

一种新的信息系统启发式属性约简算法

崔 巍¹, 徐章艳²

(1 中国地质大学(北京)人文经管学院, 北京 100083; 2 广西师范大学 计算机系, 广西 桂林 541004)

摘 要: 为降低信息系统属性约简算法的时间复杂度, 首先引入简化信息系统的概念, 然后定义了简化信息系统中属性集的区分对象数, 证明了基于简化信息系统区分对象数的属性约简与原属性约简是等价的. 进而定义了简化信息系统的属性重要性, 并设计了一个快速求属性重要性的算法. 同时, 以属性重要性为启发式信息, 设计了一个新的属性约简算法. 新算法的时间复杂度降为 $\max\{O(|C||U|), O(|C|^2|U|C|)\}$. 最后用一个实例说明了新算法的有效性.

关键词: 粗糙集; 简化信息系统; 区分对象数; 属性重要性; 属性约简; 算法复杂度

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2010)09-0080-06

A New Heuristic Attribution Reduction Algorithm Based on Information System

CUI Wei¹, XU Zhang-yan²

(1 School of Economics and Management, China University of Geosciences(Beijing), Beijing 100083, China;

2 Department of Computer, Guangxi Normal University, Guilin 541004, China)

Abstract: To reduce the time complexity of attribute reduction algorithm based on information system, the definition of the simplified information system is introduced at first. Then the number of discernibility object of attribute set is defined in the simplified information system. And it is proved that attribute reduction based on the number of discernibility object in the simplified system is equivalent to that in the old information system. Then the definition of attribute importance in the simplified information system is proposed. And an algorithm for computing attribute importance is designed. At the same time, a new attribute reduction algorithm is designed with attribute importance as heuristic. Its time complexity is cut down to. At last, an example is used to illustrated the efficiency of the new algorithm.

Key words: rough set; simplified information system; number of discernibility object; attribution importance; attribution reduction; algorithm complexity

1 引言

属性约简是粗糙集理论中的核心内容之一. 信息系统的属性约简, 就是在保持信息系统的分类能力不变的前提下, 删除其中的冗余属性. 人们希望能找到具有最少属性的约简, 即最佳约简. 然而, 找到一个最佳约简是一个很困难的问题, 通常是采用启发式搜索方法, 求出最佳或次佳约简^[1].

文献[2] 给出一个最坏时间复杂度为 $O(|C|^2$

$|U|^2)$ 的属性约简算法, 使得算法降低一个数量级; 文献[3] 通过快速排序方法又将属性约简算法的最坏时间复杂度降为 $O(|C|^2|U|\log|U|)$; 分析了文献[2] 中给出的启发发信息的不足后, 文献[4] 给出了一个更为合理的启发信息, 并在此基础上设计了一个新的属性约简算法, 其时间复杂度为 $\max\{O(|C||U|\log|U|), O(|C|^2|U|)\}$, 使得算法复杂度又降低了半个数量级.

我们认为属性约简算法的复杂度还可以进一步

收稿日期: 2010-07-13; 修回日期: 2010-09-13

基金项目: 国家自然科学基金资助项目(60463003); 广西教育厅科研项目(200626)

降低.为此,构造了信息系统的简化信息系统,给出了新的属性约简算法^[5].

2 基本概念

定义1 四元组 $S = (U, C, V, f)$ 是一个信息系统,其中 U 表示对象的非空有限集合,称为论域; C 表示属性的非空有限集; $V = \bigcup_{a \in C} V_a$, V_a 是属性 a 的值域, f 表示 $U \times C \rightarrow V$ 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 $\forall a \in C, x \in U, \text{有 } f(x, a) \in V_a$,每一个属性子集 $P \subseteq C$ 决定了一个二元不可区分关系 $\text{IND}(P)$:

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

关系 $\text{IND}(P), P \subseteq C$, 构成了 U 的一个划分,用 $U/\text{IND}(P)$ 表示,简记为 U/P , U/P 中的任何元素 $[x]_P$ 称为等价类.

定义2 在信息系统 $S = (U, C, V, f)$ 中, $a \in P \subseteq C$, 若 $\text{IND}(P - \{a\}) = \text{IND}(P)$; 则称属性 a 在 P 中是必要的.

定义3 在信息系统 $S = (U, C, V, f)$ 中, 若 $a \in P \subseteq C$ 在 P 中都是必要的, 则称 C 为独立的, 否则, 称 P 为相依的.

定义4 在信息系统 $S = (U, C, V, f)$ 中, $P \subseteq C$, 如果满足 $\text{IND}(P) = \text{IND}(C)$, 且 P 为独立的, 则称 P 为 C 的一个属性约简.

定义5 在信息系统 $S = (U, C, V, f)$ 中, $P, Q \subseteq C$, 记 $U/P = \{P_1, P_2, \dots, P_t\}, U/Q = \{Q_1, Q_2, \dots, Q_s\}$. 若 $\forall P_i \in U/P \Rightarrow \exists Q_j \in U/Q$ 使 $P_i \subseteq Q_j$, 则称 U/P 为 U/Q 的加细, 常记为 $U/P \leq U/Q$.

引理1 在信息系统 $S = (U, C, V, f)$ 中, $\forall Q \subseteq P \subseteq C$, 则有 $U/P \leq U/Q$ ^[4].

引理1说明,在信息系统 $S = (U, C, V, f)$ 中, U/C 是最小的知识粒度,即划分 U/C 的每个等价类是最小的,若在条件属性 C 中去掉某个属性,则划分 U/C 的一些等价类就会合并成新等价类,并成为一种新的知识粒度.故由引理1,我们可以从划分 U/C 的每一个等价类中取出一个代表元(等价类中任意一个元素均可)组成一个简化的信息系统.

定义6 在信息系统 $S = (U, C, V, f)$ 中, 设 $U/C = \{C_1, C_2, \dots, C_k\} = \{[x'_1]_C, [x'_2]_C, \dots, [x'_k]_C\}$, 令 $U' = \{x'_1, x'_2, \dots, x'_k\}$, 则称 $S' =$

(U', C, V, f) 为信息系统的简化信息系统.

3 属性的重要性

在基于信息系统的属性约简中,除了用定义设计属性约简算法之外,还有一种常用的方法就是差别矩阵方法^[6].用这种方法设计属性约简算法时,首先要求出差别矩阵,然而求差别矩阵的时间复杂度是 $O(|C| \cdot |U|^2)$, 故用差别矩阵设计属性约简算法时,其时间复杂度至少是 $O(|C| \cdot |U|^2)$. 相对于文献[4]中属性约简算法的时间复杂度 $\max\{O(|C| \cdot |U| \log |U|), O(|C|^2 \cdot |U|)\}$, 这个时间复杂度并不好.特别当 $|U| \gg |C|$, 即大型信息系统,用差别矩阵方法设计的属性约简算法更是不理想^[6].但是在用差别矩阵方法设计的属性约简算法中,常常用差别矩阵中差别元素个数的多少作为启发信息,这一思想给我们启发:能不能用属性集 P 导出划分所产生的差别对象的个数作为知识 P 的度量?基于这一思想,我们给出如下知识的区分对象数的定义.

定义7 设 $S = (U, C, V, f)$ 为信息系统,其中 $P \subseteq C, U/P = \{P_1, P_2, \dots, P_t\}$, 知识 P 的区分对象数定义为

$$\text{Dis}(P) = \sum_{i=1}^t |P_i| \cdot (|U| - |P_i|).$$

注:由于 $U/\emptyset = \{u\}$, 故有 $\text{Dis}(\emptyset) = 0$.

事实上,在信息系统 $S = (U, C, V, f)$ 中,若 $P \subseteq C, U/P = \{P_1, P_2, \dots, P_t\}$, 则同一个等价类中的任意两个不同的对象不会产生差别;而对于任意两个不同时属于同一等价类中的对象一定会产生差别,我们称产生差别的两个对象为差别对象.故 $P_i (i = 1, 2, \dots, t)$ 能产生 $|P_i| \cdot (|U| - |P_i|)$ 个差别对象.从而,知识 P 产生的差别对象数为 $\sum_{i=1}^t |P_i| \cdot (|U| - |P_i|)$.

引理2 在信息系统 $S = (U, C, V, f)$ 中, $P \subseteq C, \forall a \in (C - P)$, 则有 $U/(P \cup \{a\}) = \bigcup_{x \in U/P} (X/\{a\})$ ^[4].

定理1 设 $S = (U, C, V, f)$ 为信息系统, 设 $P \subseteq C, \forall a \in (C - P)$, 则有 $\text{Dis}(P \cup \{a\}) \geq \text{Dis}(P)$.

证明 设 $U/P = \{P_1, P_2, \dots, P_t\}$, 由引理2知有 $U/(P \cup \{a\}) = \bigcup_{x \in U/P} (X/\{a\})$. 故对 $\forall P_i \in U/P$, 可设 $P_i/\{a\} = \{P_{i1}, P_{i2}, \dots, P_{ir_i}\}$, 其中 $r_i \geq$

1. 故有

$$U/(P \cup \{a\}) = \{P_{11}, P_{12}, \dots, P_{1r_1}, P_{21}, P_{22}, \dots, P_{2r_2}, \dots, P_{t1}, P_{t2}, \dots, P_{tr_t}\};$$

$$|P_i| = |P_{i1}| + |P_{i2}| + \dots + |P_{ir_i}| \quad (i = 1, 2, \dots, t).$$

另一方面,有

$$\sum_{j=1}^{r_i} |P_{ij}| (|U| - |P_{ij}|) = \sum_{j=1}^{r_i} |P_{ij}| |U| - \sum_{j=1}^{r_i} |P_{ij}|^2 \geq |U| |P_i| - \sum_{j=1}^{r_i} |P_{ij}|^2 = |U| |P_i| - |P_i|^2 = |P_i| (|U| - |P_i|).$$

故有

$$\text{Dis}(P \cup \{a\}) = \sum_{i=1}^t \sum_{j=1}^{r_i} |P_{ij}| (|U| - |P_{ij}|) \geq$$

$$\sum_{i=1}^t |P_i| (|U| - |P_i|) = \text{Dis}(P).$$

定义 8 在信息系统 $S = (U, C, V, f)$ 中, $S' = (U', C, V, f)$ 为其简化信息系统. 设 $P \subseteq C$, $\forall a \in (C - P)$, $U'/P = \{P'_1, P'_2, \dots, P'_t\}$ 定义简化信息系统的区分对象数为

$$\text{SDis}(P) = \sum_{i=1}^t |P'_i| (|U'| - |P'_i|).$$

定理 2 在信息系统 $S = (U, C, V, f)$ 中, $S' = (U', C, V, f)$ 为其简化信息系统, 则有

$$\text{SDis}(C) = |U'| (|U'| - 1).$$

证明 由于 $U'/C = \{\{x'_1\}, \{x'_2\}, \dots, \{x'_k\}\} = \{\{C'_1\}, \{C'_2\}, \dots, \{C'_k\}\}$, $|U'| = |C'_1| + |C'_2| + \dots + |C'_k|$, 故有

$$\text{SDis}(C) = \sum_{i=1}^k |C'_i| (|U'| - |C'_i|) = \sum_{i=1}^k (|U'| - 1) = |U'| (|U'| - 1).$$

定理 3 在信息系统 $S = (U, C, V, f)$ 中, $S' = (U', C, V, f)$ 为其简化信息系统, 设 $P \subseteq C$, 则 $\text{SDis}(P) = \text{SDis}(C)$ 的充分必要条件是 $|U'/P| = |U'|$.

证明 设 $U'/P = \{P'_1, P'_2, \dots, P'_t\}$. 当 $|U'/P| = |U'|$ 时, 则有 $|P'_1| = |P'_2| = \dots = |P'_t| = 1$, $|U'| = |P'_1| + |P'_2| + \dots + |P'_t|$, 故有

$$\text{SDis}(P) = \sum_{i=1}^t |P'_i| (|U'| - |P'_i|) = \sum_{i=1}^t (|U'| - 1) = |U'| (|U'| - 1)$$

从而有 $\text{SDis}(P) = \text{SDis}(C)$.

当 $\text{SDis}(P) = \text{SDis}(C)$ 时, 假设 $|U'/P| \neq |U'|$ 不成立, 设 $U' = \{x'_1, x'_2, \dots, x'_k\}$, 由 U' 的定义知 $U'/C = \{\{x'_1\}, \{x'_2\}, \dots, \{x'_k\}\} = \{\{x'_1\}_C, \{x'_2\}_C, \dots, \{x'_k\}_C\}$. 由于 $|U'/P| \neq |U'|$, 故有 $U'/P \neq U'/C$. 另一方面, 由于 $P \subseteq C$, 则由引理 1 知 $U'/P < U'/C$, 即至少存在 $P'_j = [x'_{j1}]_P \in U'/P$ 使得 $[x'_{j1}]_P = [x'_{j1}]_C \cup [x'_{j2}]_C \cup \dots \cup [x'_{jh}]_C$, 其中 $h > 1$. 从而有

$$|P'_j| (|U'| - |P'_j|) = h(k - h) = hk - h^2 < hk - h = h(k - 1) = |[x'_{j1}]_C| (|U'| - |[x'_{j1}]_C|) + |[x'_{j2}]_C| (|U'| - |[x'_{j2}]_C|) + \dots + |[x'_{jh}]_C| (|U'| - |[x'_{jh}]_C|)$$

故有

$$\text{SDis}(P) = \sum_{i=1}^t |P'_i| (|U'| - |P'_i|) =$$

$$|P'_j| (|U'| - |P'_j|) + \sum_{i=1, i \neq j}^t |P'_i| (|U'| - |P'_i|) < |[x'_{j1}]_C| (|U'| - |[x'_{j1}]_C|) + |[x'_{j2}]_C| (|U'| - |[x'_{j2}]_C|) + \dots + |[x'_{jh}]_C| (|U'| - |[x'_{jh}]_C|) + \sum_{i=1, i \neq j}^t |P'_i| (|U'| -$$

$$|P'_i|) \leq \sum_{i=1}^k |[x'_i]_C| (|U'| - |[x'_i]_C|) =$$

$$\text{SDis}(C) = |U'| (|U'| - 1)$$

即有

$$\text{SDis}(P) < \text{SDis}(C)$$

这与条件矛盾. 故假设不成立, 从而命题成立.

综上所述, 命题成立.

定义 9 在信息系统 $S = (U, C, V, f)$ 中, $S' = (U', C, V, f)$ 为其简化信息系统. 设 $P \subseteq C$, 若 $\text{SDis}(P) = \text{SDis}(C)$, 且 $\forall a \in P$ 均有 $\text{SDis}(P - \{a\}) \neq \text{SDis}(C)$, 则称 P 为 C 的一个基于区分对象数的属性约简.

定理 4 在信息系统 $S = (U, C, V, f)$ 中, $S' = (U', C, V, f)$ 为其简化信息系统, 则基于区分对象数的属性约简与原信息系统的属性约简等价.

证明 设 P 为 C 的任意一个基于区分对象数的属性约简, 则有 $\text{SDis}(P) = \text{SDis}(C)$, 故由定理 3 知有 $|U'/P| = |U'|$, 由引理 1 知 $U'/P = U'/C$. 再由 U' 的定义知有 $U/P = U/C$, 即 $\text{IND}(P) = \text{IND}(C)$.

一方面, 对 $\forall a \in P$ 均有 $\text{SDis}(P - \{a\}) \neq \text{SDis}(C)$, 故由定理 3 知有 $|U'/(P - \{a\})| \neq$

$|U'|$,由引理1知 $U'/(P - \{a\}) \neq U'/C$.再由 U' 的定义知,则有 $U/(P - \{a\}) \neq U/C$.即 $IND(P - \{a\}) \neq IND(C)$.由属性约简的定义知, P 为 C 的一个属性约简.

设 P 为 C 的任意一个基于原信息系统的属性约简,则有 $IND(P) = IND(C)$,即有 $U/P = U/C$,由 U' 的定义知有 $U'/P = U'/C$,即 $|U'/P| = |U'|$,故由定理3知有 $SDis(P) = SDis(C)$.

另一方面,对 $\forall a \in P$ 均有 $IND(P - \{a\}) \neq IND(C)$,即有 $U/(P - \{a\}) \neq U/C$.由 U' 的定义知有 $U'/(P - \{a\}) \neq U'/C$,即 $|U'/(P - \{a\})| \neq |U'|$.故由定理3知有 $SDis(P - \{a\}) \neq SDis(C)$.故 P 为 C 的一各基于区分对象数的属性约简.

由 P 的任意性知,命题成立.

定义10 设 $S = (U, C, V, f)$ 为一信息系统, $S' = (U', C, V, f)$ 为其简化信息系统,设 $P \subseteq C$, $\forall a \in C - P$,属性 a 的重要性定义为

$$Sig_P(a) = SDis(P \cup \{a\}) - SDis(P).$$

由定理1易得如下的定理:

定理5 设 $S = (U, C, V, f)$ 为一信息系统, $S' = (U', C, V, f)$ 为其简化信息系统,设 $P \subseteq C$, $\forall a \in C - P$,则有 $Sig_P(a) \geq 0$.

定理6 设 $S = (U, C, V, f)$ 为一信息系统, $S' = (U', C, V, f)$ 为其简化信息系统,则有

$$Sig_P(a) = \sum_{X \in U'/P \wedge |X| \neq 1} |X|(|X| - 1) - \sum_{X \in U'/(P \cup \{a\}) \wedge |X| \neq 1} |X|(|X| - 1).$$

证明 设 $P \subseteq C, \forall a \in C - P, U'/P = \{P'_1, P'_2, \dots, P'_t\}, U'/(P \cup \{a\}) = \{Q_1, Q_2, \dots, Q_s\}$,则有

$$\begin{aligned} Sig_P(a) &= SDis(P \cup \{a\}) - SDis(P) \\ &= \sum_{i=1}^s |Q_i|(|Q_i| - 1) - \sum_{i=1}^t |P'_i|(|P'_i| - 1) \\ &= \sum_{i=1}^s |Q_i|(|Q_i| - 1) - \sum_{i=1}^s |Q_i|^2 + \sum_{i=1}^s |Q_i| - \sum_{i=1}^t |P'_i|(|P'_i| - 1) \\ &= |U'|^2 - \sum_{i=1}^s |Q_i|^2 - \left\{ |U'|^2 - \sum_{i=1}^t |P'_i|^2 \right\} \\ &= \sum_{i=1}^t |P'_i|^2 - \sum_{i=1}^s |Q_i|^2 \\ &= \sum_{i=1}^t |P'_i|(|P'_i| - 1) + \sum_{i=1}^t |P'_i| - \left\{ \sum_{i=1}^s |Q_i|(|Q_i| - 1) + \sum_{i=1}^s |Q_i| \right\} \\ &= \sum_{i=1}^t |P'_i|(|P'_i| - 1) + |U'| - \left\{ \sum_{i=1}^s |Q_i|(|Q_i| - 1) + |U'| \right\} \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^t |P'_i|(|P'_i| - 1) - \sum_{i=1}^s |Q_i|(|Q_i| - 1) \\ &= \sum_{X \in U'/P \wedge |X| \neq 1} |X|(|X| - 1) + \sum_{X \in U'/P \wedge |X| \neq 1} |X|(|X| - 1) \\ &\quad - \left\{ \sum_{X \in U'/(P \cup \{a\}) \wedge |X| \neq 1} |X|(|X| - 1) + \sum_{X \in U'/(P \cup \{a\}) \wedge |X| \neq 1} |X|(|X| - 1) \right\} \\ &= \sum_{X \in U'/P \wedge |X| \neq 1} |X|(|X| - 1) - \sum_{X \in U'/(P \cup \{a\}) \wedge |X| \neq 1} |X|(|X| - 1). \end{aligned}$$

定理6说明,对 $\forall X \in U'/P$,若 $|X| = 1$,则 X 不影响属性的重要性.这样,我们在计算属性的重要性时,可以将 X 去掉,从而可以减少对象 U' 的个数,提高属性重要性的计算效率.

4 计算属性重要性的算法和属性约简算法

4.1 快速计算简化信息系统的算法

要求出属性的重要性,首先要求出简化信息系统,而求简化信息系统的关键是求出划分 U/C .求 U/C 的时间复杂度直接影响属性约简算法的复杂度.利用基数排序的思想,给出一个计算简化信息系统的快速算法,其时间复杂度为 $O(|C| |U|)$,空间复杂度为 $O(|U|)^{[5]}$.

算法1 求简化信息系统

输入:决策表 $S = (U, C, V, f), U = \{x_1, x_2, \dots, x_n\}, C = \{c_1, c_2, \dots, c_s\}$

输出: $U', m_i, M_i (i = 1, 2, \dots, s)$

(1) 对每一个 $c_i (i = 1, 2, \dots, s)$ 统计 $f(x_j, c_i) (j = 1, 2, \dots, n)$ 的最大值和最小值,分别记为 M_i 和 m_i ;

(2) 以静态链表依次存储对象 $x_1, x_2, \dots, x_n$; 令表头指针指向 x_1 ;

(3) for $(i = 1; i < s + 1; i++)$

① 第 i 趟“分配”:建立 $M_i - m_i + 1$ 空队列,令 $front_k$ 和 $end_k (k = 0, 1, 2, \dots, M_i - m_i)$ 分别为第 k 个队列的头指针和尾指针.将链表中的对象 $x \in U$ 按链表中的次序分配到第 $f(x, c_i) - m_i$ 个队列中去.

② 第 i 趟“收集”:表头指针指向第一个非空队列的头指针,修改每一个非空队列的尾指针,令其指向下一个非空队列的队头对象,将 $M_i - m_i + 1$ 个队列重新组成一个链表.

(4) 设由第3步得到链表中的对象序列为 $x'_1, x'_2, \dots, x'_n$;

$t = 1; B_t = \{x'_1\}$;

for $(j = 2; j < n + 1; j++)$

若任一 $c_i \in C (i = 1, 2, \dots, s)$ 均有 $f(x'_j, c_i) = f(x'_{j-1}, c_i)$, 则 $B_t = B_t \cup \{x'_j\}$;
 否则 $\{t = t + 1; B_t \cup \{x'_j\}\}$;
 (5) $U' = \emptyset$;
 for ($i = 1; i < t + 1; i++$)!
 取出 B_i 中的第一个对象 y 并入 U' ;

4.2 计算属性重要性的算法

由定理 6 可以设计如下的快速计算 $\text{Sig}_p(a)$ 的算法.

算法 2 计算 $\text{Sig}_p(a)$

输入: $S_p = \{X \mid X \in U' / P \wedge |X| \neq 1\}; M_a;$
 m_a ;

输出: $S_{P \cup \{a\}} = \{X \mid X \in U' / (P \cup \{a\}) \wedge |X| \neq 1\}; \text{Sig}_p(a)$;

(1) 计算 $\sum_{X \in S_p} |X| (|X| - 1)$;
 (2) for ($j = 1; j < |S_p| + 1; j++$)
 $\forall x \in X_j$, 令 $x.\text{flag} = j$; $// X_j \in S_p$
 (3) 令 $T = X_1 \cup X_2 \cup \dots \cup X_{|S_p|} = \{x_1, x_2, \dots, x_z\}$; 其中 $X_j \in S_p (j = 1, 2, \dots, |S_p|)$
 (4) 以静态链表依次存储对象 $x_1, x_2, \dots, x_z$;
 (5) 建立 $M_a - m_a + 1$ 空队列, 令 front_k 和 $\text{end}_k (k = 0, 1, 2, \dots, M_a - m_a)$ 分别为第 k 个队列的头指针和尾指针. 将链表中的对象 $x \in T$ 按链表中的次序分配到第 $f(x, a) - m_a$ 个队列中去. $//$ 其中 M_a, m_a 分别表示原决策表中所有对象在属性 a 上取的最大值和最小值.

(6) $S_{P \cup \{a\}} = \emptyset$;

(7) 对由第 4 步得到的每一个非空对列 $\text{Col} = \{y_1, y_2, \dots, y_h\}$ 做如下处理:

$t = 1; B_t = \{y_1\}$;

for ($j = 2; j < h + 1; j++$)

{ if ($y_j.\text{flag} = y_{j-1}.\text{flag}$)

$B_t = B_t \cup \{y_j\}$;

else $\{t = t + 1; B_t = \{y_j\}\}$;

}

for ($i = 1; i < t + 1; i++$)

if ($|B_i| \neq 1$)

$S_{P \cup \{a\}} = S_{P \cup \{a\}} \cup \{B_i\}$;

(8) 计算 $\sum_{X \in S_{P \cup \{a\}}} |X| (|X| - 1)$;

(9) 输出 $\text{Sig}_p(a) = \sum_{X \in S_p} |X| (|X| - 1) -$

$\sum_{X \in S_{P \cup \{a\}}} |X| (|X| - 1)$.

算法 2 说明: 第 2 步是将输入集合 S_p 中的同一个等价类中的每一个元素做上相同的标记, 并且使得任何两个不同等价类中元素的标记不同. 第 3 步将输入集合 S_p 中的每一个等价类依次合并得到一个新的由对象组成的集合 T . 第 4 步将集合 T 的元素依次组成静态链表, 这样若相邻对象的标记相同, 则它们一定在同一个等价类中 (即在输入集合 S_p 中某一等价类中); 若相邻对象的标记不相同, 则它们一定在不同的等价类中. 第 5 步是将静态链表中的元素依次分到 $M_a - m_a + 1$ 个对列中, 即得到 $T / \{a\}$. 第 7 步中的第 1 个循环得到的 B_t 实际上是 $T / (P \cup \{a\})$ 中的一个元素, 即 $B_t \in T / (P \cup \{a\})$. 因为由第 5 步得到的每一非空对列 $\text{Col} \in T / \{a\}$, 若 Col 中相邻元素 x 和 y 的标记相同, 则 x 和 y 一定在 T / P (即输入集合 S_p) 的某一元素中 (注意在 Col 中只有相邻的元素才有可能使得它们的标记相同, 因为在第 4 步的静态链表中只有相邻元素的标记才有可能相同, 而第 5 步是按静态链表中元素依次分放到对列中的); 若 x 和 y 的标记不相同, 则 x 和 y 一定在 T / P 的不同元素中, 故由第 7 步的第一个循环求出的是 $T / (P \cup \{a\})$ 中的一个元素. 第 7 步的第二循环是输出集合 $S_{P \cup \{a\}}$ 的. 由定理 6 知输出的 $\text{Sig}_p(a)$ 即为我们计算的属性重要性.

算法 2 的复杂度分析: 第 1 步, 第 2 步和第 4 步的时间复杂度为 $O(|T|)$; 第 5 步的时间复杂度为 $O(|T|)$; 第 7 步的第一个循环的时间复杂度为 $O(|\text{Col}|)$, 第二个循环的最坏时间复杂度为 $O(|\text{Col}|)$, 故第 7 步总的时间复杂度为 $O(|T|)$. 从而算法总的时间复杂度为 $O(|T|)$. 由于 $O(|T|) \leq O(|U'|)$, 故算法最坏的时间复杂度为 $O(|U'|)$. 算法的空间复杂度为 $O(|T|)$, 同理最坏的空间复杂度为 $O(|U'|)$.

4.3 属性约简算法

由算法 1 和算法 2, 可设计如下的快速属性约简算法.

算法 3 快速属性约简算法

输入: 决策表 $S = (U, C, V, f)$, $U = \{x_1, x_2, \dots, x_n\}$, $C = \{c_1, c_2, \dots, c_s\}$;

输出: 属性约简 R ;

(1) 用算法 1 计算出 $U' = \{x'_1, x'_2, \dots, x'_k\}$, $m_i, M_i (i = 1, 2, \dots, s)$; $\text{Num} = 0$;

(2) $R = \emptyset$; $S_0 = \{\{x'_1, x'_2, \dots, x'_k\}\}$; $//$ 开始时输入集合 S 中只有一个等价类, 即 U' .

(3) 对任意 $c_i \in C - R$, 计算 $\text{Sig}_R(c_i)$;

(4) 记 $\text{Sig}_R(c_k) = \max_{c_i \in C-R} \text{Sig}_R(c_i)$, 若这样的属性不止一个时, 则任取其一;

$\text{Num} = \text{Num} + \text{Sig}_R(c_k); R = R \cup \{c_k\};$

(5) 若 $\text{Num} = |U'|(|U'| - 1)$, 则算法终止, 输出属性约简 R ; 否则转(3).

下面给出算法3的正确性分析和复杂度分析.

算法3的正确性分析: 设输出的属性约简 $R = \{r_1, r_2, \dots, r_l\} (l \leq s)$, 其中 $r_1, r_2, \dots, r_l$ 是按属性约简算法中的顺序排列的. 算法终止的条件是 $\text{Num} = |U'|(|U'| - 1)$. 由算法3知, 当 $\text{Num} = |U'|(|U'| - 1)$ 时, 有

$$\begin{aligned} |U'|(|U'| - 1) &= \text{Sig}_{R-\{r_l\}}(r_l) + \text{Sig}_{R-\{r_{l-1}, r_l\}}(r_{l-1}) + \dots + \text{Sig}_{R-\{r_2, r_3, \dots, r_l\}}(r_2) + \text{Sig}_{\emptyset}(r_1) \\ &= \text{SDis}(R) - \text{SDis}(R - \{r_l\}) + \text{SDis}(R - \{r_l\}) - \text{SDis}(R - \{r_{l-1}, r_l\}) + \dots + \text{SDis}(\{r_1, r_2\}) - \text{SDis}(\{r_1\}) + \text{SDis}(\{r_1\}) - \text{SDis}(\emptyset) \\ &= \text{SDis}(R) - \text{SDis}(\emptyset) = \text{SDis}(R). \end{aligned}$$

即算法终止时有 $\text{SDis}(R) = |U'|(|U'| - 1) = \text{SDis}(C)$. 由定理4的证明知有: $U/R = U/C$. 故 R 是 C 的属性约简.

算法3的复杂度分析: 由算法1知, 算法3第1步的时间复杂度为 $O(|C| |U|)$, 空间复杂度为 $O(|U|)$; 第3步是求出每一个属性的重要性, 由算法2的分析知, 求一个属性重要性的最坏时间复杂度和空间复杂度都是 $O(|U'|) = O(|U/C|)$, 故第3步求属性重要性的最坏时间复杂度为 $O(|U/C| |C - R|)$, 最坏空间复杂度为 $O(|U/C| |C - R|)$; 第3至第5步总的最坏时间复杂度为 $O(|U/C| |C|) + O(|U/C| (|C| - 1)) + \dots + O(|U/C|) = O(|U/C| |C|^2)$, 最坏的空间复杂度为 $O(|U/C| |C|)$. 故算法3的最坏时间算法复杂度为 $\max(O(|C| |U|), O(|C|^2 |U/C|))$, 最坏的空间复杂度为 $\max\{O(|U|), O(|C| |U/C|)\}$.

5 结束语

在基于信息系统的属性约简定义中, 其实质是

在保持分类能力不变的前提下, 尽可能多地去掉冗余属性. 它的核心概念是等价类. 在信息系统 $S = (U, C, V, f)$ 中, 由等价类的性质可知, U/C 中的任一等价类 X 是不可再分割的, 因此它是最小的知识粒度. 根据这一性质, 我们构造了简化的信息系统, 并定义了简化信息系统中属性集的区分对象数, 给出了基于区分对象数的属性约简定义, 同时证明了二种属性定义是等价的. 再由属性集的区分对象数导出属性的重要性, 并给出了快速计算属性重要性的算法. 以属性重要性为启发信息, 设计了一个新的属性约简算法, 其时间复杂度降为 $\max(O(|C| |U|), O(|C|^2 |U/C|))$. 另一方面, 由于我们设计的算法是在简化信息系统上进行的, 因而算法运行时所占的空间较小. 从计算属性重要性的算法中知, 搜索空间是不断地被缩减, 算法终止时, 搜索空间为空集, 这也大大提高了属性约简算法的效率. 实例很好地说明了新属性约简算法的优越性.

参考文献:

- [1] Miao Duoqian, Wang Jue. An information-based algorithm for reduction of knowledge[C]// IEEE ICIPS'97. China: Beijing, 1997: 1155-1158.
- [2] 梁吉业, 曲开社, 徐宗本. 信息系统的属性约简[J]. 系统工程理论与实践, 2001, 21(12): 76-80.
- [3] 刘少辉, 盛秋骛, 史忠植, 等. Rough集高效算法的研究[J]. 计算机学报, 2003, 26(5): 524-529.
- [4] 高学东, 丁军. 一种新的信息系统属性约简算法[J]. 系统工程理论与实践, 2007, 27(1): 131-136.
- [5] 徐章艳, 刘作鹏, 杨炳儒, 等. 一个复杂度为的快速属性约简算法[J]. 计算机学报, 2006, 29(3): 391-399.
- [6] Wang Jue, Wang Ju. Reduction algorithms based on discernibility matrix: the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489-504.

作者简介:

崔巍 男, (1971-), 博士. 研究方向为管理过程优化、数据挖掘.
徐章艳 (1972-), 博士, 副教授. 研究方向为数据挖掘、粗糙集理论及应用.