计算机科学
COMPUTER SCIENCE

基于异质图神经网络预训练的多标签文档分类研究

吴家伟, 方全, 胡骏, 钱胜胜

引用本文

吴家伟, 方全, 胡骏, 钱胜胜. 基于异质图神经网络预训练的多标签文档分类研究[J]. 计算机科学, 2024, 51(1): 143-149.

WU Jiawei, FANG Quan, HU Jun, QIAN Shengsheng. Pre-training of Heterogeneous Graph Neural Networks for Multi-label Document Classification [J]. Computer Science, 2024, 51(1): 143-149.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600079

基于异质图神经网络预训练的多标签文档分类研究

吴家伟[1]　方　全[2]　胡　骏[2]　钱胜胜[2]
1 郑州大学河南先进技术研究院　郑州 450002
2 中科院自动化所模式识别国家重点实验室　北京 100190
(robin.wujw@gmail.com)

摘　要　多标签文档分类是一种将文档实例与相关标签相关联的技术,近年来受到越来越多研究者的关注。现有的多标签文档分类方法尝试探索文本之外的信息的融合,如文档元数据或标签结构。然而,这些方法要么简单地利用元数据的语义信息,要么没有考虑标签的长尾分布,因此忽略了文档及其元数据之间的高阶关系和标签的分布规律等信息,从而影响到多标签文档分类的准确性。因此,文中提出一种新的基于异质图神经网络预训练的多标签文档分类方法。该方法通过构造文档与其元数据的异质图,采用两种对比学习预训练方法捕获文档与其元数据之间的关系,并通过平衡标签长尾分布的损失函数来提高多标签文档分类的准确性。在基准数据集上的实验结果表明,所提方法的准确率比 Transformer 提高了 8%,比 BertXML 提高了 4.75%,比 MATCH 提高了 1.3%。

关键词:多标签文档分类;元数据;异质图神经网络;预训练;长尾分布
中图分类号　TP391

Pre-training of Heterogeneous Graph Neural Networks for Multi-label Document Classification

WU Jiawei[1],FANG Quan[2],HU Jun[2] and QIAN Shengsheng[2]
1 Henan Institute of Advanced Technology,Zhengzhou University,Zhengzhou 450002,China
2 National Laboratory of Pattern Recognition,Institute of Automation,Chinese Academy of Sciences,Beijing 100190,China

Abstract　Multi-label document classification aims to associate document instances with relevant labels,which has received increasing research attention in recent years. Existing multi-label document classification methods attempt to explore the fusion of information beyond the text,such as document metadata or label structure. However,these methods either simply use the semantic information of metadata or do not consider the long-tail distribution of labels,thereby ignoring higher-order relationships between documents and their metadata and the distribution pattern of labels,which affects the accuracy of multi-label document classification. Therefore,this paper proposes a new multi-label document classification method based on the pre-training of heterogeneous graph neural networks. The method constructs a heterogeneous graph based on documents and their metadata,adopts two contrastive pre-training methods to capture the relationship between documents and their metadata,and improves the accuracy of multi-label document classification by balancing the problem of long-tail distribution of labels through a loss function. Experimental results on the benchmark dataset show that the proposed method outperforms Transformer BertXML and MATCH by 8%,4.75%,1.3%,respectively.

Keywords　Multi-label document classification,Metadata,Heterogeneous graph neural network,Pre-training,Long-tail distribution

1　引言

随着数字化信息的发展以及电子文档数据的急剧增加,传统的人工分类整理变得越来越困难,信息过载问题越来越严重。以科学出版物发行数量为例,每 12 年科学出版物的发行数量就翻一番[1],到 2019 年已经达到 2.4 亿。截至 2021 年 2 月,已经有 213236 篇关于新冠肺炎的论文生成[2]。而文档分类是应对信息过载问题的一种有效手段,它可以根据文档内容对文本单位(如句子、段落等)分配相应标签[3],能够为文档检索和文档推荐提供关键有效信息,从而极大地提升人们的检索效率和阅读体验。传统的文档分类主要依赖于人工整理,但在如今信息爆炸的互联网时代,手动整理文档信息将会消耗巨大的人力、物力和时间成本。因此,为文档自动分配相应的类别具有十分重要的现实意义[4]。

到稿日期:2023-06-08　返修日期:2023-10-09
基金项目:国家自然科学基金(62072456,62036012,62106262)
This work was supported by the National Natural Science Foundation of China(62072456,62036012,62106262).
通信作者:方全(qfang@nlpr.ia.ac.cn)

在传统的文档分类问题中，每个文档只对应一个类别标签，且各个类别标签之间相互独立，这种分类问题为单标签文档分类[5]。而现实生活中的大部分文档与其标签的对应关系更加复杂多样，一个文档数据往往与多个类别的标签相关，且类别标签之间存在一定的依赖与分层关系，这就引出了多标签文档分类问题[6]。相对于单标签文档分类，多标签文档分类能更全面、完整地表示文档内容，也更符合现实中文档数据的实际需求，因此成为文档分类的主流研究方向。

为了解决这一问题，早期的工作主要关注于通过各类深度学习模型学习输入文档的语义表示，再通过分类器进行分类，例如使用卷积神经网络求解多标签文档分类问题的XML-CNN[7]方法，以及使用双向 RNN 网络和标签感知的 Attention 层进行多标签文档分类的 AttentionXML 方法[8]。此外，Zhang 等[9]利用标签结构信息构建了标签结构空间以探索标签之间的相关性；Zhang 等[4]提出了结合文档元数据信息，在同一嵌入空间中学习文档及其元数据表示的方法。然而现有的方法忽略了以下两个方面的问题：

(1)文档与其元数据之间的高阶信息。如图 1 所示，对于 Bert 论文[10]，可以构造由多个作者、多个参考文献、一个发表地点和多个标签组成的异质元数据图。通过对元数据进行建模，捕获异质元数据图各节点之间的关系，获得各节点的表示向量，有助于提高最终多标签文档分类的准确性。

(2)文档标签的不平衡样本问题。现实中文本数据一般呈现长尾分布，大部分标签(头部标签)实例数据少，只有小部分标签(尾部标签)实例数据多，不利于模型对尾部标签的分类。

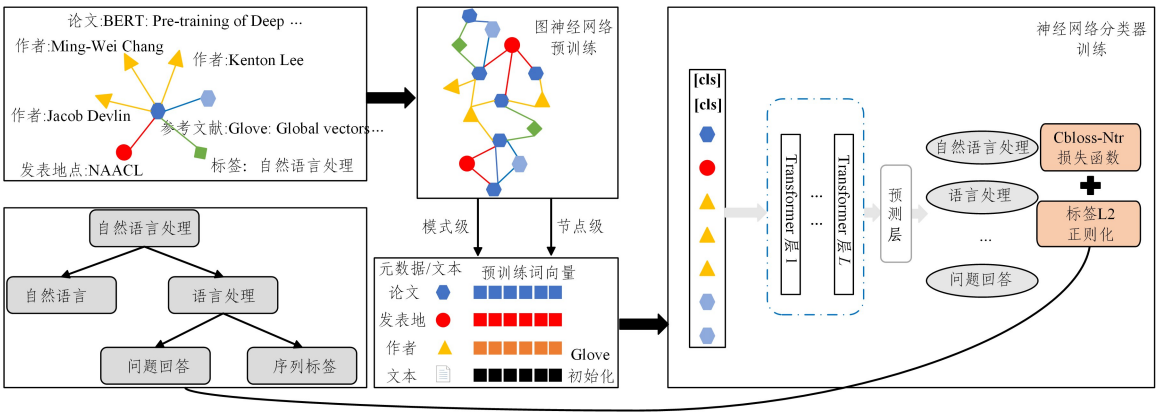

图 1 基于图神经网络预训练的多标签文档分类方法 PGMLDC 框架

Fig. 1 PGMLDC framework for pre-training of heterogeneous graph neural networks for multi-label document classification

针对以上问题，本文将文档的元数据信息，如文档的作者、发表地点、参考文献、标签等构造为异质图。经过两种对比学习预训练任务提取出文档及其元数据的词向量，输入到 Transformer 模型中学习捕获文档与元数据之间的关系，再结合平衡标签长尾分布问题的损失函数进行文档的多标签分类。

本文的主要贡献如下：

(1)提出了一种新的基于图神经网络预训练的多标签文档分类方法 PGMLDC(Pre-Training of Heterogeneous Graph Neural Networks for Multi-Label Document Classification)，构造论文及其元数据信息的异质图来捕获论文与其元数据之间的语义联系。

(2)使用平衡标签长尾分布问题的长尾分布损失函数(CBLoss-NTR[11])改善输入标签的长尾分布问题，提高多标签文档分类的准确性。

(3)在两个基准数据集上的实验结果表明，所提方法优于现有的最先进的多标签文档分类方法。

2 相关工作

目前解决多标签文档分类问题的方法主要是使用深度神经网络充分捕获文本信息来进行分类，如基于卷积神经网络的 XML-CNN 模型[7]和 AttentionXML 模型[8]使用注意力机制捕获输入文本中与每个标签最相关的部分。X-Transformer[12]提出了一种基于预训练 Transformer 的方法来捕获文本信息进行多标签文档分类。Gong 等[13]提出了 HG-Transfomer 的深度学习模型，将文本建模为一个图形结构，然后在单词、句子和图形级别使用具有多头注意机制的多层 Transfomer 结构充分捕获文本的特征。Ma 等[14]通过混合嵌入的方式，融合标签层次结构的图嵌入和标签类别的词嵌入，通过双向门控循环单元(BGRU)和混合嵌入逐级学习文本表示。

在利用文档元数据信号方面，Tang 等[15]提出了一种用于情感分析的神经网络方法，该方法把用户和产品元信息用向量空间模型表示；Kim 等[16]采用元数据信号作为额外的特征来训练一个深度神经网络分类器；Zhang 等[4]提出了一种名为 MATCH 的方法，该方法将文本和元数据在同一空间进行预训练，获取对应的表示向量，利用 Transformer 的全连接注意力机制来获得单词和不同类型的元数据之间的关系；Zhang 等[17]利用文档元数据信息，提出了一种新的对比学习方法，显著提高了零样本多标签分类问题在重排序阶段的性能。在利用标签层次结构方面，Zhang 等[9]构建了一个标签依赖图，以基于图的先验知识对标签空间的标签向量进行建模；Yang 等[18]将多标签文档分类任务转换为 Seq2Seq 任务，并预测了具有标签依赖性的标签序列；Wang 等[19]提出一种用于层次多标签分类问题的增量式超网络学习方法，通过将超网络的超边组织成相应的层次结构，使输出的预测标签能够满足标签的层次约束。然而这些方法忽略了文档及其元数据的异质结构，也没有考虑文档标签的不平衡分布等现实中普遍存在的问题。

3 问题描述

给定数据集 $D=\{(x_i,y_i)\}_{i=1}^{N}$，其中 $x_i\in X$ 代表文档数据样本，$\boldsymbol{y}_i=[1,2,3,4,\cdots,k]$ 代表所有的标签集合，标签总数为 k，对于每一个文档数据样本 x_i 都有多个标签与其相对应。多标签文档分类（Multi-Label Document Classification，MLDC）的主要任务就是学习由文档数据样本到其对应标签的映射函数 f，使得对每一个给定的文档样本 x_i 都能预测出其对应的标签集合 $\boldsymbol{y}_i$。

传统方法将文档的标题和正文当作文档的文本信息并作为输入进行预测。然而，文档的元数据 $\boldsymbol{M}_d$ 和标签的层次结构对于实际的文档多标签分类也有非常重要的作用。以图 1 中的学术出版物为例，一份文档的元数据包括其作者（例如“Jacob Devlin”）、发表地点（例如“NAACL”）、标签和参考论文。标签层次结构是根据研究主题的细粒度级别组织的，例如“自然语言处理”“语言处理”和“问答系统”。

本文不仅考虑了文档的正文信息，同时也考虑了元数据信息以及标签的分层信息。因此，在本文中，多标签文档分类任务主要学习从 $(\boldsymbol{M}_d,\boldsymbol{W}_d)\rightarrow\hat{\boldsymbol{y}}_d$ 的映射函数，$\hat{\boldsymbol{y}}_d$ 代表预测的与文档 d 最相关的标签集。

4 整体框架

本章主要介绍基于图神经网络预训练的多标签文档分类方法 PGMLDC 的整体框架，如图 1 所示。我们使用论文 BERT[10] 作为运行示例。为了更好地捕获到文档的元数据信息，我们构造了文档的元数据异构图，并使用异质图神经网络预训练模型进行训练，得到文档及其元数据的表示信息，这些表示信息被进一步输入到 Transformer[20] 模型中进行编码，并结合平衡标签长尾分布的损失函数进行文档的多标签分类。此外，本文还利用标签的层次结构信息，通过对父子标签的参数正则化，进一步提高了多标签分类的效果。

基于图神经网络预训练的多标签文档分类方法可分解为 4 个模块，分别是图神经网络预训练模块、Transformer 编码模块、平衡标签长尾分布的预测模块和标签分层正则化模块。

4.1 图神经网络预训练

本小节将介绍如何根据文档及其元数据之间的关系构建元数据异质图。异质图的模式如图 2 所示，其中包含 5 种类型的节点，如文档、作者、地点、标签和参考文献。

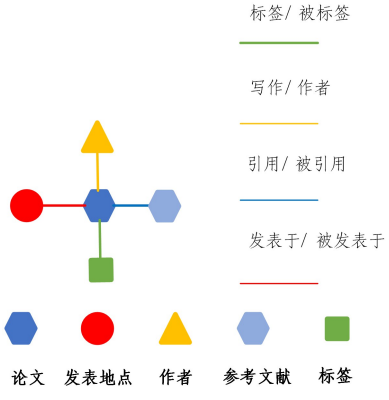

图 2 异质图的节点及其模式

Fig. 2 Nodes and schema of heterogeneous graphs

在预训练阶段，本文受最近图神经网络预训练工作的启发，主要采用节点级别和模式级别[21]两种预训练任务来帮助图神经网络更好地捕获元数据异构图的异质语义和结构信息。给定预训练图 $G=\{V,\varepsilon,A,R,\phi,\varphi\}$，其中 V 和 ε 分别代表异质图的节点集和边集，A 和 R 分别代表节点和边的类别，$|A|+|R|>2$，ϕ 和 φ 分别代表节点集和节点类别、边集和边类别的映射函数。此外，异质图的网络模式 $T_G=(A,R)$ 明确了节点和其关系的类型约束。

预训练模型的设计目标是将节点和边在模型中进行编码，以便充分表达它们之间的内在异质性，使得模型能够更好地学习异质图中的各种特征和关系。因此我们采用节点级别和模式级别两种对比学习方法来学习可转移的知识。

4.1.1 节点级别的预训练任务

本小节主要介绍异质图节点级别的对比学习预训练任务。在正样本的选择方面，我们将连接同一条边的两个节点（例如文档与其连接的标签）构建为正样本对，认为它们在隐向量空间中具有相似性，基于异质图的性质，通过 3 种方法构建负样本：

(1)同类型替换，使用同类型的节点 v^- 代替原节点 v 作为负样本。

(2)引入高阶（二阶、三阶）节点信息，使用二阶和三阶节点 v^{high} 作为负样本。

(3)非相似替换，由于相近节点替换之后可能会引入噪声，因此被替换的节点 v^- 不能与原节点过于接近。

负样本生成方法(1)和(3)可以用式(1)表达：

$$N^{\text{node}}_{\langle u,R,v\rangle}=\{\langle u,R,v^-\rangle\mid\phi(v)=\phi(v^-),(u,v^-)\notin\varepsilon,Sim(v,v^-)\leqslant\delta\}\tag{1}$$

其中，$Sim()$ 用于计算两个节点相似度；δ 是相似度阈值。为了避免构造太简单的负样本，δ 一般取较大的值。

高阶负样本选取方法表示如下：

$$N^{\text{node}}_{\langle u,R,v^{\text{high}}\rangle}=\{\langle u,R,v^{\text{high}}\rangle\mid v^{\text{high}}=v^2\bigcup v^3\}\tag{2}$$

其中：

$$v^2=(u,v)\in\varepsilon\bigcap(v,v^2)\in\varepsilon\tag{3}$$

$$v^3=(u,v)\in\varepsilon\bigcap(v,v^2)\in\varepsilon\bigcap(v^2,v^3)\in\varepsilon\tag{4}$$

节点级别的预训练任务采用式(5)所示的 InfoNCE 损失函数[22]进行训练。

$$L^{\text{node}}_{u,R}=-\log\frac{\exp\left(\frac{\boldsymbol{h}_u^{\mathrm{T}}\boldsymbol{W}_R\boldsymbol{h}_v}{\tau}\right)}{\sum\limits_{i\in\{v\}\cup\{w\mid\langle u,R,w\rangle\in N^{\text{node}}_{\langle u,R,v\rangle}\cup N^{\text{node}}_{\langle u,R,v^{\text{hard}}\rangle}\}}\exp\left(\frac{\boldsymbol{h}_u^{\mathrm{T}}\boldsymbol{W}_R\boldsymbol{h}_i}{\tau}\right)}\tag{5}$$

其中，$\boldsymbol{h}_u^T$ 和 $\boldsymbol{h}_v$ 分别表示节点的向量表示，$\boldsymbol{W}_R\in R^{d\times d}$ 是关于边的可学习的关系矩阵，τ 是温度超参数。

4.1.2 模式级别的预训练任务

节点级的预训练任务只能捕获节点之间的一阶语义信息，异质图的网络模式（Network Schema）是异质图独特的定义结构，本文中的模式信息为论文的元数据信息。为了获取节点的高阶语义和结构信息，更好地捕获文档与其对应标签之间的关联，本文利用异质图的模式信息快速从原始输入中采样论文元数据样本，因此我们又增加了模式级别的对比学习预训练任务。

在正样本的生成上，为了防止出现节点不平衡的问题，我们限制了每个类别下采样的模式样本数量，正样本采样的公式如下：

$$P_u^{\text{sche}} = \bigcup_{s \in I(u)} s \backslash \{u\} \tag{6}$$

其中，$I(u)$代表包含节点 u 的所有模式实例。

负样本的生成主要采用 3 种方式：

(1)选取一个 batch 中同样类型的不同节点，公式如下：

$$N_u^1 = \{P_u^{\text{sche}} \mid u^- \in V_B, u \neq u^-, \phi(u) = \varphi(u^-)\} \tag{7}$$

(2)采用队列利用上一个 batch 的正样本选取同类型的不同节点，其中V_B^{t-1}表示上一个 batch 中的所有节点。

$$N_u^2 = \{P_v^{\text{sche}} \mid \phi(u) = \phi(v), v \in V_B^{t-1}\} \tag{8}$$

(3)针对多标签分类任务，随机选取 n 个标签集中其他标签作为负样本。

$$N_u^3 = \{P_l^{\text{sche}} \mid \phi(u) = \phi(l), l \in L\} \tag{9}$$

最终模式级别的负样本集合是上述两种方法的并集。

$$N_u^{\text{sche}} = N_u^1 \cup N_u^2 \cup N_u^3 \tag{10}$$

我们为每种类型的节点设计一个编码器来学习节点表示。具体来说，对于节点类型 $\phi(v_i)$，通过编码器$Enc^{\phi(v_i)}$学习目标节点v_i的节点嵌入。

$$\boldsymbol{e}_{v_i} = Enc^{\phi(v_i)}(\boldsymbol{h}_{v_i}) \tag{11}$$

$$\boldsymbol{c}_u^s = Pool(\boldsymbol{e}_{v_1}, \boldsymbol{e}_{v_2}, \cdots) \tag{12}$$

然后，对于目标节点 u，通过一个池化层生成其节点嵌入 $\boldsymbol{c}_u^s$。

最终模式级别的对比学习目标如下：

$$L_u^{\text{sche}} = \sum_{s^+ \in P_u^{\text{sche}}} \log \frac{\exp\left(\frac{\boldsymbol{h}_u^{\mathrm{T}} \boldsymbol{c}^{s^+}}{\tau}\right)}{\sum_{s \in \{s^+\} \cup N_u^{\text{sche}}} \exp\left(\frac{\boldsymbol{h}_u^{\mathrm{T}} \boldsymbol{c}^{s}}{\tau}\right)} \tag{13}$$

最终的学习目标是最小化损失函数 L：

$$L = L^{\text{node}} + \lambda L^{\text{sche}} \tag{14}$$

其中，λ 是平衡系数。

4.2 Transformer 编码

本节将详细介绍如何通过 Transformer 编码器促进文档和元数据之间的信息交换。将文档中所有的元数据实例与单词实例连接起来形成编码层的输入，同时在每个输入序列开头添加[CLS]标记用作分类任务的聚合序列表示。编码层的输入为：

$$\boldsymbol{H} = [\underbrace{\boldsymbol{e}_{[CLS_1]}; \cdots; \boldsymbol{e}_{[CLS_S]}}_{[CLS]\text{标记}}; \underbrace{\boldsymbol{e}_{m_1}; \cdots; \boldsymbol{e}_{m_n}}_{\text{元数据}M_d}; \underbrace{\boldsymbol{e}_{w_1}; \cdots; \boldsymbol{e}_{w_t}}_{\text{单词}W_d}]$$

因此 $\boldsymbol{H} \in R^{\delta \times (S + |M_d| + |W_d|)}$，其中 δ 是词嵌入空间的维度。

例 1(输入序列) 假设我们得到了图 1 中的文档“BERT”。输入到 Transformer 层中的序列将是：

“[CLS1]…[CLSC][VENUE_NAACL][AUTHOR_Jacob Devlin][AUTHOR_Ming-Wei Chang]…[RE-FERENCE_Glove:Global vectors…]…[WORD_W-e][WORD_introduce][WORD_a][WORD_new]…[WORD_ modifications]”

Transformer 层主要使用多头注意力机制来捕获文档和元数据信息之间的关系，表达式如下：

$$\boldsymbol{a}_i = Attention(\boldsymbol{q}\boldsymbol{W}_i^Q, \boldsymbol{H}\boldsymbol{W}_i^K, \boldsymbol{H}\boldsymbol{W}_i^V) \tag{15}$$

$$MultiHeadAtt(\boldsymbol{q}, \boldsymbol{H}) = [\boldsymbol{a}_1 \parallel \boldsymbol{a}_2 \parallel \cdots \parallel \boldsymbol{a}_k]\boldsymbol{W}^O \tag{16}$$

其中，$\parallel$ 代表向量的拼接操作；$\boldsymbol{q} \in R^{1 \times \delta}$为查询矩阵；$\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^V$，$\boldsymbol{W}^O$均为可学习参数。

对于每一个输入的实例 $i \in H$，以及实例的预训练词向量 $\boldsymbol{e}_i$，最终每个 Transformer 层的输出如下：

$$\boldsymbol{h}_i = LayerNorm(\boldsymbol{z}_i + FFN(\boldsymbol{z}_i)) \tag{17}$$

其中$\boldsymbol{z}_i$的求解公式为：

$$\boldsymbol{z}_i = LayerNorm(\boldsymbol{e}_i + MultiHeadAtt(\boldsymbol{e}_i, \boldsymbol{H})) \tag{18}$$

其中，$LayerNorm(\cdot)$是 layer 正则化操作[23]，FFN($\cdot$)是前馈神经网络[20]，经过 L 层 Transformer 层后最终的输出$\boldsymbol{H}^{(L)}$被用作模型的预测。

4.3 平衡标签长尾分布的预测

在经过 L 层 Transformer 层之后，将所有[CLS]标签的最终状态连接起来，以获取最终的文档表示$\hat{\boldsymbol{h}}_d$。

$$\hat{\boldsymbol{h}}_d = \boldsymbol{h}_{[CLS_1]}^{(L)} \parallel \boldsymbol{h}_{[CLS_2]}^{(L)} \parallel \cdots \parallel \boldsymbol{h}_{[CLS_S]}^{(L)} \tag{19}$$

预测层的目标是给文档分配最相关的标签$\hat{\boldsymbol{y}}_d$，通过在 Transformer 层后连接全连接层，再使用 sigmoid 激活函数获取最终的多标签预测结果。

$$\hat{\boldsymbol{y}}_d = \text{Sigmoid}(\hat{\boldsymbol{h}}_d \boldsymbol{W}^L + \boldsymbol{b}) \tag{20}$$

其中，$\boldsymbol{W}^L = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{|L|}]$，任意$\boldsymbol{w}_i \in \boldsymbol{W}^L$被看作第 i 个标签的参数。

当标签存在长尾分布，即标签的一小部分(头部标签)具有多个实例，而大多数标签(尾部标签)只有少数实例时，多标签分类容易偏向于分类到头部标签。二元交叉熵损失(Binary Cross Entropy，BCE)函数常用于多标签文档分类任务中，而普通的二元交叉熵损失函数容易受到长尾分布的影响。受到文献[11]的启发，我们使用结合 Class-Balanced focal loss (CB)[24]和负容忍正则化(Negative-Tolerant Regularization，NTR)[25]的 CB-NTR 损失函数重新加权 BCE，使得模型能够更关注少数实例的尾部标签，公式如下：

$$L_{\text{CB-NTR}} = \begin{cases} -r_{CB}(1-q_d^i)^{\gamma}\log(q_d^i), & \text{if } y_d^i = 1 \\ -r_{CB}\dfrac{1}{\lambda}(q_d^i)^{\gamma}\log(1-q_d^i), & \text{otherwise} \end{cases} \tag{21}$$

其中，$y_d^i = 1$ 代表文档对应的第 i 个标签，$\hat{y}_d^i$代表模型输出预测文档具有第 i 个标签，对于尾部样本来说$q_d^i = \sigma(\hat{y}_d^i - v_i)$，对于头部样本来说$q_d^i = \sigma(\lambda(\hat{y}_d^i - v_i))$，$v_i$可以在训练开始时最小化损失函数，比例系数为 k，类别先验信息$p_i = n_i / N$，且有：

$$v_i = -k \times \hat{b}_i \tag{22}$$

$$\hat{b}_i = -\log\left(\frac{1}{p_i} - 1\right) \tag{23}$$

4.4 标签分层正则化

在多标签分类中，标签之间通常存在很强的关联性。例如，“神经网络”标签和“深度学习”标签往往一起出现，这种关联关系对于标签分类来说非常重要。通过识别和利用这些关联关系，可以更好地理解标签之间的联系，提高多标签分类的准确性和效率。

为了在模型中添加标签层次信息，我们采取 L2 正则化

的方式使得每个标签在参数上与其父标签相似。

$$\boldsymbol{f}_{\text{parameter}}=\sum_{l\in L}\sum_{l'\in\phi(l)}\frac{1}{2}\parallel\boldsymbol{w}_l-\boldsymbol{w}_{l'}\parallel^2 \tag{24}$$

其中，$\phi(l)$表示 l 的父标签集，$\boldsymbol{w}_i$为标签参数。

5 实验与结果分析

5.1 数据集介绍

本文使用了两个多标签文档分类数据集：MAG-CS 和 PubMed。

MAG-CS[2]：基于微软学术图（MAG）选取了从 1990 年到 2020 年在 105 个计算机科学顶级会议中发表的论文，包含 705 407 个文档和 15 809 个标签。

PubMed[26]：选取了从 2010 年到 2020 年 150 个顶级医学期刊里发表的论文，包括 898 546 篇文章和 17 693 个标签。

这两个数据集的文本信息由标题和摘要组成，元数据信息包括作者、发表期刊（会议）和参考文献，数据集的统计数据如表 1 所列。

表 1 数据集的统计数据

Table 1 Dataset statistics

	MAG-CS	PubMed
# Training Docs	564 340	718 837
# Validation Docs	70 534	89 855
# Testing Docs	70 533	89 854
# Labels	15 809	17 963
# Labels/Doc	5.60	7.78
Vocabulary Size	425 345	776 975
# Words/Doc	126.33	198.97
# Authors	818 927	2 201 919
# Venues	105	150
# Document-Author Edges	2 274 546	5 989 142
# Document-Venue Edges	705 407	898 546
# Document-DocumentEdges	1 518 466	4 455 702
# Edges in Taxonomy	27 288	22 842
# Layers of Taxonomy	6	15

5.2 对比方法

为了验证本文提出的模型的有效性，将本文方法与下面几种多标签文档分类方法以及基于 Transformer 的模型进行对比实验。

XML-CNN[7]：使用卷积神经网络进行多标签文档分类。

MeSHProbeNet[27]：使用递归神经网络和多个 MeSH“探针”进行多标签文档分类。

AttentionXML[8]：使用双向 RNN 网络和一个标签感知的 Attention 层进行多标签文档分类。

Transformer[20]：基于全连接 Attention 机制的多标签文档分类模型。

Star-Transformer[28]：将全连接 Attention 机制的模型改为星型结构的 Attention 多标签文档分类模型。

BERTXML[29]：受 BERT 启发的多标签文档分类模型。

MATCH[4]：元数据引导的基于 Transformer 的多标签文档分类模型。

5.3 实验参数设置

对于所有模型来说，数据的词向量维度 δ 为 100，平衡系数 λ 为 0.1，本文使用 Glove.6B.100d[30] 初始化所有的词向量，使用 Adam 优化器来最小化损失函数，同时 batch size 大小为 256，learning rate 为 2×10^{-3}，最大训练 epoch 数为 20，在运行基线模型时全部使用基线模型的默认参数。

5.4 实验评价指标

本文使用精度（p@k）和归一化折损累计增益（nDCG@k）作为性能评价指标。

$$P@k=\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{y}_d,rank(i) \tag{25}$$

$$DCG@k=\sum_{i=1}^{k}\frac{y_d,rank(i)}{\log(i+1)} \tag{26}$$

$$DCG@k=\sum_{i=1}^{k}\frac{DCG@k}{\min(k,\parallel y_d\parallel_0)\sum_{i=1}^{}\frac{1}{\log(i+1)}} \tag{27}$$

其中，$\boldsymbol{y}_d\in\{0,1\}^{|L|}$ 是文档 d 对应的真实标签向量，$rank(i)$是模型预测的文档第 i 匹配的标签。

5.5 实验结果与分析

表 2 和表 3 列出了本文模型和其他基线模型在 MAG-CS 和 PubMed 数据集上的对比结果，每个实验均进行 3 次，并取 3 次实验的平均值和标准偏差。在对比实验中我们进行了两种消融实验，其中 PGMLDC-NoGraph 代表去掉图神经网络预训练模块，PGMLDC-NoCBLoss 代表去掉平衡长尾分布损失函数模块，采用普通的交叉熵损失函数。

表 2 各模型在 MAG-CS 数据集上的结果对比

Table 2 Results comparison of each model on MAG-CS dataset

Dataset	Method	$P@1=nDCG@1$	$P@3$	$P@5$	$nDCG@3$	$nDCG@5$
MAG-CS	XML-CNN	0.8656±0.0010	0.7028±0.0010	0.5756±0.0010	0.7842±0.0009	0.7407±0.0010
	MeSH ProbeNet	0.8738±0.0016	0.7219±0.0059	0.5927±0.0075	0.8020±0.0048	0.7588±0.0067
	AttentionXML	0.9035±0.0009	0.7682±0.0017	0.6441±0.0020	0.8489±0.0016	0.8145±0.0020
	Star-Transformer	0.8569±0.0011	0.7089±0.0010	0.5853±0.0011	0.7876±0.0008	0.7486±0.0011
	BERTXML	0.9011±0.0027	0.7532±0.0015	0.6238±0.0020	0.8355±0.0025	0.7954±0.0024
	Transformer	0.8805±0.0007	0.7327±0.0006	0.6024±0.0010	0.8129±0.0008	0.7703±0.0010
	MATCH-NoHier	0.9114±0.0014	0.7634±0.0012	0.6312±0.0013	0.8486±0.0006	0.8076±0.0009
	MATCH	0.9190±0.0012	0.7763±0.0023	0.6457±0.0030	0.8610±0.0022	0.8223±0.0030
	PGMLDC-NoGraph	0.9195±0.0010	0.7775±0.0012	0.6463±0.0040	0.8606±0.0016	0.8240±0.0020
	PGMLDC-NoCBLoss	0.9208±0.0015	0.7805±0.0008	0.6493±0.0010	0.8656±0.0006	0.8270±0.0010
	PGMLDC	**0.9238±0.0007**	**0.7872±0.0010**	**0.6543±0.0013**	**0.8703±0.0009**	**0.8332±0.0013**

表 3 各模型在 PubMed 数据集上的结果对比

Table 3 Results comparison of each model on MAG-CS dataset

Dataset	Method	$P@1=nDCG@1$	$P@3$	$P@5$	$nDCG@3$	$nDCG@5$
PubMed	XML-CNN	0.9084±0.0004	0.7182±0.0007	0.5857±0.0004	0.7790±0.0007	0.7075±0.0005
	MeSHProbeNet	0.9135±0.0021	0.7224±0.0066	0.5878±0.0070	0.7836±0.0057	0.7109±0.0065
	AttentionXML	0.9125±0.0003	0.7414±0.0017	0.6169±0.0016	0.7979±0.0013	0.7341±0.0013
	Star-Transformer	0.8962±0.0023	0.6990±0.0014	0.5641±0.0008	0.7612±0.0015	0.6869±0.0011
	BERTXML	0.9144±0.0014	0.7362±0.0046	0.6032±0.0050	0.7949±0.0038	0.7247±0.0045
	Transformer	0.8971±0.0050	0.7299±0.0029	0.6003±0.0018	0.7867±0.0034	0.7178±0.0027
	MATCH-NoHier	0.9151±0.0022	0.7425±0.0041	0.6104±0.0047	0.8001±0.0037	0.7310±0.0044
	MATCH	0.9168±0.0013	0.7511±0.0029	0.6199±0.0029	0.8072±0.0027	0.7395±0.0029
	PGMLDC -NoGraph	0.9179±0.0023	0.7523±0.0025	0.6232±0.0023	0.8114±0.0025	0.7432±0.0030
	PGMLDC -NoCBLoss	0.9181±0.0014	0.7587±0.0030	0.6285±0.0019	0.8160±0.0020	0.7485±0.0023
	PGMLDC	**0.9209±0.0022**	**0.7617±0.0019**	**0.6319±0.0024**	**0.8210±0.0034**	**0.7510±0.0021**

从表 2 和表 3 中可以看出，在数据集 MAG-CS 中，所提方法的准确率比 Transformer 提高了 8%，比 BertXML 提高了 4.75%，比 MATCH 提高了 1.3%。在数据集 PubMed 中，所提方法的准确率比 Transformer 提高了 4.7%，比 BertXML 提高了 3.7%，比 MATCH 提高了 1.6%，且消融实验结果均好于各对比实验结果。

为了评估平衡系数 λ 对结果的敏感性，我们进行了敏感性测试。在测试中，我们系统地改变了 λ 的值，并观察了其在 PGMLDC 模型上对实验结果的影响。表 4 和表 5 列出了不同平衡系数 λ 下的实验结果，通过平衡系数 λ 的敏感性测试，我们最终选取 λ 为 0.1 作为实验的平衡系数。

表 4 MAG-CS 数据集上不同平衡系数 λ 的实验结果对比

Table 4 Experimental results comparison of different balancing coefficients λ on MAG-CS dataset

Dataset	λ	$P@1=nDCG@1$	$P@3$	$P@5$	$nDCG@3$	$nDCG@5$
PubMed	0.050	0.9203±0.0042	0.7628±0.0021	0.6342±0.0012	0.8513±0.0013	0.8197±0.0032
	0.750	0.9229±0.0013	0.7753±0.0012	0.6472±0.0024	0.8640±0.0011	0.8230±0.0016
	0.100	**0.9238±0.0007**	**0.7872±0.0010**	**0.6543±0.0013**	**0.8703±0.0009**	**0.8332±0.0013**
	0.125	0.9219±0.0013	0.7729±0.0020	0.6423±0.0021	0.8626±0.0013	0.8210±0.0013
	0.150	0.9191±0.0017	0.7532±0.0035	0.6310±0.0022	0.8485±0.0015	0.8123±0.0017
	0.200	0.9167±0.0011	0.7427±0.0012	0.6224±0.0036	0.8229±0.0013	0.8003±0.0024
	0.300	0.9131±0.0026	0.7313±0.0009	0.6172±0.0025	0.8135±0.0028	0.7942±0.0015

综上，PGMLDC 模型在两个数据集上均优于所有对比模型，且去掉平衡损失函数和图神经网络的消融实验结果均优于各对比实验结果，结果表明长尾分布平衡损失函数和图神经网络预训练对提高多标签文档分类准确率都是有效的。

表 5 PubMed 数据集上不同平衡系数 λ 的实验结果对比

Table 5 Experimental results comparison of different balancing coefficients λ on PubMed dataset

Dataset	λ	$P@1=nDCG@1$	$P@3$	$P@5$	$nDCG@3$	$nDCG@5$
PubMed	0.050	0.9184±0.0012	0.7328±0.0010	0.6156±0.0010	0.8042±0.0009	0.7207±0.0010
	0.750	0.9199±0.0025	0.7423±0.0016	0.6211±0.0014	0.8110±0.0021	0.7310±0.0012
	0.100	**0.9209±0.0022**	**0.7617±0.0019**	**0.6319±0.0024**	**0.8210±0.0034**	**0.7510±0.0021**
	0.125	0.9195±0.0003	0.7564±0.0027	0.6299±0.0016	0.8199±0.0023	0.7491±0.0013
	0.150	0.9162±0.0013	0.7440±0.0024	0.6141±0.0018	0.8012±0.0015	0.7169±0.0011
	0.200	0.9147±0.0024	0.7362±0.0046	0.6032±0.0050	0.7949±0.0038	0.7107±0.0045
	0.300	0.9119±0.0051	0.7292±0.0019	0.6003±0.0018	0.7867±0.0034	0.7038±0.0027

结束语 本文提出了一种基于图神经网络预训练和平衡长尾分布损失函数的模型 PGMLDC 来解决多标签文档分类问题。模型采用节点级和模式级图神经网络对比学习预训练，将元数据信息构造为异质图，有效捕获了文档及其元数据之间的语义关系，融合了各数据的异质信息，提高了文档与元数据实体嵌入的质量。模型使用平衡长尾分布的损失函数(CBLoss-NTR)，通过重加权交叉熵损失函数，充分利用了输入数据中的标签分布特征，提升了模型对尾部标签分类的准确性。

在两个数据集上进行对比实验，结果表明本文提出的方法相比现有方法具有更好的性能，消融实验的结果也表明，长尾分布平衡损失函数和图神经网络预训练对于提高多标签文档分类准确率均是有效的。实验结果表明，本文所提出的方法对于解决多标签文档分类问题具有实用性和重要性。希望这种方法能够为未来的相关研究提供参考和启示。

参 考 文 献

[1] DONG Y, MA H, SHEN Z, et al. A century of science: Globalization of scientific collaborations, citations, and innovations [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 1437-1446.

[2] WANG K, SHEN Z, HUANG C, et al. Microsoft academic graph: When experts are not enough [J]. Quantitative Science Studies, 2020, 1(1): 396-413.

[3] MINAEE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning-based text classification: a comprehensive review [J].

ACM Computing Surveys(CSUR),2021,54(3):1-40.
[4] ZHANG Y,SHEN Z,DONG Y,et al. MATCH:Metadata-aware text classification in a large hierarchy[C]// Proceedings of the Web Conference 2021. 2021:3246-3257.
[5] AGGARWAL C C,ZHAI C X. A survey of text classification algorithms[M]//Mining text data. 2012:163-222.
[6] HAO C,QIU H P,SUN Y,et al. Research Progress of Multi-label Text Classification[J]. Computer Engineering and Applications,2021,57(10):48-56.
[7] LIU J,CHANG W C,WU Y,et al. Deep learning for extreme multi-label text classification[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017:115-124.
[8] YOU R,ZHANG Z,WANG Z,et al. Attentionxml:Label tree-based attention-aware deep model for high-performance extreme multi-label text classification[C]// Advances in Neural Information Processing Systems 32:Annual Conference on Neural Information Processing Systems. 2019:5812-5822.
[9] ZHANG W,YAN J,WANG X,et al. Deep extreme multi-label learning[C]// Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. 2018:100-107.
[10] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805,2018.
[11] HUANG Y,GILEDERELI B,KÖKSAL A,et al. Balancing methods for multi-label text classi-fication with long-tailed class distribution[J]. arXiv:2109.04712,2021.
[12] CHANG W C,YU H F,ZHONG K,et al. Taming pretrained transformers for extreme multi-label text classification[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020:3163-3171.
[13] GONG J,TENG Z,TENG Q,et al. Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification[J]. IEEE Access,2020,8:30885-30896.
[14] MA Y L,LIU X F,ZHAO L J,et al. Hybrid embedding-based text representation for hierarchical multi-label text classification.[J]. Expert Systems with Applications,2022,187:115905.
[15] TANG D,QIN B,LIU T. Learning semantic representations of users and products for document level sentiment classification[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing(volume 1:long papers). 2015:1014-1023.
[16] KIM J,AMPLAYO R K,LEE K,et al. Categorical metadata representation for customized text classification[J]. Transactions of the Association for Computational Linguistics,2019,7:201-215.
[17] ZHANG Y,SHEN Z,WU C H,et al. Metadata-induced contrastive learning for zero-shot multi-label text classification[C]// Proceedings of the ACM Web Conference 2022. 2022.
[18] YANG P,SUN X,LI W,et al. SGM:sequence generation model for multi-label classification[J]. arXiv:1806.04822,2018.
[19] WANG J,CHEN Z,LI H,et al. Hierarchical multi-label classification using incremental hypernetwork [J]. Journal of Chongqing University of Posts & Telecommunications(Natural Science Edition),2019,31(4):12.
[20] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017:6000-6010.
[21] JIANG X,JIA T,FANG Y,et al. Pre-training on large-scale heterogeneous graph[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021:756-766.
[22] OORD A,LI Y,VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv:1807.03748,2018.
[23] BA J L,KIROS J R,HINTON G E. Layer normalization[J]. arXiv:1607.06450,2016.
[24] CUI Y,JIA M,LIN T Y,et al. Class-balanced loss based on effective number of samples[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:9268-9277.
[25] WU T,HUANG Q,LIU Z,et al. Distribution-balanced loss for multi-label classification in long-tailed datasets[C]// Computer Vision—ECCV 2020:16th European Conference. Springer International Publishing,2020:162-178.
[26] LU Z Y. PubMed and beyond:a survey of web tools for searching biomedical literature[J/OL]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025693/pdf/baq036.pdf.
[27] XUN G,JHA K,YUAN Y,et al. MeSHProbeNet:a self-attentive probe net for MeSH indexing[J]. Bioinformatics,2019,35(19):3794-3802.
[28] GUO Q,QIU X,LIU P,et al. Star-transformer[J]. arXiv:1902.09113,2019.
[29] XUN G,JHA K,SUN J,et al. Correlation networks for extreme multi-label textclassification[C] //Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020:1074-1082.
[30] PENNINGTON J,SOCHER R,MANNING C D. Glove:Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014:1532-1543.

WU Jiawei, born in 1998, postgraduate. His main research interests include multi-label classification, graph neural network and knowledge graph.

FANG Quan, born in 1988, associate professor. His main research interest is multimedia knowledge computing.

(责任编辑:何杨)