第35卷 第18期 计 算 机 工 程 2009年9月
Vol.35 No.18 Computer Engineering September 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)18—0191—03 文献标识码：A 中图分类号：TP18

优势关系下不协调目标信息系统的上近似约简

徐伟华[1]，张晓燕[1]，张文修[2]
(1. 重庆理工大学数理学院，重庆 400054；
2. 西安交通大学信息与系统科学研究所，西安 710049)

摘 要：针对基于优势关系下不协调目标信息系统中属性约简的复杂问题，提出基于优势关系下不协调目标信息系统上近似约简的概念，得到上近似约简的判定定理以及辨识矩阵，建立不协调目标信息系统的上近似约简的具体方法，同时通过实例验证该方法的有效性，从而为优势关系下信息系统的知识发现提供理论基础。

关键词：粗糙集；信息系统；上近似约简；辨识矩阵

Upper Approximation Reduction in Inconsistent Target Information System Based on Dominance Relations

XU Wei-hua[1], ZHANG Xiao-yan[1], ZHANG Wen-xiu[2]
(1. School of Mathematics and Physics, Chongqing University of Technology, Chongqing 400054;
2. Institute of Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049)

【Abstract】 This paper introduces the concept of upper approximation reduction in inconsistent target information systems based on dominance relations. Moreover the judgment theorem and discernable matrix are obtained, from which it can effectively provide the approach to this reduction in inconsistent systems based on dominance relations. An example strates the validity of this method. The proposed framework constructs efficient theoretical basis for knowledge discovery in information system based dominance relations.

【Key words】 rough set; information system; upper approximation reduction; discernable matrix

1 概述

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、模式识别与智能信息处理等领域[2-5]，并越来越引起了国际学术界的关注。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，通过等价关系把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[6-11]。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。因此，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[12-14]。

2 基于优势关系的信息系统

目标信息系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。

定义 1[7] 称一个五元组 $I=(U,A,F,D,G)$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，$F=\{f_k:U\to V_k,k\le p\}$，$V_k$ 是 a_k 的有限值域；

G 是 U 与 D 的关系集，$G=\{g_k:U\to V_k',k'\le q\}$，$V_k'$ 是 $d_{k'}$ 的有限值域。

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。

基金项目：重庆市教委科技基金资助项目(KJ090612)；重庆九龙区科技基金资助项目(2008Q98)
作者简介：徐伟华(1979－)，男，副教授、博士，主研方向：粗糙集理论与应用，人工智能的数学基础；张晓燕，讲师、硕士；张文修，教授、博士生导师
收稿日期：2009-03-16 **E-mail：**datongxuweihua@126.com

定义 2[7] 设 $I=(U,A,F,D,G)$为目标信息系统，对于 $B\subseteq A$，令

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

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

其中，$R_B^{\leq}$, $R_D^{\leq}$称为目标信息系统的优势关系，此时该目标信息系统称为基于优势关系下的目标信息系统。

记：

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

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

易见，优势关系有下面的性质：

命题 1[7]

(1) $R_B^{\leq}$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：$R_A^{\leq}\subseteq R_{B_2}^{\leq}\subseteq R_{B_1}^{\leq}$。

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

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

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

$$\underline{R_B^{\leq}}(X)=\{x_i\in U:[x_i]_B^{\leq}\subseteq X\}$$

$$\overline{R_B^{\leq}}(X)=\{x_i\in U:[x_i]_B^{\leq}\cap X\neq\varnothing\}$$

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，详细请参考文献[7]。

为了叙述方便，下文在没有特别说明时的信息系统都是指基于优势关系下的信息系统。

定义 3[7] 设 $I=(U,A,F,D,G)$为基于优势关系的目标信息系统，若 $R_A^{\leq}\subseteq R_D^{\leq}$，则称该基于优势关系的目标信息系统是协调的；否则，若 $R_A^{\leq}\not\subseteq R_D^{\leq}$，称该系统是不协调的。

例 1[7] 表 1 给出了一个基于优势关系的目标信息系统。

表 1　基于优势关系的目标信息系统

U	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：

$$[x_1]_A^{\leq}=\{x_1,x_2,x_5,x_6\}$$

$$[x_2]_A^{\leq}=\{x_2,x_5,x_6\}$$

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

$$[x_4]_A^{\leq}=\{x_4,x_6\}$$

$$[x_5]_A^{\leq}=\{x_5\}$$

$$[x_6]_A^{\leq}=\{x_6\}$$

$$[x_1]_d^{\leq}=[x_5]_d^{\leq}=\{x_1,x_5\}$$

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

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

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

3　不协调目标信息系统的上近似约简

由于优势关系不再是等价关系，不能形成对象集上的划分，而是一个覆盖。因此，对于优势关系下的信息系统不能采取类似于 Pawlak 近似空间的信息系统方法定义上近似函数和最大上近似函数。下面给出优势关系下的信息系统的上近似函数和最大上近似函数的定义方式。

设(U, A, F, D, G)为目标信息系统，$R_B^{\leq}$, $R_D^{\leq}$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 $B\subseteq A$，$x\in U$，记

$$U/R_B^{\leq}=\{[x_i]_B^{\leq}: x_i\in U\}$$

$$U/R_D^{\leq}=\{D_1,D_2,\cdots,D_r\}$$

$$\overline{\eta_B}=(\overline{R_B^{\leq}}(D_1),\overline{R_B^{\leq}}(D_2),\cdots,\overline{R_B^{\leq}}(D_r))$$

其中，$[x]_B^{\leq}=\{y\in U:(x,y)\in R_B^{\leq}\}$，称 $\overline{\eta}_B$为论域 U 上的关于属性子集 B 的上近似函数。

定义 4 设 $I=(U,A,F,D,G)$为目标信息系统。若对 $B\subseteq A$ 有 $\overline{\eta}_B=\overline{\eta}_A$，则称 B 是该信息系统的上近似协调集，且 B 的任何真子集不是上近似协调集，则称 B 为该系统的上近似协调约简。

显然，由上定义可直接得到下面命题。

命题 2 设 $I=(U,A,F,D,G)$为目标信息系统，$B\subseteq A$，则 B 是上近似协调集当且仅当对任意的 $D_i\in U/R_D^{\leq}$ 都有 $\overline{R_B^{\leq}}(D_i)=\overline{R_A^{\leq}}(D_i)$。

例 2 考虑例 1 给出的不协调目标信息系统。

若在该信息系统中记：

$$D_1=[x_1]_d^{\leq}=[x_5]_d^{\leq},\ D_2=[x_2]_d^{\leq}=[x_4]_d^{\leq},\ D_3=[x_3]_d^{\leq}=[x_6]_d^{\leq}$$

则有：

$$\overline{R_A^{\leq}}(D_1)=\{x_1,x_2,x_3,x_5\}$$

$$\overline{R_A^{\leq}}(D_2)=\{x_1,x_2,x_3,x_4,x_5\}$$

$$\overline{R_A^{\leq}}(D_3)=\{x_1,x_2,x_3,x_4,x_5,x_6\}$$

若取 $B=\{a_2,a_3\}$ 时，容易验证对于 $\forall x\in U$ 有：$[x]_A^{\leq}=[x]_B^{\leq}$，因此有 $\overline{\eta}_B=\overline{\eta}_A$。故 $B=\{a_2,a_3\}$ 是个上近似协调集，而且可以计算 $\{a_2\}$和$\{a_3\}$ 均不是上近似协调集，因此，$B=\{a_2,a_3\}$ 是个上近似约简。

若取 $B'=\{a_1,a_3\}$ 时，有：

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

$$[x_2]_{B'}^{\leq}=\{x_2,x_5,x_6\}$$

$$[x_3]_{B'}^{\leq}=\{x_2,x_3,x_4,x_5,x_6\}$$

$$[x_4]_{B'}^{\leq}=\{x_4,x_6\}$$

$$[x_5]_{B'}^{\leq}=\{x_2,x_5,x_6\}$$

$$[x_6]_{B'}^{\leq}=\{x_6\}$$

于是得

$$\overline{R_{B'}^{\leq}}(D_1)=\{x_1,x_2,x_3,x_5\}$$

$$\overline{R_{B'}^{\leq}}(D_2)=\{x_1,x_2,x_3,x_4,x_5\}$$

$$\overline{R_{B'}^{\leq}}(D_3)=\{x_1,x_2,x_3,x_4,x_5,x_6\}$$

即有 $\overline{\eta}_B=\overline{\eta}_A$。故 $B'=\{a_1,a_3\}$ 也是个上近似协调集，而且可以计算 $\{a_1\}$ 不是上近似协调集。因此，$\{a_1,a_3\}$ 是另一个上近似约简。

进一步可以验证 $\{a_1,a_2\}$ 不是上近似协调集。因此该目标信息系统只有 2 个上近似约简 $\{a_1,a_3\}$ 和 $\{a_2,a_3\}$。

下面具体给出不协调目标信息系统的上近似约简的判定

定理。

定理 1 设(*U*,*A*,*F*,*D*,*G*)为目标信息系统，$B \subseteq A$，则 *B* 是上近似协调集当且仅当对任意的 $D_i \in U/R_D^{\leq}$，当 $x \notin \overline{R_A^{\leq}}(D_i), y \in \overline{R_A^{\leq}}(D_i)$ 时，存在 $b \in B$ 使得 $f_b(x) > f_b(y)$。

证明：

“ $\Rightarrow$ ”反证。

假设存在某个 $D_i \in U/R_D^{\leq}$，当 $x \notin \overline{R_A^{\leq}}(D_i), y \in \overline{R_A^{\leq}}(D_i)$ 时，对任意的 $b \in B$ 使得 $f_b(x) \leq f_b(y)$。故此时有 $y \in [x]_B^{\leq}$，而 *B* 又是上近似协调集，对任意的 $D_i \in U/R_D^{\leq}$ 都有 $\overline{R_B^{\leq}}(D_i) = \overline{R_A^{\leq}}(D_i)$。

因为 $y \in \overline{R_A^{\leq}}(D_i)$，所以 $y \in \overline{R_B^{\leq}}(D_i)$，即 $[y]_B^{\leq} \cap D_i \neq \varnothing$，又 $y \in [x]_B^{\leq}$，则 $[y]_B^{\leq} \subseteq [x]_B^{\leq}$，故有 $[x]_B^{\leq} \cap D_i \neq \varnothing$。

于是，$x \in \overline{R_B^{\leq}}(D_i)$，因此可知 $x \in \overline{R_A^{\leq}}(D_i)$。矛盾。

“ $\Leftarrow$ ”若 *B* 不是上近似协调集，则一定存在某个 $D_i \in U/R_D^{\leq}$，使得 $\overline{R_A^{\leq}}(D_i) \neq \overline{R_B^{\leq}}(D_i)$，即存在 $x_0 \notin \overline{R_A^{\leq}}(D_i)$，但 $x_0 \in \overline{R_B^{\leq}}(D_i)$。故有 $[x_0]_A^{\leq} \cap D_i = \varnothing$，但 $[x_0]_B^{\leq} \cap D_i \neq \varnothing$。

又 $[x_0]_A^{\leq} \subseteq [x_0]_B^{\leq}$，故存在某个 $y_0 \in [x_0]_B^{\leq}$，且 $y_0 \in D_i$，而 $[y_0]_B^{\leq}$ 至少包含 y_0 自己一个元素，所以可知 $[y_0]_A^{\leq} \cap D_i \neq \varnothing$，于是 $y_0 \in \overline{R_A^{\leq}}(D_i)$，即有 $x_0 \notin \overline{R_A^{\leq}}(D_i)$，$y_0 \in \overline{R_A^{\leq}}(D_i)$，于是存在 $b \in B$ 使得 $f_b(x_0) > f_b(y_0)$，显然这与 $y_0 \in [x_0]_B^{\leq}$ 矛盾。

因此 *B* 是上近似协调集。定理得证明。

4 上近似约简的辨识矩阵与约简方法

第 3 节中的定理给出了不协调目标信息系统的上近似协调集的等价刻画，这是判断属性子集是否协调的理论所在，由此可进一步得出上近似约简的方法。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

$$D_{\eta}^{*} = \{(x_i, x_j) : x_i \notin \overline{R_A^{\leq}}(D_i), x_j \in \overline{R_A^{\leq}}(D_i)\},\ D_i \in U/R_D^{\leq}$$

用 $f_{a_k}(x)$ 表示属性 a_k 关于对象 *x* 的取值。定义

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

称 $D_{\bar{\eta}}(x_i, x_j)$ 为 x_i 与 x_j 的上近似可辨识属性集。矩阵 $\boldsymbol{M}_{\bar{\eta}} = (D_{\bar{\eta}}(x_i, x_j),\ x_i, x_j \in U)$ 称为该目标信息系统的上近似辨识矩阵。

定理 2 设(*U*,*A*,*F*,*D*,*G*)为不协调目标信息系统，$B \subseteq A$，则：*B* 是上近似协调集当且仅当对 $\forall (x,y) \in D_{\eta}^{*}$，都有 $B \cap D_{\bar{\eta}}(x,y) \neq \varnothing$。

证明：

“ $\Rightarrow$ ”对 $\forall (x,y) \in D_{\eta}^{*}$，则存在某个 $D_i \in U/R_D^{\leq}$，使得有 $x \notin \overline{R_A^{\leq}}(D_i), y \in \overline{R_A^{\leq}}(D_i)$。所以由定理 1 知一定存在 $b \in B$，使得 $f_b(x) > f_b(y)$。于是 $b \in D_{\bar{\eta}}(x_i, x_j)$。

因此，若 *B* 是上近似协调集，则对 $\forall (x,y) \in D_{\eta}^{*}$ 有 $B \cap D_{\bar{\eta}}(x,y) \neq \varnothing$。

“ $\Leftarrow$ ”若对 $\forall (x,y) \in D_{\eta}^{*}$ 有 $B \cap D_{\bar{\eta}}(x,y) \neq \varnothing$，则存在一个 $a_k \in B$ 使得 $a_k \in D_{\bar{\eta}}(x,y)$，故有 $f_{a_k}(x) > f_{a_k}(y)$，而此时 $x \notin \overline{R_A^{\leq}}(D_i), y \in \overline{R_A^{\leq}}(D_i)$，所以由定理 1 知 *B* 是上近似协调集。

定理得证。

定义 6 设(*U*,*A*,*F*,*D*,*G*)为不协调目标信息系统，$\boldsymbol{M}_{\delta}$ 为其上近似辨识矩阵，若记

$$F_{\bar{\eta}} = \wedge\{\vee\{a_k : a_k \in D_{\bar{\eta}}(x_i, x_j)\},\ \ x_i, x_j \in U\} = \wedge\{\vee\{a_k : a_k \in D_{\bar{\eta}}(x_i, x_j)\},\ \ x_i, x_j \in D_{\eta}^{*}\}$$

称 $F_{\bar{\eta}}$ 为该信息系统的上近似辨识公式。

定理 3 设(*U*,*A*,*F*,*D*,*G*)为不协调目标信息系统。上近似辨识公式 $F_{\bar{\eta}}$ 的极小析取范式为 $F_{\bar{\eta}} = \bigvee_{k=1}^{p}(\wedge_{s=1}^{q_k} a_s)$，若记 $B_{\eta}^{k} = \{a_s, s = 1,2,\cdots,q_k\}$，则 $\{B_{\eta}^{k}, k = 1,2,\cdots,p\}$ 是所有上近似约简形式的集合。

证明：对任意的 $(x_i, x_j) \in D_{\eta}^{*}$，由极小析取范式的定义知 $B_{\eta}^{k} \cap D_{\bar{\eta}}(x_i, x_j) \neq \varnothing$，再由定理 2 知 B_{η}^{k} 是上近似协调集。同时，$F_{\bar{\eta}} = \bigvee_{k=1}^{p}(B_{\eta}^{k})$ 在 B_{η}^{k} 中去掉一个元素形成 $B_{\eta}^{k}{}'$，则必然存在某个 $(x_i, x_j) \in D_{\eta}^{*}$ 使得 $B_{\eta}^{k'} \cap D_{\bar{\eta}}(x_i, x_j) = \varnothing$，故 $B_{\eta}^{k}{}'$ 不是上近似协调集，从而 B_{η}^{k} 是上近似约简。而上近似辨识公式中包含了所有的 $D_{\bar{\eta}}(x_i, x_j)$，因此不存在其他上近似约简。

例 3 对于例 1 给出的不协调目标信息系统的上近似辨识矩阵如表 2 所示。

表 2 例 1 中不协调目标信息系统的上近似辨识矩阵 $\boldsymbol{M}_{\bar{\eta}}$

U	x_1	x_2	x_3	x_4	x_5	x_6
x_1	∅	∅	∅	∅	∅	∅
x_2	∅	∅	∅	∅	∅	∅
x_3	∅	∅	∅	∅	∅	∅
x_4	a_1,a_3	a_3	a_1,a_3	∅	a_3	∅
x_5	∅	∅	∅	∅	∅	∅
x_6	a_1,a_3	a_3	*A*	a_1,a_2	a_3	∅

故可得：

$$F_{\bar{\eta}} = (a_1 \vee a_2 \vee a_3) \wedge (a_1 \vee a_2) \wedge (a_1 \vee a_3) \wedge a_3 = (a_1 \wedge a_3) \vee (a_2 \wedge a_3)$$

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上近似约简。这与例 2 的结果是一致的。

5 结束语

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。本文通过引入基于优势关系下不协调目标信息系统的上近似约简概念，对这一类复杂的系统中的上近似属性约简做了深入的分析，得到了上近似的判定定理以及辨识矩阵，从而建立了不协调目标信息系统的上近似的具体方法。

参考文献

[1] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. Boston, USA: Kluwer Academic Publishers, 1991.

[2] Pawlak Z. Rough Sets[J]. Communication of the ACM, 1995, 38(1): 89-95.

[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] 苗夺谦, 王 珏. 基于粗糙集的多变量决策树构造方法[J]. 软件学报, 1997, 8(6): 425-431.

[5] 米据生, 吴伟志, 张文修. 不协调目标信息系统知识约简的比较研究[J]. 模糊系统与数学, 2003, 17(3): 54-60.

(下转第 197 页)