第 34 卷 第 11 期 计 算 机 工 程 2008 年 6 月
Vol.34 No.11 Computer Engineering June 2008
·博士论文· 文章编号：1000—3428(2008)11—0018—03 文献标识码：A 中图分类号：TP301.6

# 不完备联系度粗糙集模型的知识约简

黄 兵[1,2]，李华雄[2]，周献中[2]
(1. 南京审计学院计算机科学与技术系，南京 210029；2. 南京大学工程管理学院，南京 210093)

**摘 要：**知识约简是粗糙集理论的重要研究内容之一。在不完备信息系统中，对联系度粗糙集模型的研究比较深入，但极少涉及知识约简问题。该文在进一步改进联系度粗糙集模型的基础上，研究该模型的知识约简。针对改进模型，提出多种知识约简定义，给出了这些约简之间的关系。通过定义初等分辨矩阵和属性重要度，介绍一种分配约简算法。实例分析说明了算法的有效性。

**关键词：**粗糙集；联系度；知识约简；不完备信息系统

# Knowledge Reduction for Incomplete Connection-degree-based Rough Sets Model

HUANG Bing[1,2], LI Hua-xiong[2], ZHOU Xian-zhong[2]
(1. Department of Computer Science & Technology, Nanjing Audit University, Nanjing 210029;
2. School of Engineering & Management, Nanjing University, Nanjing 210093)

**【Abstract】** Knowledge reduction is one of the most important issues in rough sets theory. In incomplete information systems, researchers study connection-degree-based rough sets model. However, knowledge reduction for this model has never been researched. In this paper, the connection-degree-based rough sets model is improved and knowledge reduction for this mode is examined. Several knowledge reductions are proposed and the relations between them are showed. Using elementary discernibility matrix and the significance of attributes defined, a knowledge reduction algorithm for incomplete information systems is presented. Example analysis shows that this algorithm is valid.

**【Key words】** rough sets; connection degree; knowledge reduction; incomplete information systems

粗糙集理论自 1982 年提出以来，已成为处理不确定信息的一种有力工具。经典粗糙集理论建立在等价关系的基础之上，通过知识约简，为从信息系统中获取有用知识提供了一种全新的方法。文献[1]将集对分析[2]中的联系度引入不完备信息系统(IIS)，建立了基于联系度的粗糙集模型，文献[3]做了一定的改进。在此基础上，很多学者做了进一步研究，但大多仍停留在模型拓展阶段，对这些模型的知识约简研究很少。本文给出了多种知识约简概念，通过定义初等分辨矩阵和属性重要度，给出了知识约简的一般方法。

## 1 基于联系度的粗糙集模型

给定信息系统 $S=(U,A=C\cup\{d\},V,f)$。其中，$U$ 为全体对象构成的集合，称为论域；$C$ 是条件属性集合；$d$ 是决策属性；$f$ 是论域 $U$ 到属性值集合 $V\cup\{*\}$ 的一个映射，其中 $f(x,c)=*$ 表示对象 $x\in U$ 在属性 $c\in C$ 上取值为空。

**定义 1**[2] 给定 2 个集合$A$和$B$，并设这 2 个集合组成集对，它们共有$N$个属性。其中，集对$A$和$B$在$S$个属性上取值相同；在$P$个属性上取值不同；在$F$个属性上取值不明确，则称比值$S/N$为集对$A$和$B$的同一度；$F/N$为$A$和$B$的差异度；$P/N$为$A$和$B$的对立度。并用 $u(A,B)=S/N+F/N\,i+P/N\,j$表示$A$与$B$的关系。简记为

$$u(A,B)=a+b\,i+c\,j$$

其中，$0\le a,b,c\le 1, a+b+c=1$。称 $u$ 为 $A$ 与 $B$ 的联系度。

对 IIS $S=(U,A=C\cup\{d\},V,f)$，$B\subseteq C$，设

$|B|=n$，$\forall x,y\in U$

令 $S=\{b\in B\,|\,f(x,b)=f(y,b)\ne *\}$，

$F=\{b\in B\,|\,f(x,b)=*\vee f(y,b)=*\}$，

$P=\{b\in B\,|\,(f(x,b)\ne f(y,b))\wedge f(x,b)\ne *\wedge f(y,b)\ne *\}$

记 $u_B(x,y)=a+b\,i+c\,j$

其中，$a=|S|/|B|$；$b=|F|/|B|$；$c=|P|/|B|$；$|X|$表示集合 $X$ 的基数。

**定义 2** $P_B^{\alpha}(x)=\{y\in U\,|\,u(x,y)=a+bi,\ a+b=1, a\ge\alpha, 0\le\alpha\le 1\}\cup\{x\}$ 称为 $x$ 关于 $B$ 的 $\alpha$ 相容类。其中，$a,b$ 分别表示对象 $x,y$ 在属性集 $B\subseteq C$ 上同一度和差异度。若 $y\in P_B^{\alpha}(x)$，则称 $x$ 与 $y$ 关于 $B$ 满足 $\alpha$ 相容关系。

说明：

(1) $y\notin P_B^{\alpha}(x)$ 有 2 种可能：1) $y$ 与 $x$ 有不相等的属性值；2) $y$ 与 $x$ 没有不相等的属性值，但是相同属性值所占比例小于 $\alpha$。

(2) $\cup\{x\}$ 项保证每个对象属于自身的 $\alpha$ 相容类。

(3) $\alpha$ 相容关系满足自反性和对称性，但不满足传递性。

(4)当 $\alpha=0$ 时，联系度容差关系即为容差关系，当 $\alpha_2>0$

**基金项目：**国家自然科学基金资助项目(70571032)；中国博士后科学基金资助项目(20060390916)；江苏省博士后科研计划基金资助项目(0601019C)；南京审计学院科研基金资助项目(NSK2006/A03)

**作者简介：**黄 兵(1972－)，男，博士、在站博士后，主研方向：粗糙集理论与应用；李华雄，博士研究生；周献中，教授、博士生导师

**收稿日期：**2007-06-12 **E-mail：**huangbing@nau.edu.cn

时排除了限制容差关系中的第 1 种情形；限制容差关系的第 2 种情形相当于联系度 $0<\alpha_2\le 1/n$ ($|B|=n$)的基于联系度容差关系；而当 $\alpha_2>1/n$ 时，有 $P_B^{\alpha_2}(x)\subseteq L_B(x)$。

相应地，可由联系度相容关系定义上、下近似运算如下：

**定义 3** 对 IIS $S=(U,A=C\cup\{d\},V,f)$，$B\subseteq A$，$X\subseteq U$，$0\le\alpha\le 1$，$\overline{B}^{\alpha}(X)$，$\underline{B}^{\alpha}(X)$ 分别表示联系度相容关系中 $X$ 关于属性子集 $B\subseteq A$ 的上、下近似集，定义如下：

$$\underline{B}^{\alpha}(X)=\{x\in U\,|\,P_B^{\alpha}(x)\subseteq X\},\ \overline{B}^{\alpha}(X)=\{x\in U\,|\,P_B^{\alpha}(x)\cap X\neq\phi\}$$

## 2 联系度粗糙集模型的知识约简定义

下面针对联系度粗糙集模型，给出 $\alpha$ 分布约简、$\alpha$ 最大分布约简、$\alpha$ 分配约简、$\alpha$ 分配序约简、$\alpha$ 下近似约简和 $\alpha$ 上近似约简定义。

对 IIS $S=(U,A=C\cup\{d\},V,f)$，为便于讨论，假设决策属性值不含空值。按照决策值的不同，可将论域划分为不同等价类，表示为 $U/R_D=\{D_1,D_2,\cdots,D_m\}$，其中，$m$ 为不同决策值的个数。对 $B\subseteq C$，$0\le\alpha\le 1$，记

$$\mu_B^{\alpha}(x)=(D_1^B(x),D_2^B(x),\cdots,D_m^B(x))$$

其中，$D_i^B(x)=\dfrac{|D_i\cap P_B^{\alpha}(x)|}{|P_B^{\alpha}(x)|},1\le i\le m$。

$$\gamma_B^{\alpha}(x)=\{D_h:D_h^B(x)=\max_{i\le m}D_i^B(x)\},$$

$$\delta_B^{\alpha}(x)=\{D_i:D_i\cap P_B^{\alpha}(x)\neq\phi\},\ \rho_B^{\alpha}(x)=(D_{i_1}\ge D_{i_2}\ge\cdots\ge D_{i_k})$$

其中，$D_{i_1}^B(x)\ge D_{i_2}^B(x)\ge\cdots\ge D_{i_k}^B(x)$，若等号成立，则按原下标从小到大排列，且 $D_{i_l}^B(x)>0,\sum_{l=1}^{k}D_{i_l}^B(x)=1$。

**定义 4** 对 IIS $S=(U,A=C\cup D,V,f)$，$B\subseteq C$，$0\le\alpha\le 1$：

(1)若 $\forall x\in U,\mu_B^{\alpha}(x)=\mu_B^{\alpha}(x)$，则称 $B$ 是 $\alpha$ 分布集。若 $B$ 是 $\alpha$ 分布集而其任意真子集都不是 $\alpha$ 分布集，则称 $B$ 为 $\alpha$ 分布约简。

(2)若 $\forall x\in U,\gamma_B^{\alpha}(x)=\gamma_B^{\alpha}(x)$，则称 $B$ 是 $\alpha$ 最大分布集。若 $B$ 是 $\alpha$ 最大分布集而其任意真子集都不是 $\alpha$ 最大分布集，则称 $B$ 为 $\alpha$ 最大分布约简。

(3)若 $\forall x\in U,\delta_B^{\alpha}(x)=\delta_B^{\alpha}(x)$，则称 $B$ 是 $\alpha$ 分配集。若 $B$ 是 $\alpha$ 分配集而其任意真子集都不是 $\alpha$ 分配集，则称 $B$ 为 $\alpha$ 分配约简。

(4)若 $\forall x\in U,\rho_B^{\alpha}(x)=\rho_B^{\alpha}(x)$，则称 $B$ 是 $\alpha$ 分配序集。若 $B$ 是 $\alpha$ 分配序集而其任意真子集都不是 $\alpha$ 分配序集，则称 $B$ 为 $\alpha$ 分配序约简。

(5)若 $\forall D_i\ (1\le i\le m)\in U/R_D$，有 $\underline{B}^{\alpha}(D_i)=\underline{C}^{\alpha}(D_i)$，则称 $B$ 是 $\alpha$ 下近似协调集。若 $B$ 是 $\alpha$ 下近似一致集而其任意真子集都不是 $\alpha$ 下近似协调集，则称 $B$ 是 $\alpha$ 下近似约简。

(6)若 $\forall D_i\ (1\le i\le m)\in U/R_D$，有 $\overline{B}^{\alpha}(D_i)=\overline{C}^{\alpha}(D_i)$，则称 $B$ 是 $\alpha$ 上近似协调集。若 $B$ 是 $\alpha$ 上近似一致集而其任意真子集都不是 $\alpha$ 上近似协调集，则称 $B$ 是 $\alpha$ 上近似约简。

由定义 4 可知，$\alpha$ 分布约简保证每个 $\alpha$ 相容类对决策类的隶属程度不变；$\alpha$ 最大分布约简保持每个 $\alpha$ 相容类的最大分布决策类不变；$\alpha$ 分配约简保持每个对象的 $\alpha$ 相容类的所有可能决策类不变，即不产生新的不一致；$\alpha$ 分配序约简保持每个 $\alpha$ 相容类对决策类隶属度大小顺序不变；$\alpha$ 下近似约简保持每个决策等价类的 $\alpha$ 下近似不变；$\alpha$ 上近似约简保持每个决策等价类的 $\alpha$ 上近似不变。

**定义 5** 对信息系统 $S=(U,C\cup D,V,f)$，给定某约简定义 $\Phi$，若 $\Phi$ 满足：$c\in C$ 是必要的 $\Rightarrow$ $c$ 是核属性。则称 $\Phi$ 具有属性保留性。

在 Pawlak 经典粗糙集意义下的保正域(或精度)不变约简具有属性保留性，但 Ziarko 定义的变精度粗糙集模型(variable precision rough sets model)知识约简不具有属性保留性。下面举例说明基于联系度粗糙集模型的约简不具有属性保留性。

表 1 给出一个不完备决策表，取 $\alpha=0.5$。

**表 1　不完备决策表**

| U\A | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $d$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | * | 1 |
| $x_2$ | 1 | 1 | * | 1 | 1 |
| $x_3$ | 1 | 0 | 3 | * | 1 |
| $x_4$ | 1 | * | 1 | * | 2 |
| $x_5$ | 1 | * | 2 | * | 1 |

$P_C^{0.5}(x_1)=\{x_1,x_2,x_4\}$，$P_C^{0.5}(x_2)=\{x_1,x_2\}$，$P_C^{0.5}(x_3)=\{x_3\}$，

$P_C^{0.5}(x_4)=\{x_1,x_4\}$，$P_C^{0.5}(x_5)=\{x_5\}$

$U/R_D=\{D_1=\{x_1,x_2,x_3,x_5\},D_2=\{x_4\}\}$

$\underline{C}^{0.5}(D_1)=\{x_2,x_3,x_5\}$，$\underline{C}^{0.5}(D_2)=\phi$

$P_{C\setminus\{c_1\}}^{0.5}(x_1)=\{x_1\}$，$P_{C\setminus\{c_1\}}^{0.5}(x_2)=\{x_2\}$，$P_{C\setminus\{c_1\}}^{0.5}(x_3)=\{x_3\}$，

$P_{C\setminus\{c_1\}}^{0.5}(x_4)=\{x_4\}$，$P_{C\setminus\{c_1\}}^{0.5}(x_5)=\{x_5\}$，而

$\underline{C\setminus\{c_1\}}^{0.5}(D_1)=D_1\neq\underline{C}^{0.5}(D_1)$，$\underline{C\setminus\{c_1\}}^{0.5}(D_2)=D_2\neq\underline{C}^{0.5}(D_2)$

因此，$c_1$ 是 0.5 下近似(分配)约简意义下的必要属性，但容易验证，$\{c_3\}$ 是一个 0.5 下近似(分配)约简，故 $c_1$ 不是核属性。

## 3 联系度粗糙集模型上的知识约简算法[4]

粗糙集理论中知识约简算法多种多样，如分辨矩阵与逻辑运算法、信息熵法、基于属性重要度方法等。但这些算法都是针对具有属性保留性的约简的。如在例 1 中，若利用分辨矩阵法则无法描述对象间的全部分辨关系；而若利用启发式算法则会得到 $c_1$ 是核属性，而 $c_1$ 并非是核属性。因此，属性保留性是设计知识约简算法的重要参考依据。

下面以 $\alpha$ 分配约简为例来分析联系度粗糙集模型上的知识约简。

Pawlak 意义下的知识约简建立在等价关系的基础之上，而基于邻域系统(将对象的相似类、相容类、$\alpha$ 相容类当作对象的邻域)的知识约简建立在邻域关系之上。当减少属性时，对象的邻域一般是单调增大的，但基于联系度的粗糙集模型上的 $\alpha$ 相容类却并非如此(见例 1)，这也是联系度粗糙集模型上的知识约简不满足属性保留性的根本原因。

在求联系度粗糙集模型上的分配 reduct 时，有些属性或属性组合是不能去除的(如例 1 中的属性 $c_3$)，而有些属性看似不能去除，但它却不含于任何约简中(如例 1 中的属性 $\alpha$)。因此，在设计求联系度粗糙集模型的分配 reduct 算法时，可先求出不能去除的属性(属性组合)，再通过逐步增加属性的方式寻找 reduct。下面给出相关定义。

**定义 6** 对 IIS $S=(U,C\cup D,V,f)$，$U=\{x_1,x_2,\cdots,x_n\}$，$0\le\alpha\le 1$，$\alpha$ 分配约简的初等分辨矩阵定义为

$$\boldsymbol{M}_{\alpha}=(m_{ij}^{\alpha})_{n\times n}=\begin{cases}\{c\in C:f(x_i,c)\neq f(x_j,c)\} & f(x_j,d)\notin\partial_C(x_i)\\ \phi & \text{else}\end{cases}$$

其中，$\partial_C(x_i)=\{f(x_j,d):x_j\in P_C^{\alpha}(x_i)\}$ 称为 $x_i$ 关于 $C$ 的 $\alpha$ 广义

决策函数。

$\alpha$ 分配约简初等分辨矩阵并不能反映信息系统联系度粗糙集模型分配约简的全部分辨关系，但它包含了各对象间原有区分关系下不相等属性值的全部信息，这保证了每个对象的 $\alpha$ 邻域排除了那些决策值不属于该对象的 $\alpha$ 广义决策函数，且与该对象有不同条件属性值的对象(即定义 2 说明中的第 1 种情形)。

由 $\alpha$ 分配约简定义，$B \subseteq C$ 是 $\alpha$ 分配 reduct 须满足 2 个条件：

(1) $\forall x \in U, y \notin P_C^{\alpha}(x) \wedge f(y,d) \notin \partial_C(x) \Rightarrow y \notin P_B^{\alpha}(x)$。

(2) $B$ 中无冗余属性。

由 $\alpha$ 分配约简初等分辨矩阵经逻辑运算求得 $\alpha$ 分配约简初等协调集 $B$ 满足：

$\forall x, y \in U, f(y,d) \notin \partial_C(x)$,若$\exists c \in C$,使得

$f(x,c) \neq f(y,c)$,则$y \notin P_B^{\alpha}(x)$

虽然 $\alpha$ 分配约简初等协调集的最小性由逻辑运算可以保证，但它并不能保证

$\forall x, y \in U, f(y,d) \notin \partial_C(x), \forall c \in C$, 当

$((f(x,c) = f(y,c)) \vee (f(x,c) = *) \vee (f(y,c) = *))$

有 $y \notin P_B^{\alpha}(x)$ 成立。

**定义 7** 对 IIS $S = (U, C \cup D, V, f)$，$U = \{x_1, x_2, \cdots, x_n\}$，$0 \le \alpha \le 1$, $B \subseteq C$。属性 $c \in C/B$ 的重要度 $Sig_B(c)$ 定义为

$$Sig_B(c) = \sum_{i=1}^{n} \left( \left| \partial_B(x_i) \oplus \partial_C(x_i) \right| - \left| \partial_{B \cup \{c\}}(x_i) \oplus \partial_C(x_i) \right| \right)$$

其中，$X \oplus Y = X \cup Y - X \cap Y$。

属性重要度反映了增加属性前后广义决策函数与原信息系统广义决策函数相异绝对量的变化大小，$Sig_B(c)$ 越大，表明 $c \in C/B$ 对 $B$ 越重要。因此，可利用属性重要度作为启发信息寻求 $\alpha$ 分配约简。下面结合 $\alpha$ 分配约简初等分辨矩阵和属性重要度给出求 $\alpha$ 分配约简的知识约简算法。

**算法 1** 基于初等分辨矩阵和属性重要度的 $\alpha$ 分配约简算法

输入 一个 IIS $S = (U, C \cup D, V, f)$，$U = \{x_1, x_2, \cdots, x_n\}$，$0 \le \alpha \le 1$

输出 一个 $\alpha$ 分配 reduct

Step1 计算 $P_C^{\alpha}(x_i)$ 及 $\partial_C(x_i), 1 \le i \le n$。

Step2 构造 $\alpha$ 分配约简的初等分辨矩阵。

Step3 通过对初等分辨矩阵中的非空项利用吸收律化简，然后将每一非空项表示为析取式，再将每个析取式表示为合取式，将合取式转化为析取式，则析取式中的每个合取式中的所有属性构成的集合即为一个 $\alpha$ 分配约简初等协调集。

Step4 选择一个 $\alpha$ 分配约简初等协调集 $B_0$，对每个 $c \in B_0$，检查 $\partial_{B_0}(x_i) = \partial_{B_0 \setminus \{c\}}(x_i), 1 \le i \le n$，若均成立，则 $B_0 = B_0 \setminus \{c\}$。

Step5 令 $B = B_0$：

(1)验证 $\partial_C(x_i) = \partial_B(x_i), 1 \le i \le n$，若均成立，则转 Step6；

(2)对每个 $c \in C \setminus B$，计算 $Sig_B(c)$；

(3)选择 $c \in \{c \in C \setminus B : Sig_B(c) = \max_{c \in C \setminus B} Sig_B(c)\}$，作 $B = B \cup \{c\}$，转(1)。

Step6 对每个属性 $c \in B \setminus B_1$，按加入 $B$ 的顺序的逆序，检查 $\partial_B(x_i) = \partial_{B \setminus \{c\}}(x_i), 1 \le i \le n$，若均成立，则 $B = B \setminus \{c\}$。

Step7 输出 $B$。

分析可得算法 1 的时间复杂度为 $O(|C|^3 |U|^2)$。

利用算法 1 可求得例 1 为 0.5 分配 reduct 为 $\{c_3\}$。

## 4 实例分析

表 2 给出了实例分析算法的有效性。

**表 2 实例分析算法的有效性**

| *Car* | *Price* | *Mileage* | *Size* | *Max-Speed* | *d* |
|---|---|---|---|---|---|
| 1 | High | High | Full | Low | Good |
| 2 | Low | * | Full | Low | Good |
| 3 | * | * | Compact | High | Poor |
| 4 | High | * | Full | High | Good |
| 5 | * | * | Full | High | Excel |
| 6 | Low | High | Full | * | Good |

其中，对象集合由 1~6 组成，条件属性集 $C = \{Price, Mileage, Size, Max-Speed\}$，$D = \{d\}$,取 $\alpha = 0.5$。则 $P_C^{\alpha}(1) = \{1\}$，$P_C^{\alpha}(2) = P_C^{\alpha}(6) = \{2,6\}$，$P_C^{\alpha}(3) = \{3\}$，$P_C^{\alpha}(4) = P_C^{\alpha}(5) = \{4,5\}$，初等分辨矩阵为

$$\boldsymbol{M}_{0.5} = \begin{bmatrix} \phi & \phi & SM & \phi & M & \phi \\ \phi & \phi & SM & \phi & S & \phi \\ SM & SM & \phi & S & S & S \\ \phi & \phi & S & \phi & \phi & \phi \\ M & S & S & \phi & \phi & \phi \\ \phi & \phi & S & \phi & \phi & \phi \end{bmatrix}$$

故得 $B_0 = \{Size, Max-Speed\}$，经验证 $\{Size, Max-Speed\}$ 即为 0.5 分配 reduct。

## 5 结束语

本文针对不同信息系统，建立相应的粗糙集模型，定义适当的知识约简，设计知识约简算法是粗糙集理论研究的一般路线和方法。本文在前文的研究基础上，定义了基于联系度的不完备信息系统的多种知识约简，并给出了基于初等分辨矩阵和属性重要度的分配约简算法。所设计的算法适用于不具有属性保留性的知识约简。这为从不完备信息系统中获取知识提供了一种新的手段和方法。

## 参考文献


[1] 黄 兵, 周献中. 基于集对分析的不完备信息系统粗糙集模型[J]. 计算机科学, 2002, 29(9): 1-3.

[2] 赵克勤. 集对分析及其初步应用[M]. 浙江: 浙江科学出版社, 2000.

[3] 黄 兵, 周献中. 不完备信息系统中基于联系度的粗集模型拓展[J]. 系统工程理论与实践, 2004, 24(1): 88-92.

[4] 周献中, 黄 兵. 基于粗集的不完备信息系统属性约简[J]. 南京理工大学学报, 2003, 27(5): 630-635.