融合路径长度的多路径层次标签分类方法

程玉胜[1,2],孙鸿飞[2],余钟萍[2]
(1. 安庆师范大学 智能感知与计算重点实验室,安徽 安庆 246133;
2. 安庆师范大学 计算机与信息学院,安徽 安庆 246133)

摘　要:标签分层分类方法从根节点出发,选择不同路径逐步细化分类。针对现有方法忽略了分类时选择不同长度路径会面临不同风险的问题,提出了一种融合路径长度的多路径层次标签分类方法(MCPL)。首先,采用自顶向下的递归方法,通过逻辑回归获得到达不同节点路径的概率;其次,根据节点的位置信息计算不同节点间的路径长度,利用路径长度为路径赋权,使用赋权后的父节点路径概率和当前节点路径概率以更新当前节点的路径概率;最后,在不同层级,依照节点间的兄弟关系在每个层级选择多个可能的粒度类别,将最后选择的多个类别经过分类器进行再次分类。在 DD、F194、Car196、VOC、CLEF 和 Bridges 数据集上进行实验,相较于六种分层分类方法中最好的结果,MCPL 的样本分类准确率指标平均提高了 2.4%,层次分类指标平均提高了 0.36%,层次结构诱导误差指标平均降低了 1.4%。实验结果表明,MCPL 能够有效提高分类性能。

关键词:层次标签;标签分类;路径长度;多路径选择;自顶向下分类
中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1673-629X(2025)07-0140-08
doi:10.20165/j. cnki. ISSN1673-629X. 2025. 0060

Multi-path Hierarchical Labeling Classification by Fusing Path Lengths

CHENG Yu-sheng[1,2], SUN Hong-fei[2], YU Zhong-ping[2]
(1. Key Laboratory of Intelligent Perception and Computing, Anqing Normal University, Anqing 246133, China;
2. School of Computer and Information, Anqing Normal University, Anqing 246133, China)

Abstract: The hierarchical labeling classification method starts from the root node and gradually refines the classification by selecting different paths. Aiming at the problem that the existing methods ignore the different risks faced when choosing paths of different lengths during classification, a Multi-path Hierarchical Labeling Classification method by Fusing Path Lengths (MCPL) is proposed. Firstly, a top-down recursive approach is employed to obtain the path probabilities to reach different nodes using logistic regression. Secondly, the path lengths between different nodes are calculated based on the positional information of the nodes, and the path length is used to assign weights to the paths. The weighted parent node path probabilities and current node path probabilities are then used to update the current node's path probability. Finally, at different levels, multiple possible granularity categories are selected based on the sibling relationships between nodes, and the final selected categories undergo further classification by a classifier. Experiments on the DD, F194, Car196, VOC, CLEF, and Bridges datasets show that compared to the best results among six hierarchical classification methods, MCPL's sample classification accuracy increases by an average of 2.4%, the hierarchical classification metric improves by an average of 0.36%, and the hierarchical structure-induced error metric decreases by an average of 1.4%. Experimental results demonstrate that MCPL can effectively improve classification performance.

Key words: hierarchical label; label classification; path length; multi-path selection; top-down classification

0　引　言

数据获取技术的发展为机器学习任务带来巨大挑战,尤其是在大规模分类任务中,这种挑战不仅体现在样本特征上,还涉及样本数量和标签数量[1]。随着数据规模持续扩张,传统方法暴露出计算复杂度飙升以及分类性能下滑的缺陷。鉴于此,依据分而治之理念构建的分层分类方法[2]应运而生,其通过将大规模数据分级处理,有效削减了分类的复杂程度,并显著提升了分类性能。

在分层分类方法的研究中,Liu 等人[3]提出标签

收稿日期:2024-12-10　　**修回日期**:2025-04-11
基金项目:安徽省教育重大项目(2024AH040175);安徽省教育重点项目(2024AH051099)
作者简介:程玉胜(1969-),男,教授,博士,通信作者,研究方向为粗糙集、数据挖掘;孙鸿飞(2000-),男,硕士研究生,研究方向为机器学习、数据挖掘。

增强的分层特征选择方法,利用损失函数和粗细粒度分布的一致性来获得标签分布。该方法选择合适特征,有效缩小语义差距,但忽略了不同细粒度类别间的语义鸿沟。Guo 等人[4]提出自底向上的特征聚类方法,为已知类别构建语义差距最小树,有效缓解细粒度类别间的语义鸿沟,却没有考虑到层间错误传播导致的语义风险问题。为此,Wang 等人[5]将分类精度精确在叶子节点的上级节点,降低层间误差传播。接着,Lin 等人[6]提出自顶向下的分层特征选择方法,实现了较为准确的细粒度分类。在进一步考虑层次结构类间独立性的基础上,Shi 等人[7]采用最大化类间独立性,最小化类内冗余的策略进行分层分类。Huang 等人[8]提出为极端多类构建分层分类的方法,即通过联合相似度和组划分动态地获取层次结构,有效地解决构建层次结构遇到的复杂度较高、参数较多的问题。

上述方法从不同角度对分层分类进行了探索与优化,但未充分利用树结构的路径信息。在分层多路径分类方面,Zheng 等人[9]为树结构的每个层次设定信息熵阈值,并根据样本数量等信息相应地分配代价敏感权重进而减少层间错误传播。然而,该方法在遇到概率相近的路径时,易出现路径选择错误的情况。鉴于此,Guo 等人[10]提出了一种改进策略,即每次选取不同数量的最大概率路径,使测试样本从根节点开始,以自顶向下的方式递归至最底层的标签节点。

这些方法取得了较优的结果,但都没有考虑树结构中不同节点间的路径长度,路径长度用节点之间的距离表示。如图 1(b),测试样本在传统的逻辑概率树中进行分类时,会先将测试样本分类为第二层概率最大的“猫科”节点。然后从猫科节点出发,在最底层分类为“猫”的概率 0.31 大于“狮子”的概率 0.3,测试样本错误分类。在使用融合路径长度概率树进行分类后,分类为“狮子”的概率 0.075 大于分类为“猫”的概率 0.071,测试样本分类准确。这是因为“猫科”节点与“猫”节点距离大于“猫科”节点与“狮子”节点。因此,该文提出一种融合路径长度的多路径层次标签分类方法(Multi-path hierarchical labeling classification by fusing path lengths,MCPL)。首先使用图 1(a)中的原始数据,通过聚类方法确定树结构中不同节点的位置信息,如图 1(b)中的位置信息树;其次根据节点的位置信息计算上下层节点之间的路径长度,如图 1(c)中不同长度路径用不同的线段表示;然后结合图 1(b)中的逻辑回归概率和路径长度信息的约束计算上层节点到下层节点不同路径的概率,如图 1(c);最后选择概率最大的前 k 个最底层节点放到分类器中分类,如图 1(c)中,将 0.075 和 0.073 两个节点放入分类器中,分类器最终将其分类为狮子。

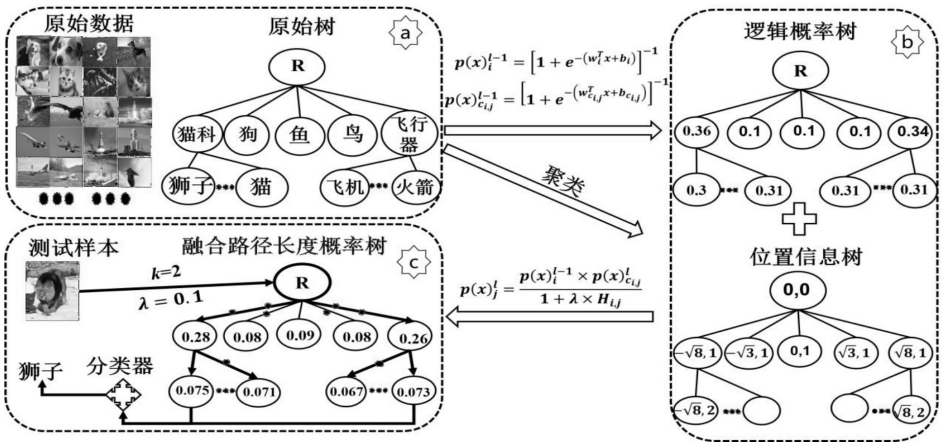

图 1　MCPL 方法框架

主要贡献如下:

(1)发现树结构中节点的位置信息。使用聚类方法,将节点包含的样本聚为一类,类中心定义为节点的位置信息。如图 1(b)中的位置信息树,每个节点都在二维平面上有具体的位置信息。

(2)使用位置信息计算节点间路径长度,计算路径概率时融合路径长度约束。如图 1(c)中 R 节点到“猫科”节点路径距离为 3,用三段线段表示。到“狗”节点路径距离为 2,用两段线段表示。最后得到融合路径长度的概率树。

(3)使用融合路径长度的概率树进行分层多路径选择,实验结果证明了 MCPL 的有效性。

1　融合路径长度多路径层次标签分类

1.1　MCPL 模型概述

定义 $\boldsymbol{X} \in \boldsymbol{R}^{m \times n}$ 为数据矩阵,其中 m 是样本数量,n 是特征数量。定义 $\boldsymbol{Y}^l = \{y_1^l, y_2^l, \cdots, y_d^l\}$ 为不同层的类别标签。其中 d 是第 l 层的标签数量。定义路径数为 k,其中 $1 \leqslant k \leqslant \min(d)$。MCPL 模型的构建主要分为三个步骤。

(1)通过逻辑回归模型计算出到达每个类节点的路径概率得到逻辑概率树,如图 1(a)中的原始树使用

逻辑回归公式得到图 1(b)中的逻辑概率树。采用传统的 k - means 聚类方法给每个节点赋予位置信息形成位置信息树。如图 1(a)中的原始数据进行聚类得到图 1(b)中的位置信息树。计算位置信息树中每条路径长度,如图 1(c)中不同长度路径用不同线段表示。然后结合逻辑概率树,计算出树结构中每条路径的概率。最后结合路径长度约束和父节点、当前节点的路径概率,更新到达当前节点的路径概率。

(2)在树结构的每一层依次选取概率最大的前 k 条路径,递归到最底层节点。

(3)将最底层选中的几个节点放到分类器中进行分类。

1.2 融合节点间路径长度的概率计算

在标签的层次树结构中,最上层节点是根节点,由上到下粒度逐渐细化,最底层节点是标签节点。每个节点相连的上一级节点是该节点的父节点。如图 1(a)中的“猫”节点的父节点是“猫科”节点。在同一层级且共享同一个父节点的一组节点为兄弟节点。比如图 1(a)中的“鱼”节点,它的兄弟节点是“狗”节点、“猫科”节点、“鸟”节点和“飞行器”节点。图 1 中“猫”节点的兄弟节点是“狮子”节点,“火箭”节点的兄弟节点是“飞机”节点。MCPL 方法的第一步中节点路径概率计算详细步骤如下:首先,使用逻辑回归计算树结构中到达每一个节点路径的逻辑回归概率[11]。逻辑回归是一种基于概率论的线性分类模型,是描述因变量与一个或多个解释变量之间关系的组成部分。具体实现如文献[10]。第 $l-1$ 层的第 i 类节点路径逻辑回归概率计算公式如下:

$$p(x)_i^{l-1} = [1 + e^{-(w_i^T x + b_i)}]^{-1} \tag{1}$$

其中,$\boldsymbol{w}$ 是特征权重矩阵,b 表示偏移量。使用极大释然估计方法对训练过程中的权重矩阵 $\boldsymbol{w}$ 和偏移量 b 进行估计。在第 $l-1$ 层,前 k 个预选类别概率集为 $\boldsymbol{P}_k^{l-1} = \{p(x)_1^{l-1}, p(x)_2^{l-1}, \cdots, p(x)_k^{l-1}\}$。对于 $l-1$ 层的第 i 类,它的第 j 个孩子节点定义为 $c_{i,j}$,所以 $c_{i,j}$ 处于第 l 层。节点 $c_{i,j}$ 的路径逻辑回归概率计算公式如下:

$$p(x)_{c_{i,j}}^{l-1} = [1 + e^{-(w_{c_{i,j}}^T x + b_{c_{i,j}})}]^{-1} \tag{2}$$

其中,处于 $l-1$ 层的第 i 类节点,它的孩子节点的集合是 $\boldsymbol{C}_i$,$c_{i,j} \in \boldsymbol{C}_i$。

然后使用传统的 k - means 聚类方法给每个节点所包含的样本聚为一类,聚类中心代表节点的位置信息。定义 h_i^{l-1} 为 $l-1$ 层的第 i 类节点的位置信息,它的孩子节点 $c_{i,j}$ 的位置信息用 h_j^l 表示。所以 $l-1$ 层的第 i 类节点到 l 层的第 j 类节点的路径长度计算公式如下:

$$H_{i,j} = |h_i^{l-1} - h_j^l|, H_{i,j} \in \boldsymbol{H}_i \tag{3}$$

其中,处于 $l-1$ 层的第 i 类节点,它与孩子节点的路径长度集合是 $\boldsymbol{H}_i$。根据邻近性原则,选择下一层节点时,路径的长度越短,被选中的概率越大。为了约束路径长度对概率的影响,引入参数 $\lambda \in [0.1, 0.5]$。λ 的大小代表了路径长度在分类时的重要性。因此结合公式 1 和公式 3,l 层的第 j 类节点最终的概率计算公式如下:

$$p(x)_j^l = \frac{p(x)_i^{l-1} \times p(x)_{c_{i,j}}^l}{1 + \lambda \times H_{i,j}} \tag{4}$$

根节点的概率为 1。例如在图 1 中设置参数 $\lambda = 0.1$, $k = 2$。根据公式 1 得到 $p(\text{猫科})_{\text{猫科,R}}^2 = 0.36$, $p(\text{狮子})_{\text{猫科,狮子}}^3 = 0.3$,通过公式 3 得到 $H_{\text{R,猫科}} = 3$, $H_{\text{猫科,狮子}} = 1$。因此根据公式 4,狮子节点的最终预测概率计算过程为:

$$p(\text{狮子})_{\text{狮子}}^3 = \frac{p(\text{R})_{\text{R}}^1}{1 + \lambda \times H_{\text{R,猫科}}} \times \frac{p(\text{猫科})_{\text{猫科}}^2 \times p(x)_{c_{\text{猫科,狮子}}}^3}{1 + \lambda \times H_{\text{猫科,狮子}}} = \frac{1}{1 + 0.1 \times 3} \times \frac{0.36 \times 0.3}{1 + 0.1 \times 1} = 0.075$$

1.3 多路径选择

MCPL 方法的第二步是在树结构的每一层依次选取 k 个概率最大的路径,递归到最底层节点。详细步骤如下:首先,使用第 l 层 i 节点与其父节点的路径概率,融合路径长度约束来更新概率集,其计算公式如下:

$$\boldsymbol{P}_{\boldsymbol{C}_i}^l = \frac{p(x)_i^{l-1} \times p(x)_{c_{i,j}}^l}{1 + \lambda \times H_{i,j}}, 1 \leqslant j \leqslant |\boldsymbol{C}_i| \tag{5}$$

然后对概率集 $\boldsymbol{P}_{\boldsymbol{C}_i}^l$ 进行排序。选取前 k 个概率最大的类节点。将所选的类节点定义为 $\boldsymbol{P}_{k_{\boldsymbol{C}_i}}^l = \max_k \boldsymbol{P}_{\boldsymbol{C}_i}^l$。通过对 $l-1$ 层选择 k 个类别,迭代地执行以上过程。通过集合的并操作获得第 l 层所有的预选类别。第 l 层有 k^2 个类别,其集合表示如下:

$$\boldsymbol{Y}_{k_{\boldsymbol{C}_1}}^l \cup \boldsymbol{Y}_{k_{\boldsymbol{C}_2}}^l \cup \cdots \cup \boldsymbol{Y}_{k_{\boldsymbol{C}_k}}^l \rightarrow \boldsymbol{Y}_{k^2}^l \tag{6}$$

$$\boldsymbol{P}_{k_{\boldsymbol{C}_1}}^l \cup \boldsymbol{P}_{k_{\boldsymbol{C}_2}}^l \cup \cdots \cup \boldsymbol{P}_{k_{\boldsymbol{C}_k}}^l \rightarrow \boldsymbol{P}_{k^2}^l \tag{7}$$

最后,当第 l 层为最底层时,将 k^2 个类别的概率进行比较,选择第 l 层的前 k 个概率最大的预选类别。公式表示如下:

$$\boldsymbol{P}_k^l = \max_k \boldsymbol{P}_{k^2}^l \tag{8}$$

通过对树结构的每一层不断地递归选择,最后在最底层选择前 k 个概率最大的叶子节点。将选取的前 k 个概率最大的叶子节点放到 RF(Random Forest)分类器中进行分类[12]。例如,在图 1(c)中定义路径 $k = 2$, $\lambda = 0.1$。从第 1 层根节点出发,在第二层选择0.28 和 0.26 这两条节点的路径。再从 0.28 节点出发,在最底层选择 0.075 和 0.071 两条路径。从 0.26 节点出发,在最底层选择 0.067 和 0.073 两条路径。最后在最底层 0.075、0.071、0.067 和 0.073 四个可能的节

点中选择排名靠前的 0.075 和 0.073 两个节点放到 RF 分类器中进行分类。算法 1 描述了 MCPL 的具体步骤。

Algorithm1:MCPL method process

Input: $\boldsymbol{Y}_k^{l-1}=\{y_1^{l-1},y_2^{l-1},\cdots,y_k^{l-1}\}$ select the top k categories at layer $l-1$. The combined probability of logistic regression and path length corresponding to them $\boldsymbol{P}_k^{l-1}=\{p(x)_1^{l-1},p(x)_2^{l-1},\cdots,p(x)_k^{l-1}\}$. The bottom layer is the l.

Output: The first k prediction class set $\boldsymbol{Y}_k^l$ and probability sets $\boldsymbol{P}_k^l$ of layer l.

1. Possible class sets $\boldsymbol{Y}_{k^2}^l$ and probability sets $\boldsymbol{P}_{k^2}^l$ at level $l-1$;
2. for $i=1$ to k do
3. Update the set of all possible child categories of $\boldsymbol{Y}_k^{l-1}$ by Eq. (4);
4. Obtain the first k most probable sets $\boldsymbol{Y}_{k_c}^l$ by $\boldsymbol{P}_{k_c}^l=\max_k \boldsymbol{P}_{k^2}^l$;
5. end for
6. $\boldsymbol{P}_{k^2}^l=\cup_{i=1}^k \boldsymbol{P}_{k_c}^l$;
7. $\boldsymbol{Y}_{k^2}^l=\cup_{i=1}^k \boldsymbol{Y}_{k_c}^l$;
8. Selection of set $\boldsymbol{P}_k^l$ from set $\boldsymbol{P}_{k^2}^l$ at level l by Eq. (8);
9. Selection of category $\boldsymbol{Y}_k^l$ from category $\boldsymbol{Y}_{k^2}^l$ at level l;
10. Return $\boldsymbol{P}_k^l$ and $\boldsymbol{Y}_k^l$;

2　实验设置

2.1　实验数据集

该文将提出的融合路径长度的多路径层次标签分类方法(MCPL)在六个数据集上进行对比实验。其中包括两个蛋白质数据集(DD,F194)和三个图像数据集(Car196,VOC,CLEF),以及一个 UCI(University of California-Irvine)数据集(Bridges)。每次实验时并未明确地划分训练集和测试集,使用十次交叉验证随机选 80% 数据作为训练集,20% 数据作为测试集。数据集的具体信息如表 1 所示。

表 1　六个数据集的具体信息

dataset	sample	feature	leaf node	node	depth
DD	3 625	473	27	32	3
F194	8 525	473	194	202	3
Car196	7 541	4 096	196	206	3
VOC	12 283	1 000	20	30	5
CLEF	9 307	80	63	88	4
Bridges	108	11	8	11	3

2.2　对比方法

自顶向下逻辑回归(TDLR):这种分层方法用于类别预测和分类器训练。通过逻辑回归方法,递归地为每一层选择最佳子节点,直到到达最底层。

分层局部贝叶斯风险最小化(HLBRM)[13]:这种方法通过设计嵌入停止策略的框架来解决分层分类中遇到的风险问题。在信息不足的情况下,它将预测样本停止在粗类别的内部节点,而不是错误预测的叶节点。

分层 N 最佳路径(HNBP-1)[14]:这种方法使用最佳路径方法来避免错误传播。将类别预测问题转化为多路径搜索问题。它需要遍历整个视觉树,计算一个级别中所有类的所有边得分,并找到具有最大联合概率的路径。

不平衡类的成本敏感分层分类(CSHCIC)[15]:这是一种解决不平衡类的成本敏感分层分类方法。它构建了一个成本敏感的因素,以平衡多数类和少数类之间的关系。即通过设置一个阈值,当节点的概率不满足阈值时,对节点进行惩罚。

基于粒计算的多路径选择层次分类(HCMP-RF)[10]:利用逻辑回归的概率,对于选择的每个节点均选择 k 条路径,递归到叶子节点,最后用 RF 分类器进行分类。

最大化类间独立性和最小化类内冗余的分层分类特征选择(HFS-MIMR)[7]:首先,利用树结构中类的层次依赖性作为结构关系的正则化项,最大化类结构中不相关类之间的独立性。其次,将特征相关性转化为特征关系正则化的数学表示,在保证稀疏性的前提下,最小化类内的冗余度。最后,将两个正则化项统一为一个层次化的特征选择方法,以权衡结构和特征之间的关系。

2.3　评价指标

在研究过程中使用三项具有代表性的评价指标,将提出的 MCPL 方法与其他相关方法展开对比分析。其中包括 F_H、TIE 两个层次分类指标和 Acc 分类准确率指标。关于这些指标的详细信息如下:

(1)层次分类指标 F1-measure(F_H)[16]:F_H 是准确率 P_H 和召回率 R_H 的一种优化整合。具体计算公式如下:

$$P_H=\frac{|Y_F\cap\bar{Y}_F|}{|Y_F|} \tag{9}$$

$$R_H=\frac{|Y_F\cap\bar{Y}_F|}{|\bar{Y}_F|} \tag{10}$$

$$F_H=\frac{2\times P_H\times R_H}{P_H+R_H} \tag{11}$$

其中,Y_F 表示真实标签的集合,$\bar{Y}_F$ 表示预测标签的集合。F_H 的值越高表示预测结果越准确。

(2)层次结构诱导误差 TIE:在层次分类中不同的误差造成的代价不一样。TIE 通过计算真实类别 Y_F 与预测类别 $\bar{Y}_F$ 的距离来判断预测结果的有效性。计算方式如下:

$$\mathrm{TIE}(y,\bar{y})=|E_H(y,\bar{y})| \tag{12}$$

其中,$|E_H(y,\bar{y})|$ 是真实标签 y 到预测标签 $\bar{y}$ 的结构

距离,值越小,分类效果越好。

(3)样本分类准确率(Acc):用来评估模型对样本进行分类的能力,计算方式如下:

$$A_{cc} = \frac{|Y_F \cap \bar{Y}_F|}{|Y_F|} \tag{13}$$

其中, $|Y_F \cap \bar{Y}_F|$ 表示分类正确的数量, $|Y_F|$ 表示总样本的数量, A_{cc} 是样本分类的准确率,值越大分类效果越好。

3 实验结果分析

3.1 不同方法的性能比较

该文分析了七种方法在六个基准数据集上的 Acc 、F_H 和 TIE 。括号内数字代表该方法在对应数据集的排名;“-”表示相应的分层分类方法不适用于对应的数据集。表标题中,标记“↑”代表该指标值越大,性能越好;“↓”表示指标值越小越好。最好的结果用粗体表示,次优结果用下划线标出。

表 2 呈现出七种方法在六个数据集上的 Acc 。MCPL 与其他六种方法在 DD、F194、Car196、VOC、CLEF 和 Bridges 数据集上最好的结果相比,分别提高了 0.61%、1.53%、0.17%、0.81%、0.95% 和 10.5%。因此证明在自顶向下分层分类时融合路径的长度信息可以提高分类的准确率。七种方法在六个数据集上的 F_H 如表 3 所示。MCPL 与其他六种方法在 DD、F194、Car196、VOC、CLEF 和 Bridges 数据集上最好的结果相比,分别提高了 0.21%、0.60%、0.16%、0.30%、0.61% 和 0.25%。由此表明在自顶向下分层分类时融合路径长度能有效减少数据分类中的假阳性问题。

七种方法在六个数据集上的 TIE 如表 4 所示。表中括号内的数值代表每个方法在该数据集上的排名,其中最后一行代表每个方法在所有数据集上的平均排名。鉴于 CSHCIC 方法在 VOC、CLEF 和 Bridges 这三个数据集上不具备适用性,故默认其在相关评估中排名处于末位。在其余方法里,通过复现所得的 HFS-MIMR 方法,因不适用于相关指标,在排名中紧随 CSHCIC 方法之后,同样处于较低位次。HLBRM 方法在 Bridges 数据集上的 TIE 优于 MCPL 方法。但是在其他数据集上,MCPL 方法均取得最优结果。综上所述,在自顶向下分层分类时融合路径的长度信息能够在大部分数据集上减少层次结构诱导误差。

表 2 七种先进方法的 Acc (↑) %

Dataset	TDLR	HLBRM	HNBP-1	HFS-MIMR	CSHCIC	HCMP-RF	MCPL
DD	75.72	67.77	74.62	81.82	76.94	84.14	**84.66**
F194	48.71	17.92	44.98	59.1	52.9	58.48	**59.38**
Car196	68.66	54.59	68.76	66.44	68.74	68.66	**68.88**
VOC	38.47	33.33	38.35	37.8	-	40.39	**40.72**
CLEF	76.29	76.4	76.48	80.42	-	79.71	**80.47**
Bridges	57.56	50	54.57	65.13	-	60.63	**67.05**

表 3 七种先进方法的 F_H (↑) %

Dataset	TDLR	HLBRM	HNBP-1	HFS-MIMR	CSHCIC	HCMP-RF	MCPL
DD	89.03	88.65	88.66	91.67	89.47	93.44	**93.64**
F194	77.31	75.52	75.72	81.16	78.89	82.39	**82.89**
Car196	83.57	77.5	83.61	78.47	83.55	83.57	**83.75**
VOC	65.2	64.92	65.19	64.43	-	65.57	**65.77**
CLEF	85.71	81.28	85.78	87.35	-	87.72	**88.26**
Bridges	78.23	78.33	76.08	82.54	-	79.32	**82.75**

表 4 七种先进方法的 TIE (↓) %

Dataset	TDLR	HLBRM	HNBP-1	HFS-MIMR	CSHCIC	HCMP-RF	MCPL
DD	238.6 (5)	222 (3)	246.6 (6)	302 (7)	229 (4)	142.8 (2)	**138.4 (1)**
F194	1 160.6 (5)	1 034 (3)	1 242 (6)	1 606.4 (7)	1 079.8 (4)	900.8 (2)	**875.2 (1)**
Car196	743.2 (3)	991 (6)	741.4 (2)	2 104.4 (7)	744.2 (5)	743.2 (4)	**735.4 (1)**
VOC	2 919.9 (4)	2 803 (3)	2 929.1 (5)	3 567.8 (6)	-(7)	2 768.1 (2)	**2 751.6 (1)**
CLEF	1 032.1 (4)	1 101 (5)	1 026.7 (3)	1 644.6 (6)	-(7)	872.3 (2)	**840.7 (1)**
Bridges	12.9 (4)	**9 (1)**	14.2 (5)	15.7 (6)	-(7)	12.1 (3)	9.2 (2)
Ave. Rank	4.1	3.5	4.5	6.5	5.6	2.5	**1.1**

3.2　Friedman 检验

为了验证实验结果的合理性,使用 Friedman 检验来验证多种算法之间是否存在显著性差异[17]。给定 k 个算法和 N 个数据集,r_j^i 是第 j 个算法在第 i 个数据集上评价指标的排名。在零假设下所有方法都是等价的,Friedman 统计量定义为:

$$F_F = \frac{(N-1)\tau_F^2}{N(k-1)-\tau_F^2} \tag{14}$$

在公式 14 中:

$$\tau_F^2 = \frac{12N}{k(k+1)}\left[\sum_{i=1}^{n} R_i^2 - \frac{k(k+1)}{4}\right] \tag{15}$$

表 5 中展示每个评价指标的 Friedman 统计值 F_F 以及相应的临界值。结果表明,在显著性水平 $\alpha = 0.05$ 时,每个指标都明确拒绝所有方法具有相同性能的原假设。为了进一步比较不同方法之间的性能差异,使用 Bonferroni－Dunn 检验[18],如图 2 所示。当 MCPL 与其它方法之间的距离超出临界值,表明两种方法之间的差异显著。临界差异(Critical Difference, CD)值定义为:

$$C_D = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{16}$$

其中,C_D 表示临界差异。

表 5　Friedman 统计每个评价指标 F_F 和临界值

Metric	F_F	Critical value
Acc	6.506 8	
TIE	9.911 2	2.420 5
F_F	12.142 9	

图 2 描绘了显著性水平为 0.05,$q_\alpha = 2.948$ 的情况下六个数据集的 Bonferroni－Dunn 检验结果。根据表 2 ~ 4 所示,$k = 7$,$N = 6$。所以 $C_D = 3.676\ 8$。

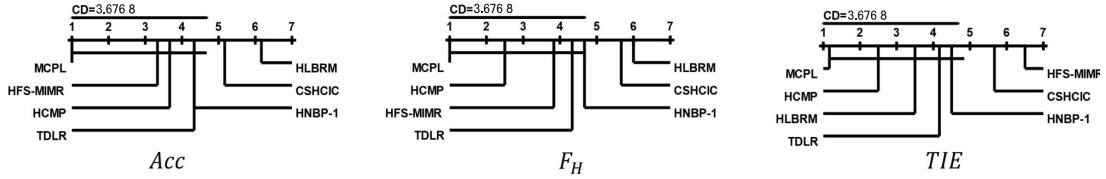

图 2　MCPL 与其他方法的比较(通过 Bonferroni－Dunn 检验)

在图 2 的 Acc 图中,MCPL 方法仅与 HLBRM 方法和 CSHCIC 方法有显著的差异性表现,但从整体性能评估的角度来看,MCPL 方法的排名处于领先地位。在图 2 的 F_H 图中,MCPL 方法与 HLBRM 方法、CSHCIC 方法之间的差异显著。特别地,MCPL 方法的 CD 线与 HNBP－1 方法的曲线相交,由此可判定 MCPL 方法和 HNBP－1 方法在特定的显著性检验标准下无显著差异。即便如此,MCPL 方法在排名中仍位居首位。在图 2 的 TIE 图中,MCPL 方法相对于 HFS－MIMR 方法和 CSHCIC 方法表现出显著的差异特性。尽管 MCPL 方法与 HCMP 方法、HLBRM 方法、TDLR 方法以及 HNBP－1 方法在显著性检验方面未表现出明显的差异结果,但从整体的效能评价与排名体系来看,MCPL 方法位列第一。

3.3　参数敏感性分析

对控制路径长度影响的参数 λ 进行敏感性分析时,设置控制路径长度的参数 $\lambda \in [0.1, 0.5]$,步长为 0.1。如图 3 所示,描述了 MCPL 方法在六个数据集上所对应的三项指标结果。在 DD、F194、Car196 和 CLEF 数据集上,设置路径数 $k = 3$,RF 分类器中分类树的数量为 30。在 VOC 数据集上,设置路径数 $k = 2$,RF 分类器中分类树的数量为 10。在 Bridges 数据集上,设置路径数 $k = 3$,RF 分类器中分类树的数量为 10。如图 3 所示,对于 Acc 指标,$\lambda = 0.1$ 时在 F194、Car196、VOC 和 CLEF 数据集上取得最优结果;$\lambda = 0.3$ 时在 DD 数据集上取得最优结果;$\lambda = 0.4$ 时在 Bridges 数据集上取得最优结果。对于 F_H 指标,$\lambda = 0.1$ 时在 F194、Car196、VOC、CLEF 和 Bridges 数据集上取得最优结果;$\lambda = 0.3$ 时在 DD 数据集上取得最优结果。对于 TIE 指标,$\lambda = 0.1$ 在 F194、Car196、VOC 和 CLEF 数据集上取得最优结果;$\lambda = 0.3$ 时在 DD 数据集上取得最优结果;$\lambda = 0.4$ 时在 Bridges 数据集上取得最优结果。

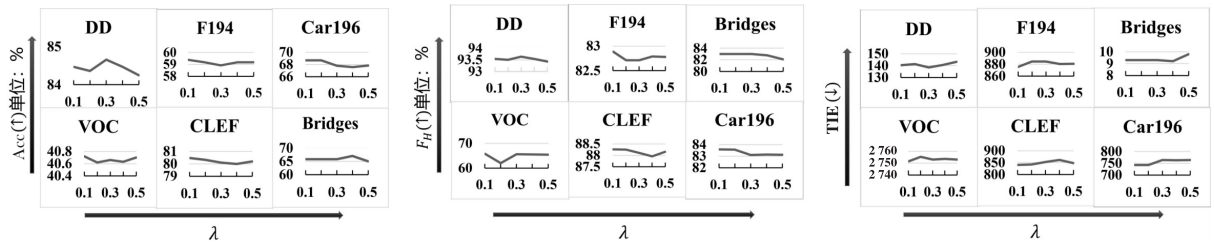

图 3　参数 λ 的敏感性分析

3.4 选择不同数量路径的结果对比

着重分析 MCPL 方法因选择不同路径数 k，在六个数据集上的三项指标结果。其中，MCPL-1 表示路径数 $k=1$ 时的 MCPL 方法，MCPL-2 表示路径数 $k=2$ 时的 MCPL 方法，MCPL-3 表示路径数 $k=3$ 时的 MCPL 方法。根据控制变量法思想，只改变路径数 k。在 DD 数据集上设置参数 $\lambda=0.3$，RF 分类器中树为 30。在 F194、Car196 和 CLEF 数据集上设置参数 $\lambda=0.1$，RF 分类器中树为 30。在 VOC 数据集上设置参数 $\lambda=0.1$，RF 分类器中树为 10。在 Bridges 数据集上设置参数 $\lambda=0.4$，RF 分类器中树为 10。

图 4 记录了 MCPL 方法在六个数据集选择不同路径数的 Acc，值越高越好。MCPL 方法在六个数据集上选择不同路径数的 F_H 如图 5 所示，值越高越好。图 6 描绘了 MCPL 方法在六个数据集上选择不同路径数的 TIE，值越低越好。由图 4～6 所示，在 DD、F194、CLEF 和 Bridges 数据集上，路径数 $k=3$ 时，取得最优结果。在 Car196 和 VOC 数据集上，路径数 $k=2$ 时取得最优结果。

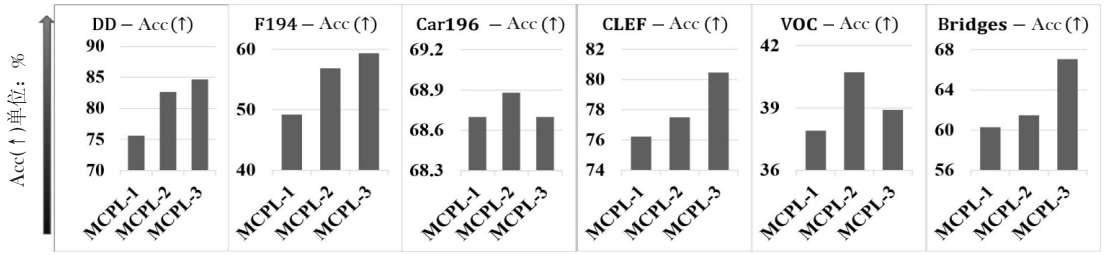

图 4 MCPL 三种路径选择方法在六个数据集上的 Acc

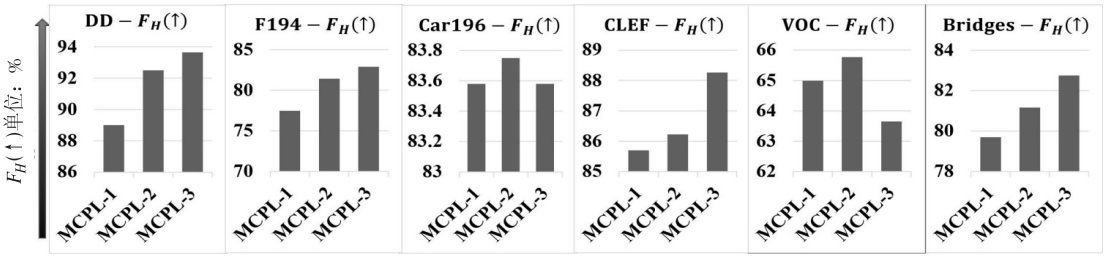

图 5 MCPL 三种路径选择方法在六个数据集上的 F_H

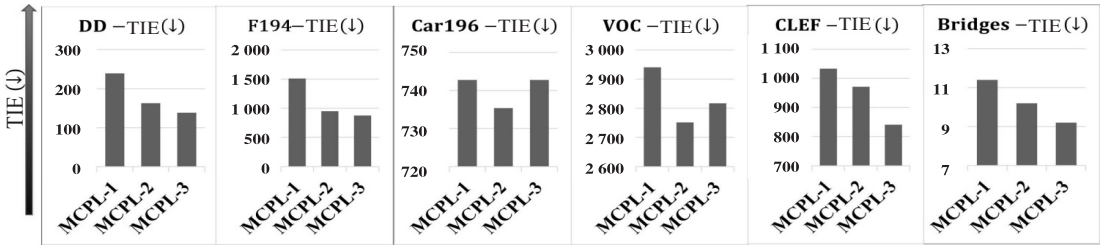

图 6 MCPL 三种路径选择方法在六个数据集上的 TIE

3.5 时间复杂度分析

该文根据相关文献的伪代码及其复现实验进行复杂度分析。MCPL 方法的时间复杂度高于 HCMP 方法和 TDLR 方法。因为 MCPL 方法相较于 HCMP 方法增加了聚类过程，相较于 HCMP 方法增加了路径选择和聚类过程。MCPL 方法的时间复杂度低于 CSHCIC 方法、HNBP 方法、HFS – MIMR 方法和 LBRM 方法。

相较于 CSHCIC 方法，MCPL 方法减少了时间复杂度较高的敏感因子更新操作；相较于 HNBP 方法，MCPL 方法省去了多层循环计算；相较于 HFS–MIMR 方法，MCPL 方法未采用特征选择的迭代计算；相较于 LBRM 方法，MCPL 方法避免了时间复杂度较高的信息熵和信息增益等操作。

3.6 RF 分类器参数分析

RF 分类器的关键参数为分类时所采用的分类树个数。在本小节中，针对不同数据集，对使用不同数量分类树的情况展开了分析，相关结果如表 6 所示，其中最优结果以粗体形式呈现。对于 DD 数据集，设置路径数 $k=3$，参数 $\lambda=0.3$。对于 F194 数据集，设置路径数 $k=3$，参数 $\lambda=0.2$。对于 Car196 数据集，设置路径数 $k=2$，参数 $\lambda=0.1$。对于 VOC 数据集，设置路径数 $k=3$，参数 $\lambda=0.1$。对于 CLEF 数据集，设置路径数 $k=2$，参数 $\lambda=0.2$。对于 Bridges 数据集，设置路径数 $k=3$，参数 $\lambda=0.4$。由表 6 中的结果可知，在 DD、F194、Car196 和 CLEF 数据集上，分类树的个数为 30 时，分类效果最好，在 VOC 和 Bridges 数据集上，分类树的个数为 10 时，分类效果最好。

表 6　MCPL 方法使用不同个数的分类树对数据集进行分类　%

指标	树个数/个	DD	F194	Car196	VOC	CLEF	Bridges
Acc (↑)/%	10	82.73	57.23	68.7	**40.72**	79.36	**67.05**
	20	84.31	58.74	68.7	40.63	80.07	64.12
	30	**84.66**	**59.38**	**68.88**	40.68	**80.47**	65.05
F_H (↑)/%	10	92.76	81.87	83.58	**65.77**	87.56	**82.75**
	20	93.44	82.54	83.58	65.7	88.08	82.19
	30	**93.64**	**82.89**	**83.75**	65.43	**88.26**	81.9
TIE (↓)	10	157.6	927.2	742.8	**2 751.6**	892.4	**9.2**
	20	142.8	893	742.8	2 757.2	854.3	9.7
	30	**138.4**	**875.2**	**735.4**	2 753	**840.7**	9.9

4　结束语

提出了一种融合路径长度的多路径层次标签分类方法 MCPL。与现有的多路径层次分类方法相比,MCPL 深度挖掘了分层结构所蕴含的节点位置信息,并通过节点位置信息计算节点间路径长度,为不同的路径赋权。在六个数据集上的实验结果表明,MCPL 有效提高了分类准确率,具有更好的层次分类性能。然而该方法也存在一些局限性。首先目前的 MCPL 方法仅适用于层次树结构,其次实验数据集的层次树结构都是已知的,未来将进一步探索更为复杂的层次结构构造和分类方法。

参考文献:

[1] 王金环,李宝敏. 基于 YOLO 模型的车流量实时采集系统研究[J]. 计算机技术与发展,2024,34(9):209-214.

[2] 于升正,程远志. 基于层次结构与多模块的海洋生物分类算法[J]. 计算机技术与发展,2024,34(11):36-42.

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

[4] GUO S,ZHAO H,YANG W. Hierarchical feature selection with multi-granularity clustering structure[J]. Information Sciences,2021,568:448-462.

[5] WANG Y,WANG Z,HU Q,et al. Hierarchical semantic risk minimization for large-scale classification[J]. IEEE Transactions on Cybernetics,2021,52(9):9546-9558.

[6] LIN Y,LIU H,ZHAO H,et al. Hierarchical feature selection based on label distribution learning[J]. IEEE Transactions on Knowledge and Data Engineering,2022,35(6):5964-5976.

[7] SHI J,LI Z,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] HUANG H,WANG Y,HU Q. Building hierarchical class structures for extreme multi-class learning[J]. International Journal of Machine Learning and Cybernetics,2023,14(7):2575-2590.

[9] ZHENG W,ZHAO H. Cost-sensitive hierarchical classification via multi-scale information entropy for data with an imbalanced distribution[J]. Applied Intelligence,2021,51(8):5940-5952.

[10] GUO S,ZHAO H. Hierarchical classification with multi-path selection based on granular computing[J]. Artificial Intelligence Review,2021,54(3):2067-2089.

[11] LIN Z. Enhanced GRU-based regression analysis via a diverse strategies whale optimization algorithm[J]. Scientific Reports,2024,14(1):25629.

[12] DJABALLAH S,SAIDI L,MEFTAH K,et al. A hybrid LSTM random forest model with grey wolf optimization for enhanced detection of multiple bearing faults[J]. Scientific Reports,2024,14(1):23997.

[13] WANG Y,HU Q,ZHOU Y,et al. Local bayes risk minimization based stopping strategy for hierarchical classification[C]//2017 IEEE international conference on data mining (ICDM). New Orleans:IEEE,2017:515-524.

[14] QU Y,LIN L,SHEN F,et al. Joint hierarchical category structure learning and large-scale image classification[J]. IEEE Transactions on Image Processing,2016,26(9):4331-4346.

[15] ZHENG W,ZHAO H. Cost-sensitive hierarchical classification for imbalance classes[J]. Applied Intelligence,2020,50(8):2328-2338.

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

[17] GONG K,LI G,GUO L,et al. Online streaming feature selection for high-dimensional small-sample data[J]. International Journal of Machine Learning and Cybernetics. DOI:10.1007/s13042-024-02416-9.

[18] 滕少华,卢建磊,滕璐瑶,等. 增强学习标签相关性的多标签特征选择方法[J]. 计算机应用研究,2024,41(7):2079-2086.