

引用格式:徐道磊,陈培林,唐轶轩,等.一种新的决策粗糙集最小化决策代价属性约简算法[J].微电子学与计算机,2020,37(8):55-60. [XU D L, CHEN P L, TANG Y X, et al. A new decision-theoretic rough set model and minimum decision cost attribute reduction algorithm [J]. Microelectronics & Computer, 2020, 37(8): 55-60.]

一种新的决策粗糙集最小化决策代价属性约简算法

徐道磊¹,陈培林²,唐轶轩¹,吴 尚¹,路 宇¹,卞显福³

(1 安徽省电力有限公司 信息通信分公司,安徽 合肥 230061;

2 中国劳动关系学院,北京 100028;3 中国科学技术大学 软件学院,安徽 合肥 230051)

摘 要: 实际应用中存在着大量的数值型数据,然而传统的决策粗糙集只能处理符号型数据,为了改善这一局限性,本文构造出一种模糊邻域决策粗糙集模型,并提出一种最小化决策代价的属性约简算法.文中首先将模糊粗糙集和邻域粗糙集融入决策粗糙集中,提出了模糊邻域决策粗糙集,使得该模型同时具有模糊粗糙集和邻域粗糙集处理数值型数据的优点;然后基于该模型,给出一种决策代价定义,并提出相应的最小化决策代价属性约简算法;最后通过实验分析表明所提出的算法具有较好的代价敏感属性约简性能.

关键词: 决策粗糙集;模糊粗糙集;邻域粗糙集;决策代价;属性约简

中图分类号: TP18

文献标识码: A

文章编号: 1000—7180(2020)08—0055—06

A new decision-theoretic rough set model and minimum decision cost attribute reduction algorithm

XU Dao-lei¹, CHEN Pei-lin², TANG Yi-xu¹, WU Shang¹,
LU Yu¹, BIAN Xian-fu³

(1 Information and Communication Branch, State Grid Anhui Electric Power Co., Ltd.,
Hefei 230061, China; 2 China Institute of Labor Relations, Beijing 100028, China;

3 School of Software, University of Science and Technology of China, Hefei 230051, China)

Abstract: There are a lot of numerical data in practical application, but traditional decision-theoretic rough sets can only deal with symbolic data. In order to improve this limitation, this paper constructs a fuzzy neighborhood decision-theoretic rough set model and proposes a minimum decision cost attribute reduction algorithm. Firstly, the fuzzy rough set and neighborhood rough set are integrated into decision-theoretic rough set, and the fuzzy neighborhood decision-theoretic rough set is proposed, which makes the model have the advantages of both fuzzy rough set and neighborhood rough set in dealing with numerical data. Then, based on the model, a definition of decision cost is given and the corresponding minimum decision cost attribute reduction is proposed. Finally, the experimental results show that the proposed algorithm has better cost-sensitive attribute reduction performance.

Key words: decision-theoretic rough set; fuzzy rough set; neighborhood rough set; decision cost; attribute reduction

收稿日期: 2019—11—01; 修回日期: 2019—12—19

基金项目: 国网合肥供电公司 2019 年智慧能源“互联网+”服务平台二期实施服务(SGAHHF00XTXX1900770); 合肥供电公司 2019 年通信网管系统软硬件运行维护项目(B112A019000Q)

1 引言

粗糙集理论^[1]是由波兰学者 Pawlak 提出的一种处理不确定性数据的数据挖掘方法,目前已广泛应用于数据挖掘^[2]、机器学习^[3]和特征选择^[4-5]等领域.传统的粗糙集模型中,由于未考虑分类错误的容差性,Yao 进一步提出了决策粗糙集模型^[6],近年来,决策理论粗糙集已不断地扩展和应用^[7-9]到各个数据挖掘领域.属性约简^[6]也称为特征选择,是粗糙集理论的重要研究内容.在决策粗糙集模型中,学者们提出了多种属性约简方法,其中属性约简方法主要分两种,一种是基于正区域的属性约简^[10-11];另一种是基于最小化决策代价的属性约简^[12-14].例如 Li 等学者在决策粗糙集模型中提出一种非单调性的属性约简算法^[15];Ma 等学者设计了三种保持决策区域分布的属性约简算法^[10];Jia 等学者在决策粗糙集中基于代价的视角提出了一种最小化决策代价的属性约简算法^[12].

然而在实际应用环境下,信息系统中存在着大量的数值型数据,为了对这类数据进行有效地处理,Li 等学者提出了基于邻域决策粗糙集模型的属性约简算法^[16];Zhao 等学者提出了基于模糊概率测度的模糊决策理论粗糙集方法^[17],这两种方法的提出,均提高了决策粗糙集的数据处理范围.

众所周知,邻域粗糙集和模糊粗糙集都是处理数值型数据的有效模型,并且在处理方式上具有各自的优越性,将这两种单一方法进行融合,则可以提高数值型数据的处理性能^[18-19].本文将采用这一思想,提出一种模糊邻域决策粗糙集模型.模糊粗糙集模型通过对数值型数据构建模糊相似关系,来构建数据的信息粒化.而邻域粗糙集模型通过邻域半径对数据对象进行邻域粒化,从而针对数值型数据进行粗糙逼近.为了融合模糊粗糙集和邻域粗糙集处理数值型数据的优点,文中在模糊粗糙集的模糊相似关系中引入一个邻域半径,通过该邻域半径构造数值型数据的模糊邻域粒化,然后基于该粒化模型重构决策粗糙集,提出模糊邻域决策粗糙集模型,基于该模型给出决策代价的定义,提出一种最小化决策代价的属性约简,最后通过实验证明所提出属性约简的有效性和优越性.

2 基本理论

2.1 决策粗糙集模型

设决策信息系统^[1] $DIS = (U, C \cup D, V, f)$,其

中 U 表示信息系统的论域,也称为对象集, C 和 D 分别称之为信息系统的条件属性集和决策属性集, V 为信息系统的值域, f 为对象在属性下关于属性值的映射,即 $\forall x \in U$ 在属性 $a \in C$ 的属性值表示为 $f(a, x)$.

对于属性子集 $\forall A \subseteq C$ 在论域 U 上的等价关系^[1] R_A 定义为 $R_A = \{(x, y) \mid f(a, x) = f(a, y), \forall a \in A\}$. 对象 x 在等价关系 R_A 下的等价类表示为 $[x]_A = \{y \in U \mid (x, y) \in R_A\}$.

在决策粗糙集中^[6], 设 $\Theta = \{X, X^c\}$ 表示的对象 x 的隶属状态,即属于 X 和不属于 X ,其中对象 x 属于 X 的概率可以计算为 $p(X \mid [x]_A) = \frac{|X \cap [x]_A|}{|[x]_A|}$, 则 $p(X^c \mid [x]_A) = 1 - p(X \mid [x]_A)$.

设含有三种行为的决策动作集合 $Ac = \{a_P, a_B, a_N\}$, 分别表示对象 x 关于决策 X 的接受、延迟和拒绝决策行为. 如下所述的是对象处于不同状态采取相应决策行为的决策代价矩阵.

$$\mathbf{M} = \begin{bmatrix} \lambda_{PP} & \lambda_{PN} \\ \lambda_{BP} & \lambda_{BN} \\ \lambda_{NP} & \lambda_{NN} \end{bmatrix}$$

式中, $\lambda_{*P} (* \in \{P, B, N\})$ 表示对象 x 属于 X 时, 分别采取 a_P, a_B, a_N 三种行动所产生的代价; $\lambda_{*N} (* \in \{P, B, N\})$ 表示对象 x 属于 X^c 时, 分别采取 a_P, a_B, a_N 三种行动所产生的代价.

那么对于 $\forall x \in U$ 采取的不同决策行为所产生的代价可表示为

$$\begin{aligned} \Gamma(a_P \mid x) &= \lambda_{PP} \cdot p(X \mid [x]_A) + \lambda_{PN} \cdot p(X^c \mid [x]_A); \\ \Gamma(a_B \mid x) &= \lambda_{BP} \cdot p(X \mid [x]_A) + \lambda_{BN} \cdot p(X^c \mid [x]_A); \\ \Gamma(a_N \mid x) &= \lambda_{NP} \cdot p(X \mid [x]_A) + \lambda_{NN} \cdot p(X^c \mid [x]_A). \end{aligned}$$

在一个合理的假设情况下,即选择正确行为所产生的代价要小于选择错误行为所产生的代价,那么有 $\lambda_{PP} \leq \lambda_{BP} \leq \lambda_{NP}$ 和 $\lambda_{NN} \leq \lambda_{BN} \leq \lambda_{PN}$. 又由于 $p(X \mid [x]_A) + p(X^c \mid [x]_A) = 1$, 所以基于贝叶斯决策理论可以得出以下最小代价的决策规则:

- (1) 接受规则: 若 $p(X \mid [x]_A) \geq \alpha$ 且 $p(X \mid [x]_A) \geq \gamma$, 则选择 $x \in POS(X)$;
- (2) 延迟规则: 若 $p(X \mid [x]_A) \leq \alpha$ 且 $p(X \mid [x]_A) \geq \beta$, 则选择 $x \in BND(X)$;
- (3) 拒绝规则: 若 $p(X \mid [x]_A) \leq \beta$ 且 $p(X \mid [x]_A) \leq \gamma$, 则选择 $x \in NEG(X)$.

这里的 α, β, γ 能够从代价矩阵中分别被计算出来,其中

$$\begin{aligned}\alpha &= \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}, \\ \beta &= \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}, \\ \gamma &= \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}.\end{aligned}$$

2.2 邻域粗糙集

邻域粗糙集模型是由 Hu 等学者提出的一种处理数值型信息系统的粗糙集模型^[19].

定义 1^[19] 给定一个数值型信息系统 $NS = (U, C \cup D, V, f)$, 属性子集 $A \subseteq C$, 对象 $\forall x \in U$ 在属性子集 A 上的 δ 邻域类定义为

$$n_B^\delta(x) = \{y \in U \mid d_A(x, y) \leq \delta\}$$

式中, $d_A(x, y)$ 为对象之间的距离度量; δ 为邻域半径, 对象 x 的邻域类表示的是与对象 x 处于 δ 半径内对象的集合, 其中 δ 用来控制这个范围的大小.

定义 2^[19] 给定数值型信息系统 $NS = (U, C \cup D, V, f)$, 对于属性子集 $A \subseteq C$ 与 $\forall X \subseteq U$, X 关于属性子集 A 的邻域下近似集和邻域上近似集分别定义为

$$\underline{N}_A(X) = \{x \in U \mid n_A^\delta(x) \subseteq X\}$$

$$\bar{N}_A(X) = \{x \in U \mid n_A^\delta(x) \cap X \neq \emptyset\}$$

3 模糊邻域决策粗糙集模型

为了融合模糊粗糙集和邻域粗糙集处理数值型数据的优点, 本节将在模糊相似关系中引入邻域半径, 通过该邻域半径构造数值型数据的模糊邻域粒化.

定义 3 给定数值型信息系统 $NS = (U, C \cup D, V, f)$, 属性子集 $A \subseteq C$ 在 U 上诱导的模糊相似关系为 $\tilde{R}_A$, $\tilde{R}_A(x, y)$ 表示 $\forall x, y \in U$ 的模糊相似度. 设 $\tilde{R}_a$ 是由 $a \in A$ 诱导的模糊相似关系, 则有 $\tilde{R}_A = \bigcap_{a \in A} \tilde{R}_a$. 对于 $\forall x \in U$ 基于 $\tilde{R}_A$ 的模糊邻域粒 $[x]_A^\delta$ 定义为:

$$[x]_A^\delta(y) = \begin{cases} 0, & R_A(x, y) < 1 - \delta \\ R_A(x, y), & R_A(x, y) \geq 1 - \delta \end{cases}$$

其中 δ 被称为模糊邻域半径, 满足 $0 < \delta \leq 1$.

类似于传统的决策粗糙集, 在模糊邻域决策粗糙集模型中, 设 $\Theta = \{\tilde{X}, \tilde{X}^c\}$ 表示的对象 x 属于模糊集 $\tilde{X}$ 和对象 x 不属于模糊集 $\tilde{X}$, 其中对象 x 属于模

糊集 $\tilde{X}$ 的概率计算为

$$P(\tilde{X} \mid [x]_A^\delta) = \frac{\sum_{i=1}^{|U|} (\tilde{X} \cap [x]_A^\delta)(x_i)}{\sum_{i=1}^{|U|} [x]_A^\delta(x_i)},$$

则对象 x 属于 $\tilde{X}^c$ 的概率即为 $p(\tilde{X}^c \mid [x]_A^\delta) = 1 - p(\tilde{X} \mid [x]_A^\delta)$. 设含有三种行为的决策动作集合 $Ac = \{a_P, a_B, a_N\}$, 分别表示对象 x 关于决策 $\tilde{X}$ 的接受、延迟和拒绝决策行为, 决策代价为 $\mathbf{M}$.

那么对于 $\forall x \in U$ 采取不同决策行为所产生的代价可表示为

$$\Gamma(a_P \mid x) = \lambda_{PP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{PN} \cdot p(\tilde{X}^c \mid [x]_A^\delta);$$

$$\Gamma(a_B \mid x) = \lambda_{BP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{BN} \cdot p(\tilde{X}^c \mid [x]_A^\delta);$$

$$\Gamma(a_N \mid x) = \lambda_{NP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{NN} \cdot p(\tilde{X}^c \mid [x]_A^\delta).$$

根据贝叶斯最小化决策规则, 有

(1) 当对象 x 采取 a_P 时的代价最小, 那么

$$\begin{aligned}\lambda_{PP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{PN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \leq \lambda_{BP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{BN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \Rightarrow p(\tilde{X} \mid [x]_A^\delta) \geq \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}\end{aligned}$$

同时

$$\begin{aligned}\lambda_{PP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{PN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \leq \lambda_{NP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{NN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \Rightarrow p(\tilde{X} \mid [x]_A^\delta) \geq \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}\end{aligned}$$

(2) 当对象 x 采取 a_B 时的代价最小,

$$\begin{aligned}\lambda_{BP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{BN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \leq \lambda_{PP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{PN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \Rightarrow p(\tilde{X} \mid [x]_A^\delta) \leq \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}\end{aligned}$$

同时

$$\begin{aligned}\lambda_{BP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{BN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \leq \lambda_{NP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{NN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \Rightarrow p(\tilde{X} \mid [x]_A^\delta) \geq \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}\end{aligned}$$

(3) 当对象 x 采取 a_N 时的代价最小,

$$\begin{aligned}\lambda_{NP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{NN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \leq \lambda_{PP} \cdot p(\tilde{X} \mid [x]_A^\delta) + \lambda_{PN} \cdot (1 - p(\tilde{X} \mid [x]_A^\delta)) \\ \Rightarrow p(\tilde{X} \mid [x]_A^\delta) \leq \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}\end{aligned}$$

同时

$$\begin{aligned} & \lambda_{NP} \cdot p(\tilde{X} | [x]_A^\delta) + \lambda_{NN} \cdot (1 - p(\tilde{X} | [x]_A^\delta)) \\ & \leq \lambda_{BP} \cdot p(\tilde{X} | [x]_A^\delta) + \lambda_{BN} \cdot (1 - p(\tilde{X} | [x]_A^\delta)) \\ & \Rightarrow p(\tilde{X} | [x]_A^\delta) \leq \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})} \end{aligned}$$

综合上述推导结果,可以得到:

(1) 当 $p(\tilde{X} | [x]_A^\delta) \geq \alpha$ 且 $p(\tilde{X} | [x]_A^\delta) \geq \gamma$, 那么 $x \in POS_A^\delta(\tilde{X})$;

(2) 当 $p(\tilde{X} | [x]_A^\delta) \leq \alpha$ 且 $p(\tilde{X} | [x]_A^\delta) \geq \beta$, 那么 $x \in BUN_A^\delta(\tilde{X})$;

(3) 当 $p(\tilde{X} | [x]_A^\delta) \leq \beta$ 且 $p(\tilde{X} | [x]_A^\delta) \leq \gamma$, 那么 $x \in NEG_A^\delta(X)$.

式中, $\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}$,

$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})},$$

$$\gamma = \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}.$$

根据上述的一系列的推导,我们可以得到模糊邻域决策粗糙集模型三个决策区域的划分.

定义 4 给定数值型信息系统 $NS = (U, C \cup D, V, f)$, 属性子集 $A \subseteq C$, 论域 U 在 A 下诱导的模糊相似关系为 $\tilde{R}_A$, 对象集 $X \subseteq U$ 关于模糊相似关系 $\tilde{R}_A$ 的模糊邻域决策下近似和模糊邻域决策上近似分别定义为

$$\underline{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{X}) = \{x \in U \mid p(\tilde{X} | [x]_A^\delta) \geq \alpha\};$$

$$\bar{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{X}) = \{x \in U \mid p(\tilde{X} | [x]_A^\delta) > \beta\}.$$

这里称 $\langle \underline{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{X}), \bar{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{X}) \rangle$ 为 X 关于模糊相似关系 $\tilde{R}_A$ 的模糊邻域决策粗糙集. 此外, 决策属性集 D 确定的决策类划分为 $U/D = \{D_1, D_2, \dots, D_m\}$, 那么决策属性集 D 关于模糊相似关系 $\tilde{R}_A$ 的三个区域表示为:

$$POS_{A,\delta}^{(\alpha,\beta)}(D) = \bigcup_{i=1}^m \underline{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{D}_i);$$

$$BUN_{A,\delta}^{(\alpha,\beta)}(D) = \bigcup_{i=1}^m \bar{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{D}_i) - \bigcup_{i=1}^m \underline{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{D}_i);$$

$$NEG_{A,\delta}^{(\alpha,\beta)}(D) = U - \bigcup_{i=1}^m \bar{N}_{A,\delta}^{(\alpha,\beta)}(\tilde{D}_i).$$

4 最小决策代价属性约简算法

属性约简是粗糙集理论的核心研究问题. 本节

在所提出的模糊邻域决策粗糙集模型中, 定义一种决策代价函数, 然后基于该决策代价提出一种最小化决策代价的属性约简算法.

定义 5 给定数值型信息系统 $NS = (U, C \cup D, V, f)$, 属性子集 $A \subseteq C$ 诱导的模糊相似关系为 $\tilde{R}_A$, 设对象集 X 关于 $\tilde{R}_A$ 在论域 U 下三个区域的划分分别为 $POS_A^\delta(\tilde{X})$, $BUN_A^\delta(\tilde{X})$ 和 $NEG_A^\delta(\tilde{X})$. 那么进行区域划分所产生的代价分别表示为

$$\Gamma_{A,\delta}^{POS}(\tilde{X}) = \sum_{\forall x \in POS_A^\delta(X)} \{\lambda_{PP} \cdot (p(\tilde{X} | [x]_A^\delta)) + \lambda_{PN} \cdot (1 - p(\tilde{X} | [x]_A^\delta))\}$$

$$\Gamma_{A,\delta}^{BUN}(\tilde{X}) = \sum_{\forall x \in BUN_A^\delta(X)} \{\lambda_{BP} \cdot (p(\tilde{X} | [x]_A^\delta)) + \lambda_{BN} \cdot (1 - p(\tilde{X} | [x]_A^\delta))\}$$

$$\Gamma_{A,\delta}^{NEG}(\tilde{X}) = \sum_{\forall x \in NEG_A^\delta(X)} \{\lambda_{NP} \cdot (p(\tilde{X} | [x]_A^\delta)) + \lambda_{NN} \cdot (1 - p(\tilde{X} | [x]_A^\delta))\}$$

对于决策属性集 D 诱导的决策类划分 $U/D = \{D_1, D_2, \dots, D_m\}$, 关于 $\tilde{R}_A$ 的区域划分为 $POS_{A,\delta}^{(\alpha,\beta)}(D)$, $BUN_{A,\delta}^{(\alpha,\beta)}(D)$ 和 $NEG_{A,\delta}^{(\alpha,\beta)}(D)$ 的决策代价分别表示为

$$\Gamma_{A,\delta}^{POS}(D) = \sum_{i=1}^m \sum_{\forall x \in POS_{A,\delta}^{(\alpha,\beta)}(D_i)} \{\lambda_{PP} \cdot (p(D_i | [x]_A^\delta)) + \lambda_{PN} \cdot (1 - p(D_i | [x]_A^\delta))\}$$

$$\Gamma_{A,\delta}^{BUN}(D) = \sum_{i=1}^m \sum_{\forall x \in BUN_{A,\delta}^{(\alpha,\beta)}(D_i)} \{\lambda_{BP} \cdot (p(D_i | [x]_A^\delta)) + \lambda_{BN} \cdot (1 - p(D_i | [x]_A^\delta))\}$$

$$\Gamma_{A,\delta}^{NEG}(D) = \sum_{i=1}^m \sum_{\forall x \in NEG_{A,\delta}^{(\alpha,\beta)}(D_i)} \{\lambda_{NP} \cdot (p(D_i | [x]_A^\delta)) + \lambda_{NN} \cdot (1 - p(D_i | [x]_A^\delta))\}$$

根据每个区域进行决策的决策代价, 我们可以进一步得到最终的决策总代价, 表示为

$$\Gamma_{A,\delta} = \Gamma_{A,\delta}^{POS}(D) + \Gamma_{A,\delta}^{BUN}(D) + \Gamma_{A,\delta}^{NEG}(D).$$

由于信息系统中不是每个属性对决策都是有利的, 有的属性会提高信息系统的决策代价, 有的属性会减少信息系统的决策代价, 因此对于提高决策代价的属性, 我们需要对他们进行删除, 从而尽可能降低整个信息系统的决策总代价, 基于这一点可以定义出信息系统的最小化决策代价属性约简.

定义 6 给定一个数值型信息系统 $NS = (U, C \cup D, V, f)$, 若 $B \subseteq C$ 是 C 的一个最小化决策代价属性约简, 则需同时满足以下两个条件

(1) $\Gamma_{A,\delta} \leq \Gamma_{C,\delta}$; (2) $\forall A' \subset A, \Gamma_{A',\delta} > \Gamma_{A,\delta}$.

根据条件(1) 和条件(2) 可知最小决策代价属性约简集是一组最小的条件属性集,并且该属性集具有最小的决策代价.在决策代价属性约简中,将不再考虑属性约简前后三个区域的大小变化,而只关注区域变化后决策代价是否减小,若删除某个属性可使得整个决策表的决策代价减小,则该属性可以进行删除.

定义 7 给定一个数值型信息系统 $NS = (U, C \cup D, V, f)$,由 $A \subseteq C$ 在论域 U 上诱导的模糊相似关系为 $\tilde{R}_A$,邻域半径设为 δ ,则对于 $\forall a \in C - A$ 在决策属性 D 下关于 A 的属性重要度定义为

$$Sig_A(a) = \frac{\Gamma_{A,\delta} - \Gamma_{A \cup \{a\},\delta}}{\Gamma_{A,\delta}}.$$

该属性重要度表示的是属性子集 A 外部属性 a 关于 A 的属性重要度,利用该属性重要度作为启发式函数,可以构造出相应的最小化决策代价属性约简算法.

算法 1. 基于模糊邻域决策粗糙集的最小化决策代价属性约简算法.

输入:数值型决策信息系统 $NS = (U, C \cup D, V, f)$,决策代价矩阵 $\mathbf{M}$,邻域半径 δ ;

输出:属性约简集 Red .

步骤 1. 令 $Red = \emptyset$.

步骤 2. 将 $\Gamma_{\emptyset,\delta}$ 定义为一个很大的常量,对于 $\forall a \in C$,计算 $Sig_{\emptyset}(a)$. 记 $a_{max} = \arg \max_{a \in C} Sig_{\emptyset}(a)$,并且 $Red \leftarrow Red \cup \{a_{max}\}$.

步骤 3. 若 $\Gamma_{Red,\delta} > \Gamma_{C,\delta}$,那么跳转至步骤 4. 若 $\Gamma_{Red,\delta} \leq \Gamma_{C,\delta}$,那么跳转至步骤 6.

步骤 4. 对于 $\forall a \in C - Red$,计算 $Sig_{Red}(a)$,记 $a_{max} = \arg \max_{a \in C - Red} Sig_{Red}(a)$. 若 $Sig_{Red}(a_{max}) > 0$,则 $Red \leftarrow Red \cup \{a_{max}\}$ 并重新跳转至步骤 4, 若 $Sig_{Red}(a_{max}) \leq 0$,则跳转至步骤 5.

步骤 5. 对于 $\forall a \in Red$,若 $\Gamma_{Red - \{a\},\delta} < \Gamma_{Red,\delta}$,则 $Red \leftarrow Red - \{a\}$. 并重复步骤 5,直到不满足时终止.

步骤 6. 返回属性约简结果 Red .

5 实验分析

为了验证本文所提出的最小化决策代价属性约简算法具有一定的优越性,本节将通过进行实验来证明.实验从 UCI 机器学习数据库中获取 7 个数据集,具体信息如表 1 所示.

表 1 数据集信息

编号	名称	对象	属性	类
1	move	360	90	15
2	iris	150	4	3
3	image	210	19	7
4	iono	351	34	2
5	wine	178	13	2
6	sona	208	60	2
7	gerbox	1 603	72	4

为了避免数据集属性量纲带来的影响,实验前将数据集的属性值采用如下公式进行归一化处理:

$$f'(a, x) = \frac{f(a, x) - a_{min}}{a_{max} - a_{min}}, \forall x \in U$$

式中, a_{min} 和 a_{max} 分别表示属性 a 中属性值的最小值和最大值; $f(a, x)$ 表示对象 x 在属性 a 下的属性值; $f'(a, x)$ 表示对象 x 在属性 a 下归一化后的值. 经过该公式进行处理,所有的属性值均归一化至 $[0, 1]$ 区间. 另外对象 x 和 y 在属性 a 下的模糊相似关系计算为

$$\tilde{R}_a(x_i, x_j) = \begin{cases} 1 - 4 \cdot |f(a, x) - f(a, y)|, & |f(a, x) - f(a, y)| \leq 0.25 \\ 0, & \text{其他} \end{cases}$$

对于本文所提出的属性约简算法,邻域半径 δ 是一个很重要的输入参数,为了能够得到优越的属性约简结果,本实验采用其他学者得到的邻域半径选取结果^[16,19],这里选择邻域半径 $\delta = 0.15$ 来进行实验.此外,算法的入参还包含了决策代价矩阵,实验中对代价函数在 $[0, 1]$ 内进行随机选择,其中选取满足关系 $\lambda_{PP} \leq \lambda_{BP} \leq \lambda_{NP}$ 和 $\lambda_{NN} \leq \lambda_{BN} \leq \lambda_{PN}$,且 $\lambda_{PP} = \lambda_{NN} = 0$. 本文所涉及的各个算法均采用 Matlab2014b 进行编程.

本实验选取了 4 种对比算法进行比较验证,分别为基于决策正区域的决策粗糙集属性约简^[11](记为算法 1),基于符号型信息系统的最小化决策代价属性约简^[12](记为算法 2),基于邻域决策粗糙集的最小化决策代价属性约简^[16](记为算法 3)和基于模糊决策粗糙集的最小化决策代价属性约简^[13](记为算法 4).将这 4 中对比算法与本文算法对表 1 中的每个数据集进行属性约简,其约简后属性数量的结果如表 2 所示.

观察表 2 可以看出,5 种属性约简算法的约简结果基本上都比数据集原始的属性数量要小,这说明这 5 种属性约简算法都是有效的.通过比较 5 种

约简算法的属性约简结果,可以看出本文所提出的属性约简算法所得到的约简集属性数量更小,因此本文所提出的约简算法更加有效.

表 2 属性约简结果比较

数据集	约简后的属性数量				
	算法 1	算法 2	算法 3	算法 4	本文算法
move	17	16	18	15	11
iris	4	4	4	4	3
image	15	10	12	13	8
iono	20	15	16	18	12
wine	8	5	6	7	4
sona	16	15	18	17	10
gerbox	17	16	18	15	14

在这 5 种算法中,有 4 种是基于决策代价的属性约简算法,表 3 所示的是各个算法属性约简决策代价的结果. 观察表 3 可以发现,算法 1 的决策代价在所有算法中是最高的,这主要是由于该算法是基于正区域的属性约简,未考虑约简结果的决策代价. 算法 2 的决策代价稍高于算法 3、4 和本文算法,这主要是由于该算法是基于符号型的信息系统,而本实验中所有的数据集均为数值型的数据集,算法 2 进行属性约简时会将数值型数据进行离散化处理,这一过程丢失了属性的分类信息,因而决策代价较高. 对比算法 3、4 和本文算法,发现这 3 种算法的决策代价相近,仔细比较可以发现除数据集 image 以外,本文算法拥有最小的决策代价,因此本文算法的性能更优,这主要是由于本文的决策粗糙集模型融合模糊和邻域两方面的度量,因而具有更高的数据相似性刻画效果.

表 3 属性约简结果决策代价比较

数据集	约简后的属性数量				
	算法 1	算法 2	算法 3	算法 4	本文算法
move	54.5	42.8	36.1	35.3	34.2
iris	32.6	28.2	27.6	27.5	26.1
image	49.4	38.6	37.9	35.3	36.7
iono	32.3	29.6	26.7	27.3	25.5
wine	34.9	27.5	26.2	28.1	23.3
sona	36.7	31.9	27.5	30.3	27.3
gerbox	44.4	36.4	31.2	32.9	29.6

在代价敏感学习中,误分类代价也是一种评估算法性能的重要指标,其中误分类代价可以定义为 $\Gamma_{\text{mis}} = n_{\text{NP}} \cdot \lambda_{\text{NP}} + n_{\text{PN}} \cdot \lambda_{\text{PN}}$. 这里的 n_{NP} 和 n_{PN} 表示的是误分类对象的数目. 表 4 所示的是各算法误分类

代价的结果. 观察表 4 可以发现,本文算法在大部分数据集下的误分类代价最小,算法 3 在数据集 wine 下拥有最小的误分类代价,算法 4 在 gearbox 下拥有最小的误分类代价,因此本文算法的属性约简结果更优一些.

表 4 误分类代价结果

数据集	约简后的属性数量				
	算法 1	算法 2	算法 3	算法 4	本文算法
move	19.5	20.8	17.5	17.3	15.9
iris	16.4	9.9	11.3	7.6	4.7
image	21.6	16.5	14	15.1	9.8
iono	11.3	8.6	9.7	7.3	5.5
wine	12.9	7.6	3.5	4.1	8.6
sona	21.5	12.8	20.4	15.1	12.4
gerbox	10.4	9.7	6.3	4.3	6.6

通过实验结果可以得出,本文所提出的最小化决策代价属性约简算法在各个指标下有着更优的结果,从而证明了本文算法的有效性.

6 结束语

决策粗糙集模型是粗糙集理论的研究热点. 为了更好地处理数值型的信息系统,本文将融合邻域粗糙集和模糊粗糙集的各自优越性,提出一种模糊邻域决策粗糙集. 文章中首先在模糊关系下增加了一个邻域半径,提出了数值型信息系统的模糊邻域粒化,进而提出了模糊邻域决策粗糙集,使得该模型同时具有模糊粗糙集和邻域粗糙集处理数值型数据的优点. 然后基于该模型给出了决策代价的定义,并提出一种最小化决策代价的属性约简算法,实验分析表明所提出的算法具有较好的属性约简性能. 接下来将进一步研究动态数据环境下最小化决策代价的属性约简问题.

参考文献:

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer Information Science, 1982, 11(5): 341-356.

[2] YAN Y T, WU Z B, DU X Q, et al. A three-way decision ensemble method for imbalanced data oversampling[J]. International Journal of Approximate Reasoning, 2019, 107:1-16

[3] YANG Y L. Research on high-precision classification algorithm for rough set data[J]. Modern electronics technology, 2018, 41(10):154-156,160.