基于二进制可分辨矩阵的快速求核算法

葛 浩[1,3],杨传健[2],李龙澍[3]
GE Hao[1,3],YANG Chuan-jian[2],LI Long-shu[3]

1.滁州学院 电子信息工程系,安徽 滁州 239012
2.滁州学院 计算机系,安徽 滁州 239012
3.安徽大学 计算机学院,合肥 230039
1.Department of Electronic and Information Engineering,Chuzhou University,Chuzhou,Anhui 239012,China
2.Department of Computer Science,Chuzhou University,Chuzhou,Anhui 239012,China
3.School of Computer Science,Anhui University,Hefei 230039,China
E-mail:togehao@126.com

GE Hao,YANG Chuan-jian,LI Long-shu.Quick computing core algorithm based on binary discernibility matrix.Computer Engineering and Applications,2009,45(7):164-166.

Abstract: At present,the algorithms of the computing core have the following shortcomings: the core acquired from these algorithms is not the core based on positive region,the time complexity and space complexity are not good.Aiming at these problems,a binary discernibility matrix and correspondence property of computing core are provided.It is proved that the core acquired from the property is equivalent to the core based on positive region.Then,the computing core algorithm is designed,its time complexity is $\max\{O(|C||U/C|^2),O(|C||U|)\}$,and its space complexity is $O(|C||U/C|^2)$.Finally,an example is given to explain the feasibility and availability of this method.

Key words: rough set;equivalence class;dicernibility matrix;core

摘 要:目前,求核算法存在以下不足:求得的核与正区域的核不一致,求核算法的时间复杂度和空间复杂度不理想。针对上述问题,给出一种二进制可分辨矩阵的定义及其求核性质,并证明了由该性质获得的核与正区域的核是等价的,然后设计求核算法,该算法的时间复杂度为 $\max\{O(|C||U/C|^2),O(|C||U|)\}$,空间复杂度为 $O(|C||U/C|^2)$。最后实例说明该方法的可行性和有效性。

关键词:粗糙集;等价类;可分辨矩阵;核
DOI:10.3778/j.issn.1002-8331.2009.07.049 **文章编号**:1002-8331(2009)07-0164-03 **文献标识码**:A **中图分类号**:TP181

1 引言

粗糙集理论[1]是波兰数学家 Z Pawlak 于 1982 年提出的,它是分析不精确、不一致、不完整信息系统的有力工具。属性约简是粗糙集理论研究的重要内容之一。目前许多约简方法是基于核属性[2-3]的,因此求核是这类方法的关键。

通过核的定义可以求得核属性,但比较繁琐。HU[4]给出一种基于可分辨矩阵的求核方法,该方法比较简单,但时间和空间复杂度均为 $O(|C||U|^2)$。支天云在文献[5]中给出一种基于二进制可分辨矩阵的属性约简算法,该方法可以减少一半的存储空间,但算法时间复杂度和空间复杂度仍均为 $O(|C||U|^2)$。无论是 HU 的方法还是支天云的方法,求得的核属性与正区域算法求得的核属性在某些情况下是不一致的。叶东毅在文献[6]中对支天云的结论提出了质疑,举例证明了该结论的问题,给出一种基于二进制可分辨矩阵的求核算法,但该算法的时间复杂度仍为 $O(|C||U|^2)$。徐章艳[7]给出一个高效的求核算法,该算法的时间复杂度和空间复杂度分别被降为 $\max\{O(|C|(|U_{pos}'||U/C|)),O(|C||U|)\}$和 $\max\{O(|U|),O(|C|(|U_{pos}'||U/C|))\}$,是目前最高效的算法。本文在文献[8]决策表信息系统定义的基础上,给出一种二进制可分辨矩阵的定义形式和求核性质,并证明了由该性质所求得的核与正区域的核是等价的。为了提高算法的效率,本文提出一种分布计数的基数排序方法求解等价类 U/C,在此基础上设计求核算法,其时间复杂度为 $\max\{O(|C||U/C|^2),O(|C||U|)\}$,空间复杂度为 $O(|C||U/C|^2)$。

2 基本概念[9]

定义 1 决策表信息系统可以定义为:

$S=(U,A,V,f)$

其中,U 为论域,是对象的集合,$U=\{x_1,x_2,\cdots,x_n\}$;A 为属性集,

基金项目:安徽省自然科学基金(the Natural Science Foundation of Anhui Province of China under Grant No.050420204);安徽高校省级自然科学研究项目(the Natural Science Foundation of Education of Anhui Province of China under Grant,No.KJ2008B117)。

作者简介:葛浩(1976-),男,讲师,主要研究方向为数据挖掘、粗糙集;杨传健(1978-),女,讲师,主要研究方向为信息集成、数据挖掘;李龙澍(1956-),男,教授,博士生导师,主要研究方向为不精确信息处理和智能软件。

收稿日期:2008-01-21 **修回日期**:2008-04-14

$A=\{a_1,a_2,\cdots,a_m\}$,A 由两个部分组成 $A=C\cup D$ 且 $C\cap D=\Phi$,C 为条件属性集,D 为决策属性集,一般情况下 D 中只含有一个属性 $D=\{d\}$;V 为属性的值域,$V=\{V_{a_1},V_{a_2},\cdots,V_{a_m}\}$;$f$ 为信息函数 $f:U\times A\rightarrow V$,对于 $a\in A$,$x\in U$,有 $f(x,a)\in V_a$。

定义 2 决策表信息系统 $S=(U,A,V,f)$,令 $R\subseteq A$,且 $R\neq\Phi$,$ind(R)=\{(x_i,x_j)|f(x_i,b)=f(x_j,b),b\in R\}$称为 S 的不可区分关系。显然不可区分关系为一个等价类,含 x 的等价类记为 $[x]_{ind(R)}$ 或 $[x]_R$。

$U/ind(R)$ 表示等价关系 $ind(R)$ 的所有等价类。在不混淆的情况下用 U/R 来代替 $U/ind(R)$。

定义 3 对于决策表信息系统 $S=(U,A,V,f)$,令 $R\subseteq A$,$R\_X=\{x\in U|[x]_R\subseteq X\}$ 称为 X 的 R 下近似集;$\overline{R}X=\{x\in U|[x]_R\cap X\neq\Phi\}$称为 X 的 R 上近似集;$POS_R(X)=R\_X$ 称为 X 的 R 的正区域。

定义 4 对于决策表信息系统 $S=(U,C\cup D,V,f)$,D 的 C 正域记为 $POS_C(D)$,定义为:$POS_C(D)=\bigcup_{X\in U/D}C\_(X)$。

定义 5 决策表信息系统 $S=(U,C\cup D,V,f)$中,$a\in C$,如果 $POS_C(D)=POS_{C-\{a\}}(D)$,则称 a 为 C 中相对 D 不必要的;否则称 a 为 C 中相对 D 必要的。C 的所有必要属性的集合称为 C 相对 D 的核,记为 $Core(C)$。

3　求核的性质

定义 6[8] 在决策表信息系统 $S=(U,C\cup D,V,f)$中,定义新的决策表 $S'=(U,C\cup D',V',f')$,其满足:

$f'(x_i,C)=f(xi,C)$

$$f'(x_i,D)=\begin{cases}MAX+1 & f(x_i,C)=f(x_j,C)\wedge f(x_i,D)\neq f(x_j,D)\\ f(x_i,D) & \text{other}\end{cases}$$

其中,对 $\forall x_i\in U$,记 $MAX=\max\limits_{i=1\cdots|U|}\{f(x_i,D)\}$。

为了降低二进制可分辨矩阵占用的存储空间,需先对 S' 进行化简。

定义 7 决策表信息系统 $S'=(U,C\cup D,V',f')$中,其中 $A=C\cup D$,设 $U/A=\{[x'_1]_A,[x'_2]_A,\cdots,[x'_m]_A\}$,记 $U'=\{x'_1,x'_2,\cdots,x'_m\}$,则化简后的决策表信息系统记为:$S''=(U',C\cup D,V',f')$。

定义 8 在简化的决策表信息系统 $S''=(U',C\cup D,V',f')$中,创建二进制可分辨矩阵 $\boldsymbol{BM}=\{m'_{((i,j),r)}\}$,其中:

$$m'_{((i,j),r)}=\begin{cases}1 & f'(x_i,a_r)\neq f'(x_j,a_r)\wedge f'(x_i,D)\neq f'(x_j,D)\\ 0 & \text{otherwise}\end{cases}$$

定义 9 对于决策表信息系统 $S''=(U',C\cup D,V',f')$,其核定义为:

$$BMCore(C)=\{a_r|a_r\in C,m'_{((i,j),r)}=1\wedge\sum_{k=1}^{|C|}m_{((i,j),k)}=1\}$$

定理 1 对于决策表信息系统 $S=(U,C\cup D,V,f)$,有 $BMCore(C)=Core(C)$。

证明 设 $U/C=\{X_1,X_2,\cdots,X_n\}$,$U/D=\{Y_1,Y_2,\cdots,Y_m\}$。

首先,证明 $BMCore(C)\subseteq Core(C)$。

只要证明 $\forall a_r\in BMCore(C)$,有 $a_r\in Core(C)$。

(1)对于 $\forall a_r\in BMCore(C)$,根据定义 8 和定义 9,存在 $m'_{((i,j),r)}=1$,则有 $x_j\notin[x_i]_C$,但 $x_j\in[x_i]_{C-\{ar\}}$。又因为 $f(x_i,D)\neq f(x_j,D)$,则设 $x_i\in Yq$,$x_j\in Yp$,$x_j\notin Yq$。因 $x_j\in[x_i]_{C-\{a_r\}}$,$x_j\notin Yq$,根据下近似定义得 $x_i\notin(C-\{a_r\})\_Yq$,因此,$x_i\notin POS_{C-\{a_r\}}(D)$。同理得 $x_j\notin(C-\{a_r\})\_Yp$,则 $x_j\notin POS_{C-\{a_r\}}(D)$。

(2)根据定义 8 和定义 9 知,$a_r\in BMCore(C)$,有三种情况:①$f'(x_i,D)=MAX+1$ 且 $f'(x_j,D)\neq MAX+1$,存在 $x_s,x_l\in[x_i]_C$ 且 $f'(x_s,D)\neq f'(x_l,D)$,对 $x_t,x_k\in[x_i]_C$ 均满足 $f'(x_t,D)=f'(x_k,D)$,则 $[x_i]_C\in Yp$,有 $x_j\in C\_Yp$,因此 $x_j\in POS_C(D)$;②$f'(x_i,D)\neq MAX+1$ 且 $f'(x_j,D)=MAX+1$,对于 $x_t,x_k\in[x_i]_C$ 均满足 $f'(x_t,D)=f'(x_k,D)$,存在 $x_s,x_l\in[x_i]_C$ 且 $f'(x_s,D)=f'(x_l,D)$,则 $[x_i]_C\in Yq$,有 $x_i\in C\_Yq$,因此 $x_i\in POS_C(D)$;③$f'(x_j,D)\neq MAX+1$ 且 $f'(x_i,D)\neq MAX+1$,则有 $x_i\in POS_C(D)$,$x_j\in POS_C(D)$。

由(1)、(2)可得 $POS_C(D)\neq POS_{C-\{a_r\}}(D)$,推导出 $a_r\in Core(C)$。则 $BMCore(C)\subseteq Core(C)$。

然后,证明 $Core(C)\subseteq BMCore(C)$。

(1)对于 $ar\in Core(C)$,有 $POS_C(D)\neq POS_{C-\{a_r\}}(D)$,则存在 x_i,有 $[x_i]_C\subseteq Yq$,但 $[x_i]_{C-\{a_r\}}\not\subset Yq$,于是存在 $x_j\in[x_i]_{C-\{a_r\}}$,但 $x_j\notin[x_i]_C$。因此有 $f(x_i,C-\{a_r\})=f(x_j,C-\{a_r\})$且 $f(x_i,D)\neq f(x_j,D)$,即 $f(x_i,a_r)\neq f(x_j,a_r)$且 $f(x_i,D)\neq f(x_j,D)$,由定义 8、9 知 $a_r\in m'_{(i,j)}$,并且 $m'_{((i,j),r)}=1$。

(2)下面要证明 a_r 是 $m'_{(i,j)}$中唯一值为 1 的属性元素。即 $\sum\limits_{k=1}^{|C|}m_{((i,j),k)}=1$,且 $m'_{((i,j),r)}=1$。

采用反证法证明。假设 a_r 不是 $m'_{(i,j)}$中唯一的属性元素,设存在 $a_s\in C$,有 $m_{((i,j),s)}=1$。

因为,$a_s\in Core(C)$且 $a_s\in m'_{(i,j)}$,则有 $f(x_i,C-\{a_s\})=f(x_j,C-\{a_s\})$且 $f(x_i,D)\neq f(x_j,D)$,此与 $f(x_i,a_r)\neq f(x_j,a_r)$相矛盾。故 a_r 是 $m'_{(i,j)}$中唯一的元素。即 $\sum\limits_{k=1}^{|C|}m_{((i,j),k)}=1$,其中 $m'_{((i,j),r)}=1$。

则 $a_r\in BMCore(C)$。因此,$Core(C)\subseteq BMCore(C)$。综上所述,得证 $BMCore(C)=Core(C)$。

4　基于二进制可分辨矩阵的求核方法

4.1　高效的等价类划分算法

求等价类的一般方法是对样本集 U 中未分类的对象进行两两比较,比较它们对条件属性集 C 每个属性取值是否相同,如果相同,则属于同一个等价类。上述方法的时间复杂度为 $O(|C||U|^2)$。

性质 1 一个决策表信息系统 $S=(U,C\cup D,V,f)$,两个样本 $x_i,x_j\in U$ 相对于属性集 C 同属于一个等价类当且仅当 $\forall a\in A$,有 $f(x_i,a)=f(x_j,a)$。

由定义 2,可以得证。

根据性质 1,可先对决策系统 S 按属性集 C 排序,然后分析排序后的决策系统 S,划分等价类。刘少辉使用了快速排序[9],使等价类划分算法的时间复杂度降低为 $O(|C||U|log|U|)$。徐章艳利用链式基数排序算法[7-8],将时间复杂度降低到 $O(|C||U|)$,这是目前效率最好的算法。提出一种分布计数的基数排序方法,按属性集 C 对决策表 S 排序,该算法的时间复杂度也为 $O(|C||U|)$,空间复杂度为 $O(|U|)$,并且该算法相对于徐章艳的方法更易于理解和实现。

对决策表采用分布计数的基数排序思想:设 $S=(U,C\cup D)$,其中 $C=\{a_i|i=1\cdots m\}$,$D=\{d\}$,决策表一行为一个数据对象,则 S 是数据对象的集合:$S=\{S_i|i=1\cdots n\}$,其中 S_i 为一个 $m+2$ 的元组:$S_i=(x_i,a_1,a_2,\cdots,a_m,d)$,其中,$S_i.x_i$ 为对象的编号,$S_i.a_j$ 表示对象 i

的 a_j 属性值，$S_i.d$ 表示对象 i 的决策属性值。

按照属性集 C 对 S 排序，即依次以每个属性 a_i 对 S 排序。首先，把需要离散化的属性 a_i 离散化，将其分布在整型区间[1…e]（其中，$0<e\leqslant|U|$）；然后，构造一个计数表 *countPos*[0…e]，*countPos* 中元素个数为 $ind(a_i)$等价类的个数，每个元素用于存放 $ind(a_i)$中每个等价类当前最后一个元素在有序决策表中的位置，根据 *countPos* 表，可以直接将每个对象 S_i 放到有序决策表最终的位置。在这个过程中，需要使用两个辅助空间：一个是 *countPos*；一个是存放有序决策表的 *sortedS*。

算法 1 采用分布计数的基数排序方法按属性集 C 对 S 排序

输入：决策系统 $S=(U,C\cup D,V,f)$，$U=\{x_i|i=1\cdots n\}$，$C=\{a_i|i=1\cdots m\}$；

输出：按属性集 C 排序后的信息系统表 *sortedS*；

步骤 1 for（$r=1;r<=|C|;r++$）

{

步骤 1.1 *countPos*[0…e]=0；

步骤 1.2 for($j=1;j<=|U|;j++$) *countPos*[$x_j.a_r$]++；

步骤 1.3 for(j=1;$j<=e;j++$) *countPos*[j]+=*countPos*[$j-1$]；

步骤 1.4 for（$j=|U|;j>0;j--$）

{

步骤 1.4.1 *sortedS*[*countPos*[$x_j.a_r$]]=x_j；

步骤 1.4.2 *countPos*[$x_j.a_r$]--；

}

}

步骤 2 输出 *sortedS*；

步骤 3 结束

在算法 1 中，步骤 1.1 的时间复杂度为 O(|e|)，最坏情况下为 O(|U|)，步骤 1.2 的时间复杂度为 O(|U|)，步骤 1.3 的时间复杂度最坏情况下为 O(|U|)，步骤 1.4 的时间复杂度为 O(|U|)。因此，步骤 1 循环体的时间复杂度为 O(|U|)+O(|U|)+O(|U|)+O(|U|)=O(|U|)，而循环次数为 O(|C|)，因而步骤 1 总的时间复杂度为 O(|C||U|)。也就是说，算法 1 的时间复杂度为 O(|C||U|)。在步骤 1.1 中辅助空间 *countPos* 其容量最大为|U|+1，步骤 1.4 中 *sortedS* 占用的空间为|U|。因而，算法 1 的空间复杂度为 O(|U|)。

在算法 1 的基础上，下面给出一种快速等价类划分算法。

算法 2 等价类划分

输入：决策系统 $S=(U,C\cup D,V,f)$，$U=\{x_i|i=1\cdots n\}$，$C=\{a_i|i=1\cdots m\}$；

输出：U/C；

步骤 1 利用分布计数的基数排序法按属性集 C 对 S 进行排序；

步骤 2 *countClassPos*[0…e]=0； //标识等价类的位置

步骤 3 $s=1$，$E_1=\{x_1\}$；*countClassPos*[s]=1； //s 计数等价类的数目

步骤 4 for($i=2;i>=|U|;i++$)

步骤 4.1 if(x_i 和 x_i-1 对于 C 的每个属性值均相等）

then {$Es=Es\cup\{x_i\}$，*countClassPos*[s]++；}

else {s++，$Es=\{x_i\}$，*countClassPos*[s]=*countClassPos*[$s-1$]+1；}

步骤 5 输出 U/C(即，集合 E)和 countClassPos；

步骤 6 结束。

在算法 2 中，步骤 1 调用了算法 1，则步骤 1 的时间复杂度为 O(|C||U|)；步骤 2 中 e 最坏情况下取值为|U|，则步骤 2 时间复杂度最坏情况下为 O(|U|)，空间复杂度为 O(|U|)；步骤 4 的时间复杂度为 O(|C||U|)。可得出，算法 2 总的时间复杂度为 O(|C||U|)+O(|C||U|)=O(|C||U|)，空间复杂度为 O(|U|)。

4.2 求核算法

算法 3 快速求核算法。

输入：决策信息系统 $S=(U,C\cup D,V,f)$，$U=\{x_i|i=1\cdots n\}$，$C=\{a_i|i=1\cdots m\}$，$D=\{d\}$；

输出：决策信息系统的核 *Core*(C)；

步骤 1 *Core*(C)=Φ；

步骤 2 采用算法 1 的分布计数的基数排序法，按属性集 C 对 S 进行排序；

步骤 3 分析排序后的有序决策表，按照定义 8 创建新的决策表 S'；

步骤 4 根据定义 9 化简决策表为 S''；

步骤 5 根据定义 10 创建可二进制可分辨矩阵 BM；

步骤 6 for（$s=1;s<=(|U/A|^2)/2;s++$）

步骤 6.1 { *flag*=1，*num*=0；

步骤 6.2 for($r=1;r<|C|;r++$)

{ if($m_{((i,j),a_r}$)) num++；

if(*num*>1){*flag*=0，break；}

if(*flag*) *Core*(C)=*Core*(C)∪{a_r}；

} //end_for_r

} //end_for_s

步骤 7 输出 Core(C)；

步骤 8 结束。

算法 3 中，步骤 2 的时间复杂度为 O(|C||U|)，步骤 3 和步骤 4 的时间复杂度为 O(|U|)，步骤 5 的时间复杂度为 O(|C||U/C|^2)，步骤 6 的时间复杂度为 O(|C||U/C|^2)。因此，算法 3 的的时间复杂度为 O(|C||U|)+2O(|U|)+2O(|C||U/C|^2)=max{O(|C||U|)，O(|C||U/C|^2)}，空间复杂度为 O(|C||U/C|^2)。其时间复杂度和空间复杂均低于 HU 方法，并且时间复杂度和空间复杂度与徐章艳的方法[7]是相当的。

4.3 实例分析

对文献[6]中的决策表 1，首先根据定义 8，建立新的决策表信息系统 S'（如表 1），然后计算 $U/A=\{\{x_1,x_2\},\{x_3,x_4\},\{x_5\}\}=\{[x_1]_A,[x_3]_A,[x_5]_A,\}$，可得 $U'=\{x_1,x_3,x_5\}$，则化简后的决策表 S''，如表 2。

表 1 化简后的决策表 S'

U	c_1	c_2	c_3	D
x_1	1	0	1	2
x_2	1	0	1	2
x_3	0	0	1	2
x_4	0	0	1	2
x_5	1	1	1	1

表 2 化简后的决策表 S''

U'	c_1	c_2	c_3	D
x_1	1	0	1	2
x_3	0	0	1	2
x_5	1	1	1	1

对 S''创建二进制可分辨矩阵 ***BM***，如表 3。该表占用 2 个空间，而采用文献[5]的方法需占用 6 个空间，可见本文算法对决策表化简后，占用的空间表明显减少。因而，对于大数据集可以更显著地降低样本对象的数目，减少空间的占用，提高求解效率。

表 3 二进制可分辨矩阵 *BM*

(x_i,x_j)	c_1	c_2	c_3
(1,5)	0	1	0
(3,5)	1	1	0

分析二进制可分辨矩阵 ***BM***，可得核 *Core*={c_2}，与正区域

（下转 177 页）