文章编号:1672-3961(2024)01-0033-12 DOI:10.6040/j.issn.1672-3961.0.2022.310

基于权重邻域熵的数值型信息系统属性约简算法

陈宝国[1],邓明[1*],陈金林[2]
(1.淮南师范学院计算机学院, 安徽 淮南 232038; 2.南京航空航天大学电子信息工程学院, 江苏 南京 211106)

摘要:在邻域粗糙集的属性约简中,每个属性被赋予相同的权重而不能更好地进行属性选择,针对这一问题,提出一种属性权重的邻域条件熵属性约简算法。通过条件属性与决策属性之间的相关系数评估条件属性的权重,基于权重方法提出一种改进的邻域关系,称为权重邻域关系,并提出相应的权重邻域粗糙集模型。以权重邻域粗糙集模型为基础,进一步提出权重邻域熵模型,理论证明权重邻域条件熵的单调性。通过权重邻域条件熵作为启发式函数提出一种新的数值型信息系统属性约简算法。试验结果表明,提出的属性约简算法具有更好的属性约简性能。

关键词:数值型信息系统;邻域粗糙集;属性约简;属性权重;邻域熵
中图分类号:TP18 **文献标志码**:A
引用格式:陈宝国,邓明,陈金林. 基于权重邻域熵的数值型信息系统属性约简算法[J]. 山东大学学报(工学版),2024,54(1):33-44.
CHEN Baoguo, DENG Ming, CHEN Jinlin. Attribute reduction algorithm of numerical information system based on weighted neighborhood entropy[J]. Journal of Shandong University (Engineering Science), 2024, 54(1):33-44.

Attribute reduction algorithm of numerical information system based on weighted neighborhood entropy

CHEN Baoguo[1], DENG Ming[1*], CHEN Jinlin[2]
(1. School of Computer Science, Huainan Normal University, Huainan 232038, Anhui, China; 2. College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, Jiangsu, China)

Abstract: In the attribute reduction of neighborhood rough set, each attribute was given the same weight and could not make better attribute selection. To solve this problem, a neighborhood conditional entropy attribute reduction algorithm with attribute weight was proposed. The weight of conditional attributes was evaluated by the correlation coefficient between conditional attributes and decision attributes. Based on the weight method, an improved neighborhood relation was proposed, which was called weighted neighborhood relation. The corresponding weighted neighborhood rough set model was also proposed. Based on the weighted neighborhood rough set model, the weighted neighborhood entropy model was further proposed, and the monotonicity of the weighted neighborhood conditional entropy was proved theoretically. A new attribute reduction algorithm for numerical information system was proposed by using the weighted neighborhood conditional entropy as the heuristic function. The experimental results showed that the proposed attribute reduction algorithm had better attribute reduction performance.

Keywords: numerical information system; neighborhood rough set; attribute reduction; attribute weight; neighborhood entropy

0 引言

属性约简又称为特征选择,是数据挖掘、机器学习和模式识别等领域的重点研究内容,属性约简的目的是去除数据集中冗余或不相关的特征(属性),而不显著降低分类器的预测精度,或产生尽可能接近原

收稿日期:2022-09-08
基金项目:安徽省高校自然科学研究重点项目(KJ2018A0469,KJ2021A0972)
第一作者简介:陈宝国(1978—),男,安徽安庆人,副教授,硕士,主要研究方向为粗糙集、粒计算、数据挖掘。E-mail:bgchen0706@163.com
*通信作者简介:邓明(1976—),男,安徽寿县人,教授,硕士生导师,博士,主要研究方向为安全信息处理、智能数据处理。
E-mail:mdeng76@163.com

始数据的类分布[1]。粗糙集理论是进行属性约简的常用工具,传统的粗糙集构建于等价关系,因此只能处理离散型数据,而真实世界中的数据集存在大量的数值型类型。例如,在医学诊断的信息系统中,年龄、体质量、血压以及各项血液指标都是数值型类型,这使得传统的属性约简算法面临一定的局限和挑战[2-5]。

近年来,针对数值型数据集,学者们研究提出了多种属性约简算法。文献[6]针对数值型信息系统提出了邻域粗糙集,利用邻域粗糙集的依赖度构建最早期的数值型信息系统属性约简算法。文献[7]将邻域粗糙集进行改进,提出最大决策邻域粗糙集模型,并利用新模型的决策依赖度设计了属性约简算法。文献[8]将传统的邻域粗糙集进行增量式学习的构造,设计一种基于邻域粗糙集的增量式属性约简算法。文献[9]对邻域粗糙集进行改进,提出了 k 近邻邻域粗糙集模型,同时提出对应的属性约简。文献[10]基于邻域粗糙集模型,提出一种非平衡数据的属性约简算法。文献[11]对邻域粗糙集进行改进,提出一种加权邻域粗糙集模型,并进一步提出加权邻域粗糙集的依赖度,同时设计出相应的属性约简算法。文献[12]在邻域粗糙集中提出邻域自信息度度量属性子集的不确定性,并且提出一种属性约简算法。文献[13]将邻域粗糙集与帝王蝶优化算法进行结合,提出一种新的数值型信息系统属性约简算法。文献[14]在邻域粗糙集中提出邻域互信息熵度量模型,并提出一种代价敏感的属性约简算法。文献[15]在邻域粗糙集中提出邻域知识粒度的概念,并设计一种属性约简算法。

目前,邻域粗糙集已成为数值型信息系统属性约简的重要方法和途径[16-20]。然而,目前所提出的各种属性约简算法大多没有考虑属性权重的问题,属性约简的过程中均按照相同的属性权重进行启发式搜索。但在实际应用中,每个属性对学习任务的权重可能是不相同的[11,21-23]。在基于邻域粗糙集的属性约简算法中,如果在启发式搜索前就考虑条件属性和决策属性之间的内部权重相关性,那么启发式搜索过程中就可以突出决策高度相关的属性,由此更容易选择具有高相关性和依赖性的属性。

针对上述问题,本研究提出一种基于属性权重的邻域粗糙集属性约简算法。首先,针对数值型信息系统的属性约简,通过条件属性与决策属性之间的相关系数评估条件属性的权重,这使得与决策属性高度相关的条件属性具有更高权重系数。同时基于权重提出一种改进的邻域关系,称之为权重邻域关系,相应地提出权重邻域粗糙集模型。其次,以权重邻域粗糙集模型为基础,进一步提出权重邻域熵度量模型,将权重邻域条件熵作为启发式函数。最后,提出一种新的属性约简算法。不同于文献[11]中加权方法的属性约简算法,本研究通过条件属性值与决策属性值的代数差值最小化求解得到每个属性的权重,提出一种基于权重的闵可夫斯基距离函数,并构造对应的权重邻域关系,在此基础上,进一步提出新的邻域熵度量、邻域联合熵度量以及邻域条件熵度量用于属性约简的设计。试验分析结果表明,本研究提出的属性约简算法具有更好的属性子集区分能力,与现有同类型属性约简算法相比,具有更高的属性约简性能。

1 邻域粗糙集模型

设一个邻域型决策信息系统表示为 $I=(U,T=C\cup D)$,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为信息系统的论域;$x_i\in U$ 为论域中第 i 个对象;T 为信息系统的全体属性集;$C=\{a_1,a_2,\cdots,a_m\}$ 为信息系统的条件属性集,论域中的对象 $\forall x\in U$ 在条件属性 $\forall a\in C$ 下的属性值表示为 $a(x)$,且 $a(x)\in\mathbf{R}$;$D=\{d\}$ 为信息系统的决策属性集,基于决策属性集 D,可以诱导出信息系统论域的决策类划分,表示为 $U/D=\{D_1,D_2,\cdots,D_r\}$。

定义 1[6] 考虑邻域型决策信息系统 $I=(U,T=C\cup D)$,条件属性子集 $A\subseteq C$ 在论域 U 下确定的邻域关系定义为:

$$N_A^{\delta}=\{(x,y)\in U\times U\mid\Delta_A(x,y)\leqslant\delta\},$$

式中:δ 为邻域关系的邻域半径,是一个非负常数;$\Delta_A(x,y)$ 为对象 x 和 y 在 A 下的闵可夫斯基距离度量函数,定义为:

$$\Delta_A(x,y)=\left(\sum_{\forall a\in A}|a(x)-a(y)|^p\right)^{1/p},$$

式中:p 为可变参数,通常取 $p=2$。

论域中对象 $\forall x \in U$ 基于邻域关系可以诱导出对应的邻域相似类,简称邻域类,表示为 $n_A^\delta(x)=\{y \in U \mid (x,y) \in N_A^\delta\}$。

定义 2[6]　考虑邻域型决策信息系统 $I=(U,T=C \cup D)$,条件属性子集 $A \subseteq C$ 确定的邻域关系为 N_A^δ,那么 $\forall X \subseteq U$ 在邻域关系 N_A^δ 下确定的邻域下近似粗糙集 $\underline{N_A^\delta}(X)$ 和邻域上近似粗糙集 $\overline{N_A^\delta}(X)$ 分别定义为:

$$\underline{N_A^\delta}(X)=\{x \in U \mid n_A^\delta(x) \subseteq X\},$$

$$\overline{N_A^\delta}(X)=\{x \in U \mid n_A^\delta(x) \cap X \neq \varnothing\}。$$

$\forall X \subseteq U$ 在邻域关系 N_A^δ 下的正区域为 $P_A^\delta(X)=\underline{N_A^\delta}(X)$,边界域为 $B_A^\delta(X)=\overline{N_A^\delta}(X)-\underline{N_A^\delta}(X)$,负区域为 $G_A^\delta(X)=U-\overline{N_A^\delta}(X)$。

论域 U 在决策属性集 D 下诱导的决策类划分为 $U/D=\{D_1,D_2,\cdots,D_r\}$,那么决策属性集 D 在邻域关系 N_A^δ 下确定的邻域决策下近似集粗糙集 $\underline{N_A^\delta}(D)$ 和邻域决策上近似粗糙集 $\overline{N_A^\delta}(D)$ 分别定义为:

$$\underline{N_A^\delta}(D)=\bigcup_{\forall D_i \in U/D} \underline{N_A^\delta}(D_i),$$

$$\overline{N_A^\delta}(D)=\bigcup_{\forall D_i \in U/D} \overline{N_A^\delta}(D_i)。$$

决策属性集 D 在邻域关系 N_A^δ 下的正区域为 $P_A^\delta(D)=\underline{N_A^\delta}(D)$,边界域为 $B_A^\delta(D)=\overline{N_A^\delta}(D)-\underline{N_A^\delta}(D)$,负区域为 $G_A^\delta(D)=U-\overline{N_A^\delta}(D)$。

2　权重邻域粗糙集与权重邻域熵

在粗糙集模型中,通常对每个属性使用相同的权重计算近似集的上近似和下近似,这使得对属性和决策之间内部关系的探索不够充分,并且使用相同的权重计算对象的相似类可能会导致在信息系统进行属性约简时更容易选择具有较大属性值的属性[23]。为了凸显属性权重在粗糙近似和属性约简中的重要性,学者们采用多种方法评估信息系统中属性的权重[21-23]。本章在邻域型信息系统下对属性的权重进行研究,提出一种改进的权重邻域关系,并基于该邻域关系进行邻域信息系统的邻域粒化,进一步构造邻域信息系统的邻域熵度量。

2.1　权重邻域粗糙集模型

考虑邻域型决策信息系统 $I=(U,T=C \cup D)$,对于 $U=\{x_1,x_2,\cdots,x_n\}$ 和 $C=\{a_1,a_2,\cdots,a_m\}$,设 U 和 C 的系数矩阵

$$\boldsymbol{\Phi}=\begin{bmatrix} a_1(x_1) & a_2(x_1) & \cdots & a_m(x_1) \\ a_1(x_2) & a_2(x_2) & \cdots & a_m(x_2) \\ \vdots & \vdots & & \vdots \\ a_1(x_n) & a_2(x_n) & \cdots & a_m(x_n) \end{bmatrix}。$$

决策属性集 D 在论域 U 下确定的决策值向量

$$\boldsymbol{\Psi}=(d(x_1)\quad d(x_2)\quad \cdots\quad d(x_n))^{\mathrm{T}},$$

式中 $d(x)$ 表示对象 x 决策属性集 D 下的属性值。

设条件属性集 $C=\{a_1,a_2,\cdots,a_m\}$ 的属性权重向量表示为 $\boldsymbol{\omega}=(\omega(a_1)\quad \omega(a_2)\quad \cdots\quad \omega(a_m))$。为确定各个属性的权重值,将该问题转化为一个最优化问题,其解为:

$$\boldsymbol{\omega}^*=\arg\min \| \boldsymbol{\Phi\omega}-\boldsymbol{\Psi} \|^2, \tag{1}$$

式中 $\| \cdot \|^2$ 表示向量的第二范数。

假设 $\boldsymbol{\omega}^*=0$,那么 $\boldsymbol{\Phi\omega}=\boldsymbol{\Psi}$,则 $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi\omega}=\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Psi}$,可以解出 $\boldsymbol{\omega}=(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Psi}$。但是,当 $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$ 不可逆时,式(1)可转换为:

$$J(\boldsymbol{\omega})=\| \boldsymbol{\Phi\omega}-\boldsymbol{\Psi} \|^2+\| \boldsymbol{\omega} \|^2,$$

式中 $J(\boldsymbol{\omega})$ 是一个凸函数。

$J(\boldsymbol{\omega})$ 导数为 0 时,函数取最小值,即 $J'(\boldsymbol{\omega})=2\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi\omega}-\boldsymbol{\Psi})+2\boldsymbol{\omega}=0$。因此 $(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\boldsymbol{E})\boldsymbol{\omega}=\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Psi}$,其中 $\boldsymbol{E}$ 为单位矩阵,$\boldsymbol{\omega}=(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\boldsymbol{E})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Psi}$。

对求解得到的权重向量 $\boldsymbol{\omega}$，其中向量第 i 个元素 ω_i 的绝对值 $|\omega_i|$ 即表示属性 a_i 和决策属性集 D 之间的关系程度。$|\omega_i|$ 越大，则关系程度越强，重要度越高，属性的权重越大。

对权重向量 $\boldsymbol{\omega}$ 中的权重值进行标准化，称为标准化的权重向量。

定义 3 考虑邻域型决策信息系统 $I=(U,T=C\cup D)$，对条件属性集 $C=\{a_1,a_2,\cdots,a_m\}$ 求解得到权重向量 $\boldsymbol{\omega}=(\omega(a_1)\quad\omega(a_2)\quad\cdots\quad\omega(a_m))$，标准化后的权重向量 $\boldsymbol{\omega}_{\text{std}}$ 表示为：

$$\boldsymbol{\omega}_{\text{std}}=(\omega_{\text{std}}(a_1)\quad\omega_{\text{std}}(a_2)\quad\cdots\quad\omega_{\text{std}}(a_m)),$$

$$\omega_{\text{std}}(a_i)=\frac{m\times|\omega(a_i)|}{\sum_{i=1}^{m}|\omega(a_i)|}。$$

根据定义 3 可以得到，

$$\omega_{\text{std}}(a_i)\geqslant 0,1\leqslant i\leqslant m,$$

$$\sum_{i=1}^{m}\omega_{\text{std}}(a_i)=m。$$

通过属性权重推导计算，可以看出属性的权重是通过使用条件属性和决策属性之间的关系程度分配，条件属性与决策属性之间的相关性越高，条件属性赋予的权重就越大。基于属性权重的定义，可以进一步在邻域型决策信息系统中提出基于权重的邻域关系。

定义 4 考虑邻域型决策信息系统 $I=(U,T=C\cup D)$，条件属性集 C 求解得到标准化权重向量 $\boldsymbol{\omega}_{\text{std}}$，对于属性子集 $A\subseteq C$ 和邻域半径 δ 确定的权重邻域关系 W_A^{δ} 定义为：

$$W_A^{\delta}=\{(x,y)\in U\times U|\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,y)\leqslant\delta\},$$

式中：$\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,y)$ 表示对象 x 和 y 基于权重的闵可夫斯基距离度量函数，定义为 $\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,y)=\left[\sum_{\forall a\in A}(\omega_{\text{std}}(a)\cdot|a(x)-a(y)|)^p\right]^{1/p}$，$p$ 通常取 2。

对象 $\forall x\in U$ 基于权重邻域关系诱导出的邻域相似类，简称权重邻域类，表示为 $W_A^{\delta}(x)=\{y\in U|(x,y)\in W_A^{\delta}\}$。

性质 1 对于 $\forall x,y,z\in U,A\subseteq C$，基于权重的闵可夫斯基距离度量函数满足：

（1）$\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,y)\geqslant 0$；

（2）$\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,x)=0$；

（3）$\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,y)=\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(y,x)$；

（4）$\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,z)\leqslant\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,y)+\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(y,z)$。

证明 类似传统的闵可夫斯基距离度量性质，可以直接得到性质 1 成立。

性质 2 若 $\forall a\in A,\omega_{\text{std}}(a)=1$，那么：

（1）$\Delta_A^{\boldsymbol{\omega}_{\text{std}}}(x,y)=\Delta_A(x,y)$；

（2）$W_A^{\delta}=N_A^{\delta}$；

（3）$w_A^{\delta}(x)=n_A^{\delta}(x)$。

证明 根据定义 1 和定义 4 可直接证明性质 2 成立。

通过性质 2 可以看出，当标准化权重向量 $\boldsymbol{\omega}_{\text{std}}$ 中所有属性的权重值都为 1，即每个属性拥有同等的权重值时，那么基于权重的闵可夫斯基距离度量将退化为传统的闵可夫斯基距离度量，权重邻域关系和权重邻域类将退化为经典的邻域关系和经典的邻域类，因此权重邻域关系是经典邻域关系的推广，经典邻域关系是权重邻域关系的特例。

通过在邻域型决策信息系统中定义的权重邻域关系，可以实现邻域信息系统论域中一种新的邻域粒化，从而进行目标近似对象集的粗糙近似计算。

定义 5 考虑邻域型决策信息系统 $I=(U,T=C\cup D)$，条件属性集 C 求解得到的标准化权重向量为 $\boldsymbol{\omega}_{\text{std}}$，$A\subseteq C$ 确定的权重邻域关系为 W_A^{δ}，那么 $\forall X\subseteq U$ 在 W_A^{δ} 下确定的权重邻域下近似粗糙集 $\underline{W_A^{\delta}}(X)$ 和权重邻域上近似粗糙集 $\overline{W}_A^{\delta}(X)$ 分别定义为：

$$\underline{W_A^{\delta}}(X)=\{x\in U|w_A^{\delta}(x)\subseteq X\},\ \overline{W}_A^{\delta}(X)=\{x\in U|w_A^{\delta}(x)\cap X\neq\varnothing\}。$$

$\forall X\subseteq U$ 在权重邻域关系 W_A^δ 下的正区域为 $P_A^\delta(X)=\underline{W_A^\delta}(X)$,边界域为 $B_A^\delta(X)=\overline{W}_A^\delta(X)-\underline{W_A^\delta}(X)$,负区域为 $G_A^\delta(X)=U-\overline{W}_A^\delta(X)$。

论域 U 在决策属性集 D 下诱导的决策类划分为 $U/D=\{D_1,D_2,\cdots,D_r\}$,决策属性集 D 在权重邻域关系 W_A^δ 下确定的权重邻域决策下近似集 $\underline{W_A^\delta}(D)$ 和权重邻域决策上近似集 $\overline{W}_A^\delta(D)$ 分别定义为:

$$\underline{W_A^\delta}(D)=\bigcup_{\forall D_i\in U/D}\underline{W_A^\delta}(D_i),\ \overline{W}_A^\delta(D)=\bigcup_{\forall D_i\in U/D}\overline{W}_A^\delta(D_i)。$$

决策属性集 D 在权重邻域关系 W_A^δ 下的正区域为 $P_A^\delta(D)=\underline{W_A^\delta}(D)$,边界域为 $B_A^\delta(D)=\overline{W}_A^\delta(D)-\underline{W_A^\delta}(D)$,负区域为 $G_A^\delta(D)=U-\overline{W}_A^\delta(D)$。

当各个属性的权重相等,那么基于权重邻域关系的上下近似集以及区域划分均退化为经典邻域关系的上下近似集和区域划分。

2.2　权重邻域熵

文献[24]提出一种邻域熵度量,并且用来评估离散型属性和连续型属性的依赖程度。本节将邻域熵进行进一步改进,提出一种基于权重邻域关系的邻域熵。

定义 6[24]　考虑邻域型决策信息系统 $I=(U,T=C\cup D)$,$U=\{x_1,x_2,\cdots,x_n\}$,条件属性子集 $A\subseteq C$ 确定的邻域关系为 N_A^δ,那么定义属性子集 $A\subseteq C$ 的邻域熵

$$E_\delta(A)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{|n_A^\delta(x_i)|}{n},$$

式中:$-\log\dfrac{|n_A^\delta(x_i)|}{n}$ 为对象 x_i 在论域下的信息量;$\dfrac{1}{n}$ 为对象 x_i 的概率,即文献[24]认为论域中的每个对象拥有相等的概率。而在本研究 2.1 节提出的权重邻域粗糙集中,每个属性都赋予一个权重,通常属性的权重并不是完全相等的,因此对每个对象赋予相同的概率并不合适,本研究将通过对象的权重邻域类占整个论域的比例评估对象的概率,进而提出一种新的邻域熵,称之为权重邻域熵。

定义 7　考虑邻域型决策信息系统 $I=(U,T=C\cup D)$,$U=\{x_1,x_2,\cdots,x_n\}$,条件属性集 C 求解得到的标准化权重向量为 $\boldsymbol{\omega}_{\text{std}}$,$A\subseteq C$ 确定的权重邻域关系为 W_A^δ,那么定义属性子集 $A\subseteq C$ 的权重邻域熵为 $E_\delta(A)=-\sum_{i=1}^{n}\dfrac{|w_A^\delta(x_i)|}{n}\log_2\dfrac{|w_A^\delta(x_i)|}{n}$。

权重邻域熵满足 $0\leqslant E_\delta(A)\leqslant\log_2 n$,当且仅当 $\forall x\in U$ 且 $w_A^\delta(x)=U$ 时,$E_\delta(A)=0$;当且仅当 $\forall x\in U$ 且 $w_A^\delta(x)=x$ 时,$E_\delta(A)=\log_2 n$。

定理 1　若邻域半径 $\delta\leqslant\delta'$,那么 $E_\delta(A)\geqslant E_{\delta'}(A)$。

证明　根据权重邻域类的定义,当 $\delta\leqslant\delta'$,有 $\forall x\in U,w_A^\delta(x)\subseteq w_A^{\delta'}(x)$,因此满足 $E_\delta(A)\geqslant E_{\delta'}(A)$。

定理 2　若属性子集 $A\subseteq B\subseteq C$,那么 $E_\delta(B)\geqslant E_\delta(A)$。

证明　根据权重邻域类的定义,当 $A\subseteq B\subseteq C$ 时,有 $\forall x\in U,w_B^\delta(x)\subseteq w_A^\delta(x)$,因此满足 $E_\delta(B)\geqslant E_\delta(A)$。

定义 8　考虑邻域型决策信息系统 $I=(U,T=C\cup D)$,$U=\{x_1,x_2,\cdots,x_n\}$,条件属性集 C 求解得到的标准化权重向量为 $\boldsymbol{\omega}_{\text{std}}$,属性子集 $A,B\subseteq C$ 的权重邻域联合熵

$$E_\delta(A,B)=-\sum_{i=1}^{n}\frac{|w_A^\delta(x_i)\cap w_B^\delta(x_i)|}{n}\log_2\frac{|w_A^\delta(x_i)\cap w_B^\delta(x_i)|}{n},$$

式中 $0\leqslant E_\delta(A,B)\leqslant\log_2 n$。

定理 3　权重邻域联合熵满足以下关系:$E_\delta(A,B)\geqslant E_\delta(A)$,$E_\delta(A,B)\geqslant E_\delta(B)$。

证明　对于 $\forall x\in U$ 满足:

$$w_A^\delta(x_i)\cap w_B^\delta(x_i)\subseteq w_A^\delta(x_i),w_A^\delta(x_i)\cap w_B^\delta(x_i)\subseteq w_B^\delta(x_i),$$

根据权重邻域联合熵的定义可以得到 $E_\delta(A,B)\geqslant E_\delta(A)$,$E_\delta(A,B)\geqslant E_\delta(B)$。

定义 9　考虑邻域型决策信息系统 $I=(U,T=C\cup D)$,$U=\{x_1,x_2,\cdots,x_n\}$,条件属性集 C 求解得到的标准化权重向量为 $\boldsymbol{\omega}_{\text{std}}$,属性子集 $B\subseteq C$ 关于 $A\subseteq C$ 的权重邻域条件熵

$$E_\delta(B|A)=-\sum_{i=1}^{n}\frac{|w_A^\delta(x_i)\cap w_B^\delta(x_i)|}{n}\log_2\frac{|w_A^\delta(x_i)\cap w_B^\delta(x_i)|}{|w_A^\delta(x_i)|}。$$

特别地,决策属性集 D 关于 $A\subseteq C$ 的权重邻域条件熵

$$E_\delta(D|A)=-\sum_{i=1}^{n}\frac{|w_A^\delta(x_i)\cap[x_i]_D|}{n}\log_2\frac{|w_A^\delta(x_i)\cap[x_i]_D|}{|w_A^\delta(x_i)|},$$

式中$[x_i]_D$ 表示对象 x_i 的等价类。

权重邻域条件熵满足如下重要的定理性质。

定理 4 若属性子集满足 $B\subseteq A\subseteq C$,那么 $E_\delta(D|B)\geqslant E_\delta(D|A)$。

证明 根据权重邻域类的定义,可以得到$\forall x\in U, w_A^\delta(x)=(w_A^\delta(x)\cap[x]_D)\cup(w_A^\delta(x)\cap(U-[x]_D))$,即 $w_A^\delta(x)=|w_A^\delta(x)\cap[x]_D|+|w_A^\delta(x)\cap(U-[x]_D)|$。

类似地,$w_B^\delta(x)=|w_B^\delta(x)\cap[x]_D|+|w_B^\delta(x)\cap(U-[x]_D)|$,设$|w_A^\delta(x)\cap[x]_D|=p_A$, $|w_A^\delta(x)\cap(U-[x]_D)|=q_A$, $|w_B^\delta(x)\cap[x]_D|=p_B$, $|w_B^\delta(x)\cap(U-[x]_D)|=q_B$,那么有 $E_\delta(D|A)=-\frac{1}{n}\sum_{i=1}^{n}p_A^i\log_2\frac{p_A^i}{p_A^i+q_A^i}$, $E_\delta(D|B)=-\frac{1}{n}\sum_{i=1}^{n}p_B^i\log_2\frac{p_B^i}{p_B^i+q_B^i}$。

定义函数 $f(x,y)=-x\log_2\frac{1}{x+y}(x>0,y>0)$且 $z=\frac{x}{x+y}$,那么

$$\begin{cases}\dfrac{\partial f}{\partial x}=-\dfrac{y}{x+y}-\log_2\dfrac{x}{x+y}=z-\log_2 z-1\\[2ex]\dfrac{\partial f}{\partial y}=\dfrac{x}{x+y}=z>0\end{cases}。$$

又由于$(z-\log_2 z-1)'=1-\frac{1}{z\ln 2}$,且 $z=\frac{x}{x+y}\in(0,1]$,即函数$\frac{\partial f}{\partial x}=z-\log_2 z-1$ 单调递减,当 $z=1$ 时取最小值 $\frac{\partial f}{\partial x}=0$,所以$\frac{\partial f}{\partial x}\geqslant 0$。因此函数 $f(x,y)$分别对于两个自变量单独递增的。

由于 $B\subseteq A\subseteq C$,那么根据权重邻域类的定义可以得到$\forall x\in U, w_A^\delta(x)\subseteq w_B^\delta(x)$,那么$|w_A^\delta(x)\cap[x]_D|\leqslant|w_B^\delta(x)\cap[x]_D|$,$|w_A^\delta(x)\cap(U-[x]_D)|\leqslant|w_B^\delta(x)\cap(U-[x]_D)|$,即 $p_A\leqslant p_B$ 且 $q_A\leqslant q_B$。

如果$\forall x\in U$ 满足 $w_A^\delta(x)=\{x\}$, $w_B^\delta(x)=\{x\}$。那么 $p_A=p_B=1$ 且 $q_A=q_B=0$,则 $E_\delta(D|A)=0$, $E_\delta(D|B)=0$; $1<p_A\leqslant p_B$ 且 $0<q_A\leqslant q_B$,那么 $f(p_A,q_A)\leqslant f(p_B,q_B)$。因此 $\frac{1}{n}\sum_{i=1}^{n}f(p_A^i,q_A^i)\leqslant\frac{1}{n}\sum_{i=1}^{n}f(p_B^i,q_B^i)$,即 $E_\delta(D|B)\geqslant E_\delta(D|A)$。

定理 4 表明,权重邻域条件熵满足属性增加而单调不增的特性。

根据定理 4 关于权重邻域条件熵的单调性,对于邻域型决策信息系统 $I=(U,T=C\cup D)$,属性子集 $A\subseteq C$,那么权重邻域条件熵满足 $0\leqslant E_\delta(D|A)\leqslant r\log_2\frac{n}{r}$。

若$\forall x\in U$ 满足 $w_A^\delta(x)=U$ 且$|[x]_D|=\frac{n}{r}$,那么此时权重邻域条件熵达到最大值 $E_\delta(D|A)=r\log_2\frac{n}{r}$。若$\forall x\in U$ 满足 $w_A^\delta(x)=\{x\}$,那么此时权重邻域条件熵达到最小值 $E_\delta(D|A)=0$。

3 基于权重邻域熵的属性约简算法

2.2 节提出了基于权重邻域关系的邻域熵度量,并理论分析了权重邻域条件熵的单调性,基于该性质可以使用贪心搜索策略寻找信息系统的一个最小属性子集,使得该属性子集与原始属性集具有相同的样本描述能力,从而对信息系统达到约简的目的[6-11]。本章将基于权重邻域条件熵提出一种邻域型信息系统的属性约简算法。

定义 10　考虑邻域型决策信息系统 $I=(U,T=C\cup D)$,条件属性集 C 求解得到的标准化权重向量为 $\boldsymbol{\omega}_{\text{std}}$,若属性子集 $A\subseteq C$ 是条件属性集 C 的一个属性约简,那么当且仅当:

$$E_\delta(D|A)=E_\delta(D|C),\tag{2}$$

$$\forall a\in A,E_\delta(D|A-\{a\})>E_\delta(D|A)。\tag{3}$$

在式(2)中,决策属性集 D 关于属性约简集 A 的权重邻域条件熵值与属性全集 C 的权重邻域条件熵值相同。式(3)中,属性约简集 A 中的所有属性都是必要的属性。因此,属性约简集 A 是一个与属性全集 C 具有相同权重邻域条件熵值的最小属性子集。

对于本章提出的属性约简方法,逐个计算每个属性子集的权重邻域条件熵值是不现实的,寻找简化算法的策略有很多种,如遗传算法、分支定界算法和贪心搜索等。本研究将通过贪心搜索算法寻找一个最优的属性子集。接下来,将基于权重邻域条件熵定义两种度量评估一个属性相对于一个属性子集的显著性。

定义 11　考虑邻域型决策信息系统 $I=(U,T=C\cup D)$,条件属性集 C 求解得到的标准化权重向量为 $\boldsymbol{\omega}_{\text{std}}$,对于属性子集 $A\subseteq C$,定义属性 $\forall a\in A$ 在属性子集 A 下的内部属性重要度 $s_{\text{in}}(a,A,D)=E_\delta(D|A-\{a\})-E_\delta(D|A)$。

定义属性 $\forall a\in C-A$ 在属性子集 A 下的外部属性重要度 $s_{\text{out}}(a,A,D)=E_\delta(D|A)-E_\delta(D|A\cup\{a\})$。

算法 1 是通过两种属性重要度函数进行启发式贪心选择搜索属性约简的算法。算法 1 中,首先从空集开始,通过外部属性重要度进行搜索属性,直到剩余的属性均为非重要的;其次通过内部属性重要度反向消除所选属性子集中的冗余属性。算法 1 中的输入参数 δ 是为了控制权重邻域类大小的阈值,需要提前设置。δ 可以由专家的先验知识设置或通过试验搜索得到,在试验环节,将对阈值 δ 的具体取值进行试验分析。

算法 1　基于权重邻域条件熵的邻域型信息系统属性约简算法。

输入　邻域型决策信息系统 $I=(U,T=C\cup D)$,邻域半径 δ;

输出　属性约简 e。

步骤 1　初始化属性约简 $e\leftarrow\varnothing$;

步骤 2　求解条件属性集 C 的标准化权重向量 $\boldsymbol{\omega}_{\text{std}}$;

步骤 3　while $C-e\neq\varnothing$ do

　for $\forall a\in C-e$ do

　　计算属性 $\forall a\in C-e$ 在属性子集 e 下的外部属性重要度 $s_{\text{out}}(a,e,D)$;

　end for

　　找出 $C-e$ 中外部属性重要度最大的属性 $a_{\max}$,即 $a_{\max}=\text{argmax}(s_{\text{out}}(a,e,D))$;

　if $s_{\text{out}}(a_{\max},e,D)>0$ then

　　$e\leftarrow e\cup\{a_{\max}\}$;

　else

　　break;

　end if

end while

步骤 4　for $\forall a\in e$ do

计算属性 $\forall a\in e$ 在属性子集 e 下的内部属性重要度 $s_{\text{in}}(a,e,D)$;

　if $s_{\text{in}}(a,e,D)=0$ then

　　$e\leftarrow e-\{a\}$;

　end if

end for

步骤 5　返回属性约简 e。

在算法 1 中,步骤 2 求解条件属性集 C 的标准化权重向量 $\boldsymbol{\omega}_{\text{std}}$所需的时间复杂度为 $O(|C|\cdot|U|)$,步骤 3 通过外部属性重要度启发式搜索属性所需的时间复杂度为 $O(|C|^2\cdot|U|^2)$,步骤 4 通过内部属性重要度进行反向冗余属性剔除的时间复杂度为 $O(|C|^2\cdot|U|^2)$,因此整个算法 1 的时间复杂度为 $O(|C|^2\cdot|U|^2)$。

4 试验分析

本试验主要分为根据算法选择出属性的数量、约简结果的分类精度以及算法的运行时间 3 部分评估算法。进行属性约简的试验数据集(这些数据集下载于 http://www.ics.uci.edu)如表 1 所示,这些数据集均为数值型类型,为了消除属性量纲对试验带来的影响,试验前将所有属性值进行归一化至[0,1]。本章所有试验内容运行在操作系统为 Windows 7 的个人电脑上,处理器为英特尔四核 CPU,主频为 3.1 GHz,内存为 8 GB。

表 1 试验数据集

Table 1 Experimental data set

编号	数据集	样本数/个	属性数量/个	类别
1	wine	178	13	3
2	ionos	351	33	2
3	wdbc	569	30	2
4	biodeg	1 055	41	2
5	mess	1 151	19	3
6	gearbox	1 603	72	5
7	segment	2 310	19	7
8	musk	6 598	166	2

在算法 1 中,输入参数 δ 控制邻域半径,不同的邻域半径诱导出不同的权重邻域类,对算法的属性约简结果起着重要的作用。为了选择合适的输入参数,本试验首先按照等距取值的方法在[0.02,0.4]内以 0.02 为间距分别取值作为邻域半径进行试验,采用十折交叉验证方法求取每个训练集的属性约简结果,然后对每个属性约简结果分别通过 SVM 分类器和 C4.5 分类器得到对应测试集的分类精度。计算出每个邻域半径下十折交叉验证试验结果的平均属性约简长度、平均 SVM 分类精度和平均 C4.5 分类精度,部分数据集的最终试验结果如图 1 所示。

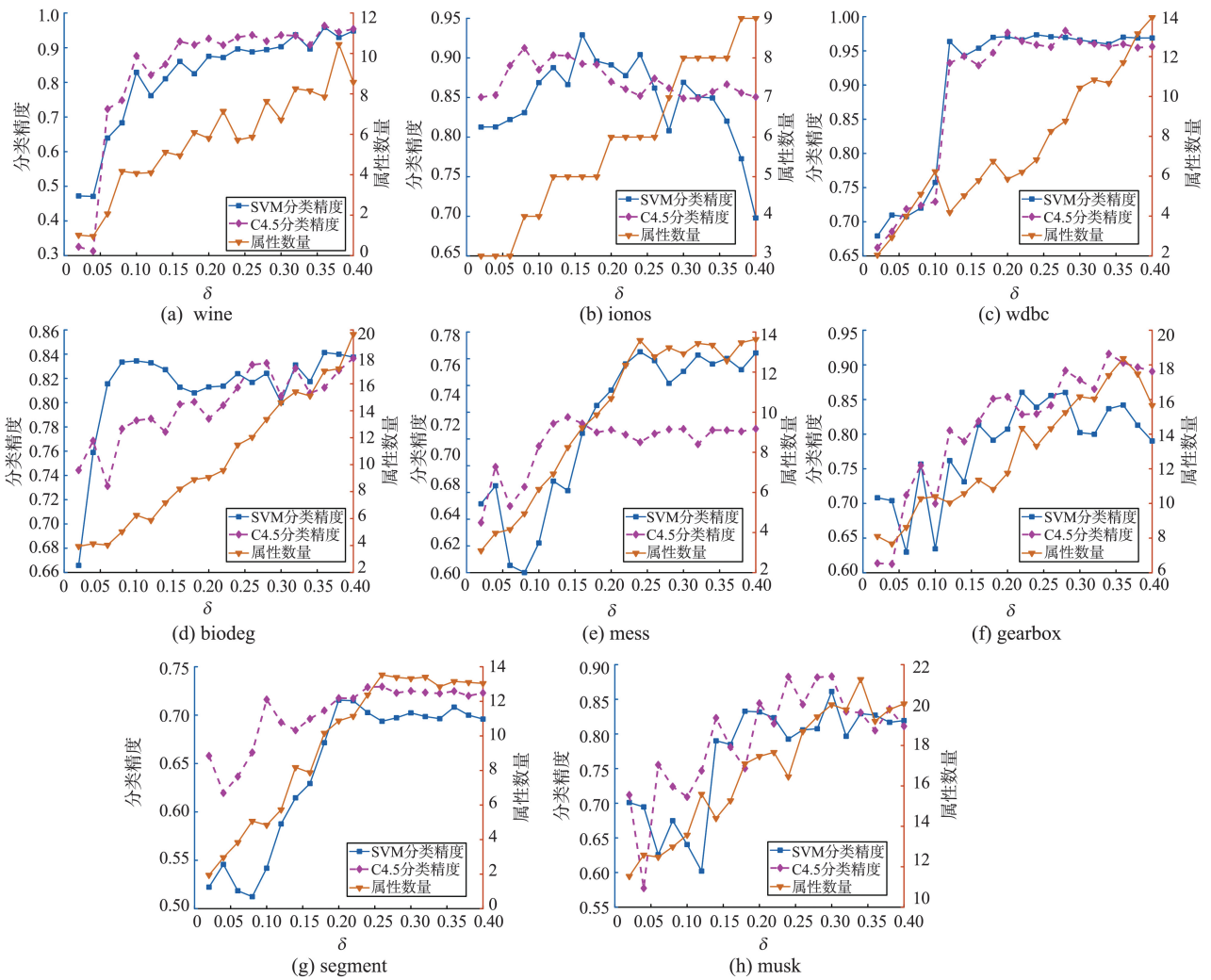

图 1 不同邻域半径下属性约简与分类精度结果

Fig.1 Results of attribute reduction and classification accuracy under different neighborhood radius

如图 1 所示,当邻域半径 δ 选取在 0.15 左右时,选择出的属性约简结果分类精度较高且属性约简集较小,因此算法选择 $\delta=0.15$ 作为最终参数进行试验。

为了进行试验参照和对比,本试验从近几年学者们提出的同类型属性约简算法中选择 4 种对比算法,分别为:基于粒决策熵的属性约简算法[2],记为比较算法 1;基于 k 近邻邻域粗糙集的属性约简算法[9],记为比较算法 2;基于距离度量学习的邻域粗糙集属性约简算法[16],记为比较算法 3;基于邻域粗糙集的一种考虑特征交互的属性约简算法[17],记为比较算法 4。

这 4 种对比算法都有对应的输入参数,本试验将选择每种对比算法其原始文献中得出的最优试验参数进行试验,其中比较算法 1 的最优邻域半径为 0.10;比较算法 2 的最优邻域半径为 0.15;比较算法 3 的最优邻域半径为 0.12;比较算法 4 的最优参数为 1.80,该参数与本研究的邻域半径有一定区别,但是具有类似于邻域半径的含义。将 4 种比较算法和本研究算法分别进行试验,4 种比较算法和本研究算法对每个数据集属性约简结果的长度比较如表 2 所示。

表 2　属性约简长度比较

Table 2　Attribute reduction length comparison

数据集	属性全集	属性约简长度				
		比较算法 1	比较算法 2	比较算法 3	比较算法 4	本研究算法
wine	13	8.2	5.0	5.0	5.0	5.0
ionos	33	10.5	5.5	6.4	6.2	5.6
wdbc	30	8.8	7.4	7.4	8.0	6.8
biodeg	41	12.6	9.3	7.2	8.2	8.0
mess	19	12.5	10.0	10.6	11.4	9.2
gearbox	72	15.4	12.8	13.3	11.2	11.0
segment	19	11.7	8.5	9.3	9.2	8.7
musk	166	23.2	18.5	20.3	17.6	16.5

从表 2 可以看出,所有属性约简算法都可以有效约简属性,例如对于 musk 数据集,5 种属性约简算法均大幅度删除了数据集中的不相关属性,简化数据的规模和结构,提高数据的知识发现性能。在大多数数据集中,比较算法 1 选择出的属性数量比其他算法多,可能是由于比较算法 1 是一种针对离散型数据的属性约简算法,进行试验时需要对数据进行离散化处理,这一过程造成了信息的丢失,降低了删除属性的数量。而对于本研究算法,大多数数据集选择出了更少的属性,这主要是由于本研究算法进行属性约简时首先进行属性权重的计算,同时邻域类的计算也建立在属性的权重上,使得最终的邻域熵包含了属性权重的信息,得到了更好的属性约简效果。

各个算法属性约简结果的 SVM 分类精度和 C4.5 分类精度比较如表 3、4 所示,其结果展示为“平均分类精度±标准差”的形式,表中粗体表示对应数据集中的分类精度最高值。

从表 3、4 可以看出:比较算法 1 在数据集 ionos 中有最高的 SVM 分类精度,比较算法 3 在数据集 mess 中有最高的 SVM 分类精度;比较算法 3 在数据集 wine 中有最高的 C4.5 分类精度,比较算法 2 在数据集 gearbox 中有最高的 C4.5 分类精度;本研究所提出的算法在大部分数据集中有更高的 SVM 分类精度和 C4.5 分类精度。这说明从整体角度分析,本研究算法所选择的约简结果具有更高的分类性能,这主要得益于本研究算法在属性约简时进行了数据集属性权重评估,使得算法在属性启发式搜索过程中能够选择属性重要度更高的属性,因此最终约简集的分类精度更高。

表 3　属性约简的 SVM 分类精度

Table 3　SVM classification accuracy of attribute reduction

数据集	属性全集	SVM 分类精度				
		比较算法 1	比较算法 2	比较算法 3	比较算法 4	本研究算法
wine	0.854 7±0.003 8	0.915 2±0.006 7	0.881 6±0.003 7	0.896 4±0.006 2	0.908 0±0.005 9	**0.938 7±0.004 4**
ionos	0.848 2±0.071 7	**0.916 3±0.082 2**	0.887 4±0.081 3	0.900 2±0.099 4	0.896 5±0.053 9	0.912 5±0.066 5
wdbc	0.875 9±0.014 4	0.941 2±0.009 2	0.947 6±0.011 2	0.934 1±0.013 2	0.961 9±0.008 4	**0.970 7±0.005 7**
biodeg	0.811 7±0.015 5	0.820 0±0.016 4	0.830 6±0.011 4	0.812 0±0.009 3	0.834 0±0.011 7	**0.841 3±0.010 3**
mess	0.689 8±0.024 8	0.740 7±0.017 9	0.766 1±0.022 7	**0.783 5±0.020 8**	0.750 7±0.021 8	0.760 2±0.015 7
gearbox	0.827 2±0.011 7	0.846 6±0.020 1	0.834 2±0.017 1	0.858 8±0.014 3	0.835 5±0.013 9	**0.865 2±0.017 4**
segment	0.548 1±0.009 5	0.682 9±0.012 9	0.673 6±0.013 6	0.703 9±0.008 1	0.686 7±0.018 9	**0.718 3±0.007 2**
musk	0.780 6±0.010 9	0.836 7±0.017 9	0.847 8±0.010 7	0.842 4±0.013 3	0.831 2±0.016 3	**0.866 8±0.019 8**

表 4 属性约简的 C4.5 分类精度

Table 4 C4.5 classification accuracy of attribute reduction

数据集	属性全集	C4.5 分类精度				
		比较算法 1	比较算法 2	比较算法 3	比较算法 4	本研究算法
wine	0.845 0±0.010 6	0.936 5±0.013 5	0.926 1±0.007 4	**0.965 2±0.006 7**	0.958 6±0.008 1	0.953 4±0.004 9
ionos	0.834 9±0.007 7	0.856 8±0.012 2	0.843 2±0.006 7	0.887 5±0.004 9	0.873 9±0.004 0	**0.902 8±0.006 7**
wdbc	0.851 2±0.003 3	0.907 2±0.006 0	0.930 3±0.006 7	0.928 9±0.007 4	0.895 7±0.008 8	**0.942 2±0.005 3**
biodeg	0.751 0±0.019 1	0.803 8±0.015 7	0.808 5±0.013 8	0.819 2±0.014 6	0.801 5±0.015 2	**0.828 4±0.010 3**
mess	0.671 5±0.016 6	0.715 5±0.012 3	0.714 9±0.004 5	0.702 7±0.013 0	0.693 6±0.012 1	**0.736 3±0.012 9**
gearbox	0.804 8±0.017 0	0.867 6±0.014 8	**0.917 4±0.012 9**	0.908 9±0.012 9	0.889 5±0.012 4	0.896 6±0.011 5
segment	0.642 8±0.013 4	0.731 7±0.017 4	0.723 4±0.014 8	0.692 5±0.013 5	0.705 8±0.011 6	**0.745 5±0.012 5**
musk	0.850 2±0.026 2	0.844 2±0.020 2	0.864 8±0.024 3	0.872 5±0.024 1	0.842 5±0.019 8	**0.892 5±0.022 9**

运行时间是验证属性约简算法有效性的另一重要指标,本试验使用 4 种对比算法与本研究算法对每个数据集进行属性约简,并重复 20 次,记录其每次约简用时,最终平均运行时间结果如表 5 所示。

表 5 属性约简的时间比较

Table 5 Time comparison of attribute reduction

数据集	时间/s				
	比较算法 1	比较算法 2	比较算法 3	比较算法 4	本研究算法
wine	2.24	3.38	3.61	2.55	2.50
ionos	30.28	43.05	57.80	37.99	34.02
wdbc	56.31	77.20	83.98	54.74	68.22
biodeg	141.52	167.16	190.07	161.73	156.22
mess	134.14	168.77	198.95	141.12	139.19
gearbox	714.67	883.94	1 095.48	823.45	739.68
segment	248.78	318.45	375.80	285.86	278.94
musk	2 975.58	3 875.05	4 272.65	3 583.60	3 102.37

由表 5 可知,比较算法 1 的平均运行时间最短,这主要是由于该算法进行属性约简时已将数据集进行了离散化,并利用等价关系构建信息系统的数据粒化,因此其计算效率更高。本研究算法的平均运行时间略长于比较算法 1,但短于其余比较算法,说明本研究算法具有较高的属性约简效率。

为了进一步比较各个属性约简算法的性能,应用 Friedman 检验和 Bonferroni-Dunn 检验验证试验结果的统计学意义[25]。Friedman 统计量的定义为:

$$\chi_F^2=\frac{12M}{k(k+1)}\left(\sum_{i=1}^{k}R_i^2-\frac{k(k+1)^2}{4}\right),$$

$$F_F=\frac{(M-1)\chi_F^2}{M(k-1)-\chi_F^2},$$

式中,M 为数据集的数量,k 为参与试验的算法数量,R_i 为第 i 个算法在所有数据集下的平均秩,F_F 服从 $k-1$ 和 $(k-1)(M-1)$ 自由度的 Fisher 分布。如果在 Friedman 检验统计量中拒绝了原假设,则可以使用 Bonferroni-Dunn 检验进一步探索哪些算法在统计方面存在显著性差异。根据本试验的结果,如果平均秩的距离超过临界距离 $d_\alpha=q_\alpha\sqrt{\frac{k(k+1)}{6M}}$(其中 q_α 为文献[25]提供的临界表值),则认为这两种算法的性能存在显著性差异。

为了探究 4 种比较算法和本研究算法在每个分类器下的分类性能是否存在显著性差异,进行了两次 Friedman 检验,Friedman 检验的原假设是所有的算法在分类性能方面都是相同的。5 种属性约简算法在 SVM 分类器和 C4.5 分类器下的秩如表 6、7 所示。

基于表 6、7 可以得到 SVM 分类器下 $F_F=7.436$ 和 C4.5 分类器下 $F_F=6.825$,而 $\alpha=0.1$ 时 $k-1$ 和 $(k-1)(M-1)$ 自由度的 Fisher 分布临界值 $F(4,28)=3.15$。因此可以拒绝 $\alpha=0.1$ 处的原假设,并接受 5 种算法在 SVM 分类器和 C4.5 分类器下具有显著性差异的替代假设。因此,进行两次 Bonferroni-Dunn 检验。在文献[25]中临界值 $q_{0.1}=3.622$,即 $d_{0.1}=2.862(k=5,M=8)$。在 SVM 分类器和 C4.5 分类器的秩结果中,Bonferroni-Dunn 检验表明本研究算法优于比较算法 1、2 和 4,而没有一致性的证据表明本研究算法与比较

算法 3 存在统计学差异。

表 6　SVM 分类精度结果的秩

Table 6　Rank of SVM classification accuracy results

数据集	秩				
	比较算法 1	比较算法 2	比较算法 3	比较算法 4	本研究算法
wine	4.500	3.000	2.000	3.500	2.000
ionos	5.000	3.500	2.500	4.000	1.500
wdbc	4.000	4.000	1.000	3.000	2.000
biodeg	3.000	3.000	2.000	3.000	1.500
mess	5.000	3.000	3.000	3.500	1.500
gearbox	5.000	3.500	2.500	4.000	1.000
segment	4.500	3.000	3.500	4.500	2.000
musk	5.500	4.000	3.000	3.500	1.500
平均	4.563	3.375	2.436	3.625	1.625

表 7　C4.5 分类精度结果的秩

Table 7　Rank of C4.5 classification accuracy results

数据集	秩				
	比较算法 1	比较算法 2	比较算法 3	比较算法 4	本研究算法
wine	4.000	3.000	2.000	2.500	2.000
ionos	5.500	4.000	2.500	3.500	1.500
wdbc	4.500	4.500	2.000	3.000	2.000
biodeg	4.000	3.500	3.000	3.500	1.500
mess	4.500	4.000	2.500	3.500	2.500
gearbox	5.000	3.000	2.500	3.000	2.000
segment	4.000	3.500	3.500	3.500	2.500
musk	5.000	3.500	4.000	4.500	2.500
平均	4.563	3.625	2.750	3.375	2.063

上述所有试验结果表明,本研究所提出的邻域型信息系统属性约简算法在属性约简集长度、分类精度以及运行时间方面均具有更高的属性约简性能,并通过统计学检验的方法进一步验证了本研究算法性能的优越性。

5　结语

对数据集进行属性约简可以提高机器学习和数据挖掘的性能和计算效率。传统的基于邻域粗糙集的属性约简算法在进行启发式搜索属性时,对每个属性分配了相同的权重,而没有预先充分挖掘属性和决策之间的内部关系,这降低了最终属性约简结果的分类性能。针对这一问题本研究提出一种改进的属性约简算法。首先,通过计算条件属性与决策属性之间的相关系数,为属性分配不同的权重;其次,基于权重的方法重构邻域关系以及邻域粗糙集模型,并提出了权重邻域熵度量;最后,利用权重邻域条件熵提出了一种启发式属性约简算法。试验结果表明,本研究算法对比其他算法具有较优的属性约简性能。基于本研究方法,接下来可以进一步探究权重邻域粗糙集的增量式属性约简问题。

参考文献:

[1] 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 Tao, LU Huiling, REN Hailing, et al. Survey on attribute reduction algorithm of rough set[J]. Acta Electronica Sinica, 2021, 49(7):1439-1449.

[2] GAO Can, ZHOU Jie, MIAO Duoqian, et al. Granular-conditional-entropy-based attribute reduction for partially labeled data with proxy labels[J]. Information Sciences, 2021, 580:111-128.

[3] ZHANG Qinli, CHEN Yiying, ZHANG Gangqiang, et al. New uncertainty measurement for categorical data based on fuzzy information structures: an application in attribute reduction[J]. Information Sciences, 2021, 580:541-577.

[4] 李明,甘秀娜,王月波. 基于集成学习的决策粗糙集特定类属性约简算法[J]. 计算机应用与软件, 2021, 38(6):262-270.
LI Ming, GAN Xiuna, WANG Yuebo. Class-specific attribute reduction algorithm for decision-theoretic rough sets based on ensemble learning[J]. Computer Applications and Software, 2021, 38(6):262-270.

[5] 姚晟,李初宴,陈悦. 基于非平衡数据下不完备混合型信息系统的属性约简[J]. 计算机应用研究, 2021, 38(5):1331-1335.
YAO Sheng, LI Chuyan, CHEN Yue. Attribute reduction of incomplete hybrid information system based on unbalanced data [J]. Application Research of Computers, 2021, 38(5):1331-1335.

[6] HU Qinghua, YU Daren, LIU Jingfu, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18):3577-3594.

[7] FAN Xiaodong, ZHAO Weida, WANG Changzhong, et al. Attribute reduction based on max-decision neighborhood rough set model[J]. Knowledge-Based Systems, 2018, 151(1):16-23.

[8] SHU Wenhao, QIAN Wenbin, XIE Yonghong. Incremental feature selection for dynamic hybrid data using neighborhood rough set[J]. Knowledge-Based Systems, 2020, 194:105516.

[9] WANG Changzhong, SHI Yunpeng, FAN Xiaodong, et al. Attribute reduction based on *k*-nearest neighborhood rough sets[J]. International Journal of Approximate Reasoning, 2019, 106:18-31.

[10] CHEN Hongmei, LI Tianrui, FAN Xin, et al. Feature selection for imbalanced data based on neighborhood rough sets[J]. Information Sciences, 2019, 483:1-20.

[11] HU M, TSANG E C C, GUO Y T, et al. A novel approach to attribute reduction based on weighted neighborhood rough sets [J]. Knowledge-Based Systems, 2021, 220:106908.

[12] WANG Changzhong, HUANG Yang, SHAO Mingwen, et al. Feature selection based on neighborhood self-information[J]. IEEE Transactions on Cybernetics, 2020, 50(9):4031-4042.

[13] 孙林,赵婧,徐久成,等. 基于邻域粗糙集和帝王蝶优化的特征选择算法[J]. 计算机应用, 2022, 42(5):1355-1366.
SUN Lin, ZHAO Jing, XU Jiucheng, et al. Feature selection algorithm based on neighborhood rough set and monarch butterfly optimization[J]. Journal of Computer Applications, 2022, 42(5): 1355-1366.

[14] 熊菊霞,吴尽昭,王秋红.邻域互信息熵的混合型数据决策代价属性约简[J]. 小型微型计算机系统, 2021, 42(8): 1584-1590.
XIONG Juxia, WU Jinzhao, WANG Qiuhong. Decision cost attribute reduction of hybrid data based on neighborhood mutual information entropy[J]. Journal of Chinese Computer Systems, 2021, 42(8):1584-1590.

[15] 陈曦,刘晶. 基于邻域关系的知识粒度增量式属性约简算法[J]. 微电子学与计算机, 2020, 37(10):1-6.
CHEN Xi, LIU Jing. Knowledge granularity incremental attribute reduction algorithm based on neighborhood relation[J]. Microelectronics & Computer, 2020, 37(10):1-6.

[16] YANG Xiaoling, CHEN Hongmei, LI Tianrui, et al. Neighborhood rough sets with distance metric learning for feature selection[J]. Knowledge-Based Systems, 2021, 224:107076.

[17] WAN Jihong, CHEN Hongmei, YUAN Zhong, et al. A novel hybrid feature selection method considering feature interaction in neighborhood rough set[J]. Knowledge-Based Systems, 2021, 227:107167.

[18] SUN Lin, WANG Tianxiang, DING Weiping, et al. Feature selection using fisher score and multilabel neighborhood rough sets for multilabel classification[J]. Information Sciences, 2021, 578:887-912.

[19] 张雨新,孙达明,李飞. 基于粒化单调的不完备混合型数据增量式属性约简算法[J]. 计算机应用与软件, 2021, 38(3): 279-286.
ZHANG Yuxin, SUN Daming, LI Fei. Incremental attribute reduction algorithm for incomplete mixed data based on granulation monotony[J]. Computer Applications and Software, 2021, 38(3):279-286.

[20] 蔡艳婧,程实,王强. 不完备混合决策粗糙集特定类多目标属性约简[J]. 计算机工程与设计, 2020, 41(11):3063-3071.
CAI Yanjing, CHENG Shi, WANG Qiang. Class-specific multi-objective attribute reduction for incomplete mixed decision-theoretic rough set[J]. Computer Engineering and Design, 2020, 41(11):3063-3071.

[21] 李小南,赵璐,易黄建. 基于加权信息熵的直觉模糊信息系统的三支决策[J]. 控制与决策, 2022, 37(10):2705-2713.
LI Xiaonan, ZHAO Lu, YI Huangjian. Three-way decision of intuitionistic fuzzy information systems based on the weighted information entropy[J]. Control and Decision, 2022, 37(10):2705-2713.

[22] 徐怡,李宝峰,李策. 基于权重分布的多粒度粗糙集模型[J]. 模糊系统与数学, 2020, 34(6):55-67.
XU Yi, LI Baofeng, LI Ce. Multi-granulation rough set model based on weight distribution[J]. Fuzzy Systems and Mathematics, 2020, 34(6):55-67.

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

[24] HU Qinghua, ZHANG Lei, ZHANG David, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information[J]. Expert Systems with Applications, 2011, 38:10737-10750.

[25] DEMIAR J, SCHUURMANS D. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006, 7(1):1-30.

(编辑:李骏)