基于二进制可辨矩阵的决策规则约简算法

王锡淮,张腾飞,肖健梅
WANG Xi-huai,ZHANG Teng-fei,XIAO Jian-mei

上海海事大学 电气自动化系,上海 200135
Department of Electrical and Automation,Shanghai Maritime University,Shanghai 200135,China
E-mail:wxh@shmtu.edu.cn

WANG Xi-huai,ZHANG Teng-fei,XIAO Jian-mei.Algorithm for decision rules reduction based on binary discernibility matrix.Computer Engineering and Applications,2007,43(27):178-180.

Abstract: The decision rules reduction,namely deleting the redundant attribute values of each rule in decision table,is an important topic in the research on data reduction based on rough set theory.It can be achieved by heuristic information after attribute reduction.A method for calculating decision rules core based on binary discernibility matrix is presented directly.And then,an algorithm for decision rules reduction is designed,which is suitable for not only consistent decision table but also inconsistent decision table.

Key words: rough set theory;binary discernibility matrix;attribute reduction;decision rules reduction;decision table

摘 要:决策规则的约简是利用决策逻辑分别消去每一条决策规则中的冗余属性值,是粗糙集理论知识约简的重要内容,一般是在属性约简之后采用启发式信息实现决策规则的约简。基于二进制可辨矩阵给出一个简单的直接求取决策规则核的方法,并提出一种决策规则的约简算法。所给算法简单直观,不但适用于相容决策表,也适用于不相容的决策表。

关键词:粗糙集;二进制可辨矩阵;属性约简;决策规则约简;决策表
文章编号:1002-8331(2007)27-0178-03 **文献标识码**:A **中图分类号**:TP18

1 引言

粗糙集理论[1-3]是一种有效的处理模糊和不确定性知识的数学工具,它在保持决策表分类能力不变的前提下,通过知识约简,导出问题的分类或决策规则。利用粗糙集理论进行知识约简,最重要的就是属性约简和规则约简。通过约简操作,降低属性的维数,得到简化的决策规则。决策规则的约简就是利用决策逻辑分别消去每一条决策规则中的冗余属性值。文献[4]描述了决策表知识约简的一般方法。由于求取决策表的所有决策规则约简是一个 NP 完全问题,采用启发式信息来得到简化的决策规则,大多是基于差别矩阵的逻辑运算[4-6]。

二进制可辨矩阵[7]是用二进制的方法对 A.Skowron 区分矩阵的改进,由于采用了二进制的表达形式,其计算比施行等价类计算要快得多,灵活得多,而且更加简单直观。基于二进制可辨矩阵的属性约简已经有了较多的研究[8,11]。为了可以更方便直观地得到属性的约简及规则的约简,本文将二进制可辨矩阵仍然表示为决策表的形式。基于这种形式的二进制可辨矩阵,文献[11]给出了一种简单的决策表属性约简算法,本文在此基础上进一步研究决策规则的约简,给出一种求取决策规则核值及约简的算法。

2 基本概念

决策表是一类特殊而重要的知识表达系统,多数决策问题都可以用决策表形式来表达,这一工具在粗糙集理论中起着重要的作用。

决策表可以根据知识表达系统定义如下:

设 $S=<U,R,V,f>$为一知识表达系统,U 是论域,$R=C\cup D$,$C\cap D=\varnothing$,C 称为条件属性集,D 称为决策属性集,V 为属性值的集合,$f:U\times R\to V$ 是一个信息函数,它指定 U 中每一个对象 x 的属性值。具有条件属性和决策属性的知识表达系统称为决策表。

定义 1 在信息系统 S 中,对于属性子集 $P\subseteq R$,不可分辨关系定义为:

$$IND(P)=\{(x,y)\in U\times U:\forall a\in P,a(x)=a(y)\}$$

在决策表中,最重要的就是决策规则的产生。令 X_i、Y_j 分别代表 U/C 与 U/D 中的各个等价类,$des(X_i)$ 表示对等价类 X_i 的描述,即等价类 X_i 对于各条件属性值的特定取值 $f(X_i,C)$,$des(Y_j)$表示对等价类 Y_j 的描述,即等价类 Y_j 对于各决策属性值的特定取值 $f(Y_i,D)$。

决策规则定义如下:

$r_{ij}:des(X_i)\to des(Y_j)$,$Y_j\cap X_i\neq\varnothing$

任何决策表 $S=<U,R,V,f>$可以看作如下形式的(广义)决策规则集[1]:

$$\wedge(c,v)\to\vee(d,w)$$

其中 $c\in C,v\in V_c,w\in V_d$,$\wedge(c,v)$称为规则的条件部分,$\vee(d,$

基金项目:上海市重点学科建设项目(No.T0602);上海市教育委员会科研项目(No.04FA02,No.05FZ06)。

作者简介:王锡淮(1961-),男,教授,博士生导师,主研方向:粗糙集理论,复杂系统建模与控制等;张腾飞(1980-),男,博士生,主要研究方向:粗糙集理论、智能信息处理和神经网络等;肖健梅(1962-),女,教授,主要研究方向:智能控制,智能信息处理等。

w)称为规则的决策部分。

可见,在决策表中,每一个 x_i 以及所对应的属性值代表了一条决策规则,若存在 $f(x_i,C)=f(x_j,C)$,但 $f(x_i,D)\neq f(x_j,D)$,则称该决策表为不相容的。

属性约简之后,决策表已经得到了简化,但此时的决策规则中仍然存在着冗余的属性值,希望通过决策规则约简得到简化的决策规则集。

为了可以方便地求出决策规则的核与约简,引入二进制可辨矩阵。考虑决策表可能存在的不相容性,这里引用文献[9]的一个记法:对于 $x_i\in U$,记

$d(x_i)=card\{f(y,D)|y\in[x_i]_C\}$

$d(x_i)$表示 U 中所有与 x_i 在条件属性 C 下等价的实例相应的决策属性构成的集合的基数。

定义 2 构造决策表 $S=<U,R,V,f>$的二进制可辨矩阵 $\boldsymbol{S}^*=<U^*,R^*,V^*,f^*>$:

$U^*=\{(x_i,x_j)\in U\times U|d(x_i)\neq d(x_j)\wedge\min\{|d(x_i)|,|d(x_j)|\}=1\}$

$A^*=\{c|c\in C\}$

$V^*=\{1,0\},\forall c\in A^*$

$f^*:U^*\times A^*\rightarrow V^*$

若 $f(x_i,c)\neq f(x_j,c)$,则 $f^*((x_i,x_j),c)=1$;否则 $f^*((x_i,x_j),c)=0$。根据二进制可辨矩阵的构造可知,对 $\forall c\in A^*,f^*((x_i,x_j),c)=1$ 表示在原决策表中属性 c 能区分对象 x_i 与 x_j。

为了便于说明,看表 1 所示的决策表。根据上面的定义,可构造二进制可辨矩阵如表 2。

表 1 决策表

U	a	b	c	d
x_1	1	0	1	1
x_2	1	0	0	1
x_3	0	0	0	0
x_4	1	1	1	0
x_5	1	1	2	2
x_6	2	2	2	2

表 2 决策表 1 对应的二进制可辨矩阵

U^*	a	b	c
(x_1,x_3)	1	0	1
(x_1,x_4)	0	1	0
(x_1,x_5)	0	1	1
(x_1,x_6)	1	1	1
(x_2,x_3)	1	0	0
(x_2,x_4)	0	1	1
(x_2,x_5)	0	1	1
(x_2,x_6)	1	1	1
(x_3,x_5)	1	1	1
(x_3,x_6)	1	1	1
(x_4,x_5)	0	0	1
(x_4,x_6)	1	1	1

由此可以看出,定义 2 中的二进制可辨矩阵实质上是和文献[9]改进的差别矩阵等价的,所不同的是二进制可辨矩阵采用更简单直观的 1 和 0 来表示对应的属性是否可以区分所在行的实例对,因此本文在这种二进制可辨矩阵基础上提出的算法完全适用于不相容决策表。

3 决策规则约简算法

根据二进制可辨矩阵的构造可知,各行中 1 的个数表示的是可以区分该行所对应的实例对的属性个数,当某行 1 的个数为 1 时,表明此 1 所对应的属性是唯一能区分该行所对应的实例对的属性,因此,这个属性是保持决策表不可分辨关系不可缺少的,即为核属性,同时它也是区分该行实例对所必须的,因此,这个属性所对应的实例对的属性值即分别为该行实例对相互区分的核值。例如,在表 2 中,第 2 行仅有一个 1,对应着属性 b,因此属性 b 是唯一能区分 x_1 和 x_4 的属性,(b=0)和(b=1)分别为 x_1 和 x_4 的一个核值。由此,可以得到如下性质。

性质 对于每条决策规则(实例),在二进制可辨矩阵中,所有包含该实例而且仅有一个 1 的行中 1 所对应的属性值集合构成了该规则的核。

决策表中不同的属性可能具有不同的重要性,在知识约简的启发式算法中,属性的重要性度量具有重要的作用。由前面的分析可知,二进制可辨矩阵中各列所含 1 的个数表示该列所对应的属性可以区分实例对的个数;属性可以区分的实例对越多,从某种程度上也表明了属性的重要性越高。

定理 1 对于一条决策规则(实例),在二进制可辨矩阵所有包含该实例的行中,如果某个属性 c_i 所对应的位置均为 1,则属性 c_i 在该实例上的取值构成该规则的一个约简。

证明 在二进制可辨矩阵所有包含该实例的行中,属性 c_i 所对应的位置均为 1,说明属性 c_i 可以将该实例与其他的所有实例区分开,因此,属性 c_i 在该实例上的取值可以构成该规则的一个约简。

根据前面给出的性质和定理,可以给出简单的直接求取决策表属性核值表的算法,以及决策规则的约简算法。

算法 1 属性核值表计算方法。

输入:属性约简后的决策表 $S=<U,R,V,f>$,$R=C\cup D$ 是属性集合,$C=\{c_i|i=1,2,\cdots,m\}$和 $D=\{d_i|i=1,2,\cdots,n\}$分别称为条件属性集和决策属性集。

输出:决策表属性核值表。

步骤 1 构造二进制可辨矩阵 $\boldsymbol{S}^*$;

步骤 2 对二进制可辨矩阵逐行扫描,如果该行中 1 的个数为 1,则将此 1 所对应的属性在该行实例对的不同取值分别添加到这两个实例的核值集,否则跳过;

步骤 3 输出决策表的属性核值表。

如果二进制可辨矩阵没有一个仅含 1 的行,则核值表为空。

算法 2 决策规则约简算法。

输入:属性约简后的决策表 $S=<U,R,V,f>$,其中,$U=\{x_i|i=1,2,\cdots,r\}$,$R=C\cup D$ 是属性集合,$C=\{c_i|i=1,2,\cdots,m\}$和 $D=\{d_i|i=1,2,\cdots,n\}$分别称为条件属性集和决策属性集。

输出:简化的决策规则集。

步骤 1 构造二进制可辨矩阵 $\boldsymbol{S}^*$;

步骤 2 对每个实例 $x_i,i=1,2,\cdots,r$ 执行如下操作:

(2.1)在二进制可辨矩阵中对含有实例 x_i 的行以及含有与 x_i 不相容实例的行进行标记;

(2.2)如果所有标记的行在某个属性 c_j 所对应的位置均为 1,则 $Reduct=\{c_j\}$,转(2.5),否则转(2.3);

(2.3)查找标记的仅含一个 1 的行计算该决策规则的核属性 $Core$,令 $Reduct=Core$;

(2.4)对标记的行进行扫描,如果某行中 1 所对应属性均不在 $Reduct$ 中,选择一个重要性较大的属性加入 $Reduct$ 中;

(2.5)由属性集 $Reduct$ 在该决策规则上的取值得到该规则的约简;清除(2.1)对行的标记;

步骤 3 对约简后相同的决策规则进行合并,输出简化的决策规则集。

4 实例计算

以决策表 1 为例,表 2 为相应的二进制可辨矩阵。在表 2 中仅含一个 1 的行如表 3 所示。根据算法 1 可以很容易得到仅包含核值的决策表,如表 4 所示。

表3 表2中仅含一个1的行

(x_1,x_4)	0	1	0
(x_2,x_3)	1	0	0
(x_4,x_5)	0	0	1

表4 仅包含核值的决策表

U	a	b	c	d
x_1	–	0	–	1
x_2	1	–	–	1
x_3	0	–	–	0
x_4	–	1	1	0
x_5	–	–	2	2
x_6	–	–	–	2

根据算法2可以得到简化的决策规则集为：

rule 1: $(b=0)(c=1) \dashrightarrow (d=1)$

rule 2: $(a=1)(b=0) \dashrightarrow (d=1)$

rule 3: $(a=0) \dashrightarrow (d=0)$

rule 4: $(b=1)(c=1) \dashrightarrow (d=0)$

rule 5: $(c=2) \dashrightarrow (d=2)$

rule 6: $(a=2) \dashrightarrow (d=2)$

由此可以得到一个最小化决策算法：

$b_0c_1 \vee a_1b_0 \rightarrow d_1$

$a_0 \vee b_1c_1 \rightarrow d_0$

$a_2 \vee c_2 \rightarrow d_2$

可见，由上述算法可以方便快捷地求出决策规则的核与约简。

5 结论

决策规则约简算法是基于粗糙集理论的知识约简研究的重要内容之一，本文通过引入决策表的二进制可辨矩阵，给出了一种简单的直接求取决策表属性核值的方法，并设计一种基于二进制可辨矩阵的决策规则约简算法。实例分析验证了算法的简单有效性。（收稿日期：2007年1月）

参考文献：

[1] 张文修.Rough集理论与方法[M].北京：科学出版社，2001.

[2] 刘清.Rough集及Rough推理[M].北京：科学出版社，2001.

[3] 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.

[4] 王珏，苗夺谦，周育健.关于Rough Set理论与应用的综述[J].模式识别与人工智能，1996，9(4)：337-344.

[5] Yang Yan-yi，Chiam T C.Rule discovery based on rough set theory[C]//12th International Symposium on Integrated Ferroelectrics，Aachen，Germany，2000：11-16.

[6] 刘文军，谷云东，李洪兴.基于区分矩阵求决策算法的约简[J].北京师范大学学报：自然科学版，2003，39(3)：311-315.

[7] Fleix R，Ushio T.Rough sets-based machine learning using a binary discernibility matrix[C]//IPMM'99 Published，1999：299-305.

[8] 支云天，苗夺谦.二进制可辨矩阵的变换及高效属性约简算法的构造[J].计算机科学，2002，29(2)：140-142.

[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] 叶东毅.一个新的二进制可辨识矩阵及其核的计算[J].小型微型计算机系统，2004，25(6)：965-967.

[11] Wang Xi-huai，Zhang Teng-fei.A new algorithm for relative attribute reduction in decision table[C]//The 6th World Congress on Intelligent Control and Automation，Dalian，P R China，June 21-23，2006：4051-4054.

（上接155页）

databases[C]//SIGMOD Conference.New York：ACM Press，2002：334-345.

[5] Raptopoulou K，Papadopoulos A，Manolopoulos Y.Fast nearest-neighbor query processing in moving-object databases[J].GeoInformatica，2003，7(2)：113-137.

[6] Saltenis S，Jensen C S，Leutenegger S T，et al.Indexing the positions of continuously moving objects[C]//SIGMOD Conference，2000：331-342.

[7] Iwerks G S，Samet H，Smith K.Continuous k-nearest neighbor queries for continuously moving points with updates[C]//VLDB，2003：512-523.

[8] Tao Y，Faloutsos C，Papadias D，et al.Prediction and indexing of moving objects with unknown motion patterns[C]//SIGMOD Conference，2004：611-622.

[9] Song Z，Roussopoulos N.Hashing moving objects[C]//Proceedings of the 2nd International Conference on Mobile Data Management.London：Springer-Verlag，2001：161-172.

[10] Kwon D，Lee S，Choi W，et al.An adaptive hashing technique for indexing moving objects[J].Data & Knowledge Engineering，2006，56(3)：287-303.

[11] Yu X，Pu K，Koudas N.Monitoring k-nearest neighbor queries over moving objects[C]//ICDE，Tokyo，2005：631-642.

[12] Brinkhoff T.A framework for generating network-based moving objects[J].GeoInformatica，2002，6(2)：153-180.

（上接170页）

引入了记忆抗体"年龄"的概念，通过实验证明该算法在解决聚类簇更新问题时是有效的。（收稿日期：2007年1月）

参考文献：

[1] Xie Y，Phoha VV.Web user clustering from access log using belief function[C]//Proceedings of the First International Conference International Conference on Knowledge Capture(K-CAP 2001).[S.l.]：ACM Press，2001：202-208.

[2] Heer J，Chi EH.Mining the structure of user activity using cluster stability[C]//Proceedings of the Workshop on Web Analytics，Second SIAM Conference on Data Mining.[S.l.]：ACM Press，2002.

[3] Abraham A.i-Miner：a web usage mining framework using hierarchical intelligent systems[C]//The IEEE International Conference on Fuzzy Systems，FUZZ-IEEE'03，2003：1129-1134.

[4] 金阳，左万利.有序概念格与Web用户访问模式的增量挖掘[J].计算机研究与发展，2003，40(5)：675-683.

[5] 董一鸿，庄越挺.基于新型的竞争型神经网络的Web日志挖掘[J].计算机研究与发展，2003，40(5)：621-667.

[6] 李涛.计算机免疫学[M].北京：电子工业出版社，2004.

[7] 刘韬，王耀才，王致杰.一种基于人工免疫的聚类算法[J].计算机工程与应用，2004，40(19)：182-184.

[8] 李春华，朱燕飞，毛宗源.一种新型的自适应人工免疫算法[J].计算机工程与应用，2004，40(22)：84-87.

[9] 蒋加伏，罗晓萍，唐贤瑛，等.数据聚类的FCM与aiNet方法[J].计算机工程与设计，2004(4).