第 38 卷 第 2 期
Vol.38 No.2

计 算 机 工 程
Computer Engineering

2012 年 1 月
January 2012

• 软件技术与数据库 •　　文章编号：1000—3428(2012)02—0051—03　　文献标识码：A　　中图分类号：TP311

# 基于顺序表的启发式属性约简算法

梁宝华，汪世义，蔡 敏

(巢湖学院计算机科学与技术系，安徽 巢湖 238000)

**摘 要：** 利用顺序表存储数据集对象，并借助基数排序按关键字“分配”思想，求解 $U/C$ 的时间复杂度为 $O(|P||U|)$、空间复杂度为 $O(U)$。在求属性约简集时，为避免存储差别矩阵所需的大量空间，利用差别矩阵的直观性，给出一种计算差别对象个数公式，并以此为启发信息，设计 2 种动态约简算法，其时间/空间复杂度分别为 $O(|C|^2|U/C|)$、$\max(O(|U/C_i|))$。理论分析与实验结果表明该算法是有效可行的。

**关键词：** 粗糙集；属性重要性；差别矩阵；顺序表；启发式

# Heuristic Attribute Reduction Algorithm Based on Order Table

LIANG Bao-hua, WANG Shi-yi, CAI Min

(Department of Computer Science and Technology, Chaohu College, Chaohu 238000, China)

**【Abstract】** Using order list to store data set objects and borrowing the idea of allocation by keys in radix sorting, its time and space complexity for $U/C$ is $O(|P||U|)$ and $O(U)$ respectively. To avoid large space to store discernibility matrix and use the intuition of it, a expressions to compute the number of discernibility objects is presented when computing attribute reduction sets. Two algorithms are designed with time and space complexity only $O(|C|^2|U/C|)$ and $\max(O(|U/C_i|))$. Theoretical analysis and experimental results show that the algorithm is effective and feasible.

**【Key words】** rough set; attribute importance; discernibility matrix; order table; heuristic

**DOI:** 10.3969/j.issn.1000-3428.2012.02.016

## 1 概述

粗糙集理论是一种处理不精确、不确定性知识的数学工具，并广泛应用于机器学习、人工智能、数据挖掘等领域[1]。属性约简是粗糙集理论的核心内容之一，是指在保持分类能力不变的前提下，以基于最少的条件属性和最小冗余的属性值，导出最少的决策规则或分类规则[2]。属性约简通常不是唯一的，如何找到最佳的约简是一个 NP-hard 问题，解决此问题常采用启发式方法。

目前，属性约简主要分为基于正区域的算法和基于分辨矩阵的约简算法。前者的属性约简算法中，最差的时间复杂度从文献[3]的 $O(|C|^3|U|^2)$，到文献[4]的 $O(|C|^2|U|^2)$，后来，采用快速排序技术，使复杂度降为 $O(|C||U|\log|U|)$[5]，甚至采用基数排序技术降为更低的 $O(|C||U|)$[6]。对于那些基于分辨矩阵的约简算法，通过对不同决策类间的对象进行逐一比较建立区分矩阵，一般算法的时间复杂度不低于 $O(|C|^2|U|^2)$。本文以顺序表方式存储数据集，结合基数排序的思想，设计一个总的时间复杂度为 $O(|C||U|)$，并在此基础上给出了 2 种动态约简算法。

## 2 Rough Set 的概念

**定义 1** 五元组 $S=(U,C,D,V,f)$ 是一个决策表[6]，其中，$U$ 为对象的集合；$C$ 为条件属性的非空有限集；$D$ 表示决策属性的非空有限集，且 $C\cap D\neq\varnothing$，$V=\bigcup_{a\in C\cup D}V_a$，$V_a$ 为所有属性的值域集合，$f:U\times(C\cup D)\to V$ 是一信息函数，$\forall a\in C\cup D, x\in U$，有 $f(x,a)\in V_a$；每个属性子集 $P\subseteq(C\cup D)$ 决定了一个二元不可区分关系 $IND(P)$：

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

其中，关系 $IND(P)$构成 $U$ 的一个划分，用 $U/IND(P)$ 表示，简记 $U/P$，则 $[x]_p=\{y\mid\forall a\in P, f(x,a)=f(y,a)\}$ 称等价类。

**定义 2** 在决策表 $S$ 中，任意属性子集 $P$，$Q\subseteq(C\cup D)$，记 $U/P=\{P_1,P_2,\cdots,P_k\}$，$U/Q=\{Q_1,Q_2,\cdots,Q_l\}$，若 $\forall P_i\in U/P\Rightarrow\exists Q_j\in U/Q$，使 $P_i\in Q_j$，则称 $U/P$ 为 $U/Q$ 的加细[6]。

**定义 3** 在决策表 $S$ 中任意属性子集 $P$，$U/P=\{P_1,P_2,\cdots,P_k\}$，$|P_i|$ 为划分子类的基数，即子类中包含对象的个数。在 $P_i$ 中每 2 个对象组成一对，共有 $C_{|P_i|}^2$ 对，简记 DistinctN。在同一划分类中的每 2 个对象不能区分，所以 $U/P$ 中不可区分对象共有 $\sum_{i=1}^{k}C_{|P_i|}^2$，简记 $Sum\_DN(P)$(其中，$k$ 为 $U/P$ 划分类的个数，若某一划分子类只有一个对象，则不参加计算，因为不属于不可区分对象范畴)。

## 3 $U/C$ 的快速计算算法

对对象集进行约简，首先是区分对象集中的元素，即求不可区分关系 $INP(P)$。对于 $INP(P)$的计算直接影响约简算法的复杂度。一般区别对象集时间复杂度为 $O(|P||U|^2)$，后来文献[5]采用快速排序方法，使计算 $INP(P)$的时间复杂度降为 $O(|P||U|\log|U|)$，文献[6]采用基数排序技术，使计算 $INP(P)$ 的时间为 $\left||P||U|+\sum_{i=1}^{s}(M_i-m_i+1)\right|\leqslant|P||U|+|P||U|$，时间复杂度降为 $O(|P||U|)$。

经过对计算 $INP(P)$的研究，采用顺序表形式存储对象集，利用顺序表的下标，结合基数排序的思想，给出一个计算 $INP(P)$时间复杂度也为 $O(|P||U|)$，但比文献[6]少了每次要按关键字收集和最后的一趟收集并判断相邻对象是否为同一类的步骤，效率有所提高。

**基金项目：** 安徽省高校重点自然科学研究计划基金资助项目(KJ2008A35ZC)

**作者简介：** 梁宝华(1973－)，男，讲师、硕士，主研方向：数据挖掘，粗糙集理论；汪世义，副教授、博士；蔡 敏，讲师

**收稿日期：** 2011-04-16　**E-mail：** liangbh426@126.com

**算法 1**

**输入** 决策表 S($U$, $P$, $D$)

$U=\{x_1,x_2,\cdots,x_n\}, P=\{p1,p2,\cdots,ps\}, U'=\varnothing$

**输出** $U/P$

**Step1** $U'=\varnothing$

**Step2** for($k$=1；$k$<$s$+1；$k$++)

{ if($U'$!=$\varnothing$)

{ 在 $UT$ 中剩下未能区分的对象中，分别划分每个子类再按属性 $pk$ 划分；//若 $U[i]_{pk}=U[j]_{pk}$，则$\{i,j\}$为同一划分子//类中的对象；

删除 $U$ 中划分基数为 1 的子类并将此对象加入到 $U'$；

} }

**Step3** 在将 $U$ 中未区分的对象，取每划分子类的第 1 个对象，并入 $U$’； //此时 $U$’为约简的决策表

(1)算法的时间复杂度分析：由算法可知，循环次数为|$P$|(|$P$|为约简属性个数)，在进行计算 $U/p_i$ 时，在原来划分的基础上进一步细划每一个子划分，由于每次进行细划要去掉基数为 1 的划分子类对象，因此每次划分的对象数目≤|$U$|(|$U$|为要约简的对象个数)，所以，Step2 的时间复杂度小于等于 $O(|P||U|)$。Step3 的时间复杂度为 $O(1)$。所以，算法 1 总的时间复杂度为 $O(|P||U|)$。与文献[6]的基于基数排序算法复杂度相比，效率略有提高，空间复杂度为 $O(|U|)$。

(2)实例分析：对如表 1 所示的所有对象，用算法 1 计算 $U/C$，其中，$C$={$a$, $b$, $c$, $d$})。

**表 1 决策表**

| $U$ | $a$ | $b$ | $c$ | $d$ | $D$ |
|---|---|---|---|---|---|
| $U$[1] | 1 | 0 | 1 | 1 | 1 |
| $U$[2] | 2 | 2 | 2 | 3 | 2 |
| $U$[3] | 3 | 2 | 3 | 1 | 1 |
| $U$[4] | 3 | 1 | 1 | 3 | 1 |
| $U$[5] | 4 | 2 | 2 | 2 | 2 |
| $U$[6] | 1 | 0 | 1 | 1 | 1 |
| $U$[7] | 3 | 1 | 2 | 3 | 3 |
| $U$[8] | 2 | 0 | 4 | 2 | 4 |
| $U$[9] | 3 | 2 | 3 | 1 | 2 |
| $U$[10] | 2 | 2 | 2 | 3 | 2 |

将 $U$ 中所有对象的 $A$1 属性值进行分类，且在同一类的对象以在顺序表中存放位置的下标代替该对象，并统计每一类中对象个数，$U/A$1 划分过程如下：

1)$U/a$：因 $U$[1]$a$=$U$[6]$a$=1，则{$U$[1],$U$[6]}为同子类，且基数为 2。同理，{$U$[2],$U$[8],$U$[10]}为一子类，基数为 3。{$U$[3],$U$[4],$U$[7],$U$[9]}为一子类，基数为 4。{$U$[5]}为一子类，且基数为 1。在 $U$ 中删除子类基数为 1 的{$U$[5]}，将之并入 $U$’，得 $U$’={$U$[5]}，此时 $U$={{$U$[1],$U$[6]}, {$U$[2],$U$[8],$U$[10]}, {$U$[3],$U$[4],$U$[7],$U$[9]}}。

2)同理可得，在剩下的未区分的每个划分子类再分别对 $b$ 划分得(限于篇幅，过程简略)：

$U$={{$U$[1],$U$[6]},{$U$[2],$U$[10]},$U$[3],$U$[9]},{$U$[4],$U$[7]}}

$U$’={$U$[5], $U$[8]}

3)在剩下的未区分的每个划分子类再分别对 $c$ 划分得：

$U$={{$U$[1],$U$[6]}, {$U$[2], $U$[10]}, {$U$[3], $U$[9]}}

$U$’={$U$[5], $U$[8], $U$[4], $U$[7]}

4)同理，对剩余的对象进行 $d$ 划分，未发现有新的基数为 1 的类产生，划分结果与(3)相同。

$U$={{$U$[1],$U$[6]}, {$U$[2], $U$[10]}, {$U$[3], $U$[9]}}

对于最终划分结果中子类基数大于 1 的划分取每个子类中的第 1 个对象，再合并 $U$’中的所有对象得简化决策表：

$T$={$U$1,$U$2,$U$3,$U$4,$U$5,$U$7,$U$8}

(3)同类算法比较：目前在求 $U/C$ 时，效率较高的是以基数排序为基础的文献[6]算法。对于表 1 对应的决策表，文献[6]需要 227 次比较，本文算法仅需 33 次比较，相对计算时间是文献[6]近似为 1/7，因为在算法过程中，采用在原有划分子类的基础上进行加细，动态删除一些基数为 1 划分子类，从而节约大量时间。

## 4 快速属性约简算法

基于差别矩阵约简算法一般时间复杂度为 $O(|C|^2|U|^2)$，且空间复杂度为 $O(|C||U|^2)$，优点是算法实现简单，且意义直观、明确。差别矩阵算法最大的缺点是要花大量空间去存储矩阵信息。经深入研究，本文利用差别矩阵的潜在信息，但不建立差别矩阵，提出基于不可区分对象启发式及可区分对象启发式算法。

要衡量某属性的重要性，即看哪个属性能够区分数据集 $U$ 的能力强。要区分数据集中的每个对象，最原始方法就是将某对象与其他所有对象一一比较，看对应的属性值是否相等，这样查找次数太多。可从反面看，找出相对于某属性的划分中，不能区分对象易找，因为只要在同一划分子类中的对象均不可区分。这样，若某个划分子类的基数为 $n$，则不可区分对象个数为 $C_n^2=n(n-1)/2$，简记 DistinctN。若某划分子类的基数为 1，则直接可区分，不参加不可区分数的计算。再对每个属性划分的子类求 $Sum\_DN(C_i)$ (见定义 4)，若某属性的 $Sum\_DN(C_i)$ 越小，说明不可区分对象数越小，或可区分对象数越大，说明此属性越重要。

属性约简目的是为了能用最小属性集区分每一个对象。

**算法 2** 基于不可区分对象启发式算法

**输入** $U$’简化决策表，条件属性 $C$

**输出** 约简属性集 $P$

**Step1** $P=\varnothing$；

**Step2** for($i$=1;$i$≤|$C$|;$i$++){

求 $U'/C_i$，并统计每个划分子类的基数，计算其 Distinct$N$，再按每个属性划分的子类求 $Sum\_D(C_i)$，找出最小 $Sum\_DN(C_i)$ 的属性，若有多个最小值，则任选一个；}

**Step3** $C=C-C_j$， $P\cup C_j$；

//其中，$c_j$ 为 Step2：得到 $Sum\_DN()$ 值最小的第 $j$ 个//属性；

**Step4** 在 $U'$ 删除基数为 1 的划分子类

While($U'\neq\varnothing$){

对 $U$’在 $C$ 剩下的每个属性求 $Sum\_DN(C_i)$，并选择最小的那个属性；

在 $U$’中删除基数为了 1 的划分子类；$C=C-C_i$；$P\cup C_j$；}

最后 $P$ 中的属性集即为最小约简集。

(1)复杂度分析：Step2 计算 $U'/C$ 和 $Sum\_D(Ci)$，共循环次数为|$C$|，时间复杂度为 $O(|C||U/C|)$，Step3 为 $O(1)$，Step4 为：$O(|C-1||U_1'/P_{k1}|)+O(|C-2||U_2'/P_{k2}|)+\cdots+O(|U_c'/P_{kc}|)$。

因为 $|U_i'/P_{ki}|\leqslant|U'/C|$，所以总的时间复杂度为：$O(|C|^2|U/C|)$，空间复杂度为 max($O(|U/C_i|)$)。

(2)实例分析：为方便叙述，以下对象均以对象在 $U$ 中的

下标代替，$P=\varnothing$：

$U'/a$ ={{1},{2,8},{3,4,7}{4}}, $Sum\_DN(a)$=4

$U'/b$={{1,8},{4,7},{2,3,5}}, $Sum\_DN(b)$=5

$U'/c$={{1,4},{2,5,7},{3},{8}},$Sum\_DN(c)$=4

$U'/d$={{1,3},{5,8},{2,4,7}},$Sum\_DN(d)$=5

所以，取 $P$={$a$}，剩下待划分的未区分对象为{{2,8},{3,4,7}}$\neq\varnothing$，进行新的划分：

{{2,8},{3,4,7}}/$b$={{2},{8},{3}{4,7}},S$um_DN(b)$=1

{{2,8},{3,4,7}}/$c$={{2},{8},{3},{4},{7}},$Sum\_DN(c)$=0

{{2,8},{3,4,7}}/$d$={{2},{8},{3}{4,7}},$Sum\_DN(d)$=1

因有一属性的 $Sum\_DN(c)$=0, $P$={$a$, $c$}，说明有一属性已完全区分所有对象，算法终止，所以得约简为{$a$,$c$}

**定理** 若 $U/C_i=\{K_1,K_2,\cdots,K_l\}$，$|K_i|$表示子划分 $K_i$ 中对象的个数，$1\leqslant i\leqslant l$，在这 $l$ 个子划分中，可产生相互区分的对象共有对数简记为：

$$Sum\_D(C_i)=|K_2|\times|K_1|+|K_3|\times(|K_2|+|K_1|)+\cdots+|K_l|\times(|K_{l-1}|+|K_{l-2}|+\cdots+|K_2|+|K_1|)$$

证明：限于篇幅，证明从略。

要衡量某属性的重要性，可直接根据某属性的划分子类中，能够相互区分对象的数目，若区分的对象数越大，说明此属性越重要。

**算法 3** 基于可区分对象启发式算法

**输入** $U'$简化决策表

**输出** 约简属性

**Step1** $P\neq\varnothing$; sum=0;//sum 为已区分对象的数目

**Step2** while($U'\neq\varnothing$)

{ 求 $U'/C_i, C_i\in\{C-P\}$；

并求出 max(fun($C_j$))；

//同时找出 fun( )值最大的属性；

$P=P\cup C_j$； // $C_j$ 为刚刚得到 fun($C_j$)最大的那个属性

删除 $U'$中基数为 1 的划分子类； }

int fun( $U'$,$P$ ){ //$U'$为对象集，$P$ 为属性

int $S_1$=0;// $s_1$ 表示所有不可区分对象对数

int $S_2$=0;// $s_2$ 表示将按每个属性划分后，子类间产生不可区分对象的对数

//设 $U'/P$ 得到 $k$ 个划分，每个子划分类的基数分别为 $p_1,p_2,\cdots,p_k$

if( $k$=1) return 0; //无需划分；

else if($k$=2) return $p_1\cdot p_2$;

else { $S_1=p_1\cdot p_2$;

$S2=p_1$;

for(int $i$=3;$i\leqslant k$;$i$++)

{ $S_2+=p_{i-1}$;

$S_1+=S2\cdot p_i$;}

return $S_1$; } }

(1)时间/空间复杂度分析

为了叙述方便，使用了 fun()函数，其实此函数可以与求 $U'/C$ 同时操作。所以，总的时间复杂度与算法一样，为 $O(|C|^2|U/C|)$，空间复杂度为 max( $O(|U/C_i|)$)。

(2)实例分析

简化决策表：

$U'$={1,2,3,4,5,7,8}, $P=\varnothing$;

求 $C$-$P$ 中所有属性的 fun()值得：

$U'/a$ ={{1},{2,8},{3,4,7},{5}}, fun($a$)=17

$U'/b$ ={{1,8},{4,7},{2,3,5}}, fun($b$)=16

$U'/c$ ={{1,4 },{2,5,7},{3},{8}}, fun($c$)=17

$U'/d$ ={{1},{5,8},{2,4,7},{3}}, fun($d$)=17

因 fun($a$)，fun($c$)，fun($d$)中均为 17，故任取一个属性，设取 $a$，$P$={$a$}；删除 $U'/a$ 划分基数为 1 的子类，得 $U'$={{2,8},{3,4,7}}重复上述过程，在 $U'$={{2,8}, {3,4,7}}中再求 fun()值：

$U'/b$ ={{2},{8},{3},{4,7}}, fun($b$)=9

$U'/c$ ={{2},{8},{3},{4},{7}}, fun($c$)=10

$U'/d$ ={{2},{8},{3},{4,7}}, fun($d$)=9

因为 fun($c$)=10 最大，所以取 $c$, $P$={$a$, $c$ }。

删除{{2},{8},{3},{4},{7}}基数为 1 的子类，得 $U'=\varnothing$，算法终止。

综合以上分析，算法 2、算法 3 得到同样的约简结果。

(3)同类算法比较：在决策表 1 简化数据集上求属性约简，经计算，算法 2 效果一样需要计算和比较 44+17=61 次，算法 3 需要计算 58+45=103 次，文献[6]需要 63 次计算和比较。算法 2 与文献[6]提出的算法相仿。算法 3 与算法 2 相比，虽然复杂度也为 $O(|C|^2|U/C|)$，但每次计算，算法 2 对于基数为 1 的划分子类不参加运算，从而减少了计算次数，为算法节省了时间。算法 2、算法 3 为求属性约简集提供了新的角度考虑问题。

## 5 结束语

本文针对差别矩阵算法的缺点，给出 2 种高效算法，提出计算区分对象个数的计算公式。这样既具有了基于差别矩阵算法的直观性，同时又避免了用大量空间去存储差别矩阵，时间复杂度为 $O(|C|^2|U/C|)$。因此，新算法无论在时间上还是空间上均优于传统的约简算法。下一步的工作是研究适用于动态增长的数据约简算法。

## 参考文献

[1] Pawlak Z. Rough Set[J]. Communication of the ACM, 1995, 38(11): 89-95.

[2] 曾黄麟. 粗集理论及其应用——关于数据推理的新方法[M]. 重庆: 重庆大学出版社, 1996.

[3] Hu Xiaohua, Nick C. Learing in Relational Database: A Rough Set Approach International[J]. Journal of Computational Intelligence, 1995, 11(2): 323-338.

[4] 韩智东, 王志良. 基于相容矩阵的改进属性约算法[J]. 计算机工程, 2010, 36(1): 49-50.

[5] 刘少辉, 盛秋戬, 吴 斌, 等. Rough 集高效算法的研究[J]. 计算机学报, 2003, 26(5): 524-529.

[6] 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.

编辑 索书志