

粗糙集的属性正域扩展方法

苗碧舟¹, 周 健²

(1 合肥工业大学 计算机与信息学院, 安徽 合肥 230009; 2 合肥工业大学 信息与网络中心, 安徽 合肥 230009)

摘 要: 针对现有正域扩展方法中存在的不足, 对现有方法进行改进, 通过引入抖动率, 使其能够有效减少扩展了的属性正域里的噪声数据. 实验说明了改进方法的有效性.

关键词: 决策表; 粗糙集; 正域扩展; 属性核; 噪音数据

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2013)11-0047-04

The Method of Extending Positive Region of the Attribute in Rough Set

MIAO Bi-zhou¹, ZHOU Jian²

(1 School of Computer and Information, Hefei University of Technology, Hefei 230009, China;

2 Information and Network Center, Hefei University of Technology, Hefei 230009, China)

Abstract: This paper focuses on shortcomings in the existing positive region extension methods; to improve the existing methods by introducing jitter rate so that it can effectively reduce noise data in the extended positive region of attributes. Finally, the experimental results illuminate the efficiency of the improved method.

Key words: decision table; rough set; positive region extension; attribute core; noise data

1 引言

粗糙集理论中主要的应用价值体现在对数据进行属性约简上. 经典粗糙集理论中属性约简的做法是只选取决策表中正区域的部分, 而全部丢弃边界域中不相容数据. 其实这种方法是不合理的, 因为, 从认知科学角度来讲, 不相容数据的出现不是偶然的, 多少也是客观世界内在规律的体现. 所以, 许多学者从 Pawlak 粗糙集中正区域着手研究, 对其正区域进行扩展, 以期从决策表中挖掘出更有效的潜在的知识.

Zhang^[1]等人定义了参数化的平均支持启发式方法, 其中给出了知道先验概率的扩展方法和不知道先验概率的扩展方法. 对于不知道先验概率的情况, 作者是按照先验概率相等来处理的. 王俊祥, 胡峰^[2]等人把正域分成两部分, 一是经典正区域, 另一个是边界正区域, 通过对边界正区域的定义, 扩展了决策表中原有的正区域的表示.

王俊祥等人提出的方法由于把所有边界域都计算进去, 虽然能扩展正区域, 但是对边界域域扩展过大, 结果可能导致扩展后的正区域中含有很多噪音数据; Ziarko W 提出的方法, 有一个缺点, 就是阈值 β 的选取需要先验知识, 在现实应用中对数据的先验知识的获取往往不精确^[3]. 冯林提出的方法^[4], 完全是根据决策表中的已知的数据, 不需要先验知识, 不需要额外计算原始决策表中的正域与边界域. 但是对其提出的能够减少噪音的提法没有效果. 本文在文献^[4]中方法的基础上, 进行了适当的改进, 使其能够改善对噪音的处理. 并通过实验说明改进方法的有效性.

2 基本理论知识

在粗糙集理论中, 信息表被定义成一个二维数组 $T = (U, A)$, 其中 U 和 A 都是非空有限集合, U 代表论域, A 是属性集合. 其中 A 由条件属性和决策属性两部分组成, 分别用 C 和 D 表示. C 和 D 满足关

系: $C \cup D = A$, 且 $C \cap D = \emptyset$.

定义 1 (不可分辨关系) 设 $P \subset A$ 是一个属性子集, 那么不可分辨关系定义为

$$IND(P) = \{(x, y) \in U \times U: \forall a \in P, a(x) = a(y)\} \quad (1)$$

式中, $a(x)$ 代表对象 x 的属性. 不可分辨关系也称为等价关系.

$U/IND(P)$ 表示在论域 U 中由 P 确定的所有等价关系 $IND(P)$. 等价关系 $U/IND(C)$ 和 $U/IND(D)$ 分别表示条件类和决策类.

对任意 $X \subseteq U$ 以及属性子集 $R \subseteq A$, X 的下近似表示为

$$R_* = \bigcup \{E \in U/IND(R): E \subseteq X\}.$$

X 的上近似表示为

$$R^* = \bigcup \{E \in U/IND(R): E \cap X \neq \emptyset\}.$$

边界域表示为

$$BND_R = R^*(X) - R_*(X).$$

相对于条件属性 C 的决策类 $U/IND(D)$ 的正区域表示为

$$POS_C(D) = \bigcup R_*(X).$$

此正区域即是经典的正区域.

3 正域扩展方法

3.1 多数优先选择方法

基于认知科学的角度, 决策表中的正区域分为 2 个部分^[2]:

(1) 经典的正区域 $POS_C(D)$;

(2) 由边界区域所确定的正区域 $CPBN_C(D)$;

于是决策表中经典正区域被定义为认知正区域 $CPOS_C(D)$, 于是

$$CPOS_C(D) = POS_C(D) \cup CPBN_C(D).$$

相应的边界区域定义为

$$CBN = BN - CPBN_C(D).$$

定义 2 在边界区域 BN 中对于条件属性集 C 的某个分类 $E_i \in U/IND(C)$, 根据决策属性集 D 对 E_i 进行划分, 假设分类为 $E_{i1}, E_{i2}, \dots, E_{ij}$ (j 为分类的个数), 则由 E_i 所确定的正区域为

$$POS(E_i) = E_{r*}: |E_{r*}| = \max\{|E_{ir}|\} \quad (1 \leq k, r \leq j) \quad (2)$$

相应的边界正区域定义为

$$CPBN_C(D) = \bigcup_{i=1}^{|BN/IND(C)|} POS(E_i) \quad (3)$$

本方法能够扩展属性的正域, 但是由于每个边界的分类都要产生一个正区域, 导致扩展后的正

域范围较大, 包含了可能的噪音数据.

3.2 文献[4]中的正域扩展方法

有决策表 $S = (U, C \cup D, V, f)$, 条件属性集和决策属性集分别为 C 和 D , 分类 $E_i \in U/IND(C)$ ($i = 1, 2, \dots, m$) 为条件分类, $X_j \in U/IND(D)$ ($j = 1, 2, \dots, n$) 为决策分类, 则对于任意条件分类 E_i , 对应集合 T_i , 满足:

$$T_i = \{E_i \cap X_j \mid \max |E_i \cap X_j| \text{ and } X_j \in U/IND(D)\} \quad (4)$$

定义 3 (决策表整体确定性和不确定性)^[5] 有决策表 $S = (U, C \cup D, V, f)$, $E_1, E_2, \dots, E_m$ 是所有的条件分类, 那么决策表整体确定性定义为

$$\mu_C = \sum_{i=1}^m (|T_i| / |U|) \quad (5)$$

决策表整体不确定性定义为

$$\mu_C = 1 - \mu_C = 1 - \mu_C = \sum_{i=1}^m (|T_i| / |U|) \quad (6)$$

定义 4 (条件分类对决策分类的确定性程度)^[5] 决策表 $S = (U, C \cup D, V, f)$, 分类 $E_i \in U/IND(C)$ ($i = 1, 2, \dots, m$) 为条件分类, $X_j \in U/IND(D)$ ($j = 1, 2, \dots, n$) 为决策分类, 则对于任意条件分类 $E_i \in U/IND(C)$ 对于决策属性分类的确定性程度定义为

$$k(E_i) = \max\{|E_i \cap E_j| / |E_i| \mid X_j \in U/IND(D)\} = |T_i| / |E_i| \quad (7)$$

定义 5 (决策表局部最小确定性和局部最大不确定性)^[5] 有决策表 $S = (U, C \cup D, V, f)$, $k(E_1), \dots, k(E_i), \dots, k(E_m)$ 是条件分类对决策分类的确定性程度, 则决策表局部最小确定性定义为 $\alpha_C = \min\{k(E_1), \dots, k(E_i), \dots, k(E_m)\}$; 决策表局部最大不确定性定义为

$$\alpha_C = 1 - \alpha_C \quad (8)$$

定义 6 (决策表局部最小确定性扩展法)^[4] 设决策表 $S = (U, C \cup D, V, f)$, 分类 $E_i \in U/IND(C)$ ($i = 1, 2, \dots, m$) 为条件分类, $X_j \in U/IND(D)$ ($j = 1, 2, \dots, n$) 为决策分类, 则 S 中 C 相对于 D 的扩展正域的方法定义如下:

$$CPOS_C(D) = \bigcup_{i=1}^m \{E_i \cap X_j \mid |E_i \cap E_j| / |E_i| \geq \alpha_C \text{ and } X_j \in U/IND(D)\} \quad (9)$$

本方法较多数优先选择方法加快了正域生成的速度, 但在噪音数据处理上不理想.

4 正域扩展的改进方法

本节将对文献[4] 中的方法进行改进,并给出下面的定义.

定义 7 (条件分类对决策表整体不确定性的支持率) 有决策表 $S = (U, C \cup D, V, f), E_1, E_2, \dots, E_m$ 是所有的条件分类,那么条件分类 E_i 对决策表整体不确定性的支持率定义为

$$\mu_{C \cup D} = \sum_{i=1}^m (|T_i| / |U|) \quad (m \neq i) \tag{10}$$

定义 8 (条件分类对决策表整体确定性的抖动率) 给定决策表 $S = (U, C \cup D, V, f), E_1, E_2, \dots, E_m$ 是所有的条件分类,那么条件分类 E_i 对决策表整体不确定性的抖动率定义为

$$p_i = \frac{\mu_C - \mu_{C \cup D}}{\mu_C} \tag{11}$$

式中, p_i 的值越大,说明决策表的整体确定性抖动越大,反之, p_i 的值越小,说明决策表的整体确定性抖动越小. 在本文中,如果 p_i 的值越小,说明其对应的条件分类 E_i 是噪声数据的可能性越大.

定义 9 (正域扩展的改进方法) 设决策表 $S = (U, C \cup D, V, f)$, 分类 $E_i \in U \mid IND(C) \ (i = 1, 2, \dots, m)$ 为条件分类, $X_j \in U \mid IND(D) \ (j = 1, 2, \dots, n)$ 为决策分类,则 S 中 C 相对于 D 的扩展正域的方法定义如下:

$$CPOS_C(D) = \bigcup_{i=1}^m \{E_i \cap E_j \mid |E_i \cap E_j| / |E_i| \geq \alpha_c \text{ and } X_j \in U \mid IND(D) \text{ and } p_i > p\} \tag{12}$$

式中, p 是待设定的抖动阈值.

通过引入抖动阈值,本方法能够有效的抑制正域里面的噪音数据.

5 改进方法的验证

5.1 一个例子

表 1 中不相容决策表有条件属性 a, b, c , 决策属性为 d ; 条件分类为 $E_1, E_2, \dots, E_5$; 决策分类为 $X_1, X_2, \dots, X_6$; E_4 和 E_6 分别为 $E_{4,1}, E_{4,2}, E_{4,3}$ 和 $E_{5,1}, E_{5,2}$.

表 1 中,由于 E_4 在整体数据中所占有的比例很小,所以,可以把 E_4 看作是噪音数据.

分别使用多数优先选择方法、文献[4] 中的正域扩展方法、本文改进的方法以及经典的正域的代表方法来对决策表表 1 进行正域计算. 得到的对比结果如表 2 所示.

表 1 不相容决策表

U	a	b	c	d	X
E_1	0	1	0	1(150px)	X_1
E_2	0	1	2	2(40px)	X_2
E_3	0	2	1	3(300px)	X_3
$E_{4,1}$	3	5	1	3(1px)	X_3
$E_{4,2}$	3	5	1	4(2px)	X_4
$E_{4,3}$	3	5	1	5(2px)	X_5
$E_{5,1}$	2	2	2	6(5px)	X_6
$E_{5,2}$	2	2	2	6(100px)	X_6

表 2 正域结果

正域扩展方法	正域扩展结果
多数优先选择方法	$\{E_1, E_2, E_3, E_{4,2}, E_{5,2}\}$
	$\{E_1, E_2, E_3, E_{4,3}, E_{5,2}\}$
文献[4] 中的正域扩展方法	$\{E_1, E_2, E_3, E_{4,2}, E_{5,2}\}$
	$\{E_1, E_2, E_3, E_{4,3}, E_{5,2}\}$
本文改进的方法(取 $p = 0.01$)	$\{E_1, E_2, E_3, E_{5,2}\}$
经典的正域	$\{E_1, E_2, E_3\}$

从表 2 中可以看出,用经典的正域得出的结果把边界区域的数据都删了,不能体现出数据潜在的规律;用多数优先选择方法以及文献[4] 中的正域扩展方法得出的结果,包含了每个边界域产生的正域,结果把可能的噪音给包含了进来;而用本文的改进方法,通过引入抖动率的概念,能够把可能的噪音数据删除.

5.2 性能实验

扩展决策表属性正域的目的是为了更精确地求得决策表的属性约简. 下面分别应用上述属性扩展方法,并用 Rosetta^[6] 软件对数据集进行约简. 为了验证属性约简的效果,我们用约简后的属性集作为输入,利用台湾大学林智仁教授等开发 LIBSVM^[11] 软件包对数据集进行分类,研究分类的正确率. 其中用数据集的一半数据作为训练集,另一半作为测试集. 每组实验做 10 次,结果取其平均值. 从 UCI 数据库中选择三个数据集 wine、glass 以及 abalone. 其中 wine 数据集中有 178 条数据,13 个条件属性以及 3 个决策;glass 数据集中有 214 条数据,10 个条件属性以及 7 个决策属性;abalone 数据集中有 4177 条数据,8 个条件属性以及 3 个决策属性. 本应用的硬件环境是:操作系统: windows; CPU: 2.39 GHz; 内存: 1.99 GB. 表 3 中给出了约简结果以及

分类正确率, 其中 N 表示约简后的条件属性个数;
 P 代表分类的正确率.

表 3 约简结果与分类正确率

数据集	2.1 节方法		2.2 节方法		改进方法		经典正域	
	P	N	P	N	P	N	P	N
wine	91.01	7	93.76	7	94.38	7	90.83	7
glass	97.26	8	98.13	8	98.19	8	97.21	8
abalone	54.38	6	54.66	6	55.76	6	53.51	6

从表 3 所得的数据可以看出, 利用本文改进的方法来进行决策表属性约简, 进而用约简后的属性进行决策表的分类, 分类的正确性上优于其他的几种方法. 说明了改进方法的有效性.

6 结束语

本文在已有正域扩展的方法上进行了改进, 通过引入抖动率, 来控制扩展的正域含有较多噪声数据的问题. 从表 2 的对比数据中可以看出, 本文改进的方法能够有效的去除正域扩展的过程中产生的噪音数据. 最后通过实验的数据对比说明了改进方法的有效性.

参考文献:

[1] Zhang M, Yao J T. A rough sets based approach to feature selection, fuzzy information[C]// Processing NAFIPS' 04. Canada: Waterloo University, 2004: 434-439.

[2] 王俊祥, 胡峰. 扩展正区域的属性约简方法[J]. 计算机工程与应用, 2008, 44(34): 145-148.

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

[4] 冯林. 一种扩展正域的属性约简方法[J]. 计算机工程, 2010, 36(21): 62-64.

[5] Wang Guoyin, Wang Yan. 3DM: Domain oriented data-driven data mining[J]. Fundamenta Informatices, 2009, 90(4): 395-426.

[6] Chang Chihchung, Lin Chihjen. LIBSVM : a library for support vector machines[EB/OL]. (2012-11-06) [2013-01-10]. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

作者简介:

苗碧舟 男, (1987-), 硕士研究生. 研究方向为网络与通信.

(上接第 46 页)

[7] Marsi S, Impoco G, Ukovich A. Using a recursive rational filter to enhance color images[J]. IEEE Transactions on Instrumentation and Measurement, 2008, 57(6): 1230-1236.

[8] 纪则轩, 陈强. 基于双边滤波的单尺度 Retinex 图像增强算法[J]. 微电子学与计算机, 2009, 26(10): 99-102.

[9] 李学明. 基于 Retinex 理论的图像增强算法[J]. 计算机应用研究, 2005, 22(2): 235-237.

[10] Kimmel R, Elad M, Shaked D, et al. A variational framework for retinex [J]. International Journal of Computer Vision, 2003, 52(1): 7-23.

[11] 戚晓伟, 陈秀宏. 改进的 Otsu 方法的双边滤波边缘检

测算法[J]. 计算机工程与应用, 2012, 48(31): 150-154.

[12] 张志强, 王万玉. 一种改进的双边滤波算法[J]. 中国图像图形学报, 2009, 14(3): 443-447.

作者简介:

刘 松 男, (1987-), 硕士研究生. 研究方向为图像增强、图像匹配等.

孙 东 男, (1982-), 博士研究生, 讲师. 研究方向为图像增强、图像识别等.

卢一相 男, (1980-), 博士研究生, 讲师. 研究方向为图像增强、图像识别等.