# 不完备混合型数据的决策粗糙集与三支决策分类算法

王 光 琼
(四川文理学院智能制造学院 四川 达州 635000)

**摘 要** 决策粗糙集是目前粗糙集理论的重要研究分支。目前的决策粗糙集很少对不完备混合型的信息系统进行研究,为了改善这一局限,提出一种扩展的决策粗糙集模型。通过引入邻域容差关系来处理不完备混合型信息系统,在其基础上定义扩展的决策粗糙集模型,同时提出相应的三支决策。在该模型的基础上设计一种最小化决策代价的属性约简算法。根据三支决策,构建出一种不完备混合型数据的三支决策分类算法。实验结果表明,该算法具有更高的数据分类精度和更小的误分类代价。

**关键词** 粗糙集 决策粗糙集 不完备混合型数据 三支决策 属性约简 分类

**中图分类号** TP18 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.11.040

# DECISION-THEORETIC ROUGH SET AND THREE-WAY DECISIONS CLASSIFICATION ALGORITHMS FOR INCOMPLETE MIXED DATA

Wang Guangqiong
(*School of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou 635000, Sichuan, China*)

**Abstract** Decision-theoretic rough set is an important branch of rough set theory. However, at present, decision-theoretic rough sets rarely study incomplete mixed information systems. In order to improve this limitation, this paper presents an extended decision-theoretic rough set model. The incomplete mixed information system was dealt with by introducing the neighborhood tolerance relation; an extended decision-theoretic rough set model was defined based on it, and the corresponding three-way decisions were also proposed; based on the proposed model, an attribute reduction algorithm was designed to minimize the decision cost; according to the proposed three-way decisions, a three-way decisions classification algorithm with incomplete mixed data was constructed. The simulation results show that the proposed classification algorithm has higher data classification accuracy and lower misclassification cost.

**Keywords** Rough set Decision-theoretic rough set Incomplete mixed data Three-way decisions Attribute reduction Classification

## 0 引 言

粗糙集理论[1]是当今人工智能和知识发现领域的一种重要模型,在处理不确定性和不完备性的数据方面发挥着尤为重要的作用。传统的粗糙集模型基于等价关系建立,通过等价关系对信息系统进行划分来达到不确定性概念的粗糙近似。然而传统的粗糙集模型对噪声数据较为敏感[1-4],不具有较好的泛化能力,为了改善这一局限,提出了一种称之为决策粗糙集的模型。

决策粗糙集最早由加拿大学者 Yao 等[5]提出,将贝叶斯决策理论融入传统粗糙集模型中,使得其最终的粗糙近似结果具有最小的决策代价,其中决策粗糙集的上下近似通过一对阈值来限定,相比传统的粗糙集模型,该模型对噪声数据具有更好的容忍效果。在决策粗糙集模型的基础上,Yao[6]进一步地提出了三支决策理论,建立了不确定性数据环境下一种新的决策方法。为了提高决策粗糙集模型的应用范围,学者们进行了大量的改进和推广,例如:在分布式数据集

收稿日期:2019-07-21。"2019 年中国物流学会/中国物流与采购联合会面上研究课题计划"面上项目(2019CSLKT3-231);四川省教育厅重点项目(18ZA0421)。**王光琼**,高级实验师,主研领域:软件工程,人工智能理论。

下,Lin 等[7]提出了一种多源信息系统的决策粗糙集模型;在不完备信息系统下,Liu 等[8]提出了一种适用不完备数据的改进决策粗糙集模型;Zhao 等[9]针对多值信息系统提出一种扩展的决策粗糙集;Feng 等[10]提出一种变精度的多粒度决策粗糙集模型;在模糊数据的环境下,Sun 等[11]提出了模糊集的决策粗糙集模型以及相关应用;Zhao 等[12]在其基础上进一步地提出了模糊区间值的决策粗糙集模型;刘久兵等[13]提出了直觉模糊信息系统的决策粗糙集模型。另一方面,数值型数据也是一种常见的数据类型,Li 等[14]提出了基于邻域关系的决策粗糙集模型。因此可以看出,目前决策粗糙集模型的研究已不断趋于完善。

混合性和不完备性是目前数据的一个典型特征,对于粗糙集理论,学者们对这种类型的数据也进行了广泛的研究[15-17]。然而目前的决策粗糙集模型还未对这种类型的数据进行探索,因此本文将在前人研究的基础上提出一种不完备混合型信息系统的决策粗糙集模型。

对于不完备混合型信息系统,Zhao 等[15]定义了邻域容差关系,对这类信息系统进行了有效处理。本文采用邻域容差关系重新对传统的决策粗糙集模型进行重构,提出不完备混合型信息系统下的决策粗糙集模型,同时基于该模型进一步地提出相应的三支决策。此外,基于最小化决策代价的原则,本文对于所提出的决策粗糙集模型设计出一种最小化决策代价的属性约简算法。另一方面,由于三支决策提供了一种新的决策思维,本文将其融入分类模型中,提出一种不完备混合型数据的三支决策分类算法,该分类算法将样本对象的类决策结果分成三种情况,比传统的分类算法增加延迟分类的情形,即对于不确定性的样本对象进行延迟处理。仿真实验结果表明,所提出的三支决策分类算法可以有效地降低分类结果的误分类代价,提高分类精度,具有更高的分类性能。

## 1　决策粗糙集模型

在粗糙集理论中,数据集表示成信息系统的形式,一个信息系统可表示为 $S=(U,At=C\cup D,V)$,其中:$U$ 为该信息系统 $S$ 的论域,即数据集的样本空间;$At$ 为信息系统的属性集;$C$ 为条件属性集;$D$ 为决策数据集;$V$ 为整个信息系统的属性值集域,根据 $V$ 中属性值的类型,通常可以将信息系统分为离散型信息系统、连续型信息系统以及混合型信息系统。通常信息系统也可简单表示为 $S=(U,At=C\cup D)$。

**定义 1**[1]　对于离散型信息系统 $S=(U,At=C\cup D)$,基于属性子集 $B\subseteq C$ 构建的等价关系 $E_B$ 定义为:

$$E_B=\{(x,y)\in U\times U\mid a(x)=a(y),\forall a\in B\}\quad(1)$$

式中:$a(x)$ 表示对象 $x$ 在属性 $a$ 下的属性值。根据等价关系 $E_B$,可以得到任意对象的等价类,即对象 $x$ 的等价类表示为 $[x]_B=\{y\in U\mid(x,y)\in E_B\}$。

在决策粗糙集模型[5-6]中,一般假设 $\Omega=\{X,\widetilde{X}\}$ 表示两种目标决策结果,那么对象 $x$ 关于对象集 $X$ 的概率可表示为:

$$p(X\mid[x])=\frac{|X\cap[x]|}{|[x]|}\quad(2)$$

类似地,对象 $x$ 与对象集 $\widetilde{X}$ 的概率可表示为 $p(\widetilde{X}\mid[x])=1-p(X\mid[x])$,其中 $\widetilde{X}$ 表示 $X$ 的补集。同时定义对象 $x$ 被分类入 $X$ 三个区域的动作表示为 $\Gamma=\{a_P,a_B,a_N\}$。设 $\Omega=\{X,\widetilde{X}\}$ 与 $\Gamma=\{a_P,a_B,a_N\}$ 之间的代价结果如表 1 所示。

**表 1　决策代价**

| 动作 | $X$ | $\widetilde{X}$ |
|---|---|---|
| $a_P$ | $\lambda_{PP}$ | $\lambda_{PN}$ |
| $a_B$ | $\lambda_{BP}$ | $\lambda_{BN}$ |
| $a_N$ | $\lambda_{NP}$ | $\lambda_{NN}$ |

表 1 中:$\lambda_{PP}$、$\lambda_{BP}$和 $\lambda_{NP}$表示对象 $x$ 原本属于 $X$ 采取 $a_P$、$a_B$和 $a_N$三种动作时的代价,$\lambda_{PN}$、$\lambda_{BN}$和 $\lambda_{NN}$表示对象 $x$ 原本属于 $\widetilde{X}$ 采取 $a_P$、$a_B$和 $a_N$三种动作时的代价。这里的代价结果通常满足 $\lambda_{PP}\leqslant\lambda_{BP}\leqslant\lambda_{NP}$且$\lambda_{NN}\leqslant\lambda_{BN}\leqslant\lambda_{PN}$。此外,表 1 所示的决策代价结果可以表示成矩阵的形式,记为决策代价矩阵:

$$\boldsymbol{C}=\begin{bmatrix}\lambda_{PP} & \lambda_{PN}\\ \lambda_{BP} & \lambda_{BP}\\ \lambda_{NP} & \lambda_{NN}\end{bmatrix}$$

根据贝叶斯决策理论,Yao 等通过最小化决策代价的原则,利用代价矩阵推导出一对阈值来进行粗糙集模型中粗糙近似的计算,使得近似的结果拥有最小的误分类代价,该模型即为决策粗糙集模型。

**定义 2**[5]　给定信息系统 $S=(U,AT)$,属性集 $A\subseteq AT$在信息系统 $S$ 下确定的等价关系为 $E_A$,设信息系统的决策代价矩阵为 $\boldsymbol{C}$,对于 $X\subseteq U$ 关于 $E_A$的决策粗糙集下近似$\underline{E}_A^{\theta^+}(X)$和决策粗糙集上近似$\overline{E}_A^{\theta^-}(X)$分别定义为:

$$\begin{cases}\underline{E}_A^{\theta^+}(X)=\{x\in U\mid p(X\mid[x]_A)>\theta^+\}\\ \overline{E}_A^{\theta^-}(X)=\{x\in U\mid p(X\mid[x]_A)\geqslant\theta^-\}\end{cases}\quad(3)$$

式中：$\theta^+ = \max\{\alpha,\beta,\gamma\}$；$\theta^- = \min\{\alpha,\beta,\gamma\}$。其中：

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

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

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

在决策粗糙集模型基础上，Yao 等进一步地提出了三支决策模型。对于一个决策对象 $x$，利用三支决策进行的决策行为可以描述为：

(1) 若 $p(X|[x]) > \theta^+$，则 $x$ 判定为 $X$；

(2) 若 $\theta^- < p(X|[x]) \leqslant \theta^+$，则 $x$ 待进一步分析；

(3) 若 $p(X|[x]) \leqslant \theta^-$，则 $x$ 判定为 $\tilde{X}$。

## 2 不完备混合型信息系统下的决策粗糙集模型

### 2.1 邻域容差关系

在粗糙集理论中，传统的模型都基于完备离散型的信息系统而建立，而现实环境下的数据类型是复杂多样，不完备混合型的信息系统便是其中常见的一种。Hu 等[18]通过在连续型数据下建立邻域关系，从而解决粗糙集理论对连续型以及混合型信息系统的处理。Kryszkiewicz[19]提出一种基于容差关系的扩展粗糙集模型，解决了不完备信息系统下的粗糙集近似。在两位学者的基础上，Zhao 等[15]提出了邻域容差关系用于对不完备混合型信息系统的处理。

**定义 3**[15] 给定不完备混合型信息系统 $S = (U, AT)$，设属性集 $A \subseteq AT$ 满足 $A = A_D \cup A_N$，其中 $A_D$ 和 $A_N$ 分别表示 $A$ 下的离散型属性集和连续型属性集，那么属性集 $A$ 在不完备混合型信息系统下确定的邻域容差关系定义为：

$$NT_A^\delta = \{(x,y) \in U \times U | \forall a \in A, (a(x) = *) \vee (a(y) = *) \vee ((a \in A_D \rightarrow a(x) = a(y)) \wedge (a \in A_N \rightarrow d(x,y) \leqslant \delta))\} \quad (4)$$

式中：$d(x,y)$ 表示对象 $x$ 与 $y$ 之间的距离度量[18]；$\delta$ 为邻域半径，是一个非负常数；$a(x)$ 表示对象 $x$ 在属性 $a$ 下的属性值，$a(x) = *$ 表示属性值为缺失的情形。

根据邻域容差关系，可以对整个不完备混合型信息系统的论域诱导出一组邻域容差粒化，邻域容差粒化的结果将是不完备混合型信息系统进行粗糙逼近的基础。

**定义 4**[15] 给定不完备混合型信息系统 $S = (U, AT)$，设混合类型属性集 $A \subseteq AT$ 在信息系统 $S$ 下确定的邻域容差关系为 $NT_A^\delta$，其中 $\delta$ 为邻域半径，那么 $\forall x \in U$ 基于邻域容差关系 $NT_A^\delta$ 的邻域容差类定义为：

$$\delta_A(x) = \{y \in U | (x,y) \in NT_A^\delta\} \quad (5)$$

同时，论域 $U$ 上所有对象邻域容差类构成的集合 $GS_A = \{\delta_A(x_1), \delta_A(x_2), \cdots, \delta_A(x_{|U|})\}$ 称为该信息系统的一个粒结构。显然，$GS_A$ 为论域 $U$ 上的一个覆盖。

### 2.2 模型设计

在 Zhao 等提出的邻域容差关系基础上，将经典的决策粗糙集进行推广，提出不完备混合型信息系统下的决策粗糙集模型，同时相应的三支决策也被提出。

在邻域量化容差关系中，将对象 $x \in U$ 的邻域容差类 $\delta(x)$ 看成与该对象属于同一类的对象集，因此在不完备混合型信息系统中，对于一个对象 $x$ 隶属于某个对象集 $X$ 的概率可表示为：

$$P(X|\delta(x)) = \frac{|X \cap \delta(x)|}{|\delta(x)|} \quad (6)$$

基于该定义框架，本文构造了不完备混合型信息系统的决策粗糙集模型以及相应的三支决策。

类似于传统的决策粗糙集模型，同样假设对象 $x$ 的两种决策结果 $\Omega = \{X, \tilde{X}\}$，其中 $\tilde{X}$ 表示 $X$ 的补集。根据邻域容差类的含义，对象 $x$ 隶属于对象集 $X$ 的程度为 $p(X|\delta(x))$；对象 $x$ 隶属于对象集 $\tilde{X}$ 的程度为 $1 - p(X|\delta(x))$。

定义对象 $x$ 关于 $X$ 的三个分类动作表示为集合 $\Gamma = \{a_P, a_B, a_N\}$，其中：$a_P$ 表示对象 $x$ 分类入 $X$ 的正区域；$a_B$ 表示对象 $x$ 分类入 $X$ 的边界域；$a_N$ 表示对象 $x$ 分类入 $X$ 的负区域。设结果集 $\Omega = \{X, \tilde{X}\}$ 与动作集 $\Gamma = \{a_P, a_B, a_N\}$ 之间的代价矩阵为 $\boldsymbol{C}$。其中 $\lambda_{PP}$、$\lambda_{BP}$ 和 $\lambda_{NP}$ 分别表示对象 $x$ 原本属于 $X$ 时采取 $a_P$、$a_B$ 和 $a_N$ 三种动作所产生的代价，即 $\lambda_{PP}$、$\lambda_{BP}$ 和 $\lambda_{NP}$ 表示对象 $x$ 原本属于 $X$ 时分类入 $X$ 正区域、边界域和负区域所产生的代价，同理 $\lambda_{PN}$、$\lambda_{BN}$ 和 $\lambda_{NN}$ 表示对象 $x$ 原本属于 $\tilde{X}$ 时分类入 $\tilde{X}$ 正区域、边界域和负区域所产生的代价。

根据贝叶斯决策理论，对于对象 $x$、结果集 $\Omega = \{X, \tilde{X}\}$ 和给定分类代价矩阵 $\boldsymbol{C}$，可以得到对象 $x$ 采取三种动作时的预期代价，记对象 $x$ 分类入 $X$ 正区域、边界域和负区域的预期代价分别为 $R_P^\delta$、$R_B^\delta$ 和 $R_N^\delta$，$\delta$ 为邻域半径。则有：

$$\begin{cases} R_P^\delta = \lambda_{PP} \cdot p(X|\delta(x)) + \lambda_{PN} \cdot p(\sim X|\delta(x)) \\ R_B^\delta = \lambda_{BP} \cdot p(X|\delta(x)) + \lambda_{BN} \cdot p(\sim X|\delta(x)) \\ R_N^\delta = \lambda_{NP} \cdot p(X|\delta(x)) + \lambda_{NN} \cdot p(\sim X|\delta(x)) \end{cases} \quad (7)$$

因此,可以得到如下三种最小代价规则:

(1) 若 $R_P^\delta \leqslant R_B^\delta$ 且 $R_P^\delta \leqslant R_N^\delta$,则 $x \in POS^\delta(X)$;

(2) 若 $R_B^\delta \leqslant R_P^\delta$ 且 $R_B^\delta \leqslant R_N^\delta$,则 $x \in BUN^\delta(X)$;

(3) 若 $R_N^\delta \leqslant R_P^\delta$ 且 $R_N^\delta \leqslant R_B^\delta$,则 $x \in NEG^\delta(X)$。

$POS^\delta(X)$、$BUN^\delta(X)$ 和 $NEG^\delta(X)$ 分别表示 $X$ 的 $\delta$ 正区域、边界域和负区域。

一般情况下,对象 $x$ 原本属于 $X$ 时分类入 $X$ 正区域的代价要小于分类入 $X$ 边界域的代价,分类入 $X$ 边界域的代价要小于分类入 $X$ 负区域的代价。类似地,对象 $x$ 原本属于 $\widetilde{X}$ 时分类入 $\widetilde{X}$ 负区域的代价要小于分类入 $\widetilde{X}$ 边界域的代价,分类入 $\widetilde{X}$ 边界域的代价要小于分类入 $\widetilde{X}$ 正区域的代价。即分类代价满足关系 $\lambda_{PP} \leqslant \lambda_{BP} \leqslant \lambda_{NP}$ 且 $\lambda_{NN} \leqslant \lambda_{BN} \leqslant \lambda_{PN}$,同时:

$$P(X|\delta(x)) + P(\widetilde{X}|\delta(x)) = 1$$

所以,可以进一步得到:

1) 对于 $R_P^\delta \leqslant R_B^\delta$ 且 $R_P^\delta \leqslant R_N^\delta$,即:

$$\lambda_{PP} \cdot p(X|\delta(x)) + \lambda_{PN} \cdot (1 - p(X|\delta(x))) \leqslant \lambda_{BP} \cdot p(X|\delta(x)) + \lambda_{BN} \cdot (1 - p(X|\delta(x))) \Rightarrow p(X|\delta(x)) \geqslant \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}$$

且:

$$\lambda_{PP} \cdot p(X|\delta(x)) + \lambda_{PN} \cdot (1 - p(X|\delta(x))) \leqslant \lambda_{NP} \cdot p(X|\delta(x)) + \lambda_{NN} \cdot (1 - p(X|\delta(x))) \Rightarrow p(X|\delta(x)) \geqslant \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}$$

2) 对于 $R_B^\delta \leqslant R_P^\delta$ 且 $R_B^\delta \leqslant R_N^\delta$,即:

$$\lambda_{BP} \cdot p(X|\delta(x)) + \lambda_{BN} \cdot (1 - p(X|\delta(x))) \leqslant \lambda_{PP} \cdot p(X|\delta(x)) + \lambda_{PN} \cdot (1 - p(X|\delta(x))) \Rightarrow p(X|\delta(x)) \leqslant \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}$$

且:

$$\lambda_{BP} \cdot p(X|\delta(x)) + \lambda_{BN} \cdot (1 - p(X|\delta(x))) \leqslant \lambda_{NP} \cdot p(X|\delta(x)) + \lambda_{NN} \cdot (1 - p(X|\delta(x))) \Rightarrow p(X|\delta(x)) \geqslant \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}$$

3) 对于 $R_N^\delta \leqslant R_P^\delta$ 且 $R_N^\delta \leqslant R_B^\delta$,即:

$$\lambda_{NP} \cdot p(X|\delta(x)) + \lambda_{NN} \cdot (1 - p(X|\delta(x))) \leqslant \lambda_{PP} \cdot p(X|\delta(x)) + \lambda_{PN} \cdot (1 - p(X|\delta(x))) \Rightarrow p(X|\delta(x)) \leqslant \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}$$

且:

$$\lambda_{NP} \cdot p(X|\delta(x)) + \lambda_{NN} \cdot (1 - p(X|\delta(x))) \leqslant \lambda_{BP} \cdot p(X|\delta(x)) + \lambda_{BN} \cdot (1 - p(X|\delta(x))) \Rightarrow p(X|\delta(x)) \leqslant \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}$$

这里令:

$$\begin{cases} \alpha = \dfrac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})} \\ \beta = \dfrac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})} \\ \gamma = \dfrac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})} \end{cases} \tag{8}$$

则有:

(1) 若 $p(X|\delta(x)) \geqslant \alpha$ 且 $p(X|\delta(x)) \geqslant \gamma$,那么 $x \in POS^\delta(X)$;

(2) 若 $p(X|\delta(x)) \leqslant \alpha$ 且 $p(X|\delta(x)) \geqslant \beta$,那么 $x \in BUN^\delta(X)$;

(3) 若 $p(X|\delta(x)) \leqslant \beta$ 且 $p(X|\delta(x)) \leqslant \gamma$,那么 $x \in NEG^\delta(X)$。

特别地,若代价函数满足如下关系:

$$\frac{\lambda_{NP} - \lambda_{BP}}{\lambda_{BN} - \lambda_{NN}} > \frac{\lambda_{BP} - \lambda_{PP}}{\lambda_{PN} - \lambda_{BN}}$$

那么此时有 $0 \leqslant \beta < \gamma < \alpha \leqslant 1$,则上述三个规则即为:

(1) 若 $p(X|\delta(x)) \geqslant \alpha$,则 $x \in POS^\delta(X)$;

(2) 若 $\beta \leqslant p(X|\delta(x)) \leqslant \alpha$,则 $x \in BUN^\delta(X)$;

(3) 若 $p(X|\delta(x)) \leqslant \beta$,则 $x \in NEG^\delta(X)$。

根据以上推导的这三条规则,可以直接得到不完备混合型信息系统下的决策粗糙集模型,同时也蕴含了不完备混合型信息系统下三支决策。

**定义 5**　给定不完备混合型信息系统 $S = (U, AT)$,混合类型属性集 $A \subseteq AT$ 在信息系统 $S$ 下确定的邻域容差关系为 $NT_A^\delta$,$\delta$ 为邻域半径,设信息系统的代价矩阵为 $\boldsymbol{C}$。对于 $X \subseteq U$ 关于邻域容差关系 $NT_A^\delta$ 的决策粗糙集下近似和决策粗糙集上近似分别定义为:

$$\begin{cases} \underline{NT_A^{\theta^+}}(X) = \{x \in U | p(X|\delta_A(x)) > \theta^+\} \\ \overline{NT_A^{\theta^-}}(X) = \{x \in U | p(X|\delta_A(x)) \geqslant \theta^-\} \end{cases} \tag{9}$$

式中:$\theta^+ = \max\{\alpha, \beta, \gamma\}$;$\theta^- = \min\{\alpha, \beta, \gamma\}$。同时 $X$ 关于邻域容差关系 $NT_A^\delta$ 的决策粗糙集正区域、边界域和负区域分别表示为:

$$\begin{cases} POS_A^{\theta^+}(X) = \underline{N}_A^{\theta^+}(X) \\ BUN_A^{(\theta^-, \theta^+)}(X) = \overline{N}_A^{\theta^-}(X) - \underline{N}_A^{\theta^+}(X) \\ NEG_A^{\theta^-}(X) = U - \overline{N}_A^{\theta^-}(X) \end{cases} \tag{10}$$

另一方面,对于不完备混合型决策信息系统 $S = (U, AT = C \cup D)$,其中 $D$ 为决策属性,设 $\frac{U}{D} = \{D_1, D_2, \cdots, D_m\}$ 表示信息系统论域 $U$ 在 $D$ 下的划分,$D_i$ $(1 \leqslant i \leqslant m)$ 表示每个决策类,即数据样本的每个类。那么

整个决策属性 $D$ 关于混合型属性集 $A\subseteq C$ 的决策粗糙集正区域、边界域和负区域分别表示为：

$$\begin{cases} POS_A^{\theta^+}\left(\dfrac{U}{D}\right)=\bigcup_{i=1}^{m}POS_A^{\theta^+}(D_i) \\ BUN_A^{(\theta^-,\theta^+)}\left(\dfrac{U}{D}\right)=U-\left(POS_A^{\theta^+}\left(\dfrac{U}{D}\right)\cup NEG_A^{\theta^-}\left(\dfrac{U}{D}\right)\right. \\ NEG_A^{\theta^-}\left(\dfrac{U}{D}\right)=\bigcup_{i=1}^{m}NEG_A^{\theta^-}\left(\dfrac{U}{D}\right) \end{cases} \tag{11}$$

在定义 5 中，通过 $\alpha$、$\beta$、$\gamma$ 三个阈值来确定不完备混合型信息系统下的粗糙集近似。基于语义的视角，在本文所提出决策粗糙集模型中，$POS_A^{\theta^+}(X)$ 表示论域中“可以”分类入 $X$ 的对象集；$BUN_A^{(\theta^-,\theta^+)}(X)$ 表示论域中“有可能”分类入 $X$ 的对象集；$NEG_A^{\theta^-}(X)$ 表示论域中“不可以”分类入 $X$ 的对象集，这种分类的可能程度通过 $\alpha$、$\beta$、$\gamma$ 来体现，由于 $\alpha$、$\beta$、$\gamma$ 基于分类代价得到，因此决策粗糙集的分类使得最终产生的代价最小。

另一方面，根据本文所提出决策粗糙集模型的三个区域划分，这里便得到了不完备混合型信息系统下的三支决策。对于目标决策结果 $X$ 和待决策的对象 $x$，那么：

（1）若对象 $x$ 满足目标决策结果 $X$ 的决策条件，即 $p(X|\delta(x))>\theta^+$，那么接受对象 $x$ 判定为 $X$，即 $x\in POS^{\theta^+}(X)$；

（2）若对象 $x$ 不满足目标决策结果 $X$ 的决策条件，即 $p(X|\delta(x))<\theta^-$，那么拒绝对象 $x$ 判定为 $X$，即 $x\in NEG^{\theta^-}(X)$；

（3）若对象 $x$ 不确定是否满足目标决策结果 $X$ 的决策条件，即 $\theta^-\leqslant p(X|\delta(x))\leqslant\theta^+$，那么延迟对象 $x$ 的判定，即 $x\in BUN^{(\theta^-,\theta^+)}(X)$，待得到更多信息后再进行确定。

**性质 1** 考虑不完备混合型信息系统 $S=(U,AT)$，混合型属性集 $A\subseteq AT$，阈值满足 $\theta_1^-\leqslant\theta_2^-<\theta_2^+\leqslant\theta_1^+$，$\delta$ 为邻域半径，对于 $X\subseteq U$ 关于邻域容差关系 $NT_A^{\delta}$ 的决策粗糙集下近似和决策粗糙集上近似满足：

$$POS_A^{\theta_1^+}(X)\subseteq POS_A^{\theta_2^+}(X) \tag{12}$$

$$NEG_A^{\theta_1^-}(X)\subseteq NEG_A^{\theta_2^-}(X) \tag{13}$$

$$BUN_A^{(\theta_2^-,\theta_2^+)}(X)\subseteq BUN_A^{(\theta_1^-,\theta_1^+)}(X) \tag{14}$$

证明：

（1）对于 $\forall x\in U$，当 $\theta_2^+\leqslant\theta_1^+$ 有：

$$p(X|\delta_A(x))>\theta_1^+\Rightarrow p(X|\delta_A(x))>\theta_2^+$$

即 $\underline{N}_A^{\theta_1^+}(X)\subseteq\underline{N}_A^{\theta_2^+}(X)$，亦即：

$$POS_A^{\theta_1^+}(X)\subseteq POS_A^{\theta_2^+}(X)$$

因此（1）成立。

（2）对于 $\forall x\in U$，当 $\theta_1^-\leqslant\theta_2^-$ 有：

$$p(X|\delta_A(x))\geqslant\theta_2^-\Rightarrow p(X|\delta_A(x))\geqslant\theta_1^-$$

即 $\overline{N}_A^{\theta_2^-}(X)\subseteq\overline{N}_A^{\theta_1^-}(X)$，由于 $NEG_A^{\theta_1^-}(X)=U-\overline{N}_A^{\theta_1^-}$，$NEG_A^{\theta_2^-}(X)=U-\overline{N}_A^{\theta_2^-}$ 所以 $NEG_A^{\theta_1^-}(X)\subseteq NEG_A^{\theta_2^-}(X)$。则（2）成立。

（3）由于：

$$BUN_A^{(\theta_1^-,\theta_1^+)}(X)=\overline{N}_A^{\theta_1^-}(X)-\underline{N}_A^{\theta_1^+}(X)$$

$$BUN_A^{(\theta_2^-,\theta_2^+)}(X)=\overline{N}_A^{\theta_2^-}(X)-\underline{N}_A^{\theta_2^+}(X)$$

综合（1）、（2）可以得到：

$$BUN_A^{(\theta_2^-,\theta_2^+)}(X)\subseteq BUN_A^{(\theta_1^-,\theta_1^+)}(X)$$

因此（3）成立。

证毕。

性质 1 表明了本文所提出决策粗糙集三个区域的单调性。基于三支决策的视角，性质 1 中的式（12）表明当决策的接受阈值 $\theta^+$ 越大，即接受的决策条件越为苛刻，那么最终可接受的对象越少。式（13）表明当决策的拒绝阈值 $\theta^-$ 越小，即拒绝的决策条件越为苛刻，那么最终拒绝的对象越少。式（14）表明接受决策阈值越大且拒绝决策阈值越小时，即接受决策和拒绝决策都比较严格时，那么延迟决策的程度就比较宽松。相反，接受决策阈值越小且拒绝决策阈值越大时，即接受决策和拒绝决策都比较宽松时，那么延迟决策的程度就比较严格，这表现出了两种不同的决策态度。由于决策粗糙集中的阈值 $\theta^-$ 和 $\theta^+$ 直接由分类代价矩阵直接确定，那么代价的取值不同就决定了三支决策的决策态度。

## 2.3 属性约简

属性约简是粗糙集理论的重要研究内容，在决策粗糙集模型中，基于最小代价的属性约简是目前的研究热点[20-22]。

**定义 6** 给定不完备混合型决策信息系统为 $S=(U,AT=C\cup D)$，设邻域半径为 $\delta$，由决策代价矩阵确定的一对阈值分别为 $\theta^-$ 和 $\theta^+$。决策属性 $D$ 关于属性集 $A\subseteq C$ 的决策正区域、边界域和负区域分别为 $POS_A^{\theta^+}\left(\dfrac{U}{D}\right)$、$BUN_A^{(\theta^-,\theta^+)}\left(\dfrac{U}{D}\right)$ 和 $NEG_A^{\theta^-}\left(\dfrac{U}{D}\right)$，那么属性集 $A$ 下的决策代价定义为：

$$\begin{aligned} Cost_A=&\sum_{x\in POS_A^{\theta^+}\left(\frac{U}{D}\right)}(1-p_x)\cdot\lambda_{PN}+\\ &\sum_{y\in BUN_A^{(\theta^-,\theta^+)}\left(\frac{U}{D}\right)}(p_y\cdot\lambda_{BP}+(1-p_y)\cdot\lambda_{BN})+\\ &\sum_{z\in NEG_A^{\theta^-}\left(\frac{U}{D}\right)}p_z\cdot\lambda_{NP} \end{aligned} \tag{15}$$

式中：$p_t=\dfrac{|[t]_D\cap\delta_A(t)|}{|\delta_A(t)|}$，$[t]_D$ 表示对象 $t$ 在决策属性

$D$ 下的等价类。

基于最小化代价的属性约简定义如下:

**定义 7**　给定不完备混合型决策信息系统为 $S=(U,AT=C\cup D)$,设邻域半径为 $\delta$,由决策代价确定的一对阈值分别为 $\theta^-$ 和 $\theta^+$。若属性子集 $A\subseteq C$ 是该信息系统的最小代价属性约简,那么当且仅当:

(1) $Cost_A\leqslant Cost_C$;

(2) $\forall A'\subset A, Cost_{A'}>Cost_A$。

在定义 7 中,条件(1)表明属性约简集的决策代价小于属性全集的决策代价;条件(2)展示了属性约简集决策代价的极小性,即属性约简集的决策代价在所有属性子集中是最小的。

启发式搜索是寻找信息系统约简集的一种常用方法,其中启发式函数的构造是该方法的核心。本节将通过决策代价 *Cost* 构造出一种属性约简的启发式函数。

给定不完备混合型决策信息系统为 $S=(U,AT=C\cup D)$,设属性集 $A\subseteq C$,对于 $\forall a\in A$ 关于属性集 $A$ 的属性重要度定义为:

$$sig_A(a)=\frac{Cost_{A-\{a\}}-Cost_A}{Cost_A} \tag{16}$$

利用 $sig_A(a)$ 作为启发式函数设计出的最小代价属性约简算法如算法 1 所示。

**算法 1**　不完备混合型信息系统下决策粗糙集模型的最小代价属性约简

输入:不完备混合型信息系统 $S=(U,AT=C\cup D)$;决策代价矩阵 $\boldsymbol{C}$,邻域半径 $\delta$。

输出:属性约简集 $R$。

**步骤 1**　初始化 $R=\varnothing$。

**步骤 2**　对于 $\forall a\in C$,计算 $a$ 的属性重要度 $sig_C(a)$,并将属性集 $C$ 按照属性重要度从大到小进行排序,排序后的属性集记为 $C'$。

**步骤 3**　选择属性集 $C'$ 中属性重要度最大的属性 $a_t$,若 $Cost_{R\cup\{a_t\}}>Cost_C$,那么进行 $C'\leftarrow C'-\{a_t\}$ 且 $R\leftarrow R\cup\{a_t\}$,并重新进入步骤 3,若 $Cost_{R\cup\{a_t\}}\leqslant Cost_C$,那么 $R\leftarrow R\cup\{a_t\}$ 并进入步骤 4。

**步骤 4**　对于属性集 $\forall r\in R$,若满足关系 $Cost_{R-\{r\}}\leqslant Cost_R$,那么进行 $R\leftarrow R-\{r\}$。

**步骤 5**　返回结果 $R$。

## 3　不完备混合型信息系统的三支决策分类算法

Hu 等[23]通过邻域粗糙集模型构造出了混合型数据的邻域分类算法,实验证明该算法具有较好的分类效果。本文在该分类算法的基础上,将三支决策思想融入其中,提出基于三支决策方法的数据分类模型。

三支决策是在经典的贝叶斯决策模型基础上的推广,它将决策对象的决策结果分成三个部分,分别为接受、拒绝和延迟,确定这三种决策结果则通过决策粗糙集模型中的阈值 $\theta^-$ 和 $\theta^+$ 来实现。把数据的分类也看成对数据类别的决策,因此利用三支决策模型来用于数据的分类,可以描述成如下形式:

对于二分类问题,设一个训练样本集为 *Data*,其中样本包含两种类别,分别记为正类别和负类别,并且 *Data* 中正类别样本集表示为 $D^+$,负类别样本集表示为 $D^-$。对于一个待标记类别的样本对象 $x$,$\delta(x)$ 为对象 $x$ 在样本集 *Data* 中的邻域类,那么基于三支决策模型对象 x 的判定规则为:

(1) 若 $p(D_+|\delta(x))>\theta^+$,$x$ 判定为正类别;

(2) 若 $p(D_+|\delta(x))\leqslant\theta^-$,$x$ 判定为负类别;

(3) 若 $\theta^-<p(D_+|\delta(x))\leqslant\theta^+$,$x$ 判定结果待进一步确定。

对于多分类情形,可以不断将其转换成多个二分类问题进行处理,因此基于三支决策模型的多分类判定规则为:

(1) 若 $p(D_{\max}|\delta(x))>\theta^+$,$x$ 判定为 $D_{\max}$;

(2) 若 $p(D_{\max}|\delta(x))\leqslant\theta^-$,$x$ 不判定为任何类;

(3) 若 $\theta^-<p(D_{\max}|\delta(x))\leqslant\theta^+$,$x$ 判定结果待进一步确定。

这里的 $p(D_{\max}|\delta(x))=\max\limits_{D_i\in\frac{U}{D}}p(D_i|\delta(x))$。

根据如上判定规则,不完备混合型信息系统的三支决策分类算法如算法 2 所示。

**算法 2**　不完备混合型信息系统的三支决策分类算法

输入:不完备混合型信息系统 $S=(U,AT=C\cup D)$,决策代价矩阵为 $\boldsymbol{C}$,邻域半径 $\delta$,$S$ 中的类别划分 $\frac{U}{D}$,待分类对象 $x$。

输出:对象 $x$ 的类别。

**步骤 1**　根据决策代价矩阵 $\boldsymbol{C}$ 计算决策阈值 $\theta^-$ 和 $\theta^+$。

**步骤 2**　根据算法 1 对原信息系统 $S$ 进行最小化代价属性约简,得到约简结果 $R$。

**步骤 3**　计算对象 $x$ 在论域 $U$ 中属性集 $R$ 下的邻域类 $\delta_R(x)$。

**步骤 4**　判断 $p(D_{\max}|\delta_R(x))$ 与 $\theta^+$ 之间的关系:

1) 若 $p(D_{\max}|\delta_R(x))>\theta^+$,那么 $x$ 判定为 $D_{\max}$;

2) 若 $\theta^-<p(D_{\max}|\delta_R(x))\leqslant\theta^+$,那么 $x$ 判定结果待进一步确定;

3) 若 $p(D_{\max}|\delta_R(x))\leqslant\theta^-$ 那么 $x$ 不判定为任何类。

**步骤 5**　返回对象 $x$ 的类别。

## 4 实 验

表 2 为实验中所使用的数据集,这 10 个数据集均来源于 UCI 机器学习数据库,其中:Mushroom 为只包含离散型属性的数据集;Wine、Sonar 和 Musk 为只包含连续型属性的数据集;其余为混合型属性的数据集。部分数据集为完备型的数据集,本实验选择其中 5% 的属性值进行删除,从而构造出不完备的数据集,同时,为了避免连续型属性量纲带来的影响,在实验前将所有数据集的连续型属性标准化至[0,1]区间。

**表 2 实验数据集**

| 数据集 | 对象 | 属性 | 类别 | 类型 |
|---|---|---|---|---|
| Mushroom | 5 644 | 22 | 3 | 离散型 |
| Wine | 178 | 13 | 2 | 连续型 |
| Sonar | 208 | 60 | 2 | 连续型 |
| Musk | 6 598 | 166 | 2 | 连续型 |
| Credit | 690 | 15 | 2 | 混合型 |
| Annealing | 798 | 38 | 6 | 混合型 |
| German | 1 000 | 19 | 2 | 混合型 |
| Abalone | 4 177 | 8 | 29 | 混合型 |
| Thyroid | 9 172 | 29 | 2 | 混合型 |
| Inconme | 48 842 | 14 | 2 | 混合型 |

### 4.1 实验设置

在本文提出的三支决策分类算法中,决策代价矩阵发挥着很重要的作用,实验采用在[0,1]之间取随机值的方法进行选取,选取的决策代价满足如下关系:

$$\begin{cases}\lambda_{PP}=0 \text{ 且 } 0<\lambda_{BP}<\lambda_{NP}\\ \lambda_{NN}=0 \text{ 且 } 0<\lambda_{BN}<\lambda_{PN}\\ \lambda_{NP}-\lambda_{BP}>\lambda_{BP}-0 \text{ 且 } \lambda_{PN}-\lambda_{BN}>\lambda_{BN}-0\end{cases} \tag{17}$$

本实验将所提出的三支决策分类算法与支持向量机分类算法(SVM)、决策树分类算法(C4.5)、朴素贝叶斯分类算法(NB)和邻域粗糙集分类算法[23](NRSC)进行实验比较,其中比较结果通过分类精度 Acc、F 度量和误分类 MCost 代价来体现,计算式表示为:

$$\begin{cases}Acc=\dfrac{n_{PP}}{n_{PP}+n_{NP}}\\ F=2\cdot\dfrac{Acc\cdot Cov}{Acc+Cov} \quad Cov=\dfrac{n_{PP}+n_{NP}}{n_{PP}+n_{NP}+n_{BP}}\\ MCost=n_{NP}\cdot\lambda_{NP}+n_{BP}\cdot\lambda_{BP}\end{cases} \tag{18}$$

式中:$n_{PP}$表示被分类正确的对象数;$n_{NP}$表示被错误分类的对象数;$n_{BP}$表示被待定的对象数。

### 4.2 邻域半径的选取

在本文所提出的三支决策分类算法中,邻域半径是一个较为关键的参数,它的取值不同对最终的实验结果将产生很大的影响,因此在进行实验之前需要对邻域半径的大小进行确定。由于连续型属性已标准化至[0,1]区间,本实验将邻域半径 $\delta$ 在区间[0,0.3]中以 0.02 为间隔分别进行取值,将选取的每个值对所有数据集进行十折交叉分类,这样便得到对应的分类精度结果。图 1 为每个数据集在不同邻域半径下得到分类精度结果。

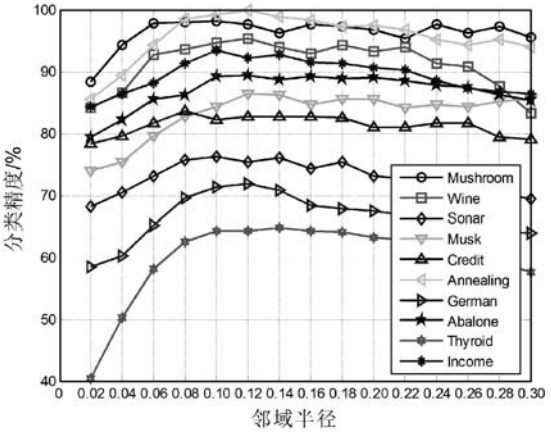


图 1 不同邻域半径下各个数据集的分类精度

观察图 1 的实验结果,可以发现当邻域半径选取为 0.10 时可以得到较好的分类结果,因此本实验选择 $\delta=0.10$ 进行实验。

### 4.3 结果分析

表 3 为本文的三支决策分类算法与 SVM、C4.5、NB 和 NRSC 算法对每个数据集通过十折交叉法得到的分类精度 Acc,结果通过“均值 ± 标准差”来表示,最高的分类精度已用粗体表示。观察表 3 的实验结果可以发现,本文算法在大部分数据集下拥有最高的分类精度;SVM 在少部分数据集下拥有最高的分类精度,例如数据集 Sonar 和 Credit;NB 算法在数据集 German 中拥有最高的分类精度。因此本文算法具有更高的分类准确度,这主要是由于分类机制的差异导致的。SVM 拥有较好的分类性能,但是它是一种二支分类模型,即对象的分类的结果只有两种,标记为特定的类或不标记为特定的类,对于处于类之间的对象,可能会出现误分类情形。而本文算法对于确定的对象直接进行分类,对于类与类之间的模糊对象,通过进行延迟处理的方式,减少误分类的情况,因而在大部分数据集下拥有更高的分类精度。

**表 3　分类精度 Acc 比较结果**

| 数据集 | 算法 | | | | |
|---|---|---|---|---|---|
| | SVM | C4.5 | NB | NRSC | 本文算法 |
| Mushroom | 0.955 9 ±0.015 4 | 0.974 2 ±0.011 6 | 0.945 3 ±0.015 7 | 0.952 0 ±0.017 4 | **0.982 4 ±0.004 2** |
| Wine | 0.938 6 ±0.009 2 | 0.917 9 ±0.014 7 | 0.924 6 ±0.009 8 | 0.914 7 ±0.010 3 | **0.946 5 ±0.018 8** |
| Sonar | **0.784 4 ±0.006 9** | 0.761 5 ±0.012 6 | 0.772 1 ±0.003 5 | 0.751 9 ±0.005 4 | 0.762 4 ±0.012 3 |
| Musk | 0.833 9 ±0.004 5 | 0.817 3 ±0.006 6 | 0.823 8 ±0.004 2 | 0.835 3 ±0.008 9 | **0.842 6 ±0.007 6** |
| Credit | **0.842 9 ±0.003 1** | 0.824 5 ±0.007 2 | 0.802 9 ±0.006 1 | 0.831 4 ±0.002 1 | 0.822 5 ±0.005 7 |
| Annealing | 0.974 0 ±0.011 6 | 0.958 5 ±0.014 8 | 0.983 4 ±0.013 9 | 0.967 3 ±0.003 2 | **0.992 4 ±0.032 5** |
| German | 0.724 9 ±0.006 8 | 0.713 6 ±0.010 3 | **0.734 2 ±0.008 5** | 0.705 5 ±0.006 9 | 0.712 4 ±0.004 6 |
| Abalone | 0.876 1 ±0.005 9 | 0.855 7 ±0.007 5 | 0.849 3 ±0.007 3 | 0.854 1 ±0.009 9 | **0.892 4 ±0.002 5** |
| Thyroid | **0.663 7 ±0.024 1** | 0.658 4 ±0.010 3 | 0.634 9 ±0.014 2 | 0.624 8 ±0.012 7 | 0.642 4 ±0.009 0 |
| Inconme | 0.924 7 ±0.003 8 | 0.908 6 ±0.012 5 | 0.926 9 ±0.005 8 | 0.908 5 ±0.007 4 | **0.933 5 ±0.006 9** |

表 4 给出了三支决策分类算法与 SVM、C4.5、NB 和 NRSC 算法对每个数据集进行分类的 F 度量结果,其中最高的结果值已用粗体表示。观察表 4 可以发现,SVM 分类算法在大部分数据集下拥有最高的度量值,而本文算法在所有数据集中都拥有较小的 F 度量结果,这主要是由于参与比较的分类算法对待分类的对象都进行了具体的类别判定,不存在延迟判定的情况,即 $n_{BP}=0$,因此 $Cov$ 始终等于 1,而本文算法会对有的对象进行延迟判别,因而 $n_{BP}\geqslant 0$,那么 $Cov\leqslant 1$,因此 F 值会偏小。

**表 4　F 度量比较结果**

| 数据集 | 算法 | | | | |
|---|---|---|---|---|---|
| | SVM | C4.5 | NB | NRSC | 本文算法 |
| Mushroom | **0.952 9 ±0.011 7** | 0.941 8 ±0.009 8 | 0.911 9 ±0.007 4 | 0.924 5 ±0.007 9 | 0.897 3 ±0.002 8 |
| Wine | 0.953 1 ±0.003 8 | 0.932 6 ±0.016 5 | **0.964 6 ±0.005 9** | 0.942 6 ±0.006 0 | 0.925 7 ±0.028 5 |
| Sonar | **0.857 2 ±0.005 4** | 0.846 1 ±0.008 2 | 0.829 5 ±0.002 6 | 0.847 4 ±0.003 8 | 0.811 7 ±0.013 8 |
| Musk | 0.879 6 ±0.003 9 | **0.904 9 ±0.004 9** | 0.885 9 ±0.003 7 | 0.867 2 ±0.005 2 | 0.854 3 ±0.008 7 |
| Credit | **0.795 6 ±0.002 3** | 0.772 5 ±0.004 6 | 0.743 8 ±0.004 2 | 0.764 5 ±0.002 4 | 0.734 4 ±0.015 3 |
| Annealing | **0.892 5 ±0.008 2** | 0.864 1 ±0.013 4 | 0.876 4 ±0.008 9 | 0.856 2 ±0.003 5 | 0.847 2 ±0.023 6 |
| German | 0.823 7 ±0.005 4 | **0.862 2 ±0.007 4** | 0.840 6 ±0.005 7 | 0.852 7 ±0.007 2 | 0.802 8 ±0.016 2 |
| Abalone | 0.784 5 ±0.006 5 | 0.761 9 ±0.006 0 | **0.804 1 ±0.007 4** | 0.776 9 ±0.006 1 | 0.745 3 ±0.009 7 |
| Thyroid | **0.743 8 ±0.011 4** | 0.736 9 ±0.017 9 | 0.726 5 ±0.008 2 | 0.709 3 ±0.010 7 | 0.694 8 ±0.014 5 |
| Inconme | **0.869 6 ±0.002 1** | 0.840 6 ±0.005 3 | 0.857 4 ±0.003 7 | 0.825 7 ±0.008 5 | 0.813 1 ±0.021 8 |

表 5 为所有算法对每个数据集分类结果的误分类代价 MCost 比较结果,其中最低的误分类代价 MCost 已用粗体表示。观察表 5 可以发现,本文算法在所有的数据集下都拥有最小的误分类代价,其他分类算法的误分类代价都高于本文算法。主要原因是本文算法增加延迟分类的判别结果,处于类边界的对象进行延迟决策,减少误分类的情况,而其他传统的分类算法对这类情形可能会将判别对象分类入其他错误的类,而进行延迟分类的代价要小于错误分类的代价,因此本文算法的误分类代价要更小。

**表 5　误分类代价 MCost 比较结果**

| 数据集 | 算法 | | | | |
|---|---|---|---|---|---|
| | SVM | C4.5 | NB | NRSC | 本文算法 |
| Mushroom | 199.120 0 ±23.894 5 | 116.490 ±13.979 | 246.980 ±29.638 | 216.730 ±19.008 | **79.601 0 ±20.875 6** |
| Wine | 8.743 4 ±1.049 2 | 11.691 0 ±2.402 9 | 10.737 0 ±1.288 4 | 12.147 0 ±3.457 6 | **5.713 8 ±2.976 4** |
| Sonar | 35.876 0 ±4.305 1 | 39.686 0 ±5.762 4 | 37.923 0 ±4.550 7 | 41.284 0 ±5.954 1 | **29.652 0 ±6.539 8** |
| Musk | 876.740 0 ±76.214 6 | 964.36 ±89.72 | 930.05 ±111.61 | 869.35 ±94.32 | **622.880 0 ±104.865 9** |
| Credit | 86.719 ±10.406 | 96.876 ±17.625 | 108.800 ±13.056 | 93.067 ±11.168 | **73.3190 0 ±19.906 7** |
| Annealing | 16.598 0 ±1.991 8 | 26.494 0 ±3.179 2 | 10.597 0 ±1.271 7 | 20.876 0 ±4.505 1 | **7.638 9 ±1.876 3** |
| German | 220.080 0 ±26.412 7 | 229.120 ±29.494 | 212.640 ±25.517 | 235.530 ±28.263 | **182.560 0 ±30.766 9** |
| Abalone | 414.020 0 ±41.683 4 | 482.190 ±67.863 | 503.580 ±50.429 | 487.240 ±48.469 | **359.670 0 ±45.125 7** |
| Thyroid | 2 467.600 0 ±186.128 5 | 2 506.50 ±200.78 | 2 679.00 ±171.47 | 2 753.10 ±230.37 | **2 067.900 0 ±157.257 4** |
| Inconme | 2 942.200 0 ±253.072 6 | 3 571.30 ±328.56 | 2 856.30 ±272.75 | 3 575.20 ±329.03 | **2 248.800 0 ±237.654 7** |

综合实验结果表明,本文提出的三支决策分类算法在不完备混合型数据下具有较好的分类效果。

## 5　结　语

决策粗糙集是目前粗糙集理论研究的重点模型。由于现实应用环境下数据往往都是不完备混合类型,本文将传统的决策粗糙集模型进行推广,提出不完备混合型信息系统下的决策粗糙集模型,构建该模型框架下的三支决策,并设计出该模型的一种最小化代价属性约简算法。基于所提出的三支决策,提出一种不完备混合型数据的三支决策分类算法。实验分析表明,所提出的三支决策分类算法比其他传统的分类算法具有更高的分类精度、较小的误分类代价和更高的优越性。动态性也是现实数据集的重要特征,因此接下来将进一步研究不完备混合型数据决策粗糙集的增量式学习问题。

## 参　考　文　献


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

[2] Yao Y. Probabilistic rough set approximations[J]. International Journal of Approximate Reasoning, 2008, 49: 255 – 271.

[3] Xu J, Miao D, Zhang Y, et al. A three-way decisions model with probabilistic rough sets for stream computing[J]. International Journal of Approximate Reasoning, 2017, 88: 1 –

22.
[4] 刘小伟, 王宁, 李天瑞, 等. 概率复合粗糙集模型的改进及其属性约简[J]. 南京大学学报(自然科学), 2018, 54(5):958 - 966.
[5] Yao Y, Zhao Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356 - 3373.
[6] Yao Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010, 180(3): 341 - 353.
[7] Lin G, Liang J, Qian Y, et al. A fuzzy multigranulation decision-theoretic approach to multi-source fuzzy information systems[J]. Knowledge-Based Systems, 2016, 91:102 - 113.
[8] Liu D, Liang D, Wang C. A novel three-way decision model based on incomplete information system [J]. Knowledge-Based Systems, 2016, 91:32 - 45.
[9] Zhao X, Hu B. Three-way decisions with decision-theoretic rough sets in multiset-valued information tables[J]. Information Sciences, 2020, 507:684 - 699.
[10] Feng T, Mi J S. Variable precision multigranulation decision-theoretic fuzzy rough sets[J]. Knowledge-Based Systems, 2016, 91:93 - 101.
[11] Sun B, Ma W, Zhao H. Decision-theoretic rough fuzzy set model and application[J]. Information Sciences, 2014, 283(1):180 - 196.
[12] Zhao X, Hu B. Fuzzy and interval-valued fuzzy decision-theoretic rough set approaches based on fuzzy probability measure[J]. Information Sciences, 2015, 298(20):534 - 554.
[13] 刘久兵, 张里博, 周献中, 等. 直觉模糊信息系统下的三支决策模型[J]. 小型微型计算机系统, 2018, 39(6): 1281 - 1285.
[14] Li W, Huang Z, Jia X, et al. Neighborhood based decision-theoretic rough set models[J]. International Journal of Approximate Reasoning, 2016, 69:1 - 17.
[15] Zhao H, Qin K. Mixed feature selection in incomplete decision table[J]. Knowledge-Based Systems, 2014, 57:181 - 190.
[16] 王映龙, 曾淇, 钱文彬, 等. 变精度下不完备混合数据的增量式属性约简方法[J]. 计算机应用, 2018, 38(10): 2764 - 2771.
[17] 史倩玉, 梁吉业, 赵兴旺. 一种不完备混合数据集成聚类算法[J]. 计算机研究与发展, 2016, 53(9):1979 - 1989.
[18] Hu Q, Yu D, Liu J, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18):3577 - 3594.
[19] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112(1 - 4): 39 - 49.
[20] Zhang Y, Jia X, Tang Z. Minimum cost attribute reduction in incomplete systems under decision-theoretic rough set model[C]//2016 12th International Conference on Natural Computation and 13th Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2016.
[21] Jia X, Liao W, Tang Z. Minimum cost attribute reduction in decision-theoretic rough set models [J]. Information Sciences, 2013, 219(10):151 - 167.
[22] Song J, Tsang E C C, Chen D, et al. Minimal decision cost reduct in fuzzy decision-theoretic rough set model [J]. Knowledge-Based Systems, 2017, 126:104 - 112.
[23] Hu Q, Yu D, Xie Z. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2):866 - 876.

(上接第 184 页)

[5] Abrishami S, Naghibzadeh M, Epema D H J. Deadline-constrained workflow scheduling algorithms for infrastructure as a service clouds [J]. Future Generation Computer Systems, 2013, 29(1):158 - 169.
[6] Chopra N, Singh S. HEFT based workflow scheduling algorithm for cost optimization within deadline in hybrid clouds [C]//4th International Conference on Computing, Communications and Network Technology. IEEE, 2013.
[7] Bossche R V D, Vanmechelen K, Broeckhove J. Online cost-efficient scheduling of deadline-constrained workloads on hybrid clouds[J]. Future Generation Computer Systems, 2013, 29(4):973 - 985.
[8] Verma A. Deadline and budget distribution based cost-time optimization workflow scheduling algorithm for cloud[C]// International Conference on Recent Advances and Future Trends in IT, 2012.
[9] 曹斌, 王小统, 熊丽荣, 等. 时间约束云工作流调度的粒子群搜索方法[J]. 计算机集成制造系统, 2016, 22(2):372 - 380.
[10] Verma A, Kaushal S. Deadline constraint heuristic-based genetic algorithm for workflow scheduling in cloud[J]. Journal of Grid & Utility Computing, 2014, 5(2):96 - 106.
[11] Verma A, Kaushal S. Budget constrained priority based genetic algorithm for workflow scheduling in cloud[C]//Fifth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2013), 2013.
[12] Chang Y, Fan C, Sheu R, et al. An agent-based workflow scheduling mechanism with deadline constraint on hybrid cloud environment[J]. International Journal of Communication Systems, 2018, 31(1):e3401.
[13] Chen W, Deelman E. WorkflowSim:a toolkit for simulating scientific workflows in distributed environments[C]//2012 IEEE 8th International Conference on E-Science. IEEE, 2012.