控制与决策
Control and Decision

基于改进邻域空间的高维混合数据特征选择算法

张腾飞, 张宇迪, 马福民

引用本文:
张腾飞,张宇迪,马福民. 基于改进邻域空间的高维混合数据特征选择算法[J]. 控制与决策, 2024, 39(3): 929−938.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0789

基于改进邻域空间的高维混合数据特征选择算法

张腾飞[1†], 张宇迪[1], 马福民[2]

(1. 南京邮电大学 自动化学院 人工智能学院，南京 210023；2. 南京财经大学 信息工程学院，南京 210023)

摘　要: 作为数据挖掘领域中一项重要的数据预处理技术,特征选择算法能够有效应对高维数据带来的“维数灾难”问题. 然而,如何对高维的混合数据进行特征选取仍然是当前研究的重点和难点之一. 基于邻域关系的邻域粗糙集模型因其能够处理名词型属性与数值型属性并存的混合数据,已成功应用于混合数据的特征选择. 但是,现有邻域粗糙集对混合数据邻域关系的度量,仍然是基于等价关系的名词型数据划分与基于相似关系的数值型数据划分的简单融合,在利用模型划分的邻域空间和预定义的评价函数对高维混合数据进行特征选取时,适应性较差. 为此,在邻域粗糙集模型的基础上,提出一种改进的邻域空间构造方法,并设计相应的邻域空间度量公式作为判别指标,自适应地调节邻域空间下邻域粒的大小;为了准确地表征高维混合数据邻域空间的判别能力,设计一种考虑边界数据和邻域空间大小的评价函数;在此基础上,提出一种启发式的高维混合数据特征选择算法. 通过UCI标准数据集验证所提出算法的有效性.

关键词: 特征选择；邻域空间；高维混合数据；邻域粗糙集；评价函数

中图分类号: TP18　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0789

引用格式: 张腾飞,张宇迪,马福民. 基于改进邻域空间的高维混合数据特征选择算法[J]. 控制与决策, 2024, 39(3): 929-938.

Improved neighborhood space based feature selection algorithm for high-dimensional mixed data

ZHANG Teng-fei[1†], *ZHANG Yu-di*[1], *MA Fu-min*[2]

(1. College of Automation & College of Artificial Intelligence，Nanjing University of Posts and Telecommunications，Nanjing 210023，China；2. College of Information Engineering，Nanjing University of Finance and Economics，Nanjing 210023，China)

Abstract: As important data preprocessing technology in the field of data mining, the feature selection algorithm can effectively deal with the “curse of dimensionality” caused by high-dimensional data. Nonetheless, how to perform feature selection on high-dimensional mixed data is still one of the focuses and difficulties of current research. Because of competently dealing with mixed data of categorical attributes and numerical attributes coexisting, the neighborhood rough set model has been widely used in feature selection of mixed data in recent years. However, existing measurement of the neighborhood relationship for mixed data still adopts the simple fusion of categorical data partition based on equivalence relationship and numerical data partition based on similarity relationship. When the features of high-dimensional mixed data are selected by the partitioned neighborhood space and predefined evaluation function, the adaptability is poor. Therefore, an improved construction method of neighborhood space is proposed on the basis of the neighborhood rough set model. Considering boundary overlapped data and the size of neighborhood space, an evaluation function is designed to characterize the discrimination ability of neighborhood space. On this basis, a heuristic feature selection algorithm considering high-dimensional mixed data is proposed. The validity and superiority of proposed algorithm are verified by the UCI standard data set.

Keywords: feature selection；neighborhood space；high-dimensional mixed data；neighborhood rough set；evaluation function

收稿日期: 2022-05-08；**录用日期:** 2022-11-09.

基金项目: 国家自然科学基金项目(62073173, 61973151)；江苏省自然科学基金项目(BK20191376, BK20191406).

责任编委: 胡清华.

†通讯作者. E-mail: tfzhang@126.com.

0 引 言

为有效应对高维数据带来的“维数灾难”问题,提升机器学习模型的性能,特征选择算法[1-3]已成为数据挖掘领域的研究热点之一. 作为一种能够有效处理不确定、不精确信息的数学工具,粗糙集理论[4-6]在无需任何先验信息的情况下,仅利用给定数据的等价类即可判定属性子集对决策的支持程度,被广泛应用于高维数据的特征选择. 然而,随着泛在感知技术的快速发展,实际工业系统需要处理的往往是名词型属性与数值型属性并存的混合数据[7-8],而基于等价关系的经典粗糙集理论已不再适用. 为此,Lin[9]利用邻域关系替代等价关系,提出了邻域系统的概念,为数值型数据的处理提供了可行的方法;Hu等[10]在此基础上依据邻域半径的大小确定邻域类,提出了邻域粗糙集模型,并引入了邻域依赖度的概念,设计了一种适用于混合型数据的特征选择算法.

邻域粗糙集模型因其能够对名词型数据和数值型数据的邻域空间进行划分,诸多专家学者对其拓展模型以及相应的特征选择算法进行了研究. 黄恒秋等[11]引入了联系度距离函数,并在此基础上构建了一种双邻域粗糙集模型,在混合数据具有较多缺失值的情况下,所选取的特征优势更为明显;徐怡等[12]依据混合数据不同的属性子集序列和邻域半径,建立了基于双重粒化准则的邻域多粒度粗糙集模型;马福民等[13]在此基础上深入分析了不同属性子集序列和邻域半径对正域的影响,提出了一种该模型下的特征快速选择算法;Yang等[14]将伪标签策略引入至邻域粗糙集模型,通过距离函数和样本生成的伪标签共同划分邻域空间,降低了混合数据特征选择时的不确定性;针对不平衡的混合数据,Chen等[15]考虑了多元分类中多数类和少数类的分布不均和边界区域不稳定等问题,设计了一种基于边界特征显著性的特征选择算法;Wang等[16]通过引入KNN的思想对邻域关系进行刻画,构造的邻域空间能够更好地表征属性子集的分类能力;Wan等[17]考虑了特征间的相互作用,将特征相关性引入至邻域信息的不确定度量,提出了一种基于邻域条件互信息的交互特征选择算法;针对现有邻域粗糙集大多使用单个预定义的距离函数来描述邻域关系,易受到噪声干扰的问题,Yang等[18]提出了一种基于度量学习的邻域粗糙集模型,改善了邻域空间在知识表达过程中的判别能力.

尽管近些年采用邻域粗糙集对混合数据进行特征选取的研究已取得较为丰富的成果,但是仍然存在不少挑战和问题.

1) 现有邻域粗糙集模型对混合数据邻域关系的度量,仍然是基于等价关系的名词型数据划分与基于相似关系的数值型数据划分的简单融合,在对混合数据的高维特征进行选取时,其构造的邻域空间难以准确刻画各特征所包含的不确定性.

2) 现有邻域粗糙集模型在划分混合数据的邻域空间时,不同属性子集间采用了相同的邻域半径,且邻域半径的选取大多依赖专家经验,因此构造的邻域空间缺乏理论依据且粒度结构复杂,不利于后续使用评价函数进行特征选取.

3) 现有邻域粗糙集模型对数值型数据的邻域关系进行描述时,并未考虑在以不同样本为中心而划分的邻域空间内,样本点空间分布的不同对属性子集判别能力造成的不利影响,因此无法满足复杂的高维混合数据对不确定信息的处理需求.

为解决上述问题,本文在邻域粗糙集模型的基础上,提出一种改进的邻域空间构造方法,并设计相应的邻域空间度量公式作为判别指标,以自适应地调节邻域空间下邻域粒的大小;设计一种考虑边界数据和邻域空间大小的评价函数,以准确地表征高维混合数据下邻域空间的判别能力;基于改进的邻域空间,提出一种启发式的高维混合数据特征选择算法,以有效地对混合数据的高维特征进行选取.

1 邻域粗糙集

定义1 设$\langle U,\Delta,\delta\rangle$为一非空度量空间. 其中:$U$为对象的非空有限集合,称为论域;以$x\in U$为中心、$\delta$为邻域半径的闭球,称为$x$的$\delta$邻域,定义[9]如下:

$$n^{\delta}(x)=\{x_i\in U|\Delta(x,x_i)\leqslant\delta\}. \tag{1}$$

其中:Δ为预定义的距离函数;邻域半径$\delta\geqslant 0$,二者共同决定了邻域的大小.

目前,距离函数Δ常采用闵可夫斯基距离,论域中的对象$x=\{x_1,x_2,\dots,x_n\}$与$y=\{y_1,y_2,\dots,y_n\}$间的闵可夫斯基距离为

$$\Delta(x,y)=\Big[\sum_{l=1}^{n}|x_l-y_l|^p\Big]^{\frac{1}{p}}, \tag{2}$$

其中p为常数,通常取1或2.

定义2 设$\mathrm{NS}=(U,\mathrm{AT},V,F,\delta)$为一个邻域信息系统. 其中:$U$为对象的非空有限集合,称为论域;AT为属性的非空有限集合;$V=\bigcup\limits_{a\in\mathrm{AT}}V_a$为属性值的集合,$V_a$为属性$a\in\mathrm{AT}$的值域;$F=\{f|U\times A\to V\}$为信息函数的集合;$\delta$为邻域半径. 对于属性子集$\forall a\in\mathrm{AT}$,在邻域关系下形成的邻域粒和邻域空间的定义[19]如下.

1) 若$g_a^{\delta}(x)=n_a^{\delta}(x)$,则$g_a^{\delta}(x)$为对象$x\in U$在属性子集$a$下生成的邻域粒.

2) 若$G_a^\delta=\{g_a^\delta(x)|x\in U\}$,则$G_a^\delta$为属性子集$a$下生成的邻域粒子群.

定义3　设$\mathrm{NS}=(U,\mathrm{AT}=A\bigcup B,V,F,\delta)$为一个邻域信息系统. 其中: A为名词型属性; B为数值型属性;对象x在属性集A、B、$A\bigcup B$上划分的邻域分别定义[10]为

$$n_A(x)=\{x_i\in U|\Delta_A(x,x_i)=0\},\tag{3}$$

$$n_B(x)=\{x_i\in U|\Delta_B(x,x_i)\leqslant\delta\},\tag{4}$$

$$n_{(A\bigcup B)}(x)=$$

$$\{x_i\in U|\Delta_A(x,x_i)=0\bigwedge\Delta_B(x,x_i)\leqslant\delta\}.\tag{5}$$

定义4　设$\mathrm{NS}=(U,\mathrm{AT},V,F,\delta)$为一个邻域信息系统,对象集$\forall X\subseteq U$,在属性子集$a$下的邻域下近似集和上近似集定义[10]如下:

$$\underline{N_a}(X)=\{x\in U|n_a(x)\subseteq X\},\tag{6}$$

$$\overline{N_a}(X)=\{x\in U|n_a(x)\bigcap X\neq\varnothing\}.\tag{7}$$

对象集X基于邻域关系N的近似精度定义[10]如下:

$$\alpha_N(X)=|\underline{N}(X)|/|\overline{N}(X)|.\tag{8}$$

2　改进的混合数据邻域空间

为了更为合理地对高维混合数据的邻域关系进行度量,在邻域粗糙集模型的基础上,提出一种改进的邻域空间构造方法: 1) 将名词型数据邻域关系的度量从等价关系拓展到相似关系,使得不论在何种类型的属性子集下,邻域空间的大小均得以被灵活调整; 2) 依据粗糙模糊k-means聚类得到的下近似集和边界集,构造双邻域粒子群下的数值型数据邻域空间,考虑了属性子集下数值分布的不同对分类能力产生的不利影响; 3) 设计了相应的邻域空间度量方法,并依据决策属性邻域空间的大小,自适应调节各属性子集下邻域半径的范围,增强了高维混合数据邻域空间内邻域粒的一致性.

2.1　邻域空间度量方法

定义5　设$\mathrm{NS}=(U,\mathrm{AT}\bigcup d,V,F,\delta)$为一个邻域信息系统,其中$d$为决策属性的集合. 对于属性子集$\forall a\subseteq\mathrm{AT}$, g_a^δ为属性子集a下的邻域粒. 属性子集下a的邻域空间度量公式如下式所示:

$$\mathrm{SM}(G_a^\delta)=\frac{S(g_a^\delta(x_i))}{|U|}=\frac{1}{|U|}\sum_{i=1}^{|U|}|g_a^\delta(x_i)|,\tag{9}$$

其中$|\cdot|$为给定集合的基数. 邻域空间的大小满足下列条件: $|g_a^\delta(x_i)|_{\min}\leqslant\mathrm{SM}(G_a^\delta)\leqslant\|g_a^\delta(x_i)|_{\max}$.

该度量公式反映了邻域空间下邻域粒的大小和粒度结构的复杂程度. 度量值越小,该邻域空间下的邻域粒越小,粒度结构越复杂;度量值越大,该邻域空间下的邻域粒越大,粒度结构越简单. 因此,决策属性d的邻域空间度量公式如下式所示:

$$\mathrm{SM}(G_d)=\frac{S\Big(\sum\limits_{i=1}^{|U/R_d|}e_d^i\Big)}{|U|}=\frac{1}{|U|}\sum_{i=1}^{|U/R_d|}|e_d^i|^2.\tag{10}$$

其中: U/R_d为属性子集d在等价关系集R下划分的等价类集合, e_d为属性子集d下的等价类. 邻域空间的大小满足下列条件: $|e_d^i|_{\min}\leqslant\mathrm{SM}(G_d)\leqslant|e_d^i|_{\max}$.

在确定决策属性的邻域空间后,可依据下式自适应地调节属性子集a下邻域半径的范围:

$$\hat{\delta}_a=\arg\min(\mathrm{SM}(G_a^{\hat{\delta}})\geqslant\mathrm{SM}(G_d)).\tag{11}$$

2.2　邻域空间构造方法

2.2.1　名词型数据邻域空间

现有模型对名词型数据的度量,其本质仍然是基于对象的等价关系. 一方面,邻域空间的大小较为固定,且与数值型数据邻域空间的粒度结构差别较大;另一方面,基于该固有的邻域空间,数据内部的隐含知识无法被充分挖掘. 为此,将名词型数据邻域关系的度量从等价关系拓展到相似关系,进而对名词型数据的邻域空间进行构造.

定义6　名词型数据相似性度量公式如下式所示:

$$\mathrm{NDM}(x,y)=\sqrt{\sum_{a=1}^{m}d_a(x,y)}.\tag{12}$$

其中: m为特征的个数; $d_a(x,y)$为数值x与y关于属性a之间的距离,其具体定义为

$$d_a(x,y)=\begin{cases}\mathrm{vdm}_a(x,y),\ a\text{为无序属性;}\\ \mathrm{odm}_a(x,y),\ a\text{为有序属性.}\end{cases}\tag{13}$$

其中: $\mathrm{vdm}_a(x,y)$和$\mathrm{odm}_a(x,y)$距离定义分别为

$$\mathrm{vdm}_a(x,y)=\Big[\sum_{c=1}^{C}\Big|\frac{N_{a,x,c}}{N_{a,x}}-\frac{N_{a,y,c}}{N_{a,y}}\Big|^m\Big]^{\frac{1}{m}}.\tag{14}$$

其中: $N_{a,x}$为在属性a上取值为x的数量, $N_{a,x,c}$为在第c个决策类中于属性a上取值为x的数量, C为决策类的个数.

$$\mathrm{odm}_a(x,y)=|x_{\mathrm{if}}-y_{\mathrm{if}}|=\Big|\frac{r_x-1}{M_f-1}-\frac{r_y-1}{M_f-1}\Big|.\tag{15}$$

式中: r_x为变量x在有序变量R_{if}上的序, $R_{\mathrm{if}}\in\{1,2,\ldots,M_f\}$; $\dfrac{r_x-1}{M_f-1}$将变量x的范围映射至$[0,1]$间.

定义7　设$\mathrm{NS}=(U,A,V,F,\delta)$为一个邻域信息系统. 其中: A为名词型属性, $a\in A$为有序型属性, $b\in A$为无序型属性. 对象x在属性子集a和b上划分的邻域定义为

$$n_a(x)=\begin{cases}\{x_i\in U|\mathrm{vdm}_a(x,x_i)=0\},\ |U/R_a|\leqslant C;\\ \{x_i\in U|\mathrm{vdm}_a(x,x_i)\leqslant\hat{\delta}_a\},\ |U/R_a|>C.\end{cases}\tag{16}$$

其中$\hat{\delta}_a=\arg\min(\mathrm{SM}(G_a^{\delta})\geqslant\mathrm{SM}(G_d))$.

$$n_b(x)=\begin{cases}\{x_i\in U|\mathrm{odm}_a(x,x_i)=0\},\ |R_{\mathrm{if}}|\leqslant C;\\ \{x_i\in U|\mathrm{odm}_a(x,x_i)\leqslant\hat{\delta}_b\},\ |R_{\mathrm{if}}|>C.\end{cases}\tag{17}$$

同理,$\hat{\delta}_b=\arg\min(\mathrm{SM}(G_b^{\delta})\geqslant\mathrm{SM}(G_d))$.

基于上述定义,名词型数据邻域关系的度量可拓展为相似关系. 结合式(11),各属性子集下的邻域半径可被进一步确定.

由此可知,在属性集a、b、$a\bigcup b$下划分的邻域空间为

$$G_a=\{n_a(x)|x\in U\},\ G_b=\{n_b(x)|x\in U\};\tag{18}$$

$$G_{a\bigcup b}=\{n_a(x)\bigcap n_b(x)|x\in U\}.\tag{19}$$

定理1 设$\mathrm{NS}=(U,A=a\bigcup b,V,F,\delta)$为一个邻域信息系统,对于任意的属性集$a$、$b$、$a\bigcup b$下划分的邻域空间,得到:

1) $G_{a\bigcup b}\subseteq G_a\subseteq G_a\bigcup G_b$;

2) $G_{a\bigcup b}\subseteq G_b\subseteq G_a\bigcup G_b$.

证明 1) 对于$\forall x_i\in U$,可知

$$(n_a(x)\bigcap n_b(x))\subseteq n_a(x)\subseteq(n_a(x)\bigcup n_b(x))\Rightarrow$$
$$\{n_a(x)\bigcap n_b(x)|x\in U\}\subseteq\{n_a(x)|x\in U\}\subseteq$$
$$\{n_a(x)\bigcup n_b(x)|x\in U\}\Rightarrow G_{a\bigcup b}\subseteq G_a\subseteq G_a\bigcup G_b.$$

2) 的证明过程同1). □

例1 以表1所示的邻域信息系统为例,进一步表明该邻域空间构造方法的优势和意义. 其中:a_1、a_2为无序型属性,b_1、b_2为有序型属性,d为决策属性.

表 1 邻域信息决策

U	a_1	a_2	b_1	b_2	d
x_1	true	red	A	very low	0
x_2	true	red	B	low	0
x_3	true	blue	B	medium	0
x_4	true	green	B	high	1
x_5	false	black	C	very high	1
x_6	false	yellow	D	high	1
x_7	false	purple	E	extreme	1
x_8	false	purple	F	extreme	1

由式(9)和(10),得到

$$\mathrm{SM}(G_{a_1})=(4+4+4+4+4+4+4+4)/8=4,$$
$$\mathrm{SM}(G_{a_2})=$$
$$(2+2+1+1+1+1+2+2)/8=1.5,$$
$$\mathrm{SM}(G_{b_1})=$$
$$(1+3+3+3+1+1+1+1)/8=1.75,$$
$$\mathrm{SM}(G_{b_2})=$$
$$(1+1+1+2+1+2+2+2)/8=1.5,$$
$$\mathrm{SM}(G_d)=$$
$$(3+3+3+5+5+5+5+5)/8=4.25\ .$$

由此可知,该决策表在等价关系下构造的邻域空间粒度结构差别较大,各属性内含的不确定性难以在不同的粒度标准下被精准地刻画.

由定义6和定义7,相似关系下形成的邻域信息系统如表2所示.

表 2 相似关系下的邻域信息决策

U	a_1	a_2	b_1	b_2	d
x_1	0	0	0	0	0
x_2	0	0	0.2	0.2	0
x_3	0	0	0.2	0.4	0
x_4	0	1	0.2	0.6	1
x_5	1	1	0.4	0.8	1
x_6	1	1	0.6	0.6	1
x_7	1	1	0.8	1	1
x_8	1	1	1	1	1

结合已知的$\mathrm{SM}(G_d)$,各属性子集下的邻域半径为$\theta_{a_1}=0,\theta_{a_2}=0,\theta_{b_1}=0.4,\theta_{b_2}=0.4$. 由此得到各属性在相似关系下形成的邻域空间,其中

$$\mathrm{SM}(G_{a_1})=(4+4+4+4+4+4+4+4)/8=4,$$
$$\mathrm{SM}(G_{a_2})=$$
$$(3+3+3+5+5+5+5+5)/8=4.25,$$
$$\mathrm{SM}(G_{b_1})=(4+5+5+5+5+3+3+2)/8=4,$$
$$\mathrm{SM}(G_{b_2})=$$
$$(2+3+4+4+5+4+4+3)/8=3.5\ .$$

可以明显看出,改进后邻域空间的邻域粒大小更为接近,便于后续在相近的粒度标准下对特征进行选取.

2.2.2 名词型数据邻域空间

现有模型在对数值型数据的邻域空间进行划分时,忽略了在以不同样本中心为基准划分的邻域空间内,样本空间分布的不同对属性子集判别能力造成的不利影响. 为此,引入一种伪标签策略,依据粗糙模糊k-means聚类得到每个决策的下近似和边界区域,构

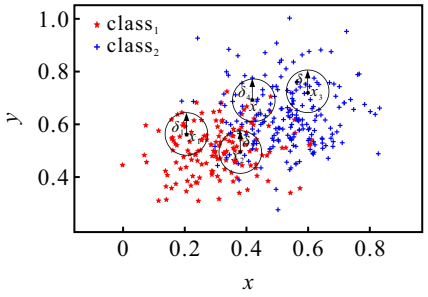

图 1 数值型数据邻域粒的划分

建双重邻域粒子群下的数值型数据邻域空间. 该步骤使得邻域空间的粒度结构更为清晰,显著降低了特征选择时的不确定性. 以图1数值型数据邻域粒的划分为例,给出具体讨论.

由图1可见: x_1 和 x_2 分别为类簇1的下近似样本和边界样本; x_3 和 x_4 分别为类簇2的下近似样本和边界样本. 在以属性集的下近似样本为中心描述邻域关系时,构造的邻域粒对决策具有较高的支持程度;而以属性集的边界样本为中心描述邻域关系时,邻域粒中却含有较多的误分类样本,对决策的支持程度较差. 众所周知,决策边界周围的样本往往具有噪声和不确定性,若以相同的权重参与混合数据的特征选择中,势必会影响算法的准确性.

粗糙模糊 k-means[20] 是一种有效的聚类算法,该算法能够在获得数据集原始雏形的基础上,有效甄别类簇边界的不确定数据. 因此,可有效适用于数值型数据邻域空间的划分.

算法中不同的数据样本对于类簇中心的隶属度公式[20] 如下式所示:

$$u_{\text{in}}=\frac{(1/d_{\text{in}})^{\frac{2}{m-1}}}{\sum_{z=1}^{k}(1/d_{\text{zn}})^{\frac{2}{m-1}}}=\frac{1}{\sum_{z=1}^{k}(d_{\text{in}}/d_{\text{zn}})^{\frac{2}{m-1}}}. \tag{20}$$

其中: u_{in} 为样本 x_n 对于第 i 个类簇的隶属度; d_{in} 为样本 x_n 与簇心 v_i 的欧氏距离; m 为模糊系数; k 为聚类的类簇个数, $k\geqslant i$.

相应地,类簇中心的迭代公式[20] 如下式所示:

$$v_i=\begin{cases} w_l\times\dfrac{\sum\limits_{x_n\in\underline{C_i}}\mu_{\text{in}}^m x_n}{\sum\limits_{x_n\in\underline{C_i}}\mu_{\text{in}}^m}+w_b\times\dfrac{\sum\limits_{x_n\in\hat{C}_i}\mu_{\text{in}}^m x_n}{\sum\limits_{x_n\in\hat{C}_i}\mu_{\text{in}}^m}, \\ \qquad \underline{C_i}\neq\varnothing\bigwedge\hat{C}_i\neq\varnothing; \\ \dfrac{\sum\limits_{x_n\in\underline{C_i}}\mu_{\text{in}}^m x_n}{\sum\limits_{x_n\in\underline{C_i}}\mu_{\text{in}}^m},\ \underline{C_i}\neq\varnothing\bigwedge\hat{C}_i=\varnothing; \\ \dfrac{\sum\limits_{x_n\in\hat{C}_i}\mu_{\text{in}}^m x_n}{\sum\limits_{x_n\in\hat{C}_i}\mu_{\text{in}}^m},\ \underline{C_i}=\varnothing\bigwedge\hat{C}_i\neq\varnothing. \end{cases} \tag{21}$$

其中: $\underline{C_i}$ 为第 i 个类簇的下近似数据集, $\hat{C}_i$ 为第 i 个类簇的边界数据集.

利用粗糙模糊 k-means 聚类得到每个样本的伪标签,由此可构建双重邻域粒子群下的邻域空间,其具体定义如下.

定义8　设 $\text{NS}=(U,B,V,F,\delta)$ 为一个邻域信息系统. 其中: B 为数值型属性,属性子集 a、$v\in B$,对象 x 在属性子集 a 上划分的邻域为

$$n_a(x)=\{x_i\in U|\Delta_a(x,x_i)\leqslant\hat{\delta}_a\}, \tag{22}$$

其中 $\hat{\delta}_a=\arg\min(\text{SM}(G_a^{\delta})\geqslant\text{SM}(G_d))$. 因此,在属性子集 a 下划分的邻域空间为

$$\underline{G_a}=\{n_a(x)|x\in\underline{\text{Apr}^a}\}\hat{G}_a=\{n_a(x)|x\in\hat{\text{Apr}}^a\}, \tag{23}$$

$$G_a=\underline{G_a}\bigcup\hat{G}_a=\{n_a(x)|x\in\underline{\text{Apr}}\bigvee x\in\hat{\text{Apr}}\}. \tag{24}$$

进一步地,在属性子集 $a\bigcup b$ 下划分的邻域空间为

$$\begin{aligned}&\underline{G_{a\cup b}}=\{n_a(x)\bigcap n_b(x)|x\notin\hat{\text{Apr}}^a\bigvee x\notin\hat{\text{Apr}}^b\},\\&\hat{G}_{a\cup b}=\{n_a(x)\bigcap n_b(x)|x\in\hat{\text{Apr}}^a\bigwedge x\in\hat{\text{Apr}}^b\},\end{aligned} \tag{25}$$

$$\begin{aligned}&G_{a\cup b}=\\&\{n_a(x)\bigcap n_b(x)|x\notin\hat{\text{Apr}}^a\bigvee x\notin\hat{\text{Apr}}^b\}\bigcup\\&\{n_a(x)\bigcap n_b(x)|x\in\hat{\text{Apr}}^a\bigwedge x\in\hat{\text{Apr}}^b\}.\end{aligned} \tag{26}$$

其中: $\underline{\text{Apr}^a}$ 为粗糙模糊 k-means 在属性子集 a 上划分的下近似集, $\hat{\text{Apr}}^a$ 为属性子集 a 上的边界集.

定理2　设 $\text{NS}=(U,B,V,F,\delta)$ 为一个邻域信息系统,其中属性子集 a、$b\in B$. 在属性子集 $a\bigcup b$ 下划分的下近似邻域空间为

$$\underline{G_{a\cup b}}=\{n_a(x)\bigcap n_b(x)|x\in\underline{\text{Apr}^a}\bigvee x\in\underline{\text{Apr}^b}\}.$$

证明　对于 $\forall x_i\in U$,可知

$$\begin{aligned}&x\in\underline{\text{Apr}^a}\bigvee x\in\underline{\text{Apr}^b}\Rightarrow\\&x\in(U-\hat{\text{Apr}}^a)\bigvee x\in(U-\hat{\text{Apr}}^b)\Rightarrow\\&x\notin\hat{\text{Apr}}^a\bigvee x\notin\hat{\text{Apr}}^b\Rightarrow\\&\underline{G_{a\cup b}}=\{n_a(x)\bigcap n_b(x)|x\in\underline{\text{Apr}^a}\bigvee x\in\underline{\text{Apr}^b}\}.\end{aligned}$$

定理3　设 $\text{NS}=(U,B,V,F,\delta)$ 为一个邻域信息系统,其中属性子集 a、$b\in B$. 在属性子集 $a\bigcup b$ 下划分的邻域空间为:

1) $G_a=\underline{G_a}\bigcup\hat{G}_a=\{n_a(x)|x\in U\}$;

2) $G_{a\cup b}=\{n_a(x)\bigcap n_b(x)|x\in U\}$.

证明　1) 对于 $\forall x\in U$,可知

$$\begin{aligned}&x\in U\Rightarrow x\in\text{apr}_1\bigvee x\in\text{apr}_2\bigvee\ldots\bigvee\text{apr}_k\Rightarrow\\&(x\in\underline{\text{apr}_1}\bigvee x\in\hat{\text{apr}}_1)\bigvee(x\in\underline{\text{apr}_2}\bigvee x\in\hat{\text{apr}}_2)\bigvee\\&\ldots\bigvee(x\in\underline{\text{apr}_k}\bigvee x\in\hat{\text{apr}}_k)\Rightarrow\\&x\in(\underline{\text{apr}_1}\bigvee\underline{\text{apr}_2}\bigvee\ldots\bigvee\underline{\text{apr}_k})\bigvee\\&x\in(\hat{\text{apr}}_1\bigvee\hat{\text{apr}}_2\bigvee\ldots\bigvee\hat{\text{apr}}_k)\Rightarrow\\&x\in\bigcup_{i=1}^{k}\underline{\text{apr}_i}\bigvee x\in\bigcup_{i=1}^{k}\hat{\text{apr}}_i\Rightarrow x\in\underline{\text{Apr}}\bigvee\\&x\in\hat{\text{Apr}}\Rightarrow G_a=\underline{G_a}\bigcup\hat{G}_a=\{n_a(x)|x\in U\}.\end{aligned}$$

2)的证明过程同1). □

2.2.3 混合型数据邻域空间

名词型数据的度量方法由基于不可分辨关系的等价类划分拓展为基于相似关系的邻域类划分,该邻域空间构造的理论依据与数值型数据相同. 因此,混合型数据的邻域空间由两者融合而成,定义如下.

定义9 设$\mathrm{NS}=(U,\mathrm{AT}=A\bigcup B,V,F,\delta)$为一个邻域信息系统. 其中:$a\in A$为名词型属性,$b\in B$为数值型属性,对象$x$在属性集$a\bigcup b$上划分的邻域为

$$n_{a\bigcup b}(x)=\begin{cases}x_i\in U|\Delta_a(x,x_i)=0\bigwedge\Delta_b(x,x_i)\leqslant\hat{\delta},\\ \qquad |U/R_a|\leqslant C\text{ or }|R_{\mathrm{if}}|\leqslant C;\\ x_i\in U|\Delta_a(x,x_i)\leqslant\hat{\delta}\bigwedge\Delta_b(x,x_i)\leqslant\hat{\delta},\\ \qquad |U/R_a|>C\text{ or }|R_{\mathrm{if}}|>C.\end{cases}\tag{27}$$

因此,在混合属性集$a\bigcup b$下划分的邻域空间为

$$\underline{G_a\bigcup b}=\{n_a(x)\bigcap n_b(x)|x\in\underline{\mathrm{Apr}^b}\},$$
$$\hat{G}_a\bigcup b=\{n_a(x)\bigcap n_b(x)|x\in\hat{\mathrm{Apr}}^b\}.\tag{28}$$

3 基于改进邻域空间的特征选择算法

为了准确地表征不同属性子集下邻域空间的判别能力,提出了一种名为邻域空间支持度的评价函数,充分考虑边界数据和邻域空间大小对高维混合数据特征选择的影响. 基于改进的邻域空间和评价函数,提出了一种启发式的高维混合数据特征选择算法.

3.1 邻域空间支持度

作为最常用的2种评价属性子集判别能力的指标,近似质量γ[21]和条件邻域熵ENT[22]近年来被广泛应用于混合数据的特征选择. 近似质量依据样本的下近似是否完全属于某一决策类来表征特征的判别能力;条件邻域熵则引入熵的不确定性来表征特征的判别能力. 然而,当面对高维的混合数据,随着特征子集的增加,动态变化的邻域空间在一定程度上会影响后续特征选择的准确性. 此外,现有评价函数尚未考虑改进后混合数据邻域空间的不同构造方法,因此提出一种名为邻域空间支持度的评价指标对混合数据的特征进行选取.

定义10 设$\mathrm{NS}=(U,\mathrm{AT}\bigcup d,V,F,\delta)$为一个邻域信息系统. 其中:$A\subseteq\mathrm{AT}$为名词型属性,$B\subseteq\mathrm{AT}$为数值型属性,$C\subseteq\mathrm{AT}$为AT的属性子集,则决策属性$d$在属性子集$\forall a\in\mathrm{AT}-C$下的邻域空间支持度为

$$\mathrm{NGSR}_a=\frac{1}{|U|}\sum_{x\in U}\frac{\|n_a(x)\bigcap[x]_d\|}{|n_a(x)|+\xi\times[x]_d},\ a\in A;\tag{29}$$

$$\mathrm{NGSR}_a=w_l\times\frac{1}{|\underline{\mathrm{Apr}^a}|}\sum_{x\in\underline{\mathrm{Apr}^a}}\frac{\|n_a(x)\bigcap[x]_d\|}{|n_a(x)|+\xi\times[x]_d}+$$
$$w_b\frac{1}{|\hat{\mathrm{Apr}}^a|}\sum_{x\in\hat{\mathrm{Apr}}^a}\frac{\|n_a(x)\bigcap[x]_d\|}{|n_a(x)|+\xi\times[x]_d},$$
$$a\in B.\tag{30}$$

其中:w_l、w_b为权重系数,且$0<w_b<w_l<1$,取值与式(21)相同;ξ为惩罚系数,且$0<\xi<1$,取值越小,条件越苛刻,所选取的特征越少. 因此,改进的启发式前向特征选择公式如下式所示:

$$\mathrm{Sig}(a,C,d)=\mathrm{NGSR}_{C\bigcup a}(d)-\mathrm{NGSR}_C(d),\tag{31}$$

其中$0\leqslant\mathrm{Sig}(a,C,d)\leqslant1$. 若$\mathrm{Sig}(a,C,d)=0$,则特征$a$是冗余的.

3.2 混合数据特征选择算法

基于改进的混合数据邻域空间和设计的邻域空间支持度函数,提出一种前向启发式的混合数据特征选择算法. 算法的步骤和伪代码如下所示.

算法1 基于改进邻域空间的混合数据特征选择算法(INMFS).

输入: 邻域信息系统$\mathrm{NS}=(U,\mathrm{AT}=A\bigcup B\bigcup d,V,F)$,权重系数$w_l$、$w_b$,惩罚函数$\xi$;

输出: 特征子集SF.

step 1: $\mathrm{SF}\leftarrow\varnothing,\mathrm{temp}=0$;利用式(10)计算$\mathrm{SM}(G_d)$.

step 2: 对于$\forall a_i\in A$

step 2.1: 由式(11)确定邻域半径$\hat{\delta}_{a_i}$;

step 2.2: 利用式(16)和(17)计算$n_{a_i}(x)$,并划分邻域空间G_{a_i};

step 2.3: 利用式(29)计算NGSR_{a_i}.

step 3: 对于$\forall b_i\in B$

step 3.1: 由式(11)确定邻域半径$\hat{\delta}_{b_i}$;

step 3.2: 利用式(22)计算$n_{b_i}(x)$,并划分邻域$\underline{G_{b_i}}$和$\hat{G}_{b_i}$;

step 3.3: 利用式(30)计算NGSR_{b_i}.

step 4: 对于$\forall\mathrm{at}_i\in\mathrm{AT}$,计算$\mathrm{Sig}(\mathrm{at}_i,\mathrm{SF},d)$.

step 5: 选择满足下式的特征at_i:

$$\mathrm{Sig}(\mathrm{red},\mathrm{SF},d)=\max(\mathrm{Sig}(\mathrm{at}_i,\mathrm{SF},d)).$$

step 6: if $\mathrm{Sig}(\mathrm{red},\mathrm{SF},d)>\mathrm{temp}$

$\mathrm{temp}=\mathrm{Sig}(\mathrm{red},\mathrm{SF},d)$

$\mathrm{SF}\leftarrow\mathrm{SF}\bigcup\{\mathrm{red}\},\mathrm{AT}\leftarrow\mathrm{AT}-\{\mathrm{red}\}$

go back to step 4

else break.

算法流程如图2所示. 算法的时间复杂度分析如下:step 1的时间复杂度为$O(1)$,step 2的时间复杂度为$O(|U|^2\times|A|)$,step 3的时间复杂度为$O(|U|\times|B|\times(|U|+|U/d|\times I))$,step 4 ~ step 6的时间复杂

为$O((|\mathrm{SF}|^2+|\mathrm{SF}|)/2)$, 其中$I$为聚类算法的迭代次数. 因此, 算法1的总体时间复杂度为$O(|U|^2\times|\mathrm{AT}|+|U|\times B\times|U/d|\times I+(|\mathrm{SF}|^2+|\mathrm{SF}|)/2+1)$.

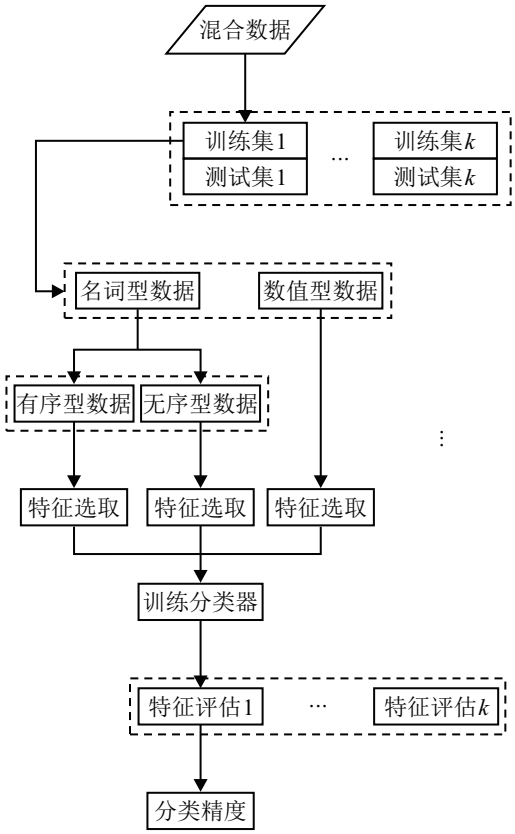

图 2　特征选择算法流程

4　实验分析

本节通过相关实验来验证所提出混合数据特征选择算法INMFS的有效性, 所选取的UCI数据集特征如表3所示. 实验的硬件运行环境为CPU AMD Ryzen 7 5 800 H, 3.2 GHz, 内存16 GB, 采用Python 3.8在软件Spyder 4.2.5上进行编译运行.

表 3　数据集

数据集	对象	名词型数据	数值型数据	分类个数
Lymphography	148	18	0	4
Breast	286	9	0	2
Wine	178	0	13	3
Ionosphere	351	0	34	2
Wdbc	569	0	30	2
Heart-Cleveland	297	8	5	5
Credit	652	9	6	2
Dermatology	366	33	1	6
German	1 000	16	4	2
Waveform	5 000	0	22	3
Segmentation	2 310	0	19	7
Adult	30 162	8	6	2

为了验证所提出算法的有效性, 分别采用邻域粗糙集模型(NRS)[10]、基于多粒度的邻域粗糙集模型(NMGRS)[23]和基于邻近的邻域粗糙集模型(KNNB)[16]与所提出基于改进邻域空间的邻域粗糙集模型进行对比. 不同算法的参数设置如下: 对于NRS和NMGRS, 依据文献[10, 15], 邻域半径$\delta=0.2$; 对于KNNB, 依据文献[16], 邻域半径$\delta=0.2$和邻近数$K=0.05N$; 对于INMFS, 无需设置邻域半径, 权重系数[20]$w_l=0.7, w_b=0.3$和惩罚系数$\xi=0.01$. 2个经典的分类器KNN、CART和10折交叉验证用于评估不同特征选择算法的性能, 算法的性能由2个指标表征: 1) 特征选择的数量; 2) KNN ($K=3$)和CART分类器的分类精度. 实验结果如表4～表6以及图3所示.

表 4　特征选择数量比较

数据集	特征数	NRS		NMGRS		KNNB		INMFS
		γ	ENT	γ	ENT	γ	ENT	NGSR
Lymphography	18	6	6	6	6	7	6	7
Breast	9	7	6	7	6	6	5	4
Wine	13	9	7	9	7	5	6	5
Ionosphere	34	12	9	12	9	5	5	7
Wdbc	30	12	7	12	7	5	5	7
Heart-Cleveland	13	11	11	10	10	7	6	4
Credit	15	12	9	7	9	6	7	4
Dermatology	34	10	10	11	11	10	9	9
German	20	9	8	10	9	6	6	4
Waveform	22	11	13	11	13	8	7	8
Segmentation	19	11	7	11	7	5	5	6
Adult	14	9	8	7	6	4	7	4
平均值	20.09	9.92	8.42	9.42	8.33	6.17	6.17	**5.75**

表5 分类器KNN($K=3$)分类精度对比

数据集	Raw Data	NRS		NMGRS		KNNB		INMFS
		γ	ENT	γ	ENT	γ	ENT	NGSR
Lymphography	77.04±10.98	73.57±7.86	73.57±7.86	73.57±7.86	73.57±7.86	77.04±7.92	77.66±10.42	**80.43±9.43**
Breast	71.45±5.08	66.73±9.11	70.68±7.37	66.73±9.11	70.68±7.37	70.31±10.51	70.31±8.63	**76.12±8.11**
Wine	94.89±4.01	97.15±2.84	97.74±2.76	97.15±2.84	97.74±2.76	92.71±7.86	96.04±5.62	**98.33±2.55**
Ionosphere	84.33±7.58	88.91±5.28	90.31±4.63	88.91±5.28	90.31±4.63	88.64±5.28	**90.88±3.79**	90.13±4.23
Wdbc	97.01±2.22	95.77±2.28	96.31±1.84	95.77±2.28	96.31±1.84	96.31±1.99	95.25±2.36	96.83±2.19
Heart-Cleveland	55.93±6.41	52.93±7.88	54.26±6.94	56.59±5.66	56.91±5.63	52.85±4.79	55.95±7.64	**58.59±3.81**
Credit	82.72±15.83	83.18±9.13	83.49±9.07	83.49±15.67	83.03±15.71	82.26±15.04	82.27±15.81	**85.94±14.83**
Dermatology	96.05±4.07	90.79±6.34	92.48±5.11	90.19±6.09	93.58±4.11	93.29±4.86	94.14±4.54	**96.09±3.08**
German	72.71±1.41	70.51±2.57	69.12±2.25	70.81±4.11	70.42±2.83	71.09±2.84	69.71±2.79	71.69±2.81
Waveform	78.52±4.42	76.21±4.37	75.57±4.01	76.21±4.37	75.57±4.01	69.83±3.49	73.83±3.21	**79.56±3.34**
Segmentation	87.62±4.36	86.67±3.56	88.57±4.36	86.67±3.56	**88.57±4.36**	86.19±5.41	83.33±4.39	88.09±4.39
Adult	78.21±2.61	77.89±2.63	77.99±2.61	76.53±1.87	78.41±2.31	77.58±1.87	78.17±2.49	**79.31±2.29**
平均值	81.37±5.75	80.02±5.32	80.84±4.91	80.21±5.73	81.26±5.29	79.84±5.99	80.62±5.97	**83.43±5.08**

表6 分类器CART分类精度对比

数据集	Raw data	NRS		NMGRS		KNNB		INMFS
		γ	ENT	γ	ENT	γ	ENT	NGSR
Lymphography	78.28±12.61	79.52±11.02	79.52±11.02	79.52±11.02	**79.52±11.02**	75.04±11.12	77.33±9.11	78.95±9.71
Breast	63.05±11.73	64.22±12.44	64.26±9.51	64.22±12.44	64.26±9.51	70.31±10.82	72.89±9.27	**76.14±8.11**
Wine	85.94±9.41	88.73±9.76	88.73±7.21	88.73±9.76	88.73±7.21	92.17±6.09	90.95±6.92	**92.19±7.52**
Ionosphere	88.59±5.26	87.49±6.18	90.31±6.18	87.49±6.18	90.31±6.18	88.03±7.21	89.17±4.76	**92.04±4.69**
Wdbc	91.92±2.95	93.31±2.35	93.31±3.93	93.31±2.35	93.31±3.93	90.34±8.27	91.91±5.95	**94.38±3.31**
Heart-Cleveland	47.53±9.84	46.19±10.85	48.81±8.14	48.16±6.03	50.17±10.08	45.49±7.48	48.89±7.68	**56.24±6.04**
Credit	82.24±13.23	79.97±8.31	80.26±7.89	80.26±15.69	79.49±11.88	81.18±10.46	78.88±14.57	**82.87±12.26**
Dermatology	93.57±3.01	93.31±4.51	94.43±4.47	92.73±4.17	93.88±3.01	91.87±3.92	95.34±3.96	**95.55±3.36**
German	67.51±3.53	65.62±4.96	67.81±6.31	68.49±4.86	68.39±6.46	65.69±4.36	64.73±2.92	**70.69±3.67**
Waveform	73.62±1.08	72.83±4.74	72.28±2.51	72.83±4.74	72.28±2.51	68.41±1.99	67.51±2.18	73.52±2.25
Segmentation	86.67±5.55	88.09±4.88	87.61±7.43	88.09±4.88	87.61±7.43	88.11±6.19	88.09±7.14	**90.48±7.96**
Adult	77.73±2.85	76.89±2.63	74.26±3.56	78.21±2.52	76.67±2.33	76.53±1.73	75.08±2.74	**80.94±1.73**
平均值	78.05±6.75	78.01±6.89	78.46±6.51	78.49±7.05	78.72±6.79	77.76±6.64	78.39±6.43	**82.01±5.88**

由实验结果可以看出,NRS在对单一类型的特征进行选取时,具有良好的效果. 然而,该模型在构造混合数据的邻域空间时,是基于等价关系的名词型数据划分与基于相似关系的数值型数据划分的简单融合,在利用评价函数对混合数据进行特征选取时,无法准确分辨出不同属性子集间判别能力的差异.

NMGRS在NRS的基础上,从多粒度的角度构建不同的属性集序列,再对它们的邻域关系进行描述. 从实验结果分析,粒计算的引入有效提升了算法的分类精度. 然而,该算法在描述不同属性集序列的邻域关系时,依然采用了相同的邻域半径,划分的邻域空间结构复杂,不利于高维混合数据的特征选取.

KNNB在划分混合数据的邻域空间时,引入了KNN的思想来确定邻域空间的大小. 从实验结果来看,该算法有效降低了混合数据邻域空间的不确定性. 然而,与数值型数据相比,名词型数据是由离散的范畴值组成,利用KNN和邻域半径共同确定其邻域空间的大小会丢失较多的信息. 因此,在处理名词型属性较多的高维混合数据时,算法存在一定的缺陷.

INMFS在处理不同类型的数据集时,均取得了较好的特征选择效果. 这主要是由于INMFS重新构造了不同属性子集下的邻域空间,充分挖掘其内在的隐含知识;INMFS还依据决策属性邻域空间的大小,自适应调节了各属性子集下邻域半径的范围,邻域

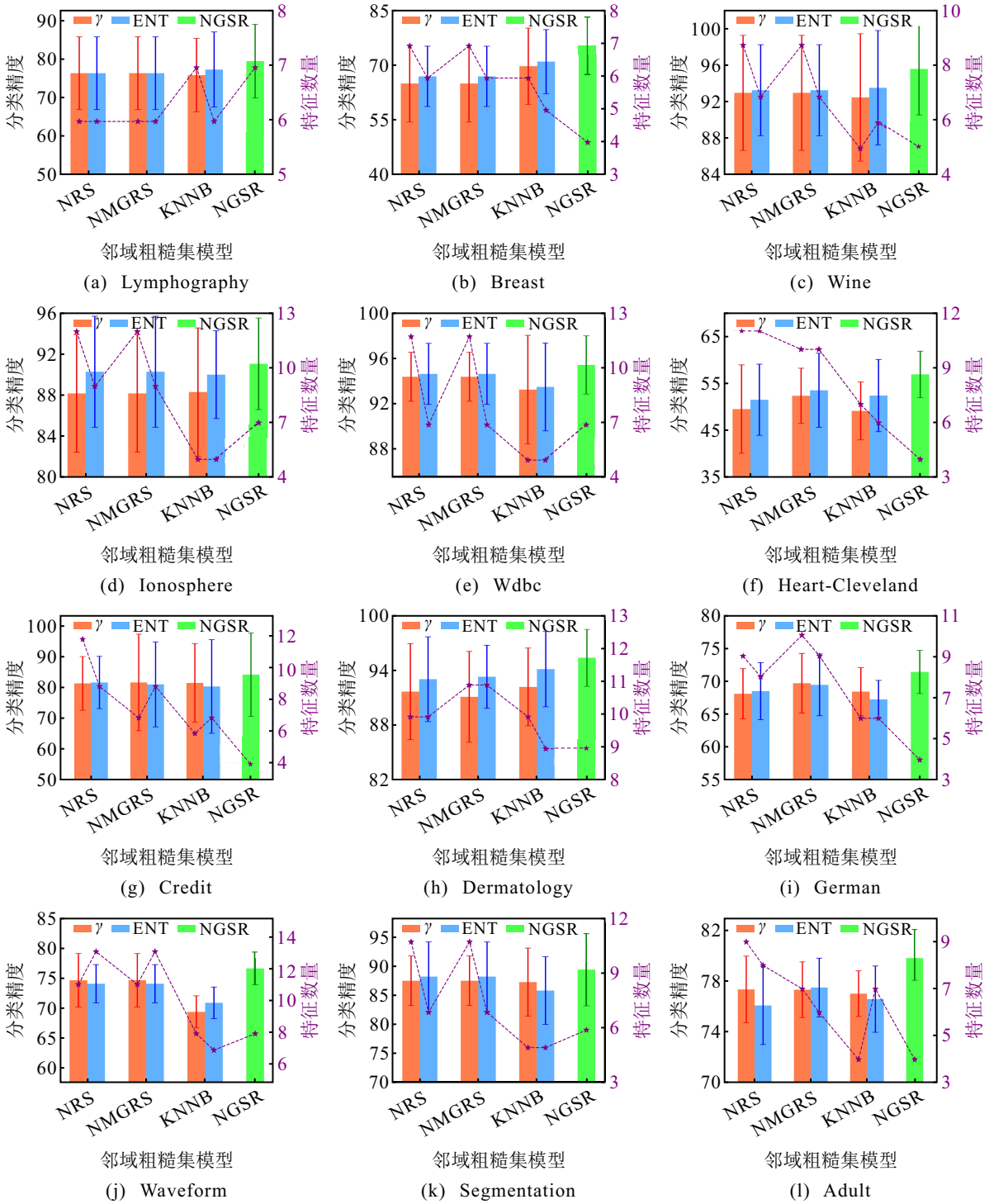

图 3　分类器 KNN ($K=3$) 与 CART 平均分类精度对比

空间的粒度结构得以被清晰地表达; 此外, 设计的邻域空间支持度函数更准确地表征了属性子集的判别能力, 进一步提升了算法的性能.

5　结　论

如何对高维的混合数据进行特征选取一直是数据挖掘领域的重要基础研究内容之一. 本文深入分析了现有邻域粗糙集模型对混合数据邻域关系度量时存在的缺陷, 提出了一种改进的混合数据邻域空间构造和度量方法; 利用设计的邻域空间支持度函数, 提出了一种启发式的高维混合数据特征选择算法. 仿真实验表明, 所提出算法能够有效应对更为复杂的数据应用场景, 进一步提升分类器的分类精度和分类效率. 本文仅考虑了静态数据的特征选择, 如何结合所提出算法的思路, 探索高维混合数据动态的增量式特征选取, 将是下一步研究工作的重点.

参考文献 (References)

[1] Ji W T, Pang Y, Jia X Y, et al. Fuzzy rough sets and fuzzy rough neural networks for feature selection: A review[J].

WIREs Data Mining and Knowledge Discovery, 2021, 11(3): 1-15.

[2] Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends[J]. Information Fusion, 2019, 52: 1-12.

[3] 李郅琴, 杜建强, 聂斌, 等. 特征选择方法综述[J]. 计算机工程与应用, 2019, 55(24): 10-19.
(Li Z Q, Du J Q, Nie B, et al. Summary of feature selection methods[J]. Computer Engineering and Applications, 2019, 55(24): 10-19.)

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

[5] Zhang P F, Li T R, Wang G Q, et al. Multi-source information fusion based on rough set theory: A review[J]. Information Fusion, 2021, 68: 85-117.

[6] 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.
(Zhou T, Lu H L, Ren H L, et al. Survey on attribute reduction algorithm of rough set[J]. Acta Electronica Sinica, 2021, 49(7): 1439-1449.)

[7] Liu J, Li T R, Xie P, et al. Urban big data fusion based on deep learning: An overview[J]. Information Fusion, 2020, 53: 123-133.

[8] Yuan Z, Chen H M, Li T R, et al. Unsupervised attribute reduction for mixed data based on fuzzy rough sets[J]. Information Sciences, 2021, 572: 67-87.

[9] Lin T Y. Granular computing on binary relations I: Data mining and neighborhood systems[J]. Rough Sets in Knowledge Discovery, 1998(1): 107-121.

[10] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[11] 黄恒秋, 曾玲, 黎利辉. 混合值不完备系统的双邻域粗糙集分类方法[J]. 控制与决策, 2018, 33(7): 1207-1214.
(Huang H Q, Zeng L, Li L H. Double-neighborhood rough set classification method in incomplete decision system with hybrid value[J]. Control and Decision, 2018, 33(7): 1207-1214.)

[12] 徐怡, 杨宏健, 纪霞. 基于双重粒化准则的邻域多粒度粗糙集模型[J]. 控制与决策, 2015, 30(8): 1469-1478.
(Xu Y, Yang H J, Ji X. Neighborhood multi-granulation rough set model based on double granulate criterion[J]. Control and Decision, 2015, 30(8): 1469-1478.)

[13] 马福民, 陈静雯, 张腾飞. 基于双重粒化准则的邻域多粒度粗集快速约简算法[J]. 控制与决策, 2017, 32(6): 1121-1127.
(Ma F M, Chen J W, Zhang T F. Quick attribute reduction algorithm for neighborhood multi-granulation rough set based on double granulate criterion[J]. Control and Decision, 2017, 32(6): 1121-1127.)

[14] Yang X B, Liang S C, Yu H L, et al. Pseudo-label neighborhood rough set: Measures and attribute reductions[J]. International Journal of Approximate Reasoning, 2019, 105: 112-129.

[15] 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.

[16] 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.

[17] Wan J H, Chen H M, Yuan Z, et al. A novel hybrid feature selection method considering feature interaction in neighborhood rough set[J]. Knowledge-Based Systems, 2021, 227: 107167.

[18] Yang X L, Chen H M, Li T R, et al. Neighborhood rough sets with distance metric learning for feature selection[J]. Knowledge-Based Systems, 2021, 224: 107076.

[19] Chen Y M, Qin N, Li W, et al. Granule structures, distances and measures in neighborhood systems[J]. Knowledge-Based Systems, 2019, 165: 268-281.

[20] Zhang T F, Ma F M, Yue D, et al. Interval type-2 fuzzy local enhancement based rough k-means clustering considering imbalanced clusters[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(9): 1925-1939.

[21] Hu Q H, Yu D R, Xie Z X. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866-876.

[22] 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.

[23] Lin G P, Qian Y H, Li J J. NMGRS: Neighborhood-based multigranulation rough sets[J]. International Journal of Approximate Reasoning, 2012, 53(7): 1080-1093.

作者简介

张腾飞(1980–), 男, 教授, 博士生导师, 从事智能信息处理、大数据分析等研究, E-mail: tfzhang@126.com;

张宇迪(1997–), 男, 硕士生, 从事数据挖掘、粒计算等研究, E-mail: zyd971001@126.com;

马福民(1979–), 女, 教授, 博士, 从事智能信息处理、智能生产系统等研究, E-mail: fmmatj@126.com.