http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0975

用 Critic 赋权法加权邻域粗糙集的属性约简算法

吴尚智[1,*]，任艺璇[2]，葛舒悦[1]，王立泰[1]，王志宁[1]

(1. 西北师范大学 计算机科学与工程学院，兰州 730070；2. 国家税务总局武安市税务局，邯郸 056300)

摘 要：邻域粗糙集相比经典粗糙集能够处理非离散型数据和高维度数据，具有获得简化数据且不降低数据处理的能力。针对邻域粗糙集中每个属性具有相同权重，且每个属性对决策的影响程度不同的问题，提出用 Critic 赋权法加权邻域粗糙集的属性约简算法。使用 Critic 赋权法为条件属性赋权，引入加权距离函数计算邻域关系，得到加权邻域关系；构建加权邻域粗糙集，采用属性依赖度和重要度评估子集的重要性，使用等距搜索寻找最佳阈值，进行属性约简找到最优属性子集；采用 UCI 库中的 10 个数据集进行实验验证，与传统邻域粗糙集的属性约简算法的性能进行比较分析。实验结果表明：所提算法可得到最小属性约简集，并可保证约简后数据的分类准确率，具有有效性和实际应用价值。

关 键 词：Critic 赋权法；邻域粗糙集；属性约简；加权邻域关系；属性依赖度

中图分类号：TP301.6

文献标志码：A **文章编号**：1001-5965(2025)01-0075-10

粗糙集理论[1]是 Pawlak 提出的一种处理知识中模糊性和不确定性问题的数学工具，属性约简是粗糙集理论研究的核心内容之一。经典粗糙集理论拥有严格的等价关系。非离散型数据需要离散化处理[2]，实际应用中，所处理数据往往为非离散型。许多学者扩展了粗糙集模型研究，如基于邻域关系的邻域粗糙集[3]、基于模糊对等关系的模糊粗糙集[4]等，这些扩展粗糙集模型已被广泛应用于图像处理[5]、数据挖掘[6]、人工智能[7]等诸多领域。

大数据时代，数据量大且数据维度很高，传统方法处理高维度数据时，很难达到较高的效率，因此，属性约简是十分必要的。属性约简是对经过信息粒化与近似逼近的原始数据集的条件属性进行约简，剔除冗余属性，得到更简洁的数据，且不降低数据处理的能力。

邻域粗糙集能更好地进行属性约简。文献 [8] 讨论了邻域算子和粗糙集近似算子的关系，提出了基于邻域系统的广义粗糙集近似算子。文献 [9] 指出，可以使用邻域模型为不同属性分配不同的阈值，进而减少数值和分类特征，并提出了基于邻域粗糙集来处理异构特征子集选择。文献 [10] 通过上下边界区域定义特征重要性，使用基于可辨别矩阵的特征选择方法，提出了基于邻域粗糙集的不平衡数据特征选择。文献 [11] 结合优势关系和邻域关系的优点，提出了基于优势邻域粗糙集中的并行属性约简。文献 [12] 结合邻域粗糙集和 F-粗糙集的优势，提出了 F-邻域粗糙集和 F-邻域并行约简。文献 [13] 结合 delta 邻域和 k 近邻的优点，提出了一种新的邻域粗糙集模型，称为 k-最近邻域粗糙集。但在邻域粗糙集研究上，研究者没有考虑属性的重要性，每种属性对结果的影响是不同的，如果提前考虑条件属性和决策之间的内部相关性，可以更容易地选择具有高相关性和依赖性的属性。因此，本文用赋权法对属性进行赋权，从而使不同属性具有

收稿日期：2022-12-15；**录用日期**：2023-05-05；**网络出版时间**：2023-05-15 09:48
网络出版地址：link.cnki.net/urlid/11.2625.V.20230512.1629.002
基金项目：国家自然科学基金 (12161082,61861039)
* **通信作者**. E-mail：wusz@nwnu.edu.cn

引用格式：吴尚智，任艺璇，葛舒悦，等. 用 Critic 赋权法加权邻域粗糙集的属性约简算法 [J]. 北京航空航天大学学报，2025，51（1）：75-84. WU S Z，REN Y X，GE S Y，et al. An attribute reduction algorithm of weighting neighborhood rough sets with Critic method [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（1）：75-84（in Chinese）.

不同的权重,进而更好地对数据做出更加高效准确的处理。

属性的权重是其重要性的体现。文献 [14] 利用决策树提出了一种多粒化区间值决策的粒化加权模型。文献 [15] 利用基于核函数的模糊粗糙集定义了加权 Parzen 函数,提出了一种新的基于加权 Parzen 函数的 k 近邻分类算法,以提高分类精度。文献 [16] 提出了一种新的基于有序加权平均的权重选择策略,以提高模糊粗糙集的抗噪声能力。赋权法主要有主观赋权法和客观赋权法。主观赋权法对属性的权重确定依赖于人的主观性,在使用时需要相关专家的介入;而客观赋权法通过对原始数据的计算便能得到权重值,被广泛使用。常用的客观赋权法有熵权法[17]、标准离差法[18] 及 Critic 赋权法[19]。吴希[20] 通过比较发现,Critic 赋权法计算权重时,既考虑了关联性对指标的影响,又考虑了变异性对指标的影响,比其他 2 种方法更加全面。

为解决邻域粗糙集的属性约简未考虑到每个属性对决策影响不同的问题,本文提出了使用 Critic 赋权法加权邻域粗糙集的属性约简算法。首先,由于条件属性对决策的影响不同,用 Critic 赋权法计算属性的变异性评价和冲突性指数,为条件属性赋权。然后,确定加权邻域关系时,使用加权距离函数得到加权邻域关系及其他相关定义和性质,构建加权邻域粗糙集。最后,将依赖度[21] 和属性重要度作为评价子集重要性的指标,进行属性约简。在 UCI 库中的 10 个数据集上进行实验验证,结果表明,用 Critic 赋权法加权邻域粗糙集的属性约简算法不仅能够降低维度,还能提高分类准确率。

1 相关理论

1.1 邻域粗糙集

给定一个信息系统 $S=(U,C,D,f)$,该系统是由 m 个属性构成的数据对象集,其中,$U=\{x_1,x_2,\cdots,x_n\}$ 为全部样本集合,$C=\{a_1,a_2,\cdots,a_m\}$ 为样本的条件属性集合,D 为决策属性集合,$f\left(x_i,a_j\right)$ 为样本 x_i 在属性 a_j 上的取值。

定义 1 对于信息系统 S,在集合 U 中的任意 $x_i\in U$,则 x_i 在 U 上的邻域 $\delta(x_i)$ 定义为

$$\delta(x_i)=\left\{x_j\in U,\Delta\left(x_i,x_j\right)\leqslant\delta\right\} \tag{1}$$

式中:δ 为邻域半径;x_j 为集合 U 中的元素;Δ 为 U 上的距离函数,且 (U,Δ) 为度量空间,对于任意的 $x_1,x_2,x_3\in U$,Δ 满足:

1) $\Delta(x_1,x_2)\geqslant 0$,当且仅当 $x_1=x_2$ 时,$\Delta(x_1,x_2)=0$;

2) $\Delta(x_1,x_2)=\Delta(x_2,x_1)$;

3) $\Delta(x_1,x_3)\leqslant\Delta(x_1,x_2)+\Delta(x_2,x_3)$。

定义 2 给定一个五元组为 $\{U,A,V,f,\delta\}$,$A=C\cup D$ 且 $C\cap D=\varnothing$。其中,U 为对象集,C 为条件属性集,D 为决策属性集,V 为属性值集合,f 为信息函数。当 D 为非空有限集合时,该系统为邻域决策系统 $[NDS]=\{U,C\cup D,V,f,\delta\}$。

定义 3 给定一个邻域决策系统 $[NDS]=\{U,C\cup D,V,f,\delta\}$,其中,$U=\{x_1,x_2,\cdots,x_n\}$ 为有限样本集合,$B\subseteq C$,$\forall x_i\in U$,称 N_B 为属性子集 B 在 U 上的邻域关系,定义为

$$N_B=\left\{\left(x_i,x_j\right)|x_i,x_j\in U\text{且}x_j\in\delta_B(x_i),\delta\geqslant 0\right\} \tag{2}$$

式中:$\delta_B(x_i)=\left\{x_j\in U,\Delta\left(x_i,x_j\right)\leqslant\delta\right\}$ 表示 x_i 在属性子集 B 上的 δ-邻域,$\delta=0$ 时,邻域关系变为等价关系。

定义 4 给定一个邻域决策系统 $[NDS]=\{U,C\cup D,V,f,\delta\}$,集合 $X_1,X_2,\cdots,X_N$ 为 U 在 D 上的划分,$B\subseteq C$ 生成 U 上的邻域关系 N_B,$\delta_B(x_i)$ 表示 x_i 在 B 上的 δ-邻域,则 D 关于属性子集 B 的上近似、下近似集定义为

$$\overline{N_B}(D)=\bigcup_{i=1}^{N}\overline{N_B}(X_i) \tag{3}$$

$$\underline{N_B}(D)=\bigcup_{i=1}^{N}\underline{N_B}(X_i) \tag{4}$$

式中:

$$\overline{N_B}(X)=\{x_i|\delta_B(x_i)\cap X\neq\varnothing,x_i\in U\} \tag{5}$$

$$\underline{N_B}(X)=\{x_i|\delta_B(x_i)\subseteq X,x_i\in U\} \tag{6}$$

另外,D 关于属性子集 B 的正域为 $[pos]_B(D)=\underline{N_B}(D)$,负域为 $[neg]_B(D)=U-\overline{N_B}(D)$,边界域为 $[bnd]_B(D)=\overline{N_B}(D)-\underline{N_B}(D)$。

定义 5 给定一个邻域决策系统 $[NDS]=\{U,C\cup D,V,f,\delta\}$,子集 B 是属性集合 C 的子集,则 D 对于属性子集 B 的依赖度为

$$\gamma_B(D)=|[pos]_B(D)|/|U|\qquad 0\leqslant\gamma_B(D)\leqslant 1 \tag{7}$$

定义 6 给定一个邻域决策系统 $[NDS]=\{U,C\cup D,V,f,\delta\}$,$B\subseteq C$,若属性子集 B 满足:

1) $\gamma_B(D)=\gamma_C(D)$;

2) $\forall a\in B,\gamma_{B-\{a\}}(D)<\gamma_B(D)$。

则称属性子集 B 是属性集合 C 的一个属性约简。

1.2 Critic 赋权法

Critic 赋权法[22] 是由 Diakoulaki 提出的一种处理多元准则的客观赋权法,其基本思想是:构造权重时,以对比强度和冲突性作为基础。其中,对比强度用标准差的形式体现,一般来说,标准差越大,各类别的差距越大。而冲突性则用各指标之间的相关系数体现,如果指标之间的负相关性较大,则冲突性较大。引入信息量,其数学计算公式为:$H_j=S_j\sum_{i=1}^{p}\left(1-r_{ij}\right)$。其中,$p$ 为属性数量,S_j 为标准差,r_{ij} 为相关系数。

H_j 越大, 第 j 个属性对决策的影响作用越大, 应给其分配更多的权重。因此, 第 j 个属性的客观权重表示为 $W_j = H_j \Big/ \sum_{j=1}^{p} H_j$。

2　属性约简算法

计算邻域关系前有必要挖掘属性和决策之间的内在相关性。如果属性与决策之间的内在相关性较高, 则属性的权重较大; 否则, 权重应更小。

2.1　Critic 赋权法计算属性权重

给定一个信息系统 $S=(U,C,D,f)$, 该信息系统是有 m 个属性构成的数据对象集。

1) 无量纲化处理。不同的属性之间数量级不同, 会对结果有不同影响, 需要对数据进行无量纲化处理, 消除不同属性之间的数量级差异。

2) 计算属性变异性评价。

$$M(a_j)=\frac{1}{n}\sum_{i=1}^{n} f(x_i,a_j) \tag{8}$$

$$S_{\mathrm{td}}(a_j)=\sqrt{\frac{\sum_{i=1}^{n}\left(f(x_i,a_j)-M(a_j)\right)^2}{n-1}} \tag{9}$$

式(8)表示属性 a_j 的平均值, 式(9)表示属性 a_j 的标准差。

在 Critic 赋权法中, 用标准差反映各属性的取值波动情况, 属性的数值之间差异越大, 则标准差越大, 越能反映更多的信息, 该属性对决策的影响强度也越强, 应赋予较大的权重。

3) 计算属性冲突性指数。

$$R(a_j)=\sum_{i=1}^{m}\left(1-r(a_i,a_j)\right) \tag{10}$$

式(10)表示属性 a_j 的冲突性指数, $r(a_i,a_j)$ 表示属性 a_i 与属性 a_j 之间的相关系数, 由相似度函数求得。

属性冲突性由属性与属性之间的相关系数决定, 与其他属性的相关性越高, 则该属性的冲突性就越小, 反映的信息相同度越高, 对决策的影响强度也越小, 应赋予较小的权重。用 Critic 赋权法时, 当标准差一定时, 属性之间冲突性越小, 所赋权重也越小; 冲突性越大, 所赋权重也越大。

4) 计算信息量。

$$H(a_j)=S_{\mathrm{td}}(a_j)R(a_j) \tag{11}$$

5) 计算属性客观权重。

$$w(a_j)=H(a_j)\Big/\sum_{j=1}^{m}H(a_j) \tag{12}$$

定义 7　给定信息系统 $S=(U,C,D,f)$, 属性的权重集合 w 定义为

$$w=\{w(a_1),w(a_2),\cdots,w(a_m)\} \tag{13}$$

2.2　构建加权邻域粗糙集

传统邻域粗糙集在确定邻域时未考虑属性对决策的影响不同, 却对评估子集重要性有着一定的影响, 因此, 用不同的权重计算不同属性的邻域相似关系。本文引入加权距离函数确定邻域关系, 构建加权邻域粗糙集。

给定一个信息系统 $S=(U,C,D,f,w)$, $U=\{x_1,x_2,\cdots,x_n\}$ 表示全部样本集合, $C=\{a_1,a_2,\cdots,a_m\}$ 表示条件属性, D 表示决策属性, $f(x_i,a_j)$ 表示样本 x_i 在属性 a_j 上的取值, $w=\{w(a_1),w(a_2),\cdots,w(a_m)\}$ 表示给定信息系统的属性权重集合。

定义 8　对于给定信息系统中, U 为有限非空样本集合, C 表示条件属性, $w(a_k)\in w$, $B\subset C$。假设 $\forall x_i,x_j\in U,\forall a_k\in B$, 则加权距离函数定义为

$$[wd]_B(x_i,x_j)=\sqrt{\sum_{k=1}^{m}\left[w(a_k)\left(f(x_i,a_k)-f(x_j,a_k)\right)\right]^2} \tag{14}$$

当 $x_1,x_2,x_3\in U$, 加权距离函数满足如下关系:

1) $[wd]_B(x_1,x_2)\geqslant 0$, 当且仅当 $x_1=x_2$ 时 $[wd]_B(x_1,x_2)=0$;

2) $[wd]_B(x_1,x_2)=[wd]_B(x_2,x_1)$;

3) $[wd]_B(x_1,x_3)\leqslant[wd]_B(x_1,x_2)+[wd]_B(x_2,x_3)$。

性质 1　对于 $B_1\subset B_2\subseteq C$, 任意 $x_i,x_j\in U$, 加权距离 $[wd]_B(x_i,x_j)$ 有: $[wd]_{B_1}(x_i,x_j)\leqslant[wd]_{B_2}(x_i,x_j)$。

证明　设 $B_1=\{a_1,a_2,\cdots,a_p\}$, $B_2=\{a_1,a_2,\cdots,a_m\}$, $p\leqslant m$, 则有

$$[wd]_{B_1}(x_i,x_j)=\sqrt{\sum_{k=1}^{p}\left[w(a_k)\left(f(x_i,a_k)-f(x_j,a_k)\right)\right]^2}$$

$$[wd]_{B_2}(x_i,x_j)=\sqrt{\sum_{k=1}^{m}\left[w(a_k)\left(f(x_i,a_k)-f(x_j,a_k)\right)\right]^2}$$

因为 $\left[w(a_k)\left(f(x_i,a_k)-f(x_j,a_k)\right)\right]^2$ 的值恒大于 0, 且 $p\leqslant m$, 所以, $[wd]_{B_1}(x_i,x_j)\leqslant[wd]_{B_2}(x_i,x_j)$。

定义 9　对于信息系统 S, $\forall x_i\in U$, $[wd]_B$ 为 U 上的加权距离函数, B 为属性集合 C 的属性子集, δ 为正实数, 则 x_i 在属性子集 B 上的 δ-加权邻域定义为

$$[w\delta]_B(x_i)=\{x_j\in U,[wd]_B(x_i,x_j)\leqslant\delta\} \tag{15}$$

$[w\delta]_B(x_i)$ 又称为 x_i 在 B 上的 δ-加权邻域信息粒, 加权距离函数 $[wd]_B(x_i,x_j)$ 和邻域半径 δ 决定了 x_i 的邻域大小。邻域半径 δ 的值越大, 邻域越大, 邻

域内点越多。邻域粒子簇$\{[w\delta]_B(x_i)|i=1,2,\cdots,n\}$构成了$U$的一个覆盖。

性质 2 给定任意一个正整数$\delta>0$, $B_1\subset B_2\subseteq C$, 对于$\forall x_i\in U$, 都有$[w\delta]_{B_2}(x_i)\subseteq[w\delta]_{B_1}(x_i)$。

证明 令$y\in[w\delta]_{B_2}(x_i)$, 则$[wd]_{B_2}(x_i,y)\leqslant\delta$, 根据性质 1 可知, $[wd]_{B_1}(x_i,y)\leqslant[wd]_{B_2}(x_i,y)\leqslant\delta$。因此, $y\in[w\delta]_{B_1}(x_i)$, 即$[w\delta]_{B_2}(x_i)\subseteq[w\delta]_{B_1}(x_i)$。

定义 10 对于信息系统S, B为属性集合C的子集, δ为正实数, 则属性子集B导出的x_i在U上的加权邻域关系定义为

$$[WN]_B=\left\{\left(x_i,x_j\right)|x_i,x_j\in U且x_j\in[w\delta]_B(x_i),\delta\geqslant 0\right\} \tag{16}$$

式中: $[w\delta]_B(x_i)=\left\{x_j\in U,[wd]_B\left(x_i,x_j\right)\leqslant\delta\right\}$表示$x_i$在属性子集$B$上的加权$\delta$-邻域, 即给定$\forall x_i,x_j\in U$, $\forall a_k\in B$, 加权邻域关系表达式为

$$[WN]_B=\left\{\left(x_i,x_j\right)\left|\sqrt{\sum_{k=1}^{m}\left[w(a_k)\left(f(x_i,a_k)-f\left(x_j,a_k\right)\right)\right]^2}\leqslant\delta\right.\right\} \tag{17}$$

计算邻域关系时, 当$w(a_k)>1$时, 则表示属性a_k在计算时较为重要; 当$w(a_k)<1$时, 则表示该属性的重要程度下降; 当$w(a_k)=1$时, $[WN]_B$则与传统邻域关系相同。即加权邻域关系是传统邻域关系的扩展, 而传统的邻域关系是加权邻域关系的特例。

性质 3 给定信息系统S, $[WN]_B$是加权邻域关系, 对于U中的任意x_i,x_j, 满足自反性和对称性, 并且根据定义 8 中加权距离函数所满足的关系易证。

性质 4 给定信息系统S, $B_1\subset B_2\subseteq C$, 对于$\delta$为给定正整数, $\forall x_i,x_j\in U$, 都有$[WN]_{B_2}\subseteq[WN]_{B_1}$。

证明 若$(x_i,x_j)\in[WN]_{B_2}$, 则$x_j\in[w\delta]_{B_2}(x_i)$, 由性质 2 得, $[w\delta]_{B_2}(x_i)\subseteq[w\delta]_{B_1}(x_i)$, 因此, $x_j\in[w\delta]_{B_1}(x_i)$, 即$(x_i,x_j)\in[WN]_{B_1}$。故有$[WN]_{B_2}\subseteq[WN]_{B_1}$。

性质 5 给定 2 个正整数δ_1和δ_2, $\delta_1<\delta_2$, 对任意的$B\subseteq C$, 都有$[WN]_B^{\delta_1}\subseteq[WN]_B^{\delta_2}$。

证明 若$(x_i,x_j)\in[WN]_B^{\delta_1}$, 则$x_j\in[w\delta]_B^{\delta_1}(x_i)$, $[wd]_B(x_i,x_j)\leqslant\delta_1<\delta_2$。因此, $x_j\in[w\delta]_B^{\delta_2}(x_i)$, 即$(x_i,x_j)\in[WN]_B^{\delta_2}$。故有$[WN]_B^{\delta_1}\subseteq[WN]_B^{\delta_2}$。

定义 11 给定信息系统S, $X\subseteq U$, $B\subseteq C$, $[w\delta]_B(x_i)$为x_i在B上的δ-加权邻域, 则X对于B的上、下近似表示为

$$\overline{[WN]_B}(X)=\{x_i|[w\delta]_B(x_i)\cap X\neq\varnothing,x_i\in U\} \tag{18}$$

$$\underline{[WN]_B}(X)=\{x_i|[w\delta]_B(x_i)\subseteq X,x_i\in U\} \tag{19}$$

当$\overline{[WN]_B}(X)=\underline{[WN]_B}(X)$时, 说明$X$可以用加权邻域关系来精确表示; 反之, 则不可以。X为U上的一个加权邻域粗糙集。

定义 12 给定信息系统S, B为条件属性C的子集, $B\subseteq C$, $U/D=\{X_1,X_2,\cdots,X_k\}$为$D$在$U$上的划分, 则$D$关于$B$的加权上、下近似表示为

$$\overline{[WN]_B}(D)=\bigcup_{i=1}^{k}\overline{[WN]_B}(X_i) \tag{20}$$

$$\underline{[WN]_B}(D)=\bigcup_{i=1}^{k}\underline{[WN]_B}(X_i) \tag{21}$$

另外, D关于B的正域、负域和边界域分别表示为: $[Wpos]_B(D)=\underline{[WN]_B}(D)$, $[Wneg]_B(D)=U-\overline{[WN]_B}(D)$, $[Wbnd]_B(D)=\overline{[WN]_B}(D)-\underline{[WN]_B}(D)$。当边界域等于零时, 即上下近似相同, 则$B$是精确的。当正域集合内的样本越多时, 则表示属性子集B和决策属性集D的相关性较高, B的近似能力和判定能力越强。

定义 13 给定信息系统S, C为条件属性, D为决策属性, B为条件属性子集, $B\subseteq C$, $[WN]_B$为U上的加权邻域相似关系, 则D对于B的加权依赖度表示为

$$[w\gamma]_B(D)=\left|[Wpos]_B(D)\right|/|U| \tag{22}$$

依赖度反映了条件属性子集和决策属性子集相关性, 依赖度的值越大, 则它们的相关性越大, 即属性子集B逼近于D的能力越强。

由性质 2 可知, 依赖度和属性子集的大小呈正相关, 当属性子集B中的属性增加时, δ-加权邻域越小, D关于B的下近似集越大, 依赖度越大。同理, 由性质 5 可知, 与邻域半径的大小呈现负相关, δ越大, 依赖度越小。

依赖度除了与属性子集大小及邻域半径相关外, 也与属性的权重有很大的关系。通过实例对比加权邻域粗糙集依赖度和传统邻域粗糙集的依赖度, 证明权重对属性子集的依赖度有影响。

实例如下: 表 1 给定一个信息系统, 其中, U为样本空间, $C=\{a_1,a_2,a_3,a_4\}$为条件属性集, $D=\{d\}$为决策属性集。样本可以被分为 2 类: $D_1=\{x_1,x_2,x_3,x_4,x_5,x_6\}$和$D_2=\{x_7,x_8,x_9,x_{10},x_{11},x_{12}\}$。

给定 2 个属性集合$B_1=\{a_1,a_2\}$, $B_2=\{a_3,a_4\}$, 邻域半径δ为 0.01, 则属性子集B_1、B_2下的δ-邻域如表 2 所示。可以看出, D关于属性子集B_1和B_2的正域均为U, 则$\gamma_1=\gamma_2=1$, 即B_1近似于D的能力与B_2近似于D的能力相同。

相同的信息系统中, 属性子集B_1、B_2下的加权邻域如表 3 所示。可以得到: $[Wpos]_{B_1}(D)=\{x_2,x_3,x_4,x_{10},x_{11},x_{12}\}$, $[Wpos]_{B_2}(D)=\{x_2,x_3,x_5,x_6,x_7,x_9,x_{11},x_{12}\}$, 此时$[w\gamma]_1=0.5$, $[w\gamma]_2=0.67$。通过加权邻域粗糙

集的依赖度可知, B_2 近似于 D 的能力优于 B_1 近似于 D 的能力。

表 1　某信息系统统计信息

Table 1　Statistical information of an information system

U	a_1	a_2	a_3	a_4	d
x_1	0.06	0.91	0.11	0.08	1
x_2	0.03	0.92	0.32	0.10	1
x_3	0.03	0.43	0.31	0.14	1
x_4	0.06	0.41	0.45	0.11	1
x_5	0.05	0.68	0.67	0.13	1
x_6	0.09	0.12	0.81	0.14	1
x_7	0.08	0.69	0.23	0.03	2
x_8	0.12	0.21	0.39	0.07	2
x_9	0.10	0.89	0.61	0.03	2
x_{10}	0.14	0.93	0.78	0.09	2
x_{11}	0.13	0.41	0.74	0.07	2
x_{12}	0.14	0.43	0.91	0.03	2

表 2　B_1、B_2 的邻域

Table 2　Neighborhood of B_1 and B_2

U	B_1	B_2
x_1	$\{x_1\}$	$\{x_1\}$
x_2	$\{x_2\}$	$\{x_2\}$
x_3	$\{x_3\}$	$\{x_3\}$
x_4	$\{x_4\}$	$\{x_4\}$
x_5	$\{x_5\}$	$\{x_5\}$
x_6	$\{x_6\}$	$\{x_6\}$
x_7	$\{x_7\}$	$\{x_7\}$
x_8	$\{x_8\}$	$\{x_8\}$
x_9	$\{x_9\}$	$\{x_9\}$
x_{10}	$\{x_{10}\}$	$\{x_{10}\}$
x_{11}	$\{x_{11}\}$	$\{x_{11}\}$
x_{12}	$\{x_{12}\}$	$\{x_{12}\}$

表 3　B_1、B_2 的加权邻域

Table 3　Weighted neighborhood of B_1 and B_2

U	B_1	B_2
x_1	$\{x_1, x_2, x_5, x_7, x_9\}$	$\{x_1, x_2, x_8\}$
x_2	$\{x_1, x_2, x_5\}$	$\{x_1, x_2, x_4\}$
x_3	$\{x_3, x_4, x_5\}$	$\{x_3, x_5\}$
x_4	$\{x_3, x_4\}$	$\{x_2, x_4, x_5, x_{10}\}$
x_5	$\{x_1, x_2, x_3, x_5, x_7\}$	$\{x_3, x_4, x_5, x_6\}$
x_6	$\{x_6, x_8\}$	$\{x_5, x_6\}$
x_7	$\{x_1, x_5, x_7\}$	$\{x_7, x_9\}$
x_8	$\{x_6, x_8, x_{11}, x_{12}\}$	$\{x_1, x_8, x_{10}, x_{11}\}$
x_9	$\{x_1, x_7, x_9, x_{10}\}$	$\{x_7, x_9, x_{12}\}$
x_{10}	$\{x_9, x_{10}\}$	$\{x_4, x_8, x_{10}, x_{11}\}$
x_{11}	$\{x_8, x_{11}, x_{12}\}$	$\{x_8, x_{10}, x_{11}\}$
x_{12}	$\{x_8, x_{11}, x_{12}\}$	$\{x_9, x_{12}\}$

使用邻域粗糙集时, B_1 和 B_2 近似于 D 的能力是相同的, 通过实验得出结果如表 4 所示, 选择 B_1 作为特征子集时的准确率没有选择 B_2 作为特征子集时的准确率高。然而, 用加权邻域粗糙集对比加权依赖度, 可知 B_2 近似于 D 的能力优于 B_1 近似于 D 的能力。结果与实验结果相同。

表 4　实验结果

Table 4　Experimental results

属性子集	准确率/%	
	KNN	SVM
$\{a_1, a_2\}$	75	83.3
$\{a_3, a_4\}$	83.3	91.7

2.3　算法描述

定义 14　给定一个信息系统 S, 任意一个属性子集 $B \subseteq C$, 属性 $a \in C - B$, 则属性 a 对决策属性 D 的加权属性重要度为

$$[Wsig](a, B, D) = [w\gamma]_{B+\{a\}}(D) - [w\gamma]_B(D) \qquad (23)$$

属性 a 对决策属性 D 的加权属性重要度表示原条件属性集 B 中没有属性 a, 则增加属性 a 后, 决策属性 D 对条件属性 B 的加权依赖度增加的程度。

用 Critic 赋权法加权邻域粗糙集的属性约简算法描述如下:

输入: 数据集、邻域半径。

输出: 约简属性子集 Red。

步骤 1　对数据集进行标准归一化处理。

步骤 2　使用式(8)和式(9)计算每个属性的平均值和标准差。

步骤 3　使用式(10)计算属性之间的相关系数及每个属性的冲突性指数。

步骤 4　使用式(11)计算每个属性的信息量, 式(12)计算每个属性的权重, 构造属性的权重集合(式(13))。

步骤 5　初始化约简属性子集 Red 值为空。

步骤 6　用式(22)计算每个属性的加权依赖度。

步骤 7　使用式(23)计算属性的加权属性重要度, 选择加权属性重要度最大的属性。

步骤 8　当加权属性重要度不等于零时, 执行步骤 9, 否则跳转到步骤 11。

步骤 9　把属性重要度最大的属性加入到约简属性子集 Red 中。

步骤 10　遍历未选择属性集合的每个属性 a_i, 使用式(22)计算 $[w\gamma]_{\mathrm{Red}+\{a_i\}}(D)$, 返回步骤 7。

步骤 11　算法结束。

算法中, 先输入数据集、邻域半径, 经过计算得到约简后的属性子集。在步骤 1～步骤 4 中, 用 Critic

赋权法计算每个属性的权重值，步骤6和步骤8中用加权距离函数(式(14))计算每个样本的加权邻域类(式(15))，根据加权邻域类计算每个分类的加权下近似集(式(19))，再整理合并所有下近似集(式(21))，计算出加权依赖度(式(22))。

由邻域半径过大会导致无法得到有效特征属性集合，邻域半径过小会导致特征属性集合为空。为此，本文实验选择不同的数据集确定不同的邻域半径参数，提高准确率。

3 实验结果及分析

实验数据选用UCI数据库中10个数据集，数据集的基本信息如表5所示，即每个数据集的样本数和属性数各不相同。各数据集中的属性最少的数据集中有6个属性，属性最多的数据集中有54个属性，各数据集均无缺失数据。

为验证本文算法的有效性，用约简后的属性对数据构建分类进一步验证约简后数据的分类能力。

表5 数据集信息

Table 5 Dataset information

数据集	样本数	属性数
wine	178	13
glass	214	9
wpbc	198	33
diab	768	8
bupa	345	6
horse	366	27
seeds	210	7
divorce	169	54
hcv	1 385	14
german	1 000	20

3.1 邻域半径

在邻域粗糙集的邻域半径选择中，有些学者[23]使用标准差计算邻域半径，计算公式为

$$\delta = S_{\mathrm{td}}(a_i)/\lambda \tag{24}$$

式中：$S_{\mathrm{td}}(a_i)$为标准差函数，a_i为条件属性；λ为设定参数，调整邻域大小。

本文中，将样本数据根据十折交叉验证[24]随机划分为2部分，其中，90%作为训练集训练模型，10%作为测试集测试模型效果。分别选取λ为0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9进行实验，找到每个数据集的最优参数。对于本文中提出的属性约简算法，不同参数对应的属性约简率和分类准确率如图1～图10所示。

由图2、图5、图7可知，λ不为0时也会出现约简率为0的情况，说明此时无法得到有效特征属性集合。从图1～图10可知，不同的数据集分别在不同的λ值下得到最佳分类准确率。实验中，λ的取值为最佳准确率所对应的数值，进而得到最佳分类准确率。

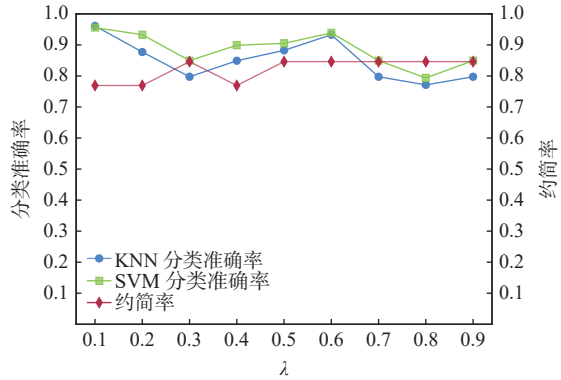

图1 不同λ值对应的属性约简率及分类准确率(wine)

Fig. 1 Attribute reduction rate and classification accuracy corresponding to different λ values (wine)

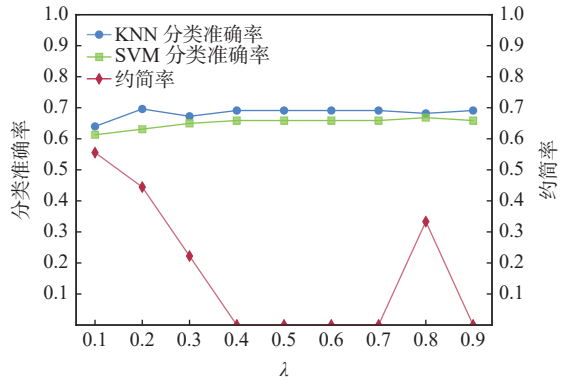

图2 不同λ值对应的属性约简率及分类准确率(glass)

Fig. 2 Attribute reduction rate and classification accuracy corresponding to different λ values (glass)

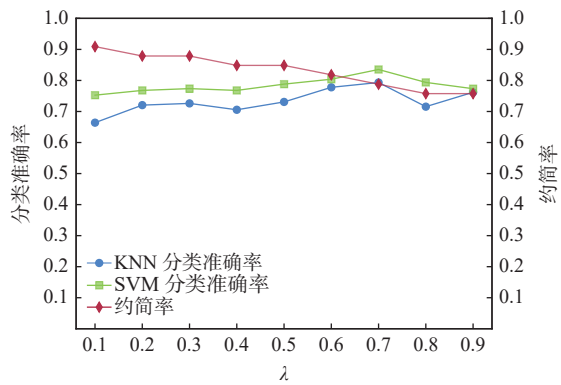

图3 不同λ值对应的属性约简率及分类准确率(wpbc)

Fig. 3 Attribute reduction rate and classification accuracy corresponding to different λ values (wpbc)

3.2 结果分析

使用KNN[25]和SVM[26]分类算法，用基于邻域粗糙集(NRS)、基于信息熵[27](Entropy)的属性约简算法与本文提出的属性约简算法进行对比分析，且

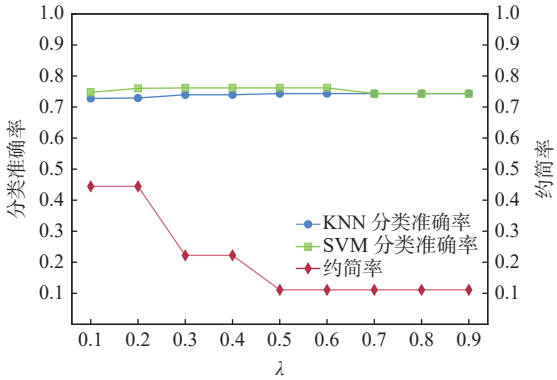

图 4　不同 λ 值对应的属性约简率及分类准确率(diab)

Fig. 4　Attribute reduction rate and classification accuracy corresponding to different λ values (diab)

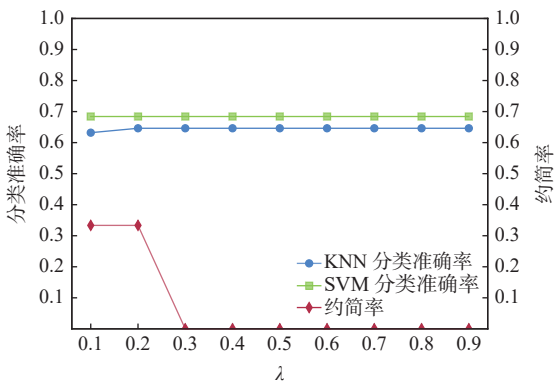

图 5　不同 λ 值对应的属性约简率及分类准确率(bupa)

Fig. 5　Attribute reduction rate and classification accuracy corresponding to different λ values (bupa)

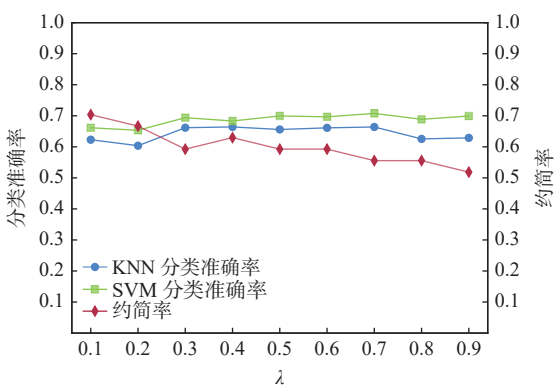

图 6　不同 λ 值对应的属性约简率及分类准确率(horse)

Fig. 6　Attribute reduction rate and classification accuracy corresponding to different λ values (horse)

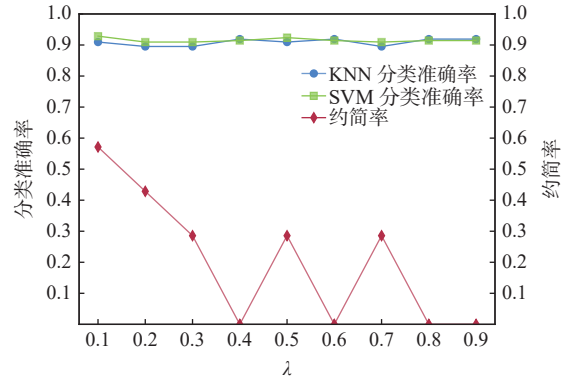

图 7　不同 λ 值对应的属性约简率及分类准确率(seeds)

Fig. 7　Attribute reduction rate and classification accuracy corresponding to different λ values (seeds)

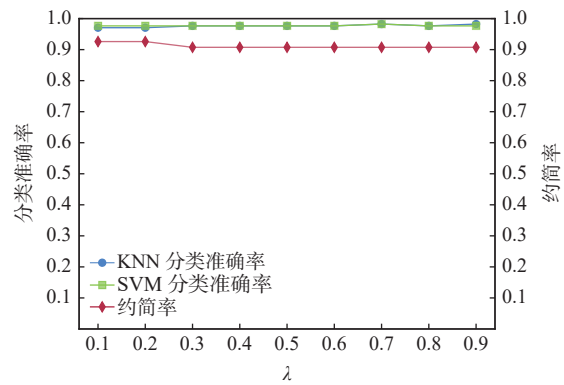

图 8　不同 λ 值对应的属性约简率及分类准确率(divorce)

Fig. 8　Attribute reduction rate and classification accuracy corresponding to different λ values(divorce)

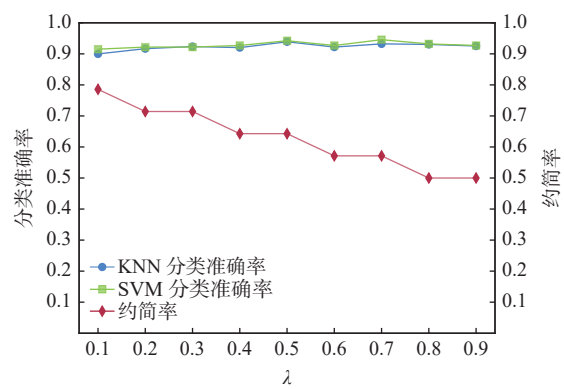

图 9　不同 λ 值对应的属性约简率及分类准确率(hcv)

Fig. 9　Attribute reduction rate and classification accuracy corresponding to different λ values (hcv)

将分类准确率和所选属性个数作为评价算法有效性指标。

表 6 为原始数据集和 3 种算法约简后数据集分别使用 2 种分类算法得到的分类准确率, 其中, 粗体突出不同约简算法中的最佳性能。表 6 的所有数据集(除 seeds 和 diab 数据集外)中, 属性约简后的数据集做分类预测实验时, 实验效果更好, 说明数据集中有冗余属性, 删除这些属性并不会削弱信息系统的分类性能。10 个数据集使用 KNN 算法和 SVM 算法进行实验, 其中, 原始数据集实验得到的平均分类准确率分别为 77.75% 和 79.45%, 用 3 种算法约简后数据集实验得到的平均分类准确率分别为 81.03%、80.70%、81.13% 和 81.83%、81.99%、82.94%。与原始数据得到的结果相比可发现, 使用 KNN 分类算法进行实验, 用 3 种约简算法后得到的分类准确率分别提高了 3.28%、2.95% 和 3.38%; 使

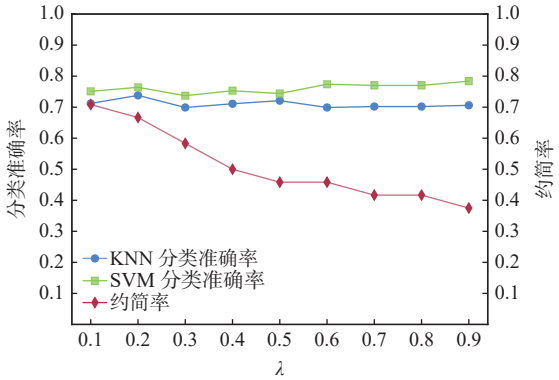

图 10 不同 λ 值对应的属性约简率及分类准确率(german)

Fig.10Attribute reduction rate and classification accuracy corresponding to different λ values (german)

表 6 分类准确率实验结果

Table 6 Experimental results of classification accuracy %

数据集	算法	Original	NRS	Entropy	本文算法
wine	KNN	94.92	97.22	96.11	95.52
	SVM	97.72	98.33	98.30	97.75
glass	KNN	66.15	71.95	70.06	69.61
	SVM	67.69	65.89	65.45	68.82
wpbc	KNN	70.63	76.89	76.79	79.34
	SVM	74.77	77.95	80.47	80.39
diab	KNN	74.35	74.74	74.34	74.94
	SVM	76.83	76.57	76.17	76.17
bupa	KNN	58.55	64.61	63.67	64.62
	SVM	57.68	66.41	67.41	68.41
horse	KNN	66.03	68.97	69.15	69.41
	SVM	70.27	72.15	71.57	73.77
seeds	KNN	93.65	91.90	91.90	91.90
	SVM	92.06	92.86	91.43	92.86
divorce	KNN	94.11	97.65	98.23	98.24
	SVM	94.11	98.14	98.23	98.24
hcv	KNN	88.70	93.55	93.79	93.89
	SVM	89.83	93.04	94.06	94.57
german	KNN	70.40	72.80	73.00	73.80
	SVM	73.50	77.00	76.80	78.40

用 SVM 分类算法进行实验,性能分别提高了 2.38%、2.54% 和 3.49%,其中,本文算法提高幅度最大。

表 6 中,50% 以上的数据集使用本文算法得到的结果最优,其中,divorce 数据集中,仅使用了原始数据集不到 1/10 的属性,分类准确率提高了 4%。有 4 个数据集使用传统邻域粗糙集时分类准确率最高,但与使用本文算法所得到的结果差距不是很大,甚至本文算法在得到属性个数最少的同时,分类效果更好。从分类准确率分析,本文算法在大多数情况下表现最好,且在 2 种分类器下,平均分类准确率均为最佳。

属性约简是找到一个信息量小的属性子集,所选属性子集的大小如表 7 所示。与其他约简算法相比,本文算法选择的属性子集最小。实验表明,本文算法可实现剔除冗余属性,得到更加精准的约简属性子集。

表 7 约简后的属性个数

Table 7 Number of attributes after reduction

数据集	算法	Original	NRS	Entropy	本文算法
wine	KNN	13	6	5	4
	SVM	13	6	6	6
glass	KNN	9	7	6	5
	SVM	9	8	8	6
wpbc	KNN	33	9	7	7
	SVM	33	8	7	6
diab	KNN	8	7	8	8
	SVM	8	4	4	7
bupa	KNN	6	6	6	4
	SVM	6	6	6	4
horse	KNN	27	10	12	10
	SVM	27	8	13	12
seeds	KNN	7	7	7	7
	SVM	7	5	7	3
divorce	KNN	54	3	5	5
	SVM	54	8	5	5
hcv	KNN	14	7	5	5
	SVM	14	4	5	6
german	KNN	20	9	10	8
	SVM	20	14	15	14

综上,用 Critic 赋权法加权邻域粗糙集属性约简算法对 diab 数据集的约简效果不太好,减少属性的同时降低了分类准确率,原因可能是因为 diab 数据属性较少,样本较多,经过属性约简后剔除了重要属性,导致分类准确率不理想。但该约简算法在大多数的数据集上有较好的效果,既降低数据维度,又保证约简后数据集的分类准确率,相对于经典邻域粗糙集模型有优势。

4 结 论

传统的基于邻域粗糙集的属性约简在计算邻域关系时没有考虑属性的权重问题,本文提出了一种基于 Critic 赋权法加权邻域粗糙集的属性约简算法。

1) 用 Critic 赋权法为条件属性分配不同的权重,使得构建的加权邻域粗糙集的粒度更细。

2) 可实现剔除冗余属性,得到更好的约简属性集,具有高效的处理能力。例如,在 divorce 数据集中,仅使用 5 个属性,分类准确率却得到 98% 以上。

3) 具有较高的约简率,在降维方面具有明显的优势。对于大维度且冗余的数据集,该算法可快速

选择划分能力较强的属性, 获得最精简的属性集, 进一步提高分类性能。

但本文算法还有些局限性, 在计算权重的时候进行累加操作, 增加了运行时间, 因此, 在后续研究中可以考虑并行计算等方法, 解决此类问题。

参考文献（References）

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

[2] 苑红星, 卓雪雪, 竺德, 等. 基于矩阵的混合型邻域决策粗糙集增量式更新算法[J]. 控制与决策, 2022, 37(6): 1621-1631.
YUAN H X, ZHUO X X, ZHU D, et al. Rough set incremental update algorithm for mixed neighborhood decision based on matrix [J]. Control and Decision, 2022, 37(6): 1621-1631(in Chinese).

[3] YAO Y Y. Relational interpretations of neighborhood operators and rough set approximation operators[J]. Information Sciences, 1998, 111(1-4): 239-259.

[4] YUAN Z, CHEN H M, XIE P, et al. Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions[J]. Applied Soft Computing, 2021, 107: 107353.

[5] 饶梦, 苗夺谦, 罗晟. 一种粗糙不确定的图像分割方法[J]. 计算机科学, 2020, 47(2): 72-75.
RAO M, MIAO D Q, LUO S. Rough uncertain image segmentation method[J]. Computer Science, 2020, 47(2): 72-75(in Chinese).

[6] WANG G Q, LI T R, ZHANG P F, et al. Double-local rough sets for efficient data mining[J]. Information Sciences, 2021, 571: 475-498.

[7] 冀俊忠, 龙腾, 杨翠翠. 基于邻域决策粗糙集的脑功能连接生物标记物识别[J]. 控制与决策, 2023, 38(4): 1092-1100.
JI J Z, LONG T, YANG C C. Identifying brain functional connectivity biomarkers based on neighborhood decision rough set[J]. Control and Decision, 2023, 38(4): 1092-1100(in Chinese).

[8] KONDO M. On topologies defined by neighbourhood operators of approximation spaces[J]. International Journal of Approximate Reasoning, 2021, 137: 137-145.

[9] 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.

[10] CHEN H M, LI T R, FAN X, et al. Feature selection for imbalanced data based on neighborhood rough sets[J]. Information Sciences, 2019, 483: 1-20.

[11] CHEN H M, LI T R, CAI Y, et al. Parallel attribute reduction in dominance-based neighborhood rough set[J]. Information Sciences, 2016, 373: 351-368.

[12] 邓志轩, 郑忠龙, 邓大勇. F-邻域粗糙集及其约简[J]. 自动化学报, 2021, 47(3): 695-705.
DENG Z X, ZHENG Z L, DENG D Y. F-neighborhood rough sets and its reduction[J]. Acta Automatica Sinica, 2021, 47(3): 695-705 (in Chinese).

[13] WANG C Z, SHI Y P, FAN X D, et al. Attribute reduction based on k-nearest neighborhood rough sets[J]. International Journal of Approximate Reasoning, 2019, 106: 18-31.

[14] GUO Y T, TSANG E C C, XU W H, et al. Adaptive weighted generalized multi-granulation interval-valued decision-theoretic rough sets[J]. Knowledge-Based Systems, 2020, 187: 104804.

[15] TSANG E C C, HU Q H, CHEN D G. Feature and instance reduction for PNN classifiers based on fuzzy rough sets[J]. International Journal of Machine Learning and Cybernetics, 2016, 7(1): 1-11.

[16] VLUYMANS S, MAC PARTHALÁIN N, CORNELIS C, et al. Weight selection strategies for ordered weighted average based fuzzy rough sets[J]. Information Sciences, 2019, 501: 155-171.

[17] KUMAR R, SINGH S, BILGA P S, et al. Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: Acritical review[J]. Journal of Materials Research and Technology, 2021, 10: 1471-1492.

[18] YOU C S, YANG S Y. A simple and effective multi-focus image fusion method based on local standard deviations enhanced by the guided filter[J]. Displays, 2022, 72: 102146.

[19] 徐宇恒, 程嗣怡, 庞梦洋. 基于 CRITIC-TOPSIS 的动态辐射源威胁评估[J]. 北京航空航天大学学报, 2020, 46(11): 2168-2175.
XU Y H, CHENG S Y, PANG M Y. Dynamic radiator threat assessment based on CRITIC-TOPSIS[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(11): 2168-2175 (in Chinese).

[20] 吴希. 三种权重赋权法的比较分析[J]. 中国集体经济, 2016(34): 73-74.
WU X. Comparative analysis of three weight empowerment methods[J]. China's Collective Economy, 2016(34): 73-74(in Chinese).

[21] ZIELOSKO B, STAŃCZYK U. Condition attributes, properties of decision rules, and discretisation: Analysis of relations and dependencies[J]. Procedia Computer Science, 2021, 192: 3922-3931.

[22] BHADRA D, DHAR N R, SALAM M A. Sensitivity analysis of the integrated AHP-TOPSIS and CRITIC-TOPSIS method for selection of the natural fiber[J]. Materials Today: Proceedings, 2022, 56: 2618-2629.

[23] 李冬. 基于邻域粗糙集的属性约简算法研究及应用[D]. 成都: 成都信息工程大学, 2020.
LI D. Research and application of attribute reduction algorithm based on neighborhood rough set[D]. Chengdu: Chengdu University of Information Technology, 2020(in Chinese).

[24] 吴尚智, 王旭文, 王志宁, 等. 利用粗糙集和支持向量机的银行借贷风险预测模型[J]. 成都理工大学学报 (自然科学版), 2022, 49(2): 249-256.
WU S Z, WANG X W, WANG Z N, et al. Prediction model of bank lending risk using rough set and support vector machine[J]. Journal of Chengdu University of Technology (Science & Technology Edition), 2022, 49(2): 249-256(in Chinese).

[25] XIONG L, YAO Y. Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm[J]. Building and Environment, 2021, 202: 108026.

[26] AKRAM-ALI-HAMMOURI Z, FERNÁNDEZ-DELGADO M, ALBTOUSH A, et al. Ideal kernel tuning: Fast and scalable selection of the radial basis kernel spread for support vector classification[J]. Neurocomputing, 2022, 489: 1-8.

[27] HUANG Y Y, GUO K J, YI X W, et al. Matrix representation of the conditional entropy for incremental feature selection on multi-source data[J]. Information Sciences, 2022, 591: 263-286.

An attribute reduction algorithm of weighting neighborhood rough sets with Critic method

WU Shangzhi[1,*], REN Yixuan[2], GE Shuyue[1], WANG Litai[1], WANG Zhining[1]

(1. College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China;
2. Wu'an Tax Bureau, State Taxation Administration, Handan 056300, China)

Abstract: Compared with classical rough sets, neighborhood rough sets can process non-discrete and high-dimensional data, and get simplified data without reducing the ability of data processing. An attribute reduction approach of weighting neighborhood rough sets using the Critic method is proposed, aiming at the problem that every attribute in neighborhood rough sets has the same weight and every attribute has varied influence on decision making. Firstly, the Critic method is used to weigh the conditional attributes, the weighted distance function is introduced to calculate the neighborhood relationship, and then the weighted neighborhood relationship is obtained. Secondly, the weighted neighborhood rough sets are constructed, the attribute dependency and importance are used to evaluate the importance of the subset, the isometric search is used to find the best threshold, attribute reduction is carried out, and the optimal attribute subset is found.Finally, the experimental verification is carried out with 10 data sets in the UCI database, and the performance of the attribute reduction algorithm is compared with that of traditional neighborhood rough sets.The outcomes of the experiment demonstrate that the algorithm is able to guarantee the classification accuracy of the reduced data in addition to obtaining the minimum attribute reduction set.It has effectiveness and practical application value.

Keywords: Critic method; neighborhood rough set; attribute reduction; weighted neighborhood relationship; attribute dependence degree

Received: 2022-12-15; **Accepted**: 2023-05-05; **Published Online**: 2023-05-15 09:48

URL: link.cnki.net/urlid/11.2625.V.20230512.1629.002

Foundation items: National Natural Science Foundation of China (12161082, 61861039)

* **Corresponding author**. E-mail: wusz@nwnu.edu.cn