Advances in Applied Mathematics 应用数学进展, 2024, 13(5), 2141-2147
Published Online May 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.135202

Hans汉斯

基于Bert的层次多标签文本分类

林　娜

中国地质大学(武汉)数学与物理学院，湖北　武汉

收稿日期：2024年4月28日；录用日期：2024年5月21日；发布日期：2024年5月30日

摘　要

层次多标签文本分类(Hierarchical Multi-label Text Classification, HMTC)是自然语言处理领域(Natural Language Processing, NLP)一项重要的任务。在其由浅至深的标签层次结构中，深层标签更能精确地代表文本所属的标签类别。然而，深层标签的样本实例较少且彼此之间语义接近，导致其难以被正确分类。针对上述的问题，文章提出了基于Bert的层次多标签文本分类方法，先利用Bert构建优越的文本表示，再以自上而下逐层的方式利用浅层级的标签信息引导深层级标签的分类，有效地提升了分类精度。实验结果表明所提模型与其它基线模型相比具有更好的分类性能。

关键词

层次多标签文本分类，Bert，双向长短期记忆网络

Hierarchical Multi-Label Text Classification Based on Bert

Na Lin

School of Mathematics and Physics, China University of Geosciences (Wuhan), Wuhan Hubei

Received: Apr. 28th, 2024; accepted: May 21st, 2024; published: May 30th, 2024

Abstract

Hierarchical Multi-label Text Classification (HMTC) is an important task in the field of natural language processing (NLP). In its shallow-to-deep label hierarchy, deep labels can more accurately represent the label categories to which the text belongs. However, there are fewer sample instances of deep labels and they are semantically close to each other, making it difficult to classify them correctly. To address the above problem, this article proposes a hierarchical multi-label text classification method based on Bert. First, it uses Bert to construct a superior text representation, and then uses the shallow-level label information to guide the classification of deep-level labels in

文章引用: 林娜. 基于 Bert 的层次多标签文本分类[J]. 应用数学进展, 2024, 13(5): 2141-2147.
DOI: 10.12677/aam.2024.135202

a top-down layer-by-layer manner, effectively improving the classification accuracy. The experimental results show that the proposed model has better classification performance compared to other baseline models.

Keywords

HMTC, Bert, Bi-LSTM

1. 引言

随着手机电脑等电子设备的普及以及互联网技术的高速发展，迎来了大数据的时代，越来越多的人选择在互联网上发布信息进行交流，文本是重要的信息载体之一，这使得互联网上文本数据的数量大大增长。如何从互联网上海量且持续增长的文本数据中，高效又准确地获取到有价值的信息，是一个非常值得关注的研究内容，也是自然语言处理领域(NLP)中一个非常重要的研究方向。文本分类是常用的处理文本数据的手段之一，也是自然语言处理领域一项基础且重要的任务。

现实生活中，文本单元(句子、段落、文档等)包含多个标签，且标签之间具有层次结构，指标签可以被组织为树型(Tree)或者有向无环图型(Directed Acyclic Graph, DAG) [1]。将文本分配到上述具有层次结构的标签上，称之为层次多标签文本分类(HMTC)，是多标签文本分类(MTC)的一个重要分支。近年来，人们对层次多标签文本分类越来越感兴趣，其应用场景非常广泛，例如国际专利分类、产品标注、广告推荐等。

层次多标签文本分类任务中不同层级标签之间存在“父子”关系，例如标签“亚洲历史”是“世界历史”的子类。同时，越浅层级的标签越有可能是样本实例较多的父节点标签，而越深层级的标签越有可能是样本量较小且相似度高的叶子节点标签，而模型训练往往趋向于样本量较多的浅层标签，导致深层标签难以被分类。为了提高深层标签的分类精度，提出了基于 Bert 的层次多标签文本分类方法(Hierarchical Multi-label Text Classification Based on BERT)，简称 ML-Bert，ML 表示多个全连接层。ML-Bert 模型利用 Bert 强大的特征学习能力构建优越的文本词嵌入，再以自上而下逐层输出和逐层传递的方式构建多个全连接层对 Bert 做针对性微调，利用父标签信息辅助子标签进行分类，最后通过实验证明了 ML-Bert 模型的有效性和优越性。

2. 相关研究

现有的层次多标签文本分类研究方法大多关注到了层次信息的提取，Aly 等人[1]首次将胶囊网络(Capsule-network)引入到层次多标签文本分类任务中，利用胶囊和动态路由算法对标签特征进行编码。Wehrmann 等人[2]首次提出了混合分类的神经网络模型 HMCN，同时对局部输出和全局输出进行优化。Huang 等人[3]在 HMCN 基础上提出了模型 HARNN，先对文本做 Word2vec 词嵌入，再以自上而下方式注意力机制的方式对文本与每个标签类别的依赖关系做典型注意力建模，捕捉了文本与标签的关联。但 HARNN 在局部和全局选用了相同的文本嵌入表示，影响了模型的分类性能。Zhang 等人[4]在 HARNN 的基础上提出了模型 LA-HCN，以 Glove 作为词嵌入方法，并引入了组件作为中间桥梁搭建标签与文本的表示，采用基于标签的注意力机制建模标签与文本之间的关系，捕捉到了与标签最相关的文本特征信

息。但上述方法都只采取简单了的文本编码方式(静态词向量 Word2vec 和 Glove)，无法捕获上下文语境信息，无法解决“一词多义”的问题。而层次多标签文本分类是多层级多粒度的分类任务，只有构建一开始好的文本表示，才能取得分类所需的层次文本特征，最终取得好的分类效果。

3. 模型介绍

3.1. 层次多标签文本分类定义

给定一系列文档 $D=\{X_1,X_2,\cdots,X_K\}$ 和对应的标签组 $\mathcal{Y}=\{Y_1,Y_2,\cdots,Y_K\}$，其中每篇文档 X_i 由一系列词 $X_i=\{w_1,w_2,\cdots,w_N\}$ 组成，每个标签组由多个标签类别 $Y_i=\{l_1,l_2,\cdots\}$ 组成，标签类别 $\mathcal{L}=\{l_1,l_2,\cdots,l_C\}$ 彼此之间具有层次结构关系 γ，C 表示这一系列文档所包含的标签类别总数。层次多标签文本分类的目标是学习文档 $\mathcal{X}$ 到标签 $\mathcal{L}$ 之间的映射 $\mathcal{P}$，即 $\mathcal{P}:\mathcal{X}\to\mathcal{L}$，通过分析文档内容以及标签的层次结构，再利用映射 $\mathcal{P}$ 对新文档的层次结构标签信息进行预测。

3.2. 模型框架

本节主要介绍 ML-Bert 模型的实现方式。如图 1 所示，ML-Bert 模型主要由特征编码，层级信息整合，混合预测等三个模块组成。下面将通过三小节来具体介绍这三个模块的细节。

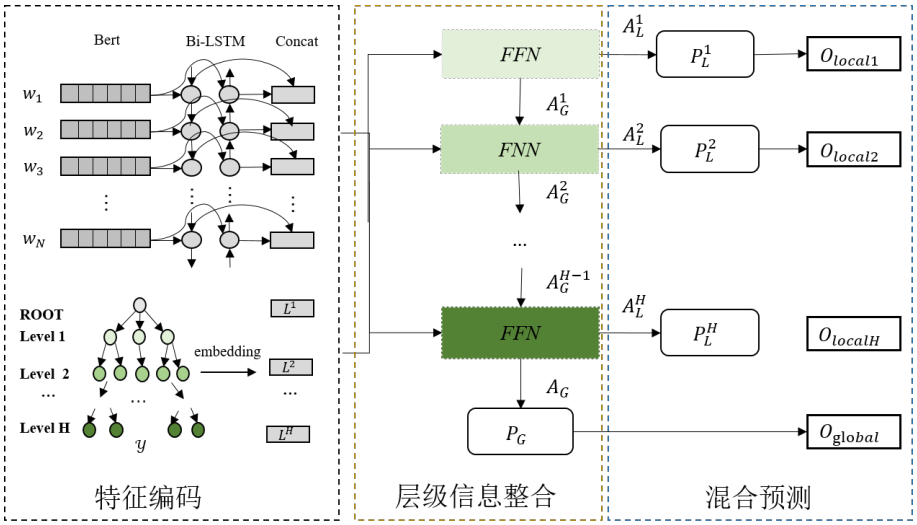

Figure 1. ML-Bert model
图 1. ML-Bert 模型

在特征编码模块，本文利用目前最强大的预训练语言模型 Bert [5]提取文本的上下文关系来构建文本嵌入表示。Bert 通过在大量文本语料上进行无监督学习，学习语言的统计规律和语义表示。在预训练阶段，Bert 的预训练任务包括两个主要部分：遮蔽模型(Masked Language Model, MLM)和下一句预测任务(Next Sentence Prediction, NSP)。与传统的单向语言模型或两个单向语言模型的浅层拼接不同，Bert 采用 MLM 进行预训练，使得每个单词都能同时依赖其上下文进行词向量表示。在 MLM 任务中，Bert 随机遮盖 15%的单词，其中的 80%变成真正的“mask”，10%替换成其他单词，10%保持不变。随后，模型致力于预测这些被遮盖单词的原始内容。Bert 通过上述方式要求模型解构和重新构建文本，以双向学习的方式深入挖掘单词间的依赖关系，从而生成高质量的词向量表示。在 NSP 任务中，Bert 判断两个句子是否连续出现，这有助于模型理解句子之间的关联和上下文信息。

本节具体做法如下，首先，将文本分成一个一个的单词 $D=(w_1,w_2,\cdots,w_N)$，输入到 BERT 中，学习

每个单词的词向量表示 H_i，如下式所示：

$$H_i = BERT\left(w_i\right) \tag{1.1}$$

再将 H_i 输入到 Bi-LSTM 中做语义增强操作得到 $E = \left\{E_1, E_2, \cdots, E_N\right\} \in R^{N\times 2u}$，$u$ 是单向隐藏层大小，$\vec{E}_n$ 和 $\overleftarrow{E}_n$ 分别是 Bi-LSTM 在第 n 处的前向和后向隐藏向量，计算公式如下所示：

$$\vec{E}_n = \overrightarrow{LSTM}\left(\vec{E}_{n-1}, H_i\right) \tag{1.2}$$

$$\overleftarrow{E}_n = \overleftarrow{LSTM}\left(\overleftarrow{E}_{n-1}, H_i\right) \tag{1.3}$$

$$E_n = \left[\vec{E}_n, \overleftarrow{E}_n\right] \tag{1.4}$$

对所有的标签 $L = \left\{l_1, l_2, \cdots, l_C\right\}$ 用正态分布进行随机初始化嵌入，设标签层次结构的总层数为 H，得到每个标签的随机初始化向量，在第 h 层的标签表示为 $L = \left\{l_1^h, l_2^h, \cdots, l_{\left|L^h\right|}^h\right\} \in R^{\left|L^h\right|\times q}$，$\left|L^h\right|$ 表示第 h 层的标签数目，q 表示嵌入的维数。

在层级信息整合模块，利用多个全连接前馈神经网络(FCN)对标签结构中的各个层级进行建模，提取层级文本特征并传递给下一层，辅助深层标签分类。每一层的输入是上一层的文本表示 A_G^{h-1} 和原始文本语义矩阵 E 的拼接，每一层都输入 E 是为了避免原始特征随着层级的加深被稀释，遗漏部分重要的信息。同时也将 A_G^h 输入到局部分类器进行逐层分类，全局信息流 A_G^h 经过一个全连接层的计算公式如下：

$$A_G^h = \varphi\left(W_G^h\left(A_G^{h-1} \oplus E\right) + b_G^h\right) \tag{1.5}$$

其中，$A_G^{h-1} \in R^{\left|L^{h-1}\right|\times d}$ 是上一个层级的信息，E 是原始文本语义矩阵，W_G^h 和 b_G^h 分别是该层的权重矩阵和偏置向量，$\oplus$ 表示向量拼接，将 A_G^h 输入到局部分类器，得到层级文本表示 A_L^h：

$$A_L^h = \varphi\left(W_T^h A_G^h + b_T^h\right) \tag{1.6}$$

将 A_G^h 做平均池化操作再逐层进行拼接，得到用于全局分类的全局文本表示 A_G，公式如下：

$$A_G = avg\left(A_G^1\right) \oplus avg\left(A_G^2\right) \oplus \cdots \oplus avg\left(A_G^H\right) \tag{1.7}$$

混合预测模块旨在同时考虑局部和全局预测的结果，做损失优化计算，以捕捉全面的层次信息。

对于局部分类，通过单层 MLP 计算出局部层级文本表示 A_L^h 与第 h 层标签 L^h 相关的概率 P_L^h，Y_L^h 表示真实的标签情况，计算公式如下：

$$P_L = sigmoid\left(A_L, L^h\right) \tag{1.8}$$

$$O_L = \varepsilon\left(P_L^h, Y_L^h\right) \tag{1.9}$$

其中，$\varepsilon\left(\cdot\right)$ 是二元交叉熵损失(BCE)。对于全局分类，通过双层 MLP 计算出全局文本表示 A_G 与所有标签 L 相关的概率 P_G，计算公式如下：

$$P_G = sigmoid\left(A_G, L\right) \tag{1.10}$$

$$O_G = \varepsilon\left(P_G, Y_G\right) \tag{1.11}$$

最终的概率和损失如下式所示，α 是衡量局部和全局预测重要性的权重超参数。根据以往的经验，将超参数 α 设为 0.5。

$$P_F = \alpha\left(P_1 \oplus P_2 \oplus \cdots + P_H\right) + \left(1-\alpha\right)P_G \tag{1.12}$$

$$O_{all} = O_{Local1} + O_{local2} + \cdots + O_{localH} + O_G \tag{1.13}$$

4. 实验

4.1. 数据集

本文在三个基准数据集 BGC、WIPO 和 WOS 上进行实验，下面具体介绍各个数据集的情况。

Enron 是一个包含大量数字的短篇电子邮件数据集，包含 1648 个文档数，56 个标签，三层结构。其中，第一层包含 3 个标签，第二层包含 40 个标签，第三层包含 13 个标签。Enron 的最细粒度标签不一定是叶子节点标签，即某篇文档实例的标签不一定包含第三层级的标签。

BGC 是由书籍简介和一些与书籍相关的元信息构成的数据集，包含 91,892 个文档数，146 个标签，四层结构。其中，第一层包含 7 个标签，第二层包含 46 个标签，第三层包含 77 个标签，第四层包含 16 个标签。BGC 的最细粒度标签不一定是叶子节点标签，也就是某篇文档实例的标签不一定包含第四层级的标签。

WIPO-alpha 是关于国际专利分类的数据集，包含 75,177 个文档数，5229 个标签，四层结构。其中，第一层包含 8 个标签，第二层包含 114 个标签，第三层包含 451 个标签，第四层包含 4656 个标签。WIPO-alpha 的最细粒度标签必然是叶子节点，但 WIPO-alpha 的标签数量比较大。

WOS 是由科学网(Web of Science)已发表论文的摘要构成的数据集，包含 46,985 个文档数，141 个标签，两层结构。其中，第一层包含 7 个标签，第二层包含 134 个标签。

4.2. 实验设置与评价指标

本文在 bert-bert-uncased [5]上进行文本编码，由 12 个 Transformer encoder 层组成，词嵌入维度是 768，Bi-LSTM 的隐藏层大小设为 256。所有全连接层的神经元数量设定为 512。层次多标签文本分类模型的输出是整个层次结构中所有标签类别的预测概率值。为此，若采用诸如 Precision、Recall 或 F1 分数等评价指标，需要对这些标签类别概率值先设定阈值。然而，实际上最佳阈值的选择是十分困难且主观的。因此，本文采用平均精确率–召回率曲线下的面积 $AU\left(\overline{PRC}\right)$ 作为评价指标。

4.3. 结果分析

本文模型 ML-Bert 与六个基线模型的对比实验结果如表 1 所示，可以看出，本文模型 ML-Bert 在三个数据集(BGC、WIPO、WOS)中表现均最佳，$AU\left(\overline{PRC}\right)$ 值分别达到 83.95%、60.42%和 88.25%，超过了其它所有基线模型，说明 ML-Bert 模型提取到了更充分的深层文本特征信息，利用上一层的先验知识有效地帮助下层进行分类，充分挖掘了标签间的层次结构关系，擅长处理复杂的层次结构数据。

常用的基线模型 Capsule-network、HMCN-F、HMCN-R 在这三个数据集上的性能表现相差不是很大。Capsule-network 利用 fast-text 作词嵌入表示，但仅考虑全局输出的结果。HMCN-F、HMCN-R 分别通过构建混合前馈神经网络层和混合循环神经网络层建模标签结构的关系，同时考虑了局部输出和全局输出的结果，但忽略了文本特征的词嵌入表示。与上述模型相比，本文模型 ML-Bert 在三个数据集上均取得了显著的性能提升，具有更好的泛化能力和综合性能。

基于注意力机制的混合方法模型 HARNN 和 LA-HCN，利用注意力机制挖掘标签与文本特征之间的相关性，并且同时关注标签层次结构的局部和全局信息，相比前三个基线模型，一度占据了较好的水平。LA-HCN 模型是在 HARNN 模型的基础上进行的改进，其对局部分类和全局分类采用了不同的文本表示，得到了更好于 HARNN 的效果。而本文模型 ML-Bert 相较性能表现较好的 LA-HCN，在 BGC、WIPO 和 WOS 三个数据集上分别提高了 1.18%，0.8%和 2.77%的 $AU\left(\overline{PRC}\right)$ 值，进一步证实了本文的方法相较于其他基线方法有显著的优势。

Table 1. Experimental result
表 1. 实验结果

Model	BGC	WIPO	WOS
Capsule-network	78.27	-	-
HMCN-F	79.53	54.87	-
HMCN-R	79.68	55.12	-
HARNN	82.14	58.17	84.23
LA-HCN	82.77	59.65	85.48
Bert	83.10	58.61	87.39
ML-Bert	**83.95**	**60.42**	**88.25**

Bert 模型在 BGC、WIPO 和 WOS 这三个数据集上都达到较好水平，原因在于其利用大规模的语料进行预训练，并且能实时捕捉单词的上下文信息，而本文模型相比于 Bert 有着进一步的提升，$AU\left(\overline{PRC}\right)$ 值分别提高了 0.85%、1.81%和 0.86%，一方面是因为 ML-Bert 在 Bert 的基础上引入了层级信息整合模块，做了针对层次多标签文本分类任务层次结构特点的微调，使得模型更好地学习具有层次结构性的文本特征，另一方面关注到了标签不平衡问题，关注于深层级文本特征的学习，从而提升了模型的分类效果。

从表 1 还可以看到，对于 WIPO 数据集而言，其分类效果远远低于 WOS 和 BGC 数据集。这是由于 WIPO 数据集自身复杂的标签层次结构，如何处理这样复杂的数据集也是层次多标签文本分类任务中的一大难点，而 ML-Bert 利用浅层标签信息辅助深层标签分类，又同时关注局部和全局的情况，充分地对 Bert 进行了针对层次多标签文本分类任务的微调，相比其他基线模型取得了更好的分类性能。

为了进一步验证层级信息整合模块的作用，还对 ML-Bert 的变体 ML-Local 和 ML-Global 进行了消融实验，前者是仅仅考虑了局部输出的结果，后者是仅仅考虑了全局输出的结果，这两者都遗漏了部分层次结构信息。如表 2 所示，ML-Global 利用全局信息流将上一层的先验知识辅助下一层标签进行分类，减少了错误传播问题的发生，分类效果好于 ML-Local。但仅仅是只考虑局部或者全局的方法都不如两者都考虑了的 ML-Bert 分类效果好，这是因为 ML-Bert 综合考虑了局部输出和全局输出的结果，学习到了更全面的层次结构的信息，有效地提高分类模型的性能。

Table 2. Ablation experiment
表 2. 消融实验

Model	BGC	WIPO	WOS
ML-Local	83.29	60.20	87.65
ML-Global	83.81	60.31	88.01
ML-Bert	**83.95**	**60.42**	**88.25**

5. 结论

优化文本表示是层次多标签文本分类任务一直以来的研究重点。为提取更充分的文本特征信息，本文提出了模型 ML-Bert，其利用多个 FCN 层构建层级信息整合模块对 Bert 进行微调，以浅层标签信息帮助深层标签分类，并同时进行局部和全局损失优化，充分地挖掘文本特征信息和标签的层次结构信息。最后通过实验验证了所提模型的有效性和优越性。

参考文献

[1] Aly, R., Remus, S. and Biemann, C. (2019) Hierarchical Multi-Label Classification of Text with Capsule Networks. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*: *Student Research Workshop*, Florence, July 2019, 323-330. https://doi.org/10.18653/v1/P19-2045

[2] Wehrmann, J., Cerri, R. and Barros, R.C. (2018) Hierarchical Multi-Label Classification Networks. *Proceedings of the* 35*th International Conference on Machine Learning*, Stock Holmsmässan, July 2018, 5225-5234. https://doi.org/10.1145/3019612.3019664

[3] Huang, W., Chen, E., Liu, Q., *et al*. (2019) Hierarchical Multi-Label Text Classification: An Attention-Based Recurrent Network Approach. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, New York, November 2019, 1051-1060. https://doi.org/10.1145/3357384.3357885

[4] Zhang, X., Xu, J., Soh, C., *et al*. (2021) LA-HCN: Label-Based Attention for Hierarchical Multi-Label Text Classification Neural Network. *Expert Systems with Applications*, **187**, Article ID: 115922. https://doi.org/10.1016/j.eswa.2021.115922

[5] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1 (Long and Short Papers), Minneapolis, June 2019, 4171-4186.