doi:10.3969/j.issn.1001-4616.2019.04.009

基于编码解码器与深度主题特征抽取的多标签文本分类

陈文实,刘心惠,鲁明羽

(大连海事大学信息科学技术学院,辽宁 大连 116026)

[**摘要**] 本文提出了一种基于编码解码器与深度主题特征的模型,实现了多标签文本分类. 针对传统多标签文本分类方法的特征语义缺失的问题,采用一种长短时记忆(long short-term memory,LSTM)网络提取文本的局部特征与主题模型(latent dirichlet allocation,LDA)提取文本的全局特征的深度主题特征提取模型(deep topic feature extraction model,DTFEM),得到具有文本深层语义特征的语义编码向量,并将该编码向量作为解码器网络的输入. 解码器网络将多标签文本分类的任务看作序列生成的过程,解决了多标签文本分类的标签相关性的问题,并加入 attention 机制,计算注意力分布概率,突出关键输入对输出的作用,改进了由于输入过长导致的语义缺失问题,最终实现多标签文本分类. 实验结果表明,该模型能够获得比传统的多标签文本分类系统更优的结果. 另外,实验证明使用深度主题特征的方法可以提高多标签文本分类的性能.

[**关键词**] 多标签文本分类,深度主题特征,标签相关性,编码解码器,attention 机制
[**中图分类号**]TP311　[**文献标志码**]A　[**文章编号**]1001-4616(2019)04-0061-08

Multi-label Text Classification Based on Seq2Seq Model and Deep Topic Feature Extraction

Chen Wenshi,Liu Xinhui,Lu Mingyu

(School of Information Science and Technology,Dalian Maritime University,Dalian 116026,China)

Abstract:In this paper,a model based on seq2seq model and deep topic feature extraction is proposed to realize multi-label text classification. Aiming at the problem of feature semantics loss in traditional multi-label text classification method,a model is proposed to extract the local features of texts by using the Long Short-term Memory(LSTM)network and extract the global features of texts by using topic model(Latent Dirichlet Allocation,LDA)named Deep Topic Feature Extraction Model(DTFEM),and then obtain the semantic coding vector with deep semantic feature,and the vector is used as the input of the decoder network. The decoder network regards the task of multi-label text classification as the process of sequence generation,solves the problem of label correlation of multi-label text classification,and adds the attention mechanism to calculate the probability distribution of attention,highlights the effect of key input on the output,improves the semantic missing problem due to excessive input,and realizes the final multi-label text classification. The experimental results show that the model can obtain better results than the traditional multi-label text classification system. In addition,the experiments have shown that the use of deep topic features can improve the performance of multi-label text classification.

Key words:multi-label text classification,deep topic feature extraction,label correlation,seq2seq,attention mechanism

文本分类是指按照预先定义的标签类别,通过一定的学习机制,在对带有类别标签的训练文本进行学习的基础上,给未知文本分配一个或多个类别标签的过程,它是自然语言处理中的一个重要的分支领域. 传统的文本分类主要是单标签的分类,每个样本只能用一个类别来表示,分类的粒度较粗,然而随着电子文档信息量的急剧增长,文本内容的多样化,确定文本的单标签分类技术已经难以满足人们对文本分

收稿日期:2019-06-25.
基金项目:国家自然科学基金(61073133).
通讯联系人:鲁明羽,博士,教授,博士生导师,研究方向:自然语言处理. E-mail:lumingyu@dlmu.edu.cn

类的需求,在互联网数据爆炸增长的今天,数据的内容更加丰富、类别的粒度也越来越细,一个样本可能和多个类别标签相关联,类别标签的数目也越来越多. 多标签文本分类技术的相关研究已经成为近年来国际学术界的研究热点. 多标签文本分类技术已经被广泛应用于解决现实世界的问题,如信息检索[1]、情感计算[2]、情绪分析[3]、电子邮件垃圾邮件检测[4]等.

文本的特征提取是运用学习文本的关键特征对文本进行向量化表示的过程,对信息检索、机器翻译、文本分类等自然语言处理任务具有重要的意义. 传统的文本特征都是人工设计的特征,特征设计和提取的过程也非常耗费人力,同时还可能带来积累误差和噪音,其大都依据简单的数理统计思想,并且认为特征词之间是相互独立的,忽略了文本的结构和语义对于特征词选取的重要性,导致了语义因素无法在提取特征词的过程中发挥作用. 2003 年 Blei 等人[5]提出了对文本生成过程进行建模,挖掘文本的隐含语义信息,采用隐含狄利克雷分布(latent Dirichlet allocation,LDA)主题模型,用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文本的主题概率分布,实现了文本数据的特征提取,被广泛应用于推荐系统[6]、文本聚类[7]、话题检测等[8]. 近年来,深度学习作为一个能够通过大量数据自动学习特征的技术框架,在文本处理领域也得到了广泛应用. 从 2013 年 Google 发表的 word2Vec 开始,深度学习被广泛应用于文本处理中的机器翻译、文档摘要、阅读理解、关系抽取、情感分析、文档分类等任务上,并取得了很多重要的进展. 循环神经网络(recurrent neural network,RNN)作为自然语言处理中的文本序列神经网络模型,改变了传统神经网络的连接方式,加入了循环结构,可以表示上文对下文的影响,但针对其存在的梯度消失、梯度爆炸等问题,Hochreiter 等人[9]提出了长短时记忆单元(long-short term memory,LSTM),在 RNN 的基础上增加了一个记忆单元,可以保留或者丢弃历史信息,使信息可以长期保留,取得了比 RNN 更好的效果.

在多标签文本分类任务中,标签与标签之间不是相互独立的,通常具有较强的相关性,随着类别标签数目的增加,输出空间的大小会呈现指数增长,严重地影响着多标签文本分类的性能. 2014 年 Sutskever 等人[10]提出序列到序列模型(sequence to sequence,Seq2seq),该模型使用了两个循环神经网络(RNN)分别构建编码器(Encoder)和解码器(Decoder),可以将可变长度的输入序列映射到可变长度的输出序列. Cho 等人将 Encoder-Decoder 模型用于机器翻译[11],Google 将 Encoder-Decoder 模型应用于对话生成[12],除此之外 Encoder-Decoder 模型也被用于自然语言处理(nature language processing,NLP)的其他领域,如文本摘要[13]、多任务学习[14]等. Bahdanau 等人[15]在 Encoder-Decoder 基础上加入了 Attention 机制,有效解决了由于输入语句过长而导致的语义信息丢失问题,该模型的提出使机器翻译的效果显著提升.

综上所述,本文针对多标签文本分类中存在的问题:(1)提出了一种深度主题特征的提取方法,将深度学习模型的有监督的特征提取方法与主题模型 LDA 无监督的特征提取方法相结合,从而使模型能够同时表达文本的全局特征和局部特征,实现了文本不同层次的特征提取;(2)将多标签文本分类任务视为序列到序列问题,将编码解码器网络应用于多标签文本分类任务中,利用标签序列的依赖性来解决类别标签之间的相关性问题;(3)将 Attention 机制引入到编码解码器网络中,通过考虑不同文本内容对类别标签的贡献差异,对输入序列赋予不同的权重,改进了由于输入序列过长导致的语义缺失问题;(4)利用 3 个公共可用的数据集,即 Reuters-21578、AAPD 和 IMDB,将我们的方法与几个现有模型进行比较.

本文的第 1 节对多标签文本分类的问题进行了定义并总结了目前的多标签文本分类的方法,第 2 节对主要模型进行了详细的描述,包括长短时记忆网络、主题模型、序列到序列模型以及本文提出的多标签文本分类模型,第 3 节将通过实验评估本文方法的有效性,展示了实验数据和训练参数设置,并进行了模型对比分析,第 4 节梳理总结本文的研究工作.

1 问题定义

假设 $X=\mathbf{R}^d$ 表示在实数域 $\mathbf{R}$ 上有 D 维的样本空间,$Y=\{y_1,y_2,\cdots,y_q\}$ 表示包含 q 个标签空间的类别标签集合,多标签分类是指通过训练数据集 $D=\{(x_i,Y_i)\mid 1\leqslant i\leqslant m\}$ 得到一个映射函数 $f:X\to 2^Y$,其中 $x_i\in X$ 即 x_i 是输入空间 X 中的一个训练数据,$Y_i\in Y$ 即 Y_i 是样本 x_i 的标签集合[16]. 输入一个待分类样本 $x\in X$ 时,通过映射函数 f 得到 x 的预测标签集合 $P_x\subset Y$,使 P_x 与样本 x 的真实标签集合 Y_x 最接近. 单标签分类问题是多标签分类问题的特例,当 $|Y_i|=q=1$ 时,多标签分类转化为单标签分类.

多标签分类算法分为两种,一种方法将多标签分类问题转化为传统的分类问题即问题转换法,包括

Binary Relevance 算法[17]将每个标签作为单独的类别进行分类;LabelPowerset 算法[18]把问题转化为一个多类问题,一个多类分类器在训练数据中将所有不同的类别标记子集转化为不同的类别进行训练;Classifier Chains 算法[19]将多标签分类问题转化为二分类子问题,每个标签都对应于一个子问题. 另一种方法是调整现有的算法使其适应于多标签分类即算法适应法,包括 ML-KNN 算法[20]、RankSVM 算法[21]等,这两种算法分别对 k 近邻算法和支持向量机算法进行改进,使其适应于多标签分类算法. 多标签分类中的标签与标签之间不是相互独立的,而是具有一定的关系,在分类时如果考虑标签相关性,那么会使分类准确率得到提升,上述的 Classifier Chains 算法便是利用标签之间的相关性进行分类.

本文将多标签文本分类任务视为序列到序列问题,将 seq2seq 模型(Encoder-Decoder)用于多标签文本分类任务中,模型由编码器网络和解码器网络两部分组成,在编码器(Encoder)网络中,利用 LSTM 神经网络提取文本的局部特征表示 $\boldsymbol{L}_t=q(\{h_1,h_2,\cdots,h_n\})$,借助主题模型提取文本的全局特征表示 $D=[d_1,d_2,\cdots,d_n]$,并加以融合,得到文本的特征输出 $\boldsymbol{M}=[\boldsymbol{L}_t;D]$,融合的方法采用合并拼接的方式进行,从而实现了不同层次的文本特征提取;在解码器(Decoder)网络中采用 LSTM 神经网络来处理标签序列的依赖性,解决类别标签之间的相关性问题. Attention 机制最早被应用在图像处理领域,用于图片主题生成[22]、字符识别[23]等. Bahdanau 等人在 Encoder-Decoder 基础上加入了 attention 机制,有效地解决了由于输入语句过长而导致的语义信息丢失问题,该模型的提出使机器翻译的效果显著提升. 本文在 seq2seq 模型的 Decoder 阶段也加入了 attention 机制,考虑了不同文本内容对类别标签的贡献差异,对输入序列赋予不同的权重,改进了由于输入序列过长导致的语义缺失问题.

2 模型框架

本文提出的基于序列模型的多标签文本分类模型如图 1 所示,由 seq2seq 模型和 attention 机制组成. 该模型包括 3 个部分:编码器(Encoder)网络、解码器(Decoder)网络和连接 Encoder-Decoder 的中间语义向量 $\boldsymbol{M}$. 编码器网络和解码器网络分别对应输入序列和输出序列的两个神经网络,输入序列通过 Encoder 网络及其主题模型对其编码,形成一个中间语义向量 $\boldsymbol{M}$,把这个向量传递给 Decoder 网络,Decoder 阶段通过 LSTM 神经网络捕获类别标签之间的相关性,加入 attention 机制的 seq2seq 模型不仅可以考虑类别标签之间的相关性,还可以在预测不同类别标签时,通过计算注意力概率分布,突出关键输入对输出的作用,自动选择贡献大的单词赋予较大的权重,解码输出一系列类别标签.

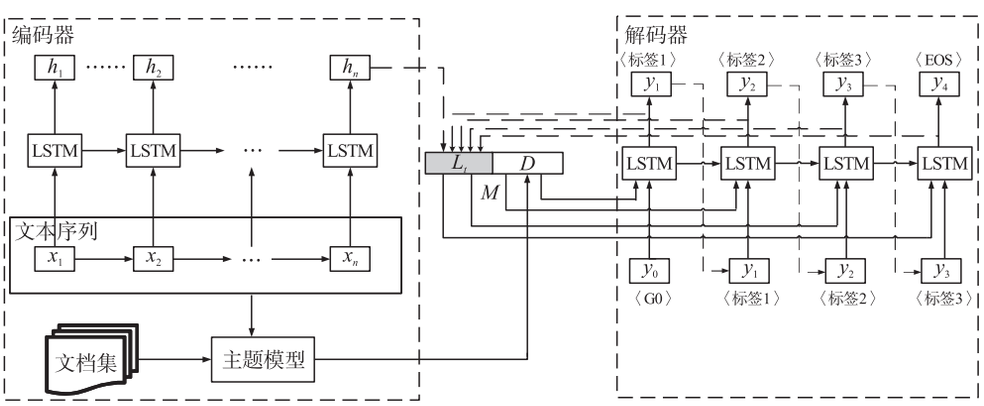

图 1 多标签文本分类的研究框架

Fig. 1 Multi-label text classification research framework

2.1 Encoder 模型

Encoder:将输入序列编码成一个中间语义向量 $\boldsymbol{M}$,该向量由两部分组成,一部分为通过 LSTM 神经网络得到的文本的局部特征表示 L_t,另一部分为利用主题模型得到的文本的全局特征表示 $\boldsymbol{D}$,即文本的主题向量.

(1)长短时记忆网络(LSTM)

LSTM 可以解决 RNN 梯度消失等问题,标准的 LSTM 网络包含 3 层结构:输入层、隐藏层、输出层,其中隐藏层包含了输入门(Input Gate)、输出门(Output Gate)、遗忘门(Forget Gate)、记忆单元(Memory Cell),通过 3 种门结构来确定保留或者遗忘记忆单元中的信息.

$$i_t=\sigma(\boldsymbol{W}_i\cdot|h_{t-1},x_t|+b_i),\tag{1}$$

$$f_t=\sigma(\boldsymbol{W}_f\cdot|h_{t-1},x_t|+b_f),\tag{2}$$

$$o_t=\sigma(\boldsymbol{W}_o\cdot|h_{t-1},x_t|+b_o),\tag{3}$$

$$\tilde{C}_t=\tanh(\boldsymbol{W}_c\cdot|h_{t-1},x_t|+b_c),\tag{4}$$

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t,\tag{5}$$

$$h_t=o_t*\tanh(C_t),\tag{6}$$

式中,i,f,o 分别表示输入、遗忘、输出门,x,h,c 表示输入层、隐藏层、记忆单元,$\boldsymbol{W},b$ 表示权重矩阵、偏置.

我们将输入序列表示为每篇文档的词向量序列 $\boldsymbol{x}=(x_1,\cdots,x_n)$,则编码过程为:

$$h_t=f(x_t,h_{t-1}),\tag{7}$$

$$\boldsymbol{L}=q(\{h_1,h_2,\cdots,h_n\}),\tag{8}$$

式中,h_t 为 t 时刻的隐层单元的状态,f 为神经网络单元,通常采用 RNN(recurrent neural network)的变体 LSTM(long short-term memory)或 GRU(gated recurrent unit)替代 RNN,来避免 RNN 梯度消失等问题,本文使用 LSTM 神经网络,q 为线性函数. 输入句子的编码向量作为编码器的最后一个状态向量,得到了文本的局部特征表式向量 $\boldsymbol{L}_t$

$$\boldsymbol{L}_t=q(\{h_1,h_2,\cdots,h_n\}).\tag{9}$$

(2)主题模型

Blei 等人在 2003 年提出了隐含狄利克雷分布(latent Dirichlet allocation,LDA)主题模型,用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文本的主题概率分布. 主题模型具有良好的特征降维功能,可以把高维度、稀疏的文本-词分布转换为低维度的文本-主题概率分布. LDA 主题模型由文档、主题、单词 3 个层次的概率分布组成,在文档-单词特征层次中加入主题,捕获文本的全局底层语义信息,实现文本在主题潜在空间上特征的良好表达. LDA 主题模型如图 2 所示,α 为文档-主题分布的先验参数,β 为主题-单词分布的先验参数,$W_{m,n}$ 表示第 m 篇文档中的第 n 个单词,$Z_{m,n}$ 表示第 m 篇文档中第 n 个单词所对应的主题,通过概率推断方法学习得到每一个主题对应的词项分布 φ_k 以及每一篇文档对应的主题概率分布 θ_m.

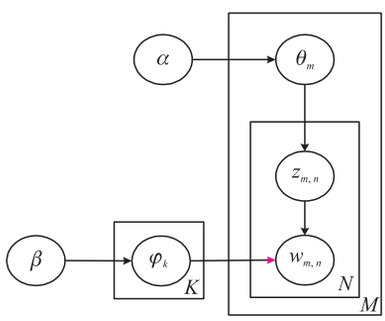

图 2　LDA 主题模型图

Fig. 2　The LDA topic model

模型利用语料库中的文本进行训练,获得 LDA 主题模型,构建当前文本的全局特征表示 $\boldsymbol{D}$,为了实现与 LSTM 局部特征具有同等的作用,实验中设置了 LDA 全局特征与 LSTM 局部特征具有相同的维度 128,并将其与 LSTM 网络获得的文本的局部特征表示向量 $\boldsymbol{L}_t$ 加以融合,得到编码器中间语义向量 $\boldsymbol{M}=[\boldsymbol{L}_t;\boldsymbol{D}]$.

2.2　Decoder 模型

Decoder:假设输出序列是类别标签 $y=(y_1,\cdots,y_t)$,时刻 t 的输出由之前的标签输出和语义向量 $\boldsymbol{M}=[\boldsymbol{L}_t;\boldsymbol{D}]$ 决定,对向量 $\boldsymbol{M}$ 计算类别标签的概率值

$$p(y)=\prod_{t=1}^{T}p(y_t|\{y_1,\cdots,y_{t-1}\},\boldsymbol{M}),\tag{10}$$

$$p(y_t|\{y_1,\cdots,y_{t-1}\},\boldsymbol{M})=g(y_{t-1},s_t,\boldsymbol{M}),\tag{11}$$

$$s_t=f(s_{t-1},y_{t-1},s_t,\boldsymbol{M}),\tag{12}$$

式中,g 表示 LSTM 神经网络单元,s_t 表示时刻 t 的 LSTM 隐层单元的状态.

Attention 机制:

结合文本分类任务的特性,在 Decoder 阶段加入 attention 机制,通过计算注意力概率分布,可以突出关键输入对输出的作用. 将 Encoder 部分的每一个隐藏层都进行加权处理,从而 Decoder 拥有多个语义信息来得到输入的语义表示.

在 Decoder 阶段,

$$p(y_t|\{y_1,\cdots,y_{t-1}\},[\boldsymbol{L}_t;\boldsymbol{D}])=g(y_{t-1},s_t,[\boldsymbol{L}_t;\boldsymbol{D}]),\tag{13}$$

$$s_t = f(s_{t-1}, y_{t-1}, s_t, [\boldsymbol{L}_t; \boldsymbol{D}]), \tag{14}$$

式中,s_t 表示时刻 t 时 Decoder 的隐藏层单元状态,与基本的 seq2seq 模型不同的是,语义编码向量 $\boldsymbol{L}_t$ 不仅仅是 Encoder 阶段最后一步的输出,而是把 Encoder 阶段所有的隐藏层($h_1, h_2, \cdots, h_n$)加权求和,可以得到新状态的输入. 时刻 t 编码向量 $\boldsymbol{L}_t$ 可以表示为:

$$\boldsymbol{L}_t = \sum_{i=1}^{n} \alpha_{ti} h_i, \tag{15}$$

式中,h_i 表示时刻 t 的 Decoder 隐藏层状态,α_{ti} 表示权重,对于每一个隐藏层状态 h_i,权重 α_{ti} 的计算公式为:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{n} \exp(e_{tk})}, \tag{16}$$

$$e_{ti} = a(s_{t-1}, h_i). \tag{17}$$

e_{ti} 通过 Decoder 由时刻 $t-1$ 的隐藏层单元状态和时刻 i 的隐藏层单元状态 h_i 得到.

本文的序列模型由一个 Encoder 和一个具有 attention 机制的 Decoder 组成,Encoder 部分如图 1 左侧所示,输入层将待分类文本统一为同等长度序列后传递给 Embedding 层,Embedding 层采用词嵌入方法将每个单词转换为一个向量,将其输入到 LSTM 神经网络中,得到文本的局部特征向量并与主题模型获得的全局特征向量融合,获得文本的编码向量. 模型预测不同的类别标签时,不同的输入单词会做出不同的贡献,attention 机制通过对输入文本中的不同的单词赋予不同的权重产生中间语义向量,对输出类别标签贡献大的单词赋予较大的权重,否则赋予较小的权重. 权重计算如公式 16,从而当模型预测不同类别标签时,attention 机制可以考虑输入文本的单词贡献差异.

本文模型的 Decoder 部分如图 1 右侧所示,Decoder 的输入序列把<GO>作为类别标签序列的头部,输出序列把<EOS>作为预测类别标签的尾部. 通过 LSTM 处理标签序列的依赖性来考虑类别标签之间的相关性,顺序生成标签,得到类别标签的输出.

3 实验

3.1 实验数据

Reuters-21578:分布在 22 个文件中,从 reut2-000.sgm 到 reut2-020.sgm,每个文件包含 1 000 个文档,reut2-021.sgm 包含 578 个文档,总计 21 578 个文档,共分为有 672 个类别,本实验选取文章数最多的 20 个类别标签.

AAPD:从 arXiv 网站上关于计算机科学领域 55 840 篇论文的摘要和主题,每篇论文可能涉及多个科目,共有 54 个类别,可以根据摘要内容预测学术论文对应的类别.

IMDB:包含有 117 352 篇电影简介,其中英文的电影简介包括有 117 190 篇,每部电影可与 27 个类别相关.

3.2 评价指标

(1)微平均(Micro-averaging):对数据集中每个样本不区分类别进行统计建立混淆矩阵,然后计算相应指标.

$$\text{Micro\_P} = \frac{\sum_{i=1}^{|D|} \text{TP}_i}{\sum_{i=1}^{|D|} \text{TP}_i + \sum_{i=1}^{|D|} \text{FP}_i} \times 100\%, \tag{18}$$

$$\text{Micro\_R} = \frac{\sum_{i=1}^{|D|} \text{TP}_i}{\sum_{i=1}^{|D|} \text{TP}_i + \sum_{i=1}^{|D|} \text{FN}_i} \times 100\%, \tag{19}$$

$$\text{Micro\_F} = \frac{2 \times \text{Micro\_P} \times \text{Micro\_R}}{\text{Micro\_P} + \text{Micro\_R}} \times 100\%. \tag{20}$$

(2)汉明损失(Hamming Loss,HL):是样本中真实结果与预测结果间的异或,表示样本标签对被错分类的次数. 该值越小,模型的性能越好.

$$\mathrm{HL}=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{xor(x_i,y_i)}{|L|}. \tag{21}$$

(3)子集准确率(Subset Accuracy):计算所有样本中完全预测正确的比例,完全预测正确的样本要求预测得到的标签集合和正确的标签集合完全相同.

$$\mathrm{SA}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}\times 100\%. \tag{22}$$

式中,$|D|$表示数据集中的样本数;真正例(true positive,TP)表示真实类别为正例,预测类别为正例;假正例(false positive,FP)表示真实类别为负例,预测类别为正例;假负例(false negative,FN)表示真实类别为正例,预测类别为负例;真负例(true negative,TN)表示真实类别为负例,预测类别为负例;$|L|$表示标签总数;x_i 和 y_i 分别表示预测的标签结果和真实标签,xor 表示异或运算.

3.3 模型超参

由于原始的文本中存在不规范的字符(HTML 标记、xml 标记等)、多余的标点符号、无意义的停用词等,因此需要去掉文本中的这些噪声数据,形成纯英文文本语料. 每个文本中包含单词数长短不一,单词数过短会导致无法准确判断文本所属类别,单词数过长则会导致空间浪费,因此去除单词数少于 20 个单词的文本,选取 70%文本的最大长度作为输入长度,少于该单词数的文本用<pad>进行填充,多于该长度则将文本截断. 主题模型中的主题数目设置为 128. 为了评估模型,我们将数据集划分为两个部分,2/3 的数据用于训练,1/3 的数据用于测试,采用微平均精确率(Micro_P,记为 MiP)、微平均召回率(Micro_R,记为 MiR)、微平均综合指标 F1 值(Micro_F,记为 MiF)、汉明损失(hamming loss,HL)、子集准确率(subset accuracy,记为 SA)5 个指标来比较不同分类方法的分类效果.

本文模型均使用 ADAM(adaptive moment estimation)作为优化器函数来优化神经网络,损失函数使用 binary_crossentropy,为防止神经网络产生过拟合现象,我们在编码器中采用 dropout,值为 0.3,epoch 设置为 10.Encoder 和 Decoder 中的 LSTM 层数为 2,Encoder 隐藏层单元数目为 128,Decoder 隐藏层单元数目为 128.

Reuters-21578:最大文本输入长度为 915,70%文本的最大长度为 83,文本的最大标签数 8,70%文本的标签数 2. 训练集文本数为 8 870,测试集文本数为 2 218,批处理大小为 16.

AAPD:最大文本输入长度为 336,70%文本的最大长度为 111,文本的最大标签数 8,70%文本的标签数 3. 训练集文本数为 29 387,测试集文本数为 7 347.

IMDB:最大文本输入长度为 1106,70%文本的最大长度为 65,文本的最大标签数 12,70%文本的标签数 3. 训练集文本数为 62 304,测试集文本数为 15 577.

3.4 实验结果与分析

本文实现了基于序列模型的多标签文本分类算法,将机器学习方法 Binary Relevance(BR)、Classifier Chains(CC)、MLkNN(ML-KNN)、深度学习的基于序列模型 Seq2Seq 以及未加入 Attention 机制的 Seq2Seq+LDA(LDASeq2Seq)与本文提出的模型基于序列模型的 Seq2Seq+LDA+attention(LDASeq2Seq_A)模型在 MiP、MiR、MiF、HL、SA 结果进行对比,各数据集实验结果如表 1~3 所示,“+”表示该值越大,模型性能越好,“-”表示该值越小,模型性能越好.

表 1 Reuters-21578 数据集实验结果

Table 1 Experiment results of dataset Reuters-21578

Models	BR	CC	ML-KNN	Seq2Seq	LDASeq2Seq	LDASeq2Seq_A
iP(+)	73.44	79.24	77.55	81.52	85.54	85.87
MiR(+)	65.28	53.85	50	79.24	79.69	80.07
MiF(+)	69.12	64.12	60.8	80.37	82.51	82.87
HL(-)	0.254 5	0.284 8	0.296 9	0.031 1	0.027 2	0.026 6
SA(+)	24.24	24.24	27.27	76.04	78.31	78.81

由表 1 可知,在 Reuters-21578 数据集上,本文提出的模型在各项指标上均优于机器学习模型以及深度学习模型 Seq2Seq,其中 Binary Relevance、MLkNN 未考虑标签相关性,Classifier Chains 及 Seq2Seq 模型

考虑了标签相关性,各项指标均有提升,F1 值较 Classifier Chains 提升了 18.75%,SA 提升了 54.57%;较 Seq2Seq 模型提升了 2.5%,SA 提升了 2.77%.

表 2 AAPD 数据集实验结果

Table 2 Experiment results of dataset AAPD

Models	BR	CC	MLKNN	Seq2Seq	LDASeq2Seq	LDASeq2Seq_A
MiP(+)	65.28	68.83	58.11	73.14	73.66	73.27
MiR(+)	63.51	69.74	60.56	68.48	68.70	68.59
MiF(+)	64.38	69.28	59.31	70.72	71.08	70.83
HL(-)	0.351 2	0.284 8	0.357 6	0.024 7	0.024 4	0.024 6
SA(+)	27.27	18.18	9.09	43.13	44.05	43.85

由表 2 可知,在 AAPD 数据集上,本文提出的模型在各项指标上均优于机器学习模型以及深度学习模型 Seq2Seq,结果表明本文提出的模型各项指标均有提升,F1 值较 Classifier Chains 提升了 4.44%,SA 提升了 25.67%;较 Seq2Seq 模型提升了 0.13%,SA 提升了 0.72%.

表 3 IMDB 数据集实验结果

Table 3 Experiment results of dataset IMDB

Models	BR	CC	MLKNN	Seq2Seq	LDASeq2Seq	LDASeq2Seq_A
MiP(+)	65.57	76.19	69.35	80.12	80.38	80.30
MiR(+)	50.63	79.01	55.13	76.62	76.54	76.63
MiF(+)	57.14	77.58	61.43	78.35	78.40	78.41
HL(-)	0.363 6	0.224 2	0.327 3	0.032 6	0.032 3	0.032 3
SA(+)	9.09	30.30	15.15	59.67	60.10	60.26

由表 3 可知,在 IMDB 数据集上,本文提出的模型在各项指标上均优于机器学习模型以及深度学习模型 Seq2Seq,各项指标均有提升,F1 值较 Classifier Chains 提升了 4.11%,SA 提升了 29.96%;较 Seq2Seq 模型提升了 0.18%,SA 提升了 0.59%.

由表 1~3 可知,考虑了标签相关性的模型 Classifier Chains、Seq2Seq、LDASeq2Seq 以及 LDASeq2Seq_A 各项指标大都优于未考虑标签相关性的算法,且 LDASeq2Seq、LDASeq2Seq_A 模型的实验结果优于传统的机器学习算法以及深度学习的 Seq2Seq 模型,在各项指标上都有提升. 因此在进行多标签文本分类研究时,应考虑标签相关性对分类结果的影响. 由于输入序列的不同部分可能对预测不同的类别标签有不同的贡献,加入了 attention 机制的 seq2seq 模型的分类性能大都优于未加入 attention 机制的 seq2seq 模型,attention 机制在多标签文本分类中展现出了一定的效果.

4 结语

本文的主要工作是将多标签分类问题看作是序列到序列的问题,采用 Seq2Seq 模型进行建模,编码器(Encoder)网络利用 LSTM 神经网络提取文本的局部特征,借助主题模型提取文本的全局特征,并加以融合,实现了不同层次的文本特征提取;解码器网络中采用 LSTM 顺序生成标签,以解决标签之间的相关性问题. 加入 attention 机制的 seq2seq 模型自动找到输入文本和预测类别标签之间的对应关系,并对贡献大的单词赋予较大的权重. 实验结果表明,本文提出的模型在精确率、召回率、F1 值、汉明损失和子集准确率等指标上都取得了较好的结果,相比于基线模型,效果有了明显的提升.

[参考文献]

[1] GONG Y,KE Q,ISARD M,et al. A multi-view embedding space for modeling internet images,tags and their semantics[J]. International journal of computer vision,2014,106(2):210-233.

[2] PORIA S,CAMBRIA E,BAJPAI R,et al. A review of affective computing:from unimodal analysis to multimodal fusion[J]. Information fusion,2017,37:98-125.

[3] CAMBRIA E. Affective computing and sentiment analysis[J]. IEEE intelligent systems,2016,31(2):102-107.

[4] CARRERAS X, MARQUEZ L. Boosting trees for anti-spam email filtering[C]//Proceeding of the 4th International

Conference on Recent Advances in Natural Language Processing. Tzigov Chark,Bulgaria,2001.
[5] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research,2003,3:993-1022.
[6] 黄泽明. 基于主题模型的学术论文推荐系统研究[D]. 大连:大连海事大学,2013.
[7] 王李冬,魏宝刚,袁杰. 基于概率主题模型的文档聚类[J]. 电子学报,2012,40(11):2346-2350.
[8] 邱云飞,郭弥纶,邵良杉. 基于主题树的微博突发话题检测[J]. 计算机应用,2014,34(8):2332-2335.
[9] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural computation,1997,9(8):1735-1780.
[10] SUTSKEVER I,VINYALS O,LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing systems. Montreal,Canada,2014.
[11] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha,Qatar,2014.
[12] VINYALS O,LE Q. A neural conversational model[J]. arXiv preprint arXiv:1506.05869,2015.
[13] RUSH A M,CHOPRA S,WESTON J. A neural attention model for abstractive sentence summarization[J]. arXiv preprint arXiv:1509.00685,2015.
[14] LUONG M T,LE Q V,SUTSKEVER I,et al. Multi-task sequence to sequence learning[C]//Proceedings of ICLR. San Juan, Puerto Rico,2015.
[15] BAHDANAU D,CHO K,BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473,2014.
[16] ZHANG M L,ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE transactions on knowledge and data engineering, 2014,26(8):1819-1837.
[17] LUACES R ó,DíEZ P J,BARRANQUERO T J,et al. Binary relevance efficacy for multilabel classification[J]. Progress in artificial intelligence,2012,1(4):303-313.
[18] CHERMAN E A,MONARD M C,METZ J. Multi-label problem transformation methods:a case study[J]. CLEI electronic journal,2011,14(1):4-4.
[19] READ J,PFAHRINGER B,HOLMES G,et al. Classifier chains for multi-label classification[J]. Machine learning,2011, 85(3):333.
[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] ELISSEEFF A,WESTON J. A kernel method for multi-labelled classification[C]//Proceedings of the 14th International Conference on Neural Information Proceesing Systems:Natural and Synthetic. Vancouver,British Columbia,Canada,2001.
[22] XU K,BA J,KIROS R,et al. Show,attend and tell:Neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille,France,2015.
[23] LEE C Y,OSINDERO S. Recursive recurrent nets with attention modeling for ocr in the wild[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las vegas,2016.

[责任编辑:顾晓天]