文章编号:1671-9352(2024)03-0061-10 DOI:10.6040/j.issn.1671-9352.7.2023.1073

基于样本相关性的层次特征选择算法

史春雨[1,2],毛煜[1,2*],刘浩阳[1,2],林耀进[1,2]
(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重点实验室(闽南师范大学), 福建 漳州 363000)

摘要:提出了基于样本相关性的层次特征选择算法(hierarchical feature selection algorithm based on instance correlations, HFSIC)以进一步提高分层分类特征选择算法的性能。在使用稀疏正则项去除不相关特征之后,将层次结构中的父子关系与特征空间中样本之间的重构关系相结合,学习同一子树下各类别的样本相关性,利用递归正则优化输出特征权重矩阵。在衡量样本相关性时,将重构系数矩阵整合到训练模型中,同时利用 $l_{2,1}$ 范数去除不相关的和冗余的特征。使用加速近端梯度法解决所提模型的优化问题,并在多个评价指标下评估所提算法的优越性。实验结果表明,所提方法在 5 个数据集上的表现优于其他算法,验证了该算法的有效性。

关键词:特征选择;层次结构;样本相关性;递归正则化
中图分类号:TP391 **文献标志码**:A
引用格式: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.

Hierarchical feature selection algorithm based on instance correlations

SHI Chunyu[1,2], MAO Yu[1,2*], LIU Haoyang[1,2], LIN Yaojin[1,2]
(1. School of Computer Science, Minnan Normal University, Zhangzhou 363000, Fujian, China; 2. Key Laboratory of Data Science and Intelligence Application(Minnan Normal University), Zhangzhou 363000, Fujian, China)

Abstract: A hierarchical feature selection algorithm based on instance correlations (HFSIC) is proposed to further improve the performance of the hierarchical feature selection algorithm. After using sparse regularization items to remove irrelevant features, the parent-child relationship in the hierarchical structure with the reconstruction relationship between samples in the feature space are combined. The correlation of samples of each category under the same subtree are learned. Recursive regularization to optimize the output features weight matrix is used. When measuring the sample correlation, the reconstructed coefficient matrix is integrated into the training model, and the norm is used to remove irrelevant and redundant features. The optimization problem of the proposed model is solved using the accelerated proximal gradient method, and the superiority of the proposed algorithm is evaluated under multiple evaluation metrics. The experimental results show that the proposed method outperforms the other algorithms on five datasets. The test verifies the effectiveness of the proposed algorithm.

Key words: feature selection; hierarchical structure; instance correlation; recursive regularization

0 引言

在大数据时代,数据化样本和特征数量急剧增加,数据空间所包含的类别标记数量繁多,例如恒星光谱有数百类,ImageNet 图像数据和网页数据的类别则达到了数万个[1-3]。传统的方法很难为区分众多的类别找到一个统一且紧凑的全局特征子集,因此,受到分治策略的启发,将分类任务中的众多类别用层次结构来

收稿日期:2023-04-29; **网络出版时间**:2023-10-31 16:28:14
网络出版地址:https://link.cnki.net/urlid/37.1389.N.20231031.0903.004
基金项目:国家自然科学基金资助项目(62076116);福建省自然科学基金资助项目(2022J01914)
第一作者:史春雨(1997—),女,硕士研究生,研究方向为数据挖掘. E-mail:shichunyuuu@ 163.com
*通信作者:毛煜(1985—),男,讲师,硕士生导师,博士,研究方向为数据挖掘. E-mail:maoyu_bit@ 163.com

管理,这种针对层次结构的分类任务被称为分层分类[4]。

众所周知,高维数据空间通常包含大量冗余、不相关的特征,不仅极大地增加了数据分析的存储和计算成本,而且学习算法的效率也会因此降低。特征选择是重要的数据预处理技术,是一种有效的降维方法[5],通过保留相关特征,去除冗余特征,从而简化学习模型,缩短模型的训练时间。传统特征选择方法假定所有类别都是相互独立的,并未考虑到类别间的层次结构[6]。

为了解决在大规模分类学习中具有层次结构的数据所带来的挑战,学者们提出了许多基于层次结构的特征选择方法。Freeman 等[7]提出一种通过联合特征选择和遗传算法的分层分类器,该算法通过将几个"基分类器"排列成一个树形结构来构造一个层次分类器,每个基分类器都将数据集分离为一个逐渐缩小的类别集合;Freeman 等[8]进而提出了一种在不同分类任务中独立的选取不同的特征子集的算法;Grimaudo 等[9]为层次结构的每一层节点独立地选择不同的特征子集并提出了针对分层文本的特征选择算法,但缺乏考虑层次结构中的父子关系以及兄弟关系节点之间的深层联系;Zhao 等[10]提出了基于递归正则化的分层特征选择算法,为不同子分类任务选择不同的特征子集;Tuo 等[11]提出基于子树的图正则化层次特征选择,进一步挖掘层次结构中不同类别之间的双向依赖。

虽然以上特征选择方法都充分利用了类别间存在的层次结构,有效地提升了分类器的性能,但并未考虑样本间关系等问题。近年来,越来越多的学者将样本相关性引入特征选择算法的研究。de Abreu 等[12]提出利用标记空间将样本相关性纳入多标记分类中,并认为属于相同或相似标记集的样本之间存在相关性;Huang 等[13]认为具有相似标记向量的样本通常共享相同的相关性;Huang 等[14]提出利用成对的标记相关性来学习标记特有特征和共有特征,并利用基于 Fisher 判别的正则化项来最小化标记的类内距离和最大化标记的类间距离;Li 等[15]提出基于相关信息的共有和特有特征的多标记学习算法,认为在特征空间共享相似标记的子集中,任何 2 个相关的样本在多标记学习中都具有关键作用。上述方法考虑了样本间的关系,相较于大多数未考虑样本间相关性的多标记特征选择算法获得了更加优异的性能;然而,在层次结构数据集中,样本仅具有单一类别,因此上述方法无法解决层次结构数据集的特征选择问题。

针对上述问题,本文提出了基于样本相关性的层次特征选择算法(hierarchical feature selection algorithm based on instance correlations, HFSIC)以进一步提高分层分类特征选择算法的性能。本文在使用稀疏正则项去除不相关特征之后,将层次结构中的父子关系与特征空间中样本之间的重构关系相结合,学习同一子树下各类别的样本相关性,最后,利用递归正则优化输出特征权重矩阵。实验结果表明,所提方法在 5 个数据集上的表现优于其他算法,验证了该算法的有效性。

1 相关理论

1.1 基于类别的层次结构

层次结构主要分为树结构和有向无环图结构,树结构中的从属关系有 3 个特性:不可逆性、反自反性和传递性[2]。用$(\boldsymbol{Y},\prec)$定义层次结构,$\boldsymbol{Y}$ 表示标记集合,"$\prec$"表示从属关系。

(1) 不可逆性:若 $t_i \prec t_j$, $\forall t_i, t_j \in \boldsymbol{Y}$,则 $t_j \nprec t_i$。

(2) 反自反性:$\forall t_i \in \boldsymbol{Y}$,有 $t_i \nprec t_i$。

(3) 传递性:若 $t_i \prec t_k$ 且 $t_k \prec t_j$,对 $\forall t_i, t_j, t_k \in Y$,则 $t_i \prec t_j$。

1.2 基于层次结构的稀疏学习

设 $\boldsymbol{X} \in \mathbf{R}^{n \times d}$ 是样本矩阵,n 和 d 分别表示样本数和特征数。令类别的层次结构中所有非叶子节点即内部结点数为 $N+1$,则样本 $\boldsymbol{X}$ 划分为 $\boldsymbol{X}_0$、$\boldsymbol{X}_1$、…、$\boldsymbol{X}_N$,其中,$\boldsymbol{X}_i = [x_i^1 \quad x_i^2 \quad \cdots \quad x_i^{n_i}] \in \mathbf{R}^{n_i \times d}$ $(0 \leqslant i \leqslant N, n_i \leqslant n)$ 表示内部节点 i 的样本矩阵。令类别标记矩阵为 $\boldsymbol{Y}_0$、$\boldsymbol{Y}_1$、…、$\boldsymbol{Y}_N$,其中 $\boldsymbol{Y}_i = [y_i^1 \quad y_i^2 \quad \cdots \quad y_i^{n_i}] \in \mathbf{R}^{n_i \times d_{\max}}$,并且 $y_j = \{0, 1\}^{d_{\max}}$ $(1 \leqslant j \leqslant n_i)$,其中 $d_{\max}$ 表示内部节点中类别数目的最大值。

在有监督的特征选择中,稀疏学习因良好的可解释性而被证明是一种有效的技术手段[16]。这种方法的目的是最小化拟合误差和稀疏正则化项[17]。基于稀疏学习的模型的一般形式为

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

其中 $L(\cdot)$是一个损失函数,常用的损失函数包括最小二乘损失、铰链损失和逻辑损失等,本文采用被大多数基于稀疏学习的方法广泛使用的最小二乘损失[10-11],因此损失函数定义为

$$L(\boldsymbol{W};\boldsymbol{X},\boldsymbol{Y})=\|\boldsymbol{XW}-\boldsymbol{Y}\|_{\mathrm{F}}^{2}。 \tag{2}$$

$\Gamma(\boldsymbol{W})$是 $\boldsymbol{W}$ 上的稀疏正则化项。在稀疏学习方法中 $l_{2,1}$范数正则化是为特征选择添加结构化稀疏性,以便简单地去除不相关特征。将式(2)与 $l_{2,1}$范数组合可以将式(1)重新表示为

$$\min_{\boldsymbol{W}} \sum_{i=0}^{N} (\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_{\mathrm{F}}^{2}+\lambda\|\boldsymbol{W}_i\|_{2,1}), \tag{3}$$

其中 $\boldsymbol{W}_i=[w_i^1 \quad w_i^2 \quad \cdots \quad w_i^d]\in\mathbf{R}^{d\times d_{\max}}$表示内部结点 i 的权重矩阵。

1.3　基于父子关系的层次学习

在层次树结构中主要有 2 种类型的关系:父子关系和兄弟关系。一般来说,来自同一子树的类别比来自不同子树的类别共享更多的邻域信息。父子关系是类别树形层次结构中最相邻的结点,会共享某些特征,因此,将父子关系引入到层次特征选择中的正则化项为

$$\sum_{i=1}^{N}\|\boldsymbol{W}_i-\boldsymbol{W}_{p_i}\|_{\mathrm{F}}^{2}, \tag{4}$$

其中 p_i 表示结点 i 的父节点。

2　算法模型

本节首先描述样本相关性模型。在此基础上,定义模型的目标函数。最后,介绍模型的优化算法以及算法的伪代码。

2.1　样本相关性模型

由于特征空间中存在噪声和冗余特征,使用传统的余弦相似度评估样本相关性可能不够准确,无法反映样本之间的复杂关系,因此本文提出利用特征空间中样本之间的重构关系,对一个样本和其他样本之间的关系进行建模学习得到重构系数矩阵 $\boldsymbol{S}\in\mathbf{R}^{n\times n}$,其中 S_{ij}表示第 j 个样本对第 i 个样本的重构贡献。$\boldsymbol{X}\in\mathbf{R}^{n\times d}$是样本矩阵,$n$ 和 d 分别表示样本数和特征数。系数矩阵 $\boldsymbol{S}$ 可以通过求解以下优化问题得到

$$\min_{S}\frac{1}{2}\|\boldsymbol{SX}-\boldsymbol{X}\|_{\mathrm{F}}^{2}+\gamma\sum_{i=1}^{n}\|S_i\|_1 \quad \text{s.t. } \mathrm{diag}(\boldsymbol{S})=0, \tag{5}$$

其中:第一项为所有样本的线性重构误差;第二项为控制每个样本重构系数稀疏性的 l_1 范数;参数 γ 表示平衡正则项的相对重要性;$\mathrm{diag}(\boldsymbol{S})=0$ 表示任何样本对自身重构没有贡献。

为了将特征空间中训练样本之间的重构关系保留在标记空间,对于每一个样本,可以由其他样本根据 $\boldsymbol{S}$ 中的重构系数进行重构,并给出一个度量样本相关性的正则项,使得样本相关性的度量误差尽可能小。其中,正则项为

$$\sum_{i=1}^{n}\|\boldsymbol{XW}-S_i\boldsymbol{XW}\|_{\mathrm{F}}^{2}=\|(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{XW}\|_{\mathrm{F}}^{2}, \tag{6}$$

其中 $S_i=(s_{i1},s_{i2},\cdots,s_{in})\in\mathbf{R}^{n\times n}$表示系数矩阵 $\boldsymbol{S}$ 的第 i 行,$S_{ii}=0$。

最终目标函数定义为

$$\min_{\boldsymbol{W}}\sum_{i=0}^{N}\left(\frac{1}{2}\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_{\mathrm{F}}^{2}+\lambda\|\boldsymbol{W}_i\|_{2,1}+\beta\|\boldsymbol{W}_i-\boldsymbol{W}_{\boldsymbol{p}_i}\|_{\mathrm{F}}^{2}+\alpha\sum_{j=1}^{n}\|\boldsymbol{X}_i\boldsymbol{W}_i-S_j\boldsymbol{X}_i\boldsymbol{W}_i\|_{\mathrm{F}}^{2}\right)。 \tag{7}$$

其中:第一项是损失函数,用于衡量预测标记与真实标记之间的距离;第二项为 $l_{2,1}$正则化项,用于得到稀疏解,满足特征稀疏性的要求;第三项考虑结点间父子关系;第四项用于保持特征空间中训练样本之间的重构关系;α、β、γ、λ 是平衡因子。

2.2　模型优化与算法伪代码

为了方便起见,式(7)中的目标函数为

$$T(\boldsymbol{W})=\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{XW}-\boldsymbol{Y}\|_{\mathrm{F}}^{2}+\lambda\|\boldsymbol{W}\|_{2,1}+\beta\|\boldsymbol{W}-\boldsymbol{W}_{p}\|_{\mathrm{F}}^{2}+\alpha\|(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{XW}\|_{\mathrm{F}}^{2},\tag{8}$$

其中 $\boldsymbol{I}\in\mathbf{R}^{n\times n}$是单位矩阵。

由于 $l_{2,1}$正则化项具有非光滑性,因此本文利用加速近端梯度法来解决函数 $T(\boldsymbol{W})$的最小化问题。加速近端梯度法通常用于解决非光滑、凸函数的优化问题,可以用一个通用的优化框架表示为

$$\min_{\boldsymbol{W}\in\boldsymbol{H}}T(\boldsymbol{W})=f(\boldsymbol{W})+g(\boldsymbol{W}),\tag{9}$$

其中 $\boldsymbol{H}$ 为特征权重矩阵, $f(\boldsymbol{W})$是 Lipschitz 连续的,即 $f(\boldsymbol{W})$满足条件

$$\|\nabla f(\boldsymbol{W}_{i}')-\nabla f(\boldsymbol{W}_{i})\|_{2}\leqslant L_{\mathrm{lip}}\|\Delta\boldsymbol{W}\|_{2}(\forall\boldsymbol{W}_{i}',\boldsymbol{W}_{i}),\tag{10}$$

其中 $1\leqslant i\leqslant q,\Delta\boldsymbol{W}=\boldsymbol{W}_{i}'-\boldsymbol{W}_{i}$, L_{lip}是 Lipschitz 常数。

由式(8)、(9)得 $f(\boldsymbol{W})$和 $g(\boldsymbol{W})$分别为

$$\begin{cases}f(\boldsymbol{W})=\dfrac{1}{2}\|\boldsymbol{XW}-\boldsymbol{Y}\|_{\mathrm{F}}^{2}+\beta\|\boldsymbol{W}-\boldsymbol{W}_{p}\|_{\mathrm{F}}^{2}+\alpha\|(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{XW}\|_{\mathrm{F}}^{2},\\ g(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_{2,1},\end{cases}\tag{11}$$

因此,得到 $f(\boldsymbol{W})$相对于 $\boldsymbol{W}$ 的梯度为

$$\nabla f(\boldsymbol{W})=\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{XW}-\boldsymbol{Y})+2\beta(\boldsymbol{W}-\boldsymbol{W}_{p})+\alpha\boldsymbol{E}^{\mathrm{T}}\boldsymbol{EW},\tag{12}$$

其中 $\boldsymbol{E}=(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}$。

由式(10)、(12)得

$$\begin{aligned}&\|\nabla f(\boldsymbol{W}')-\nabla f(\boldsymbol{W})\|_{\mathrm{F}}^{2}=\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\Delta\boldsymbol{W}+2\beta(\Delta\boldsymbol{W}-\boldsymbol{W}_{p})+\alpha\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\Delta\boldsymbol{W}\|_{\mathrm{F}}^{2}\\ &\leqslant 3(\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\Delta\boldsymbol{W}\|_{\mathrm{F}}^{2}+\|2\beta(\Delta\boldsymbol{W}-\boldsymbol{W}_{p})\|_{\mathrm{F}}^{2}+\|\alpha\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\Delta\boldsymbol{W}\|_{\mathrm{F}}^{2}\\ &\leqslant 3(\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\|_{\mathrm{F}}^{2}\|\Delta\boldsymbol{W}\|_{\mathrm{F}}^{2}+\|2\beta\boldsymbol{I}\|_{\mathrm{F}}^{2}\|\Delta\boldsymbol{W}\|_{\mathrm{F}}^{2}+\|\alpha\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\|_{\mathrm{F}}^{2}\|\Delta\boldsymbol{W}\|_{\mathrm{F}}^{2})\\ &=3(\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\|_{\mathrm{F}}^{2}+\|2\beta\boldsymbol{I}\|_{\mathrm{F}}^{2}+\|\alpha\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\|_{\mathrm{F}}^{2})\|\Delta\boldsymbol{W}\|_{\mathrm{F}}^{2},\end{aligned}\tag{13}$$

因此,Lipschitz 常数为

$$L_{\mathrm{lip}}=\sqrt{3(\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\|_{\mathrm{F}}^{2}+\|2\beta\boldsymbol{I}\|_{\mathrm{F}}^{2}+\|\alpha\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\|_{\mathrm{F}}^{2}}}。\tag{14}$$

考虑 $f(\boldsymbol{W})$的二阶泰勒级数在参数向量 $\boldsymbol{W}^{(t)}$ 的当前估计为

$$\begin{aligned}f(\boldsymbol{W})&=f(\boldsymbol{W}^{(t)})+\langle\nabla f(\boldsymbol{W}^{(t)}),\boldsymbol{W}-\boldsymbol{W}^{(t)}\rangle+\frac{1}{2}\|\boldsymbol{W}-\boldsymbol{W}^{(t)}\|_{\mathrm{F}}^{2}\\ &=\frac{L_{\mathrm{lip}}}{2}\|\boldsymbol{W}-(\boldsymbol{W}^{(t)}-\frac{1}{L_{\mathrm{lip}}}\nabla f(\boldsymbol{W}^{(t)}))\|_{\mathrm{F}}^{2}+c_{\mathrm{const}},\end{aligned}\tag{15}$$

其中: c_{const}是与 $\boldsymbol{W}$ 无关的常数; $\langle\cdot,\cdot\rangle$表示内积。在 $\boldsymbol{W}^{(t+1)}$上得到式(15)的最小值为

$$\boldsymbol{W}^{(t+1)}=\boldsymbol{W}^{(t)}-\frac{1}{L_{\mathrm{lip}}}\nabla f(\boldsymbol{W}^{(t)})。\tag{16}$$

迭代得到 $\boldsymbol{W}$ 的最优解为

$$\boldsymbol{W}^{(t+1)}=\arg\min_{\boldsymbol{W}}\frac{L_{\mathrm{lip}}}{2}\|\boldsymbol{W}-\boldsymbol{Z}\|_{\mathrm{F}}^{2}+g(\boldsymbol{W}),\tag{17}$$

其中, $\boldsymbol{Z}=\boldsymbol{W}^{(t)}-\frac{1}{L_{\mathrm{lip}}}\Delta f(\boldsymbol{W})$。由文献[18]可知,设置 $\boldsymbol{W}^{(t)}=\boldsymbol{W}^{(t)}+\frac{b^{(t-1)}-1}{b^{(t)}}(\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(t-1)})$,对于序列 $b^{(t)}$ 满足 $(b^{(t)})^{2}-b^{(t)}\leqslant(b^{(t-1)})^{2}$, $\boldsymbol{W}^{(t)}$是第 t 次迭代时 $\boldsymbol{W}$ 的结果。

式(17)的封闭解可通过软阈值法计算,定义

$$\boldsymbol{W}_{ij}^{(t+1)}=\begin{cases}\boldsymbol{Z}_{ij}^{(t)}-\lambda/L_{\mathrm{lip}}, & \lambda/L_{\mathrm{lip}}<\boldsymbol{Z}_{ij}^{(t)},\\ 0, & |\boldsymbol{Z}_{ij}^{(t)}|\leqslant\lambda/L_{\mathrm{lip}},\\ \boldsymbol{Z}_{ij}^{(t)}+\lambda/L_{\mathrm{lip}}, & \boldsymbol{Z}_{ij}^{(t)}<-\lambda/L_{\mathrm{lip}}。\end{cases}\tag{18}$$

利用式(9)—(18)的加速近端梯度法也可解决式(5)中关于重构矩阵 $\boldsymbol{S}$ 的优化问题。为了满足约束条件 $\mathrm{diag}(\boldsymbol{S})=0$,本文在每次迭代更新后都将 $\boldsymbol{S}$ 中的所有对角线元素设置为 0。

根据式(7)的最终目标函数以及式(9)—(18)的优化过程,给出本文所提算法的伪代码如算法 1 所示。通过算法 1,得到将各权重降序排序的权重矩阵,并选择权重矩阵中权重值较大的特征完成对各内部节点的特征选择任务。

算法 1　基于样本相关性的层次特征选择算法。

输入　输入数据$\boldsymbol{X}_i \in \mathbf{R}^{n_i \times d}$,

标签$\boldsymbol{Y}_i \in \{0,1\}^{n_i \times d}$,其中 $i=0,1,\cdots,N$,

正则化参数 α、β、γ、λ。

输出　权重矩阵 $\boldsymbol{W} \in \mathbf{R}^{d \times d_{\max}}$。

1. 随机初始化矩阵$\boldsymbol{W}_i \in \mathbf{R}^{d \times d_{\max}}$;
2. $\boldsymbol{W}=[\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_N]$;
3. 初始化参数: b^0, $b^1 \leftarrow 1$, $\boldsymbol{W}^0$, $\boldsymbol{W}^1 \leftarrow (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\gamma\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, $t \leftarrow 0$;
4. 根据式(5)计算重构系数矩阵 $\boldsymbol{S}$;
5. REPEAT:
6. $\boldsymbol{W}^{(t)}=\boldsymbol{W}^{(t)}+\dfrac{b^{(t-1)}-1}{b^{(t)}}(\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(t-1)})$计算第 t 次迭代时 $\boldsymbol{W}$ 的结果;
7. $\boldsymbol{Z}^{(t)}=\boldsymbol{W}^{(t)}-\dfrac{1}{L_{\mathrm{lip}}}\nabla f(\boldsymbol{W}^{(t)})$;
8. 通过软阈值法计算 $\boldsymbol{W}$ 的最优解:

$$\boldsymbol{W}_{ij}^{(t+1)}=\begin{cases}\boldsymbol{Z}_{ij}^{(t)}-\lambda/L_{\mathrm{lip}}, & \lambda/L_{\mathrm{lip}}<\boldsymbol{Z}_{ij}^{(t)},\\ 0, & |\boldsymbol{Z}_{ij}^{(t)}|\leqslant\lambda/L_{\mathrm{lip}},\\ \boldsymbol{Z}_{ij}^{(t)}+\lambda/L_{\mathrm{lip}}, & \boldsymbol{Z}_{ij}^{(t)}<-\lambda/L_{\mathrm{lip}};\end{cases}$$

9. $b^{(t+1)} \leftarrow \dfrac{1+\sqrt{4((b^{(t)})^2+1)}}{2}$;
10. $t \leftarrow t+1$;
11. UNTIL:到达终止条件
12. RETURN $\boldsymbol{W}$;

3　实验分析

本节将依次介绍实验所使用的数据集、评价指标、对比算法、实验设置并对实验结果进行了分析。

3.1　数据集

实验使用 5 个具有层次结构的数据集,包括 2 个蛋白质数据集:protein data data set(DD)、F194,3 个图像数据集:imageNet large scale visual recognition challenge(ILSVRC65)、the PASCAL visual object classes(VOC)、cross language evaluation forum(CLEF)。表 1 给出了数据集的详细信息。

表 1　数据集描述

Table 1　Data set description

序号	数据集	训练集数	测试集数	特征数	节点数	叶子节点数	层数
1	DD	3 020	605	473	32	27	3
2	F194	7 105	1 420	473	202	194	3
3	VOC	7 178	5 105	1 000	30	20	5
4	CLEF	8 368	939	80	88	63	4
5	ILSVRC65	12 346	11 845	4 096	65	57	4

3.2　评价指标

除了传统衡量精度的指标之外,实验还采用 2 个在传统评价指标基础上改进的分层算法特有的评价

指标。

树诱导损失[19](tree induced error, TIE)指标能够反映样本在树结构上的错误程度,即

$$T_{\mathrm{TIE}}(y,\hat{y})=\sum_{E}(y,\hat{y}), \tag{19}$$

其中:y 代表真实标记;$\hat{y}$ 代表预测标记;$\sum_{E}(y,\hat{y})$ 表示真实标记和预测标记节点之间的总边数。

基于增广集合的分层 F_1[20](hierarchical-F_1 measure, F_{H})指标综合考虑分层准确率和召回率,度量错误发生的程度,即

$$F_{\mathrm{H}}=\frac{2\cdot P_{\mathrm{H}}\cdot R_{\mathrm{H}}}{P_{\mathrm{H}}+R_{\mathrm{H}}}, \tag{20}$$

其中,

$$P_{\mathrm{H}}=\frac{|Y_{\mathrm{aug}}\cap\hat{Y}_{\mathrm{aug}}|}{|Y_{\mathrm{aug}}|},\ R_{\mathrm{H}}=\frac{|Y_{\mathrm{aug}}\cap\hat{Y}_{\mathrm{aug}}|}{|\hat{Y}_{\mathrm{aug}}|}。$$

用 Anc(y)表示真实标记 y 的祖先节点集合,Anc($\hat{y}$)表示预测标记 $\hat{y}$ 的祖先节点集合,则真实标记扩展集和预测标记扩展集表示为:$Y_{\mathrm{aug}}=y\cup\mathrm{Anc}(y)$,$\hat{Y}_{\mathrm{aug}}=\hat{y}\cup\mathrm{Anc}(\hat{y})$。

3.3 对比算法

为了证明所提算法的有效性,选择了 5 个层次特征选择算法进行比较,所选择的 5 个对比算法如下:

(1) 基于 $l_{2,1}$范数最小化的高效鲁棒的特征选择(efficient and robust feature selection method to employ joint $l_{2,1}$-norm minimization, HierFSNM)算法[21]。在损失项和正则项的基础上联合 $l_{2,1}$范数最小化的对数据点的异常值具有鲁棒性的分层特征选择方法。

(2) 最大相关性最小冗余性(minimal-redundancy-maximal-relevance, HiermRMR)算法[22]。根据互信息最大统计依赖准则选择候选特征集,引入最小冗余性和最大相关性并结合其他复杂特征选择器提出的一个具有两阶段的分层特征选择算法。

(3) 层次特征选择(hierarchical feature selection, Hier-FS)算法[10]。只考虑内部节点之间的层次结构而不考虑节点之间的亲子关系和兄弟关系特征选择。

(4) 基于家庭关系的递归正则化层次特征选择(family relationship based hierarchical feature selection with recursive regularization, HiRRfam-FS)算法[10]。不仅考虑内部节点之间的层次关系,也考虑节点之间的父子和兄弟关系来进行特征选择。

(5) 基于数据和知识驱动的鲁棒的层次特征选择(robust hierarchical feature selection driven by data and knowledge, HFSDK)算法[4]。由数据和知识驱动,通过分割原始的大标记空间来生成紧凑的特征子集,对数据异常值具有鲁棒性的分层特征选择算法。

3.4 实验设置

该实验的基分类器统一使用线性支持向量机,然后用十折交叉验证的方法验证算法的准确性。根据文献[10]描述,Hier-FS 的参数被设置为 10,HiRRfam-FS 的参数分别设置为 $\alpha=1$,$\beta=1$,$\lambda=10$。为了与算法 Hier-FS 和 HiRRfam-FS 保持一致,本实验为蛋白质数据集和图像数据集分别选择前 10%和 20%的特征。

3.5 实验结果与分析

3.5.1 完整数据集上的性能比较

表 2、3 分别给出了 5 个对比算法在 5 个数据集上不同指标的实验结果,其中,表 2 给出各算法在不同数据集上的 T_{TIE},表 3 给出各算法在不同数据集上的 Hierarchical-F_1 measure 结果。“↓”表示 TIE 取值数值越小越好,“↑”表示 Hierarchical-F_1 measure 取值越大越好,黑色粗体表示最好的结果。由表 2、3 给的 T_{TIE}和 Hierarchical-F_1 measure 分析可得,本实验所提算法在数据集 DD 上的结果略低于 HiRRfam-FS 算法,这是由数据集 DD 的特征具有极稀疏性所致,然而在 F194、ILSVRC65、VOC、CLEF 4 个数据集上均取得最好的结果,这一结果证明了样本相关性对于标记特定特征的选择十分重要。另外,HFSIC 算法通过同时考虑标记相关性和样本相关性在具有层次结构的数据集上可以取得更好的预测效果。

表 2　不同特征选择算法在不同数据集上的标准 TIE 结果(↓)

Table 2　Standard TIE results of different feature selection algorithms on different data sets(↓)

数据集	T_{TIE}					
	HierFSNM	HiermRMR	Hier-FS	HiRRfam-FS	HFSDK	HFSIC
F194	0.212 3(6)	0.180 0(5)	0.174 6(3)	0.173 0(2)	0.175 2(4)	**0.166 0(1)**
DD	0.088 6(5)	0.091 9(6)	0.085 0(3)	**0.083 6(1)**	0.086 3(4)	0.083 9(2)
ILSVRC65	0.035 0(5)	0.033 5(6)	0.032 8(2)	0.032 8(2)	0.032 9(4)	**0.032 6(1)**
VOC	0.214 4(5)	0.2188(6)	0.214 3(3)	0.214 3(3)	0.212 6(2)	**0.208 7(1)**
CLEF	0.207 7(6)	0.182 5(3)	0.182 6(4)	0.182 6(4)	0.174 5(2)	**0.173 5(1)**
平均排名	5.4	5.2	3	2.4	3.2	1.2

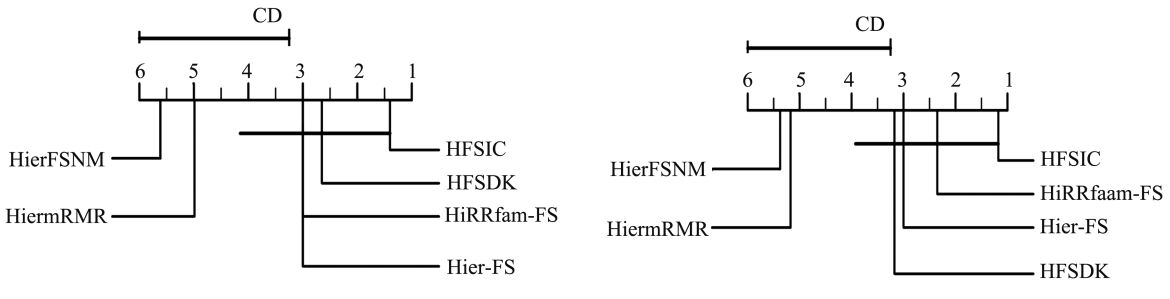
表 3　不同特征选择算法在不同数据集上 Hierarchical-F_1 measure 结果(↑)

Table 3　Hierarchical-F_1 measure results of different feature selection algorithms on different data sets(↑)

数据集	F_{H}					
	HierFSNM	HiermRMR	Hier-FS	HiRRfam-FS	HFSDK	HFSIC
F194	0.646 2(6)	0.700 0(5)	0.708 9(3)	0.711 2(2)	0.707 5(4)	**0.712 7(1)**
DD	0.852 4(5)	0.846 8(6)	0.858 4(3)	**0.860 6(1)**	0.859 0(2)	0.858 4(3)
ILSVRC65	0.956 3(6)	0.958 1(5)	0.959 1(2)	0.958 8(4)	0.958 9(3)	**0.959 2(1)**
VOC	0.673 9(5)	0.666 9(6)	0.675 4(3)	0.675 8(3)	0.677 2(2)	**0.682 0(1)**
CLEF	0.739 6(6)	0.763 5(3)	0.763 1(4)	0.762 3(5)	0.774 2(2)	**0.775 5(1)**
平均排名	5.6	5	3	3	2.6	1.4

为了进一步验证实验结果,引入统计测试。通过 Friedman 测试[23]来检测不同算法之间是否存在显著性差异。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 j 个数据集的平均序值为 $R_j=\frac{1}{N_1}\sum_{i=1}^{N_1} r_i^j$,假设所有算法的性能都相同的情况下,通常使用 $F_F=\frac{(N_1-1)\chi_F^2}{N_1(k-1)-\chi_F^2}$ 来进行统计比较,其中 $\chi_F^2=\frac{12N_1}{k(k+1)}\left(\sum_{i=1}^{k}R_j^2-\frac{k(k+1)^2}{4}\right)$。通过表 3 给出的不同算法的 Hierarchical-F_1 measure 序值,可求得其 F_F 值为 4.838,对于 6 个算法和 5 个数据集的临界值 $F(6-1,(6-1)\times(5-1))=F(5,20)$,大于 $\alpha=0.05$ 时的 F 检验临界值 2.711,因此拒绝“所有算法性能都相同这一假设”。使用 Bonferroni-Dunn 后续检验来准确比较不同算法性能差异。由文献[23]的表 5 可知 $k=6$ 时,$q_{0.10}=2.326$。通过测试计算平均值序值差别的临界值(critical difference, CD)$D_{\mathrm{CD}_\alpha}=q_\alpha\sqrt{\frac{k(k+1)}{6N}}$,因此,可得 $D_{\mathrm{CD}_{0.10}}=2.752$。

同理,通过表 2 给出的不同算法的标准 T_{TIE} 序值,可求得 $F_F=4.373$,大于 $\alpha=0.05$ 时的 F 检验临界值 2.711。由文献[23]的表 5 可得,$k=6$ 时,$q_{0.10}=2.326$,因此,$D_{\mathrm{CD}_{0.10}}=2.752$。图 1 分别为 Hierarchical-$F_1$ measure和 TIE 评价指标下通过 Bonferroni-Dunn 检验($\alpha=0.1$)进一步比较不同算法的性能的检验结果。检验结果表明,HFSIC 在 2 种评价指标上明显优于 HierFSNM 和 HiermRMR 算法。



(a) Hierarchical-F1 measure　　(b) TIE

图 1　通过 Bonferroni-Dunn 检验比较 HFSIC 算法与其他算法的性能

Fig.1　Comparing the performance of HFSIC algorithm with other algorithms through Bonferroni-Dunn test

3.5.2 消融实验

通过消融实验验证 HFSIC 算法中父子关系正则项和样本相关性正则项的有效性。式(7)中各部分的组合表示如下：

(1) HFSIC-α。该函数由式(7) 中的损失函数和惩罚函数以及父子关系正则项组成,即

$$\min_{W}\sum_{i=0}^{N}\left(\frac{1}{2}\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_{\mathrm{F}}^{2}+\lambda\|\boldsymbol{W}_i\|_{2,1}+\beta\|\boldsymbol{W}_i-\boldsymbol{W}_{p_i}\|_{\mathrm{F}}^{2}\right)。\tag{21}$$

(2) HFSIC-β。该函数由式(7) 中的损失函数和惩罚函数以及样本相关性正则项组成,即

$$\min_{W}\sum_{i=0}^{N}\left(\frac{1}{2}\|\boldsymbol{X}_i\boldsymbol{W}_i-\boldsymbol{Y}_i\|_{\mathrm{F}}^{2}+\lambda\|W_i\|_{2,1}+\alpha\sum_{j=1}^{n}\|\boldsymbol{X}_i\boldsymbol{W}_i-S_j\boldsymbol{X}_i\boldsymbol{W}_i\|_{\mathrm{F}}^{2}\right)。\tag{22}$$

图 2 展示了在不同领域的数据集 F194 和 VOC 中,比较 HFSIC 算法不同组合的 Hierarchical-F_1 measure 结果。如图 2 所示,由于同时考虑同一子树下各类别间的相关性和特征空间中样本之间的重构关系的,因此所提算法 HFSIC 优于 HFSIC-α 和 HFSIC-β 算法。

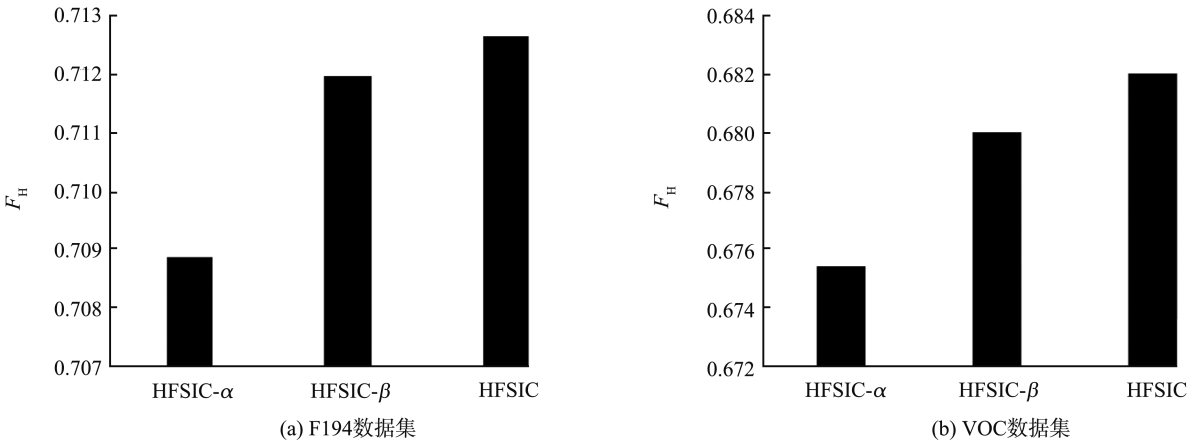


图 2 基于 F194 和 VOC 数据集的消融实验结果

Fig.2 Ablation results based on F194 and VOC data sets

3.5.3 参数敏感性分析

对 HFSIC 算法进行参数敏感性分析,共有参数 α、β、λ,其中 λ 控制特征的稀疏程度,α 和 β 分别控制样本相关性正则项和父子关系正则项。采用网格搜索法在一定范围内调整参数 α、β 和 λ。λ 和 β 从集合 $\{10^{-2},10^{-1},10^{0},10^{1}\}$ 中选择,α 从集合 $\{10^{-3},10^{-2},10^{-1},1\}$ 中选择。实验将在上述范围内选择参数的值,通过改变一个参数固定另外 2 个参数来观察算法 Hierarchical-F_1 measure 值,以此得出算法对变化参数的敏感性。

图 3、4 分别给出 F194 和 VOC 数据集上算法 HFSIC 的参数敏感性分析结果。对于样本相关性的惩罚程度不应太大,父子关系的惩罚程度对图像数据集的影响要远大于蛋白质数据集,而稀疏程度对 2 类数据集的影响基本一致。

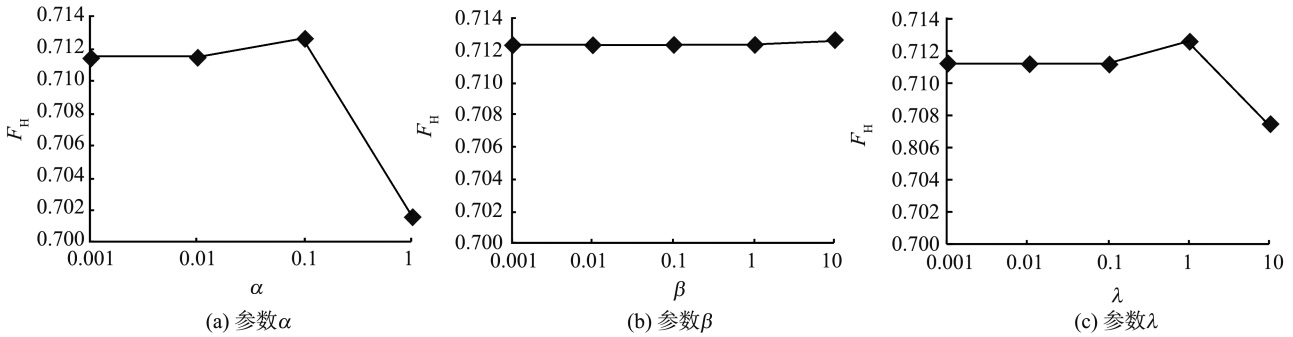
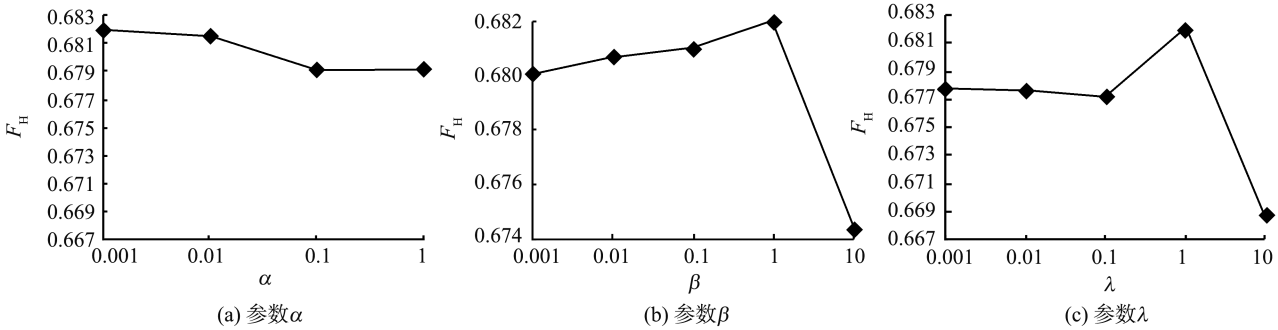


图 3 基于 F194 数据集的参数敏感性分析

Fig.3 Parameter sensitivity analysis based on F194 data set



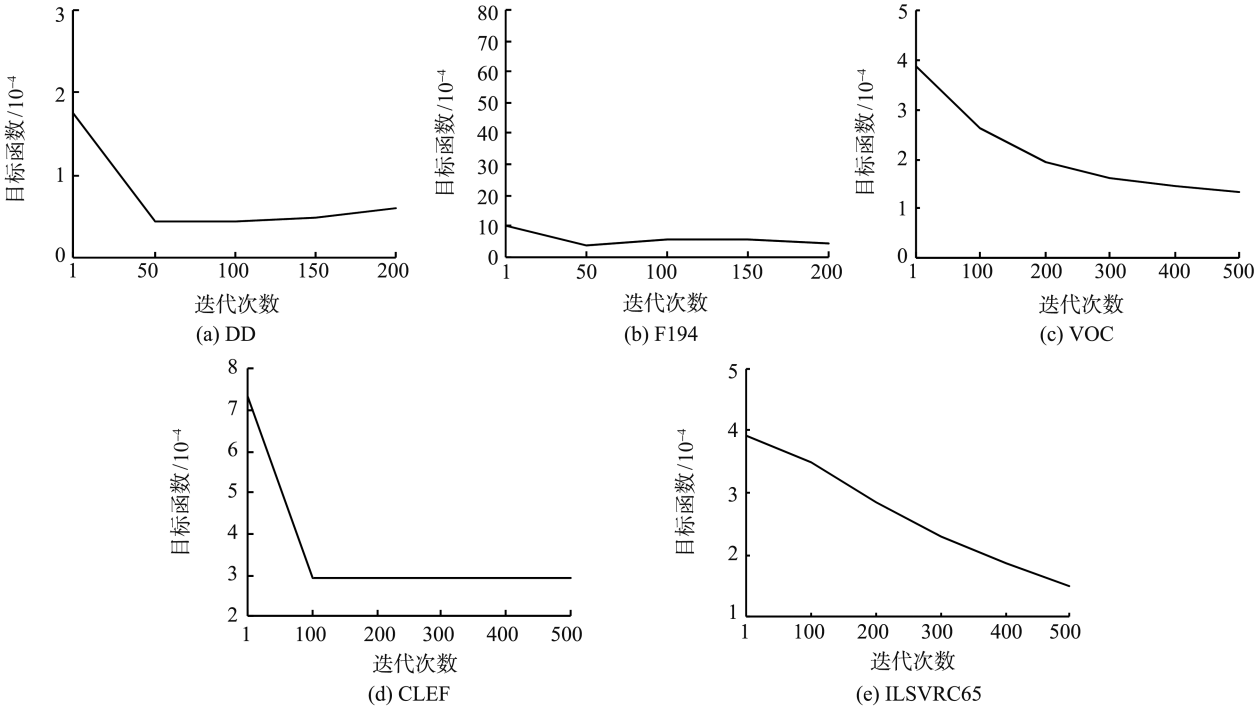
(a) 参数α　　(b) 参数β　　(c) 参数λ

图 4　基于 VOC 数据集的参数敏感性分析

Fig.4　Parameter sensitivity analysis based on VOC data set

3.5.4　模型收敛性分析

本节对所提算法 HFSIC 进行收敛性分析,在式(7) 所提供的目标函数基础上的收敛性曲线如图 5 所示。其中,在蛋白质数据集上设置最大迭代次数 $T=200$,在图像数据集上设置最大迭代次数 $T=500$。实验表明,对于 2 个蛋白质数据集 DD 和 F194,目标函数在 $T=50$ 时取极小值,另外 3 个图像数据集目标函数单调递减并在不超过 500 次内收敛。



(a) DD　　(b) F194　　(c) VOC

(d) CLEF　　(e) ILSVRC65

图 5　目标函数值的收敛曲线

Fig.5　Convergence curve of objective function values

4　小结

本文提出一种基于层次结构和样本相关性的特征选择算法,同时考虑层次结构和样本相关性。算法利用递归正则化项学习层次结构中类别标记之间存在的父子关系和每个类别标记内部样本之间的相关性,选择针对相应标记的标记特定特征,特征空间中的样本相关性通过加速近端梯度法迭代优化得到,最终达到较优的分类效果。实验针对 5 个常见的层次数据集,使用 5 个针对具有层次结构数据的特征选择算法进行对比,结果表明所提的算法可以有效地提高层次结构数据标记预测的准确率。今后将针对所提模型在优化方法上进行改进,同时更好地改进模型使其可以处理有向无环图结构。

参考文献:

[1] 王忠伟,陈叶芳,钱江波,等. 基于 LSH 的高维大数据 k 近邻搜索算法[J]. 电子学报,2016,44(4):906-912.

WANG Zhongwei, CHEN Yefang, QIAN Jiangbo, et al. LSH-based algorithm for k nearest neighbor search on bigdata[J]. Acta Electronica Sinica, 2016, 44(4):906-912.

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

[3] DUDA R O, HART P E, STORK D G. Pattern classification[M]. Hoboken: Wiley, 2000.

[4] LIU Xinxin, ZHOU Yucan, ZHAO Hong. Robust hierarchical feature selection driven by data and knowledge[J]. Information Sciences, 2021, 551:341-357.

[5] WANG Jian, ZHANG Huaqing, WANG Junze, et al. Feature selection using a neural network with group lasso regularization and controlled redundancy[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(3):1110-1123.

[6] 林耀进,白盛兴,赵红,等. 基于标签关联性的分层分类共有与固有特征选择[J]. 软件学报, 2022, 33(7):2667-2682.
LIN Yaojin, BAI Shengxing, ZHAO Hong, et al. A label correlation based common and specific feature selection for large-scale hierarchical classification[J]. Journal of Software, 2022, 33(7):2667-2682.

[7] FREEMAN C, KULIC D, BASIR O. Joint feature selection and hierarchical classifier design[C]//2011 IEEE International Conference on Systems, Man and Cybernetics. Waterloo: IEEE, 2011:1728-1734.

[8] FREEMAN C, KULIC D, BASIR O, et al. Feature-selected tree-based classification[J]. IEEE Transactions on Cybernetics, 2013, 43(6):1990-2004.

[9] 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). Copenhagen: IEEE, 2012:463-468.

[10] ZHAO Hong, HU Qinghua, ZHU Pengfei, 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.

[11] TUO Qianjuan, ZHAO Hong, HU Qinghua. Hierarchical feature selection with subtree based graph regularization[J]. Knowledge-Based Systems, 2018, 163(1):996-1008.

[12] DE ABREU I B M, MANTOVANI R G, CERRI R. Incorporating instance correlations in multi-label classification via label-space[C]//2017 International Joint Conference on Neural Networks (IJCNN). Anchorage: IEEE, 2017:581-588.

[13] HUANG Shengjun, ZHOU Zhihua. Multi-label learning by exploiting label correlations locally[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Toronto, Ontario: AAAI, 2012, 26(1):949-955.

[14] HUANG Jun, LI Guorong, HUANG Qingming, et al. Joint feature selection and classification for multilabel learning[J]. IEEE Transactions on Cybernetics, 2018, 48(3):876-889.

[15] LI Junlong, LI Peipei, HU Xuegang, et al. Learning common and label-specific features for multi-label classification with correlation information[J]. Pattern Recognition, 2022, 121:108-259.

[16] LI Jundong, CHENG Kewei, WANG Suhang, et al. Feature selection: a data perspective[J]. ACM Computing Surveys (CSUR), 2017, 50(6):1-45.

[17] 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.
LIU Haoyang, LIN Yaojin, LIU Jinghua, et al. Hierarchical feature selection from coarse to fine[J]. Acta Electronica Sinica, 2022, 50(11):2778-2789.

[18] LIN Zhouchen, GANESH A, WRIGHT J, et al. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix[J]. Computational Advances, 2009, 10:1-18.

[19] DEKEL O, KESHET J, SINGER Y. Large margin hierarchical classification[C]//Proceedings of the Twenty-first International Conference on Machine Learning. New York: ACM, 2004:1-8.

[20] SILLA C N, FREITAS A A. A survey of hierarchical classification across different application domains[J]. Data Mining & Knowledge Discovery, 2011, 22(1/2):31-72.

[21] NIE Feiping, HUANG Heng, CAI Xiao, 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. Kyoto: IEEE, 2010:1813-1821.

[22] PENG Hanchuan, LONG Fuhui, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8):1226-1238.

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

(编辑:陈丽萍)