可变精度粗集模型约简异常分析*)

王加阳　陈松乔　罗　安
(中南大学信息科学与工程学院　长沙 410083)

摘　要　粗糙集理论一直致力于研究不确定或不精确信息的数据分析问题。本文基于可变精度粗糙集模型相关概念,对其约简异常进行了深入研究,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,提出了消除异常的基本思想,从而完善了基于可变精度粗集模型的约简。

关键词　可变精度,粗糙集模型,约简,异常

The Analysis of Reduct Anomaly on Variable Precision Rough Set Model

WANG Jia-Yang　CNEH Song-Qiao　LUO An
(College of Information Science and Engineering, Central South University, Changsha 410083)

Abstract　Rough set theory has being aimed at data analysis problems involving uncertain or imprecise information. Based on the relative concepts of variable precision rough set model, the paper makes a lucubration on reduct anomaly, analyzes the reduct anomaly when inclusion degree vibrates and positive area changes. It also presents some basic ideas to eliminate reduct anomaly and develops the variable precision rough set model.

Keywords　Variable precision, Rough set model, Reduct, Anomaly

1　引言

来源于现实的数据集合,大多存在着数据的不确定性或不精确性。经典粗糙集合理论采用精确集合概念定义下、上近似集合,描述数据之间的相关性。由于对数据的要求过于严格,导致存在一些不足之处,主要体现在:缺乏对噪声数据的适应能力,抗干扰能力差;分类只有严格的"包含"和"不包含",缺乏柔性或鲁棒性;对于边缘区域,不能区分等价类与集合的重叠度,没有体现程度上的差别等。

Ziarko 提出的可变精度粗集模型[1],允许上近似和下近似存在一定的分类误差,对经典粗糙集理论进行了扩展,主要分析了属性间统计意义上的数据模式或者存在概率上的不确定关系时的分类问题,而不是严格意义上的属性函数依赖关系,增强了粗糙集模型的数据分析能力。

2　可变精度粗集模型

定义 1[1]　设 X 和 Y 为论域 U 的非空子集,$0\leqslant\beta\leqslant 1$,定义 β 包含度关系:

$$Y\overset{\beta}{\supseteq}X$$

其中,$\beta=(|X|\cap|Y|)/|X|$,当 $|X|>0$;$\beta=0$,当 $|X|=0$。

当 $0.5<\beta\leqslant 1$,则定义了 Y 对 X 的 β 多数包含度关系,即 X 中有 50% 以上的元素被 Y 包含(或 X 与 Y 的公共元素占 X 的 50% 以上)。

定义 2[2,3]　给定论域 U,不可分辨关系 $R\subseteq U\times U$,$X\subseteq U$,$\beta\in(0.5,1]$,则:

$$\underline{apr_\beta}(X)=\bigcup\{[x]_R\mid\frac{|[X]_R\cap X|}{|[X]_R|}\geqslant\beta\}$$

$$\overline{apr_\beta}(X)=\bigcup\{[x]_R\mid\frac{|[X]_R\cap X|}{|[X]_R|}>1-\beta\}$$

分别称为 X 关于 R 的 β 下近似,X 关于 R 的 β 上近似。

可变精度粗糙集模型的近似定义是基于多数包含的,β 下近似定义体现了多数包含关系,即 $[x]_R$ 中至少有 $\beta*100\%$ 以上的元素被 X 包含,则 $[x]_R$ 属于 X 的下近似。上近似定义提高了包含的程度要求,即 $[x]_R$ 中至少有 $(1-\beta)*100\%$ 以上的元素被 X 包含,则 $[x]_R$ 属于 X 的上近似,不仅仅是相交非空即可。

与标准粗糙集合类似,各可变精度粗糙集合区域也是由满足不同条件的等价类合并而成。

定义 3[1]　给定论域 U,不可分辨关系 $R\subseteq U\times U$,$X\subseteq U$,$\beta\in(0.5,1)$,则:

β 正区域:$\mathrm{POS}_\beta(X)=\underline{\mathrm{apr}_\beta}(X),\frac{|[X]_R\cap X|}{|[X]_R|}\geqslant\beta$

β 负区域:$\mathrm{NEG}_\beta(X)=U-\overline{\mathrm{apr}_\beta}(X),\frac{|[X]_R\cap X|}{|[X]_R|}\leqslant 1-\beta$

β 边界域:$\mathrm{BND}_\beta(X)=\overline{\mathrm{apr}_\beta}(X)-\underline{\mathrm{apr}_\beta}(X)$,$1-\beta<\frac{|[X]_R\cap X|}{|[X]_R|}<\beta$

β 值体现了近似空间的"包含度"或"精确度",其值的变化直接影响到各区域的变化。由此可定义信息系统的正区域与分类率。

信息系统 $S=(U,C\cup D)$,U 为论域,C 为条件属性集合。D 为决策属性集合,给定 $\beta\in(0.5,1)$,决策属性集 D 对条件属性集 C 的 β 近似依赖或基于 β 的分类率定义为:

$$\gamma(C,D,\beta)=|\mathrm{POS}(C,D,\beta)|/|U|$$

其中:$\mathrm{POS}(C,D,\beta)=\bigcup_{Y\in U/D}\underline{C}_\beta(Y)$ 为 β 正区域。

$\gamma(C,D,\beta)$ 定义了在特定 β 值下,论域 U 中基于决策类能被确定分类的对象比率,即所有决策类 β 正域中对象的个数与整个论域中的对象个数之比,$\beta=1$ 时就是标准粗糙依赖度。

对于一致的信息系统,无论 β 取何值,分类率 γ 总为 1。

*)国家自然科学基金资助项目(60474041)。王加阳　主要研究领域为粗糙集理论与方法、决策支持。

信息的不一致性引起了分类率的变化，从统计概率的角度，决策表的不一致程度越大，分类率呈减小趋势，β取值对γ的影响越大。在可变精度粗糙集模型中，考虑的是动态的β值，其变化直接影响到整个论域的分类，从而影响到分类率的值。一方面，在某一个β取值区间内，γ是保持不变的；另一方面，随着β取值增加，γ呈下降趋势。

Ziarko最先给出了一个关于可变精度粗糙集合模型的定义，它是基于预先给定β值的约简，描述了对固定β值的变精度约简。

定义4[2] 给定决策信息系统$S=(U, C\cup D)$，U为论域，C为条件属性集合。D为决策属性集合，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$RED(C,D,\beta)$，且满足：

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

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

Ziarko探讨了β最优值的选择问题[4]，这些研究都是努力寻求一个特定的β最优值。然而，要确定一个最优的β值是困难的。在相同γ值下，满足约简条件的β值通常为一个区间范围。仅通过上述两条原则来产生β约简可能会产生约简异常，主要体现在约简β区间和正区域的动态变化性，从而导致约简信息系统不能正确体现原信息系统的基本特征。

3 约简异常

在不同分类率γ下，信息系统的β取值区间是不同的，因而表示的信息存在着差异，不存在可比性。根据不同的分类率γ，对应的β区间是分段的，进行各自的约简。在相同的分类率γ下，约简后β取值区间可能产生差异，即与约简前的属性集合比较，约简属性集的β取值范围可能不同。

Beynon分析了约简异常情况[5]，但没有考虑约简过程中的β区间动态性以及β约简对正区域元素的影响，由此导致分类异常。因而，必须进一步考虑这方面的影响，全面体现基于可变精度粗集模型的约简。

定义5 决策信息系统$S=(U, C\cup D)$，U为论域，C为条件属性集合，D为决策属性集合。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)根据条件属性集C对U的分类称为条件分类，表示为$U/C=\{X_1,X_2,\cdots,X_{|U/C|}\}$，其中每一个成员$X_i$为一个条件类。

(2)根据决策属性集D对U的分类称为决策分类，表示为$U/D=\{Y_1,Y_2,\cdots,Y_{|U/D|}\}$，其中每一个成员$Y_j$为一个决策类。

(3)给定条件类$X\in U/C$，令$H_X=\underset{j=1}{\overset{|U/D|}{\mathrm{MAX}}}\frac{|X\cap Y_j|}{|X|}$，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，或简称$X$的条件类阈值。给定决策信息系统[5]$S=(U, C\cup D)$如表1。

表1 决策信息系统

U	a_1	a_2	a_3	a_4	a_5	a_6	D
o_1	1	1	1	1	1	1	M
o_2	1	0	1	0	1	1	M
o_3	0	0	1	1	0	0	M
o_4	1	1	1	0	0	1	F
o_5	1	0	1	0	1	1	F
o_6	0	0	0	1	1	0	F
o_7	1	0	1	0	1	1	F

分析β与γ的关系。当$\gamma=1$时，$\beta\in(0.5,0.667]$；$\gamma=0.571$时，$\beta\in(0.667,1.0]$。表2给出了决策信息系统的条件类阈值。

表2 条件类阈值

$0.667/X1=\{o2,o5,o7\}$	$1.0/X2=\{o1\}$	$1.0/X3=\{o3\}$	$1.0/X4=\{o4\}$	$1.0/X5=\{o6\}$

对给定的分类率，按Ziarko定义规则来计算其约简，可得到在给定分类率γ下β区间和β正区域发生变化的约简，其异常可分成下面的情况。

3.1 区间动态性

表3体现了约简过程中产生第一次条件类归并时的信息系统。这时条件分类也必定发生变化，约简1在保持分类率γ不变情况下，约简后的β区间为$(0.5,0.75]$。表4为约简前后的各种参数情况。

表3 第一次归并的约简信息系统

U	a_3	a_4	a_6	D
o_1	1	1	1	M
o_2	1	0	1	M
o_3	1	1	0	M
o_4	1	0	1	F
o_5	1	0	1	F
o_6	0	1	0	F
o_7	1	0	1	F

表4 第一次归并约简分析

	条件属性	条件分类及阈值	决策分类	分类率γ	β正区域	β区间	异常
约简前	属性集C	$0.667/X1=\{o2,o5,o7\}$ $1.0/X2=\{o1\}$ $1.0/X3=\{o3\}$ $1.0/X4=\{o4\}$ $1.0/X5=\{o6\}$	$\{o1,o2,o3\}\{o4,o5,o6,o7\}$	1.0	U	$(0.5,0.667]$	上限扩张
约简后	$\{a3,a6\}$	$0.60/X1=\{o2,o4,o5,o7\}$ $1.0/X2=\{o1\}$ $1.0/X2=\{o3\}$ $1.0/X2=\{o6\}$		1.0	U	$(0.5,0.75]$	

表5体现了约简过程中产生第二次条件类归并时的信息系统，这时形成了信息系统的一个约简。在保持分类率γ不变情况下，约简后的β区间为$(0.5,0.60]$。表6为约简前后的各种参数情况。

表5 第二次归并的约简信息系统

U	a_3	a_6	D
o_1	1	1	M
o_2	1	1	M
o_3	1	0	M
o_4	1	1	F
o_5	1	1	F
o_6	0	0	F
o_7	1	1	F

表 6 第一次归并约简分析

	条件属性	条件分类及阈值	决策分类	分类率 γ	β 正区域	β 区间	异常
约简前	属性集 C	$0.667/X1=\{o2,o5,o7\}$ $1.0/X2=\{o1\}$ $1.0/X3=\{o3\}$ $1.0/X4=\{o4\}$ $1.0/X5=\{o6\}$	$\{o1,o2,o3\}\{o4,o5,o6,o7\}$	1.0	U	(0.5, 0.667]	上限收缩
约简后	$\{a_3,a_6\}$	$0.60/X1=\{o1,o2,o4,o5,o7\}$ $1.0/X2=\{o3\}$ $1.0/X2=\{o6\}$		(0.5, 0.60]	1.0	(0.5, 0.60]	

从约简的本质看，在可变精度粗糙集模型下，约简前后的信息系统应根据相同的 γ 和 β 来比较其信息的一致性。显然，相同分类能力 γ 下，约简后的信息系统在与原信息系统具有相同分类率的情况下，β 取值范围出现了差异，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。

3.2 分类异常

以上异常体现在 β 区间发生了变化，但约简后的正区域没有改变。由 Ziarko 约简定义，表 7 体现了另一约简前后的各种参数情况。

表 7 分类异常约简分析

	条件属性	条件分类及阈值	决策分类	分类率 γ	β 正区域	β 区间	异常
约简前	属性集 C	$0.667/X1=\{o2,o5,o7\}$ $1.0/X2=\{o1\}$ $1.0/X3=\{o3\}$ $1.0/X4=\{o4\}$ $1.0/X5=\{o6\}$	$\{o1,o2,o3\}\{o4,o5,o6,o7\}$	4/7= 0.571	$\{o1,o3,o4,o6\}$	(0.667, 1.0]	分类变化
约简后	$\{a_3\}$	$0.667/X1=\{o2,o4,o5,o7\}$ $1.0/X2=\{o1,o3,o6\}$		4/7= 0.571	$\{o2,o4,o5,o7\}$	(0.667, 0.75]	

显然，约简后 $POS(C,D,\beta)\neq POS(RED,D,\beta)$，约简前后分类率虽然保持不变，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，出现了分类异常，将使决策信息系统产生不同的分类决策规则。

可见，Ziarko 定义的 β 约简不能保证约简前后规则一致，$\beta=1$ 与 $\beta<1$ 在对分类的影响上有着很大的本质区别，特别在约简上将产生新的概念模式。

4 异常消除

约简实质上即泛化过程。随着属性被约简掉，约简过程中伴随着条件类的归并，使得条件分类的粒度增大，分类率整体呈非递增趋势，使分类率保持不变的归并体现了一个约简的可行性，约简过程为条件分类中条件类不断归并的过程，形成粒度更大的条件分类。

要讨论的是基于可变精度粗糙集模型的约简归并过程，在保持某一确定 γ 值的情况下，β 是如何变化的，归并后的信息系统是否为一个约简。

在基于可变精度粗集模型下，β 约简是根据分类率来定义的，其约简过程的本质体现在保持分类率的不变性。但要保证前后信息的一致性，对分类率描述，必须是基于相同的包含度和正区域，否则分类率间的比较就没有相同的准则。根据上述异常的体现，对可变精度粗糙集模型约简定义必须有更多的限制，从而消除约简异常。

定义 6 给定信息系统 $S=(U, C\cup D)$，U 为论域，C 为条件属性集合，D 为决策属性集合。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 $RED(C,D,\beta_{RED})$，且满足：

(1) $POS(C,D,\beta')=POS(RED(C,D,\beta_{RED}),D,\beta_{RED})$

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

(3) $(\beta_{RED}=\bigcap_P \beta_P)\neq\Phi$

β 约简问题的关键在于约简集合能否表达与原属性集完全一致的信息。定义描述了在可变精度粗糙集模型下，给定分类率的 β 区间约简。条件 (1) 体现了约简正区域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 β 约简过程中，由于 β 值区间在约简过程中可能发生变化，条件 (3) 中的 $\bigcap_P \beta_P$ 表示了属性约简中每一步 β 区间的交集合。β 约简的区间性描述了在一个区间范围内，任意取定一个 β 值，都对应着相同分类率的约简，从而把对给定 β 的约简扩展到 β 区间的约简。

结束语 可变精度粗糙集分析的数据模型表示了统计趋势而非函数依赖。大部分决策信息系统属性之间并不一定存在严格的函数依赖关系，而只是表现出近似依赖的关系。这种关系的不确定性导致了分类的异常，在可变精度粗糙集约简模型下，必须深入地考虑 β 区间的变化，以及约简归并导致的正区域元素集合变化才能避免约简异常的产生。

参 考 文 献

1 Ziarko W. Variable precision rough set model. Journal of Computer and System Science, 1993, 46: 39～59

2 AN A, Shan N, Chan C, et al. Discovering rules for water demand prediction: an enhanced rough-set approach. Engineering Application and Artificial Intelligence, 1996, 9(6): 645～653

3 Ziarko W. Analysis of uncertain information in the framework of variable pecision rough sets. Foundations of Computing and Decision Sciences, 1993, 18: 381～396

4 Ziarko W. Decision making with probabilistic decision tables. In: Zhong N, Skowron A, Ohsuga S, eds. New Directions in Rough Sets, Data Mining, and Granular-Soft Computing. Proc. of the Seventh International Workshop, RSFDGrC'99, Yamaguchi, Japan, 1999. 463～471

5 Beynon M. Reducts within the variable precision rough set model: a further investigation. European Journal of Operation research, 2001, 134: 592～605