软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(7):2667−2682 [doi: 10.13328/j.cnki.jos.006335]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于标签关联性的分层分类共有与固有特征选择*

林耀进[1,2], 白盛兴[1,2], 赵 红[1,2], 李绍滋[3], 胡清华[4]

[1](闽南师范大学 计算机学院, 福建 漳州 363000)

[2](数据科学与智能应用福建省高校重点实验室(闽南师范大学), 福建 漳州 363000)

[3](厦门大学 人工智能系, 福建 厦门 361005)

[4](天津大学 智能与计算学部, 天津 300072)

通信作者: 林耀进, E-mail: zzlinyaojin@163.com

摘 要: 在大数据时代, 数据的样本数量、特征维度和类别数量都在急剧增加, 且样本类别间通常存在着层次结构. 如何对层次结构数据进行特征选择具有重要意义. 近年来, 已有相关特征选择算法提出, 然而现有算法未充分利用类别的层次结构信息, 且忽略了不同类节点具有共有与固有属性的特点. 据此, 提出了基于标签关联性的分层分类共有与固有特征选择算法. 该算法利用递归正则化对层次结构的每个内部节点选择对应的固有特征, 并充分利用层次结构分析标签关联性, 进而利用正则化惩罚项学习各子树的共有特征. 该模型不仅能够处理树结构层次化数据, 也能直接处理更为复杂常见的有向无环图结构的层次化数据. 在 6 个树结构数据集和 4 个有向无环图结构数据集上的实验结果, 验证了该算法的有效性.

关键词: 特征选择; 分层分类; 共有特征; 固有特征; 递归正则化

中图法分类号: TP18

中文引用格式: 林耀进, 白盛兴, 赵红, 李绍滋, 胡清华. 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. http://www.jos.org.cn/1000-9825/6335.htm

英文引用格式: Lin YJ, Bai SX, Zhao H, Li SZ, Hu QH. Label-correlation-based Common and Specific Feature Selection for Hierarchical Classification. Ruan Jian Xue Bao/Journal of Software, 2022, 33(7): 2667–2682 (in Chinese). http://www.jos.org.cn/1000-9825/6335.htm

Label-correlation-based Common and Specific Feature Selection for Hierarchical Classification

LIN Yao-Jin[1,2], BAI Sheng-Xing[1,2], ZHAO Hong[1,2], LI Shao-Zi[3], HU Qing-Hua[4]

[1](School of Computer Science, Minnan Normal University, Zhangzhou 363000, China)

[2](Key Laboratory of Data Science and Intelligent Application (Minnan Normal University), Zhangzhou 363000, China)

[3](Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China)

[4](College of Intelligence and Computing, Tianjin University, Tianjin 300072, China)

Abstract: In the era of big data, the sizes of data sets in terms of the number of samples, features, and classes have dramatically increased, and the classes usually exists a hierarchical structure. It is of great significance to select features for hierarchical data. In recent years, relevant feature selection algorithms have been proposed. However, the existing algorithms do not take full advantage of the information of the hierarchical structure of classes, and ignore the common and specific features of different class nodes. This study proposes a label-correlation-based feature selection algorithm for hierarchical classification with common and specific features. The algorithm uses recursive regularization to select the corresponding specific features for each internal node of the hierarchical structure, and makes full use of the hierarchical structure to analyze the label correlation, and then utilizes regularized penalty to select the common features of each subtree. Finally, the proposed model not only can address hierarchical tree data, but also can address more complex hierarchical

* 基金项目: 国家自然科学基金(62076116, 61672272, 61925602, 61732011)
收稿时间: 2020-11-27; 修改时间: 2021-01-27; 采用时间: 2021-03-09

DAG data directly. Experimental results on six hierarchical tree data sets and four hierarchical DAG data sets demonstrate the effectiveness of the proposed algorithm.

Key words: feature selection; hierarchical classification; common features; specific feature; recursive regularization

在大数据时代，数据的样本数量、特征维度和类别数量都在快速增长[1]. 例如，在 ImageNet[2]的数千万个图像样本中包含着数万个类别，且其中每个样本都由成千上万个属性进行描述. 值得注意的是，随着类别数量的急剧增加，类别之间通常存在语义结构，该结构通常用层次结构进行表示. 在机器学习与数据挖掘领域，这种针对层次结构数据的分类任务被称为分层分类[3].

特征选择作为一种处理特征维数灾难的重要手段，已有大量的研究成果[4−9]. 近年来，针对层次结构数据的特征选择方法越来越受到关注[10−14]. 现有传统的特征选择方法虽然可以获得分类学习任务中特征较为紧致的表示[10,13,14]，然而这类算法假定所有样本的类别是相互独立的，完全忽略了类别之间存在着层次语义结构，对所有类别都选择统一的特征子集. 随着层次结构数据中样本类别的急剧增加，会导致已有模型所选特征数目依然巨大，显然无法解决维度灾难问题.

为了应对层次结构数据的特征选择问题，一些借助层次结构、对不同类别分别挑选各自具有判别性特征子集的分层特征选择算法相继被提出. Grimaudo 等人[11]基于 mRMR (minimum redundancy and maximal relevance)特征评价策略，针对类别层次树结构数据，为每一层类别各自选取不同特征子集的算法. 尽管这些工作通过利用类别间的层次结构降低了特征维度，且提高了每个类别的分类任务准确性，但并没有考虑到不同类别之间的依赖关联性. 基于此，Zhao 等人[14]提出了基于递归正则化的分层特征选择框架，将当前类别和父类别、兄弟类别之间的关系分别作为评估特征重要性的正则项. 这些算法都充分利用了层次树结构中类别的关系，相较于传统特征选择算法获得了更高的分类性能.

在实际应用场景如文本分类[15]、功能基因组学[16]和目标识别[17]中，类别的层次关系不仅仅是简单的树结构关系，即一个类别仅归属于一个父节点. 通常，类标签之间的依赖关系是使用有向无环图结构而不是树结构给出的，即一个类别同时属于一个或多个父节点. 例如，在功能基因组学中，存在一个基因不仅属于一个功能类别，而且也直接属于这个类别的祖先功能类别. 此外，在分层分类任务的类别层次结构中，各祖先类别往往与其子孙类别共享着相同的特征，可称为共有特征；而同一层次的兄弟类别在包含其祖先类别的共有特征外，各自还具有相应不同的特征来用于区分其他兄弟类别，这些同层兄弟类别的区分性特征与共有特征一起组成了该类别的固有特征. 如图 1 关于 6 种传染病毒分层分类信息所示[18,19]，整个层次结构的共有特征为根类别传染病毒的固有特征(用蓝色色块表示)，如非细胞结构、具传染性和个体小等. 通过引入新的区分性特征，如含 DNA 核酸(用橙色色块表示)或含 RNA 核酸(用红色色块表示)，又可将传染病毒划分为 DNA 传染病毒与 RNA 传染病毒. 而 RNA 传染病毒的固有特征也将作为该类别子树的共有特征.

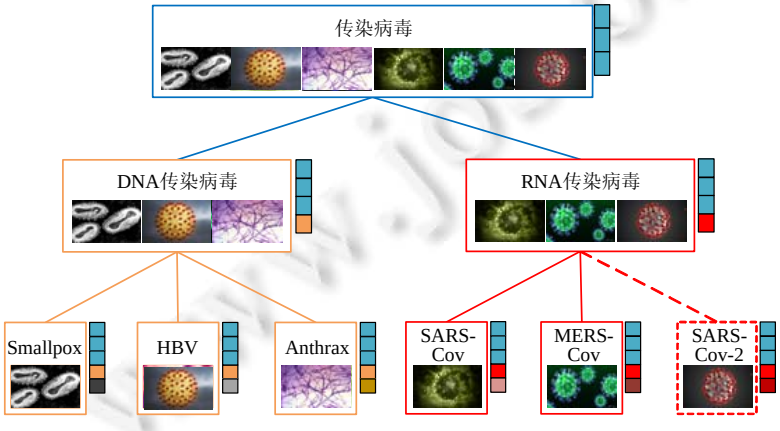


图 1　6 种传染病毒分层分类信息

在分层数据集中, 样本仅具有单一类别, 所以该类数据建模问题本质为分层分类单标记特征选择问题[5], 区分于扁平分类单标记特征选择问题(不考虑类别的层次结构). 同时, 分层分类也可以作为一种特殊的多标记或多输出问题[3]. 如图 1 中, 新型冠状病毒(SARS-Cov-2)同时也是 RNA 传染病毒和传染病毒. 故也可将分层分类特征选择问题作为特殊的多标记特征选择问题. 借鉴多标记特征选择算法对标签关联性的处理, 可以进一步量化层次结构中类别的关联性. 从分类学习的一致性思想出发, 若层次结构中两个标签类别越相似, 那么它们共有特征也越多, 所选的特征也越相似; 同时, 对于同一层次兄弟标签类别, 其各自应包含的固有特征要足以区分于其他兄弟标签类别.

基于以上思想, 本文提出一种基于标签关联性的分层分类共有与固有特征选择算法, 通过检索文献所知, 本文为首次提出解决包含有向无环图结构数据分层特征选择模型的相关文献. 为此, 本文首先对类别层次结构中每个内部节点构建相关损失函数, 用以学习相应节点的固有特征. 其次, 根据类别的层次结构, 计算每个内部节点标签类别的相似度, 进而得到相似度矩阵. 利用相似度矩阵构建特征权重损失函数, 学习层次结构中各子树共有特征. 该模型可以同时处理树结构和有向无环图结构的层次化数据. 最后, 选取 6 个树结构数据集和 4 个有向无环图结构数据集, 将所提算法与近几年提出的其他分层特征选择算法进行实验对比, 以验证所提算法的有效性.

本文第 1 节将介绍相关工作. 第 2 节将提出基于标签关联性的共有与固有分层特征选择算法. 第 3 节对实验结果进行讨论, 并对所提出的算法有效性进行分析. 第 4 节总结全文.

1 相关工作

特征选择是机器学习和数据挖掘的一项重要技术, 具有多个优点, 如加快模型训练速度、降低对过拟合的风险、降低数据分析过程中的存储、内存和处理要求[20]. 传统的特征选择有很多代表性的算法, 如 Fisher Score[6], FSNM[7], mRMR[8], Laplacian Score[9]等. 然而这类算法假定所有样本的类别是相互独立的, 完全忽略了类别之间存在着层次语义结构, 对所有类别都选择统一的特征子集. 实际上, 所选的一些特征仅对识别一个或几个类有作用. 因此, 随着层次结构数据中样本类别的急剧增加, 会导致已有模型所选特征数目依然巨大, 显然无法解决维度灾难问题.

在许多现实应用中, 类的层次结构普遍存在于不同的分类问题中[3,21]. 为了应对层次结构数据的特征选择问题, 一些研究人员提出了利用类别的层次树结构的方法. Grimaudo 等人[11]基于 Peng 等人[8]提出的 mRMR 特征选择算法, 结合类别层次结构, 提出了为每一层类别各自选取不同特征子集的算法. Freeman 等人[12]提出了采用遗传算法联合优化分层分类器和特征选择来提高分类器精度的方法, 并对分层分类的不同分类任务挑选不同的特征子集. Song 等人[13]提出了一种分层文本分类的特征选择算法. 这些工作通过利用类别间的层次结构降低了特征维度, 对不同的节点选择不同的特征子集, 且提高了每个类别的分类任务准确性. 但现有算法并未考虑到不同类别之间的依赖关系.

类别的层次结构间通常存在着依赖关系, 如父子关系、兄弟关系等. 如 Zhao 等人[14]提出了基于递归正则化的分层特征选择框架, 将当前类别和父类别、兄弟类别之间的关系分别作为评估特征重要性的正则项. Tuo 等人[22]提出了基于层次结构子图正则化的分类特征选择算法, 其同时考虑当前类别与祖先类别、子孙类别的关系. 这些算法利用层次结构中类别的关系, 进一步提高了分类任务的准确性. 这些算法进一步利用层次树结构中类别的关系. 但实际应用中, 类别间还存在着更具一般性的有向无环图结构.

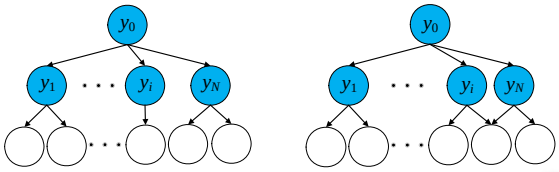
因此, 本文充分考虑并利用类别的层次结构, 提出了基于标签关联性的共有与固有分层特征选择算法, 同时处理包含树结构或有向无环图结构的层次化结构数据.

2 基于标签关联性的共有与固有分层特征选择

本节首先对目标问题进行陈述, 其次提出了基于标签关联性的共有特征学习方法, 进而提出基于标签关联性的共有与固有分层特征选择模型, 最后对该模型进行优化, 并给出算法伪代码与收敛性分析.

2.1 问题陈述

在层次结构数据中一般存在巨大的类别数量，表现为层次结构中节点数量巨多. 类别的层次结构一般存在树结构和有向无环图两种[23]，分别如图 2(a)与图 2(b)所示，其中，树结构可以看作特殊的有向无环图结构.



(a) Tree (b) DAG

图 2 类别的层次结构

为了更具有一般性的定义，树和有向无环图层次结构类别间的‘从属’关系可统一归纳为不可逆性、反自反性和传递性等特性[3]. 用序对$(Y,\prec)$可表示层次结构，其中，Y是标签的集合，$\prec$表示从属关系. 则 3 个特性的形式化描述为:

(1) 不可逆性: 若 $y_i \prec y_j$, $\forall y_i,y_j \in Y$, 则 $y_j \not\prec y_i$.

(2) 反自反性: $\forall y_i \in Y$, 有 $y_i \not\prec y_i$.

(3) 传递性: 若 $y_i \prec y_k$ 且 $y_k \prec y_j$, 对$\forall y_i,y_j,y_k \in Y$, 则 $y_i \prec y_j$.

在功能基因组学和目标识别等领域的层次结构数据集一般拥有超多类别，已有传统的扁平分类特征选择算法[4,6,8]对超多类数据集进行一次性建模，显然无法达到理想的结果，所以本文借助层次结构对每个内部节点进行递归建模，在每一层上分解为更少类别的子问题，有效降低了建模难度. 给定 $X \in R^{n\times m}$ 为样本矩阵，其中，n 和 m 分别为样本和特征的数量. 令类别层次结构(树或有向无环图)的内部节点(非叶子节点)数为 N+1，将样本集 X 划分为 $X_0,X_1,...,X_N$，其中，$X_i=[x_i^1;x_i^2;...;x_i^{n_i}] \in R^{n_i\times m}(0 \leqslant i \leqslant N, n_i \leqslant n)$ 表示为内部节点 i 的样本集合. 同时令 $Y_0,Y_1,...,Y_N$ 表示为对应的标签类别集合，其中，$Y_i=[y_i^1;y_i^2;...;y_i^{n_i}] \in R^{n_i\times d_{\max}}$ 且 $y_i^j=\{0,1\}^{d_{\max}}(1 \leqslant j \leqslant n_i)$，$d_{\max}$ 是内部节点中类别数目的最大值. 令 $W_i=[w_i^1;w_i^2;...;w_i^m] \in R^{m\times d_{\max}}$ 表示为内部节点 i 的权重矩阵，则可对内部节点 i 构建经验损失项 $L(X_i,W_i,Y_i)$，用以学习内部节点 i 的固有特征. 同时，考虑到节点间存在的共有特征关系，可将其作为正则化项包含于 $\Gamma(W_i)$中，对共有特征的依赖性进行惩罚. 最后，为了保证特征的稀疏性，$\Gamma(W_i)$中还应包含稀疏性正则项，则 $\min\limits_{W_i} L(X_i,W_i,Y_i)+\Gamma(W_i)$ 可作为内部节点 i 的子分类任务的目标函数.

综上，考虑类别层次结构时，可将整体目标优化求解转换为对每个内部节点分类损失求和，其目标函数表述如公式(1)所示.

$$F(W_0,W_1,...,W_N)=\min_{W_0,W_1,...,W_N}\sum_{i=0}^{N}(L(X_i,W_i,Y_i)+\Gamma(W_i)) \tag{1}$$

其中，$W_0,W_1,...,W_N$ 分别为内部节点中根节点和每个中间节点的特征权重矩阵.

2.2 基于标签关联性的共有特征学习

在分层分类任务的类别层次结构中，各祖先类别往往与其子孙类别共享着相同的特征，可称为共有特征. 为了获得各类别子树的共有特征，本文通过分析类别的层次结构得到类别间的相似度矩阵，利用相似度矩阵构建特征权重损失函数，来对层次结构中各子树共有特征进行学习. 若类别 y_i 与类别 y_j 越相似，则显然其共享的特征数量也越多，特征权重 W_i 与 W_j 也越相似. 所以，可构建共有特征关系的正则项如公式(2)所示.

$$\Gamma_1(W_i)=C_{ij}W_i^TW_j \tag{2}$$

其中，C_{ij}=1−S_{ij}, S_{ij} 表示为各类别 y_i 与其他类别 y_j 的相似度，且 $S \in R^{N\times N}$.

如图 3 所示，本文基于类别的层次结构，利用二元关联的方式将分层分类问题转为多标签问题，并使用余

弦相似度[24]计算标签的关联性. 当类别 y_i 与类别 y_j 越不相似时, y_i 与 y_j 的余弦相似度 S_{ij} 取值也越小, 则其标签关联性 C_{ij} 取值越大, 表明这两个类别的共有特征较少. 此时, 目标函数的优化将使得 $W_i^TW_j$ 的取值尽量小, 也即特征权重 W_i 与 W_j 也尽量不相似. 理想情况下, 最终所得的特征权重间的余弦相似度将逼近标签类别间的余弦相似度.

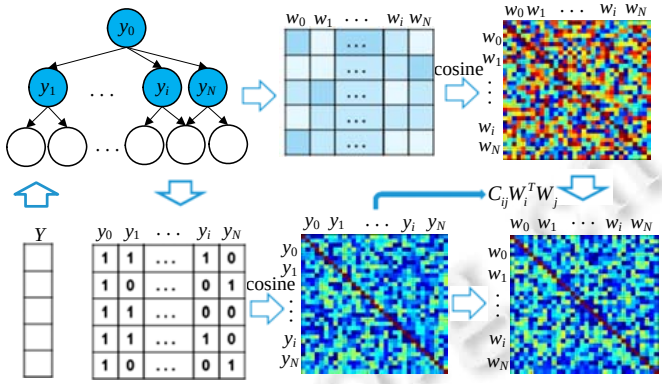


图 3 基于标签关联性的共有特征学习

2.3 基于标签关联性的共有与固有分层特征选择

基于公式(1), 对于固有特征的学习, 损失项可用最小二乘损失、铰链损失或逻辑损失等. 为了便于求得闭合形式解, 本文采用最小二乘损失作为损失函数. 令 $\|\cdot\|_F$ 表示为矩阵的 F 范数, 则内部节点 i 最小二乘损失如公式(3)所示.

$$\| X_iW_i - Y_i \|_F^2 \tag{3}$$

对于特征的稀疏性学习, 有几种基于不同规范的正则化项, 如 ℓ_0 范数、ℓ_1 范数、ℓ_2 范数、$\ell_{2,0}$ 范数和 $\ell_{2,1}$ 范数等. 其中, 基于 ℓ_0 范数的正则化虽然简单直接地做稀疏性学习, 但是函数非凸、非光滑且不连续, 导致模型不好求解; 基于 ℓ_1 范数的正则化虽然可以实现特征稀疏性, 但不能发现类的组结构的稀疏性; 基于 ℓ_2 范数的正则化不能实现特征稀疏性; 基于 $\ell_{2,0}$ 范数的正则化可以实现特征稀疏性, 但基于该正则化优化模型如 ℓ_1 范数一样, 不便于求解; 相比之下, 基于 $\ell_{2,1}$ 范数的正则化是凸的, 可以实现特征稀疏性, 且很容易根据文献[25]中的方法进行优化. 故本文采用 $\ell_{2,1}$ 范数作为稀疏性学习的正则项. 对内部节点 i 稀疏性学习如公式(4)所示.

$$\Gamma_2(W_i)=\|W_i\|_{2,1} \tag{4}$$

结合考虑公式(2)的固有特征学习、公式(3)的共有特征学习和公式(4)特征稀疏性学习, 可得最终目标函数如公式(5)所示.

$$F(W_0,W_1,...,W_N)=\min_{W_0,W_1,...,W_N}\sum_{i=0}^{N}\left(\| X_iW_i - Y_i \|_F^2 +\lambda\| W_i \|_{2,1} +\alpha\sum_{j=0,j\neq i}^{N} C_{ij}W_i^TW_j\right) \tag{5}$$

其中, λ和α为两个非负的参数, 分别控制着特征的稀疏性和共有特征依赖性的惩罚程度.

2.4 模型优化与算法伪代码

由于 $\ell_{2,1}$ 范数的非光滑性, 很难直接推导出公式(5)中优化问题的封闭解. 根据文献[25], 可以用另一种方法解决这个问题如下. $\|W\|_{2,1}$ 对 W 的导数可如公式(6)所示.

$$\frac{\partial\| W \|_{2,1}}{\partial W}=\frac{\partial Tr(W^TDW)}{\partial W}=2DW \tag{6}$$

其中, $D\in R^{d_{\max}\times d_{\max}}$ 是对角矩阵; 第 j 个对角元素为 $d_{jj}^i\in\frac{1}{2\| w_i^j \|_2}$, 若 $w_i^j=0$, 则 $D_{jj}=\varepsilon\varepsilon$.

依据公式(6), 可将公式(5)重新表述为优化问题, 如公式(7)所示.

$$F(W_0,W_1,...,W_N)=\min_{W_0,W_1,...,W_N}\sum_{i=0}^{N}\left(\|X_iW_i-Y_i\|_F^2+\lambda Tr(W_i^TD_iW_i)+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^TW_j\right) \tag{7}$$

对于各个内部节点, 将公式(7)关于 W_i 的导数设置为 0, 如公式(8)所示.

$$\frac{\partial F}{\partial W_i}=2X_i(X_iW_i-Y_i)+2\lambda D_iW_i+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_j=(2X_iX_i+2\lambda D_i)W_i-2X_iY_i+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_j=0 \tag{8}$$

由此可以求得 W_i, 如公式(9)所示.

$$W_i=(X_iX_i+\lambda D_i)^{-1}\left(X_iY_i-\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_j\right) \tag{9}$$

根据目标函数公式(7)及其优化结果公式(9), 给出所提算法的伪代码如算法 1 所示. 通过算法 1, 可以得到权重矩阵 $W=[W_0,W_1,...,W_N]$, 将各个权重进行降序排序, 选择权重值较大的特征来完成对各个内部节点进行特征选择的任务.

算法 1 中, 每次迭代的时间复杂度主要取决于特征权重 W 的计算与更新, 每个内部节点特征权重一次迭代所需的时间复杂度为 $O(m^3+Nm^2d_{\max}+m^2n_i+mn_id_{\max})$, 其中, N 表示内部节点数, m 表示特征数, $d_{\max}$ 表示内部节点最大类别数, n_i 是第 i 个内部节点的样本数. 由于 $X_i^TX_i$ 与 $X_i^TY_i$ $(i=1,2,...,N)$ 仅需要计算 1 次, 时间复杂度为 $O(m^2n_i+mn_id_{\max})$, 所以所有内部节点共需要的时间复杂度为 $O(m^2n+mnd_{\max})$, 其中, n 表示样本总数. 令 T 表示迭代总次数, 则算法 1 的时间复杂度为 $O(T(m^3+Nm^2d_{\max})+m^2n+mnd_{\max})$.

算法 1. 基于标签关联性的共有与固有分层特征选择(label correlation based common and specific hierarchical feature selection, LCCSHFS).

输入: 输入数据 $X_i\in R^{n_i\times m}$, 标签 $Y\in\{0,1\}^{n_i\times m}$, 其中, $i=0,1,\ldots,N$, 正则化参数λ, α, 迭代次数 T.

输出: 权重矩阵 $W\in R^{m\times d_{\max}(N+1)}$.

1: 初始化 $d_{\max}$ 为内部节点的最大类别数, $t=0$

2: $W=[W_0,W_1,...,W_N]$, 随机初始化 $W_i\in R^{m\times d_{\max}}$

3: **WHILE** $t<T$ **DO**

4: **FOR** $i=0:N$ **DO**

5: 通过计算 $d_{jj}^i=1/(2\|w_j^i\|_2)$, 求得矩阵 $D_i^{(t)}$

6: **END FOR**

7: **FOR** $i=0:N$ **DO**

8: 更新 W_i 为 $W_i^{(t+1)}=(X_i^TX_i+\lambda D_i^{(t+1)})^{-1}\left(X_i^TY_i-\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_j^{(t+1)}\right)$

9: **END FOR**

10: $W^{(t+1)}=[W_0,W_1,...,W_N]$

11: $t=t+1$

12: **END WHILE**

13: 返回 W

2.5 收敛性分析

本节将依据定理 1 对算法 1 的收敛性进行分析.

定理 1[7]. 对于任意两个正数 a 和 b, 可以得到不等式如公式(10)所示.

$$a-\frac{a^2}{2b}\leqslant b-\frac{b^2}{2b} \tag{10}$$

证明: 已知$(a-b)^2\geqslant 0$ 知, 可得 $a^2+b^2\geqslant 2ab$, 移项得 $2ab-a^2\leqslant b^2$, 所以可推出 $2a-\frac{a^2}{b}\leqslant 2b-b$, 因此, $a-\frac{a^2}{2b}\leqslant b-\frac{b^2}{2b}$. □

定理 2. 算法 1 在每一次迭代中单调递减公式(5)目标值, 并收敛.

证明: 在第 t 次迭代中, 任取内部节点 i 的权重更新为

$$W_i^{(t+1)}=\arg\min_{W_i}\|X_iW_i^{(t)}-Y_i\|_F^2+\lambda Tr((W_i^{(t)})^T D_iW_i^{(t)})+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^{(t)T}W_j^{(t)} \tag{11}$$

因此, 可推出结果如公式(12)所示.

$$\begin{aligned}&\|X_iW_i^{(t+1)}-Y_i\|_F^2+\lambda Tr((W_i^{(t+1)})^T D_iW_i^{(t+1)})+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^{(t+1)T}W_j^{(t+1)}\leqslant\\&\|X_iW_i^{(t)}-Y_i\|_F^2+\lambda Tr((W_i^{(t)})^T D_iW_i^{(t)})+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^{(t)T}W_j^{(t)}\end{aligned} \tag{12}$$

由 $\|W_i^{(t+1)}\|_{2,1}=\sum_{j=1}^{m}\|(w_i^j)^{(t+1)}\|_2$, 可重写公式(12)如公式(13)所示.

$$\begin{aligned}&\|X_iW_i^{(t+1)}-Y_i\|_F^2+\lambda\sum_{j=1}^{m}\frac{\|(w_i^j)^{(t+1)}\|_2^2}{2\|(w_i^j)^{(t)}\|_2}+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^{(t+1)T}W_j^{(t+1)}\leqslant\\&\|X_iW_i^{(t)}-Y_i\|_F^2+\lambda\sum_{j=1}^{m}\frac{\|(w_i^j)^{(t)}\|_2^2}{2\|(w_i^j)^{(t)}\|_2}+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^{(t)T}W_j^{(t)}\end{aligned} \tag{13}$$

根据定理 1, 令 $a=\|(w_i^j)^{(t+1)}\|_2$, $b=\|(w_i^j)^{(t)}\|_2$, 可得结果如公式(14)所示.

$$\|(w_i^j)^{(t+1)}\|_2-\frac{\|(w_i^j)^{(t+1)}\|_2^2}{2\|(w_i^j)^{(t)}\|_2}\leqslant\|(w_i^j)^{(t)}\|_2-\frac{\|(w_i^j)^{(t)}\|_2^2}{2\|(w_i^j)^{(t)}\|_2} \tag{14}$$

因此, 可得结果如公式(15)所示.

$$\lambda\sum_{j=1}^{m}\left(\|(w_i^j)^{(t+1)}\|_2-\frac{\|(w_i^j)^{(t+1)}\|_2^2}{2\|(w_i^j)^{(t)}\|_2}\right)\leqslant\lambda\sum_{j=1}^{m}\left(\|(w_i^j)^{(t)}\|_2-\frac{\|(w_i^j)^{(t)}\|_2^2}{2\|(w_i^j)^{(t)}\|_2}\right) \tag{15}$$

结合公式(12)与公式(15), 可得结果如公式(16)所示.

$$\|X_iW_i^{(t+1)}-Y_i\|_F^2+\lambda\|W_i^{(t+1)}\|_{2,1}+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^{(t+1)T}W_j^{(t+1)}\leqslant\|X_iW_i^{(t)}-Y_i\|_F^2+\lambda\|W_i^{(t)}\|_{2,1}+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}W_i^{(t)T}W_j^{(t)} \tag{16}$$

因此, 推广公式(16)如公式(17)所示.

$$\begin{aligned}&\sum_{i=0}^{N}\left(\|X_iW_i^{(t+1)}-Y_i\|_F^2+\lambda\|W_i^{(t+1)}\|_{2,1}+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}(W_i^{(t+1)})_i^TW_j^{(t+1)}\right)\leqslant\\&\sum_{i=0}^{N}\left(\|X_iW_i^{(t)}-Y_i\|_F^2+\lambda\|W_i^{(t)}\|_{2,1}+\alpha\sum_{j=0,j\neq i}^{N}C_{ij}(W_i^{(t)})_i^TW_j^{(t)}\right)\end{aligned} \tag{17}$$

公式(17)显示, 算法 1 单调递减公式(5)目标值, 并收敛. □

3 实验结果与分析

本节中, 选取 10 个类别包含层次结构的蛋白质数据集、图像数据集和基因数据集用于显示算法 LCCSHFS 的性能. 其中,

- 6 个为树结构的数据集, 分别是蛋白质数据集 DD[26]和 F194[27]以及图像数据集 VOC[28], ILSVRC65[29], SUN[30,31]和 Cifar100[32].
- 其余 4 个为有向无环图结构的基因数据集, 分别为 Eisen[33], Derisi[33], Cellcycle[33]和 Gasch[33].

表 1 给出了数据集的相关描述信息.

表 1 数据集描述

序号	数据集	训练集	测试集	特征数	节点数	叶子节点数	层数	结构类型
1	DD	3 020	605	473	32	27	3	Tree
2	F194	7 105	1 420	473	202	194	3	Tree
3	VOC	7 178	5 105	1 000	30	20	5	Tree
4	ILSVRC65	12 346	11 845	4 096	65	57	4	Tree
5	SUN	45 109	22 556	4 096	343	324	4	Tree
6	Cifar100	50 000	10 000	4 096	121	100	3	Tree
7	Eisen	1 048	820	80	3 574	1 707	11	DAG
8	Derisi	1 598	1 255	64	4 120	2 037	12	DAG
9	Cellcycle	1 612	1 274	78	4 126	2 041	12	DAG
10	Gasch	1 632	1 275	53	4 132	2 043	12	DAG

3.1 实验评价指标及环境设置

为了评价所提算法的优劣，除了采用传统的预测精度，针对层次结构中错分程度的描述，额外引入两种分层分类评价指标 Tree Induced Error(TIE)[34]和 Hierarchical-$F1$ measure[35]作为评价算法性能的指标.

令 y，$\tilde{y}$ 分别表示样本标记和预测标记，而它们各自的分层分类扩展标记表示为

$$Y_{aug}=y\cup Anc(y),\ \tilde{Y}_{aug}=\tilde{y}\cup Anc(\tilde{y}),$$

其中，$Anc(y)$和 $Anc(\tilde{y})$ 分别表示样本标记 y 和预测标记 $\tilde{y}$ 的祖先节点集合.

• Tree Induced Error

表示样本的预测标记 $\tilde{y}$ 和样本标记 y 在层次结构中节点之间的总边数，如公式(18)所示.

$$TIE(y,\tilde{y})=\sum_{E}(y,\tilde{y}) \tag{18}$$

• Hierarchical-F_1 measure

表示分层准确率和召回率的调和平均，如公式(19)所示.

$$Hierarchical\text{-}F_1\ measure=\frac{2\cdot P_H\cdot R_H}{P_H+R_H} \tag{19}$$

其中，$P_H=\frac{|Y_{aug}\cap\tilde{Y}_{aug}|}{|\tilde{Y}_{aug}|}$, $R_H=\frac{|Y_{aug}\cap\tilde{Y}_{aug}|}{|Y_{aug}|}$.

对于这两个性能评价指标, TIE 指标取值越小越好，而 Hierarchical-F_1 measure 指标的取值越大反而越好.

此外，本文选择了 5 种分层特征选择算法作为对比算法，即 HierFisher (由 Fisher score[6]修改为分层特征选择算法), HierFSNM (由 FSNM[7]修改为分层特征选择算法), HierMRMR[11], Hier-FS[14]和 HiRRfam-FS[14]. 其中, Hier-FS 和 HiRRfam-FS 这两个算法的参数λ设置为 10, HiRRfam-FS 的参数α和β设置为 1.

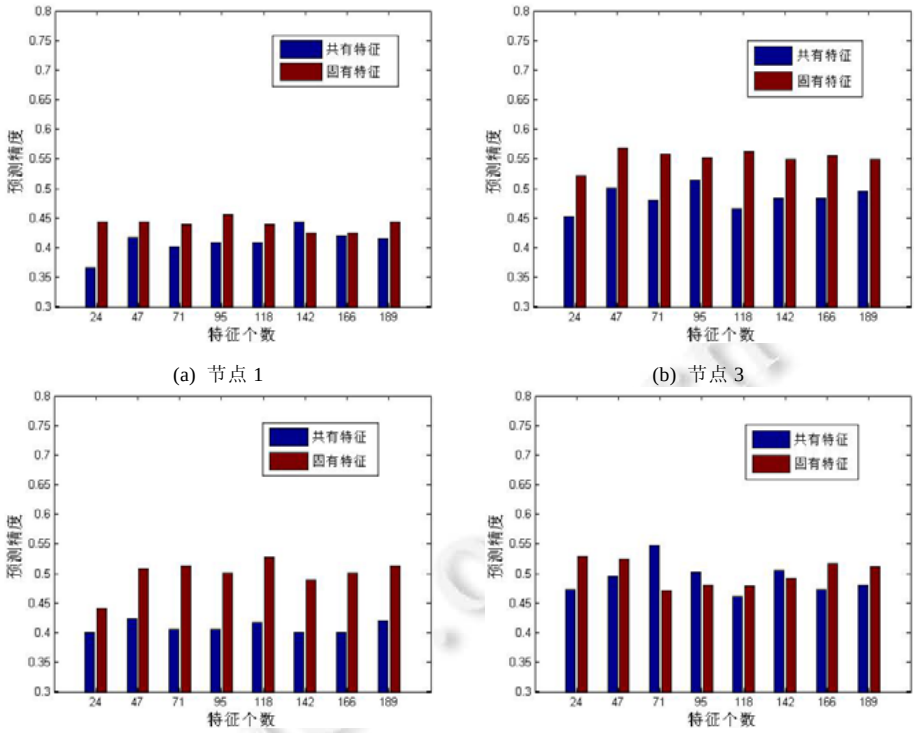
在实验中，采用线性支持向量机作为基分类器. 参数λ和α从网格{0.001,0.01,0.1,1,10,100,1000}中搜索. 针对包含树结构的数据集，参数λ设置为 10，参数α设置为 0.1；针对包含有向无环图结构的数据集，参数λ设置为 100，参数α设置为 0.001.

本文实验系统环境为：一台配有 16 GB 内存、1.90 GHz 的 Intel Core i7-8665U CPU 和 Windows 10 系统的个人主机. 编程环境为 Matlab 2016a 软件.

3.2 在树结构数据上的性能分析

3.2.1 各节点共有特征与固有特征分类性能比较

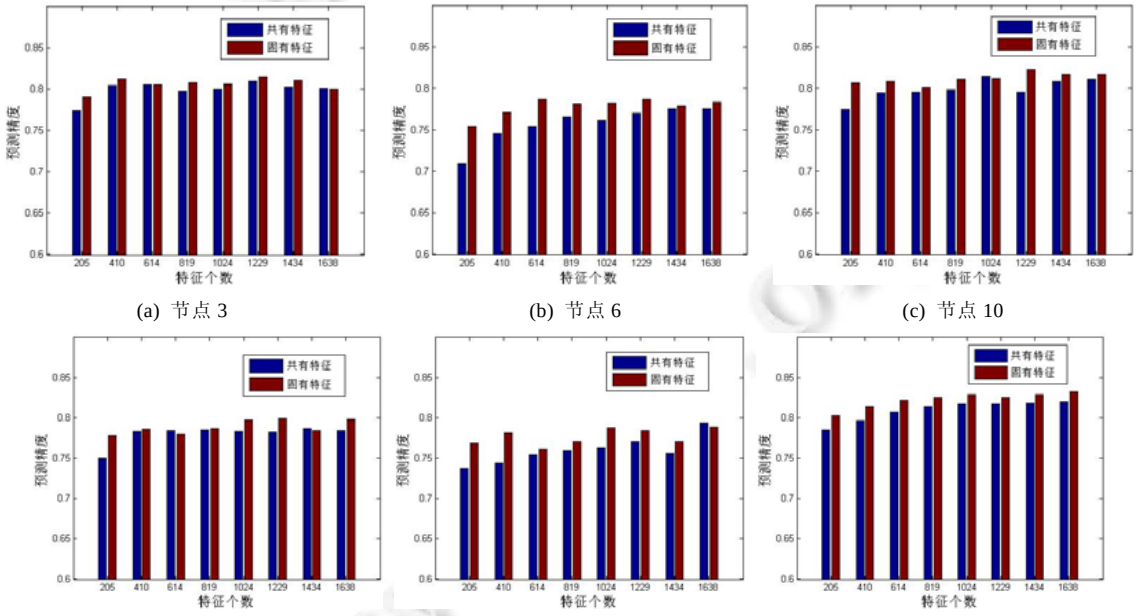
为了验证 LCCSHFS 所提取的共有特征与固有特征的有效性，本节分别对比了各子树节点采用不同数量的共有特征与固有特征时的分类精度. 其中，每个子树节点使用线性支持向量机为基分类器. 为了使实验更有代表性，本节从 6 个包含层次树结构的蛋白质数据集与图像数据集中各选出一个节点数与特征数最多的数据集 F194 与 SUN 作为实验数据集，所使用的特征数量占总数量的 5%, 10%, 15%, 20%, 25%, 30%, 35%和 40%. 最终的实验结果如图 4 和图 5 所示.



(a) 节点 1 (b) 节点 3

(c) 节点 4 (d) 节点 7

图 4 F194 的不同子节点使用不同数量的共有特征和固有特征时的分类精度



(a) 节点 3 (b) 节点 6 (c) 节点 10

(d) 节点 13 (e) 节点 15 (f) 节点 18

图 5 SUN 的不同子节点使用不同数量的共有特征和固有特征时的分类精度

图 4 展示了从蛋白质数据集 F194 的 8 个内部节点中, 随机挑选的 4 个节点采用不同数量的共有特征与固有特征时的分类精度. 从图中可以看出, 各节点共有特征的分类精度与固有特征的分类精度差别都比较大.

随着特征数量的增加, 除了节点1与节点7存在少量共有特征的分类精度大于包含固有特征的分类精度外, 大部分情况下, 都是固有特征的分类性能比较好.

图 5 展示了从图像数据集 SUN 的 21 个内部节点中, 随机挑选的 6 个节点采用不同数量的共有特征与固有特征时的分类精度. 从图中可以看出, 各节点共有特征的分类精度与固有特征的分类精度差别都比较小. 随着特征数量的增加, 除了节点 13 与节点 15 存在少量共有特征的分类精度大于固有特征的分类精度外, 大部分情况下, 同样都是固有特征的分类性能比较好.

综上, 蛋白质数据集每个层次的兄弟标签类别固有特征的差异性比图像数据集明显, 同时, 固有特征的分类性能一般都比共有特征好, 那是因为描述蛋白质数据的特征比图像数据的特征具有更明确的物理意义.

3.2.2 各算法的性能对比

为了进一步分析算法 LCCSHFS 在包含层次树结构数据上的性能, 本节使用自顶向下的线性支持向量机作为分类器, 对采用不同分层特征选择算法的情况下, 对样本的最终类别的分类性能进行分析. 其中, LCCSHFS 使用固有特征作为特征选择的结果; 同时, 为了与 HiRRfam-FS 和 Hier-FS 保持一致, 数据集 DD 和 F194 选择 10%的特征, 其他数据集选择 20%的特征.

表 2 列出了不同分层特征选择算法在不同数据集上的 TIE 值(将结果除以样本数量做归一化处理), 其中最好的结果用粗体显示. 实验结果表明, 所提算法 LCCSHFS 在所有数据集上显著优于其他分层特征选择算法. LCCSHFS 在大多数情况下, 比 HiRRfam-FS 和 Hier-FS 表现得更好、更稳定.

表 2 不同特征选择算法在不同数据集上的标准化 TIE 结果

数据集	HierFisher	HierFSNM	HierMRMR	Hier-FS	HiRRfam-FS	LCCSHFS
DD	0.135 5	0.088 6	0.091 9	0.085 0	**0.083 6**	0.084 3
F194	0.194 5	0.212 3	0.180 0	0.174 6	0.173 0	**0.156 8**
VOC	0.227 1	0.214 4	0.218 8	0.214 3	0.213 8	**0.213 2**
ILSVRC65	0.033 6	0.035 0	0.033 5	**0.032 8**	0.032 9	**0.032 8**
SUN	0.134 1	–	0.132 2	0.128 0	**0.127 1**	0.127 7
Cifar100	0.128 5	–	0.127 3	0.126 9	0.127 2	**0.126 5**

表 3 列出了不同分层特征选择算法在不同数据集上的 Hierarchical-F_1 measure 值. 从实验结果中可以得出与在 TIE 评价指标上相同的结论. 为更直观地对比 LCCSHFS 和 5 个对比算法之间分类性能差异, 引入统计方法 Friedman 检验[36]与 Bonferroni-Dunn 检验[37].

表 3 不同特征选择算法在不同数据集上的 Hierarchical-F_1 measure 值

数据集	HierFisher	HierFSNM	HierMRMR	Hier-FS	HiRRfam-FS	LCCSHFS
DD	0.7741 (6)	0.8524 (4)	0.8468 (5)	0.8584 (3)	**0.8606 (1)**	0.8596 (2)
F194	0.6758 (5)	0.6462 (6)	0.7000 (4)	0.7089 (3)	0.7117 (2)	**0.7387 (1)**
VOC	0.6576 (6)	0.6739 (4)	0.6669 (5)	0.6754 (3)	0.6758 (2)	**0.6769 (1)**
ILSVRC65	0.9580 (5)	0.9563 (6)	0.9581 (4)	**0.9591 (1)**	0.9588 (3)	0.9590 (2)
SUN	0.8324 (5)	–(6)	0.8348 (4)	0.8400 (3)	**0.8411 (1)**	0.8404 (2)
Cifar100	0.7859 (5)	–(6)	0.7879 (4)	0.7885 (2)	0.7880 (3)	**0.7891 (1)**
平均数	**5.33**	**5.33**	**4.33**	**2.5**	**2**	**1.5**

给定 k 个算法和 N 个数据集, r_i^j 是第 j 个算法在第 i 个数据集上的序值, 第 i 个数据集的平均序值为 $R_i=\frac{1}{N}\sum_{i=1}^{N}r_i^j$, 假设所有算法的性能都相同的情况下, 通常使用 $F_F=\frac{(N-1)\chi_F^2}{(\chi_F^2N(k-1)-\chi_F^2)}$ 来进行统计比较, 其中, $\chi_F^2=\frac{12N}{k(k+1)}\left(\sum_{i=1}^{k}R_i^2-\frac{k(k+1)^2}{4}\right)$. 表 4 给出了不同算法的 Hierarchical-F_1 measure 值的序值, 可求得其 F_F=30.79, 大于显著性水平 α=0.1 时的临界值为 $F(6-1,(6-1)\times(6-1))=F(5,25)$=2.09, 因此拒绝所有算法性能相同的假设.

由此, 进一步采用 Bonferroni-Dunn 检验来准确比较不同算法性能差异, 通过测试计算平均值序值差别的临界值域 $CD_\alpha=q_\alpha\sqrt{\frac{k(k+1)}{6N}}$, 在显著性水平 α=0.1 下有 q_α=2.326, 因此可计算出 CD=2.5124 (k=6,N=6).

图 6 显示了对 6 个数据集进行 α=0.1 的 Bonferroni-Dunn 检验结果. 结果表明, LCCSHFS 的 Hierarchical-F_1

measure 值在统计上优于 HierFSNM, HierFisher 和 HierMRMR. 没有一致的证据表明, HiRRfam-FS, Hier-FS 在 Hierarchical-F_1 measure 评价指标上和 LCCSHFS 之间存在统计学差异.

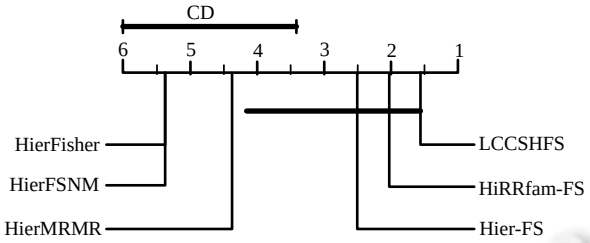


图 6 用 Bonferroni-Dunn 检验比较 LCCSHFS 算法与其他算法的性能

3.3 在有向无环图结构数据上的性能分析

本节分析算法 LCCSHFS 在有向无环图数据上的性能. 其中, 使用自顶向下的线性支持向量机作为分类器, 对采用不同分层特征选择算法的情况下, 对样本的最终类别的分类性能进行分析. LCCSHFS 使用固有特征作为特征选择的结果; 所使用的特征数量分别占总数量的 20%, 30%和 40%. 最终的实验结果如图 7 所示.

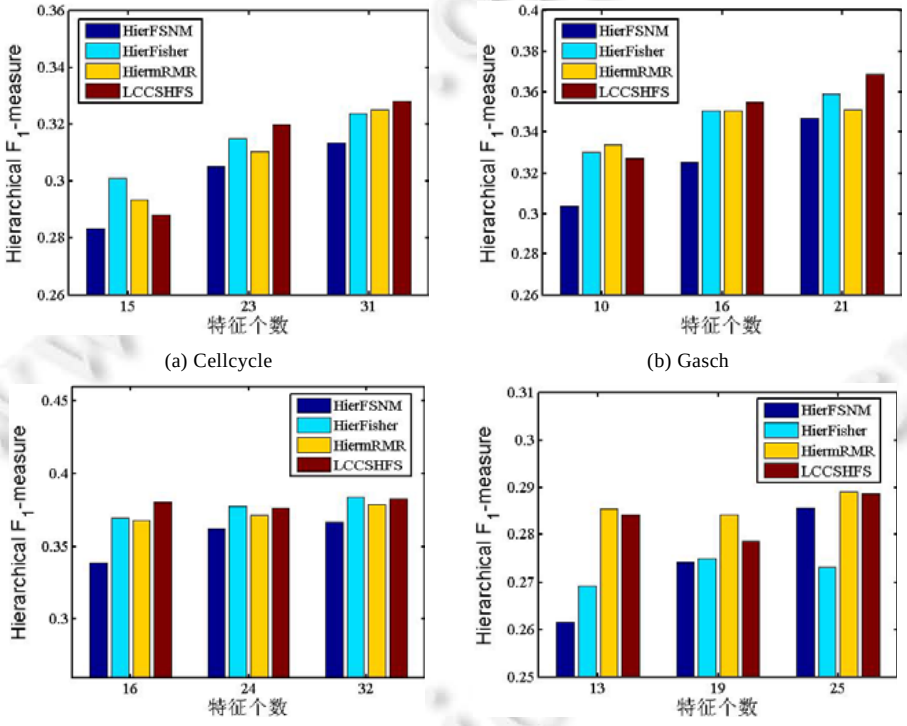


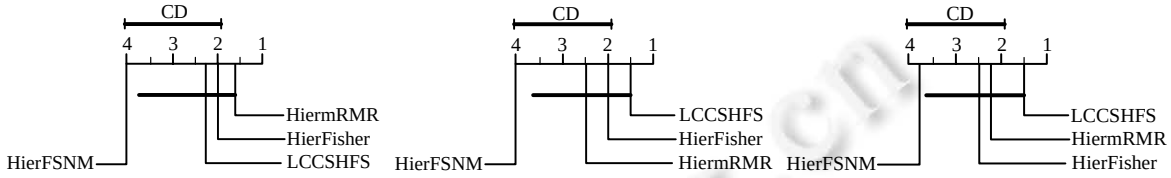
图 7 不同数据集使用不同数量的固有特征时的 Hierarchical-F_1 measure 值

从图 7(a)和图 7(b)可以发现, 在数据集 Cellcycle 和 Gasch 中, 当特征数量选择为 20%时, LCCSHFS 的 Hierarchical-F_1 measure 值比 HierFisher 与 HiermRMR 稍差; 当所选特征数量增加到 30%和 40%时, LCCSHFS 的 Hierarchical-F_1 measure 则取得最好的排名.

从图 7(c)可以发现, 在数据集 Eisen 中, 当特征数量选择为 20%时, LCCSHFS 的 Hierarchical-F_1 measure 值排名最高; 而当所选特征数量增加到 30%和 40%时, LCCSHFS 的 Hierarchical-F_1 measure 值比 HierFisher 稍差.

从图 7(d)可以发现，在数据集 Derisi 中, LCCSHFS 的 Hierarchical-F_1 measure 值仅比 HiermRMR 稍差.

为进一步探索这 4 种特征选择算法在选择不同特征数量的 Hierarchical-F_1 measure 值是否有显著差异，本节进行了 Friedman 检验. 分别计算出选取 20%, 30%和 40%特征数量时, F_F 值分别为 5, 7 和 3.32，大于显著性水平α=0.1 时的临界值为 $F(4-1,(4-1)\times(4-1))=F(3,9)=2.81$，因此拒绝所有算法性能相同的假设. 进一步使用 Bonferroni-Dunn 检验，在显著性水平α=0.1 下有 q_α=2.218，因此可计算出 $CD=2.0247(k=4,N=4)$. 从图 8 可知，在不同特征数量上，LCCSHFS 的 Hierarchical-F_1 measure 值与算法 HierFSNM 有显著性差异，而与算法 HierFisher 和 HiermRMR 相差不大.



(a) 20%特征数 (b) 30%特征数 (c) 40%特征数

图 8 不同数据集使用不同数量的固有特征时，基于 Bonferroni-Dunn 检验比较 LCCSHFS 算法与其他算法的 Hierarchical-F_1 measure 值

综上，由于所使用的包含图结构的数据集所包含的节点都在 3 500 个以上，即需要分类的类别非常多，且所含的特征与样本数量相对较少，对于不考虑层次结构关系的扁平特征选择算法修改而来的分层特征选择算法 HierFSNM, HierFisher, HiermRMR 而言，反而减轻了每个内部节点的分类任务，因而也取得了不错的分类性能. 但从最终结果可以得出：在包含有向无环图结构的数据集上，LCCSHFS 比不考虑类别间分层关系的分层特征选择算法表现得更好且更稳定.

3.4 参数敏感性分析

本节分析所提算法 LCCSHFS 的参数λ和α进行敏感性分析，其中，λ控制着所选特征稀疏程度，α控制着共有特征惩罚程度. 表 4 与表 5 分别列出了包含层次树结构的数据集 ILSVRC65 与包含有向无环图结构的数据集 Eisen 选取 20%的特征时，计算不同参数所得的 Hierarchical-F_1 measure 值. 采用网格搜索方法，在一定范围内对参数λ和α进行调整. 参数λ和α从集合{0.001,0.01,0.1,1,10,100,1000}中选择.

表 4 基于 ILSVRC65 的参数敏感性分析

λ	α						
	0.001	0.01	0.1	1	10	100	1 000
0.001	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8
0.01	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8
0.1	0.948 2	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8
1	0.951 1	0.951 2	0.955 8	0.955 8	0.955 8	0.955 8	0.955 8
10	0.959 0	0.958 8	0.959 0	0.955 8	0.955 8	0.955 8	0.955 8
100	0.960 2	0.960 2	0.960 4	0.959 7	0.955 8	0.955 8	0.955 8
1 000	0.959 0	0.959 0	0.959 0	0.959 0	0.959 3	0.955 8	0.955 8

表 5 基于 Eisen 的参数敏感性分析

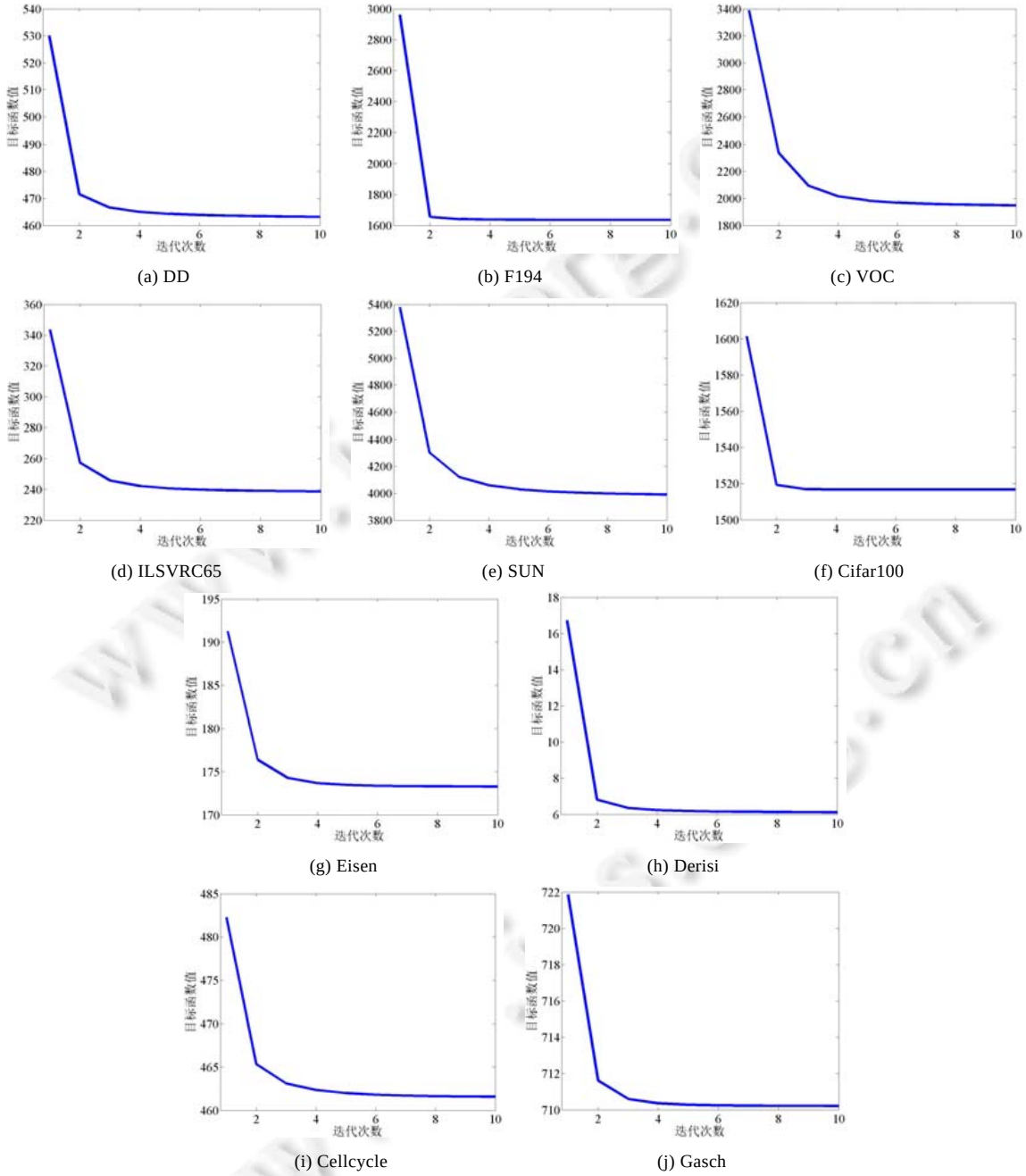
λ	α						
	0.001	0.01	0.1	1	10	100	1 000
0.001	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0
0.01	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0
0.1	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0
1	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0
10	0.369 2	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0
100	0.375 0	0.368 6	0.311 0	0.311 0	0.311 0	0.311 0	0.311 0
1 000	0.363 8	0.363 8	0.363 8	0.311 0	0.311 0	0.311 0	0.311 0

从表 4 与表 5 中可以发现，不论在包含树结构的数据集还是包含有向无环图结果的数据集，当$\alpha\geqslant\lambda$时，

Hierarchical-F_1 measure 值在变差，所以可以适当增大对特征稀疏性的惩罚，不应当对共有特征施加过多惩罚.

3.5 模型的收敛分析

本节研究算法 1 中提出的 LCCSHFS 的收敛性. 所有数据集基于公式(5)中目标函数值的收敛曲线如图 9 所示. 实验中，在所有数据集上设置了最大的迭代次数 T=10. 该图表明，对于所有数据集，目标函数值单调递减，并在不超过 10 次迭代内收敛.



(a) DD (b) F194 (c) VOC

(d) ILSVRC65 (e) SUN (f) Cifar100

(g) Eisen (h) Derisi

(i) Cellcycle (j) Gasch

图 9 目标函数值的收敛曲线

4 总 结

本文提出了一种基于标签关联性的分层分类共有与固有特征选择算法. 利用递归正则化对层次结构的每个内部节点选择对应的固有特征, 充分利用类别的层次结构分析标签关联性, 进而利用正则化惩罚项学习各子树的共有特征. 与现有的分层特征选择算法相比, 本文更充分地利用类别的层次结构所提供的信息, 为层次结构中的每个节点选择不同的特征子集, 且模型具有处理更具复杂性的有向无环图结构数据集的能力. 在 10 个分层数据集上对所提算法与不同的特征选择方法进行了比较. 实验结果验证了所提出算法的有效性. 未来将着重改进标签关联性的度量方法与模型所用的损失函数, 更好地提升分类性能.

References:

[1] Babbar R, Partalas I, Gaussier E, Amini MR, Amblard C. Learning taxonomy adaptation in large-scale classification. Journal of Machine Learning Research, 2016, 17(1): 3350−3386. [doi: 10.5555/2946645.3007051]

[2] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248−255. [doi: 10.1109/CVPR.2009.5206848]

[3] Silla CN, Freitas AA. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, 2011, 22(1-2): 31−72. [doi: 10.1007/s10618-010-0175-9]

[4] Lin YJ, Hu QH, Liu JH, Li JJ, Wu XD. Streaming feature selection for multi-label learning based on fuzzy mutual information. IEEE Trans. on Fuzzy Systems, 2017, 25(6): 1491−1507. [doi: 10.1109/TFUZZ.2017.2735947]

[5] Hu QH, Yu DR, Xie ZX. Numerical attribute reduction based on neighborhood granulation and rough approximation. Ruan Jian Xue Bao/Journal of Software, 2008, 19(3): 640−649 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/640.htm [doi: 10.3724/SP.J.1001.2008.00640]

[6] Gu QQ, Li ZH, Han JW. Generalized fisher score for feature selection. In: Proc. of the ACM Conf. on Uncertainty in Artificial Intelligence. Virginia: AUAI, 2012. 266−273. [doi: 10.5555/3020548.3020580]

[7] Nie FP, Huang H, Cai X, Ding C. Efficient and robust feature selection via joint $L_{2,1}$-norms minimization. In: Proc. of the ACM Int'l Conf. on Neural Information Processing Systems. New York: Curran Press, 2010. 1813−1821. [doi: 10.5555/2997046.2997098]

[8] Peng HC, Long FH, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226−1238. [doi: 10.1109/TPAMI.2005.159]

[9] He XF, Cai D, Niyogi P. Laplacian score for feature selection. In: Proc. of the ACM Int'l Conf. on Neural Information Processing Systems. Massachusetts: MIT, 2005. 507−514. [doi: 10.5555/2976248.2976312]

[10] Hu QH, Wang Y, Zhou YC, Zhao H, Qian YH, Liang JY. Review on hierarchical learning methods for large-scale classification task. Scientia Sinica Informations, 2018, 48(5): 487−500 (in Chinese with English abstract). [doi: 10.1360/N112017-00246]

[11] Grimaudo L, Mellia M, Baralis E. Hierarchical learning for fine grained Internet traffic classification. In: Proc. of the IEEE Int'l Wireless Communications and Mobile Computing Conf. Limassol: IEEE, 2012. 463−468. [doi: 10.1109/IWCMC.2012.6314248]

[12] Freeman C, Kuli D, Basir O. Feature-selected tree-based classification. IEEE Trans. on Cybernetics, 2013, 43(6): 1990−2004. [doi: 10.1109/TSMCB.2012.2237394]

[13] Song J, Zhang PZ, Qin SJ, Gong JP. A method of the feature selection in hierarchical text classification based on the category discrimination and position information. In: Proc. of the Int'l Conf. on Industrial Informatics-computing Technology, Intelligent Technology, Industrial Information Integration. Wuhan: IEEE, 2015. 132−135. [doi: 10.1109/ICIICII.2015.116]

[14] Zhao H, Hu QH, Zhu PF, Wang Y, Wang P. A recursive regularization based feature selection framework for hierarchical classification. IEEE Trans. on Knowledge and Data Engineering, 2021, 33(7): 2833−2846. [doi: 10.1109/TKDE.2019.2960251]

[15] Rousu J, Saunders C, Szedmak S, Shawe-Taylor J. Kernel-based learning of hierarchical multi-label classification models. Journal of Machine Learning Research, 2006, 7: 1601−1626. [doi: 10.1007/s10450-006-0008-8]

[16] Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. Bioinformatics, 2006, 22(7): 830−836. [doi: 10.1093/bioinformatics/btk048]

[17] Stenger B, Thayananthan A, Torr PHS, Cipolla R. Estimating 3D hand pose using hierarchical multi-label classification. Image and Vision Computing, 2007, 25(12): 1885−1894. [doi: 10.1016/j.imavis.2005.12.018]

[18] Zhang ZX, Wang Y. Classification of Viruses. Beijing: Higher Education Press, 2006. 10−15 (in Chinese).

[19] Zhang WH. COVID-19 from Basics to Clinical Practices. Shanghai: Fudan University Press, 2020. 63−73 (in Chinese).

[20] Liu H, Motoda H. Computational Methods of Feature Selection. New York: Chapman and Hall, 2007.

[21] Kosmopoulos A, Partalas I, Gaussier E, Paliouras G, Androutsopoulos I. Evaluation measures for hierarchical classification: A unified view and novel approaches. Data Mining and Knowledge Discovery, 2015, 29(3): 820−865. [doi: 10.1007/s10618-014-0382-x]

[22] Tuo QJ, Zhao H, Hu QH. Hierarchical feature selection with subtree based graph regularization. Knowledge Based Systems, 2019, 163(1): 996−1008. [doi: 10.1016/j.knosys.2018.10.023]

[23] Wu FH, Zhang J, Honavar V. Learning classifiers using hierarchically structured class taxonomies. In: Proc. of the ACM Int'l Conf. on Abstraction, Reformulation and Approximation. Berlin: Springer, 2005. 313−320. [doi: 10.1007/11527862_24]

[24] Zhu SW, Wu JJ, Xiong H, Xia GP. Scaling up top-*K* cosine similarity search. Data and Knowledge Engineering, 2011, 70(1): 60−83. [doi: 10.1016/j.datak.2010.08.004]

[25] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. In: Proc. of the IEEE Int'l Conf. on Neural Information Processing Systems. Massachusetts: MIT, 2006. 41−48. [doi: 10.1007/s10994-007-5040-8]

[26] Ding CHQ, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 2001, 17(4): 349−358. [doi: 10.1093/bioinformatics/17.4.349]

[27] Li DP, Ju Y, Zou Q. Protein folds prediction with hierarchical structured SVM. Current Proteomics, 2016, 13(2): 79−85. [doi: 10.2174/1570164613021605140009 40]

[28] Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. Int'l Journal of Computer Vision, 2010, 88(2): 303−338. [doi: 10.1007/s11263-009-0275-4]

[29] Deng J, Krause J, Berg AC, Li FF. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3450−3457. [doi: 10.1109/CVPR.2012.6248086]

[30] Xiao JX, Hays J, Ehinger KA, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 3485−3492. [doi: 10.1109/CVPR.2010.5539970]

[31] Xiao JX, Ehinger KA, Hays J, Torralba A, Oliva A. SUN database: Exploring a large collection of scene categories. Int'l Journal of Computer Vision, 2016, 119(1): 3−22. [doi: 10.1007/s11263-014-0748-y]

[32] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Technical Report, TR2009, Department of Computer Science, University of Toronto, 2009.

[33] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 2000, 25(1): 25−29. [doi: 10.1038/75556]

[34] Dekel O, Keshet J, Singer Y. Large margin hierarchical classification. In: Proc. of the ACM Int'l Conf. on Machine Learning. New York: Machinery Press, 2004. [doi: 10.1145/1015330.1015374]

[35] Kosmopoulos A, Gaussier E, Paliouras G, Aseervatham S. The ECIR 2010 large scale hierarchical classification workshop. ACM SIGIR Forum, 2010, 44(1): 23−32. [doi: 10.1145/1842890.1842894]

[36] 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. [doi: 10.1214/aoms/1177731944]

[37] Dunn OJ. Multiple comparisons among means. Publication of the American Statistical Association, 1961, 56(293): 52−64. [doi: 10.1080/01621459.1961.10482090]

附中文参考文献:

[5] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报, 2008, 19(3): 640−649. http://www.jos.org.cn/1000-9825/19/640.htm [doi: 10.3724/SP.J.1001.2008.00640]

[10] 胡清华, 王煜, 周玉灿, 赵红, 钱宇华, 梁吉业. 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. [doi: 10.1360/N112017-00246]

[18] 张忠信, 王瑶. 病毒分类学. 北京: 高等教育出版社, 2006. 10−15.

[19] 张文宏. 2019 冠状病毒病: 从基础到临床. 上海: 复旦大学出版社, 2020. 63−73.

林耀进(1980－), 男, 博士, 教授, 主要研究领域为机器学习, 数据挖掘.

白盛兴(1995－), 男, 硕士生, 主要研究领域为机器学习, 数据挖掘.

赵红(1979－), 女, 博士, 教授, CCF 专业会员, 主要研究领域为粒计算, 机器学习.

李绍滋(1963－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 机器学习.

胡清华(1976－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 数据挖掘.