第 25 卷第 1 期　　苏州科技学院学报(自然科学版)　　Vol.25 No.1
2008 年 3 月　　Journal of University of Science and Technology of Suzhou (Natural Science)　　Mar. 2008

基于关联规则挖掘的粗糙集属性值约简算法研究

杜　跃, 王治和, 景永霞
(西北师范大学 数学与信息科学学院, 甘肃 兰州 730070)

摘　要: 粗糙集理论中约简是一个重要的研究课题, 它包括属性约简和属性值约简两方面内容。针对目前属性值约简只能实现约简, 而不能计算各个规则的出现次数的问题, 结合关联规则和粗糙集两方面的优点, 对冗余规则和不一致规则进行处理, 获得具有实际意义的约简表。实验证明, 此算法是有效的。

关键词: 关联规则; 粗糙集; 值约简; 属性; 决策表
中图分类号: TP301.6　　MR(2000) Subject Classification: 65L07
文献标识码: A　　文章编号: 1672-0687(2008)01-0016-04

1　引言

粗糙集理论(Rough 集理论)是波兰人 Z.Pawlak 在 20 世纪 80 年代提出的一种新的数学工具, 它通过严格的数学公式来处理不精确、不确定的问题, 具有演绎、归纳和常识推理 3 种能力, 因此, Rough 集理论很快就在机器学习和模式识别等方面得到了广泛应用[1,2]。在粗糙集理论中约简是一个重要的研究课题, 它包括属性约简和属性值约简两方面内容, 属性值约简的目的是在保持规则集分类能力的条件下删除多余的属性值, 以便得到更简洁的规则集[3]。

文献[4]中提到的约简方法对于数据量较少的数据库是有效的, 但不适合决策表是非常大型的数据库, 所得到的规则数目也很多的情况。文献[5]中的算法是对文献[4]的改进, 但是由于没有筛选所得规则, 得到很多冗余规则。冗余规则不仅会引起资源的浪费(需要存储空间和处理时间), 而且干扰人们作出正确的决策。文献[6]中的约简方法虽然对冗余规则进行了删除, 但是对于不一致决策表却无能为力。笔者在对以上算法研究的基础上, 提出了一种利用改进的关联规则进行粗糙集属性值约简的算法。

2　基础知识[7,8]

定义 1　一个知识表示系统 S 是一个四元组 $S=(U, A, V, f)$, 其中 U 为对象的非空有限集, 称为论域; $A=C\cup D$ 是有限个属性的非空集合, 子集 C 和 D 分别为条件属性和决策属性。$V=\cup V_a$, V_a 为属性 a 的值域; $f: U\times A\to V$ 是一个信息函数。

定义 2　知识表达系统 $S=(U, A, V, f)$ 对应一张二维表, 表中每一列描述对象的一种属性 $A=C\cup D$, $C\cap D=\varnothing$, A 为非空有限集称为属性集合, C 称为条件属性集, D 称为决策属性集。每一行表示一个对象的一条由观察或测量得到的信息即该对象的各个属性值, 这样得到的表称为决策表。决策表对应一个信息库。

定义 3　相同的条件属性蕴含的决策值必须相同, 决策值函数依赖于条件属性值。因此相同的条件属性值不能决策多个决策值, 相同的决策值却可以依赖于多个不相同的条件属性值, 这样才能保证决策规则的一致性以及决策表的一致性。

[收稿日期] 2007-10-04
[作者简介] 杜　跃(1982-), 女, 辽宁阜新人, 硕士研究生, 研究方向: 数据库、数据挖掘、粗糙集。

定义 4　设 U 是一个论域, P 是定义在 U 上的一个等价关系簇, $R\in P$。如果 $IND(P\setminus\{R\})=IND(P)$, 则称关系 R 在 P 中是绝对不必要的(多余的); 否则 R 在 P 中是绝对必要的。

定义 5　在决策表中并不是每个条件属性的属性值都对该对象的分类归属起作用, 有些属性值是多余的。属性值约简的目的就是分别消去每条规则中多余的属性值, 同时在保持规则集的分类能力的条件下, 进一步简化规则集。

定义 6　决策表中支持度 s 与置信度 c 的计算公式如式(1), 其中 card( ) 表示集合中元素的个数。

$$\text{support}=\frac{\text{card}([C]\cap[D])}{\text{card}([U])}\qquad \text{confidence}=\frac{\text{card}([C]\cap[D])}{\text{card}([C])}\tag{1}$$

定义 7　决策表中得到的规则是形如 $C\Rightarrow D[s, c]$的逻辑蕴含式, 其中支持度 s 不小于用户定义的支持度阈值(min_sup), 置信度满足用户定义的置信度阈值(min_conf)。

3　实现算法描述

3.1　算法思想

3.1.1　利用关联规则挖掘算法　有时不仅要实现决策表约简, 而且要计算出各个规则的出现次数。事实上研究者不关心那些出现几率很小的决策规则。如果在属性值约简时加入用户主观信息, 将会大大增强约简的有效性。因此, 这里引入关联规则挖掘中的支持度和置信度的概念, 并对这两个概念进行了重新定义。提出一种基于关联规则挖掘思想的属性值约简算法, 删除支持度小于用户给定的支持度阈值和置信度阈值的那些规则, 得到更有效的约简表。

3.1.2　删除冗余规则　在实际应用中大家希望每一个决策类所决定的决策规则能够用最少的属性表达。但是规则 $C_1\Rightarrow D$ 和规则 $C_2\Rightarrow D$ 且有 $C_1\subset C_2$, 显然 $C_2\Rightarrow D$ 是冗余规则, 规则 $C_1\Rightarrow D$ 更具有概括性。所以在文中算法中每产生一条规则就删除决策表中能够利用该规则作出正确决策的记录, 这些被删除的记录将不再参与最简规则集的求取(频繁项集的所有非空子集都必须也是频繁的[9])。这样决策表中的记录不断减少, 可以较大的减少计算量, 加快整个决策表的属性值约简速度。在进行属性扩张时每次只能增加一个属性这样避免了属性的遗漏。

3.1.3　处理不一致规则　在进行决策表的属性值约简时, 可能得到像 $C\Rightarrow D_1$ 和 $C\Rightarrow D_2$ 的两条规则, 它们是前提重复的两条规则, 但是结论却不一致, 这样的规则会对作出正确的决定产生误导。这时可以把置信度低的规则去掉, 避免不一致规则的产生。

3.2　算法描述

第一步: 对决策表的数据进行预处理。

第二步: 关联规则的挖掘是基于布尔型的, 如果决策表的属性值是连续的, 那么要经过聚类分析对属性值归类, 然后将约简得到的决策表数据转化成布尔型的[9]。

第三步: 利用改进的关联规则进行粗糙集属性值约简, 得到所需的正确而且简洁的规则。文中改进的关联规则算法实现如下:

Input: 决策表 $S=(U, C\cup D, V, f)$; min_sup: 最小支持度阈值; min_conf: 最小置信度阈值;
Output: 最简规则集 R
$R=\phi$; //R 存放规则集
A_1=find_frequent_1- itemsets(S); //获得频繁一项集
for($i=1$; $A_i\neq\phi$; $i=i+1$)
{ if support($\cup_i C\Rightarrow D$) $\geq$ min_conf then
if (inconsistent rule$\in$rule sets) then
select the rule of the highest confidence;
move $\cup_i C$ from A_i;

add $\cup_i C \Rightarrow D$ to R;
for each item $a_1 \in A_i$
for each item $a_2 \in A_i$
{ if $a_1[1]=a_2[1] \wedge a_1[2]=a_2[2] \wedge \cdots \wedge a_1[i-1]=a_2[i-1] \wedge a_1[i]<a_2[i]$ then
$b=a_1 \infty a_2$;
if $b \geq$ min_sup then
add b to A_{i+1};
}}
return R;

第四步:输出约简表 R。

4 示例演示

为了验证算法的有效性,使用表 1 经过预处理的试验数据做实验,其中 X, Y, Z 为条件属性, M 为决策属性[7],设 min_sup=0.2, min_conf=0.6。由于关联规则的挖掘是基于布尔型的,所以将表 1 转化成如表 2 所示的布尔型的数据。对表 2 的数据进行属性值约简(过程见图 1)可得到化简结果(表 3)。

表 1 预处理后的决策表

U	X	Y	Z	M
1	1	0	1	1
2	1	0	0	1
3	0	0	0	0
4	1	1	1	0
5	1	1	2	2
6	2	1	2	2
7	2	2	2	2

表 2 布尔量化后的决策表

U	X			Y			Z			M		
	x_1	x_2	x_3	y_1	y_2	y_3	z_1	z_2	z_3	m_1	m_2	m_3
1	0	1	0	1	0	0	0	1	0	0	1	0
2	0	1	0	1	0	0	1	0	0	0	1	0
3	1	0	0	1	0	0	1	0	0	1	0	0
4	0	1	0	0	1	0	0	1	0	1	0	0
5	0	1	0	0	1	0	0	0	1	0	0	1
6	0	0	1	0	1	0	0	0	1	0	0	1
7	0	0	1	0	0	1	0	0	1	0	0	1

表 3 约简表

U	X	Y	Z	M
1	2	*	*	2
2	*	0	*	1
3	*	1	*	2
4	*	*	2	2

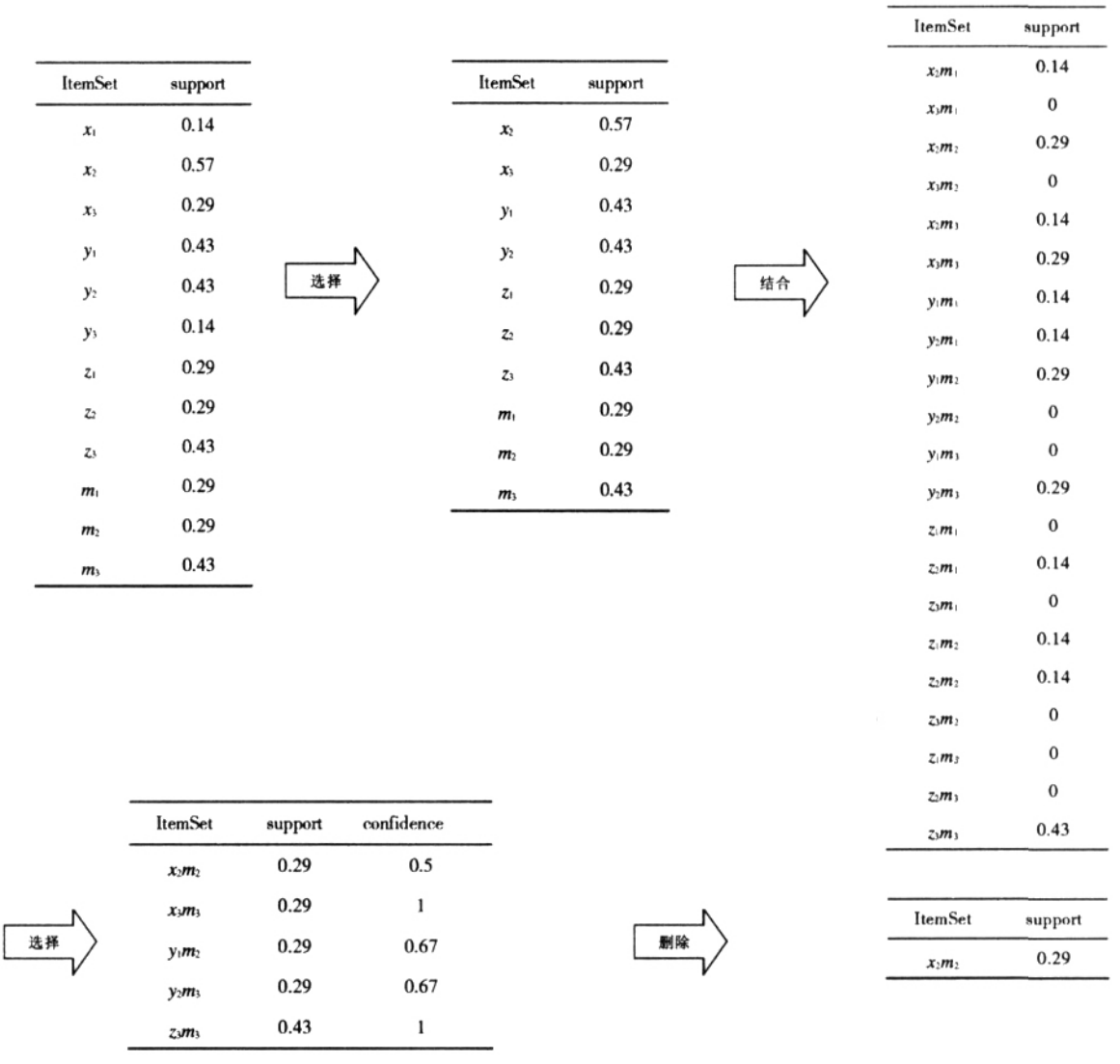

图 1　规则约简示意图

5　结语

粗糙集理论中约简是一个重要的研究课题,它包括属性约简和属性值约简两方面内容,属性值约简的目的是在保持规则集分类能力的条件下删除多余的属性值,以便得到简洁的规则集。但有的时候研究者不仅要得到最简的规则集,而且要了解每个规则的出现次数,以便作出正确的决策。所以文中提出了一种利用改进的关联规则进行粗糙集属性值约简的算法,对冗余规则和不一致规则进行处理。文中的重点不在于求得最佳属性值约简,而在于求得满足用户需求的最佳属性值约简。

参考文献:

[1] Pawlak Z. Rough Sets[J]. International Journal of Information and Computer Sciences, 1982, 11(5): 341- 356.

[2] Pawlak Z. Rough Sets- theoretical Aspects of Reasoning about Data[M]. Boston: Kluwer Academic Publishers, 1991.

[3] Wang J. Reduction algorithems based on discerniblity matrix: The ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489- 504.

[4] 韩祯祥, 张琦, 文福拴. 粗糙集理论及其应用综述[J]. 控制理论与应用, 1999, 16(2): 153- 157.

[5] 白秀玲, 王平, 普杰信. 一种粗糙集值约简算法及其应用[J]. 微计算机信息, 2006, 22(11): 207- 208.

(下转第 44 页)

参考文献:

[1] 郭程轩,甄坚伟. 基于 TM 图像的城市生态绿地格局分析与评价[J]. 国土资源遥感, 2003, (3) : 33- 36.

[2] 陈晔,赵纯勇,魏兴萍. 基于 RS、GIS 重庆市都市区生态绿地分析[J]. 曲阜师范大学学报, 2005, 31(3) : 115- 118.

[3] 周文佐,潘剑君,房世波,等. 应用 TM 影像分析南京城市生态绿地格局[J]. 城市环境与城市生态, 2002, 15(1) : 4- 6.

[4] 王延乔,高峻. 城市绿化遥感信息快速提取及其景观格局分析[J]. 中国园林, 2002, (1) : 8- 11.

[5] 孟昭山,杨士伟. 卫星遥感技术在城市绿地调查方面的应用[J]. 东北测绘, 2003, 26(2) : 54- 56.

[6] 陈颖彪,吴志峰,程炯,等. 遥感与 GIS 支持下的城市绿地信息提取方法研究- 以深圳市为例[J]. 生态环境, 2004, 13(3) : 362- 364.

Analysis of Greenbelt Pattern of Suzhou Based on Remote Sensing Images

WANG Ying[1], YAN Yong[1,3], FAN Ling- yun[2]

(1.School of Environmental Science and Engineering, USTS, Suzhou 215011, China; 2.School of Achitechture and Urban Planning, USTS, Suzhou 215011, China; 3.National Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China)

Abstract: The city greenbelt information of Suzhou is extracted by the following means: adopting TM images and the China Brazil Earth Resources Satellite (CBERS- 1) data and using advanced remote sensing image processing means such as band combination, image rectification, image enhancement, neural network classification, with the aids of the field survey, airphoto and relief map, ERDAS IMAGINE and ENVI software. On the basis of these processing means, the author obtained the dynamic investigation result of the greenbelt overlay in 1986, 1998 and 2004. Finally, the paper discusses the greenbelt distributing and the changing situation through multiple analyses of these results wholly and locally, and summarizes the characteristics of the greenbelt pattern distribution of Suzhou.

Key words: TM images; CBERS- 1; image classification; information extraction; greenbelt pattern analysis

责任编辑:谢金春

(**上接第 19 页**)[6] 陈炼,邓少波,万芳,等. 基于二进制的 Rough 集决策表约简[J]. 计算机工程, 2007, 33(16) : 193- 195.

[7] 林杰斌,刘明德,陈湘. 数据挖掘与 OLAP 理论与实务[M]. 北京:清华大学出版社, 2003.

[8] 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.

[9] 范明,孟小峰. 数据挖掘:概念与技术[M]. 北京:机械工业出版社, 2001.

A Study of the Value Reduction Algorithm in Rough Set Based on Association Rules Mining

DU Yue, WANG Zhi- he, JING Yong- xia

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China)

Abstract: Reduction is a very important issue in Rough Set, which consists of attribute reduction and value reduction. At present, only reduction can be realised in attribute reduction and the frequency of occurrence for each rule can not be calculated. To resolve the problem, this paper presents a new algorithm which takes the advantages of both association rule and rough set, dealing with redundant rules and inconsistent rules, so that a practical reduction table is obtained. Experiments show that the algorithm is efficient.

Key words: association rule; rough set; value reduction; attribute; decision table **责任编辑:蔡熹芸**