* 1

分级粗糙集和分级知识约简

袁修久，　高生强，　杨　宇
(空军工程大学　理学院，陕西　西安　710051)

摘　要：Pawlak 粗糙集模型认为一个元素要么属于一个集合，要么不属于该集合，要么可能属于该集合，把可能属于该集合的元素的全体称为边界。Pawlak 粗糙集模型对边界的研究较少。文章认为对边界的隶属度差别较小的元素以同一个量级属于边界，从而可按一个对象对边界的隶属量级对边界进行划分。基于这一思想提出了分级粗糙集模型和分级最大分布约简、分级分布约简的概念。给出了这两种约简的判定定理及辨识矩阵以及相应的核属性的等价条件。分级粗糙集模型推广了 Pawlak 粗糙集及变精度粗糙集模型。

关键词：粗糙集；分级粗糙集；分级约简
中图分类号：TP18　　**文献标识码**：A　　**文章编号**：1009 - 3516(2009)02 - 0091 - 04

Pawlak 粗糙集[1]把不能用已有知识进行分类的对象都归为边界。文献[2]提出的变精度模型对 Pawlak 粗糙集的边界进行了“压缩”，缩小了边界，它是对边界进行“部分认识”的最早的文献，后来的文献[3 - 4]提出的熵约简和分布约简等概念也在一定的程度上考虑了边界信息。粗糙隶属函数[5]是由数据自身提供的，但是 Pawlak 粗糙集并未充分地利用这一信息。分级粗糙集模型推广了 Pawlak 粗糙集及变精度粗糙集模型。属性约简是粗糙集的主要内容，常见的约简有分布约简[3]、熵约简[4]和最大分布约简[6]、基于变精度粗糙集模型的约简[7-8]和广义决策约简[9]等。文献[10]已证明了分布约简和熵约简是等价的。分布约简要求约简后的决策表同原决策表的各对象的分布相等。由于实际中数据常常含有噪音，所以该约简的条件很“苛刻”。本文提出的分级分布约简和分级最大分布约简放宽了分布约简和最大分布约简定义的条件，能够适用包含一定噪音数据的数据集。

1　Pawlak 粗糙集和变精度粗糙集模型的不足

Pawlak 粗糙集利用上近似和下近似近似一个集合。设 X 是论域 U 上的任意一个子集，R 是 U 上的等价关系，Rough 集中的粗糙隶属函数定义如下：

$$\mu_X^R(x) = |X \cap [x]_R| / |[x]_R| \tag{1}$$

式中：$|\cdot|$ 表示集合的基数；$[x]_R$ 是 x 的等价类。利用粗糙隶属函数集合 X 的下、上近似可以定义为：

$$\begin{cases} \underline{R}(X) = \{x \in U \mid \mu_X^R(x) = 1\} \\ \overline{R}(X) = \{x \in U \mid \mu_X^R(x) > 0\} \end{cases}$$

X 的边界定义为：

$$BN_R(X) = \{x \in U \mid 0 < \mu_X^R(x) < 1\}$$

粗糙隶属函数被理解为一个系数，它表示 x 是 X 的成员的不精确性。同一等价类的元素具有相同的不精确性，边界点上的对象都有不精确性，然而它们的不精确性是不相同的。粗糙集推广了经典集合论，它认为一个对象要么属于某个集合，要么不属于某个集合，要么可能属于某个集合。可能属于某个集合的对象就

* **收稿日期**：2007 - 10 - 08
基金项目：国家自然科学基金资助项目(60663003)
作者简介：袁修久(1966 -)，男，陕西旬阳人，教授，主要从事数据挖掘研究. E - mail：yuanxiujiu@ sohu. com

构成了边界,从隶属函数的观点看同一等价类上的对象隶属于某个集合的程度是相同的,而不同等价类中的对象属于某个集合的程度则可能是不相同的。然而 Pawlak 粗糙集不考虑这种的差别,统一认为它们"都可能属于某个集合"。这显然不利于我们清晰地认识边界。Ziarko 在允许一定程度的错误分类率存在的基础上,对 Pawlak 粗糙集模型进行了推广,提出了变精度粗糙集模型。

定义 1 设 R 是论域 U 上的等价关系,$X\subseteq U$,$\beta\in(0.5,1]$。

$$\begin{cases}\underline{R}^{\beta}(X)=\{x\mid\mu_X^R(x)\geqslant\beta\}\\ \overline{R}^{\beta}(X)=\{x\mid\mu_X^R(x)\geqslant 1-\beta\}\end{cases}$$

变精度模型边界 $BN_R^{\beta}(X)=\{x\mid 1-\beta<\mu_X^R(x)<\beta\}$,同 Pawlak 粗糙集边界定义相比,它实质上缩小了边界,即对部分边界元素进行了明确分类。但是对于边界 $BN_R^{\beta}(X)$ 仍然没有深入的认识。

例 1 设论域 $U=\{x_1,x_2,\cdots,x_{10000}\}$。记 $X_1=\{x_1,x_2,\cdots,x_{1000}\}$,$X_2=\{x_{1001},x_{1002},\cdots,x_{2000}\}$,$X_3=\{x_{2001},x_{2002},\cdots,x_{2100}\}$,$X_4=\{x_{2101},x_{2102},\cdots,x_{2200}\}$,$X_5=\{x_{2201},x_{2202},\cdots,x_{2300}\}$,$X_6=\{x_{2301},x_{2302}\}$,$X_7=\{x_{2303},x_{2304},\cdots,x_{10000}\}$,设 $X=\{x_1,x_2,\cdots,x_{250};x_{1001},x_{1002},\cdots,,x_{1255};x_{2001},x_{2002},\cdots,x_{2050};x_{2101},x_{2102},\cdots,x_{2152};x_{2201},x_{2202},\cdots,x_{2295};x_{2301},x_{2302}\}$,显然 $X_1,X_2,X_3,X_4,X_5,X_6,X_7$ 是论域 U 的一个划分,则根据式(1)$\mu_X^R(X_1)=0.25$,$\mu_X^R(X_2)=0.255$,$\mu_X^R(X_3)=0.5$,$\mu_X^R(X_4)=0.52$,$\mu_X^R(X_5)=0.95$,$\mu_X^R(X_6)=1$。

若取 $\beta=0.9$,则:

$$BN_R^{0.9}(X)=\{x\mid 0.1<\mu_X^R(x)<0.9\}=X_1\cup X_2\cup X_3\cup X_4$$

X_1,X_2,X_3,X_4 中元素属于边界的程度不相同。x_1 属于边界的程度为 0.25,x_{1001} 属于边界的程度为 0.255,x_{2001} 属于边界的程度为 0.5,x_{2101} 属于边界的程度为 0.52,我们认为 x_1 和 x_{1001} 对边界的隶属程度处在一个数量级上,x_{2001} 和 x_{2101} 也处在一个数量级上,因此不加区别。即可以认为 $X_1\cup X_2$ 中的元素属于边界的程度处于同一个数量级上。$X_3\cup X_4$ 中的元素属于边界的程度也在一个数量级上,而 $X_1\cup X_2$ 与 $X_3\cup X_4$ 中的元素属于边界的程度有差别。基于这样的考虑,引入分级粗糙集的概念。

2 分级粗糙模型

定义 2 设 $\alpha_0,\alpha_1,\cdots,\alpha_n$ 是给定的常数,并且满足 $0\leqslant\alpha_0<\alpha_1<\cdots<\alpha_n\leqslant 1$,定义 X 的分级近似为:

$$\begin{cases}\underline{R}^{\alpha_n}(X)=\{x\mid\mu_R^X(x)\geqslant\alpha_n\}\\ BN_{\alpha_0}^{\alpha_1}(X)=\{x\mid\alpha_0<\mu_R^X(x)<\alpha_1\}\\ BN_{\alpha_{i-1}}^{\alpha_i}(X)=\{x\mid\alpha_{i-1}<\mu_R^X(x)<\alpha_i\},i=2,3,\cdots,n\\ \overline{R}_{\alpha_0}^{\alpha_n}(X)=\bigcup_{i=1}^{n}BN_{\alpha_{i-1}}^{\alpha_i}(X)\cup\underline{R}^{\alpha_n}(X)\end{cases}$$

命题 1)$BN_{\alpha_{i-1}}^{\alpha_i}(X)\cap BN_{\alpha_{j-1}}^{\alpha_j}(X)=\varnothing,i\neq j,\quad i,j=1,2,\cdots,n$;

2)$BN_{\alpha_{i-1}}^{\alpha_i}(X)\cap\underline{R}(X)=\varnothing,i=1,2,\cdots,n$;

3)$BN_R(X)=\bigcup_{i=1}^{n}BN_{\alpha_{i-1}}^{\alpha_i}(X)$。

当 $n=1$,$\alpha_n=1$,$\alpha_0=0$ 时,分级粗糙集模型就是 Pawlak 粗糙集模型。当 $n=1$,$0.5<\alpha_n<1$,$\alpha_0=1-\alpha_n$,则分级粗糙集模型就是变精度粗糙集模型。由命题 1),3)分级粗糙集模型对边界进行了划分。分级粗糙集模型可以解释为对论域中的任意 x,要么 x 属于 X 的程度超过 α_0,要么不超过 a_0,要么介于 α_{i-1} 和 α_i 之间($i=1,2,\cdots,n$)。

3 分级属性约简

设$(U,A\cup\{d\})$是一个不协调决策表,$B\subseteq A$,$R_B=\{(x,y)\mid a(x)=a(y),a\in B\}$,$[x]_B=\{y\mid(x,y)\in R_B\}$,$R_d=\{(x,y)\mid d(x)=d(y)\}$,$U/R_B=\{[x]_B\mid x\in U\}$,$U/R_d=\{[x]_d\mid x\in U\}=\{D_1,D_2,\cdots,D_m\}$,$\mu_B^i(x)=\dfrac{|[x]_B\cap D_i|}{|[x]_B|},i=1,2,\cdots,m$。$\mu_B(x)=(\mu_B^1(x),\mu_B^2(x),\cdots,\mu_B^m(x))$是决策属性 d 在等价

类$[x]_B$上的分布。

记$\mu_{B0}(x)=\max\limits_{1\leqslant i\leqslant m}\{\mu_B^i(x)\}$,它表示$d$在$[x]_B$上的最大分布。设$[0,1]$区间的一个划分$0=a_0<a_1<\cdots<a_n=1$把区间$[0,1]$划分成$[a_0,a_1]$,$(a_1,a_2]$,$\cdots$,$(a_{n-1},a_n]$,则任给$x\in U$,$\mu_{B0}(x)$必落入上面的某个区间,可将这个区间记为$I_{Bx}$。又设$\mu_A^i(x)\in I_{Ax}^i$,记$I_x=(I_{Ax}^1,I_{Ax}^2,\cdots,I_{Ax}^m)$。

定义 3　设$(U,A\cup\{d\})$是一个决策表,$B\subseteq A$,且B非空。

1)对于任意的$x\in U$,若$\mu_{A0}(x)\in I_{Ax}$,总有$\mu_{B0}(x)\in I_{Ax}$,则称B为分级最大分布协调集。若B为分级最大协调集且B的任何非空真子集不是B的分级最大分布协调集,则称B为决策表的分级最大分布约简。

2)对于任意的$x\in U$,若$\mu_A(x)\in I_x$,一定有$\mu_B(x)\in I_x$,则称B是一个分级分布协调集。如果B是一个分级分布协调集且B的任何真子集不是B的分级分布协调集,则B称为分级分布约简。

定理 1　设$(U,A\cup\{d\})$是一个决策表,$B\subseteq A$,且B非空。

1)B是分级最大分布协调集的充分必要条件为:若对于任意的$x,y\in U$,$\mu_{A0}(x)\in I_{Ax}$,$\mu_{A0}(y)\in I_{Ay}$,且$I_{Ax}\neq I_{Ay}$,则$[x]_B\cap[y]_B=\varnothing$。

2)B是分级分布协调集的充分必要条件为:对于任意的$x,y\in U$,$\mu_A(x)\in I_x$,$\mu_A(y)\in I_y$,且$I_x\neq I_y$,则$[x]_B\cap[y]_B=\varnothing$。

证明　只证明 1)。设B是分级最大分布协调集,且$\mu_{A0}(x)\in I_{Ax}$,$\mu_{A0}(y)\in I_{Ay}$,$I_{Ax}\neq I_{Ay}$,则$\mu_{B0}(x)\in I_{Ax}$,$\mu_{B0}(x)\in I_{Ay}$。假如$[x]_B\cap[y]_B=\varnothing$,则$[x]_B=[y]_B$。从而$\mu_{B0}(x)=\mu_{B0}(y)$,故$\mu_{B0}(x)$和$\mu_{B0}(y)$只能属于$I_{Ax}$和$I_{Ay}$中的一个,这同$\mu_{B0}(x)\in I_{Ax}$,$\mu_{B0}(y)\in I_{Ay}$,且$I_{Ax}\neq I_{Ay}$相矛盾。

反过来,假若存在$x\in U$,当$\mu_{A0}(x)\in I_{Ax}$,但是$\mu_{B0}(x)\notin I_{Ax}$,则必有$[x]_A\neq[x]_B$,故一定存在$y\in[x]_B$,使得$[x]_A\neq[y]_A$,设$[x]_B=\bigcup\limits_{y\in[x]_B}[y]_A$,由于

$$\mu_B^i(x)=\frac{|[x]_B\cap D_i|}{|[x]_B|}=\sum_{y\in[x]_B}\frac{|[y]_A|}{|[x]_B|}\frac{|[y]_A\cap D_i|}{|[y]_A|}=\sum_{y\in[x]_B}\frac{|[y]_A|}{|[x]_B|}\mu_A^i(y)$$

假如对所有的$\mu_A^i(y)$都有$\mu_A^i(y)\in I_{Ax}$,则$\mu_B^i(y)\in I_{Ax}$,故必有某个$\mu_A^i(y)\notin I_{Ax}$。设$\mu_A^i(y)\in I_{Ay}$,则$I_{Ax}\neq I_{Ay}$,按假设有$[x]_B\cap[y]_B=\varnothing$,这同$y\in[x]_B$相矛盾。

推论 1　设$B_1\subseteq B_2\subseteq A$,若$B_1$是分级最大分布协调集(分级分布协调集),则$B_2$也是分级最大分布协调集(分级分布协调集)。

定义 4　设$(U,A\cup\{d\})$是一个决策表,记:

$$D_1(x,y)=\begin{cases}\{a\in A\mid a(x)\neq a(y)\}, & \mu_{A0}(x)\in I_{Ax},\mu_{A0}(y)\in I_{Ay}\text{ 且 }I_{Ax}\neq I_{Ay}\\ A, & \text{其他}\end{cases}$$

$$D_2(x,y)=\begin{cases}\{a\in A\mid a(x)\neq a(y)\}, & \mu_{A0}(x)\in I_x,\mu_{Ay}(y)\in I_{Ax}\text{ 且 }I_x\neq I_y\\ A, & \text{其他}\end{cases}$$

分别称$\boldsymbol{D}_i=(D_i(x,y))$,$i=1,2$,为分级最大分布约简和分级分布约简的辨识矩阵。

分别称$M_l=\wedge\{\vee D_l(x,y)\}$,$l=1,2$,为分级最大分布约简和分级分布约简的辨识公式。

定理 2　设$(U,A\cup\{d\})$是一个决策表,M_l,$l=1,2$的极小析取范式为:

$$M_l=\bigwedge_{k=1}^{p_l}\left(\bigvee_{s=1}^{q_k}a_{ls}\right),l=1,2\text{ 。}$$

记$B_{lk}=\{a_{ls}\mid s=1,2,\cdots,q_k\}$,则$\{B_{lk}\mid k=1,2,\cdots,p_l\}$,$l=1,2$分别是分级最大分布约简和分级分布约简的集合。

记$\mathrm{cor}_l=\bigcap\limits_{k=1}^{p_l}B_{lk}$,$l=1,2$,分别称$\mathrm{cor}_l$,$l=1,2$为分级最大分布约简和分级分布约简的核属性集合。

定理 3　1)$a\in\mathrm{cor}_1$当且仅当存在$x\in U$,当$\mu_{A0}(x)\in I_{Ax}$,$\mu_{(A-\{a\})0}(x)\notin I_{Ax}$;

2)$a\in\mathrm{cor}_2$当且仅当存在$x\in U$,当$\mu_A(x)\in I_x$,$\mu_{A-\{a\}}(x)\notin I_x$。

4　结束语

Pawlak 粗糙集模型利用上、下近似近似一个集合,它并没有充分地利用数据所提供的信息。本文利用粗糙隶属函数定义了分级粗糙集和分级约简,并且讨论了分级约简的性质。分级粗糙集推广了 Pawlak 粗糙

集模型和变精度粗糙集模型。

参考文献:

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

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

[3] Kryszkiewcy M. Comparative Study of Alternative Type of Knowledge Reduction in Inconsistent System [J]. Interernational Journal of Intelligent Systems , 2001, 16:105 - 120.

[4] 苗多谦,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):381 - 384.
MIAO Duoqian, HU Guirong. A Heuristic Algorithm for Reduction of Knowledge [J]. Journal of Computer Research & Development, 1999,36(6):381 - 384. (in Chinese)

[5] 刘 清. Rough 集与 Rough 推理[M]. 北京:科学出版社,2001.
LIU Qing. Rough Sets and Rough Reasoning [M]. Beijing: Science Press, 2001. (in Chinese)

[6] 张文修,吴伟志,米据生. 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.
ZHANG Wenxiu , WU Weizhi, MI Jusheng. Knowledge Reductions in Inconsistent Information Systems[J]. Journal of Computers, 2003,26(1): 12 - 18. (in Chinese)

[7] Beynon. Reducts within the Variable Precision Rough Sets Model : A Further Investigation[J]. European Journal of Operational Precision Research,2001,134:592 - 605.

[8] 袁修久,张文修. 变精度粗糙集下约简和一致决策表约简的关系[J]. 模式识别与人工智能,2004,17(2):196 - 200.
YUAN Xiujiu, ZHANG Wenxiu. The Relationship between Attribute Reduction based on Variable Precision Rough Set Model and Attribute Reduction in Consistent Decision Tables[J]. Pattern Recognition and Artificial Intelligence,2004,17(2):196 - 200. (in Chinese)

[9] 袁修久,杨合俊,张小水. 广义决策约简同相对约简的关系[J]. 空军工程大学学报:自然科学版,2005,6(1):44 - 47.
YUAN Xiujiu, YANG Hejun, ZHANG Xiaoshui. Relationships between Generalized Decision Reduction and Relative Reduction[J]. Journal of Air Force Engineering University:Natural Science Edition,2005,6(1):44 - 47. (in Chinese)

[10] 袁修久,张文修. 分布约简与严格下凸函数约简的等价性[J]. 系统工程,2003,21(5):5 - 7.
YUAN Xiujiu, ZHANG Wenxiu. Studies on Equivalence of the Distribution Reduction and the Strictly Convex Function based Reduction in Decision Tables[J]. Systems Engineering, 2003, 21(5):5 - 7(in Chinese)

(编辑:徐楠楠)

Band Rough Set and Band Knowledge Reduction

YUAN Xiu - jiu , GAO Sheng - qiang , YANG Yu
(Science Institute, Air Force Engineering University, Xi'an 710051, China)

Abstract: In Pawlak rough set model, an element is in a set, or not in the set, or possibly in the set. A subset of objects, possibly in the set is called boundary of the set. In the research on Pawlak rough set model, less is paid to the research of the boundary issue. In this paper, two objects with a few minor differences in the degree of membership in a set are viewed in the same level. From this view a partition of the boundary of a rough set is obtained and the concepts of band rough set and band distribution reduction and maximum distribution reduction are presented. The judgment theorems, discernibility matrices, equivalence condition of core attribute association with band distribution reduction and band maximum distribution reduction are given . Band rough set model is a generalization of the Pawlak rough set and of the variable precision rough set model.

Key words: rough set; band rough set; band reduction