软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(12):5686−5703 [doi: 10.13328/j.cnki.jos.006802]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

多粒度信息关系增强的多标签文本分类*

李芳芳[1], 苏朴真[1], 段俊文[1], 张师超[1], 毛星亮[2]

[1](中南大学 计算机学院, 湖南 长沙 410038)

[2](湖南工商大学 大数据与互联网创新研究院, 湖南 长沙 410205)

通信作者: 段俊文, E-mail: jwduan@csu.edu.cn; 张师超, E-mail: zhangsc@csu.edu.cn

摘　要: 基于深度学习的多标签文本分类方法存在两个主要缺陷: 缺乏对文本信息多粒度的学习, 以及对标签间约束性关系的利用. 针对这些问题, 提出一种多粒度信息关系增强的多标签文本分类方法. 首先, 通过联合嵌入的方式将文本与标签嵌入到同一空间, 并利用 BERT 预训练模型获得文本和标签的隐向量特征表示. 然后, 构建 3 个多粒度信息关系增强模块: 文档级信息浅层标签注意力分类模块、词级信息深层标签注意力分类模块和标签约束性关系匹配辅助模块. 其中, 前两个模块针对共享特征表示进行多粒度学习: 文档级文本信息与标签信息浅层交互学习, 以及词级文本信息与标签信息深层交互学习. 辅助模块通过学习标签间关系来提升分类性能. 最后, 所提方法在 3 个代表性数据集上, 与当前主流的多标签文本分类算法进行了比较. 结果表明, 在主要指标 Micro-*F*1、Macro-*F*1、*nDCG@k*、*P@k* 上均达到了最佳效果.

关键词: 注意力机制; 多标签文本分类; 标签关系; 多粒度信息

中图法分类号: TP18

中文引用格式: 李芳芳, 苏朴真, 段俊文, 张师超, 毛星亮. 多粒度信息关系增强的多标签文本分类. 软件学报, 2023, 34(12): 5686–5703. http://www.jos.org.cn/1000-9825/6802.htm

英文引用格式: Li FF, Su PZ, Duan JW, Zhang SC, Mao XL. Multi-label Text Classification with Enhancing Multi-granularity Information Relations. Ruan Jian Xue Bao/Journal of Software, 2023, 34(12): 5686–5703 (in Chinese). http://www.jos.org.cn/1000-9825/6802.htm

Multi-label Text Classification with Enhancing Multi-granularity Information Relations

LI Fang-Fang[1], SU Pu-Zhen[1], DUAN Jun-Wen[1], ZHANG Shi-Chao[1], MAO Xing-Liang[2]

[1](School of Computer Science and Engineering, Central South University, Changsha 410038, China)

[2](Institute of Big Data and Internet Innovation, Hunan University of Technology and Business, Changsha 410205, China)

Abstract: Multi-label text classification methods based on deep learning lack multi-granularity learning of text information and the utilization of constraint relations between labels. To solve these problems, this study proposes a multi-label text classification method with enhancing multi-granularity information relations. First, this method embeds text and labels in the same space by joint embedding and employs the BERT pre-trained model to obtain the implicit vector feature representation of text and labels. Then, three multi-granularity information relations enhancing modules including document-level information shallow label attention (DISLA) classification module, word-level information deep label attention (WIDLA) classification module, and label constraint relation matching auxiliary module are constructed. The first two modules carry out multi-granularity learning from shared feature representation: the shallow interactive learning between document-level text information and label information, and the deep interactive learning between word-level text information and label information. The auxiliary module improves the classification performance by learning the relation between labels. Finally, the comparison with current mainstream multi-label text classification algorithms on three representative datasets shows that the proposed

* 基金项目: 国家自然科学基金 (62172449, 61836016, 71790615, 62006251, 62172441); 湖南省自然科学基金 (2021JJ30870, 2021JJ40783); 长沙市自然科学基金 (kq2014134); 国防科技重点实验室基金 (6142101190302)
收稿时间: 2022-06-07; 修改时间: 2022-08-29; 采用时间: 2022-09-21; jos 在线出版时间: 2023-03-15
CNKI 网络首发时间: 2023-03-17

method achieves the best performance on main indicators of Micro-*F*1, Macro-*F*1, *nDCG@k*, and *P@k*.

Key words: attention mechanism; multi-label text classification; label relation; multi-granularity information

文本分类作为自然语言处理方向的一个重要研究任务, 被运用于多个下游任务中. 例如, 在医学大数据领域, 依托于健康医疗信息化, 人们在出现健康问题时往往会通过在线咨询、网络搜索等手段获得相关解答, 而健康医疗系统在处理健康问题时只有获得精准分类健康问句所归属的主题, 才能更好地给出令提问者满意的回答; 在司法大数据领域, 裁判文书作为一种重要信息载体, 在法院公开审判, 给出裁判理由、依据和结果时发挥着极其关键的作用, 对裁判文书中争议焦点的精准提取则是解决纠纷、矛盾的核心问题. 争议焦点的自动识别和检测是我国法治建设中的重要一环, 如何根据裁判文书的内容, 进行争议焦点的自动识别和检测则显得尤为重要. 基于以上, 如何设计一种高效的分类算法来应对海量的数据则显得尤为重要.

现有分类算法可以分为单标签分类算法和多标签分类算法, 在单标签分类算法中, 其针对研究目标的某一特征进行学习与预测, 例如谣言检测[1]、情感极性分析[2]等任务. 而在真实情况下研究目标通常具有多个特征, 多种粒度的信息, 例如针对法律文书的研究, 一篇裁判文书中所涉及的内容通常具有多个相关的罪名, 这样一来, 仅依靠单标签分类算法是无法有效解决此类问题的, 因此多标签分类算法便应运而生. 多标签分类算法包含多种应用场景, 其中应用最为广泛的则是多标签文本分类. 多标签文本分类算法通常应用于推荐系统[3]、意图识别[4]、信息检索[5]、情感分析[6]等方向.

早期的研究中, 研究者们将多标签文本分类机械地视为多个单标签文本分类任务的组合[7], 通过设置多个单标签文本分类算法以应对研究目标不同的特征, 然而这种直接的方式则完全抛弃了标签与标签之间的信息联系, 同时对于标签集合较大的任务, 这无疑加重了任务的计算复杂度. 随着深度学习研究的不断推进, 深度神经网络被广泛应用于多标签文本分类任务中, 并逐渐成为主流方法. 通过深层的神经网络结构来优化文本的特征表示进而提升模型对标签的推理能力. 然而这些方法存在以下问题: 1) 建模粒度过于单一, 缺乏对文本信息进行多粒度地学习, 从而导致在模型获取的信息中存在粗细粒度失衡的问题, 例如基于 CNN 的多标签文本分类方法仅关注局部信息而忽视全局信息, 进而导致模型对于细粒度信息产生过度依赖; 2) 忽略了文本与标签以及标签之间的相关性. 缺乏对文本与标签间相关性和标签间相关性建模, 没有将文本与标签有机地联系起来, 未统筹利用文本标签间相互关系和标签相互间约束性关系. 结合以上分析, 本文归纳总结了当前多标签文本分类领域尚未进行很好处理的两个问题: 1) 如何更好地学习到文本与标签之间的关系以提升算法推理能力; 2) 如何在多标签文本分类中充分利用标签之间的关联性信息辅助算法进行推理.

针对上述两个问题, 本文提出了一种基于多维信息关系增强的多标签文本分类方法, 在预训练语言模型 (BERT) 的基础上, 首先提出文本-标签联合嵌入方式, 通过文本与标签联合嵌入并输入预训练语言模型的方式, 将文本和标签投射到同一语义向量空间中, 得到文本以及每个标签的隐向量表示. 在此基础上提出 3 个子模块, 其中包括文档级信息浅层标签注意力 (document-level information shallow label attention, DISLA) 下的多标签文本分类模块、词级信息深层标签注意力 (word-level information deep label attention, WIDLA) 下的多标签文本分类模块以及标签约束性关系匹配 (label constraint relation matching, LCRM) 辅助模块并行地学习, 其中, 两级多标签文本分类从不同的粒度学习文本与标签之间的相互关系, 而标签约束性关系匹配模块则通过学习标签之间的约束性关系, 隐式地辅助两级多标签文本分类进行推理预测, 最后按照比例加权求和两级多标签分类模块的输出作为预测结果.

本文在 AAPD、LAIC-争议焦点识别提取、医疗公共健康问句数据集上进行了实验, 结果表示本文提出的方法在主要评价指标上均取得了最佳效果, 同时针对文本-标签联合嵌入方式和标签约束性关系匹配模块进行了消融实验, 结果表明本文上述两个部分对于提升多标签文本分类性能具备有效性. 本文的主要贡献如下.

1) 提出了一种多粒度信息关系增强的多标签文本分类方法, 通过文本-标签联合嵌入方式在同一向量空间下学习了文本与标签的隐向量表示及相互关系, 并且通过多粒度学习模块丰富了文本表示, 加强了标签相关性学习.

2) 提出了两个针对不同文本粒度的多标签文本分类模块 (DISLA 和 WIDLA) 来增强文本与标签之间关系, 从

而提升模型对于标签的推理能力.

3) 本文在分别在不同语言, 不同标签量级的代表性数据集上与多个基线算法和前沿算法进行了全面的比较评估, 实验结果表明, 本文提出的方法具备有效性和通用性并且在主要指标上均取得了最好的效果.

1 相关工作

对于多标签文本分类任务, 研究者们的研究重点主要分为文本特征表示学习和标签相关性学习两种类型.

在文本特征表示学习上, 基于深度神经网络的多标签文本分类方法[7]取得了优秀的效果, 这类方法均在未利用标签信息的前提下, 仅使用文本信息作为多标签文本分类的推理依据, 诸如 CNN[8]、RNN[9]、CNN-RNN[10]等方法充分利用了文本信息并生成了丰富的语义表示, 基于注意力机制的方法[11]则从人们观察客观事物的角度出发, 通过从文本序列对每一个元素进行重要程度的学习, 然后根据学习到的重要程度将元素合并, 从而使得算法更加关注文本序列中的重要信息, 以便考虑局部与整体的联系. 随着 BERT[12]预训练语言模型的提出, 自然语言处理方向中各类任务均迎来了关键的转折点, BERT 采用掩码语言模型对双向的 Transformer 结构进行预训练, 从而产生深层的双向语义表示. 基于 BERT 进行多标签分类的方法, 仅在其特殊标记 CLS 后连接一层全连接层进行分类即可超过大部分多标签文本分类算法. 然而这些基于文本特征表示进行学习的多标签文本分类方法, 缺乏对文本的多粒度信息进行学习, 基于 CNN 的方法过于依赖关键词信息, 而缺乏利用全局信息进行标签推理的能力, 当文本中关键词特征不显著时, 其性能则会急剧下降. 相对地, 基于 BERT 的方法则更注重全局信息, 却会使得模型对细粒度词级信息敏感度不足从而影响模型对于关键词与标签之间关系的学习.

在标签相关性学习上, 由于最初的研究中, 由于 binary relevance (BR)[7]其简单易用的特点被研究者们广泛使用在多标签文本分类任务上, 它通过将多标签任务机械地拆解为多个不相关的单标签任务, 而忽略了标签之间的相关性. 逐渐意识到标签之间相关问题的重要性, 研究者们开始探索利用标签相关性学习进行多标签文本分类, label powerset (LP)[13]通过将每组标签划分成新的标签子集的形式将多标签任务转化为单标签多类别任务. classifier chain (CC)[14]通过构建标签链式结构在 BR 算法的基础之上, 利用标签链中前一个标签分类器的预测结果作为当前标签分类器的输入. 随着研究者们对于标签相关性学习的不断深入以及 Seq2Seq 结构[15]的提出, Nam 等人[16]通过将 Seq2Seq 结构引入多标签任务中, 将标签组合视为序列, 通过利用编码器对输入的文本序列进行编码, 再利用解码器依照顺序对标签序列进行解码的方式, 将多标签任务转变为一个标签序列生成任务. 由 Yang 等人[17]提出的 SGM 结构则在引入 Seq2Seq 结构的基础之上, 在解码器部分加入了 attention 机制. 然而上述这些基于链式规则和序列生成的算法和结构, 都一定程度上存在对标签序列顺序过于敏感从而导致其具有泛化能力较弱且易过拟合的特点, 难以预测未存在训练集中的标签顺序和组合.

近些年, 研究者们开始通过标签信息与文本相结合以提升多标签文本分类的效果. LEAM[18]将标签嵌入应用于多标签文本分类任务中, 利用标签所对应的文本描述来为文本和标签生成同一个向量空间的嵌入表示. LSAN[19]则是利用自注意力机制和标签注意力机制来学习标签特定的文本表示. LASA[20]在标签语义信息的基础上学习单词重要性并通过集成重要词汇信息进行多标签文本分类. 这些工作仅重点关注了文本与标签间的交互, 而没有更深入地考虑对于标签间约束性关系的学习. 因此, 本文在前人研究的基础之上从多个粒度对文本与标签、标签与标签之间相互关系进行综合考察, 生成文本与标签深入交互的语义表示以进一步提升多标签文本分类的效果.

2 多粒度信息关系增强的多标签文本分类

2.1 问题描述

多标签文本分类任务是自然语言处理中文本分类方向中一个极具挑战性的子方向, 其具体目标可以描述为: 给定多标签文本分类训练集 $D=\{(D_i,Y_i)|1\leqslant i\leqslant N\}$, 其中 D_i 为输入文本序列, Y_i 为文本对应的标签序列, N 为训练集样本数. 具体地, 将文本序列 D_i 表示为由 d 个字构成的 $D_i=\{x_1,x_2,\ldots,x_d\}$, 将标签序列 Y_i 表示为由 l 个标签类别构成的 $Y_i=\{y_1,y_2,\ldots,y_l\}$. 多标签文本分类的任务目标则是学习一种预测函数 $F:D\to Y=\{(y_1,y_2,\ldots,y_l)|y_i\in\{0,1\}\}$,

来预测输入文本序列的对应标签集合.

2.2 模型框架

本文提出的层级模型结构 MIRE 如图 1 所示, 共分为两层. 其中底层结构为特征表示共享层, 其目标是将文本和标签嵌入到相同的空间中, 再将得到的文本-标签联合嵌入表示输入到 BERT 预训练语言模型中, 得到共享特征表示作为后续模块的共同输入. 顶层结构则为 3 个并行子模块, 由左至右依次为两级多标签文本分类 (分类模块) 和标签约束性关系匹配 (辅助模块). 在分类模块中, 为了充分学习文本信息以及在不同粒度下文本与标签之间的相关性, 本文提出文档级信息浅层标签注意力分类模块, 通过文档级信息提取层直接获取文本与标签在预训练语言模型中初步交互下的文本语义信息 (浅层交互). 词级信息深层标签注意力分类模块则通过标准化嵌入层和文本-标签注意力层, 将特征表示共享层输出的文本标签特征表示通过标准化操作以减少模型过拟合的风险, 并进一步利用深层注意力机制获取词与标签进一步交互下的语义信息 (深层交互). 而在辅助模块中, 则通过标签关系提取层学习不同标签之间的约束性关系, 为分类模块提供标签之间的隐式信息. 整体上, 模型在两级多标签文本分类模块和辅助模块的共同作用下, 能隐式地学习并优化每个标签对应的标签特征表示和标签之间的约束性关系. 最终由两级分类模块得到多标签文本分类的结果.

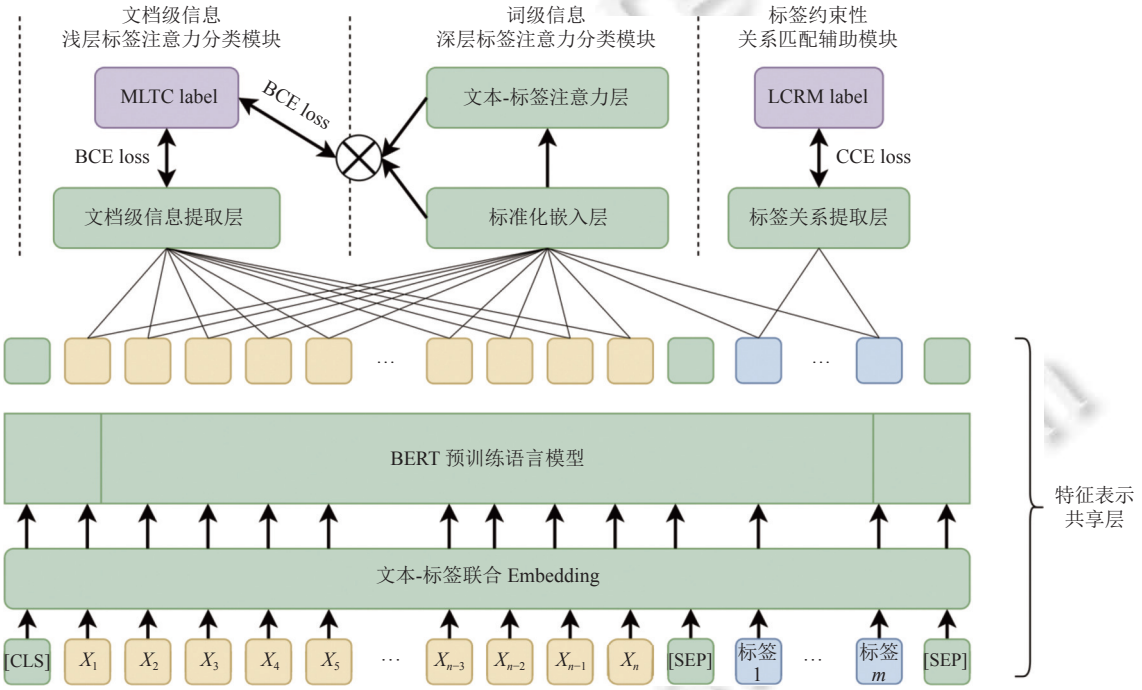


图 1 MIRE 模型框架

2.3 文本-标签联合嵌入

对于自然语言处理任务来说, 文本特征表示所包含语义信息的质量决定着任务效果的好坏, 而随着 BERT 等预训练语言模型的出现, 研究者们能够利用预训练语言模型获得具有丰富语义信息的文本特征表示, 在这些预训练语言模型的基础上, 解决多标签文本分类任务的研究重点逐渐转变为如何充分利用标签的信息来最大化模型进行多标签文本分类的能力. 由于 BERT 等预训练语言模型在进行文本嵌入时受到了输入长度的限制, 以及部分多标签文本分类数据集中标签并不具有明确的文本定义, 这些原因则导致研究者们在处理标签时无法将标签直接嵌入到输入序列中, 以至于无法利用标签的信息提高模型多标签文本分类能力. 基于以上考量, 本文采用文本-标签联合嵌入的方式, 将每一个标签视为独立的字并将其抽象为标签唯一映射 Token, 对于任一标签 $y_i \in \{y_1, y_2, \ldots, y_l\}$, 生成与之唯一对应的 $T_{[\text{Label}_i]}$, 且在标签 Token 嵌入时采用 $T_{[\text{SEP}]}$ 对 Token 嵌入输入进行划分. 故对于任一训练样

本 (D_i, Y_i) 生成 $\{T_{[CLS]}, T_{x_1}, \ldots, T_{x_d}, T_{[SEP]}, T_{[Label_1]}, \ldots, T_{[Label_l]}, T_{[SEP]}\}$ 作为预训练语言模型的 Token 嵌入输入, 从而大幅减少引入标签所占用的序列长度, 同时 BERT 预训练语言模型可以在进行训练时逐步学习到文本-标签、标签-标签之间的关系.

2.4 共享特征表示

共享特征表示层的目标是将文本和标签嵌入到同一向量空间中. 共享特征表示方法如图 2 所示, 首先将文本 X 与标签 Y 通过文本-标签联合嵌入的方式生成 Token 嵌入并与 Segment 嵌入和 Position 嵌入求和得到 BERT 预训练语言模型的输入 $E([X,Y])$.

$$E([X,Y]) = E_{tok}([X,Y]) + E_{seg}([X,Y]) + E_{pos}([X,Y]) \tag{1}$$

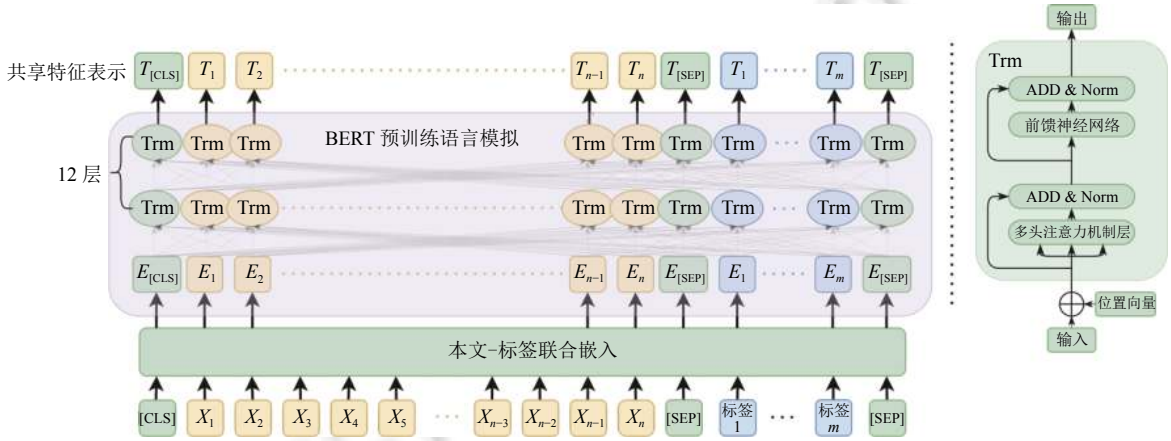


图 2 共享特征表示

然后 BERT 预训练语言模型中首个 Transformer 层会根据初始输入 $E([X,Y])$ 生成 3 个权重矩阵: 查询矩阵 W_{query}、键矩阵 W_{key} 和值矩阵 W_{value}, 并利用这 3 个矩阵分别于输入 $E([X,Y])$ 相乘得到对应向量 $Q_{(0)}, K_{(0)}, V_{(0)}$, 其中:

$$Q_{(0)} = E([X,Y])W_q, K_{(0)} = E([X,Y])W_k, V_{(0)} = E([X,Y])W_v \tag{2}$$

再经过注意力机制计算当前 Transformer 层的输出 $T^{(0)}_{output}$ 并将其作为下一层 Transformer 的输入 $T^{(1)}_{input}$ 以参与新一轮的计算. 过程如公式 (3)、公式 (4) 所示:

$$T^{(1)}_{input} = T^{(0)}_{output} = Softmax\left(\frac{Q_{(0)}K_{(0)}^{\mathrm{T}}}{\sqrt{d_k}}V_{(0)}\right) \tag{3}$$

$$Q_{(1)} = T^{(1)}_{input}W_{query},\ K_{(1)} = T^{(1)}_{input}W_{key},\ V_{(1)} = T^{(1)}_{input}W_{value} \tag{4}$$

其中, $W_{query}, W_{key}, W_{value}$ 均为可训练参数矩阵, d_k 是矩阵 W_{key} 的维度, 其作用在于防止点乘的结果过大而造成梯度过小. 得益于 BERT 中的 12 层 Transformer 结构, 输入中每一个信息单元都与其余信息单元进行了充分的信息交互, 即文本与文本之间、文本与标签之间、标签与标签之间的浅层关系都会经由 Transformer 结构中的多头注意力机制进行计算从而获得交互后的丰富特征表示. 由 BERT 生成的特征表示将作为 3 个并行模块的共享特征表示, 供各模块共同使用、优化.

2.5 两级多标签文本分类模块

在多标签文本分类模块中, 本文针对文档-标签、词-标签两个粒度进行学习, 在文档-标签上, 虽然 BERT 在预训练时于词嵌入中附加了一个特殊且可学习的分类标记 [CLS], 但由于 BERT 预训练时的任务为 NSP (next sentence prediction) 和 MLM (masked language model), 故当嵌入形式为文本-标签联合嵌入时, [CLS] 标记主要作用为判断由 [SEP] 标记分隔的两个部分是否存在前后文关系, 即输入文本序列与输入标签序列是否存在前后文关系. 故直接将 BERT 对应 [CLS] 标记的各层输出进行平均池化来作为文档-标签级特征表示的方式无法有效提供高

质量的语义信息. 基于以上, 本文提出了将共享特征表示中, 文本信息对应位置的特征表示通过文档级信息提取层进行信息集成以实现文档级信息浅层标签注意力 (DISLA) 下的多标签文本分类. 而在词-标签上, 由于观察到人为进行多标签文本分类时, 对于标签的判断往往是取决于文档中某一关键词与标签的关联度, 而仅针对文档-标签单一粒度进行学习则会在判断时丢失细粒度信息, 故本文将 BERT 词级输出与标签输出进行交互以实现词级信息深层标签注意力 (WIDLA) 下的多标签文本分类.

2.5.1 文档级信息浅层标签注意力 (DISLA) 分类模块

DISLA 的目标在于使模型具备利用文档级全局信息进行多标签文本分类的能力. 在 DISLA 中, 由于文本信息与标签信息已经通过 Transformer 结构在预训练语言模型中进行了粗粒度的信息交互, 故本文通过文档级信息提取层对文本信息进行集成, 具体结构如图 3 所示, 首先本文通过 15 层膨胀率为 [1,2,4,8,1,2,4,8,1,2,4,8,1,1,1] 的空洞门卷积以对文本信息进行学习. 在此基础上, 将空洞门卷积的输出输入至多头注意力机制和卷积层中, 最后通过全局最大池化层得到文档级信息浅层标签注意力特征表示 DI_{SLA}, 并以此进行文档级多标签文本分类.

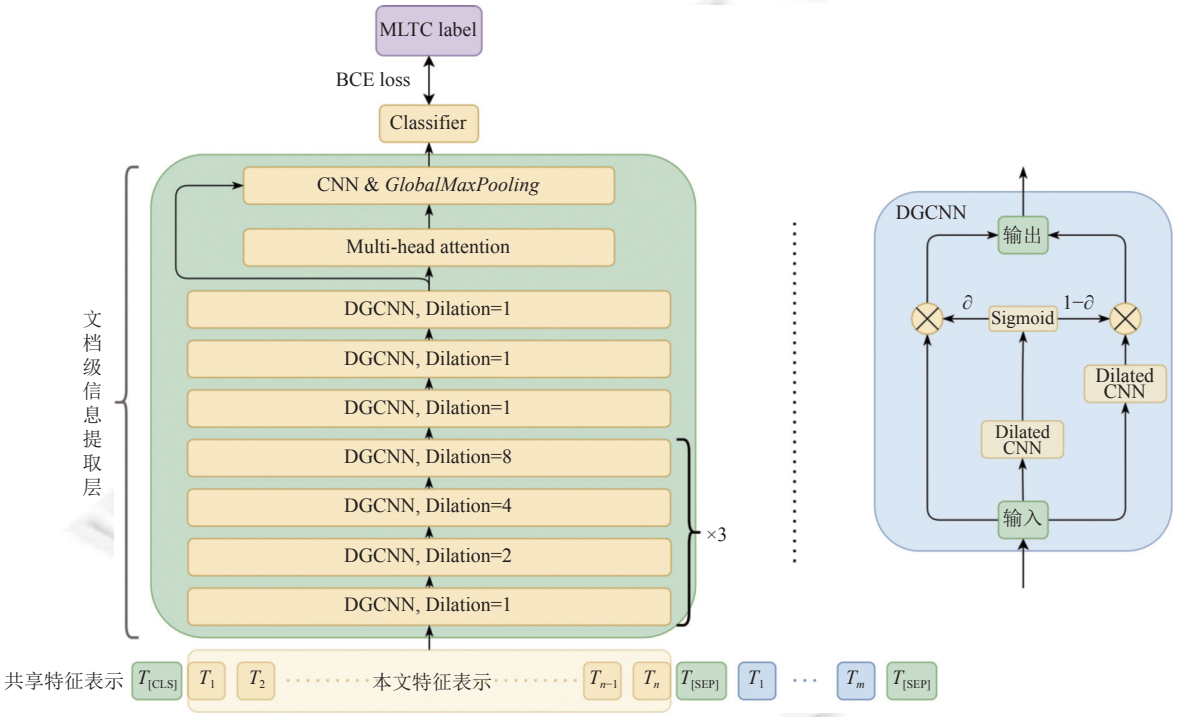


图 3 文档级信息浅层标签注意力多标签文本分类模块结构

空洞卷积是指通过进行卷积时依照膨胀率跳过部分输入, 以实现扩大感受野的目的, 也得益于其不采用池化操作改变特征图大小, 在一定程度上避免了丢失信息等问题. 此外, 由于空洞卷积存在网格效应, 即不采用多层次且相互交错的空洞卷积时, 将会存在导致局部输入丢失的缺点, 同时, 空洞卷积的膨胀率通常按照几何级数增长. 故为了捕获文本与标签信息及其之间的相关性, 本文在设置膨胀率时根据经验采用了 [1, 2, 4, 8], 粒度由细至粗多次学习以捕获多粒度信息及相互关系, 最终由 3 层膨胀率为 1 的空洞门卷积进行细粒度调整以确保信息的完整性.

2.5.2 词级信息深层标签注意力 (WIDLA) 分类模块

WIDLA 的目标在于使模型具备利用词级局部信息进行细粒度多标签文本分类的能力. 具体结构如图 4 所示, 为了加快模型收敛速度, 本文在标准化嵌入层中通过 L2 标准化操作将词级特征表示 $\boldsymbol{T}_{\text{Word}}^{\text{raw}} \in \mathbb{R}^{M\times H}$ 与标签特征表示 $\boldsymbol{T}_{\text{Label}}^{\text{raw}} \in \mathbb{R}^{L\times H}$ 进行标准化处理, 以得到标准化后的词级特征表示 $\boldsymbol{T}_{\text{Word}} \in \mathbb{R}^{M\times H}$ 和标签特征表示 $\boldsymbol{T}_{\text{Label}} \in \mathbb{R}^{L\times H}$, 其中 M 为文本序列长度, L 为标签序列长度, H 为特征表示向量维度:

$$T_{\text{Word}} = L2\_Norm\left(T_{\text{Word}}^{\text{raw}}\right) \tag{5}$$

$$T_{\text{Label}} = L2\_Norm\left(T_{\text{Label}}^{\text{raw}}\right) \tag{6}$$

$$L2\_Norm([x_1, x_2, \ldots, x_n]) = \left[\frac{x_1}{\sqrt{\sum_{i=1}^{H} x_{1i}^2}}, \frac{x_2}{\sqrt{\sum_{i=1}^{H} x_{2i}^2}}, \ldots, \frac{x_n}{\sqrt{\sum_{i=1}^{H} x_{ni}^2}}\right] \tag{7}$$

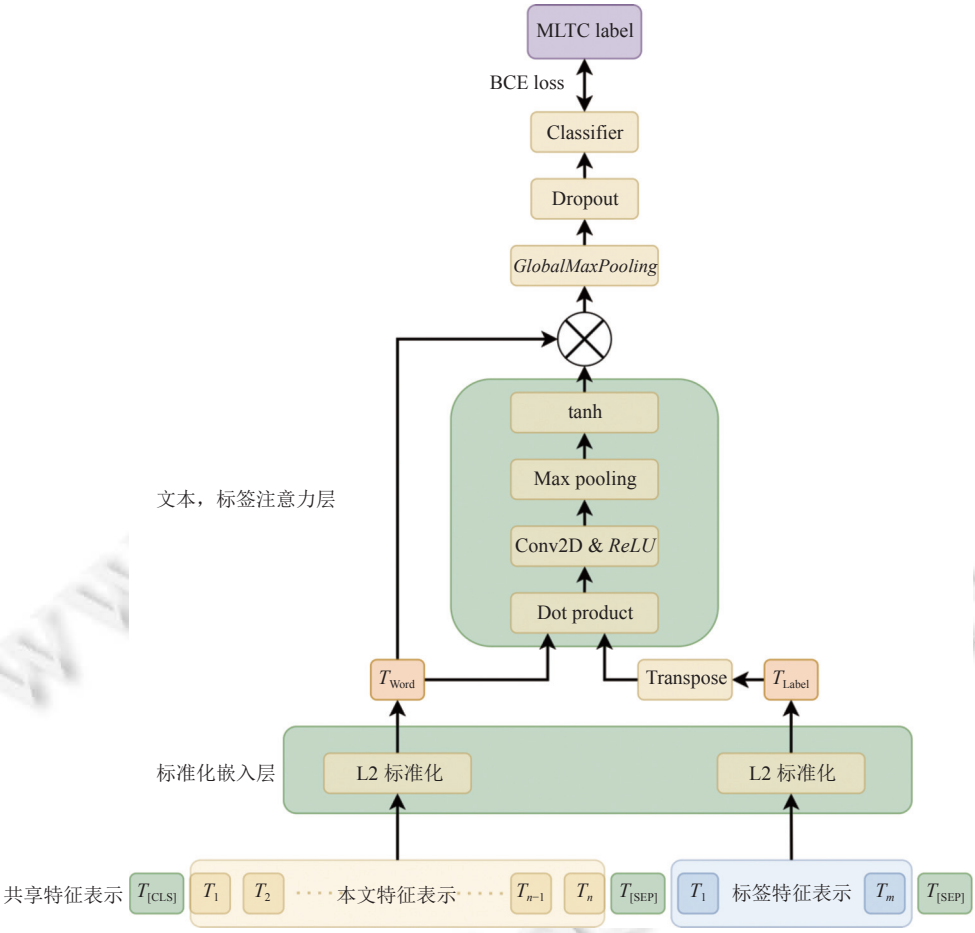


图 4 词级信息深层标签注意力多标签文本分类模块结构

考虑到文本中各词与各标签之间关联强度存在差异, 为了深入挖掘细粒度文本标签关系信息, 本文在文本-标签注意力层中, 通过点积操作:

$$W_{DLA} = T_{\text{Word}} T_{\text{Label}}^{\text{T}} \tag{8}$$

使文本信息与标签信息充分交互, 以此得到词级信息标签注意力特征表示 $W_{DLA} \in \mathbb{R}^{M \times L}$, 接着使用激活函数为 *ReLU* 的卷积神经网络从特征表示 W_{DLA} 中提取深层交互信息 $\vec{H}$, 以满足充分利用特征表示 W_{DLA}, 并突出每个词在文本中不同的重要程度以及与标签关联程度的目的.

最后, 通过 MaxPooling 函数和 tanh 函数计算每个词对标签预测的重要性得分 $\vec{a}$, 最后将 $\vec{a}$ 与 T_{Word} 相乘结合 *GlobalMaxPooling* 函数得到词级信息深层标签注意力特征表示 $WI_{DLA} \in \mathbb{R}^{1 \times H}$, 并以此进行词级信息深层标签注意

力多标签文本分类:

$$\overrightarrow{H} = ReLU\left(Conv2D\left(W_{DLA}\right)\right) \tag{9}$$

$$\overrightarrow{a} = \tanh\left(\overrightarrow{H}\right) \tag{10}$$

$$WI_{DLA} = GlobalMaxPooling(\overrightarrow{a} \cdot T_{\text{Word}}) \tag{11}$$

2.6 标签约束性关系匹配 (LCRM) 辅助模块

标签约束性关系匹配 (label constraint relation matching, LCRM) 辅助模块的目标是通过学习标签间约束性关系隐式地辅助分类模块进行多标签文本分类. 在 LCRM 中, 本文首先统计标签集合的分布, 依次统计每个标签与剩余标签同时出现的次数, 对于给定的标签集合 $Y = \{y_1, y_2, \ldots, y_l\}$, 通过公式 (9) 得到标签关联性矩阵:

$$M_{\text{Label}} = \sum_{i=0}^{l}\sum_{j=0}^{l} Count\left(y_i, y_j\right) \tag{12}$$

其中, M_{ij} 表示标签 y_i 与标签 y_j 在训练数据集中同时出现的次数, 根据得到的标签关联性矩阵 M_{Label} 按行依次进行排序, 得到每个标签与其余标签同时出现次数的递增序列 $M_i^{\text{sort}} = \{M_{ia}, \ldots, M_{ib}, M_{ii}\}$, 其中 $0 \leqslant a, b, i \leqslant l$, $0 \leqslant M_{ia} \leqslant M_{ib} \leqslant M_{ii}$, 并根据平均值

$$M_i^{\text{mean}} = \frac{1}{l-1}\sum_{j=0, j\neq i}^{l} M_{ij} \tag{13}$$

对递增序列 M_i^{sort} 进行划分得到低频次数序列 $M_i^{\text{lowfreq}} = \{M_{ia}, \ldots, M_{ib}\}$ 以及对应标签序列 $Y_i^{\text{lowfreq}} = \{y_a, \ldots, y_b\}$, 其中 $0 \leqslant a, b \leqslant l$, $0 \leqslant M_{ia} \leqslant M_{ib} \leqslant M_i^{\text{mean}}$.

为了更好地使模型学习标签与标签之间的约束性关系, 本文根据文本标签相关性和标签共同出现频率将标签划分为 Y^{+*}、Y^{+}、Y^{-} 这 3 个集合, 其中对于 3 个集合的定义如下.

1) 标签 $y_i^{+*} \in Y^{+*}$ 当且仅当 y_i^{+*} 与文本相关, 即对于输入样本 $\left(D_j, Y_j\right)$, $y_i^{+*} \in Y_j = \{y_1, y_2, \ldots, y_l\}$, 且 $y_i^{+*} = 1$.

2) 标签 $y_i^{+} \in Y^{+}$ 当且仅当 y_i^{+} 与文本无关, 但与当前文本对应标签 y_i^{+*} 有着高频共同出现率, 即对于输入样本 $\left(D_j, Y_j\right)$, $y_i^{+} \in Y_j = \{y_1, y_2, \ldots, y_l\}$, 且 $y_i^{+} = 0$, $M_{y_i^{+*}}^{\text{mean}} < M_{y_i^{+*}y_i^{+}}$.

3) 标签 $y_i^{-} \in Y^{-}$ 当且仅当 y_i^{-} 与文本无关, 但与当前文本对应标签 y_i^{+*} 有着低频共同出现率, 即对于输入样本 $\left(D_j, Y_j\right)$, $y_i^{-} \in Y_j = \{y_1, y_2, \ldots, y_l\}$, 且 $y_i^{-} = 0$, $M_{y_i^{+*}}^{\text{mean}} > M_{y_i^{+*}y_i^{-}}$, 其中 $y_i^{-} \in Y_{y_i^{+*}}^{\text{lowfreq}}$.

基于以上定义, 本文从输入样本中与文本相关的标签集合 Y^{+*} 中随机抽取一个标签 y_i 作为匹配样本 A, 再随机从剩余标签中随机抽取一个标签 y_j 作为匹配样本 B. 若匹配样本 $B \in Y^{+*}$, 则将 LCRM 标签记作 2, 直接相关; 若匹配样本 $B \in Y^{+}$, 则将 LCRM 标签记作 1, 间接相关; 若匹配样本 $B \in Y^{-}$, 则将 LCRM 标签记作 0, 不相关. 如图 1 所示, 本文通过标签关系提取层完成上述过程, 将 A 和 B 的特征表示进行拼接输入至激活函数为 *Softmax* 的全连接层中得到预测结果.

2.7 损失函数及标签预测

MIRE 在两个多标签分类模块上均采用二元交叉熵损失 (binary cross entropy loss)[21]作为损失函数, 在标签相关性匹配模块上采用多类交叉熵损失 (categorical cross entropy loss)[22]作为损失函数, 两类函数均在分类任务中广泛使用, 二元交叉熵损失和多分类交叉熵损失定义如下:

$$L_{\text{bce}} = -\sum_{i=1}^{N}\sum_{j=1}^{l} y_{ij}\log\left(\widehat{y}_{ij}\right) + \left(1-y_{ij}\right)\log\left(1-\widehat{y}_{ij}\right) \tag{14}$$

$$L_{\text{cce}} = -\sum_{k=1}^{c} y_k \log\left(\widehat{y}_k\right) \tag{15}$$

其中, N 为样本总数, l 为多标签文本分类模块中标签总数, c 为标签相关性模块中标签总数, $y_{ij} \in \{0,1\}$, $\widehat{y}_{ij} \in [0,1]$,

$y_k \in \{0,1,2\}$, $\widehat{y_k} \in [0,2]$ 依次为多标签文本分类中第 i 个样本对应标签集合 Y_i 中第 j 个标签的真实值和预测值, 以及标签相关性模块中第 k 个样本对应标签的真实值和预测值. 并行子模块在训练时的整体损失函数定义如下:

$$L_{\text{all}} = \alpha L_{\text{DISLA}} + (1-\alpha) L_{\text{WIDLA}} + \beta L_{\text{LCRM}} \tag{16}$$

其中, α,β 为超参数, 通过多次实验, 最终设置 $\alpha = 0.4, \beta = 0.15$. L_{LCRM}, L_{DISLA} 和 L_{WIDLA} 分别为标签相关性匹配模块, 文档级信息浅层标签注意力多标签文本分类模块和词级信息深层标签注意力多标签文本分类模块的损失. 此外, 本文在标签的预测阶段, 将 DISLA 和 WIDLA 预测的结果分别乘以 α 和 $1-\alpha$ 再求和作为最终的预测结果.

3 实验分析

本文在 3 个数据集 (AAPD, LAIC-争议焦点识别提取数据集, 医疗公众健康问句数据集) 上进行实验. 在 AAPD 数据集上同算法 XML-CNN、LASA、SGM、BERT(cls)、BERT(pooled) 使用评价指标 $P@k$、$nDCG@k$、Micro-$F1$、Macro-$F1$ 进行评估对比, 在 LAIC-争议焦点识别提取数据集和医疗公众健康问句数据集上同算法 XML-CNN、BERT(cls)、BERT(pooled) 使用 $P@k$、$nDCG@k$、Micro- $F1$、Macro-$F1$ 进行评估对比.

3.1 实验设置

3.1.1 数据集

本文所采用的 3 个多标签文本分类数据集为: LAIC-争议焦点识别提取数据集、AAPD、医疗公众健康问句数据集, 具体如表 1 所示.

表 1 实验数据集

Dataset	train	test	Labels	Avg. document	Avg. label
AAPD	53 840	1 000	54	163.42	2.41
LAIC	10 000	3 690	99	550.4	5.05
医疗公共健康问句	4 000	1 000	6	110.7	1.31

LAIC-争议焦点识别提取数据集 (http://data.court.gov.cn/pages/laic2021.html): LAIC 竞赛数据集, 来自由浙江省高级法院提供并标注的法院裁判文书, 包含13 690 篇裁判文书以及人工额外标注的 99 个争议焦点, 其中所涉及的裁判文书均为民事判决书, 涉及的案由包括民间借贷、离婚、机动车交通事故责任、金融借款合同等.

AAPD (https://git.uwaterloo.ca/jimmylin/Castor-data/-/tree/master/datasets/AAPD): 计算机科学领域的大型数据集, 包含 arXiv 上 55 840 篇论文摘要, 共 54 个标签.

医疗公共健康问句数据集 (https://www.heywhale.com/home/competition/5f2d0ea1b4ac2e002c164d82): 医学数据挖掘算法评测大赛数据集, 包含 5 000 条健康有关的中文问句, 对问句的主题进行分类, 共包含 6 个标签.

3.1.2 评价指标

在本文中, 我们采用常用微/宏观 $F1$ (Micro/Macro-$F1$)、微宏观精度 (Micro/Macro-P)、微宏观召回率 (Micro/Macro-R)、精度 (precision at k, $P@k$)、归一化折损累积增益 (normalized discounted cumulative gain at k, $nDCG@k$)[23]作为性能比较的评价指标. 具体定义如下:

$$\text{Micro-}F1 = \frac{\sum_{i=1}^{L} 2TP_i}{\sum_{i=1}^{L} 2TP_i + FP_i + FN_i} \tag{17}$$

$$\text{Macro-}F1 = \frac{1}{L}\sum_{i=1}^{L} \frac{2TP_i}{2TP_i + FP_i + FN_i} \tag{18}$$

其中, i 表示标签集合中第 i 类标签, TP_i, FP_i, FN_i 分别代表真正例, 假正例, 假负例.

$$P@k = \frac{1}{k}\sum_{i \in \text{True}(\bar{y})}^{k} y^i \tag{19}$$

$$DCG@k = \sum_{i\in \mathrm{True}(\hat{y})}^{k} \frac{y^i}{\log(i+1)} \tag{20}$$

$$nDCG@k = \frac{DCG@k}{\sum_{i=1}^{\min(k,|y|)} \frac{1}{\log(i+1)}} \tag{21}$$

其中, $\mathrm{True}(\hat{y})$ 为预测结果中预测值最高的前 k 个标签, $|y|$ 为真实标签中相关标签的个数, 对于测试集的 $P@k$ 、$nDCG@k$, 本文对所有样本进行求值后取均值.

3.1.3 对比算法

为了验证本文所提出方法的有效性以及通用性, 本文在 AAPD 数据集上选择 BR、LP、CC、XML-CNN、SGM、Seq2set、LEAM、LSAN、OCD、HTTN、ML-R、LASA、BERT(cls)、BERT(pooled) 这 14 种算法作为对比算法以验证有效性, 在剩余两个中文数据集上选择 XML-CNN、BERT(cls)、BERT(pooled) 这 3 种算法作为对比算法以验证有效性和通用性.

BR (binary relevance, BR)[7]: 将多标签分类问题进行分解, 转换为多个独立的二分类问题, 其中每个二分类问题对应一个标签空间中待遇测的标签.

LP (label powerest, LP)[13]: 将多标签分类转化为多分类的集成, 每个成分学习器面向标签集合的一个子集, 通过 label powerset 训练一个多分类器.

CC (classifier chains, CC)[14]: 将多标签学习转化为一个链式的二分类问题, 其中链路上后续的二分类器建于先前二分类器的预测之上.

XML-CNN[24]: 使用卷积神经网络并利用动态池化提取特征进行多标签文本分类, 并在和输出层增加隐藏单元, 建模复杂组合关系, 是使用卷积神经网络处理文本分类的代表性算法.

SGM[17]: 将多标签分类任务视为一个序列生成问题, 同时应用序列生成模型以及解码器结构来捕捉标签之间相关性.

Seq2set[25]: 在 SGM 算法的基础上增添 Set 解码器, 通过利用 Set 的无序性来降低错误的标签排序对多标签文本分类带来的影响.

LEAM[18]: 以注意力机制为基础, 将文本和标签进行联合嵌入, 利用两者之间的相关性构建文本表示, 从而获得更具识别性的文本表征.

LSAN[19]: 提出了一种自适应融合策略将文本自注意力机制和标签注意力机制所获得的表示进行融合, 并以此为每个文档建立标签特定的表示进行分类.

OCD[26]: 提出了一种在训练阶段采用 BR 解码器和 RNN 解码器联合训练, 并采用 OCD 对 Seq2Seq 模型进行优化, 同时在推理阶段组合两个解码器输出概率作为预测的新框架, 从而缓解暴露偏差.

HTTN[27]: 通过语义提取器将元知识从数据丰富的头标签迁移到数据贫乏的尾标签, 从而解决长尾标签问题.

ML-R[28]: 使用二元分类器同时预测所有标签, 并应用一种新的迭代推理机制以有效利用标签间信息, 其中每个推理实例将所有标签的先前预测似然度作为附加输入.

LASA[20]: 基于标签语义信息进行单词重要性学习, 得到文本中各单词的重要性, 并通过注意力机制得到文档与各标签间的匹配得分, 最后将得分与各文档中单词表示相结合以进行多标签文本分类的算法.

BERT(cls)[12]: 利用 BERT 预训练语言模型输出中的特殊标志 cls 进行多标签文本分类的算法.

BERT(pooled)[12]: 将 BERT 预训练语言模型输出中每个词的信息平均池化后, 进行多标签文本分类的算法.

3.1.4 参数设置

对于 AAPD 数据集, 本文采用 12 层的 BERT-Base-Cased 作为预训练语言模型, 其中向量维度为 768, 共 110 M 个参数, 对于其他两个中文数据集, 采用 12 层的 BERT-Base-Chinese 作为预训练语言模型, 其中向量维度为 768, 共 110 M 个参数. 模型采用 Adam[29]优化器进行训练, 学习率为 5E–5, 在 AAPD 数据集上 Batch size 设置为 32, 最

大输入序列长度为 320, 在其他两个中文数据集上 Batch size 设置为 16, 最大输入序列长度为 512. 通过模型在验证集上的 Micro-*F*1 得分表现来结束训练模型, 如果连续 10 个 epoch 后 Micro-*F*1 得分仍没有增加, 则停止训练. 此外, 针对 LAIC 数据集, 本通过分析文本信息分布特征, 将最大文本长度设置为 410, 分类截断超出最大长度的文档: 1) 长度在 [411, 616] 的文档按照首部截断, 仅保留文档前 410 个词; 2) 长度在 [617, +∞) 的文档按照首尾截断, 将文档前 205 个词和最后 205 个词进行拼接保留.

3.2 实验结果与分析

本文提出的 MIRE 在 AAPD 数据集上与其他 14 个算法进行对比, 在 LAIC、医疗健康公共问句数据集上与具有代表性且表现优异的算法进行对比, 评价指标情况见表 2–表 4. 每个评价指标中最好的结果以粗体标出. 其中 *P* 和 *R* 分别代表 precision 和 recall 值, (+) 表示值越高则模型效果越好, 标有*的算法为复现后得到的结果, 未标记*的为直接引用文献的结果, “—”为文献中未给出对应评价指标的结果.

表 2 AAPD 数据集实验结果

算法	Micro (+)			Macro (+)			*nDCG* (+)			*P* (+)		
	P	*R*	*F*1	*P*	*R*	*F*1	@1	@3	@5	@1	@3	@5
BR	0.644	0.648	0.646	—	—	—	—	—	—	—	—	—
LP	0.662	0.608	0.634	—	—	—	—	—	—	—	—	—
CC	0.657	0.651	0.654	—	—	—	—	—	—	—	—	—
XML-CNN*	0.777	0.600	0.677	0.640	0.408	0.465	0.800	0.760	0.799	0.800	0.574	0.393
SGM*	0.746	0.659	0.664	0.560	0.435	0.475	0.757	0.724	0.762	0.757	0.567	0.366
Seq2set	0.739	0.674	0.705	—	—	—	—	—	—	—	—	—
LEAM	0.765	0.596	0.670	0.524	0.403	0.456	—	—	—	—	—	—
LSAN	0.777	0.646	0.706	**0.676**	0.472	0.535	—	—	—	—	—	—
LASA	—	—	—	—	—	—	0.839	0.798	0.837	0.839	0.600	0.408
OCD	—	—	0.720	—	—	0.585	—	—	—	—	—	—
HTTN	0.778	0.617	0.688	0.481	0.379	0.416	—	—	—	—	—	—
ML-R	0.726	0.718	0.722	—	—	—	—	—	—	—	—	—
BERT(cls)*	0.739	0.718	0.729	0.600	0.568	0.574	0.823	0.797	0.831	0.823	0.607	0.409
BERT(pooled)*	**0.779**	0.685	0.729	0.673	0.551	0.582	0.847	0.809	0.848	0.847	0.613	0.418
MIRE	0.771	**0.720**	**0.744**	0.653	**0.579**	**0.595**	**0.862**	**0.823**	**0.860**	**0.862**	**0.624**	**0.424**

表 3 LAIC 数据集实验结果

算法	Micro (+)			Macro (+)			*nDCG* (+)			*P* (+)		
	P	*R*	*F*1	*P*	*R*	*F*1	@1	@3	@5	@1	@3	@5
XML-CNN*	**0.657**	0.346	0.453	**0.555**	0.202	0.267	0.562	0.522	0.516	0.562	0.359	0.264
BERT(cls)*	0.599	0.490	0.539	0.499	0.359	0.396	0.622	**0.699**	0.717	0.622	0.402	0.297
BERT(pooled)*	0.547	0.485	0.514	0.447	0.379	0.402	0.580	0.552	0.548	0.580	0.379	0.280
MIRE	0.608	**0.520**	**0.560**	0.527	**0.388**	**0.422**	**0.636**	0.683	**0.729**	**0.636**	**0.409**	**0.301**

表 4 医疗公共健康问句数据集实验结果

算法	Micro (+)			Macro (+)			*nDCG* (+)			*P* (+)		
	P	*R*	*F*1	*P*	*R*	*F*1	@1	@3	@5	@1	@3	@5
XML-CNN*	0.741	0.734	0.737	0.610	0.445	0.487	0.790	0.768	0.768	0.790	0.412	**0.263**
BERT(cls)*	**0.878**	0.855	0.866	**0.709**	0.639	0.668	0.900	0.889	0.889	0.900	0.427	**0.263**
BERT(pooled)*	0.867	0.870	0.868	0.674	0.675	0.674	0.903	0.890	0.890	0.903	0.425	**0.263**
MIRE	0.869	**0.891**	**0.880**	0.686	**0.696**	**0.691**	**0.923**	**0.958**	**0.964**	**0.923**	**0.431**	**0.263**

根据 AAPD 数据集上的实验结果可知, MIRE 在主要指标上如 Micro-*F*1、Macro-*F*1 以及 *nDCG@k* 与 *P@k* 上均达到了最佳的效果, 同时可以得出以下分析结论.

1) 诸如 BR, LP 此类机器学习算法在进行多标签文本分类时的效果相较于深度学习算法均稍逊一筹. 而基于预训练语言模型的算法在多个评价指标上的效果全面优于大部分深度学习算法, 其原因在于, 相比与机器学习算法, 深度学习算法以及基于预训练语言模型的算法能通过训练数据集充分学习文本中的语义信息为多标签文本分类提供高质量的推理依据, 而基于预训练语言模型的算法则通过预训练进一步学习了更为丰富的语义信息, 从而在表现上更为优异.

2) 观察深度学习算法中的 XML-CNN 算法, 可以发现 XML-CNN 算法在 Micro-*P*、Macro-*P* 上的表现超过大部分深度学习算法, 甚至超过基于预训练语言模型的 BERT(cls) 算法并且接近 BERT(pooled) 的表现, 而其他指标上的效果均表现不佳, 其原因在于, XML-CNN 算法在进行计算时往往是针对局部特征进行卷积, 并且由于其中的最大池化部分会进一步地扩大算法对于局部特征的关注, 如此一来 XML-CNN 算法对于标签的判定变得十分敏感, 一旦出现近似与标签相关的词, 即被判定为相关标签, 故 XML-CNN 在 Micro-*R*、Macro-*R* 上的表现均不佳.

3) 通过将基于预训练语言模型的算法与各类深度学习算法进行比较, BERT(cls) 与 BERT(pooled) 算法在 Micro-*F*1 上的效果优于各类深度学习算法, 在 Macro-*F*1 上的效果仅略低于 OCD 算法, 而 BERT(pooled) 算法在 *nDCG@k* 、*P@k* 两个指标上均有着超越各类深度学习算法的效果. 同时值得注意的是, 对比两类基于预训练语言模型的算法可以发现, BERT(cls) 更关注全局信息, 而 BERT(pooled) 更关注局部信息, 根据两类算法的 Micro/Macro-*P*、Micro/Macro-*R* 指标上可以看出这一显著区别. 故综合来看, 基于预训练语言模型的算法在多标签文本分类任务上的效果相比于深度学习算法更具代表性、有效性, 因此本文在后续中文数据集上仅于深度学习算法中关注局部特征的 XML-CNN 以及两类基于预训练语言模型的算法进行对比以验证 MIRE 的有效性和通用性.

整体来说, 本文提出的 MIRE 相比于各类算法均存在优势, 主要体现在以下几点.

1) 重点关注文本信息部分的 XML-CNN, 在 Micro-*F*1、Macro-*F*1 上相对提升了 9.9% 和 27.9%, 以及在 *nDCG@k* (*k* =1, 3, 5)、*P@k* (*k* =1, 3, 5) 上相对提升了 7.8%、8.3%、7.6% 和 7.8%、8.2%、7.9%, 充分说明了 MIRE 的两级多标签文本分类模块在文本信息特征提取上对于提升分类综合性能更为有效.

2) 同基于 Seq2Seq 模型的算法如 SGM、Seq2set 算法相比较, 在各项指标上均有着最佳的表现, 说明本文提出的 LCRM 辅助模块在标签信息及其相关性的学习对于分类效果的提升更具优势.

3) 同文本标签联合嵌入的 LEAM 相比, 在 Micro-*F*1 与 Macro-*F*1 上相对提升了 11.0% 和 30.5%, 说明本文中标签-文本联合方式更具备优越性.

4) 同学习了标签语义信息的算法如 LSAN、LASA 相比较, 相比于 LSAN 算法, 在 Micro-*F*1 与 Macro-*F*1 上相对提升了 5.4% 和 11.2%. 相比与 LASA 算法, 在 *nDCG@k* (*k* =1, 3, 5)、*P@k* (*k* =1, 3, 5) 上相对提升了 2.7%、3.1%、2.7% 和 2.7%、4.0%、3.9%, 说明本文提出的两级多标签文本分类模块对于整体模型性能提升的有效性.

5) 同 OCD、HTTN、ML-R 此类为了解决标签问题的深度学习算法相比, 在 Micro-*F*1 与 Macro-*F*1 上得到最佳效果, 说明本文提出的算法能学习到标签之间的信息以解决长尾标签、暴露偏差等问题.

从 LAIC 数据集与医疗公共健康问句数据集上得出的结果, 可以看出本文提出的 MIRE 在主要评价指标 Micro/Macro-*F*1、*nDCG@k* (*k* =1, 5)、*P@k* (*k* =1, 3, 5) 上均取得了最高得分, 充分说明了本文提出的 MIRE 在中文数据集、不同标签量的数据集上仍然具备有效性, 进而验证了模型的通用性.

此外, 可以通过对评价指标 Micro/Macro-*P* 的变化在两个数据集上的不同表现进行深入分析, 当面临标签数量较多 (标签数为 99) 的 LAIC 数据集, 本文提出的 MIRE 能较大幅度提升模型在 Micro/Macro-*P* 评价指标上的效果, 而对于标签类别较少 (标签数为 6) 的医疗公共健康问句数据集, 模型对于 Micro/Macro-*P* 评价指标的提升则并不明显, 造成上述问题的具体原因将通过对 LCRM 模块的消融实验展开进一步研究和说明. 综合 3 个数据集的实验结果, 可以明显发现 MIRE 相较于综合表现同样优秀的基于预训练语言模型的算法仍取得了显著优势, 进而可以充分地得出 MIRE 在多标签文本分类任务上具有较强的有效性和通用性, 同时相比与先进的对比算法在主要评价指标上仍具备较大优势.

3.3 消融实验

为了进一步验证本文提出模型的有效性及合理性, 本文在 3 个数据集上分别对各个模块进行了一系列消融实验, 其中包括: 验证文本-标签联合嵌入、LCRM 模块的有效性.

3.3.1 文本-标签联合嵌入消融实验

针对文本-标签联合嵌入方式有效性的验证, 如表 5 所示, 本文在 3 个数据集上对比了 BERT(cls) 和 BERT(pooled) 在进行多标签文本分类时不同的嵌入方式对结果的影响, 其中带*的为采用了文本-标签联合嵌入的方法, 每个评价指标中最佳效果由粗体标出.

表 5 文本-标签联合嵌入消融实验结果

数据集	评价指标	BERT(cls)	BERT(cls)*	BERT(pooled)	BERT(pooled)*
AAPD	*P*@1 (+)	0.823	0.849	0.847	**0.851**
	P@3 (+)	0.607	**0.616**	0.613	0.615
	P@5 (+)	0.409	0.415	**0.418**	0.417
	nDCG@1 (+)	0.823	0.849	0.847	**0.851**
	nDCG@3 (+)	0.797	**0.812**	0.809	**0.812**
	nDCG@5 (+)	0.831	0.846	0.848	**0.849**
	Micro-*P* (+)	0.739	0.748	0.779	**0.800**
	Micro-*R* (+)	**0.738**	0.720	0.685	0.670
	Micro-*F*1 (+)	0.729	**0.734**	0.729	0.729
	Macro-*P* (+)	0.600	0.640	0.673	**0.683**
	Macro-*R* (+)	0.568	**0.577**	0.551	0.509
	Macro-*F*1 (+)	0.574	**0.596**	0.582	0.558
LAIC	*P*@1 (+)	**0.622**	**0.622**	0.580	0.597
	P@3 (+)	0.402	**0.403**	0.379	0.384
	P@5 (+)	0.297	**0.298**	0.280	0.287
	nDCG@1 (+)	**0.622**	**0.622**	0.580	0.597
	nDCG@3 (+)	0.669	**0.671**	0.552	0.563
	nDCG@5 (+)	0.717	**0.718**	0.548	0.559
	Micro-*P* (+)	**0.599**	0.589	0.547	0.549
	Micro-*R* (+)	0.490	**0.502**	0.485	0.493
	Micro-*F*1 (+)	0.539	**0.542**	0.514	0.520
	Macro-*P* (+)	**0.499**	0.496	0.447	0.443
	Macro-*R* (+)	0.359	**0.371**	0.379	0.385
	Macro-*F*1 (+)	0.396	**0.406**	0.402	0.403
医疗公共健康问句	*P*@1 (+)	0.900	0.908	0.903	**0.913**
	P@3 (+)	0.427	0.422	0.425	**0.429**
	P@5 (+)	**0.263**	**0.263**	**0.263**	**0.263**
	nDCG@1 (+)	0.889	0.908	0.903	**0.913**
	nDCG@3 (+)	0.889	0.896	0.890	**0.898**
	nDCG@5 (+)	0.889	0.896	0.890	**0.898**
	Micro-*P* (+)	**0.878**	0.866	0.867	0.873
	Micro-*R* (+)	0.855	**0.880**	0.870	0.871
	Micro-*F*1 (+)	0.866	**0.873**	0.868	0.872
	Macro-*P* (+)	**0.709**	0.693	0.674	0.697
	Macro-*R* (+)	0.639	0.666	**0.675**	0.656
	Macro-*F*1 (+)	0.668	**0.677**	0.674	0.675

根据结果可以发现, 在 AAPD 数据集中的主要评价指标上, 未采用文本-标签联合嵌入的方法仅在 *P*@5 上有着最佳效果, 而其余主要评价指标中最佳效果分别为采用文本-标签联合嵌入的 BERT(cls)*和 BERT(pooled)*方法, 其中 BERT(cls)*在 *P*@*k* 和 *nDCG*@*k* 上均在*k* =3 时有着最佳效果, 在 Micro/Macro-*F*1 上有着最佳效果, 而

BERT(pooled)*在 *P@k* 上仅当 k =1 时有着最佳效果, 在 *nDCG@k* 上效果为最佳, 这说明采用文本-标签联合嵌入的方法能够更好地使预训练语言模型利用标签信息以及学习文本与标签、标签与标签之间的关系.

在 LAIC 数据集中的主要评价指标上, 采用文本-标签联合嵌入的方法均有着最佳效果, 同时可以注意到 BERT(pooled)*方法相比与 BERT(pooled) 方法在各项评价指标上均有提升. 此外, BERT(pooled) 方法无论是否采用文本-标签联合嵌入方式均在各项指标上劣于 BERT(cls) 方法, 其主要原因在于, LAIC 数据集中文本长度通常大于最大输入序列长度 512, 同时由于法律文书中存在大量噪音文本, 如此一来仅是单纯地将每个词的信息平均池化后进行多标签文本分类的效果则远不如 BERT(cls) 方法, 这也进一步说明了本文提出的 MIRE 中, 对于文本信息的提取方法对于提升分类效果更具有效性.

在医疗健康问句数据集中的主要评价指标上, 采用文本-标签联合嵌入的方法同样均有着最佳效果, 其中 BERT(cls)*在 Micro/Macro-*F*1 上均有着最佳效果, 而 BERT(pooled)*在 *P@k* 与 *nDCG@k* 上效果均为最佳. 根据表 5 的结果可以证明文本-标签联合嵌入方式能够通过将标签与文本嵌入同一空间中以有效提高多标签文本分类的综合性能.

3.3.2 LCRM 模块消融实验

针对 LCRM 模块有效性的验证, 如表 6 所示, 本文在 3 个数据集上分别对比了 MIRE 与 $\text{MIRE}_{(\text{-LCRM})}$ 的区别, 即验证添加 LCRM 模块是否有助于提升多标签文本分类任务的综合性能, 其中 MIRE 为添加了 LCRM 模块, $\text{MIRE}_{(\text{-LCRM})}$ 则表示仅有两级多标签文本分类模块, 每个评价指标中最佳效果由粗体标出.

表 6 LCRM 模块消融实验结果

数据集	评价指标	$\text{MIRE}_{(\text{-LCRM})}$	MIRE	相对提升率 (%)
AAPD	*P*@1 (+)	0.848	**0.862**	1.651
	P@3 (+)	0.619	**0.624**	0.808
	P@5 (+)	0.412	**0.424**	2.913
	nDCG@1 (+)	0.848	**0.862**	1.651
	nDCG@3 (+)	0.815	**0.823**	0.982
	nDCG@5 (+)	0.853	**0.860**	0.821
	Micro-*P* (+)	0.770	**0.771**	0.130
	Micro-*R* (+)	0.709	**0.720**	1.551
	Micro-*F*1 (+)	0.738	**0.744**	0.813
	Macro-*P* (+)	**0.659**	0.653	−0.910
	Macro-*R* (+)	0.571	**0.579**	1.401
	Macro-*F*1 (+)	0.594	**0.595**	0.168
LAIC	*P*@1	0.629	**0.636**	1.113
	P@3	0.405	**0.409**	0.988
	P@5	0.300	**0.301**	0.333
	nDCG@1	0.629	**0.636**	1.113
	nDCG@3	0.677	**0.683**	0.886
	nDCG@5	0.725	**0.729**	0.552
	Micro-*P* (+)	0.599	**0.608**	1.503
	Micro-*R* (+)	0.520	**0.520**	0.000
	Micro-*F*1 (+)	0.557	**0.560**	0.539
	Macro-*P* (+)	0.513	**0.527**	2.729
	Macro-*R* (+)	0.387	**0.388**	0.258
	Macro-*F*1 (+)	0.417	**0.422**	1.129
医疗健康问句	*P*@1 (+)	0.914	**0.923**	0.985
	P@3 (+)	0.430	**0.431**	0.233
	P@5 (+)	**0.263**	**0.263**	0.000
	nDCG@1 (+)	0.914	**0.923**	0.985
	nDCG@3 (+)	0.954	**0.958**	0.419
	nDCG@5 (+)	0.960	**0.964**	0.417

表 6 LCRM 模块消融实验结果 (续)

数据集	评价指标	$MIRE_{(-LCRM)}$	MIRE	相对提升率 (%)
医疗健康问句	Micro-*P* (+)	0.867	**0.869**	0.231
	Micro-*R* (+)	0.877	**0.891**	1.596
	Micro-*F*1 (+)	0.872	**0.880**	0.917
	Macro-*P* (+)	0.684	**0.686**	0.292
	Macro-*R* (+)	0.691	**0.696**	0.724
	Macro-*F*1 (+)	0.688	**0.691**	0.436

根据结果可以发现, 加入 LCRM 模块对于主要评价指标 Micro/Macro-*F*1、*nDCG@k* 、*P@k* 的得分均有提升, 我们能够得出 LCRM 模块可以通过学习标签之间的约束性关系以提升整体模型多标签文本分类综合性能的提升. 同时通过对 Micro/Macro-*P* 评价指标上的得分在不同数据集上的差异进行更深入的分析可以发现, 相较于 BERT(cls)、BERT(pooled) 的表现, LCRM 模块在标签数量较多 (标签数为 99) 的 LAIC 数据集上的提升非常显著, 而在标签数量较少 (标签数为 6) 的医疗公共健康问句数据集上的提升较为微弱.

造成上述问题的原因在于, LCRM 模块对于整体模型多标签文本分类综合性能的提升依托于学习标签之间的约束性关系, 当标签数量较少时, 标签之间的约束性关系并不显著, 同时基于预训练语言模型的算法本身所具备的语义知识也已经足够进行有效的多标签文本分类任务, 而当标签数量较多时, 标签之间相关性、互斥性等约束性关系表现得更为明显, 同时基于预训练语言模型的算法在面对标签数量较多的任务时仅利用本身的语义知识将无法支撑其进行有效的多标签文本分类.

3.4 算法在不同频率标签下的性能对比

为了充分探讨在不同频率标签下, MIRE 均能有效利用标签之间的约束性关系, 本文针对没有引入标签信息的算法 XML-CNN、BERT(cls)、BERT(pooled) 与 $MIRE_{(-LCRM)}$、MIRE 在 AAPD 数据集上进行对比实验. 通过对 AAPD 数据集的标签频率进行分析, 由图 5 可知, AAPD 数据集的标签频率服从长尾分布. 本文按照标签频率将所有标签划分为 4 组, 超低频组、低频组、中频组、高频组 (Group 1–Group 4), 并分别在 4 组频率标签集上考察 5 种算法的 Micro-*F*1 效果表现. 根据图 6 的实验结果, 可以得出以下结论.

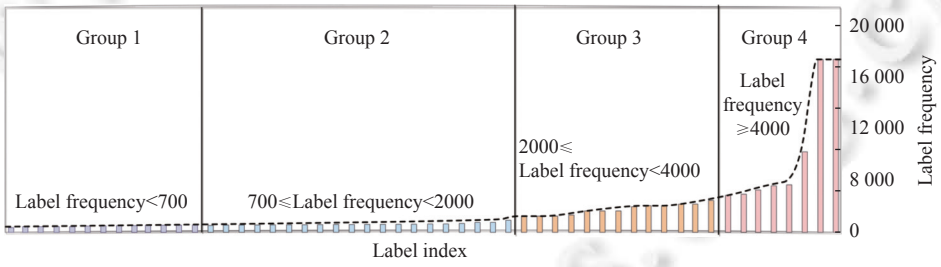


图 5 AAPD 数据集标签频率分组图

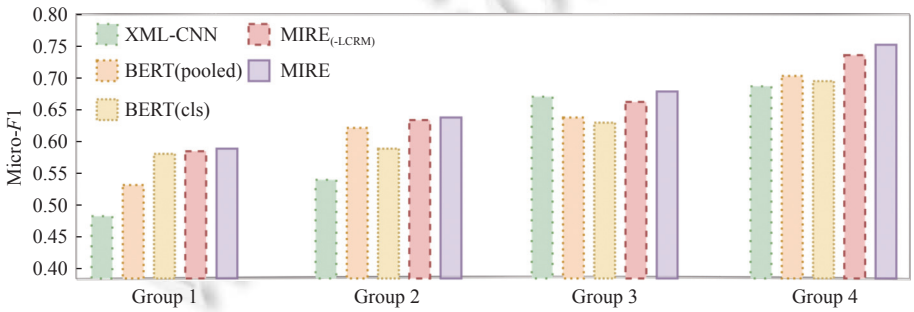


图 6 AAPD 数据集中 4 个标签频率分组下的 Micro-*F*1 值

1) $MIRE_{(-LCRM)}$、MIRE 在 4 个标签频率分组上的表现均优于其他没有引入标签信息的算法, 进而得以验证本文提出的文本-标签联合嵌入方式能够利用预训练语言模型来隐式地学习标签之间的关系来提升模型进行多标签文本分类的综合能力.

2) 相比于 $MIRE_{(-LCRM)}$, 添加了 LCRM 模块的 MIRE 在 4 个标签频率分组上的表现均更为优秀, 可以看出在通过预训练语言模型来隐式地学习标签之间关系的基础上进一步学习标签之间的约束性关系的方法能够更好地提升模型对于多标签文本分类的综合效果, 从而证明了 LCRM 模块的有效性.

3) 随着标签频率的不断降低, 所有算法的 Micro-$F1$ 值均随之产生了减益效果, 因此长尾问题仍然是困扰多标签文本分类任务的一重点问题, 亟待针对性地进一步解决.

4 总 结

本文针对不同粒度的文本信息以及标签之间的约束性关系提出了一种多粒度信息关系增强的多标签文本分类方法, 分别从文档和词两种粒度出发, 综合不同层次上的信息交互对文本-标签之间关系进行考察, 并充分利用了文本本身所携带的信息, 同时也对标签之间的约束性关系进行了学习, 通过增强对文档-标签、词-标签、标签-标签之间关系的利用, 提升了算法在多标签文本分类上的效果. 本文通过在 AAPD、LAIC、医疗公共健康问句 3 个数据集上进行实验, 验证了 MIRE 的有效性和通用性. 在消融实验部分中, 证明了本文提出的文本-标签联合嵌入方式和 LCRM 模块对于提升整体模型进行多标签文本分类综合性能的有效性, 并探究了算法在不同频率标签下的性能变化, 进一步验证了文本-标签联合嵌入方式和 LCRM 模块能提升模型在低频标签下的性能表现.

致谢 感谢中南大学高性能计算中心提供的计算资源.

References:

[1] Shelke S, Attar V. Rumor detection in social network based on user, content and lexical features. Multimedia Tools and Applications, 2022, 81(12): 17347–17368. [doi: 10.1007/s11042-022-12761-y]

[2] Tang DY, Qin B, Feng XC, Liu T. Effective LSTMs for target-dependent sentiment classification. arXiv:1512.01100, 2016.

[3] Guo L, Jin B, Yu RY, Yao CL, Sun CL, Huang DG. Multi-label classification methods for green computing and application for mobile medical recommendations. IEEE Access, 2016, 4: 3201–3209. [doi: 10.1109/ACCESS.2016.2578638]

[4] Papanikolaou Y, Dimitriadis D, Tsoumakas G, Laliotis M, Markantonatos N, Vlahavas I. Ensemble approaches for large-scale multi-label classification and question answering in biomedicine. In: Proc. of the 2014 CLEF (Working Notes). 2014. 1348–1360.

[5] Gopal S, Yang YM. Multilabel classification with meta-level features. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Geneva: Association for Computing Machinery, 2010. 315–322. [doi: 10.1145/1835449.1835503]

[6] Zhang DW, Yang PF, Xu YF. Research of Chinese comments sentiment classification based on Word2Vec and SVMperf. Computer Science, 2016, 43(6A): 418–421 (in Chinese with English abstract). [doi: 10.11896/j.issn.1002-137X.2016.6A.099]

[7] Boutell MR, Luo JB, Shen XP, Brown CM. Learning multi-label scene classification. Pattern Recognition, 2004, 37(9): 1757–1771. [doi: 10.1016/j.patcog.2004.03.009]

[8] Kim Y. Convolutional neural networks for sentence classification. arXiv:1408.5882, 2014.

[9] Liu PF, Qiu XP, Huang XJ. Recurrent neural network for text classification with multi-task learning. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York: AAAI, 2016. 2873–2879.

[10] Chen GB, Ye DH, Xing ZC, Chen JS, Cambria E. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In: Proc. of the 2017 Int'l Joint Conf. on Neural Networks (IJCNN). Anchorage: IEEE, 2017. 2377–2383. [doi: 10.1109/IJCNN.2017.7966144]

[11] Li BH, Xiang YX, Feng D, He ZC, Wu JJ, Dai TL, Li J. Short text classification model combining knowledge aware and dual attention. Ruan Jian Xue Bao/Journal of Software, 2022, 33(10): 3565–3581 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6630.htm [doi: 10.13328/j.cnki.jos.006630]

[12] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2019.

[13] Tsoumakas G, Katakis I. Multi-label classification: An overview. Int'l Journal of Data Warehousing and Mining, 2007, 3(3): 1–13. [doi:

10.4018/jdwm.2007070101]

[14] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. In: Proc. of the 20th European Conf. on Machine Learning and Knowledge Discovery in Databases. Bled: Springer, 2009. 254–269. [doi: 10.1007/978-3-642-04174-7_17]

[15] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 3104–3112.

[16] Nam J, Mencía EL, Kim HJ, Fürnkranz J. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5419–5429.

[17] Yang PC, Sun X, Li W, Ma SM, Wu W, Wang HF. SGM: Sequence generation model for multi-label classification. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 3915–3926.

[18] Wang GY, Li CY, Wang WL, Zhang YZ, Shen DH, Zhang XY, Henao R, Carin L. Joint embedding of words and labels for text classification. arXiv:1805.04174, 2018.

[19] Xiao L, Huang X, Chen BL, Jing LP. Label-specific document representation for multi-label text classification. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 466–475. [doi: 10.18653/v1/D19-1044]

[20] Xiao L, Chen BL, Huang X, Liu HF, Jing LP, Yu J. Multi-label text classification method based on label semantic information. Ruan Jian Xue Bao/Journal of Software, 2020, 31(4): 1079–1089 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5923.htm [doi: 10.13328/j.cnki.jos.005923]

[21] Nam J, Kim J, Mencía EL, Gurevych I, Fürnkranz J. Large-scale multi-label text classification—Revisiting neural networks. In: Proc. of the 2014 European Conf. on Machine Learning and Knowledge Discovery in Databases. Nancy: Springer, 2014. 437–452. [doi: 10.1007/978-3-662-44851-9_28]

[22] Rubinstein R. The cross-entropy method for combinatorial and continuous optimization. Methodology & Computing in Applied Probability, 1999, 1(2): 127–190. [doi: 10.1023/A:1010091220143]

[23] Bhatia K, Jain H, Kar P, Varma M, Jain P. Sparse local embeddings for extreme multi-label classification. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 730–738.

[24] Liu JZ, Chang WC, Wu YX, Yang YM. Deep learning for extreme multi-label text classification. In: Proc. of the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Shinjuku: Association for Computing Machinery, 2017. 115–124. [doi: 10.1145/3077136.3080834]

[25] Yang PC, Luo FL, Ma SM, Lin JY, Sun X. A deep reinforced sequence-to-set model for multi-label classification. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 5252–5258. [doi: 10.18653/v1/P19-1518]

[26] Tsai CP, Lee HY. Order-free learning alleviating exposure bias in multi-label classification. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(4): 6038–6045. [doi: 10.1609/aaai.v34i04.6066]

[27] Xiao L, Zhang XL, Jing LP, Huang C, Song MY. Does head label help for long-tailed multi-label text classification. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(16): 14103–14111. [doi: 10.1609/aaai.v35i16.17660]

[28] Wang R, Ridley R, Su X, Qu WG, Dai XY. A novel reasoning mechanism for multi-label text classification. Information Processing and Management, 2021, 58(2): 102441. [doi: 10.1016/j.ipm.2020.102441]

[29] Kingma DP, Ba J. ADAM: A method for stochastic optimization. arXiv:1412.6980, 2017.

附中文参考文献:

[6] 张冬雯, 杨鹏飞, 许云峰. 基于Word2Vec和SVMperf的中文评论情感分类研究. 计算机科学, 2016, 43(6A): 418–421. [doi: 10.11896/j.issn.1002-137X.2016.6A.099]

[11] 李博涵, 向宇轩, 封顶, 何志超, 吴佳骏, 戴天伦, 李静. 融合知识感知与双重注意力的短文本分类模型. 软件学报, 2022, 33(10): 3565–3581. http://www.jos.org.cn/1000-9825/6630.htm [doi: 10.13328/j.cnki.jos.006630]

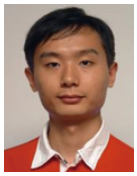
[20] 肖琳, 陈博理, 黄鑫, 刘华锋, 景丽萍, 于剑. 基于标签语义注意力的多标签文本分类. 软件学报, 2020, 31(4): 1079–1089. http://www.jos.org.cn/1000-9825/5923.htm [doi: 10.13328/j.cnki.jos.005923]

李芳芳(1983－), 女, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 自然语言处理, 文本挖掘.

苏朴真(1999－), 男, 硕士生, CCF 学生会员, 主要研究领域为机器学习, 自然语言处理.

段俊文(1990－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为自然语言处理, 信息抽取.

张师超(1962－), 男, 博士, 教授, 博士生导师, 主要研究领域为数据挖掘, 知识发现.



毛星亮(1979－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理, 文本挖掘.