

差别矩阵 HU 属性约简的几种等价表示

曾凡智^a, 黄国顺^b, 文 翰^b

(佛山科学技术学院 a. 计算机系; b. 理学院, 广东 佛山 528000)

摘 要: 对于一致决策表, 现有基于相对粒度、相对划分粒度、知识量和同可区分度的属性约简与分别代数约简是等价的, 但对于不一致决策表, 它们与代数约简并不等价。为此, 针对不一致决策表, 建立相对粒度与新条件信息熵、知识量和同可区分度之间的线性关系, 从而得出结论: 现有基于相对粒度、相对划分粒度、知识量和同可区分度的属性约简本质上仅与基于差别矩阵的 HU 属性约简等价, 并通过设计一个不一致决策表验证该结论的正确性。

关键词: 差别矩阵; HU 属性约简; 条件信息熵; 相对粒度; 知识量; 同可区分度

Some Equivalent Representations of HU's Attribute Reduction for Discernibility Matrix

ZENG Fan-zhi^a, HUANG Guo-shun^b, WEN Han^b

(a. Department of Computer; b. Science School, Foshan University, Foshan 528000, China)

【Abstract】 For consistent decision table, the attribute reductions based on relative granularity, relative partition granularity, knowledge quantity and common discernibility degree are equivalent with the algebraic reduction respectively. But they are inconsistent while the decision table is inconsistent. For inconsistent decision table, it is proved that the relationship between new conditional information entropy, knowledge quantity, common discernibility degree and relative granularity is linear. It is followed that the attribute reductions based on relative granularity, relative partition granularity, knowledge quantity and common discernibility degree are just equivalent with the one based on HU's discernibility matrix. An inconsistent decision table is designed to illustrate the correctness of conclusion.

【Key words】 discernibility matrix; HU's attribute reduction; conditional information entropy; relative granularity; knowledge quantity; common discernibility degree

DOI: 10.3969/j.issn.1000-3428.2011.16.022

1 概述

由于基于差别矩阵求决策表的属性约简方法^[1]需要构造差别矩阵, 从而大大降低算法的计算效率。在一段时间以来, 人们对它的研究仅集中在它与代数约简和条件信息熵约简的异同及改进上^[2]。近年来, 随着国内外学者对粗糙集属性约简研究的深入, 提出各种不同的约简方法和概念, 尤其对基于粒度计算的属性方法研究, 得到不少新的研究成果, 如基于相对粒度的属性约简、基于相对划分粒度的属性约简、基于知识量的属性约简、基于同可区分度的属性约简等^[3-6]。与基于差别矩阵的约简方法相比, 上述约简方法都是启发式约简算法, 具有较好的计算效率, 但这些方法所得到的约简结果到底是一种什么类型的约简, 并未引起大家的注意和重视。另一方面, 文献[7]提出一种新条件信息熵及基于该条件信息熵的属性约简, 文献[8]对这种新条件信息熵进行研究, 指出基于差别矩阵的 HU 属性约简与基于新条件信息熵约简是等价的。本文在文献[8]的基础上, 通过理论分析, 证明基于相对粒度的属性约简、基于相对划分粒度的属性约简、基于知识量的属性约简、基于同可区分度的属性约简本质上即是一个基于差别矩阵的 HU 属性约简。

2 相关概念

定义 1 决策表 $S = \langle U, V, f, CUD \rangle$ 是一类特殊的信息系统, 其中, U 是一组对象的非空有限集合, 称为论域; C 为有限的条件属性集; D 为有限的决策属性集 $C \cap D = \emptyset$;

$V = \bigcup_{a \in C} V_a$, V_a 为属性 a 的值域; $f: U \times (CUD) \rightarrow V$ 是信息函数, 对 U 上的任意属性集 $B \subseteq CUD$, 定义不可分辨关系 $ind(B) = \{(x, y) \in U^2 \mid \forall a \in B, f(x, a) = f(y, a)\}$, 关系 $ind(B)$ 构成 U 的一个划分, 记作 $U/ind(B)$, 简记为 U/B 。 U/B 中的任何元素 $[x]_B = \{y \mid \forall a \in B, f(x, a) = f(y, a)\}$ 称为等价类。

对于决策表 $S = \langle U, V, f, CUD \rangle$, 文献[1]给出如下差别矩阵定义。

定义 2 给定决策表 $S = \langle U, V, f, CUD \rangle$, 定义 HU 差别矩阵为 $M_H(S) = (m_H(i, j))_{n \times n}$, 其中, 元素 $m_H(i, j)$ 定义如下, $i, j = 1, 2, \dots, n$:

$$m_H(i, j) = \begin{cases} \{a \in C : f(x_i, a) \neq f(x_j, a)\} & (x_i, x_j) \notin ind(D) \\ \emptyset & \text{其他} \end{cases}$$

定义 3 给定决策表 $S = \langle U, V, f, CUD \rangle$, $M_H(S) = (m_H(i, j))_{n \times n}$ 是 HU 差别矩阵, 对任意的 $P \subseteq C$, 如果满足: (1) 对任意的 $\emptyset \neq m_H(i, j) \in M_H$, 则有 $P \cap m_H(i, j) \neq \emptyset$; (2) 对任意的 $b \in P$, $P - \{b\}$ 不满足(1), 则称 P 是 C 相对于 D 的基于差别矩阵的 HU 属性约简, 简称为 HU 属性约简。

基金项目: 广东省自然科学基金资助项目(10452800001004185)

作者简介: 曾凡智(1964—), 男, 副教授、博士, 主研方向: 数据库理论, 粗糙集, 数据挖掘; 黄国顺, 副教授、博士; 文 翰, 讲师、博士研究生

收稿日期: 2011-03-09 **E-mail:** fszsfz@126.com

针对决策表, 梁吉业等提出一种新条件信息熵的计算公式及基于该新条件信息熵的属性约简。

定义 4^[7] 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, $U/P = \{P_1, P_2, L, P_m\}$, $U/D = \{D_1, D_2, L, D_r\}$, 那么 D 相对于 P 的新条件熵 $I(D/P)$ 定义为:

$$I(D/P) = \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|}{|U|} \frac{|P_i \cap D_j^c|}{|U|}$$

其中, D_j^c 表示 D_j 的补集。

定义 5 给定决策表 $S = \langle U, V, f, CUD \rangle$, 对条件属性集 $P \subseteq C$, 如果有 $I(D/P) = I(D/C)$, 且对任意的 $P' \subset P$ 有 $I(D/P') \neq I(D/C)$, 则称 P 为 C 相对于 D 的新条件信息熵属性约简, 简称为新条件信息熵约简。

文献[8]证明了新条件信息熵随条件属性集合的增大而减小, 并证明了 HU 属性约简与新条件信息熵约简是等价的, 即如定理 1。

定理 1 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 那么 P 为 C 相对于 D 的 HU 属性约简当且仅当 P 为 C 相对于 D 的新条件信息熵约简。

为刻画属性区分能力, 文献[9]针对一般信息系统给出属性集的知识粒度概念。

定义 6 给定信息系统 $S = \langle U, V, f, A \rangle$, 对任意属性集 $P \subseteq A$, 记 $U/P = \{P_1, P_2, L, P_m\}$, 则 P 的知识粒度 $GK(P)$ 定义为:

$$GK(P) = \sum_{i=1}^m |P_i|^2 / |U|^2$$

定义 7^[3] 给定信息系统 $S = \langle U, V, f, A \rangle$, 对任意属性集 $P, Q \subseteq A$, 则 Q 关于 P 的相对粒度 $GK_P(Q)$ 定义为:

$$GK_P(Q) = GK(P) - GK(P \cup Q)$$

其中, $GK(P)$ 和 $GK(P \cup Q)$ 分别表示属性集 P 和 $P \cup Q$ 所对应的知识粒度。

定义 8^[10] 给定信息系统 $S = \langle U, V, f, A \rangle$, $P \subseteq A$, 记 $U/P = \{P_1, P_2, L, P_m\}$, 称:

$$E(P) = \begin{cases} \sum_{i=1}^m |P_i| p(P_i) = \sum_{i=1}^m |P_i| \frac{|P_i|}{|U|} & P \neq \emptyset \\ |U| + 1 & P = \emptyset \end{cases}$$

为知识 P 的划分粒度。

划分粒度 $E(P)$ 表示划分 U 的综合颗粒大小。颗粒越大, 对应的知识粒度 $GK(P)$ 越大, 不可区分能力越大, 可区分能力越小; 反之, 如果颗粒越小, 则 $GK(P)$ 越小, 区分能力越大。

定义 9^[10] 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 称:

$$E(P; D) = E(P) - E(P \cup D)$$

为 P 相对 D 的相对划分粒度。

定义 10^[6] 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P, Q \subseteq C$, $U/P = \{P_1, P_2, L, P_m\}$, 则 Q 相对于 P 的知识量 $K_P(Q)$ 定义为:

$$K_P(Q) = \frac{\sum_{i=1}^m \sum_{j=1}^{S_i} |P_i^j| |P_i - P_i^j|}{|U|(|U| - 1)}$$

其中, $P_i^j \in P_i/Q = \{P_i^1, P_i^2, L, P_i^{S_i}\}$; $S_i = |P_i/Q|$ 表示 P_i 在属性集 Q 上进一步细分所得到的等价类个数。

定义 11^[5] 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 记 $U/P = \{P_1, P_2, L, P_m\}$, $U/D = \{D_1, D_2, L, D_r\}$, $U/(P \cup D) = \{H_1, H_2, L, H_k\}$, 则 D 和 P 的可区分度 $|DIS(D; P)|$ 定义为:

$$|DIS(D; P)| = |U|^2 - \sum_{i=1}^m |P_i|^2 - \sum_{j=1}^r |D_j|^2 + \sum_{s=1}^k |H_s|^2$$

3 与 HU 属性约简的等价性证明

给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 那么条件信息熵 $I(D/P)$ 与相对粒度 $GK_P(D)$ 是等价的。

定理 2 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 假设 $U/P = \{P_1, P_2, L, P_m\}$, $U/D = \{D_1, D_2, L, D_r\}$, 那么 $I(D/P) = GK_P(D)$ 。

证明:

$$\begin{aligned} I(D/P) &= \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|}{|U|} \frac{|P_i \cap D_j^c|}{|U|} = \\ &= \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|}{|U|} \frac{|P_i - D_j|}{|U|} = \\ &= \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|}{|U|} \frac{(|P_i| - |P_i \cap D_j|)}{|U|} = \\ &= \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i| |P_i \cap D_j|}{|U|^2} - \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|^2}{|U|^2} \end{aligned}$$

由于 $\sum_{j=1}^r \frac{|P_i \cap D_j|}{|U|} = \frac{|P_i|}{|U|}$, 因此:

$$I(D/P) = \sum_{i=1}^m \frac{|P_i|^2}{|U|^2} - \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|^2}{|U|^2} = GK(P) - GK(P \cup D) = GK_P(D)$$

定理 3 给定决策表 $S = \langle U, V, f, CUD \rangle$, 设 $P \subseteq C$, $U/P = \{P_1, P_2, L, P_m\}$, 则 $E(P; D) = |U| GK_P(D)$ 。

证明:

$$\begin{aligned} E(P; D) &= E(P) - E(P \cup D) = \\ &= |U|(GK(P) - GK(P \cup D)) = |U| GK_P(D) \end{aligned}$$

定理 4 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 假设 $U/P = \{P_1, P_2, L, P_m\}$, $U/D = \{D_1, D_2, L, D_r\}$, 那么 $K_P(D)$ 与 $GK_P(D)$ 有如下关系成立:

$$K_P(D) = \frac{|U|}{|U| - 1} GK_P(D)$$

证明: 由于在 $K_P(D)$ 中, $P_i^j \in P_i/D$ 是 P_i 经决策属性 D 的进一步细分, 因此 P_i 可表示成 $P_i \cap D_j$ 的形式。从而有:

$$\begin{aligned} K_P(D) &= \frac{\sum_{i=1}^m \sum_{j=1}^r |P_i \cap D_j| |P_i - (P_i \cap D_j)|}{|U|(|U| - 1)} = \\ &= \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i| |P_i \cap D_j|}{|U|(|U| - 1)} - \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|^2}{|U|(|U| - 1)} \end{aligned}$$

由于 $\sum_{j=1}^r \frac{|P_i \cap D_j|}{|U|} = \frac{|P_i|}{|U|}$, 因此:

$$\begin{aligned} K_P(D) &= \sum_{i=1}^m \frac{|P_i|}{|U|(|U| - 1)} \sum_{j=1}^r \frac{|P_i \cap D_j|}{|U|} - \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|^2}{|U|(|U| - 1)} = \\ &= \sum_{i=1}^m \frac{|P_i|^2}{|U|(|U| - 1)} - \sum_{i=1}^m \sum_{j=1}^r \frac{|P_i \cap D_j|^2}{|U|(|U| - 1)} = \frac{|U|}{|U| - 1} GK_P(D) \end{aligned}$$

定理 5 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 那么 D 和 P 的可区分度 $|DIS(D; P)|$ 与它们的相对粒度 $GK_P(D)$ 有如下关系成立:

$$|DIS(D; P)| = |U|^2 (1 - GK(D) - GK_P(D))$$

证明: 设 $U/P = \{P_1, P_2, L, P_m\}$, $U/D = \{D_1, D_2, L, D_r\}$, $U/P \cup D = \{H_1, H_2, L, H_k\}$, 由于任意的 $H_s \in U/P \cup D$ 可表示成 $P_i \cap D_j$ 的形式。因此:

$$\begin{aligned}
|DIS(D;P)| &= |U|^2 - \sum_{i=1}^m |P_i|^2 - \sum_{j=1}^r |D_j|^2 + \sum_{s=1}^k |H_s|^2 = \\
&= |U|^2 - \sum_{i=1}^m |P_i|^2 - \sum_{j=1}^r |D_j|^2 + \sum_{i=1}^m \sum_{j=1}^r |P_i \cap D_j|^2 = \\
&= |U|^2 - |U|^2 GK(P) - |U|^2 GK(D) + |U|^2 GK(PUD) = \\
&= |U|^2 (1 - GK(D) - GK_P(D))
\end{aligned}$$

对于给定的决策表, 由于 $|U|$ 和 $GK(D)$ 取非零的固定值, 根据定理 1~定理 5 知, 基于相对粒度、划分粒度、知识量和同可区分度的属性约简与 HU 属性约简是等价的。

定义 12 给定决策表 $S = \langle U, V, f, CUD \rangle$, 对条件属性集 $P \subseteq C$, 如果有 $GK_P(D) = GK_C(D)$, 且对任意的 $P' \subset P$ 有 $GK_{P'}(D) \neq GK_C(D)$, 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。

定义 13 给定决策表 $S = \langle U, V, f, CUD \rangle$, 对条件属性集 $P \subseteq C$, 如果有 $K_P(D) = K_C(D)$, 且对任意的 $P' \subset P$ 有 $K_{P'}(D) \neq K_C(D)$, 则称 P 为 C 相对于 D 的基于相对知识量的属性约简, 简称为相对知识量约简。

定义 14 给定决策表 $S = \langle U, V, f, CUD \rangle$, 对条件属性集 $P \subseteq C$, 如果有 $E(P;D) = E(C;D)$, 且对任意的 $P' \subset P$ 有 $E(P';D) \neq E(C;D)$, 则称 P 为 C 相对于 D 的基于相对划分粒度属性约简, 简称为相对划分粒度约简。

定义 15 给定决策表 $S = \langle U, V, f, CUD \rangle$, 对条件属性集 $P \subseteq C$, 如果有 $|DIS(D;P)| = |DIS(D;C)|$, 且对任意的 $P' \subset P$ 有 $|DIS(D;P')| \neq |DIS(D;C)|$, 则称 P 为 C 相对于 D 的基于同可区分度属性约简, 简称为同可区分度约简。

定理 6 给定决策表 $S = \langle U, V, f, CUD \rangle$, $P \subseteq C$, 那么 P 是 C 的 HU 属性约简当且仅当 P 为 C 相对于 D 的相对粒度(相对知识量或相对划分粒度或同可区分度)约简。

证明: 根据定理 1, P 是 C 的 HU 属性约简当且仅当 P 为 C 相对于 D 的新条件信息熵约简, 即 $I(D/P) = I(D/C)$, 且对任意的 $P' \subset P$ 有 $I(D/P') \neq I(D/C)$, 又因定理 2 及定义 12 知, P 是 C 的 HU 属性约简当且仅当 P 是 C 相对于 D 的相对粒度约简。

类似地, 根据定理 1 到定理 5 知, P 是 C 的 HU 属性约简当且仅当 P 是 C 相对于 D 的相对知识量或相对划分粒度或同可区分度约简。

4 数值算例

当决策表一致时, 现有研究结论表明, 基于差别矩阵的 HU 属性约简与代数约简是等价的, 但当决策表不一致时, HU 属性约简与代数约简有时并不相同。下面构造一个不一致决策表, 分别用 HU 差别矩阵、代数约简方法和文献[3-6]的方法计算其属性约简。

例 1 给定决策表 S_1 如表 1 所示, 其中, $U = \{x_1, x_2, \dots, x_{10}\}$; $C = \{a, b, c, e, f\}$; $D = \{d\}$ 。

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

根据定义 3 知, 其 HU 属性约简为 $\{a, c, e, b\}$ 和 $\{a, c, e, g\}$;

另外, 可计算到其代数约简分别为 $\{a, e, b\}$ 和 $\{a, e, g\}$, 可见 HU 属性约简与代数约简有时确实不相同。下面分别用文献[2-5]中的方法计算其属性约简。限于篇幅只选取部分数据罗列如下:

令 $P = \{a, c, e, b\}$, 因为 $GK_P(D) = GK_C(D) = 6/100$, 且可验证对任意的 $P' \subset P$ 有 $GK_{P'}(D) \neq GK_C(D)$, 所以 $P = \{a, c, e, b\}$ 也是决策表的一个相对粒度约简。

类似地, 因为 $|DIS(D;P)| = |DIS(D;C)| = 56$ 且可验证对任意的 $P' \subset P$ 有 $|DIS(D;P')| \neq |DIS(D;C)|$, 所以 $P = \{a, c, e, b\}$ 是决策表的一个同可区分度约简。

由于 $E(P;D) = E(C;D) = 6/10$, 并且可验证对任意的 $P' \subset P$ 有 $E(P';D) \neq E(C;D)$, 因此 $P = \{a, c, e, b\}$ 是决策表的一个相对划分粒度约简。

由于 $K_P(D) = K_C(D) = 6/90$ 并且可验证对任意的 $P' \subset P$ 有 $K_{P'}(D) \neq K_C(D)$, 因此 $P = \{a, c, e, b\}$ 是决策表的一个基于相对知识量的属性约简。

如果利用文献[3]的启发式约简算法搜索约简结果, 求到的仍是一个 HU 属性约简。先计算其核属性集, 得到核属性集为 $C_0 = \{a, e, c\}$, 此时因 $GK_{C_0}(D) = 18/100 \neq GK_C(D)$, 进一步添加属性, 分别计算添加属性 b, f, g 后的相对粒度, 得到 $GK_{C_0 \cup \{b\}}(D) = 6/100$, $GK_{C_0 \cup \{f\}}(D) = 10/100$, $GK_{C_0 \cup \{g\}}(D) = 6/100$, 由于有 2 个属性同时达到最小值, 任取一个, 如选取属性 b , 令 $C_0 = \{a, e, c, b\}$, 因 $GK_{C_0}(D) = GK_C(D)$, 最后得到约简结果 $C_0 = \{a, e, c, b\}$, 是其中的一个 HU 属性约简, 若选取属性 g , 得到另一个 HU 属性约简 $\{a, c, e, b\}$ 。

5 结束语

由于基于差别矩阵的属性约简方法需要构造差别矩阵, 从而降低算法的计算效率。如何快速地计算出属性约简是基于粗糙集面向实际应用必须解决的问题。基于启发式信息的约简方法能提高算法效率, 采用不同的启发式信息, 得到的约简结果一般也不同。对这些采用不同启发式信息所得的“不同”约简结果是什么类型的约简也必须澄清和予以重视。尽管在具体的判断指标上, 基于相对粒度的属性约简、基于相对划分粒度的属性约简、基于知识量的属性约简和基于同可区分度的属性约简方法各不相同, 但本文结论表明, 它们得到的约简结果都是一个基于差别矩阵的 HU 属性约简。本文结论说明了 HU 属性约简与知识量、粒度计算及同可区分度约简方法之间存在联系, 具有一定的理论意义。

参考文献

- [1] Hu Xiaohua, Nick C. Learning in Relational Databases: A Rough Set Approach[J]. International Journal of Computation Intelligence, 1995, 11(2): 323-338.
- [2] 王加阳, 高 灿. 改进的基于差别矩阵的属性约简算法[J]. 计算机工程, 2009, 35(3): 66-67.
- [3] 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.
- [4] 冯琴荣, 苗谦谦, 程 映. 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.
- [5] 腾书华, 德 才, 孙即祥, 等. 基于同可区分度的属性约简算法[J]. 模式识别与人工智能, 2010, 23(5): 630-638.
- [6] 梁春华, 张海云. 基于知识量的决策表约简算法[J]. 山西农业大学学报: 自然科学版, 2007, 27(2): 214-217.

(下转第 71 页)