一种改进的规则分辨矩阵及其属性值约简方法

舒 芬,王加阳
SHU Fen,WANG Jia-yang

中南大学 信息科学与工程学院,长沙 410083
College of Information Science and Engineering,Central South University,Changsha 410083,China
E-mail:shufen_928@163.com

SHU Fen,WANG Jia-yang.Improvement of rule discernibility matrix and method for attributes value reduction.Computer Engineering and Applications,2007,43(32):77-79.

Abstract: Attributes value reduction is one of the important parts researched in rough set theory.In order to correct the errors of computing reduction of the values in a decision table based on discernibility matrix,this paper puts forward an improved discernibility matrix definition together with a method for values reduction.It takes the new inconsistency into consideration caused in the values reduction.Therefore,the matrix discerns the rules that have the same values of decisive attributes with the inconsistent rules in order to avoid the possibility of including wrong rules after values reduction.

Key words: rule discernibility matrix;attributes value reduction;inconsistency

摘 要:属性值约简是粗糙集理论的重要研究内容之一,针对利用分辨矩阵求值约简的错误,提出了一种改进的规则分辨矩阵和值约简方法,主要考虑属性值约简可能导致新的不一致性问题。该矩阵区分与不一致规则的决策值相同的规则,从而避免了值约简出现错误规则的可能。

关键词:规则分辨矩阵;属性值约简;不一致性
文章编号:1002-8331(2007)32-0077-02 **文献标识码**:A **中图分类号**:TP391

1 引言

粗糙集理论是波兰数学家 Pawlak 在 20 世纪 80 年代提出的一种能有效处理不精确,不确定和含糊信息的数学理论[1],经过二十余年的发展,它已在数据挖掘、机器学习、模式识别与智能信息处理等领域得到了较为广泛的应用[2,3]。在粗糙集理论中,属性值约简是重要研究内容之一,也是知识获取的关键步骤,因此属性值约简倍受粗糙集研究者的关注,也取得了很大的进展[4,5]。而有些值约简方法是基于分辨矩阵的,分辨矩阵的定义成了属性值约简的关键步骤,因而探索研究分辨矩阵的有效定义具有重要的使用价值。

文献[4]根据 HU 提出的分辨矩阵[6],通过分辨函数的化简得到决策信息系统的值约简。但该方法对于一致的决策表,可以达到完整的值约简,但是在不一致的决策表中,将会得到一些不正确规则。文献[7]在 HU 的分辨矩阵定义基础上,提出新的分辨矩阵,该方法在保持相对正域不变的基础上考虑了不一致性情况下的异常, 因此可以得到不一致性决策表的代数约简,文献[8]也提出了一种规则矩阵,由此可以得到不一致性表的信息论的核及其约简,但以上两种定义都没有考虑到值约简所产生的新的不一致。

本文给出了改进的规则分辨矩阵和值约简方法,对于一致性等价类和不一致性等价类,该矩阵分别考虑了其在规则分辨矩阵中的元素。首先根据定义得到决策表的规则分辨矩阵,进而化简每条规则的值约简函数,得到每条规则的值核属性以及最简产生式规则。

2 粗糙集概念

定义 1 信息系统(或决策表)定义为四元组 $IS=(U,Q=C\cup D,V,f)$。其中,论域 $U=\{x_1,x_2,\cdots,x_{|U|}\}$,$|U|$表示 U 中包含对象的数目,$Q=C\cup D$ 为属性集合,C 和 D 分别为条件属性集和决策属性集;$V=\bigcup_{a\in Q}V_a$,为属性 a 的值域集;f 是 $U\times Q\to V$ 的映射。

定义 2 属性集 $R(R\subseteq Q)$,不可分辨关系定义为:$IND(R)=\{(x,y)\in U^2|\forall a\in R,f(x,a)=f(y,a)\}$论域 U 在属性集 C 上形成的划分 $U/IND(C)=\{X_1,X_2,\cdots,X_m\}$称为条件分类集,论域 U 在属性集 D 上形成的划分 $U/IND(D)=\{Y_1,Y_2,\cdots,Y_n\}$称为决策分类集。

定义 3 设 $X\subseteq U$ 为论域的一个子集,$B\subseteq C$,X 的关于 B 的下近似为 $\underline{B}X=\{x\in U:[x]_B\subseteq X\}$,$X$ 关于 B 的上近似为 $\overline{B}X=\{x\in U:[x]_B\cap X\neq\varnothing\}$,$X$ 的边界 $BN_C(X)=\overline{B}X-\underline{B}X$,其中$[x]_B$ 表示 U 中所有与 x 在关系 $IND(B)$下是等价的元素构成的集合。

定义 4 分辨矩阵 $\boldsymbol{M}=\{m_{ij}\}$定义为:

$$m_{ij}=\begin{cases}\{a\in C:f(x_i,a)\neq f(x_j,a)\},f(x_i,D)\neq f(x_j,D)\\ \varnothing(\text{空集}),\text{其他}\end{cases}$$

定义 5 信息系统(或决策表)$IS=(U,Q=C\cup D,V,f)$,条件

基金项目:湖南省自然科学基金(the Natural Science Foundation of Hunan Province of China under Grant No.06JJ20075[MS1])。
作者简介:舒芬(1982-),女,硕士,主要研究方向:粗糙集理论及应用;王加阳(1963-),男,博士,教授,研究方向:智能计算与信息融合。

属性 $a\in C$ 是规则 $L\in U/(C\cup D)$不可缺少的属性，当且仅当去除属性 a 将引起新的规则与 L 发生冲突，相应的属性值称为规则的值核。为叙述方便，称与一条决策规则值核对应的属性为该规则的值核属性，记为 $CORE(L)$。

定义 6 在信息系统（或决策表）$IS=(U,Q=C\cup D,V,f)$中，对于任一给定的规则 $L\in U/(C\cup D)$，根据任一属性集 $B\subseteq C\cap D$ 对 U 进行分类，$[L]_{IND(B)}$表示满足规则 L 的 B 条件的一个 B等价类，若 L 满足$[L]_{IND(C)}\subseteq[L]_{IND(D)}$，则 L 是一致规则，否则 L 是不一致规则。

3 现有的分辨矩阵求值约简的缺陷

定义 7 值约简函数为：$\delta_i=\wedge\{\vee m_{ij}|1\leqslant j\leqslant n,m_{ij}\neq\varnothing\}$，$m_{ij}$ 为分辨矩阵中的单元。

若把每一个对象看成一条规则，那么一条决策规则可以简单地通过一个值约简函数来泛化。值约简函数 δ_i 是一个布尔函数，由分辨矩阵的第 i 行构造，把出现在第 i 行中的每个属性作为一个布尔变量，并且对每个矩阵分量进行先析取后合取的布尔运算，其结果就是一条规则的值约简。根据这一点，可以通过分辨矩阵进行值约简。

实例 1 表 1 所示的是一张数据表，其中共有 5 个元素（x_1~x_5）和 4 个属性，$C=\{C_1,C_2,C_3\}$为条件属性集，D 为决策属性。

表 1 数据表

U	C_1	C_2	C_3	D
x_1	1	0	1	1
x_2	1	0	1	0
x_3	0	0	1	1
x_4	0	0	1	0
x_5	1	1	1	1

根据定义 4 构造的分辨矩阵如表 2 所示。

表 2 分辨矩阵

U	x_1	x_2	x_3	x_4	x_5
x_1		$\varnothing$	$\varnothing$	C_1	$\varnothing$
x_2	$\varnothing$		C_1	$\varnothing$	C_2
x_3	$\varnothing$	C_1		$\varnothing$	$\varnothing$
x_4	C_1	$\varnothing$	$\varnothing$		C_1,C_2
x_5	$\varnothing$	C_2	$\varnothing$	C_1,C_2	

化简后得到规则如下：

规则 1 $C_{1(1)}\rightarrow D_{(1)}$

规则 2 $C_{1(1)}\wedge C_{2(0)}\rightarrow D_{(0)}$

规则 3 $C_{1(0)}\rightarrow D_{(1)}\vee D_{(0)}$

规则 4 $C_{2(1)}\rightarrow D_{(1)}$

其中 $C_{i(j)}$表示属性 C_i 取值为 j。

可以发现，规则 1 和规则 2 被判定为一致性规则，而根据决策表有 $C_{1(1)}\rightarrow D_{(0)}$，$C_{1(1)}\wedge C_{2(0)}\rightarrow D_{(1)}$，因此这两条规则是不正确的。

出现这种不正确规则，是因为分辨矩阵没有考虑不一致性对象。当两个对象决策相同，且属于不同条件类，值约简将有可能导致新的不一致性。具体而言，当出现一个不一致性对象与一个一致性对象喊有部分相同的属性值且决策也相同时（如表 1 中的对象 x_1 和 x_5），就会不可避免地出现这种不正确规则。因此，该分辨矩阵不能得到本实例的值约简。

4 改进的规则分辨矩阵及其应用

根据条件属性和决策属性的联合进行分类，这样就能节省存储空间并减少矩阵运算量。所划分的等价类同时考虑了条件属性值和决策属性值，实际上每个等价类可看作是一条规则，这样定义的矩阵元素比较的是两条规则，可以称为规则分辨矩阵。

定义 8 给定信息系统 $IS=(U,Q=C\cup D,V,f)$，属性集 Q 在论域 U 上的划分 $U/Q=\{L_1,L_2,\cdots,L_N\}$，新的规则分辨矩阵定义为：

$$m_{ij}=\begin{cases}a\in C:f(L_i,a)\neq f(L_j,a),f(L_i,D)\neq f(L_j,D)\\ a\in C:f(L_i,a)\neq f(L_j,a),f(L_i,D)=f(L_j,D),\max\{D(L_i),D(L_j)\}>1\\ \varnothing,\text{其他}\end{cases}$$

其中 $D(L_i)=card(f(L_i,D):L_j\subset[L_i]_{IND(C)})$。$[L_i]_{IND(C)}$为规则 L_i 关于条件属性集 C 上的等价类，$D(L_i)$表示与 L_i 关于条件属性集上的等价类相应的决策属性值构成的集合的基数。当 $D(L_i)=1$ 时，表示$[L_i]_{IND(C)}\subseteq Y_k,k\in[1,n]$，而当 $D(L_i)>1$ 表示$[L_i]_{IND(C)}$中的元素$[L_i]_{IND(C)}\not\subset Y_k,k\in[1,n]$，即数据存在不相容性。

考虑到值约简会产生新的冲突，因此当 $\max\{D(L_i),D(L_j)\}>1,f(L_i,D)=f(L_j,D)$时，其对应的条件属性不能一概都不区分。若 $L_i\subseteq\sum_{k=1}^{n}\underline{C}Y_k,L_j\subseteq U-\sum_{k=1}^{n}\underline{C}Y_k$，即$|D(L_i)|=1,|D(L_j)|>1$，而且 $L_j\not\subset[L_i]_{IND(C)},L_j\subset[L_i]_{IND(C-\{a\})}$，那么去除属性 a 后 L_i,L_j 条件等价类合并，存在 $L_r\subset[L_j]_{IND(C)}$与规则 L_i 冲突；若 $L_i\not\subset\sum_{k=1}^{n}\underline{C}Y_k,L_j\not\subset\sum_{k=1}^{n}\underline{C}Y_k$，即$|D(L_i)|>1,|D(L_j)|>1$，，如果去除属性 a 后 L_i,L_j 条件等价类发生合并，那么也会存在 $L_r\subset[L_j]_{IND(C)}$与 L_i 冲突，因此，在这两种情况下，L_i,L_j 对应的条件属性需要区分，在分辨矩阵中规则 L_i、L_j 对应的属性组合 m_{ij} 为值核属性，不能被删除。

定理 1 对于决策信息系统 $IS=(U,Q=C\cup D,V,f)$，如果记 $IDM(L_i)=\{m_{ij}:m_{ij}$ 为第 i 行的单个属性$\}$，则有 $IDM(L_i)=CORE(L_i)$。即当且仅当某个 m_{ij} 为单个属性时，该属性就是规则 L_i 的值核属性 $CORE(L_i)$。

证明 首先证明 $IDM(L_i)\subseteq CORE(L_i)$。任取一个 $a\in IDM(L_i)$，由定义 8 可知至少存在矩阵 $\boldsymbol{M}$ 中的一个元素，不妨设为 m_{ij} 使得 $m_{ij}=\{a\}$。于是，由定义知存在 L_j 使得 $L_j\not\subset[L_i]_{IND(C)}$，但 $L_j\subset[L_i]_{IND(C-\{a\})}$。若 $f(L_i,D)\neq f(L_j,D)$，可以分三种情况考虑：(1)$L_i\subseteq\sum_{k=1}^{n}\underline{C}Y_k,L_j\subseteq\sum_{k=1}^{n}\underline{C}Y_k$，由于 $f(L_i,D)\neq f(L_j,D)$，因此对于划分$\{Y_1,Y_2,\cdots,Y_n\}$来说，L_i、L_j 不属于其中的同一个划分子集。不妨设 $L_i\subseteq Y_s,L_j\subseteq Y_t,1\leqslant s,t\leqslant n,s\neq t$。由 $L_j\subset[L_i]_{IND(C-\{a\})},L_j\not\subset Y_s$，以及下近似的定义可知：$L_i\not\subset\sum_{k=1}^{n}\underline{C-\{a\}}Y_k$，同样 $L_j\not\subset\sum_{k=1}^{n}\underline{C-\{a\}}Y_k$，因此，去除 a 后决策信息系统出现了新的冲突规则，故 $a\in CORE(L_i)$。(2)$L_i\subseteq\sum_{k=1}^{n}\underline{C}Y_k,L_j\subseteq\sum_{k=1}^{n}\underline{C}Y_k$，由于 $f(L_i,D)\neq f(L_j,D)$，则 $L_i\not\subset\sum_{k=1}^{n}\underline{C-\{a\}}Y_k$，从而去除属性 a 使 L_i 与 L_j 产生了新的冲突，因此 $a\in CORE(L_i)$。(3)$L_i\not\subset\sum_{k=1}^{n}\underline{C}Y_k,L_j\not\subset\sum_{k=1}^{n}\underline{C}Y_k$，去除属性 a 后，存在 $L_r\subset[L_j]_{C-\{a\}}$与 L_i 相冲突，从而决策信息系统也会出现新的冲突规则，故 $a\in CORE(L_i)$。若 $f(L_i,D)=f(L_j,D)$，且 $\max\{|D(L_i)|,|D(L_j)|\}>1$，可以分两种情况：①$L_i\subseteq\sum_{k=1}^{n}\underline{C}Y_k,L_j\subseteq\sum_{k=1}^{n}\underline{C}Y_k$，那么去除属性 a 使得 $L_i\not\subset\sum_{k=1}^{n}\underline{C-\{a\}}Y_k$，从而在$[L_j]_{IND(C)}$

中存在 L_r:$f(L_r,D)\neq f(L_i,D)$,与规则 L_i 产生新冲突,故 $a\in CORE(L_i)$;②$L_i\not\subset\sum_{k=1}^{n}\underline{C}Y_k$,$L_j\not\subset\sum_{k=1}^{n}\underline{C}Y_k$,去除属性 a 后,与 L_i 和 L_j 相冲突的规则数增加,从而决策信息系统出现了新的冲突规则,故 $a\in CORE(L_i)$。由此证得 $SDM(L_i)\subseteq CORE(L_i)$。

下面证明反包含 $SDM(L_i)\supseteq CORE(L_i)$成立,采用反证法。假设对某个 $a\in CORE(L_i)$,不存在 m_{ij},使得 $m_{ij}=\{a\}$。取 $L_i\subseteq\underline{C}Y_k$,由下近似的定义知,$[L_j]_{IND(C)}\subseteq Y_k$,现任取 $L_j\subset[L_i]_{IND(C-\{a\})}$,来说明一定有 $L_j\subseteq Y_k$。为此,分两种情形讨论:(1)如果 $j=i$,则显然 $L_j\subseteq Y_k$。(2)考虑 $j\neq i$ 的情形。如果 $L_j\not\subset Y_k$,则 $L_j\not\subset[L_i]_{IND(C)}$,可以分两种情况来考虑:①如果 $f(L_i,D)\neq f(L_j,D)$,则根据定义 8 有 $m_{ij}=\{a\}$,这与假设矛盾;②如果 $f(L_i,D)=f(L_j,D)$,此时如果 $\max\{|ID(L_i)|,|ID(L_j)|\}\leqslant 1$,则 $|ID(L_i)|=1$,$|ID(L_j)|=1$,而 $L_j\subset[L_i]_{IND(C-\{a\})}$,$f(L_i,D)\neq f(L_j,D)$,这与 $L_i\subseteq\underline{C}Y_k$ 矛盾,因此 $L_j\subseteq Y_k$。再由 L_i 选取的任意性,可得 $[L_i]_{IND(C-\{a\})}\subseteq Y_k$,即 $L_i\subseteq\underline{C-\{a\}}Y_k$,因此,去除属性 a 决策信息系统中没有新的规则与规则 L_i 冲突,由此可得 $a\notin CORE(L_i)$,这与 $a\in CORE(L_i)$的假设矛盾。因此,反设不成立。故 $CORE(L_i)\subseteq SDM(L_i)$。定理由此得证。

证毕。

对实例 1,按定义 8 建立规则分辨矩阵如表 3:

表 3　规则分辨矩阵

U	L_1	L_2	L_3	L_4	L_5
L_1		Ø	C_1	C_1	C_2
L_2	Ø		C_1	C_1	C_2
L_3	C_1	C_1		Ø	Ø
L_4	C_1	C_1	Ø		C_1,C_2
L_5	C_2	C_2	C_1,C_2	C_1,C_2	

由定理可得规则值核为:

$CORE(L_1)=\{C_1,C_2\}$　$CORE(L_2)=\{C_1,C_2\}$

$CORE(L_3)=\{C_1\}$　$CORE(L_4)=\{C_1\}$

$CORE(L_5)=\{C_2\}$

C_2 是规则 L_1 的值核属性,因此在值约简过程中若去除属性 C_2,那么信息系统中肯定会有新的规则与 L_1 冲突,故在分辨矩阵中 $m_{15}=\{C_2\}$不能被删除。

可以得到每个联合分类后的等价类的值约简函数为:

$\delta(L_1)=C_1\wedge C_1\wedge C_2=C_1\wedge C_2$　$\delta(L_2)=C_1\wedge C_1\wedge C_2=C_1\wedge C_2$

$\delta(L_3)=C_1\wedge C_1\wedge(C_1\vee C_2)=C_1$　$\delta(L_4)=C_1\wedge C_1\wedge(C_1\vee C_2)=C_1$

$\delta(L_5)=C_2\wedge C_2\wedge(C_1\vee C_2)=C_2$

值约简表如表 4 所示:

表 4　值约简表

U	C_1	C_2	D
L_1	1	0	1
L_2	1	0	0
L_3	0	*	1
L_4	0	*	0
L_5	*	1	1

获取规则:

规则 1　$C_{1(1)}\wedge C_{2(0)}\rightarrow D_{(1)}\vee D_{(0)}$

规则 2　$C_{1(0)}\rightarrow D_{(0)}\vee D_{(1)}$

规则 3　$C_{2(1)}\rightarrow D_{(1)}$

其中规则 1,规则 2 为不一致性规则,规则 3 为一致性规则,这样就排除了类似实例 1 中 $C_{1(1)}\rightarrow D_{(0)}$和 $C_{1(1)}\wedge C_{2(0)}\rightarrow D_{(1)}$的情况。

5　结语

本文提出了一种改进的规则分辨矩阵,主要考虑了基于分辨矩阵的值约简问题。针对一致性等价类和不一致性等价类,该矩阵分别考虑了它们在规则分辨矩阵中的元素,从而完善了值约简方法。(收稿日期:2007 年 5 月)

参考文献:

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

[2] Pawlak Z.Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research,1994,72(3):443-459.

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

[4] Walczak B,Massart D L.Rough sets theory [J].Chemometrics and Intelligent Laboratory System,1999,47(1):1-16.

[5] 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,1999,10(11):1206-1211.

[6] Hu Xiao-hua,Cercone N.Learing in relational databases:a rough set approach[J].Computational Intelligence,1995,11(2):323-337.

[7] 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.

[8] 李嘉,王加阳.基于 Rough 集的规则分辨矩阵研究[J].计算机工程与应用,2006,42(11):27-31.

(上接 14 页)

[2] 李建中,李金宝,石胜飞.传感器网络及其数据管理的概念、问题与进展[J].软件学报,2003,14(10):1717-1727.

[3] 任丰原,黄海宁,林闯.无线传感器网络[J].软件学报,2003,14(2).

[4] 孙利民,李建中,陈渝,等.无线传感器网络[M].北京:清华大学出版社,2005.

[5] Chen P,Dea B O,Callaway E.Energy efficient system design with optimum transmission range for wireless ad hoc networks[C]//IEEE International Conference on Comm 2002,2002,2:945-952.

[6] Sankara Y,Akyildiz I F,McLaughlin S W.Energy efficiency based packet size optimization in wireless sensor networks[C]//Proc 1st IEEE International Workshop on Sensor Network Protocols and Applications(SNPA),Anchorage AK,2003.

[7] Ye W,Heidemann J,Estrin D.An energy-efficient MAC protocol for wireless sensor networks[C]//INFOCOM 2002,New York,2002:1567-1576.

[8] Achir M,Ouvry L.QoS and energy estimation in wireless sensor networks using CSMA/CA[C]//SENET'05:International Conference on Sensor Networks,Montreal,Canada,August 2005.

[9] Sigh S,Ragha vendra C S.PAMAS-Power Aware Multi-Access protocol with signaling for ad hoc networks[J].ACM SIGCOMM Computer Communication Review,1998,28(3):5-26.

[10] Dam T V,Langendoen K.An adaptive energy-efficient MAC protocol for wireless sensor networks[C]//SenSys'03,Los Angeles,2003:171-180.

[11] Haapola J,Shelby Z,CA Pomalaza-Rúez,et al.Multihop medium access control for WSNs:an energy analysis model[J].EURASIP Journal of Wireless Communications and Networking,2005(4):523-540.

[12] ASH Transceiver Designer's Guide FRM-1000 Transceiver[J/OL].[2001].http://www.rfm.com.