智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

特定类的代价敏感近似属性约简

胡军, 黄小涵

引用本文:
胡军, 黄小涵. 特定类的代价敏感近似属性约简[J]. 智能系统学报, 2024, 19(6): 1468-1478.
HU Jun, HUANG Xiaohan. Cost sensitive approximate attribute reduction for specific classes[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1468-1478.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202309032

DOI: 10.11992/tis.202309032
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240909.1354.014

特定类的代价敏感近似属性约简

胡军[1,2], 黄小涵[1,2]
（1. 重庆邮电大学 计算智能重庆市重点实验室, 重庆 400065; 2. 重庆邮电大学 计算机科学与技术学院, 重庆 400065）

摘　要: 特定类属性约简指针对特定决策类提供对应约简集的属性约简, 现有特定类属性约简方法过于严苛, 限制其在一些场景下的应用。针对存在噪声的数据, 提出一种特定类的代价敏感近似属性约简方法。该方法首先结合正域与边界域信息定义特定类的相对不确定度, 然后利用相对不确定度与测试代价计算属性重要度, 进而根据属性重要度选择属性, 并通过放松相对不确定度来避免冗余属性的加入, 最后给出了特定类的代价敏感近似启发式属性约简算法。实验结果表明, 所提方法与同类方法相比能够在保持甚至提升约简质量的同时获得更精简的约简集, 并且约简集的测试代价相对更小。

关键词: 粗糙集; 不确定信息; 特定类; 相对不确定度; 属性重要度; 测试代价敏感; 近似属性约简; 启发式算法
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)06−1468−11

中文引用格式: 胡军, 黄小涵. 特定类的代价敏感近似属性约简 [J]. 智能系统学报, 2024, 19(6): 1468–1478.
英文引用格式: HU Jun, HUANG Xiaohan. Cost sensitive approximate attribute reduction for specific classes[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1468–1478.

Cost sensitive approximate attribute reduction for specific classes

HU Jun[1,2], HUANG Xiaohan[1,2]
(1. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Class-specific attribute reduction refers to reducing attributes that are provided specifically for a given decision class. Existing class-specific attribute reduction methods are often too strict, which limits their applicability in certain scenarios. For noisy data, this paper proposes a cost-sensitive approximate attribute reduction method tailored for specific classes. First, the method combines information from the positive and boundary regions to define the relative uncertainty for a specific class. Then, attribute importance is calculated using relative uncertainty and test cost, allowing for attribute selection based on importance and avoiding the inclusion of redundant attributes by relaxing the relative uncertainty. Finally, the study introduces a cost-sensitive approximate heuristic attribute reduction for specific classes. Experimental results show that the proposed method can maintain or even improve the reduction quality while achieving a more streamlined reduction compared to other methods, with a relatively lower test cost for the reduction set.

Keywords: rough set; uncertain Information; specific class; relative uncertainty; attribute importance; test-cost-sensitive; approximate attribute reduction; heuristic algorithm

粗糙集是由 Pawlak[1] 提出的一种处理不精确、含糊或不一致数据的不确定性分析理论, 现已在数据挖掘、知识发现和决策分析等多个领域中[2-7] 取得了广泛的应用。属性约简是其中的一

收稿日期: 2023−09−18.　**网络出版日期:** 2024−09−09.
基金项目: 国家自然科学基金项目（62221005,62276038）; 重庆市自然科学基金项目（cstc2021ycjh-bgzxm0013）; 重庆市教委重点合作项目（HZ2021008）.
通信作者: 胡军. E-mail: hujun@cqupt.edu.cn.

个重要研究问题, 其在分析信息系统以挖掘其隐藏知识的方面有重要的作用。

属性约简是在保持信息系统分类能力不变的前提下, 去除其中的冗余和不相关属性, 实现对信息系统的优化处理。属性约简根据相对目标有 2 种类型, 即面向决策的属性约简[8] 和面向特定类的属性约简[9-10]。其中, 面向决策的属性约简通常为所有决策类选择相同的属性约简。Pawlak[8] 将面向决策的属性约简定义为保持系统正域不变的最小属性子集。在一些实际应用场景下, 有时只关注某一个决策类, 而不关注其他决策类。比如, 在医疗诊断中, 判断一个病人是否患有某种疾病并不需要做完所有检查, 只需检查部分项目即可, 且不同的疾病所需的检查项目也各不相同。因此, 有必要研究面向特定类的属性约简, 即允许为不同的决策类提供不同的属性约简。Stepaniuk[9] 引入了基于决策类正域的属性约简概念。Yao 等[10] 提出了特定类属性约简的名称来重新考虑决策类正域保留的属性约简, 并探讨了面向决策的属性约简与面向特定类的属性约简之间的关系。

目前面向特定类的属性约简主要从三支决策和信息度量 2 个角度进行研究。从三支决策的角度来看, Ma 等[11] 提出了基于三支决策的特定类的正域和负域属性约简; Zhang 等[12] 提出了基于概率粗糙集模型正域保留、负域保留以及正域和负域保留的定量三支特定类的属性约简。从信息度量的角度来看, Zhang 等[13] 提出了信息角度下的特定类属性约简; 吴婉琳等[14] 利用粗糙集的不确定度来进行特定类的属性约简, 探讨了 2 种属性约简之间的关系; 陈阳等[15] 探讨了在不完备决策系统中的特定类分布约简; Zhang 等[16] 提出了基于邻域粗糙集的特定类属性约简; Ma[17] 构造了基于模糊熵的单调测度进行特定类的属性约简; Zhang 等[18] 利用代数和信息理论这 2 个视角以及 2 种约简类型进行了 4 种组合的讨论, 并探讨了 4 种组合之间的横向联系与层次联系。总之, 这些研究为针对特定决策类的属性约简提供了方法, 但基本没有考虑实际情况下的代价。

代价敏感属性约简是传统属性约简的自然推广, 其代价一般有误分类代价和测试代价, 其中误分类代价指错误决策造成的代价, 测试代价则是指获取新信息所需要的代价。许多学者探讨了如何利用误分类代价和测试代价来评估属性, 进而通过代价最小化准则得到约简属性子集。Li 等[19-20] 将决策粗糙集模型扩展到代价敏感的三支决策模型; Yao 等[21] 讨论了决策粗糙集模型中属性约简的代价准则; Min 等[22] 提出了一种最小化测试代价的约简算法; Yang 等[23] 建立了测试代价敏感的多粒度粗糙集模型, 并提出了一种测试代价最小的粒度结构选择回溯算法; Ju 等[24] 通过将测试代价引入到不可分辨关系中, 建立了代价敏感粗糙集模型; Ju 等[25] 同时考虑误分类代价和测试代价, 构建了一种新的多粒度代价敏感粗糙集模型。Ma 等[26] 在三支特定类属性约简中同时考虑了误分类代价和测试代价, 并通过平衡 2 种代价得到特定类的最小代价约简。可见, 在结合特定类属性约简与代价敏感问题的研究中, 相关的研究工作还相对较少。

目前的许多针对特定类的属性约简方法一般都存在对约简子集的生成过于严苛的问题, 这可能会限制其在一些特定场景下的应用, 比如数据中存在噪声, 这些方法就不能适用。并且, 部分冗余属性会因为过于严苛的约简条件进入到约简子集中, 从而影响到约简属性子集的分类精度。考虑到实际问题中属性的测试代价, 本研究提出了一个关于特定类的代价敏感近似属性约简算法。主要贡献有以下 3 点:

1) 结合依赖度和不确定度定义了特定类下的相对不确定度, 利用相对不确定度和测试代价评价属性的重要性, 为属性约简时选择分辨能力强且测试代价低的属性提供了依据。

2) 利用近似属性约简的思想给出了特定类近似属性约简的定义, 通过放松相对不确定度来避免冗余属性的加入, 提高了特定类属性约简对噪声数据的适应能力。

3) 提出了特定类的代价敏感近似属性约简算法, 实验结果表明, 所提算法与同类算法相比能够在保持甚至提升约简质量的同时获得更精简的约简集, 并且约简集的测试代价相对更小。

1　预备知识

粗糙集属性约简的几个基本概念。

定义 1[1,10]　决策表可表示为 $T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\},I=\{I_a|a\in A_{\mathrm{T}}\})$, U 是一个非空的有限论域, $A_{\mathrm{T}}=C\cup D$ (C 是一组条件属性, D 是决策属性, 且 $C\cup D=\varnothing$) 是一个非空的有限属性集, V_a 是一个属性 $a\in A_{\mathrm{T}}$ 的值域, $I_a:U\to V_a$ 是一个相应的信息函数, 将 U 中的一个对象映射到 V_a 中的一个值。

给定一个条件属性子集 $B\subseteq A_{\mathrm{T}}$, 等价关系的定义如下:

$$E_B=\{(x,y)\in U\times U|\forall a\in B, I_a(x)=I_a(y)\}$$

$[x]_B=\{y\in U|(x,y)\in E_B\}$表示等价类, $\pi_B=U/E_B=\{[x]_B|x\in U\}$是$U$的一个划分。对于决策属性$D$, 可以得到划分$\pi_D=\{D_1,D_2,\cdots,D_j,\cdots,D_m\}$, 称为$U$的一个决策分类, 任意的$D_j\in\pi_D$称为一个决策类。

定义 2[10,14] 给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\}, I=\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$, 关于条件属性子集$B\subseteq C$, 决策类D_j的上近似$\overline{B}(\ )$、下近似$\underline{B}(\ )$分别定义为

$$\overline{B}(D_j)=\{x|[x]_B\cap D_j\neq\varnothing\}$$
$$\underline{B}(D_j)=\{x|[x]_B\subseteq D_j\}$$

于是, 决策类D_j的正域 (POS, 记为P_{OS})、负域 (NEG, 记为N_{EG})、边界域 (BND, 记为B_{ND}) 如下

$$P_{\mathrm{OS}}(D_j|\pi_B)=\underline{B}(D_j)$$
$$N_{\mathrm{EG}}(D_j|\pi_B)=U-\overline{B}(D_j)$$
$$B_{\mathrm{ND}}(D_j|\pi_B)=\overline{B}(D_j)-\underline{B}(D_j)$$

进而, 决策类D_j的近似精度(α)、粗糙度 (ρ) 和不确定度 (UNC, 式中记为U_{NC}) 定义为

$$\alpha(D_j|\pi_B)=\frac{|\underline{B}(D_j)|}{|\overline{B}(D_j)|}$$
$$\rho(D_j|\pi_B)=1-\alpha(D_j|\pi_B)$$
$$U_{\mathrm{NC}}(D_j|\pi_B)=\frac{|D_j|}{|U|}\rho(D_j|\pi_B)$$

定义 3[12] 给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\}, I=\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$, 决策类D_j关于条件属性子集$B\subseteq C$的依赖度γ_B定义为

$$\gamma_B(D_j|\pi_B)=\frac{|P_{\mathrm{OS}}(D_j|\pi_B)|}{|U|}$$

可见, 决策类D_j的依赖度是其关于条件属性子集所确定的正域集合在论域中所占的比例。

2 特定类的代价敏感近似属性约简

首先定义了相对不确定度, 随后引入测试代价构造了属性重要度, 最后提出了一种基于相对不确定度的近似属性约简以及相应的启发式属性约简算法。

现有的使用信息度量的特定类属性约简方法一般只单独考虑来自正域的信息或来自边界域的不确定信息, 有可能会缺失一部分有用信息。为了同时有效利用来自正域和边界域的有用信息, 结合依赖度和不确定度, 构造了一个新的度量, 称为相对不确定度。

定义 4 给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\}, I=\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$, 决策类D_j关于条件属性子集$B\subseteq C$的相对不确定度 (RUNC, 记为R_{UNC}) 定义如下

$$R_{\mathrm{UNC}}(D_j|\pi_B)=(1-\gamma_{\mathrm{B}}(D_j|\pi_B))\times U_{\mathrm{NC}}(D_j|\pi_B)$$

性质 1 给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\}, I=\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$, 当决策类D_j一致, 即对于任意条件属性子集$B\subseteq C$, 都有$P_{\mathrm{OS}}(D_j|\pi_B)=D_j$, $U_{\mathrm{NC}}(D_j|\pi_B)=0$时, 取最小值 0; 当决策类$D_j$完全不一致, 即对于任意条件属性子集$B\subseteq C$, 都有$P_{\mathrm{OS}}(D_j|\pi_B)=\varnothing$, $U_{\mathrm{NC}}(D_j|\pi_B)=\dfrac{|D_j|}{|U|}$时, 取最大值$\dfrac{|D_j|}{|U|}$。

性质 2 给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\}, I=\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$, $\forall P,Q\subseteq C$若$P\supseteq Q$, 则$R_{\mathrm{UNC}}(D_j|\pi_P)\leqslant R_{\mathrm{UNC}}(D_j|\pi_Q)$。

证明 因为$\forall P,Q\subseteq C$, 且$P\supseteq Q$, 则$\pi_P\preceq\pi_Q$,

所以$\forall x\in D_j,[x]_P\subseteq[x]_Q([x]_P\in\pi_P,[x]_Q\in\pi_Q)$。

1) $\forall X\subseteq D_j,\underline{P}(X)=\{x|[x]_P\subseteq X\},\underline{Q}(X)=\{x|[x]_Q\subseteq X\}$

因此对$\forall x\in D_j,\forall X\subseteq D_j$, 若$x\in\underline{Q}(X),[x]_Q\subseteq X$, 则$[x]_P\subseteq[x]_Q\subseteq X$。

所以$x\in\underline{P}(X)$,

所以$\forall X\subseteq D_j$, 有$\underline{Q}(X)\subseteq\underline{P}(X)$。

2) $\forall X\subseteq D_j,\overline{P}(X)=\{x|[x]_P\cap X\neq\varnothing\}$

$\overline{Q}(X)=\{x|[x]_Q\cap X\neq\varnothing\}$

因此对$\forall x\in D_j,\forall X\subseteq D_j$, 若$x\in\overline{P}(X),[x]_P\cap X\neq\varnothing$, 则因为$[x]_P\subseteq[x]_Q$, 所以$[x]_Q\cap X\neq\varnothing$

所以$x\in\overline{Q}(X)$,

所以$\forall X\subseteq D_j$, 有$\overline{P}(X)\subseteq\overline{Q}(X)$。

因此, $\forall x\in D_j,\rho(D_j|\pi_P)=1-\dfrac{|\underline{P}(D_j)|}{|\overline{P}(D_j)|}\leqslant\rho(D_j|\pi_Q)=1-\dfrac{|\underline{Q}(D_j)|}{|\overline{Q}(D_j)|}$, $|P_{\mathrm{OS}}(D_j|\pi_P)|=\underline{P}(D_j)\geqslant|P_{\mathrm{OS}}(D_j|\pi_P)|=\underline{Q}(D_j)$

所以, $R_{\mathrm{UNC}}(D_j|\pi_P)=\left(1-\dfrac{|P_{\mathrm{OS}}(D_j|\pi_P)|}{|U|}\right)\times U_{\mathrm{NC}}(D_j|\pi_P)\leqslant R_{\mathrm{UNC}}(D_j|\pi_Q)=\left(1-\dfrac{|P_{\mathrm{OS}}(D_j|\pi_Q)|}{|U|}\right)\times U_{\mathrm{NC}}(D_j|\pi_Q)$.

证毕。

性质 1 给出了相对不确定度的值域及最值, 性质 2 说明相对不确定度具有单调性, 即属性越多, 其相对不确定度越小。

在基于特定类的属性约简算法中, 通常要求经过约简得到的属性子集能够保持甚至提升属性全集下的度量。然而, 这样的做法忽视了样本中存在噪声数据的可能性。因此, 在属性约简后期, 少量的属性对度量贡献值很小, 为了保持甚至提升度量值的要求会拟合可能存在的一些噪声数据, 添加一些冗余属性到约简子集中, 从而引发约简质量不高的问题。为此, 本研究引入了近似属性约简的概念。

定义 5 给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\}, I=$

$\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$及参数$\varepsilon(\varepsilon\in(0,1])$, 属性子集$B\subseteq C$是决策类$D_j\in D$的一个$\varepsilon$–近似属性约简, 当且仅当以下条件成立

$$(S)\varepsilon\leqslant\frac{R_{\mathrm{UNC}}(D_j|\pi_C)}{R_{\mathrm{UNC}}(D_j|\pi_B)}$$

$$(N)\forall a\in B,\varepsilon>\frac{R_{\mathrm{UNC}}(D_j|\pi_C)}{R_{\mathrm{UNC}}(D_j|\pi_{B-\{a\}})}$$

式中: ε为一个参数, 其取值范围为$(0,1]$。当ε等于 1 时, 近似属性约简子集即为度量保持约简, 所以现有的度量保持约简是ε–近似属性约简在ε 等于 1 时的特例。

定义 6　给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\},I=\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$, 决策类D_j下属性$a\in C-B$, $B\subseteq C$关于条件属性子集B的属性影响度定义为

$$I(a,B,D_j)=R_{\mathrm{UNC}}(D_j|\pi_B)-R_{\mathrm{UNC}}(D_j|\pi_{B\cup\{a\}})$$

$I(a,B,D_j)$值越大, 表明属性a相对属性子集B的重要度越大。

考虑到各个属性的测试代价会有所不同, 为了使最后得到的特定类的近似属性约简子集在保持约简精度的同时能够具有较低的测试代价, 结合测试代价和属性影响度来综合评价属性的重要度。对于$\forall a\in C$, 将其测试代价定义为$t_c(a)$, 目标是在每一轮属性选择中选取属性影响度较高并且测试代价较小的属性。

定义 7　给定决策表$T=(U,A_{\mathrm{T}},\{V_a|a\in A_{\mathrm{T}}\},I=\{I_a|a\in A_{\mathrm{T}}\})$, $A_{\mathrm{T}}=C\cup D$, 决策类D_j下属性$a\in C-B$, $B\subseteq C$关于条件属性子集B的属性重要度 (SIG, 式中记为S_{IG}) 定义如下

$$S_{\mathrm{IG}}(a,B,D_j)=I(a,B,D_j)t_c(a)^{\lambda}$$

式中λ是惩罚指数。通过惩罚指数λ, 可以根据需要平衡属性影响度以及测试代价, 在属性选择时能够选出属性影响度较大并且测试代价较小的属性。

基于相对不确定度的特定类的代价敏感近似属性约简算法 (class-specific cost-sensitive approximate attribute reduction based on relative uncertainty, RUNC-CSCAR) 过程如算法 1 所示。

算法 1　基于相对不确定度的特定类的代价敏感近似属性约简算法 (RUNC-CSCAR)

输入　决策表T

输出　特定类D_j的近似属性约简B_j

1) 遍历每一个特定决策类D_j, 计算 $R_{\mathrm{UNC}}(D_j|\pi_C)$, 做步骤 2)—步骤 5)的操作;

2) 设置$B_j=\varnothing$;

3) $\forall a\in(C-B_j)$, 计算属性重要度 $S_{\mathrm{IG}}(a,B_j,D_j)$, 选择最大的条件属性加入约简集$B_j$中, $B_j\leftarrow B_j\cup\{a\}$;

4) 若$\varepsilon R_{\mathrm{UNC}}(D_j|\pi_{B_j})>R_{\mathrm{UNC}}(D_j|\pi_C)$, 转 3);
　若$\varepsilon R_{\mathrm{UNC}}(D_j|\pi_{B_j})\leqslant R_{\mathrm{UNC}}(D_j|\pi_C)$, 转 5);

5) 输出决策类D_j的近似属性约简B_j,

6) 算法结束。

算法 1 使用了添加的策略进行属性约简, 针对每一个特定的决策类, 先向空集中不停地添加属性重要度高的属性, 直至约简子集的相对不确定度达到近似参数ε约束下的属性全集度量标准, 此时便得到了一个针对特定类的近似属性约简。

假设有$|U|$个样本, $|C|$个条件属性, m个决策类, 算法 1 遍历了每一个决策类, 同时每一轮选取最优属性时, 计算最优属性所需的最多时间复杂度为$O(|U||C|^2)$, 所以算法 1 最终的时间复杂度就是$O(m|U||C|^2)$。

3　实验结果与分析

对所提方法的有效性进行了实验分析。实验在 Windows 10 操作系统, Intel(R) Core(TM) i7-9700 CPU、3.00 GHz、16 GB RAM 处理器上完成, 数据集使用 WEKA 软件进行数据预处理。

3.1　实验数据集

从 UCI 数据集中选取了 7 个数据集, 并使用 WEKA 软件对数据集进行了预处理, 包括缺失值填充和连续数据离散化处理。经过预处理的数据集见表 1。

表 1　实验数据集

Table 1　Datasets used in experiments

数据集	实例数	条件属性	决策类
Cardio	2 126	21	3
Credit	30 000	23	2
Turkiye	5 820	31	3
CMC	1 473	9	3
Biodeg	1 055	41	2
Flare	610	12	6
Optical	5 620	64	10

3.2　结果分析与比较

实验对比了 3 种特定类属性约简算法: 特定类正域约简 (CSB2CSR)[10]、特定类不确定度约简 (UNC-CSR)[14] 和特定类互信息约简 (MI-CSR)[18]。具体比较了各个算法在分类精度、测试代价以及约简子集规模 3 个方面的表现。实验使用 2 种不同的分布函数来生成测试代价: 均匀分布和正态分布。其中, 均匀分布的测试代价取值服从$X\sim(1,50)$, 正态分布的测试代价取值服从$N\sim(25,5)$。

首先对测试代价惩罚指数λ进行了实验。实验中，λ分别设定为{−1, −1/2, −1/4, −1/6, −1/8}5 种取值，近似参数ε设定为 1，在 2 种测试代价分布情况下针对每个决策类得到约简子集，计算平均每个约简子集的测试代价以及分类精度，其中分类精度是由约简子集经过 KNN 分类器 (K=5)10 次十折交叉验证所得到的。图 1 为平均测试代价以及平均分类精度随惩罚指数λ的变化曲线。研究发现，随着惩罚指数λ的变大，代价有一定的提升，而精度有一定的下降。结合定义 7 可知，λ的取值越大，对测试代价的惩罚力度也就越大，属性重要度中测试代价的权重也随之变大，因此每一轮属性选择更倾向于选择测试代价小的属性，而该属性包含的信息却可能不是很多。根据这样的实验结果，在接下来的实验中，将λ取值设定为−1/2。

当近似属性约简设定的近似参数ε不同时，约简的约束条件将会相应地放宽或加强，由于实验中低于 0.7 的取值得出的约简子集精度普遍偏低，故而将取值设定在 [0.7,1] 以 0.05的步长进行递增，并记录了在不同ε设置下，针对各个决策类所得到的约简子集的长度。如图 2 和图 3 所示分别是在均匀分布和正态分布下各数据集中针对每个决策类所得到的约简子集的属性数随ε的变化曲线。其中，每个子图表示一个数据集，图例中D_j表示该数据集的各决策类。

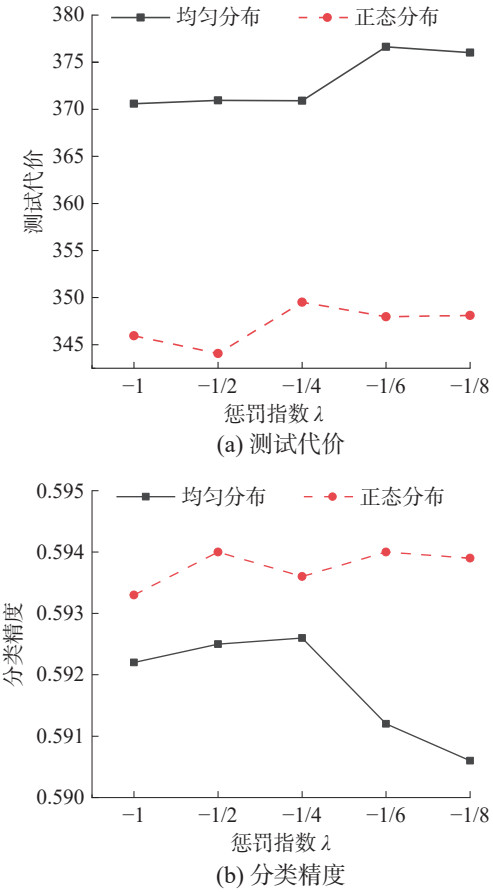
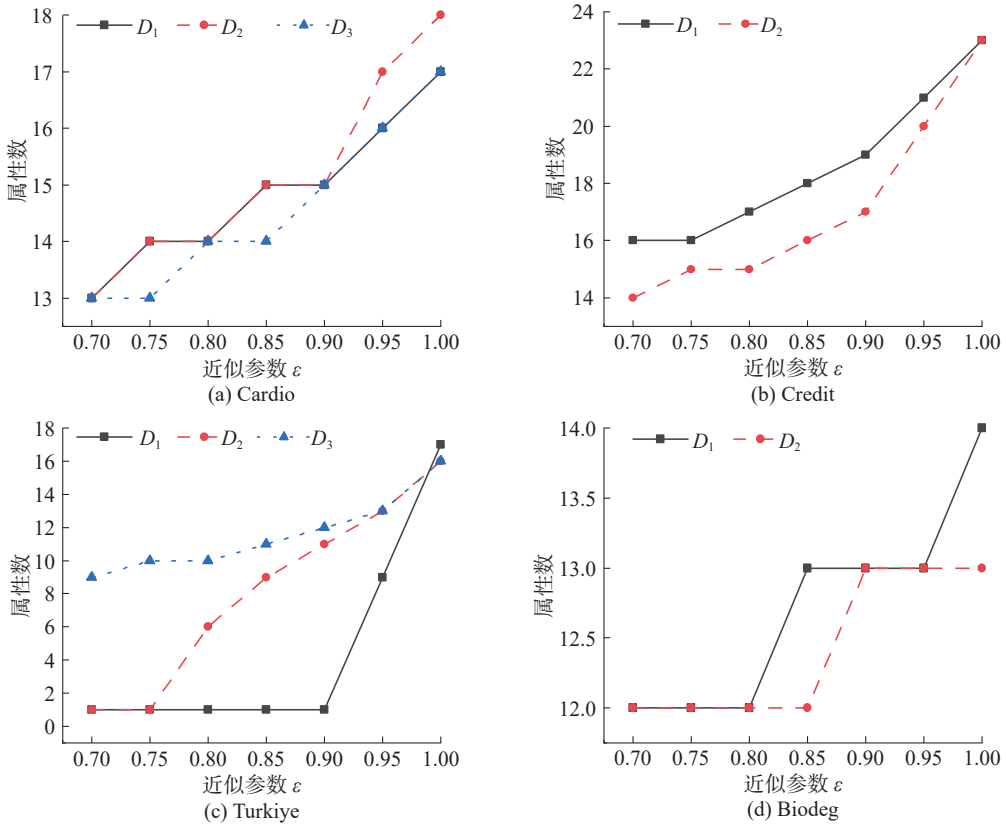


图 1　2 种分布下特定类约简子集的平均代价和平均精度随着惩罚指数λ的变化曲线

Fig. 1　Average cost and average accuracy of class-specific reduct varying with the punishment index λ under two kinds of distribution



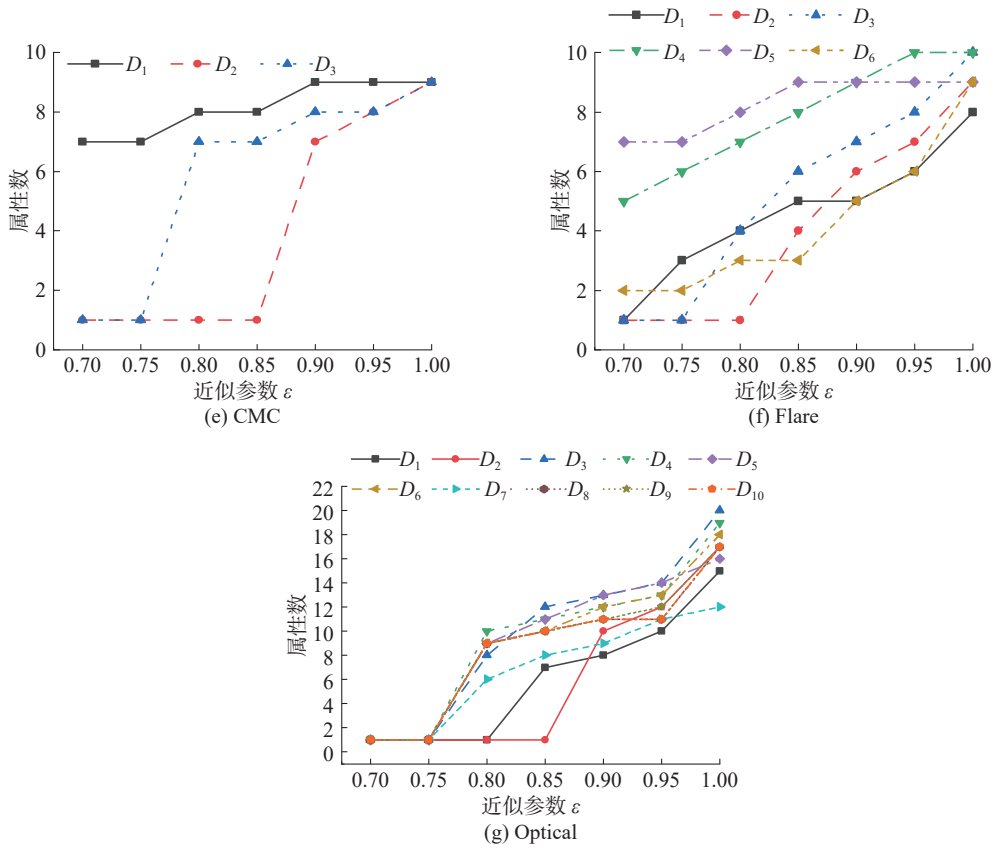
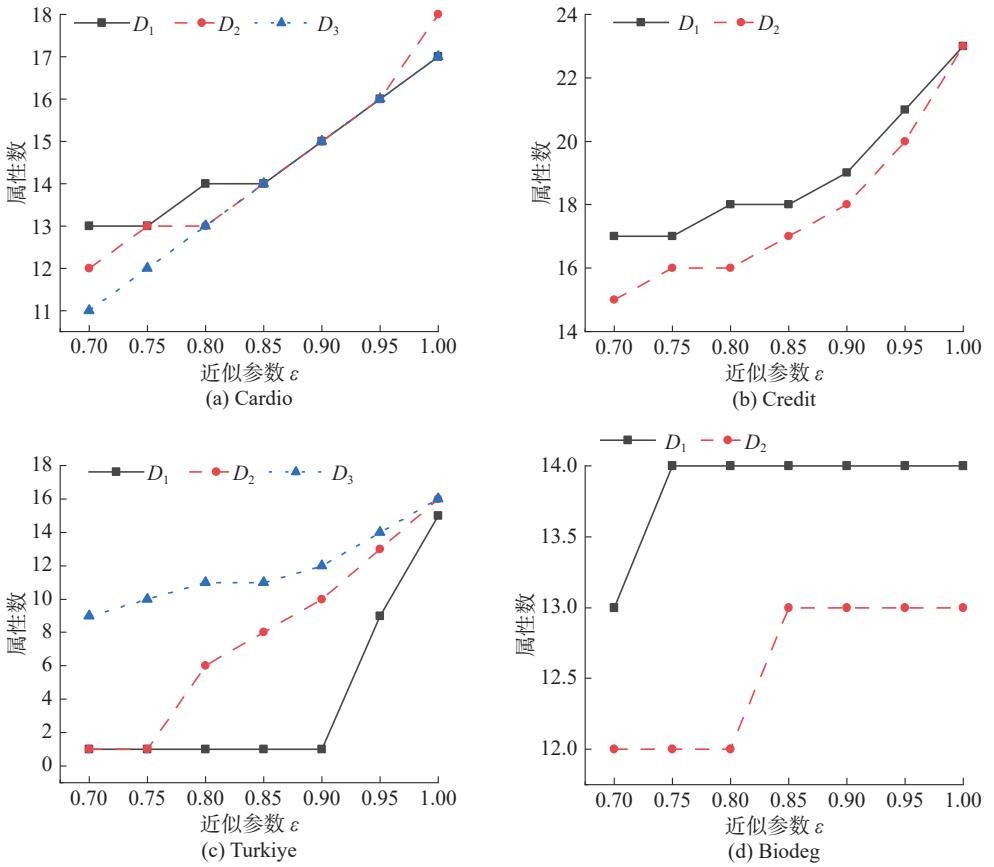


图 2　均匀分布下特定类约简子集属性数随着近似参数 (ε) 的变化曲线

Fig. 2　Number of attributes of class-specific reduct varying with the approximate parameter ε under uniform distribution



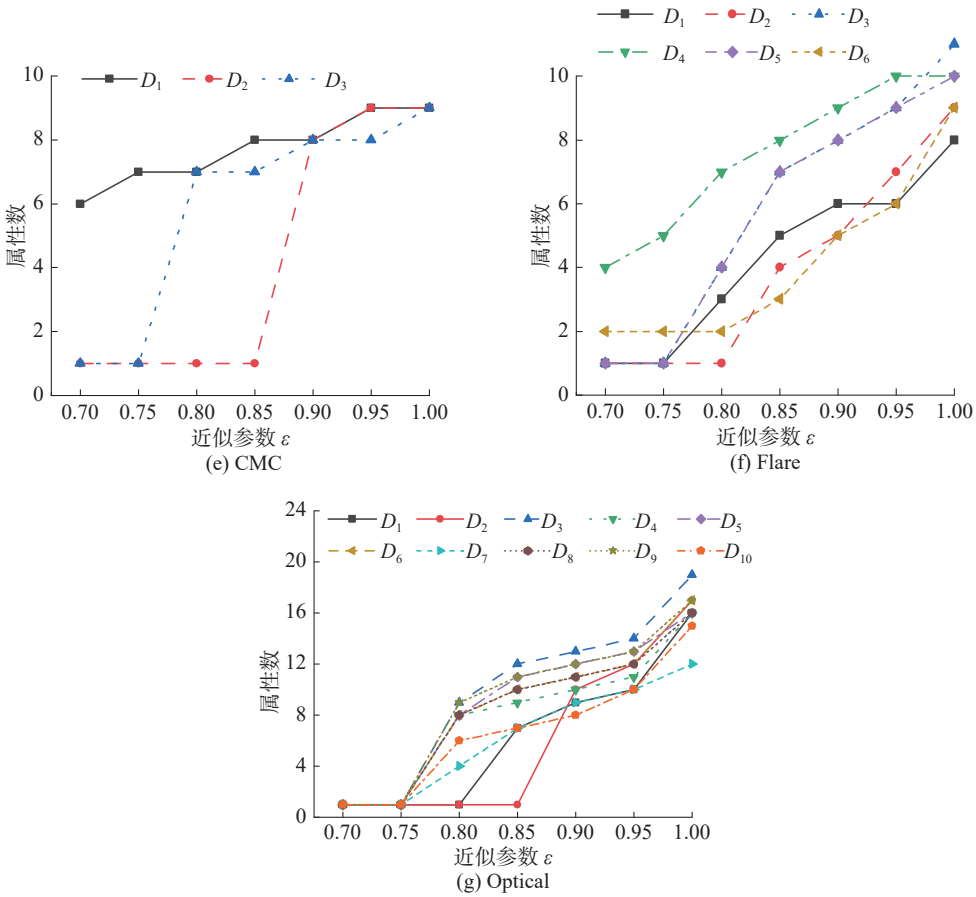


图 3　正态分布下特定类约简子集属性数随着近似参数 (ε) 的变化曲线

Fig. 3　Nnumber of attributes o f class-specific reduct varying with the approximate parameter ε under normal distribution

通过图 2 和 3 可以看出，随着近似参数ε逐渐增大，约简子集中属性的数目也呈递增的趋势。这是因为，近似参数越大，约简的度量标准就会越高，为了满足度量标准就会有更多的属性加入到约简子集中，这也验证了所提出的相对不确定度具有单调性。

考虑到近似参数过小时，所得到的约简子集很容易缺失关键信息，而如果近似参数过大，得到的约简子集就会趋于度量保持约简，在后续实验中将近似参数ε设置为 0.95。

表 2 是在均匀分布和正态分布下各算法所选取的约简子集经过 KNN 分类器 (K=5)10 次十折交叉验证所得到的平均分类精度，其中每个数据集名称后的数字代表的该数据集对应的决策类，例如 Cardio(1) 代表 Cardio 数据集的第 1 个决策类，RAW 列表示原始数据上的结果。

表 2　2 种分布下 4 种算法得到的约简集分类精度

Table 2　The classification accuracy of reduct obtained by four algorithms under two kinds of distribution

数据集	RAW	CSB2CSR	UNC-CSR	MI-CSR	RUNC-CSCAR	
					均匀分布	正态分布
Cardio(1)	**0.958 0**	0.958 1	0.956 9	0.957 5	**0.960 1**	0.957 6
Cardio(2)	0.644 4	0.633 7	0.628 3	0.637 6	**0.649 8**	**0.653 1**
Cardio(3)	**0.708 2**	0.664 5	0.689 9	0.696 1	0.669 5	0.698 4
Credit(1)	0.923 1	0.923 1	0.923 1	0.923 1	**0.923 5**	**0.924 4**
Credit(2)	**0.310 8**	**0.310 8**	**0.310 8**	**0.310 8**	0.308 3	**0.310 8**
Turkiye(1)	0.717 4	0.719 1	0.723 4	0.729 6	**0.808 3**	**0.839 7**
Turkiye(2)	0.196 7	0.197 1	**0.202 8**	0.193 8	**0.203 4**	0.154 9
Turkiye(3)	0.277 9	**0.289 5**	0.286 1	0.280 3	0.266 2	0.270 7

续表 2

数据集	RAW	CSB2CSR	UNC-CSR	MI-CSR	RUNC-CSCAR	
					均匀分布	正态分布
CMC(1)	**0.810 1**	**0.810 1**	**0.810 1**	**0.810 1**	**0.810 1**	**0.810 1**
CMC(2)	0.233 3	**0.233 3**	**0.233 3**	**0.233 3**	**0.247 4**	**0.233 3**
CMC(3)	0.152 7	**0.152 7**	**0.152 7**	**0.152 7**	**0.154 2**	0.133 8
Biodeg(1)	**0.791 7**	0.765 5	0.783 2	0.785 0	**0.795 0**	0.773 9
Biodeg(2)	**0.856 0**	0.783 3	0.808 6	0.855 7	0.804 9	0.823 4
Flare(1)	0.812 5	0.814 3	0.805 9	**0.822 8**	0.765 0	0.804 8
Flare(2)	0.510 3	0.450 6	**0.516 9**	0.489 5	0.420 6	0.416 3
Flare(3)	0.431 8	0.465 9	0.456 5	0.462 3	**0.486 9**	**0.479 5**
Flare(4)	0.809 7	0.818 9	0.817 6	0.814 4	**0.829 6**	**0.823 3**
Flare(5)	0.636 5	0.630 2	0.633 3	0.629 6	**0.637 4**	**0.642 2**
Flare(6)	**0.573 8**	0.529 4	0.563 1	0.563 7	**0.577 7**	0.556 3
Optical(1)	**0.665 4**	0.660 9	0.626 6	0.663 2	0.591 4	0.595 1
Optical(2)	0.574 6	0.563 4	0.554 3	0.572 6	**0.587 4**	**0.596 5**
Optical(3)	0.558 1	0.554 3	0.546 3	0.551 8	0.512 7	**0.559 7**
Optical(4)	**0.557 1**	0.524 8	0.533 3	0.550 0	**0.567 7**	0.475 6
Optical(5)	0.532 1	0.554 0	0.573 7	**0.603 2**	0.554 5	0.514 3
Optical(6)	**0.563 1**	0.527 5	0.518 0	0.558 4	0.453 7	**0.612 9**
Optical(7)	0.598 9	0.586 9	0.526 1	0.606 2	**0.618 2**	**0.613 3**
Optical(8)	0.571 3	0.580 3	0.564 4	**0.610 6**	0.564 7	0.530 4
Optical(9)	0.441 6	0.445 1	0.444 6	0.447 9	**0.457 5**	**0.472 7**
Optical(10)	0.508 2	**0.537 4**	0.485 0	0.536 4	**0.547 3**	0.422 3

注: 粗体表示不同算法取得的最高精度。

在表 2 中, 其他几种算法由于没有考虑代价, 故而在 2 种测试代价分布下, 都得到了相同的属性约简集, 所得分类精度也相同。可以发现, CSB2-CSR 算法在多数决策类上的精度相对较差, UNC-CSR 算法与 MI-CSR 算法的表现相对较好一点, 本研究方法 RUNC-CSCAR 比其他 3 种算法结果都要更好。这主要是由于 CSB2CSR 算法首先进行了面向决策的属性约简, 并将该约简子集作为所有决策类共同的初始约简集, 随后利用删除策略对各个决策类进行属性约简, 因此每个决策类所得到的约简集大部分是相同的, 故而影响到多数决策类的分类精度。UNC-CSR 算法与 MI-CSR 算法则是基于各个决策类进行的属性约简, 故而与 CSB2CSR 算法的精度相比表现较好, 但由于过于严苛的属性约简而对噪声数据缺乏良好的适应能力, 故而相较 RUNC-CSCAR 算法表现较弱。本研究提出的 RUNC-CSCAR 算法引入了近似属性约简的思想, 通过放松相对不确定度避免部分冗余属性进入约简集, 对噪声数据具有良好的适应能力, 故而在大多数的决策类以及 2 种代价分布下相较于其他算法都取得了更高的分类精度。

表 3 是 2 种分布下各个算法所得到的特定类约简子集的规模, 即条件属性的数目。与分类精度实验结果相似, 其他几种算法由于没有考虑代价, 故而在 2 种测试代价分布情况下都取得了相同的属性约简集。研究发现本研究提出的 RUNC-CSCAR 算法所获得的约简集规模要比 CSB2CSR 算法、UNC-CSR 算法以及 MI-CSR 算法的约简集规模都更加精简。CSB2CSR 算法是在面向决策的属性约简的基础上针对各决策类利用删除策略进行约简的, 故而得到的约简属性数目较多, UNC-CSR 算法以及 MI-CSR 算法则是因为严格的约简要求, 在约简后期使一些对度量贡献值很小的冗余属性加入约简子集, 因此获得的约简属性数也相对较多。而本研究所提出的 RUNC-CSCAR 方法利用了近似属性约简的思想, 使用近似参数 ε 来指导属性约简的放松, 避免更多的冗余属性被添加到约简集中, 所以相比于其他 3 种算法能够获得更精简的约简子集。

表 3　2 种分布下 4 种算法得到的约简属性数

Table 3　Number of attributes in reduct obtained by four algorithms under two kinds of distribution

数据集	RAW	CSB2CSR	UNC-CSR	MI-CSR	RUNC-CSCAR	
					均匀分布	正态分布
Cardio(1)	21	17	17	17	**16**	**16**
Cardio(2)	21	18	19	**17**	**17**	**16**
Cardio(3)	21	**15**	17	16	16	16
Credit(1)	23	23	23	23	**21**	**21**
Credit(2)	23	23	23	23	**20**	**20**
Turkiye(1)	31	15	15	14	**9**	**9**
Turkiye(2)	31	16	16	14	**13**	**13**
Turkiye(3)	31	16	16	15	**13**	**14**
CMC(1)	**9**	**9**	**9**	**9**	**9**	**9**
CMC(2)	9	9	9	9	**8**	**9**
CMC(3)	9	9	9	9	**8**	**8**
Biodeg(1)	41	15	14	**12**	13	14
Biodeg(2)	41	15	14	**13**	**13**	**13**
Flare(1)	12	9	10	8	**6**	**6**
Flare(2)	12	8	10	9	**7**	**7**
Flare(3)	12	10	10	10	**8**	**9**
Flare(4)	12	**10**	**10**	**10**	**10**	**10**
Flare(5)	12	**9**	11	**9**	**9**	**9**
Flare(6)	12	**1**	10	9	6	6
Optical(1)	64	21	15	14	**10**	**10**
Optical(2)	64	21	14	15	**12**	**12**
Optical(3)	64	21	**14**	**14**	**14**	**14**
Optical(4)	64	21	17	14	**13**	**11**
Optical(5)	64	18	17	**12**	14	13
Optical(6)	64	20	17	**13**	**13**	**12**
Optical(7)	64	18	13	**11**	**11**	**10**
Optical(8)	64	21	15	14	**11**	**12**
Optical(9)	64	20	16	14	**12**	**13**
Optical(10)	64	21	17	14	**11**	**10**

注：粗体表示不同算法约简属性数最少。

表 4 是 2 种分布下各个算法所得到的特定类约简子集的测试代价。本研究所提出的 RUNC-CSCAR 算法相比于其他算法在绝大部分决策类中都获得了更小的测试代价。这是由于其他 3 种算法并没有在属性选择时将测试代价作为参考标准，而是单一地将属性重要度作为评判标准，并且这 3 种算法所获得的属性约简集规模相较于本研究方法所获得的属性约简集规模更大。本研究考虑到现实中可能希望获取的属性约简集既能够保持约简精度，又能够同时具有较低的测试代价，将测试代价引入到属性重要度的评判中，且利用近似属性约简的思想避免了冗余属性的加入，故而相比于其他 3 种算法最终能够获取测试代价较低的约简。

表 4　2 种分布下 4 种算法得到的约简集代价

Table 4　Cost of reduct obtained by four algorithms under two kinds of distribution

数据集	均匀分布					正态分布				
	RAW	CSB2CSR	UNC-CSR	MI-CSR	RUNC-CSCAR	RAW	CSB2CSR	UNC-CSR	MI-CSR	RUNC-CSCAR
Cardio(1)	684	561	516	559	**468**	538	435	429	436	**400**
Cardio(2)	684	583	588	543	**500**	538	463	463	439	**409**

续表 4

数据集	均匀分布					正态分布				
	RAW	CSB2CSR	UNC-CSR	MI-CSR	RUNC-CSCAR	RAW	CSB2CSR	UNC-CSR	MI-CSR	RUNC-CSCAR
Cardio(3)	684	467	500	495	**454**	538	**379**	467	413	410
Credit(1)	729	729	729	729	**632**	577	577	577	577	**527**
Credit(2)	729	729	729	729	**591**	577	577	577	577	**494**
Turkiye(1)	722	310	242	319	**79**	769	371	348	372	**193**
Turkiye(2)	722	341	256	314	**147**	769	404	372	355	**276**
Turkiye(3)	722	341	278	345	**208**	769	404	376	374	**332**
CMC(1)	286	286	286	286	**286**	224	224	224	224	**224**
CMC(2)	286	286	286	286	**254**	224	224	224	224	**224**
CMC(3)	286	286	286	286	**254**	224	224	224	224	**200**
Biodeg(1)	1 067	359	363	297	**211**	1 044	380	349	**299**	345
Biodeg(2)	1 067	359	384	283	**204**	1 044	380	355	321	**315**
Flare(1)	396	314	343	272	**202**	320	242	271	219	**169**
Flare(2)	396	275	343	314	**228**	320	211	271	242	**175**
Flare(3)	396	343	343	343	**292**	320	271	271	271	**241**
Flare(4)	396	343	343	343	**343**	320	271	271	271	**271**
Flare(5)	396	314	355	314	**314**	320	242	296	242	**242**
Flare(6)	396	**49**	343	314	197	320	**26**	271	242	153
Optical(1)	1 649	433	277	300	**73**	1 604	549	381	358	**191**
Optical(2)	1 649	433	331	332	**156**	1 604	549	346	370	**264**
Optical(3)	1 649	433	274	390	**198**	1 604	549	376	345	**326**
Optical(4)	1 649	433	474	325	**172**	1 604	549	424	333	**230**
Optical(5)	1 649	335	367	243	**151**	1 604	473	434	**293**	300
Optical(6)	1 649	400	437	305	**136**	1 604	523	415	326	**260**
Optical(7)	1 649	335	329	284	**128**	1 604	473	368	241	**237**
Optical(8)	1 649	433	350	412	**105**	1 604	549	369	383	**273**
Optical(9)	1 649	400	335	283	**127**	1 604	523	419	353	**289**
Optical(10)	1 649	433	426	317	**92**	1 604	549	408	363	**213**

注: 粗体表示不同算法约简集代价最小。

总的来说, 本研究所提出的特定类的代价敏感近似属性约简算法利用了近似属性约简的思想, 放宽了属性约简的条件, 避免了更多冗余属性进入约简子集, 提高了对噪声数据的适应能力, 同时也保留了绝大部分的有效信息, 并且倾向于选择测试代价更小的条件属性。因此对大部分决策类来说本研究的方法能够在获取更精简的约简子集和更小的测试代价的同时提升约简子集的分类精度。

4 结束语

现有面向特定类的属性约简要求通常过于严苛, 限制了其在一些场景下的应用, 缺乏对噪声数据的适应能力。因此, 本研究在近似参数的指导下放宽了属性约简的要求, 提出了面向特定类的代价敏感近似属性约简方法, 并由此给出了相应的属性约简启发式算法。通过多个 UCI 数据集上的实验, 讨论了惩罚指数和近似参数对近似属性约简算法约简结果的影响, 同时与其他特定类属性约简算法所得约简的分类精度、测试代价和约简集规模进行了对比。实验结果表明, 本研究方法可以有效减小约简属性子集的规模并提升分类精度, 同时还能获得相对更小的测试代价。在一些问题中, 数据中的各个类别有可能是不平衡的, 特别是少数类样本数量过少可能导致属性约简集存在过拟合的问题, 本研究尚未对此进行探讨, 如何消除数据不平衡带来的问题是后续需要进一步研究的问题。

参考文献：

[1] PAWLAK Z. Rough sets[J]. International journal of computer & information sciences, 1982, 11(5): 341−356.

[2] ZHANG Pengfei, LI Tianrui, WANG Guoqiang, et al. Multi-source information fusion based on rough set theory: a review[J]. Information fusion, 2021, 68: 85−117.

[3] DIXIT V, CHAUDHURI A, SRIVASTAVA R K. Assessing value of customer involvement in engineered-to-order shipbuilding projects using fuzzy set and rough set theories[J]. International journal of production research, 2019, 57(22): 6943−6962.

[4] DHAL K G, DAS A, RAY S, et al. An analytical review on rough set based image clustering[J]. Archives of computational methods in engineering, 2022, 29(3): 1643−1672.

[5] HALDER B, MITRA S, MITRA M. Development of cardiac disease classifier using rough set decision system[C]// ABRAHAM A, DUTTA P, MANDAL J, et al. Emerging Technologies in Data Mining and Information Security. Singapore: Springer, 2019: 775−785.

[6] MAO Hua, WANG Shengyu, LIU Chang, et al. Hypergraph-based attribute reduction of formal contexts in rough sets[J]. Expert systems with applications, 2023, 234: 121062.

[7] DAI Jianhua, HUANG Weiyi, WANG Weisi, et al. Semi-supervised attribute reduction based on label distribution and label irrelevance[J]. Information fusion, 2023, 100: 101951.

[8] PAWLAK Z. Rough sets: theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[9] STEPANIUK J. Approximation spaces, reducts and representatives[M]// Rough Sets in Knowledge Discovery 2. Heidelberg: Physica, 1998: 109−126.

[10] YAO Yiyu, ZHANG Xianyong. Class-specific attribute reducts in rough set theory[J]. Information sciences, 2017, 418/419: 601−618.

[11] MA Xiao, YAO Yiyu. Three-way decision perspectives on class-specific attribute reducts[J]. Information sciences, 2018, 450: 227−245.

[12] ZHANG Xianyong, TANG Xiao, YANG Jilin, et al. Quantitative three-way class-specific attribute reducts based on region preservations[J]. International journal of approximate reasoning, 2020, 117: 96−121.

[13] ZHANG Xianyong, YANG Jilin, TANG Lingyu. Three-way class-specific attribute reducts from the information viewpoint[J]. Information sciences, 2020, 507: 840−872.

[14] 吴婉琳, 张贤勇, 莫智文. 基于粗糙集不确定度的特定类属性约简[J]. 四川师范大学学报 (自然科学版), 2021, 44(6): 840−846.
WU Wanlin, ZHANG Xianyong, MO Zhiwen. Class-specific attribute reducts based on uncertainty degrees of rough sets[J]. Journal of Sichuan normal university (natural science), 2021, 44(6): 840−846.

[15] 陈阳, 张楠, 孙雪姣, 等. 基于特定类不完备决策系统的分布约简[J]. 计算机应用研究, 2020, 37(9): 2659−2664.
CHEN Yang, ZHANG Nan, SUN Xuejiao, et al. Class-specific distribution reduction in incomplete decision systems[J]. Application research of computers, 2020, 37(9): 2659−2664.

[16] ZHANG Xianyong, FAN Yunrui, YAO Yuesong, et al. Class-specific attribute reducts based on neighborhood rough sets[J]. Journal of intelligent & fuzzy systems, 2022, 43(6): 7891−7910.

[17] MA Xiao. Fuzzy entropies for class-specific and classification-based attribute reducts in three-way probabilistic rough set models[J]. International journal of machine learning and cybernetics, 2021, 12(2): 433−457.

[18] ZHANG Xianyong, YAO Hong, LYU Zhiying, et al. Class-specific information measures and attribute reducts for hierarchy and systematicness[J]. Information sciences, 2021, 563: 196−225.

[19] 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.

[20] LI Huaxiong, ZHOU Xianzhong, HUANG Bing, et al. Cost-sensitive three-way decision: a sequential strategy[C]// Rough Sets and Knowledge Technology: 8th International Conference, RSKT 2013, Berlin: Springer Berlin Heidelberg, 2013: 325−337.

[21] YAO Yiyu, ZHAO Yan. Attribute reduction in decision-theoretic rough set models[J]. Information sciences, 2008, 178(17): 3356−3373.

[22] MIN Fan, HE Huaping, QIAN Yuhua, et al. Test-cost-sensitive attribute reduction[J]. Information sciences, 2011, 181(22): 4928−4942.

[23] 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.

[24] JU Hengrong, YANG Xibei, YU Hualong, et al. Cost-sensitive rough set approach[J]. Information sciences, 2016, 355/356: 282−298.

[25] JU Hengrong, LI Huaxiong, YANG Xibei, et al. Cost-sensitive rough set: a multi-granulation approach[J]. Knowledge-based systems, 2017, 123: 137−153.

[26] MA Xiao, ZHAO Xuerong. Cost-sensitive three-way class-specific attribute reduction[J]. International journal of approximate reasoning, 2019, 105: 153−174.

作者简介：

胡军，教授，博士，主要研究方向为多粒度认知计算、人工智能安全和图分析与挖掘。发表学术论文 80 余篇。E-mail: hujun@cqupt.edu.cn。

黄小涵，硕士研究生，主要研究方向为粒计算、粗糙集。E-mail: 13417-56280@qq.com。