

# 基于粗糙集理论的条件属性动态约简算法

覃伟荣, 秦亮曦

(广西大学 计算机与电子信息学院, 广西 南宁 530004)

**摘要:**粗糙集理论是一种新的处理含糊和不确定性问题的数学工具,可以有效地分析和处理不完备信息。条件属性约简是粗糙集理论算法研究的重点。在启发式条件属性约简算法的基础上提出了动态条件属性约简算法,算法以一个信息大的属性作为基础,不断添加条件属性,并对新增加的条件属性进行修正,找到约简条件属性,目的是为了进行遥感数据的动态分类做基础。文中在 VC++ 6.0 开发环境下实现了两种算法,用 HSV 和 Iris 数据验证了算法的有效性,并分析了算法的时间和空间复杂度。

**关键词:**粗糙集;动态约简;信息量;条件属性约简

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2008)08-0023-03

## An Algorithm of Condition Attribute Dynamic Reduct Based on Rough Set

QIN Wei-rong, QIN Liang-xi

(School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

**Abstract:** Rough set theory is a relatively new soft computing tool to deal with vagueness and uncertainty. It has received much attention of the researchers around the world. Condition attribute reduct algorithm is the key point of rough set research. A new dynamic condition attribute reduct is established based on the heuristic condition attribute reduct, and it realizes two kinds algorithm under the VC++ 6.0 development environments. A trial was made to confirm the validity of the algorithm by the use of HSV and Iris data. Also analyzes the time and spatial complexity of the algorithm.

**Key words:** rough set; dynamic reduct; information quantity; condition attribute reduct

### 0 引言

粗糙集理论(Rough Set)由波兰科学家 Z. Pawlak<sup>[1]</sup>在 1982 年首先提出。粗糙集理论是数据挖掘研究的重点热点问题,并成为数据挖掘新的处理数据的数学工具。数据知识与发现的主要研究对象是关系型数据库,关系表可被看作为粗糙集理论中的决策表,这给粗糙集方法的应用带来极大的方便。粗糙集理论与其它方法的比较的优点,如神经网络不能自动地选择合适的属性集,而利用粗糙集方法进行预处理,去掉多余属性,可提高发现效率,降低错误率。粗糙集方法比模糊集方法或神经网络方法在得到的决策规则和推理过程方面更易于被证实和检测。

粗糙集理论研究中,条件属性约简是研究的重点。

条件属性约简的算法很多,其中动态约简在某种意义上是给定决策表中最稳定的约简,它们是在从给定的决策表中随机抽样形成的子表中最常出现的约简,动态约简能够有效地增强约简的抗噪音能力<sup>[2]</sup>。

动态约简计算过程较为简明,主要对决策表采样,然后对采样后的决策表计算所有的约简,在所有的子表中保持不变或者近似保持不变的约简就是动态约简。动态属性约简具有约简稳定和抗噪音等优点。粗糙集的条件属性的约简对噪声非常的敏感,为了处理这个问题,将数据库从一个很小的不完备的数据库开始建立属性之间的关系,随着数据库的扩展逐渐修正属性关系就显得十分必要,这就建立“动态”模型。动态算法就是在增加条件属性的基础上对原有的条件属性的关系进行修正,静态算法就是增加条件属性的时候需要重新刷新数据,把数据表当成一个新的表格来处理,增加了搜索时间,动态算法的优点就是减少计算的重复,提高算法的搜索效率。

文中对粗糙集理论中的静态数据库出现的问

收稿日期:2007-11-22

基金项目:广西自然科学基金(桂科自 0728032)

作者简介:覃伟荣(1981-),女,硕士研究生,主要研究方向为数据挖掘;秦亮曦,博士,硕士生导师,主要研究方向为数据挖掘、进化计算、信息管理系统。

题<sup>[3]</sup>,还有日前动态约简算法体现在增加记录的基础上进行动态叠加运算,刘振华<sup>[4]</sup>和彭黎黎<sup>[5]</sup>等对条件信息的约简算法研究取得一定的效果,但针对条件属性的算法很少,从而提出了条件属性动态约简算法,并与启发式条件属性约简算法进行比较,在动态约简的基础上提供算法的改进方案。

## 1 粗糙集理论的概念

### 1.1 信息系统

定义 1 信息系统。粗糙集理论中的信息系统可用一个四元组来表示:  $S = \{U, A, V, f\}$ 。如果该属性表同时为决策表,则在  $A$  属性中可以进一步分为条件属性  $C$  和决策属性  $D$ ,  $A = C \cup D$ ;  $V$  是属性值组成的集合;  $f$  是属性和记录的函数,  $f(a, e)$  的值确定记录  $e$  关于属性  $a$  的取值。

### 1.2 约简与核

定义 2 约简。设  $S = \{U, A, V, f\}$  是一个信息系统,  $B \subseteq A$ , 且属性  $a \in B$ 。若  $\text{IND}(B) = \text{IND}(B - \{a\})$ , 则称属性  $a$  在族集  $A$  中是可以省略 (Dispensable) 的, 否则就是不可以省略的。如族集  $A$  中的每个属性  $a$  都是不可以省略的, 则称  $R$  是独立的 (Indispensable), 否则就是依赖的或不独立的。进一步定义, 若  $Q \subseteq P$  是独立的, 并且  $\text{IND}(Q) = \text{IND}(P)$ , 则称  $Q$  是关系族集  $P$  的一个约简 (Reduce)。在族集  $P$  中所有不可省的关系的集合称为  $P$  的核 (Core), 用  $\text{Core}(P)$  来表示。不难看出,  $P$  有多个约简。且  $\text{Core}(p) = \bigcap \text{red}(P)$ , 其中的  $\text{red}(P)$  是所有  $P$  约简的族集。

定义 3 核。相对于属性集合  $D$ , 属于所有属性集合  $C$  的所有规约的交集的属性集合称为属性集  $C$  的核, 记为  $\text{Core}(C, D)$ 。用核作为计算规约集的起点, 可以化简属性规约集。为简化计算核, 一般通过分辨矩阵进行。

定义 4 正区域  $\text{POS}_C(D) = \{L_x \mid x \in U/\text{IND}(B)\}$

### 1.3 差别矩阵

定义 5 差别矩阵由华沙大学数学家 Skowron 提出, 对一个新信息系统  $S\{U, A, V, f\}$ ,  $a(x)$  是  $x$  在属性  $a$  上的值, 属性集  $B \subseteq A$ ,  $|U/\text{IND}(B)| = n$ , 则差别矩阵 (Discernibility Matric) 为  $M_D(B) = \{m_{i,j}\}_{n \times n}$ ,  $1 \leq i, j \leq n$ , 其中,  $\{m_{i,j}\}$  是  $x_i$  与  $x_j$  存在差别的所有属性构成的集合。

$$(C_{i,j}) = \begin{cases} a \in A : a(x_i) \neq a(x_j), D(x_i) \neq D(x_j) \\ 0 : \emptyset, D(x_i) = D(x_j) \\ -1 : a(x_i) = a(x_j), D(x_i) \neq D(x_j) \end{cases} \quad (1)$$

### 1.4 属性的重要性

在属性约简表格中, 条件属性集  $C$  和决策属性集  $D$  之间的依赖程度可以定义为:

$$\gamma_C(D) = \frac{\text{card}(\text{POS}_C(D))}{\text{card}(C)} \quad (2)$$

### 1.5 知识的条件信息量

定义 6 信息量定义为:

$$I(D' \mid C) = \sum_{i=1}^n \frac{|X_i|}{|U|} \sum_{j=1}^m \frac{|X_i \cap Y_j|}{|X_i|} (1 - \frac{|X_i \cap Y_j|}{|X_i|}) \quad (3)$$

其中:  $X_i$  为条件属性的子集,  $Y_j$  为决策属性的子集,  $C$  是条简属性,  $D$  为决策属性。等价类的定义:

$$U/\text{IND}(C) = \{X_1, X_2, \dots, X_n\}$$

$$U/\text{IND}(D) = \{Y_1, Y_2, \dots, Y_m\}$$

## 2 启发式条件属性约简算法

算法 1: 信息系统表有多个条件属性和一个决策属性值, 杜晓昕等人<sup>[6]</sup>利用属性的重要度作为启发式信息快速求得条件属性的约简集。

① 令初始属性集  $P = \emptyset$ , 计算决策属性对每一个条件属性的依赖性, 按照依赖的大小对属性进行排序, 将依赖性最大的属性  $S$  加入属性约简集,  $P = P \cup \{S\}$ , 如果有多个属性的依赖性相等, 则选择属性值少的属性加入  $P$  中, 如果属性值的个数都一样, 就按顺序加入  $P$  中。

② 若  $\text{POS}_P(D) = \text{POS}_C(D)$ , 则结束运算, 取  $P$  为一个属性约简集, 否则, 计算  $P$  之外的属性加入到  $P$  的重要属性, 按重要性大小对属性进行排序, 得一排序集  $M$ 。

③ 从  $M$  中取重要的属性  $S$  加入属性集  $P = P \cup \{S\}$ , 如果有多个属性的重要性相等, 则选择属性值少的属性加入  $P$  中, 若  $\text{POS}_P(D) = \text{POS}_C(D)$ , 则结束运算。否则, 转入步骤 ③ 继续计算。

## 3 条件属性动态约简算法

算法 2: 根据上述的算法, 提出条件属性动态约简算法;

// 输入一个有条件属性和决策属性的表格

// 输出一个属性表约简表, 包括约简条件属性和决策属性

① 先用差分矩阵计算整个表的核条件属性;

② 计算  $U/\text{IND}(C) = \{X_1, X_2, \dots, X_n\}$ ,  $U/\text{IND}(D) = \{Y_1, Y_2, \dots, Y_m\}$ , 并计算  $I(D' \mid C) = \sum_{i=1}^n \frac{|X_i|}{|U|} \sum_{j=1}^m \frac{|X_i \cap Y_j|}{|X_i|} (1 - \frac{|X_i \cap Y_j|}{|X_i|})$

计算  $I(D|C)^{[7]}$ :首先列出  $|X_i \cap Y_j| = \{|X_1 \cap Y_j|, |X_2 \cap Y_j|, \dots, |X_n \cap Y_m|\}$ , 加入方程进行计算。

③ 计算不是核的条件属性对决策属性的依赖度;  
// 如何计算  $\gamma_c(D)$ ,  $U/IND(C) = \{X_1, X_2, \dots, X_n\}$ ,  
 $U/IND(D) = \{Y_1, Y_2, \dots, Y_m\}$ , 计算  $X_i \subseteq Y_j$  的个数  
 $|X_i \subseteq Y_j|$ , 即正区域  $POS_C(D)$ ,  $\gamma_c(D) = \frac{\text{card}(POS_C(D))}{\text{card}(C)}$

④ 计算  $I(D|\text{core}(C, D))$ //这是核条件属性与决策属性的互信息量, 添加新的条件属性就修正  $I(D|\text{core}(C, D))^{[8]}$

⑤ 如果  $I(D|\text{core}(C, D)) = I(D|C)$ , 说明它们有同等的分类能力, 结束计算, 否则, 将属性依赖度大的条件属性加入核条件属性表;

⑥ 重复步骤 ⑤, 直到结束。

#### 4 算法的测试

用 VC++ 6.0 开发软件实现了启发式属性约简算法和动态条件属性约简算法的平台。所有的数据采用文本文件的型式保存, 最后的结果除了在 VC 结果环境下可视以外, 也保存在文本文件中, 目的是为了进一步做规则分类提取算法。使算法具有更好的容错性, 以及针对海量数据的特点, 采用了动态存储方式。将 HSV 和 Iris 两种数据用记事本保存, 对两种算法进行了测试, 两个算法分别对 HSV 和 Iris 数据约简表一样。两种算法的时间复杂度和空间复杂度如表 1 所示。 $r$  是数据记录行数,  $m$  ( $m = \text{column}$ ) 是每次的比较次数,  $a, b$  代表每次的比较次数。

动态条件属性约简算法刚开始可以不采用差分矩阵来求核条件属性, 这里用差分矩阵的目的是为进一步的规则分类做铺垫。互信息的公式计算可以采用迭代<sup>[8]</sup>的方式计算, 可以保留上一次的信息, 只计算新增加的信息, 这里保留原始的计算在差分矩阵中差别, 差别取反就是两两记录的交集。这样可节约重复的计算

(上接第 22 页)

备字库开发能力的人也能够通过技术手段从字库中删除签名部分, 再修改字库, 甚至重新对字库进行签名。加密和压缩可以解决上述问题, 防止对字库的再次签名和修改, 但这仍然有待于未来 Windows 操作系统的支持。

#### 参考文献:

- [1] Stallings W. 密码编码学与网络安全——原理与实践[M]. 第3版. 刘玉珍, 王丽娜, 傅建明译. 北京: 电子工业出版社

时间。

#### 5 结束语

在启发式属性约简算法的基础上, 提出了动态条件属性约简算法, 并在 VC++ 6.0 环境下实现了两个算法, 动态条件属性的算法是利用了新增加的条件属性列对原有的互信息进行修正。实验证明算法是实用有效的。

表 1 两种算法的比较

	启发式属性约简算法	动态属性约简算法
空间复杂度	$O( C )$	$O( C )$
时间复杂度	$r(\frac{r}{2} - 1)m^2 + 2Xbr(\frac{r}{2} - 1)$	$r(\frac{r}{2} - 1)m + Xar(\frac{r}{2} - 1)$

#### 参考文献:

- [1] Pawlak Z. Rough set[J]. International Journal of Information and Computer Science, 1982(11):341-356.
- [2] Bazan J G, Skowron A, Synak P. Dynamic reducts as a tool for extracting lows from decisions tables [M]//Ras Z W, Zemankiva M. Methodologies for Intelligent Systems. Berlin: Springer-Verlag, 1994:346-355.
- [3] 韩斌, 吴铁军, 杨明晖. 结合粗集理论的动态属性约简研究[J]. 系统工程理论与实践, 2002(6):67-73.
- [4] 刘振华, 刘三阳, 王珏. 基于信息量的一种条件属性约简算法[J]. 西安电子科技大学学报: 自然科学版, 2003(6):834-838.
- [5] 彭黎黎, 刘山. 基于信息量的动态属性约简[J]. 计算机工程, 2005(7):104-105.
- [6] 杜晓昕, 徐慧, 任长伟, 等. 基于粗糙集的属性约简在数据挖掘中的应用[EB/OL]. 2006. 中国科技论文在线. <http://www.paper.edu.cn>.
- [7] 王加阳, 陈松乔, 罗安. 粗集动态约简研究[J]. 小型微型计算机系统, 2006(11):2056-2060.
- [8] 刘山, 张慧. 基于条件信息量的动态属性约简方法[J]. 计算机工程, 2007(6):182-183.

社, 2004.

- [2] 贺卫红, 曹毅. RSA 公钥密码体制在数字签名中的应用[J]. 微机发展, 2003, 13(9):49-52.
- [3] 屈喜龙. 基于数字证书的数字签名系统的设计与实现[J]. 计算机工程与应用, 2006, 42(15):189-192.
- [4] Microsoft Corporation. OpenType specification v. 1.2 [DB/OL]. 2002. <http://www.microsoft.com/typography/ot-spec/otff.htm>.
- [5] Microsoft Corporation. TrueType 字型核心技术[M]. 北京: 学苑出版社, 1993.