文章编号: 1001 - 2486(2003)05 - 0098 - 05

# 属性约简与近似集的关系研究*

李克文, 吴孟达
(国防科技大学理学院, 湖南 长沙 410073)

摘 要: 上近似和下近似是在粗集理论中最基本的概念, 而粗集理论的主要思想又是在保持分类能力不变的前提下通过对属性的约简导出问题的决策和分类规则。本文讨论了属性的约简和相对约简与上近似和下近似的关系, 并在此基础上提出了一种新的相对约简 —— 保近似约简。

关键词: 粗集; 相对约简; 保近似约简
中图分类号: TP18 文献标识码: A

# Study on the Relationship between the Reduction of Attributes and Approximation Sets

LI Ke-wen, WU Meng-da
(College of Science, National Univ. of Defense Technology, Changsha 410073, China)

**Abstract**: Upper approximation and lower approximation are the most elementary concepts in rough set theory, while the primary idea in rough set theory is to deduct the decision and rule of classification through reducing the attributes and keeping the ability of classification unchanged. This paper discusses the relationship of reduct and relative reduct and the upper approximation and lower approximation. On the basis of the relationship we propose a novel relative reduct-holding approximation reduct.

**Key words**: rough set; relative reduct; holding approximation reduct

粗集理论是一种新的处理不确定性知识的数学工具[1~3], 基于粗集理论的数据分析无需提供所需处理的数据集合之外的任何先验信息, 而是“让数据自己说话[2]”。上、下近似是粗集理论中最基本的概念, 属性约简则是粗集理论的核心[5], 关于属性约简目前已有很多成熟的方法[6]。本文通过对属性约简与近似集的关系的研究, 揭示已定义的相对约简实际上只是在保持所关心范畴( 论域的子集) 的下近似不变的前提下对属性约简, 而在粗集理论中却是同时利用上、下近似来对一个范畴进行描述, 从而现有的相对约简过程可能会导致有用信息的丢失。

## 1 约简与近似集的关系

定理 **1** 对于信息系统 $S = (U, A)$, 如果非空集合 $Q \subseteq P \subseteq A$ 独立, 则如下等价:

(1) $Q$ 是 $P$ 的一个约简;

(2) $\forall X \in U/P, \underline{Q}X = \underline{P}X$;

(3) $\forall X \in U/P, \overline{Q}X = \overline{P}X$;

(4) $\forall X \subseteq U, \underline{Q}X = \underline{P}X$;

(5) $\forall X \subseteq U, \overline{Q}X = \overline{P}X$。

证明: “1⇒4, 5”: 若 $Q$ 是 $P$ 的一个约简, 于是 $Q$ 和 $P$ 具有相同的基本范畴, 这样显然有 4, 5 成立; “2, 3⇒1”: 如果 $\forall X \in U/P, \underline{Q}X = \underline{P}X$ 或 $\overline{Q}X = \overline{P}X$ 之一成立, 由于 $P$ 基本范畴的 $P$ - 下近似或 $P$ - 上近似都是其本身, 也就是说任意的 $P$ 基本范畴的 $Q$ - 下近似或 $Q$ - 上近似也为其本身, 即 $P$ 基本范畴都是 $Q$ 基本范畴的并, 由 $Q$ 独立可知 $Q$ 是 $P$ 的一个约简。

* 收稿日期: 2003 - 03 - 14
基金项目: 国防科技大学基础研究项目
作者简介: 李克文(1977—), 男, 硕士生。

对属性的约简是在保持属性集分类能力的前提下删除不重要的属性。从上述定理可以看出, 保持对论域的分类能力和使得论域中任何子集(特别地, $P$ 基本范畴) 的上近似集或下近似集不改变是等价的。

定理 **2** 对于信息系统 $S = (U, A)$, 如果有 $A$ 的非空子集 $B$、$C$、$D$, 且 $B \subseteq C$ 是相对 $D$ 独立的, 则如下等价:

(1) $B$ 是 $C$ 的一个 $D$ 约简;

(2) $\forall X \in U/D, \underline{B}X = \underline{C}X$。

证明 "1⇒2": 由 $B \subseteq C$ 知 $U/C \subseteq U/B$, 于是 $\forall X \subseteq U/D, \underline{B}X \subseteq \underline{C}X$, 因 $B$ 是 $C$ 的一个 $D$ 约简, 故 $\mathrm{pos}_B(D) = \bigcup_{X \in U/D} \underline{B}X = \mathrm{pos}_C(D) = \bigcup_{X \in U/D} \underline{C}X$, 由 $U/D$ 是 $U$ 的划分易知: $\underline{B}X = \underline{C}X$。

"2⇒1": $\forall X \subseteq U/D, \underline{B}X = \underline{C}X$, 于是 $\mathrm{pos}_B(D) = \bigcup_{X \in U/D} \underline{B}X = \bigcup_{X \in U/D} \underline{C}X = \mathrm{pos}_C(D)$, 由 $B$ 是相对 $D$ 独立的得 $B$ 是 $C$ 的一个 $D$ 约简。

推论 **1** 对于信息系统 $S = (U, A)$, 如果有 $A$ 的非空子集 $B$、$C$、$D$、$D'$, 且 $U/D = U/D'$, 则如下等价:

(1) $B$ 是 $C$ 的一个 $D$ 约简;

(2) $B$ 是 $C$ 的一个 $D'$ 约简。

推论 **2** 对于信息系统 $S = (U, A)$, 如果非空集合 $B \subseteq C \subseteq A$, 则如下等价:

(1) $B$ 是 $C$ 的一个 $C$ 约简;

(2) $B$ 是 $C$ 的一个约简。

证明: 由"定理 2" 知 $B$ 是 $C$ 的一个 $C$ 约简等价 $\forall X \subseteq U/C, \underline{B}X = \underline{C}X$, 再由"定理 1" 得 $B$ 是 $C$ 的一个约简。

定理 **3** 对于信息系统 $S = (U, A)$, 如果非空集合 $B, C, D \subseteq A$ 且 $B$ 是 $C$ 的一个 $D$ 约简, 则:

$$BX \supseteq CX \ (\forall X \subseteq U)$$

由"定理 2" 知, $C$ 的 $D$ 约简是在 $D$ 基本范畴的下近似不变的前提下对来对 $C$ 进行约简, 这样对于相同 $D$ 基本范畴内的不同 $C$ 基本范畴, 以及不属于任意 $D$ 基本范畴的不同 $C$ 基本范畴之间的区分信息将变得不重要, 显然, 后者可能导致"定理 3" 结论中真包含的情形成立。

从上面的讨论可以看出, 对属性集进行约简是在每个范畴的上、下近似不改变的前提下删除不重要的属性, 这不会改变范畴的近似精度[25]。然而在相对约简中却只保持所关心范畴的下近似保持不变, 而上近似则可能变大, 从而使其近似精度减小。

## 2 保近似约简

定义 **1** 对于信息系统 $S = (U, A)$, 如果非空集合 $B, D \subseteq A$, 称 $a \in B$ 是 $B$ 中 $D$ 保近似不必要的, 如果:

(1) $\forall X \in U/D, \underline{(B - \{a\})}X = \underline{B}X$;

(2) $\forall X \in U/D, \overline{(B - \{a\})}X = \overline{B}X$。

否则称 $a$ 是 $B$ 中 $D$ 保近似必要的; 如果 $\forall a \in B$ 都是 $D$ 保近似必要的, 称 $B$ 是相对 $D$ 保近似独立的; $B$ 中所有 $D$ 保近似必要的属性构成的集合称为 $B$ 的 $D$ 保近似核, 记为 $\mathrm{hcore}_D(B)$。

定理 **4** 对于信息系统 $S = (U, A)$, 如果有 $A$ 的非空子集 $B$、$D$, 且 $B$ 是相对 $D$ 独立的, 则 $B$ 是相对 $D$ 保近似独立的。

事实上, 如果 $B$ 是相对 $D$ 独立的, 则由 $\forall a \in B, \mathrm{pos}_B(D) \neq \mathrm{pos}_{(B-\{a\})}(D)$, 显然可得"定义 1" 中的 (1) 不被满足, 即 $\forall a \in B$ 都是 $D$ 保近似必要的, 从而 $B$ 是相对 $D$ 保近似独立的。

定理 **5** 对于信息系统 $S = (U, A)$, 如果有 $A$ 的非空子集 $B$、$D$ 且 $\mathrm{pos}_B(D) = U$, 则如下等价:

(1) $B$ 是相对 $D$ 独立的;

(2) $B$ 是相对 $D$ 保近似独立的。

事实上,当 $\mathrm{pos}_B(D) = U$ 时 $\forall X \in U/D$ 为 $B$ 可定义,"定义1"中的(1),(2)变成一回事。

定义**2** 对于信息系统 $S = (U, A)$,如果有 $A$ 的非空子集 $B, C, D$ 且 $B \subseteq C$ 是相对 $D$ 保近似独立的,我们称 $B$ 是 $C$ 的一个 $D$ 保近似约简,若:

(1) $\forall X \in U/D, \underline{BX} = \underline{CX}$;

(2) $\forall X \in U/D, \overline{BX} = CX$。

保近似核和保近似约简有如下关系:

定理**6** $\mathrm{hcore}_D(C) = \bigcap \mathrm{hred}_D(C)$,其中 $\mathrm{hred}_D(C)$ 是所有 $C$ 相对 $D$ 的保近似约简构成的集合。

证明:( )$\mathrm{hcore}_D(C) \subseteq \bigcap \mathrm{hred}_D(C)$。设 $a \in \mathrm{hcore}_D(C)$,需要证明 $a \in \bigcap \mathrm{hred}_D(C)$。反设 $a \notin \bigcap \mathrm{hred}_D(C)$,则存在某个 $B \in \mathrm{hred}_D(C)$ 使得 $a \notin B$。从而有 $B \subseteq C - \{a\}$,于是:

$$\forall X \in U, \underline{BX} \subseteq \underline{(C - \{a\})X} \subseteq \underline{CX}, \overline{BX} \supseteq \overline{(C - \{a\})X} \supseteq CX$$

另一方面,因为 $B$ 是 $C$ 相对 $D$ 的一个保近似约简,即 $\forall X \in U/D, \underline{BX} = \underline{CX}, \overline{BX} = CX$,从而可以得到 $\forall X \in U/D, \underline{(C - \{a\})X} = \underline{CX}, \overline{(C - \{a\})X} = CX$,这说明 $a$ 是 $C$ 中 $D$ 保近似不必要的,即 $a \notin \mathrm{hcore}_D(C)$,矛盾。

( ) $\bigcap \mathrm{hred}_D(C) \subseteq \mathrm{hcore}_D(C)$。设 $a \in \bigcap \mathrm{hred}_D(C)$,需要证明 $a \in \mathrm{hcore}_D(C)$。反设 $a \notin \mathrm{hcore}_D(C)$,这说明 $a$ 是 $C$ 中 $D$ 保近似不必要的,即 $\forall X \in U/D, \underline{(C - \{a\})X} = \underline{CX}, \overline{(C - \{a\})X} = CX$。另一方面,因为保近似约简总是存在的,我们知道 $C - \{a\}$ 有一个相对 $D$ 的保近似约简 $B_0$。从 $B_0 \subseteq C - \{a\}$ 知,$a \notin B_0$。$C - \{a\}$ 有一个相对 $D$ 的保近似约简 $B_0$ 满足:

(1) $\forall X \in U/D, \underline{(C - \{a\})X} = \underline{B_0X}, \overline{(C - \{a\})X} = \overline{B_0X}$;

(2) 如果 $B \subset B_0$, $\forall X \in U/D, \underline{(C - \{a\})X} = \underline{BX}, \overline{(C - \{a\})X} = BX$ 不同时成立。这两个条件可以重写为

(1′) $\forall X \in U/D, \underline{CX} = \underline{B_0X}, CX = \overline{B_0X}$;

(2′) 如果 $B \subset B_0$, $\forall X \in U/D, \underline{CX} = \underline{BX}, CX = BX$ 不同时成立。

注意到 $B_0 \subseteq C - \{a\} \subseteq C$,我们知道 $B_0$ 是一个 $C$ 相对 $D$ 的保近似约简,即 $B_0 \in \mathrm{hred}_D(C)$,由 $a \notin B_0$ 可知,$a \notin \bigcap \mathrm{hred}_D(C)$,矛盾。

定理**7** 对于信息系统 $S = (U, A)$,如果有 $A$ 的非空子集 $B$、$C$、$D$,且 $B$ 为 $C$ 的一个约简,则存在 $B'$ 为 $C$ 的 $D$ 约简,$B''$ 为 $C$ 的 $D$ 保近似约简满足:

$$B \supseteq B'' \supseteq B'$$

由前面的讨论,$C$ 的约简保持论域所有范畴的上、下近似,而保近似约简只需保持 $D$ 基本范畴的上、下近似,$C$ 的 $D$ 约简仅需保持 $D$ 基本范畴的下近似,由此所需要的区分信息相应减少,从而可得定理的结果。

定理**8** 对于信息系统 $S = (U, A)$,如果有 $A$ 的非空子集 $B$、$C$、$D$,且 $B \subseteq C$, $\mathrm{pos}_C(D) = U$,则如下等价:

(1) $B$ 是 $C$ 的一个 $D$ 约简;

(2) $B$ 是 $C$ 的一个 $D$ 保近似约简。

证明 如果(1)成立,即 $\forall X \in U/D, \underline{BX} = \underline{CX}$,由 $\mathrm{pos}_C(D) = U$ 知 $\forall X \in U/D$ 都是 $B$ 和 $C$ 可定义的,即 $\underline{BX} = \underline{CX} = X = CX = BX$,从而 $B$ 是 $C$ 的一个 $D$ 保近似约简。如果(2)成立,由"定理4"易知(1)成立。

由此,协调决策系统 $S = (U, A = C \cup D)$ 条件属性 $C$ 对决策属性 $D$ 的约简和保近似约简是一致的。

$C$ 相对 $D$ 的保近似约简对同一 $D$ 基本范畴内的不同 $C$ 基本范畴之间的区分信息不予考虑,因为这并不改变所关心的 $D$ 基本范畴的上、下近似,这与相对约简是一致的。在对不属于任意 $D$ 基本范畴的不同 $C$ 基本范畴间的区分信息的处理上,在相对约简中同样认为它们是不重要的。然而,当与不同的由 $D$

基本范畴构成的集合中, 每个 $D$ 基本范畴交都非空的 $C$ 基本范畴( 这些 $C$ 基本范畴不属于任意 $D$ 基本范畴且它们的 $D$ 上近似不同) 间的区分信息被忽略而使它们合并成为一个基本范畴时, 这将增大部分 $D$ 基本范畴的上近似, 从而减小其近似精度, 这正是保近似约简中要避免的情况, 也是相对约简和保近似约简的区别所在。类似求相对约简的方法[5], 通过构造区分矩阵[4] 可以求关于属性集的保近似约简。

设信息系统 $S = (U, A)$ 的区分矩阵是一个 $n \times n$ 矩阵, 其任一元素为:

$$a(x, y) = \{a \in C \mid a(x) \neq a(y) \text{ 且 } w(x, y)\}$$

对于 $x, y \in U$, $w(x, y)$ 满足如下之一:

(1) $x \in \mathrm{pos}_C D$ 且 $y \notin \mathrm{pos}_C D$;

(2) $x \notin \mathrm{pos}_C D$ 且 $y \in \mathrm{pos}_C D$;

(3) $x, y \in \mathrm{pos}_C D$ 且 $(x, y) \notin \mathrm{ind}(D)$;

(4) $x, y \notin \mathrm{pos}_C D$ 且 $D([x]_C) \neq D([y]_C)$。

也可以按如下的方法来确定区分矩阵中任一元素:

$$a(x, y) = \begin{cases} \phi & \text{当选取 } v(x, y); \\ \{a \in C \mid a(x) \neq a(y)\} & \text{否则。} \end{cases}$$

对于 $x, y \in U$, $v(x, y)$ 满足如下之一:

(1) $x, y \in \mathrm{pos}_C D$ 且 $(x, y) \in \mathrm{ind}(D)$;

(2) $x, y \notin \mathrm{pos}_C D$ 且 $D([x]_C) = D([y]_C)$。

$S$ 的区分函数定义为

$$\Delta = \prod_{(x, y) \in U \times U} \sum a(x, y)$$

$S$ 区分函数的极小析取范式中的所有合取式是 $C$ 的所有的相对 $D$ 的保近似约简。易知, 如果 $B \subseteq C$ 是满足: $B \cap a(x, y) \neq \phi$( $\forall a(x, y) \neq \phi$) 的极小子集, 则 $B$ 是 $C$ 的一个 $D$ 保近似约简。$D$ 核( 保近似核) 是 $S$ 的区分矩阵中所有单个元素组成的集合。

通过以上的论述, 我们容易得到如下结果:

**定理 9** 对于信息系统 $S = (U, A)$, 如果有 $A$ 的非空子集 $B$、$C$、$D$, 且 $B$ 是 $C$ 的一个 $D$ 保近似约简, 则如下成立:

$$D([x]_B) = D([x]_C) \qquad \forall x \in U$$

也就是说, 对于决策系统的条件属性进行保近似约简不会使论域中每一个对象的导出的可能决策增多, 这在下面的实例中将得到具体的体现。

## 3 实例

设有决策系统( 表 1) $S = (U, A = C \cup D)$, 其中 $C = \{a, b, c\}$, $D = \{d\}$。求保近似约简的区分矩阵由表 2 给出, 矩阵带阴影的元素在一般的相对约简中为空。由表 2 对应的区分函数为

$$\Delta = b(a \vee b \vee c)(a \vee c)(a \vee b)(b \vee c) = b(a \vee c) = ab \vee bc.$$

表 1 决策系统

Tab.1 A decisien system

| | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|
| 1 | 2 | 0 | 0 | 0 |
| 2 | 2 | 2 | 0 | 1 |
| 3 | 2 | 2 | 0 | 2 |
| 4 | 1 | 2 | 2 | 0 |
| 5 | 1 | 2 | 2 | 1 |
| 6 | 0 | 0 | 0 | 0 |
| 7 | 2 | 0 | 1 | 0 |

表 2 区分矩阵

Tab.2 Discernibility matrix

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | $b$ | | | | | | |
| 3 | $b$ | | | | | | |
| 4 | $a, b, c$ | $a, c$ | $a, c$ | | | | |
| 5 | $a, b, c$ | $a, c$ | $a, c$ | | | | |
| 6 | | $a, b$ | $a, b$ | $a, b, c$ | $a, b, c$ | | |
| 7 | | $b, c$ | $b, c$ | $a, b, c$ | $a, b, c$ | | |

因此, $C$ 的有两个 $D$ 保近似约简 $\{a, b\}$ 和 $\{b, c\}$, $D$ 保近似核是 $\{b\}$。易得该决策系统中 $C$ 的 $D$ 约简

和相对核均为{ $b$}。表 3 为 $D$ 基本范围关于相对约简和保近似约简结果的上近似:

表 **3** $U/D$ 的上近似

Tab. 3 The upper approximation of $U/D$

| $U/d$ | {1, 4, 6, 7} | {3} | {2, 5} |
|---|---|---|---|
| { $b$} | {1, 2, 3, 4, 5, 6, 7} | {2, 3, 4, 5} | {2, 3, 4, 5} |
| { $a$, $b$} | {1, 4, 5, 6, 7} | {2, 3} | {2, 3, 4, 5} |
| { $b$, $c$} | {1, 4, 5, 6, 7} | {2, 3} | {2, 3, 4, 5} |

相对约简所得到的决策规则为:

(1) ( $b$ = 0) $\overset{\rightarrow}{|}$ ( $d$ = 0);

(2) ( $b$ = 2) $\overset{\rightarrow}{|}$ ( $d$ = 0 ∨ 1 ∨ 2)。\ 覆盖对象为论域中的{2, 3, 4, 5}

而保近似约简得到的规则为( 不妨取{ $a$, $b$} 为约简结果):

(3) ( $b$ = 0) $\overset{\rightarrow}{|}$ ( $d$ = 0);

(4) ( $a$ = 2) ∧ ( $b$ = 2) $\overset{\rightarrow}{|}$ ( $d$ = 1 ∨ 2); \ 覆盖对象为论域中的{2, 3}

(5) ( $a$ = 1) ∧ ( $b$ = 2) $\overset{\rightarrow}{|}$ ( $d$ = 0 ∨ 1)。\ 覆盖对象为论域中的{4, 5}

## 4 结束语

对于协调的决策系统条件属性集相对决策属性集的保近似约简和相对约简是等价的。但通常的决策系统并不一定是协调, 于是不确定规则变得十分重要, 然而相对约简却往往会改变不确定规则的确定性, 如例中“规则( 2)” 较之“规则( 4, 5)” 明显损失了信息, 更加不确定。在例子中, 条件属性集是独立的( 而相对决策属性集保近似不独立), 但利用全部的条件属性 $C$ = { $a$, $b$, $c$} 也只能够得到保近似约简后所得到的同样确定的决策规则。也就是说, 保近似约简既能约简掉多余的属性, 又能够最大限度地保留知识。由此, 借助保近似约简代替相对约简来处理问题在实际中是有意义的。

## 参 考 文 献:

[1] Pawlak Z. Rough Sets—Theoretical Aspects of Reasoning About Data[M]. Kluwer Academic Pub, 1991.

[2] Ivo D ntsch, G nther Gediga. Rough set data analysis[M]. In Encyclopedia of Computer Science and Technology, Mercel Dekker, 2000, 43: 28 - 301.

[3] Ivo D ntsch, G nther Gediga. Roughian-Rough Information Analysis[J]. International Journal of Intelligent Systems, 46: 121 - 147, 2001.

[4] Skowron A, Rauszer C. The discernibility matrices and functions in information systems[J]. Fundamenta Informaticae, 15(2): 331 - 362, 1991.

[5] 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.

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