文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

基于邻域决策系统的三因素融合属性约简

吴慧佳
中国民用航空飞行学院 理学院
DOI:10.32629/acair.v3i4.17925

[摘　要] 针对邻域粗糙集中属性重要度相同的多候选属性问题,本文提出了基于邻域决策系统的三因素融合属性约简方法。在邻域决策系统中,在属性依赖度的基础上,引入邻域粒距离以度量属性的分类差异性,结合交互信息来衡量条件属性间的组合互补性,设计了一个三因素融合属性约简算法。通过使用6个UCI公开数据集进行实验分析,结果表明所提算法能有效提升约简效率和约简集的分类准确率。
[关键词] 属性约简；邻域决策系统；属性依赖度；邻域粒距离；交互信息
中图分类号：TP18　文献标识码：A

Attribute Reduction for Three-Factor Fusion Based on Neighborhood Decision Systems

Huijia Wu
School of Science, Civil Aviation Flight University of China

[Abstract] To address the problem of multiple candidate attributes with identical importance in neighborhood rough sets, this paper proposes an attribute reduction method for three-factor fusion based on neighborhood decision systems. Within the neighborhood decision system, neighborhood granular distance is introduced based on attribute dependency to measure the classification diversity of attributes. Combined with interaction information to assess the combinatorial complementarity among conditional attributes, a three-factor fusion attribute reduction algorithm is designed. Experimental analysis using six UCI public datasets demonstrates that the proposed algorithm effectively enhances reduction efficiency and classification accuracy of the reduced attribute set.
[Key words] attribute reduction; neighborhood decision systems; attribute dependency; neighborhood granularity distance; interaction information

引言

作为粗糙集研究的核心内容,属性约简[1]旨在保持数据集分类能力的前提下,从高维数据集中筛选出重要属性,降低目标任务描述的数据集维度,提高挖掘任务的计算效率,减少过拟合现象。属性约简已在机器学习[2]、生物信息学[3]和金融风控[4]等研究中被广泛应用,提高了模型性能、知识发现的准确性和可靠性。

基于邻域粗糙集的属性约简算法常采用前向启发式方法进行设计,一般会分为三类：基于信息论、差别矩阵和属性依赖度[5]。其中,基于属性依赖度的约简方法,是将重要的属性逐步加入候选约简集,直到满足约简终止条件时停止。其计算方式简单,但大部分属性约简算法都只考虑将条件属性相对于决策属性的重要度或依赖度大的属性添加到约简中,没有考虑对不确定性度量相等的不同属性子集情形的处理,因此,约简结果可能仍然含有冗余属性。为此,从统计学角度出发,姚晟等引入秩相关系数来度量条件属性间的相关性[6]。翟俊海等从信息论角度引入互信息来度量条件属性之间的相关程度,对重要度进行改进,提出了基于最小相关性和最大依赖度准则的属性约简算法[7]。然而,当其相关性判定阈值过大或过小时,都无法得到最优约简集。为此,文献[8]利用属性互信息均值作为相关性阈值,但是,它又忽略了条件属性与决策属性间内在的相关性,影响了约简效率和分类精度。由此,文献[9]引入交互信息概念,在度量属性相关性时考虑决策属性所提供的信息。

为此,针对属性重要度相同的多候选属性问题,本文在前向启发式搜索策略的基础上,基于属性依赖度,利用邻域粒距离及交互信息对属性重要度进行优化,进而构建了一个基于邻域决策系统的三因素融合属性约简算法。

1 相关知识

对于邻域信息系统[10] $NIS=(U,A,V,f,\delta)$，$U=\{x_1,x_2,x_3,\cdots,x_n\}$ 是非空有限对象集，$V=\bigcup_{a_i\in A}V_{a_i}$ 是非空有限属性集,是所有对象的值域 V_{a_i} 的并，$a_i\in A$，有 $f(x,a_i)\in$

V_{a_i} 。当由条件属性子集 $C=\{a_1,a_2,\cdots,a_m\}$ 和决策属性子集 D 构成时且 C∩ D=∅ NIS ，为邻域决策信息系统，记 $NDS=(U,C\cup D,V,f,\delta)$ 。记决策属性关于论域 U 的划分 $U/D=\{D_1,D_2,\ldots,D_r\}$ 。

定义1(混合欧氏重叠距离函数[11])在 $NDS=(U,C\cup D,V,f,\delta)$ 中 $B\subseteq C$, $B=\{a_1,a_2,\cdots,a_k\}$，且 $k\le|C|$，对于 $\forall x,y\in U$ ，距离函数定义为：

$$HEOM_B(x,y)=\sqrt{\sum_{a_i\in B}d_{a_i}(x,y)^2} \tag{1}$$

其中，$d_{a_i}(x,y)$ 是对象 x,y 在 a_i 下的距离，$d_{a_i}(x,y)$ 定义为：

$$d_{a_i}(x,y)=\begin{cases}1, & \text{对象}x\text{或}y\text{在}a_i\text{下未知}\\ 0, & \text{若}a_i\text{是符号型且}f(x,a_i)=f(y,a_i)\\ 1, & \text{若}a_i\text{是符号型且}f(x,a_i)\neq f(y,a_i)\\ f(x,a_i)-f(y,a_i), & \text{若}a_i\text{是数值型}\end{cases}$$

基于 $HEOM_B(x,y)$，以自适应邻域半径为基础，对论域 U 进行粒化，从而形成适用于混合数据的邻域类及邻域关系，构建以对象邻域为基础的邻域信息系统。

定义2(δ- 邻域类[12])对 B，δ- $N_B^\delta(x_i)$ 基于 B 的 δ- 邻域类 $N_B^\delta(x_i)$ 定义为：

$$N_B^\delta(x_i)=\{y\in U\mid d_B(x_i,y)\le\delta\} \tag{2}$$

其中，δ 是邻域半径且 $\delta=\sum_{a_i\in B}\delta_{a_i}/|B|$，$\delta_{a_i}$ 是在 a_i 下的邻域半径，定义为：

$$\delta_{a_i}=\begin{cases}0, & \text{如果}a_i\text{是符号型属性}\\ \lambda*\mathrm{std}(a_i), & \text{如果}a_i\text{是数值型属性}\end{cases}$$

其中，λ 是邻域半径调节参数且 $\lambda>0$ 。

公式(2)中 $N_B^\delta(x_i)$ 聚集了对象 B 的描述下，距离不超过邻域半径的数据对象，又被称为邻域粒。显然，$N_B^\delta(x_i)\neq\varnothing$ 。

定义3(邻域关系[12])给定 $NDS=(U,C\cup D,V,f,\delta)$，$B\subseteq C$，$B$ 下邻域关系如下：

$$NR^\delta(B)=\{(x,y)\in U\times U\mid d_B(x,y)\le\delta\} \tag{3}$$

显然，U 中对象邻域关系符合自反性和对称性，但不满足传递性。特别地，当 $\delta=0$ 时，邻域关系退化为等价关系。

定义4(邻域上下近似集[12])给定 $NDS=(U,C\cup D,V,f,\delta)$ $B\subseteq C$，$\forall X\subseteq U$，那么 B 关于 X 的邻域上近似集和下近似集分别为：

$$\overline{NR}_B^\delta(X)=\{x_i\in U\mid N_B^\delta(x_i)\cap X\neq\varnothing\} \tag{4}$$

$$\underline{NR}_B^\delta(X)=\{x_i\in U\mid N_B^\delta(x_i)\subseteq X\} \tag{5}$$

定义5(属性正域依赖度[12])给定 $NDS=(U,C\cup D,V,f,\delta)$，$\forall B\subseteq C$，$U/D=\{D_1,D_2,\cdots,D_r\}$，$D$ 关于 B 的分类正域定义为：

$$POS_B^\delta(D)=\underline{NR}_B^\delta(D) \tag{6}$$

故 D 关于 B 的属性正域依赖度定义为：

$$\gamma^\delta(B,D)=\frac{|POS_B^\delta(D)|}{|U|} \tag{7}$$

可见，属性正域依赖度表示 U 在条件属性子集 B 下，能正确划分到决策属性子集 D 的 U/D 的论域准确率，即度量了条件属性与决策属性间的相关性。易知，$\gamma^\delta(B,D)$ 值越大，D 越依赖于 B，表明 B 的分类能力越强。

当数据集中存在噪声数据时，使用标准严苛的经典属性约简可能会导致模型拟合噪声数据。降低算法对数据集的整体分类效果。为了降低这类噪声数据带来的属性约简负面影响，本文采取近似属性约简[13]，即通过牺牲较小的约简精度，有效地降低这类噪声带来的影响，最终提升约简算法的约简效率。

定义6(近似属性约简[13])对 $NDS=(U,C\cup D,V,f,\delta)$，$\delta\ge0$，假设近似约简的调参值 $\kappa\in(0,1)$，$R\subseteq C$，称 R 是 C 的近似属性约简当且仅当以下条件成立：

(1) $(1-\kappa)\gamma^\delta(C,D)\le\gamma^\delta(R,D)$ ；

(2) $\gamma^\delta(R-\{a\},D)<(1-\kappa)\gamma^\delta(C,D)$ ，.

定义7(邻域互信息[14])给定 $NDS=(U,C\cup D,V,f,\delta)$，$B_1,B_2\subseteq C$，$\forall x_i\in U$，则 B_1 和 B_2 的邻域互信息熵为：

$$NE_\delta(B_1;B_2)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|N_{B_1}^\delta(x_i)|\cdot|N_{B_2}^\delta(x_i)|}{|U|\cdot|N_{B_1\cup B_2}(x_i)|} \tag{8}$$

定义8(邻域条件互信息[15])给定 $NDS=(U,C\cup D,V,f,\delta)$，

$B_1, B_2, B_3 \subseteq C$，$\forall x_i \in U$，则定义 B_3 下 B_1 和 B_2 的邻域条件互信息熵为：

$$NE_\delta(B_1;B_2 \mid B_3) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|N^\delta_{B_1\cup B_3}(x_i)|\cdot|N^\delta_{B_2\cup B_3}(x_i)|}{|N^\delta_{B_3}(x_i)|\cdot|N^\delta_{B_1\cup B_2\cup B_3}(x_i)|} \quad (9)$$

定义9(邻域交互信息[15])给定 $NDS=(U,C\bigcup D,V,f,\delta)$，$B_1, B_2, B_3 \subseteq C$，$\forall x_i \in U$，则定义 B_3、B_1 和 B_2 的邻域交互信息熵为：

$$NE_\delta(B_1;B_2;B_3) = NE_\delta(B_1;B_2 \mid B_3) - NE_\delta(B_1;B_2) \quad (10)$$

2 基于邻域决策系统的三因素融合属性约简算法

表1 样表

U	a_1	a_2	a_3	d
x_1	0.7	Y	H	Y
x_2	0.8	N	M	Y
x_3	0.9	Y	M	Y
x_4	0.7	Y	M	N
x_5	0.7	Y	M	N
x_6	0.1	N	L	N

表2 在单属性下邻域粒化结果表（$\delta=0.1$）

U	a_1	a_2	a_3	d
x_1	$\{x_1,x_2,x_4,x_5\}$	$\{x_1,x_3,x_4,x_5\}$	$\{x_1\}$	$\{x_1,x_2,x_3\}$
x_2	$\{x_1,x_2,x_3,x_4,x_5\}$	$\{x_2,x_6\}$	$\{x_2,x_3,x_4,x_5\}$	$\{x_1,x_2,x_3\}$
x_3	$\{x_2,x_3\}$	$\{x_1,x_3,x_4,x_5\}$	$\{x_2,x_3,x_4,x_5\}$	$\{x_1,x_2,x_3\}$
x_4	$\{x_1,x_2,x_4,x_5\}$	$\{x_1,x_3,x_4,x_5\}$	$\{x_2,x_3,x_4,x_5\}$	$\{x_4,x_5,x_6\}$
x_5	$\{x_1,x_2,x_4,x_5\}$	$\{x_1,x_3,x_4,x_5\}$	$\{x_2,x_3,x_4,x_5\}$	$\{x_4,x_5,x_6\}$
x_6	$\{x_6\}$	$\{x_2,x_6\}$	$\{x_6\}$	$\{x_4,x_5,x_6\}$

对于基于邻域的前向启发式约简中的多属性候选问题，利用邻域粗糙集对论域进行粒化，针对于表1的邻域粒化结果如表2所示。此时，属性 a_1 和 a_3 的正域分别为 $\{x_3,x_6\}$ 和 $\{x_1,x_6\}$，其属性正域依赖度均为 $\gamma^\delta(a_1,D)=\gamma^\delta(a_3,D)\approx 0.33$，它只对属性进行了基于统计量的属性正域依赖度评价，当统计量相同时，要合理选择 a_1 或 a_3 进入约简集就变得困难。如果能融合其他辅助因素，如不同对象邻域对属性分类能力的影响，将有助于合理、准确地选出候选约简属性。

为了解决上述问题，下面将考虑从数据对象邻域和条件属性的关系出发，改进基于属性正域依赖度的属性重要度评价方式。

2.1条件属性的分类能力差异性

基于邻域粗糙集模型属性约简，对具有相同属性重要度的条件属性子集，其内在知识结构不一定相同，导致分类能力可能不同。如表1，a_1 和 a_3 的属性正域依赖度相同，其知识结构为 $U/\{a_1\}=\{\{x_1,x_2,x_4,x_5\},\{x_1,x_2,x_3,x_4,x_5\},\{x_2,\ x_3\},\{x_6\}\}$ 和 $U/\{a_3\}=\{\{x_1\},\ \{x_2,x_3,x_4,x_5\},\{x_6\}\}$。显然，两个属性的知识结构有差异。如果能刻画这种知识结构差异，对约简属性选择是有益的。实际上，可以引入邻域粒距离[16]来表示一个知识结构的整体表示能力差异。

定义10(R-C邻域粒距离)给定 $NDS=(U,C\bigcup D,V,f,\delta)$ $R\subseteq C$，是条件属性子集，R 到 C 的邻域粒距离可以定义为：

$$NKD^\delta(R,C) = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|N^\delta_R(x_i)\oplus N^\delta_C(x_i)|}{|U|} \quad (11)$$

其中，$N^\delta_R(x_i)\oplus N^\delta_C(x_i)=N^\delta_R(x_i)\bigcup N^\delta_C(x_i)-N^\delta_R(x_i)\bigcap N^\delta_C(x_i)$。

显然，$0\le NKD^\delta(R,C)\le 1$。$NKD^\delta(R,C)$ 中两种邻域粒间交集越小，交集基数越小，而并集基与分母不变，分子越大，整体的值就越大。因此，$NKD^\delta(R,C)$ 值越大，表示 R 与 C 的邻域粒或知识结构差异越大。反之，$NKD^\delta(R,C)$ 值越小，R 与 C 的两种邻域粒结构差异就越小，进一步，R 与 C 的知识结构差异越小，由 R 与 C 重构的信息系统的相似度就越高，也说明 R 与 C 的分类能力越接近。因此，R-C邻域粒距离概念可以用于度量条件属性子集的分类能力大小。

例如，由表1和表2知，$NKD^\delta(a_1,C)=0.39$，$NKD^\delta(a_3,C)=0.33$，表明 a_3 到 C 的邻域粒距离比 a_1 小，即 a_3 的知识结构比 a_1 更接近原系统。因此，可以认为，属性 a_3 的分类能力比属性 a_1 强。

2.2条件属性间的组合互补性

对于分类能力描述，在高维数据空间中，可以利用各分类属性所提供的分类信息的互补性来补充刻画属性分类能力。这里的互补性是指，两个属性共同提供的分类信息的共性特征和差异特征的相互支撑性。随着条件属性数量增加，若只孤立地考虑单属性的冗余性和相关性，会丢失一些重要的分类信息，由此获得的约简的分类能力会有损失。

设表3的邻域信息系统如表4所示。其中，$U/D=\{\{x_1,x_2\},\{x_3\}\}$，由表3和表4知，属性 a_1 和 a_2 的邻域互信息熵为：$NE_\delta(a_1;D)=NE_\delta(a_2;D)=0.2516$。从表征分类能力上讲，单个的 a_1 和 a_2 有可能被误判为分类能力不强的冗余属性。但是，a_1 和 a_2 的组合邻域互信息熵 $NE_\delta(\{a_1,a_2\};D)=0.9183>NE_\delta(a_1;D)+NE_\delta(a_2;D)$，表明属性集的组合可以提供更多的分类信息，因此，$a_1$ 和 a_2 并不是真正的冗余属性。可见，在高维数据空间中，属性约简可以

文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

综合考虑属性间的互补性信息带来的组合分类能力评估的准确性。为此,可用平均交互信息[9]度量属性组合互补性。

表3 针对属性间的互补性的实例样本数据表

U	a_1	a_2	D
x_1	Y	N	1
x_2	N	Y	1
x_3	Y	Y	2

表4 对表3的邻域粒化结果表

U	a_1	a_2
x_1	$\{x_1,x_3\}$	$\{x_1\}$
x_2	$\{x_2\}$	$\{x_2,x_3\}$
x_3	$\{x_1,x_3\}$	$\{x_2,x_3\}$

定义11(平均交互信息熵)给定 $NDS=(U,C\cup D,V,f,\delta)$，$R$ 是一个条件属性子集，$a\in C-R$，那么，a 、R 和 D 之间的平均交互信息熵可以定义为：

$$I(a;R;D)=\frac{1}{|R|}\sum_{a_i\in R}NE_\delta(a;a_i;D) \tag{12}$$

$I(a;R;D)>0$，表示属性 a 和 R 结合时提供的分类信息量大于两者单独提供的分类信息量；$I(a;R;D)=0$，表示 a 和 R 所能提供的分类信息量相互独立；$I(a;R;D)<0$，表示属性 a 和 R 结合时所提供的分类信息量小于两者单独提供的分类信息量,说明 a 和 R 之间有冗余的分类信息。

2.3三因素融合属性重要度

为了能在基于属性正域依赖度相同的多候选属性集中选出分类能力更强的分类属性,下面给出了一种将R-C邻域粒距离、交互信息与属性正域依赖度相结合的融合型属性重要度评价方法。

定义12(三因素融合属性重要度)给定 $NDS=(U,C\cup D,V,f,\delta)$，$R$ 是条件属性集的一个约简集，$a\in C-R$，那么，a 在 R 下的三因素融合属性重要度可以定义为：

$$DISIG(a,R,D)=\gamma^\delta(R\cup\{a\},D)-NKD^\delta(R\cup\{a\},C)+\beta\cdot I(a;R;D) \tag{13}$$

其中，β 是平衡因子,用于平衡条件属性与决策属性的相关性和条件属性间的交互性的影响,定义为 $\beta=|R|/(2|C|)$ 。

显然，$DISIG(a,R,D)$ 值越大,属性 a 越重要,即 a 的分类能力越强。分析可知,属性正域依赖度与邻域粒距离变化关系是反向的,依赖度越大,邻域粒距离越小。交互信息与依赖度和邻域粒距离变化无关,且数值比较小,因此,在属性重要度中只起到调节作用。

3 基于邻域决策系统的三因素融合属性约简算法

针对经典邻域粗糙集模型下属性重要度相同的多候选属性问题,根据改进的三因素融合属性重要度可以克服此问题,并以近似约简为基础,提升了邻域粗糙集模型的约简效率。基于邻域决策系统的三因素融合属性约简算法(Attribute Reduction for Three-Factor Fusion Based on Neighborhood Decision Systems,ARTFNDS)过程如算法1所示。

算法1基于邻域决策系统的三因素融合属性约简算法(ARTFNDS)

输入：$NDS=(U,C\cup D,V,f,\delta)$，以及邻域半径调节参数 λ
输出：约简集 R

(1) 初始化：约简子集 $R=\varnothing$，$\gamma^\delta(R,D)=0$；
(2) 先属性集 C 下的对象邻域,再计算 $\gamma^\delta(C,D)$；
(3) *while* $\gamma^\delta(R,D)/(\gamma^\delta(C,D)\le(1-\kappa)$
(4) *if* $R=\varnothing$
(5) 设 $I(a_i;R;D)=0$，计算 $\gamma^\delta(R\cup\{a_i\},D)$ 和 $NKD^\delta(R\cup\{a_i\},C)$；
(6) *else* $\forall a_i\in C-R$，计算 $\gamma^\delta(R\cup\{a_i\},D)$、$NKD^\delta(R\cup\{a_i\},C)$ 和 $I(a_i;R;D)$；
(7) *end if*
(8) 计算 $\forall a_i\in C-R$ 的属性重要度 $DISIG(a_i,R,D)$；
(9) 选择 $DISIG(a_i,R,D)$ 最大的属性 a_k；
(10) $R=R\cup\{a_k\}$； //将属性 a_k 加入 R 中；
(11) 再计算 $\gamma^\delta(R,D)$；
(12) *end while*
(13) *return R*

4 实验分析

为了更好验证本文提出的约简算法的有效性,将在UCI数据集[17]上进行消融实验和对比实验,并进行详细分析。

4.1实验过程的设计和数据集的选择

为验证基于邻域决策系统的三因素融合属性约简算法(ARTFNDS)的有效性,从属性约简结果、约简率和分类准确率三个方面展开消融实验和对比实验。在消融实验中,为验证算法中属性正域依赖度、属性分类能力差异性和多条件属性间组合互补性对约简结果的影响,假设A算法只考虑属性正域依赖度,B算法融合属性正域依赖度和属性分类能力差异性,C算法融合属性分类能力差异性和条件属性间的组合互补性。为了进一步验证算法的有效性,将与不完备邻域粗糙集的不确定性度量和属性约简算法(ARNME)[18]和基于信息粒度和交互信息的属性约简算法(IG2IAR)[9]进行对比实验分析。其中,ARNME通过在不完备邻域粗糙集上定义邻域混合熵,来代替正域属性重要度。IG2IAR针对相同重要度下的多候选问题,提出基于信息粒度的候选属性选择优化策略,并引入交互信息度量条件属性间的相关性。

实验选取6个数据类型不同的UCI数据集,其基本信息和属性结构信息描述如表5所示。所有算法的计算都假设在标准化后的数据集上展开,以减少各类型属性可能由于度量的量纲不同带来的影响。其中,表5的第 $|C|$ 列括号内数字表示数据集包含的数值型属性数。

表5 ARTFNDS算法的实验数据基本信息表

数据集	缩写	$\|C\|$	$\|U\|$	数据类型	类别数
Credit	Cre	15(7)	690	混合型	2
Heart	Hea	13(6)	270	混合型	2
Lymphography	Lym	18(0)	148	符号型	4
Molecular	Mol	56(0)	106	符号型	2
Vowelsori	Vow	12(12)	1456	数值型	2
Cardioori	Car	21(21)	1831	数值型	2

实验参数设置：假设近似约简参数 $\kappa=0.05$，设置调整参数取值范围为[0.01,1]，可以确保邻域半径取值范围为 $\delta\in[0,2]$。为了验证约简结果的分类能力，以KNN[19]和SVM[20]分类器为标准，采用五折交叉验证法，以分类准确率作为算法分类能力评价指标。其中，在KNN分类器中，k设置为5。

4.2实验结果及分析

表6展示了ARTFNDS与3个消融实验的算法、2个对比算法的约简结果，其中，“/()”中的括号内数字代表约简结果属性数；图1展示了各个算法的约简率；在KNN分类器下，表7中约简结果的分类准确率。表8展示了在SVN分类器下的分类准确率。综合观察图1和表6~8，得出以下分析和结论：

(1)由表6和图1可见，6种算法都能进行有效的属性约简。总体上而言，ARTFNDS约简率优于或者接近于消融实验的算法和对比算法。在Lym和Mol数据集上略低，但结合表7和表8可知，分类准确率是高于其他算法的，因此，从另外一个侧面说明ARTFNDS所得的约简集更能提供数据的分类能力。同时，观察表6的消融实验易知，ARTFNDS与A的约简集共性相对较多，与B的共性较少，而C是共性最少的；因此，可以认为，A是基础，B、C及ARTFNDS是改进。在改进算法中，B和C主要起约简调节作用。

(2)由表7和表8可知，在KNN和SVM分类器上，ARTFNDS所得约简结果的平均分类准确率高于原数据、3个消融实验的算法和2个对比算法。同时，ARTFNDS对6个数据集的约简分类准确率也均达到最佳分类准确率。

表6 ARTFNDS算法的消融实验约简结果对比表

数据集	消融算法			对比算法		ARTFNDS
	A 算法	B 算法	C 算法	ARNME	IG2IAR	
Cre	1,2,3,6,8,9,14,15/(8)	2,3,6,8,9,11,14/(7)	1,4,6,7,9,10,12,13/(8)	1,2,3,6,7,8,9,10,14/(9)	1,2,4,6,7,8,9,10,12,14,15/(11)	1,2,3,6,9,12,14/(7)
Hea	2,3,4,5,7,9,11,12,13/(9)	1,2,3,7,9,11,12,13/(8)	1,2,3,6,7,9,10,11,13/(9)	1,3,4,5,8,10,12/(7)	1,2,3,4,5,7,8,10,12,13/(10)	2,3,7,9,11,12,13/(7)
Lym	2,13,14,15,18/(5)	2,13,14,15,18/(5)	2,3,6,12,14,16,17,18/(8)	1,2,13,14,15,16,18/(7)	2,13,14,15,16,18/(6)	2,3,8,11,14,15,16,18/(8)
Mol	3,15,17,39/(4)	15,17,39/(3)	17,39,51,56/(4)	4,15,17,24,39,54/(6)	3,15,17,39/(4)	15,17,24,39/(4)
Vow	1,2,3,4,7,8,10,11/(8)	1,2,3,4,6,7,8,10,11/(9)	1,2,3,4,7,8,11,12/(8)	1,3,4,5,8,9,11/(7)	1,2,3,4,5,9/(6)	1,2,3,4,7,8,10,12/(8)
Car	2,4,7,8,10,11,17,20/(8)	2,4,7,8,10,12,19/(7)	1,2,4,5,6,7,8,10,11,14,15,16,20/(13)	1,2,4,8,9,12,15,19/(8)	1,2,4,7,8,13,17/(7)	2,4,8,10,14,19,20/(7)

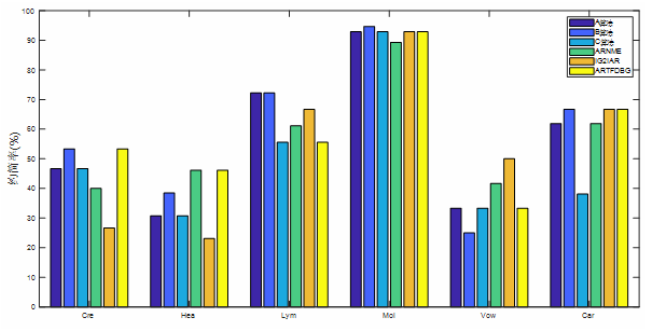

图1 ARTFNDS算法的消融实验和对比实验的约简率对比图

综上，以消融实验而言，A、B和C都能有效约简，但约简率和约简结果的分类准确率有待提升，而融合属性分类能力和多属性组合互补能力的ARTFNDS算法能够使得约简率和约简结果的分类能力得到明显提升。针对对比实验，在KNN和SVM分类器下，ARTFNDS的分类能力优于IG2IAR和ARNME，说明ARTFNDS能有效处理多属性候选难题。

表7 消融实验和对比实验在KNN分类器的分类准确率对比表

数据集	消融算法				对比算法		ARTFNDS (%)
	原始数据(%)	A 算法(%)	B 算法(%)	C 算法(%)	ARNME(%)	IG2IAR(%)	
Cre	81.72	85.40	84.99	81.90	83.08	82.07	86.13
Hea	80.02	82.25	83.28	81.49	82.60	82.59	84.46
Lym	76.34	76.47	76.47	74.92	76.99	77.19	83.11
Mol	74.56	76.39	79.22	66.96	75.27	76.39	80.35
Vow	98.62	98.85	98.13	98.54	98.16	98.49	98.91
Car	97.43	98.79	98.74	98.64	97.87	98.25	98.80
平均	84.78	86.36	86.81	83.74	85.66	85.83	88.63

表8 消融实验和对比实验在SVM分类器的分类准确率对比表

数据集	消融算法				对比算法		ARTFNDS (%)
	原始数据(%)	A 算法(%)	B 算法(%)	C 算法(%)	ARNME(%)	IG2IAR(%)	
Cre	86.65	86.63	86.62	86.61	86.63	86.64	86.72
Hea	82.25	84.10	83.32	81.12	82.62	83.33	85.55
Lym	76.33	75.91	75.91	73.77	76.28	77.19	83.97
Mol	74.74	77.39	79.22	42.55	75.36	77.39	79.35
Vow	98.19	98.13	98.27	97.43	96.70	96.78	98.51
Car	97.76	97.81	97.16	97.87	97.11	95.75	98.25
平均	85.99	86.66	86.75	79.89	85.78	86.18	88.73

5 结束语

在邻域决策系统下，针对多候选属性的问题，本文在属性正域依赖度的基础上，运用邻域粒距离和交互信息概念改进属性重要度，提出了基于邻域决策系统的三因素融合属性约简方法。通过多组消融实验和多个对比算法分析，展示了ARTFNDS算法的约简有效性以及约简结果的分类准确性。但存在邻域半径阈值的自适应性不太好的问题，之后将继续研究构建一个自适应性较强的拓展粗糙集模型。

[基金项目]
校级青年基金(25CAFUC05012)。

[参考文献]

[1]BAUTISTA R,MILLAN M,DIAZ J F.An efficient implementation to calculate relative core and reducts[C].18th International Conference of the North American Fuzzy Information Processing Society-NAFIPS. IEEE,1999:791-794.

[2]CUI S G, LI G S,SANG B B,el.Distance metric learning-based multi-granularity neighborhood rough sets for attribute reduction. Appl[J].Soft Comput,2024:159.

[3]刘富,张潇,侯涛,等.基于粗糙集的基因信号属性约简[J].吉林大学学报(工学版),2015,45(2):624-629.

[4]张宇敬,王柳,齐晓娜,等.基于信息熵的商业银行客户画像属性约简研究[J].河北大学学报(自然科学版),2022,42(1):98-104.

[5]周涛,陆惠玲,任海玲,等.基于粗糙集的属性约简算法综述[J].电子学报,2021,49(7):1439-1449.

[6]姚晟,汪杰,徐风,陈菊.不一致邻域粗糙集的不确定性度量和属性约简[J].小型微型计算机系统,2018,39(4):700-706.

[7]翟俊海,万丽艳,王熙照.最小相关性最大依赖度属性约简[J].计算机科学,2014,41(12):148-150,154.

[8]毛华,赵书峰.最小相关性最大依赖度属性约简的改进算法[J].河北大学学报(自然科学版),2019,39(3):225-229.

[9]张清华,李新太,赵凡,等.基于信息粒度与交互信息的属性约简改进算法[J].闽南师范大学学报(自然科学版),2021,34(2):68-78.

[10]MIAO D,LI D.Rough sets theory algorithms and applications[D].Press of Tsinghua University,2008.

[11]WILSON D R,MARTINEZ T R.Improved heterogeneous distance functions[J].Journal of Artificial Intelligence Research,1997,6(1):1-34.

[12]胡清华,于达仁,谢宗霞.基于邻域粒化和粗糙逼近的数值属性约简[J].软件学报,2008,19(3):640-649.

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

[14]HU Q H,PAN W,AN S,et al.An efficient gene selection technique for cancer recognition based on neighborhood mutual information[J].International Journal of Machine Learning and Cybernetics,2010,1(1-4):63-74.

[15]SUN L,XU J C,Feature selection using mutual information based uncertainty measures for tumor classification[J].Bio-Medical Materials and Engineering,2014,24(1):763-770.

[16]QIAN Y H,LIANG J Y,DANG C Y,et al. Knowledge granulation and knowledge distance in a knowledge base[J].International Journal of Approximate Reasoning,2011,19(2):263-264.

[17]BAY S D.The UCI KDD repository[J].http://kdd.ics.uci.edu.

[18]姚晟,汪杰,徐风,等.不完备邻域粗糙集的不确定性度量和属性约简[J].计算机应用,2018,38(1):97-103.

[19]HANG S,LI X,ZONG M,et al.Efficient kNN classification with different numbers of nearest neighbors[J].IEEE transactions on neural networks and learning systems,2017,29(5):1774-1785.

[20]WANG R,LI W,LI R,et al.Automatic blur type classification via ensemble SVM[J].Signal processing: image communication,2019,71:24-35.

作者简介：

吴慧佳(1998--),女,汉族,四川雅安人,硕士,助教,研究方向：粗糙集与数据挖掘。