Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 310-322
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144164

Hans汉斯

混合决策信息系统下基于加权k近邻的属性约简

马　达

长安大学理学院，陕西　西安

收稿日期：2025年3月14日；录用日期：2025年4月7日；发布日期：2025年4月15日

摘　要

属性约简是粗糙集理论的一项重要应用，大多数的邻域粗糙集模型只关注邻域粒完全包含在某些决策类中的对象，而忽略邻域粒不包含在任何决策类中的边界对象的可分性。本文采用k近邻来描述混合信息系统对象的粒化，利用边界对象邻域粒中对象的决策信息以及在各自决策类中的分布情况构造对象的评价函数，并基于此定义近似正域。进而在混合型信息系统上提出基于加权k近邻的约简定义，给出属性的重要性度量，以及基于加权k近邻的属性约简方法。最后在3个混合数据集上对约简结果的分类精度进行比较。与传统邻域粗糙集的属性约简算法相比，该算法能保证约简后数据有较高的分类精度。

关键词

邻域粗糙集，k近邻，边界域，属性约简

Attribute Reduction Based on Weighted k-Nearest Neighborhood in Hybrid Decision Information System

Da Ma

School of Sciences, Chang'an University, Xi'an Shaanxi

Received: Mar. 14th, 2025; accepted: Apr. 7th, 2025; published: Apr. 15th, 2025

Abstract

Attribute reduction is an important application of rough set theory. Most neighborhood rough set models only focus on the objects whose neighborhood granules are completely contained in some decision classes, but ignore the divisibility of boundary objects whose neighborhood granules are not contained in any decision classes. In this paper, the k-nearest neighborhood is used to describe the granulation of hybrid information system objects. The evaluation function of objects is

文章引用: 马达. 混合决策信息系统下基于加权 k 近邻的属性约简[J]. 应用数学进展, 2025, 14(4): 310-322.
DOI: 10.12677/aam.2025.144164

constructed by using the decision information of objects in the neighborhood of boundary objects and their distribution in their decision classes. Then, the definition of reduction based on weighted *k*-nearest neighborhood is proposed in hybrid information system, the importance measure of attribute and the attribute reduction method based on weighted *k*-nearest neighborhood are given. Finally, the classification accuracy of the reduction results is compared on 3 mixed data sets. Compared with the traditional neighborhood rough set attribute reduction algorithm, the proposed algorithm in this paper can ensure the high classification accuracy of the reduced data.

Keywords

Neighborhood Rough Set, *k*-Nearest Neighborhood, Boundary Domain, Attribute Reduction

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

属性约简是数据预处理的关键步骤[1]，在模式识别和机器学习中有着广泛的应用[2] [3]。属性约简要求在保持原始数据集的决策能力不变的前提下，删除其中不相关的、冗余的属性信息，从而起到提高决策效率的作用。通过属性约简，能有效地简化决策规则和减少储存需求，从而降低计算复杂度[4]。

粗糙集理论[5]是一种处理不确定知识的数据分析方法。经典粗糙集理论是建立在等价关系上的，只适用于离散数据。但是，当面对较复杂的实际问题时，可能呈现连续落在某个区间或者数值型与分类型混合的数据。因此，研究人员陆续提出多种信息系统[6]-[9]用于存放复杂数据。粗糙集理论在处理连续数据时，必须先将原始数据离散化，而数据离散化会导致大量的信息丢失。为了解决这一问题，Hu 等[10]在连续型数据集上建立了邻域关系，并提出了邻域粗糙集模型，该模型可以对连续型数据直接进行处理。

目前已经提出了多种邻域粗糙集模型及相应的属性约简算法。文献[11]针对数值型信息系统提出了邻域粗糙集，并利用邻域粗糙集的依赖度构建了数值型信息系统属性约简算法；邻域粗糙集模型只关注邻域完全包含在某些决策类中的对象，因此文献[12]提出了一种最大决策邻域粗糙集模型，增加邻域与某些决策类交集最大的对象来扩展正域，并决策依赖度设计了属性约简算法；文献[13]提出了一种优势邻域粗糙集模型，并构造了一种并行属性约简算法来提高计算速度；文献[14]针对数值型信息系统，利用模糊邻域的概念定义对象的模糊决策，利用模糊邻域与模糊决策之间的关系，构造了一种模糊邻域粗糙集模型；为了处理动态混合信息系统，文献[15]对传统的邻域粗糙集进行了增量式学习的构造，提出了一种基于邻域熵的增量式属性约简算法；文献[16]在计算邻域关系之前充分挖掘属性与决策之间的相关性，对相关性高的属性赋予更高的权重，提出一种加权邻域粗糙集模型，同时设计出相应的属性约简算法；当对象之间存在类内差异时，选择单个属性子集来表示整个数据集可能会降低性能，为此文献[17]对每个类进行再分类，利用邻域粗糙集对每个类进行局部特征选择；文献[18]考虑了边界域的模糊性而导致分类错误的概率和类的不均匀分布，提出了一种基于邻域粗糙集理论的非平衡混合数据特征选择方法；文献[19]基于模糊粗糙集理论，对不同类型的属性定义不同的模糊关系，提出了一种基于模糊粗糙集的信息熵用于混合数据集的特征选择。

邻域粗糙集的目的是利用邻域信息粒来区分属于不同决策的对象。文献[20]结合 δ 单位邻域信息粒和 k 近邻信息粒，提出了一种在实值信息系统上的新的 k 近邻信息粒。在属性约简中，如何选择属性是一个至关重要的问题。依赖度通过比较正域和论域的基数比例来衡量邻域决策系统的分类能力。然而，

基于正域的依赖度忽略了边界域中对象的可分性，为了提高分类能力，通过对边界域中的对象进行再划分来扩大正域。

针对混合决策信息系统，本文引入 k 近邻来描述对象的粒化，通过定义边界域中对象的评价函数来判断其可分性，定义近似正域来反映邻域决策系统的分类能力。由于提出的近似正域不是单调的，进而提出一种具有单调性的近似正域，并定义新的属性依赖度，基于此设计出一种启发式属性约简算法。通过实验验证本文提出的属性约简方法的有效性。

2. 预备知识

称 $DIS=(U,A=C\cup D,V,f)$ 为决策信息系统[21]，其中 U 表示非空有限对象集，称为论域；C 是条件属性集，D 是决策属性集，且 $C\cap D=\varnothing$；信息函数 $f:U\times A\to\bigcup_{a\in A}V_a$，$V_a$ 是属性 a 的值域。

定义 1 [20] 给定决策信息系统 $DIS=(U,A=C\cup D,V,f)$。对任意 $B\subseteq C$，$x_i\in U$，x_i 关于属性子集 B 的 δ 邻域信息粒定义为：

$$\delta_B(x_i)=\{x_j\in U\mid \Delta_B(x_i,x_j)\le\delta\}. \tag{1}$$

其中 $\Delta_B(x_i,x_j)=\sqrt{\sum_{a\in B}\left(\frac{f(x_i,a)-f(x_j,a)}{\sigma_a}\right)^2}$，$\sigma_a$ 表示 V_a 的标准差。

由 δ 邻域信息粒可得到 δ 邻域关系 $N_B^\delta=\{(x_i,x_j)\in U\times U\mid x_j\in\delta_B(x_i)\}$，$N_B^\delta$ 满足自反性和对称性，但不满足传递性。δ 邻域信息粒 $\delta_B(x_i)$ 构成 U 的一个覆盖。

定义 2 [20] 给定决策信息系统 $DIS=(U,A=C\cup D,V,f)$。对任意 $B\subseteq C$，$x_i\in U$，x_i 关于属性子集 B 的 k 近邻信息粒定义为：

$$\kappa_B(x_i)=\{x_i^1,x_i^2,\cdots,x_i^k\mid\forall x_j\in U,x_j\ne x_i^h,\Delta_B(x_i,x_j)>\Delta_B(x_i,x_i^h),h=1,2,\cdots,k\}. \tag{2}$$

$\kappa_B(x_i)$ 表示与 x_i 距离最近的 k 个对象集合。由于 $x_i\in\kappa_B(x_i)$，因此 k 近邻信息粒 $\{\kappa_B(x_i)\mid x_i\in U\}$ 构成 U 的一个覆盖。

文献[20]结合 δ 单位邻域信息粒和 k 近邻信息粒，引入了 x_i 关于属性子集 B 的新的 k 近邻信息粒：

$$\tau_B(x_i)=\{x_j\in U\mid x_j\in B(x_i)\cap\kappa_B(x_i)\}. \tag{3}$$

其中 $B(x_i)=\{x_j\in U\mid\Delta_B(x_i,x_j)\le1\}$。

同样，新的 k 近邻信息粒 $\{\tau_B(x_i)\mid x_i\in U\}$ 也构成了 U 的一个覆盖。

定义 3 [20] 给定决策信息系统 $DIS=(U,A=C\cup D,V,f)$。对任意 $B\subseteq C$，$D_j\in U/D$，D_j 关于属性子集 B 的下、上近似定义为：

$$\underline{KN}_B(D_j)=\{x_i\mid\tau_B(x_i)\subseteq D_j\},\ \overline{KN}_B(D_j)=\{x_i\mid\tau_B(x_i)\cap D_j\ne\varnothing\}. \tag{4}$$

决策属性 D 关于属性子集 B 的下、上近似定义为：

$$\underline{KN}_B(D)=\bigcup_{j=1}^{r}\underline{KN}_B(D_j),\ \overline{KN}_B(D)=\bigcup_{j=1}^{r}\overline{KN}_B(D_j). \tag{5}$$

决策属性 D 关于属性子集 B 的正域[20]表示为：

$$KPOS_B(D)=\underline{KN}_B(D). \tag{6}$$

决策属性 D 关于属性子集 B 的依赖度为：

$$\gamma_B(D)=\frac{\left|POS_B(D)\right|}{|U|}. \tag{7}$$

其中$|\cdot|$表示集合的基数。$\gamma_B^{\delta}(D)$反映了B近似D的能力，$\gamma_B^{\delta}(D)$越大表示B近似D的能力越强。

3. 基于加权 k 近邻的属性约简

本节将文献[20]中信息粒的构造方法引入到混合决策信息系统中，提出混合决策信息系统上新的k近邻信息粒。同时关注边界域中对象邻域的决策信息，定义对象的评价函数来判断边界域对象的可分性，将边界域中能够区分的对象近似地划分到正域中，基于此提出近似正域的概念，进而讨论混合决策信息系统上的属性约简方法。

称$HDIS=(U,A=C\cup D,V,f)$为混合决策信息系统[11]，其中U表示非空有限对象集，称为论域；C是条件属性集，$C=C_c\cup C_n$，C_c表示分类型属性集，C_n表示数值型属性集，D是决策属性集，且$C\cap D=\varnothing$；信息函数$f:U\times A\to\bigcup_{a\in A}V_a$，$V_a$是属性$a$的值域。

定义4 [18] 给定混合决策信息系统$HDIS=(U,A=C\cup D,V,f)$。对任意$B\subseteq C$且$B=B_c\cup B_n$，$x_i\in U$，邻域半径$\delta\in[0,1]$，x_i关于属性子集B的δ邻域信息粒定义为：

$$\delta_B(x_i)=\left\{x_j\in U\mid d_B(x_i,x_j)\le\delta\right\}. \tag{8}$$

其中

$$d_B(x_i,x_j)=\sqrt{\frac{1}{|B|}\left(\sum_{a_k\in B_n}\left|\hat{f}(x_i,a_k)-\hat{f}(x_j,a_k)\right|^2+\sum_{a_l\in B_c}\varphi_{a_l}(x_i,x_j)\right)}, \tag{9}$$

$$\hat{f}(x_i,a_k)=\frac{f(x_i,a_k)-\min(V_{a_k})}{\max(V_{a_k})-\min(V_{a_k})},\forall a_k\in B_n,\quad \varphi_{a_r}(x_i,x_j)=\begin{cases}0, f(x_i,a_r)=f(x_j,a_r)\\ 1, f(x_i,a_r)\neq f(x_j,a_r)\end{cases},\forall a_r\in B_c。$$

定义5 给定混合决策信息系统$HDIS=(U,A=C\cup D,V,f)$。对任意$B\subseteq C$且$B=B_c\cup B_n$，$x_i\in U$，邻域半径$\delta\in[0,1]$，x_i关于属性子集B的k近邻信息粒定义为：

$$\kappa_B(x_i)=\left\{x_i^1,x_i^2,\cdots,x_i^k\mid\forall x_j\in U,x_j\neq x_i^h,d_B(x_i,x_j)>\Delta_B(x_i,x_i^h),h=1,2,\cdots,k\right\}. \tag{10}$$

x_i关于属性子集B新的k近邻信息粒定义为：

$$\tau_B(x_i)=\left\{x_j\in U\mid x_j\in\delta_B(x_i)\cap\kappa_B(x_i)\right\}. \tag{11}$$

例1 考虑混合决策信息系统$HDIS=(U,A=C\cup D,V,f)$，其中$U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$，$C=\{a_1,a_2,a_3,a_4,a_5,a_6,a_7\}$，$C_n=\{a_1,a_2,a_3,a_4,a_5\}$，$C_c=\{a_6,a_7\}$，$D=\{d\}$ (见表1)。

Table 1. Mixed decision information system $HDIS=(U,A=C\cup D,V,f)$

表1. 混合决策信息系统$HDIS=(U,A=C\cup D,V,f)$

U	a_1	a_2	a_3	a_4	a_5	a_6	a_7	d
x_1	0.22	1.02	0.30	0.81	1.52	1	男	健康
x_2	0.24	1.11	0.52	0.39	0.31	1	女	健康
x_3	2.10	0.06	0.95	1.01	2.98	2	女	健康
x_4	0.63	5.04	1.52	1.12	0.64	2	男	不健康

续表

x_5	1.01	1.53	0.71	0.52	1.24	1	男	健康
x_6	0.10	2.05	1.13	0.73	0	1	女	健康
x_7	1.43	3.98	1.36	1.39	1.22	2	男	不健康
x_8	1.21	2.49	0.30	0.88	0.93	1	女	不健康

当 $\delta=0.4$，$k=3$ 时，$\tau_C(x_1)=\{x_1,x_2,x_5\}$，$\tau_C(x_2)=\{x_2,x_6,x_8\}$，$\tau_C(x_3)=\{x_3\}$，$\tau_C(x_4)=\{x_4,x_7\}$，$\tau_C(x_5)=\{x_1,x_5\}$，$\tau_C(x_6)=\{x_2,x_6\}$，$\tau_C(x_7)=\{x_4,x_7\}$，$\tau_C(x_8)=\{x_2,x_8\}$。根据下、上近似的定义可得：$\underline{KN}_C(D)=\{x_1,x_3,x_4,x_5,x_6,x_7\}$，即 $KPOS_C(D)=\{x_1,x_3,x_4,x_5,x_6,x_7\}$。

根据正域的定义，在例 1 中，x_1 的 k 近邻信息粒中对象决策都相同，我们认为 x_1 属于正域，x_2 的 k 近邻信息粒中对象决策不完全相同，我们认为 x_2 属于边界域。但是，边界域中可能存在一些可分的对象被误分，所以认为需要对边界域进行再划分。

本文通过分析邻域粒中对象的决策信息，寻找一种描述可分性的方式，找出存在于边界域中的可分对象。文献[22]中在数值型信息系统中，将对象的决策信息以及在各自决策类中的分布情况考虑到对象的评价标准中，为 k 近邻对象加权。本文借鉴文献[22]中对象 x 的评价准则，提出评价函数来量化边界域中对象决策的差异程度，从而刻画对象的可分性。

定义 6 给定混合决策信息系统 $HDIS=(U,A=C\cup D,V,f)$。对任意 $B\subseteq C$ 且 $B=B_c\cup B_n$，$x\in U$，x 在 B 上的评价函数为：

$$E_B(x)=\sum_{x_i\in[x]_D\cap\tau_B(x)}\omega_B(x_i)\cdot Sim_B(x,x_i)-\sum_{x_i\in([x]_D)^C\cap\tau_B(x)}\omega_B(x_i)\cdot Sim_B(x,x_i) \tag{12}$$

其中 $\sum_{x_i\in[x]_D\cap\tau_B(x)}\omega_B(x_i)\cdot Sim_B(x,x_i)$ 表示 x 和它 k 近邻对象中与 x 决策相同的对象的加权相似度，$\sum_{x_i\in([x]_D)^C\cap\tau_B(x)}\omega_B(x_i)\cdot Sim_B(x,x_i)$ 表示 x 和它 k 近邻对象中与 x 决策不相同的对象的加权相似度。

$Sim_B(x,x_i)=1-d_B(x,x_i)$，其中 $d_B(x,x_i)$ 表示 x 的 k 近邻对象 x_i 和 x 的公式 (9) 中的距离，$\omega_B(x_i)=1-\upsilon_B(x_i)$，其中

$$\upsilon_B(x_i))=\left(\frac{1}{|B|}\sum_{a_k\in B}\left(\upsilon_{a_k}(x_i)\right)^2\right)^{\frac{1}{2}}$$

$$\upsilon_{a_k}(x_i)=\begin{cases}\hat{f}(x_i,a_k)-\overline{f}(x_i,a_k),a_k\in B_n\\ \dfrac{\left|[x_i]_D\cap\left([x_i]_{a_k}\right)^c\right|}{|[x_i]_D|},a_k\in B_c\end{cases}$$

$$\overline{f}(x_i,a_k)=\frac{1}{|[x_i]_D|}\sum_{x_j\in[x_i]_D}^{|[x_i]_D|}\hat{f}(x_j,a_k)$$

在 x 的 k 近邻中，与 x 决策相同的对象越多，其对应的 $E_B(x)$ 越大，表示对象 x 的邻域中与对象 x 决策相同的对象和决策不同的对象之间差异程度越大。因此，根据对象 x 的质量评价值，把评价函数值大的对象近似看作正域对象，认为和正域对象一样是可分的。

定义 7 给定混合决策信息系统 $HDIS=(U,A=C\cup D,V,f)$。对任意 $B\subseteq C$，决策属性 D 在 X 上关于

B 的近似可分集表示为：

$$S_B(D)=\{x\mid E_B(x)\geq\alpha\} \tag{13}$$

则 D 关于 B 的近似正域表示为：

$$APOS_B(D)=\underline{KN}_B(D)=\bigcup_{j=1}^{r}\left(\underline{KN}_B(D_j)\cup S_B(D_j)\right) \tag{14}$$

由于近似正域即为决策属性 D_j 关于属性集 B 的下近似和近似可分集的并集，且 k 近邻 $\tau_B(x_i)$ 的大小相对于 B 不是单调的，因此随着属性集 B 的增大，上面定义的近似正域不是单调的。但是单调函数更有利于设置属性约简的终止条件[20]为了简化属性约简，我们提出了一种计算近似正域的迭代策略，并将近似可分集考虑到计算中。

定义 8 给定混合决策信息系统 $HDIS=(U,A=C\cup D,V,f)$。$B_0=\varnothing$，$B_i=B_{i-1}\cup\{b\}$，$b\in C-B_{i-1}$，$i=1,2,\cdots,|C|$。子集族 $\{B_0,B_1,\cdots,B_{|C|}\}$ 为 C 的集合序列，记为 $I(C)$。决策类 D_j 的近似正域定义为：

$$APOS_{B_i}(D_j)=\bigcup_{k=1}^{i}\underline{KN}_{B_i}(D_j,X_k)\cup S_{B_i}\left(D_j,\overline{KN}_{B_i}(D_j,W_i)-\bigcup_{k=1}^{i}\underline{KN}_{B_i}(D_j,X_k)\right). \tag{15}$$

其中 $X_1=D_j$，$X_k=D_j-\bigcup_{h=1}^{k-1}\underline{KN}_{B_h}(D_j,X_h)\cup S_{B_h}\left(D_j,\overline{KN}_{B_h}(D_j,W_i)-\bigcup_{h=1}^{k}\underline{KN}_{B_h}(D_j,X_h)\right)$，$W_1=U$，$W_i=\overline{KN}_{B_{i-1}}(D_j)$，$k=1,2,\cdots,i$ 且 $i=1,2,\cdots,|A|$。

则对任意 $B_i\in I(C)$，$APOS_{B_i}(D)=\bigcup_{j=1}^{r}APOS_{B_i}(D_j)$。

性质 1 对任意 $B_1,B_2\in I(C)$，$X\subseteq U$，若 $B_1\subseteq B_2$，则有 $APOS_{B_1}(D)\subseteq APOS_{B_2}(D)$。

证明：令 $B_1=\{b_{l_1},b_{l_2},\cdots,b_{l_k}\}$，$B_2=B_1\cup\{b_{m_1},b_{m_2},\cdots,b_{m_t}\}$，根据定义 8 可得：

$\underline{KN}_{B_1}(D_j)=\bigcup_{k=1}^{s}\underline{KN}_{\{b_{l_1},b_{l_2},\cdots,b_{l_k}\}}(D_j,X_k)$，其中 $X_1=D_j$，$X_k=D_j-\bigcup_{h=1}^{k-1}\underline{R}_{B_h}(D_j,X_h)$，$k\geq2$，则有

$$\begin{aligned}\underline{KN}_{B_2}(D_j)&=\bigcup_{k=1}^{s}\underline{KN}_{\{b_{l_1},b_{l_2},\cdots,b_{l_k}\}}(D_j,X_k)\cup\bigcup_{k=1}^{t}\underline{KN}_{B_1\cup\{b_{m_1},b_{m_2},\cdots,b_{m_k}\}}(D_j,X_{s+k})\\&=\underline{KN}_{B_1}(D_j)\cup\bigcup_{k=1}^{t}\underline{KN}_{B_1\cup\{b_{m_1},b_{m_2},\cdots,b_{m_k}\}}(D_j,X_{s+k}).\end{aligned}$$

其中 $X_{s+k}=D_j-X_k-\bigcup_{h=1}^{k-1}\underline{KN}_{B_1\cup\{b_{m_1},b_{m_2},\cdots,b_{m_k}\}}(D_j,X_{s+h})$，$X_k=D_j-\bigcup_{h=1}^{k-1}\underline{KN}_{B_h}(D_j,X_h)$。

因此 $\bigcup_{k=1}^{s}\underline{KN}_{B_1}(D_j,X_k)\subseteq\bigcup_{k=1}^{s+t}\underline{KN}_{B_2}(D_j,X_k)$。

又有 $S_{B_i}\left(D_j,\overline{KN}_{B_i}(D_j,W_i)\right)-\bigcup_{k=1}^{i}\underline{KN}_{B_i}(D_j,X_k)\subseteq\bigcup_{k=1}^{t}\underline{KN}_{B_1\cup\{b_{m_1},b_{m_2},\cdots,b_{m_k}\}}(D_j,X_{s+k})$，

所以 $APOS_{B_1}(D_j)\subseteq APOS_{B_2}(D_j),\forall D_j\in U/D$。即 $APOS_{B_1}(D)\subseteq APOS_{B_2}(D)$。

定义 9 给定混合决策信息系统 $HDIS=(U,A=C\cup D,V,f)$。对任意 $B\subseteq C$，决策属性 D 对于属性子集 B 的依赖度为：

$$\gamma_B(D)=\frac{|APOS_B(D)|}{|U|} \tag{16}$$

性质 2 给定混合决策信息系统 $MDIS=(U,A=C\cup D,V,f)$。对任意 $B_{i-1},B_i\in I(C)$，且 $B_{i-1}\subseteq B_i$，

$\gamma_{B_{i-1}}^{\alpha}(D)$ 和 $\gamma_{B_i}^{\alpha}(D)$ 分别是 D 对于属性子集 B_{i-1} 和 B_i 的依赖度，则有 $\gamma_{B_{i-1}}^{\alpha}(D) \le \gamma_{B_i}^{\alpha}(D)$。

性质 2 表明，随着属性子集的增加，决策属性集 D 对于属性子集 B 的依赖度逐渐增大。这意味着在已有的属性子集中添加一个新属性，决策属性 D 对于属性子集 B 的依赖度的值不会减小。这为下文构造基于加权 k 近邻的属性约简算法提供了理论依据。

定义 10 给定混合决策信息系统 $HDIS=(U,A=C\cup D,V,f)$，对任意 $B\subseteq C$，若 B 满足 $\gamma_B^{\delta}(D)=\gamma_C^{\delta}(D)$，且对任意 $a\in B$，$\gamma_{B-\{a\}}^{\delta}(D)\neq\gamma_C^{\delta}(D)$，称 B 是混合决策信息系统的一个约简。

定义 11 给定混合决策信息系统 $HDIS=(U,A=C\cup D,V,f)$，对任意 $B\subseteq C$，$a\in C-B$，则属性 a 关于属性子集 B 的重要度定义为：

$$Sig(a,B,D)=\gamma_{B\cup\{a\}}^{\alpha}(D)-\gamma_B^{\alpha}(D) \tag{17}$$

我们可以得到 $0\le Sig(a,B,D)\le 1$。

根据上述结论，可以构造下述启发式属性约简选择算法。该算法从空集开始，在每个步骤中添加一个重要度最大的属性，直到依赖度不变。具体过程见算法 1。

算法 1 加权 k 近邻的属性约简

输入：混合决策信息系统 $MDIS=(U,C,f,D,g)$；

输出：约简结果 red。

Step 1 设置初始值 $B_k\leftarrow\varnothing$；

Step 2 $k=0$；

Step 3 计算属性全集的 $\gamma_C^{\alpha}(D)$；

Step 4 从 $C-B_k$ 中选择 $Sig(a,B_k,D)$ 最大的属性 a；

Step 5 $k\leftarrow k+1$，$B_k\leftarrow B_{k-1}\cup\{a\}$，计算 $\gamma_{B_k}^{\alpha}(D)$；

Step 6 若 $\gamma_{B_k}^{\alpha}(D)\ge\gamma_C^{\alpha}(D)$，则转步骤 8；

Step 7 否则转步骤 4；

Step 8 $red\leftarrow B_k$

Step 9 输出约简结果 red。

4. 实验

4.1. 实验环境

为验证本文算法的可行性，本文选取 UCI 上的 3 个数据集进行实验，3 组数据均是含有数值型和分类型属性的混合决策数据集，但是部分存在缺失，在本文实验前已对数据进行预处理，表 2 所示为数据集的基本信息。

本文借鉴文献[20]，采用 KNN (K = 3)和 RBF-SVM 两个经典分类器对属性约简算法进行了评价。在实验中，SVM 的所有参数都被设置为默认值。实验比较采用十倍交叉验证进行。也就是说，一个数据集被随机分成十个部分。对于每个循环，一个部分用于测试，其他部分用于执行属性约简。对简化后的测试数据，计算了 3NN 和 SVM 的分类精度。经过 10 次循环后，计算分类精度的平均值作为最终结果。

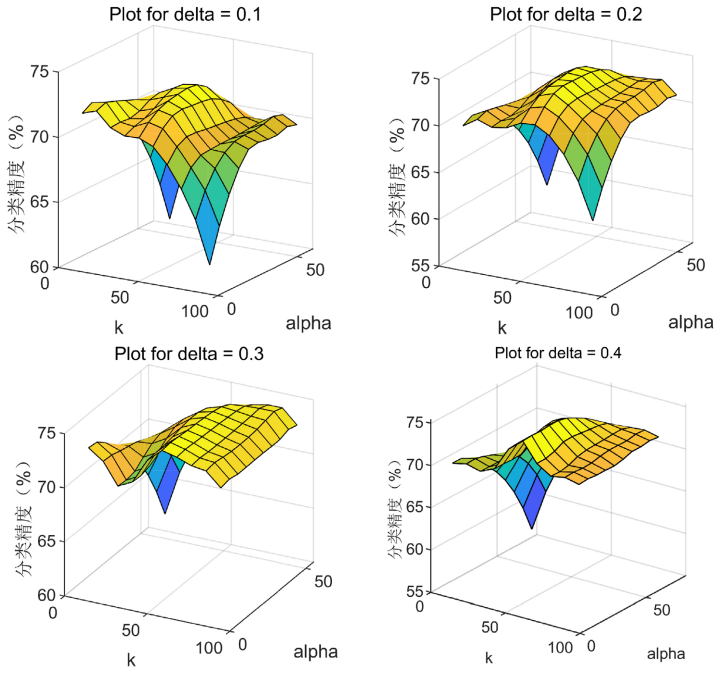
Table 2. Experimental data sets
表 2. 实验数据集

数据集	对象数	数值型属性数	符号型属性数	决策类数
German	1000	7	13	2
Heart1	270	7	6	2
Heart2	303	6	7	5

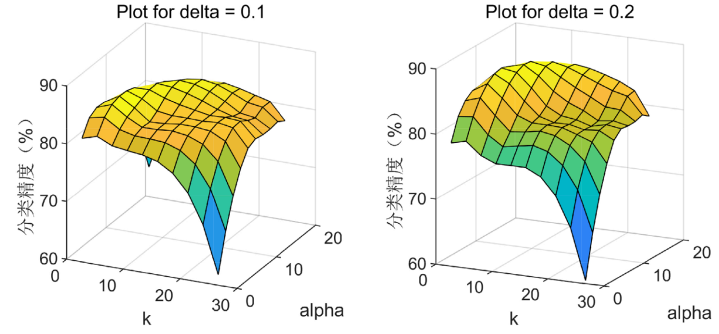
4.2. 参数分析

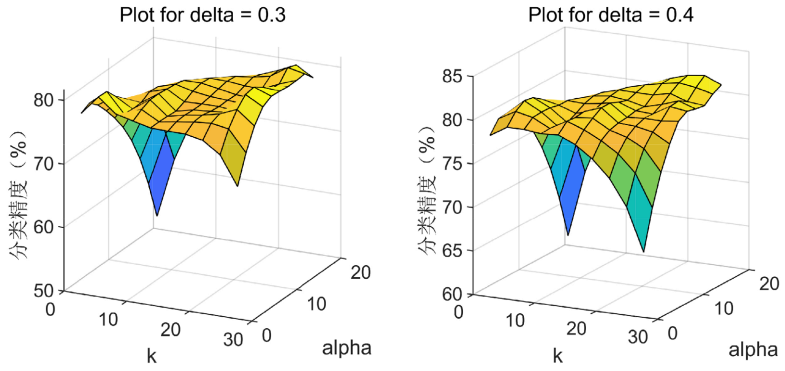
在本文提出的算法中，包含了邻域半径 δ， k 和阈值 α 三个参数。由于不同的参数取值对约简结果有不同的影响，为分析在不同数据集下 δ， k 和阈值 α 对分类准确率的影响，本组实验以 0.1 为步长在区间 $[0.1,0.4]$ 内选取邻域半径 δ，以 0.01 N 为步长在区间 $[0.01N,0.1N]$ 内选取 k，以 1 为步长在区间 $\left[1,\frac{2k}{3}\right]$ 内选取 α，评估 δ， k 和 α 在不同取值下的约简结果在 KNN 和 SVM 两个分类器上对数据集分类精度的影响。

图 1 为数据集在 KNN 分类器上的分类精度；图 2 为数据集在 SVM 分类器上的分类精度。

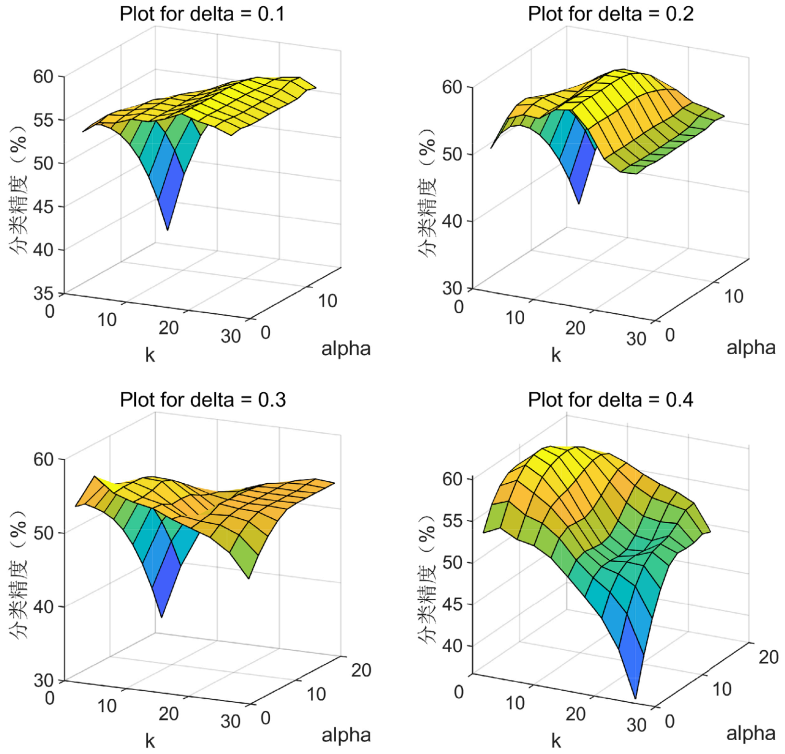


(a) German



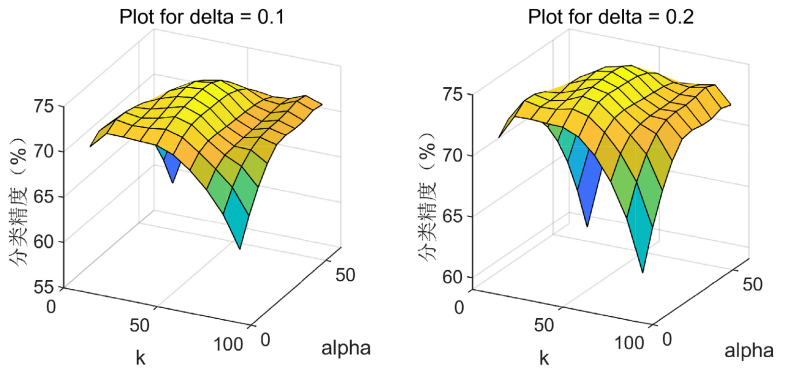


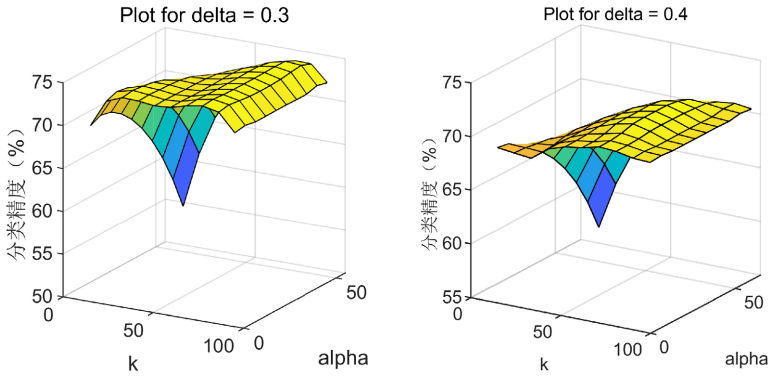
(b) Heart1



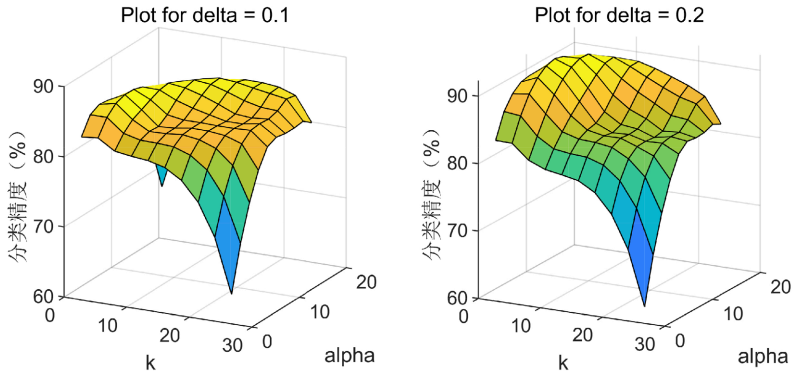
(c) Heart2

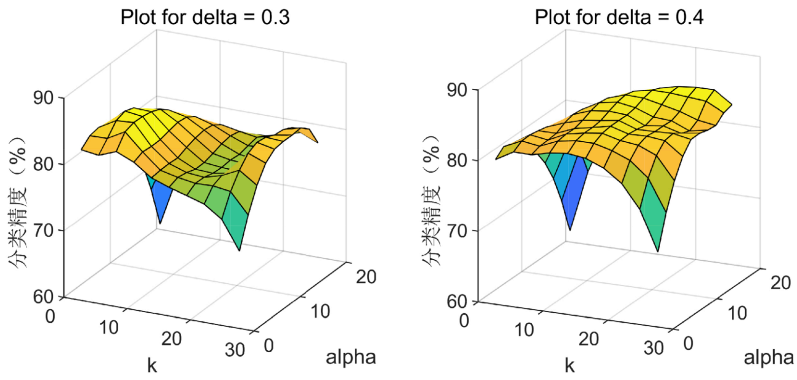
Figure 1. KNN classification accuracy
图 1. KNN 分类器分类精度



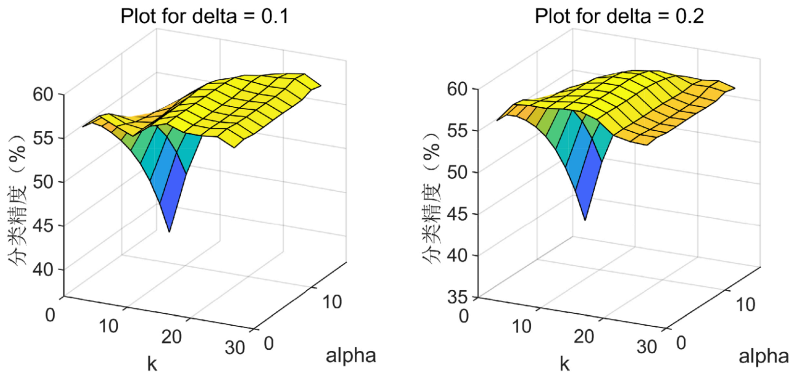


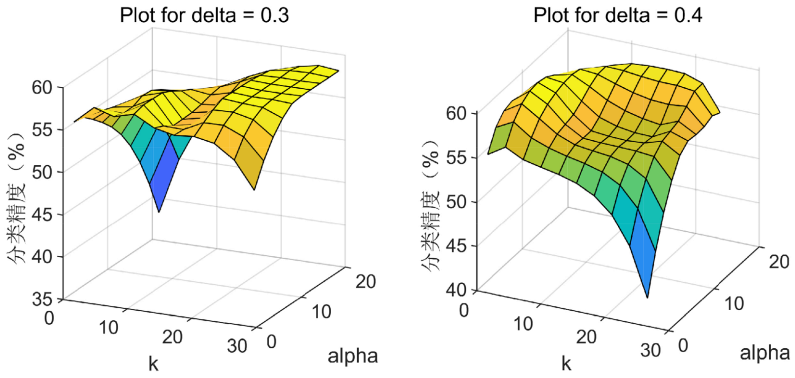
(a) German





(b) Heart1





(c) Heart2

Figure 2. SVM classification accuracy
图 2. SVM 分类器分类精度

4.3. 对比分析

本实验将所提出的属性约简算法与文献[11]的 NRS 属性约简算法和文献[22]的 NNRS 属性约简算法分别进行实验，通过属性约简的长度和约简结果的分类精度来比较各个算法。

Table 3. Comparison of the length for attribute reductions
表 3. 属性约简长度比较

数据集	原始数据	NRS		NNRS		本文算法	
		KNN	SVM	KNN	SVM	KNN	SVM
German	20	11	12	12	11	**10**	**10**
Heart1	13	12	11	10	12	**7**	**9**
Heart2	13	10	10	10	9	**9**	**7**

表 3 给出约简结果属性数量的比较。从表中可以看出，本文算法相对于原始属性集有效地删除了冗余属性。相较于其他算法，本文算法在所选数据集下得到的属性数量更少一些。这是由于本文的算法充分考虑了边界域的有效信息，在属性约简时会提高一些属性的选择能力，所以结果较好一些。

Table 4. KNN classification accuracy of attribute reductions
表 4. 属性约简的 KNN 分类精度

数据集	原始数据	NRS	NNRS	本文算法
German	70.3000	71.3000	71.5000	**72.7000**
Heart1	78.5185	79.6296	80.3704	**81.8519**
Heart2	50.6552	55.092	55.8034	**56.8276**

Table 5. SVM classification accuracy of attribute reductions
表 5. 属性约简的 SVM 分类精度

数据集	原始数据	NRS	NNRS	本文算法
German	73.7000	74.4000	73.9000	**74.6000**
Heart1	81.4815	83.3333	84.0741	**84.8148**
Heart2	56.023	57.5287	58.092	**58.1954**

通过表 4 和表 5 观察可得，本文算法不管是在 KNN 分类器还是 SVM 分类器上都有较高的分类精度，因为本文的算法考虑了边界域中对象的有效信息，在属性约简时会提高一些属性的选择能力，有效地删除冗余的、无关的属性，并且不会降低邻域决策系统原有的分类性能，效果更好一些。

5. 结束语

本文针对混合决策信息系统，引入 k 近邻来描述混合决策信息系统中对象的粒化，通过定义边界域中对象的评价函数来判断其可分性，定义具有单调性的近似正域来反映邻域决策系统的分类能力，基于此定义新的属性依赖度，设计出一种启发式属性约简算法。本文主要针对完备的混合决策信息系统，之后将进一步研究不完备混合决策信息系统的属性约简。

参考文献

[1] Qian, Y., Liang, J., Pedrycz, W. and Dang, C. (2010) Positive Approximation: An Accelerator for Attribute Reduction in Rough Set Theory. *Artificial Intelligence*, **174**, 597-618. https://doi.org/10.1016/j.artint.2010.04.018

[2] Lin, Y., Hu, Q., Liu, J., Li, J. and Wu, X. (2017) Streaming Feature Selection for Multilabel Learning Based on Fuzzy Mutual Information. *IEEE Transactions on Fuzzy Systems*, **25**, 1491-1507. https://doi.org/10.1109/tfuzz.2017.2735947

[3] Zhu, P., Zhu, W., Hu, Q., Zhang, C. and Zuo, W. (2017) Subspace Clustering Guided Unsupervised Feature Selection. *Pattern Recognition*, **66**, 364-374. https://doi.org/10.1016/j.patcog.2017.01.016

[4] Yao, S., Xu, F., Zhao, P., *et al*. (2017) Feature Selection Algorithm Based on Neighborhood Valued Tolerance Relation Rough Set Model. *Pattern Recognition and Artificial Intelligence*, **30**, 416-428.

[5] Pawlak, Z. (1982) Rough Sets. *International Journal of Computer & Information Sciences*, **11**, 341-356. https://doi.org/10.1007/bf01001956

[6] Guan, Y. and Wang, H. (2006) Set-Valued Information Systems. *Information Sciences*, **176**, 2507-2525. https://doi.org/10.1016/j.ins.2005.12.007

[7] Leung, Y., Fischer, M.M., Wu, W. and Mi, J. (2008) A Rough Set Approach for the Discovery of Classification Rules in Interval-Valued Information Systems. *International Journal of Approximate Reasoning*, **47**, 233-246. https://doi.org/10.1016/j.ijar.2007.05.001

[8] Dubois, D. and Prade, H. (1990) Rough Fuzzy Sets and Fuzzy Rough Sets. *International Journal of General Systems*, **17**, 191-209. https://doi.org/10.1080/03081079008935107

[9] Zeng, A., Li, T., Liu, D., Zhang, J. and Chen, H. (2015) A Fuzzy Rough Set Approach for Incremental Feature Selection on Hybrid Information Systems. *Fuzzy Sets and Systems*, **258**, 39-60. https://doi.org/10.1016/j.fss.2014.08.014

[10] HU, Q., YU, D. and XIE, Z. (2008) Neighborhood Classifiers. *Expert Systems with Applications*, **34**, 866-876. https://doi.org/10.1016/j.eswa.2006.10.043

[11] Hu, Q., Yu, D., Liu, J. and Wu, C. (2008) Neighborhood Rough Set Based Heterogeneous Feature Subset Selection. *Information Sciences*, **178**, 3577-3594. https://doi.org/10.1016/j.ins.2008.05.024

[12] Fan, X., Zhao, W., Wang, C. and Huang, Y. (2018) Attribute Reduction Based on Max-Decision Neighborhood Rough Set Model. *Knowledge-Based Systems*, **151**, 16-23. https://doi.org/10.1016/j.knosys.2018.03.015

[13] Chen, H., Li, T., Cai, Y., Luo, C. and Fujita, H. (2016) Parallel Attribute Reduction in Dominance-Based Neighborhood Rough Set. *Information Sciences*, **373**, 351-368. https://doi.org/10.1016/j.ins.2016.09.012

[14] Wang, C., Shao, M., He, Q., Qian, Y. and Qi, Y. (2016) Feature Subset Selection Based on Fuzzy Neighborhood Rough Sets. *Knowledge-Based Systems*, **111**, 173-179. https://doi.org/10.1016/j.knosys.2016.08.009

[15] Shu, W., Qian, W. and Xie, Y. (2020) Incremental Feature Selection for Dynamic Hybrid Data Using Neighborhood Rough Set. *Knowledge-Based Systems*, **194**, Article 105516. https://doi.org/10.1016/j.knosys.2020.105516

[16] Hu, M., Tsang, E.C.C., Guo, Y., Chen, D. and Xu, W. (2021) A Novel Approach to Attribute Reduction Based on Weighted Neighborhood Rough Sets. *Knowledge-Based Systems*, **220**, Article 106908. https://doi.org/10.1016/j.knosys.2021.106908

[17] Sewwandi, M.A.N.D., Li, Y. and Zhang, J. (2024) Granule-Specific Feature Selection for Continuous Data Classification Using Neighborhood Rough Sets. *Expert Systems with Applications*, **238**, Article 121765. https://doi.org/10.1016/j.eswa.2023.121765

[18] Chen, H., Li, T., Fan, X. and Luo, C. (2019) Feature Selection for Imbalanced Data Based on Neighborhood Rough Sets. *Information Sciences*, **483**, 1-20. https://doi.org/10.1016/j.ins.2019.01.041

[19] Zhang, X., Mei, C., Chen, D. and Li, J. (2016) Feature Selection in Mixed Data: A Method Using a Novel Fuzzy Rough Set-Based Information Entropy. *Pattern Recognition*, **56**, 1-15. https://doi.org/10.1016/j.patcog.2016.02.013

[20] Wang, C., Shi, Y., Fan, X. and Shao, M. (2019) Attribute Reduction Based on K-Nearest Neighborhood Rough Sets. *International Journal of Approximate Reasoning*, **106**, 18-31. https://doi.org/10.1016/j.ijar.2018.12.013

[21] 张文修, 梁怡, 吴伟志. 信息系统与知识发现[M]. 北京: 科学出版社, 2003.

[22] Wang, N. and Zhao, E. (2024) A New Method for Feature Selection Based on Weighted k-Nearest Neighborhood Rough Set. *Expert Systems with Applications*, **238**, Article 122324. https://doi.org/10.1016/j.eswa.2023.122324