第41卷 第6期 西 安 交 通 大 学 学 报 Vol.41 №6
2007年6月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Jun. 2007

一种新的快速求核算法

周江卫，冯博琴，刘洋
（西安交通大学计算机科学与技术系，710049，西安）

摘要：为了解决核影响属性约简算法的速度和效率等问题，提出了一种基于正区域的求核算法.采用基数排序思想计算正区域，分别得到决策属性正区域的条件属性集和除决策属性正区域的一个条件属性之外的条件属性集，并且计算这2种属性集的基数之差，以判断该条件属性是否是核属性，依次判断所有条件属性，从而快速获得所需要的核.基于正区域求核算法的时间复杂度为$O(|C||U|)$.实验结果表明，利用该算法求核，所耗时间将随对象数的增加呈线性增长，且当对象数最大时，求核所耗时间仅为对比算法的0.6%，同时证明了该算法对各种数据集均有很好的适应性.

关键词：属性约简；基数排序；正区域；核
中图分类号：TP18 **文献标识码**：A **文章编号**：0253-987X(2007)06-0688-04

New Algorithm for Quick Computing Core

Zhou Jiangwei, Feng Boqin, Liu Yang
(Department of Computer Science and Technology, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: In order to improve the speed and efficiency of reduction algorithm of core influence attributes, a new core computation algorithm based on positive region is provided. The positive region based on radix sorting is used to get the positive region condition attributes set of decision attributes and the condition attribute set which excludes one of the condition attributes of decision attribute positive region. Then the difference between the two radices of positive regions is calculated to judge whether the condition attribute is a core attribute, thereby all the condition attributes are judged and the required core is quickly acquired. The time complexity of the proposed algorithm is $O(|C||U|)$. Experimental results show that the core computation time increases linearly with the increasing number of entries, and the computation time is only 0.6% of the contrastive algorithm when the entries is maximum. Meanwhile, it is shown that the algorithm is well suitable for various kinds of data sets.

Keywords: attribute reduction; radix sorting; positive region; core

粗糙集理论[1]于20世纪80年代由波兰科学家Pawlak提出，它是一种处理不精确、不完全与不相融的数学工具.在粗糙集理论中，属性约简是重要的研究内容之一.研究者们提出了许多属性约简算法，而大部分属性约简算法是在核的基础上进行的[2]，因而快速得到核以提高属性约简算法的效率至关重要.

求核方法一般是先求出决策表中的所有不可缺少属性，如基于差别矩阵的求核算法[3-5]，其复杂度均比较高，基于简化差别矩阵的求核算法[6]，其复杂度为$\max(O(|C||U/C|^2),O(|C||U|))$，基于改进的差别矩阵的核增量式更新算法[5]，其复杂度为$O(|C||U|^2)$，这样的时间复杂度是不理想的.

文献[7]中提出了基数排序思想，使得求取正区域的时间复杂度为$O(|C||U|)$.本文在基数排序思想的基础上，首先对所有对象按条件属性依次排序，

收稿日期：2006-10-13. 作者简介：周江卫(1975～)，男，博士生；冯博琴(联系人)，男，教授，博士生导师. 基金项目：国家高技术研究发展计划资助项目(2006AA01Z210).

计算正区域 $POS_C(D)$以及所有的 $POS_{C-\{c_i\}}(D)$，然后得到 $POS_C(D)$与每个 $POS_{C-\{c_i\}}(D)$之间的基数之差，从而得到所有不可缺少的属性，进而获得所需的核.

1　粗糙集相关概念

定义 1[1]　决策表 $S=(U,C,D,V,f)$，其中：$U=\{x_1,x_2,\cdots,x_n\}$是论域；$C=\{c_1,c_2,\cdots,c_r\}$为条件属性集；$D=\{d_1,d_2,\cdots,d_l\}$为决策属性集；$f:U\times(C\cup D)\to V$是信息函数，$F=C\cup D$，$C\cap D=\varnothing$；$V=\bigcup V_a$，$a\in F$，V_a表示属性 a 的值域. 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

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

而 $\mathrm{IND}(P)$构成了 U 的一个划分，用 $U/\mathrm{IND}(P)$表示，简记为 U/P，　U/P 中的任何元素 $[x]_P=\{y\mid \forall a\in P, f(x,a)=f(y,a)\}$称为等价类.

定义 2[8]　设 $X\subseteq U$ 为论域的一个子集，$P\subseteq C$，X 的关于 P 的下近似为

$$\underline{P}X=\{x\in U:[x]_P\subseteq X\}$$

式中：$[x]_P$ 表示 U 中所有与 x 在关系 $\mathrm{IND}(P)$下由等价的元素构成的集合.

定义 3　在 $S=(U,C,D,V,f)$中，用 $U/D=\{D_1,D_2,\cdots,D_k\}$表示由决策属性 D 对论域 U 的划分，$U/P=\{P_1,P_2,\cdots,P_m\}$表示条件属性集 $P(P\subseteq C)$对论域 U 的划分，则称 $POS_P(D)=\bigcup\limits_{D_i\in U/D}\underline{P}(D_i)$为 P 关于 D 的正区域.

定义 4[8]　设 $P\subseteq C$，则对划分$\{\psi_1,\psi_2,\cdots,\psi_k\}$的 P-近似精度 $\gamma_P=\sum\limits_{i=1}^{k}|\underline{P}\psi_i|/|U|$.

定义 5[8]　设 $P\subseteq C$，若 $\gamma_P=\gamma_C$，且不存在 $R\subseteq P$，使得 $\gamma_R=\gamma_C$，则称 P 为 C 的一个(相对于决策属性 D 的)属性约简，所有 C 的属性约简的交称为 C 的核，记为 $\mathrm{Core}(C)$.

定义 6[8]　如果属性 $c\in C$ 满足 $\gamma_{C-\{c\}}<\gamma_C$，则称属性 c 为不可缺少的，否则称属性 c 为冗余的.

性质 1[8]　属性 $c\in\mathrm{Core}(C)$，当且仅当 c 是不可缺少的属性.

2　基于正区域的求核算法

定理 1　对于条件属性 $c_i\in C(i=1,2,\cdots,r)$，$c_i\in\mathrm{Core}(C)$的充要条件是

$$|POS_C(D)|-|POS_{C-\{c_i\}}(D)|>0$$

证明　由性质 1 可知，属性 $c_i\in\mathrm{Core}(C)$，当且仅当 c_i 是不可缺少的属性. 由定义 6 知，若属性 c_i 为不可缺少的属性，须满足 $\gamma_{C-\{c_i\}}<\gamma_C$.

根据定义 4，$\gamma_{C-\{c_i\}}=\sum\limits_{i=1}^{k}|\underline{(C-\{c_i\})}\psi_i|/|U|=|POS_{C-\{c_i\}}(D)|/|POS_C(D)|$，从而 $\gamma_C-\gamma_{C-\{c_i\}}=1-|POS_{C-\{c_i\}}(D)|/|POS_C(D)|$. 要使 $\gamma_{C-\{c_i\}}<\gamma_C$，即要 $\gamma_C-\gamma_{C-\{c_i\}}=1-|POS_{C-\{c_i\}}(D)|/|POS_C(D)|>0$，亦即 $|POS_C(D)|-|POS_{C-\{c_i\}}(D)|>0$.

根据定理 1，若要判断 $c_i\in C(i=1,2,\cdots,r)$是否是核属性，就要计算 $|POS_C(D)|$与 $|POS_{C-\{c_i\}}(D)|$之间的差值. 若差值大于 0，则 c_i 是核属性，否则就不是核属性. 文献[7]给出了基于基数排序计算 $POS_P(D)(P\subseteq C)$的算法，复杂度为 $O(|C||U|)$，故用基数排序算法计算 $|POS_C(D)|$、$|POS_{C-\{c_i\}}(D)|$的时间复杂度均为 $O(|C||U|)$. 若对每个 $c_i\in C(i=1,2,\cdots,r)$分别计算 $|POS_{C-\{c_i\}}(D)|$，则计算所有 $c_i\in C(i=1,2,\cdots,r)$的时间复杂度为 $O(|C|^2\cdot|U|)$，这并不是所期望的.

在计算 $|POS_{C-\{c_i\}}(D)|$时，要对 U 中的对象依 $c_1,c_2,\cdots,c_{i-1},c_{i+1},\cdots,c_r$ 进行排序，计算 $|POS_{C-\{c_{i+1}\}}(D)|$时，要对 U 中的对象依 $c_1,c_2,\cdots,c_i,c_{i+2},\cdots,c_r$ 进行排序. 在这 2 次排序中可以发现，有 $r-2$ 次排序操作是重复的，只有一次操作与之不同，所以依 $c_1,c_2,\cdots,c_i,c_{i+2},\cdots,c_r$ 进行排序时，可以从上次排序的结果出发，只要依 c_i 进行一次排序就可以得到本次的排序结果，从而对所有 $c_i\in C(i=1,2,\cdots,r)$进行排序的时间复杂度仅为 $O(2|C|\cdot|U|)=O(|C||U|)$.

2.1　基数排序算法

对于待排序的 U 和 $c_i\in C$，统计 $f(x_j,c_i)(j=1,2,\cdots,n)$在 U 中的最大值 M_i 和最小值 m_i，以静态链表依次存储对象 $x_1,x_2,\cdots,x_n$，并令表头指针指向 x_1. 在分配时，建立 M_i-m_i+1 个空队列，令 front_k和 $\mathrm{end}_k(k=0,1,\cdots,M_i-m_i)$分别为第 k 个队列的头指针和尾指针. 将链表中的对象 $x\in U$ 按链表的次序分配到 $f(x,c_i)-m_i$ 个队列中. 在收集时，表头指针指向第 1 个非空队列的头指针，修改每一个非空队列的尾指针，并令其指向下一个非空队列的对头对象，这样表头指针指向的链表就是经过排序的链表. 由于最大值 M_i 和最小值 m_i 是统计结果，所以每个队列总不会为空，同时用 $S(U,c_i)$表示依条件属性 $c_i\in C$ 排序后的链表. 基数排序算法的复杂度主要包括 2 部分，第 1 部分是在分配时的时

间复杂度 $O(|U|)$，第 2 部分是在收集时的时间复杂度 $O(M_i-m_i+1)$，故总的时间复杂度为 $O(|U|)+O(M_i-m_i+1)$. 通常情况下，M_i-m_i+1 远小于 $|U|$，所以复杂度为 $O(|U|)$.

2.2 基于正区域求核算法

用 $S(U,P)$ 表示 U 中的对象依 $P(P\subseteq C)$ 的条件属性依次排序后的结果.

定理 2 $S(U,C-\{c_{i+1}\})$ 等价于 $S(U,C-\{c_i\})\cdot S(U,\{c_i\})$，其中 $c_i\in C$.

证明 $S(U,C-\{c_i\})=S(U,\{c_{i+1}\})\cdots S(U,\{c_r\})\cdot S(U,\{c_1\})S(U,\{c_2\})\cdots S(U,\{c_{i-1}\})$，而 $S(U,C-\{c_{i+1}\})=S(U,\{c_{i+2}\})\cdots S(U,\{c_r\})S(U,\{c_1\})\cdot S(U,\{c_2\})\cdots S(U,\{c_i\})=\overline{S}(U,\{c_{i+1}\})\ S(U,C-\{c_i\})S(U,\{c_i\})$，其中 $\overline{S}(U,\{c_{i+1}\})$ 表示 $S(U,\{c_{i+1}\})$ 对应的无序状态. 在计算 $\mathrm{POS}_{C-\{c_{i+1}\}}(D)$ 时，由于不关心条件属性 c_{i+1} 的排列顺序，所以 $S(U,C-\{c_{i+1}\})$ 与 $S(U,C-\{c_i\})S(U,\{c_i\})$ 等价.

性质 2 $S(U,C-\{c_1\})$ 等价于 $S(U,C)$.

定理 3 对于 $x,y\in U$，如果 $f(x,D)\neq f(y,D)$，且 $\forall c_i\in P$ 均有 $f(x,c_i)=f(y,c_i)$，那么 x、y 均不属于 P 关于 D 的正区域.

证明 由已知 $\forall c_i\in P$ 均有 $f(x,c_i)=f(y,c_i)$，则有 $x,y\in[x]_P$，而由已知 $f(x,D)\neq f(y,D)$，则有 x、y 分别属于决策属性 D 对论域 U 的 2 个不同的等价类. 所以，根据定义 3 得 x、y 均不属于 P 关于 D 的正区域.

对于排序后的 $S(U,P)$，如果相邻 2 个对象 x_i、x_{i+1} 不属于 P 的关于 D 的同一个等价类，那么 x_{i+1} 与 x_i 之前的所有对象不属于 P 的关于 D 的同一个等价类，从而只需遍历一次 $S(U,P)$ 且根据定理 3 就可以得到 $|\mathrm{POS}_P(D)|$.

算法 1 基于正区域的快速求核算法.

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

输出：$\mathrm{Core}(C)$.

步骤 1：$\mathrm{Core}(C)=\varnothing$//初始化核.

步骤 2：利用基数排序思想得到 $S(U,C)$，并且计算 $|\mathrm{POS}_C(D)|$.

步骤 3：根据性质 2，$S(U,C-\{c_1\})$ 等价于 $S(U,C)$，计算 $|\mathrm{POS}_{C-\{c_1\}}(D)|$，如果 $|\mathrm{POS}_{C-\{c_1\}}(D)|<|\mathrm{POS}_C(D)|$，则根据定理 1 知 c_1 为核属性，即

$$\mathrm{Core}(C)=\mathrm{Core}(C)\cup\{c_1\}$$

步骤 4：根据定理 2，依次计算 $S(U,C-\{c_i\})$，$i=2,3,\cdots,r$，并且分别计算得到 $|\mathrm{POS}_{C-\{c_i\}}(D)|$，如果 $|\mathrm{POS}_{C-\{c_i\}}(D)|<|\mathrm{POS}_C(D)|$，则 $\mathrm{Core}(C)=\mathrm{Core}(C)\cup\{c_i\}$.

下面分析算法复杂度.

根据步骤 2 计算 $S(U,C)$，其时间复杂度 $O(|C||U|)$，计算 $|\mathrm{POS}_C(D)|$ 其时间复杂度为 $O(|U|)$，步骤 3 的时间复杂度为 $O(|U|)$，步骤 4 是一个 $|C-1|$ 次的循环，其时间复杂度为 $2|C-1|\cdot(O(|U|))$，所以总的复杂度为 $O(|C||U|)+O(|U|)+2|C-1|(O(|U|))=3O(|C||U|)$，故基于正区域的求核算法的时间复杂度为 $O(|C||U|)$.

3 实验结果

为了进一步验证算法的正确性和高效性，本文用 VC 实现了文献[3]中提出的求核算法（以下用算法 A 表示）和本文提出的基于正区域的求核算法（用算法 B 表示），并进行测试，同时利用美国加利福尼亚大学提供的知识库（UCI）[9]（Connect-4 Opening Database）进行实验. 将 Connect-4 Opening Database 看作决策表，依次取其中的前 10 000、20 000、30 000、40 000、50 000、60 000 个对象分别用算法 A 和算法 B 进行求核. 在求核之前，对所有数据进行数值化预处理，即将字母或字符串表示的属性值转换成大于等于 0 的整数属性值，实验结果如图 1 所示.

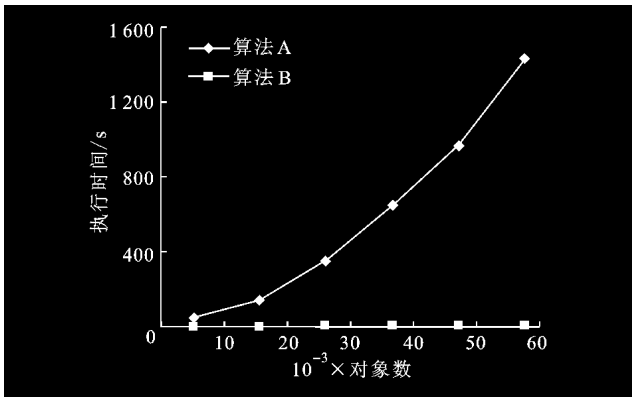

图 1 算法 A 与算法 B 执行时间比较

利用 UCI 知识库[9] 中的 CHESS-END-GAME、DNA、Mushroom、Connect-4 Opening Database 的所有属性和数据对象，以及 Adult 数据中的 8 个离散属性、1 个条件属性和所有数据对象，分别用算法 A 和算法 B 求核. 在求核之前，对所有数据进行数值化预处理. 实验结果如表 1 所示.

表 1　2 种求核算法执行时间比较

决策表	对象数	条件属性个数	算法 A 执行时间/s	算法 B 执行时间/s
DNA	2 000	180	1.359	1.062
CHESS-END-GAME	3 196	36	4.125	0.390
Mushroom	8 124	22	16.672	0.563
Adult	32 561	8	13.469	0.594
Connect-4 Opening Database	67 557	42	1 755.547	9.343

从图 1 和表 1 的实验结果可以看出，本文算法明显优于文献[7]算法. 在属性个数不变的情况下，随着对象个数的增加，本文算法的执行时间呈线性上升趋势，即执行时间与对象个数成正比，而文献[7]算法则呈二次曲线上升趋势，且随着对象个数的增加，执行时间迅速增加. 从表 1 的实验数据和结果可以得到如下结论：本文的算法和对象个数与条件属性个数的乘积成正比，这与算法时间复杂度分析的结果是一致的.

4　结　论

研究发现，粗糙集理论可以应用于医疗数据分析，声音识别和金融、图像处理等许多方面[10]，然而在实际应用中却相当困难，其关键原因是时间复杂度不能满足要求. 属性约简是粗糙集应用的重要方面，有效提高属性约简的效率有助于推广粗糙集理论在各个领域的应用. 相当多的属性约简算法是基于核的，因此本文提出了新的求核算法，以大大提高求核效率. 所提算法的时间复杂度仅与决策表的对象个数和属性个数的乘积成正比，从而克服了以往求核算法的时间复杂度随着对象个数和属性个数的增加呈指数上升的缺陷，因而它更适用于高维属性空间以及大数据集决策表的求核.

参考文献：

[1] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science, 1982,11(5):341-356.

[2] 杨明. 一种基于改进差别矩阵的核增量式更新算法 [J]. 计算机学报，2006,29(3):57-63.
Yang Ming. An incremental updating algorithm of the computation of a core based on the improved discernibility matrix [J]. Chinese Journal of Computers, 2006,29(3):57-63.

[3] 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.
Yang Ming, Sun Zhihui. Improvement of discernibility matrix and the computation of a core [J]. Journal of Fudan University,2004, 43(5): 865-868.

[4] Hu X H, Cercone N. Learning in relational databases: a rough set approach [J]. Computational Intelligence, 1995,11(2):323-338.

[5] 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.
Ye Dongyi, Chen Zhaojiong. A new discernibility matrix and the computation of a core [J]. Acta Electronica Sinica,2002,30(7):1086-1088.

[6] 徐章艳，杨炳儒. 一个基于差别矩阵的快速求核算法 [J]. 计算机工程与应用，2006,42(6):4-6.
Xu Zhangyan, Yang Bingru. Quick computing core algorithm based on discernibility matrix [J]. Computer Engineering and Applications,2006,42(6):4-6.

[7] 徐章艳，刘作鹏. 一个复杂度为 max($O(|C|^2|U/C|)$, $O(|C||U|)$)的快速属性约简算法 [J]. 计算机学报，2006,29(3):391-399.
Xu Zhangyan, Liu Zuopeng. A quick attribute reduction algorithm with complexity of max($O(|C|^2|U/C|)$, $O(|C||U|)$)[J]. Chinese Journal of Computers, 2006,29(3):391-399.

[8] Pawlak Z. Rough set approach to multi-attribute decision analysis [J]. European Journal of Operational Research,1994,72:443-459.

[9] John G, Shapiro A. Connect-4 opening database, chess databases, mushrooms database, adult database [EB/OL]. [2006-07-01]. http://www.ics.uci.edu/~mlearn/MLSummary.html.

[10] Pawlak Z. Some issues on rough sets [J]. Transactions on Rough Sets: I, 2004 (LNCS 3100):1-58.

（编辑　苗凌）