

测试代价敏感的粗糙集方法

鞠恒荣^{1,4}, 周献中^{1,2}, 杨佩^{1,3}, 李华雄¹, 杨习贝⁴

(1. 南京大学 工程管理学院, 南京 210093; 2. 南京大学 智能装备新技术研究中心, 南京 210093;
3. 南京大学 软件新技术国家重点实验室, 南京 210023; 4. 江苏科技大学 计算机科学与工程学院, 镇江 212003)

摘要 在粗糙集模型中, α 量化不可分辨关系是强与弱不可分辨关系的推广形式。然而值得注意的是, 基于这三种不可分辨关系的粗糙集并未考虑数据中属性的测试代价。为解决这一问题, 提出了测试代价敏感的 α 量化粗糙集模型, 从二元关系的角度使得粗糙集模型代价敏感, 并将新模型与基于强不可分辨、弱不可分辨以及传统 α 量化不可分辨关系的粗糙集模型进行了对比分析。进一步地, 通过分析传统启发式算法在求解约简的过程中未考虑降低代价这一不足之处, 提出一种新的属性适应性函数, 并将其应用于基于遗传算法的约简求解中。实验结果表明该方法不仅可以降低由边界域所带来的不确定性而且同时降低了约简后的测试代价。

关键词 粗糙集; α 量化不可分辨关系; 测试代价敏感; 属性约简

Test-cost-sensitive based rough set approach

JU Hengrong^{1,4}, ZHOU Xianzhong^{1,2}, YANG Pei^{1,3}, LI Huaxiong¹, YANG Xibei⁴

(1. School of Management and Engineering, Nanjing University, Nanjing 210093, China; 2. Research Center for Novel Technology of Intelligent Equipments, Nanjing University, Nanjing 210093, China; 3. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China; 4. School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract In rough set model, α quantitative indiscernibility relation is a generalization of both strong and weak indiscernibility relations. However, such three indiscernibility relations based rough sets do not take the test costs of the attributes into consideration. To solve this problem, a test-cost-sensitive α quantitative indiscernibility relation based rough set is proposed. From the viewpoint of the binary relation, the new rough set is then sensitive to test costs. Moreover, the relationships among strong, weak, α quantitative and test-cost-sensitive α quantitative indiscernibility relations based rough sets are explored. Finally, it is noticed that the traditional heuristic algorithm does not take the decreasing of cost into account. Therefore, not only a new fitness function is proposed, but also such fitness function is carried out in genetic algorithm for obtaining reduct with minor test cost. The experimental results show that such approach not only decreases the uncertainty comes from boundary region, but also decreases the cost of reduct.

Keywords rough set; α quantitative indiscernibility relation; test-cost-sensitive; attribute reduction

收稿日期: 2015-06-17

作者简介: 鞠恒荣 (1989-), 男, 汉, 江苏泰兴人, 博士研究生, 研究方向: 粒计算, 代价敏感, 粗糙集, E-mail: justjuhengrong@126.com; 通信作者: 周献中 (1962-), 男, 汉, 江苏泰兴人, 教授, 博士生导师, 研究方向: 粗糙集理论, 智能信息处理, 系统工程理论及应用等, E-mail: zhouxz@nju.edu.cn; 杨佩 (1977-), 女, 汉, 江苏南京人, 讲师, 博士, 研究方向: 智能 agent、多agent 系统, E-mail: yangpei@nju.edu.cn; 李华雄 (1977-), 男, 汉, 江西万年人, 讲师, 博士, 研究方向: 粗糙集, 数据挖掘, 机器学习, E-mail: huaxiongli@nju.edu.cn; 杨习贝 (1980-), 男, 回, 江苏镇江人, 副教授, 博士 (后), 研究方向: 粗糙集、粒计算、知识发现, E-mail: zhenjiangyangxibei@163.com.

基金项目: 国家自然科学基金 (61572242, 71671086, 61473157); 江苏省普通高校研究生科研创新计划项目 (KYLX16_0021)

Foundation item: National Natural Science Foundation of China (61572242, 71671086, 61473157); Postgraduate Innovation Foundation of Jiangsu Province of China (KYLX16_0021)

中文引用格式: 鞠恒荣, 周献中, 杨佩, 等. 测试代价敏感的粗糙集方法 [J]. 系统工程理论与实践, 2017, 37(1): 228-240.

英文引用格式: Ju H R, Zhou X Z, Yang P, et al. Test-cost-sensitive based rough set approach[J]. Systems Engineering — Theory & Practice, 2017, 37(1): 228-240.

1 引言

粗糙集理论^[1]是 Pawlak 提出的一种刻画不确定问题的数学工具。在经典的粗糙集方法中, 不可分辨关系占据重要地位, 利用不可分辨关系, Pawlak 给出了下、上近似的概念来刻画目标问题。目前, 粗糙集理论已被广泛应用于模式识别、知识发现、决策支持、人脸识别等众多研究领域^[2-6]。

众所周知, 在人类思维和现实世界里存在十分复杂的信息粒化以及近似逼近模式^[7], 所以将粗糙集方法应用于复杂问题求解就必须拓展粗糙集理论中的相关概念。为此, 国内外众多学者做出了巨大努力, 例如: 为了解决数据的噪声问题, Ziarko^[8]提出了变精度粗糙集; 为了分析有序决策的不一致性, Greco 等人^[9]提出了优势关系粗糙集; 为了应对多源信息处理的需求, Qian 等人^[10]提出了多粒化粗糙集并得到了广泛应用^[11-13]; 将清晰的信息粒化拓展为模糊信息粒化, Dubois 等人研究了模糊粗糙集^[14]方法; 当考虑数据集中存在的缺损值问题时, 很多学者提出了各种类型的拓展二元关系^[15]以替代不可分辨关系; 在文[16]中, Zhao 和 Yao 认为 Pawlak 的不可分辨关系要求过于严格, 为此, 他们提出了弱不可分辨关系以及 α 量化不可分辨关系, 即通过阈值 α 来调节对象间具有相同属性值的个数。

显然, 上述的各种粗糙集模型本身并未考虑数据的代价问题, 但由于代价敏感学习^[17-19]在数据挖掘等领域具有举足轻重的地位, 所以研究基于代价敏感的粗糙集模型对于粗糙集理论的进一步发展是有实际意义的。就粗糙集本身的研究现状来看, 代价大致可分为决策代价和测试代价。例如, 决策理论粗糙集方法^[20-22]就充分考虑了数据中的决策代价。另一方面, “天下没有免费的午餐”, 在现实社会的工程应用中, 数据的获取是需要付出一些成本或代价的, 称其为测试代价, 如在临床系统中, 病人在选择医生和检查的时候就需要考虑就诊费用、时间等众多因素, 而这些就构成了做出一次决策的代价^[23]。针对该现象, Min 等人^[24-26]率先将测试代价引入到粗糙集的约简问题中。然而遗憾的是, Min 等人的研究未能将测试代价引入到粗糙集本身的近似模型上, 这表示粗糙集对于不确定性的刻画依然与测试代价无关。系统工程的观点认为, 数学模型是将某系统的相依关系或逻辑关系, 用形式化的数学语言概括地或近似地表述成一个数学结构。测试代价作为数据的一个内在性质, 在粗糙集数据建模中理应考虑近似刻画与测试代价之间的相依关系和逻辑关系。

为了解决上述问题, 笔者将测试代价引入到 α 量化不可分辨关系中, 提出了测试代价敏感的 α 量化粗糙集模型, 并讨论了基于 Pawlak 不可分辨关系、弱不可分辨关系、 α 量化不可分辨关系的粗糙集模型与新提出模型之间的关系。此外, 因为属性约简是粗糙集理论中研究最为广泛的分支, 所以需进一步考虑测试代价敏感的 α 量化粗糙集的约简问题。在测试代价敏感的 α 量化粗糙集模型中, 每个属性都被赋予了测试代价, 这为属性约简提出了新的要求, 即怎样获得具有相对较小代价的约简。因为传统的启发式算法在求解约简的过程中并未考虑代价因素, 所以最直接的办法就是利用穷举法, 求得数据的所有约简, 然后筛选出其中具有较小测试代价的约简。然而该方法需穷举所有可能的属性子集, 实践表明其计算复杂度过高。同时, 考虑到具有一定阈值 α 的下近似集合为空或者包含的对象很少, 在此情形下保持下近似不发生变化这一度量没有实际意义。为解决这些问题, 本文将属性约简问题转化为优化问题, 设计新的属性适应性函数并采用遗传算法进行问题求解。

2 相关基本概念

2.1 Pawlak 粗糙集

形式化地, 一个信息系统可表示为四元组形如 $IS = \langle U, AT, V, f \rangle$, 其中 $U = \{x_1, x_2, \dots, x_m\}$ 为研究对象的有限集合, 称为论域; AT 为描述对象的全部属性所组成的集合; $V = \cup_{a \in AT} V_a$ 为属性集合 AT 的值域, 其中 V_a 为属性 a 的值域; $f : U \times AT \rightarrow V$ 为信息函数, 表示对每一个 $x \in U, a \in AT, f(x, a) \in V_a$ 。特别地, 当信息系统中属性集 $AT = C \cup D$ 且 $C \cap D = \emptyset$ (其中 C 为条件属性集合, D 为决策属性集合) 时, 信息系统也被称为决策系统。

定义 1 令 IS 为决策系统, $\forall A \subseteq C$, 可定义一个二元关系:

$$IND(A) = \{(x, y) \in U^2 : f(x, a) = f(y, a), \forall a \in A\} \quad (1)$$

称 $IND(A)$ 为由属性集 A 生成的不可分辨关系。

定义 2 令 IS 为决策系统, $\forall A \subseteq C, \forall X \subseteq U, X$ 基于不可分辨关系的下近似集合 $\underline{A}_S(X)$ 与上近似集合 $\overline{A}_S(X)$ 分别定义为:

$$\underline{A}_S(X) = \{x \in U : [x]_A \subseteq X\} \quad (2)$$

$$\overline{A}_S(X) = \{x \in U : [x]_A \cap X \neq \emptyset\} \quad (3)$$

其中 $[x]_A = \{y \in U : (x, y) \in IND(A)\}$ 表示 x 的等价类.

2.2 弱不可分辨关系粗糙集

由定义 1 可以发现论域中任意的两个对象满足不可分辨关系当且仅当这两个对象在所有的属性上都具有相同的属性值, 这是一种非常严格的二元关系, Zhao 与 Yao 将其称为强不可分辨关系^[16]. 在深入研究 Pawlak 不可分辨关系的基础上, Zhao 与 Yao 提出了弱不可分辨关系的概念^[16]. 弱不可分辨关系对要求在所有属性上的取值相等这一条件进行了弱化, 认为任意两个对象只要在一个或一个以上的属性上具有相同的属性值, 那么这两个对象就满足弱不可分辨关系, 具体定义如下:

$$WIND(A) = \{(x, y) \in U^2 : f(x, a) = f(y, a), \exists a \in A\}.$$

显然, 弱不可分辨关系满足自反性、对称性, 因而是一个相容关系.

定义 3 令 IS 为决策系统, $\forall A \subseteq C, \forall X \subseteq U, X$ 基于弱不可分辨关系的下近似集合 $\underline{A}_W(X)$ 与上近似集合 $\overline{A}_W(X)$ 分别定义为:

$$\underline{A}_W(X) = \{x \in U : [x]_A^W \subseteq X\} \quad (4)$$

$$\overline{A}_W(X) = \{x \in U : [x]_A^W \cap X \neq \emptyset\} \quad (5)$$

其中 $[x]_A^W = \{y \in U : (x, y) \in WIND(A)\}$ 表示 U 中所有与 x 具有弱不可分辨关系的对象的合集.

2.3 α 量化粗糙集

强不可分辨关系和弱不可分辨关系分别代表着两种极端情况, 即强不可分辨关系要求在所有属性上的取值都相等, 而弱不可分辨关系则要求在一个或一个以上的属性上具有相同的属性值. 由此可见两者都不能充分说明具有相同属性值的属性数目在不可分辨关系中扮演的重要角色. 为此, Zhao 与 Yao^[16] 从量化的角度对不可分辨关系进行了深入研究.

定义 4 令 IS 为决策系统, $\forall A \subseteq C, \forall (x, y) \in U^2$, 对象之间的不可分辨程度 $ind_A(x, y)$ 定义为:

$$ind_A(x, y) = |\{a \in A : f(x, a) = f(y, a)\}| / |A| \quad (6)$$

其中 $|X|$ 表示集合 X 的基数.

定义 5 令 IS 为决策系统, $\forall A \subseteq C, \alpha$ 量化不可分辨关系定义如下:

$$ind_\alpha(A) = \{(x, y) \in U^2 : ind_A(x, y) \geq \alpha\} \quad (7)$$

其中 $\alpha \in (0, 1]$.

根据以上定义, U 中所有与 x 具有 α 量化不可分辨关系 $ind_\alpha(A)$ 的对象的合集则可表示为 $[x]_A^\alpha$, 即 $[x]_A^\alpha = \{y \in U : (x, y) \in ind_\alpha(A)\}$.

为了便于理解 α 量化不可分辨关系的定义, 文中假定了一个决策表实例, 具体内容如例 1 所示.

例 1 给定一个决策信息表 $IS = < U, C \cup D, V, f >$, 如表 1 所示, 其中 $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ 为论域, $C = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ 为条件属性集, $D = \{d\}$ 为决策属性, 假定 $\alpha = 0.5$, 根据定义 5 可得到 $[x_1]_C^\alpha = \{x_1, x_3, x_4, x_5, x_6\}$, $[x_2]_C^\alpha = \{x_2, x_3\}$, $[x_3]_C^\alpha = \{x_1, x_2, x_3\}$, $[x_4]_C^\alpha = [x_5]_C^\alpha = [x_6]_C^\alpha = \{x_1, x_4, x_5, x_6\}$.

表 1 决策表示例

	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	0	0	0	0	0	1	0
x_2	1	1	1	1	1	1	0
x_3	0	1	0	1	1	1	0
x_4	1	1	0	0	0	0	1
x_5	0	0	0	0	0	0	1
x_6	0	1	1	0	0	0	1

根据 α 量化不可分辨关系, 不难构建基于 α 量化不可分辨关系的粗糙集模型, 其定义如下所示.

定义 6 令 IS 为决策系统, $\forall A \subseteq C, \forall X \subseteq U, X$ 基于 α 量化不可分辨关系的下近似集合 $\underline{A}_\alpha(X)$ 与上近似集合 $\overline{A}_\alpha(X)$ 分别定义为:

$$\underline{A}_\alpha(X) = \{x \in U : [x]_A^\alpha \subseteq X\} \quad (8)$$

$$\overline{A}_\alpha(X) = \{x \in U : [x]_A^\alpha \cap X \neq \emptyset\} \quad (9)$$

$[\underline{A}_\alpha(X), \overline{A}_\alpha(X)]$ 则被称为 X 的 α 量化粗糙集.

3 测试代价与 α 量化粗糙集

上述的 α 不可分辨关系本质上是从条件属性的个数角度进行关系的构造, 这种方法对于没有实际工程背景的数据处理暂且有效. 但是, 在实际工程应用中, 我们获取的数据往往具备一些特殊的、专业的含义, 即对象所拥有的属性及属性值是具有现实意义的. 表 2 给出了一个胃病检查结果的实例决策表, 其中讨论的论域为 6 位医院患者, 条件属性为医院检查项目, 项目分为常规检查和专项检查, 其中 {体温、血白细胞、红血球、尿检} 为常规检查, {钡餐透视、胃镜检查} 为检查是否患有胃病的专项检查. 若根据定义 5, 则患者 P1 和 P5 属于同一类别, 显然这是不合理的, 因为患者 P5 在胃镜检查中成阴性. 为此, 在处理实际工程数据时, 传统的 α 量化不可分辨关系显得无能为力. 与此同时, 上节所涉及到的三种粗糙集模型并未考虑数据的代价问题, 然而在现实工程应用中, 数据的获取并不是免费的. 为了解决这个问题, Min 等人^[24] 将测试代价引入到信息系统中, 具体的描述见定义 7.

表 2 胃病检查实例决策表

	体温	血白细胞	红血球	尿检	钡餐透视	胃镜检查	是否患有胃病
P1	正常	正常	正常	正常	正常	反常	是
P2	偏高	偏多	偏多	反常	反常	反常	是
P3	正常	偏多	正常	反常	反常	反常	是
P4	偏高	偏多	正常	正常	正常	正常	否
P5	正常	正常	正常	正常	正常	正常	否
P6	正常	偏多	偏多	正常	正常	正常	否

3.1 测试代价敏感的 α 量化粗糙集

定义 7 测试代价敏感决策系统 CS 是一个五元组: $CS = < U, C \cup D, V, f, c^* >$, 其中: $U, C \cup D, V$ 和 f 的含义与第一节所示相同, $c^* : C \rightarrow \mathbb{R}^+ \cup \{0\}$ 为测试代价函数 (\mathbb{R}^+ 表示正实数集), 即 $c^*(C) = \sum_{a \in C} c^*(a)$ 其中 $c^*(a)$ 表示单个属性 a 的测试代价.

本文假设所有属性的测试代价都大于 0. 为了在测试代价敏感决策系统中研究对象之间的相似程度, 需引入如下所示的特征函数概念.

定义 8 令 CS 为测试代价敏感决策系统, 其中 $A \subseteq C, \forall x, y \in U, \forall a \in A$, 定义特征函数如下所示:

$$F_a(x, y) = \begin{cases} 1 & : f(x, a) = f(y, a) \\ 0 & : f(x, a) \neq f(y, a) \end{cases}$$

定义 9 令 CS 为测试代价敏感决策系统, $\forall A \subseteq C, \forall (x, y) \in U^2$, 对象之间的不可分辨程度定义为:

$$ind_A^{c^*}(x, y) = \frac{\sum_{a \in A} c^*(a) \cdot F_a(x, y)}{c^*(A)} \quad (10)$$

根据定义 9, 可以定义测试代价敏感决策系统中的 α 量化不可分辨关系, 具体形式如定义 10 所示.

定义 10 令 CS 为测试代价敏感决策系统, $\forall A \subseteq C, \alpha$ 量化不可分辨关系定义如下:

$$ind_\alpha^{c^*}(A) = \{(x, y) \in U^2 : ind_A^{c^*}(x, y) \geq \alpha\} \quad (11)$$

其中 $\alpha \in (0, 1]$.

根据以上定义, 测试代价敏感决策系统中所有与 x 具有 α 量化不可分辨关系的对象的合集则可表示为 $[x]_A^{\alpha, c^*}$, 即 $[x]_A^{\alpha, c^*} = \{y \in U : (x, y) \in ind_\alpha^{c^*}(A)\}$.

定义 9 和 10 将测试代价融合到二元关系的构建中, 这样一种融合机制基于一假定, 即属性的测试代价越大表明该属性越重要. 该假定是合理的, 例如在表 2 中, 专项检查所需要的费用必定远高于常规检查

所需要的费用; 胃镜检查的费用也比钡餐透视的高很多, 因为胃镜检查的结果更准确. 表 3 假定了这几种检测所需要的费用, 根据定义 9 和 10, 则可得到 $[P1]_C^{\alpha,c^*} = [P2]_C^{\alpha,c^*} = [P3]_C^{\alpha,c^*} = \{P1, P2, P3\}$, $[P4]_C^{\alpha,c^*} = [P5]_C^{\alpha,c^*} = [P6]_C^{\alpha,c^*} = \{P4, P5, P6\}$. 在该例中, 虽然患者 P1 检查的常规项目结果都是正常的, 但是在胃镜检查结果却是不理想的, 因此他必须接受医生建议, 平时注意饮食或采取一些治疗.

表 3 胃病检查项目所需的费用

a	体温	血白细胞	红血球	尿检	钡餐透视	胃镜检查
$c^*(a)$	2 ¥	10 ¥	10 ¥	15 ¥	50 ¥	100 ¥

根据定义 10, 可构建测试代价敏感的 α 量化粗糙集模型, 其具体定义如下所示.

定义 11 令 CS 为测试代价敏感决策系统, $\forall A \subseteq C, \forall X \subseteq U, X$ 基于 α 量化不可分辨关系的下近似集合 $\underline{A}_\alpha^{c^*}(X)$ 与上近似集合 $\overline{A}_\alpha^{c^*}(X)$ 分别定义为:

$$\underline{A}_\alpha^{c^*}(X) = \{x \in U : [x]_A^{\alpha,c^*} \subseteq X\} \quad (12)$$

$$\overline{A}_\alpha^{c^*}(X) = \{x \in U : [x]_A^{\alpha,c^*} \cap X \neq \emptyset\} \quad (13)$$

$[\underline{A}_\alpha^{c^*}(X), \overline{A}_\alpha^{c^*}(X)]$ 称为 X 的测试代价敏感的 α 量化粗糙集.

定理 1 令 CS 为测试代价敏感决策系统, $\forall A = \{a_1, a_2, \dots, a_n\} \subseteq C$, 若 $c^*(a_1) = c^*(a_2) = \dots = c^*(a_n)$, 则 $\forall X \subseteq U$, 有:

$$\underline{A}_\alpha^{c^*}(X) = \underline{A}_\alpha(X), \quad \overline{A}_\alpha^{c^*}(X) = \overline{A}_\alpha(X) \quad (14)$$

证明 因为 $c^*(a_1) = c^*(a_2) = \dots = c^*(a_n)$, 不妨设 $c^*(a_1) = c^*(a_2) = \dots = c^*(a_n) = c$, 那么 $\forall x, y \in U$, 根据定义 5 和定义 9 可知:

$$\begin{aligned} y \in [x]_A^{\alpha,c^*} &\Leftrightarrow \text{ind}_A^{c^*}(x, y) \geq \alpha \\ &\Leftrightarrow \frac{\sum_{a \in A} c^*(a) \cdot F_a(x, y)}{c^*(A)} \geq \alpha \\ &\Leftrightarrow \frac{\sum_{a \in A} c \cdot F_a(x, y)}{c \cdot |A|} \geq \alpha \\ &\Leftrightarrow \frac{\sum_{a \in A} F_a(x, y)}{|A|} \geq \alpha \\ &\Leftrightarrow \frac{|\{a \in A : f(x, a) = f(y, a)\}|}{|A|} \geq \alpha \\ &\Leftrightarrow y \in [x]_A^\alpha. \end{aligned}$$

根据上述推导可知 $[x]_A^\alpha = [x]_A^{\alpha,c^*}$, 所以再根据粗糙集的定义, $\underline{A}_\alpha^{c^*}(X) = \underline{A}_\alpha(X)$ 与 $\overline{A}_\alpha^{c^*}(X) = \overline{A}_\alpha(X)$ 显然成立.

定理 1 说明了若测试代价敏感决策系统中所有属性的测试代价都相同, 那么测试代价敏感的 α 量化粗糙集就退化为 Zhao 与 Yao 提出的 α 量化粗糙集.

定理 2 令 CS 为测试代价敏感决策系统, $\forall A \subseteq C, \forall X \subseteq U$, 有:

$$1) \alpha > 0 \Rightarrow \underline{A}_\alpha^{c^*}(X) \supseteq \underline{A}_W(X), \quad \overline{A}_\alpha^{c^*}(X) \subseteq \overline{A}_W(X);$$

$$2) \alpha = 1 \Rightarrow \underline{A}_\alpha^{c^*}(X) = \underline{A}_S(X), \quad \overline{A}_\alpha^{c^*}(X) = \overline{A}_S(X).$$

证明 $\forall x, y \in U$, 因为 $\alpha > 0$ 且本文所讨论的测试代价都假设大于 0, 所以就有

$$\begin{aligned} y \in [x]_A^{\alpha,c^*} &\Rightarrow \text{ind}_A^{c^*}(x, y) \geq \alpha \\ &\Rightarrow \frac{\sum_{a \in A} c^*(a) \cdot F_a(x, y)}{c^*(A)} > 0 \\ &\Rightarrow \sum_{a \in A} F_a(x, y) > 0 \\ &\Rightarrow \exists a \in A, f(x, a) = f(y, a) \\ &\Rightarrow y \in [x]_A^W. \end{aligned}$$

由上述讨论可知 $[x]_A^{\alpha,c^*} \subseteq [x]_A^W$, 所以再根据粗糙集的定义, $\underline{A}_\alpha^{c^*}(X) \supseteq \underline{A}_W(X), \quad \overline{A}_\alpha^{c^*}(X) \subseteq \overline{A}_W(X)$ 显然成立. 类似地, 不难证得结论 2.

由定理2可以看出, 测试代价敏感的 α 量化粗糙集模型是 Pawlak 粗糙集和基于弱不可分辨关系粗糙集的扩展。当 $\alpha > 0$ 时, 测试代价敏感的 α 量化下近似集包含基于弱不可分辨关系粗糙下近似集, 测试代价敏感的 α 量化上近似集包含于基于弱不可分辨关系粗糙上近似集; 当 $\alpha = 1$ 时, 测试代价敏感的 α 量化下/上近似集与 Pawlak 下/上近似集相等。

图1给出了文中所述四种粗糙集模型的格结构, 自底向上表示集合之间存在包含关系。图中每个点代表近似集合或者被近似的目标。经观察, 不难得出如下结论:

1) 测试代价敏感的 α 量化粗糙下近似集与 α 量化粗糙下近似集之间不存在包含关系; 类似地, 测试代价敏感的 α 量化粗糙上近似集与 α 量化粗糙上近似集之间也不存在包含关系; 2) 在近似逼近目标集方面, Pawlak 粗糙集、 α 量化粗糙集和测试代价敏感的 α 量化粗糙集都优于基于弱不可分辨关系的粗糙集模型。

定理3 令 CS 为测试代价敏感决策系统, $\forall A \subseteq C, \forall X \subseteq U$, 若 $0 < \alpha_1 < \alpha_2 \leq 1$, 则有:

$$\underline{A}_{\alpha_1}^{c^*}(X) \subseteq \underline{A}_{\alpha_2}^{c^*}(X), \quad \overline{A}_{\alpha_1}^{c^*}(X) \supseteq \overline{A}_{\alpha_2}^{c^*}(X) \quad (15)$$

证明 $\forall x, y \in U$, 因为 $0 < \alpha_1 < \alpha_2 \leq 1$, 所以有 $y \in [x]_A^{\alpha_2, c^*} \Rightarrow ind_A^{c^*}(x, y) \geq \alpha_2 \Rightarrow ind_A^{c^*}(x, y) \geq \alpha_1 \Rightarrow y \in [x]_A^{\alpha_1, c^*}$, 即 $[x]_A^{\alpha_2, c^*} \subseteq [x]_A^{\alpha_1, c^*}$ 。再根据粗糙集的定义, $\underline{A}_{\alpha_1}^{c^*} \subseteq \underline{A}_{\alpha_2}^{c^*}$ 与 $\overline{A}_{\alpha_1}^{c^*} \supseteq \overline{A}_{\alpha_2}^{c^*}$ 显然成立。

定义12 令 CS 为测试代价敏感决策系统, $\forall A \subseteq C, U/IND(D) = \{X_1, X_2, \dots, X_t\}$ 是由决策属性集 D 诱导出的论域上的划分, 那么近似质量可定义为:

$$\gamma(A, \alpha, D) = |\cup \{\underline{A}_\alpha^{c^*}(X_j) : 1 \leq j \leq t\}|/|U| \quad (16)$$

定理4 令 CS 为测试代价敏感决策系统, $\forall A \subseteq C$, 若 $0 < \alpha_1 < \alpha_2 \leq 1$, 则有:

$$\gamma(A, \alpha_1, D) \leq \gamma(A, \alpha_2, D) \quad (17)$$

证明 根据定理3的结果, 定理4显然成立。

3.2 实验对比

本小节将通过实验, 从近似质量的角度对基于强不可分辨关系、弱不可分辨关系、传统量化不可分辨关系以及测试代价敏感的四种粗糙集模型进行对比分析。

表4列出了实验中使用的6组测试数据的基本信息, 所有数据集均源于UCI数据集。对于每个数据集生成满足泊松分布的10组不同的测试代价。实验运行环境为Windows 7 & Matlab R2012b。

实验结果如表5所示, 在表5中分别由四种粗糙集模型计算6组实验数据集所得到的近似质量, 由于量化粗糙集模型由阈值调节, 因此在本组实验中分别选取了10个不同的 α 值, 并计算在不同 α 值下的近似质量。为了简化表格, 表5将基于强不可分辨关系的粗糙集模型简写为PRS, 基于弱不可分辨关系的粗糙集模型简写为WRS, 在量化粗糙集模型表示中, 本文提出的测试代价敏感的粗糙集模型记为TCS, 传统非测试代价敏感的粗糙集模型记为NTCS。由表5的实验结果不难得出如下结论:

1) 在四种粗糙集模型中, PRS能够得到最大的近似质量; 相反, WRS能得到最小的近似质量。当 $\alpha = 1$ 时, 量化粗糙集得到的近似质量与由Pawlak粗糙集得到的近似质量相同, 该结论与定理2的理论结果是吻合的。2) 在量化粗糙集模型中, 随着阈值 α 的不断增大, 所得到的近似质量也越来越大, 该结论也与定理4的理论结果相吻合。3) 由TCS和NTCS的近似质量对比可发现, 基于TCS的粗糙集获得的近似质量等于或微弱高于NTCS, 由此可见, 将测试代价引入粗糙集模型在一定程度上可以提升模型的近似质量。

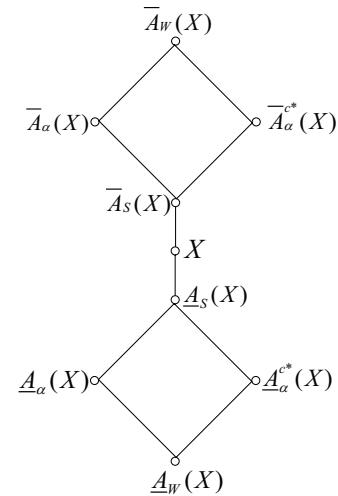


图1 四种粗糙集模型之间的关系

表4 实验数据基本信息				
序号	数据集	样本个数	属性个数	决策类个数
1	Adult	1605	14	2
2	Dermatology	366	34	6
3	Soybean	307	35	4
4	Spect Heart	267	22	2
5	Wdbc	569	30	4
6	Zoo	101	16	7

表 5 四种粗糙集模型的近似质量比较

序号	WRS	不同 α 取值下的 α 量化粗糙集										PRS	
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0		
1	0	TCS	0	0.0006	0.0012	0.0012	0.0050	0.0604	0.2978	0.6517	0.8729	0.9794	0.9794
		NTCS	0	0	0.0012	0.0012	0.0019	0.0561	0.1869	0.6710	0.8611	0.9794	
2	0	TCS	0	0	0	0	0	0.1011	0.4727	0.7951	1.0000	1.0000	1.0000
		NTCS	0	0	0	0	0	0.0820	0.4426	0.8005	0.9891	1.0000	
3	0	TCS	0	0	0	0.0033	0.1173	0.4300	0.7459	0.9674	1.0000	1.0000	1.0000
		NTCS	0	0	0	0.0033	0.0717	0.3388	0.6808	0.9316	1.0000	1.0000	
4	0	TCS	0	0	0	0	0.0112	0.0524	0.1610	0.3408	0.6404	0.8165	0.8165
		NTCS	0	0	0	0	0	0.0599	0.1536	0.3221	0.5843	0.8165	
5	0	TCS	0	0	0	0	0.0158	0.0264	0.2619	0.9824	0.9965	0.9965	0.9965
		NTCS	0	0	0	0	0	0.0228	0.0422	0.8717	0.9965	0.9965	
6	0	TCS	0	0	0	0	0	0.0198	0.2277	0.7624	0.9406	1.0000	1.0000
		NTCS	0	0	0	0	0	0.0099	0.2376	0.6238	0.9406	1.0000	
平均值	0	TCS	0	0.0001	0.0002	0.00075	0.0249	0.1150	0.3612	0.7499	0.9084	0.9654	0.9654
		NTCS	0	0	0.0002	0.00075	0.0245	0.0949	0.2909	0.7035	0.8953	0.9654	

4 属性约简

属性约简是粗糙集理论的主要研究内容之一。传统属性约简的思路是在寻求保持某些度量不变的情况下, 通过删除冗余属性, 来得到简化的数据。然而, 寻找决策表的最小约简已被证明是一个 NP-hard 问题, 在处理大规模数据时计算时间代价很大。针对这一问题, 许多学者提出了许多高效的约简算法^[27], 启发式搜索方法就是其中的一个典型代表。

4.1 启发式算法

考察定义 9 可发现, 在测试代价敏感决策系统中, 本文提出的量化粗糙集模型的近似质量并不一定随着属性集中属性的增多(减少)而单调增加(减小)。为此本节引入近似分布约简的思想用以实现粗糙集模型的下近似分布不发生变化, 进而实现保持近似质量不变的目的。

定义 13 令 CS 为测试代价敏感决策系统, $\alpha \in (0, 1]$, $U/IND(D)$ 是由决策属性集 D 诱导出的论域上的划分, 则基于条件属性集 C 下近似分布集可定义为: $L_\alpha(C) = \{\underline{C}_\alpha^{c^*}(X_1), \underline{C}_\alpha^{c^*}(X_2), \dots, \underline{C}_\alpha^{c^*}(X_t)\}$, 对于任意的属性子集 $A \subseteq C$, A 为测试代价敏感决策系统 CS 的下近似分布约简当且仅当:

- 1) A 为测试代价敏感决策系统 CS 的下近似分布协调集, 即 $L_\alpha(A) = L_\alpha(C)$;
- 2) 对于 A 的任意真子集 A' , A' 不为测试代价敏感决策系统 CS 的下近似分布协调集, 即 $L_\alpha(A') \neq L_\alpha(C)$.

令 CS 为测试代价敏感决策系统, $\alpha \in (0, 1]$, $\forall A \subseteq C$, $\forall a_i \in A$, a_i 的重要度定义为:

$$Sig_{in}(a_i, A) = \frac{\sum_{j=1}^t \{|\underline{A}_\alpha^{c^*}(X_j) \oplus A - \{a_i\}_\alpha^{c^*}(X_j)|\}}{|U|} \quad (18)$$

其中 $X \oplus Y$ 表示集合 X 与集合 Y 的对称差, 由上式可以看出, $Sig_{in}(a_i, A)$ 反映了将 a_i 从当前条件属性集 A 中删除后下近似的变化程度, 相应地, 也可定义

$$Sig_{out}(a_i, A) = \frac{\sum_{j=1}^t \{ |A \cup \{a_i\}_\alpha^{c^*}(X_j) \oplus \underline{A}_\alpha^{c^*}(X_j)| \}}{|U|} \quad (19)$$

其中, $a_i \in C - A$, $Sig_{out}(a_i, A)$ 用以度量向属性集 A 增加属性 a_i 后下近似的变化程度。根据上述属性的重要度可以设计启发式属性约简如算法 1 所示。

算法 1. 基于启发式的属性约简算法 (HAAR).

输入: 测试代价敏感决策系统 CS , α ;

输出: 约简 Red 及其测试代价 $c^*(Red)$.

步骤 1. 计算下近似分布集 $L_\alpha(C)$;

步骤 2. $Red \leftarrow \emptyset$;

- 步骤 3.** $\forall a_i \in C$, 计算属性 a_i 的重要度 $Sig_{in}(a_i, C)$;
- 步骤 4.** 若 a_j 满足 $Sig_{in}(a_j, C) = \max\{\forall a_i \in C : Sig_{in}(a_i, C)\}$, 则 $Red \leftarrow a_j$, 计算 $L_\alpha(Red)$;
- 步骤 5.** 若 $L_\alpha(Red) \neq L_\alpha(C)$, 则重复以下循环, 否则转步骤 6;
- 1) $\forall a_i \in C - Red$, 计算 $Sig_{out}(a_i, Red)$;
 - 2) 若 $Sig_{out}(a_j, Red) = \max\{Sig_{out}(a_i, Red) : a_i \in C - Red\}$, 则 $Red = Red \cup \{a_j\}$;
 - 3) 计算 $L_\alpha(Red)$;
- 步骤 6.** $\forall a_i \in Red$, 若 $L_\alpha(Red - a_i) = L_\alpha(C)$, 则 $Red = Red - a_i$;
- 步骤 7.** 输出 Red 及 $c^*(Red)$.

4.2 遗传优化算法

算法 1 将下近似的变化程度作为衡量属性是否重要的标准. 然而在测试敏感代价决策系统中, 往往希望将具有较小测试代价的属性集作为约简. 虽然笔者提出的粗糙集模型对测试代价敏感, 但在上文的启发式算法过程中并未凸显降低约简测试代价这一目的. 一个最简单的寻求较小测试代价的办法是穷举法, 即求得决策系统的所有约简, 然后挑选出其中具有较小测试代价的约简, 如回溯算法和分辨矩阵方法, 然而这些方法需要穷举决策系统中所有可能的属性子集, 计算复杂度过高, 难以适应于实际工程应用的需要. 针对这一情况, Min 等人采用启发式算法设计了一种基于信息增益的加权约简算法 (IGWAR)^[25]. 此外, 由表 2 可以发现, 当阈值 α 较小时, 近似质量的值为 0 或者很小, 即 α 量化粗糙集模型的下近似为空或者包含的对象很少. 在此情形下, 保持下近似集不发生变化显得意义不大. 顺理成章的, 人们寄希望于寻求一个属性子集使得下近似尽可能增大同时其测试代价较小. 为实现这一目标, 本文将属性约简视为属性优化问题. 从优化角度考虑, 寻求使得下近似尽可能增大并且具有较小测试代价的属性子集. 在优化问题中, 适应性函数发挥着关键作用, 适应性函数设计的好与坏将会直接影响优化的效果. 对于条件属性集 C 的任意子集 A , 本文定义适应性函数如下:

$$Fit(A) = \gamma(A, \alpha, D) + \frac{c^*(C) - c^*(A)}{c^*(C)} \quad (20)$$

由式 (20) 可知, 该适应性函数同时兼顾了下近似集以及测试代价, 求解满足条件的属性子集即可转化为求使 $Fit(A)$ 最大的属性集 A .

本文选取遗传算法实现优化问题, 在遗传算法中, 每条染色体由长度为 $|C|$ 的二进制序列表示, 其中“1”表示相应的属性存在于该染色体中, 相反“0”表示该属性不在该染色体中. 基于遗传算法的属性约简算法如算法 2 所示:

算法 2. 基于遗传优化的属性约简算法 (GAAR).

输入: 测试代价敏感决策系统 CS, α ;

输出: 约简及其测试代价.

- 步骤 1.** 初始化: 产生最初种群数和最大演化代数;
- 步骤 2.** 计算适应性函数, 对适应性函数进行评估, 若满足停止条件, 转第 6 步, 否则执行第 3 步;
- 步骤 3.** 选择: 采用转盘赌选择方式选择对象产生下一代;
- 步骤 4.** 交叉与变异: 对生成的新一代进行交叉操作和变异操作;
- 步骤 5.** 评估新一代的适应性函数;
- 步骤 6.** 从当前种群中选择最优的约简及其代价.

4.3 评价指标

为了直观地对算法的效果进行比较, 须制定客观合理的算法效果评价指标. 对于测试代价敏感粗糙集模型的属性约简而言, 决策规则和测试代价是两个重要的方面. 决策规则与粗糙集模型的下、上近似集合紧密联系而测试代价则与条件属性关联. 本文将从这两个方面出发, 制定如下四种评价指标.

- 1) 规则变化指标 (RC)

在粗糙集理论中, 根据目标的下、上近似集, 可以将论域划分为三个互不相交的区域, 即正域、边界域和

负域. $\forall X \subseteq U, \forall A \subseteq C$, 由测试代价敏感的 α 量化粗糙集模型所诱导的三个区域可分别表示为:

$$POS_\alpha(A, X) = \underline{A}_\alpha^{c^*}(X) \quad (21)$$

$$BND_\alpha(A, X) = \overline{A}_\alpha^{c^*}(X) - \underline{A}_\alpha^{c^*}(X) \quad (22)$$

$$NEG_\alpha(A, X) = U - \overline{A}_\alpha^{c^*}(X) \quad (23)$$

由于遗传算法得到的约简致力于寻求最大的近似质量, 而近似质量是由下近似决定的, 因此应考察采用不同约简算法约简后的正域、边界域和负域的变化, 分别记为 PRC_α 、 BRC_α 和 NRC_α . 假设属性集 $A \subseteq C$ 由为算法得到的约简, $U/IND(D) = \{X_1, X_2, \dots, X_t\}$ 为决策属性诱导出的划分, 则:

$$PRC_\alpha = \frac{\sum_{j=1}^t (|POS_\alpha(A, X_j)| - |POS_\alpha(C, X_j)|)}{|U|} \quad (24)$$

$$BRC_\alpha = \frac{\sum_{j=1}^t (|BND_\alpha(A, X_j)| - |BND_\alpha(C, X_j)|)}{|U|} \quad (25)$$

$$NRC_\alpha = \frac{\sum_{j=1}^t (|NEG_\alpha(A, X_j)| - |NEG_\alpha(C, X_j)|)}{|U|} \quad (26)$$

如上定义的 PRC_α 、 BRC_α 和 NRC_α 即规则变化指标 RC 的具体体现. $PRC_\alpha > 0$ 表明约简后数据集生成的正域规则数多于原始数据的正域规则; $PRC_\alpha = 0$ 表明约简后数据集生成的正域规则与原始数据的正域规则相同; $PRC_\alpha < 0$ 则表明约简后数据集生成的正域规则数少于原始数据的正域规则. 类似的解释适用于 BRC_α 和 NRC_α .

2) 约简测试代价指标 (RTC)

测试代价是本文的核心概念, 由于其定量化, 可直接采取测试代价为一种评价指标. 对于一个有测试代价的数据集, 假设属性子集 A 为一算法进行属性约简工作后得到的约简, 则可定义约简测试代价指标为:

$$RTC = c^*(A) = \sum_{a \in A} c^*(a) \quad (27)$$

3) 近似质量测试代价综合指标 ($AQTC$)

本文 3.2 节定义的优化适应性函数融合了近似质量和测试代价, 以达到在降低约简测试代价的情况下尽可能提高近似质量. 这一情况极可能会出现测试代价降低的很少, 近似质量却减少很多的极端情况. 为此, 必须制定体现近似质量和测试代价的综合指标. 假设属性子集 $A \subseteq C$ 为一算法进行属性约简工作后得到的约简, 则可定义 $AQTC$ 指标如下:

$$AQTC = \gamma(A, \alpha, D) - \gamma(C, \alpha, D) + \frac{c^*(C) - c^*(A)}{c^*(C)} \quad (28)$$

考察该定义, 一方面, 由于 $\gamma(C, \alpha, D) \in [0, 1]$, $\gamma(A, \alpha, D) \in [0, 1]$, 那么 $(\gamma(A, \alpha, D) - \gamma(C, \alpha, D)) \in [-1, 1]$; 另一方面, $\frac{c^*(C) - c^*(A)}{c^*(C)} \in [0, 1]$ 显然成立. 因此可得到 $AQTC \in [-1, 2]$. $AQTC < 0$ 表示近似质量损失了很多, 算法效果不好; $AQTC > 0$ 表示算法总体达到了降低测试代价和提高近似质量这一综合目标.

4) 约简长度指标 (RL)

评价约简算法的效果也可从约简的长度进行考虑. 一般而言, 约简长度指数据集在经过约简工作后属性子集包含的属性个数. 假设属性子集 $A \subseteq C$ 为一算法进行属性约简工作后得到的约简, 则可定义 RL 指标如下:

$$RL = |A| \quad (29)$$

4.4 实验分析

本节将通过实验对比分析本文提出的启发式算法 (HAAR) 和遗传算法 (GAAR), 为了说明本文算法的有效性, 本文基于测试代价敏感粗糙集模型复现了 Min 等人提出的 IGWAR 算法并采用评价指标进行了对比分析. 本次实验同样选取表 3 给出的 6 组数据集进行计算, 对于每个数据集生成满足泊松分布的 10 组不同的测试代价.

表 6 至表 11 列出了 6 组数据的实验结果. 每一个表对应一个数据集在 10 组不同测试代价下的平均实验结果. 为了避免经验主义, 在每张表中分别计算了该数据集在 10 组不同的阈值下得到约简的正域、边界域、负域的变化, 以及约简长度, 即 PRC_α 、 BRC_α 、 NRC_α 和 RL 指标值. 考察文中所示表可得到如下结论:

表6 约简RC和RL指标的比较(Adult数据)

α	PRC_α			BR_α			NRC_α			RL		
	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR
0.1	0	0.034	0	0	-0.068	0	0	0.034	0	1	1	2
0.2	0	0.023	0	0	-0.446	0	0	0.423	0	3.5	1.8	3.8
0.3	0	0.030	0	0	-0.261	0	0	0.231	0	8.1	1.9	8.6
0.4	0	0.026	0	0	-0.252	0	0	0.223	0	7	1.3	7.3
0.5	0	0.020	0	0	-0.441	0	0	0.420	0	9.1	2	10.1
0.6	0	0.117	0	0	-0.234	0	0	0.117	0	12.4	2.1	13.4
0.7	0	0.079	0	0	-0.158	0	0	0.079	0	14	2.8	13.9
0.8	0	0.053	0	0	-0.106	0	0	0.053	0	13.9	4.1	13.9
0.9	0	0.024	0	0	-0.048	0	0	0.024	0	13.7	6.1	13.9
1.0	0	-0.109	0	0	0.217	0	0	-0.109	0	10	5.9	10

表7 约简RC和RL指标的比较(Dermatology数据)

α	PRC_α			BR_α			NRC_α			RL		
	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR
0.1	0	0.096	0	0	-0.193	0	0	0.096	0	1	1.3	1
0.2	0	0.107	0	0	-0.214	0	0	0.107	0	1	1.4	1.2
0.3	0	0.137	0	0	-0.273	0	0	0.137	0	1	1.2	1.3
0.4	0	0.140	0	0	-0.280	0	0	0.140	0	1	1.6	1
0.5	0	0.134	0	0	-0.267	0	0	0.134	0	7.1	2	7.1
0.6	0	0.097	0	0	-0.194	0	0	0.097	0	13.6	2	15.8
0.7	0	0.028	0	0	-0.056	0	0	0.028	0	16.8	2	16
0.8	0	-0.042	0	0	0.084	0	0	-0.042	0	17.5	4.3	17.9
0.9	0	-0.089	0	0	0.177	0	0	-0.089	0	19.3	7.2	20.9
1.0	0	-0.154	0	0	0.317	0	0	-0.158	0	20	9.7	19.1

表8 约简RC和RL指标的比较(Soybean数据)

α	PRC_α			BR_α			NRC_α			RL		
	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR
0.1	0	0.300	0	-0.978	-1.688	-1.124	0.978	1.388	1.124	2	2.4	1.8
0.2	0	0.303	0	-0.904	-1.671	-1.085	0.904	1.368	1.085	2	2.6	2
0.3	0	0.366	0	-0.561	-1.483	-0.539	0.561	1.117	0.539	2	2.9	2.2
0.4	0	0.279	0	0.260	-0.816	-0.100	-0.260	0.537	0.100	4.3	4.8	7.1
0.5	0	0.319	0	0.169	-0.776	0.141	-0.169	0.456	-0.141	19.6	6	20.4
0.6	0	0.526	0	-0.072	-1.448	-0.068	0.072	0.922	0.068	26.1	6.2	26.5
0.7	0	0.261	0	-0.006	-0.669	-0.005	0.006	0.409	0.005	28.8	7.1	31.5
0.8	0	0.025	0	0	-0.048	0	0	0.023	0	28.6	5.1	26.6
0.9	0	-0.016	0	0	0.039	0	0	-0.023	0	16	5.2	11.8
1.0	0	-0.014	0	0	0.029	0	0	-0.015	0	4	5.4	4.4

表9 约简RC和RL指标的比较(Spect Heart数据)

α	PRC_α			BR_α			NRC_α			RL		
	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR
0.1	0	0.096	0	0	-0.193	0	0	0.096	0	1	1.3	1
0.2	0	0.107	0	0	-0.214	0	0	0.107	0	1	1.4	1.2
0.3	0	0.137	0	0	-0.273	0	0	0.137	0	1	1.2	1.3
0.4	0	0.140	0	0	-0.280	0	0	0.140	0	1	1.6	1
0.5	0	0.134	0	0	-0.267	0	0	0.134	0	7.1	2	7.1
0.6	0	0.097	0	0	-0.194	0	0	0.097	0	13.6	2	15.8
0.7	0	0.028	0	0	-0.056	0	0	0.028	0	16.8	2	16
0.8	0	-0.042	0	0	0.084	0	0	-0.042	0	17.5	4.3	17.9
0.9	0	-0.089	0	0	0.177	0	0	-0.089	0	19.3	7.2	20.9
1.0	0	-0.154	0	0	0.317	0	0	-0.158	0	20	9.7	19.1

表 10 约简 RC 和 RL 指标的比较 (Wdbc 数据)

α	PRC_α			$BR C_\alpha$			NRC_α			RL		
	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR
0.1	0	0.619	0	0	-1.244	0	0	0.624	0	2.2	1.9	2.3
0.2	0	0.511	0	0	-1.026	0	0	0.515	0	2	2.3	1.6
0.3	0	0.701	0	0	-1.407	0	0	0.706	0	2.3	2.8	2.2
0.4	0	0.714	0	0	-1.436	0	0	0.721	0	2.3	3.8	2.5
0.5	0	0.878	0	0	-1.763	0	0	0.885	0	10.2	4.8	11.2
0.6	0	0.952	0	0	-1.911	0	0	0.959	0	20.8	4.6	18.6
0.7	0	0.752	0	0	-1.507	0	0	0.755	0	28.9	3.9	29.1
0.8	0	0.146	0	0	-0.294	0	0	0.147	0	26.9	3.2	25.3
0.9	0	-0.012	0	0	0.024	0	0	-0.012	0	2.4	3.5	3.1
1.0	0	-0.008	0	0	0.016	0	0	-0.008	0	3	3.4	2.7

表 11 约简 RC 和 RL 指标的比较 (Zoo 数据)

α	PRC_α			$BR C_\alpha$			NRC_α			RL		
	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR	HAAR	GAAR	IGWAR
0.1	0	0.385	0	-2.128	-3.311	-1.447	2.128	2.926	1.447	1	1.2	1
0.2	0	0.385	0	-2.128	-3.311	-1.751	2.128	2.926	1.751	1	1.2	1.5
0.3	0	0.385	0	-2.127	-3.309	-1.352	2.127	2.924	1.352	1	1.4	1.8
0.4	0	0.365	0	-1.998	-3.022	-1.463	1.998	2.656	1.462	1	1.2	1.6
0.5	0	0.385	0	-1.412	-2.594	-0.946	1.412	2.209	0.946	1	1.1	1.4
0.6	0	0.536	0	-0.005	-2.685	0.077	0.005	2.149	-0.077	10.5	2.3	10.1
0.7	0	0.484	0	0.032	-1.760	0.055	-0.033	1.276	-0.055	14.2	2.6	13.9
0.8	0	0.196	0	0.074	-0.354	0.084	-0.074	0.158	-0.084	14.4	3.1	13.8
0.9	0	0.022	0	-0.001	-0.038	0.051	0.001	0.017	-0.051	11.5	4.9	10.7
1.0	0	0.022	0	0	0.044	0	0	-0.022	0	6	5.3	6

- 由 PRC_α 的结果可以发现, HAAR 算法和 IGWAR 算法下的约简 PRC_α 值均为 0, 这表明 HAAR 算法和 IGWAR 算法下的约简生成的正域决策与原始数据集保持一致. 在 Zoo 数据集上, GAAR 算法下的约简 PRC_α 值和 NRC_α 值均大于 0, 当 $\alpha < 1$ 时, 其他 5 组数据集上 GAAR 下的约简 PRC_α 值也均大于 0. 这表明 GAAR 算法使得数据的正域得以增大.
- 由 $BR C_\alpha$ 的结果可以发现, GAAR 算法的 $BR C_\alpha$ 值在多数情况下都小于 0, 这表明 GAAR 算法可以使边界域相对减少; 而 HAAR 算法和 IGWAR 算法下的约简的 $BR C_\alpha$ 值在有的数据集上小于 0, 在有的数据集上等于 0. 此现象说明 GAAR 算法在降低由边界域造成的不确定性方面优于其他两种算法.
- 综合 PRC_α 、 $BR C_\alpha$ 和 NRC_α 值可发现, 正域和负域增加的值等于边界域减少的值, 即 $PRC_\alpha + NRC_\alpha + BR C_\alpha = 0$. 此现象表明正域和负域增大的部分均来自于边界域, 换言之, 基于本文定义的属性适应性函数下的遗传算法压缩了边界域, 使原先一些不可确定属于哪一类的对象得到正确的分类. 同时, 应当注意到在遗传算法中, 当 $\alpha = 1$ 时, 正域和负域的一些对象落入了边界域中, 这是由于在该阈值下遗传算法陷入了局部最优, 无法得到最大的正域值, 此时可利用启发式算法解决该问题.
- 从约简长度角度来看, 当阈值较小时, GAAR 算法的 RL 值稍大于其他两种算法, 但随着近似质量的增加, GAAR 算法的 RL 值则远小于其他两种算法.

图 2 列出了三种算法得到的约简测试代价比较, 由图 2 的实验结果可发现, 在一些数据集的实验结果中, 当阈值较小时, GAAR 算法得到的 RTC 值即约简测试代价稍大于其他两种算法, 而当阈值设置的较大时, GAAR 算法得到的约简的 RTC 值远小于其他两种算法. 此现象与约简长度的实验结果类似. 这种现象是由于数据本身的结构导致的. 由表 5 可知, 当阈值较小时, 很多数据集的正域为空集, 为保持下近似不发生变化, 对于启发式算法而言, 任何对应正域为空集的单个属性都可以作为最终的约简, 这就导致约简中只有单个属性从而降低了约简的测试代价. 在这一情况下, 虽遗传算法得到的约简代价和长度稍大于启发式算法的约简, 但其正域得到了增大, 这就意味着决策者获得了更多的正规则. 因此, GAAR 算法得到的结果是更有意义的.

表 12 列出了 6 组数据在 GAAR 算法下的 $AQTC$ 指标值, 显示 6 组数据在 10 个不同的阈值下分别得

到的 AQTC 值均为正值, 这表明 GAAR 算法总体达到了降低测试代价和提高近似质量这一综合目标.

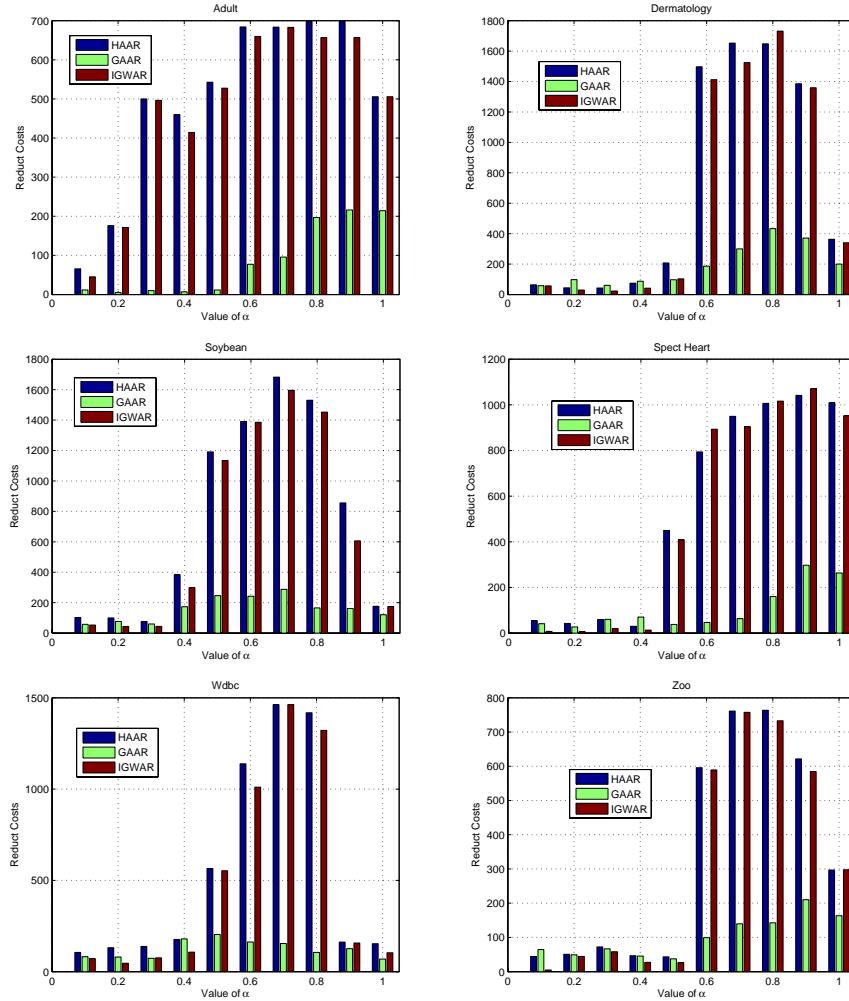


图 2 约简的测试代价比较

表 12 GAAR 算法 AQTC 指标值

数据集	α									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Adult	+	+	+	+	+	+	+	+	+	+
Dermatology	+	+	+	+	+	+	+	+	+	+
Soybean	+	+	+	+	+	+	+	+	+	+
Spect Heart	+	+	+	+	+	+	+	+	+	+
Wdbc	+	+	+	+	+	+	+	+	+	+
Zoo	+	+	+	+	+	+	+	+	+	+

5 结论

根据应用需求, 将经典粗糙集模型进行扩展对于粗糙集理论的发展具有重要的现实意义. 本文考虑数据的测试代价, 提出了基于测试代价敏感的量化粗糙集模型, 并分别从理论和实验角度, 将其与基于强不可分辨关系、弱不可分辨关系和传统 α 量化不可分辨关系的粗糙集模型进行了对比分析. 进一步地, 通过分析传统启发式约简算法未考虑降低属性测试代价以及追求保持下近似不发生变化这两点不足之处, 本文将传统属性约简问题转化为获取具有较小测试代价和较大下近似集的优化问题, 提出了一种属性适应性函数, 并通过遗传优化算法对其进行验证. 实验结果表明, 基于该适应性函数得到的约简不仅降低了由边界域所带来的不确定性, 同时亦降低了约简的代价和约简的长度. 在本文研究工作的基础上, 下一步的研究重点将集中在以下三个方面: 1) 测试代价敏感粗糙集模型中阈值的学习机制; 2) 在综合考虑数据的错分类代价以及测试

代价的基础上设计高效的属性约简算法; 3) 基于 α 量化不可分辨关系的粗糙分类器的设计.

参考文献

- [1] Pawlak Z. Rough sets-theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic, 1991.
- [2] Park I K, Choi G S. Rough set approach for clustering categorical data using information-theoretic dependency measure[J]. Information Systems, 2015, 48: 289–295.
- [3] Hu Q H, Che X J, Zhang L, et al. Rank entropy based decision trees for monotonic classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(11): 2052–2064.
- [4] Guo Y G, Jiao L C, Wang S, et al. A novel dynamic rough subspace based selective ensemble[J]. Pattern Recognition, 2015, 48: 1638–1652.
- [5] 黄兵, 魏大宽. 基于距离的直觉模糊粗糙集模型及应用 [J]. 系统工程理论与实践, 2011, 31(7): 1356–1362.
Huang B, Wei D K. Distance-based rough set model in intuitionistic fuzzy information systems and its application[J]. Systems Engineering — Theory & Practice, 2011, 31(7): 1356–1362.
- [6] Li H X, Zhang L B, Huang B, et al. Sequential three-way decision and granulation for cost-sensitive face recognition[J]. Knowledge-Based Systems, 2016, 91: 241–251.
- [7] 胡清华, 于达任. 应用粗糙计算 [M]. 北京: 科学出版社, 2012.
Hu Q H, Yu D R. Applied rough computing[M]. Beijing: Science Press, 2012.
- [8] Ziarko W. Variable precision rough set model[J]. Journal of Computer and System Science, 1993, 46(1): 39–59.
- [9] Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis[J]. European Journal of Operational Research, 2002, 129(1): 1–47.
- [10] Qian Y H, Zhang H, Sang Y L, et al. Multigranulation decision-theoretic rough sets[J]. International Journal of Approximate Reasoning, 2013, 55: 225–237.
- [11] Ju H R, Yang X B, Dou H L, et al. Variable precision multigranulation rough set and attributes reduction[C]// Transactions on Rough Set XVIII, Springer, 2014: 52–68.
- [12] Ju H R, Yang X B, Qi Y S, et al. Dynamic updating multigranulation fuzzy rough set: Approximations and reducts[J]. International Journal of Machine Learning and Cybernetics, 2014, 5(6): 981–990.
- [13] Yang X B, Qi Y, Yu H L, et al. Updating multigranulation rough approximations with increasing of granular structures[J]. Knowledge-Based Systems, 2014, 64: 59–69.
- [14] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems, 1990, 17(2–3): 191–209.
- [15] Yang X B, Yang J Y. Incomplete information system and rough set theory: Models and attribute reductions[M]. Beijing: Science Press & Springer, 2012.
- [16] Zhao Y, Yao Y Y, Luo F. Data analysis based on discernibility and indiscernibility[J]. Information Sciences, 2007, 177(22): 4959–4976.
- [17] Palacios A M, Sánchez L, Couso I. Linguistic cost-sensitive learning of genetic fuzzy classifiers for imprecise data[J]. International Journal of Approximate Reasoning, 2011, 52(6): 841–862.
- [18] Yang Q, Ling C, Chai X Y, et al. Test-cost sensitive classification on data with missing values[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(5): 626–638.
- [19] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63–77.
- [20] Yao Y Y. The superiority of three-way decisions in probabilistic rough set models[J]. Information Sciences, 2011, 181: 1080–1096.
- [21] Jia X Y, Liao W H, Tang Z M, et al. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2013, 219: 151–167.
- [22] Li W T, Xu W H. Double-quantitative decision-theoretic rough set[J]. Information Sciences, 2015, 316: 54–67.
- [23] Ju H R, Yang X B, Yu H L, et al. Cost-sensitive rough set approach[J]. Information Sciences, 2016, 355–356: 282–298.
- [24] Min F, He H P, Qian Y H, et al. Test-cost-sensitive attribute reduction[J]. Information Sciences, 2011, 181(22): 4928–4942.
- [25] Min F, Zhu W. Attribute reduction of data with error ranges and test costs[J]. Information Sciences, 2012, 211: 48–67.
- [26] Min F, Hu Q H, Zhu, W. Feature selection with test cost constraint[J]. International Journal of Approximate Reasoning, 2014, 55(1): 167–179.
- [27] 张显勇. 精度与程度逻辑差双量化粗糙集模型的属性约简 [J]. 系统工程理论与实践, 2015, 35(11): 2925–2931.
Zhang X Y. Attribute reduction for the double-quantitative rough set model based on logical difference of precision and grade[J]. Systems Engineering — Theory & Practice, 2015, 35(11): 2925–2931.