

一种基于条件熵的增量式属性约简算法

刘 薇 梁吉业 魏 巍 钱宇华

(山西大学计算机与信息技术学院计算机智能与中文信息处理教育部重点实验室 太原 030006)

摘 要 粗糙集是一种处理不确定、不完全知识的数学工具,属性约简是粗糙集理论的重要研究内容之一。提出了一种基于条件熵的快速增量约简方法,主要分析了在对象动态增加情况下信息熵的变化机制。该算法通过判断更新前决策表的约简属性对新增对象的区分情况来计算新的条件熵值,可以快速求解出更新后的决策表的属性约简结果。实验结果也进一步验证了该方法的有效性。

关键词 条件熵,增量式,属性约简,决策表

中图法分类号 TP18 **文献标识码** A

Incremental Algorithm for Attribute Reduction Based on Conditional Entropy

LIU Wei LIANG Ji-ye WEI Wei QIAN Yu-hua

(Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing, School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract Rough set theory is a mathematic tool to deal with incomplete and uncertain information, in which attribute reduction is one of important issues. The changing mechanism of condition entropy was analyzed when a new object was added to the original decision table. Based on this mechanism, a new incremental algorithm for attribute reduction was proposed. In this algorithm we divided the added objects into three cases. Furthermore, by these different cases incremental attribute reduces could be calculated quickly. At last, the validity of the proposed algorithm was depicted by an experiment.

Keywords Conditional entropy, Incremental, Attribute reduction, Decision table

1 引言

波兰数学家 Pawlak 在 20 世纪 80 年代初提出了粗糙集 (Rough Set, RS) 是一种处理不精确、不相容和不完全等各种不完备信息的数学理论^[1]。在粗糙集理论中,属性约简是粗糙集应用的主要研究内容之一,也是挖掘和简化知识的关键步骤^[2,3]。因此,近年来属性约简得到了研究人员的广泛关注,并且取得了较为瞩目的研究成果。

在现实生活中,数据总是保持不断增长的趋势,这就要求我们要将数据原来包含的信息与新产生的信息相结合,以保证知识发现过程中的实时性和有效性。但是,在此过程中对原有信息的计算将会产生大量的冗余。同样,在计算增量式更新的决策表的属性约简时,如果通过反复计算更新后决策表的划分来求解属性约简结果将会带来计算时间的冗余。为此,在不计算原有约简的基础上对增量决策表的属性约简进行实时有效的更新,具有重要的研究意义和应用价值。现有的属性约简大体上可以分为基于差别矩阵或在此基础上改进的属性约简算法^[4]、基于正域的属性约简算法^[5,6]以及基于启发式信息的属性约简算法^[7]。但这些都只针对静态的信息

系统或者决策表。针对增量式的属性约简算法,文献[8]提出一种基于邻域差别矩阵的属性约简增量式算法,主要考虑连续型数据对象动态增加的情况下的属性约简问题。文献[9]提出了基于正域的属性约简的增量式算法,其提高了属性约简的更新效率,但对对象增加引起的不一致情况还需改进。文献[10]提出一种基于差别矩阵的属性约简的增量式更新算法。以上算法都是通过动态更新差别矩阵,从而得到增加新对象的属性核,进而获得新的属性约简。

本文提出一种基于条件熵的属性约简更新方法,该方法利用文献[11]提出的决策表的决策属性关于条件属性的条件熵变化原理,分析了加入新对象后决策表条件熵的变化对其属性约简的影响,将属性约简的更新分为 3 种情况研究,并提出了相应的求解方法。该算法充分利用了更新前决策表的信息熵和属性约简等信息,在一定程度上缩短了约简的更新时间,保证了属性约简的实时性。

2 基本概念

定义 1^[3] 五元组 $S=(U, CUD, V, f)$ 是一个决策表,其中 $U=\{x_1, x_2, \dots, x_n\}$ 表示对象的非空有限集合,称为论域; C

到稿日期:2010-03-09 返修日期:2010-06-18 本文受国家自然科学基金(No. 60773133, 70971080, 60903110), 山西省自然科学基金(No. 2008011038, 2009021017-1)资助。

刘 薇(1985—),女,硕士生,主要研究方向为粗糙集理论、智能信息处理, E-mail: sdcsluwei@gmail.com; 梁吉业(1962—),男,教授,博士生导师,主要研究方向为粗糙集理论、数据挖掘与知识发现; 魏 巍(1980—),男,博士生,主要研究方向为智能决策、粗糙集理论; 钱宇华(1976—),男,博士生,讲师,主要研究方向为数据挖掘、粒度计算。

表示条件属性的非空有限集合; D 表示决策属性的非空有限集合且 $C \cap D = \Phi$, $V = \bigcup_{a \in C \cup D} V_a$; V_a 是属性 a 的值域, $f: U \times C \cup D \rightarrow V$ 是一个信息函数, 它对一个对象的每一个属性赋予一个信息值, 即 $\forall a \in C \cup D, x \in U$, 有 $f(x, a) \in V_a$; 每一个属性子集 $P \subseteq (C \cup D)$ 决定了一个二元不可区分关系 $IND(P)$: $IND(P) = \{(x, y) \in U \times U \mid \forall a \in P, f(x, a) = f(y, a)\}$ 。关系 $IND(P)$ 构成了 U 的一个划分, $U/IND(P)$ 简记为 U/P , U/P 中的任何元素 $[x]_P = \{y \mid \forall a \in P, f(x, a) = f(y, a)\}$ 都称为等价类。

令 $B \subseteq C$, 有 $C \leq B$, 表示对任意 $X \in U/B$, 存在 $Y \in U/C$, 使得 $X \supseteq Y$, 这意味着 C 的划分比 B 更精细。 $C < B$ 表示存在 $X_0 \in U/B, Y_0 \in U/C, X_0 \not\supseteq Y_0$ 。

定义 2^[12,13] 给定决策表 $S = (U, C \cup D, V, f)$, C 是条件属性, D 是决策属性集, $U/C = \{X_1, X_2, \dots, X_m\}$, $U/D = \{Y_1, Y_2, \dots, Y_n\}$ 。决策表的条件熵定义为

$$E_U(D|C) = \sum_{i=1}^m \sum_{j=1}^n \frac{|Y_i \cap X_j|}{|U|} \frac{|Y_i^c - X_j^c|}{|U|}$$

式中, $E_U(D|C)$ 的下标 U 表示论域, X_j^c 表示 X_j 的补集, Y_i^c 表示 Y_i 的补集。

定理 1^[3] 设 $B \subseteq C$, 若 $E_U(D|B) = E_U(D|C)$, 且不存在 $R \subseteq B$, 使得 $E_U(D|R) = E_U(D|C)$, 则称 B 为 C 的一个相对于决策属性 D 的属性约简。

定理 2^[3] 给定决策表 $S = (U, C \cup D, V, f)$, $B < C$, 则有 $E(D|C) < E(D|B)$ 。

定义 3^[3] 给定决策表 $S = (U, C \cup D, V, f)$, $B \subseteq C$, 对于 $\forall a \in C - B$, a 相对于 B 的重要度定义为 $sig(a, B, D) = E(D|B) - E(D|B \cup \{a\})$ 。

定义 4^[10] 若 $B \subseteq C$, 且两个不同的对象 x 和 y , 对 $\forall a \in B$ 满足 $f(x, a) = f(y, a)$, $f(x, D) \neq f(y, D)$, 则称 x 和 y 为 B 不一致的。

3 新增对象后条件熵的变化机制

增加新的对象后, 决策表的条件熵将会发生相应变化, 这可能会引起属性约简的变化。本节主要分析加入新对象后决策表的条件熵的变化, 为研究属性约简的更新方法提供理论支撑。

定义 5 给定决策表 $S = (U, C \cup D, V, f)$, C 为条件属性集, D 为决策属性集, $B \subseteq C$ 。新对象 x 加入决策表后, 对于 $\forall y \in U, \forall a \in B, f(a, x) \neq f(a, y)$, 则称属性集 B 可以区分新对象 x 。

对于给定的决策表 $S = (U, C \cup D, V, f)$, 新对象 x 加入后, 对原约简属性集的区分情况的影响可以分为以下 3 种:

(1) 原约简属性集 B 可以区分新对象, 条件属性集 C 也可以区分新对象;

(2) 原约简属性集 B 不可以区分新对象, 条件属性集 C 可以区分新对象;

(3) 原约简属性集 B 不可以区分新对象, 条件属性集 C 也不可以区分新对象。

对情况 1 和情况 2, 由定义 4 可知, 它们分别为 B 一致和 C 一致的。而对于情况 3, 所有的条件属性集都不能区分新对象 x , 所以它是不一致的。

定理 3^[11] 给定决策表 $S = (U, C \cup D, V, f)$, C 为条件属

性集, D 为决策属性集, $U/C = \{X_1, X_2, \dots, X_m\}$, $U/D = \{Y_1, Y_2, \dots, Y_n\}$ 。增加一个对象 x 后, $x \in X_q'$ 且 $x \in Y_p'$ (X_q' 是 $(U \cup \{x\})/C$ 中的条件类, Y_p' 是 $(U \cup \{x\})/D$ 中的决策类), 则新的条件熵为

$$E_{U \cup \{x\}}(D|C) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(D|C) + 2|X_q' - Y_p'|)$$

这两种情况均可以通过以上公式证明条件熵的变化情况。

定理 4 给定决策表 $S = (U, C \cup D, V, f)$, C 为条件属性集, D 为决策属性集, B 为决策表 S 中 C 关于 D 的约简。加入新对象 x 后, 若原决策表约简属性集 B 可以区分新对象, 条件属性集 C 也可以区分新对象, 则 B 为 S 加入 x 后的新决策表的约简。

证明: 不妨设 $U/C = \{X_1, X_2, \dots, X_m\}$, $U/B = \{X_1, X_2, \dots, X_t\}$, $U/D = \{Y_1, Y_2, \dots, Y_n\}$, 加入新对象 x 后, $U \cup \{x\}/C = \{X_1, X_2, \dots, X_r\}$, $U \cup \{x\}/B = \{X_1, X_2, \dots, X_t\}$, $U \cup \{x\}/D = \{Y_1, Y_2, \dots, Y_s\}$ 。

由于原决策表中约简 B 可以区分新对象, 而 C 也可以区分新对象, 可得 $X_p' = X_p'' = \{x\}$, ($p \leq \min(r, t)$), 包含 x 的决策类为 Y_p' 。

进一步根据定理 4, 可得

$$E_{U \cup \{x\}}(D|B) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(D|B) + 2|X_p' - Y_p'|)$$

$$E_{U \cup \{x\}}(D|C) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(D|C) + 2|X_q' - Y_p'|)$$

所以, $E_{U \cup \{x\}}(D|B) = E_{U \cup \{x\}}(D|C)$ 。

对于 $\forall a \in B, B - \{a\} > B$, 由定理 3 可得 $E_U(D|B) < E_U(D|B - \{a\})$ 。

又因为 $X_p' \subseteq X_p^{(B-\{a\})'}$, 所以

$$|X_p' - Y_p'| \leq |X_p^{(B-\{a\})'} - Y_p'|$$

由定理 4 得

$$E_{U \cup \{x\}}(D|B - \{a\}) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(D|B - \{a\}) + 2|X_p^{(B-\{a\})'} - Y_p'|)$$

$$E_{U \cup \{x\}}(D|B) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(D|B) + 2|X_p' - Y_p'|)$$

所以 $E_{U \cup \{x\}}(D|B) \neq E_{U \cup \{x\}}(D|B - \{a\})$ 。

由定理 1 得 B 为约简。

定理 5 给定决策表 $S = (U, C \cup D, V, f)$, C 为条件属性集, D 为决策属性集, B 为决策表 S 中 C 关于 D 的约简。加入新对象 x 后, 若原决策表约简属性集 B 不可以区分新对象, 条件属性集 C 可以区分新对象, 则 B 不为 S 加入 x 后的新决策表的约简。

证明: 不妨设 $U/C = \{X_1, X_2, \dots, X_m\}$, $U/B = \{X_1, X_2, \dots, X_t\}$, $U/D = \{Y_1, Y_2, \dots, Y_n\}$, 加入新对象 x 后, $U \cup \{x\}/C = \{X_1, X_2, \dots, X_r\}$, $U \cup \{x\}/B = \{X_1, X_2, \dots, X_t\}$, $U \cup \{x\}/D = \{Y_1, Y_2, \dots, Y_s\}$ 。

由于原约简属性集 B 不能区分新对象, 设 $X_p' = X_p'' \cup \{x\}$, $X_{m+1}' = \{x\}$ 。

对于 $\forall i < m, i \neq p$, 有 $X_i' = X_i''$, 包含 x 的决策类为 Y_p' , 显然, $|X_p' - Y_p'| \geq |X_p'' - Y_p'|$ 。

当 $X_p' = Y_p'$ 时, 等号成立。

$$E_{U \cup \{x\}}(D|B) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(Y|X) + 2|X_p' - Y_p'|)$$

$$\begin{aligned} &\geq \frac{1}{(|U|+1)^2} (|U|^2 E_U(Y|X) + 2|X_p^C - Y_p'|) \\ &= E_{U \cup \{x\}}(D|C) \end{aligned}$$

所以, $E_{U \cup \{x\}}(D|B) \geq E_{U \cup \{x\}}(D|C)$ 。

由定理 1 得 B 不为约简。证毕。

定理 6 给定决策表 $S=(U, C \cup D, V, f)$, C 为条件属性集, D 为决策属性集, B 为决策表 S 中 C 关于 D 的约简。加入新对象 x 后, 若原决策表约简属性集 B 不可以区分新对象, 条件属性集 C 也不可以区分新对象, 则 B 为 S 加入 x 后的新决策表的约简。

证明: 设决策表 $S=(U, C, D, V, f)$ 中协议的部分构成决策表 $S=(U', C, D, V, f)$, 不协调的部分构成的决策表 $S=(U'', C, D, V, f)$, 设 $U'/B=\{X_1^B, X_2^B, \dots, X_a^B\}$, $U''/B=\{X_{a+1}^B, X_{a+2}^B, \dots, X_b^B\}$, $U'/C=\{X_1^C, X_2^C, \dots, X_v^C\}$, $U''/C=\{X_{v+1}^C, X_{v+2}^C, \dots, X_t^C\}$, 因为若原决策表约简属性集 B 不可以区分新对象, 条件属性集 C 也不可以区分新对象, 所以有 $X_a^B = X_{a+1}^B \cup \{x\}$, $X_b^C = X_{v+1}^C \cup \{x\}$, ($u < a < r, v < b < t$)。显然, $X_a^B = X_b^C$, 包含 x 的决策类为 Y_p' , 所以

$$E_{U \cup \{x\}}(D|B) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(Y|X) + 2|X_a^B - Y_p'|)$$

$$E_{U \cup \{x\}}(D|C) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(Y|X) + 2|X_b^C - Y_p'|)$$

故有 $E_{U \cup \{x\}}(D|B) = E_{U \cup \{x\}}(D|C)$ 。

对于 $\forall a \in B, B - \{a\} > B$, 由定理 3 可得 $E_U(D|B) < E_U(D|B - \{a\})$ 。

又因为 $X_a^B \subseteq X_{a+1}^{B-\{a\}}$, 所以

$$|X_a^B - Y_p'| \leq |X_{a+1}^{B-\{a\}} - Y_p'|$$

由定理 4 得

$$E_{U \cup \{x\}}(D|B - \{a\}) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(D|B - \{a\}) + 2|X_{a+1}^{B-\{a\}} - Y_p'|)$$

$$E_{U \cup \{x\}}(D|B) = \frac{1}{(|U|+1)^2} (|U|^2 E_U(D|B) + 2|X_a^B - Y_p'|)$$

所以, $E_{U \cup \{x\}}(D|B) \neq E_{U \cup \{x\}}(D|B - \{a\})$ 。

4 决策表中属性约简求解算法

4.1 非增量属性约简求解算法(ARC)

根据文献[15]中的 CEBARKCC 算法的思想, 我们给出了一种基于 Liang 条件熵非增量约简算法。

算法 1 非增量属性约简求解算法

输入: 1) 决策表 $S=(U, C \cup D, V, f)$;

2) 新增的对象 x ;

输出: 决策表新的约简 $red_{U \cup \{x\}}(B)$

步骤 1 将新对象放入决策表中, 划分等价类 $U \cup \{x\}/C = \{X_1, X_2, \dots, X_m\}$, $U \cup \{x\}/D = \{Y_1, Y_2, \dots, Y_m\}$ 。

步骤 2 根据公式 $E_U(D|C) = \sum_{i=1}^m \sum_{j=1}^m \frac{|Y_i \cap X_j|}{|U|} \frac{|Y_j - X_i^C|}{|U|}$ 计算条件熵。

步骤 3 $\forall a \in C$, 如果 $E_{U \cup \{x\}}(D|C) < E_{U \cup \{x\}}(D|C - \{a\})$, 则 $Core = Core \cup \{a\}$ 。

步骤 4 如果 $E_{U \cup \{x\}}(D|Core) = E_{U \cup \{x\}}(D|C)$, 则 $red_{U \cup \{x\}}(B) = Core$, 结果输出。否则 $\forall a \in C - Core$, $sig(a, Core, D) = E_{U \cup \{x\}}(D|Core \cup \{a\}) - E_{U \cup \{x\}}(D|B)$ 。当 $a = \text{Max}(sig(a, Core, D))$, 则 $Core = Core \cup \{a\}$, 重新计算 $E_{U \cup \{x\}}(D|B)$ 。

4.2 增量式约简算法(IARC)

通过对对象动态增加决策表原约简对新对象的 3 种区分情况, 我们设计了一种新的增量约简求解算法。

算法 2 增量式约简算法(IARC)

输入: 1) 决策表 $S=(U, C \cup D, V, f)$, 现有决策表的约简 $red_U(B)$, 决策表的核属性 $Core$;

2) 新增的对象 x ;

输出: 决策表新的约简 $red_{U \cup \{x\}}(B)$

步骤 1 将决策表 $S=(U, C \cup D, V, f)$, 按照属性集 B 划分等价类 $U'/B = \{X_1^B, X_2^B, \dots, X_a^B\}$, $U''/B = \{X_{a+1}^B, X_{a+2}^B, \dots, X_b^B\}$; 按照条件属性集划分等价类 $U'/C = \{X_1^C, X_2^C, \dots, X_v^C\}$, $U''/C = \{X_{v+1}^C, X_{v+2}^C, \dots, X_t^C\}$ 。

步骤 2 在 $U'/B, U''/B$ 中查找和 x 的 $red_U(B)$ 属性集完全相同的等价类 X_a^B , 如果不存在, $red_{U \cup \{x\}}(B) = red_U(B)$, 结果输出; 若存在, 则 $X_a^B = X_a^B \cup \{x\}$, 转到步骤 3。

步骤 3 在 $U'/C, U''/C$ 中找和 C 条件属性集完全相同的等价类 X_b^C 。如果不存在, 转到步骤 4; 若存在, 有下面两种状况:

(1) $X_b^C \in U''/C$ 或 $X_a^B \in U'/B$, $X_b^C \in U'/C$, $red_{U \cup \{x\}}(B) = red_U(B)$, 结果输出;

(2) 如果 $X_a^B \in U'/B$, $X_b^C \in U'/C$, 转到步骤 4。

步骤 4 从 $Core$ 开始, $Core = B$, 如果 $E_{U \cup \{x\}}(D|B) = E_{U \cup \{x\}}(D|C)$, 则 B 为属性约简, $red_{U \cup \{x\}}(B) = red_U(B)$ 结果输出; 否则, 对所有的属性 $a \in C - B$, 计算 $sig(a, B, D) = E_{U \cup \{x\}}(D|B \cup \{a\}) - E_{U \cup \{x\}}(D|B)$, 当 $a = \text{Max}(sig(a, B, D))$, 则 $B = B \cup \{a\}$, 重新计算 $E_{U \cup \{x\}}(D|B)$ 。

4.3 实例说明

实例 1 表 1 所列的是一张二值数据表, 其中有 4 个对象和 5 个属性, $C = \{C_1, C_2, C_3, C_4\}$ 为条件属性集, D 为决策属性。对于实例 1, 约简属性集为 $\{1, 2\}$ 。

表 1 数据表

元素	C_1	C_2	C_3	C_4	D
x_1	1	0	1	0	0
x_2	1	0	1	0	1
x_3	1	1	0	0	0
x_4	0	1	0	0	1

为进一步说明算法, 对于实例 1, 下面通过增加不同的对象来说明对象插入后属性约简更新的不同情况 $U'/B = \{\{x_3\}, \{x_4\}\}$, $U''/B = \{\{x_1, x_2\}\}$, $U'/C = \{\{x_3\}, \{x_4\}\}$, $U''/C = \{\{x_1, x_2\}\}$ 。

(1) 对实例 1, 若新增对象 x 为 $\{0, 0, 1, 0, 0\}$, 原属性约简属性集 B 可以区分新对象, 约简属性保持不变。

(2) 对实例 1, 若新增对象 x 为 $\{1, 0, 1, 1, 1\}$, 原属性集 B 不能区分新对象, C 可以区分新对象, 重新计算约简, 约简为 $\{1, 2, 3\}$ 。

(3) 对实例 1, 若新增对象 x 为 $\{1, 1, 0, 0, 0\}$, 原约简属性集 B 可以区分新对象, C 也可以区分新对象, 约简保持不变。

(4) 对实例 1, 若新增对象 x 为 $\{0, 1, 0, 0, 0\}$, 原约简属性集 B 不可以区分新对象, C 也不可以区分新对象, 约简保持不变。

4.4 实验结果

为进一步验证算法的性能, 在硬件配置 CPU 为 3.4GHz, 内存为 1GB 的台式机上, 用 MATLAB 实现了非增量和增量的条件熵约简算法。利用 UCI 数据集 mushroom 来进行实

(下转第 239 页)

参考文献

[1] Shi Kaiquan. P-sets and its applications[J]. An International Advances in Systems Science and Applications,2009,9(2):209-219

[2] 史开泉. P-集合[J]. 山东大学学报:理学版,2008,43(11):77-84

[3] 史开泉,张丽. 内 P-集合与数据外-恢复[J]. 山东大学学报:理学版,2009,44(4):8-14

[4] 汤积华,陈保会,史开泉. P-集合与(F,F)-数据生成-辨识[J]. 山东大学学报:理学版,2009,44(9):50-55

[5] 于秀清. P-集合的识别与筛选[J]. 山东大学学报:理学版,2010,45(1):94-98

[6] 张飞,陈萍,张丽. P-集合的 P-分离与应用[J]. 山东大学学报:理

学版,2010,45(3):18-22

[7] Shi Kaiquan,LI Xiuhong. Camouflaged informatiation and its on identification and its applications[J]. An International Journal Advances in System Science and Applications,2010,10(2):208-216

[8] Zhang li, Cui Yuquan. Outer P-sets and data internal-recover [J]. An International Journal Advances in System Science and Applications,2010,10(2):229-236

[9] Lin Hong-kang, Li Yu-ying. P-sets and its P-separation theorems[J]. An International Journal Advances in System Science and Applications,2010,10(2):245-251

[10] Zhang Guan-yu,LI En-zhong. Information gene and its information knock-out/knock-in[J]. An International Journal Advances in System Science and Applications,2010,10(2):267-27

(上接第 231 页)

验,去掉其中非完备的对象,该数据库共有 5644 个对象,记录蘑菇的有 22 个条件属性、1 个决策属性。将 mushroom 数据库看作决策表,选择 4644 个对象作为基准决策表,从剩下的 1000 个对象中依次增加 100,200,300,400,500,600,700,800,900,1000 个对象作为增量,实验结果如表 2 所列。

表 2 非增量与增量约简算法的实验结果对比表

对象数	ARC		IARC	
	约简结果	时间(S)	约简结果	时间(S)
4744	{2,3,4,6,8,10,13,14,15,16,21,22}	7774.39	{2,3,4,6,8,10,13,14,15,16,21,22}	5564.67
4844	{2,3,4,6,8,10,13,14,15,16,21,22}	8035.61	{2,3,4,6,8,10,13,14,15,16,21,22}	5827.67
4944	{2,3,4,6,8,10,13,14,15,16,21,22}	8459.110	{2,3,4,6,8,10,13,14,15,16,21,22}	6151.89
5044	{2,3,4,6,8,10,13,14,15,16,21,22}	9311.061	{2,3,4,6,8,10,13,14,15,16,21,22}	6447.88
5144	{2,3,4,6,8,10,13,14,15,16,21,22}	9659.641	{2,3,4,6,8,10,13,14,15,16,21,22}	6772.91
5244	{2,3,4,6,8,10,13,14,15,16,21,22}	10279.69	{2,3,4,6,8,10,13,14,15,16,21,22}	7123.34
5344	{2,3,4,6,8,10,13,14,15,16,21,22}	10750.16	{2,3,4,6,8,10,13,14,15,16,21,22}	7507.74
5444	{2,3,4,6,8,10,13,14,15,16,21,22}	11527.95	{2,3,4,6,8,10,13,14,15,16,21,22}	7865.13
5544	{2,3,4,6,8,10,13,14,15,16,21,22}	12480.84	{2,3,4,6,8,10,13,14,15,16,21,22}	8340.89
5644	{2,3,4,6,7,8,10,13,14,15,16,21,22}	12908.11	{2,3,4,6,7,8,10,13,14,15,16,21,22}	8829.38

从图 1 可见,本文的算法优于传统的非增量算法,原因正是本文的算法可以有效利用已经计算过的条件熵和约简结果,通过更新已有的约简结果来计算新约简。

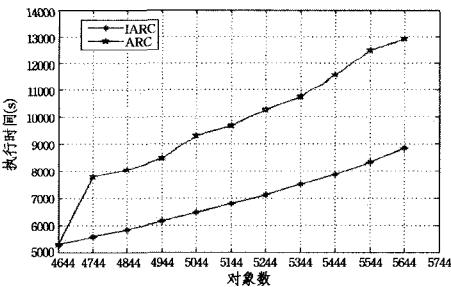


图 1 算法执行时间

结束语 本文通过分析决策表对象动态增加时条件熵关

于条件属性的变化机制,提出了一种基于条件熵的属性约简增量式更新算法。算法充分利用了原决策表的约简属性和条件熵的信息,明显地减少了增量更新属性约简的计算量,为动态更新决策表的属性约简提供了一种快速有效的新途径。

参考文献

[1] Pawlak Z. Rough sets theoretical aspects of reasoning about data [M]. Kluwer Academic Publisher,1991

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

[3] 梁吉业,李德玉. 信息系统中的不确定性与知识获取[M]. 北京:科学出版社,2005

[4] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domain for neural networks[J]. Computational Intelligence,1995,11(2):339-347

[5] 刘少辉,盛秋戡,吴斌,等. Rough 集高效算法的研究[J]. 计算机学报,2003,26(5):524-529

[6] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligences,1998,105(1/2):77-103

[7] Woblewski J. Finding minimal reducts using genetic algorithm [R]. Warsaw University of Technology; ICS Research Report, 16/95,1995

[8] 林俊伟,叶东毅. 基于领域辨识矩阵的属性约简增量式算法[J]. 计算机应用,2009,29(11):119-121

[9] Hu F, Wang G Y, Huang H, et al. Incremental attribute reduction based on elementary sets[C]//Processing of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Regina, Canada,2005:183-193

[10] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报,2007,30(5):815-823

[11] 梁吉业,魏巍,钱宇华. 一种基于条件熵的增量核求解方法[J]. 系统工程理论与实践,2008,28(4):81-89

[12] Liang J Y, Xu Z B. The algorithm on knowledge reduction in incomplete information system[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based System,2002,10(1):95-103

[13] 梁吉业,钱宇华. 信息系统中的信息粒与熵理论[J]. 中国科学(E 辑),2008,38(12):2048-2065

[14] 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