变精度粗糙集模型中 β取值范围的确定

庾慧英,刘文奇
(昆明理工大学 理学院,云南 昆明 650093)

摘要: 分析了可信度阈值 β与近似分类质量关系,给出了由近似分类质量阈值 r来确定 β的取值范围的两种算法,并在给定近似分类质量阈值 r的基础上,讨论了两种算法的时间复杂度. 实例证明,通过这两种算法能够得到可信度阈值 β的有效取值范围.

关键词: 变精度粗糙集;阈值
中图分类号: TP18　　**文献标识码**: A　　**文章编号**: 1007 - 855X(2005)06 - 0109 - 03

Confirming the Range of β Value in the Variable Precision Rough Sets Model

YU Hui-ying, LIU Wen-qi
(Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: The relation between β threshold value of reliability and approximate quality of classification is analyzed in this paper, and two kinds of algorithms to confirm the range of β value by the r threshold value of approximate quality of classification are provided. In addition, the complexity of the two kinds of algorithms is discussed based on a certain r value. The example is used to show the effective range of β value by the two kinds of algorithms.

Key words: variable precision rough sets; threshold value

0 引 言

Ziarko[1]提出的变精度粗糙集模型是对 Pawlak[2]粗糙集理论的扩充,它是在基本粗糙集模型的基础上引进 $\beta(0 \leq \beta < 0.5)$即允许一定程度的错误分类率存在. 一方面完善了近似空间的概念,另一方面也有利于粗糙集理论从认为不相关的数据中发现相关的数据. 变精度粗糙集模型的主要任务是解决属性间无函数或不确定关系的数据分类问题,这给研究者处理由于噪声所引起的数据不一致性问题提供了很好的方法. 后来 An等[3]人又将 β定义为正确分类率 $(0.5 < \beta \leq 1)$,并称之为强化粗糙集,在此基础上大量的文献[4,5]讨论了 β的近似约简及规则提取. 设 P为条件属性集,Q为决策属性集,用近似分类质量 $\gamma^{\beta}(P,Q) = \frac{|POS(P,Q,\beta)|}{|U|}$来表示属性集 Q与 P的 β依赖性,并以此作为近似约简的一个准则(保持近似分类质量不变). 近似分类质量 $\gamma^{\beta}(P,Q)$表示的是条件属性类以不低于 β的正确分类率(可信度)划入决策类的对象的百分比,代表了决策表的分类能力. 上述的文献中都是在假定某个特定的阈值 β的基础上来讨论的,在现实生活中,决策分析者往往并不知道 β的取值,但有可能知道用户所要求的质量不得低于某一阈值 r. 如何根据 r来确定 β的取值范围是本文研究的重点.

1 基本概念

定义 1[3]　设 $T = (U, A = C \cup D)$为一决策信息系统, $P \subseteq C$为条件属性集, $Q \subseteq D$为决策属性集,

收稿日期: 2004 - 11 - 09.
第一作者简介:庾慧英 (1979～),女,在读硕士研究生. 主要研究方向:粗糙集理论及应用.
E - mail: huiying1977@163.com

分类 $U/P=\{X_1,\cdots,X_n\}$, $U/Q=\{Y_1,\cdots,Y_m\}$, $0.5<\beta\le 1$,对任意的 $Y\in U/Q$,定义:

$POS_p^{\beta}(Y)=\cup\{X_i\in U/P\mid pr(Y/X_i)\ge\beta\}$ 为 β-正域

$NEG_p^{\beta}(Y)=\cup\{X_i\in U/P\mid pr(Y/X_i)\le 1-\beta\}$ 为 β-负域

$BND_p^{\beta}(Y)=\cup\{X_i\in U/P\mid 1-\beta<pr(Y/X_i)<\beta\}$ 为 β-边界域

其中 $pr(Y/X_i)=\dfrac{|Y\cap X_i|}{|X_i|}$ 表示条件类分配之决策类的正确分类率,可以理解为可信度,β可理解为可信度阈值.近似分类质量定义为 $\gamma^{\beta}(P,Q)=\dfrac{|POS(P,Q,\beta)|}{|U|}$,其中 $POS(P,Q,\beta)=\bigcup_{Y\in U/Q}POS_p^{\beta}(Y)$.

由定义1我们可以得到如下的定理:

定理1 (1)设 $0.5<\beta_1\le\beta\le 1$,若 $X\in POS_p^{\beta}(Y)$,则 $X\in POS_p^{\beta_1}(Y)$.

(2)设 $0.5<\beta\le\beta_1\le 1$,若 $X\notin POS_p^{\beta}(Y)$,则 $X\notin POS^{\beta_1}(Y)$.

定理2 当 $0.5<\beta_1\le\beta_2\le 1$时,有 $\gamma^{\beta_2}(P,Q)\le\gamma^{\beta_1}(P,Q)$.

证明 因为当 $0.5<\beta_1\le\beta_2\le 1$时,对任意的 $Y\in U/Q$,有 $POS_p^{\beta_2}(Y)\le POS_p^{\beta_1}(Y)$,从而有 $POS(P,Q,\beta_2)=\bigcup_{Y\in U/Q}POS_p^{\beta_2}(Y)\le\bigcup_{Y\in U/Q}POS_p^{\beta_1}(Y)=POS(P,Q,\beta_1)$,而 $\gamma^{\beta_2}(P,Q)=\dfrac{|POS(P,Q,\beta)|}{|U|}$,所以有 $\gamma^{\beta_2}(P,Q)\le\gamma^{\beta_1}(P,Q)$.

由定理2可知:β的选择会影响近似分类质量的高低;而当用户对决策表的分类能力提出要求时,即要求近似分类质量不得低于某值 r(近似分类质量阈值),那么 r就会影响到β值的选择.

定理3 设 r为近似分类质量阈值,β_1为满足 r要求的可信度阈值,即 $\gamma^{\beta_1}(P,Q)\ge r$,则对任意的 $\beta\in(0.5,\beta_1]$,有 $\gamma^{\beta}(P,Q)\ge r$.

证明 由定理2易得.

对给定的 r值,满足 r要求的β最大值称为可信度上限,记为 β^r,$\beta\in(0.5,\beta^r)$.

定理4 设 r_1, r_2为近似分类质量阈值,若 $r_1\le r_2$,则 $\beta^{r_1}\ge\beta^{r_2}$.

证明 因有 $\gamma^{\beta^{r_1}}(P,Q)\ge r_1$, $\gamma^{\beta^{r_2}}(P,Q)\ge r_2$,且 $r_1\le r_2$,从而有 $\gamma^{\beta^{r_2}}(P,Q)\ge r_2\ge r_1$,即 β^{r_2}也是满足 r_1要求的可信度阈值,所以 $\beta^{r_1}\ge\beta^{r_2}$.

2 根据近似分类质量阈值 r确定β取值范围的算法研究

由前面讨论的 r与β的关系可知,当用户所要求的 r值较高时,我们所考虑的β值相对要低一些,当用户所要求的 r值较低时,我们相应地可以考虑选择较高的β值.这样即能满用户对 r的要求,又相应地可以提高规则的可信度.

设 $T=(U,A=P\cup Q)$为一决策信息系统,P为条件属性集,Q为决策属性集,$U/P=\{X_1,\cdots,X_n\}$, $U/Q=\{Y_1,\cdots,Y_m\}$,令 $M=\{pr(Y_j/X_i)\mid\forall_i\le 0.5, j\le m, pr(Y_i/X_i)>0.5\}$,对 M中的值按从小到大的顺序排序,设为 $M=\{p_1,p_2,\cdots,p_l\}$ $(1\le l\le mn)$.

算法1:

⑴令 $k=1$,若 $\gamma^{p_1}(P,Q)$,转(2);否则,返回用户要求降低 r值或提供新的信息以扩充决策表.

⑵令 $k=k+1$,若 $k\le l$,继续;否则,输出 $\beta\in(0.5,p_{k-1}]$.

⑶计算 $\gamma^{p_k}(P,Q)$,若 $\gamma^{p_k}(P,Q)\ge r$,转(2);否则,输出 $\beta\in(0.5,p_{k-1}]$.

算法2:

(1)令 $k=l$,若 $\gamma^{p_l}(P,Q)<r$,转(2);否则,输出 $\beta\in(0.5,p_l]$.

(2)令 $k=k-1$,若 $k\ge 1$,继续;否则停止.

(3)计算 $\gamma^{p_k}(P,Q)$,若 $\gamma^{p_k}(P,Q)<r$,转(2);否则,输出 $\beta\in(0.5,p_k]$.

以上两种算法都可由 r计算出β可允许的取值范围,算法1是从最小值开始搜索,直到不满足要求为止;算法2是从最大值开始搜索.那么在 r值给定的情况下选取哪一种算法最合理,时间复杂度最低?

当 l为奇数时,取 $mid = (l+1)/2$,如果 $r > \gamma^{p_{mid}}(P,Q)$,选择算法 1时,循环时间复杂度为 $o((l+1)/2)$,选择算法 2时,循环时间复杂度为 $o(l)$,所以此时应选择算法 1;如果 $r < \gamma^{p_{mid}}(P,Q)$,选择算法 1时,循环时间复杂度为 $o(l)$,选择算法 2时,循环时间复杂度为 $o((l+1)/2)$,所以此时应选择算法 2. 如果 $r = \gamma^{p_{mid}}(P,Q)$, $\beta \in (0.5, p_{mid}]$.

当 l为偶数时,取 $mid_1 = l/2, mid_2 = l/2+1$,如果 $r > \gamma^{p_{mid_1}}(P,Q)$,选择算法 1时,循环时间复杂度为 $o(l/2)$,选择算法 2时,循环时间复杂度为 $o(l)$,所以此时应选择算法 1;如果 $r < \gamma^{mid_2}(P,Q)$,选择算法 1时,循环时间复杂度为 $o(l)$,选择算法 2时,循环时间复杂度为 $o(l/2+1)$,所以此时应选择算法 2. 如果 $\gamma^{mid_2}(P,Q) < r \leq \gamma^{mid_1}(P,Q)$, $\beta \in (0.5, p_{mid_1}]$;如果 $r = \gamma^{mid_2}(P,Q)$, $\beta \in (0.5, p_{mid_2})$.

3 算 例

决策表 $T = (U, A = P \cup Q)$如表 1所示,条件属性集 $P = \{a, b, c, d\}$,决策属性集 $Q = \{e\}$, f_r为频数 (总体样本数为 50).

$U/P = \{X_1, X_2, \cdots, X_8\}$, $X_1 = \{x_1\}$, $X_2 = \{x_2\}$, $X_3 = \{x_3, x_4\}$, $X_4 = \{x_5\}$, $X_5 = \{x_6\}$, $X_6 = \{x_7, x_8\}$, $X_7 = \{x_9, x_{10}\}$, $X_8 = \{x_{11}, x_{12}\}$.

$U/P = \{Y_1, Y_2, Y_3\}$, $Y_1 = \{x_1, x_4, x_5, x_6, x_{12}\}$, $Y_2 = \{x_2, x_3, x_8, x_{10}, x_{11}\}$, $Y_3 = \{x_7, x_9\}$.

计算 $M = \{p_1 = 2/3, p_2, 3/4, p_3 = 4/5, p_4 = 12/13, p_5 = 1\}$

若 $r = 0.8, mid = (l+1)/2 = 3, \gamma^{p_3}(P,Q) = 0.72 < 0.8$.

选择算法 1,取 $k = 1, \gamma^{p_1}(P,Q) = 1 > 0.8$, 再取 $k = 2, \gamma^{p_2}(P,Q) = 0.88 > 0.8$;

再取 $k = 3, \gamma^{p_3}(P,Q) = 0.72 < 0.8$,所以 $\beta \in (0.5, p_2]$,即 $\beta \in (0.5, 3/4]$. 循环次数为 3. 若选择算法 2则要循环 4次才能得到结果 .

同理,若 $r = 0.6, mid = (l+1)/2 = 3, \gamma^{p_3}(P,Q) = 0.72 > 0.6$,选择算法 2求得 .

表 1　决策表

Tab. 1　Decision chart

U	f_r	条件属性			决策属性	
		a	b	c	d	e
x_1	3	0	0	1	1	0
x_2	12	0	1	0	2	1
x_3	12	0	2	1	2	1
x_4	1	0	2	1	2	0
x_5	2	1	0	1	2	0
x_6	1	1	1	0	1	0
x_7	4	1	2	0	1	2
x_8	2	1	2	0	1	1
x_9	4	2	0	0	1	2
x_{10}	1	2	0	0	1	1
x_{11}	6	2	1	2	0	1
x_{12}	2	2	1	2	0	0

4 结 论

变精度粗糙集模型对处理带有噪声的数据是十分有效的,该模型通过引入一个可信度阈值 $0.5 < \beta \leq 1$,从而具有一定的容错能力 . 我们以前都是在假定某个特定的阈值 β,进而讨论 β的近似约简及规则提取 . 当用户对决策表的分类能力提出要求,即提供了近似分类质量的阈值 r时,对假定的某个阈值 β可能就不满足 r的要求了 . 本文给出了由 r确定 β的两种算法,这样得到的 β值既能满用户对 r的要求,又相应的可以提高规则的可信度 (取 β^r时规则的可信度最高);并讨论了根据 r的取值来选择不同的算法 . 实例证明这两种算法是有效的 .

参考文献:

[1] Ziarko W. Variable Precicsion Rough Set Model[J]. Journal of Computer and System Sciences, 1993, 46(1): 39～59.

[2] Pawlak Z. Rough Sets[J]. Int. J. Coumput. Inf. Sci, 1982, (11): 341～356.

[3] An A, Shan N, Chan C, et al. Discovering rules for water demand prediction: An enhanced rough - set approach[J]. Engineering Application and Artifical Intelligence, 1996, 9(6): 645～653.

[4] 陶志,许宝栋,汪定伟,等 . 基于可变精度粗糙集理论的粗糙规则挖掘算法[J]. 信息与控制, 2004, 33(1): 18～30.

[5] 张宏宇,梁吉业 . 不完备信息系统下的变精度粗糙集模型及其知识约简算法[J]. 计算机科学, 2003, 30(4): 153～155.