文章编号：1672-3961(2026)01-0035-14 DOI：10.6040/j.issn.1672-3961.0.2024.200

基于数据权重的鲁棒性模糊粗糙集与属性约简

李璐[1]，王鑫[2]
（1.安徽建筑大学数理学院，安徽 合肥 230601；2.安徽建筑大学经济与管理学院，安徽 合肥 230601）

摘要：提出一种鲁棒性模糊粗糙集模型和属性约简算法。考虑数据样本的局部密度并进行量化，利用量化结果评估样本在数据集整体中的噪声程度；通过噪声程度衡量样本的权重，定义一种样本集之间的距离度量，并将样本之间的模糊相似性替换为样本集之间的模糊相似性，提升模糊相似关系的鲁棒性，建立一种鲁棒性模糊粗糙集模型；基于所提出的鲁棒性模糊粗糙集定义属性与类之间的依赖度，以评估属性子集的显著性，并设计一种鲁棒性模糊粗糙集的属性约简算法。试验结果表明，所设计的属性约简算法比现有的算法具有更强的鲁棒性和优越性。

关键词：模糊粗糙集；鲁棒性；噪声数据；样本权重；模糊依赖度；属性约简
中图分类号：TP181 **文献标志码**：A
引用格式：李璐，王鑫. 基于数据权重的鲁棒性模糊粗糙集与属性约简[J]. 山东大学学报(工学版)，2026，56(1)：35-48.
LI Lu，WANG Xin. Robust fuzzy rough set and attribute reduction based on data weights[J]. Journal of Shandong University（Engineering Science），2026，56(1)：35-48.

Robust fuzzy rough set and attribute reduction based on data weights

LI Lu[1]，WANG Xin[2]
（1. School of Mathematics and Physics，Anhui Jianzhu University，Hefei 230601，Anhui，China；2. School of Economics and Management，Anhui Jianzhu University，Hefei 230601，Anhui，China）

Abstract：A robust fuzzy rough set model and attribute reduction algorithm were proposed in this paper. This article considered the local density of data samples and quantifies them，using the quantization results to evaluate the noise level of the samples in the overall dataset. The weight of the samples was measured by the level of noise，and a distance measure between sample sets was defined. The fuzzy similarity between samples was replaced by the fuzzy similarity between sample sets，which improved the robustness of the fuzzy similarity and established a robust fuzzy rough set model. Based on the proposed robust fuzzy rough set，the dependency between attributes and classes was defined to evaluate the significance of attribute subsets，and a robust fuzzy rough set attribute reduction algorithm was designed. The experimental results showed that the designed attribute reduction algorithm had stronger robustness and superiority than existing algorithms.

Keywords：fuzzy rough set；robustness；noise data；sample weight；fuzzy dependency degree；attribute reduction

0 引言

属性约简又称特征选择，是从原始数据集中选择有效的属性子集，以提升后续学习任务的性能和质量[1-2]。当前属性约简方法已广泛应用于聚类学习、模式识别和数据挖掘等领域[3-4]。文献[5]提出的模糊粗糙集理论通过分析数值型数据中的模糊不确定性，成为属性约简的一种重要工具和方法。

然而，由于模糊粗糙集在模型计算过程中对噪声数据具有很强的敏感性，导致它的应用性存在很大不足。近年来关于模糊粗糙集的鲁棒性研究受到较多关注，基于不同的鲁棒原理提出多种鲁棒性模糊粗糙集模型，如变量精度测度、鲁棒距离测度、鲁棒统计测度和概率分布等。文献[6]将变精度思想加入模糊粗糙集中，提出变精度模糊粗糙集，该

收稿日期：2024-08-10
基金项目：安徽省高校省级自然科学研究项目基金资助项目(KJ2021JD20)
第一作者简介：李璐(1980—)，女，副教授，硕士，主要研究方向为数据挖掘、粗糙集。E-mail：lilu@ ahjzu.edu.cn

模型成为模糊粗糙集容噪性的一种典型模型;文献[7]针对数据的噪声情况,提出多种鲁棒性模糊粗糙集模型;文献[8]对模糊相似关系下的变精度模糊粗糙集进行推广和改进,提升模糊粗糙集的容噪性能;文献[9]通过样本的概率密度定义对象的模糊相似度,提出一种新的鲁棒性模糊粗糙集属性约简;文献[10]提出一种特殊距离测度的模糊相似度关系,并重建模糊粗糙集的上下近似集,同时为减少噪声的影响,采用最近邻算子定义决策对属性的依赖性以及属性约简;文献[11]通过定义样本的局部密度评估样本的噪声程度,并重新定义模糊粗糙集的鲁棒性,同时也设计出对应的特征选择算法;文献[12]将重叠函数与变精度参数引入模糊粗糙集,提出基于重叠函数的变精度模糊粗糙集,进一步提升模糊粗糙集的鲁棒性;文献[13]对变精度模糊粗糙集进行类似的推广,并提出一种鲁棒性三支决策模型;在文献[7]的基础上,文献[14]提出改进的核模糊相似关系,提出一种新的鲁棒性核模糊粗糙集模型以及特征选择算法;文献[15]针对优势关系的含噪声数据,提出鲁棒性模糊优势粗糙集模型,通过自适应属性权重的方法设计优势关系的属性约简。

一些模糊粗糙集模型采用不同的鲁棒性原理处理噪声。这些措施可以总结为利用数据分布感知噪声[6,9,11-13,15]和改进模糊相似度自适应噪声[7-8,10,14],然而这种处理策略也存在一定的不足。利用数据分布感知噪声仅局限于数据的整体性,无法精确到单个数据样本的周边分布情况,即单个样本的噪声程度被整体数据中和,通过改进一些模糊相似度距离方法进行评估噪声数据,都是基于样本个体进行度量,而忽略周边样本对模糊相似度的影响。

针对目前鲁棒性模糊粗糙集存在的一些不足,本研究提出一种改进的鲁棒性模糊粗糙集模型。首先,针对数据样本的分布情况,定义单个样本的局部密度概念,对数据集的每个样本评估出对应的噪声程度,用于定义样本的权重;其次,将样本之间的距离度量替换为样本集之间的距离度量,并且样本集之间的距离通过每个对象的加权和实现,其中对象的权重就使用所定义的样本局部密度因子,使得距离度量结果能够自适应样本的噪声程度,利用这种新的距离度量刻画对象的模糊相似关系,提出一种鲁棒性模糊粗糙集模型;最后,根据提出的改进模型,定义属性与类之间的鲁棒性模糊粗糙依赖度,并设计出一种属性约简算法,该算法可以有效降低噪声对选定属性子集的影响,通过数值试验验证了该模型和算法的鲁棒性和优越性。

1 基本理论

设(U,C,D)为一个模糊决策信息系统。其中U称为非空论域,是一个对象(样本)的集合;C是条件属性的集合,$C=C_c\cup C_n$(C_n是数值型属性集,C_c是离散型属性集);D称为决策属性集,$D=\{d\}$(d是决策属性)。$v_a(x)$表示对象$x(x\in U)$在属性$a(a\in C)$的属性值。$\tilde{X}$称为U上的模糊集,$\tilde{X}(x)$表示对象$x(x\in U)$对于$\tilde{X}$的隶属度。设$F(U)$表示U的模糊幂集,即$\tilde{X}\in F(U)$,精确集是模糊集的一种特殊情况。设$N:[0,1]\to[0,1]$是一个单调递减函数,如果$N(0)=1$,$N(1)=0$并且$N(N(t))=t(t\in[0,1])$,那么N又称为对合负数映射[16]。

给定(U,C,D)和$B\subseteq C$,$\tilde{R}_B:U\times U\to[0,1]$是属性集$B$在$U$上诱导的模糊关系。对于模糊集$\tilde{X}\in F(U)$,$\forall x,y\in U$,$\tilde{R}_B(x,y)$表示对象$x$和$y$的模糊相似度,那么模糊集$\tilde{X}\in F(U)$在$\tilde{R}_B$的模糊粗糙集下、上近似分别定义为

$$\underline{\tilde{R}_B}\tilde{X}(x)=\inf_{y\in U}\max\{N(\tilde{R}_B(x,y)),\tilde{X}(y)\},$$

$$\overline{\tilde{R}_B}\tilde{X}(x)=\sup_{y\in U}\min\{\tilde{R}_B(x,y),\tilde{X}(y)\},$$

式中$\underline{\tilde{R}_B}\tilde{X}$和$\overline{\tilde{R}_B}\tilde{X}$称为模糊集$\tilde{X}$的模糊粗糙集[17]。

论域U在决策属性集D上的决策类划分表示为$U/D=\{L_1,L_2,\cdots,L_r\}$。在模糊粗糙集中$D$关于属性子集$B\subseteq C$的模糊正区域定义为$\mathrm{POS}_B(D)=\cup_{L\in U/D}\underline{\tilde{R}_B}L$,$D$关于属性子集$B\subseteq C$的模糊依赖度定义为$\gamma_B(D)=\dfrac{|\mathrm{POS}_B(D)|}{|U|}$,这里的$|\circ|$表示集合的基数。

2 基于样本权重策略的鲁棒性模糊粗糙集模型

经典模糊粗糙集模型中对象之间的近似算子采用最大值和最小值的运算策略[2]。该方法对分类数据中的噪声非常敏感,例如对于含离群点的噪声数据,最大值和最小值运算直接拉高和拉低了对象之间的相似度,影响对象之间的模糊性刻画。本节将提出一种样本权重的概念,对于不同噪声程度的数据进行权重刻画,利用对象的权重重新定义对象之间的模糊相似性,最终提出一种鲁棒性的模糊

粗糙集模型。

2.1　基于噪声程度的样本权重评估

在邻域粗糙集模型中,邻域是对象周围相近区域对象的集合,利用对象间的距离函数衡量不同对象之间的相似性[18]。针对数值型和离散型的混合数据,对象之间的混合欧氏距离函数定义如下。

定义 1　给定(U,C,D)和$C=C_c\cup C_n$, $B\subseteq C$的混合欧氏距离定义为

$$\Delta_B(x,y)=\sqrt{\sum_{\forall a\in B} d_a(x,y)^2},\ x,y\in U,$$

其中

$$d_a(x,y)=\begin{cases}\mathrm{abs}(v_a(x)-v_a(y)),\ a\in C_n\\ 1,\ a\in C_c,\ v_a(x)=v_a(y)\\ 0,\ a\in C_c,\ v_a(x)\neq v_a(y)\end{cases},$$

式中 abs(∘)表示绝对值运算。

通过定义的距离函数,接下来可以计算样本对象的邻域,根据邻域中的对象数量可以评估当前样本对象在整个论域中的分布情况以及局部密度情况。

定义 2　给定(U,C,D)和$B\subseteq C$,决策类划分$U/D=\{L_1,L_2,\cdots,L_r\}$,对于$\forall x\in L_i$, $\phi_k(x)$表示对象 x 的 k 个最近邻对象,$\varphi_\delta(x)$表示 $\phi_k(x)$中对象 x 以 δ 为半径的邻域对象($\varphi_\delta(x)$中不包含对象 x)。定义$\forall x\in L_i$的样本局部密度为

$$l_B(x)=\frac{\sum_{y\in\varphi_\delta(x)}\Delta_B(x,y)}{\sum_{y\in\phi_k(x)}\Delta_B(x,y)}。$$

样本局部密度满足$0\leqslant l_B(x)\leqslant 1$。如果对象 x 为非噪声数据,例如满足 $\varphi_\delta(x)=\phi_k(x)$,那么此时$l_B(x)=1$。如果对象 x 是一个噪声数据,那么意味着 $\varphi_\delta(x)\subset\phi_k(x)$,即 $l_B(x)<1$;随着噪声程度加剧,$l_B(x)$越来越小,当 $\varphi_\delta(x)=\varnothing$时,$l_B(x)=0$。因此,可以利用决策类内部对象的局部密度反映样本的分布信息。

然而,当数据集中对象的密度分布变化较大时,局部密度并不能直接反映样本是否为噪声。因此,接下来引入局部密度因子识别噪声。

定义 3　给定(U,C,D)和$B\subseteq C$,对于$\forall x\in L_i$的局部密度为$l_B(x)$,定义对象 x 的局部密度因子为

$$\chi_B(x)=\frac{\sum_{y\in\phi_k(x)}\frac{l_B(y)}{l_B(x)}}{|\phi_k(x)|},\ (l_B(x)\neq 0)。$$

对象的信息与其相邻对象的信息密切相关。在定义 3 中,将对象的局部密度与相邻对象的局部密度进行比较,并利用它们的比值确定当前对象是否为噪声数据。如果当前对象的数据分布与相邻对象的数据分布有显著差异,即可以认为对象 x 是一个噪声数据,那么对象 x 的局部密度因子将大于 1 或小于 1,即对象的局部密度因子偏离 1 的程度越大,是噪声数据的程度越明显。

举例　给定一个信息系统,如表 1 所示,其中论域$U=\{x_1,x_2,\cdots,x_{20}\}$,属性集$C=\{a,b\}$,$a$ 和 b 为属性集 C 的两个属性。

表 1　信息系统

Table 1　Information systems

论域	a	b
x_1	0.26	0.58
x_2	0.22	0.71
x_3	0.08	0.67
x_4	0.23	0.54
x_5	0.14	0.58
x_6	0.19	0.42
x_7	0.18	0.58
x_8	0.25	0.63
x_9	0.25	0.58
x_{10}	0.17	0.54
x_{11}	0.14	0.46
x_{12}	0.26	0.53
x_{13}	0.25	0.88
x_{14}	0.31	0.92
x_{15}	0.17	0.46
x_{16}	0.20	0.50
x_{17}	0.27	0.53
x_{18}	0.17	0.46
x_{19}	0.22	0.58
x_{20}	0.21	0.63

考察对象 x_5、x_{14}和 x_{19},设 $\delta=0.08$, $k=4$,根据定义 2 可以得到

$$\varphi_\delta(x_5)=\{x_7,x_{10},x_{19}\},$$
$$\phi_k(x_5)=\{x_7,x_{10},x_{19},x_{20}\},$$
$$\varphi_\delta(x_{14})=\{x_{13}\},$$
$$\phi_k(x_{14})=\{x_{13},x_2,x_8,x_{20}\},$$
$$\varphi_\delta(x_{19})=\{x_1,x_4,x_5,x_7,x_8,x_9,x_{10},x_{12},x_{17},x_{20}\},$$
$$\phi_k(x_{19})=\{x_9,x_1,x_4,x_7\},$$

则对象 x_5、x_{14}和 x_{19}的样本局部密度分别为

$$l_C(x_5)=\frac{\Delta_C(x_5,x_7)+\Delta_C(x_5,x_{10})+\Delta_C(x_5,x_{19})}{\Delta_C(x_5,x_7)+\Delta_C(x_5,x_{10})+\Delta_C(x_5,x_{19})+\Delta_C(x_5,x_{20})}=0.65,$$

$$l_C(x_{14})=\frac{\Delta_C(x_{14},x_{13})}{\Delta_C(x_{14},x_{13})+\Delta_C(x_{14},x_2)+\Delta_C(x_{14},x_8)+\Delta_C(x_{14},x_{20})}=0.07,$$

$$l_C(x_{19})=$$

$$\frac{\Delta_C(x_{19},x_9)+\Delta_C(x_{19},x_1)+\Delta_C(x_{19},x_4)+\Delta_C(x_{19},x_7)}{\Delta_C(x_{19},x_9)+\Delta_C(x_{19},x_1)+\Delta_C(x_{19},x_4)+\Delta_C(x_{19},x_7)}=1.00。$$

进一步地,根据定义 3 可以得到对象 x_5、x_{14}和 x_{19}的局部密度因子分别为

$$\chi_C(x_5)=\frac{\frac{l_C(x_7)}{l_C(x_5)}+\frac{l_C(x_{10})}{l_C(x_5)}+\frac{l_C(x_{19})}{l_C(x_5)}+\frac{l_C(x_{20})}{l_C(x_5)}}{|\phi_k(x_5)|}=1.54,$$

$$\chi_C(x_{14})=\frac{\frac{l_C(x_{13})}{l_C(x_{14})}+\frac{l_C(x_2)}{l_C(x_{14})}+\frac{l_C(x_8)}{l_C(x_{14})}+\frac{l_C(x_{20})}{l_C(x_{14})}}{|\phi_k(x_{14})|}=7.46,$$

$$\chi_C(x_{19})=\frac{\frac{l_C(x_9)}{l_C(x_{19})}+\frac{l_C(x_1)}{l_C(x_{19})}+\frac{l_C(x_4)}{l_C(x_{19})}+\frac{l_C(x_7)}{l_C(x_{19})}}{|\phi_k(x_{19})|}=1.00。$$

对象的局部密度因子偏离 1.00 的程度越大,是噪声数据的程度越明显。将表 1 中的所有对象绘制成散点图,如图 1 所示。由图 1 可以直观地观察到,对象 x_5、x_{14}和 x_{19}的噪声程度与 $\chi_C(x_5)$、$\chi_C(x_{14})$和 $\chi_C(x_{19})$数值结果大致相符合。

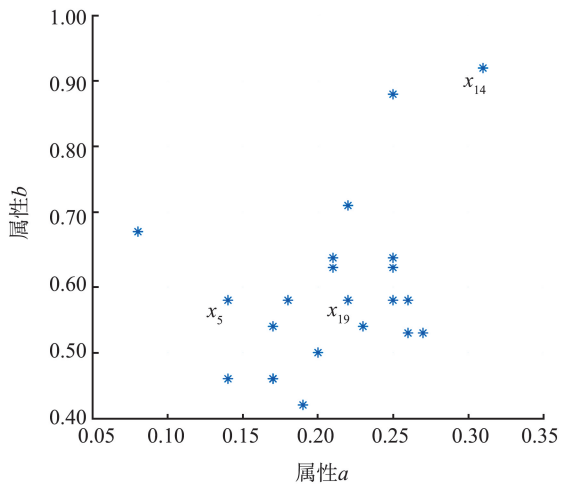


图 1 信息系统对象的散点图

Fig.1 Scatter plot of information system objects

综合定义 2、3 可以看出,样本对象局部密度因子的主要计算量集中在样本 k 近邻对象的搜索,因此计算复杂度较低。对于大规模样本,可以对样本集进行采样,利用采样之后的样本子集进行权重计算和属性约简。

定义 3 的局部密度因子描述了信息系统对象数据的噪声程度,接下来用来衡量对象在信息系统中的样本权重。

定义 4 给定(U,C,D)和 $B\subseteq C$, $\forall x\in U$ 的样本权重定义为

$$\omega_B(x)=\begin{cases}0, & |\varphi_\delta(x)|=0\\ 1, & |\varphi_\delta(x)|=k\\ \dfrac{1}{|\ln\chi_B(x)|+1}, & \text{其他}\end{cases},$$

下文中,在不引起混淆的情况下,将 $\omega_B(x)$ 简记为 ω_x,同时对象$x_i\in U$ 的样本权重 $\omega_B(x_i)$简记为 ω_i。

在定义 4 中,当对象 x_i 的$|\varphi_\delta(x_i)|=0$ 时,即对象 x_i 的 k 近邻范围内没有 δ 邻域对象,说明对象 x_i 是一个噪声程度比较大的数据,因此设定样本权重 $\omega_i=0$;当对象 x_i 的$|\varphi_\delta(x_i)|=k$ 时,即对象 x_i 的 k 近邻对象均为 δ 邻域对象,说明对象 x_i 处于样本密集区域,因此设定样本权重 $\omega_i=1$。处于前述二者中间区域时,$\chi_B(x_i)>0$,并且 $\chi_B(x_i)=1$ 满足$\varphi_\delta(x_i)=\phi_k(x_i)$,通过$\frac{1}{|\ln\chi_B(x_i)|+1}$可以衡量样本的权重。

2.2 基于样本权重的对象集距离

根据定义 4 所示的样本权重定义,接下来给出一种对象集之间的距离度量。

定义 5 给定(U,C,D)和 $B\subseteq C$,设对象 $x\in U$ 和对象集 $Y=\{y_1,y_2,\cdots,y_m\}$,y_i 的样本权重为 ω_i,定义 $x\in U$ 与对象集 Y 的样本权重距离为

$$d_B(x,Y)=\frac{\omega_1\Delta_B(x,y_1)+\omega_2\Delta_B(x,y_2)+\cdots+\omega_m\Delta_B(x,y_m)}{\sum_{i=1}^{m}\omega_i},$$

特别地,当$\sum_{i=1}^{m}\omega_i=0$时,定义 $d_B(x,Y)=+\infty$,表示对象 $x\in U$ 与对象集 Y 之间的距离为无限远。

在定义 5 中,对象集 $Y=\{y_1,y_2,\cdots,y_m\}$ 中每个对象的权重均为 1 时,即每个对象都不为噪声数据,那么此时

$$d_B(x,Y)=\frac{1}{m}\sum_{i=1}^{m}\Delta_B(x,y_i)。$$

文献[9]中作者提出一种概率粒距离度量方法,样本集之间的距离也是采用单个对象的加权和实现,其中每个对象的权重使用的是单个对象的概率分布,得到最终的距离度量结果,从而适应数据粒的分布。本研究与之类似,使用对象的噪声权重对距离度量结果进行加权,因此可以根据数据的噪声程度自适应地评估距离结果,可以看出所提出的样本权重距离具有很好的鲁棒性。

定义 6 给定(U,C,D)和 $B\subseteq C$,设对象集 $X=\{x_1,x_2,\cdots,x_n\}$ 和对象集 $Y=\{y_1,y_2,\cdots,y_m\}$,定义对象集 X 和对象集 Y 的样本权重距离为

$$d_B(X,Y)=\begin{cases}0, & X=Y\\ \boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{M}_B(X,Y)\boldsymbol{W}_2, & X\neq Y\end{cases},$$

同时

$$\boldsymbol{W}_1^{\mathrm{T}}=[\omega_1^1 \quad \omega_2^1 \quad \cdots \quad \omega_n^1],\ \omega_i^1=\frac{\omega_{x_i}}{\sum_{i=1}^{n}\omega_{x_i}},$$

$$\boldsymbol{W}_2^{\mathrm{T}}=[\omega_1^2 \quad \omega_2^2 \quad \cdots \quad \omega_m^2],\ \omega_i^2=\frac{\omega_{y_i}}{\sum_{i=1}^{m}\omega_{y_i}},$$

$$\boldsymbol{M}_B(X,Y)=\begin{bmatrix}d_{11} & d_{12} & \cdots & d_{1m}\\ d_{21} & d_{22} & \cdots & d_{2m}\\ \vdots & \vdots & & \vdots\\ d_{n1} & d_{n1} & \cdots & d_{nm}\end{bmatrix},$$

$d_{ij}=\Delta_B(x_i,y_j)$, $i=1,2,\cdots,n$, $j=1,2,\cdots,m$。

在定义 6 中,通过考虑样本权重的策略计算两对象集之间的距离,不同权重的对象对最终的对象集距离具有不同的贡献程度。

根据定义 6 可以进一步推导得到

$$d_B(X,Y)=\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{M}_B(X,Y)\boldsymbol{W}_2=$$

$$[\omega_1^1 \quad \omega_2^1 \quad \cdots \quad \omega_n^1]\begin{bmatrix}d_{11} & d_{12} & \cdots & d_{1m}\\ d_{21} & d_{22} & \cdots & d_{2m}\\ \vdots & \vdots & & \vdots\\ d_{n1} & d_{n2} & \cdots & d_{nm}\end{bmatrix}\begin{bmatrix}\omega_1^2\\ \omega_2^2\\ \vdots\\ \omega_m^2\end{bmatrix}=$$

$$\begin{bmatrix}\omega_1^1d_{11}+\omega_2^1d_{21}+\cdots+\omega_n^1d_{n1}\\ \omega_1^1d_{12}+\omega_2^1d_{22}+\cdots+\omega_n^1d_{n2}\\ \vdots\\ \omega_1^1d_{1m}+\omega_2^1d_{2m}+\cdots+\omega_n^1d_{nm}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\omega_1^2\\ \omega_2^2\\ \vdots\\ \omega_m^2\end{bmatrix}=$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_i^1\omega_j^2d_{ij}。$$

性质 1　设对象集 $X=\{x_1,x_2,\cdots,x_n\}$ 和对象集 $Y=\{y_1,y_2,\cdots,y_m\}$,满足 $d_B(X,Y)=d_B(Y,X)$。

证明　当 $X=Y$, $d_B(X,Y)=d_B(Y,X)=0$ 成立。当 $X\neq Y$ 时,根据定义 6 有

$$d_B(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_i^1\omega_j^2d_{ij},$$

$$d_B(Y,X)=\sum_{i=1}^{m}\sum_{j=1}^{n}\omega_i^2\omega_j^1d_{ij},$$

即 $d_B(X,Y)=d_B(Y,X)$,证毕。

性质 2　设对象集 $X=\{x_1,x_2,\cdots,x_n\}$ 和对象集 $Y=\{y_1,y_2,\cdots,y_m\}$,满足

$$d_B(X,Y)=\omega_1^2d_B(X,y_1)+\omega_2^2d_B(X,y_2)+\cdots+$$
$$\omega_m^2d_B(X,y_m)=\omega_1^1d_B(x_1,Y)+\omega_2^1d_B(x_2,Y)+\cdots+$$
$$\omega_n^1d_B(x_n,Y)。$$

证明

$$d_B(X,Y)=\begin{bmatrix}\omega_1^1d_{11}+\omega_2^1d_{21}+\cdots+\omega_n^1d_{n1}\\ \omega_1^1d_{12}+\omega_2^1d_{22}+\cdots+\omega_n^1d_{n2}\\ \vdots\\ \omega_1^1d_{1m}+\omega_2^1d_{2m}+\cdots+\omega_n^1d_{nm}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\omega_1^2\\ \omega_2^2\\ \vdots\\ \omega_m^2\end{bmatrix}=$$

$$[d_B(X,y_1) \quad d_B(X,y_2) \quad \cdots \quad d_B(X,y_m)]\begin{bmatrix}\omega_1^2\\ \omega_2^2\\ \vdots\\ \omega_m^2\end{bmatrix}=$$

$$\omega_1^2d_B(X,y_1)+\omega_2^2d_B(X,y_2)+\cdots+\omega_m^2d_B(X,y_m)=$$

$$d_B(X,Y)=\begin{bmatrix}\omega_1^1\\ \omega_2^1\\ \vdots\\ \omega_n^1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\omega_1^2d_{11}+\omega_2^2d_{12}+\cdots+\omega_m^2d_{1m}\\ \omega_1^2d_{21}+\omega_2^2d_{22}+\cdots+\omega_m^2d_{2m}\\ \vdots\\ \omega_1^2d_{n1}+\omega_2^2d_{n2}+\cdots+\omega_m^2d_{nm}\end{bmatrix}=$$

$$[\omega_1^1 \quad \omega_2^1 \quad \cdots \quad \omega_n^1]\begin{bmatrix}d_B(x_1,Y)\\ d_B(x_2,Y)\\ \vdots\\ d_B(x_n,Y)\end{bmatrix}=$$

$$\omega_1^1d_B(x_1,Y)+\omega_2^1d_B(x_2,Y)+\cdots+\omega_n^1d_B(x_n,Y),$$

证毕。

对于两个对象集之间的距离度量,常用的有最小距离、最大距离以及平均距离[4],然而这 3 种距离度量对噪声数据非常敏感。定义 6 中利用样本噪声程度的权重参与对象集之间距离的度量,噪声程度小的对象对计算距离的贡献程度较大,噪声程度大的对象对计算距离的贡献程度较小,因此对度量结果具有很好的鲁棒性。同时也为粗糙近似空间中粒与粒之间的距离评估提供一种新的解决方案。

2.3　鲁棒性模糊粗糙集模型

本节提出一种样本权重距离度量的鲁棒性模糊粗糙集模型。

定义 7　给定 (U,C,D) 和 $B\subseteq C$, $\forall x,y\in U$ 对应的邻域类分别为 $\varphi_\delta(x)$ 和 $\varphi_\delta(y)$,定义鲁棒性模糊相似度

$$\tilde{R}_B(x,y)=\exp(-d_B(\varphi_\delta(x),\varphi_\delta(y))),$$

式中, $\exp(\cdot)$ 为自然指数函数。由于 $d_B(\varphi_\delta(x),\varphi_\delta(y))\geqslant 0$,因此 $0<\tilde{R}_B(x,y)\leqslant 1$。同时,由于 $d_B(\varphi_\delta(x),\varphi_\delta(y))=d_B(\varphi_\delta(y),\varphi_\delta(x))$,因此 $\tilde{R}_B(x,y)=\tilde{R}_B(y,x)$ 成立。

定义 8　给定 (U,C,D) 和 $B\subseteq C$,对于模糊集 $\tilde{X}\in F(U)$, $\tilde{R}_B(x,y)$ 表示对象 x 和 y 的鲁棒性模糊相似度,模糊集 $\tilde{X}$ 在 $\tilde{R}_B$ 下的鲁棒性模糊粗糙集下、

上近似分别定义为

$$\underline{\tilde{R}}_B\tilde{X}(x)=\inf_{y\in U}\max\{1-\tilde{R}_B(x,y),\tilde{X}(y)\},$$

$$\overline{\tilde{R}}_B\tilde{X}(x)=\sup_{y\in U}\min\{\tilde{R}_B(x,y),\tilde{X}(y)\},$$

式中,$\underline{\tilde{R}}_B\tilde{X}$ 和$\overline{\tilde{R}}_B\tilde{X}$ 称为模糊集 $\tilde{X}$ 的鲁棒性模糊粗糙集。

对于分类数据集,类别 L 是一个精确集,因此

$$L(y)=\begin{cases}1, & y\in L\\ 0, & y\notin L\end{cases},$$

那么下近似集可以进一步表示为

$$\underline{\tilde{R}}_BL(x)=\inf_{y\in U}\max\{1-\tilde{R}_B(x,y),L(y)\}=$$
$$\inf_{y\in L}\max\{1-\tilde{R}_B(x,y),1\}\wedge\inf_{y\notin L}\max\{1-\tilde{R}_B(x,y),0\}=$$
$$1\wedge\inf_{y\notin L}\max\{1-\tilde{R}_B(x,y),0\}=\inf_{y\notin L}\max\{1-\tilde{R}_B(x,y)\},$$

上近似集可以进一步表示为

$$\overline{\tilde{R}}_BL(x)=\sup_{y\in U}\min\{\tilde{R}_B(x,y),L(y)\}=$$
$$\sup_{y\in L}\min\{\tilde{R}_B(x,y),1\}\vee\sup_{y\notin L}\min\{\tilde{R}_B(x,y),0\}=$$
$$\sup_{y\in L}\min\{\tilde{R}_B(x,y)\}\vee0=\sup_{y\in L}\min\{\tilde{R}_B(x,y)\}。$$

本研究提出的鲁棒性模糊粗糙集模型在计算上下近似时,减少了信息不可靠数据样本的影响,使用样本权重的方法计算对象之间的模糊关系,在计算模糊近似的过程中忽略噪声样本,实现对噪声数据的鲁棒性。

鲁棒性模糊粗糙集模型满足如下性质。

性质 3 给定(U,C,D)和 $B\subseteq C$, $U/D=\{L_1,L_2,\cdots,L_r\}$,那么

$$\underline{\tilde{R}}_B(\bigcap_{L\in U/D}L)=\bigcap_{L\in U/D}\underline{\tilde{R}}_BL,\tag{1}$$

$$\overline{\tilde{R}}_B(\bigcup_{L\in U/D}L)=\bigcup_{L\in U/D}\overline{\tilde{R}}_BL。\tag{2}$$

证明 根据定义 8,满足

$$\underline{\tilde{R}}_B(\bigcap_{L\in U/D}L)(x)=$$
$$\inf_{y\in U}\max\{1-\tilde{R}_B(x,y),\min_{L\in U/D}(L(y))\}=$$
$$\min_{L\in U/D}\{\inf_{y\in U}\max\{1-\tilde{R}_B(x,y),L(y)\}=$$
$$\min_{L\in U/D}\{\underline{\tilde{R}}_BL(x)\}=\bigcap_{L\in U/D}\underline{\tilde{R}}_BL(x),$$

因此式(1)成立,同理可以得到式(2)成立。

性质 4 给定(U,C,D)和 $B\subseteq C$, $U/D=\{L_1,L_2,\cdots,L_r\}$,那么

$$1-\underline{\tilde{R}}_BL(x)=\overline{\tilde{R}}_B(1-L(x)),\tag{3}$$

$$1-\overline{\tilde{R}}_BL(x)=\underline{\tilde{R}}_B(1-L(x))。\tag{4}$$

证明 根据定义 8,满足

$$1-\underline{\tilde{R}}_BL(x)=$$
$$1-\inf_{y\in U}\max\{1-\tilde{R}_B(x,y),L(y)\}=$$
$$\sup_{y\in U}\min\{\tilde{R}_B(x,y),1-L(y)\}=$$
$$\overline{\tilde{R}}_B(1-L(x)),$$

因此式(3)成立,同理可以得到式(4)成立。

性质 5 给定(U,C,D)和 $B\subseteq C$, L_1, $L_2\in U/D$ 且 $L_1\subseteq L_2$,那么

$$\underline{\tilde{R}}_BL_1\subseteq\underline{\tilde{R}}_BL_2,\tag{5}$$

$$\overline{\tilde{R}}_BL_1\subseteq\overline{\tilde{R}}_BL_2。\tag{6}$$

证明 根据鲁棒性模糊粗糙集模型的定义可以直接得到性质 5 成立。

上述性质为属性约简算法的构建提供了理论基础。

2.4 鲁棒性模糊粗糙集的属性约简算法

在本节中,提出一种基于鲁棒性模糊粗糙集的属性约简算法。类似于传统的模糊粗糙集模型,接下来提出鲁棒性模糊粗糙集的正区域以及依赖度定义。

定义 9 给定(U,C,D)和 $B\subseteq C$,鲁棒性模糊粗糙集决策属性集 D 关于条件属性集 B 的鲁棒性模糊正区域定义为

$$\mathrm{POS}_{\tilde{R}_B}(D)(x)=\sup_{\forall L\in U/D}\underline{\tilde{R}}_BL(x),$$

式中,$\mathrm{POS}_{\tilde{R}_B}(D)(x)$表示样本对象 x 被正确分类的程度,通过它可以进一步定义鲁棒性模糊粗糙集中属性集的模糊依赖度。

定义 10 给定(U,C,D)和 $B\subseteq C$,决策属性集 D 关于条件属性集 B 的模糊正区域为 $\mathrm{POS}_{\tilde{R}_B}(D)(x)$,定义 D 关于 B 的鲁棒性模糊依赖度为

$$\gamma_{\tilde{R}_B}(D)=\frac{\sum_{\forall x\in U,\omega_x\neq0}|\ \mathrm{POS}_{\tilde{R}_B}(D)(x)\ |}{|\ U\ |}。$$

定义 10 中,鲁棒性模糊依赖度描述了属性子集 B 和决策属性集 D 之间的关系程度,同时鲁棒性模糊依赖度的计算过滤了样本权重为 0 的样本,排除了噪声数据对依赖度的贡献。通过鲁棒性模糊依赖度可以评估属性的重要度。

定义 11 给定(U,C,D)和$B\subseteq C$, $\forall a\in B$ 关于 $B\subseteq C$ 的鲁棒性内部属性重要度 $s_{\mathrm{in}}(a,B,D)$和

$\forall b \in C-B$ 关于 $B \subseteq C$ 的鲁棒性外部属性重要度 $s_{\text{out}}(a,B,D)$ 分别定义为

$$s_{\text{in}}(a,B,D)=\gamma_{\tilde{R}_B}(D)-\gamma_{\tilde{R}_{B-\{a\}}}(D),$$

$$s_{\text{out}}(a,B,D)=\gamma_{\tilde{R}_{B\cup\{b\}}}(D)-\gamma_{\tilde{R}_B}(D)。$$

利用鲁棒性内外属性重要度作为信息系统贪心搜索的启发式函数,接下来提出一种基于鲁棒性模糊粗糙集的属性约简算法。

算法 1　信息系统鲁棒性模糊依赖度的计算算法

输入　模糊决策信息系统 (U,C,D),最近邻参数 k,邻域半径 δ,属性子集 $B \subseteq C$。

输出　鲁棒性模糊依赖度 $\gamma_{\tilde{R}_B}(D)$。

(1) 对于 $x \in U$,计算 k 近邻对象集 $\phi_k(x)$ 和邻域对象集 $\varphi_\delta(x)$;

(2) 根据定义 2 和定义 3 计算样本局部密度 $l_B(x)$ 和样本局部密度因子 $\chi_B(x)$;

(3) 根据定义 4 计算 x 的样本权重 $\omega_B(x)$;

(4) 重复步骤(1)~(3),计算每个对象 $\forall x \in U$ 的样本权重 $\omega_B(x)$;

(5) 对于 $x,y \in U$,计算 $\varphi_\delta(x)$ 和 $\varphi_\delta(y)$ 的样本权重距离 $d_B(\varphi_\delta(x),\varphi_\delta(y))$,并进一步计算对象 $x,y \in U$ 的鲁棒性模糊相似度 $\tilde{R}_B(x,y)=\exp(-d_B(\varphi_\delta(x),\varphi_\delta(y)))$;

(6) 对于 $U/D=\{L_1,L_2,\cdots,L_r\}$,计算鲁棒性模糊粗糙下近似集 $\underline{\tilde{R}}_B L_i(1 \leqslant i \leqslant r)$,根据性质 3 得到鲁棒性模糊正区域 $\text{POS}_{\tilde{R}_B}(D)$;

(7) 根据定义 10 计算鲁棒性模糊依赖度 $\gamma_{\tilde{R}_B}(D)$,返回最终结果。

算法 1 所示的是信息系统鲁棒性模糊依赖度计算算法,其计算复杂度集中在步骤(2)~(4),因此整个算法 1 的时间复杂度为 $O(|C|\cdot|U|^2)$。

算法 2　基于鲁棒性模糊依赖度的属性约简算法

输入　模糊决策信息系统 (U,C,D),最近邻参数 k,邻域半径 δ。

输出　属性约简子集 κ。

(1) $\kappa \leftarrow \varnothing$, $\gamma_{\tilde{R}_\kappa}(D)=0$;

(2) 遍历条件属性集 C 中每个属性 c_i, $1 \leqslant i \leqslant |C|$,如果 $\gamma_{\tilde{R}_{|c_i|}}(D)>0$,将当前属性 c_i 记录到 κ 中,即 $\kappa \leftarrow \kappa \cup \{c_i\}$;

(3) 如果 $\gamma_{\tilde{R}_\kappa}(D) \neq \gamma_{\tilde{R}_C}(D)$,遍历属性集 $C-\kappa$ 中每个属性 c_i, $1 \leqslant i \leqslant |C-B|$,计算鲁棒性外部属性重要度 $s_{\text{out}}(c_i,\kappa,D)$,并寻找出属性重要度最大的属性 $c_{\max}=\underset{c_i=C-\kappa}{\arg} s_{\text{out}}(c_i,\kappa,D)$,当 $s_{\text{out}}(c_{\max},\kappa,D)>0$,记录 $\kappa \leftarrow \kappa \cup \{c_{\max}\}$,重复步骤(3);

(4) 遍历属性集 κ 中每个属性 c_i, $1 \leqslant i \leqslant |\kappa|$,计算鲁棒性内部属性重要度 $s_{\text{in}}(c_i,\kappa,D)$,若 $s_{\text{in}}(c_i,\kappa,D)=0$,则 $\kappa \leftarrow \kappa-\{c_i\}$,并重复步骤(4),直到所有属性的鲁棒性内部属性重要度均不为 0;

(5) 返回属性集 κ。

算法 2 的流程图如图 2 所示。在算法 2 中,计算量集中在步骤(2)和步骤(4),根据算法 1 的时间复杂度结果,可以得到算法 2 的时间复杂度为 $O(|C|^2\cdot|U|^2)$。

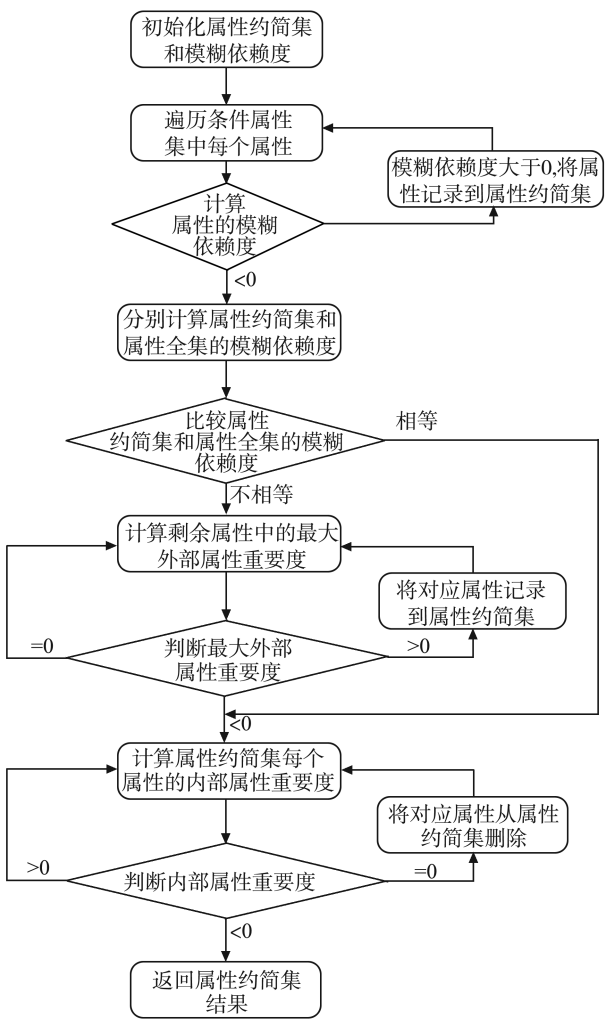


图 2　算法 2 的流程图

Fig.2　Flowchart of algorithm 2

3　试验分析

本章分别进行一系列试验评估所提出的模糊粗糙集属性约简算法的鲁棒性和有效性。这些试验在 Windows 操作系统环境中仿真实现,硬件环

境为 Intel(R)酷睿 i5-6500 3.20 GHz 四核心四线程的 CPU,内存 8 GB,试验算法使用 MATLAB2020 进行开发实现。本试验在 UCI 机器学习数据库中获取了 10 个应用数据集用于试验测试,这些数据集信息如表 2 所示。编号 9 和 10 为大样本高维度的数据集。这些数据集均为离散型和连续型混合类型,所有的连续型属性均标准化至[0,1]。

表 2 试验数据集

Table 2 Experimental dataset

编号	数据集	对象个数	属性个数	类别
1	Sonar	208	60	2
2	Ionosphere	351	34	2
3	Dermatology	366	34	6
4	Credit	690	16	2
5	Statlog	1 000	20	2
6	Waveform	1 000	21	3
7	Sick	3 772	29	2
8	Pendigits	7 494	16	10
9	Swarm	24 017	2 400	2
10	Consumption	130 000	21 000	7

本试验中,将提出的属性约简算法与 5 种近几年提出的属性约简算法进行比较,这些对比算法如下。

(1) 基于改进变精度模糊粗糙集的属性约简算法[12]。该算法通过变精度的策略来容忍数据的噪声,提升模糊粗糙集的属性约简性能。

(2) 基于概率粒距离度量的模糊粗糙集属性约简算法[9]。该算法使用概率粒距离度量来评估样本集的近似程度,使得距离度量结果可以自适应数据的概率分布,提升噪声数据环境的属性约简性能。

(3) 基于噪声感知的模糊粗糙集属性约简算法[11]。该算法建立在能够感知噪声的模糊粗糙集上,使得属性约简结果具有很好的鲁棒性。

(4) 基于核模糊相似度的模糊粗糙集属性约简算法[14]。该算法通过核相似关系评估对象之间的模糊相似性,使得提出的属性约简方法具有更好的属性选择效果。

(5) 基于特征子集划分的模糊粗糙属性约简算法[19]。该算法通过对数据集的原始特征集进行划分子集,每个特征子集赋以对应的权重,最终属性约简时可以提升算法的鲁棒性。

这些算法分别记录为对比算法 1~5。为了进行更全面的算法比较,本试验使用 2 种分类器来评估所选属性子集的分类性能,即朴素贝叶斯(naïve Bayes, NB)分类器和支持向量机(support vector machine, SVM)分类器。

在本试验中,参与试验的属性约简算法应用于每个数据集,并且试验中对每个数据集都采用 10 倍交叉验证,数据集被随机分为 10 个子集。这 10 个子集依次分别用作测试集,其余的 9 个子集用作训练集。即参与试验的属性约简算法在每个数据集上重复属性约简试验 10 次,并得到 10 个对应的属性子集结果。利用每个属性约简算法对应的属性子集在训练集上建立一个分类模型。通过这些分类器对测试集进行分类结果预测,记录和整理每个属性约简算法的平均分类精度,用于各个算法的分析和比较。

3.1 属性约简算法的有效性验证和分析

10 个数据集在所有属性约简算法下所选择的属性子集平均长度结果如表 3 所示。

表 3 属性约简子集平均长度

Table 3 Average length of attribute reduction subsets

算法	平均长度/个									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	60.0	34.0	34.0	16.0	20.0	21.0	29.0	16	2 400.0	21 000.0
对比算法 1	38.5	15.0	27.8	15.0	16.0	15.0	22.0	13	39.5	53.8
对比算法 2	47.0	14.0	26.0	15.0	13.0	16.0	19.0	9	42.6	59.3
对比算法 3	45.0	17.5	29.0	15.0	15.0	14.0	25.0	12	44.3	58.4
对比算法 4	22.7	13.4	20.5	12.6	14.6	15.0	20.0	13	33.2	54.6
对比算法 5	50.4	20.0	35.2	15.0	19.0	18.0	26.4	15	48.6	69.2
本研究算法	25.6	11.2	18.6	10.0	12.0	14.0	15.2	9	35.4	47.8

所选属性子集的 NB 和 SVM 分类精度结果分别如表 4、5 所示,其中粗体数据表示所有算法的最优试验结果。从表 4、5 的结果中可以看出,使用属性约简算法后属性子集的分类性能优于数据集全体属性集(原始属性集)的分类性能。这表明在数据集中有些属性是冗余的,所选的属性有助于提高

分类性能。对于两种分类器,本研究算法在 NB 分类器下在 8 个数据集上取得了更好的结果,分类精度比原始数据集和对比算法 1~5 分别提升17.0%、3.4%、3.7%、2.5%、4.8%和 3.3%;在 SVM 分类器下也在 8 个数据集上取得更好的结果,分类精度比原始数据集和对比算法 1~5 分别提升18.9%、3.1%、2.3%、1.7%、3.5%和 3.8%。从属性子集长度的角度来看,本研究算法所选择的平均属性数量是最少的,比原始数据集和对比算法 1~5 分别降低 99.2%、22.4%、23.9%、28.0%、9.6%和 38.4%。

表 4　属性约简子集 NB 分类精度

Table 4　Classification accuracy of attribute reduction subset (NB)

算法	分类精度									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	0.722 5	0.801 8	0.867 7	0.784 8	0.689 0	0.833 3	0.862 0	0.671 0	0.752 7	0.702 6
对比算法 1	0.863 4	0.867 2	0.974 7	0.842 4	0.738 0	0.939 0	0.963 7	0.790 3	0.833 9	0.887 3
对比算法 2	0.841 6	0.884 0	0.965 5	0.842 4	0.744 7	0.942 1	0.971 5	0.803 4	0.817 9	0.861 5
对比算法 3	0.886 5	0.898 4	0.970 8	0.842 4	0.762 5	0.932 8	0.963 5	0.772 0	0.852 4	0.894 2
对比算法 4	0.786 3	0.879 4	0.961 5	0.857 0	0.759 0	0.943 7	0.964 3	0.785 0	0.804 9	0.845 7
对比算法 5	0.865 2	0.866 4	0.952 3	0.867 4	0.719 0	0.928 7	0.951 8	0.803 4	0.847 3	0.912 4
本研究算法	0.892 9	0.922 0	0.985 9	0.895 8	0.758 4	0.958 2	0.983 7	0.829 0	0.865 7	0.906 9

表 5　属性约简子集 SVM 分类精度

Table 5　Classification accuracy of attribute reduction subset (SVM)

算法	分类精度									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	0.742 9	0.780 4	0.876 8	0.764 7	0.674 0	0.765 4	0.838 7	0.603 2	0.685 9	0.716 4
对比算法 1	0.814 6	0.872 8	0.963 5	0.865 3	0.742 0	0.877 3	0.938 8	0.805 6	0.848 6	0.865 9
对比算法 2	0.826 2	0.862 0	0.971 6	0.865 3	0.724 7	0.894 8	0.946 8	0.803 2	0.865 9	0.895 7
对比算法 3	0.836 7	0.885 4	0.967 2	0.865 3	0.737 0	0.875 9	0.938 8	0.816 5	0.871 3	0.917 5
对比算法 4	0.794 9	0.872 1	0.952 4	0.877 9	0.729 0	0.872 0	0.948 7	0.829 7	0.839 4	0.847 3
对比算法 5	0.808 1	0.896 8	0.953 8	0.854 2	0.735 0	0.854 8	0.915 5	0.784 5	0.855 8	0.882 9
本研究算法	0.842 5	0.884 9	0.981 5	0.899 0	0.757 4	0.884 5	0.968 8	0.834 6	0.885 6	0.923 6

3.2　属性约简算法的鲁棒性验证和分析

为了分析和验证本研究所提出属性约简算法的鲁棒性能,分别对每个数据集随机选择 5% 和 15%的属性值设置为噪声数据,通过取随机数的方式设置。若属性值介于[0,0.5),将其修改为[0.5,1.0]的随机值;若属性值介于[0.5,1.0],将其修改为[0,0.5)的随机值。通过这种方式生成对应新的数据集,然后将所有的属性约简算法对生成后的数据集进行试验。5%噪声数据下所有属性约简算法得到的属性子集平均长度结果如表 6 所示,5%噪声数据下所选属性子集的 NB 和 SVM 分类精度结果如表 7、8 所示,15%噪声数据下所有属性约简算法得到的属性子集平均长度结果如表 9 所示,15%噪声数据下所选属性子集的 NB 和 SVM 分类精度结果如表 10、11 所示。

表 6　属性约简子集平均长度(5%噪声数据)

Table 6　Average length of attribute reduction subsets (5% noisy data)

算法	平均长度/个									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	60.0	34.0	34.0	16.0	20.0	21.0	29.0	16.0	2 400.0	21 000.0
对比算法 1	39.5	16.0	29.2	16.0	16.0	16.6	24.0	14.0	41.7	55.3
对比算法 2	49.0	16.5	28.0	16.0	15.0	18.0	21.0	10.0	45.2	63.8
对比算法 3	46.0	18.5	29.0	16.0	16.0	15.0	26.0	13.0	46.8	61.5
对比算法 4	25.7	15.0	22.5	13.6	16.2	17.2	22.0	14.2	37.0	56.4
对比算法 5	52.4	23.0	33.2	16.0	20.0	19.8	28.8	15.0	51.2	70.6
本研究算法	24.6	13.6	22.6	14.0	14.0	15.0	17.0	10.0	37.6	49.4

表 7 属性约简子集 NB 分类精度(5%噪声数据)

Table 7 Classification accuracy of attribute reduction subset(NB, 5% noisy data)

算法	分类精度									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	0.707 9	0.786 9	0.853 2	0.769 1	0.674 2	0.818 8	0.847 4	0.656 3	0.737 5	0.685 6
对比算法 1	0.849 1	0.853 0	0.959 0	0.828 0	0.722 5	0.924 4	0.949 3	0.775 0	0.822 5	0.869 3
对比算法 2	0.826 0	0.869 8	0.950 4	0.828 1	0.730 0	0.926 8	0.965 9	0.788 7	0.801 4	0.846 9
对比算法 3	0.872 2	0.882 8	0.956 6	0.827 7	0.746 8	0.918 6	0.948 5	0.757 2	0.844 6	0.878 0
对比算法 4	0.771 5	0.865 2	0.947 1	0.842 4	0.743 7	0.927 8	0.949 2	0.769 5	0.793 5	0.823 8
对比算法 5	0.849 5	0.852 2	0.938 2	0.852 8	0.703 4	0.913 6	0.937 1	0.787 8	0.832 2	0.901 6
本研究算法	0.877 5	0.907 0	0.971 2	0.880 6	0.742 8	0.943 0	0.959 1	0.814 0	0.853 5	0.895 3

表 8 属性约简子集 SVM 分类精度(5%噪声数据)

Table 8 Classification accuracy of attribute reduction subset(SVM, 5% noisy data)

算法	分类精度									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	0.728 6	0.765 7	0.862 4	0.749 3	0.659 5	0.749 7	0.822 9	0.588 3	0.657 6	0.703 9
对比算法 1	0.799 3	0.858 5	0.948 0	0.850 5	0.726 7	0.862 4	0.923 5	0.790 2	0.835 6	0.845 1
对比算法 2	0.811 3	0.847 8	0.957 2	0.850 3	0.710 1	0.860 3	0.931 2	0.788 4	0.844 6	0.877 9
对比算法 3	0.821 2	0.870 5	0.951 7	0.850 0	0.722 4	0.881 6	0.924 3	0.801 8	0.866 8	0.904 5
对比算法 4	0.780 3	0.856 5	0.938 2	0.863 5	0.713 8	0.856 4	0.933 3	0.815 1	0.829 3	0.835 4
对比算法 5	0.792 7	0.881 2	0.938 5	0.839 3	0.720 1	0.839 9	0.900 6	0.769 4	0.834 7	0.871 5
本研究算法	0.827 5	0.869 5	0.965 9	0.884 1	0.741 8	0.869 4	0.953 2	0.818 9	0.873 0	0.913 8

表 9 属性约简子集平均长度(15%噪声数据)

Table 9 Average length of attribute reduction subsets (15% noisy data)

算法	平均长度/个									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	60.0	34.0	34.0	16.0	20.0	21.0	29.0	16.0	2 400.0	21 000.0
对比算法 1	41.2	18.0	32.4	16.0	18.0	19.6	26.0	15.0	43.2	57.0
对比算法 2	52.0	17.5	29.0	16.0	17.0	20.0	23.0	13.0	47.0	64.4
对比算法 3	47.0	20.0	31.0	16.0	18.0	18.0	28.0	14.0	47.6	63.2
对比算法 4	29.7	17.0	26.5	14.6	17.2	19.2	26.0	15.2	38.4	58.8
对比算法 5	53.4	26.0	36.2	16.0	20.0	20.8	28.8	15.0	53.0	71.4
本研究算法	26.5	14.2	23.6	14.0	14.5	17.0	18.0	12.0	38.8	50.9

表 10 属性约简子集 NB 分类精度(15%噪声数据)

Table 10 Classification accuracy of attribute reduction subset(NB, 15% noisy data)

算法	分类精度									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	0.668 3	0.747 0	0.813 7	0.728 4	0.634 4	0.779 3	0.807 8	0.616 6	0.717 5	0.675 6
对比算法 1	0.809 8	0.813 7	0.918 3	0.788 6	0.682 1	0.884 7	0.909 8	0.734 7	0.804 8	0.846 3
对比算法 2	0.785 4	0.830 6	0.910 3	0.788 8	0.690 3	0.886 5	0.925 3	0.749 0	0.793 4	0.825 4
对比算法 3	0.832 9	0.842 2	0.917 4	0.788 0	0.706 1	0.879 4	0.908 5	0.717 4	0.822 3	0.857 1
对比算法 4	0.731 7	0.826 0	0.907 6	0.802 8	0.703 4	0.887 0	0.909 1	0.729 0	0.772 8	0.812 4
对比算法 5	0.808 8	0.812 9	0.899 0	0.813 2	0.662 8	0.873 5	0.897 4	0.747 2	0.817 2	0.887 6
本研究算法	0.857 1	0.887 0	0.951 5	0.860 4	0.722 2	0.922 8	0.939 5	0.794 0	0.838 5	0.882 7

表 11　属性约简子集 SVM 分类精度(15%噪声数据)

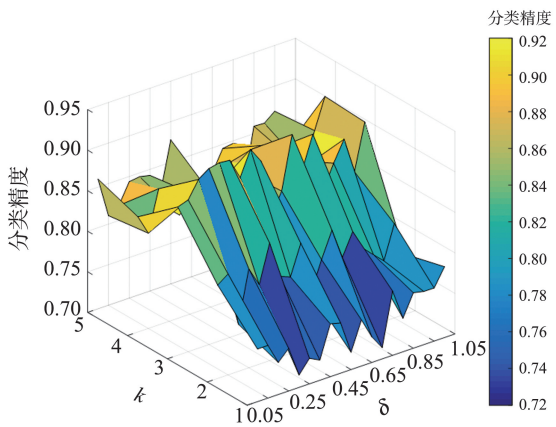
Table 11　Classification accuracy of attribute reduction subset(SVM, 15% noisy data)

算法	分类精度									
	Sonar	Ionosphere	Dermatology	Credit	Statlog	Waveform	Sick	Pendigits	Swarm	Consumption
原始属性集	0.689 2	0.726 0	0.823 0	0.708 9	0.620 0	0.709 1	0.782 1	0.548 4	0.644 2	0.683 7
对比算法 1	0.759 1	0.819 1	0.907 6	0.810 7	0.686 4	0.822 5	0.883 2	0.749 8	0.814 6	0.826 8
对比算法 2	0.771 4	0.808 5	0.917 8	0.810 3	0.670 4	0.820 8	0.890 6	0.748 6	0.825 8	0.851 4
对比算法 3	0.780 8	0.830 6	0.911 2	0.809 8	0.682 7	0.842 3	0.884 8	0.762 1	0.854 8	0.883 6
对比算法 4	0.740 7	0.815 9	0.899 0	0.824 0	0.673 6	0.815 8	0.893 0	0.775 5	0.809 6	0.821 4
对比算法 5	0.752 3	0.840 6	0.898 2	0.799 4	0.680 2	0.800 0	0.860 6	0.729 3	0.812 8	0.856 5
本研究算法	0.807 5	0.849 1	0.945 3	0.864 2	0.721 2	0.849 3	0.932 6	0.798 3	0.867 9	0.905 9

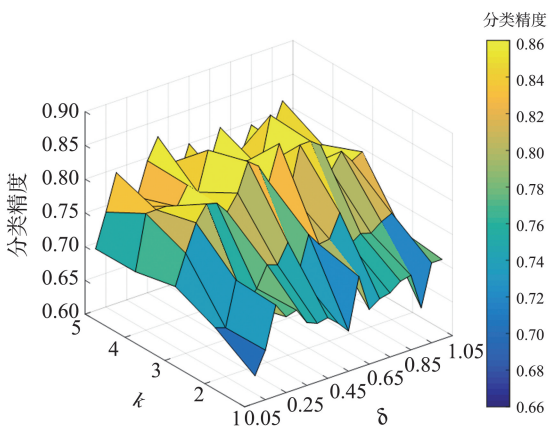
通过将表 7、8 和表 10、11 与表 4、5 进行比较,可以发现当数据集中存在噪声数据时,该噪声影响了数据的分类性能,其分类精度对应降低。对比 5%和 15%噪声数据的分类精度结果,在噪声比例为 5%时,本研究算法在少部分数据集下约简子集不是最小的,少部分数据集 NB 和 SVM 分类精度不是最高的,当噪声提升至 15%后,本研究算法在所有数据集下的约简子集均是最小的,在大部分数据集下的 NB 和 SVM 分类精度均是最高的。因此综合比较可以得出,本研究算法具有更好的鲁棒性,对于噪声数据具有更低的敏感性。

3.3　本研究属性约简算法的参数分析

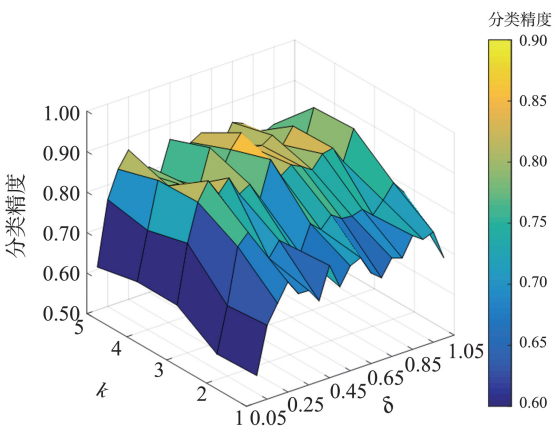
本研究所提出的鲁棒性属性约简算法包括 2 个参数,分别为最近邻参数 k 和邻域半径 δ。这 2 个参数会影响到本研究算法的有效性,本小节试验对这 2 个参数进行分析。为了验证参数的范围及其影响,本试验执行了网格搜索,δ 从 0.05 至0.95分别取值,步长为 0.05,k 从 2 至 6 分别取值,步长为 1。利用相应参数进行属性约简得到对应的平均分类精度,将部分数据集试验结果绘制成三维图如图 3 所示。



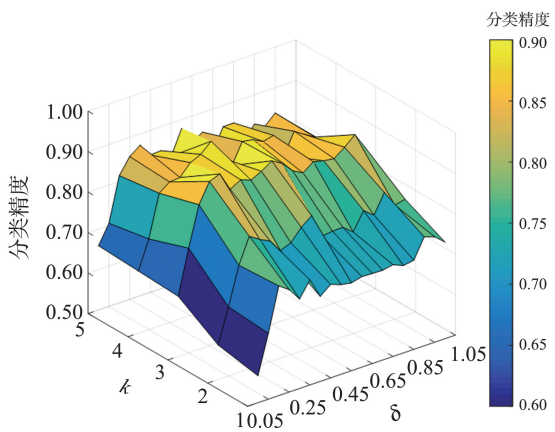
(a) Sonar(NB)



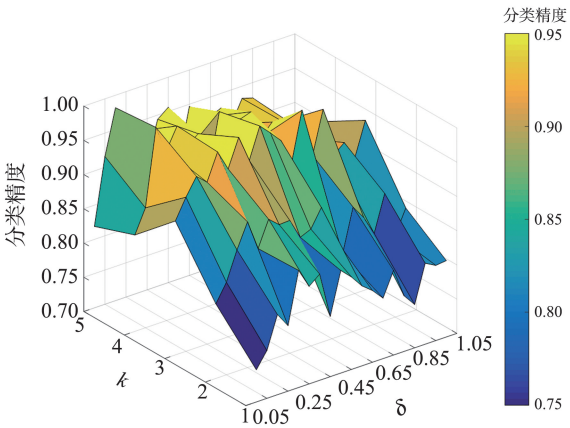
(b) Sonar(SVM)



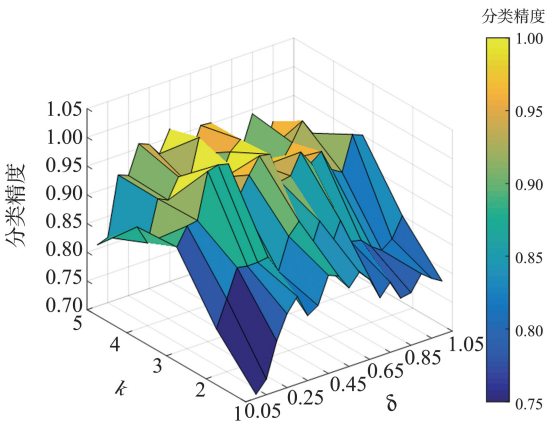
(c) Ionosphere(NB)



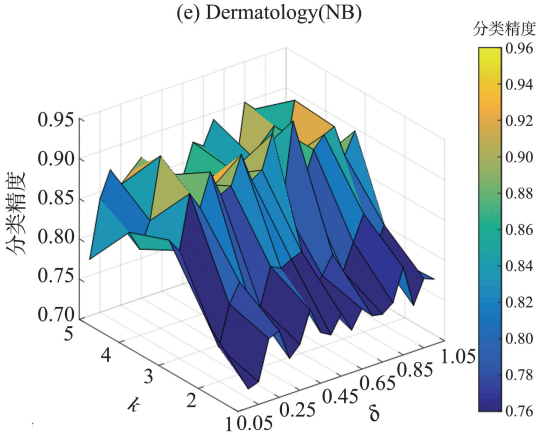
(d) Ionosphere(SVM)



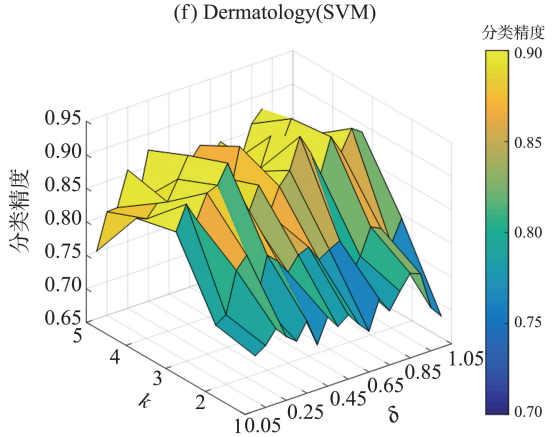
(e) Dermatology(NB)



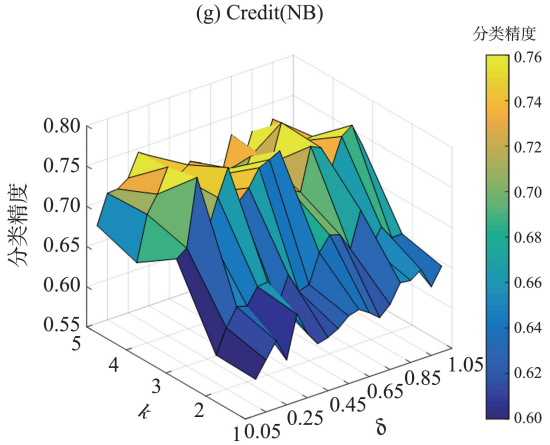
(f) Dermatology(SVM)



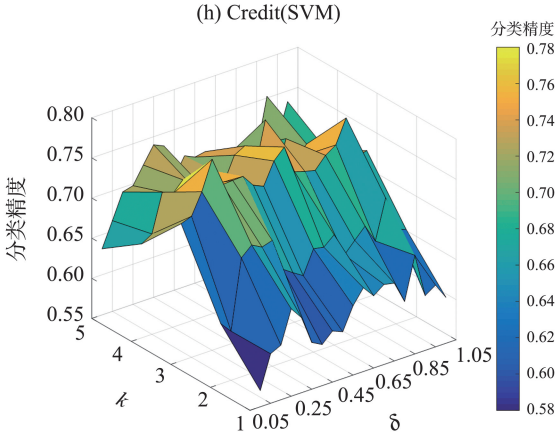
(g) Credit(NB)



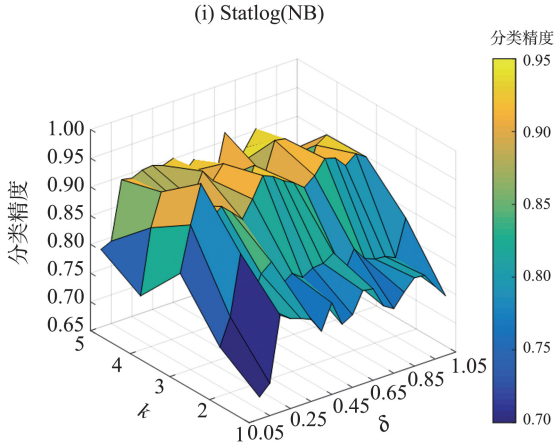
(h) Credit(SVM)



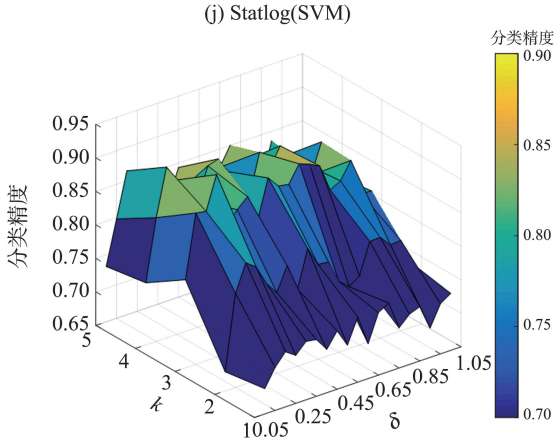
(i) Statlog(NB)



(j) Statlog(SVM)



(k) Waveform(NB)



(l) Waveform(SVM)

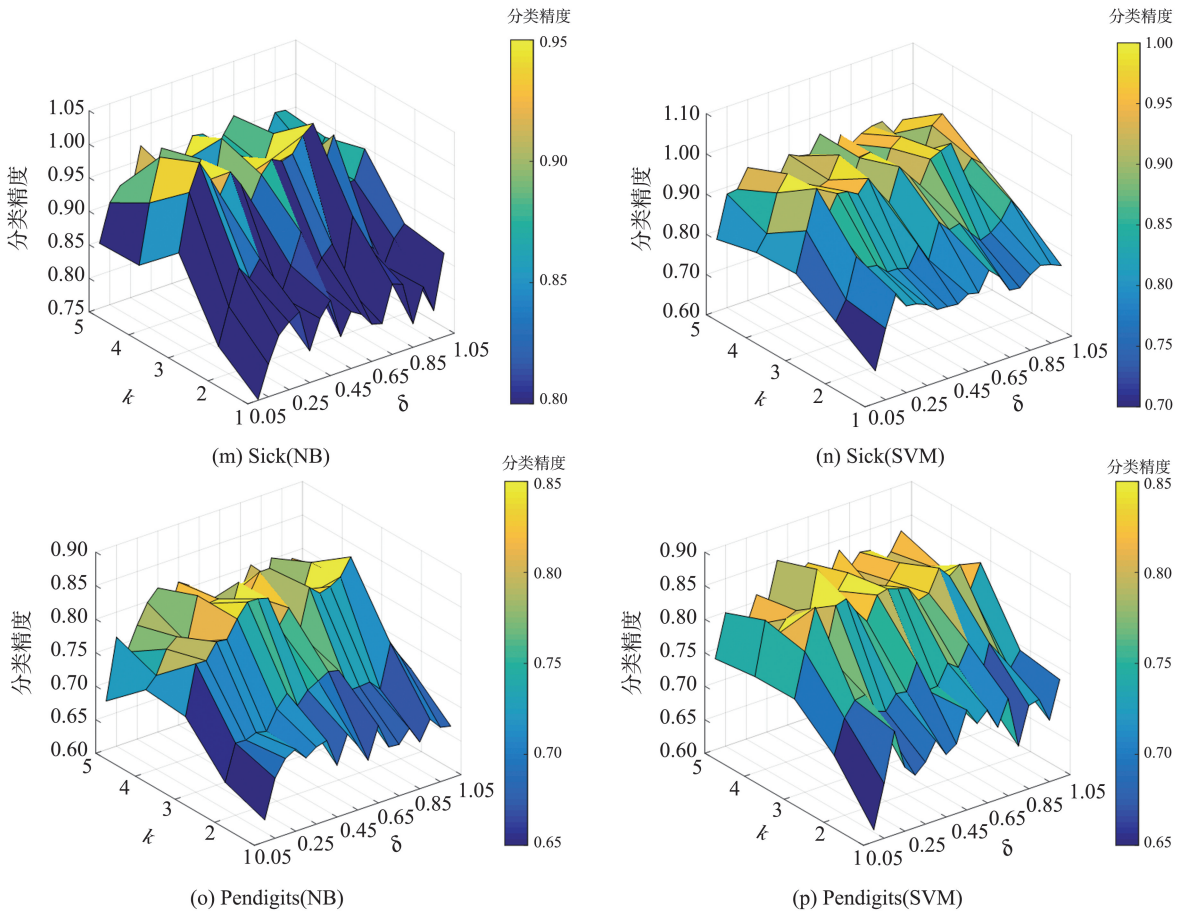


图 3　部分数据集不同参数分类精度结果

Fig.3　Classification accuracy results of different parameters for partial dataset

由图 3 可知,参数 k 和 δ 对这些数据集的分类精度有一定影响,当 δ 增加时,对于属性约简集的分类精度,有的数据集呈现先增加后减小的趋势,有的数据集呈现先增加后稳定的趋势,参数 k 的增加,呈现出逐渐增加后趋于平稳的趋势。当 δ 为 $[0.3,0.4]$,k 为 4 或 5 时,所选属性子集的分类精度在大多数数据集上都可以达到最优的分类精度。因此,本研究算法的最优参数可以在该范围进行选择,同时本研究算法在 3.1 节和 3.2 节的试验也是按照此范围进行设置。

4　结论

本研究通过样本的噪声评估样本权重,提出一种鲁棒性模糊粗糙集属性约简算法,本研究的贡献体现在如下 3 方面。(1)本研究通过样本的分布密度评估样本的噪声程度,通过噪声程度定义样本在数据集中的权重,并定义对象集的权重距离。该距离能有效降低噪声对经典距离测量的影响。(2)提出鲁棒性模糊粗糙集模型。在该模型中,通过对象集的权重距离定义对象之间的模糊相似关系,代替传统的模糊相似关系,可以有效降低噪声数据对模糊粗糙集上下近似逼近的影响;(3)基于鲁棒性模糊粗糙集提出一种属性约简算法,试验结果表明,比当前已有的属性约简算法具有更高的鲁棒性。

在接下来的研究工作中,将进一步探索该鲁棒性属性约简算法的增量式计算问题,以应对动态变化的数据场景。

参考文献:

[1] 宋苏洋, 叶军, 曾广财, 等. 基于优化可辨识矩阵的多粒度粗糙集属性约简算法[J]. 山东大学学报(理学版), 2024, 59(5): 52-62.
SONG Suyang, YE Jun, ZENG Guangcai, et al. Multi-granularity rough set attribute reduction algorithm based on optimized discernibility matrix[J]. Journal of Shandong University (Natural Science), 2024, 59(5): 52-62.

[2] YANG J, QIN X D, WANG G Y, et al. Attribute reduction for hierarchical classification based on improved fuzzy rough set[J]. Information Sciences, 2024, 677: 120900.

[3] CHEN Y P, DING W P, JU H R, et al. A distributed

attribute reduction based on neighborhood evidential conflict with Apache Spark[J]. Information Sciences, 2024, 668: 120521.

[4] CUI S G, LI G S, SANG B B, et al. Distance metric learning-based multi-granularity neighborhood rough sets for attribute reduction[J]. Applied Soft Computing, 2024, 159: 111656.

[5] DUBOIS D, PRADE H. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems, 1990, 17(2/3): 191-209.

[6] KOU G, YANG P, PENG Y, et al. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods[J]. Applied Soft Computing, 2020, 86: 105836.

[7] HU Q H, ZHANG L, AN S, et al. On robust fuzzy rough set models[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4): 636-651.

[8] WANG C Y, WAN L J. New results on granular variable precision fuzzy rough sets based on fuzzy (co) implications[J]. Fuzzy Sets and Systems, 2021, 423: 149-169.

[9] AN S, HU Q H, WANG C Z. Probability granular distance-based fuzzy rough set model[J]. Applied Soft Computing, 2021, 102: 107064.

[10] WANG C Z, HUANG Y, SHAO M W, et al. Fuzzy rough set-based attribute reduction using distance measures[J]. Knowledge-Based Systems, 2019, 164: 205-212.

[11] YANG X L, CHEN H M, LI T R, et al. A noise-aware fuzzy rough set approach for feature selection[J]. Knowledge-Based Systems, 2022, 250: 109092.

[12] ZHANG X H, OU Q Q, WANG J Q. Variable precision fuzzy rough sets based on overlap functions with application to tumor classification[J]. Information Sciences, 2024, 666: 120451.

[13] ZOU D D, XU Y L, LI L Q, et al. Novel variable precision fuzzy rough sets and three-way decision model with three strategies[J]. Information Sciences, 2023, 629: 222-248.

[14] LIANG P, LEI D F, CHIN K S , et al. Feature selection based on robust fuzzy rough sets using kernel-based similarity and relative classification uncertainty measures[J]. Knowledge-Based Systems, 2022, 255: 109795.

[15] SANG B B, CHEN H M, WAN J H, et al. Self-adaptive weighted interaction feature selection based on robust fuzzy dominance rough sets for monotonic classification[J]. Knowledge-Based Systems, 2022, 253(11): 109523.

[16] BAI H X, JING J H, LI D Y, et al. A fuzzy rough sets-based data-driven approach for quantifying local and overall fuzzy relations between variables for spatial data[J]. Applied Soft Computing, 2024, 162: 111848.

[17] THEERENS A, CORNELIS C. On the granular representation of fuzzy quantifier-based fuzzy rough sets[J]. Information Sciences, 2024, 665: 120385.

[18] SHU T X, LIN Y J, GUO L. Online hierarchical streaming feature selection based on adaptive neighborhood rough set[J]. Applied Soft Computing, 2024, 152: 111276.

[19] WANG Z H, CHEN H M, YANG X L, et al. Fuzzy rough dimensionality reduction: a feature set partition-based approach[J]. Information Sciences, 2023, 644: 119266.

(编辑:郭少华)