# 基于程度粗糙集的知识约简方法

成蓉华, 陈世联
(昆明理工大学 理学院, 云南 昆明 650093)

摘要: 给出了基于程度粗糙集模型上知识的近似约简及近似相对约简, 引入了目标信息系统上的 $k$ 上(下)近似约简及 $k$ 上(下)分布约简的概念, 并讨论了它们之间的关系, 得到了几个重要结论, 可为知识发现或数据挖掘技术奠定一些基础.

关键词: 知识约简; 信息系统; 粗集理论
中图分类号: TP18 文献标识码: A 文章编号: 1007- 855X(2006)01- 0119- 03

## Knowledge Reduction Based on Graded Rough Set

CHENG Rong-hua, CHEN Shi-lian
(Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China)

**Abstract**: The main objective is to introduce some new concepts of knowledge reduction based on the graded rough set theory, such as upper (lower) approximation reduction and upper (lower) distribution reduction. The relationships among alternative reduction in information systems are discussed. These results are meaningful both in theory and practice.

**Key words**: knowledge reduction; information system; rough set theory

## 0 引 言

知识约简是知识发现的重要课题, 也是粗糙集理论的核心问题之一. 目前, 信息系统的知识约简大多是在 Pawlak粗糙集模型下进行的, Pawlak粗糙集模型的一个局限是它处理的分类必须是完全正确的或是肯定的, 因而它的分类是精确的, 亦即只考虑完全"包含"与"不包含", 而没有某种程度上的"包含"与"属于". Pawlak粗糙集模型的另一个局限性是它所处理对象是已知的, 且从模型中得到的结论仅适合于这些对象. Pawlak粗糙集模型的这些局限性限制了它的应用. 为了弥补这些局限, 许多学者从不同角度推广了这一模型, 如程度粗糙集模型、变精度粗糙集模型、模糊粗糙集模型等等.

本文给出程度粗糙集模型中四种知识约简的概念, 即 $k$ 上(下)近似约简, $k$ 上(下)分布约简. $k$ 上(下)近似约简是保持决策类的 $k$ 上(下)近似中的对象总数不变的属性集, 与原系统产生的命题规则可能不同; $k$ 上(下)分布约简则是保持每个决策类的 $k$ 上(下)近似不变的属性集, 与原系统产生的命题规则相容. 本文得到的结论可为信息系统的知识发现或数据挖掘奠定一些基础.

## 1 程度粗糙集模型

程度粗糙集模型是 Pawlak粗糙集模型的推广, 先给出有关概念和术语.

定义 1[2] 设 $A=(U, R)$为广义近似空间, $k$为非负整数, 定义 $X$ 关于近似空间 $A$ 依程度 $k$ 的下近似和上近似分别为

$\underline{apr_k}X=\{x\in U: |R_s(x)|-|R_s(x)\cap X|\leqslant k\}$, $\overline{apr_k}X=\{x\in U: |R_s(x)\cap X|>k\}$, 在目标信息系统

收稿日期: 2005- 01- 20.
第一作者简介: 成蓉华(1980. 11~), 女, 在读硕士研究生. 主要研究方向: 粗集理论与决策分析.
E-mail: kmchrh@126.com

中, 上述定义为:

定义 **2** 设 $(U, A, F, D, G)$为目标信息系统, $B\subseteq A$, $R_B$ 和 $R_D$ 分别为 $U$ 上由 $B$ 与 $D$ 导出的等价关系, 它们分别产生的 $U$ 上的分划为 $U/R_B$与 $U/R_D$, 记

$$U/R_B=\{[x]_B: x\in U\}=\{B_1, B_2, \cdots, B_m\},\quad U/R_D=\{[x]_D: x\in U\}=\{D_1, D_2, \cdots, D_r\}.$$

其中

$[x]_B=\{y\in U: (x, y)\in R_B\}$, $[x]_D=\{y\in U: (x, y)\in R_D\}$ 分别是 $x$ 关于 $B$ 与 $D$ 的等价类.

$\forall x\in U$, 记 $\underline{B}(X)=\{x\in U: [x]_B\subseteq X\}=\bigcup\{[x]_B: [x]_B\subseteq X\}$,

$\overline{B}(X)=\{x\in U: [x]_B\cap X\neq\varnothing\}=\bigcup\{[x]_B: [x]_B\cap X\neq\varnothing\}$.

则 $\underline{B}(X)$ 与 $\overline{B}(X)$分别称为 $X$ 关于 $B$ 的 Pawlak意义下的下近似集和上近似集, $X$ 的下近似是根据知识 $B$ 肯定属于 $X$ 的对象全体, 而 $X$ 的上近似是根据知识 $B$ 可能属于 $X$ 的对象全体.

显然, $\underline{B}(X)$与 $\overline{B}(X)$满足下列性质:

$$\underline{B}(X)\subseteq X\subseteq\overline{B}(X),\ \underline{B}(X)\subseteq\underline{B\cup\{a\}}(X),\ \overline{B}(X)\supseteq\overline{B\cup\{a\}}(X).$$

这说明增加属性会减少对象是否属于 $X$ 的不确定程度. 称 $(U, R_A, \underline{A}, \overline{A})$为 Pawlak粗糙集模型.

定义 **3** 设 $(U, A, F)$为信息系统, $B\subseteq A$, $X\subseteq U$, 对 $k\in[0, \partial/2)$, 其中 $\partial=\max\limits_{1\leqslant i\leqslant m}\{|B_i|\}$, $|B_i|$表示 $B_i$的基数, 记

$$\underline{B}_k(X)=\{x\in U: |[x]_B|-|[x]_B\cap X|\leqslant k\},\quad \overline{B}_k(X)=\{x\in U: |[x]_B\cap X|>k\}$$

则$\underline{B}_k(X)$与$\overline{B}_k(X)$分别称为 $X$ 关于程度 $k$的下近似与上近似; 基于程度 $k$的下近似与上近似的粗糙集模型称为程度粗糙集模型.

当 $k=0$时, $\underline{B}_k(X)=\underline{B}(X)$, 且$\overline{B}_k(X)=\overline{B}(X)$. 因此, 程度粗糙集模型是 Pawlak粗糙集模型的推广.

设 $(U, A, F, D, G)$为目标信息系统, $U/R_D=\{[x]_D: x\in U\}=\{D_1, D_2, \cdots, D_r\}$, 容易验证$\underline{B}_k$与$\overline{B}_k$具有以下性质:

(1)$\underline{B}_k(X)=\sim\overline{B}_k(\sim X)$, $\overline{B}_k(X)=\sim\underline{B}_k(\sim X)$;

(2)$\underline{B}_k(D_i)\cap\underline{B}_k(D_j)=\varnothing\quad (i\neq j)$;

(3)$\bigcup\limits_{j=1}^{r}\underline{B}_k(D_j)\subseteq\bigcup\limits_{j=1}^{r}\overline{B}_k(D_j)\subseteq U$, 等号未必成立;

(4)$\overline{B}_k(D_i)\cap\overline{B}_k(D_j)=\varnothing\quad (i\neq j)$, 一般不成立.

## **2** 程度粗糙集模型上的知识约简

基于程度粗糙集理论, 本文讨论其 $k$近似约简及 $k$分布约简.

定义 **4** 设 $(U, A, F, D, G)$为目标信息系统, $B\subseteq A$, 记

$$\mu_B(x)=\{|[x]_B\cap D_1|, |[x]_B\cap D_2|, \cdots, |[x]_B\cap D_r|\}$$

若 $\forall x\in U$, 有 $\mu_B(x)=\mu_A(x)$, 则称 $B$ 是分布协调集; 若 $B$ 是分布协调集, 但 $B$ 的任何真子集不是分布协调集, 则称 $B$ 为分布约简.

定义 **5** 设 $(U, A, F, D, G)$为目标信息系统, $B\subseteq A$, 记

$$\sigma_B^k=\frac{|pos(B, D, K)|}{|U|},\quad \lambda_B^k=1-\frac{|neg(B, D, K)|}{|U|},$$

$L_B^k = (\underline{B}_k(D_1), \cdots, \underline{B}_k(D_r))$, $H_B^k = (\overline{B}_k(D_1), \cdots, \overline{B}_k(D_r))$.

(1)若 $\sigma_B^k = \sigma_A^k$, 则称 $B$ 是 $k$ 下近似协调集; 若 $B$ 是 $k$ 下近似协调集, 但 $B$ 的任何真子集都不是 $k$ 下近似协调集, 则称 $B$ 是 $k$ 下近似约简.

(2)若 $\lambda_B^k = \lambda_A^k$, 则称 $B$ 是 $k$ 上近似协调集; 若 $B$ 是 $k$ 上近似协调集, 但 $B$ 的任何真子集都不是 $k$ 上近似协调集, 则称 $B$ 是 $k$ 上近似约简.

(3)若 $L_B^k = L_A^k$, 则称 $B$ 是 $k$ 下分布协调集; 若 $B$ 是 $k$ 下分布协调集, 但 $B$ 的任何真子集都不是 $k$ 下分布协调集, 则称 $B$ 是 $k$ 下分布约简.

(4)若 $H_B^k = H_A^k$, 则称 $B$ 是 $k$ 上分布协调集; 若 $B$ 是 $k$ 上分布协调集, 但 $B$ 的任何真子集都不是 $k$ 上分布协调集, 则称 $B$ 是 $k$ 上分布约简.

$k$ 上(下)分布协调集是保持每个决策类的 $k$ 上(下)近似不变的属性集, 它与 $A$ 产生相容的命题规则, 即在原系统和约简系统中, 由同一对象所产生的命题规则的决策部分相同; 而 $k$ 上(下)近似协调集则保持决策类的 $k$ 上(下)近似中的对象总数不变, 由它产生的命题规则与由 $A$ 产生的命题规则可能冲突, 但支持这些规则的对象个数不变.

**定理 1**　设 $(U, A, F, D, G)$ 是目标信息系统, 则 $k$ 下分布协调集必为 $k$ 下近似协调集; $k$ 上分布协调集必为 $k$ 上近似协调集.

证明　由定义即可得.

**例 1**　$k$ 下近似协调集未必是 $k$ 下分布协调集, 给出目标信息系统(表 1).

记　$A = \{a_1, a_2, a_3, a_4\}$, $B = \{a_2, a_3\}$, $D = \{d\}$,

$D_1 = \{x_1, x_5, x_6\}$, $D_2 = \{x_2, x_3, x_4\}$, 则

$\underline{A}_1(D_1) = \{x_1, x_2, x_3, x_4, x_5, x_6\}$

$\underline{A}_1(D_2) = \{x_1, x_2, x_4\}$

$\underline{B}_1(D_1) = \{x_1, x_3, x_5, x_6\}$

$\underline{B}_1(D_2) = \{x_1, x_2, x_4\}$

$$\sigma_A^1 = \frac{|pos(A, D, 1)|}{|U|} = 1, \quad \sigma_B^1 = \frac{|pos(B, D, 1)|}{|U|} = 1$$

表 1　目标信息系统

Tab. 1　Target information system

| $U$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $d$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 0 | 0 | 0 | 1 |
| $x_2$ | 0 | 1 | 1 | 1 | 2 |
| $x_3$ | 0 | 1 | 0 | 0 | 2 |
| $x_4$ | 0 | 1 | 1 | 0 | 2 |
| $x_5$ | 0 | 1 | 0 | 0 | 1 |
| $x_6$ | 0 | 1 | 0 | 0 | 1 |

而 $L_A^1 \neq L_B^1$, 因此, $B = \{a_2, a_3\}$ 是 1下近似协调集, 但不是 1下分布协调集.

同样, $k$ 上近似协调集也未必是 $k$ 上分布协调集.

**定理 2**　设 $(U, A, F, D, G)$ 为目标信息系统, 则分布协调集必为 $k$ 上、下分布协调集.

证明　设 $B$ 是分布协调集, 则对于任意 $x \in U$, 有 $\mu_B(x) = \mu_A(x)$. 即对于任意 $j \leqslant r$, 有 $|[x]_B \cap D_j| = |[x]_A \cap D_j|$. 因此, $y \in \underline{B}_k(D_j)$ 当且仅当 $y \in \underline{A}_k(D_j)$. 于是 $\underline{B}_k(D_j) = \underline{A}_k(D_j)$ $(\forall j \leqslant r)$, 即 $B$ 是下分布协调集.

同理可证, $B$ 是上分布协调集.

**定理 3**　设 $(U, A, F, D, G)$ 为目标信息系统, 则 $B$ 是 0上近似协调集的充分必要条件是 $B$ 是 0上分布协调集.

证明　"⇒"设 $B$ 为 0上近似协调集, 则 $\frac{|neg(B, D, 1)|}{|U|} = \frac{|neg(A, D, 1)|}{|U|}$, 即

$$\left|U - \bigcup_{D_i \in U/D} \overline{B}_0(D_i)\right| = \left|U - \bigcup_{D_i \in U/D} \overline{A}_0(D_i)\right|, \text{从而} \left|\bigcup_{i=1}^{r} \overline{B}_0(D_i)\right| = \left|\bigcup_{i=1}^{r} \overline{A}_0(D_i)\right|,$$

又因为 $\forall j \leqslant r$, 有 $\overline{B}_0(D_j) = \overline{B}(D_j) \supseteq \overline{A}(D_j) = \overline{A}_0(D_j)$, 因此 $\overline{B}_0(D_j) = \overline{A}_0(D_j)$, 即 $B$ 是 0上分布协调集;

(下转第 124页)

(2) $\forall x, y \in U, y \in F(x)$, 则 $F(x) \subseteq F(y)$;

(3) $\overline{F}(X) \subseteq \underline{F}\overline{F}(X)$ $(\forall X \subseteq U)$;

(4) $\overline{F}\underline{F}(X) \subseteq \underline{F}(X)$ $(\forall X \subseteq U)$.

证明 (1)⇔(2) 显然.

(2)⇒(3) 设 $x \in \overline{F}(X)$, 由上近似的定义可得 $F(x) \cap X \neq \varnothing$, 对于 $\forall y \in F(x)$ 由(2)得 $F(x) \subseteq F(y)$, 故 $F(y) \cap X \neq \varnothing$, 即 $y \in \overline{F}(X)$, 由 $y \in F(x)$ 的任意性知 $F(x) \subseteq \overline{F}(X)$, 由下近似的定义可得 $x \subseteq \underline{F}\overline{F}(X)$, 所以 $\overline{F}(X) \subseteq \underline{F}\overline{F}(X)$.

(3)⇒(1) 设 $y \in F(x)$, 并且 $z \in F(x)$, 由 $z \in F(x)$ 知 $F(x) \cap \{z\} \neq \varnothing$, 故 $x \in \overline{F}\{z\}$, 由(3)成立得到 $x \in \underline{F}\overline{F}\{z\}$, 由下近似的定义知 $F(x) \subseteq \overline{F}\{z\}$, 结合 $y \in F(x)$ 得 $y \in \overline{F}\{z\}$, 再由上近似的定义有 $F(y) \cap \{z\} \neq \varnothing$, 即 $z \in F(y)$, 这样我们就从 $y \in F(x)$ 和 $z \in F(x)$ 得到了 $z \in F(y)$.

(3)⇒(4) 可由引理 1的性质(1)得到.

## 3 结束语

集值映射的粗糙集模型是经典 Pawlak 粗糙集模型的推广, 是研究随机集与证据理论的基础, 本文对集值映射近似算子的性质进行了深入研究, 得到了一些重要的结论, 这些结论可为进一步研究随机集、证据理论以及粗集代数奠定一些理论基础.

参考文献:

[1] PAWLAK Z Rough Sets [J]. International Journal of Computer and Information Science, 1982, 11(5): 341~356.

[2] 刘少辉, 盛少戬, 吴斌, 等. 计算机学报 [J]. 2003, 26(5): 524~529.

[3] 张文修, 梁怡, 吴伟志. 信息系统与知识发现 [M]. 北京: 科学出版社, 2003 134~140.

[4] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法 [M]. 北京: 科学出版社, 2001. 41~53.

---

(上接第 121页)

"⇐" 由定理 1即得.

定理 **4** 设 $(U, A, F, D, G)$ 为目标信息系统, 则

1) 0上分布协调集必为 0下分布协调集;

2) 0上近似协调集必为 0下近似协调集.

证明 由上述定理可得.

参考文献:

[1] PAWLAK Z Rough Sets: Theoretical Aspects of Reasoning about Pata[M]. Boston: Kluwer Academic Publishers, 1991.

[2] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法 [M]. 北京: 科学出版社, 2001.

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

[4] 米据生, 吴伟志, 张文修. 基于变精度粗糙集理论的知识约简方法 [J]. 系统工程理论与实践, 2004, (1): 76~81.