第 62 卷　第 1 期
2026 年 1 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 1
Jan., 2026

DOI:10.13232/j.cnki.jnju.2026.01.013

基于标记模糊化的层次特征选择

龚匡丰[1,2*],李国和[1,3],郭凌云[4],林耀进[5]

(1.中国石油大学(北京)人工智能学院,北京,102249;2.龙岩学院数学与信息工程学院,龙岩,364000;
3.新疆油气智能勘探与开发重点实验室,中国石油大学(北京)克拉玛依校区,克拉玛依,834000;
4.河南师范大学软件学院,新乡,453007;5.闽南师范大学计算机学院,漳州,363000)

摘　要:分层分类任务通常面临高维特征空间、复杂的类别层次结构以及标记稀疏等多重挑战,其中,标记稀疏性导致监督信息不足,进而削弱特征选择的效果.针对该问题,提出一种新的层次特征选择方法——基于标记模糊化的层次特征选择方法(Hierarchical Feature Selection Based on Label Fuzzification,HFSLF),其核心思想是通过增强稀疏标记的语义表达能力来改善监督信息.具体地,HFSLF首先利用兄弟关系构建类别间的模糊相似性,并利用该相似性将样本的原始标记转化为标记分布,这一转化过程有效扩展了监督信息的覆盖范围,增强了稀疏场景下的语义监督表达.进一步,所提算法以特征与标记分布之间的互信息为监督信号,引导特征权重逼近其对应的互信息值,从而增强模型对高相关性特征的选择偏好.在六个层次数据集上的实验证明了所提算法的有效性.

关键词:特征选择,分层分类学习,标记模糊化,互信息

中图分类号:TP181　　**文献标志码**:A

Hierarchical feature selection based on label fuzzification

Gong Kuangfeng[1,2*],Li Guohe[1,3],Guo Lingyun[4],Lin Yaojin[5]

(1. College of Artificial Intelligence,China University of Petroleum－Beijing,Beijing,102249,China;
2. School of Mathematics and Information Engineering,Longyan University,Longyan,364000,China;
3. Xinjiang Key Laboratory of Intelligent Petroleum Exploration and Engineering,
China University of Petroleum－Beijing at Karamay,Karamay,834000,China;
4. College of Software,Henan Normal University,Xinxiang,453007,China;
5. School of Computer Science,Minnan Normal University,Zhangzhou,363000,China)

Abstract: Hierarchical classification tasks typically face multiple challenges, such as high-dimensional feature space, a complex label hierarchy, and label sparsity. Among these, label sparsity can lead to insufficient supervision, thereby degrading the effectiveness of feature selection. To address this issue, this paper proposes a novel hierarchical feature selection method: Hierarchical Feature Selection Based on Label Fuzzification (HFSLF). The core idea of this method is to improve supervision by enhancing the semantic expressiveness of sparse labels. Specifically, HFSLF first uses sibling relationships to construct fuzzy similarities among categories and transforms the original sample labels into label distributions. This transformation effectively expands the coverage of supervisory information and strengthens semantic supervision in sparse scenarios. Then, the proposed algorithm employs the mutual information between features and label distributions as a supervisory signal, guiding the feature weights to approximate their corresponding mutual information values, thereby enhancing the model's

基金项目:国家自然科学基金(62576158),中国石油大学(北京)克拉玛依校区科研基金(RCYJ2016B-03-001,XQZX20240032),克拉玛依科技计划(2020CGZH0009)

收稿日期:2025－10－30

*通信联系人,E-mail:fjgongkf@126.com

preference for highly relevant features. Experiments on six hierarchical datasets demonstrate the effectiveness of the proposed algorithm.

Keywords: feature selection, hierarchical classification learning, label fuzzification, mutual information

在大数据时代,面对日益增长的数据,传统的扁平化建模方式正面临严峻挑战:一方面,样本的特征空间通常具有高维性,并可能伴随显著稀疏性,这对模型的表达能力和泛化能力提出了更高的要求;另一方面,样本的标记空间往往蕴含丰富的语义结构,类别之间通过父子、兄弟等关系形成层次化体系. 这种结构化的先验知识不仅反映了人类对知识的自然组织方式,也为处理大规模、细粒度分类任务提供了关键线索. 从 18 世纪卡尔·冯·林奈建立的生物分类体系[1]到当代 ImageNet 的语义层级结构[2]和层次化文本分类[3],层次化建模的思想一脉相承,广泛应用于各类复杂分类场景. 研究表明,当样本规模庞大且类别间存在较强语义关联时,分层分类建模已成为提升模型泛化能力和预测一致性的主流范式之一. 因此,有效利用标记空间的层次结构进行建模已成为当前研究的热点.

在分层分类任务建模过程中,随着任务涉及的类别规模从最初的二类分类到网页数据的万类级别[4],学习模型正面临日益严峻的多重挑战. 高维特征导致存储压力与计算开销显著增加,类别数量的急剧增长进一步加剧了模型的泛化难度. 为了缓解高维性问题,特征选择作为一种关键的数据降维技术受到广泛关注. 然而,传统特征选择方法大多基于扁平化分类框架设计,忽略标记空间存在的层次关系,导致特征选择过程难以充分利用标记空间中的先验知识,限制了对样本信息的充分挖掘. 因此,面对超多类别且类别间存在层次化语义关联时,选出最具判别力的特征子集已成为当前亟待解决的关键问题.

粒计算[5−6]是模拟人类多层次认知机制的计算范式,其理论框架强调通过构建多层次的信息结构,实现对复杂数据的结构化分析与建模. 模糊粗糙集[7]作为粒计算的重要理论工具,能够基于模糊相似关系对数据进行粒化,形成不同粒度层次的信息表示,有效处理数据中的模糊性与不确定性. Wang et al[8]利用模糊邻域的概念定义了样本的模糊决策,将原始标记转化为模糊标记,通过引入参数化模糊关系对模糊信息粒进行刻画,在重构模糊决策下近似和上近似的基础上提出一种模糊粗糙模型. Wang et al[9]提出一种有向模糊粗糙集模型,将类别子空间的分布信息融合到有向模糊二元关系中,由此开发了一种启发式特征选择算法. Deng et al[10]通过标记分布的相关性定义样本间的模糊等价关系并评估被划分为同一类别的概率,构建了一种新的邻域模糊粗糙集. 然而,上述算法主要面向扁平标记空间,没有充分考虑标记空间存在层次关系.

在分层分类任务中,利用模糊粗糙集模型构建与类别层次相协调的信息粒,有助于挖掘数据中潜在的信息[11],目前已有一些基于模糊粗糙集理论构建的层次特征选择算法被提出. Zhao et al[12]将层次结构嵌入模糊粗糙集,采用包含策略和兄弟策略为层次分类重新定义了下近似和上近似,由此设计了一种基于模糊粗糙集的层次分类特征选择算法. Qiu and Zhao[13]根据标记语义的层次结构,将特征选择任务分解为粗粒度和细粒度任务,采用 Hausdorff 距离的模糊粗糙集方法,给出一种基于粒计算的层次特征选择方法. Bai et al[14]通过在线重要性选择和在线冗余分析,构建了一种基于核模糊粗糙集的层次流特征选择框架. 已有方法提升了特征选择的性能,但仍然存在对类别语义层次利用不充分等问题. 具体地,样本与邻近类别(如兄弟类)的潜在语义关联常被忽略,导致信息粒划分过于刚性,造成样本类别的监督信息不够充分,因此,更细腻地刻画样本与层次化标记之间的语义关联,成为提升模型泛化能力的关键.

为此,本文引入标记模糊化机制,将样本原有的标记扩展为对多个相关类别的模糊隶属度表示. 通过模糊相似关系和兄弟策略,计算研究样本与兄弟类别样本间的相似关系,进一步为每个

样本分配其在不同层次上的局部隶属度，构建更细腻的标记空间．这种通过模糊化生成的标记分布能捕捉样本与语义相似类别之间的潜在关联，为后续特征选择模型的构建提供更丰富的语义支持．

综上，本文提出基于标记模糊化的层次特征选择算法（Hierarchical Feature Selection Based on Label Fuzzification，HFSLF）．首先，利用兄弟策略重新构建模糊相似关系将样本的原始标记转化为标记分布；随后，在目标函数中嵌入特征与标记分布之间的互信息作为正则化项，以增强所选特征与层次化语义结构之间的关联性．最后，通过实验验证了本文所提算法的有效性．

1　准备知识

1.1　类别的层次结构　层次结构主要有两种类型：树结构和有向无环图结构[15]．本文主要关注树结构关系，树结构的“从属”关系存在三个特性，即不可逆性、反自反性和传递性[16]．用$(L, \prec)$表达层次结构，其中，L为标记集合，“$\prec$”表示从属关系，则上述特性可形式化地表达为：

(1)不可逆性：若$l_i \prec l_j, \forall l_i, l_j \in L$，则$l_j \not\prec l_i$；

(2)反自反性：$\forall l_i \in L$，则$l_i \not\prec l_i$；

(3)传递性：若$l_i \prec l_j$且$l_j \prec l_k$，对$\forall l_i, l_j, l_k \in L$，则$l_i \prec l_k$.

1.2　分层分类的类别关系　在分层分类任务中，可根据不同的策略对目标样本的同类和异类进行刻画．如表1第1行所示，假定目标样本的类别为L_p，则根据排斥策略[17]得到异类为非L_p．在分层分类中，利用类别之间的父子关系和兄弟关系可得到包含策略和兄弟策略[17−18]下的同类样本和异类样本（见表1第2行和第3行）．

例1　图1展示了数据集VOC[19]类别层次结构的局部信息．以Farmyard类为例，在排斥策略下，Farmyard类的同类仅包含其自身，其余所有类别均为其异类．在包含策略下，Farmyard类的同类包括其自身及其子类Cow，Horse和Sheep，其余类别为其异类．在兄弟策略下，Farmyard类的同类为其自身，而其异类仅包含其兄弟类别Domestic.

表1　三种策略下的同类样本和异类样本

Table 1　Positive and negative samples under three strategies

策略	同类	异类
排斥策略	L_p	Not L_p
包含策略	$L_p + des(L_p)$	$\mathrm{Not}(L_p + des(L_p))$
兄弟策略	L_p	$\mathrm{sib}(L_p)$

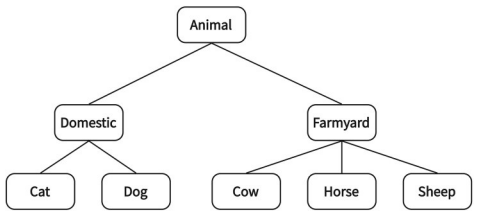

图1　VOC数据集的类别层次结构

Fig. 1　Category hierarchy of the VOC dataset

1.3　基于稀疏学习的层次特征选择框架　首先，设样本矩阵为$X \in R^{n \times m}$，其中，n表示样本数，m表示特征数．定义层次结构中的非叶子结点的个数（也称内部结点）为$N+1$，则样本矩阵可划分为$X_0, X_1, \cdots, X_N$，其中，$X_i = [x_i^1, x_i^2, \cdots, x_i^{n_i}] \in R^{n_i \times m}$表示第$i$个内部结点的样本矩阵．其次，定义$Y_0, Y_1, \cdots, Y_N$为内部结点的标记矩阵：

$$Y_i = \left[y_i^1, y_i^2, \cdots, y_i^{n_i}\right] \in R^{n_i \times d_{\max}}$$

$$y_k = \{0,1\}^{d_{\max}}, 1 \leqslant k \leqslant n_i$$

其中，$d_{\max}$代表内部结点标记数量的最大值．再者，定义$W_i = \left[w_i^1; w_i^2; \cdots; w_i^m\right] \in R^{m \times d_{\max}}$为每个内部结点的权重矩阵．根据已有经验，稀疏学习被证明是一种有效的特征选择方法[20]，通常可表达为如下形式[21]：

$$\min_W L(W; X, Y) + \lambda \Gamma(W) \tag{1}$$

其中，$L(\cdot)$表示损失函数，通常是最小二乘损失、铰链损失等．本文采用最小二乘损失作为损失函数，损失函数可定义为：

$$L(W; X, Y) = \| XW - Y \|_F^2 \tag{2}$$

对于稀疏正则化项$\Gamma(W)$，由于$l_{2,1}$范数的正则化是凸的，且容易根据Argyriou et al[22]的方法进行优化，因此本文采用该范数来构建模型．结合式(2)和$l_{2,1}$范数得到基于稀疏学习的层次特征选择的基本框架为：

$$J = \min_W \left(\sum_{i=0}^{N} \left(\| X_i W_i - Y_i \|_F^2 + \lambda \| W_i \|_{2,1} \right) \right) \tag{3}$$

2 基于标记模糊化的层次特征选择

首先介绍单标记环境下的标记模糊化模型,然后将其推广至层次化结构的数据场景,在此基础上,将该模型嵌入到所提算法框架中.

2.1 模糊标记 称 $FDIS=\langle U,A,f,D,q\rangle$ 为模糊决策信息系统,其中,U 为非空论域,A 为非空条件属性集,D 为决策属性集. $f:U\times A\to\bigcup_{i=1}^{m}V_i$,其中,$V_i$ 表示任意属性 $a_i\in A$ 的值域. $q:U\times D\to\bigcup_{i=1}^{m}V_i^d$,其中,$V_i^d$ 表示决策属性 $d_i\in D$ 的值域[23].

给定论域 $U=\{x_1,x_2,\cdots,x_n\}$,设 $B\subseteq A$ 是一组实值属性的子集,这些属性在 U 上诱导出一个模糊二元关系 R_B. 如果 R_B 满足以下条件,则称其为模糊相似关系[8].

(1)自反性:$R_B(x,x)=1,\forall x\in U$.

(2)对称性:$R_B(x,y)=R_B(y,x),\forall x,y\in U$.

对于任意 $x\in U$,x 的模糊邻域 $[x]_B$ 定义为:

$$[x]_B=R_B(x,y),\forall y\in U$$

给定论域 U,D 是论域上的一个决策属性,并将 U 划分为 k 个清晰的等价类 $U/D=\{D_1,D_2,\cdots,D_k\}$,引入样本的模糊划分和模糊决策的概念.

定义 1[8] 给定论域 U,$\{\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_k\}$ 是 U 上的一组模糊集,如果满足:

$$\sum_{i=1}^{k}\tilde{D}_i=1,\forall x\in U \tag{4}$$

则称 $\{\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_k\}$ 为一个模糊划分.

定义 2[8] 给定一个决策属性 D,且 $U/D=\{D_1,D_2,\cdots,D_k\}$,$R_B$ 是由属性集 B 在 U 上诱导出来的模糊相似关系. 对于任意 $x\in U$,x 的模糊决策,定义为:

$$\tilde{D}_i(x)=\frac{\left|[x]_B\cap D_i\right|}{\left|[x]_B\right|},i=1,2,\cdots,k \tag{5}$$

对任意 $x\in U,B\subseteq A,\delta\in(0,1)$,样本 x 的模糊信息粒 $[x]_B^{\delta}$ 定义为:

$$[x]_B^{\delta}(y)=\begin{cases}0, & R_B(x,y)<\delta\\ R_B(x,y), & R_B(x,y)\geqslant\delta\end{cases} \tag{6}$$

2.2 分层分类任务下的模糊标记 2.1 给出了单标记场景下的模糊邻域粒和模糊决策的定义. 为了适应分层分类任务下的样本粒化,利用兄弟策略重新定义样本的模糊信息粒.

称 $HFDIS=\langle U,A,f_H,\tilde{D},H,q_H\rangle$ 为层次模糊决策信息系统,其中,U 为非空论域,A 为非空条件属性集,$\tilde{D}$ 为模糊决策属性集,H 表示标记空间的层次结构关系. $f_H:U\times A\to[0,1]$,$q_H:U\times\tilde{D}\to[0,1]$ 表示决策属性 $d_i\in\tilde{D}$ 的值域在 $[0,1]$.

为了适应分层分类的应用场景,在定义 1 的基础上,利用兄弟策略可得如下模糊划分的定义.

定义 3 给定 $HFDIS=\langle U,A,f_H,\tilde{D},H,q_H\rangle$,$\{\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_k\}$ 是 U 上的一组模糊集,如果 $\{\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_k\}$ 之间互为兄弟关系,且满足:

$$\sum_{i=1}^{k}\tilde{D}_i=1,\forall x\in U \tag{7}$$

则称 $\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_k$ 为一个分层分类任务下的模糊划分.

陈曦等[24]在多标记模糊信息系统中利用高斯核函数定义了特征子集的模糊关系,本节在此基础上进行扩展,利用兄弟策略重新定义层次模糊信息系统下的模糊关系.

定义 4 给定 $HFDIS$,对任意 $B\subseteq A$,$\forall x,y\in U$ 且 $L(y)\in sib(L(x))$,其中,$L(x)$ 表示样本 x 的类标记. 利用高斯核函数定义属性集 B 的模糊关系 $\tilde{R}_B^H(x,y)$:

$$\tilde{R}_B^H(x,y)=\exp\left\{-\frac{1}{2\sigma^2}\sum_{b\in B}\left|f(x,b)-f(y,b)\right|\right\} \tag{8}$$

其中,参数 σ 表示高斯核宽度.

由此,得到任意样本 $x\in U$ 关于属性子集 $B\subseteq A$ 的模糊信息粒 $[x]_B^{sib}(y)$ 为:

$$[x]_B^{sib}(y)=\begin{cases}0, & L(y)\notin sib(L(x))\\ R_B(x,y), & L(y)\in sib(L(x))\end{cases} \tag{9}$$

基于定义 4,进一步提出样本的模糊决策概念,其定义如下.

定义 5 给定一个决策属性 D,且 $U/D=\{D_1,D_2,\cdots,D_k\}$,$\tilde{R}_B^H(x,y)$ 是由属性集 B 在 U 上诱导出来的模糊相似关系. 对于任意 $x\in U$,x 的基

于兄弟关系的模糊决策定义为：

$$\tilde{D}_i^{sib}(x)=\frac{\left|[x]_B^{sib}\cap D_i\right|}{\left|[x]_B^{sib}\right|},i=1,2,\cdots,k \tag{10}$$

根据定义5，可将样本的原始标记转化为标记分布．相较于原始标记，式(10)得到的标记分布充分融合了类别间的兄弟关系，增强了稀疏标记下的监督信息表达．基于以上分析，设计了一种面向层次分类任务的标记模糊化算法，具体步骤如下．

算法1　面向分层分类任务的标记模糊化算法(Label Fuzzification for Hierarchical Classification Tasks)

输入：训练样本特征矩阵 $X_i\in R^{n_i\times m}$，训练样本标记矩阵 $Y_i\in R^{n_i\times d_{\max}}$，参数 σ

输出：样本的标记分布

初始化：$\tilde{D}_i(x)=Y_i$

1. for $i=0:N$ do
2.　for $j=0:n_i$ do
3.　　for $q=0:n_i$ do
4.　　通过式(8)计算 $\tilde{R}_B^H$
5.　　end
6.　end
7.　for $k=0:n_i$ do
8.　　for $q=0:d_i$ do
9.　　通过式(10)计算样本的标记分布
10.　　end
11.　end
12.　更新 $\tilde{D}_i^{sib}=\left[d_0;d_1;\cdots;d_{n_i}\right]$
13.　更新 $\tilde{D}^{sib}=\left[\tilde{D}_1^{sib},\tilde{D}_2^{sib},\cdots,\tilde{D}_N^{sib}\right]$
14. end
15. return $\tilde{D}^{sib}$

在获得标记分布矩阵 $\tilde{D}^{sib}$ 后，将分层特征选择的基本模型重新定义为：

$$J=\min\sum_{i=0}^{N}\left(\left\|X_iW_i-\tilde{D}_i^{sib}\right\|_F^2+\lambda\left\|W_i\right\|_{2,1}\right) \tag{11}$$

2.3　融合互信息的层次特征选择框架　互信息广泛用于度量变量之间的统计依赖关系，因此，为探索特征与标记分布之间的关系，本文引入互信息作为正则项．采用Kraskov et al[25]的方法计算特征与标记之间的互信息如下所示：

$$I\left(X_{\cdot p},\tilde{D}_{\cdot l}^{sib}\right)=\psi(z)+\psi(n)-\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(n_{X_{ip}}\right)+\psi\left(n_{\tilde{D}_{il}^{sib}}\right)\right) \tag{12}$$

其中，$I\left(X_{\cdot p},\tilde{D}_{\cdot l}^{sib}\right)$ 表示特征向量 $X_{\cdot p}$ 与标记分布向量 $\tilde{D}_{\cdot l}^{sib}$ 之间的互信息．

将 $\sqrt{\left(X_{ip}-X_{jp}\right)^2+\left(\tilde{D}_{il}^{sib}-\tilde{D}_{jl}^{sib}\right)^2}$ 作为样本 X_i 与 X_j 之间的距离，设 τ_i 表示 X_i 与其 z 近邻之间的距离．根据文献建议，取 $z=3$，则式(12)中的 $n_{X_{ip}}$ 和 $n_{\tilde{D}_{il}^{sib}}$ 分别满足 $\left|X_{ip}-X_{jp}\right|<\tau_i$ 和 $\left|\tilde{D}_{il}^{sib}-\tilde{D}_{jl}^{sib}\right|<\tau_i$ 的样本数量．$\psi(\cdot)$ 为 *digamma* 函数，满足：

$$\psi(u+1)=\begin{cases}\psi(u)+\dfrac{1}{u}, & u>0\\ -C, & u=0\end{cases} \tag{13}$$

其中，$C=0.58$ 表示欧拉常数．

通过式(12)计算特征与标记分布之间的互信息，由此构造矩阵 $M\in R^{m\times d'}$，其中，d' 表示当前内部节点所对应的标记数量，其元素 $M_{pl}=I\left(X_{\cdot p},\tilde{D}_{\cdot l}^{sib}\right)$．由此构造如下正则项以约束特征权重：

$$\left\|W-M\right\|_F^2 \tag{14}$$

结合式(11)和(14)，可得最终的目标函数：

$$J=\min\sum_{i=1}^{N}\left(\left\|X_iW_i-\tilde{D}_i^{sib}\right\|_F^2+\lambda\left\|W_i\right\|_{2,1}+\gamma\left\|W_i-M_i\right\|_F^2\right) \tag{15}$$

其中，λ 和 γ 为平衡因子．

2.4　模型优化与算法伪代码　式(15)中，由于 $l_{2,1}$ 的非光滑性，根据Argyriou et al[22]进行推导：

$$\frac{\partial\left\|W\right\|_{2,1}}{\partial W}=\frac{\partial\mathrm{Tr}\left(W^{\mathrm{T}}AW\right)}{\partial W}=2AW \tag{16}$$

其中，$A\in R^{d_{\max}\times d_{\max}}$ 为对角矩阵，其第 j 个对角元素是 $a_{jj}^i=\dfrac{1}{2\left\|W_i^j\right\|_2}$，如果 $W_i^j=0$，设 $a_{jj}^i=\epsilon$.

依据式(16)，将目标函数重新表达为：

$$J=\min\sum_{i=0}^{N}\left(\left\|X_iW_i-\tilde{D}_i^{sib}\right\|_F^2+\lambda\mathrm{Tr}\left(W^{\mathrm{T}}AW\right)+\gamma\left\|W_i-M_i\right\|_F^2\right) \tag{17}$$

对于各个内部结点，将式(17)关于 W_i 的导数设置为0，得到：

$$\frac{\partial J}{\partial W}=$$
$$2X_i^{\mathrm{T}}\left(X_iW_i-\tilde{D}_i^{sib}\right)+2\lambda A_iW_i+2\gamma(W_i-M_i)=0 \tag{18}$$

由此可得:

$$W_i=\left(X_i^{\mathrm{T}}X_i+\lambda A_i+\gamma I_i\right)^{-1}\left(X_i^{\mathrm{T}}\tilde{D}_i^{sib}+\gamma M_i\right) \tag{19}$$

根据式(18)和式(19),给出所提算法的伪代码,如算法2所示.通过算法2可到特征的权重矩阵 W,对权重矩阵进行排序之后,选取权重值较大的特征即可完成特征选择任务.

算法2 基于标记模糊化的层次特征选择(Hierarchical Feature Selection Based on Label Fuzzification)

输入:训练样本特征矩阵 $X_i\in R^{n_i\times m}$,参数 λ,γ,迭代次数 T

输出:特征权重矩阵集合 $W\in R^{m\times d_{\max}}$

初始化:$\tilde{D}_i(x)=Y_i$

1. 初始化 $d_{\max}$ 为内部节点的最大类别数,$t=0$
2. 随机初始化:$W^{(0)}=W_0,W_1,\cdots,W_N\in R^{m\times d_{\max}}$
3. 计算训练样本的标记分布矩阵 $\tilde{D}_i^{sib}\in R^{n_i\times d_{\max}}$
4. 根据式(12)计算互信息 M
5. while $t<T$ do
6. for $i=0:N$ do
7. 计算 $a_{jj}^i=\dfrac{1}{2\left\|W_i^j\right\|_2}$,求矩阵 $A_i^{(t)}$
8. end for
9. for $i=0:N$ do
10. 通过式(19)更新 W_i
11. end for
12. $W^{(t+1)}=[W_0,W_1,\cdots,W_N]$
13. $t=t+1$
14. end while
15. 返回 W

算法2中包含算法1的标记分布生成过程和互信息的计算过程,因此在分析算法2基于标记模糊化的层次特征选择算法HFSLF的时间复杂度时,需分析算法1和互信息计算过程的时间开销.算法1的时间复杂度为 $O\left(Nn^2(m+d_{\max})\right)$,其中,$N$ 是节点个数,n 是样本总数.互信息的时间复杂度为 $O\left(Nn^2md_{\max}\right)$.对于算法2,其迭代域的时间复杂度主要取决于特征权重的计算和更新,每个内部节点迭代权重矩阵的时间复杂度为 $O\left(m^3+Nm^2d_{\max}+m^2n_i+mn_id_{\max}\right)$,其中,$n_i$ 为第 i 个内部节点的样本数.观察式(19),其中,$X_i^{\mathrm{T}}X_i$ 与 $X_i^{\mathrm{T}}\tilde{D}_i^{sib}(i=1,2,\cdots,N)$ 只要计算一次,时间复杂度表达为 $O\left(m^2n_i+mn_id_{\max}\right)$,可知所有内部节点所需的时间复杂度为 $O\left(m^2n+mnd_{\max}\right)$.算法迭代次数为 T,可知这部分时间复杂度为:

$$O\left(T\left(m^3+Nm^2d_{\max}\right)+m^2n+mnd_{\max}\right)$$

综上,所提算法的时间复杂度为:

$$O\left(T\left(m^3+Nm^2d_{\max}\right)+m^2n+mnd_{\max}+Nn^2(m+d_{\max}+md_{\max})\right)$$

3 实验与分析

对实验结果进行系统分析,包括三个方面:数据集与评价指标、对比算法与参数设置和算法性能分析.其中,性能分析进一步涵盖性能指标、参数敏感性、消融实验与模型收敛等内容.

3.1 数据集与评价指标 为了验证算法的有效性,选取六个具有层次结构的数集进行实验,包括两个蛋白质数据集DD[26]和F194[27]、四个图像数据集AWA[28],CLEF[29],ILSVRC65[30]和VOC[19].表2为数据集的相关描述.

评价指标包括预测精度、树诱导损失(Tree Induced Error,*TIE*)[31]和基于增广集合的分层 $F1$(Hierarchical-$F1$ measure)[32].其中,预测精度的计算方法与传统算法一致,而 *TIE* 和Hierarchical-$F1$ measure是为了评估层次结构中的错分程度而引入的.

令 y 和 $\hat{y}$ 分别代表样本真实标记和预测标记,$Anc(y)$ 和 $Anc(\hat{y})$ 为 y 和 $\hat{y}$ 的祖先结点集合,则 y

表2 层次数据集信息

Table 2 Hierarchical datasets information

序号	数据集	训练集	测试集	特征数	结点数	叶子结点数	层数
1	AWA	6405	3202	252	17	10	3
2	CLEF	8368	939	80	88	63	4
3	DD	3020	605	473	32	27	3
4	F194	7105	1420	473	202	194	3
5	ILSVRC65	12346	11845	4096	65	57	4
6	VOC	7178	5105	1000	30	20	5

和 $\hat{y}$ 的分层分类扩展标记分别表示为 $Y_{\text{aug}}=y\cup Anc(y)$ 和 $\hat{Y}_{\text{aug}}=\hat{y}\cup Anc(\hat{y})$.

TIE 指标通过计算预测标记 $\hat{y}$ 到真实标记 y 在层次结构中结点之间的总边数来反映分错程度：

$$TIE(y,\hat{y})=\left|E_H(y,\hat{y})\right| \tag{20}$$

其中，$E_H(y,\hat{y})$ 表示从 y 到 $\hat{y}$ 结点之间边的集合.

Hierarchical-$F1$ measure 的计算如式(21)所示：

$$F_H=\frac{2\times P_H\times R_H}{P_H+R_H} \tag{21}$$

其中，

$$P_H=\frac{\left|Y_{\text{aug}}\cap\hat{Y}_{\text{aug}}\right|}{\left|\hat{Y}_{\text{aug}}\right|},R_H=\frac{\left|Y_{\text{aug}}\cap\hat{Y}_{\text{aug}}\right|}{\left|Y_{\text{aug}}\right|}$$

TIE 指标取值越小越好，Hierarchical-$F1$ measure指标取值则越大越好.

3.2　对比算法和参数设置　将所提算法与五个分层特征选择算法进行比较.

(1) HRelief[33]：由Relief扩展而来.

(2) HFSNM[34]：根据FSNM修改而来.

(3) HIFSRR[35]：是基于层次结构中的粒度关系，同时考虑父子关系和兄弟关系的优化特征选择算法.

(4) HSDFS[36]：集成迹比目标和结构化稀疏子空间约束，以获取特征子集.

(5) HFSDK[37]：基于粗粒度和细粒度类之间的相似性来约束所选的层间特征，并依赖于有上限的铰链损失来消除数据异常值.

所提算法涉及三个参数，分别是标记模糊化过程中的高斯核宽度参数 σ、优化目标中的平衡因子 λ 和 γ. 其中，σ 在 $[0.5,2]$ 以步长为0.5进行调整，λ 固定为10，γ 的取值在 $\{10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2}\}$ 调整，并记录最优实验结果. 采用自上而下的支持向量机分类器对所提算法的有效性进行评估. 对于蛋白质数据集，采用排名靠前40个特征进行训练；对于图像数据集，采用前20%的特征进行训练. 采用10折交叉验证. 对于对比算法，根据相应文献中的建议搜索参数.

3.3　算法性能分析　从以下方面对实验进行了评估：(1)算法性能分析；(2)参数敏感性分析；(3)消融实验；(4)收敛性分析.

3.3.1　算法性能分析　表3～5分别给出六种算法在六个数据集上的 Acc，F_H 和 *TIE* 三个指标的实验结果，其中，符号“↓”表示指标的取值越小越好，符号“↑”表示指标的取值越大越好，表中黑体字表示对应指标下的最优结果.

从实验结果看，所提算法在所有评价指标上，均有半数的数据集上达到最优性能，体现出其在不同数据特性下的稳定性和适应性. 以数据集DD为例，其特征空间具有高稀疏性，而本文算法在该数据集上的优异表现，验证了将原始标记转化为标记分布后，通过计算标记空间与特征空间互信息的有效性，这主要得益于标记分布能够提供更丰富的监督信息. 以上结果初步说明，所提算法具有较强的鲁棒性和泛化能力.

为了进一步评估算法的统计显著性，采用

表3　不同算法在各数据集上的*Acc*对比(↑)

Table 3　*Acc* of different algorithms on various datasets (↑)

数据集	HRelief	HFSNM	HIFSRR	HSDFS	HFSDK	HFSLF
AWA	0.2174	0.2411	0.2402	0.2336	0.2386	**0.2427**
CLEF	0.5741	0.5751	0.6039	0.5910	0.6230	**0.6284**
DD	0.4381	0.6696	0.6893	0.5125	0.6877	**0.6926**
F194	0.2218	0.2451	**0.3430**	0.2430	0.3303	0.3310
ILSVRC	0.8489	0.8431	0.8492	0.8501	**0.8531**	0.8507
VOC	0.4029	**0.4253**	0.4200	0.4141	**0.4253**	0.4212
Avg. rank	5.8333	3.7500	2.8333	4.5000	2.4167	1.6667

表4　不同算法在各数据集上的F_H对比(↑)

Table 4　F_H of different algorithms on various datasets (↑)

数据集	HRelief	HFSNM	HIFSRR	HSDFS	HFSDK	HFSLF
AWA	0.5625	0.5713	0.5718	0.5690	0.5707	**0.5726**
CLEF	0.7344	0.7396	0.7625	0.7493	**0.7742**	0.7721
DD	0.7747	0.8518	0.8606	0.7736	0.8590	**0.8650**
F194	0.6704	0.6507	**0.7164**	0.6744	0.7075	0.7092
ILSVRC	0.9579	0.9563	0.9586	0.9583	0.9589	**0.9590**
VOC	0.6585	0.6739	0.6740	0.6690	**0.6772**	0.6746
Avg. rank	5.5000	4.6667	2.3333	4.6667	2.3333	1.5000

表 5 不同算法在各数据集上的 *TIE* 对比(↓)

Table 5 *TIE* of different algorithms on various datasets (↓)

数据集	HRelief	HFSNM	HIFSRR	HSDFS	HFSDK	HFSLF
AWA	0.3500	0.3430	0.3425	0.3448	0.3434	**0.3419**
CLEF	0.2037	0.2007	0.1831	0.1931	**0.1745**	0.1761
DD	0.1352	0.0889	0.0836	0.1359	0.0846	**0.0810**
F194	0.1977	0.2096	**0.1701**	0.1954	0.1755	0.1745
ILSVRC	0.0337	0.0350	0.0331	0.0333	0.0329	**0.0328**
VOC	0.2237	0.2144	0.2151	0.2187	**0.2126**	0.2150
Avg. rank	5.5000	4.3333	2.5000	4.6667	2.3333	1.6667

Friedman检验[38]作为算法排序的评估标准. 给定 P 个算法和 Q 个数据集,第 i 个算法的平均排名表示为 r_i,则Friedman检验使用式(22)进行统计:

$$F_F=\frac{(Q-1)\chi_F^2}{Q(P-1)-\chi_F^2} \tag{22}$$

其中,

$$\chi_F^2=\frac{12Q}{P(P+1)}\left(\sum_{i=1}^{P}r_i^{\,2}-\frac{P(P+1)^2}{4}\right)$$

Friedman检验的结果和各项指标对应的阈值见表6. 结果表明,显著性水平 $\alpha=0.05$ 时,每个指标的 F_F 值大于 F 检验临界值,所有算法性能相同的原假设被拒绝.

表 6 各指标的 F_F 及相应的临界值

Table 6 F_F of each indicator and their corresponding critical values

评价指标	F_F	临界值
Acc	9.5497	
F_H	16.5753	2.6030
TIE	10.2913	

进一步,使用Nemenyi测试[39]来检验所提算法与对比算法之间的性能差异. 该方法首先计算两种算法在平均排名上的差异,然后利用临界差异值(*CD*)来评估该差异是否具有统计显著性. 临界值域 *CD* 的值由式(23)给出:

$$CD_\alpha=q_\alpha\sqrt{\frac{P(P+1)}{6Q}} \tag{23}$$

在显著性水平 $\alpha=0.05$ 下,Nemenyi检验得到 $q_\alpha=2.850$,得到 $CD_\alpha=3.0784$.

图2显示了各指标下通过Nemenyi检验来比较各算法性能的检验结果. 由图可见,所提算法的 *Acc* 和 *TIE* 两个指标明显优于HRelief算法;F_H 指标明显优于HRelief,HSDFS和HFSNM算法. 但没有一致的证据表明HFSDK,HIFSRR的评价指标和HFSLF算法之间存在统计学差异.

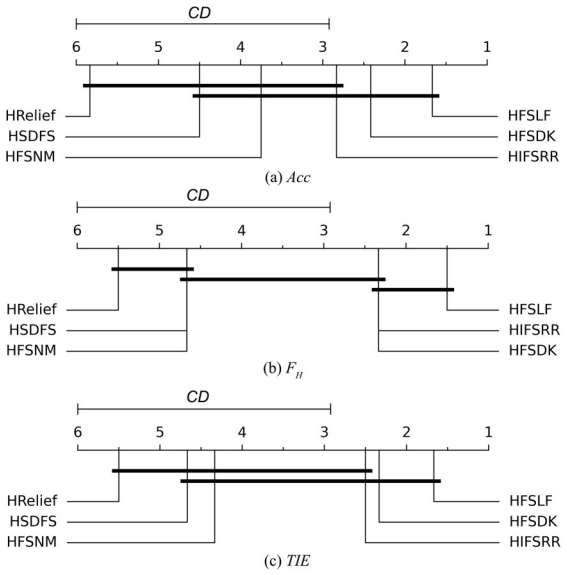

图 2 通过Nemenyi检验比较HFSLF算法与其他算法的性能

Fig.2 Performance of the HFSLF algorithm with other algorithms by the Nemenyi test

3.3.2 参数敏感性分析 HFSLF的三个参数,λ 固定为10,高斯核宽度参数 $\sigma\in[0.5,2]$ 以步长为0.5进行调整,γ 取值于集合 $\{10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2}\}$. 由于 λ 已固定,重点对 σ 和 γ 进行参数敏感性分析. 采用控制变量法,固定其中一个参数,调节另一个参数,观察其Hierarchical-$F1$的变化,以评估算法性能对参数变化的敏感性.

图3和图4展示了数据集CLEF和F194的分析结果. 由图可见,在CLEF上,模型性能在 $\sigma=1$ 时达到峰值;在F194上,最优性能出现在 $\sigma=0.5$. 由于本文采用绝对值距离作为相似性度量,较小的 σ 意味着只有在特征差异较小的情况下才赋予高相似权重,这一结果表明,F194需要更严格的局部近邻性才能实现最佳匹配,反映出其局部结构上的高敏感性.

对于参数 γ,其性能在 $10^{-3}\sim10^{1}$ 波动较小,表现出良好的鲁棒性;当 $\gamma=10^{2}$ 时出现较明显

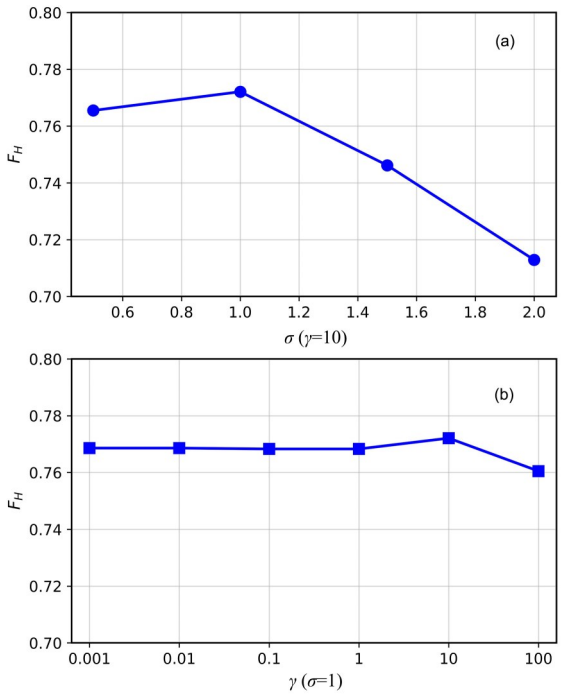

图3　基于CLEF数据集的参数敏感性分析

Fig. 3　Parameter sensitivity analysis based on the CLEF dataset

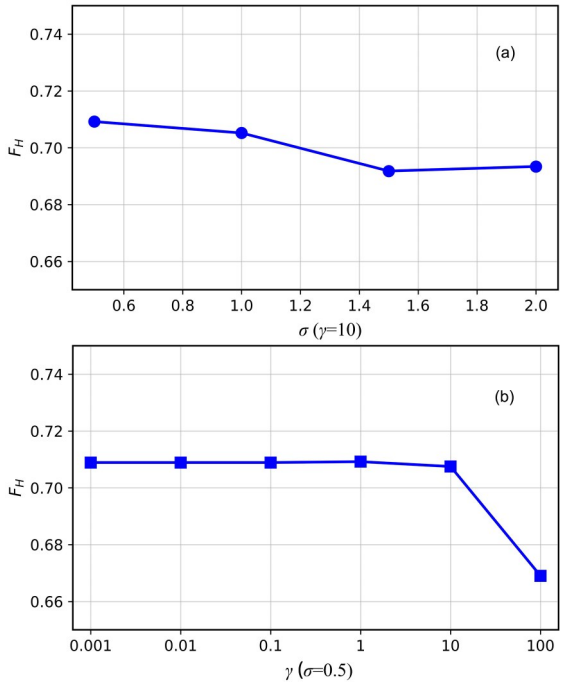

图4　基于F194数据集的参数敏感性分析

Fig.4　Parameter sensitivity analysis based on the F194 dataset

的下降.这可能是互信息正则项权重过大,导致模型的学习能力受限,进而可能引发欠拟合.

3.3.3　消融实验　通过消融实验来分析标记分布和互信息正则化项对HFSLF特征选择的影响.实验设置了以下模型.

(1) HFSLF-Base:仅基于稀疏学习正则化,不包含标记分布和互信息正则化项.

(2) HFSLF-FL:将原始标记转化为标记分布,但不包含互信息正则化项.

(3) HFSLF:完整模型,包含标记分布与互信息正则化项.

图5展示了HFSLF及其两个比较版本在六

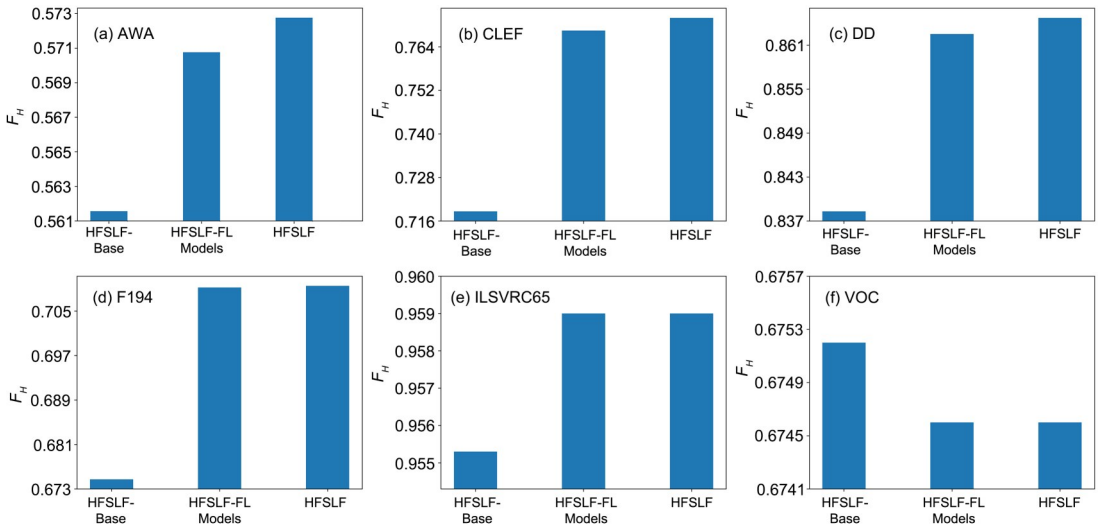

图5　消融实验

Fig.5　Ablation study

个层次数据集上的结果,由图可得以下结论.

(1) 与 HFSLF-Base 相比,HFSLF-FL 在五个数据集上的性能均得到了提升,由此可知,本文提出的标记分布机制对模型性能有积极的影响. 然而,在 VOC 数据集上,应用标记分布后,效果却略有下降,这可能是由于标记分布本身对该数据集引入了噪声或弱化了关键语义信息,干扰了模型的学习效果.

(2) 与 HFSLF-FL 相比,HFSLF 在绝大多数数据集上的表现更优,这表明本文所提的基于互信息的正则化项是有效的.

综上,标记分布与互信息正则项共同构成了 HFSLF 的核心贡献,在多数情况下提升了模型的泛化能力.

3.3.4 收敛性分析 对所提算法 HFSLF 进行收敛性分析,所有数据集基于目标函数基础上的收敛曲线如图 6 所示. 实验中,在所有数据集上设置了最大的迭代次数为 10. 实验表明,所有数据集目标函数单调递减并在不超过 10 次内收敛.

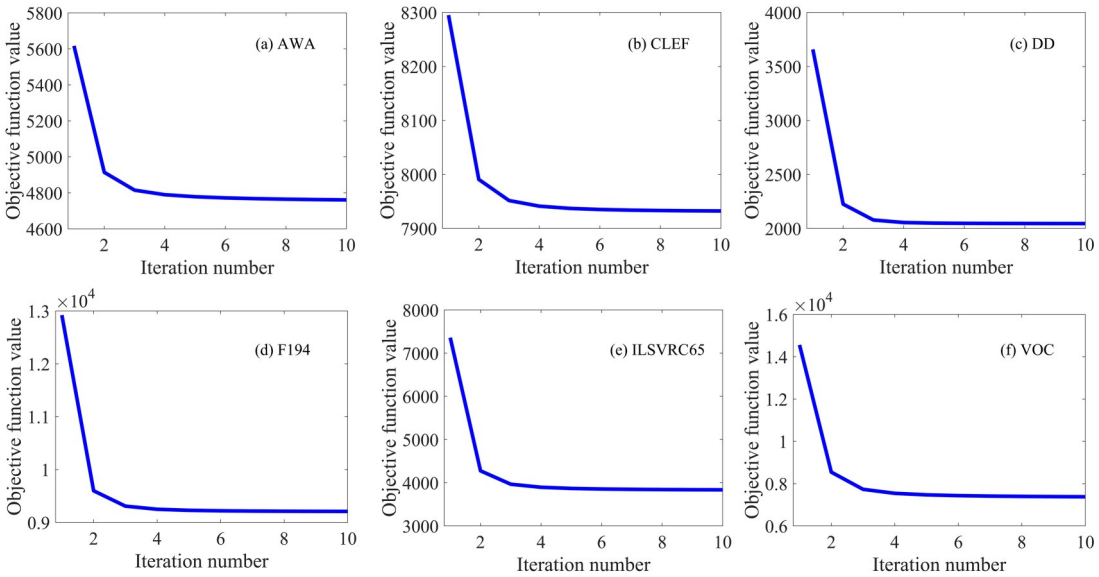

图 6 目标函数值的收敛曲线

Fig.6 The convergence curve of the objective function value

4 结论

本文提出一种结合标记分布建模与互信息度量的层次特征选择方法. 通过标记分布建模,充分挖掘类别间的层次结构信息,并将特征与标记分布之间的互信息融合到特征选择框架,以增强特征排序的相关性与判别性. 实验结果表明,该方法在多个层次数据集上具有良好的有效性与鲁棒性.

本研究为复杂结构标记下的特征选择提供了新思路,未来将探索更精细的标记增强方法,进一步提升分类性能.

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. 中国科学(信息科学),2018,48(5):487－500.

[2] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248－255.

[3] Sun A X, Lim E P. Hierarchical text classification and evaluation//Proceedings 2001 IEEE International Conference on Data Mining. San Jose, CA, USA: IEEE, 2001: 521－528.

[4] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211－252.

[5] Zadeh L A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems, 1997, 90(2): 111－127.

[6] 张超,丁雨欣,李文涛,等. 新一代人工智能背景下粒计算研究现状与展望. 南京理工大学学报,2025, 49(3):265－277.

[7] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. International Journal of General System, 1990,17(2/3):191－209.

[8] Wang C Z, Qi Y L, Shao M W, et al. A fitting model for feature selection with fuzzy rough sets. IEEE Transactions on Fuzzy Systems, 2017, 25(4): 741－753.

[9] Wang C Y, Wang C Z, An S, et al. Feature selection and classification based on directed fuzzy rough sets. IEEE Transactions on Systems, Man, and Cybernetics:Systems,2025,55(1):699－711.

[10] Deng Z X, Li T R, Zhang P F, et al. Feature selection for label distribution learning based on neighborhood fuzzy rough sets. Applied Soft Computing, 2025, 169: 112542.

[11] 折延宏,黄婉丽,贺晓丽,等. 面向层次结构数据的增量特征选择. 计算机科学与探索, 2023, 17(12): 2928－2941.

[12] Zhao H, Wang P, Hu Q H, et al. Fuzzy rough set based feature selection for large-scale hierarchical classification. IEEE Transactions on Fuzzy Systems, 2019,27(10):1891－1903.

[13] Qiu Z Y, Zhao H. A fuzzy rough set approach to hierarchical feature selection based on Hausdorff distance. Applied Intelligence, 2022, 52(10): 11089－11102.

[14] Bai S X, Lin Y J, Lü Y, et al. Kernelized fuzzy rough sets based online streaming feature selection for large-scale hierarchical classification. Applied Intelligence, 2021,51(3):1602－1615.

[15] Wu F H, Zhang J, Honavar V. Reformulation and approximation//Zucker J D, Saitta L. Abstraction, Reformulation and Approximation. Heidelberg: Springer,2005:313－320.

[16] Silla C N, Freitas A A. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, 2011, 22(1): 31－72.

[17] Eisner R, Poulin B, Szafron D, et al. Improving protein function prediction using the hierarchical structure of the gene ontology//2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. La Jolla, CA, USA: IEEE,2005:1－10.

18] Ceci M, Malerba D. Classifying web documents in a hierarchy of categories: A comprehensive study. Journal of Intelligent Information Systems, 2007, 28 (1):37－78.

[19] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision,2010,88(2): 303－338.

[20] Li J D, Cheng K W, Wang S H, et al. Feature selection: A data perspective. ACM Computing Surveys,2017,50(6):1－45.

[21] 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. 软件学报,2022,33(7): 2667－2682.

[22] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning//Proceedings of the 20th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2006: 41－48.

[23] 张文修. 信息系统与知识发现. 北京:科学出版社, 2003.

[24] 陈曦,马建敏,刘权芳. 基于模糊依赖决策熵的多标签特征选择. 昆明理工大学学报(自然科学版),2024, 49(2):62－72.

[25] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 2004, 69(6 Pt 2): 066138.

[26] Ding C H, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics,2001,17(4):349－358.

[27] Li D P, Ju Y, Zou Q. Protein folds prediction with hierarchical structured SVM. Current Proteomcs, 2016,13(2):79－85.

[28] Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009:951－958.

[29] Dimitrovski I, Kocev D, Loskovska S, et al. Hierarchical annotation of medical images. Pattern Recognition, 2011, 44(10/11): 2436－2449.

[30] Deng J, Krause J, Berg A C, et al. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012: 3450－3457.

[31] Dekel O, Keshet J, Singer Y. Large margin hierarchical classification// Proceedings of the 21st International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2004: 27.

[32] Kosmopoulos A, Gaussier E, Paliouras G, et al. The ECIR 2010 large scale hierarchical classification workshop. ACM SIGIR Forum, 2010, 44(1): 23－32.

[33] Kira K, Rendell L A. A practical approach to feature selection// Proceedings of the 9th International Workshop on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992: 249－256.

[34] Nie F P, Huang H, Cai X, et al. Efficient and robust feature selection via joint $\ell 2,1$-norms minimization// Proceedings of the 24th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2010: 1813－1821.

[35] Zhao H, Zhu P F, Wang P, et al. Hierarchical feature selection with recursive regularization// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, 2017: 3483－3489.

[36] Wang Z, Nie F P, Tian L, et al. Discriminative feature selection via a structured sparse subspace learning module// Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Yokohama, Japan: IJCAI, 2020: 3009－3015.

[37] Liu X X, Zhou Y C, Zhao H. Robust hierarchical feature selection driven by data and knowledge. Information Sciences, 2021, 551: 341－357.

[38] Friedman M. A comparison of alternative tests of significance for the problem of M rankings. The Annals of Mathematical Statistics, 1940, 11(1): 86－92.

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

(责任编辑 杨可盛)