文章编号: 1001-0920 (2016) 07-1199-07 DOI: 10.13195/j.kzyjc.2015.0656

基于属性重要度的风险决策粗糙集属性约简

张清华[1a], 胡荣德[1a,2], 姚龙洋[1b], 谢万成[1a]

(1. 重庆邮电大学 a. 理学院，b. 计算机科学与技术学院，重庆 400065；2. 北京邮电大学 计算机学院，北京 100876)

摘 要: 基于 Pawlak 粗糙集的属性约简一般保持决策表的正区域不变, 然而由于现实中不同用户对不同约简精度的需求, 获取属性值的实际代价与个人偏好可能不同. 针对决策者主观个人偏好、客观约简精度、获取属性值的实际代价和决策表各区域的误判代价等综合情况, 提出新的约简算法, 并讨论约简代价与约简精度间的关系. 通过遗传算法, 采用启发式方法搜索出局部最优约简子集. 仿真实验表明, 所提出的算法操作性强, 更适合处理实际决策问题.

关键词: 决策粗糙集；属性重要度；代价函数；用户偏好；属性约简

中图分类号: TP18 **文献标志码:** A

Risk DTRS attribute reduction based on attribute importance

ZHANG Qing-hua[1a], *HU Rong-de*[1a,2], *YAO Long-yang*[1b], *XIE Wan-cheng*[1a]

(1a. School of Science，1b. School of Computer Science and Technology，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；2. School of Computer Science，Beijing University of Posts and Telecommunications，Beijing 100876，China．Correspondent：ZHANG Qing-hua，E-mail：zhangqh@cqupt.edu.cn)

Abstract: Generally, when talking about attribute reduction of a decision table, it usually keeps the positive region unchanged based on the Pawlak's rough sets theory. However, the needs may be different for different precision of the reduction in real life as well as the actual cost to obtain attribute values and personal preferences. Based on the risk of personal preference for the subjective aspect, the accuracy of reduction, the actual cost of obtaining attribute value, and the risk of interval misjudgment for the objective aspects, a novel attribute reduction algorithm is proposed. Then, the relationship between the reduction cost and the reduction accuracy is discussed. Based on the genetic algorithm, a heuristic method for searching the local optimal reduction subset is proposed. Simulation experiments show that the algorithm is feasible, and more realistic to deal with practical decision-making problems.

Keywords: decision-theoretic rough set；attribute importance；cost function；user preferences；attribute reduction

0 引 言

粗糙集 (RS) 理论在 20 世纪 80 年代由波兰科学家 Pawlak[1]提出, 是一种新的处理不精确、不完全和不相容知识的数学工具. 目前, 该理论已广泛应用于数据挖掘、模式识别、人工智能和决策支持与分析等众多领域[2]. 经典粗糙集理论是通过定义上、下近似集来描述数据的不确定性, 但过于严格的代数关系导致正区域决策规则对知识分类具有敏感性, 从而使其缺乏容错能力, 因此学者们相继提出了一系列概率粗糙集模型. 在此背景下, Yao 等[3]引入贝叶斯决策相关理论, 提出了决策粗糙集模型. 通过对粗糙集理论中正区域、负区域和边界域的语义分析, 及决策者给出的风险代价值来确定最小风险原则下的概率阈值[4].

属性约简是粗糙集理论重点研究内容之一, 现有的属性约简大致可分为代数观角度和信息观角度两大类[5]. 从代数观角度看, 主要有基于正区域的属性约简算法[6]和基于差别矩阵及其改进的属性约简算法[7-8]; 从信息观角度看, 可分为基于信息熵的属性约简算法[9]、基于互信息的决策表属性约简算法[10]和基于条件信息熵的决策表约简算法[11]等. 当然, 也有与群智能算法相结合的属性约简算法[12]. 在决策粗糙集下, Zhao 等和 Yao 等[13-14]提出了保持决策表正区域不变的属性约简算法, 并在此基础上, 从三支决策角度提出了保持决策不变和当决策发生变化时的属性约简[14], 从而使属性约简得到泛化.

然而, 决策粗糙集与经典粗糙集的不同之处在

收稿日期: 2015-05-25；**修回日期:** 2015-12-04.

基金项目: 国家自然科学基金项目(61472056)；大学生科研训练计划项目(A2014-45).

作者简介: 张清华(1974−), 男, 教授, 博士, 从事不确定信息处理、粗糙集与粒计算等研究；胡荣德(1992−), 男, 硕士生, 从事物联网、多媒体信息处理与不确定信息处理的研究.

于, 决策区域和决策规则与属性增减之间并不具备单调性, 且约简后决策表的各区域可能发生变化, 从而导致一定的潜在风险. 贾修一等[15]讨论了基于最小决策代价的属性约简算法; 于洪等[16]以此为基础, 提出基于属性重要度的决策风险最小化的属性约简算法, 综合考虑了属性区间误判和属性重要度两方面因素, 并以此为启发式信息进行约简.

实际生活中, 由于所讨论对象涉及的论域不尽相同, 噪声的普遍性等因素, 造成条件属性形成的等价类完全包含于决策类的知识较少. 另外, 考虑到现实生活中决策者主观个人偏好以及客观约简精度, 获取属性值的实际代价, 以及决策表各区域的误判产生的代价等情况, 本文提出一种基于属性重要度的风险决策粗糙集属性约简算法, 从决策优化的角度解决问题, 并通过遗传算法搜索出局部最优约简子集来逼近全局最优属性约简子集. 最后, 进一步讨论了约简代价与约简精度之间的关系. 理论分析和实验对比分析验证了所提出算法的有效性.

1 决策风险最小化属性约简

1.1 决策粗糙集基本理论

本节将介绍决策粗糙集和决策风险最小化属性约简的相关定义.

定义 1[14] 决策表信息系统是一个四元组, $S=(U, At=C\bigcup D, \{V_a|a\in At\}, \{I_a|a\in At\})$. 论域为 $U=\{x_1, x_2, \cdots, x_n\}$, 条件属性为$C$, 决策属性集为 D, V_a 为属性 $a\in At$ 取值的非空集合, $I_a: U\to V_a$ 为从 U 到 V_a 的映射函数. 通常假设 I_a 是单值的, 任意对象 $x\in U$ 在属性 $a\in At$ 上的取值可以表示为 $I_a(x)$.

在信息表中, 对于给定属性集合的子集 $A\subseteq At$, 等价关系为

$$E_A=\{(x, y)\in U\times U|\forall\, a\in A(I_a(x)=I_a(y))\}.$$

设 (U, E_A) 是定义在属性集合 A 上的近似空间, U/E_A 是基于等价关系 E_A 对 U 的一个划分, 通常表示为 π_A. 包含对象 x 的等价类表示为

$$[x]_{E_A}=[x]_A=\{y\in U|(x, y)\in E_A\}.$$

定义 2[17] 给定决策表信息系统 $S=\langle U, At, V, f\rangle$, $At=C\bigcup D$ 是属性集合, 子集 C 和 D 分别是条件属性集和决策属性集. 条件属性集合 C 关于决策属性集合 D 的近似分类质量为

$$K_C(D)=\frac{|\mathrm{POS}_C(D)|}{|U|}.$$

C 对 D 的近似分类质量表示通过属性子集 C 得到的知识对对象进行分类时, 能够确定决策的对象在论域中所占比例. 本文以近似分类质量作为约简精度.

定义 3[18] 给定决策表信息系统 $S=\langle U, At, V, f\rangle$, $At=C\bigcup D$ 是属性集合, 子集 C 和 D 分别是条件属性集和决策属性集. 条件属性子集 $B\subseteq C$ 关于 D 的重要度为

$$\delta_B(D)=K_C(D)-K_{C-B}(D).$$

定义 4[19] 等价类 $[x]_A$ 的条件概率函数定义为

$$P(X|[x]_A)=\frac{|X\bigcap[x]_A|}{|[x]_A|}.$$

1.2 决策风险最小化的属性约简

设 $\Omega=\{\omega_1, \omega_2, \cdots, \omega_s\}$ 表示 s 个状态的集合; $A=\{a_1, a_2, \cdots, a_m\}$ 表示所有可能决策; x 表示论域中的对象; $P(\omega_j|x)$ 表示对象 x 具有状态 ω_j 的条件概率; $\lambda(a_i|\omega_j)$ 表示在状态 ω_j 的情况下, 作出决策 a_i 的风险损失函数. 则当采取决策 a_i 时, 其期望风险为[15]

$$R(a_j|x)=\sum_{j=1}^{s}\lambda(a_i|\omega_j)P(\omega_j|x).$$

一般来说, 决策粗糙集中可将决策规则看作描述对象 x 的函数 $\tau(x)$, 即对于任意一种描述对象 x 都存在唯一的决策 $\tau(x)$ 与之对应, 所以决策规则的风险损失代价即为决策函数 $\tau(x)$ 的期望风险, 表示为

$$R=\sum_x R(\tau(x)|x)P(x).$$

为了描述清晰, 根据决策粗糙集模型, 对于对象 $x\in X$, λ_{PP}、λ_{NP} 和 λ_{BP} 分别表示将对象 x 划分至 $\mathrm{POS}(X)$、$\mathrm{NEG}(X)$ 和 $\mathrm{BND}(X)$ 区域时的损失函数; 反之, λ_{PN}、λ_{NN} 和 λ_{BN} 分别表示当对象 $x\notin X$ 时, 将 x 划分至 $\mathrm{POS}(X)$、$\mathrm{NEG}(X)$ 和 $\mathrm{BND}(X)$ 区域时的损失函数, 对应的决策代价损失矩阵如表 1 所示[4].

表 1 区间误判的损失矩阵

类别	$\mathrm{POS}(X)$	$\mathrm{BND}(X)$	$\mathrm{NEG}(X)$
X	λ_{PP}	λ_{BP}	λ_{NP}
$\neg X$	λ_{PN}	λ_{BN}	λ_{NN}

考虑一种常见的情况, 假设损失函数满足

$$\lambda_{\mathrm{PP}}\leqslant\lambda_{\mathrm{BP}}<\lambda_{\mathrm{NP}},\ \lambda_{\mathrm{NN}}\leqslant\lambda_{\mathrm{BN}}<\lambda_{\mathrm{PN}}. \tag{1}$$

现实意义为: 对于一个原本属于 X 的对象 x, 将其划分至 X 的正区域所带来的风险小于等于将其划分至边界域的风险, 且两者的风险均小于将其划分至 X 的负区域的风险[4]. 令

$$\begin{cases}\alpha=\dfrac{\lambda_{\mathrm{PN}}-\lambda_{\mathrm{BN}}}{(\lambda_{\mathrm{PN}}-\lambda_{\mathrm{BN}})+(\lambda_{\mathrm{BP}}-\lambda_{\mathrm{PP}})},\\ \beta=\dfrac{\lambda_{\mathrm{BN}}-\lambda_{\mathrm{NN}}}{(\lambda_{\mathrm{BN}}-\lambda_{\mathrm{NN}})+(\lambda_{\mathrm{NP}}-\lambda_{\mathrm{BP}})},\\ \gamma=\dfrac{\lambda_{\mathrm{PN}}-\lambda_{\mathrm{NN}}}{(\lambda_{\mathrm{NP}}-\lambda_{\mathrm{PP}})+(\lambda_{\mathrm{PN}}-\lambda_{\mathrm{NN}})}.\end{cases} \tag{2}$$

若损失函数满足式(1), 则根据Yao三支决策语义规则[4], 可推导出 $\alpha \in (0,1]$, $\gamma \in (0,1)$, $\beta \in [0,1)$. 另外, 决策粗糙集下的正区域、边界域和负区域的决策规则分别表示如下[19]:

$$p(D_{\max}([x]_A)|[x]_A) \geqslant \alpha,$$
$$[x]_A \subseteq \mathrm{POS}_{(\alpha,\beta)}(\pi_D|\pi_A);$$
$$\beta < p(D_{\max}([x]_A)|[x]_A) < \alpha,$$
$$[x]_A \subseteq \mathrm{BND}_{(\alpha,\beta)}(\pi_D|\pi_A);$$
$$p(D_{\max}([x]_A)|[x]_A) \leqslant \beta,$$
$$[x]_A \subseteq \mathrm{NEG}_{(\alpha,\beta)}(\pi_D|\pi_A).$$

其中 $D_{\max}([x]_A) = \arg\max\limits_{D_i \in \pi_D} \left(\dfrac{|[x]_A \bigcap D_i|}{|[x]_A|}\right)$, 表示等价类被划分至具有最大概率的那个决策类.

由于决策粗糙集决策的不确定性, 其存在一定的风险. 令 $p(D_{\max}([x]_A)|[x]_A) = l$, 则各个规则的风险表示如下[21]:

正规则的风险损失

$$l\lambda_{\mathrm{PP}} + (1-l)\lambda_{\mathrm{PN}};$$

边界规则的风险损失

$$l\lambda_{\mathrm{BP}} + (1-l)\lambda_{\mathrm{BN}};$$

负规则下风险损失

$$l\lambda_{\mathrm{NP}} + (1-l)\lambda_{\mathrm{NN}}.$$

定义 5[15]　在信息表 $S = (U, At = C \bigcup D, V, f)$ 中, 属性集合 $R \subseteq C$ 的决策风险定义为

$$\begin{aligned}\mathrm{Cost}_R = &\sum_{x_i \in \mathrm{POS}_{(\alpha,\beta)}(\pi_D|\pi_A)} (l_i\lambda_{\mathrm{PP}} + (1-l_i)\lambda_{\mathrm{PN}}) + \\ &\sum_{x_j \in \mathrm{BND}_{(\alpha,\beta)}(\pi_D|\pi_A)} (l_j\lambda_{\mathrm{BP}} + (1-l_j)\lambda_{\mathrm{BN}}) + \\ &\sum_{x_k \in \mathrm{NEG}_{(\alpha,\beta)}(\pi_D|\pi_A)} (l_k\lambda_{\mathrm{NP}} + (1-l_k)\lambda_{\mathrm{NN}}).\end{aligned}$$

定义 6[19]　给定信息表 $S = (U, At = C \bigcup D, V, f)$, 当且仅当:

1) $R = \arg(\min_{R \subseteq C}(\mathrm{Cost}_R))$;

2) $\forall R' \subset R$, $\mathrm{Cost}_{R'} > \mathrm{Cost}_R$

时, 属性集合 $R \subseteq C$ 是 C 的一个决策属性约简. 这说明在约简过程中, 仅考虑属性增删后决策表决策风险是否减小, 而不再考虑约简前后区域的变化.

2　基于属性重要度的风险决策粗糙集属性约简

现实生活中, 虽然用户偏好和获取属性值的代价有一定的关联性, 但两个概念的内涵却完全不同. 用户偏好是用户认知、心理感受及理性的经济学权衡的综合结果, 用以表达个人喜好和倾向性意见方面的意向; 获取属性值的代价指现实生活中为了获得某些属性值而需要付出的经济或其他方面的代价. 所以, 不能简单地以属性代价的高低来衡量用户对该属性偏好程度的高低. 例如, 在医疗系统中, 考虑到病人的经济条件, 倘若病人十分富裕, 那么该病人在选取采用何种检测方式的时候, 既可能选取虽然费用很高但检验结果相对准确的检测方式, 也可能选取大众化的检测方式, 但此类人群大多会优选前者. 另外, 需要考虑到在属性约简过程中可能发生误判分类区域的情况而造成一定的风险损失. 因此, 综合考虑决策者主观方面的个人偏好选择产生的风险以及客观方面的约简精度、获取属性值的实际代价和决策表区域的误判产生的风险等情况, 提出了一种基于属性重要度风险决策粗糙集属性约简算法.

现实生活中, 不同的人对同一问题有不同的个人偏好, 进而影响其决策. 因此, 讨论决策问题时, 确定用户对某特定问题偏好程度的大小就显得尤为重要. 下面给出用户偏好程度的定义.

定义 7(用户偏好程度)　用户 x_i 对属性 c_j 的主观偏好程度 ω_{ij} 可用主观偏好矩阵描述如下:

$$(\omega_{ij})_{m\times n} = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{m1} & \omega_{m2} & \cdots & \omega_{mn} \end{bmatrix},$$

$$\sum_{j=1}^{n} \omega_{ij} = 1.$$

对于现实生活中不同背景下的海量数据, 通过统计分析可以评估出不同财富阶层的人群对特定信息系统中属性的偏好程度, 进而得到用户偏好矩阵. 针对某一特定人群, 假设具有相似水平的用户偏好, 记为 $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$.

定义 8(综合权重函数)　假设向量 $T = (t_1, t_2, \cdots, t_n)$, 各分量分别为获取条件属性 $C = (c_1, c_2, \cdots, c_n)$ 属性值所需的经济代价, 归一化后为 $T' = (t'_1, t'_2, \cdots, t'_n)$. 其综合权重定义为 $p_i = u(t'_i, \omega_i)$. 特别地, 选取线性函数

$$p_i = \mu t'_i + (1-\mu)\omega_i,$$

其中 $0 \leqslant \mu \leqslant 1$ 为调节系数.

定义 9(改进属性重要度)　给定决策表信息系统 $S = \langle U, At, V, f\rangle$, $At = C \bigcup D$ 是属性集合, 子集 C 和 D 分别是条件属性集和决策属性集. $\forall c_i \in C = \{c_1, c_2, \cdots, c_n\}$, 其改进属性重要度定义为 $S_i = h(p_i, \delta_{c_i}(D))$. 特别地, 选取线性函数 $S_i = kp_i + (1 -$

$k)\delta_{c_i}(D)$, 代入 p_i 可得

$$S_i = k\mu t_i' + k(1-\mu)\omega_i + (1-k)\delta_{c_i}(D),$$

其中 $k(0 \leqslant k \leqslant 1)$ 为属性权值和属性重要度的调节系数.

当 $\mu = 0, k = 1$ 时, 仅考虑用户偏好;

当 $\mu = 1, k = 1$ 时, 仅考虑获取属性实际代价;

当 $k = 0$ 时, 仅考虑属性重要度;

当 $0 < \mu < 1, 0 < k < 1$ 时, 综合考虑用户偏好、获取属性实际代价和属性重要度.

定义 10 (代价函数) 为了度量用户偏好, 获取属性实际代价、约简精度以及区间误判产生的风险大小, 定义每个属性的代价函数为

$$f_i = f(K_C(D), \mathrm{Cost}_R, S_i).$$

为了简化模型, 取 3 个变量的加权平均值函数, 即

$$f_i = \lambda_1 K_C(D) + \lambda_2 \mathrm{Cost}_R + \lambda_3 S_i.$$

其中

$$\lambda = (\lambda_1, \lambda_2, \lambda_3);$$

$$\lambda_i > 0,\ i = 1, 2, 3;$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1.$$

该函数中的 λ_1、λ_2、λ_3 表示不同因数对应的权重 ($\lambda_1 + \lambda_2 + \lambda_3 = 1$), 这 3 个参数的设置需要一定的领域经验值, 它们对属性约简的结果有一定影响. 对于不同领域的问题, 因为考虑的因素不同, 3 个参数的取值也可能有一定差异, 但为了避免因数之间的“大吃小”现象, 一般要求这 3 个参数的数量级相同, 即相差不要太大. 由此, 全局的属性约简算法如下.

算法 1 全局属性约简算法.

Input: $S = (U, At = C \bigcup D, \{V_a | a \in At\}, \{I_a | a \in At\})$, 调节系数 μ、k, 损失矩阵 $(\lambda_{ij})_{2\times 3}$, 用户偏好矩阵 $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$, $\lambda = (\lambda_1, \lambda_2, \lambda_3)$;

Output: 约简属性子集、约简精度和约简代价.

Step 1: 计算 IND(C) 和 IND(D), 对各个属性从 $\{0, 0, \cdots, 0\}$ 至 $\{1, 1, \cdots, 1\}$ 进行编码, 1 代表选择该属性, 0 代表不选择该属性, 共 2^n 个组合.

Step 2: 1) 计算各对象 $x \in X$ 的各个组合情况下的 $P(X|[x]_A)$;

2) 根据式 (2) 计算出 α、β 的值, 确定对象 x_i 所在的区域 (正区域、负区域和边界域);

3) 根据定义 5 计算 Cost_R.

Step 3: 计算出 $K_C(D)$、p_i 和 $S_i,\ i = 1, 2, \cdots, n$.

Step 4: 对每种情况, 通过定义 10 计算出 f_i.

Step 5: 对 f_i 排序, 选择最优约简属性集.

3 基于遗传算法的局部最优约简子集的搜索

从理论上分析, 可以通过算法 1 遍历所有属性集来寻找全局最优属性约简集合, 但由于属性组合呈指数增长的特点, 属于 NP-hard[20] 问题, 本节考虑通过遗传算法, 采用启发式方法搜索出局部最优约简子集来逼近全局最优属性约简子集.

采用定长的二进制向量表示染色体, 其长度为决策信息表条件属性的个数, 且染色体的每个基因位与条件属性相对应, 即 1 和 0 分别表示选择和不选择该条件属性. 例如, 设决策表中有 7 个条件属性 $\{c_1, c_2, \cdots, c_7\}$, 则染色体 0111000 对应的条件属性为 $\{c_2, c_3, c_4\}$.

算法 2 局部最优约简子集的搜索算法.

Input: $S = (U, At = C \bigcup D, \{V_a | a \in At\}, \{I_a | a \in At\})$, 调节系数 μ、k, 损失矩阵 $(\lambda_{ij})_{2\times 3}$, 种群大小 N, 最大迭代次数 $|I - \max|$, 用户偏好矩阵 $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$, $\lambda = (\lambda_1, \lambda_2, \lambda_3)$;

Output: 约简属性子集、约简精度和约简代价.

Step 1: 计算 IND(C)、IND(D) 和近似分类质量 $K_C(D)$.

Step 2: 1) 计算各对象 $x \in X$ 的各个组合情况下的 $P(X|[x]_A)$;

2) 根据式 (2) 计算出 α、β 的值, 确定对象 x_i 所在的区域 (正区域、负区域和边界域);

3) 根据定义 5 计算 Cost_R.

Step 3: 计算出 p_i 和 $S_i,\ i = 1, 2, \cdots, n$.

Step 4: 随机产生 m 个长度为 n (条件属性的个数) 的二进制串组成初始群体; 通过交叉和变异操作, 以代价函数 f_i 为适应度函数, 计算每个属性个体的适应值, 并选择适应值最小的个体作为最优个体, 通过定义 10 计算出 f_i.

Step 5: 如果连续进行 N 代的最优个体适应度函数值不再减小或迭代次数达到最大值, 则终止计算, 否则回到 Step 4.

4 实验与对比分析

4.1 算法的实验

本节针对文中提出的算法 1 和算法 2 给出实验分析. 数据源自 UCI 数据库中的 Spect Heart, Contraceptive Method Choice, Car Evaluation 数据集, 经过一定的数据预处理, 通过 Matlab 进行实验仿真, 分别作出约简精度和约简代价的散点图, 并采用最小二乘法进行线性拟合. 由属性区间误判造成的损失矩阵如表 2 所示.

表 2　区间误判的损失矩阵

类别	POS(X)	BND(X)	NEG(X)
X	0	0.2	0.8
$\neg X$	0.9	0.2	0

设 Best_Set 为最优约简属性集, 给出属性约简率 η 如下:

$$\eta = \frac{|U| - |\text{Best\_Set}|}{|U|} \times 100\,\%.$$

选取 $\lambda = (\lambda_1, \lambda_2, \lambda_3) = (0.25, 0.45, 0.30)$, $\mu = 0.5$, $k = 0.5$, 以 Spect Heart 共 22 个条件属性的数据集进行第 1 组实验, 生成随机数, 归一化后作为用户偏好向量.

$$\omega = \{0.060\,2, 0.066\,9, 0.009\,4, 0.067\,5, 0.046\,7,$$
$$0.007\,2, 0.020\,6, 0.040\,4, 0.070\,8, 0.071\,3,$$
$$0.035\,9, 0.011\,6, 0.071\,7, 0.070\,7, 0.059\,1,$$
$$0.010\,5, 0.031\,2, 0.067\,7, 0.058\,5, 0.070\,9,$$
$$0.048\,5, 0.002\,6\}.$$

获取属性的实际代价为

$$T' = \{0.073\,6, 0.080\,9, 0.058\,8, 0.065\,7, 0.064\,4,$$
$$0.034\,0, 0.056\,8, 0.014\,8, 0.061\,2, 0.002\,8,$$
$$0.071\,4, 0.024\,0, 0.004\,0, 0.008\,4, 0.060\,2,$$
$$0.027\,5, 0.082\,3, 0.003\,0, 0.038\,0, 0.033\,1,$$
$$0.066\,3, 0.068\,9\}.$$

根据算法 2, 得到的结果如表 3 所示.

表 3　Spect Heart 数据集实验结果

约简精度	约简属性数	约简率 η/ %	约简代价 f
0.930 5	1	4.55	0.174 3
0.925 1	4	18.18	0.179 9
0.887 7	5	22.73	0.173 3
0.834 2	7	31.82	0.161 0
0.909 1	5	22.73	0.175 6
0.914 4	3	13.64	0.174 5
0.960 0	5	22.73	0.184 1
0.861 5	8	36.36	0.168 2

由表 3 可知, 在约简精度相对较高的情况下 (如 0.960 0), 其对应的约简代价同样很高, 为 0.184 1, 约简率达到 22.73 %. 在损失一定精度的情况下 (如约简精度为 0.861 5), 约简代价为 0.168 2, 明显减小.

以 Contraceptive Method Choice 共 9 个条件属性的数据集进行第 2 组实验, 生成随机数, 归一化后作为用户偏好向量.

$$\omega = \{0.038\,345, 0.179\,885, 0.080\,071,$$
$$0.115\,806, 0.075\,862, 0.180\,309,$$
$$0.157\,683, 0.158\,716, 0.013\,322\}.$$

获取属性的实际代价为

$$T' = \{0.186\,636, 0.137\,461, 0.036\,682,$$
$$0.167\,173, 0.089\,521, 0.020\,32,$$
$$0.073\,602, 0.231\,648, 0.056\,957\}.$$

根据算法 2, 得到的结果如表 4 所示.

表 4　Contraceptive Method Choice 数据集实验结果

约简精度	约简属性数	约简率 η/ %	约简代价 f
0.845 2	2	22.22	0.168 5
0.881 2	1	11.11	0.179 3
0.773 9	2	22.22	0.154 9
0.903 6	1	11.11	0.178 0
0.832 3	1	11.11	0.166 0
0.984	2	22.22	0.196 5
0.992	1	11.11	0.198 7
0.934	1	11.11	0.186 2

以 Car Evaluation 共 6 个条件属性的数据集进行第 3 组实验, 生成随机数, 归一化后作为用户偏好向量.

$$\omega = \{0.233\,4, 0.259\,5, 0.036\,4, 0.261\,7, 0.181\,2, 0.027\,9\}.$$

获取属性的实际代价为

$$T' = \{0.071\,9, 0.141\,1, 0.247\,0, 0.248\,9, 0.040\,7, 0.250\,4\}.$$

根据算法 2, 得到的结果如表 5 所示.

表 5　Car Evaluation 数据集实验结果

约简精度	约简属性数	约简率 η/ %	约简代价 f
0.936	2	33.33	0.197 2
0.984	1	16.67	0.204 3
0.808	1	16.67	0.170 3
0.69	1	16.67	0.145 1
0.96	1	16.67	0.197 5
0.88	3	50.00	0.187 0
0.92	1	16.67	0.193 4
0.72	1	16.67	0.151 9

4.2　实验分析

通过对比属性约简率、约简精度和约简代价来说明算法的有效性. 由上述 3 组实验可知, 属性约简的结果不止一种, 不同的约简结果分别对应不同的约简精度. 对约简精度和约简代价散点图进行线性拟合, 如图 1 ~ 图 3 所示.

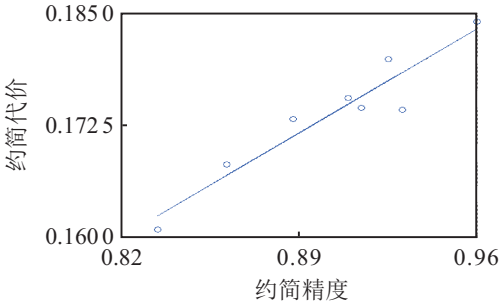

图 1　Spect Heart 数据集精度与约简代价变化图

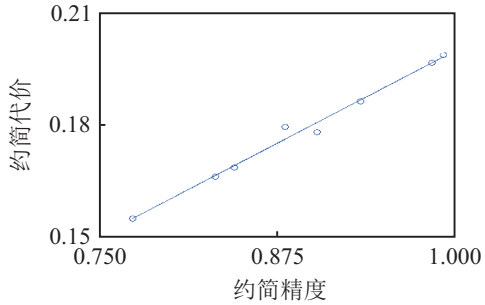

图 2 Contraceptive Method Choice 数据集精度与约简代价变化图

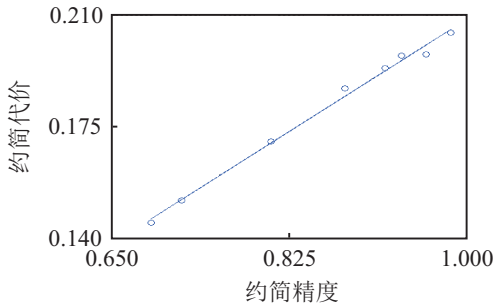

图 3 Car Evaluation 数据集精度与约简代价变化图

由图 1 ～图 3 可以发现, 约简代价与约简精度之间有正相关的关系. 随着约简精度的提高, 约简代价也会明显地增加, 但并不能保证约简代价相对于约简精度成正比地增加. 例如, 在医疗系统中, 大多数情况下, 医生在未能完全确诊的情况下便需要进行相关的治疗, 而病情的确定程度与病情检测实际代价以及病人对疾病的检测方式的偏好程度之间要有一个权衡. 另外, 如经济领域的风险决策等都在一定程度上有此规律. 因此, 本文实验结果可定量地指导用户, 以便根据自身的偏好和经济能力作出适当的决策. 实验分析进一步说明本文算法更符合现实生活, 对现实生活中的不确定、不精确事件的决策有一定指导意义.

5 结 论

本文针对 Pawlak 粗糙集决策信息系统, 特别是不一致决策信息系统[20-22]过于严格且缺乏容错能力等问题, 结合决策粗糙集理论, 考虑到不同用户对不同约简精度的需求以及获得属性值的实际代价和个人偏好等因素, 综合考虑了决策者主观个人偏好以及客观约简精度、获取属性值的实际代价、决策表各区域的误判产生的代价等情况, 提出了一种基于属性重要度的风险决策粗糙集属性约简算法, 并讨论了约简精度与约简代价之间的关系. 由于寻找全局最优属性约简集合时, 属性组合具有呈指数增长的特点, 不能对大数据集进行约简. 因此, 本文结合遗传算法, 采用启发式方法搜索出局部最优约简子集来逼近全局最优属性约简子集, 进而得到属性约简集合和风险代价值. 在 UCI 数据集上的实验分析显示, 属性约简代价与约简精度之间有正相关的关系, 在约简精度有一定损失的情况下, 约简代价也会明显地减小, 但并不能保证约简代价相对于约简精度成正比地减小. 综上可知, 所提算法具有操作性强、更加符合实际等优点. 未来的工作将会考虑到由于现实生活中不确定性的普遍存在, 决策者在不确定确切结果或未达到风险代价最小化的情况下便要做出相应决策等情况. 因此, 如何科学平衡约简精度、约简代价与用户偏好等问题是我们关注的重点.

参考文献(References)

[1] Pawlak Z. Rough sets[J]. Int J of Computer & Information Sciences, 1982, 11(5): 341-356.

[2] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1246.
(Wang G Y, Yao Y Y, Yu H. A survey on rough set theory and applications[J]. Chinese J of Computers, 2009, 32(7): 1229-1246.)

[3] Yao Y Y, Wong S K M. A decision theoretic framework for approximating concepts[J]. Int J of Man-machine Studies, 1992, 37(6): 793-809.

[4] Yao Y Y. The superiority of three-way decisions in probabilistic rough set models[J]. Information Sciences, 2011, 181(6): 1080-1096.

[5] Wang G Y, Zhao J, An J J, et al. Theoretical study on attribute reduction of rough set theory: comparison of algebra and information views[C]. Proc of the 3rd IEEE Int Conf on Cognitive Informatics. Victoria: IEEE, 2004: 148-155.

[6] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligence, 1998, 105(1/2): 77-103.

[7] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报, 2007, 30(5): 815-822.
(Yang M. An incremental updating algorithm for attribute reduction based on improved discernibility matrix[J]. Chinese J of Computers, 2007, 30(5): 815-822.)

[8] Skowron A, Rauszer C. The discernibility matrices and functions in information systems[M]. Intelligent Decision Support. Berlin: Springer Netherlands, 1992: 331-362.

[9] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.
(Miao D Q, Wang J. An information representation of the concepts and operations in rough set theory[J]. J of Software, 1999, 10(2): 113-116.)

[10] 苗夺谦, 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): 681-684.
(Miao D Q, Hu G R. A heuristic algorithm for reduction of knowledge[J]. J of Computer Research and Development, 1999, 36(6): 681-684.)

[11] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766.
(Wang G Y, Yu H, Yang D C. Decision table reduction based on conditional information entropy[J]. Chinese J of Computers, 2002, 25(7): 759-766.)

[12] 廖建坤, 叶东毅. 基于免疫粒子群优化的最小属性约简算法[J]. 计算机应用, 2007, 27(3): 550-552.
(Liao J K, Ye D Y. Minimal attribute reduction algorithm based on particle swarm optimization with immunity[J]. J of Computer Applications, 2007, 27(3): 550-552.)

[13] Zhao Y, Wong S K M, Yao Y Y. A note on attribute reduction in the decision-theoretic rough set model[M]. Transactions on Rough Sets XIII. Berlin: Springer, 2011: 260-275.

[14] Yao Y Y, Zhao Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356-3373.

[15] Jia X Y, Li W W, Shang L, et al. An optimization viewpoint of decision-theoretic rough set model[M]. Rough Sets and Knowledge Technology. Berlin: Springer, 2011: 457-465.

[16] 于洪, 姚园, 赵军. 一种有效的基于风险最小化的属性约简算法[J]. 南京大学学报: 自然科学版, 2013, 49(2): 210-216.
(Yu H, Yao Y, Zhao J. An attribute reduction algorithm based on risk minimization[J]. J of Nanjing University: Natural Sciences, 2013, 49(2): 210-216.)

[17] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001: 29-30.
(Wang G Y. Rough set theory and knowledge acquisition[M]. Xi'an: Xi'an Jiaotong University Press, 2001: 29-30.)

[18] 何明, 冯博琴, 马兆丰, 等. 一种基于 Rough 集理论的属性约简启发式算法[J]. 小型微型计算机系统, 2005, 26(3): 356-359.
(He M, Feng B Q, Ma Z F, et al. Heuristic algorithm for reduction of attributes based on rough set theory[J]. Mini-micro Systems, 2005, 26(3): 356-359.)

[19] 贾修一, 商琳, 陈家骏. 决策风险最小化属性约简[J]. 计算机科学与探索, 2011, 5(2): 155-160.
(Jia X Y, Shang L, Chen J J. Attribute reduction based on minimum decision cost[J]. J of Frontiers of Computer Science and Technology, 2011, 5(2): 155-160.)

[20] Wong S K M, Ziarko W. On optimal decision rules in decision tables[J]. Bulletin of the Polish Academy of Sciences, 1985, 33(11/12): 693-696.

[21] Yao Y Y. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3): 341-353.

[22] 陈泽华, 张裕, 谢刚. 不一致决策表规则获取的粒计算方法[J]. 控制与决策, 2015, 30(4): 709-714.
(Chen Z H, Zhang Y, Xie G. GrC method of rule acquisition for inconsistent decision table[J]. Control and Decision, 2015, 30(4): 709-714.)

(责任编辑：齐　霁)