计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2026/20(05)-1465-13
doi: 10.3778/j.issn.1673-9418.2507027

# 面向结构化特征选择的层次感知样本加权方法

张鑫茹[1],王一宾[1,2],程玉胜[1,2+]
1. 安庆师范大学 计算机与信息学院,安徽 安庆 246011
2. 安徽省高校智能感知与计算重点实验室(安庆师范大学),安徽 安庆 246011
+ 通信作者 E-mail: chengyshaq@163.com

**摘 要:**针对现有层次化特征选择中因类别语义重叠、样本分布失衡与跨层噪声干扰导致的特征评估偏差问题,提出层次感知样本加权方法(HASW),通过三类针对性策略重构数据分布,增强特征选择过程的稳健性与层次一致性。针对高层语义模糊性问题,设计动态聚类压缩策略(HASW-DCC),基于类内样本相似性动态调整聚类粒度,通过加权抑制冗余样本对细粒度特征学习的干扰;面向样本分布不均衡挑战,构建树感知元加权策略(HASW-TAMW),结合层次结构的元特征相关性优化样本权重分布,缓解模型对多数类的偏好偏差;针对跨层噪声破坏语义一致性的问题,提出层次离群双重过滤策略(HASW-HODF),联合全局与局部相关性阈值实现多粒度噪声过滤,增强父子节点的语义连续性。在6个层次数据集上,分别将3类策略集成至6种典型特征选择算法中进行对比实证。实验结果表明,HASW为层次化数据提供了即插即用的鲁棒预处理框架,有效提升特征选择稳定性与分类性能,同时强化了层次结构的语义一致性,为复杂层级数据建模提供了系统性优化策略。

**关键词:**特征选择;层次感知;样本加权;语义重叠;样本分布不平衡
**文献标志码:**A **中图分类号:**TP181

# Hierarchy-Aware Sample Weighting for Structured Feature Selection

ZHANG Xinru[1], WANG Yibin[1,2], CHENG Yusheng[1,2+]
1. College of Computer and Information, Anqing Normal University, Anqing, Anhui 246011, China
2. The University Key Laboratory of Intelligent Perception and Computing of Anhui Province (Anqing Normal University), Anqing, Anhui 246011, China

**Abstract:** Existing hierarchical feature selection methods suffer from biased feature importance evaluation due to semantic overlap between categories, unbalanced sample distribution, and cross-layer noise interference. To address these challenges, the hierarchy-aware sample weighting (HASW) method for structured feature selection is proposed. Through three targeted strategies, the data distribution is restructured to enhance the robustness and hierarchical consistency of the feature selection process. Firstly, the dynamic clustering compression strategy (HASW-DCC) is introduced to mitigate high-level semantic ambiguity. This strategy adjusts the clustering granularity based on intra-class sample similarity. It also applies weighting to suppress the interference of redundant samples on fine-grained feature learning. Secondly, the tree-aware meta-weighting strategy (HASW-TAMW) is developed to tackle sample distribution imbalance. Sample weight distribution is optimized based on the correlation of meta-features within the hierarchical structure. This reduces the bias towards the majority class. Lastly, the hierarchical outlier dual-filtering strategy (HASW-HODF) is proposed to reduce the impact of cross-layer noise on semantic consistency. This strategy uses both global and local correlation thresholds for multi-granularity

**基金项目:**安徽省教育厅重大项目(2024AH040175);安徽省教育厅重点项目(2024AH051099)。
This work was supported by the Major Project of Department of Education of Anhui Province (2024AH040175), and the Key Project of Department of Education of Anhui Province (2024AH051099).
**收稿日期:**2025-07-09 **修回日期:**2025-09-16

noise filtering, enhancing continuity between parent and child nodes. Three types of strategies are integrated into six feature selection algorithms and evaluated on six hierarchical datasets. Results demonstrate that HASW provides a robust, plug-and-play preprocessing framework for hierarchical data. It significantly improves feature selection stability, enhances classification performance, and maintains semantic consistency within the hierarchical structure. This approach offers a systematic optimization strategy for modeling complex hierarchical data.

**Key words:** feature selection; hierarchy-aware; sample weighting; semantic overlap; imbalanced sample distribution

随着数据规模的增长，特征选择对提升模型的泛化能力和降低复杂度变得愈加重要[1]。传统方法假设数据扁平且类别独立[2]，难以处理生物、图像等领域的层次结构数据[3]，而分层特征选择能利用类别间内在关联提升分类效果。例如，Li等人[4]提出了一种半监督局部特征选择方法，针对每个类的特征重要性进行学习，从而为每个类选择最具辨识力的特征子集。Zhao等人[5]提出的Hier-FS（hierarchical feature selection）算法通过层次结构选择特征子集，从而提升分类性能。

然而，这些方法仅在基本层面上处理了分层结构，未能有效应对层次化分类中的复杂挑战。例如，Zhao等人[5]提出的HiRRfam-FS（hierarchical recursive regularization framework for feature selection）方法利用递归正则化框架优化特征选择，考虑父子节点和兄弟节点关系，从而缓解冗余特征的影响。该方法主要依赖于层次结构关系，未能有效解决类别语义重叠和样本分布不均衡问题。Lin等人[6]提出的HFSLDL（hierarchical feature selection with local label distribution learning）方法，通过标签增强，缓解了语义缺口问题，减少了高层次分类中的误差传播。该方法在处理类别语义重叠和样本分布不均衡时仍存在不足。Shi等人[7]提出的HFS-MIMR（hierarchical feature selection via maximizing inter-class independence and minimizing intra-class redundancy）方法，通过最大化类间独立性和最小化类内冗余，利用结构与特征关系有效减少类别语义重叠，从而提高分类性能。Liu等人[8]提出的HFSLE（hierarchical feature selection with label enhancement）方法，通过标签增强和父子节点一致性约束，缓解语义缺口问题，提升分类精度。该方法在处理层次间噪声和跨层次特征关系时仍面临挑战。Lin等人[9]提出的HFSNIS（hierarchical feature selection via neighbor interaction search）算法采用粗粒度搜索策略改进高维数据集中的层次特征选择，解决了细粒度层次中丢失重要特征的问题，优化特征子集选择。Wang等人[10]提出的GLUFS-HSC（global-local unified feature selection with hierarchical structure constraints）算法通过全局-局部统一优化框架与节点间约束机制，进一步增强了层次一致性。然而，该算法的优化过程仍直接作用于原始数据分布，未能显式处理样本层面的分布失衡与噪声扰动，而这是造成特征评估出现偏差的关键因素。例如，其父子节点双向约束的性能会因多数类样本主导或跨层噪声干扰而显著下降，难以充分发挥约束作用。

尽管这些方法在层次化特征选择中取得了一定进展，但仍需进一步优化，以应对类别语义重叠[11]、样本分布不平衡[12]和跨层噪声[13]等问题。本文提出了层次感知样本加权方法（hierarchy-aware sample weighting，HASW），通过三种策略优化数据分布：动态聚类压缩抑制冗余特征，树感知元加权增强层次语义约束，层次离群双重过滤提升选择准确性。实验结果表明，HASW方法显著提高了层次分类准确率，减少了特征冗余，并减轻了噪声干扰。

## 1 相关工作

在层次分类任务[14]中，输入数据的质量直接决定特征选择的鲁棒性与分类器的泛化性能[15]。然而，现有预处理方法（如标准化、降维等）通常基于样本的低阶统计特性[16]，缺乏对类间层级结构及语义依赖关系的建模能力[17]。当面临语义重叠、样本不均衡以及跨层噪声耦合等复杂干扰时，这类方法难以维持稳定性与判别性，呈现出显著的性能瓶颈。本章将通过具体实例，分析现有预处理策略在复杂层次语义场景下的适应性局限。

### 1.1 全局特征主导下的细粒度语义重叠问题

在图像分类任务中，传统预处理方法多侧重于提取图像的全局特征，以捕捉其宏观语义。然而，这种全局特征的主导性往往掩盖了细粒度特征的作用。由于细粒度特征在统计上方差较小，常在降维或标准化过程中被忽略[18]，导致模型对类别间微小但关键的差异缺乏感知能力。进一步而言，不同类别样本在全局特征空间中可能呈现高度相似性，从而使模型难以识别它们的本质差异。这种特征表达上的模糊性加剧了语义重叠，进而显著削弱了模型在复杂分类任务中的判别能力。全局特征虽能提供粗粒度的类别信息，但真正定义类别边界的，往往是细粒度的局部特征。如果二者在特征空间中未被有效解耦，模型便容易混淆不同语义层级，从而影响最终分类性能。

如图1上图所示，猫（蓝）与狗（橙）在毛色和体型等全局属性上的特征区间高度重叠，形成密集混合区域，导致模型难以正确区分二者。图1下图则显示出猫爪、狗嗅觉等细粒度特征对于类别区分的重要性。然而，当全局特征占主导地位时，这些关键的细节信息往往被忽略。因此，应采用能有效解耦全局与细粒度特征的模型，以提升分类判别能力，更全面地捕捉类别之间的本质差异。

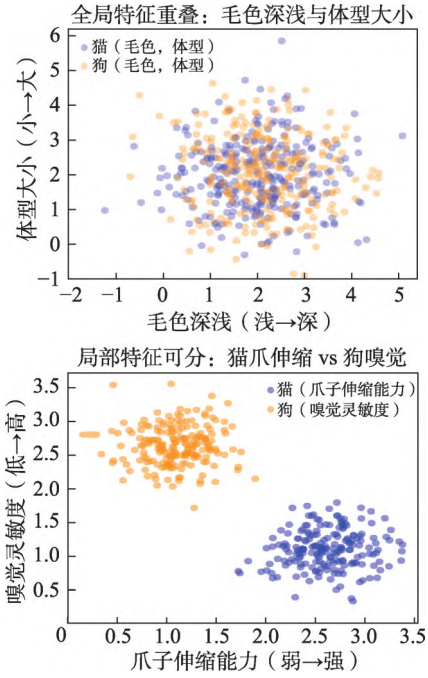



**图1 语义重叠：全局特征主导下细粒度特征被忽略**

Fig.1 Semantic overlap: fine-grained features neglected under dominance of global features

## 1.2 多数类样本主导下的特征抑制问题

在类别分布不平衡的场景中，多数类样本由于其数量优势，往往在特征空间中占据主导地位。这种主导地位不仅导致特征分布呈现中心化偏移，还使得特征选择过程更加倾向于保留多数类的主导特征，从而忽视对少数类具有判别性的局部特征结构[19]。理论上，当多数类样本密集分布于特征空间的核心区域时，其所占据的高频统计模式在特征评估指标中具有更高的显著性评分[20]。而少数类样本由于数量稀少，往往位于特征空间的边缘区域，其特征分布呈现出局部稀疏、高离散性的性质，难以在全局特征排序中获得足够的评分。这种结构性不对称会导致特征选择过程中出现“特征抑制效应”，即少数类的关键特征被误判为异常值或噪声而被剔除，进而影响其在分类决策中的贡献。图2可视化了该现象，橙色区域为多数类样本在“独立性”与“运动需求”两个维度上的密集分布，形成特征空间的核心。紫色点代表的

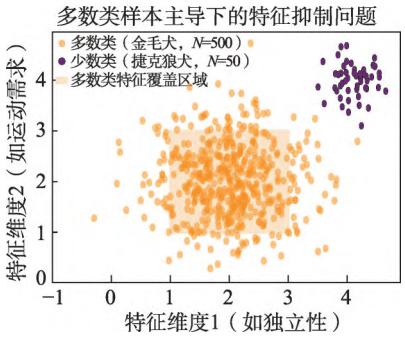



**图2 样本分布不平衡：多数类主导特征选择**

Fig.2 Sample distribution imbalance: majority class dominates feature selection

少数类捷克狼犬，其分布显著偏离中心区域，且局部样本密度较低。

在此背景下，传统特征选择方法难以准确识别这些区域所蕴含的局部判别信息，容易将其误识为离群值而忽略，从而导致与少数类高度相关的特征未被纳入最终特征子集。这不仅削弱了模型对少数类的识别能力，也加剧了训练阶段损失函数与决策边界对多数类的偏向。因此，如何从特征分布层面构建对少数类结构更敏感的特征评估机制，成为提升分类模型性能的关键问题之一。

## 1.3 跨层噪声引发层次结构断裂与特征混淆

在层次化分类任务中，噪声样本和层次结构的不一致会严重影响模型的分类性能和层次结构的完整性[21]。理论上，父节点的特征应由所有子节点的关键属性共同决定。然而，当噪声样本被错误归类至不匹配的子节点时，这些异常样本会改变父节点的特征空间，使其偏离该类别的本质特征[22]。同时，同一父节点下的兄弟节点可能因噪声干扰而被迫共享非判别性特征，从而弱化兄弟节点之间的本质差异。这种特征偏离现象在层次结构中具有累积效应。父节点特征的变化可能进一步影响其子节点之间的区分能力，导致模型错误地将非关键特征视为分类依据。例如，图3展示了海豚（哺乳动物）错误归类为“鲨鱼”（鱼类）的子节点的情况。这一错误使得父节点“鱼类”被赋予了哺乳动物的属性（如肺呼吸），与其实际类别定义不符。

海豚与鲨鱼因错误分类被模型视为兄弟节点，模型可能倾向于强化二者共享的非判别性特征（如流线型外形），而忽略关键属性的本质差异（如呼吸方式）。这一现象不仅削弱了模型在局部分类任务中的可靠性，还通过层次结构逐步扩展，导致整体分类逻辑的不稳定。因此，识别并抑制噪声干扰，同时确保层次结构连贯性，是提升模型性能和可解释性的关键。

针对上述局限，本文提出层次感知样本加权方法

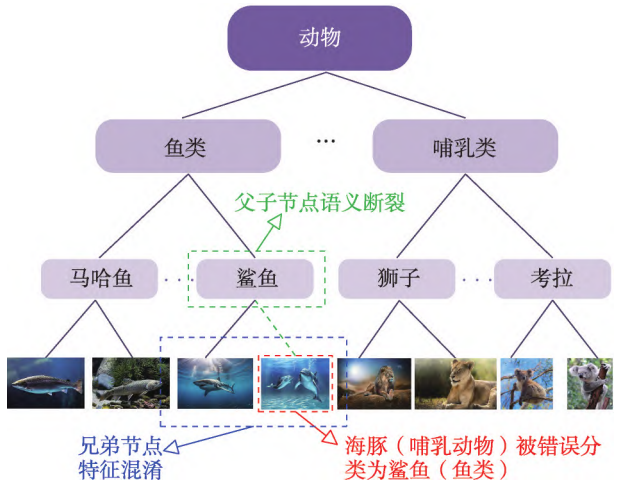



**图3 复杂层次结构中的跨层噪声:语义断裂与特征混淆**

Fig.3 Cross-layer noise in complex hierarchical structures: semantic discontinuity and feature confusion

(HASW),通过三个核心策略优化特征选择过程,将动态样本权重调整与层次结构结合,以提升在复杂层次数据场景下的特征选择准确性与分类性能。

## 2 层次感知样本加权在结构化特征选择中的优化策略

本章介绍层次感知样本加权方法(HASW),其总体框架如图4所示,核心包含动态聚类压缩(dynamic clustering compression strategy,HASW-DCC)、树感知元加权(tree-aware meta-weighting strategy,HASW-TAMW)与层次离群双重过滤(hierarchical outlier dual-filtering strategy,HASW-HODF)三种策略在结构化特征选择中的应用,下文详细阐述每种策略的基本模型,并提供算法伪代码,同时探讨三种核心策略与特征选择的整合机制。

### 2.1 动态聚类压缩与冗余抑制(HASW-DCC)

在层次化特征学习过程中,高层语义的模糊性可能导致类内特征分布差异显著,噪声样本带来梯度干扰,从而影响模型对判别性特征的建模。为此,本策略引入动态聚类机制,通过结合相关性加权,保留高代表性样本,优化输入数据质量,并抑制冗余样本的干扰。

#### 2.1.1 数学模型

(1)动态聚类数量计算

设原始样本集为 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,其中 $N$ 为原始样本数。动态聚类树 $k$ 由 $k_s\in(0,1)$ 控制聚类粒度,以平衡聚类后的样本密度,避免过拟合:

$$k=\text{round}(N\times k_s) \tag{1}$$

通过交叉验证选择为0.6或0.8,其中0.6适用于高冗余数据。

(2)K-means聚类优化

目标是最小化类内样本与聚类中心的欧氏距离平方和,公式为:

$$\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k\}=\arg\min_{\boldsymbol{\mu}_i}\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i}\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2 \tag{2}$$

式中,$C_i$ 为第 $i$ 聚类簇,$\boldsymbol{\mu}_i$ 为第 $i$ 个簇的中心向量。通过迭代优化将样本划分为 $k$ 个类,每个类中心 $\boldsymbol{\mu}_i$ 代表局部特征分布。

(3)样本相关性加权与截断

皮尔逊相关系数 $R_{ij}$ 对特征尺度不敏感,适用于多模态数据,用于衡量样本 $\boldsymbol{x}_j$ 与其所属类中心 $\boldsymbol{\mu}_i$ 的线性相关性:

$$R_{ij}=\frac{\text{cov}(\boldsymbol{x}_j,\boldsymbol{\mu}_i)}{\sigma_{\boldsymbol{x}_j}\sigma_{\boldsymbol{\mu}_i}} \tag{3}$$

动态阈值截断,$\theta_{\text{dec}}=\text{profile}(R,60\%)$,以所有样本与所属簇中心皮尔逊相关系数的60%分位数作为阈值。

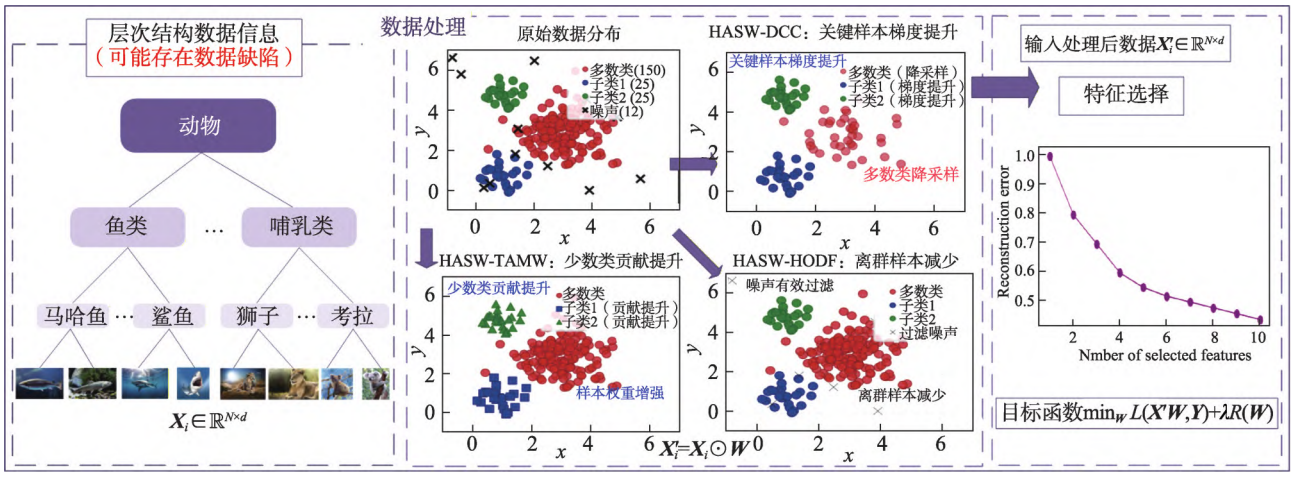



**图4 HASW方法框架**

Fig.4 Framework of HASW method

此阈值设定遵循统计假设检验的基本原理，将相关性分布于尾部的样本视为可能偏离主分布的噪声或异常值。在经过 $k_s$ 压缩得到的代表性更强的样本子集上执行此操作，能以较高的置信度保留蕴含类别判别信息的核心样本，显著增强特征学习过程的鲁棒性。$\boldsymbol{R}$ 为所有样本的相关系数矩阵，精准分离高低相关样本。引入权重因子 $w_j$ 进行权重赋值，具体定义如下：

$$w_j=\begin{cases}\dfrac{R_{ij}}{\max(\boldsymbol{R})}, R_{ij}\geqslant\theta_{\text{dec}}\\ 0.01, R_{ij}<\theta_{\text{dec}}\end{cases} \tag{4}$$

其用于调节不同样本在后续计算中的贡献大小，进一步突出关键样本的特征表达能力，该权重机制依据皮尔逊相关系数 $R_{ij}$ 分配样本重要性：相关性高于阈值 $\theta_{\text{dec}}$ 的样本将获得归一化权重，从而增强其在特征聚合过程中的影响；而低相关性的样本被赋予极小权重，以抑制噪声干扰。权重机制使关键样本梯度幅度提升，显著增强特征表达。

### 2.1.2 算法伪代码

**算法 1** 动态聚类压缩算法

输入：原始样本集 $X$，压缩比例 $k_s$。

输出：加权后的代表性样本集 $X'$。

**步骤 1** 动态聚类，根据式(1)计算聚类数 $k$，并对样本集 $X$ 进行聚类，得到各聚类中心 $\boldsymbol{\mu}_i$。

**步骤 2** 根据式(3)计算每个样本 $\boldsymbol{x}_j$ 与所属类中心 $\boldsymbol{\mu}_i$ 的皮尔逊相关系数 $R_{ij}$。

**步骤 3** 动态截断，引入权重因子 $w_j$ 进行权重赋值，保留 $R_{ij}\geqslant\theta_{\text{dec}}$ 样本，其余赋低权重。

**步骤 4** 特征缩放，对每个样本 $\boldsymbol{x}_j$（第 $j$ 个样本）的特征值乘以权重 $w_j$，即 $\boldsymbol{x}_j'=\boldsymbol{x}_j\times w_j$。

**步骤 5** 返回加权后的代表性样本集 $X'$。

## 2.2 树感知元加权(HASW-TAMW)

在样本分布不均衡的情况下，模型易偏向多数类，导致决策边界偏移。这种偏移还表现为类间特征的区分度降低，导致混淆样本增多，影响模型的准确性。为了解决这一问题，本策略引入层次结构信息，并结合元特征加权来平衡样本的贡献，从而缓解这些不均衡问题。

### 2.2.1 数学模型

(1)元特征生成

设父节点样本集为 $X_p$，$C=\{C_1,C_2,\cdots,C_K\}$ 为其子节点集合，首先计算每个子节点 $k\in C$ 的均值向量 $\boldsymbol{\mu}_k$，以反映其局部特征分布：

$$\boldsymbol{\mu}_k=\frac{1}{|C_k|}\sum_{\boldsymbol{x}\in C_k}\boldsymbol{x}, k=1,2,\cdots,K \tag{5}$$

式中，$C_k$ 为父节点的第 $k$ 个子节点样本集，反映子节点局部特征分布。

为缓解类别规模差异带来的偏差，采用基于样本量的加权平均策略计算全局元特征 $\boldsymbol{\mu}_m$，其计算公式为：

$$\boldsymbol{\mu}_m=\frac{1}{|X_p|}\sum_{k\in C}|C_k|\times\boldsymbol{\mu}_k \tag{6}$$

该加权策略确保了样本规模更大的子类别对全局元特征的形成贡献更大，从而更稳健地捕捉层次结构的全局模式。

(2)动态相关性阈值生成

为解决固定阈值在层次结构动态变化与极端不平衡场景下的适应性问题，本文提出一种自适应的动态阈值机制。该阈值 $\theta_{\text{meta}}$ 由基础值、不平衡调整因子和深度调整因子共同决定，其计算公式如下：

$$\theta_{\text{meta}}^{(i)}=\theta_{\text{base}}\times\alpha^{(i)}\times\delta^{(i)} \tag{7}$$

其中，$\theta_{\text{base}}$ 为基础阈值参数，经实验验证取值为0.4。$\alpha^{(i)}$ 为第 $i$ 个节点的不平衡调整因子，用于应对节点内类别分布失衡。假设该节点所有子节点中样本数量最多和最少的类别样本数分别为 $N_{\max}^{(i)}$ 和 $N_{\min}^{(i)}$，则不平衡比率 $IR^{(i)}=N_{\max}^{(i)}/N_{\min}^{(i)}$。$\alpha^{(i)}$ 的计算公式为：

$$\alpha^{(i)}=\begin{cases}1+\dfrac{\ln IR^{(i)}}{10}, IR^{(i)}>\gamma\\ 1, IR^{(i)}\leqslant\gamma\end{cases} \tag{8}$$

其中，$\gamma$ 为判定极端不平衡的阈值，取 $\gamma=50$。$\delta^{(i)}$ 为层次深度调整因子，$d^{(i)}$ 为当前节点 $i$ 在树中的深度，$D$ 为树的总深度。深层节点语义更细粒度，样本更稀疏，需放宽阈值以保留更多判别性信息，其计算公式为：

$$\delta^{(i)}=1-0.2\times\frac{d^{(i)}}{D} \tag{9}$$

(3)自适应相关性加权

计算样本 $\boldsymbol{x}_j$ 与全局元特征 $\boldsymbol{\mu}_m$ 的皮尔逊相关系数 $corr_j$：

$$corr_j=\frac{\operatorname{cov}(\boldsymbol{x}_j,\boldsymbol{\mu}_m)}{\sigma_{\boldsymbol{x}_j}\sigma_{\boldsymbol{\mu}_m}} \tag{10}$$

该系数用于衡量样本 $\boldsymbol{x}_j$ 与全局元特征 $\boldsymbol{\mu}_m$ 之间的一致性，高相关性样本对全局模式贡献更大。使用动态阈值 $\theta_{\text{meta}}^{(i)}$ 进行加权：

$$w_j'=\begin{cases}1, corr_j\geqslant\theta_{\text{meta}}^{(i)}\\ \varepsilon\times\left(\dfrac{corr_j}{\theta_{\text{meta}}^{(i)}}\right)^2, corr_j<\theta_{\text{meta}}^{(i)}\end{cases} \tag{11}$$

其中，$\varepsilon=0.01$ 用于抑制低相关性样本的贡献。该加权策略对略低于阈值的样本进行平滑过渡，有效缓解极端不平衡数据集上的过度过滤问题，增强策略的层次适应性与鲁棒性。

### 2.2.2 算法伪代码

**算法 2** 树感知元加权算法

输入：父节点 $X_p$ 及其子节点样本集 $\{C_1,C_2,\cdots,C_K\}$，层次树结构 $T$。

输出:加权后的代表性样本集 $X''$。

**步骤1** 根据式(5)计算子节点均值 $\boldsymbol{\mu}_k$。

**步骤2** 根据式(6)生成全局元特征 $\boldsymbol{\mu}_m$。

**步骤3** 获取节点 $X_p$ 在树 $T$ 中的深度 $d$ 及树 $T$ 的总深度 $D$。

**步骤4** 计算节点下的不平衡比率 $IR$。

**步骤5** 根据式(8)计算不平衡调整因子 $\alpha$。

**步骤6** 根据式(9)计算深度调整因子 $\delta$。

**步骤7** 根据式(7)计算动态阈值 $\theta_{\text{meta}}$。

**步骤8** 计算样本 $\boldsymbol{x}_j'$ 与 $\boldsymbol{\mu}_m$ 的相关性。

**步骤9** 使用动态阈值 $\theta_{\text{meta}}$ 进行加权计算 $w_j'$。

**步骤10** 对每个样本 $\boldsymbol{x}_j'$(第 $j$ 个样本)的所有特征值乘以 $w_j'$ 进行特征缩放:$\boldsymbol{x}_j''=\boldsymbol{x}_j'\times w_j'$。

**步骤11** 返回加权后的代表性样本集 $X''$。

## 2.3 层次离群双重过滤(HASW-HODF)

跨层噪声会破坏语义连续性,表现为异常样本引起父子节点特征空间断裂,局部噪声进一步传播至全局层次结构。此外,单一阈值过滤方法难以兼顾全局分布与局部语义。本策略通过双重相关性截断,结合全局均值和类内均值,实现了多粒度的噪声过滤,有效缓解了上述问题。

### 2.3.1 数学模型

(1)均值计算

全局均值 $\boldsymbol{\mu}_g$ 用来衡量样本与类内分布的协调程度,反映数据整体分布,捕获类内分布模式。

$$\boldsymbol{\mu}_g=\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}_j \tag{12}$$

局部均值(类内均值)$\boldsymbol{\mu}_l$ 为当前类别 $C$ 的均值向量,针对当前类别优化局部特征,确保样本符合层次语义结构,保证层次语义连贯。

$$\boldsymbol{\mu}_l=\frac{1}{|X_C|}\sum_{\boldsymbol{x}\in C}\boldsymbol{x} \tag{13}$$

其中,$X_C$ 为类别 $C$ 的样本集,$|X_C|$ 为样本数量。

(2)双重相关性截断

样本 $\boldsymbol{x}_j$ 与全局均值 $\boldsymbol{\mu}_g$ 的皮尔逊相关系数:

$$R_{jg}=\frac{\operatorname{cov}(\boldsymbol{x}_j,\boldsymbol{\mu}_g)}{\sigma_{\boldsymbol{x}_j}\sigma_{\boldsymbol{\mu}_g}} \tag{14}$$

样本 $\boldsymbol{x}_j$ 与类内均值 $\boldsymbol{\mu}_l$ 的皮尔逊相关系数:

$$R_{jl}=\frac{\operatorname{cov}(\boldsymbol{x}_j,\boldsymbol{\mu}_l)}{\sigma_{\boldsymbol{x}_j}\sigma_{\boldsymbol{\mu}_l}} \tag{15}$$

样本 $\boldsymbol{x}_j$ 需同时满足与全局分布和局部分布的一致性,其筛选条件为:

$$S'=\{\boldsymbol{x}_j\mid R_{jg}\geqslant\theta_{\text{global}}\wedge R_{jl}\geqslant\theta_{\text{local}}\} \tag{16}$$

通过实验验证,选取全局离群阈值 $\theta_{\text{global}}=0.3$,避免跨层噪声,保留高协调性样本;$\theta_{\text{local}}=0.5$ 为局部离群阈值,防止过度过滤,确保语义连续性。平衡严格性与覆盖率,联合阈值实现粗-细粒度过滤。

(3)特征缩放

对保留的高质量样本集 $S'$ 重新计算截断后均值,以强化语义一致性:

$$\boldsymbol{\mu}_{\text{true}}=\frac{1}{|S'|}\sum_{\boldsymbol{x}\in S'}\boldsymbol{x} \tag{17}$$

为进一步强化高相关性样本、抑制离群点影响,对保留样本进行特征缩放。

$$\boldsymbol{x}_j'=\boldsymbol{x}_j\times\frac{\operatorname{corr}(\boldsymbol{x}_j,\boldsymbol{\mu}_{\text{true}})}{\max(\operatorname{corr}(\boldsymbol{x}_j,\boldsymbol{\mu}_{\text{true}}))} \tag{18}$$

其中,高相关性样本的特征值被相应放大,低相关性样本的特征值被抑制。

### 2.3.2 阈值协同优化理论

在层次离群双重过滤策略(HASW-HODF)的理论框架中,全局阈值 $\theta_{\text{global}}$ 与局部阈值 $\theta_{\text{local}}$ 的协同作用可通过联合概率密度严格建模。给定样本 $\boldsymbol{x}_j$,设其全局置信度为 $\rho_g(\boldsymbol{x}_j)$,局部置信度为 $\rho_l(\boldsymbol{x}_j)$,其保留概率模型为:

$$P(\boldsymbol{x}_j)=\exp\left\{-\frac{(\rho_g(\boldsymbol{x}_j)-\boldsymbol{\mu}_g)^2}{2\sigma_g^2}\right\}\times\left[1-\operatorname{erf}\left(\frac{\rho_l(\boldsymbol{x}_j)-\boldsymbol{\mu}_l}{\sigma_l\sqrt{2}}\right)\right] \tag{19}$$

其中,$\sigma_g$ 为全局置信度标准差,反映跨类别预测波动性;$\sigma_l$ 为类内置信度标准差均值,表征类内一致性;$\boldsymbol{\mu}_g=0.3$,$\boldsymbol{\mu}_l=0.5$ 为经验最优中心值:

$$\sigma_g=\sqrt{\frac{1}{N}\sum_{j=1}^{N}(\rho_g(\boldsymbol{x}_j)-0.3)^2} \tag{20}$$

$$\sigma_l=\frac{1}{C}\sum_{c=1}^{C}\sqrt{\frac{1}{|c|}\sum_{\boldsymbol{x}_j\in c}(\rho_l(\boldsymbol{x}_j)-0.5)^2} \tag{21}$$

阈值关系的最优性证明基于统计理论推导。通过 Fisher-Z 变换建立方差分解方程:

$$\Delta\theta=z_{1-\alpha/2}\times\sqrt{\frac{1}{n_c-3}-\frac{1}{N-3}} \tag{22}$$

该式揭示了全局与局部阈值的理论差值仅依赖于样本量,与数据分布无关。协同效应的存在性由二阶偏导数判定:

$$\frac{\partial^2 P}{\partial\theta_g\partial\theta_l}>0 \tag{23}$$

表明双重阈值存在正向交互作用。精度曲面的极值点满足:

$$\nabla P(\theta_g,\theta_l)=0\rightarrow\theta_l=k\theta_g \tag{24}$$

其中,比例系数 $k=1.67$ 为理论推导值。

### 2.3.3 算法伪代码

**算法3** 层次离群双重过滤算法

输入:某类别样本集 $X_C$。

输出:过滤后的样本集 $X_C'$。

**步骤1** 根据式(12)计算全局均值 $\boldsymbol{\mu}_g$。

**步骤2** 基于当前类别标签计算类内均值 $\boldsymbol{\mu}_l$。

**步骤3** 根据式(14)和式(15)计算每个样本 $\boldsymbol{x}_j$ 与 $\boldsymbol{\mu}_g$ 和 $\boldsymbol{\mu}_l$ 的双重相关系数。

**步骤4** 双重相关性筛选，保留同时满足式(16)条件的样本。

**步骤5** 根据式(17)计算截断后均值 $\boldsymbol{\mu}_{\text{true}}$，并根据式(18)进行特征缩放。

**步骤6** 返回过滤后的样本集 $X_C'$。

## 2.4 层次感知样本加权策略与特征选择算法整合机制

为了从理论层面验证策略有效性，将3类策略集成至6种典型特征选择算法中。本节以HFS-MIMR算法为例进行分析，该算法核心公式为：

$$J(\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N)=\sum_{i=0}^{N}(\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_{\text{F}}^2+\lambda\|\boldsymbol{W}_i\|_{2,1}^2+$$
$$2\beta(\|\boldsymbol{W}_i\boldsymbol{W}_i^{\text{T}}\|_1-\|\boldsymbol{W}_i\|_2^2))+\alpha\sum_{i=1}^{N}\sum_{j\in S_i}(\|\boldsymbol{W}_j^{\text{T}}\boldsymbol{W}_i-\boldsymbol{E}_m\|_{\text{F}}^2) \quad (25)$$

其中，$\boldsymbol{W}_i\in\mathbf{R}^{d\times m}$ 表示第 $i$ 个非叶节点的特征权重矩阵，$d$ 为特征维度，$m$ 为当前节点的类别数；单位矩阵为 $\boldsymbol{E}_m\in\mathbf{R}^{m\times m}$，用于约束兄弟节点权重的正交性。$S_i$ 为第 $i$ 个非叶节点的兄弟节点集。目标函数包含四项：数据拟合项，用于保证特征权重与标签的一致性；$L_{2,1}$ 范数约束实现特征级稀疏；$\beta$ 项控制权重矩阵的平滑性；$\alpha$ 项通过正交约束增强兄弟节点间的区分性。$\|\cdot\|_{\text{F}}^2$ 表示范数，$\|\cdot\|_{2,1}$ 为行稀疏范数。

(1)HASW-DCC(动态聚类压缩)

该策略通过降低数据冗余，减少噪声对数据拟合项的干扰。原始数据矩阵 $\boldsymbol{X}_i\in\mathbf{R}^{N\times d}$ 通过剔除冗余样本与低相关性样本，得到加权后的代表性样本集 $\boldsymbol{X}_i'\in\mathbf{R}^{N\times d}$，其中压缩比例为 $k_s\in(0,1)$。经此处理，数据拟合项因样本冗余度降低而获得更小的拟合误差；稀疏正则项因特征区分度增强，更容易满足稀疏性约束；同时，算法计算复杂度由 $O(Tk_sN^2d)$ 降至 $O(Tk_s^2N^2d)$。

(2)HASW-TAMW(树感知元加权)

通过元特征 $\boldsymbol{\mu}_m$ 加权 $\boldsymbol{X}_i'=\boldsymbol{X}_i\boldsymbol{W}_{\text{meta}}$，使少数类关键样本权重提升，多数类样本权重抑制。原始数据拟合项受到多数类主导，经加权处理后实现均衡化，提升了少数类的贡献。$\|\boldsymbol{W}_j^{\text{T}}\boldsymbol{W}_i-\boldsymbol{E}_m\|_{\text{F}}^2$ 层次一致性项因兄弟节点特征分布对齐，差异减少；结构约束项因权重矩阵平滑性提升，协方差更稳定。

(3)HASW-HODF(层次离群双重过滤)

该策略通过过滤全局与局部噪声，提升特征选择的鲁棒性。保留满足式(16)的样本，剔除全局离群点和局部离群点后，根据式(18)进行特征缩放，生成过滤后的高质量样本集 $\boldsymbol{X}_j'$。数据拟合项因噪声样本被剔除，误差项显著降低；稀疏正则项因特征区分度增强，非判别性特征的权重趋近于零；离群样本的减少使得层次一致性项中的正交约束更易满足。

# 3 实验设置

本章将从实验数据集、评价指标以及实验参数设置几个方面描述实验设置。

## 3.1 实验数据集描述

实验中使用了6个具有层次结构的公共数据集，包含蛋白质[23]和图像数据集[24-27]。这些数据集都是单标签数据，所有样本都被分配到叶节点上。表1列出了这些数据集的介绍。

## 3.2 评价指标

采用层次分类指标和平面分类指标对各方法的有效性进行评价。

(1)基于最小共同祖先的 $F1$ 方法 ($F_{\text{LCA}}$)[28]：衡量层级分类中预测标签与真实标签语义接近程度的指标，具有对共享祖先类别数量不敏感的特性。

(2)分类精度 ($Acc$)[29]：正确分类的样本数除以所有测试样本。

## 3.3 参数设置

动态聚类压缩策略包含两个核心参数：

(1)压缩比例 $k_s$，控制每个聚类中保留的样本比例，典型值范围为 $0.5\leqslant k_s\leqslant 0.9$。增大 $k_s$ 会保留更多样本，但可能引入冗余，计算开销增加。减小 $k_s$ 会降低计算复杂度，但可能丢失重要样本，导致特征选择偏差。该参数范围的设定遵循机器学习中经典的偏差-方差权衡准则。下限0.5由统计学习理论中的采样有效性决定，旨在确保聚类中心估计的稳定性，防止因样本过少

**表1 具有层次结构的6个实验数据集描述**

Table 1 Description of 6 experimental datasets with hierarchical structure

| 数据集 | 领域 | 特征数 | 层次深度 | 训练集 | 测试集 | 类别数 | 稀疏度 | 特征相关性 | 不平衡比 |
|---|---|---|---|---|---|---|---|---|---|
| Bridges | 图像 | 12 | 3 | 86 | 79 | 6 | 0.21 | 0.24 | 4.4 |
| CLEF | 图像 | 80 | 4 | 8 368 | 939 | 63 | 0.00 | 0.14 | 383.6 |
| DD | 蛋白质 | 473 | 3 | 3 020 | 605 | 27 | 0.64 | 0.12 | 19.5 |
| VOC | 图像 | 1 000 | 5 | 7 178 | 5 105 | 20 | 0.06 | 0.20 | 13.7 |
| Car196 | 图像 | 4 096 | 3 | 8 144 | 7 541 | 196 | 0.78 | 0.06 | 2.8 |
| ILSVRC57 | 图像 | 4 096 | 4 | 12 346 | 11 845 | 57 | 0.73 | 0.07 | 1.4 |

而导致估计方差过大;上限0.9则保证了预处理的意义,在保留绝大部分原始信息的同时减少计算成本。针对样本高度冗余的场景,经实验验证最优值为 $k_s=0.6$。在如蛋白质折叠(DD)等高稀疏度数据中,类别内样本间常因功能或结构相似性而呈现出高度的分布聚集性,即高样本冗余。此设置能有效压缩冗余样本,突显类别核心分布模式。

(2)相关性阈值 $\theta_{dec}$ 用于判定样本间的结构相似度,提升压缩后样本的代表性,抑制潜在噪声干扰,对蛋白质类等高稀疏、低相关性数据尤为关键。高稀疏性是数据结构特性,而低特征相关性表明特征间独立性较好;同时,其类别内常因功能收敛性存在大量相似样本(即高样本冗余),这使得抑制冗余的收益非常显著。对于样本分布本就均匀、冗余度低的数据集,该策略的增益会相应减弱,这与4.2节消融实验在不同数据集上的结论相互印证。

树感知元加权策略的核心参数:元特征相关性阈值 $\theta_{meta}$,截断与元特征相关性低的样本,仅保留高相关性样本参与特征选择。增大 $\theta_{meta}$ 会强化对高一致性样本的偏好,增强类内判别性,但可能忽略低频重要特征;减小 $\theta_{meta}$ 可增加样本包容性,但可能引入噪声。

层次离群双重过滤策略采用 $\theta_{global}$ 全局相关性阈值和局部相关性阈值 $\theta_{local}$ 控制机制。$\theta_{global}$ 用于跨层级噪声的初步剔除,$\theta_{local}$ 则侧重于在类内进一步保证结构一致性。增大 $\theta_{global}$ 或 $\theta_{local}$ 会使过滤更严格,数据纯净度提高,但可能过度过滤;减小 $\theta_{global}$ 或 $\theta_{local}$ 会保留更多样本,但噪声可能增加。

## 4 实验结果及分析

本章将HASW方法中的3种策略集成到典型特征选择算法中,并在多个领域的数据集上开展对比实验与结果分析。此外,本文还在3类具有代表性的数据集上对基线方法与集成方法进行了消融实验,分析方法中关键参数的敏感性,最后对算法的收敛性进行了评估。

### 4.1 对比实验及结果分析

为了评估3种策略在不同领域数据集上与不同算法整合机制的性能,本实验在6个层次数据集上,分别将3类策略集成至6种典型特征选择算法,通过对比实验及评价指标对各方法的性能表现进行实证分析。具体性能表现见表2、表3。实验结果表明,HASW-TAMW在CLEF极端不平衡数据集上表现最优,在深层级数据

**表2 3类策略集成至6种分层特征选择方法的 *Acc* 结果**

Table 2 *Acc* results of integrating 3 types of strategies into 6 hierarchical feature selection methods

| 算法 | *Acc*(排名) | | | | | | 平均排名 |
|---|---|---|---|---|---|---|---|
| | Bridges | DD | CLEF | VOC | Car196 | ILSVRC57 | |
| Hier-FS | 0.508 9(4) | 0.686 1(4) | 0.602 8(4) | 0.420 0(3) | 0.645 9(4) | 0.849 2(4) | 3.83 |
| HASW-DCC | 0.518 9(3) | **0.688 2(1)** | 0.608 0(3) | 0.415 0(4) | **0.663 0(1)** | 0.849 8(3) | 2.50 |
| HASW-TAMW | **0.543 9(1)** | 0.688 0(2) | **0.620 2(1)** | **0.428 0(1)** | 0.656 0(2) | **0.852 0(1)** | **1.33** |
| HASW-HODF | 0.528 9(2) | 0.687 5(3) | 0.612 0(2) | **0.428 0(1)** | 0.653 0(3) | 0.849 9(2) | 2.17 |
| HiRRfam-FS | 0.508 9(4) | 0.689 4(4) | 0.606 0(4) | 0.422 9(3) | 0.647 5(4) | 0.850 6(4) | 3.83 |
| HASW-DCC | 0.513 9(3) | **0.691 2(1)** | 0.610 5(3) | 0.418 0(4) | **0.665 0(1)** | 0.851 0(3) | 2.50 |
| HASW-TAMW | **0.538 9(1)** | 0.691 0(2) | **0.623 0(1)** | 0.429 0(2) | 0.658 0(2) | **0.852 0(1)** | **1.50** |
| HASW-HODF | 0.523 9(2) | 0.690 5(3) | 0.614 0(2) | **0.430 0(1)** | 0.655 0(3) | 0.851 2(2) | 2.17 |
| HFSLDL | 0.508 9(4) | 0.689 4(4) | 0.607 1(4) | 0.425 3(3) | 0.649 9(4) | 0.850 9(4) | 3.83 |
| HASW-DCC | 0.518 9(3) | **0.691 2(1)** | 0.613 0(3) | 0.420 0(4) | **0.667 0(1)** | 0.851 3(3) | 2.50 |
| HASW-TAMW | **0.543 9(1)** | 0.690 3(3) | **0.622 0(1)** | 0.432 0(2) | 0.660 0(2) | **0.852 8(1)** | **1.67** |
| HASW-HODF | 0.528 9(2) | 0.690 5(2) | 0.616 0(2) | **0.433 0(1)** | 0.657 0(3) | 0.851 5(2) | 2.00 |
| HFSLE | 0.508 9(4) | 0.689 4(4) | 0.623 0(4) | 0.423 5(3) | 0.646 9(4) | 0.850 8(4) | 3.83 |
| HASW-DCC | 0.513 9(3) | **0.691 2(1)** | 0.628 0(3) | 0.418 0(4) | **0.664 0(1)** | 0.851 2(3) | 2.50 |
| HASW-TAMW | **0.538 9(1)** | 0.691 0(2) | **0.641 0(1)** | 0.431 0(2) | 0.657 0(2) | **0.852 7(1)** | **1.50** |
| HASW-HODF | 0.528 9(2) | 0.690 5(3) | 0.635 0(2) | **0.432 0(1)** | 0.654 0(3) | 0.851 9(2) | 2.17 |
| HFS-MIMR | 0.508 9(4) | 0.687 8(4) | 0.602 8(4) | 0.417 0(3) | 0.670 1(4) | 0.853 5(4) | 3.83 |
| HASW-DCC | 0.533 9(3) | 0.689 4(3) | 0.614 5(3) | 0.405 9(4) | **0.687 0(1)** | 0.854 3(3) | 2.83 |
| HASW-TAMW | **0.588 9(1)** | **0.691 0(1)** | **0.622 8(1)** | **0.433 8(1)** | 0.680 0(3) | **0.855 8(1)** | **1.33** |
| HASW-HODF | 0.583 9(2) | **0.691 0(1)** | 0.619 8(2) | 0.430 8(2) | 0.682 0(2) | 0.854 8(2) | 1.83 |
| GLUFS-HSC | 0.571 4(4) | 0.691 1(4) | 0.643 2(4) | 0.432 9(3) | 0.671 1(4) | 0.854 5(4) | 3.83 |
| HASW-DCC | 0.581 2(3) | **0.701 1(1)** | 0.646 2(3) | 0.427 9(4) | **0.683 1(1)** | 0.855 1(3) | 2.50 |
| HASW-TAMW | **0.601 4(1)** | 0.695 1(3) | **0.658 2(1)** | 0.442 9(2) | 0.677 1(2) | **0.856 8(1)** | **1.67** |
| HASW-HODF | 0.591 0(2) | 0.696 6(2) | 0.650 2(2) | **0.446 9(1)** | 0.674 1(3) | 0.855 5(2) | 2.00 |

表3 3类策略集成至6种分层特征选择方法的 $F_{LCA}$ 结果

Table 3 $F_{LCA}$ results of integrating 3 types of strategies into 6 hierarchical feature selection methods

| 算法 | $F_{LCA}$（排名） | | | | | | 平均排名 |
|---|---|---|---|---|---|---|---|
| | Bridges | DD | CLEF | VOC | Car196 | ILSVRC57 | |
| Hier-FS | 0.730 7(4) | 0.825 2(4) | 0.730 5(4) | 0.631 5(3) | 0.784 9(4) | 0.922 3(4) | 3.83 |
| HASW-DCC | 0.740 0(3) | **0.827 8(1)** | 0.735 5(3) | 0.628 0(4) | **0.800 0(1)** | 0.922 8(3) | 2.50 |
| HASW-TAMW | **0.757 0(1)** | 0.826 5(2) | **0.744 0(1)** | 0.638 0(2) | 0.793 0(2) | **0.925 2(1)** | **1.50** |
| HASW-HODF | 0.748 0(2) | 0.826 0(3) | 0.738 0(2) | **0.638 5(1)** | 0.790 0(3) | 0.923 0(2) | 2.17 |
| HiRRfam-FS | 0.730 7(4) | 0.827 2(4) | 0.732 1(4) | 0.633 5(3) | 0.785 6(4) | 0.923 0(4) | 3.83 |
| HASW-DCC | 0.735 0(3) | **0.829 5(1)** | 0.735 0(3) | 0.630 0(4) | **0.801 0(1)** | 0.923 5(3) | 2.50 |
| HASW-TAMW | **0.750 0(1)** | 0.828 5(2) | **0.746 0(1)** | 0.639 0(2) | 0.794 0(2) | **0.925 8(1)** | **1.50** |
| HASW-HODF | 0.745 0(2) | 0.828 0(3) | 0.738 0(2) | **0.640 0(1)** | 0.791 0(3) | 0.923 6(2) | 2.17 |
| HFSLDL | 0.730 7(4) | 0.827 4(4) | 0.733 5(4) | 0.634 7(3) | 0.788 0(4) | 0.923 1(4) | 3.83 |
| HASW-DCC | 0.740 0(3) | **0.829 8(1)** | 0.737 0(3) | 0.631 0(4) | **0.802 0(1)** | 0.923 6(3) | 2.50 |
| HASW-TAMW | **0.756 0(1)** | 0.828 7(2) | **0.745 0(1)** | 0.641 0(2) | 0.795 0(2) | **0.926 0(1)** | **1.50** |
| HASW-HODF | 0.748 0(2) | 0.828 0(3) | 0.740 0(2) | **0.642 0(1)** | 0.792 0(3) | 0.923 8(2) | 2.17 |
| HFSLE | 0.730 7(4) | 0.827 1(4) | 0.743 1(4) | 0.633 8(3) | 0.785 7(4) | 0.923 1(4) | 3.83 |
| HASW-DCC | 0.735 0(3) | **0.829 3(1)** | 0.746 0(3) | 0.630 0(4) | **0.801 5(1)** | 0.923 5(3) | 2.50 |
| HASW-TAMW | **0.753 0(1)** | 0.828 2(2) | **0.757 0(1)** | 0.642 0(2) | 0.794 5(2) | **0.924 9(1)** | **1.50** |
| HASW-HODF | 0.748 0(2) | 0.828 0(3) | 0.750 0(2) | **0.643 0(1)** | 0.791 5(3) | 0.924 2(2) | 2.17 |
| HFS-MIMR | 0.730 7(4) | 0.826 5(4) | 0.733 1(4) | 0.630 2(3) | 0.804 2(4) | 0.924 5(4) | 3.83 |
| HASW-DCC | 0.744 5(3) | 0.826 8(3) | 0.739 4(3) | 0.622 4(4) | **0.820 0(1)** | 0.924 8(3) | 2.83 |
| HASW-TAMW | **0.769 0(1)** | **0.828 6(1)** | **0.741 6(1)** | **0.641 0(1)** | 0.813 0(3) | **0.925 9(1)** | **1.33** |
| HASW-HODF | 0.762 0(2) | 0.827 6(2) | **0.741 6(1)** | 0.637 0(2) | 0.815 0(2) | 0.924 9(2) | 1.83 |
| GLUFS-HSC | 0.762 0(4) | 0.828 2(4) | 0.755 0(4) | 0.640 5(3) | 0.805 2(4) | 0.925 0(4) | 3.83 |
| HASW-DCC | 0.772 0(3) | **0.833 2(1)** | 0.757 0(3) | 0.635 5(4) | **0.817 1(1)** | 0.925 5(3) | 2.50 |
| HASW-TAMW | **0.792 0(1)** | 0.832 2(3) | **0.767 0(1)** | 0.650 5(2) | 0.811 2(2) | **0.927 2(1)** | **1.67** |
| HASW-HODF | 0.782 0(2) | 0.833 2(2) | 0.761 0(2) | **0.654 5(1)** | 0.808 2(3) | 0.925 8(2) | 2.00 |

集VOC上HASW-HODF效果最显著，对高稀疏数据DD、Car196数据集有效验证了HASW-DCC策略的冗余抑制能力，提升稳定。3种策略在所有评估指标上整体表现优异，特别是在大多数据集上达到了最优性能，同时与最新算法GLUFS-HSC的策略集成对比表明，HASW预处理框架能够有效弥补其对于样本层面偏差的不敏感性，为其提供了稳定的性能增益，充分证明了其在层次结构数据集上的适应性和优势。

为科学验证3种策略在分层特征选择算法中的适应性，本研究采用分层统计检验框架。以每个基础算法为独立分析单元，在保持算法架构一致性的前提下，通过Friedman检验[30]（$\alpha=0.5$）判断策略间性能差异的统计显著性，用Bonferroni-Dunn事后检验[31]识别差异来源。表4的Friedman检验量化验证了策略有效性，在Hier-FS算法中，*Acc*与$F_{LCA}$指标的Friedman统计量远超临界值3.290（$K=4,N=6$），显著拒绝零假设；证明策略提升具有跨指标稳健性。

如图5所示，以代表性算法Hier-FS为例，进行

表4 每个评估指标Friedman统计量（$K=4,N=6$）结果

Table 4 Friedman statistic ($K=4,N=6$) results for each evaluation metric

| 指标 | Friedman统计量 $F_F$ | 临界值（$\alpha=0.5$） |
|---|---|---|
| *Acc* | 4.603 | 3.290 |
| $F_{LCA}$ | 6.780 | |

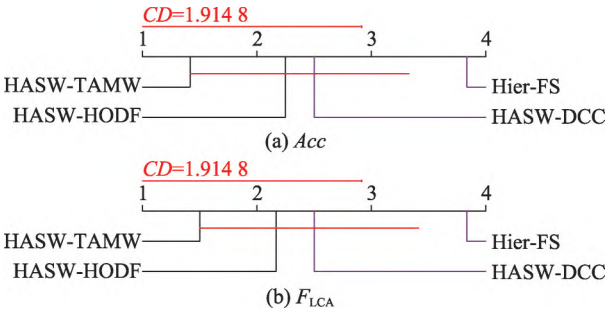



图5 用Bonferroni-Dunn检验集成至Hier-FS算法的HASW 3种策略有效性

Fig.5 Effectiveness of 3 HASW strategies integrated into Hier-FS algorithm via Bonferroni-Dunn test

Bonferroni-Dunn检验（$CD=1.914\,8$）[32]。结果证明所有策略均显著优于基础算法，最优策略与次优无显著差异，表明二者在算法中具有相近的适应性优势。尽管本文仅展示了Hier-FS算法的详细实验结果，但对全部6组算法的检验均表明，3种策略能够普适性地提升分层特征选择算法的适应性，其中TAMW与HODF策略展现出最优的架构融合能力。

### 4.2 消融分析

为验证3种策略的有效性，本文在CLEF、VOC和DD 3个具有代表性的数据集上进行消融实验。

如图6所示，分析结果表明针对CLEF数据集中极端类别不平衡问题，HASW-TAMW策略通过树感知元加权显著提升了各算法在 $F_{LCA}$ 指标上的表现，其中HFSLE算法提升最显著。在层次结构较深、噪声干扰严重的VOC数据集上，HASW-HODF策略通过双重过滤机制有效增强了鲁棒性。而HASW-DCC在该场景下效果不佳，说明过度聚类可能损失判别特征。

对于高度稀疏且语义重叠显著的DD数据集，HASW-DCC通过动态聚类压缩实现了稳定提升，其中HiRRfam-FS算法表现最佳，但整体提升有限，表明蛋白质类数据仍需更复杂策略加以建模。

### 4.3 参数敏感分析

为进一步验证3种策略在具体任务中的适用性与参数稳定性，本文将其分别整合进典型层次结构特征选择算法Hier-FS，并在对应优势场景的数据集上进行关键参数的敏感性分析，以探讨参数变化对模型性能的影响和策略的鲁棒性。

如图7所示，在DD数据集上对HASW-DCC策略中参数压缩比例 $k_s$ 和相关性阈值 $\theta_{dec}$ 进行了敏感性测试。结果表明，当 $k_s=0.6$ 时，$F_{LCA}$ 性能达到最优，能够有效权衡冗余抑制与信息保留。偏离该值则会导致性能显著下降，比例过小易造成信息丢失，过大则引入噪声冗余。同样，$\theta_{dec}=0.6$ 是最佳截断点，能较为准确地区分高相关样本与无关噪声。两参数的性能曲线均呈现倒U型趋势，说明推荐值处于性能最优区间，具备一定鲁棒性。此外，聚类粒度具备生物学合理性，有助于在蛋白质数据中保留关键结构域特征。参数性能曲线呈现出的显著倒U型趋势表明，HASW-DCC策略在推荐值（$k_s=0.6$，$\theta_{dec}=0.6$）附近存在一个性能高台区，说明该方法对参数扰动不敏感，具备良好的鲁棒性。这一特性可有效降低实际应用过程中对参数进行精细调优的依赖，减少因参数设置差异导致的性能波动风险，显著增强了方法的泛化能力。

在CLEF数据集上分析了HASW-TAMW策略中元特征阈值 $\theta_{meta}$ 对模型性能的影响。实验表明，最佳效果为 $\theta_{meta}=0.4$。该参数直接影响样本权重分布及类别间平衡程度：较高阈值增强层次一致性但牺牲样本多样性，较低阈值则提升包容性但可能引入噪声。在CLEF极端类别不平衡场景下，阈值为0.4有助于提升少数类权重，矫正决策边界偏移。

在VOC数据集上分析了HASW-HODF策略中全局阈值 $\theta_{global}$ 与局部阈值 $\theta_{local}$ 的协同影响。最佳组合为 $\theta_{global}=0.3,\theta_{local}=0.5$，能实现有效的分层噪声过滤：前者用于消除跨层级干扰，后者维持类内一致性。参数敏感性分析还显示，图像数据需较高的局部阈值以保障语义连贯性，而层次结构更复杂的数据集则需更宽松的全局阈值以适应语义跨度。

图8展示的 *Acc* 曲面揭示了阈值协同作用的深层机制。曲面在 $\theta_{global}=0.3$ 附近呈现明显的“鞍形”结构，这一特征源于全局置信度的双峰分布特性。模型参数空间中的最优流形对应于沿 $\theta_{local}=1.67\theta_{global}$ 方向形成的脊

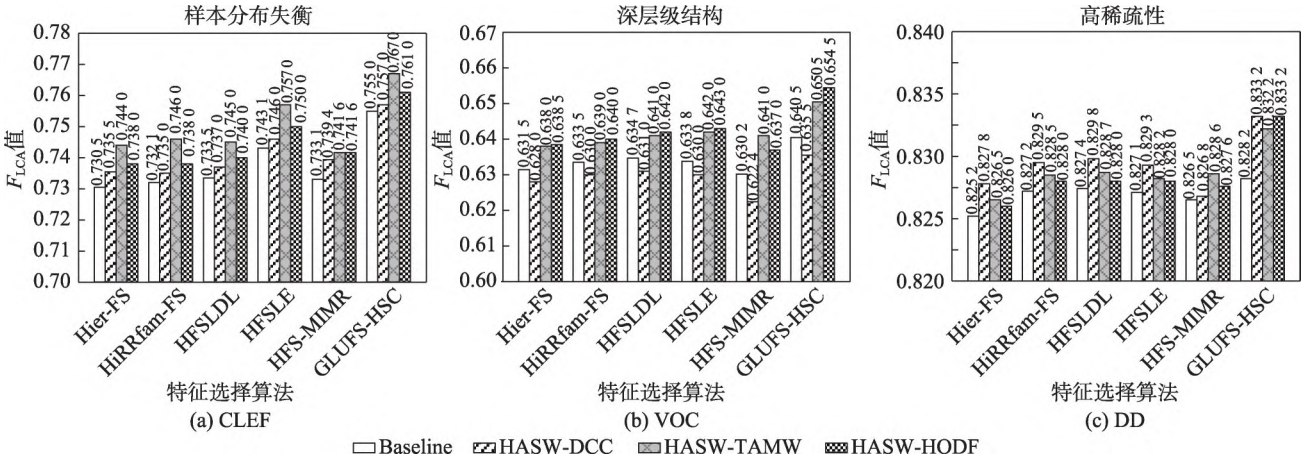



**图6 HASW消融性分析**

Fig.6 Ablation analysis of HASW

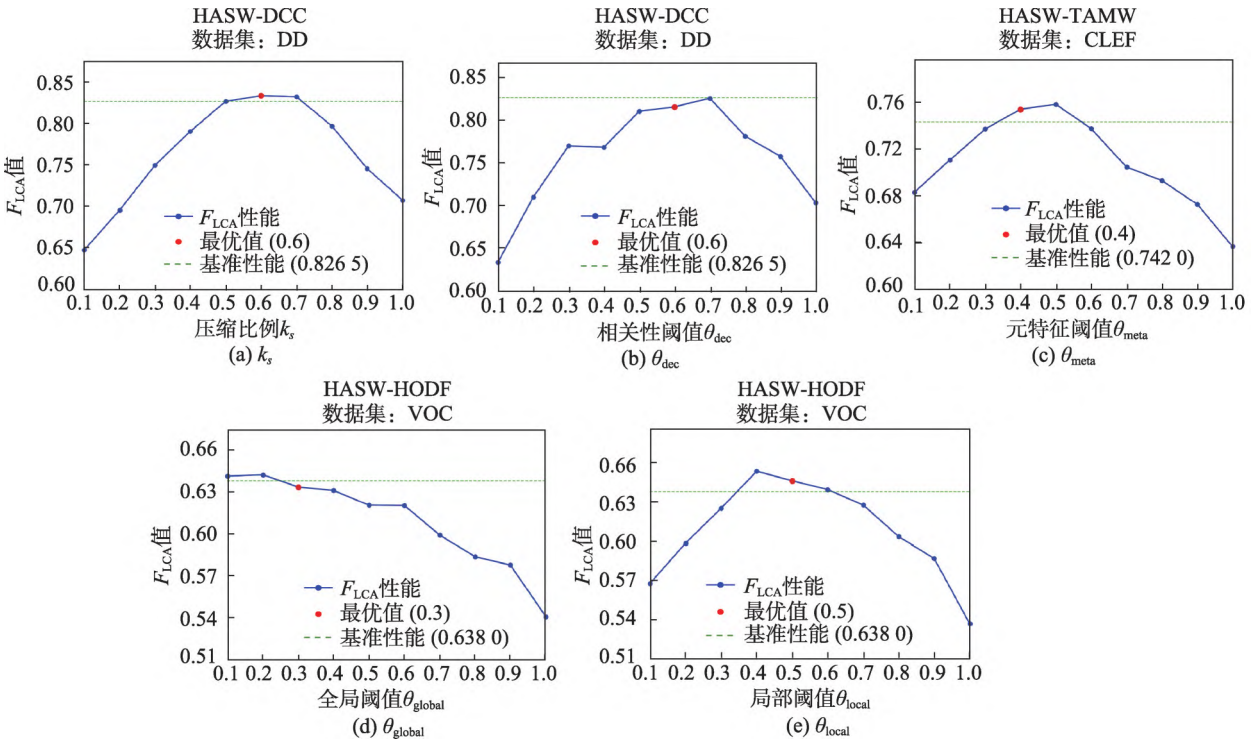



**图7 HASW参数敏感性分析**

Fig.7 Parameter sensitivity analysis of HASW

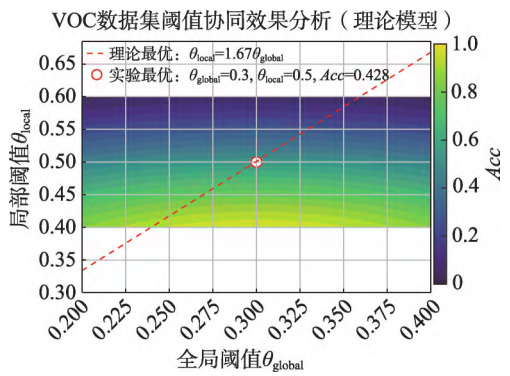



**图8 HASW-HODF策略在VOC数据集阈值协同效果分析**

Fig.8 HASW-HODF strategy: analysis of threshold collaboration effect on VOC dataset

线。曲面的梯度变化规律验证了理论预测，在区域$\theta_{global}<0.25$，梯度方向主要指向全局阈值增大方向，反映跨层次过滤不足；在$\theta_{local}>0.55$区域，梯度转为指向局部阈值减小方向，表明类内过滤。最优点位于这两个临界状态的平衡位置。特别值得注意的是，曲面的等高线在最优比例线附近呈现平行特征，这一几何性质说明当阈值组合沿最优比例线微调时，模型精度保持稳定。该现象为阈值选择的鲁棒性提供了理论保障。

## 4.4 收敛性分析

为验证3种策略集成至不同算法后目标函数仍为凸优化问题，图9以HFS-MIMR算法为例，展示了其在

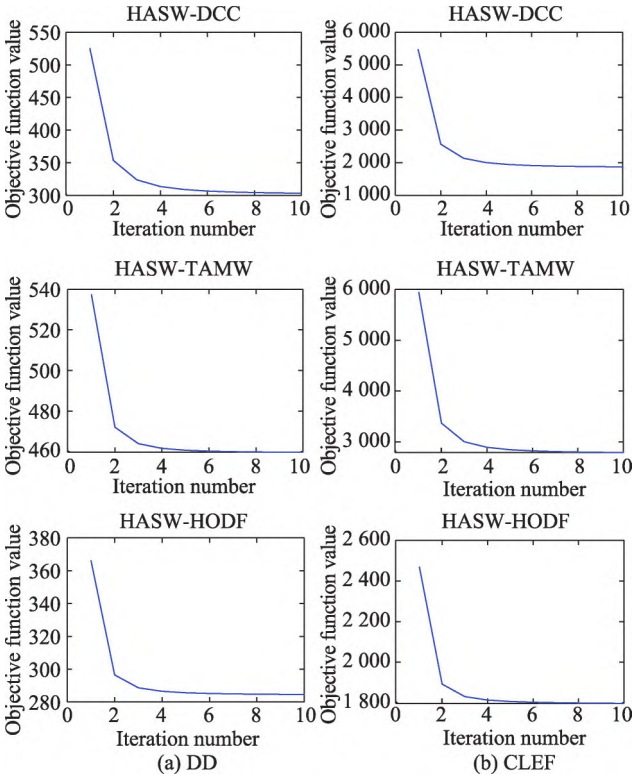



**图9 HASW收敛性研究**

Fig.9 Convergence study of HASW

DD与CLEF数据集上的收敛趋势。

结果表明，无论是在较小的蛋白质数据集还是较大

的图像数据集上，目标函数值随着迭代次数的增加逐渐稳定。实验证明了3个策略在实现全局最优解方面的实用性和有效性。

## 5 结束语

本文针对层次分类任务中的数据冗余、噪声干扰及层次不一致性问题，提出了三种即插即用预处理策略（HASW-DCC、HASW-TAMW、HASW-HODF），通过动态聚类压缩抑制样本冗余，树感知加权优化权重分布，层次离群双重过滤消除跨层噪声，分别解决了语义重叠导致的特征敏感性不足、样本失衡引发的多数类偏好以及层次结构破坏问题，显著提升了特征选择鲁棒性与分类性能。值得注意的是，虽然这些策略在蛋白质分类和图像识别等具有明确层次语义的任务中表现优异，但在处理多父继承的有向无环图（directed acyclic graph，DAG）结构时会面临挑战。这主要源于TAMW策略的单一父节点假设限制了多父特征的融合能力，以及HODF策略的路径依赖性在复杂拓扑中可能导致过度过滤。当前研究仍依赖人工经验调参，且预处理与特征选择的理论耦合性有待深化。未来工作将聚焦于将预处理机制形式化为可微正则化项，嵌入特征选择目标函数，构建端到端的联合优化理论框架；同时探索基于图神经网络的适应性改进，使方法能够更好地处理DAG结构和非严格层次数据。

## 参考文献：


[1] CAI J, LUO J W, WANG S L, et al. Feature selection in machine learning: a new perspective[J]. Neurocomputing, 2018, 300: 70-79.

[2] SAEYS Y, INZA I, LARRAÑAGA P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19): 2507-2517.

[3] BOLÓN-CANEDO V, REMESEIRO B. Feature selection in image analysis: a survey[J]. Artificial Intelligence Review, 2020, 53(4): 2905-2931.

[4] LI Z C, TANG J H. Semi-supervised local feature selection for data classification[J]. Science China Information Sciences, 2021, 64(9): 192108.

[5] ZHAO H, HU Q H, ZHU P F, et al. A recursive regularization based feature selection framework for hierarchical classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(7): 2833-2846.

[6] LIN Y J, LIU H Y, ZHAO H, et al. Hierarchical feature selection based on label distribution learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(6): 5964-5976.

[7] SHI J, LI Z Y, ZHAO H. Feature selection via maximizing inter-class independence and minimizing intra-class redundancy for hierarchical classification[J]. Information Sciences, 2023, 626: 1-18.

[8] LIU H Y, LIN Y J, WANG C X, et al. Semantic-gap-oriented feature selection in hierarchical classification learning[J]. Information Sciences, 2023, 642: 119241.

[9] LIN Z L, LIN Y J. Hierarchical feature selection based on neighborhood interclass spacing from fine to coarse[J]. Neurocomputing, 2024, 575: 127319.

[10] WANG Y B, ZHANG X R, CHENG Y S. A global and local unified feature selection algorithm based on hierarchical structure constraints[J]. Expert Systems with Applications, 2025, 282: 127535.

[11] 辛宇, 杨静, 汤楚蘅, 等. 基于局部语义聚类的语义重叠社区发现算法[J]. 计算机研究与发展, 2015, 52(7): 1510-1521.
XIN Y, YANG J, TANG C H, et al. An overlapping semantic community detection algorithm based on local semantic cluster[J]. Journal of Computer Research and Development, 2015, 52(7): 1510-1521.

[12] 张永清, 卢荣钊, 乔少杰, 等. 一种基于样本空间的类别不平衡数据采样方法[J]. 自动化学报, 2022, 48(10): 2549-2563.
ZHANG Y Q, LU R Z, QIAO S J, et al. A sampling method of imbalanced data based on sample space[J]. Acta Automatica Sinica, 2022, 48(10): 2549-2563.

[13] 刘芬, 范洪强, 吕涛, 等. 基于卡尔曼滤波的含噪声小样本数据处理方法[J]. 上海大学学报(自然科学版), 2022, 28(3): 427-439.
LIU F, FAN H Q, LU T, et al. Kalman filter based method for processing small noisy sample data[J]. Journal of Shanghai University (Natural Science Edition), 2022, 28(3): 427-439.

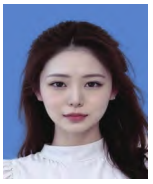
[14] KUMAR A, KAUR A, SINGH P, et al. Efficient multiclass classification using feature selection in high-dimensional datasets[J]. Electronics, 2023, 12(10): 2290.

[15] JIAO R W, NGUYEN B H, XUE B, et al. A survey on evolutionary multiobjective feature selection in classification: approaches, applications, and challenges[J]. IEEE Transactions on Evolutionary Computation, 2024, 28(4): 1156-1176.

[16] KANG M, TIAN J. Machine learning: data pre-processing [J]. Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, 2018, 24: 111-130.

[17] NIE F P, DONG X, TIAN L, et al. Unsupervised feature selection with constrained-norm and optimized graph[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(4): 1702-1713.

[18] SHI Y L, HONG Q H, YAN Y, et al. LDH-ViT: fine-grained visual classification through local concealment and feature selection[J]. Pattern Recognition, 2025, 161: 111224.
[19] DHAL P, AZAD C. A comprehensive survey on feature selection in the various fields of machine learning[J]. Applied Intelligence, 2022, 52(4): 4543-4581.
[20] ELRAHMAN S M A, ABRAHAM A. A review of class imbalance problem[J]. Journal of Network and Innovative Computing, 2013, 1: 9.
[21] YANG C, ROTTENSTEINER F, HEIPKE C. A hierarchical deep learning framework for the consistent classification of land use objects in geospatial databases[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 177: 38-56.
[22] BLANCO V, JAPÓN A, PUERTO J. Robust optimal classification trees under noisy labels[J]. Advances in Data Analysis and Classification, 2022, 16(1): 155-179.
[23] 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.
[24] DIMITROVSKI I, KOCEV D, LOSKOVSKA S, et al. Hierarchical annotation of medical images[J]. Pattern Recognition, 2011, 44(10/11): 2436-2449.
[25] 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.
[26] WANG P, XUE B, LIANG J, et al. Multiobjective differential evolution for feature selection in classification[J]. IEEE Transactions on Cybernetics, 2023, 53(7): 4579-4593.
[27] XIAO J X, HAYS J, EHINGER K A, et al. SUN database: large- scale scene recognition from abbey to zoo[C]//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2010: 3485-3492.
[28] SCHIEBER B, VISHKIN U. On finding lowest common ancestors: simplification and parallelization[J]. SIAM Journal on Computing, 1988, 17(6): 1253-1262.
[29] BALDI P, BRUNAK S, CHAUVIN Y, et al. Assessing the accuracy of prediction algorithms for classification: an overview[J]. Bioinformatics, 2000, 16(5): 412-424.
[30] FRIEDMAN M. A comparison of alternative tests of significance for the problem of m rankings[J]. The Annals of Mathematical Statistics, 1940, 11(1): 86-92.
[31] DUNN O J. Multiple comparisons among means[J]. Journal of the American Statistical Association, 1961, 56(293): 52-64.
[32] DEMŠAR J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006, 7: 1-30.



**张鑫茹**(2001—)，女，硕士研究生，CCF学生会员，主要研究方向为机器学习、分层分类。
**ZHANG Xinru**, born in 2001, M.S. candidate, CCF student member. Her research interests include machine learning and hierarchical classification.

**王一宾**(1970—)，男，硕士，教授，CCF高级会员，主要研究方向为多标签学习、机器学习、软件安全等。
**WANG Yibin**, born in 1970, M.S., professor, CCF senior member. His research interests include multi- label learning, machine learning, software security, etc.

**程玉胜**(1969—)，男，博士，教授，CCF会员，主要研究方向为机器学习、数据挖掘。
**CHENG Yusheng**, born in 1969, Ph.D., professor, CCF member. His research interests include machine learning and data mining.