多标签文本分类研究进展

郝 超,裘杭萍,孙 毅,张超然
陆军工程大学 指挥控制工程学院,南京 210007

摘 要:文本分类作为自然语言处理中一个基本任务,在20世纪50年代就已经对其算法进行了研究,现在单标签文本分类算法已经趋向成熟,但是对于多标签文本分类的研究还有很大的提升空间。介绍了多标签文本分类的基本概念以及基本流程,包括数据集获取、文本预处理、模型训练和预测结果。介绍了多标签文本分类的方法。这些方法主要分为两大类:传统机器学习方法和基于深度学习的方法。传统机器学习方法主要包括问题转换方法和算法自适应方法。基于深度学习的方法是利用各种神经网络模型来处理多标签文本分类问题,根据模型结构,将其分为基于CNN结构、基于RNN结构和基于Transformer结构的多标签文本分类方法。对多标签文本分类常用的数据集进行了梳理总结。对未来的发展趋势进行了分析与展望。

关键词:自然语言处理;多标签文本分类;深度学习
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.2101-0096

Research Progress of Multi-label Text Classification

HAO Chao, QIU Hangping, SUN Yi, ZHANG Chaoran
Command & Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China

Abstract: As a basic task in natural language processing, text classification has been studied in the 1950s. Now the single-label text classification algorithm has matured, but there is still a lot of improvement on multi-label text classification. Firstly, the basic concepts and basic processes of multi-label text classification are introduced, including data set acquisition, text preprocessing, model training and prediction results. Secondly, the methods of multi-label text classification are introduced. These methods are mainly divided into two categories: traditional machine learning methods and the methods based on deep learning. Traditional machine learning methods mainly include problem transformation methods and algorithm adaptation methods. The methods based on deep learning use various neural network models to handle multi-label text classification problems. According to the model structure, they are divided into multi-label text classification methods based on CNN structure, RNN structure and Transformer structure. The data sets commonly used in multi-label text classification are summarized. Finally, the future development trend is summarized and analyzed.

Key words: natural language processing; multi-label text classification; deep learning

文本作为信息的一种重要载体,通过各种社交APP、各大新闻门户网站等多种方式流入互联网。这些文本信息在主题上多种多样,在规模上也表现出很大的差异,如何对这些文本信息进行高效处理是一个具有重大研究意义的问题,推动了自动文本分类技术的快速发展。

文本分类是自然语言处理(Natural Language Processing,NLP)中重要且经典的问题[1]。在传统的文本分类问题中,每个样本只有一个类别标签,并且各个类别标签之间相互独立,分类粒度比较粗略,称为单标签文本分类。随着文本信息日益丰富,分类粒度细化程度越来越高,一个样本与多个类别的标签相关,同时类别标签之间存在一定的依赖关系,称为多标签文本分类[2]。比如一篇新闻可能被同时认为是与“体育”和“教育”相关的新闻。

多标签文本分类问题是多标签分类的重要分支之一,目前已经广泛应用于标签推荐[3]、信息检索[4]和情感分析[5]等领域。本文将多标签文本分类方法分为两大类:传统机器学习方法和基于深度学习的方法。传统机

基金项目:国家部委科技创新特区计划项目。
作者简介:郝超(1996—),男,硕士研究生,研究领域为自然语言处理、多标签文本分类,E-mail:2331192415@qq.com;裘杭萍(1965—),女,博士,教授,CCF会员,研究领域为系统工程;孙毅(1993—),男,博士研究生,研究领域为信息检索、自然语言处理;张超然(1994—),男,硕士研究生,研究领域为深度学习、机器阅读理解。
收稿日期:2021-01-06 **修回日期:**2021-03-09 **文章编号:**1002-8331(2021)10-0048-09

器学习方法包括问题转换方法和算法自适应方法。基于深度学习的方法是利用各种神经网络模型来处理多标签文本分类问题,根据网络的结构将其分为基于卷积神经网络(Convolutional Neural Network,CNN)结构、基于循环神经网络(Recurrent Neural Network,RNN)结构和基于Transformer结构的多标签文本分类方法。对该领域常用的数据集进行了梳理总结,最后对未来的发展趋势进行了分析与展望,可以为该领域研究提供一定的参考。

1　多标签文本分类

1.1　基本概念

多标签文本分类的主要任务是:将一个待分类的文本通过特定的分类器对该文本给定多个标签。可以用特定的数学符号来表示该任务,假定 $D=\{(x_i,y_i)|1\leqslant i\leqslant m\}$ 是训练集中的样本,利用设计的模型学习到一个映射 $f:X\rightarrow Y$,其中 $x_i\in X$ 是一个实例, $y_i\in Y$ 是实例 x_i 所对应的类别标签。该映射如图1所示。

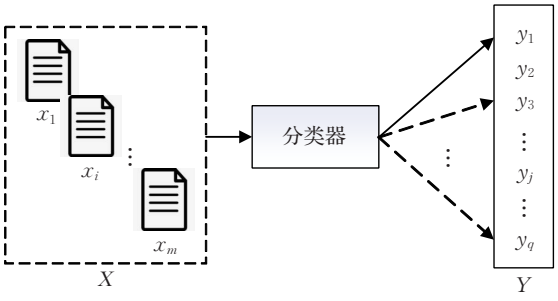

图1　多标签文本分类概念

图1实例空间 X 中包含 m 个实例,标签空间 Y 中包含 q 个类别标签,通过数据集训练得到分类器模型。测试过程中,每一个实例通过分类器模型得到相对应的标签,标签是一个或者多个,获得标签的过程就叫作多标签文本分类。

1.2　多标签文本分类流程

多标签文本分类的具体流程包括数据集获取、文本预处理、模型训练和预测结果,如图2所示。

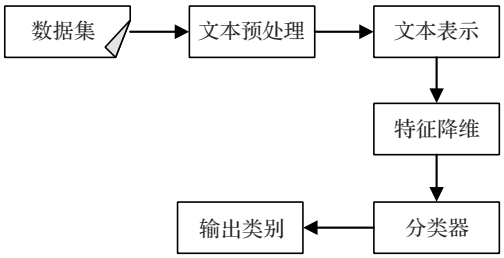

图2　多标签文本分类流程

(1)数据集

首先要找到需要的数据集。数据集一般分为训练集、测试集和验证集,文中第三部分列举了多标签文本分类领域常用的数据集。

(2)文本预处理

文本预处理是自然语言处理任务的重要环节,将文本转换为结构化的数据形式,以便计算机处理。文本预处理一般有固定的流程,包括分词、词干提取、词性还原等。

(3)文本表示

文本的信息是非结构化的,计算机无法直接处理这种非结构化的信息,因此在完成了预处理之后的文本要进行向量化表示:将输入的文本数据通过一定的方法转换为计算机能够识别的数字数据,良好的文本表示形式可以极大地提升算法效果。文本向量化主要分为两类方法:第一类是离散表示,主要方法有One-hot编码、词袋(Bag of Words,BOW)模型等;第二类方法是分布式表示,主要方法包括共现矩阵、Word2Vec[6]、Glove[7]等。Word2Vec和Glove是第一代预训练模型(Pre-trained Models,PTM),通常采用浅层模型来学习词嵌入;新一代PTM专注于学习上下文的词嵌入,如ELMo[8]、OpenAI、GPT[9]和BERT[10],学习更合理的词表征,包括了上下文信息[11]。

(4)特征降维

特征降维也称特征提取。通过文本向量化处理后得到的特征比较稀疏,维度较高。特征提取就是在保证文本语义表达完整的前提下,去除无用特征,保留有效特征,进行特征降维。常用的特征选择方式有词频-逆向文件频率[12](Term Frequency-Inverse Document Frequency,TF-IDF)、卡方检验、深度神经网络等。在预训练模型提出之后,大多数预训练模型采取Transformer结构作为特征提取模块。

(5)分类器和输出类别

将预处理之后的文本(训练集)送入特定的分类器(模型)中进行训练,得到分类器模型。通过验证集和测试集输出类别的预测,利用F1值等相关指标来评判模型的优劣。

2　多标签文本分类方法

近年来,多标签文本分类得到了快速的发展,涌现出大量多标签文本分类方法,这些方法可以分为两大类:传统机器学习方法和基于深度学习方法。具体分类如图3所示。

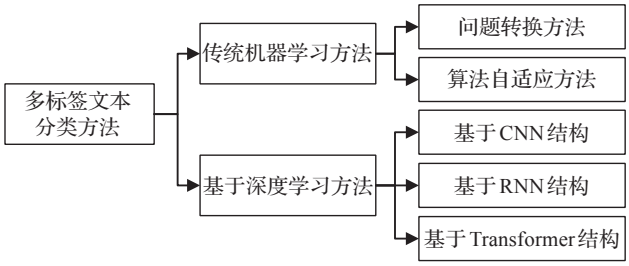

图3　多标签文本分类方法分类

2.1 传统机器学习方法

传统机器学习方法根据解决策略角度,可以分为问题转换方法和算法自适应方法[13]。

2.1.1 问题转换方法

问题转换方法是最简单的方法,将多标签文本分类任务转换为其他已经成熟的方案,比如将多标签文本分类问题转换为多个二分类问题。Boutell等人[14]提出的二元相关(Binary Relevance,BR)方法就是典型的问题转换方法,它直接忽略标签之间的相关性,并为每个标签建立一个单独的分类器,以此来达到多标签文本分类的效果,但该模型的性能较低。为了捕获标签之间的依赖问题,Tsoumakas等人[15]提出标签幂集分解(Label Powerset,LP)方法,该方法通过为每个标签组合使用唯一的二进制分类器,将任务转变为标签组合的多分类问题。Read等人[16]对BR方法进行改进,提出了分类器链(Classifier Chain,CC)方法,将任务转换为二进制分类问题链,其中后续的二进制分类器基于先前的预测,如果前面的标签预测错误就会对后面的标签产生影响。Tsoumakas等人[17]提出了名为Random k-labelsets的算法,主要是将多标签学习任务转换为多类分类任务。

2.1.2 算法自适应方法

算法自适应方法通过采用合适的算法来直接处理多标签数据以解决多标签学习问题。代表性的算法包括Clare等人[18]提出的ML-DT(Multi-Label Decision Tree)方法,它通过构造决策树来执行分类。Elisseeff等人[19]提出排名支持向量机(Ranking Support Vector Machine,Rank-SVM),采用类似于学习系统的支持向量机(Support Vector Machine,SVM)来处理多标签问题,其中优化了一组线性分类器来最小化经验ranking loss,并且能够用核技巧处理非线性情况。Zhang等人[20]提出了一个多标签 K 最近邻(Multi-Label K-Nearest-Neighbor,ML-KNN)方法,该方法是基于 KNN 算法改进的,通过 K 近邻来处理多标签数据,其中最大后验(Maximum a Posteriori,MAP)规则用于通过推理包含在邻居中的标签信息来进行预测。

综上所述,问题转换方法的关键是使数据适合算法,而算法自适应方法的关键是使算法适应数据。

2.2 基于深度学习方法

由于深度学习的快速发展,深度学习模型在计算机视觉(Computer Vision,CV)和语音识别(Speech Recognition,SR)领域取得了很好的效果。在自然语言处理领域中,许多深度学习方法也得到了广泛的应用。深度学习在文本分类中取得了很好的效果,比较有代表性的有Kim[21]提出的TextCNN模型。该方法首次将CNN结构用于文本分类,利用CNN来进行句子级别的分类,基于Word2Vec进行了一系列实验,但是该模型无法避免使用CNN中固定窗口的缺点,因此无法建模更长的序列信息。Lai等人[22]提出了TextRCNN模型,该方法主要针对传统分类方法存在忽略上下文的问题以及针对CNN卷积窗口设置的问题,结合RNN和CNN的优点提出了RCNN模型。当时的网络都是针对单一任务进行训练,缺少标注数据,因此Liu等人[23]提出了TextRNN模型,将多个任务联合起来训练,以此来对网络进行改善。Yang等人[24]将Attention机制加入到TextRNN中,提出一个分层注意力网络模型HAN,采用“词-句子-文章”的层次化结构来表示一篇文本,具有很好的可解释性。随着Transformer和BERT的提出,Sun等人[25]将BERT应用到文本分类中,介绍了一些调参以及改进的方法,进一步挖掘BERT在文本分类中的应用。

在多标签文本分类领域,深度神经网络也得到了广泛的应用,并且取得了不错的效果。Zhang等人[26]早在2006年就提出了名为BP-MLL的算法,这是首次将神经网络应用到多标签文本分类上。该方法源于BP算法,通过使用一种新的误差函数来捕获多标签学习的特征,即属于一个实例的标签要比不属于该实例的标签排名高。Nam等人[27]改进了BP-MLL算法,用交叉熵损失函数代替ranking loss,并且使用了Dropout、AdaGrad和ReLUs。上面的两种模型只是用了简单的神经网络,无法说明文本信息的完整性并且不会保留单词顺序。针对这些缺点,后面又陆续提出了大量的基于CNN、RNN和Transformer的多标签文本分类模型。

下面按照网络结构的不同,将基于深度学习的多标签文本分类算法分为三大类,包括基于CNN、基于RNN和基于Transformer的多标签文本分类。

2.2.1 基于CNN的多标签文本分类

CNN首先是应用在图像领域,特别是在计算机视觉领域取得了不错的效果,比如图像分类、目标检测和图像分割等。在CNN[28]中,主要包括卷积层、池化层和全连接层。用来处理文本分类任务的典型CNN结构如图4所示,其在图像领域取得了巨大成功。在TextCNN模型提出后,越来越多的基于CNN的分类模型被提出。

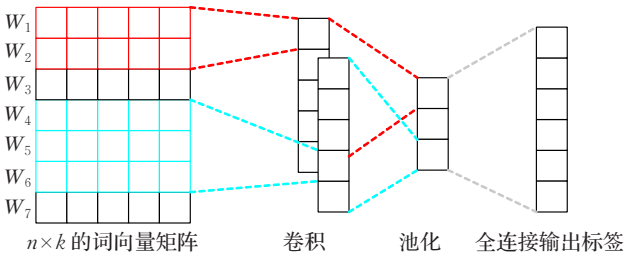

图4 CNN结构

Berger[29]提出了一种分别将TextCNN和门控循环单元(Gate Recurrent Unit,GRU)与Word2Vec词向量使用的方法来解决大规模多标签文本分类问题,通过保留单词顺序和使用语义词向量来保留较大语义的词汇,最后根据一个阈值 α 来确定样本是否属于某个类别。Baker等人[30]设计了一种基于CNN结构的标签共现的多标签文本分类方法,该方法主要是通过初始化神经网络模型

的最终隐藏层来利用标签共现关系。Kurata等人[31]提出了一种利用标签共现来改进基于CNN结构的多标签分类方法,主要的改进在于提出了一种新的网络初始化方法来利用标签共现信息。Liu等人[32]基于TextCNN结构进行了改进,提出了XML-CNN模型,该模型不同于TextCNN的方面在于池化操作时使用了动态池化,改进了损失函数,采用了二元交叉熵损失函数,并在池化层和输出层之间加了一个隐藏层,能够将高维标签映射到低维,以此来减少计算量。Shimura等人[33]提出了一种针对短文本多标签文本的分层卷积神经网络结构HFT-CNN,该方法的主要思想是利用预训练加微调的思想,并且利用类别之间的层次关系解决短文本数据稀疏问题。Yang等人[34]提出了一种针对数据不平衡的多标签文本分类的孪生CNN网络HSCNN,主要用孪生网络的结构来处理少样本的问题,利用混合机制来解决极端不平衡多标签文本分类问题,针对头标签采用单一的网络结构,针对尾标签采用少样本孪生网络方法。

基于CNN的多标签文本分类方法都是对CNN结构改进,以此来适应多标签文本分类。虽然这种方法比较简单,并且也不需要花费巨大的计算代价,但是利用CNN的池化操作时,会造成语义信息的丢失,并且当文本过长时,CNN不利于捕获前后文的关系而造成语义的偏差。

2.2.2　基于RNN的多标签文本分类

CNN无法处理以序列形式出现的输入,然而在自然语言处理中,大多数输入都是序列数据,比如一个句子就是一个序列数据。为了处理这些序列输入的要求,RNN也得到了快速的发展,在文本分类领域也得到了广泛的应用。RNN类似于所有的深层架构,网络越深,梯度消失和梯度爆炸问题也就越明显,无法掌握长时间跨度非线性关系,因此在采用RNN的时候往往会采用改进的RNN结构,包括长短时记忆网络(Long Short-Term Memory,LSTM)[35]和GRU[36]来解决长期依赖问题。这些深度神经网络处理的都是定长序列的问题,即输入和输出的大小是固定不变的。为了解决这个问题,Sutskever等人[37]提出了序列到序列(Sequence to Sequence,Seq2Seq)的结构,其网络结构如图5所示。采用了两个RNN组合的方式构成网络,主要思想是用多层的LSTM来进行编码,然后用另一个深层的LSTM来解码。Seq2Seq模型的提出首先是为了解决机器翻译的问题,后面也迁移到了各个自然语言处理任务中,包括多标签文本分类。

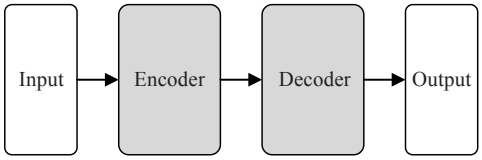

图5　Seq2Seq结构

Nam等人[38]利用RNN来代替分类器链,并使用基于RNN的Seq2Seq去建模,这种方法借助RNN依次产生标签序列来捕获标签之间的相关性。这是首次将Seq2Seq模型应用在多标签文本分类上,在这之后,有更多的Seq2Seq模型被提出并用于处理多标签文本分类。Chen等人[39]提出CNN-RNN模型,该模型将CNN和RNN进行融合,先将词向量送入到CNN中得到文本特征序列,然后将该特征输入到RNN中得到相应的预测标签。但是该模型受训练集大小影响较大,如果训练集过小,可能会产生过拟合。Yang等人[40]提出了引入注意力机制的SGM模型,也是一种Seq2Seq结构的模型,该模型将多标签分类任务视为序列生成问题,以此来考虑标签之间的相关性,也是首次将序列生成的思想应用到多标签文本分类中。编码部分采用Bi-LSTM来获取单词的序列信息,并且提出了一种具有注意力(Attention)机制的解码器结构的序列生成模型,该解码器在预测的时候能够自动选择最有信息量的单词。该模型利用生成的思想考虑标签之间的相关性,这会带来误差的累积。针对这一缺点,Yang等人[41]针对SGM模型进行了改进,主要是在SGM的基础上加了一个Set Decoder,利用Set的无序性,降低错误标签带来的影响。Qin等人[42]沿用了序列生成的思想,提出了自适应的RNN序列模型,提供一个新的训练目标,以便RNN模型能够发现最佳标签顺序。

注意力机制首先在图像领域取得成功之后,在多标签文本分类领域,也有越来越多的模型引入了Attention机制。Lin等人[43]提出多级扩展卷积,是通过在原始编码器LSTM生成表示法的基础上,应用多层卷积神经网络,通过捕获单词之间的局部相关性和长期依赖性来生成语义单元表示,进而增强Seq2Seq的效果,并且将高层的Attention和词级别的Attention做了整合,提出混合注意力(Hybrid Attention)来兼顾各个级别表示的信息。该模型有来自LSTM编码器的注释和来自MDC的语义单元表示,解码器部分首先关注的是来自MDC的语义单元表示,然后关注的是LSTM编码器的源注释。You等人[44]提出了基于标签树的Attention-XML模型,该模型使用Bi-LSTM来捕获单词之间的长距离依赖关系,以及使用多标签注意来捕获文本中与每个标签最相关的部分,针对长尾标签,提出了概率标签树(Probability Label Tree,PLT),能够高效处理上百万级别的标签。Yang等人[45]基于"并行编码,串行解码"策略,提出一种新的序列到序列模型,该模型将CNN和并行自注意力机制结合作为编码器,从源文本中提取局部邻域信息和全局交互信息,设计了一个分层解码器来解码并生成标签序列。

基于RNN的多标签文本分类方法大多都是采用Seq2Seq结构来实现,利用序列生成来考虑标签间的关系,后一个标签往往是依赖于前一个标签的,因此错误标签带来的影响往往就会叠加,虽然有一些方法提出了

改进,但还是存在着缺陷。并且利用这种方法虽然提升了结果,但是能否很好地学习到标签之间的相关性还有待商榷。

2.2.3　基于Transformer的多标签文本分类

Google提出了经典的网络结构Transformer[46],具体结构如图6。该结构只采用了Attention机制,不像传统的编码-解码的模型需要结合RNN或者CNN来使用。Transformer的提出给自然语言处理领域带来了极大的影响,之后的预训练模型GPT-2和BERT都是基于Transformer结构提出的,预训练模型的提出在各项自然语言处理任务都取得了很好的效果。BERT的提出可以说是自然语言处理领域的里程碑,其证明了一个非常深的模型可以显著提高自然语言处理任务的准确率,而这个模型可以从无标记数据集中预训练得到。

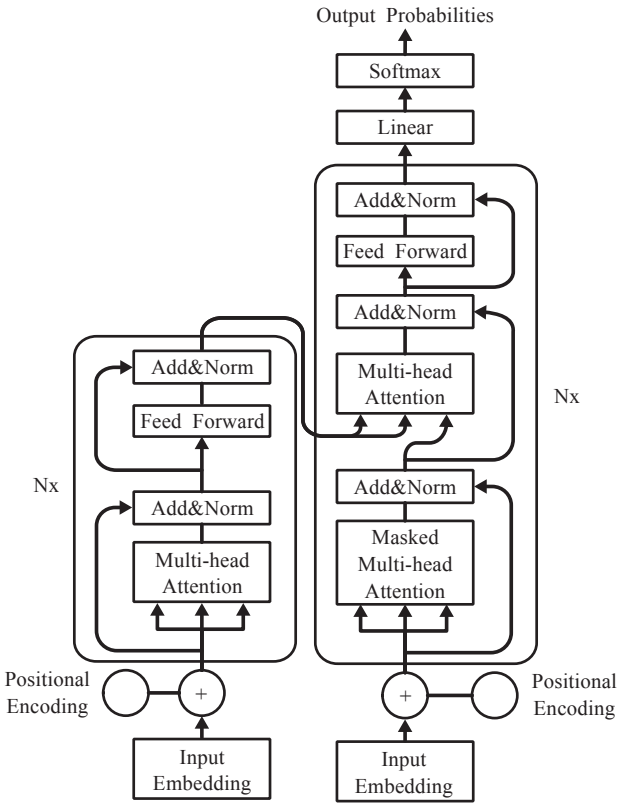

图6　Transformer结构

在多标签分类领域,也有很多采用Transformer的模型被提出。Yarullin等人[47]首次尝试BERT并探索其在多标签设置和分层文本分类中的应用,提出应用在多标签文本分类领域的序列生成BERT模型。Chang等人[48]提出X-Transformer模型,该模型是由三部分组成,包括语义标签序列组件、深度神经匹配组件和整体排名组件。语义标签序列组件通过标签聚类将棘手的极端多标签文本分类(Extreme Multi-label Text Classification,XMC)问题分解为一组输出空间较小的可行子问题,从而减轻标签稀疏性问题;深度神经匹配组件针对语义标签序列引起的每个XMC子问题微调Transformer模型,从而使得输入文本到标签簇集有更好的映射;最后,对整体排名组件进行有条件的训练,包括实例-群集分配和来自Transformer的神经嵌入,并用于组合从各种语义标签序列引起的子问题中得出的分数,以进一步提高性能。Gong等人[49]提出HG-Transformer的深度学习模型,该模型首先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Transformer结构以充分捕获文本的特征,最后利用标签的层次关系来生成标签的表示形式,并基于标签的语义距离设计加权损失函数。

基于Transformer结构的多标签文本分类模型的效果往往会优于基于CNN和基于RNN结构的模型,但是基于Transformer结构的模型比起前两种结构来说,参数量往往是巨大的,并且网络结构比较复杂,在实际场景中难以应用。

传统机器学习方法包括问题转换方法和算法自适应方法,虽然相对基于深度学习方法来说比较简单,但是在预测效果上往往不能达到很好的效果。除此之外,传统的机器学习在特征提取的时候往往需要人工提取,这会加大人工的花费,并且人工提取的特征并不能得到保障,因此在此过程中会出现很多差错,也会直接影响算法和模型的效果。随着深度学习在自然语言处理领域广泛应用,在单标签文本分类中已经取得不错的效果,目前也已经应用在多标签文本分类中,表1列举了部分基于深度学习的方法。深度学习的方法可以自动提取特征,大大减少了花费,使得算法的鲁棒性更强,不过对于设备和硬件要求以及设备计算能力要求也大大提升,并且在数据规模上要求更大;深度学习在可解释性上不如机器学习,它能够给出一个结果,但是中间的过程相当于一个黑盒子;深度学习的算法虽然大大提高了多标签文本分类的效果,但还是有很大的提高空间。

3　数据集

多标签文本分类虽然已经取得了快速的发展,但是在这方面的公开数据集并不是很多。本文收集了一些多在标签文本分类领域中常用的数据集,根据标签数量的多少可以将其分为小型数据集(标签数0~10 000)、中型数据集(标签数10 000~100 000)和大型数据集(标签数超过100 000)。本文从标签数、文本的数量等方面进行了统计,具体信息如表2所示。

对数据集的详细说明如下:

(1)Ren-CECps1.0[50]:该数据集是由Quan等人提供的,是一个多标签的中文情感语料库,它包含了37 678个中文博客的句子和11种情感标签,其中每句话被赋予一种或多种情感。

(2)Reuters-21578(https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection):该数据集是由路透社的新闻组成的,是路透社金融新闻服

表1　模型简介

模型	文献	简单介绍	模型结构	年份
TextCNN	[21]	利用基本CNN结构来进行分类,该模型将词向量组成的句子矩阵作为CNN的输入,利用CNN卷积器来提取特征	CNN	2014
改进BP-MLL	[27]	针对BP-MLL算法的不足,提出了改进,用交叉熵损失函数代替了ranking loss	简单神经网络	2014
XML-CNN	[32]	XML-CNN模型是基于TextCNN结构改进的,该模型不同于TextCNN的方面在于池化操作时使用了动态池化,改进了损失函数,采用了二元交叉熵损失函数	CNN	2017
HSCNN	[34]	主要用孪生网络的结构来处理少样本的问题,利用了混合机制来解决极端不平衡多标签文本分类问题,针对头标签采用单一的网络结构,针对尾标签采用少样本孪生网络方法	CNN 孪生网络	2020
CNN-RNN	[39]	该模型将CNN和RNN进行融合,先将词向量送入到CNN中得到文本特征序列,然后将该特征输入到RNN中得到相应的预测标签	Seq2Seq	2017
SGM	[40]	把多标签分类问题变为一个序列生成问题,提出一种新的解码器用于序列生成模型	Seq2Seq Attention	2018
MDC	[43]	基于LSTM的Seq2Seq,使用附加的多级展开卷积组件提取高级语义信息,并使用相应的混合注意力	Seq2Seq Attention	2018
BERT+SGM	[47]	提出了序列生成BERT模型(BERT+SGM)和一个混合模型,是vanilla BERT和BERT+SGM模型的集合	Transformer	2019
X-Transformer	[48]	该模型包含了三部分,包括语义标签序列组件(SLI)、深度神经匹配组件和整体排名组件	Transformer	2020
HG-Transformer	[49]	该模型将文本建模为一个图像结构,在单词、句子和图形级别引入了具有多头注意机制的多层Transformer结构来捕获特征,利用标签的层次关系来生成标签的表示形式	Transformer	2020

表2　数据集相关信息

数据集	样本总数	标签数	样本平均单词数	样本平均标签数
Ren-CECps1.0	37 687	11	24.71	2.36
Reuters-21578	10 788	90	—	—
AAPD	55 840	54	163.42	2.41
RCV1-V2	804 414	103	123.94	3.24
EUR-Lex	19 314	3 956	1 239.49	5.30
AmazonCat-13K	1 493 021	13 330	448.57	5.04
Amazon-670K	643 474	670 091	244.27	5.45
Amazon-3M	2 460 406	2 812 281	104.13	36.04

务进行分类的常用数据集,它包含了7 769个训练文本和3 019个测试文本,包含多个标签和单个标签。

(3)AAPD[40]:该数据集是由Yang等人提供的,是从网络上收集了55 840篇论文的摘要和相应学科类别,一篇学术论文属于一个或者多个学科,总共由54个学科组成,目的是根据给定的摘要来预测学术论文相对应的学科。

(4)RCV1-V2[51]:该数据集是由Lewis等人提供的,是由路透社新闻专栏报道组成,共有804 414篇新闻,每篇新闻故事分配有多个主题,共有103个主题。

(5)EUR-Lex[52]:该数据集是由Mencía等人提供的,是由欧盟法律组成的,里面包含了许多不同类型的文件,包括条约、立法、判例法和立法提案,共有19 314个文档,3 956个分类。

(6)AmazonCat-13K[53]:该数据集来自于亚马逊,其中包括评论(评分、文字、帮助性投票),产品元数据(描述、类别信息、价格、品牌和图像特征)和链接(可以查看/购买的图表),在做多标签文本分类时主要考虑的是类别信息。

(7)Amazon-670K[53]:该数据集数据的来源是亚马逊商品的评论、产品的数据,和AmazonCat-13K的数据有类似之处,只是规模和商品不一样。

(8)Amazon-3M[53]:该数据集的数据也来源于亚马逊,包含的是产品的信息、链接以及产品的评论。

由上面的分析可知,在公开的多标签文本分类数据集中,中文的数据集很少,以上的8个数据集中只有Ren-CECps1.0数据集是中文,其他的都是英文。大部分都是来自于亚马逊网站,都是商品的评论,因此适合用来做短文本分类。

4　多标签文本分类性能评价

4.1　评价指标

在多标签文本分类中,常用的评价指标通常包括汉明损失(Hamming Loss,HL)、Micro-F1值。

(1)汉明损失

Schapire等人[54]在1999年就提出了汉明损失,简单来说就是衡量被错分的标签的比例大小,正确的标签没有被预测正确以及错误标签被预测的标签占比,就是两个标签集合的差别占比,汉明损失的值越小,预测结果就越好。计算公式如下:

$$HL=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{XOR(x_i,y_i)}{|L|} \tag{1}$$

其中,$|D|$是样本的数量,$|L|$是标签的总数,x_i表示标签,y_i表示真实标签,XOR是异或运算。

(2)Micro-precision、Micro-recall和Micro-F1

对单标签文本分类而言,精准率(Precision)是针对预测结果而言的,表示预测为正的样本中有多少是真正的样本,一种是把正类预测为正类(TP),另一种就是把负类预测为正类(FP),它反映了模型的查准率。召回率(Recall)是针对样本来说,表示样本中的正样本有多少

被预测正确了,一种是把原来的正样本预测为正类(TP),另一种就是把原来的正样本预测为负类(FN),它反映了模型的查全率。具体可以见表3。

表3　评价指标F1值

混淆矩阵		预测值	
		Positive	Negative
实际值	Positive	TP	FN
	Negative	FP	TN

多标签文本分类将文本赋予多个标签,标签数量也不是固定的,通常使用Micro-precision和Micro-recall,考虑所有标签的整体精确率和召回率,在理想情况下是两者都越高越好,但实际情况往往会产生矛盾,因此在多标签文本分类领域,采用Micro-F1来评价。Micro-F1是Micro-precision和Micro-recall的调和平均,其计算公式[55]如下所示(其中 L 代表类别标签总数):

$$Micro\text{-}precision=\frac{\sum_{j=1}^{L}TP_j}{\sum_{j=1}^{L}(TP_j+FP_j)} \tag{2}$$

$$Micro\text{-}recall=\frac{\sum_{j=1}^{L}TP_j}{\sum_{j=1}^{L}(TP_j+FN_j)} \tag{3}$$

$$Micro\text{-}F1=\frac{\sum_{j=1}^{L}2TP_j}{\sum_{j=1}^{L}(2TP_j+FP_j+FN_j)} \tag{4}$$

4.2　结果分析

早期的多标签文本分类方法原理是基于传统机器学习方法来实现的,实现过程相对来说是比较简单的,但是效果还是不够理想。深度学习的发展,也大大促进了多标签文本分类的发展。表4对相关多标签文本分类模型在AAPD、RCV1-V2、EUR-Lex等数据集上的结果进行了总结。

表4　模型结果对比

模型	文献	数据集	Micro-F1	年份
TextCNN	[21]	RCV1-V2	0.829	2014
		AAPD	0.674	
		Ren-CECps1.0	0.565	
HAN	[24]	AAPD	0.708	2016
改进BP-MLL	[27]	RCV1-V2	0.784	2014
		EUR-Lex	0.575	
SGM	[40]	RCV1-V2	0.815	2018
		AAPD	0.710	
MDC	[43]	RCV1-V2	0.882	2018
		Ren-CECps1.0	0.590	
BERT+SGM	[47]	RCV1-V2	0.846	2019
		AAPD	0.718	

模型在AAPD、RCV1-V2、EUR-Lex等数据集上的结果显示,Micro-F1值逐渐提升,在RCV1-V2数据集上Micro-F1值从0.784提升到0.893,在AAPD数据集上Micro-F1值从0.674提升到0.725,提升效果明显。但还有很大的上升空间,特别是在预训练模型提出后,在各项任务上都取得了不错的效果,比如BERT的提出在11项NLP任务中都取得了很好的效果。

5　总结与展望

文本分类作为有效的信息检索和挖掘技术在关于文本管理方面发挥着重大的作用。虽然在单标签文本分类领域已经取得了不错的效果,但还是无法使模型像人一样从语义层面理解文本信息。多标签文本分类相较于单标签文本分类来说更加复杂,还存在着很多的挑战,主要体现在以下几点:

(1)特定领域的数据集缺失问题。目前公开的多标签文本分类领域的数据集,大部分是针对新闻领域的,对于特定领域的数据集非常匮乏,比如医疗领域、金融领域和法律领域。因此,需要构建特定领域的多标签文本分类数据集。

(2)极端多标签文本分类问题。极端多标签文本分类[48]目的是学习一个分类器,该分类器能够从大量标签中自动选择最相关的标签来对数据进行归类[56]。极端多标签文本分类的难点在于标签集的数目非常多,包含数十万、甚至成百上千万的标签。目前多标签文本分类模型的内存占用、模型大小都随着标签空间的变大而线性变大,在面对极端多的标签时,无法成功部署甚至训练。因此,如何设计出一个高效的模型来解决极端多标签文本分类问题是未来亟待解决的一个难点。

(3)标签间的相关性研究问题。多标签文本分类的标签之间是存在内在联系的,比如属于“人工智能”的文本往往跟“深度学习”是相关联的。传统的一些方法在处理多标签文本分类问题上,往往没有考虑标签之间的相关性,这也严重影响了模型的效率。后面虽然提出了一些方法来研究标签之间的相关性,比如Baker等人[30]提出了一种分层的多标签文本分类方法来得到标签间的共现关系,但只是考虑了标签之间浅层次的关系,忽略了标签之间深层次的关系。因此,如何高效捕捉标签间的关系也是多标签文本分类任务未来的一大研究重点。

(4)数据集标签长尾问题。对于多标签文本分类领域存在的数据集,都是由文本集和标签集构成的,对于标签集来说就会有分布不均衡的问题存在,部分标签与很多文本样本相关联,而还有的一些标签就非常少,甚至说没有与文本样本相关联,可以理解为标签“长尾”的问题[57]。用不平衡的数据训练出来的模型会导致样本少的种类预测性能很差,甚至无法预测。因此,如何解决标签长尾问题也是多标签文本分类领域一个重要的研究问题。

参考文献:

[1] ALI T,ASGHAR S.Multi-label scientific document classification[J].Journal of Internet Technology,2018,19(6):1707-1716.

[2] 刘心惠.基于改进seq2seq模型的多标签文本分类研究[D].辽宁大连:大连海事大学,2020.

[3] FÜRNKRANZ J,HÜLLERMEIER E,MENCÍA E L,et al.Multilabel classification via calibrated label ranking[J].Machine Learning,2008,73(2):133-153.

[4] GOPAL S,YANG Y.Multilabel classification with meta-level features[C]//33rd International ACM SIGIR Conference on Research and Development in Information Retrieval,2010:315-322.

[5] CAMBRIA E,OLSHER D,RAJAGOPAL D.SenticNet 3:a common and common-sense knowledge base for cognition-driven sentiment analysis[C]//28th AAAI Conference on Artificial Intelligence,2014:1515-1521.

[6] MIKOLOV T,CHEN K,CORRADO G,et al.Efficient estimation of word representations in vector space[J].arXiv:1301.3781,2013.

[7] PENNINGTON J,SOCHER R,MANNING C D.Glove:global vectors for word representation[C]//2014 Conference on Empirical Methods in Natural Language Processing,2014:1532-1543.

[8] PETERS M E,NEUMANN M,IYYER M,et al.Deep contextualized word representations[C]//2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2018:2227-2237.

[9] RADFORD A,NARASIMHAN K,SALIMANS T,et al.Improving language understanding by generative pre-training[EB/OL].(2018)[2020-11-30].https://s3-us-west-2.amazonaws.com/openaiassets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf.

[10] DEVLIN J,CHANG M W,LEE K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language,2019:4171-4186.

[11] QIU X P,SUN T X,XU Y G,et al.Pre-trained models for natural language processing:a survey[J].arXiv:2003.08271,2020.

[12] GHOSH S,DESARKAR M S.Class specific TF-IDF boosting for short-text classification:application to short-texts generated during disasters[C]//Companion of the the Web Conference 2018,Lyon,2018:1629-1637.

[13] ZHANG M L,ZHOU Z H.A review on multi-label learning algorithms[J].IEEE Transactions on Knowledge and Data Engineering,2013,26(8):1819-1837.

[14] BOUTELL M R,LUO J,SHEN X,et al.Learning multi-label scene classification[J].Pattern Recognition,2004,37(9):1757-1771.

[15] TSOUMAKAS G,KATAKIS I.Multi-label classification:an overview[J].International Journal of Data Warehousing and Mining,2007,3(3):1-13.

[16] READ J,PFAHRINGER B,HOLMES G,et al.Classifier chains for multi-label classification[J].Machine Learning,2011,85(3):333.

[17] TSOUMAKAS G,VLAHAVAS I.Random k-labelsets:an ensemble method for multilabel classification[C]//European Conference on Machine Learning.Berlin,Heidelberg:Springer,2007:406-417.

[18] CLARE A,KING R D.Knowledge discovery in multi-label phenotype data[C]//European Conference on Principles of Data Mining and Knowledge Discovery.Berlin,Heidelberg:Springer,2001:42-53.

[19] ELISSEEFF A,WESTON J.A kernel method for multi-labelled classification[C]//Advances in Neural Information Processing Systems,2002:681-687.

[20] ZHANG M L,ZHOU Z H.ML-KNN:a lazy learning approach to multi-label learning[J].Pattern Recognition,2007,40(7):2038-2048.

[21] KIM Y.Convolutional neural networks for sentence classification[J].arXiv:1408.5882,2014.

[22] LAI S,XU L,LIU K,et al.Recurrent convolutional neural networks for text classification[C]//29th AAAI Conference on Artificial Intelligence,2015.

[23] LIU P,QIU X,HUANG X.Recurrent neural network for text classification with multi-task learning[J].arXiv:1605.05101,2016.

[24] YANG Z,YANG D,DYER C,et al.Hierarchical attention networks for document classification[C]//2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2016:1480-1489.

[25] SUN C,QIU X,XU Y,et al.How to fine-tune bert for text classification?[C]//China National Conference on Chinese Computational Linguistics.Cham:Springer,2019:194-206.

[26] ZHANG M L,ZHOU Z H.Multilabel neural networks with applications to functional genomics and text categorization[J].IEEE Transactions on Knowledge and Data Engineering,2006,18(10):1338-1351.

[27] NAM J,KIM J,MENCÍA E L,et al.Large-scale multi-label text classification—revisiting neural networks[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases.Berlin,Heidelberg:Springer,2014:437-452.

[28] ZAGORUYKO S,KOMODAKIS N.Learning to compare image patches via convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern

Recognition,2015:4353-4361.
[29] BERGER M J.Large scale multi-label text classification with semantic word vectors[R].Stanford University,2015.
[30] BAKER S,KORHONEN A L.Initializing neural networks for hierarchical multi-label text classification[C]//BioNLP 2017,Association for Computational Linguistics,2017:307-315.
[31] KURATA G,XIANG B,ZHOU B.Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence[C]//2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2016:521-526.
[32] LIU J,CHANG W C,WU Y,et al.Deep learning for extreme multi-label text classification[C]//40th International ACM SIGIR Conference on Research and Development in Information Retrieval,2017:115-124.
[33] SHIMURA K,LI J,FUKUMOTO F.HFT-CNN:learning hierarchical category structure for multi-label short text categorization[C]//2018 Conference on Empirical Methods in Natural Language Processing,2018:811-816.
[34] YANG W,LI J,FUKUMOTO F,et al.MSCNN:a Monomeric-Siamese convolutional neural network for extremely imbalanced multi-label text classification[C]//2020 Conference on Empirical Methods in Natural Language Processing,2020:6716-6722.
[35] HOCHREITER S,SCHMIDHUBER J.Long short-term memory[J].Neural Computation,1997,9(8):1735-1780.
[36] CHUNG J,GULCEHRE C,CHO K H,et al.Empirical evaluation of gated recurrent neural networks on sequence modeling[J].arXiv:1412.3555,2014.
[37] SUTSKEVER I,VINYALS O,LE Q V.Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems,2014:3104-3112.
[38] NAM J,MENCÍA E L,KIM H J,et al.Maximizing subset accuracy with recurrent neural networks in multi-label classification[C]//Advances in Neural Information Processing Systems,2017:5413-5423.
[39] CHEN G,YE D,XING Z,et al.Ensemble application of convolutional and recurrent neural networks for multi-label text categorization[C]//2017 International Joint Conference on Neural Networks,2017:2377-2383.
[40] YANG P,SUN X,LI W,et al.SGM:sequence generation model for multi-label classification[J].arXiv:1806.04822,2018.
[41] YANG P,MA S,ZHANG Y,et al.A deep reinforced sequence-to-set model for multi-label text classification[J].arXiv:1809.03118,2018.
[42] QIN K,LI C,PAVLU V,et al.Adapting RNN sequence prediction model to multi-label set prediction[J].arXiv:1904.05829,2019.
[43] LIN J,SU Q,YANG P,et al.Semantic-unit-based dilated convolution for multi-label text classification[J].arXiv:1808.08561,2018.
[44] 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,2019:5820-5830.
[45] YANG Z,LIU G.Hierarchical sequence-to-sequence model for multi-label text classification[J].IEEE Access,2019,7:153012-153020.
[46] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//Advances in Neural Information Processing Systems,2017:5998-6008.
[47] YARULLIN R,SERDYUKOV P.BERT for sequence-to-sequence multi-label text classification[J].2019.
[48] CHANG W C,YU H F,ZHONG K,et al.Taming pre-trained transformers for extreme multi-label text classification[C]//26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2020:3163-3171.
[49] 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.
[50] QUAN C,REN F.A blog emotion corpus for emotional expression analysis in Chinese[J].Computer Speech & Language,2010,24(4):726-749.
[51] Lewis D D,Yang Y M,Rose T G,et al.RCV1:a new benchmark collection for text categorization research[J].Journal of Machine Learning Research,2004,5:361-397.
[52] MENCIA E L,FÜRNKRANZ J.Efficient pairwise multi-label classification for large-scale problems in the legal domain[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases.Berlin,Heidelberg:Springer,2008:50-65.
[53] MCAULEY J,LESKOVEC J.Hidden factors and hidden topics:understanding rating dimensions with review text[C]//7th ACM Conference on Recommender Systems,2013:165-172.
[54] SCHAPIRE R E,SINGER Y.Improved boosting algorithms using confidence-rated predictions[J].Machine Learning,1999,37(3):297-336.
[55] SCHÜTZE H,MANNING C D,RAGHAVAN P.Introduction to information retrieval[M].Cambridge:Cambridge University Press,2008.
[56] LIU W,SHEN X,WANG H,et al.The emerging trends of multi-label learning[J].arXiv:2011.11197,2020.
[57] WU T,HUANG Q,LIU Z,et al.Distribution-balanced loss for multi-label classification in long-tailed datasets[C]//European Conference on Computer Vision.Cham:Springer,2020:162-178.