

基于区分函数与强等价集理论的属性约简算法^{*}

陶 志^{1,2}, 许宝栋¹, 赵春元³

(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110004;

2. 沈阳航空工业学院, 辽宁 沈阳 110034;

3. 沈阳工程学院, 辽宁 沈阳 110036)

摘 要: 提出了基于区分函数与强等价集理论的属性相对约简算法。在决策表的相对约简过程中采用区分矩阵来表达知识, 并在区分函数的化简过程中引进强等价集的概念, 通过去除强等价集的方法, 使约简效率提高。现在已经证明, 寻找决策表最小相对约简是典型的 NP-hard 问题, 采用本文所提供的算法可降低问题复杂度, 同时又可节省大量的计算存储空间。通过实例分析, 证明该算法是求解全部属性相对约简的快速有效的方法。

关键词: 粗糙集理论; 决策表; 相对约简; 区分函数; 强等价集

中图分类号: TP18 文献标识码: A 文章编号: 0529-6579 (2004) 02-0013-04

粗糙集理论是由波兰学者 Pawlak 于 1982 年提出的^[1], 它作为一种刻划具有不完整性和不确定性信息的全新数学工具, 已经成为人工智能领域的一个新的学术热点。其主要思想是, 在保持知识库的分类能力不变的前提下, 通过属性(知识)约简, 导出问题的决策或分类规则^[2]。目前, 粗糙集理论已被成功地应用于机器学习、过程控制、决策分析、模式识别与数据挖掘等领域^[3,4]。属性约简是粗糙集理论中的核心内容之一。所谓属性约简, 就是在保持知识库分类能力不变的条件下, 删除不相关或不重要的冗余属性。特别是, 当知识库中的数据是随机采集时, 其属性的冗余性尤为普遍, 属性约简就显得更为重要。

一般来讲, 一个知识库中的属性约简不是惟一的, 计算属性约简的复杂性是随着决策表的增大呈指数增长的, 是一个典型的 NP 完全问题^[5]。当然, 实际问题中没有必要求出所有约简, 通常人们关心的是寻找知识库中所含条件属性最少的约简, 即最小约简。

本文首先给出了决策表中属性相对约简的概念, 并采用构造区分矩阵与区分函数的方法求取最小属性相对约简。由于约简算法的效率主要依赖于区分函数的化简, 计算的复杂程度也体现在对区分函数的化简过程, 因此本文运用了强等价集替代的方法, 此方法可使计算复杂性大幅度降低, 同时也大大提高了约简效率。实际计算结果表明, 该方法可节省大量计算存储空间, 在现有计算机硬件环境

下会使求解问题的规模增大。

1 决策表与属性的相对约简

定义 1 四元组 $S = (U, A, V, f)$ 是一个决策表, 其中 U 表示对象的非空有限集合, 称为论域; $A = C \cup D$, $C \cap D = \emptyset$, C 称为条件属性集, D 称为决策属性集; $V = \bigcup_{a \in A} V_a$, V_a 是属性 a 的值域; f 表示 $U \times A \rightarrow V$ 是一个信息函数, 它为每个对象的每个属性赋予一个信息值, 即 $\forall a \in A, x \in U, f(x, a) \in V_a$ 。每个属性子集 $B \subseteq A$ 决定了一个二元不可区分关系^[6] $\text{IND}(B)$, 即:

$$\text{IND}(B) = \{(x, y) \in U \times U \mid \forall a \in B, f(x, a) = f(y, a)\}$$

关系 $\text{IND}(B)$ 构成了 U 的一个划分, 用 $U/\text{IND}(B)$ 表示, 简记为 U/B 。 $U/\text{IND}(B)$ 中的任何元素 $[x]_B$ 称为等价类。 $U/\text{IND}(B)$ 表示知识表达系统 $S = (U, A, V, f)$ 中与 B 相关的知识, 它包含知识 B 的基本模块。 $\text{IND}(B)$ 的补集对应一个 B 的二元可区分关系 $\text{Dis}(B)$, 即:

$$\text{Dis}(B) = U \times U - \text{IND}(B)$$

定义 2 设 $S = (U, A, V, f)$ 是一个决策表, 对于每个子集 $X \subseteq U$ 和一个属性子集 $B \subseteq A$, 称子集 $\underline{B}X = \{x \in U \mid [x]_B \subseteq X\}$ 为 X 的 B 下近似集。 B 下近似集也称作 X 的 B 正域, 记作 $\text{pos}_B(X)$, 即 $\text{pos}_B(X) = \underline{B}X$ 。

* 收稿日期: 2003-04-02

基金项目: 国家自然科学基金资助项目(70171056); 国家重点科技攻关资助项目(975620107)

作者简介: 陶志(1963年生), 男, 博士生, 副教授, E-mail: t86543213@163.com

不同的属性子集可能确定不同的分类。在应用中,一个分类对另一个分类的关系十分重要,因此我们给出相对正域和相对约简的概念。

定义3 设 $S = (U, A, V, f)$ 是踺个决策表, $P, Q \subseteq A$, 则称子集 $\text{pos}_P(Q) = \bigcup_{X \in U/Q} PX$ 为 Q 的 P 正域。

定义4 设 $S = (U, A, V, f)$ 是一个决策表, $P, Q \subseteq A, R \in P$, 若 $\text{pos}_P(Q) = \text{pos}_{(P-\{R\})}(Q)$, 则称 R 为 P 中 Q 不必要的; 否则 R 为 P 中 Q 必要的。

不必要的属性在决策系统中是多余的, 若将它从系统中去掉, 不会改变系统分类能力。

定义5 设 $S = (U, A, V, f)$ 是一个决策表, $P, Q \subseteq A$, 若每个 $R \in P$ 在 P 中都是 Q 必要的, 则 P 为 Q 独立的; 否则, P 为 Q 依赖的。

对于相依赖的属性来说, 其中包含有多余的属性, 可以对其约简。

定义6 设 $S = (U, A, V, f)$ 是一个决策表, $P, Q \subseteq A, K \subseteq P$, 如果 (1) $\text{pos}_K(Q) = \text{pos}_P(Q)$; (2) K 是 Q 独立的, 则称 K 是 P 的一个 Q 约简。 P 的 Q 约简称为相对约简; 显然, 相对约简不是惟一的。

2 区分矩阵与区分函数

区分矩阵是华沙大学数学家 Skowron 首先提出的, 它为求取属性约简提供了很好的基础。

定义7 决策表 $S = (U, C \cup \{d\}, V, f)$ 的区分矩阵是一个对称的 $|U| \times |U|$ 阵, 矩阵的每一项 c_{ij} 表示为

$$c_{ij} = \{a \in C \mid f(x_i, a) \neq f(x_j, a)\}$$

若 $f(x_i, d) \neq f(x_j, d)$; 否则 $c_{ij} = \phi$ (1)

可以证明, 如果 $C' \subseteq C$ 是满足条件 $C' \cap c_{ij} \neq \phi, \forall c_{ij} \neq \phi$ 的极小子集, 则 C' 是 C 的 D 约简^[7]。为求属性相对约简, 先引入一个布尔函数 χ_B 如下:

$$\chi_B(C') = \begin{cases} 1 & \text{若 } B \cap C' \neq \phi \\ 0 & \text{若 } B \cap C' = \phi \end{cases}$$

由此构造决策表 S 的区分函数 f_A 如下:

$$f_A(C') = \left[\bigwedge_{1 \leq i, j \leq n, c_{ij} \neq \phi} \chi_{c_{ij}} \right](C') = \prod_{1 \leq i, j \leq n, c_{ij} \neq \phi} \chi_{c_{ij}}(C')$$

这里, c_{ij} 代表 S 的区分矩阵第 i 行第 j 列元素, 也称为区分函数 f_A 的索引集。

性质 1^[8] (吸收律) 设 $\phi \neq C \subseteq D \subseteq A, B \subseteq A$, 若 $\chi_C(B) = 1$, 则 $\chi_D(B) = 1$ 。

性质 2^[8] (扩展律) 设 $f_A = \chi_{C_1} \wedge \dots \wedge \chi_{C_k} \wedge \chi_{C_{k+1}} \wedge \dots \wedge \chi_{C_s}$, 并对 $a \in A$ 有 $\chi_{C_i}(\{a\}) = 1, i = 1, \dots, k$ 和 $\chi_{C_i}(\{a\}) = 0, i = k+1, \dots, s$ 成立, 则

$$f_A = f_1 \vee f_2 = (\chi_{\{a\}} \wedge (\chi_{C_{k+1}} \wedge \dots \wedge \chi_{C_s})) \vee$$

$$((\chi_{C_1-\{a\}} \wedge \dots \wedge \chi_{C_k-\{a\}}) \wedge (\chi_{C_{k+1}} \wedge \dots \wedge \chi_{C_s}))$$

性质 2 说明区分函数 f_A 可以分解成一些布尔函数 f_i 的析取, 而 f_i 是由一些布尔函数 χ 的合取构成的。

定义8 如果区分函数 f_A 有如下分解式 $f_A = f_1 \vee \dots \vee f_k$, 且对每个 f_i 均有 $f_i = \chi_{C_1^i} \wedge \dots \wedge \chi_{C_{k_i}^i}$ 。这里 $C_l^i \cap C_m^i = \phi (l \neq m)$, 则称布尔函数集 $\{f_1, f_2, \dots, f_k\}$ 是区分函数 f_A 的一个简单覆盖。

反复应用性质 2, 容易证明, 任何区分函数 f_A 均可分解成它的一个简单覆盖的析取范式。关于区分函数 f_A 与属性约简之间的关系, 有如下定理。

定理 1^[8] 若 f_A 是一个区分函数, 则 B 是 A 的一个约简的充分必要条件是: ① $f_A(B) = 1$; ② $C \subset B \Rightarrow f_A(C) = 0$ 。

定理 1 从区分函数的角度提供了条件属性对决策属性相对约简的定义, 同时为以下定义提供了理论依据。

定义 9 设 $\{f_1, f_2, \dots, f_k\}$ 是区分函数 f_A 的一个简单覆盖, 如果子集 $B \subseteq A$ 满足: ① $f_i(B) = 1$; ② $C \subset B \Rightarrow f_i(C) = 0$, 则称 B 是 f_i 的一个约简。 f_i 的所有约简的全体记作 $\text{Red}(f_i)$, 这里 $B \in \text{Red}(f_i)$ 。显然, 如果 B 是 f_i 的一个约简, 则 B 必是 A 的一个约简。

在实际应用中, 可首先由性质 1 和性质 2 将 f_A 分解成一个简单覆盖的析取范式, 然后再根据定理 1 所给的条件求解每个 f_i 的相对约简, 进而可求得整个决策表的所有相对约简。

3 强等价集与属性相对约简算法

定义 10 若 $a \in A$, 则称子集 $[a] = \{b \in A \mid \text{Dis}(\{a\}) = \text{Dis}(\{b\})\}$ 为 a 的等价类。

定义 11 设 $a \in A$, 如果存在某一集合 $B \subseteq [a]$, 则称集合 B 是属性 a 的强等价集。

显然, 强等价集有如下属性:

$$\text{Dis}(B) = \text{Dis}(\bigcup_{b \in B} \{b\}) = \bigcup_{b \in B} \text{Dis}(\{b\}) = \text{Dis}(\{a\})$$

因此, 强等价集所提供的论域对象特征与单个属性所提供的对象特征相同, 这也说明与强等价集有关联的知识是冗余的知识, 可以对其约简。本文在约简过程中用单个属性标识符代替强等价集, 因此在实际计算中大量降低了内存的开销, 同时也减小了计算的复杂程度。

由定义 11 容易得出如下定理 2。

定理 2 A 的子集 $B \subseteq A$ 是强等价集的充分必要条件是： B 被区分矩阵中 2 个或 2 个以上项同时包含，且 B 与区分矩阵中其它项的交为空集。

下面证明关于强等价集的另一个结论。

定理 3 如果 $\{f_1, f_2, \dots, f_k\}$ 是区分函数 f_A 的一个简单覆盖， B 是强等价集， C 是 f_i 的一个约简，且 $\{a, b\} \in B$ ，则 $\{a, b\} \not\subset C \in \text{Red}(f_i)$ 。

证明(反证法) 假定 C 是 f_i 的一个约简，且有 $\{a, b\} \subset C \in \text{Red}(f_i)$ 。因为 $\{a, b\} \in B$ ，且 B 是强等价集，所以有 $\text{Dis}(\{a\}) = \text{Dis}(\{b\})$ ，于是得 $\text{Dis}(C - \{a\}) = \text{Dis}(C)$ ，即 $\text{Ind}(C - \{a\}) = \text{Ind}(C)$ ，这与 C 是 f_i 的一个约简相矛盾。因此应有 $\{a, b\} \not\subset C \in \text{Red}(f_i)$ 成立。

定理 3 表明， f_i 的任何一个约简至多只能包含强等价集中的一个属性，此结论进一步说明强等价集所含的属性是可以约简的，也为下面区分函数化简过程中强等价集的可替代性提供了理论基础。

由于区分矩阵是对称阵，所以在实际计算中只要求出下三角部分即可。若假定原始区分函数已经给定： $f_A = \prod_{1 \leq i, j \leq n, c_{ij} \neq \emptyset} \chi_{c_{ij}}$ ，这里 c_{ij} 代表下三角区分矩阵中第 i 行第 j 列元素，则求属性相对约简的算法步骤如下。

步骤 1 首先对每一个 f_i 由性质 1，若存在索引集 $C_i^j \subseteq C_m^j$ ，则在 f_i 中去除所有 $\chi_{c_m^j}$ 。

步骤 2 由定理 2 在 f_i 中找出所有强等价集，并取单独属性标识符替换每个强等价集。

步骤 3 对每个 f_i ，查找在索引集 $C_j^i (j = 1, \dots, k_i)$ 中出现频数最高的属性(要求频数 ≥ 2)，然后应用性质 2 将 f_i 分解成 $f_i = f_{i1} \vee f_{i2}$ 。

步骤 4 重复执行步骤 1—3 直至将 f_A 分解成一个简单覆盖时为止。

步骤 5 对 f_A 简单覆盖中的每个 f_i ，再用强等价集替换回原相应的属性标识，即用 χ_c 替换 χ_C ，其中 $C = \bigcup \{[a] : a \in C\}$ 。并在 f_i 中查找是否存在某一索引集 $C_j^i \subset C$ ，如果是则去除相应的 χ_C 。

步骤 6 计算 f_i 的所有约简 $\text{Red}(f_i)$ 。

步骤 7 在 $\text{Red}(f_i) i = 1, \dots, k$ 中，选择所含属性最少的约简作为最小相对约简，其中 $f_A = f_1 \vee \dots \vee f_k$ 。

4 算 例

例 1 给定一个决策表 $S = (U, A, V, f)$ ，其中 $A = C \cup D, C \cap D = \emptyset, U = \{x_1, x_2, \dots, x_6\}, C =$

$\{a, b, c, d, e, f\}, D = \{g\}$ ，信息函数 f 由表 1 给出。求该决策表的最小相对约简。

表 1 决策表
Tah 1 Decision table

U	条件属性						决策属性
	a	b	c	d	e	f	g
x_1	1	2	0	2	1	1	0
x_2	1	1	0	2	2	1	1
x_3	0	1	1	1	2	0	2
x_4	0	0	1	1	0	0	1
x_5	1	2	1	1	0	1	2
x_6	2	1	2	1	0	2	2

表 2 区分矩阵
Tah 2 Discernibility matrix

	x_1	x_2	x_3	x_4	x_5	x_6
x_1						
x_2	b, e					
x_3	a, b, c, d, e, f	a, c, d, f				
x_4	a, b, c, d, e, f		b, e			
x_5	c, d, e	b, c, d, e		a, b, f		
x_6	a, b, c, d, e, f	a, c, d, e, f		a, b, c, f		

由 (1) 式构造决策表 1 的区分矩阵，如表 2 所示。表 2 对应的区分函数为： $f_A = \{b, e\} \wedge \{a, b, c, d, e, f\} \wedge \{a, b, c, d, e, f\} \wedge \{c, d, e\} \wedge \{a, b, c, d, e, f\} \wedge \{a, c, d, f\} \wedge \{b, c, d, e\} \wedge \{a, c, d, e, f\} \wedge \{b, e\} \wedge \{a, b, f\} \wedge \{a, b, c, f\}$ 。

(为方便计算上式没有明确标注 χ)

(1) 因为 $\{b, e\} \subset \{a, b, c, d, e, f\}, \{b, e\} \subset \{b, c, d, e\}, \{c, d, e\} \subset \{a, c, d, e, f\}, \{a, b, f\} \subset \{a, b, c, f\}$ ，所以由性质 1，在 f_A 中去除第 2, 3, 5, 7, 8, 9, 11 项，得等价的区分函数：
 $f_A = \{b, e\} \wedge \{c, d, e\} \wedge \{a, c, d, f\} \wedge \{a, b, f\}$

(2) 由定理 2 可知 $\{a, f\}$ 是强等价集，所以可在 f_A 中由一个单独的属性标识符 h 替代 $\{a, f\}$ ，于是区分函数变为

$f_A = \{b, e\} \wedge \{c, d, e\} \wedge \{h, c, d\} \wedge \{h, b\}$

(3) 属性 e 是出现频数最大的属性，于是由性质 2 可将 f_A 分解为：

$f_A = f_1 \vee f_2 = (\{e\} \wedge \{h, c, d\} \wedge \{h, b\}) \vee (\{b\} \wedge \{c, d\} \wedge \{h, c, d\} \wedge \{h, b\})$
在 f_2 中有 $\{b\} \subset \{h, b\}, \{c, d\} \subset \{h, c, d\}$ ，应用性质 1 可将 f_A 继续化简为

$f_A = f_1 \vee f_2 = (\{e\} \wedge \{h, c, d\} \wedge \{h, b\}) \vee (\{b\} \wedge \{c, d\})$

(4) 对 f_1, f_2 继续应用性质 2 化简为: $f_A = f_{11} \vee f_{12} \vee f_{21} = (\{h\} \wedge \{e\}) \vee (\{c, d\} \wedge \{b\} \wedge \{e\}) \vee (\{b\} \wedge \{c, d\})$, 此时的 f_A 即是一个简单覆盖 $\{f_{11}, f_{12}, f_{21}\}$ 的析取范式。

(5) 用强等价集替代 f_{ij} 中与其等价的单个属性标识。即有:

$$f_A = f_{11} \vee f_{12} \vee f_{21} = (\{a, f\} \wedge \{e\}) \vee (\{c, d\} \wedge \{b\} \wedge \{e\}) \vee (\{b\} \wedge \{c, d\})$$

(6) 求所有 f_{ij} 的约简 $\text{Red}(f_{ij})$ 。即有:

$$\text{Red}(f_{11}) = \{\{a, e\}, \{e, f\}\},$$

$$\text{Red}(f_{12}) = \{\{c, b, e\}, \{d, b, e\}\},$$

$$\text{Red}(f_{21}) = \{\{b, c\}, \{b, d\}\}$$

(7) 将第 (6) 步计算结果求并, 即:

$$\bigcup_{i,j} \text{Red}(f_{ij}) =$$

$$\{\{a, e\}, \{e, f\}, \{c, b, e\}, \{d, b, e\}, \{b, c\}, \{b, d\}\}$$

因此, 据以上计算结果所求最小相对约简应为:

$$\text{Red}(A) = \{\{a, e\}, \{e, f\}, \{b, c\}, \{b, d\}\}$$

此结果恰是本问题的最优解。

5 结 语

本文给出了区分函数和强等价集的概念, 并对它们的特性进行了深入讨论, 根据强等价集与单个属性具有相同分类能力这一特征, 在区分函数的化简过程中采用了去除强等价集的方法, 这一方法使化简的时间大量缩短, 也节省了大量存储空间, 算法效率明显提高。但是, 在一些实际问题中也有不

含强等价集的情况, 应用此算法反而会增加部分计算量。大量实验结果表明, 出现强等价集的频率远大于不含强等价集的情况, 因此从统计学观点来看, 此算法是提高约简效率的实际有效方法。

参考文献:

- [1] PAWLAK Z. Rough sets-theoretical aspects of reasoning about data[M] .Dordrecht:Kluwer Academic Publishers 1991: 9—30.
- [2] 韩祯祥, 张琦, 文福栓. 粗糙集理论及其应用综述[J] .控制理论与应用, 1999, 16(2): 153—157.
- [3] PAWLAK Z. Rough set theory and its application to data analysis[J] . Cybernetics and Systems 1998, 29(9): 661—668.
- [4] HU X H. Mining knowledge rules from databases—a rough set approach[A] . Proceedings of IEEE International Conference on Data Engineering[C] . Los Alamitos: IEEE Computer Society Press 1996: 96—105.
- [5] WANG S K M, ZIARKO W. On optimal decision rules in decision tables[J] . Bulletin of Polish Academy of Sciences 1985, 33(6): 663—676.
- [6] DUNTSCH I, GEDIGA G. Statistical evaluation of rough set dependency analysis[J] . Intemational J Human-Computer Study, 1997, 46(5): 589—604.
- [7] 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.
- [8] STARZYK J A. A mathematical foundation for improved reduction generation in information systems[J] . Knowledge and Information Systems, 2000, 2: 131—146.

Attributes Relative Reduction Approach Based on Discernibility Function and Strong Compressible Set Theory

TAO Zhi^{1,2}, XU Bao dong¹, ZHAO Chun yuan³

(1. School of Information Science & Engineering, Northeastern University, Shenyang 110004, China;

2. Shenyang Institute of Aeronautical Engineering, Shenyang 110034, China;

3. Shenyang Institute of Engineering, Shenyang 110036, China)

Abstract: A kind of attributes relative reduction approach based on discernibility function and strong compressible set is proposed. In the process of relative reduction of decision table, knowledge is expressed by discernibility matrix. The concept of strong compressible set is introduced when discernibility function reduced. Efficiency of the reduction is improved by means of eliminating strong compressible set. It is shown that finding the minimal reduction of a decision table is an NP hard problem. By the approach of this paper, the complicated degree of the problem could be reduced and a great deal of memory space is saved. The practical results show that the approach is quick and effective in solving all reduction of decision table.

Key words: rough set theory; decision table; relative reduction; discernibility function; strong compressible set