基于优势关系下不协调目标信息系统的知识约简 *)

徐伟华　张文修

(西安交通大学理学院信息与系统科学研究所　西安 710049)

摘　要　在基于优势关系下不协调目标信息系统中引入了分配约简和近似约简的概念,并讨论了它们二者之间的关系,进一步给出了知识约简的判定定理和辨识矩阵,从而提供了在优势关系下不协调目标信息系统知识约简的具体操作方法。

关键词　粗糙集,信息系统,知识约简,辨识矩阵

Knowledge Reductions in Inconsistent Information Systems Based on Dominance Relations

XU Wei-Hua　ZHANG Wen-Xiu

(Institute of Information and System Sciences, Faculty of Science, Xi'an Jiaotong University, Xi'an 710049)

Abstract　Knowledge reduction is one of the most important problems in rough set theory. However, most of information systems are not only inconsistent, but also based on dominance relations because of various factors. To acquire brief decision rules from inconsistent systems based on dominance relations, knowledge reductions are needed. The main aim of this paper is to study the problem. The assignment reduction and approximation reduction are introduced in inconsistent systems based on dominance relations, and relationships between them are examined. The judgment theorem and discernable matrix are obtained, from which we can provide new approach to knowledge reductions in inconsistent systems based on dominance relations.

Keywords　Rough set, Information system, Knowledge reduction, Discernibility matrix

1　引言

粗糙集理论[1]是近年来发展起来的一种处理不精确性、不确定性和模糊知识的软计算工具,它已被成功地应用于人工智能、数据挖掘、模式识别与之能信息处理等领域[2~5],并越来越引起了国际学术界的关注。其在保持信息系统分类能力不变的前提下,通过知识约简,导出问题的分类能力或决策规则。

经典粗糙集是以完备信息系统为研究对象,以等价关系(满足自反性、对称性、传递性)为基础,通过等价关系对论域分成互不相交的等价类,划分越细,知识越丰富,信息越充分。

知识约简是粗糙集理论的核心问题之一。在实际的知识库中描述知识的属性并不是同等重要的,甚至其中有些属性是冗余的。所谓知识约简,就是在保持知识库分类能力不变的条件下,删除其中不相关或不重要的属性。通过知识约简去掉不必要的属性,可以使知识表示简化,又不丢失基本信息。目前,许多学者通过不同的方法从不同的角度对知识约简做了深入的研究,并取得了很多成果[6~10]。

然而,这些研究主要是在等价关系下的信息系统进行的,在实际问题中有许多信息系统由于各种原因(如噪声、信息缺损等)是基于优势关系的,而且是不协调的。要想从这种复杂的基于优势关系的不协调信息系统中获取简洁的不确定性命题,就必须对系统进行知识约简。因而,对于优势关系下的不协调目标信息系统知识约简的研究是非常有意义的。为此,本文对这一问题进行了探讨研究,给出了优势关系下不协调信息系统的分配约简和近似约简的概念,并证明了它们二者是等价的,进一步给出了知识约简的判定定理和辨识矩阵,从而提供了知识约简的具体操作方法。

2　基于优势关系的目标信息系统

目标信息统是既有条件属性又有目标属性(决策属性)的一种特殊信息系统。目标信息系统主要是研究条件属性和目标属性之间的关系问题。为了方便理解,下面先给出一些基本概念和术语。

定义 1　称一个五元组(U,A,F,D,G)为一个目标信息系统,其中(U,A,F)是信息系统,A称为条件属性集,D称为目标属性集,即:

U是有限对象集,$U=\{x_1,x_2,\cdots,x_n\}$;

A是有限条件属性集,$A=\{a_1,a_2,\cdots,a_p\}$;

D是有限目标属性集,$D=\{d_1,d_2,\cdots,d_q\}$;

F是U与A的关系集,$F=\{f_k:U\to V_k,k\leqslant p\}$,$V_k$是$a_k$的有限值域;

G是U与D的关系集,$G=\{g_{k'}:U\to V'_k,k'\leqslant q\}$,$V'_k$是$d_{k'}$的有限值域。

我们知道,在Pawlak近似空间意义下的信息系统,对每个属性集和目标属性集就决定了一个二元不可区分关系,即等价关系。然而,在实际生活中有许多系统并不是基于等价关系的,有不少是基于优势关系的,即对每个属性值域和目标属性值域有一个偏序关系,如一个班级的各科成绩情况等问题。这时就需要建立基于优势关系下的信息系统。

定义 2　设(U,A,F,D,G)为目标信息系统,对于$B\subseteq A$,令

$R_B=\{(x_i,x_j)\in U\times U: f_l(x_i)\leqslant f_l(x_j),\forall a_l\in B\}$,

$R_D=\{(x_i,x_j)\in U\times U: g_m(x_i)\leqslant g_m(x_j),\forall d_m\in D\}$,

*)国家"九七三"计划项目资助(项目号:2002CB31200)。**徐伟华**　博士研究生,主要研究人工智能、粗糙集;**张文修**　博导,教授,主要研究模糊集、粗糙集、人工智能。

R_B，R_D 称为目标信息系统的优势关系，此时该目标信息统称为是基于优势关系下的目标信息系统。

记：$[x_i]_B=\{x_j\in U:(x_i,x_j)\in R_B\}=\{x_j\in U:f_l(x_i)\leqslant f_l(x_j),\forall a_l\in B\}$，

$[x_i]_D=\{xj\in U:(x_i,x_j)\in R_D\}=\{x_j\in U:g_m(x_i)\leqslant g_m(x_j),\forall d_m\in D\}$。

易见，优势关系有下面性质：

命题 1 (1)R_B 是自反的和传递的，未必是对称的，因而一般不再是等价关系。

(2)当 $B_1\subseteq B_2\subseteq A$ 时有：$R_A\subseteq R_{B_2}\subseteq R_{B_1}$。

(3)当 $B_1\subseteq B_2\subseteq A$ 时有：$[x_i]_A\subseteq[x_i]_{B_2}\subseteq[x_i]B_1$。

(4)当 $x_j\in[x_i]_B$ 时有：$[x_j]_B\subseteq[x_i]_B$。

对于任意 $X\subseteq U$，定义 X 关于优势关系下 R_B 的下近似和上近似分别为：

$\underline{R_B}(X)=\{x_i\in U:[x_i]\subseteq X\}$，$\overline{R_B}(X)=\{x_i\in U:[x_i]\cap X=\varnothing\}$。

优势关系下的上、下近似也满足类似于 Pawlak 近似空间中的许多性质，详细请参考文[7]。

定义 3 设(U,A,F,D,G)为基于优势关系的目标信息系统，若 $R_A\subseteq R_D$，则称该基于优势关系的目标信息系统是协调的，或该目标信息系统在优势关系下是协调的，否则是不协调的。

例 1 表 1 给出了一个基于优势关系的目标信息系统。

表 1

$U\times(A\cup D)$	a_1	a_2	a_3	d
x_1	1	2	1	3
x_2	3	2	2	2
x_3	1	1	2	1
x_4	2	1	3	2
x_5	3	3	2	3
x_6	3	2	3	1

于是，按照优势关系的定义有：

$[x_1]_A=\{x_1,x_2,x_5,x_6\}$，$[x_2]_A=\{x_2,x_5,x_6\}$，$[x_3]_A=\{x_2,x_3,x_4,x_5,x_6\}$，

$[x_4]_A=\{x_4,x_6\}$，$[x_5]_A=\{x_5\}$，$[x_6]_A=\{x_6\}$

$[x_1]_d=[x_5]_d=\{x_1,x_5\}$，$[x_2]_d=[x_4]_d=\{x_1,x_2,x_4,x_5\}$，

$[x_3]_d=[x_6]_d=\{x_1,x_2,x_3,x_4,x_5,x_6\}$

显然，$R_A\not\subset R_d$。因此该目标信息系统在优势关系下是不协调的。

注：为了叙述方便，下文中在未作特殊说明的情况下，目标信息系统均指基于优势关系的目标信息系统。

3 不协调目标信息系统的知识约简理论

设(U,A,F,D,G)为目标信息系统，R_B，R_D 分别为属性集 A 和目标属性集 D 生成的 U 上的优势关系，对于 $B\subseteq A$，记

$U/R_B=\{[x_i]_B:x_i\in U\}$，

$U/R_D=\{D_1,D_2,\cdots,D_r\}$，

$\sigma_B(x)=\{D_j:D_j\cap[x]_B=\varnothing,x\in U\}$，

$\eta_B=-\frac{1}{U}\sum_{j=1}^{r}|\overline{R_B}(D_j)|$

其中，$[x]_B=\{y\in U:(x,y)\in R_B\}$

显然由以上可立即得到

命题 2 (1)$\overline{R_B}(D_j)=\bigcup\{[x]_B:D_j\in\sigma_B(x)\}$。

(2)当 $B\subseteq A$ 时，有 $\sigma_B(x)\supseteq\sigma_A(x)$，$\forall x\in U$。

(3)当$[x]_B\supseteq[y]_B$ 时，有 $\sigma_B(x)\supseteq\sigma_B(y)$，$\forall y\in U$。

定义 4 设(U,A,F,D,G)为目标信息系统。

(1)若 $\forall x\in U$，有 $\sigma_B(x)=\sigma_A(x)$，则称 B 是分配协调集，且 B 的任何真子集不是分配协调集，则称 B 为分配协调约简。

(2)若 $\forall x\in U$，有 $\eta_B=\eta_A$，则称 B 是近似协调集，且 B 的任何真子集不是近似协调集，则称 B 为近似协调约简。

由上定义可知，分配协调集保持所有对象的可能决策不变，而近似协调集保持每一个决策类的上近似不变。

例 2 对于表 1 给出的不协调目标信息系统

若记 $D_1=[x_1]_d=[x_5]_d$，$D_2=[x_2]_d=[x_4]_d$，$D_3=[x_3]_d=[x_6]_d$。

由上面定义可以得到：

$\sigma_A(x_1)=\sigma_A(x_2)=\sigma_A(x_3)=\sigma_A(x_5)=\{D_1,D_2,D_3\}$，

$\sigma_A(x_4)=\{D_2,D_3\}$，$\sigma_A(x_6)=\{D_3\}$

若取 $B=\{a_2,a_3\}$时，容易验证对于 $\forall x\in U$ 有：$[x]_A=[x]_B$，因此有 $\sigma_B(x)=\sigma_A(x)$，故 $B=\{a_2,a_3\}$是分配协调集，而且可以计算$\{a_2\}$，$\{a_3\}$均不是分配协调集，因此 $B=\{a_2,a_3\}$是一个分配约简。

若取$B'=\{a_1,a_3\}$时，

$[x_1]_{B'}=\{x_1,x_2,x_3,x_4,x_5,x_6\}$，$[x]_{B'}=\{x_2,x_5,x_6\}$，$[x_3]_{B'}=\{x_2,x_3,x_4,x_5,x_6\}$。

于是可知

$\sigma_{B'}(x_1)=\sigma_{B'}(x_2)=\sigma_{B'}(x_3)=\sigma_{B'}(x_5)=\{D_1,D_2,D_3\}$，

$\sigma_{B'}(x_4)=\{D_2,D_3\}$，$\sigma_{B'}(x_6)=\{D_3\}$

因此有 $\sigma_{B'}(x)=\sigma_A(x)$，$(x\in U)$。因而 $B'=\{a_1,a_3\}$是分配协调集，而且可以计算$\{a_1\}$也不是分配协调集，因此 $B'=\{a_1,a_3\}$是一个分配约简。进一步容易验证$\{a_1,a_2\}$不是分配协调集，因此该目标信息系统只有两个约简：$\{a_1,a_3\}$，$\{a_2,a_3\}$。

定理 1 设(U,A,F,D,G)为目标信息系统，则 $B\subseteq A$ 是分配协调集当且仅当 B 是近似协调集。

证明："⇒"设 B 是分配协调集，即 $\forall x\in U$，有 $\sigma_B(x)=\sigma_A(x)$，于是 $\forall j\leqslant r$ 有 $x\in\bar{R}_B(D_j)\Leftrightarrow[x]_B\cap D_j\neq\varnothing\Leftrightarrow D_j\in\sigma_B(x)\Leftrightarrow D_j\in\sigma_A(x)\Leftrightarrow[x]_A\cap D_j\neq\varnothing\Leftrightarrow x\in\bar{R}_A(D_j)$。

于是有 $\bar{R}(D_j)=\bar{R}_A(D_j)$，因此 $\eta_B=\eta_A$，即 B 是近似协调集。

"⇐"设 B 是近似协调集，则有 $\eta_B=\eta_A$，即$\sum_{j=1}^{r}|\overline{R_B}(D_j)|=\sum_{j=1}^{r}|\overline{R_A}(D_j)|$，又由于 $\forall j\leqslant r$，有 $\bar{R}_B(D_j)\supseteq\bar{R}_A(D_j)$，故 $\bar{R}_B(D_j)=\bar{R}_A(D_j)$，于是对 $\forall x\in U$ 有 $D_j\in\sigma_B(x)\Leftrightarrow[x]_B\cap D_j\neq\varnothing\Leftrightarrow x\in\bar{R}_B(D_j)\Leftrightarrow x\in\bar{R}_A(D_j)\Leftrightarrow\Leftrightarrow[x]_A\cap D_j\neq\varnothing\Leftrightarrow x\in\sigma_A(x)$。因此有 $\forall x\in U$，$\sigma_B(x)=\sigma_A(x)$成立，从而 B 是分配协调集。 □

推论 1 设(U,A,F,D,G)为目标信息系统，则 $B\subseteq A$ 是分配约简当且仅当 B 是近似约简。

定理 2 设(U,A,F,D,G)为目标信息系统，$B\subseteq A$，则 B 是分配协调集当且仅当对 $\forall x,y\in U$，当 $\sigma_A(x)\cap\sigma_A(y)\neq\sigma_A(y)$时有$[x]_B\cap[y]_B\neq[y]_B$。

证明："⇒"反证。

假设当 $\sigma_A(x)\cap\sigma_A(y)\neq\sigma_A(y)$时$[x]_B\cap[y]_B\neq[y]_B$ 不成立。故此时有$[x]_B\cap[y]_B=[y]_B$，即有$[x]_B\supseteq[y]_B$，由命题 2(3)知 $\sigma_B(x)\supseteq\sigma_B(y)$，而 B 又是分配协调集，于是有 $\sigma_A(x)\supseteq\sigma_A(y)$。即 $\sigma_A(x)\cap\sigma_A(y)=\sigma_A(y)$，矛盾。

"⇐"由命题 2(2)知，只需证 $\sigma_B(x)\subseteq\sigma_A(x)$即可。

由条件知，对$\forall x,y\in U$，若$\sigma_A(x)\cap\sigma_A(y)\neq\sigma_A(y)$成立，则$[x]_B\cap[y]_B\neq[y]_B$成立，因此对$\forall x,y\in U$，若$[x]_B\cap[y]_B=[y]_B$成立，则$\sigma_A(x)\cap\sigma_A(y)=\sigma_A(y)$成立，即$[x]_B\supseteq[y]_B$成立可推出$\sigma_A(x)\supseteq\sigma_A(y)$成立。

另外，任取$D_k\in\sigma_B(x)$，于是有$[x]_B\cap D_k\neq\varnothing$。不妨设$y\in[x]_B\cap D_k$，则$y\in D_k$且$y\in[x]_B$，故由命题1(4)知$[x]_B\supseteq[y]_B$成立。因此可得$\sigma_A(x)\supseteq\sigma_A(y)$。又$y\in[y]_A$，所以$y\in[y]_A\cap D_k$，即有$[y]_A\cap D_k\neq\varnothing$。因此$D_k\in\sigma_A(y)$，所以$D_k\in\sigma_A(x)$，故$\sigma_B(x)\subseteq\sigma_A(x)$成立。

推论2 设(U,A,F,D,G)为目标信息系统，$B\subseteq A$，则B是近似协调集当且仅当对$\forall x,y\in U$，当$\sigma_A(x)\cap\sigma_A(y)\neq\sigma_A(y)$时有$[x]_B\cap[y]_B\neq[y]_B$。

4 不协调目标信息系统的知识约简方法

第3节中的定理2给出了不协调目标信息系统的分配协调集的等价刻画，这是判断属性子集是否协调的理论所在，由此我们可进一步得出知识约简的方法。下面先给出辨识属性矩阵的概念。

定义5 设(U,A,F,D,G)为不协调目标信息系统，记

$$D^*=\{(x_i,x_j):\sigma_A(x_i)\not\subseteq\sigma_A(x_j)\}$$

用$f_{a_k}(x)$表示属性a_k关于对象x的取值。定义

$$D(x_i,x_j)=\begin{cases}\{a_k\in A,f_{a_k}(x_i)>f_{a_k}(x_j)\}, & (x_i,x_j)\in D^*\\ A, & (x_i,x_j)\notin D^*\end{cases}$$

称$D(x_i,x_j)$为x_i与x_j的分配可辨识属性集。矩阵$M=(D(x_i,x_j),x_i,x_j\in U)$称为该目标信息系统的分配辨识矩阵。

定理3 设(U,A,F,D,G)为不协调目标信息系统，$B\subseteq A$，则：B是分配协调集当且仅当$B\cap D(x,y)\neq\varnothing$，对$\forall(x,y)\in D^*$。

证明："⇒"对$\forall(x,y)\in D^*$，则有：$\sigma_A(x)\not\subseteq\sigma_A(y)$，于是有$\sigma_A(x)\cap\sigma_A(y)\neq\sigma_A(y)$，而$B$又是分配协调集，故由定理2有$[x]_B\cap[y]_B\neq[y]_B$。因此$[x]_B$与$[y]_B$的关系有三种情况：(1)$[x]_B\not\subseteq[y]_B$，(2)$[x]_B\cap[y]_B=\varnothing$，(3)$[x]_B\cap[y]_B\not\subseteq[x]_B$且$[x]_B\cap[y]_B\not\subseteq[y]_B$。下面证明在这三种情况下均有$B\cap D(x,y)\neq\varnothing$成立。

(1)若$[x]_B\not\subseteq[y]_B$时，则至少存在一个$z\in[y]_B$，但$z\notin[x]_B$。由$z\notin[x]_B$可知至少存在一个$a_k\in B$，使得$f_{a_k}(x)>f_{a_k}(z)$，而$z\in[y]_B$，则$f_{a_k}(y)\leqslant f_{a_k}(z)$。于是有：$f_{a_k}(x)>f_{a_k}(y)$，因此$a_k\in D(x,y)$，即有$B\cap D(x,y)\neq\varnothing$。

(2)若$[x]_B\cap[y]_B=\varnothing$时，事实上，此时必然至少存在一个$a_k\in B$，使得$f_{a_k}(x)>f_{a_k}(y)$，即$B\cap D(x,y)\neq\varnothing$。否则，若对$\forall a_l\in B$都有$f_{a_l}(x)\leqslant f_{a_l}(y)$，则$y\in[x]_B$，这与$[x]_B\cap[y]_B=\varnothing$矛盾。

(3)若$[x]_B\cap[y]_B\not\subseteq[x]_B$且$[x]_B\cap[y]_B\not\subseteq[y]_B$时，证明同(1)。因为此时也至少存在一个$z\in[y]_B$，但$z\notin[x]_B$。

因此，若B是分配协调集，则对$\forall(x,y)\in D^*$有$B\cap D(x,y)\neq\varnothing$。

"⇐"若对$\forall(x,y)\in D^*$有$B\cap D(x,y)\neq\varnothing$，则存在一个$a_k\in B$使得$a_k\in D(x,y)$，故有$f_{a_k}(x)>f_{a_k}(y)$，所以$y\notin[x]_B$，而又$y\in[y]_B$。因此有$[x]_B\cap[y]_B\neq[y]_B$，而另一方面$(x,y)\in D^*$，则$\sigma_A(x)\not\subseteq\sigma_A(y)$，于是有$\sigma_A(x)\cap\sigma_A(y)\neq\sigma_A(y)$。故有当$\sigma_A(x)\cap\sigma_A(y)\neq\sigma_A(y)$成立时，$[x]_B\cap[y]_B\neq[y]_B$也成立。由定理2知$B$是分配协调集。 □

定义6 设(U,A,F,D,G)为不协调目标信息系统，$M=(D(x_i,z_j),x_i,x_j\in U)$为其分配辨识矩阵，称

$$F=\wedge\{\vee\{a_k:a_k\in D(x_i,x_j)\},x_i,x_j\in U\}$$
$$=\wedge\{\vee\{a_k:a_k\in D(x_i,x_j)\},x_i,x_j\in D^*\}$$

为分配辨识公式。

定理4 设(U,A,F,D,G)为不协调目标信息系统，辨识公式F的极小析取范式为

$$F=\bigvee_{k=1}^{p}(\bigwedge_{s=1}^{q}a_s)$$

记

$B_k=\{a_s,s=1,2,\cdots,q_k\}$，则$\{B_k=1,2,\cdots,p\}$是所有分配约简形式的集合。

证明：对任意的$(x_i,x_j)\in D^*$，由极小析取范式的定义知$B_k\cap D(x_i,x_j)\neq\varnothing$，再有定理3知$B_k$是分配协调集。同时，$F=\bigvee_{k=1}^{p}$在$B_k$中去掉一个元素形成$B_k'$，则必然存在某个$(x_i,x_j)\in D^*$使得$B_k'\cap D(x_i,x_j)=\varnothing$，故$B_k'$不是分配协调集，从而$B_k'$是分配约简。而辨识公式中包含了所有的$D(x_i,x_j)$，因此不存在其他分配约简。

例3 对于表1给出的不协调目标信息系统的辨识矩阵如表2所示。

表2

x_i,x_j	x_1	x_2	x_3	x_4	x_5	x_6
x_1	A	A	A	A	A	A
x_2	A	A	A	A	A	A
x_3	A	A	A	A	A	A
x_4	a_1,a_3	a_3	a_1,a_3	A	a_3	A
x_5	A	A	A	A	A	A
x_6	a_1,a_3	a_3	A	a_1,a_2	a_3	A

故可得：

$$F=(a_1\vee a_2\vee a_3)\wedge(a_1\vee a_3)\wedge(a_1\vee a_2)\wedge a_3$$
$$=(a_1\wedge a_3)\vee(a_2\wedge a_3)$$

因此$\{a_1,a_3\}$与$\{a_2,a_3\}$是该不协调目标信息系统的所有分配约简。这与例2的结果是一致的。

结论 我们知道要想从复杂的基于优势关系的不协调信息系统中获取简洁的不确定性命题，就必须对系统进行知识约简。因此，对于优势关系下的不协调目标信息系统的知识约简的研究是非常有意义的。本文定义了优势关系下的不协调目标信息系统的分配约简和近似约简，并证明了它们是等价的，进一步给出了知识约简的判定定理和辨识矩阵，从而提供了在优势关系下不协调目标信息系统知识约简的具体操作方法。

参考文献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Boston: Kluwer Academic Publishers,1991

2 Pawlak Z. Rough Sets. Communication of the ACM, 1995, 38(1): 89~95

3 王国胤. Rough集理论与知识获取. 西安:西安交通大学出版社,2001

4 苗夺谦,王珏. 基于粗糙集的多变量决策树构造方法. 软件学报,1997,8(6): 425~431

5 王珏,王任,苗夺谦,等. 基于Rough Set理论的"数据浓缩". 计算机学报,1998,21(5): 393~400

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

7 张文修,梁怡,吴伟志. 信息系统与知识发现. 北京:科学出版社,2003

8 张文修,米据生,吴伟志. 不协调目标信息系统的知识约简. 计算机学报,2003,26(1): 12~18

9 Kryszkiewicz M. Comparative Studies of Alternative of Knowledge Reduction in Inconsistent Systems. Intelligent Systems, 2001,16(1): 105~120

10 Grecos,Matarazzo B, Slowinski R. Rough Approximation of Preference Relation by Dominance Relations. European Journal of Operational Research,1999,117:63~68