基于代价敏感的序贯三支决策最优粒度选择方法

张清华* 庞国弘 李新太 张雪秋

(重庆邮电大学计算智能重庆市重点实验室 重庆 400065)

摘 要：最优粒度选择是序贯三支决策领域研究的热点之一，旨在通过合理的粒度选择来对复杂问题进行求解。在现阶段最优粒度选择中，代价敏感是影响决策的重要因素之一。针对这个问题，该文首先基于信息增益和卡方检验提出一种新的属性重要度计算方法；其次，为了更好地符合实际应用场景，在构建多粒度空间时将代价参数与粒度大小相结合，设置了相应的惩罚规则，并分析了决策阈值的变化规律；最后，为了消除测试代价和决策代价量纲不一致所带来的影响，借助变异系数设计了一种客观的代价计算方法。实验结果表明，该模型适用于现有代价认知场景，能在给定代价情况下选出代价最小的最优粒层。

关键词：序贯三支决策；属性重要度；惩罚函数；变异系数；最优粒度选择

中图分类号：TP301.6 **文献标识码**：A **文章编号**：1009-5896(2021)10-3001-09

DOI: 10.11999/JEIT200821

Optimal Granularity Selection Method Based on Cost-sensitive Sequential Three-way Decisions

ZHANG Qinghua PANG Guohong LI Xintai ZHANG Xueqiu

(*Chongqing Key Laboratory of Computational Intelligence*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

Abstract: Optimal granularity selection is one of the hotspots in the research of sequential three-way decisions. It aims to solve complex problems through reasonable granularity selection. At present, in the field of optimal granularity selection, cost sensitivity is one of the important factors affecting decision making. To solve this problem, firstly, based on information gain and chi-squared test, a novel method to measure the attribute significance is proposed when constructing the multi-granularity space in this paper. Then, to better conform the practical application, the corresponding penalty rule is set by combining the cost parameters and the granularity, and the variation rule of the decision threshold is analyzed. Finally, to eliminate the influence of the dimensional difference between the test cost and the decision cost, an objective cost calculation method is given by the coefficient of variation. The experimental results show that the proposed algorithm can be used in existing cost cognition scene, and the optimal granular layer with the lowest cost can be obtained under the given cost scene.

Key words: Sequential three-way decisions; Attribute significance; Penalty function; Coefficient of variation; Optimal granularity selection

1 引言

在实际决策中，如何处理代价敏感问题一直是研究的热点之一。代价通常分为决策代价(误分类代价和延迟代价)和测试代价(测试成本)。一般地，决策代价随着信息逐步增加而降低，而测试代价随着信息增加而增加，两者呈负相关关系且量纲不同。例如，在医疗诊断中，若患者偏好高精度的诊断，会选择成本较高的检查项目；相反，若患者偏好普通的诊断，往往会选择成本低的检查项目。这两种情况都广泛发生在实际应用中，因此如何实现代价最小的决策是值得研究的。

现阶段许多专家学者将代价敏感研究运用于机器学习理论中，并取得了重要的研究成果[1,2]。目前，代价敏感方面的研究方法主要分为以下3个方面：从决策代价敏感的角度来看，Li等人[3]结合序

收稿日期：2020-09-21；改回日期：2021-07-19；网络出版：2021-08-18
*通信作者： 张清华 zhangqh@cqupt.edu.cn
基金项目：国家重点研发计划(2020YFC2003502)，国家自然科学基金(61876201)

Foundation Items: The National Key Research and Development Program of China (2020YFC2003502), The National Natural Science Foundation of China (61876201)

贯三支决策提出了一种最小化代价的决策模型；Zhang等人[4]基于邻域覆盖方法，根据损失函数改变覆盖半径，来减小分类损失；Jia等人[5]通过定义一种新的属性约简方法使模型的决策代价最小。同时，在降低测试代价方面，Yang等人[6]提出了一种测试代价最优的粒度结构选择回溯算法。Min等人[7]在测试代价中引入代价敏感决策系统的层次结构。另外，在同时考虑决策代价和测试代价的研究中，广大学者也进行了相应的工作[8,9]。

序贯三支决策[10]是近年发展起来的一种处理不确定性决策的方法。作为粒计算[11–13]概念下的具体模型，其目标是提供一种灵活的机制和方法，帮助用户在信息粒化过程中做出合适的决策。目前在图像分析、属性约简、语音识别等方面均已取得了较大的成果[14–17]。代价敏感的序贯三支决策从粒计算的角度提高了三支决策的有效性，实现了粗粒度到细粒度渐进式的决策过程。但在最优粒度选择方面，仍存在一些问题需要改进。首先，在构建多粒度空间过程中，从属性重要度选择方法上来看，存在没有充分考虑数据中有冗余属性或不相关属性的问题，这样可能会增加额外的测试代价或有损模型的性能。其次，随着获取信息的增多，针对两类错误分类和两类不确定性分类[18]的代价参数是保持不变的，使得代价参数在序贯三支决策渐进计算过程中缺乏一定的自适应性，导致在粗粒层产生较低的分类精度，从而影响模型的最优粒度选择。此外，在现有计算总代价的方法中，未能考虑测试代价与决策代价测量尺度或量纲不统一所带来的影响，从而丢失部分关键因素，导致直接进行计算得到的结果不准确。针对这些问题，本文首先利用卡方检验剔除高相关性的条件属性，再借助信息增益计算属性重要度并根据得到的属性重要度序列进行多粒度空间的构建。其次，针对两类错误分类和两类不确定性分类[18]的代价参数缺乏自适应性，结合渐进计算的思想，借助惩罚函数来对代价参数设置相应的惩罚规则，有效提升了模型的分类精度。最后，利用变异系数构建了一种合理的代价结构，实现了同量纲下的代价计算，从而可以有效利用测试代价和决策代价的信息。实验表明所提出的模型在不同的代价场景下能够产生合理的多粒度空间结构，同时所得到的代价最小的粒度空间也更符合实际应用场景代价最小的需求。

2 基础知识

定义1[19,20] 给定决策信息系统$S=(U,C\cup D,V,f)$，其中U表示非空有限论域；C和D分别表示条件属性集和决策属性集，且$C\cap D=\varnothing$；V表示属性值的集合；$f:U\times C\to V$表示一个信息函数，用于指定U中每一个对象x的属性值。

定义2[19,20] 给定决策信息系统$S=(U,C\cup D,V,f)$，对于任意属性子集$A\subseteq C$，等价关系E_A定义为

$$E_A=\{(x,y)\in U\times U|\forall a\in A,a(x)=a(y)\}。\tag{1}$$

等价关系可形成论域U上的一个划分，记为U/E_A，简记为U/A。给定对象$x\in U$，$[x]_{E_A}$表示在属性子集A所形成的等价关系下的等价类，简记为$[x]_A$或$[x]$。

相比于二支决策，三支决策理论的关键在于引入了延迟决策，即当决策对象的信息不足时采用延迟决策，等待收集更多有用信息后再重新进行决策。这种对决策对象的认识从粗粒度向细粒度转化，使边界域中的对象逐渐被正确决策，进而形成一种序贯决策方法。下面介绍序贯三支决策的一些基本概念。

定义3[10] 给定决策信息系统$S=(U,C\cup D,V,f)$，假定$A_1,A_2,\cdots,A_n$表示一组条件属性集，且满足$A_1\subset A_2\subset\cdots\subset A_n\subseteq C$。对于$\forall x\in U$，有

$$E_{A_n}\subset\cdots\subseteq E_{A_2}\subseteq E_{A_1}\tag{2}$$

$$[x]_{A_n}\subseteq\cdots\subseteq[x]_{A_2}\subseteq[x]_{A_1}\tag{3}$$

定义4[10] 给定决策信息系统$S=(U,C\cup D,V,f)$，设$A_1,A_2,\cdots,A_n$表示一组条件属性集，且满足$A_1\subset A_2\subset\cdots\subset A_n\subseteq C$。在这种条件属性集的序贯情形下多粒度空间记为GS，在第$i(i=1,2,\cdots,n)$层，GS的粒度结构记为GL_i，GL_i和GS定义为

$$\mathrm{GL}_i=(U_i,A_i\cup D,V_i,f_i)\tag{4}$$

$$\mathrm{GS}=(\mathrm{GL}_1,\mathrm{GL}_2,\cdots,\mathrm{GL}_n)\tag{5}$$

在多粒度空间中，给定第i层的阈值(α_i,β_i)，则第i层的接受域、延迟域和拒绝域可以表示为

$$\mathrm{POS}_{(\alpha_i,\beta_i)}(X_i)=\{x\in U_i|\Pr(X_i|[x]_{A_i})\geq\alpha_i\}\tag{6}$$

$$\mathrm{BND}_{(\alpha_i,\beta_i)}(X_i)=\{x\in U_i|\beta_i<\Pr(X_i|[x]_{A_i})<\alpha_i\}\tag{7}$$

$$\mathrm{NEG}_{(\alpha_i,\beta_i)}(X_i)=\{x\in U_i|\Pr(X_i|[x]_{A_i})\leq\beta_i\}\tag{8}$$

其中，U_i表示第i层的论域，$X_i(X_i\subseteq U_i)$表示第i层的目标概念。

经过GS的第i层决策后，得到边界域$\mathrm{BND}_{(\alpha_i,\beta_i)}(X_i)$，对于$\mathrm{BND}_{(\alpha_i,\beta_i)}(X_i)$中的对象，在第$i+1$层重新进行决策，因此$U_{i+1}=\mathrm{BND}_{(\alpha_i,\beta_i)}(X_i)$，满足$U_n\subset\cdots\subset U_2\subset U_1$且$U_1=U$。此外，第$i+1$层的目标概念$X_{i+1}=X_i\cap\mathrm{BND}_{(\alpha_i,\beta_i)}(X_i)$，满足$X_n\subset\cdots\subset X_2\subset X_1$且$X_1=X$。

粗糙集理论为序贯三支决策奠定了理论基础，从多粒度的角度来看，随着属性的增加，等价类会被进一步的细分。依据条件属性集构建的多粒度空间可以用树形结构来表示，最顶层表示论域的信息，即最粗粒层，随着属性的逐步加入，信息粒度逐步变细。因此，序贯三支决策的决策过程能够构成一个多粒度空间。图1简要介绍了多粒度的构造过程示意图。

3　代价敏感的序贯三支决策最优粒度选择模型

3.1　基于信息增益和卡方检验的属性重要度选择方法

多粒度空间的构建与属性重要度的选择是紧密相连的，如果充分考虑条件属性内在的关系和条件属性与决策属性之间的关系来进行属性重要度选择，所得到的多粒度空间往往会更优。因为数据集中有些条件属性是冗余甚至是不相关的。冗余属性的存在会增加额外的测试代价，而不相关的属性会有损模型的性能。因此，对条件属性进行相关性分析是有必要的，从而使模型泛化能力更强。

卡方检验是一种用途很广的计数资料的假设检验方法，属于非参数检验，主要是比较两个及两个以上样本率(构成比)以及两个分类变量的关联程度。其主要思想在于比较理论频数和实际频数的吻合程度或者拟合优度，用来描述两个事件的独立性。卡方值χ^2越大，说明两个事件的相互独立性越弱。

定义5(卡方分布[21])　设s个相互独立的随机变量$Y_1, Y_2, \cdots, Y_s$，且符合标准正态分布$N(0,1)$，则这s个随机变量的平方和$Q=\sum_{i=1}^{s} Y_i^2$为服从自由度为s的卡方分布，记为$Q \sim \chi^2(s)$。

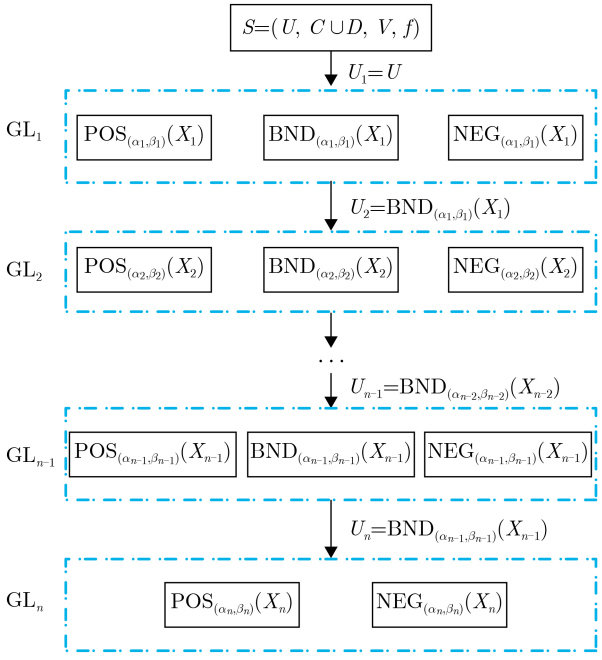

图 1 多粒度空间的构造过程

定义6(卡方检验[21])　给定数据的实际值A和理论值T，则卡方检验的公式为

$$\chi^2=\sum \frac{(A-T)^2}{T} \tag{9}$$

理论上，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值为0，表明理论值与数据的实际值完全符合。因此，通过卡方检验可以更好地剔除条件属性集中的冗余属性，减小测试代价。

同时，多粒度空间的构建与条件属性的划分能力是紧密相连的，如果充分考虑条件属性的划分能力来进行论域的划分，所得到的多粒度空间往往会更优。目前，属性重要度选择的方法大多基于熵。熵是用来描述论域中不确定性的一种度量方法。熵越大，论域的不确定性就越大。因此可以使用信息增益(论域集合划分前后熵的差值)来衡量使用当前属性对于论域划分效果的好坏。

定义7(信息增益[22,23])　给定决策信息系统$S=(U, C\cup D, V, f)$, $B\subseteq C$。假设论域U在等价关系E_B和E_D下的划分分别为$U/B=\{B_1, B_2, \cdots, B_m\}$和$U/D=\{D_1, D_2, \cdots, D_p\}$，信息增益$\mathrm{Gain}(D,B)$可定义为

$$\mathrm{Gain}(D,B)=H(D)-H(D|B) \tag{10}$$

其中，$H(D)=-\sum_{i=1}^{p}\frac{|D_i|}{|U|}\log_2\frac{|D_i|}{|U|}$，$H(D|B)=-\sum_{i=1}^{m}P(B_i)\sum_{j=1}^{p}P(D_j|B_i)\log_2 P(D_j|B_i)$。

基于信息增益的属性重要度做出选择的规则是：对于待划分的论域，在划分前的熵是一定的，而划分后的熵是不定的，且划分后的熵越小说明使用此属性划分所得到的子集的不确定性越小，即纯度越高，因此划分前后熵值差异越大，说明使用当前属性划分论域，其不确定性越小。以信息增益作为划分论域的属性选择的标准，在属性选择上更倾向于选择取值较多的属性，这样在多粒度空间构建的过程中粒度空间往往能够朝着最快到达最细粒度空间的方向发展，因此可以选择使得信息增益最大的属性来划分当前论域。

3.2　惩罚规则下代价参数和阈值的变化规律

因为基于决策粗糙集的三支决策存在一定的容错能力，所以3个域中都可能存在不确定性进而产生相应的代价。在序贯三支决策中，随着属性的增加，等价类被进一步细分，信息粒度逐步变细，对象之间的区分也越明显，边界域中的对象可能会被重新分类，分类精度会进一步的提升，所以针对错

误分类和不确定性分类应该给予更高的代价惩罚。本文借助文献[24]中的思想，考虑损失函数在随着粒度变化的情况下，利用惩罚函数对其进行相应的修改。因为在实际应用中，通常可以通过加大惩罚力度的方式来获取“优秀”的目标对象。同时，惩罚力度会随着惩罚次数的增加而增加，因此，惩罚函数必定是一个单调递增函数。进一步地，在序贯三支决策中，通过惩罚规则对代价参数进行修改，进而调整决策阈值(即α值的增大或β值的减小)，这样可以使等价类得到更准确的分类。同时，代价参数的值增大，即错分代价和延迟代价也会增高。所以，通过引入惩罚规则，利用代价参数值的增大进而提高决策精度。

考虑到采取不同行动会产生不同的损失，记λ_{BP}^k和λ_{NP}^k表示在第k层，x属于X时采取行动a_B和a_N下的损失；相似地，记λ_{PN}^k和λ_{BN}^k表示在第k层，x不属于X时采取行动a_P和a_B下的损失；另外，代价参数λ_{PP}和λ_{NN}表示正确划分下的代价，不产生代价损失。代价参数矩阵可以描述为表1。

表 1 代价参数矩阵

	X	$\neg X$
a_P	0	λ_{PN}^k
a_B	λ_{BP}^k	λ_{BN}^k
a_N	λ_{NP}^k	0

代价参数λ_σ^k, σ={NP, BP, PN, BN}可以表示为$\lambda_\sigma^k=\lambda_\sigma^{k-1}+\phi(\lambda_\sigma^{k-1})$，其中$\phi(x)$是单调递增的凹函数，粒度越细(即$k$越大)，$\phi(x)$的值越大。因此，根据上述规律，可以得到$\lambda_\sigma^k$与$\lambda_\sigma^1$的关系

$$\left.\begin{aligned}\lambda_\sigma^k&=\lambda_\sigma^{k-1}+\phi(\lambda_\sigma^{k-1})\\ \lambda_\sigma^{k-1}&=\lambda_\sigma^{k-2}+\phi(\lambda_\sigma^{k-2})\\ &\cdots\\ \lambda_\sigma^2&=\lambda_\sigma^1+\phi(\lambda_\sigma^1)\end{aligned}\right\} \tag{11}$$

将式(11)进行累加，可以得到：$\lambda_\sigma^k=\lambda_\sigma^1+\sum_{i=1}^{k-1}\phi(\lambda_\sigma^i)$。

因此，λ_σ^k可以表示为

$$\lambda_\sigma^k=\begin{cases}\lambda_\sigma^1, & k=1\\ \lambda_\sigma^1+\sum\limits_{i=1}^{k-1}\phi(\lambda_\sigma^i), & k>1\end{cases} \tag{12}$$

根据贝叶斯决策理论，将属于目标集合的对象分类到接受域的代价要小于等于将其分类到延迟域和拒绝域中的代价。相似地，将不属于目标集合的对象分类到拒绝域的代价要小于等于将其分类到延迟域和接受域中的代价。基于这两种规则，可以得到代价参数之间存在以下规律，$\lambda_{\mathrm{NP}}^k>\lambda_{\mathrm{BP}}^k\geq\lambda_{\mathrm{PP}}^k$, $\lambda_{\mathrm{PN}}^k>\lambda_{\mathrm{BN}}^k\geq\lambda_{\mathrm{NN}}^k$。因此决策阈值可以表示为

$$\alpha=\frac{(\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k)}{(\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k)+(\lambda_{\mathrm{BP}}^k-\lambda_{\mathrm{PP}}^k)} \tag{13}$$

$$\gamma=\frac{(\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{NN}}^k)}{(\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{NN}}^k)+(\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{PP}}^k)} \tag{14}$$

$$\beta=\frac{(\lambda_{\mathrm{BN}}^k-\lambda_{\mathrm{NN}}^k)}{(\lambda_{\mathrm{BN}}^k-\lambda_{\mathrm{NN}}^k)+(\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k)}。 \tag{15}$$

一般地，随着属性的增加，粒度变细，形成的等价类将发生变化，代价参数值增大，阈值也会相应地发生改变。

定理1 在多粒度空间中，任意相邻两个粒层GL_{k+1}和GL_k上的代价参数分别为λ_σ^{k+1}和λ_σ^k，且$\lambda_\sigma^{k+1}=\lambda_\sigma^k+\phi(\lambda_\sigma^k)$。相邻两个粒层$\mathrm{GL}_{k+1}$和$\mathrm{GL}_k$之间的阈值$(\alpha_{k+1},\beta_{k+1})$和$(\alpha_k,\beta_k)$存在以下4种关系：

(1) 如果满足$\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k>\lambda_{\mathrm{BP}}^k$和$\lambda_{\mathrm{BN}}^k>\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k$，则阈值$\alpha_{k+1}>\alpha_k$, $\beta_{k+1}>\beta_k$。

(2) 如果满足$\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k>\lambda_{\mathrm{BP}}^k$和$\lambda_{\mathrm{BN}}^k<\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k$，则阈值$\alpha_{k+1}>\alpha_k$, $\beta_{k+1}<\beta_k$。

(3) 如果满足$\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k<\lambda_{\mathrm{BP}}^k$和$\lambda_{\mathrm{BN}}^k>\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k$，则阈值$\alpha_{k+1}<\alpha_k$, $\beta_{k+1}>\beta_k$。

(4) 如果满足$\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k<\lambda_{\mathrm{BP}}^k$和$\lambda_{\mathrm{BN}}^k<\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k$，则阈值$\alpha_{k+1}<\alpha_k$, $\beta_{k+1}<\beta_k$。

因上述4种情形证明过程类似，故本文仅证明情形(1)。

证明 对于相邻两个粒层的阈值$\alpha_k=\dfrac{\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k}{\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k+\lambda_{\mathrm{BP}}^k}$和$\alpha_{k+1}=\dfrac{\lambda_{\mathrm{PN}}^{k+1}-\lambda_{\mathrm{BN}}^{k+1}}{\lambda_{\mathrm{PN}}^{k+1}-\lambda_{\mathrm{BN}}^{k+1}+\lambda_{\mathrm{BP}}^{k+1}}$，$\alpha_{k+1}-\alpha_k=\dfrac{\lambda_{\mathrm{BP}}^k(\lambda_{\mathrm{PN}}^{k+1}-\lambda_{\mathrm{BN}}^{k+1})-\lambda_{\mathrm{BP}}^{k+1}(\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k)}{(\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k+\lambda_{\mathrm{BP}}^k)(\lambda_{\mathrm{PN}}^{k+1}-\lambda_{\mathrm{BN}}^{k+1}+\lambda_{\mathrm{BP}}^{k+1})}$。因为$\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k>\lambda_{\mathrm{BP}}^k$，所以$\dfrac{\lambda_{\mathrm{PN}}^{k+1}-\lambda_{\mathrm{BN}}^{k+1}}{\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k}>\dfrac{\lambda_{\mathrm{BP}}^{k+1}}{\lambda_{\mathrm{BP}}^k}$，则$\alpha_{k+1}>\alpha_k$。

对于相邻两个粒层的阈值$\beta_k=\dfrac{\lambda_{\mathrm{BN}}^k}{\lambda_{\mathrm{BN}}^k+\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k}$和$\beta_{k+1}=\dfrac{\lambda_{\mathrm{BN}}^{k+1}}{\lambda_{\mathrm{BN}}^{k+1}+\lambda_{\mathrm{NP}}^{k+1}-\lambda_{\mathrm{BP}}^{k+1}}$，$\beta_{k+1}-\beta_k=\dfrac{\lambda_{\mathrm{BN}}^{k+1}(\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k)-\lambda_{\mathrm{BN}}^k(\lambda_{\mathrm{NP}}^{k+1}-\lambda_{\mathrm{BP}}^{k+1})}{(\lambda_{\mathrm{BN}}^k+\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k)(\lambda_{\mathrm{PN}}^{k+1}-\lambda_{\mathrm{BN}}^{k+1}+\lambda_{\mathrm{BP}}^{k+1})}$。因为$\lambda_{\mathrm{BN}}^k>\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k$，所以$\dfrac{\lambda_{\mathrm{NP}}^{k+1}-\lambda_{\mathrm{BP}}^{k+1}}{\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k}>\dfrac{\lambda_{\mathrm{BN}}^{k+1}}{\lambda_{\mathrm{BN}}^k}$，则$\beta_{k+1}>\beta_k$。 证毕

定理2 在多粒度空间中，任意相邻两个粒层GL_{k+1}和GL_k上的代价参数分别为λ_σ^{k+1}和λ_σ^k，且$\lambda_\sigma^{k+1}=\lambda_\sigma^k+\phi(\lambda_\sigma^k)$。如果满足$\lambda_{\mathrm{PN}}^k-\lambda_{\mathrm{BN}}^k=\lambda_{\mathrm{BP}}^k$和$\lambda_{\mathrm{BN}}^k=\lambda_{\mathrm{NP}}^k-\lambda_{\mathrm{BP}}^k$，则阈值$\alpha_{k+1}=\alpha_k$, $\beta_{k+1}=\beta_k$。

定理1与定理2同理可证。

因此，通过引入惩罚函数来处理实际决策过程中的代价参数变化，使得多粒度空间具有更好的适应性，能够动态地进行决策。

3.3 序贯三支决策模型的代价结构设计

在序贯三支决策中主要存在两种代价，第1种是因对象误分类或者需要延迟决策而产生的决策代价，第2种是因获得新的属性而产生的测试代价，即获取某些属性值的成本。在实际应用场景中，这两种代价都应该被考虑。因此，如何合理地结合决策代价和测试代价来解决问题具有重要意义。为了寻求决策代价和测试代价的最优平衡点，本文设计了一个启发式函数用来综合决策代价和测试代价。

因为产生测试代价的因素(时间、金钱、复杂度等)的维度不同，很难将各因素综合起来考虑。一般地，属性重要度越高的属性，它所拥有的分类能力越强，测试成本越高。

定义8　给定决策信息系统$S=(U,C\cup D,V,f)$，条件属性$c(c\in C)$对决策结果的影响度可以定义为

$$I(c)=H(D|C-\{c\})-H(D|C) \tag{16}$$

其中，$I(c)$的值越大，该决策属性对属性c的依赖程度越高，说明属性c的影响度越大。属性影响度作为启发式信息来度量某一属性的分类能力，区分能力越大，带来的测试代价越高。因此，测试代价与属性重要度呈现正相关关系，所以条件属性c的测试代价可以定义为

$$\mathrm{TC}_c=\eta\times I(c) \tag{17}$$

其中，η是一个常数。

一般地，若两个条件属性对决策属性的影响度一致(即划分能力一致)，那么这两个条件属性具有一样的测试代价。

定义9　在多粒度空间$\mathrm{GS}=(\mathrm{GL}_1,\mathrm{GL}_2,\cdots,\mathrm{GL}_n)$中，第$i$层的决策代价可以定义为

$$\begin{aligned}\mathrm{DC}_{\mathrm{GL}_i}=&\mathrm{COST}(\mathrm{POS}_{(\alpha_i,\beta_i)}(X_i))\\&+\mathrm{COST}(\mathrm{BND}_{(\alpha_i,\beta_i)}(X_i))\\&+\mathrm{COST}(\mathrm{NEG}_{(\alpha_i,\beta_i)}(X_i))\end{aligned} \tag{18}$$

其中，GL_i表示GS的第i粒层，$\mathrm{COST}(\mathrm{POS}_{(\alpha_i,\beta_i)}(X_i))$表示产生第1类分类错误带来的代价，$\mathrm{COST}(\mathrm{NEG}_{(\alpha_i,\beta_i)}(X_i))$表示产生第2类分类错误带来的代价，$\mathrm{COST}(\mathrm{BND}_{(\alpha_i,\beta_i)}(X_i))$表示产生不确定性分类带来的代价。

因为测试代价和决策代价呈现负相关关系且量纲不相同，所以不能将其直接进行计算。为了更好地计算总代价，本文引入变异系数的概念，并基于变异系数定义一种综合客观的评价函数进行总代价计算的方式

$$\mathrm{TotalCOST}_{\mathrm{GL}_i}=\theta_1\times\mathrm{TC}'_{\mathrm{GL}_i}+\theta_2\times\mathrm{DC}'_{\mathrm{GL}_i} \tag{19}$$

其中，$\mathrm{TotalCOST}_{\mathrm{GL}_i}$表示第$i$粒层上的总代价，$\mathrm{TC}'_{\mathrm{GL}_i}$和$\mathrm{DC}'_{\mathrm{GL}_i}$是标准化后的测试代价和决策代价，

$$\theta_1=\frac{\mathrm{C.V}_{\mathrm{TC}_{\mathrm{GL}_i}}}{\mathrm{C.V}_{\mathrm{TC}_{\mathrm{GL}_i}}+\mathrm{C.V}_{\mathrm{DC}_{\mathrm{GL}_i}}},\quad \theta_2=\frac{\mathrm{C.V}_{\mathrm{DC}_{\mathrm{GL}_i}}}{\mathrm{C.V}_{\mathrm{TC}_{\mathrm{GL}_i}}+\mathrm{C.V}_{\mathrm{DC}_{\mathrm{GL}_i}}} \tag{20}$$

C.V表示变异系数。

变异系数是衡量各组数据变异程度的一种统计量。在统计学中，如果两组数据的测量尺度相差太大，或者数据量纲不同，直接使用标准差来进行综合计算不合适，此时就应当消除测量尺度和量纲的影响，而变异系数可以做到这一点，它是原始数据标准差与原始数据平均数的比。因为变异系数没有量纲，因此得到结果是一个标量，可以客观地将决策代价与测试代价相结合。

4　实验对比及分析

4.1 实验设计

为了更好地说明所提模型的有效性和实用性，本文选取美国加州大学欧文分校(University of California Irvine, UCI)数据库的6个标准数据集进行了对比实验，并且每个数据集在两种不同的代价环境下进行实验。数据集的详细信息如表2所示。实验环境为8GB RAM, 3.2 GHz CPU, Windows 10 system，编程语言是Python。

本文算法的框架如图2所示，可以分为3个过程：属性重要度选择、多粒度空间构建和最优粒度

表 2　数据集的描述

序号	数据集	属性特征	数目	条件属性个数
1	Balance-scale	Categorical	625	4
2	Breast Cancer Wisconsin	Integer	699	9
3	Tic-Tac-Toe Endgame	Categorical	958	9
4	Car Evaluation	Categorical	1728	6
5	Nursery	Categorical	12960	8
6	Chess	Categorical, Integer	28056	6

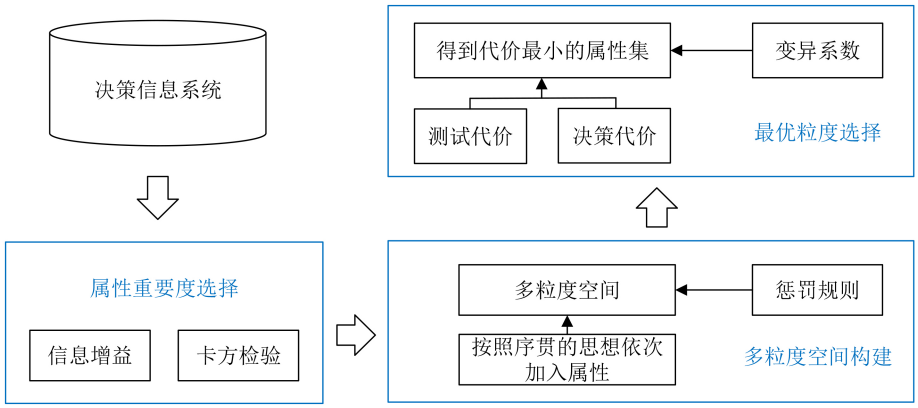

图 2 算法框架

选择。其中属性重要度选择部分由信息增益和卡方检验构成；在多粒度空间构建时，为代价参数设置了惩罚规则；最后利用变异系数消除测试代价与决策代价量纲的差异。

在计算算法的时间复杂度时，往往以最坏情况计算。根据上述实验步骤，算法的时间复杂度主要取决于多粒度空间构建，从图1中可知，多粒度空间是一个自顶向下且具有偏序关系的层级结构，层数是由条件属性集的基数(属性个数)所决定的。因属性重要度的选择方法是由卡方检验和信息增益所构成，因此需要对所有的属性进行计算：第1步属性重要度选择过程的时间复杂度为$O(n)$；多粒度空间的构建是基于经过属性重要度方法计算后条件属性集的属性个数的，所以构建多粒度空间的时间复杂度为$O(n)$，同时在每一粒层上借助惩罚规则对代价参数进行修改的时间复杂度为$O(1)$，因此第2步构建多粒度空间的时间复杂度为$O(n)$；第3步在最优粒度选择过程中，需要对全部粒层进行遍历计算，同样时间复杂度为$O(n)$。因为算法中3个步骤是递进关系，所以该算法整体的时间复杂度为$O(n)$，其中n表示序贯三支决策的条件属性集中属性的个数。

4.2 实验结果分析

本节对4.1节所选的UCI数据集进行了实验，为了方便研究，首先将数据集中的字符型数据转化为整数型数据；其次给出2组代价参数，其数值均满足第4节中定义并通过代价参数计算决策阈值对(α,β)，如表3所示；此外，为了体现最优化的思想，设计惩罚函数对代价参数进行惩罚。本文所选的惩罚函数是$\phi(x)=\log_2(1+0.1\times k)\times\lambda_\sigma$，其中$\sigma$={NP, BP, PN, BN}。

通过实验发现，运用上述的算法均可以得到不同数据集的代价最小的最优粒层，验证了算法的实用性。图3和图4给出了不同代价参数下的各数据集的代价变化以及最优粒层。另外，表4和表5分别列出了各数据集最优粒层的详细数据。从图3、图4和表4、表5中清楚地看出，所选的最优粒度较符合人类的认知。同时，所提出的代价结构利用标准化和变异系数进行处理能够消除因测试代价和决策代价尺度和量纲不同所带来的影响。

表 3 代价参数

	λ_{PP}	λ_{BP}	λ_{NP}	λ_{PN}	λ_{BN}	λ_{NN}
第1组	0	1	4	5	2	0
第2组	0	2	6	7	3	0

具体地，针对Breast Cancer Wisconsin数据集，通过使用最优粒度选择算法，将在不同代价参数环境下寻找一个总代价最小的粒度空间。从实验结果可以看出，在第1组代价参数下，代价最小的最优粒度空间由$\{c_2,c_3,c_6,c_7,c_5,c_8,c_4,c_9\}$诱导而得到并且构造多粒度空间的顺序是$c_2\to c_3\to c_6\to c_7\to c_5\to c_8\to c_4\to c_9$。此时构建的粒度空间总代价最小，为0.3684(标准化后)；在第2组代价参数下，代价最小的最优粒度空间$\{c_2,c_3,c_6,c_7,c_5\}$由诱导而得到，并且构造多粒度空间的顺序是$c_2\to c_3\to c_6\to c_7\to c_5$。此时构建的粒度空间总代价最小，为0.4459(标准化后)。

从以上6个数据集的实验结果可以看出，选取不同的代价参数时，所得到的最优粒层不一定是相同的，即便是改变一个代价参数也可能引起整个序贯三支决策粒层结构的改变，进而得到代价最小的最优粒层可能也是不一样的。相比于第1组代价参数，第2组代价参数值更大，所得到的最优属性子集中属性个数更少，这种所得到的代价最小的最优粒层是较为符合人类认知的。同时，两组代价参数通过定理1可以得到$\alpha_{k+1}>\alpha_k$, $\beta_{k+1}<\beta_k$，随着粒度空间的细化，每一粒层上的决策标准更为严格，分类到接受域(或延迟域)中对象的准确率更高，这与现实生产中的实际情况也是相吻合的。

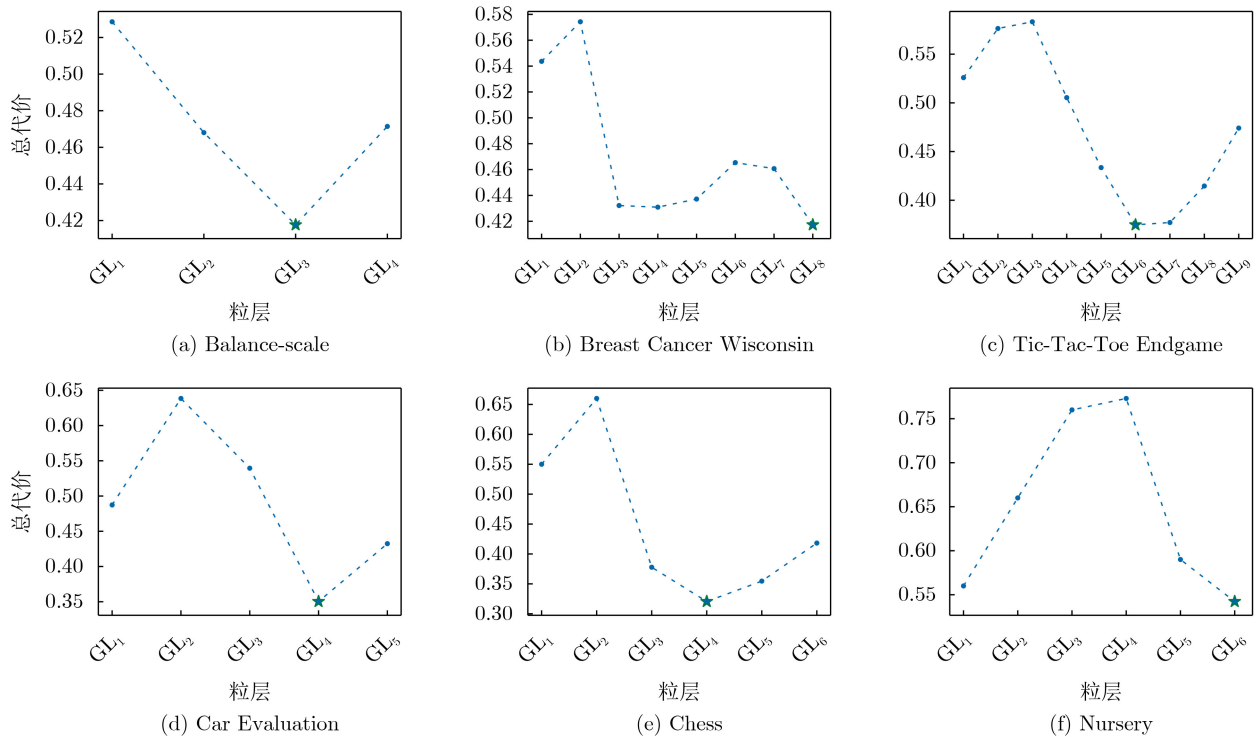

图 3 第1组代价参数下各数据集最优粒层的代价变化

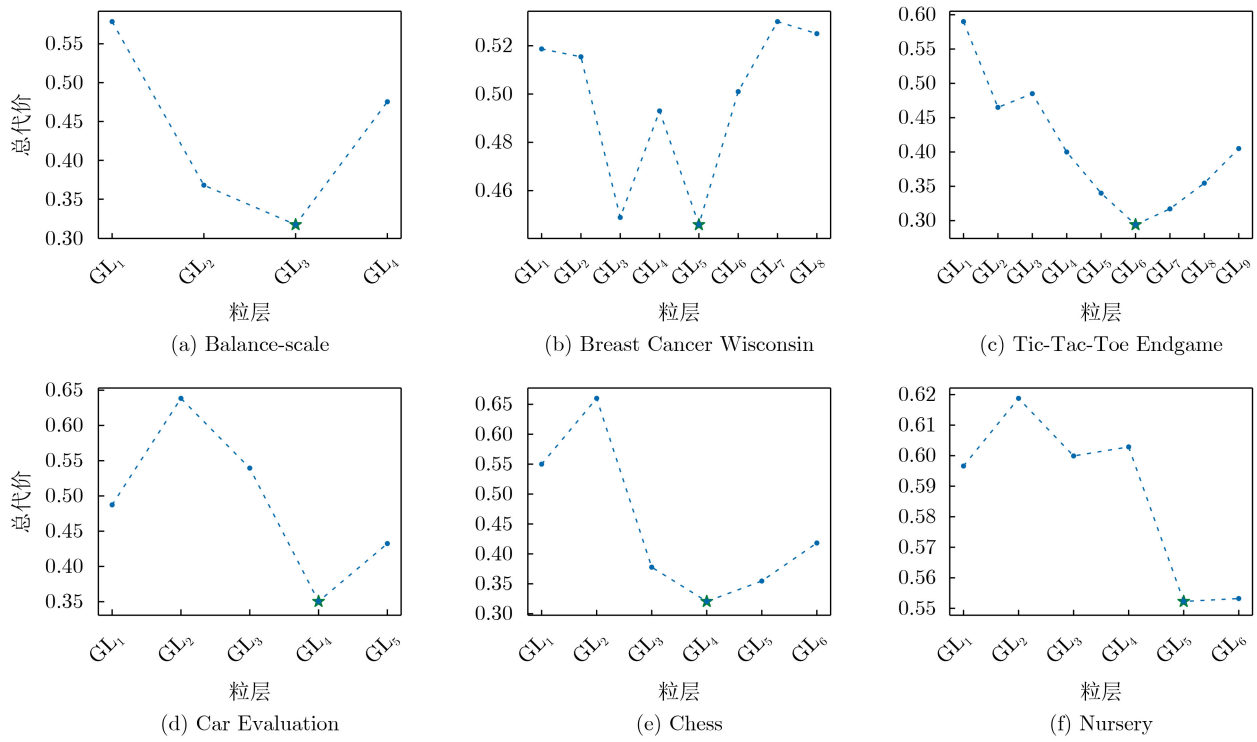

图 4 第2组代价参数下各数据集最优粒层的代价变化

此外，为了说明惩罚规则的有效性，将所提模型(模型1)与不加惩罚规则的最优粒层选择模型(模型2)在第1组代价参数下进行对比，实验结果如表6所示。从表中可以发现，模型1和模型2均可以得到代价最小的粒层。相比于模型2，模型1所得到的粒层比模型2所得到的最优属性子集中属性个数更多，即当前模型1所得的粒层能够获取的信息更多。通过实验说明，利用惩罚函数对代价参数进行合理的修改，在选取最优粒层的时候逐步提高了阈值要求，能够有效地防止选择测试代价较小同时精

表 4 第1组代价参数下各个数据集最优粒层信息

数据集	最优粒层	测试代价	决策代价	权重	总代价
Balance-scale	3	696.6	118.0	(0.46,0.54)	0.4172
Breast Cancer Wisconsin	8	323.7	522.6	(0.52,0.48)	0.3684
Tic-Tac-Toe Endgame	6	424.0	314.3	(0.49,0.51)	0.4453
Car Evaluation	4	636.0	35.2	(0.52,0.48)	0.4032
Chess	4	1997.3	0.0	(0.50,0.50)	0.4818
Nursery	6	998.9	8677.1	(0.54,0.46)	0.5423

表 5 第2组代价参数下每个数据集最优粒层信息

数据集	最优粒层	测试代价	决策代价	权重	总代价
Balance-scale	3	696.6	73.0	(0.47,0.53)	0.3172
Breast Cancer Wisconsin	5	227.9	652.1	(0.48,0.52)	0.4459
Tic-Tac-Toe Endgame	6	424.0	147.4	(0.40,0.60)	0.2941
Car Evaluation	4	636.0	372.1	(0.42,0.58)	0.3132
Chess	4	1997.3	14029.1	(0.41,0.59)	0.3705
Nursery	5	731.3	17162.085	(0.55,0.45)	0.5522

表 6 最优粒层比较

数据集	模型	最优粒层	冗余属性	最优属性子集
Balance-scale	模型1	3	ϕ	$\{c_4, c_3, c_2\}$
	模型2	3	ϕ	$\{c_4, c_3, c_2\}$
Breast Cancer Wisconsin	模型1	8	$\{c_1\}$	$\{c_2, c_3, c_6, c_7, c_5, c_8, c_4, c_9\}$
	模型2	5	$\{c_1\}$	$\{c_2, c_3, c_6, c_7, c_5\}$
Tic-Tac-Toe Endgame	模型1	6	ϕ	$\{c_8, c_6, c_4, c_2, c_9, c_7\}$
	模型2	6	ϕ	$\{c_8, c_6, c_4, c_2, c_9, c_7\}$
Car Evaluation	模型1	3	$\{c_2\}$	$\{c_4, c_1, c_6\}$
	模型2	3	$\{c_2\}$	$\{c_4, c_1, c_6\}$
Chess	模型1	6	ϕ	$\{c_3, c_5, c_2, c_1, c_4, c_6\}$
	模型2	5	ϕ	$\{c_3, c_5, c_2, c_1, c_4\}$
Nursery	模型1	6	$\{c_3, c_6\}$	$\{c_8, c_2, c_1, c_7, c_5, c_4\}$
	模型2	6	$\{c_3, c_6\}$	$\{c_8, c_2, c_1, c_7, c_5, c_4\}$

度较差的粒层。因此，所提出的模型具有更好的实用性。

在一定程度上，本文所提模型在实验过程中给定的代价参数需要在满足一定约束条件下进行随机选择，不同的代价参数组合得到的结果可能不一致。一般地，所给出的代价参数满足$\lambda_{\mathrm{PN}}-\lambda_{\mathrm{BN}}>\lambda_{\mathrm{BP}}$和$\lambda_{\mathrm{BN}}<\lambda_{\mathrm{NP}}-\lambda_{\mathrm{BP}}$等条件较为合理，在惩罚规则下，阈值$\alpha$会逐渐增大，阈值$\beta$会逐渐减小，每一粒层上分类时的标准更为严格，接受域或拒绝域中的对象精度越大。

5 结论

序贯三支决策作为粒计算概念下的产物，其目标是提供一个灵活的机制和方法，使得用户在信息粒化过程中做出合适的决策，因此如何通过合理的粒度选择，来对复杂问题进行求解是值得研究的。本文介绍了一种新的序贯三支决策中最优粒度选择的方法，其思想是首先通过信息增益对属性的分类能力进行排序，再利用卡方检验进行属性之间的相似度检验，去除冗余属性。其次，设计惩罚函数对代价参数进行处理，使其能够随着粒度自适应变化。进一步地，通过测试代价和决策代价的变异系数建立了一种客观的综合度量代价的方法，消除两种代价量纲不一致带来的影响，实现同量纲下的评价。最后，通过UCI上的标准数据集对本文所提方法进行了验证，实验结果表明了所提方法选取的最优粒度空间具有一定的实用性。

参 考 文 献

[1] LIAO Shujiao, ZHU Qingxin, QIAN Yuhua, *et al*. Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs[J]. *Knowledge-Based Systems*, 2018, 158: 25–42. doi: 10.1016/j.knosys.2018.05.020.

[2] YANG Jie, WANG Guoyin, 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. doi: 10.1016/j.knosys.2018.08.019.

[3] 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. doi: 10.1016/j.ijar.2017.03.008.

[4] ZHANG Yanping, ZOU Huijin, CHEN Xi, *et al*. Cost-sensitive three-way decisions model based on CCA[C]. The 9th International Conference on Rough Sets and Current Trends in Computing, Granada and Madrid, Spain, 2014: 172–180. doi: 10.1007/978-3-319-08644-6_18.

[5] JIA Xiuyi, LIAO Wenhe, TANG Zhenmin, *et al*. Minimum cost attribute reduction in decision-theoretic rough set models[J]. *Information Sciences*, 2013, 219: 151–167. doi: 10.1016/j.ins.2012.07.010.

[6] YANG Xibei, QI Yunsong, SONG Xiaoning, *et al*. Test cost sensitive multigranulation rough set: Model and minimal cost selection[J]. *Information Sciences*, 2013, 250: 184–199. doi: 10.1016/j.ins.2013.06.057.

[7] MIN Fan and LIU Qihe. A hierarchical model for test-cost-sensitive decision systems[J]. *Information Sciences*, 2009, 179(14): 2442–2452. doi: 10.1016/j.ins.2009.03.007.

[8] JU Hengrong, LI Huaxiong, YANG Xibei, *et al*. Cost-sensitive rough set: A multi-granulation approach[J]. *Knowledge-Based Systems*, 2017, 123: 137–153. doi: 10.1016/j.knosys.2017.02.019.

[9] JU Hengrong, YANG Xibei, YU Hualong, *et al*. Cost-sensitive rough set approach[J]. *Information Sciences*, 2016, 355/356: 282–298. doi: 10.1016/j.ins.2016.01.103.

[10] YAO Yiyu and DENG Xiaofei. Sequential three-way decisions with probabilistic rough sets[C]. IEEE 10th International Conference on Cognitive Informatics and Cognitive Computing (ICCI-CC'11), Banff, Canada, 2011: 120–125. doi: 10.1109/COGINF.2011.6016129.

[11] ZHANG Qinghua, CHEN Yuhong, YANG Jie, *et al*. Fuzzy entropy: A more comprehensible perspective for interval shadowed sets of fuzzy sets[J]. *IEEE Transactions on Fuzzy Systems*, 2020, 28(11): 3008–3022. doi: 10.1109/tfuzz.2019.2947224.

[12] ZHANG Qinghua, ZHAO Fan, YANG Jie, *et al*. Three-way decisions of rough vague sets from the perspective of fuzziness[J]. *Information Sciences*, 2020, 523: 111–132. doi: 10.1016/j.ins.2020.03.013.

[13] 张清华, 幸禹可, 周玉兰. 基于粒计算的增量式知识获取方法[J]. 电子与信息学报, 2011, 33(2): 435–441. doi: 10.3724/SP.J.1146.2010.00217.
ZHANG Qinghua, XING Yuke, and ZHOU Yulan. The incremental knowledge acquisition algorithm based on granular computing[J]. *Journal of Electronics & Information Technology*, 2011, 33(2): 435–441. doi: 10.3724/SP.J.1146.2010.00217.

[14] FANG Yu, GAO Cong, and YAO Yiyu. Granularity-driven sequential three-way decisions: A cost-sensitive approach to classification[J]. *Information Sciences*, 2020, 507: 644–664. doi: 10.1016/j.ins.2019.06.003.

[15] LI Huaxiong, ZHANG Libo, HUANG Bing, *et al*. Sequential three-way decision and granulation for cost-sensitive face recognition[J]. *Knowledge-Based Systems*, 2016, 91: 241–251. doi: 10.1016/j.knosys.2015.07.040.

[16] QIAN Jin, LIU Caihui, MIAO Duoqian, *et al*. Sequential three-way decisions via multi-granularity[J]. *Information Sciences*, 2020, 507: 606–629. doi: 10.1016/j.ins.2019.03.052.

[17] 陈泽华, 马贺. 基于粒矩阵的多输入多输出真值表快速并行约简算法[J]. 电子与信息学报, 2015, 37(5): 1260–1265. doi: 10.11999/JEIT141129.
CHEN Zehua and MA He. Granular matrix based rapid parallel reduction algorithm for MIMO truth table[J]. *Journal of Electronics & Information Technology*, 2015, 37(5): 1260–1265. doi: 10.11999/JEIT141129.

[18] ZHANG Qinghua, XIA Deyou, and WANG Guoyin. Three-way decision model with two types of classification errors[J]. *Information Sciences*, 2017, 420: 431–453. doi: 10.1016/j.ins.2017.08.066.

[19] 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: 17–18.

[20] PAWLAK Z. Rough sets[J]. *International Journal of Computer & Information Sciences*, 1982, 11(5): 341–356. doi: 10.1007/BF01001956.

[21] PLACKETT R L. Karl Pearson and the chi-squared test[J]. *International Statistical Review*, 1983, 51(1): 59–72. doi: 10.2307/1402731.

[22] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 75.

[23] QUINLAN J R. Induction of decision trees[J]. *Machine Learning*, 1986, 1(1): 81–106. doi: 10.1023/A:1022643204877.

[24] LIU Dun, LI Tianrui, and LIANG Decui. Three-way decisions in dynamic decision-theoretic rough sets[C]. The 8th International Conference on Rough Sets and Knowledge Technology, Halifax, Canada, 2013: 291–301. doi: 10.1007/978-3-642-41299-8_28.

张清华：男，1974年生，教授，博士，博士生导师，研究方向为不确定信息处理、粗糙集与粒计算等.

庞国弘：男，1994年生，硕士生，研究方向为不确定信息处理、粗糙集与三支决策.

李新太：男，1995年生，硕士生，研究方向为不确定信息处理、数据挖掘与机器学习.

张雪秋：女，1993年生，硕士生，研究方向为不确定信息处理、粗糙集与多尺度.

责任编辑：余　蓉