文章编号: 1001-0920(2007)01-0078-03

基于变精度粗糙集阈值的选取

赵越岭[1,2], 王建辉[1], 顾树生[1]
(1. 东北大学 信息科学与工程学院, 沈阳 110004; 2. 辽宁工学院 信息科学与工程学院, 辽宁 锦州 121001)

摘 要: 针对变精度粗糙集阈值选取缺乏可预见性的问题,提出了基于集合可辨性的阈值选取方法. 在变精度粗糙集模型研究的基础上,首先在满足集合可辨性条件下,选定可分辨类相关的阈值的上界;然后在近似分类质量保持不变的前提下,选定阈值区间;最后依据变精度粗糙集近似约简标准,确定阈值范围. 实例计算验证了该方法的有效性.

关键词: 变精度粗糙集; 分类质量; 集合可辨性; 阈值
中图分类号: TP181 **文献标识码**: A

Choice of threshold value based on variable precision rough sets

ZHAO Yue-ling[1,2], *WANG Jian-hui*[1], *GU Shu-sheng*[1]
(1. College of Information Science and Engineering, Northeastern University, Shenyang 110004, China. 2. College of Information Science and Engineering, Liaoning Institute of Technology, Jinzhou 121001, China. Correspondent: ZHAO Yue-ling, E-mail: zhao7267@163.com)

Abstract: To the problem that the variable precision rough set (VPRS) lacks a feasible method to determine a parameter of threshold value, a choice method of threshold value based on discernibility of set is presented. The upper boundary of threshold value with the resolution class is selected under satisfying discernibility of set by studying the model of VPRS. The section of threshold value is chosen under maintaining the invariability of approximation quality of classification. Finally, the scope of threshold value is defined based on approximate reduct criterion of VPRS. The example of computing and analyzing with the proposed method shows its effectiveness.

Key words: Variable precision rough set; Quality of classification; Discernibility of set; Threshold value

1 引 言

粗糙集理论以其简单实用的特点,在诸多领域获得了成功应用[1,2]. 但是,由于实际决策过程的复杂性和不确定信息的多样性,粗糙集理论在某些条件下具有一定的局限性. 粗糙集理论缺乏对复杂系统的处理机制,对于不确定性概念的边界区域,刻画过于简单,缺乏对噪音数据的适应能力. 尤其在数据集中存在噪音等干扰情况下,经典理论会由于对数据的过拟合而使其对新对象的预测能力大为降低. 在实际应用中,噪音是难免的,为增强粗糙集合模型的抗干扰能力, Ziarko 提出了一种变精度粗糙集模型[3]. 它是在基本粗糙集模型的基础上引入了 β $(0 \leq \beta < 0.5)$,即允许存在一定程度的错误分类率. Aijun 提出了一种变精度粗糙集模型[4],引入了 β $(0.5 < \beta \leq 1)$ 作为正确率. 在变精度粗糙集研究的文献中,大都没有具体阐述有关阈值 β 的选取问题. Ziarko 依据决策者的经验来选取特定的 β 值[3]; Beynon 提出在满足近似质量分类的前提下,以一个区域间隔选取 β 值[5]. Katzberg 等进一步提出了不对称边界的变精度粗糙集模型[6-8],即存在两个参数 l 和 u,从而使粗糙集模型更加一般化.

本文在对 Aijun 变精度粗糙集模型 $(0.5 < \beta \leq 1)$ 研究的基础上,针对阈值 β 选取问题,提出了基于集合可辨性的阈值 β 选取的方法. 在选定可分辨类相关的阈值上界的条件下,依据 β 的约简标准,确定阈值 β 的范围.

2 变精度粗糙集模型

2.1 变精度粗糙集模型[4]

设 U 为有限对象构成的论域, R 为 U 上的等价关系,其构成的等价类为 $U/R = \{X_1, X_2, \cdots, X_n\}$.

对于任意 $Z \subseteq U$, $P \subseteq C$,定义 Z 关于概率近似空间 (U, R, P_r) 依阈值参数 $0.5 < \beta \leq 1$ 的下近似、

收稿日期: 2005-09-19; **修回日期**: 2005-11-25.
基金项目: 国家自然科学基金项目(60274024, 60474040).
作者简介: 赵越岭(1972—),男(回族),辽宁凌海人,博士生,从事智能控制、粗糙集理论等研究; 顾树生(1939—),男,黑龙江绥化人,教授,博士生导师,从事智能控制理论及应用等研究.

上近似和边界域如下:

β的下近似

$$\underline{\mathrm{apr}}_P^{\beta}(Z) = \bigcup_{\Pr(Z|X_i)\geq\beta}\{X_i \in E(P)\}. \quad (1)$$

式中: $E(P)$ 表示条件属性 P 的等价类, $\Pr(Z|X_i) = \mathrm{card}(Z\cap X_i)/\mathrm{card}(X_i)$ 表示事件 X_i 发生时 Z 出现的条件概率.

β的上近似

$$\overline{\mathrm{apr}}_P^{\beta}(Z) = \bigcup_{\Pr(Z|X_i)>1-\beta}\{X_i \in E(P)\}. \quad (2)$$

β的边界域

$$\mathrm{bnr}_P^{\beta}(Z) = \bigcup_{1-\beta<\Pr(Z|X_i)<\beta}\{X_i \in E(P)\}. \quad (3)$$

当 $\beta = 1$ 时,变精度粗糙集即为标准粗糙集. 随着 β减小,变精度粗糙集的近似边界区域变窄, 即变精度粗糙集意义下的不确定区域变小. 因此,变精度粗糙集对数据不一致性具有一定的容忍度,在某些场合可以更好地抗噪声,增强产生规则的鲁棒性.

2.2　近似分类质量

设 $P \subseteq C$,分类质量 $\gamma^{\beta}(P,D)$ 定义为

$$\gamma^{\beta}(P,D) = \frac{\mathrm{card}\left(\bigcup_{\Pr(Z|X_i)\geq\beta}\{X_i \in E(P)\}\right)}{\mathrm{card}(U)}, \quad (4)$$

式中 $\gamma^{\beta}(P,D)$ 仅依赖于 β值. 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.

2.3　近似约简

在考虑正确率 $\beta(0.5 < \beta \leq 1)$ 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

$$\gamma^{\beta}(C,D) = \gamma^{\beta}(\mathrm{red}(C,D),D); \quad (5)$$

从 $\mathrm{red}(C,D)$ 中去掉任何一个属性,都将使式(5) 不成立.

$\mathrm{red}(C,D)$ 是指条件属性 C 对决策属性 D 的一个近似约简. 显然,用近似约简进行归类与用整个条件属性集归类在数目上能够保持相同.

3　变精度粗糙集阈值 β选取

3.1　变精度粗糙集阈值依据 β的约简选取标准

1) 近似分类质量尽可能大;

2) 在满足 1) 的前提下,阈值 β有最大上界;

3) 在满足 2) 的前提下,β的约简中属性最少;

4) 在满足 3) 的前提下,阈值 β有最大间隔.

3.2　基于集合可辨性 β选取

3.2.1　集合的相对可辨性[3]

集合边界的可辨别概念是相对的,如果允许一个大的分类误差,则在假定的分类误差限内,集合可能有较大的可辨别性.

如果集合 Z 的 β的边界域 $\mathrm{BNR}_P^{\beta}(Z) = \varnothing$,或 $\underline{\mathrm{apr}}_P^{\beta}(Z) = \overline{\mathrm{apr}}_P^{\beta}(Z)$,则称集合 Z为 β可辨别,否则称 β不可辨别.

定理 1[9]　若 Z 在阈值 $0.5 < \beta \leq 1$ 上是可辨别的,则 Z 在任何 $0.5 < \beta_1 < \beta$上也是可辨别的.

定理 2[9]　若 Z 在阈值 $0.5 < \beta \leq 1$ 上是不可辨别的,则 Z在任何$\beta < \beta_1 < 1$ 上也是不可辨别的.

定理 1 和定理 2 表明,与每一个可分辨类相关的是 β值的上界,等于或小于 β值的上界时可分辨,超过这个上界则不可分辨.

3.2.2　集合可辨性阈值 β选取

对每一个相对粗糙集 X,都存在一个阈值 β,使得集合 X 在这个阈值水平上是可辨别的. 令

$$\mathrm{ndis}(R,X) = \{0.5 < \beta \leq 1 \mid \mathrm{BNR}_P^{\beta}(Z) \neq \varnothing\}, \quad (6)$$

其中 $\mathrm{ndis}(R,X)$ 是满足 X 是不可分辨的所有β值的全体. 满足 X 是可分辨的阈值 β的最大值称为可辨别的阈值. 依据定理 1 和定理 2 可知,这个阈值等于 $\mathrm{ndis}(R,X)$ 的最小上界,即

$$\xi(R,X) = \inf \mathrm{ndis}(R,X), \quad (7)$$

$$\beta = \xi(R,X) = \min(m_1,m_2), \quad (8)$$

$$m_1 = 1 - \max\{\Pr(Z|X_i) \mid Z \subseteq U, \Pr(Z|X_i) < 0.5\}, \quad (9)$$

$$m_2 = \min\{\Pr(Z|X_i) \mid Z \subseteq U, \Pr(Z|X_i) > 0.5\}. \quad (10)$$

4　算　　例

数据集[4] 如表 1 所示. 其中: U 为有限对象构成的论域; $a_0, a_1, \cdots, a_{13}$ 为条件属性; D 为决策属性.

表 1　信息系统

U	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	D
O_1	6	6	7	7	9	7	1	1	0	1	3	2	6	0	0
O_2	1	5	7	6	9	9	0	0	0	0	1	5	1	0	0
O_3	6	6	7	7	9	7	1	1	0	1	3	2	6	0	0
O_4	3	7	7	7	5	2	0	0	0	0	1	3	1	2	1
O_5	3	7	7	7	8	2	0	0	0	0	1	7	1	9	1
O_6	3	7	7	7	5	2	0	0	0	0	1	3	1	2	1
O_7	6	6	8	7	3	2	0	0	0	0	6	0	6	2	2
O_8	3	7	7	7	5	2	0	0	0	0	1	3	1	2	2

Step1　选定可分辨类阈值的上界. 基于条件属性 $C = \{a_0, a_1, \cdots, a_{13}\}$ 的等价类: $X_1 = \{O_1, O_3\}$, $X_2 = \{O_2\}$, $X_3 = \{O_4, O_6, O_8\}$, $X_4 = \{O_5\}$, $X_5 = \{O_7\}$. 基于决策属性 D 的等价类: $Z_0 = \{O_1, O_2, O_3\}$, $Z_1 = \{O_4, O_5, O_6\}$, $Z_2 = \{O_7, O_8\}$. 当 $Z_0 = \{O_1, O_2, O_3\}$ 时,有

$$P(Z_0 \mid X_1) = \frac{\mathrm{card}(Z_0 \cap X_1)}{\mathrm{card}(X_1)} = 1.$$

同理, $P(Z_0 \mid X_2) = 1$, $P(Z_0 \mid X_3) = 0$, $P(Z_0 \mid X_4) = 0$, $P(Z_0 \mid X_5) = 0$. 由式(9)计算得 $m_1 = 1 - \max(0,0,0) = 1$;由式(10)计算得 $m_2 = \min(1,1) = 1$;由式(8)计算得 $\beta_0 = \min(1,1) = 1$.

类似计算,当 $Z_1 = \{O_4, O_5, O_6\}$ 时,$\beta_1 = 0.667$;当 $Z_2 = \{O_7, O_8\}$ 时,$\beta_2 = 0.667$. 因此,满足决策属性 D 的可分辨阈值上界 $\beta = 0.667$.

Step2 在不同的阈值 β 区间计算近似质量. 由式(5)可计算:当 $\beta \in (0.5, 0.667]$ 时,$\gamma_C(D) = 1$;当 $\beta \in (0.667, 1)$ 时,$\gamma_C(D) = 0.625$,如图 1 所示.

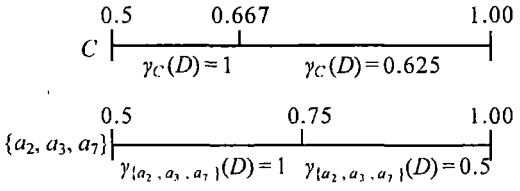

图 1 β 取值范围及近似质量

下面仅对表 1 的决策表的一个约简 $\{a_2, a_3, a_7\}$ 的阈值 β 选取进行讨论.

表 2 条件属性 $\{a_2, a_3, a_7\}$ 信息系统

U	a_2	a_3	a_7	D
O_1	7	7	1	0
O_2	7	6	0	0
O_3	7	7	1	0
O_4	7	7	0	1
O_5	7	7	0	1
O_6	7	7	0	1
O_7	8	7	0	2
O_8	7	7	0	2

重复 Step1 和 Step2,计算出满足决策属性 D 的可分辨阈值上界及不同 β 区间的近似质量. 基于条件属性 $C = \{a_2, a_3, a_7\}$ 的等价类

$$U/C = \{X_1, X_2, X_3, X_4\} = \{\{O_1, O_3\}, \{O_2\}, \{O_4, O_5, O_6, O_8\}, \{O_7\}\},$$

基于决策属性 D 的等价类

$$U/D = \{Z_0, Z_1, Z_2\} = \{\{O_1, O_2, O_3\}, \{O_4, O_5, O_6\}, \{O_7, O_8\}\}.$$

同理能计算出在条件属性 $C = \{a_2, a_3, a_7\}$ 下,满足决策属性 D 的可分辨阈值 $\beta = 0.75$. 当 $\beta \in (0.5, 0.75]$ 时,$\gamma_{\{a_2, a_3, a_7\}}(D) = 1$;当 $\beta \in (0.75, 1)$ 时,$\gamma_{\{a_2, a_3, a_7\}}(D) = 0.5$. 如图 1 所示.

Step3 依据近似约简标准及集合可辨性,确定阈值范围. 由近似约简条件可知,$\gamma_C(D) = \gamma_{\{a_2, a_3, a_7\}}(D) = 1$ 满足式(5),阈值 β 为 $(0.5, 0.667] \subset (0.5, 0.75]$,即阈值 $\beta \in (0.5, 0.667]$ 范围内,条件属性 $\{a_2, a_3, a_7\}$ 是依据 β 的约简.

上述分析表明,β 值与近似分类质量逆相关. 一方面,在特定的 β 取值区间内,近似分类质量保持不变;另一方面,随着 β 取值增加,近似分类质量呈下降趋势,如图 1 所示.

5 结 语

变精度粗糙集引入阈值 $\beta(0.5 < \beta \leq 1)$ 参数后,扩充了基本粗糙集理论,更好地体现了数据分析中的数据相关性,为获取近似决策规则奠定了基础. 本文在对变精度粗糙集研究的基础上,提出了基于集合可辨性的阈值选取方法. 在选定可分辨类相关阈值上界的前提下,依据变精度粗糙集近似约简标准确定阈值范围. 计算实例验证了该方法的有效性.

参考文献(References)

[1] Pawlak Zdzislaw. Rough sets: Theoretical aspects of reasoning about data [M]. London: Kluwer Academic Publishers, 1991.

[2] Pawlak Zdzislaw. Rough set theory and its applications to data analysis[J]. Cybernetics and Systems, 1998, 29(7): 661-688.

[3] Ziarko Wojciech. Variable precision rough set model [J]. J of Computer and System Sciences, 1993, 46(1): 39-59.

[4] Aijun An, Ning Shan, Christine Chan, et al. Discovering rules for water demand prediction: An enhanced rough-set approach [J]. Engineering Applications of Artificial Intelligence, 1996, 9(6): 645-653.

[5] Beynon Malcolm J. Griffiths Benjamin. An expert system for the utilization of the variable precision rough sets model [C]. Rough Sets and Current Trends in Computing. Heidelbeyg: Springer-Verlag, 2004: 714-720.

[6] Katzberg J D, Ziarko W. Variable precision rough sets with asymmetric bounds[C]. Proc of the Int Workshop on Rough Sets and Knowledge Discovery (RSKD '93). Heidelberg: Springer-Verlag, 1993: 167-177.

[7] Beynon Malcolm J. The introduction and utilization of (l, u)-graphs in the extended variable precision rough sets model[J]. Int J of Intelligent Systems, 2003, 18(10): 1035-1055.

[8] Beynon Malcolm J. Degree of dependency and quality of classification in the extended variable precision rough sets model[C]. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Heidelberg: Springer-Verlag, 2003: 287-290.

[9] Beynon M. Reducts within the variable precision rough sets model: A further investigation[J]. European J of Operational Research, 2001, 134(3): 592-605.