基于有序差别集和属性重要性的属性约简

张迎春[1] 王宇新[2] 郭 禾[1]
(大连理工大学软件学院 大连 116024)[1] (大连理工大学计算机科学与技术学院 大连 116024)[2]

摘 要 针对粗糙集理论的属性约简问题,提出新的差别矩阵简化算法,该算法在无需排序和较少遍历次数的情况下简化了差别矩阵,明显提高了简化速度并最终得到简化的有序差别集。实验验证了该算法的高效性;给出度量属性重要性的新标准,即根据属性所在差别矩阵元素的权重、在差别集中出现的频数和吸收能力3方面来度量其重要性;在上述两者基础上,提出一种基于有序差别集和属性重要性的属性约简新方法,理论分析证明新方法的最坏时间复杂度低于其它基于差别矩阵的属性约简算法。大量实验结果也表明,新方法的有效性甚至可以在很大程度上得到最小属性约简。

关键词 粗糙集,属性约简,简化差别矩阵,差别集,属性重要性
中图法分类号 TP311 **文献标识码** A

Attribute Reduction Based on Ordered Discernibility Set and Significance of Attribute

ZHANG Ying-chun[1] WANG Yu-xin[2] GUO He[1]
(School of Software Technology, Dalian University of Technology, Dalian 116024, China)[1]
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)[2]

Abstract A new improved algorithm for the simplified discernibility-matrix was proposed on the subject of attribute reduction in rough set theory. Discernibility-matrix is being simplified without being sorted and at fewer cost of traversing. This can notably raise the speed of being simplified discernibility-matrix and ultimately obtain the ordered and simplified discernibility set. The comparative experiments on computational efficiency show that this new algorithm is more efficient than the homogeneous ones. A new criterion of significance of attribute was put forward based on the three aspects which are the weight of element containing the attribute, the frequency and the absorptive ability of the attribute in the discernibility set. Therefore a new method for attribute reduction was introduced on the basis of the above two points, the theoretical analysis proves that the worst time complexity of the new method is less than the other ones based on discernibility-matrix. In addition, lots of comparative experiments in attribute reduction display that this new method is effective and can largely find out a minimal attribute reduction.

Keywords Rough set, Attribute reduction, Simplified discernibility-atrix, Discernibility set, Significance of attribute

粗糙集(Rough sets)理论[1]是一种处理模糊和不确定性问题的数学工具,已在知识发现、人工智能、模式识别等方面得到广泛应用[2-4]。属性约简是研究粗糙集理论的重要内容,是在保持信息系统分类能力不变的条件下删除其中冗余属性。

目前,已经有许多关于粗糙集理论的属性约简方法[5-12],一般可分为3种:第一种是基于正域的属性约简方法;第二种是基于差别矩阵的属性约简方法;第三种是基于信息熵的属性约简方法。其中,Skowron提出经典的差别矩阵方法[4]为属性约简提供了很好的思路,该方法直观明了、约简计算比较简单。很多属性约简算法都是基于差别矩阵的或在此基础上进行改进的算法[6,9-12]。基于差别矩阵求属性约简的改进算法大体上可分为两种:一种是首先构造出最简化的差别矩阵[6,9],若有核属性就将核属性作为初始约简集,然后计算除去核属性集的剩余条件属性在差别矩阵中出现的频数,依次选择频数最大的属性加入约简集,直到产生一个属性约简为止,最后通过属性频数的大小来确定属性的重要性;另一种是首先构造最简化的差别矩阵,而后依次选择必要属性加入约简集,实际上是在每次迭代过程中都将必要属性当成是重要性最大的属性加入属性约简集,直到产生一个属性约简为止[11,12]。文献[6]和文献[9]的区别在于构造差别矩阵的方法不同。文献[11]和文献[12]的区别在于文献[12]在每次迭代中找不到必要属性时将去除在差别矩阵中出现频率小的属性,而文献[11]在找不到必要属性时则随机去除一个属性,为下一次迭代找到必要条件属性做准备。基于差别矩阵的属性约简算法的最坏时间复杂度是$O(|C|^2|A|)$,其中$|C|$和$|A|$

到稿日期:2010-11-08 返修日期:2011-03-21

张迎春(1980—),女,博士生,主要研究方向为粗糙集理论、图像处理、计算机视觉等,E-mail:zhangyingchun1871@163.com;王宇新(1973—),男,讲师,主要研究方向为计算机系统结构、图像识别与理解、粗糙集理论等;郭 禾(1955—),男,教授,博士生导师,主要研究方向为分布式计算、图像处理、计算机视觉、粗糙集理论等。

分别为条件属性个数和差别矩阵元素个数。为了降低基于差别矩阵求属性约简的时间复杂度和尽可能得到最小属性约简，本文提出新的属性约简方法。

大部分学者在利用差别矩阵求属性约简时，为了求出最简化的差别矩阵，通常先求出决策表的属性核，以核属性为启发信息约去包含核属性的差别矩阵元素来简化差别矩阵，具有较好的时间性能，如文献[9]。但并不是所有决策表都含有核属性。当决策表不含核属性时，其时间性能跟经典算法的时间性能是一样的。文献[10]针对这种情况提出了边生成差别矩阵元素、边简化差别矩阵的简化差别矩阵方法，但是这种方法有需改进之处，当生成差别矩阵元素时，每次都需要从差别矩阵的首元素检索是否有元素能包含于新生成的差别矩阵元素。为此本文提出了快速简化差别矩阵的改进算法。

1 简化差别矩阵方法

1.1 基本概念

定义 1[6] 决策表 $S=(U,C\cup D,V,F)$。其中 C 是条件属性集合，$U=\{x_1,x_2,\cdots,x_n\}$；D 是决策属性集合，一般只含有一个属性 $D\neq\phi$；$V=\bigcup_{a\in(C\cup D)}V_a$，$V_a$ 为属性 a 的值域集合；F：$U\times(C\cup D)\rightarrow V$ 是一个信息函数，指定 U 中每一个对象 x 的属性值。

定义 2[6] 决策表 $S=(U,C\cup D,V,F)$ 中对于 $IND(B)=\{(x,y)\in U\times U\mid \forall b\in B$ 有 $F(x,b)=F(y,b)\}$，$B\subseteq C\cup D$，称 $IND(B)$ 为一个等价关系，$U/IND(C)=\{U_1,U_2,\cdots,U_l\}$ 表示条件属性集 C 对论域 U 的划分，记为 U/C，其中称 $U_i(i=1,2,\cdots,l)$ 为一个等价类，$U/IND(D)=\{D_1,D_2,\cdots,D_p\}$ 表示有决策属性集 D 对论域 U 的划分，记为 U/D。

定义 3[6] 给定相容决策表 S，差别矩阵 $M=(m_{ij})_{n\times n}$ 的元素定义为 $m_{ij}=\{a\in C\mid F(x_i,a)\neq F(x_j,a)\wedge F(x_i,D)\neq F(x_j,D)\}$，在其它情况下，$m_{ij}$ 为空集。

定义 4[11] 给定决策表 S，差别矩阵为 M，M 中的所有非空元素组成的集合 δ_M 成为 S 的差别集。

定义 5[11] δ_M 为决策表 S 的差别集，$a\in C$，若 $\{a\}\in\delta_M$，则称 a 为 δ_M 必要属性，即核属性。

定义 6 给定差别矩阵 M，在不考虑差别矩阵元素为空集的情况下，简化的差别矩阵定义为 $M'=\{m\mid m\in M$，且不存在 $m'\in M$ 使得 $m'\subset m\}$。

定义 7 给定决策表 S，简化的差别矩阵为 M'，M' 中的所有非空差别矩阵元素按照所含属性的个数排序组成的集合 $\delta_{M'}$ 称为 S 的有序的简化的差别集。

定义 8(逻辑属性桶) 设 $T[|C|-1]$ 为数组链表，用 LB_k 表示逻辑属性桶：

$$LB_k=\begin{cases}m=\{x\mid x\in M,|x|=k\}\\ First(LB_k)=T[k-1]\\ Last(LB_k)=T[k]\end{cases},0<k<|C|$$

1.2 简化差别矩阵的算法

算法 1 基于属性桶的简化差别矩阵的算法

输入：决策表 $S=(U,C\cup D,V,F)$；

Step 1 按决策属性取值对 S 进行等价类排序，$U/D=\{X_1,X_2,\cdots,X_{dS}\}$，$1<dS<|U|$；

Step 2 依次对从 X_1 到 X_{dS} 的各等价类进行删除重复对象处理，获得新决策表 $S'=\langle U',C\cup D,V',F'\rangle$，$U'/D=\{Y_1,Y_2,\cdots,Y_{ds}\}$，$1<ds<|U'|$；

Step 3 建立数组链表 $T[|C|-1]$，相当申请 $|C|-1$ 个逻辑空桶，再申请一个临时空结点 Lp 用于存储某个差别矩阵元素；

Step 4 For $k=0$ to $ds-1$ do

For $i=D[k]$ to $D[k+1]-1$ do

//$D[k]$存储第 $k+1$ 个等价类的首元素在 S' 中的序列号

For $j=D[k+1]$ to $|U'|$ do

Step 4.1 让 S' 中第 i 和第 j 个元素进行比较，获得差别矩阵元素 m_{ij}，并存入 Lp 中，记 es 为 m_{ij} 中所含属性的个数；

Step 4.2 比较逻辑属性桶 es 内是否存在着跟 m_{ij} 相同的差别矩阵元素；如果存在就执行 Step4.3，否则执行步骤 Step4.4；

Step 4.3 将 Lp 置空，跳出本次循环；

Step 4.4 分别与逻辑桶 1 到逻辑桶 $es-1$ 内差别矩阵元素作比较，如果存在着差别矩阵元素被 m_{ij} 包含，就执行 Step4.3，否则执行 Step4.5；

Step 4.5 将 Lp 加入逻辑桶 es 中，再将逻辑桶 $es+1$ 到逻辑桶 $|C|-1$ 之间包含 m_{ij} 的差别矩阵元素删除，之后重新建立一个临时空结点 Lp；

Step 5 删除 $|C|-1$ 个逻辑桶。

输出：有序的简化的差别集。

在每一个差别矩阵元素加入差别矩阵之前，Step4.2 避免了每次判断含有差别矩阵元素是否被新生成的差别矩阵元素所包含都需要从头开始搜索。Step4.5 使得只将符合条件生成的新元素加入了差别矩阵再申请临时空间结点，即在每次生成差别矩阵元素的时候不是必须申请临时空间结点。这就大大减少了搜索时间、判断时间以及存储空间。

循环完毕后删掉所有桶标记，就得到了一个最简化差别矩阵，同时得到一个按属性个数升序表示的简化差别集。

1.3 简化差别矩阵的算法时间复杂度

为了计算本文的简化差别矩阵的时间复杂度，首先给出如下定理的证明。

定理 1 若 $|U'|=\sum_{i=1}^{dS}|Y_i|$，那么 $\sum_{1\leqslant i<j\leqslant dS}|Y_i||Y_j|=\frac{1}{2}(|U'|^2-\sum_{i=1}^{dS}|Y_i|^2)$

$$\begin{aligned}\because |U'|^2&=(\sum_{1\leqslant i\leqslant dS}|Y_i|)^2=(|Y_1|+|Y_2|+\cdots+|Y_{dS}|)^2\\&=(|Y_1|+|Y_2|+\cdots+|Y_{dS}|)\times(|Y_1|+|Y_2|+\cdots+|Y_{dS}|)\\&=|Y_1|^2+|Y_1||Y_2|+\cdots+|Y_1||Y_{dS}|+|Y_2||Y_1|+|Y_2|^2+\cdots+|Y_2||Y_{dS}|+\cdots+|Y_{dS}||Y_1|+|Y_{dS}||Y_2|+\cdots+|Y_{dS}|^2\\&=2\sum_{1\leqslant i<j\leqslant ds}|Y_i||Y_j|+\sum_{i=1}^{dS}|Y_i|^2\\\therefore \sum_{1\leqslant i<j\leqslant dS}|Y_i||Y_j|&=\frac{1}{2}(|U'|^2-\sum_{i=1}^{dS}|Y_i|^2)\end{aligned}$$

定理 1 得证。

Step1 算法复杂度为 $O(|U|)$，Step2 算法复杂度为 $O(|C||U|)$，因为采用基数排序法；Step3、Step4.1 及 Step4.3 的算法复杂度为 $O(|C|)$；因为差别矩阵元素之间的比较采用基本函数 strcmp 来实现，所以算法复杂度为 O(1)。而 Step4.2、Step4.4 及 Step4.5 每执行一次的算法复杂度取决于当时差别矩阵中含有的差别矩阵元素个数和差别矩阵元素的分布情况，但是在 Step4 过程中的所有循环执行完毕后，差别矩阵中元素的插入和删除操作执行的总次数是有一定范围限定的，若设插入到差别矩阵中的元素个数是 $|A^1|$，简化后差别

矩阵中元素个数是$|A|$，插入和删除操作的执行总次数就是$2|A^1|-|A|$。Step4 的总循环次数是$\sum_{1\leqslant i<j\leqslant d'}|Y_i||Y_j|$，因为生成的差别矩阵元素有一部分根本没有插入到差别矩阵中，故$|A|<|A^1|<\sum_{1\leqslant i<j\leqslant d'}|Y_i||Y_j|$，而$\sum_{1\leqslant i<j\leqslant d'}|Y_i||Y_j|=\frac{1}{2}(|U'|^2-\sum_{i=1}^{d'}|Y_i|^2)$，所以 Step4 的算法复杂度是$O((|U'|^2-\sum_{i=1}^{d'}|Y_i|^2)|C|)$；Step5的算法复杂度是$O(|A|)$；因此本文的简化差别矩阵的算法时间复杂度是$O((|U'|^2-\sum_{i=1}^{d'}|Y_i|^2)|C|)$。这要比经典的求差别矩阵算法时间复杂度$O(|U|^2|C|)$低很多。尤其是当决策表的等价类的个数很少且每个等价类所含的元素个数很多时，本文的简化差别矩阵的算法时间复杂度会很小。

2 相对属性约简方法

2.1 属性约简方法的思想

首先给出差别矩阵元素权重和属性的频数及属性吸收能力的定义。

定义 9 属性C_i在差别矩阵M中出现的频数$FRE(C_i,M)=|\{m|C_i\in m\wedge m\in M\}|$。

定义 10 差别矩阵元素的权重$WGT(m_i)=|C|-|m_i|$。

性质 1 如果差别矩阵元素m_i所含的属性个数$|m_i|$小于差别矩阵元素m_j所含的属性个数$|m_j|$，那么差别矩阵元素m_i要比差别矩阵元素m_j重要。

定义 11 属性的相对吸收能力$SRB(C_i,M)=\frac{\sum|m_k|}{|C||M|}$，其中$m_k\in\{m|C_i\in m\wedge m\in M\}$。

性质 2 如果某条件属性C_i和C_j在差别集中出现的频数$FRE(C_i,M)=FRE(C_j,M)$，且$SRB(C_i,M)>SRB(C_j,M)$，那么属性C_i比C_j可以吸收更多的相对不重要的差别矩阵元素。

在基于差别矩阵的属性约简算法中，通常采用属性在差别矩阵中出现的频数作为属性重要性的度量标准，或者采用必要属性作为最重要属性。而本文的属性重要性的度量标准分 3 个步骤：首先找出含属性个数最少的差别矩阵元素，即优先选择权重最大的差别矩阵元素；再从这个差别矩阵元素中选取出现频率最高的属性加入属性约简集合中，如果出现两个以上属性的频数相同，那么将选择吸收能力大的属性；同时将含有该属性的差别矩阵元素删除，直到所有的差别矩阵元素处理完毕。

新属性重要性标准产生的原因分析：决策表的属性约简的结果可能不止一个，核属性存在于每个属性约简的结果中[10]，所以核属性是最重要的属性。若差别矩阵元素只含有一个属性，那么这个属性是核属性[11]，只含有一个属性的元素就可认为是最重要的差别矩阵元素。同理可认为元素所含属性越少，这个差别矩阵元素越重要。而频数大的属性能删除更多的差别矩阵元素[6,9]，但有可能把一些更重要的差别矩阵元素删掉。为了避免这种情况，将选择重要的差别矩阵元素中频数最大属性作为最重要的属性。在属性约简的过程中，可能出现属性所在差别元素的权重、频数都相同的情况。如果出现这种情况，就将吸收能力强的属性视为重要属性。因为吸收能力强的属性可以删掉更多的相对不重要的差别矩阵元素，所以本文从属性所在差别矩阵元素的权重和属性所在差别矩阵中出现的频数及属性的吸收能力这 3 个方面依次度量属性的重要性。

2.2 属性约简的算法设计

算法 2 基于有序的差别集和新属性重要性度量标准的属性约简

输入：有序简化的差别集

输出：一个属性约简集

Step1 初始化指针 Zhead 指向当前有序简化差别集的首元素，属性约简集合 R 为空；

Step2 搜索 Zhead 所指差别元素中属性重要性最大的属性

Step2.1 令某条件属性的标号 $Cr=0$，频数 $Cf=0$，吸收能力 $Cs=0$；

Step2.2 For $i=0$ to $|C|-1$ do

Step2.2.1 if(Zhead 所指的元素第 i 个位的值==1) then 计算 $FRE(C_i,M')$，$SRB(C_i,M')$；

Step2.2.2 if($FRE(C_i,M')>Cf$) OR ($FRE(C_i,M')==Cf$ AND $SRB(C_i,M')>Cs$) then 执行 Step2.2.4；else {if($FRE(C_i,M')==Cf$ AND$SRB(C_i,M')==Cs$) then 执行 Step2.2.3；if($FRE(C_i,M')<Cf$) then 跳出本次循环；}

Step2.2.3 if ($FRE(C_i,M)>FRE(C_{Cr},M)$ OR ($FRE(C_i,M)==FRE(C_{Cr},M)$ AND $SRB(C_i,M)>SRB(C_{Cr},M)$)) then 执行 Step2.2.4；else 跳出本次循环；

Step2.2.4 令 $Cr=i$；$Cf=FRE(C_i,M')$；$Cs=SRB(C_i,M')$；

Step3 将 Cr 对应的属性加入属性约简集 R；

Step4 删除所有含有 Cr 对应的属性的差别矩阵元素，删除过程结束时 Zhead 指向当前差别矩阵首元素；

Step5 指针 Zhead 不指向空，就执行 Step2，否则就停止。

其中 Zhead 始终指向一个不断缩小的差别集的第一个元素，因为输入的简化差别集是有序的，根据性质 1 可知，Zhead 指向的差别矩阵元素相对当前差别矩阵总是权重最大的，这样就省去了计算和寻找权重最大的差别矩阵元素的时间。

2.3 相对属性约简算法复杂度分析

$|R|$表示属性约简集合含有的属性个数，若$|R|=|C|$，即每个条件属性都是核属性，则算法的时间复杂度是$O(|C|)$；若$1\leqslant|R|<|C|$算，法最坏时间复杂度是$O(|R||C-R+1||A|)$，因为约简集有$|R|$个属性就意味着算法的最外层 Step2 循环语句最多执行$|R|$次。一般情况下，对步进 for 循环语句只需要考虑循环体中语句的执行次数。而循环体的执行次数由条件语句来控制，也即由差别矩阵元素所含属性个数来确定。假设算法中产生第一个约简属性的差别矩阵元素所含属性个数是$|A_1|$，那么循环体的执行次数是$|A_1|$，从而依次类推到第$|R|$个约简属性对应的差别矩阵元素所含属性个数为$|A_R|$，循环体的执行次数是$|A_R|$。由于本文的简化差别矩阵方法最终得到的差别集是按属性个数升序排列，因此$|A_1|\leqslant|A_2|\leqslant|A_R|$。而当求取第 R 个约简属性时 for 语句的循环体执行次数是$|A_R|$，而$|A_R|\leqslant|C-R+1|$，所以外层 Step2 语句执行一次 for 语句的循环体最多执行次数是$|C-R+1|$。for 语句的循环体每执行一次的最坏时间复杂度是$O(|A|)$，因此本文的最小属性约简算法的最坏时间复杂度是$O(R||C-R+1||A|)$，这比文献[6,9,11,12]基于属性在差别矩阵中的频数和必要性求最小属性约简的最坏时间复杂度$O(|C|^2|A|)$要低。

3 实验

实验条件是 CPU 为 1.5GHz、内存为 256MB 的笔记本电脑。实验用的数据集全部选用 UCI 机器学习数据库[13]中的数据集。因为在求出简化差别矩阵后，本文的属性约简算法的最坏时间复杂度能很清楚地与其他算法作比较，所以实验的目的是着重考查本文的简化差别矩阵算法的运算效率和本文的相对属性约简算法的有效性。

3.1 简化差别矩阵的效率对比实验

文献[10]对不相容决策表数据的预处理方式以及求完简化差别矩阵后的属性约简算法与本文不一致，为了保证实验的公平性，实验中用到的数据集都是相容决策表，并且在求完简化差别矩阵后统一采用本文的属性约简算法。实验如表 1 所列。

表 1 简化差别矩阵运算效率对比

数据集	对象数	条件属性数	算法 a			算法 1			简化后差别矩阵元素个数
			运行时间(s)	最大内存使用(MB)	CPU 占用率	运行时间(s)	最大内存使用(MB)	CPU 占用率	
promoter	106	57	7.062	43.420	0.97	0.578	1.132	0.11	2762
vehicle	846	18	1.360	1.032	0.35	0.328	0.876	0.17	431
waveform	5000	41	25.719	4.456	0.99	9.187	4.412	0.99	908
DNA1	100	180	5.750	3.384	0.27	0.797	1.240	0.12	3041
DNA2	300	180	425.969	26.000	0.99	80.016	7.340	0.99	26531
DNA3	400	180	1383.093	46.220	0.99	300.984	12.060	0.99	46847

表 1 中算法 a 是文献[10]的边生成差别矩阵元素、边简化差别矩阵的求最简化差别矩阵的方法，算法 1 是本文的借助属性桶构造简化差别矩阵的算法。图 1 和图 2 是表 1 另一种表示形式。为了验证本文算法的高效性，特意在 UCI 数据库中挑选了条件属性个数和数据对象个数比较多的 6 个数据集。实验表明，算法 1 的最大内存使用都不高于算法 a，尤其是简化差别矩阵的过程中当差别矩阵元素个数很大时，相对于算法 a 本文的算法 1 所用内存就非常小。数据 promoter 就是一个典型说明。并且在属性个数很大的情况下随着数据集中对象个数的增加，算法 a 运行时间和内存的使用急剧增大。而算法 1 的运行时间和内存使用的增加相对平缓，通过表 1、图 1 和图 2 很明显得出本文算法 1 的运行效率高于算法 a 的结论。

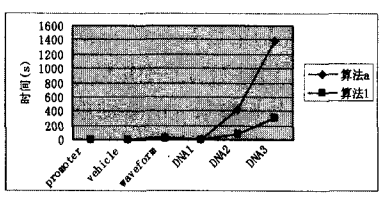

图 1 两种算法时间运行时间对比

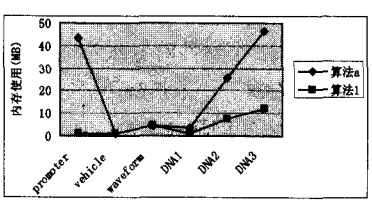

图 2 两种算法最大内存使用对比

3.2 相对属性约简结果对比实验

为了验证求属性约简结果的有效性，分别与文献[6,9,11,12]作对比。实验结果表明，本文的属性约简方法是有效的，可以最大程度地获得最小属性约简。结果如表 2 所列。

在表 2 中，约简结果的数值表示决策表中条件属性的序号，例如数值 2 表示条件属性集中第 3 个属性。约简结果中的数值个数代表约简集中属性的个数。本文中所有属性约简结果都可以在国外粗糙集软件 ROSETTA[14]中得到验证。表 2 中全部数据集由算法 2 求得的属性约简集所含属性个数都小于等于文献[6,9,11,12]。尤其是 letter 数据集与其他文献相比，本文算法求出的属性约简集所含的属性个数最小，这就进一步说明了本文算法 2 能够最大程度求出最小属性约简。

表 2 约简结果对比

数据集	文献[6,9]	文献[11]	文献[12]	本文算法 2
Letter	1,2,3,5,6,7,8,9,10,11,14,15	3,5,6,7,8,9,10,11,12,13,14,15	1,3,4,5,6,7,8,9,10,13,14,15	3,4,5,6,7,8,9,11,12,13,14
Zoo	2,3,5,7,12	3,5,9,11,12,13	3,5,8,11,12	3,5,8,11,12
Vehicle	0,11,15	13,14,15,16,17	3,11,12	11,12,15
Promoter	2,14,16,38	52,53,54,55,56	5,15,16,38	14,38,47,50
Waveform	0,2,10	37,38,39	10,14,16	10,14,16
Mushroom	2,4,10,21	4,19,20,21	4,10,19,20	3,7,10,19
Auto-mpg	0,2,4	4,5	0,2,4	4,5
Heart-statlog	0,3,7	4,9,10,11	0,4,7	0,4,7
Car	0,3,4,8	3,4,5,8	3,4,5,8	3,4,5,8

结束语 本文提出的简化差别矩阵新算法的最坏时间复杂度在理论上低于文献[2]、不高于文献[9,10]。实验结果也验证了此算法是高效的，算法在数据集不存在核属性的情况下可以高效率地简化差别矩阵。最小属性约简问题本身是一个 NP 完全问题，因此属性约简新方法并不能保证一定能求出最小属性约简。但是实验说明本方法能够最大程度求出最小属性约简，并且约简算法的时间复杂度低于同类基于差别矩阵求属性约简算法，如文献[6,9,11,12]。属性的重要性新标准的提出也会为进一步进行粗糙集理论研究和实际应用提供有意义的参考。

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. Rough sets and Boolean reasoning [M]. Pedrycz Wed. Granular Computing: An Emerging Paradigm. New York: Physica-Verlag, 2001: 95-124

[3] Wang Guo-Yin. Rough Set Theory and Knowledge Acquisition [M]. Xi'an: Xi'an Jiaotong University Press, 2001(in Chinese)

[4] Wang Guo-Yin, Yao Yin-Yu, YU Hong. A Survey on Rough Set Theory and Applications [J]. Chinese Journal of Computer, 2009, 32(7): 1229-1246

[5] Skowron A, Rauszer C. The discernibility matrics and functions

in information system[M]//Slowinski R, ed. Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Therory. Dordrecht: Kluwer Academic Publishers, 1992: 331-362

[6] Hu Xiao-hua, Cercone N. Learning in relational databases: A rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338

[7] 刘少辉,盛秋戬,史忠植. 一种新的快速计算正域的方法[J]. 计算机研究与发展,2003,40(5):637-642

[8] 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

[9] 史君华,胡学钢. 一种基于粗集的决策表属性约简新算法[J]. 计算机科学,2006,33(11):83-85

[10] 田卫东,周创德,胡学钢,等. 基于简化分辨矩阵的粗糙集属性约简算法[J]. 计算机科学,2008,35(3):209-212

[11] 王兵,陈善本. 一种基于差别矩阵的属性约简完备算法[J]. 上海交通大学学报,2004,38(1):43-46

[12] 蒋瑜,王燮,叶振. 基于差别矩阵的 Rough 集属性约简算法[J]. 系统仿真学报,2008,20(14):3717-3725

[13] http://www.ics.uci.edu/~mlearn/databases/

[14] http://www.cs.uregina.ca/~roughset

(上接第 222 页)

决策类 D_2 的可变粒度粗糙集的下近似。

3.2 变粒度粗糙集的度量

从粗糙集近似的角度来看,知识的粗糙性(不确定性)是由于边界域的存在而造成的,边界域越大,则粗糙程度就越大。下面从粗糙精度、属性依赖度和概念的可精确近似程度等方面来度量变粒度粗糙集。

定义 6 设 $IS=\langle U,AT,V,f\rangle$,令 $A_1,A_2,\cdots,A_m$ 是 AT 的 m 个属性子集,$A=\{A_1,A_2,\cdots,A_m\}$,$0<\beta\leqslant 1$,$\forall X\subseteq U(X\neq\varnothing)$关于 A 的乐观多粒度粗糙集、悲观多粒度粗糙集和可变粒度粗糙集的粗糙度因子分别定义为 α_O、α_P 和 α_β。

$$\alpha_O\left(\sum_{i=1}^m A_i,X\right)=\frac{\left|\underline{\sum_{i=1}^m A_i}^O(X)\right|}{\left|\overline{\sum_{i=1}^m A_i}^O(X)\right|}$$

$$\alpha_P\left(\sum_{i=1}^m A_i,X\right)=\frac{\left|\underline{\sum_{i=1}^m A_i}^P(X)\right|}{\left|\overline{\sum_{i=1}^m A_i}^P(X)\right|}$$

$$\alpha_\beta\left(\sum_{i=1}^m A_i,X\right)=\frac{\left|\underline{\sum_{i=1}^m A_i}^\beta(X)\right|}{\left|\overline{\sum_{i=1}^m A_i}^\beta(X)\right|}$$

定义 7 设 $DIS=\langle U,C\cup\{d\},V,f\rangle$,令 $A,A_2,\cdots,A_m$ 是 C 的 m 个属性子集,$A=\{A_1,A_2,\cdots,A_m\}$,$0<\beta\leqslant 1$,$D=\{D_1,D_2,\cdots,D_i\}$是由决策属性集$\{d\}$在论域 U 上导出的划分,则决策类 D 对属性子集 A 的粗糙分类能力,也即对 A 的属性依赖程度记为 γ,且乐观多粒度、悲观多粒度和可变粒度粗糙集的依赖度因子分别记为 γ_O、γ_P 和 γ_β。

$$\gamma_O\left(\sum_{i=1}^m A_i,D\right)=\frac{\sum\{|\underline{\sum_{i=1}^m A_i}^O(D_i)|:D_i\in D\}}{|U|}$$

$$\gamma_P\left(\sum_{i=1}^m A_i,D\right)=\frac{\sum\{|\underline{\sum_{i=1}^m A_i}^P(D_i)|:D_i\in D)\}}{|U|}$$

$$\gamma_\beta\left(\sum_{i=1}^m A_i,D\right)=\frac{\sum\{|\underline{\sum_{i=1}^m A_i}^\beta(D_i)|:D_i\in D\}}{|U|}$$

定义 8 设 $IS=\langle U,AT,V,f\rangle$,令 $A_1,A_2,\cdots,A_m$ 是 AT 的 m 个属性子集,$A=\{A_1,A_2,\cdots,A_m\}$,$0<\beta\leqslant 1$,$\forall X\subseteq U(X\neq\varnothing)$在属性子集 A 下可以被精确近似的程度定义为 π,且乐观多粒度、悲观多粒度和可变粒度粗糙集的精确近似程度因子分别记为 π_O、π_P 和 π_β。

$$\pi_O\left(\sum_{i=1}^m A_i,X\right)=\frac{\left|\underline{\sum_{i=1}^m A_i}^O(X)\right|}{|X|}$$

$$\pi_P\left(\sum_{i=1}^m A_i,X\right)=\frac{\left|\underline{\sum_{i=1}^m A_i}^P(X)\right|}{|X|}$$

$$\pi_\beta\left(\sum_{i=1}^m A_i,X\right)=\frac{\left|\underline{\sum_{i=1}^m A_i}^\beta(X)\right|}{|X|}$$

性质 1 设 $IS=\langle U,AT,V,f\rangle$,令 $A_1,A_2,\cdots,A_m$ 是 AT 的 m 个属性子集,$A=\{A_1,A_2,\cdots,A_m\}$,$0<\beta\leqslant 1$,$\forall X\subseteq U(X\neq\varnothing)$关于属性子集 A 的变粒度粗糙精度因子 α_β、可精确近似程度 π_β 与多粒度糙集对应的度量因子有如下的关系成立。

$$\alpha_P\left(\sum_{i=1}^m A_i,X\right)\leqslant\alpha_\beta\left(\sum_{i=1}^m A_i,X\right)\leqslant\alpha_o\left(\sum_{i=1}^m A_i,X\right)$$

$$\pi_P\left(\sum_{i=1}^m A_i,X\right)\leqslant\pi_\beta\left(\sum_{i=1}^m A_i,X\right)\leqslant\pi_o\left(\sum_{i=1}^m A_i,X\right)$$

性质 2 设 $DIS=\langle U,C\cup\{d\},V,f\rangle$,令 $A_1,A_2,\cdots,A_m$ 是 C 的 m 个属性子集,$A=\{A_1,A_2,\cdots,A_m\}$,$0<\beta\leqslant 1$,$D=\{D_1,D_2,\cdots,D_i\}$是由决策属性集$\{d\}$在论域 U 上导出的划分,则决策类 D 对属性子集 A 的依赖度因子 γ_β 与多粒度糙集对应的度量因子有如下的关系成立。

$$\gamma_P\left(\sum_{i=1}^m A_i,X\right)\leqslant\gamma_\beta\left(\sum_{i=1}^m A_i,X\right)\leqslant\gamma_O\left(\sum_{i=1}^m A_i,X\right)$$

说明:性质(1)和性质(2)由定理 6 很容易得到。

结束语 多粒度粗糙集模型是 Qian 从粒计算的角度出发提出的一种新的粗糙集模型,笔者在深入地研究了乐观和悲观多粒度粗糙集性质的基础上,提出了可变粒度粗糙集模型,该模型是 Qian 的两种多粒度粗糙集模型的泛化。笔者将进一步研究可变粒度粗糙集的知识获取和约简等问题。

参 考 文 献

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1984, 11(5): 341-356

[2] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112(1-4): 39-49

[3] 王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展,2002,39(10):1238-1243

[4] Stefanowski J, Tsoukias A. Incomplete information tables and rough classification[J]. Computational Intelligence, 2001, 17(3): 545-566

[5] 苗夺谦,王国胤,刘清,等. 粒计算:过去、未来和展望[M]. 北京:科学出版社,2007

[6] Qian Y H, Liang J Y, Yao Y Y C, et al. MGRS: a multigranulation rough set[J]. Information Sciences, 2010, 180(5): 949-970

[7] Qian Y H, Liang J Y, Dang C Y. Incomplete multigranulation rough set[J]. IEEE Transactions on Systems, Man and Cybernetics, Part A, 2010, 40(2): 420-431

[8] Qian Y H, Liang J Y, Wu W. Pessimistic rough decision[C]// Second International Workshop on Rough Sets Theory. Zhoushan, P. R. China, October 2010: 440-449