ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 1,Jan. 2025,pp. 117－133
DOI:10. 16337/j. 1004-9037. 2025. 01. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

混合层次依赖度下的邻域粗糙集多目标特征选择算法

骆公志，张尚蕾

(南京邮电大学管理学院,南京 210003)

摘　要：精度和效率是评判特征选择算法性能的关键指标,分别对应邻域粗糙集的属性依赖度和约简规模,而已有的特征选择算法通常以属性约简的最大依赖度为导向进行寻优,忽略了约简规模的重要性。现实中,随着数据特征维度的增加和类别层次结构的出现,导致类别信息复杂且结构关系混乱,传统属性依赖度计算未有效利用类别层次结构信息,使得分类性能不佳。鉴于此,本文构造了一种综合考量属性重要度和类别层次结构关系的混合层次依赖度,将混合层次依赖度和约简规模作为两个独立的优化目标,引入多目标进化算法对其分别进行优化,从属性依赖度和属性规模两方面提升所得属性约简的性能,以得到满足目标约束的约简结果。数据实验分析结果表明,所提算法能够在目标约束内得到更高质量的约简结果,并且能够提高分类精度。

关键词：多目标特征选择;邻域粗糙集;层次结构;混合层次依赖度;属性约简

中图分类号：TP181　　**文献标志码**:A

Multi-objective Feature Selection Algorithm for Neighborhood Rough Set Under Mixed Hierarchical Dependence

LUO Gongzhi, ZHANG Shanglei

(School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Accuracy and efficiency are the key metrics for evaluating the performance of feature selection algorithms. They correspond to the attribute dependence and reduction scale of neighborhood rough sets respectively. Conventional feature selection algorithms often optimize solely based on maximum attribute dependence reduction, overlooking the significance of reduction scale. However, as data feature dimensions increase and category hierarchies emerge, category information becomes complex and structural relationships become chaotic. Traditional attribute dependency calculations fail to effectively utilize category hierarchy information, leading to suboptimal classification performance. In response to this, a mixed hierarchical dependency that considers the relationship between attribute importance and category hierarchy structure is constructed. This treats mixed hierarchical dependency and reduction scale as two independent optimization objectives, and introduces a multi-objective evolutionary algorithm to optimize them independently. This approach improves attribute reduction performance from both the attribute dependency and attribute scale perspectives, resulting in reduction results that meet target constraints.

基金项目:国家自然科学基金(72171124);江苏高校哲学社会科学研究重大项目(2021SJZDA129);江苏省研究生科研创新计划项目(KYCX22_0884)。
收稿日期:2024-01-22;**修订日期**:2024-05-15

Experimental results demonstrate that the proposed algorithm achieves higher-quality reduction results within target constraints, leading to the improvement of classification accuracy.

Key words: multi-objective feature selection; neighborhood rough set; hierarchical structure; mixed hierarchical dependence; attribute reduction

引　言

大数据时代背景下,数据样本量和特征维数正在爆炸式增长[1],并且数据类别标签间的层次结构呈多样化,此类数据在图像数据、Web数据、音频数据、地理数据和生物数据等现实中广泛存在[2]。这些类别信息众多且结构关系复杂的高维数据导致分类学习任务面临维数灾难和类别不平衡的挑战。

最小属性约简的核心目标是从原始特征集合中提取出最具代表性的特征子集,旨在缩减特征数量的同时,保持提升分类精度的性能。近年来,邻域粗糙集属性约简算法备受研究者的关注。当前,该领域的研究重点在于通过精心构建属性重要度来提高约简集合的分类准确性,并致力于优化正域的计算效率。然而,最小属性约简作为一个NP难题[3],其计算复杂性随着特征数量的增加而呈指数级上升。因此,在实际应用中,更倾向于寻求在一定目标条件下满足需求的次优约简方案。

针对类别层次结构数据,已有基于邻域粗糙集的层次分类在线流特征选择算法[4]和基于层次类别邻域粗糙集的在线流特征选择算法[5]等,这些算法从不同角度进行特征选择并取得了一定的效果。然而,以上研究主要关注于寻求最大的属性依赖度,没有充分考虑属性约简的规模。属性依赖度和属性约简规模是评判属性约简性能的重要指标,分别对应分类精度和属性约简效率。如果只追求最大的属性依赖度,可能会导致属性规模过大,失去约简冗余属性的意义。特别是对于包含类别层次结构的高维数据,过多的属性不仅会占用额外的存储空间,还会增加计算的负担。传统的属性依赖度计算方法通常直接对超多类别的数据集进行一次性建模[6],没有充分利用层次结构信息,导致在处理大规模层次结构数据集时,传统粗糙集算法的分类性能低下。因此,在处理属性约简问题时,需要同时考虑如何利用层次类别信息来提高分类精度和如何控制属性约简规模。

特征选择问题可以视为一个非线性的组合优化问题[7]。为了解决这个问题,将属性约简拓展为多目标优化问题,并采用多目标进化算法MOEA_D(Multi-objective evolutionary algorithm based on decomposition)进行求解。由于传统的属性依赖度难以满足包含类别层次结构的高维数据的分类需求,构建了一种新的度量函数——混合层次属性依赖度。该度量函数充分利用层次结构间的依赖关系,从决策类、决策层和决策分类3个角度逐步集成得到,结合MOEA_D算法,提出了一种混合层次依赖度下的邻域粗糙集多目标特征选择算法MNMHD(Multi-objective feature selection based on neighborhood rough set mixed hierarchical dependence)。MNMHD算法不仅考虑了如何利用层次类别信息提高分类精度,还考虑了如何控制属性约简规模。通过将特征选择问题转化为多目标优化问题,MNMHD算法能够同时优化属性依赖度和属性约简规模两个目标,从而在保证分类精度的同时,提高特征选择的效率。这种算法不仅适用于处理包含类别层次结构的高维数据,还能为相关领域的特征选择提供一种新的思路和方法。具体研究思路如图1所示,在邻域决策信息系统的微观底层(D_j^l)、中观中层(U^l)和宏观高层(D)上实施层次构建,决策类、决策层和决策分类3个角度对应的属性依赖度为($SR_B(D_j^l), SR_B(U^l), SR_B(D)$),考虑到决策空间中的决策类粒度大小,构建了决策分辨度($S(U^l), S(D_j^l)$),将层次依赖度和决策依赖度异质融合,构造了混合层次依赖度$SR_B(D)$,与多目标优化算法MOEA_D结合,得到算法MNMHD。

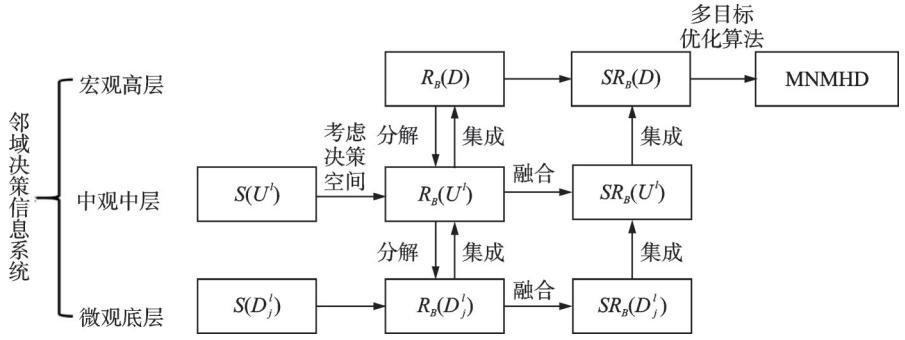

图1 本文方法研究思路

Fig.1 Research ideas of the proposed method

1 基本理论

1.1 类别层次结构

树结构的“从属”关系可归纳为:不可逆性、反自反性和传递性等[8],使用序$(D, \prec)$来定义层次结构,其中D为决策属性集,$\prec$代表从属关系,层次结构具有的3个特性描述如下:

(1)不可逆性:若$D_i \prec D_j, \forall D_i \subseteq D, \forall D_j \subseteq D$,则$D_j \not\prec D_i$;

(2)反自反性:$\forall D_i \subseteq D$,有$D_i \not\prec D_i$;

(3)传递性:若$\forall D_i \subseteq D, \forall D_j \subseteq D, \forall D_k \subseteq D, D_i \prec D_j$且$D_j \prec D_k$,则有$D_i \prec D_k$。

1.2 邻域粗糙集

粗糙集理论中,任意n维实数空间都可以表示成邻域决策信息系统$\mathrm{NDIS}=(U, A, V, f)$,其中,$U=\{x_1, x_2, \cdots, x_n\}$称为论域,$A=\{c_1, c_2, \cdots, c_m\}$为属性的有限非空集合,分为条件属性集$C$和决策属性集$D$,即$A=C \cup D, C \cap D=\varnothing, V=\cup V_a$,$V$为信息函数$f$的值域,$V_a$为属性$a$的值域,$a \in A$。

定义1(度量计算)[9] 设邻域决策信息系统为$\mathrm{NDIS}=(U, A, V, f)$,对于论域$U$中任意一点$x_i, B \subseteq C$,邻域半径为$\delta$,属性集$B$诱导的邻域定义为:$\delta_B(x_i)=\left\{x_j \middle| \triangle_B(x_i, x_j) \leqslant \delta\right\}, \delta \geqslant 0$,其中$\triangle(x, y)$为距离函数。

定义2(邻域决策信息系统及上下近似)[10] 设邻域决策信息系统为$\mathrm{NDIS}=(U, A, V, f), A=C \cup D$,决策属性$D$将论域$U$划分为$m$个决策划分类:$U/D=D_1, D_2, \cdots, D_m, \forall B \in C$,定义决策属性$D$关于条件属性$B$的下近似、上近似和边界区分别定义为

$$\underline{N_B D}=\bigcup_{j=1}^{m} \underline{N_B D_j}, \ \overline{N_B D}=\bigcup_{j=1}^{m} \overline{N_B D_j} \tag{1}$$

$$\mathrm{BND}_B(D)=\overline{N_B D}-\underline{N_B D} \tag{2}$$

$$\underline{N_B D_j}=\left\{x_i \middle| \delta_B(x_i) \subseteq D_j, \ x_i \in U\right\} \tag{3}$$

$$\overline{N_B D_j}=\left\{x_i \middle| \delta_B(x_i) \cap D_j \neq \varnothing, \ x_i \in U\right\} \tag{4}$$

定义3[11] 设邻域决策信息系统为$\mathrm{NDIS}=(U, A, V, f), A=C \cup D$,决策属性$D$对于条件属性子集$B \subseteq C$的依赖度为

$$R_B(D)=\frac{|N_B D|}{|U|} \tag{5}$$

2 混合层次依赖度下的邻域粗糙集模型

为了避免邻域大小的选择,结合层次结构中的粗细粒度类别信息,引入类别层次结构下邻域粒度计算方法。

定义 4 设邻域决策信息系统为 $\mathrm{HNDS}=(U,A,V,f)$, $A=C\cup D$,样本空间 U 的层次关系为 $U=U^1\cup U^2\cup\cdots\cup U^N$, N 为层级结构的层数。第 l 层的类别结构为 $U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$, $U^l\subseteq U$, $l\in[1,N]$, m 为该层决策类别的数量。$\forall x_{pj}\in D_j^l$, $\forall B\subseteq C$,在条件属性集 B 下样本 x_{pj} 的邻域点集为

$$\delta_B^l(x_{pj})=\begin{cases}\{x_{pj}\} & \mathrm{dis}_2^l(x_{pj})-\mathrm{dis}_1^l(x_{pj})>0\\ \varnothing & \text{其他}\end{cases}\tag{6}$$

在第 l 层的样本空间下,$\mathrm{dis}_1^l(x_{pj})$ 表示利用兄弟策略找到与样本与 x_{pj} 距离最近的同类样本,即

$$\mathrm{dis}_1^l(x_{pj})=\min\left(\Delta_B(x_{pj},x_{ij})\right)\quad \forall x_{pj}\in D_j^l\wedge\forall x_{ij}\in D_j^l-\{x_{pj}\}\tag{7}$$

$\mathrm{dis}_2^l(x_p)$ 表示在样本空间 U^l 中利用兄弟策略找到与样本与 x_{pj} 距离最近的异类样本,即

$$\mathrm{dis}_2^l(x_p)=\min\left(\Delta_B(x_{pj},y_{ij})\right)\quad \forall x_{pj}\in D_j^l\wedge\forall y_{ij}\in D_c^l\wedge D_c^l\subseteq U_l-D_j^l\tag{8}$$

若 $\mathrm{dis}_2^l(x_p)-\mathrm{dis}_1^l(x_{pj})>0$,说明样本 x_{pj} 是可区分的,所以将样本 x_{pj} 加入邻域点集 $\delta_B^l(x_{pj})$;否则,$\delta_B^l(x_{pj})$ 为空集。

定义 5 给定邻域决策信息系统 $\mathrm{HNDS}=(U,A,V,f)$, $A=C\cup D$,样本集 U 的决策类别存在层次关系 $U=U^1\cup U^2\cup\cdots\cup U^N$, N 为层级结构的层数。在类别层次结构上第 l 层的样本集为 $U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$, $l\in[1,N]$,其中 $U^l\subseteq U$,条件属性集 $B\subseteq C$ 关于 U 和 U^l 的下近似和上近似表示为

$$\underline{N_BD}=\bigcup_{l=1}^{N}\underline{N_BU^l}\tag{9}$$

$$\overline{N_BD}=\bigcup_{l=1}^{N}\overline{N_BU^l}\tag{10}$$

$$\underline{N_BU^l}=\bigcup_{j=1}^{m}\underline{N_BD_j^l}\tag{11}$$

$$\overline{N_BU^l}=\bigcup_{j=1}^{m}\overline{N_BD_j^l}\tag{12}$$

式中:$\underline{N_BD_j^l}=\left\{x_{pc}\middle|\,\delta_B^l(x_{pj})\subseteq D_j^l,\delta_B^l(x_{pj})\neq\varnothing,x_{pc}\in U^l\right\}$; $\overline{N_BD_j^l}=\left\{x_{pc}\middle|\,\delta_B^l(x_{pj})\cap D_j^l\neq\varnothing,x_{pc}\in U^l\right\}$; $j=1,2,\cdots,m$。

属性依赖度是一种重要不确定性度量方法[11],由于层次结构数据的决策类别众多且决策类别间存在语义结构关系,很难通过传统属性依赖度的一次度量完成条件属性重要度计算。受文献[12]所述的“三层思想”启发,构造了3种不同层面的属性依赖度计算方法:(1)决策分类层面(即 D 的属性依赖度 $R_B(D)$);(2)在层次决策分类层面(即 U^l)的属性依赖度 $R_B(U^l)$;(3)在层次决策类层面(即 D_j^l)的属性依赖度 $R_B(D_j^l)$。3种度量层次分别关联于层次结构邻域决策信息系统的宏观高层、中观中层和微观底层。

定义 6 (1)令类别层次结构的层数为 N,在层次决策分类层面的属性依赖度定义为

$$R_B(D)=\frac{1}{N}\sum_{l=1}^{N}R_B(U^l)\tag{13}$$

式中 $R_B(U^l)$ 为第 l 层的属性依赖度。

(2)在层次决策类层面的属性依赖度定义为

$$R_B(U^l)=\frac{1}{m}\sum_{j=1}^{m}R_B(D_j^l) \tag{14}$$

为了实施中层到高层的集成,定义6给出属性依赖度的两次度量分解,$R_B(D)$位于层次决策高层,按照层次结构分层分解为$R_B(U^l)$, $R_B(U^l)$位于层次决策中层,进一步可以析出位于层次决策底层的$R_B(D_j^l)$,这种中层度量自然集成对应高层度量的优良性质和相关特征。同理,底层度量同样继承了中层相应的性质和特征。由此,先在底层计算$R_B(D_j^l)$,再集成至中层$R_B(U^l)$,最后层次集成至高层度量$R_B(D)$最终的功能和性质也能有效实现。

层次依赖度的集成取决于决策类别D,而层次邻域决策信息系统中决策类别D的空间结构复杂,决策类别间存在父子关系和兄弟关系。如图2所示,层次结构有两层,其中D^0为虚拟节点,D_1^1和D_2^1属于第1层,D_1^2和D_2^2属于第2层且依赖于D_2^1。从决策类分类角度出发,D_1^1和D_2^1的分类错误所产生的负面影响显然不同,D_2^1一旦分类发生错误,会影响到D_1^2和D_2^2的分类结果,这种差异无疑会影响属性重要度的计算精度。

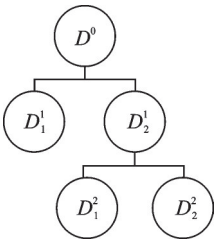

图2 样本决策类别的层次结构

Fig.2 Hierarchical structure of sample decision categories

为了解决这一问题,引入决策分辨度的概念来区分同一层次间不同依赖关系的决策类别。

定义7 邻域决策信息系统为$\mathrm{HNDS}=(U,A,V,f)$,$A=C\cup D$,D_j^l为第l层的第j个决策类别,D_j^l包含的决策类别数为$Q(D_j^l)$,第l层包含的决策类别数定义为

$$Q(U^l)=\sum_{j=1}^{m}Q(D_j^l) \tag{15}$$

D_j^l的决策分辨度可以定义为

$$S(D_j^l)=\frac{Q(D_j^l)}{Q(U^l)} \tag{16}$$

U^l的决策分辨度可以定义为

$$S(U^l)=\frac{Q(U^l)}{\sum_{i=1}^{M}Q(U^l)} \tag{17}$$

$Q(D_j^l)$描述D_j^l及所包含的子类别数量和,$Q(U^l)$描述第l层的决策类别数量和包含的子类别数量,$Q(D_j^l)$的取值范围为$Q(D_j^l)\in\left[1,Q(U^l)-1\right]$,$Q(U^l)\geqslant 2$。$S(D_j^l)$和$S(U^l)$主要描述微观底层和宏观中层所对应决策空间的层次结构,而$R(D_j^l)$和$R(U^l)$表示条件属性对决策属性的依赖程度,两者的融合能够产生更有效的不确定性度量。从依赖度的角度来看,引入不同层的决策分辨度更适用于层次结构决策信息系统的条件属性重要度计算,更好地实施层次结构决策信息系统的刻画与应用,故构造了一种混合层次依赖度。

定义8 邻域决策信息系统为$\mathrm{HNDS}=(U,A,V,f)$,$A=C\cup D$,层数为N,$B\subseteq C$,微观底层的混合层次依赖度定义为

$$SR_B(D_j^l)=R_B(D_j^l)S(D_j^l)=\frac{\left|\underline{N_B}D_j^l\right|}{|U|}\frac{Q(D_j^l)}{Q(U^l)} \tag{18}$$

中观中层的混合层次依赖度定义为

$$SR_B(U^l)=\frac{1}{m}\sum_{j=1}^{m}SR_B(D_j^l)=\frac{1}{m}\sum_{j=1}^{m}\frac{\left|\underline{N_BD_j^l}\right|}{|U|}\frac{Q(D_j^l)}{Q(U^l)} \tag{19}$$

进而宏观高层的混合层次依赖度定义为

$$SR_B(D)=\frac{1}{N}\sum_{l=1}^{N}SR_B(U^l)S(U^l)=\frac{1}{N}\sum_{l=1}^{N}\frac{\left|\underline{N_BD^l}\right|}{|U|}\frac{Q(U^l)}{Q(U)} \tag{20}$$

定理 1 设邻域决策信息系统为 $\mathrm{HNDS}=(U,A,V,f)$, $A=C\cup D$, 论域 U 的层次结构为 $U=U^1\cup U^2\cup\cdots\cup U^N$, N 为论域 U 的类别结构层数, 其第 l 层的类别结构为 $U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$, 这里, D_m^l 为第 l 层中第 m 个决策类别, $l\in[1,N]$。$\forall U^l\subseteq U, B\subseteq C$, 则有

$$SR_B(D)\in\left[0,\frac{Q(U^l)-1}{Q(U^l)}\right]$$

证明 由定义 7 可知 $Q(D_j^l)\in\left[1,Q(U^l)-1\right]$, 从而可得 $\frac{Q(D_j^l)}{Q(U^l)}\in\left[\frac{1}{Q(U^l)},\frac{Q(U^l)-1}{Q(U^l)}\right]$, 又因为 $S(D_j^l)=\frac{Q(D_j^l)}{Q(U^l)}$, 从而有 $S(D_j^l)\in\left[\frac{1}{Q(U^l)},\frac{Q(U^l)-1}{Q(U^l)}\right]$。根据定义 3 有 $R_B(D_j^l)\in[0,1]$, 从而有 $S(D_j^l)R_B(D_j^l)\in\left[0,\frac{Q(U^l)-1}{Q(U^l)}\right]$, $SR_B(D_j^l)=S(D_j^l)R_B(D_j^l)$, 即 $SR_B(D_j^l)\in\left[0,\frac{Q(U^l)-1}{Q(U^l)}\right]$, 进而 $\frac{1}{m}\sum_{j=1}^{m}SR_B(D_j^l)\in\left[0,\frac{1}{m}\times m\frac{Q(U^l)-1}{Q(U^l)}\right]$, 又因为 $SR_B(U^l)=\frac{1}{m}\sum_{j=1}^{m}SR_B(D_j^l)$, 所以有 $SR_B(U^l)\in\left[0,\frac{Q(U^l)-1}{Q(U^l)}\right]$。由定义 7 可得 $S(U^l)\in\left[\frac{2}{Q(U)},1\right]$, 故 $SR_B(U^l)S(U^l)\in\left[0,\frac{Q(U^l)-1}{Q(U^l)}\times 1\right]$, 进一步计算得到 $\frac{1}{N}\sum_{l=1}^{N}SR_B(U^l)S(U^l)\in\left[0,\frac{1}{N}\times N\frac{Q(U^l)-1}{Q(U^l)}\times 1\right]$, 由于 $SR_B(D)=\frac{1}{N}\sum_{l=1}^{N}SR_B(U^l)S(U^l)$, 可得 $SR_B(D)\in\left[0,\frac{Q(U^l)-1}{Q(U^l)}\right]$, 证毕。

决策分辨度刻画决策属性 D 的空间结构, 与特征子集 $B\subseteq C$ 无关, 和层次依赖度进行乘积混合集成, 没有改变混合层次依赖度的非负性, 可以更好地适用于层次邻域决策信息系统的条件属性重要度计算。

定理 2 设邻域决策信息系统为 $\mathrm{HNDS}=(U,A,V,f)$, $A=C\cup D$, U 的层次关系为 $U=U^1\cup U^2\cup\cdots\cup U^N$, N 为论域 U 的类别结构层数, 第 l 层的类别结构为 $U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$, $U^l\subseteq U, D_m^l$ 为第 l 层中第 m 个决策类别, $l\in[1,N]$。若 $\delta_{B_1}(x)=\delta_{B_2}(x)$, $\forall x\in U, B_1\subseteq C, B_2\subseteq C$, 则 $SR_{B_1}(D)=SR_{B_2}(D)$。

证明 由于 B_1、$B_2\subseteq C, \delta_{B_1}(x)=\delta_{B_2}(x)$, 根据定义 5 可得 $\underline{N_{B_1}D}=\underline{N_{B_2}D}$, 进而由定义 3 可得 $R_{B_1}(D_j^l)=R_{B_2}(D_j^l)$, $S(D_j^l)R_{B_1}(D_j^l)=S(D_j^l)R_{B_2}$, 即 $SR_{B_1}(D_j^l)=SR_{B_2}(D_j^l)$, 进一步集成可得

$\frac{1}{m}\sum_{j=1}^{m}SR_{B_1}(D_j^l)=\frac{1}{m}\sum_{j=1}^{m}SR_{B_2}(D_j^l)$, 有 $SR_{B_1}(U^l)=SR_{B_2}(U^l)$, $SR_{B_1}(U^l)S(U^l)=SR_{B_2}(U^l)S(U^l)$, 又有 $\frac{1}{N}\sum_{l=1}^{N}SR_{B_1}(U^l)S(U^l)=\frac{1}{N}\sum_{l=1}^{N}SR_{B_2}(U^l)S(U^l)$,故可得$SR_{B_1}(D)=SR_{B_2}(D)$,证毕。

基于决策系统的三层思想[12],$SR_{B_1}(D)$在决策类底层进行依赖度和决策分辨度的混合,再实施决策层中层混合度量的层次构建,最后集成到决策分类高层,这种操作带来的分层混合属性依赖度自然继承高层度量的优良性质,满足粒化不变性和取值非负性,更适用于实施层次邻域决策系统的特征选择。

定义 9 给定邻域决策信息系统为 $HNDS=(U,A,V,f)$, $A=C\cup D$,类别层次结构为 $U=U^1\cup U^2\cup\cdots\cup U^N$, N为论域U的类别结构层数,第l层的类别结构为 $U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$。条件属性集$B\subseteq C$在第$l$层的正区域定义为

$$\mathrm{Pos}_B(U^l)=\underline{N_BU^l}=\bigcup_{j=1}^{m}\underline{N_BD_j^l} \tag{21}$$

式中 $\underline{N_BD_j^l}=\left\{x_{pc}\middle|\delta_B^i(x_{pj})\subseteq D_j^l\wedge\delta_B^i(x_{pj})\neq\varnothing,\forall x_{pj}\in D_j^l\right\}$。

定理 3 设邻域决策信息系统为 $HNDS=(U,A,V,f)$, $A=C\cup D$, U的层次结构为 $U=U^1\cup U^2\cup\cdots\cup U^N$, N为论域U的类别结构层数,第l层的类别结构为 $U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$, $U^l\subseteq U$。若$\forall B_1\subseteq C$, $\forall B_2\subseteq C$, $B_1\subseteq B_2$, $SR_{B_1}(D)\leqslant SR_{B_2}(D)$。

证明 由于$B_1\subseteq B_2$,不妨令 $B_1=B_2-\{b\}$, $b\in B_2$,有 $\delta_{B_2-(b)}(x)\leqslant\delta_{B_2}(x)$,条件属性越多,对论域的分类越细,下近似的对象数量增加,而根据定义5,又有 $\underline{N_{B_2-b}D_j^l}\leqslant\underline{N_{B_2}D_j^l}$, $\mathrm{Pos}_{B_2-\{b\}}(D_j^l)\leqslant\mathrm{Pos}_{B_2}(D_j^l)$,因此有 $\frac{\left|\mathrm{Pos}_{B_2-\{b\}}(D_j^l)\right|}{|U|}\leqslant\frac{\left|\mathrm{Pos}_{B_2}(D_j^l)\right|}{|U|}$,进而 $S(D_j^l)\frac{\left|\mathrm{Pos}_{B_2-\{b\}}(D_j^l)\right|}{|U|}\leqslant S(D_j^l)\frac{\left|\mathrm{Pos}_{B_2}(D_j^l)\right|}{|U|}$,即 $SR_{B_2-\{b\}}(D_j^l)\leqslant SR_{B_2}(D_j^l)$;依据微观底层集成到中观中层的概念,有 $\frac{1}{m}\sum_{j=1}^{m}SR_{B_2-\{b\}}(D_j^l)\leqslant\frac{1}{m}\sum_{j=1}^{m}SR_{B_2}(D_j^l)$,进而 $SR_{B_2-\{b\}}(U^l)\leqslant SR_{B_2}(U^l)$,因此有 $\frac{1}{|N|}\sum_{l=1}^{|N-1|}SR_{B_2-\{b\}}(U^l)S(U^l)\leqslant\frac{1}{|N|}\sum_{l=1}^{|N-1|}SR_{B_2}(U^l)S(U^l)$,进而$SR_{B_2-\{b\}}(D)\leqslant SR_{B_2}(D)$,即$SR_{B_1}(D)\leqslant SR_{B_2}(D)$,证毕。

$S(U^l)$为决策空间的统计特征,不改变$SR_B(U^l)$的单调性,且$SR_B(U^l)$是层次依赖度$R_B(U^l)$和决策分辨度$S(U^l)$的集成,满足取值非负性、粒化不变性和粒化单调性,相关性质来源于基度量$R_B(D)$的对应性质。相较于$R_B(D)$,$SR_B(D)$具有度量融合更充分和度量更深入的特点,更适用于高精度的融合需要。

定理 4 设邻域决策信息系统 $HNDS=(U,A,V,f)$, $A=C\cup D$,类别层次结构为 $U=U^1\cup U^2\cup\cdots\cup U^N$, N为层数,第l层的类别结构为 $U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$。若$N=1$,则$SR_B(D)$退化为$\frac{1}{m^2}R_B(D)$。

证明 $N=1$,层次结构数据退化到扁平数据,可得

$$U^1=U,D_j^1=D_j$$

$$S(D_1^1)=S(D_2^1)=\cdots=S(D_j^1)=\frac{1}{m}\qquad j=1,2,\cdots,m$$

在特征子集B的条件下,依赖度为

$$SR_B(D)=\frac{1}{N}\sum_{l=1}^{N}S(U^l)SR_B(U^l)=\frac{1}{1}\sum_{l=1}^{1}S(U^l)SR_B(U^l)=$$

$$S(U^1)SR_B(U^1)=\frac{1}{m}\sum_{j=1}^{m}S(D_j^1)SR_B(D_j^1)=\frac{1}{m}\sum_{j=1}^{m}S(D_j)SR_B(D_j)=$$

$$\frac{1}{m}\sum_{j=1}^{m}\frac{1}{m}\frac{\left|\underline{N_BD_j}\right|}{|U|}=\frac{1}{m^2}\frac{\sum_{j=1}^{m}\left|\underline{N_BD_j}\right|}{|U|}=\frac{1}{m^2}\frac{\left|\underline{N_BD_j}\right|}{|U|}=\frac{1}{m^2}R_B(D)$$

证毕。

由此可知,混合层次依赖度退化为一般属性依赖度,也可以退化应用于扁平数据信息系统。

定理5 给定邻域决策信息系统为$\mathrm{HNDS}=(U,A,V,f)$,$A=C\cup D$,$B\subseteq C$,类别层次结构为$U=U^1\cup U^2\cup\cdots\cup U^N$,$N$为论域$U$的类别结构层数。若第$l$层的类别结构为$U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$,则不同类别层次上的正区域具有如下关系

$$l_1<l_2\Rightarrow\mathrm{Pos}_B(U^{l_2})\subseteq\mathrm{Pos}_B(U^{l_1}) \tag{22}$$

证明 邻域决策信息系统$\mathrm{HNDS}=(U,A,V,f)$中,$U^l=D_1^l\cup D_2^l\cup\cdots\cup D_m^l$,若$l_1<l_2$,由于底层决策类$U^{l_2}$是上一层决策类$U^{l_1}$的分解,则$\forall D_j^{l_1}\subseteq U^{l_1}$,$\exists D_j^{l_2}\subseteq U^{l_2}$,$D_j^{l_2}\preccurlyeq D_j^{l_1}$,而对于$\forall x_{pj}\in\mathrm{Pos}_B(U^{l_2})$,根据定义9有$\delta_B^{l_2}(x_{pj})\subseteq D_j^{l_2}$,从而有$\delta_B^{l_2}(x_{pj})\subseteq D_j^{l_1}$,再由$x_{pi}\in\mathrm{Pos}_B(U^{l_1})$,可得$\mathrm{Pos}_B(U^{l_2})\subseteq\mathrm{Pos}_B(U^{l_1})$,证毕。

由定理5可知,自顶向下的层次样本集中决策类不断细化,底层特征选择的本质是对上一层的正区域进一步分类。因此,将上层的正区域作为底层的论域可以减少重复计算,同时更好地度量在细粒度上特征的分类性能。这里提供1个包含层次结构的决策信息表实例,说明混合层次依赖度的概念和具体计算过程。

例1 如表1所示,邻域决策信息系统中包含了9个对象$U=\{x_1,x_2,\cdots,x_9\}$,C为条件属性,D为决策属性。

表1 层次结构样本数据实例

Table 1 Sample data example of hierarchical structure

U	C	D
x_1	0.10	1
x_2	0.20	1
x_3	0.40	2
x_4	0.60	3
x_5	0.80	4
x_6	0.82	3
x_7	0.81	4
x_8	0.60	2
x_9	0.82	4

决策属性D中的类别层次结构如图2所示,总层数N为2。由表1可得,$U^1=9$,U^1包含两个决策类$U^1/D^1=\{D_1^1,D_2^1\}$,$U^1/D_1^1=\{x_1,x_2\}$和$U^1/D_2^1=\{x_3,x_4,\cdots,x_9\}$,$U^2=5$,$U^2$被划分为两类$U^2/D^2=\{D_1^2,D_2^2\}$,其中$D_1^2=\{x_4,x_6\}$和$D_2^2=\{x_5,x_7,x_9\}$。$D_1^1$为子类别,$Q(D_1^1)=1$,$D_2^1$为父类别,$D_1^2$、$D_2^2$依赖于$D_2^1$,故$Q(D_2^1)=3$,由式(14)可得,第1层决策类别数为:$Q(U^1)=4$;同理,$Q(D_1^2)=1$,$Q(D_2^2)=1$,$Q(U^2)=2$,根据式(16)可得,第1层的决策分辨度为:$S(U^1)=\dfrac{Q(U^1)}{Q(U^1)+Q(U^2)}=\dfrac{2}{3}$,同理可得,$S(U^2)=\dfrac{1}{3}$;由式(16)可得,决策类别$D_1^1$的分辨度为:$S(D_1^1)=\dfrac{Q(D_1^1)}{Q(U^1)}=\dfrac{1}{4}$。同理可得:$S(D_2^1)=\dfrac{Q(D_2^1)}{Q(U')}=\dfrac{3}{4}$,$S(D_1^2)=\dfrac{Q(D_1^2)}{Q(U^2)}=\dfrac{1}{2}$,$S(D_2^2)=$

$\frac{Q(D_2^2)}{Q(U^2)}=\frac{1}{2}$。

第1层中,条件属性集C上样本x_1的最近异类为x_3,最近同类为x_2,由式(6)计算可得:$H_{pj}^l-M_{pj}^l=0.2>0$。所以$\delta_C^1(x_1)=\{x_1\}$。同理可得

$\delta_C^1(x_2)=\{x_2\},\delta_C^1(x_3)=\varnothing,\delta_C^1(x_4)=\{x_4\},\delta_C^1(x_5)=\{x_5\},\delta_C^1(x_6)=\{x_6\},\delta_C^1(x_7)=\{x_7\},$

$\delta_C^1(x_8)=\{x_8\},\delta_C^1(x_9)=\{x_9\}$

由定理1可得,D_1^1和D_2^1的下近似为

$$\underline{N_CD_2^1}=\{x_4,x_5,x_6,x_7,x_8,x_9\},\ \underline{N_CD_1^1}=\{x_1,x_2\}$$

根据混合层次依赖度计算式(17)可得

$$SR_C(D_1^1)=\frac{\left|\underline{N_CD_1^1}\right|}{|U|}\frac{Q(D_1^1)}{Q(U^1)}=\frac{1}{18}$$

同理可得,$SR_C(D_2^1)=\frac{7}{8}$,根据式(18)可得

$$SR_C(U^1)=\frac{1}{2}\sum_{j=1}^{2}SR_C(D_j^1)=\frac{5}{12}$$

根据定理3可得,D_i^2和D_2^2的下近似为

$$\underline{N_CD_1^2}=\{x_4\},\ \underline{N_CD_2^2}=\{x_5,x_7,x_9\}$$

进一步可得

$$SR_C(D_1^2)=R_C(D_1^2)S(D_1^2)=\frac{1}{4}$$

$$SR_C(D_2^2)=R_C(D_2^2)S(D_2^2)=\frac{1}{2}$$

$$SR_C(U^2)=\frac{1}{2}\sum_{j=1}^{2}SR_C(D_j^2)=\frac{3}{8}$$

根据式(19)可得

$$SR_C(D)=\frac{1}{2}\sum_{l=1}^{2}SR_C(U^l)S(U^l)=\frac{19}{48}$$

即决策类别D以$SR_C(D)=\frac{19}{48}$依赖于条件属性集C。

3 本文算法

3.1 多目标特征选择问题

在图像识别和生物学等领域中,数据集通常拥有超多类别且类别之间存在层次结构[13],而传统的特征选择算法没有有效地利用层次结构信息,导致分类性能不理想,从而无法处理大规模数据,所以文中将混合层次依赖度引入基于邻域粗糙集的特征选择问题,然而只基于混合层次依赖度这一个目标开展粗糙集算法设计,忽略属性约简规模的重要性,会导致属性约简个数的增加,从而增大分类时的计算代价。因此,综合考虑利用层次结构来最大化分类性能和最小化属性约简个数,基于属性依赖度和属性约简个数,多目标特征选择问题可表示为

$$\begin{cases} f_1 = \max RS(a_1, a_2, \cdots, a_m) \\ f_2 = \max \dfrac{1}{\sum\limits_{i=1}^{m} a_i} \\ \text{s.t.} \quad \sum\limits_{i=1}^{m} a_i \leqslant m, a_i = 0, 1 \end{cases} \tag{23}$$

式中：a_i为0-1变量，$a_i=1$表示条件属性c_i为关键属性，否则$a_i=0$表示条件属性c_i为冗余属性，可以被约简；f_1表示最大化层次依赖度，f_2表示最小化属性约简的规模。进一步，将式(23)表示为

$$\begin{cases} \max F(X) = [f_1(X), f_2(X)] \\ \text{s.t.} \quad X \in S \end{cases} \tag{24}$$

式中：X为条件属性集的解；S表示为条件属性集的解空间。

定义10（条件属性集的支配关系）[14] 设条件属性集$X_A \subseteq S, X_B \subseteq S, f_i(X_A) \geqslant f_i(X_B), i=1,2,\cdots,n, \exists j \in \{1,2,\cdots,n\}$，使得$f_i(X_A) > f_i(X_B), X_A \prec X_B$，$X_B$可定义为支配条件属性集，即条件属性集$X_A$支配$X_B$。

定义11（非支配条件属性集）[14]若条件属性集$X_B \in S$，且X_B不被任何条件属性集支配，则X_B可定义为非支配条件属性集。

定义12（属性约简的Pareto解集）[14] 设$X_B \in S$为非支配条件属性集，属性约简的Pareto解集由解空间S中所有非支配条件属性集的目标函数值的集合组成，Pareto解集PF定义为

$$\mathrm{PF} = \{f_1(X_B), f_2(X_B)\} \tag{25}$$

在定义8中，为了衡量邻域粗糙集的属性重要度，设计了混合层次依赖度这一有效的度量函数。该函数能够综合考虑数据集各个特征在不同层次上的贡献程度，从而在处理具有复杂类别层次结构的邻域决策信息系统时具有更高的适用性。混合层次依赖度和属性约简规模设定为两个相互独立的优化目标，属于非支配关系。优化非支配关系问题的常用多目标优化方法有MOEA_D和NSGA-Ⅱ(Non-dominated sorting genetic algorithm-Ⅱ)算法[15]。在高维数据集的特征选择问题中，相较于NSGA-Ⅱ算法，MOEA_D算法获得相似质量的PF解集的计算复杂度更低。由于考虑到包含类别层次结构关系的数据集的规模较大，文中的多目标优化方法选用MOEA_D算法，将混合层次依赖度下的邻域粗糙集模型与其结合，以此构建了混合层次依赖度下的邻域粗糙集多目标特征选择模型。这一模型在优化过程中能够同时考虑到属性依赖度和属性约简规模这两个关键因素，从而在特征选择过程中取得了更好的效果。

3.2 算法设计

由定义6给出的混合层次依赖度，引入多目标进化算法MOEA_D来设计MNMHD，具体步骤如下所示。

输入：邻域决策信息系统$\mathrm{HNDS}=(U, A, V, f)$，子问题的个数$P$，均匀分布的$K$个权重向量$\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2, \cdots, \boldsymbol{\lambda}^P$，每个权重向量的邻居个数$T$。

输出：Pareto解集PF。

(1) 计算任意两个权向量之间的欧氏距离，然后求解每个权向量的T个邻居。对于$i=1,2,\cdots,K$，即索引集合$B(i)=\{i_1, i_2, \cdots, i_T\}$，$\boldsymbol{\lambda}^{i_1}, \boldsymbol{\lambda}^{i_2}, \cdots, \boldsymbol{\lambda}^{i_T}$为离$\boldsymbol{\lambda}^i$最近的$T$个向量；

(2) 种群初始化。根据2.2节,随机生成一个规模POP,维度为$|C|$的初始种群$X^1, X^2, \cdots, X^{POP}$,其中$X_i=\{a_1, a_2, \cdots, a_m\}$, $a_i\in[0,1]$;令$FV^i=F(X^i)$,初始化理想点$z=(z_1, z_2, \cdots, z_K)^T$;//初始化

(3) for $i=1, 2, \cdots, K$

(4) 基因重组:从$B(i)$中随机选择2个索引k和1利用均匀交叉方法产生新的解Y;

(5) For $j=1, 2, \cdots, m$

(6) if $z_j<f_j(Y)$

(7) then $z_j=f_j(Y)$// 更新z

(8) end

(9) for $\forall j\in B(i)$

(10) If $F\left(Y|\lambda^j, z\right)\leqslant F\left(X^j|\lambda^j, z\right)$

(11) then $X^j=Y$ $FV^j=F(Y)$

(12) end

(13) 从PF中移除$F(Y)$的支配向量,若PF中不存在支配$F(Y)$的向量,则将$F(Y)$加入PF;//更新PF

(14) end

(15) end

(16) end

(17)输出,若符合终止条件,则终止计算,输出Pareto解集PF,否则转到(3)。

MNMHD算法的主要步骤包括权重向量的产生、分解策略和子问题求解。首先,随机生成一个规模POP,维度为$|C|$的初始种群$X^1, X^2, \cdots, X^{POP}$,初始化种群的时间复杂度为$O(POP)$。根据MOEA_D的分解策略,将多目标问题分解为$P$个子问题,每个子问题使用$K$个权重向量进行求解,分解策略的时间复杂度为$O(P)$。接下来,每个子问题使用$K_K$个权重向量进行求解,每个权重向量有$T$个邻居解,故子问题求解的时间复杂度为$O(P\times K\times T)$。最后,将每个子问题的最优解组合起来形成原问题的解,时间复杂度为$O(P\times|C|)$。综合以上步骤,MNMHD的时间复杂度为$O(POP)+O(P)+O(P\times K\times T)+O(P\times|C|)=O(POP+PKT+P|C|)$。

4 实验分析

实验运行的硬件环境为:AMD R7-5800H CPU和16 GB RAM,算法均在Windows 10环境下的Matlab R2020a进行编程实现。MNMHD的参数设置如下:种群数量为100,算法迭代50次,交叉概率$CR=0.9$。

4.1 数据集介绍

为了验证算法MNMHD的优越性,本节设置了3组实验,选取了5个包含层次结构的公共数据集进行实验分析,具体信息如表2所示。Bridges是关于桥梁识别的标准数据集,包含108个样本,来源于美国加州大学创建的UCI数据库;DD[16]是1个蛋白质数据集,拥有27个真实类别,包含3 020个训练蛋白质序列,605个测试蛋白质序列;F194[17]是1个包含473个特征和194个真实类别的蛋白质数据集,包含7 015个蛋白质序列;VOC[18]是1个关于视觉对象的分类识别监测的数据集,拥有88个真实类别和7 178个样本,具有4层树结构;SAIAPR[19]是图像分割数据集,拥有512个真实特征,包含了5 000条样本。

表 2 数据集描述

Table 2 Data set description

序号	数据集	样本数	特征数	内部节点数	叶子节点数	层数
1	Bridges	108	12	8	6	2
2	DD	3 625	473	32	27	3
3	F194	7 015	473	202	194	3
4	VOC	7 178	1 000	30	88	4
5	SAIAPR	5 000	512	56	200	2

4.2 层次评价指标

除了传统的分类精度,针对层次结构中的错分程度,额外引入两种评价算法性能的层次分类指标:树诱导损失(Tree induced error, TIE)、最近公共(Lowest common ancestor-F_1, F_{LCA})。对于TIE和F_{LCA}两个层次分类评价指标,TIE值越小越好,而F_{LCA}取值越大越好。

不同的分类错误导致不同程度的惩罚,TIE利用类别在树结构中的距离来定义惩罚,设正确的类别标签为$\hat{d}$,预测的类别标签为d,TIE可以计算为:$\mathrm{TIE}(d,\hat{d})=|\mathrm{EH}(d,\hat{d})|$,其中$\mathrm{EH}(d,\hat{d})$为层次结构中$d$到$d$路径上的边集,$|\cdot|$为元素的个数。图2中树诱导损失分别计算为$\mathrm{TIE}(D_1^2,D_2^2)=2$, $\mathrm{TIE}(D_1^1,D_1^2)=3$。为了方便观察,实验统一取平均值。

TIE仅考虑预测类别与真实类别之间的关系。不同于TIE的度量,F_{LCA}考虑了真实类别和预测类别的最近共同祖先增广集,其度量过程主要分为两个阶段:

(1)利用类的层次关系来求增广集,真实标签是d,预测标签是$\hat{d}$,其增广集分别表示为

$$d_{\mathrm{aug}}^{\mathrm{LCA}}=d\cup\mathrm{LCA}(d,\hat{d})$$
$$\hat{d}_{\mathrm{aug}}^{\mathrm{LCA}}=\hat{d}\cup\mathrm{LCA}(d,\hat{d})$$

(2)通过增广集来度量分类错误的惩罚,分层精度、召回率以及分层的F_{LCA}的定义分别如下

$$P_{\mathrm{LCA}}=\frac{\left|d_{\mathrm{aug}}^{\mathrm{LCA}}\cap\hat{d}_{\mathrm{aug}}^{\mathrm{LCA}}\right|}{\left|\hat{d}_{\mathrm{aug}}^{\mathrm{LCA}}\right|}$$
$$R_{\mathrm{LCA}}=\frac{\left|d_{\mathrm{aug}}^{\mathrm{LCA}}\cap\hat{d}_{\mathrm{aug}}^{\mathrm{LCA}}\right|}{|d_{\mathrm{aug}}^{\mathrm{LCA}}|}$$
$$F_{\mathrm{LCA}}=\frac{2P_{\mathrm{LCA}}R_{\mathrm{LCA}}}{P_{\mathrm{LCA}}+R_{\mathrm{LCA}}}$$

将例1作为示例,结合图2中的类别层次关系,当$d=D_1^2$、$\hat{d}=D_2^2$时,$P_{\mathrm{LCA}}=\frac{1}{2}$, $R_{\mathrm{LCA}}=\frac{1}{2}$, $F_{\mathrm{LCA}}=\frac{1}{2}$。

为了有效评价MNMHD的分层分类效果,设计了两种层次邻域决策信息系统的特征选择算法作为对比算法:MNHD(Multi-objective feature selection algorithm based on neighborhood rough set and hierarchical dependence)是基于邻域粗糙集层次依赖度的多目标特征选择算法,该算法只采用层次依赖度评估特征而不考虑决策分辨度来验证混合层次依赖度的合理性;MNMHD-Ⅱ是基于邻域粗糙集混合层次依赖度和NSGA-Ⅱ的多目标特征选择算法。根据文献[16]可知,相较于NSGA-Ⅱ算法,MOEA_D算法有更少的计算复杂度。为了验证MOEA_D算法对混合层次依赖度和属性约简规模的优化效果,

在分类精度方面将MNMHD和MNMHD-Ⅱ进行对比,以验证基于邻域粗糙集混合层次依赖度与MOEA_D算法结合的合理性。

4.3 分类效果对比

本节采用线性支持向量机(Linear support vector machine, LSVM)和K-近邻分类器(K-nearest neighbor, KNN)来评价算法MNMHD、MNHD、MNMHD-Ⅱ所选特征的分类精度。根据文献[4]可知,KNN分类器中邻近点的个数为10时,各算法表现效果较好,故文中设置$k=10$。给出在3个数据集上不同算法的分类性能比较。为了保证公平,在保持其他参数不变的情况下改变所选特征的数量,将平均分类精度作为最终结果[20]。图3展示了应用于3个数据集的各个特征选择算法的分类精度。

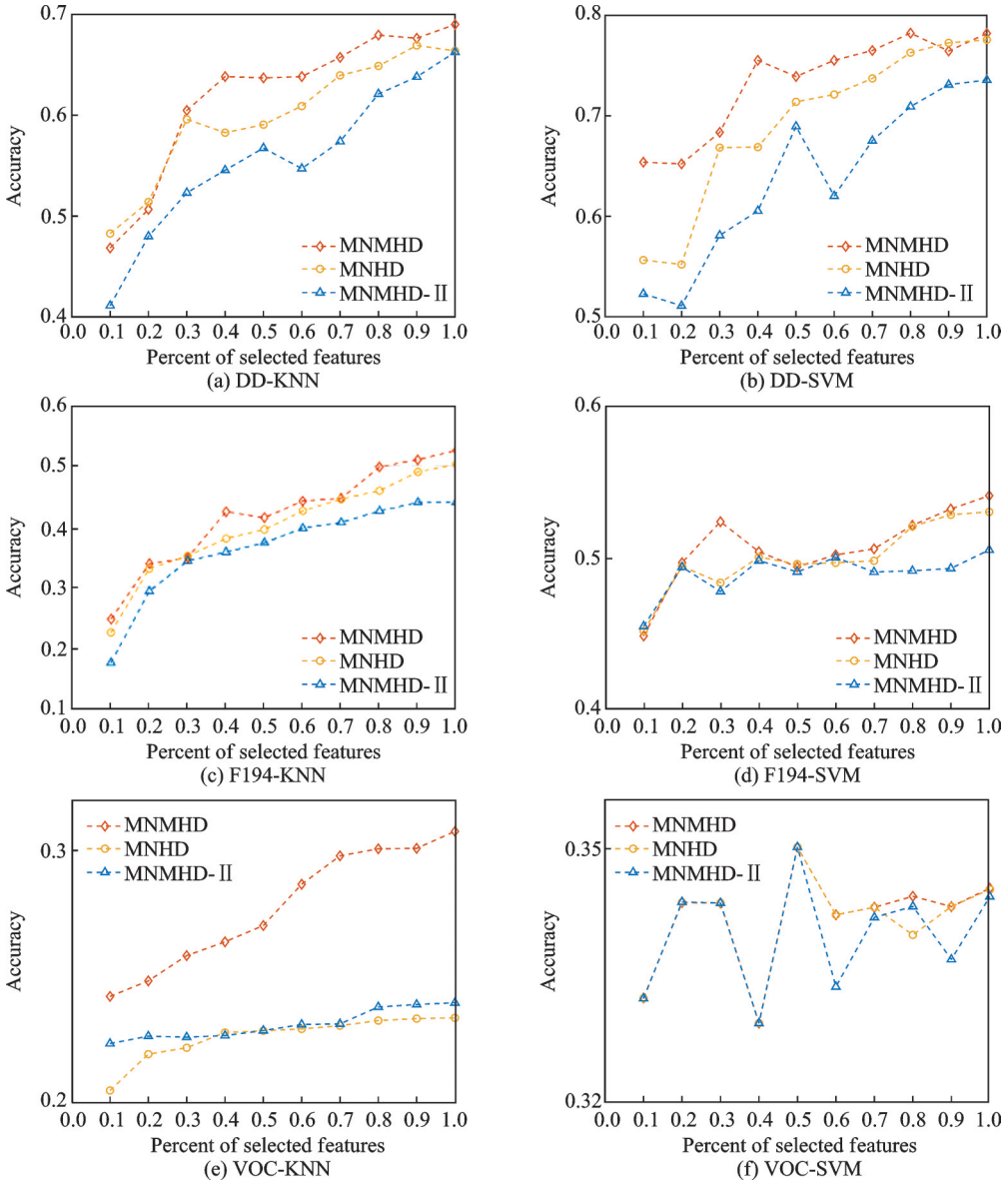

图3 MNMHD、MNHD和MNMHD-Ⅱ在3个数据集上选择不同特征数量的分类精度

Fig.3 Classification accuracy of MNMHD, MNHD, and MNMHD-Ⅱ with different feature selections on three datasets

观察图3得到以下结论：(1)随着在不同数据集上特征数量增加，MNMHD的收敛速度比MNMHD-Ⅱ更快，这是因为MOEA_D算法使用少量的样本能够产生少量但均匀分布的解集，在包含较少特征的数据集中，MNMHD获得属性约简的质量更高。MNMHD在DD数据集上的分类精度略高于MNMHD-Ⅱ，在VOC数据集中分类精度值总体范围为0.32～0.35，数据比例较少的情况下，3种算法精度相差不大，随着数据比例增大，MNMHD的分类精度高于MNMHD-Ⅱ。但在F194数据集上，MNMHD的分类精度要远高于MNMHD-Ⅱ，这是因为F194和VOC数据集的类别层次结构信息增加，导致目标函数计算更复杂。因此，选取MOEA_D算法与邻域粗糙集的层次依赖度相结合更具有合理性。(2)从整体上看，MNMHD分类精度的最小值、最大值都要高于MNHD，这是因为混合层次依赖度中的决策分辨度提供了决策属性的空间结构信息，提升了衡量条件属性重要度的准确性。在包含较少特征的数据集中，相比于MNHD，MNMHD拥有更多的层次结构信息，经计算所获得的属性约简具有更好的分类性能。因此，决策分辨度与层次依赖度的结合有利于提升算法的稳定性。

综上所述，在LSVM、KNN两个分类器中，MNMHD在DD和F194数据集上的分类精度最高，在数据集VOC上的分类精度与最优值相差不大，在每个数据集上的分类性能都较稳定。这是因为MNMHD在处理含有复杂层次结构信息的数据集时，能够充分利用层次结构信息，从而更准确地计算出条件属性重要度，进而获得更好的分类性能。相比其他算法，MNMHD在分类精度、最大值、最小值以及收敛速度等方面表现得更好。

4.4 实验结果分析

为了更全面地比较算法在层次结构数据的实际性能，在不同数据集上采取固定的特征数量，采用分类精度、TIE和F_{LCA}三种评价指标来衡量算法在层次结构数据中的分类效果。此外，还选取两种面向层次结构的邻域粗糙集特征选择算法(OHS-NDER[4]和OHFS[5])，两种传统的邻域粗糙集特征选择算法(FOSFS[21]和OFSD[22])行对比实验。在5个数据集上的3种指标值对比如表3～8所示，表中黑体字表示最优值。

表3 各算法在KNN分类器上的分类精度

Table 3 Classification accuracy of each algorithm on KNN classifier

数据集	MNMHD	MNHD	MNMHD-Ⅱ	OHS-NDER	OHFS	FOSFS	OFSD
Bridge	0.575 6	0.569 2	0.530 9	0.595 6	**0.630 0**	0.417 8	0.507 8
DD	**0.690 1**	0.663 9	0.662 6	0.429 2	0.624 2	0.405 8	0.558 4
F194	**0.525 6**	0.502 6	0.440 4	0.447 0	0.469 3	0.289 3	0.221 7
VOC	**0.307 6**	0.233 6	0.239 6	0.234 7	0.284 1	0.202 6	0.223 9
SAIAPR	**0.206 7**	0.194 8	0.177 6	0.172 8	0.198 3	0.104 2	0.171 8
平均值	**0.501 1**	0.432 8	0.410 2	0.375 8	0.441 1	0.283 9	0.336 7

表4 各算法在KNN分类器上的F_{LCA}值

Table 4 F_{LCA} values of each algorithm on KNN classifier

数据集	MNMHD	MNHD	MNMHD-Ⅱ	OHS-NDER	OHFS	FOSFS	OFSD
Bridge	0.785 6	0.789 3	0.778 4	**0.798 1**	0.788 9	0.674 1	0.715 7
DD	**0.826 2**	0.825 6	0.808 0	0.721 6	0.792 4	0.659 8	0.749 9
F194	**0.731 2**	0.717 9	0.681 0	0.615 5	0.698 1	0.600 5	0.563 0
VOC	**0.720 8**	0.504 8	0.508 6	0.522 2	0.537 6	0.479 3	0.497 4
SAIAPR	**0.487 5**	0.461 8	0.453 7	0.426 6	0.473 9	0.424 5	0.412 3
平均值	**0.710 26**	0.659 88	0.645 94	0.616 8	0.658 18	0.567 64	0.587 66

表5 各算法在KNN分类器上的TIE值

Table 5 TIE values of each algorithm on KNN classifier

数据集	MNMHD	MNHD	MNMHD-Ⅱ	OHS-NDER	OHFS	FOSFS	OFSD
Bridge	1.070 2	1.151 9	1.241 4	1.027 8	**0.981 5**	1.481 5	1.314 8
DD	**0.864 1**	0.937 3	0.954 2	1.670 1	0.988 1	1.705 4	1.234 2
F194	**1.328 6**	1.395 1	1.589 6	2.307 1	1.497 7	1.951 0	2.130 9
VOC	**1.380 7**	2.955 5	2.929 9	3.098 0	2.735 4	3.154 9	2.995 8
SAIAPR	**1.904 8**	2.428 7	2.704 1	2.782 5	2.425 8	3.565 2	3.418 6
平均值	**1.160 9**	1.610 0	1.678 8	2.025 75	1.550 675	2.073 2	1.918 9

表6 各算法在LSVM分类器上的分类精度

Table 6 Classification accuracy of each algorithm on LSVM classifier

数据集	MNMHD	MNHD	MNMHD-Ⅱ	OHS-NDER	OHFS	FOSFS	OFSD
Bridge	0.575 6	0.588 7	0.538 3	**0.632 2**	0.600 0	0.630 0	0.564 4
DD	**0.782 2**	0.775 8	0.735 8	0.422 5	0.711 8	0.370 7	0.307 9
F194	**0.541 1**	0.530 5	0.505 3	0.414 7	0.496 0	0.253 7	0.101 0
VOC	0.341 1	0.341 6	0.340 3	0.257 6	**0.360 8**	0.267 1	0.289 8
SAIAPR	**0.212 8**	0.201 0	0.195 8	0.196 0	0.177 6	0.158 4	0.143 2
平均值	**0.490 6**	0.487 5	0.463 1	0.384 6	0.469 2	0.336 0	0.281 3

表7 各算法在LSVM分类器上的F_{LCA}值

Table 7 F_{LCA} values of each algorithm on LSVM classifier

数据集	MNMHD	MNHD	MNMHD-Ⅱ	OHS-NDER	OHFS	FOSFS	OFSD
Bridge	0.783 8	0.770 5	0.742 7	**0.797 2**	0.770 4	0.785 2	0.752 8
DD	**0.854 5**	0.853 1	0.819 4	0.659 2	0.839 6	0.633 9	0.574 3
F194	**0.735 2**	0.723 7	0.718 9	0.544 6	0.717 1	0.580 9	0.452 6
VOC	0.510 6	0.506 8	0.506 3	0.517 6	**0.589 8**	0.523 6	0.541 1
SAIAPR	**0.473 2**	0.469 0	0.451 8	0.470 9	0.421 8	0.418 3	0.417 6
平均值	**0.671 5**	0.664 6	0.647 8	0.597 9	0.667 7	0.588 4	0.547 7

表8 各算法在LSVM分类器上的TIE值

Table 8 TIE values of each algorithm on LSVM classifier

数据集	MNMHD	MNHD	MNMHD-Ⅱ	OHS-NDER	OHFS	FOSFS	OFSD
Bridge	1.088 6	1.061 5	1.188 2	**0.944 4**	1.074 1	1.009 3	1.138 9
DD	**1.674 5**	1.819 4	1.910 4	1.683 8	0.779 1	1.875 9	2.339 9
F194	**1.274 1**	1.505 9	1.843 0	2.252 0	1.375 2	2.041 8	2.971 3
VOC	2.437 3	2.873 9	2.874 8	2.819 2	**2.390 6**	2.787 7	2.680 0
SAIAPR	**2.160 9**	2.550 6	2.678 8	2.609 9	2.918 9	3.025 8	3.073 2
平均值	1.727 1	1.962 3	2.099 0	2.061 9	**1.707 6**	2.148 1	2.440 7

由表3~8可知,MNMHD在5个数据集上的平均性能排名第一。在3个关键评价指标上,超过一半以上的数据集展现出了MNMHD的最优性能。在其他数据集上,其性能也全部达到了次优水平。传统的邻域粗糙集特征选择方法忽略了类别层次结构的信息,导致FOSFS和OFSD在各个评价指标上的表现不佳。而OHS-NDER和OHFS虽然考虑到了类别层次信息,在数据集Bridge上表现较好,但在复杂数据集DD、F194和VOC上的表现较差。这是因为OHS-NDER和OHFS计算不同层次的属性依赖度并进行累加,未考虑不同层次结构之间的空间信息,导致条件属性重要度的准确度较低。

相比之下,MNMHD既考虑了类别层次信息又考虑了空间信息,因此能够选择更有区分性的特征。在分类器KNN中,MNMHD表现最为出色,其分类性能指标在5个数据集上均排名第一,充分证明了MNMHD的有效性。此外,在LSVM分类器中,MNMHD在DD、F194和SAIAPR数据集上的分类性能也表现出色,各项指标均达到最优。在其他数据集上,其分类性能与最优值之间的差距微小,表现出较高的稳定性。综上所述,MNMHD算法在指标评价上具有优越的性能,进一步证明了MNMHD算法的有效性。

5 结束语

为了满足邻域决策系统的高精度分类需求,本文提出了一种混合层次依赖度下的邻域粗糙集多目标特征选择算法MNMHD,利用类别层次结构信息,从决策类、决策层和决策分类3个关键层面来衡量属性依赖度,构造了混合层次属性依赖度来计算条件属性重要度。与现有的分层特征选择算法相比,所提算法更充分利用类别间层次结构所提供的信息,提供兼顾分类性能和属性约简规模的特征子集,新模型更具有处理包含复杂性层次结构的数据集的能力。在5个数据集上的实验结果验证了所提算法的有效性,下一步将尝试设计混合层次依赖度下的增量式邻域粗糙集多目标特征选择算法。

参考文献:

[1] 胡清华,王煜,周玉灿,等. 大规模分类任务的分层学习方法综述[J]. 中国科学:信息科学, 2018, 48(5): 487-500.
HU Qinghua, WANG Yu, ZHOU Yucan, et al. Review on hierarchical learning methods for large-scale classification task[J]. Scientia Sinica Informationis, 2018, 48(5): 487-500.

[2] ZHAO H, HU Q, ZHU P, et al. A recursive regularization based feature selection framework for hierarchical classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(7): 2833-2846.

[3] WANG Chuang, XIAO Jianmei. Study on power system transient stability assessment based on rough intensive reduction method[J]. International Core Journal of Engineering, 2022, 8(5): 135-145.

[4] 曾艺祥,林耀进,范凯钧,等. 基于层次类别邻域粗糙集的在线流特征选择算法[J]. 南京大学学报(自然科学), 2022, 58(3): 506-518.
ZENG Yixiang, LIN Yaojin, FAN Kaijun, et al. Online streaming feature selection method based on hierarchical class neighborhood rough set[J]. Journal of Nanjing University (Natural Science), 2022, 58(3): 506-518.

[5] 白盛兴,林耀进,王晨曦,等. 基于邻域粗糙集的大规模层次分类在线流特征选择[J]. 模式识别与人工智能, 2019, 32(9): 811-820.
BAI Shengxing, LIN Yaojin, WANG Chenxi, et al. Large-scale hierarchical classification online streaming feature selection based on neighborhood rough set[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(9): 811-820.

[6] ZHAO H, WANG P, HU Q, et al. Fuzzy rough set based feature selection for large-scale hierarchical classification[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(10): 1891-1903.

[7] 方波,陈红梅,王生武. 基于粗糙集和果蝇优化算法的特征选择方法[J]. 计算机科学, 2019, 46(7): 157-164.
FANG Bo, CHEN Hongmei, WANG Shengwu. Feature selection algorithm based on rough sets and fruit fly optimization[J]. Computer Science, 2019, 46(7): 157-164.

[8] QIU Z, ZHAO H. A fuzzy rough set approach to hierarchical feature selection based on Hausdorff distance[J]. Applied

Intelligence, 2022, 52(10): 11089-11102.
[9] LI J, LIU X, WANG X L. Financial decision knowledge acquisition based on neighborhood rough set and ensemble classifiers with grid search[J]. Data Analysis and Knowledge Discovery, 2019, 3(1): 85-94.
[10] 杨洁,匡俊成,王国胤,等. 代价敏感的多粒度邻域粗糙模糊集的近似表示[J]. 计算机科学, 2023, 50(5): 137-145.
YANG Jie, KUANG Juncheng, WANG Guoyin, et al. Cost-sensitive multigranulation approximation of neighborhood rough fuzzy sets[J]. Computer Science, 2023, 50(5): 137-145.
[11] 陈于思,艾志华,张清华. 基于三角不等式判定和局部策略的高效邻域覆盖模型[J]. 计算机科学, 2022, 49(5): 152-158.
CHEN Yusi, AI Zhihua, ZHANG Qinghua. Efficient neighborhood covering model based on triangle inequality check and local strategy[J]. Computer Science, 2022, 49(5): 152-158.
[12] 王茜,张贤勇,吕智颖. 不完备决策信息系统的混合条件熵与多属性决策[J]. 系统工程理论与实践, 2022, 42(12): 3401-3411.
WANG Qian, ZHANG Xianyong, LYU Zhiying. Hybrid conditional entropy and multi-attribute decision making of incomplete decision information system[J]. System Engineering Theory &Practice, 2022, 42(12): 3401-3411.
[13] 林耀进,白盛兴,赵红,等. 基于标签关联性的分层分类共有与固有特征选择[J]. 软件学报, 2022, 33(7): 2667-2682.
LIN Yaojin, BAI Shengxin, ZHAO Hong, et al. Label-correlation-based common and specific feature selection for hierarchical classification[J]. Journal of Software, 2022, 33(7): 2667-2682.
[14] ZHITAO W, HAO L, JIAN Z, et al. An improved MOEA/D algorithm for the solution of the multi-objective optimal power flow problem[J]. Processes, 2023, 11(2): 337-337.
[15] SANTANA-QUINTERO L V, HERNANDEZ-DIAZ A G, MOLINA J, et al. DEMORS: A hybrid multi-objective optimization algorithm using differential evolution and rough set theory for constrained problems[J]. Computers & Operations Research, 2010, 37(3): 470-480.
[16] LAMPERT C H, NICKISCH H, HARMELING S. Learning to detect unseen object classes by between-class attribute transfer[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2009: 951-958.
[17] WEI L Y, LIAO M H, GAO X, et al. An improved protein structural classes prediction method by incorporating both sequence and structure information[J]. IEEE Transactions on NanoBioscience, 2015, 14(4): 339-349.
[18] LI D, JU Y, ZOU Q. Protein folds prediction with hierarchical structured SVM[J]. Current Proteomics, 2016, 13(2): 79-85.
[19] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.
[20] 余鹰,张志强,钱进,等. 基于标记补充的多标记特征选择算法[J]. 数据采集与处理, 2023, 38(3): 539-548.
YU Ying, ZHANG Zhiqiang, QIAN Jin, et al. Multilabel feature selection algorithm based on label supplement[J]. Journal of Data Acquisition and Processing, 2023, 38(3): 539-548.
[21] ZHOU P, HU X G, LI P P. A new online feature selection method using neighborhood rough set[C]//Proceedings of 2017 IEEE International Conference on Big Knowledge. Hefei, China: IEEE, 2017: 135-142.
[22] WU X, YU K, DING W, et al. Online feature selection with streaming features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(5): 1178-1192.

作者简介:

骆公志(1972-),男,博士,教授,研究方向:粗糙集理论及应用,E-mail: lgzlyg@163.com。

张尚蕾(1999-),**通信作者**,女,硕士研究生,研究方向:粗糙集理论及应用,E-mail:13951617689@163.com

(编辑:张黄群)