

基于序关系的快速计算正区域核的算法

徐章艳^{1,2} 舒文豪¹ 钱文彬¹ 杨炳儒²

(广西师范大学计算机系 桂林 541004)¹ (北京科技大学信息工程学院 北京 100083)²

摘要 目前设计基于正区域的求核算法的主要方法是差别矩阵方法。该方法通过搜索差别矩阵的所有差别元素来得到核,故比较耗时。为此,在简化决策表和简化差别矩阵的基础上,若将其对象按条件属性值看成一个数,则对象是有序的。利用这个序,可将具有核属性的差别元素集映射到一个较小的搜索空间上,故只需判断简化差别矩阵的少量差别元素就可以找到核属性集。在此基础上,利用基数排序的思想,设计了一个高效求核算法,其时间复杂度为 $O(|C|^2|U/C|)+O(|C||U|)$,空间复杂度为 $O(|U|)$ 。由于新算法只需判断简化差别矩阵的少量差别元素就可以找到核属性集,故算法的效率得到了改善。

关键词 粗糙集,简化决策表,正区域,核,算法复杂度

中图法分类号 TP18 文献标识码 A

Quick Algorithm for Computing Core of the Positive Region Based on Order Relation

XU Zhang-yan^{1,2} SHU Wen-hao¹ QIAN Wen-bin¹ YANG Bing-ru²

(Department of Computer, Guangxi Normal University, Guilin 541004, China)¹

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)²

Abstract At present, the main method of designing algorithm for computing the core based on the positive region is discernibility matrix. In this method, the core is found by discovering all discernibility elements of discernibility matrix. So this method is very time consuming. On the foundation of the simplified decision table and simplified discernibility matrix, if the objects with condition attribute value of simplified decision table are looked as numbers, they are order. Using the order, the discernibility elements including the core may be turned into a small quantity of research space. So the core based on the positive region can be found only by searching a small quantity of discernibility elements of the simplified discernibility matrix. On this condition, an efficient computing core was designed by integrating the idea of radix sorting. The time complexity and space complexity are $O(|C|^2|U/C|)+O(|C||U|)$ and $O(|U|)$ respectively. Since the core can be found by searching a small quantity of discernibility elements of the simplified discernibility matrix in the new algorithm, the efficiency of the new algorithm is improved.

Keywords Rough set, Simplified decision table, Positive region, Core, Algorithm complexity

粗糙集理论是 Pawlak^[1]于1982年提出的,目前,被广泛应用于人工智能、模式识别、数据挖掘和智能决策等领域。在粗糙集理论中,属性约简和核^[2,3]是重要研究内容之一。在很多属性约简算法中,一般都要先求出核,然后再由核通过启发式知识扩展到最小约简。此外,核可有效地用于多变量决策树的构造^[4,5]。因此,提高求核算法的效率是一件很有意义的工作。

求核算法最早是由 Skowron 和 HU 提出的^[2,3]。该算法首先要求出 Skowron 差别矩阵,然后再求核。该算法的时间复杂度为 $O(|C||U|^2)$,但是其求出的核与基于正区域的核是不一致的^[6-8]。叶东毅教授等给出了基于正区域的差别矩阵^[7],并证明了利用该差别矩阵设计的求核算法所求出的核

与基于正区域的核是一致的,其算法的时间复杂度仍为 $O(|C||U|^2)$ ^[7,8]。文献[9]基于决策表的一致性,定义了“关键属性”的概念,并在此基础上,结合快速排序的思想,提出一种新的计算正区域的核的算法,其时间复杂度为 $O(|C|^2|U|\log|U|)$,但它要求决策表是一致的,无法用于不一致的决策表。文献[10]提出了基于正区域的简化差别矩阵,为设计效率更好的基于正区域的属性约简算法提供了一种新的思路,但是该文献的简化差别矩阵不是最简化的。为提高属性约简和求核的算法效率,我们在文献[11-14]中引入了简化决策表和相应的简化差别矩阵、简化二进制差别矩阵,并用基数排序的思想设计了一个快速求划分 U/C 的算法,在此基础上设计了两个基于正区域的求核算法,其时间复杂度均为 $\max\{O(|C||U|$

到稿日期:2009-08-14 返修日期:2009-10-13 本文受国家自然科学基金重点项目(69835001),广西教育厅基金项目(200807MS015),西师大博士启动基金项目资助。

徐章艳(1972—),男,博士,副教授,主要研究方向为粗糙集理论与应用、决策树、数据挖掘;舒文豪(1985—),女,硕士生,主要研究方向为粗糙集理论与应用;钱文彬(1984—),男,硕士生,主要研究方向为粗糙集理论与应用;杨炳儒(1943—),男,教授,博士生导师,主要研究方向为知识发现与智能系统、柔性建模与集成技术。

$\log|U|), O(|C||U||POS_C(D)|)$ 。为进一步提高求核算法的效率,我们又提出区分对象对的概念,以区分对象对为基础,设计了一个新的求核算法,其时间复杂度为 $\max\{O(|U||C|), O(|Dis(c_n)||C|), O(|Dis(c_w)||C|)\}$, 它优于 $\max\{O(|C||U|\log|U|), O(|C||U||POS_C(D)|)\}^{[14]}$ 。

在进一步的研究中,若将文献[11-14]中得到的简化决策表中的对象按其条件属性的取值看成一个数,则这个数按条件属性值有序。利用这个序,只要计算简化差别矩阵中少量的差别元素即可求出核。其思想是利用简化决策表中的对象关于属性值有序和核在简化差别矩阵中的差别元素个数为1这两个性质,得到一个新的性质:只需判断简化差别矩阵的少量差别元素就可以找到核属性集。利用这个新的性质,并结合基数排序的思想,我们设计了一个新的算法,其时间复杂度为 $O(|C|^2|U|)$ 。对于大型决策表,往往有 $|U| \gg |C|$, 故新算法在处理大型决策表时较好。而且由于新算法只判断简化差别矩阵的少量差别元素就可以找到核属性集,故新算法的效率较文献[14]的效率好。最后用一个实例说明了新算法的高效性。

1 基于正区域的简化决策表

定义 1^[1] 五元组 $S=(U, C, D, V, f)$ 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,称为论域; C 表示条件属性的非空有限集; D 表示决策属性的非空有限集; $C \cap D = \emptyset$; $V = \bigcup_{a \in C \cup D} V_a$, 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都赋予一个信息值,即 $\forall a \in C \cup D, x \in U$, 有 $f(x, a) \in V_a$, 每一个属性子集 $P \subseteq (C \cup D)$ 都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,用 $U/IND(P)$ 表示,简记为 U/P , 划分 U/P 中的任何元素 $[x]_P = \{y \mid \forall a \in P, f(x, a) = f(y, a)\}$ 称为等价类。

定义 2^[1] 在决策表 $S=(U, C, D, V, f)$ 中, $\forall R \subseteq C \cup D$, $X \subseteq U$, 记 $U/R = \{R_1, R_2, \dots, R_t\}$, 则称 $R_-(X) = \bigcup \{R_i \mid R_i \in U/R, R_i \subseteq X\}$ 为 X 关于 R 的下近似集。

定义 3^[1] 在决策表 $S=(U, C, D, V, f)$ 中, 设 $U/D = \{D_1, D_2, \dots, D_k\}$ 表示由决策属性集 D 对论域 U 的划分, $U/C = \{C_1, C_2, \dots, C_m\}$ 表示由条件属性集 C 对论域 U 的划分, 其中 $C_i (i=1, 2, \dots, m)$ 称为基本块, 称 $POS_C(D) = \bigcup_{D_i \in U/D} C_-(D_i)$ 为 C 关于 D 的正区域。

定义 4^[11] 设在决策表 $S=(U, C, D, V, f)$ 中, 记 $U/C = \{[x'_1]_C, [x'_2]_C, \dots, [x'_m]_C\}$, $U' = \{x'_1, x'_2, \dots, x'_m\}$ 。设 $POS_C(D) = [x'_1]_C \cup [x'_2]_C \cup \dots \cup [x'_t]_C$, 其中 $\{x'_1, x'_2, \dots, x'_t\} \subseteq U'$ 且 $|[x'_s]_C/D| = 1 (s=1, 2, \dots, t)$; 记 $U'_{pos} = \{x'_1, x'_2, \dots, x'_t\}$, $U'_{neg} = U' - U'_{pos}$; 令 $\forall x \in U'_{neg}$ 有 $f(x, D) = *$; 称 $S'=(U', C, D, V, f)$ 为简化决策表。

定义 5^[12] 在决策表 $S=(U, C, D, V, f)$ 中, 决策表的核 $Core(C) = \{a \mid a \in C, POS_{C-\{a\}}(D) \neq POS_C(D)\}$ (基于正区域的核定义)。

定义 6^[12] 在决策表 $S=(U, C, D, V, f)$ 中, 其简化决策表为 $S'=(U', C, D, V, f)$, 定义该简化决策表的差别矩阵为 (简称简化差别矩阵): $M^S = (m_{ij})$, 其元素定义如下:

$$m_{ij} = \begin{cases} \{a \mid a \in C, f(x_i, a) \neq f(x_j, a), \\ \text{当 } x_i, x_j \text{ 只有一个在 } U'_{pos} \text{ 中}; \\ f(x_i, a) \neq f(x_j, a), f(x_i, D) \neq \\ f(x_j, D), \text{当 } x_i, x_j \text{ 都在 } U'_{pos} \text{ 中}\} \\ \emptyset; \text{否则} \end{cases}$$

定义 7^[12] 在决策表 $S=(U, C, D, V, f)$ 中, 其简化决策表为 $S'=(U', C, D, V, f)$, $M'=(m'_{ij})$ 表示简化差别矩阵, 称 $SDCore(C)$ 为简化差别矩阵的核, $SDCore(C)$ 定义如下:

$$SDCore(C) = \{a \mid \exists m'_{ij} = \{a\}, m'_{ij} \in M'\}.$$

定理 1^[12] 在决策表 $S=(U, C, D, V, f)$ 中, $SDCore(C) = Core(C)$ 。

2 求简化决策表的快速算法

求简化决策表的计算时间主要花在计算不可区分关系 U/C 上。一般求划分 U/C 算法的时间复杂度为 $O(|C||U|^2)$ 。对计算 U/C 的方法进行深入研究后, 利用基数排序的思想, 给出了一个计算 U/C 的时间复杂度为 $O(|C||U|)$ 的算法^[11]。

算法 1 求简化决策表

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

输出: U'

- 对每一个 $c_i (i=1, 2, \dots, r)$ 统计 $f(x_j, c_i) (j=1, 2, \dots, n)$ 的最大和最小值, 分别记为 M_i 和 m_i ;
- 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 ;
- for ($i=1; i \leq r+1; i++$)
 - 第 i 趟“分配”: 建立 $M_i - m_i + 1$ 空对列, 令 $front_k$ 和 $end_k (k=0, 1, 2, \dots, M_i - m_i)$ 分别为第 k 个对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 $f(x, c_i) - m_i$ 个对列中去。
 - 第 i 趟“收集”: 表头指针指向第一个非空对列的头指针, 修改每一个非空对列的尾指针, 令其指向下一个非空对列的对头对象, 这样将 $M_i - m_i + 1$ 个对列重新组成一个链表;
- 设由第 3 步得到的对象序列为 $x'_1, x'_2, \dots, x'_n$;
 $t=1; B_t = \{x'_1\};$
 for ($j=2; j \leq n+1; j++$)

若任一 $c_i \in C (i=1, 2, \dots, r)$ 均有 $f(x'_j, c_i) = f(x'_{j-1}, c_i)$, 则

 $B_t = B_t \cup \{x'_j\};$

否则 $t=t+1; B_t = \{x'_j\};$
- $U' = \emptyset;$
 for ($i=1; i \leq t+1; i++$)

若 B_i 中所有对象在决策属性上取值均相同, 则取出 B_i 中的第一个对象并入 U' ;

否则将 B_i 中的第一个对象 x 并入 U' ; 令 $f(x, D) = *$;

算法 1 的时间复杂度为 $O(|C||U|)^{[11,12]}$ 。

3 快速求核算法

在第 2 节中, 对 $\forall x \in U'$, 记 $F(x) = f(x, c_r) f(x, c_{r-1}) \dots f(x, c_1)$, 若将 $f(x, c_i) (i=r, r-1, \dots, 2, 1)$ 看成一个数的第 i 位, 则 U' 中的对象是按其条件属性取值单调增加有序的, 如后面表 2 所列 (表 2 是决策表 1 的简化决策表), $F(X1) = 1212 < F(X6) = 1213 < F(X2) = 1222$ 等。也就是说简化决策表中的对象取值是按某种方式有序, 故有简化决策表的性质

如下:

性质 1 在决策表 $S=(U, C, D, V, f)$ 中, 其简化决策表 $S'=(U'=\{y_1, y_2, \dots, y_{|U/C|}\}, C, D, V, f)$ 是由条件属性的某种序得到的, 则有 $F(y_1) < F(y_2) < \dots < F(y_{|U/C|})$ 。

根据这一特性, 我们提出如下新的核性质。

定理 2 在决策表 $S=(U, C, D, V, f)$ 中, 其简化决策表 $S'=(U'=\{y_1, y_2, \dots, y_{|U/C|}\}, C, D, V, f)$ 是按序 $c_1 < c_2 < \dots < c_r$ 得到的, 若 $c_1 \in \text{Core}(C)$, 则存在 $y_i \in U' (i=1, 2, \dots, |U/C|-1)$ 使得 $m_{i,i+1} \in M^S \wedge m_{i,i+1} = \{c_1\}$ 。

证明: 由于 $c_1 \in \text{Core}(C)$, 故由定理 1 知, $\text{Core}(C) = \text{SD-}\text{Core}(C)$, 由简化差别矩阵的核定义知, 一定有 $m_{i_0 j_0} \in M^S \wedge m_{i_0 j_0} = \{c_1\} (i_0 < j_0)$ 。当 $j_0 = i_0 + 1$ 时, 只需令 $i = i_0$, 命题即成立。

当 $j_0 > i_0 + 1$ 时, 由于 $m_{i_0 j_0} \in M^S \wedge m_{i_0 j_0} = \{c_1\} (i_0 < j_0)$, 故有 $f(y_{i_0}, c_i) = f(y_{j_0}, c_i) (i=r, r-1, \dots, 2)$ 且 $f(y_{i_0}, c_1) < f(y_{j_0}, c_1)$ (由性质 1 可得), 同时 $y_{i_0}, y_{j_0} \in U'_{\text{pos}} \wedge f(y_{i_0}, D) \neq f(y_{j_0}, D)$ 或者 y_{i_0} 和 y_{j_0} 有且仅有一个在 U'_{pos} 中, 而另一个在 U'_{neg} 。

由性质 1 可知, 对 $\forall k (i_0 < k < j_0)$ 均有: $f(y_{i_0}, c_1) < f(y_k, c_1) < f(y_{j_0}, c_1)$ 且 $f(y_{i_0}, c_i) = f(y_{j_0}, c_i) (i=r, r-1, \dots, 2)$ 。当 y_{i_0} 和 y_{j_0} 有且仅有一个在 U'_{pos} 中, 而另一个在 U'_{neg} , 不妨设 $y_{i_0} \in U'_{\text{pos}}$ 且 $y_{j_0} \in U'_{\text{neg}}$, 则由简化差别矩阵的定义知: 对 $\forall k (i_0 \leq k < j_0)$ 均有 $m_{k, k+1} \in M^S \wedge m_{k, k+1} = \{c_1\}$ 。当 $y_{i_0}, y_{j_0} \in U'_{\text{pos}} \wedge f(y_{i_0}, D) \neq f(y_{j_0}, D)$ 时, 则一定有 $\exists k (i_0 \leq k < j_0)$ 使得 $f(y_k, D) \neq f(y_{k+1}, D)$, 否则对于 $\forall k (i_0 \leq k < j_0)$ 均有 $f(y_k, D) = f(y_{k+1}, D)$, 从而有 $f(y_{i_0}, D) = f(y_{j_0}, D)$, 但是 $f(y_{i_0}, D) \neq f(y_{j_0}, D)$ 矛盾, 即 $\exists k (i_0 \leq k < j_0)$ 使得 $f(y_k, D) \neq f(y_{k+1}, D)$ 。另一方面有: $f(y_{i_0}, c_1) < f(y_k, c_1) < f(y_{j_0}, c_1)$ 且 $f(y_{i_0}, c_i) = f(y_{j_0}, c_i) (i=r, r-1, \dots, 2)$, 故由简化差别矩阵的定义知: $m_{k, k+1} \in M^S \wedge m_{k, k+1} = \{c_1\}$ 。

综上所述, 命题成立。

注: 该定理(核的性质)说明, 要在简化决策表中寻找核属性, 只需比较 $f(y_i, c_1)$ 和 $f(y_{i+1}, c_1)$ 的值, 并且若 $f(y_i, c_1) \geq f(y_{i+1}, c_1)$, 则对象 y_i 和 y_{i+1} 不可能产生核属性 c_1 , 换句话说, 若 c_1 为核属性, 那么它一定能在对象 y_i 和 $y_{i+1} (f(y_i, c_1) < f(y_{i+1}, c_1))$ 中产生。这为我们快速判断 c_1 是否为核属性提供了理论基础。

为了判断其它属性是否为核属性, 给出如下的定理。

定理 3 在决策表 $S=(U, C, D, V, f)$ 中, 其简化决策表 $S'=(U'=\{y_1, y_2, \dots, y_{|U/C|}\}, C, D, V, f)$ 是按序 $c_1 < c_2 < \dots < c_r$ 得到的, 设 $A = \{c_1, c_2, \dots, c_{i-1}\}, U'/A = \{A_1, A_2, \dots, A_i\}$, 若 $c_i \in \text{Core}(C) (i=2, 3, \dots, r)$, 则一定存在 $A_{i_0} \in U'/A$ 使得 $y_k, y_{k+1} \in A_{i_0}$ 满足 $m_{k, k+1} \in M^S \wedge m_{k, k+1} = \{c_i\}$ 。

证明: 由于 $c_i \in \text{Core}(C)$, 故由定理 1 知, $\text{Core}(C) = \text{SD-}\text{Core}(C)$, 由简化差别矩阵的核定义知, 一定有 $m_{i_0 j_0} \in M^S \wedge m_{i_0 j_0} = \{c_i\} (i_0 < j_0)$, 故有 $f(y_{i_0}, c_k) = f(y_{j_0}, c_k) (k \in \{r, r-1, \dots, 2, 1\} - \{i\})$ 且 $f(y_{i_0}, c_i) < f(y_{j_0}, c_i)$ (由性质 1 可得), 同时 $y_{i_0}, y_{j_0} \in U'_{\text{pos}} \wedge f(y_{i_0}, D) \neq f(y_{j_0}, D)$ 或者 y_{i_0} 和 y_{j_0} 有且仅有一个在 U'_{pos} 中, 而另一个在 U'_{neg} 。显然有 $y_{i_0}, y_{j_0} \in U'/A$ (因为 $f(y_{i_0}, c_k) = f(y_{j_0}, c_k) (k=1, 2, \dots, i-1)$)。由性质 1 可知, 对 $\forall k (i_0 < k < j_0)$ 均有: $f(y_{i_0}, c_i) < f(y_k, c_i) < f(y_{j_0}, c_i)$ 且 $f(y_{i_0}, c_i) = f(y_{j_0}, c_i) (k \in \{r, r-1, \dots, 2, 1\} -$

$\{i\})$ 。由定理 2 的证明过程可知命题成立。

注: 定理 3 为判断其它的条件属性提供了理论依据, 即若 $c_i \in \text{Core}(C) (i=2, 3, \dots, r)$, 设 $A = \{c_1, c_2, \dots, c_{i-1}\}$, 则核一定能在 $y_k, y_{k+1} \in A_{i_0} \in U'/A (f(y_k, c_i) < f(y_{k+1}, c_i))$ 的对象中产生。

根据定理 2、定理 3 和算法 1, 可设计如下基于序关系的快速求正区域核的算法。

算法 2 求核算法

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

输出: 决策表的核 $\text{Core}(C)$

1. 由算法 1 求出: $U' = \{y_1, y_2, \dots, y_{|U/C|}\};$
 $\text{core}(C) = \emptyset; A = \emptyset; i = 0;$
2. 对任意 $A_k = \{z_1, z_2, \dots, z_t\} \in U'/A$ 做如下处理:
 - 2.1. 若 U'/A 为空集, 则转向第 4 步; 若 $|A_k|$ 为 1, 则删除 A_k ;
 - 2.2. for ($j = 1; j < t; j++$)

if ($f(z_j, c_i) < f(z_{j+1}, c_i)$)
 if ($(f(z_j, D) = *) \wedge f(z_{j+1}, D) = *)$) 不成立
 { 计算 $m_{j, j+1};$
 若 $|m_{j, j+1}| = 1$, 则 $\text{Core}(C) = \text{Core}(C) \cup \{c_i\};$
 算法转向第 3 步;
 }
3. $i = i + 1$; 若 $i > r$, 转向第 4 步; 否则 $A = A \cup \{c_i\}$; 用算法 1 的思想计算 U'/A , 算法转向第 2 步;
4. 终止算法, 输出核 $\text{Core}(C)$ 。

定理 4 由算法 2 求出的核是基于正区域的核。

证明: 算法的第 1 步求出的是原决策表的简化决策表。在第 2 步中, 2.1 步说明 A_k 只有一个对象, 由定理 3 知, 在 A_k 中不能产生核; 由划分的概念知, 对 A_k 进行进一步划分更不可能产生核, 故可以删除 A_k 。这样可以加快求核算法的速度。由定理 3 知, 2.2 步能求出所有的核属性。故算法 2 能求出基于正区域的核。

算法 2 的复杂度分析: 算法 2 的第 1 步的时间复杂度由算法 1 得, 为 $O(|C| |U|)$; 第 2.2 步的最坏时间复杂度为 $O((|C| - i) |A_k|)$, 故第 2 步总的最坏时间复杂度为 $O((|C| - i) |U'|)$ 。第 3 步的时间复杂度为 $O(|U'|)$ 。第 2 步到第 3 步总的最坏时间复杂度为 $\sum_{i=1}^r O((|C| - i) |U'|) = O(|C|^2 |U/C|)$ 。故新算法的时间复杂度为 $O(|C|^2 |U/C|) + O(|C| |U|)$, 易知算法 2 的空间复杂度为 $O(|U|)$ 。

4 实例

为了更好地说明新算法的高效性, 以决策表 1 为例来说明新算法。

表 1 决策表 1

U	a	b	c	d	D
X1	2	1	2	1	0
X2	2	2	2	1	1
X3	2	1	2	1	0
X4	2	3	2	3	0
X5	2	2	2	1	1
X6	3	1	2	1	0
X7	1	2	3	2	2
X8	4	3	1	2	3
X9	3	1	2	1	1
X10	1	2	3	2	2

X11	3	1	2	1	1
X12	4	3	1	2	3
X13	4	3	4	2	1
X14	1	2	3	2	3
X15	4	3	4	2	2

对决策表 1 的 15 个对象 U , 由算法 2 的第 1 步, 即用算法 1 计算得到 $U/\{a, b, c, d\}$ 如下:

$\{X1, X3\}, \{X6, X9, X11\}, \{X2, X5\}, \{X8, X12\}, \{X7, X10, X14\}, \{X13, X15\}, \{X4\}$.

$U' = \{X1, X6, X2, X8, X7, X13, X4\}$;

得到的简化决策表如表 2 所列。

表 2 决策表 1 的简化决策表

U	a	b	c	d	D
X1	2	1	2	1	0
X6	3	1	2	1	*
X2	2	2	2	1	1
X8	4	3	1	2	3
X7	1	2	3	2	*
X13	4	3	4	2	*
X4	2	3	2	3	0

由算法 2 的第 2 步可知输入为:

$U'/O = \{\{X1, X6, X2, X8, X7, X13, X4\}\}$

由第 2.2 步得到: $f(x_1, a) < f(x_6, a)$, 且由于 $f(x_1, D) = 0 \neq *$, 计算 $m_{1,6} = \{a\}$, 故 $Core(C) = \{a\}$, 算法转向第 3 步。

此时有: $i=1, i < 5$; 故有 $A = \{a\}$, 用算法 1 的思想计算 $U'/\{a\} = \{\{X1, X2, X4\}, \{X6\}, \{X8, X13\}, \{X7\}\}$ 。算法转向第 2 步。

对于 $\{X1, X2, X4\} \in U'/\{a\}$, 由算法 2 的第 2.2 步得到: $f(x_1, b) < f(x_2, b)$, 且由于 $f(x_1, D) = 0 \neq *$, 计算 $m_{1,2} = \{b\}$, 故 $Core(C) = \{a, b\}$, 算法转向第 3 步。

此时有: $i=2, i < 5$; 故有 $A = \{a, b\}$, 用算法 1 的思想计算 $U'/\{a, b\} = \{\{X1\}, \{X2\}, \{X4\}, \{X6\}, \{X8, X13\}, \{X7\}\}$ 。算法转向第 2 步。

由算法 2 的第 2.1 步, 删除了 $\{X1\}, \{X2\}, \{X4\}, \{X6\}$ 。由算法 2 的第 2.2 步知: 对于 $\{X8, X13\} \in U'/\{a, b\}$, $f(x_8, c) < f(x_{13}, c)$, 且 $f(x_8, D) = 3 \neq *$, 计算 $m_{8,13} = \{c\}$, 故 $Core(C) = \{a, b, c\}$, 算法转向第 3 步。

此时有: $i=3, i < 5$; 故有 $A = \{a, b, c\}$, 用算法 1 的思想计算 $U'/\{a, b, c\} = \{\{X8\}, \{X13\}, \{X7\}\}$ 。算法转向第 2 步。

由算法 2 的第 2.1 步, 删除了 $\{X8\}, \{X13\}, \{X7\}$ 。由算法 2 的第 2.2 步知: 无计算对象, 算法转向第 3 步。

此时有: $i=4, i < 5$; 故有 $A = \{a, b, c, d\}$, 用算法 1 的思想计算 $U'/\{a, b, c, d\} = \emptyset$ 。算法转向第 2 步。

由算法 2 的由第 2.1 步知: 算法终止, 核 $Core(C) = \{a, b, c\}$ 。

由实例可知, 在新的算法中, 要计算的简化差别矩阵元素只占简化差别矩阵的一小部分, 因而新求核算法的效率较文

献[12-14]的效率高。

结束语 在现有基于正区域的求核算法中, 常利用差别矩阵来设计求核算法, 在这种方法中, 一般总是先求出所有的差别元素, 再从这些差别元素中去求核。这种方法简单易懂, 但计算量较大。为降低求核算法的时间复杂度, 经研究发现简化差别矩阵的对象按条件属性值有序, 利用这个序和基数排序的思想, 将具有核属性的差别元素集映射到一个较小的搜索空间上, 这样新算法只需判断简化差别矩阵中极小部分差别元素, 即可找到所有的核属性集。新算法的时间复杂度为 $O(|C|^2|U/C|) + O(|C||U|)$, 空间复杂度为 $O(|U|)$ 。该文的思想也为我们下一步设计高效的属性约简算法提供了一个新的思路与视角。

参考文献

- [1] Pawlak Z. Rough Sets[J]. International Journal of Computer and information Science, 1982, 11(5): 341-356
- [2] Skowron A, Rauszer C. The Discernibility Matrices and Functions in Information Systems[CM// Slowinski R, ed. Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, 1992: 331-362
- [3] Hu X H, Cercone N. Learning in relational databases: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-337
- [4] 苗夺谦, 王珏. 基于粗糙集的多变量决策树的构造方法[J]. 软件学报, 1997, 8(6): 425-431
- [5] 张霞, 施国生, 梁道雷. 基于粗糙集的多变量决策树方法研究[J]. 浙江理工大学学报, 2008, 25(4): 438-441
- [6] 徐章艳, 杨炳儒, 宋威, 等. 几种不同属性约简的比较[J]. 小型微型计算机系统, 2008, 29(5): 848-853
- [7] 叶东毅, 陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报, 2002, 30(7): 1086-1088.
- [8] 叶东毅, 陈昭炯. 一个新的二进制可辨识矩阵及其核的计算[J]. 小型微型计算机系统, 2004, 25(6): 965-967
- [9] 赵军, 王国胤, 吴中福, 等. 一种高效的属性核计算方法[J]. 小型微型计算机系统, 2003, 24(11): 1950-1953
- [10] 杨明, 孙志挥. 改进的差别矩阵及其求核方法[J]. 复旦学报: 自然科学版, 2004, 43(5): 865-868
- [11] 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
- [12] 徐章艳, 杨炳儒, 蔡卫东, 等. 一个基于正区域的快速求核算法[J]. 系统工程与电子技术, 2006, 28(12): 1902-1905
- [13] 徐章艳, 杨炳儒, 宋威. 基于简化的二进制差别矩阵的快速求核算法[J]. 小型微型计算机系统, 2006, 27(9): 1711-1714
- [14] 徐章艳, 杨炳儒, 宋威, 等. 基于区分对象对集的快速求核算法[J]. 系统工程与电子技术, 2008, 30(4): 731-734