信息系统知识约简简便算法

邵明文[1] 张文修[1] 吴伟志[2]

(西安交通大学理学院信息与系统科学研究所 西安710049)[1] (浙江海洋学院数学系)[2]

Simplified Algorithms for Knowledge Reductions in Information Systems

SHAO Ming-Wen[1] ZHANG Wen-Xiu[1] WU Wei-Zhi[2]

(Institute of Information and System Science, Science College of Xi'an Jiaotong University, Xi'an 710049)[1]

(Department of Mathematics, Zhejiang Ocean University)[2]

Abstract In this paper, by using discernibility attribute matrices, give the sufficient and necessary conditions for justfying an attribute is absolute, relative and superfluous, and then simplified approaches to knowledge reductions in information systems and consistent target information systems are presented, from which the cost for computing the reductions is reduced.

Keywords Information systems, Consistent target information systems, Attributes, Knowledge reductions

1 引言

粗糙集[1]理论是一种新的处理模糊和不确定知识的数学工具，其主要思想是，在保持信系统的分类能力不变的前提下，通过知识约简，导出问题的决策或分类规则。近20年来，粗糙集理论已经在理论和应用上取得了长足的发展，特别是由于20世纪90年代在知识发现等领域得到了成功的应用而受到国际学术界广泛关注，基于 Rough 集方法的知识发现与知识约简被许多专家所研究。目前，它正在被广泛应用于机器学习、决策分析、过程控制、模式识别和数据挖掘等领域[2~4]。知识约简是粗糙集理论的核心内容之一[1]。对一个信息系统 (U,A,F) 来说，知识库中属性并不是同等重要的，甚至其中某些属性是冗余的。所谓知识约简就是在知识库分类能力不变的条件下，删除其中不重要或不相关的属性。特别是，当信息系统中的数据是随机采集时，其冗余性更为普遍。

一般地讲，一个信息系统的属性不是唯一的，人们希望找到具有最少属性的约简，即最小约简。然而，要找到一个信息系统的最小约简是一个 NP-hard 问题[5]。不过，在实际应用中，要求得到相对属性约简就可以了。许多研究人员已提出了属性约简算法[6~8]本文利用可辨识属性矩阵，确定了信息系统的核心属性和去掉绝对不必要属性，并给出一个由可辨识属性矩阵求信息系统的一个约简的简便算法。

2 信息系统的知识约简

定义2.1 称 (U,A,F) 为一个信息系统，或者数据库系统。其中 U 为对象集，即：

$$U=\{x_1,x_2,\cdots,x_n\}$$

U 中的每个 x_i，$(i\leqslant n)$，称为一个对象。而 A 为属性集，即

$$A=\{a_1,a_2,\cdots,a_m\}$$

A 中的每个 a_j，$(j\leqslant m)$，称为一个属性。F 为 U 和 A 的关系集。即

$$F=\{f_j:J\leqslant m\}$$

其中 $f_J:U\to V_j$，$(j\leqslant m)$，V_j 为属性 a_j 的值域。

定义2.2 设 (U,A,F) 为信息系统，其中 U 为有限对象集，称 $B\subseteq A$ 是 (U,A,F) 的约简，是指 $R_B=R_A$，且对任意 $b\in B$，$R_{B-\{b\}}\neq R_A$。

要在 A 的所有子集中选择满足上述条件的约简属性 B，在计算上是很复杂的。这是因为，如果 A 有 m 个元素，这种子集个数为 2^m 个，一般说来是不可能去一一验证的，因此如何降低算法的复杂性是知识约简的一个重要课题。

定理2.1[9] 对于任何信息系统 (U,A,F)，约简总是存在的。

一般说来，信息系统的约简不一定是唯一的，且两个不同的约简中含有属性的个数也不一定是相等的。

例2.1 表1给出了一个信息系统。

表1 一般信息系统

U	a_1	a_2	a_3	a_4
x_1	2	1	3	2
x_2	3	2	1	1
x_3	2	1	3	2
x_4	1	1	4	3
x_5	1	1	2	3
x_6	1	1	4	3
x_7	1	2	3	2
x_8	1	2	3	2

由属性集 $A=\{a_1,a_2,a_3,a_4\}$ 生成的等价关系为：$R_A=\{(x_1,x_1),(x_1,x_3),(x_2,x_2),(x_3,x_1),(x_3,x_3),(x_4,x_4),(x_4,x_6),(x_5,x_5),(x_6,x_4),(x_6,x_6),(x_7,x_7),(x_7,x_8),(x_8,x_7),(x_8,x_8)\}$。$B_1=\{a_1,a_3\}$ 对于属性集 $B_1=\{a_1,a_3\}$ 和 $B_2=\{a_2,a_3\}$，易见：$R_{B_1}=R_{B_2}=R_A$，但是 $R_{a_1}\neq R_A$，$R_{a_2}\neq R_A$，$R_{a_3}\neq R_A$。于是属性集 B_1 和 B_2 都是信息系统的约简，且容易验证 A 的其它真子集都不是信息系统的约简。

如果用 $\{B_i:i\leqslant l\}$ 表示信息系统的所有约简，则 $K=\bigcap_{i=1}^{l}B_i$ 为信息系统的核心。一般说来，核心不一定构成约简，若核心非空，则任何约简都包含核心，即核心中的元素是任何约简都

邵明文 博士.张文修 博导，西安交通大学理学院教授.吴伟志 博士，浙江海洋学院数学系副教授.

必需的，因而是绝对必要属性。$B'=\bigcup_{i=1}^{l}B_i-\bigcap_{i=1}^{l}B_i$，$B'$中的元素不出现在任何约简中，但出现在某些约简中，是相对必要属性。$B''=-\bigcup_{i=1}^{l}B_i$是绝对不必要属性，它不出现在任何约简中。

在例2.1中，由于信息系统共有两个约简$B_1=\{a_1,a_3\}$和$B_2=\{a_2,a_3\}$，且$B_1\cap B_2=\{a_3\}$，因此a_3是核心元素。它是绝对必要属性，而a_1和a_2是相对必要属性，a_4是绝对不必要属性。

定义2.3 设(U,A,F)为信息系统，由R_A确定的分划为：$A=U/R_A=\{C_i:i\leqslant l\}$，用$f_l(C_i)$表示属性$a_l$关于$C_i$中的对象的值，称：$D(C_i,C_j)=\{a_l\in A:f_l(C_i)\neq f_l(C_j)\}$为$C_i$与$C_j$的可辨识属性集，称：$\partial=(D(C_i,C_j):i,j\leqslant l)$为信息系统的可辨识属性矩阵[10]。

性质2.1 信息系统(U,A,F)的可辨识性矩阵∂，有以下性质：(1)$D(C_i,C_j)=\emptyset(\forall\ i\leqslant l)$；(2)$D(C_i,C_j)=D(C_j,C_i)$$(\forall\ i,j\leqslant l)$；(3)$D(C_i,C_j)\neq\emptyset(\forall\ i\neq j)$；(4)$D(C_i,C_j)\subseteq D(C_i,C_k)\cup D(C_k,C_j)(\forall\ i,k,j\leqslant l)$。

证明：(1)与(2)是显然的，现证(3)(4)。

证明(3)：若某$D(C_i,C_j)=\emptyset(i\neq j)$则$C_i=C_j$，因为$i\neq j$，这与$\{c_i:i\leqslant l\}$为$U$的分划矛盾。□

证(4)：若：$a_l\notin D(C_i,C_k)\cup D(C_k,C_j)$，则$a_l\notin D(C_i,C_k)$，且$a_l\notin D(C_k,C_j)$，于是$f_l(C_i)=f_l(C_k)$且$f_l(C_k)=f_l(C_j)$，从而$f_l(C_i)=f_l(C_j)$，这样就有$a_l\notin D(C_i,C_j)$。□

如果记$d(C_i,C_j)=|D(C_i,C_j)|$，则d是分划A上的一个度量，即满足：

(1)$d(C_i,C_j)=0(\forall\ i\leqslant l)$；

(2)$d(C_i,C_j)=d(C_j,C_i)(\forall\ i,j\leqslant l)$；

(3)$d(C_i,C_j)\leqslant d(C_i,C_k)+d(C_k,C_j)(\forall\ i,k,j\leqslant l)$。

例2.2 在例2.1中，通过等价关系R_A可以将U分成5类：$A=\{C_1,C_2,C_4,C_4,C_5\}$，对应的信息系统可以简化为表2。

表2 简化后的信息系统

A	a_1	a_2	a_3	a_4
$C_1=\{x_1,x_3\}$	2	1	3	2
$C_2=\{x_2\}$	3	2	1	1
$C_3=\{x_7,x_8\}$	1	2	3	2
$C_4=\{x_4,x_6\}$	1	1	4	3
$C_5=\{x_5\}$	1	1	2	3

可辨识属性矩阵见表3。

表3 可辨识属性矩阵∂

A	C_1	C_2	C_3	C_4	C_5
C_1	$\emptyset$	A	$\{a_1,a_2\}$	$\{a_1,a_3,a_4\}$	$\{a_1,a_3,a_4\}$
C_2	A	$\emptyset$	$\{a_1,a_3,a_4\}$	A	A
C_3	$\{a_1,a_2\}$	$\{a_1,a_3,a_4\}$	$\emptyset$	$\{a_2,a_3,a_4\}$	$\{a_2,a_3,a_4\}$
C_4	$\{a_1,a_3,a_4\}$	A	$\{a_2,a_3,a_4\}$	$\emptyset$	$\{a_3\}$
C_5	$\{a_1,a_3,a_4\}$	A	$\{a_2,a_3,a_4\}$	$\{a_3\}$	$\emptyset$

定理2.2[9] ∂为信息系统(U,A,F)的可辨识属性集，若$B\subseteq A$使$B\cap D(C_i,C_j)\neq\emptyset,(\forall\ i\neq j,i,j\leqslant l)$当且仅当$R_B=R_A$。

定理2.3 $a\in A$，a为核心元素，当且仅当：存在$C_i,C_j(i\neq j)$，使$D(C_i,C_j)=a$。

证明(⇒)为证a为核心元素，我们只须证$R_A\neq R_{A-\{a\}}$，事实上，由于$D(C_i,C_j)=\{a\}$，则当$x_i\in C_i,x_j\in C_j$时有$f_l(x_i)=f_l(x_j),(\forall\ a_l\in A-\{a\})$。于是$[x_i]_{A-\{a\}}=C_i\cup C_j=[x_j]_{A-\{a\}}$。从而$U/R_{A-\{a\}}\neq U/R_A$，即$R_A\neq R_{A-\{a\}}$。

(⇐)用反证法。假设不存在$C_i,C_j(i\neq j)$使$D(C_i,C_j)=\{a\}$。即：$\forall\ \ C_i,C_j(i\neq j),2\leqslant d(C_i,C_j),\subseteq A-a\subset A$，我们在每一个$D(C_i,C_j)$中取一个异于$a$的元素生成集合$B$，显然$B\subseteq A-a\subset A$，且$B\cap D(C_i,C_j)\neq\emptyset,(\forall\ \ i\neq j,i,j\leqslant l)$。由定理2.2知$R_B=R_A$，故$R_{A-a}=R_A$，这与$a$为核心元素矛盾。□

注：由定理2.2可知信息系统(U,A,F)的属性集A的核心元素的个数就是可辨识属性矩阵$\partial=(D(C_i,C_j):i,j\leqslant l)$中单点集的个数。

推论2.1[9] 记$\partial_0=\{D(C_i,C_j):i\neq j\}$，$R_A=R_B$，当且仅当对于任意$D\subseteq A,B\cap D=\emptyset$，必有$D\notin\partial_0$。

证明，理2.2易证。□

定理2.4 若存在$C_i,C_j(i\neq j),d(C_i,C_j)=2$，且$D(C_i,C_j)$不含单点集中的元素，则$D(C_i,C_j)$中的元素皆为相对必要属性。

证明：用反证法。不妨设$D(C_i,C_j)=\{a,b\}$，我们假设a为绝对不必要属性，即A的任何约简都不含a。令B为A的一个约简，由定理2.2知$B\cap D(C_i,C_j)\neq\emptyset$；因为$a\notin B$，故$b\in B$，即$A$的任何约简都包含$b$，则$b$为绝对必要属性这与条件和定理2.3知这是一个矛盾。因此a为相对必要属性；同理可证b为相对必要属性。□

定理2.5 信息系统(U,A,F)中，$c\in A$为绝对不必要属性当且仅当：$\forall$包含c元素的$D(C_i,C_j)$，存在$D(C_m,C_n)$，$c\notin D(C_m,C_n)$，满足$D(C_m,C_n)\subset D(C_i,C_j)$。

证明：(⇒)我们在每一个$D(C_i,C_j),(j<i\leqslant l)$中取一个元素合起来生成集合$B$，其中含有$c$的$D(C_i,C_j)$只能选取$c$，这样$B\cap D(c_i,C_j)\neq\emptyset(j<i\leqslant l)$，由定理2.2知$R_B=R_A$，因为$c$为绝对不必要属性，则$R_{B-\{c\}}=R_A$，即：$\{B-\{c\}\}\cap D(C_i,C_j)\neq\emptyset(j<i\leqslant l)$。假设对任意一个含有$c$元素$D(C_i,C_j)$，在所有不含$c$的$D(C_i,C_j)$中不存在$D(C_m,C_n)$，满足$D(C_m,C_n)\subset D(C_i,C_j)$，我们在每一个不含$c$的$D(C_i,C_j)$中取一个不包含于$D(C_m,C_n)$的元素，含有$c$的$D(C_i,C_j)$选取$c$，合起来生成集合$B$，这样$B\cap D(C_i,C_j)\neq\emptyset(j<i\leqslant l)$，由定理2.2知$R_B=R_A$，因为$c$为绝对不必要属性，则$R_{B-\{c\}}=R_A$，即：$\{B-\{c\}\}\cap D(C_i,C_j)\neq\emptyset$而由我们的选取知：$D(C_m,C_n)\cap\{B-\{c\}\}=0$，矛盾。故对任意一个含有$c$元素的$D(C_i,C_j)$，存在一个$D(C_m,C_n)$，满足$D(C_m,C_n)\subset D(C_i,C_j)$，进一步若$D(C_m,C_n)$还有其它绝对不必要属性，用同样的方法一定可以找到$D(C_m,C_n)\subset D(C_i,C_j)$，且$D(C'_m,C'_n)$中不含完全不必要属性。

(⇐)设$B\subseteq A$为A的任意一个约简，我们证明$c\notin B$。若$c\in B$，因为$B\cap D(C_i,C_j)\neq\emptyset$且任意一个包含$c$的$D(C_i,C_j)$，存在$D(C_m,C_n)\subset D(C_i,C_j)$。所以$B\cap D(C_i,C_j)\neq\emptyset$，故$R_{B-\{c\}}=R_A$，这与$B$为$A$的一个约简矛盾，即$c\notin B$，所以为绝对不必要属性。□

定理2.6 信息系统(U,A,F)中，记$A=\{D(C_i,C_j)|b\in D(C_i,C_j)\}$，$B=\{D(C_i,C_j)|b\notin D(C_i,C_j)\}$，$b\in A$不是核心元素，则$b$为相对必要属性当且仅当：存在$D(C_i,C_j)\in\Phi$，满足任意$D(C_m,C_n)\in B$，$D(C_m,C_n)\not\subset D(C_i,C_j)$。

证明：由定理2.5即得。□

由辨识公式求信息系统的约简计算量是很大的，为了降低计算量我们首先由辨识矩阵利用定理2.3和定理2.4确定核心属性和去掉绝对不必要属性，然后再利用辨识公式计算。

实际上，如果$B\subseteq A$是满足条件$B\cap D(C_i,C_j)\neq\emptyset(i>j)$为极小子集(关于包含)，则$B$是$A$的一个约简。换句话说约

简是满足能区别由整个属性集区别的所有对象的属性极小子集。

下边给出一个由可辨识属性矩阵求信息系统的一个约简的简便算法：

定理2.7 设(U,A,F)为信息系统，$\partial=(D(C_i,C_j):i,j\leqslant l)$为信息系统的可辨识属性矩阵，则如下所取集合$B$为$(U,A,F)$的一个最小属性约简：

第一步：首先把可辨识属性矩阵中包含单点集的$D(C_i,C_j)$用该单点集替换，记这时的信息系统的可辨识属性矩阵为$\partial_1=(D_1(C_i,C_j):i,j\leqslant l)$。

第二步：在$\partial_1=(D_1(C_i,C_j):i,j\leqslant l)$的非单点集中取$D_1(C_i,C_j)$使$D(C_i,C_j)$为极小，取$b\in D_1(C_i,C_j)$，使在可辨识属性矩阵中含有属性$b$的可辨识属性集的个数为极大。然后，把含有$b$的$D_1(C_i,C_j)$全部用单点集$b$替换，记这时得到的可辨识属性矩阵为$\partial_2=(D_2(C_i,C_j):i,j\leqslant l)$。

第三步：重复第二步。由于A为有限集，所以经过有限步以后，得可辨识属性矩阵$\partial_n=(D_n(C_i,C_j):i,j\leqslant l)$。满足$D_n(C_i,C_j)(i,j\leqslant l)$皆为单点集。

第四步：把所有$D_n(C_i,C_j)(i,j\leqslant l)$中的元素取并集得集合$B$。

证明：由定理2.5知B中不存在绝对不必要属性。由取法知B中不含有重复元素，在B中任意去掉一个元素形成的B'，总存在$i\neq j$，使$B'\cap D(C_i,C_j)=\emptyset$。因此$B'$为信息系统的约简。由于在第二步中，每次所取的元素$b$都使在可辨识属性矩阵中含有属性$b$的可辨识属性集的个数为极大，故$B$为$(U,A,F)$的一个最小属性约简。□

例2.3 对例2.1的可辨识属性矩阵∂，由表3：

第一步：把含有a_3的$D(C_i,C_j)$全部替换为$\{a_3\}$，得表4：

表4 可辨识属性矩阵$\partial_1=(D_1(C_i,C_j):i,j\leqslant l)$

A	C_1	C_2	C_3	C_4	C_5
C_1	$\emptyset$	A	$\{a_1,a_2\}$	$\{a_3\}$	$\{a_3\}$
C_2	a_3	$\emptyset$	$\{a_3\}$	a_3	a_3
C_3	$\{a_1,a_2\}$	$\{a_3\}$	$\emptyset$	$\{a_3\}$	$\{a_3\}$
C_4	$\{a_3\}$	$\{a_3\}$	$\{a_3\}$	$\emptyset$	$\{a_3\}$
C_5	$\{a_3\}$	$\{a_3\}$	$\{a_3\}$	$\{a_3\}$	$\emptyset$

第二步：在$D_1(C_1,C_3)=\{a_1,a_2\}$中，取a_1，然后把表4中含有a_1的$D_1(C_i,C_j)$全部替换为$\{a_1\}$，得表5。

表5 可辨识属性矩阵$\partial_2=\{D_2(C_i,C_j):i,j\leqslant l)$

A	C_1	C_2	C_3	C_4	C_5
C_1	$\emptyset$	A	$\{a_1\}$	$\{a_3\}$	$\{a_3\}$
C_2	$\{a_3\}$	$\emptyset$	$\{a_3\}$	$\{a_3\}$	$\{a_3\}$
C_3	$\{a_1\}$	$\{a_3\}$	$\emptyset$	$\{a_3\}$	$\{a_3\}$
C_4	$\{a_3\}$	$\{a_3\}$	$\{a_3\}$	$\emptyset$	$\{a_3\}$
C_5	$\{a_3\}$	$\{a_3\}$	$\{a_3\}$	$\{a_3\}$	$\emptyset$

第三步：把所有$\partial_2=(D_2(C_i,C_j):i,j\leqslant l)$中得元素取并集得集合$B=\{a_1,a_3\}$，$B$即为原信息系统得一个约简。

3 协调目标信息系统的知识约简

设(U,A,F,d,G)为目标信息系统，其中(U,A,F)为信息系统，$G:U\to V_d$。对于任意$B\subseteq A$，$A(B)=\{[x]_B:x\in U\}$。

同样记$R_D=\{(x_i,x_j):f_l(x_i)=f(x_j),\forall(d_l\in D)\}$，$A(D)=\{[x]_B:x\in U\}$。其中$[x]_B=\{y:(x,y)\in R_B\}$，$[x]_D=\{y:(x,y)\in R_D\}$。

定义3.1 设(U,A,F,d,G)为目标信息系统，若$R_A\subseteq R_D$，即$A(A)\leqslant A(D)$，则称目标信息系统是协调的，否则称目标信息系统是不协调的。

定义3.2 对于协调的目标信息系统，若存在$B\subseteq A$，使$R_B\subseteq R_D$，即$A(B)\leqslant A(D)$，且对于任意$b\in B$，$R_{B-\{b\}}\not\subseteq R_D$，即$A(B-\{b\})\leqslant A(D)$不成立，则称$B$是目标信息系统的一个约简。

定理3.1 协调的目标信息系统的约简总是存在的。

证明：类似于定理2.1可证。□

一般来说协调的目标信息系统的约简不一定是唯一的。如果用$\{B_i,i\leqslant l\}$表示协调的目标信息系统的所有约简，则$B=\bigcap_{i=1}^{l}B_i$为核心。一般说来，核心不一定存在。即使存在也不一定构成目标信息系统的约简，但这时所有的约简都包含核心。若核心非空，则核心中的任何元素是目标信息系统的约简都必需的，因而是绝对必要属性。$B'=\bigcup_{i=1}^{l}B_i-\bigcap_{i=1}^{l}B_i$，是相对必要属性集，它们可能出现在部分约简中，而$B''=A-\bigcup_{i=1}^{l}B_i$是绝对不必要属性，它不出现在任何约简中。

在应用中，一个分类相对于另一个分类的关系十分重要，因此我们定义一个分类相对于另一个分类的正域。

令P和Q为U中的等价关系，Q的P正域记为$POS_P(Q)$[11]，即：$POS_P(Q)=\bigcup_{X\in U/Q}\underline{P}X$。$Q$的$P$正域是所有根据分类$U/P$的信息准确的划分到关系$Q$的价类中去的对象集合。

定义3.3 设(U,A,F,E,G)为协调目标信息系统，$D_E(C_i,C_j)=\{a\in A|G_i(C_i)\neq G_i(C_j)$，且$W(C_i,C_j)\}$，对于$x,y\in U$，$W(x,y)$满足：$x\in POS_A(E)$且$y\notin POS_A(E)$；或者$x\notin POS_A(E)$且$y\in POS_A(E)$；或者$x,y\in POS_A(E)$且$(x,y)\notin ind(D)$，称$D_E(C_i,C_j)$为$C_i$与$C_j$关于$E$的可辨识属性集。称$\partial_E=(D_E(C_i,C_j):i,j\leqslant l)$为目标信息系统的可辨识属性矩阵。

性质3.1 协调目标信息系统(U,A,F,E,G)的可辨识属性矩阵∂_E具有以下性质：(1)$D_E(C_i,C_j)=\emptyset(\forall\ i\leqslant l)$；(2)$D_E(C_i,C_j)=D_E(C_j,C_i)(\forall\ i,j\leqslant l)$；(3)$D_E(C_i,C_j)\subseteq D_E(C_i,C_k)\cup D_E(C_k,C_j)(\forall\ i,k,j\leqslant l)$。

证明：类似于性质2.1可证。□

例3.1 表6是一个病历诊断系统：a_1,a_2,a_3为三种症状，10个病例：$x_1\cdots,x_{10}$，d为目标属性；

表6 决策表

U	a_1	a_2	a_3	d
x_1	2	1	3	1
x_2	3	2	1	2
x_3	2	1	3	1
x_4	2	2	3	2
x_5	1	1	4	3
x_6	1	1	2	3
x_7	3	2	1	2
x_8	1	1	4	3
x_9	2	1	3	1
x_{10}	3	2	1	2

表7给出了例3.1中目标信息系统的可辨识属性矩阵∂_d。

定理3.2 设$D_E(C_i,C_j)(\forall\ i,j\leqslant l)$为协调目标信息系统$(U,A,F,E,G)$的可辨识属性集，若$B\subseteq A$，使$\forall\ D_E(C_i,C_j)\neq$

$\emptyset$,($\forall\ i,j\leqslant l$)满足 $B\bigcap D_E(C_i,C_j)\neq\emptyset$当且仅当:$R_B=R_A$。

证明:类似于定理2.2可证。 □

表7 表6的可辨识属性矩阵

A	C_1	C_2	C_3	C_4	C_5
C_1	$\emptyset$	$\emptyset$	$\{a_2\}$	$\{a_1,a_3\}$	$\{a_1,a_3\}$
C_2	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_3	$\{a_2\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_4	$\{a_1,a_3\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_5	$\{a_1,a_3\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

定理3.3 $D_E(C_i,C_j)$($\forall\ \ i,j\leqslant l$)为协调目标信息系统$(U,A,F,E,G)$的可辨识属性集$a\in A$,$a$为核心元素,当且仅当:存在$C_i,C_j(i\neq j)$,使$D_E(C_i,C_j)=a$。

证明:类似于定理2.3可证。 □

推论3.1 $D_E(C_i,C_j)$($\forall\ \ i,j\leqslant l$)为协调目标信息系统$(U,A,F,E,G)$的可辨识属性集,记$\partial_0=\{D_E(C_i,C_j):i\neq j\}$,$R_A=R_B$,当且仅当对于任意$D\subseteq A$,$E\bigcap D=\emptyset$,必有$D\notin\partial_0$。

证明:由定理3.2易证。 □

定理3.4 协调目标信息系统(U,A,F,E,G)中,$c\in A$为绝对不必要属性当且仅当:$\forall$包含c元素的$D_E(C_i,C_j)$,存在$D_E(C_m,C_n)$,满足$c\notin D_E(C_m,C_n)$,$D_E(C_m,C_n)\subset D_E(C_i,C_j)$。

证明:类似于定理2.5可证。 □

定理3.5 协调目标信息系统(U,A,F,E,G)中,记$A=\{D_E(C_i,C_j)\mid b\in D_E(C_i,C_j)\}$,$B=\{D_E(C_i,C_j)\mid b\notin D_E(C_i,C_j)\}$,$b\in A$不是核心元素,$b$为相对必要属性当且仅当:存在$D_E(C_i,C_j)\in A$,使任意$D(C_m,C_n)\in B$,$D_E(C_m,C_n)\not\subset D_E(C_i,C_j)$。

证明:由定理3.4即得。 □

由辨识公式求协调目标信息系统的约简计算量是很大的,为了降低计算量我们首先由辨识矩阵利用定理3.2和定理3.4确定核心属性和去掉绝对不必要属性,然后再利用辨识公式计算。

实际上,如果$B\subseteq A$是满足条件$B\bigcap D_E(C_i,C_j)\neq\emptyset$($D_E(C_i,C_j)\neq\emptyset$)的极小子集(关于包含),则$B$是$A$的一个约简。

下边给出一个由可辨识属性矩阵求协调目标信息系统的一个约简的简便算法:

定理3.6 设(U,A,F,E,G)为协调目标信息系统,$\partial_E=(D_E(C_i,C_j):i,j\leqslant l)$为协调目标信息系统的可辨识属性矩阵,则如下所取集合B为(U,A,F,E,G)的属性约简:

第一步:首先把可辨识属性矩阵中包含单点集的$D_E(C_i,C_j)$用该单点集替换,记这时的目标信息系统的可辨识属性矩阵为$\partial_1=(D_{E1}(C_i,C_j):i,j\leqslant l)$。

第二步:在$\partial_1=(D_{E1}(C_i,C_j):i,j\leqslant l)$的非单点集中取$D_{E1}(C_i,C_j)$使$d(C_i,C_j)$为极小,取$b\in D_{E1}(C_i,C_j)$使在可辨识属性矩阵中含有属性$b$的可辨识属性集的个数为极大。然后,把含有$b$的$D_{E1}(C_i,C_j)$全部用单点集$b$替换,记这时得到的可辨识属性矩阵为$\partial_2=(D_{E2}(C_i,C_j):i,j\leqslant l)$。

第三步:重复第二步。由于A为有限集,所以经过有限步以后,直至$|D_{En}(C_i,C_j)|=1$或0,($\forall\ \ i,j\leqslant l$)。

第四步:把所有$D_{En}(C_i,C_j)(i,j\leqslant l)$中的元素取并集得集合$B$。

证明:类似于定理2.7可证。 □

例3.2 对表7所给例3.1中目标信息系统的可辨识属性矩阵∂_E,我们用定理3.6求协调目标信息系统的属性约简:

第一步:把含有a_2的$D_E(C_i,C_j)$全部替换为$\{a_2\}$,得表8:

表8 可辨识属性矩阵$\partial_1=(D_{E1}(C_i,C_j):i,j\leqslant l)$

A	C_1	C_2	C_3	C_4	C_5
C_1	$\emptyset$	$\emptyset$	$\{a_2\}$	$\{a_1,a_3\}$	$\{a_1,a_3\}$
C_2	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_3	$\{a_2\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_4	$\{a_1,a_3\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
c_5	$\{a_1,a_3\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

第二步:在$D_{E1}(C_1,C_4)=\{a_1,a_3\}$中,取$a_1$,然后把表8中含有$a_1$的$D_{E1}(C_i,C_j)$全部替换为$\{a_1\}$,得表9:

表9 可辨识属性矩阵$\partial_2=(D_{E2}(C_i,C_j):i,j\leqslant l)$

A	C_1	C_2	C_3	C_4	C_5
C_1	$\emptyset$	$\emptyset$	$\{a_2\}$	$\{a_1\}$	$\{a_1\}$
C_2	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_3	$\{a_2\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_4	$\{a_1\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
C_5	$\{a_1\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

第三步:把所有$\partial_2=(D_{E2}(C_i,C_j):i,j\leqslant l)$中的元素取并集得集合$B=\{a_1,a_2\}$,$B$即为原协调目标信息系统得一个约简。

结语 本文利用可辨识属性矩阵,确定了信息系统的核心属性和去掉绝对不必要属性,并给出一个由可辨识属性矩阵求信息系统的一个约简的简化算法。实际计算结果表明,该算法容易实现,计算量相对较少,而且在任何情况下总能求得一个最小属性约简。

参考文献

1 Pawlak Z. Rough sets. International Journal of Computer and Information Science, 1982, 11: 341～356

2 Pawlak Z. Rough Sets Theory and It's Application to Data Analysis [J]. Cybernetics Systems, An International Journal, 1998, 29: 661～688

3 Aijun An, et al. Applying Knowledge Discovery to Predict Water-Supply Consumption[J]. IEEE Expert, 1997. 72～78

4 Pawlak Z, Slowinski R. Rough set approach to multiattribute decision analysis, invited review [J]. European Journal of Operational Research, 1994, 72: 443～459

5 Wong S K M, Ziarko W. On optional decision rules in decision tables[J]. Bulletin of Polish Academy of Science, 1985, 33: 693～696

6 Hu X H, Cercone N. Learning in Relational Database: a Rough Set Approach[J]. Computational Intelligence, 1995, 11(2): 323～338

7 Pawlak Z, Grzymala-busse J, Slowinski, et al. Rough Set [J]. Communications of the ACM, 1995, 38(11): 89～95

8 常梨云,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

9 张文修,梁怡,吴伟志. 信息系统与知识发现. 北京:科学出版社,将出版

10 Skowron A. The rough sets theory and evidecne theory. Fundamenta Informatica. , XIII, 1990. 145～162

11 张文修,吴伟志,梁吉业,李德玉. 精糙集理论与方法. 北京:科学出版社,2001