智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向混合数据的代价敏感三支决策边界域分类方法

周阳阳, 钱文彬, 王映龙, 彭莉莎, 曾武序

引用本文:
周阳阳,钱文彬,王映龙,彭莉莎,曾武序. 面向混合数据的代价敏感三支决策边界域分类方法[J]. 智能系统学报, 2022, 17(2): 411–419.
ZHOU Yangyang,QIAN Wenbin,WANG Yinglong,PENG Lisha,ZENG Wuxu. Classification method of cost–sensitive three–way decision boundary region for hybrid data[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(2): 411–419.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202012048

DOI: 10.11992/tis.202012048
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20211012.1841.002.html

面向混合数据的代价敏感三支决策边界域分类方法

周阳阳[1], 钱文彬[1,2], 王映龙[1], 彭莉莎[3], 曾武序[1]
(1. 江西农业大学 计算机与信息工程学院, 江西 南昌 330045; 2. 江西农业大学 软件学院, 江西 南昌 330045; 3. 南京大学 工程管理学院, 江苏 南京 210046)

摘 要: 针对现有三支决策模型的研究对象多为单一性数据的决策系统, 对于混合数据边界域样本处理的研究相对较少, 本文面向混合数据提出了基于核属性的代价敏感三支决策边界域分类方法。该方法基于正域约简计算混合邻域决策系统的核属性集, 在此基础上计算混合邻域类, 并利用三支决策规则分别将对象划分到各决策类的正域、边界域和负域; 提出了一种基于代价敏感学习的三支决策边界域分类方法, 并构造了误分类代价的计算方法, 以此划分边界域中的对象。通过对 UCI 上的 10 个数据集进行实验对比与分析, 进一步验证了本文方法, 为处理边界域样本提供了一种可行有效的方法。

关键词: 三支决策; 粒计算; 代价敏感; 混合数据; 正域约简; 边界域样本处理; 粗糙集; 核属性
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2022)02−0411−09

中文引用格式: 周阳阳, 钱文彬, 王映龙, 等. 面向混合数据的代价敏感三支决策边界域分类方法 [J]. 智能系统学报, 2022, 17(2): 411–419.
英文引用格式: ZHOU Yangyang, QIAN Wenbin, WANG Yinglong, et al. Classification method of cost-sensitive three-way decision boundary region for hybrid data[J]. CAAI transactions on intelligent systems, 2022, 17(2): 411–419.

Classification method of cost-sensitive three-way decision boundary region for hybrid data

ZHOU Yangyang[1], QIAN Wenbin[1,2], WANG Yinglong[1], PENG Lisha[3], ZENG Wuxu[1]
(1. School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China; 2. School of Software, Jiangxi Agricultural University, Nanchang 330045, China; 3. School of Engineering Management, Nanjing University, Nanjing 210046, China)

Abstract: The research objects of existing three-way decisions models are mostly decision-making systems with single data. Relatively few studies on the boundary region sample processing of mixed data have been conducted. To address this issue, a classification method of a cost-sensitive three-way decision boundary region based on core attributes for hybrid data is proposed in this study. This method computes the core attribute set of the hybrid neighborhood decision system based on positive domain reduction. On this basis, the hybrid neighborhood class is calculated, and the objects are divided into the positive, boundary, and negative regions of each decision-making class through three-way decision rules. The classification method of the three-way decision boundary region based on cost-sensitive learning is proposed. Then, a calculation method of the misclassification cost is constructed to divide the objects in the boundary region. Experiments and analyses are performed on 10 datasets of UCI, which show the feasibility and the effectiveness of the proposed method for the processing of boundary region samples.

Keywords: three-way decisions; granular computing; cost sensitive; hybrid data; positive domain reduction; boundary region sample processing; rough set; core attribute

三支决策是加拿大学者 Yao[1-2] 提出的一种“化繁为简”决策理论, 它从粒计算视角将论域划分为三个互不相交的论域子空间, 并对其分别采取不同的应对策略, 这种分而治之的思想, 可有

收稿日期: 2020−12−28. **网络出版日期**: 2021−10−13.
基金项目: 国家重点研发计划项目 (2020YFD1100605);国家自然科学基金项目 (61966016); 江西省自然科学基金项目 (20192BAB207018);江西省研究生创新专项基金项目(YC2020-S236).
通信作者: 钱文彬. E-mail: qianwenbin1027@126.com.

效提高决策准确度，降低误分类代价。三支决策理论模拟人类认知、学习和决策的过程，可处理决策过程中出现的不确定性问题。近年来，三支决策理论引起了许多研究者的关注，已成为了粒计算和知识发现领域中的一个重要研究方向。目前，三支决策在众多应用领域中得到广泛的应用，如人脸识别[3]、推荐系统[4-5]、决策系统[6]和邮件过滤[7]等；为了处理复杂的应用场景，提出了不同的计算模型，如序贯三支决策[3,8]、优化三支决策[9]、前景三支决策[10]、三支模糊集[11]和三支约简[12]等。

在实际应用中，代价是影响三支决策划分的重要因素之一。代价敏感学习能够有效缓解分类过程中的数据不平衡问题，其主要作用是处理决策过程和结果产生的各类代价问题。代价敏感学习主要研究两种代价：误分类代价（结果代价）和测试代价，两者互相关联，呈负相关。如在医疗诊断中，患者想要获得更高的诊断准确率（即决策代价越低），就需要做更多的检查（即测试代价越高）。由于代价是数据的内在特征，将其与知识发现结合会使得问题更具有普适性，目前，代价敏感学习已经应用到现实生活中的许多领域，如：人脸识别[13]、价格预测[14]和客户信用评价[15]等。

因此，基于代价敏感的三支决策算法与模型引起了许多学者的关注和研究，已取得重要的研究成果。Fang 等[8]将信息粒度纳入决策分析过程，同时考虑决策过程和决策结果的代价，分别设计了两种不同的算法以最小化决策过程和决策结果代价。Fang 等[16]提出了一种三支决策和可分辨矩阵的框架，在此框架下分别设计了基于删除和增加的代价敏感近似属性约简算法。Jia 等[17]构造了一种可以直接应用于传统的代价敏感学习问题的三支决策模型，在此基础上，提出基于多类三支决策模型的多阶段代价敏感学习方法。Li 等[18]为从输入图像中顺序提取分层粒度结构，提出了一种基于 DNN 的顺序粒度特征提取方法，在此基础上，提出一种代价敏感的序贯三支决策模型。Yang 等[19]考虑了用户需求，提出一种基于模糊粗糙集的序贯三支决策模型的优化机制，用来实现对代价敏感的最优粒度选择。Ma 等[20]定义了三支特定类的最低代价约简，分别设计了基于添加−删除策略和删除策略来构建特定类的最小代价约简算法。以上算法与模型能够最小化结果代价或过程代价。而在许多应用领域中往往需要从代价敏感视角来分析三支决策边界域样本，目前三支决策的研究对象多为单一性数据的决策系统，对于混合数据边界域样本处理的研究相对较少。

为此，本文提出了一种面向混合数据的代价敏感三支决策边界域分类方法。首先，基于正域约简，提出了面向混合数据的属性约简模型；然后，提出了一种基于代价敏感的三支决策边界域样本处理方法，在贝叶斯最小风险的基础上构造误分类代价公式，划分边界域中的对象。最后，对 UCI 上的 10 个数据集进行实验，结果表明该方法能够降低误分类代价，而且能较准确地划分边界域中的对象；这为三支决策的边界域样本处理提供了一种可借鉴的方法。

1　基本知识

1.1　邻域粗糙集

在粗糙集理论[21]中，给定一个四元组决策系统：

$$\mathrm{DS}=\{U,A_t=C\cup D,\{V_a|a\in A_t\},\{I_a|a\in A_t\}\}$$

其中$U=\{x_1,x_2,\cdots,x_n\}$表示有限非空的对象全集，称为论域或者对象空间；A_t表示有限非空的属性全集，由条件属性和决策属性共同组成；$C=\{a_1,a_2,\cdots,a_n\}$表示有限非空的条件属性全集；D表示决策属性；V_a表示$a\in C$的属性值集；$I_a|U\times A_t\to V$是一个信息函数，能给每个对象的每个属性赋值，即$I_a(x)\to V_a$。

定义 1[22]　给定混合邻域决策系统$\mathrm{DN}=\{U,F^D\cup F^C,D,V_a,I_a,\delta\}$，距离度量函数$\Delta N:U\times U$，给定属性子集$B\subseteq C$和邻域参数$\delta$，则对象$x$和$y$基于$B$的邻域关系为

$$\mathrm{NR}^{\delta}(B)=\{(x,y)\in U\times U|\Delta N_B(x,y)\leqslant\delta\}$$

对$\forall x\in U$，x的邻域粒度可表示为

$$\delta_B(x)=\{y|x,y\in U,\Delta N_B(x,y)\leqslant\delta\}$$

式中：F^D为离散属性集合；F^C为连续属性集合；δ是邻域参数。

1.2　三支决策粗糙集

三支决策粗糙集[23]通过 2 个状态集和 3 个动作集来描述其决策过程。其中，状态集$S=\{X,\neg X\}$分别表示对象属于概念X和不属于概念X，动作集$A=\{a_P,a_B,a_N\}$表示对于不同状态分别采取接受、延迟和拒绝 3 种不同的动作。由于采取不同动作会产生不同的损失，记λ_{PP}、λ_{BP}、λ_{NP}表示当$x\in X$时，分别采取动作a_P、a_B和a_N产生的风险损失值；同样地，记λ_{PN}、λ_{BN}、λ_{NN}表示当$x\in\neg X$时，分别采取动作a_P、a_B和a_N产生的风险损失值；损失之间的关系满足：$\lambda_{PP}<\lambda_{BP}<\lambda_{NP}$，$\lambda_{NN}<\lambda_{BN}<\lambda_{PN}$。在实际应用中，这些损失值通过专家的经验获取。

定义 2[1]　在决策系统$\mathrm{DS}=\{U,C\cup D,V_a,I_a\}$中，令$X$为论域$U$基于决策属性$D$的划分，$\alpha$和$\beta$为三支

决策的阈值, $P(X|[x])$表示对象x的条件概率, 对于$\forall x \in U$, 根据贝叶斯决策过程, 计算得到最小成本准则的三支决策规则:

$$\mathrm{POS}(X)=\{x\in U|\alpha \leqslant P(X|[x]) \leqslant 1\}$$
$$\mathrm{BND}(X)=\{x\in U|\beta \leqslant P(X|[x]) \leqslant \alpha\}$$
$$\mathrm{NEG}(X)=\{x\in U|0 \leqslant P(X|[x]) \leqslant \beta\}$$

其中, $P(X|[x])=\dfrac{|X\cap[x]|}{|[x]|}$, $|\cdot|$表示对象的个数;

$$\alpha=\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})-(\lambda_{BP}-\lambda_{PP})}$$
$$\beta=\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})-(\lambda_{NP}-\lambda_{BP})}$$

其中, 正域$\mathrm{POS}(X)$、负域$\mathrm{NEG}(X)$和边界域$\mathrm{BND}(X)$分别对应三支决策规则中的接受、拒绝和不承诺规则, 且满足: $\mathrm{POS}(X)\cup\mathrm{BND}(X)\cup\ \mathrm{NEG}(X)=X$; 仅当$X=U$时, $\mathrm{POS}(X)\cup\mathrm{BND}(X)\cup\mathrm{NEG}(X)=U$。

1.3　代价敏感学习

代价敏感学习主要研究误分类代价和测试代价, 由于本文中考虑了其误分类代价, 误分类代价表示对对象错误划分后的一种惩罚。用$\boldsymbol{C}_{k\times k}$表示误分类代价矩阵, 其中$k$表示$k$分类问题。为方便理解, 以二分类代价矩阵$\boldsymbol{C}_{2\times 2}=\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$为例; 其中$c_{11}$表示将类别为 1 的对象划分到类别 1 中, 因此c_{11}的值为 0, 同理c_{22}的值也为 0; c_{12}表示将类别为 1 的对象划分到类别 2 中, 此时属于误分类, 在划分中需付出惩罚代价, 因此$c_{12}>0$, 同理$c_{21}>0$。

2　基于正域约简的代价敏感三支决策边界域分类方法

2.1　面向混合邻域决策系统的正域约简

由于基于三支决策的粒计算方法大多是处理连续型数据或离散型数据等单一型数据, 但是在现实生活的应用领域中数据类型通常是既含有连续型数据又含有离散型数据的混合数据, 为此需对混合数据的三支决策模型展开研究。

定义 3　给定混合邻域决策系统$\mathrm{DN}=\{U,F^D\cup F^C,D,V_a,I_a,\delta\}$, $V_a(x)$表示对象x在属性a上的属性值:

对于$\forall x,y\in U,\forall a\in F^D$, 则$x$和$y$基于$F^D$的距离为

$$\Delta N_{F^D}(x,y)=\begin{cases} 0, & V_a(x)=V_a(y) \\ 1, & V_a(x)\neq V_a(y) \end{cases}$$

对于$\forall x,y\in U,\forall a\in F^C$, 则$x$和$y$基于$F^C$的距离为

$$\Delta N_{F^C}(x,y)=\left(\sum_{k=1}^{m}|V_a(x)=V_a(y)|^p\right)^{\frac{1}{p}}$$

其中, 当$p=1$时, $\Delta N_{F^C}(x,y)$为曼哈顿距离; 当$p=2$时, $\Delta N_{F^C}(x,y)$为欧氏距离; 当$p\to\infty$时, $\Delta N_{F^C}(x,y)$为切比雪夫距离。

定义 4　给定混合邻域决策系统$\mathrm{DN}=\{U,F^D\cup F^C,D,V_a,I_a,\delta\}$, 令$D_i$为论域$U$基于决策属性$D$的划分, 则混合邻域决策系统的上下近似表示为:

$$\underline{\mathrm{AN}}(D)=\{x\in U|\delta_C(x)\subseteq D_i\}$$
$$\overline{\mathrm{AN}}(D)=\{x\in U|\delta_C(x)\cap D_i\neq\varnothing\}$$

通过上下近似集, 可知特征子集B上的正域如下:

$$\mathrm{POS}_C(D)=\underline{\mathrm{AN}}(D)=\{x\in U|\delta_C(x)\subseteq D_i\}$$

定义 5　给定混合邻域决策系统$\mathrm{DN}=\{U,F^D\cup F^C,D,V_a,I_a,\delta\}$, 令属性$a_i\in C$, 则混合邻域决策系统中基于三支决策的核属性集定义为:

$$\mathrm{CORE}(C)=\{a_i||\mathrm{POS}_C(D)|-|\mathrm{POS}_{C-\{a_i\}}(D)|>0\}$$

以表 1 为例, 给出一个混合邻域决策系统, 其中, $U=\{x_1,x_2,\cdots,x_{10}\}$为对象集, $C=\{a_1,a_2,\cdots,a_6\}$为条件属性集, 决策类$U/D=\{D_1,D_2\}$, 分别为$D_1=\{x_1,x_3,x_5,x_6,x_7,x_9\}$, $D_2=\{x_2,x_4,x_8,x_{10}\}$。

表 1　混合邻域决策系统 DN
Table 1　Hybrid neighborhood decision system DN

U	a_1	a_2	a_3	a_4	a_5	a_6	D
x_1	1	0	0.93	0.73	0.80	0.68	d_1
x_2	1	1	0.64	0.50	0.33	0.48	d_2
x_3	1	1	0.58	0.39	0.38	0.29	d_1
x_4	0	1	0.21	0.12	0.11	0.18	d_2
x_5	1	0	0.63	0.80	0.48	0.58	d_1
x_6	0	0	0.74	0.78	0.42	0.62	d_1
x_7	1	0	0.85	0.80	0.50	0.73	d_1
x_8	1	1	0.50	0.62	0.38	0.44	d_2
x_9	0	0	0.42	0.62	0.49	0.50	d_1
x_{10}	0	0	0.39	0.58	0.29	0.50	d_2

根据定义 5 可计算出混合邻域决策系统的核属性集, 具体的计算过程为: 首先, 根据定义 3, 利用$p=2$时的欧氏距离计算全体对象的混合邻域粒度, 再根据定义 5 计算出$\mathrm{POS}_C(D)=\{x_1,x_4,x_5,x_6,x_7\}$, 同理可计算出$\mathrm{POS}_{C-\{a_1\}}(D)=\{x_1,x_4,x_5,x_6,x_7\}$, 因为$\mathrm{POS}_C(D)=\mathrm{POS}_{C-\{a_1\}}(D)$, 所以属性$a_1\notin\mathrm{CORE}(C)$, 同理可求出$\{a_2,a_3,a_5,a_6\}\notin\mathrm{CORE}(C)$, 只有属性$a_4\in\mathrm{CORE}(C)$。由此可知核属性集为$\mathrm{CORE}(C)=\{a_4\}$。下面将在此基础上, 提出了代价敏感下的三支决策边界域分类方法。

2.2　基于核属性集的代价敏感三支决策边界域分类方法

定义 6　给定混合邻域决策系统$\mathrm{DN}=\{U,F^D\cup F^C,D,V_a,I_a,\delta\}$, 设属性子集$B\subseteq C$, α和β为三支

决策的阈值，D_i表示不同的决策属性，则不同属性子集下的三支决策规则定义为：

$$\mathrm{POS}_B(D_i)=\{x\in U|\alpha\leqslant P(D_i|\delta_B(x))\leqslant 1\}$$
$$\mathrm{BND}_B(D_i)=\{x\in U|\beta<P(D_i|\delta_B(x))<\alpha\}$$
$$\mathrm{NEG}_B(D_i)=\{x\in U|0\leqslant P(D_i|\delta_B(x))\leqslant\beta\}$$

其中，$P(D_i|\delta_B(x))=\dfrac{|D_i\cap\delta_B(x)|}{|\delta_B(x)|}$。

以表 1 为例，可给出混合邻域决策系统代价矩阵，如表 2 所示。结合定义 2 和表 2，可求出三支决策的阈值$\alpha=7/9,\beta=1/3$。

表 2　误分类代价矩阵

Table 2　Misclassification cost matrix

状态/动作	$X(P)$	$\neg X(N)$
a_P	$\lambda_{PP}=0$	$\lambda_{PN}=8$
a_B	$\lambda_{BP}=2$	$\lambda_{BN}=1$
a_N	$\lambda_{NP}=4$	$\lambda_{NN}=0$

令$B=\mathrm{CORE}(C)=\{a_4\}$,根据定义 3 可计算出核属性子集$B$下的对象之间的邻域粒度；再根据定义 6 计算出核属性集下决策类D_1的的正域、负域和边界域，具体的计算过程为：由定义 3 可计算出核属性集B下x_1的邻域粒度$\delta_B(x_1)=\{x_1,x_2,x_5,x_6,x_7,x_8,x_9,x_{10}\}$,由此求出$x_1$的条件概率$P(D_1|\delta_B(x_1))=5/8<\alpha$，所以$x_1\in\mathrm{BND}_B(D_1)$，同理$\{x_2,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}\in\mathrm{BND}_B(D_1)$，即$\mathrm{BND}_B(D_1)=\{x_1,x_2,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$。通过相同的计算可求出：

$$\mathrm{POS}_B(D_1)=\varnothing,\quad \mathrm{NEG}_B(D_1)=\{x_3\}$$

定义 7　在混合邻域决策系统$\mathrm{DN}=\{U,F^D\cup F^C,D,V_a,I_a,\delta\}$中，$D_i$为论域$U$基于决策属性$D$的划分，给定属性子集$B\subseteq C$，为了简化公式，用$\mathrm{CP}^r$和$(1-\mathrm{CP})^r$分别代替$1\Big/P(D_i\big|\delta_B(x_j))$和$1\Big/(1-P(D_i\big|\delta_B(x_j)))$，对于$\forall x_j\in\mathrm{BND}_B(D_i)$，样本简化后的误分类代价计算公式如下：

$$\mathrm{PC}_B(D_i|x_j)=\frac{\mathrm{CP}^r\times\lambda_{PN}}{(\mathrm{CP}^r\times\lambda_{PN})+((1-\mathrm{CP})^r\times\lambda_{NP})}$$
$$\mathrm{NC}_B(D_i|x_j)=\frac{(1-\mathrm{CP})^r\times\lambda_{NP}}{((1-\mathrm{CP})^r\times\lambda_{NP})+(\mathrm{CP}^r\times\lambda_{PN})}$$

其中，$\mathrm{PC}_B(D_i|x)$表示在决策类D_i下将对象x划分到正域产生的误分类代价，同理，$\mathrm{NC}_B(D_i|x)$表示在决策类D_i下将对象x划分到负域产生的误分类代价。λ_{NP}和λ_{PN}是代价矩阵中的风险损失值，$P(D_i|\delta_B(x))$表示在决策类 D_i下对象 x的条件概率。

性质 1　在混合邻域决策系统$\mathrm{DN}=\{U,F^D\cup F^C,D,V_a,I_a,\delta\}$中，$D_i$是对决策属性$D$的划分，假设属性子集$B\subseteq C$，对于$\forall x\in\mathrm{BND}_B(D_i)$，可得出如下推论：

1) 如果$\mathrm{PC}_B(D_i|x)>\mathrm{NC}_B(D_i|x)$，则$x\in\mathrm{NEG}_B(D_i)$；

2) 如果$\mathrm{PC}_B(D_i|x)\leqslant\mathrm{NC}_B(D_i|x)$，则$x\in\mathrm{POS}_B(D_i)$。

以表 1 为例，令$B=\mathrm{Core}(C)=\{a_4\}$，已知$D_1=\{x_1,x_3,x_5,x_6,x_7,x_9\}$和$\mathrm{BND}_B(D_1)=\{x_1,x_2,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$，根据定义 7 和性质 1 可将边界域中的对象划分到正域和负域，具体的计算过程如下：

对于$\forall x\in\mathrm{BND}_B(D_1)$，根据定义 7 可求出划分对象$x_1$产生的两种误分类代价$\mathrm{PC}_B(D_1|x_1)=6/11$，$\mathrm{NC}_B(D_1|x_1)=5/11$，因为$\mathrm{PC}_B(D_1|x_1)>\mathrm{NC}_B(D_1|x_1)$，所以$x_1\in\mathrm{NEG}_B(D_1)$，同理可得$\{x_2,x_4,x_6,x_8,x_9,x_{10}\}\in\mathrm{NEG}_B(D_1)$和$\{x_5,x_7\}\in\mathrm{POS}_B(D_1)$。由此可知，该混合邻域决策系统的正域为$\mathrm{POS}_B(D_1)=\{x_5,x_7\}$，负域为$\mathrm{NEG}_B(D_1)=\{x_1,x_2,x_3,x_4,x_6,x_8,x_9,x_{10}\}$。

3　算法描述及复杂度分析

针对混合邻域决策系统，为了有效划分其三支决策边界域中的对象，本文提出了一种面向混合数据的代价敏感三支决策边界域分类方法，该算法主要分为三个部分。首先，针对混合邻域决策系统中的数据，通过混合邻域计算公式计算每个对象的混合邻域粒度，得到混合邻域决策表的正域对象集合，由此基于启发式策略计算核属性集。其次，在此基础上，计算混合邻域决策表中每个对象的邻域粒度，从而计算出每个对象属于不同决策类的条件概率，利用三支决策规则将对象分别划分到不同决策类的正域、边界域和负域中；最后，针对边界域中的对象，分别计算其划分到正域和负域所产生的误分类代价，通过比较这两种代价的大小，将边界域中的对象划分到正域或负域中，为此，算法的流程图 1 所示。

算法　面向混合数据的代价敏感三支决策边界域分类方法

输入　混合邻域决策系统DN，邻域参数δ和阈值α,β；

输出　核属性集下对不同决策类的正域和负域。

1）对混合邻域决策系统DN做归一化处理；

2）计算决策类$D_i\subseteq U/D$；

3）计算邻域粒度$\delta_C(x)$,初始化$\mathrm{CORE}_C(D)=\varnothing$；

4）对于$\forall x\in U$，若满足$\delta_C(x)\subseteq D_i$，则将对象x存入到正域$\mathrm{POS}_C(D)\leftarrow\mathrm{POS}_C(D)\cup\{x\}$；

5）对于$\forall a_i\in C$，分别计算去除每个对象之后的特征子集的正域集合$\mathrm{POS}_{C-\{a_i\}}(D)$，若满足$\mathrm{POS}_C(D)\neq\mathrm{POS}_{C-\{a_i\}}(D)$，则将属性$a_i$存入到核属性集$\mathrm{CORE}_C(D)\leftarrow\mathrm{CORE}_C(D)\cup\{a_i\}$；

6）基于核属性集$\mathrm{CORE}_C(D)$，计算对象的邻域粒度$\delta_{\mathrm{CORE}_C(D)}(x)$；

7）对于$\forall x\in U$，计算对象x属于决策类D_i的条件概率$P(D_i|\delta_{\mathrm{CORE}_C(D)}(x))$：

①若$\alpha \leqslant P(D_i|\delta_{\mathrm{CORE}_C(D)}(x)) \leqslant 1$, 则将对象$x$划分到决策类$D_i$的正域$\mathrm{POS}_{\mathrm{CORE}_C}(D_i)$;

②否则, 若$0 \leqslant P(D_i|\delta_{\mathrm{CORE}_C(D)}(x)) \leqslant \beta$, 则将对象$x$划分到决策类$D_i$的负域$\mathrm{NEG}_{\mathrm{CORE}_C}(D_i)$;

③否则将对象x划分到决策类D_i的边界域$\mathrm{BND}_{\mathrm{CORE}_C}(D_i)$;

8)对于$\forall x_b \in \mathrm{BND}_{\mathrm{CORE}_C}(D_i)$计算$\mathrm{PC}_{\mathrm{CORE}_C}(D_i|x_j)$和$\mathrm{NC}_{\mathrm{CORE}_C}(D_i|x_j)$:

①若满足$\mathrm{PC}_{\mathrm{CORE}_C}(D_i|x_j) > \mathrm{NC}_{\mathrm{CORE}_C}(D_i|x_j)$, 则将对象$x_j$划分到决策类$D_i$的负域$\mathrm{NEG}_{\mathrm{CORE}_C}(D_i)$;

②否则将对象x_b划分到决策类D_i的正域$\mathrm{POS}_{\mathrm{CORE}_C}(D_i)$;

9)输出划分结果正域$\mathrm{POS}_{\mathrm{CORE}_C}(D_i)$, 负域$\mathrm{NEG}_{\mathrm{CORE}_C}(D_i)$。//算法结束。

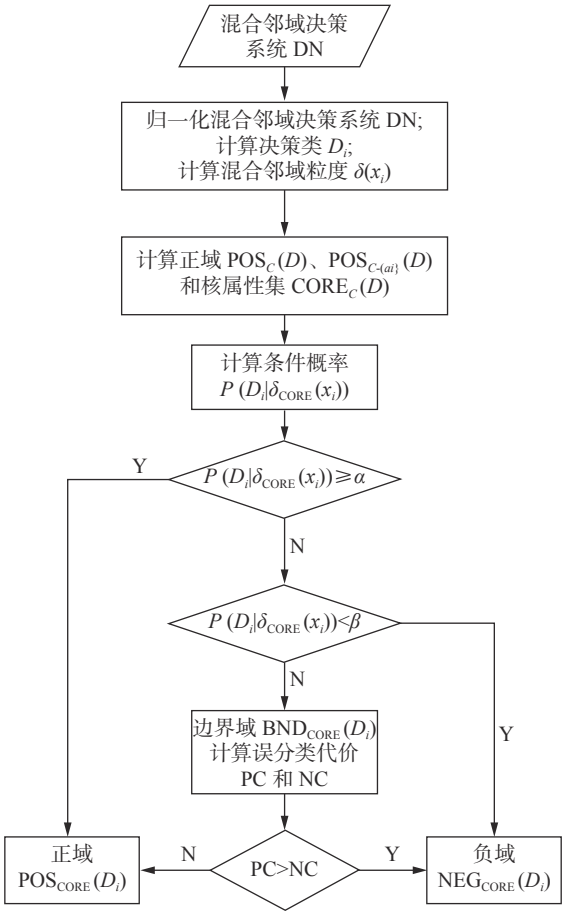


图 1　算法流程图

Fig. 1　The flowchart of algorithm

算法时间复杂度分析:

1)算法的时间复杂度为$O(|U||C|)$; 2)划分决策类所需的时间复杂度为$O(|U|)$; 3)在属性全集下, 通过混合邻域计算公式得出每个对象的混合邻域粒度, 其时间复杂度为$O(|U|^2|C|)$; 4)计算正域对象的时间复杂度为$O(|U|)$; 5)计算核属性集的时间复杂度为$O(|U|^2|C|)$; 6)在核属性集CORE下, 计算每个对象的混合邻域粒度, 其时间复杂度为$O(|U|^2|\mathrm{CORE}_C(D_i)|)$; 7)计算各决策类正域、边界域和负域, 其时间复杂度为$O(|U|)$; 8)结合代价敏感划分边界域中的对象, 其时间复杂度为$O(|\mathrm{BND}_{\mathrm{CORE}_C}(D_i)|)$。综上所述, 算法最坏情况下的时间复杂度是$O(|U|^2|C|)$; 由于存储空间主要用于存放数据, 因此算法的空间复杂度为$O(|U||C|)$。

4　实验比较与分析

为了验证本文方法对边界域对象划分的可行性和有效性, 实验从 UCI 中选取了 10 个混合数据集进行实验测试与分析; 选用分类准确率、权衡因子、误分类损失和时间作为评价指标, 对实验结果进行对比与分析。

4.1　数据集与实验设置

为了更好地说明所提出算法的普适性, 本文根据数据集的来源和规模两个方面, 从国际公开的机器学习 UCI 数据库中选取了 10 个数据集进行实验结果的对比和分析, 数据集的信息描述如表 3 所示。表中 Speaker Accent 和 Ionosphere 数据集中包含连续型数据, Phishing Websites 和 Student Evaluation 数据集中包含离散型数据; 其余数据集均包含连续型和离散型数据; 这些数据集来自欺诈分析、医学诊断、信号处理和教育评价等应用领域。同时为了消除量纲的影响, 对所有数据集中的连续型数据进行归一化处理。本次实验的运行环境为: Win10, Intel(R)Core(TM), i5-6 500 CPU @ 3.20 GHz 3.19 GHz 和 8 GB 内存, 用 Python 编程语言实现算法设计。

4.2　评价指标

实验将从准确率、权衡因子、误分类损失和运行时间 4 种度量指标[24]对划分结果进行分析, 定义如下:

准确率: $\mathrm{Acc} = \dfrac{\mathrm{POS}(D_i) \cap D_i}{\mathrm{POS}(D_i)}$

权衡因子: $F = 2 \times \dfrac{\mathrm{Acc} \times \mathrm{Cov}}{\mathrm{Acc} + \mathrm{Cov}}$

误分类损失: $\mathrm{Cost} = n_b \times \lambda_{bp} + n_n \times \lambda_{np}$

式中: $\mathrm{POS}(D_i)$和D_i表示正域和决策类, n_b和n_n分别表示边界域、负域中的对象个数; λ_{bp}和λ_{np}分别表示将属于某一决策类的对象错误划分到该类别的边界域和负域中产生的损失; 由于本文算法的输出只包含正域和负域, 因此$\mathrm{Cov} = 1$。本实验的风险损失参数为$\lambda_{bp} = 0.3$, $\lambda_{np} = 0.7$。

4.3　实验结果与分析

4.3.1　参数λ_{PN}和λ_{NP}对划分结果的影响

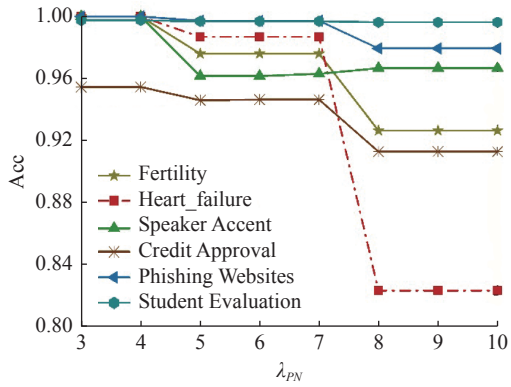
在混合邻域决策系统中, 参数λ_{PN}和λ_{NP}通过影

响阈值对(α,β)的大小来影响三支决策的划分。因此，为了详细分析参数λ_{PN}和λ_{NP}的值对划分准确度的影响。本小节中，为了一般性，从上述数据集中选取 6 个作为代表进行实验分析，分别将λ_{PN}和λ_{NP}的值从 3 到 10，且每次步长变化 1 进行实验。实验结果如图 2 所示。

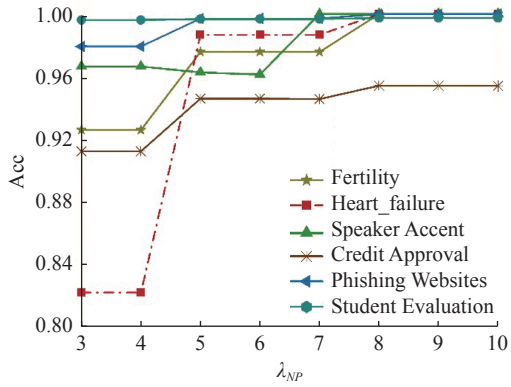
表 3　数据集的基本信息

Table 3　Basic information of the data set

数据集	样本	属性类别		决策类	决策类的样本数
		连续属性	离散属性		
Fertility	100	2	7	2	{88, 12}
Heart failure	299	7	5	2	{96, 203}
Speaker Accent	329	12	0	6	{29, 30, 30, 30, 45, 165}
Ionosphere	351	34	0	2	{225, 126}
Thoraric Surgery	470	2	14	2	{400, 70}
Credit Approval	690	7	8	2	{307, 383}
Audit Data	773	11	15	2	{305, 468}
Diabetic	1 151	16	3	2	{540, 611}
Phishing Websites	2 456	0	30	2	{1 094, 1 362}
Student Evaluation	5 820	0	33	5	{812,560,1 612,1 695,1 141}



(a) 代价 λ_{PN} 对划分准确率的影响



(b) 代价 λ_{NP} 对划分准确率的影响

图 2　参数λ_{PN}和λ_{NP}对准确率的影响

Fig. 2　Influence of parameters λ_{PN} and λ_{NP} on the accuracy

在图 2(a) 中，当λ_{PN}的取值区间在[4,5]时，Credit Approval 等 5 个数据集的准确率随代价的增加而下降，且变化趋势较为平缓；当λ_{PN}的取值区间在[7,8]时，这些数据集的准确率随代价的增加而下降，且变化趋势较为显著。在图 2(b) 中，当λ_{NP}的取值区间在[4,5]时，Credit Approval 等 5 个数据集的准确率随代价的增加而上升，且变化趋势较为显著；当λ_{NP}的取值在[6,7]区间时，数据集 Speaker Accent 的准确率随代价的增加而升高，进而达到平稳状态；当λ_{NP}的取值在[7,8]时，Credit Approval 等个 5 数据集的准确率随代价的增加而升高，且变化趋势较为平缓；当代价λ_{PN}和λ_{NP}的取值在 [8,10] 时，准确率达到平稳状态，所有数据集的准确率不再随着代价的变化而变化。

综上所述，从整体上看，代价λ_{PN}和λ_{NP}对分类准确度的影响呈负相关，数据集的准确率随着代价λ_{PN}的增加，呈现出整体下降的趋势；而随着代价λ_{NP}的增加，整体呈现上升的趋势。从局部上看，当代价的取值在[4,5]和[7,8]这两个区间时，数据集的准确率随着代价的增加而发生变化，当代价的取值在其他区间时，数据集的准确率趋于稳定的状态。由此，在实际的决策过程中，可结合上述分析的结论，并根据数据集的分布和代价敏感学习构造合适的代价矩阵。

4.3.2　本文模型与不同三支决策模型的对比分析

本节主要分析不同三支决策模型对分类性能的影响，表 4~7 给出了 3 种粗糙集模型下准确率 Acc、权衡因子 F、误分类损失 Cost 和运行时间 Time 的实验结果。其中，NCTM (neighborhood rough set based cost-sensitive three-way decision boundary region processing model) 是基于邻域粗糙集[25] 设计考虑了代价敏感的三支决策边界域处理模型，PCTM (pawlak rough set based cost-sensitive three-way decision boundary region processing

model) 是基于经典粗糙集[22]设计考虑了代价敏感的三支决策边界域处理模型, MCTM (mixed-neighborhood rough set based cost-sensitive three-way decision boundary region processing model) 代表本文基于混合邻域粗糙集的代价敏感三支决策边界域处理模型。在 PCTM 模型中对数据集进行离散化预处理, 在 NCTM 和 MCTM 模型中对数据集进行了归一化预处理, 另外, 为了使距离处于同一量纲下, 在 NCTM 模型中采取平均距离度量, 同时 Acc、F、Cost 和 Time 的值均为数据集所有决策类的平均值。实验结果如表 4~7 所示, 其中, 符号↑表示度量指标的值越大越好, 符号↓表示度量指标的值越小越好, 加粗字体表示算法在所对应的数据集上的最优值。

表 4　三种粗糙集模型的准确率 Acc(↑)对比

Table 4　Comparison of accuracy Acc(↑) under three kinds of rough set models

数据集	NCTM	PCTM	MCTM
Fertility	0.9318	0.9385	**1.0000**
Heart_failure	0.8863	0.9646	**1.0000**
Speaker Accent	**1.0000**	0.8558	**1.0000**
Ionosphere	0.9241	**0.9884**	0.9241
Thoraric Surgery	0.9314	0.9345	**0.9962**
Credit Approval	0.9144	0.9326	**0.9620**
Audit Data	0.9370	**0.9870**	0.9750
Diabetic	0.8582	**0.8965**	0.8590
Phishing Websites	0.9709	0.9897	**0.9996**
Student Evaluation	0.6582	0.9257	**0.9979**

如表 4 所示, 使用本文模型的分类准确率高于其他 2 种模型, 例如, 其在数据集 Credit Approval 上的准确率比 NCTM 和 PCTM 模型分别提高了 4.8% 和 2.9%, 由于 MCTM 能够针对不同的数据类型采取不同的分类方法, 且具有更低的错误率, 因此其划分准确率能整体上高于 NCTM 和 PCTM。此外, 在数据集 Ionosphere 上, PCTM 模型的优势更加明显, 而在数据集 Speaker Accent 上, 本文模型和 NCTM 模型的准确率相同, 由此可知, 本文模型能有效地提高分类准确率, 且在数据集上整体表现良好。

如表 5 所示, 对权衡因子而言, 由其度量公式可知, 权衡因子由准确率和覆盖率共同决定, 由于本文中的三支决策最终转换成二支决策, 因此覆盖率Cov = 1, 在本文中权衡因子 F 很大程度上取决于准确率 Acc 的值。对比表 4 和表 5 的实验结果可知, 权衡因子 F 的值略高于准确率 Acc 的值, 但是整体上的变化趋势和 Acc 相同。

表 5　3 种粗糙集模型的权衡因子F(↑)对比

Table 5　Comparison of trade-off factor F(↑) under three kinds of rough set models

数据集	NCTM	PCTM	MCTM
Fertility	0.9647	0.9676	**1.0000**
Heart_failure	0.9393	0.9819	**1.0000**
Speaker Accent	**1.0000**	0.9220	**1.0000**
Ionosphere	0.9602	**0.9941**	0.9602
Thoraric Surgery	0.9632	0.9661	**0.9981**
Credit Approval	0.9553	0.9651	**0.9806**
Audit Data	0.9664	**0.9935**	0.9872
Diabetic	0.9174	**0.9448**	0.9234
Phishing Websites	0.9852	0.9948	**0.9998**
Student Evaluation	0.7616	0.9599	**0.9989**

如表 6 所示, 使用本文模型的误分类损失整体上明显低于其他 2 种模型, 例如, 在数据集 Student Evaluation 中, 本文模型的误分类损失比 NCTM 和 PCTM 分别降低了 478.1 和 287.0。从不同的模型角度分析, 针对混合邻域决策系统, PCTM 对划分的要求较为苛刻, 而 NCTM 对划分的要求较于放松, 容错率低, 导致划分错误率提高; 本文模型 MCTM 可灵活应用于不同类型的决策系统, 容错率高, 所以具有更低的误分类代价。

表 6　3 种粗糙集模型的误分类损失 Cost (↓) 对比

Table 6　Comparison of misclassification loss Cost (↓) under three kinds of rough set models

数据集	NCTM	PCTM	MCTM
Fertility	4.2000	1.4000	**0.0000**
Heart_failure	20.3000	7.0000	**0.0000**
Speaker Accent	**0.0000**	23.8000	**0.0000**
Ionosphere	13.3000	**3.5000**	13.3000
Thoraric Surgery	44.1000	21.7000	**2.1000**
Credit Approval	33.6000	28.7000	**16.8000**
Audit Data	46.2000	**7.0000**	14.7000
Diabetic	**13.3000**	23.8000	32.2000
Phishing Websites	42.0000	16.8000	**0.7000**
Student Evaluation	485.1000	294.0000	**7.0000**

如表 7 所示, 从整体上看, 3 种粗糙集粒计算模型所消耗的时间较少且随着数据规模的增大而增多; 从部分上看, NCTM 模型耗时相对较长, 主要是由于 NCTM 是用邻域关系计算邻域类, 每两个对象之间都要计算, 导致其时间复杂度较高。而 PCTM 模型和 MCTM 模型在耗时方面差异性不大, 且差异性随数据规模的增大而减小。

综上所述, 与其他 2 种不同的粗糙集模型进行实验对比和分析可知, 本文模型总体上具有较高的分类准确度和较低的误分类损失, 因此, 用其对混合邻域决策系统进行划分较为合理。

表 7　3 种粗糙集模型的运行时间 Time(↓) 对比

Table 7　Comparison of operation hours Time(↓) under three kinds of rough set models

数据集	NCTM	PCTM	MCTM
Fertility	0.1467	**0.0185**	0.1112
Heart_failure	1.4253	**0.1860**	1.1989
Speaker Accent	1.5592	**0.2207**	1.5177
Ionosphere	1.6507	**0.2217**	1.4853
Thoraric Surgery	3.9171	**0.9209**	2.7756
Credit Approval	7.9367	**1.1905**	6.7061
Audit Data	9.0060	**1.6071**	6.6358
Diabetic	30.0278	**10.4657**	26.6360
Phishing Websites	135.5500	16.6810	**14.1239**
Student Evaluation	868.6762	82.2837	**73.5983**

4.3.3　本文模型和序贯三支决策模型的边界域分类方法对比

为了进一步验证本文模型的有效性，本小节将本文模型与序贯三支决策的方法进行实验对比和分析。其中，MSTM (mixed-neighborhood rough set based sequential three-way decision boundary region processing model) 是基于经典序贯三支决策[8]改造的基于混合邻域粗糙集的序贯三支决策边界域处理模型。实验结果如表 8 所示，分别给出了 MCTM 和 MSTM 的分类准确度、权衡因子、误分类损失和时间的对比。

由表 8 的实验结果可知，在数据集 Ionosphere 和 Audit Data 上，本文模型 MCTM 的分类性能与 MSTM 相同，而在另外 8 个数据集上，本文模型 MCTM 的分类性能要优于序贯三支决策模型 MSTM。从理论上分析，由于 MSTM 直接由代价矩阵计算的阈值划分边界域对象，而本文在此基础上进一步考虑条件概率和误分类代价来划分边界域中的对象，因此本文模型 MCTM 在 Acc、F、Cost 和 Time 上表现较优。为此，在同等条件下，对于混合邻域决策系统，本文基于属性约简的混合代价敏感三支决策边界域分类方法为处理边界域对象提供了一种可借鉴的分析方法。

表 8　不同边界域处理模型的实验结果对比

Table 8　Comparison of experimental results of different boundary domain processing models

数据集	Acc(↑)		F(↑)		Cost(↓)		Time(↓)	
	MCTM	MSTM	MCTM	MSTM	MCTM	MSTM	MCTM	MSTM
Fertility	**1.0000**	0.9880	**1.0000**	0.9939	**0.0000**	1.4000	0.1112	**0.1092**
Heart_failure	**1.0000**	0.9854	**1.0000**	0.9926	**0.0000**	2.8000	**1.1989**	1.2184
Speaker Accent	**1.0000**	0.9889	**1.0000**	0.9943	**0.0000**	0.7000	**1.5177**	1.5807
Ionosphere	**0.9241**	**0.9241**	**0.9602**	**0.9602**	**13.3000**	**13.3000**	**1.4853**	1.50835
Thoraric Surgery	**0.9962**	0.9924	**0.9981**	0.9962	**2.1000**	4.2000	2.7756	**2.6520**
Credit Approval	**0.9620**	0.9540	**0.9806**	0.9764	**16.8000**	20.9999	**6.7061**	6.9013
Audit Data	**0.9750**	**0.9750**	**0.9872**	**0.9872**	**14.7000**	**14.7000**	**6.6358**	6.7636
Diabetic	**0.8590**	0.8408	**0.9234**	0.9131	**32.2000**	42.0000	26.6360	**24.9780**
Phishing Websites	**0.9996**	0.9979	**0.9998**	0.9989	**0.7000**	3.5000	**14.1239**	14.3478
Student Evaluation	**0.9979**	0.9971	**0.9989**	0.9985	**7.0000**	9.1000	**73.5983**	74.0003

5　结束语

近年来三支决策理论成为热点研究问题，其研究对象多为单一型决策系统，然而，在许多的应用领域中，数据往往呈现混合类型的特点，目前三支决策对混合数据边界域样本处理的研究相对较少。为划分混合决策系统中的边界域对象，本文提出了基于混合数据的属性约简方法；并在此基础上，提出了一种基于核属性的代价敏感三支决策边界域分类方法。通过在不同的数据集上进行实验对比与分析，验证了本文方法的可行性和有效性，获得了一种相对合理的边界域对象的划分方法。由于序贯三支决策更加符合现实生活中的决策过程及人类的认知，下一步工作将研究基于代价敏感的序贯三支决策的粒化问题。

参考文献：

[1] YAO Yiyu. Three-way decisions with probabilistic rough sets[J]. Information sciences, 2010, 180(3): 341–353.

[2] YAO Yiyu. Three-way decision and granular computing[J]. International journal of approximate reasoning, 2018, 103: 107–123.

[3] LI Huaxiong, ZHANG Libo, HUANG Bing, et al. Sequential three-way decision and granulation for cost-sens-

itive face recognition[J]. Knowledge-based systems, 2016, 91: 241–251.

[4] ZHANG Hengru, MIN Fan, SHI Bing, et al. Regression-based three-way recommendation[J]. Information sciences, 2017, 378: 444–461.

[5] HUANG Jiajin, WANG Jian, YAO Yiyu, et al. Cost-sensitive three-way recommendations by learning pair-wise preferences[J]. International journal of approximate reasoning, 2017, 86: 28–40.

[6] CHEN Yufei, YUE Xiaodong, FUJITA H, et al. Three-way decision support for diagnosis on focal liver lesions [J]. Knowledge-based systems, 2017, 127: 85–99.

[7] ZHOU Bing, YAO Yiyu, LUO Jigang. Cost-sensitive three-way email spam filtering[J]. Journal of intelligent information systems, 2014, 42(1): 19–45.

[8] FANG Yu, GAO Cong, YAO Yiyu. Granularity-driven sequential three-way decisions: a cost-sensitive approach to classification[J]. Information sciences, 2020, 507: 644–664.

[9] LIU Jiubing, LI Huaxiong, ZHOU Xianzhong, et al. An optimization-based formulation for three-way decisions [J]. Information sciences, 2019, 495: 185–214.

[10] WANG Tianxing, LI Huaxiong, ZHOU Xianzhong, et al. A prospect theory-based three-way decision model[J]. Knowledge-based systems, 2020, 203: 106129.

[11] YAO Yiyu, WANG Shu, DENG Xiaofei. Constructing shadowed sets and three-way approximations of fuzzy sets[J]. Information sciences, 2017, 412-413: 132–153.

[12] MA Xi'ao, YAO Yiyu. Three-way decision perspectives on class-specific attribute reducts[J]. Information sciences, 2018, 450: 227–245.

[13] WAN Jianwu, WANG Yi. Cost-sensitive label propagation for semi-supervised face recognition[J]. IEEE transactions on information forensics and security, 2019, 14(7): 1729–1743.

[14] MA Chao, LIU Zhenbing, CAO Zhiguang, et al. Cost-sensitive deep forest for price prediction[J]. Pattern recognition, 2020, 107: 107499.

[15] XIAO Jin, ZHOU Xu, ZHONG Yu, et al. Cost-sensitive semi-supervised selective ensemble model for customer credit scoring[J]. Knowledge-based systems, 2020, 189: 105118.

[16] FANG Yu, MIN Fan. Cost-sensitive approximate attribute reduction with three-way decisions[J]. International journal of approximate reasoning, 2019, 104: 148–165.

[17] JIA Xiuyi, LI Weiwei, SHANG Lin. A multiphase cost-sensitive learning method based on the multiclass three-way decision-theoretic rough set model[J]. Information sciences, 2019, 485: 248–262.

[18] LI Huaxiong, ZHANG Libo, ZHOU Xianzhong, et al. Cost-sensitive sequential three-way decision modeling using a deep neural network[J]. International journal of approximate reasoning, 2017, 85: 68–78.

[19] YANG Jie, WANG Guoying, ZHANG Qinghua, et al. Optimal granularity selection based on cost-sensitive sequential three-way decisions with rough fuzzy sets[J]. Knowledge-based systems, 2019, 163: 131–144.

[20] MA Xi'ao, ZHAO Xuerong. Cost-sensitive three-way class-specific attribute reduction[J]. International journal of approximate reasoning, 2019, 105: 153–174.

[21] PAWLAK Z, SKOWRON A. Rough sets: some extensions[J]. Information sciences, 2007, 177(1): 28–40.

[22] HU Qinghua, YU Daren, XIE Zhongxia. Neighborhood classifiers[J]. Expert systems with applications, 2008, 34(2): 866–876.

[23] YAO Y Y, WONG S K M. A decision theoretic framework for approximating concepts[J]. International journal of nan-machine studies, 1992, 37(6): 793–809.

[24] XU Yi, TANG Jingxin, WANG Xusheng. Three sequential multi-class three-way decision models[J]. Information sciences, 2020, 537: 62–90.

[25] HU Qinghua, YU Daren, LIU Jinfu, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information sciences, 2008, 178(18): 3577–3594.

作者简介：

周阳阳，硕士研究生，主要研究方向为粒计算与知识发现。

钱文彬，副教授，博士，主要研究方向为知识发现与机器学习。主持国家自然科学基金项目 2 项、江西省自然科学基金项目 2 项。发表学术论文 30 余篇。

王映龙，教授，博士，主要研究方向为知识发现与数据挖掘。参与国家自然科学基金项目 2 项，主持江西省自然科学基金项目 3 项。发表学术论文 20 余篇。