山西大学学报(自然科学版)48(1):144－157,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024137

基于正交约束和最大类内特征判别性的分层分类特征选择算法

殷金钊[1],郑文利[2],钱婷[2,3*],折延宏[2,3]
(1.西安石油大学 计算机学院,陕西 西安 710065;
2.西安石油大学 理学院,陕西 西安 710065;
3.西北大学 概念、认知与智能研究中心,陕西 西安 710127)

摘　要:在大数据时代,数据正呈现出指数级增长趋势。数据间的类别层次结构使得分类学习任务更有效率。现有的分层分类特征选择算法未充分体现出类内特征的判别性,因此本文提出了一种基于正交约束和最大化类内特征判别性的分层分类特征选择算法(Hierarchical Classification Feature Selection Algorithm Based on Orthogonal Constraints and Intra-class Maximum Feature Discriminability,HFSOC)。该算法在使用稀疏正则化项去除不相关特征后,利用改进后的正交约束公式来度量类间独立性,并将每个内部节点特征矩阵的各个列向量互相正交,以提高类内特征的判别性。最后,利用递归正则项优化输出特征权重矩阵。实验结果表明,本文所提算法在5个数据集上取得了一定的效果,其分类准确率在DD数据集上相比于HFisher算法提高约17%,在F194数据集和CLEF数据集上相比于基于$\ell_{2,1}$范数最小化的高效鲁棒的特征选择算法(HFSNM)均提高约10%,在ILSVRC数据集上相比于HFSNM算法提高约1%。

关键词:特征选择;稀疏学习;层次树结构;正交约束;递归正则化
中图分类号:TP311　　**文献标志码**:A　　**文章编号**:0253-2395(2025)01-0144-14

Hierarchical Classification Feature Selection Algorithm Based on Orthogonal Constraints and Intra-class Maximum Feature Discriminability

YIN Jinzhao[1], ZHENG Wenli[2], QIAN Ting[2,3*], SHE Yanhong[2,3]
(1. *College of Computer, Xi'an Shiyou University, Xi'an* 710065*, China;*
2. *College of Science, Xi'an Shiyou University, Xi'an* 710065*, China;*
3. *Institute of Concepts, Cognition and Intelligence, Northwest University, Xi'an* 710127*, China*)

Abstract: In the era of big data, data is showing an exponential growth trend. The hierarchical structure of categories among data makes the classification learning task more efficient. However, existing hierarchical classification feature selection algorithms do not fully reflect the discriminative nature of intra-class features. This paper proposes a hierarchical classification feature selection algorithm based on orthogonal constraints and maximizing the discriminative nature of intra-class features(HFSOC). The algorithm utilizes an improved orthogonality constraint formula to measure inter-class independence and orthogonalizes the individual column

收稿日期:2024-06-14;**接受日期**:2024-10-24
基金项目:国家自然科学基金(12171388;61976244;12101478;12171294);陕西省自然科学基础研究计划项目(2023JCYB027);浙江海洋大学海洋大数据挖掘与应用重点实验室(OBDMA202101);陕西数理基础科学研究项目(23JSZ008);研究生创新项目(YCX2413143)
作者简介:殷金钊(2000－),女,山东聊城人,硕士研究生,研究方向为分层分类。E-mail:yin2991802968@163.com
* **通信作者**:钱婷(QIAN Ting),E-mail:qiant2000@126.com
引文格式:殷金钊,郑文利,钱婷,等.基于正交约束和最大类内特征判别性的分层分类特征选择算法[J].山西大学学报(自然科学版),2025,48(1):144-157. DOI:10.13451/j.sxu.ns.2024137.

vectors of each internal node's feature matrix to each other to improve the discriminative property of intra-class features after using sparse regularization terms to remove irrelevant features. Finally, the output feature weight matrix is optimized using recursive regularization terms. The experimental results show that the algorithm proposed in this paper achieves certain results on five datasets, and its classification accuracy is improved by about 17% compared to HFisher's algorithm on the DD dataset, by about 10% compared to efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization algorithm (HFSNM) on both the F194 dataset and the CLEF dataset, and by about 1% compared to HFSNM's algorithm on the ILSVRC dataset.

Key words: feature selection; sparse learning; hierarchical tree structure; orthogonal constraint; recursive regularization

0 引言

在大数据时代,数据正呈现出指数级增长趋势,数据样本急剧增加,数据种类也由最初的数百数千发展至数十万数百万乃至更多[1]。例如,在ImageNet中图像数据的类别就有上万个,图像数据样本更是达到了上千万之多[2]。在超大规模的数据下,人们面临的分类学习任务也遇到了新的挑战[3]。面对这种情况,学者通常采用分而治之策略在类别间建立层次结构进行分类学习。这种具有层次结构的分类任务被称为分层分类[4]。此后,许多学者针对分层分类问题展开了研究。Wang等[5]认为在层次结构分类时每个非叶节点存在保守风险和冒进风险,提出了局部Bayes风险最小化的分类模型,即通过在每个非叶节点比较保守风险和冒进风险来决定停止或者继续向下进行分类。Zhou等[6]提出了一种深度超类学习模型来解决长尾分布式数据分类问题,该模型将组稀疏与卷积神经网络相结合,类别之间的视觉相似度越高,它们出现在同一组中的可能性越大,进而提高分类精度。分层分类学习已成为一个热点研究问题[7-12]。

在分类任务中,特征选择是一项不可或缺的环节[13-14]。研究人员提出了各种分层分类特征选择算法。Freeman等[15]提出了一种多类分类模型,该模型一边构建层次树结构,一边对局部分类器进行特征选择,二者同时进行。该模型可以为每个分类子任务选取互相独立的特征子集。Grimaudo等[16]提出了基于层次结构的分类器,在每个非叶节点设立分类器,每个分类器相互独立,利用mRMR特征选择算法为每个分类器选择有代表性的特征。Song等[17]提出了一种基于信息增益和特征位置信息及其频率分布的分层分类特征选择算法。该算法使用类别的语义结构进行特征降维,并且为每个类别选取适当且互相独立的特征子集。上述算法解决了传统特征选择中只能选择统一特征子集的问题,但均未考虑层次结构间的各种关系。

父子关系及兄弟关系是层次结构中常被考虑的类别关系[3]。Tuo等[18]为了探索不同类别之间的双向依赖关系,提出了基于子树的图正则化分层分类特征选择算法。Zhao等[19]提出了类别的层次结构与递归正则化结合的特征选择算法。该算法利用父子节点之间的相关性以及兄弟节点之间的独立性分别构造正则项,进而递归地为每个非叶节点选择不同的特征子集。Lin等[20]认为层次结构中的内部节点之间存在共有特征和固有特征,进而提出利用损失函数以及类别相似度矩阵构建特征选择目标函数来学习层次结构中不同类别的固有特征和共有特征。但上述这些算法均未考虑类内特征之间的关系。Shi等[21]提出了基于类间最大独立性和类内最小冗余性的分层分类特征选择算法,具有地,其利用兄弟关系正交约束提高类间独立性,用内积项度量类内相关性进而减小类内特征冗余性。

运用正交约束理论刻画特征关系是一种有效的手段,但我们发现Shi等[21]提出的正交约束会导致兄弟节点之间选取的特征较为一致,影响兄弟节点之间的独立性。例如对于特征权重矩阵为2×2的布尔矩阵,为使得$\left\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i-\boldsymbol{E}\right\|_F^2$取最小,而极端情况下的最小值就是$\left\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i-\boldsymbol{E}\right\|_F^2=0$,我们通过穷举法发现,只有$\boldsymbol{W}_i=\begin{pmatrix}1&0\\0&1\end{pmatrix}\boldsymbol{W}_j=\begin{pmatrix}1&0\\0&1\end{pmatrix}$或者$\boldsymbol{W}_i=\begin{pmatrix}0&1\\1&0\end{pmatrix}\boldsymbol{W}_j=\begin{pmatrix}0&1\\1&0\end{pmatrix}$这两种组合情况能满足最小约束,而这两种组合情况均说明兄弟节点之间所选特征一致,故不能较好地反映兄弟节点

之间特征的独立性。

其次Shi等[21]通过利用行与行之间的正交关系考虑类内特征冗余性,降低了特征维数,但其忽略了分类的需求。我们仍以2×2的布尔矩阵位列,假设非叶节点i的类内特征权重矩阵$\boldsymbol{W}_i$为$\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$,从矩阵可以看出两个特征对两个类别都重要且重要程度相同,并且二者之间的相关性强,冗余性大。经过内积项对特征(即行与行)之间的相关性进行惩罚之后,非叶节点i的特征权重矩阵$\boldsymbol{W}_i$被约束为$\begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix}$。从该矩阵中我们可以看出特征的冗余性变小。但是所选特征仍然不能进行有效分类,这就是所选类内特征判别性小。基于选择特征的最终目的是分类,所以类内特征选择时首要考虑筛选后的特征具有最大判别性。

因此本文提出基于正交约束和最大化类内特征判别性的分层分类特征选择算法(Hierarchical Classification Feature Selection Algorithm Based on Orthogonal Constraints and Intra-class Maximum Feature Discriminability, HFSOC)。并做了相关对比实验,实验结果表明,本文所提出的HFSOC算法具有一定的有效性。

本文相关工作如下:第1节介绍相关理论知识;第2节介绍基于正交约束和最大化类内特征判别性的分层分类特征选择算法;第3节将所提算法与其他算法进行比较,并使用多种评价指标对结果进行分析;第4节进行总结与展望。

1 相关理论

1.1 类别的层次树结构

类别的层次结构主要分为两种类型,树型和有向无环图型[22]。为了便于理解,本论文重点讨论树型层次结构。层次树结构将类别标签组织成树状结构,以表示类别之间的一种“从属”关系[4]。用$(Y,\prec)$表示层次结构,Y表示标签集合,“$\prec$”表示“从属”关系。类别间“从属”关系的三种性质描述如下:

(1)不可逆性:若$i\prec j$,则对于$\forall i,j\in Y$,有$j\not\prec i$;

(2)反自反性:对$\forall i\in Y$,有$i\not\prec i$;

(3)传递性:若$i\prec k$且$k\prec j$,则对$\forall i,j,k\in Y$,有$i\prec j$。

1.2 基于稀疏学习的特征选择

在有监督的特征学习中,由于稀疏学习具有良好的可解释性,所以其常为特征选择首选的技术手段[23]。分层特征选择的目标是确保在一定的精度损失下挑选出少数有代表性的特征,因此将分层特征选择与稀疏学习相结合可以达到更好的效果[24]。设$\boldsymbol{X}\in R^{n\times m}$是样本矩阵,其中$n$表示样本数量,$m$表示特征数量。设$T=\{0,1,\cdots,N\}$表示类别层次树结构的非叶子节点集,当$N=0$时表示根节点。令$\boldsymbol{X}_0,\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N$表示每个非叶节点的样本矩阵,其中$\boldsymbol{X}_i=[x_i^1,x_i^2,\cdots,x_i^{n_i}]\in R^{n_i\times m}(0\leqslant i\leqslant N,n_i\leqslant n)$;$\boldsymbol{Y}_0,\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_N$表示每个非叶节点对应的标签矩阵,其中$\boldsymbol{Y}_i=[y_i^1,y_i^2,\cdots,y_i^{n_i}]\in R^{n_i\times d}$,并且$y_i^j=\{0,1\}^d(1\leqslant j\leqslant n_i)$,$d$是非叶节点中类别数目的最大值。

基于稀疏学习的特征选择目标函数[25]的一般形式为

$$\min_{w} L(\boldsymbol{XW},\boldsymbol{Y})+\lambda R(\boldsymbol{W}), \tag{1}$$

其中$L(\cdot)$是一个损失项,$R(\boldsymbol{W})$是一个稀疏正则项,$\lambda>0$。在常见的损失函数[26]中,最小二乘损失具有计算简单、容易实现、最优解唯一等优点[19],因此本文所提算法的损失函数使用最小二乘损失进行计算,如公式(2):

$$L(\boldsymbol{XW},\boldsymbol{Y})=\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2。 \tag{2}$$

在稀疏学习方法中,常见的正则项包括ℓ_0范数、ℓ_1范数、ℓ_2范数、$\ell_{2,0}$范数和$\ell_{2,1}$范数等[20]。其中

ℓ_0范数可以实现特征稀疏化,但由于其非凸、不光滑且不连续的特性,以至于不方便求解;ℓ_1范数不能进行类的结构稀疏化;ℓ_2范数不能实现特征稀疏化;$\ell_{2,0}$范数预先定义了所选特征的数量,容易陷入局部最优;相比之下$\ell_{2,1}$范数可以实现类的结构稀疏化,并且是凸的,便于优化[27]。基于此,本文采用$\ell_{2,1}$范数作为稀疏学习的正则项。因此稀疏正则项可表示为

$$R(\boldsymbol{W})=\|\boldsymbol{W}\|_{2,1}。\tag{3}$$

综上,结合公式(2)和公式(3)可将基于稀疏学习的特征选择目标函数公式(1)重新表示为

$$\min_{W}\sum_{i=0}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\|\boldsymbol{W}_i\|_{2,1}\right),\tag{4}$$

其中$\boldsymbol{W}_i$表示非叶节点i的权重矩阵,并且$\boldsymbol{W}_i=\left[\boldsymbol{w}_i^1,\boldsymbol{w}_i^2,\cdots,\boldsymbol{w}_i^m\right]\in R^{m\times d}$。

2 算法模型

2.1 基于正交约束的兄弟关系正则项

类别层次结构中兄弟关系是节点间的重要依赖关系[3]。一般来说,兄弟节点来自同一个父节点,相比于来自不同父节点的类别共享更多的特征,但是每个内部节点都有自己的子树,在每个内部节点挑选的特征都应具备判别子类别的特性。因此兄弟节点间应减少共享特征,增加判别性特征,进而加大类间的独立性。

正交约束是指在求解优化问题时,强制让变量在某个规范化的空间中正交。在机器学习中,正交约束优化方法将权重矩阵分解为两个正交矩阵的乘积,这样做有助于减少模型中的冗余信息,提高模型的泛化能力和稳定性。本文将修改Shi等[21]所提出的结构关系正则化项,使得经过正交约束后兄弟节点之间的特征更具有独立性,以及避免出现上述所分析的极端情况。设S_i为第i个非叶节点的兄弟节点,$j\in S_i$表示第i个非叶节点的第j个兄弟节点。因此兄弟关系正则化定义为

$$\sum_{j\in S_i}\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i\|_F^2。\tag{5}$$

基于以上讨论,将兄弟关系正则项加入稀疏学习特征选择公式(4)中,以便提高特征的独立性。因此目标函数定义为

$$T(\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N)=\min_{\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N}\sum_{i=0}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\lambda\|\boldsymbol{W}_i\|_{2,1}\right)+\alpha\sum_{i=1}^{N}\sum_{j\in S_i}\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i\|_F^2,\tag{6}$$

其中第一项是损失函数,用来度量预测标签和真实标签之间的误差程度;第二项是稀疏正则项,用于最小化特征权重值,增强模型的稀疏性,避免过拟合;第三项是兄弟关系正则项,用于加大类间的独立性,α是惩罚兄弟节点间依赖度的非负参数。

2.2 基于正交约束的类内判别正则项

类别层次结构中每个非叶节点都有自己的子类别,因此在非叶节点i的特征子集应该具有最大程度判别子类别的作用,即具备最大类内特征判别性。已知特征权重矩阵$\boldsymbol{W}_i$的每个列向量$\boldsymbol{w}_i^j$对应着非叶节点i的第j个子类别的特征的权重,因此$\boldsymbol{W}_i$中列与列之间的相关性即表示非叶节点i的子类别之间的相关性。相关性越大,选取的特征之间越相似,对子类别的判别性越弱;相关性越小,选取的特征之间越独立,对子类别的判别性越强。受正交约束方法的启发,将特征权重矩阵$\boldsymbol{W}_i$的每个列向量互相正交,可以惩罚非叶节点i上子类别之间的相关性,进而提高子类别特征之间的独立性,使得在非叶节点i挑选出的特征可以最大化地判别子类别。因此类内特征关系正则化定义为

$$\|\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i-\boldsymbol{E}\|_F^2,\tag{7}$$

其中$\boldsymbol{E}$是单位矩阵,减$\boldsymbol{E}$是为了忽略各个列向量与自身的正交约束。

综上讨论,我们在稀疏学习特征选择公式(4)中添加类内判别正则项来加强类内特征的判别性,因此目标函数定义为

$$T(\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N)=\min_{\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N}\sum_{i=0}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\lambda\|\boldsymbol{W}_i\|_{2,1}+\beta\|\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i-\boldsymbol{E}\|_F^2\right),\tag{8}$$

其中前两项分别是损失函数和稀疏正则项;第三项是类内判别正则项,用于提高类内特征的判别性,β是惩罚类内特征相关性的非负参数。

2.3 基于正交约束和最大化类内特征判别性的分层分类特征选择算法

在本节中,结合考虑兄弟关系公式(5)、类内特征关系公式(7)和稀疏学习特征选择公式(4),可得最终目标函数公式

$$T(\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N)=\min_{\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N}\sum_{i=0}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\lambda\|\boldsymbol{W}_i\|_{2,1}+\beta\|\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i-\boldsymbol{E}\|_F^2\right)+\alpha\sum_{i=1}^{N}\sum_{j\in S_i}\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i\|_F^2。\tag{9}$$

通过优化公式(9),得到第i个非叶节点的特征权重矩阵$\boldsymbol{W}_i$,我们对相应任务中的特征权重按降序排序,选择排名靠前的特征作为最优特征。

2.4 算法优化

由于$\ell_{2,1}$范数的非光滑性导致优化时很难推导出封闭解。针对这个问题,Argyriou等[27]提出了解决方法,当$w_i\neq 0$,$i=1,\cdots,d$时,$\|\boldsymbol{W}\|_{2,1}$对$\boldsymbol{W}$的导数为

$$\frac{\partial\|\boldsymbol{W}\|_{2,1}}{\partial\boldsymbol{W}}=\frac{\partial Tr(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{W})}{\partial\boldsymbol{W}}=2\boldsymbol{D}\boldsymbol{W},\tag{10}$$

其中$\boldsymbol{D}\in R^{d\times d}$是对角矩阵,第$j$个对角元素是$D_{jj}=\dfrac{1}{2\|w_j\|_2}$,若$w_j=0$,则$D_{jj}=\varepsilon$。

本文对$\ell_{2,1}$范数使用上述优化方法,由于在层次结构中根节点没有兄弟节点,因此根节点和非根节点的内部节点需要分开优化计算,因此最终目标函数公式(9)可变换为

$$T(\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N)=\min_{\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N}\left(\|\boldsymbol{X}_0\boldsymbol{W}_0-\boldsymbol{Y}_0\|_F^2+\lambda\|\boldsymbol{W}_0\|_{2,1}+\beta\|\boldsymbol{W}_0^{\mathrm{T}}\boldsymbol{W}_0-\boldsymbol{E}\|_F^2\right)+$$
$$\sum_{i=1}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\lambda\|\boldsymbol{W}_i\|_{2,1}+\alpha\sum_{j\in S_i}\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i\|_F^2+\beta\|\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i-\boldsymbol{E}\|_F^2\right)。\tag{11}$$

对于根节点,目标函数为

$$T(\boldsymbol{W}_0)=\min_{\boldsymbol{W}_0}\|\boldsymbol{X}_0\boldsymbol{W}_0-\boldsymbol{Y}_0\|_F^2+\lambda\|\boldsymbol{W}_0\|_{2,1}+\beta\|\boldsymbol{H}_0\boldsymbol{W}_0-\boldsymbol{E}\|_F^2。\tag{12}$$

为了方便维度计算,令$\boldsymbol{H}_0=\boldsymbol{W}_0^{\mathrm{T}}$。然后将(12)式中$\boldsymbol{W}_0$的导数设为0,即

$$\frac{\partial T}{\partial\boldsymbol{W}_0}=2\boldsymbol{X}_0^{\mathrm{T}}(\boldsymbol{X}_0\boldsymbol{W}_0-\boldsymbol{Y}_0)+2\lambda\boldsymbol{D}_0\boldsymbol{W}_0+2\beta\boldsymbol{H}_0^{\mathrm{T}}(\boldsymbol{H}_0\boldsymbol{W}_0-\boldsymbol{E})=$$
$$2\left(\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{X}_0+\lambda\boldsymbol{D}_0+\beta\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{H}_0\right)\boldsymbol{W}_0-2\left(\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{Y}_0+\beta\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{E}\right)=0。\tag{13}$$

因此,求得$\boldsymbol{W}_0$为

$$\boldsymbol{W}_0=\left(\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{X}_0+\lambda\boldsymbol{D}_0+\beta\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{H}_0\right)^{-1}\left(\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{Y}_0+\beta\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{E}\right)。\tag{14}$$

对于中间节点,目标函数为

$$T(\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N)=\min_{\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N}\sum_{i=1}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\lambda\|\boldsymbol{W}_i\|_{2,1}+\alpha\sum_{j\in S_i}\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i\|_F^2+\beta\|\boldsymbol{H}_i\boldsymbol{W}_i-\boldsymbol{E}\|_F^2\right)。\tag{15}$$

同样为了便于优化,令$\boldsymbol{H}_i=\boldsymbol{W}_i^{\mathrm{T}}$。然后将(15)式中$\boldsymbol{W}_i$的导数设为0,即

$$\frac{\partial T}{\partial\boldsymbol{W}_i}=2\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i)+2\lambda\boldsymbol{D}_i\boldsymbol{W}_i+2\alpha\sum_{j\in S_i}\boldsymbol{W}_j\left(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i\right)+2\beta\boldsymbol{H}_i^{\mathrm{T}}(\boldsymbol{H}_i\boldsymbol{W}_i-\boldsymbol{E})=$$
$$2\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\lambda\boldsymbol{D}_i+\alpha\sum_{j\in S_i}\boldsymbol{W}_j\boldsymbol{W}_j^{\mathrm{T}}+\beta\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{H}_i\right)\boldsymbol{W}_i-2\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Y}_i+\beta\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{E}\right)=0,\tag{16}$$

因此,求得$\boldsymbol{W}_i$为

$$\boldsymbol{W}_i=\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\lambda\boldsymbol{D}_i+\alpha\sum_{j\in S_i}\boldsymbol{W}_j\boldsymbol{W}_j^{\mathrm{T}}+\beta\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{H}_i\right)^{-1}\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Y}_i+\beta\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{E}\right)。\tag{17}$$

算法伪代码如算法1 HFSOC所示,非叶节点的特征权重矩阵更新计算在第2行到第12行,对角矩阵在第3行到第5行更新计算,根节点的特征权重矩阵 $\boldsymbol{W}_0$ 在第6行更新计算,中间节点的特征权重矩阵 $\boldsymbol{W}_i$ 在第7行到第9行更新计算。算法1的时间复杂度主要取决于特征权重矩阵 $\boldsymbol{W}$ 的更新时间,每个中间节点的 $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i$ 与 $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Y}_i$ 各需要计算一次,时间复杂度分别为 $O(m^2n_i)$,$O(mdn_i)$;其中 m 表示特征数,n_i 表示第 i 个内部节点的样本数,d 表示内部节点的最大类别数。因此,所有中间节点更新的时间复杂度为 $O(m^2n+mdn)$;根节点更新的时间复杂度为 $O(m^3+m^2d)$。假设总迭代次数为 K,那么HFSOC算法的总时间复杂度为 $O\left(K(m^3+m^2d)+m^2n+mdn\right)$。

算法1 HFSOC

输入:样本矩阵 $\boldsymbol{X}_i\in R^{n_i\times m}$,标签 $Y\in\{0,1\}^{n_i\times d}$,其中 $i=0,1,\cdots,N$,d 是内部节点的最大类别数,正则化参数 λ,α,β,迭代次数 F

输出:特征权重矩阵 $\boldsymbol{W}\in R^{m\times d}$

1:设置迭代次数 f,随机初始化矩阵 $\boldsymbol{W}\in R^{m\times d}$,$\boldsymbol{W}=\left[\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N\right]$

2:WHILE $f<F$ DO

3: FOR $i=0:N$ DO

4: 通过 $D_{jj}=1\Big/\left(2\left\|w_j\right\|_2\right)$ 求得矩阵 $\boldsymbol{D}_i^{(f)}$

5: END FOR

6: 更新 $\boldsymbol{W}_0$:$\boldsymbol{W}_0^{(f+1)}=\left(\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{X}_0+\lambda\boldsymbol{D}_0^{(f)}+\beta\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{H}_0\right)^{-1}\left(\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{Y}_0+\beta\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{E}\right)$

7: FOR $i=1:N$ DO

8: 更新 $\boldsymbol{W}_i$:$\boldsymbol{W}_i^{(f+1)}=(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\lambda\boldsymbol{D}_i^{(f)}+\alpha\sum_{j\in S_i}\boldsymbol{W}_j\boldsymbol{W}_j^{\mathrm{T}}+\beta\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{H}_i)^{-1}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Y}_i+\beta\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{E})$

9: END FOR

10: 更新 $\boldsymbol{W}^{(f+1)}=\left[\boldsymbol{W}_0^{(f+1)},\boldsymbol{W}_1^{(f+1)},\cdots\boldsymbol{W}_N^{(f+1)}\right]$

11: $f=f+1$

12:END WHILE

13:返回 $\boldsymbol{W}$

3 实验分析

3.1 数据集

本实验使用了5个数据集,所有数据集均具有层次结构,其中有两个蛋白质数据集:DD[28]和F194[29],三个图像数据集:CLEF[30]、VOC(The PASCAL Visual Object Classes dataset)[31]和ILSVRC[32]。表1给出了数据集的详细信息。

3.2 评价指标

为了能更准确地衡量算法,除了传统的度量精度的方法以外,实验还引用了以下两种分层分类特有的评价指标:

表1 数据集描述

Table 1 Description of the datasets

数据集	训练集	测试集	特征数	节点数	叶子节点数	层数
DD	3 020	605	473	32	27	3
F194	7 105	1 420	473	202	194	3
CLEF	8 368	939	80	88	63	3
VOC	7 178	5 105	1 000	30	20	5
ILSVRC	12 346	11 845	4 096	65	57	4

树诱导误差(Tree Induced Error, TIE)[33]能够反映测试样本在层次结构上的误差程度,用真实标签节点和预测标签节点在层次结构中的总边数表示,即节点i的TIE表示为:

$$i_{\mathrm{TIE}}(y,\hat{y})=\sum_{E}(y,\hat{y})。\tag{18}$$

Hierarchical-$F1$ measure[25]在传统的$F1$ measure的基础上充分考虑了真实类别和预测类别的祖先和后代,它代表了算法在准确率和召回率上的总体分类表现。用准确率和召回率的调和平均数表示,即

$$F_H=\frac{2\cdot P_H\cdot R_H}{P_H+R_H}。\tag{19}$$

准确率P_H和召回率R_H分别为

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

其中Y_{aug}表示真实标签扩展集,$\hat{Y}_{\mathrm{aug}}$表示预测标签扩展集,且有$Y_{\mathrm{aug}}=y\cup Anc(y)$,$\hat{Y}_{\mathrm{aug}}=\hat{y}\cup Anc(\hat{y})$,$Anc(y)$和$Anc(\hat{y})$分别表示真实标签$y$和预测标签$\hat{y}$的祖先节点集合。

3.3 对比算法

在本实验中选取的六种对比算法如下所示:

(1)HFisher[10]:Fisher分数是一种用于多元分类和数据降维的有效方法。HFisher算法是在Fisher分数的基础上进行改进后与层次结构结合的分层分类特征选择算法。

(2)HFSNM(Efficient and Robust Feature Selection via Joint $\ell_{2,1}$-norms Minimization)[11]:基于$\ell_{2,1}$范数最小化的高效鲁棒的特征选择算法,该算法对损失函数和正则化联合使用$\ell_{2,1}$范数最小化,即对所有数据进行$\ell_{2,1}$范数正则化,选择具有联合稀疏性的特征,实现了鲁棒的特征选择目标。

(3)HFSDK(Robust Hierarchical Feature Selection Driven by Data and Knowledge)[12]:基于数据和知识驱动的鲁棒的分层分类特征选择算法,该算法采用自顶向下的方式从粗粒度到细粒度为每个非叶节点选择具有鲁棒性和判别性的局部特征子集。

(4)Hier-FS(Hierarchical Feature Selection)[19]:基于层次结构的特征选择算法,该算法只考虑节点间的层次结构不考虑层次结构间的依赖关系。

(5)HiRRfam-FS(Family Relationship Based Hierarchical Feature Selection with Recursive Regularization)[19]:基于家庭关系的递归正则化层次特征选择算法,该算法提出父子关系正则项和兄弟关系正则项,进而为每个节点选取不同的特征子集。

(6)HFS-MIMR(Feature Selection via Maximizing Inter-class Independence and Minimizing Intra-Class Redundancy for Hierarchical ClassifiCation)[21]:基于最大化类间独立性和最小化类内冗余性的分层分类特征选择算法,该算法不仅充分考虑类间结构关系同时也考虑类内特征关系来进行特征选择。

3.4 实验设置

本实验采用线性支持向量机作为分类器。实验参数λ设置为10,参数α和β的设置:蛋白质数据集的α和β分别是0.1和1,图像数据集的α和β分别是1和10。据文献[10-12,19,21],本实验所有对比算法均使用其调整好的参数,为了与对比算法保持一致,本实验保持分别采用10%的蛋白质数据集和20%的图像数据集的特征。另外,为了保证算法的精度,本实验采用十折交叉验证法作为模型评估方法。

实验环境:主机环境为16 GB内存、3.30 GHz的AMD Ryzen 5 5600H CPU和Windows 11系统,编程环境为Matlab 2021a软件。

3.5 实验结果及分析

3.5.1 性能比较

本节分别从三个评价指标列表分析,表2、表3和表4分别给出了7个算法在各个数据集上的准确率、Hierarchical-F1 measure和TIE。“↑”表示评价指标结果数值越大越好,“↓”表示评价指标结果数据越小越好。表中黑色粗体数值表示在各个数据集上最好的结果。

从三个表中可以得出,本论文所提算法在大部分数据集上都能表现出良好效果。例如在DD数据集上,本论文所提算法比HFisher的准确率高出0.178 4,Hierarchical-F1 measure高出0.090 9。与其他算法相比,本论文所提算法在VOC数据集上表现略差于HFSDK和HFS-MIMR,这可能是因为所提算法实验设置的参数不适合VOC数据集,并且VOC数据集中的噪声数据较多,算法具有鲁棒性对VOC数据集分类学习更有优势。总体来说,实验结果证明本论文所提算法在分类任务的特征选择方面具有良好的性能,并且能够在分类过程中带来更好的预测效果。

为了评价算法性能,引入统计学检验Friedman检验和Bonferroni-Dunn后验检验[34]。非参数检验Friedman检验的统计量表示为

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

其中k和N分别是算法和数据集的数量;R_i表示给定算法在所有数据集中的平均排名;F_F服从Fisher分布,自由度为$F(k-1,k-1(N-1))$。

利用表3的Hierarchical-F1 measure值进行排序,在每个数据集上性能最好的算法排名为1,第

表2 算法HFSOC与6个对比算法在各个数据集上的分类准确率(↑)

Table 2 Classification accuracy of algorithm HFSOC versus six comparison algorithms on each dataset

数据集	HFisher	HFSNM	HFSDK	Hier-FS	HiRRfam-FS	HFS-MIMR	HFSOC
F194	0.257 7	0.245 1	0.335 2	0.331 0	0.338 0	0.345 1	**0.345 8**
DD	0.520 7	0.682 8	0.687 7	0.682 7	0.686 1	0.687 8	**0.699 1**
CLEF	0.580 4	0.524 0	0.625 1	0.621 9	0.616 6	0.618 8	**0.629 4**
VOC	0.393 4	0.403 3	**0.430 2**	0.419 0	0.424 3	0.428 8	0.425 1
ILSVRC	0.850 1	0.843 1	0.850 9	0.850 9	0.849 7	0.851 0	**0.852 1**

注:↑表示评价指标结果数值越大越好。

表3 算法HFSOC与6个对比算法在各个数据集上的Hierarchical-F1 measure(↑)

Table 3 Hierarchical-F1 measure of algorithm HFSOC versus six comparison algorithms on each dataset

数据集	HFisher	HFSNM	HFSDK	Hier-FS	HiRRfam-FS	HFS-MIMR	HFSOC
F194	0.675 8	0.646 2	0.709 2	0.708 9	0.711 7	0.714 6	**0.716 2**
DD	0.774 1	0.852 4	0.859 0	0.858 4	0.860 6	0.861 2	**0.865 0**
CLEF	0.739 5	0.710 0	0.772 5	0.765 7	0.762 8	0.766 9	**0.773 9**
VOC	0.657 6	0.673 9	0.677 7	0.675 4	0.675 8	**0.679 4**	0.678 3
ILSVRC	0.958 0	0.956 3	0.959 3	0.959 1	0.958 8	0.958 9	**0.959 5**

注:↑表示评价指标结果数值越大越好。

表4 本文所提算法HFSOC与6个对比算法在各个数据集上的TIE(↓)

Table 4 TIE of algorithm HFSOC versus six comparison algorithms on each dataset

数据集	HFisher	HFSNM	HFSDK	Hier-FS	HiRRfam-FS	HFS-MIMR	HFSOC
F194	1.945	2.123	1.750	1.746	1.730	1.728	**1.721**
DD	1.230	0.830	0.850	0.850	0.836	0.826	**0.818**
CLEF	2.010	2.240	1.760	1.796	1.758	1.794	**1.746**
VOC	2.279	2.220	2.120	2.143	2.138	**2.115**	2.126
ILSVRC	0.336	0.350	0.334	0.328	0.329	0.327	**0.323**

注:↓表示评价指标结果数值越小越好。

二好的算法排名为2,以此类推。在相等数据值的情况下,分配平均排名。按照以上规则排好序后计算每个算法在所有数据集上的平均排名。原假设为所有特征选择算法性能都相同,通过计算得出F_F=16.270。对于七个算法和五个数据集的临界值$F(7-1,(7-1)\times(5-1))=F(6,24)$,大于$\alpha$=0.05时的$F$检验临界值2.508,因此拒绝原假设。即被比较的七种算法并不相同并且算法之间存在显著差异。

本文采用Bonferroni-Dunn后验检验来进一步比较算法之间的差异。Bonferroni-Dunn后验检验的临界值计算方式为$CD_\alpha=q_\alpha\sqrt{\dfrac{k(k+1)}{6N}}$。通过计算,在显著性水平$\alpha=0.1$时,有$q_\alpha=2.394$,因此可计算出$CD$=3.270 8($k$=7,$N$=5)。为了直观地展示这些差异,使用带有临界值的特殊图形来连接这些算法。图1为使用Bonferroni-Dunn后续检验比较所提算法HFSOC与其他六种对比算法性能的检验结果。

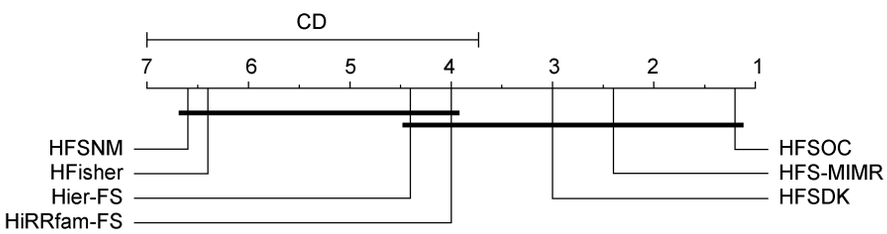


图1 通过Bonferroni-Dunn检验验证HFSOC算法性能

(图中数字横轴表示各个算法的平均性能排名,CD线表示两个算法间在统计上可被认为有显著差异的最小差异值。图中如果两个算法之间没有连线,表示这两个算法之间存在显著差异;反之,如果存在连线,则差异不显著。本图说明HFSOC算法与HFisher算法和HFSNM算法之间存在显著差异。)

Fig. 1 Verification of HFSOC algorithm performance by Bonferroni-Dunn test

(The numerical horizontal axis of the figure indicates the average performance ranking of each algorithm, and the CD line indicates the value of the smallest difference between two algorithms that can be considered statistically significant. If there is no line between two algorithms in the figure, it means that there is a significant difference between these two algorithms; conversely, if there is a line, the difference is not significant. This figure illustrates that there is a significant difference between the HFSOC algorithm and the HFisher and HFSNM algorithms.)

3.5.2 消融实验

本节通过消融实验验证算法中兄弟关系正则化和类内特征关系正则化的有效性,公式(11)各部分可以重新组合表示如下:

(1)Hspar:该公式为公式(11)减去兄弟关系正则化和特征关系正则化,仅剩下稀疏学习公式,即

$$\min_{W}\sum_{i=0}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\|\boldsymbol{W}_i\|_{2,1}\right)。\tag{20}$$

(2)HFSOC-α:该公式为公式(11)减去兄弟关系正则化,由稀疏学习公式和类内特征关系正则化组合表示,即

$$\min_{\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N}\sum_{i=0}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\lambda\|\boldsymbol{W}_i\|_{2,1}+\beta\|\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_i-\boldsymbol{E}\|_F^2\right)。\tag{21}$$

(3)HFSOC-β:该公式为公式(11)减去类内特征关系正则化,由稀疏学习公式和兄弟关系正则化组合表示,即

$$\min_{\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N}\sum_{i=0}^{N}\left(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_F^2+\lambda\|\boldsymbol{W}_i\|_{2,1}\right)+\alpha\sum_{i=1}^{N}\sum_{j\in S_i}\|\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_i\|_F^2。\tag{22}$$

图2中(a—e)分别展示了在各个数据集中HFSOC算法不同组合的Hierarchical-F1 measure结果。从图中我们可以看出,类间独立性正则项对F194数据集起到重要作用,而CLEF数据集和

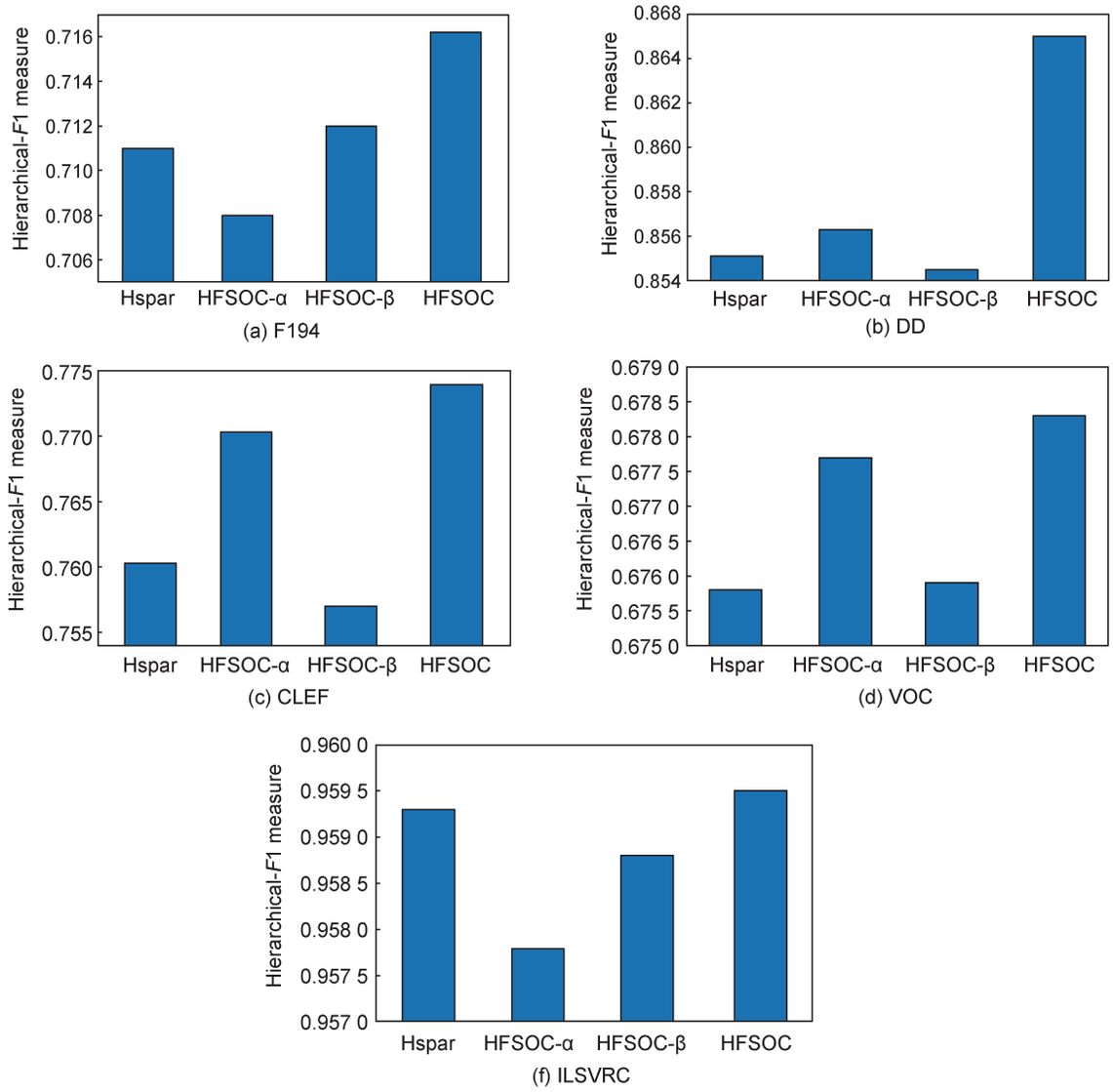


图2 各个数据集上的消融实验结果

((a—d)和(e)分别表示在F194数据集、DD数据集、CLEF数据集、VOC数据集和ILSVRC数据集上对不同模型组合的Hierarchical-F1 measure值的统计对比图,进而验证HFSOC算法的有效性。(a—d)和(e)中横轴均表示四个不同模型组合的公式,纵轴均表示Hierarchical-F1 measure值。)

Fig. 2 Results of ablation experiments on various datasets

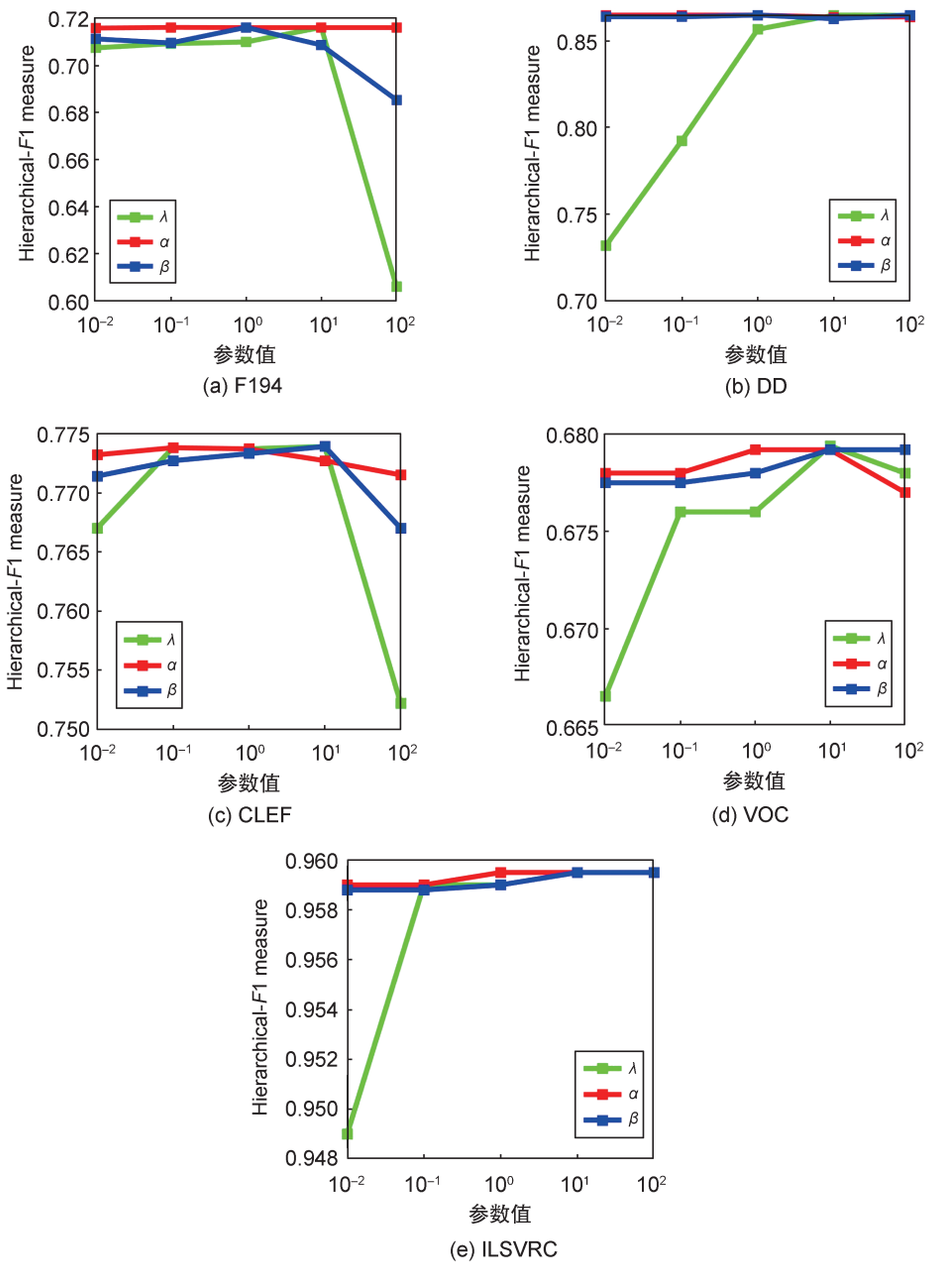
((a—d) and (e) represent the statistical comparison plots of the Hierarchical-1 measure values for different model combinations on the F194 dataset, the DD dataset, the CLEF dataset, the VOC dataset and the ILSVRC dataset, respectively, which in turn validate the effectiveness of the HFSOC algorithm. The horizontal axes in (a—d), and (e) all denote the formulas for the four different model combinations, and the vertical axes all denote the Hierarchical-F1 measure values.)

VOC数据集对类内特征判别性正则项更为敏感。DD数据集在重新组合的三个公式上表现均不佳,ILSVRC数据集的类别独立性高,易于分类,因此ILSVRC数据集在稀疏学习公式上表现与HFSOC不相上下。总体来说,HFSOC算法明显优于其他组合。

3.5.3 参数敏感性分析

本节进行参数敏感性分析,在本文所提算法HFSOC中共有三个参数,分别是λ、α和β。其中λ控制着稀疏正则项,α和β分别控制着兄弟关系正则项和类内特征关系正则项。我们采用网格搜索法在$\{0.01, 0.1, 1, 10, 100\}$中调整三个参数。实验通过固定两个参数改变一个参数来得出Hierarchical-F1 measure的结果,并根据此结果验证本文所提算法对变化参数的敏感性。

图3给出了各个数据集上的参数敏感性分析结果,可以得出当λ过小时,特征的稀疏性约束小,选择的特征较多,影响分类效果;当λ太大时,特征的稀疏性约束较大,选择的特征较少并且特征重要性降低,无法达到想要的效果。α和β与λ相比较为稳定,但当α和β过小时,兄弟节点之间选择的特征和类内选择的特征的独立性都较小,冗余性较大;当α和β太大时,无法选出紧凑的特征子集,影响分类效果。总体而言,在各个数据集上进行参数调整时,所提算法HFSOC的分类效果不受太大影响。



注:图中绿线表示固定参数α和β,调整参数λ;红线表示固定参数λ和β,调整参数α;蓝线表示固定参数λ和α,调整参数β。

图3 参数敏感性分析

(图(a—d)和(e)分别表示在F194数据集、DD数据集、CLEF数据集、VOC数据集和ILSVRC数据集上的参数敏感性分析,各个图中横轴均表示参数值,纵轴均表示Hierarchical-F1 measure值。)

Fig. 3 Parameter sensitivity analysis

(Figures (a—d), and (e) represent the parameter sensitivity analysis on the F194 dataset, the DD dataset, the CLEF dataset, the VOC dataset, and the ILSVRC dataset, respectively, and the horizontal axis of each figure represents the parameter values, and the vertical axis represents the Hierarchical-F1 measure values.)

3.5.4 收敛性分析

本节我们重点研究了HFSOC算法的收敛性,实验结果如图4所示。本实验设置每个数据集的最大迭代次数为10。从图4中可以看出,所有数据集的图像都是递减的,并且都在10次迭代以内收敛。

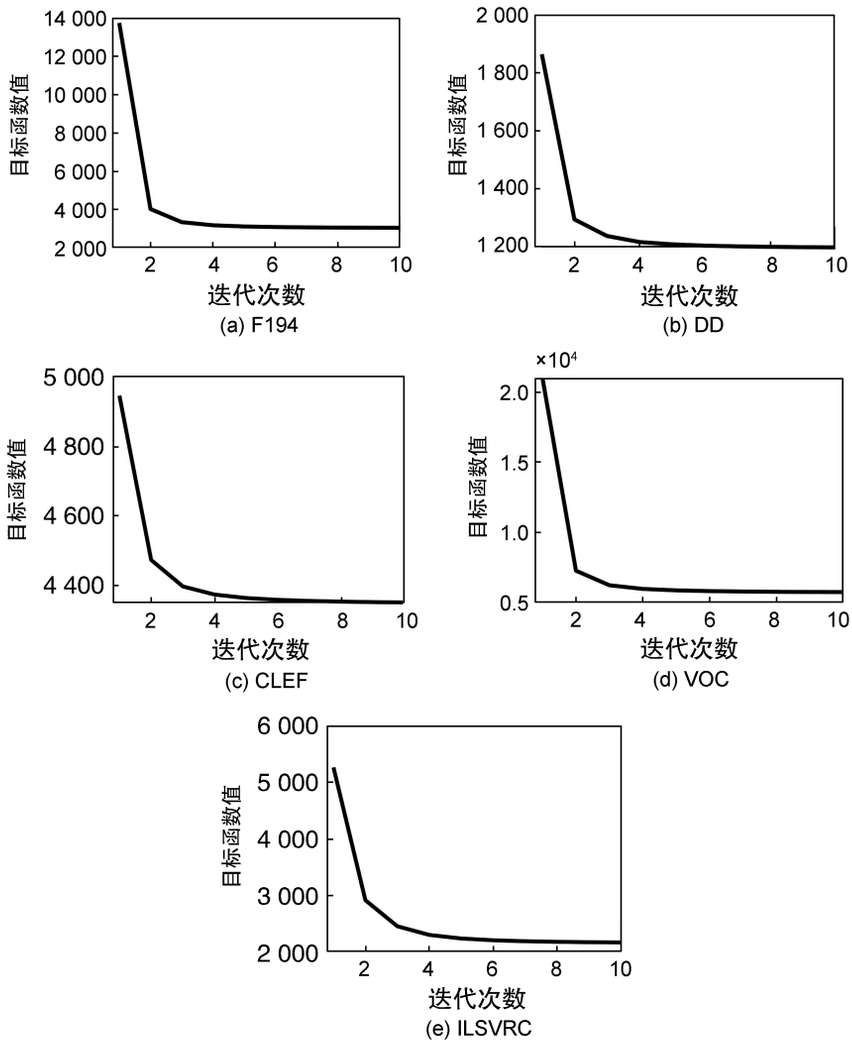


图4 目标函数值的迭代次数

(图(a—d)和(e)分别表示在F194数据集、DD数据集、CLEF数据集、VOC数据集和ILSVRC数据集上目标函数的收敛性,各个图中横轴均表示迭代次数,纵轴均表示目标函数值。)

Fig. 4 Number of iterations for the value of the objective function

(Figures (a—d) and (e) represent the convergence of the objective function on the F194 dataset, the DD dataset, the CLEF dataset, the VOC dataset, and the ILSVRC dataset, respectively, and the horizontal axes of the individual plots denote the number of iterations, and the vertical axes denote the objective function values.)

4 总结与展望

本文提出了一种基于正交约束和最大化类内特征判别性的分层分类特征选择算法。本文使用稀疏正则化项去除不相关特征,同时考虑类间特征独立性和类内特征判别性,利用递归正则化优化输出特征权重矩阵。通过7种算法在5个数据集上的对比实验显示,本文所提出的HFSOC算法能够做出良好表现。在本文的基础上,后续将在类内同时联合构造最小化冗余性和最大化判别性的正则项,并考虑去除噪声数据,设计鲁棒性特征选择算法。在所提算法中,手动调整参数较为

耗时,后续将建立自适应调参模型选择最优参数。并且在损失函数和优化方法上进行改进,使该模型能适用于有向无环图结构。

参考文献:

[1] BENGIO S, WESTON J, GRANGIER D. Label Embedding Trees for Large Multi-class Tasks[J]. *Adv Neural Inf Process Syst*, 2010, **23**: 163−171. DOI: US20120082371 A1.

[2] DENG J, DONG W, SOCHER R, *et al*. ImageNet: a Large-scale Hierarchical Image Database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248−255. DOI: 10.1109/CVPR.2009.5206848.

[3] 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. DOI: 10.1360/N112017-00246.
HU Q H, WANG Y, ZHOU Y C, *et al*. Review on Hierarchical Learning Methods for Large-scale Classication Task[J]. *Sci Sin Inform*, 2018, **48**(5): 487−500. DOI: 10.1360/N112017-00246.

[4] SILLA C N, FREITAS A A. A Survey of Hierarchical Classification across Different Application Domains[J]. *Data Min Knowl Discov*, 2011, **22**(1): 31−72. DOI: 10.1007/s10618-010-0175-9.

[5] WANG Y, HU Q H, ZHOU Y C, *et al*. Local Bayes Risk Minimization Based Stopping Strategy for Hierarchical Classification[C]//2017 IEEE International Conference on Data Mining (ICDM). New York: IEEE, 2017: 515−524. DOI: 10.1109/ICDM.2017.61.

[6] ZHOU Y C, HU Q H, WANG Y. Deep Super-class Learning for Long-tail Distributed Image Classification[J]. *Pattern Recognit*, 2018, **80**: 118−128. DOI: 10.1016/j. patcog.2018.03.003.

[7] LIN Y J, LIU H Y, ZHAO H, *et al*. Hierarchical Feature Selection Based on Label Distribution Learning[J]. *IEEE Trans Knowl Data Eng*, 2023, **35**(6): 5964−5976. DOI: 10.1109/TKDE.2022.3177246.

[8] WANG Y, HU Q H, ZHU P F, *et al*. Deep Fuzzy Tree for Large-scale Hierarchical Visual Classification[J]. *IEEE Trans Fuzzy Syst*, 2020, **28**(7): 1395−1406. DOI: 10.1109/TFUZZ.2019.2936801.

[9] WANG Y, WANG Z, HU Q H, *et al*. Hierarchical Semantic Risk Minimization for Large-scale Classification[J]. *IEEE Trans Cybern*, 2022, **52**(9): 9546−9558. DOI: 10.1109/TCYB.2021.3059631. [PubMed]

[10] DUDA R O, HART P E. Pattern Classification[M]. Hoboken: John Wiley & Sons, 2006. DOI: 10.1007/978-1-4471-0409-4_3.

[11] NIE F P, HUANG H, CAI X, *et al*. Efficient and Robust Feature Selection *via* Joint $\ell_{2,1}$-Norms Minimization[C]// Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. 2010: 1813−1821. DOI: 10.1007/978-3-319-10690-8_12.

[12] LIU X X, ZHOU Y C, ZHAO H. Robust Hierarchical Feature Selection Driven by Data and Knowledge[J]. *Inf Sci*, 2021, **551**: 341−357. DOI: 10.1016/j.ins.2020.11.003.

[13] 张子宁, 单甘霖, 段修生, 等. 基于改进遗传算法的支持向量机特征选择[J]. 电子产品世界, 2010, **17** (Z1): 45−47+51. DOI: 10.3969/j.issn.1005-5517.2010.1.008.
ZHANG Z N, SHAN G L, DUAN X S. Feature Selection for SVM Based on Improved Genetic Algorithm. [J]. *Electron Eng Prod World*, 2010, **17** (Z1): 45−47+51. DOI: 10.3969/j.issn.1005-5517.2010.1.008.

[14] PENG H C, LONG F H, DING C. Feature Selection Based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy[J]. *IEEE Trans Pattern Anal Mach Intell*, 2005, **27**(8): 1226−1238. DOI: 10.1109/TPAMI.2005.159.

[15] FREEMAN C, KULIĆ D, BASIR O. Joint Feature Selection and Hierarchical Classifier Design[C]//2011 IEEE International Conference on Systems, Man, and Cybernetics. New York: IEEE, 2011: 1728−1734. DOI: 10.1109/ICSMC.2011.6083921.

[16] GRIMAUDO L, MELLIA M, BARALIS E. Hierarchical Learning for Fine Grained Internet Traffic Classification[C]//2012 8th International Wireless Communications and Mobile Computing Conference (IWCMC). New York: IEEE, 2012: 463−468. DOI: 10.1109/IWCMC.2012.6314248.

[17] SONG J, ZHANG P Z, QIN S J, *et al*. A Method of the Feature Selection in Hierarchical Text Classification Based on the Category Discrimination and Position Information [C]//2015 International Conference on Industrial Informatics−Computing Technology, Intelligent Technology, Industrial Information Integration. New York: IEEE, 2015: 132−135. DOI: 10.1109/ICIICII.2015.116.

[18] TUO Q J, ZHAO H, HU Q H. Hierarchical Feature Selection with Subtree Based Graph Regularization[J]. *Knowl Based Syst*, 2019, **163**: 996−1008. DOI: 10.1016/j.knosys.2018.10.023.

[19] ZHAO H, HU Q H, ZHU P F, *et al*. A Recursive Regularization Based Feature Selection Framework for

Hierarchical Classification[J]. *IEEE Trans Knowl Data Eng*, 2021, **33**(7): 2833−2846. DOI: 10.1109/TKDE.2019.2960251.

[20] 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. DOI: 10.13328/j.cnki.jos.006335.
LIN Y J, BAI S X, ZHAO H, *et al*. Label-correlation-based Common and Specific Feature Selection for Hierarchical Classification[J]. *J Softw*, 2022, **33**(7): 2667−2682. DOI: 10.13328/j.cnki.jos.006335.

[21] SHI J, LI Z Y, ZHAO H. Feature Selection via Maximizing Inter-class Independence and Minimizing Intra-class Redundancy for Hierarchical Classification[J]. *Inf Sci*, 2023, **626**: 1−18. DOI: 10.1016/j.ins.2023.01.048.

[22] WU F H, ZHANG J, HONAVAR V. Learning Classifiers Using Hierarchically Structured Class Taxonomies [M]//ZUCKER J D, SAITTA L, eds. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 313−320. DOI: 10.1007/11527862_24.

[23] LI J D, CHENG K W, WANG S H, *et al*. Feature Selection[J]. *ACM Comput Surv*, 2018, **50**(6): 1−45. DOI: 10.1145/3136625.

[24] 刘浩阳, 林耀进, 刘景华, 等. 由粗到细的分层特征选择[J]. 电子学报, 2022, **50**(11): 2778−2789. DOI: 10.12263/DZXB.20211263.
LIU H Y, LIN Y J, LIU J H, *et al*. Hierarchical Feature Selection from Coarse to Fine[J]. *Acta Electron Sin*, 2022, **50**(11): 2778−2789. DOI: 10.12263/DZXB.20211263.

[25] LI X P, WANG Y D, RUIZ R. A Survey on Sparse Learning Models for Feature Selection[J]. *IEEE Trans Cybern*, 2022, **52**(3): 1642−1660. DOI: 10.1109/TCYB.2020.2982445.

[26] 史春雨, 毛煜, 刘浩阳, 等. 基于样本相关性的层次特征选择算法[J]. 山东大学学报(理学版), 2024, **59**(3): 61−70. DOI: 10.6040/j.issn.1671-9352.7.2023.1073.
SHI C Y, MAO Y, LIU H Y, *et al*. Hierarchical Feature Selection Algorithm Based on Instance Correlations[J]. *J Shandong Univ Nat Sci*, 2024, **59**(3): 61−70. DOI: 10.6040/j.issn.1671-9352.7.2023.1073.

[27] ARGYRIOU A, EVGENIOU T, PONTIL M. Convex Multi-task Feature Learning[J]. *Mach Learn*, 2008: 41−48. DOI: 10.1007/s10994-007-5040-8.

[28] DING C H Q, DUBCHAK I. Multi-class Protein Fold Recognition Using Support Vector Machines and Neural Networks[J]. *Bioinformatics*, 2001, **17**(4): 349−358. DOI: 10.1093/bioinformatics/17.4.349.

[29] LI D P, JU Y, ZOU Q. Protein Folds Prediction with Hierarchical Structured SVM[J]. *Curr Proteom*, 2016, **13**(2): 79−85. DOI: 10.2174/1570164613021605140009 40.

[30] DIMITROVSKI I, KOCEV D, LOSKOVSKA S, *et al*. Hierarchical Annotation of Medical Images[J]. *Pattern Recognit*, 2011, **44**(10/11): 2436−2449. DOI: 10.1016/j.patcog.2011.03.026.

[31] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, *et al*. The Pascal Visual Object Classes (VOC) Challenge[J]. *Int J Comput Vis*, 2010, **88**(2): 303−338. DOI: 10.1007/s11263-009-0275-4.

[32] JIA D, KRAUSE J, BERG A C, *et al*. Hedging your Bets: Optimizing Accuracy-specificity Trade-offs in Large Scale Visual Recognition[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2012: 3450−3457. DOI: 10.1109/CVPR.2012.6248086.

[33] DEKEL O, KESHET J, SINGER Y. Large Margin Hierarchical Classification[C]//Twenty−first international conference on Machine learning−ICML '04. New York: ACM, 2004. DOI: 10.1145/1015330.1015374.

[34] DEMSAR J. Statistical Comparisons of Classifiers over Multiple Data Sets[J]. *J Mach Learn Res*, 2006, **7**: 1−30.