计算机科学
COMPUTER SCIENCE

一种基于正域的三支近似约简

王志成，高灿，邢金明

引用本文

王志成，高灿，邢金明. 一种基于正域的三支近似约简[J]. 计算机科学, 2022, 49(4): 168-173.

WANG Zhi-cheng, GAO Can, XING Jin-ming. Three-way Approximate Reduction Based on Positive Region[J]. Computer Science, 2022, 49(4): 168-173.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210500067

一种基于正域的三支近似约简

王志成　高　灿　邢金明
深圳大学计算机与软件学院　广东 深圳 518060
深圳大学智能信息处理重点实验室　广东 深圳 518060
(wzc2802005420@163.com)

摘　要　属性约简是三支决策理论的重要研究内容之一。然而,现有基于三支决策的属性约简方法过于严格,限制了其属性约简的效率。文中提出了一种基于正域的三支近似属性约简方法。具体地,属性约简被视为根据条件属性与决策属性的相关性,将所有属性划分为正域、负域或边界域3类的过程。首先通过保留正域度量来去除负域属性,然后通过放松正域度量来迭代地排除一些边界属性,最后将剩余属性构成一个近似约简。UCI数据实验结果显示,与其他代表性的方法相比,所提方法能在保持甚至提升性能的同时获得更小的属性约简,说明了所提方法的有效性。

关键词:粗糙集;三支决策;属性约简;正域;近似约简
中图法分类号　TP391

Three-way Approximate Reduction Based on Positive Region

WANG Zhi-cheng,GAO Can and XING Jin-ming
College of Computer Science and Software Engineering,Shenzhen University,Shenzhen,Guangdong 518060,China
Key Laboratory of Intelligent Information Processing,Shenzhen,Guangdong 518060,China

Abstract　Attribute reduction is one of the most important research topics in the theory of three-way decision. However,the existing attribute reduction methods based on three-way decision are too strict,which limit the efficiency of attribute reduction. In this paper,a three-way approximate attribute reduction method based on the positive region is proposed. More specifically,attribute reduction is considered as the process of determining attributes as positive,boundary,or negative ones according to their correlation to the decision attribute. The negative attributes are first removed by retaining the measure of the positive region. Then,some of the boundary attributes are iteratively excluded by relaxing the positive region measure. Finally,an approximate reduction is formed by the remaining attributes. Extensive experiments on UCI data sets demonstrate that the proposed method can achieve much smaller reducts with the same or even better performance in comparison with other representative methods,showing the effectiveness in attribute reduction.

Keywords　Rough set,Three-way decision,Attributes reduction,Positive region,Approximate reduction

1　引言

现实应用领域(如图像分类、文本或基因分析)[1]中的样本包含成千上万个属性,而过多的属性会导致学习器学习过慢和过拟合问题,并进一步导致泛化能力变差[2]。属性约简[3-9](也称为特征选择)作为一种去除冗余和不相关属性的有效方法,已成为机器学习、模式识别和数据挖掘中的重要预处理步骤。

粗糙集理论是Pawlak[10]提出的一种处理含糊、不精确或不确定数据的有效方法,而属性约简是粗糙集理论的重要研究内容之一。在Pawlak经典粗糙集模型中,下近似是由等价关系下所有完全包含于概念中的等价类构成的,但对于实际应用问题,等价关系过于严格,尤其针对含有噪声的数据。三支决策[11-12](TWD)是由Yao提出的一种决策支持和近似推理的有效方法。作为Pawlak粗糙集模型的概率扩展,三支决策模拟了人类在不确定性和风险下的决策过程,不仅考虑了决策的置信度,还考虑了决策行为导致的成本,原有“是”与“否”的二元决策演变为“接受”“拒绝”和“延迟决策”三元决策。三支决策为解决复杂问题提供了一种有效的方法[13],并在理论、模型和应用方面都取得了较大进展,如属性约简[14]、

到稿日期:2021-05-10　返修日期:2021-10-15
基金项目:国家自然科学基金(61806127,62076164);佛山市教育局项目(2019XJZZ05)
This work was supported by the National Natural Science Foundation of China(61806127,62076164) and Bureau of Education of Foshan(2019XJZZ05).
通信作者:高灿(2005gaocan@163.com)

冲突分析[15]和认知计算[16]等。

基于三支决策的属性约简方法主要分为度量保持和度量优化两类[17]。度量保持方法要求得到的约简度量指标必须完全保持甚至有所提升。Li 等[18]提出一种基于正域样本数目度量的属性约简方法。Ma 等[19-20]定义了基于概率分布的不确定性度量,并提出了一种保持条件信息或信息熵的启发式算法。Gao 等[21]提出粒度化最大决策熵的不确定性度量及相关的启发式约简算法,度量优化方法将属性约简视为不确定性度量的优化问题。Yao 等[14]研究了正域、非负域、规则置信度、规则覆盖度和规则代价等不确定性度量,并给出了最优约简的一般定义。Zhao 等[17]介绍了最优约简的正决策、正域扩张和基于非负区域的不确定性度量,并提出一种基于差别矩阵的最优约简构造方法。Zhang 等[22-23]讨论了 4 种属性约简不确定性度量,并由此提出了 1 种通用约简和 3 种最优约简。Jia 等[24-25]从决策代价最小化出发,提出一种利用遗传、模拟退火或粒子群技术的最小成本属性约简算法。Liao 等[26]研究了具有最小决策代价和测试代价的属性约简问题。

度量保持和度量优化属性约简方法要求过于严格,如实际待处理的数据包含噪声样本时,约简效率可能较低。Slezak[27]在决策表中提出一种类信息熵测度,将概率与通用方法结合并引入局部信息,要求只要特征子集的类熵测度接近原数据的熵测度即可。Slezak[28]利用信息熵测度来解释粗糙集属性约简的基本原理,并提出近似熵约简原理(AERP),即任何保持条件熵近似的属性约简都应该减少其先验熵。基于条件熵的属性约简对噪声样本非常敏感,得到的属性子集在某些情况下会包含冗余属性。Yang 等[29]提出一种基于决策表中条件信息熵的近似约简算法,可以通过适当地去除冗余属性来提高对噪声的鲁棒性。

受上述工作的启发,本文提出一种新的基于正域的三支近似属性约简方法。该方法能在保持甚至提高分类性能的情况下获得属性数目更少的约简,UCI 数据集实验显示了其有效性。

本文第 2 节介绍了粗糙集和三支决策的基本概念;第 3 节介绍了三支决策属性约简,并提出一种三支决策近似约简的定义和相应的启发式算法;第 4 节给出实验结果和分析;最后总结全文并展望未来。

2 相关知识

本节主要介绍粗糙集理论与三支决策的相关概念,更多细节请参见文献[13-14,30-32]。

定义 1[10] 决策信息系统(或决策表)可表示为 $S=(U,A=C\cup D,V,f)$。其中,U 是样本集合;A 为非空属性集合,分为条件属性集合 C 和决策属性集合 D;V 是属性的值域;f 为信息函数。

定义 2[10] 给定决策表 $S=(U,A=C\cup D,V,f)$,可定义属性子集 $B\subseteq A$ 的不可辨别关系 $IND(B)=\{(x,y)\in U\times U|\forall a\in B,f(x,a)=f(y,a)\}$。

$IND(B)$对 U 形成一个划分,简记为 U/B。包含样本 x 的基本划分块称为基本集或基本粒,表示为$[x]_B$。

定义 3[10] 给定决策表 $S=(U,A=C\cup D,V,f)$,集合 $X\subseteq U$ 关于属性子集 B 的下、上近似定义分别为:

$$\underline{B}(X)=\bigcup\{x\in U:[x]_B\subseteq X\}$$
$$\overline{B}(X)=\bigcup\{x\in U:[x]_B\cap X\neq\varnothing\} \tag{1}$$

定义 4[10] 给定决策表 $S=(U,A=C\cup D,V,f)$,条件属性集 C 相对于决策 D 的正域、边界域和负域分别定义为:

$$POS_C(D)=\bigcup_{Y_i\in U/D}\underline{C}(Y_i)$$
$$BND_C(D)=\bigcup_{Y_i\in U/D}(\overline{C}(Y_i)-\underline{C}(Y_i)) \tag{2}$$
$$NEG_C(D)=U-\bigcup_{Y_i\in U/D}\overline{C}(Y_i)$$

定义 5[10] 给定决策表 $S=(U,A=C\cup D,V,f)$,P 为 C 的一个属性子集,P 相对于决策属性 D 的依赖度定义为:

$$\gamma_P(D)=\frac{|POS_P(D)|}{|D|} \tag{3}$$

令 $\Omega=\{X,X^C\}$ 表示样本 x 是否属于 X 集合,$\Lambda=\{a_P,a_B,a_N\}$为决定样本 x 属于 $POS(X)$,$BND(X)$或 $NEG(X)$的决策行动集。则对于样本 x 采取不同决策行动的代价函数如表 1 所列[31]。

表 1 不同状态下样本采取不同决策的代价函数

Table 1 Cost function in different decisions with different states

	a_P	a_B	a_N
X	λ_{PP}	λ_{BP}	λ_{NP}
X^C	λ_{PN}	λ_{BN}	λ_{NN}

表 1 中,λ_{PP},λ_{BP},λ_{NP}表示将属于 X 的样本划分为正域、边界域或负域的代价。而λ_{PN},λ_{BN},λ_{NN}表示将不属于 X 的样本划分为正域、边界域或负域的代价。

给定样本 x,采取不同决策的期望代价可以表示为[31]:

$$\begin{aligned}R(a_P|[x])&=\lambda_{PP}P(X|[x])+\lambda_{PN}P(X^C|[x])\\R(a_B|[x])&=\lambda_{BP}P(X|[x])+\lambda_{BN}P(X^C|[x])\\R(a_N|[x])&=\lambda_{NP}P(X|[x])+\lambda_{NN}P(X^C|[x])\end{aligned} \tag{4}$$

其中,$P(X|[x])$和 $P(X^C|[x])$分别表示样本 x 属于 X 和X^C的概率,且 $P(X|[x])=1-P(X^C|[x])$。

根据贝叶斯决策理论,最小风险决策如下[31]:

(P)若 $R(a_P|[x])\leqslant\min\{R(a_B|[x]),R(a_N|[x])\}$,则 $x\in POS(X)$;

(B)若 $R(a_B|[x])\leqslant\min\{R(a_P|[x]),R(a_N|[x])\}$,则 $x\in BND(X)$;

(N)若 $R(a_N|[x])\leqslant\min\{R(a_P|[x]),R(a_B|[x])\}$,则 $x\in NEG(X)$。

当代价函数不等式$(\lambda_{PN}-\lambda_{BN})(\lambda_{NP}-\lambda_{BP})>(\lambda_{BP}-\lambda_{PP})(\lambda_{BN}-\lambda_{NN})$成立时,决策规则可以进一步简化为[31]:

(P)若 $P(X|[x])\geqslant\alpha$,则 $x\in POS(X)$;

(B)若 $\beta<P(X|[x])<\alpha$,则 $x\in BND(X)$;

(N)若 $P(X|[x])\leqslant\beta$,则 $x\in NEG(X)$。

其中:

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

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

定义 6[31] 给定决策表 $S=(U,A=C\cup D,V,f)$及代价函数,集合 $X\subseteq U$ 关于属性子集 B 的下、上近似定义分别为:

$$\underline{B}_{(\alpha,\beta)}(X)=\{x\in U|\mu_B(x)\geqslant\alpha\}$$
$$\overline{B}_{(\alpha,\beta)}(X)=\{x\in U|\mu_B(x)>\beta\} \tag{5}$$

定义 7[31] 给定决策表 $S=(U,A=C\cup D,V,f)$及代价函数,条件属性集 C 相对于决策 D 的正域、边界域和负域定义为:

$$POS_C^{(\alpha,\beta)}(D)=\{x\in U|\mu_{[x]_C}(D)\geqslant\alpha\}$$
$$BND_C^{(\alpha,\beta)}(D)=\{x\in U|\beta<\mu_{[x]_C}(D)<\alpha\} \tag{6}$$
$$NEG_C^{(\alpha,\beta)}(D)=\{x\in U|\mu_{[x]_C}(D)\leqslant\beta\}$$

其中,$\mu_{[x]_C}(D)=P(D_{\max}([x]_C)|[x]_C)$,$D_{\max}([x]_C)=\arg\max_{D_i\in U/D}\{P(D_i|[x]_C)\}$。

3 三支近似约简

本节主要介绍近似约简的概念,并提出了一种基于正域的近似约简定义以及由三支决策引导的启发式属性约简算法。

3.1 度量保持约简

定义 8[17] 给定决策表 $S=(U,A=C\cup D,V,f)$,属性子集 P 是 C 的一个度量保持约简,当且仅当以下条件成立:

(1)$MES(P,D)\geqslant MES(C,D)$;

(2)对于任意属性 $a\in P$,$MES(P-\{a\},D)<MES(C,D)$。

其中,MES 是属性约简的度量标准,符号"≥"表示"等于或优于"。度量标准可以有多种选择,如依赖度、互信息和最小代价等。本文采用依赖度作为度量。

3.2 度量近似约简

在基于正域的属性约简算法中,生成最多正域样本的属性会被优先选择。因此,在属性约简过程中,前面选择的属性较为重要,而后面的属性的重要性则较低。在属性选择的开始阶段,正域样本数较少,而随着属性数量的增加,正域样本数逐渐增加,直至与条件属性集 C 下的正域样本数相同。当属性约简接近结束时,绝大多数样本都在正域中,只有少量样本依旧是不确定的。此时,基于度量保持原则的算法会继续添加属性,只为了使这些少量的样本也处于正域中。

然而,度量保持方法忽视了样本数据在采集过程中的随机性,即样本本身可能含有噪声。在属性约简已经选择大量属性后,可能仍存在少量非正域样本,而这些样本可能是噪声样本。为了拟合这些噪声样本,度量保持的方法会增加在本质上无关的属性。为了避免出现上述情况,本文引入了近似约简的概念。

定义 9 给定决策表 $S=(U,A=C\cup D,V,f)$及参数 ε,$\eta(\varepsilon\in[0,1],\eta\in[0,1])$,属性子集 P 是 C 的一个(ε,η)-近似约简,当且仅当以下条件成立:

(1)$MES(P,D)\geqslant\varepsilon MES(C,D)$;

(2)对于任意属性 $a\in P$,$MES(P,D)-MES(P-\{a\},D)\geqslant\eta MES(C,D)$。

其中,第一项表示 P 近似于 C,即近似约简的描述能力接近原条件属性集;第二项表示从 P 中删除任意属性后,度量准则都会大幅下降,保证了约简中个体属性的重要性。在实际应用中,可通过需求经验设置参数。一般地,ε 越大,η 越小,近似约简就越接近度量保持约简。

在近似约简思想下获得的约简只需要在一定程度上接近条件属性集 C 即可。但近似的标准需要人为指定,如何确定合适的近似标准是一个尚待解决的问题。

基于三支决策和近似约简的思想考虑将所有属性分为 3 类,即正域、边界域和负域属性。负域是度量保持约简算法未选择的属性,正域是算法首先选择的部分属性,边界域是算法最后选择的属性。本文通过参数 ε 和 η 来指导近似约简,通过删除部分边界域的属性,来达到在保留大多数重要属性的同时删除部分冗余和不相关属性的目的。

定义 10 给定决策表 $S=(U,A=C\cup D,V,f)$及参数 ε,$\eta(\varepsilon\in[0,1],\eta\in[0,1])$,属性子集 P 是 C 的一个(ε,η)-近似约简,当且仅当以下条件成立:

(1)$\gamma_P(D)\geqslant\varepsilon\gamma_C(D)$;

(2)对于任意属性 $a\in P$,$\gamma_P(D)-\gamma_{P-\{a\}}(D)\geqslant\eta\gamma_C(D)$。

其中,条件 1 决定了近似约简 P 的最小依赖度,条件 2 决定了 P 是满足参数 ε 和 η 约束下的最小近似约简。

3.3 基于三支决策的近似约简算法

由于最优属性约简是 NP 难题,因此一般采用启发式搜索策略。基于三支决策的启发式近似约简算法的描述如算法 1 所示。

算法 1 基于正域的三支近似约简启发式算法

输入:决策表 S=(U,A=C∪D,V,f)和参数 ε,η

输出:近似约简 P

1. P←{};
2. 若$\gamma_P(D)\geqslant\gamma_C(D)$,则转步骤 5;
3. 从 C−P 中选择最大依赖度提升的属性a_{opt};
4. 若 P←P∪{a_{opt}},则转步骤 2;
5. 从 P 中依次选择最后加入的属性 a;
6. $\gamma_{P-\{a\}}(D)<\varepsilon\gamma_C(D)$或 $\gamma_P(D)-\gamma_{P-\{a\}}(D)\geqslant\eta\gamma_C(D)$,则转步骤 8;
7. 若 P←P−{a},则转步骤 5;
8. 输出近似约简 P,算法结束。

算法 1 先贪婪地选择依赖度提升最大的属性,直至达到与条件属性集 C 相同的依赖度,再在度量保持约简中依次删除重要度较小的部分属性,并确保依赖度的下降满足约束条件。通过进一步分析发现,在算法 1 的步骤 2 中,随着属性的添加,整体的依赖度逐渐接近并超过$\varepsilon\gamma_C(D)$,而每次增加的依赖度则逐渐减少并低于$\eta\gamma_C(D)$。因此,可将算法进行改进,省去先添加再删除部分属性的步骤,具体如算法 2 所示。

算法 2 算法 1 的加速版本

输入:决策表 S=(U,A=C∪D,V,f)和参数 ε,η

输出:条件属性集 C 的近似约简 P

1. P←{};
2. 从 C−P 中选择最大依赖度提升的属性a_{opt};
3. 若$\gamma_P(D)\geqslant\varepsilon\gamma_C(D)$并且$\gamma_{P\cup\{a\}}(D)-\gamma_P(D)<\eta\gamma_C(D)$,则转步骤 5;
4. 若 P←P∪{a_{opt}},则转步骤 2;
5. 输出近似约简 P,算法结束。

算法 2 省略了算法 1 中多余的添加再删除属性的步骤，单向地添加属性直到整体的依赖度足够大，并且新增属性带来的依赖度提升非常有限时算法即停止。

算法 1 和算法 2 的时间主要花费在计算最优依赖度提升的属性上。假设有$|U|$个样本和$|C|$个属性，在每次迭代过程中，计算最优属性的时间代价为 $O(|C||U|)$。最坏情况下，算法需要运行$|C|$次，即最坏时间复杂度为 $O(|C|^2|U|)$，空间复杂度为 $O(|C||U|)$。

4 实验分析

实验的目的有两个：1)验证新方法属性约简效果；2)与其他方法对比属性约简后的质量，即分类性能。实验均在 Windows 10 操作系统、Intel(R) Core(TM) i5-9500 CPU @ 3.00 GHz 处理器、8 GB 内存的计算机上进行，所有代码用 Python 3.8 实现。

4.1 数据集和实验设计

实验选用了 14 个 UCI 数据集[33]，详细信息如表 2 所列。

表 2 所选 UCI 数据集

Table 2 Selected UCI data sets

Data Sets	$\|C\|$	$\|U\|$	$\|d\|$
credit-g(credit)	20(7)	1 000	2
lymph(lymph)	18(3)	148	4
molecular-biology-promotersnew(molecular)	57(0)	106	2
polish-companies-bankruptcy-2year(polish)	64(64)	10 173	2
quality-assessment-schiller(quality)	62(62)	92	2
seismic-bumps(seismic)	18(14)	2 584	2
solar-flare-1(solar)	12(0)	323	6
spectf-test(spectf)	44(44)	269	2
splice(splice)	60(60)	3 190	3
sponge(sponge)	44(0)	76	3
turkiye-student-evaluation-generic(turkiye)	31(31)	5 820	3
waveform-5000(wave)	40(40)	5 000	3
wine(wine)	13(13)	178	3
zoo(zoo)	16(1)	101	7

表 2 中，第一列括号内的内容表示在后文中使用的数据集缩写，第二列表示离散属性数和连续属性数(括号中)。数据集中所有缺失值都使用相应属性的频率最大值(或平均值)来填充。针对连续属性，使用三等频原则离散化为离散属性。

4.2 属性约简效率分析

当使用不同参数时，将放宽或加强近似约简的约束条件，因此产生不同的约简。若减少 ε，则获得的约简将具有更低的依赖度和更小的规模。若减小 η，则会保留依赖度提升较弱的属性并增加属性子集的规模，反之亦然。需要注意的是，η 应当满足 $\eta \leqslant 1-\varepsilon$，这是因为不存在一个属性使得属性子集在添加后依赖度大于 1。根据近似约简的原理，算法的参数不应当设置得太大或太小。如果 ε 值过大，算法会趋于度量保持约简，如果 ε 值过小，生成的约简可能会因规模过小而缺失关键信息。若 η 过大，则那些较为重要的属性可能会被误删，若 η 过小，则任何无关的属性也会被认为是不可删除的。本文将两个参数分别设置为 $\varepsilon=0.95$ 和 $\eta=0.05$，实验结果表明在此参数下可以获得较为理想的效果。

在参数 $\varepsilon=0.95$、$\eta=0.05$ 的情况下，测试了所提出的属性约简算法在 14 个数据集上的效果，并对比了基于正域的前向增加算法(Positive Region-based Method with Forward Adding Strategy，PRFA)和基于正域的后向删除算法(Positive Region-based Method with Backward Adding Strategy，PRBD)。另外，为了体现约简效率，原始数据属性信息(RAW)亦被列出，实验结果如表 3 所列。

表 3 所选方法的属性约简结果($\varepsilon=0.95$，$\eta=0.05$)

Table 3 Atttribute reducts obtained by selected methods ($\varepsilon=0.95$，$\eta=0.05$)

Data sets	RAW	PRFA	PRBD	Proposed
credit	20	10	13	8
lymph	18	7	9	6
molecular	57	4	5	3
polish	64	26	25	12
quality	62	6	7	5
seismic	18	14	13	11
solar	12	10	10	9
spectf	44	7	9	6
splice	60	10	11	8
sponge	44	3	4	2
turkiye	31	31	31	19
wave	40	13	13	11
wine	13	5	6	4
zoo	16	10	7	9
平均	35.64	11.14	11.64	8.07

表 3 列出了各个数据集的原属性数量和经过 PRFA、PRBD 和所提方法得到的属性约简属性数量。最后一行展示了各方法在所有数据集上所得约简的平均规模。从表 3 可见，大多数情况下 PRFA 取得的约简规模小于或等于 PRBD，这可能归因于 PRBD 采用的是后向删除策略。PRBD 是从所有属性的集合中删除相对冗余的属性，而在开始删除阶段有较多的属性相对于全体属性来说都是冗余的，算法可能会删除较重要的相对冗余属性，导致需要保留更多属性才能满足约简保持度量。从表中可见，本文方法在所有数据集上都取得了优于 PRFA 和 PRBD 的属性约简率。特别是在 turkiye 数据集上，约简规模从 31 下降到了 19，减少了 12 个属性。这是因为新方法使用了近似约简的思想，放宽了约简的度量标准。约简的依赖度达到条件属性集的 95%即可，稍微宽松的条件使得约简的规模可以大幅减小。

4.3 约简性能分析

为了验证新方法的属性约简质量，实验对比了各属性约简方法生成的约简在不同分类器下的性能。具体地，各数据集先去除约简以外的冗余属性，然后在 KNN($K=3$)和 SVM (LinearSVC)分类器[34]下进行 10 重交叉验证。考虑到样本的次序对 10 重交叉验证的影响，实验进行了 10 次随机 10 重交叉验证，取 10 次结果的平均值作为最终的结果。

表 4 和表 5 列出了 10 次 10 重交叉验证的平均性能，以“误分率±方差”的形式表示。不同方法之间的最优性能用粗体表示，所有数据集上的平均性能显示在最后一行。

在 KNN 分类器下，PRFA 与 RAW 的性能比较接近，而 PRBD 在多数数据集上的效果相对较差。PRFA 是 RAW 的度量保持约简，具有与 RAW 相同的依赖度，保留了 RAW 的

绝大部分信息，因此 PRFA 的性能和 RAW 在大多数数据集上非常接近。PRBD 由于在算法开始阶段的删除操作具有较大的随机性，容易删除重要的相关属性，因此其性能在大多数数据集上较差。仅在 seismic 和 sponge 数据集上，PRBD 略优于其他几种方法。新方法在 10 个数据集上都取得了最好的性能，并且平均性能也是最优的。新方法在放宽约简条件的同时保留了绝大部分的有效信息，同时避免了选择大量冗余或无关的属性以致泛化性能差的问题。因此，在获得更高的约简率的同时也能取得较好的分类结果。

表 4 KNN 分类器下各方法的约简性能

Table 4 Performance of reducts on KNN classifier

Data sets	RAW	PRFA	PRBD	Proposed
credit	31.23±0.58	30.66±0.59	33.45±0.84	**30.28 ± 0.87**
lymph	23.7±1.43	23.4±1.96	23.57±1.77	**22.95 ± 1.33**
molecular	16.43±1.67	12.04±1.45	40.08±2.97	**9.52 ± 1.67**
polish	4.09±0.04	3.78±0.03	3.85±0.06	**3.75 ± 0.05**
quality	28.93±1.65	31.34±2.22	34.78±2.39	**22.82 ± 2.31**
seismic	8.28±0.33	8.45±0.3	**8.03 ± 0.24**	8.17±0.27
solar	34.92±1.82	34.05±1.68	34.42±1.56	**33.94 ± 0.85**
spectf	24.4±1.01	17.59±0.74	24.9±1.48	**17.18 ± 0.78**
splice	34.01±0.33	28.33±0.36	51.29±0.34	**24.17 ± 0.28**
sponge	8.88±0.81	8.23±0.59	7.82±0.21	**7.68 ± 0.85**
turkiye	47.81±1.63	47.81±1.63	47.81±1.63	**47.54 ± 1.0**
wave	**28.99 ± 0.19**	36.36±0.5	67.34±0.52	34.39±0.43
wine	**5.41 ± 0.35**	5.95±0.56	7.53±0.56	5.84±1.12
zoo	**5.06 ± 0.29**	7.81±1.32	9.04±2.0	5.91±0.98
平均	21.58	21.13	28.13	**19.58**

表 5 SVM 分类器下各方法的约简性能

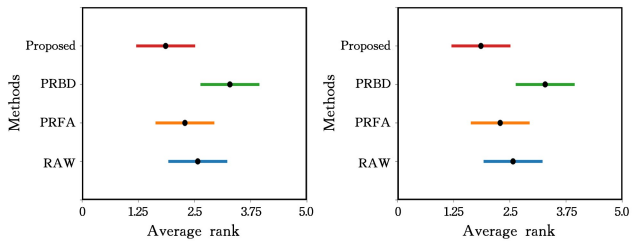
Table 5 Performance of reducts on SVM classifier

Data sets	RAW	PRFA	PRBD	Proposed
credit	28.91±0.24	**27.93±0.26**	30.62±0.24	30.5±0.40
lymph	**17.23 ± 1.36**	21.26±1.17	26.64±1.49	21.69±1.45
molecular	19.75±1.49	**18.53±1.41**	53.51±3.57	19.69±0.62
polish	**3.93±0.0**	3.93±0.0	3.93±0.0	3.93±0.0
quality	28.59±3.29	25.91±1.71	29.68±2.03	**25.07±1.01**
seismic	6.58±0.0	6.58±0.0	6.58±0.0	**6.58±0.0**
solar	30.79±0.63	30.63±0.94	**30.45±0.47**	30.52±1.15
spectf	20.97±1.11	16.1±0.44	19.78±0.66	**15.99±0.49**
splice	**17.92±0.20**	34.81±0.09	48.12±0.0	33.84±0.21
sponge	11.38±0.75	8.34±0.82	**5.2±0.15**	8.05±0.33
turkiye	38.17±0.06	38.17±0.06	38.17±0.06	**38.02±0.06**
wave	**16.7±0.09**	27.79±0.1	66.45±0.26	27.77±0.13
wine	5.35±0.40	**3.32±0.17**	7.14±0.90	4.32±0.35
zoo	5.54±1.1	**3.83±0.28**	9.48±0.82	5.41±0.64
平均	**17.99**	19.08	26.84	19.38

在 SVM 分类器上，虽然新方法的平均性能比 PRFA 和 RAW 差，但在大部分数据集上仍具有较为优异的表现。在 polish 数据集中，RAW 有 64 个属性，PRFA 和 PRBD 分别选取了 26 和 25 个属性，而新方法仅用 12 个属性，并在性能上达到了最优。在 turkiye 数据集上，PRFA 和 PRBD 都未能起到属性约简的目的，而新方法在 38.7% 的约简率下仍然取得了最好的性能。在平均性能上，新方法与 PRFA 近似，相比 RAW 较差，远高于 PRBD。

4.4 统计显著性分析

实验进一步进行了统计显著性分析，以验证新方法在统计学意义上是否优于其他方法。经过 Friedman 检验和 Nemenyi 检验，所选用的几种方法具有一定的显著性差异，其结果如图 1 所示。



(a)KNN classifier (b)SVM classifier

图 1 各种方法在不同分类器上的 Friedman 检验结果

Fig. 1 Friedman test on different methods with different classifiers

如图 1(a)所示，在 KNN 分类器上，新方法与 RAW，PRFA 和 PRBD 无相交部分，这表明新方法在统计意义上优于另外 3 种方法。

而在图 1(b)所示的 SVM 分类器上，由于 RAW 和 PRFA 的性能并不稳定，而新方法在 12 个数据集上的表现较好，因此新方法具有最高的平均 rank。新方法与 PRBD 具有显著性的优势，与 RAW 和 PRFA 并无显著性差异。

结束语 现有的大多数基于三支决策的属性约简方法都严格依赖度量准则的精确保持，一定程度上限制了其约简效率和泛化性能。本文提出了一种新的基于正域的三支近似约简方法，并由此提出了两种启发式属性约简算法。由于放宽了度量准则的限制，新方法具有更快的运行速度。同时，UCI 实验分析表明，新方法在属性约简效率和约简质量上比其他方法更具优越性。本文方法仅适用于离散型数据，连续型数据需要进行离散化预处理。模糊粗糙集可同时处理离散型和连续性数据，下一步将引入模糊粗糙集理论以扩展本文方法的适应范围。另外，大规模数据的近似约简亦值得进一步深入研究。

参 考 文 献

[1] LI Y, LI T, LIU H. Recent advances in feature selection and its applications [J]. Knowledge and Information Systems, 2017, 53(3): 551-577.

[2] BISHOP C M. Pattern recognition and machine learning [M]. New York: Springer, 2006.

[3] ARMANFARD N, REILLY J P, KOMEILI M. Local feature selection for data classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(6): 1217-1227.

[4] MIAO D Q, ZHAO Y, YAO Y Y, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model[J]. Information Sciences, 2009, 179(24): 4140-4150.

[5] LI F, MIAO D Q, PEDRYCZ W. Granular multi-label feature selection based on mutual information [J]. Pattern Recognition, 2017, 67: 410-423.

[6] YAO Y Y, ZHAO Y. Discernibility matrix simplification for constructing attribute reducts [J]. Information Sciences, 2009, 179(7): 867-882.

[7] LAI Z H, YONG X, JIAN Y, et al. Rotational invariant dimensionality reduction algorithms [J]. IEEE Transactions on Cybernetics, 2016, 47(11): 3733-3746.

[8] WANG X,PENG Z H,LI J Y,et al. Method of Concept Reduction Based on Concept Discernibility Matrix [J]. Computer Science,2021,48(1):125-130.

[9] ZENG H K,MI J S,LI Z L. Dynamic Updating Method of Concepts and Reduction in Formal Context [J]. Computer Science,2021,48(1):131-135.

[10] PAWLAK Z. Rough sets [J]. International Journal of Computer & Information Sciences,1982,11(5):341-356.

[11] YAO Y Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences,2010,180(3):341-353.

[12] YAO Y Y. Three-way decision and granular computing [J]. International Journal of Approximate Reasoning,2018,103:107-123.

[13] YAO Y Y. Three-way granular computing,rough sets,and formal concept analysis [J]. International Journal of Approximate Reasoning,2020,116:106-125.

[14] YAO Y Y,ZHAO Y. Attribute reduction in decision-theoretic rough set models [J]. Information Sciences,2008,178(17):3356-3373.

[15] YAO Y Y. Three-way conflict analysis:Reformulations and extensions of the Pawlak model [J]. Knowledge-Based Systems,2019,180:26-37.

[16] YAO Y Y. Three-way decisions and cognitive computing [J]. Cognitive Computation,2016,8(4):543-554.

[17] ZHAO Y,WONG S K M,YAO Y Y. A note on attribute reduction in the decision-theoretic rough set model [C]// Transactions on Rough Sets XIII. Berlin:Springer,2011:260-275.

[18] LI H X,ZHOU X Z,ZHAO J B,et al. Non-monotonic attribute reduction in decision-theoretic rough sets [J]. Fundamenta Informaticae,2013,126(4):415-432.

[19] MA X A,WANG G Y,YU H. Heuristic method to attribute reduction for decision region distribution preservation [J]. Ruan Jian Xue Bao/Journal of Software,2014,25(8):1761-1780.

[20] MA X A,WANG G Y,YU H,et al. Decision region distribution preservation reduction in decision-theoretic rough set model [J]. Information Sciences,2014,278:614-640.

[21] GAO C,LAI Z H,ZHOU J,et al. Maximum decision entropy-based attribute reduction in decision-theoretic rough set model [J]. Knowledge-Based Systems,2018,143:179-191.

[22] ZHANG X Y,MIAO D Q. Region-based quantitative and hierarchical attribute reduction in the two-category decision theoretic rough set model [J]. Knowledge-Based Systems,2014,71:146-161.

[23] ZHANG X Y,MIAO D Q. Reduction target structure-based hierarchical attribute reduction for two-category decision-theoretic rough sets [J]. Information Sciences,2014,277:755-776.

[24] JIA X Y,LIAO W H,TANG Z M,et al. Minimum cost attribute reduction in decision-theoretic rough set models [J]. Information Sciences,2013,219:151-167.

[25] JIA X Y,TANG Z M,LIAO W H,et al. On an optimization representation of decision-theoretic rough set model [J]. International Journal of Approximate Reasoning,2014,55(1):156-166.

[26] LIAO S J,ZHU Q X,FAN M. Cost-sensitive attribute reduction in decision-theoretic rough set models [J]. Mathematical Problems in Engineering,2014,35(1):1-9.

[27] SLEZAK D. Approximate reducts in decision tables [C]// Proceedings of IPMU' 96. Granada:Spain,1996:1159-1164.

[28] SLEZAK D. Approximate entropy reducts [J]. Fundamental Informaticae,2002,53(3/4):365-390.

[29] YANG M. Approximate reduction based on conditional information entropy in decision tables [J]. Acta Electronica Sinica,2007,35(11):2156-2160.

[30] YANG X,LI T R,LIU D,et al. A unified framework of dynamic three-way probabilistic rough sets [J]. Information Sciences,2017,420:126-147.

[31] ZHANG Q H,XIE Q,WANG G Y. A survey on rough set theory and its applications [J]. CAAI Transactions on Intelligence Technology,2016,1(4):323-333.

[32] YAO Y Y,WONG S K M. A decision theoretic framework for approximating concepts [J]. International Journal of Man-machine Studies,1992,37(6):793-809.

[33] LICHMAN M. UCI machine learning repository [DB/OL]. University of California,Irvine,CA,USA,2013. http://archive.ics.uci.edu/ml.

[34] PEDREGOSA F,VAROQUAUX G,GRAMFORT A,et al. Scikit-learn:Machine learning in Python [J]. Journal of Machine Learning Research,2011,12:2825-2830.

WANG Zhi-cheng, born in 1998, postgraduate. His main research interests include machine learning and granular computing.

GAO Can, born in 1983, Ph.D, assistant professor, master supervisor. His main research interests include machine learning and computer vision.

(责任编辑:李亚辉)