t_8$ 进行同样的分析,可以发现,由于变迁 t_6 是不可调度的,因此,该序列成为不可调度。

表 1　变迁的时间约束

变迁	EFT	LFT
t_1	1	2
t_2	1	2
t_4	2	4
t_5	1	2
t_6	5	7
t_7	1	2
t_8	2	4

表 2　可调度性分析过程

序号	时间戳	状态 M	D_i	AD_i	RT	TS
0	(0,0)	$\{P_0\}$	$\{t_1(1,2)\}$	$\{t_1(1,2)\}$	(1,2)	t_1 : (1,2)
1	(1,2)	$\{P_2\}$	$\{t_2(1,2), t_4(2,4)\}$	$\{t_2(2,4), t_4(3,6)\}$	(2,2)	t_4 : (3,4)
2	(3,6)	$\{P_4, P_5\}$	$\{t_5(1,2), t_6(5,7)\}$	$\{t_5(4,8), t_6(8,13)\}$	(1,2)	t_5 : (4,8)
3	(4,8)	$\{P_5, P_6\}$	$\{t_6(3,6)\}$	$\{t_6(7,14)\}$	(3,6)	t_6 : (7,14)
4	(7,14)	$\{P_6, P_7\}$	$\{t_7(1,2)\}$	$\{t_7(8,16)\}$	(1,2)	t_7 : (8,16)
5	(8,16)	$\{P_8\}$	$\{t_8(2,4)\}$	$\{t_8(10,20)\}$	(2,4)	t_8 : (10,20)

5　结　论

将时间参数引入工作流模型中并对模型进行时间维上的分析是工作流技术的重要研究内容。文中采用工作流的时间约束 Petri 网模型,提出了工作流过程可调度性分析的方法和具体的实现算法,将可调度性分析与活动执行时间的估计结合起来,为工作流的高效执行提供了理论和方法上的支持。

参考文献:

[1] 罗海滨,范玉顺,吴　澄. 工作流技术综述[J]. 软件学报,2000,11(7):899 - 907.

[2] 潘启澍,姜　兵. 基于 Petri 网的工作流建模技术及应用[J]. 清华大学学报,2000,4(9):86 - 89.

[3]] 李慧芳, 范玉顺. 工作流系统时间管理[J]. 软件学报,2002,13(4):1552 - 1557.

[4] Eder J, Panagos E. Time management in workflow systems [C]//Proc. International conference in business information systems. Heidelberg, Berlin: Springer - Verlag, 1999: 265 - 280.

[5] 李建强,范玉顺:工作流模型时间有界性验证与分析研究[J]. 计算机集成制造系统,2002,8(10):770 - 775.

[6] Ver Der Aalst W M P. Verification of workflow task structures [J]. Information systems,2000,25(1):43 - 69.

[7] Bowden E D J. A brief survey and synthesis of the roles of time in Petri nets[J]. Mathematical and computer modeling,2000, 31(10 - 12):55 - 68.

[8] Tsai J J P. Yang S J. Timing constraint Petri nets and their application to schedulability analysis of real - time system specifications[J]. Transactions on the software engineering,1995,21 (1):32 - 49.

(上接第 33 页)

但利用文中所提出的方法,首先利用公式(2) 找到相对核 Core(C) = $\{x_2x_3x_5\}$,依次文中的算法,上面的区分矩阵可化简如表 3 所示。

表 3　简化后的区分矩阵

	1	2	3	4	5	6	7	8	9	10
1										
2	0									
3		0								
4	0		0							
5	0		0							
6		0		0	x1x4					
7		0		0	0					
8	0		0			0	x1x4			
9	0		0			0	x1x4			
10		0		0	0			x1x4	0	

则得到的析取逻辑表达式得到极大简化,为:$x_1 \vee x_4$。

最后把所有的析取逻辑表达式进行合取运算则仍为:$x_1 \vee x_4$。

可以很容易地得到此决策表的属性约简为:$\{x_1, x_2, x_3, x_5\}$ 和$\{x_4, x_2, x_3, x_5\}$。

4　结束语

基于 Rough Set 的机器学习理论是数据挖掘的一个很重要的方法。简化和改善 Rough Set 的属性约简和决策规则约简,是提高数据挖掘能力和效率的重要途径。

文中在区分矩阵的基础上,提出了一种新的改进的计算信息系统属性约简的方法。该方法把数学逻辑运算运用于区分矩阵中相对核属性以外的其它属性组合,并以求核为起点化简了区分矩阵,从而使属性约简计算量大幅度减小,快速得到给定要求下的属性约简。

参考文献:

[1] Dutsch I. A Logic for Rough Sets[J]. Theoretical Computer Science(B),1997,179:427 - 436.

[2] 韩祯祥,张　琦,文福拴. 粗集理论及其应用综述[J]. 控制理论与应用,1999,16(2):153 - 156.

[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] 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.

[5] Wang Hui, Duntsch I, Gediga G. Classificatory filtering in decision systems[J]. International Journal of Approximate Reasoning,2000,23:111 - 136.

[6] 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.