文章编号: 1001-0920 (2014) 08-1354-09 DOI: 10.13195/j.kzyjc.2013.0839

代数约简的知识粒度表示及其高效算法

黄国顺[a], 曾凡智[b], 文 翰[a]

(佛山科学技术学院 a. 理学院，b. 电子信息工程学院，广东 佛山 528000)

摘 要: 首先提出了修正相对粒度计算公式, 给出其单调性证明以及等号成立的充要条件; 然后证明了保持修正相对粒度不变是保持正区域不变的充要条件, 并给出代数约简的知识粒度表示; 最后讨论了现有相对粒度与修正相对粒度之间的关系, 利用修正相对粒度的单调性给出计算属性重要性定义及其递归计算公式, 进而利用基排序思想计算等价类, 设计出一种计算决策表代数约简的高效算法. 实验结果表明该算法是可行且高效的.

关键词: 知识粒度；相对粒度；属性重要性；代数约简

中图分类号: TP18 **文献标志码:** A

Knowledge granularity representation and efficient algorithm of algebraic reduction

HUANG Guo-shun[a], *ZENG Fan-zhi*[b], *WEN Han*[a]

(a. Science School，b. Electronics and Information Engineering School，Foshan University，Foshan 528000，China. Correspondent：HUANG Guo-shun，E-mail：fshgs_72@163.com)

Abstract: Firstly, a modified relative knowledge granularity is proposed. Its monotonicity is proved, and the necessary and sufficient conditions for equality are given. It is demonstrated that remaining the modified relative knowledge granularity and positive region unchanged is a necessary and sufficient condition for each other. Then the main concepts of algebraic reduction are described by knowledge granularity. The relation between existing relative knowledge granularity and its improvement is discussed. By modified relative knowledge granularity, an attribute relative significance is defined, and its recursive computing formula is presented. Then a heuristic attribute reduction algorithm based on this significance is designed, whose equivalence is computed by radix sort. The experimental results show that the algorithm is feasible and efficient.

Key words: knowledge granularity；relative granularity；attribute significance；algebraic reduction

0 引 言

属性约简是粗糙集研究的核心问题之一. Pawlak[1]首先提出了基于正区域的属性约简, 这种约简是基于集合代数运算的, 所得约简被称为代数约简. 由于在代数表示下的各种概念不容易被理解, 国内外学者分别从不同的角度对粗糙集理论的概念与运算进行重新定义, 提出了各种不同的约简概念 (方法). 常见的约简主要有: Hu 等[2]提出的基于 Skowron 差别矩阵的属性约简, 简称为 HU 属性约简; 王国胤等[3-4]基于条件信息熵提出的属性约简, 简称为信息熵约简. 但上述约简在决策表不一致时与代数约简不等价[5-6]. 因此对它们进行改进以保证能求到原始决策表的代数约简是目前的热点问题之一[7-9].

随着研究的不断深入, 人们发现可以将知识粒度[10-11]应用到决策表属性约简中. 如徐久成等[12]、陈玉明等[13]提出的相对粒度属性约简; 冯琴荣等[14]提出的相对划分粒度属性约简; Qian 等[15]提出的基于组合粒度的属性约简等. 上述基于知识粒度的约简也面临着同样的问题, 即当决策表不一致时, 它们与代数约简也不等价. 因此如何建立与代数约简相适应的知识粒度表示及其高效算法仍是一个有待解决的难点问题. 曾凡智等[16]证明了相对粒度 (相对划分粒度) 属性约简仅与 HU 属性约简等价; 黄国顺等[17]从相对知识量出发给出了一种计算 HU 属性约简的高

收稿日期: 2013-06-24；**修回日期:** 2013-12-15.

基金项目: 广东省自然科学基金项目(10452800001004185).

作者简介: 黄国顺(1972−), 男, 副教授, 博士, 从事粗糙集、粒度计算等研究； 曾凡智(1965−), 男, 教授, 博士, 从事数据库理论、数据挖掘等研究.

效算法.

本文在分析代数约简语义的基础上, 提出一种考虑到不一致对象集影响的修正相对粒度计算公式. 从理论上证明了保持修正相对粒度不变与保持正区域不变相互等价, 从而得到代数约简的知识粒度表示. 证明了修正相对粒度的单调性并利用它设计出一种计算属性重要性的递归计算公式, 给出了一种基于知识粒度计算代数约简的高效算法. 与现有约简算法相比, 本文方法是完备的, 同时能得到一些较好的约简结果. 如果采用徐章艳等[18]提出的基排序方法, 则时间复杂度与文献 [18] 相当.

1　基本概念

本节的相关概念和定理引自文献 [1-2, 10, 12-13, 16-17], 直接陈述如下.

定义 1　决策表 $S = \langle U, V, f, C \bigcup D\rangle$ 是一四元组. 其中: U 为非空有限对象集, 称为论域; C 为有限条件属性集, D 为有限决策属性集, $C \bigcap D = \varnothing$; $V = \bigcup\limits_{a\in C} V_a$, V_a 为属性 a 的值域; $f : U \times (C \bigcup D) \to V$ 是信息函数. 对于 U 上的任意属性集 $B \subseteq C \bigcup D$, 定义不可分辨关系 $\mathrm{ind}(B) = \{(x, y) \in U^2 | \forall a \in B, f(x, a) = f(y, a)\}$, 关系 $\mathrm{ind}(B)$ 构成 U 的一个划分, 记作 $U/\mathrm{ind}(B)$, 简记为 U/B. U/B 中的任何元素 $[x]_B = \{y | \forall a \in B, f(x, a) = f(y, a)\}$ 称为等价类.

定义 2　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, 定义 HU 差别矩阵为 $M_H(S) = (m_H(i, j))_{n\times n}$, 其中 $m_H(i, j)$ 定义如下:

$$m_H(i,j) = \begin{cases} \{a \in C | f(x_i, a) \neq f(x_j, a)\}, \\ \quad (x_i, x_j) \notin \mathrm{ind}(D); \\ \varnothing, \text{ others.} \end{cases}$$

$$i, j = 1, 2, \cdots, n. \tag{1}$$

定义 3　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, $M_H(S) = (m_H(i, j))_{n\times n}$ 是 HU 差别矩阵, 对于任意的 $P \subseteq C$, 若满足: 1) 对于任意的 $\varnothing \neq m_H(i, j) \in M_H$, 有 $P \bigcap m_H(i, j) \neq \varnothing$; 2) 对于任意的 $b \in P$, $P - \{b\}$ 不满足 1). 则称 P 是 C 相对于 D 的基于 HU 差别矩阵的属性约简, 简称为 HU 属性约简.

为便于讨论, 将定义 3 中满足条件 1) 的 P 称为 C 的 HU 协调集.

定义 4　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, 对于 $\forall P \subseteq C \bigcup D, Y \subseteq U$, 记 $U/P = \{P_1, P_2, \cdots, P_m\}$, 分别称 $\underline{P}Y = \{x \in U : [x]_P \subseteq Y\}$, $\overline{P}Y = \{x \in U : [x]_P \bigcap Y \neq \varnothing\}$ 为 Y 关于 P 的下近似集和上近似集.

定义 5　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, 设 $P \subseteq C$, $U/D = \{D_1, D_2, \cdots, D_r\}$, $U/P = \{P_1, P_2, \cdots, P_m\}$, 称 $\mathrm{POS}_P(D) = \bigcup\limits_{D_i\in U/D} \underline{P}D_i$ 为 P 关于 D 的正区域.

设 $P_i \in U/P$, 若对于任意的 $x, y \in P_i$, $x \neq y$, 有 $f(x, D) = f(y, D)$, 则称 P_i 是 U/P 中的一致等价类, 否则称为 U/P 的不一致等价类. 称 $U_1 = \mathrm{POS}_C(D)$ 为决策表的一致对象集, $U_2 = U - U_1$ 为决策表的不一致对象集.

显然, $\mathrm{POS}_P(D)$ 是 U/P 中一致等价类的并, $U - \mathrm{POS}_P(D)$ 是 U/P 中不一致等价类的并.

定义 6　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, 设 $P \subseteq C$, 若 $\mathrm{POS}_P(D) = \mathrm{POS}_C(D)$, 则称 P 是 C 的代数协调集; 若 P 是 C 的代数协调集且对于任意的 $P' \subset P$ 有 $\mathrm{POS}_{P'}(D) \neq \mathrm{POS}_C(D)$, 则称 P 是 C 相对于 D 的代数约简.

定理 1　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle, a \in C$ 是代数约简下的核属性当且仅当 $\mathrm{POS}_{C-\{a\}}(D) \neq \mathrm{POS}_C(D)$[17].

定义 7　给定信息系统 $S = \langle U, V, f, A\rangle$, $P \subseteq A$, 记 $U/P = \{P_1, P_2, \cdots, P_m\}$, 则 P 在对象集 U 上的知识粒度 $\mathrm{GK}(P)$ 定义为

$$\mathrm{GK}(P) = \sum_{P_k\in U/P} |P_k|^2/|U|^2. \tag{2}$$

定义 8　给定信息系统 $S = \langle U, V, f, A\rangle$, $P, Q \subseteq A$, 则 Q 关于 P 的相对粒度 $\mathrm{GK}(Q/P)$ 定义为

$$\mathrm{GK}(Q/P) = \mathrm{GK}(P) - \mathrm{GK}(P \bigcup Q). \tag{3}$$

定义 9　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, 设 $P \subseteq C$, 若 $\mathrm{GK}(D/P) = \mathrm{GK}(D/C)$, 则称 P 是 C 的相对粒度协调集; 若 P 是 C 的相对粒度协调集且对于任意的 $P' \subset P$ 有 $\mathrm{GK}(D/P') \neq \mathrm{GK}(D/C)$, 则称 P 为 C 相对于 D 的基于相对粒度属性约简, 简称为相对粒度约简.

文献 [16] 证明了相对粒度约简仅与HU属性约简等价, 文献 [17] 进一步给出了相对粒度的一种递归计算方法.

定理 2　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, $P \subseteq C$, 则 P 是 C 的 HU 属性约简当且仅当 P 是 C 的相对粒度约简. P 是 C 的 HU 协调属性集当且仅当 P 是 C 的相对粒度协调集[16].

定理 3　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, $P \subseteq C, U/P = \{P_1, P_2, \cdots, P_m\}$, $a \in C - P$, 则有

$$\mathrm{GK}(D/(P \bigcup \{a\})) = \sum_{P_k\in U/P} \mathrm{GK}_{P_k}(D/\{a\}), \tag{4}$$

其中 $\mathrm{GK}_{P_k}(D/\{a\})$ 表示在等价类 P_k 上 D 相对于 $\{a\}$ 的相对粒度[17].

2 代数约简概念的知识粒度表示

当决策表不一致时, 基于知识粒度的相对粒度约简与代数约简并不等价. 具体考察例 1 的算例.

例 1 给定决策表 S_1 如表 1 所示. 其中: $U_1 = \{x_1, x_2, \cdots, x_{10}\}$, $C = \{a, b, c, e, f, g\}$, $D = \{d\}$.

表 1 决策表 S_1

U	a	b	c	e	f	g	d
x_1	0	0	0	0	0	0	0
x_2	1	0	0	0	0	0	2
x_3	0	1	1	1	1	1	1
x_4	0	0	0	1	0	0	2
x_5	0	1	1	1	0	1	1
x_6	0	0	1	1	0	0	0
x_7	0	0	1	1	0	0	1
x_8	0	1	1	1	0	1	1
x_9	0	0	0	1	0	0	1
x_{10}	1	0	0	0	0	0	0

决策表 S_1 的相对粒度约简分别为 $\{a, b, c, e\}$ 和 $\{a, c, e, g\}$, 其代数约简分别为 $\{a, b, e\}$ 和 $\{a, e, g\}$, 两者并不相同.

由于决策表的代数约简是保持正区域不变的极小条件属性集, 其本质是在 C 中删除不必要的属性, 同时保持正区域不变, 但在删除属性时可能会导致等价类合并. 因此如果要保持正区域不变, 则这种合并只能属于以下两种情形: 1) 具有相同决策属性值的等价类合并, 即等价类本身在合并前已经是一致的且它们具有相同决策属性值; 2) 与正区域无关的等价类(即 U_2 中的等价类) 合并. 另一方面, 当决策表一致时, $\mathrm{GK}(D/C) = 0$; 当决策表不一致时, $\mathrm{GK}(D/C) > 0$. 因此, 现有相对粒度很难统一刻画出一致和不一致决策表的数字特征. 又因决策表一致时, 现有相对粒度约简与代数约简是等价的, 这意味着 $\mathrm{GK}(D/C) = 0$ 是合理的, 故要求新提出的修正相对粒度在决策表不一致时也能取到 0 值. 同时, 当不一致决策表退化为一致决策表时, 它将退化为现有相对粒度公式. 基于上述分析, 提出如下修正相对粒度公式.

定义 10 给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, $P \subseteq C$, 记 $U/P = \{P_1, P_2, \cdots, P_m\}$, $U/D = \{D_1, D_2, \cdots, D_r\}$, 定义修正相对粒度公式为

$$\mathrm{MGK}(D/P) = \mathrm{GK}(D/P) - \mathrm{GK}_{U_2}(D/P), \quad (5)$$

其中 $\mathrm{GK}_{U_2}(D/P) = \mathrm{GK}_{U_2}(P) - \mathrm{GK}_{U_2}(P \bigcup D)$. 用下式分别表示 P 和 $P \bigcup D$ 在 U_2 上的知识粒度:

$$\mathrm{GK}_{U_2}(P) = \sum_{P_k \in U/P} |P_k \bigcap U_2|^2/|U|^2,$$

$$\mathrm{GK}_{U_2}(P \bigcup D) = \sum_{P_k \in U/P} \sum_{D_j \in U/D} |P_k \bigcap U_2 \bigcap D_j|^2/|U|^2.$$

特别地, 当决策表一致时, $U_2 = \varnothing$, $\mathrm{MGK}(D/P)$ 将退化为 $\mathrm{GK}(D/P)$, 因此式 (5) 是对 (3) 的推广, 而且与式 (3) 相比, 它同时考虑到不一致对象的影响.

引理 1 给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, 设 $P \subseteq Q \subseteq C$, $U/Q = \{Q_1, Q_2, \cdots, Q_s\}$, $U/D = \{D_1, D_2, \cdots, D_r\}$, $U/P = \{Q_1, Q_2, \cdots, Q_{i-1}, Q_{i+1}, \cdots, Q_{k-1}, Q_{k+1}, \cdots, Q_i \bigcup Q_k\}$, 则

$$|Q_i||Q_k| - \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j| \geqslant 0, \quad (6)$$

并且等号成立的充要条件是, 存在 $D_{j_0} \in U/D$, 使得 $(Q_i \bigcup Q_k) \subseteq D_{j_0}$.

证明 因为 $|Q_i \bigcap D_j| \leqslant |Q_i|$ 且 $\bigcup_{j=1}^{r} D_j = U$, 所以

$$\sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j| \leqslant |Q_i| \sum_{D_j \in U/D} |Q_k \bigcap D_j| = |Q_i||Q_k|.$$

下面证明等号成立的充要条件. 因为 $|Q_k \bigcap D_j| = |Q_k| - |Q_k \bigcap D_j^C|$ (D_j^C 表示 D_j 的补集), 所以

$$|Q_i||Q_k| - \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j| = \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j^C|.$$

利用对称性, 类似地有

$$|Q_i||Q_k| - \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j| = \sum_{D_j \in U/D} |Q_k \bigcap D_j||Q_i \bigcap D_j^C|.$$

因此式 (6) 等号成立的充要条件是, 对于任意的 $D_j \in U/D$, $|Q_i \bigcap D_j||Q_k \bigcap D_j^C| = 0$ 且 $|Q_k \bigcap D_j||Q_i \bigcap D_j^C| = 0$. 分两种情况讨论.

1) 若 $|Q_i \bigcap D_j| = 0$, 则 $Q_i \bigcap D_j = \varnothing$, 从而 $|Q_i \bigcap D_j^C| = |Q_i| \neq 0$, 又因 $|Q_k \bigcap D_j||Q_i \bigcap D_j^C| = 0$ 同时成立, 故 $Q_k \bigcap D_j = \varnothing$, $(Q_i \bigcup Q_k) \bigcap D_j = \varnothing$.

2) 若 $|Q_k \bigcap D_j^C| = 0$, 则 $Q_k \subseteq D_j$, 类似地可推出 $Q_i \subseteq D_j$, 从而 $(Q_i \bigcup Q_k) \subseteq D_j$.

综合 1) 和 2) 的讨论可知: 对于任意 $D_j \in U/D$, 或者 $D_j \bigcap (Q_i \bigcup Q_k) = \varnothing$, 或者 $(Q_i \bigcup Q_k) \subseteq D_j$; 又因 $\bigcup_{j=1}^{r} D_j = U$, 所以必存在某个 $D_{j_0} \in U/D$, 使得 $(Q_i \bigcup Q_k) \subseteq D_{j_0}$. 反之结论也成立. □

定理 4 给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, $P \subseteq Q \subseteq C$, $U/Q = \{Q_1, Q_2, \cdots, Q_s\}$, $U/D = \{D_1, D_2, \cdots, D_r\}$, $U/P = \{Q_1, Q_2, \cdots, Q_{i-1}, Q_{i+1}, \cdots, Q_{k-1}, Q_{k+1}, \cdots, Q_i \bigcup Q_k\}$, 则 $\mathrm{GK}(D/P) \geqslant \mathrm{GK}(D/Q)$, 且等号成立的充要条件是, 存在 $D_{j_0} \in U/D$, 使得 $(Q_i \bigcup$

$Q_k) \subseteq D_{j_0}$.

证明　因为 $Q_i \bigcap Q_k = \varnothing$, 所以 $|Q_i \bigcup Q_k| = |Q_i| + |Q_k|$, $|(Q_i \bigcup Q_k) \bigcap D_j| = |Q_i \bigcap D_j| + |Q_k \bigcap D_j|$, 由定义 8 知

$$\mathrm{GK}(D/P) - \mathrm{GK}(D/Q) = 2\Big[|Q_i||Q_k| - \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j|\Big]/|U|^2.$$

由引理 1 有 $\mathrm{GK}(D/P) \geqslant \mathrm{GK}(D/Q)$, 且等号成立的充要条件是, 存在 $D_{j_0} \in U/D$, 使得 $(Q_i \bigcup Q_k) \subseteq D_{j_0}$. □

定理 4 表明, 当且仅当具有相同决策属性值的等价类合并才能保证现有相对粒度不变, 否则都会引起它变大, 包括与正区域无关的两等价类合并的情形, 这说明现有相对粒度与正区域在语义上不一致.

定理 5　给定决策表 $S = \langle U, V, f, C \bigcup D\rangle$, $P \subseteq Q \subseteq C$, $U/Q = \{Q_1, Q_2, \cdots, Q_s\}$, $U/D = \{D_1, D_2, \cdots, D_r\}$, $U/P = \{Q_1, Q_2, \cdots, Q_{i-1}, Q_{i+1}, \cdots, Q_{k-1}, Q_{k+1}, \cdots, Q_i \bigcup Q_k\}$, 则 $\mathrm{MGK}(D/P) \geqslant \mathrm{MGK}(D/Q)$, 并且等号成立的充要条件是, $(Q_i \bigcup Q_k) \subseteq \mathrm{POS}_P(D)$ 或 $(Q_i \bigcup Q_k) \subseteq U_2$.

证明　因为 $P \subseteq Q \subseteq C$, 所以 $\mathrm{POS}_P(D) \subseteq \mathrm{POS}_Q(D) \subseteq \mathrm{POS}_C(D)$, $U - \mathrm{POS}_C(D) \subseteq U - \mathrm{POS}_Q(D) \subseteq U - \mathrm{POS}_P(D)$. 又因 $Q_i \bigcap Q_k = \varnothing$, 故 $|Q_i \bigcup Q_k| = |Q_i| + |Q_k|$, $|(Q_i \bigcup Q_k) \bigcap D_j| = |Q_i \bigcap D_j| + |Q_k \bigcap D_j|$, $|(Q_i \bigcup Q_k) \bigcap U_2| = |Q_i \bigcap U_2| + |Q_k \bigcap U_2|$, $|(Q_i \bigcup Q_k) \bigcap U_2 \bigcap D_j| = |Q_i \bigcap U_2 \bigcap D_j| + |Q_k \bigcap U_2 \bigcap D_j|$. 因此有

$$\begin{aligned}&\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = \\ &2\Big[|Q_i||Q_k| - |Q_i \bigcap U_2||Q_k \bigcap U_2| - \\ &\sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j| + \\ &\sum_{D_j \in U/D} |Q_i \bigcap U_2 \bigcap D_j||Q_k \bigcap U_2 \bigcap D_j|\Big]\Big/|U|^2.\end{aligned}$$

对于 Q_i, Q_k 的取值, 可分为下列 4 种情形之一: 1) 两个都在 $\mathrm{POS}_Q(D)$ 中且有相同决策属性值; 2) 两个都在 $\mathrm{POS}_Q(D)$ 中但决策属性值不同; 3) 一个在 $\mathrm{POS}_Q(D)$ 中, 另一个在 $U - \mathrm{POS}_Q(D)$ 中; 4) 两个都在 $U - \mathrm{POS}_Q(D)$ 中.

对于情形 1), 存在 $D_{j_0} \in U/D$ 使得 $Q_i \subseteq D_{j_0}$, $Q_k \subseteq D_{j_0}$, 从而 $Q_i \bigcap U_2 = \varnothing$, $Q_k \bigcap U_2 = \varnothing$, $\sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j| = |Q_i||Q_k|$, 所以 $\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = 0$.

对于情形 2), 存在 $D_{j_1}, D_{j_2} \in U/D$ 且 $D_{j_1} \neq D_{j_2}$, 使得 $Q_i \subseteq D_{j_1}, Q_k \subseteq D_{j_2}$, 从而 $\sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j| = 0$, 又因 $|Q_i \bigcap U_2| = |Q_k \bigcap U_2| = 0$, 所以 $\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = 2|Q_i||Q_k|/|U|^2 > 0$.

对于情形 3), 不妨设 $Q_i \subseteq \mathrm{POS}_Q(D)$, $Q_k \subseteq U - \mathrm{POS}_Q(D)$, 则有 $Q_i \subseteq U_1$, 从而 $|Q_i \bigcap U_2| = 0$, $\sum_{D_j \in U/D} |Q_i \bigcap U_2 \bigcap D_j|^2 = 0$, 所以由引理 1 有

$$\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = 2\Big[|Q_i||Q_k| - \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j|\Big]\Big/|U|^2 > 0.$$

对于情形 4), 因为 $Q \subseteq C$, 所以 $U_2 \subseteq U - \mathrm{POS}_Q(D)$, 且 $U - \mathrm{POS}_Q(D)$ 的等价类是由 U/C 中的若干个等价类合并而成, 因此, 如果 Q_i, Q_k 都在 $U - \mathrm{POS}_Q(D)$ 中, 又需分 4 种情况分别讨论.

① 都在 U_2 中, 从而 $|Q_i \bigcap U_2||Q_k \bigcap U_2| = |Q_i||Q_k|$, $|Q_i \bigcap U_2 \bigcap D_j||Q_k \bigcap U_2 \bigcap D_j| = |Q_i \bigcap D_j||Q_k \bigcap D_j|$, 易得 $\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = 0$;

② 至少有一个与 U_2 相交为空, 不妨设 $Q_i \bigcap U_2 = \varnothing$, 从而

$$\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = 2\Big[|Q_i||Q_k| - \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_k \bigcap D_j|\Big]\Big/|U|^2.$$

根据引理 1, $\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) > 0$.

③ 一个在 U_2 中, 另一个与 U_2 相交不空但又不完全包含于 U_2, 不妨设 $Q_i \subseteq U_2$, $Q_k \bigcap U_2 \neq \varnothing$ 且 $Q_k \not\subset U_2$. 显然, Q_k 是由 U_1 中的若干等价类与 U_2 中的若干等价类合并而成, 记其 U_1 中等价类的并为 Q_{k_1}, U_2 中等价类的并为 Q_{k_2}, 则 $Q_k = Q_{k_1} \bigcup Q_{k_2}$, $Q_{k_1} \subseteq U_1$, $Q_{k_2} \subseteq U_2$, 从而有

$$\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = 2\Big[|Q_i||Q_{k_1}| - \sum_{D_j \in U/D} |Q_i \bigcap D_j||Q_{k_1} \bigcap D_j|\Big]\Big/|U|^2.$$

根据引理 1, $\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) > 0$.

④ 两个都与 U_2 相交不空, 且都不完全包含于 U_2. 类似于情形 ③ 的分析, 可设 $Q_i = Q_{i_1} \bigcup Q_{i_2}$, $Q_k = Q_{k_1} \bigcup Q_{k_2}$. 其中: $Q_{i_1} \subseteq U_1$, $Q_{i_2} \subseteq U_2$, $Q_{k_1} \subseteq U_1$, $Q_{k_2} \subseteq U_2$. 则有

$$\begin{aligned}&\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) = \\ &2\Big[|Q_{i_1}||Q_{k_1}| - \sum_{D_j \in U/D} |Q_{i_1} \bigcap D_j||Q_{k_1} \bigcap D_j| + \\ &|Q_{i_1}||Q_{k_2}| - \sum_{D_j \in U/D} |Q_{i_1} \bigcap D_j||Q_{k_2} \bigcap D_j| + \\ &|Q_{i_2}||Q_{k_1}| - \sum_{D_j \in U/D} |Q_{i_2} \bigcap D_j||Q_{k_1} \bigcap D_j|\Big]\Big/|U|^2.\end{aligned}$$

根据引理 1 有 $\mathrm{MGK}(D/P) - \mathrm{MGK}(D/Q) > 0$.

综上所述, 不管 Q_i、Q_k 取值如何, 只要它们合并, 都将导致修正相对粒度单调不减, 当且仅当参与合并

的等价类Q_i、Q_k都属于正区域且具有相同决策属性值, 或者Q_i、Q_k都在U_2中时其值才保持不变. □

定理 6 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq Q\subseteq C$, 则$\mathrm{POS}_P(D)=\mathrm{POS}_C(D)$的充要条件是$\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$.

证明 先证明$\mathrm{MGK}(D/C)=0$. 假设$U/C=\{X_1,X_2,\cdots,X_n\}$, $U/D=\{D_1,D_2,\cdots,D_r\}$, 记

$$\mathrm{GK}(C)=\sum_{X_i\in U/C}|X_i|^2/|U|^2,$$

$$\mathrm{GK}(C\bigcup D)=\sum_{X_i\in U/C}\sum_{D_j\in U/D}|X_i\bigcap D_j|^2/|U|^2.$$

当$X_i\subseteq U_1$时, $\sum\limits_{D_j\in U/D}|X_i\bigcap D_j|^2/|U|^2=|X_i|^2/|U|^2$, 从而

$$\mathrm{GK}(C)-\mathrm{GK}(C\bigcup D)=\sum_{X_i\subseteq U_2}|X_i|^2/|U|^2-\sum_{X_i\subseteq U_2}\sum_{D_j\in U/D}|X_i\bigcap D_j|^2/|U|^2.$$

又因$X_i\subseteq U_1$时, $|X_i\bigcap U_2|=0$, $X_i\subseteq U_2$时, $|X_i\bigcap U_2|=|X_i|$, 故

$$\mathrm{GK}_{U_2}(C)-\mathrm{GK}_{U_2}(C\bigcup D)=\sum_{X_i\subseteq U_2}|X_i|^2/|U|^2-\sum_{X_i\subseteq U_2}\sum_{D_j\in U/D}|X_i\bigcap D_j|^2/|U|^2.$$

所以$\mathrm{MGK}(D/C)=0$.

下面证明定理6中的结论. 对于任意$P_k\in U/P$, 当$P_k\in\mathrm{POS}_P(D)$时, $\sum\limits_{D_j\in U/D}|P_k\bigcap D_j|^2/|U|^2=|P_k|^2/|U|^2$. 根据题设, 若$\mathrm{POS}_P(D)=\mathrm{POS}_C(D)$, 则$U-\mathrm{POS}_P(D)=U_2$, 从而

$$\mathrm{GK}(P)-\mathrm{GK}(P\bigcup D)=\sum_{P_k\subseteq U_2}|P_k|^2/|U|^2-\sum_{P_k\subseteq U_2}\sum_{D_j\in U/D}|P_k\bigcap D_j|^2/|U|^2=\mathrm{GK}_{U_2}(P)-\mathrm{GK}_{U_2}(P\bigcup D).$$

所以$\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$.

反之, 如果$\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$, 则一定有$\mathrm{POS}_P(D)=\mathrm{POS}_C(D)$成立. 现假设结论不成立. 因为$P\subseteq C$, 所以有$\mathrm{POS}_P(D)\subset\mathrm{POS}_C(D)$, 这意味着至少存在一个$x_0\in\mathrm{POS}_C(D)$, 但$x_0\notin\mathrm{POS}_P(D)$, 从而在$[x_0]_P$中至少存在两个相异元素$x_s,x_t\in[x_0]_P$, 使得$f(x_s,D)\neq f(x_t,D)$. 又因$[x_0]_P$是由若干个$U/C$的等价类合并而成, 且一定包含等价类$[x_0]_C$, 所以$[x_0]_P$不可能是定理5中等号成立的两种情形, 从而$\mathrm{MGK}(D/P)>0$且$\mathrm{MGK}(D/C)=0$与假设矛盾. 因此结论成立. □

定理6表明, 保持正区域不变与保持修正相对粒度不变相互等价, 因此可通过修正相对粒度来刻画正区域的变化情况以及代数约简的各种概念.

定义 11 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, 对于任意的$a\in P$, 若$\mathrm{MGK}(D/P-\{a\})=0$, 则称$a$为$P$中在代数约简意义下的不必要属性, 否则称$a$为$P$中必要的属性.

定义 12 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, 对于任意的$a\in C$, 若$\mathrm{MGK}(D/C-\{a\})>0$, 则称$a$为决策表的代数约简核属性, 简称为代数核.

根据定理6和定义6可以给出代数约简的知识粒度描述.

定义 13 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, 若$\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$, 则称$P$是$C$的代数协调集. 如果$P$是$C$的代数协调集, 且对于任意的$P'\subset P$, $\mathrm{MGK}(D/P')>\mathrm{MGK}(D/C)$, 则称$P$是$C$的代数约简.

对于现有的相对粒度与修正相对粒度有如下关系.

定理 7 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, 若$\mathrm{GK}(D/P)=\mathrm{GK}(D/C)$, 则一定有$\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$.

证明 由于$P\subseteq C$, U/P中的等价类都可由U/C中若干等价类合并而成. 如果$\mathrm{GK}(D/P)=\mathrm{GK}(D/C)$, 则根据定理4, 对于任意参与合并的两个等价类$X_i,X_k\in U/C$, 一定存在$D_{j_0}\in U/D$, 使得$(X_i\bigcup X_k)\subseteq D_{j_0}$. 再由定理5有$\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$. □

反之则不成立. 如例1的S_1, $\mathrm{MGK}(D/\{a,b,e\})=\mathrm{MGK}(D/C)$, 但$\mathrm{GK}(D/\{a,b,e\})\neq\mathrm{GK}(D/C)$.

利用定理7可进一步解释清楚HU属性约简与代数约简之间的关系. 根据定义9和定理2可先给出HU属性约简的知识粒度表示.

定义 14 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, 对于任意的$a\in P$, 若$\mathrm{GK}(D/P-\{a\})=0$, 则称a为P中在HU属性约简意义下的不必要属性, 否则称a为P中必要的属性.

定义 15 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, 如果$\mathrm{GK}(D/P)=\mathrm{GK}(D/C)$, 则称$P$是$C$的HU差别矩阵属性协调集, 简称为HU协调集. 如果P是C的HU协调集且对于任意的$P'\subset P$, 有$\mathrm{GK}(D/P')>\mathrm{GK}(D/C)$, 则称$P$是$C$的HU属性约简.

定义 16 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, 对于任意的$a\in C$, 若$\mathrm{GK}(D/C-\{a\})>\mathrm{GK}(D/C)$, 则称$a$为决策表的HU属性约简核属性, 简称为HU核.

推论 1 给定决策表$S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, 若P是C的HU协调集, 则P是C的代数协调集.

证明 根据定理7、定义13和定义15即知结论成立. □

推论 2　给定决策表 $S=\langle U,V,f,C\bigcup D\rangle$, 若 $a\in C$ 是代数核, 则它一定是 HU 核.

证明　令 $P=C-\{a\}$, 根据定理 7 的逆否命题即知结论成立. □

例 2　仍然使用例 1 的决策表 S_1 来阐述如何利用知识粒度求取代数约简并结合定理 5 给出相应的解释.

由于 $U/C=\{\{x_1\},\{x_2,x_{10}\},\{x_3\},\{x_4,x_9\},\{x_5,x_8\},\{x_6,x_7\}\}$, $U/D=\{\{x_1,x_6,x_{10}\},\{x_2,x_4\},\{x_3,x_5,x_7,x_8,x_9\}\}$, 易得 $U_1=\{x_1,x_3,x_5,x_8\}$, $U_2=\{x_2,x_4,x_6,x_7,x_9,x_{10}\}$. 若设 $P=\{a,b,e\}$, 则根据代数约简的定义可以验证它是 C 的一个代数约简.

另一方面, 根据 U/C 和 $U/(C\bigcup D)=\{\{x_1\},\{x_2\},\{x_{10}\},\{x_3\},\{x_4\},\{x_9\},\{x_5,x_8\},\{x_6\},\{x_7\}\}$, 可得 $\mathrm{GK}(D/C)=6/100$, 又因 $\mathrm{GK}_{U_2}(D/C)=6/100$, 所以 $\mathrm{MGK}(D/C)=0$.

若设 $P=\{a,b,e\}$, 根据 $U/P=\{\{x_1\},\{x_2,x_{10}\},\{x_3,x_5,x_8\},\{x_4,x_6,x_7,x_9\}\}$, $U/(P\bigcup D)=\{\{x_1\},\{x_2\},\{x_{10}\},\{x_3,x_5,x_8\},\{x_4\},\{x_6\},\{x_7,x_9\}\}$, 可得 $\mathrm{GK}(D/P)=12/100$, $\mathrm{GK}_{U_2}(D/P)=12/100$, 因此 $\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$, 由定理 6 的结论有 $\mathrm{POS}_P(D)=\mathrm{POS}_C(D)$. 更进一步, 可以验证, 对于任意 $P'\subset P$, $\mathrm{MGK}(D/P')\neq\mathrm{MGK}(D/C)$, 同样可由修正相对粒度变化情况判定 P 是 C 的代数约简.

再深入分析会发现, 由于 $P\subseteq C$, 对于任意的 $P_i\in U/P$ 都可由 U/C 的等价类合并而成. 根据定理 5, 当参与合并的等价类具有相同的决策属性值, 或者能同时被 U_2 所包含时, 它们的合并不会导致修正相对粒度变大. 例如本例中, U/P 中的 $\{x_3,x_5,x_8\}$ 是由 $\{x_3\}$ 和 $\{x_5,x_8\}$ 合并而成, $\{x_4,x_6,x_7,x_9\}$ 由 $\{x_4,x_9\}$ 和 $\{x_6,x_7\}$ 合并得到. 由于 $\{x_3\}$ 和 $\{x_5,x_8\}$ 具有相同决策属性值, 它们的合并不会引起修正相对粒度变大; 同理, $\{x_4,x_9\}\subseteq U_2$, $\{x_6,x_7\}\subseteq U_2$, 它们的合并也不会引起修正相对粒度的变大. 从而有 $\mathrm{MGK}(D/P)=\mathrm{MGK}(D/C)$, 继而保证 $\mathrm{POS}_P(D)=\mathrm{POS}_C(D)$ 成立, 但 $\mathrm{GK}(D/P)\neq\mathrm{GK}(D/C)$. 这说明 P 是 C 的代数协调集但不是 HU 协调集.

若令属性集 $Q=\{a,b,c,e\}$, 则由于 $U/Q=\{\{x_1\},\{x_2,x_{10}\},\{x_3,x_5,x_8\},\{x_4,x_9\},\{x_6,x_7\}\}$, 它也是由 U/C 中某些等价类合并而得到, 但实际上只发生了一个合并事件, 即 $\{x_3\}$ 与 $\{x_5,x_8\}$ 合并成 $\{x_3,x_5,x_8\}$. 由于它们具有相同的决策属性值, 根据定理 4, $\mathrm{GK}(D/Q)=\mathrm{GK}(D/C)$. 更进一步, 可以验证 Q 是 C 的 HU 属性约简. 但 $P=\{a,b,e\}$ 则不同, 它不仅会引起具有相同决策属性值的等价类 $\{x_3\}$ 与 $\{x_5,x_8\}$ 合并成 $\{x_3,x_5,x_8\}$, 而且会引起不一致对象集 $\{x_4,x_9\}$ 与 $\{x_6,x_7\}$ 合并成 $\{x_4,x_6,x_7,x_9\}$, 根据定理 4, 必然会引起相对粒度变大, 因此有 $\mathrm{GK}(D/P)\neq\mathrm{GK}(D/C)$.

核属性方面, 由于 $U/C-\{c\}=\{\{x_1\},\{x_2,x_{10}\},\{x_3\},\{x_5,x_8\},\{x_4,x_6,x_7,x_9\}\}$, 它引起不一致对象集 $\{x_4,x_9\}$ 与 $\{x_6,x_7\}$ 合并成 $\{x_4,x_6,x_7,x_9\}$. 根据定理 4, $\mathrm{GK}(D/C-\{c\})>\mathrm{GK}(D/C)$, 但根据定理 5, 这种合并不会引起修正相对粒度变化, 因为 $\{x_4,x_9\}\subseteq U_2$, $\{x_6,x_7\}\subseteq U_2$, 所以属性 c 是 HU 核但不是代数核. 对于属性 a, 因为 $U/C-\{a\}=\{\{x_1,x_2,x_{10}\},\{x_3\},\{x_4,x_9\},\{x_5,x_8\},\{x_6,x_7\}\}$, 它引起一致等价类 $\{x_1\}$ 与不一致等价类 $\{x_2,x_{10}\}$ 合并成 $\{x_1,x_2,x_{10}\}$, 所以 $\mathrm{GK}(D/C-\{a\})>\mathrm{GK}(D/C)$, $\mathrm{MGK}(D/C-\{a\})>\mathrm{MGK}(D/C)$, 因此属性 a 既是代数核也是 HU 核.

3　基于粒度计算的代数约简高效算法

上节给出了一个与代数约简相适应的修正相对粒度 $\mathrm{MGK}(D/P)$, 它满足单调递减性. 本节将利用这种单调递减性给出刻画属性重要性的计算公式及相应的高效约简算法.

定义 17　给定决策表 $S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, 对于任意 $a\in C-P$ 关于 P 的相对重要性为

$$\mathrm{Sig}_P(a)=\mathrm{MGK}(D/P)-\mathrm{MGK}(D/(P\bigcup\{a\})).\quad(7)$$

定理 8　给定决策表 $S=\langle U,V,f,C\bigcup D\rangle$, $P\subseteq C$, $U/P=\{P_1,P_2,\cdots,P_m\}$, $a\in C-P$, 有

$$\mathrm{MGK}(D/(P\bigcup\{a\}))=\sum_{P_k\in U/P}\mathrm{MGK}_{P_k}(D/\{a\}).\quad(8)$$

其中: $\mathrm{MGK}_{P_k}(D/\{a\})=\mathrm{GK}_{P_k}(D/\{a\})-\mathrm{GK}_{U_2\bigcap P_k}(D/\{a\})$ 表示在对象集合 P_k 上的修正相对粒度, $\mathrm{GK}_{P_k}(D/\{a\})$ 表示在对象集合 P_k 上 D 相对于 a 的相对粒度, $\mathrm{GK}_{U_2\bigcap P_k}(D/\{a\})$ 表示在集合 $U_2\bigcap P_k$ 上 D 相对于 a 的相对粒度.

证明　假设 $U/P=\{P_1,P_2,\cdots,P_m\}$, $U/\{a\}=\{R_1,R_2,\cdots,R_s\}$, $U/(P\bigcup\{a\})=\{Q_1,Q_2,\cdots,Q_n\}$, 由于 $U/(P\bigcup\{a\})=\bigcup\limits_{R_j\in U/\{a\},P_k\in U/P}(R_j\bigcap P_k)$, 有

$$\mathrm{GK}_{U_2}(P\bigcup\{a\})=\sum_{P_k\in U/P}\sum_{R_j\in U/\{a\}}|R_j\bigcap P_k\bigcap U_2|^2/|U|^2=\sum_{P_k\in U/P}\mathrm{GK}_{P_k\bigcap U_2}(\{a\}).$$

类似地, 有

$$\mathrm{GK}_{U_2}(D\bigcup P\bigcup\{a\})=\sum_{P_k\in U/P}\mathrm{GK}_{P_k\bigcap U_2}(D\bigcup\{a\}),$$

从而

$$\mathrm{GK}_{U_2}(D/(P\bigcup\{a\}))=\sum_{P_k\in U/P}\mathrm{GK}_{P_k\bigcap U_2}(D/\{a\}).$$

由定义 10 和式 (4), 有

$$\mathrm{MGK}(D/(P\bigcup\{a\})) =$$

$$\sum_{P_k\in U/P}[\mathrm{GK}_{P_k}(D/\{a\}) - \mathrm{GK}_{U_2\cap P_k}(D/\{a\})]. \quad \square$$

与文献[17]的算法类似, 应优先选取修正相对粒度最小的属性, 然后再依次根据属性重要性大小, 挑选最重要的属性. 由于对于任意的$a\in C-P$, MGK(D/P)是固定的, 根据式(7), 只需挑选出$\mathrm{MGK}(D/(P\bigcup\{a\}))$取值最小的属性即可满足属性重要性最大原则. 如此循环直至为0, 最后得到基于修正相对粒度的约简算法.

算法 1 基于修正相对粒度的启发式算法.

输入: 决策表S, 阈值ε;

输出: 代数约简R.

Step 1: 计算正区域U_1和矛盾对象集$U_2 = U - U_1$, 同时令$R = \varnothing$.

Step 2: 对于任意$a\in C$, 计算其修正相对粒度$\mathrm{MGK}(D/\{a\})$, 从中挑选取值最小的属性, 记为a', 其值记为W.

Step 3: 更新$R = R\bigcup\{a'\}$, $C = C-\{a'\}$, $U/R = \{P_1, P_2, \cdots, P_s\}$.

Step 4: 如果$|W|\geqslant\varepsilon$, 则转Step 5, 否则转Step 6.

Step 5: 对于任意$P_i\in U/R$, 分别计算其上的相对粒度$\omega_{i_1} = \mathrm{GK}_{P_i}(D/\{a\})$, 限制在$U_2$上的相对粒度$\omega_{i_2} = \mathrm{GK}_{U_2\cap P_i}(D/\{a\})$. 令$\mathrm{MGK}(D/R\bigcup\{a\}) = \sum_{i=1}^{s}(\omega_{i_1}-\omega_{i_2})$, 取计算结果最小的属性, 仍记为$a'$(有相同的任取一个), 令$W = \mathrm{MGK}(D/R\bigcup\{a\})$, 转Step 3.

Step 6: 对于任意$a\in R$, 从后往前依次检验其在R中是否可约, 若$|W|<\varepsilon$, 则把它从R中删除.

如果采用文献[18]提出的基排序方法, 算法1的时间复杂度分析与文献[17]类似, 结果相当: 对于只含单决策属性的决策表, 其时间复杂度为$O(|C^2||U|)$.

例 3 利用例1阐明算法1的计算过程.

Step 1: 计算正区域和矛盾对象集, 得到$U_1 = \{x_1, x_3, x_5, x_8\}$, $U_2 = \{x_2, x_4, x_6, x_7, x_9, x_{10}\}$, $R = \varnothing$.

Step 2: 对于任意$x\in C$, 计算其修正相对粒度$\mathrm{MGK}(D/\{x\})$, 结果如表2所示.

表 2 修正相对粒度$\mathrm{MGK}(D/\{x\})$

x					
a	b	c	e	f	g
24/100	8/100	12/100	14/100	28/100	8/100

由于b和g的取值同时达到最小, 任选其中一个, 这里选择b, 记$W = 8/100$.

Step 3: 更新$R = \{b\}$, $C = \{a, c, e, f, g\}$, 记$U/R = \{P_1, P_2\}$. 其中: $P_1 = \{x_1, x_2, x_4, x_6, x_7, x_9, x_{10}\}$, $P_2 = \{x_3, x_5, x_8\}$. 由于$W = 8/100 > 0$, 跳到Step 5, 对于任意$x\in C-R$, 分别在P_i上计算其修正相对粒度$\mathrm{MGK}_{P_i}(D/\{x\})$, 然后求和即可得到在$R = \{b\}$上增加属性$x$的修正相对粒度$\mathrm{MGK}(D/(\{b\}\bigcup\{x\}))$, 结果如表3所示.

表 3 修正相对粒度$\mathrm{MGK}(D/(\{b\}\bigcup\{x\}))$

x				
a	c	e	f	g
6/100	6/100	2/100	8/100	8/100

因$\mathrm{MGK}(D/(\{b\}\bigcup\{e\}))$最小, 故选取属性$e$, 更新$W = 2/100$; 转Step 3, 更新$R = \{b, e\}$, $C = \{a, c, f, g\}$, 此时$U/R = \{P_{11}, P_{12}, P_2\}$. 其中: $P_{11} = \{x_1, x_2, x_{10}\}$, $P_{12} = \{x_4, x_6, x_7, x_9\}$, $P_2 = \{x_3, x_5, x_8\}$. 分别在U/R的各等价类P_i上计算修正相对粒度$\mathrm{MGK}_{P_i}(D/\{x\})$, 求和即得增加该新属性后的修正相对粒度, 结果如表4所示.

表 4 修正相对粒度$\mathrm{MGK}(D/(\{b,e\}\bigcup\{x\}))$

x			
a	c	f	g
0/100	6/100	8/100	8/100

由于$\mathrm{MGK}(D/(\{b,e\}\bigcup\{a\}))$取值最小, 选取属性$a$, 更新$W = 0$, $R = \{b, e, a\}$, $C = \{c, f, g\}$. 此时$W = 0$, 跳转到Step 6, 最后得到$R = \{a, b, e\}$是C的一个代数约简. 如果在计算过程第1步选取属性g, 则最后将得到代数约简$R = \{a, e, g\}$.

4 实验比较

由于只有不一致决策表的代数约简与HU属性约简才有可能不一致, 为体现修正相对粒度的有效性, 首先对已报道过的不一致决策表分别用文献[17]的算法与本文算法1进行计算. 表5摘录出3个约简结果和核属性不同的算例及其文献出处. 因数据集太小, 程序运算时间都非常小, 相差也不大, 故此处省略.

在表5中, 本文S_1和文献[20]中的表1计算得到的相对粒度约简与基于修正相对粒度约简算法计算得到的结果都不同, 而且可以验证, 通过相对粒度约简算法得到的约简结果确实是HU属性约简, 而通过修正相对粒度约简算法得到的是一个代数约简. 在核属性方面, 表5中3个决策表的HU核和代数核都不同, 且代数核是HU核的子集.

进一步, 选用UCI中部分数据集在PC机上展开实验(型号及参数为: Intel Xeon 2.0 G, 1 G RAM, Windows Server 2003专业版SP2, C# 编程, 运行在dotnetfx3.0上), 并分别与文献[18, 21]的算法(分别记作算法A和算法B)进行比较, 本文算法记作算法C, 3个约简算法的实验结果如表6所示.

表 5　相对粒度与修正相对粒度约简结果比较

决策表	对象个数	条件属性个数	是否一致	一致对象个数	不一致对象个数	相对粒度约简算法		修正相对粒度约简算法	
						约简属性集	HU 核	约简属性集	代数核
本文的 S_1	10	6	否	4	6	$\{a,b,c,e\}$	$\{a,c,e\}$	$\{a,b,e\}$	$\{a,e\}$
文献 [19] 表 3	6	3	否	1	5	$\{a,b\}$	$\{a,b\}$	$\{a,b\}$	$\{b\}$
文献 [20] 表 1	18	6	否	1	17	$\{a,b,c,e,f\}$	$\{a,e,c,f\}$	$\{a,e\}$	$\{a,e\}$

表 6　3 个约简算法的比较

决策表	对象个数	条件属性个数	核属性数目	算法 A		算法 B		算法 C	
				约简后属性数	执行时间/ms	约简后属性数	执行时间/ms	约简后属性数	执行时间/ms
S_1	10	6	2	5	<0.01	3	<0.01	3	<0.01
Voting	435	16	7	10	15.625	9	15.625	9	15.625
Tic-tac-toe	958	9	0	8	15.625	8	31.25	8	15.625
Credit	1 000	20	1	8	1 640.625	8	9 828.125	8	1 687.5
Chess-end-game	3196	36	27	31	406.25	29	687.5	29	625
Mushroom	8 124	22	0	5	171.875	5	484.375	**4**	203.125

由表 6 可以看出: 对于算法 A, 由于它是不完备的, 在约简结果方面稍微差些, 有时得到的约简结果比算法 B 和算法 C 的约简集大些, 如 Voting、Chess-end-game 和 Mushroom 数据集. 其中算法 A 在 Voting 和 Mushroom 上即使增加回溯步骤也不能得到更好的约简结果, Chess-end-game 可以经回溯步骤缩减到只含 29 个属性的约简集. 算法 C 所得约简结果最好, 在各个测试数据集上得到的都是最优约简结果. 运行时间方面, 在小数据集上, 上述 3 种算法相差不大; 随着数据集的增大, 3 种约简算法的执行时间都增大. 有几个现象值得注意: 1) 当数据集没有核属性时, 算法 B 的耗时大于算法 A 和算法 C, 尤其在数据集 Credit 上, 这种现象更加明显; 算法 B 的总约简时间为 9 828.125 ms, 其中求核时间为 8 140.625 ms, 如果不从核属性出发, 则算法 B 与其他两种算法所花的时间差不多. 2) 虽然 Mushroom 的数据对象和条件属性较多, 但 3 种算法的求约简时间都不大, 特别是算法 A, 只用时 171.875 ms, 这可能与 Mushroom 的约简结果含属性较少有关 (只有 4 个或 5 个属性). 算法 B 用时多些, 同样是因为求核花去大部分时间, 但 Mushroom 没有核属性. 在 Tic-tac-toe 和 Chess-end-game 数据集上, 由于核属性较多, 算法 B 与算法 A、算法 C 相比, 其差别不像在 Credit 数据集上那么明显. 3) 如果不求核, 则上述 3 种约简算法在各数据集上计算时间差不多, 几乎相当, 算法 A 快些, 但也仅是快一点点, 这可能与它采用简化决策表有关系.

5　结　论

为了克服现有相对粒度约简与基于正区域的代数约简不等价问题, 本文首先对如何保持正区域不变进行语义分析, 给出了现有相对粒度保持不变的充要条件, 即只有当具有相同决策属性值的等价类发生合并时才不会引起现有相对粒度变大, 其他情形的合并都会导致其变大, 包括与正区域无关的不一致等价类的合并, 从而在语义上阐述了现有相对粒度约简与代数约简的不一致性问题. 为了给出代数约简的知识粒度描述, 本文提出了一种修正相对粒度计算公式, 从理论上证明了保持修正相对粒度不变与保持正区域不变等价, 继而给出了代数约简的知识粒度表示. 利用修正相对粒度的单调性, 设计了一种计算属性重要性的递归计算公式, 得到一种计算代数约简的高效算法. 与现有约简算法相比, 本文方法是完备的, 同时能够得到一些较好的约简结果. 在时间复杂度方面, 本文方法与已有的高效算法从理论上相当, 实际运算则慢些, 但相差不大. 与现有其他保证能求到决策表代数约简的改进方法相比, 如文献 [7, 9] 的方法, 本文方法最大的优势在于不必构造差别矩阵, 也不必先将不一致决策表转化为一致决策表这一中间过程, 本文方法是直接针对不一致决策表进行计算, 其计算效率只依赖于等价类基数的计算效率. 本文的后续工作将重点研究基于条件信息熵约简的知识粒度表示, 将本文方法推广到更多的约简概念上, 尝试建立一种统一的基于知识粒度表示的属性约简理论.

参考文献(References)

[1] Pawlak Z. Rough set[J]. Commucation of the ACM, 1995, 38(11): 89-95.

[2] Hu X H, Nick Cerene. Learning in relational databases: A rough set approach[J]. Int J of Computation Intelligence, 1995, 11(2): 323-338.

[3] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766.

(Wang G Y, Yu H, Yang D H. Decision table reduction based on conditional information entropy[J]. Chinese J of Computers, 2002, 25(7): 759-766.)

[4] Wang G Y, Zhao J, An J J, et al. A comparative study of algebra viewpoint and information viewpoint in attribute reduction[J]. Fundamenta Informaticae, 2005, 68(3): 289-301.

[5] 黄国顺, 刘云生. 不一致决策表各种属性约简的不一致性分析与转化[J]. 小型微型计算机系统, 2008, 29(4): 703-708.
(Huang G S, Liu Y S. Inconsistency analysis and translation of different types of attribute reduction for inconsistent decision tables[J]. J of Chinese Computer Systems, 2008, 29(4): 703-708.)

[6] 徐章艳, 杨炳儒, 宋威, 等. 几种不同属性约简的比较研究[J]. 小型微型计算机系统, 2008, 29(5): 848-853.
(Xu Z Y, Yang B R, Song W, et al. Comparative research of different attribute reduction definitions[J]. J of Chinese Computer Systems, 2008, 29(5): 848-853.)

[7] 葛浩, 李龙澍, 杨传健. 新的可分辨矩阵及其约简方法[J]. 控制与决策, 2010, 25(12): 1891-1895.
(Ge H, Li L S, Yang C J. New discernibility matrix and attribute reduction method[J]. Control and Decision, 2010, 25(12): 1891-1895.)

[8] 杨传健, 葛浩, 李龙澍. 垂直划分二进制可分辨矩阵的属性约简[J]. 控制与决策, 2013, 28(4): 563-568.
(Yang C J, Ge H, Li L S. Attribute reduction of vertically partitioned binary discernibility matrix[J]. Control and Decision, 2013, 28(4): 563-568.)

[9] 刘启和, 李凡, 闵帆, 等. 一种基于新的条件信息熵的高效知识约简算法[J]. 控制与决策, 2005, 20(8): 878-882.
(Liu Q H, Li F, Min F, et al. An efficient knowledge reduction algorithm based on new conditional information entropy[J]. Control and Decision, 2005, 20(8): 878-882.)

[10] 苗夺谦, 范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践, 2002, 22(1): 48-56.
(Miao D Q, Fan S D. The calculation of knowledge granulation and its application[J]. Systems Engineering-Theory & Practice, 2002, 22(1): 48-56.)

[11] Liang Jiye, Wang Junhong, Qian Yuhua. A new measure of uncertainty based on knowledge granulation for rough sets[J]. Information Sciences, 2009, 179(4): 458-470.

[12] 徐久成, 史进玲, 孙林. 一种基于相对粒度的决策表约简算法[J]. 计算机科学, 2009, 36(3): 205-207.
(Xu J C, Shi J L, Sun L. Attribute reduction algorithm based on relative granularity in decision tables[J]. Computer Science, 2009, 36(3): 205-207.)

[13] 陈玉明, 吴克寿, 谢荣生. 基于相对知识粒度的决策表约简[J]. 山东大学学报: 工学版, 2012, 42(6): 8-12.
(Chen Y M, Wu K S, Xie R S. Reduction for decision table based on relative knowledge granularity[J]. J of Shandong University: Engineering Science, 2012, 42(6): 8-12.)

[14] 冯琴荣, 苗夺谦, 程昳. 决策表属性约简的相对划分粒度表示[J]. 小型微型计算机系统, 2008, 29(12): 2305-2308.
(Feng Q R, Miao D Q, Cheng Y. Presentation of relative granularity of attribute reduction for decision tables[J]. J of Chinese Computer Systems, 2008, 29(12): 2305-2308.)

[15] Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: An accelerator for attribute reduction in rough set[J]. Artificial Intelligence, 2010, 174(9/10): 597-618.

[16] 曾凡智, 黄国顺, 文翰. 差别矩阵HU属性约简的几种等价表示[J]. 计算机工程, 2011, 37(16): 65-67.
(Zeng F Z, Huang G S, Wen H. Some equivalent representations of HU's attribute reduction based on discernibility matrix[J]. Computer Engineering, 2011, 37(16): 65-67.)

[17] 黄国顺, 曾凡智, 陈广义, 等. 一种HU差别矩阵属性约简的高效算法[J]. 华中科技大学学报: 自然科学版, 2012, 40(4): 8-12.
(Huang G S, Zeng F Z, Chen G Y, et al. Efficient algorithm of the attribute reduction using HU's discernibility matrix[J]. J of Huazhong University of Science and Technology: Natural Science Edition, 2012, 40(4): 8-12.)

[18] 徐章艳, 刘作鹏, 杨炳儒, 等. 一个复杂度为$\max(O(|C||U|), O(|C|^2|U/C|))$的快速属性约简算法[J]. 计算机学报, 2006, 29(3): 391-399.
(Xu Z Y, Liu Z P, Yang B R, et al. A quick attribute reduction algorithm with complexity of $\max(O(|C||U|), O(|C|^2|U/C|))$[J]. Chinese J of Computers, 2006, 29 (3): 391-399.)

[19] 王国胤. 决策表核属性的计算方法[J]. 计算机学报, 2003, 26(5): 611-615.
(Wang G Y. Calculation methods for core attribute of decision table[J]. Chinese J of Computers, 2003, 26(5): 611-615.)

[20] 李凡, 刘启和, 叶茂, 等. 不一致决策表的知识约简方法研究[J]. 控制与决策, 2006, 21(8): 857-862.
(Li F, Liu Q H, Ye M, et al. Approaches to knowledge reductions in inconsistent decision tables[J]. Control and Decision, 2006, 21(8): 857-862.)

[21] 葛浩, 李龙澍, 杨传健. 基于冲突域的高效属性约简算法[J]. 计算机学报, 2012, 35(2): 342-350.
(Ge H, Li L S, Yang C J. An efficient attribute reduction algorithm based on conflict region[J]. Chinese J of Computers, 2012, 35(2): 342-350.)

(责任编辑: 李君玲)