变精度粗糙集的属性核和最小属性约简算法

陈 昊[1),2)] 杨俊安[1),2),3)] 庄镇泉[3)]

[1)](解放军电子工程学院 合肥 230037)

[2)](安徽省电子制约技术重点实验室 合肥 230037)

[3)](中国科学技术大学电子科学与技术系 合肥 230026)

摘 要 文中深入研究了变精度粗糙集的属性约简问题,给出了3种属性约简的概念,针对不同概念的属性约简,分别提出了两种不同的求解变精度粗糙集最小属性约简算法:基于容差矩阵和属性核的最小约简.提出了变精度粗糙集的属性核思想,对其进行了形式化描述,说明了变精度粗糙集的属性核真正具备了核的本质特征,从而更深层地提出了基于属性核的启发式约简以求解最小约简.理论分析和实例表明,所提出的两种最小约简算法可以减小属性约简的搜索空间,提高约简的效率,使得变精度粗糙集的属性约简具有了实用性.

关键词 变精度粗糙集;属性约简;属性核;容差矩阵;最小约简

中图法分类号 TP18 **DOI号**: 10.3724/SP.J.1016.2012.01011

The Core of Attributes and Minimal Attributes Reduction in Variable Precision Rough Set

CHEN Hao[1),2)] YANG Jun-An[1),2),3)] ZHUANG Zhen-Quan[3)]

[1)](*Electronic Engineering Institute, Hefei* 230037)

[2)](*Key Laboratory of Anhui Electronic Restriction, Hefei* 230037)

[3)](*Department of Electronic Science and Technology, University of Science and Technology of China, Hefei* 230026)

Abstract The attributes reduction in Variable Precision Rough Set (VPRS) is researched by this paper thoroughly. We define different attributes reduction and propose two methods of calculating minimal reduction based on tolerance matrix and core of attributes. The core attributes concept is presented. We discuss some properties of core attributes, which means that attributes core has the essential character about feature of core and makes the attributes reduction in VPRS practical. The theoretical analysis and example demonstrate two methods of calculating minimal reduction proposed in this paper can reduce space of attributes reduction to improve the efficiency of calculating it.

Keywords variable precision rough set; attributes reduction; attributes core; tolerance matrix; minimal reduction

1 引 言

由Pawalk等人[1]提出的粗糙集理论,作为一种新的处理模糊和不确定性知识的数学工具,在决策分析、模式识别及数据挖掘等领域取得了很大的成功.对经典粗糙集理论的扩充目前主要有基于容差关系、相似关系和限制容差关系等方法[2].然而利

收稿日期:2008-04-08;最终修改稿收到日期:2010-01-14.本课题得到国家自然科学基金(60872113)资助.**陈 昊**,男,1982年生,博士研究生,研究方向为粗糙集、智能计算.E-mail: ritian99@21cn.com.**杨俊安**,男,1965年生,教授,博士生导师,研究领域为智能计算、遗传算法、机器学习.**庄镇泉**,男,1938年生,教授,博士生导师,中国计算机学会(CCF)高级会员,研究领域为粗糙集,智能计算.

用这些模型获得的知识大多都只是一个层次上的，难以从多个级别上对原有系统进行分析处理.为此，文献[3-11]提出了一些变精度粗糙集(Variable Precision Rough Set,VPRS)模型，研究了 VPRS 模型中的参数关系[5-6,8,10]，并提出了基于 VPRS 的知识获取方法[3-4,7,9,11].

属性约简是 VPRS 理论最重要的研究内容之一，它是得到精简且完备的决策规则集合的前提.求解 VPRS 的最小属性约简是 NP 问题，主要原因是属性的组合爆炸.所以，减小属性约简的搜索空间，提高约简的效率，对于 VPRS 的属性约简具有重要的意义.

但是，由于正确分类率 β 的引入，VPRS 的属性约简非常复杂.文献[3-4,11]详细给出了 VPRS 三种约简的概念，分析了 VPRS 模型约简异常出现的原因，结合 VPRS 模型特征，将特定 β 值上的约简扩展为区间约简，并从分类质量、β 相对正域和决策类 3 个层次分别对约简进行了描述，研究了它们与约简异常之间的关系.但是，文献[3-4,11]并没有提出如何求解 VPRS 的最小属性约简问题.我们知道，一个决策信息系统可能存在多个属性约简集合，相对于决策属性集合和条件属性集合的所有约简的交集称为属性核.核中的属性是约简的极限粒度.用核作为计算约简集的起点，可以简化计算约简集，提高属性约简的效率.文献[12-14]对经典粗糙集理论的属性核进行了深入研究.由于 VPRS 约简的复杂性，至今还没有相关文献具体提出其核的计算方法.

本文深入研究了基于分类率不变、正域不变、下近似不变的 VPRS 属性约简，提出了 VPRS 属性核思想，指出基于分类率不变、正域不变的 VPRS 属性约简不存在属性核，而只有基于下近似不变的 VPRS 属性约简才有属性核的存在.并且针对基于分类率不变、正域不变的 VPRS 属性约简问题，提出了一种基于容差矩阵的最小属性约简算法；针对基于下近似不变的 VPRS 属性约简，在属性核的思想下，更深层地提出了基于属性核的启发式约简以求解其最小约简.通过理论分析和实例验证，从算法复杂度角度考虑，所提出的两种最小属性约简算法可以减小 VPRS 属性约简的搜索空间，提高约简的效率.

2 VPRS 模型的基本概念和理论

VPRS 是对标准粗糙集理论的一种推广，它通过设置参数 β，放松标准粗糙集对近似边界的要求.下面给出 VPRS 的有关概念.

定义 1[11]. 设 X 是有限集合，$F=\{Y\mid Y\subseteq X\}$，$\subseteq$ 为 F 上的偏序关系，对任意 $A,B\in F$，记

$$\beta=\begin{cases}\dfrac{|B\cap A|}{|A|}, & A\neq\varnothing\\ 1, & A=\varnothing\end{cases},$$

β 为 A 关于 B 的可信度阈值，即 B 包含 A 的程度，这里 $|A|$ 表示集合 A 的基数.

定义 2[11]. 给定论域 U，不可分辨关系 $R\subseteq U\times U$，$X\subseteq U$，$\beta\in(0.5\quad 1]$，则

$$R_\beta(X)=\bigcup\left\{[x]_R\mid\frac{|[x]_R\cap X|}{|[x]_R|}\geq\beta\right\},$$

$$R^\beta(X)=\bigcup\left\{[x]_R\mid\frac{|[x]_R\cap X|}{|[x]_R|}>1-\beta\right\}$$

分别称为 X 的 R 下 β 近似，X 的 R 上 β 近似.

β 正区域：$POS_\beta(X)=R_\beta(X)$，$\dfrac{|[x]_R\cap X|}{|[x]_R|}\geq\beta$.

β 负区域：$NEG_\beta(X)=U-R^\beta(X)$，$\dfrac{|[x]_R\cap X|}{|[x]_R|}\leq 1-\beta$.

β 边界域：$BND_\beta(X)=R^\beta(X)-R_\beta(X)$，$1-\beta<\dfrac{|[x]_R\cap X|}{|[x]_R|}<\beta$.

显然，在 VPRS 模型下，近似区域与 β 取值有着十分紧密的关系，将随着 β 的调整而变化.

定义 3[11]. 给定决策信息系统 $S=(U,Q=C\cup D,V,F)$，U 为论域，C 为条件属性集，D 为决策属性集.由条件属性和决策属性定义的不可分辨关系对 U 产生不同的分类.

(1)根据条件属性对 U 的分类称为条件分类，为 $U/C=\{X_1,X_2,\cdots,X_{|U/C|}\}$.其中每个成员为 X 的一个条件类.根据决策属性对 U 的分类称为决策分类，为 $U/D=\{Y_1,Y_2,\cdots,Y_{|U/D|}\}$.其中每个成员为 Y 的一个决策类.

(2)给定条件类 $X\in U/C$，令 $H_X=\max\limits_{j=1}^{|U/D|}\dfrac{|X\cap Y_j|}{|X|}$.则 H_X 为条件类 X 相对所有决策类的最大被包含度，称为条件类 X 的包含度阈值.

定义 4[11]. 给定决策信息系统 $S=(U,Q=C\cup D,V,F)$，U 为论域，C 为条件属性集，D 为决策属性集.给定 $\beta\in(0.5\quad 1]$，决策属性集 D 与条件属性集 C 的 β 近似依赖或基于 β 的分类率为

$$\gamma(C,D,\beta)=\frac{|POS(C,D,\beta)|}{|U|}.$$

其中 $POS(C,D,\beta)=\bigcup_{Y_j\in U/D} C_\beta Y_j$ 为 β 的相对正域，此外 $C_\beta Y_j$ 表示了决策类 Y_j 相对于条件属性集 C 的 β 下近似. 近似分类质量度量了论域中给定某一 β 值时，可能正确的分类知识在现有知识中的百分比.

记 $DP(C,D,\beta)=\{C_\beta D_1, C_\beta D_2, \cdots, C_\beta D_{|U/D|}\}$，DP 为所有决策类 β 下近似构成的集合，是各决策类关于 U/C 的概率分布，称 DP 为决策类下近似分布.

定理 1.　当 $0.5<\beta_1\leq\beta_2\leq 1$ 时，有 $\gamma^{\beta_2}(C,D)\leq\gamma^{\beta_1}(C,D)$.

证明. 因为 $0.5<\beta_1\leq\beta_2\leq 1$，所以 $POS_C^{\beta_2}(Y)\leq POS_C^{\beta_1}(Y)$，从而有

$$\gamma^{\beta_2}(C,D)=\frac{|POS(C,D,\beta_2)|}{|U|}=\frac{\bigcup_{Y\in U/D} POS_C^{\beta_2}(Y)}{|U|}\leq \frac{\bigcup_{Y\in U/D} POS_C^{\beta_1}(Y)}{|U|}=\frac{POS(C,D,\beta_1)}{|U|}=\gamma^{\beta_1}(C,D).$$

证毕.

由定理 1 易得到定理 2.

定理 2.　对于给定的 γ 值，满足 γ 要求的 β 最大值为可信度上限，记为 β^γ，$\beta\in(0.5\quad\beta^\gamma]$.

根据定理 1 和定理 2，γ 与 β 间的关系计算如下.

输入：决策信息系统 $S=(U,Q=C\cup D,V,F)$，U 为论域，C 为条件属性集，D 为决策属性集

输出：γ 值与相应 β 间的区间关系

1. $\gamma/\beta=\varnothing$，$n=|U/C|$；

2. 各条件类的元素为 $|X_1|,|X_2|,\cdots,|X_n|$；

3. 求各条件类包含度阈值，设从小到大的排序为 $X_1, X_2,\cdots,X_n$，对应的包含度阈值从小到大为 $H_{X_1}, H_{X_2},\cdots, H_{X_n}$；

4. 取 $\beta=\min\{H_{X_i}\mid H_{X_i}>0.5\}$，则 $(0.5\quad\beta]$ 区间的分类 $\gamma=\frac{|X_i|+|X_{i+1}|+\cdots+|X_n|}{|U|}$，令 $\gamma/\beta=\gamma/\beta+\langle\gamma,(0.5\quad\beta]\rangle$；

5. while $i<n$
   { $i=i+1$
     if $H_{X_i}>\beta$
     then
       $\gamma=\frac{|X_i|+|X_{i+1}|+\cdots+|X_n|}{|U|}$；
       $\gamma/\beta=\gamma/\beta\cup\langle\gamma,(\beta\quad H_{X_i}]\rangle$，$\beta=H_{X_i}$；
     end if
   }
   end while

6. if $\beta<1$
   then $\gamma/\beta=\gamma/\beta\cup\langle 0,(\beta\ 1]\rangle$；
   end if

由上面算法可以得到 γ 值与相应 β 间的区间关系如图 1 所示.

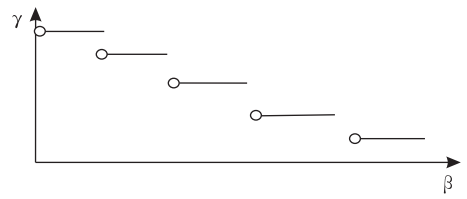

图 1　γ 与 β 间的关系

所以，不同的正确分类率对应不同的分类质量；两个正确分类率之间的分类质量是保持不变的. 根据以上关系就可以确定信息系统的条件属性相对与决策属性所有不同的分类质量以及每个分类质量所对应的参数 β 的范围. 在对集合进行分类时，β 取值的区间性，从而将对特定 β 值的包含关系扩展为 β 区间域.

3　近似约简与 VPRS 的属性核

在考虑正确分类率 $\beta(0.5<\beta\leq 1)$ 存在的情况下，依据近似分类质量的标准对属性进行约简.

条件属性 C 关于决策属性 D 的近似约简应满足：

$$\gamma^\beta(C,D)=\gamma^\beta(red(C,D),D)\qquad(1)$$

从 $red(C,D)$ 中去掉任何一个属性，都会使式(1)不成立. $red(C,D)$ 是指条件属性 C 关于决策属性 D 的一个近似约简.

在相同分类率 γ 下，满足约简条件的 β 值通常为一个区间范围. 约简后的决策信息系统与原决策信息系统具有相同的分类率，β 取值范围出现了波动差异，不能提供与属性子集 C 完全一致的信息，亦即约简区间变化产生异常，约简前后的决策信息系统在这种情况下，γ 值相同而 β 不同. 出现这种情况的原因在于约去某个属性后，产生了条件类的合并，合并后新的条件类将产生新的包含度阈值，从而引起 β 区间的扩张或收缩. 因此必须把约简的 β 值从点扩展到区间.

约简前后分类率虽然保持不变，即正区域的大小不变，但正区域中的元素发生了变化，即约简改变了原决策系统的分类，出现了正区域元素的变化. 要保持分类质量不变，正区域中的元素前后一致，必须考虑对 β 约简过程的动态描述.

定义 5[11].　决策信息系统 $S=(U,Q=C\cup D,V,F)$，U 为论域. C 为条件属性集，D 为决策属性集. 条件属性 C 关于决策属性 D 的 β 约简定义为 C 的一个最小属性子集 $RED(C,D,\beta)$，$\beta\in(0.5\quad 1]$，

且满足：

(1) $\gamma(C,D,\beta)=\gamma(RED(C,D,\beta),D,\beta)$；

(2) 从 $RED(C,D,\beta)$ 中去掉任何一个属性，条件(1)不成立.

条件(1)体现了约简中保持分类率 $\gamma(C,D,\beta)$ 不变；条件(2)体现了约简的最小性.

定义 6[11]. 决策信息系统 $S=(U,Q=C\cup D,V,F)$，U 为论域. C 为条件属性集，D 为决策属性集. 在既定分类率 γ 条件下，条件属性 C 关于决策属性 D 的 β 定义为 C 的一个最小子集 $RED((C,D,\beta),D,\beta)$，

(1) $POS(C,D,\beta)=POS(RED(C,D,\beta),D,\beta)$；

(2) 从 $RED(C,D,\beta)$ 中去掉任何一个属性，条件(1)不成立.

条件(1)体现了约简正区域 $POS(C,D,\beta)$ 中元素必须具有前后一致性，同时保持分类率不变；条件(2)体现了约简的最小性.

在 VPRS 近似约简中，核的描述以及性质是个值得探讨的问题.

根据 Ziarko 约简定义，VPRS 模型下属性具有不稳定性，使得约简过程产生“是约简→不是约简→是约简”的跳跃过程，即：某此约去属性后是决策表的约简，再约去一个属性后不是原决策表的约简，但继续约简后又变成原决策表的约简. 以决策表 1 和决策表 3 为例.

给定决策信息系统 $S=(U,Q=C\cup D,V,F)$，如表 1 所示，其中 $C=\{a_1,a_2,a_3,a_4,a_5\}$ 为条件属性集，$D=\{d\}$ 为决策属性集.

表 1 决策表 1

U	a_1	a_2	a_3	a_4	a_5	d
o_1	1	1	1	1	1	Y
o_2	1	1	0	1	1	Y
o_3	0	0	1	0	0	Y
o_4	1	1	2	1	1	Y
o_5	1	1	0	1	0	N
o_6	1	1	0	1	1	N
o_7	0	0	1	2	1	N
o_8	1	1	0	1	1	N
o_9	1	1	2	1	1	N

根据定义 5，分别约去属性 a_1、a_2、a_3、a_4、a_5 后分类率的变化，如表 2.

表 2 决策表 1 的约简

条件属性	等价类	包含度及分类质量	β 相对正域
C	$0.5/X_1=\{o_4,o_9\}$ $0.667/X_2=\{o_2,o_6,o_8\}$ $1.0/X_3=\{o_1\}$，$1.0/X_4=\{o_3\}$ $1.0/X_5=\{o_5\}$，$1.0/X_6=\{o_7\}$	$\beta\in(0.5\ 0.667]$ $\gamma=7/9$	$\{o_1,o_2,o_3,o_5,o_6,o_7,o_8\}$
$C-\{a_1\}$	与 C 相同	与 C 相同	与 C 相同
$C-\{a_2\}$	与 C 相同	与 C 相同	与 C 相同
$C-\{a_3\}$	$0.5/X_1=\{o_1,o_2,o_4,o_6,o_8,o_9\}$ $1.0/X_2=\{o_3\}$，$1.0/X_3=\{o_5\}$ $1.0/X_4=\{o_7\}$	$\beta\in(0.5\ 1]$ $\gamma=3/9$	$\{o_3,o_5,o_7\}$
$C-\{a_4\}$	与 C 相同	与 C 相同	与 C 相同
$C-\{a_5\}$	$0.5/X_1=\{o_4,o_9\}$ $0.75/X_2=\{o_2,o_5,o_6,o_8\}$ $1.0/X_3=\{o_1\}$，$1.0/X_4=\{o_3\}$ $1.0/X_5=\{o_7\}$	$\beta\in(0.5\ 0.75]$ $\gamma=7/9$	$\{o_1,o_2,o_3,o_5,o_6,o_7,o_8\}$
$C-\{a_5,a_4,a_3\}$	$0.5/X_1=\{o_3,o_7\}$ $0.57/X_2=\{o_1,o_2,o_4,o_5,o_6,o_8,o_9\}$	$\beta\in(0.5\ 0.57]$ $\gamma=7/9$	$\{o_1,o_2,o_4,o_5,o_6,o_8,o_9\}$

从表 2 可以看出，从分类质量考虑，属性 a_3 是属性核. 但是在属性逐个约简过程中，其分类质量、包含度、β 相对正域变化如表 2 所示，同时约去属性 a_3,a_4,a_5 后，分类质量与原决策系统相同，故属性 a_3 又可以约简. 所以根据定义 5，VPRS 属性约简不存在属性核.

给定决策信息系统 $S=(U,Q=C\cup D,V,F)$，如表 3 所示，其中 $C=\{a_1,a_2,a_3,a_4,a_5\}$ 为条件属性集，$D=\{d\}$ 为决策属性集.

表 3 决策表 2

U	a_1	a_2	a_3	a_4	a_5	d
o_1	1	1	1	2	1	Y
o_2	1	1	1	0	1	Y
o_3	1	1	0	1	0	Y
o_4	0	0	1	2	1	Y
o_5	1	1	1	0	0	N
o_6	1	1	1	0	1	N
o_7	0	0	2	1	1	N
o_8	1	1	1	0	1	N
o_9	0	0	1	2	1	N

根据定义 6，分别约去属性 a_1、a_2、a_3、a_4、a_5 后 β 相对正域的变化，如表 4. 根据属性核的定义，a_4 是属性核. 同时约去属性 a_3，a_4，a_5 后，β 相对正域又与原决策系统相同，故属性 a_4 又可以约简. 所以根据定义 6，VPRS 属性约简不存在属性核.

表 4　决策表 2 的约简

条件属性	等价类	包含度及分类质量	β 相对正域
C	$0.5/X_1=\{o_4,o_9\}$ $0.667/X_2=\{o_2,o_6,o_8\}$ $1.0/X_3=\{o_1\}$，$1.0/X_4=\{o_3\}$ $1.0/X_5=\{o_5\}$，$1.0/X_6=\{o_7\}$	$\beta\in(0.5\ 0.667]$ $\gamma=7/9$	$\{o_1,o_2,o_3,o_5,o_6,o_7,o_8\}$
$C-\{a_1\}$	与 C 相同	与 C 相同	与 C 相同
$C-\{a_2\}$	与 C 相同	与 C 相同	与 C 相同
$C-\{a_3\}$	与 C 相同	与 C 相同	与 C 相同
$C-\{a_4\}$	$0.5/X_1=\{o_1,o_2,o_6,o_8\}$ $1.0/X_2=\{o_3\}$，$0.5/X_1=\{o_4,o_9\}$ $1.0/X_4=\{o_5\}$，$1.0/X_5=\{o_7\}$	$\beta\in(0.5\ 1]$ $\gamma=3/9$	$\{o_3,o_5,o_7\}$
$C-\{a_5\}$	$0.5/X_1=\{o_4,o_9\}$ $0.75/X_2=\{o_2,o_5,o_6,o_8\}$ $1.0/X_3=\{o_1\}$，$1.0/X_4=\{o_3\}$ $1.0/X_5=\{o_7\}$	$\beta\in(0.5\ 0.75]$ $\gamma=7/9$	$\{o_1,o_2,o_3,o_5,o_6,o_7,o_8\}$
$C-\{a_5,a_4,a_3\}$	$0.5/X_1=\{o_4,o_9\}$ $0.57/X_2=\{o_1,o_2,o_3,o_5,o_5,o_6,o_7\}$	$\beta\in(0.5\ 0.57]$ $\gamma=7/9$	$\{o_1,o_2,o_3,o_5,o_6,o_7,o_8\}$

由上面的论述可知，根据定义 5、6 的属性约简其属性核是不存在的. 但是它们均存在着最小属性约简. 本文将信息系统的所有属性约简以及每个约简的参数 β 范围结合相考虑，提出了一种基于容差矩阵的最小属性约简算法. 该算法从中选出所含属性个数最少而且参数 β 范围最大的约简，作为最小约简.

算法思想：首先，将决策表中条件属性的所有组合形式通过二进制编码表示出来；其次，将各种组合中的属性集相对于决策属性的分类质量与条件属性相对于决策属性的分类质量相比较，若相等则将该属性集作为矩阵的一行；再次，根据准则删除那些子集已经是约简的行以及全为零的行，最后得到信息系统的所有约简组成的矩阵，具体算法如下.

算法 1.　基于容差矩阵的最小属性约简算法.

输入：决策信息系统 $S=(U,Q=C\cup D,V,F)$，U 为论域，C 为条件属性集，D 为决策属性集. $C=\{C_1,C_2,\cdots,C_m\}$，正确分类率 β，分类率 γ

输出：最小约简 y

1. 令 $w=2^m-1$，$\boldsymbol{y}$ 为 $w\times m$ 的零矩阵，并且令 $j=1$；

2. 对 j 进行二进制编码，计算每个二进制编码中 1 所对应的属性组合 p 的分类质量 $\gamma(p,D,\beta)$ 以及所对应的 β 范围；

3. 若 $\gamma(p,D,\beta)=\gamma(C,D,\beta)$ 成立，则用 p 来替换 $\boldsymbol{y}$ 的第 j 行，否则继续；

4. $j=j+1$，若 $j<w$，返回到步 2，否则继续；

5. 删除冗余的行，并且删除 $\boldsymbol{y}$ 中全为零的行，最后得到 VPRS 的所有 β 约简以及各个约简所对应的 β 范围.

整个算法的时间复杂度为 $O(m^2)$.

根据上面算法，可以求得表 1 和表 3 的最小约简为 $\{a_1,a_3\}$，$\{a_2,a_4\}$.

在 VPRS 理论中，分类质量 γ 与 β 相对正域随着属性数目的减少，均会出现跳跃现象，但是，下近似分布 DP 随着属性数目的减少不会出现跳跃现象.

定理 3.　给定决策信息系统 $S=(U,Q=C\cup D,V,F)$，U 为论域，C 为条件属性集，D 为决策属性集. 正确分类率 $\beta\in(0.5\ 1]$，条件分类为 $U/C=\{X_1,X_2,\cdots,X_{|U/C|}\}$，决策分类为 $U/D=\{Y_1,Y_2,\cdots,Y_{|U/D|}\}$. 若属性 $a\in C$，满足 $DP(C,D,\beta)\neq DP(C-\{a\},D,\beta)$，则 $\forall B\subset C-\{a\}$，均有 $DP(B,D,\beta)\neq DP(C,D,\beta)$.

证明.　设 $D_i\in U/D$，$C_\beta D_i\neq(C-\{a\})_\beta D_i$，这说明约去属性后条件等价类至少存在 C_i 和 C_j 合并. 其中 $C_i,C_j\in U/C$ 且 $C_i\subseteq C_\beta D_i$，$C_j\not\subset C_\beta D_i$，$\forall B\subset C-\{a\}$，均有 $U/(C-\{a\})$ 细分 U/B，则合并后的条件类 $C_i\cup C_j$ 在条件集 B 下都不会被细分. 若 $C_i\cup C_j\subseteq B_\beta D_i$，则 $B_\beta D_i\neq C_\beta D_i$；若 $C_i\cup C_j\not\subset B_\beta D_i$，则也有 $B_\beta D_i\neq C_\beta D_i$，从而说明 $\forall B\subset C-\{a\}$，$DP(B,D,\beta)\neq DP(C,D,\beta)$.　证毕.

由于以分类质量与 β 相对正域是否改变为定义的约简，非单调递减特征的打破，跳跃现象的出现，通过约简前后分类质量或 β 相对正域变化不能判定一个属性是否可约，约简过程具有不稳定性. 所以，它们不具有属性核. 但是，以下近似分布是否改变定义的属性约简，如果一个属性不可约，它在整个约简过程中都是不可约的，与约去属性之间的次序无关，不受其它属性的影响，约简具有稳定性，存在着属性

核.下面我们定义 VPRS 的稳定约简以及其属性核.

定义 7. 决策信息系统 $S=(U,Q=C\cup D,V,F)$,U 为论域.C 为条件属性集,D 为决策属性集.在既定正确分类率 β 和分类率 γ 条件下,C 关于 D 的下近似分布为 $DP(C,D,\beta)$,C 关于 D 的 β 的一个最小约简子集 $RED((C,D,\beta),D,\beta)$,C 关于 D 的 β 的所有最小约简集合为 R,C 关于 D 的 β 的属性核为 $CORE(C,D,\beta)$.根据定理 3,

(1) $DP(C,D,\beta)=DP(RED(C,D,\beta),D,\beta)$;

(2) 从 $RED(C,D,\beta)$ 中去掉任何一个属性,(1) 不成立;

(3) $CORE(C,D,\beta)=\bigcap R$.

基于核的启发式求解最小属性约简算法思想:对一个决策表中所有的条件属性集 C 的依赖度大小排序,依赖度最高的条件属性必然存在于核中;再依次将其它条件属性加入到依赖度最高的条件属性中,直到整个集合关于决策属性集 D 的下近似分布等于 $DP(C,D,\beta)$.有关 C 关于 D 的依赖度 $sig(C,D,\beta)$ 计算参见文献[13].基于核的启发式求解最小属性约简具体算法如下.

算法 2. 基于属性核最小属性约简算法.

输入:决策信息系统 $S=(U,Q=C\cup D,V,F)$,U 为论域,C 为条件属性集,D 为决策属性集.$C=\{C_1,C_2,\cdots,C_m\}$,正确分类率 β,分类率 γ,C 关于 D 的依赖度 $sig(C,D,\beta)$,C 关于 D 的下近似分布为 $DP(C,D,\beta)$

输出:最小属性约简 $RED(C,D,\beta)$

1. 以属性依赖度对 C 中的属性排序,假设顺序为 $sig(C_1,D,\beta)>sig(C_2,D,\beta)>\cdots>sig(C_m,D,\beta)$;

2. $RED(C,D,\beta)=C_1$;

3. for $i=1:m$

  if $DP(C,D,\beta)=DP(RED(C,D,\beta),D,\beta)$

  then $RED(C,D,\beta)=C_1\cup C_2\cup\cdots\cup C_i$

  end

 end

整个算法的时间复杂度为 $O(m^2/|2\times U|)$.

根据定义 7 和算法 3,可以求得表 1 的最小约简为 $\{a_2,a_3,a_4\}$,$\{a_1,a_2,a_5\}$,属性核为 $\{a_2\}$;表 3 的最小约简为 $\{a_1,a_4\}$,$\{a_1,a_5\}$,属性核为 $\{a_1\}$.

4 结 论

本文深入分析了 VPRS 属性约简问题,给出了正确分类率 β 与分类率 γ 的区间关系,研究了 β 值区间对分类率的影响,探讨了基于分类率 γ、β 相对正域和下近似分布 3 种概念的约简,分析了包含度区间的动态变化和正区域变化引起的约简异常,提出了基于容差矩阵的最小属性约简算法;针对基于下近似分布的属性约简,提出了 VPRS 属性核的概念以及基于核的启发式约简算法.理论分析和实例表明,本文所提出的针对 3 种不同概念的求解最小属性约简的算法时间复杂度低,有很强的实用性.

参 考 文 献

[1] Pawlak Z. Rough-Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic Publisher, 1991

[2] Pawlak Z. Rough sets: Some extension. Information Sciences, 2007, 177(1): 28-40

[3] Mi J S, Wu W Z, Zhang W X. Approaches to knowledge reduction based on variable precision rough set model. Information Sciences, 2004, 159(3-4): 255-272

[4] Inuiguchi M. Several approaches to attribute reduction in variable precision rough set model//Proceeding of the Modeling Decisions for Artificial Intelligence. Tsukuba, Japan, 2005: 215-226

[5] Su C T, Hsu J H. Precision in the variable precision rough sets model: An application. Omega, 2006(34): 149-157

[6] Su Chao-Ton, Hua Jigh-Hwa. Precision parameters in the variable precision rough sets model: an application. The International Journal of Management Science, 2006, 34(2): 149-157

[7] Cheng Yu-Sheng, Zhang You-Sheng, Hu Xue-Gang. The relationships between variable precision value and knowledge reduction based on variable precision rough set model//Proceedings of the RSKT2006. Chongqing, China, 2006

[8] Hong Tzung-Pei, Wang Tzu-Ting, Wang Shyue-Liang. Mining fuzzy β-certain and β-possible rules from quantitative data based on the variable precision rough set model. Expert Systems with Application, 2007(32): 223-233

[9] Gang Xie, Jin Long-Zhang, Lai K K, Yu Lean. Variable precision rough set for decision-making — An application. International Journal of Approximate Reasoning, 2008, 49: 331-343

[10] Ji Yang-Sheng, Shang Lin. Incremental computation of variable precision value in variable precision rough set. Computer Science, 2008, 35(3): 228-230(in Chinese)
(吉阳生,商琳. 可变精度粗糙集 β 值的增量计算. 计算机科学, 2008, 35(3): 228-230)

[11] Wang Jia-Yang, Zhou Jie. Research of Reduct features in the variable precision rough set model. Neruocomputing, 2009, 72(2): 2643-2648

[12] Ye Dong-Yi, Chen Zhao-Jiong. A new discernibility matrix and the computation of a core. Acta Electronic Sinica, 2002, 30(7): 1086-1088(in Chinese)
(叶东毅,陈昭炯. 一个新的差别矩阵及其求核方法. 电子学报, 2002, 30(7): 1086-1088)

[13] Wang Guo-Yin. Calculation methods for core attributes of decision table. Chinese Journal of Computers, 2003, 26(5): 611-615(in Chinese)

(王国胤. 决策表核属性的计算方法. 计算机学报，2003，26(5)：611-615)

[14] Ge Hao，Li Long-Shu，Yang Chuan-Jian. Quick algorithm for core attribute. Control and Decision，2009，24(5)：738-740(in Chinese)
(葛浩，李龙澍，杨传健. 一种核属性快速求解算法. 控制与决策，2009，24(5)：738-740)

CHEN Hao，born in 1982，Ph. D. candidate. His research interests include rough sets and intelligent computing etc.

YANG Jun-An，born in 1965，professor，Ph. D. supervisor. His research interests include intelligent computing，genetic algorithm and machine learning etc.

ZHANG Zhen-Quan，born in 1938，professor，Ph. D. supervisor. His research interests include rough sets and intelligent computing etc.

Background

This paper researched the attributes reduction in VPRS thoroughly. VPRS is extension of Rough Set，which is ineffective when processing database including noise. Attributes reduction in VPRS has been researched for long time. Until now，researchers have researched features of Reduct，reduction mergence，hierarchy of interval reduct in order to find weak dependence relationship，more general association and decision rules. They all focused on the attribution reduction anomalies from quality of classification，relative positive region and lower approximation，but didn't solve how to seek minimal attributes reduction. This paper defined different attributes reduction and proposed two methods of calculating minimal reduction based on tolerance matrix and core of attributes. The core attributes concept was presented. The theoretical analysis and example demonstrated two methods of calculating minimal reduction proposed in this paper can reduce space of attributes reduction to improve the efficiency of calculating it.

This paper was supported by project of Natural Science Foundation of China under Grant No. 60872113.