计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009771] [CSTR: 32024.14.csa.009771]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向标签共现和长尾分布的层级文本分类①

智　媛, 雷海卫, 张斌龙

(中北大学 计算机科学与技术学院, 太原 030051)
通信作者: 雷海卫, E-mail: lhw0312@nuc.edu.cn

摘　要: 针对当下层级文本分类模型尚未充分利用层级实例的标签信息以及缺乏对类别分布不平衡的处理这两方面问题, 本文提出一种面向标签共现和长尾分布的层级文本分类方法 (hierarchical text classification for label co-occurrence and long-tail distribution, LC-LTD), 对基于共享标签的文本全局语义和面向长尾分布的平衡损失函数进行研究. 首先, 设计一种基于共享标签的对比学习目标, 使具有更多共享标签的文本表示在特征空间中的语义距离更近, 引导模型生成具有判别性的语义表征; 其次, 引入分布平衡损失函数替换二进制交叉熵损失, 缓解层级分类固有的长尾分布问题, 提高模型的泛化能力. 在 WOS、BGC 两个公开数据集上将 LC-LTD 与当前多个主流模型进行比较, 结果表明所提方法具有更好的分类性能, 更适合处理层级文本分类任务.

关键词: 层级文本分类; 标签共现; 长尾分布; 对比学习; 平衡损失

引用格式: 智媛,雷海卫,张斌龙.面向标签共现和长尾分布的层级文本分类.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9771.html

Hierarchical Text Classification for Label Co-occurrence and Long-tail Distribution

ZHI Yuan, LEI Hai-Wei, ZHANG Bin-Long

(School of Computer Science and Technology, North University of China, Taiyuan 030051, China)

Abstract: There are two problems in existing hierarchical text classification model: underutilization of the label information across hierarchical instances, and lack of handling unbalanced label distribution. To solve these problems, this study proposes a hierarchical text classification method for label co-occurrence and long-tail distribution (LC-LTD) to study the global semantic of text based on shared labels and balanced loss function for long-tail distribution. First, a contrastive learning objective based on shared labels is devised to narrow the semantic distance between text representations with more shared labels in feature space and to guide the model to generate discriminative semantic representations. Second, the distribution balanced loss function is introduced to replace binary cross-entropy loss to alleviate the long-tail distribution problem inherent in hierarchical classification, improving the generalization ability of the model. LC-LTD is compared with various mainstream models on WOS and BGC public datasets, and the results show that the proposed method achieves better classification performance and is more suitable for hierarchical text classification.

Key words: hierarchical text classification (HTC); label co-occurrence; long-tail distribution (LTD); contrastive learning; balanced loss

在当前信息过载的时代, 随着文本数据指数级激增[1], 对这些错综复杂的数据进行有效组织和分类管理的标签规模也日益庞大[2]. 层级标签结构由于能够直观建模类别之间的隶属关系[3], 以及有助于为各种应用程

① 收稿时间: 2024-07-29; 修改时间: 2024-08-20; 采用时间: 2024-09-03; csa 在线出版时间: 2024-11-28

序和终端用户提供高效的信息检索[4]等优势, 在各大领域得到了广泛应用. 国际专利分类 (international patent classification, IPC)、学术论文分类、电商平台的商品分类等都是以层级结构组织的类别体系, 层级文本分类任务得到了广泛关注. 当面对主题多级化的大规模分类任务时, 现有分类方法难以满足实际生产和生活中对自动分类技术的应用需求. 本研究期望为实现具有层级标签的自动分类提供参考依据, 进一步提高层级文本分类的效率和准确性.

层级文本分类 (hierarchical text classification, HTC)[5]是文本分类领域中一项具有挑战性的任务, 旨在将文本分类到一组以层次化结构组织的标签集合中. 根据类别之间的关联, 层级分类的标签结构被建模为树 (tree), 其中每个结点对应一个标签, 上层 (浅层) 的标签表示粗粒度的概念, 下层 (深层) 的标签表示细粒度的概念. 分类时, 给定一条文本, HTC 预测该文本所属的所有类别, 通常匹配树结构中自上而下的一条或多条路径. 以图 1 展示的 WOS 论文分类任务为例, 一条文本被定位到类别“CS”和“Computer graphics”.

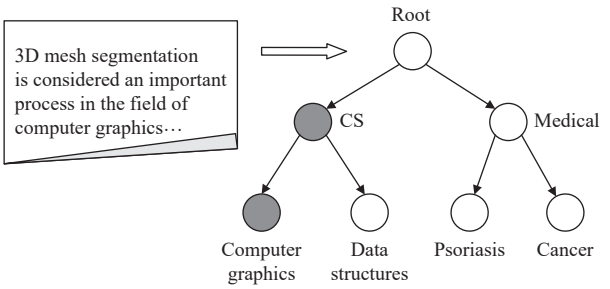

图 1 WOS 论文分类实例

由于单个文本通常与多个具有依赖关系的标签相关联, HTC 的关键在于如何利用标签层级指导模型更好地进行分类[6]. 目前主流的层级文本分类方法主要采用双编码器架构的模式, 利用文本编码器进行文本特征提取, 构建结构编码器对整体标签结构进行建模, 通过标签与文本特征的交互完成分类任务. 其中, HGCLR[7]在双编码器架构的基础上, 通过对比学习机制将标签层次集成到文本编码器, 使文本编码器可以学习独立地生成层级感知的文本表示, 提供了解决层级分类问题的另一种思路, 达到了较好的分类性能, 但其存在两方面问题有待进一步优化.

一方面, HGCLR 忽略了包含共享标签的实例对之间应该具有的语义相似性. 层级文本分类中, 实例对的共享标签数量越多, 其文本表示在特征空间中的语义距离应该越接近[8]. 因此, 不同于 HGCLR 只学习文本局部语义 (词特征) 的方法, 本文进一步研究基于共享标签的文本全局语义, 设计一种对比学习目标来更加充分地捕捉层级实例之间的高阶相关性.

另一方面, HGCLR 没有针对层级分类固有的类别不平衡问题加以优化. 由于父类标签的实例数量远远多于子类标签的实例数量, 层级分类问题的类别标签呈现长尾分布[9]. 如图 2, 分别统计了 WOS、BGC 各层标签在训练集中的实例分布情况, 展示了极端的长尾分布现象. 因此, 本文引入分布平衡损失替换常用的二进制交叉熵损失, 避免网络模型过多偏向头类标签, 提高下层尾类标签对文本的鉴别能力.

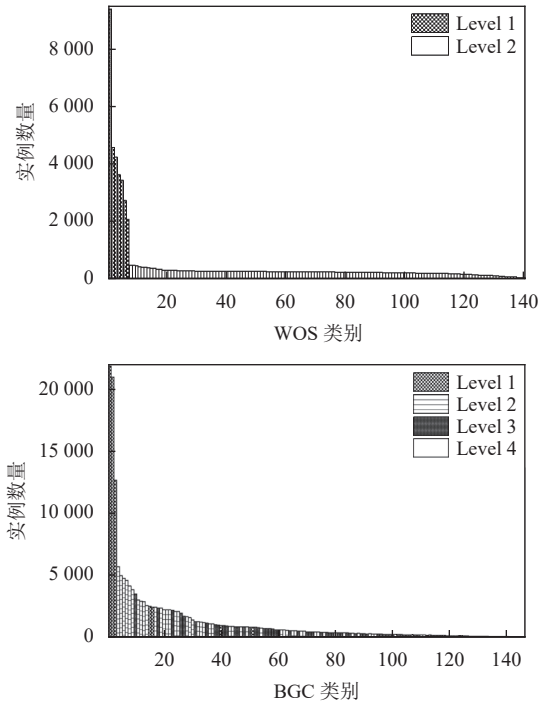

图 2 WOS、BGC 训练集分布

基于以上两方面的考虑, 本文在 HGCLR 模型的基础上提出一种面向标签共现和长尾分布的层级文本分类方法 (hierarchical text classification for label co-occurrence and long-tail distribution, LC-LTD). 聚焦于层级分类问题的两大特征, 充分利用训练实例中的丰富知识, 使用对比学习目标和面向长尾分布的平衡损失目标训练模型, 帮助模型进行更准确的预测.

最终, 将 LC-LTD 与目前主流的 HTC 模型在公开数据集 WOS、BGC 上进行一系列实验, 结果显示该模

型显著地超过其他模型. 与基线模型 HGCLR 相比, LC-LTD 在 WOS、BGC 数据集的 *Macro-F*1 分别提升了 0.74%、0.64%. 与 BERT 相比, LC-LTD 在两个数据集的 *Macro-F*1 分别提升了 0.7%、2.5%, *Micro-F*1 分别提升了 1.2%、0.61%. 此外, 通过分层表现分析, 验证了该模型在提高下层长尾标签的分类性能方面具有优势. 并且, 通过消融实验证明了各个模块的有效性.

1 相关工作

依据是否利用以及如何利用标签的整体层级结构, 层级文本分类的解决方法大致可以分为 3 种[10]: 平面分类、局部分类和全局分类.

平面分类方法将任务简化为一般的多标签分类问题, 仅对标签体系的最后一层进行分类[11,12], 忽略了分类层次的整体结构信息, 在 HTC 任务中表现不佳.

局部分类方法采取“分而治之”的思想, 将 HTC 分解为多个局部子问题, 一般以单个结点 (local classifier per node, LCN)[13]、单个层级 (local classifier per level, LCL)[14]或者单个父结点 (local classifier per parent node, LCP) 为单位, 自顶向下构建多个分类器, 通过聚合多个分类器的不同预测来获取最终分类结果. 然而, 其上层分类器的预测结果通常会影响下层分类器的预测, 不可避免地导致错误传播问题, 并且模型的可训练参数通常很大.

全局分类方法只包含一个分类器, 采用端到端的训练方式, 将标签层级结构隐式或显式地编码到网络结构中[15], 对所有标签进行一次性预测. Zhou 等人[16]提出了基于多标签注意力的 HiAGM-LA 和基于文本特征传播的 HiAGM-TP 两种层级感知全局模型, 构建文本编码器和结构编码器分别提取文本特征和标签特征, 再通过注意力机制或特征传播模块生成特定标签的文本特征. Chen 等人[17]提出 HiMatch, 将文本与标签的关系重新表述为语义匹配问题, 引入联合嵌入损失和匹配学习损失来显式约束文本语义与不同层级标签语义之间的相关性. Deng 等人[18]提出 HTCInfoMax, 引入文本-标签互信息最大化和标签先验匹配过滤不相关的信息, 学习更好的层级感知表示. Wang 等人[7]提出 HGCLR, 考虑到结构编码器对所有实例提供完全相同的层级标签表示, 尝试将层次结构集成到文本编码器, 在预测过程中直接丢弃结构编码器. 与局部分类方法相比, 上述全局方法有效捕获了标签的层级结构关系, 通过文本与标签特征的交互提高了层级分类的总体性能. 然而, 这些模型对实例标签信息的利用不够充分, 并且缺乏对长尾分布问题的处理, 在许多情况下具有局限性.

2 面向标签共现和长尾分布的层级文本分类方法

2.1 问题描述

定义$S=\{(D_i,Y_i)\}_{i=1}^{N}$是由 N 个“文本-标签对”组成的层级文本分类实例集合. D_i 表示第 i 个实例文本, $Y_i=(y_{i1}, y_{i2}, \cdots, y_{iC})\in\{0, 1\}^C$ 表示对应的多热标签向量. Y_i 中不止一个分量的值为 1, $y_{ij}=1$ 表示该实例的其中一个标签为 $l_j\in L$. HTC 的目标是学习一个分类模型, 在给定任意一条输入文本 D 的情况下, 从预定义的标签集 L 中预测出该条文本对应的多热标签向量 Y, 实现从输入文本到相关标签的映射 $(D, L)\rightarrow Y$.

将具有层级结构的候选标签集$L=\{l_1,\cdots,l_C\}$组织为预定义的有向无环图或者树结构$G=(V,B)$. 其中, $V=\{v_1,\cdots,v_C\}$是标签结点的集合, 非根结点 v_j 对应于标签 l_j , C 表示标签总数. 边集 B 表示标签结点之间的层级关系, 每个非根结点 v_j 与唯一的父结点 v_p 相关联.

2.2 模型总体架构

本文提出的 LC-LTD 模型总体架构如图 3 所示, 主要包括文本表示与标签层级建模、文本全局特征增强、文本局部特征增强、分类与目标函数 4 个部分. 该方法属于全局分类方法, 深入应用两种有监督的对比学习及其对层级分类表征的影响, 在训练阶段利用真实标签信息指导标签层级特征、文本全局特征和局部关键词特征的学习优化, 使得在预测阶段整个模型仅使用一个文本编码器就可以得到文本对应每个标签的分类概率, 获得良好的层级分类性能.

具体而言, 为了实现有效的特征学习, LC-LTD 模型利用标签与文本、标签与单词两种角度的关联性进行优化训练. 首先, 文本编码器 BERT 用来生成文本的全局语义表示. 与此同时, 将层级结构中的类别建立为标签结点, 类别名称初始化标签语义, 使用结构编码器 Graphormer 捕获标签结构关系, 学习聚合层级信息的标签表示. 其次, 设计基于共享标签的对比学习目标, 对文本全局特征进行约束, 使得实例对之间共享标签的数量越多, 其在特征空间中的语义距离越近. 再次,

通过注意力机制识别出文本中对真实标签重要的词构成增强文本, 并利用局部增强的对比学习, 提高原始文本与增强文本的特征相似性, 明确引导文本表征关注对分类重要的局部单词信息. 最后, 分类损失函数采用面向长尾分布的分布平衡损失, 与两个对比学习损失相结合作为训练目标来优化网络参数. 至此, 让模型自适应地学习文本、标签、单词的语义信息, 促进彼此的特征优化.

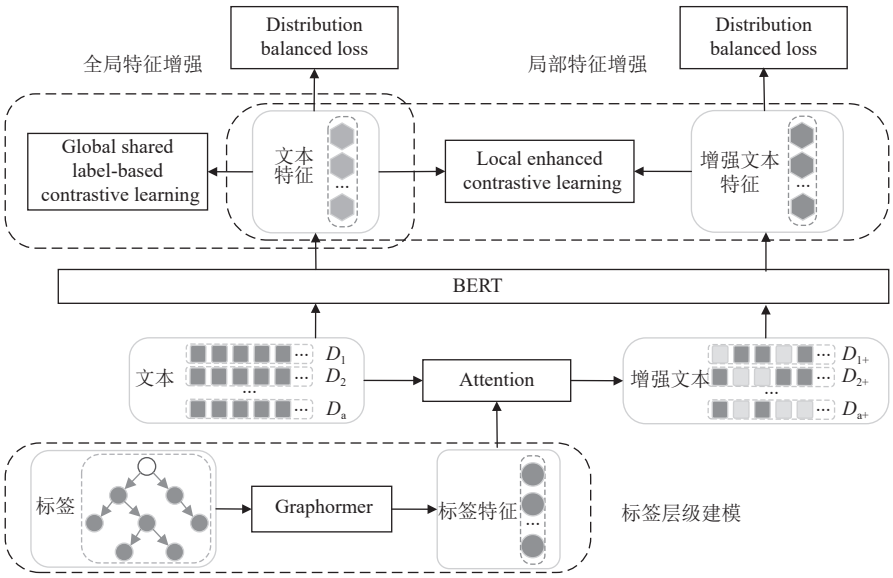

图 3 LC-LTD 模型总体框架

2.3 文本表示与标签层级建模

模型采用预训练 BERT[19]作为文本编码器. 给定输入文本序列 D_i = {[CLS], w_1,⋯, w_{n-2}, [SEP]}, BERT 对其进行编码, 生成 n 个分词的最终隐藏表示$H_i \in R^{n\times d}$.

$$H_i = \mathrm{BERT}(D_i) \tag{1}$$

结合上下文的文本全局语义表示为 $h_i = h_{[\mathrm{CLS}]}$.

对标签层级结构$G = (V, B)$的建模通过采用 Graphormer[20]作为结构编码器来实现, 能够有效捕获图中的层级结构信息和标签结点特征:

$$H_v' = \mathrm{Graphormer}(H_v) \tag{2}$$

其中, H_v = {h_{v_1},⋯, h_{v_C}} ∈ $R^{C\times d}$ 是初始标签特征. 标签结点 v_j 的初始特征 h_{v_j} 由名称嵌入 $name\_emb(v_j)$ 和标签嵌入 $label\_emb(v_j)$ 表示:

$$h_{v\_j} = name\_emb\left(v_j\right) + label\_emb\left(v_j\right) \tag{3}$$

其中, $name\_emb(v_j)$ 是标签名称的 BERT 分词嵌入平均, $label\_emb(v_j)$ 是用于存储标签依赖关系的可学习嵌入.

2.4 全局特征增强

为了帮助模型学习到具有判别性的全局语义表示以实现文本特征增强, 本文设计一种基于共享标签的对比学习目标, 对层级实例之间细粒度的相关性进行建模, 引导模型为具有更多共享标签的实例对生成更加接近的特征表示, 并推开不具有相同标签的实例对. 如图 4 所示, 若两个实例文本 D_1、D_2 的标签对应于{“CS”, “Computer graphics”}, D_3 对应于{“CS”, “Data structures”}, D_4 对应于{“Medical”, “Cancer”}, 则特征空间中应有距离不等式:

$$d(h_1,h_2) < d(h_1,h_3) < d(h_1,h_4) \tag{4}$$

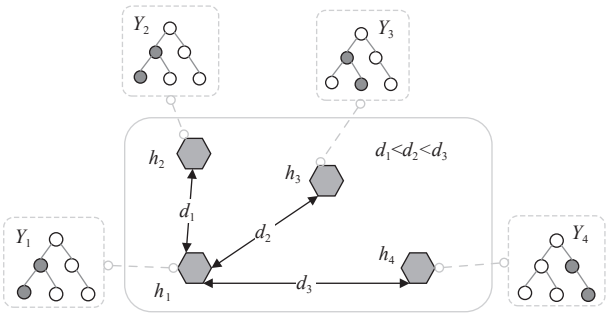

图 4 文本语义相似性

受 Su 等人[8]的启发, 设一个批次的实例数量为 a, 将实例 i 除自身外的其他实例定义为 $g(i) = \{k|k \in \{1,2,\cdots,a\}, k \neq i\}$, 每个实例对 (i, j) 之间基于共享标签

的对比学习损失L_{ij}^{G}可以计算为:

$$L_{ij}^{G}=-\beta_{ij}\log\frac{\exp\left(-d\left(h_i,h_j\right)/\tau\right)}{\sum_{k\in g(i)}\exp\left(-d\left(h_i,h_k\right)/\tau\right)} \tag{5}$$

$$\beta_{ij}=\frac{C_{ij}}{\sum_{k\in g(i)}C_{ik}},\ C_{ij}=Y_i^{\mathrm{T}}\cdot Y_j \tag{6}$$

其中, β_{ij} 是基于标签相似性的动态系数, $d(h_i, h_j)$ 是两者之间的欧氏距离, h_i 是实例文本 D_i 的全局语义表示, τ 是对比学习的温度系数. C_{ij} 表示实例 i 和实例 j 共享标签的数量.

整个批次内, 基于共享标签的对比损失 L^G 是所有实例对的和:

$$L^{G}=\sum_i\sum_{j\in g(i)}L_{ij}^{G} \tag{7}$$

对于实例对 (i, j), 其共享标签的数量 C_{ij} 越多, 动态系数 β_{ij} 越大, 从而导致更大的损失L_{ij}^{G}, 距离 $d(h_i, h_j)$ 将被优化得更近; 反之, 如果共享的标签较少或没有共享标签, 距离将相对优化得更远.

2.5 局部特征增强

由于文本的不同部分与各个标签的相关程度不同, 因此利用标签与单词语义的关联可进一步实现 BERT 词向量的特征优化. 依托 HGCLR[7]的正样本生成及对比学习机制, 对于文本序列 D_i, 首先通过注意力机制和 Gumbel_*Softmax*[21]提取与真实标签 Y_i 最相关的关键分词组成增强文本子序列 D_{i+}. D_{i+} 的特征表示采用与原始文本共享参数的文本编码器. 然后利用 *NT-Xent* 损失[22]使原始文本表示 h_i 与增强文本表示 h_{i+} 在降维后的特征空间中彼此靠近, 引导文本表征更关注对分类重要的词汇, 达到局部特征增强的作用. 此时, 一个批次的局部增强对比损失 L^L 为:

$$L^{L}=\frac{1}{2a}\sum_{i=1}^{2a}NT\text{-}Xent(\tilde{h}_i,\tilde{h}_{i+}) \tag{8}$$

$$\tilde{h}_i=W_2ReLU(W_1h_i) \tag{9}$$

其中, $\tilde{h}_i$、$\tilde{h}_{i+}$ 分别为投影到对比学习潜在空间的原始文本表示和增强文本表示. 由于每条原始文本对应一条增强文本, 故构成$\{i, i+\}_{i=1}^{a}$共 $2a$ 条文本.

2.6 分类与目标函数

2.6.1 总训练目标

模型将原始文本表示 h_i 和增强文本表示 h_{i+} 送入全连接层, 分别得到其对于标签的逻辑输出 z_i 和 z_{i+}, 进而完成分类工作.

$$z_i=h_iW^C+b^C \tag{10}$$

模型的总训练目标 L 是分类损失和两种对比学习损失的总和.

$$L=L_{DB}+L_{DB+}+\lambda_GL^G+\lambda_LL^L \tag{11}$$

其中, L_{DB}、L_{DB+} 分别是由原始文本和增强文本得到的分类损失, λ_G、λ_L 分别是全局基于共享标签的对比学习和局部增强对比学习的权重因子.

2.6.2 分布平衡损失

分布平衡损失 (distribution balanced loss, DBLoss)[23]通过对二进制交叉熵损失 (binary cross-entropy loss, BCELoss) 进行重平衡权重和负容忍正则化两处改进, 允许罕见的“文本-标签对”直观获得合理的“关注”, 能够有效缓解长尾分布问题, 提高模型的泛化能力.

其一, 重平衡权重方法基于类别频率对每个实例的单个真实标签重新加权, 显式地为“类别频率高”的实例标签赋予较低的权重, 使得原始的数据分布接近平衡分布.

令 n_j 表示包含标签 j 的实例数量. 实例 i 基于其单个标签 y_{ij} 被采样的概率记为 $P(x_{ij})$. 同时, 由于层级分类的实例包含多个标签, i 被采样的总概率记为 $P(x_i)$. 具体计算方式如下:

$$P(x_i)=\frac{1}{C\times\sum_{j\in\{y_{ij}=1\}}n_j} \tag{12}$$

$$P\left(x_{ij}\right)=\frac{1}{C\times n_j},\ n_j=\sum_{i=1}^{N}y_{ij} \tag{13}$$

那么, 实例 i 中类 y_{ij} 的重平衡权重 r_{ij} 可以由采样概率的归一化计算. 另外, 为了使优化过程稳定, 利用平滑函数将 r_{ij} 映射到适当的值范围 $[\alpha, \alpha+1]$ 内.

$$r_{ij}=\frac{P\left(x_{ij}\right)}{P(x_i)},\ \hat{r}_{ij}=\alpha+Sigmoid\left(\beta\times\left(r_{ij}-\mu\right)\right) \tag{14}$$

其二, 负容忍正则化方法以不同方式处理同一实例的正负类. 设计固定的类别偏置 ε_j 以降低尾部类的阈值, 并引入缩放因子 λ 对负类进行线性缩放, 控制对 z_{ij} 的“容忍度”, 避免负标签的过度抑制.

将分类器的逻辑输出表示为 $z_i=[z_{i1},\cdots,z_{iC}]$, 实例 i 属于类别 j 的概率 q_{ij} 定义为:

$$\begin{cases}q_{ij}=Sigmoid\left(z_{ij}-\varepsilon_j\right), & y_{ij}=1\\ q_{ij}=Sigmoid\left(\lambda\left(z_{ij}-\varepsilon_j\right)\right), & y_{ij}=0\end{cases} \tag{15}$$

其中, 类别偏置 ε_j 通过类 j 在数据集中的出现频率 $p_j = n_j/N$ 和比例因子 κ 计算:

$$b_j = -\log\left(\frac{1}{p_j}-1\right),\ \varepsilon_j = -\kappa b_j \tag{16}$$

最终, 基于广泛使用的 BCE 损失, 分布平衡损失 L_{DB} 将重平衡权重和负容忍正则化方法相结合:

$$L_{DB} = -\sum_{i=1}^{a}\sum_{j=1}^{C}\left[y_{ij}\log\left(q_{ij}\right)+\frac{1}{\lambda}\left(1-y_{ij}\right)\log\left(1-q_{ij}\right)\right]\times\hat{r}_{ij} \tag{17}$$

L_{DB} 即一个批次内 a 条原始文本的分类损失. 同时, 对于所构造的增强文本, 其分类损失 L_{DB+} 可以通过在式 (17) 中将 i 替换为 $i+$ 类似地进行计算.

3 实验

3.1 数据集

实验在 Web of Science (WOS-46985)[24]、blurb genre collection (BGC)[25]这两个来自不同领域的公共数据集上进行评估. WOS 由 Web of Science 已发表学术论文的摘要组成, BGC 由书籍简介组成. WOS 用于单路径 HTC, BGC 用于多路径 HTC. 表 1 列出了两个数据集的相关信息.

表 1 数据集信息

数据集	WOS	BGC
层级深度	2	4
标签总数	141	146
各层标签数	7-134	7-46-77-16
每个实例的平均标签数	2.0	3.01
训练集大小	30070	58715
验证集大小	7518	14785
测试集大小	9397	18394

3.2 评价指标

由于层级分类问题本质上属于多标签分类的特例, 许多研究人员选择使用在多标签分类场景中广泛使用的标准评估指标 *Micro-F*1 和 *Macro-F*1, 来衡量层级文本分类模型的性能[26]. 具体计算方式如下:

(1) 对于单个标签 $l_j \in L$:

$$F1_j = 2\times\frac{Precision_j\times Recall_j}{Precision_j + Recall_j} \tag{18}$$

$$Precision_j = \frac{TP_j}{TP_j+FP_j},\ Recall_j = \frac{TP_j}{TP_j+FN_j} \tag{19}$$

其中, TP 表示标签被正确预测的次数, $(TP+FP)$ 代表标签在模型预测中出现的次数, $(TP+FN)$ 代表标签在真实标签中出现的次数.

(2) *Micro-F*1 考虑所有实例的总体精确度和召回率, 为频繁标签赋予更多权重.

$$Micro\text{-}F1 = 2\times\frac{Precision_{\mathrm{mi}}\times Recall_{\mathrm{mi}}}{Precision_{\mathrm{mi}}+Recall_{\mathrm{mi}}} \tag{20}$$

$$Precision_{\mathrm{mi}} = \frac{\sum_{j=1}^{C}TP_j}{\sum_{j=1}^{C}(TP_j+FP_j)} \tag{21}$$

$$Recall_{\mathrm{mi}} = \frac{\sum_{j=1}^{C}TP_j}{\sum_{j=1}^{C}(TP_j+FN_j)} \tag{22}$$

(3) *Macro-F*1 计算所有标签的平均 *F*1 分数, 对每个标签赋予相同的权重.

$$Macro\text{-}F1 = \frac{1}{C}\sum_{j=1}^{C}F1_j \tag{23}$$

3.3 实验设置

本文使用 Python 3.7 作为开发语言, PyTorch 1.11.0 作为训练框架, 在一台显卡配置为 NVIDIA GeForce RTX 3090 的 Ubuntu 20.04 服务器上进行实验. BERT 预训练模型采用“BERT-base-uncased”, 词嵌入维度 d 为 768 维, 文本长度 n 设置为 512. 对于所有数据集, 选择 Adam 优化器来最小化损失函数, 学习率设置为 3E–5. 批量大小 batch_size 设置为 12. 为防止过拟合, 使用早停策略. 实验将分类阈值设置为 0.5, 概率高于阈值的标签即被判定为预测结果. 对于 WOS 数据集, 设置 $\lambda_G = 0.0005$, $\lambda_L = 0.05$; 对于 BGC 数据集, 设置 $\lambda_G = 0.0001$, $\lambda_L = 0.3$. 对比学习的温度系数 τ 设置为 1. 分布平衡损失中, 重平衡方法的超参数 $\alpha = 0.1$, $\beta = 10.0$, $\mu = 0.05$, 负容忍正则化的超参数 $\lambda = 2.0$.

3.4 实验结果

为了验证 LC-LTD 在层级文本分类任务上预测性能的提升, 将其与目前主流的层级文本分类模型在 WOS、BGC 数据集上进行了详细对比实验, 具体实验结果如表 2 所示. 参与对比的模型主要分为两类: 一类是基于静态词向量的层级感知模型, 另一类是基于 BERT 的模型. 表中标有*的为原论文引用结果, 其他为复现结果.

从表 2 中可以看出, 本文提出的模型 LC-LTD 在 *Macro-F*1、*Micro-F*1 两个指标方面均优于其他模型, 取得了良好的效果, 突出了在层级分类问题中考虑标签共现和长尾分布的重要性.

从上下两部分看, 基于 BERT 的模型总体上优于基

于静态词嵌入的模型, 这是由于 BERT 强大的文本表示能力能够提升模型对数据特征的提取能力和学习能力.

表 2 LC-LTD 与其他模型的比较 (%)

类型	模型	WOS		BGC	
		*Macro-F*1	*Micro-F*1	*Macro-F*1	*Micro-F*1
Hierarchy-aware models	HiAGM-LA	78.77	84.63	55.71	75.70
	HiAGM-TP	79.68	85.67	56.06	76.39
	HiMatch	80.06	85.98	58.95	76.79
	HTCInfoMax	78.31	84.44	56.44	76.10
	DLAC (2023)[6]	81.02*	86.80*	58.80*	76.51*
Pretrained language models	BERT	80.16	86.05	61.41	79.06
	HGCLR	81.12	86.85	63.17	80.19
	HiTIN (2023)[27]	81.57*	87.19*	—	—
	LC-LTD	**81.86**	**87.25**	**63.81**	**80.47**

相比只利用简单的 BERT 模型, LC-LTD 在 WOS 数据集上的 *Macro-F*1、*Micro-F*1 值分别提升了 0.7%、1.2%; 在 BGC 数据集提升了 2.5%、0.61%. 这是因为 LC-LTD 利用真实标签与文本、单词之间的相关性, 充分挖掘文本语义信息和标签结构特征, 引导 BERT 词向量的优化, 使其更适用于特定领域的分类任务.

相比基线模型 HGCLR, LC-LTD 在 WOS 数据集上的 *Macro-F*1、*Micro-F*1 值分别提升了 0.74%、0.4%, 在 BGC 数据集上提升了 0.64%、0.28%. 这是因为模型在学习文本特征的时候进一步引入了全局基于共享标签的对比学习, 可以更好地学习文本的全局语义, 又采用了缓解长尾分布问题的策略, 用分布平衡损失替换二进制交叉熵损失, 进一步提升了模型最终的分类效果.

另外, LC-LTD 在两个数据集的 *Macro-F*1 相较其他模型有了显著提高, *Micro-F*1 的改进相对较小. 这一现象可以从 *Macro-F*1 和 *Micro-F*1 的计算性质来解释. *Macro-F*1 值平等地对待每个类别, 被认为更适合于处理标签不平衡问题; 而 *Micro-F*1 偏向于频繁出现的标签, 可能不能准确反映模型在低频率标签上的表现. 因此, 对于类别不平衡的层级分类, *Macro-F*1 更能反映模型的性能. LC-LTD 对 *Macro-F*1 的影响比对 *Micro-F*1 的影响大, 说明它在处理标签不平衡问题上的有效性.

综上, LC-LTD 可以成为层级分类问题的一种有效解决方法.

3.5 性能分析

3.5.1 消融实验

除了与各基准方法的整体性能进行比较, 为了更好地评估 LC-LTD 中不同组件的有效性, 本文在 WOS、BGC 数据集上进行了消融实验, 结果展示在表 3 中, r.m. CON 代表移除基于共享标签的对比学习, r.p. LOSS 代表将 DBLoss 替换为原始的 BCELoss.

从表 3 中可以观察到, 在 WOS、BGC 这两个数据集上, 完整的模型 LC-LTD 总体性能为最佳, 去掉或者替换任意一个模块都会导致一定程度的性能下降. 移除基于共享标签的对比学习, 会使得模型缺乏对文本全局特征的约束, 降低文本聚类的可能性. 将 DBLoss 替换为 BCELoss, 模型容易受到标签分布的影响, 导致子类标签嵌入质量差, 对文本的鉴别能力弱, 进而影响模型整体的分类性能.

表 3 消融实验 (%)

模型	WOS		BGC	
	*Macro-F*1	*Micro-F*1	*Macro-F*1	*Micro-F*1
r.m. CON	81.60	87.11	63.57	80.40
r.p. LOSS	81.74	87.15	63.35	80.36
LC-LTD	**81.86**	**87.25**	**63.81**	**80.47**

3.5.2 分层表现分析

层级文本分类任务中, 标签自上而下从宽泛到具体的概念进行组织, 随着层级越深, 标签的实例数量越少, 分类难度越大. 因此, 模型在关注整体分类精度的同时, 每一层级的分类性能也相当重要. 本文在长尾分布数据集 WOS、BGC 的每个层级上都将 LC-LTD 与 HGCLR、BERT 做了比较, 分别计算各层的 *Macro-F*1、*Micro-F*1 分数, 具体情况展示在图 5 中.

WOS 数据集的标签分为 2 层, 如图 5(a)、(b) 分别展示了 LC-LTD 在 WOS 数据集上的逐级 *Macro-F*1、*Micro-F*1 性能. 图中, LC-LTD 在第 1 层级的性能高于 BERT, 略低于 HGCLR. 在第 2 层级, LC-LTD 的 *Macro-F*1、*Micro-F*1 性能分别比 HGCLR 高 0.79%、0.83%, 比 BERT 高 1.75%、1.69%.

BGC 数据集的标签分为 4 层, 如图 5(c)、(d) 分别展示了 LC-LTD 在 BGC 数据集上各个层级的 *Macro-F*1、*Micro-F*1 分数. 可以看出, 在 1、2 层级, LC-LTD 与 HGCLR 的性能相差无几, 均优于 BERT. 在第 3、4 层级, LC-LTD 表现优异, 其 *Macro-F*1 性能分别比 HGCLR 高 0.73%、0.66%, *Micro-F*1 性能分别比 HGCLR 高 0.44%、2.46%.

总而言之, LC-LTD 在两个数据集的多个层级上都表现出了卓越的性能, 尤其是对下层尾类标签的预测性能有了明显提升, 表明 LC-LTD 有效捕获了复杂的标签层次, 更适合处理层级文本分类问题, 其引入的分布平衡损失在缓解长尾分布问题上是有效的.

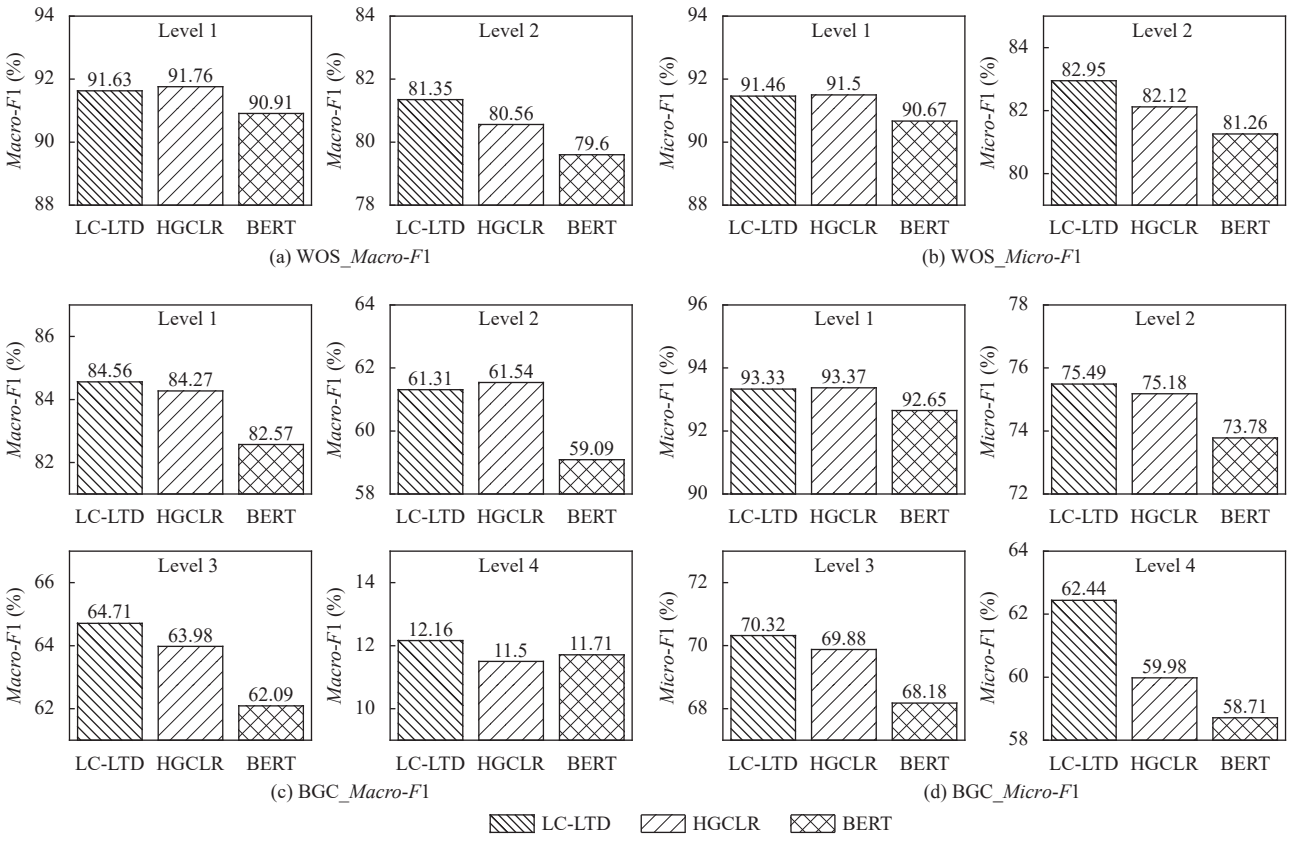

图 5 WOS、BGC 数据集的层级性能

4 结论与展望

本文介绍了面向标签共现和长尾分布的层级文本分类模型 LC-LTD, 针对以 HGCLR 为代表的现有分类方法存在的两方面问题进行优化. 考虑到层级文本分类的标签共现特征, 设计基于共享标签的对比学习目标建模实例对之间细粒度的语义相关性, 引导模型生成更好的判别性语义表征; 考虑到层级文本分类的长尾分布特征, 引入分布平衡损失替换二进制交叉熵损失, 缓解类别分布不平衡对模型造成的影响. 通过一系列实验, 展示了 LC-LTD 在 WOS、BGC 两个公开数据集上的性能优于现有方法, 并分析了模型在不同层级上的表现, 对 LC-LTD 的关键部件也进行了验证, 突出了本文所提方法的有效性. 未来, 考虑到直接利用标签名称这样的粗粒度描述可能难以与文本中细粒度的词汇建立联系, 我们希望进一步研究标签语义的更丰富表达, 以此强化文本与标签的交互作用, 提高分类性能.

参考文献

1 Zhang Y, Shen ZH, Dong YX, *et al*. MATCH: Metadata-aware text classification in a large hierarchy. Proceedings of the Web Conference 2021. Ljubljana: ACM, 2021. 3246–3257.

2 黄威. 层次化多标签分类方法及其应用研究 [博士学位论文]. 合肥: 中国科学技术大学, 2023.

3 郭豪. 基于标签嵌入的层级多标签分类方法设计 [硕士学位论文]. 重庆: 重庆邮电大学, 2022.

4 Kumar A, Toshinwal D. HLC: Hierarchically-aware label correlation for hierarchical text classification. Applied Intelligence, 2024, 54(2): 1602–1618. [doi: 10.1007/s10489-023-05257-1]

5 Zangari A, Marcuzzo M, Schiavinato M, *et al*. Hierarchical text classification: A review of current research. Expert Systems with Applications, 2023, 224.

6 Cao YK, Wei ZY, Tang YJ, *et al*. Hierarchical label text classification method with deep-level label-assisted classification. Proceedings of the 12th IEEE Data Driven Control and Learning Systems Conference. Xiangtan: IEEE, 2023. 1467–1474.

7 Wang ZH, Wang PY, Huang LZ, *et al*. Incorporating hierarchy into text encoder: A contrastive learning approach for hierarchical text classification. arXiv:2203.03825v2, 2022.

8 Su XA, Wang R, Dai XY. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 672–679.

9 王嫄, 徐涛, 王世龙, 等. 层级标签语义引导的极限多标签文本分类策略. 中文信息学报, 2021, 35(10): 110–118. [doi: 10.3969/j.issn.1003-0077.2021.10.013]

10 赵海燕, 曹杰, 陈庆奎, 等. 层次多标签文本分类方法. 小型微型计算机系统, 2022, 43(4): 673–683.

11 Li SB, Hu J, Cui YX, *et al*. DeepPatent: Patent classification with convolutional neural networks and word embedding. Scientometrics, 2018, 117(2): 721–744. [doi: 10.1007/s11192-018-2905-5]

12 Li YD, Zhang YQ, Zhao Z, *et al*. CSL: A large-scale Chinese scientific literature dataset. Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: ACL, 2022. 3917–3923.

13 Banerjee S, Akkaya C, Perez-Sorrosal F, *et al*. Hierarchical transfer learning for multi-label text classification. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6295–6300.

14 Shimura K, Li JY, Fukumoto F. HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 811–816.

15 滕思洁. 基于图神经网络的层级文本分类 [硕士学位论文]. 合肥: 中国科学技术大学, 2022.

16 Zhou J, Ma CP, Long DK, *et al*. Hierarchy-aware global model for hierarchical text classification. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 1106–1117.

17 Chen HB, Ma QL, Lin ZX, *et al*. Hierarchy-aware label semantics matching network for hierarchical text classification. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). ACL, 2021. 4370–4379.

18 Deng ZF, Peng H, He DX, *et al*. HTCInfoMax: A global model for hierarchical text classification via information maximization. arXiv:2104.05220v1, 2021.

19 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. arXiv:1810.04805v2, 2019.

20 Ying CX, Cai TL, Luo SJ, *et al*. Do transformers really perform bad for graph representation? Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 2212.

21 Jang E, Gu SX, Poole B. Categorical reparameterization with gumbel-softmax. arXiv:1611.01144v5, 2017.

22 Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 1597–1607.

23 Wu T, Huang QQ, Liu ZW, *et al*. Distribution-balanced loss for multi-label classification in long-tailed datasets. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 162–178.

24 Kowsari K, Brown DE, Heidarysafa M, *et al*. HDLTex: Hierarchical deep learning for text classification. Proceedings of the 16th IEEE International Conference on Machine Learning and Applications. Cancun: IEEE, 2017. 364–371.

25 Aly R, Remus S, Biemann C. Hierarchical multi-label classification of text with capsule networks. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence: ACL, 2019. 323–330.

26 Peng H, Li JX, He Y, *et al*. Large-scale hierarchical text classification with recursively regularized deep graph-CNN. Proceedings of the 2018 World Wide Web Conference. Lyon: International World Wide Web Conferences Steering Committee, 2018. 1063–1072.

27 Zhu H, Zhang C, Huang JJ, *et al*. HiTIN: Hierarchy-aware tree isomorphism network for hierarchical text classification. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: ACL, 2023. 7809–7821.

(校对责编: 王欣欣)