基于粗集边界域的快速约简算法

黎 敏[1,2] 冯圣中[2] 樊建平[2] 刘 清[3]
(南昌工程学院信息工程学院 南昌 330099)[1] (中国科学院深圳先进技术研究院 深圳 518055)[2]
(南昌大学信息工程学院 南昌 330047)[3]

摘 要 属性约简是粗糙集研究的核心内容之一。已有的大多数属性约简算法都是采用基于正域的贪心算法求决策表的代数约简。事实上,对于不一致决策表,代数约简改变了决策类族原有的Pawlak拓扑结构,造成决策类的不确定性扩大。为此,提出了一种新的基于粗集边界域的约简模型,它能够保持决策类族原有的Pawlak拓扑结构。依据新模型,提出了一种高效率的基于粗集边界域的属性约简算法。理论分析和实验表明,所提算法是有效可行的。

关键词 粗糙集,不一致决策表,不确定性,属性约简,边界域
中图法分类号 TP301 **文献标识码** A

Quick Attribute Reduction Based on Rough Boundary Region

LI Min[1,2] FENG Sheng-zhong[2] FAN Jian-ping[2] LIU Qing[3]
(Institute of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, China)[1]
(Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China)[2]
(Institute of Information Engineering, Nanchang University, Nanchang 330047, China)[3]

Abstract Attribute reduction is one of the core research content of Rough set. Most of the existing greedy reduction algorithm is based on positive region to find out an algebraic reduct. In fact, for an inconsistency decision table, algebra reduct changes the original Pawlak topology and expands the uncertainty degree of decision table. Therefore, in this paper, a novel reduction modal based on rough boundary region was introduced, which can keep the original Pawlak topology. Based on this model, an efficient algorithm for attribute reduction based on rough boundary region was proposed. Theoretical analysis and experimental results show that the algorithm of this paper is effective and feasible.

Keywords Rough set, Inconsistent decision table, Uncertainty, Attribute reduction, Boundary region

粗糙集(Rough Set)[1]是20世纪80年代波兰数学家Pawlak提出的一种新的处理不精确、不确定知识的数学工具,目前已经在机器学习、知识与数据发现、专家系统、模式识别与分类等方面得到了广泛的应用[2-5]。属性约简是粗糙集理论研究的核心内容之一,现有的属性约简方法主要包括基于正区域的算法[6-8]、基于差别矩阵及其改进的属性约简算法[9-11]和启发式信息熵的属性约简算法3种[12-14],这些属性约简算法建立在等价关系的基础之上,适用于离散化的数据集。连续值信息系统的属性约简方法也得到众多关注和研究。例如文献[16]提出一种基于邻域的属性约简算法,该算法需要反复计算和保存各对象的领域,时间和空间开销较大,且存在领域参数选择问题。文献[17]提出一种基于一致性准则的属性约简算法,该算法选择一个相似性函数(距离函数)且仅需计算不同类对象之间的相似性,这在一定程度上减少了计算量,但该算法仍然是计算密集型的。

代数约简的意义在于在保持原有分辨力的前提下,删除冗余条件属性,这对一致决策表是成立的。然而对于非一致决策表,代数约简可能改变原有的Pawlak拓扑[4]结构。由于现实信息的不完整或某些数据预处理步骤的原因,如连续属性离散化,导致不一致决策表客观存在。对于不一致决策表,代数约简仅保持了原正区域对象原有拓扑结构,但可能造成边界域扩大。针对不一致决策表,文献[19]给出了分布约简和分配约简两种定义。分布约简有过于苛刻的要求。文献[20]给出了最大分布约简的概念,但该约简可能导致正区域缩小,而且其推理过程较为繁琐,文献没有给出具体的形式化算法。本文引入基于粗集边界域的约简模型并给出高效的约简算法,其优势在于能够保持所有决策类原有的Pawlak拓扑结构。

1 基础概念

下面简要论述粗糙集的基本概念。

设$U=\{u_1,u_2,\cdots,u_n\}$是一个非空集合,称为论域,R是U

到稿日期:2011-02-13 返修日期:2011-04-03 本文受国家863项目(2007AA120502),江西省科技厅科技支撑项目(2010ZDG03100),江西省教育厅科研项目(GJJ11631)资助。

黎 敏(1975-),男,博士生,副教授,主要研究方向为粗糙集理论、数据挖掘等;冯圣中(1968-),男,研究员,主要研究方向为生物信息学、高性能计算等;樊建平(1963-),男,研究员,博士生导师,主要研究方向为体系结构、并行处理、操作系统;刘 清(1938-),男,教授,主要研究方向为人工智能、粗糙集和粒计算理论。

上的等价关系。记$U/R=\{X_1,\cdots,X_i,\cdots,X_s\}$为$U$的$R$划分，其中$X_i$是$U$关于$R$的一个等价类。对于$U$上的任意子集$X$，记$\underline{R}(X)=\bigcup_{X_i\subseteq X}X_i$和$\overline{R}(X)=\bigcup_{X_i\cap X\neq\phi}X_i$，分别称为$X$的$R$下近似和上近似。即下近似$\underline{R}(X)$是根据知识$R$确定属于$X$的对象全体，上近似$\overline{R}(X)$是根据知识$R$可能属于$X$的对象全体，显然$\underline{R}(X)\subseteq X\subseteq\overline{R}(X)$。如果$\underline{R}(X)=\overline{R}(X)$，称$X$是$R$可定义的，否则称$X$是$R$不可定义的。$X$的$R$边界域为$BN_R(X)=\overline{R}(X)-\underline{R}(X)$，边界域越小，集合$X$的不确定性越小。

决策表$DT=(U,C\cup D,V,f)$是一个四元组，$A=C\cup D$是属性集合，其中C和D分别称为条件属性集和决策属性集，$D\neq A$；V是属性值的集合，f是$U\times A\to V$的映射，它指定了U中每个对象u_i的属性值。对于$\forall B\subset C$，B相对于决策属性D的正区域为$POS_B(D)=\bigcup_{X\in U/D}\underline{B}(X)$。

定义1 给定一个决策表$DT=(U,C\cup D,V,f)$，若$POS_C(D)=U$，称DT是一致决策表，否则称DT为不一致决策表。

定义2 给定一个决策表$DT=(U,C\cup D,V,f)$，$a\in B\subset C$，若有$POS_{B-\{a\}}(D)=POS_B(D)$，称$a$是$B$中不必要的，否则$a$是$B$中必要的。若$B$中的每个属性都是必要的，称$B$为独立的。

粗糙集理论认为，在信息系统或决策表中包含了冗余信息。例如，对于一个信息系统在保持原有分类或决策能力不变的条件下，可能并不需要全部的条件属性。粗糙集理论把能够维持原有的一致性所需要的独立属性子集称为约简。下面是它的形式化定义。

定义3 给定一个决策表$DT=(U,C\cup D,V,f)$，$B\subseteq C$，若$POS_B(D)=POS_C(D)$且B是独立的，称B是C的一个相对约简，也称代数约简。

定义4 给定一个决策表$DT=(U,C\cup D,V,f)$，对于任意$X\subseteq U\wedge X\neq\phi$，$B\subseteq C$，$X$的$B$近似质量定义为：

$$r_B(X)=\frac{|\underline{B}(X)|}{|X|}\tag{1}$$

对于多数分类应用而言，只需找出决策表的一个约简。许多文献采用基于属性重要度的前向贪心搜索算法来寻找决策表的一个相对约简[6,17]。一个被广泛采用的属性重要度定义如下[5]：

$$SGF(a,B,D)=\gamma_{B\cup\{a\}}(D)-\gamma_B(D)\tag{2}$$

$$\gamma_B(D)=\frac{|POS_B(D)|}{|U|}\tag{3}$$

式(2)表示的是目前所选属性集为B时属性a的重要度。式(3)定义了条件属性集B相对决策属性D的依赖度，我们称之为正域依赖度。正域依赖度等于正域和全域的基数之比，实际上也等于各个决策类的加权近似质量之和，如式(4)所示：

$$\gamma_B(D)=\frac{|POS_B(D)|}{|U|}=\sum_{X_i\in U/D}\frac{|X_i|}{|U|}r_B(X_i)\tag{4}$$

式(3)由于比式(4)有着更高的计算效率而被实际使用。

根据式(2)，一个广泛采用的爬山法求决策表的一个代数约简的大致步骤是：开始时约简$red=\phi$，每一次从$C-red$中选择使得$SGF(a,red,D)$最大的条件属性加入到red，直到$POS_{red}(D)=POS_C(D)$。

决策表的一个代数约简是能够保持原有分辨力的一个独立条件属性子集，这对一致决策表是成立的。对于非一致决策表，代数约简仅保持了一致性决策类的Pawlak拓扑结构而无法保证含有不一致对象的决策类保持原有的Pawlak拓扑结构，从而改变这些决策类的近似精度(式(5))。造成这种现象的原因是原有知识划分的多个不一致等价类在代数约简下形成更大的不一致等价类，从而使得含有不一致对象的决策类的上近似扩大。下面的例子说明了这种现象。

例1 设$DT=(U,C\cup D,V,f)$是一个决策表，论域$U=\{u_1,u_2,\cdots,u_{15}\}$；决策类族$U/D=\{X_1,X_2,X_3\}$，其中$X_1=\{u_1,u_2,u_3,u_4\}$，$X_2=\{u_5,u_6,u_7,u_8,u_9,u_{10},u_{11}\}$，$X_3=\{u_{12},u_{13},u_{14},u_{15}\}$；$C$对$U$产生的划分为$U/C=\{E_1,E_2,E_3,E_4,E_5\}$，$E_1=\{u_1,u_2,u_3\}$，$E_2=\{u_4,u_5\}$，$E_3=\{u_6,u_7,u_8,u_{12}\}$，$E_4=\{u_9,u_{10},u_{11}\}$，$E_5=\{u_{13},u_{14},u_{15}\}$。

若存在$B\subset C$，$U/B=\{Z_1,Z_2,Z_3,Z_4\}$，其中$Z_1=E_1$，$Z_2=E_2\cup E_3$，$Z_3=E_4$，$Z_4=E_5$，则$\underline{B}(X_1)=\underline{C}(X_1)=E_1$，$\underline{B}(X_2)=\underline{C}(X_2)=E_4$，$\underline{B}(X_3)=\underline{C}(X_3)=E_5$。若$B$又是独立的，则从定义3可知$B$是一个相对约简。但是上近似$\overline{C}(X_1)=E_1\cup E_2$，$\overline{B}(X_1)=E_1\cup E_2\cup E_3$，$\overline{C}(X_3)=E_3\cup E_5$，$\overline{B}(X_3)=E_2\cup E_3\cup E_5$，因此$BN_B(X_1)\supset BN_C(X_1)$，$BN_B(X_1)\supset BN_C(X_1)$，即相对约简$B$引起决策类$X_1$和$X_2$的边界域扩大。而集合的不确定性是由于边界域的存在而引起的，集合的边界域越大，其精确性越低。因此相对约简B扩大了决策类的不确定性，即改变了决策类原有的Pawlak拓扑结构。

由上述分析可知，对于不一致决策表，代数约简能够保持决策类的下近似不发生变化。但是由于知识的粒度变大，使得含有不一致对象的决策类的边界区域扩大，因此造成这些决策类的近似精度降低。为此，我们引入基于粗集边界域的约简模型。对于非一致决策表，模型能够同时保证所有决策类的上近似和下近似不发生变化，从而保持决策类原有的粗糙近似空间。

2 基于边界域的约简

式(2)定义的属性重要度计算了正域的变化量，仅涉及粗糙集的下近似的概念。但是，若一个对象集合X在两个知识集上所产生的下近似相同但上近似不同，这两个知识集对X的分辨能力是不同的。近似精度[22]综合上下近似考察知识的分类能力，对于任意$X\subseteq U(X\neq\phi)$和$B\subseteq C$，X的B近似精度定义如下：

$$\alpha_B(X)=\frac{|\underline{B}(X)|}{|\overline{B}(X)|}\tag{5}$$

显然$0\leqslant\alpha_B(X)\leqslant 1$。当$\alpha_B(X)=1$时，$\underline{B}(X)=\overline{B}(X)$，$X$的$B$边界域$BN_B(X)=\phi$，集合$X$是$B$可定义的；当$\alpha_B(X)<1$时，$\underline{B}(X)\subset\overline{B}(X)$，$X$的$B$边界域$BN_B(X)\neq\phi$，集合$X$是$B$不可定义的。

近似精度$\alpha_B(X)$表达了知识B对集合X描述的精确程度。由于集合的边界域越大，其不确定性越大，因此X的B粗糙度定义如下：

$$\rho_B(X)=\frac{|BN_B(X)|}{|\overline{B}(X)|}=\frac{|\overline{B}(X)-\underline{B}(X)|}{|\overline{B}(X)|}=\frac{|\overline{B}(X)|-|\underline{B}(X)|}{|\overline{B}(X)|}=1-\alpha_B(X)\tag{6}$$

性质1 对于$\forall X\subseteq U$和$B\subseteq P\subseteq C$，有$\rho_P(X)\leqslant\rho_B(X)$。

证明：对于 $B\subseteq P$，有 $\underline{B}(X)\subseteq\underline{P}(X)\subseteq X\subseteq\overline{P}(X)\subseteq\overline{B}(X)$，即有 $|\underline{B}(X)|\leqslant|\underline{P}(X)|\leqslant|X|\leqslant|\overline{P}(X)|\leqslant|\overline{B}(X)|$，因此 $\frac{|\underline{P}(X_i)|}{|\overline{P}(X_i)|}\geqslant\frac{|\underline{B}(X_i)|}{|\overline{B}(X_i)|}$，即有 $\alpha_P(X)\geqslant\alpha_B(X)$，因此 $\rho_P(X)\leqslant\rho_B(X)$。

性质 1 说明集合的粗糙度随着知识的增加而减小。

在粗糙度定义的基础上，我们给出决策的不确定度的定义。

定义 5 给定一个决策表 $DT=(U,C\cup D,V,f)$ 和 $B\subseteq C$，D 相对 B 的不确定度定义为：

$$UNC_B(D)=\sum_{X_i\in U/D}\frac{|X_i|}{|U|}\rho_B(X_i) \tag{7}$$

即 D 相对 B 的不确定度等于各个决策类的加权粗糙度之和。决策类的权重越大，其粗糙度对整个不确定度的影响越大。

显然，$0\leqslant UNC_B(D)\leqslant 1$。当 $\forall X_i\in U/D\wedge\rho_B(X_i)=0$ 时，$UNC_B(D)$ 获得最小值 0；当 $\forall X_i\in U/D\wedge\rho_B(X_i)=1$ 时，$UNC_B(D)$ 获得最大值 1。D 的 B 不确定度越小，B 对 D 的解释能力就越好。

命题 1 一个决策表 $DT=(U,C\cup D,V,f)$ 是一致决策表，当且仅当 $UNC_C(D)=0$。

证明：若 DT 是一致性决策表，则 $POS_C(D)=U$，并且 $\forall X_i\in U/D$ 有 $\underline{C}(X_i)=\overline{C}(X_i)=X_i$（否则与一致决策表相矛盾），因此 $\forall X_i\in U/D$ 有 $\rho_C(X_i)=0$。从而 $UNC_B(D)=\sum_{X_i\in U/D}\frac{|X_i|}{|U|}0=0$。反过来，若 $UNC_C(D)=0$，则根据定义 5 可知 $\forall X_i\in U/D$ 有 $\rho_C(X_i)=0$（否则 $UNC_C(D)>0$），即 $\forall X_i\in U/D$ 有 $BN_C(X_i)=\phi$，也就是 $\underline{C}(X_i)=\overline{C}(X_i)=X_i$，因此 $POS_C(D)=U$，所以 DT 是一致决策表，即证。

命题 1 说明一致决策表的不确定度是 0。

性质 2 对于 $B\subseteq P\subseteq C$，有 $UNC_P(D)\leqslant UNC_B(D)$。

证明：由定义 5 和性质 1 直接可得。

性质 2 说明，决策的不确定度随着知识的增加而减小。

定义 6 给定一个决策表 $DT=(U,C\cup D,V,f)$，我们称 $B\subseteq C$ 是 C 的一个近似约简，如果 B 满足：(1) $UNC_B(D)=UNC_C(D)$；(2) $\forall B'\subset B, UNC_{B'}(D)>UNC_C(D)$。

上述定义说明，若 B 是 C 的近似约简，则 D 的 B 不确定度等于 D 的 C 不确定度，而对于 B 的任意真子集 B'，D 的 B' 不确定度大于 D 的 C 不确定度。

命题 2 对于一致性决策表 $DT=(U,C\cup D,V,f)$，若 B 是 C 的近似约简，则 B 也是 C 的代数约简。反之亦然。

证明：要证明 B 是 C 的代数约简，只需证明(1) $POS_B(D)=POS_C(D)$ 和 (2)($\forall B'\subset B, POS_{B'}(D)<POS_C(D)$)。

因为 DT 是一致决策表，由命题 1 可知 $UNC_C(D)=0$。若 B 是 C 的近似约简，则 $UNC_B(D)=UNC_C(D)$，即有 $UNC_B(D)=0$，当且仅当 $\forall X_i\in U/D$ 有 $\underline{B}(X_i)=\overline{B}(X_i)=X_i$ 成立，因此 $POS_B(D)=U=POS_C(D)$。又因为 $\forall B'\subset B$ 有 $UNC_{B'}(D)>UNC_C(D)$，即 $\forall B'\subset B$ 有 $UNC_{B'}(D)>0$，而 $UNC_{B'}(D)=0$ 当且仅当 $POS_{B'}(D)=U$，所以 $POS_{B'}(D)\subset U$。因此 B 是 C 的一个代数约简。

反过来，若 B 是 C 的一个代数约简，则 $POS_B(D)=POS_C(D)=U$，因此 $\forall X_i\in U/D$ 有 $\underline{B}(X_i)=\overline{B}(X_i)=\underline{C}(X_i)=\overline{C}(X_i)=X_i$，所以 $UNC_B(D)=UNC_C(D)=0$；另外 B 是 C 的代数约简，则 $\forall B'\subset B$ 有 $POS_{B'}(D)\subset POS_C(D)$，即 $POS_{B'}(D)\subset U$，因此 $\exists X_i\in U/D$ 使得 $\underline{B'}(X_i)\subset X_i$，即有 $\rho_{B'}(X_i)>\rho_C(X_i)$，所以 $UNC_{B'}(D)>UNC_C(D)$。因此 B 是 C 的一个近似约简。

性质 3 对于非一致决策表 $DT=(U,C\cup D,V,f)$，若 B 是 C 的近似约简，则 $\forall X_i\in U/D$ 有 $\underline{B}(X_i)=\underline{C}(X_i)$ 和 $\overline{B}(X_i)=\overline{C}(X_i)$ 成立。

证明：因为 B 是近似约简，则 $UNC_B(D)=UNC_C(D)$，根据定义 5 和性质 1，可以得到 $\forall X_i\in U/D$，有 $\rho_B(X_i)=\rho_C(X_i)$（否则 $UNC_B(D)>UNC_C(D)$）成立，因此 $\alpha_B(X_i)=\alpha_C(X_i)$ 成立。又因为 $\forall B\subseteq C$ 和 $\forall X_i\in U/D$，有 $\underline{B}(X_i)\subseteq\underline{C}(X_i)\subseteq X_i\subseteq\overline{C}(X_i)\subseteq\overline{B}(X_i)$ 成立，因此 $|\underline{B}(X_i)|\leqslant|\underline{C}(X_i)|\leqslant|X_i|\leqslant|\overline{C}(X_i)|\leqslant|\overline{B}(X_i)|$ 成立。再根据式(5)可知 $\alpha_B(X_i)=\alpha_C(X_i)$ 成立必有 $\underline{B}(X_i)=\underline{C}(X_i)$ 和 $\overline{B}(X_i)=\overline{C}(X_i)$ 成立，即证。

性质 3 给出了近似约简的重要特性。若 B 是 C 的近似约简，则 B 不仅维持了所有决策类的下近似不变，还维持了其上近似不变，这是与代数约简显著不同的特点。即决策属性 D 在近似约简 B 和全部条件属性集 C 上产生完全相同的 Rough 近似空间。对于一致决策表，这个结论也显然是成立的。

命题 3 给定一个决策表 $DT=(U,C\cup D,V,f)$ 和 $B\subseteq C$，若 B 是 C 的近似约简，则 B 必定包含 C 的一个代数约简。

证明：若 DT 是一致决策表，由命题 2 可知 B 也是 C 的一个代数约简。

若 DT 是非一致决策表，根据性质 3，$\forall X_i\in U/D$ 有 $\underline{B}(X_i)=\underline{C}(X_i)$ 成立，因此 $POS_B(D)=POS_C(D)$，所以 B 必定包含一个代数约简，即证。

可以证明，若 B 是非一致决策表上的一个近似约简，则 B 不一定是代数约简。只有当 $\forall B'\subset B\wedge\exists X_i\in U/D\rightarrow\alpha_B(X_i)=1\wedge\alpha_{B'}(X_i)\neq 1$ 时，B 才是一个代数约简。

定义 6 设 $B\subset C$ 且 $a\in C-B$，定义属性 a 相对于 B 的重要度为

$$SIG'(a,B,D)=UNC_B(D)-UNC_{B\cup\{a\}}(D) \tag{8}$$

即属性 a 相对于 B 的重要性体现在把属性 a 并入 B 之后与并入之前的不确定度的减小量。该值越大，则 a 相对于 B 的重要性越大，可以据此作为前向搜索属性选择的依据。

3 基于边界域的前向贪心约简算法

根据式(8)可以采用前向贪心搜索策略，通过测试加入新的候选属性后的不确定度的变化，选择使不确定度减少量最大的属性加入到近似约简变量。式(8)的计算涉及到根据式(7)计算属性子集的相对不确度。式(7)需要计算所有决策类相对条件属性子集的粗糙度，而每个决策类的粗糙度的计算涉及到计算该决策类的上下近似基数，因此它的计算复杂度直接关系到求近似约简的效率。下面提出一种高效率地求相对不确定度的算法，该算法和文献[7]中的求正区域的效率相当。

算法 1 求决策属性 D 相对条件属性子集 B 的不确定度

输入：$DT=(U,C\cup D,V,f)$ 和 $B\subseteq C$

输出：$UNC_B(D)$

步骤 1 根据 B 对对象集 U 进行排序；

步骤 2 对于每一个决策类 $X_j\in U/D$，为其定义一个上近似计数变量 $Upper_{X_j}$ 和下近似计数变量 $Lower_{X_j}$，并赋初值 $Upper_{X_j}=$

$0, Lower_{X_j}=0$;

步骤 3

3.1 令 $s=1, g=1, list=\varphi$，把对象 u_1 的决策值加入到集合 $list$；

3.2 for $i=2$ to $|U|$ do

if u_i 与 u_s 对于 B 中的每个属性都相等，则 g=g+1；

if u_i 的决策值 $\notin list$，则把 u_i 的决策值加入到集合 $list$；

end if

else //u_i 与 u_s 的条件属性值不相等

if $|list|=1$ 则更新 $list$ 集合所含每个决策值对应的决策类 X_j 的上下近似计数：

$Upper_{X_j}=Upper_{X_j}+g$；

$Lower_{X_j}=Lower_{X_j}+g$；

else 对 $list$ 集合所包含的决策值对应的决策类 X_j，更新其上近似计数：

$Upper_{X_j}=Upper_{X_j}+g$；

end if

s=i; g=1；

把集合 $list$ 清空；

把 u_i 的决策值加入到集合 $list$；

end if

步骤 4 根据步骤 3 得到的所有决策类的上下近似计数，直接计算 $UNC_B(D)$。

步骤 1 在快速排序下的时间复杂度为 $O(|B||U|\log_2|U|)$；步骤 2 的时间复杂度为 $O(|U/D|)$；步骤 3 的执行只需对排序后的对象集合扫描一遍，其时间复杂度为 $O(|B||U|)$；步骤 4 的时间复杂度为 $O(|U/D|)$；因此算法 1 的时间复杂度主要集中在步骤 1 根据 R 的排序上，在快速排序下整个算法的时间复杂度为 $O(|B||U|\log_2|U|)$。算法 1 的时间效率是相当高的，它计算出了所有决策类的上下近似基数，但其时间复杂度与属性排序下计算正域[7]的时间复杂度相同。依据算法 1 和定义 6 的属性重要度的概念，基于边界域的快约简算法具体描述如下。

算法 2 基于边界域的属性约简算 ARBR(Attribute Reduction based on Boundary Region)

输入：$DT=(U,C\cup D,V,f)$

输出：C 相对于 D 的一个近似约简 red

步骤 1 $red=\phi$；

步骤 2 计算 $UNC_C(D)$；

步骤 3

3.1 在 $C-red$ 中找出使 $SIG'(a,red,D)$ 最大的条件属 a；

3.2 把 a 并入 red 的尾部，$red=red\cup\{a\}$；

3.3 若 $UNC_{red}>UNC_C(D)$ Goto 步骤 3.1；

步骤 4 从后向前遍历 red 中的每一个属性 a，若 $UNC_{red-\{a\}}=UNC_{red}(D)$ 则 $red=red-\{a\}$；

步骤 5 输出 red。

步骤 2 的时间复杂度为 $O(|C||U|\log_2|U|)$；步骤 3 最坏情况下需要计算 $|C|(|C|+1)/2$ 次属性重要度，采用渐近式计算 $SIG(a,red,D)$ 的时间复杂度为 $O(|D||U|\log_2|U|)$，故步骤 3 总的时间复杂度为 $O(|C|^2|D||U|\log_2|U|)$；步骤 4 最坏情况下的时间复杂度为 $O(|C|^2|D||U|\log_2|U|)$。因此 ARBR 算法的时间复杂度为 $O(|C|^2|D||U|\log_2|U|)$。对于只含一个决策属性的决策表，利用 ARBR 算法求一个近似约简的时间复杂度为 $O(|C|^2|U|\log_2|U|)$。

4 与其它不一致决策表的知识约简比较

文献[19]给出了分布约简和分配约简两种定义，但未给出有效的约简方法。在文献[19]的基础上，文献[20]给出了最大分布约简的定义，并给出了可辩识矩阵的约简方法，其过程较为繁琐。下面给出相关定义。

定义 7 给定决策表 $DT=(U,C\cup D,V,f)$，设 $U/D=(X_1,X_2,\cdots,X_s)$，$P(D_j/[x]_B)=\frac{|D_j\cap[x]_B|}{[x]_B}$，$u_B(x)=(P(D_1/[x]_B),P(D_2/[x]_B),\cdots,P(D_s/[x]_B))$，$\delta_B(x)=\{D_j|[x]_B\cap D_j=\varphi\}$，$\gamma_B(x)=\max\limits_{j\leqslant s}P(D_j/[x]_B)$，$\eta_B=\frac{1}{|U|}\sum\limits_{X\in U/\mathrm{Ind}(D)}|\underline{B}(X)|$。$\forall x\in U, B\subseteq C$，

(1)若 $u_B(x)=u_C(x)$ 且 $\forall B'\subset B, u_{B'}(x)\neq u_C(x)$，则称 B 为分布约简；

(2)若 $\delta_B(x)=\delta_C(x)$ 且 $\forall B'\subset B, \delta_{B'}(x)\neq\delta_C(x)$，则称 B 为分配约简；

(3)若 $\gamma_B(x)=\gamma_C(x)$ 且 $\forall B'\subset B, \gamma_{B'}(x)\neq\gamma_B(x)$，则称 B 为最大分布约简；

(4)若 $\eta_B=\eta_C$ 且 $\forall B'\subset B, \eta_B\neq\eta_C$，则称 B 为近似约简。

由上面定义可知，分配约简保证每个对象可能的决策类不变，近似约简保证所有决策类的上近似不变。文献[20]证明了分配约简和近似约简是等价的。其实这里的近似约简还蕴含了所有决策类的下近似不变，因为如果近似约简导致某个决策类的下近似缩小(某些下近似中的对象成为边界域的对象)，则必然引起某个(某些)其它相关决策类的上近似扩大。因此该近似约简和本文中定义的近似约简也是等价的。

几种约简对非一致决策表的正区域、边界域以及边界域对象隶属度的影响如表 1 所列。

表 1 几种约简的比较

	最大分布约简	代数约简	近似约简(分配约简)	分布约简
正域	可能缩小	不变	不变	不变
边界域	可能变大	可能变大	不变	不变
边界域对象隶属度	可能改变	可能改变	可能改变	不变

和经典的基于正区域的前向贪心代数约简算法相比，ARBR 算法有着相同的算法时间复杂度。对于一致决策表，两者都能找出一个经典的代数约简。对于非一致决策表，代数约简可能造成原有决策类的边界域扩大，但 ARBR 算法能够同时保持所有决策类的上下近似不发生变化，从而保持数据原有的 Pawlak 拓扑结构。

分布约简需要保持对象在每个决策类的隶属程度不变，其条件相对苛刻且得到的约简基数相对较大。最大分布约简保持每个对象的最大分布决策类不变，其条件又过于宽松，可能导致原正区域缩小。而近似的约简介于两者之间，其维护了决策类族原有的 Rough 近似空间不变，尽管决策类边界区域对象的隶属程度可能改变，但其边界区域的大小保持不变，因此决策类的近似精度保持不变。另外，分布约简和最大分布约简的求解过程较为繁琐，而 ARBR 算法简单清晰，效率较高。

5 实验结果及分析

为验证 ARBR 约简算法的有效性，采用了 8 个 UCI(http://www.ics.uc.i edu/~mlearn/MLRepository.html)数据集进行实验，各数据集的描述见表 2。实验中，我们对经典的正域属性重要度的约简算法(POS_RED)和 ARBR 算法的约

简结果进行比较。由于两者都需要在离散化的决策表上进行约简，因此我们采用 CAIM[21] 离散化算法对表 2 中的数据集进行离散。

表 2　数据描述

Datasets	Abbreviation	Samples	Condition attributes	classes
Glass Identification	Glass	214	9	3
Wine recognition	Wine	178	13	3
Sonar, Mines vs. Rocks	Sonar	208	60	2
Jhons Hopkins University Ionosphere	Ions	351	34	2
Statlog Heart	Heart	270	13	2
Wisconsin diagnostic breast cancer	Wdbc	569	30	2
Wisconsin prognostic breast cancer	Wpbc	198	32	2
Statlog project satellite image	Sat	6435	36	6

表 3 中的第 2 列显示各数据集离散后的不确定度，其中等于 0 说明离散后的数据集是一致决策表，大于 0 说明离散后的数据集是非一致决策表。表 3 的第 4－5 列显示出了两种算法得到的约简含有的条件属性个数，总体上 ARBR 得到的近似约简含有的条件属性个数要相对多一些。两种算法的执行时间非常接近，取决于约简的基数。

表 3　约简结果比较

DataSet	不确定度	Attributes		约简时间(s)	
		POS_RED	ARBR	POS_RED	ARBR
Wine	0	5	6	0.162	0.184
Sonar	0	8	9	2.112	2.229
WPBC	0	9	10	1.296	1.328
WDBC	0.01051	14	13	2.221	2.128
Sat	0.00674	30	30	82.812	82.892
Glass	0.063	7	8	0.118	0.118
Ions	0.11146	11	13	2.093	2.312
Heart	0.064726	11	12	0.259	0.275

表 4 和表 5 分别显示了 C4.5 和 KNN($K=3$)归纳学习算法以 10 折交叉验证的方式对两种约简的数据集和约简前的数据集进行分类的精度。小括号中的数字代表各种数据集的分类性能排名，Mean rank(排名均值)提供了 8 个数据集分类性能排名均值。可见，ARBR 约简算法诱导出的分类器的性能总体上优于由全部属性集得到的分类器的性能。对于不确定度等于或非常接近于 0 的数据集 Wine, Sonar, WPBC, WDBC 和 Sat 而言，ARBR 约简算法诱导出的分类精度略强于基于正域属性重要度算法诱导出的分类精度；对于不确定度相对较大的数据集 Glass, Ions 和 Heart，由 ARBR 算法诱导出的分类器性能明显优于基于正域属性重要度算法诱导的分类器性能。这也再一次证实了对于不一致决策表而言，近似约简比代数约简有着更好的分类优势。

表 4　C4.5 分类精度比较

DataSet	不确定度	C4.5		
		Un_reduced	POS_RED	ARBR
Wine	0	94.97±4.11 (2)	**95.52±3.53** (1)	94.93±5.61 (3)
Sonar	0	**81.76±7.96** (1)	78.83±0.36 (3)	79.83±9.15 (2)
WPBC	0	78.29±5.75 (2)	77.29±6.74 (3)	**85.26±8.59** (1)
WDBC	0.01051	95.08±2.16(2.5)	95.08±2.96 (2.5)	**95.78±1.88** (1)
Sat	0.00674	86.51±0.75(2.5)	86.51±1.20 (2.5)	**86.76±1.39** (1)
Glass	0.063	**73.86±8.97** (1)	71.02±6.84 (3)	72.92±5.06 (2)
Ions	0.11146	89.98±4.45 (3)	91.14±4.11 (2)	**93.46±2.97** (1)
Heart	0.064726	78.52±0.15(3)	80.37±7.42(2)	**80.74±6.49**(1)
Mean rank		2.125	2.375	1.5

表 5　KNN(K=3)精度比较

DataSet	不确定度	C4.5		
		Un_reduced	POS_RED	ARBR
Wine	0	94.97±4.11 (2)	**95.52±3.53** (1)	94.93±5.61 (3)
Sonar	0	**81.76±7.96** (1)	78.83±0.36 (3)	79.83±9.15 (2)
WPBC	0	78.29±5.75 (2)	77.29±6.74 (3)	**85.26±8.59** (1)
WDBC	0.01051	95.08±2.16(2.5)	95.08±2.96 (2.5)	**95.78±1.88** (1)
Sat	0.00674	86.51±0.75(2.5)	86.51±1.20 (2.5)	**86.76±1.39** (1)
Glass	0.063	**73.86±8.97** (1)	71.02±6.84 (3)	72.92±5.06 (2)
Ions	0.11146	89.98±4.45 (3)	91.14±4.11 (2)	**93.46±2.97** (1)
Heart	0.064726	78.52±0.15(3)	80.37±7.42(2)	**80.74±6.49**(1)
Mean rank		2.125	2.375	1.5

结束语　借助粗糙集的边界域引入了决策表的不确定度概念，提出了近似约简的定义。从理论上分析和证明了对于一致决策表，近似约简是代数约简；对于非一致决策表，近似约简能够同时保持所有决策类原有的上下近似不变，从而保持了决策表原有的 Pawlak 拓扑结构。提出了高效率的计算相对不确定度的算法和求近似约简的算法。理论分析和实验结果表明，近似约简比代数约简有着更强的鲁棒性，这突出地表现在对于不一致决策表，近似约简比代数约简有着更好的分类优势。

参 考 文 献

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

[2] Pawlak Z. Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research, 1994, 72(3): 443-459

[3] 刘清. Rough 集及 Rough 推理[M]. 北京：科学出版社，2001

[4] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

[5] Swiniarski R W, Skowron A. Rough set methods in feature selection and recognition[J]. Pattern Recognition Letters, 2003, 24: 833-849

[6] Jelonek J, Krawiec K, Slowinski R. Rough Set reduction of attributes and their domains for neural networks[J]. Computational Intelligence, 1995, 11(2): 339-347

[7] 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

[8] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligences, 1998, 105 (1/2): 77-103

[9] Skowron A, Rauszer C. The discernibility matrices and functions in information system[M] // Slowinski R. Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publishers, 1992: 331-362

[10] Wang Jue, Wang Ju. Reduction algorithm based on discernibility matrix the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489-504

[11] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报，2007, 30(5): 815-822

[12] 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展，1999, 36(6): 681-684

[13] 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报，2002, 25(7): 759-766

(下转第 247 页)

表3 测试用的相关参数

参数符号	含义	设置大小
N	项的数目	300
M	项的属性个数	2
|T|	事务的个数	20K-100KB
V	项集的平均长度	20
F	最大频繁项集的平均长度	15
ε	最小支持度	0.001～0.01
δ	选择约束率	0～100%

实验结果如图1－图3所示。其中图1是测试没有满足约束的频繁项集所占比例对运行时间产生的影响，图2是测试事务大小对可扩展性的影响，图3是不同的支持度对运行时间的影响。

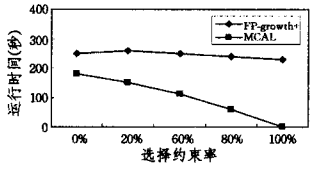

图1 不同选择约束率和运行时间对应图

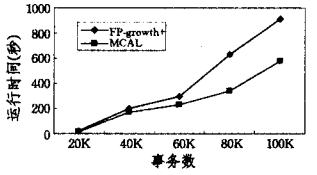

图2 不同事务数和运行时间的对应图

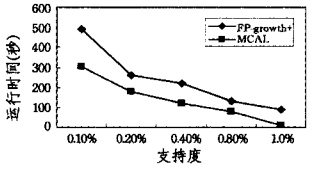

图3 不同支持度和运行时间的对应图

从实验结果可以看出，MCAL算法是有效的，且扩展性能较好，它充分利用非单调性约束和单调性约束的性质并与有效的剪枝技术相结合，大大减少了不必要的扫描，使其在运行时间和可扩展性方面均优于FP-growth+算法。

结束语 本文给出一个多约束关联挖掘算法。通过实验对比，无论从运行时间还是可扩展性来说，MCAL算法均获得了较好的性能。约束关联挖掘是一种重要的关联挖掘，以有效的剪枝策略，结合约束的性质来挖掘是主要的方法。基于动态、多关系、分布式数据的多维、多约束挖掘是今后研究的一个主要方向。

参考文献

[1] Ng R T, Lakshmannan L V S, Han Jia-wei. Exploratory mining and pruning optimizations of constrained associations rules[C]// Proceedings ACM SIGMOD International Conference on Management of Data. Seattle, Washington, USA, June 1998

[2] Srikant R, Vu Q, Agrawal R. Mining association rules with item constraints[C]// Proceedings of ACM International Conference on Knowledge Discovery and Data Mining. 1997: 67-73

[3] Lakshmanan L, Ng R, Han J, et al. Optimization of constrained frequent set queries with 2-variable constraints[C]// ACM SIGMOD Conference on Management of Data. 1999: 157-168

[4] Pie J, Han J. Can we push more constraints into frequent pattern mining? [C]// ACM SIGKDD Conference. 2000: 350-354

[5] Pie J, Han J, Lakshmanan L. Mining frequent itemsets with convertible constraints[C]// IEEE ICDE Conference. 2001: 433-442

[6] Bucila C, Gehrke J, Kifer D, et al. Dualminer: Adual-pruning algorithm for itemsets with constraints[C]// EightACM SIGKDD Internationa Conf. on Knowledge Discovery and Data Mining. Edmonton, Alberta, August 2002: 42-51

[7] Ting R M, Bailey J, Ramamohanarao K. Paradualminer: An efficient parallel implementation of the dualminer algorithm[C]// Eight Pacific-Asia Conference (PAKDD 2004). Sydney, Australia, May 2004: 96-105

[8] Bonchi F, Giannotti F, Mazzanti A, et al. Examiner: Optimized level-wise frequent pattern mining with monotone constraints [C]// IEEE ICDM. Melbourne, Florida, November 2004

[9] Bonchi F, Lucchese C. On closed constrained frequent pattern mining[C]// IEEE International Conference on Data Mining Brighton. UK, November 2004

[10] Bonchi F, Lucchese C, Trasarti R. Pushing tougher constraints in frequent pattern mining[C]// 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Hanoi, Vietnam, May 2005

[11] Laks V, Lakshmanan S, Ng R T. Efficient dynamic mining of constrained frequent sets[J]. ACM Transactions on Database Systems, 2003, 28(4)

[12] Anthony J, Lee T, Lin Wan-chuen, et al. Mining association rules with multi-dimensional constraints[J]. The Journal of Systems and Software, 2006(79): 79-92

[13] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]// Proceedings of International Conference on Very Large Data Bases. 1994: 487-489

(上接第227页)

[14] Wang Guo-yin, et al. Theoretical study on attribute reduction of rough set theory: comparison of algebra and information views [C]// Proceedings of the Third IEEE International Conference on Cognitive Informatics. Canada: IEEE Computer Society, 2004: 148-155

[15] Fleuret F. Fast Binary Feature Selection with Conditional Mutual Information[J]. Journal of Machine Learning Research, 2004, 5: 1531-1555

[16] 胡清华，于达仁，谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报，2008，19(3)：640-649

[17] 杨明. 一种基于一致性准则的属性约简算法[J]. 计算机学报，2010，33(2)：231-239

[18] Hu Qing-hua, Yu D, Xie Zong-xia, et al. EROS: Ensemble rough subspaces[J]. Pattern Recognition, 2007: 3728-3739

[19] Kryzkiewicz M. Comparative study of alternative types of knowledge reduction in inconsistent systems [J]. International Journal of Intelligent Systems, 2001, 16: 105-120

[20] 张文修，米据生，吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报，2003，26(1)：12-18

[21] Kurgan L A, Cios K J. CAIM Discretization Algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153

[22] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. Dordrecht: Kluwer Academic Publishers, 1991