中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

融合粒度分组与Pareto最优的属性选择

印振宇, 王平心, 杨习贝, 于化龙, 钱宇华

引用本文:
印振宇,王平心,杨习贝,于化龙,钱宇华. 融合粒度分组与Pareto最优的属性选择[J]. 控制与决策, 2024, 39(9): 2959−2968.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1505

融合粒度分组与Pareto最优的属性选择

印振宇[1], 王平心[2], 杨习贝[1†], 于化龙[1], 钱宇华[3]

(1. 江苏科技大学 计算机科学与工程学院，江苏 镇江 212100；
2. 江苏科技大学 数理学院，江苏 镇江 212100；
3. 山西大学 计算机与信息技术学院，太原 030006)

摘 要: 利用某一给定度量作为属性评价指标以及启发式算法的约束条件,是大量属性选择方案的关键. 然而,属性相似性评价的缺失与朴素的逐个选择机制,使属性遍历存在冗余,故时间消耗巨大. 此外,单一度量限制了属性评价视角,难以挖掘出高学习性能的属性. 鉴于此,提出一种属性选择框架,其中: 1)利用属性粒度及属性间的知识距离对属性分组,组内属性具有明显差异性且组间属性具有较强区分能力,使属性遍历以组为单位,有效压缩候选属性搜索空间,提升属性选择效率; 2)利用提出的受限Pareto最优原则,对属性组进行迭代选取,最终得到期望的属性子集. 在12组UCI 数据集上,通过注入4种不同比例的属性噪声进行实验,结果表明: 相较于8 种流行方法,所提出方法得到的属性选择结果,在分类稳定性这一指标上平均提升了5.89 %,在分类准确率这一指标上平均提升了12.28 %,在时间消耗这一指标上平均降低了59.27 %.

关键词: 属性选择；粒度；启发式算法；启发式信息；邻域粗糙集；Pareto最优

中图分类号: TP182 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1505

引用格式: 印振宇,王平心,杨习贝,等. 融合粒度分组与Pareto最优的属性选择[J]. 控制与决策, 2024, 39(9): 2959-2968.

Attribute selection based on granularity grouping and Pareto optimality

YIN Zhen-yu[1], *WANG Ping-xin*[2], *YANG Xi-bei*[1†], *YU Hua-long*[1], *QIAN Yu-hua*[3]

(1. School of Computer，Jiangsu University of Science and Technology，Zhenjiang 212100，China；2. School of Science, Jiangsu University of Science and Technology，Zhenjiang 212100，China；3. School of Computer and Information Technology，Shanxi University，Taiyuan 030006，China)

Abstract: The key to numerous attribute selection methods lies in the utilization of a given measure as the attribute evaluation criterion, along with the constraints of heuristic algorithms. However, the absence of attribute similarity evaluation and the simplistic sequential selection mechanism result in redundant attribute traversal, leading to significant time consumption. Additionally, the use of a single measure limits the perspective of attribute evaluation, making it difficult to unearth attributes with high learning performance. In view of this, a framework for attribute selection is proposed, where: 1) Attribute grouping is performed based on attribute granularity and knowledge distance between attributes. Within each group, the attributes exhibit significant differences, while between groups, the attributes possess strong discriminative power. This allows attribute traversal to be conducted at the group level, effectively compressing the search space of candidate attributes and improving attribute selection efficiency. 2) The proposed restricted Pareto optimality principle is utilized to iteratively select attribute groups, ultimately obtaining the desired subset of attributes. In experiments conducted on 12 UCI datasets by injecting four different levels of attribute noise, the results show that compared to 8 popular methods, the proposed approach yields attribute selection results with an average improvement of 5.89 % in classification stability, an average improvement of 12.28 % in classification accuracy, and an average reduction of 59.27 % in time consumption.

Keywords: attribute selection；granularity；heuristic algorithm；heuristic information；neighborhood rough set；Pareto optimality

收稿日期: 2022-08-23；**录用日期:** 2023-09-14.
基金项目: 国家自然科学基金项目(62076111)；江苏省研究生实践创新计划项目(SJCX22_1905).
†通讯作者. E-mail: jsjxy_yxb@just.edu.cn.

0 引 言

属性约简[1-2]是一种基于粗糙集[3-4]的属性选择方法.在机器学习领域中,可视其为一种特征降维技术.一般而言,属性约简的任务为:在原始属性合集的范畴内,设计一个搜索过程,通过给定的约束条件,尽可能多地去除质量较低的属性,从而获得一个满足约束条件的最小属性子集.作为一种有效的数据预处理方法,属性约简具有以下优点:利用较少的属性揭示原始属性合集上的结构信息,继而简化后续的数据分析过程,以提升学习任务的泛化能力.

属性约简的研究主要有两方面:

1)提升获取约简的效率[5];

2)改善约简所具备的性能[6].

需指出的是,输出的约简子集由两部分共同决定:搜索策略与属性评价. 在求解约简常用的启发式搜索策略[7]中,利用启发式信息指导属性评估与搜索方向,从而避免属性的组合爆炸,最终得到一个满足约束条件的属性子集. 此外,从不同需求与约简定义的角度出发,多数学者根据不同的属性评价视角,提出了不同的度量准则,如依赖度[8]、条件熵[9]、决策错误率[10] 等.

基于上述背景的约简求解策略往往具有以下缺点:

1)忽略了约简池中已有属性与候选属性之间的相似性;

2)在每次迭代过程中,都需要评价所有侯选属性.

显然,当此类逐个尝试策略面临高维数据时,其求解约简的时间消耗将显著上升,求解效率因此降低. 其原因有两方面:

1)对于约简池中的属性,在算法执行的多次迭代过程中,重复评判了与其强相关的侯选属性;

2)利用固定的度量对属性进行评价,可能会忽视与该度量弱关联的、具备优秀分类性能的属性,这将在一定程度上阻碍后续的学习任务.

为了解决上述问题,本文提出一类基于粒度分组与Pareto最优相结合的约简搜索策略. 不同于使用固定度量的启发式约简求解方法,首先,为了高效评估属性自身所具备的鉴别信息,利用属性的粒度信息对侯选属性进行排序;其次,在排序的基础上,考虑到属性之间的相关性,利用属性间的结构信息,对属性进行分组;最后,在分组的基础上,利用新提出的受限Pareto 最优策略进行属性选取. 本文工作有如下优势:在属性选择的过程中,属性的遍历以组而非单个属性为基本单位,从而有效地压缩候选属性的搜索空间,进而在不降低、甚至提升约简分类性能的情况下,同时提高属性选择的效率.

1 基础知识

1.1 邻域粗糙集

邻域粗糙集中,邻域决策系统表示为$\mathrm{DS}=\langle U,\mathrm{AT}\bigcup\{d\}\rangle$:$U$为一个非空样本集;$\mathrm{AT}=\{a_k|1\leqslant k\leqslant m\}$为非空条件属性集;$d$用以记录样本标签,称为决策属性,表示为$L=\{l_p|1\leqslant p\leqslant q\}$,$\forall x_i\in U$,$d(x_i)\in L$表示样本$x_i$的标签. 通过$d$可以构建一个形如$\mathrm{IND}(d)=\{(x_i,x_j)\in U^2|d(x_i)=d(x_j)\}$的等价关系,此时$U$被划分为一组不相交且形如$U/\mathrm{IND}(d)=\{X_1,X_2,\ldots,X_q\}$的等价类. $\forall X_p\in U/\mathrm{IND}(d)$,$X_p$表示所有标签为$l_p$的样本所构成的集合.

定义1 给定一个DS,$\forall A\subseteq\mathrm{AT}$,$A$诱导出一个邻域关系为

$$\delta_A=\{(x_i,x_j)\in U^2|\Delta_A(x_i,x_j)\leqslant\delta\}. \tag{1}$$

其中:$\Delta_A(x_i,x_j)$为样本x_i与x_j之间的距离,$\delta>0$为给定的邻域半径.

根据邻域关系δ_A,样本x_i的邻域记为$\delta_A(x_i)=\{x_j|x_j\in U,\Delta(x_i,x_j)\leqslant\delta_A\}$.

定义2 给定一个DS,$\forall A\subseteq\mathrm{AT}$,$\forall X_p\in U/\mathrm{IND}(d)$,$X_p$关于$A$的上、下近似分别定义为

$$\overline{\delta_A}(X_p)=\{x_i\in U|\delta(x_i)\bigcap X_p\neq\varnothing\}, \tag{2}$$

$$\underline{\delta_A}(X_p)=\{x_i\in U|\delta(x_i)\subseteq X_p\}. \tag{3}$$

根据式(2)和(3),可得到d的邻域上、下近似为

$$\overline{\delta_A}(d)=\bigcup_{p=1}^{q}\overline{\delta_A}(X_p), \tag{4}$$

$$\underline{\delta_A}(d)=\bigcup_{p=1}^{q}\underline{\delta_A}(X_p). \tag{5}$$

1.2 依 赖 度

粗糙集理论中,下近似用以刻画数据的确定性程度. 可定义一个称为依赖度的度量.

定义3 给定一个DS,$\delta>0$,$\forall A\subseteq\mathrm{AT}$,决策属性$d$关于$A$的依赖度为

$$\gamma^{\delta}(A,U)=\frac{|\underline{\delta_A}(d)|}{|U|}. \tag{6}$$

基于式(6),文献[11-12]提出了局部依赖度的概念.

定义4 给定一个DS和邻域半径$\delta>0$,$\forall A\subseteq$

AT, $\forall l_p \in L$, 标签l_p的局部依赖度为

$$\gamma^{\delta}(A, l_p) = \frac{|\delta_A(X_p)|}{|X_p|}. \tag{7}$$

根据式(7), 可得到一个局部依赖度矩阵 $\boldsymbol{M}_{\gamma} = \{\gamma^{\delta}(a_k, l_p)|1 \leqslant k \leqslant m, 1 \leqslant p \leqslant q\}$.

1.3 属性约简

属性约简意在减少冗余或者不相关的条件属性, 一般定义如下.

定义5 给定一个DS和邻域半径$\delta > 0, \forall \text{red} \subseteq \text{AT}$, red为DS中的一个$\epsilon$-约简当且仅当:

1) $\gamma^{\delta}(\text{red}, U)/\gamma^{\delta}(\text{AT}, U) \geqslant \epsilon$;

2) $\forall \text{red}' \subset \text{red}, \gamma^{\delta}(\text{red}', U)/\gamma^{\delta}(\text{AT}, U) < \epsilon$.

依赖度值随条件属性个数的减少而单调递减. 一般使用阈值$\epsilon \in [0, 1]$来控制这种递减程度.

搜索策略[13-15]对获取高质量约简至关重要, 核心为: 如何评估或衡量候选属性的重要性. 借助依赖度的变化, 可对重要性进行如下度量.

定义6 给定一个DS和邻域半径$\delta > 0, \forall A \subseteq \text{AT}, \forall a_k \in \text{AT} - A$, 属性$a_k$相对于条件属性集合$A$的重要性为

$$\text{Sig}(U, A, a_k, d) = \gamma^{\delta}(A \bigcup \{a_k\}, U) - \gamma^{\delta}(A, U). \tag{8}$$

式(6)中的值越高, 意味着候选属性a_k越重要.

1.4 约简的前向贪心求解

目前, 为兼顾属性约简求解的效率和性能, 前向贪心搜索是一种常用的技术. 一般流程如下[16].

算法1 前向贪心搜索.

step 1: 初始化参数;

step 2: 逐步向约简集red中添加候选属性中重要性最大的属性, 直到$\gamma^{\delta}(\text{red}, U)/\gamma^{\delta}(\text{AT}, U) > \epsilon$;

step 3: 若$\gamma^{\delta}(\text{red}, U)/\gamma^{\delta}(AT, U) > \epsilon$在去除red中任一属性$a_k$后仍成立, 则去除该属性, 直到red不再改变或$|\text{red}| = 1$.

2 研究方法

2.1 粒 度

近年来, 粒度的表示与计算得到了深入的探索, 基于粒度的工具也被引进到许多复杂的机器学习任务中[17-21].

定义7 给定一个二元组$S = \langle U, R\rangle$, R是U上的一个二元关系, $\forall x_i \in U$, x_i的R-相关集记为

$$R(x_i) = \{x_j \in U|(x_i, x_j) \in R\}. \tag{9}$$

定义8 给定一个二元组$S = \langle U, R\rangle$, 二元关系R诱导出的粒度定义为

$$G_R(U) = \frac{\sum\limits_{x_i \in U} |R(x_i)|}{|U|^2}. \tag{10}$$

由式(9)可知$0 \leqslant G_R(U) \leqslant 1$. 二元关系$R$对应的粒度清楚地揭示了信息粒化结果(所有的$R$-相关集)的可区分性. $G_R(U)$的值越小, R具有越强的区分性, 这主要是因为较小的$G_R(U)$对应的二元关系R包含较少的有序对, 此时U中的大多数样本都可以被区分开来.

因为式(1)所示的邻域关系δ_A也是一种二元关系, 所以由定义8不难得到如下所示的基于邻域关系的粒度.

定义9 给定一个邻域决策系统DS和邻域半径$\delta > 0, \forall A \subseteq \text{AT}$, 邻域关系$\delta_A$的粒度为

$$G_{\delta}(A, U) = \frac{\sum\limits_{x_i \in U} |\delta_A(x_i)|}{|U|^2}. \tag{11}$$

2.2 知识距离

虽然粒度能够评价条件属性集在样本空间上的粒化能力, 但为了进一步考虑如何量化地区分不同条件属性集的粒化能力, Qian等[22-23]提出了如下所示的知识距离的概念.

定义10 给定一个DS和邻域半径$\delta > 0, \forall A, B \subseteq \text{AT}$, A与B之间的邻域知识距离可定义为

$$D(A, B) = \frac{\sum\limits_{x_i \in U} \dfrac{|\delta_A(x_i) \bigcup \delta_B(x_i)| - |\delta_A(x_i) \bigcap \delta_B(x_i)|}{|U|}}{|U|}. \tag{12}$$

由式(13)可知$0 \leqslant D(A, B) \leqslant 1$. 知识距离清晰地刻画了由条件属性集合$A$与$B$分别引导的两种知识结构的差异性程度. $D(A, B)$的值越大, 表示信息粒化结果之间的差异性越大. 这主要是因为较大的$D(A, B)$对应的两种样本邻域包含更小的交集, 交集的基数越小, 并集基数不变, 于是分子部分就越大, 故而信息粒化结果的差异性越大, 知识距离值越大.

2.3 Pareto最优与受限Pareto最优

Pareto最优概念[24-25]源于经济学, 现已被应用于至各种科学领域. 假设有一组等待分配资源的对象与一组待分配的资源, 执行从一种分配方案到另一种分配方案的进程中, 在没有使任何对象变坏的情况下, 使至少一个对象变好, 这称为Pareto改进. 而没有Pareto改进余地的状态被称为Pareto最优.

一般而言, Pareto最优问题存在两个关键点: 分配方案与维度. 通过在不同维度上对不同分配方案

的比较,可得到Pareto最优的核心概念:支配方案与非支配方案. 据此,可以进行如下假设:1) $\mathrm{AT}=\{a_k|1\leqslant k\leqslant m\}$,所有属性的合集是分配方案集;2) $L=\{l_p|1\leqslant p\leqslant q\}$,所有标签的合集是用于在不同的分配方案之间进行比较的维度集.

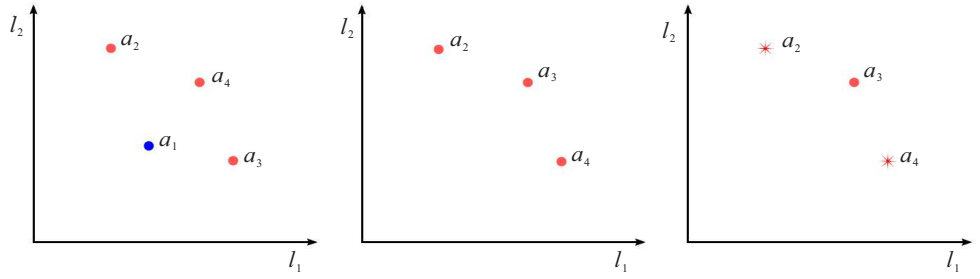
根据Pareto最优的基本定义,可得到:

1) $\forall a_k\in\mathrm{AT}$,如果在$\mathrm{AT}-\{a_k\}$中,存在比$a_k$在所有维度上(所有标签上)表现都要优秀的分配方案,则称a_k为非支配方案;

2) $\forall a_k\in\mathrm{AT}$,如果$a_k$在至少一个维度上(一个标签上)表现是最优的或者在$\mathrm{AT}-\{a_k\}$中不存在比$a_k$在所有维度上都要优秀的分配方案,则称$a_k$为支配方案.

Pareto最优分配方案集即为所有支配方案的合集. 图1(a)给出了二维空间(两个标签)上的支配方案(红色点)与非支配方案(蓝色点)的例子,不同的点表示不同的分配方案,一个点在一个维度上的值越大,说明它在这个维度上越优. 需要注意的是,可能存在以下特殊情况:初始条件属性均隶属于Pareto最优分配方案集条件属性的选择失去意义(图1(b)). 鉴于此,提出受限Pareto最优的概念:将那些至少在一个维度上已经是最优的分配方案从Pareto最优分配方案集中选出,加入受限Pareto最优分配方案集(图1(c)).

图1(b)中,分配方案a_2与a_4分别在l_2与l_1上是最优的,且不存在一个分配方案在两个维度上都优于a_3,故a_2、a_3、a_4构成了Pareto最优分配方案集. 图1(c)中,根据受限Pareto最优的概念,在图1(b)基础上剔除了a_3,由a_2、a_4构成受限Pareto最优分配方案集. 进一步地,可将1.2节所示的矩阵$\boldsymbol{M}_\gamma$以上述形式映射到多维空间中:点表示条件属性,不同维度表示矩阵中的不同决策类(标签),各个点在不同维度上的投影即为不同属性关于不同标签的局部依赖度,在此基础上应用受限Pareto最优原则可以筛选条件属性.



(a) 支配与非支配方案 (b) Pareto 最优分配方案集 (c) 受限Pareto 最优分配方案集

图 1 Pareto最优与改进

3 融合粒度分组与Pareto最优的属性约简

虽然算法1能够快速地完成属性约简的任务,但其存在以下问题:1)当某候选属性与已选属性具有较高的相似度时,冗余遍历将带来不必要的时间消耗,且不会带来分类性能的显著提升;2)固定的度量阻碍了从不同视角对条件属性进行评价.

针对上述问题,提出一种采用粒度分组与Pareto最优相结合的约简搜索策略,具体过程为:

1)计算每个条件属性所对应的粒度,并按粒度值从小到大对条件属性进行排序.

2)将产生最小粒度值的条件属性放置到最新空组的第1个位置.

3)依次计算剩余未入组的条件属性与该组第1个条件属性之间的知识距离,计算它们的平均值,将知识距离值大于该平均值的条件属性添加至该组中.

4)另起一组在剩余的条件属性中重复步骤2)和步骤3),直至所有条件属性都被分到不同的组中.

5)从第1组开始,将每组的条件属性分批次加入到约简池中,并判断约简池中的属性在当前批次下是否满足给定约束,若满足,则转步骤6).

6)为防止所得约简是超约简这一情形,在约简池中利用受限Pareto最优挑选出相应的属性,判断这些属性所构成的合集是否满足给定约束,若满足,则进一步执行删除属性的操作;若不满足,则从候选属性中继续挑选,加入重要属性直至满足约束条件,再进一步执行删除属性的操作.

与启发式贪心搜索策略不同的是,分组机制使每轮迭代无需遍历所有侯选属性,在一定程度上减少了时间消耗. 此外,使用知识距离衡量属性间的相似性程度,可以在同一组中聚集差异性更大的属性. 粒度与知识距离两种度量能够从不同视角审视评价候选属性. 流程如下.

算法2 融合粒度分组与Pareto最优的约简求解.

输入: DS, $\delta > 0, \epsilon \in [0,1]$;

输出: 约简red.

step 1: Group $= \varnothing$, red $= \varnothing$, $i = 1, j = 1$, Temp $=$ AT, $\gamma_{\text{Group}}(d) = -\infty$.

step 2: 计算AT中粒度值并升序排列.

step 3: while Temp $\neq \varnothing$ do

1) 将Temp中具有最小粒度值的属性b_k添加至Group$_i$中, 计算b_k与剩余属性知识距离及平均值mean$_D$, 将知识距离大于平均值的属性添加进该组;

2) Temp $=$ Temp $-$ Group$_i$, $i = i + 1$.

end

step 4: while $\gamma^{\delta}(\text{Group}, U)/\gamma^{\delta}(\text{AT}, U) < \epsilon$ do

1) Group $=$ Group$\bigcup$Group$_j$, 使用受限Pareto最优原则从Group中选出属性加入red.

if $\gamma^{\delta}(\text{red}, U)/\gamma^{\delta}(\text{AT}, U) \geqslant \epsilon$ then

2) 若$\gamma^{\delta}(\text{red}, U)/\gamma^{\delta}(\text{AT}, U) \geqslant \epsilon$在去除red中任一属性$a_k$后仍成立, 则去除该属性, 直到red不再改变或$|\text{red}| = 1$.

else

3) $j = j + 1$;

end

end

step 5: return red.

算法2的时间复杂度包含3个部分:

1) step 1和step 2, 计算粒度值并升序排列, 遍历次数为$|\text{AT}|(|\text{AT}| - 1)/2$;

2) step 3, 最坏情况下, 每个属性为一组, 遍历次数也为$|\text{AT}|(|\text{AT}| - 1)/2$;

3) step 4, 最坏情况下, 需要遍历所有组以添加属性, 遍历次数为$|\text{AT}|$.

综上, 算法2的时间复杂度为$\mathcal{O}(|\text{AT}|^2)$. 此外, 算法2只有在分组时开辟了临时存储空间, 故空间复杂度为$\mathcal{O}(|\text{Group}|)$, 即$\mathcal{O}(1)$.

4　实验分析

4.1　实验配置

本节选取了包含“Forest Type Mapping”“Statlog (Landsat satellite)”等在内的12组UCI数据集进行实验对比分析, 所有算法均采用Matlab2017b实现. 实验平台的操作系统为Windows 10, CPU为Intel® Core(TM) i9-10885H, 内存为16.00 GB.

本节实验均采用10折交叉验证测试算法的性能. 即将数据按照样本数量分为10等份, 每次取其中9份进行约简求解, 1份作为测试集, 以测试所求得约简的分类性能, 分类器采用CART与KNN. 在实验中, 对于每一个数据集, 通过设置百分比β为属性注入噪声, β的取值分别为0, 10%, 20%, 30%和40%. 例如, 当$\beta = 10\%$时, 随机选取10%的条件属性, 注入高斯白噪声, 扰动这些条件属性上的真实数据. 此外, 所有实验均基于邻域粗糙集模型实现, 其中邻域半径设置为0.02, 0.04, …, 0.40共20个.

将本文方法与以下8种流行的约简求解方法进行对比分析:

1) 基于知识变化率的约简求解[26];

2) 传统启发式约简求解[27];

3) 改进森林优化约简求解[28];

4) 基于自信息的约简求解[29];

5) 属性簇快速约简求解[30];

6) 集成选择器约简求解[31];

7) 基于新适应度的约简求解[32];

8) 鲁棒属性约简求解[33].

4.2　分类稳定性

在本小节的实验中, 根据不同算法求得的约简结果, 采用CART[34]与KNN[35]这两种分类器在测试样本上进行分类学习. 首先对比本文方法与4.1节中所示8种方法求得的约简, 它们在测试集上得到的分类结果的稳定性[31]及具体实验结果如图2和图3所示.

观察图2和图3, 不难得出以下结论:

1) 当数据集未注入属性噪声($\beta = 0$)时, 在大部分数据集上, 本文方法与对比方法所求得的分类稳定性差异性不大. 这说明在原始数据集上, 算法2中以粒度排序、知识距离分组的策略具备与目前流行的约简方法相似的分类能力.

2) 对于注入了属性噪声后的数据集, 利用本文方法得到的约简在面临分类任务时, 能够得到更高的分类稳定性. 例如, 对于数据集“QSAR biodegradation (ID-8)”, 本文方法在两个分类器上所求得的分类稳定性均高于所有对比方法求得的分类稳定性, 而当噪声比例β为40%时, 这一优势更加显著. 这说明本文所提出方法求得的约简, 在面临分类任务时, 具有较好的抗噪性能.

3) 随着噪声百分比的增大, 在9种约简方法所对应的分类稳定性上, 性能都有下降趋势, 但该趋势并不是严格单调的. 例如在数据集“Database on

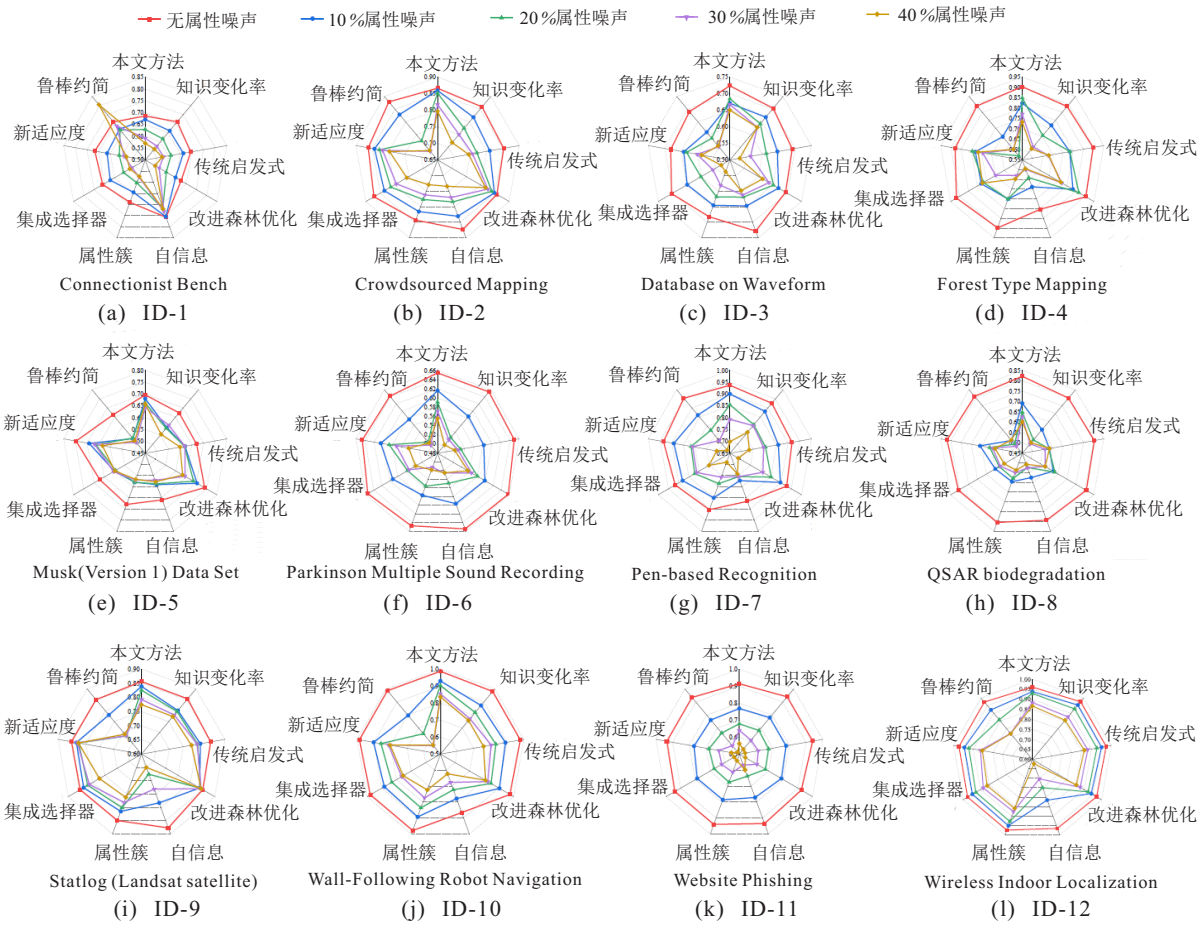


图 2 CART 分类稳定性

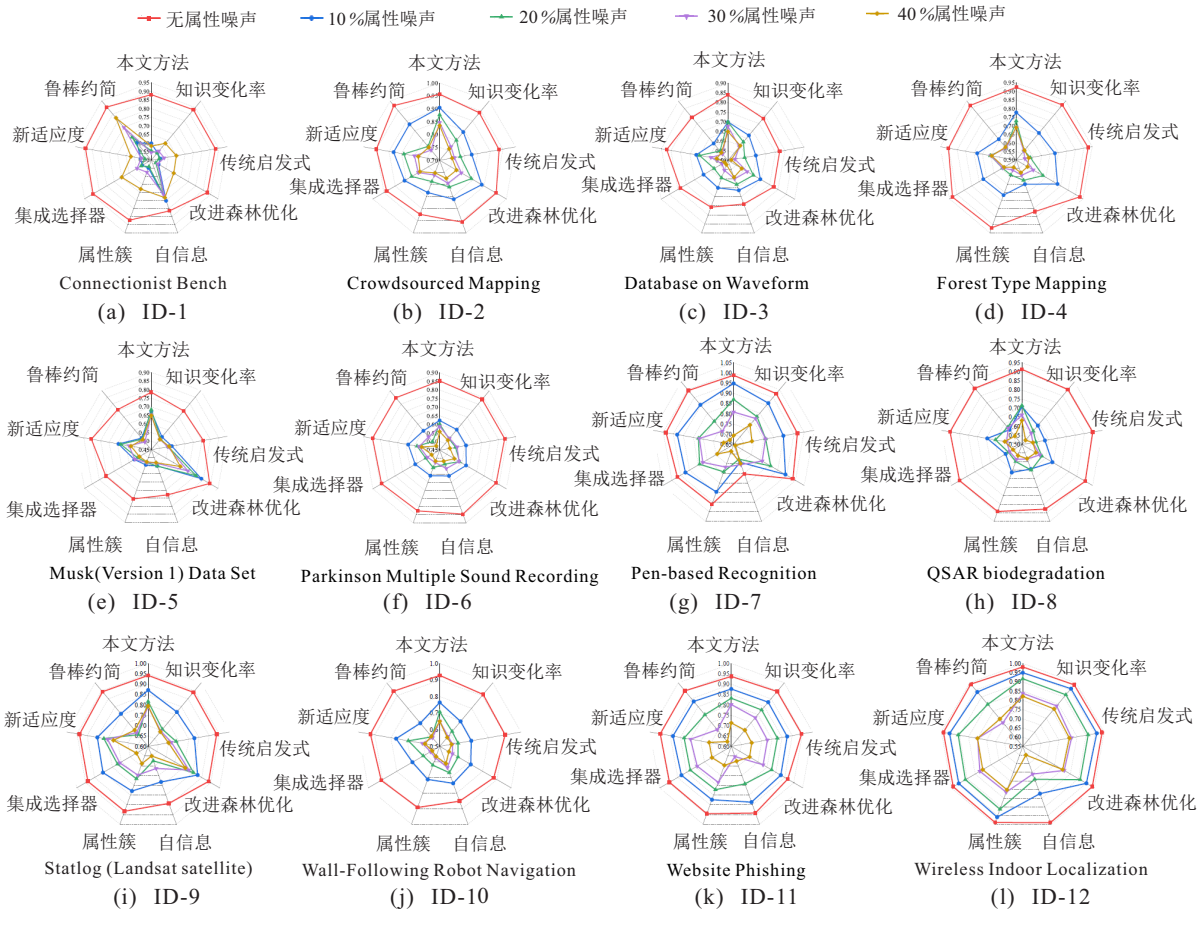


图 3 KNN 分类稳定性

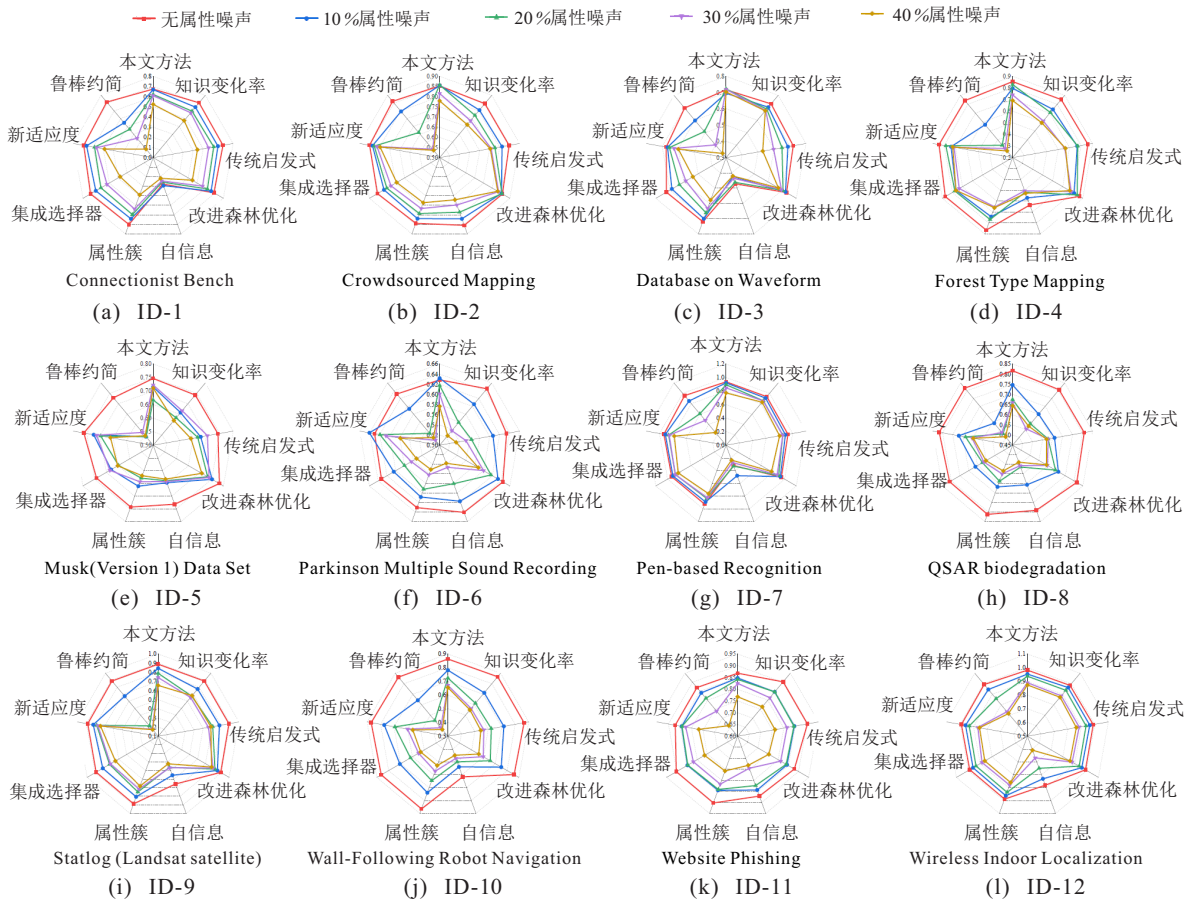


图 4　CART 分类准确率

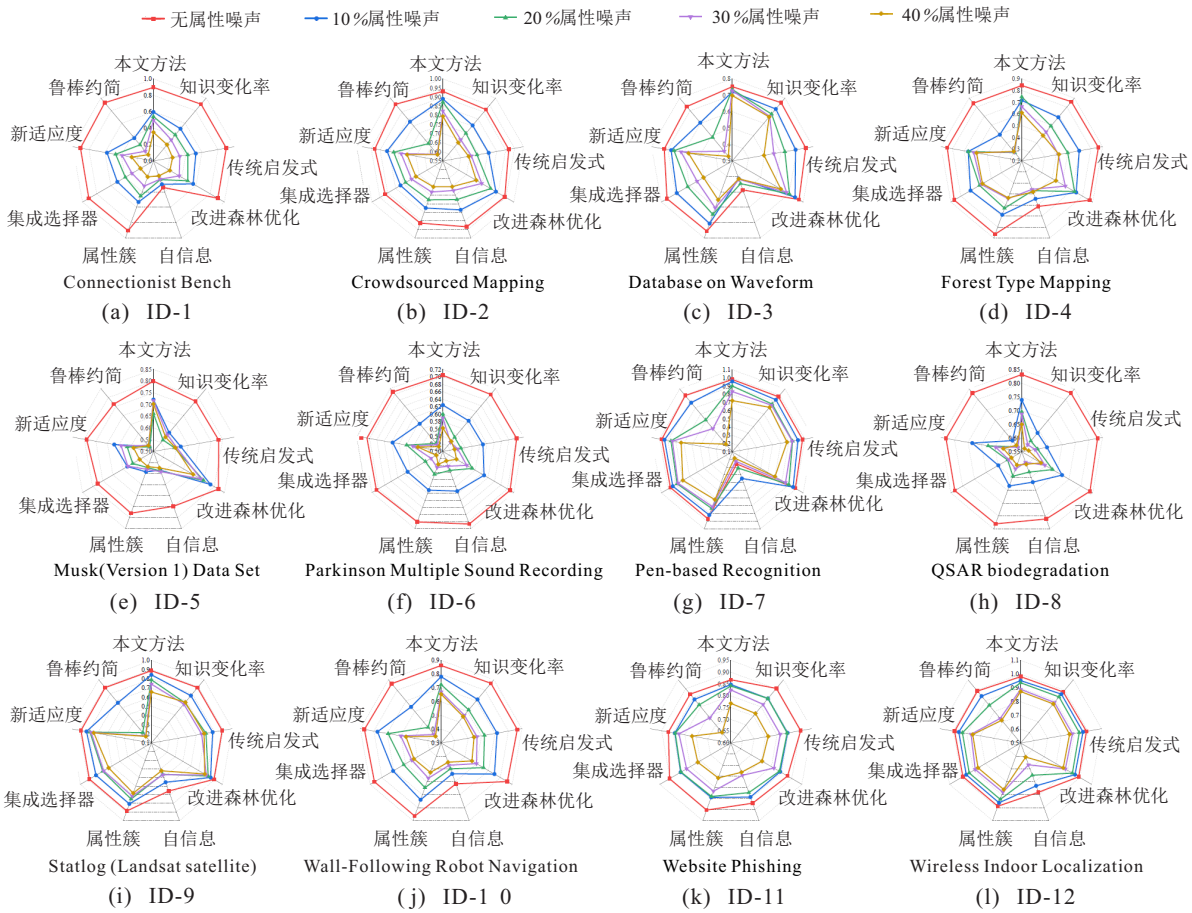


图 5　KNN 分类准确率

Waveform (ID-3)”上, 利用CART分类器, 本文方法在面临$\beta = 20\%$噪声比例时, 对应的分类稳定性高于面临$\beta = 10\%$噪声比例时所对应的分类稳定性, 在数据集“Statlog (Landsat satellite) (ID-9)”上, 利用KNN分类器, 鲁棒属性约简求解方法[33]在面临$\beta = 40\%$噪声比例时, 对应的分类稳定性高于面临$\beta = 30\%$噪声比例时所对应的分类稳定性.

综上所述, 本文方法相较于其他8种对比方法, 在CART分类器上, 分类稳定性平均提升了5.37%, 在KNN分类器上, 分类稳定性平均提升了6.41%. 因此, 总体的分类稳定性提升了5.89%.

4.3 分类准确率

本文方法与其他8种方法在测试集上得到的分类结果的准确率, 如图4和图5所示.

根据图4、图5的结果, 不难得出以下结论:

1) 当数据未注入属性噪声($\beta = 0$)时, 由与之对应的图4、图5中的红色线条可知, 本文方法较对比方法并非完全占优, 但在10个数据集上, 利用CART分类器, 本文方法与基于自信息的约简方法[29]相比, 能够得到更高的分类准确率. 对于一些注入了属性噪声的数据集, 以数据集“Connectionist Bench(ID-1)”为例, 利用本文方法得到的约简在面临分类任务时, 能够得到最高的分类准确率, 说明通过本文方法得到的约简, 在注入噪声的数据集上具备更好的分类性能.

2) 随着噪声百分比的增大, 在9种约简方法所对应的分类准确率上, 性能都有下降趋势. 与分类稳定性类似, 该趋势也不是严格单调的, 这说明噪声比例也会影响约简对应的分类准确率. 然而, 对于一些数据集, 例如“Database on Waveform(ID-3)”, 无论是使用CART还是KNN分类器, 属性噪声比例的改变没有对本文方法的约简结果所对应的分类准确率产生重大影响. 例如, 利用CART分类器, 本文方法在面临$\beta = 10\%, 20\%, 30\%, 40\%$噪声比例时的分类准确率分别为69.55%, 72.22%, 71.91%和70.36%. 鉴于此, 从分类准确率这项指标来看, 本文方法也具有较好的抗噪性能.

综上而言, 本文方法相较于其他8种对比方法, 在CART分类器上, 分类准确率平均提升了11.17%; 在KNN分类器上, 分类准确率平均提升了13.39%. 因此, 总体的分类准确率提升了12.28%.

4.4 时间消耗

在实验比对时, 为确保公平性, 9种算法均遵循以下步骤: 使用10折交叉验证方法在训练集中搜索约简, 记录相应的时间消耗, 并计算其平均值. 表1是本文方法与8种对比方法的平均时间消耗, 耗时最短的用斜粗体表示.

表1 不同方法求解约简的平均时间消耗 单位: s

β	本文方法	文献[26]方法	文献[27]方法	文献[28]方法	文献[29]方法	文献[30]方法	文献[31]方法	文献[32]方法	文献[33]方法
$\beta = 0$	***10.5719***	102.4571	25.8298	433.1520	31.5561	10.5905	14.6648	19.3302	105.1387
$\beta = 10\%$	***10.6330***	100.4277	26.8901	164.5372	32.9597	10.6508	14.2690	20.3774	92.4739
$\beta = 20\%$	***10.7377***	93.8406	28.8925	77.7917	33.9303	10.7892	14.5272	21.7056	69.6513
$\beta = 30\%$	8.7223	70.5664	29.6423	66.3752	26.9757	***8.3134***	10.7827	19.6230	55.4457
$\beta = 40\%$	***7.7014***	70.6504	26.0687	53.8072	26.7115	7.7071	10.3404	19.4632	50.4879

由表1结果可以得出以下结论: 当噪声比逐步提高时, 在12个数据集上的结果表明, 平均时间消耗呈现整体下降的趋势. 此外, 在面临$\beta = 0, 10\%, 20\%, 40\%$噪声比例时, 本文方法所需平均时间消耗的均值在所有方法中都最低; 在面临30%噪声比例时, 虽然本文方法时间消耗的均值不是最低, 但其值(8.7223 s)与最小值(8.3134 s[30])之间的差异性不大, 这说明无论噪声注入与否, 本文方法都能够提高求解约简的效率.

综上而言, 本文方法相较于其他8种对比方法, 总体的时间消耗平均降低了59.27%.

5 结 论

目前已知的启发式属性约简求解算法往往消耗较长, 其使用固定且单一的度量进行属性评价, 这促使本文考虑在提升约简求解效率的同时对属性进行多元评价的可能性. 首先, 利用邻域知识粒度对属性排序; 然后, 利用邻域知识距离对排序后的属性分组, 不仅为后续选择受限Pareto最优属性压缩了筛选间, 提升了求解约简的时间效率, 而且达到了属性多元评价的目的; 最后, 根据受限Pareto最优原则, 依照分组

顺序选择出符合约束条件的属性子集. 实验分析表明,本文算法相较于其他算法而言,时间消耗得以降低,且得到的约简比其他算法得到的约简具有更好的分类表现. 在此基础上,将进一步探讨以下问题: 1)将所提出方法进一步引入多标签及标签分布数据集中的降维问题上; 2)在面临标签噪声等弱监督数据时,如何进一步完善所提出方法.

参考文献(References)

[1] Liu K Y, Li T R, Yang X B, et al. Granular cabin: An efficient solution to neighborhood learning in big data[J]. Information Sciences, 2022, 583: 189-201.

[2] 贾鹤鸣, 姜子超, 李瑶. 基于改进秃鹰搜索算法的同步优化特征选择[J]. 控制与决策, 2022, 37(2): 445-454.
(Jia H M, Jiang Z C, Li Y. Simultaneous feature selection optimization based on improved bald eagle search algorithm[J]. Control and Decision, 2022, 37(2): 445-454.)

[3] 冯锋, 万喆, 徐泽水, 等. 基于软粗糙集的犹豫模糊三支决策方法[J]. 控制与决策, 2023, 38(3): 834-842.
(Feng F, Wan Z, Xu Z S, et al. Hesitant fuzzy three-way decision method based on soft rough sets[J]. Control and Decision, 2023, 38(3): 834-842.)

[4] 李雪岩, 李学伟, 蒋君. 基于知识粒度特征的多目标粗糙集属性约简算法[J]. 控制与决策, 2021, 36(1): 196-205.
(Li X Y, Li X W, Jiang J. Multi objective rough set attribute reduction algorithm based on characteristics of knowledge granularity[J]. Control and Decision, 2021, 36(1): 196-205.)

[5] Chen Y, Yang X B, Li J H, et al. Fusing attribute reduction accelerators[J]. Information Sciences, 2022, 587: 354-370.

[6] Fan X D, Chen X Y, Wang C Z, et al. Margin attribute reductions for multi-label classification[J]. Applied Intelligence, 2022, 52(6): 6079-6092.

[7] Liu Y X, Gong Z C, Liu K Y, et al. A Q-learning approach to attribute reduction[J]. Applied Intelligence, 2023, 53(4): 3750-3765.

[8] Zhang X, Mei C L, Chen D G, et al. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition, 2016, 56: 1-15.

[9] Xu J C, Yang J, Ma Y Y, et al. Feature selection method for color image steganalysis based on fuzzy neighborhood conditional entropy[J]. Applied Intelligence, 2022, 52(8): 9388-9405.

[10] Liu K Y, Yang X B, Fujita H, et al. An efficient selector for multi-granularity attribute reduction[J]. Information Sciences, 2019, 505: 457-472.

[11] Chen D G, Zhao S Y. Local reduction of decision system with fuzzy rough sets[J]. Fuzzy Sets and Systems, 2010, 161(13): 1871-1883.

[12] Zhang X Y, Yao H, Lv Z Y, et al. Class-specific information measures and attribute reducts for hierarchy and systematicness[J]. Information Sciences, 2021, 563: 196-225.

[13] 李军. 基于贪心核特征提取方法的中期峰值负荷预测[J]. 控制与决策, 2014, 29(9): 1661-1666.
(Li J. Greedy kernel feature extraction method for medium term electricity peak load forecasting[J]. Control and Decision, 2014, 29(9): 1661-1666.)

[14] An S, Liu J Y, Wang C Z, et al. A relative uncertainty measure for fuzzy rough feature selection[J]. International Journal of Approximate Reasoning, 2021, 139: 130-142.

[15] Huang Y Y, Guo K J, Li Z, et al. Matrix representation of the conditional entropy for incremental feature selection on multi-source data[J]. Information Sciences, 2022, 591: 263-286.

[16] Rao X S, Yang X B, Yang X, et al. Quickly calculating reduct: An attribute relationship based approach[J]. Knowledge-Based Systems, 2020, 200: 106014.

[17] Shah N, Ali M I, Shabir M, et al. Uncertainty measure of Z-soft covering rough models based on a knowledge granulation[J]. Journal of Intelligent & Fuzzy Systems, 2020, 38(2): 1637-1647.

[18] Wei W, Wang D, Liang J Y. Accelerating relief using information granulation[J]. International Journal of Machine Learning and Cybernetics, 2022, 13(1): 29-38.

[19] Guo H Y, Wang L D, Liu X D, et al. Information granulation-based fuzzy clustering of time series[J]. IEEE Transactions on Cybernetics, 2021, 51(12): 6253-6261.

[20] 李金海, 贺建君. 多粒度形式背景的不确定性度量与最优粒度选择[J]. 控制与决策, 2022, 37(5): 1299-1308.
(Li J H, He J J. Uncertainty measurement and optimal granularity selection for multi-granularity formal context[J]. Control and Decision, 2022, 37(5): 1299-1308.)

[21] 张思源, 王国胤, 刘群, 等. 基于多粒度特征融合的边缘一致性图像补全[J]. 控制与决策, 2022, 37(12): 3240-3250.
(Zhang S Y, Wang G Y, Liu Q, et al. Edge consistent image completion based on multi-granularity feature fusion[J]. Control and Decision, 2022, 37(12): 3240-3250.)

[22] Qian Y H, Cheng H H, Wang J T, et al. Grouping

granular structures in human granulation intelligence[J]. Information Sciences, 2017, 382/383: 150-169.

[23] Qian Y H, Liang J Y, Dang C Y. Knowledge structure, knowledge granulation and knowledge distance in a knowledge base[J]. International Journal of Approximate Reasoning, 2009, 50(1): 174-188.

[24] Atta S, Mahapatra P R S, Mukhopadhyay A. A multi-objective formulation of maximal covering location problem with customers' preferences: Exploring Pareto optimality-based solutions[J]. Expert Systems with Applications, 2021, 186: 115830.

[25] Li G H, Li Y, Zheng Y F, et al. A novel feature selection approach with Pareto optimality for multi-label data[J]. Applied Intelligence, 2021, 51(11): 7794-7811.

[26] Jin C X, Li F C, Hu Q H. Knowledge change rate-based attribute importance measure and its performance analysis[J]. Knowledge-Based Systems, 2017, 119: 59-67.

[27] Hu Q H, Pedrycz W, Yu D R, et al. Selecting discrete and continuous features based on neighborhood decision error minimization[J]. IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics, 2010, 40(1): 137-150.

[28] 刘兆赓, 李占山, 王丽, 等. 森林优化特征选择算法的增强与扩展[J]. 软件学报, 2020, 31(5): 1511-1524.
(Liu Z G, Li Z S, Wang L, et al. Enhancement and extension of feature selection using forest optimization algorithm[J]. Journal of Software, 2020, 31(5): 1511-1524.)

[29] Wang C Z, Huang Y, Shao M W, et al. Feature selection based on neighborhood self-information[J]. IEEE Transactions on Cybernetics, 2020, 50(9): 4031-4042.

[30] Chen Y, Liu K Y, Song J J, et al. Attribute group for attribute reduction[J]. Information Sciences, 2020, 535: 64-80.

[31] Yang X B, Yao Y Y. Ensemble selector for attribute reduction[J]. Applied Soft Computing, 2018, 70: 1-11.

[32] Ye D Y, Chen Z J, Ma S L. A novel and better fitness evaluation for rough set based minimum attribute reduction problem[J]. Information Sciences, 2013, 222: 413-423.

[33] Dong L J, Chen D G, Wang N L, et al. Key energy-consumption feature selection of thermal power systems based on robust attribute reduction with rough sets[J]. Information Sciences, 2020, 532: 61-71.

[34] Gey S. Risk bounds for CART classifiers under a margin condition[J]. Pattern Recognition, 2012, 45(9): 3523-3534.

[35] Wang Y K, Pan Z B, Dong J. A new two-layer nearest neighbor selection method for kNN classifier[J]. Knowledge-Based Systems, 2022, 235: 107604.

作者简介

印振宇 (1997–), 男, 硕士生, 从事粗糙集、粒计算等研究, E-mail: Y_Tenssy@163.com;

王平心 (1980–), 男, 副教授, 硕士生导师, 从事三支决策、粒计算等研究, E-mail: wangpingxin@just.edu.cn;

杨习贝 (1980–), 男, 教授, 博士生导师, 从事粗糙集、粒计算等研究, E-mail: jsjxy_yxb@just.edu.cn;

于化龙 (1982–), 男, 教授, 博士生导师, 从事机器学习、数据挖掘等研究, E-mail: yuhualong@just.edu.cn;

钱宇华 (1976–), 男, 教授, 博士生导师, 从事人工智能、数据挖掘与机器学习等研究, E-mail: jinchengqyh@126.com.